

Enhanced Synchronous Packet Switching for IP Packets

Peter HOMAN¹ and Janez BESTER¹, Associate Members

SUMMARY Fast packet switches for variable-size packets have become an everyday necessity with the rapid growth in the volume of Internet traffic. Such switches can be designed in two different ways, either by segmenting packets into smaller fixed-size cells and designing packet switches for such cells, or by designing generic packet switches for variable-size packets, where packet segmentation and reassembly can be omitted. The second option is investigated in this paper. The synchronous operation mode with time-limited bulk service is selected. The switching fabric is assumed to be internally non-blocking and provided with input queues. A previous maximum switch throughput analysis has been done under the assumption that the length of the time slot is fixed set to its minimum allowed value (T_{min}). In this work, a so-called time-slot *stretch factor* (SF) is introduced. The actual time-slot length is determined by multiplying T_{min} with the SF , where $SF \geq 1$. Next, a so-called Internet traffic-source model is proposed based on findings from real IP traffic measurements. The performance implications of the proposed time-slot length modification are analyzed by discrete-event computer simulation. The maximum switch throughput is increased by increasing the SF value, e.g. for uniform packet size distribution and $SF = 10$, the maximum switch throughput is increased from 75% to 97%. The influence of the traffic-source characteristics on the maximum switch throughput is decreased when SF value is increased. In order to prevent any possible throughput degradations, it is advisable to use integer SF values. Packet delay analysis has revealed that by increasing the SF value, the mean packet delay is also increased. Nevertheless, it is shown that the number of switch input and output ports is the most important factor to be considered when packet delay is at stake. Service class differentiation inside investigated packet switch is possible and is not affected by the increasing SF value. Such a packet switch is suitable for implementation in wide area networks, due to high transmission speeds and the small number of switch ports.

key words: internet protocol, synchronous packet switching, time-slot stretch factor

1. Introduction

In recent years, the increased popularity of multimedia applications, combined with the rapidly growing number of users, has resulted in dramatically increased traffic volumes [1-3]. IP packets through the worldwide Internet network carry the majority of this traffic.

The throughput of all network elements has to be increased in order to cope with such traffic volumes. Ever since the introduction of optical transmission lines, the throughput limitations have been set by the limited capabilities of packet switches. It is increasingly clear that the Internet network is *de facto* becoming the universal, worldwide spread and widely accepted packet network [1-3]. Packet switch developers are thus faced with a challenge to design fast packet switches for variable-size packets.

We are interested in studying the performance of packet switching architecture for variable-size packets whose specific characteristics are connectionless operation; each packet carries its own full user address. A packet switch can be thought of as operating according to an asynchronous or synchronous mode. In the former case,

the switch changes its inlets/outlets connection pattern at arbitrary time instants, based on the outlet addresses requested by the packets arriving at the switch inlets. In the latter case, the packet switch keeps its inlets/outlets connection pattern for a time period called slot, lasting T time units. In general, the chance of output conflict is always higher in an asynchronous (continuous-time) operation mode than in a synchronous (discrete-time) operation mode. Generally speaking, a discrete-time packet switch provides better performance than a continuous-time packet switch with the same resources, because contention (which causes output conflict) is less random [4].

IP packet switches can be designed in two different ways. The first is by segmenting variable-size packets into appropriate number of smaller fixed-size packets, usually called cells. Packet switches can then be designed for transfer of fixed-size packets. Such switches are most suitable for synchronous operation mode [5]. Unfortunately, fast segmentation and reassembly capabilities have proved troublesome for vendors to implement [6]. To alleviate this problem, packets are usually segmented only once at the network segment entry point, then transferred across multiple switches, cell by cell, and finally reassembled at the network segment exit point. Two examples of such networking concepts are multiprotocol label switching [6] and A/I Net [7].

The second way to design IP packet switches is by designing generic packet switches for variable-size packets. No packet segmentation and reassembly is required in this case. Packets are queued and transferred across the switching fabric in one piece. There is, however, a problem that has to be resolved. Since packets have variable lengths, it takes a different amount of time to transfer them across the switching fabric. Such switches are thus most suitable for asynchronous operation mode; no packet length constraints hold in the case of asynchronous operations. An example of such a packet switch is described in [8]. In this paper, the generic packet switches for variable-size packets are investigated.

Switching fabric can be designed in several ways. Different switching architectures can adopt either a non-blocking network with input and/or output queuing, or a blocking network with various queuing options. We will restrict ourselves to a non-blocking switching architecture, most often implemented by a crossbar switching network. That is, we assume that the switch fabric is capable of transmitting any set of packets offered at its inputs without internal blocking, when the destination ports of the packets are all distinct. Output queuing is impractical for switches with high line rates and/or with a large number of ports, since the fabric and memory of an $N \times N$ switch must run

¹ The authors are with the Faculty of Electrical Engineering, University of Ljubljana, Trzaska 25, SI-1000 Ljubljana, Slovenia.

N times as fast as the line rate. Unfortunately, at high line rates, memories with sufficient bandwidth are simply not available. On the other hand, the fabric and the memory of an input-queued switch need only run as fast as the line rate. This makes input queuing very appealing for switches with fast line rates or with a large number of ports. For this reason, the highest-performance switches and routers use input-queued crossbar switches [9].

In this work, we further explore the studies carried out in [10], in which the maximum throughput analysis of a datagram switch for broadband networks was performed. We assume a synchronous connectionless switching fabric that is internally non-blocking and is provided with input queues to cope with the conflicts between packets addressing the same outgoing switch termination. No packet segmentation is accomplished. Due to a combination of synchronous and connectionless operation, at least one queued packet must be transferred across the switching fabric during each time slot T . In order to meet this requirement, a so-called time-limited bulk service was introduced. A synchronous packet switch with embedded time-limited bulk service will be referred to as a synchronous bulk (SB) packet switch.

We are interested in packet lengths on the order of 1000 bytes and more, so as to carry in a packet the user information block generated according to, e.g., Ethernet and TCP/IP standards. The duration of the time slot cannot be simply defined as these packets have various lengths. The idea of the time-limited bulk service is, that the largest possible packet determines the length of the time slot. If C is the switching fabric's transmission bit rate, and B is the size of the largest possible packet, then the time-slot length is set to be B/C time units. It is obvious that in the case of smaller packets, more than one packet can be transferred during each time slot, as shown in Fig. 1. However, it can also be seen that in most cases, the time slot cannot be fully utilized. If the transfer of the next packet would be completed after the current time slot has ended, then this packet must be left for transfer in the next time slot. In [10], the maximum switch throughput analysis of a SB packet switch was derived under the assumption that the time-slot length T is fixed set to the aforementioned B/C seconds. It was shown analytically that the maximum switch throughput could not reach 100%.

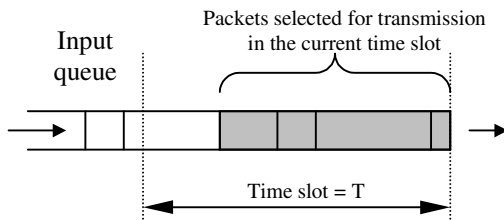


Fig. 1 Time-limited bulk service architecture

We propose to investigate the influence of enlarged time-slot length on overall SB packet switch performance. The aforementioned B/C time units represent the minimum allowed time-slot length T_{min} . This value must not be lowered, but it certainly can be increased. We introduce a so-called time-slot *stretch factor* SF , where $SF \geq 1$. The minimum allowed time-slot length B/C is then multiplied

by the stretch factor SF and so a new, larger time-slot T is derived, as depicted in Eq. (1).

$$T = T_{min} \cdot SF = \frac{B}{C} \cdot SF, \quad SF \geq 1 \quad (1)$$

This paper is organized as follows. In Sect. 2 the methodology used during the switch performance analysis is presented, including the proposed Internet traffic-source model. In Sect. 3 the simulation results are presented. The main implications are summarized in Sect. 4.

2. Methodology

2.1 Packet Switch Architecture

An input-queued packet switch was used in our simulations, as presented in Fig. 2. A non-blocking switching fabric is assumed. There are N input ports and N output ports. All individual queues are organized according to the FIFO (First In First Out) principle.

Choosing the right queue configuration is of vital importance. Our simulations showed that if a single FIFO queue per input port is used, the maximum switch throughput would drop below 10%, due to the well-known HOL (Head Of Line) blocking effect. Usually more than one packet is transferred during a single time slot. The connection through the switching fabric is set to satisfy the first packet in the queue at the beginning of each time slot. It is therefore very likely that any subsequent packet will be destined to a different destination and will thus block the transfer for the rest of the time slot.

Using separate FIFO queues for each output port can resolve this problem. This is a so-called VOQ (Virtual Output Queuing) queue configuration [10, 11] and it is depicted in Fig. 2. Now we have N queues at each input port, which all together amounts to N^2 queues.

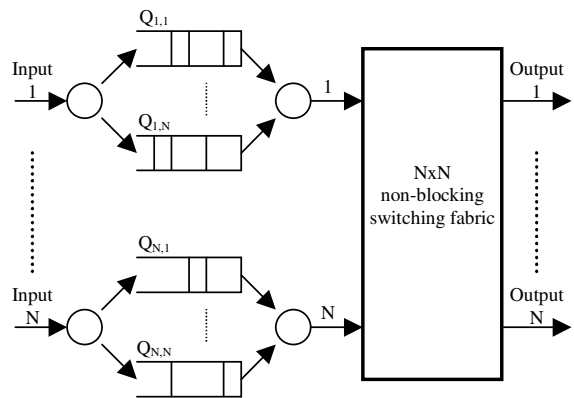


Fig. 2 Packet switch architecture

2.2 Scheduling Mechanisms

At each time slot, at most one queue among the N of the same input port can transmit packets. Moreover, at each time slot, an output port can receive packets from at most one input port. We define as *conflict-free* a set of N queues (chosen among the N^2 available) such that the previous

conditions are met. It is the task of a scheduling mechanism to select a conflict-free set of queues at the beginning of each time slot.

The simplest scheduling mechanism uses a *Rotation* algorithm [10]. In this case, the queues are served one after another, regardless of the current queue occupancy status. For example, the two queues at each input port are served alternately in a switch with two input and two output ports. In the first time slot, the first input is connected to the first output and the second input is connected to the second output. In the second time slot, the first input is connected to the second output and the second input is connected to the first output. In the following time slots the procedure is repeated from the beginning. It can be readily seen that by increasing the number of input and output ports, the time gap between serving each individual input queue is increased considerably. It is also to be noted that the Rotation algorithm assumes a uniform traffic pattern, i.e. all input ports receive the same amount of traffic and the destinations of individual packets are uniformly distributed among all output ports.

Another scheduling mechanism applied in our simulations is called OPF (Oldest Packet First) [11]. As the name itself implies, the packets are served according to the time spent inside the input queues – the higher the delay the higher the priority. A bipartite graph method [11] is usually used to select a conflict-free set of input queues at the beginning of each time slot.

2.3 Traffic Source Models

In our simulations, new packets arrive at inlets following N identical and independent source processes. A uniform reference pattern of arrivals is assumed, i.e. the destinations of packets are uniformly distributed over the set of N outlets.

Packet inter-arrival times are exponentially distributed around mean value, which is calculated accordingly to the specified offered switch load and average packet size. It should be noted however that packet arrival rate is limited by the switch input line rate. Therefore, if input line is already occupied, a newly generated packet is delayed inside a traffic source until the input line becomes available. Limited input line rate can be considered as a simple traffic shaper.

We selected two different methods to determine the size of newly generated packets. The first method, a so-called *truncated-exponential source*, was introduced in [10]. In this case, the length of newly generated packets is defined by the truncated-exponential probability density function, which is given by Eq. (2). The value of parameter a determines the actual shape of probability function. For $a > 0$ the packet's length is a negative exponential; smaller packets are more likely to be generated than larger ones. On the other hand, for $a < 0$, the packet's length is a positive exponential and, therefore, larger packets are more frequent than smaller ones. In this particular case, for $a = 0$, a uniform packet size distribution is used. In the case of Eq. (2), it is assumed that packets can occupy any size between 0 and 1. By applying the appropriate translation function, packet sizes can be distributed between arbitrary minimum and maximum values.

$$p(x) = \begin{cases} \frac{ae^a}{e^a - 1} e^{-ax}, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The second method defines a so-called *Internet source*. We conceived this traffic source based on reported measurements of real IP traffic [12]. The keystone was provided by the fact that there are basically two packet sizes generated by real IP traffic sources. It was established [12] that 80% of all packets can be classified as small and they provide 20% of total Internet traffic. The remaining 20% of IP packets can be classified as large and they provide 80% of total Internet traffic. It is also known that smallest IP packets occupy little less than 50 bytes and largest IP packets are usually limited to approx. 1500 bytes. We can start building our Internet traffic-source model by setting small and large packet sizes to 64 bytes and 1024 bytes, respectively. A binom probability function is used to select the actual size of a new packet. On average, one large packet is generated after four small ones. Packets can occupy more than these two sizes in reality. To enhance our model, a so-called *packet-size smearing factor* has been introduced. Now, the size of a newly generated packet is initially defined to be either small or large by using binom probability function. By doing so, the mean packet size is set to be either 64 bytes or 1024 bytes, respectively. In the next step, the actual packet size is determined by using a uniform distribution around the selected mean value. The boundaries of the uniform distribution smearing function are defined by the aforementioned smearing factor. A smearing factor of 50% has been selected for our Internet source. This means that small packet sizes are uniformly distributed between 32 bytes and 96 bytes and large packet sizes are uniformly distributed between 512 bytes and 1536 bytes. The joint packet-size probability density function is depicted in **Fig. 3**. One can readily see that our Internet traffic source fairly accurately models the previously mentioned characteristics of real Internet traffic.

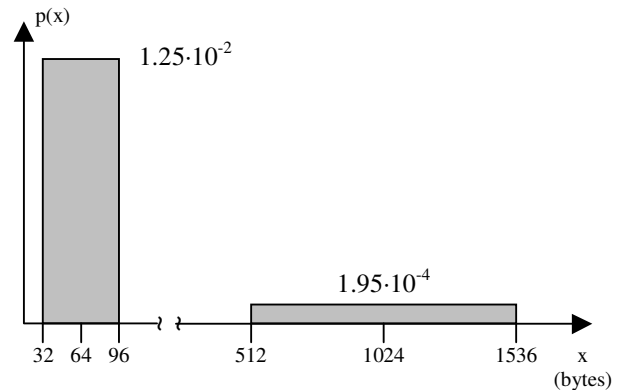


Fig. 3 Packet-size probability density function for the Internet traffic-source model

2.4 Normalized Units

Throughout our simulations, normalized units are used. Normalized transmission speeds of input lines, input ports,

switching fabric, output ports, and output lines are all set to 1.

For truncated-exponential source, the minimum and the maximum packet size were set to 1 and 10, respectively. Therefore, a newly generated packet can occupy any size between 1 and 10.

For our Internet source, the normalized sizes of an average small and an average large packet were set to 1 and 16, respectively. Therefore, considering a smearing factor of 50%, small packets can occupy any size between 0.5 and 1.5, and large packets can occupy any size between 8 and 24.

Queue sizes are also expressed in normalized units. The size of each individual input queue in our simulations was set to 1000 normalized size units. For example, this implies that each queue can concurrently deposit 100 packets with normalized size 10.

A normalized time unit can be readily defined by using normalized transmission speed and normalized packet size. For example, a packet with normalized size 1 requires 1 normalized time unit to be fully transferred across the switching fabric.

2.5 Performance Measures

The following performance measures are used:

- Maximum (normalized) throughput, defined as the sum of normalized sizes of all received packets per outlet at the maximum offered load, divided by the normalized duration of the simulation
- Packet loss probability, defined as the probability that a new packet arrives at an inlet when the buffer for that inlet is full
- Mean packet delay, defined as the average of the time intervals from when a new packet arrives at an inlet to when it is delivered to its outlet

2.6 Statistical Data

All simulation results presented in this paper were obtained using Modsim III, an object-oriented simulation language from CACI. Simulation results were obtained by simulating the steady-state behavior of these systems and by applying the replication method, as explained in [13]. Simulations were repeated, until performance measures achieved or exceeded the required level of relative precision (1%), at a 95% confidence level.

3. Simulation Results

3.1 Maximum Switch Throughput

Due to partially unused time slots, as mentioned before, the maximum throughput of a SB packet switch cannot reach 100%. It is known however that the maximum switch throughput depends upon the selected packet-size distribution function [10]. Our intention was to investigate the influence of both the selected packet-size distribution function and the selected SF parameter value on the maximum switch throughput.

The switch queues were first saturated with packets by setting the switch load to 100% to assure the correct

results. Our research showed that both Rotation and OPF scheduling mechanisms yield the same throughput results. We also discovered that the packet switch size (number of input and output ports) had no noticeable effect on throughput results.

a	SF							
	1	1.25	1.5	1.75	2	2.25	2.5	2.75
-10	90.9	72.7	61.7	61.3	91.0	81.1	75.4	80.2
-8	88.7	71.2	61.9	64.6	88.9	80.0	76.5	83.0
-6	85.3	69.6	63.2	69.1	86.2	79.4	78.6	85.2
-4	80.6	68.9	66.5	73.9	83.5	80.2	81.7	86.1
-2	76.1	70.3	71.6	78.0	82.4	82.4	84.4	86.5
0	74.7	74.3	77.3	81.3	83.8	85.2	86.8	88.0
2	77.0	79.3	82.3	84.9	86.7	88.1	89.3	90.2
4	80.8	83.8	86.3	88.2	89.7	90.8	91.6	92.4
6	84.3	87.1	89.2	90.7	91.8	92.7	93.4	94.0
8	86.8	89.4	91.1	92.3	93.3	94.0	94.6	95.1
10	88.5	90.8	92.3	93.4	94.2	94.8	95.3	95.8

Table 1 Maximum switch throughput (%) for truncated-exponential traffic source, $SF < 3$

a	SF							
	3	4	5	6	7	8	9	10
-10	91.1	91.2	91.6	92.1	92.7	93.4	94.2	95.0
-8	89.3	89.9	90.7	91.8	92.9	93.9	94.8	95.5
-6	87.4	89.0	90.7	92.3	93.6	94.5	95.2	95.7
-4	86.5	89.3	91.5	93.0	94.0	94.8	95.4	95.8
-2	87.2	90.3	92.3	93.6	94.5	95.2	95.7	96.1
0	88.9	91.7	93.3	94.4	95.2	95.8	96.2	96.6
2	91.0	93.2	94.6	95.4	96.1	96.6	96.9	97.2
4	93.0	94.8	95.8	96.4	96.9	97.3	97.5	97.8
6	94.5	95.8	96.7	97.2	97.5	97.8	98.0	98.1
8	95.5	96.6	97.2	97.6	97.9	98.1	98.3	98.4
10	96.1	97.0	97.6	97.9	98.2	98.3	98.5	98.6

Table 2 Maximum switch throughput (%) for truncated-exponential traffic source, $SF \geq 3$

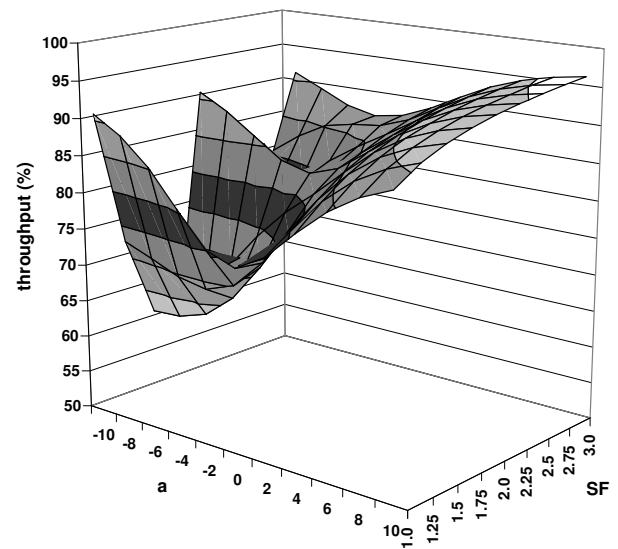


Fig. 4 Maximum switch throughput for truncated-exponential traffic source, $SF \in \mathbf{R}$

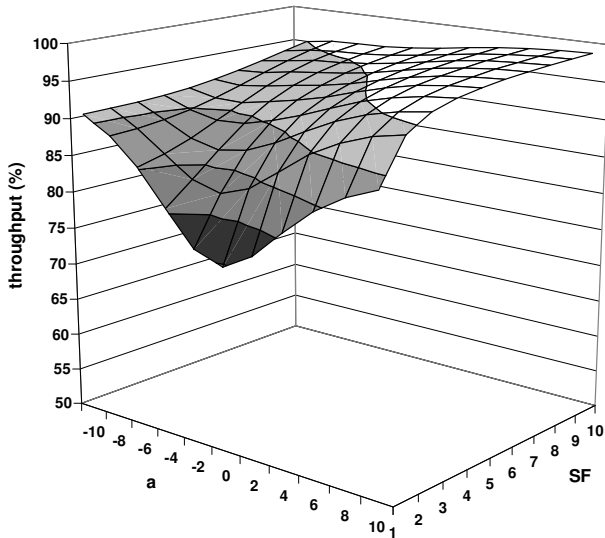


Fig. 5 Maximum switch throughput for truncated-exponential traffic source, $SF \in \mathbb{N}$

In our first experiment, the truncated-exponential traffic source was used. In **Table 1** and **Table 2**, the maximum switch throughput is shown for various values of the parameters a and SF . The same results are presented graphically in **Fig. 4** and **Fig. 5**. Two sub-experiments were performed with the time-slot stretch factor value. First, SF value was increased by a non-integer step, as shown in **Fig. 4**. It can be clearly seen that the use of non-integer SF values result in heavy variations of the maximum switch throughput. The traffic sources with predominantly larger packets are especially affected. This can be easily explained by the fact that if the time slot is increased by only a small fraction, there will not be enough space for any additional packet, and therefore, SF will only increase the percentage of unused time slot and thus lower the maximum switch throughput. Next, the SF value was increased by integer step, as shown in **Fig. 5**. In this case, the maximum switch throughput is a monotonically increasing function of SF , regardless of the selected traffic-source characteristics. In [10] it was shown that the worst throughput results are obtained when uniform packet-size distribution is used. In **Fig. 5** we can easily see that while increasing the SF value, the highest throughput increment is achieved if uniform packet-size distribution is used. Indeed, we can see that for high SF values the maximum switch throughput becomes fairly independent of the packet-size distribution function. Traffic characteristics in real packet networks are usually poorly defined. To avoid any possible throughput degradations, it is highly advisable to use only integer SF values.

In our second experiment, the Internet traffic source was used. In **Fig. 6** and **Fig. 7**, the maximum switch throughput is shown for various values of the SF parameter. Again, SF value was first increased by a non-integer step, which resulted in variations of the maximum switch throughput. The maximum throughput curve begins monotonically increasing by using an integer step. One can readily see the similarities with the truncated-exponential source when the value of the parameter a is set somewhere between 0 and 2.

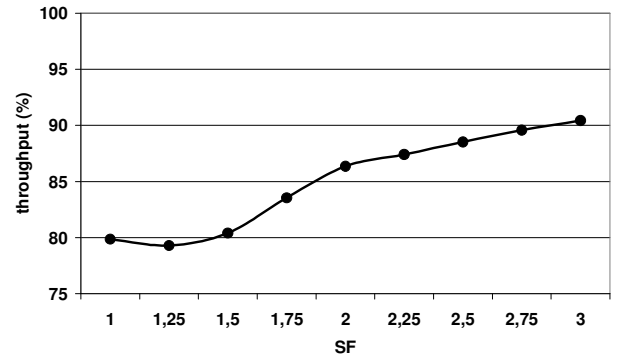


Fig. 6 Maximum switch throughput for Internet traffic source, $SF \in \mathbb{R}$

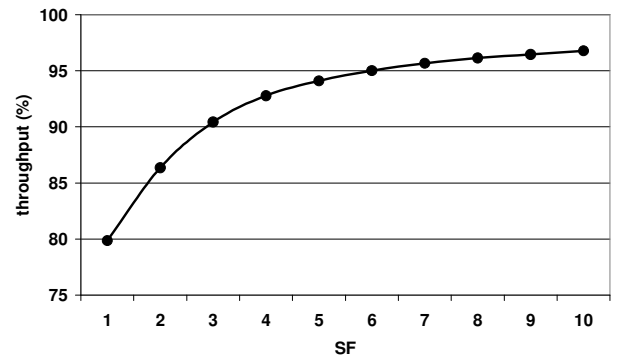


Fig. 7 Maximum switch throughput for Internet traffic source, $SF \in \mathbb{N}$

We can conclude that by using integer SF values, the maximum switch throughput can be noticeably increased, regardless of the traffic-source characteristics. The higher the SF values, the higher the maximum switch throughput.

3.2 Mean Packet Delay

Our Internet traffic-source model was selected to study the mean packet delay in SB packet switch. Packet delay is influenced by many parameters. The selected traffic-source type is the first one. Further, the delay depends upon the selected offered switch load, SF value, scheduling mechanism and switch size (N). In the case of switch congestion, queue sizes also play an important role. We performed simulations for both Rotation and OPF scheduling mechanisms, and for packet switch with 2, 10 and 30 input and output ports. The results are presented in **Fig. 8-Fig. 13**.

The most noticeable delay occurs when the offered switch load is set to 80% and SF takes values of 1 or 1.5. This was to be expected, since with such parameters set, the offered switch load is higher than the maximum switch throughput and, therefore, switch congestion occurs. All other delay results were derived with switch operating in congestion-free mode with no detected packet loss event.

The switch size N has the highest influence on mean packet delay values. Roughly speaking we can say that the increment of mean packet delay values is in linear relation with the increment of the switch size, i.e. doubling the packet switch size also doubles the mean packet delay values. The reason is hidden in the switch queue

configuration. The number of queues at each input port is equal to the number of output ports with the selected VOQ configuration. Since only a single queue can be served at each input at a time, it becomes clear that the higher the number of output ports the longer the interval between queue service opportunities.

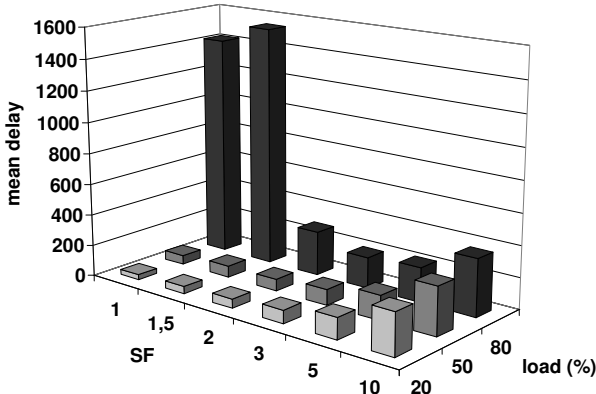


Fig. 8 Mean packet delay, Rotation scheduling, $N = 2$

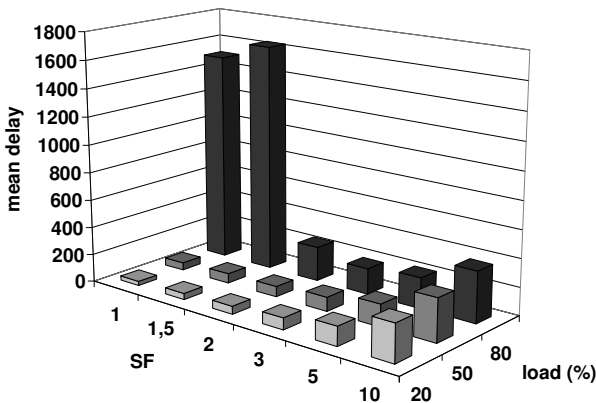


Fig. 9 Mean packet delay, OPF scheduling, $N = 2$

As was expected, the results for small switches ($N = 2$) are fairly independent of the selected scheduling mechanism. On the contrary, for larger switches ($N = 30$), a difference does appear. OPF scheduling performs somewhat better, especially for low SF values. This is due to the shorter time-slot intervals, which enable OPF scheduler to make more frequent and thus better decisions. On the other hand, for higher SF values (e.g. $SF = 10$), the time-slot interval becomes too long for OPF scheduler to exert any noticeable influence on mean packet delay values. In this case, both Rotation and OPF schedulers deliver approx. the same mean packet delay.

The advantage of OPF scheduler clearly appears when the switch is heavily loaded (load = 80%). Let's look at the mean packet delay of large switch with $SF = 2$, as presented in Fig. 12 and Fig. 13. Here we can readily see that the mean packet delay of Rotation scheduler is approx. twice the mean packet delay of OPF scheduler. OPF scheduler has other advantages as well. For example, our finding, not presented in the accompanying figures, is that the maximum packet delay produced by OPF scheduler is

approx. three times lower in comparison to the Rotation scheduler.

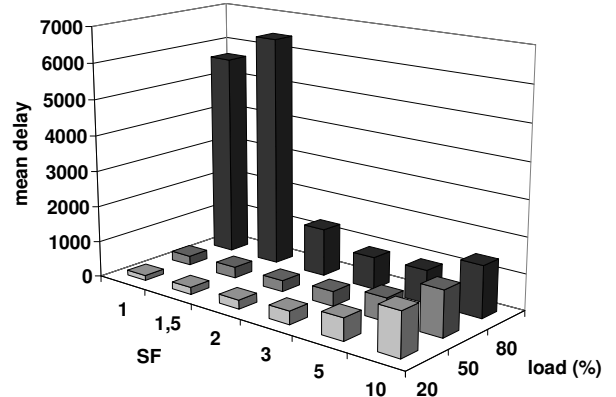


Fig. 10 Mean packet delay, Rotation scheduling, $N = 10$

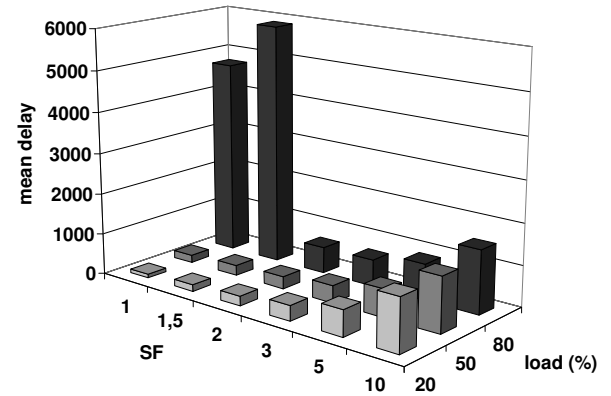


Fig. 11 Mean packet delay, OPF scheduling, $N = 10$

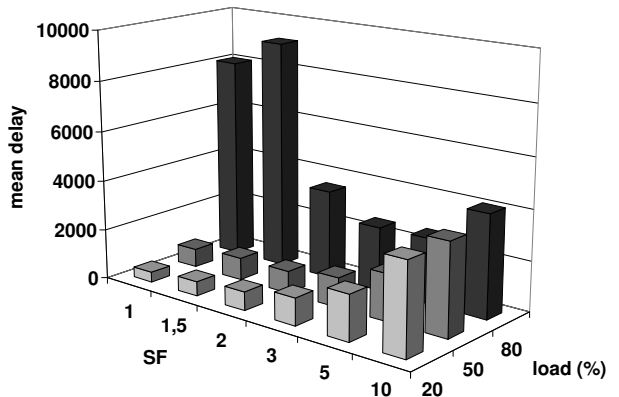


Fig. 12 Mean packet delay, Rotation scheduling, $N = 30$

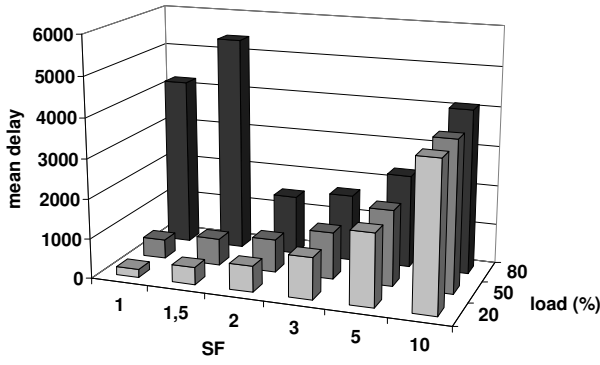


Fig. 13 Mean packet delay, OPF scheduling, $N = 30$

3.3 Service Class Differentiation Capabilities

Here we investigate the influence of SF parameter on the capability of SB packet switch to differentiate between two service classes, C1 and C2. In all simulations, the Internet traffic source and the OPF scheduling mechanism were used. Each service class provides 50% of the offered switch load. The weight of the first service class ($W1$) is set fixed to 1, while the weight of the second service class ($W2$) varies between 1 and 5. The current packet's priority is calculated by multiplying the current packet's age with the weight of appropriate service class. Therefore, a higher weight value provides a better chance to be served promptly.

Separate queues are used for each service class in order to assure independent queuing. This means that the total number of input queues is increased to $2N^2$. This in turn inevitably doubles the mean packet delay in comparison to a single service class. Nevertheless, it does enable service class separation.

Simulation results are presented of Fig. 14-Fig. 17. On the x -axis, together with the SF parameter, the service class identification is presented. Readers can thus easily compare the results of both service classes for the same SF value. By observing the results we can clearly see that service class separation performs well in all simulated experiments, regardless of the SF value, switch size parameter N or offered switch load. The influence of parameters SF and N on mean packet delay values can be explained in the same way as for a single service class.

In cases with no congestion, by increasing the $W2$ value, the delay of high priority traffic is lightly reduced, as shown in Fig. 14 and Fig. 15. At the same time, the delay of low priority traffic is increased considerably. Thus, the overall switch performance is decreased. On the other hand, in the case of switch congestion, the situation is inverted, as shown in Fig. 16. The delay of low priority traffic cannot be greatly increased since the queues are already full. But the delay of high priority traffic can be lowered considerably. At the same time, the packet loss probability of high priority traffic is also considerably reduced and can be completely eliminated, as shown in Fig. 17. Here we can also see that the overall packet loss probability is reduced when SF value is increased, due to the increment of maximum switch throughput.

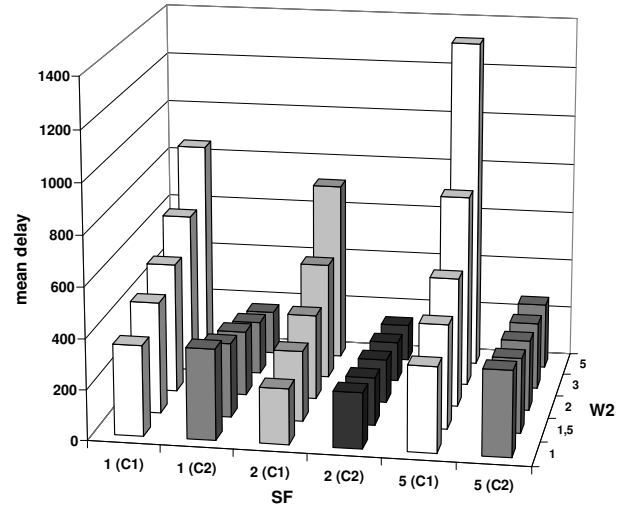


Fig. 14 Mean packet delay for two service classes, load = 75%, $N = 2$

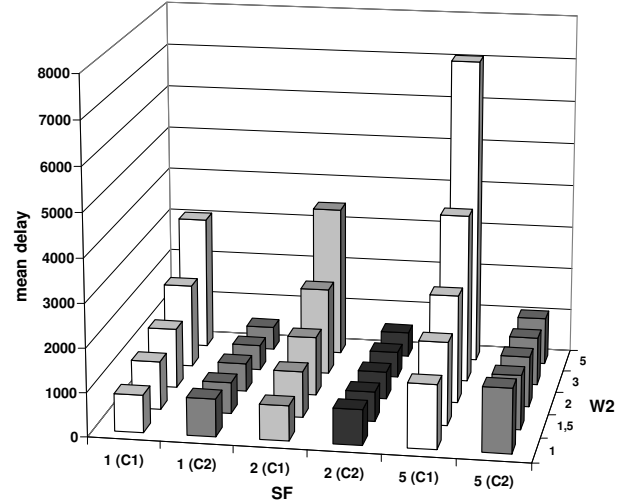


Fig. 15 Mean packet delay for two service classes, load = 75%, $N = 10$

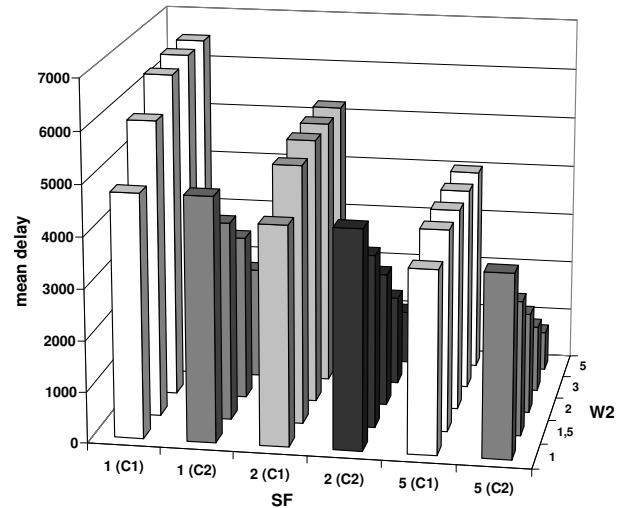


Fig. 16 Mean packet delay for two service classes, load = 100%, $N = 2$

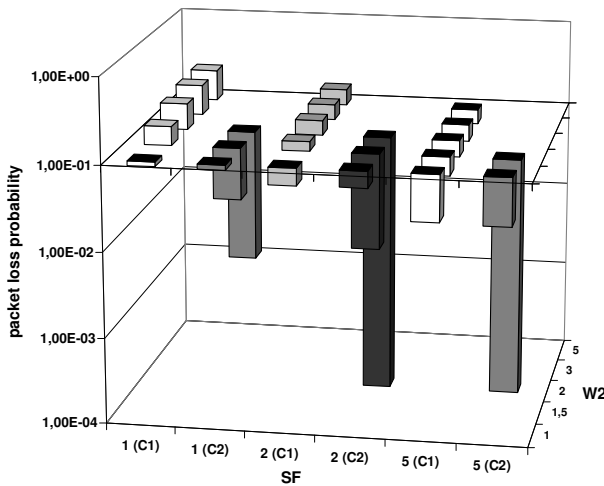


Fig. 17 Packet loss probability for two service classes, load = 100%, $N = 2$

4. Conclusions

We began this paper by describing the time-limited bulk service and how it can be used to build a synchronous packet switch for variable-size packets. The main purpose of this paper was to introduce the time-slot stretch factor SF and to investigate its influence on the overall performance of the SB packet switch. By using a discrete-event simulation engine, maximum switch throughput values were derived. By using the same method, mean packet delay and packet loss probability values were derived as well. Furthermore, we examined the effectiveness of switch capability to distinguish between two service classes.

There is no doubt that by using the proposed time-slot stretch factor SF , the maximum switch throughput is increased. Theoretically speaking, one could achieve a 100% maximum throughput by using infinite SF value. Unfortunately, by doing so the packet transmission delays would also become infinite and such a switch would be useless. The rationale of these words is that a compromise must be negotiated between the desired maximum switch throughput and the acceptable packet delays. As it was shown, for $SF = 10$ the maximum switch throughput is increased to at least 95% or higher, depending upon the selected traffic-source characteristics. In the case of a uniform packet-size distribution, the maximum throughput is increased from 75% to 97%. It is important to use integer SF numbers, especially for smaller SF values, in order to prevent any possible switch throughput degradation. The influence of the traffic-source characteristics on the maximum switch throughput is decreased when SF value is increased.

Our simulations revealed, that in addition to the SF parameter, there is an even more influential parameter to be considered as far as packet delays are concerned. We are talking about the number of packet switch input and output ports. Our simulations showed that the packet delays increase proportionally with the increment of switch size N . In previous chapters, we have shown that using an

advanced packet scheduling mechanism cannot eliminate this process. At best it can be somewhat alleviated. Therefore, the question arises whether it is reasonable to implement complex, and thus costly scheduling mechanisms, especially if a single service class is used. However, we must not forget that use of a simple Rotation algorithm is limited to a uniform traffic distribution pattern. For non-uniform traffic, we can assume that the maximum switch throughput would decrease, because some queues could be empty while others would be filled up. Some mechanism would have to be added to the Rotation algorithm in order not to serve the empty queues. This is a subject for further study.

As we have seen, service class differentiation is possible and is not affected by the increased SF value. Nevertheless, we have to consider two factors. First, by using separate queues for each service class, the total number of queues increases rapidly, which in turn increases the overall packet delay due to increased time intervals between queue service opportunities. Second, when the switch is in a congestion-free state, service class separation is not very productive – little is gained but much more is lost. Traffic separation could be useful in the case of switch congestion; i.e. it could be used to prevent important traffic from losing packets. In any case, in order to enable service class differentiation, an advanced scheduling mechanism must be implemented. This is also a subject for further study.

Based on simulation results we can conclude that the presented SB packet switch is suitable for implementation in wide area networks (WAN) because of high transmission speeds and the small number of switch ports. High transmission speeds allow higher SF values, which in turn increase the maximum switch throughput and thus switch utilization. Small number of switch ports is required in order to keep packet delays under control. We must not forget that by using the SB packet switch the need for packet segmentation and reassembly is eliminated. In local area networks (LAN), with low transmission speeds and larger numbers of switch ports, other packet switch types may be considered.

References

- [1] E. Traupman, P. O'Connell, J. Minnis, M. Jadul, and H. Mario, The Evolution of the Existing Carrier Infrastructure, *IEEE Commun. Mag.*, vol.37, no.6, pp.134-139, June 1999.
- [2] H. Hadama, K. Kitami, K. Kinoshita, and H. Yamaguchi, GMN: A Broadband Multimedia Communications Service Concept and Field Trials, *IEEE Commun. Mag.*, vol.37, no.10, pp.64-70, October 1999.
- [3] L. Mathy, C. Edwards, and D. Hutchison, The Internet: A Global Telecommunications Solution?, *IEEE Network*, vol.14, no.4, pp.46-57, July/August 2000.
- [4] D.-J. Ma, The Maximum Throughput of a Nonblocking Packet Switch with Window Policy, *IEICE Trans. Commun.*, vol.E80-B, no.4, pp.573-580, April 1997.
- [5] I. Iliadis and W. E. Denzel, Analysis of Packet Switches with Input and Output Queuing, *IEEE Trans. Commun.*, vol.41, no.5, pp.731-740, May 1993.
- [6] G. Armitage, MPLS: The Magic Behind the Myths, *IEEE Commun. Mag.*, vol.38, no.1, pp.124-131, January 2000.
- [7] C.-T. Lea, C.-Y. Tsui, B. Li, L. C.-Y. Kwan, S. K.-M. Chan, and A. H.-W. Chang, A/I Net: A Network that Integrates ATM and IP, *IEEE Network*, vol.13, no.1, pp.48-55, January/February 1999.

- [8] C. Bison and A. Pattavina, Connectionless switching by asynchronous banyan networks, Proc. of GLOBECOM 94, San Francisco, CA, pp.441-447, Nov. 1994.
- [9] C. Partridge, P. P. Carvey, E. Burgess, I. Castineyra, T. Clarke, L. Graham, M. Hathaway, P. Herman, A. King, S. Kohalmi, T. Ma, J. Mcallen, T. Mendez, W. C. Milliken, R. Pettyjohn, J. Rokosz, J. Seeger, M. Sollins, S. Storch, B. Tober, G. D. Troxel, D. Waitzman, and S. Winterble, A 50-Gb/s IP Router, IEEE/ACM Trans. Network., vol.6, no.3, pp.237-248, June 1998.
- [10] P. Giacomazzi, Maximum Throughput Analysis of a Datagram Switch for Broadband Networks, IEICE Trans. Commun., vol.E81-B, no.2, pp.354-362, February 1998.
- [11] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, Achieving 100 % Throughput in an Input-Queued Switch, IEEE Trans. Commun., vol.47, no.8, pp.1260-1267, August 1999.
- [12] K. Thompson, G. J. Miller, and R. Wilder, Wide-area Internet traffic patterns and characteristics, IEEE Network, vol.11, no.6, pp.10-23, November/December 1997.
- [13] H. Akimaru and K. Kawashima, Teletraffic: Theory and Applications, Springer-Verlag, London, 1999.

Peter Homan graduated in 1995. He received his M.S. and Ph.D. from the University of Ljubljana, Slovenia, in 1998 and 2001, respectively. All degrees are in the field of telecommunications. At present, he is employed on the Faculty of Electrical Engineering in Ljubljana as Senior Researcher. His previous work involved development of telecommunication signaling software (CCS No. 7), with emphasis on intelligent networks (IN). His current work and study focuses on IP, B-ISDN and ATM networks, particularly on quality of service (QoS) issues. He is a member of IEEE and an associate member of IEICE.

Janez Bester graduated in 1979. He received his M.S. and Ph.D. from the University of Ljubljana, Slovenia, in 1982 and 1995, respectively. All degrees are in the field of telecommunications. Currently he is Head of the Laboratory for Telecommunications on the Faculty of Electrical Engineering in Ljubljana. He is also Assistant Professor for the following subjects: Telecommunication Basics II, Commutation Systems and Networks, Communication Networks and Services, Design, Modeling and Management of Telecommunication Networks and Intelligent Networks. His twenty years of research, development and pedagogical activities have focused on the field of planning, realization and management of telecommunication systems and services, and use of information technologies in education. His current work involves research and engineering in the fields of IP, ATM, TM (Telemanagement) and IIN (Integrated Intelligent Networks). He is a member of IEEE and an associate member of IEICE.

The impact of bursty traffic on FPCF packet switch performance

P. Homan*, J. Bester, A. Kos, T. Slivnik

Faculty of Electrical Engineering, University of Ljubljana, Trzaska 25, SI-1000 Ljubljana, Slovenia

Received 9 January 2002; revised 21 May 2002; accepted 12 June 2002

Abstract

This paper analyses and compares the performance of forward planning conflict-free (FPCF), virtual output queuing-partitioned (VOQ-P) and virtual output queuing-shared (VOQ-S) packet switches. The influence of packet burst size, offered switch load and packet switch size are all investigated. Packet loss probability, average packet delay and 99.9% packet delay values are also derived. In comparison tests, FPCF packet switches perform better than VOQ-P packet switches and yield similar results to VOQ-S packet switches. The advantage of VOQ-S switches is manifested in somewhat lower probability of packet loss, while the advantage of FPCF switches is revealed in the limited values of their maximum packet delays.

© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Packet switch; Forward planning conflict-free; Virtual output queuing; Bursty traffic; Performance analysis

1. Introduction

It has become clear of late that the mode of packet transfer shall dominate the future telecommunication networks. The increased popularity of multimedia applications combined with the rapidly growing number of users, has resulted in dramatically increased traffic volumes [1–3]. The throughput of all network elements has to be increased in order to cope with such traffic volumes. Ever since the introduction of optical transmission lines, the throughput limitations have been set by the limited capabilities of packet switches.

One of the major reasons for the success of packet technology has been the introduction of so-called statistical multiplexing [4,5]. It is well known that the volume of any individual traffic flow varies over time. These variations can be quite high, thus requiring the provision of high-speed transmission links to accommodate traffic bursts. The idea of statistical multiplexing is to combine multiple individual traffic flows into a single collective traffic flow. The probability that all individual traffic flows would transmit bursts at the very same time remains low. The collective transmission line capacity can therefore be lower than the sum of all the individual transmission line capacities.

Indeed, this principle is performing well in the case of delay and loss insensitive traffic.

Nevertheless, a traffic congestion situation can occur when too many traffic flows transmit bursts at the same time. Let's take a look at what is going on inside a typical packet switch. New packets are delivered across input lines to multiple input ports: they are then transferred across the switching fabric to appropriate output ports and thence finally transmitted across output lines. A situation can occur when all multiple newly received packets at different input ports want to go to the same output port at the same time. Two problems must be resolved in this case: Firstly, an agreement must be found to assure the conflict-free transfer of packets across the switching fabric. This is a so-called output-contention problem. Secondly, another agreement must be found to assure the conflict-free transmission of packets across each output line. This is a so-called line-contention problem.

The foundation of statistical multiplexing is based on a presumption of typical traffic flow characteristics (e.g. Poisson, Batch–Poisson, Markov–Modulated Poisson Process, Fluid Flow models, etc.) [6]. It is a fact that through multiplexing multiple such traffic sources, a smoothed-out aggregate traffic with reduced burstiness is produced. However, it has become clear that the characteristics of traffic flows in real packet networks tend to have self-similar properties [6–9]. Self-similar traffic is characterised by burstiness across an extremely wide range of time scales.

* Corresponding author. Tel.: +386-1-4768-810; fax: +386-1-4768-732.
E-mail address: peter.homan@fe.uni-lj.si (P. Homan).

It is a fact that by aggregating self-similar traffic flows the intensity of burstiness is retained. The design and management of packet switches is thus additionally complicated.

Packet switches can be generically classified into three types, according to the packet queuing strategies. There are no input queues in output-queued (OQ) packet switches. The switching fabric must run at high speed in order to transfer all newly received packets within a single time-slot [4,10]. This way all packets can be transferred to the same output port almost simultaneously. The output-contention problem is thus entirely eliminated in output-queued packet switches. The line-contention problem is still present however. There are no output queues in input-queued (IQ) packet switches. The switching fabric is operating at the same speed as output lines and therefore no switching fabric speedup is required. Unfortunately, the line-contention and the output-contention problems are both unified into a single, rather complicated packet-scheduling problem. Queues are distributed at both ends in combined input-output queued (CIOQ) packet switches. Both, output-contention and line-contention problems exist. However, with careful design a combination of both problems can be solved more easily than the aforementioned unified problem which exist in IQ packet switches, this for the moderate cost of switching fabric speedup [10,11]. CIOQ packet switches thus represent a compromise between the switching fabric speedup requirements of OQ packet switches and the complicated scheduling mechanisms of IQ packet switches.

Switching fabric can be designed in several ways. Different switching architectures can adopt either a non-blocking network with input and/or output queuing, or a blocking network with various queuing options. This paper restricts itself to non-blocking switching architecture, which is most often implemented by a crossbar-switching network. That is, it is assumed that when the destination ports of the packets are all distinct, the switch fabric is capable of transmitting any set of packets offered at its inputs without any internal blocking. Output queuing is impractical for switches with high line rates and/or a large number of ports, since the fabric and memory of an $N \times N$ switch must run N times as fast as the line rate. Unfortunately, at high line rates, memories with sufficient bandwidth are simply not available. On the other hand, the fabric and the memory of an input-queued switch need only run as fast as the line rate. This makes input queuing very appealing for switches with fast line rates or with a large number of ports. For this reason, the highest-performance switches and routers use input-queued crossbar switches [12].

This paper deals with IQ packet switches. In most existing IQ packet switches first-in first-out (FIFO) queues are used. A single FIFO queue per input port is used in generic FIFO input-queued packet switch architecture. The received packets are simply inserted at the end of appropriate FIFO queues. The packet scheduler is invoked at the beginning of each time-slot. Only packets at the front of FIFO queues can be scheduled, which leads to

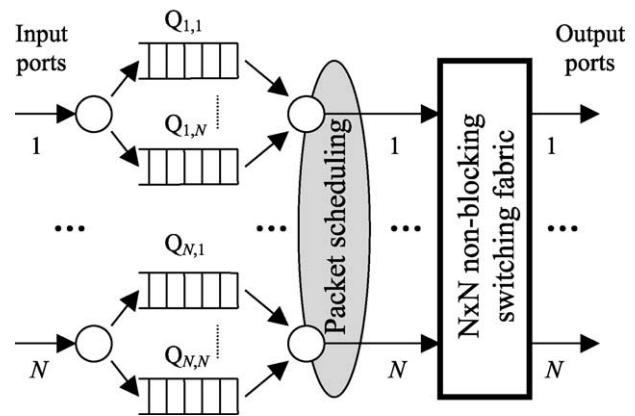


Fig. 1. VOQ—Virtual output queuing FIFO packet switch architecture.

the well-known head-of-line (HOL) blocking events. Due to HOL blocking, the maximum throughput of such switches is limited to no more than 58%. In order to prevent HOL blocking events, a so-called virtual output queuing (VOQ) queue configuration is usually implemented, i.e. rather than maintaining a single FIFO queue for all packets, each input maintains a separate FIFO queue for each output [10]. Fig. 1 presents a VOQ queue configuration.

Another way to achieve high throughput in IQ packet switches is through the use of forward planning conflict-free (FPCF) queuing policy [13]. Fig. 2 depicts a generic FPCF packet switch architecture. The essential idea is that by using a FPCF algorithm, the output-contention problem at the input to the switching fabric can be entirely eliminated. Any possible conflicts are resolved before the newly received packets are inserted into appropriate FPCF queues. Newly received packets can be inserted at any position inside FPCF queues. All packets inside FPCF queues are periodically shifted toward the switching fabric at the end of each time-slot. FPCF switches have one distinctive advantage over FIFO IQ switches—the total packet delay is known immediately after the packet is inserted into the FPCF queue and the worst case packet delay is limited by

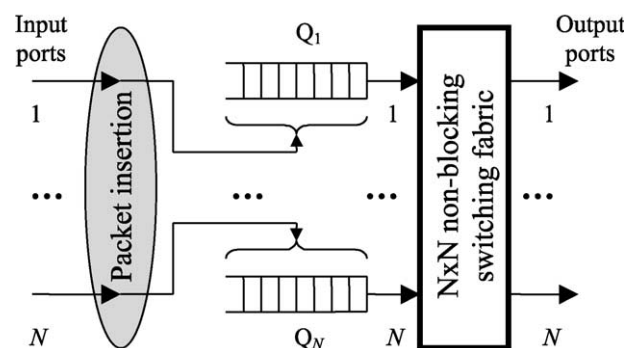


Fig. 2. Packet switch architecture for FPCF queuing policy.

the FPCF queue size B and the switching fabric transfer speed.

A performance analysis of the FPCF packet switch was conducted by Yau and Pawlikowski [13]. They compared the results with those obtained by the VOQ FIFO packet switch. It was their conclusion that the FPCF packet switch performs just as well as the VOQ packet switch with an optimal scheduling algorithm. However, all their results [13] were obtained using simple homogenous and uniform traffic sources, without any packet bursts. It has been illustrated [6–9] that the characteristics of real traffic flows do not comply with the aforementioned assumptions of homogeneity and uniformity. Real traffic flows frequently indicate self-similar properties. This can cause severe packet switch performance degradations, especially if IQ packet switches are implicated—the output contention problem becomes even more complicated, resulting in yet lower throughput, higher packet loss and increased packet delays.

It was our intention to study and compare the performance of both VOQ and FPCF packet switches under non-uniform and non-homogenous traffic conditions. We decided to apply so-called bursty traffic sources. All packets within an individual burst are generated at a single instance of time and all are destined for the same destination. Bursty traffic sources appear to be non-uniform and non-homogenous on a temporal scale, which is comparable to the duration of an average burst. The size of the time scale is thus directly related to the size of the average burst size—the larger the bursts, the larger the time scale. By using different burst sizes, the performance of a packet switch can be tested under different traffic conditions.

This paper is organised as follows. Section 2 is a presentation of the simulation environment and methodology used during the switch performance analysis. The simulation results are presented in Section 3, and the main implications are summarised in Section 4.

2. Simulation environment

2.1. Scheduling mechanisms

For each time-slot in VOQ packet switches, only one queue at most among the N of the same input port can transmit packets. Moreover, at each time-slot, an output port can receive packets from at most one input port. A set of N queues (chosen among the N^2 available) such that the previous conditions are met is defined as conflict-free. It is the task of a scheduling mechanism to select a conflict-free set of queues at the beginning of each time-slot.

There exist many scheduling mechanisms for input-queued FIFO packet switches [14–17]. In line with our own experience [18], the so-called oldest packet first (OPF) scheduling mechanism was selected for the simulations. As

the name itself implies, in OPF scheduling the packets are served according to the time they have spent inside the input queues—the longer the delay the higher their priority. A bipartite graph method [14] is usually used to select a conflict-free set of input queues at the beginning of each time-slot.

In FPCF packet switches, a scheduling mechanism is embedded inside the packet insertion function. A simple linear search algorithm with rotating input port priority [13] was applied in our simulations.

2.2. Traffic source models

In our simulations, new packet bursts arrive at inlets following N identical and independent source processes. A uniform reference pattern of arrivals is assumed, i.e. the destinations of bursts are uniformly distributed over the set of N outlets. All packets have the same fixed size.

Burst inter-arrival times are exponentially distributed around a mean value, which is calculated according to the specified offered switch load. It should be noted, however, that the packet inter-arrival rate is limited by the limited capacity of the switch input line rate. Therefore, if the input line is already occupied, a following packet is delayed inside a traffic source until the input line becomes available. The limited input line rate can be considered as a simple traffic shaper.

Burst sizes are determined by the specified average burst size and by the Poisson distribution function. This means that not all bursts have the same size. The size of the average burst determines the duration of non-homogenous and non-uniform time periods.

2.3. Normalised units

Normalised units were used throughout our simulations. Normalised transmission speeds of input lines, input ports, switching fabric, output ports, and output lines are all set to 1.

The packet size of all packets is also set to 1. An average burst size is expressed in an average number of packets inside a burst.

Queue sizes are also expressed in normalised units. The total size of all queues at each input port is fixed set at 2000 normalised units. For example, the size of each individual FIFO queue in a VOQ packet switch with 10 input and 10 output ports (i.e. $N = 10$) is set to 200 units (10 FIFO queues per input port). There is always a single FPCF queue per input port in FPCF packet switches. The size of each individual FPCF queue is therefore predetermined at 2000 units.

A normalised time unit can be readily defined through implementing a normalised transmission speed and a normalised packet size. For example, a packet (with a normalised size of 1) requires 1 normalised time unit to be fully transferred across the switching fabric.

2.4. Measurements of performance

The following measurements of performance were employed:

- *packet loss probability*: defined as the probability that a new packet arrives at an inlet when the queue for that inlet is full;
- *mean packet delay*: defined as the average time interval from when a new packet arrives at an inlet to when it is delivered to its outlet; and
- *99.9% Packet delay*, the definition of which is given in the last paragraph of this section.

In many of the published simulation results, it is common practice to measure mean packet delay and its standard deviation. This information is usually insufficient to make reasonable assessments of packet switch QoS capabilities. For delay sensitive traffic, the maximum permissible packet delay is usually pre-specified. Network operators must know the probability of a packet being delayed for longer than the preordained period of time—i.e. the chances that it will be late. Late packets are no longer of use and can thus to all intents and purposes be regarded as lost.

More informative results were gathered in our simulations. Each individual packet delay was measured and noted during the simulation run-time. The gathered data was subsequently processed at the end of each simulation run. The notion of *packet delay type* was introduced. Already familiar with terms such as *minimum*, *mean* and *maximum packet delay*, we introduced a new family of new delay types which are referred to as *truncated maximum packet delays*.

There can be a virtually unlimited number of different types of truncated maximum packet delay. The delay analysis begins with a complete population of all measured packet delay values. The basic notion is to exclude $X\%$ of most delayed packets and then to find the highest delay value within the remaining population. For example, by ignoring those 0.1% of most delayed packets, the maximum delay of 99.9% of all received packets is measured. This is a so-called *99.9% delay type*.

2.5. Statistical data

All simulation results presented in this paper were obtained using *Modsim III*, an object-oriented simulation language from CACI. Simulation results were obtained by simulating the steady-state behaviour of these systems and thence by applying the replication method, as explained in Ref. [19]. Simulations were repeated until switch throughput and mean packet delay measures achieved or exceeded the requisite level of relative precision (within 1%) at a 95% confidence level. During each simulation execution at least 10^6 packets were generated, setting the minimum detectable packet loss probability at 10^{-6} .

3. Simulation results

The performances of both FPCF and VOQ packet switches were investigated in our simulations. Two parameters have been varied in all experiments. The first one is the *offered switch load*, which was iterated from 50% to 100%, in steps (graduations) of 5%. The second parameter is the *average burst size*, which was repeated through a series of degrees—1, 2, 5, 10, 20, 50 and 100. Due to the pairing of twin two-dimensional input parameters, simulation results are presented in a three-dimensional format.

Three sets of experiments were performed in our simulations. The first set was dedicated to the study of *packet loss probability*. The second and the third set of experiments were dedicated to the study of *average packet delay* and, respectively, *99.9% packet delay*.

Two types of VOQ packet switches were simulated. The first one being virtual output queuing-partitioned (VOQ-P). This is a form of VOQ packet switch where the total buffer space within each input port is evenly divided in static fashion between the individual input queues. This means that the size of each input queue is limited by the pre-determined amount of buffer space, which can thus be used by one and only one input queue. Buffer sharing between individual input queues in VOQ-P packet switches is not permitted. The second type of VOQ packet switch is called virtual output queuing-shared (VOQ-S). This is a form of VOQ packet switch in which the total buffer space within each input port is dynamically distributed between individual input queues. This means that each input queue can use any free buffer space within its input port. Buffer space within input ports is thus fully shared among all input queues in VOQ-S packet switches. However, the sharing of buffer space between input ports is not facilitated.

Three experiments, each with different packet switch sizes, were performed within each set of experiments. Thus,

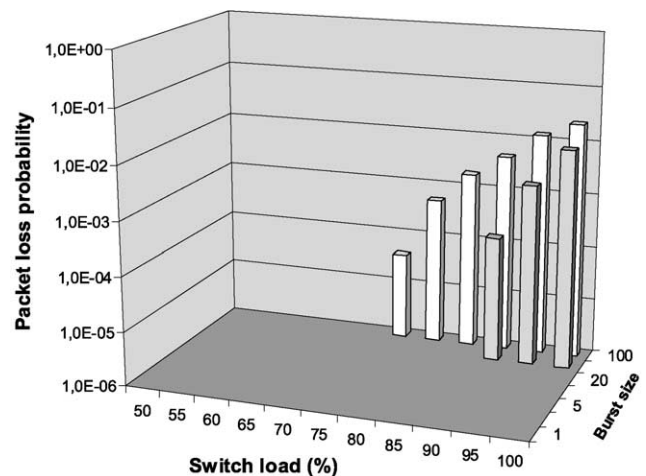


Fig. 3. Packet loss probability against offered switch load and versus average burst size, VOQ-P packet switch where $N = 3$ (three input/output port packet switch with partitioned virtual output queuing).

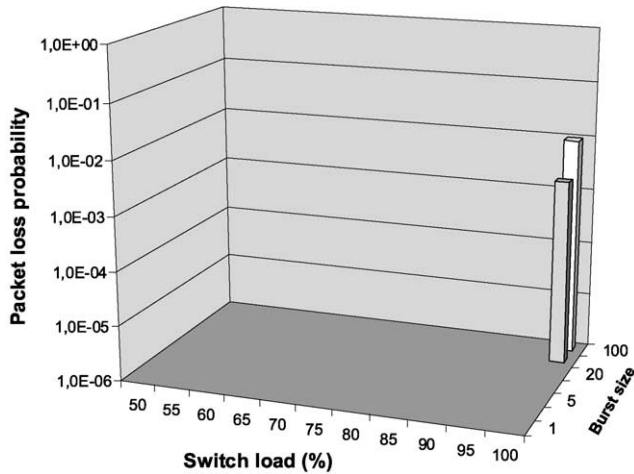


Fig. 4. Packet loss probability against offered switch load and versus average burst size, VOQ-S packet switch, $N = 3$ (three input/output port packet switch with shared virtual output queuing).

the influence of packet switch size was also investigated in our simulations. A small packet switch, with only three input and three output ports ($N = 3$), was used in *Experiment 1*. The predetermined size of each FIFO queue in the VOQ-P packet switch was thus set at 667 units ($2000 \div 3$). A packet switch with 10 input and 10 output ports ($N = 10$) was used in *Experiment 2*. The predetermined size of each FIFO queue in VOQ-P packet switch was thus set at 200 units ($2000 \div 10$). A large packet switch with 30 input and 30 output ports ($N = 30$) was used in *Experiment 3*. The predetermined size of each FIFO queue in the VOQ-P packet switch was thus set at 67 units ($2000 \div 30$). The total size of all FIFO queues within each input port in VOQ-S packet switches was limited by the amount of buffer space within said input port, which itself was limited to 2000 units in all experiments.

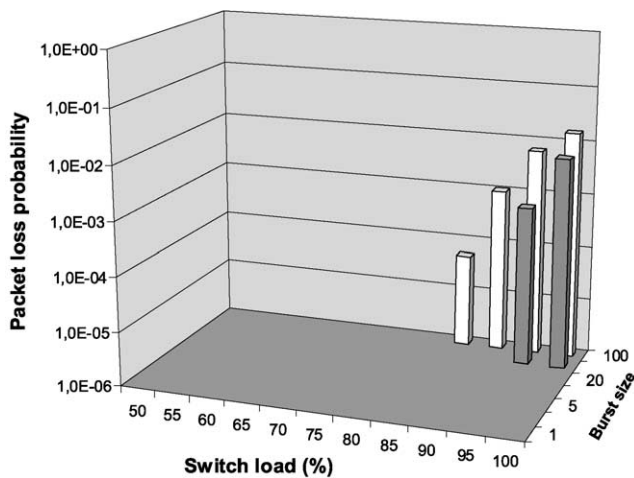


Fig. 5. Packet loss probability against offered switch load and versus average burst size, FPCF packet switch, $N = 3$ (three input/output port packet switch with FPCF switching policy).

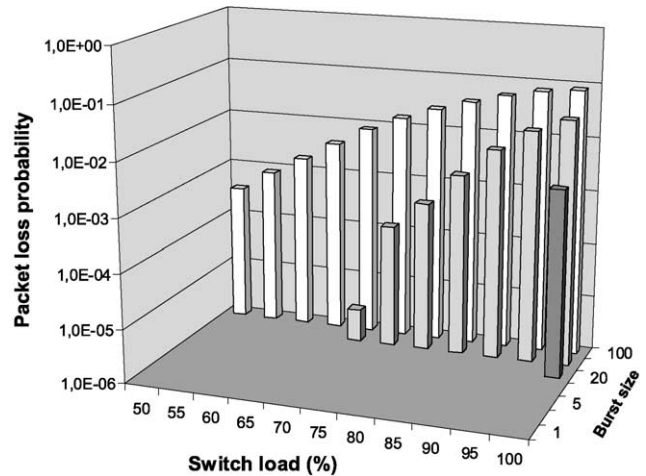


Fig. 6. Packet loss probability against offered switch load and versus average burst size, VOQ-P packet switch, $N = 10$ (10 input/output port packet switch with partitioned virtual output queuing).

The predetermined size of each FPCF queue in FPCF packet switches was likewise set at 2000 units in all experiments.

3.1. Packet loss results

The results of the first experiment ($N = 3$) are presented in Figs. 3–5. They revealed that the packet loss probability is low in all cases, although the initial packet losses do appear a bit sooner in the VOQ-P and FPCF packet switches. For this reason it may be concluded that buffer space is being best utilised in the VOQ-S packet switch. The results of the further two experiments ($N = 10$, $N = 30$) clearly show that the packet loss probability in VOQ-P packet switches increases considerably when the packet switch size is increased, as is illustrated through

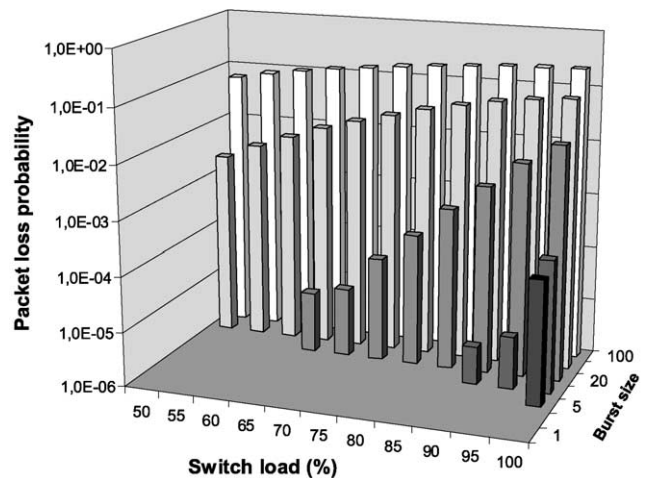


Fig. 7. Packet loss probability against offered switch load and versus average burst size, VOQ-P packet switch, $N = 30$ (30 input/output port packet switch with partitioned virtual output queuing).

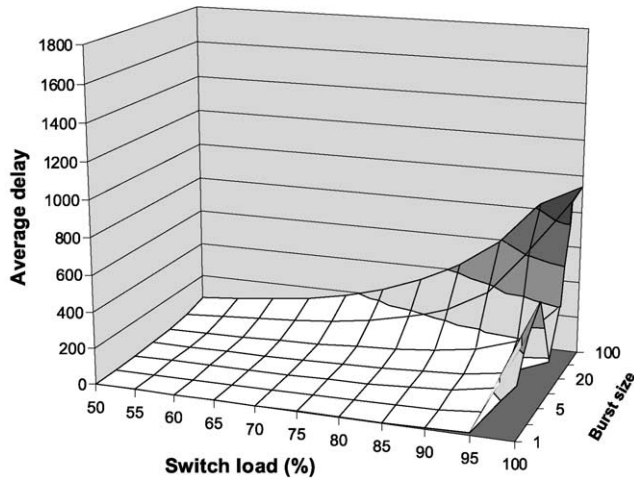


Fig. 8. Average packet delay against offered switch load and versus average burst size, VOQ-P packet switch, $N = 3$ (three input/output port packet switch with partitioned virtual output queuing).

a comparison of Fig. 3 with Figs. 6 and 7. This is mainly due to the small size of individual input queues in these switches, which itself is a direct consequence of partitioned queue configuration. The last two experiments also revealed that the packet loss probability in VOQ-S and FPCF packet switches is fairly independent of the switch size.

3.2. Average packet delay results

The results of the first experiment ($N = 3$) revealed that the measured average packet delay values are similar in all three packet switch types, as can be seen by comparing Figs. 8–10. It is also important to emphasise that the variation of offered switch load and average burst size yielded similar results in all the simulated packet switches. It should be noted, however, that when both the offered switch load and average burst size are set at a high levels, the delay values in VOQ-S packet switches soar to high

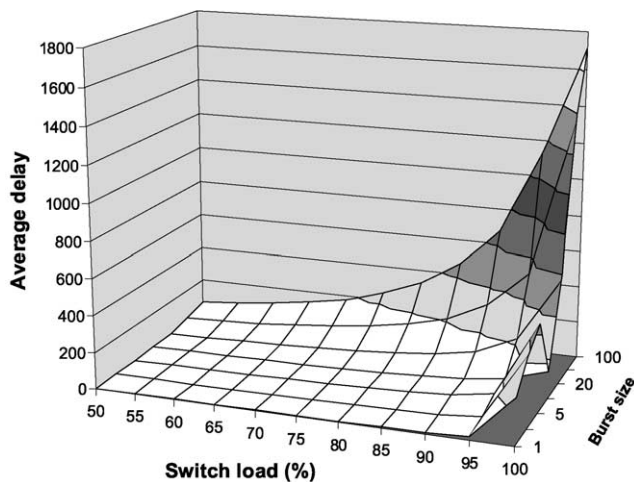


Fig. 9. Average packet delay against offered switch load and versus average burst size, VOQ-S packet switch, $N = 3$ (three input/output port packet switch with shared virtual output queuing).

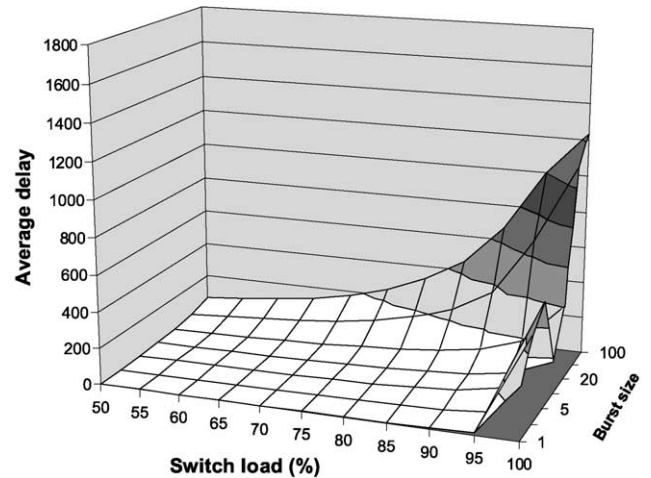


Fig. 10. Average packet delay against offered switch load and versus average burst size, FPCF packet switch, $N = 3$ (three input/output port packet switch with FPCF switching policy).

levels. The results of the further two experiments ($N = 10$, $N = 30$) revealed that the average packet delay values are higher in VOQ-S and FPCF packet switches, especially when the switch is heavily loaded with large packet bursts. The average packet delay values in both VOQ-S and FPCF packet switches turned out to be fairly independent of the switch size. On the other hand, such is not the case with the VOQ-P packet switch. By increasing the switch size, average packet delay values in VOQ-P packet switches decrease noticeably, which is a direct consequence of a markedly higher packet loss probability in the VOQ-P packet switch. However, all simulated packet switches yield similar results for small average burst sizes.

3.3. 99.9% Packet delay results

The results of the first experiment ($N = 3$) revealed that, in most cases, 99.9% packet delays are similar in all three

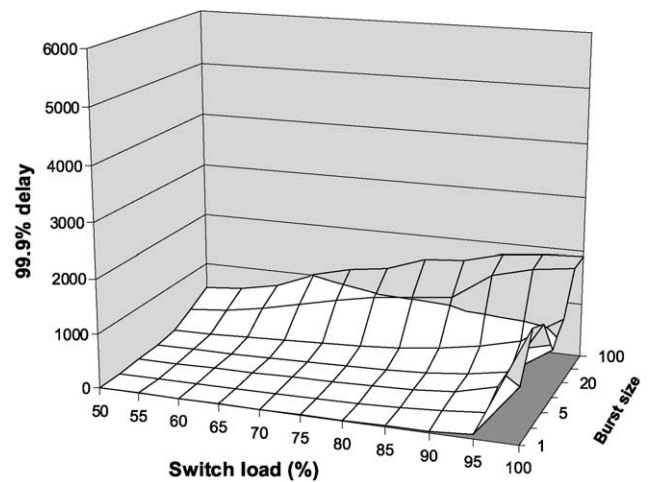


Fig. 11. 99.9% Packet delay against offered switch load and versus average burst size, VOQ-P packet switch, $N = 3$ (three input/output port packet switch with partitioned virtual output queuing).

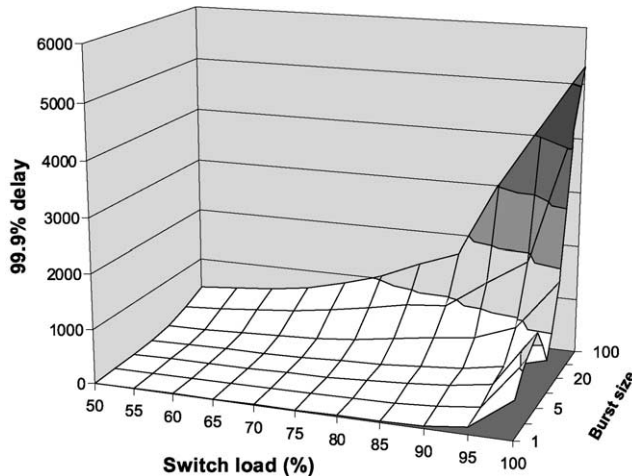


Fig. 12. 99.9% Packet delay against offered switch load and versus average burst size, VOQ-S packet switch, $N = 3$ (three input/output port packet switch with shared virtual output queuing).

simulated packet switch types, as can be seen by comparing Figs. 11–13. It is also important to emphasise that the variation of offered switch load and average burst size yields similar results in all the simulated packet switches. In instances where both offered switch load and average burst size are set at high levels, delays in the VOQ-S packet switch soar to very high levels. This is a direct consequence of the somewhat higher packet loss probability in VOQ-P and FPCF packet switches—lost packets are, in the case of unlimited buffer space, invariably the ones, which would have experienced the highest delay values. Fig. 13 also reveals a typical characteristic of FPCF packet switches—the maximum packet delay of any packet is limited by both the FPCF queue size and the switching fabric transfer speed. In our experiments no packet could ever be delayed for more than 2000 normalised time units

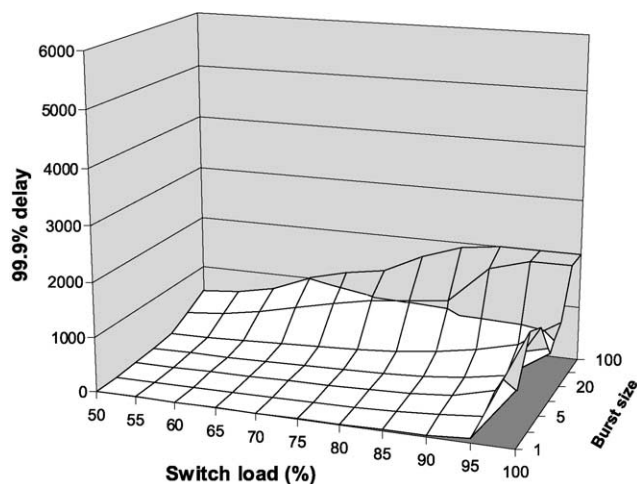


Fig. 13. 99.9% Packet delay against offered switch load and versus average burst size, FPCF packet switch, $N = 3$ (three input/output port packet switch with FPCF switching policy).

within the FPCF packet switch. The results of the further two experiments ($N = 10$, $N = 30$) revealed that 99.9% packet delay values in the FPCF packet switches are independent of the switch size. Similarly, 99.9% delay values in the VOQ-S packet switch are fairly independent of the switch size, although the highest values do decrease when packet switch size is increased. On the other hand, in the case of VOQ-P packet switches, 99.9% packet delay values do decrease considerably when the packet switch size is increased. However, as regards small average burst sizes, all simulated packet switches yield similar results in all three experiments.

4. Conclusions

This paper began with a description of the basic characteristics of traffic flows in packet networks, and was followed by a presentation of the fundamental packet switch structures. The main purpose of this work has been to investigate the influence of packet burst size, offered switch load and packet switch size on the overall performance of FPCF packet switches. For reasons mainly of comparison, we were also interested in the influence of the aforementioned parameters on the VOQ-P and VOQ-S packet switches. By the use of a discrete-event simulation engine, packet loss probability, average packet delay and 99.9% packet delay values were derived.

The main results can be summarised through the following statements:

- All simulated packet switches yield similar results when small average packet bursts (a few packets in a row) are generated by traffic sources. This statement holds true for all packet switch sizes and is independent of the offered switch load.
- All simulated packet switches yield similar results when small packet switches (low number of input and output ports) are used. This statement holds true for all average packet burst sizes and is independent of the offered switch load.
- The performance characteristics of FPCF packet switches are not influenced by packet switch size. This remains true in all types of traffic flow.
- The influence of packet switch size on VOQ-S packet switch performance characteristics is relatively small. This is a consequence of a fully shared buffer space within each input port. It can be stated that VOQ-S packet switches are fairly independent of the selected switch size. This remains true in all types of traffic flow.
- The characteristics of VOQ-P packet switches deteriorate when packet switch size is increased. Due to the limited buffer capacity per input port and an increasing number of FIFO queues per input port, it is unavoidable that the performance of a VOQ-P packet switch

deteriorates when the switch size is increased. This is especially true if bursty traffic flows are being transferred.

- Packet delays (both average and 99.9%) are increased proportionally with the increment of average packet burst size and exponentially with the increment of offered switch load, provided that no packet is lost.

Generally speaking we can say that FPCF packet switches perform better than VOQ-P packet switches. This is especially true for packet switches with a large number of switch ports. FPCF packet switches are appropriate for a modular structure, where individual input and output ports can be added or removed, as is necessary, without affecting the overall characteristics of the packet switch.

VOQ-S packet switches yield similar results when directly compared to FPCF packet switches. The advantage of VOQ-S packet switches is exhibited in lower packet loss probability, while the advantage of FPCF packet switches lies in the relatively low level of their maximum packet delays. VOQ-S packet switches are, therefore, more appropriate for loss sensitive non-real-time transfers, whereas FPCF packet switches perform better in instances where real-time traffic flow transfers need to be insured.

We can conclude this paper by saying that the main performance characteristics of VOQ-S and FPCF packet switches are very much alike. The major difference occurs in the structure of these two switch types. It is, therefore, our belief that issues such as complexity and cost of implementation would play pivotal roles in any final decision as to which type should be used in any given instance.

References

- [1] E. Traupman, P. O'Connell, J. Minnis, M. Jadul, H. Mario, The evolution of the existing carrier infrastructure, *IEEE Commun. Magn.* 6 (1999) 134–139.
- [2] H. Hadama, K. Kitami, K. Kinoshita, H. Yamaguchi, GMN: a broadband multimedia communications service concept and field trials, *IEEE Commun. Magn.* 10 (1999) 64–70.
- [3] L. Mathy, C. Edwards, D. Hutchison, The Internet: a global telecommunications solution?, *IEEE Network* 4 (2000) 46–57.
- [4] N. Giroux, S. Ganti, *Quality of Service in ATM Networks*, Prentice-Hall, New Jersey, 1999.
- [5] W. Stallings, *High-Speed Networks: TCP/IP and ATM Design Principles*, Prentice-Hall, New Jersey, 1998.
- [6] B. Tsybakov, N.D. Georganas, On self-similar traffic in ATM queues: definitions, overflow probability bound, and cell delay distribution, *IEEE/ACM Trans. Network* 3 (1997) 397–408.
- [7] M.E. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: evidence and possible causes, *IEEE/ACM Trans. Network* 6 (1997) 835–846.
- [8] Z. Sahinoglu, S. Tekinay, On multimedia networks: self-similar traffic and network performance, *IEEE Commun. Magn.* 1 (1999) 48–52.
- [9] J. Cao, W.S. Cleveland, D. Lin, D.X. Sun, On the nonstationarity of Internet traffic, *Proc. ACM SIGMETRICS Cambridge MA* (2001) 102–112.
- [10] P. Krishna, N.S. Patel, A. Charny, R.J. Simcoe, On the speedup required for work-conserving crossbar switches, *IEEE J. Select Areas Commun.* 6 (1999) 1057–1065.
- [11] C. Mikenberg, T. Engbersen, A combined input and output queued packet switched system based on PRIZMA switch-on-a-chip technology, *IEEE Commun. Magn.* 12 (2000) 70–77.
- [12] C. Partridge, P.P. Carvey, E. Burgess, I. Castineyra, T. Clarke, L. Graham, M. Hathaway, P. Herman, A. King, S. Kohalmi, T. Ma, J. Mcallen, T. Mendez, W.C. Milliken, R. Pettyjohn, J. Rokosz, J. Seeger, M. Sollins, S. Storch, B. Tober, G.D. Troxel, D. Waitzman, S. Winterble, A 50-Gb/s IP router, *IEEE/ACM Trans. Network* 3 (1998) 237–248.
- [13] V. Yau, K. Pawlikowski, An algorithm that uses forward planning to expedite conflict-free traffic assignment in time-multiplex switching systems, *IEEE Trans. Commun.* 11 (1999) 1757–1765.
- [14] N. McKeown, A. Mekkittikul, V. Anantharam, J. Walrand, Achieving 100% throughput in an input-queued switch, *IEEE Trans. Commun.* 8 (1999) 1260–1267.
- [15] N. McKeown, The iSLIP scheduling algorithm for input-queued switches, *IEEE/ACM Trans. Network* 2 (1999) 188–201.
- [16] M.A. Marsan, A. Bianco, E. Leonardi, L. Milia, RPA: a flexible scheduling algorithm for input buffered switches, *IEEE Trans. Commun.* 12 (1999) 1921–1932.
- [17] N.D. Serpanos, P.I. Antoniadis, FIRM: high-speed distributed scheduling for ATM switches with multiple input queues, *IEE Electron Lett.* 22 (1999) 1915–1916.
- [18] P. Homan, *Queue Configuration and Control for Assuring Qualitative Telecommunications Services in Packet Switches*, PhD Dissertation, University of Ljubljana, 2001.
- [19] H. Akimaru, K. Kawashima, *Teletraffic: Theory and Applications*, Second ed., Springer, Berlin, 1999.

Peter Homan graduated in 1995. He received his M.Sc. and PhD degrees in the field of telecommunications from the University of Ljubljana, Slovenia, in 1998 and 2001, respectively. At present, Peter Homan is employed as a Senior Researcher at the Faculty of Electrical Engineering in Ljubljana. His previous work has involved the development of telecommunication signalling software (CCS No. 7), with emphasis on intelligent networks (IN). His current work and study focuses on IP, B-ISDN and ATM networks, with particular emphasis on quality of service (QoS) issues. Peter Homan is a member of the IEEE and an associate member of the IEICE.

Janez Bester graduated in 1979. He received his M.Sc. and PhD degrees in the field of telecommunications from the University of Ljubljana, Slovenia, in 1982 and 1995, respectively. Currently Janez Bester is Head of the Laboratory for Telecommunications at the Faculty of Electrical Engineering in Ljubljana. He is also an Assistant Professor for such subjects as Telecommunication Basics II, Commutation Systems and Networks, Communication Networks and Services, Design, Modelling and Management of Telecommunication Networks and Intelligent Networks. Bester's 20 years of research and development activities as well as teaching have focused on the field of planning, realisation and management of telecommunication systems and services, together with the application of information technologies in education. His current work involves research and engineering in the fields of IP, ATM, TM (Telemanagement) and IIN (Integrated Intelligent Networks). Janez Bester is a member of the IEEE and ACM as well as an associate member of the IEICE.

Andrej Kos graduated from the University of Ljubljana, Slovenia, in 1996 and was awarded his M.Sc. degree in the field of telecommunications in 1999. At present he is employed as a Senior Researcher at the Faculty of Electrical Engineering in Ljubljana and is currently working on his PhD. Andrej Kos has extensive research and industrial experience in the design and analysis of advanced telecommunications systems and networks. His current work and study focuses on the next generation element and network designs, especially in the field of packet switching and Internet protocol suite, including multi-protocol label switching and asynchronous transfer mode networks. Other fields of Kos' expertise include mobile communications, telemanagement and telecommunication signalling software (CCS No. 7). He is a member of the IEEE and TM Forum.

Tomaz Slivnik received his B.Sc., M.Sc. and PhD degrees in the field of Electrical Engineering from the University of Ljubljana, Slovenia in 1971, 1973 and 1975, respectively. In 1975 he also gained a B.Sc. in Mathematics at the University of Ljubljana. Currently he is a Professor of Mathematics and Electrical Engineering at the Faculty of Electrical Engineering in Ljubljana. A member of the IEEE, Tomaz Slivnik's research interests include mathematical analysis, electromagnetic field calculations, microprocessor applications as well as analysis of telecommunication networks.

PAPER

Performance Evaluation of a Synchronous Bulk Packet Switch Under Real Traffic Conditions

Andrej KOS^{†a)}, *Regular Member*, Peter HOMAN[†], and Janez BEŠTER[†], *Associate Members*

SUMMARY Real traffic flows are captured in various network environments and their statistical properties are analyzed. Based on real traffic flows, MWM (Multifractal Wavelet Model) and Poisson equivalent synthetic traffic flows are generated. Performance analysis of a SB (Synchronous Bulk) packet switch is joined with different types of traffic. Maximum throughput performance of the SB packet switch for various real traffic flows and appropriate MWM and Poisson equivalent synthetic traffic flows are evaluated by using discrete-event simulations. Different flow persistence, SF (Stretch Factor) and scheduling mechanisms are used in order to assess their influence on SB packet switch performance. Traffic asymmetry, either input or output based, has a major influence on SB packet switch performance. By increasing the level of asymmetry, maximum throughput values decrease considerably, especially if the ROT (Rotation) scheduling mechanism is applied. Traffic asymmetry also decreases the influence of the SF parameter on maximum switch throughput. As a general rule of thumb, SF values of no more than 5 must be used if asymmetrical traffic is switched. It is also advisable that OPF (Oldest Packet First) scheduling mechanism is used in such cases. The influence of burstiness and scaling of traffic flows turns out to be relatively insignificant for the SB packet switch maximum throughput results, if the OPF scheduling mechanism is used. Larger throughput discrepancies are detected, if ROT scheduling is used.

key words: *synchronous bulk packet switch, real traffic, traffic asymmetry and burstiness, switch utilization*

1. Introduction

In recent years it has become clear that the mode of packet transfer will dominate the future of telecommunication networks. There are many reasons for the success of packet technology [1], one of the major being the introduction of so-called statistical multiplexing [2]. The idea of statistical multiplexing is to combine multiple individual traffic flows into a single collective traffic flow. The probability that all individual traffic flows transmit bursts at the very same time remains low. The collective transmission line capacity can therefore be lower than the sum of all the individual transmission line capacities.

The foundation of statistical multiplexing is based on a presumption of simple traffic flow characteristics (such as Poisson, Batch-Poisson, Markov-Modulated Poisson Process, Fluid Flow models, etc.) [3], [4],

where through multiplexing multiple traffic sources, a smoothed-out aggregate traffic with reduced burstiness is produced. However, it has become clear that real telecommunication traffic is not so simple. Indeed, it is so complex that even its analysis and modelling is a daunting task considering the amount and variety of available data [5].

The important factors in evaluating modern packet switch performance are the invariants, i.e., characteristics of traffic that are observed with some reproducibility and independently of the precise setting of the network under consideration. Examples of two such important invariants are long-range dependence and heavy tail marginal distributions [6].

Analysis of high-quality traffic measurements confirms the prevalence of long-range dependent features and scaling in traffic processes loading packet switching communications networks [5], [7]–[11], where aggregation of traffic is really not successful in smoothing out traffic burstiness. Only recently, has analysis suggested that in contrast to network edge traffic, highly aggregated core traffic slowly tends back to Poisson [12], [13].

Burstiness is important because from the field of performance analysis of switches, it is known that increased burstiness results in lower levels of resource utilization [14], [15]. Considering quality of service parameters, this impacts available bandwidth, different measures of delay, jitter, and packet loss. In the fields of network and service management, it influences on call admission control, congestion control, and policies for fairness [5]. Considering all of the above, the design and management of packet switches is therefore even more complicated.

In most existing IQ (Input Queued) packet switches FIFO (First-In First-Out) queues are used [16]–[18]. A single FIFO queue per input port is used in generic FIFO input-queued packet switch architecture. The received packets are simply inserted at the end of appropriate FIFO queues. The packet scheduler is invoked at the beginning of each time-slot. Only packets at the front of FIFO queues can be scheduled, which leads to well-known HOL (Head of Line) blocking events. Due to HOL blocking, the maximum throughput of such switches is limited to no more than 58%. In order to prevent HOL blocking events, a so-called VOQ (Virtual Output Queuing) queue configuration is usually implemented—i.e. rather than maintain a sin-

Manuscript received August 15, 2002.

Manuscript revised November 19, 2002.

[†]The authors are with the University of Ljubljana, Faculty of Electrical Engineering, Tržaška c. 25, SI-1000 Ljubljana, Slovenia.

a) E-mail: andrej.kos@fe.uni-lj.si

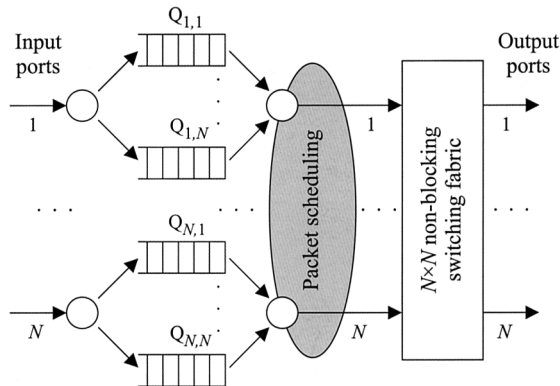


Fig. 1 Virtual output queuing queue configuration.

gle FIFO queue for all packets, each input maintains a separate FIFO queue for each output. Figure 1 presents a VOQ configuration.

SB (Synchronous Bulk) packet switches are a special variant of VOQ packet switches [19]. An enhanced version of SB packet switches with increased performances was introduced in [20]. SB packet switches use synchronous switching fabric and are nevertheless capable of effective switching of variable-size packets. This makes them suitable for switching contemporary packet traffic, predominantly composed of variable-size IP (Internet Protocol) packets. Earlier performance analyses of SB packet switches were performed using homogenous and uniform synthetically generated traffic sources. These traffic sources obviously do not recreate everyday real-life traffic conditions. As noted, many scientific papers from the field of telecommunications traffic engineering deal with analysis and characterization of real traffic.

Joining performance analysis of a SB packet switch with different types of real traffic is one of original ideas presented in the paper. Another original idea in the paper is to generate and apply equivalent MWM and Poisson synthetic traffic flows in order to improve the validity of packet switch simulation results.

Therefore, it is our intention to study and compare the performance of various SB packet switch configurations loaded with different real traffic sources and equivalent synthetic traffic sources.

This paper is organized as follows. In Sect. 2, SB packet switch architecture is briefly described. Section 3 depicts the process of capturing, modeling, and analyzing traffic data. Methodology and simulation results are presented in Sect. 4 and Sect. 5, respectively. The main implications are summarized in Sect. 6.

2. Synchronous Bulk Packet Switch

The simulated SB packet switch is an IQ packet switch with VOQ queue configuration. It is composed of N input ports, a synchronous switching fabric, and N output ports. All individual queues are organized accord-

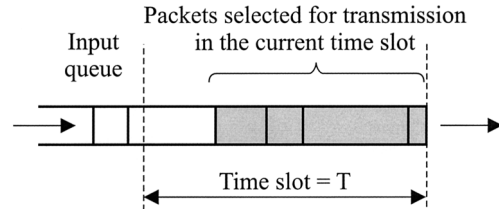


Fig. 2 Time limited bulk service.

ing to the FIFO principle.

2.1 Time Limited Bulk Service

The switching of variable-size packets through a synchronous switching fabric is not a trivial task. One solution is to segment incoming packets into smaller, fixed-size cells, which can then be easily transferred through the synchronous fabric. Cells are later re-assembled back into packets before leaving the switch.

In another solution, packets can be switched through a synchronous switching fabric without segmentation. In order to do this, the switching fabric time slot T must be at least as long as it takes to transfer the largest possible packet. This is a so-called TLBS (Time Limited Bulk Service). A synchronous VOQ packet switch with embedded TLBS functionality is referred to as a SB packet switch [20].

The idea of TLBS is that the largest possible packet determines the length of the time slot T . If C is the switching fabric's transmission bit rate, and B is the size of the largest possible packet, then the time-slot length is set to be B/C time units. It is obvious that in the case of smaller packets, more than one packet can be transferred during each time slot, as shown in Fig. 2. However, in most cases, the time slot cannot be fully utilized. If the transfer of the next packet is completed after the current time slot has ended, then this packet must be left for transfer in the next time slot. In [19], the maximum switch throughput analysis of a SB packet switch is derived under the assumption that the time-slot length T is fixed set to B/C seconds. It is shown analytically that the maximum switch throughput in this case cannot reach 100%.

2.2 Stretch Factor

In [21] a stretch factor SF is introduced, where $SF \geq 1$. The minimum allowed time-slot length B/C is multiplied by the SF in order to derive larger time-slot T , as depicted in Eq. (1). This enables higher SB packet switch utilization, as the relative percentage of unused slot space is decreased. The influence of enlarged time-slot length on overall SB packet switch performance is evaluated in [21] for uniform and homogenous synthetically generated traffic. It is shown that the maximum switch throughput increases toward 100% as the SF

factor increases. It is also established that only integer SF factor values must be used.

$$T = T_{min} \cdot SF = \frac{B}{C} \cdot SF, \quad SF \geq 1 \quad (1)$$

2.3 Scheduling

For each time slot in SB packet switches, only one queue at most, among the N of the same input port, can transmit packets. Moreover, at each time slot, an output port can receive packets from at most one input port. A set of N queues (chosen among the N^2 available) such that the previous conditions are met is defined as conflict-free. It is the task of a scheduling mechanism to select a conflict-free set of queues at the beginning of each time slot. There exist many scheduling mechanisms for IQ FIFO packet switches [22], [23]. In line with [20], the so-called OPF (Oldest Packet First) and ROT (Rotation) scheduling mechanism is selected for the simulations. As the name itself implies, in OPF scheduling the packets are served dynamically, according to the time they have spent inside the input queues—the longer the delay, the higher their priority. A bipartite graph method is usually used to select a conflict-free set of input queues at the beginning of each time slot. In ROT scheduling input queues are served periodically in a fixed set rotational manner.

3. Traffic Data

Traffic sources for the simulator are provided in the form of trace files. All trace files consist of one line per IP packet arrival. Each line contains five fields: consecutive number of a packet, time stamp, source IP address, destination IP address, and packet size. Each input port is fed by a separate trace file.

3.1 Traffic Flows

Determining how to make a map from real destination IP addresses to virtual addresses of packet switch output ports is not a trivial task. Flows are used in order to provide the transformation. That is, real traffic is characterized by a series of packets, which are generated by the same source in a relatively short period of time and have the same destination. Time period T_f is defined, which determines the maximum time gap between two consecutive packets having the same destination if they are to be treated as belonging to the same traffic flow. In simulations the value of T_f is set to 3 s.

Changing the value of T_f makes it possible to control the persistence of traffic flows. In a special case, when $T_f = 0$ s, flows are ignored. In this case there is no relation between the packet's destination IP address and its assigned output switch port. Incoming packets are uniformly distributed among all switch output

ports and thus output symmetry is achieved.

3.2 Capturing Real Traffic

Real traffic flows, conducted in five different network environments, were used.

Traffic data on ATM (Asynchronous Transfer Mode) cables were conducted at the Telekom Slovenije facility in April 2002. The HP BSTS (Broadband Series Test System) network analyzer was used, which has the ability to synchronously capture traffic running in both directions with 10 ns time stamping. An additional special tool for post processing and trace files generation was provided by Modsim III.

Traffic captured on this link is leased line aggregate traffic from various small and medium sized companies that are FRATM (Frame Relay over ATM) or ATM connected to the Internet service provider Siol, and from there to the Internet. Both directions were captured synchronously.

Traffic data on Ethernet cables was conducted at the Internet service provider Siol's facility in April 2002 and at the Faculty of Electrical Engineering in April and May 2002. Records were collected with Linux workstation and 100 Mbit/s Ethernet card. Traffic was captured using a TCPDump program with 1 μ s time stamping. An additional special tool for post processing and trace files generation was developed in Perl.

Two types of traffic were captured at the Siol facility. The first, denoted as Dial-Up, is traffic from an access router with four primary rate access links on the user side flowing towards the core network. The second, here denoted as Transit, is traffic generated by other smaller internet service providers, to which Siol provides connection to Internet.

At the Faculty of Electrical Engineering aggregate traffic from Metulj to the faculty network was captured. Metulj is an academic backbone network that connects all Ljubljana University members.

All time stamps are expressed to 8 places after the decimal point in the case of BSTS and to 6 places in the case of Linux, giving the appearance of 10 ns and 1 μ s resolution. The actual resolution, due to clock accuracy, jitter, and possible inner code loops, is 100 ns and 10 μ s respectively (or better). The later for Linux was confirmed by comparing its time stamping with a HP network analyzer with 100 Mbit/s Ethernet interface and 1 μ s resolution.

In addition, traces of core traffic from OC12c PoS (Packet over Synchronous optical network) interface with PPP (Point to Point Protocol) encapsulation on the AIX (Ames Internet Exchange) was used [24]. These traces were captured in June 2002. The accuracy of time stamp is 1 μ s.

Basic information about traces used in the simulations is shown in Table 1. Trace files are listed in the following format < day, month, year, hour, min.trc >.

Table 1 Traces used in simulations.

Name	Trace file	Link, Frame, Encapsulation	Average throughput [Mbit/s]	Average IAT [ms]	Number of destinations
Academic	220420021135.trc	Ethernet, 100 Mbit/s	4.162832	1.147229	1653
Core	AIX-1025225859-1.tsh.01.trc	OC12c, PoS, PPP	80.226788	0.056188	5531
Dial-Up	0904020958.trc	Ethernet, 100 Mbit/s	0.849022	3.102264	2172
Transit	100420020807.trc	Ethernet, 100 Mbit/s	15.342815	0.273883	5668
Leased Line	ip-ak02s3.trc	STM-1, ATM/AAL5, LLC/SNAP	21.990519	0.238568	333

Original trace files consist of more than 10^6 packets. In order to use them in the simulator, we broke them up into smaller, more manageable files of 10^5 packets. A sequence number was added at the end of these trace files. Poisson equivalent synthetic traffic trace files have P as the last letter in the filename. In the case of MWM (Multifractal Wavelet Model) equivalent synthetic trace files, the letter W is added.

To get a general idea about the traces' average throughput, average inter-arrival time, and number of destinations are also given for each of the first of the smaller cut up trace files.

3.3 Synthetic Traffic Generation

To generate synthetic trace files, real traffic trace files are used as a basis. Poisson equivalent traffic and wavelet based MWM equivalent traffic are generated.

Poisson traffic has been used extensively in the field of packet switch simulations due to its properties of being analytically defined and therefore simple to generate and use. Poisson models have served a mayor role in modeling classical circuit switched voice telecommunications and due to that they still represent a solid reference point (although they don't capture all the richness and complexity of contemporary real packet traffic) [25]. Two steps are required in order to produce equivalent synthetic Poisson traffic.

Real traffic trace file is analyzed in the first step. As a result the following parameters are retrieved: number of packets, average throughput, minimum, maximum, and average packet size, average packet inter-arrival time, and number of different destinations. In addition, a special filter vector is created for packet destination addresses. It is the purpose of the destination vector to enable proper distribution of synthetic packets (generated in the second step) between the list of available destination addresses. The filter vector is used in combination with uniform random distribution, in order to properly transform the original uniform distribution into a distribution, which is in accordance with the distribution of destination addresses in the original (real) trace file.

In the second step, a new synthetic traffic trace file is generated in such a way that all the above measured statistical parameters remain unchanged. New packet inter-arrival times are produced in a random manner by using negative exponential probability distribution

with a known mean. New packet sizes are also produced in a random manner. However, due to specified minimum and maximum packet size, a truncated exponential probability distribution is applied. Thus, statistical values for minimal, average, and maximal packet size are properly preserved. To select a destination IP address of a new packet, a uniform probability distribution is applied and then transformed by using a filter vector. The newly generated synthetic Poisson trace file thus retains the statistical properties of the original real trace file. Detailed steps to produce equivalent Poisson traffic are presented in [26].

Equivalent MWM synthetic traffic generation is implemented according to [27] where the scheme is derived to match the model to real data observations. In our case, only packet inter-arrival times are changed, whereas the rest of the data is used from the original trace.

3.4 Traffic Analysis

Matlab routines are used for traffic data analysis. Statistical properties of packet lengths and the gaps between packets in terms of histograms, cumulative distributions, burstiness, moments, autocorrelation functions, and scaling, including LRD (Long-Range Dependence), SS (Self-Similar) and multifractals are analyzed. In addition, statistics about the traffic passing through links on different time scales τ ranging from 100 μ s to 0.1 s is analyzed. In this paper, due to space limitations, we restrict our attention to just some of these.

Packet inter-arrival time and packet size histograms, in addition to cumulative packet size distribution from a sample of 5.6 s time period of Core traffic and its corresponding Poisson synthetic traffic trace, are shown in Fig. 3. Corresponding MWM equivalent traffic (not shown in the figure) yields an identical packet size histogram and its cumulative distribution as real traffic. The MWM IAT (Inter-arrival Time) histogram is also almost indistinguishable from real traffic. Figures (a) and (d) show linear-scale histograms of IATs. Figures (b) and (e) show linear-scale histograms of packet sizes. Figures (c) and (f) show the cumulative distribution plot for the same packet size data. As it is seen the IAT distribution of the real traffic model is roughly exponential, whereas IAT distribution of Poisson traffic is exponential, as explained above.

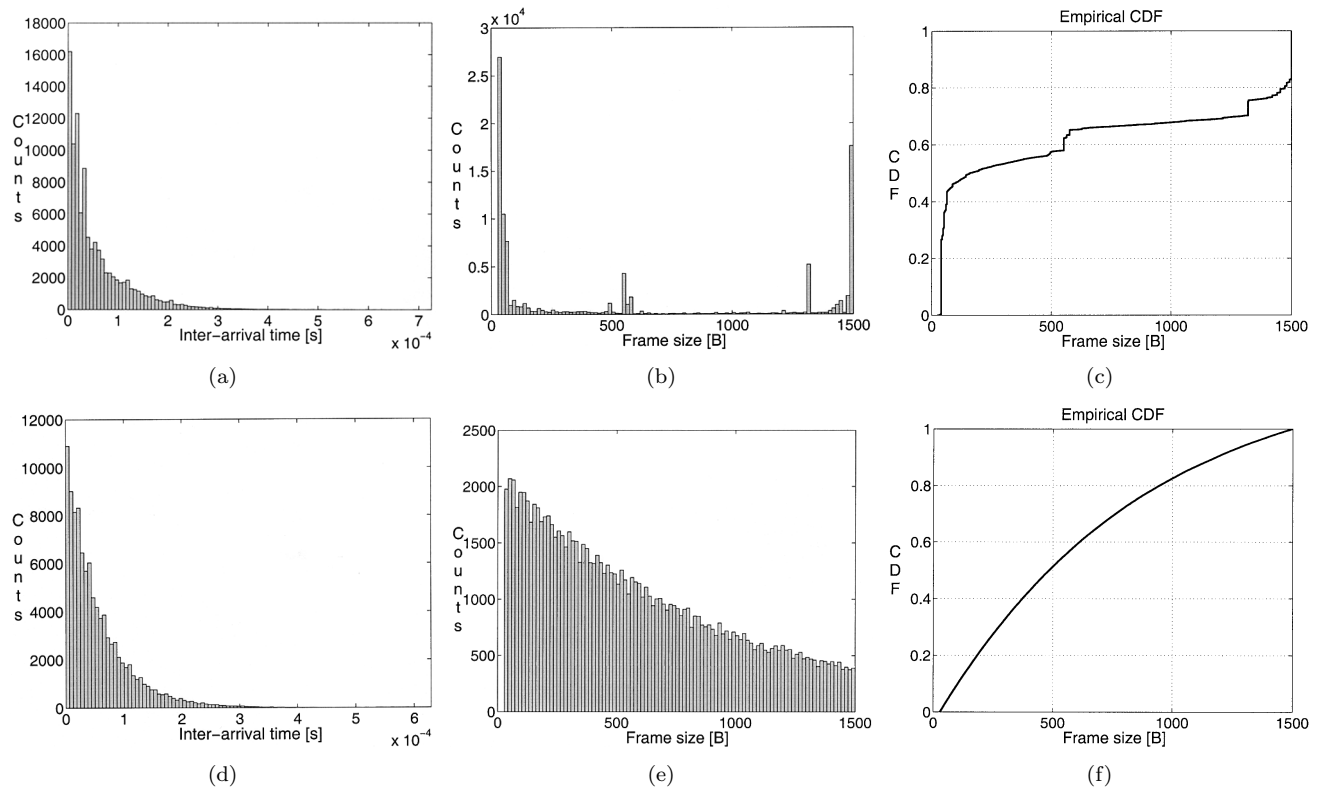


Fig. 3 Packet IAT histograms, packet size histograms and cumulative packet size distribution from a sample of Core traffic (a, b, c) and its corresponding synthetic Poisson traffic model (d, e, f).

Figures (a) and (b) for real traffic show the expected predominance of small packets, with peaks at the common sizes of 40, 44, 552, 576, and 1500 bytes. Almost 60% of the packets are smaller than 552 bytes. Nearly half of the packets are 40 to 44 bytes in length. All packets are 1504 bytes or smaller. The above characteristics were examined also for other traffic sources. No significant differences were found, except a minor one. For Transit traffic, a slightly larger percentage of 576 byte packets was found. This is the consequence of the fact that some older TCP (Transmission Control Protocol) implementations do not implement path MTU (Maximum Transmission Unit) discovery, yielding a 552 or 576 byte packet size for nonlocal IP destinations [28]. Therefore we can conclude that the behaviors in Fig. 3 are representative for all our traffic sources.

Autocorrelation and burstiness are also two important traffic characteristics. Burstiness for Academic, Core, Dial-Up, Transit, and Leased Line traffic is shown in Figs. 4(a), (b), (c), (d), and (e). Examples of MWM and Poisson models of Leased Line traffic are shown in Figs. (f) and (g). Burstiness is shown for different time scales. Every time scale is divided into 100 samples. Each point in the graph represents the average bit rate of the sample, normalized by average bit rate of that time scale. All graphs have the same ordinate axis scale to make comparisons easier.

It is evident that Academic traffic is the most bursty and Core traffic is the least bursty. Dial-Up and Leased Line traffic are not so bursty due to the fact that they are multiplexed from many sources with an upper limit of bandwidth. In the case of Dial-Up this limit was 64 kbit/s, whereas in the Leased Line case, all sources had 2 Mbit/s or less bandwidth available. As expected Transit traffic has behavior that is very similar to Core traffic behavior. The MWM model based on Leased Line traffic has burstiness similar to the original trace, whereas the Poisson model, as expected, shows good deal less burstiness. Intuitively, burstiness at higher time scales (long-term bursts) is more difficult for a packet switch to handle, than burstiness at lower time scales (short-term bursts). Smaller average bandwidth supports higher burstiness. On the other hand, a filled traffic link means lower burstiness.

Autocorrelation for the same traces as in Fig. 4 is shown in Fig. 5. The behavior of the autocorrelation function is also defined in various time scales. Traffic trace is cut into many smaller parts of time duration τ . Transmitted bits in each τ are calculated. Normalized autocorrelation function for lags from 0 to 5 is defined according to the formulae below [13].

$$r(\tau, \Delta t) = \frac{1}{N \cdot \sigma^2(\tau)} \sum_{i=1}^N \left(\Delta l_i(t_i, \tau) - m^{(1)}(\tau) \right)$$

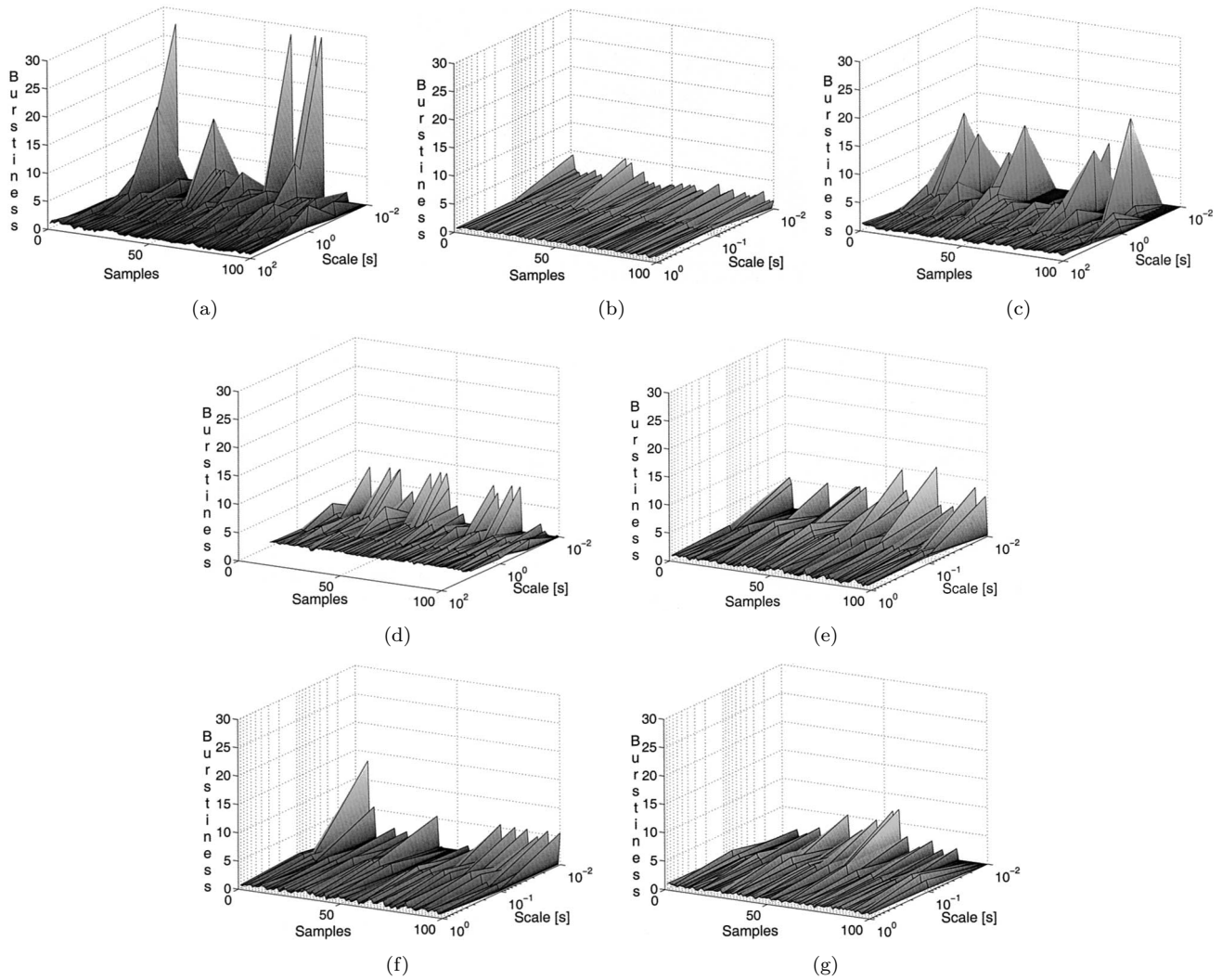


Fig. 4 Burstiness for Academic (a), Core (b), Dial-Up (c), Transit (d) and Leased Line (e) traces. Burstiness of MWM (f) and Poisson models (g) of Academic traffic.

$$\cdot \left(\Delta l_i(t_i + \Delta t, \tau) - m^{(1)}(\tau) \right) \quad (2)$$

The moments are calculated according to

$$m^{(k)}(\tau) = \frac{1}{N} \sum_{i=1}^N (\Delta l_i(t_i, \tau))^k \quad (3)$$

$$\sigma^2(\tau) = m^{(2)} - \left(m^{(1)} \right)^2 \quad (4)$$

where $\Delta l_i(t_i, \tau)$ is the amount of traffic transmitted during the i th time slot window from t_i to $t_i + \tau$ of time scale length τ . N is the number of observations.

Autocorrelation as defined above averages the process during each window τ . It measures how correlated these spaced apart windows are (in a window averaged sense) with increasing τ . The appropriate size of the window τ also depends on average throughput of each flow.

For Dial-Up, Transit, and the Poisson model, it

can be seen that the level of correlation even at one time lag does not exceed 0.1. In contrast, Core, Academic, Leased Line and the MWM model of Academic traffic show prolonged higher-levels of correlation. The MWM and Poisson models of traffic are representative of both extreme cases. The MWM equivalent traffic model shows strong correlations that monotonically increase with increasing time scale τ . On the other hand, packet arrivals in the Poisson equivalent traffic model show no correlations.

Scaling is an important parameter that says much about traffic behavior. Straight lines in logscale diagrams are evidence for the existence of scaling in analyzed data, while the range of scales involved gives information of its precise nature (self-similarity, long-memory). Estimation of scaling exponents was carried out by means of wavelet-based tools [5], [29]. Due to space limitations, second-order logscale and multiscale diagrams are not shown. Second order parameters with

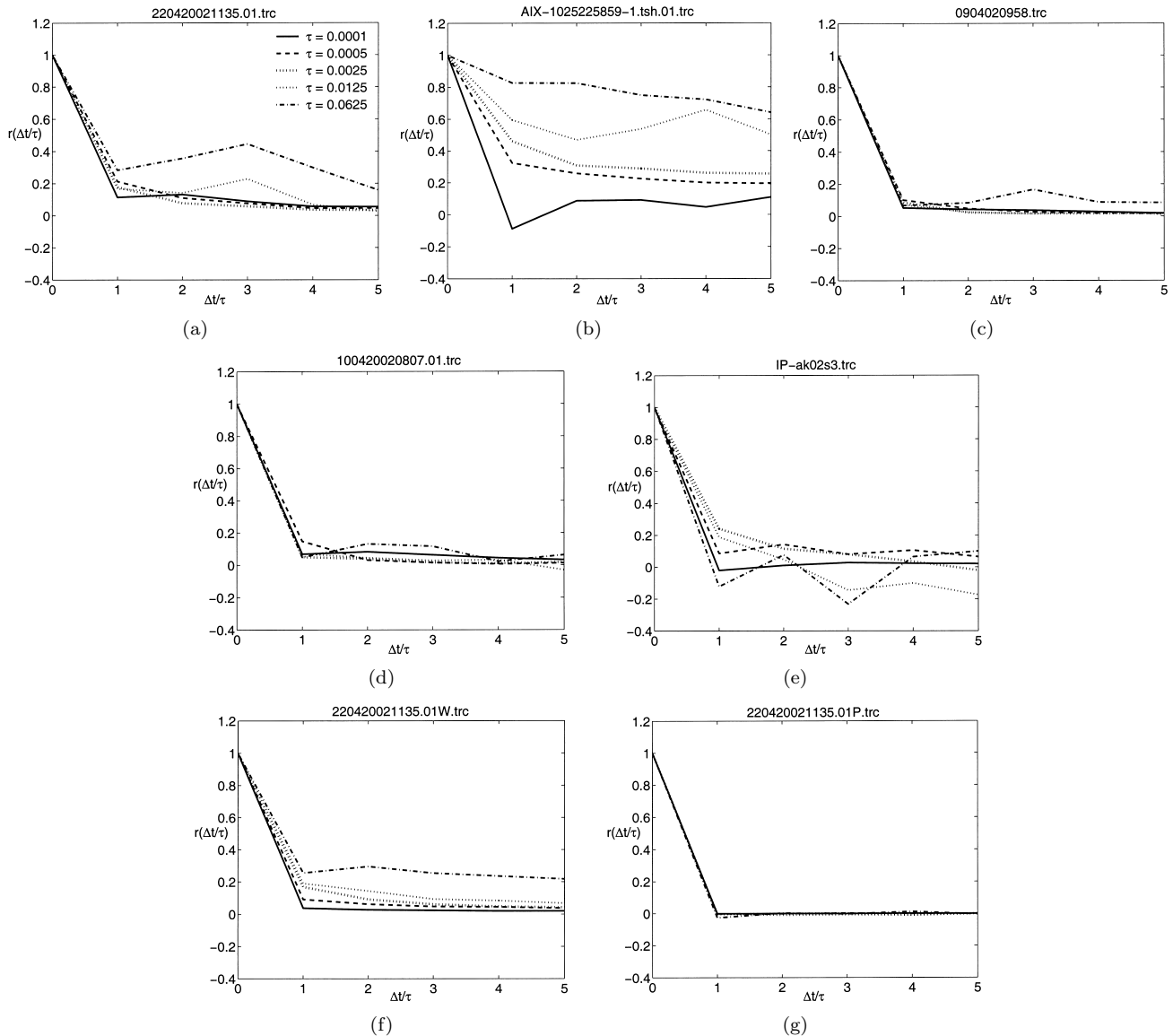


Fig. 5 Autocorrelation of Academic (a), Core (b), Dial-Up (c), Transit (d) and Leased Line (e) traces. Autocorrelation of MWM (f) and Poisson models (g) of Academic traffic.

corresponding confidence intervals and analysis ranges are shown in Table 2. Poisson and MWM equivalent models yield similar behavior for all traffic sources. Therefore only data for Academic traffic is shown.

4. Methodology and Simulation Environment

4.1 Performance Measures

It is our goal to determine the maximum throughput of a SB packet switch. It is defined as the sum of sizes of all received packets per outlet at the maximum offered load, divided by the duration of the simulation. The capacity of a switching fabric is determined by summarizing the average bit rates of all input trace files. A switch is thus automatically saturated with maximum

offered load during the simulation execution. Different input trace files, require for different switching fabric capacity settings.

4.2 Statistics

Simulation results are obtained by simulating the steady-state behavior of the SB packet switch. The initial statistical results are reset at the end of each warm-up period. The number of available packets in each simulation execution is limited by the number of lines in each input trace file. When the first input trace file becomes exhausted, simulation execution is terminated and final statistical calculations are performed.

Table 2 Estimated second order parameters.

Trace	H	95% CI	α	95% CI	Analysis range [j1, j2]
Academic	0.892	[0.843, 0.953]	0.784	[0.661, 0.907]	[8, 14]
Core	0.574	[0.543, 0.604]	0.147	[0.086, 0.208]	[6, 11]
Dial-Up	0.819	[0.788, 0.850]	0.637	[0.575, 0.700]	[6, 11]
Transit	0.582	[0.534, 0.631]	0.146	[0.067, 0.261]	[7, 11]
Leased Line	0.686	[0.559, 0.813]	0.372	[0.118, 0.625]	[8, 12]
Academic (MWM equivalent)	0.729	[0.672, 0.787]	0.459	[0.345, 0.573]	[8, 12]
Academic (Poisson equivalent)	0.500	[0.496, 0.505]	0.000	[-0.009, 0.009]	[1, 14]

4.3 Switch Configuration

A SB packet switch with 8 input and 8 output ports is used. The capacity of each FIFO buffer is set to 50 kB. The Simulator is controlled by a configuration file, which contains the necessary input parameters, including the names of input trace files.

8 trace files are used in each simulation execution. Each set of 8 captured real traffic trace files is supplemented by a set of 8 Poisson equivalent trace files and a set of 8 MWM equivalent trace files. Each simulation execution is repeated three times, once with each set of input trace files.

In combination with each set of input trace files, the interest is in the influence of SF parameter, scheduling, and traffic flow settings on the maximum switch throughput performance. All simulations are therefore repeated

- 5 times, with the following SF values: 1, 2, 3, 5, and 10,
- 2 times, once with OPF scheduling and once with ROT scheduling,
- 2 times, once with flows turned ON ($T_f = 3$ s) and once with flows turned OFF ($T_f = 0$ s).

To summarize the above, 20 consecutive simulation executions are performed with each set of input trace files. Furthermore, considering that three sets of input trace files are produced from one original set of real traffic trace files, it amounts to 60 consecutive simulation executions for each original set of captured real traffic trace files.

4.4 Input and Output Asymmetry

The switch load can be asymmetrically distributed among individual packet switch input and output ports. The interest is in long-term asymmetry, where the average bit rates at different input or output ports are compared.

Input asymmetry is caused by asymmetrically distributed input load. In the actual case of input asymmetry, comparison of average bit rates of trace files at different input ports shows considerable deviations from one trace file to another. Input asymmetry cannot be eliminated other than by using different input trace

files.

Asymmetrically distributed packets' destination addresses cause output asymmetry, which leads to saturation of individual output ports. Output asymmetry is only possible if traffic flows are enabled ($T_f > 0$ s). Output asymmetry can easily be eliminated by disabling the flows (setting $T_f = 0$ s).

5. Simulation Results

During simulation traffic sources are sent through the SB packet switch. Simulation results are presented in the form of 3D graphs. Each graph represents simulation results for one type of input traffic (Academic, Dial-Up, Leased Line, Transit, Core). Within each graph, the X -axis represents the type of simulation scenario. Each simulation scenario is presented in the form of a three-word segment. The first word segment indicates traffic subtype. RT stands for the original trace file, which was obtained by capturing real traffic. MWM stands for the MWM equivalent trace file, which was derived from the original trace file. Poiss represents the Poisson equivalent trace file, which was also generated on the basis of the statistical properties of the original trace file. The second word segment indicates the type of scheduling mechanism (either OPF or ROT) used by the SB packet switch during the simulation execution. The third word segment indicates whether traffic flows are accounted for (ON, $T_f = 3$ s) or not (OFF, $T_f = 0$ s). The Y -axis represents SF values, which are used by the SB packet switch during consecutive simulation executions. The following SF values are used: 1, 2, 3, 5, 10. The Z -axis represents measured switch throughput values, which are reported at the end of each simulation execution. Considering the SB packet switch is saturated with input traffic, these values represent the maximum switch throughput values.

5.1 Academic

Results for Academic traffic are presented in Fig. 6. First the results for original trace files (RT) are evaluated. The first thing to notice is a huge difference in SB packet switch performance, if different scheduling mechanisms are used and flows are turned ON. OPF scheduling is a clear winner in this case. On the other

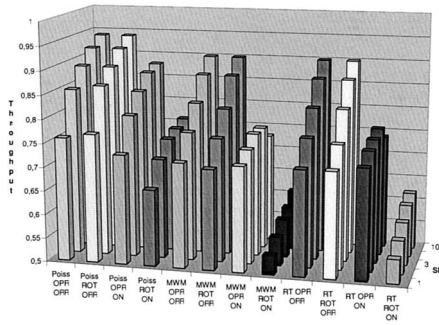


Fig. 6 Maximum switch throughput for Academic traffic source.

hand, when flows are turned OFF, both scheduling mechanisms yield approximately the same results. This is a clear indicator of input traffic having input symmetry (indicated by similar results in the case of flows being turned OFF) and output asymmetry (indicated by the different results in the case of flows being turned ON). Previously obtained maximum switch throughput results [19], [20] are independent of the selected switch scheduling mechanism. This is a consequence of using simple homogenous and uniform traffic sources. Our simulations reveal the effect of traffic asymmetry. It is clearly visible that in case of traffic asymmetry, different maximum switch throughput values are obtained by using different scheduling mechanisms. Indeed, by using an OPF scheduling mechanism 20% higher maximum switch throughput, as a result a 20% higher SB packet switch utilization is achieved.

The effect of increased SF factor value depends on the selected scheduling mechanism. In the case of ROT scheduling, it is useful to use high SF values. This is, however, not the case if OPF scheduling is used. The increment of SF value, although it does increase the possibility of better utilizing the individual connections through the switching fabric, also increases the time gaps between consecutive interventions of the OPF scheduling mechanism and thus reduces the possibility of necessary changes in switching fabric configuration.

The results of MWM equivalent input trace files are examined next. It is evident that these results are very similar to the results of original trace files. From this fact it can be concluded that MWM equivalent trace files represent a very good imitation of original real trace files.

The main difference when comparing results for Poisson equivalent trace files and results of original real trace files is in increased maximum switch throughput. This is a direct consequence of the decreased burstiness of Poisson equivalent trace files. The effect of reduced burstiness is especially noticeable in the case of the ROT scheduling mechanism, where the maximum switch throughput is increased by 10% when flows are turned ON.

Simulations confirm that Academic traffic is more

bursty than other traffic and has output asymmetry. It represents a typical example of access-level traffic.

Three main findings can be drawn from the above results. First, that in cases of such input traffic, the OPF scheduling mechanism is superior in comparison with the ROT scheduling mechanism, at least as far as maximum switch throughput capabilities are concerned. This is a completely new finding, not observed in previous simulations. Indeed, when using simple homogenous and uniform traffic sources, both scheduling mechanisms yield the same maximum switch throughput results [19], [20].

The second finding is that the increment of the SF parameter value does not necessarily result in the increment of the maximum switch throughput. This, too, is a completely new finding, not detected in previous simulations, when simple homogenous and uniform traffic sources were applied.

The third important finding is that in the case of output asymmetry and high level of burstiness, SF values that are too high can cause maximum switch throughput to degrade. According to these results it is advisable not to use excessively high SF values, e.g. $SF_{\max} = 5$.

The reason for problems with the SF factor can be seen by comparing results RT OPF ON and Poiss OPF ON. The difference between real and Poisson equivalent traffic is in the decreased burstiness of the later, thus high level of burstiness is one of the main reasons for ineffective, and sometimes even a destructive influence of the SF factor. This is especially noticeable when the OPF scheduling mechanism is used, which is by itself capable of providing for an effective scheduling procedure, thus eliminating the influence of traffic bursts. By using larger SF values, the OPF scheduling mechanism slowly loses control over the scheduling procedure, due to increased time gaps between subsequent scheduling executions. The same is not true if a simple ROT scheduling mechanism is used. In this case, the increase in SF value has a positive effect on maximum switch throughput results.

5.2 Dial-Up

Results for Dial-Up traffic are presented in Fig. 7. This traffic demonstrates the characteristics of access-level traffic, although it has reduced burstiness in comparison to Academic traffic. This is expected as Dial-Up represents a multiplex of traffic flows from a larger number of dial-up internet users.

In the case when flows are turned ON, the difference between maximum throughput results of the OPF and ROT scheduling mechanisms is evident and is approximately 10%. This difference is eliminated if flows are turned OFF. The reason is that traffic has input symmetry and output asymmetry. Again, the OPF scheduling proves to be more adaptable, capable

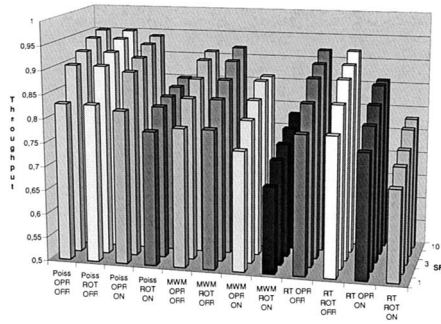


Fig. 7 Maximum switch throughput for Dial-Up traffic source.

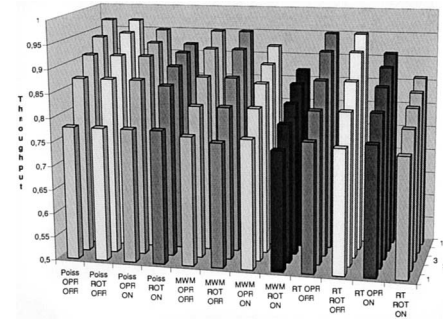


Fig. 9 Maximum switch throughput for Transit traffic source.

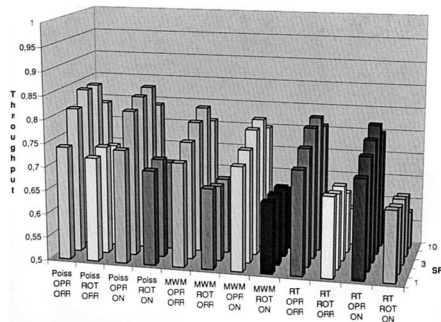


Fig. 8 Maximum switch throughput for Leased Line traffic source.

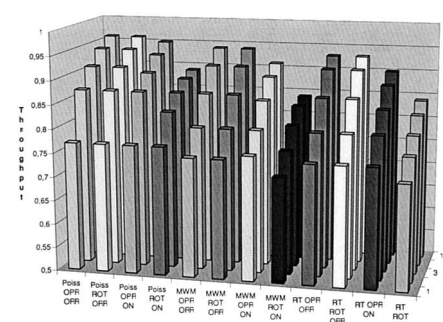


Fig. 10 Maximum switch throughput for Core traffic source.

of effectively scheduling asymmetric traffic flows.

The results of Poisson equivalent traffic reveal a slightly higher maximum throughput values in comparison to the results of real traffic, on the order of a few percents. This indicates a small impact of moderate burstiness, which is present in real traffic, but eliminated in Poisson equivalent traffic.

No negative influences from increased SF factor values are detected. Nevertheless, it is evident that values larger than 10 must not be used, since this does not necessarily result in increased maximum switch throughput, while it does lead to increased packet delay values.

5.3 Leased Line

Leased line traffic results are presented in Fig. 8. This figure is the most distinctive of the presented simulation results, due to traffic flows having strong input asymmetry.

In these difficult traffic conditions with distinctive input asymmetry, the OPF scheduling mechanism again proves to be a better choice, especially if used in combination with a SF factor value other than one. The effect of input asymmetry is so powerful that it is clearly visible even in case of Poisson equivalent traffic simulation results.

The use of the ROT scheduling mechanism is not recommended, especially not in combination with larger SF values. In any case, the use of SF values

larger than 5 must be eliminated, as they degrade switch throughput characteristics ($SF_{\max} = 5$).

The effect of output asymmetry is small, as there is hardly any difference when comparing the results with flows turned ON and OFF. The effect of burstiness is negligible if OPF scheduling is used. On the other hand, a 5% difference in switch throughput performance is detected if ROT scheduling is used.

5.4 Transit

The results for Transit traffic are presented in Fig. 9. Real traffic sources with flows turned ON are the most important results in this figure. The OPF scheduling mechanism provides for higher maximum switch throughput in comparison to the ROT scheduling mechanism. However, the throughput difference is much smaller in comparison to the difference in results in the case of Academic traffic sources. Maximum throughput results are almost identical in cases when flows are turned OFF. Traffic sources thus have input symmetry and a limited degree of output asymmetry.

The switch throughput results of Poisson equivalent trace files are very similar to the results of real trace files. This is an important finding, as it reveals that the burstiness and self-similar properties of backbone traffic tend to decrease. Such traffic can then more easily be simulated by using Poisson equivalent traffic. This fact also vindicates for the use of Poisson-based analytical models.

An important finding is that in the case of Transit traffic, there is no negative effect of increased SF parameter value. The increment in the SF parameter value always results in the increase of maximum switch throughput value. Therefore, high SF values can be used in order to achieve high switch throughput results. Transmission speeds are high in backbone networks, which require large switch processing capabilities in order to execute scheduling mechanism in each cycle. By increasing the SF value, the time gap between two consecutive executions of scheduling mechanism is increased. The processing requirements for the SB packet switch are thus decreased by factor SF , e.g. one tenth of original processing requirements is sufficient when SF is set at 10. By increasing the SF parameter value, packet delay values also increase. However, the absolute value of delay is kept low, due to high transmission speeds in backbone networks.

5.5 Core

The results for Core traffic are presented in Fig. 10. Simulation results are similar to the results of Transit traffic. Both Core and Transit traffic demonstrate a high level of input and output symmetry. They can both be classified as backbone-level traffic. There is a difference in the level of burstiness, which is higher in the case of Transit traffic. Nevertheless, the absolute level of burstiness in both cases is low compared to other traffic sources.

Considering the above findings, the results of real backbone traffic flows are very similar to the results of Poisson equivalent traffic flows. This is in line with [12], [13]. The influence of increased SF value is favorable in all simulated cases. An interesting point is that the influence of slight output asymmetry is best demonstrated by a moderate decreased influence of increased SF value on the maximum switch throughput. From that it is evident that traffic asymmetry reduces the effectiveness of the SF factor.

6. Conclusions

Real traffic flows were captured in various network environments and their statistical properties were analyzed. Based on real traffic flows, MWM and Poisson equivalent synthetic traffic flows were generated. These are important as a validation points when evaluating packet switch performance results under real traffic conditions.

Maximum throughput performance of the SB packet switch was evaluated using two steps. In the first step, maximum throughput was obtained using various real traffic flows. In the second step, corresponding maximum throughput of the same SB packet switch was obtained using MWM and Poisson equivalent traffic flows.

Different SF parameter values, scheduling mechanisms, and flow settings were applied in order to evaluate the influence of different SB packet switch settings on its performance.

Traffic asymmetry, either input or output based, proved to have the greatest influence on SB packet switch performance. By increasing the level of asymmetry, maximum throughput values decrease considerably, especially if the ROT scheduling mechanism is used. Traffic asymmetry also decreases the influence of the SF parameter on maximum switch throughput. Large SF values can even cause maximum switch throughput to decrease. Special care must be taken when ROT scheduling is used. As a general rule of thumb, SF values of no more than 5 must be used if asymmetrical traffic is being relayed across the SB packet switch. It is also advisable that the OPF scheduling mechanism is used in such cases, as it is capable of, at least partially, compensating for traffic asymmetry.

An interesting finding is that the influence of burstiness and scaling of traffic flows is less significant for the SB packet switch maximum throughput results, if the OPF scheduling mechanism is being used. The maximum measured difference, in comparison to Poisson equivalent traffic flows, was in this case on the order of a few percents. Larger discrepancies were detected, if ROT scheduling was used. The differences here can amount to approximately 10%. However, it must be emphasized that the increased level of burstiness has substantial influence on both packet delay values and packet loss ratio [3], [15], [30].

Whatever configuration is used, SB packet switches proved to be suitable for switching backbone-level traffic flows. From the packet switch's point of view, these traffic flows exhibit Poisson-like statistical characteristics and satisfactory input and output symmetry. High transmission rates are used in backbone networks, which demands a great deal of switch processing power. In these cases, the ROT scheduling mechanism, in combination with larger SF values, e.g. 5 or 10, is advisable. Maximum switch throughput values of more than 80% can then be expected. In backbone-level networks the influence of large SF parameter values on packet delay is compensated for by high transmission speeds. Maximum switch throughput values can in addition be increased by approximately 5% if the OPF scheduling mechanism is used. This also increases the resilience of SB packet switch on possible traffic flow anomalies, at the cost of additional switch processing power.

SB packet switches can also be used for switching access-level traffic flows. In this case, however, the OPF scheduling mechanism is advisable and the maximum value of the SF parameter should not exceed the value of 5. This provides for resilience of the SB packet switch on common access-level traffic flow anomalies, such as traffic asymmetry, scaling and high level of burstiness.

Maximum switch throughput of approximately 75% can be expected in this case. The larger processing requirements of the OPF packet scheduling mechanism are not critical due to relatively low transmission speeds in access-level networks. Also, due to low transmission speeds, larger SF parameter values should be avoided in order to keep packet delay values as low as possible.

References

- [1] P. Baran, "The beginnings of packet switching: Some underlying concepts," *IEEE Commun. Mag.*, vol.40, no.7, pp.42–48, 2002.
- [2] W. Stallings, *High-speed Networks: TCP/IP and ATM Design Principles*, Prentice-Hall, 1998.
- [3] B. Tsybakov and N.D. Georganas, "On self-similar traffic in ATM queues: Definitions, overflow probability bound and cell delay distribution," *IEEE/ACM Trans. Networking*, vol.5, no.3, pp.397–408, 1997.
- [4] H. Hlavacs, G. Kotsis, and C. Steinkellner, "Traffic source modelling," Technical Report TR-99101, Institute of Applied Computer Science and Information Systems, University of Vienna, 1999.
- [5] P. Abry, R. Baraniuk, P. Flandrin, R. Reidi, and D. Veitch, "Multiscale nature of network traffic," *IEEE Signal. Process. Mag.*, vol.19, no.3, pp.28–46, 2002.
- [6] O. Cappe, E. Moulines, J.C. Pesquet, A. Petropolu, and X. Yang, "Long-range dependence and heavy-tail modeling for teletraffic data," *IEEE Signal. Process. Mag.*, vol.19, no.3, pp.14–27, 2002.
- [7] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol.2, no.1, pp.1–15, 1994.
- [8] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson," *IEEE/ACM Trans. Networking*, vol.3, no.3, pp.226–244, 1995.
- [9] J. Beran, R. Sherman, M.S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Trans. Commun.*, vol.43, pp.1566–1579, Feb./March/April 1995.
- [10] M.E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic evidence and possible causes," *IEEE/ACM Trans. Networking*, vol.5, no.6, pp.835–846, 1997.
- [11] A. Feldmann, A. Gilbert, and W. Willinger, "Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic," *Proc. ACM SIGCOMM'98*, pp.43–55, 1998.
- [12] J. Cao, W.S. Cleveland, D. Lin, and D.X. Sun, "Internet traffic tends toward Poisson and independent as the load increases, in *Nonlinear Estimation and Classification*, Springer, New York, 2002.
- [13] A. Stepanenko, C.C. Constantinou, T.N. Arvanitis, and K. Baughan, "Statistical properties of core network Internet traffic," *Electron. Lett.*, vol.38, no.7, pp.350–351, 2002.
- [14] Z. Sahinoglu and S. Tekinay, "On multimedia networks: Self-similar traffic and network performance," *IEEE Commun. Mag.*, vol.37, no.1, pp.48–52, 1999.
- [15] P. Homan, J. Bešter, A. Kos, and T. Slivnik, "The impact of bursty traffic on FPCF packet switch performance," article in press, corrected proof, *Computer Communications*, available online 25, Oct. 2002.
- [16] M. Schwartz, *Broadband Integrated Networks*, Prentice-Hall PTR, New York, 1996.
- [17] D.J. Ma, "The maximum throughput of a nonblocking packet switch with window policy," *IEICE Trans. Commun.*, vol.E80-B, no.4, pp.537–580, April 1997.
- [18] J. Turner and N. Yamanaka, "Architectural choices in large scale ATM switches," *IEICE Trans. Commun.*, vol.E81-B, no.2, pp.120–136, Feb. 1998.
- [19] P. Giacomazzi, "Maximum throughput analysis of a datagram switch for broadband networks," *IEICE Trans. Commun.*, vol.E81-B, no.2, pp.354–362, Feb. 1998.
- [20] P. Homan and J. Bešter, "Enhanced synchronous packet switching for IP packets," *IEICE Trans. Commun.*, vol.E85-B, no.1, pp.247–256, Jan. 2002.
- [21] P. Homan, *Configuration and Control of Queues for Assuring Quality of Telecommunication Services in Packet Switches*, PhD Thesis, University of Ljubljana, 2001, in Slovene language.
- [22] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," *IEEE Trans. Commun.*, vol.47, no.8, pp.1260–1267, 1999.
- [23] N. McKeown and T.E. Anderson, "A quantitative comparison of scheduling algorithms for input-queued switches," *Comput. Netw.*, vol.30, no.24, pp.2309–2326, 1998.
- [24] <http://pma.nlanr.net/PMA/>
- [25] N.W. Macfadyen, "Traffic characterisation and modelling," *BT Technology Journal*, vol.20, no.3, pp.14–30, 2002.
- [26] P. Homan, A. Kos, and J. Bešter, "Poisson equivalent of real packet traffic," *Proc. Electrotechnical and Computer Conference, ERK 2002*, pp.107–110, Portorož, Slovenia, 2002.
- [27] R.H. Riedi, M.S. Crouse, V.J. Ribeiro, and R.G. Baraniuk, "A multifractal wavelet model with application to network traffic," *IEEE Trans. Inf. Theory*, vol.45, no.3, pp.1–51, 1999.
- [28] K. Thompson, G.J. Miller, and R. Wilder, "Wide area Internet traffic patterns and characteristics," *IEEE Network*, vol.11, no.6, pp.10–23, 1997.
- [29] D. Veitch and P. Abry, "A wavelet based joint estimator of the parameters of long-range dependence," *IEEE Trans. Inf. Theory*, vol.45, no.3, pp.878–897, 1999.
- [30] T. Hagiwara, H. Majima, T. Matsuda, and M. Yamamoto, "Impact of packet delay self-similarity on TCP performance," *IEICE Trans. Commun.*, vol.E84-B, no.11, pp.1–4, Nov. 2001.



Andrej Kos graduated from the University of Ljubljana, Slovenia, in 1996 and was awarded his M.Sc. degree in the field of telecommunications in 1999. At present he is employed as a Senior Researcher at the Faculty of Electrical Engineering in Ljubljana and is currently working on his Ph.D. He has extensive research and industrial experience in the analysis, modelling, and design of advanced telecommunications elements, networks, and systems. His current work and research focuses on next generation packet switching and management. His other fields of expertise include mobile communications and intelligent networks. He is a member of the IEEE and Telemanagement Forum.



Peter Homan graduated in 1995. He received his M.Sc. and Ph.D. degrees in the field of telecommunications from the University of Ljubljana, Slovenia, in 1998 and 2001, respectively. At present, he is employed as a Senior Researcher at the Faculty of Electrical Engineering in Ljubljana. His previous work has involved research, analysis, modeling and design of advanced telecommunications systems and networks, including the development

of telecommunication software. His current work and study focuses on next generation broadband networks, with emphasis on switching and quality of service issues. Peter Homan is a member of the IEEE.



Janez Bešter graduated in 1979. He received his M.Sc. and Ph.D. degrees in the field of telecommunications from the University of Ljubljana, Slovenia, in 1982 and 1995, respectively. Currently, he is Head of the Laboratory for Telecommunications and Assistant Professor at the Faculty of Electrical Engineering in Ljubljana. His twenty years of research and development activities, as well as teaching, have focused on the field of planning,

realization and management of telecommunication systems and services, together with the application of information technologies in education. Janez Bešter is a member of the IEEE and ACM.

Article

Developing a multilingual, personalised medical library portal: use of MyLibrary in Slovenia

*Anamarija Rozic-Hristovski,
Iztok Humar and Dimitar Hristovski*

The authors

Anamarija Rozic-Hristovski is Director of Central Medical Library, Faculty of Medicine, University of Ljubljana, Slovenia. E-mail: anamarija.rozic-hristovski@mf.uni-lj.si

Iztok Humar is Teaching Assistant at the Faculty of Electrical Engineering, University of Ljubljana, Slovenia. E-mail: iztok.humar@fe.uni-lj.si

Dimitar Hristovski is Research Assistant at the Institute of Biomedical Informatics, Faculty of Medicine, University of Ljubljana, Slovenia. E-mail: dimitar.hristovski@mf.uni-lj.si

Keywords

Medical libraries, Libraries, Portals, Slovenia

Abstract

The Central Medical Library (CMK) at the Faculty of Medicine of the University of Ljubljana, Slovenia, created its Web site in 1997 and has since then been actively involved in its maintenance and improvement. The analysis of Web site usage, which was performed using different data-mining methods, revealed an increasing interest among the Slovene biomedical community. Patrons have often complained to librarians about information overload and difficulty in following all the Web site's enhancements. This situation called for a dynamic restructuring of the CMK Web site. Such problems are often solved with customisable and personalised library portals. Having limited financial and human resources, it was decided to customise and extend an existing library portal software solution (MyLibrary from North Carolina State University). The MyLibrary portal for the CMK was not only customised, but some important new functionality, most notably multilingual support, was added. This experience of developing a personalised library portal could be of interest to most libraries that offer information in at least two languages.

Electronic access

The Emerald Research Register for this journal is available at
<http://www.emeraldinsight.com/researchregister>

The current issue and full text archive of this journal is available at
<http://www.emeraldinsight.com/0033-0337.htm>

Introduction

Academic medical libraries provide information resources and services to students, faculty, health professionals and researchers in an environment that supports learning, teaching, health care and research. The rapid development of information technology is transforming library services at a spectacular rate. By connecting to the Internet and World Wide Web, the academic library expands its access to information resources, some of which appear as a digital library. According to Lesk (1997):

Digital libraries are organised collections of digital information. They combine the structuring and gathering of information, which libraries and archives have always done, with the digital representation that computers have made possible.

Due to technology advances and the wide dissemination of information, users suffer from information overload and expect their library to select the most valuable information and organise it effectively for their personal consumption. Top LITA (Library and Information Technology Association, a division of the American Library Association) experts identified a number of important future trends for technologies in libraries (LITA Top Tech Trends, 1999). They identified Trend 1 as "library users who are Web users, a growing group, expect customisation, interactivity, and customer support". They stated that approaches that were library-focused instead of user-focused would be increasingly irrelevant and mentioned The University of Washington's MyGateway (www.lib.washington.edu/resource/help/MyGateway.html) and North Carolina State University's MyLibrary@NCState (<http://my.lib.ncsu.edu/>) are examples of customised library portals.

A Web portal, also commonly called a gateway, is now the standard interface to aggregate library resources and services through a single access and management point for users. The concepts of personalisation and customisation have gained interest among librarians, especially those at universities and institutions with

The authors are grateful to Mr Gaj Vidmar of the Institute of Biomedical Informatics of the University of Ljubljana for reading the manuscript and improving its language and style.

large collections, growing numbers of online resources, and user populations with varying levels of computer and research expertise. Customisable and personalised library portals tend to be customer focused, responding to customer needs, and empowering users to create personal information systems that are responsive to their individual needs (Lakos and Gray, 2000).

Because there are no clear and unambiguous definitions of the concepts “personalisation” and “customisation” in the context of Web portals, we use both terms interchangeably or together to describe the ability of a Web portal:

- to deliver content based on the profile a user creates;
- to allow the user to change her/his profile; and
- to analyse the user’s behaviour and provide different content to each user.

Library portals have been defined as “systems which gather a variety of useful information resources into a single, one-stop Web page, helping the user to avoid being overwhelmed by ‘infoglut’ or feeling lost on the Web” (Looney and Lyman, 2000). Portals are user-centric, while home pages are owner-centric. Portals allow the users to customise their Web information based on personal preferences and the librarians to develop customised gateways to specialised resources of value. Customised Web portals can bring to light the range of resources being offered by a library and will allow the user to create a personalised service.

The MyLibrary option appearing on many library portals is part of a larger trend towards customisation and personalisation. Consumers are being trained to expect that the world will be focused on their needs, wants, and desires. This expectation translates into MyThis and MyThat on the Web. Implementers of MyLibrary projects seek to provide a customised view of the library that reduces information overload and administrative overheads. Some early adopters of MyLibrary concept are:

- University of Washington
(www.lib.washington.edu/resource/help/MyGateway.html);
- NCSU – North Carolina State University
(<http://my.lib.ncsu.edu/>);
- Cornell University
(<http://mylibrary.cornell.edu>);

- University of Utah
(<http://medstat.med.utah.edu/library/personalize.html>); and
- New York University
(<http://library.med.nyu.edu/library/mlsetup.html>).

Many developed a system in direct response to cries of information excess by users encountering a megalibrary with thousands of resources. MyLibrary appeals to individual users to create a personally relevant library (Ketchell, 2000).

Customisation is only one part of the MyLibrary trend. A library portal must be more than a method to select from a set of links; it must also include database and application windows.

Scope

The Central Medical Library (CMK) is a department of the Faculty of Medicine of the University of Ljubljana, Slovenia. It is the main biomedical library in the country and serves students, faculty, health professionals, researchers and laymen. CMK started to build a library Web site (www.mf.uni-lj.si/cmk/) which included a guide to library services and resources in 1997 (Rozic-Hristovski *et al.*, 1999). Since then, the contents of the Web site have doubled, users have become more aware of its importance, and users’ computer literacy, computer equipment and Internet connections have improved significantly.

The planned content of the Web site had crucially influenced the decisions made regarding its structure. The CMK Web site is built as an information entity embedded in uniform graphic design that encompasses three levels of menus, two levels of headers, the footer and background. It is possible to choose between eight sub-menus that provide some key information needed for effective use of CMK and access to information resources. Desired functionality is obtained by a menu/hypertext link structure with two frames displaying the current menu and active text, whereby information screens and hypertext links allow cross-menu navigation. It is possible to scroll each frame independently of the other. That enables the user to change quickly between different options inside a menu.

In 1998 and 1999 staff at CMK performed an analysis of Web site usage behaviour to obtain concrete knowledge about the way visitors navigate the Web site (Rozic-Hristovski *et al.*, 2002). The Web site access evaluation was conducted by analysing the CMK Web-server log-files. The results revealed that the visitors most frequently started and ended their visits on the CMK home page. Reference pages with lists of printed and electronic information resources were also frequent starting and ending points in exploring the Web site. It seemed necessary to ensure greater visibility of these pages and a more convenient navigation path to them. A more intuitive design for some pages was also needed so that visitors would be able to access information more quickly and easily with fewer clicks. Log-file analysis adequately revealed overall usage patterns, but could only provide rough estimates of individual user characteristics. Because of the problems with log-file analysis we were not able to discover the identity of many frequent users.

In 2000 we decided to explore user navigation patterns in the CMK Web site using a Web usage mining tool Web utilisation miner (WUM) (Spiliopoulou, 2000). We found several interesting patterns, some of which came as a surprise to the Web site designer (Rozic-Hristovski *et al.*, 2001). The discovered navigation patterns gave us a better understanding of our users and their needs, as well as an opportunity to redesign and optimise our site according to the users' preferences. Unfortunately, very frequent requests for links to external Web sites such as electronic journals, databases and search engines were not recorded in the Web log file. Thus, we were unable to analyse the paths to these links.

The results of our studies revealed that users have very diverse interests. The most visited Web pages were dispersed among different sub-menus and subject categories. Hence, the users probably found the whole Web site to be of potential interest. Today the users have to cope with the rapidly expanding digital library and continuous improvement of library services. Such a situation demanded a dynamic restructuring of the CMK Web site because of limitations of a static Web design. Often such problems are solved with customisable and personalised library portals because they address the modern library role directly by being customer focused, by

responding to customer needs and by empowering users to create personal information systems that are responsive to their individual needs. That is why we started to think about a personalised medical library portal. As we have limited financial and human resources, we decided to customise and extend an existing library portal software solution.

We started a pilot project with the aim of selecting the most appropriate portal software from the higher education community. We were looking for a solution distributed under the GNU (GNU's not Unix – see www.gnu.org) public licence as open source code and wanted to customise it to the needs of MyLibrary CMK.

CMK portal design

The number of the CMK Web site visitors is steadily growing and it is approaching 10,000 per year. The majority of them are from Slovenia, while some are from abroad. The patrons from the Faculty of Medicine usually use our Web site most heavily. At the moment the Web site comprises a large number of pages on general information and services, more than 1,500 electronic journals and about 500 links to Internet resources by topic. Web site maintenance and development is supported by a part-time electrical engineer and two librarians, in occasional collaboration with other staff who simultaneously take various other responsibilities.

In such a library environment, considering the results of our analyses, we decided to offer our users a personalised medical library portal. Because of limited human and financial resources we wanted to adopt an existing library portal solution that best suited our needs and the given situation. We wanted the portal to enable MyLibrary features and a powerful search engine. The portal should support the main functions of an academic medical library and its users' needs. Patrons should have the possibility to select their favourite journals, Internet resources and frequently used library links. The number of clicks and time needed to locate a variety of library services and resources should be minimised. Identifying frequent users and reviewing their needs and preferences using registration forms and cookies was perceived as an important portal function. We expected

to establish a connection with the Slovenian national online public library catalogue, COBISS, and establish lists of journals and new books which would facilitate a current awareness service and save staff time (Seljak and Seljak, 2002).

The difference between a static Web site, used only for delivering information via the World Wide Web, and a portal is that the latter presents a standard interface to aggregate the library's resources and services through a single access and management point for the users. Using dynamic page generation, personalisation and customisation, a portal can help users to avoid being overwhelmed with information. Technically speaking, a library portal is a database application, accessible from a Web interface, and governed by the principles and practices of librarianship. The distinctive element of such a system is not the technology driving it but the interactive assistance and services it provides via librarians.

In the first phase of the portal design process, we had to find out how to customise the existing portal sections to the scope of CMK resources and services, and determine which information needed to be personalised to meet the needs of library patrons. We can divide the information that the library provides into two types. Static information, such as opening hours, service charges, location, staff information and circulation policy, does not change often and can be a part of a static Web page. The second type is the library information resources that grow permanently and therefore change very often. This type of information has to be updated simply by a librarian, without interaction of a system administrator, whereby all the portal sections containing information not of interest to all users require the personalisation approach.

If a library decides to provide personalised access to its resources and services, nowadays the system no longer needs to be developed from scratch. Several software projects of this kind have been developed and some of them are willing to share their code. Technical infrastructure requirements – hardware, software and needs for local configuration and customisation – may vary considerably, although when deploying an already developed system, the majority of the hard work has already been done. Some of these projects are quite sophisticated, but

unfortunately some of them are quite dependent on, and linked to, the library's online catalogue system, which means different standards in different countries. Customisation of such systems may require more than just an installation and deployment of readily developed portal software, while the development of the library-dependent intersystem interfaces could take considerable time and demand significant programming experience.

During our platform-choosing process we reviewed the documentation of the systems listed below.

- MyLibrary service at Virginia Commonwealth University (Gharphery and Ream, 2000);
- My Gateway system at the University of Washington (Jordan, 2000);
- MyLibrary personalised electronic services in the Cornell University Library (Cohen *et al.*, 2000);
- MyLibrary project at Los Alamos National Laboratory (Di Giacomo *et al.*, 2001); and
- MyLibrary system at North Carolina State University (Morgan, 1999).

We decided to select MyLibrary@NCState using the short review of these systems, provided below, which we made during our decision-making process.

MyLibrary service at Virginia Commonwealth University (MyLibrary@VCU) is fairly simple both in presentation and infrastructure. Users select items for display from an initial set-up screen listing possible resources, and can later adjust them by clicking the edit button on the personalised page. On the server side, a Perl program is used to store all the user passwords, logins and e-mail addresses in a single text file. Text files are also used to store information about each user's profile. The system can be ported to various platforms. The source code is free but still requires localisation and customisation for use within the library.

My Gateway system at the University of Washington offers an extensive array of resources. Users can add and delete categories of resources using the "category manager". The URLs can be drawn from the library's URL database that resides in the "digital registry" or can be the users' own URLs. The records of the resources are

loaded from the library's INNOPAC system into My Gateway's local database, which enables resources to have far more detailed records than resources at competing systems. The server side is implemented using Microsoft technology. Server side scripting is done by Active Server Pages (ASP), used for dynamic generation of HTML pages using the data stored in the Microsoft SQL server database. The operating system must be a member of the Microsoft server family.

MyLibrary personalised electronic services from Cornell University Library consists of several tools for collecting and organising resources (MyLinks) and informing users of new resources provided by the library (MyUpdates). Users can sort the library's and their own links into folders, accessible from everywhere with only a Web browser. The system is built in a way that encourages users to communicate with the library and thus reduces common pitfalls in library-patron communication. The system enables offline integration with the NOTIS, the online mainframe-based catalogue software used at Cornell. The technology behind the system uses an Oracle database for storing information, Kerberos for authentication of user identification and password combination, and Java servlets for implementing access to the database over JDBC (Java database connectivity) and providing dynamic Web content.

The MyLibrary project at Los Alamos National Laboratory (MyLibrary@LANL) provides digital library users (as individuals or groups) with a personalised Web environment while offering tools that extend the functionality of a Web browser (bookmarklet). The personalised fields consist of the list of databases, electronic journals, general reference, Web resources as well as personal Web links recommender system integration and Web link checking mechanism. The technical solution adopted for MyLibrary@LANL is object-oriented Perl and Javascript, and the database used is Open Source SQL Database MySQL. The operating system is Linux, but software should be portable to other Unix-originated systems. The system is complex and consists of several modules: such as library manager, folder manager, link manager, database manager, style sheet manager and bookmarklet manager.

We found the MyLibrary system at North Carolina State University (MyLibrary@NCState) to be the most suitable application for the previously mentioned needs of our library. It is very well documented, open source (easily upgradeable) and freely available on the Web. A more detailed description follows below.

MyLibrary@NCState: system description

The MyLibrary system at North Carolina State University (NCState) is probably the best-documented implementation of a user-centred, customisable interface to the library's collection of information resources. It is well developed and offers an intuitive interface for both users and administrators. MyLibrary@NCState as an entry point offers its users several information and resource collections divided into the following sections:

- Global message;
- Message from librarian;
- Your librarians;
- Library and university links;
- Bibliographic databases;
- Electronic journals; and
- Reference shelf.

Users have an opportunity to select or deselect items they want to have displayed in these sections on the starting screen using the "customise" option. The last three sections are discipline dependent. This allows users quicker access to the resources from the field they are particularly interested in. Once configured, MyLibrary@NCState can store information about the user on the local computer using Web cookies. This way, the user does not have to login repeatedly unless he or she wants to use previously defined settings on a different computer.

The administrative part of the system, which is also very important from the library's point of view, is divided into four categories covering functions for:

- (1) content providers – librarians;
- (2) Webmasters;
- (3) system administrators; and
- (4) report generation.

Content management functions allow authorised librarians to review, add, delete or change the current content, such as list of disciplines, names, links and descriptions of databases, journals, libraries and universities, the names and contact information of librarians and texts provided by the system.

The administration system provides creation of reports on broken links and is used for reviewing items of all resources linked with a specific discipline. Webmasters are provided with a powerful system, which uses a proprietary scripting language for describing the data that reside in the MyLibrary's database. Web page design, such as changing background and text colours or selections arrangements, can also be done through the Webmaster part of the administration system. The appearance of the system can be controlled via cascading style sheet commands directly entered into forms provided in the Webmaster sub-menu. The system administrators have overall system control by means of setting the global preferences, defining new disciplines, assigning librarians to disciplines and controlling the texts used for user-system communication.

It is evident that this system is much more than just a bookmark manager. It allows users to regularly receive and search lists of new books added to the library's collections. The user selects the preferred discipline and, according to that choice, he/she is referred to an appropriate librarian and a collection manager who is specialised in the given subject area. Since library users do not always have time to hunt constantly for new and better resources, librarians can use the "Message from the librarian" service to keep patrons informed and make announcements and suggestions in order to help users stay in touch with interesting information in their selected discipline.

At the North Carolina State University, the MyLibrary@NCState system resides on a Unix computer, but for the same purpose an open source and free operating system from the Unix family could also be used, such as Linux or FreeBSD. The database server is MySQL, and is also freely available, with the ability to run on multiple computers and an application programmer's interface support for both the Perl and C programming languages. MyLibrary@NCState works with any (free) HTTP server which co-operates with a Unix family operating system, as long as it has common gateway interface (CGI) scripts support. The application part uses Perl to extract data from the database and dynamically format the HTML output according to the user's profile. Perl was chosen since it is an interpreted language,

easy to debug, free, works on almost every platform and comes with API (application program interface) support for MySQL.

The customisation and functionality extension of MyLibrary@NCState

The installation process of MyLibrary@NCState is straightforward and, due to very good installation instructions, it is a simple task for a skilled Unix/Linux administrator. The CMK administrator chose Intel-Pentium based hardware to serve the application. He opted for the Linux operating system and chose MyLibrary as it was the only application hosted on the selected system.

The following applications, usually shipped with Linux distribution, were chosen during the operating system installation:

- Apache as an open source and free Web server with CGI support;
- Perl as the scripting language (freely distributed at CPAN (comprehensive Perl archive network)); and
- MySQL as the required database.

The operating system set-up was followed by installation of several modules and packages that were deliberately left out during operating system setup because of not being included in the adopted distribution of Linux. After that step, the system was ready for MyLibrary@NCState installation, which consists of configuring and testing the MyLibrary SQL database and publishing the HTML content and Perl scripts on the Web.

Since the database shipped with the distribution of the system is filled with examples, the system can be tested immediately after installation.

We customised the portal by:

- extending the database structure to enable the multilingual functionality needs;
- redeveloping Perl scripts to support multilingual functionality for the user part;
- loading the database with the contents that CMK wants to provide to its users; and
- providing the portal with appropriate design to suit the existing CMK Web site.

The majority of the content we wanted to make available through the CMK portal is already accessible as part of the static CMK Web site. It is published in a static HTML format and contains the resources needed to

be loaded into the database. In order not to have to manually rewrite all the resources, we developed Perl tools that reformatted all the gathered information and automatically inserted it into the database. We are also planning to develop Perl scripts to help our librarians provide a current awareness service using the national online library catalogue system COBISS.

A disadvantage of the readily developed portal systems is that they support only English as the language of user-library communication. None of the abovementioned systems provides either the possibility for translation or for multilingual support.

Since most of our Web site users originate from Slovenia, and there is also a non-negligible percentage of users from other countries (such as Croatia), our primary goal was to upgrade and change the functionality of the MyLibrary@NCState system to support language dependence and later on multilingual support. The languages supported in the first version of the CMK portal are Slovenian (for local users) and English (for the international users). The libraries in countries with several official languages could use systems with multilingual support to provide the users with access in their native language (e.g. a Swiss library could be offering the contents in German, French and Italian).

The multilingual support is offered primarily at the user-oriented part of the MyLibrary@NCState system. There is no reason to add this kind of functionality to the administrative part since the content providers and the administrators usually know the system in detail and have no problems using a computer system in the English language.

Choosing the language is an important part of the personalisation process. At the time of account creation, the user provides the information about the language that he/she wants to use (i.e. selects the language from the list of languages the system supports). The information about the user is stored in the system as a user profile and is read and reconfigured during the log-on process. For example, if a user who uses the system in the Slovene language, logs on to MyLibrary, the information provided to him/her should be in Slovene. When the user logs off and another user logs on from the same computer, and if

the second user's selected language is English, the system should respond in English.

The process of adding the described functionality was not trivial – it required redesign of the entire system. First, the entire system had to be inspected for language-dependent data. The system has most of the data continuously used for generating HTML communication stored in one table in the database. The other language-dependent features are the description texts of links, electronic journals, universities and databases. This information is stored in the database's tables linked with the sections they support. There is some general and frequently used data which are also language dependent (e.g. name of sections, such as "quick searches", headers and footers); this information is also stored in the database. Unfortunately, there are some pieces of information (used most frequently for the communication forms, such as the words: "username" and "password") that are hard-coded into the Perl module and the executable Perl script.

Every part of the system that comprises language-dependent data required a different approach to enable the multilingual functionality. At the database level, all the tables that contain language-dependent data had to be expanded with an additional field, which indicates the language of record's information. The primary key was changed from the ID column to the conjunction of the ID and the language column. The user's data also had to be extended to include the information on the language the user has selected during the account creation phase. The translation of hard-coded language-dependent information in the Perl module could not be done directly. Fortunately, these texts present a minor portion of all of the text displayed by MyLibrary@NCState. They are mainly presented in forms and have not yet been completely translated in the pilot version of the system with integrated multilingual functionality. We suggested to the system developers to move these texts to the database, where the multilingual functionality can be added in the same way as for general data. In the next version of the MyLibrary@NCState system a different approach would be helpful for the data-presentation layer, using language-dependent XML files and XSL transformation. This solution would not only make multilingual

design easier – it is also recommended as a state-of-the-art portal building solution and would also make the system portable to handheld or wireless devices.

To accomplish multilingual functionality at the application layer, Perl scripts and modules had to be changed to be able to use the new extensions of the database layer. This was the most demanding task. After completing it, some extensions to the forms and dialogues had to be made to enable users to change their language of choice.

At the end of the multilingual features implementation, the database had to be filled with the data in the languages the system was going to support. In the case of CMK, previously chosen records, needed for communication with users, had to be translated into Slovene and stored in the correct fields in the database. We did not develop our own administrative user interface for this purpose; instead, we installed MyODBC (my open database connectivity) drivers to the translators' computers and gave the translators direct access to a copy of the MyLibrary database. This way, the librarians in charge of translation could use the program they are already familiar with – Microsoft® Access – to enter the translated texts directly into the database. The system administrator had to review all the inputs and copy them into the original MyLibrary database.

MyLibrary CMK functionality

CMK is an academic medical library that provides biomedical information resources to its clientele. The organisation, role and function of CMK is somewhat different from that of the NCSU libraries. Therefore the help text for users, the contents of MyLibrary and the list of disciplines needed appropriate customisation and changes. CMK cataloguers classify library information into 60 medical disciplines, which are included in discipline-specific sections: reference shelf; databases; and electronic journals.

All the text in the records that help the user to create and use MyLibrary CMK was translated into the Slovene language. Records considering databases, reference shelf, new book list, librarians, technical details and some other topics needed customisation to the CMK environment.

Resources on the user's MyLibrary CMK page include information about the system, messages from librarians, links to librarians,

university resources, personal list of Internet resources, reference resources, databases, electronic journals, as well as access to the search engine and display of news. The authorised librarian can create, modify or delete the names and contact information of librarians, links and description of information resources, help texts and messages from librarians.

The usage of MyLibrary CMK usually starts with the creation of a new account. The user has to fill in the displayed form and choose the appropriate language from it. Afterwards, the MyLibrary page is displayed in Slovene or English and all subsequent communication with MyLibrary is done in the selected language (Figures 1 and 2).

The users have a possibility to customise their profiles, including changing the language (Figure 3).

The message from the librarian is intended to be displayed only to people who have chosen a particular discipline. Every user of MyLibrary is associated with a medical discipline and each of these disciplines is associated with a text message. The user cannot customise message functions.

- *Your librarians* section lists names, telephone numbers and e-mail addresses of the librarians associated with the user's chosen discipline. In CMK, all reference librarians serve all medical disciplines therefore all of them should be associated with all disciplines. This section is not directly customisable by the user.
- *Library links* section enables the selection of some of the most important items about library services, policy and directions. Users can select only those items needed for the regular use of the library.
- *University links* section provides access to some important items on the Faculty of Medicine home page and the University of Ljubljana home page. It is useful to have all these links in the same list and the possibility to choose the preferred ones.
- *Personal links* section allows the users to save links to Internet resources that they often use and are not included in other sections.
- *Reference shelf* section includes a digest of about 500 quality Internet resources

Figure 1 Login and MyLibrary CMK page are displayed in the language that was selected in the user's profile, e.g. in Slovene



UNIVERZA V Ljubljani
MEDICINSKA FAKULTETA

[Domov](#)
[O MojKnjižnici](#)
[CMK](#)

[Nova prijava](#)
[Prijava](#)
[Profil](#)
[Izločed](#)
[Odjava](#)

Prijava v MojoKnjižnico

Na tej strani se lahko prijavite v MojoKnjižnico.

Predmet	Izbor	Opis
Up. ime	matej_turk	Vnesite uporabniško ime
Geslo	*****	Vnesite geslo.

(Na začetek)

CMK © 2001
Vprašanja pošljite na: moja.knjiznica@mfn.uni-lj.si



UNIVERZA V Ljubljani
MEDICINSKA FAKULTETA

[Domov](#)
[O MojKnjižnici](#)
[CMK](#)

[Nova prijava](#)
[Prijava](#)
[Profil](#)
[Izločed](#)
[Odjava](#)

MojaKnjižnica

Centralna medicinska knjižnica

Prijavite se na konferenco s nuklearne medicine 28.7.2002!

Hitro iskanje (o | po men)

CMK homepage

Vaši knjižničarji (o)

Vaši referenčni knjižničarji so: [Nana Turk](#) (543-7432, nana.turk@mfn.uni-lj.si).

Knjižnične povezave (o | po men)

- [Delovni čas](#)
- [Fotokopiranje](#)
- [Izposoja](#)
- [Knjižracni red](#)
- [Seznam tiskanih revij](#)

Univerzitetne povezave (o | po men)

- [Clanice na UL](#)
- [Knjižnice UL](#)
- [Medicinska fakulteta \(MF\)](#)
- [Študij na UL](#)

Matej - povezave (o | po men)

- [All resources](#)
- [Nuclear medicine - design](#)
- [Nuclear power](#)

Referenčna polica (o | po men)

- [All resources](#)
- [Konference](#)
- [Navodila za pisanje članikov](#)
- [Slovani](#)

Baze podatkov (o | po men)

- [All resources](#)
- [AMIA Home Page](#)
- [MerckMedicus Using Technology](#)
- [University of Washington Radiology Webserver](#)

Elektronske revije (o | po men)

- [BioMed Central - BMC nuclear medicine](#)
- [Human Molecular Genetics](#)
- [Medical Informatics and the Internet Medicine](#)
- [Nucleic Acid Research](#)

Novosti (o | po men)

- [Novosti februar 2002](#)
- [Novosti januar 2002](#)

(Na začetek)

CMK © 2001
Vprašanje pošljite na: moja.knjiznica@mfn.uni-lj.si

listed by medical discipline. Users can choose items from an alphabetically ordered list or from a discipline-specific list.

- *Databases* section includes various medical databases: bibliographic; factual; and electronic textbooks. The list of databases is limited; therefore the discipline-specific selection is not necessary even though it can be enabled. The selection from alphabetic list seems to be the useful choice.
- *E-journals* is the most rapidly expanding section in CMK. At the moment, it provides access to more than 1,500 titles.

Hence, personalisation according to the preferred discipline could be a very useful tool for busy faculty or clinicians. Selection from an alphabetic list is also enabled (Figure 4).

- *News* section is intended to offer a list of new books. For the creation of this list, we plan to establish a connection with the national online library catalogue COBISS.
- *The quick searches* section provides direct access to the Google search engine and the CMK catalogue within COBISS. A searching tool for the whole CMK Web site is under development.

Figure 2 Login and MyLibrary CMK page are displayed in the language which was selected in the user's profile, e.g. in English language

UNIVERSITY OF LJUBLJANA
cmk
FACULTY OF MEDICINE

Login to MyLibrary

Use this page to login into MyLibrary.

Item	Selection	Description
Username	john_white	Enter your username. (More help)
Password	*****	Enter your password. (More help)

[Login](#)

[\(Go to top\)](#)

CMK © 2001
Send comments and questions to: maja.kojzarica@mf.uni-lj.si

UNIVERSITY OF LJUBLJANA
cmk
FACULTY OF MEDICINE

MyLibrary

Central Medical Library

Submit the papers to the Nuclear medicine Conference due the 28.7.2002!

Quick Searches ([about](#) | [customize](#))

CMK Homepage [Search](#)

Your Librarian(s) ([about](#))

Your reference librarians are: [Nana Turk](#) (543-7432, nana.turk@mf.uni-lj.si).

Library Links ([about](#) | [customize](#))

- [Circulation policies](#)
- [Hours of operation](#)
- [Library policies](#)
- [Photocopy services](#)
- [Textbook catalog](#)

University Links ([about](#) | [customize](#))

- [Faculty of Medicine \(MF\)](#)
- [UL Libraries](#)
- [UL Members](#)
- [Students](#)

John - Links ([about](#) | [customize](#))

- [All resources](#)
- [Nuclear medicine - design](#)
- [Nuclear power](#)

Reference Shelf ([about](#) | [customize](#))

- [All resources](#)
- [Conferences](#)
- [Dictionaries](#)
- [Instructions to Authors](#)

Databases ([about](#) | [customize](#))

- [All resources](#)
- [AMIA Home Page](#)
- [MerckMedicus Using Technology](#)
- [University of Washington Radiology Webserver](#)

E-journals ([about](#) | [customize](#))

- [BioMed Central - BMC nuclear medicine](#)
- [Human Molecular Genetics](#)
- [Medical Informatics and the Internet Medicine](#)
- [Nucleic Acid Research](#)

News ([about](#) | [customize](#))

- [News february 2002](#)
- [News january 2002](#)

[\(Go to top\)](#)

CMK © 2001
Send comments and questions to: maja.kojzarica@mf.uni-lj.si

Plans for further work

We plan to finish our pilot version of the CMK portal during 2003. To ensure full functionality, some further elaboration is still needed in certain sections. Afterwards, the existing pilot portal will be evaluated from the users' perspective through a usability study. The study will focus on what the users think about the portal, how they figure out how to interact with the portal and how they acquire the information essential to the portal's operation. We plan to include both faculty and students in the study.

The detailed usability study will be performed by different approaches: we will carry out a survey among users and at the

same time we will analyse the Web server log files and database records about users. Since the users who decide to use personalisation features of the library's portal have registered the data on their status, education and fields of interest, we are able to correlate the way they use the portal with the appearance of their portal. We were not able to perform such analysis using only Web log files since these files offer no possibility to recognise distinct users.

Later, we plan to automatically personalise the user's portal based on her/his behaviour and on the behaviour of other users with similar interests. To achieve this we will use data mining methods to discover user navigation patterns.

Figure 3 Customisation of the user's profile includes the change of language

Customize your Profile

Your Profile is personal information related to your account and your interest areas. Use this page to customize or change your information.

Item	Selection	Description
Username (ID)	john_white (551)	Here are your username and ID number. You cannot change them.
First name	John	Enter your first name. (Help)
Last name	White	Enter your last name. (Help)
Discipline	Nuclear medicine	Select your primary interest area. (Help)
Language	english	Select your primary language choice. (Help)
Status	slovenscina english	What is your rank. (Help)
Email address	john.white@hotmail.com	Enter your email address. (Help)
Can contact	☒ Yes ☐ No	May we send you email? (Help)
Telephone	(01) 543 - 7784 ext.	What is your telephone number? (Help)

[Submit](#) [Reset](#)

[\(Go to top\)](#)

CMK © 2001
Send comments and questions to: voja.krznica@mf.uni-lj.si

Figure 4 Selection of electronic journals which user wants to have displayed on his/her MyLibrary CMK page

Select electronic journals

Use the form below to select electronic journals to add to your library.

[Submit](#) [Reset](#)

Selection	Item
☐	Annals of surgery
☐	Archives of disease in childhood
☐	Archives of surgery
☐	Annals of otology, rhinology & laryngology
☐	Annals of thoracic surgery
☐	Biochemical pharmacology
☒	Blood
☐	Cell and tissue research
☐	Circulation
☐	Circulation research
☐	CMAJ Canadian medical association journal
☒	Head and neck
☒	Human pathology
☐	JAMA
☐	JCU Journal of clinical ultrasound
☐	Journal of infectious diseases
☐	Journal of nuclear medicine
☐	Journal of ophthalmology

Conclusion

The exponential growth of the amount of information and its availability in electronic form makes users concerned about their ability to assimilate and manage this bounty. Library users have high expectations based on the broader commercial marketplace. The development of library Web portals will inevitably make libraries and their staff more customer-centred. Trends in portal development include customisation and personalisation. The personalisation process is provided by portals using dynamic page generation and server-resided data storage. But there are benefits not only for the users – such systems also facilitate the content-updating procedures through the user interfaces and the tools they provide.

The pilot development of MyLibrary CMK is a logical step in the CMK Web site developments from 1997 onwards. It is based on an existing portal solution with extended functionality, which is adjusted to the functions of the academic medical library and its users' needs. One of the most important needs of visitors from Slovenia and abroad is multilingual support, which means that the users can select a language in which the portal interface is presented to them. This multilingual support is the key extension we have developed for the MyLibrary CMK portal.

References

- Cohen, S., Ferreira, J., Horne, A., Kibbee, B. and Smith, A. (2000), "MyLibrary – personalised electronic services in the Cornell University Library", *D-Lib Magazine*, Vol. 6 No. 4, available at: www.dlib.org/dlib/april00/mistlebauer/04mistlebauer.html
- Di Giacomo, M., Mahoney, D., Bollen, J., Monroy-Hernandez, A. and RuizMeraz, M.C. (2001), "MyLibrary, a personalisation service for digital library environments", *ERICIM Workshop Proceedings – No. 01/W03*, Dublin City University, Ireland, 18-20 June 2001, available at: www.ericim.org/publication/ws-proceedings/DelNoe02/Giacomo.pdf
- Gharphery, J. and Ream, D. (2000), "VCU's my library: librarians love it . . . users? Well maybe", *Information Technology and Libraries*, Vol. 19 No. 4, pp. 186-90.
- Jordan, W. (2000), "My gateway at the University of Washington libraries", *Information Technology and Libraries*, Vol. 19 No. 4, pp. 180-5.
- Ketchell, D.S. (2000), "Too many channels: making sense out of portals and personalisation", *Information Technology and Libraries*, Vol. 19 No. 4, pp. 175-9.
- Lakos, A. and Gray, C. (2000), "Personalised library portals as an organizational culture change agent", *Information Technology and Libraries*, Vol. 19 No. 4, pp. 169-74.
- Lesk, M. (1997), *Practical Digital Libraries: Books, Bytes and Bucks*, Morgan Kaufman, San Francisco, CA.
- LITA TopTech Trends (1999), "Technology and libraries users: LITA experts identify trends to watch", paper presented at the ALA Midwinter Meeting, January 1999, available at: www.lita.org/committee/toptech/mw99.htm
- Looney, M. and Lyman, P. (2000), "Portals in higher education", *EDUCAUSE Review*, Vol. 35, p. 30.
- Morgan, E. (1999), "MyLibrary@NCState: the implementation of a user-centred, customisable interface to a library's collection of information resources", *Proceedings of the Customised Information Delivery Workshop, SIGIR '99*, Berkeley, CA, 19 August, pp. 12-18.
- Rozic-Hristovski, A., Todorovski, L. and Hristovski, D. (1999), "Developing a medical library Website at the University of Ljubljana, Slovenia", *Program*, Vol. 33 No. 4, pp. 313-25.
- Rozic-Hristovski, A., Hristovski, D. and Todorovski, L. (2001), "Discovering patterns in library website usage: a data mining approach", paper presented at the EAHIL workshop 2001 Cyberspace Odyssey, 7-9 June, Alghero, available at: <http://pacs.unica.it/alghero2001/proceedings/043.htm>
- Rozic-Hristovski, A., Hristovski, D. and Todorovski, L. (2002), "User's information-seeking behavior on a medical library Website", *Journal of the Medical Library Association*, Vol. 90 No. 2, pp. 210-17.
- Seljak, M. and Seljak, T. (2002), "The development of the COBISS system and services in Slovenia", *Program*, Vol. 36 No. 2, pp. 89-98.
- Spiliopoulou, M. (2000), "Web usage mining for Web site evaluation", *Communications of the Association for Computing Machinery*, Vol. 43, pp. 127-34.

Routing in ISL Networks Considering Empirical IP Traffic

Ales Svigelj, Mihael Mohorcic, *Member, IEEE*, Gorazd Kandus, *Member, IEEE*, Andrej Kos, *Member, IEEE*, Matevž Pustišek, *Member, IEEE*, and Janez Bešter, *Member, IEEE*

Abstract—Next-generation satellite networks are expected to provide a variety of applications with diverse performance requirements, which will call for the development of adaptive routing procedures supporting different levels of services. In this paper, we propose traffic class dependent (TCD) routing, which has the potential to differentiate between traffic classes using different optimization criteria in route calculation. The performance of TCD routing is evaluated for different traffic scenarios using an empirical traffic source model derived from the real backbone Internet traffic trace and compared with results obtained with equivalent Poisson traffic as a reference point. In addition, TCD routing is compared with a simple single service routing procedure, which does not make any distinction between traffic classes. Performance analysis, in terms of average packet delay, normalized data throughput, and normalized link load, reveals improved routing resulting from traffic class differentiation, regardless of the traffic scenario considered. The performance measures based on aggregate traffic flow show no significant difference between routing of empirical and equivalent Poisson traffic.

Index Terms—Backbone Internet protocol (IP) traffic, inter-satellite links, low earth orbit (LEO), traffic class dependent (TCD) routing.

I. INTRODUCTION

DUE TO THE possibility of providing different types of traffic to a large geographical coverage, satellite networks are expected to be an essential component of the next-generation Internet. They are best suited for supporting asymmetric applications such as data, audio and video streaming, bulk data transfer, and multimedia applications with limited interactivity, and also for providing broadband access to remote users beyond densely populated areas. Constellations with satellites in low earth orbit (LEO) interconnected with intersatellite links (ISLs) are particularly attractive for global broadband communication networks, since they provide capacity and delays comparable to terrestrial networks.

High initial investment, long development and launching cycles, and the physically inaccessible location of satellites, as well as different and fast changing service requirements, de-

mand an even more accurate and well considered planning phase for satellite networks than for terrestrial networks. Furthermore, the complexity and nature of satellite systems preclude testing or closed-form mathematical analysis, so most of the system design and network dimensioning relies on service scenario and computer simulation, the latter at least profiting significantly from recent advances in processing power and advanced simulation tools.

ISL networks based on LEO constellations are characterized by significant dynamics of topology and traffic load due to variation of distance between satellites in adjacent orbit planes and changing satellite coverage area on the surface of the earth. Under such operating conditions it is essential to implement efficient adaptive routing. Most of the existing ISL routing studies address the problems of routing in connection-oriented networks, either on packet or, predominantly, on connection level. Consideration of Internet protocol (IP)-like per-hop packet routing, which is addressed in this paper, is only possible in connectionless networks. Performance evaluation of such routing, however, needs to be carried out on the packet level, taking into account the actual status of the network, and requires the use of suitable traffic sources. Due to their simple implementation and the possibility of mathematical presentation with analytical formulae, basic traffic sources, such as Poisson or uniform, are most commonly applied in computer simulations of network performance. Thus, our previous studies of adaptive per-hop routing in ISL network were all based on the use of Poisson traffic sources. Assuming completely homogeneous traffic load conditions in the ISL network, we studied the inherent routing characteristics of a simple shortest path routing in nonequatorial LEO constellation [1], demonstrating the effect of interplane traffic concentrating at higher latitudes due to shorter distance between neighbouring orbit planes. This effect motivated the research in the direction of traffic load sharing using forwarding policies based on alternate link routing (ALR) [2]. In particular, we proposed alternate link routing with deflection in the source node and alternate link routing with deflection in all nodes. Both ALR forwarding policies efficiently handle traffic load sharing among alternative routes, reducing peak values of link load by as much as 50% in comparison with the reference forwarding policy employing shortest path routing (SPR). They also significantly suppress the effect of interplane traffic concentration at higher latitudes. Next, we extended our research to nonhomogeneous traffic flow scenarios with homogeneous [3] and demographically weighted nonhomogeneous [4] distribution of traffic sources over the landmasses, assuming different traffic flow patterns between different continental regions. In [3], we investigated

Manuscript received December 15, 2002; revised July 1, 2003 and September 20, 2003.

A. Svigelj, M. Mohorcic, and G. Kandus are with the Department of Digital Communications and Networks, Jozef Stefan Institute, Ljubljana SI-1000, Slovenia (e-mail: ales.svigelj@ijs.si; miha.mohorcic@ijs.si; gorazd.kandus@ijs.si).

A. Kos, M. Pustišek, and J. Bešter are with the Faculty of Electrical Engineering, University of Ljubljana, Ljubljana SI-1000, Slovenia (e-mail: andrej.kos@fe.uni-lj.si; matevz.pustisek@fe.uni-lj.si; bestej@fe.uni-lj.si).

Digital Object Identifier 10.1109/JSAC.2003.819974

effects of different traffic flow scenarios on adaptive routing, in particular the impact of increasing nonhomogeneity of traffic flows, showing that the effect of interplane traffic concentration at higher latitudes remains also in more realistic traffic load conditions. In [4], we demonstrated that alternate link routing reduces the effect of interplane traffic concentration and, thus, improves the routing performance by more uniform load sharing among links also in nonhomogeneous traffic load conditions. To overcome drawbacks of conventional single service routing we proposed in [5] the traffic class dependent (TCD) routing, which has the potential to satisfy diverse performance requirements by taking into account different traffic classes, each with its particular optimisation criteria. We showed that efficient traffic class differentiation requires implementation of advanced scheduling policies, such as weighted round robin (WRR) and not only a simple first come first serve (FCFS) suitable for a conventional single service routing. Finally, considering different traffic flow scenarios, we demonstrated that the higher the traffic load and the more nonhomogeneous the traffic scenario, the higher the gain of introducing the TCD routing.

As has been said earlier, our previous studies were conducted using Poisson traffic, however, traffic sources that are tractable by pure mathematics often do not conform to traffic patterns present in real communication networks. For this reason, in this paper, we enhance our adaptive routing studies with consideration of empirical traffic source derived from the real backbone Internet traffic trace. The performance of routing using an empirical traffic source model is compared with results obtained with Poisson traffic.

The rest of this paper is organized as follows. Section II discusses traffic scenarios. It includes analysis of real backbone Internet traffic and, based on this, defines an adequate model of empirical traffic. Furthermore, different traffic classes with diverse performance requirements are introduced and traffic flow models are described briefly for geographical and temporal variation of traffic load on satellites in the ISL network. Section III states the assumptions regarding the satellite constellation and ISL network topology, and TCD routing is introduced in Section IV. In Section V, ISL network simulator implementation issues are discussed, and relevant modeling assumptions and simulation parameters are summarized. Finally, the simulation results are presented and discussed in Section VI.

II. TRAFFIC SCENARIOS

Traffic analysis and modeling are integral parts of engineering broadband telecommunications networks, including satellite networks with ISLs. The analysis is concerned with obtaining statistical properties of the traffic and analytical solutions for the evaluating performance of networks, in terms of different measures such as call blocking probability or packet delay. Among the most important traffic properties for circuit-switched networks are the connection duration distribution and the average number of connection requests per time unit. By contrast, in the case of packet-switched networks, traffic characteristics are given typically by packet lengths and packet interarrival times (in the form of distributions or histograms), burstiness, moments, autocorrelations, and

scaling (including long-range dependence, self-similarity, and multifractals) [6]–[8].

In recent years, it has become clear that the IP-based protocols making use of packet technology will play a crucial role in the development of future broadband communication networks. The success of packet technology is attributed to many reasons [9], such as robustness, reliability, flexibility and, especially, statistical multiplexing [10]. Statistical multiplexing is based on combining multiple individual traffic flows into a single aggregate traffic flow. The capacity needed in case of a single aggregate traffic flow is lower than the sum of individual capacities needed in the case of multiple traffic flows. Thus, the result of statistical multiplexing is a smoothed-out traffic flow with reduced burstiness. In the case of Poisson-like traffic, aggregation yields fast decrease in burstiness. For real, predominantly self-similar traffic, the decrease of burstiness is slower, due to the higher correlation between individual traffic flows [11]. In this paper, we assess the impact of using different traffic source models on the overall performance of adaptive routing in ISL networks. Traffic, that represents aggregate traffic flows between satellites in the ISL network, is generated according to the selected statistical or mathematical properties. The generation of traffic is modulated with a suitable traffic flow model describing geographic and time dependent variation of traffic intensity. Within the aggregate traffic flows, however, it is appropriate to differentiate between packets belonging to different traffic classes, thus supporting the introduction of routing procedures capable of providing different levels of service.

A. Traffic Sources

Analytical solutions for circuit switched voice networks have been one of the most successful applications of mathematical techniques in the industry [12]. This is possible because voice traffic has the property of being relatively homogenous and predictable, and it spans long time scales. Data traffic, on the other hand, is much more variable, since in a typical multimedia application it contains a mix of packets from various sources. For this reason, the applicability of traffic analysis based on mathematical tractability is diminishing, as purely mathematical traffic sources cannot capture the traffic characteristics in real networks to the extent which would allow detailed performance evaluation of the network. Consequently, the importance of computer simulation has grown considerably, but poses different requirements for traffic source models [13]. A suitable traffic source model should represent real traffic and the possibility of mathematical description will be less important. Ideally, the model should capture the essential characteristics of traffic that have significant impact on network performance with only a small number of parameters (i.e., a parsimonious model) and should allow fast generation of packets.

With the lack of suitable traffic models for multimedia traffic sources that meet the above requirements, two alternatives exist for simulating network performance. We can either decide for a trace driven simulation, by applying real traffic traces obtained by measurement, or we can use these measurements to develop an appropriate traffic source model which conforms to empirical distributions of real traffic. Trace-driven simulations are

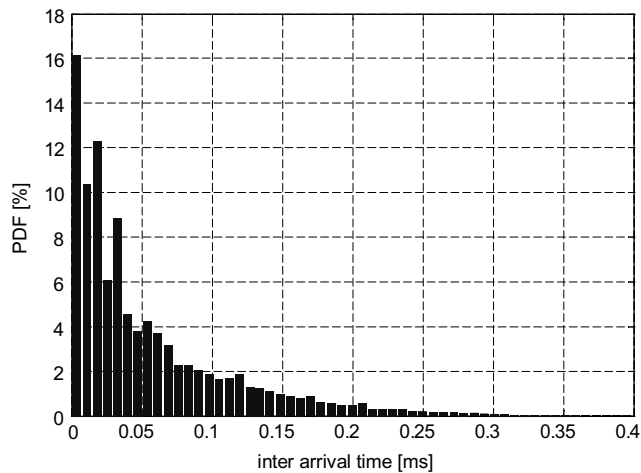


Fig. 1. Packet interarrival time distribution in a sample of a real backbone traffic trace.

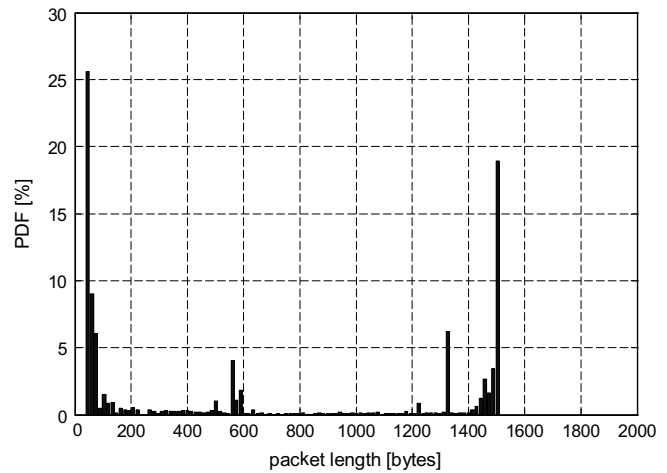


Fig. 2. Packet length distribution in a sample of a real backbone traffic trace.

recommended for model validation, but suffer from two drawbacks. First, the simulation can only reproduce something that has happened in the past and secondly, there is seldom enough data to simulate all possible scenarios, since the extreme situations are particularly hard to capture. Simulations with traffic source models, whose empirical distributions conform to those of real traffic, avoid these shortcomings, but can sometimes deviate considerably from real situations.

In the field of research, traffic source models based on empirical distributions of real traffic are becoming increasingly popular in simulations and are replacing basic traffic generators, since they are also relatively simple to implement and at the same time allow for high flexibility. We, therefore, decided to use a traffic source model that resembles the empirical distributions of real backbone Internet traffic. To validate this model, the results are compared with those obtained by using Poisson traffic.

In packet-switched networks, two random processes are used to define the traffic properties. These are the interarrival time series $\{T_i\}$, where T_i denotes the i th interarrival time between two successive packet arrivals, and the packet length sequence, $\{L_i\}$, where L_i represents the length of the i th packet. As there is no operating broadband LEO satellite network, the traffic properties needed to build a suitable traffic source model were extracted from a real traffic trace that was captured on the 622 Mbit/s backbone Internet link, carrying 80 Mbit/s traffic [14]. The selected traffic trace comprises aggregate traffic from a large number of individual sources and its characteristics yield similar results, in terms of packet lengths and represented protocols and applications, as in [15]. Such a traffic trace resembles the traffic load experienced by the satellite, both from numerous traffic sources within its coverage area, and from aggregate flows transferred over broadband intersatellite links.

Packet interarrival time and packet length distributions of the selected real traffic trace are shown in Figs. 1 and 2, respectively. From the latter, it is evident that almost 45% of packets are shorter than 100 bytes, with 26% constituting a distinctive peak of packets containing 44 bytes. Approximately 16% of all packets are located in the range between 100 bytes and 600 bytes

with 7% concentrated between 550 and 600 bytes. Another peak comprising 6% of packets is at 1320 bytes, while nearly 26% are between 1450 and 1500 bytes, the latter being the maximum length with a distinctive peak comprising 19% of packets. The average packet length in the selected backbone Internet traffic trace is 670 bytes.

We have determined different characteristics of the real traffic trace by the traffic analysis, i.e., probability density function distributions of packet lengths and packet interarrival times, burstiness, and scaling [16]. Traffic analysis has demonstrated that not all the statistical properties of real traffic have a crucial impact on network performance. Scaling, for example, has a negligible impact on maximum throughput and cell loss probability, in agreement with [17]. Therefore, for the traffic source model which resembles IP traffic in the backbone network, and is required for the simulation of routing in the ISL network, we reproduced only the properties that have major impact on network performance, i.e., interarrival time and packet length distribution. In this way, we have developed a traffic generator which fully preserves the shape of the packet interarrival time distribution depicted in Fig. 1. A traffic generator, that uses a lookup table with normalized values, allows packet interarrival time values to be scaled, so as to achieve the desired total traffic load. The packet interarrival time distribution of this traffic generator is shown in Fig. 3. Furthermore, the traffic generator produces packets of three distinctive lengths corresponding to lengths of the most frequent packets in the backbone network. Thus, short packets are represented in the traffic generator by a packet length of 44 bytes, intermediate packets by 576 bytes, and long packets by 1500 bytes. The proportion of packets with a particular length is defined in such a manner that the average packet length and the ratio between short and long packets are the same as in the real traffic trace. The packet length distribution obtained with this traffic generator is shown in Fig. 4. As the traffic generated in such a way resembles empirical distributions of the real traffic trace, we refer to it in the rest of the paper as empirical traffic.

In this paper, the performance of routing procedures involving empirical traffic has been compared with that obtained with equivalent Poisson traffic. In this respect, a Poisson traffic

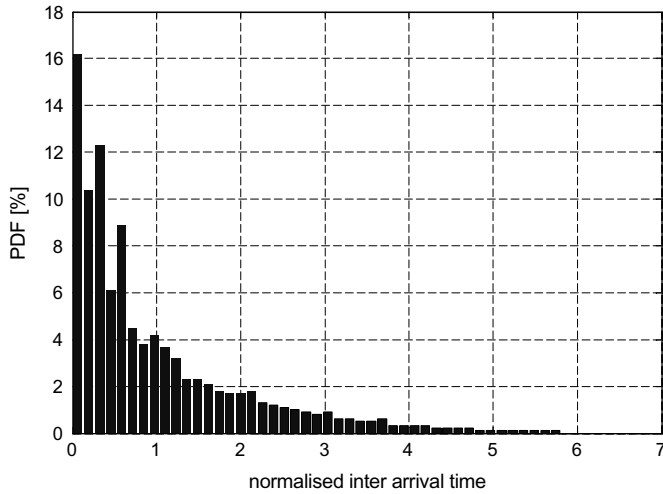


Fig. 3. Packet interarrival time distribution in empirical traffic.

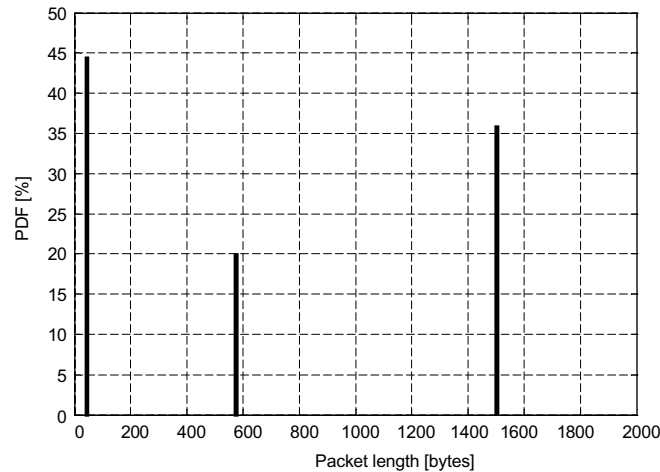


Fig. 4. Packet length distribution in empirical traffic.

source model has been used with a negative exponential probability distribution of interarrival times and a truncated negative exponential probability distribution of packet lengths. Packets are truncated if greater than the specified maximum value and a new packet with the length of the residue is generated. The statistical properties of minimum, average, and maximum packet lengths are selected so that the overall traffic load in the network is the same as in the case of empirical traffic, i.e., they are set to 20, 670, and 2000 bytes, respectively. Probability density functions for packet interarrival time and for packet length from an equivalent Poisson traffic source model are shown in Figs. 5 and 6, respectively. The effect of packet truncation is best demonstrated in Fig. 6, which shows that all packets larger than 2000 bytes are truncated, which explains large number of packets with the maximum length. As shown in Fig. 6, approximately 5% of packets generated with the Poisson generator using the above parameters are originally larger than the maximum length and, hence, truncated. The Poisson traffic source is modeled in such a way that it resembles some of the first-order statistical properties of the real traffic trace, i.e., minimum, average, and maximum packet length and average packet interarrival time. In the rest of the paper, therefore, we

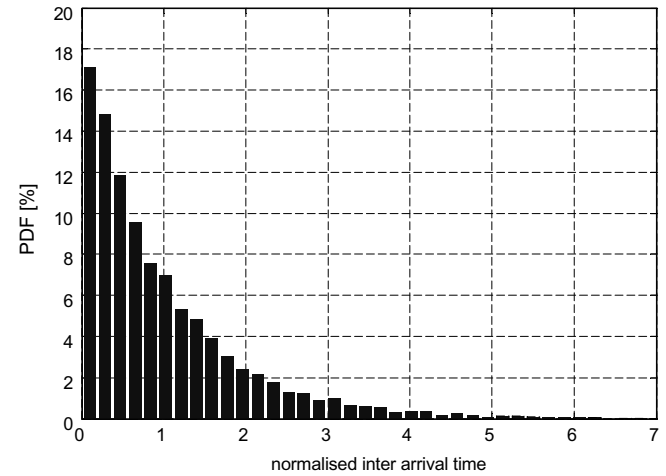


Fig. 5. Packet interarrival time distribution in Poisson traffic.

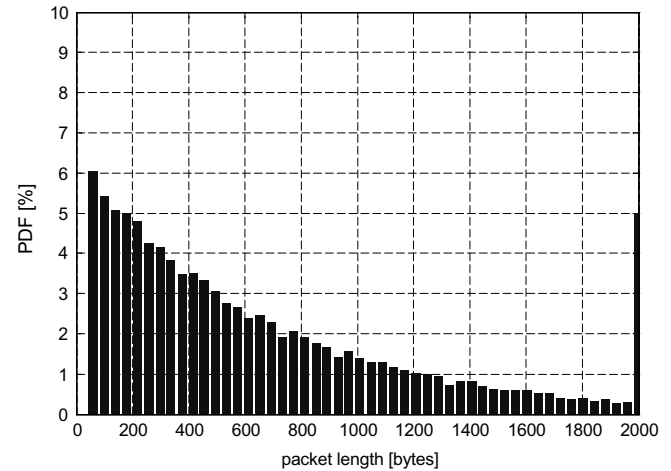


Fig. 6. Packet length distribution in Poisson traffic.

refer to traffic generated with it as equivalent Poisson traffic or, in short, Poisson traffic. Although Poisson traffic never fully resembles the characteristics of real traffic, it can serve as a reference point to compare simulation results, when different scenarios are used.

B. Traffic Class Differentiation

The emergence of applications with different requirements of throughput, loss, or delay calls for adaptive routing procedures supporting different levels of services, as opposed to a single best-effort service. In order to evaluate the performance of such routing procedures in the ISL network, we have to introduce different traffic classes with diverse performance requirements, each being routed according to its particular optimization criteria. Thus, three representative traffic classes are considered in this paper.

- Traffic class A: typical applications belonging to this traffic class include interactive real-time applications, such as voice-over-IP (VoIP) and interactive video applications, which require delay to be minimized.
- Traffic class B: representative applications of this traffic class are video-on-demand (VoD) and large file distribution, which require the throughput to be maximized.

- Traffic class C: this traffic class represents best-effort service as known in Internet and is meant for applications without any specific requirements. This traffic class is expected to be available at low price in exchange for the reduced quality-of-service (QoS).

In the simulation model, packets belonging to different traffic classes A, B, and C are generated, using three independent traffic generators in each satellite, but all of them have the same characteristics, i.e., we are considering either solely empirical or solely Poisson traffic.

C. Traffic Flows

In this paper, the traffic source models are used to generate the aggregate traffic flows between satellites in the ISL network. These generators, however, have to be modulated with respect to traffic intensity in the coverage areas of satellites and in traffic flows between different regions. Geographical and time dependent variation of traffic intensity can be represented with a suitable traffic flow model that takes into account a particular distribution of traffic sources and destinations on the surface of the earth, the variation of their traffic intensity with time, and the traffic flow pattern between them [3].

In order to evaluate the basic routing performance without the unpredictable effects of nonhomogeneous traffic load, we have defined a reference traffic flow model assuming network-uniform (NET-UNI) traffic load. For the NET-UNI traffic flow model, we specify a distribution of sources and destinations in satellite footprints, which translates into a homogeneous accumulated source/destination traffic demand on all satellites. Assuming the same flows between all pairs of satellites, this results in completely uniform network traffic, i.e., the same normalized traffic load between each end pair of satellites, regardless of their position relative to each other and to the underlying earth.

Two more realistic distributions of sources and destinations on the surface of the earth can be considered. One assumes homogeneous distribution over the landmasses (considered as continents and major islands [3]), and the other nonhomogeneous distribution, taking into account the more realistic geographic distribution, where different levels of geographical granularity may be adopted [4]. The geographic and time dependent intensity of traffic sources is mapped to the currently serving satellites with respect to their actual coverage areas on the surface of the earth. Once having determined the aggregate traffic load on each satellite, the destinations of traffic flows are assigned in accordance with the customized traffic flow pattern that resembles the flow characteristic of a certain type of service between the earth's six continental regions. For the performance analysis of a routing procedure under nonhomogeneous traffic load conditions, we have selected an LM-HS traffic flow dynamics model. The LM-HS model considers a homogeneous distribution of traffic sources over landmasses (LM) and hot spot (HS) traffic flow pattern, which is based on the assumption that, regardless of the source region, the highest traffic volume is directed towards North America and Europe, as indicated in Table I, except for traffic originated in Asia, which mostly terminates within the region.

TABLE I
PERCENTAGE OF TOTAL TRAFFIC FLOW BETWEEN SOURCE AND DESTINATION REGIONS IN HS TRAFFIC FLOW PATTERN

		Destination region					
		N. America	S. America	Europe	Africa	Asia	Oceania
Source region	N. America	60 %	10 %	15 %	2 %	10 %	3 %
	S. America	35 %	40 %	12 %	2 %	8 %	3 %
	Europe	40 %	5 %	40 %	2 %	10 %	3 %
	Africa	40 %	2 %	30 %	20 %	5 %	3 %
	Asia	30 %	2 %	10 %	2 %	50 %	6 %
	Oceania	40 %	2 %	10 %	2 %	12 %	34 %

III. SATELLITE CONSTELLATION AND ISL NETWORK ASSUMPTIONS

In satellite communication networks, different restrictions and requirements on different links have to be taken into account. Thus, resource management is typically divided into routing in the terrestrial segment, access control on up/down links, and routing in the ISL network. The ISL network dynamics constitute the main challenge for routing in satellite networks with respect to terrestrial networks. These dynamics require the adaptation of existing and development of new routing algorithms to take into account the characteristics of nongeostationary satellite networks. In this paper, we have focused on adaptive routing in ISL network. Empirical and Poisson traffic source models are used to generate the aggregate traffic load on the satellite according to the traffic intensity of sources in its coverage area. Thus, the problem of adaptive routing in ISL network can be studied independently of the terrestrial segment and up/down links.

The performance of an adaptive routing algorithm depends largely on the selected satellite constellation and the ISL topology. With regard to the latter it is advantageous to select a constellation that can maintain permanent connections between satellites in the same orbit plane (intraplane ISL) and in neighboring orbit planes (interplane ISL). Such permanent ISL connections can be provided by Walker-type satellite constellations with moderate inclination of orbit planes to the equatorial plane [18]. As a reference constellation, we have selected a typical representative of LEO Walker-type satellite constellations with ISLs, the CELESTRI constellation [19] with 63 satellites in 7 orbit planes at an altitude of 1400 km and inclination angle of 48°. The selected constellation assumes permanent ISL topology with two intraplane and two interplane ISLs per satellite, thus forming the regular skew toroidal mesh topology shown in Fig. 7. Such a topology is well suited for the provision of future ISL backbone networks [18], as it satisfies the requirements for high capacity and low propagation delay.

In spite of assuming permanent connectivity, the ISL topology is still subject to continuous variation of interplane ISL lengths, caused by the revolution of satellites around the earth. The ISL topology dynamics are deterministic and peri-

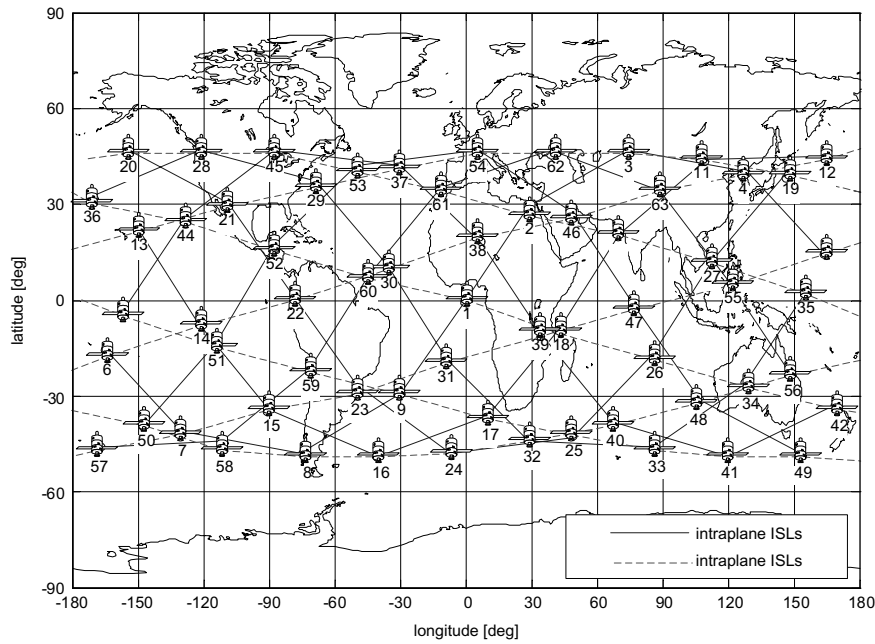


Fig. 7. Snapshot of ISL network topology with intraplane and interplane links.

odic, except for satellite or link failures, and can be computed in advance. Taking into account these properties, we considered the ISL topology through a set of time-discrete snapshots of satellite positions. The position of a satellite at a given snapshot is identified by the respective longitude and latitude values of the subsatellite point on the surface of the earth. Time steps Δt of 10 s between snapshots of the ISL topology were considered. This value provides a reasonably smooth sequence of snapshots, since in this period a satellite in the selected constellation moves by only 1.2% of its footprint diameter, which is in the range of the granularity a global traffic model can take into account.

IV. TRAFFIC CLASS DEPENDENT (TCD) ROUTING

Routing in current packet networks typically supports only one type of service and it is optimized for a single additive link-cost metric, such as delay or hop count, which is not adequate for all types of services. Conversely, the main task of TCD routing is to find suitable paths that can accommodate different types of services using different optimization criteria, such as delay or bandwidth, although still not guarantying the provision of any minimum requirements as in QoS routing [5]. In general, with TCD routing, different classes of traffic can be routed along different routes.

Any routing algorithm computes a routing table, taking into account the cost of links in the network. The computation of link cost involves two stages: 1) monitoring the link status between routing table update periods and 2) using the acquired information in a link cost computation. Selection of a suitable link-cost metric has the greatest influence on the performance of routing procedure. The parameters considered in the link-cost metric should directly represent the dynamics of the network status and the fundamental network characteristics affecting a particular traffic class, with the aim of responding to changes in the network. They should exhibit orthogonality to each other in order

to avoid interdependence and to eliminate redundant information. Since different traffic classes have different optimization requirements, different parameters have to be taken into account in their respective link-cost metrics. At the end of each routing table update interval, the link costs for all traffic classes and for all links are calculated and different routing tables are then computed, one for each traffic class.

Due to the motion of satellites in their orbit planes and the rotation of the earth there are two dynamically changing parameters in ISL network which have a significant effect on the routing performance - the length of ISLs and the traffic load on a particular satellite. These parameters are particularly well suited for a link-cost metric for the delay-sensitive traffic class A, and we use them also for traffic class C. We consider the length of ISLs through the propagation delay (T_{prop}). For the selected constellation the propagation delay is changing periodically on the interplane ISLs from 10.3 to 19.7 ms in a quarter of an orbit period (28.5 min), while on the intraplane ISLs it is equal 17.7 ms. The traffic load on the particular link is monitored through the number of packets in the outgoing queues separately for both traffic classes. The expected queueing delay, T_{exp} , is calculated using (1), where L_{average} is the average packet length, C the link capacity, and $n(t)$ the number of packets in the queue at time t .

$$T_{\text{exp}}(t) = n(t) \cdot \frac{L_{\text{average}}}{C}. \quad (1)$$

The expected queueing delay can be calculated locally, and, thus, does not require any distribution of link status among neighboring nodes. This enables a very fast response to congestion on the link, which can be best utilized by advanced forwarding policies making use of alternative paths [2]. In the link-cost function, however, we cannot take into account the continuously changing expected queueing delay, so the average expected queueing delay ($\tilde{T}_{\text{exp}}$) is calculated at the end of each

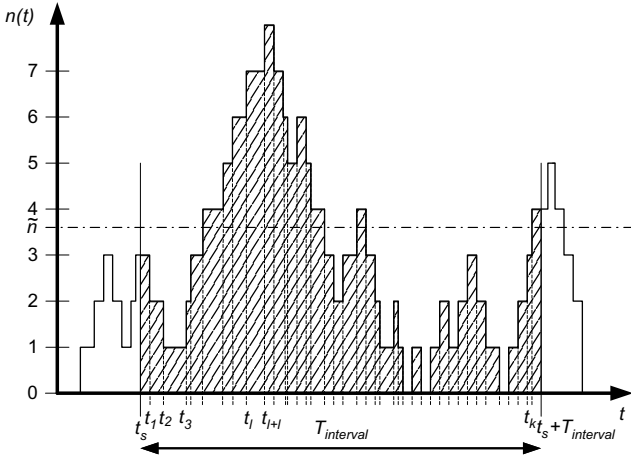


Fig. 8. Calculation of average expected queueing delay.

routing table update interval, denoted in (2) by T_{interval} and, subsequently, used in calculating link cost

$$\tilde{T}_{\text{exp}} = \frac{1}{T_{\text{interval}}} \cdot \int_{t_s}^{t_s + T_{\text{interval}}} n(t) \cdot \frac{L_{\text{average}}}{C} \cdot dt. \quad (2)$$

Fig. 8 is a graphical presentation of the average expected queueing delay calculation. As the number of packets in the queue changes only in discrete steps, the integral in (2) can be transformed into a sum denoted in (3), shown at the bottom of the page.

Using the selected metric, the link costs for traffic classes A and C are calculated as shown in (4) and (5), respectively

$$LC_A = T_{\text{prop}} + \tilde{T}_{\text{exp } A} \quad (4)$$

$$LC_C = T_{\text{prop}} + \tilde{T}_{\text{exp } C}. \quad (5)$$

For the throughput-sensitive traffic class B, on the other hand, the available bandwidth is a more suitable optimization parameter than the packet delay. Thus, on each link the lengths of the traversing packets are monitored between consecutive routing table updates. Based on this the link utilization (LU) is calculated according to (6), where L_r is the length of the r th traversing packet. In our case, the time interval during which the sum of the packet lengths was obtained equals the routing table update interval T_{interval}

$$LU = \frac{\sum_r L_r}{T_{\text{interval}} \cdot C}. \quad (6)$$

The link cost for the traffic class B is the normalized available bandwidth on the link. It is calculated according to (7).

$$LC_B = 1 - LU. \quad (7)$$

In order to increase the efficiency of the simulation model, a mechanism for the distribution of link cost has not been imple-

mented. Instead, the link costs are calculated in the simulator at one central location, assuming that the information about the network is updated in all nodes before new routing tables are calculated.

Link-cost metrics for the traffic classes A and C are typical additive metrics and, thus, the shortest routes are calculated using the Dijkstra algorithm. The main feature of an additive metric is that, if $m(n1, n2)$ is an additive metric for link $(n1, n2)$, the total cost for any path $P = (n1, n2, n3, \dots, ni, nj)$ is a sum of costs of intermediate links $m(P) = m(n1, n2) + m(n2, n3) + \dots + m(ni, nj)$. On the other hand, the link cost for the traffic class B is a concave metric. Thus, if $m(n1, n2)$ is a concave metric for link $(n1, n2)$, then the total cost for any path $P = (n1, n2, n3, \dots, ni, nj)$ equals the one on the link with minimum cost $m(P) = \text{MIN}\{m(n1, n2), m(n2, n3), \dots, m(ni, nj)\}$. The optimization criterion for the traffic class B is to find the paths within minimum hop count with the maximum available bandwidth. Minimum hop count is an additional constraint, which is used to minimize the use of resources. The Bellman–Ford shortest path algorithm was adopted to compute paths of the maximum available bandwidths within a minimum hop count [20]. It is a property of the Bellman–Ford algorithm that, at its h th iteration, it identifies the optimal path (in our context the path with the maximum available bandwidth) between the source and each destination not more than h hops away. However, because the Bellman–Ford algorithm progresses by increasing the hop count, it also provides the hop count of a path as a side result, which can be used as a second optimization criterion.

In this paper, we also compare the performance of TCD routing to that of a single service (SS) routing that does not make any distinction between packets belonging to different traffic classes. In this case, all packets are routed according to the routing tables calculated for traffic class C providing only best-effort service.

V. ISL NETWORK SIMULATOR

The complexity and dynamics of ISL networks preclude performance evaluation of routing algorithms using a closed-form analytic expression. For testing and analyzing various adaptive routing algorithms, we have, therefore, developed an ISL network simulator schematically illustrated in Fig. 9. It is built on the packet level, which increases the simulator complexity and computational effort, but on the other hand allows the study of routing algorithms that consider the actual status of the network. The simulator is not restricted to any communication protocol or connection mode; however, in this paper, we focus on the connectionless per-hop TCD routing.

The ISL network simulator was implemented in a modular approach in the discrete event driven simulation tool OPNET modeler. The required functionality for the routing study was

$$\tilde{T}_{\text{exp}} = \frac{(t_1 - t_s) \cdot n(t_s) + \sum_{i=1}^{K-1} (t_{i+1} - t_i) \cdot n(t_i) + (t_s + T_{\text{interval}} - t_K) \cdot n(t_s + T_{\text{interval}})}{T_{\text{interval}}} \cdot \frac{L_{\text{average}}}{C} \quad (3)$$

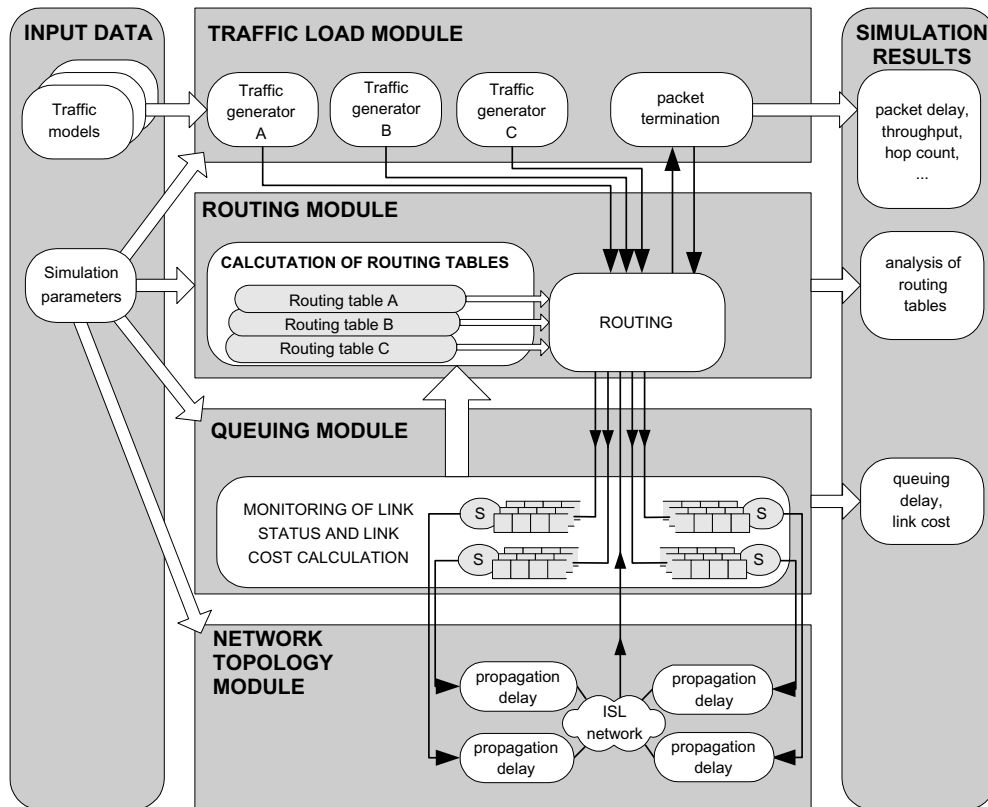


Fig. 9. ISL network simulation model.

distributed among four main modules. In general, telecommunication network in OPNET modeler is modeled on three hierarchical levels referred as the *modeling domains* [21]. The topology of an ISL network is defined in the network domain, where each satellite corresponds to communicating entity called node. The node domain provides the modeling of individual communication devices that can be deployed and interconnected. Node models consist of smaller building blocks, which correspond to particular functions of a satellite (e.g., queues, traffic generators, routing module). The behavior of each building block is specified in process domain using finite state machines and extended high-level language.

Traffic generators were implemented in the **traffic load module** to generate packets according to the selected traffic source model, i.e., generating empirical or Poisson traffic. Regardless of the selected model, three independent packet generators were used in each satellite, one for each traffic class. Different interarrival times were considered for the three traffic classes, resulting in fixed proportions of 20%, 30%, and 50% of generated packets belonging to the traffic classes A, B, and C, respectively. Packet lengths for all traffic classes are distributed as described for both traffic source models in Section II.

Packet generators are modulated by the imported normalized traffic load matrix, which takes into account geography and time dependent intensities of traffic sources. Thus, each satellite generates packets according to the traffic intensity between satellite pairs for the particular traffic class and fills their headers with all the necessary data for routing and subsequent analysis.

The **routing module** is concerned with updating routing tables and forwarding packets to the next node on the route to their final destination (per-hop routing). Routing tables are updated every 30 s. The next node on the route is determined from the currently valid routing table, calculated with a centralized version of the Dijkstra shortest path algorithm for the traffic classes A and C, and with Bellman–Ford algorithm for the traffic class B.

Routing algorithms use the information about the link cost acquired in the **queuing module**, which monitors the link status and collects data from incoming packets. Packets received from the routing module are placed into an appropriate first-in–first-out (FIFO) queue according to the traffic class denoted in the header. Three separate FIFO queues are implemented for each outgoing link, one for each traffic class. Three different scheduling policies are used to evaluate the performance of routing procedures. Priority queueing (PRI) assumes different priorities for different traffic classes. The highest priority is assigned to the delay-sensitive traffic class A and the lowest priority to the traffic class C which has no specific requirements, i.e., $\text{Pri}(A) > \text{Pri}(B) > \text{Pri}(C)$. Thus, PRI scheduling policy provides an upper limit for the routing performance for traffic class A. With a weighted round robin (WRR) scheduling policy, the minimum shares of bandwidth allocated to the A, B, and C traffic classes are 0.5, 0.3, and 0.2, respectively. Hence, WRR policy provides some minimum bandwidth to the traffic class C, even in a heavily loaded network. To compare performance with that of conventional single service routing, packets can also be scheduled with a FCFS (*first come first served*) scheduling policy.

TABLE II
AVERAGE RELATIVE DELAY DEVIATION AND AVERAGE NORMALIZED DATA
THROUGHPUT FOR DIFFERENT SIMULATION SCENARIOS

Traffic model	Routing	Traffic Class	Scheduling policy	average RDD [%]		average NDT [%]	
				Empirical traffic	Poisson traffic	Empirical traffic	Poisson traffic
NET-UNI	SS	A, B, C	FCFS	3.11	3.77	84.35	83.68
	TCD	A	WRR	2.49	3.09	85.94	85.49
			PRI	2.23	2.79	86.28	85.85
		B	WRR	9.58	10.22	87.16	86.32
			PRI	9.49	10.13	87.25	86.37
		C	WRR	2.82	3.46	85.51	84.96
			PRI	2.99	3.71	85.34	84.66
	SS	A, B, C	FCFS	2215.46	1966.65	61.77	61.75
LM-HS	TCD	A	WRR	5.70	6.09	76.06	76.83
			PRI	4.36	4.76	77.25	78.25
		B	WRR	13.85	14.26	78.18	78.63
			PRI	12.85	13.38	78.70	79.16
		C	WRR	593.28	132.50	70.45	71.08
			PRI	567.98	131.04	70.26	70.72
	SS	A, B, C	FCFS	2215.46	1966.65	61.77	61.75
	TCD	A	WRR	5.70	6.09	76.06	76.83
			PRI	4.36	4.76	77.25	78.25
		B	WRR	13.85	14.26	78.18	78.63
			PRI	12.85	13.38	78.70	79.16
		C	WRR	593.28	132.50	70.45	71.08
			PRI	567.98	131.04	70.26	70.72

In the **network topology module**, which defines the topology of the ISL network, the path length is taken into account via the propagation delay between satellites connected with ISLs. The propagation delays are computed in advance at the beginning of each snapshot.

The total network traffic has a very significant impact on simulation complexity, thus both the number of packets generated per second and the link capacity have been scaled down to reduce the otherwise overwhelming simulation requirements. The results given in the following section were obtained for a total network traffic load of 8000 packets/s, while the ISL link capacity C was set to 4 Mb/s. The results obtained with such assumptions are not suitable for dimensioning link capacity or length of queues, but they can be used for direct comparisons between different simulation runs with various simulation parameters, provided that the length of queues is adequate to prevent any overflow. Simulation for each combination of simulation parameters was run for one orbit period which, in the reference satellite constellation, is equal to 114 min.

VI. SIMULATION RESULTS

In this section, representative simulation results are presented for the adaptive TCD routing in different traffic scenarios. In particular, the performance of routing procedures in empirical traffic is evaluated and compared with that obtained in Poisson traffic. In addition, the impact of TCD routing on the network performance is compared with that of SS routing.

Simulation results for all three traffic classes are given in terms of average packet delay in the network, which is best suited to evaluate the performance of traffic class A, and average normalized data throughput, which is the most suitable performance measure of traffic class B. The results for the average packet delay (D_P) reflect only the delay component, that is a result of the propagation delay and the queueing delay experienced by the packet when it traverses the network from a source node to a destination node. For a proper comparison of results obtained with different traffic scenarios we defined the relative delay deviation (RDD), given by (8), where $D_{P_{\min}}$ denotes the delay the packet would experience if it traversed the

shortest paths in an empty network. Thus, RDD is a relative measure showing the efficiency of a routing algorithm under given traffic load conditions (i.e., performance decrease due to queueing delay and actual propagation delay on the path relative to the shortest path)

$$RDD = \frac{D_P - D_{P_{\min}}}{D_{P_{\min}}} \cdot 100\%. \quad (8)$$

Normalized data throughput (NDT) is defined by (9), where DT denotes the data throughput, calculated as the quotient of the packet length (L) and the difference between the time when the packet was transmitted and the time when it was inserted into the queue

$$NDT = \frac{DT}{C} \cdot 100\%; \quad \text{where } DT = \frac{L}{T_{\text{PacketTransmitted}} - T_{\text{PacketEnqueued}}}. \quad (9)$$

Results for RDD depend on the path length of a particular packet and vary with time so, in this paper, we present these results averaged over all packets in the network and over the simulation time, i.e., as average RDD. Likewise, results for NDT are calculated for a given packet on a particular link while, to compare different simulation runs, we present the average value over all packets on all links for the entire simulation time, i.e., as average NDT. The simulation results for average RDD and average NDT for different types of traffic, traffic flow models, scheduling policies and routing procedures are summarized in Table II.

Simulation results obtained for the homogeneous NET-UNI traffic flow model show that the routing procedure performs slightly better with empirical traffic than Poisson traffic. However, the difference is very small in terms of both average RDD and average NDT, because the link cost, and consequently the computed routing tables, depends largely on the long term average (over routing table update time) of traffic intensity on a particular link. This long term average of traffic intensity depends on the total network traffic load and the distributions of traffic flows and is, thus, similar for empirical and Poisson traffic. Furthermore, in the case of the NET-UNI model, with

homogeneously distributed traffic flows and relatively low link load, the difference between different scheduling policies and routing procedures is also very small. Still, the best performance in terms of average RDD is achieved with TCD routing for traffic class A, since routing tables for this traffic class are optimized to end-to-end packet delay. As expected, the lower bound of RDD (2.23% for chosen simulation parameters) is achieved with the PRI scheduling policy. Similarly, traffic class B achieves the best average NDT, because its routing tables are optimized according to the available bandwidth.

For the nonhomogeneous LM-HS traffic flow model, the differences between different simulation scenarios increase. SS routing makes no distinction between different traffic classes, so that the average packet delay for Poisson traffic is almost 20 times greater, and for empirical traffic even 22 times greater, than on the shortest paths in an empty network. Under such conditions the introduction of advanced TCD routing with differentiation of traffic classes becomes essential. As shown in Table II, TCD routing performs similarly for empirical and Poisson traffic for traffic classes A and B. For traffic class C, on the other hand, TCD routing performs much better for Poisson traffic than for empirical traffic. In particular, the average RDD in case of priority queueing scheduling deteriorates from 131% in the case of Poisson traffic to 567% in the case of empirical traffic. In the case of both the NET-UNI and the LM-HS traffic flow models, the best performance, in terms of average RDD is achieved with TCD routing for traffic class A, while the best performance in terms of average NDT is obtained for traffic class B.

The comparison of routing performance using WRR and PRI scheduling policies shows that PRI scheduling policy performs better for traffic classes A and B, both in terms of average RDD and average NDT. In the case of the nonhomogeneous LM-HS traffic flow model, the PRI scheduling policy performs better in terms of average RDD for traffic class C also. With nonhomogeneous traffic load and PRI scheduling policy, some links are so heavily loaded with packets belonging to traffic classes A and B that traffic class C packets with the lowest priority experience very large delays or no throughput at all. Such a situation will be reflected in the updated traffic class C routing table so that packets will be routed along the paths with fewer packets of traffic classes A and B. Conversely, in the case of WRR scheduling policy, some minimum bandwidth is guaranteed for traffic class C, so in a heavily loaded region all traffic classes will be distributed similarly between all links, and traffic class C cannot find an alternative, lightly loaded link in the area of congestion.

Fig. 10 summarizes the results in terms of average RDD and average NDT for various traffic scenarios, and routing procedures TCD with PRI scheduling policy, and SS with FCFS scheduling policy. These results indicate that the implementation of TCD routing significantly improves the performance of traffic classes A and B, especially in more realistic nonhomogeneous LM-HS traffic scenarios. In these traffic scenarios, the performance of traffic class C is also significantly improved. The best performance in terms of average packet delay is achieved for traffic class A and in terms of average normalized data throughput, for traffic class B. In the case of NET-UNI traffic scenarios, however, the benefit of introducing TCD routing is small, and the performance is almost the same for both Poisson and empirical traffic.

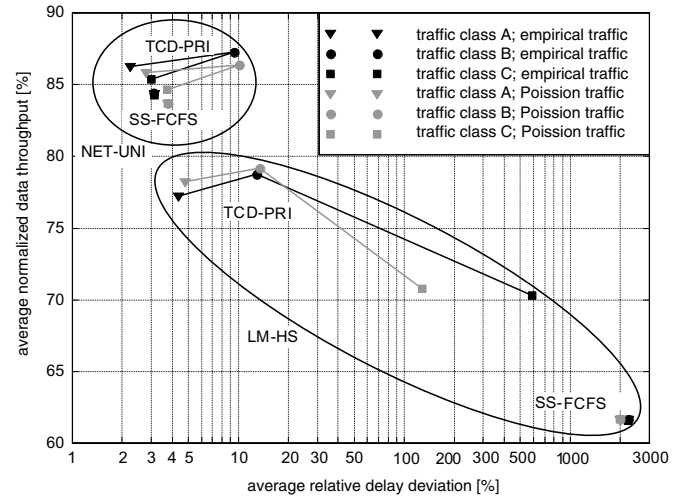


Fig. 10. Average normalized data throughput versus average relative delay deviation for empirical and Poisson traffic.

In this paper, the effect of different traffic source models and routing procedures on network performance is studied also from the network point of view, focusing on the link load and corresponding positions of connected satellites. The simulation results for link load are calculated from the accumulated traffic volume (ATV) over a certain time period (τ), as shown in (10), and normalized with respect to the link capacity (C) so that the highest value does not exceed one. Time period τ is set to 10 s, which means that the normalized link load (NLL) is measured three times within one routing table update interval T_{interval}

$$\text{NLL} = \frac{\text{ATV}}{\tau \cdot C}. \quad (10)$$

From the network point of view it is most important to identify the most heavily loaded links. In our study, we assume that the link is heavily loaded if the normalized link load is higher than 0.95. The simulation results depicted in Figs. 11 and 12 show the subsatellite points of the satellites with at least one heavily loaded link for SS and TCD routing procedures, respectively. These results were obtained with empirical traffic in nonhomogeneous LM-HS traffic flow model and PRI scheduling policy, but similar results were obtained for Poisson traffic and WRR scheduling policy.

Fig. 11 shows that, in the case of SS routing procedure, traffic tends to use links at higher latitudes where interplane ISLs are considerably shorter than at lower latitudes. As SS routing does not differentiate between different traffic classes, all of them are routed according to the same link-cost metrics as defined for traffic class C along the same routes, which results in congestion in shorter links at higher latitudes. On the other hand, the use of advanced TCD routing leads to significant reduction of heavily loaded links, as depicted in Fig. 12. Still, all heavily loaded links remain in the Northern Hemisphere, which is a consequence of the traffic flow model being considered. The LM-HS traffic flow model assumes homogeneous distribution of traffic sources on landmasses, most of which are in the Northern Hemisphere, and also the highest traffic volume is directed towards North America and Europe, which coincides with the locations of heavily loaded links.

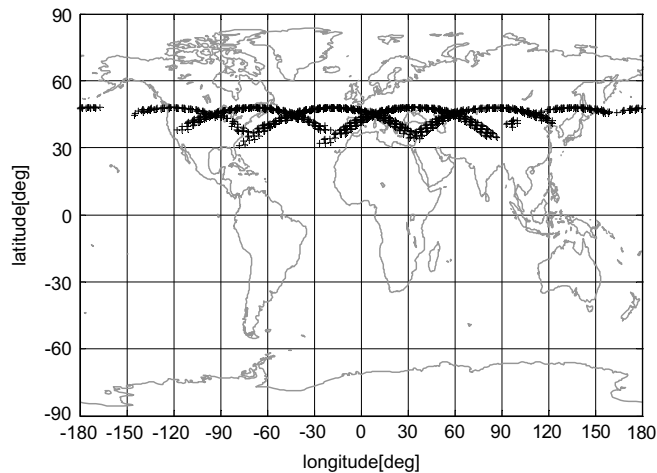


Fig. 11. Positions of satellites with NLL higher than 0.95 for LM-HS and SS routing procedure.

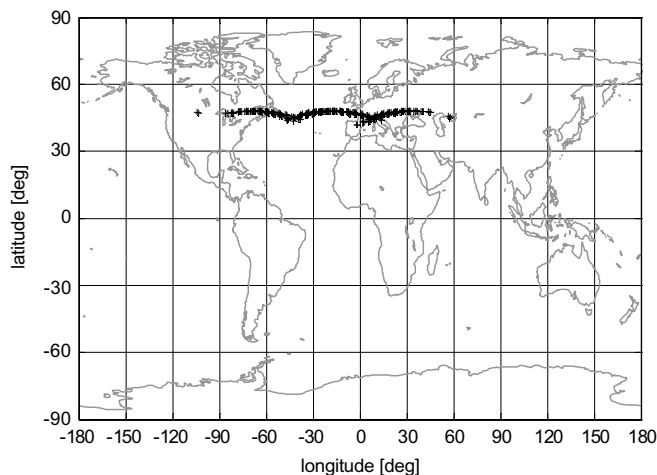


Fig. 12. Positions of satellites with NLL higher than 0.95 for LM-HS and TCD routing procedure.

VII. CONCLUSION

Satellite networks are best suited for covering large geographical areas and providing a variety of mainly asymmetric services and applications, and are expected to play an important role in the next-generation Internet. TCD routing, proposed in this article, is one of the mechanisms that allow differentiation of services with different performance requirements. In order to evaluate the performance of TCD routing, we defined three representative traffic classes, the delay-sensitive traffic class A, the throughput-sensitive traffic class B, and the best-effort traffic class C, each with its own link-cost metrics and optimization criteria used for calculation of routing tables.

The performance evaluation has been carried out using an ISL network simulator, assuming a reference constellation based on the orbital parameters of the CELESTRI constellation. In the simulator, TCD routing was implemented with priority and weighted round robin scheduling policies. For comparison reasons a reference single service routing was also implemented using FCFS scheduling policy. Packets of different traffic classes were generated using a specifically developed empirical traffic source model that resembles distributions of interarrival times

and packet lengths obtained from the real backbone Internet traffic trace. Results obtained with empirical traffic were compared with those obtained using equivalent Poisson traffic. In both cases, we considered homogeneous and nonhomogeneous traffic flows between satellites, defined by a suitable model of geography and time dependent variation of traffic intensity.

The performance of routing algorithms was evaluated in terms of average relative delay deviation, average normalized data throughput and normalized link load. The simulation results have shown that the introduction of TCD routing significantly improves network performance in comparison with a conventional single service routing for all traffic classes, regardless of traffic generators and traffic flow models used. However, in the more realistic nonhomogeneous LM-HS traffic flow model, TCD routing leads to significantly higher performance improvement than in a reference homogeneous NET-UNI traffic flow model. In addition, TCD routing significantly reduces the number of heavily loaded links in the ISL network, which is important from the network operator perspective.

The simulation results presented for empirical and Poisson traffic exhibit similar performance. There are several reasons for this similarity. First, the link costs used in the routing table are calculated according to the long term average of traffic intensity on a particular link. Thus, it depends only on the total network traffic load and the distribution of traffic flows, and not on the short term statistical properties of the traffic source generator. Second, traffic source models in the ISL network simulator are used for generating aggregate traffic flows on satellites, and not traffic flows generated by individual applications, where consequent statistical multiplexing may lead to heavy tail distribution of aggregate traffic flow. Finally, the comparison is made with performance measures defined on the level of aggregate traffic flows, which are suitable for evaluating routing procedures but do not show the effects of using different traffic source models on the level of particular traffic flow, such as burstiness and delay jitter. The only notable difference is obtained for best-effort traffic class C under nonhomogeneous traffic load conditions, where the performance of TCD routing of empirical traffic deteriorates significantly due to the more unfavorable characteristics of real backbone Internet traffic than of Poisson traffic.

REFERENCES

- [1] M. Mohorcic, A. Svigelj, G. Kandus, and M. Werner, "Performance evaluation of adaptive routing algorithms in packet switched intersatellite link networks," *Int. J. Satell. Commun.*, vol. 20, pp. 97–120, 2002.
- [2] M. Mohorcic, M. Werner, A. Svigelj, and G. Kandus, "Alternate link routing for traffic engineering in packet-oriented ISL networks," *Int. J. Satell. Commun.*, vol. 19, pp. 463–480, 2001.
- [3] —, "Adaptive routing for packet-oriented inter satellite link networks: Performance in various traffic scenarios," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 808–818, Oct. 2002.
- [4] M. Mohorcic, A. Svigelj, G. Kandus, Y. F. Hu, and R. E. Sheriff, "Demographically weighted traffic flow models for adaptive routing in packet switched non-geostationary satellite meshed networks," in *Computer Networks*, pp. 113–131, Apr., 16 2003.
- [5] A. Svigelj, M. Mohorcic, and G. Kandus, "Traffic class dependent routing in ISL networks considering various traffic scenarios," in *Mobile and Personal Satellite Communications 5: Proc. 5th European Workshop on Mobile/Personal Satcoms (EMPS)*, E. Del, Ed., Baveno, Italy, Sept. 24–26, 2002, pp. 45–52.
- [6] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, "The changing nature of network traffic: Scaling phenomena," *Comput. Commun. Rev.*, vol. 28, no. 2, pp. 5–29, 1998.

- [7] O. Cappe, E. Moulines, J. C. Pesquet, A. Petropolu, and X. Yang, "Long-range dependence and heavy-tail modeling for teletraffic data," *IEEE Signal Processing Mag.*, vol. 19, pp. 14–27, May 2002.
- [8] P. Abry, R. Baraniuk, P. Flandrin, R. Reidi, and D. Veitch, "Multiscale nature of network traffic," *IEEE Signal Processing Mag.*, vol. 19, pp. 28–46, May 2002.
- [9] P. Baran, "The beginnings of packet switching: Some underlying concepts," *IEEE Commun. Mag.*, vol. 40, pp. 42–48, July 2002.
- [10] W. Stallings, *High-Speed Networks: TCP/IP and ATM Design Principles*. Englewood Cliffs, NJ: Prentice-Hall, 1998.
- [11] N. W. MacFadyen, "Traffic characterisation and modelling," *BT Technol. J.*, vol. 20, no. 3, pp. 14–30, 2002.
- [12] W. Willinger and V. Paxson, "Where mathematics meets the internet," *Notices Amer. Math. Soc.*, vol. 45, no. 8, pp. 961–970, 1998.
- [13] B. Ryu, "Modeling and simulation of broadband satellite networks: Part II—traffic modeling," *IEEE Commun. Mag.*, vol. 37, pp. 48–56, July 1999.
- [14] J. Miceel, *Passive Measurement and Analysis* (National Laboratory for Applied Network Research), June 2003.
- [15] K. Thompson, G. J. Miller, and R. Wilder, "Wide area internet traffic patterns and characteristics," *IEEE Network*, vol. 11, pp. 10–23, 1997.
- [16] A. Kos, P. Homan, and J. Bester, "Performance evaluation of a synchronous bulk packet switch under real traffic conditions," *IEICE Trans. Commun.*, vol. E86-B, no. 5, pp. 1612–1624, 2003.
- [17] B. Ryu and A. Elwalid, "The importance of the long-range dependence of VBR video traffic in ATM traffic engineering: Myths and realities," presented at the ACM SIGCOMM'96, Stanford, CA, Aug. 1996.
- [18] M. Werner, J. Frings, F. Wauquiez, and G. Maral, "Topological design, routing and capacity dimensioning for ISL networks in broadband LEO satellite systems," *Int. J. Satell. Commun.*, vol. 19, pp. 499–527, 2001.
- [19] M. D. Kennedy and P. L. Malet, *Application for Authority to Construct, Launch and Operate the Celestri Multimedia LEO System*. Washington, DC: Filing to FCC, 1997.
- [20] G. Apostolopoulos, D. Williams, S. Kamat, R. Guerin, A. Orda, and T. Przygienda, *QoS Routing Mechanisms and OSPF Extensions*, 1999. RFC 2676.
- [21] OPNET Modeler (2003, June). [Online]. Available: www.opnet.com



Ales Svigelj received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the University of Ljubljana, Ljubljana, Slovenia, in 1997, 2000, and 2003, respectively.

He is a Research Fellow in the Department of Digital Communications and Networks, Jozef Stefan Institute, Ljubljana. From 2000 to 2001, he spent one year at Leeds Metropolitan University, Leeds, U.K., where he worked as a Research Assistant in the Mobile Networks and Applications Group. His research interests concern the development of telecommuni-

cations systems, network protocols and architectures for satellite and terrestrial mobile communication systems. He participated in several national and international projects including COST Actions, INCO-Copernicus project ATMIS-TMS, and EU funded R&D IST project SUITED. Currently, he is active in COST Action 272.



Mihael Mohorcic (M'02) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the University of Ljubljana, Ljubljana, Slovenia, in 1994, 1998, and 2002, respectively, and the M.Phil. degree in electrical engineering from University of Bradford, Bradford, U.K., in 1998.

He is a Research Fellow in the Department of Digital Communications and Networks, Jozef Stefan Institute, Ljubljana. From 1996 to 1997, he spent 12 months as a Visiting Scientist with Satellite Mobile Group, University of Bradford. His research inter-

ests include development and performance evaluation of network protocols and architectures for mobile and wireless communication systems, and resource management in satellite networks with the emphasis on adaptive routing algorithms and traffic engineering. He has participated in several international projects considering terrestrial and satellite mobile communications and stratospheric telecommunication systems, including COST Actions and EU funded 4FP and 5FP R&D projects. He is a national delegate to the COST 272 Action and a national representative in IST project IDEALIST-5FP.



Gorazd Kandus (M'85) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the University of Ljubljana, Ljubljana, Slovenia, in 1971, 1974, and 1991, respectively.

Currently, he is Head of the Department of Digital Communications and Networks, Jozef Stefan Institute, Ljubljana, and an Associate Professor with the Faculty of Electrical Engineering, Computer Science and Information Technology, University of Maribor, Maribor, Slovenia. He spent a year at Worcester Polytechnic Institute, Worcester, MA, as

a Fulbright Fellow and five months as a Visiting Scientist at the University of Karlsruhe, Karlsruhe, Germany. His main research interests include design and simulation of mobile and wireless communication systems and development of new telecommunication services. He participated in various national and international projects considering mobile and wireless communication systems and services including several COST Actions, INCO-Copernicus projects, and European 5FP IST projects. He is a national delegate to the COST 273 and COST 279 Actions.

Dr. Kandus is a member of Upsilon Pi Epsilon.



Andrej Kos (S'98–M'03) graduated in 1996 and received the M.Sc. degree in telecommunications from the University of Ljubljana, Ljubljana, Slovenia, in 1999. He is a Senior Researcher with the Faculty of Electrical Engineering, University of Ljubljana, where he is working toward the Ph.D. degree.

He has extensive research and industrial experience in the analysis, modelling, and design of advanced telecommunications elements, networks, and systems. His current work and research focuses on next-generation packet switching and management.

His other fields of expertise include mobile communications and intelligent networks.

Mr. Kos is a member of the Telemanagement Forum and the Institute of Electrical, Information, and Communications Engineers (IEICE).



Matevž Pustišek (M'01) was born in Ljubljana, Slovenia, in 1969. He received the Dipl.Ing. degree in electrical engineering and the M.S. degree from the University of Ljubljana, Ljubljana, Slovenia, in 1992 and 1997, respectively.

Currently, he is with the Faculty of Electrical Engineering, University of Ljubljana, as a Teaching Assistant. He dedicates himself to the study of broadband systems and services, including telemanagement, QoS mechanisms in network elements and networks and IP based systems. He sets a special

focus on the research and simulation of packet switches.



Janez Bešter (M'93) graduated in 1979. He received the M.Sc. and Ph.D. degrees in telecommunications from the University of Ljubljana, Ljubljana, Slovenia, in 1982 and 1995, respectively.

Currently, he is Head of the Laboratory for Telecommunications and Assistant Professor, Faculty of Electrical Engineering, University of Ljubljana. His 20 years of research and development activities, as well as teaching, have focused on the field of planning, realization and management of telecommunication systems and services, together with the application of information technologies in education.

Dr. Bešter is a Member of the Association for Computing Machinery (ACM) and an Associate Member of the Institute of Electrical, Information, and Communications Engineers (IEICE).

Optimization of Multi-attribute User Modeling Approach

Matevž Pogačnik, Jurij Tasič, and Andrej Košir

Abstract: In this paper we present a novel approach to user modeling for a personalized selection of TV programmes. The proposed solution requires minimal or no explicit feedback and preliminary results show that it works well with varying quality of content descriptions including poor genre classification and arbitrary number of descriptonal attributes. The basis of the proposal is a hierarchical user model structure. It has been developed for the TV-Anytime standard but can be adapted for use elsewhere.

Keywords: User modeling, Personalization, TV Anytime, Interactive TV, Support vector machines, Hierarchical user model

1. Introduction

One of the major problems encountered by users, when searching for multimedia content and information, is the amount of available content. An everyday example of it is checking the evenings schedule for 50 or more available television (TV) channels. It is a time consuming effort, which no user really likes to do. Consequently, users tend to check only some of the most popular TV channels and therefore miss many interesting TV programmes.

The solution to such problems is automatic filtering of content listings (Electronic Programme Guides-EPGs), based on user modelling. The process is also known as personalized content retrieval. Basic procedure involves the evaluation of a content item based on its descriptonal attributes (meta-data) and its comparison with the user model. The result of this process is classification of the content item into classes of 'liked' and 'disliked' items.

A lot of research effort has been dedicated to this field, starting with user modelling for selection of Web documents and newsgroup postings. Many different approaches have been used, but the most popular among them are term vectors [1–4], Naive Bayesian classifier [1, 2], decision trees [5], Latent Semantic Analysis (Indexing) [2] and some other machine learning techniques. These techniques have been extended and used in different fields, among which is personalized selection of TV programmes [6–11].

The field of personalization, focused on TV programmes, is of special interest to us. Firstly, it is targeting the widest possible audience and secondly, it is experiencing major changes in terms of content description stan-

dardization and evolution of the TV watching paradigm. The standardization effort is being addressed by the TV-Anytime forum [12, 13], which is setting up a standard for multi attribute content descriptions. The importance of the TV-Anytime standard is evident from the list of its member organizations (major broadcasters, hardware and software manufacturers, Telcos, etc.). For details see Section 2.1. The changing of the TV watching paradigm is a consequence of a fast development in the field of consumer electronics. The advent of Personal Video Recorders (PVRs) has brought massive storage devices, which will soon be capable of storing around 100 hours of multimedia content for a reasonable price [12]. The prototypes are already capable of running Java applications on built-in Multimedia Home Platform (MHP) and have Internet connections, which provide access to additional information for the user oriented applications [14], like EPGs for example. Among the most interesting applications in this domain are undoubtedly personalized TV programme recommenders.

Making the personalized selections of content as relevant as possible requires good content descriptions, good choice of selection algorithms and appropriate structure of the user model. The TV-Anytime standard [15] provides content descriptions with a number of attributes including programme genre, title, synopsis, actors, production country etc. The genre attribute takes a value from a predefined hierarchical taxonomy, which provides information on relatedness of content genres.

Since a lot of the attributes in the TV-Anytime standard are arbitrary, the quality of content description may vary. The user model and selection algorithm should enable good performance, regardless of the quality of content description. The system should provide some recommendations shortly after the beginning of user-system interaction, with the capability of frequent, computationally undemanding updates of user profiles. The last requirement is especially important for multi-user systems such as web services. The concept of the user interaction with the system, should not require explicit ratings for the adaptation of the system too often, if at all.

In this paper we propose a novel approach to classification of multimedia content described by the TV-Anytime meta-data. The two-class classification (liked, disliked) is based on calculations of similarity between content description and the user model, which is calculated separately for each of the descriptonal attributes. Calculated attribute similarities are then combined into total similarity measure using Support Vector Machines (SVM), which is, to the best of our knowledge, a novel approach in this field. For computations of attribute similarities we use a unique structure of hierarchical partial

Received October 23, 2003. Revised January 13, 2004.

Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, Ljubljana, Slovenia

Correspondence to: M. Pogačnik.

E-mail: matevz.pogacnik@fe.uni-lj.si

user profiles, which can handle poor content descriptions including poor genre classification and missing description attributes. The adaptation of the user profiles is based on implicit feedback.

We have tested our model on a Web based implementation of the personalization system. We are aware of the fact that such an approach simulates a 3rd party personalization service well, but it may be different when used on a home PVR. However, the feedback from the usage history of a PVR is even more informative in terms of expressed user preferences. Therefore we believe that the results obtained would be improved if this service would be used on a PVR.

In Section 2 some related work and state-of-the-art is presented, including the TV-Anytime standard. Section 3 gives the problem statement, while Section 4 presents our novel approach. It includes detailed description of mathematical tools and models used, the mechanisms for calculations of similarity between user model and content description and finally the optimization task. Section 5 presents the experimental results, their analysis and comments. We resume the discussion in Section 6 with some concluding remarks.

2. Related work

Generally there are two approaches to content filtering and selection. The first is individual, also known as content based filtering (CBF), the other is collaborative filtering (CBR). Although our system was designed to use both approaches, the focus of this paper will be only CBF.

There are a lot of approaches in the domain of CBF, but they differ in many ways. Some of the systems for example, generate recommendations of TV programmes, which are watched at the time of their broadcast, therefore information about the day and time of broadcast are important [7–9]. The TV Anytime approach [12] is not concerned with the day and time related issues. It assumes that users have a massive storage PVR, able to record many hours of programmes, which can be recorded in advance and watched at any time.

Another issue is the problem of different meta-data standards and description schemes, which are being used for classification (selection) of TV programmes. Description schemes used in existing approaches include: only keywords [16], a combination of attributes like genre, daytime and channel [8], genre, channel and actors [10], genre and keywords [11] and genre, actors, directors, keywords, daytime and channel in [7, 9]. One of the approaches [6] is using attributes similar to the TV-Anytime standard (genre, actors, director, synopsis, production country, production year) and is from this point of view the most similar to ours. Although, in [17] they use the TV-Anytime description standard, but they make recommendations based only on the genre attribute.

Similarly, a number of mathematical tools are used for personalized content filtering. These include: Naive-Bayes classifier [7, 9, 11], decision trees [7, 9], Bayesian

Belief networks [8] and weighted sum of attribute similarities [6, 10]. In the last approach, the similarity between content and the user model is first computed for each of the discretionary attributes and then combined into the overall similarity value [6, 10].

All the above-mentioned issues make a sensible comparison of results difficult. Nevertheless, some authors have compared a number of approaches [7, 9] and report that the Naive-Bayesian classifier often outperforms others. In [18] some comparisons have been made between decision trees, Bayesian belief networks (BBN), correlation methods and vector methods. Authors report that BBN and correlation methods performed best in terms of classification accuracy. In terms of processing power and memory requirements BBN outperformed correlation methods during the content classification process, but were very demanding during training and building of the belief network structure. This means that updating of the BBN is not a trivial task especially if the service should run on a central server, serving a big number of users and regularly updating their profiles.

We should also mention, that most of the authors don't report the details about mathematical basis for their algorithms which makes comparison of approaches difficult.

Another important issue is the type of feedback that systems get from users. Some approaches use explicit feedback for training of the system [6, 10, 11], others implicit feedback [7, 9, 17] and some even combine both [8]. It is clear that explicit feedback gives better results, since user preferences are explicitly stated, but on the other hand this approach is much more annoying for users. Systems using implicit feedback are user-friendlier, but have to make decisions relying on incomplete and uncertain information. According to [7] most of the users require only minimum interaction with the system, which provides them with programme recommendations, and are not willing to give much feedback.

Therefore our goal was to build a user-friendly personalization system, mostly using implicit feedback for adaptation of the user model with a possibility for improvement if minor amount of explicit feedback is given. In the near future it should be possible to adapt our system to run on a PVR with MHP support and in this way get access to the PVR usage history (play, record, stop, delete). This information is undoubtedly richer than information about selected and ignored programmes, which should prove as an upgrade to the existing approach. In fact, such an approach has, in a limited way, already been used in [17] and is also being standardized by the MPEG-7 [19].

2.1 TV-Anytime meta-data standard

The TV-Anytime idea and its genre hierarchy form an important part of our work. Therefore, before we focus on the proposed approach, let's take a brief look at the TV-Anytime standard.

TV-Anytime is a forthcoming standard for digital interactive television and a result of work of the TV-

Anytime Forum [13]. Member organizations come from Europe, the USA, and Asia, and are drawn from a wide variety of industries. Therefore it is reasonable to expect that use of the standard will become widespread. Standard specifications cover many areas including Rights Management and Protection, Metadata and Content Referencing. We will focus only on the usage of metadata.

2.1.1 Metadata structure

TV-Anytime standard [15] enables simple and complex descriptions of content. Complex descriptions may include a multidimensional classification system, where each dimension is used to describe content from a single viewpoint. However, it is more probable that most of the programmes will be described using simpler, BasicContentDescriptionType, which specifies a number of description fields or attributes, as we call them. Among them are mandatory, recommended and optional attributes. The mandatory attribute is 'Title', while recommended attributes include 'Synopsis', 'Genre', 'Language', 'Member Of' and 'Credits'. 'Credits' attribute contains lists of actors, director(s), writer(s), music composers, etc. 'Member Of' attribute contains a reference to a group of related programmes like 'Star Trek' series or all 'Euroleague' football matches, for example. It is also called 'Group CRID' attribute and this notation will be used throughout this paper. Among optional attributes are 'Awards list', 'Production date', 'Production location', 'Parental guidance' and others.

The possible values that the genre attribute may take are hierarchically organized in a tree structure. The tree structure gives additional information about relatedness between genres and is as such more advanced than a linear list of genres. An example of the genre tree is presented in Fig. 1.

The genre hierarchy, used in our experiments was taken from the myTV project [14] and is less detailed as the present TV-Anytime genre hierarchy [15]. This is due to the fact that the majority of TV programme descriptions available to us at the beginning of our work had originated from the work on the myTV project. Nevertheless, our proposed concept works independently of any genre hierarchy used and is as such universal.

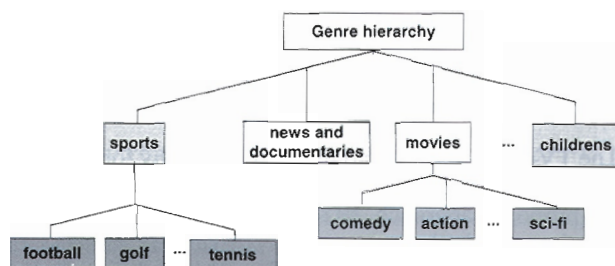


Fig. 1. Example of the genre hierarchy.

3. Problem statement

From the previous section we can conclude that the following problems exist in the domain of user modelling for TV programmes.

Firstly, content descriptions are generated manually or semi automatically, therefore it is not realistic to expect that all existing attributes will always be included in the descriptions. Furthermore, some of the content, like movies for example, contain by itself more attributes than others (e.g. actors, director, etc.). The TV programme descriptions, that were available to us during the testing, contained very heterogeneous levels of description details.

Another problem is labelling of TV programmes with genre description that are too general. As was stated in the introduction, the genre categories are hierarchically organized in a tree structure (see Fig. 1), but the genre classifications, assigned to programmes, are very often too general. An example of it is a football match, which should obviously be labelled as 'Sports:Football', but is instead labelled as Sports:Unknown or 'Sports:Other'. This problem has also been reported by other authors [8].

The next problem is the so called 'cold start problem'. The algorithms and the structure of the user model should, on one hand, be able to provide some recommendations early in the learning process, while on the other hand it should be relatively easy to update the user model at any time in the process of usage.

Last but not least, there is a problem of the feedback mechanism. The dilemma between implicit and explicit feedback is always present, but we believe that users should not be required to give too much explicit feedback, if any at all, because this may lead to functional but unpopular solutions.

The goal of this paper is therefore to propose a solution for personalized selection of TV programmes, which can cope with poor genre descriptions, arbitrary number of descriptive attributes and is able to provide some recommendations early in the process without constantly bothering the user for explicit feedback.

4. Proposed user modeling approach

Our aim was to build a suitable personalization algorithm, which can cope with above-mentioned issues and which is capable of frequent adjustments to user preferences through time. The algorithm should require relatively low amount of processing in order to insure near real-time user model update and estimation of content suitability. It should also work with implicit feedback obtained from users. We should note that the algorithms and structures were designed for personalized selection of any content with multiple attribute description. Therefore we will no longer refer to 'TV programmes'; instead we will use the term 'content'.

In our novel approach the similarity between user model and content is first computed separately for each

descriptive attribute (genre, keywords, actors, director(s), etc.) and then combined into the final similarity estimate. Based on the final similarity estimate, the content is classified either as ‘liked’ or ‘disliked’. For classification purposes we tried out some of the existing approaches, like expert defined weighted sum [6], linear LMS [20], Support Vector Machines (SVM) [21]. We established that the SVM approach, using Radial-Basis-Function (RBF) kernel, outperforms others.

In addition to personalized calculations of attribute similarities the classification, based on the attribute similarities, can also be adapted to every individual user. Our novel idea is to use different SVM models (see Section 4.4.1) for prediction of classification in different stages of system usage. In the early stages a universal SVM model is used, while in the later stages a custom SVM model is built, improving the prediction efficiency. We should note that regardless of the SVM model used, the attribute similarities are always calculated using the personal user model, while using the custom SVM model for classification prediction represents just an additional step in the process of personalization.

Although the attribute similarities are calculated independently of each other, some dependence between attributes is taken into account. The genre attribute is undoubtedly the attribute carrying the most information. Therefore it is considered the principal attribute and all other attribute values in the user model are stored under the genre notation. Genres, used in the experimental user model are compliant with the concept of the TV-Anytime genre hierarchy (taxonomy). The details of the proposed approach are described in following sections.

4.1 Notations and mathematical tools

This section briefly defines our notations. Mathematical tools applied to perform user modelling are described shortly, complete descriptions of the underlying concepts are given elsewhere [20–22].

4.1.1 Basic terms and notations

Sets and maps introduced along with the terminology are based on TV-Anytime metadata standard representation, but can be used for any multi-attribute description standard. The hierarchical taxonomy of the main attribute is recommendable, but not mandatory.

To facilitate understanding of the rather complicated data structures introduced here, a brief outline of the user modelling procedure together with data structures is given first (see Fig. 4), while the exact definitions follow later. It is divided into five steps:

1. The programme genre hierarchy is represented by a tree (T). The same structure is used for the user model (UM).
2. The partial user model PUM is built by labelling the tree T with n -tuples of metadata entries Z^n .
3. The set of multimedia content is H .

4. Similarity maps Sim are applied to compute the “similarity” between content metadata and corresponding partial user models. A similarity map is computed separately for each descriptive attribute.
5. The classification of content h is approximated based on calculated similarity maps and is called a classification map.

Now we will describe these five steps in more detail:

Step 1

A directed tree (directed graph with no cycles, for reference see [22]) $T = (V(T), E(T))$ of genres is defined by the hierarchical genre taxonomy. The set of tree vertices is a set of genres $V(T) = \{g_1, \dots, g_m\}$. A hierarchical genre ordering is given by the set of directed tree edges $E(T) = \{(g_i, g_j) : g_i, g_j \in V(T)\}$, i.e. the directed edge (g_i, g_j) denotes that the genre g_j is a sub-genre of the genre g_i . An example of the genre hierarchy is presented in Fig. 1.

The attributes of metadata are grouped into n classes $A_1, \dots, A_n$ (see Fig. 2); a single attribute value is denoted by $a_j^i \in A_i$. The attribute classes used in our experiment are genre classification, keywords (derived from the title and synopsis), actors, director(s), country of origin, year of origin and group CRID, as described in Section 2.1. Consequently, the typical attribute values are: name of the actor, keyword of the title, country of origin, etc.

The metadata entry is a triple (v, w, l) , where $v \in Y$ is a string describing an attribute value a_j^i from class A_i for some $1 \leq i \leq n$, $w \in [-1, 1]$ is the weight of a_j^i , and $l \in N$ is the number of times that the user has interacted with the content items containing the attribute value a_j^i . Examples of metadata entries for all attributes are presented in Fig. 2. A set Y is a set of strings; the property when a string $v \in Y$ describes the attribute of class A_i is denoted by a predicate $P(v, A_i)$. A set of metadata entries is denoted by $Z = Y \times [-1, 1] \times N$.

Step 2

A n -tuple of sets of metadata entries Z is assigned to each user U and each genre $g \in V(T)$ called “partial user model” and denoted by $PUM(U, g) = \{(v, w, l) \in Z : P(v, A_i), 1 \leq i \leq n\}$. For a particular user U , a map

$$PUM(U, \cdot) : V(T) \rightarrow \mathcal{P}(Z)^n$$

is labelling of the tree T . The set $\mathcal{P}(Z)$ is a set of all subsets of the set Z . An example of $PUM(U, g)$, where genre g is “Movies:Drama”, is presented in Fig. 2.

A set of “partial user models” assigned to each user U is denoted by

$$UM(U) = \{PUM(U, g) : g \in V(T)\},$$

where $UM(U)$ is the user model of user U .

Step 3

A class of multimedia content items (TV programmes) is denoted by:

$$H = \{h : \text{metadata content}\}.$$

Genre ($g=A_1$)	Movies:Drama 77#25
Keywords (A_2)	murder:15#7, police:22#9, lover:6#3, family:9#3, innocent:5#5, ...
Actors (A_3)	Robert de Niro:20#9, Michael Caine:10#7 Nicole Kidman:9#4, ...
Directors (A_4)	Stanley Cubrick:9#3, Woody Allen:8#3 Joze Pogacnik:9#4, ...
Country (A_5)	USA:50#25, France:12#4, Italy:5#2,...
Production year (A_6)	-2010:9#3, -2000:30#12, -1990:35#16, -1980:22#12,...
Group CRID (A_7)	

Fig. 2. Example of a partial user model (PUM).

MD^n is a class of metadata items. Each content item $h \in H$ carries its own metadata denoted by $md(h) \in MD^n$ and belongs to a certain genre denoted by $g(h) \in E(T)$. Metadata $md(h)$ describes each of the attribute classes (genre, title and synopsis, actors, country of origin, etc.), where title and synopsis attributes both describe the keywords class. We can note this with: $md(h) = (md_1(h), \dots, md_n(h))$, where $P(md_i(h), A_i)$. An example of a metadata item is shown in Fig. 3.

Step 4

When the content h is to be “examined by the user model UM ”, for each index $1 \leq i \leq n$ a similarity among the pertained metadata $md_i(h)$ and corresponding “partial user model” $PUM(U, g(h))$ describing a class of attributes A_i is measured. The similarity is measured by a map

$$Sim_i : \mathcal{P}(Z) \times MD \rightarrow \mathbb{R},$$

Genre	Movies:Drama
Title & Synopsis	Mona Lisa George, after getting out of prison, begins looking for a job, but his time in prison has reduced his stature in the criminal underworld..
Actors	Bob Hoskins, Cathy Tyson, Michael Caine, Robbie Coltrane, Clarke Peters, ...
Director(s)	Neil Jordan
Country	UK
Production year	1986

Fig. 3. Example of a metadata item.

where $\mathbb{R}$ is a set of real numbers. Therefore, similarity map Sim is a Cartesian product of maps Sim_i ,

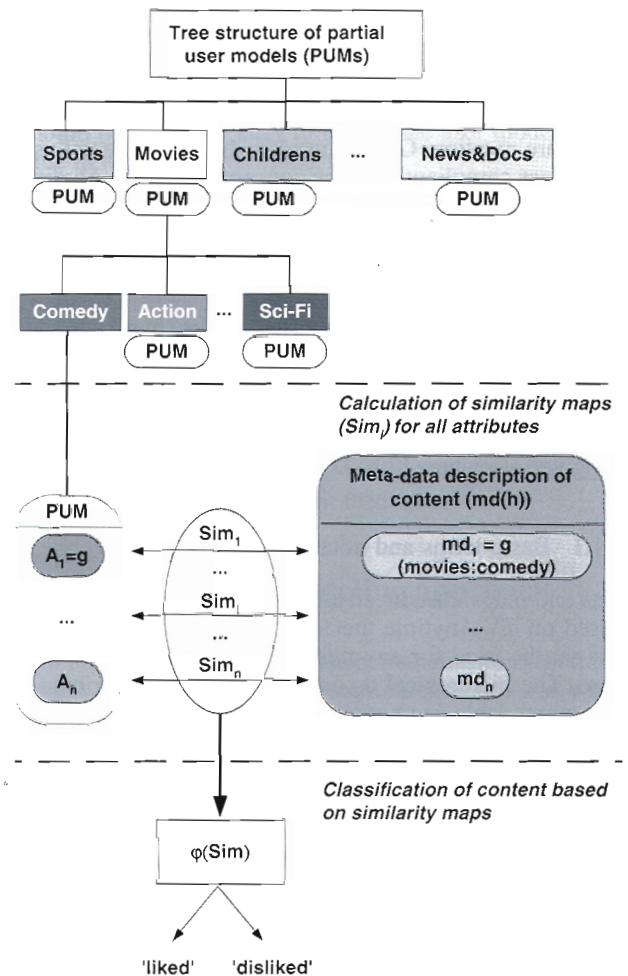
$$Sim = Sim_1 \times \dots \times Sim_n : \mathcal{P}(Z)^n \rightarrow \mathbb{R}^n.$$

Step 5

A set of decisions $Dc = \{dc_1, \dots, dc_p\}$ each user can take about a given content $h \in H$ is introduced. The classes of decisions are represented by real numbers $dc_j \in [0, 1]$. In our case, only two classes were chosen, $dc_1 = 0$ and $dc_2 = 1$. The class dc_1 represents the rejection of the content and the class dc_2 represents its acceptance.

To classify a given multimedia content $h \in H$ into decision classes according to Dc regarding all n classes of attributes, a classification map $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined.

A number of existing techniques can be used to tackle the problem of classification. Among them are the LMS-criteria and p-norm, which can be used for an estimation of linear combinations. A very promising technique which enables linear and non-linear classification is Support Vector Machines [21]. For more details see Section 4.4.1.

Fig. 4. The flow chart of the classification of content (h) through comparison with the $PUM(g(h))$.

Unfortunately, it cannot be expected to model the classification map φ accurately since the process of content selection is more complex than the models used. Therefore, the values and models obtained by classification map φ only approximate the value of the decision class dc_j . In order to perform the classification based on the value of the classification maps, a neighborhood interval of each decision class is defined, namely $I(dc_j) = [dc_j - t_j, dc_j + t_j]$. Values t_j are called decision class thresholds. The predicted decision of a user U regarding the multimedia material $h \in H$ is the decision class dc_j if and only if $\varphi(\text{Sim}(PUM(U, g(h)), md(h))) \in I(dc_j)$.

To track the content classification decisions made by the user U of the multimedia system, a set of decisions is assigned to each user, $b = (b_1, \dots, b_l)$, where $b_i \in Dc$.

Predicted user decisions are modelled by the estimation $\hat{\varphi} : \mathbb{R}^n \rightarrow \mathbb{R}$ of the classification map $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$.

User decisions are needed for the training of the classification procedure based on calculated attribute similarities and also for the performance evaluation of our approach. They were given by users during the test phase. To enable exact evaluation of the system performance they were obtained using the explicit feedback but in a typical usage scenario with the system running on the home PVR they would be obtained from the usage history of the PVR (play, record, delete content item, etc.). However, the classification training should not be confused with the training of the user model which is based on implicit feedback only. As we will show in 4.4.1 and 5.3 the classification procedure used for the active user can also be trained on a set of decisions b obtained from arbitrary users and therefore no explicit feedback from the active user is actually required. However, through the analysis of PVR usage history or occasional explicit feedback, a custom classification map (custom SVM model) can be obtained and used for performance improvement (see Section 5.3). The exact description of the classification map estimation is given in the next subsections. Map $\hat{\varphi}$ is then used in the sense of predicting user decisions.

The diagram of maps depicting the classification procedure can be summarized as follows:

$$T \times H \xrightarrow{PUM(U, \cdot) \times md} \mathcal{P}(Z)^n \times MD^n \xrightarrow{Sim} \mathbb{R}^n \xrightarrow{\varphi} \mathbb{R}.$$

4.2 Similarity maps

The calculation of similarity map $Sim_i(PUM(U, g(h)), md_i(h))$ for the i -th attribute of content item h is made using two well-known approaches from the personalization domain that have been adapted for our purpose, namely the Naive-Bayes [2, 5] and TF-IDF (Term Frequency Inverse Document Frequency) approach [2].

The similarity maps for the attributes of genre, actors, director and group CRID are calculated by using a modified Naive-Bayes approach [11], while similarity maps for attributes keywords, country of origin and production year are computed using the TF-IDF approach [2].

The Naive-Bayes approach predicts the classification probability $Pr(dc_i | a_j^i)$ of a content item h to the class of liked items dc_i , considering the presence of a particular attribute value a_j^i in the metadata item $md_i(h)$. This means the probability of content item h being accepted by the user, considering the presence of a particular actor, director, keyword, etc., in the description of the content. The modification of this approach compares on the one hand the number of occurrences, where a particular attribute value has been present in a content item that was selected or ignored by the user, with on the other hand the number of occurrences of any attribute value present in the selected or ignored content items. The modification is expressed in the form of the *sig* function (see (1)) [11]. In (1) parameter k_1 is the number of times the attribute value a_j^i occurs in the set of programme descriptions of genre $g(h)$ that the user chose to watch, whereas k_2 is the number of times the attribute value a_j^i occurs in the set of programmes that the user ignored. n_1 and n_2 are, respectively, the sum of occurrences of all attribute values in the sets of chosen and ignored programmes of genre $g(h)$.

$$sig = \frac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{\left(\frac{k_1+k_2}{n_1+n_2}\right) \cdot \left(1 - \frac{k_1+k_2}{n_1+n_2}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (1)$$

The relevance of the attribute value a_j^i is then expressed by weight u_j calculated by using (2).

$$u_j = (Pr(dc_i | a_j^i) - 0.5) \cdot (fabs(sig)) \quad (2)$$

The similarity map for the i -th attribute Sim_i is finally computed as the normalized sum of weights u_j for all the attribute values in the description of the content item. We find this approach very suitable because Naive-Bayes is known to give good results [2, 11] and because its computational requirements are independent of the number of training items. Furthermore, the *sig* function assigns higher weight to those attribute values whose conditional probabilities are calculated from more numerous samples and whose conditional probability value is therefore more relevant.

The TF-IDF approach was selected for calculations of the similarity maps for keyword, country origin and production year attributes. This is due to the fact that the distribution of attribute values for these attributes is very specific. Namely, the vast majority of movies are produced in the USA and in the years between 1980 and 2000. Similarly, some keywords are used very often, while others are less frequent.

We wanted to use an approach which would penalize attribute values with high occurrence frequency. The TF-IDF approach has exactly this characteristic and was therefore used for it. Even though it is simple, it is known to give good results [2, 18]. We have adapted it for the calculation of the similarity maps through attribute value relevance. The relative weight w of the attribute value a_j^i in the user model is multiplied by its TF-IDF factor, presented in (3), where x_{1g} is the total number of programmes, belonging to genre $g(h)$; x_{2jg} is the number of

programmes of the same genre containing the attribute value a_j^i in its description. y_{1ij} is the number of times the attribute value a_j^i is present in the programme description $md_i(h)$, while y_{2j} is the number of occurrences of all attribute values (strings) in programme description $md_i(h)$.

$$TFIDF(a_j^i) = \frac{y_{1ij}}{y_{2j}} \cdot \left(\log_2 \frac{x_{1g}}{x_{2ijg}} + 1 \right) \quad (3)$$

We should note that in the case of the ‘production year’ attribute there is only one attribute value, and the parameters y_{1ij} and y_{2j} always take the value of 1. A similar situation occurs with the ‘country of origin’ attribute, unless the programme is a co-production of two or more countries. Therefore, this part of the equation mostly brings variance in the case of the (multi-valued) ‘key-words’ attribute.

Similarity map Sim_i for attribute A_i of the content item h is then calculated as the sum of attribute value weights $TF-IDF(a_j^i)$ over all attribute values in the metadata item $md_i(h)$.

Our approach has the additional benefit in the case of poor genre classification of content. As was already stated in the introduction and reported by other authors [8], the content is very often assigned poor genre descriptions ($g(h) = \text{‘Sports:Unknown’}$ or $g(h) = \text{‘Movies:Other’}$, for example). In such cases the similarity maps are calculated using the partial user profiles of the broader (parent) genres represented in $PUM(U, g^p)$ ($g^p(h) = \text{‘Sports’}$ or $g^p(h) = \text{‘Movies’}$). Due to the limited size of this paper we are unable to present in more detail the propagation of the partial user model updates to the upper nodes in the genre hierarchy. However, we should note that these updates are made simultaneously with the updates of the lower nodes in the genre hierarchy.

4.3 Classification training

The history of decisions of one or more users U recorded in their personal user models $UM(U)$ is the input data for estimation procedure based on the classification map.

This data is based on their previous decisions when examining multimedia content items $H = \{h_1, \dots, h_l\}$. It is structured as a training set $D = \{(x_1, b_1), \dots, (x_l, b_l)\}$ where

$$x_j = Sim(PUM(U, g(h_j)), md(h_j))$$

are similarity vectors and $b_j \in Dc$ are decisions made by user U about the multimedia content $h_j \in H$.

The optimization methods that are used in order to estimate the classification map φ , which is based on training set D , are given in the following subsections.

4.4 Optimization task (classification maps)

The classification map φ is estimated on the basis of training set D obtained either from a number of users

or a single user. The criterion used in estimation of the classification map is the minimization of the error vector $\varepsilon(\varphi) = (\varphi(x_1) - b_1, \dots, \varphi(x_l) - b_l)$, $(x_j, b_j) \in D$. Recall the role of the decision map,

$$\varphi(x_j) \in I(dc_j) \Leftrightarrow b_j \in dc_j,$$

$$x_j = Sim(PUM(U, g(h_j)), md(h_j)).$$

Therefore, the optimization task is

$$\hat{\varphi} = \arg \min_{\varphi} \{||\varepsilon(\varphi)||\} \quad (4)$$

The choice of the vector norm $||\cdot||$ and the optimization procedure applied to gain the estimation $\hat{\varphi}$ is discussed in the next subsection.

4.4.1 Support vector machines

It is generally assumed that the best classification accuracy is achieved when the largest component of the error vector $||\varepsilon||$ is minimized. The optimization procedure suitable for solving the optimization task is called Support vector machines (SVM). We do not enter into the details here; see literature [21].

The SVM optimization procedure is determined by the choice of the kernel $K(x, y)$; see literature [21]. A number of available kernels were tested: linear, polynomial and the Radial Basis Function (RBF) kernel. The RBF kernel performed best and was therefore chosen for our application.

The soft classifier is used in this paper and output of the SVM procedure is given by

$$f(x) = v \left(\sum_i \alpha_i K(x_i, x) + b \right) \quad (5)$$

where $v: \mathbb{R} \rightarrow \mathbb{R}$ is a certain function, vector $\alpha = (\alpha_1, \dots, \alpha_p)$ and constant b are results of the SVM optimization task; see literature [21].

The input data to the SVM optimization procedure is the training set $D = \{(x_1, b_1), \dots, (x_l, b_l)\}$; see Section 4.3. The result is approximated classification map $\hat{\varphi} = f$, which is the solution of the optimization task; see also Section 5. It can also be seen as the SVM model represented as a collection of support vectors obtained from the training process using training set D . The SVM model can be either ‘universal’, i.e. obtained in advance from training sets of a number of users or ‘custom made’, i.e. obtained from the training set of the target user. In a typical scenario, the universal SVM model would be used as classification map $\hat{\varphi}$ in the early stages, while the custom SVM model would be built gradually in the process of user interaction with the system.

5. Experimental results

5.1 System architecture and meta-data

In our test group were 20 users. 7 of them were female and 13 of them male, ranging between 24 and 55 years of age.

The personalization service was implemented as a web service, using Java servlets, with TV programme meta-data and user profiles stored in a relational database on the server. The architecture of the system is presented in Fig. 5.

The majority of TV-Anytime meta-data were obtained from TVData Media Tribune company, which provided TV programme descriptions for 50 television channels, covering a broadcast period of 2 months. Since the meta-data contained little information about contributors (actors, directors, etc.), production country and production year, they were additionally enriched with data from the Internet Movie DataBase (IMDB) [23]. A minority of meta-data used in the evaluation was obtained from the database used during the myTV project [14]. Altogether the database contained descriptions of more than 20.000 TV programmes. The genre hierarchy used for programme descriptions are based on the myTV version of genre hierarchy.

5.2 User interaction and testing

At first login, users were asked to fill in a simple initialization form, expressing their preferences for the genres. They could fill in an arbitrary number of fields or even skip the entire initialization process. After initialization, they used the search engine, searching for TV programmes based on their descriptive attributes. In this

way they were able to look for particular actors, directors, titles, genres or use a simple keyword-based search. After every query they were presented with a list of matching TV programmes. Another possibility was browsing lists of programmes broadcast during a certain time period. If users liked any of the presented programmes, they marked their choice by pressing the 'I like it' button. In this way we tried to simulate a near-future use of such a service, where the information about selected programmes would be sent to the PVR at home and the programmes would be automatically recorded¹.

The selected and ignored programmes were used as the training set for building and adapting the user model. We should mention that in this way the user profiles were built with arbitrary training sets obtained as a result of user queries and their browsing through available TV programme listings, without unnecessary and unwanted explicit evaluations required from users.

For evaluation purposes, users were additionally asked to classify over 150 TV programme descriptions into following categories: 'I will watch it', 'I may watch it' and 'I won't watch it'. The first two were considered as the classification into the group of liked programmes, while the last one was considered as classification into disliked programmes. One half of the test set programmes were movies, whereas the other half were non-movies (sports, entertainment, news and documentaries, etc.)

5.3 Evaluation of results

We evaluated the system performance by comparing predicted programme classifications with those specified by the users.

In the prediction process, the test set was divided into two test sets. After training and updating of the user models, the attribute similarities for programmes of the first test set were calculated for all users in the test group as was described in Section 4.2. They were combined into classification prediction (liked, not liked) using the SVM mapping (see Section 4.4.1). The kernel function that performed best at SVM mapping was the radial basis function [21] with parameter values $C = 32$ and $\gamma = 0.25$. The best parameter values were assessed using the optimization tool [21]. The basis for the mapping was a SVM model (see Section 4.4.1), constructed from ratings b and programme attribute similarities Sim_i of all of the users. We therefore call it the universal SVM model. For details regarding construction of the SVM model see [21]. The ratings used for the construction of the universal SVM model were taken from the second test set. In the next step, we switched the roles of both test sets and made predictions using the identical procedure.

Our idea was that shortly after the beginning of system usage, programme classifications $\hat{\phi}$ should be predicted by the mapping of calculated attribute similarities

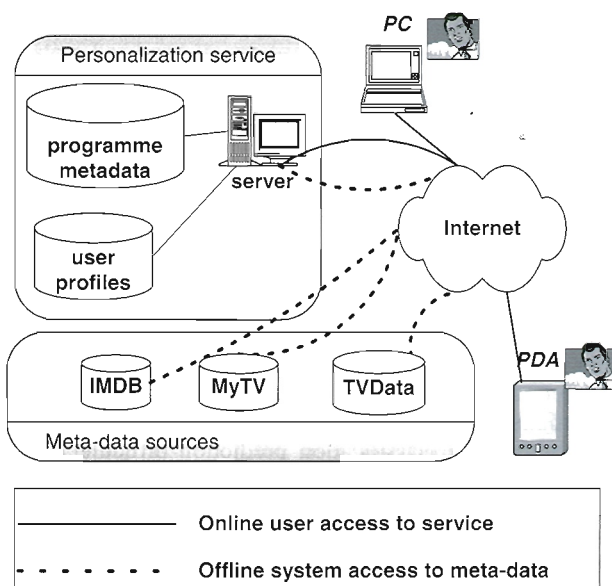


Fig. 5. Test system architecture.

¹ Remote control of the PVR through Web and Wap services was developed during the myTV project [14]

Sim_i with the universal SVM model. After a period of time the system can build a custom SVM model which should perform even better. This implies that we can perform personalization on two levels: firstly, when calculating the attribute similarities $Sim_i(PUM(U, g(h)), m_i(h))$ for content item h and, secondly, when mapping attribute similarities into prediction of classification $dc_j \in [0, 1]$. Unfortunately, building the custom SVM model requires some form of explicit feedback in order for it to be built correctly. Although the system should avoid requiring explicit feedback as much as possible, we should note that the training of the custom SVM model requires a fairly low amount of explicit feedback (for around 50–70 programmes). In addition, this feedback can be spread over an arbitrary period of time. Until the custom SVM model is fully operational, the universal SVM model is being used for prediction of classification. Some set-top box systems like TiVo [24] have already designed simple one-button mechanisms ('thumbs up'–'thumbs down'), which on the one hand are not burdening the user and on the other hand facilitate (the obtaining of explicit feedback).

The efficiency of the proposed approach was evaluated by comparing the predicted classifications with classifications given by users. Two well-known performance factors were calculated: precision and recall [2]. Precision represents the ratio between the number of relevant items retrieved and the total number of retrieved items, while recall represents the ratio between the number of relevant items retrieved and the number of relevant items in the programme collection (test set). Ideally, the goal is to achieve both high recall and precision. However, in reality we must strike a compromise. A lower threshold for class decision results in higher recall at the cost of precision and the other way round. In the case of a TV programme recommender it is preferable to achieve relatively high precision for the sake of the system's trustworthiness. A lower recall is less of a problem, because of the high number of available programmes, in particular because there are usually more 'recommendable' programmes than a user could watch.

The results are presented in Fig. 6 and Fig. 7, which show averaged values from both test-set trials.

5.4 Comment of results

In Fig. 6 we can observe that the recall values differ quite significantly for different genres. Higher values for movies (recall around 70%) are to be expected, since movies are often described with all 7 attributes (title, genre, synopsis, actors, director(s), production country and production year), while non-movies (recall around 40%) are described with only 3 attributes (title, genre and synopsis). This makes the identification of liked programmes more difficult. The precision values, on the one hand, are not that different for movies and non-movies (around 80%). This result can be attributed to the fact that a relatively small percentage of non-movies were chosen as preferable (30%) and therefore the higher drop in recall

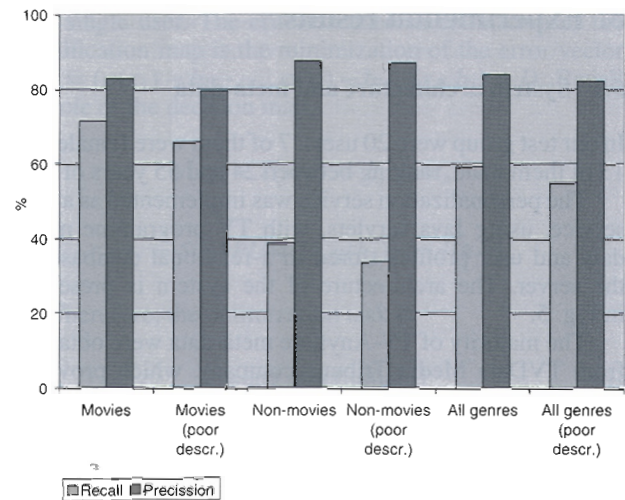


Fig. 6. Recall and precision for different genres and different genre description qualities.

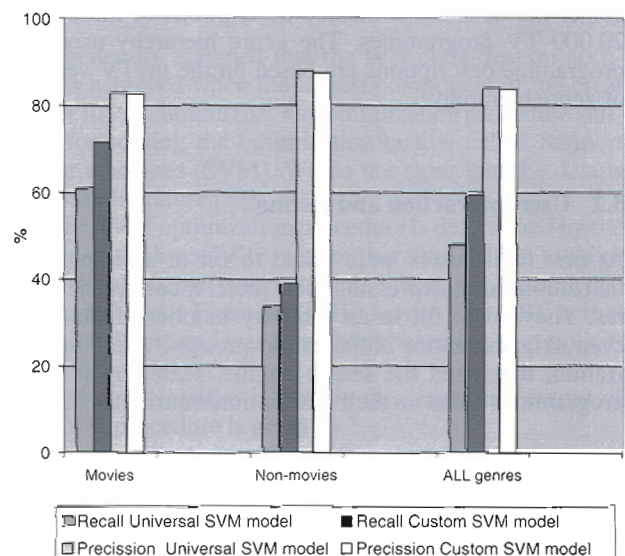


Fig. 7. Comparison of recall and precision values for universal and custom SVM models.

(higher number of false negatives) was not accompanied by the lower precision value.

The same Fig. 6 also demonstrates the advantage of our system with respect to programmes with poor content descriptions. As was stated earlier in the paper, the programmes are very often assigned poor genre descriptions like 'Sports:Unknown' or 'Movies:Unknown', which makes a classification prediction difficult. Using our hierarchical user modelling approach we were able to make classification predictions also for programmes with poor description (see Section 4.4.1). The results show that the decrease in recall and the decrease in precision values for programmes with poor descriptions is around 5% for both movies and non-movies,

which demonstrates the efficiency of our hierarchical user model.

Fig. 7 compares the efficiency of the universal and custom SVM model. As anticipated, the recall values were higher when the custom SVM model was used, for both movies and non-movies. Improvement was in the range of around 10%. The precision values were not improved, but since they were already high enough (over 80%) there is not that much room for improvement anyway.

6. Conclusions

This paper presents the design and results of a TV classification programme that combines a novel hierarchical user model for calculation of the attribute similarity with the support vector machines for class prediction. This combination (design) has produced good results in that, according to our preliminary study, the average precision values reached over 80% with recall at around 60%. The recall values for movies were even slightly higher (70%). A particular value of this approach lies in the fact that it handles well programme classification with poor genre descriptions and that the drop in recall and precision remains within acceptable limits. According to our preliminary results the drop in recall was around 5% and the drop in precision was even lower. These results are significant bearing in mind that a lot of content has poor genre classification.

Usage of Support vector machines for class prediction from attribute similarities has also proven its value. We had some doubts about the classification efficiency using the universal SVM model, but the results turned out to be satisfactory. The average precision value for all genres was again over 80% with recall slightly under 50%, while in case of movies recall reached 60%. We can conclude that the performance of the universal user model is good enough in the early stages of system usage. After the system has been in use for some time, the custom SVM model can be trained and used. The results have shown that using the custom SVM model in the later stage the average recall value for all genres can be raised to almost 60% in case of movies even to 70%.

Since the preliminary results were satisfactory, we intend to make some enhancements to the existing system. In particular we intend to combine the present content-based filtering with collaborative-filtering approach. Furthermore, we will transfer our system to a PVR with MHP support and in this way get access to the PVR usage history (play, record, stop, delete). We believe that the conclusions about user preferences, based on information about content usage, are more reliable. Related to it is the issue of the user interface (GUI) design, which can greatly improve the user-system interaction and consequently the performance of the system. Therefore we will carefully design the structure and functionality of the future GUIs. In order to prevent overfitting to some programme genres, actors, etc., we will enhance our algorithms with a forgetting function, applied

to the weights of the metadata entries in the partial user models.

This paper does not deal with the security and privacy issues. However, we are aware of them and believe that they can be solved with the standard approaches, which should prevent the misuse and compromising of the user data.

Acknowledgement. This work would not have been possible without the grant of Slovenian Ministry of education, science and sport No. S-64303 and work on the myTV project (IST-1999-11702).

References

- [1] Billsus, D.; et al.: A hybrid user model for news story classification. Proceedings of the Seventh International Conference on User Modeling, University of California, Pattaya, USA, 1999.
- [2] Hand, D.J.; Mannilla, H.; et al.: Principles of data mining. Cambridge, Massachusetts, USA: The MIT Press, 2001.
- [3] Gudivada, N.; Venkat, Raghavan V. Vijay, et al.: Information retrieval on the world wide web. IEEE Internet Computing 1 (1997), 58–68.
- [4] Salton, G.; et al.: An introduction to modern information retrieval. McGraw-Hill, 1983.
- [5] Michalski, R.S.; Bratko, I.; et al.: Machine learning and data mining-methods and applications. Chichester, England: John Wiley and Sons, 1998.
- [6] Bezzera, B.; Carvalho, F.; et al.: Speeding up recommendation systems. In proceedings of the AH2002 Workshop on personalisation in future TV, Malaga, Spain, 2002.
- [7] Buczak, A.; Zimmerman, J.; et al.: Personalization: improving ease-of-use, trust and accuracy of a tv show recommender. In proceedings of the AH2002 Workshop on personalisation in future TV, Malaga, Spain, 2002.
- [8] Difino, A.; Negro, B.; et al.: A multi agent system for a personalized electronic program guide. In proceedings of the AH2002 Workshop on personalisation in future TV, Malaga, Spain, 2002.
- [9] Kurapati, K.; Gutta, S.; et al.: A multi agent recommender. Proceedings of 8th International Conference on user modeling UM 2001, Sonthofen, Germany, 2001. Springer.
- [10] Smyth, B.; et al.: Surfing the digital wave: Generating personalised TV listings using collaborative, case-based recommendation. Lecture Notes in Computer Science, Vol. 1650, 1999.
- [11] Uchyigit, G.; et al.: An agent based electronic program guide. In proceedings of the AH2002 Workshop on personalisation in future TV, Malaga, Spain, 2002.
- [12] Draper, S.; Earnshaw, N.; et al.: TV-Anytime. Paper presented at the opening of the IBC99 conference, 1999.
- [13] TV-Anytime forum home page. <http://www.tv-anytime.org/>, 2003.
- [14] myTV project – final report. <http://www.extra.research.philips.com/euprojects/mytv/>, 2002.
- [15] TVAnytime: Specification series: S-3 metadata, SP0003v1.3, part A. <ftp://tva:tva@ftp.bbc.co.uk/pub/Specifications/SP003v13.zip>, 2002.
- [16] Ahanger, G.; et al.: Data semantics for improving retrieval performance of digital news video systems. IEEE Transactions on knowledge and data engineering 13 (2001).

- [17] Lee, H.; Nam, J.; et al.: Personalized contents guide and browsing based on user preference. In proceedings of the AH2002 Workshop on personalisation in future TV, Malaga, Spain, 2002.
- [18] Breese, J.S.; Heckerman, D.; et al.: Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the fourteenth conference on uncertainty in artificial intelligence, Madison, USA, 1998. Morgan Kaufmann Publisher.
- [19] Manjunath, B.S.; Salembier, P.; et al.: Introduction to mpeg-7, multimedia description interface. England: John Wiley & sons Ltd., 2002.
- [20] Stark, H.; et al.: Probability and random processes with applications to signal processing. New Jersey: Prentice Hall, 2002.
- [21] Chang, C.; et al.: Libsvm: a library for support vector machines. electronic version <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>, 2003.
- [22] Korte, B.; et al.: Combinatorial optimization. Berlin: Springer-Verlag, 1991.
- [23] Internet movie database web site. <http://www.imdb.com>, 2003.
- [24] TiVo: The official TiVo web site. <http://www.tivo.com>, 2003.



Matevž Pogačnik is a researcher at the Faculty of Electrical engineering, University of Ljubljana. He received his M.Sc. degree in agent technologies in 2000 and his Ph.D. degree in user modeling in 2004. His research interests include personalized selection of multimedia content and user modeling applied to heterogeneous environments such as mobile devices, set-top boxes connected in standard and P2P networks.



Andrej Košir received the M.Sc. degree in probability theory in 1996, and the Ph.D. degree in electrical engineering from University of Ljubljana, in 1999. He is currently engaged in the research of digital image understanding, digital data coding, stochastic optimization methods and statistical data modelling. His current interests include the statistical data modelling for communication and higher level image understanding using formal systems and fuzzy logic.



Jurij Tasič is a Professor of System theory and Computing at University of Ljubljana. From 1983 he began working on communication systems and use of numerical methods in DSP, including the use of system theory in communications and communication systems. His current interests are the advanced algorithms in communication systems, multi-dimensional signal processing and parallel algorithms. He is the author of more than 60 research papers in Journals or on Conferences.

Integrated Component Web-Based Interactive Learning Systems for Engineering

Iztok Humar, *Member, IEEE*, Anton R. Sinigoj, Janez Bešter, *Member, IEEE*, and Marion O. Hagler, *Fellow, IEEE*

Abstract—Dynamic web-based learning tools are indispensable in modern teaching, especially considering their capability for interaction on demand as a means of stimulating and engaging students. The experience on which the results reported here are based has grown out of progressive testing of different approaches for publishing technical sketches and mathematical notations from the field of electromagnetics on the web, starting in 1997 with simple static web pages of solved examples. The encouraging results from using and evaluating this educational material and the needs expressed by students for learning from more dynamic and interactive web learning materials that offer the possibility of changing parameters in online calculations motivated a search for new approaches to publishing interactive learning materials on the web. Integration of already developed components for building animations, presenting mathematical equations, and performing online computations with browser scripting led to development of a learning environment where animations are synchronized with corresponding derivations of equations and supported by dynamic, parametric-driven calculations and visualizations that can be integrated with sound and video. The system supports tests and examinations in which the answers are provided as mathematical notations. The use of Mathematical Markup Language (MathML) permits equations in the learning materials to be copied and pasted into most of the popular mathematical software tools for algebraic manipulation or numerical computation. Examples demonstrate how the system can be used with a course on electromagnetics, although the basic approach is applicable in other fields of engineering and natural science.

Index Terms—Interactive technical sketches, mathematical notations, network traffic modeling, online calculations/visualizations, system access evaluation, web-based education.

I. INTRODUCTION

MODERN web-based technologies [1]–[4] can function as a bridge between difficult educational programs on one side and students who need considerable motivation on the other. These technologies are welcome additions both to the classic educational process and for autonomous asynchronous study. They can incorporate dynamic displays, interactivity, online calculations, audio, and video. All these attributes can have a stimulating effect on students [4], [5]. At the same time, thoughtfully prepared content can represent a stepping-stone

for further research by students and for their study of supplementary literature.

In engineering, a basic element of explanation is a good technical sketch. In addition, technical writings often include equations, special notations, and graphs. Today, web-based technologies efficiently support these elements. In comparison with traditional printed books, web-based technologies offer the advantage of allowing individual parts of technical sketches to appear gradually on the screen through animation, synchronized with, for example, corresponding steps in the derivation of equations. Furthermore, navigational and dynamic elements can provide interactivity. For example, mathematical tools with web support can be integrated into the web-based learning environment so that students can see the results of changing the values of parameters in calculations and graphs.

Using such web-based educational environments, students can gain better insight into the usually demanding and often difficult-to-imagine role and meaning of theory in a particular discipline. Moreover, students can access a vast spectrum of practical examples that preview even more complex structural problems with which they will have to cope during further study, engineering practice, and research. The web-based system also can support pre-laboratory learning. Offering guided step-by-step instructions through assignments enables students to learn the concept of problem solving so that later they can come to appropriate solutions by themselves.

In previous work, authors have developed and implemented proprietary classical [6]–[11] or web-based [12]–[17] applications tightly related to a specific course. Work by Fisher and Michielssen, for example, incorporates some of the engineering learning content elements in a web-based simulation environment built with the Mathematica gateway [15]. In this system, the Mathematica kernel is accessed through a web-based graphical user interface by using scripts based on the Common Gateway Interface (CGI). This system provides interactive computational and visualization capabilities, but does not address the presentation of other necessary elements, such as interactive equations, technical sketches, and animations. Also, the users of the system need some knowledge of Mathematica.

Such applications are compelling individually, but can be very hard or even impossible to employ in other courses. Often, the environments include only limited capabilities for visualization or presentation of equations. Usually, modifications and improvements require further development of the software. To reduce entanglements with specific course material, contemporary web-based learning environments can rely on contemporary standard-based or widely available commercial components, which now support essentially all

Manuscript received August 2, 2004; revised June 22, 2005. This work was supported in part by the Slovenian Ministry of Information Society, under grant no. 2811-04-000060: "Interactive multimedia web-based educational material of Electromagnetics, based on state-of-the-art Information Technologies."

I. Humar, A. R. Sinigoj, and J. Bešter are with the Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia (e-mail: iztok.humar@fe.uni-lj.si).

M. O. Hagler is with the Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS 39762 USA (e-mail: hagler@ece.msstate.edu).

Digital Object Identifier 10.1109/TE.2005.858396

necessary engineering learning content elements. Specifically, such components support integration of technical sketches, animations, mathematical equations, graphs, sound and video, dynamic displays, interactivity, and server side computation and visualization.

The basis of the results reported in this paper is experience in testing different approaches for publishing technical sketches and mathematical notations from the field of electromagnetics on the web that began in 1997, with publishing of solved exercises using static Hypertext Markup Language (HTML) and Graphics Interchange Format (GIF) images. Continuing efforts have produced a digital library that includes more than 1000 solved exercises and is growing through the addition of 100–150 additional exercises per year [18]. Since the end of 1999, when the website was restructured into a portal, the Acrobat Portable Document Format (PDF) has been used for publishing the content. Section II discusses both the HTML + GIF image approach and the Acrobat PDF approach with their advantages and disadvantages.

In 2003, an evaluation of this website was made in two different ways [5]. First, a survey was carried out among students to find out their opinions. Second, the web server log files and the network traffic were analyzed. The analysis revealed a great and constantly increasing interest for this educational material among Slovenian students. A surprisingly high percentage of students claimed that they use the system very frequently to study for an exam, and most of them indicated that they would like to use a similar system in other courses. In addition, students indicated that they would like to see the exercises solved in a more detailed and structured way in dynamic and interactive web material provided by teachers and would like to investigate phenomena themselves by changing parameters in online calculations.

These student interests motivated a search for new approaches to publishing the material and to extending the educational process by using animation synchronized with corresponding equations, sound, and other multimedia elements. Lately, many software components that offer partial support for incorporating these educational content elements in engineering study material have been developed and have become popular and frequently used. Thus, development of new approaches and new *plug-ins* to present animations, equations, visualizations, and tests on the web was neither necessary nor reasonable. Consequently, the system described here relies as much as possible on these widely utilized components.

The interactive web-based examples presented in this paper provide supplementary pre-laboratory background for the laboratory experiments of the course Fundamentals of Electromagnetics. The subjects covered by this course include electric fields, electrostatic forces, movement of particles in electric fields, breakdown, magnetic fields, magnetic forces (rotation), induction, electrical circuits, Kirchhoff's laws in circuits, Thevenin's theorem, ac circuits, and power.

After describing and comparing the approaches used for web presentation of mathematical notation, Section II demonstrates the advantages of equation reusability with MathML and the possibility of using equations as answers in web-based tests. Section III introduces and reviews the tools for preparing

TABLE I
COMPARISON OF CONCEPTS FOR PUBLISHING MATHEMATICAL NOTATIONS

	HTML+GIF	PDF	MathML
Simple content building	No	Yes	Partial ^a
Font size problems	Yes	No	Partial ^b
Printing quality	Low	High	High
Interactivity	Partial ^c	No	Yes
Combination with other media	Yes	No	Yes
Reusability	No	No	Yes
Notations as answer to questions	No	No	Yes

^a Documents can be exported directly from some word processors; no interactivity can be achieved with this approach, however.

^b MathML notations can be scaled separately.

^c Interactivity can be achieved using standard HTML elements.

technical sketches and interactive animations and describes an approach for integrating animations with the corresponding derivation of equations. In Section IV, mathematical packages providing web-based access to the server-side kernel of a mathematical package are described, and taxonomy that aids comparison is provided. In Section V, the system design concept is presented, and the integration of the various components is demonstrated by reviewing examples from the course on electromagnetics. Section VI discusses experiences in using the web-based tools as a supplement to traditional teaching methods in an introductory course on electromagnetics.

II. MATHEMATICAL NOTATIONS ON THE WEB

This section describes three basic concepts for publishing mathematical notations on the web—their advantages and drawbacks. The first two are described only briefly because they provide no interactivity and reusability and, consequently, are of little interest in dynamic, web-based environments. More extensive description of them appears in [19]. A comparison of the three approaches is summarized in Table I.

A. HTML + GIF

The earliest, simplest and most naive approach to publishing mathematical notations on the web is by using native HTML capabilities to combine text and symbol fonts with tables and black and white GIF pictures for mathematical notations and technical sketches. There are several drawbacks when using this approach in practice. 1) *Equations do not display the same on all computers and in all browsers.* The size of the notations depends on the resolution of the screen and is, therefore, bigger when the resolution is low and the other way around. Moreover, adjusting the text size in the browser settings cannot change the display size. Both of these problems lead to inequality of size in the symbols used in notations and those used in the text. The resulting appearance of the document is displeasing. 2) *Equations and pictures do not print in suitable size.* They may print either too small or too large. If they print too large, the low-resolution GIF images, usually meant for display on a screen, appear jagged. 3) *Complex documents can generate hundreds of GIF files.* 4) *The published material suffers the lack of interactivity.* The content of notations cannot be changed through user interaction, an important disadvantage in modern web-based learning material. 5) *The equations cannot be reused directly for further computations by the students.*

The main advantage of this approach is its simplicity in requiring no additional *plug-ins* or special software to view the content in a browser.

B. Acrobat PDF

Simplicity in document creation and publishing with the Acrobat PDF format has led to its widespread use among those who publish technical documentation, research papers, and educational material on the web. A PDF file, created simply by printing to a file, is a precise copy of the original document, which can be prepared in Microsoft Word, TEX, Mathematica, MathCad, etc. The author, therefore, has total control over the formatting. This approach ensures very accurate printing of the text and notations in the content published on the web. However, the PDF format requires users to install a special *plug-in* to display the documents in a browser. Unfortunately, this approach provides little interactivity beyond simple linking and provides no possibility of reusing published mathematical notations for further calculation, plotting, or simulation.

C. Mathematical Markup Language

MathML is an XML language that defines the structure of mathematical expressions so that they can be displayed, manipulated, and shared over the web [20], [21]. This language nowadays represents a standardized link between different applications that provide mathematical content and programs/browsers used to present it. A carefully encoded MathML expression can be evaluated in a computer algebra system, rendered in a web browser, edited in an equation processor, printed, or even articulated by a speech synthesizer. Most Mathematical software vendors have already implemented MathML support, thus MathML is becoming the lingua franca of scientific publication on the web. However, the presentation of MathML is not natively supported in all web browsers (Table II, [22]), although Design Science MathPlayer—a free MathML rendering component—enables Internet Explorer to display MathML.

Mathematical notations can be encoded as MathML separately and then inserted into the HTML-based educational material, or they can be composed using WYSIWYG editors, such as MathType or webEQ [23], [24], a comprehensive Java toolkit for building web pages that include dynamic math. A very popular way to produce mathematical notations in a web-based education system is by using server side tools for mathematical computations, discussed later in Section IV, that can directly produce MathML output for displaying mathematical content.

The introduction of MathML eliminates most of the problems previously described in incorporating mathematical notations in learning materials for engineering students. For this reason, and the important pedagogical advantages discussed below, MathML presentation of equations in contemporary web-based education material offers compelling advantages to content developers and was chosen for the work here.

1) *Reusability of Mathematical Notations*: Mathematical notations written in MathML can be used for further mathematical calculation, graphing, searching, and analysis. Major computer algebra systems, such as Mathematica, Maple, and MathCad, support copy-and-paste of a MathML expression from web-based learning material into their documents (Fig. 1).

TABLE II
BROWSER REQUIREMENTS TO SUPPORT MATHML

Browser	Requirements for rendering MathML
Internet Explorer 5.0 ^{a,b}	Techexplorer plug-in
Internet Explorer 5.5, 6.0 ^a	MathPlayer or Techexplorer plug-in
Netscape 6.1 ^{a,c}	Techexplorer plug-in
Netscape 7.0 ^{a,c}	MathML supported
Mozilla ^{a,b,c}	MathML supported
Amaya ^{a,c}	Presentation MathML supported

^a Windows, ^b Macintosh, ^c Linux

The possibility of copying and modifying equations and parameters can encourage students to make further calculations by themselves and thus investigate a problem more seriously.

2) *Questions Answered With Mathematical Expressions*: In the past, automated examination and self-testing of students' knowledge in the domain of mathematical notations was hard to perform. Students usually were offered a choice of several answers prepared by the authors in advance. Using MathML as a protocol for communication between a web client and an application for evaluating the input, students can answer questions by writing mathematical expressions. Integration of graphical equation editor applets, such as the webEQ Input Control [23], into an educational web page with questions can enable the examinee simply to write the answer to a given problem in a WYSIWYG manner (see Fig. 2). In the background, the answers are compiled in MathML format; after clicking the "submit" button, they are sent for evaluation in a way similar to that in which values are collected by other HTML form elements. The processing can be done either with JavaScript code embedded in the same web page [25] or by an application at the web server side.

Students are not required to enter their results in simplified format. For example, they can respond with $(r^2 - 4)$ or with $(r - 2)(r + 2)$ —both are treated as the same result. To verify and evaluate correctly a given expression on the client side with JavaScript, an Evaluation Control applet [23] can be added to the document. This control utilizes sophisticated mathematical algorithms to compare MathML expressions and determines when two expressions are equivalent. The other possibility for verification of answers written in MathML is to use server-side applications to process the MathML, simplify the expression, and perform the assessment.

III. ANIMATIONS

The importance of technical sketches as an element in educational material for engineers was discussed in the Introduction. In contrast to the black and white static technical sketches usually used in books and journals, modern web-based technologies offer the possibility of using colorful and dynamic two- and three-dimensional (2-D and 3-D) sketches and animations. The main advantage of the latter from a pedagogical perspective is their ability for providing user interaction. They can be triggered by time or by specific events, such as a mouse click on the "Next" button. This capability enables gradual presentation of information to learners so that new items can appear individually, step-by-step. Each step can be synchronized with corresponding equations, which are usually derived from previous

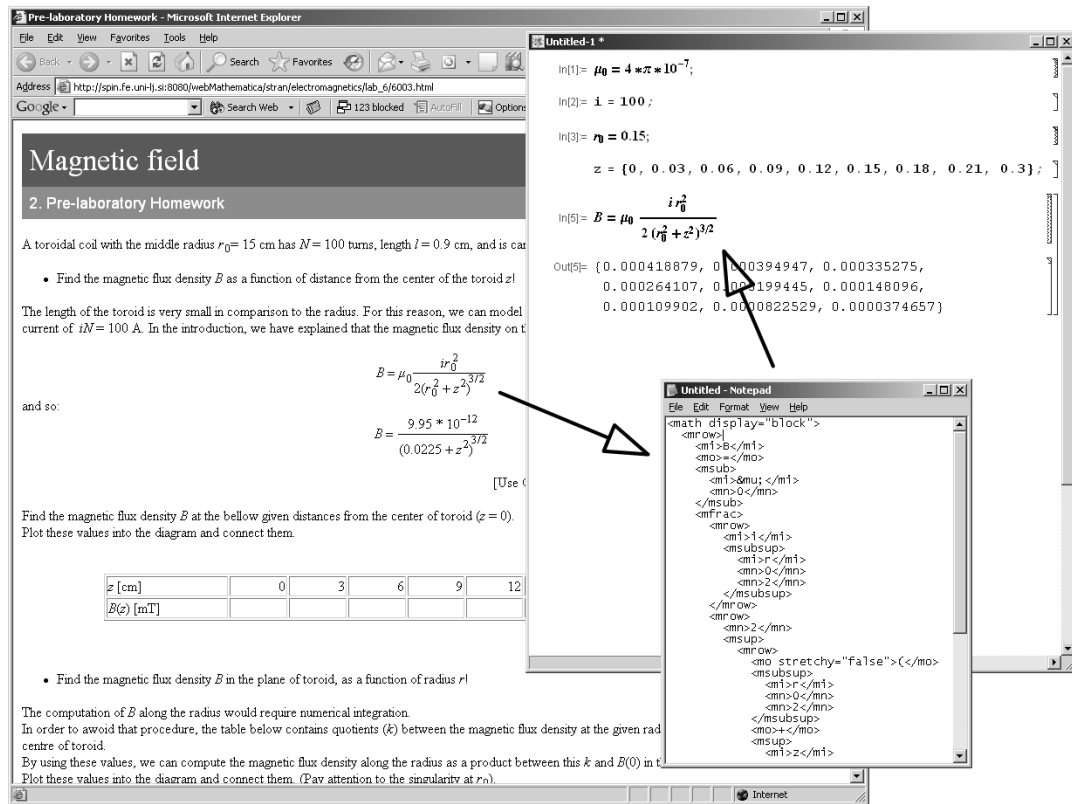


Fig. 1. Copy-and-paste mathML from a web browser to the mathematical program for further calculations.

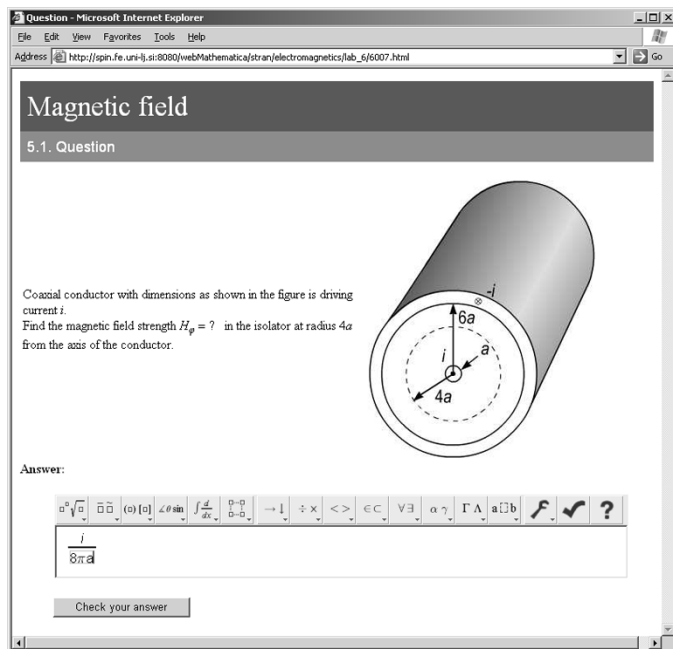


Fig. 2. Entering and evaluating a mathematical expression as the answer to a given problem.

equations. Explanatory comments can be added with additional text or voice. Beyond explaining derivations, animations with integrated speech or video clips also can be an effective means of explaining some concepts and phenomena.

A complete and widely used package for animation is Macromedia Flash, which provides powerful features for animation,

multimedia, and application development, features that allow designers and developers to create rich e-learning courses [26], [27]. An example of its use for preparing 2-D animations and visualization of phenomena is shown in Fig. 3.

Developers of e-learning content interested in describing functions in space from the fields of engineering and natural science might wish to prepare 3-D animations carried out by using MindAvenue Axel [28]. Axel Educator's Program offers excellent capabilities for modeling, animating, texturing, lighting, previewing, integrating, and publishing interactive educational contents to the web without the need for programming. Although this tool makes possible composition of almost any content imaginable, preparation requires considerable time and effort.

A. JavaScript as a Link Between Interactive Elements

The possibility of moving through animations step-by-step with navigational elements integrated into a web page was noted earlier. Alternatively, the animation steps can serve as triggers to move through successive corresponding steps in the derivation of an equation, to perform online calculation with particular values, to draw graphs with selected parameters, to start sound clips with explanations, or to run video clips.

In practice, the integration of these technologies is accomplished by client-side scripting, e.g., JavaScript. For example, to achieve the step-by-step derivation of equations, the MathML code of all levels of derived equations is prepared in advance and stored into several JavaScript variables. When an appropriate event takes place (mouse click on a "Next" button, animation triggered event, etc.), the content of equations is replaced by the

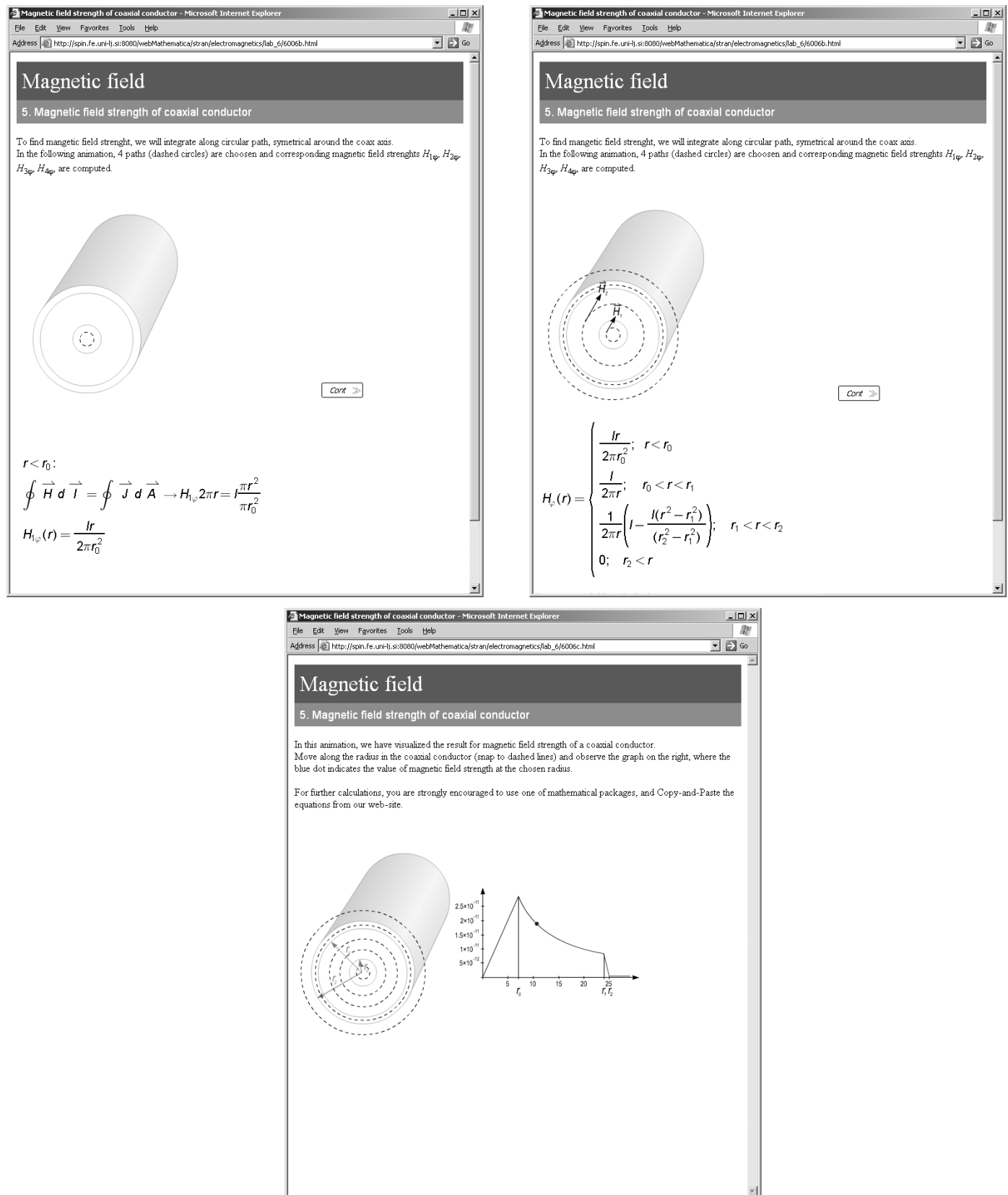


Fig. 3. Three steps of solving a problem in electromagnetics, presented using flash animation dynamically synchronized with corresponding equations written in MathML.

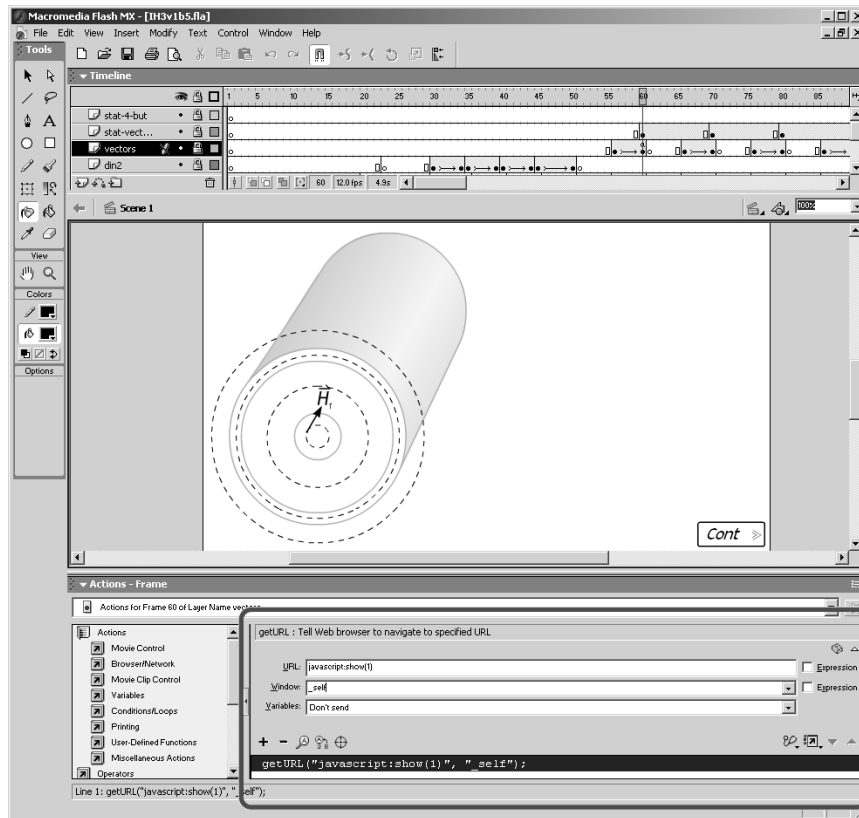


Fig. 4. The usage of ActionScript for synchronization between animations and mathematical equations.

next-level JavaScript variable, resulting in a new equation presented to the user [see Fig. 3].

To control the web-based content by sensible elements within animations, suitable scripting must be integrated during their compilation. Macromedia Flash and MindAvenue Axel both support the usage of JavaScript. For example, Fig. 4 shows how ActionScripting is used for triggering browser scripts in Flash. These scripts dynamically control MathML content that is synchronously previewed in MathPlayer. With this approach, MathPlayer can be employed for dynamic presentations of notations in coordination with animations, an approach that represents an excellent opportunity for structured presentation of engineering problem solutions.

IV. ONLINE CALCULATIONS AND VISUALIZATIONS

The benefits of educational material that contains animations synchronized with corresponding derivation of equations and other multimedia elements were noted in earlier work [29]. This section focuses on an approach that provides on-line visualization of relations governing electromagnetic fields in space and in time rather than derivation of equations. According to the analysis mentioned earlier [5], students not only want to learn from examples prepared in advance, but also want to be able to change the parameters in calculations, graphs, and animations. This ability allows them to find answers to their questions and to understand certain terms better.

Portability across operating systems, strong support for visualization, analytical and numerical computations, and presentation of equations with MathML can result from employing one

of several powerful mathematical packages that have become widely available. Today, most of the companies that make mathematical packages for analytical and numerical arithmetic offer supplements that provide access to the mathematical kernel of the program through a web interface. These supplements allow building of examples that help students to carry out efficiently their own limited analyzes and more complex analyzes prepared in advance by tutors. Students do not need any special knowledge about the software tools for mathematical analysis or even knowledge about which mathematical tools are being used in the background of the educational system.

During the process of choosing a platform, the documentation for several mathematical systems that offer the capability of computation and visualization through a web interface was reviewed as follows:

- Wolfram Research webMathematica [30];
- The MathWorks Matlab webServer [31];
- Maplesoft mapleNet [32];
- Mathsoft Mathcad Application Server [33].

Wolfram Research webMathematica and The MathWorks Matlab webServer were installed and tested. In related work [35], an extensive review and comparison is conducted of all the systems listed. To summarize results from the review, taxonomy is given in Table III. In regard of its capabilities, ease of installation, effectiveness of integration with other components of the web-based education system, and experiences from the past, webMathematica was chosen as the mathematical package for driving the online calculations of the web-based educational system described here.

TABLE III
TAXONOMY OF MATHEMATICAL SYSTEMS USED FOR ONLINE CALCULATIONS AND VISUALIZATIONS

	webMathematica	Matlab Webserver	mapleNet	Mathcad Appl. Server
Server platform	Windows, Unix/Linux	Windows, Unix/Linux	Windows, Linux, Macintosh	Windows
Client Requirements	JavaScript Browser, MathML Player	JavaScript Browser ^a	JavaScript Browser, MathML Player	JavaScript Browser, MathML Player
Symbolic Calculations	Yes	Yes ^b	Yes	Yes
Numerical Calculations	Yes	Yes	Yes	Yes
MathML Support	Yes	Partial ^c	Yes	Yes
Visualization	Yes	Yes	Yes	Yes
3D Preview	LiveGraphics3D	JPG images	mapleNet applet	PNG images

^a no MathML support by default

^b by using Symbolic Math Toolbox

^c by using MathML Visualization tool [34]

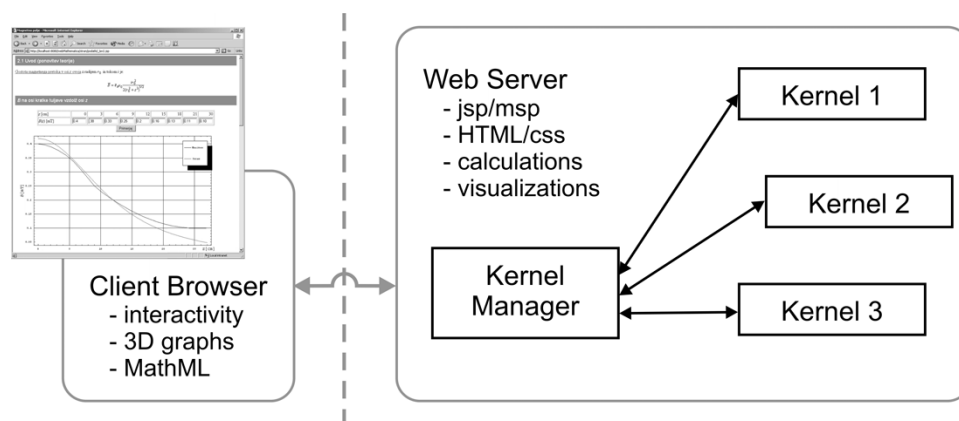


Fig. 5. Architecture of webMathematica deployment.

A. webMathematica

Wolfram Research webMathematica adds interactive calculations and visualizations to a web-based educational site by integrating Mathematica with web server technology, such as the Apache web Server. This extension makes possible the construction, test, and deployment of specialized web services for computation and visualization, all powered by a large library of Mathematica commands on the server side.

The system is based on two standard Java technologies: Java Servlets and JavaServer Pages (JSP). Page code, similar to HTML, is enhanced by the addition of Mathematica commands. When a request is made for one of these pages the Mathematica commands are evaluated, and the computed result is placed in the page. This process is performed with the standard Java templating mechanism, JavaServer Pages, making use of special tags and the request/response standard followed by web servers. The input is usually entered through HTML forms, applets, JavaScript calls, or interactive animations (as shown in previous section) although data files or Mathematica Notebooks also can be sent to a webMathematica server for processing. Output can be presented in many different formats—the most interesting for web-based education are HTML, MathML, images (2-D/3-D graphs), XML, and text, although the output can also be a Mathematica notebook, SVG, PostScript, or PDF.

An important part of this system is the kernel manager, which calls Mathematica in a robust, efficient, and secure manner. The manager maintains a pool of one or more Mathematica kernels,

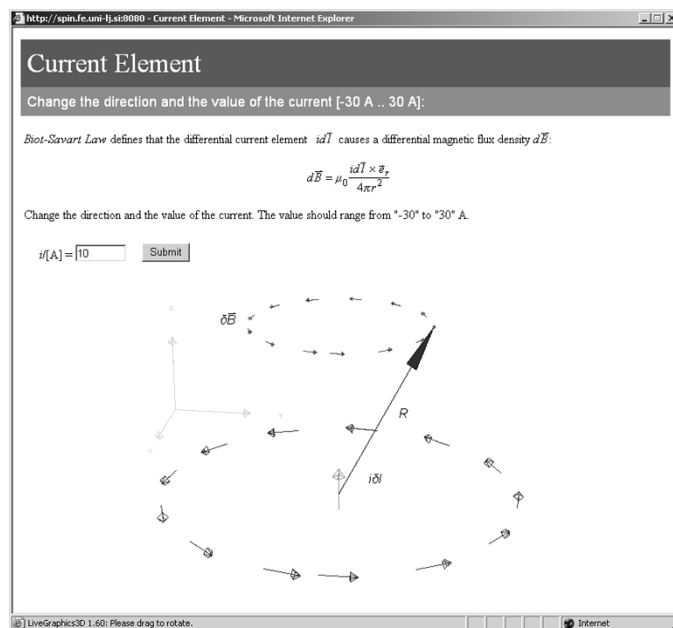


Fig. 6. An example of visualization with webMathematica.

which can process several requests at a time. An overview of deploying webMathematica is shown in Fig. 5.

The aim of webMathematica and JSP technology is to reduce the amount of specialized knowledge required of those who use the site and to minimize the time needed for developing the

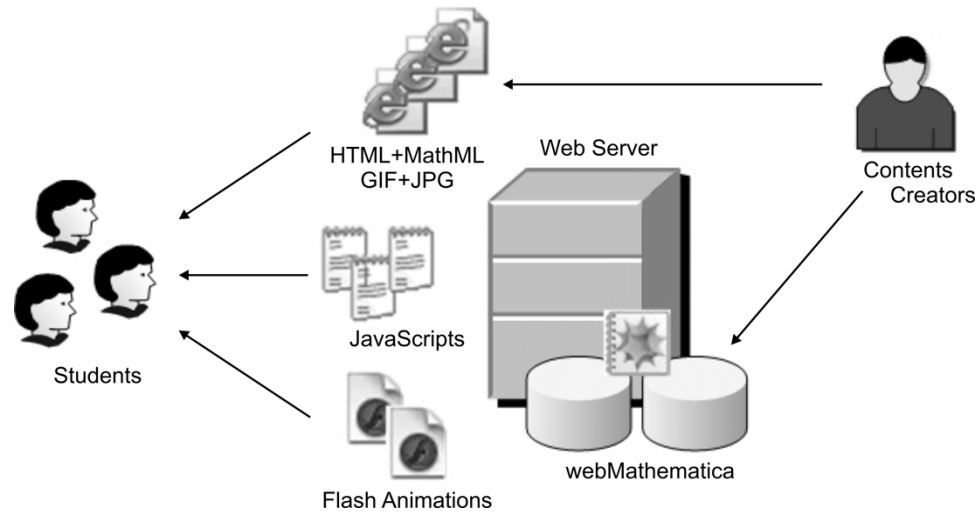


Fig. 7. The system architecture.

```

<html xmlns:m="http://www.w3.org/1998/Math/MathML">
<head><meta http-equiv="Content-Type" content="text/html; charset=windows-1250">
<object id="MathPlayer" classid="clsid:32F66A20-7614-11D4-BD11-00104BD3F987" ><?import namespace="m"
implementation="#MathPlayer" ?>
<script type="text/javascript">
function show(nspointer) {
if (nspointer == 0) {
document.wbqNS.setEquation("<m:math><m:semantics><m:mrow><m:mstyle displaystyle='true'><m:mrow><m:mo>...</m:mrow></m:math>");
if (nspointer == 1) {
document.wbqNS.setEquation("<m:math><m:semantics><m:mrow><m:mstyle displaystyle='true'><m:mrow><m:mo>...</m:mrow></m:math>");
}
...
}
</script>

<title>Magnetic field</title><link href="Classroom.css" rel="stylesheet" type="text/css" />
</head>
<body>

<object classid="clsid:D27CDB6E-AE6D-11cf-96B8-444553540000" codebase="http://download.macromedia.com/pub/...
<param name="movie" value="6006b.swf"> <param name="quality" value="high"> <param name="bgcolor" value="#FFFFFF>
<embed src="6006b.swf" quality="high" bgcolor="#FFFFFF" width=700 height=400 type="application/x-shockwave-flash"...

<br>

<applet code="webq3.ViewerControl" codeBase="../../classes" height=200 name=wbqNS width=700 MAYSCRIPT>
<applet name="color" value="#FFFFFF"><param name="valign" value="top"><param name="align" value="left">
<param name="selection" value="true"><param name="controls" value="true"><param name="size" value="20">
<param name="eq" value="<m:math><m:semantics><m:mrow><m:mstyle displaystyle='true'><m:mrow><m:mo>...</m:mrow></m:math>";
</applet>

</body>
</html>

```

a) Integrating JavaScript

b) Flash Animation

c) Dynamic Notations in MathML

Fig. 8. HTML code with JavaScript for integration of flash animation and webEQ viewer to achieve the step-by-step derivation of equations depicted in Fig. 3.

educational content. In practice, authors must know something about HTML and Mathematica although they need no extensive knowledge of Java. Basic knowledge of JavaScript is needed for performing integration with other educational elements. For example, integration of Flash and webMathematica is possible by using JavaScript. In this way, graphs can be presented as parts of animations, and interactive communication with animations can be redirected to server-side mathematical tools as parameters for computation and visualization.

Incorporation of the webMathematica LiveGraphics3D applet in the web-based educational material permits easy integration of webMathematica results into a web page. This applet displays Mathematica 3-D graphics and provides features, such as interactive rotation and resizing of elements. The applet works in a wide range of different Java-enabled browsers.

An example of visualization with webMathematica is given in Fig. 6, which shows the magnetic field of a current element with different strength/direction of the current as a parameter.

Through more detailed consideration of examples introduced earlier in this paper, the next section illustrates the integration of Mathematica into a learning environment with Flash animation, MathML (both for display and for user input of mathematical expressions), JavaScript, and visualizations (static and dynamic).

V. DESIGN CONCEPT AND IMPLEMENTATION

A. The Design Concept of the System

The central element of the system is the web server, used for delivering the web-based content. The web server natively delivers the contents as HTML, GIF, JPG, and Flash Animations. The JavaScripts are parts of the HTML documents. To perform online computation, webMathematica must be installed on the server. The architecture of the system is shown on Fig. 7. Providing other types of content requires installation of different features, i.e., installation of Windows Media Services for video streaming, etc.

```

<%@ page language="java" %>
<%% taglib uri="/webMathematica-taglib" prefix="msp" %>
<html xmlns:m="http://www.w3.org/1998/Math/MathML">
<head><meta http-equiv="Content-Type" content="text/html; charset=windows-1250">
<object id="MathPlayer" classid="clsid:32F66A20-7614-11D4-BD11-00104BD3F987" ></object> <?import namespace="m"
implementation="#MathPlayer" ?> <title>Tokovni element</title><link href="Classroom.css" rel="stylesheet"
type="text/css" />
</head>
<body>

<p class="Title">Current element</p>
<p><BR><I>Biot-Savart Law</I> defines the magnetic field of ...</p> <m:math display='block'>
<m:mrow>
<m:mi>d</m:mi><m:mi><m:mover accent='true'>
<m:mi>B</m:mi>
...
</m:math>

<APPLET ARCHIVE="live.jar" CODE="Live.class" WIDTH=800 HEIGHT=800 ALIGN=LEFT>
<PARAM NAME=BGCOLOR VALUE=#FFFFFF><PARAM NAME=MAGNIFICATION VALUE=1.1>
<PARAM NAME=INPUT_FILE VALUE="currentelement_m.m">
</APPLET>

<msp:allocateKernel><form action="currentelement2.jsp" method="post">
<p>i=</p>
<input type="text" name="i" size="5" value = "<msp:evaluate> MSPValue[ $$i, "10"] </msp:evaluate>">
...
<input type="submit" name="btnSubmit" value="Visualize"><br> <msp:evaluate>
Get["Utilities\tokovnielement"]; MSPBlock[ {$$i}, objekt[ $$i] ]</msp:evaluate>
</form></msp:allocateKernel>
</body>
</html>

```

a) Text + MathML

b) Static visualization

c) Interactive visualization

Fig. 9. JSP code integrating static and interactive visualizations, produced by mathematica on the server-side.

The implementation described here uses the Windows 2000 Server. The system is integrated into E-CHO (e-learning system at the Faculty of Electrical Engineering, University of Ljubljana), a system that permits the authentication of users, the collection of statistics, network traffic measurement, support for video streaming, and use of e-learning study tools, such as forums, email correspondence with tutors, videoconferencing, etc.

B. Analysis of the Examples

This section shows source code for the examples shown in figures of previous sections and discusses its functionality. These examples can be found at [36].

Fig. 8 shows the structure of an HTML document that integrates a Flash object [see Fig. 8(b)] for animation and a webEQ Viewer Control applet [see Fig. 8(c)] for presentation of the dynamic derivation of equations. The integration is performed by using JavaScript [see Fig. 8(a)]. The successive steps of equations are stored as separate variables in MathML format and appear gradually in the viewer applet when triggered by an ActionsScript from the Flash animation (see Fig. 4).

Fig. 9 reveals the Java Server Pages code of web material that includes visualizations, such as the one in Fig. 6. The static visualizations can be generated in advance by Mathematica and only displayed in the web course [see Fig. 9(b)] or can be calculated on-line at the server side by using the parameters entered by the user with a form [see Fig. 9(c)].

Fig. 10 shows the code of a questionnaire that allows entering a mathematical expression as an answer to a given problem, such as the one shown in Fig. 2. The questionnaire document is composed of a static image of a problem [see Fig. 10(b)], a question composed of text that includes MathML [see Fig. 10(c)], and a call of the webEQ Input Control applet [see Fig. 10(d)]. In this example, some of the possible answers are predicted and stored in JavaScript variables [see Fig. 10(a)], and JavaScript also evaluates the answer. However, authors ordinarily should take advantage of the more versatile and robust evaluation features of the Evaluation Control (see Section II-C2) or the server side mathematical tools (see Section IV) for verification of answers.

VI. USAGE EVALUATION

System usage statistics and student opinions have been collected during the seven years since 1997, when publishing web-based educational material for the field of electromagnetics commenced. At the start of the year 2003, however, a more detailed study began that used two different approaches: analysis of the web server log files and network traffic load characteristics, and a simultaneous survey among students.

A. Website Access Evaluation

The website access evaluation was conducted by analyzing the web server log files. The logs indicate which material the students review, how frequently students use the system, how the usage is distributed over the day, week, month or year, etc. The log analyzes give an opportunity to build models of system usage, which show exponentially distributed lesson intervals during which the data is being sent in bursts of nearly uniform, distributed length. The reader is referred to [38] for more detailed information about user access modeling.

The website evaluation was completed with webTrends Log Analyzer, which generates several predefined reports. The reports show the number of visits in certain periods, amount of transferred bytes, most/least frequently used pages, etc.

Access analysis for the year 2002 revealed 187 220 accesses during this period. The number of accesses in December 2002 increased 25% over December 2001, as shown in Fig. 11. The peak day of access was 10 June (2351). The monthly distribution of requests shows that interest increases during the exam periods in January, April, June, September, and December, while the lowest interest in the study material occurs during the summer holidays (July, beginning of August).

In the last three years, the number of users has remained steady, probably because almost all the first year students, for whom the material is intended, are using the material. Recently, the system has also been introduced to teachers of more advanced courses. They are expected to adopt this approach in the near future. Consequently, the number of users is going to increase.

```
<html xmlns:m="http://www.w3.org/1998/Math/MathML">
<head><meta http-equiv="Content-Type" content="text/html; charset=windows-1250">
<object id="MathPlayer" classid="clsid:32F66A20-7614-11D4-BD11-00104BD3F987" ></object> <?import namespace="m"
implementation="#MathPlayer" ?>
<script type=javascript>
function check() {
    var answer_MathML = document.myApplet.getPackedMathML();
    var answer1 = "<math><mrow><mfrac><mrow><mi>i</mi></mrow><mrow><mn>8</mn></mrow></mfrac></mrow></math>";
    var answer2
    "<math><mrow><mfrac><mrow><mi>i</mi></mrow><mrow><mi><math><mi>a</mi></mrow></mfrac></mrow></math>";
    var answer3 = "<math><mrow><mfrac><mrow><mi>i</mi></mrow><mrow><mi><math><mi>a</mi></mrow></mfrac></mrow></math>";
    if (answer_MathML == answer1 || answer_MathML == answer2 || answer_MathML == answer3) {alert("Correct!");}
    else {alert("Wrong!");}}
</script>

<title>Question</title> <link href="Classroom.css" rel="stylesheet" type="text/css" />
</head>

<body>
    <p class="Title">Magnetic field of coaxial conductor</p>

    <p></p>

    <p>Coaxial conductor with dimensions as shown in the figure is driving current <i>I</i>. Find the magnetic ...
    <m:math><m:semantics><m:mrow><m:sub><m:mi>H</mi><m:mi><math><mi>a</mi></mrow></m:sub><m:mo>=</m:mo><m:mo>?</m:mo></m:math>
    </m:math>
    <br>in the isolator at radius 4<i>a</i> from the axis of the conductor.</p>
    <p><b>Answer:</b></p>

    <blockquote>
        <applet codeBase=http://www.mathtype.com/en/reference/webmath/classes/
        height=100 width=700 code=webq3.editor.InputControl.class name=myApplet></applet>
        <form><input onclick=check() type=button value="Check your answer"></form>
    </blockquote>
</body>
</html>
```

Fig. 10. Code of a questionnaire based on webEQ input control. The result is evaluated by JavaScript.

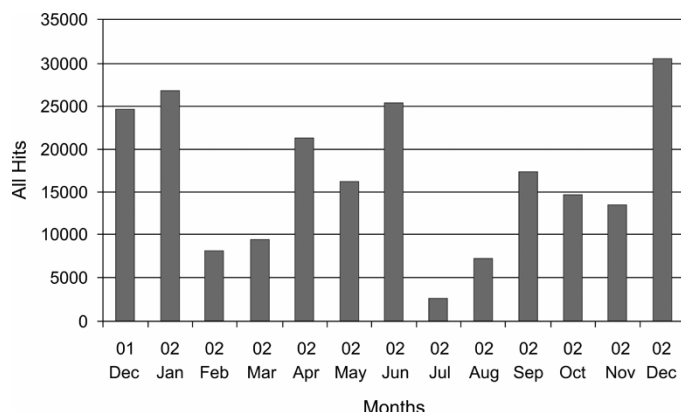


Fig. 11. The distribution of requests during the year.

B. Network Traffic Measurement

To understand the system load, ensure Quality of Service, and predict required network resources for the future, the network traffic is continuously being measured and analyzed [37] at the e-learning system that provides the described web-based study material.

Over time, the network traffic characteristics of the e-learning environment are changing for the following reasons: 1) as described in Section VI-A, both the number of users and their interest are continuously growing; 2) the majority of users study from their homes, where the low-bandwidth modem internet connections have recently been replaced with high-speed broadband technologies, such as cable or ADSL; 3) the study material has been enhanced with multimedia elements, particularly video and audio; 4) the modern study material offers a high level of interactivity: the students can change the values of parameters in online calculations and visualizations or answer the questions with mathematical equations.

Based on the measurements, the traffic models for different e-learning components were built and are provided in more details in [38]. When study material was composed of HTML and GIF files, the communication traffic was low bit-rate, based

on asymmetric bidirectional Transmission Control Protocol (TCP) packets with long periods between consecutive packet exchanges, caused by user's access to next portion of material. The introduction of PDF files and growth of user's connection bandwidth resulted in more "bursty" elastic network traffic. The enhanced multimedia elements, such as streaming audio and video, introduces unidirectional, more aggressive User Datagram Protocol (UDP) traffic with high constant bit-rate requirements that can significantly influence network traffic resources and system load, especially in the case of several simultaneous users. The consecutive requests of users working with modern interactive study material also cause an increased amount of TCP traffic carrying large amount of data in downstream, a result of online computations that produce large portions of data used for visualization and animations, as described in Section IV.

C. Survey Among Students

Students who participated in the survey [39] were asked to describe the process of working with the web-based material, its usefulness, the purpose of usage, and the way in which they used it. The survey was carried out among 280 first year undergraduate students. All students except one knew about the web-based material. The majority (75%) of them were told about the material by their teacher, and 20% by their colleagues. Surprisingly, 5% of the students claimed to have found it via searchers, such as Google or Yahoo.

Fifteen percent of the students affirmed that they studied all available web-based material. The majority used it to consolidate knowledge before exams. Fifteen percent claimed that the web-based material was their primary resource in preparing for the exam. Twenty percent of the students read directly from the screen, while 80% printed the material. Ten percent of students also printed the materials for their colleagues. Seventy percent used the system at home, while others used it on campus (23%) or somewhere else. The great majority of students found the system very useful and would like to have access to a similar collection of materials for other courses.

Most of the comments by students were very positive and encouraging. Nevertheless, they requested the capability for solving problems in a more detailed and structured way with dynamic and interactive web material, including the possibility of changing parameters in the online calculations.

D. Assessment of the Interactive Web-Based Methodology

The interactive web-based methodology has been employed in the course, Fundamentals of Electromagnetics, to provide prelaboratory background for the laboratory experiments [40] and also to support some online postlaboratory calculations. The material has been used in the winter semester of 2004/05 as a supplement to the traditional theoretical explanations in the classroom. The content was delivered through the e-learning system (E-CHO), which enables the collection of statistics on content usage for individual users. Student opinions also were collected.

A significant parameter is the number of students using the web-based materials in comparison to the attendance in the classroom (which is obligatory for students). From a group of 47 students, 38 (81%) logged on the system. Three students (6%) studied only the first few pages and stopped using the system (two of them dropped the laboratory course). Nearly half of the students (47%) visited all the prepared materials. Four students skipped the instructions for one of the laboratory experiments, while the rest studied approximately 90% of the total material. The average time students spent on the system studying the background of one laboratory experiment was 49 min, with individual times ranging from 21–150 min.

An index of student success was the number of students who submitted correctly solved laboratory course homework on time. Statistics from previous years indicate that approximately 85% of students who intended to attend the laboratory course submitted the laboratory on time. Around 60% of these students completed the homework correctly. During the winter semester of 2004/2005 when the system was available as a supplement to the traditional theoretical explanations in the classroom, in contrast, 96% of students submitted their work on time, and 89% of students had no mistakes in the homework they submitted.

The students also were asked to express their opinion about this new method of learning. The initial response was favorable. Most of them share the opinion that the web-based material offered them an opportunity to learn the hard-to-envision phenomena more quickly and effectively than if they received only handwritten materials. The motivation of the students also was increased because of their enthusiasm for web-based learning. They liked the opportunity to study at home at any time and to use the system interactively to complete and verify the results required for homework. Five of nine students who had never used the system claimed that they had no Internet access at home. They did not want to spend additional time at the University; some of them used printed versions of the results obtained from their colleagues, however.

The faculty staff did not notice a drop in attendance at the traditional class meetings. Some students studied the material in advance and brought printed versions of the material to the

classes to seek answers to their questions. Nevertheless, the increased percentage of correct homework reduced the time required to review the submitted work. The faculty staff also observed a noticeable improvement, in comparison to previous years, in the students' knowledge about performing laboratory experiments.

VII. CONCLUSION

This paper describes how to integrate state-of-the-art web component technologies through browser scripting and server-side applications into a modern web-based learning system that supports the presentation of key elements of engineering education: dynamic and interactive animations synchronized with the derivation of equations, parametric driven online computations and visualizations, and on-line tests that can be answered with equations. Such learning environments can be used for building self-study courses or supplements to traditional teaching methods and are excellent tools for stimulating and engaging students in further research.

The examples demonstrate the use of web-based educational material built with widely available technology components in a first year undergraduate course entitled Fundamentals of Electromagnetics. Because the effort required for developing the materials through integrating mainly existing components is reasonable and because students have expressed interest and requested additional features and material, development of the interactive and dynamic web-based learning environments described here will continue.

The application of the basic approach is not limited to the field of electromagnetics only and is useful in other fields of engineering and natural science.

ACKNOWLEDGMENT

I. Humar would like to thank Prof. P. Kokelj, Faculty of Electrical Engineering, University of Ljubljana, for his ideas and encouragement and T. Strahija, A. Kovač, and students for contributions in developing the content.

REFERENCES

- [1] P. Penfield Jr and R. C. Larson, "Education via advanced technologies," *IEEE Trans. Educ.*, vol. 39, pp. 436–443, Aug. 1996.
- [2] D. Harris and A. DiPaolo, "Advancing asynchronous distance education using high-speed networks," *IEEE Trans. Educ.*, vol. 39, pp. 444–449, Aug. 1996.
- [3] M. O. Hagler and W. M. Marcy, "Computer-Aided instruction," in *Encyclopedia of Electrical and Electronics Engineering*, J. G. Webster, Ed. New York: Wiley, 1999, vol. 3, pp. 648–668.
- [4] M. F. Iskander, "Technology-Based electromagnetic education," *IEEE Trans. Microw. Theory Techn.*, vol. 50, no. 3, pp. 1015–1020, Mar. 2002.
- [5] I. Humar, E. Bulić, and A. R. Sinigoj, "Solved problems of fundamentals of electromagnetics on the web: System usage evaluation," *Electrotech. Rev.*, vol. 71, no. 1–2, pp. 40–44, 2004.
- [6] A. Žemva, A. Trost, and B. Zajc, "A rapid prototyping environment for teaching digital logic design," *IEEE Trans. Educ., Rapid Publication Suppl.*, vol. 41, no. 4, pp. 1–8, Nov. 1998.
- [7] F. Buret, D. Muller, and L. Nicolas, "Computer-Aided education for magnetostatic," *IEEE Trans. Educ.*, vol. 42, pp. 45–49, Feb. 1999.
- [8] D. Consonni and A. C. Seabra, "A modern approach to teaching basic experimental electricity and electronics," *IEEE Trans. Educ.*, vol. 44, pp. 5–15, Feb. 2001.

- [9] M. Castro, A. Hilario, S. Acha, J. Pérez, A. Colmenar, P. Losada, I. Rivilla, and J. Peire, "Multimedia design and development for distance teaching of electronics," in *Proc. IEEE Frontiers in Education Conf.*, Reno, NV, Oct. 2001.
 - [10] M. A. Mora, R. Moriyón, and F. Saiz, "Building mathematics learning applications by means of CONSMATH," in *Proc. IEEE Frontiers in Education Conf.*, Boulder, CO, Nov. 2003.
 - [11] D. Millard and G. Burnham, "Innovative interactive media for electrical engineering education," in *Proc. IEEE Frontiers Educ. Conf.*, Reno, NV, Oct. 2001.
 - [12] Y. Katsuyama and O. Kouama, "Experiment & calculation supporting system over WWW for problem-based learning in electrical engineering," in *Proc. Information Technology Based Higher Education and Training*, Kumamoto, Japan, Jul. 2001.
 - [13] V. Uskov and C. Etaugh, "Developing innovative web-lecturing technology and online courses at bradley university," in *Proc. Int. Conf. on Computers in Education*, Auckland, New Zealand, Dec. 2002.
 - [14] J. Švajger and V. Valenčič, "Discovering electricity by computer based experiments," *IEEE Trans. Educ.*, vol. 46, pp. 502–507, Dec. 2003.
 - [15] S. E. Fisher and E. Michielssen, "Mathematica[®] assisted web-based antenna education," *IEEE Trans. Educ., Rapid Publication Suppl.*, vol. 41, no. 4, pp. 1–15, Nov. 1998.
 - [16] J. Goerke, R. Beemagani, M. Hoffmann, and W. Strasser, "Interactive courseware for teaching field theory," in *Proc. Int. Conf. on Engineering Education*, Valencia, Spain, Jul. 2003.
 - [17] M. de Magistris, "A MATLAB-based virtual laboratory for teaching introductory quasistationary electromagnetics," *IEEE Trans. Educ.*, vol. 48, pp. 81–88, Feb. 2005.
 - [18] —, (1997–2005) Collection of Solved Problems of Fundamentals of Electromagnetics. [Online]. Available: http://torina.fe.uni-lj.si/oe/index_e.html
 - [19] I. Humar, A. R. Sinigoj, and E. Bulič, "Dynamic educational material of electromagnetics with integrated mathematical notation for the web," in *Proc. 4th Int. Conf. on Information Technology Based Higher Education and Training*, Marrakech, Morocco, Jul. 2003.
 - [20] M. O. Hagler, "Mathematics and equations on the WWW," in *Proc. Frontiers in Education*, Tempe, AZ, Nov. 1998.
 - [21] —, "Application of recent developments in mathematics on the WWW with MathML," in *Proc. 4th Int. Conf. Information Technology Based Higher Education and Training*, Marrakech, Morocco, Jul. 2003.
 - [22] —, (2005) Putting Mathematics on the web With MathML. [Online]. Available: <http://www.w3.org/Math/XSL/>
 - [23] —, (2005) Design Science Products: MathType, WebEQ, WebEQ Input Control, WebEQ Evaluation Control. [Online]. Available: <http://www.mathtype.com>
 - [24] R. Miner and P. Topping, (2002) Math on the Web: A Status Report (2001, 2002), Design Science. [Online]. Available: <http://www.dessci.com/webmath/status>
 - [25] M. Hagler, "Pedagogic coaching with web browsers and JavaScript," in *Proc. Information Technology Based Higher Education and Training*, Istanbul, Turkey, May 2004.
 - [26] —, (2005) Macromedia Flash. [Online]. Available: <http://www.macromedia.com/software/flash/>
 - [27] E. Ferré, W. S. Lau, B. Ngo, E. A. Riskin, M. Soma, R. Christie, J. Harris, L. J. Collins, R. E. Lee, and M. Campion, "Flash animation in introductory EE courses," in *Proc. IEEE Frontiers in Education Conf.*, Boston, MA, Nov. 2002.
 - [28] —, (2005) Mindavenue Axel. [Online]. Available: <http://www.mindavenue.com/en/products/>
 - [29] I. Humar, M. Pustišek, and J. Bešter, "Developing dynamic educational material with integrated mathematical notation for web-based learning system," in *Proc. IEEE Frontiers in Education Conf.*, Boulder, CO, Nov. 2003.
 - [30] —, (2005) Wolfram Research webmathematica. [Online]. Available: <http://www.wolfram.com/products/webmathematica/>
 - [31] —, (2005) The MathWorks Matlab webServer. [Online]. Available: <http://www.mathworks.com/products/webserver/>
 - [32] —, (2005) Maplesoft MapleNet. [Online]. Available: <http://www.maplesoft.com/maplenet/>
 - [33] —, (2005) Mathsoft Mathcad Application Server. [Online]. Available: http://www.mathcad.com/products/Mathcad_Application_Server.asp
 - [34] M. Kempf and M. Reinhold, (2003) MathML Visualization. [Online]. Available: <http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=3049#>
 - [35] I. Humar, A. R. Sinigoj, and M. O. Hagler, "Mathematical tools for supporting web-based education," in *Proc. Information Technology Based Higher Education and Training*, Istanbul, Turkey, May 2004.
 - [36] I. Humar and A. R. Sinigoj, (2005) Some Examples of Interactive web-Based Material From the Course Fundamentals of Electromagnetics. [Online]. Available: <http://spin.fe.uni-lj.si/electromagnetics/>
 - [37] I. Humar, M. Golja, and J. Bešter, "IP network traffic measurement," *Electrotech. Rev.*, vol. 72, no. 2, pp. 1–7, 2005.
 - [38] I. Humar, M. Papić, and J. Bešter, "Network traffic models of multimedia enhanced interactive e-learning environments," in *Final Report Faculty of Electrical Engineering*: University of Ljubljana, 2005.
 - [39] I. Humar, (2003) Questionnaire. [Online]. Available: <http://torina.fe.uni-lj.si/oe/usage/questionnaire.pdf>
 - [40] P. Kokelj, "Experimenting with the high voltage testing board," *Electrotech. Rev.*, vol. 68, no. 1, pp. 40–44, 2001.
- Iztok Humar** (S'00–M'04) received the B.Sc. and M.Sc. degrees in electrical engineering from University of Ljubljana, Faculty of Electrical Engineering, Slovenia, in 2000 and 2003, respectively. Currently, he is working towards the Ph.D. degree.
- He is currently a Teaching Assistant at the University of Ljubljana, Faculty of Electrical Engineering, Slovenia, where he lectures on "The Fundamentals of Electromagnetics." His main research topics are from the field of Communications; his work also includes the research and development of modern communications and information technologies for the process of web-based education.
- Mr. Humar is a Member of the Electrical Association of Slovenia. He serves as IEEE Communication Society of Slovenia Chapter Chair and IEEE University of Ljubljana Student Branch Counselor.
- Anton R. Sinigoj** was born in Ljubljana, Slovenia, in 1955. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Ljubljana in 1978, 1982, and 1989, respectively.
- He began his career as an Assistant Professor at the Faculty for Electrical Engineering, University of Ljubljana in 1989. His research interests include electromagnetic fields, numerical methods in electromagnetics, antennas and propagations in optical fibers, electric and magnetic shielding, and induction warming. He is an author of the books: *Fundamentals of Electrical Engineering and Electromagnetic Fields* (Ljubljana, Slovenia: FE and FRI Publishing, 1994 and 1996, respectively).
- Janez Bešter** received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Ljubljana, Faculty of Electrical Engineering, Slovenia, in 1979, 1982, and 1995, respectively.
- He is the Head of the Laboratory for Telecommunications and Associate Professor at the Faculty of Electrical Engineering in Ljubljana. His 20 years of research and development and teaching activities have focused on the field of planning, development, and management of telecommunication systems and services and web-based learning systems.
- Dr. Bešter is a Member of the Association for Computing Machinery (ACM) and Associate Member of the Institute of Electronics, Information and Communication Engineers (IEICE).
- Marion O. Hagler** (S'61–M'72–SM'79–F'80) received the B.A. and B.S.E.E. degrees from Rice University, Houston, TX, in 1962 and 1963, respectively, and the M.S.E.E. and Ph.D. degrees in electrical engineering from the University of Texas, Austin, in 1964 and 1967, respectively.
- He was P. W. Horn Professor of Electrical and Computer Engineering at Texas Tech until 2001, when he was named the Robert D. Guyton Chair in Teaching Excellence and Professor of Electrical and Computer Engineering at Mississippi State University.
- Dr. Hagler was elected Fellow of the Optical Society of America (OSA) in 1981 and Fellow of the Society for Design and Process Science in 1996. In 2000, he received an IEEE Third Millennium Medal. He is Past President of the IEEE Education Society and has served as Chair of the IEEE Awards Board and as a Member of the IEEE Educational Activities Board, the IEEE Technical Activities Board, and of the IEEE Publication Services and Products Board. He served as Guest Editor for CD-ROM issues of the IEEE TRANSACTIONS ON EDUCATION published in 1996, 1997, and 1998.



Personal Content Recommender Based on a Hierarchical User Model for the Selection of TV Programmes

MATEVZ POGACNIK, JURIJ TASIC, MARKO MEZA and ANDREJ KOSIR
*University of Ljubljana, Faculty of Electrical Engineering, Trzaska 25, 1000 Ljubljana,
Slovenia. e-mail: matevz.pogacnik@fe.uni-lj.si*

(Received: 14 June 2004; accepted in final form 5 March 2005)

Abstract. In this paper we present our approach to user modeling for a personalized selection of multimedia content tested on a corpus of TV programmes. The idea of this approach is to classify content (TV programmes) based on the calculation of similarities between the description of content and the user model for each description attribute. Calculated similarities are then combined into a classification decision using the Support Vector Machines. The basis for the calculation of similarities is a hierarchical structure of the user model, overlaid upon a taxonomy of TV programme genres. Preliminary results show that it works well with a varying quality of content descriptions including incomplete genre classification and arbitrary number of description attributes. The evaluation of the system performance was based on content described using the TV-Anytime standard, but the approach can be adapted for search of other types of content with multi-attribute descriptions.

Key words. hierarchical user model, multimedia, support vector machines, TV programmes, updating of the user model, user modeling

1. Introduction

The advent of digital devices, especially in the field of consumer electronics, is about to bring to our homes a number of massive storage devices with an adequate processing power and Internet connections. According to (Greenspan, 2003), by 2006 the number of Personal Digital Videorecorders (PVRs) will increase tenfold and will reach between 10% and 15% of European households. As these boxes become part of a connected network (cable, ADSL, etc.), the step from storage to the sharing of content is a small one. The connectivity of home devices with different networks like peer-to-peer, for example, results in new means and sources of content search and acquisition. The success of existing, although compromised, peer-to-peer systems, such as Napster, Kazzaa, etc., is clearly indicating that such networks are interesting for people and that various kinds of multimedia content are being exchanged among users. Additional evidence of such trends are the research efforts of the European IST projects such as Share it! (Share it! project, 2001), in which the Personal Digital Video Recorders (PVRs) connected with a network of sharing devices are being developed.

Clearly, the number of sources of content will increase further in the future. However, such rapid development also has a downside. More content may mean more choice, but very often the user gets lost among the available choices. The appropriate solution to this problem is personalized selection of content. The properties of the existing approaches are shortly discussed in Section 4.

In this paper we propose a recommendation approach that supports the selection of content. Since we wanted to ensure that our novel personalization algorithms could also be used in devices which do not have access to information about other users' preferences, we first focused on the development of a content-based recommender. This approach is also applicable to those users who may not be willing to give away information about their preferences and will therefore not get any information in return. Nevertheless, content-based filtering still requires a large knowledge effort which can be alleviated by using the metadata information supplied to the Personal Digital Video Recorder. For example, through the Digital Video Broadcast (DVB) stream. Therefore, this approach should be seen as the first stage in the development of a more extensive recommender system. In the future we intend to enhance it by combining it with collaborative approaches.

Making personalized recommendations as relevant as possible by using a content-based approach requires good content descriptions, appropriate structure of the user model and effective selection algorithms. The TV Anytime standard (TV Anytime S-3, 2002), for example, has foreseen content descriptions with a number of attributes including programme genre, title, synopsis, actors, production country, etc. These descriptions are at the moment intended for TV programmes, but could in the future be used for descriptions of other content as well (TV Anytime R-1, 2003). In this description standard the genre attribute takes a value from a predefined hierarchical taxonomy that provides information on relatedness of content genres. From our and other authors' experience (Ardissono et al., 2004) it is evident that the quality of content descriptions may vary, especially in terms of the genre categorization and the number of descriptive attributes. Namely, a lot of TV programme descriptions that are contained in EPG bundles have incomplete genre denotation, like "Sports" instead of "Sports:Football" or "Movies" instead of "Movies:Drama", for example. Nevertheless, the user model and selection algorithms should enable good performance, regardless of the quality of content description. They should not require too much processing power and the system should also be capable of frequent, computationally undemanding updates of the user models. The last two requirements are especially important for multi-user systems such as third-party web services or personalization systems running on PVRs (Lee et al., 2002; O'Sullivan et al., 2004; Manjunath et al., 2002).

In the paper we present our approach to the classification of content that is described with multi-attribute metadata and the updating procedure of the user model. For the evaluation of our approach we used TV programmes that are described with TV Anytime metadata. It should be noted that this approach is suitable for classification of content based on any multi attribute description

standard. The goal is the classification of content into two classes: recommended and not recommended.

The classification is based on calculations of similarity between the content description and the user model. Similarity is calculated separately for each of the description attributes (genre, keywords, actors, etc.). The calculated attribute similarities are then combined into a total similarity measure using the Support Vector Machines (SVM). The novelty of this approach is the usage of the Support Vector Machines for this purpose, although combining separately calculated attribute similarities into a total similarity value is not a novel approach by itself (see Smyth and Cotter, 1999 for example). One of the advantages of this approach is the possibility to train the classification procedure separately for every user. In this way the individual relevance of every attribute is taken into account for every user. For computations of attribute similarities we used a novel structure of the hierarchical partial user models that is able to handle inadequate content descriptions including incomplete genre classification (e.g. Movies:Unknown instead of Movies:Drama) and missing description attributes.

In addition we present our approach to the updating of the user model based on descriptions of a selected or ignored content.

The quality of our approach was tested in a number of ways: we made a classification comparison using the universal and the custom SVM model (see Section 3.6.1), then we compared classification quality for content with detailed and less detailed genre denotation, investigated the usage of lexical relations between keywords (Wordnet, see Section 3.5.1.3) and analyzed the classification quality based on single attributes such as genre and keywords.

After the presentation of background in Section 2, we describe the basic structure of our user model, define the content metadata and present the calculation of attribute similarities in Section 3. In Section 3 we also explain how the SVM-based content classification works and describe the updating of the user model. Section 3 concludes with a description of the testing system architecture and of the testing procedure. Section 4 presents related work in the field of personalized selection of TV programmes. Section 5 gives results, their analysis and comments. We offer some concluding remarks in Section 6. The abbreviations used in this paper appear in Table I.

Table I. Table of abbreviations

Abbreviation	Explanation
BBN	Bayesian Belief Networks
CBF	Content Based Filtering
CF	Collaborative Filtering
CRID	Content Reference IDentifier
DVB	Digital Video Broadcast

Table I. continued

Abbreviation	Explanation
EPG	Electronic Programme Guide
IMDB	Internet Movie DataBase
LMS	Least Mean Square
MHP	Multimedia Home Platform
PVR	Personal (digital) VideoRecorder
PUM	Partial User Model
RBF	Radial Basis Function
SVM	Support Vector Machines
TF-IDF	Term Frequency – Inverse Document Frequency
UM	User Model
WAP	Wireless Access Protocol

2. Background

According to Kobsa (2001), the first user modeling approaches appeared in the late 1970s. At the time they were general and modular in structure, containing explicit assumptions about users' preferences. Over time they became more topic oriented, focusing on commercial applications. The complex reasoning capabilities and assumptions of generic user modeling systems were gradually replaced by simpler approaches. On the other hand, the introduction of the client-server architecture in these systems has enabled the possibility of comparing users among each other (Kobsa, 2001). Today we can speak about a number of recommendation approaches: collaborative recommenders, content-based recommenders, demographic recommenders, utility-based recommenders, knowledge-based recommenders and also group-oriented recommenders. They all have their advantages and drawbacks and are not universally suitable for all fields of usage. Therefore, they are often combined into hybrids (see Burke, 2002 for an overview).

Our goal has been to build an efficient recommender system by using a computationally undemanding process for updating the user model. It should mostly be based on implicit feedback, with the possibility of future usage on a PVR with MHP (Multimedia Home Platform) support. MHP is a technical solution intended for user terminals such as digital TVs or PVRs which enables execution of applications on these devices.¹ It is Java based and offers functionalities such as development of graphical user interfaces, reception of user's commands, database access, internet connectivity, playback and manipulation of audio/video content, etc. In this way we should be able to get access to the PVR usage history (play, record, stop, delete). This information is undoubtedly richer than the information about selected and ignored programmes and should prove as an upgrade to the existing approach. In fact, such an approach has, in a limited way, already been used in (Lee et al., 2002; O'Sullivan et al., 2004) and is also being standardized by the MPEG-7 (Manjunath et al., 2002).

¹<http://www.mhp.org/>

3. Recommendation of Content Items

The design of our system has been based on the idea that the system for a personalized selection of TV programmes should be able to cope with an incomplete genre classification and an arbitrary number of description attributes and should provide some recommendations without constantly bothering the user for explicit feedback. The algorithm should require a relatively low amount of processing in order to ensure near real-time user model updating and estimation of content suitability on devices with limited processing power. It should be noted that the algorithms and structures have been designed for personalized selection of any content with a multiple attribute description. Therefore we will not refer to “TV programmes” and will use instead the term “content”.

In the following subsections we will briefly examine the type and properties of the content descriptions used, give an insight into the user modeling procedure and updating of the user model and explain the testing procedure.

3.1. CONTENT DESCRIPTIONS

The content descriptions that have been used in the evaluation process of our system are based on the TV Anytime metadata standard (TV Anytime S-3, 2002). This means that they are described with a number of description attributes like “genre”, “title”, “synopsis”, “actors”, “directors”, “production country”, “production year” and “group CRID”.² Not all programmes had all of the mentioned attributes in the description; non-movies mostly had two or three (genre, title, synopsis, group CRID), while movies had six or seven (genre, title, synopsis, actors, directors, production country, production year). Most of the content descriptions had been obtained from the TVData Media Tribune company and were then additionally enriched with the information from the Internet Movie Database (IMDB). A minor part of the content descriptions came from the database used during the myTV project (myTV final report, 2000). Altogether our database contains descriptions of some 20000 TV programmes. They include descriptions of the same programmes broadcast on different channels or at different times. We must point out that many of these descriptions are of poor quality. The possible values that the genre attribute may take are hierarchically organized in a tree structure. The tree structure gives additional information about the relatedness between genres and is as such more advanced than a linear list of genres (see Figure 1).

3.2. GENERAL OVERVIEW OF CONTENT RECOMMENDATION

Our approach is organized in two steps: First, the similarity between the user model and content is estimated separately for each description attribute (genre,

²“Group CRID” (Content Reference IDentifier) attribute contains a reference to a group of related programmes like “Star Trek” series or all “Euroleague” basketball matches, for example.

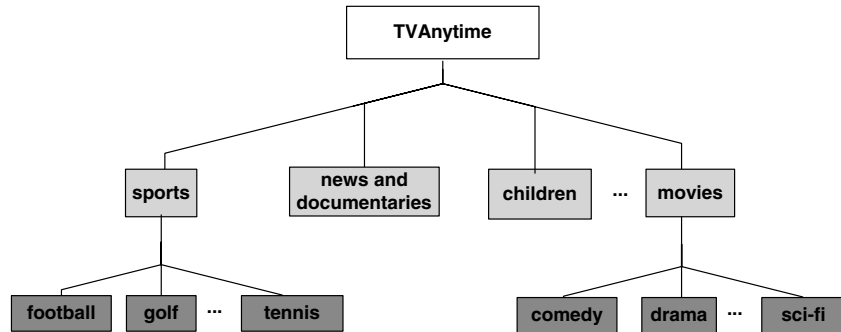


Figure 1. Example of the genre hierarchy.

keywords, actors, director(s), etc.). In the second step the content is classified as either recommendable or not recommendable. The classification based on the calculated attribute similarities was made using the Support Vector Machines (SVM) (Chang and Lin, 2003). We have established that the SVM approach, using Radial-Basis-Function (RBF) kernel, outperforms other approaches like expert-defined weighted sum (Bezerra et al., 2002) or linear LMS (Stark and Wood, 2002).

Despite the fact that the attribute similarities are calculated independently of each other, some dependence between attributes is taken into account. Our assumption is that the desired attribute values, such as directors, actors, etc. depend to some extent on the genre value. Therefore, they are stored beneath each particular genre notation in the user model, keeping the information about the genre of the programme in which a director, actor, etc. was present. Genres used in our experimental user model were taken from the genre hierarchy (taxonomy) used during the myTV project (see Figure 1). Our approach works with an arbitrary genre taxonomy. The hierarchical taxonomy of the main attribute (genre) is recommendable, but not mandatory. Details of the proposed approach are described in sections that follow.

3.3. CONTENT METADATA

A class of multimedia content items (TV programmes) is denoted by H and a multimedia content item is denoted by $h \in H$. Multimedia content item h is a given multimedia item, for instance Michael Moore's documentary "Bowling for Columbine".

Each content item $h \in H$ is described with a metadata item $md(h)$. In this way $md(h)$ describes each of the attribute classes. The attributes of metadata are grouped into n classes $A_1, \dots, A_n$; a single attribute value is denoted by $a_j^i \in A_i$. According to the chosen subset of the TV Anytime standard used in MyTV project (myTV final report, 2000) and our experiment, $n = 7$. It contains the following attribute classes: genre classification: A_1 , keywords (obtained from the title and

synopsis): A_2 , actors: A_3 , director(s): A_4 , country of origin: A_5 , year of origin: A_6 and group CRID: A_7 , as described in section 3.1. Consequently, the typical attribute values are: name of an actor a_j^3 , keyword of the title a_j^2 , country of origin a_j^5 , etc. The first class of attributes (A_1) is a set of vertices of the genre hierarchy tree T_G , i.e. $A_1 = V(T_G)$, explained in the next paragraph. In order to maintain our formal presentation, for $1 \leq i \leq n$ the extraction of the attribute $a^i \in A_i$ is denoted by md_i .

A hierarchical genre taxonomy is presented by a directed tree (directed graph with no cycles, for reference see (Korte and Vygen, 1991)) $T_G = (V(T_G), E(T_G))$ of genres. The set of tree vertices is a set of genres $V(T_G) = \{g_1, \dots, g_m\}$. An example of a genre is “Movies:Drama”. A hierarchical genre ordering is given by a set of directed tree edges $E(T_G) = \{(g_i, g_j) : g_i, g_j \in V(T_G)\}$, i.e. the directed edge (g_i, g_j) denotes that the genre g_j is a sub-genre of the genre g_i . Symbolically, “Movies:Drama” $\equiv g_j$, “Movies” $\equiv g_i$ and *drama is a subgenre of the genre movie* $\equiv (g_i, g_j) \in E(T_G)$. Figure 1 presents an example of the genre hierarchy.

Keywords are extracted from the title and the synopsis and are processed using the elimination of “stop words” and “stemming”. “Stop words” are common words such as “where” and “how”, as well as certain single letters and digits. Their lists are widely available on the Internet. “Stemming” is a well-known procedure which replaces different word endings with a common ending for the word type (Porter, 1980). For example, with stemming, the words “boating”, “boats” and “boat” would get the same form – “boat”.

An example of a metadata item is shown in Figure 2. In this example, the third attribute of the movie drama “Mona Lisa”: h is a list of actors (Bob Hoskins, Cathy Tyson, Michael Caine, etc.) $\equiv md_3(h) \subseteq A_3$, the fifth attribute is the movie’s country of origin (UK) $\equiv md_5(h) \subseteq A_5$, etc. Here $md_1, \dots, md_n$ are the coordinates of the metadata item $md(h)$ ($md(h) = (md_1(h), \dots, md_n(h))$) for content h . We should again note that the first attribute $md_1(h)$ is the genre attribute sometimes also denoted as $g \in V(T_G)$.

3.4. HIERARCHICAL STRUCTURE OF THE USER MODEL

3.4.1. Basic User Model Structures and Definitions

The user model is a data structure that contains the system’s assumptions about the user’s preferences. The proposed user model is rather complex. In order to provide an exact presentation of the user model and clarify the process of its updating, a formal presentation introducing sets and maps is given in this section.

A hierarchical genre taxonomy described in Section 3.3 is used as a model for organization of the user model. In fact, the user model is an overlay model based on the hierarchical genre taxonomy. Its purpose is to enable estimation of similarity for content with incomplete genre notification.

Genre	Movies:Drama
Title & Synopsis	Mona Lisa George, after getting out of prison, begins looking for a job, but his time in prison has reduced his stature in the criminal underworld..
Actors	Bob Hoskins, Cathy Tyson, Michael Caine, Robbie Coltrane, Clarke Peters, ...
Director(s)	Neil Jordan
Country	UK
Production year	1986

Figure 2. Example of a metadata item.

The user's preferences towards particular actors, genres, keywords, etc. are kept in metadata entries. The metadata entry is a 3-tuple (a_j^i, w, l) :

- $a_j^i \in A_i$ is the j -th attribute value from class A_i .
- $w \in [-100, 100]$ is the weight of a_j^i , for some $1 \leq i \leq n$.
- $l \in \mathbb{N}$ is the natural number indicating how many times the user interacted with the content items containing the attribute value a_j^i .

Weight w implicitly indicates the user's preferences towards the attribute value a_j^i . We should note that the default value of the weight w for any attribute value is 0 and its value changes depending on the user feedback in a negative or positive direction (see Section 3.7). Typical metadata entries for all attributes are presented in Figure 3. In this example we can find a metadata entry for the attribute class "Actors" (A_3) referring to the attribute value "Robert de Niro", with a weight value 19 obtained from nine interactions in which the user was given a possibility to choose or ignore a movie drama, with Robert de Niro in its list of actors. A set of metadata entries pertaining to a class of attributes A_i is denoted by $Z_i = A_i \times [-100, 100] \times N$.

3.4.2. Partial User Models

Partial user models ($PUM(U, g)$) are used to store the user's preferences towards the attribute values (actors, directors, keywords, etc.) recorded under the notation of genre g . Formally, PUM is a set of metadata entries.

An example of $PUM(U, g)$, where genre g is "Movies:Drama", is presented in Figure 3.

A n -tuple of sets of metadata entries Z_i , $1 \leq i \leq n$ is assigned to a pair (U, g) , where U is a user and $g \in V(T_G)$ is a genre. It is called a "partial user model"

Genre (A ₁)	(Movies:Drama, 55, 25)
Keywords (A ₂)	(murder, 13, 7), (police, 19, 9), (lover, - 5, 3), (innocent, 7, 5), (family, - 5, 3),...
Actors (A ₃)	(Robert de Niro, 19, 9), (Michael Caine, 5, 3) (Nicole Kidman, 9, 3), ...
Directors (A ₄)	(Stanley Cubrick, 9, 3), (Woody Allen, 5, 3) (Igor Sterk, 4, 4), ...
Country (A ₅)	(USA, 35, 25), (France, 12, 4), (UK, 17, 7),...
Product year (A ₆)	(2010, 1, 3), (2000, - 4, 12), (1990, 20, 16), (1980, 20, 12),...
Group CRID (A ₇)	

Figure 3. Example of a partial user model ($PUM(U, g)$) for genre “Movies:Drama”.

and is denoted by

$$PUM(U, g) = \left(\{(a^1, w, l): a^1 \in A_1\}, \dots, \{(a^n, w, l): a^n \in A_n\} \right).$$

A set of “partial user models” assigned to each user U is denoted by

$$UM(U) = \{PUM(U, g): g \in V(T_G)\},$$

where $UM(U)$ is the complete user model of the user U , composed of the partial user models ($PUM(U, g)$) for all genres g in the genre hierarchy. As such it is an overlay model, based on the genre hierarchy (taxonomy).

3.5. ATTRIBUTE SIMILARITY

When the content h is to be “compared to the user model UM ”, for each index $1 \leq i \leq n$ a similarity is calculated among the pertained metadata entry $md(h)$ and the corresponding partial user model $PUM(U, md_1(h))$ describing past “treatment” of multimedia content items of genre $md_1(h)$ by user U . Let us point out that each partial user model $PUM(U, md_1(h))$ describes the user’s preferences in terms of all classes of attributes $md_1(h), \dots, md_n(h)$. The first attribute genre $md_1(h)$ is distinguished and used in the notation of the partial user model since partial user models are organized by genre $md_1(h) \in A_1$.

The similarity among partial user model $PUM(U, md_1(h))$ and metadata entry $md(h)$ is calculated by a map

$$Sim_i: \mathcal{P}(Z_i) \times A_i \rightarrow \mathbb{R}, \quad 1 \leq i \leq n$$

where $\mathbb{R}$ is a set of real numbers and $\mathcal{P}(Z_i)$ is a power set (a set of all subsets) of metadata entries. Similarity maps $Sim_1, \dots, Sim_n$ measure similarities based on attributes $A_1, \dots, A_n$ and are combined into a similarity map of the metadata entry $md(h)$ and the partial user model $PUM(U, md(h))$:

$$Sim(PUM(U, md_1(h)), md(h)) \in \mathbb{R}^n$$

Therefore, the similarity of multimedia content item h and the user model representing the preferences of user U is given by a vector of n real numbers.

3.5.1. Approaches Used for Calculations of Similarity Maps

Calculations of similarity map $Sim_i(PUM(U, md_1(h)), md_i(h))$ for the attributes of content item h are made using the following steps:

- First the relevance weight for an individual attribute value (such as an actor, a keyword, etc.) within an attribute category is calculated.
- Then the similarity value for the i -th attribute (such as the actors attribute) is calculated as a normalized sum of the relevance weights for the individual attribute values.

The relevance weights are calculated using two well-known approaches from the personalization domain that have been adapted for this purpose, namely the Bayesian approach and the *TF-IDF* (Term Frequency Inverse Document Frequency) approach (Hand et al., 2001).

These two approaches are being used separately for each of the description attributes. For the attributes genre, actors, director and group CRID the similarity maps are calculated using the modified Bayesian approach (Uchyigit and Clark, 2002), while similarity maps for attributes keywords, country of origin and production year are calculated using the *TF-IDF* approach. We decided to use the latter approach in the calculation of similarity for the attributes whose distribution of attribute values is noticeably unbalanced (unequal frequency of different keywords in textual descriptions of content, unequal frequency of countries of origin and production years). The use of the modified Bayesian approach was motivated by its simplicity in terms of computational requirements and because it gives good results.

The basis for our calculations is in both cases the information contained in metadata entries (a_j^i, w, l) . For the *TF-IDF* approach the value of weight w is directly used in calculations, so no recalculations of additional parameters are necessary. For the Bayesian approach the information about the number of occurrences when a particular attribute value a_j^i is selected or ignored is important. These two values are calculated from weight

value w and the value of parameter l of the metadata entry, taking into account that the default weight value is 0.

3.5.1.1. Calculation of Similarity Maps Using the Bayesian Approach. First we introduce a set of decisions $Dc = \{dc_1, dc_2\}$ each user can take about a given content h . Here, dc_1 is the class of ignored or not recommendable content items, while dc_2 is the class of recommendable items. The Bayesian approach predicts the classification probability $Pr(dc_2|a_j^i)$ of a content item h to the class of recommendable items dc_2 by considering the presence of a particular attribute value a_j^i in metadata item $md_i(h)$ using:

$$Pr(dc_2|a_j^i) = \frac{Pr(a_j^i|dc_2)Pr(dc_2)}{Pr(a_j^i|dc_2)Pr(dc_2) + Pr(a_j^i|dc_1)Pr(dc_1)} \quad (1)$$

The probabilities used in (1) are calculated from the statistical data, obtained from the user's past interactions.

This means that we estimate the probability of a content item h being accepted by the user by considering the presence of a particular actor, director, etc., in the description of the content item. We used a modification of this approach as suggested in (Uchyigit and Clark, 2002). The relevance of an attribute value a_j^i is expressed by relevance weight u_j^i , calculated as a conditional probability value using (2) (Uchyigit and Clark, 2002).

$$u_j^i = \left(Pr(dc_2|a_j^i) - 0.5 \right) (fabs(sig)) \quad (2)$$

The modification was applied using the $sig(h_1, h_2, n_1, n_2)$ function (see (3)). The parameters h_1 and h_2 represent the number of occurrences when an attribute value (e.g. movie actor "Robert De Niro", director "Steven Spielberg", etc.) was selected (h_1) or ignored (h_2). n_1 and n_2 represent the total number of occurrences when any attribute value of this attribute class (any actor, any director, etc.) was selected (n_1) or ignored (n_2).³ The $sig(h_1, h_2, n_1, n_2)$ function returns higher values for the statistical data of those attribute values whose conditional probabilities differ from the conditional probabilities calculated over the entire data corpus. Apart from that, it also gives an emphasis to those attribute values whose conditional probabilities are calculated from more numerous samples. The $fabs(x)$ function is the absolute value function of the real number x . Typical values of the $fabs(sig(h_1, h_2, n_1, n_2))$ function are in the range from 0 to around 1, although theoretically they can be higher. The value of the function $fabs(sig(h_1, h_2, n_1, n_2))$ is 0 when the ratio selected/ignored is exactly the same for a particular attribute value (e.g. actor "Robert De Niro") and for the overall corpus of attribute values of the same class (e.g. all actors). This value is around 1 when there is a significant difference between the mentioned ratios.

³In our terminology an attribute value is selected/ignored when the content item which has this attribute value in its description is selected/ignored.

$$\text{sig}(h_1, h_2, n_1, n_2) = \frac{\frac{h_1}{n_1} - \frac{h_2}{n_2}}{\sqrt{\left(\frac{h_1+h_2}{n_1+n_2}\right) \left(1 - \frac{h_1+h_2}{n_1+n_2}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (3)$$

Let us illustrate this with an example. A user's history of interactions with the content contains a number of selected and ignored movies and consequently a number of selected and ignored actors, directors, genres, etc. Let us assume that the number of "selected" actors in his/her partial user model for movies is 100 and the number of "ignored" actors is 500. If a particular actor (e.g. "Robert De Niro") were "selected" five times and "ignored" once, the value of $\text{fabs}(\text{sig}(5, 1, 100, 500))$ would be 1.05. On the other hand, if this actor were "selected" only once and "ignored" five times, then the value of $\text{fabs}(\text{sig}(1, 5, 100, 500))$ would be 0. In this way the $\text{fabs}(\text{sig}(h_1, h_2, n_1, n_2))$ function diminishes the effect of those attribute values whose statistical preferences in the user model do not differ from the overall statistical preferences.

Let us demonstrate the calculation of the attribute value relevance with an example. This example represents a small part in the entire calculation of attribute similarities between the metadata items of the movie "Mona Lisa" (see Figure 2) and the partial user model of a fictional user U , presented in Figure 3. We will demonstrate the calculation of the attribute relevance for one attribute value from the attribute class A_3 (actors). In our case it will be the relevance of the actor "Michael Caine". Since "Mona Lisa" is a movie drama, the calculation of similarity will be calculated using the partial user model for the genre "Movies:Drama":

1. First we calculate $\text{Pr}(dc_2|a_j^i)$ – the probability that the movie in question belongs to the category of "selected" content items (dc_2), given that actor "Michael Caine" (a_j^i) is starring in the movie. Figure 3 contains the metadata entry in partial user model $PUM(U, md_1(h))$, $md_1(h) = \text{"Movies:Drama"}$, for actor "Michael Caine", which is $(\text{MichaelCaine}, 5, 3)$. As mentioned above, weight w and parameter l are used to calculate the number of occurrences when content containing attribute value a_j^i is selected or ignored. Knowing the procedure of updating the attribute values in the partial user model (see Section 3.7), it turns out that user U has twice "selected" and once "ignored" the movies with "Michael Caine". In the user model we can also find information that user U has previously "selected" four movie dramas and "ignored" seven of them. Using the Bayes' theorem we first calculate the probability of this movie belonging to the category of "selected" content items ($\text{Pr}(dc_2|a_j^i)$). The calculated value is 0.667.
2. Now we calculate the value of the $\text{fabs}(\text{sig})$ function for which we additionally need the information about n_1 and n_2 parameters from the user model. In our case the values are $n_1 = 412$ and $n_2 = 754$. Knowing that $h_1 = 2$ and $h_2 = 1$, we use (3) and calculate the $\text{fabs}(\text{sig})$ value, which turns out to be 0.10. In the last step we calculate relevance weight value u_j^i using (2) and get the value of 0.0167.

The complete similarity map Sim_i for the i -th attribute (actors) is finally computed as a normalized sum of relevance weights u_j^i for all attribute values (in this case all actors) in the description of the content item. If an attribute value from the description of content (an actor, director, etc.) is not present in the user model (not yet rated), its relevance weight value cannot be calculated and is therefore ignored.

We find this approach suitable because it is simple and gives good results and because its computational requirements are independent of the number of training items.

3.5.1.2. Calculation of Similarity Maps Using the TF-IDF Approach. The *TF-IDF* approach was selected for calculations of the similarity maps for the keyword, the country origin and the production year attributes. This is due to the fact that the distribution of attribute values for these attributes is very uneven. The unequal distribution of keywords is a well-known fact, in fact for that very reason the *TF-IDF* approach was invented. According to the Internet Movie Data Base (IMDB) (IMDB, 2003), which holds data for videos, movies and TV movies from 1890 until today, over 142,000 of them originate from the USA, while other countries' quotas are much lower (UK 28,000, France 24,600, Germany 24,500, Italy 14,750, Mexico 10,522 etc.). When looking at the year of production, we get a similar picture. In the last 25 years (1980–2004) almost 206,000 videos, movies and TV movies have been filmed, while only 153,000 of them were filmed in the 50 years period before that (1930–1980). We are aware of the fact that the IMDB does not entirely reflect the situation in the field of TV production, but we believe that this data represents a relevant approximation.

We wanted to use an approach which would penalize attribute values with a high occurrence frequency. The *TF-IDF* approach has exactly this characteristic and was therefore used. Even though it is simple, it is known to give good results (Hand et al., 2001; Breese et al., 1998). We adapted it for the calculation of the similarity maps through attribute value relevance. The attribute value weights u_j^i are calculated by the multiplication of relative weight w of attribute value a_j^i in the user model by its *TF-IDF* factor (see (4)). Here x_{1g} is the total number of programmes belonging to genre $md_1(h)$, x_{2ijg} is the number of programmes of genre $md_1(h)$ containing attribute value a_j^i in its description, y_{1ij} is the number of times attribute value a_j^i is present in programme description $md_i(h)$, and y_{2j} is the number of occurrences of all attribute values (strings) in programme description $md_i(h)$.

$$u_j^i = w(a_j^i) \cdot \frac{y_{1ij}}{y_{2j}} \cdot \left(\log_2 \frac{x_{1g}}{x_{2ijg}} + 1 \right) \quad (4)$$

This calculation will be illustrated with an example. Again, the following example represents a small part in the entire calculation of attribute similarities between the metadata items of the movie “Mona Lisa” (see Figure 2) and the user

model presented in Figure 3. The calculation will be demonstrated for the attribute country of origin (“UK”).

The metadata entry in the partial user model of the user U for movie dramas ($PUM(U, md_1(h), md_1(h) = \text{“Movies:Drama”})$ is $(UK, 17, 7)$. For the calculation of relevance weight u_j^i using (4) we also need information about the values of other parameters, which are obtained from the database of the available content descriptions. In this example they have the following values: $y_{1ij} = 1$, $y_{2j} = 1$, $x_{1g} = 78$ and $x_{2ijg} = 9$. The y_{1ij} and y_{2j} values tell us that the attribute value “UK” appears only once in the description of the country of origin and that it is evidently the only country of origin. These two values are usually different when calculating similarity for keywords. Furthermore, we can see that there are 78 movie dramas in the database, out of which nine were made in the “UK”. Using (4) we calculate the value of the relevance weight u_j^i , which is 69.9631. This value is around four times higher than the profile weight value (the *TF-IDF* multiplication factor is 4.1), which is the consequence of the fact that there are only 11% of the “UK” movie dramas in the database. In case of the American movies (attribute value “USA”), which represent the vast majority of movie dramas, the *TF-IDF* multiplication factor would be only around 1.3.

The exact impact of weight value(s) w on the classification decision is not trivial to estimate. The weight w in (4) describes the user’s preference towards an attribute value and the *TF-IDF* factor tunes the impact of this preference, given the (statistical) specificity of the attribute value. Therefore, we can say that the higher the weight values the more probable the classification into the class of recommended items.

Similarity map Sim_i for attribute A_i of content item h is then calculated as a normalized sum of attribute value weights u_j^i over all attribute values in metadata item $md_i(h)$. This step is only necessary when calculating the attribute similarities whose descriptions contain more than one attribute value, which is always true for keywords (attribute class A_2) and sometimes for the country of origin (attribute class A_5), but only in the case where the content is made by a co-production of a number of countries. If an attribute value from the description of content (a keyword, country of origin, etc.) is not present in the user model (not yet rated), its relevance weight value cannot be calculated and is therefore ignored.

3.5.1.3. Extending the Calculation of Similarity by Exploiting the Lexical Relations between Keywords. We tried to further improve the quality of classification by taking into account the lexical relations between keywords in the process of calculating the attribute similarity for keywords. For this purpose we used the WordNet (Miller et al., 1993) lexical database that contains lexical relations for over 100,000 concepts and is accessible in a digital form. The lexical (semantic) relations between concepts available in WordNet are synonyms/antonyms, hyponyms/hypernyms and meronyms/holonyms. It includes concepts from different categories such as nouns, verbs, adverbs and adjectives. In our approach we limited our

scope to nouns only, taking into account synonyms, hyponyms and hypernyms. We used the following procedure: for every keyword in the description of the content we tried to find a matching keyword attribute value a_j^i in the user model. If a match was found, we used the procedure described in subsection 3.5.1.2. In the opposite case we tried to find some lexical relationship between the keyword from the content description and the keyword attribute values in the user model. We looked for synonyms, hyponyms and hypernyms. While searching for synonyms is pretty straightforward, we limited the search for hypernyms and hyponyms to the search over two levels in the ontological structure (see Figure 4). In case a lexically related attribute value was found, we calculated its relevance weight u_j^i using (4) and multiplied it by a relationship factor. We set the values of relationship factors to 0.8 for synonyms, 0.5 for direct hyper/hyponyms (one level distance in the WordNet ontological structure) and 0.3 for indirect hyper/hyponyms (two levels distance). These values were set based on our intuitive estimation. As evident from the results in section 5, this approach does not improve the classification of content.

3.6. CONTENT CLASSIFICATION ON THE BASIS OF SIMILARITY VECTORS

The calculations of attribute similarities having been made, the similarity values need to be transformed into a classification decision. A number of techniques can be used to approach the problem of classification, among which are the LMS-criteria and the p-norm (Stark et al., 2002). They can be used for estimation of linear combinations. Linear (weighted) combinations of attribute similarities have been used before in the TV domain by (Smyth and Cotter, 1999), while (Bezerra et al., 2002) explicitly sets the importance of individual attributes on a 0 to 1 scale. However, the relevance of particular attributes is not easy to define and may differ from one user to another. Furthermore, if the attribute similarities are calculated using different approaches (TF-IDF, Bayes) as we did in our case, more advanced techniques are required, since the range of similarity values can differ significantly. Such a promising technique is Support Vector Machines (SVM), which enables linear and non-linear classification. It is the technique used in our experiment. SVM outperforms other methods due to the fact that the optimization minimizes the structural risk, aiming at minimizing an upper bound on the approximation error (Chang and Lin, 2003). As maximum of the error is

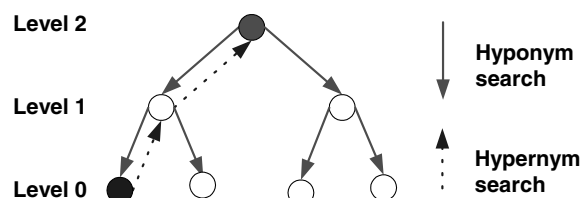


Figure 4. Search for hypernyms and hyponyms over two levels of lexical ontology.

minimized, the classification into classes is better controlled. In addition, the SVM approach trades accuracy of the approximation at the training points for smoothness of the approximating classification function and is therefore less sensible to overfitting. The differences in similarity values are handled using the normalization procedure, which maps the similarity values into the interval $[0, 1]$. For more details on SVM, see Section 3.6.1.

A set of decisions $Dc = \{dc_1, dc_2\}$ each user can take about a given content $h \in H$ has already been introduced. In our experiment, the following decision classes were used: the class of recommended content ($dc_2 = 1$) and the class of not recommended content ($dc_1 = 0$). For the training of the classification procedure and also for a later evaluation of the system's classification quality, the tracking of the user's classification decisions was necessary. A set of decisions was assigned to each user U : $b = (b_1, \dots, b_l)$, where $b_r = 0$ when a content item was ignored and $b_r = 1$ when it was selected. They were given by users during the test phase using explicit feedback to enable exact evaluation of the system performance. In a typical usage scenario with the system running on the home PVR they can be obtained from the usage history of the PVR (play, record, delete content item, etc.).

To classify a given multimedia content $h \in H$ into decision classes according to Dc regarding all n classes of attributes, a classification map $\varphi: \mathbb{R}^n \rightarrow \mathbb{R}$ is defined.

3.6.1. Classification Map Using SVM

Classification map φ is estimated on the basis of the training set with l samples $D = \{(x_1, b_1), \dots, (x_l, b_l)\}$.

- $x_r = \text{Sim}(UM(U), md(h_r))$ is a vector of n real numbers, where each number represents attribute similarity Sim_i for each of the n description attributes A_k (see Section 3.5.1).
- b_r represents the user's classification decision regarding the content item h_r .

The result is approximated classification map $\hat{\varphi} = f$, which is the solution of the optimization task. The criterion used in the estimation of the classification map is the minimization of the error vector $\varepsilon(\varphi) = (\varphi(x_1) - b_1, \dots, \varphi(x_l) - b_l)$, $(x_j, b_j) \in D$ and the optimization task is

$$\hat{\varphi} = \arg \min_{\varphi} \{\|\varepsilon(\varphi)\|\}$$

It is generally assumed that the best classification accuracy is achieved when the largest component of the error vector $\|\varepsilon\|$ is minimized, that is when the $\|\cdot\|_{\infty}$ is chosen, because in this way the classification into classes (recommended, not recommended) is better controlled. Minimization of the largest error component in practice means that the system will not likely give recommendations that are highly undesired by the user. Furthermore, such a system will probably not miss those content items that are very much liked by the user. The optimization procedure

suitable for solving the optimization task is called the Support Vector Machines (SVM) procedure.

The main idea of SVM classification is to separate the training set data with a linear decision surface (hyperplane) and maximize the margin between the different classes (recommended, not recommended). The hyperplane can also be expressed in terms of vectors which support the hyperplane, hence the name of the procedure. In real cases the two classes may not be separable, and both the hyperplane and the support vectors are obtained from the solution of a problem of constrained optimization. Another possibility is to use non-linear SVM classification. Here, the training data is first projected into a higher dimensional space, where it is separated by a hyperplane. The exact positioning of the hyperplane is ruled by the SVM kernel, which is used for later classification of the test data. We should note that the construction of a separating hyperplane in the higher dimensional space leads to a non-linear decision surface separating the two classes in the original space. However, the usage of kernels simplifies the computation of the classification procedure, by projecting the test data into a higher dimensional space, where they are classified using the approximated classification map (5); see literature (Chang and Lin, 2003).

Not all kernels are best suited to different statistical properties of the input data. We tested the linear approach and two of the available kernels: the polynomial and the Radial Basis Function (RBF) kernel. The RBF kernel performed best in our experiments and was therefore chosen for our application. It is given by:

$$K(y, z) = \exp(-\gamma \|y - z\|^2)$$

where $\gamma = 2\sigma$ is a constant calculated from the distribution of samples in the training set D , whereas y and z are vectors of an arbitrary dimension. The result of the optimization procedure using the Lagrange functionals (see Chang and Lin, 2003 for the details) is a set of support vectors x_s , their corresponding Lagrange multipliers α_s and a constant t , which together represent the hyperplane separating the two classes.

The results of the optimization (training) procedure (x_s, α_s, t) are used for the classification of content represented as vector x of the real valued attribute similarities, given by the approximated classification map:

$$f(x) = v \left(\sum_s \alpha_s K(x_s, x) + t \right) \quad (5)$$

where $v: \mathbb{R} \rightarrow \mathbb{R}$ is the *sign* function. Those samples whose function value $f(x)$ is positive are classified as “recommended” items, while others are classified as “not recommended”.

As previously mentioned, the SVM model can be either ‘universal’, i.e. obtained in advance from training sets of a number of users, or ‘custom’, i.e. obtained only from the training set of the target user. In a typical scenario, the universal SVM

model would be used as the classification map $\hat{\phi}$ in the early stages, while the custom SVM model would be built gradually in the process of user interaction with the system.

When the custom SVM model is used, the relevance of a particular attribute for the target user is better expressed in the classification model, since it indirectly depends on the correlation between the calculated attribute similarity values and the corresponding classification decisions in the training set. In fact, the results show that the universal SVM model performs well enough, but the performance can be further improved by using the custom SVM model (see Section 5).

Figure 5 presents the flow chart of the content classification.

3.7. UPDATING THE USER MODEL

In order to enable a quality similarity calculation and later classification for content with incomplete genre categorization, the hierarchical user model needs to be updated regularly and on all levels of the hierarchy. The user model is updated with metadata information $md(h)$ about content h that has been either selected or ignored by the user. Actual updates are performed on partial user models $PUM(U, md_1(h))$, belonging to genre $g = md_1(h)$ and partial user model $PUM(U, g_p)$ belonging to parent genre $g_p(h)$ of genre $md_1(h)$. In this way the information about the user's preferences is propagated to all levels of the user model hierarchy. The idea of hierarchical updating of the user model is presented in Figure 6. The example presents the updating of the user model with a metadata description of a comedy movie. First, the partial user model belonging to genre "Movie:Comedy" is updated and after that also the partial user model belonging to its parent genre "Movies". The dashed arrow depicts the propagation of information about attribute values a_j^i from metadata description $md(h)$ through the hierarchy of partial user models.

Let's now take a look at the updating procedure of a single partial user model $PUM(U, md_1(h))$ with metadata description $md(h)$. In this process the metadata entry values of the partial user model are being modified or added. When the user interacts with a content item h , the following procedure is executed. For all the attribute values a_j^i of the item's metadata description $md(h)$, we first try to find such metadata entry $(v, w, l) \in Z$ where $v = a_j^i$. If such metadata entry (v, w, l) exists, we modify its weight value w and increase the value of the counter l by 1. The modification of the weight value depends on whether the content h has been selected or ignored. In case h has been selected, the weight value w is increased by 3, while in the case it has been ignored the weight value w is decreased by 1. Using different weight changes for selected and ignored content enables simple calculation of the number of occasions when a particular attribute value was selected or ignored, which is necessary for similarity calculations using the Bayesian approach. Furthermore, in this way the positive preferences have a

higher influence on the selection of content using the *TF-IDF* approach, which we believe is important in defining user preferences. If there is no metadata entry in the partial user model where $v = a_j^i$, we simply create one with the default values $w = 3$ (selected) or $w = -1$ (ignored), $l = 1$ and with the attribute value $v = a_j^i$. This updating procedure is repeated for all the attribute values $a_j^i \in A_i$ of all attribute classes A_i that are present in the metadata description $md(h)$.

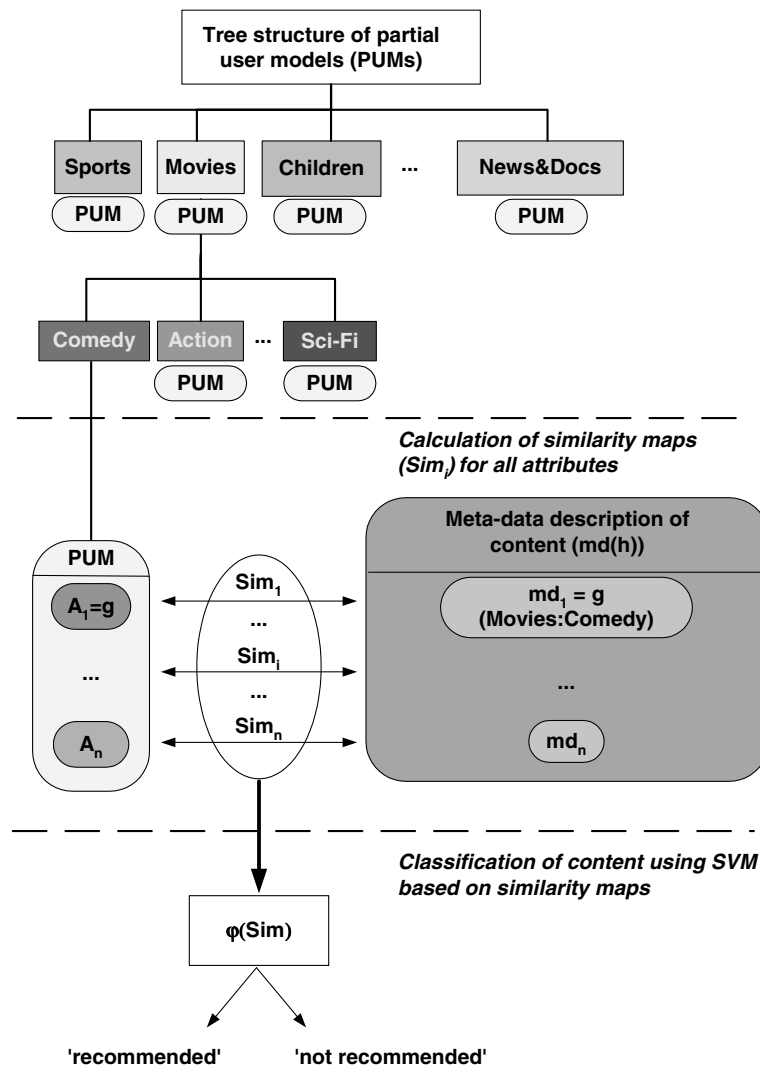


Figure 5. The flow chart of the content classification through comparison with the PUM ($U, md_1(h)$).

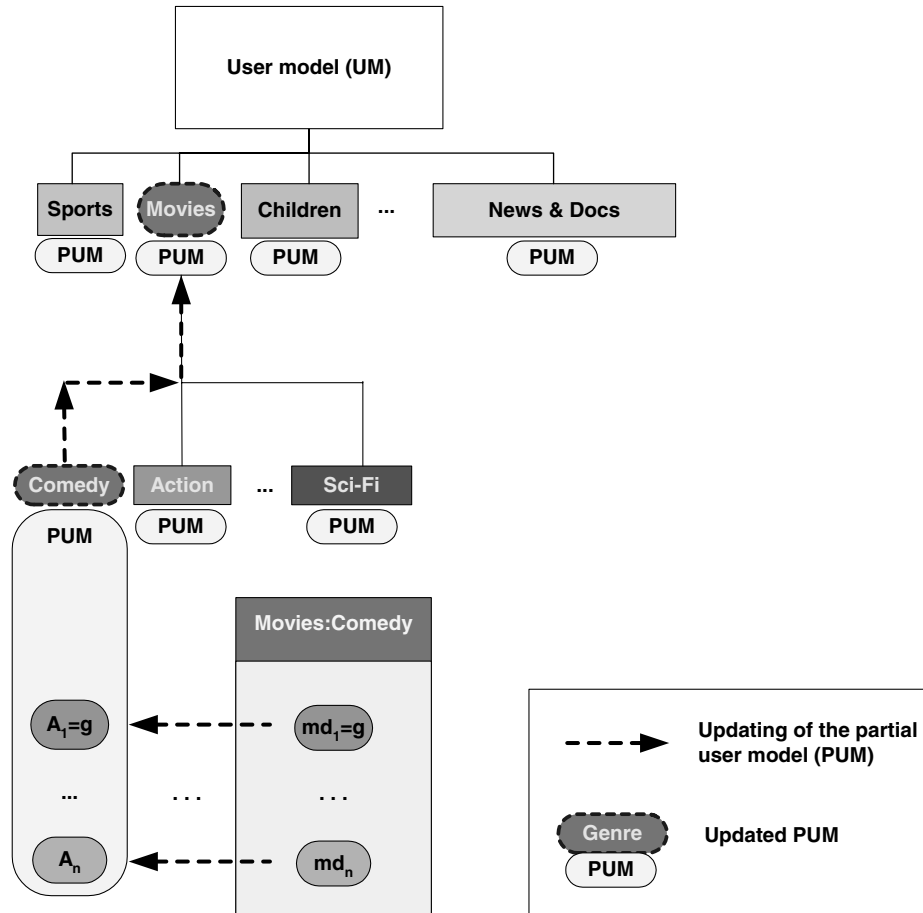


Figure 6. The updating of the user model with the metadata description of a comedy movie.

This approach seems very suitable because it is not very demanding and can be done off-line whenever the device is not busy with other processes. It is also efficient, which is proven by the quality of classification. (see results in Section 5).

3.8. IMPLEMENTATION OF THE TEST SYSTEM

Our test personalization service was implemented as a web service, using Java servlets. The TV programme metadata and user models were stored in a relational database on the server. The architecture of the system is presented in Figure 7.

TV programme descriptions covered a broadcast period of 2 months for 50 television channels. Altogether the database contained descriptions of more than 20,000 TV programmes. For details see Section 3.1. The genre hierarchy used for

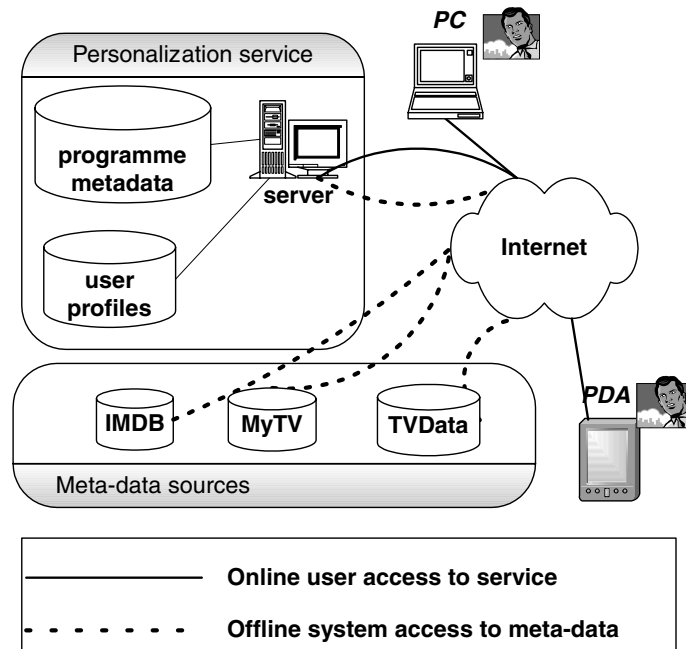


Figure 7. Test system architecture.

programme descriptions are based on the myTV version of genre hierarchy, see Figure 1.

In our test group there were 41 users. 16 of them were female and 25 of them male, ranging between 14 and 55 years of age. The average age of the male subjects was 26.7 years ($\sigma = 6.74$) and the average age of the female subjects was 32.1 years ($\sigma = 8.26$). Although the average values of male and female population differ, we believe that the tests were relevant to some extent as the age between 25 and 35 years represents the user population for recommender systems.⁴

During the testing phase, we wanted to evaluate a number of different approaches and algorithms for personalized selection of TV programmes. The test system was therefore not designed as a typical personalization service, learning and giving recommendations to users, because this would not enable us to test multiple approaches to personalized selection. Instead we decided for an evaluation procedure with two phases. The first one was the training phase used for updating of the user model and was based on implicit feedback. It was followed by the evaluation phase in which users gave explicit ratings for the test set of 150 TV programs. These programmes were then classified by our system, using a number of different settings and approaches. In this way we were able to compare classifications given

⁴Testing of statistical significance showed that the average values for the male and female subjects differ significantly. This was established by checking the hypothesis that the average values of the two testing populations do not differ significantly using the *t*-test ($p = 0.0286$).

by users with those made by a number of personalization approaches and to make a comparative analysis (see Section 5).

3.9. USER INTERACTION AND TESTING

At the first login, users were asked to fill in a simple initialization form expressing their preferences for the genres. They had the possibility to skip the entire initialization process or fill in an arbitrary number of fields. After initialization, they used the search engine enabling them to search for TV programmes based on some of the description attributes (genre, keywords in title and synopsis, actors and directors). They had an additional possibility to browse lists of programmes broadcast during a certain time period. After every query, they were presented with a list of matching TV programmes. If users liked any of the presented programmes, they marked their choice by pressing the 'I like it' button. Their decisions were based on the descriptions of programmes displayed on the screen. In this way we tried to simulate a near-future use of such a service, where the information about selected programmes would be sent to the PVR at home and the programmes would be automatically recorded.⁵

The selected and ignored programmes were used as the training set for building and updating the user model. In this way the user models were built with arbitrary training sets obtained as a result of user queries and their browsing through available TV programme listings, without unnecessary and unwanted explicit evaluations required from users.

In the evaluation phase users were asked to additionally classify over 150 TV programme descriptions into the following categories: "*I will watch it*", "*I may watch it*" and "*I won't watch it*". The first one was considered as a classification into the group of liked programmes, while the last one was considered as a classification into disliked programmes. The programmes classified as "*I may watch it*" were discarded. In this way we were able to extract those programmes for which the users had stronger preferences (negative or positive). If the users didn't know any of the programmes in the test set, they were able to choose the option "*I don't know*". One half of the test set programmes were movies, whereas the other half were non-movies (sports, entertainment, news and documentaries, children's programmes, musical programmes and category called "Other").

Figure 8 presents the design of the user interface (GUI). It is very simple and it enables only basic functionalities. Among them there are the search based on time and channel of broadcast and the search based on description attributes (keywords, genre, actors, directors, etc.) We are aware of the fact that it should be further optimized especially for giving explicit feedback. The figure presents GUI enabling the search based on description attributes. On the left hand side there are search fields while on the right hand side there are matching query results. The search for

⁵Remote control of the PVR through Web and Wap services was developed during the myTV project (myTV final report, 2000).

Figure 8. Graphical user interface.

the content based on the channel selection and the time period is available in a different operational mode and is not presented in this figure.

4. Related Work

There are many approaches in the domain of personalized selection of TV programmes. Most of them are either content-based (CBF) (Bezerra et al., 2002; Zimmerman et al., 2004; Lee et al., 2002), collaborative (CF) (Ardissono et al., 2004; Kurapati et al., 2002) or combine both approaches (Smyth and Cotter, 1999, Uchyigit and Clark, 2002). A good overview of their properties is made in (Burke, 2002). The advantages of collaborative systems are independence of the representation of content items and capability of recommending content items from cross-genre niches (music, movies, books...). The drawbacks are problems of new users and new content items. New users have not yet rated enough content items, while new content items have not yet been rated by a sufficient number of users. The content-based approach similarly suffers from the problem of new users and is not capable of cross-genre recommendations. Both approaches also need large historical data sets in order to provide quality recommendations. The mentioned approaches make recommendations to individual users, while another approach, which has come up recently, makes recommendations of TV programmes for groups of users (viewers) (Masthoff, 2004). TV viewing is often a group activity and the selection of content items (TV programs) for a group of users is not a trivial task. Authors employ a number of strategies to combine item preference ratings of individual group members in order to make suggestions to the group as a whole. From this point of view, our approach is not directly suitable for group recommendations as the content similarity is based on individual user models.

However, it can estimate the content ratings for individual users and can be used as one of the inputs for a group recommender system.

Despite the same goal – selecting a list of TV programmes for the user – existing approaches differ in many ways. The first difference is in the watching paradigm itself. Some of the systems generate recommendations of TV programmes watched at the time of their broadcast, therefore information about the day and time of broadcast is important (Kurapati et al., 2001; Ardissono et al., 2004; Zimmerman et al., 2004). Other approaches, like the TV Anytime approach (Draper et al., 1999) for example, are not concerned with the day and time related issues. They assume that users have a massive storage PVR able to record many hours of programmes recorded in advance and watched at any time.

Another issue is the problem of different metadata standards and description schemes used in the process of classification (selection) of TV programmes. Description schemes used in existing approaches include: only keywords (Ahanger and Little, 2001), a combination of attributes like genre, daytime and channel (Ardissono et al., 2004), genre, channel and actors (Smyth and Cotter, 1999), genre and keywords (Uchyigit and Clark, 2002) and genre, actors, directors, keywords, daytime and channel in (Kurapati et al., 2001; Zimmerman et al., 2004). One of the approaches (Bezerra et al. 2002) uses attributes similar to the TV Anytime standard (genre, actors, director, synopsis, production country, production year) and is from this point of view the most similar to ours. Although Lee et al. (2002) use the TV-Anytime description standard, they make recommendations based only on the genre attribute combined with the information about the day and the time of broadcast.

Similarly, a number of mathematical tools are used for personalized content filtering. These include Naive–Bayes classifier (Kurapati et al., 2001; Uchyigit and Clark, 2002, Zimmerman et al., 2004, decision trees (Kurapati et al., 2001; Zimmerman et al., 2004 Bayesian belief networks (Ardissono et al., 2004), symbolic analysis (Bezerra et al., 2002) and weighted sum of attribute similarities (Smyth and Cotter, 1999; Bezerra et al. 2002). In the last approach, the similarity between content and the user model is first computed for each of the description attributes and then combined into the overall similarity value.

All the aforementioned issues make a sensible comparison of results difficult. Nevertheless, some authors have compared a number of approaches (Kurapati et al., 2001; Zimmerman et al., 2004) and reported that the Naive–Bayesian classifier often outperforms others. In Breese et al. (1998) some comparisons have been made between decision trees, Bayesian Belief Networks (BBN), correlation methods and vector methods. Authors report that BBN and correlation methods performed best in terms of classification accuracy. In terms of processing power and memory requirements BBN outperform correlation methods during the content classification process, but are very demanding during training and building of the belief network structure. This means that updating BBN is not a trivial task. The problem is even worse if the service is supposed to run on a third-party

central server, serving a lot of users and regularly updating their models. The processing power requirements are similar in case of a home Personal Video Recorder (PVR) with lower processing capabilities.

An interesting issue is the type of feedback that systems get from users. Some approaches use explicit feedback for the revision of the user model (Smyth and Cotter, 1999; Bezerra et al., 2002; Uchyigit and Clark, 2002), others implicit feedback (Kurapati et al., 2001; Lee et al., 2002; Zimmerman et al., 2004) and some even combine both (Ardissono et al., 2004). One would normally expect better results from explicit feedback, although some authors report improved results when using implicit feedback (O'Sullivan et al., 2004) obtained through the analysis of the PVR usage history. As most of the users are not willing to spend much time to provide explicit feedback, new approaches to existing recommender strategies are being devised. In Baudisch (1998), authors suggest the introduction of the so-called "opinion leaders". These are the users which make an extra effort in rating content items and allow publishing their selected programs along with their names. Ordinary users can use suggestions of their favorite opinion leaders, which should in return be somehow rewarded for their effort. Although this approach is less powerful than explicit feedback, it improves the overall quality of the system. It is clear that the question of the feedback mechanism differs from one field of usage to another and as Baudisch (1998) states: "the higher the required effort the more potential users will quit". Therefore simple feedback mechanisms not requiring more than a touch of a button (thumbs-up, thumbs-down) should be used. They should be further enhanced with the employment of the PVR usage history, which can give good results as it represents the user's natural behavior in the selection process.

5. Results

Since the content was classified into two classes, we used for the evaluation of results two well-known performance measures: precision and recall (Hand et al., 2001). As mentioned earlier in the paper, we evaluated and compared several approaches:

- comparison of the classification quality for TV programmes with adequate and incomplete genre categorization.
- usage of the universal and custom SVM model (see Section 3.6.1).
- analysis of the system performance in which only one description attribute at a time was used for classification. This analysis was made for two description attributes: keywords and genre.
- approach in which we used a lexical database (WordNet).

The results of these four approaches are presented in Figures 9–12.

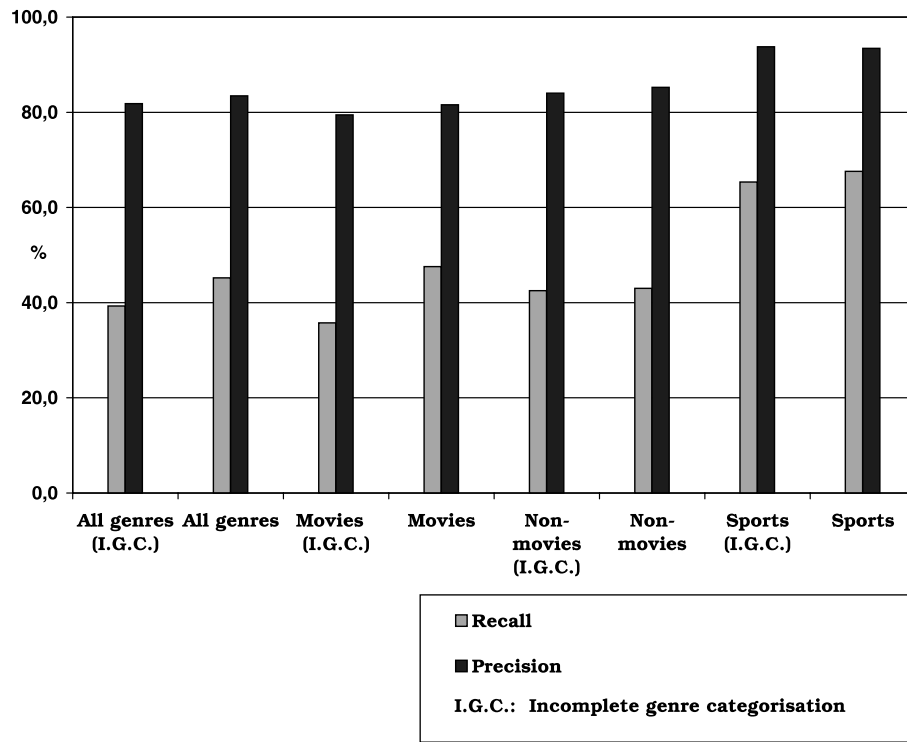


Figure 9. Recall and precision for different genres and different genre description qualities.

5.1. COMMENTS OF RESULTS

It should be noted that unless otherwise stated the presented results were obtained using custom SVM model, with no consideration for lexical relations between words.

Figure 9 presents a comparison of the classification quality for content with adequate genre categorization and content with incomplete genre categorization. First of all we can observe that in the case of adequate categorization recall values do not differ much for movies and non-movies. Average recall values are around 45%, while precision values are around 80%. Slight differences in the classification quality between movies and non-movies are present, but they remain under 5% for both precision and recall values. As expected, the best classification results were obtained for sports (precision around 90%, recall around 65%), since sports subgenres are thematically very differentiated and predicting user interests in a particular sport is therefore easier. Also to be noted is that the keywords in the descriptions of sports programmes are probably very informative from the point of view of identifying the user's interest (competition type, teams, location, etc.).

Figure 9 also demonstrates the advantage of our system regarding programmes with incomplete genre categorization in their content descriptions. The results

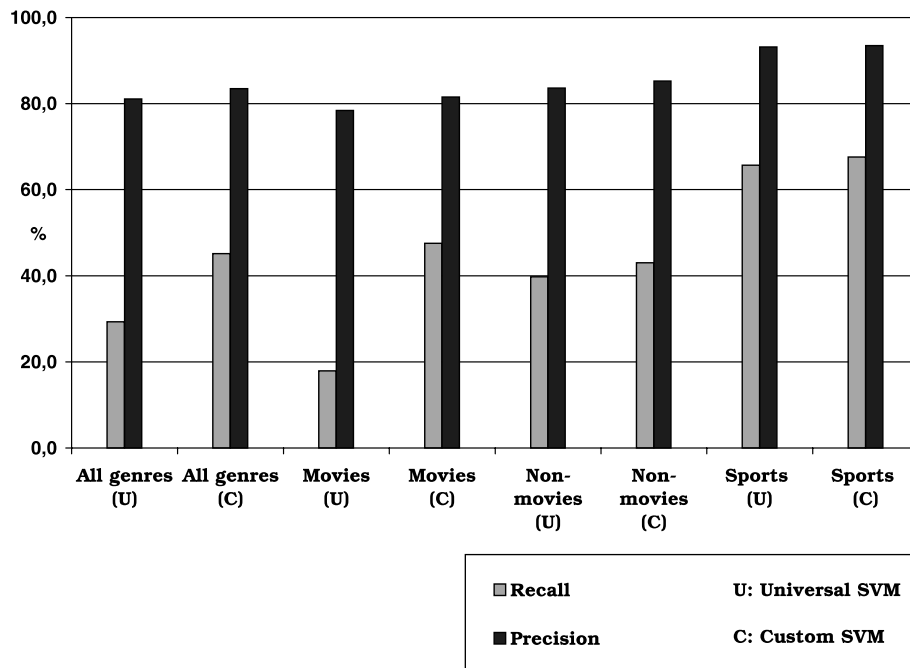


Figure 10. Comparison of recall and precision for universal and custom SVM model for different genres.

show that the decrease in precision values for such programmes is small for both movies and non-movies, which demonstrates the quality of the hierarchical user model. The drop in recall is higher for movies than for non-movies, which is most probably due to the high information value of the keywords in case of non-movies. This assumption is proven by the classification results based on individual attributes, discussed later in this section.

In Figure 10, we see a comparison of the classification quality for the universal and custom SVM model, presented for a number of genres. The results are satisfactory when either universal or custom SVM model is used. The differences between both approaches appear mostly for recall values. The classification quality when using the universal SVM model instead of the custom SVM model is 30% lower in case of movies, and only 5% lower in case of non-movies, while precision values remain almost unchanged at around 80%. In our opinion, this is a consequence of the higher dimensional classification space in case of movies described with a higher number of attributes (usually six) thus making SVM classification more complex than in the case of non-movies, which are usually described with two or three description attributes.

As some of the users tend to check the lists of authors and directors and some look at the country of origin, etc., the classification model trained on the users' own training set can reflect the information about the relevance of a particular

attribute. Therefore, the SVM model built from the training set belonging to a high number of different users cannot classify content items as well as the one built by using only the training set of the target user. At this point we should mention that our classification system was a priori biased towards high precision values at the expense of lower recall values. This was achieved through the tuning of the cost parameters (C) for both classes (recommended, not recommended) during the SVM training procedure. The consequence of such a biased tuning was ‘drifting’ of the content items into the class of not recommended items during the classification procedure. We believe that it is more important for a personal recommender system to give recommendations with higher precision, because a high number of false recommendations may deter users from using the system. On the other hand, the high number of available content ensures that the number of recommendable content will be high enough even at lower recall values.

Figure 11 shows an analysis of the classification quality using different combinations of description attributes. We compared three different scenarios: classification based on attribute similarities of all description attributes, classification based only on information about the genre and classification based only on the keywords. It is understandable that the system performs best when all attributes are used in the classification process. Nevertheless, we can make some interesting observations when we take a look at particular genre types. We can notice that in case of movies the classification is better when it is based on genres than when based on keywords. The classification based only on keywords makes practically no sense, since precision values are at around 50%. On the other hand, the classification based on keywords works better in case of non-movies and is practically not changed when information about the content genre is also used. This may be surprising at first glance, but not after some thought. The fact is that title keywords are often repeated in descriptions of non-movies, and are therefore more distinctive. Typical examples are television series, game shows, sports programs, etc., whose main part of the title is always the same (“Star Trek”, “Late night with Conan O’Brien”, “Football Euroleague”, for example). This fact is additionally supported by the classification quality in case of sports programmes. It also explains good classification performance for non-movies with incomplete genre categorization of which we spoke earlier in this section (see Figure 9).

We tried to improve additionally the classification results by taking into account the lexical relations between keywords. For this purpose we used the WordNet lexical database, See Section 3.5.1.3. However, the results show no improvement for any of the genre types (see Figure 12). The reason might be the standardized terminology in descriptions of TV programmes. An additional reason is probably the universal nature of the WordNet database, which does not cover much details of heterogeneous genre topics such as television genres. Usage of specialized, thematic lexical databases might perform better. We will try to investigate this issue in our future work.

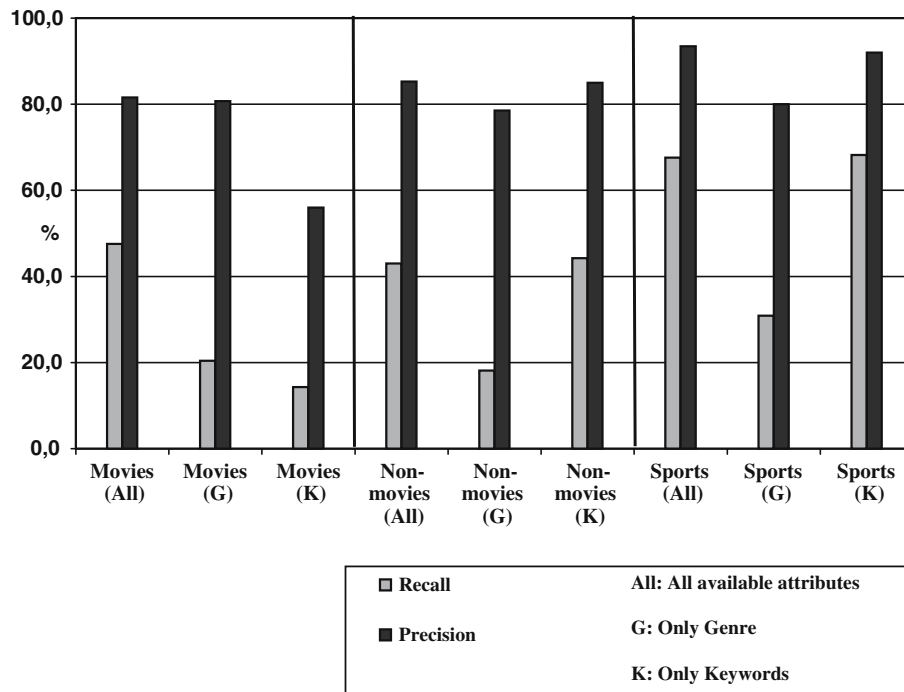


Figure 11. Comparison of recall and precision using a different number of description attributes (all attributes, only genre, only keywords).

6. Conclusions and Future Work

In this paper, we presented a concept and results of a user modeling approach for classification of multimedia content. We tested it in the domain of personalized selection of TV programmes, but this approach is applicable to classification of other content with multi attribute descriptions. The test study proved the suitability of our approach reaching the average precision value around 80%, with recall around 40%. Our novel approach is based on a hierarchical user model using calculations of the attribute similarities with the support vector machines for classification. A particular value of this approach lies in the fact that it handles well the programme classification with genre descriptions that are incomplete and that the drop in recall and precision remains within acceptable limits. A successful classification of such content is the result of the hierarchical structure of the user model and of the updating procedure.

The experiment with the Support Vector Machines showed that it is possible to further improve the classification results through training and usage of a “custom” SVM classification model. This experiment indicates that individual attributes do not have the same importance for different users and that it is possible to get improved classification results by using the same attribute similarities with

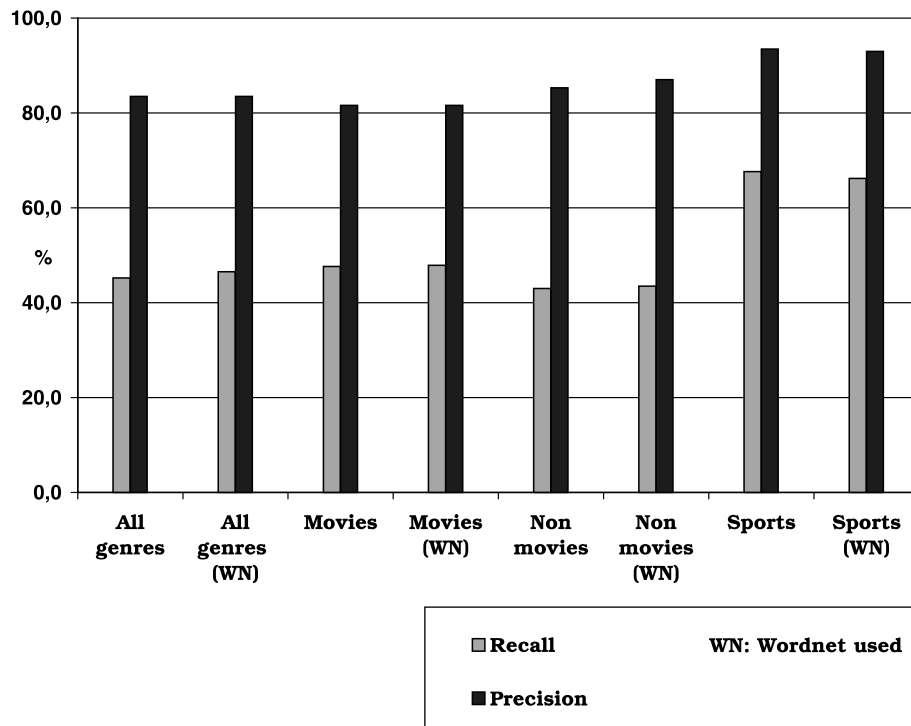


Figure 12. Comparison of recall and precision using lexical relations between keywords (WordNet).

a better adapted classification model. This assumption was proven especially for movies that have a higher number of description attributes.

The experiment with the classification based on single attributes showed that the information value of individual attributes is different for different genre types. The classification based on genres is more successful than the classification based on keywords in case of movies, and the other way around for non-movies.

Surprising were the classification results obtained through the usage of the WordNet lexical database. They showed no improvement at all. Despite the high number of concepts it was apparently unable to provide a sufficient amount of lexically related keywords and improve the classification performance.

Another issue, namely the variability of suggestions, is an important property of recommender systems (O'Sullivan et al., 2002). Content-based systems usually suffer from this problem since content is selected through comparison with the user profile that indirectly contains metadata information of past selections. This can negatively affect the variability of future suggestions. We did not directly address this issue in our approach, but since we intend to enhance it with a collaborative approach, this should also improve on the problem of variability of suggestions. However, collaborative approaches suffer from other problems such as sparsity and latency (O'Sullivan et al., 2002; Burke, 2002), which requires additional

optimization efforts. One could also improve the variability of suggestions by selecting the top N items of every genre that the user “likes” to some extent. As our user model also contains information about genre preferences, this solution would be possible to implement. However, this would improve the variability of suggestions only partially as the popularity of a particular genre is also related to the history of past content selections.

We also intend to transfer our system to the personal video recorder with MHP support and in this way get access to content usage history. We believe that conclusions about user preferences based on information about the content usage are more reliable than the current EPG based interaction with the system. Since the used algorithms are not computationally demanding, their usage on computationally weaker devices like personal video recorders should present no major problem. Evaluation of our approach in real life scenarios will finally test its real value and appropriateness.

Acknowledgements

This work would not have been possible without the grant of the Slovenian Ministry of Education, Science and Sport, No. S-64303 and our work on the myTV project (IST-1999-11702).

References

- Ahanger, G. and Little, T.: 2001, Data semantics for improving retrieval performance of digital news video systems. *IEEE Transactions on knowledge and data engineering* **13**(3), 352–360.
- Ardissono, L., Gena, C., Torasso, P., Bellifemine, F., Chiarotto, A., Difino, A. and Negro, B.: 2004, User Modeling and Recommendation Techniques for Personalized Electronic Program Guides. *Personalized Digital Television. Targeting Programs to Individual Viewers*. Kluwer Academic Publishers, Human-Computer Interaction Series, Vol. 6, pp. 3–26.
- Balabanovic, M. and Shoham, Y.: 1997, Combining content-based and collaborative recommendation. *Communications of the ACM* **40**(3), 66–72.
- Baudisch, P.: 1998, Recommending TV programs How far can we get at zero user effort. *Recommender Systems: Papers from the 1998 AAAI Workshop*, Technical Report WS-98-08, Madison Wisconsin, pp. 16–18.
- Bezerra, B., Carvalho, F., Ramalho, G. and Zucker, J. D. Speeding up Recommender Systems with Meta-prototypes. *Lecture Notes in Computer Science, Advances in Artificial Intelligence*, 16th Brazilian Symposium on Artificial Intelligence – SBIA 2002, Brazil, Springer-Verlag GmbH, pp. 227–236, November 2002.
- Billsus, D. and Pazzani, M. J.: 1999, A hybrid user model for news story classification. In: *Proceedings of the Seventh International Conference on User Modeling*, Banff, Canada, pp. 99–108.
- Breese, J. S., Heckerman, D. and Kadie, C.: 1998, Empirical analysis of predictive algorithms for collaborative filtering. In: *Proceedings of the fourteenth conference on uncertainty in artificial intelligence*, Madison, USA, Morgan Kaufmann Publisher, pp. 43–52.
- Burke, R.: 2002, Hybrid recommender systems survey and experiments. *User Modeling and User Adapted Interaction*, **12**, 331–370.

- Chang, C. and Lin, C.: 2003, Libsvm: a library for support vector machines. ELECTRONIC VERSION (<http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>).
- Draper, S., Earnshaw, N., Montie, E., Parnall, S. et al.: 1999, TV-Anytime. Paper presented at the opening of the IBC99 conference.
- Greenspan R. Remote Power – Can PVRs kill TV spots? *CyberAtlas, The newsletter for Web marketers*, Ver 5.47, 22 September 2003, http://cyberatlas.internet.com/markets/advertising/article/05941_3080851.00.html
- Gudivada, N. V., Raghavan, V. V., Grosky, I. W. and Kasanatogu, R.: 1997, Information retrieval on the world wide web. *IEEE Internet Computing*, **1**(5) 58–68.
- Hand, D. J., Mannilla, H. and Smyth, P.: Principles of data mining, The MIT Press, Cambridge, Massachusetts, USA, pp. 449–485, 2001.
- Internet movie database web site (INDB), 2003, <http://www.imdb.com>.
- Korte, B. and Vygen, J.: 1991, *Combinatorial Optimization*. Springer-Verlag, Berlin.
- Kobsa, A., 2001, Generic user modeling systems, *User Modeling and user adapted interaction*, **11**, 49–63.
- Kurapati, K. and Gutta, S.: 2002, TV personalization through stereotypes. In *Proceedings of the AH2002 Workshop on personalisation in future TV*, Malaga, Spain.
- Kurapati, K., Gutta, S., Schaffer, D., Martino, J. and Zimmerman J.: 2001, A multi agent TV recommender. In: *Proceedings of the UM 2001 Workshop on Personalization in Future TV*, Sonthofen, Germany.
- Lee, H., Lee, H., Nam, J. and Bae, B.: 2002, Personalized contents guide and browsing based on user preference. In: *Proceedings of the AH2002 Workshop on personalisation in future TV*, Malaga, Spain.
- Masthoff, J.: 2004 Group modeling: Selecting a sequence of television items to suit a group of viewers. *User Modeling and user adapted interaction*, **14**, 37–85.
- Manjunath, B. S., Salembier, P. and Sikore, T.: 2002, *Introduction to MPEG-7, multimedia Description Interface*, John Wiley & sons Ltd., England, pp. 163–175.
- Michalski, R. S., Bratko, I. and Kubat, M.: 1988, *Machine Learning and Data Mining- Methods and Applications*. John Wiley and Sons, Chichester, England.
- Miller, G. A., Beckwith, R., Fellbaum C., Gross, D. and Miller K. J.: 1933, Introduction to WordNet: An online lexical database. Technical Report CSL 43, Cognitive Science Laboratory, Princeton University.
- myTV project IST-1999-11702. Personalised services for digital television – Final report, 2000, ELECTRONIC VERSION (<http://www.extra.research.philips.com/euproject/mytv/>).
- O'Sullivan, D., Smyth, B., Wilson, D.C., McDonald, K. and Smeaton A.: 2004, Improving the quality of the personalized electronic program guiden. *User Modeling and User Adapted Interaction*, **14**, 5–36.
- Porter, M. F.: 1980 An algorithm for suffix stripping. *Program, Automated Library and Information Systems*, **14**, 130–137.
- Salton, G. and McGill, M.: 1983, *An Introduction to Modern Information Retrieval*. McGraw-Hill.
- Share it! project IST-2000-28703. Public project description, Deliverable No.0. 2001, ELECTRONIC VERSION (<http://www.extra.research.philips.com/euprojects/share-it/deliverables/share-it-del00.pdf>).
- Smyth, B. and Cotter, P.: 1999, Surfing the digital wave: Generating personalised TV listings using collaborative case-based recommendation. In: *Proceedings of 3rd International Conference on Case-Based Reasoning*, Lecture Notes in Artificial Intelligence, Vol. 1650, pp. 561–571.
- Stark, H. and Woods, J. W.: 2002, *Probability and Random Processes with Applications to Signal Processing*. Prentice Hall, New Jersey.

- O'Sullivan, D., Wilson, D. and Smyth, B.: 2002, Improving case-based recommendation: A Collaborative Filtering Approach. In: *Proceedings of the 6th European conference on case-based reasoning, ECCBR 2002*, Springer Verlag, pp. 278–291.
- TiVo: The official TiVo web site. <http://www.tivo.com>, 2003.
- TVAnytime: Specification series: S-3 metadata, SP0003 v1.3, part A, 2002, ELECTRONIC VERSION (<ftp://tva:tva@ftp.bbc.co.uk/pub/Specifications/SP003v13.zip>).
- TVAnytime: Requirements series: R-1 Business models, RQ001v20, final release, August 2003 ELECTRONIC VERSION (<ftp://tva:tva@ftp.bbc.co.uk/pub/Specifications/RQ0001v20.zip>).
- Uchyigit, G. and Clark, K.: 2002, An agent based electronic program guide. In: *Proceedings of the AH2002 Workshop on personalisation in future TV*, Malaga, Spain.
- Zimmerman, J., Kurapati, K., Buczak, A., Schaffer, D., Gutta, S. and Martino, J.: 2004, TV Personalization system. Design of a TV Show Recommender Engine and Interface. *Personalized Digital Television. Targeting Programs to Individual Viewers*. Kluwer Academic Publishers, Human-Computer Interaction Series, Vol. 6.

Authors' vitae

Dr. Matevž Pogačnik is a researcher at the Faculty of Electrical engineering, University of Ljubljana. He received his M.Sc. degree in agent technologies in 2000 and his Ph.D. in personalized content retrieval in 2004. His research interests include user modeling and knowledge-based data mining. Currently he is working on European IST projects, in which he is developing mechanisms for personalized selection of multimedia content applied to heterogeneous environments (p2p networks, mobile devices, set-top-boxes, PDAs, etc.).

Dr. Jurij Tasič is a Professor of system theory and computing at the University of Ljubljana. In 1983 he began working on communication systems and the usage of numerical methods in DSP, including the usage of system theory in communications and communication systems. His current interests are advanced algorithms in communication systems, multidimensional signal processing and parallel algorithms. He is the author of more than 60 research papers published in journals and conferences.

Marko Meža is a researcher and a Ph.D. candidate in Telecommunications at the University of Ljubljana. He received his B.A. in Telecommunications in 2001 from the University of Ljubljana. His primary research interests lie in the areas of networking, and informatics in medical processes. His work includes development of a tele-medical system for the Blood Transfusion Center in Slovenia.

Dr. Andrej Košir received his M.Sc. degree in probability theory in 1996 and his Ph.D. degree in electrical engineering in 1999, both from the University of Ljubljana. He is currently engaged in research of digital image understanding, digital data coding, stochastic optimization methods and statistical data modeling. His current interests include statistical data modeling for communication, and higher level image understanding, using formal systems and fuzzy logic.

Robust power line modem synchronization using signal phase measurement

J Ambrozic¹, A Kos² and J Bester²

¹ Iskraemeco, d.d., Savska loka 4, Kranj, Slovenia

² Faculty of Electrical Engineering, University of Ljubljana, Trzaska 25, 1000 Ljubljana, Slovenia

E-mail: janez.ambrozic@iskraemeco.si, andrej.kos@fe.uni-lj.si and janez.bester@fe.uni-lj.si

Received 21 February 2007, in final form 23 May 2007

Published 17 July 2007

Online at stacks.iop.org/MST/18/2686

Abstract

Intelligent automatic meter reading and automatic meter management systems that collect, communicate and process data from different meters, i.e. electricity, gas, water, etc, form an important building block of smart and efficient environments. When using power line modems for communications, achieving fast and accurate synchronization is one of the key issues. This paper provides a novel contribution of phase measurements of the power line signal for robust synchronization purposes. Phase calculation is based on an extended Goertzel algorithm providing zero crossing information and reducing the negative effects of noise and disturbances. With relatively small complexity and low implementation costs it assures sufficient accuracy and speed of convergence. Practical realization on a low-cost microprocessor is discussed. The whole solution has been field-tested and measurements proved that it is highly efficient and very suitable for large quantity deployment.

Keywords: phase measurement, zero-crossing, phase jitter, synchronization, Goertzel algorithm, automatic meter reading, automatic meter management

1. Introduction

Improvement of the processing power of different smart devices enables their connection to the network, thus opening an opportunity for new value added services. A group of devices that will dramatically increase the number of communicating units in networks are also different kinds of meters (electricity, water, gas, heat, etc). Increasing energy prices, carbon emissions, electricity and water supply constraints on the one hand and liberalization of the energy markets [9–11] and value added services on the other hand increase the importance of AMR (automatic meter reading) systems and their development to AMM (automatic meter management) systems. The basic element of such a system is an intelligent electricity system meter that collects information from other meters (for example, the gas meter, the water meter, etc), communicates with other equipment on this level (i.e. the customer display unit, smart house home appliances [7, 8],

etc), drives the customer loads and even is able to disconnect the customer from the mains. At the same time it can gather customer's alarms and information regarding the status of an electricity distribution network. The information is then sent from the meter through different communication channels (i.e. PSTN—public switched telephone network, ISDN—integrated services digital network, ADSL—asymmetric digital subscriber line, cable, GSM—global system for mobile communications, GPRS—general packet radio service, DLC—distribution line carrier, etc) to the upper layers of the system structure. A typical AMR/AMM system is shown in figure 1.

The choice of the communication channel highly depends on the owner of the AMR/AMM system and its business model. For obvious reasons, the preferred communication channel for DGOs—distribution grid operators—is DLC. An original and robust implementation of DLC modem synchronization is presented in this paper.

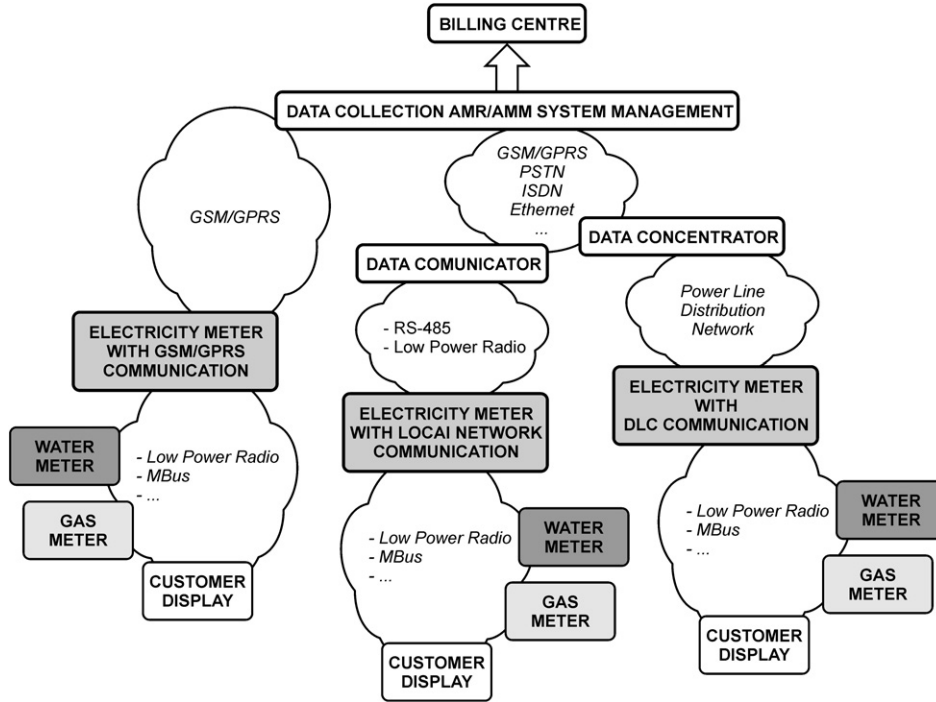


Figure 1. Typical AMR/AMM system [23].

1.1. Communications over power line

The power line mains are optimized for the distribution of electrical power and as such they present a harsh communication environment [1, 2, 5, 6]. Standard [12] regulates frequency bands, maximum transmission levels and other general requirements concerning communications over the power line network. Different modulation schemes can be used. Standardized solutions are based on FSK (frequency shift keying) and SFSK (spread frequency shift keying) modulation [13, 14]. The SFSK modulation shows performance advantages over the FSK modulation [3]. Spread spectrum modulations are also preferred by the standard [12]. Standard [14] describes a client-server communication topology, where the client (data frame that includes the preamble) is synchronized to the zero crossing of the mains frequency (50 Hz or 60 Hz). The mains frequency is not a high precision reference signal [15], but synchronization errors due to propagation can be neglected over a distance of a few kilometres [2]. The server synchronizes itself to the preamble of the frame from the client. At the same time it measures the time (t_s) between the last zero crossing and the start of the preamble of the received frame. From this value the phase of the server regarding the phase of the client can be defined. After the synchronization, the server is locked to the client and it uses the value t_s for receiving and sending frames. With the described mechanism the client and servers stay synchronized regardless of the change of frequency of the reference signal.

1.2. Synchronization and power line noise

The zero crossing of the reference signal is detected by a zero crossing detector. The problem that significantly affects successful synchronization of the described data

communication system is the noise superposed onto the reference signal. Zero crossing detectors that just detect the sign of the observed signal [2] are not efficient because they produce many false triggers near the 'crossing-point'. Adding an analogue filter that filters out higher frequency components in front of such zero-crossing detectors improves the number of false detections, but on the other hand it introduces a phase shift that is not constant. It depends on the values and tolerances of the elements that it is built from. It also increases the element count and overall complexity of the system. The solution against white noise would be a detector with an appropriate hysteresis, but also in that case an additional changeable phase shift is introduced and has to be compensated. As such a detector, the Schmitt trigger inverter could be used. From [16, 17] it can be seen that the hysteresis of the Schmitt trigger inverter in the worst case could be as low as 0.4 V (at 3.0 V voltage supply), the positive going threshold voltage values are from 1.5 V to 2.2 V and the negative going threshold voltage values are from 0.6 V to 1.35 V. The voltage slope of the sinus signal around the zero crossing is

$$\frac{dv}{dt} \approx A \cdot 2\pi \cdot f \quad (1)$$

where A is the amplitude and f is the frequency of the signal. If the zero crossing detector is triggered by a non-attenuated signal (for example, if the signal is filtered with a low-pass RC filter followed by a simple voltage limiter), the jitter of the threshold uncertainty could be neglected, but even the noise with a signal-to-noise ratio higher than 55 dB would cause false triggering. If the zero crossing detector is triggered by an attenuated signal (if the RC filter's cut-off frequency is appropriately low), the jitter made by the threshold uncertainty would be in a range of hundreds of microseconds. On the other hand, the presence of the disturbance sinus signal with an amplitude of a few volts with the frequency in the range from

100 Hz to 500 Hz and a changeable phase [1] (as in the case of ripple control signals [22], ripple control is used to control consumers in the power grid by switching on/off different loads, e.g., storage heating, direct controlled heating, etc, as well as tariff switching, monthly demand reset in electricity meters, and the control of other consumers like street lighting, etc; for control signals ASK modulation is used; the frequency of the signals is from 110 Hz to 2 kHz) can significantly shift the zero crossing of the fundamental signal. So in the case of high noise or single frequency distortions, the detector with hysteresis fails to operate properly.

To minimize the phase jitter of the zero-crossing detector and to eliminate the described problems, and, at the same time, to maintain simplicity, a new method of detecting zero-crossing utilizing the phase measurement has been developed in this paper.

2. Signal phase measurement

The signal phase could be evaluated using DFT (discrete Fourier transform) or one of its more efficient derivatives (FFT—fast Fourier transform, chirp z-transform, etc). A very efficient algorithm [18, 19] that enables numerical evaluation of the magnitude and phase of a single spectral line, as it would be calculated (for a particular target frequency f) by performing complete FFT, was published by Goertzel [20]. A single spectral line could also be calculated utilizing direct calculation of DFT, but the main advantage of the Goertzel algorithm is that fewer multiplications and additions are required. Also the number of pre-calculated coefficients needed in these mathematical operations is, in the case of the Goertzel algorithm, reduced to just 2 per spectral line (other coefficients are implicitly computed by the recursive algorithm). It is also not necessary that the number of samples is of the power of 2 (as in the case of FFT). As long as the number of frequencies of interest is lower than $\log_2 N$ (where N is the number of samples), the Goertzel algorithm is the most efficient method [18]. The real ($\text{Re } X$) and imaginary ($\text{Im } X$) frequency components can be evaluated (as with regular or FFT) after gathering a block of N samples (a_i):

$$\text{Re } X = Q_1 \cos(\omega) - Q_2 \quad (2)$$

$$\text{Im } X = Q_1 \sin(\omega) \quad (3)$$

where Q_i is defined as

$$Q_i = 2 \cos(\omega) \cdot Q_{i+1} - Q_{i+2} + a_i \quad \text{for } i = N, N-1, \dots, 1 \quad (4)$$

$$Q_{N+2} = Q_{N+1} = 0 \quad (5)$$

and k , ω , $\cos(\omega)$ and $\sin(\omega)$ are constants determined by equations (6) and (7):

$$k = N \frac{f}{f_{\text{sampling}}} \quad (6)$$

$$\omega = \frac{2\pi}{N} k. \quad (7)$$

Here f_{sampling} is the sampling frequency of the A/D converter and k is a constant. To achieve the best result the target frequency f should be centred in its respective bin, in which

case k would be an integer constant. From (2) and (3) the signal phase of interest can be calculated:

$$\varphi = \arctan\left(\frac{\text{Im } X}{\text{Re } X}\right). \quad (8)$$

The arctan function can then be calculated using Maclaurin series [24],

$$\arctan(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{2n+1} = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \dots, \quad (9)$$

or can be calculated in advance and stored in a table.

3. Practical exploitation of the Goertzel algorithm

As shown in the previous section, the first step in determining the phase of the signal with frequency f is to define sampling frequency (f_{sampling}) and sample block size (N). In practice, sampling frequency is often predefined by the clock of the A/D converter or microprocessor used. k in equation (6) should be an integer and sample block size (N) should be selected to satisfy that condition. With k and N being known constants ω , $\cos(\omega)$ and $\sin(\omega)$ are also determined. Using equations (4), (2) and (3) on a known block of samples ($a_N, a_{N-1}, \dots, a_1$), the real and imaginary components are evaluated. The ratio between $\text{Im } X$ and $\text{Re } X$ is defined as

$$R(X) = \frac{\text{Im } X}{\text{Re } X}. \quad (10)$$

Finally, the signal phase can be determined by equation (8). It is not very efficient to calculate the arctan using (9) on a processor with limited resources. Using table values is preferred. The results (2), (3) and (10) correspond to the signal phase at the point when the last sample from the block was gathered.

With regard to the microprocessor's 32-bit number limitation for precise phase shift evaluation (table 1) the ratio $R(X)$ should not be around 1 (one of the components, real or imaginary, should be much greater than the other). So, the most precise estimate of the phase can be made if the last sample (a_1) of the signal was sampled at the point when the signal phase was around:

$$\varphi = j \frac{\pi}{2} \quad \text{for } j = 0, 1, \dots \quad (11)$$

The period of sampling the whole block can be smaller than the period of the target signal. If the process of signal phase determination is continuous, the beginning of a sampling block can be shifted along the time axis of the periodic signal to gather the last sample of block N at points defined by (11) and to gain the most precise estimate of the signal phase. When the first block is sampled, there is no reference for block sampling start, and the first phase estimate is far from synchronization. The block sampling start is then positioned relatively to that first phase estimate. The new phase estimate is calculated and the process is continued. The phase estimate rapidly converges to a minimal phase shift and the system is synchronized.

But if the frequency of the signal is changing fast, the last sample should be taken just before the decision, the synchronization point, to get most actual data (depending on the time needed for final calculations). A typical example

Table 1. Tabulated function arctan. The first and last columns are the same as used in the microprocessor.

Timer frequency = 14 745 600 Hz Target frequency = 50 Hz Number of ticks per period = 294 912					
Table index	R	Arctan(R)	Arctan(R) (deg)	delta = arctan(R) (deg)/360	Delta (in ticks) (rounded values)
39	1.911×10^{-6}	1.91×10^{-6}	0.000 109 283	3.04×10^{-7}	0
38	3.81×10^{-6}	3.81×10^{-6}	0.000 218 566	6.07×10^{-7}	0
37	7.63×10^{-6}	7.63×10^{-6}	0.000 437 132	1.21×10^{-6}	0
36	1.53×10^{-5}	1.53×10^{-5}	0.000 874 264	2.43×10^{-6}	1
35	3.05×10^{-5}	3.05×10^{-5}	0.001 748 528	4.86×10^{-6}	1
34	6.10×10^{-5}	6.10×10^{-5}	0.003 497 057	9.71×10^{-6}	3
33	0.000 122 07	0.000 122 07	0.006 994 114	1.94×10^{-5}	6
32	0.000 244 141	0.000 244 141	0.013 988 227	3.89×10^{-5}	11
31	0.000 488 281	0.000 488 281	0.027 976 453	7.77×10^{-5}	23
30	0.000 976 563	0.000 976 562	0.055 952 892	0.000 155 425	46
29	0.001 953 125	0.001 953 123	0.111 905 677	0.000 310 849	92
28	0.003 906 25	0.003 906 23	0.223 810 5	0.000 621 696	183
27	0.007 812 5	0.007 812 341	0.447 614 171	0.001 243 373	367
26	0.015 625	0.015 623 729	0.895 173 71	0.002 486 594	733
25	0.031 25	0.031 239 833	1.789 910 608	0.004 971 974	1 466
24	0.062 5	0.062 418 81	3.576 334 375	0.009 934 262	2 930
23	0.125	0.124 354 995	7.125 016 349	0.019 791 712	5 837
22	0.25	0.244 978 663	14.036 243 47	0.038 989 565	11 498
21	0.5	0.463 647 609	26.565 051 18	0.073 791 809	21 762
20	1	0.785 398 163	45	0.125	36 864
0	1	0.785 398 163	45	0.125	36 864
1	2	1.107 148 718	63.434 948 82	0.176 208 191	51 966
2	4	1.325 817 664	75.963 756 53	0.211 010 435	62 230
3	8	1.446 441 332	82.874 983 65	0.230 208 288	67 891
4	16	1.508 377 517	86.423 665 63	0.240 065 738	70 798
5	32	1.539 556 493	88.210 089 39	0.245 028 026	72 262
6	64	1.555 172 598	89.104 826 29	0.247 513 406	72 995
7	128	1.562 983 986	89.552 385 83	0.248 756 627	73 361
8	256	1.566 890 097	89.776 189 5	0.249 378 304	73 545
9	512	1.568 843 204	89.888 094 32	0.249 689 151	73 636
10	1024	1.569 819 765	89.944 047 11	0.249 844 575	73 682
11	2048	1.570 308 046	89.972 023 55	0.249 922 288	73 705
12	4096	1.570 552 186	89.986 011 77	0.249 961 144	73 717
13	8192	1.570 674 256	89.993 005 89	0.249 980 572	73 722
14	16 384	1.570 735 292	89.996 502 94	0.249 990 286	73 725
15	32 768	1.570 765 809	89.998 251 47	0.249 995 143	73 727
16	65 536	1.570 781 068	89.999 125 74	0.249 997 571	73 727
17	131 072	1.570 788 697	89.999 562 87	0.249 998 786	73 728
18	262 144	1.570 792 512	89.999 781 43	0.249 999 393	73 728
19	524 288	1.570 794 419	89.999 890 72	0.249 999 696	73 728

The first and last columns of the table are actually used in the microprocessor. The arctan(R) function from this table returns values from 0 to 90°. The final calculation of the phase is done after determining of the quadrant (from signs of Real X and Imag X).

of fast signal frequency changing is the mains voltage. The instants of the zero crossing of the mains voltage can deviate from nominal value due to interference, e.g., large loads switch on/off.

For power line frequency with 50 Hz nominal value standard [15] defines the maximal allowed deviation from 47 Hz to 52 Hz. Measurement showed mains voltage period deviations of more than 100 μ s inside a few period cycles.

4. Realization of the synchronization on a low-cost microprocessor

A practical realization of zero crossing detection utilizing the Goertzel algorithm was implemented on a low-cost 32-bit ARM microprocessor [21]. The development prototype is shown in figure 2.

The signal was first filtered with a simple RC low-pass filter with cut-off frequency 240 Hz. The worst case analysis (done by the authors using PSpice) showed that the tolerance of elements (1% for R and 5% for C) of the RC filter can cause a jitter of around 10 μ s. The voltage of the signal was limited to the value acceptable for the A/D converter. A 10-bit A/D converter on the ARM microprocessor with a sampling frequency of 1 kHz was used for sampling. The sampling was triggered by software. The constants $\cos(\omega)$ and $\sin(\omega)$ were pre-calculated and stored in a non-volatile memory. After every sample variable Q , equation (3), was calculated. The first calculation (after the first sample in the block was gathered) was performed considering initial conditions (5). The calculations of the real and imaginary components of the signal were executed every 20th sample (the size of the block for the Goertzel algorithm was 20,

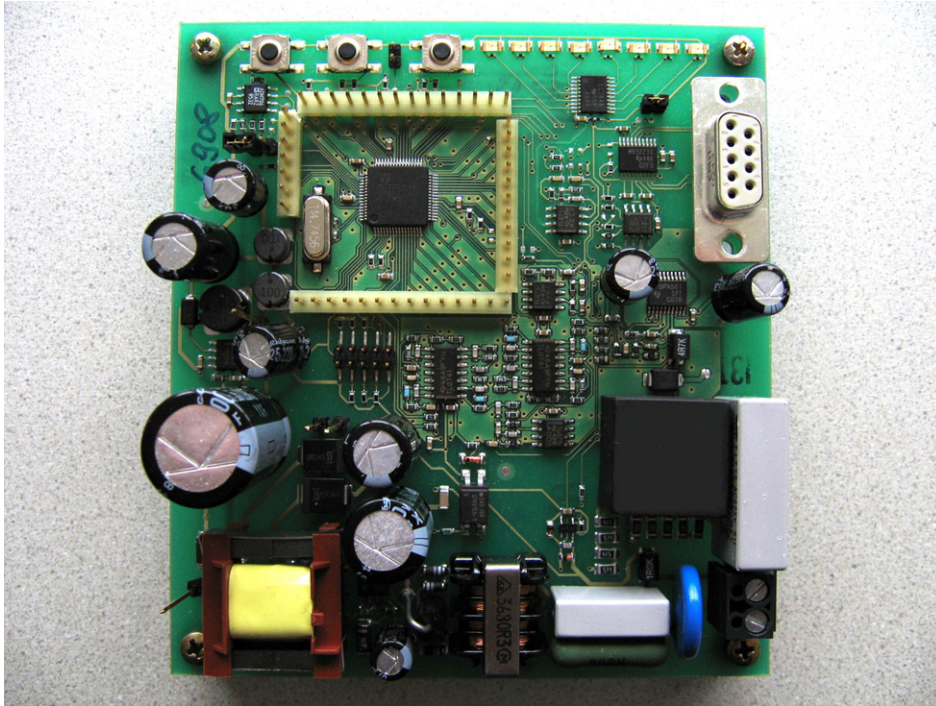


Figure 2. The development prototype with a 32-bit ARM microcontroller.
(This figure is in colour only in the electronic version)

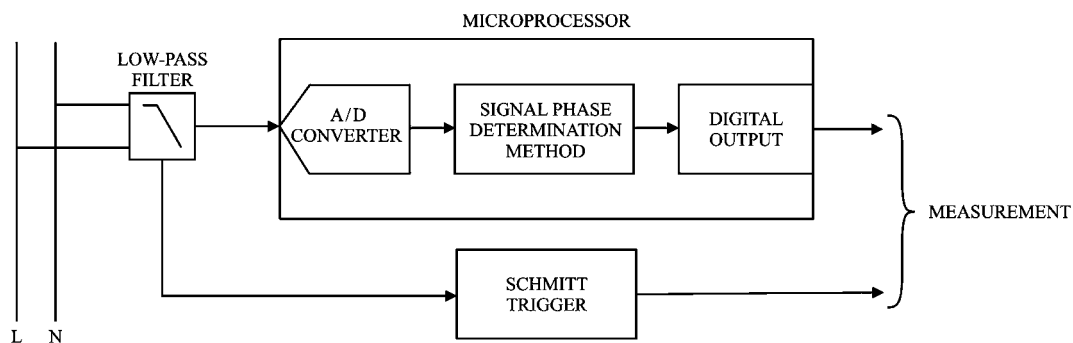


Figure 3. Block diagram of the measurement setup.

the constant k was an integer). After determining the ratio between the imaginary and real components, the arctan value from table 1 was obtained. To keep the process of determining this ratio as simple as possible, only values of the power of two were used. To obtain the final estimate of the signal phase at the point where the last sample in the block was gathered, the actual quadrant was determined in accordance with signs of the imaginary and real components, and the phase was calculated. To verify the successfulness of the method in the real environment (utilizing the oscilloscope), the digital output transition (pin toggle) is set corresponding to the calculated zero crossing. A block diagram of the measurement setup is shown in figure 3. The processor's internal clock was 60 MHz. The clock for the A/D converter was sourced from a 14.7456 MHz oscillator. The frequency tolerance of the crystal was 30 ppm, the temperature tolerance over an extended temperature range (-40°C to $+85^{\circ}\text{C}$) was 50 ppm and the aging effect was less than ± 5 ppm for the first year. In the worst case the effect of

these tolerances on the sampling jitter was in the range of 1–10 microseconds and therefore negligible. One channel of the 10-bit A/D converter, one 32-bit timer with match interrupt and less than 0.75% of CPU time (internal clock was 60 MHz) were needed for signal phase estimation.

4.1. Pre-evaluation using simulations

During the implementation of the real system simulations of such a system were also performed in Matlab[®]. Different kinds of disturbances (single frequency noise, white noise and harmonics) that are common [2, 5, 6] for power lines were applied to the basic 230 VAC_{RMS}, 50 Hz signal to generate an input signal for the system. All characteristics of the real processing system (10 bit A/D, 32-bit word length, sampling frequency, etc) were followed as close as possible. Simulations were performed for five different sets of noise (no noise, two different levels of white noise, single frequency noise, predetermined harmonics). The simulated signal phase was

Table 2. Simulation results.

Phase of the input signal	Maximal absolute error (deg)				
	No disturbances	White noise (SNR = 20 dB)	White noise (SNR = 40 dB)	Single frequency noise ($f = 100$ Hz, $V_{amp} = 330$ V)	Harmonics
−180.0	0.0073	1.7896	0.1123	0.0281	0.448
−179.9	0.0438	1.8896	0.0123	0.0123	0.548
−135.0	18.4351	18.4351	18.4351	18.4351	18.4351
−120.0	3.4351	3.4351	3.4351	3.4351	3.4351
−95.0	1.4233	2.1252	1.4233	1.4233	4.1052
−91.0	0.1052	0.541	0.1052	0.1052	2.7896
−90.0	0.0073	0.2234	0.1123	0.0281	1.7896
−45.0	18.4351	18.4351	18.4351	18.4351	18.4351
−10.0	2.8748	2.8784	2.8748	2.8748	2.8748
−2.0	0.2104	1.5767	0.2104	0.2104	1.1052
−1.0	0.1052	1.8948	0.1052	0.1052	0.8877
−0.5	0.0520	0.7234	0.2766	0.052	0.5281
−0.1	0.0438	1.8896	0.1234	0.0438	0.548
0.0	0.0073	1.7896	0.1123	0.0281	0.448
0.1	0.0438	1.6896	0.2123	0.0123	0.0123
0.2	0.0877	1.0948	0.2288	0.0234	0.0234
0.5	0.0520	1.2896	0.052	0.052	0.3948
1.0	0.1052	2.7896	0.1052	0.1052	0.7896
5.0	1.4233	3.2104	1.4233	1.4233	1.4233
10.0	2.8747	2.8748	2.8748	2.8748	2.8748
45.0	18.4531	18.4351	18.4351	18.4351	18.4351
89.9	0.0438	1.6896	0.1234	0.0123	7.2252
90.0	0.0073	0.8948	0.1123	0.0073	7.1252
90.1	0.0123	0.7948	0.2034	0.0438	7.0251
135.0	0.4351	18.4351	18.4351	18.4351	18.4351
179.9	0.0438	1.8896	0.2123	0.0438	0.348
180.0	0.0073	1.7896	0.1123	0.0281	0.448

The results of simulations have been gathered in this table. In the first column, the phase of the useful signal is shown. The second column shows the simulation of maximal absolute error when no noise was added to the useful signal. In the third and fourth columns there are results of simulations when different levels of white noise were added to the useful signal. The next column shows maximal absolute error when a signal with high amplitude was added as noise. The last column presents simulation results in the presence of harmonics.

changed from -180° to 180° . The results of simulations are shown in table 2. In the second column are given the results of simulations with an input signal with no distortion. It can be seen that if the phase of the input signal was around values defined by (11), the error made by the phase estimation process was low. For example, in phases -180° , -90° , 0° , 90° and 180° , the phase estimation differed from the actual phase by only $\pm 0.0073^\circ$ (approx. $\pm 2.03 \times 10^{-7}\%$ or $0.5 \mu s$). In this case, one component (real or imaginary) was 8192 times bigger than the other one. If the signal phase was around

$$\varphi = m \frac{\pi}{4} \quad \text{for } m = 1, 3, 5, \dots, \quad (12)$$

the error of the phase estimation process was high. The reason is that the absolute values of the real and imaginary components are almost the same.

The second simulation (column 3 of table 2) was done by adding white noise with SNR = 20 dB. In this case, the phase estimation error in phases defined by (11) was increased to maximal $\pm 1.7896^\circ$ (approx. $\pm 0.5\%$ or $\pm 100 \mu s$).

The third simulation was similar to the second one except that, in this case, the SNR was 40 dB. In this case, the maximal phase estimation error in phases defined by (11) was $\pm 0.1123^\circ$ ($= \pm 3.12 \times 10^{-6}\%$ or $\pm 6 \mu s$).

In the fourth simulation (column 5 of table 2), an additional sinus signal was added to the useful signal. The amplitude of this signal was 330 V and the frequency was 100 Hz. The maximal phase estimation error in phases defined by (10) was $\pm 0.0281^\circ$ ($\pm 8 \times 10^{-7}\%$ or $1.6 \mu s$).

The last simulation was done with the signal and its harmonics. The definitions of the harmonics are in table 3. The values were obtained with experimental field measurements done by Iskraemeco d.d. during the development of the ripple control receiver and different DLC modems.

The fourth and fifth simulations, single frequency noise and harmonics, also included phase change between the useful signal and noise signals, since noise signals always had the same phase (except that the phases of harmonics signals were random from 0° to 60°), but the phase of the useful signal was in the range from -180° to $+180^\circ$. That is also the reason why there are differences in error of phase estimation in phases defined by (11) in the case of harmonic distortion. The maximal phase estimation error if the phase was 0° , 180° or -180° was $\pm 0.448^\circ$ ($\pm 0.13\%$ or $\pm 25 \mu s$). The fourth simulation does not reflect the real possible conditions (the maximum real amplitudes of the harmonics are in the range of 10%—around 30 V—of the fundamental frequency), but it shows the resistance of the method to such interference.

Table 3. Harmonics definition for simulation.

Basic signal frequency $f_0 = 50$ Hz Basic signal amplitude $A_0 = 324$ V		
Harmonic frequency	Harmonic amplitude	Harmonic phase
$2f_0 = 100$ Hz	2% of A_0	0–60° (random)
$3f_0 = 150$ Hz	7% of A_0	0–60° (random)
$4f_0 = 200$ Hz	1.5% of A_0	0–60° (random)
$5f_0 = 250$ Hz	8% of A_0	0–60° (random)
$6f_0 = 300$ Hz	1% of A_0	0–60° (random)
$7f_0 = 350$ Hz	7% of A_0	0–60° (random)
$8f_0 = 400$ Hz	0.8% of A_0	0–60° (random)
$9f_0 = 450$ Hz	1.2% of A_0	0–60° (random)
$10f_0 = 500$ Hz	0.7% of A_0	0–60° (random)
$11f_0 = 550$ Hz	5% of A_0	0–60° (random)
$13f_0 = 650$ Hz	5% of A_0	0–60° (random)
$15f_0 = 750$ Hz	0.5% of A_0	0–60° (random)
$17f_0 = 850$ Hz	2% of A_0	0–60° (random)
$19f_0 = 950$ Hz	2% of A_0	0–60° (random)
$23f_0 = 1150$ Hz	1.5% of A_0	0–60° (random)
$25f_0 = 1250$ Hz	1.5% of A_0	0–60° (random)
$29f_0 = 1450$ Hz	0.8% of A_0	0–60° (random)
$31f_0 = 1550$ Hz	0.8% of A_0	0–60° (random)
$35f_0 = 1750$ Hz	0.7% of A_0	0–60° (random)
$37f_0 = 1850$ Hz	0.7% of A_0	0–60° (random)

In one part of simulations, harmonic distortion of the useful signal was also simulated. Frequencies, amplitudes and phases of added harmonic components are defined in this table.

During all simulations the amplitude change of the useful signal in the range $\pm 10\%$ was also simulated and the results showed no effect on phase estimation.

4.2. Measurement setups and results

The measurements were performed simultaneously comparing the successfulness of the zero crossing detector based on the phase measurement (ZC-PM) and the zero crossing detector based on the Schmitt trigger (ZC-ST) and included three different measurement setups. The first one connected zero-crossing detectors to the real mains where relatively high white noise but no continuous single frequency or harmonic components were present. The second measurement setup was similar to the first one except that in this case additionally the ripple control transmitter was connected to the real mains and its control signal amplitude was set to 5 V and the frequency to 130 Hz. In the third measurement the ripple control signal was substituted by 10% of the third harmonic with a 30° phase.

All measurements were performed on a set of two pairs of ZC-PM and ZC-ST with a common RC filter. We measured the jitter or delta delay (measurements performed on a LeCroy oscilloscope WaveRunner 6100A with built-in statistical functions—mean, max, min, standard deviation) between the rising edges of the zero crossing signals of the ZC-PM detectors (second columns of tables 4 and 5) and of the ZC-ST detectors (fourth columns of tables 4 and 5). At the same time the widths of the pulse outputs from the detectors (third and fifth columns of tables 4 and 5) were measured with the same oscilloscope. Normal widths should be half of the mains frequency period—10 ms. The delta delay or jitter is important

for mass deployment and it shows the dependence of the zero crossing detection in regard to element tolerances. The width of the pulse is the base for synchronization and its stability is in direct relation to the successfulness of the communication. For the power line conditions and disturbances indication also a mains frequency (sixth columns of tables 4 and 5) was measured. All results are based on a population of 5000 measurements.

The results of the first measurement (table 4) show that the performances of the ZC-PM and ZC-ST detectors in a white noise environment are very similar.

The delay between ZC-PM detectors was around ± 40 μ s, with practically zero mean value and standard deviation around 6 μ s. On the other hand, the delay between the ZC-ST detectors was from -2 μ s to -41 μ s with a mean value of -15 μ s (these values depend on the characteristic tolerances of the used Schmitt triggers if the ZC-ST detectors were swapped, the mean delay value was 15 μ s and the delay was between 2 μ s and 41 μ s). It can also be noted that the pulse width stability of the ZC-PM detector was about ten times better than that of the ZC-ST detector (the standard deviation of the delay between the ZC-PM detectors was around 2.5 μ s and between the ZC-ST detectors was more than 25 μ s). The standard deviation of the mains frequency was 30 mHz.

A graphical presentation of the measured results from table 4 is shown in figure 4. In the upper graph of figure 4 the delay measurements are presented. The swapped case is shown in dotted outline. The performances of both zero-crossing detectors regarding delay measurement were practically identical.

In the lower graph pulse width measurements are presented. The ZC-PM detectors showed much higher stability than the ZC-ST detectors.

In the second measurement (table 5), in the presence of the disturbing ripple control signal, the zero crossing point of the mains signal was shifted (regarding the phase of the disturbing signal at the zero crossing).

The standard deviation of the measured mains frequency was almost 0.5 Hz. The ZC-PM detector showed no deterioration regarding the first measurement, but the performance of the ZC-ST detector decreased significantly. The delay between the ZC-ST detectors remained the same, but the standard deviation of pulse width rose to more than 130 μ s.

A graphical presentation of the measurement results from table 5 is shown in figure 5. It can be seen that in the case of additional noise (ripple control signal disturbance) the performance of the ZC-ST detectors became worse. The delay measurements showed just a slight deterioration, but the pulse width measurements showed significant performance decrease. The performance of the ZC-PM detectors remained similar to that in the first case.

In the presence of the higher harmonic component the zero crossing point was shifted, but the pulse width remained the same (the standard deviation of mains frequency was 30 mHz)—the measured results of both kinds of detectors were very similar to the results from the first measurement setup.

The performance of the ZC-ST detectors is highly dependent on the characteristic parameters of the Schmitt

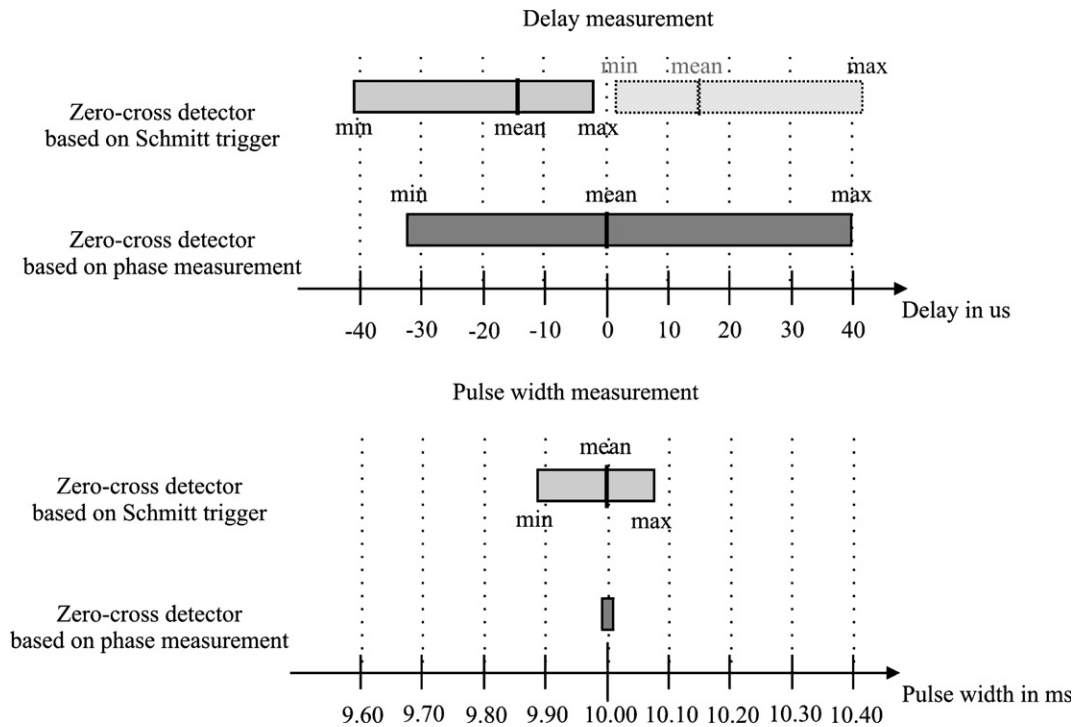


Figure 4. Zero-crossing detection comparison—no additional noise.

Table 4. Measurements results—no additional interference.

	ZC-PM detector (delay)	ZC-PM detector (pulse width)	ZC-ST detector (delay)	ZC-ST detector (pulse width)	Mains frequency (50 Hz)
Mean	362.900 ns	10.003 ms	-14.797 μ s	9.971 ms	49.984 Hz
Min	-34.000 μ s	9.994 ms	-40.970 μ s	9.872 ms	49.865 Hz
Max	39.980 μ s	10.013 ms	-2.000 μ s	10.076 ms	50.133 Hz
SD	6.279 μ s	2.424 μ s	5.721 μ s	25.829 μ s	32.347 mHz

The measurements were performed comparing the successfulness of the zero crossing detector based on the phase measurement (ZC-PM) and the zero crossing detector based on the Schmitt trigger (ZC-ST) connected to the mains where relatively high white noise but no continuous single frequency or harmonic components were measured. In the second, fourth and sixth columns the jitter or delta delay (also statistically – mean, max, min, standard deviation) between the rising edges of the zero crossing signal of ZC-PM detectors, of ZC-ST detectors and of the mains frequency are shown respectively. In the third and fifth columns the widths of the pulses of ZC-PM detectors and ZC-ST detectors are shown respectively.

Table 5. Measurements results—ripple control signal added.

	ZC-PM detector (delay)	ZC-PM detector (pulse width)	ZC-ST detector (delay)	ZC-ST detector (pulse width)	Mains frequency (50 Hz)
Mean	436.4 ns	10.002 ms	-17.204 μ s	9.970 ms	49.997 Hz
Min	-31.02 μ s	9.994 ms	-45.030 μ s	9.678 ms	49.383 Hz
Max	38.03 μ s	10.012 ms	-1.94 μ s	10.200 ms	50.545 Hz
Std. Deviation	6.979 μ s	2.850 μ s	6.8945 μ s	130.424 μ s	487.723 mHz

The measurements were performed comparing the successfulness of the zero crossing detector based on the phase measurement (ZC-PM) and the zero crossing detector based on the Schmitt trigger (ZC-ST) connected to the mains with added single frequency noise (in this case ripple control signal with 5 V amplitude and frequency 130 Hz).

trigger used. The ZC-PM detectors showed high repeatability. They perform similarly (slightly better) in a white noise

environment and are far more efficient in single signal distortions with changeable phase than ZC-ST.

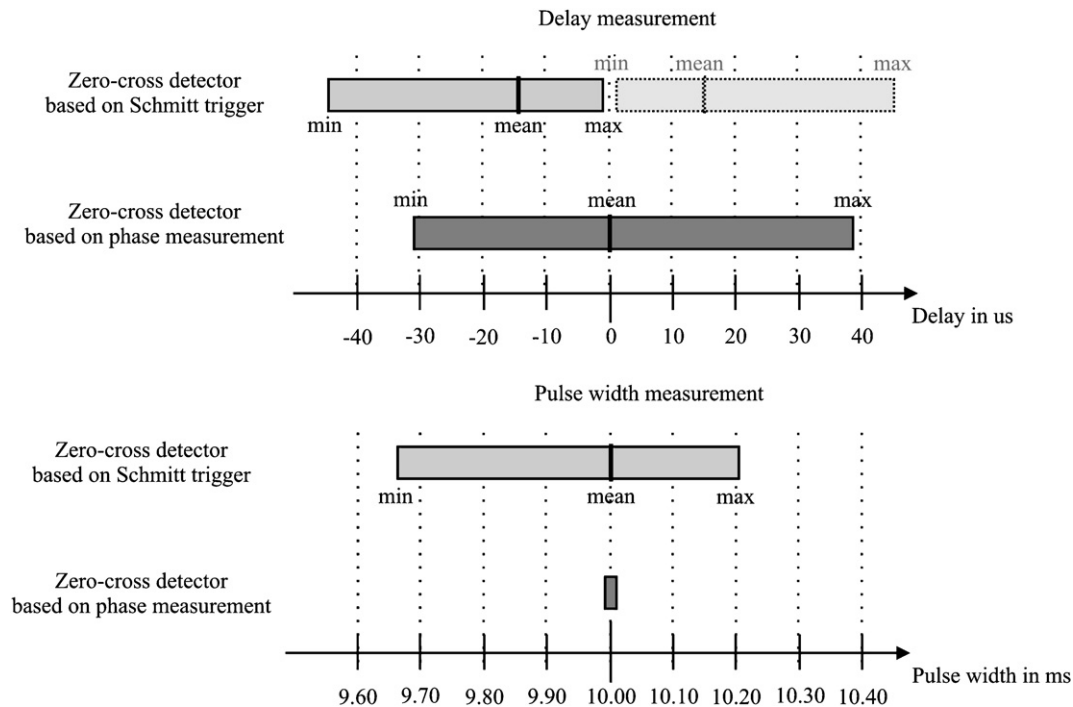


Figure 5. Zero-crossing detection comparison—ripple control signal added.

5. Conclusion

AMR/AMM systems are growing in importance. The main reasons for that are restricted energy and water resources and their management, ecology problems, smart house services and comfort.

In this paper the realization of DLC modem synchronization utilizing phase measurement is discussed. By utilizing the Goertzel algorithm it is possible to implement simple, process effective phase measurement of the observed signal and use that information for synchronization of the system to the zero crossing of the signal. It is possible to synchronize it to any phase selected. The phase estimation algorithm takes less than one period of observed signal, so phase estimation can be performed at every period (every 20 ms).

This method shows similar efficiency in White noise conditions and is far more efficient in single signal distortions with changeable phase than other methods (for example zero-crossing detector with hysteresis). It does not employ any special hardware, but just the available peripherals on the microprocessor and is, due to stable parameters, very suitable for the production of large quantities.

This solution increases DLC communication robustness in environments where different communication solutions are present (different DLC systems, ripple control system, etc) and has also influence on the communication speed. In some cases, an anti-aliasing filter will be necessary in front of the A/D converter. However, the measured frequency and sampling frequency (also cut-off frequency) can be set far enough in order not to add any unpredictable, changeable phase error due to filter element tolerance. Predictable phase error can be compensated in software.

The same principle can also be used in many different applications, using information about the signal phase.

The results were field-tested in test pilots at Iskraemeco. The phase measurement method of synchronization proved to have the right combination of complexity, accuracy, speed, simplicity and cost. This robust power line modem synchronization solution enabled further work on DLC modem speed improvements which is now in progress.

References

- [1] Zhang X H *et al* 2006 The design of virtual low-voltage power line noise generator *J. Phys.: Conf. Ser.* **48** 686–90
- [2] Dostert K 2001 *Power-line Communications* (New York: Prentice-Hall)
- [3] Schaub T 1994 Spread frequency shift keying *IEEE Trans. Commun.* **42** Feb/Mar/Apr 1056–64
- [4] NordGreen E 2006 Vattenfall Eldistribution AB, Sweden Project AMR – Automatic Meter Reading *Metering Europe 2006* (Copenhagen, Denmark, 10–12 Oct.)
- [5] Hooijen O G 1998 *Aspects of Residential Power Line Communications* (Aachen: Shaker)
- [6] O'Neal J B Jr 1986 The residential power circuit as a communication medium *IEEE Trans. Consum. Electron.* **32** 567–77
- [7] Brackmann L 1997 Powerline applications with European Home System (EHS) *ISPLC* (Essen, Germany, 2–4 April)
- [8] Konnex, <http://www.konnex.org>
- [9] DIRECTIVE 2001/77/EC on the Promotion of Electricity Produced from Renewable Energy Sources in the Internal Electricity Market (Brussels, 27 Sept.)
- [10] DIRECTIVE 2003/54/EC Concerning Common Rules for the Internal Market in Electricity (Brussels, 26 June)
- [11] DIRECTIVE 2006/32/EC on Energy End-use Efficiency and Energy Services (Brussels, 5 April)

- [12] EN 50065-1:2001 Signalling on low-voltage electrical installations in the frequency range 3 kHz to 148.5 kHz—Part 1: General requirements, frequency bands and electromagnetic disturbances
- [13] BS IEC TR 61334-5-2:1998-12-15, Distribution automation using distribution line carrier systems—Lower layer profiles—Frequency shift keying (FSK) profile
- [14] EN 61334-5-1:2001 Distribution automation using distribution line carrier systems—Part 5-1: Lower layer profiles—The spread frequency shift keying (S-FSK) profile
- [15] European standard EN50160:1999 Voltage characteristics of electricity supplied by public distribution systems
- [16] Fairchild Semiconductors January 1996, rev. August 2004 NC7S14 TinyLogic HS Inverter with Schmitt Trigger Input www.fairchild.com
- [17] Texas Instruments April 1999, rev. September 2005 SN74 LVC1G14 Single Schmitt-Trigger Inverter www.ti.com
- [18] Oppenheim A V, Schaffer R W and Buck J R 1998 *Discrete-Time Signal Processing* (Englewood Cliffs, NJ: Prentice-Hall)
- [19] Sirin S 2004 A DSP algorithm for frequency analysis www.embedded.com
- [20] Goertzel G Jan 1958 An algorithm for the evaluation of finite trigonometric series *Am. Math. Mon.* **65** 34–35
- [21] NXP 24 June 2005 LPC213x User Manual rev.01, www.nxp.com
- [22] Elster Metering, http://www.elstermetering.com/en/load_management_ripple_control.shtml
- [23] Iskraemeco Homepage, www.iskraemeco.si
- [24] <http://mathworld.wolfram.com/InverseTangent.html>

Optimisation of combined collaborative recommender systems

Matevž Kunaver*, Tomaž Požrl, Matevž Pogačnik, Jurij Tasič

Faculty of Electrical Engineering, Tržaška 25, SI-1000 Ljubljana, Slovenia

Received 8 August 2006; accepted 3 April 2007

Abstract

A new approach to collaborative user modelling is presented in this paper. We have developed a framework that can be used for easy testing of different concepts. We have also introduced three different areas where collaborative modelling can be further improved. For the first phase of testing, we have created a hybrid system based on three different collaborative recommender techniques. Since this system implements multiple collaboration techniques, we decided to call this approach Combined Collaborative Recommender. Although each prediction technique can produce adequate results, we have proved that the combination of these techniques into a unified system provides a much more stable system. It should also be pointed out that these analyses were done using a very large dataset (more than 2 million ratings) providing reliable results. Results of these optimisations are presented along with pointers for further development.

© 2007 Elsevier GmbH. All rights reserved.

Keywords: User modelling; Personalisation; Collaborative recommendation; Hybrid recommender systems

1. Introduction

One of the most common problems encountered by users, when using the internet or other information sources to search for multimedia content, is manipulating a sheer volume of available information. If we concentrate on the field of video distribution, for example, we find that the amount of available video content items is almost limitless and when users have to choose among thousands of items they might easily get lost, bored or even reluctant to use such systems.

The solution to this problem is using a recommender system that is able to automatically generate appropriate information and filter out items that are less important for the user. But such a system presents additional requirements, because it must be able to “learn” the users’ likes or dislikes and act accordingly. This process is known as user modelling. The basic version of user modelling consists of using

the user’s feedback (the ratings the user gave to accessed items) in order to learn what type of content the user prefers and is more likely to accept, if suggested by the system.

We have developed a framework that can be used for testing of different collaborative algorithms for user modelling. The framework runs in Java programming environment. The framework consists of modules enabling a simple user interface and data retrieval from existing SQL databases. Algorithms are implemented as modules that are integrated into the recommender system. Using this approach, we were able to test different algorithms by simply switching modules or combining modules into a hybrid recommender system. Using a dataset provided by Digital Equipment Corporation [1], we have made several different experiments with the aim to obtain information on how to optimise the working of a collaborative recommender system. While we used some algorithms that are already known in the field of user modelling (such as Pearson’s correlation coefficient and True Bayesian prediction), we also introduced some new approaches in order to compensate for the shortcomings of the original

* Corresponding author.

E-mail address: matevz.kunaver@fe.uni-lj.si (M. Kunaver).

algorithms. Because the database was large enough, we were able to make a relevant analysis of recommender efficiency.

We are also introducing our idea of a combined recommender system that can incorporate all of the tested algorithms into a single recommender system. We examine some of the collaborative-based approaches and describe a few ideas for further development in this area.

2. Related work

The area of user modelling is a widely explored field that is continuously expanding. A number of user modelling approaches have been published so far. When dealing with content filtering, two of the more distinct approaches are content-based recommendation and collaborative-based recommendation. A number of other approaches exist (demographic, utility based, etc.), but they are not so widely used as the two mentioned above.

When using collaborative recommendation, one of the main advantages is that the system does not need to use any content descriptors and is therefore description independent. Because of this, it is also possible to use the system in an environment containing different content types, for example movies, documents and images. This makes the system much more flexible and provides the user with a wider choice of recommended items. Integrating new items into the database is one of the problems of this system because the system processes only those items that have received ratings from users, so until the item has been rated, it is “invisible” to the system.

Content-based recommenders, on the other hand, do not interact with data from other users but instead track the metadata of items that the user has rated. If the user prefers a certain type of genre, this will probably manifest itself in positive ratings for items of this genre. The advantage of such a system is that it is independent of other users and that integration of new items is very simple if the necessary metadata are available. The disadvantage of using this type of recommender is that it can only work with the selected metadata standard and therefore does not include items that do not have the necessary metadata or are of a different type.

Advanced recommender systems combine different collaborative and content-based techniques into hybrid systems that switch between different modules according to circumstances in order to produce better results. There are many different possibilities how to combine these modules [2]. The most interesting examples are weighted, switching or cascade hybrid systems.

When using weighted hybrid system, the recommendations of each module are calculated separately and then combined to produce a single recommendation.

A switching hybrid system works by using only one recommendation module at a time. At the beginning of the process, the system determines which module is best suited for the task and then generates recommendations using this module.

The last hybrid system – cascade – works by running recommendation modules sequentially one after another, the output of the first one being the input for the next one. This is very useful if the first recommendation module produces a very large list of possible items for suggestion.

Many of these approaches come from the domain of personalised selection of video and TV programmes. Most of them are either content based [3,4], collaborative [5] or a combination of both approaches [6,7]. A good overview of their properties is available in [2,8]. Additional information on collaborative recommenders can also be found in [9], while [10] provides interesting information on alternate approaches in this area.

It should also be mentioned that these approaches make recommendations to individual users, while another approach, which has come up recently, makes recommendations of TV programmes for groups of users (viewers) [11]. TV viewing is considered as a group activity and the selection of content items for a group of users is not a trivial task. The authors applied a number of strategies to combine item preference ratings of individual group members in order to make suggestions to the group as a whole. Despite the same goal, creating recommendations for the user, existing approaches differ in many ways. This is also one of the reasons we have decided to test some of them in order to see how they perform in the same environment.

While many other existing approaches also work in other fields of recommendations, not just TV programmes and movies, we have pointed out these approaches because the database we used for prediction generation and evaluation contains movie items and user ratings of these items. We have therefore based our system on these approaches.

3. Combined collaborative approach

The goal of our research was to develop a combined recommendation system by using a number of collaborative recommendation algorithms and also to introduce a few optimisations that can be used to increase the reliability and accuracy of generated predictions.

We have decided to test some of the already published approaches and have created a new recommender system based on those results. The preliminary testing was based on collaborative recommendation techniques. The next phase was further development of a new system by creating additional content-based recommendation modules and integrating them into a hybrid recommender system. Since we were creating a system that implements multiple collaboration techniques, we decided to call this approach “Combined Collaborative Recommender”.

3.1. System design

Recommendation of content items using a collaborative approach is a two-stage process. First, a number of nearest

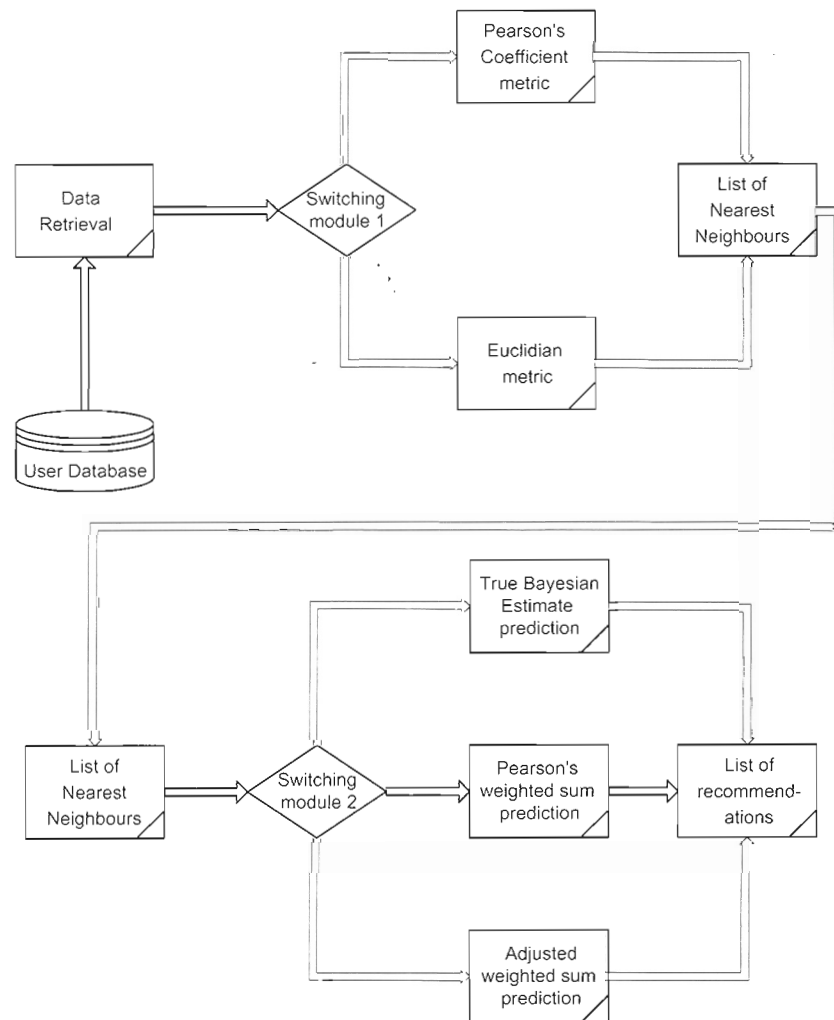


Fig. 1. System design.

neighbours is identified using distance calculation, followed by the prediction of ratings for content items. The design of our system is presented in Fig. 1.

We selected two different techniques for distance calculation/neighbour identification. The first is the Pearson's coefficient [12], which is often used in existing research. The other is the Euclidean distance, which we selected due to shortcomings demonstrated by Pearson's module.

The prediction technique that was used in our work is based on two existing techniques, with the addition of our own original technique, developed during the first stages of our research. We used the True Bayesian estimate prediction [13], the Pearson's weighted sum and our modified weighted sum prediction technique.

Our analyses were focused on three different areas. The first area is the optimum number of nearest neighbours required for the system to work properly. The second area was focused on the decision of which distance metric to use, based upon the user's information. The third area was the evaluation of the prediction technique results in

order to decide which one works best in given circumstances.

Our system is designed to be easily updated in order to add new modules in the future. The system runs in Java environment to enable inter-operability and access to mobile devices. The modules are optimised to permit easy reconfiguration for implementation in a specific environment. These modules are then integrated into a framework with the ability of easy integration of further modules and also a simple way of connecting them into a hybrid system. A module with the capability of switching modules (functions) is also needed. These switching modules take care of switching between different distance metrics for neighbour identification as well as switching between different recommendation techniques and can thus create a hybrid environment.

The system is designed to include a module for gathering information from the database (MySQL in our case), two different modules for nearest-neighbour selection, three modules for recommendation generation and a final module for generating output data and system efficiency calculation.

One of the stages of our work was to optimise this system to provide a high degree of confidence.

The system retrieves active user's information (his/her ID, ratings and his/her average rating) in the first stage. The first switching module then selects the best possible distance metric based on statistical character of the user's ratings (see Section 3.2 for more details). Once the distance metric is selected, the appropriate module then identifies the pre-set number of nearest neighbours for the active user. These neighbours are then sorted in a list that is going to be used for prediction generation. The next switching module selects one of the available prediction techniques. This module processes neighbour's information and creates a list of predicted items together with the predicted rating for each of these items. In order to evaluate the new system, the list of the generated recommendations is compared to the ratings found in the test database.

3.2. Nearest-neighbour selection

As we have already pointed out in Section 2, one of the main advantages of collaborative recommenders is that they are independent of content type. The principal mechanism used in such a system is the search for nearest neighbours. One of the most often used techniques used to achieve this goal is calculation of the correlation between the main user and each possible neighbour. This correlation is calculated using Eq. (1) where $w(a, i)$ represents the Pearson's coefficient, $v_{i,j}$ the rating that user i gave to item j and $\bar{v}_i$ represents the average rating of user i . This correlation is also known as the Pearson correlation coefficient. The smaller the value of the coefficient, the more similar the users are [12]. Distance is then represented as the absolute value of the correlation coefficient

$$w(a, i) = \frac{\sum_j (v_{a,j} - \bar{v}_a)(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_j (v_{a,j} - \bar{v}_a)^2 \sum_i (v_{i,j} - \bar{v}_i)^2}}. \quad (1)$$

While this distance metric returns very accurate results, our early research has indicated that there are some cases where it fails significantly.

The problem with the Pearson coefficient occurs in cases when the user's ratings do not deviate enough from his/her average rating $\bar{v}_a$. In such a case the resulting coefficient is always an extremely large number. This is the result of the numerical nature of the operations needed to calculate this correlation. We intend to try and fix this issue in later versions by implementing an alternative way of calculating correlation in such cases.

Ignoring the users who rate all items identically is disputable. A good example of such an exception would be a situation, where a new user is introduced into the system and wants to help the system by inputting all the items that he/she likes very much, giving all of them the same (highest) rating. Such a user has provided more than enough information for the system to work with, yet because he/she

registered only the best ratings he/she does not deviate enough from his/her average rating for the system to work using Pearson's correlation coefficient.

We have, therefore, set out to correct this problem by introducing an additional metric and configured the system to switch to the appropriate module according to the circumstances. Our proposed solution to this situation is an additional metric that would substitute the Pearson coefficient in situations where the user's deviation from his/her average rating falls below a set threshold. The metric that we suggest is the Euclidean distance calculated using Eq. (2). While this metric produces overall results of slightly lower quality, it has the benefit of being independent of the users' average rating and therefore works in any situation,

$$w(a, i) = \sum_j \sqrt{(v_{a,j} - v_{i,j})^2}. \quad (2)$$

Another benefit of using this metric is that it produces results in the same scope i.e. lowest being best and therefore no transformation is necessary for the prediction step.

These distance measures are realised as two different modules. A special switching module is introduced to detect which metric is correct for the current user and switch to the appropriate module as necessary.

3.3. Prediction generation

Once the nearest neighbours are identified the prediction step commences. Basically, it means checking which items were given a high rating by the neighbours that the main user has not accessed yet. By using several different equations, it is possible to calculate the predicted rating of the main user for these items. Once the predicted ratings are calculated, they are ordered according to their predicted vote. Those that have the highest predicted rating are most likely to be positively accepted by the user if suggested by the system.

One of the most frequently suggested methods for the calculation of predicted ratings is the weighted sum method [12], calculated using Eq. (3) where $p_{a,j}$ represents the predicted vote for item j calculated for user a , $\bar{v}_a$ the average rating of the active user and k a constant user for normalising the total sum with the number of ratings registered by neighbours for this item:

$$p_{a,j} = \bar{v}_a + k \sum_{i=1}^n w(a, i)(v_{i,j} - \bar{v}_i). \quad (3)$$

While this method generates good results, our research indicated that, in some situations, it falls short and fails to generate a good range of possible suggestions. The most obvious case is when a user has put only very high or very low ratings into the system. His/her average rating is then either very high or very low. In either case, the problem is the sum of weights from the nearest neighbours. The system relies on this sum to indicate whether the item will be appropriate or not. The problem with these two situations is that this

sum will be much smaller than the user's average rating. This means that the final predicted rating will have almost the same value as the user's average rating. In the case of a high average rating, this means that all the items recommended by neighbours will be automatically recommended and in the case of a low average rating no recommendations will be made because no object will be deemed appropriate.

In order to correct this situation, we propose an additional way of calculating predicted ratings using

$$p_{a,j} = k \sum_{i=1}^n w(a,i)(v_{i,j} - \bar{v}_i). \quad (4)$$

This is essentially just a modified version of Eq. (3). The logic behind is that in the above-mentioned situations the value of predicted vote is less important than finding out whether the content will be appropriate or not. Therefore, we concentrate on the sum of weights obtained from the nearest neighbours. If a neighbour liked the content his/her weights will be positive (they rated the content with a rating that is higher than their average rating) and negative if they disliked it. If the sum of all the weights is positive that means that most neighbours liked the content and it should therefore be recommended. If the sum is negative, it means that they either did not like it at all or did not consider it to be very interesting. Therefore, it can be discarded. Our experiment has shown that this measure is best suited for situations where the average rating of the user is very high or low, while in other cases the results produced are slightly worse than normal.

The final prediction method implemented in our framework is the true Bayesian prediction estimate [13] using

$$w = \frac{v}{v+m} R + \frac{m}{v+m} C. \quad (5)$$

In this equation, we calculate the weighted rank of each possible item and then suggest those with the highest rank. In this equation, R represents the average rating of the item calculated from ratings made by nearest neighbours, while C represents the item's average rating calculated from all the registered ratings. The number of neighbours who have rated the items is represented by v , while m represents a set threshold. This method is also quite popular on movie-oriented web-sites so we decided to see what results can be obtained by using it on our recommender system.

3.4. Framework optimisation

After creating our framework modules, we have discovered three different aspects that can be optimised further.

3.4.1. Distance metric switching

The first is the module used for switching between the two different distance metric modules for nearest-neighbour selection. As noted in Section 3.2, the main parameter is the user's deviation from his/her average rating. The

performance of the system is dependent on the threshold that determines how low deviation is still good enough to use the Pearson coefficient. We have, therefore, conducted a series of runs to measure the optimum threshold.

3.4.2. Number of nearest neighbours

The second aspect is the number of neighbours that are to be selected by the system. This number has a large impact on the quality of prediction. If the number is too small, the neighbours selected will be the best match possible but there will be few new items for the system to consider as possible recommendations. If the number is too high, all of the neighbours will not match the user's taste very well and the system will be overloaded with possible items for recommendations. Also, processing time will drastically increase if a larger number of neighbours is required. In order to find this optimum, we have conducted a series of tests where we used a different number of neighbours each time. The results of these tests should indicate what range could be considered as optimum.

3.4.3. Prediction method switching

The third aspect is switching between different prediction methods. If the system needed to calculate three different prediction sets each time a user logged on, the process would require a large amount of resources and the waiting time would increase. Since most users are not prepared to spend too much time waiting for the system to provide new suggestions, this aspect should be optimised as well. Ideally, research should point out those attributes that would identify which method is the most appropriate in the current situation. At this stage, our research acknowledges this problem and we are planning on conducting further experiments to identify these attributes, but no results related to this issue are available at this point.

4. Results

For our experiments, we have programmed individual modules containing parts of our hybrid recommender developed in Java. For the evaluation, we have used the Each-Movie database provided by Digital Equipment Corporation. The evaluation has been realised in three stages.

4.1. Framework design

The framework consists of different modules, where each module represents one of the basic functions. The current setup provides two modules for distance metric calculation – the Euclidean distance metric and the Pearson coefficient metric. An additional module is used to switch between these two metrics. For the purpose of prediction generation, three modules were created. The first one generates predictions using the Pearson's coefficient method (3). The next module

we designed creates predictions using our modified variant of the Pearson's weighted sum (4), while the last implements the true Bayesian prediction using Eq. (5).

The advantage of this framework design is that it is very easy to switch between different modules. Creating a system that also works with just one or two of these modules is also very simple and therefore experiments can be performed without many complications. We have decided to use the Java environment to enable easy integration on other operating system.

4.2. Experimental database structure

The EachMovie database [1] consists of three different files. The first file *Movie.txt* provides information on contents rated by the users. The file contains movie titles, URLs to websites providing movie metadata, release dates and genre information. The second file *Person.txt* provides user information such as user identification, age and gender. The last file *Vote.txt* contains all ratings given by the users and the corresponding dates.

For our purposes, we used only the information about user's ratings because collaborative filtering is focused only on finding nearest neighbours with the help of ratings the user has made. In the near future, we are also planning to implement the content-based recommendation modules so the information from the first two files will also be needed.

The database contains 61265 user profiles and 1628 different items that were rated by these users. Users were rating items using a 10-level scale, from 0 to 1, with 0.1 increment. The average number of ratings registered by a single user was 100. The whole database contains 2,811,718 registered ratings. The average rating in the database is 0.596.

When working with these data we first moved the data set into a MySQL database. The database allows future extensions in case of new functionalities. Results generated by the framework are written into files in order to allow further processing by parsing these files into other applications.

4.3. Result evaluation technique

After a set of predicted votes is generated, it is compared to the actual votes made by the user. Based on this comparison, we are able to calculate Precision and Recall for each user based on [14]

$$\begin{aligned} \text{recall} &= \frac{\text{true positive}}{\text{true positive} + \text{false negative}}, \\ \text{precision} &= \frac{\text{true positive}}{\text{true positive} + \text{false positive}}. \end{aligned} \quad (6)$$

For our evaluation purposes we set the threshold for determining true positives to 0.7 meaning that if a content was rated 0.7 or higher, it was considered to be accepted by the user. For the final result comparison, we used the *F*-measure

calculated from precision and recall

$$F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (7)$$

We believe this to be a good evaluation metric, because it represents both measures together and therefore provides a better overview of the recommender efficiency overall.

4.4. Selection of the optimum number of nearest neighbours

In order to find the optimal number of nearest neighbours to use in a collaborative recommendation system, we ran a series of tests in which we alternated the number of selected nearest neighbours and compared the results for each set. We also ran these tests for each of the two different distance metrics (Euclidean and Pearson's) in order to see the effect of the number of nearest neighbours on each of them.

In our test, we calculated predictions using 5, 20, 30, 50, 70 and 90 neighbours. When calculating predictions based upon these neighbours, we used the true Bayesian prediction module for predicted rating generation. We decided to use this prediction technique based upon our early results that indicated that the true Bayesian prediction results in best Precision and Recall values in most of the cases. This can also be seen in Figs. 5 and 6. An additional test also revealed that the number of neighbours is independent of the prediction algorithm selected for the system. We, therefore, decided to use the algorithm that would produce consistent recommendations throughout the whole experiment. Each of the users was first checked to determine whether he/she had enough votes registered in the system for the algorithm to work. In our case, we have decided that the minimum number of these votes is 40 since every user's ratings need to be divided into a training set and a test set, while in the final implementation of this system this number would be 20. If we set the number to a lower value, there would be no significant improvement in the test results. If we set the number to a higher value, we would lose important information about how the system performs at the very beginning when the user's profile does not contain a large amount of ratings. The user's data set was then split in two profiles. The first contained half of his/her ratings that were used as a training set while the second contained the remaining ratings to be used as a test set. At the end of this filtering process, 21,994 user profiles turned out to be suitable for testing. After calculating results for each of the metrics, we calculated a logarithmic trend line for each of the subsets. Results are shown in Figs. 2 and 3.

If we compare the results in Figs. 2 and 3, it can be clearly seen that a too many or too few neighbours produce very bad results. In our case that was the situation when using 5 or 90 neighbours, where the *F*-measure remained below 0.60 when using the Pearson's metric. When using the Euclidean metric (Fig. 3) these numbers were

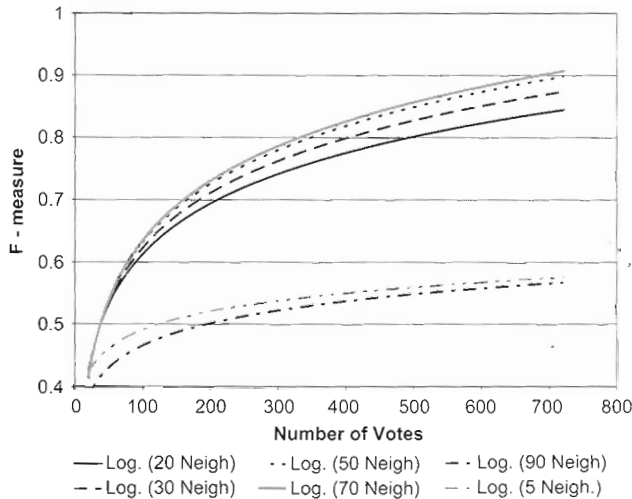


Fig. 2. Optimum number of neighbours for the Pearson's distance metric.

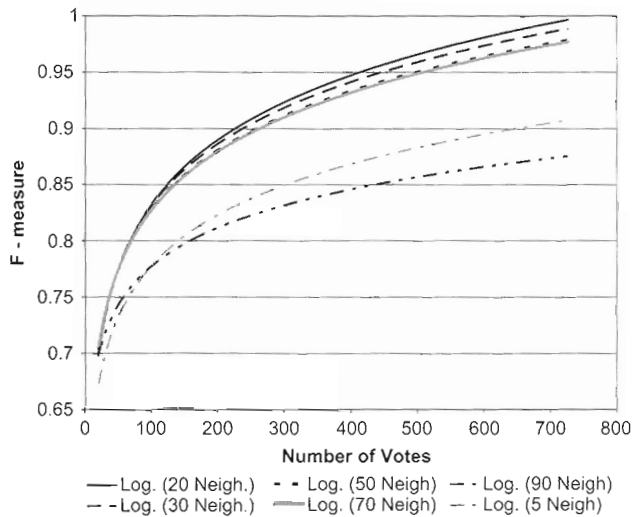


Fig. 3. Optimum number of neighbours for the Euclidean distance metric.

higher (F -measure around 0.85) but still significantly lower (0.10–0.15) when compared to results obtained when using 20–50 neighbours. Interestingly, there is a difference in the optimal number of neighbours when using different metrics. In the case of the Euclidean distance the optimum was reached very fast (20 neighbours), while in the case of the Pearson's distance the optimum number was significantly higher (70 neighbours). If the system was to be constructed using only one of these distance metrics, these numbers could be considered as the optimum. Since we are designing a system that uses both distance metrics, that means that it would be necessary to change the number of neighbours for the active user each time the system detects that a change of metric is required. Such a change also requires a complete recalculation of the user's profile. Such a system would be

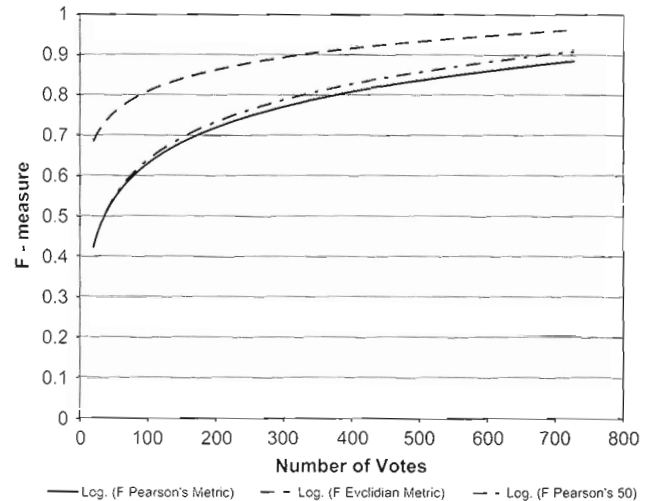


Fig. 4. Distance metrics comparison.

very inefficient and would require a lot of additional time to process each user. We have, therefore, decided for a trade-off achieved by using 50 neighbours. As can be seen from the results this does not represent any significant loss in the case of the Euclidean distance, yet still gives a significant increase in the case of the Pearson's distance.

4.5. Distance metric evaluation

As we outlined in the previous sections, part of our experiment was also trying to discover which distance metric produced better results or was best suited for certain situations. We already noticed in the previous experiment that each metric produced different results for the same user.

In the first phase of this experiment, we tried to find out which metric produced better results when used on the whole database. To produce results we ran our recommender system using either only the Euclidean distance metric or only the Pearson's coefficient distance metric. In the first run, we calculated each set for 30 nearest neighbours. Because we also had access to results from the first phase of our experiments, we then calculated an additional set where we used the Pearson's coefficient distance metric using 50 and 70 nearest neighbours to allow a better comparison. For the actual comparison we again used the F -measure to calculate the effectiveness of our algorithm. Results from these experiments are shown in Fig. 4.

If we observe these results in Fig. 4, we see that in the early stages (when the user has a small number of registered ratings) the Euclidean distance metric outperforms the Pearson's distance metric. For example, when dealing with a user that has made 50 ratings, recommendations made using the Pearson's metric results in F -measure of 0.55, while the Euclidean metric results in F -measure of 0.75. This means that in the early stages the Euclidean metric outperforms the Pearson's metric by as much as 0.2. In the later

stages this difference somewhat decreases. When processing an user with 400 ratings this difference is lowered to only 0.10 (Euclidean metric F -measure = 0.91, Pearson's metric F -measure = 0.81). In the final stages (i.e. when the user has registered a very large amount of ratings) this difference becomes even smaller dropping to about 0.05.

In order to get additional information, we then compared each individual user's F -measure for the Euclidean metric and the Pearson's metric. These results have shown that although the Euclidean distance gives best results overall there are cases when the Pearson's coefficient works better. In our test set, the Euclidean metric was better in 18,820 cases while the Pearson's metric performed better in 3174 cases. This means that the optimum system would be one that includes both metrics and switches between them (as already mentioned above).

In the final stage of this experiment, we then tried to determine what kind of switching mechanism would produce the best overall results. The best solution we found was by using a threshold function that determined which metric to use based on user's deviation from his/her average rating. We decided to use a threshold value of 0.1 using the Pearson's coefficient metric when the user's deviation is greater than this value and the Euclidean metric in other cases. This produced improved results but we believe that further improvements can be made by using a more complex switching function. We are still in the process of discovering other relevant parameters that would enable us to construct such a function.

4.6. Prediction method evaluation

In the final stage of our experimentation, we tested three different prediction methods in order to determine what kind of results could be expected from them in given circumstances. In order to generate prediction, we first identified nearest neighbours for each user using previously discussed distance metric modules and threshold function for switching. After the neighbours were identified, we separately generated results for each of the prediction methods. For result comparison we once again used F -measure to determine which method performed better.

We used three different methods – the true Bayesian prediction, prediction using the Pearson's coefficients and a weighted sum prediction that we developed by modifying Pearson's coefficients. In the first part of this experiment, we compared how each of these methods performed when using a different distance metric to select nearest neighbours. Results are shown in Figs. 5 and 6.

The first conclusion that we can make from these results is that when dealing with a single-distance metric, we can see that the true Bayesian prediction outperforms other methods by 0.05–0.10 if using F -measure to compare results. The next conclusion that we can draw is that prediction based on neighbours selected using the Euclidean distance metric

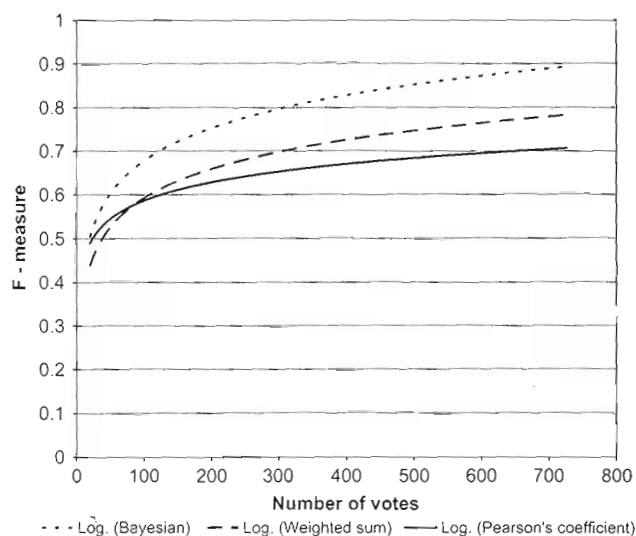


Fig. 5. Comparison of prediction techniques using the Euclidean distance metric.

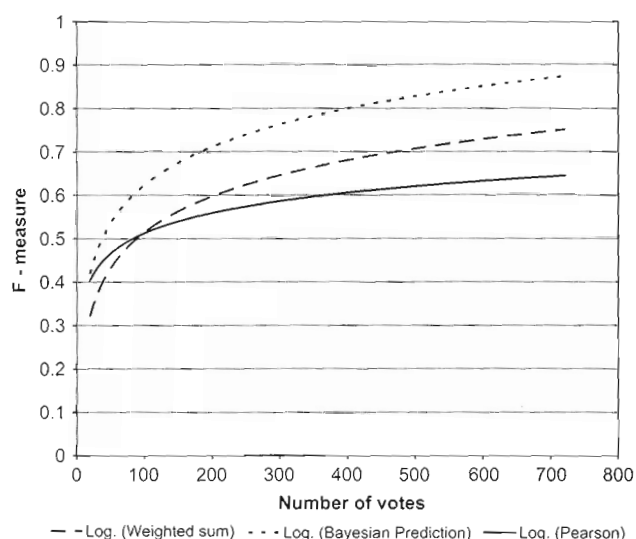


Fig. 6. Comparison of prediction techniques using the Pearson's distance metric.

provides better results. From the two figures (Figs. 5 and 6), we can observe that the F -measures increases at a slower pace when using the Pearson's distance metric. This effect is stronger when using weighted sum based prediction methods, while the true Bayesian prediction still achieves a reasonably high F -measure (greater than 0.70) fairly quickly. We can compare different combinations of distance metrics and prediction algorithms in several different ways. If we, for example, focus on a user who has registered 100 ratings we can see that the best choice would be a system that uses the Euclidean distance metric for neighbour selection and the true Bayesian estimate for prediction generation. Such a system typically results in a F -measure with a value of 0.68.

Table 1. Prediction method comparison results

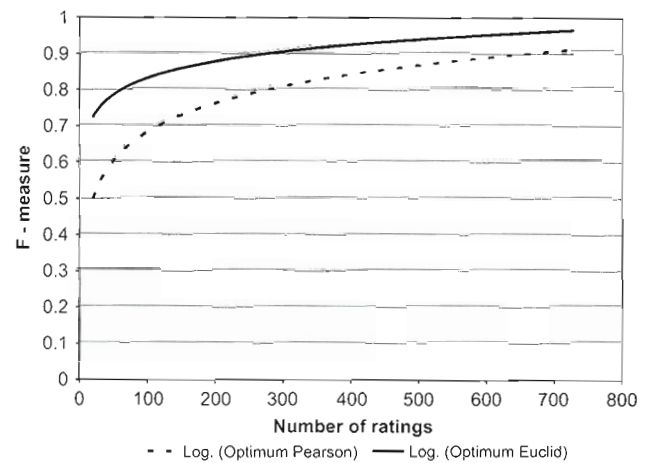
Distance metric	Prediction algorithm	<i>F</i> -measure at no. of ratings		
		60	100	250
Euclidian	True Bayes	0.60	0.68	0.77
Euclidian	Pearson's WS	0.54	0.58	0.65
Euclidian	Adjusted WS	0.51	0.58	0.67
Pearson's coeff.	True Bayes	0.55	0.63	0.74
Pearson's coeff.	Pearson's WS	0.47	0.51	0.58
Pearson's coeff.	Adjusted WS	0.44	0.51	0.63

If we decide to use the Pearson's coefficient metric for neighbour selection while still using the true Bayesian estimate for prediction generation the results are slightly worse, resulting in a *F*-measure of 0.63. Other combinations of prediction algorithms do not differ much among themselves, the only difference being based upon the distance metric used. Combinations using the Euclidean distance metric for neighbour selection result in *F*-measure value of 0.58, while combinations using the Pearson's coefficient result in *F*-measure value of 0.51. If we make the same comparison at a later stage (i.e. when the user has registered more ratings) the differences become more obvious. If we make this comparison at 250 registered ratings we see the following results. System using the Euclidean distance metric for neighbour selection and the true Bayesian prediction for prediction generation works best, resulting in *F*-measure value of 0.77. The next best system is one based on the Pearson's coefficient distance metric for neighbour selection and the true Bayesian prediction results in an *F*-measure of 0.74. Also differences between the Pearson's weighted sum and our weighted sum become apparent. These results are shown in Table 1.

From this table, we can see that our weighted sum performs better than the Pearson's weighted sum, and this difference is growing with the number of ratings the user has registered. While this indicates that a system could be designed using only the Euclidean distance for neighbour identification and the true Bayesian estimate for prediction generation, this is not necessarily true as we have already pointed out in Section 4.5.

Because we want to create a system that would always produce optimal results, we compared these sets and compiled them into a single "optimum" set that represents how the system should perform in the final design. The final design means a system that would successfully switch between different distance metrics in order to provide best results. To create this design we used the best results for each user based on *F*-measure values. These results are shown in Fig. 7.

We decided to create this combined system because a more thorough inspection of results showed that while the true Bayesian prediction usually gives better predictions, there are still cases when other techniques perform better.

**Fig. 7.** Optimum recommender results.

In our case, we found that the true Bayesian prediction gave best results for 12,890 users out of total 21,978 users in our database. The Pearson's weighted sum performed better for 5220 users, while our weighted sum gave better predictions for 3860 users. This means that the true Bayesian prediction was the better choice for 59% of our users, Pearson's weighted sum for 24% and our weighted sum for 17%. These numbers led us to believe that it would be sensible to create a combined recommender in order to improve prediction results for those 41% of users who would be disregarded if we only used the true Bayesian prediction.

We created this optimised recommender using additional switching modules that enabled us to decide which combination of distance metrics and predictions algorithms to use for each user. At this early stage, the switching module selects the distance metric based upon the user's deviation from his average rating. While this approach has already produced good results, we hope to further improve the results in later stages of our research where we will try to find additional parameters that should be considered when switching between modules.

5. Conclusions

We have developed a new combined collaborative recommender system and tested it in many different situations.

We have used three different sets of results, each of them helping us to draw a separate conclusion. When comparing different distance metrics for the purpose of nearest-neighbour identification we have shown that the Euclidean distance metric performs better than the Pearson's coefficient metric in most cases, but we have also demonstrated that although this situation is quite often true there are still some exceptions that make combining these two metrics into a single switching system seem worthwhile. We have pointed out some weaknesses that can occur when using the Pearson's coefficient distance metric. One of these weaknesses

that must be considered thoroughly is this – recommender systems should work by becoming better and better at predicting what the user might like. Should the user rate only items suggested by the system, his/her deviation from average rating would likely become smaller and thus creating the situation where the Pearson's metric fails as we pointed out in Section 3.2.

We have also analysed what information is relevant when trying to decide which of these metrics to use. The first results show that switching between distance metrics based upon user's deviation from his/her average rating works reasonably well, but we believe that these results can be further improved with the help of a more detailed analysis of available user information.

We have also compared three different prediction techniques and tried to determine which one performs best. We have demonstrated that the true Bayesian prediction gives best results when viewing the overall results but there are still exceptions. We have also introduced our prediction technique – adjusted weighted sum that gives better results than the Pearson's weighted sum. The results are inferior to the true Bayesian estimate but, it should be pointed out that our prediction generates results by using only information provided by nearest neighbours thus requiring less processing time. The true Bayesian prediction requires additional information about all the users who have rated the selected item, thus requiring a longer time to generate predictions. We have also demonstrated that the true Bayesian estimate does not always produce best result, which is why we believe that our combined collaborative system represents an improved recommendation system that should provide more reliable recommendations and should also be more flexible and easier to adapt to new requirements.

Although each prediction technique can produce adequate results, we have proved that the combination of these techniques into a unified system provides a much more stable system. It should also be pointed out that these analyses were done using a very large dataset (more than 2 million ratings) providing reliable results.

For future development, we have decided to further improve our framework based on our experimental results. In the first stage, the system will be configured to work with 50 nearest neighbours to provide best possible results for both distance metrics without having to change the number of neighbours each time we switch the distance metric.

In addition to the improvements mentioned in previous sections, we will introduce some major improvements. Our testing has also shown that calculating nearest neighbours for each separate user is very time consuming and therefore inefficient if used in a system with a large number of registered users as can be seen in Fig. 8. Therefore, we have decided to introduce clustering whereby we would simply analyse user profiles and see how many clusters of similar users exist in the system. Each user would then be assigned to a specific cluster and all other users present in this cluster would be considered his nearest neighbours. That would

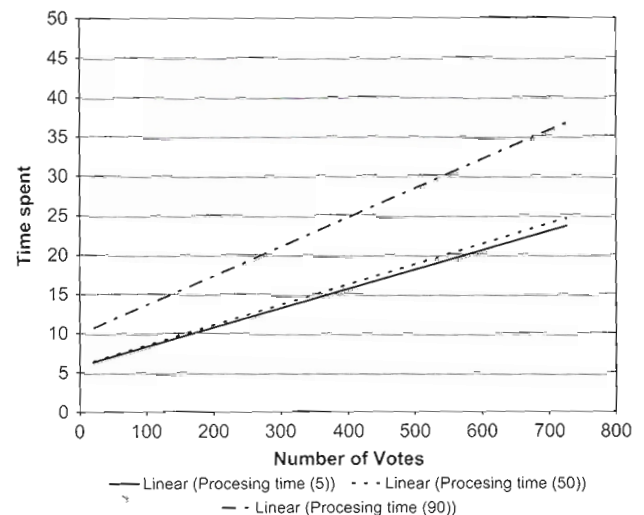


Fig. 8. Processing times for different numbers of neighbours.

drastically decrease the amount of time necessary to create suggestions for each user.

This would also lead to a solution of the “Cold start” problem that the system encounters when trying to generate predictions for new users who have not registered enough ratings yet. Since the system would now consist of clusters each cluster could be identified by some representative items that would typically be highly rated only by users from this cluster. By using these unique items, it would be possible to create a questionnaire for new users. A new user would then just answer these questions and would be assigned to the appropriate cluster according to his/her answers. This means that predictions could be generated immediately after the user is assigned to a cluster.

The last goal of our research is to improve our combined collaborative recommender by fusing it with other recommenders, preferably content based thus creating a true hybrid recommender system. This will be done by evaluating several content based recommendation techniques. A selection of those that can work with our datasets will be used to create a hybrid recommender system. This will also present a challenge from the point of view of constructing a hybrid system because we also want to analyse which hybrid approach works best in our environment.

Acknowledgements

This work would not have been possible without the Grants of the Slovenian Research Agency nos. P2-0246 and R-74305.

References

- [1] Eachmovie collaborative filtering dataset. Provided by Digital Equipment Corporation. (<http://research.compaq.com/SRC/eachmovie>), accessed on October 2004.

- [2] Burke R. Hybrid recommender systems: survey and experiments. *User Model User-Adap Interact* 2002;12: 331–70.
- [3] Bezzera B, Carvalho F, Ramalho G, Zucker J-D. Speeding up recommender systems with meta-prototypes. *Lecture notes in computer science, advances in artificial intelligence*. Berlin: Springer; 2002. p. 227–36.
- [4] Zimmerman J, Kurapati K, Buczak AL, Schaffer D, Martino J, Gutta S. TV personalization system: design of a TV show recommender engine and interface. Dordrecht: Kluwer Academic Publishers; 2004.
- [5] Kurapati K, Gutta S. TV personalization through stereotypes. In: *Proceedings of the AHÂ'2002 workshop on personalization in future TV*, 2002.
- [6] Uchyigit G, Clark K. An agent based electronic program guide. In: *Proceedings of the AHÂ'2002 workshop on personalization in future TV*, 2002.
- [7] Salter J, Antonopoulos N. Cinemascreen recommender agent: combining collaborative and content-based filtering. *IEEE Intell Syst* 2006;21:35–41.
- [8] Pogacnik M, Tasic J, Tomazic S. Personalized content retrieval. In: *Proceedings of the international symposium on telecommunications VITEL*. Ljubljana: Electrotechnical Society of Slovenia; 2004.
- [9] Sarwar B, Karypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms.
- [10] Goldberg K, Roeder T, Gupta D, Perkins C. *Eigenstate: a constant time collaborative filtering algorithm*; 2000.
- [11] Mashhoff J. Group modeling: selecting a sequence of television items to suit a group of viewers. *User Model User-Adap Interact* 2004;14:37–85.
- [12] Breese JS, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: *Proceedings of the fourteenth conference on uncertainty in artificial intelligence*, Madison, USA: Morgan Kaufmann Publisher; 1998.
- [13] Hand D, Smyth HMP. *Principles of data mining*. Cambridge: The MIT Press; 2001.
- [14] Raaijmakers S, den Hartog J, Baan J. Multimodal topic segmentation and classification of news video, 2002.



Matevž Kunaver is a researcher at the Faculty of Electrical engineering, University of Ljubljana. He received his B.Sc. degree in 2004. His research interests include user modelling and hybrid system design. Currently, he is working on European IST projects, in which he is developing and testing collaborative-based recommendation algorithms and their integration into hybrid recommenders systems.



Tomaž Počrl is a researcher at the Faculty of Electrical engineering, University of Ljubljana. He received his B.Sc. degree in telecommunications in 2005. His research interests include user modelling, data mining and ontology-based multimedia applications. He is currently working on European IST projects, in which he is developing mechanisms for content-based selection and recommendation of multimedia content.



Matevž Pogačnik is a researcher at the Faculty of Electrical engineering, University of Ljubljana. He received his M.Sc. degree in agent technologies in 2000 and his Ph.D. in personalised content retrieval in 2004. His research interests include user modelling and knowledge-based data mining. Currently, he is working on European IST projects, in which he is developing mechanisms for personalised selection of multimedia content applied to heterogeneous environments (p2p networks, mobile devices, set-top-boxes, PDAs, etc.).



Jurij F. Tasič is a Professor of system theory and computing at the University of Ljubljana. In 1983, he began working on communication systems and the usage of numerical methods in DSP, including the usage of system theory in communications and communication systems. His current interests are advanced algorithms in communication systems, multi-dimensional signal processing and parallel algorithms. He is the author of more than 60 research papers published in journals and conferences.



Number of microvesicles in peripheral blood and ability of plasma to induce adhesion between phospholipid membranes in 19 patients with gastrointestinal diseases

Rado Janša^a, Vid Šuštar^b, Mojca Frank^c, Petra Sušan^j, Janez Bešter^d,
Mateja Manček-Keber^e, Mojca Kržan^f, Aleš Iglič^{d,*}

^a Department of Gastroenterology, University Medical Centre Ljubljana, Ljubljana, Slovenia

^b Laboratory of Clinical Biophysics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

^c Department of Rheumatology, University Medical Centre Ljubljana, Ljubljana, Slovenia

^d Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia

^e Department of Biotechnology, National Institute of Chemistry, Ljubljana, Slovenia

^f Institute of Pharmacology and Experimental Toxicology, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

ARTICLE INFO

Article history:

Submitted 29 January 2008

Available online 1 April 2008

(Communicated by J.F. Hoffman, Ph.D.,
29 January 2008)

Keywords:

Microvesicles
Microparticles
Cancer, gastric
Tumor progression
Procoagulant

ABSTRACT

It was recently shown that the plasma protein-mediated attractive interaction between phospholipid membranes could in the budding process cause adhesion of the bud to the mother membrane [J. Urbanija, N. Tomšič, M. Lokar, A. Ambrožič, S. Čučnik, M. Kanduđer, B. Rozman, A. Iglič, V. Kralj-Iglič, Coalescence of phospholipid membranes as a possible origin of anticoagulant effect of serum proteins, *Chem. Phys. Lipids* 150 (2007) 49–57]. Since in the *in vivo* conditions the budding of cell membranes leads to the release of microvesicles into the circulation, a hypothesis was put forward that the ability of plasma to cause adhesion between membranes suppresses the microvesiculation process. In the present work, this hypothesis was tested in a population of 19 patients with gastrointestinal diseases. The number of microvesicles in peripheral blood of patients was determined by flow cytometry while the ability of plasma to cause adhesion between membranes was determined by adding patient's plasma to the suspension of giant phospholipid vesicles created by electroformation method, and measuring the average effective angle of contact between the adhered vesicles. Statistically significant negative correlations between the number of microvesicles and the average effective angle of contact (Pearson coefficient -0.50 , $p=0.031$) and between the number of microvesicles per number of platelets and the average effective angle of contact (Pearson coefficient -0.64 , $p=0.003$) were found, which is in favor of the above hypothesis. Patients with gastrointestinal cancer had larger number of microvesicles (difference 140%, statistical significance 0.033) and smaller average effective angle of contact (difference 20%, statistical significance 0.013) compared to patients with other gastrointestinal diseases.

© 2008 Elsevier Inc. All rights reserved.

Introduction

Microvesicles (MVs) can be considered as extracellular organelles [2] which convey communication between distant cells. MVs found in circulation originate from different cells [3–6], but mostly from blood cells [7]. They are final products of a process starting with lateral redistribution of membrane components, proceeding with development of membrane buds and ending with pinching off the buds from the mother membrane [8]. Increased number of MVs was found in peripheral blood of patients suffering from cardiovascular disorders [3,4], infection [4,9–11], autoimmune diseases [12,13] and cancer [14–16]. Tumor-derived MVs are present also in other body fluids of patients with cancer [17].

Due to the importance of microvesiculation of membranes, the underlying mechanisms were studied in different systems. A simple

system which could reveal physical mechanisms involved are giant phospholipid vesicles (GPVs). GPVs possess many important properties of cell membranes and can be observed live under an optical microscope. Vesiculation of GPVs can be induced by different mechanisms such as an increase of the temperature and manipulation of the solution in contact with the outer layer of the GPV membrane. In such studies it was found that the addition of plasma proteins (in particular, beta 2 glycoprotein I and antiphospholipid antibodies) into the suspension of GPVs may cause budding of GPVs [18], coalescence of GPVs [19] and adhesion of buds to the mother membrane [1], (Fig. 1). The accompanying theoretical studies on protein-mediated interactions between membranes indicate that plasma proteins can induce attractive interaction between membranes even when they are like-charged, due to positional and orientational distribution of protein molecules with spatially distributed charge [20].

It was suggested on the basis of the above experimental and theoretical results that plasma protein-mediated attractive interaction between membranes may prevent the release of MVs into the

* Corresponding author. Fax: +386 1 4768 850.
E-mail address: ales.iglic@fe.uni-lj.si (A. Iglič).

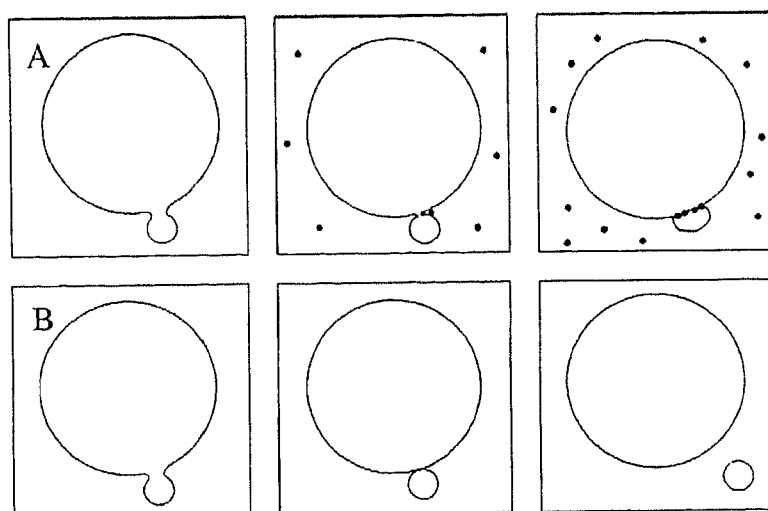


Fig. 1. Scheme of the budding and vesiculation of the membrane in the presence (A) and in the absence (B) of adhesion mediators. If adhesion mediators are present, the bud adheres to the mother cell membrane (A) while if adhesion mediators are absent the bud pinches off the mother membrane and becomes a microvesicle (B).

circulation [1]. A hypothesis was put forward that in subjects with plasma which induces a more pronounced adhesion between membranes the number of MVs in the peripheral blood will be smaller. A method for determination of the extent of plasma-induced adhesion between membranes was proposed by assessing the average effective angle of contact between GPVs which adhered due to the addition of plasma to the suspension of GPVs [21]. It was assumed that larger average effective angle of contact corresponds to a more pronounced adhesion [21]. The proposed trend was confirmed by a pilot study involving three healthy subjects [21]. However, in order to obtain decisive evidence regarding the validity, the hypothesis should be tested on different populations involving larger number of subjects. This work presents a study of a population of 19 subjects with various gastrointestinal diseases. At this point it is of interest to determine a general trend of the dependence of the number of MVs on the average effective angle of contact between GPVs. Therefore, it is preferable to include in the study the subjects within a wide range of the number of MVs, including patients with cancer, in whom the number of MVs is expected to be high. As among 19 of the included patients 5 of them were diagnosed with cancer we could also assess possible differences between the group formed by these patients and the group formed by all other patients, regarding the number of MVs and the average effective angle of contact between GPVs.

Patients

We originally included in the study 21 patients from Department of Gastroenterology, University Medical Centre Ljubljana, who were admitted in 2007 due to various gastrointestinal diseases (Tables 1–3). Patients gave a written consent to participate in the study. Two experiments were performed (the first experiment included 15 patients and the second experiment included 6 patients). In two patients, the sample with GPVs was of insufficient quality to determine the average effective angle of contact, so these two patients were included only in the analysis of the number of MVs. To compare the patients with diagnosed gastrointestinal cancer with other patients we formed two respective groups. The group A (patients diagnosed with cancer) consisted of 5 patients and the group B (patients with other gastrointestinal diseases) consisted of 16 patients (for analysis of MVs) or 14 patients (for analysis of the average effective angles of contact between GPVs).

Materials and methods

Preparation of human plasma samples and microvesicle isolation

2 ml of venous blood was collected into vacutubes containing trisodium citrate and processed within 15 minutes (min). Following the centrifugation of blood (1500 $\times$ g, 15 min, 20 °C, SIGMA 3K18 Centrifuge, Sigma), two plasma samples for each patient (250 μ l each) were used for isolation of MVs according to the method described in Diamant et al. [7] while 300 μ l of plasma was used for experiments with GPVs. To isolate MVs, 250 μ l of plasma was centrifuged at 17,570 $\times$ g for 30 min at 20 °C (SIGMA 3K30 Centrifuge, Sigma) and 225 μ l of supernatant was removed. After adding 225 μ l of phosphate buffer saline (136.9 mmol/l NaCl, 2.7 mmol/l KCl, 7.8 mmol/l $\text{Na}_2\text{HPO}_4 \times 2\text{H}_2\text{O}$ and 1.5 mmol/l KH_2PO_4 , pH 7.4) containing 10.9 mmol/l trisodium citrate, centrifugation at 17,570 $\times$ g and the removal of supernatant was repeated. For analysis by flow cytometer, the samples containing MVs were diluted 4 times: 975 μ l of citrated phosphate buffer saline was added to 25 μ l of the pellet. Isolation of MVs and experiments with GPVs were performed simultaneously, immediately after the separation of plasma from the cells.

Flow cytometry

Samples containing isolated MVs were analyzed using a Coulter EPICS Altra flow cytometer (Beckman Coulter Electronics). 7000 events were recorded. The number of MVs was given relatively to the number of Flow-count fluorospheres (1.05×10^6 /ml in solution, Beckman Coulter). The samples for measurement of the number of MVs were prepared from 585 μ l of diluted MVs and 15 μ l of fluorospheres in solution.

Platelet count

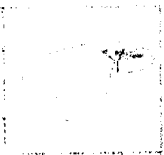
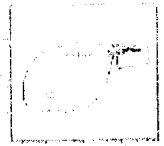
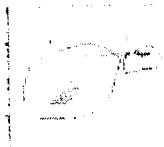
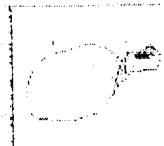
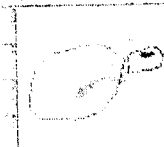
For determination of the platelet count, 3 ml of venous blood was collected into vacutubes containing K_3EDTA anticoagulant. Platelets were counted by Siemens Advia 120 hematologic counter.

Preparation and observation of giant phospholipid vesicles

GPVs were prepared by the modified electroformation method, originally proposed by Angelova et al. [22]. For neutral GPVs, 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC) and cholesterol (Avanti

Table 1

Data on sex, age, forward scatter vs. side scatter distribution of MVs(I)/fluorospheres (J) as measured by flow cytometer, diagnosis, platelet concentration, number of MVs, number of MVs with respect to number of platelets and the average effective angle of contact between adhered GPVs, pertaining to the patients with gastrointestinal cancer (group A)

#	Sex, year of birth	Forward scatter/side scatter of MVs (I) and fluorospheres (J)	Diagnosis	Platelet concentration $\times 10^9/l$	MVs/fluoro spheres	MVs/fluoro spheres/platelets ($10^{-9} l$)	Average effective angle of contact (degrees)
1	F, 1963		metastatic pancreatic cancer	411	3.82	0.0271	84
2	F, 1928		locally advanced pancreatic cancer, hyperlipidemia	349	3.64	0.0104	97
3	M, 1922		carcinoma papillae Vateri	231	7.13	0.0308	103
4	M, 1942		initial hepatocellular carcinoma, liver cirrhosis	130	0.4	0.0031	116
5	M, 1949		advanced hepatocellular carcinoma	63	2.05	0.0325	80

Polar Lipids, Inc.), dissolved in a 2:1 chloroform/methanol mixture at concentration of 1 mg/ml, were combined in the proportion of 4:1 (v/v). Cholesterol was added to POPC to increase the longevity of vesicles.

20 μ l of the lipid mixture was applied to each of the two platinum electrodes shaped as rods (approximate length 4 cm and diameter 1 mm). The electrodes were left in a low vacuum for 2 h for solvent to evaporate. The lipid-coated electrodes were thereafter placed into a microcentrifuge tube filled with 1.8 ml of 0.2 M sucrose solution to form an electroformation chamber. An AC electric current with an amplitude of 5 V and a frequency of 10 Hz was applied to the electrodes for 2 h, which was followed by 2.5 V and 5 Hz for 15 min, 2.5 V and 2.5 Hz for 15 min and finally 1 V and 1 Hz for 15 min. After the electroformation, 600 μ l of sucrose solution containing GPVs and 1 ml of 0.2 M glucose solution were pipetted in three 2 ml plastic microcentrifuge tubes which were sealed with parafilm band to prevent entrance of air and to protect the solution from microorganisms. The vesicles were left for sedimentation under gravity in a low vacuum at room temperature for 1 day.

The GPVs were observed by an inverted microscope Nikon Eclipse T-300 with phase contrast optics. For the experiments 70 μ l CoverWell™ Perfusion Chambers (Grace Bio-Labs), which allowed simultaneous performance of four experiments, were used. 7 μ l of citrated plasma sample was added through a circular opening to the 63 μ l of GPV suspension in the well of the perfusion chamber. For all experiments, the

images of adhered GPVs were taken 30 min after the addition of plasma at 4 mm distance from the circular opening where plasma was introduced. The experiments were performed at room temperature.

Determination of the mediating effect of plasma

To assess the adhesion between GPVs after the addition of plasma, all clearly visible effective angles of contact between adhered GPVs from an image (see Figs. 2–4) were measured using Corel Draw software. An average from all the angles determined from a given frame was calculated for each patient.

Statistical analysis

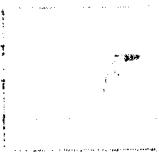
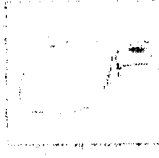
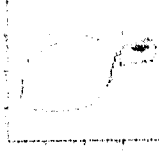
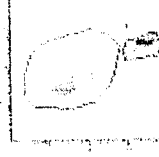
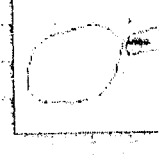
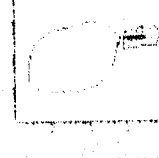
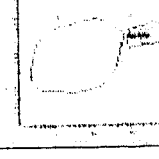
Methods of descriptive statistics were used. The statistical significance of the correlation was determined by the Pearson coefficient while the statistical significance of the difference between the two groups of patients was determined by the Student's *t*-test. Power analysis was performed by the Power and Precision V3 software.

Results

Fig. 2 shows the GPV samples 30 min after the addition of plasma pertaining to patients from group A (patients with gastrointestinal

Table 2

Data on sex, age, forward scatter vs. side scatter distribution of MVs(I)/fluorospheres (J) as measured by flow cytometer, diagnosis, platelet concentration, number of MVs, number of MVs with respect to number of platelets and the average effective angle of contact between adhered GPVs, pertaining to the patients from group B (patients marked by 6–11, 20 and 21)

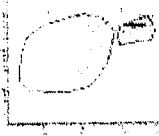
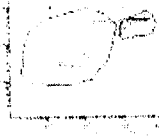
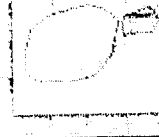
#	Sex, year of birth	Forward scatter/side scatter of MVs (I) and fluorospheres (J)	Diagnosis	Platelet concentration $\times 10^9/l$	MVs/fluorospheres	MVs/fluorospheres/platelets ($10^{-9} l$)	Average effective angle of contact (degrees)
20	F, 1945		ulcus ventriculi	389	0.62	0.0016	
21	M, 1944		acute pancreatitis, hyperlipidemia	256	0.41	0.0016	
6	M, 1919		hemorrhagic angiodysplasia of the upper gut	261	0.28	0.0011	95
7	M, 1978		ulcerative colitis	1052	0.58	0.0005	108
8	F, 1945		liver cirrhosis	338	1.34	0.0029	103
9	M, 1932		ulcus bulbi duodeni, infection	668	1.14	0.0017	122
10	M, 1958		liver cirrhosis	97	0.36	0.0037	110
11	M, 1965		liver cirrhosis	137	0.56	0.0041	145

cancer). Correspondingly, Figs. 3 and 4 show the GPV–plasma samples pertaining to patients from group B (patients with other gastrointestinal diseases). The clearly visible effective angles of contact were measured as shown in Figs. 2,3 and 4 and the average for each patient

was determined. Differences between GPV–plasma samples pertaining to different patients can be observed, such as the presence of "debris" in some samples and larger differences between refraction indexes of the vesicles and of the surrounding solution as exhibited in

Table 3

Data on sex, age, forward scatter vs. side scatter distribution of MVs(I)/fluorospheres (J) as measured by flow cytometer, diagnosis, platelet concentration, number of MVs, number of MVs with respect to number of platelets and the average effective angle of contact between adhered GPVs, pertaining to the patients from group B (patients marked by 12–19)

#	Sex, year of birth	Forward scatter/side scatter of MVs (I) and fluorospheres (J)	Diagnosis	Platelet concentration $\times 10^9/l$	MVs/fluoro spheres	MVs/fluoro spheres/platelets ($10^{-9} l$)	Average effective angle of contact (degrees)
12	M, 1951		liver cirrhosis, infection	78	0.27	0.0035	110
13	F, 1941		liver cirrhosis, acute cholecystitis	100	0.44	0.0044	117
14	F, 1940		liver cirrhosis	114	0.59	0.0051	118
15	F, 1984		autoimmune hepatitis	105	0.66	0.0063	118
16	M, 1934		liver cirrhosis, infection	174	0.53	0.0030	119
17	M, 1955		liver cirrhosis, hemoragic ulcer	78	0.40	0.0051	128
18	F, 1951		acute pancreatitis, hyperlipidemia	437	1.37	0.0031	135
19	F, 1914		hemorrhagic ulcer, infection	396	0.35	0.0008	106

the halo effect. The observed samples contained different populations of GPVs with respect to their size and density.

We found a negative, statistically significant correlation (Pearson coefficient = -0.50 , $p=0.031$) between the number of MVs in periph-

eral blood and the ability of plasma to induce coalescence between membranes – represented by the average effective angle of contact between adhered GPVs (Fig. 5A). Statistical significance of the correlation was even higher if the number of MVs was calculated

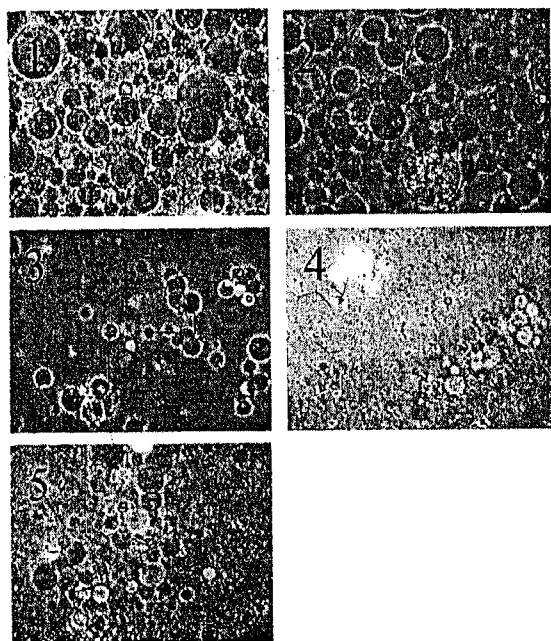


Fig. 2. Images of GPV samples 30 min after the addition of patients' plasma (group A, patients marked by 1–5). The effective angles of contact between the adhered GPVs are depicted. Small circles indicate the contribution of 360° by three adjacent effective angles of contact.

with respect to the number of platelets (Pearson coefficient = -0.64 , $p=0.003$), (Fig. 5B). This result is in favor of the hypothesis that in subjects with plasma which induces a more pronounced adhesion

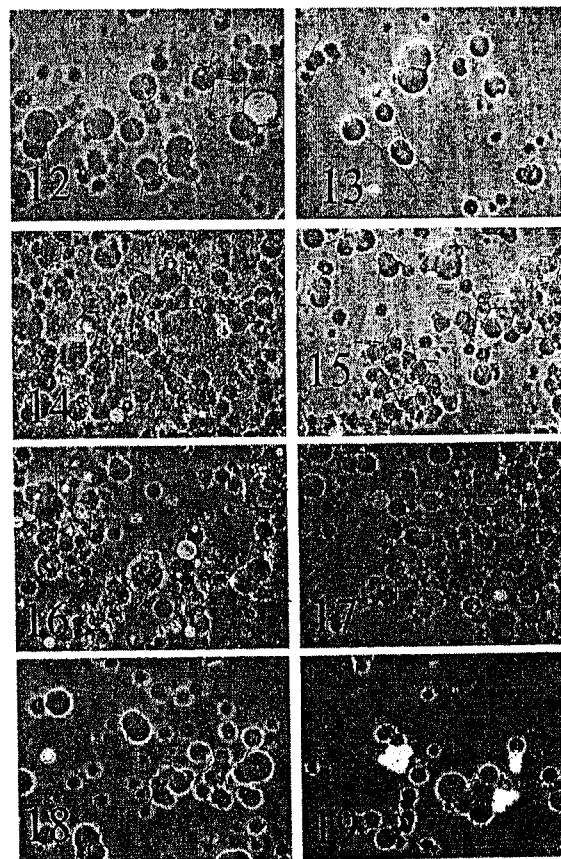


Fig. 4. Images of GPV samples 30 min after the addition of patients' plasma (group B, patients marked by 12–19). The effective angles of contact between the adhered GPVs are depicted. Small circles indicate the contribution of 360° by three adjacent effective angles of contact.

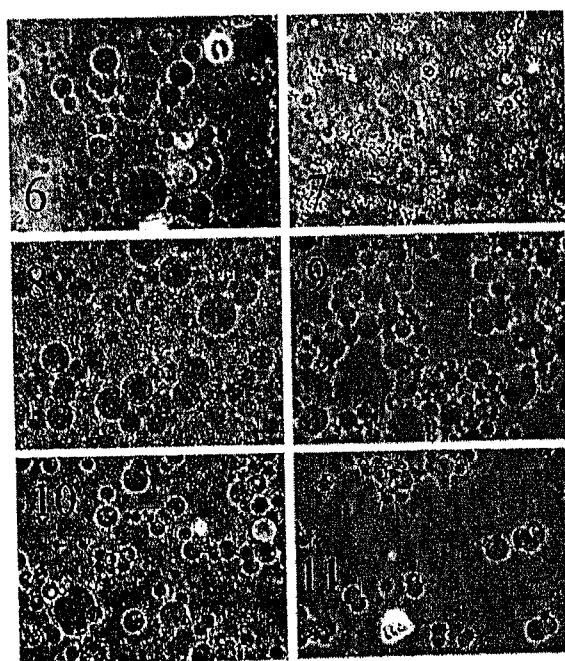


Fig. 3. Images of GPV samples 30 min after the addition of patients' plasma (group B, patients marked by 6–11). The effective angles of contact between the adhered GPVs are depicted. Small circles indicate the contribution of 360° by three adjacent effective angles of contact.

between membranes the number of MVs in the peripheral blood is smaller.

By comparing patients diagnosed with cancer (group A) with patients having other gastrointestinal diseases (group B) we found a large (140%) and statistically significant ($p=0.033$) difference between groups A and B regarding the number of MVs in peripheral blood (Table 4) while the difference between the two groups regarding the average effective angles of contact between GPVs was smaller than the difference in MVs, but still considerable (20%) and statistically significant ($p=0.013$) (Table 4). Further statistical analysis yielded power 100% for MVs at $\alpha=0.05$ while for the average effective angles of contact, the power at $\alpha=0.05$ was 90%, which is excellent regarding that there were only 5 subjects in group A. On the basis of these results we can conclude that considerable and statistically significant differences in the number of MVs and in the ability of plasma to cause adhesion of membranes between the two groups are indicated.

Discussion

The main result of this work is the correlation between the number of MVs in peripheral blood of patients and the average effective angle of contact between GPVs after the addition of plasma to GPVs (Fig. 5). Our study presents the evidence in favor of the hypothesis that plasma which mediates attractive interaction between membranes may cause suppression of microvesiculation (Fig. 1) which results in a smaller number of MVs shed from the vesiculating pool. We think that the presented results are promising, however, further studies should be

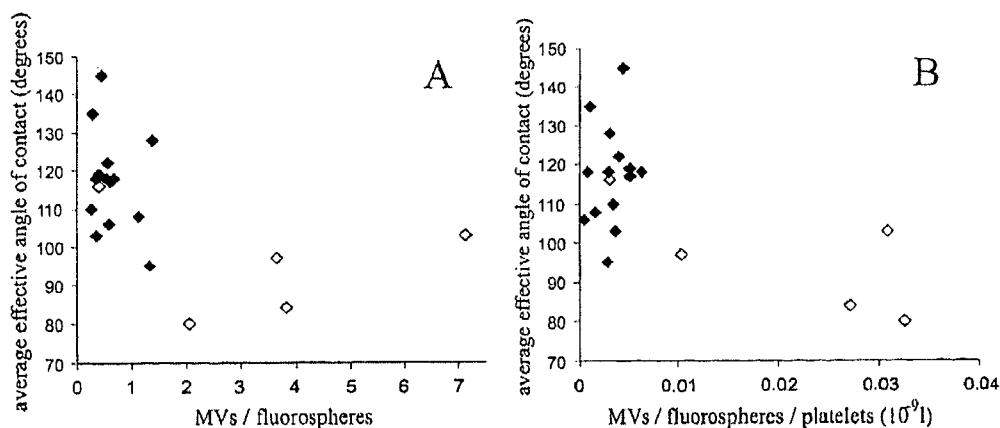


Fig. 5. Correlations between the average effective angle of contact between adhered GPVs and the number of MVs in peripheral blood (A) and between the average effective angle of contact between adhered GPVs and number of MVs per number of platelets (parameter f), for 19 patients with gastrointestinal diseases. Empty diamonds correspond to patients from group A (patients diagnosed with gastrointestinal cancer), while full diamonds correspond to patients from group B (patients with other gastrointestinal diseases).

made to validate the hypothesis. Also, some shortcomings of the method should be overcome.

We have improved the method for preparation of GPVs with respect to the previously reported study [21] where it was argued that the yield of GPVs was too low to analyse larger number of samples. In the modified protocol which was used in the present study, a larger part of lipid-coated electrodes was in contact with sucrose solution in the electroformation chamber, so that a larger lipid/solution interface is available in the electroformation process compared to previous protocol. During sedimentation samples were better protected against microorganisms. Also, according to the new protocol, the GPVs were rinsed from the electrodes by removing the electrodes from the solution and re-inserting them several times. In the previous protocol, the electrodes were rinsed by pouring the solution out of the chamber. By using the improved method of electroformation we can now test about 20 samples of plasma with the same yield of GPVs. In this respect, we could in principle analyze all samples from our study with the same yield of GPVs, however, we were limited by the limited number of blood samples that our centrifuge can process at the same time. Thus, the results presented were obtained from two experiments on two different days, whereby for each experiment, GPVs were produced a day before. In any case, even with the same yield, the population of GPVs is a heterogeneous one with respect to size and shape. Also, fluctuations in the density of GPVs in different samples are considerable (Figs. 2–4). Addition of plasma may cause bursting of GPVs, thereby diminishing their number in the sample. For each sample, we took the image at a position that was chosen previously for all samples – at the same distance from the insertion of plasma into the sample. These points may vary considerably in density of GPVs. In the future, further improvements of the method for production of GPVs and other possible methods for studying the ability of plasma to cause adhesion between membrane structures should be considered.

We have found differences in the number of MVs and in the average effective angle of contact between patients diagnosed with gastrointestinal cancer (group A) and patients diagnosed with other gastrointestinal diseases (group B) (Table 4). Four out of five patients from group A form a separate group in Fig. 5 at large numbers of MVs and small average effective angles of contact between GPVs. In these patients the disease had advanced. The patient #1 died two weeks after the blood sample was taken, due to cardiac arrest following coma. In the fifth patient with cancer who had a small number of MVs and a large average effective angle of contact (and was with respect to these parameters indistinguishable from patients from group B) the tumor was smaller and the stage of the disease was described as initial. Also note that in diagrams showing the size and the forward scatter vs. side scatter distributions of MVs (Table 1) in patients with advanced cancer there is a pronounced region of small MVs while there is no pronounced region of small MVs in the patient in whom the disease was in the initial stage. Large difference in the number of MVs between the groups A and B (Table 4) would therefore be ascribed to the advanced stage of cancer in four out of five patients from group A. These results agree with the results of Kim et al. [16] who found an increased number of MVs in peripheral blood of patients with gastric cancer with respect to normal controls, the number of MVs increasing with the advanced stage of the disease.

Circulating MV concentration is influenced by the size of the pool of cells that are able to vesiculate, the intrinsic properties of cell membranes, the mediating effect of plasma to suppress microvesiculation and the ability of the clearance system to remove MVs from circulation. In order to better assess the effect of the ability of plasma to suppress microvesiculation, it was suggested [21] that the relevant parameter would be f – the amount of MVs in plasma/the size of the vesiculating pool, rather than the concentration of MVs. In healthy subjects, the major contribution of the vesiculating pool contributing to MVs in peripheral blood (around 80%) are platelets [7], therefore in

Table 4

Statistical analysis of the differences in the number of MVs and the ability of plasma to cause coalescence of membranes (represented by the average effective angle of contact between the adhered GPVs) between patients with gastrointestinal cancer (group A) and patients with other gastrointestinal diseases (group B)

	Group A	Group B	Difference between groups	Number of subjects	Statistical significance of difference	Tails	Power at $\alpha=0.05$
MVs/fluorospheres	0.62	3.41	-2.79	21	0.033	1	1.00
st. dev.	0.35	2.5					
average effective angle of contact (degrees)	117	96	21	19	0.013	1	0.90
st. dev. (degrees)	13	14					

the pilot study [21] the parameter f was defined as the ratio between the concentration of MVs and the concentration of platelets. A better correspondence was found in the dependence of the parameter f on the average effective angle of contact than in the dependence of MV concentration on the average effective angle of contact [21]. Similar results were obtained by applying this parameter to our population: the Pearson coefficient representing the strength of the correlation was higher for the parameter f (Fig. 5B) than for the MV number (Fig. 5A, Table 4), mostly on account of the patients with diagnosed cancer. In line with the hypothesis, this indicates that also in these patients, platelets represent an important vesiculating pool. Indeed, it was found that an increased number of tissue factor-bearing MVs presumably derived from platelets was found in the blood of patients with Duke's D colorectal cancer [23]. It was suggested that the tissue factor might have been transported to the platelet membrane from monocytes and macrophages by means of MVs [24].

The tissue factor, an integral membrane protein of the vessel wall and the principal initiator of blood coagulation was found on circulating MVs [30–32]. It was reported that the tissue factor-bearing MVs originate from lipid rafts of the monocyte or macrophage membrane, the rafts being enriched in tissue factor and cholesterol [23,24,32], outlining the importance of redistribution of membrane constituents [33] and interactions between membrane constituents [34] in the budding process [8]. Moreover, tumor-released MVs of patients with cancer are likely to be involved in tumor progression [25–29]. In patients with cancer, MVs can therefore be considered as procoagulant and cancer-promoting.

Suppression of processes leading to the release of MVs into circulation may therefore be of benefit as to prevent or slow down the development of the above pathological processes.

In this work we address a physical mechanism that is common in all membranes, regardless to accompanying chemical processes which finally lead for example to blood clot formation or to protein synthesis. Suppression of the release of MVs by the mediating effect of proteins in solution adjacent to the membrane would apply to the budding of any membrane that is likely to shed MVs. Focusing on this mechanism, natural or artificial suppressors of microvesiculation would act simultaneously as anticoagulants and cancer invasive potential decelerators. It would therefore be of interest to establish which plasma constituents can mediate the attractive interaction between membranes. In previous studies, such constituent was found to be the plasma protein beta 2 glycoprotein I [19,1]. In the future, the list of possible candidates should be expanded and their clinical relevance should be assessed.

Conclusion

Results involving a population of 19 patients with gastrointestinal diseases support the hypothesis that in subjects with plasma which induces a more pronounced adhesion between membranes the number of MVs in the peripheral blood is smaller.

Acknowledgments

We thank V. Kralj-Iglič, B. Rozman and R. Jerala for fruitful discussions, B. Mavčič for help with statistical analysis and K. Schara, M. Lokar, V. Janša, N. Tomšič and J. Pavlič for help with experiments. Collaboration with J. Košir, J. Polanec and S. Ukmar is acknowledged. Experiments with giant phospholipid vesicles were performed at the Institute of Cell Biology, Faculty of Medicine, University of Ljubljana for which we are indebted to K. Jezernik and P. Veranič. Electroformation of vesicles was performed at the Laboratory of Biocybernetics, Faculty of Electrical Engineering, University of Ljubljana, for which we thank D. Miklavčič and M. Kandušar. This work was supported by the Slovenian Research Agency, grants No. P2-0232-1538 and J3-9219-0381-06.

References

- [1] J. Urbanija, N. Tomšič, M. Lokar, A. Ambrožič, S. Čučnik, M. Kandušar, B. Rozman, A. Iglič, V. Kralj-Iglič, Coalescence of phospholipid membranes as a possible origin of anticoagulant effect of serum proteins, *Chem. Phys. Lipids* 150 (2007) 49–57.
- [2] D. Choi, J. Lee, G.W. Park, H. Lim, J.Y. Bang, Y. Kim, K. Kwon, H.J. Kwon, K.P. Kim, Y.S. Cho, Proteomic analysis of microvesicles derived from human colorectal cancer cells, *J. Proteome Res.* 6 (2007) 4646–4655.
- [3] M. Diamant, M.E. Tushuizen, A. Sturk, R. Nieuwland, Cellular microparticles: new players in the field of vascular disease? *Eur. J. Clin. Invest.* 34 (2004) 392–401.
- [4] J.H.W. Distler, D.S. Pisetsky, L.C. Huber, J.R. Kalden, S. Gay, O. Distler, Microparticles as regulators of inflammation, *Arthritis Rheum.* 52 (2005) 3337–3348.
- [5] T.J. Greenwalt, The how and why of exocytic vesicles, *Transfusion* 46 (2006) 143–152.
- [6] J. Ratajczak, M. Wysoczynski, F. Hayek, A. Janowska-Wieczorek, M.Z. Ratajczak, Membrane-derived microvesicles: important and underappreciated mediators of cell to cell communication, *Leukemia* 20 (2006) 1487–1495.
- [7] M. Diamant, R. Nieuwland, R.F. Pablo, A. Sturk, J.W. Smit, J.K. Radder, Elevated numbers of tissue-factor exposing microparticles correlate with components of the metabolic syndrome in uncomplicated type 2 diabetes mellitus, *Circulation* 106 (2002) 2442–2447.
- [8] H. Hagerstrand, L. Mrowczynska, U. Salzer, R. Prohaska, A.K. Michelsen, V. Kralj-Iglič, A. Iglič, Curvature dependent lateral distribution of raft markers in the human erythrocyte membrane, *Mol. Membr. Biol.* 23 (2006) 277–288.
- [9] J.A. Lopez, I. Del Conde, C.N. Shrimpton, Receptors, rafts, and microvesicles in thrombosis and inflammation, *J. Thromb. Haemost.* 3 (2005) 1737–1744.
- [10] N. Coltel, V. Combes, S.C. Wassmer, G. Chimini, G.E. Grau, Cell vesiculation and immunopathology: implications in cerebral malaria, *Microbes Infect.* 8 (2006) 2305–2316.
- [11] R.J. Berckmans, R. Nieuwland, P.P. Tak, A.N. Böing, F.P. Romijn, M.C. Kraan, F.C. Breedveld, C.E. Hack, A. Sturk, Cell-derived microparticles in synovial fluid from inflamed arthritic joints support coagulation exclusively via a factor VII-dependent mechanism, *Arthritis Rheum.* 46 (2002) 2857–2866.
- [12] F. Dignat-George, L. Camoin-Jau, F. Sabatier, D. Arnoux, F. Anfosso, N. Bardin, V. Velt, V. Combes, S. Gentile, V. Moal, M. Sanmarco, J. Sanpoh, Endothelial microparticles: a potential contribution to the thrombotic complications of the antiphospholipid syndrome, *Thromb. Haemost.* 91 (2004) 667–673.
- [13] O. Morel, L. Jesel, J.M. Freyssinet, F. Toti, Elevated levels of procoagulant microparticles in a patient with myocardial infarction, antiphospholipid antibodies and multifocal cardiac thrombosis, *Thromb. J.* 3 (2005) 15.
- [14] A. Ginestra, M.D. La Placa, F. Saladini, D. Cassara, H. Nagase, M.L. Vittorelli, The amount and proteolytic content of vesicles shed by human cancer cell lines correlates with their in vitro invasiveness, *Anticancer Res.* 18 (1998) 3433–3437.
- [15] K. Shedden, X.T. Xie, P. Chandaroy, Y.T. Chang, G.R. Rosania, Expulsion of small molecules in vesicles shed by cancer cells: association with gene expression and chemosensitivity profiles, *Cancer Res.* 63 (2003) 4331–4337.
- [16] H.K. Kim, K.S. Song, Y.S. Park, Y.H. Kang, Y.J. Lee, K.R. Lee, H.K. Kim, K.W. Ryu, J.M. Bae, S. Kim, Elevated levels of circulating platelet microparticles, VEGF, IL-6 and RANTES in patients with gastric cancer: possible role of a metastatic predictor, *Eur. J. Cancer* 39 (2003) 184–191.
- [17] D.D. Taylor, H.D. Homesley, G.J. Doellgast, Membrane-associated immunoglobulins in cysts and ascites fluids of ovarian cancer patients, *Am. J. Reprod. Immunol.* 3 (1983) 7–11.
- [18] A. Ambrožič, B. Božič, T. Kveder, J. Majhenc, V. Arrigler, S. Svetina, B. Rozman, Budding, vesiculation and permeabilization of phospholipid membranes—evidence for a feasible physiologic role of beta2-glycoprotein I and pathogenic actions of anti-beta2-glycoprotein I antibodies, *Biochim. Biophys. Acta* 1740 (2005) 38–44.
- [19] A. Ambrožič, S. Čučnik, N. Tomšič, J. Urbanija, M. Lokar, B. Babnik, B. Rozman, A. Iglič, V. Kralj-Iglič, Interaction of giant phospholipid vesicles containing cardiolipin and cholesterol with beta2-glycoprotein-I and anti-beta2-glycoprotein-I antibodies, *Autoimmun. Rev.* 6 (2006) 10–15.
- [20] K. Bohinc, A. Iglič, S. May, Interaction between macroions mediated by divalent rod-like ions, *Europhys. Lett.* 68 (2004) 494–500.
- [21] M. Frank, M. Manžek-Keber, M. Kržan, S. Sodini-Semrl, R. Jerala, A. Iglič, B. Rozman, V. Kralj-Iglič, Prevention of microvesiculation by adhesion of buds to the mother cell membrane — a possible anticoagulant effect of plasma, *Autoimmunity Rev.* 7 (2008) 240–245.
- [22] M.I. Angelova, S. Soléau, P. Méléard, J.F. Faucon, P. Bothorel, Preparation of giant vesicles by external AC electric fields. Kinetics and applications, *Prog. Colloid & Polym. Sci.* 89 (1992) 127–131.
- [23] G. Hron, M. Kollars, H. Weber, V. Sagaster, P. Quehenberger, S. Eichinger, P.A. Kyrle, A. Weltermann, Tissue factor positive microparticles — cellular origin and association with coagulation activation in patients with colorectal cancer, *Thromb. Haemost.* 97 (2007) 119–123.
- [24] I. Müller, A. Klocke, M. Alex, M. Kotzsch, T. Luther, E. Morgenstern, S. Ziesenis, S. Zahler, K. Preissner, B. Engelmann, Intravascular tissue factor initiates coagulation via circulating microvesicles and platelets, *FASEB J.* 17 (2003) 476–478.
- [25] M. Baj-Krzyworzeka, R. Szatanek, K. Weglarczyk, J. Baran, B. Urbanowicz, P. Branski, M.Z. Ratajczak, M. Zembala, Tumour-derived microvesicles carry several surface determinants and mRNA of tumour cells and transfer some of these determinants to monocytes, *Cancer Immunol. Immunother.* 55 (2005) 808–818.
- [26] A. Janowska-Wieczorek, M. Wysoczynski, J. Kijowski, L. Marques-Curtis, B. Machalinski, J. Ratajczak, M.Z. Ratajczak, Microvesicles derived from activated platelets induce metastasis and angiogenesis in lung cancer, *Int. J. Cancer* 113 (2005) 752–760.

- [27] A. Janowska-Wieczorek, L.A. Marquez-Curtis, M. Wysoczynski, M.Z. Ratajczak, Enhancing effect of platelet-derived microvesicles on the invasive potential of breast cancer cells, *Transfusion* 46 (2006) 1199–1209.
- [28] V. Huber, S. Fais, M. Iero, L. Lugini, P. Canese, P. Squarcina, A. Zaccheddu, M. Colone, G. Arancia, M. Gentile, E. Seregni, R. Valenti, G. Ballabio, F. Belli, E. Leo, G. Parmiani, L. Rivoltini, Human colorectal cancer cells induce T-cell death through release of proapoptotic microvesicles: role in immune escape, *Gastroenterology* 128 (2005) 1796–1804.
- [29] R. Valenti, V. Huber, M. Iero, P. Filipazzi, G. Parmiani, L. Rivoltini, Tumor-released microvesicles as vehicles of immunosuppression, *Cancer Res.* 67 (2007) 2912–2915.
- [30] U. Rauch, S. Antoniak, Tissue factor-positive microparticles in blood associated with coagulopathy in cancer, *Thromb. Haemost.* 97 (2007) 9–18.
- [31] I. del Conde, L.D. Bharwani, D.J. Dietzen, U. Pendurthi, P. Thiagarajan, J.A. Lopez, Microvesicle-associated tissue factor and Trousseau's syndrome, *J. Thromb. Haemost.* 5 (2007) 70–74.
- [32] I. del Conde, C.N. Shrimpton, P. Thiagarajan, J.A. Lopez, Tissue-factor-bearing microvesicles arise from lipid rafts and fuse with activated platelets to initiate coagulation, *Blood* 106 (2005) 1604–1611.
- [33] V. Kralj-Iglič, V. Heinrich, S. Svetina, B. Žekš, Free energy of closed membrane with anisotropic inclusions, *Eur. Phys. J. B* 10 (1999) 5–8.
- [34] V. Kralj-Iglič, P. Veranič, Curvature-induced sorting of bilayer membrane constituents and formation of membrane rafts, in: A. Leitmanova Liu (Ed.), *Advances in Planar Lipid Bilayers and Liposomes*, vol. 5, Elsevier, 2007, pp. 129–149.

Bringing Click-to-Dial Functionality to IPTV Users

Urban Sedlar, Luka Zebec, Janez Bester, and Andrej Kos, *University of Ljubljana*

ABSTRACT

Architecture for click-to-dial over IP television is discussed, utilizing Parlay X Web services to communicate with telephony network elements. The architecture and individual elements of the system are presented in detail, including a gateway to call servers and telephony switches. The implemented system enables IPTV users to establish calls over mobile or fixed phones by pushing a button on their IPTV remote control. Additionally, we describe a method for automatic phone number recognition from live video that communicates the data to IPTV users via Web services architecture, enabling them to initiate a call easily to any number displayed on the video.

INTRODUCTION

The idea to transmit a video signal over an IP network, dubbed IP television, has existed for years, but only lately has it become a prevailing trend and a strong focus of telecom operators, Internet service providers (ISPs), and content providers alike.

IP television can denote either a closed infrastructure video transmission system, managed by an operator (usually called IPTV), or any system that can deliver TV content over the public Internet (commonly termed Internet television). With Internet television, anyone can participate as a content publisher, service provider, or a consumer on a global basis. On the other hand, the IPTV communication channel is managed by an operator to provide adequate quality of service. Thus, currently, the resulting quality of experience is much better and more suited for lean-back living room consumption and can directly compete with existing means of TV broadcasting; whereas Internet television remains in the domain of technically more advanced users. Because of that and for the reason of better standardization, our focus will be mainly on IPTV.

IPTV offers many advantages over terrestrial, satellite, and cable television, most notably an inherent return channel, which paves the way for interactive services.

Due to demanding network and hardware requirements, there are significant costs involved in the deployment of an IPTV system. Integration with existing back-office systems increases

its complexity and can affect the speed and success of the deployment.

The wide availability of digital subscriber line (DSL) technologies in recent years has helped alleviate some of the problems and has sped up IPTV evolution significantly by reducing infrastructure costs.

Innovative services are becoming the key differentiator between competing service providers. Many proposals and pilot deployments of such services have emerged, ranging from electronic program guides (EPGs) and simple games to social network-based and community-driven services over TV. To our knowledge, however, no generally available telephony support for IPTV users exists at present. The reason behind this appears to be limited capability of customer premises equipment. Although voice-over-IP (VoIP) and even video conferencing functionalities are already available directly through terminal equipment, this approach involves notable economic issues.

At the same time, an independent telephony system is already in place, usually owned by a different service provider. With the support for open Web services-based interfaces, the existing telephony system is becoming easier to control by third party application logic. One such standard interface is Parlay X, which facilitates application development and enables developers to leverage their applications with an entire existing telephony network and its users.

We present and discuss a solution, which was implemented without significant costs, and ties the existing IPTV and telephony systems together. It enables IPTV users to easily establish calls over mobile or fixed phones with the push of a button on their remote control.

The solution is based on Parlay X and computer supported telecommunications applications (CSTA) standards, the latter being a common interface to existing telephony systems.

This article is organized as follows: first, the general architecture of IPTV systems is presented followed by an introduction to telephony systems. Next, we describe selected interfaces to the telephony system, including Parlay X and CSTA. The architecture of the converged system is presented, along with generic Parlay X gateway architecture and a presentation of automatic video-stream object recognition. Finally, possible usage scenarios and implementation details of a click-to-dial solution for IPTV users are presented.

IPTV SYSTEMS

IPTV is defined as multimedia services, such as television, video, audio, text, graphics, and data delivered over IP-based networks that are managed to provide the required level of quality of service (QoS)/quality of experience (QoE), security, interactivity, and reliability [1].

Transmission of video streams over the network is an essential functionality of an IPTV system. It serves as a basis for elementary services of broadcast TV, content-on-demand, and network personal video recorders and enables numerous supplementary services.

The general architecture of an IPTV system consists of four entities: content provider, service provider, network provider, and the customer.

The content provider owns the rights to the content and delivers it to the service provider, either as a live stream or in a pre-recorded form.

The IPTV service provider delivers the content to end users by sending the data from the head-end over an IP-based network. To facilitate the transmission, the content must be compressed; this is done by encoders that compress live video signal in real-time. If the content is offered on-demand, it is usually stored on a server in a form suitable for streaming over the network. In addition to content preparation and management, the service provider must take adequate security measures to protect the content from improper use (digital rights management) and to protect its users from fraud, as well as to provide a means for service management, billing, and customer relationship management. Complexity of the entire workflow can be reduced by a software system called *middleware* that serves as an abstraction layer over the IPTV system. It connects different components of the system and provides high-level control to administrators and application programming interfaces (APIs) to the developers.

The network provider owns the communication channel between the service provider and the end users [2]. The network must be managed to meet QoS requirements and is usually segmented into core, metro, and access networks. Broadcast TV is transmitted to end users using IP multicast, which reduces bandwidth requirements in the core and access networks. Content-on-demand is transmitted by unicasting it to individual users. Although the network provider and IPTV service provider are technically independent, it is not uncommon to find them together in a single ISP.

After the video stream is sent over the network to the customer, it must be properly received, decoded, and displayed. Although this can be done by any networked computer with adequate processing power to decode video, usually a separate device called a set-top box (STB) is used. This acts as an intermediate between the IP network and an analog or digital video input of the TV receiver. Apart from video decoding capability, the STB requires a client application to communicate with the IPTV system components at the head-end and manage the visual presentation of the content on the screen. This application can be regarded as an extension of the middleware to the STB.

Many separate areas of IPTV systems are covered by standards, from networking to video encoding, and a number of standardizing bodies are involved. Due to the lack of an overall framework, encompassing all IPTV system components, the International Telecommunication Union — Telecommunication Standardization Sector (ITU-T) IPTV Focus Group was established with the aim of coordinating the development of global IPTV standards, taking into account the existing work of the standardizing bodies.

IPTV systems are currently in commercial operation through a number of telecommunications service providers worldwide, whereas many are involved in trial deployments and the introduction of new services. The most commonly used video encoding is MPEG-2, which is gradually being replaced by H.264/MPEG-4 advanced video coding (AVC) or in some cases, Microsoft's proprietary VC-1. Both have approximately two times better compression efficiency than MPEG-2. Codecs are network agnostic, tied only to the choice of encoders and decoding support of the terminal equipment.

Middleware offers even more possibilities than video encoding; because it can be implemented from scratch in simple pilot deployments, many solutions exist. However, there is a notable trend in the growth in popularity of the hypertext markup language (HTML) and JavaScript-based presentation engine, as it allows backward compatibility with standard Web sites and rapid service development using off-the-shelf development tools and existing Web development frameworks.

TELEPHONY SYSTEM

Today, telephony systems can be viewed as a mix of traditional and emerging networks, including PSTN, ISDN, GSM, VoIP, GPRS, UMTS, cdma2000, HSDPA, and WiMAX, where gateways are used to provide connectivity and interoperability. The hybrid network of today is expected to converge to IP on the network layer.

Although the majority of classic telephony services still reside on the call control layer, with only a few of them dislocated in special network elements, newer services usually are provided using application servers and service gateways.

The traditional approach to service development and deployment was based on the concepts of intelligent networks (IN) and customized applications for mobile networks enhanced logic (CAMEL). These architectures used intelligent network application part (INAP), CSTA, Camel application part (CAP), mobile application part (MAP), and other protocols. This approach has proven to be stable and reliable, and currently, there are many systems deployed that are based on these architectures. However, a thorough understanding of the protocols and architectures is required for service development, which was previously always kept in the domain of telecom vendors and operators. Thus, only a small number of innovative and converged services were developed and used in the past.

Next generation network (NGN) architecture presents a different approach, with the broad-

There is a notable trend in the growth in popularity of HTML and JavaScript-based presentation engine, as it allows backward compatibility with standard Web sites and rapid service development using off-the-shelf development tools and existing Web development frameworks.

A Web service is a collection of standards and protocols that enables interoperable machine-to-machine interaction and data exchange over a network. It is standardized by the World Wide Web Consortium.

band QoS-enabled IP network at its core [3]. Such networks can support all services, because in NGN, service-related functions are independent from underlying transport-related technologies.

NGN service control is based on concepts such as IP multimedia subsystem (IMS) [4] and Parlay, using different protocols and interfaces, including Web services, Session Initiation Protocol (SIP), common object request broker architecture (CORBA), and Java. Their main characteristic is openness. Together with the decoupling of service and transport layers, this enables flexible and easy service development and deployment for both service providers and third party developers.

INTERFACES

Three interfaces are presented: Web services in general as a basis for Parlay X, Parlay X itself, and the CSTA standard as an introduction to the gateway implementation described in the following sections.

WEB SERVICES

A Web service is a collection of standards and protocols that enables interoperable machine-to-machine interaction and data exchange over a network. It is standardized by the World Wide Web Consortium (W3C) [5].

The interface for communication is described in the Web services description language (WSDL). The communication itself is performed by exchanging Simple Object Access Protocol (SOAP) messages [6] that are based on the extensible markup language (XML) format. The messages usually are transferred over Hypertext Transfer Protocol (HTTP), although other underlying protocols can be used as well.

To facilitate Web service discovery, universal description discovery and integration (UDDI) protocol was developed, which enables the creation of Web service directories.

PARLAY X

Parlay X is one of the standards for NGN service development. It has been standardized by the European Telecommunications Standards Institute (ETSI) [7] and offers a set of simple-to-use, high-level Web services-based APIs that enable applications to access the telephony network functionalities. Being one of the service deployment choices for the emerging all-over-IP IMS network, Parlay X is, in our opinion, one of the most promising concepts for the future of service delivery. Due to the abstract nature of Parlay X APIs, they are not tied to any specific telephony protocol and can be used in traditional, NGN, and IMS networks.

The APIs usually are implemented on a gateway and positioned between application servers and network elements. The communication between applications and the Parlay X interface of the gateway is performed by exchanging SOAP messages over the HTTP protocol.

Parlay X version 2.1 consists of 13 Web services, including third party call, that enable call initiation and can be used for click-to-dial applications and audio call, which introduces a flexi-

ble way to provide vocal message delivery using text-to-speech, audio playback, or rendering voice XML.

The remaining Web services are mentioned only for consistency, as their names are sufficiently self-explanatory: call notification, call handling, short messaging service, multimedia messaging service, multimedia conference, presence, address list management, terminal status, terminal location, payment, and account management.

CSTA

CSTA is a European Computer Manufacturers Association (ECMA) standard, intended for use in computer telephony integration (CTI). It provides an abstraction layer for telecommunications applications that is independent of underlying protocols and devices. CSTA exposes advanced features of a communications platform to applications developers, providing them with comprehensive access to the call handling mechanisms.

Currently, three versions of the CSTA standard exist. Phase I provides basic call control services [8], whereas Phase II and III further extend these services. To the best of our knowledge, currently, CSTA Phase I prevails in real-life implementations.

CSTA protocol messages are structured using abstract syntax notation - one (ASN.1). In addition to these binary encoded messages, CSTA Phase III also enables a Web service-based implementation that covers all Phase III features and facilitates their use by Web developers. The WSDL for CSTA Phase III is specified in the ECMA-348 standard and builds upon the XML data types and messages (CSTA XML) defined in ECMA-323 [9].

One of the main reasons why not many pure CSTA applications are available is the required low-level conceptual understanding of telephony systems architecture and functionalities. This was further aggravated by the complexity of the ASN.1 scheme before the introduction of CSTA XML.

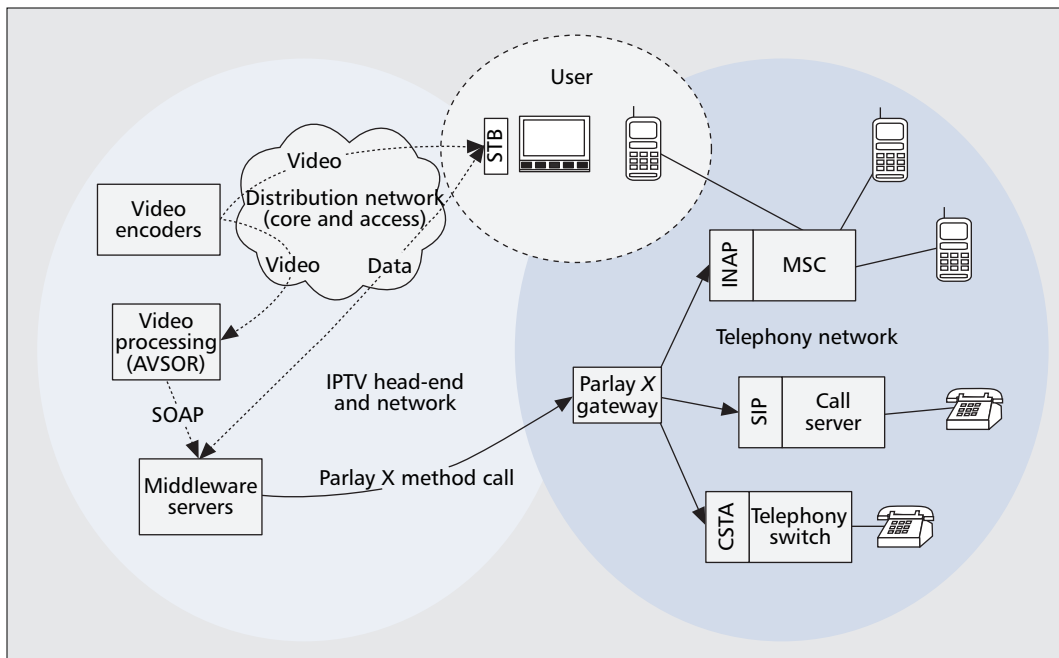
Furthermore, although being more powerful for telephony applications than Parlay X, CSTA does not address many of today's popular telecom services, such as presence and video streaming.

SYSTEM ARCHITECTURE

Architecture for connecting an IPTV system to the telephony system, allowing an IPTV user to establish a call between her mobile or fixed telephone and any valid phone number, is presented in Fig. 1.

Integration of both domains is based on a gateway that provides connectivity to traditional telephony network elements using Web services-based Parlay X API. A server-side system for extracting telephone numbers from a live video stream was added and enables users to easily initiate a call to any number displayed on the TV screen.

It should be noted that the described architecture also could be adapted to apply to Internet television; however, modifications would



■ Figure 1. IPTV click-to-dial system architecture.

have to be made because of different video delivery system architecture.

PARLAY X GATEWAY ARCHITECTURE

Due to the fact that traditional telephony interfaces are much more widely available than the highly abstracted Parlay X API, a translating gateway is required to allow a Parlay X client to connect to telephony network elements. The general architecture of the gateway is shown in Fig. 2 and is described using our preferred implementation.

The gateway comprises a Parlay X framework for open Web services-based communication and a generic telephony framework (GTF) that ultimately provides interfaces to supported traditional interfaces, such as CSTA.

The Parlay X framework consists of three sub-modules. The *Parlay X interface* presents an entry point for SOAP Parlay X method calls. These calls are realized through communication with either of the remaining two modules, depending on the nature of the Parlay X method. The module, *ActiveCalls*, provides a store for information about currently active calls and is accessed from the interface module in the case when only information about an active call is needed. If the Parlay X method requires a message to be sent to the network element API through the protocol adapter, the request is forwarded to the GTF using a GTF client module. This happens by a remote procedure call (RPC) in the server module of the GTF.

The GTF serves as an abstraction layer for generic telephony functionalities. It comprises two main sub-modules and an arbitrary number of protocol adapters. The GTF server receives method calls from the client in the Parlay X framework, processes them, and forwards them to the appropriate protocol adapter using request routing. A protocol adapter provides mappings to the supported protocol. Each

adapter module runs as a separate thread and communicates with the corresponding traditional interface of a network element, which can be a call server, telephony switch, class5 switch, or a mobile switching center (MSC).

AUTOMATIC OBJECT RECOGNITION

IPTV click-to-dial architecture can be highly leveraged by the automated means of phone number recognition from live video. This is accomplished by detecting a general set of objects (letters, symbols) in the video. We define a concept of automatic video stream object recognition (AVSOR) that allows analysis of the video stream and extraction of desired information using appropriate input modules. We focus on the recognition of text, allowing immediate application in the proposed system.

In general, detecting and recognizing text appearing as visual images is termed optical character recognition (OCR). Recognition of typewritten text on a uniform background has already become highly reliable and is used routinely to scan books and documents.

Solutions for recognizing text in video streams rely on the decomposition of video into individual frames; for sequences of static images, a standard OCR engine can be used. However, the textual information on complex (non-homogenous) backgrounds is difficult to recognize without appropriate preprocessing [10]. Different detection and recognition techniques must be used to provide satisfactory results.

The AVSOR concept has additional uses in the IPTV domain beyond telephone number recognition, such as automatic subtitle translation and bar code recognition that would enable a *lean-back* experience of product bookmarking using a remote control.

To communicate the telephone numbers detected by AVSOR to the middleware server, we define and use a Web service for its ease of

Due to the fact that traditional telephony interfaces are much more widely available than the highly abstracted Parlay X API, a translating gateway is required to allow a Parlay X client to connect to telephony network elements.

implementation and cross-platform compatibility. By doing so, we obtain an open, formally described interface that can be accessed easily from a variety of development platforms. Because Web services can be regarded as a *lingua franca* of the Internet, the interoperability between AVSOR and different middleware server implementations can be supported easily.

SYSTEM INTEGRATION

The components are integrated as seen in Fig. 1. At the IPTV head-end, a video processing system receives the selected video streams over the core network of the provider. Video streams are analyzed using the AVSOR method, and occurrences of telephone numbers are communicated to the middleware server using a Web service. The application logic of the middleware server stores the number in a database the moment it is received. From the database, the numbers are transmitted to any STB client application requesting them. As the STB client application receives the number, it renders it on-screen and enables the user to initiate the call. In case the user chooses to ignore it, it disappears shortly

after it is not detected in the video anymore. If the user initiates the call, the middleware server is notified first. Its application logic triggers a Parlay X method call on the Parlay X gateway, which translates the request to the selected protocol and sends it to a telephony network element. The network element consequently establishes the call and reports its status back to the gateway, which always holds the latest information about the status of the call.

When the client application of the STB requests the status of the call to refresh the information displayed on-screen, the request is received again by the middleware server. The latter calls a different Parlay X method on the gateway, inquiring about the status of the call. The gateway responds immediately, providing the status of the call from internal memory.

SCENARIOS

An integrated system allows different usage scenarios that are all based on a single underlying mechanism, presented in Fig. 3. The call is initiated by the user through the IPTV system and returns back to the user through the telephony system.

In the following section, we describe three real-life scenarios that illustrate the use of integrated IPTV click-to-dial architecture.

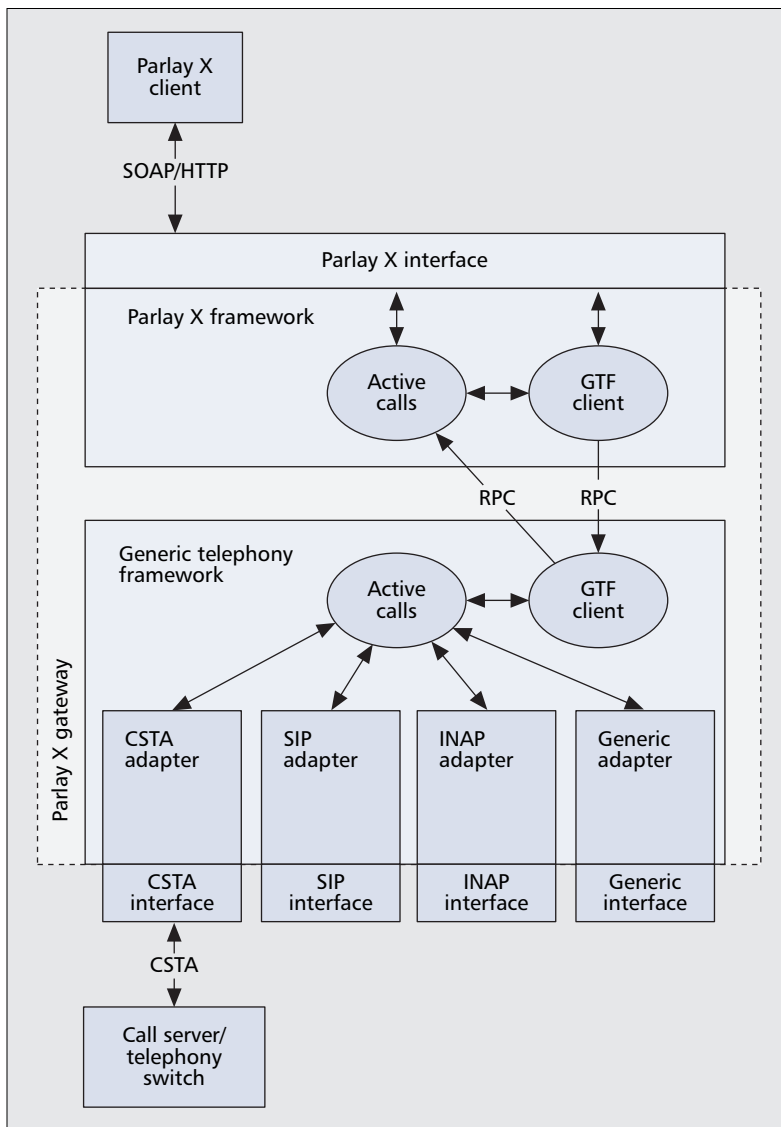
Alice, an IP television user, is watching her favorite TV show and wishes to respond to a survey by dialing a number on the screen and casting a vote. When a phone number is shown in the TV station's on-screen graphic, a large *dial* button appears next to it. As soon as Alice pushes the dial button using her remote control, a call is initiated to the number she provided earlier. Her phone rings and when she answers it, a connection to the voting service is established. The voting service terminates the connection as soon as the vote is recorded.

Bob, watching a different channel, where a call-in show is running, wishes to dial a number that is displayed in the on-screen graphic in order to state an opinion. By using his remote control and pushing the dial button, Bob establishes a connection and can monitor the status and duration of the call on the screen, overlaying the TV picture. He decides to dismiss the on-screen status display and resumes watching while he talks on the phone.

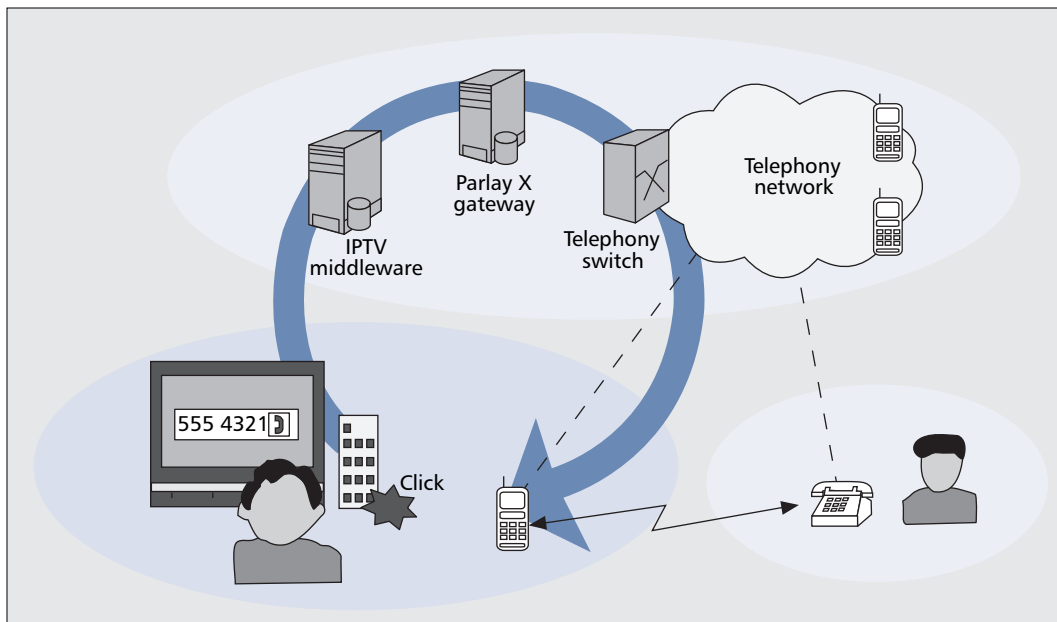
Charlie catches a glimpse of a TV advertisement while channel surfing that captures his attention and on the spur of the moment, he decides he wants to order the product by dialing the phone number in the advertisement. The advertisement disappears, but the dial button remains on the screen for several seconds and enables Charlie to establish the call to the previously displayed number. When the call is established, Charlie monitors its status on the TV screen; later he decides to terminate the call using his remote control.

IMPLEMENTATION

A working implementation of the click-to-dial system for SiOL IPTV users was created, based on an Iskratel SI2000 telephony switch with a



■ Figure 2. Parlay X gateway architecture.



■ Figure 3. IPTV click-to-dial scenario.

A Parlay X gateway with modular architecture was developed in Java and supported with a Tomcat Web server and an Axis framework for Web services integration. In addition, a single protocol adapter was implemented, providing a functional Parlay X to CSTA transformation.

CSTA interface. To communicate with the switch, the overall gateway architecture with a selected protocol adapter was implemented. Additionally, the AVSOR system for phone number detection was assembled, which feeds the numbers to the middleware server. A modified STB client application displays the numbers on the screen. The details of individual elements are presented in the following sections.

PARLAY X/CSTA GATEWAY

A Parlay X gateway with modular architecture was developed in Java and supported with a Tomcat Web server and an Axis framework for Web services integration. In addition, a single protocol adapter was implemented, providing a functional Parlay X to CSTA transformation.

The mappings between Parlay X and CSTA protocols are based on similarities that can be identified between the call control functionalities of both protocols [11]. These include *third party call*, *call notification* (including *call direction*), and a limited set of *call handling* and *multimedia conference* functions. The implemented protocol adapter provides mappings of Parlay X Web services onto appropriate CSTA methods, as seen in Table 1. Parlay X method calls are translated to ASN.1 messages, understood by CSTA-enabled network elements.

The implemented CSTA protocol adapter is only one of the possibilities. It should be noted, however, that telephony network elements are interconnected by a signaling network that allows the requests to be further transferred to elements with no compatible interfaces.

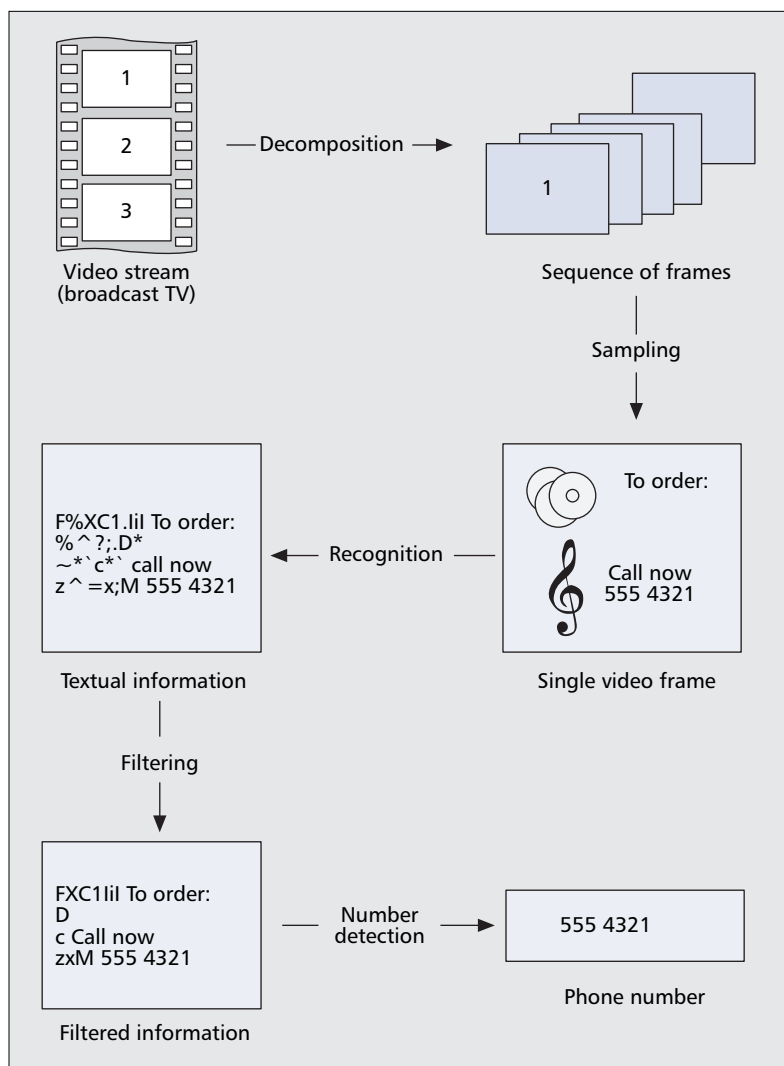
PHONE NUMBER DETECTION

The OCR system is used as a basis for our AVSOR implementation. The system is implemented using freely available software solutions (*mplayer*, *djpeg*, and *tesseract*) as seen in Fig. 4. To reduce processing requirements for the OCR,

Parlay X method	CSTA service
Third party call Web service	
makeCall	Monitor Start Make Call
getCallInformation	Call Event Reports
endCall	Clear Connection Monitor Stop
cancelCall	
Multimedia conference Web service	
createConference	/
inviteParticipant	Monitor Start Make Call Consultation Call Conference Call
getConferenceInfo	Call Event Reports
endConference	Clear Connection Monitor Stop
Call handling Web service	
setRules	Monitor Start Set Feature
clearRules	Monitor Stop

■ Table 1. Parlay X to CSTA mappings.

a single channel is selected, and the video is sampled with a temporal resolution of two seconds (at the rate of 25 frames/s, one frame in 50 is kept). Since phone numbers usually are displayed for longer than that, this is assumed to be



■ **Figure 4.** Automatic video stream object recognition for phone number detection.



■ **Figure 5.** TV capture of implemented interface

enough for detection. Each acquired frame is fed to the OCR engine. No prior preprocessing is performed to further reduce the complexity of the process. After the extraction of textual data, non-alphanumeric characters are filtered out. This removes a substantial amount of noise caused by misinterpretations of the image due to the absence of preprocessing. Finally, the filtered data is matched against the most common phone number formats for a local region. Valid phone numbers obtained in the process are pushed to the middleware server using a Web service to invoke a server-side method that stores the number in the database, together with an exact timestamp of its appearance.

IPTV MIDDLEWARE INTEGRATION

The presented architecture is tied to the network of one of the world's first IPTV service providers (SiOL), which uses an HTML and JavaScript-based presentation engine as part of its middleware. The communication between middleware servers and the client is based on HTTP serving simple HTML pages, enriched with client-side scripting code to provide interactivity.

Client-side application logic on the STB is modified to include a separate module for click-to-dial functionality, performing periodic lookups to the middleware server, where the time-stamped telephone number information is stored in the database. When a number is obtained from the database, the client-side application logic displays it in a large semitransparent rectangle on the television screen, along with a clickable dial button (Fig. 5). At the press of the button using a remote control, a method on the middleware server is called that relays the method call to the Parlay X gateway, invoking the *makeCall* method.

After the call is initiated, the client-side logic in the STB polls the middleware server at periodic intervals for the status of the call. Each call is forwarded to the Parlay X gateway, invoking a *getCallInformation* method there and returning the call status to the middleware server, where the relevant information is extracted. Finally, the information is returned to the STB client, where it is rendered on-screen to reflect the status of the call.

CONCLUSIONS

IP television is currently a hot topic with service providers. Presenting a convergent platform for multimedia and communication services, it also allows integration with independent systems such as the telephony system and smart environments. In this paper, an innovative service is presented based on a novel approach for providing IPTV users with click-to-dial telephony using an existing telephony system. It is further improved by incorporating automatic number recognition from video streams. The key communication paths of the system are based on Web services architecture, which allows cross-platform and language-neutral access to data and applications over the Internet. We believe that the described application presents only the beginning in the era of convergent IPTV applications.

ACKNOWLEDGEMENTS

The authors would like to thank project partner companies, Iskratel and SiOL.

This work was supported in part by the Center of Excellence for Information and Communication Technologies; the Ministry of Higher Education, Science, and Technology of Slovenia; and the Slovenian Research Agency.

REFERENCES

- [1] ITU-T IPTV Focus Group; <http://www.itu.int/ITU-T/IPTV/>
- [2] S. Vedantham, S.-H. Kim, and D. Kataria, "Carrier-Grade Ethernet Challenges for IPTV Deployment," *IEEE Commun. Mag.*, vol. 44, no. 6, July 2006.
- [3] M. Carugi, B. Hirschman, and A. Narita, "Introduction to the ITU-T NGN Focus Group Release 1: Target Environment, Services, and Capabilities," *IEEE Commun. Mag.*, vol. 43, no. 10, Oct. 2005.
- [4] 3GPP, "Technical Specification Group Services and System Aspects: TS 23.228: IP Multimedia Subsystem (IMS)."
- [5] Web Services Architecture; <http://www.w3.org/TR/ws-arch/>
- [6] SOAP; <http://www.w3.org/TR/soap/>
- [7] Parlay X Working Group, "Parlay X Web Services Specification, Version 2.1," June 2006.
- [8] Services for Computer Supported Telecommunications Applications (CSTA) Phase I, ECMA-179, June 1992.
- [9] Web Services Description Language (WSDL) for CSTA Phase III, 3rd edition, ECMA-348, Dec. 2006.
- [10] R. Lienhart, "Video OCR: A Survey and Practitioner's Guide," *Video Mining*, Kluwer, Oct. 2003, pp. 155–84.
- [11] L. Zebec *et al.*, "NGN Service Development — Overview and Parlay X Implementation," *Electrotech. Rev.*, vol. 72, no. 1, May 2005.

BIOGRAPHIES

URBAN SEDLAR [S] graduated from the Faculty of Electrical Engineering at the University of Ljubljana, Slovenia, in 2004. He is currently working toward his Ph.D. as a researcher at the Laboratory for Telecommunications of the Faculty of Electrical Engineering. His research focuses on Web technologies, Web-based services, and multimedia.

LUKA ZEBEC received his M.Sc. degree in electrical engineering from the University of Ljubljana, Faculty of Electrical Engineering. He joined the Laboratory for Telecommunications in 1999. He is employed as a research associate. He has extensive research and industrial experience in the analysis and design of convergent Internet applications and next generation telecommunications systems and services.

JANEZ BEŠTER [M] received his Ph.D. in the field of telecommunications from the University of Ljubljana. He is currently head of the Laboratory for Telecommunications and is an associate professor in the Faculty of Electrical Engineering. His work focuses on planning, realization, and management of telecommunication systems and services, as well as applying information and communication technologies to education.

ANDREJ KOS [M] graduated from the University of Ljubljana in 1996 and was awarded his Ph.D. degree in telecommunications in 2003. At present, he works as an assistant professor at the Faculty of Electrical Engineering. He has extensive research and industrial experience in the analysis, modeling, and design of advanced telecommunications elements, networks, systems, and services. His current work and research focus on managed broadband packet switching and next generation intelligent converged services.

IP television is currently a hot topic with service providers. Presenting a convergent platform for multimedia and communication services, it also allows integration with independent systems such as the telephony system and smart environments.

Quality-Assured Provisioning of IPTV Services within the NGN Environment

Mojca Volk, Joze Guna, Andrej Kos, and Janez Bester, University of Ljubljana

ABSTRACT

IPTV represents a solution for interactive television-like services. As the NGN has gained attention for the IP multimedia service delivery platform, IPTV has been recognized as the way to provide the key value-added services. However, IPTV differs from typical NGN-based voice and data services by the fact that it combines three conceptually unfamiliar (until now) components: streamed video, Web services, and NGN-based service control. Another difference is in the sense of the quality-assured service delivery that in the case of IPTV is much stricter for two reasons: it is more challenging to meet an end user's satisfaction in the case of television services, and there are issues of quality-assured provisioning of real-time multimedia services in an environment that is best-effort in its nature.

In this article, possible approaches to NGN-based IPTV services assurance are presented from the QoE and QoS viewpoints. Current environments and an overview of the standardization efforts are given. A proposal for a fully NGN-integrated quality-assured IPTV provisioning model is presented with an associated converged profile structure. The service-aware quality assurance approach is argued.

The article investigates further NGN service delivery enhancements for quality-assured provisioning of IPTV services that until now remained unresolved. The objective of this article is to present the design of a realistic quality assurance model, establish the associated framework for NGN-based IPTV services delivery, and contribute to discussions and research activities. At the same time, the conclusions and the proposed models will serve as a framework for future efforts in our laboratory NGN testbed.

INTRODUCTION

Internet Protocol television (IPTV) represents the provisioning of real-time television services above IP core and access networks with various mechanisms implemented to ensure the appropriate level of quality [1]. Today, the IPTV system and services are seen as one of the key new next-generation network (NGN) services and

thus, are also heavily involved in the NGN standardization process. In the case of IPTV, the latter are composed of three unrelated components: streamed video provisioning, NGN service control and provisioning, and Web services, that is, they are components stemming from three conceptually unrelated service environments that must be interlaced and blended to achieve the end user's satisfaction.

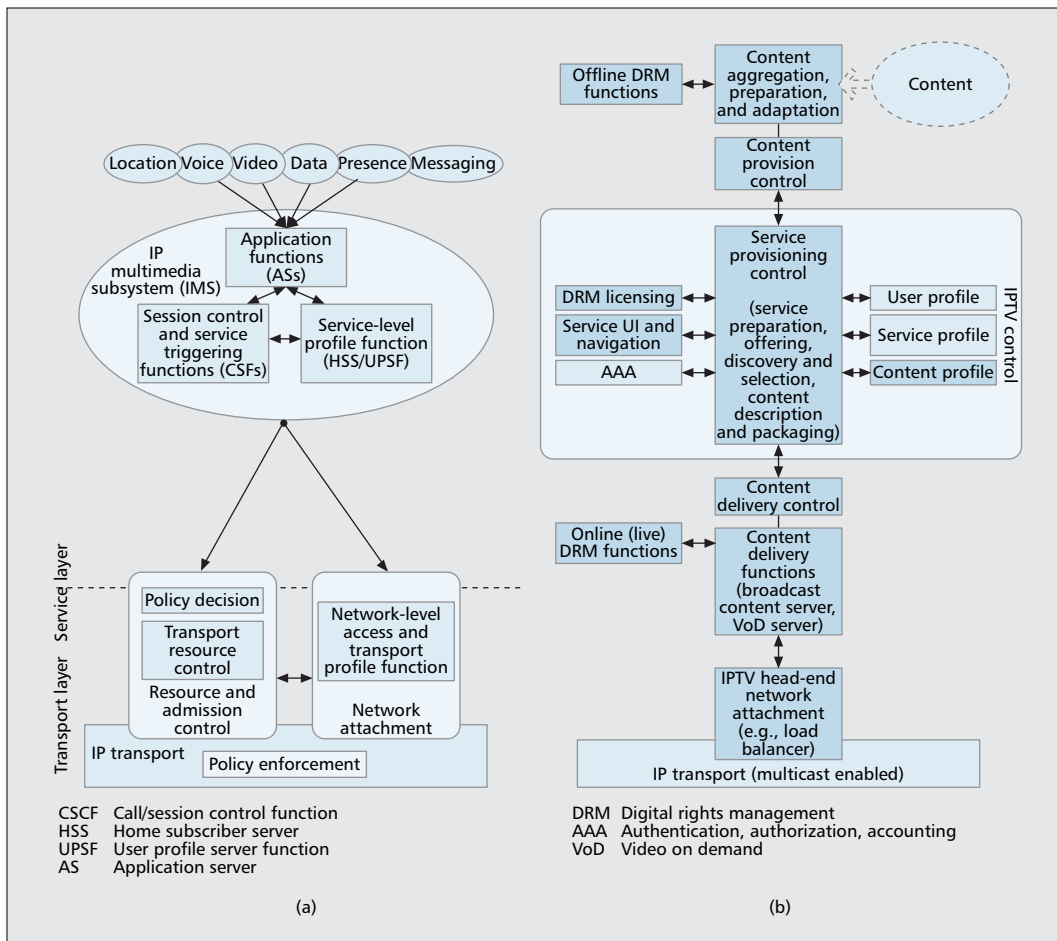
However, the provisioning of real-time multimedia IPTV services within the NGN introduces challenges regarding quality assurance due to the following factors:

- The inherent nature of the IP-based next generation environment, originating from the Internet domain, is best-effort and therefore requires several additional mechanisms to meet the appropriate quality and availability levels.
- An extensive range of different media-rich services presents challenges to resource allocation due to diverse performance requirements and issues of interactivity and personalization (e.g., real-time or near-real-time delivery, priority treatment).
- Even when not strictly classified as emergency services, some services are historically characterized as highly available and come with expectations of assured quality from an end user's viewpoint and therefore, come with the same expectations for the NGN environment as well. The IPTV services are such services.

There are two different viewpoints of system performance, related to quality-assured service delivery [2]. Quality of experience (QoE) describes the overall system performance from an end user's perspective and is related to service layer and the ability of a service to meet the end user's expectations. On the other hand, quality of service (QoS) represents a measure of performance from the network perspective and involves a set of QoS mechanisms, employed to manage network conditions and deliver differentiated services to selected traffic flows and selected end users. The relation between the two is the issue of mapping the requirements of the end user and the service to the transport parameters and mechanisms.

For complete service delivery, converged and

This article proposes a novel scenario of converged IPTV service delivery in the NGN from a systematic QoE and QoS viewpoint, provided by both service and transport layer service-aware quality mechanisms.



■ Figure 1. a) The generic NGN model; b) the generic vertical IPTV model.

overall QoE and QoS assurance is required [1, 3]. Many works [1, 4] propose network-level upgrades to assure the appropriate QoS mechanism for IPTV service delivery within NGNs. However, limiting the QoE and QoS solutions to the transport (network) layer omits the desired service-level dependencies when providing blended services. It is suggested more and more frequently that the applicable solution would be service-aware and policy-based. For these reasons, this article proposes a novel scenario of converged IPTV service delivery in the NGN from a systematic QoE and QoS viewpoint, provided by both service and transport layer service-aware quality mechanisms.

This article is organized as follows. In the first section, the current approach to the IPTV system is presented. A generic NGN model is briefly discussed, giving insight into admission control and quality assurance issues from a service delivery and integration viewpoint. In the second section, an overview of the current status of NGN-based IPTV architecture proposals from QoE and QoS viewpoints is given and standardization efforts are summarized. IPTV is introduced to the NGN environment. In the third section, a study on QoE and QoS assurance procedures is presented for NGN-based delivery of IPTV services. A converged proposal of an NGN-based IPTV QoE and QoS assurance model and the associated extended profile

structure are given. Finally, concluding remarks are presented and open issues for further discussion and research work are identified.

CURRENT ENVIRONMENTS

NGN ARCHITECTURE

The NGN concept has been addressed, thoroughly discussed, and well defined from both the research and development sector and the standardization and regulatory bodies; among these are foremost, the International Telecommunication Union-Telecommunication (ITU-T), the Third Generation Partnership Project (3GPP), and the European Telecommunications Standards Institute (ETSI) TISPAN [4–6]. There are several architectural proposals available; nevertheless, they all share certain basic characteristics, represented within the following generic NGN model (Fig. 1a):

- IP multimedia subsystem (IMS), providing core session control, service triggering, and authentication, authorization, and accounting (AAA) mechanisms
- Network attachment subsystem (NASS) for the end user's device initialization and network attachment procedures
- Resource and admission control subsystem (RACS) for policy enforcement, admission control, and resource management

In the following sections, the engaged con-

There have been several activities involved in NGN-based IPTV architecture considerations. The greatest efforts are being made by the ITU-T FG IPTV, ATIS-IIF and ATIS-TMOC, ETSI TISPAN, DVB Consortium, PacketCable, and OMA BCAST standardization organizations and bodies.

	Area of contribution (NGN IPTV)	Architectural approach
ITU-T FG IPTV	General	Non-NGN IPTV functional architecture NGN-based non-IMS IPTV functional architecture NGN IMS-based IPTV functional architecture
ATIS IIF ATIS TMOC	General	Non-IMS approach for IPTV Core IMS approach for IPTV in the NGN framework
ETSI TISPAN	NGN oriented	Dedicated IPTV subsystem within the NGN
DVB Consortium	DVB technology oriented	DVB IPTV system, services, home platforms, content
CableLabs PacketCable	Cable network oriented	Voice over IP architecture for cable systems (PacketCable 1.x) NGN IMS-based functional architecture for cable systems (PacketCable 2.0) Architecture for multimedia services delivery in cable systems (PacketCable Multimedia)
OMA BCAST	Interactive mobile TV oriented	Mobile IPTV framework On-demand video services

■ **Figure 2.** Summarized efforts of the key organizations and bodies involved in the NGN-based IPTV standardization.

cepts and mechanisms refer to the previously presented model.

VERTICAL IPTV ARCHITECTURE

From the architectural viewpoint, current IPTV deployments are typically proprietary-based vertical solutions, comprising four segments (Fig. 1b).

- Content provisioning, where content and the associated metadata are ingested, aggregated, and prepared through adapting processes and digital rights management (DRM)
- IPTV control, implementing service provisioning control and AAA functionalities (middleware)
- Content delivery, where content is packed into services and delivered to the end users
- Customer premises, represented by a set of content acquisition and processing functionalities and service execution functionalities within end user equipment

The content provisioning, IPTV control, and content delivery segments together are known as the head-end.

One notable advantage of a vertical approach is a dedicated platform for a chosen service type, namely, IPTV. Two notable disadvantages are the maintenance of separate parallel vertical networks — typically, there are dedicated networks for voice over Internet protocol services, Web services and video streaming services — and the incapability of blending voice, data, video, presence, and messaging services due to separate and conceptually incompatible provisioning platforms. Moreover, a vertical IPTV solution was proven to be considerably walled and proprietary-based, which presents issues of interconnectivity, multi-vendor environments, and third party provisioning [7].

The possible solution to these challenges is offered through the deployment of IPTV services within the NGN environment. This topic is thoroughly addressed, foremost by the ITU-T, which is very active through the contributions of

its IPTV focus group, FG IPTV, which addresses the possible approaches of converged NGN-IPTV deployments [7, 8]. The focus of the rest of this article is on the quality issues for integrating IPTV into the NGN and vice versa. The proposed framework relates substantially to the ITU-T proposal in [9], as follows.

INTEGRATING IPTV INTO THE NGN

STANDARDIZATION EFFORTS

There have been several activities involved in NGN-based IPTV architecture considerations. The greatest efforts are being made by the ITU-T FG IPTV, ATIS-IIF and ATIS-TMOC, ETSI TISPAN, DVB Consortium, PacketCable, and OMA BCAST standardization organizations and bodies (Fig. 2).

Whereas the main standardization emphasis is on IPTV architecture and service requirements and their integration into the NGN environment, other issues such as end-to-end content protection, QoS and QoE functionalities, end user's premises equipment, and interoperability standards and testing requirements are being addressed as well. In addition to general QoE and QoS requirements, specific IPTV services and their requirements are being defined.

Considering the architectural approach to the IPTV integration, the ITU-T FG IPTV proposed three different approaches [9]:

- The non-NGN IPTV functional architecture
 - The NGN-based non-IMS IPTV functional architecture
 - The NGN IMS-based IPTV functional architecture
- The ATIS-IIF proposes two approaches:
- The non-IMS approach for IPTV in the NGN framework
 - The core IMS approach for IPTV in the NGN framework

The ATIS-IIF states that both approaches should be able to co-exist on the same NGN framework and should use common components where possible.

The ETSI TISPAN defines a separate IPTV subsystem and argues that the NGN services shall not be impacted by the IPTV services; however interworking between the two should be possible.

OMA BCAST focuses heavily on creating a standard for interactive mobile TV, as well as on-demand video services. The CableLabs PacketCable is involved in solutions related to cable networks. The DVB Consortium is focused on DVB technology and has created several standards for the delivery of DVB services over IP-based networks.

There is a strong demand for a standardized solution for IPTV service. The implications of the standardization survey show that there are prevailing arguments for NGN-based delivery of IPTV services. However, the latter demonstrates specific requirements within all segments of the environment that should be addressed appropriately.

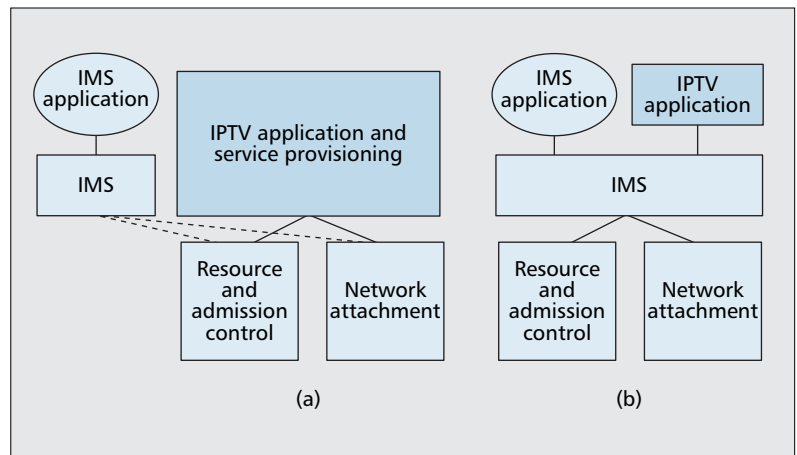
ARCHITECTURAL APPROACHES

The comparison of the IPTV and the NGN architectures (Fig. 1) reveals several functional analogies. Both architectures are based on the separation of session control and service triggering functionalities from application logic and content, that is, the functionalities of the IMS and the IPTV head-end are similar, the application logic is placed on designated application servers, and specialized media servers for content provisioning are engaged. Based on these findings, it is reasonable to accomplish the integration of IPTV services within the NGN environment.

The NGN-based IPTV implementation is applicable in two ways. IPTV could be implemented as a separate subsystem with independent session control, service triggering, profile repositories and authentication mechanisms, but with a dependence on NASS and RACS to complete the service delivery. In contrast to this partially migrated solution that is focused chiefly on the transport layer, the second approach anticipates integrally implemented IPTV services — within and above the IMS subsystem, NASS, and RACS — as specialized applications implemented in parallel to other NGN services. In the latter case, which is the focus of the rest of this article, the delivery of IPTV services is reflected in the service and transport layers and in all NGN subsystems. Both approaches are presented in Fig. 3.

FULLY INTEGRATED NGN-BASED IPTV SOLUTION

The proposed NGN IMS-based approach is presented in Fig. 4. The IPTV is introduced into the NGN architecture as a composition of specialized application logic, placed on dedicated application servers (IPTV AS), and content delivery functions performed by dedicated media servers. Additionally, specialized servers are required for ingestion, preparation, and provisioning of content, electronic programme guide (EPG) functionalities, and DRM. The session control, service triggering, and AAA mechanisms are performed uniformly by the IMS, namely, call



■ **Figure 3.** a) Partial; b) fully integrated approach to NGN-based IPTV solution.

session control functions (CSCFs) and the home subscriber server/user profile server function (HSS/UPSF). The RACS and the NASS are engaged for access, admission, and transport control.

Compared to the similar ITU-T NGN IMS-based architecture in [9]; in the proposed architecture, the application layer is generalized in the sense of the fully-converged delivery of IPTV services in conjunction with various telephony and non-telephony NGN services. Because the complexity of such a service environment is prone to increase, it is our opinion that service coordination and interaction managers (SCIM) will be required.

A generalized service discovery and selection mechanism (SD&S) is employed, extending the limited IPTV location-content discovery scheme with other mechanisms to also include services (e.g., Web-based mechanisms, GRID technology). IPTV subscriber management is realized through IMS-based subscriber-related functionalities, whereas in [9], these are implemented as separate IPTV-based functionalities.

THE PROPOSED MODEL OF NGN-BASED IPTV QOE AND QoS ASSURANCE

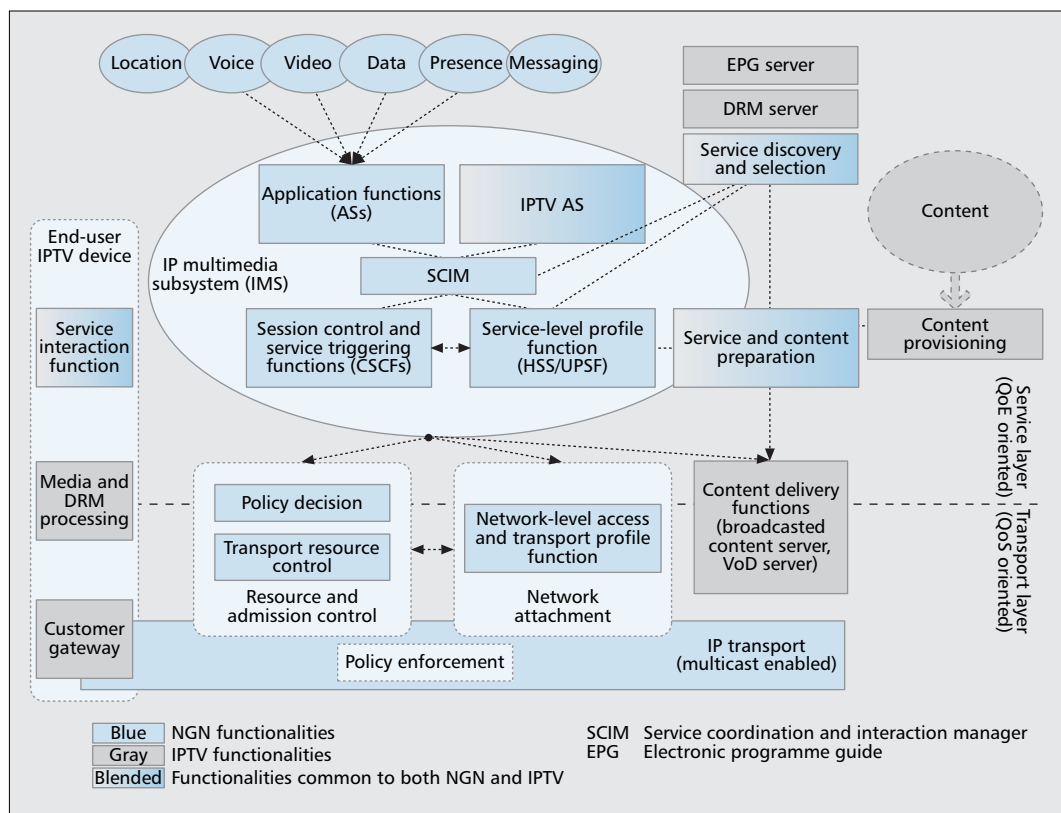
To the best of our knowledge, works have addressed NGN-based IPTV deployment issues [10] and IPTV transport QoS issues [1]; nevertheless the overall approach to systematic quality assurance within NGN-based IPTV solutions has not been addressed appropriately. In this section, a proposal for a fully-integrated quality assurance procedure is presented and discussed.

SERVICE-AWARE QUALITY ASSURANCE WITHIN NGN AND IPTV

As briefly discussed in the introduction, the measure of user satisfaction is the issue of both transport-layer and service-layer quality performance.

In the NGN, the QoS mechanisms are technology dependent and extend vertically across

The proposed procedure relates to the proposed fully-integrated NGN IMS-based model. It is based on the experience gained from the IMS-based NGN testbed that we set up in the laboratory and will serve as the framework for further research.



■ Figure 4. Fully integrated IPTV services delivery within the NGN architecture.

the transport layer and transport control section of the service layer, whereas the QoE mechanisms are technology independent and involve service control and application functionalities. The two viewpoints of meeting the end user's needs in a system are tightly interlaced.

In [8], a plastic representation of quality assurance procedures in the NGN is given. In summary, throughout the service delivery, quality assurance is provided in two stages:

- In the session setup stage, signaling and dynamic negotiation between the two communicating parties occurs to reach an agreement on communications characteristics, followed by resource allocation.
- At any time during the session, further renegotiation is possible as a reaction to eventual changes on the end user, service, or network side.

Three groups of mechanisms establish a QoE and QoS assurance solution:

- Service layer mechanisms — sustaining pathways for traffic transmission, including admission control, QoS routing and resource reservation, profile management, and AAA mechanisms
 - Transport layer — dealing directly with the end user traffic, for example, buffer management, congestion avoidance, packet marking, queuing and scheduling, traffic classification, traffic policing, and traffic shaping
 - Management mechanisms — providing traffic restoration, metering and recording, and policy
- Specialized arbitrating functionalities are

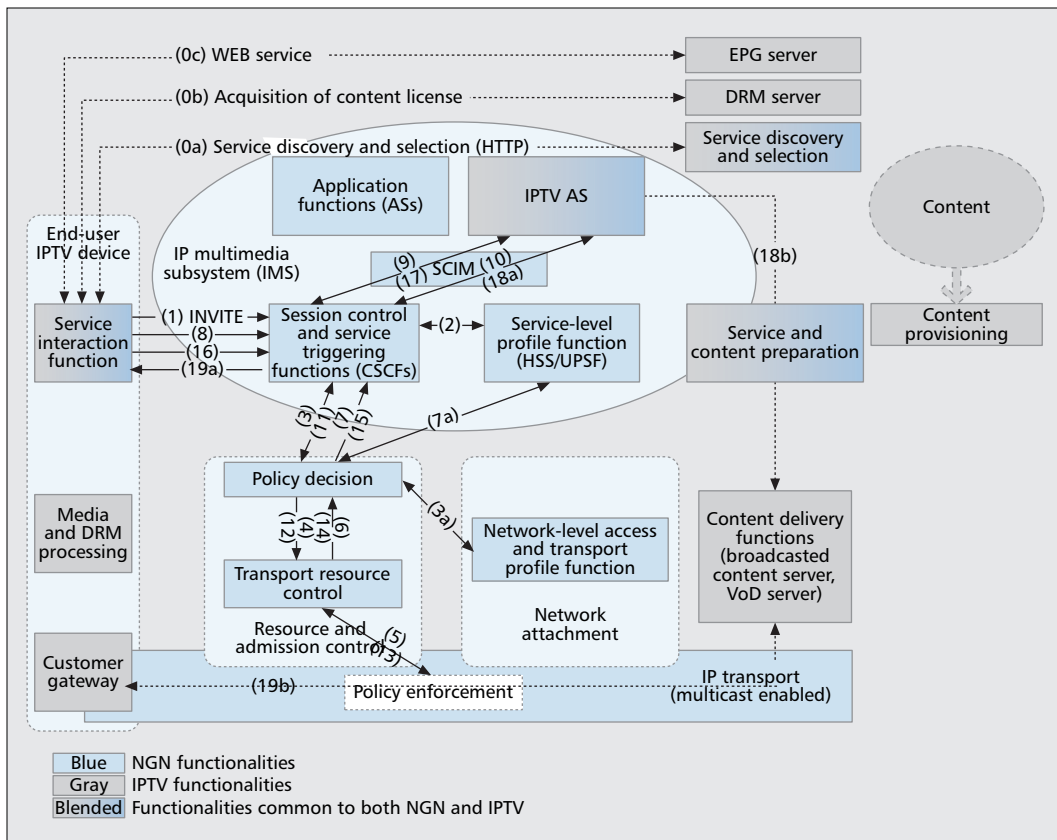
engaged, namely, RACS and NASS functionalities [3], which represent the linkage between the service and the transport layer, taking into account the service requirements associated with the requesting user and current network conditions. Thereby, dynamic service-aware admission control and resource reservation is ensured.

In vertical IPTV solutions, QoE and QoS are ensured in a proprietary way. Conceptually, there are some similarities to the NGN model for quality-supported service assurance although in IPTV solutions, it is presumed that transport-layer QoS is ensured. Access and core transport domains are considered closed and controlled, and the associated transport-layer parameters that affect IPTV service provisioning are pre-confirmed. For these reasons, no mutual negotiation and approval regarding quality mechanisms is provided, and such deployments are limited strictly to homogeneous and controlled environments. Despite the feasibility to provide IPTV services, such solutions present a problem to NGN-based concepts of access agnosticism, end user terminal transparency, and global end-user and service mobility and therefore, should be reconsidered.

THE PROPOSED SERVICE-LAYER QUALITY ASSURANCE PROCEDURE

The proposed procedure relates to the proposed fully-integrated NGN IMS-based model (Fig. 4). It is based on the experience gained from the IMS-based NGN testbed that we set up in the laboratory and will serve as the framework for further research.

Considering service-aware QoE and QoS assurance, the notion of profile is important. It introduces a centralized and technologically independent logical representation of the involved parties and applicable processes among them.



■ Figure 5. The proposed fully integrated NGN-based IPTV quality assurance model.

Referring to Fig. 5, the end user issues a service request, containing either the end user's profile or a reference to the end user's profile repository (HSS/ UPSF). Received by the CSCF entities, these perform authentication and authorization procedures based on information retrieved from HSS/UPSF (1–2). If successful, the CSCF entities issue a request for resource authorization and reservation to a policy decision entity, containing service-layer profile parameters (3).

The process of resource authorization and reservation is composed of two parts. The first part involves transport-layer profile lookup and authentication, followed by an authorization procedure with respect to service and end user requirements (subscriber and terminal equipment, access network). The second part involves the network resources allocation. The following steps (3a–6) compare the received service-layer parameters (comprising end user, service, and content requirements) against transport-layer parameters, retrieved from the NASS database, and against current network capabilities. Based on the reconciled requirements, transport-layer authorization is completed and network resources are allocated. Next, optimization processes are applied to accomplish the resource allocation based on the negotiated service-layer parameters and the current network conditions. At this point, transport-layer QoS mechanisms are invoked.

The response with the associated information is issued to the CSCF (7). In addition, possible changes in end user, service, and content related

information are reported to the profile server (7a). The CSCF forwards the response to the end user's device for parameter negotiation (8).

In case the proposed parameters are confirmed by the end user's device, the CSCF forwards the negotiated parameters to the application server, hosting the application logic of the requested service (9). The application server determines the final feasible operation configuration for the service, based on which resource reservation procedure is invoked (10, 11).

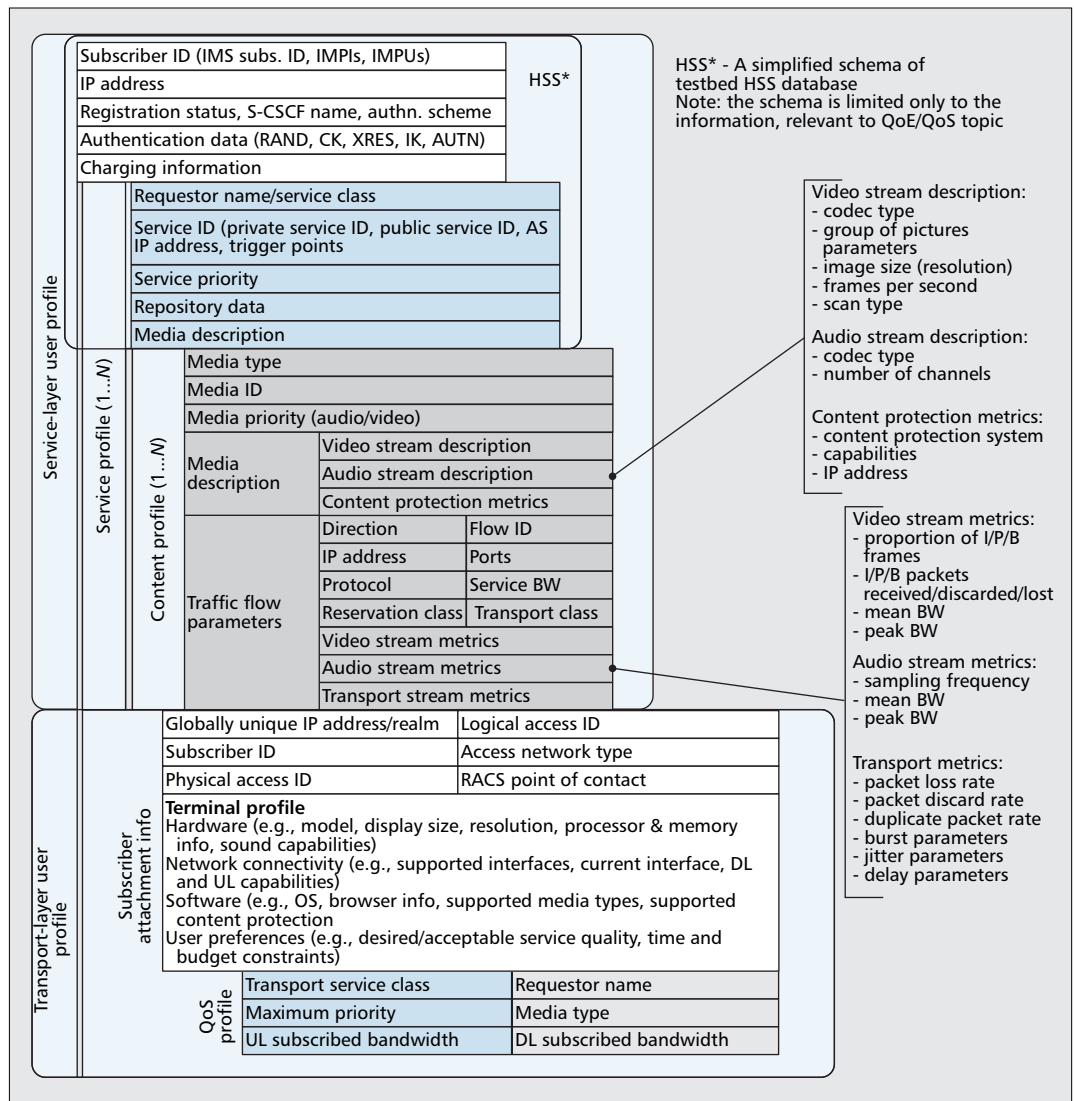
At this point, resources in the network are reserved that correspond to the requirements for final service delivery (12–14). When the CSCF receives the confirmation of the resource reservation (15), it informs the end user's device (16) and the serving application server (17). The application server issues instructions for service/content delivery to the media server (18b), and service/content is delivered to end user's device (19b). The service admittance signaling also is completed (18a, 19a).

After the session set up is completed, further renegotiations may be invoked based on session modification requests from the end user or service side or due to resource availability changes.

THE PROPOSED CONVERGED PROFILE STRUCTURE

Considering service-aware QoE and QoS assurance, the notion of profile is important. It introduces a centralized and technologically independent logical representation of the

The information in the content profile is provided by content preparation functionalities (content metadata) and media delivery servers; therefore, the media server is also an optional location for content profile storage.



■ **Figure 6.** The proposed quality-related profile structure for fully integrated NGN-based IPTV services delivery.

involved parties and applicable processes among them. The sustained intelligence information, corresponding to the underlying physical implementation, is organized into transport-layer and service-layer profiles, representing the end user, the services, and the content. The profiles are provided through HSS/UPSF and NASS database entities [6].

For the purpose of QoE and QoS assurance within the previously discussed model, we propose the following quality-related profile structure, represented in Fig. 6. The structure reflects the two-layer quality assurance model and comprises service-layer QoE parameters mapped to transport-layer QoS parameters. It represents a model of interlaced end user, service, and content profiles. Note that the profile information in Fig. 6 is limited to parameters, associated to QoE and QoS assurance. The HSS* represents the profile entity currently implemented in our NGN testbed.

Referring to the profile parameters, we identify three metrics categories: perceptual quality metrics assessing QoE (e.g., several types of the

mean opinion score), and video/audio stream metrics, and transport metrics, describing objective QoS. The perceptual quality metrics represent a subjective description of the end user experience based on the applied QoS parameters, associated with human-related factors.

However, there are several issues left to this proposal.

In our proposal, the content profile is integrated within the profile functionalities, which represents the final convergence stage. The information in the content profile is provided by content preparation functionalities (content metadata) and media delivery servers; therefore, the media server is also an optional location for content profile storage. In the latter case, content delivery requirements cannot be negotiated within the initial negotiation procedure (1–2). For this reason, a renegotiation procedure with reallocation, end user confirmation, and reservation should be conducted after the application server issues the content delivery instruction to the media server (18b).

As indicated with (0a), prior to a session set-

up request, there should be an SD&S procedure enabled, through which the end user selects the desired service. Thereby, the end user's IPTV terminal also acquires the Session Initiation Protocol (SIP) and the Real Time Streaming Protocol/Internet Group Management Protocol (RTSP/IGMP) address of the server hosting the requested content. These functionalities can be implemented independently (as indicated in our proposal) or integrated within the profile server functionalities hosting the content information.

Also, DRM and conditional access to the content should be provided as indicated with (0b). There are two types of content protection: offline, for content on demand, and online (in real time), for broadcasted content. In the former case, when content protection is applied during the session set-up procedure, the protection mechanisms can be implemented on a dedicated server or even within session control entities. In the latter case, where content is protected in real time while served, these functionalities are expected within or near media servers.

On the transport control layer, we propose a vertical approach, involving both transport-layer and service-layer profile inquiries prior to configuration procedures. This extends the direct configuration and authentication interaction between the end user's device and the NASS, as proposed in [9]. The main reason for such an approach is in the service-aware delivery, considering the user-inflicted service configuration and service-imposed requirements of convergent services aside from the connectivity requirements and current delivery conditions.

There are two mechanisms for IPTV service delivery: unicast and multicast. Whereas unicast mechanisms are used for personalized IPTV services on demand (e.g., video on demand [VoD]) and NGN services in general, multicast mechanisms are engaged for broadcast IPTV services (e.g., live television service). In this case, multicast transport capabilities are required; multicast service and transport profiles should be defined; and multicast agents should be implemented on the service layer to control multicast application delivery and quality assurance. Because multicast delivery is reflected in all layers of the NGN, further considerations are required.

Furthermore, there is a question of collocation or separation of the application logic, implementing content delivery control, and media servers hosting and serving the content. NGN anticipates two logically separated functionalities (as considered in our proposal), although dedicated content application servers are also an option.

TRANSPORT-LAYER QUALITY ASSURANCE

Finally, the QoE requirements are reflected in strict transport requirements. In this section, an overview of the most important issues is given; we discuss the traffic admittance and admission control parameters and procedures that relate to the quality assurance model discussed previously. The network mechanisms and the associated QoS parameters, related to guaranteed connectivity and transfer of media data within a QoS-enabled transport domain, are discussed thoroughly in [8].

The traffic admittance is based on the criticality of the service, the service level agreement (SLA), and the current condition monitoring of the network resources.

Admission control is used to meet service reliability/availability requirements using policy-defined operation and priority levels. The reservation itself is dependent on the available network resources, whereas the exact nature of the reservation depends on network performance requirements and the specific approach to meeting these (e.g., IntServ, DiffServ, MPLS).

Network conditions are defined through metrics; the most important parameters are bandwidth, packet loss, delay, and jitter. Whereas data services typically depend only on bandwidth and packet loss, in the case of IPTV, delay and jitter parameters also are important. Exact QoS parameters for broadcast IPTV and VoD services heavily depend on the technology used, where the choice of codec is paramount. According to [2,] delay should be under 200 ms, jitter under 50 ms, and the packet loss rate should be approximately 10^{-6} . In theory, employing the buffer techniques, almost any of these parameters are acceptable. In practice, however, the increase in buffer size has a negative impact on the QoE as the start-up time of service tends to increase to unacceptable values. In broadcast IPTV, channel switching time (i.e., channel zapping) is important and dependent on buffer size and according to [2] should be under two seconds.

CONCLUSIONS AND FUTURE WORK

There are strong reasons for the adoption of policy-based quality assurance within NGNs, that is, the possibility of differentiated and quality-assured delivery of blended media rich IPTV services on top of an IP platform. Hence, the very important quality-related issues should be resolved in a timely and consistent manner.

In this article, the QoE and QoS viewpoints of integrating IPTV services into the NGN environment were presented. We argued that it is reasonable to consider the fully-integrated NGN-based IPTV deployment.

This approach intensifies the issues of QoE and QoS assurance due to stricter requirements of real-time multimedia IPTV services and brand new service components, namely, streamed video and Web services. Therefore, we proposed a model of service-aware quality assurance procedures for IPTV service delivery within the NGN environment, taking into account the already standardized and newly proposed functionalities from different players. The proposed solution is based on the recognized IMS, RACS, and NASS subsystems of the NGN environment. The associated converged profile structure is presented, based on which the authentication, authorization, and admittance decisions are met in both the service and transport layers of the proposed framework.

Many challenges are evoked in this article. Further research is required for a precise, case-based definition of the negotiation and admittance mechanisms, content distribution and delivery, as well as to resolve the service coordi-

There are strong reasons for the adoption of policy-based quality assurance within NGNs, that is, the possibility of differentiated and quality-assured delivery of blended media rich IPTV services on top of an IP platform.

nation functionalities of such a complex real-time multimedia service delivery environment. Laboratory testbeds and professional implementations are required. Based on the proposed model, a fully-integrated quality assurance framework for NGN-based IPTV services delivery will be established within our laboratory testbed.

ACKNOWLEDGMENTS

The research and development work was supported by the Slovenian Research Agency.

REFERENCES

- [1] K. Kerpez *et al.*, "IPTV Service Assurance," *IEEE Commun. Mag.*, vol. 44, no. 9, Sept. 2006, pp. 166–72.
- [2] DSL Forum TR-126, "Triple-Play Services Quality of Experience (QoE) Requirements," Dec. 2006.
- [3] J.Y. Kim *et al.*, "NGN Architecture for IPTV Service without Effect on Conversational Services," *ICACT 2006*, vol. 1, Feb. 2006, pp. 465–69.
- [4] ITU-T Rec. Y.2001, "Next Generation Networks — Frameworks and Functional Architecture Models: General Overview of NGN," Dec. 2004.
- [5] 3GPP TS 23.228, "Technical Specification Group Services and System Aspects — IP Multimedia Subsystem (IMS)," Rel. 7, Dec. 2006.
- [6] ETSI ES 282.001, "TISPAN — NGN Functional Architecture: Release 1," Aug. 2005.
- [7] C.-S. Lee, "IPTV over Next Generation Networks in ITU-T," *IEEE/IFIP BcN2007*, May 2007, pp. 1–18.
- [8] J. Song *et al.*, "Overview of ITU-T NGN QoS Control," *IEEE Commun. Mag.*, vol. 45, no. 9, Sept. 2007, pp. 116–23.
- [9] ITU-T FG IPTV-DOC-0062, "Working Document: IPTV Architecture," Jan. 2007.

- [10] S. Q. Khan, R. Gaglianella, and M. Luna, "Experiences with Blending HTTP, RTSP, and IMS," *IEEE Commun. Mag.*, vol. 45, no. 3, Mar. 2007, pp. 122–28.

BIOGRAPHIES

MOJCA VOLK (mojca.volk@fe.uni-lj.si) received her B.Sc. degree from the Faculty of Electrical Engineering, University of Ljubljana, in 2004. She is currently with the Laboratory for Telecommunications as a Ph.D. candidate. Her main research interests are next-generation networks, protocol and mechanisms, and design and deployment of broadband IP multimedia services within service delivery platforms.

JOZE GUNA (joze.guna@ltfe.org) received his M.Sc. degree from the Faculty of Electrical Engineering, University of Ljubljana, in 2005. He is currently with the Laboratory for Telecommunications as a researcher. His research focuses on networking and Internet technologies, IP multimedia services, digital rights management, and smart home technologies.

ANDREJ KOS (andrej.kos@fe.uni-lj.si) received his Ph.D. from the Faculty of Electrical Engineering, University of Ljubljana, in 2003, where he is an affiliated assistant professor. His research work is focused on analysis, design, and development of fixed and mobile telecommunications systems and services.

JANEZ BESTER (janez.bester@fe.uni-lj.si), Ph.D., is a professor and head of the Laboratory for Telecommunications at the Faculty of Electrical Engineering, University of Ljubljana. His work focuses on planning, realization, and management of telecommunications systems and services; implementation and application of information technologies in education; and economic opportunities for knowledge-based societies.

CostGlue: Simulation Data Exchange in Telecommunications

Dragan Savić¹, Francesco Potorti², Francesco Furfari²,

Janez Bešter¹, Matevž Pustišek¹, and Sašo Tomažič¹

¹University of Ljubljana, Faculty of Electrical Engineering,
1000 Ljubljana, Tržaška 25, Slovenia

²ISTI-CNR, Pisa, Italy

dragan.savic@fe.uni-lj.si (Dragan Savić)

The final, definitive version of this paper has been published in SIMULATION, Volume 84, Issue 4 of April, 2008 by Sage Publications Ltd, All rights reserved. \copyright The Society for Modeling and Simulation International (Simulation Councils Inc), 2002. It is available at: <http://online.sagepub.com/>

Abstract

Exchanging simulation data among simulation practitioners is to a great extent hindered by the use of different kinds of data formats in simulation software packages. The purpose of the CostGlue project is to facilitate the exchange of simulation data in the field of telecommunications. We propose a common data interchange format and a data exchange model for raw simulation data, metadata and post-processing data. Based on this model, we additionally propose a framework, CostGlue, designed for packaging simulation output data into the common interchange format, launching post-processing plugins and exporting data into input formats for various third party tools. As a proof of concept we have implemented the framework as a software package and released it as free software.

1. Introduction

Due to the complexity, size and heterogeneity of telecommunication networks, simulation represents an indispensable, universal and cost efficient approach to the development of new networks and consequently to the development of new services and applications. A number of simulators are currently in use for this purpose. Information about the behavior and operation of a modeled system is gathered from the simulation data, i.e., the output of individual simulation programs. A common simulation data exchange format is required to facilitate the sharing of this information. At present, simulation data exchange is to a great extent hindered by many different data formats used in simulation software packages. This topic has been addressed within the framework of the European COST 285 Action "Modeling and

Simulation Tools for Research in Emerging Multi-service Telecommunications" [1], a forum where European researchers periodically meet to address issues related to simulation of communications systems. COST 285 observed that no general solution exists for exchanging big quantities of simulation data from different sources and in different formats. The need was expressed both for a common format for exchanging data, and for transforming this data for use with different data analysis tools, each requiring a different input format.

Conversion between different tools for simulation data analysis generally requires a conversion program for each pair of formats [2]. This means that a total of $N(N-1)/2$ different conversion programs are needed, when N different formats are used. A common interchange format would simplify data exchange: the number of conversions required to share the data reduces from $N(N-1)/2$ to N , as only one conversion tool is needed for each format.

The process of delivering valuable information about the simulated environment involves several activities such as modeling and simulation, analysis and presentation of simulation results, which are generally performed by several different independent applications. We define a reference model of the process involving data collection, simulation and result analysis, in order to provide a general and systematic overview of the simulation procedure.

Following the reference model, this paper describes the project CostGlue [3], which is intended to facilitate simulation data exchange. The project includes the CostGlue *data exchange model*, the *metadata XML description*, the *CostGlue framework*, and a *prototype implementation*. We propose a data exchange model as an abstract model that describes how data is represented and used. The model considers three types of data: raw simulation data, metadata and post-processing data. Special attention is given to the latter two, whose structure is based on the Open Archival Information System (OAIS) reference model [4], modified with a more detailed description of simulation data collection related to telecommunications. All three types of data are stored in an archive based on the common data interchange format. When dealing with a large number of archives, it is possible to aggregate them in catalogues to be published in repositories for scientific communities. For this reason, we use the taxonomy from the Council for the Central Laboratory of the Research Councils (CCLRC) Scientific Metadata Model (CSMD) [5]. The CostGlue prototype, which implements this CostGlue framework, is able to package simulation output data into the common format, launch post-processing plugins and export data into various formats. The framework consists of a core, a database, an API (Application Programming Interface), and an arbitrary number

of specialized plugins. The core is the sole responsible for accessing the database. Specific functions, such as import and export of data and different mathematical calculations, are provided by self-describing plugins, which are loaded on demand. The plugins access the database by interacting with the core through a well-defined API. The prototype CostGlue software package should be viewed as a proof of concept of the feasibility of the proposed data exchange model and the framework. Since measurements and simulation data have very similar format, they can be stored into a common database, which allows easy comparison, a valuable feature in the process of modeling and simulation [6].

The paper is organized as follows. In Section 2 we provide a brief overview of the related work. The simulation process model is described in Section 3. It is followed by a short discussion of the simulation tools and data formats in Section 4. The CostGlue data exchange model and the metadata XML description are presented in Section 5 and the accompanying CostGlue framework with available plugins is presented in Section 6. Some concluding remarks are given in Section 7.

2. Related work

To the best of our knowledge, no general solution for simulation data exchange exists in the field of telecommunications. However, there are some relevant examples that partially address the idea of simulation data exchange among researchers. Starting with those most related to the field of telecommunications, several online catalogues with measurement data for different types of telecommunication networks are available. One of them is the Community Resource for Archiving Wireless Data At Dartmouth (CRAWDAD) [7], a wireless network data resource for the research community. It hosts several data sets in the form of trace files (snmp, tcpdump, syslog, etc.) in different data formats. The data sets are accompanied by several analysis tools. In addition, the CRAWDAD archive contains metadata with the description of CRAWDAD data, tools, related papers and their authors. Despite the efforts of the CRAWDAD community, the diversity of different file formats and tools for each format still hinders data exchange among the researchers.

Multilayer Network Description (MND) [8] aims to assist multilayer modeling by improving the data interchange between various tools. This is accomplished by defining a generic data structure for the multilayer environment which is also present in communication networks. It uses eXtensible Markup Language (XML) to increase flexibility and to improve the visibility of relevant data. The drawback of the MND is that it represents only one of the many possible descriptions of a simulated environment, therefore not enabling general

applicability in all of the simulated environments relevant to the domain of telecommunications.

A software tool called DataSpork [9] developed at the University of Illinois comes from a different research area. It is distributed as a part of the Material Computing Center software archive. DataSpork was designed in order to perform statistical analyses common to most simulation methods, as well as to allow for an extension to other data types and analysis tasks. Although it is a standalone tool, its development is a part of an overall effort to improve storage and exchange of simulation data in the area of Materials Computation. It enables the creation of new data readers, which means that the data can be imported from any kind of data source as long as the necessary extensions are developed. This tool has no capability of data export, multidimensional data layout abstraction, flexible data viewing or support for structured metadata.

The American Institute of Aeronautics and Astronautics (AIAA) Modeling and Simulation Technical Committee has proposed a standard [10] for the interchange of simulation modeling data of a vehicle or an aircraft between different simulation facilities. The purpose of the standard is to maximize the efficiency of data models exchange by providing a well-defined set of information, definitions, data tables and axis systems. The standard is implemented in XML. Unfortunately, the descriptions of models within the XML files are specific to the field of aeronautics and can not be used in the field of telecommunications.

The CCLRC e-Science Centre has developed a framework, called AgentX [11], which allows simple and automated exchange of information between components of a scientific workflow. The AgentX framework is being used in the eMinerals project [11], which focuses on studying environmental processes at the molecular level by using a range of atomistic simulation tools. From this project we borrowed CSMD, the taxonomy developed by CCLRC for aggregating the archives into catalogues.

3. Simulation process model

In order to obtain a general and systematic overview of creation, flow and processing of data, we define a reference simulation process model. The model provides a layered decomposition of the main functions encountered in both the simulation and measurement processes. This three layered model with one optional sub-layer is shown in Fig. 1.

The first layer - *source layer* - provides the raw simulation output, describing a simulation run to the smallest detail. Raw data is generated by a simulator: usually, one or more records are created for each event during the simulation run at the source layer. The structure and

format of the data at this point depends entirely on the simulator (e.g., ns-2, Opnet). Most frequently data is in the form of large tabular traces, in ASCII or binary format. An optional *source recoding sub-layer* handles the raw source data. Its main purpose is to convert between different formats (e.g. from ASCII to binary or vice versa), to compress data (e.g., Gzip, Bzip2) and to remove private information (e.g. header lines) from simulation traces. The source layer supports both raw simulation data and real measurements data. In fact, during our research, we found that nearly the same model can be applied to the analysis of real network traffic traces. In this case the raw data is not a result of simulation, but rather, for example, the data traces captured in a network link. Apart from the different tool that generates the raw data (traffic capture tools like Tcpdump or Ethereal instead of a simulator) all the functions of the upper layers remain the same.

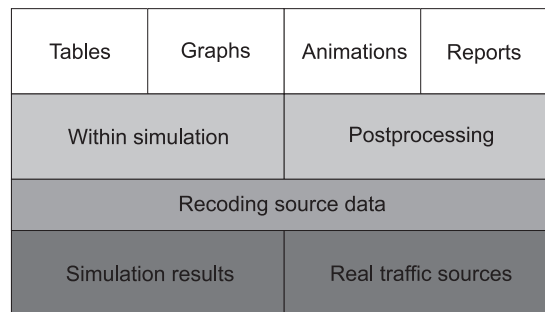


Figure 1: a reference model of a simulation process.

The *processing layer* is responsible for the analysis of simulation data. At this level, cumulative results can be derived from the raw data (e.g. the mean packet delay can be calculated). It is possible to determine the statistical confidence of the results and to conduct additional simulation runs if necessary. An important characteristic of the data processing layer is that the amount of data received from the source layer is usually much larger than the amount of results of post-processing.

The *presentation layer* is the final stage where the results are organized in a form useful for exchanging the most important findings with other simulation practitioners. In the case of simple tabular result printouts, this layer is empty or only slightly modifies the data (e.g. changes in number formats, column spacing). However, data is frequently shown in the form of 2- or 3-dimensional graphs (e.g. as a part of scientific reports or research papers, web pages, etc.) or even presented in animated form (e.g. ns-2 NAM - Network AniMator). At this layer the predominant requirement is the flexibility of presentation and a possibility to create new or modified presentation objects from new or changed simulation results without reformatting.

We can map the functionality of particular tools used in simulation to the layers of our model. Usually, a single tool provides more than one functional layer or even all of them. In the most favorable situation it would encompass all the functions needed and implement them adequately to meet all the researcher's needs. In practice this occurs very rarely and there is usually a set of complementary tools that covers the required scope of functions within the model. The selection of tools is based on arbitrary conditions, such as the capabilities and performance of the individual tools, the researcher's past experience with a particular tool, or the availability of tools. In the case of simulated results, raw data can be generated by discrete event simulators (e.g., ns-2, Opnet). Raw data can be captured in real networks, with sniffers, such as Tcpdump or Ethereal. Source recoding can be done with small dedicated tools (e.g., Gzip, Tcpdpriv) or different proprietary shell scripts. Data analysis can be performed with generic tools for mathematical computation (e.g., Octave, Matlab, Mathematica, Excel), special statistical tools (SPSS, R), or proprietary and dedicated programs or scripts. Often the simulation package provides the functionality for data processing and analysis and it is up to the researcher to decide whether this is adequate or an additional more powerful and flexible tool should be used. Besides dedicated graphing or animation tools, presentation ability can be provided in generic mathematical tools and sometimes even in the simulators themselves.

4. Simulation tools and data formats

In order to provide an overview of tools and formats generally used by telecommunications systems practitioners, we addressed a specific questionnaire to the participants of the COST 285 project, representatives of more than ten European nations. The information we gathered can be summarized in the following observations:

- No single simulation tool has a dominant position. On the contrary, there is a great variety of simulation tools in use.
- Apart from tabular data, other types of elaborate structural formats are seldom used .
- Most of the time the data is used for statistics or graphing. Other uses such as data mining are rare.
- The most common method for evaluating the statistical accuracy of the simulation data is to use independent replications: runs with different simulation times are combined in order to assure stationary intervals of appropriate length. An alternative method using the *equivalent correlation length* obtained by a single simulation run is seldom used.

- A single simulation run produces anywhere from 1 MB to 2 GB of data and a simulation campaign requires from 1 to 100 runs. A measurement campaign requires from 1 to 5 runs, each generating from 100 MB to 50 GB of data.
- The required storage varies from up to 1 GB for short-term storage, to anywhere from 10 MB to 10 GB and more for long-term storage.
- The employed metadata include type, date, parameter values and their description, version tracking, configurations, simulation scripts, and location.
- The metadata is stored in different locations: coded into directory and file names, in separate files, in different storage location inside files (e.g., under the root directory, in shared directories), and in databases.
- Among the simulation tools which use a predefined output format, the most common appears to be the network simulator ns-2.
- Among the generic tools for mathematical computation and running simulations, Matlab appears to be used by most practitioners.
- A large part of the simulators is composed by standard scripting or programming languages and, in general, by ad hoc simulators.
- A great variety of tools is used for post-processing and/or graphing.

These observations, while limited in scope, show that some sort of ASCII format with tabular data is used relatively often.

The variety of tools and data formats used in simulations calls for a general way of reading from and writing to different formats. For this reason we suggested a modular architecture for the CostGlue framework.

5. CostGlue data exchange model

Most simulation data in telecommunications are basically sets of tables with numeric data. Each simulation run generates a table with a few columns and a large number of rows. Each table is associated with certain parameters specific for the simulation, and is uniquely identified by the values of these parameters. We are interested in defining a database structure that is able to efficiently accommodate this type of data.

In the next subsection a structural description of the data exchange model is presented together with the accompanying detailed description of the structure of metadata and post-processing data. This is followed by an overview of data manipulation.

5.1. Structural description of the data exchange model

The questionnaire presented in the previous section revealed that, most commonly, simulation data is organized into a hierarchical structure accessed via a set of parameters. This is usually done by hierarchically organizing the directories of a file system, with raw data usually located in the leaf directories. Each directory is named after the value of the parameter corresponding to its depth. Generally, this means that the number of tree levels depends on the number of different simulation parameters, and the number of directories at each tree level depends on the number of values of the parameters.

A more user-friendly method is to organize the simulation data in the form of a multidimensional array, where parameter values are used as indices into the array. By appropriately indexing the multidimensional array, it is possible to easily extract slices of the whole set of data. Hence the simulation parameters act as the primary interface to raw data, which simplifies raw data querying for users. This is especially useful when the simulation data is collected from multiple simulation runs, which is the common case.

Fig. 2 presents the multidimensional structure of the simulation data. The rows of the parameter table point to data groups and datasets, where the raw simulation data, metadata and post-processing data are stored.

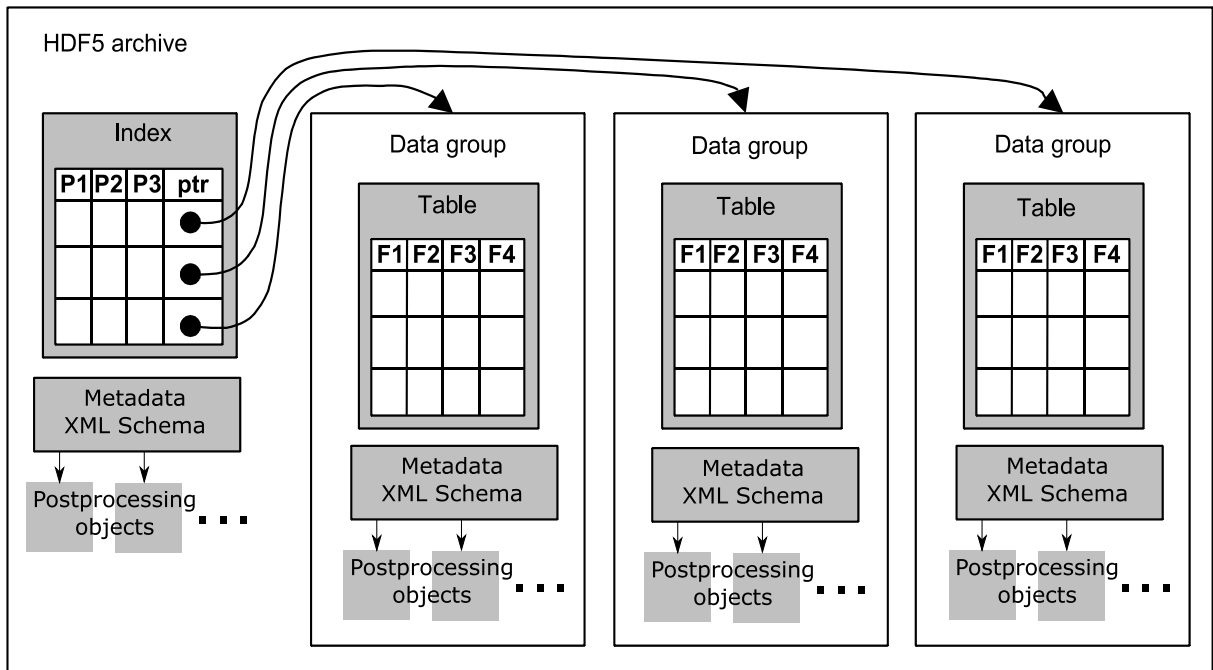


Figure 2: CostGlue multidimensional structure of the data organization.

Each table, together with *metadata* and *post-processing data*, is attached to a data group which usually holds the data produced during a single simulation run. Data groups are indexed with vectors of *parameters* representing an individual simulation run. The index into

the data groups is a 2-dimensional array, also referred to as the *parameter table*, where the parameters relative to each data group are stored. Each column in the parameter table corresponds to a different parameter, and each row contains the values of the parameter relative to a data group. Therefore a parameter table is used as the data structure for accessing a data group, while an array of parameter values relative to that data group is used as the key. A table with raw simulation data is attached to each data group.

The overall structure is a collection of 2-dimensional tables indexed by arrays of P parameters, as shown in Fig. 2. This can be logically seen as a matrix with $P+2$ dimensions, where the first P dimensions are sparse and the last 2 dimensions ($P+1$ and $P+2$) are dense. The first P indices are defined as parameters identifying a data group. As for the last two indices, the first one represents the record (row) number in the data group table, and the second one is the field (column) number of the data group table.

5.2. Metadata and post-processing data

In recent years increasing attention has been devoted to metadata in all application domains. The XML Schema [12] provides a means for defining the structure, contents and semantics of an XML document and is therefore widely used to collect metadata, that is, data about data. In order to insert metadata we have defined a metadata XML Schema whose document instances can be saved together with the simulation data, as shown in Fig. 2. Metadata can be associated to every data group; metadata referring to the archive as a whole is saved together with the parameter table, while metadata for a single simulation run is saved in the related data group. Metadata can also refer to any kind of additional data, labeled as *post-processing objects* in Fig.2. Such cases are, for example, the statistics on the raw simulation data, charts, images, and any other type of data produced from or relevant to the raw simulation data.

The metadata XML Schema is derived from the Information Model defined in the OAIS (Open Archival Information System) reference model [4] and uses parts of the Scientific Data Model (CSMD) [5]. The OAIS reference model is a technical recommendation enabling permanent or indefinite long-term preservation of digital information. The objective of the CSMD model is to enhance interoperability of scientific information systems among research organizations. The adoption of a common XML schema from the CSMD model could facilitate further aggregation of telecommunication archives in catalogues to be published in repositories for the scientific communities [7, 13].

Three main elements are present in the metadata XML Schema: one for the metadata relative to the root (*Study*), one for the metadata relative to a data group (*DataGroup*), and one that combines them together (*Archive*), as shown in Fig. 3.

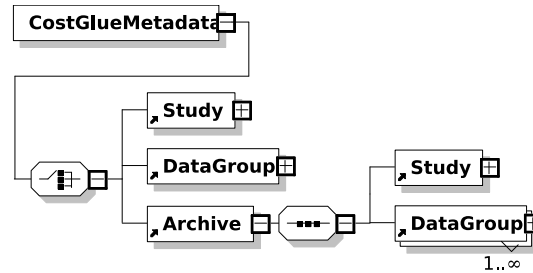


Figure 3: Main element of the CostGlue metadata XML Schema.

The metadata relative to the root is stored in the root of the archive together with the parameter table while the metadata relative to each data group is stored in the data group together with the data table.

More details on the CostGlue XML Schema with additional pictures are available at [14]. Most of the elements in the schema are optional. The choice is motivated by the need not to impose an excessive burden on the experimenter. However, the drawback of this choice is the possibility to have vastly incomplete metadata. On the other hand, it is possible to envision that a certain plugin can certify variable degrees of completeness of the metadata. This way the repositories can accept only the archives that comply with a certain degree of metadata completeness.

The metadata is separated from the rest of the data allowing for easily describing the complete archive. Metadata requires little data storage and can also be made part of the repositories, because metadata includes pointers to the archive location. Data group elements can be extracted separately for efficiency and flexibility.

The metadata Schema does not include a way to serialize the archive data. This choice was made because we see the possibility of exporting the whole archive (simulation data together with metadata) in the XML format as a feature at a different level. Specifically, the HDF5 file format used in the implementation of the software prototype includes the XML schema [15] and the tool [16] to convert a whole HDF5 binary file into the XML file.

5.3. Data manipulation

The manipulation part of the data exchange model includes updating and querying the data contained in the database. Updating the data involves importing different simulation outputs

which often have different file formats. After conversion, the raw simulation data is stored into tables, with the user having to specify data types of the table fields.

When working with raw data in post-processing stage the idea is to take advantage of the logical multidimensional data layout abstraction, where the query output is a slice of data spanning one or more simulation runs. Since the whole database can be seen as a sparse matrix with $P+2$ dimensions the slices of data can be extracted either using *selectors* written in *index notation* or assigning different conditions to arbitrary fields of the table. Both types of queries produce two-dimensional arrays, which can be used for further computations or plotting.

6. CostGlue framework

Besides the data exchange model, we have also designed a corresponding CostGlue framework for simulation data exchange. The architecture of the framework, shown in Fig. 4, consists of a core, database, API and several specialized plugins.

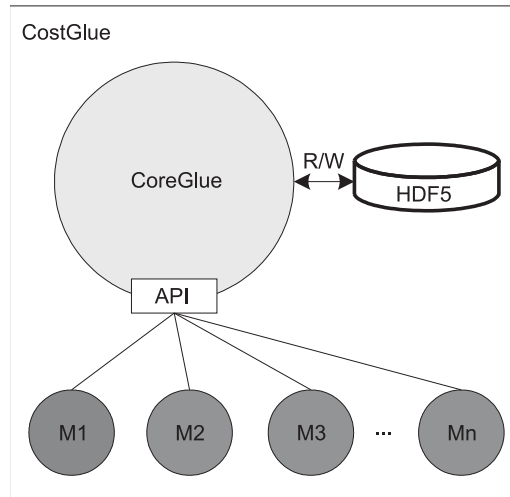


Figure 4: The CostGlue framework.

The core (see Fig. 5) is responsible for reading from and writing to the database, as well as for the dynamic loading of the plugins devoted to different specialized tasks. The core contains three different queues: command, result and report queue. They are used for exchanging different types of messages with the plugins.

For the database we have chosen the HDF5 (Hierarchical Data Format) [17] as the format to store the described data structure in Chapter 5. HDF5 consists of two primary objects - dataset and data group. A dataset represents a multidimensional array of data elements, which can hold different types of data. The data stored in datasets can be either homogeneous (only one data type used within a single dataset - simple datasets) or compound (different data types

within one dataset - compound datasets). Since tabular data collected from simulators often contains data in different forms (e.g. integer, float, char), we used compound datasets for our framework. A data group is a structure containing zero or more objects hierarchically organized by means of a tree-like structure, where an arbitrary number of objects are derived from the main "root" data group. Data groups and datasets have a logical counterpart in directories and files in a hierarchical file system and, similarly to a file system, one can refer to an object in the simulation archive by its full path name.

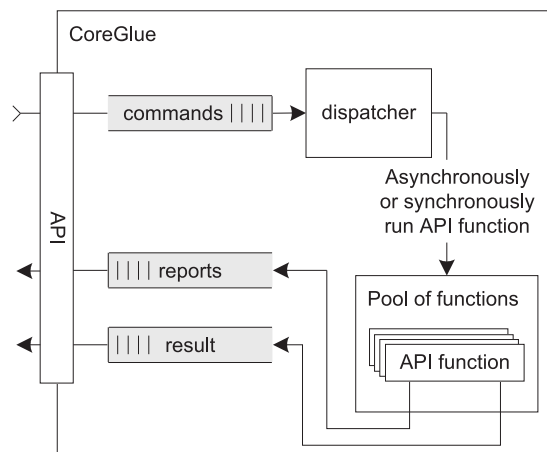


Figure 5: CoreGlue: the core of the CostGlue framework.

The data stored in the database, which is represented as a $P+2$ dimensional matrix, can be queried by using selectors written in index notation. In this notation, each of the $P+2$ indices can be "1", indicating the smallest index; "end", indicating the highest index; "n:m", indicating all the indices between n and m included; ":", indicating the whole range from the smallest to the highest index, "n:s:m", indicating the range from n to m in steps of s; or an array like "[1 5 6 8]", indicating the selected indices.

When using the framework for simulation data exchange, the first step is to call the core, together with a plugin name and its optional arguments. A given plugin can work either alone, e.g. by performing a batch job, or it can depend on other plugins, e.g. when a command line interface calls a specialized plugin to perform a task. Plugins also differ in the way they are built. The simple ones wait for the results after sending the command to the core, and are therefore unable to perform any additional tasks. On the other hand, complex plugins run asynchronously in relation with the core, so they can periodically check the result queue for results and, in the meantime, inform the user about the current work progress by reading report messages from the report queue. Plugins are run as separate threads inside the core. Examples of tasks a plugin can perform are:

- data import/export to/from the database,

- statistical computations over data stored in the database,
- data extraction from the database with complex filters,
- transformations of the data contained in the database,
- graphical output created from the data in the database,
- a generic Graphical User Interface (GUI) for exploring the data in the database, doing simple import/export or computations and running other available plugins,
- a plugin-specific GUI, etc.

6.1. Overview of the API for plugins

The core of CostGlue framework can look for available plugins and query them one by one to get to know their capabilities. This can be used, for instance, to build a menu for a GUI. Plugins are able to describe the parameters they need and the type of output they produce. The core provides the required parameters to a plugin with or without input from the user. In the case where user input is required, the core checks the provided parameters for consistency. A GUI can also use this information and present the choices to the user. The plugins communicate with the core through a well-defined API which contains all the necessary classes and methods for interacting with the database. The methods allow a plugin to manipulate the data group index in order to add or remove data groups, and to manipulate the data groups in order to add or remove data, metadata and post-processed data. Below are a few examples of commands provided by the API:

- Archive open (filename, flags) - opens an archive on disk: arguments are the same as in the C stdio library, returns an archive object.
- Group open (paramvals, exists) - paramvals is a list of parameter values; if exists is true, returns the existing data group with the given parameters, else return a newly created group with the given parameters.
- FieldList add (fieldlist) - adds fields (that is, columns) to a table by taking a list of field names, types (e.g. Int8, Float32) and sizes of the fields.

6.2. Implementation of the CoreGlue framework

As we are interested in storing simulation data into a common database, we made a thorough analysis of the different possibilities of data storage. We focused mainly on the scientific data formats, as they are most commonly used in the scientific community. After the

analysis we narrowed our research down to a specific set of scientific data formats. These were: HDF4, HDF5, netCDF, ODB, FITS and OpenDX. In addition to the scientific data formats, we also considered using plain text formats, relational and object oriented databases. In the end we decided to use the HDF5 [17] data format. The reason for such a decision was that HDF5 meets all the requirements of data organization, (i.e. separation of raw data and metadata) and different requirements of contemporary computer system architectures. These are: managing huge quantities of data, offering a general data model, supporting complex data structures, portability among different computer platforms, parallel data access and processing, diversity of physical file storage media, and sustained development and maintenance. As a confirmation that we have made a good decision, the netCDF format itself has recently decided to evolve towards using HDF5 as its underlying storage format.

Another important issue was the representation of the data. The HDF5 is a binary file format, which means that we need a compatible application capable of reading the file format to view the data. This is a disadvantage, since plain text file formats are much easier to be read. However, this problem is easily solved by using a specialized HDF5 tool [16] provided by the HDF5 development group, which converts binary files into XML text files.

Converting large binary files into XML text file format has some disadvantages. Data stored in XML format requires much more space than the same data stored in HDF5 binary format. Thus, the conversion results in a high data overhead. Data lookup also slows down, since XML parsers are not very efficient when dealing with large amounts of data, compared with the data read performance of the HDF5 library. Nevertheless, this option is available and the choice is left to the users.

For easy and efficient data manipulation we chose PyTables, a Python library built on top of the HDF5 library uses NumPy as a support package, allowing for sophisticated scientific computations.

The extraction of raw simulation data for post-processing is supported by using the Matlab-like index notation. By this it is possible to take complex orthogonal slices of the multidimensional matrix composed of all the data in the database.

The core of the software framework is written in Python. This means that the plugins for specialized tasks also have to be written in Python. To enable plugins to be written in other scripting or programming languages, a wrapper should be built around the core. In addition, we use Psyco, a Python extension module, which speeds up the execution of Python code significantly.

In order to give a flavor of the performance of the prototype application we imported and exported an ns-2 trace, which is a mixture of numbers and strings in ASCII tabular form. In this test both the read and write speeds were around 1 MB/s. For a reference, Matlab on the same machine reads and writes a big matrix of numbers (no strings) with a speed of around 3 MB/s. Matlab has similar speeds both when reading from ASCII and writing to binary or the other way around.

6.3. Examples of plugins

Plugins are meant to be used for specialized tasks such as the import, export, computation, and visualization of simulation data. We have developed a number of plugins, which are briefly described below.

A plugin for ns-2 and tcpdump – this plugin is used for importing ns-2 simulation data and tcpdump measurements into a common table format in order to enable the comparison of these two types of data. This proved to be a valuable feature, especially in the process of modeling and simulation [6]. Two data formats are available: a *full* one, that preserves all the input data, and a *minimum* one, that stores only those data that can be safely converted from ns-2 to tcpdump and vice versa. Both formats can be converted into an input file for NAM, the ns-2 network animator, for a movie-like view of the network behavior at the packet level.

The command line and the HTML GUI plugins – the interactive command line plugin is used for debugging the core and other plugins. Additionally, it can be used to import and export tabular data, whereby the user is required to specify options and arguments. A specialized plugin can automatically recognize the type of trace and automatically provide suitable naming, and possibly, filtering. The basic feature of the HTTP GUI is that when the core can act as an HTTP server, providing a graphical user interface accessible with any web browser. Through this interface, the user can look at the list of available plugins together with their description, the input they require and the output they provide.

The plugins for antenna measurements and simulations – two different import plugins were designed with the help of the researchers from SSR (Signals, Systems and Radio) group at the University of Madrid. The first plugin deals with antenna measurements made in a special antenna chamber. The measurements need to be converted in order to be stored into HDF5 database in the form of tables. A proprietary visualization plugin from SSR is used for visualizing the antenna's directivity diagrams. The second specialized plugin was developed in order to import data from the GRASP8 antenna simulator. The whole process (import, export and visualization) is quite similar to the one described above.

A very useful and practical feature is the ability to compare the measurement results with the simulated ones. The visualization plugin supports this activity as it allows plotting diagrams from both types of data.

6.4. Examples of a real-world simulation

Let us now describe how the data of a real-world simulation can be stored in the described structure. In the first example, taken from [18], we simulate the behavior of packet switches using the ns-2 simulator. Each simulation run is characterized by several parameters. We are interested in the following simulation parameters: architecture type (one of OQ – Output Queuing, VOQ – Virtual Output Queuing, and FPCF - Forward Planning Conflict-Free), buffer size, number of I/O ports, traffic load, transport protocol (TCP and UDP), and flow size distribution (one of heavy-tailed Pareto or constant flow size). Each simulation run differs in at least one of these parameters.

Referring to data exchange model when storing the results of one simulation run in the archive, we populate a new row in the parameter table: values in each row uniquely identify a simulation run. Each row includes the full path to a data group, containing the table with the results of the simulation run, metadata and post-processing data, as shown in Fig. 2. Metadata stores information about the type of scripts used to generate that simulation run, the type of network topology, traffic patterns, etc.

line	arch. type	buffer	ports	load	protocol	flow dist.
1	voq	60	4	30	udp	pareto
2	voq	60	4	40	udp	pareto
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
202	oq	100	4	90	tcp	const
203	oq	200	4	90	tcp	const
204	oq	400	4	90	tcp	const

Legend:

line – row number of the parameter table

arch. type – architecture type of the packet switch (voq – virtual output queuing, fpcf – forward planning conflict free, oq – output queuing)

buffer – buffer size of the packet switch queues

ports – number of input/output ports of the packet switch

load – traffic load of the packet switch

protocol – transport protocol (udp – user datagram protocol, tcp – transmission control protocol)

flow dist. – distribution of flow sizes (pareto or constant distribution)

Figure 6: Parameter table with few sets of different parameters.

By using the CostGlue multidimensional data access via framework's API, it is possible to obtain the raw data for computational tasks by just specifying the parameter values of interest. Using the simulation parameters as a primary interface to raw data comes in handy especially when we are dealing with more than one observation. In Fig. 6 we can see that 204

simulations are driven with different values of parameters. The obtained slices are further processed and the results are fed to the visualization plugin for plotting graphs, which are depicted in Fig. 7.

Post-processing data in this particular case includes packet loss probability (PLP) for three different observations (see Fig. 7). In the first observation we analyze the PLP versus the type of the transport protocols for different packet switch architectures. The load of the packet switch ranges from 30 to 100%. All other simulation parameters remain constant. The second observation analyzes PLP versus flow size distribution for different types of transport protocols where the load varies from 10 to 100%. The third observation analyzes the influence of buffer size (from 10 to 400 packets) on PLP for different transport protocols and packet switch architectures.

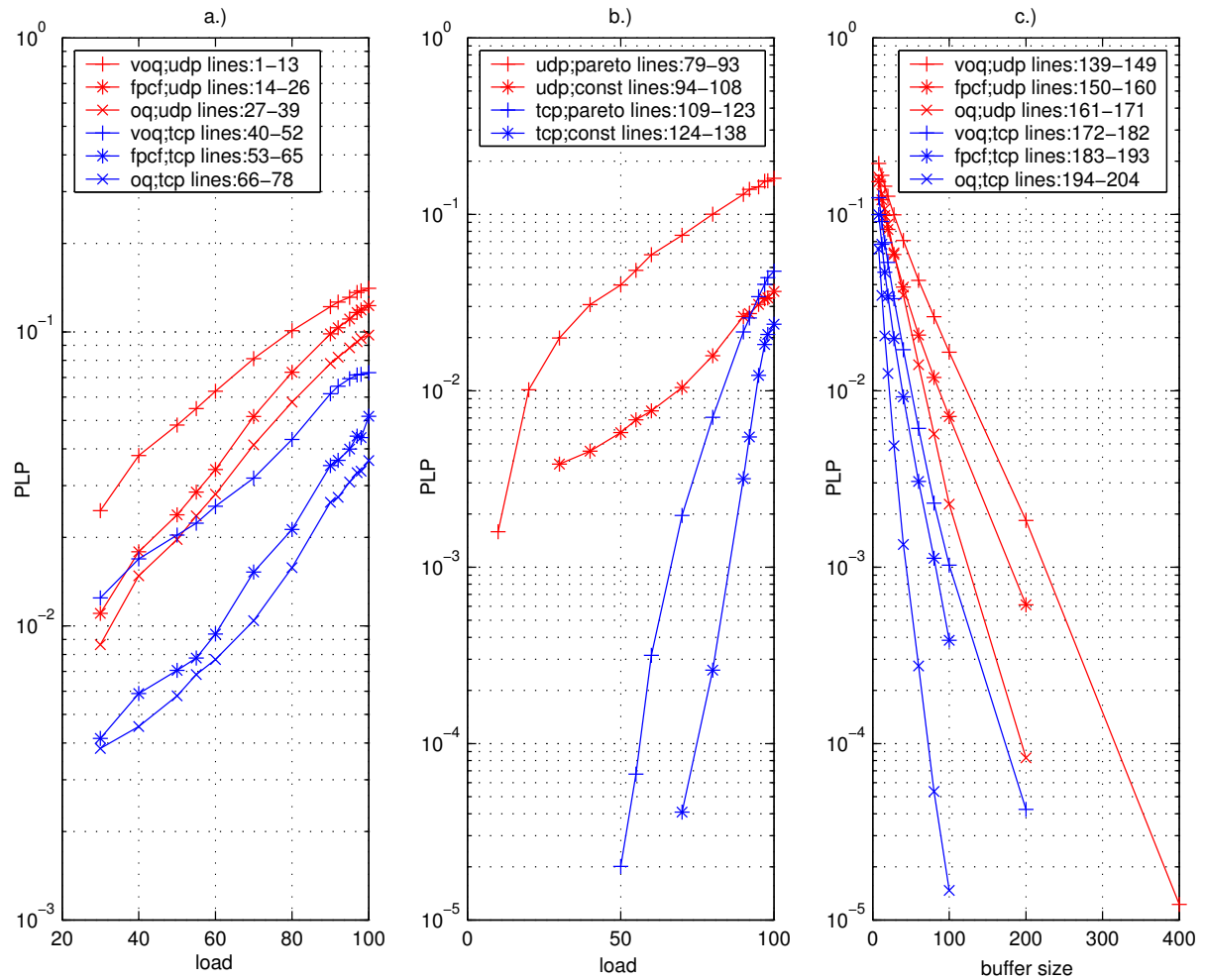


Figure 7: Post-processing results for packet switches.

In the second example we compare ns-2 simulation and measurements for two different network scenarios. Both simulation results and measurement data are saved into the common CostGlue format. Specifically, we compare the packet flow of the ns-2 sending node with the

packet flow of a real sending host where we used a packet sniffer to intercept and log originating traffic (see Fig. 8).

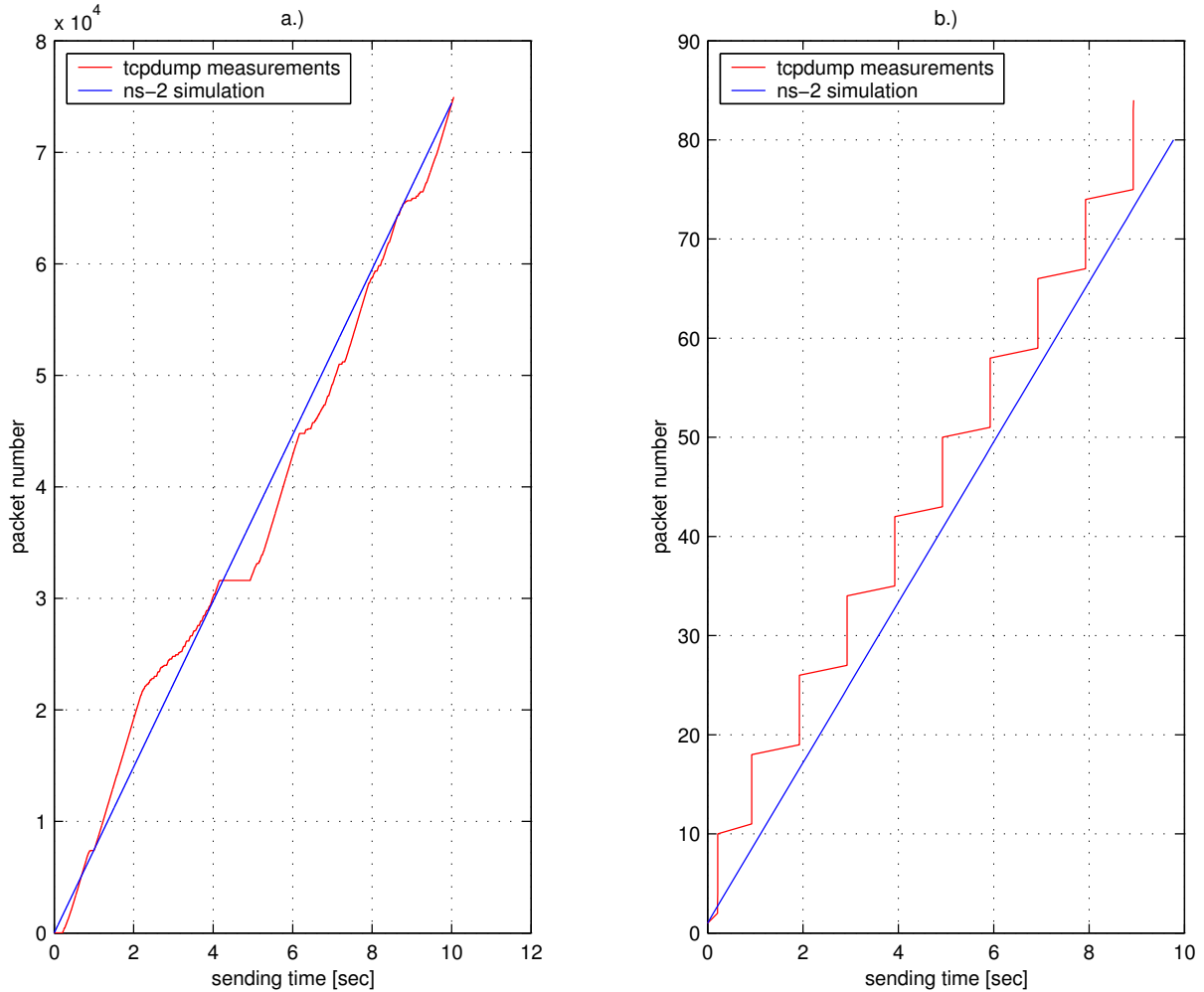


Figure 8: Comparison of simulation and measurements.

The network topology is very simple: it includes two connected nodes acting as FTP server and client respectively. In the first scenario we made a transfer of 100 Mb file from server to client side. The link bandwidth between nodes was set to 10 Mb/s in ns-2, whereas in the case of real data transfer we used a 100Mbit/s Ethernet link. In the second scenario we lowered the bandwidth link in ns-2 to 1 kb/s and used a 10 KB file for transfer. In the real network, we set a download bandwidth limit on the FTP server to 1 kb/s. The sending times for both scenarios are depicted in Fig. 5. Due to the bandwidth restriction set on the FTP server, in the second scenario we got a ramp curve instead of a near linear curve. This was something we did not anticipate. Being able to make this kind of comparisons helps us in the process of verifying models by directly validating them against experiments.

7. Conclusions

The purpose of the CostGlue project is to facilitate the exchange and management of simulation data among simulation practitioners and to ease the task of using different simulation, data processing and visualization tools, all of which have different input and output data formats. A prototype program implementing the CoreGlue framework, based on the CoreGlue data exchange model, has already been developed. We sincerely hope that due to its modular structure the prototype will be enhanced by other simulation practitioners, thanks to the use of a free software license [19]. One possible extension of our work includes the use of web services, where CostGlue would act as a black box capable of receiving commands via web service messages and generating results which would then be forwarded back to the user. This could result in an even more flexible exchange of simulation data.

8. Acknowledgments

We are grateful to the participants of the COST 285 Action for their helpful cooperation. This work was supported in part by the grant from the COST 285 action, by the CNR/MIUR program "Legge 449/97" (project IS-Manet), and by the Ministry of Higher Education, Science and Technology of the Republic of Slovenia under grant no. P2-0246.

9. References

- [1] BRAGG, A.: Observations and thoughts on the possible approaches for addressing the tasks specified in the COST 285 work-plan, COST 285 temporary document TD/285/03/15, CNUCE-CNR (IT), 2004.
- [2] Gravitz, P. D., Sheehan, J., and McLean, T.: Common Activities in Data Interchange Format (DIF) Development, report, McLeod Institute of Simulation Sciences, http://www.ecst.csuchico.edu/~hla/LectureNotes/99S_177SIWPaper.pdf, 1999.
- [3] CostGlue project web site: <http://lt.fe.uni-lj.si/costglue/>
- [4] CCSDC 650.0-r-2: Reference Model for an Open Archival Information System, Blue Book, Issue 1, ISO 14721, 2003
- [5] Sufi S., Mathews B., CCLRC scientific metadata model: Version 2, CCLRC technical report DL-TR-2004-001, September 2004.

- [6] Balci O.: Verification, Validation, and Certification of Modeling and Simulation Applications, In Proceedings of the 2003 Winter Simulation Conference (New Orleans, LA), IEEE, Piscataway, NJ, pp. 150-158, 2003.
- [7] Kotz D. and Henderson T.: CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth, IEEE Pervasive Computing, vol. 4, 4, pp. 12-14, Oct. 2005.
- [8] Rumley S., Gaumier C.: Multilayer Description of Large Scale Communication Networks, to be published in the proceedings of the COST 285 final symposium, March 2007.
- [9] DataSpork analysis toolkit: <http://www.mcc.uiuc.edu/dataspork/>
- [10] Jackson E., Hildreth B.: Flight dynamic model exchange using xml, AIAA Modeling and Simulation Technologies Conference and Exhibit, Monterey, California, Aug. 5-8, 2002
- [11] Tyler RP *et al*: Metadata management and grid computing within the eMinerals project, Proc. UK All Hands Meeting 2007 (AHM2007), Nottingham, UK, Sept. 10-13, 2007
- [12] XML schema: <http://www.w3.org/XML/Schema>
- [13] DRIVER – Digital Repository Infrastructure Vision for European Research: <http://www.driver-repository.eu/>
- [14] Furfari F., Potortì F., and Savić D.: The CostGlue XML schema, Tech. Rep. cnr.isti/2008-TR-03, CNR-ISTI, via Moruzzi, 1, January 2008.
- [15] HDF5 XML Schema: <http://www.hdfgroup.org/DTDs/HDF5-File.xsd.txt>
- [16] H5dump, a tool for converting binary HDF5 file into a XML file: <http://www.hdfgroup.org/HDF5/doc/Tools.html>
- [17] Folk M., McGrath R., Yeager N.: HDF: an update and future directions, In International Geoscience and Remote Sensing Symposium (IGARSS'99), IEEE, Ed., vol. 1, pp. 273–275, 1999

[18] Pustišek M., Savić D., Humar I., Bešter J.: Transport Protocol Dependent Communications in Different Packet Switch Architectures, IEEE Electrotechnical Conference MELECON 2006, pp 704- 708, May 2006

[19] Potortì, F.: Free software and research, in proceedings of the International Conference on Open Source Systems (OSS), M. Scotto and G. Succi, Eds., ECIG Edizioni Culturali Internazionali Genova, pp. 270–271, July 2005.



Contents lists available at ScienceDirect

International Journal of Industrial Ergonomics

journal homepage: www.elsevier.com/locate/ergon

The impact of color combinations on the legibility of a Web page text presented on CRT displays

Iztok Humar^{a,*}, Mirko Gradišar^b, Tomaž Turk^b^a University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia^b University of Ljubljana, Kardeljeva Ploščad 17, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:

Received 28 February 2006

Received in revised form

28 February 2008

Accepted 13 March 2008

Available online 16 May 2008

Keywords:

Color combinations

Computer display

Legibility

Luminance contrast

Color difference

Polarity

ABSTRACT

This paper describes the impact of color combinations on the legibility of a Web page text presented on a calibrated cathode ray tube display. Four hundred and sixty-eight participants tested the legibility of 56 color combinations, compounded of elementary Web safe colors, with a visual performance task. In the task, the participants were asked to identify 21 uppercase alphabetic characters, selected and presented in conformance with the Snellen chart. The number of correctly identified characters was recorded as objective-dependent measure for each color combination. The impacts of color combination, luminance contrast, color difference, and polarity on the legibility were studied. According to the analysis of variance test, color combinations significantly affected the number of correctly identified characters. The obtained results showed that more than 90-year-old Le Courier legibility table is not appropriate for integrative color computer displays. Consequently, a new legibility table is proposed. The influences of both luminance contrast and color difference were investigated by means of the logarithmic and the second-degree polynomial regression model. A second-degree polynomial model proved as appropriate describing the legibility in dependence of luminance contrast. A second-degree polynomial model and logarithmic model proved as appropriate describing the legibility in dependence of color difference. The analysis of influence of polarity on legibility showed that, generally, dark backgrounds lead to better results for colors with greatest luminance contrast.

Relevance to industry

Although the essential consideration when developing Web pages is a modern and fashion-driven design, the results of this study demonstrate that the role of color combinations is crucial at improving users' visual performance. These research findings may not only improve the Web page design where the legibility of displayed text is of vital importance but they could also enhance users' satisfaction by taking into account the preferred color combinations.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Human information processing when reading, learning and making decisions depends on the way information is presented. Acceptance of information depends on the form of information presentation (Ghani and Lusk, 1981; Suh, 1999). The importance of information presentation on a computer display has not become evident until 1973 (Mason and Mitroff, 1973) and systematic research on this topic, particularly on the role of a color as an additional information dimension, started even two more years afterwards (Christ, 1975; Teichner, 1979; Gremillion and Jenkins,

1981; Tullis, 1981; Silverstein, 1982; Benbasat et al., 1986). This can be supported by at least two reasons: first, as a natural phenomenon, color is strongly incorporated into the system of human interactions with the environment. In the esthetic function color is much more effective than in the functionally rational category. Therefore there has been no need for man-computer interactions for a research of possible effects of colors on the human information-processing capacity. Second, the use of a color depended on the level of information technology development. In the first period the focus of development was mostly directed towards technical and economic aspects. Ergonomic and personal aspects were neglected.

With advances in computer and communication technology, a growing amount of information is delivered in a digital format. With the expansion of World Wide Web (shorter Web), the presentation of information on computer displays becomes of key

* Corresponding author. Tel.: +386 14768 806; fax: +386 14768 732.

E-mail addresses: iztok.humar@fe.uni-lj.si (I. Humar), miro.gradisar@ef.uni-lj.si (M. Gradišar), tomaz.turk@ef.uni-lj.si (T. Turk).

importance. Unfortunately, the graphical user interface design of Web pages is usually fashion-driven with a main objective to attract attention of visitors and reflect a graphical image of organization. High legibility and readability are two important factors for effective acceptance of the presented information and are rarely treated as most important.

There is a long history of research on the impact of color combinations to visual performance. It is evident from the review of related work given below, that first studies were carried out using printed material. A more recent research work is mainly focused on computer displays. Lately, there are a few studies on readability of text/background color combinations specifically aimed at Web pages.

1.1. Readability and legibility

Readability is the ability to recognize the form of a word or a group of words for contextual purposes (ANSI/HFS 100-1988). It is the property that permits an individual to read sentences from the stimulus material easily irrespective of their meanings (Fukuzumi et al., 1998). Normally it is concerned with continuous texts. Common measures of readability include reading rate, identification of misspelled words, searching for pre-specified letters/words within word lists or passages, etc. However, as readability is considered to be a psychological response, there are several factors influencing its performance. It is difficult to isolate these factors when measuring readability performance (Gradisar et al., 2005). One of principal requirements for efficient readability is the legibility of presented information (Legge et al., 1987; Sanders and McCormick, 1993). In this paper, legibility is understood as a group of visual properties of a character or symbol that determine the ease with which it can be recognized, with respect to ANSI/HFS 100-1988 standard. In principle, legibility is affected only by observer's spatial vision abilities. Up to some asymptotic level, legibility is enhanced by high luminance and color contrast, larger targets, and increased inter- and intra-target spacing (Sanders and McCormick, 1993).

1.2. Subtractive (printing) colors

The first known research about which color combinations of the posters make them most visible from a distance was published by Le Courier (Le Courier du Livre, 1912; Luckiesh, 1923). Sheldons Limited House in Leeds, a poster printing company, performed an experiment in which they put different color posters on wooden signs. Each poster contained two rows of letters. One row had well-defined letters, the other had less-defined letters like *i* and *j*. The posters were exposed to sunlight and a group of people was asked to rank the legibility of the letters while reading the posters from different distances. Apparently, the most legible poster from far distance was the poster with black letters on a yellow background. They tested 13 color combinations and provided the results as a list of combinations, ordered from the most legible to the least legible, known as Le Courier legibility table (Table 1).

Surprisingly, the most widely used combination of printed text, black letters on a white background, is only in the sixth position of legibility. However, the scope of difference between ranks is not given. Consequently, one does not know how much worse the red on green combination is than the black on yellow combination or any other combination. Detailed statements regarding the conditions of the experiment such as the number of subjects, kind of ink and paper used, size of type, line width, text used, etc., are omitted. For these reasons more definite information was needed as a guide to improved advertising practice.

Table 1
Le Courier legibility table

Rank	Text/background
1	Black/yellow
2	Green/white
3	Red/white
4	Blue/white
5	White/blue
6	Black/white
7	Yellow/black
8	White/red
9	White/green
10	White/black
11	Red/yellow
12	Green/red
13	Red/green

Between 1928 and 1963 M. A. Tinker and D.G. Paterson carried out a comprehensive research of legibility and readability (Tinker and Paterson, 1928, 1929a–e, 1940, 1942, 1944, 1946; Tinker, 1955, 1963). Among other parameters of printed material they studied the influence of colors as well. They used 10 color combinations. Eight of them were comparable to the combinations from Le Courier table, while another two were chosen from the available colored paper stocks. Students were tested with the Chapman–Cook speed of reading test. The obtained results differ from those of Le Courier table in five out of eight cases. The most important difference is the first place of the black on white combination and the fourth of the black on yellow combination. They argued that readability does not depend on color as such but on brightness differences. In fact the results of Tinker and Paterson's research are hardly comparable with those from Le Courier table since they tested readability and not the legibility alone. Despite those differences in legibility and readability tables the main common characteristic is that, generally, both recommended dark characters on a light background.

1.3. Integrative (display) colors

The computer displays differ from printed texts on optical characteristics: the printed colors are subtractive, while cathode ray tube (CRT) displays produce integrative colors. In a CRT display, the image is produced by an energized beam of electrons bombarding a thin layer of phosphorus material. The beam is activated while scanning across the visible portion of the image and deactivated when scanning where there is no image. This results in a flickering picture on a CRT display. Previous studies (Gould et al., 1987; Dillon, 1992) report that image quality of integrative-color display is inferior to the subtractive-color prints. It has been shown that the workers performed tasks about 30% slower with a CRT display than with paper. The workers also complained more about visual fatigue and visual strain.

Because of different optical characteristics of printed texts and images presented on the display, the effects of color combinations on visual performances described in the previous subsection may not directly apply to CRT displays. This conclusion gives motivation to further investigate the influence of a color combination on visual performance using CRT displays.

Although some of the early work failed to identify specific color combinations that perform better in readability from the computer displays (Radl, 1980; Pace, 1984), it was shown that a color combination of text and background is an important characteristic of visual stimuli that may affect visual performance. Some studies (Bruce and Foster, 1982; Murch, 1985; Matthews and

Mertins, 1987) found that inappropriate use of color can result in poor performance and a higher incidence of visual discomfort. They suggested avoiding a combination of red, green and blue.

Many studies tried to explain the differences in visual performance by luminance contrast. Bruce and Foster (1982) compared 42 color combinations on a Viewdata display and found a positive correlation between luminance contrast and the rank order of reading speed for all background colors except green. Small (1982) performed an experiment that directly observed the influence of luminance contrast on the performance of detecting spelling errors in the text displayed on a screen. There was a strong positive correlation between this measure and the logarithm of the luminance contrast. The hypothesis that reading ability is sensitive to luminance contrast and insensitive to chromatic contrast was also supported by the results of Legge and Rubin (1986). They measured the reading rate of four color combinations always with black as background and four different chromaticities as text color. The colors were produced by insertion of color filters in front of the monochrome display. All of the colors were equiluminant. The results showed no influence of chromatic contrast on reading performance. However, the extensions of results from this study to other displays need to be made cautiously, since the luminance of the text used in the experiment was only 6 cd/m², which is much less than 35 cd/m², recommended by ANSI/HFS 100-1988 standard, or 200 cd/m², used on contemporary computer CRT displays. Also, the results were obtained by using only black color as background and cannot be extended to a variety of different chromatic colors that are nowadays very popular for backgrounds of Web pages. In his study, Mathews (Matthews and Mertins, 1987) used a visual search task in order to compare seven color combinations. The results indicated that the color (hue) had very little influence on visual performance and supported the hypothesis that luminance contrast is an important predictor of readability. Two very extensive experiments were conducted by Pastoor (1990). In broader experiment, 32 subjects were asked to rank the esthetic appearance of the presented screen pages on a six-point scale. A total of 440 color combinations with negative polarity and 352 combinations with positive polarity were tested. The colors were chosen from 10 different chromaticity levels using at least two different luminance contrasts. Although there were no statistically significant differences between the results of different color combinations, the instructions on how to use saturated and desaturated text/background color combinations were given. In the second experiment, a heavily reduced set of 18 color combinations was used to measure reading time, perform searching task and preference ratings. However, no statistically significant effect of color combination to the reading speed and visual search task was found.

Although luminance contrast was found to be of major importance affecting visual performance on CRT displays in the abovementioned studies, it was not sufficient to explain the results of subsequent studies. Travis et al. (1990) performed an experiment to further investigate the influence of chromatic contrast on readability. They employed 33 subjects comparing 36 color stimuli on a white background by measuring detection and discrimination of pre-specified strings among words and non-sensical anagrams presented on the screen for a short time. The results show that although the luminance contrast between text color and background color was zero, a near-perfect reading was still possible. This important finding means that purely chromatic differences may be sufficient for the visual system to maintain word identification. The same hypothesis was also supported by Shieh et al. (1997), where the percentage of correctly recognized Chinese characters was measured for five different color combinations. Shieh cautions, however, not to directly generalize these

findings, since the structure of Chinese characters is quite different from that in English and the visual processes involved may not be the same. In the second study, Shieh and Lin (2000) analyzed the visual performance of 12 color combinations compounded of yellow, red, blue, and purple by asking the participants to write down the capital English characters displayed in the central area of the screen for a very short time. The subjective preference for each color combination was also collected. Color combinations significantly affected both, visual performance and subjective preference. The results of both experiments were related and show that a blue on yellow combination performed the best, while a purple on red combination was the worst. An extensive study was carried out by Wu and Yuan (2003), in which 136 undergraduate students compared color combinations by performing the speed of reading test and subjective rating. Thirty-six combinations were composed of colors that are most frequently used in the popular application software packages (red, orange, yellow, green, blue, purple, white, and black), in three different luminance levels and three different chroma levels. The results show that reading speed was higher but the visual preference was lower when foreground luminance or chroma was lower than that of the background. Different hue combinations did affect the reading speed and visual preference, but did not override the dominant impact of relative luminance contrast and chroma contrast. They found that the impact of luminance, chroma and hue on visual preference was not consistent with their impact on reading speed; moreover, sometimes the effects were opposite. The influences of color and luminance contrast were compared in the study of Legge et al. (1990). They found that plots of reading rate versus color and luminance contrast had same shape in log-log coordinates, which differed only by a contrast scale factor. With the character size 1°, the decline in reading rate occurs at a higher contrast for color than for luminance, while the color and luminance curves for the size of 6° were nearly superimposed without any contrast scaling.

More recent studies focused on the impact of the color combinations used on the World Wide Web. For example, in their research, Hill and Scharff (1997, 1999) asked the participants to search for specific words in a text with black foreground color and different background colors (blue, yellow, light/dark gray, and white). The experiments were performed via Web browser. They found significantly better performance with gray backgrounds than with a white, yellow, and blue background. This finding, again, is not consistent with the expected effect of luminance contrast, as already discussed in some of the abovementioned studies (Travis et al., 1990; Shieh et al., 1997). In a series of two experiments, conducted by van Schaik (Ling and van Schaik, 2002; Pearson and van Schaik, 2003), subjective rating and visual search task of information in navigation bar were employed to compare the effects of different color combinations. In the first study 29 participants tested six color combinations: black on white, blue on white, black on yellow, yellow on black, red on green and green on red. In terms of speed, the yellow on black combination showed the best performance while the black on white combination was ranked highest in the subjective preference rating. Green on red was poor in both aspects. They have found some congruence between subjective and objective measures. In the second study, they compared only two different color combinations, concluding that traditional blue links should be retained as responses for blue color on white background were found to be significantly quicker than responses for red. An interesting study was conducted by Wang and Chen (2003), where a scrolling text was examined. Among different factors that were varied, the influence of color combinations was also studied. The combinations with positive polarity resulted in better performance. Lastly, Hall and Hanna (2004) examined the impact of Web page text-background color

combination on readability, retention, esthetic and behavioral intention, by measuring subjective opinion with questionnaires. Four color combinations (black on white, white on black, light blue on dark blue and cyan on black) were ranked by 36 students answering five questions on a 10-point Likert scale. The major findings were: colors with greater luminance contrast are generally ranked as more readable, while color combinations do not significantly affect the retention.

The studies cited above can be classified in three groups with regard to the methods employed for measuring visual performance. The first group comprises the studies, in which readability was measured through visual search tasks (Small, 1982; Pace, 1984; Pastoor, 1990; Ling and van Schaik, 2002; Pearson and van Schaik, 2003). Participants had to either identify misspelled words or find special words in a paragraph of non-sensical anagrams. The results were statistically significant only if a small number (two to six) of color combinations were used in the experiment. However, in these studies, it is plausible that subtle contributions from the chromatic channels may have been swamped by using large luminance contrasts or low chromatic contrasts. Hence, it may be misleading to inform designers that the visual system is more sensitive to luminance than to chromatic contrast.

The studies in a second group (Bruce and Foster, 1982; Pastoor, 1990; Wu and Yuan, 2003) estimated the readability by measuring the time needed to read a certain amount of text. However, the results did not show a significant effect of a color combination on the speed of reading. As readability is considered to be a complex psychological response it is difficult if not impossible to isolate the disturbing parameters. In many studies (Murch, 1985; Pastoor, 1990; Shieh et al., 1997; Wu and Yuan, 2003; Hall and Hanna, 2004) the participants have been asked to subjectively rate the readability of color combinations. However, it has been shown (Pastoor, 1990; Shieh et al., 1997; Wu and Yuan, 2003) that subjective ratings do not coincide with performance results. There may be an explanation for this: subjective opinion is very frequently based on the esthetic (Hall and Hanna, 2004) and stereotypic presumptions and may thus differ from the objectively measured performances.

The third group consists of the experiments (Travis et al., 1990; Shieh et al., 1997; Shieh and Lin, 2000; Wang and Chen, 2003), in which visual performance was measured as the percentage of correctly recognized characters or words. The tested stimuli were shown to the participants either for a very short time or in a relatively small size and the participants were either asked to identify them aloud or write down as many of them as possible. Although these experiments were closer to the measurement of legibility than readability, there were still other psycho-cognitive parameters influencing the results.

With respect to the effects of color on visual performance, the available experimental results of all three groups of study are inconclusive. Therefore some authors proposed further research directions. In the guidelines for further work, Hall and Hanna (2004) expressed a need for conducting more controlled systematic study of color combinations. The hues should be selected to better represent the wavelengths across the spectrum. It should be noted, that same opinion was shared by Ling and van Schaik (2002), stressing that the future research should also study the effects of simultaneously varying color combinations in the content areas of Web pages. In the conclusion of his paper, Lin (2003) also suggested to further investigate the visual performance with respect to chromaticity and luminance contrast. In their paper, Pett and Wilson (1996) pointed out that examining legibility and preference of colored lettering on colored backgrounds would provide useful information for designers, because such combinations are commonly used on instructional materials, video titles and video bulletin boards. Today, this information

would also be of great help to a large community of Web page designers. Pett and Wilson (1996) also suggested that contrary to the previously performed researches, a study with significantly more subjects and colors should be carried out.

It is evident from the above given review that a majority of authors studied readability of color combinations. The results of studies on this phenomenon are very different and sometimes in collision. Authors disagree even about the influence of luminance contrast, which seems to be the most important characteristic of color combination in regard to readability. In other words, a single scientifically proved and commonly accepted readability table of integrative color combinations does not exist and it might be that it never will. Therefore we can draw a conclusion that reading is very complex and sensitive psychological response influenced by many interrelated factors, which cannot be easily identified and controlled (Gradisar et al., 2005). It is difficult if not impossible to isolate these factors when measuring readability performance. Therefore, we decided to focus our research primarily on legibility, which is one of principal requirements for efficient readability of the presented information and is independent of esthetics, content, context, etc. of text. Such a decision is also in accordance with recommendations of some authors for further research especially when more subjects and colors were employed. Another reason for this decision was that according to our knowledge no research examined legibility of integrative colors since now and that more than 90-year-old Le Courier legibility table, which is still in use today, is probably not appropriate for computer displays. Consequently, the aim of this study was to investigate the impact of color combinations, compounded of eight colors across the whole spectrum, on legibility of information displayed as Web pages on modern CRT monitor (Resinovic et al., 1999).

The remainder of the paper is organized as follows: the description of the method is followed by the results and the discussion.

2. Method

2.1. Experimental design

This study examined legibility of different color combinations of text and background through a repeated measures design. The participants were asked to identify the characters displayed in the center of a CRT display aloud. The independent variable was a color combination of a text color and background color. The dependent variable was the number of correctly identified characters, as a measure of legibility performance.

2.2. Color combinations, luminance contrast and color difference

The colors used in the experiments must be exactly defined. This was not the case in some early studies (Bruce and Foster, 1982; Pace, 1984). Color names such as “green” or “red” can incorporate a great variety of color casts—from a light one to dark one. In some experiments, the brightness and contrast were not controlled or were adjusted by individual users to their preferred levels (Pace, 1984). In more recent studies (Shieh and Chen, 1997; Shieh and Lin, 2000; Wang and Chen, 2003) the colors were defined by means of the CIE (1931) chromaticity coordinates that are not very comprehensible for the designers. In the latest studies (Ling and van Schaik, 2002; Pearson and van Schaik, 2003; Hall and Hanna, 2004), the colors were given in RGB triplets. However, the gamma values and color temperatures of white points of a display were not given.

Table 2

RGB triplets, CIE chromaticity coordinates (x , y), luminance (L), and UCS coordinates (u' , v') of eight colors used in this study

Coordinate	Color							
	White	Yellow	Red	Magenta	Blue	Cyan	Green	Black
RGB (hex)	FFFFFF	FFFF00	FF0000	FF00FF	0000FF	00FFFF	00FF00	000000
x	0.315	0.430	0.636	0.332	0.144	0.210	0.287	0.340
y	0.320	0.483	0.320	0.168	0.073	0.330	0.598	0.320
L (cd/m ²)	235	220	63	84	22	193	181	1
u'	0.203	0.217	0.457	0.305	0.161	0.128	0.120	0.221
v'	0.464	0.548	0.517	0.347	0.183	0.454	0.561	0.468

In their study Hoadley and Jenkins (1987) found that solid colors without any patterning were the most effective to utilize in multi-color information presentations on a CRT display. In accordance with this finding and to achieve the reproduction of colors on a broad variety of different monitors, color settings, and browsers, the colors used in this study were chosen from a non-dithering Web-safe color palette (Lehn and Stern, 2000), which consists of 216 different colors. Although a very large number of color combinations might be utilized in an experiment of this kind, it was necessary to limit the present study to a smaller number of well-defined colors. The colors chosen for the experiment were the elementary colors representing the wavelengths across the spectrum: (1) white, (2) yellow, (3) red, (4) magenta, (5) blue, (6) cyan, (7) green, and (8) black. The RGB triplets,¹ CIE (1931) chromaticity coordinates (x , y) and lightness (L), measured on the display, and the corresponding CIE (1986) Uniform Color Space coordinates (u' , v') of these eight colors are given in Table 2.

Each of eight colors was combined with every other different color to make 56 text/background color combinations. Table 3 depicts color combinations, their luminance contrasts and color differences as well as their polarities.

At this point, it should be noted that there are several definitions of luminance contrast and color difference. ANSI/HFS 100-1988 standard defines the luminance contrast ratio (C_R) as a quotient of higher and lower luminance of text and background: $C_R = L_{\max}/L_{\min}$. In the same standard, luminance contrast (C) is defined as $C = (L_{\max} - L_{\min})/L_{\min}$. On the other hand, in their calculations, Shieh et al. (1997) evaluated luminous contrast as $\Delta Y/Y_{\max}$, which is equivalent to $(L_{\max} - L_{\min})/L_{\max}$. In this work, the luminance contrast was evaluated by the third (Shieh) definition. These values were used when examining the influence of luminance contrast on legibility with the regression models in Section 3.2.

There are also many definitions for color difference, based on different color spaces. The CIE colorimetry relies on the international XYZ system (CIE, 1931) that mathematically defines the absolute stimulation rates of red (X), green (Y) and blue (Z) retinal receptors for any color stimulus. The X , Y , and Z primary color axes are not orthogonal and may be transformed into mathematically orthogonal luminance (Y) and chromaticity (x , y) axes that correspond to color attributes of brightness, hue and saturation. In order to achieve perceptual uniformity, CIE Uniform Color Space has been adopted in 1976. The coordinates, u' and v' , are derived from CIE (1931) stimulus values by the equations: $u' = 4X/(X+15Y+3Z)$ and $v' = 9X/(X+15Y+3Z)$. Based on these coordinates, two candidates for uniform color spaces have been proposed by CIE in 1976 (CIE, 1986): L^* , u^* , v^* and L^* , a^* , b^* . The

Table 3

Luminance contrast, color difference, and polarity of color combinations used in this study

No.	Grp.	N	Text/bckg	Lum. contrast	Color diff.	Polarity
1	A	90	White/black	1.00	154.48	Negative
2	A	90	Black/white	1.00	154.48	Positive
3	A	90	Yellow/black	1.00	154.31	Negative
4	A	90	Black/yellow	1.00	154.31	Positive
5	A	90	Cyan/black	0.99	157.88	Negative
6	A	90	Black/cyan	0.99	157.88	Positive
7	A	90	Green/black	0.99	158.56	Negative
8	A	90	Black/green	0.99	158.56	Positive
9	A	90	White/blue	0.91	141.39	Negative
10	A	90	Blue/white	0.91	141.39	Positive
11	B	90	Yellow/blue	0.90	141.10	Negative
12	B	90	Blue/yellow	0.90	141.10	Positive
13	B	90	Cyan/blue	0.89	137.88	Negative
14	B	90	Blue/cyan	0.89	137.88	Positive
15	B	90	Magenta/black	0.99	156.26	Negative
16	B	90	Black/magenta	0.99	156.26	Positive
17	B	90	Green/blue	0.88	137.07	Negative
18	B	90	Blue/green	0.88	137.07	Positive
19	B	90	Red/black	0.98	175.44	Negative
20	B	90	Black/red	0.98	175.44	Positive
21	C	80	White/red	0.73	146.83	Negative
22	C	80	Red/white	0.73	146.83	Positive
23	C	80	Yellow/red	0.71	141.44	Negative
24	C	80	Red/yellow	0.71	141.44	Positive
25	C	80	White/magenta	0.64	106.44	Negative
26	C	80	Magenta/white	0.64	106.44	Positive
27	C	80	Magenta/yellow	0.62	101.20	Positive
28	C	80	Yellow/magenta	0.62	101.20	Negative
29	C	80	Cyan/red	0.67	159.47	Negative
30	C	80	Red/cyan	0.67	159.47	Positive
31	D	80	Red/green	0.65	159.81	Positive
32	D	80	Green/red	0.65	159.81	Negative
33	D	80	Magenta/cyan	0.56	108.95	Positive
34	D	80	Cyan/magenta	0.56	108.95	Negative
35	D	80	Blue/black	0.95	149.64	Negative
36	D	80	Black/blue	0.95	149.64	Positive
37	D	80	Magenta/blue	0.74	126.13	Negative
38	D	80	Blue/magenta	0.74	126.13	Positive
39	D	80	Magenta/green	0.54	107.46	Positive
40	D	80	Green/magenta	0.54	107.46	Negative
41	E	64	Red/blue	0.65	148.40	Negative
42	E	64	Blue/red	0.65	148.40	Positive
43	E	64	White/green	0.23	46.96	Negative
44	E	64	Green/white	0.23	46.96	Positive
45	E	64	Yellow/green	0.18	45.02	Negative
46	E	64	Green/yellow	0.18	45.02	Positive
47	E	64	White/cyan	0.18	38.91	Negative
48	E	64	Cyan/white	0.18	38.91	Positive
49	F	64	Yellow/cyan	0.12	37.59	Negative
50	F	64	Cyan/yellow	0.12	37.59	Positive
51	F	64	Magenta/red	0.25	67.88	Negative
52	F	64	Red/magenta	0.25	67.88	Positive
53	F	64	Cyan/green	0.06	10.27	Negative
54	F	64	Green/cyan	0.06	10.27	Positive
55	F	64	White/yellow	0.06	11.17	Negative
56	F	64	Yellow/white	0.06	11.17	Positive

N defines the number of times a particular combination was tested in this study; Grp. defines the group to which a particular combination belongs to.

definitions of coordinates of these color spaces are given in the Appendix A.

Different color difference formulas were developed for particular estimations. The definition of color difference $\Delta E(L^*, u^*, v^*)$ in L^* , u^* , v^* color space might be considered as the Pythagorean sum of luminance and chromatic contrast (see Appendix A). ANSI/HFS 100-1988 suggested using this color difference for assessing the color discriminability. In term of legibility, a simplified color difference formulation, proposed by Lippert (1986), outperformed other definitions of color difference. Consequently, this metric is recommended by ANSI/HFS 100-1988 for assessing character

¹ In Hypertext Markup Language (HTML), colors are specified as hexadecimal RGB triplets, as defined in "A Standard Default Color Space for the Internet", <http://www.w3.org/Graphics/Color/sRGB>.

legibility and was also employed in this study for evaluating the influence of color difference on legibility in the regression models (Section 3.3). The color difference is given as: $\Delta E(Y, u', v') = ((155\Delta Y/Y_{\max})^2 + (367\Delta u')^2 + (167\Delta v')^2)^{0.5}$, where ΔY is the difference in luminance of a text and background, $Y_{\max}$ is the maximum luminance of a text or background, $\Delta u'$ is the difference between u' coordinates of a text and background, and $\Delta v'$ is the difference between v' coordinates of a text and background. Many other color difference formulas have been developed in the color difference research, i.e. CMC (Clarke et al., 1984), CIE94 (CIE, 1995), BFD (Luo and Rigg, 1987), and CIEDE2000 (Luo et al., 2001), but are not taken in consideration in this work.

The polarity is a property of a color combination defined by the luminance ratio between background and text colors. If the luminance of a background is higher than text luminance, the polarity is considered positive, otherwise it is negative.

Since the number of color combinations was too large for a single participant to comfortably view in a single session, the combinations were sorted by decreasing luminance contrast and divided into six groups, denoted with letters from A to F. The first four groups (from A to D) consisted of 10 color combinations while the last two groups (E and F) contained 8 color combinations each. A black on white combination was added to each group as a reference. Having a limited number of participants, the first two groups of most contrast combinations were tested by 90 participants, the second two groups were tested by 80 participants and the last two groups by 64 participants.

2.3. Participants

In response to advertisements at the introductory course of Informatics at the first-year of study at the University of Ljubljana, 477 students were recruited as volunteers. The group consisted of 209 males and 268 females. The average age of the participants was 19 (ranging from 18 to 21). All participants had normal or corrected to normal visual acuity, and were tested with the Ishihara test for color blindness to identify those with color vision deficiencies (protanopia, deuteranopia, and tritanopia). Nine participants (six male and three female) failed on this test. Their data were collected, yet not included in the results examination. All participants possessed at least the basic level of computer literacy and had experience with the Web.

2.4. Apparatus, materials, environment and viewing conditions

The experimental tasks were presented on a 21" Dell CRT P1130 Trinitron display. The screen resolution was 1280×1024 pixels without interpolation, with pixel pitch of 0.24 mm. The refresh rate was 85 Hz (non-interlaced), Chromatic resolution was 24 bit. The chromaticity of white point was set to D_{65} , gamma value to 2.0, and the luminance level of white point was greater than 200 cd/m^2 . The display was calibrated with the X-Rite Color Monitor Optimizer Colorimeter (Monitor Optimizer Software Guide). The luminance values of the ambient and the CIE color values were measured using the Minolta Color Analyzer XY-DC/XY-1.

To assure adequate and equal testing environment to all participants, the viewing conditions were arranged in conformance with the ANSI/HFS 100-1988, ISO 9241-3 (1992) and ISO 12646 (2004) standards. Following these standards, the ambient was neutral (light brown) with no areas causing glares or reflections on the monitor screen. The mean ambient illumination was below 300 lx . The only light source in the room was a shielded lamp on the ceiling, while other light sources had been curtained.

The participants were seated in a position where the distance between the screen and the participants' eyes was 1.6 m. This

distance was chosen as an extrapolation of Kroemer's (1993) suggestion that users shall seat at a distance exceeding 1 m from a 15" display. The screen center was slightly below the participants' eyes height forming a viewing angle of approximately 15° . The inclination angle of the monitor was 105° in accordance with the previous studies of Shieh et al. (1997) and Shieh and Lin (2000).

2.5. Task and procedure

This experiment measured the legibility of different color combinations through a visual identification task. The participants, who tested the color combinations from the first four groups (from A to D) were exposed to a sequence of eleven color combinations, while those, who tested the color combinations from the last two groups (E and F) were exposed to nine color combinations. In each sequence, the black on white combination was added to the beginning for a reference. The sequence of color combinations was ascribed to each participant according to the Latin Square design.

The participants were asked to identify the characters displayed in the center of a CRT display aloud (Fig. 1; Humar and Gradisar, 2003). In each test, 21 uppercase alphabetic characters were presented in six lines, organized in conformance with the Snellen chart (Snellen, 1862): one character in the first row, followed by two characters in the second row, three characters in the third row, four characters in the fourth row, five characters in the fifth row, and six characters in the last row. The sizes of characters decreased from 3.8 mm in the top row to 1.4 mm in the bottom row with line spacing ranging from 5.8 mm to 3.9 mm and one blank space between the characters. The font used in this experiment was Times New Roman, since it outperformed other typefaces in terms of legibility in many studies (Tinker and Paterson, 1932; Poulton, 1972). The characters used as stimuli were randomly selected from a group of characters which—in previous testing (Snellen, 1862)—had distinguished as unlikely to be confounded: "E", "F", "P", "T", "O", "Z", "L", "C", "D", and "B".

Before performing the task, the participants were presented with a set of instructions. They also received one practice trial to familiarize with the test. At the beginning of each test, a participant had to press a key on the keyboard to proceed. As mentioned earlier, the black on white color combination was tested first as a reference, followed by a sequence of other color combinations. Between testing of two consecutive color combinations, the participants were exposed to a screen with a relaxing

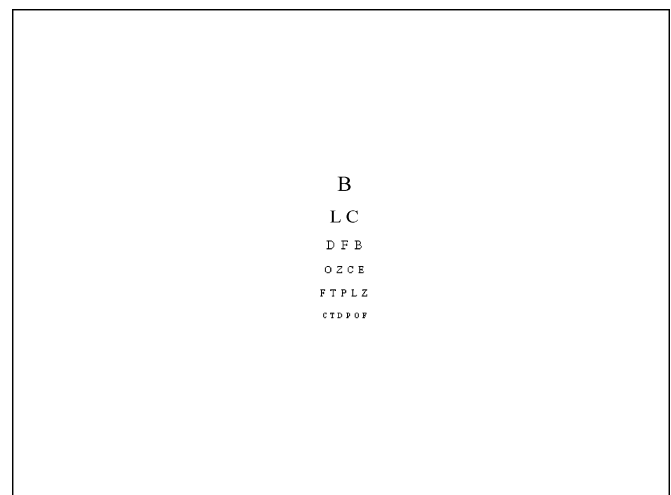


Fig. 1. Visual performance test: the participants were asked to identify the character displayed in the center of a CRT display aloud.

light gray color to neutralize the effect of the previous color combination. Each participant identified as many characters as he/she could see. The experimenter carefully recorded the characters.

The participants took approximately 20 min to complete the task. Upon completion the participants were thanked and then fully debriefed.

2.6. Dependent measures and data analysis

In this study, the legibility was measured through a visual performance task. The independent variable was a color combination of a text color and background color. The dependent variable was the number of correctly identified characters as a measure of legibility performance. In order to eliminate the influence of a participant's visual acuity, the results of an individual participant were normalized—defined by dividing the number of correctly identified characters in a particular color combination with the result of the reference (black on white) combination. In the next section, all analyses were performed by using normalized results for a given color combination (shorter addressed as scores), unless otherwise noted.

A separate analysis of variance (ANOVA) was firstly conducted to assess the effect of a color combination in terms of differences in mean scores. Furthermore, *t*-test was used to study the differences in legibility among pairs of color combinations. The regression analysis was performed to estimate the strength and the nature of the relationship between legibility and two of the color characteristics, namely luminance contrast and color difference. Each of these two characteristics was examined separately. A closer inspection of definitions of these two characteristics (reviewed in Section 2.2) reveal that a high level of correlation between them can be expected (for instance, both characteristics include luminance), asserting against the inclusion of both measures in the same model, since they include similar information. The preliminary calculations confirmed this problem and showed a high level of multicollinearity between luminance contrast and color difference. Lastly, the influence of polarity of color combinations was thoroughly explored by using the analysis of variance in comparing the observed scores between specific groups of color combinations according to their color characteristics.

3. Results

3.1. The legibility through the numbers of correctly identified characters

The mean values of normalized results and standard deviations from the mean values for all color combinations under consideration are presented in Table 4, ordered by descending mean scores. The mean values for various color combinations ranged from 0.172 to 1.119. A closer inspection of the results shows that the best results were achieved with yellow on black, cyan on black, white on blue, black on yellow, white on black, and green on black color combinations. On the other hand, the results for the color pairs of black and blue, red and magenta, green and cyan, and yellow and white were relatively low.

A separate analysis of variance was conducted to assess the influence of a color combination in terms of differences in mean scores. According to the ANOVA test, color combinations significantly affected the number of correctly identified characters. The estimated *F* value of 154.977 gave $p < 0.001$ with (56, 4835) degrees of freedom.

A closer inspection of scores is possible with a comparison of mean value pairs, where a mean score for a given color

Table 4

The mean values of results and standard deviations for all color combinations under consideration

No.	Text/bckg	Result		Score	
		Mean	Std. dev.	Mean	Std. dev.
3	Yellow/black	16.76	3.26	1.119	0.218
5	Cyan/black	16.84	3.44	1.118	0.212
9	White/blue	16.66	3.59	1.100	0.181
4	Black/yellow	16.63	3.82	1.096	0.209
1	White/black	16.52	3.55	1.094	0.211
7	Green/black	16.44	3.87	1.082	0.215
2	Black/white	16.48	4.04	1.077	0.180
11	Yellow/blue	16.42	3.91	1.071	0.171
10	Blue/white	16.26	4.13	1.060	0.182
12	Blue/yellow	16.37	4.29	1.053	0.158
8	Black/green	15.98	4.26	1.041	0.203
15	Magenta/black	15.79	3.74	1.039	0.194
17	Green/blue	15.88	4.12	1.031	0.171
18	Blue/green	15.89	4.76	1.007	0.197
6	Black/cyan	15.66	4.74	1.005	0.214
14	Blue/cyan	15.88	4.96	1.001	0.212
19	Red/black	15.29	4.14	0.994	0.184
13	Cyan/blue	15.56	4.65	0.993	0.188
24	Red/yellow	17.20	4.16	0.990	0.158
26	Magenta/white	17.21	4.15	0.989	0.160
21	White/red	16.85	4.06	0.978	0.191
22	Red/white	17.04	4.28	0.978	0.149
27	Magenta/yellow	16.99	4.32	0.973	0.154
42	Blue/red	17.72	3.36	0.968	0.128
38	Blue/magenta	17.15	3.94	0.967	0.180
30	Red/cyan	16.81	4.43	0.967	0.186
20	Black/red	15.01	4.54	0.963	0.189
33	Magenta/cyan	17.16	4.45	0.960	0.161
16	Black/magenta	14.87	4.37	0.951	0.175
23	Yellow/red	16.44	4.25	0.948	0.189
25	White/magenta	16.44	4.24	0.946	0.162
31	Red/green	16.80	4.51	0.936	0.163
39	Magenta/green	16.64	4.58	0.929	0.166
37	Magenta/blue	16.50	4.08	0.929	0.146
28	Yellow/magenta	15.88	4.52	0.909	0.193
41	Red/blue	16.20	4.47	0.871	0.188
44	Green/white	15.86	4.39	0.859	0.192
48	Cyan/white	15.70	4.62	0.846	0.203
46	Green/yellow	15.59	4.44	0.844	0.195
50	Cyan/yellow	15.48	4.55	0.832	0.195
29	Cyan/red	14.30	4.66	0.824	0.215
43	White/green	14.38	4.36	0.776	0.192
34	Cyan/magenta	14.00	4.90	0.770	0.194
32	Green/red	13.89	4.87	0.767	0.196
47	White/cyan	13.44	4.43	0.725	0.195
45	Yellow/green	13.39	4.36	0.721	0.189
40	Green/magenta	12.64	4.71	0.700	0.199
49	Yellow/cyan	12.20	4.83	0.647	0.213
35	Blue/black	10.53	5.27	0.576	0.239
36	Black/blue	10.39	5.83	0.557	0.268
52	Red/magenta	5.23	3.50	0.274	0.169
51	Magenta/red	4.31	2.69	0.224	0.131
53	Cyan/green	4.19	2.54	0.218	0.123
54	Green/cyan	3.41	2.33	0.177	0.114
56	Yellow/white	3.41	2.31	0.177	0.112
55	White/yellow	3.25	1.88	0.172	0.090

combination is compared with mean scores for other color combinations. Since 56 different color combinations were investigated, a corresponding table contains 55 rows and columns. Table 5 shows the significances from corresponding *t*-tests, and provides additional guidance when choosing between two particular color combinations, especially when used and compared with the scores from Table 4. As evident from Table 5, the majority of statistically significant differences in means were found when comparing the color combinations from the last three groups (D–F) with those in the first three groups (A–C).

Table 5
The significances for differences in means for pairs of color combinations

no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	no.	
grp. A	A	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B	B	C	C	C	C	C	C	C	C	grp.	
1 A																													A	1
2 A	1.17																												A	2
3 A	0.41	1.38																											A	3
4 A	0.04	1.21	0.30																										A	4
5 A	0.38	1.40	0.01	0.29																									A	5
6 A	0.91	0.07	1.09	1.01	1.13																								A	6
7 A	0.16	1.02	0.51	0.18	0.57	0.82																							A	7
8 A	0.64	0.57	0.91	0.75	1.03	0.57	0.62																						A	8
9 A	0.10	1.40	0.31	0.06	0.27	1.03	0.28	0.80																					A	9
10 A	0.43	0.90	0.79	0.49	0.85	0.71	0.37	0.32	0.60																				A	10
11 B	-0.14	0.48	-0.28	-0.16	-0.28	0.43	-0.07	0.19	-0.18	0.07																			B	11
12 B	-0.25	0.37	-0.40	-0.27	-0.39	0.32	-0.18	0.08	-0.29	-0.04	0.22																		B	12
13 B	-0.66	-0.05	-0.80	-0.67	-0.79	-0.09	-0.58	-0.33	-0.69	-0.45	0.87	0.76																	B	13
14 B	-0.59	0.01	-0.73	-0.61	-0.73	-0.03	-0.52	-0.27	-0.63	-0.39	0.70	0.58	0.09																B	14
15 B	-0.35	0.27	-0.49	-0.36	-0.48	0.23	-0.27	-0.02	-0.38	-0.14	0.37	0.17	0.57	0.38															B	15
16 B	-0.96	-0.37	-1.10	-0.98	-1.10	-0.40	-0.89	-0.64	-1.00	-0.76	1.31	1.17	0.47	0.58	1.05														B	16
17 B	-0.40	0.22	-0.54	-0.41	-0.54	0.18	-0.32	-0.07	-0.43	-0.19	0.47	0.31	0.45	0.33	0.09	1.03													B	17
18 B	-0.56	0.05	-0.70	-0.57	-0.70	0.01	-0.49	-0.23	-0.60	-0.36	0.72	0.55	0.15	0.05	0.33	0.68	0.31												B	18
19 B	-0.65	-0.04	-0.79	-0.66	-0.79	-0.08	-0.58	-0.32	-0.69	-0.45	0.92	0.72	0.01	0.08	0.57	0.46	0.13	0.13											B	19
20 B	-0.88	-0.28	-1.01	-0.89	-1.01	-0.31	-0.80	-0.55	-0.92	-0.68	1.27	1.06	0.34	0.42	0.85	0.14	0.74	0.44	0.47										B	20
21 C	-1.48	-0.98	-1.60	-1.50	-1.60	-0.99	-1.42	-1.20	-1.52	-1.32	-1.38	-1.28	-0.92	-0.96	-1.19	-0.64	-1.15	-1.00	-0.92	-0.71									C	21
22 C	-1.50	-0.99	-1.62	-1.51	-1.61	-1.00	-1.43	-1.21	-1.54	-1.33	-1.39	-1.29	-0.93	-0.97	-1.20	-0.65	-1.16	-1.01	-0.93	-0.72	0.01								C	22
23 C	-1.70	-1.21	-1.81	-1.71	-1.81	-1.21	-1.64	-1.42	-1.74	-1.53	-1.59	-1.50	-1.14	-1.19	-1.41	-0.87	-1.37	-1.22	-1.15	-0.94	0.41	0.41							C	23
24 C	-1.41	-0.90	-1.53	-1.42	-1.53	-0.91	-1.34	-1.12	-1.45	-1.24	-1.30	-1.20	-0.83	-0.88	-1.11	-0.55	-1.07	-0.92	-0.84	-0.63	0.20	0.22	0.73						C	24
25 C	-1.73	-1.24	-1.84	-1.74	-1.84	-1.24	-1.66	-1.45	-1.77	-1.56	-1.62	-1.53	-1.17	-1.21	-1.44	-0.89	-1.40	-1.25	-1.18	-0.97	0.50	0.55	0.05	0.75					C	25
26 C	-1.42	-0.91	-1.54	-1.43	-1.54	-0.91	-1.35	-1.13	-1.46	-1.25	-1.31	-1.21	-0.84	-0.89	-1.12	-0.56	-1.08	-0.93	-0.85	-0.64	0.15	0.19	0.58	0.02	0.74				C	26
27 C	-1.53	-1.02	-1.65	-1.54	-1.64	-1.03	-1.46	-1.25	-1.57	-1.36	-1.42	-1.33	-0.96	-1.01	-1.23	-0.68	-1.20	-1.04	-0.96	-0.75	0.07	0.07	0.38	0.34	0.42				C	27
28 C	-1.99	-1.53	-2.10	-2.00	-2.10	-1.52	-1.93	-1.72	-2.03	-1.83	-1.89	-1.81	-1.45	-1.49	-1.71	-1.18	-1.68	-1.53	-1.46	-1.25	0.87	0.93	0.67	1.26	0.64	1.10	0.89		C	28
29 C	-2.67*	-2.28*	-2.76*	-2.68*	-2.76*	-2.24	-2.61*	-2.43*	-2.71*	-2.54*	-2.59*	-2.52*	-2.19	-2.22	-2.43*	-1.94	-2.40*	-2.26	-2.19	-2.00	1.53	1.57	1.41	1.71	1.34	1.67	1.54	1.01	C	29
30 C	-1.57	-1.07	-1.68	-1.58	-1.68	-1.07	-1.50	-1.29	-1.61	-1.40	-1.46	-1.37	-1.00	-1.05	-1.28	-0.72	-1.24	-1.09	-1.01	-0.80	0.20	0.18	0.34	0.51	0.35	0.37	0.12	0.99	C	30
31 D	-1.80	-1.32	-1.91	-1.81	-1.91	-1.31	-1.74	-1.53	-1.84	-1.64	-1.70	-1.61	-1.25	-1.29	-1.52	-0.97	-1.48	-1.33	-1.25	-1.05	-0.31	-0.31	-0.09	-0.39	-0.07	-0.38	-0.27	0.20	D	31
32 D	-3.18*	-2.86*	-3.26*	-3.19*	-3.27*	-2.79*	-3.13*	-2.97*	-3.23*	-3.07*	-3.12*	-3.06*	-2.75*	-2.77*	-2.97*	-2.52*	-2.95*	-2.81*	-2.76*	-2.58*	-1.78	-1.80	-1.59	-1.87	-1.58	-1.86	-1.77	-1.31	D	32
33 D	-1.62	-1.13	-1.74	-1.63	-1.74	-1.13	-1.56	-1.34	-1.66	-1.46	-1.52	-1.42	-1.06	-1.10	-1.33	-0.78	-1.29	-1.14	-1.06	-0.86	-0.13	-0.13	0.08	-0.21	0.11	-0.20	-0.10	0.38	D	33
34 D	-3.15*	-2.82*	-3.24*	-3.16*	-3.24*	-2.76*	-3.10*	-2.94*	-3.20*	-3.04*	-3.09*	-3.03*	-2.72*	-2.74*	-2.94*	-2.49*	-2.92*	-2.78*	-2.72*	-2.55*	-1.75	-1.76	-1.56	-1.84	-1.55	-1.83	-1.73	-1.27	D	34
35 D	-5.04**	-5.01**	-5.06**	-5.04**	-5.07**	-4.84**	-5.01**	-4.95**	-5.09**	-5.02**	-5.05**	-5.03**	-4.85**	-4.84**	-4.96**	-4.73**	-4.97**	-4.87**	-4.86**	-4.75**	-3.98**	-4.02**	-3.84**	-4.06**	-3.86**	-4.06**	-4.00**	-3.64**	D	35
36 D	-5.20**	-5.20**	-5.22**	-5.20**	-5.23**	-5.03**	-5.17**	-5.12**	-5.25**	-5.19**	-5.22**	-5.21**	-5.04**	-5.02**	-5.13**	-4.94**	-5.15**	-5.06**	-5.05**	-4.95**	-4.18**	-4.23**	-4.06**	-4.27**	-4.08**	-4.26**	-4.21**	-3.86**	D	36
37 D	-1.85	-1.38	-1.97	-1.87	-1.97	-1.37	-1.79	-1.58	-1.89	-1.69	-1.75	-1.66	-1.30	-1.35	-1.57	-1.03	-1.54	-1.38	-1.31	-1.10	-0.36	-0.36	-0.15	-0.44	-0.13	-0.43	-0.33	0.15	D	37
38 D	-1.56	-1.07	-1.68	-1.58	-1.68	-1.07	-1.50	-1.28	-1.60	-1.40	-1.46	-1.36	-1.00	-1.04	-1.27	-0.72	-1.23	-1.08	-1.00	-0.80	-0.08	-0.07	0.14	-0.16	0.16	-0.15	-0.04	0.43	D	38
39 D	-1.85	-1.37	-1.96	-1.86	-1.96	-1.36	-1.79	-1.58	-1.89	-1.69	-1.75	-1.66	-1.30	-1.34	-1.57	-1.03	-1.53	-1.38	-1.30	-1.10	-0.36	-0.36	-0.14	-0.44	-0.12	-0.43	-0.32	0.15	D	39
40 D	-3.81**	-3.57**	-3.87**	-3.82**	-3.88**	-3.47*	-3.76**	-3.63**	-3.85**	-3.72**	-3.77**	-3.72**	-3.45*	-3.46*	-3.63**	-3.25*	-3.62**	-3.50*	-3.45*	-3.30*	-2.50*	-2.52*	-2.32*	-2.58*	-2.31*	-2.57*	-2.49*	-2.05	D	40
41 E	-3.28*	-3.09*	-3.34*	-3.29*	-3.34*	-3.00*	-3.25*	-3.14*	-3.32*	-3.21*	-3.25*	-3.21*	-2.98*	-2.99*	-3.14*	-2.82*	-3.13*	-3.02*	-2.99*	-2.86*	-2.16	-2.18	-2.01	-2.24	-2.01	-2.23	-2.16	-1.79	E	41
42 E	-2.68*	-2.39*	-2.75*	-2.69*	-2.75*	-2.34*	-2.63*	-2.49*	-2.72*	-2.58*	-2.62*	-2.57*	-2.30*	-2.33*	-2.49*	-2.11	-2.48*	-2.36*	-2.31*	-2.16	-1.47	-1.48	-1.30	-1.55	-1.30	-1.54	-1.46	-1.06	E	42
43 E	-3.94**	-3.86**	-3.97**	-3.95**	-3.98**	-3.73**	-3.91**	-3.84**	-3.98**	-3.90**	-3.94**	-3.91**	-3.73**	-3.72**	-3.84**	-3.60**	-3.85**	-3.76**	-3.73**	-3.63**	-2.94*	-2.97*	-2.81*	-3.02*	-2.82*	-3.01*	-2.95*	-2.62*	E	43
44 E	-3.36*	-3.18*	-3.41*	-3.37*	-3.42*	-3.09*	-3.33*	-3.22*	-3.40*	-3.30*	-3.33*	-3.29*	-3.07*	-3.08*	-3.22*	-2.91*	-3.21*	-3.11*	-3.07*	-2.95*	-2.25	-2.27*	-2.10	-2.33*	-2.10	-2.32*	-2.25	-1.88	E	44
45 E	-4.35**	-4.34**	-4.37**	-4.35**	-4.37**	-4.19**	-4.32**	-4.27**	-4.39**	-4.33**	-4.36**	-4.35**	-4.19**	-4.18**	-4.28**	-4.10**	-4.30**	-4.21**	-4.20**	-4.11**	-3.44*	-3.48*	-3.33*	-3.51*	-3.34*	-3.51*	-3.46*	-3.15*	E	45
46 E	-3.47*	-3.30*	-3.52*	-3.47*	-3.52*	-3.21*	-3.43*	-3.33*	-3.51*	-3.41*	-3.44*	-3.41*	-3.19*	-3.20*	-3.33*	-3.04*	-3.33*	-3.23*	-3.19*	-3.07*	-2.38*	-2.40*	-2.23	-2.45*	-2.23	-2.45*	-2.38*	-2.01	E	46
47 E	-4.31**	-4.30**	-4.33**	-4.32**	-4.34**	-4.15**	-4.29**	-4.24**	-4.36**	-4.30**	-4.33**	-4.31**	-4.16**	-4.14**	-4.25**	-4.06**	-4.26**	-4.18**	-4.16**	-4.07**	-3.40*	-3.44*	-3.29*	-3.47*	-3.30*	-3.47*	-3.42*	-3.11*	E	47
48 E	-3.44*	-3.28*	-3.49*	-3.45*	-3.50*	-3.18*	-3.41*	-3.31*	-3.49*	-3.38*	-3.42*	-3.38*	-3.17*	-3.17*	-3.31*	-3.01*	-3.31*	-3.21*	-3.17*	-3.05*	-2.35*	-2.38*	-2.21	-2.43*	-2.21	-2.42*	-2.35*	-1.99	E	48
49 F	-4.89**	-4.99**	-4.89**	-4.90**	-4.90**	-4.81**	-4.88**	-4.87**	-4.94**	-4.91**	-4.93**	-4.94**	-4.83**	-4.80**	-4.88**	-4.78**	-4.90**	-4.84**	-4.84**	-4.78**	-4.13**	-4.18**	-4.05**	-4.20**	-4.07**	-4.20**	-4.16**	-3.90**	F	49
50 F	-3.55*	-3.40*	-3.59**	-3.56*	-3.60**	-3.30*	-3.52*	-3.42*	-3.59**	-3																				

no.	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	no.
grp.	C	C	D	D	D	D	D	D	D	D	D	D	E	E	E	E	E	E	E	E	F	F	F	F	F	F	F	F	grp.
1 A																													A 1
2 A																													A 2
3 A																													A 3
4 A																													A 4
5 A																													A 5
6 A																													A 6
7 A																													A 7
8 A																													A 8
9 A																													A 9
10 A																													A 10
11 B																													B 11
12 B																													B 12
13 B																													B 13
14 B																													B 14
15 B																													B 15
16 B																													B 16
17 B																													B 17
18 B																													B 18
19 B																													B 19
20 B																													B 20
21 C																													C 21
22 C																													C 22
23 C																													C 23
24 C																													C 24
25 C																													C 25
26 C																													C 26
27 C																													C 27
28 C																													C 28
29 C																													C 29
30 C	1.54																												C 30
31 D	0.94	-0.23																											D 31
32 D	-0.58	-1.71	1.48																										D 32
33 D	1.11	-0.05	0.28	1.60																									D 33
34 D	-0.54	-1.68	1.53	0.04	1.69																								D 34
35 D	-3.04*	-3.93**	2.05	1.30	2.16	1.55																							D 35
36 D	-3.29*	-4.14**	2.06	1.37	2.15	1.56	0.20																						D 36
37 D	0.89	-0.28	0.08	1.44	0.43	1.46	2.14	2.06																					D 37
38 D	1.16	0.00	0.34	1.51	0.08	1.51	1.97	2.06	0.49																				D 38
39 D	0.89	-0.28	0.07	1.39	0.38	1.44	2.11	2.10	0.00	0.38																			D 39
40 D	-1.35	-2.43*	1.71	0.63	1.86	0.75	1.08	1.20	1.99	1.89	1.84																		D 40
41 E	-1.19	-2.11	-1.96	-0.71	-2.09	-0.74	0.29	0.34	-1.93	-2.12	-1.92	-0.04																	E 41
42 E	-0.43	-1.41	-1.24	0.07	-1.38	0.03	0.45	0.49	-1.20	-1.42	-1.19	0.74	1.09																E 42
43 E	-2.06	-2.90*	-2.77*	-1.62	-2.89*	-1.65	0.65	0.93	-2.75*	-2.91*	-2.74*	-0.97	0.89	1.54															E 43
44 E	-1.29	-2.20	-2.05	-0.82	-2.18	-0.85	0.27	0.32	-2.02	-2.21	-2.01	-0.14	0.15	1.15	0.94														E 44
45 E	-2.64*	-3.40*	-3.30*	-2.23	-3.40*	-2.26	0.01	0.31	-3.29*	-3.41*	-3.27*	-1.61	1.29	1.72	0.65	1.35													E 45
46 E	-1.43	-2.33*	-2.18	-0.96	-2.31*	-0.99	1.31	1.57	-2.15	-2.33*	-2.14	-0.29	0.31	1.17	0.68	0.25	1.26												E 46
47 E	-2.59*	-3.36*	-3.26*	-2.18	-3.36*	-2.21	0.06	0.36	-3.24*	-3.37*	-3.23*	-1.56	1.33	1.71	0.61	1.37	0.07	1.16											E 47
48 E	-1.40	-2.30*	-2.16	-0.93	-2.28*	-0.96	1.33	0.29	-2.13	-2.31*	-2.12	-0.27	0.31	1.16	0.73	0.18	1.34	0.04	1.25										E 48
49 F	-3.46*	-4.11**	-4.04**	-3.12*	-4.12**	-3.15*	-0.98	-0.67	-4.03**	-4.12**	-4.01**	-2.55*	-2.15	-2.73*	-1.42	-2.07	-0.88	-1.95	-0.92	-1.97									F 49
50 F	-1.54	-2.42*	-2.28*	-1.07	-2.41*	-1.10	1.20	1.46	-2.25	-2.43*	-2.24	-0.40	-0.31	-0.97	0.49	-0.22	1.04	-0.10	0.99	-0.12	1.52								F 50
51 F	-9.71**	-9.02**	-9.29**	-10.15**	-9.15**	-10.15**	-10.56**	-10.20**	-9.38**	-9.04**	-9.31**	-10.51**	-9.05**	-8.79**	-9.40**	-9.08**	-9.60**	-8.15**	-8.03**	-9.55**	-9.08**	1.82	1.95						F 51
52 F	-8.83**	-8.37**	-8.58**	-9.14**	-8.47**	-9.14**	-9.00**	-8.63**	-8.66**	-8.38**	-8.60**	-9.31**	-8.00**	-7.91**	-8.13**	-8.01**	-8.15**	-8.03**	-8.12**	-7.99**	1.70	1.91	0.94						F 52
53 F	-9.81**	-9.10**	-9.37**	-10.27**	-9.23**	-10.27**	-10.75**	-10.39**	-9.47**	-9.12**	-9.40**	-10.65**	-9.17**	-8.90**	-9.56**	-9.21**	-9.77**	-9.26**	-9.72**	-9.21**	1.85	1.97	0.11	0.98					F 53
54 F	-10.47**	-9.59**	-9.90**	-11.04**	-9.73**	-11.03**	-11.97**	-11.60**	-10.01**	-9.61**	-9.94**	-11.56**	-9.96**	-9.56**	-10.54**	-10.02**	-10.90**	-10.10**	-10.83**	-10.04**	1.90	2.00	0.84	1.19	0.83				F 54
55 F	-10.61**	-9.69**	-10.00**	-11.20**	-9.83**	-11.19**	-12.22**	-11.86**	-10.12**	-9.71**	-10.04**	-11.75**	-10.13**	-9.69**	-10.75**	-10.19**	-11.14**	-10.28**	-11.07**	-10.22**	1.91	2.00	0.95	1.33	0.98	0.13			F 55
56 F	-10.48**	-9.60**	-9.91**	-11.05**	-9.74**	-11.04**	-11.98**	-11.62**	-10.02**	-9.62**	-9.94**	-11.57**	-9.97**	-9.57**	-10.55**	-10.03**	-10.91**	-10.11**	-10.85**	-10.05**	1.88	1.99	0.92	1.40	0.95	0.01	0.16		F 56
grp.	C	C	D	D	D	D	D	D	D	D	D	D	E	E	E	E	E	E	E	E	F	F	F	F	F	F	F	F	grp.
no.	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	no.

Note: *significantly different at 0.05 level; **0.001 level.

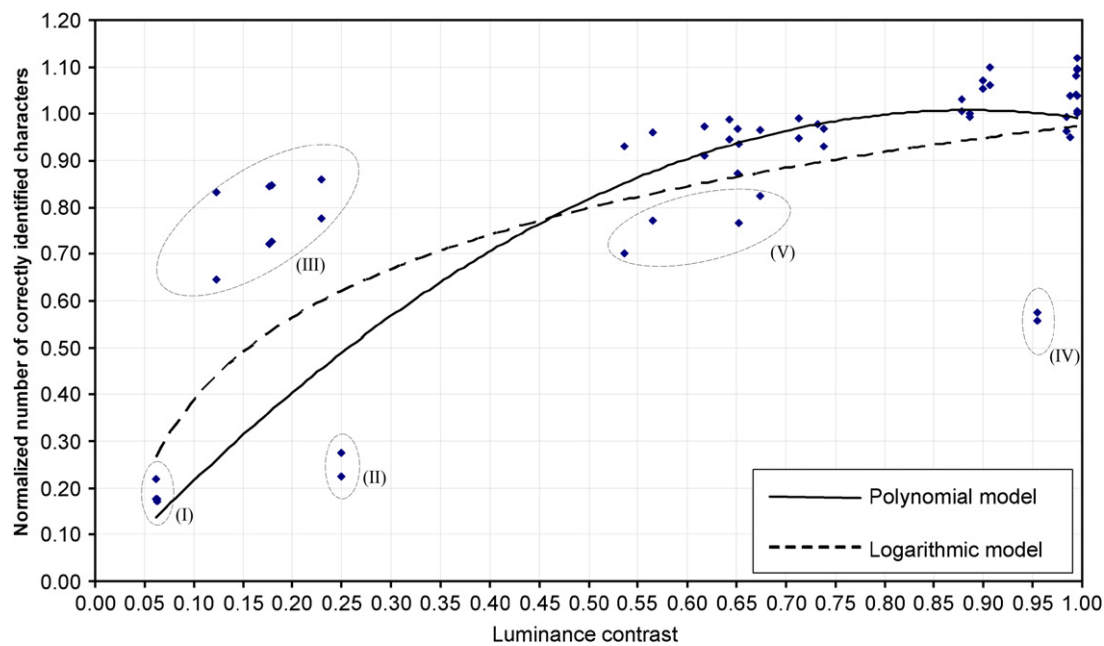


Fig. 2. Scatterplot of normalized number of correctly identified characters and luminance contrast with predicted values of polynomial and logarithmic model.

3.2. The impact of luminance contrast on legibility

The influence of luminance contrast on legibility was estimated by means of regression models. From the literature review, summarized in Section 1.3, we can draw a conclusion that larger luminance contrast leads to better legibility performance. However, the researchers rarely investigated the relation between the legibility performance and luminance contrast. In his work, Small (1982) found a strong positive correlation between the number of spelling mistakes detected per minute, and the logarithm of the contrast between foreground and background colors. Logarithm dependency was also assumed by Legge et al. (1990). These conclusions were drawn from the graphical data examination and were not supported by a model estimation of any kind. They are based on a relatively small number of color combinations taken in consideration and a small number of tested participants. Also, no other regression models were tested to verify whether they proved better than the logarithmic model.

Before the calculations were made, a data screening procedure was performed. A possible problem of heteroscedasticity was noticed, which could lead to biased results if estimates are calculated in the usual way by ordinary least-squares technique (OLS; Grimm and Yarnold, 2004). This technique relies on several assumptions, one being the equal variance of residuals for any value of independent variable (i.e. homoscedasticity), where residuals are defined as differences between values of dependent variable—luminance contrast and its values predicted by the regression model. For each regression model in this paper, the Breusch–Pagan–Godfrey test for heteroscedasticity (Godfrey, 1978; Breusch and Pagan, 1979) was conducted. If necessary, the calculations of regression models using weighted least-squares technique (WLS) were performed, which yielded unbiased estimators even when the assumption on homoscedasticity was not satisfied.

In this study, the firstly conducted regression model resembles the formulation given by Small (1982). However, it is not clear from Small's study whether the model should include an intercept term (β_0) or not. The absence of this term would introduce the assumption that luminance contrast is the only factor system-

atically influencing legibility, providing the model “through the origin”.

The calculation for the model with the intercept term was firstly conducted:

$$Y_i = \beta_0 + \beta_1 \ln C_i + u_i \quad (1)$$

The above model explained only 40.2% of variance of the dependent variable (adj. r^2 was 0.402). Besides, the calculations revealed that the estimated value for intercept (β_0) was not significantly different from 0, which supported the assumption that the luminance contrast is the only factor systematically influencing legibility. The log-normal model without the intercept term of the form:

$$\ln Y_i = \beta_1 \ln C_i + u_i \quad (2)$$

resulted in raw r^2 value of 0.496.² The estimated F value of 59.101 gave $p < 0.001$ with (1, 60) degrees of freedom, yielding an estimated value of β_1 to be 0.400 at $p < 0.001$.

A much better regression fit (raw r^2 closer to 1 and a higher F value) was achieved with the second-degree polynomial regression model of the form:

$$Y_i = \beta_1 C_i + \beta_2 C_i^2 + u_i \quad (3)$$

which obtained these estimates:

$$\hat{Y}_i = 2.277C_i - 1.287C_i^2 \quad (4)$$

The raw r^2 was 0.979. The corresponding ANOVA gave $F(2, 59) = 1399.876$ with $p < 0.001$. t -Tests for estimated regression coefficients in Eq. (4) showed that both values were significantly different from 0 (with $p < 0.001$).

² Raw coefficient of determination r^2 is a measure of goodness of fit similar to the coefficient of determination r^2 , but for regression models through the origin. It measures the proportion of variability in the Y 's about the origin explained by regression, and not of variability in the Y about its mean value as the measure of goodness does, so it is not directly comparable to r^2 . Similarly to measure of goodness r^2 , raw coefficient of determination r^2 is defined in the range $0 < \text{raw } r^2 < 1$.

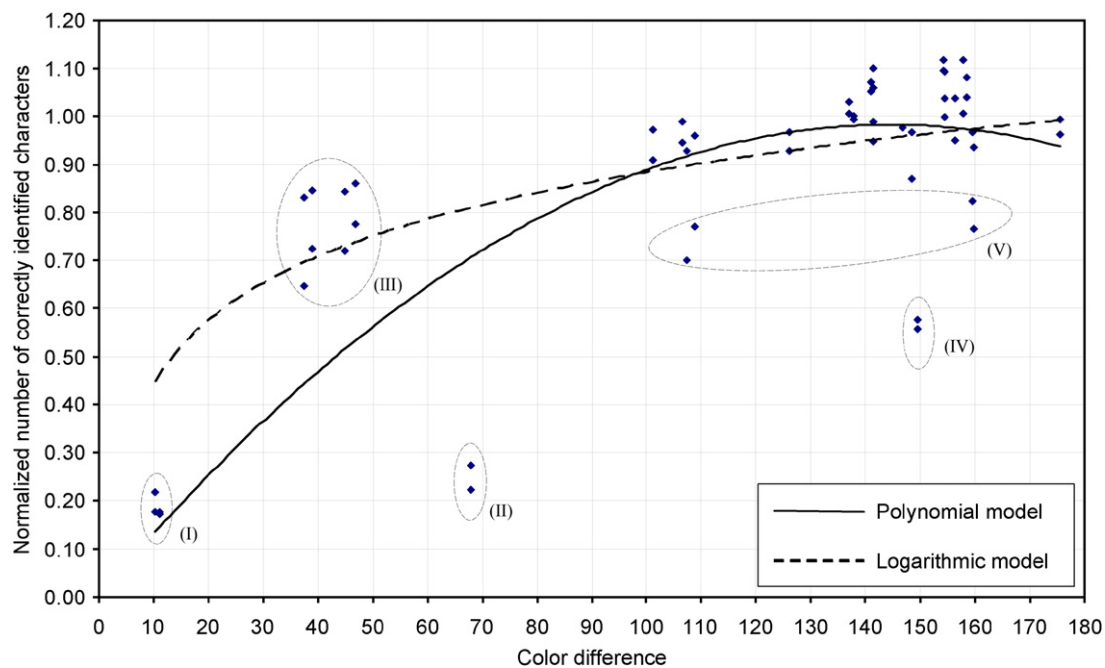


Fig. 3. Scatterplot of normalized number of correctly identified characters and color difference with predicted values of polynomial and logarithmic model.

Fig. 2 shows a scatterplot of normalized number of correctly identified characters and luminance contrast, together with predicted values of dependent variable across the luminance contrast for the polynomial model in a form of parabola and for the logarithmic model.

3.3. The impact of color difference on legibility

Using a similar approach as presented in the previous section, the influence of color difference on legibility was estimated. Again, the conclusions from literature survey in Section 1.3 indicate that larger color difference leads to better legibility results. Two regression models were under investigation, the logarithmic model and the second-degree polynomial model.

The logarithmic model of the form:

$$Y_i = \beta_1 \ln(\Delta E)_i + u_i \quad (5)$$

gave the best estimates for logarithmic models, with raw r^2 value of 0.974. The estimated F value of 2256.162 gave $p < 0.001$ with (1, 60) degrees of freedom, yielding estimated value of β_1 to be 0.192 at $p < 0.001$.

The second-degree polynomial model gave the following estimates for regression coefficients:

$$\hat{Y}_i = 0.014(\Delta E)_i - 0.000047(\Delta E)_i^2 \quad (6)$$

The raw coefficient of determination (r^2) was 0.976. The corresponding ANOVA gave $F(2, 59) = 1242.070$ with $p < 0.001$. t -Tests for estimated regression coefficients showed that both values were significantly different from 0 (with $p < 0.001$).

Fig. 3 shows a scatterplot of normalized number of correctly identified characters and color difference, together with predicted values of a dependent variable across the difference in color for the polynomial model and for the logarithmic model.

In case of color difference (as defined in Section 2.2), the results about the nature of its relationship to the legibility clearly show that this relationship is not linear. Both models (logarithmic and polynomial) gave good fits to measured values.

Table 6

Mean values of normalized numbers of correctly identified characters for positive and negative polarity for particular groups of color combinations

Group	N	Score		P
		Neg. polarity	Pos. polarity	
A	900	1.103	1.056	<0.001
B	900	1.025	0.995	<0.05
C	800	0.921	0.979	<0.001
D	800	0.749	0.870	<0.001
E	512	0.773	0.879	<0.001
F	512	0.315	0.365	<0.05
All	4424	0.861	0.896	<0.001

3.4. The impact of polarity on legibility

The effect of polarity was thoroughly explored by using the analysis of variance in comparing observed scores between specific groups of color combinations according to their color characteristics.

The results showed a statistically significant ($F(1, 4422) = 14.341$ ($p < 0.001$)) effect of polarity on normalized number of identified characters. Overall, a lighter text on a darker background resulted in a lower mean score (0.861) than a darker text on lighter backgrounds (0.896).

In a further analysis, the effects of polarity on legibility were studied within particular groups (A–F) of color combinations. The arrangement of color combinations to groups is described in Section 2.2 and evident from Table 3.

In Table 6, the mean scores for positive polarity were compared with those for negative polarity for a particular group. In the first two groups (A and B) with higher luminance contrast, the color combinations with negative polarity significantly outperformed the ones with positive polarity. In the other four groups of color combinations with lower luminous contrast (from C to F), the results of combinations with positive polarity were significantly better than those with negative polarity.

A further analysis was given to the combinations with non-chromatic³ color for either text or background. Firstly, only color combinations with either black or white color for the background and one of the remaining seven colors for the text were taken into consideration. The mean score for both non-chromatic backgrounds were compared. Statistically significant results ($F(1, 1618) = 30.658$ ($p < 0.001$)) showed that white color for a background resulted in a lower mean score (0.936) than black color for a background (1.010). Moreover, the color combinations with either black or white color for the text and one of the remaining seven colors for the background were analyzed. Statistically significant ($F(1, 1618) = 75.802$ ($p < 0.001$)) results showed that the mean score of combinations with black text color (0.979) was higher than the mean score of combinations with white text color (0.862).

4. Discussion

This study investigated the impact of a color combination on the legibility of Web page text presented on a CRT display, measured through the number of correctly identified characters. The dependences of legibility on luminance contrast and color differences, and the influence of polarity were investigated in more details.

The results of this study show that legibility was significantly affected by a color combination. This is in accordance with many of the latest studies (Shieh and Lin, 2000; Ling and van Schaik, 2002; Wu and Yuan (2003); Wang and Chen, 2003; Hall and Hanna, 2004), where authors compared a much smaller number (from two to mostly 12) of color combinations. However, to our knowledge, this is the first study that successfully proved the significant effect of color combinations on legibility for a large set of color combinations, with color chromaticities distributed over the entire spectrum. On the other side, the results of this research contradict the findings by Pace (1984), Shieh et al. (1997), and Lin (2003), who were not able to find a statistically significant influence of color combinations on visual performance. The non-significant results of those studies can be explained by at least two reasons: (a) the abovementioned studies employed a small number of participants (from 20 to 40) for testing a relatively large set of color combinations (from five to 25), while this study investigated the legibility by much larger population (400 and 68); (b) those studies measured visual performance by using methods that were not capable of isolating the influence of a color combination from other disturbing parameters.

A comparison of the results of this study with those of the past studies (Le Courier du Livre, 1912; Luckiesh, 1923) was done to explore the differences between subtractive (printing) and integrative colors. A different order of color combinations, shown in Table 7, suggests that more than 90-year-old Le Courier legibility table, which is still in use today, is not appropriate for integrative color computer displays. Consequently, we suggest that the list of color combinations, ordered by the mean scores, as given in Table 4, serves as a new legibility table, aiming at a Web page text on CRT displays.

As indicated by many previous studies (Small, 1982; Legge and Rubin, 1986; Matthews and Mertins, 1987; Pastoor, 1990; Legge et al., 1990), the luminance contrast affected visual performance. However, the small number of observed color combinations in the majority of those studies did not enable the authors to investigate the dependence of visual performance on luminance contrast. In spite of that, the logarithmic model was assumed by Small (1982)

Table 7

Comparison of legibility tables: this study and Le Courier

Text/background	Study	
	This	Le Courier
Yellow/black	1	7
White/blue	2	5
Black/yellow	3	1
White/black	4	10
Black/white	5	6
Blue/white	6	4
Red/yellow	7	11
White/red	8	8
Red/white	9	3
Red/green	10	13
Green/white	11	2
White/green	12	9
Green/red	13	12

and Legge et al. (1990), but was not investigated into more details. Similarly, none of the previous researches investigated the dependence of visual performance on color difference in detail. In this research, we investigated both dependences by using regression models. The assumptions of the abovementioned studies motivated the examination of the logarithmic regression model, although the results of the polynomial regression model, which was investigated for a comparison, are also presented. The results in Section 3.2 showed that the second-degree polynomial model proved to be more successful in describing the dependence of legibility on luminance contrast than the logarithmic model. The results of further investigation of color differences, presented in Section 3.3, showed that both models (logarithmic and polynomial) successfully described the dependence of legibility on color difference.

From a more detailed investigation of scatterplots in Figs. 2 and 3 a reader could notice that there were groups of color combinations, which noticeably deviated from the predicted models. A group (I) of four color combinations with a very low luminance contrast performed in very poor legibility. The combinations of this group were composed of the following pairs of colors: yellow and white, green and cyan. These colors provided the reader with very low legibility due to very small differences in both—luminance contrast and color difference. Similarly, the combination of red and magenta (II) also performed very low in scores, since these two colors have a small difference in color at relatively low luminance contrast. At this point it should be noted that there are color combinations (III) (yellow and cyan, white and cyan, yellow and green, white and green) that have lower luminance contrast than red on magenta (II), but performed much better in legibility due to high difference in color. This conclusion was congruent with the observation of Travis et al. (1990), who showed that purely chromatic differences may be used by visual system to extract information about form, and consequently, word or letter identification. Our study also showed that chromatic differences play an important role in legibility performance when luminance contrast is relatively low. It is the most difficult to comment the color combination of blue and black (IV). As expected, it performed relatively low in legibility, but using the provided models for luminance contrast and color difference, this combination had a high value of luminance contrast and color difference. The explanation can be found in the definitions of luminance contrast and color difference (provided in Section 2.2), which are very sensitive to luminance value of the color ($L_{\min}$) with lower luminance. It is very difficult to accurately define the luminance and chromaticity of the black color on the screen, since it depends greatly on the ambient

³ Non-chromatic color is white or black color.

conditions (illumination, glares, and reflections of different color objects), which cannot be completely eliminated in the experiment, and are also very common in real-life environments.

Somewhat more attention was also attracted by the color combinations that performed poorer in legibility than expected by the model. The following four color combinations (V) deviated most noticeably: cyan on magenta, green on magenta, cyan on red, and green on red. The poor performance of these color combinations can be explained by the effect of chromatic aberration causing additional visual strain to the users. This result is in accordance with some of the previous studies (Murch, 1985; Shieh et al., 1997; Ling and van Schaik, 2002), which also took red and green combinations in consideration. It is interesting that despite the same luminance contrast and color difference, the reverse combinations (i.e. magenta on cyan, red on green, etc.) performed with significantly better results. This observation supports the conclusion presented later in this section that positive polarity performs in better legibility for the colors with lower luminance contrast.

As already mentioned, some guidelines on the usage of colors, particularly on color differences, are provided in ANSI/HFS 100-1988. The standard recommends to use at least $100 \Delta E(Y, u', v')$ for color difference in order to achieve adequate legibility. The standard does not directly address luminance contrast. The guidelines are based on the referenced study and suggest using luminance contrast of at least 3:1 (0.667).⁴ The recommendation that luminance contrast should be greater than 7:1 (0.875) was also given by Snyder (1988), while Shieh et al. (1997) found that the green on red color combination with luminance contrast close to 0.667 performed worse in legibility than other color combinations. Our results confirmed these guidelines and extended them with the following conclusions: (a) there was a group of colors (Fig. 3, III) that performed relatively high in legibility (approximately 70% of maximal visual performance) although the combinations had relatively low values of color difference, around $40 \Delta E(Y, u', v')$: yellow and cyan, white and cyan, yellow and green, white and green. (b) these colors performed well although their luminance contrast was very low, (i.e. below 0.25). (c) there were some color combinations, which lead to worse results, although color difference exceeds $100 \Delta E(Y, u', v')$. These combinations were composed of a green or cyan text on a magenta or red background (V), blue on black and black on blue combinations (IV). The possible reasons causing these low performances were already described.

The scatterplots in Figs. 2 and 3 also imply that the experiments in future research on the influence of luminance contrast and color difference on legibility should cover not only the most interesting foreground-background color combinations (which leads to "groups" in the scatterplot) but also the combinations between these groups for instance in the range from 0.25 to 0.50 of luminance contrast).

The effect of the polarity was studied as presented in Section 3.4. The result showed that, in general, the negative polarity is preferred for better legibility. Although this finding is somewhat surprising, it is congruent with the results of Hall and Hanna's (2004) study, who analyzed only four color combinations. A large set of color combinations enabled us to perform some further investigation. Six groups of colors (A–F) were defined by splitting the colors, sorted by luminance contrast, equally among the groups, as described in Section 2.2 and evident from Table 4. When comparing the influence of polarity within groups, negative polarity was significantly better for the first two groups (A and B) with greatest contrast ratio, while for the remaining four groups

(C–F), positive polarity led to greater legibility. The preference of negative polarity can be explained by a flickering nature of a picture, which is more evident in case of positive polarity. On the other hand, it should be pointed out that positive polarity reduces the effects of distracting reflections from the surface of a display. Such reflections definitely exist in real environments, although they were eliminated when arranging the viewing conditions for the experiment of this study. It was also shown that black color as background outperformed white as background, when combined with other colors for the text. Similarly, the color combinations with a black text outperformed the ones with a white text, when combined with other colors as background. The explanation of these observations might be based on the fact that chromatic colors, chosen as stimuli in this research, had a relatively high luminance component, since such colors are relatively frequently used on the Web. Some further investigation should also take the colors with lower luminance into consideration, which will also reduce deficiencies described in the previous paragraph.

5. Conclusions and future work

In conclusion, this study shows that legibility of a Web page text presented on a CRT display is significantly affected by color combinations. The Le Courier legibility table for subtractive colors is not appropriate for integrative color computer displays. Consequently, a new legibility table is proposed. A second-degree polynomial model proved as appropriate describing the legibility in dependence of luminance contrast. A second-degree polynomial model and logarithmic model proved as appropriate describing the legibility in dependence of color difference. Color difference played a more important role for legibility, when luminance contrast was low. Negative polarity is significantly better for the color combinations with greatest contrast ratio while for other combinations the positive polarity performs better.

Future work will track the following directions: (a) comparing the scatterplots in Figs. 2 and 3, we notice some similarities based on the definition of color difference, which relies on luminance contrast. In the future, it might be reasonable to develop a new model for color difference, based only on differences in hue and saturation of the colors in combination. (b) Further studies should also observe other types of display, i.e. a thin-film transistor liquid crystal display (TFT LCD) that is becoming a more and more popular replacement for a CRT computer's display nowadays or a transfective LCD, which is a popular type of display in modern handheld devices. Both technologies have significantly different optical characteristics from a CRT display and may thus require reinvestigation of the influence of colors on legibility by using a large set of color combinations tested with a large number of participants. (c) In future research, subjective preferences for legibility, esthetics and retention should also be rated. It would be interesting to connect and compare these subjective results with the results objectively measured for legibility as visual performance task. (d) Our study tested the legibility of 56 color combinations, compounded of elementary Web safe colors, with chromaticities distributed over the entire spectrum. Future work should focus on color combinations that are the most frequently used in modern, fashion-driven designs, as well as on colors with different chromaticities or achromatic levels and saturations, such as brown, gray or light red.

Acknowledgments

The authors are grateful to professor Milton Jenkins from the University of Baltimore for valuable comments and suggestions.

⁴ The reader is referred to Section 2.2 to learn about the transformation of luminance contrast ratio to luminance contrast, used in this study.

This research was supported by the Ministry of Higher Education, Science and Technology of the Republic of Slovenia under Grant no. P2-0037 and Grant no. pp-0586-501.

Appendix A. The color coordinates and color differences

The coordinates of CIE L^* , u^* , v^* color space (CIE, 1986) are defined as

$$u^* = 13L^*(u' - u'_n) \quad \text{and} \quad v^* = 13L^*(v' - v'_n)$$

where L^* is metric lightness, u' and v' are UCS coordinates of a color, and u'_n and v'_n are the coordinates of reference white.

The coordinates of CIE L^* , a^* , b^* color space (CIE, 1986) are defined as

$$a^* = 500((X/X_n)^{1/3} - (Y/Y_n)^{1/3}), \quad \text{and} \\ b^* = 200L^*((Y/Y_n)^{1/3} - (Z/Z_n)^{1/3}), \\ (X/X_n), (Y/Y_n), (Z/Z_n) > 0.01,$$

where X , Y , and Z are the tristimulus values of a color, and X_n , Y_n , and Z_n are those of the reference white.

The color difference ΔE is defined with the Pythagorean distance formula for CIE L^* , u^* , v^* and CIE L^* , a^* , b^* color spaces:

$$\Delta E(L^*, u^*, v^*) = \sqrt{(\Delta L^*)^2 + (\Delta u^*)^2 + (\Delta v^*)^2} \\ \Delta E(L^*, a^*, b^*) = \sqrt{(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2}$$

References

- ANSI/HFS 100-1988, 1988. American National Standard for Human Factors Engineering of Visual Display Terminal Workstations. The Human Factors Society, Inc.
- Benbasat, I., Dexter, A.S., Todd, P., 1986. The influence of color and graphical information presentation in a managerial decision simulation. *Human-Computer Interaction* 2 (1), 65–92.
- Breusch, T., Pagan, A., 1979. A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47, 1287–1294.
- Bruce, M., Foster, J.J., 1982. The visibility of colored characters on colored backgrounds in viewdata displays. *Visible Language* 16 (4), 382–390.
- Christ, R.E., 1975. Review and analysis of color coding research for visual displays. *Human Factors* 17, 542–570.
- CIE, 1931. *Proceedings, International Congress on Illumination*. Cambridge University Press, Cambridge.
- CIE Pub. 116, 1995. *Industrial Colour-Difference Evaluation*. CIE Central Bureau, Vienna.
- CIE Pub. 15.2, 1986. *Colorimetry*, second ed. CIE Central Bureau, Vienna.
- Clarke, F.J.J., McDonald, R., Rigg, B., 1984. Modification on the JPC79 colour-difference formula. *Journal of the Society of Dyers and Colourists* 104, 128–132.
- Dillon, A., 1992. Reading from paper versus screen: a critical review of empirical literature. *Ergonomics* 32, 1297–1326.
- Fukuzumi, S., Yamazaki, T., Kamijo, K., Hayashi, Y., 1998. Physiological and psychological evaluation for visual display colour readability: a visual evoked potential study and a subjective evaluation study. *Ergonomics* 41 (1), 89–108.
- Ghani, J., Lusk, E.J., 1981. The Impact of Information Presentation and Modification on Decision Performance. WP, The Sloan School, MIT.
- Godfrey, L., 1978. Testing for multiplicative heteroscedasticity. *Journal of Econometrics* 8, 227–236.
- Gould, A.J.D., Alfaro, L., Barnes, V., Finn, R., Grischkowsky, N., Minuto, A., 1987. Reading is slower from CRT displays than from paper: attempts to isolate a single-variable explanation. *Human Factors* 29 (3), 269–299.
- Gradisar, M., Humar, I., Turk, T., 2005. The Impact of Colors on the Speed of Reading Web Pages on a CRT monitor Working Paper No. 174. University of Ljubljana, Faculty of Economics.
- Gremillion, L.L., Jenkins, A.M., 1981. The Effects of Color Enhanced Information Presentations Discussion Paper #173. Indiana University, Bloomington, Indiana.
- Grimm, L.G., Yarnold, P.R., 2004. *Reading and Understanding More Multivariate Statistics*. American Psychological Association, Washington.
- Hall, R., Hanna, P., 2004. The impact of Web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & Information Technology* 23 (3), 183–195.
- Hill, A.L., Scharff, L.V., 1997. Readability of screen displays with various foreground/background color combinations, font styles, and font types. In: *Proceedings of the 11th National Conference on Undergraduate Research*, Austin, Texas, pp. 742–746.
- Hill, A.L., Scharff, L.V., 1999. Legibility of computer displays as a function of colour, saturation, and texture backgrounds. In: Harris, D. (Ed.), *Engineering Psychology and Cognitive Ergonomics*. Ashgate, Sydney, pp. 123–130.
- Hoadley, E., Jenkins, A.M., 1987. The effects of color on performance in an information extraction task using varying forms of information presentation: pilot studies IRMIS Working Paper #W713. Indiana University Institute for Research on the Management of Information Systems, Graduate School of Business, Bloomington.
- Humar, I., Gradisar, M., 2003. Color test Available <<http://spin.fe.uni-lj.si/colors/>>.
- ISO 12646, 2004. *Graphic Technology—Displays for Colour Proofing—Characteristics and Viewing Conditions*. International Organization for Standardization.
- ISO 9241-3, 1992. *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), Part 3: Visual Display Requirements*. International Organization for Standardization.
- Kroemer, K.H.E., 1993. Locating the computer screen: how high, how far? *Economics in Design*, October 1993, 7–8.
- Le Courrier du Livre, 1912. *Lisibilité des Affiches en Couleurs*. Sheldons Limited House, Cosmos, p. 255.
- Legge, G.E., Rubin, G.S., 1986. Psychophysics of reading: IV. Wavelength effects in normal and low vision. *Journal of the Optics Society of America, A* (3), 40–51.
- Legge, G.E., Rubin, G.S., Luebker, A., 1987. Psychophysics of reading: V. The role of contrast in normal vision. *Vision Research* 27 (7), 1165–1177.
- Legge, G.E., Parish, D.H., Luebker, A., Wurm, L.H., 1990. Psychophysics of reading: XI. Comparing color contrast and luminance contrast. *Journal of Optical Society of America A* 7 (10), 2002–2010.
- Lehn, D., Stern, H., 2000. Death of the Web Safe Color Palette? Available <<http://webmonkey.wired.com/webmonkey/00/37/index2a.html?tw=desin>>.
- Lin, C.C., 2003. Effects of contrast ratio and text color on visual performance with TFT-LCD. *International Journal of Industrial Ergonomics* 31, 65–72.
- Ling, J., van Schaik, P., 2002. The effect of text and background colour on visual search of Web pages. *Displays* 23, 223–230.
- Lippert, M.T., 1986. Color Difference Prediction of Legibility Performance for CRT Raster Imagery. In: *Proceedings of the Society for Information display International Symposium Digest of Technical Papers XVI*. Palisades Institute for Research Services, New York, pp. 86–89.
- Luckiesh, M., 1923. *Light and Color in Advertising and Merchandising*. H. Van Nostrand Co., New York, pp. 246–251.
- Luo, M.R., Rigg, B., 1987. BFD (lc) colour difference formula. Part 1—Development of the formula. *Journal of the Society of Dyers and Colourists* 103, 86–96.
- Luo, M.R., Cui, G., Rigg, B., 2001. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research and Application* 26 (5), 340–350.
- Mason, R.O., Mitroff, I.I., 1973. A program for research on management information systems. *Management Science* 19 (5), 475–487.
- Matthews, M.L., Mertins, K., 1987. The influence of color on visual search and subjective discomfort using CRT displays. In: *Proceedings of the Human Factors Society 31st Annual Meeting*. HFES, Santa Monica, CA, pp. 1271–1275.
- Monitor Optimizer Software Guide, X-Rite Colorimeter, Available: <<http://www.xrite.com/Products/Product.asp?Show=Description&id=11>>
- Murch, G.M., 1985. Color graphics—blessing or ballyhoo? *Computer Graphics Forum* 4, 127–135.
- Pace, B.J., 1984. Color combinations and contrast reversals on visual display units. In: *Proceedings of the Human Factors Society 28th Annual Meeting*. HFES, Santa Monica, CA, pp. 326–330.
- Pastoor, S., 1990. Legibility and subjective preference for color combinations in text. *Human Factors* 32 (2), 157–171.
- Pearson, R., van Schaik, P., 2003. The effect of spatial layout of and link colour in Web pages on performance in a visual search task and an interactive search task. *International Journal on Human-Computer Studies* 59, 327–353.
- Pett, D., Wilson, T., 1996. Color research and its application to the design of instructional materials. *Educational Technology Research & Development* 44 (3), 19–35.
- Poulton, E.C., 1972. Size, style, and vertical spacing in the legibility of small typefaces. *Journal of Applied Psychology* 56 (2), 156–161.
- Radl, G.W., 1980. Experimental investigations for optimal presentation-mode and colours of symbols on the CRT-screen. In: Grandjean, E., Vigliani, E. (Eds.), *Ergonomic Aspects of Visual Display Terminals*. Taylor & Francis, London, UK, pp. 127–135.
- Resinovic, G., Jenkins, M.A., Gradisar, M., Jaklic, J., 1999. Legibility and visibility issues in color enhanced information presentation Working Paper No. 88. University of Ljubljana, Faculty of Economics.
- Sanders, M.S., McCormick, E.J., 1993. *Human Factors in Engineering and Design*. McGraw-Hill, Singapore.
- Shieh, K.K., Chen, M., 1997. Effects of screen color combination, work-break schedule, and workplace on VDT viewing distance. *International Journal of Industrial Ergonomics* 20 (1), 11–18.
- Shieh, K.K., Lin, C.C., 2000. Effects of screen type, ambient illumination, and color combination on VDT visual performance and subjective preference. *International Journal of Industrial Ergonomics* 26 (5), 527–536.
- Shieh, K.K., Chen, M., Chuang, J.H., 1997. Effects of color combinations and typography on identification of characters briefly presented on VDT's. *International Journal of Human-Computer Interaction* 9 (2), 169–181.
- Silverstein, L.D., 1982. Human factors for color CRT displays. In: *Proceedings of the Society for Information Display: Seminar Lecture Notes* 2, pp. 1–41.

- Small, P.L., 1982. Factors influencing the legibility of text/background color combinations on the IBM 3279 color display station. Report No. HF066, IBM Hursley Park, Winchester, England.
- Snellen, H., 1862. Proebuchstaben zur Bestimmung der Sehschärfe. P.W. van de Weijer, Utrecht.
- Snyder, H.L., 1988. Image quality. In: Helander, M. (Ed.), *Handbook of Human-Computer-Interaction*. Elsevier, North-Holland, Amsterdam, pp. 437–474.
- Suh, K.S., 1999. Impact of communication medium on task performance and satisfaction: an examination of media-richness theory. *Information & Management* 35 (5), 295–312.
- Teichner, W.R., 1979. Color and visual information coding. In: *Proceedings of the Society for Information Display* 20(1), pp. 3–9.
- Tinker, A.M., 1955. *Tinker Speed of Reading Test*. University of Minnesota Press, Minneapolis.
- Tinker, A.M., 1963. *Legibility of Print*. Iowa State University Press, Ames, IA.
- Tinker, A.M., Paterson, G.D., 1928. Influence of type form on speed of reading. *Journal of Applied Psychology* 12, 359–368.
- Tinker, A.M., Paterson, G.D., 1929a. Studies of typographical factors influencing speed of reading: II. Size of type. *Journal of Applied Psychology* 13 (2), 120–130.
- Tinker, A.M., Paterson, G.D., 1929b. Studies of typographical factors influencing speed of reading. III. Length of line. *The Journal of Applied Psychology* 13 (3), 205–219.
- Tinker, A.M., Paterson, G.D., 1929c. Studies of typographical factors influencing speed of reading: V. Simultaneous variations of type size and line length. *The Journal of Applied Psychology* 13 (2), 72–78.
- Tinker, A.M., Paterson, G.D., 1929d. Studies of typographical factors influencing speed of reading: VI. Black type versus white type. *The Journal of Applied Psychology* 13 (2), 241–247.
- Tinker, A.M., Paterson, G.D., 1929e. Studies of typographical factors influencing speed of reading. VII. Variations in color of print and background. *Journal of Applied Psychology* 13 (2), 471–479.
- Tinker, A.M., Paterson, G.D., 1932. Studies of typographical factors influencing speed of reading: X. Style of typeface. *The Journal of Applied Psychology* 16, 605–613.
- Tinker, A.M., Paterson, G.D., 1940. *How to Make Type Readable: A Manual for Typographers, Printers, and Advertisers, Based on Twelve Years of Research Involving Speed of Reading Tests Given to 33,031 Persons*. Harper & Brothers, New York.
- Tinker, A.M., Paterson, G.D., 1942. Reader preferences and typography. *Journal of Applied Psychology* 26, 38–40.
- Tinker, A.M., Paterson, G.D., 1944. Eye movements in reading black print on white background and red print on dark green background. *American Journal of Psychology* 57, 93–94.
- Tinker, A.M., Paterson, G.D., 1946. Readability of mixed type forms. *Journal of Applied Psychology* 30, 631–637.
- Travis, D.S., Bowles, S., Seton, J., Peppe, R., 1990. Reading from color displays: a psychophysical model. *Human Factors* 32 (2), 147–156.
- Tullis, T.S., 1981. An evaluation of alphanumeric, graphic and color information displays. *Human Factors* 23 (5), 541–550.
- Wang, A.H., Chen, C.H., 2003. Effects of screen type, Chinese typography, text/background color combination, speed, and jump length for VDT leading display on users' reading performance. *International Journal of Industrial Ergonomics* 31 (4), 249–261.
- Wu, J.H., Yuan, Y.F., 2003. Improving searching and reading performance: the effect of highlighting and text color coding. *Information & Management* 40 (7), 617–637.

Broadband Access Network Planning Optimization Considering Real Copper Cable Lengths

Blaž Peternel,^{†*a)} *Nonmember*, and Andrej Kos[†], *Member*

SUMMARY Broadband access network planning strategies with techno-economic calculations are important topics, when optimal broadband network deployments are considered. This paper analyzes optimal deployment combination of digital subscriber line technologies (xDSL) and fiber to the home technologies (FTTx), following different user bandwidth demand scenarios. For this reason, optimal placement of remote digital subscriber line multiplexer (RDSLAM) is examined. Furthermore, the article also discusses the economy of investments, depending on certain investment threshold and the reach of different xDSL technologies. Finally, the difference between broadband network deployment in a characteristic urban and rural area in Republic of Slovenia, in terms of required optical cable dig length per household is shown. A tree structure network model of a traditional copper access network is introduced. A dynamic programming logic, with recursion as a basis of a tree structure examination and evaluation of optimal network elements placement is used. The tree structure network model considers several real network parameters (e.g.: copper cable lengths, user coordinates, node coordinates). The main input for the optimization is a local loop distance between each user and a candidate node for RDSLAM placement. Modelling of copper access networks with a tree structure makes new extensions in planning optimization of broadband access networks. Optimization of network elements placement has direct influence on efficiency and profitability of broadband access telecommunication networks.

Key words: broadband, copper network, remote DSLAM, xDSL, FTTx, optimization, tree structure, dynamic programming

1. Introduction

Next generation broadband services evolution (e.g.: IP TV, high definition TV, video on demand, video streaming, IP telephony) and consequently much higher user bandwidth demands, are the main drivers for fixed operators to deploy high speed broadband access networks. The traditional asymmetric digital subscriber line technology (ADSL) with less than 8 Mbit/s download channel capacity, has become inappropriate for the next generation users (Okumura, 2007 [1], Carpenter, 2001 [2]). Development of very high bit rate digital subscriber line technologies (VDSL) is one of the solutions how to deliver a high-bandwidth capacity to end users. The main drawback of the VDSL and VDSL2

technologies is their short range limitation for high bandwidth delivery, which is less than 2 km over copper lines, as shown in Fig. 4 [3]. The other solution, which is currently deployed by a majority of European fixed operators [4], is fiber to the home access network deployment. The main disadvantage of various fiber access network solutions is the return of investment (ROI).

Deployment of access network with certain combinations of xDSL and FTTx technologies may be optimal, considering short range of VDSL and VDSL2 technologies, and vast investment needed for point to point fiber to the home solution (PtP FTTH). In the paper, the FTTH abbreviation is used to indicate point to point FTTH technical solution (e.g. active optical network (AON)). If the VDSL technologies are used for delivering high-bandwidth capacity through the last kilometer of existing copper lines, and optical fiber lines are required only for the RDSLAMs, the investment in the optical fiber dig length can be reduced and the performance of VDSL technologies can meet the needs of end users, as shown in Fig 1.

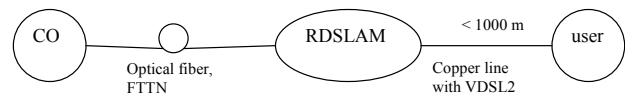


Fig. 1 Local loop shortening using FTTN and VDSL2 technology combination.

For optimal access network planning with combination of xDSL and PtP FTTH technologies and reuse of existing copper network, the existing copper network and end users in the target area have to be simulated. Different methods can be used to find optimal RDSLAM placement in the access network, generally the knowledge of dynamic programming is applied. Several projects deal with simulation of existing copper network, but do not propose optimal RDSLAM positions [5], [6], [7]. Their focus is on techno-economical comparison of various broadband access technologies deployment combinations, based on statistical user density presumptions. Different methods for cable and duct lengths modelling are proposed (e. g.: star topology, double star topology) [5], [6]. In this paper, the real copper cable lengths are used in computational comparisons. Node placement and sizing for copper broadband access networks with two real examples is analyzed in detail (Carpenter, [2]). The advantage of the proposed method is real input data

Manuscript received November 2007.

Manuscript revised March 1, 2007.

[†]The authors are with the University of Ljubljana, Faculty of electrical Engineering, Tržaška c. 25, SI-1000 Ljubljana, Slovenia,

^{*}The author is with Mobitel, Telecommunications services, PLC, Vilharjeva c. 23, SI-1000 Ljubljana, Slovenia.

a) E-mail: blaz.peternel@ltfe.org

with existing copper network considered. The cost-benefit analysis is carried out only for certain RDSLAM infrastructure without considering optical dig lengths from the central office to the RDSLAM position. Moreover, the method is inappropriate for analyzing FTTH solution.

In this paper, a different method is used for network simulation – graph theory application [8], which is up to certain extend appropriate also for FTTH solution analyses. The economy of investment is calculated in terms of required optical fiber dig length to the RDSLAM, while cables, ducts and the civil work represent the main part of the capital expenditure (CAPEX). Estimation of planned optical fiber dig length per user is the key factor for comparison of investments in urban and rural areas, as shown in next chapters of this paper.

2. Copper Access Network Tree Structure

A traditional copper cable access network (CAN) typically has a tree structure architecture, where the root of the tree represents a central office (CO), vertexes represent joints (J) in a multi-pair copper cable network and leaves represent distribution points (DP). A multi-pair copper cable access network is called local cable network (LCN). Copper cables between distribution points and users (U) make a distribution network (DN). To summarize, the cable access network (CAN) consists of the following of two parts, as shown in Fig. 2:

- the local cable network (LCN) and
- the distribution network (DN).

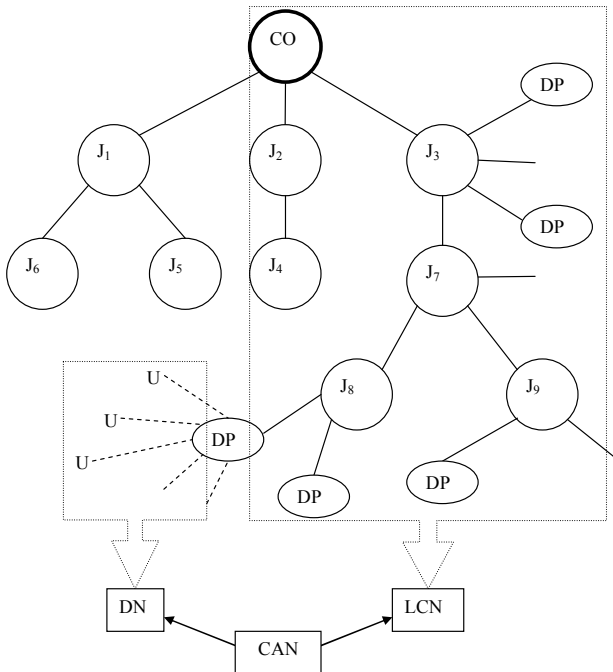


Fig.2 Topology of copper access network.

3. Optimization Methodology

Optimization tree access network model is implemented with Wolfram Research Mathematica version 6.0.0. For tree structure characteristic operations (e.g.: Breadth-First Search and SetGraphOptions) the Combinatorica extension to Mathematica is used [8], [9]. Calculations have been running on the Intel Pentium 4 personal computer with 3 GHz CPU and 1 Gbit RAM.

A tree is a connected graph with no cycles. A graph is a set of vertexes with a set of edges, where each edge is defined by a pair of vertexes [8]. Basic graph characteristics are edge weights, vertex weights and edge directions. A cable access network (CAN) is an undirected graph, a tree with the edge and vertex weights.

Edge weights are distances between joints in a local cable network (LCN). Vertex weights have to be dynamically calculated during the optimization process, and are represented with a vector notation as follows:

$$J_i = \{x, y, i, o\}, \quad (1)$$

where J_i represents the joint with a label i , x and y are coordinates of the joint i , and o is an optional set of parameters, which are attributes of the joint i . The optional parameters are calculated dynamically, during tree elements traversal, and are key factors for optimization (e.g.: maximal cable distance from the joint i to belonging users, number of users belonging to the joint i and statistical estimation of their generated traffic).

Optimization algorithm runs in two steps. In the first step a Breadth-First Search (BFS) algorithm is used for calculation of all optional parameters [8]. The BFS algorithm also gives us a list of joint labels, i , in the order they have been visited. This order is of a great use in the second step of optimization.

The second part of optimization is based on dynamic programming solution with bottom-up approach [8]. Comparison of joint vectors, recalculation of optional parameters and decision-making if a RDSLAM will be placed at a certain joint in a characteristic fragment of a tree (see Fig. 3), is recursively repeated until the root (CO) of the tree is reached. If the order, in which joints were visited in the first part of optimization is used, it is assured that all children vertexes are visited before their parent vertex. This is the crucial point of the optimization – the information between all children vertexes and a parent vertex has to be transferred before recursion step is made and the parent vertex becomes a child of another parent in lower level of a tree.

Decision-making if a RDSLAM will be placed at certain joint position in the tree, depends on thresholds which are chosen by application user. Thresholds, which can be chosen, are the reach of a xDSL technology and the investment threshold. The reach of the xDSL technology is expressed in meters of cable length (R). The investment

threshold is expressed by the number of potential users, which are required for acceptable economy of investment (N). A RDSLAM is placed at a certain joint if both thresholds are exceeded.

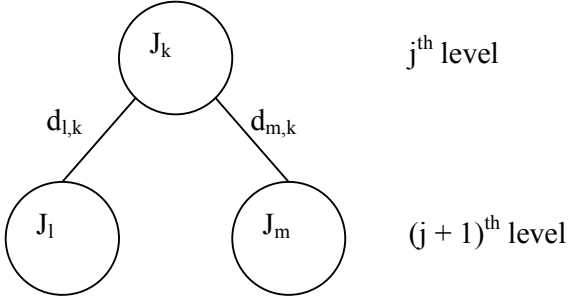


Fig. 3 Characteristic fragment of a tree.

Assuming that a RDSLAM is placed at the position of the joint m (see Fig. 3), the RDSLAM »covers« n_m belonging users with the maximal belonging user distance r_m . Also threshold conditions are assumed:

$$r_m \geq R, \quad (2)$$

$$n_m \geq N. \quad (3)$$

In general, optional parameters of the parent joint k are recalculated, following the equations below:

$$J_{k_{rec}} = \{x, y, k, r_{k_{rec}}, n_{k_{rec}}\}, \quad (4)$$

$$o_{k_{rec}} = (r_{k_{rec}}, n_{k_{rec}}), \quad (5)$$

$$r_{k_{rec}} = \max(r_l + d_{l,k}, r_m + d_{m,k}, r_k), \quad (6)$$

$$n_{k_{rec}} = n_l + n_m + n_k, \quad (7)$$

where rec denotes recalculated parameter.

If a RDSLAM is placed at the position of the joint m (see Fig. 3), optional parameters »maximal belonging user distance« (r_m) and »number of belonging users« (n_m) of the joint m are put to zero. Afterwards, the recalculation is made, following equations:

$$\begin{aligned} r_m &= 0, \\ n_m &= 0, \end{aligned} \quad (8)$$

$$\begin{aligned} d_{m,k} &= 0, \\ r_{k_{rec}} &= \max(r_l + d_{l,k}, r_k), \end{aligned} \quad (9)$$

$$n_{k_{rec}} = n_l + n_k. \quad (10)$$

After the recalculation is made, the joint k has optional parameters $r_{k_{rec}}$ and $n_{k_{rec}}$ and in the next recursion step, the

joint k becomes a child of another parent. The recursive calculations are repeated until the root (CO) of the tree is reached.

The basic result of optimization is a list of RDSLAMs with belonging co-ordinates, the »covered« user number and the maximal user distance for every RDSLAM. As additional result of optimization, estimations, like user traffic load at the RDSLAMs positions and dig length from the central office to the RDSLAM positions, are made. These additional estimations are key factors for techno-economic calculations of different FTTx deployment scenarios.

3.1 Input Parameters

General input parameters for the optimization model are as follows:

- the co-ordinate of the central office (CO),
- co-ordinates of joints (J),
- co-ordinates of distribution points (DP),
- co-ordinates of users (U),
- cable lengths between neighbouring vertexes in the tree (CO, J, DP),
- the investment threshold (N) and
- the reach of xDSL technologies (R).

Coordinates and cable lengths for characteristic copper cable access networks have been provided by Slovenian national operator, Telecom Slovenia [10]. Calculations are made for several different investment thresholds and different reaches of xDSL technologies. Available broadband bandwidth as planned at certain household can be directly mapped from the diagram (see Fig. 4). If the length of the household copper local loop is known, the available broadband bandwidth can be estimated for different xDSL technologies, as shown in Fig. 4.

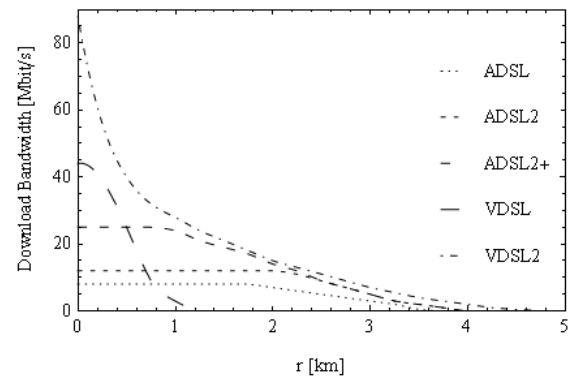


Fig. 4 Reach of xDSL technologies.

Additional input parameters are as follows:

- statistical presumptions of the offered user traffic and
- broadband penetration estimation.

Post and Electronic Communications Agency of the Republic of Slovenia reports [11], Statistical Office of the Republic of Slovenia [12] and Annual Report of Telecom Slovenia Group [10] are used for statistical presumptions of offered user traffic and broadband penetration estimation. Table 1 shows the additional input parameters, which are used in optimization process. It is defined that a high speed internet service (HSI) is “always-on” broadband connection for Internet browsing, with the average 1.012 kbit/s downlink speed and 256 kbit/s uplink speed. IP telephony and IP TV are considered as services, running over xDSL technology.

Table 1 Additional input parameters.

Input	Percentage of households
Total broadband penetration	45,0 %
xDSL technology penetration	32,3 %
HIS service	32,3 %
IP telephony service	7,7 %
IP TV service	5,4 %

Statistical presumption of the offered traffic for characteristic household is made, as shown in Table 2. It is considered that an average household uses one IP TV service with an average bit rate 5 Mbit/s, one IP telephony service (codec G.711) with an average bit rate 115 kbit/s. Oversubscription factors for IP telephony service and best effort (BE) HSI service are 0,18 Erl and 1/10, respectively. Internet group management protocol (IGMP) request message is used for user channel switching request and is a negligible quantity in offered traffic estimations.

Table 2 Generated traffic per characteristic household.

Service	Downlink traffic	Uplink traffic	Oversubscription	Total traffic
HSI	1.012 kbit/s	256 kbit/s	1/10	126,8 kbit/s
IP telephony	115 kbit/s	115 kbit/s	0,18 Erl	41,4 kbit/s
IP TV	5.000 kbit/s	IGMP request	-	5.000 kbit/s
Sum	6.127 kbit/s	371 kbit/s	-	5.168,2 kbit/s

Table 3 Presumed maximal generated traffic per household, Scenario 1.

Service	Downlink traffic	Uplink traffic	Oversubscription	Total traffic
HSI	5.060 kbit/s	2.024 kbit/s	1/10	708,4 kbit/s
2 x IP telephony	230 kbit/s	230 kbit/s	0,18 Erl	82,8 kbit/s
2 x IP TV	10.000 kbit/s	IGMP request	-	10.000 kbit/s
1 x HDTV	15.000 kbit/s	IGMP request	-	15.000 kbit/s
Sum	30.920 kbit/s	2.254 kbit/s	-	25.791,2 kbit/s

4. Optimization Results

Three different user demand scenarios are considered for the optimization calculations. In the first scenario, a “near future proof” 30 Mbit/s offered downlink speed for the majority of households is presumed, as shown in Table 3. It is presumed that an average household uses two IP TV services with an average bit rate 5 Mbit/s, one IP high definition TV (HDTV) service with an average bit rate 15 Mbit/s and two IP telephony services with an average bit rate 115 kbit/s. An average HSI service user profile has 5.060 kbit/s downlink speed and 2.024 kbit/s uplink speed. Oversubscription factors are considered, as shown in Table 2.

The driving factor for required local loop length estimation is 30.920 kbit/s peak downlink traffic. Considering the Fig. 4, the corresponding xDSL technologies are VDSL and VDSL2 with the “reach of xDSL technologies” input parameter (R), 400 m and 900 m, respectively.

In Scenario 2, a different HSI service user profile is presumed, as shown in Table 4. In this case, the driving factor for required local loop length estimation is 65.710 kbit/s peak downlink traffic. Considering the Fig. 4, the corresponding xDSL technology is VDSL2 with the “reach of xDSL technologies” input parameter (R), 200 m.

Scenario 3 considers FTTH solution, which is a future proof from the offered traffic point of view. In this case the “reach of xDSL technologies” input parameter (R) in the optimization algorithm is set to zero and “investment threshold” is set to one – the fiber has to be reachable from every household.

Scenario 2 has two sub-cases for different input parameters, R . To summarize, there are four scenarios, three with combined VDSL/VDSL2 and FTTN deployment and one with point-to-point FTTH solution.

Table 4 Presumed maximal generated traffic per household, Scenario 2.

Service	Downlink traffic	Uplink traffic	Oversubscription	Total traffic
HSI	40.480 kbit/s	20.240 kbit/s	1/10	6.072,4 kbit/s
2 x IP telephony	230 kbit/s	230 kbit/s	0,18 Erl	82,8 kbit/s
2 x IP TV	10.000 kbit/s	IGMP request	-	10.000 kbit/s
1 x HDTV	15.000 kbit/s	IGMP request	-	15.000 kbit/s
Sum	65.710 kbit/s	20.470 kbit/s	-	31.155,2 kbit/s

3.1 Urban Copper Cable Access Network

The characteristic urban central office (CO) area consists of up to six major multi-pair copper cables, each with more than one thousand connected users. Each of these multi-pair cables represents a CAN, as shown in Fig. 2. The optimization planning for the residential urban CAN with 1.125 connected households (users) has been made. The belonging local cable network (LCN) is shown in Fig. 5. The circle with label CO represents the central office and black dots represent labelled joints in the LCN. A household density in the corresponding urban municipality is 298 households per square kilometer.

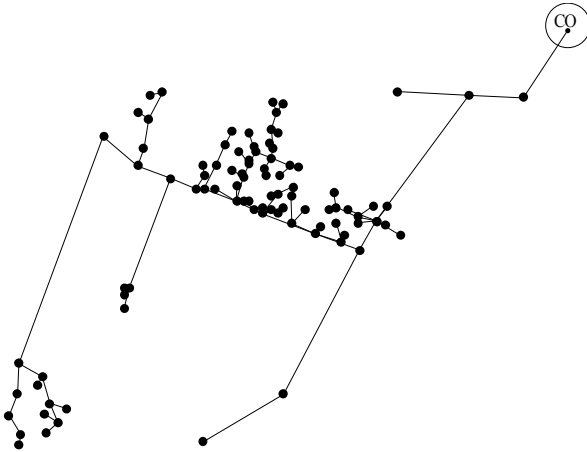
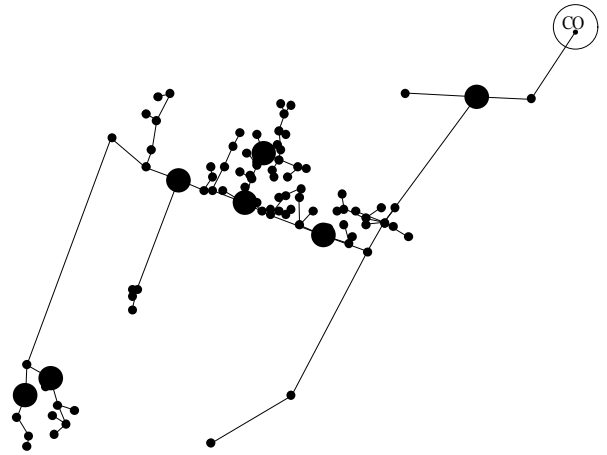


Fig. 5 Characteristic urban LCN.

Optimization scenario results are shown in Table 5. According to different scenario presumptions (1a, 1b, 2) the output parameters are the following: the number of required RDSLAMS, the average optical fiber dig length per household, the average number of households per RDSLAM and the average total traffic at the RDSLAM. The average total traffic at the RDSLAM is calculated considering generated traffic per characteristic household at the present time (Table 2). For all calculations with possible xDSL and FTTN combinations (Scenario 1a, 1b and 2) the input parameter, N , is 24 households (users). The input parameter, N , for the FTTH Scenario 3, is set to 1, because the fiber has to be reachable from every household. Also the output for Scenario 3 is slightly different – in the FTTH case, the RDSLAMS are not

required, therefore only the average optical fiber dig length per household is calculated.

Proposed remote DSLAM installation places for Scenario 1a are presented as larger marked joints in LCN network, as shown in Fig. 6

Fig. 6 RDSLAM placement for urban Scenario 1a ($R = 900$ m, $N = 24$).

Almost linear dependence between the reach of xDSL technologies, R , and the required RDSLAM number, is an interesting characteristic of urban CAN. If the R is doubled, we can plan approximately three RDSLAMS less. The difference between an average optical fiber dig length in Scenario 2 and Scenario 3, and the average number of households per RDSLAM, are important parameters from the economic point of view. In the urban scenario, the average dig length jump per household is 7,4 m or 355 %, if migration from VDS2/FTTN solution to FTTH scenario is considered.

As an additional result of optimization, the dependence between the reach of xDSL technologies, R , and the average optical fiber dig length for different investment thresholds, N , can be presented, as shown in Fig. 7. The investment in optical access network deployment in urban areas, concerning the optical dig length, starts rapidly increasing when the investment threshold, N , is less than 48 households (users) and the reach of xDSL technologies, R , is shorter than 300 m.

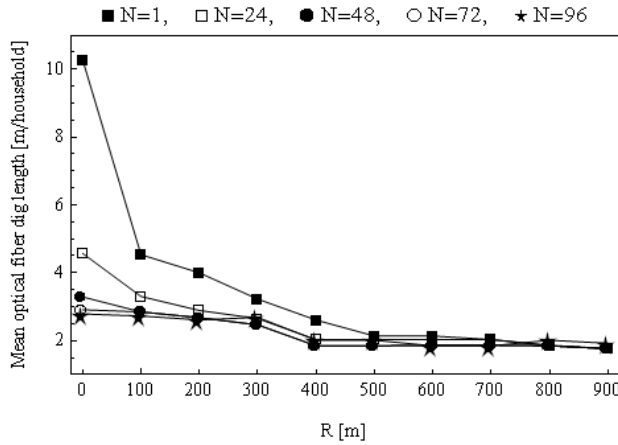


Fig. 7 Dependence between the average optical dig length per household and the reach of xDSL technologies (R) for different investment thresholds (N).

Shortening the reach of xDSL technologies, R , the average calculated total traffic, according to statistical presumptions, decreases, as shown in Fig. 8. Dependence between the average total traffic at RDSLAM and the reach of xDSL technologies, R , for different investment thresholds, N , has almost linear characteristics. That is in accordance with almost linear dependence between reach of xDSL technologies, R , and the required RDSLAM number, N , as shown in Table 5 ($N = 24$).

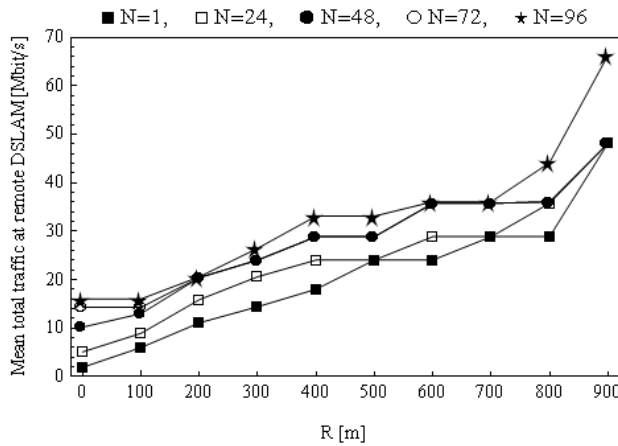


Fig. 8 Dependence between the average total traffic at RDSLAM and the reach of xDSL technologies (R) for different investment thresholds (N).

In the next chapter similar analysis is made for characteristic rural network.

3.2 Rural Copper Cable Access Network

The characteristic rural central office (CO) area consists of up to three major multi-pair copper cables, each with less than five hundred connected users. Each of these multi-pair cables represents a CAN, as shown in Fig. 2. The

optimization planning for the rural CAN with 338 connected households (users) has been made. The belonging local cable network (LCN) is shown in Fig. 9. The circle with label CO represents the central office and black dots represent labeled joints in the LCN. A household density in the corresponding rural municipality is 19 households per square kilometer.

Optimization scenario results are shown in Table 6. According to different scenario presumptions (1a, 1b, 2, 3) the output parameters are the same as in urban network case in the previous chapter. The first important difference between the urban and the rural network analysis is that in the rural case the investment threshold, N , of 24 households (users), excludes the influence of the reach of xDSL technologies on the optimization result. In other words, the RDSLAM is always planned for 33,8 households on average. In this case, many households are not served with corresponding scenario broadband services. The investment threshold, N , is therefore lowered to 6 households (users), as shown in Table 7.

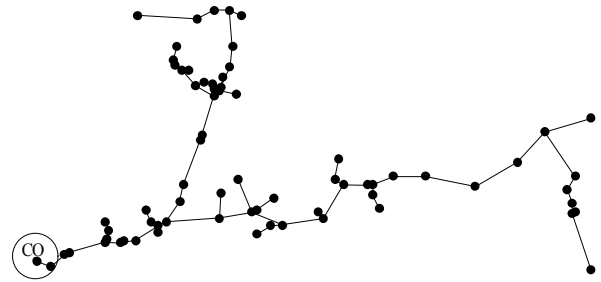


Fig. 9 Characteristic rural LCN.

The dependence between the reach of xDSL technologies, R , and the required RDSLAM number, has an almost linear characteristic – as in the urban CAN network. If the R is doubled, we can plan approximately six RDSLAMs less. In the rural case, the average dig length jump per household is 31,2 m or 42 %, if migration from VDSL2/FTTN solution to FTTH scenario is considered.

The dependence between the reach of xDSL technologies, R , and the average optical fiber dig length for different investment thresholds, N , is presented in Fig. 10. The investment in optical access network deployment in rural areas, concerning the optical dig length, starts rapidly increasing when the investment threshold, N , is less than 12 households (users) and the reach of xDSL technologies, R , is shorter than 300 m.

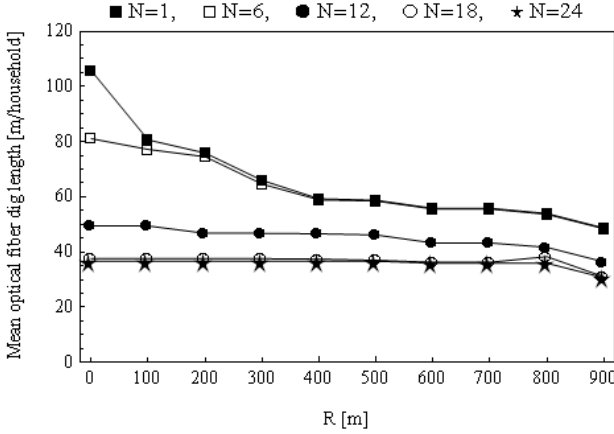


Fig. 10 Dependence between the average optical dig length per household and the reach of xDSL technologies (R) for different investment thresholds (N).

Shortening the reach of xDSL technologies, R , the average calculated total traffic, according to statistical presumptions, decreases, as shown in Fig. 11. Dependence between the average total traffic at RDSLAM and the reach of xDSL technologies, R , for different investment thresholds, N , has almost linear characteristics. That is in accordance with almost linear dependence between the reach of xDSL technologies, R , and the required RDSLAM number, N , as shown in Table 7 ($N = 6$). It can be noticed that the optimization calculation limits constant value, as shown in Table 6. The average total traffic at RDSLAM and the reach of xDSL technologies, R , for investment thresholds, N , higher than 18 households (users), become independent.

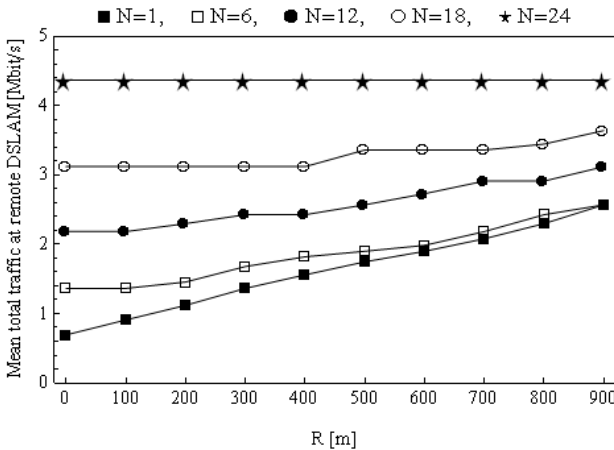


Fig. 11 Dependence between the average total traffic at remote DSLAM and the reach of xDSL technologies (R) for different investment thresholds (N).

If the average considered broadband services (Table 2) are made available to all households (users), the next important issue to be considered is in rural CAN. For average services offered, as shown in Table 2, the 6.127

kbit/s peak download speed is required. That can be offered through local loops, shorter than 3 km, as shown in Fig. 4. Proposed RDSLAM installation places for scenarios offering average services are shown in Fig. 12, where the investment threshold, N , is set to 24, as in the urban case.

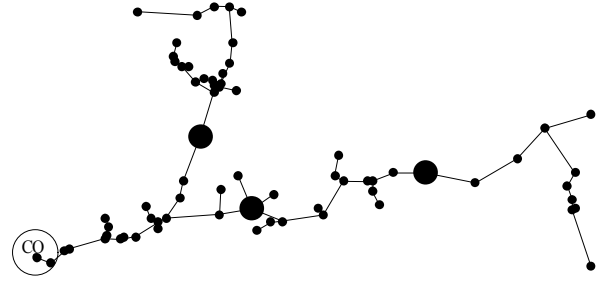


Fig. 12 RDSLAM placement for rural scenario offering average services ($R = 3.000$ m, $N = 24$).

For offering average services through rural CAN, three RDSLAMs are required. In this case the average optical fiber dig length is 22,2 m, the average number of households per RDSLAM is 90,7 and the average total traffic at RDSLAM is 11,7 Mbit/s.

5. Conclusion

Optimization of two presented characteristic broadband access network planning has shown some important results and gives us several techno-economical clues for effective broadband deployment in underserved rural areas. Understanding of difference between rural and urban scenarios is a key issue for bridging the geographical digital divide.

It has been shown that CAPEX of the rural study with FTTH scenario is approximately ten-times higher than CAPEX of the urban study with FTTH scenario, when offering average estimated services and the optical fiber dig length is applied. Urban and rural networks are similar in terms of dependence between the reach of xDSL technologies parameter and average total traffic estimation at RDSLAM if the investment threshold is set below certain value – for urban and rural scenario, $N = 96$ and $N = 12$, respectively. In the urban scenario with FTTH solution and $N = 24$, the average dig length per household is approximately five-times lower than average dig length per household in the rural scenario with FTTH solution and $N = 6$. The CAPEX for VDSL(2)/FTTN rural deployments are from twenty five to thirty-times higher than the CAPEX for the same urban deployments if the optical fiber dig length is applied. If the average presumed services have to be delivered to the majority of households, the investment threshold has to be four-times lower for the rural scenario than the urban one. The cost of

RDSLAM installation is therefore much higher in rural scenarios and the average total traffic at RDSLAM is ten to twenty-times lower than in urban ones. Low population density areas are therefore appropriate for wireless technology deployments (e. g.: WiMAX).

Further research work will involve method improvement for fixed broadband wireless systems planning and implementation of techno-economical

parameters in output parameters (e. g.: internal rate of return, net present value). Input parameters will be extended using infrastructure costs for economic evaluations. One of the important future studies is statistical comparison of different copper access networks from various operators with focus on optimal technology selection for bridging geographical digital divide in sparsely populated areas.

Table 5 Optimization results for urban scenarios ($N = 24$).

Scenario	R	N	# of remote DSLAMs	Average optical fiber dig length per household	Average # of households per remote DSLAM	Average total traffic at remote DSLAM
1a	900 m	24	3	1,8 m	372,3	48,1 Mbit/s
1b	400 m	24	6	2,0 m	186,2	24,1 Mbit/s
2	200 m	24	9	2,9 m	123,0	15,9 Mbit/s
3	FTTH	1	FTTH	10,3 m	FTTH	FTTH

Table 6 Optimization results for rural scenarios ($N = 24$).

Scenario	R	N	# of remote DSLAMs	Average optical fiber dig length per household	Average # of households per remote DSLAM	Average total traffic at remote DSLAM
1a	900 m	24	10	30,7 m	33,8	4,4 Mbit/s
1b	400 m	24	10	36,7 m	33,8	4,4 Mbit/s
2	200 m	24	10	36,7 m	33,8	4,4 Mbit/s
3	FTTH	1	FTTH	105,7 m	FTTH	FTTH

Table 7 Optimization results for rural scenarios ($N = 6$).

Scenario	R	N	# of remote DSLAMs	Average optical fiber dig length per household	Average # of households per remote DSLAM	Average total traffic at remote DSLAM
1a	900 m	6	17	48,4 m	19,8	2,5 Mbit/s
1b	400 m	6	24	58,9 m	14,1	1,8 Mbit/s
2	200 m	6	30	74,5 m	11,3	1,5 Mbit/s
3	FTTH	1	FTTH	105,7 m	FTTH	FTTH

Acknowledgments

The authors would like to thank Telecom Slovenia for cooperation. The work was supported in part by the Ministry of Higher Education, Science and Technology of Slovenia and the Slovenian Research Agency.

References

- [1] Y. Okumura, "Cost analysis of optical access network migration scenarios to broadcast service", IEICE Trans. Commun., Vol.E90-B, no.5, pp.1071-1078, May 2007
- [2] T. Carpenter, M. Eiger, D. Shallcross, P. Seymour, "Node placement and sizing for copper broadband access networks", Annals of Operational Research 106, pp.199-228, 2001
- [3] DSL Forum, www.dslforum.org
- [4] BREAD Project, BRoadband in Europe for All: a multy Disciplinary approach, www.ist-bread.org
- [5] M. Falch, "Techno-economic study", BREAD Project Report, Deliverable no.FP6-IST-507554/COM/R/Pub/D2.4-3.4, May 2006
- [6] G. Vespasiano, M. S. Iacobucci, P. Palma, "Broadband services in the access network: a technical-economic comparison of wired and wireless systems, Proc. 24th Euromicro Conf., Vol. 2, Issue 25-27, pp. 597-603, August 1998
- [7] N. K. Elnegaard, "Techno-economic framework for fixed broadband network deployment in urban and rural areas", ECOSYS Project Deliverable 8, March 2005
- [8] D. Narsingh, Graph theory with application to engineering and computer science, Prentice-Hall, 1974
- [9] S. Pemmaraju, S. Skiena, Computational discrete mathematics, Combinatorics and graph theory with mathematica, Cambridge university press, 2003

- [10] Telecom Slovenia Group, www.telekom.si
[11] Post and Electronic Communications Agency of the Republic of Slovenia, www.apek.si
[12] Statistical Office of the Republic of Slovenia, www.stat.si



Blaž Peternel graduated from the University of Ljubljana, Slovenia in 2003. He received the M.Sc. degree in 2005. At present he is with Mobitel, Telecommunications services, PLC, and works as junior researcher at research group Laboratory for telecommunications at the Faculty of electrical engineering in Ljubljana. His research interests include

wired and wireless broadband access networks optimization with focus on methods for bridging the geographical digital divide in Europe.



Andrej Kos graduated from the University of Ljubljana, Slovenia in 1996 and was awarded his Ph. D. degree in telecommunications in 2003. At present, he works as an Assistant Professor at the Faculty of Electrical Engineering in Ljubljana. He has extensive research and industrial experience in the analysis, modelling and design of advanced

telecommunications elements, networks, systems, and services. His current work and research focus on managed broadband packet switching and next generation intelligent converged services.

Bidirectional Artificial Neural Networks for Mobile-Phone Fraud Detection

Andrej Krenker, Mojca Volk, Urban Sedlar, Janez Bešter, and Andrej Kos

ABSTRACT—We propose a system for mobile-phone fraud detection based on a bidirectional artificial neural network (bi-ANN). The key advantage of such a system is the ability to detect fraud not only by offline processing of call detail records (CDR), but also in real time. The core of the system is a bi-ANN that predicts the behavior of individual mobile-phone users. We determined that the bi-ANN is capable of predicting complex time series (Call_Duration parameter) that are stored in the CDR.

Keywords—Bidirectional artificial neural networks (bi-ANN), fraud detection, mobile telecommunications.

I. Introduction

Fraud in information and communication technologies (ICT) occurs whenever a perpetrator uses deception to receive ICT services free of charge or at a reduced rate [1]. ICT fraud is a global problem that by the estimation of the European Communication Fraud Control Association represents approximately 5% of ICT revenue in developed countries [2] and in some countries even up to 20% [3]. ICT fraud is rising, and with the increased migration of everyday activities into the cyber-world, there is a vital need for more secure and trusted ICT services. The areas in which ICT fraud occurs are extensive. For the purpose of simpler detection and prevention, these are divided into several groups according to their similarities. Different fraud types and their detection methods can be found in [4]–[6].

In this study, we propose a model for mobile-phone fraud detection based on bidirectional artificial neural networks (bi-ANNs) that predict time series representing the behavior of

individual users. The acquired information allows us to predict user behavior and compare it in real time with the monitored real-life behavior. Previous works on predicting time series with a bi-ANN can be found in [7]–[9]. Although time series can be predicted with several other methods [10], we concluded that, in the observed cases, prediction with a bi-ANN delivers better results.

II. Detection Model

Most existing solutions [4]–[6] for detecting ICT fraud give satisfactory results, but are limited to detecting only one type of fraud and provide fraud detection only by offline processing.

To overcome these two issues we propose a system that is able to detect changes in a user's behavior (see Fig. 1.) The first benefit of our approach is that, by detecting changes in a user's behavior, any type of potential fraud can be identified. Another benefit is the ability to monitor these changes in real time. The proposed model applies fraud detection in three steps: monitoring the user, predicting the user's behavior, and

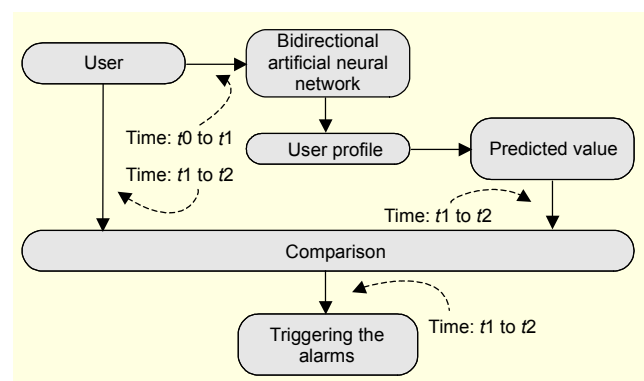


Fig. 1. Proposed model for mobile-phone fraud detection.

Manuscript received Aug. 12, 2008; revised Nov. 3, 2008; accepted Dec. 11, 2008.

Andrej Krenker (phone: + 386 1 476 81 12, email: andrej.krenker@ltfe.org), Mojca Volk (email: mojca.volk@ltfe.org), Urban Sedlar (email: urban.sedlar@ltfe.org), Janez Bešter (email: janez.bester@fe.uni-lj.si), and Andrej Kos (andrej.kos@fe.uni-lj.si) are with the Laboratory for Telecommunications, University of Ljubljana, Ljubljana, Slovenia.

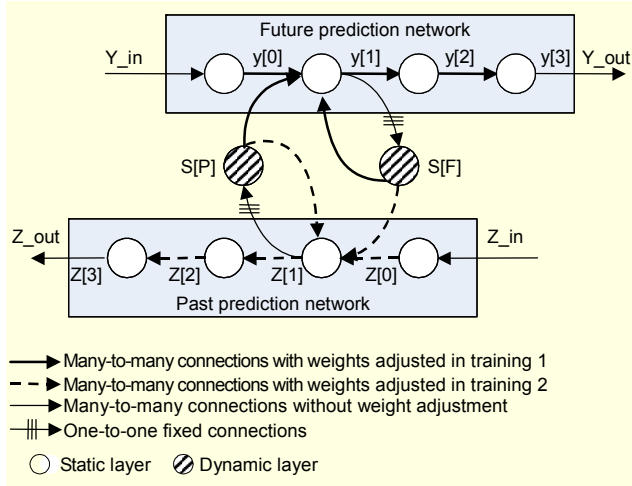


Fig. 2. Bidirectional artificial neural network.

comparing the predicted and monitored behavior. In the case of notable discrepancy between the predicted and monitored behavior, the system triggers an alert. A study of the bi-ANN and its ability to predict time series is presented in the following sections.

III. Bidirectional Artificial Neural Network

In the proposed model, we use the bi-ANN architecture that was first described in [11]. Its use for prediction of time series is described in [12]. Detailed descriptions of the bi-ANN can be found in [7], [9]–[12].

The bi-ANN model shown in Fig. 2 consists of two unidirectional ANNs that are connected through dynamic neurons. The upper ANN predicts future values, and the lower ANN predicts past values [10]. The architecture of the individual ANN is $[1-N-N-1]$, where N is the number of the neurons in the layers. In our model, we varied N from 1 to 7.

IV. Methodology

We obtained anonymized call detail records (CDRs) for 200 users from a Slovenian mobile operator. Our data set consisted of 1,082,588 calls made in time span of 12 weeks with an average call duration of 72 s (minimal call duration was 1 s and maximum call duration was 6,682 s). A CDR holds 40 different call parameters (see Fig. 3). For the purpose of analysis we observed the Call_Duration parameter.

With our model, we attempted to predict the Call_Duration parameter using the following variables: users, data pre- and post-processing methods, training scenarios, training functions, lengths of input time series, time slots of the predicted time series, length of time slots of the predicted time series, and

```

25706747, 170001, 3682978854, 1, 1, 0, 1,0, 974406530,
974406530, 2004-12-13, 00:00:01, 0, 81037355531995,
BKIEVM, 19, BAMSTR, 26,, 0, 0, 7, 127, 0, 34, 7, 00:00:02,
00:00:00,,0, 0, 0, 0, 0, 0, 0, 0,, i170001_20041212_0109.ama
  
```

Fig. 3. Sample of call detail record of randomly selected user.

number of neurons used in hidden layers. For each combination of the above variables and input signal (representing Call_Duration value in time), we performed 10 simulations of the prediction procedure in Matlab. We completed 1,728,000 simulations with different bi-ANN topologies and input time series.

Before we introduced input data to the bi-ANN, we pre-processed it. We applied two different types of pre- (and post-) processing methods, one based on minimum and maximum values and the other based on standard deviation on all 3 data subsets, namely, the training subset, validation subset, and testing subset.

Next, we formed input time series of different lengths, starting at 50 samples and increasing the length by 50 up to 400 samples. After acquiring the resulting predicted time series, we divided these into thirds, and each third was further divided into 5 equal parts. We used different lengths of input and predicted time series to estimate the performance of each model with the average relative variance (ARV) index [12]. ARV is defined as

$$ARV = \frac{1}{\sigma^2 T} \sum_{t=1}^T \{x(t) - \hat{x}(t)\}^2 = \frac{e}{\sigma^2 T}, \quad (1)$$

where $x(t)$ is the desired output series, $\hat{x}(t)$ is the actual output time series, σ^2 is the variance, T is the length of the time series, and e is the total square error. The ARV values between 0 and 1 represent satisfactory prediction, and the smaller the value, the better the prediction. Values above 1 represent unsatisfactory prediction.

V. Results

Based on the obtained results, we have come to the following conclusions. The quality of prediction (the calculated ARV value) was not affected by the selection of users. The study of pre- and post-processing methods showed that methods based on minimum and maximum values resulted in better prediction. Different training scenarios ([10] and [13]) did not affect prediction. However, the applied training function did affect the quality of the prediction task: of the six different training functions (*trainlm*, *trainbfg*, *traincgb*, *traincgf*, *traincgp*, *trainb*), two of them (*trainb* and *traincgp*) had a negative effect on prediction. We also tested and observed how the lengths of the input time series and the predicted time series

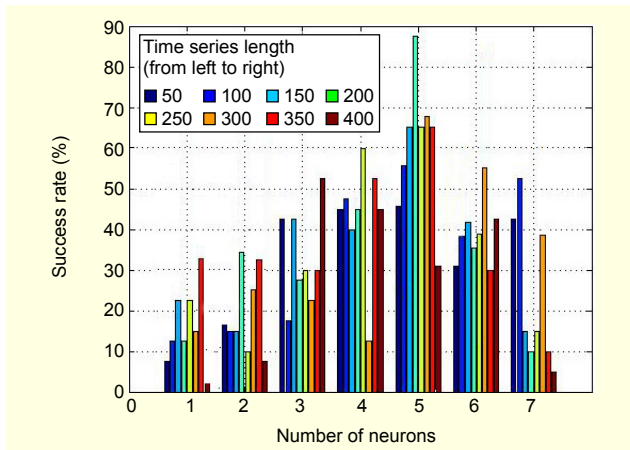


Fig. 4. Combination of parameters resulting in 90% success rate.

influenced prediction. We determined that longer input time series and shorter predicted time series gave better results. The number of neurons used in hidden layers had a major effect on prediction. That is, the bi-ANNs with a larger number of neurons in hidden layers gave better prediction results. However, using too many neurons in hidden layers and short input time series resulted in overtraining, which caused the prediction task to fail. In the case of adjusting variables individually, (see section IV) the percentage of correct predictions was below 50%.

Next, we performed the measurements adjusting all variables at once. The respective results led us to the following conclusion. The combination using ARV as a measure resulted in a 90% success rate (see Fig. 4). By using a combination of five neurons in hidden layers and an input time series length of 200 samples, the best prediction was achieved while using only 20% of the first third of the predicted time series.

VI. Conclusion

In this paper, we presented a novel bi-ANN-based approach for generic mobile-phone fraud detection capable of detecting fraud in real time. The analyses were accomplished using real-life CDR data, obtained from a Slovenian mobile operator. The focus of our study was to determine whether the bi-ANN is capable of predicting time series that describe the behavior of a mobile-phone user. The overall finding of our study is that the bi-ANN is capable of predicting these time series, resulting in 90% success rate in optimal configuration.

In the study, we based the prediction on the Call_Duration parameter in the CDRs. In the future, we intend to extend the prediction to other relevant parameters in order to create a complete mobile phone fraud detection system. Additionally, the ratio between the required prediction accuracy and its consumption of time and resources has to be optimized.

References

- [1] V. Blavette, *Application of Intelligent Techniques to Telecommunication Fraud Detection*, <http://www.eurescom.eu/public/projects/p1000-series/P1007/default.asp>, May 2001.
- [2] h71028.www7.hp.com/ERC/downloads/4AA0-8765ENW.pdf, *HP Fraud Management Systems (FMS) Solution*, June 2005.
- [3] R.J. Bolton and D.J. Hand, "Statistical Fraud Detection: A Review," *Statistical Science*, vol. 3, 2002, pp. 235-255.
- [4] J. Hollmen, *User Profiling and Classification for Fraud Detection in Mobile Communication Networks*, PhD thesis, Helsinki University of Technology, Department of Cognitive and Computer Science and Engineering, Espoo, 2000.
- [5] T. Fawcett and F. Provost, "Adaptive Fraud Detection," *Datamining and Knowledge Discovery*, vol. 1, 1997, pp. 1-28.
- [6] http://www.theregister.co.uk/2005/12/19/terror_phone_clone_scam/, Dec. 2005.
- [7] H. Wakuya, K. Shida, and J.M. Zurada, "Time series prediction by a neural network model based on bi-directional computation style: A study on generalization performance with the computer-generated time series 'Data Set D'," *Systems and Computers in Japan*, vol. 34, no. 10, 2003, pp. 64-75.
- [8] H. Wakuya and K. Shida, "Bi-directionalization of Neural Computing Architecture for Time Series Prediction. III: Application to Laser Intensity Time Record 'Data Set A'," *Proc. INNS-IEEE, Int'l Joint Conf. on Neural Networks*, 2001.
- [9] H. Wakuya, K. Shida, and J.M. Zurada, "Bi-directionalization of Neural Computing Architecture for Time Series Prediction: Application to Computer Generated Series 'Data Set D'," *6th Int'l Conf. Soft Computing*, 2000.
- [10] H. Wakuya and J.M. Zurada, "Bi-directional Computing Architecture for time Series Prediction," *Neural Networks*, vol. 14, no. 9, 2001, pp. 1307-1321.
- [11] H. Wakuya, R. Futami, and N. Hoshimiya, "A Bi-directional Neural Network Model for Generation and Recognition of Temporal Patterns," *Trans. the Institute of Electronics, Information, and Communication Engineers*, 1994, pp. 236-243.
- [12] H. Wakuya and J.M. Zurada, "Time Series Prediction by a Neural Network Model Based on the Bidirectional Computation Style," *Proc. IEEE-INNS-ENNS Int'l Joint Conf. Neural Networks*, vol. 2, 2002, pp. 225-230.
- [13] H. Wakuya and K. Shida, "Time Series Prediction with Neural Network Model Based on the Bi-directional Computation Style: An Analytical Study and Its Estimation on Acquired Signal Transformation," *Trans. IEE Japan*, vol. 122-C, no. 10, 2002, pp. 1794-1802.