híradástechnika

1945 VOLUME LXIII. 2008

info-communications-technology

Multicast Trees in WDM Networks

Security API Analysis

Tactile Sensing Arrays

GRID Applications

Selected Papers **2008/1**

Journal of the Scientific Association for Infocommunications with co-operation with
the National Council of Hungary for Information and Communications Technology

Contents

Editoral Office (Subscription and Advertisements)**:**
Scientific Association for Infocommunications
H-1055 Budapest V., Kossuth Lajos tér 6-8.
Phone: +36 353 1027, Fax: +36 353 0451
e-mail: info@hte.hu • www.hte.hu

Articles can be sent also to the following address:
Budapest University of Technology and Economics
Department of Telecommunications
Tel.: +36 463 3261, Fax: +36 463 3263
e-mail: szabo@hit.bme.hu

Publisher: PÉTER NAGY • Manager: ANDRÁS DANKÓ

HU ISSN 0018-2028 • Layout: MATT DTP Bt. • Printed by: Regiszter Kft.

Foreword

szabo@hit.bme.hu

Our journal is continuing with the practice of publishing English issues regularly, at present twice a year, in July and in January. As before, most of the content is edited from English versions of reviewed research papers, carefully selected from the preceeding five Hungarian issues. In general, we also consider papers from open call, therefore the editors would like to encourage prospective authors to submit their results specifically for the English issues.

Being a selection, the papers' topics span a wide range of issues of current interest as the reader can see from the short summaries below.

Marcell Perényi, Péter Soproni and Tibor Cinkler consider dynamically changing multicast trees in two-layer, grooming-capable optical networks. The continuous changes in the tree members (users) causes a degradation of the tree. Therefore, a huge amount of network resources can be spared by periodically repeated reconfigurations. In this paper, the benefits of reconfiguration are investigated for different multi-cast routing algorithms and reconfiguration periods.

The paper by *László H. Németh and Róbert Szabó* deals with incentives framework for voluntary autonomous cooperation in distributed networks. Today's communication networks are becoming dynamic and have a high degree of autonomy, and they often behave in a selfish way. To eliminate selfish behaviour from the network, a distributed framework has to be defined, that incites network nodes to communicate and cooperate. The paper describes a novel framework to solve this problem.

Vulnerabilities of hardware security modules at Application Programming Interfaces (API) level represent a serious threat, thus, discovering and patching security holes in APIs are important. In the paper by *Levente Buttyán and Ta Vinh Thong*, the authors argue and demonstrate that the application of formal verification methods is a promising approach for API analysis. In particular, an API verification method is proposed which is based on process algebra. The proposed method seems to be extremely well-suited for API analysis.

Tactile sensors are commonly used in industrial, medical or virtual-reality applications. *Gábor Vásarhelyi et al* present a novel tactile sensing array that processes all three components (normal and shear) of the tactile information at every sensory element (taxel, tactile pixel). The processing technology of the integrated micro-sensors is described along with the information coding behaviour of its elastic cover. The paper concludes with a robotic application example, where the three-component force measurements play a fundamental role.

The paper by *Ferenc Riesz* et al presents original research in the field of Makyoh topography, a method based on an ancient principle. The method's application is the qualitative and quantitative study of semiconductor wafers and other mirror-like surfaces.

The paper by *Ágoston Németh et al* presents one of the largest facilities of the solar cell research and development in Hungary – the Solar Cell Innovation Center. The R&D equipment is an integrated vacuum system designed and built for the preparation of thin film Copper Indium Gallium diSelenide (CIGS) solar cell layer structures. The paper reviews the layout of the solar cell structure and the equipment for its preparation, introduces the main materials science issues raising in the CIGS system and presenting challenges for the research.

Sándor Molnár and Géza Szabó present in their paper a comprehensive scaling analysis of the traffic of the four most popular Massively Multiplayer On-line Role Playing Games is presented. The examined MMORPG-games are World of Warcraft, Guild Wars, Eve Online and Star Wars Galaxies. Both server and client generated traffic are analyzed in detail. The study reveals the basic statistical properties of the investigated games focusing on the correlation and scaling behavior.

The aim of the paper by *István Tétényi et al* is to elaborate on Electronic NUmber Mapping (ENUM) technology. An ENUM measuring method is introduced, and several determining parameters are identified and it is shown how these parameters influence the performance of ENUM. Finally, the Hungarian voice communication profile is compared with the measured ENUM performance in order to have sizing guidelines for ENUM related services.

Attila Kertész in his paper examines and compares different research directions followed by researchers in the field of Grid Resource Management, in order to establish Grid Interoperability. The author proposes a meta-brokering approach, which means a higher level resource management by enabling communication among existing Grid Brokers and utilizing them.

Péter Dóbé, Richárd Kápolnai and Imre Szeberényi present a toolkit called Saleve for developing parallel Grid applications, which helps the migration of existing parameter study applications into grid environment. Programs linked against the Saleve library can be integrated into grids using different middleware systems, so the application developer need not deal with the technical details of the middleware.

László Zombory
President of the Editorial Board

Csaba A. Szabó
Editor-in-Chief

Periodic reconfiguration of groomed multicast trees in WDM networks

MARCELL PERÉNYI, PÉTER SOPRONI, TIBOR CINKLER

*Budapest University of Technology and Economics,
Department of Telecommunication and Media Informatics
{perenyim, soproni, cinkler}@tmit.bme.hu*

Reviewed

Keywords: dynamic multicast, light-tree, optical network, grooming, reconfiguration, WDM, performance evaluation, simulation

In a typical multicast scenario the tree members (users attached to the tree) change all the time. New users join the tree, while some existing users leave it. Here we consider these dynamically changing multicast trees in two-layer, grooming-capable, optical networks. The continuous changing of the tree members (users) causes the degradation of the tree. Therefore a huge amount of network resources can be spared by periodically repeated reconfiguration. In this paper the benefits of reconfiguration are investigated for different multi-cast routing algorithms and reconfiguration periods.

1. Introduction

In recent years the traffic due to multipoint network-based applications keeps on growing in transport networks. Multipoint applications include very important broadband services such as digital media broadcasting (e.g. IP-TV, IP-Radio, etc.), VoD streaming, distance learning, virtual private LAN services, etc [1].

In spite of its benefits in terms of bandwidth savings, today the multicast service is not made available to the end users by most commercial ISPs due to a number of practical reasons. This means that today a huge amount of bandwidth is wasted due to multipoint delivery based on application-layer multicasting (ALM) i.e. unicast-based distribution. In this sense, a recent application that may impel the operators to open the multicast service is TV peer-casting. This application is starting to take an unnecessarily high share of the network capacity as the same streaming information comes in and out of the network for thousands of users by unicast relaying.

Nonetheless, even though not directly available to end users, the multicast service is an essential feature present in the core of the transport network because it is the key to the scalable implementation of the triple-play concept: TV channels are usually multicast from a content distributor to local caches/relays near the end users.

In general, it can be said that it is less costly to implement multicast in the lowest layers of the network hierarchy; however, when the underlying technology is connection-oriented – as it is the case of optical networks – the number of supported connections becomes a strict bound. In the case of wavelength-routed optical networks, this limit is set by: the number of lambdas, the amount of multipoint units in optical nodes, their fan-out and the optical power budget. Given this limitation, optimizing light tree construction is quite a relevant challenge in next generation multicast-capable optical networks.

In this paper we investigate the problem of dynamic multicast trees, where the member tree leaves are continually changing. New destination nodes may log in to the tree to receive the content, while other nodes may leave the tree and return at a later time. This corresponds to a scenario where IP membership drives optical tree set up. In a real setting, the tree would be “optical” due to the aggregation of multiple multicast sessions or it could be given by a selected set of individual ultra broadband multicast sessions. Several multicast trees can exist in the network at the same time. If the trees have sub-lambda bandwidths, grooming can be applied to make network utilization more efficient.

A typical application can be a digital media distribution service, where the audience is varying in time. New customers appear, who subscribe for the content, and other customers with expired subscription leave the network. In this case a customer does not necessarily mean an individual home user, rather a local provider (e.g. a local cable-TV provider).

Another example can be a virtual LAN service, where LAN broadcast has to be delivered to all endpoints. In contrast to the previous scenario, this application is less sensitive to minor interruptions in transmission caused by reconfiguration of the multicast tree.

The continuous changing of tree members causes the degradation of the multicast tree in the sense of network and resource costs. This degradation can be cured by regular reconfiguration of the tree. Reconfiguration results in significant spare of network resources (and of the cost), which is clearly beneficial for the operator: resources (including link capacities) that are freed up can be reused.

However, there are also some drawbacks of reconfiguration:

It may consume lots of computation time as computing the Steiner tree is an NP-complete problem. However, considerable saving can be achieved by using faster heuristic methods trading-off speed and optimality.

Reconfiguration can cause a short disruption in the data transmission flow or cause packet reordering, which is sometimes not acceptable by the application and should be avoided. Furthermore, reconfiguration implies an additional signaling overhead.

1.1. Surviving to tree reconfiguration

Although our paper does not intend to solve this problem, we suggest some techniques to show that it is feasible.

A solution for an interruption-sensitive application (e.g. media streaming) is a soft switch-over from one light tree to the new one. In this case the updated light tree is already set up, before the old one is torn down. There is a short period when both trees exist and are able to transmit data at the same time. In order to prevent loss of sequence during the change of the tree, the transmission can be held for a short time at the ingress to guarantee that all the packets are flushed out of the original tree. Alternatively, the first packets that travel through the new tree are buffered at the egress node until an end-of-transmission signaling packet arrives through the old tree. However, smooth reconfiguration needs extra resources from the network. In our simple network model if one free wavelength is available in every link the reconfiguration of one light-tree can be performed – for example by *ILP* (Integer Linear Programming) optimization. In a DWDM network with at least 30 WLs per link this extra capacity is acceptable (especially compared to the huge cost gain, that the optimization results). However, it is not guaranteed that this extra capacity is always available.

1.2. Other publications in the area

Quite a few papers were published in the field of optimizing the cost of multicast routing (light-trees) in optical networks. Since the problem of routing the demands optimally is often infeasible or time-consuming, several heuristic approaches were proposed and their performance was compared with ILP-based optimal solutions.

The problem of static multicast for optical wavelength routing was investigated for ring and mesh networks among others in [2] and [3], respectively. The authors of [3] presented an analytical model of grooming problem represented as non-linear programming formulation and compared the results with heuristic approaches. Heuristic optimization algorithms are proposed in [4-6]. The authors of [7] use an ILP formulation to solve the optimal routing and wavelength assignment problem, and show that a network with only a few splitters and wavelength converters can efficiently transfer multicast demands. Mustafa et al. [8] also presented an ILP formulation and heuristic solutions assuming grooming for minimizing the number of electronic-layer equipments and the number of wavelengths.

In recent time the optimization of dynamically changing multicast trees attracted more attention. In the dynamic case the goal is usually to minimize the blocking ratio, not to route all demands (according to some constraints) as in the static case. This problem in general is even more resource- and computation-intensive than the static version. We found, however, that some subproblems of routing (e.g. optimization of a single tree, or several trees separately) can be solved optimally by ILP. Therefore, it is worth to compare the performance of dynamic routing algorithms to the optimal solution, and to calculate the benefit.

Several provisioning methods of dynamic trees (assuming grooming) are discussed in [9-11].

In [12] traffic engineering is performed through dynamic traffic grooming in grooming-capable WDM networks in the unicast scenario.

The authors of [13] proposed a dynamic wavelength assignment algorithm for multicast to minimize call blocking probability by maximizing the remaining network capacity in each step. Chowdhary et al. addressed in [14] a similar problem by provisioning on-line multicast demands with the objective of increasing the resource utilization and minimizing the blocking probability for the future arriving requests.

Boworntummarat et al. introduced light-tree-based protection schemes against single link failure in [15]. ILP formulations were developed to measure and compare the minimum spare capacity requirement of the proposed protection strategies.

According to our knowledge no work was published analyzing the effect of regular reconfiguration of light-trees, investigating the degradation of dynamic routing algorithms, and comparing the dynamically changing costs to the optimum.

2. Problem formulation

A two-layer network is assumed, where the upper, electronic layer is time switching capable while the lower, optical layer is a wavelength (space) switching capable one. The electronic layer can perform traffic grooming, i.e. multiplexing low bandwidth demands into a single WL channel. The two layers are assumed to be either interconnected according to the peer model [16] or vertically integrated, i.e. the control plane has information on both layers and both layers take part in accommodating a demand.

The network topology and the number of fibers are assumed given as well as the parameters (distribution of inter-arrival time and holding time) of dynamic traffic demands. The capacity of WL channels and the cost of routing, (e.g. space switching, optical to electronic conversion, etc.) can also be given in advance.

We assume dynamic traffic consisting of multicast traffic demands. As explained before, these demands may correspond to an individual ultra-high speed IP multicast session or to a set of aggregated sessions that share most of the leaves. The heuristics for aggregating multiple sessions into a single light tree fall out of the scope of this paper. The same consideration is made regarding joint optimization of light trees and light tree merging: for the sake of simplicity, in this paper light-trees are optimized separately, although, joint optimization could yield a higher cost gain at a higher computational cost.

A multicast tree consists of multiple so-called sub-demands, which can share resources in the network (e.g. their bandwidths are not additional). One sub-demand is assigned to each destination node (member) of the tree. The source of every sub-demand is the single source node of the multicast tree. Destination nodes of the tree change dynamically: new nodes can log in the tree or existing nodes can log out at any time. Paths of new sub-demands have to be calculated online while paths of leaving nodes need to be torn down as carefully as possible not to affect other sub-demands.

Both the active session time (holding time) and the idle (inter-arrival) time for every destination node have an exponential distribution. The traffic load can be determined by appropriately setting the rate parameters (λ) of the distributions. The objective is to reach all current destination nodes from the source in each time step.

3. Network model

We use a wavelength graph model for routing in two layer networks with grooming and with different types of nodes. The model handles any regular mesh topology and supports the peer-model. The WL graph that corresponds to the logical network is derived from the physical network considering the topology and capabilities of physical devices.

A simpler version of the model has been first proposed in [17]. ILP formulation of the static RWA problem with grooming and protection has been given in [18].

The network consists of nodes and links connecting the nodes. Both ends of an optical link (fiber) are attached to an interface (IF) of a physical device, which determines the number of supported WLs in the fiber. Every physical device contains an internal switching fabric and some IFs. Each link and every physical device has a specific logical representation in the WL graph.

A physical link is derived to as many logical edges as the number of available WLs in the link. The logical sub-graph of a physical device depends on the capabilities of the device. Every edge in the graph has a capacity and a cost of usage. The capacity of the edge usually equals to the WL capacity, which depends on the used carrier (typically 2.5 Gbps – which was assumed in our simulations – or 10 Gbps). The cost of the edge is determined by its functionality (WL edge, O/E conversion, etc.).

The WL graph model (together with our ILP framework) can support devices with different capabilities appearing in the network at the same time. The model is easily extendable; the type of devices can be changed later if new internal models are introduced.

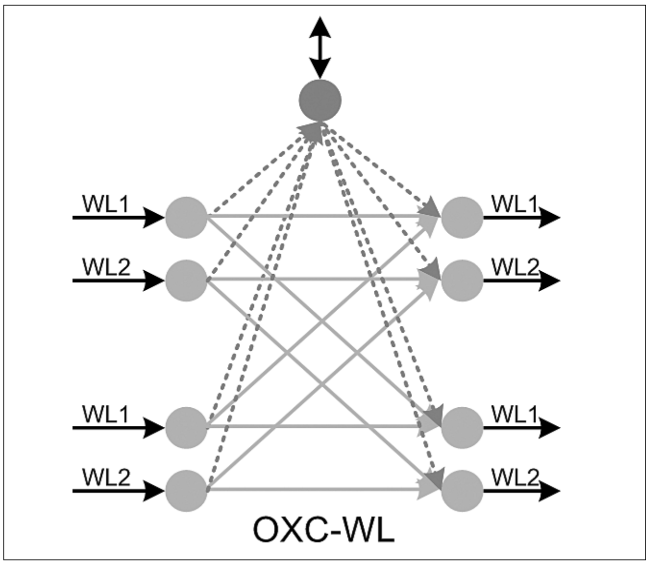


Figure 1.
Sub-graph of an OXC-WL device in the wavelength graph

A sub-graph of a versatile physical device is depicted in *Fig. 1.* The equipment is a combination of an OXC with WL-conversion and an OADM that can originate and terminate traffic demands, as well as perform space switching. WL-conversion and splitting (branching) of light-trees can only be performed in the electronic layer. We will use this complex node in our simulations.

4. Routing algorithms

We applied several algorithms to route the demands in the network. We wanted to compare their costs and performances. A simple example illustrating the different outcome of the algorithms is shown in *Fig. 2.*

4.1. ILP routing and formulation

ILP always provides the optimal cost of routing the current demands in the system, thus it serves as a baseline for comparison. However, this does not necessarily mean that the numbers of certain resources (e.g. wavelengths, O/E, E/O converter ports) are all minimal as well. On the other hand ILP routing usually consumes much time. Fortunately, the routing time of one multicast tree is still acceptable even for larger networks. This time varied from 3 seconds to 180 seconds on a 2.8 GHz Pentium for COST266 network [19], which consists of 28 nodes and 41 links. If we want to route several trees together by introducing grooming much more cost can be spared, however, the solution time becomes unacceptably high. So it is only possible to route different trees separately one after the other.

An important disadvantage of ILP routing is that the consecutive configurations are very dissimilar, thus reconfiguration of the paths of demands (including switching devices along the path) is unavoidable.

We used the ILP formulation introduced in [22] and [23] to route multiple multicast trees in the network.

This formulation is able to route unicast and multicast demands as well, or even demands from both types at the same time.

4.2. Accumulative shortest path (Dijsktra's algorithm)

Accumulative shortest path algorithm is fast and simple. It can be applied for routing a new demand by not interrupting the current active sub-demands in the network. On the other hand this algorithm is rather costly.

The accumulative shortest path algorithm works as follows: routes are calculated between the source and the destination nodes one after the other. The algorithm operates directly on the logical network (wavelength graph). The source and the destination nodes of a sub-demand are the electronic nodes of the corresponding physical device. The cost of already reserved edges of the graph is set to zero, which means it can be used for free.

Paths to leaving destination nodes are cleared. Edges that are not used by the multicast tree anymore are de-allocated (i.e. this sub-demand was the last one that used these edges). Dijsktra's algorithm never modifies paths of existing sub-demands, which unfortunately often results in longer paths.

4.3. Minimal Path Heuristic (MPH)

The MPH algorithm transforms the original wavelength graph into a virtual graph and applies Prim's algorithm [20] to form a minimum cost spanning tree. A virtual graph is a full mesh, in which only the single source and all the destination electronic nodes are presented. The weight of an edge in the virtual graph expresses the cost of the shortest path in the original wavelength graph (which implies that the shortest path has to be calculated for every node pair in back and forth).

Prim's algorithm is applied in this "upper-layer" virtual graph. After the minimal cost spanning tree is found the paths are traced back into the original wavelength graph. Already used edges of the virtual graph are equal to zero when updating the spanning tree after a new destination node logs in. This ensures that paths of existing sub-demands are not modified. Details of MPH algorithm can be found in [21].

4.4. Tree routing

This algorithm is similar to the MPH algorithm, except that it operates in the wavelength-graph, not in a derived "upper layer", virtual graph. It applies the same Prim's algorithm to determine a minimal cost spanning tree in the WL graph. Updating the tree and modification of the edge costs are also similar to the former case.

A phenomenon can occur in case of both, tree routing and MPH that needs attention: trees can branch in such nodes, where splitting is not allowed (i.e. in non-electronic nodes). These forbidden branches need to be corrected by a post-processing. In fact it is pretty simple to solve the problem by moving both branched paths up to the electronic layer.

5. Results

The simulations were carried out on the COST 266 European reference network [26] with the same traffic demands used in case of all algorithms.

In *Fig. 3.* the cost of routing is plotted as a function of elapsed events. Every change of the light-tree (i.e. a destination node enters or exits the tree) is considered as an "event". In Fig. 3 the lower curve marked as ILP represents the optimal cost in every step, while the upper one (marked as Dijkstra with no reconfiguration) stands for the case when no reconfiguration was applied. The middle curve shows the effect of the regular reconfiguration in every 20th event.

In our experiment "Dijsktra without reconfiguration" exceeds the optimal solution by more than 60 percent

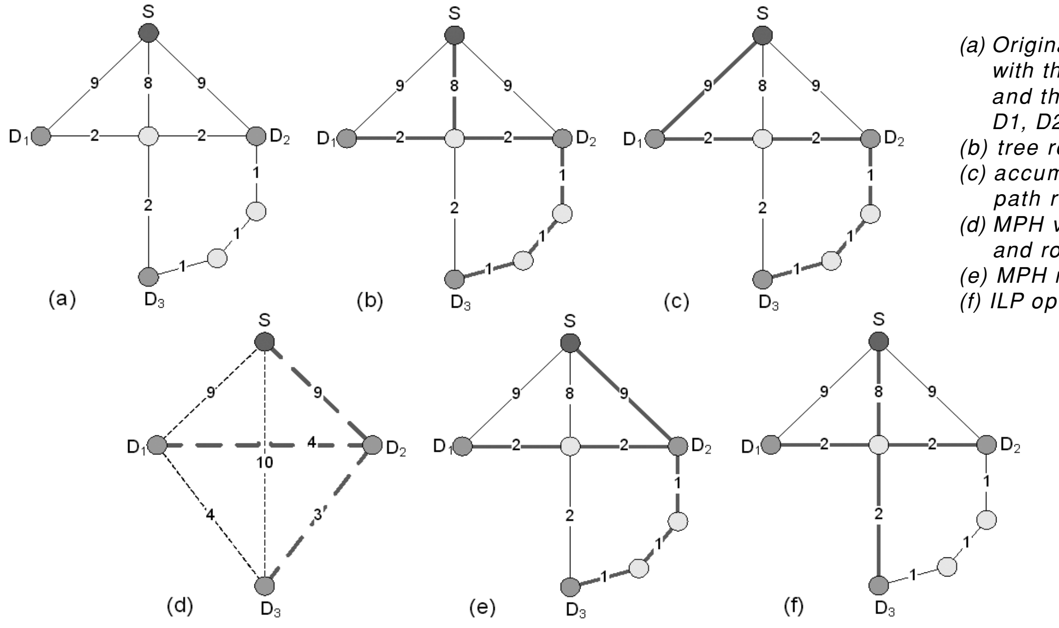


Figure 2.
(a) Original topology with the source node S and three leave nodes D1, D2 and D3,
(b) tree routing,
(c) accumulative shortest path routing,
(d) MPH virtual topology and routing,
(e) MPH routing,
(f) ILP optimal routing

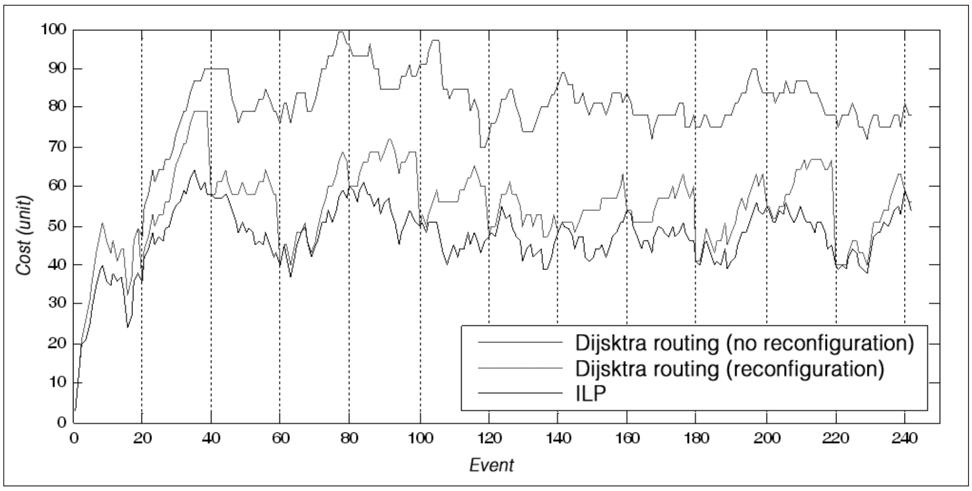
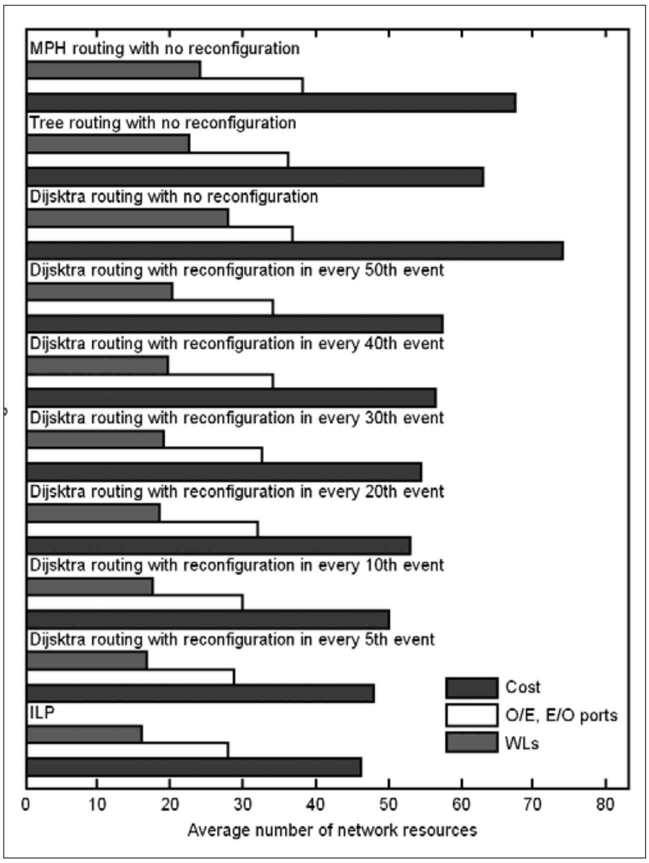


Figure 3.
The cost of routing as a function of elapsed events for Dijsktra's algorithm with (middle curve) and without (upper curve) reconfiguration compared with optimal ILP solution (lower curve)

on average. The reconfiguration curve usually diverges rapidly from the optimal curve. It has the same cost, though, as the optimal one in every 20th event because of the reconfiguration. Although reconfiguration is clearly beneficial (according to Fig. 3), it surely depends on the network topology, the applied dynamic routing algorithm and the reconfiguration period as well.

Therefore we also investigated the cost of different routing algorithms (described in Section 4), and accumulative shortest path routing (Dijsktra) with different reconfiguration periods.

Figure 4.
The average routing cost, conversion ports (O/E, E/O) usage and WL usage of different algorithms and (Dijsktra's) shortest path algorithm with different reconfiguration periods

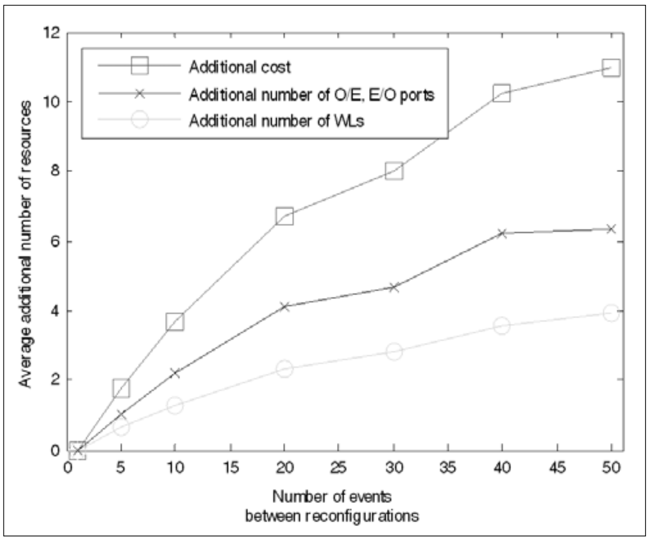


The results are depicted in *Fig. 4.* It is clear, that all of the algorithms (without reconfiguration) are far from optimal: in the current simulation the additional cost is around 34 to 57 percent compared to the optimum. Much cost can be spared by regular reconfiguration. As expected, the shorter the period of reconfiguration, the closer the average cost approaches the optimal value. However, we should know that reconfiguration can be computation-demanding and has other disadvantages as well (see Section 1.1). These drawbacks are not taken into account in the cost.

The results are very similar for network resources necessary to realize the routing: i.e. the number of required O/E and E/O conversion units and the number of wavelengths (Fig. 4).

One interesting fact is that Dijsktra's algorithm without reconfiguration has an outstanding WL usage, while the usage of opto-electronic converters is behind MPH routing. Both WL and conversion port usage approach optimal value by decreasing the length of reconfiguration period.

Figure 5.
The average additional cost of routing (upper curve), number of O/E, E/O conversion ports (middle curve) and number of WLs (lower curve) as a function of the length of reconfiguration period



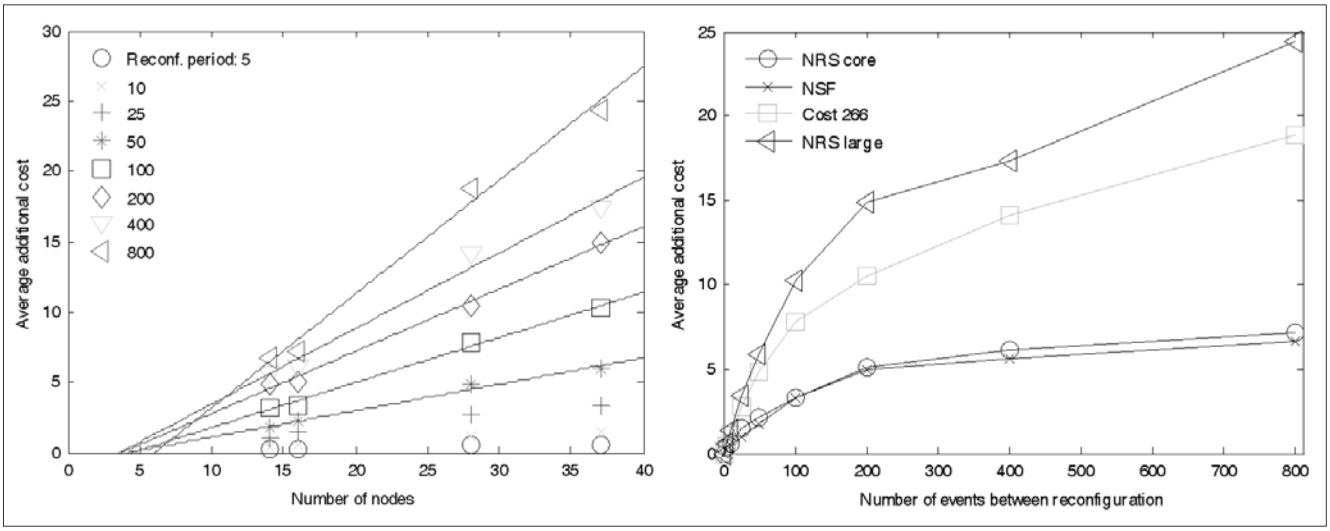
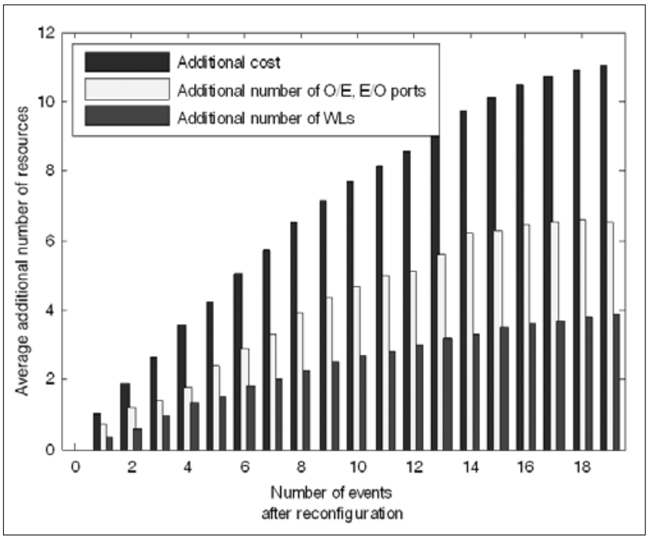


Figure 6.
The average additional cost for different network topologies (left),
and the average additional cost as a function of the number of nodes in the network (right)

We also wanted to investigate how the length of the reconfiguration period affects the cost gain. The average additional cost of routing as a function of the length of reconfiguration period is depicted in *Fig. 5.* The figure shows a saturating curve with decreasing slope. This means, that if we want to reach high cost gain, frequent reconfiguration is necessary. There is not much difference between cost gains, when the periods are long. The required number of WLs and conversion ports follow the same rule, both have a decreasing slope.

We repeated the same measurement for several reference networks to study how the additional cost curve (as a function of the reconfiguration period) looks like in case of different topologies. The same amount of traffic was injected in all of the networks. We obtained similar saturating curves again for all topologies (see *Fig. 6., left).*

Figure 7.
The average additional cost of routing (higher bar),
number of O/E, E/O conversion ports (middle bar)
and number of WLs (lower bar)
after reconfiguration as a function of elapsed events



However the slopes of the curves differ. For larger networks the additional cost rises more rapidly as the length of the reconfiguration period is increased. Therefore we depicted the additional cost as a function of the number of nodes in the network *(Fig. 6, right)*. The symbols mean different lengths of reconfiguration periods; the linear regression was also computed for most of the data series to show the clear linear trend. We found similar relationship between the average additional cost and the number of links in the network. However, the trend is not obviously linear in that case.

Topology	Number of nodes	Number of links
NRS core [24]	16	23
NSF net [25]	14	21
Cost 266 [26]	28	41
NRS large [24]	37	57

Table 1. Reference networks used in the simulations

Based on this experiment it can be assumed that the additional cost is proportional (as expected) to the number of nodes and to the number of links in the network, which means that the larger the network is, the more frequent reconfiguration is required.

Fig. 7. shows how fast the cost of the optimized reconfigured light-tree diverges from the optimal curve. This one is also a saturating curve with decreasing slope, similar to the left one. This suggests that in the first few steps the cost of the tree quickly diverges from the optimal curve, then during the next few events this divergence is slowing down. This kind of divergence is true in terms of conversion ports and WLs as well: after reconfiguration the multicast tree quickly uses more network resources compared to the optimal topology.

The next figure *(Fig. 8)* displays the cost of routing as a function of the number of destination nodes of the light-tree. Each data point corresponds to one time-step in the simulation. The figure compares shortest path rout-

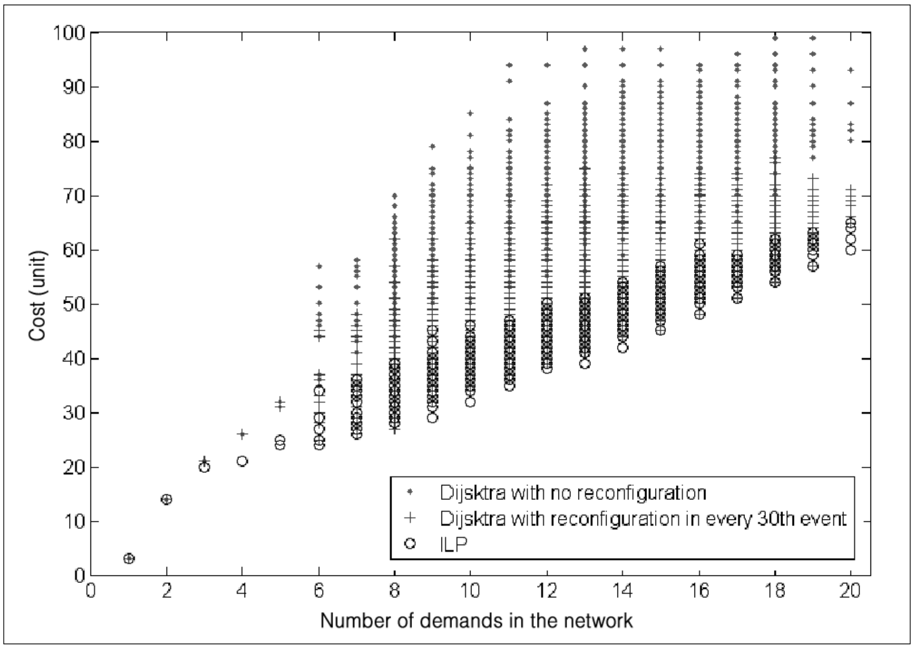


Figure 8. The cost of routing as a function of the number of destination nodes of the light-tree

ing with and without reconfiguration to the optimal solution. As expected, the routing cost naturally raises as the number of the destinations increases. The signs show the typical ranges of the dynamically changing cost for the routing methods. It is noticeable that the range of shortest path with reconfiguration is somewhere between the optimum- and the "no-reconfiguration" range.

In our last experiment we are considering multiple trees (5) at the same time with specific bandwidths. Note, that in this case all trees were optimized separately by ILP in a certain order (in decreasing order of tree size), which does not provide the global optimum. These bandwidths are set so that grooming should be applicable. The routing cost (including conversion port and WL usage) of shortest path routing and ILP are compared. The figures suggest that reconfiguration is more beneficial in case of higher bandwidths, since grooming is less useful in such a case. This observation is true for both the necessary number of conversion ports and WLs, and for the total cost as well *(Fig. 9-11)*.

6. Conclusion

In this paper we showed that reconfiguration of dynamic light-trees is clearly beneficial for the transport network operator. Lots of cost (including network resources, e.g. O/E converter units and wavelength capacity) can be spared by restoring the optimal topology of the tree. Since after the reconfiguration the tree diverges quite quickly from the optimal one frequent reconfiguration is required.

In this paper we have tried to measure the cost saving and the dynamics for periodic reconfiguration with several heuristics. The results show that reconfiguration can be a cost-effective option if the average time between events (subscriptions or leaves) is enough to take advantage of the WLs saving achieved. In this case the saved resources make up for the reconfiguration cost. Still, a number of technical challenges must be addressed to make reconfiguration practical, like the seamless switch over of traffic from the old to the new tree.

Acknowledgements

This work was supported by the EC within the IST FP6 NoE e-Photon/ONe+ (www.e-photon-one.org) and IP NOBEL II (www.ist-nobel.org) research frameworks.

References

[1] B. Quinn and K. Almeroth, "IP multicast applications: Challenges and solutions", IETF RFC 3170, Sep. 2001.

[2] Madhyastha et al., "Grooming of multicast sessions in WDM ring networks" OptiComm 2003, Nov. 2003.

[3] G. V. Chowdhary and C. S. R. Murthy, "Grooming of Multicast Sessions in WDM Mesh Networks", WS on Traffic Grooming, 2004.

[4] X. Zhang et al., "Constrained Multicast Routing in WDM Networks with Sparse Light Splitting", Journal of Lightwave Technology, Vol. 18, Issue 12, p.1917., Dec. 2000.

[5] X. H. Jia et al., "Optimization of Wavelength Assignment for QoS Multicast in WDM Networks", IEEE Transactions on Communications, Vol. 49, No. 2, Feb. 2001.

[6] Fatih Köksal and Cem Ersoy, "Multicasting for all-optical multifiber networks", Journal of Optical Networking, Vol. 6, Issue 2, Jan. 2007.

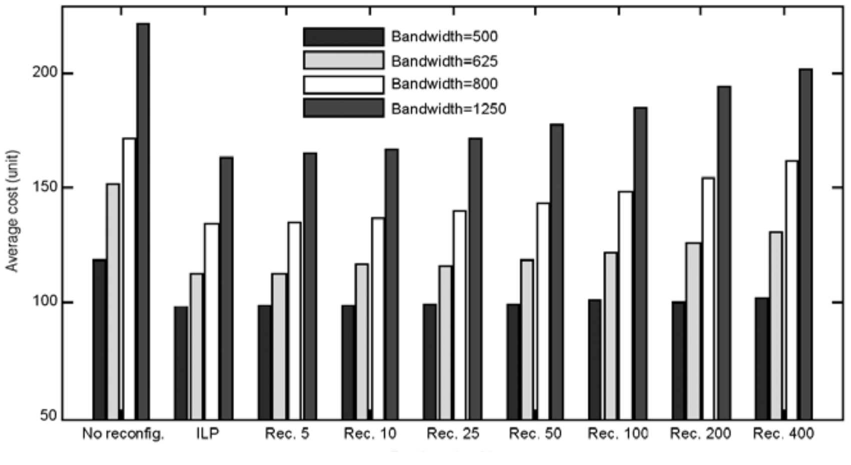


Figure 9.
Average routing cost
for different bandwidth of demands

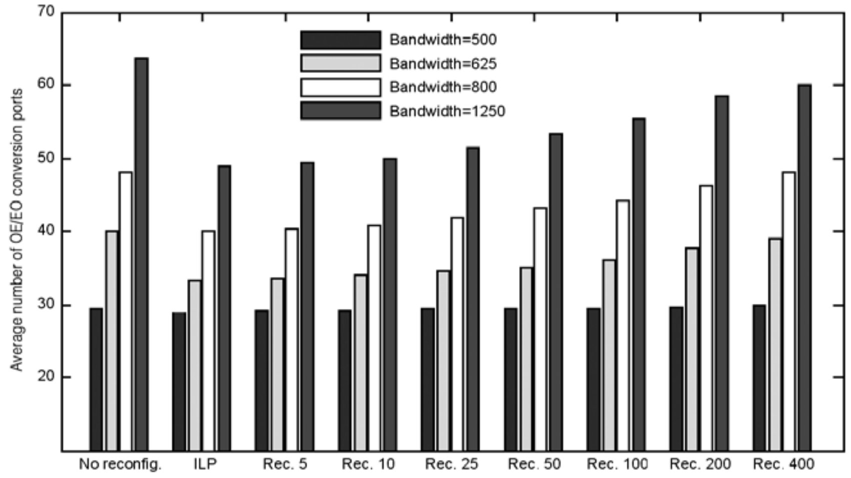


Figure 10.
Average number of
converter ports (O/E, E/O)
for different bandwidth of demands

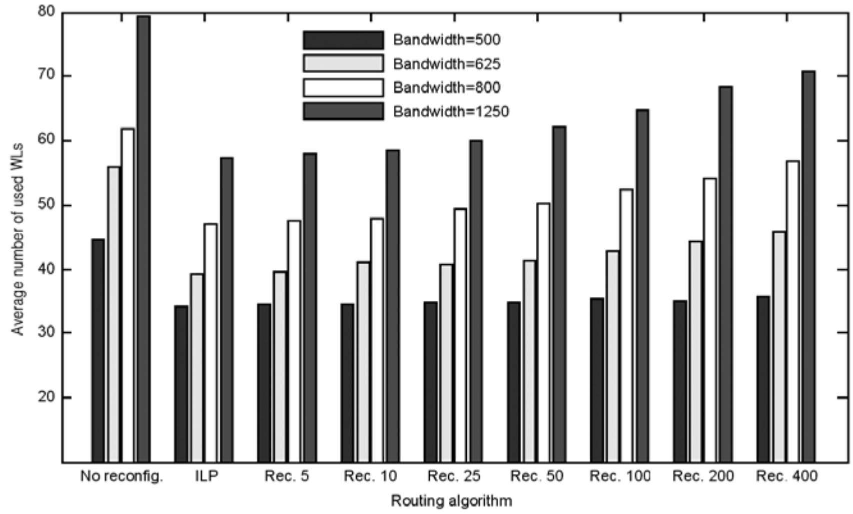


Figure 11.
Average number of used WLs
for different bandwidth of demands

[7] D. Yang and W. Liao,
"Design of light-tree based logical topologies for multicast streams in wavelength routed optical networks," in Proc. IEEE INFOCOM, San Francisco, CA, Apr. 2003.

[8] R. Mustafa and A.E. Kamal,
"Design and provisioning of WDM networks with multicast traffic grooming",
IEEE Journal on Selected Areas in Communications, Vol. 24, Issue 4, 2006.

[9] X. Huang, F. Farahmand and J.P.Jue,
"Multicast traffic grooming in wavelength-routed WDM mesh networks using dynamically changing light-trees",
Journal of Lightwave Technology,
Vol. 23, No. 10, Oct. 2005.

[10] Ahmed E. Kamat et al.,
"Algorithms for multicast traffic grooming in WDM mesh networks",

IEEE Communications Magazine, Vol. 44, Issue 11, Nov. 2006.

[11] A. Khalil et al., "Dynamic provisioning of low-speed unicast/multicast traffic demands in mesh-based WDM optical networks", Journal of Lightwave Technology, Vol. 24, Issue 2, Feb. 2006.

[12] Keyao Zhu et al., "Traffic Engineering in Multi-granularity Heterogeneous Optical WDM Mesh Networks Through Dynamic Traffic Grooming", IEEE NETWORK, Vol. 17, No. 2, Mar./Apr. 2003.

[13] J. Wang and B. Chen, "Dynamic Wavelength Assignment for Multicast in All-Optical WDM Networks to Maximize the Network Capacity", IEEE Journal on Selected Areas in Communication, Vol. 21, No. 8, Oct. 2003.

[14] G. Chowdhary and C. S. R. Murthy, "Dynamic multicast traffic engineering in WDM groomed mesh networks", WS on Traffic Grooming, 2004.

[15] C. Boworntummarat et al., Light-tree based protection strategies for multicast traffic in transport WDM mesh networks with multifiber systems, IEEE Int. Conf. on Communications, Jun. 2004.

[16] E. Dotaro, M. Vigoureux, D. Papadimitriou: "Multi-Region Networks: Generalized Multi-Protocol Label Switching as Enabler for Vertical Integration", Alcatel Technology White Paper, Feb. 2005.

[17] T. Cinkler et al., "Configuration and Re-Configuration of WDM networks" NOC'98, European Conference on Networks and Optical Communications, Manchester, UK, 1998.

[18] T. Cinkler, "ILP formulation of Grooming over Wavelength Routing with Protection", 5th Conf. on Optical Network Design and Modeling, Wien, Feb. 2001.

[19] A. Betker et al., "Reference transport network scenarios", Technical report, BMBF-Project MultiTeraNet, 2003. http://www.pt-it.pt-dlr.de/_media/MTN_Referenz_Netze.pdf

[20] Thomas H. Cormen et al., "Introduction to Algorithms", Second Edition, MIT Press and McGraw-Hill, 2001. Section 23.2: "The algorithms of Kruskal and Prim", pp.567–574.

[21] M. Ali and J.S. Deogun, "Cost-effective implementation of multicasting in wavelength-routed networks", Journal of Lightwave Techn., Vol. 18, No. 12, 2000.

[22] P. Soproni, M. Perényi, T. Cinkler, "Grooming-Enhanced Multicast in Multilayer Networks", ONDM 2007, Athens, May 2007.

[23] M. Perényi, P. Soproni, T. Cinkler, "Multicast fák rendszeres újrakonfigurálása többrétegű optikai hálózatokban", Híradástechnika 2007/8.

[24] NRS network topologies, http://www.ibcn.intec.ugent.be/INTERNAL/NRS/index.html

[25] D. Hart, "A Brief History of NSF and the Internet", Aug. 2003. http://www.nsf.gov/news/news_summ.jsp?cntn_id=103050

[26] R. Inkret et al., Advanced Infrastructure for Photonic Networks: Extended Final Report of COST Action 266, Faculty of Electrical Engineering and Computing, University of Zagreb, 2003.

Authors

Marcell Perényi received his M.Sc. degree in Computer Science from the Budapest University of Technology and Economics (BUTE), Hungary, in 2005. He is currently a Ph.D. student at the Department of Telecommunication and Media Informatics. He has participated in several research projects supported by the EU and the Hungarian government. He is a member of IEEE and the secretary of the HTE Education Committee. His research interests include simulation, algorithmic optimization and planning of optical networks, as well as identification and analysis of traffic of IP networks, especially P2P, VoIP and other multimedia applications. He has experience in planning and maintaining of database systems, web services and Microsoft infrastructures.

Péter Soproni is a graduate M.Sc. student at Budapest University of Technology and Economics (BUTE), Department of Telecommunication and Media Informatics. He has participated in several research projects supported by the EU and the Hungarian government. His research interests include algorithmic optimization, simulation and planning of optical networks. He has experience in .NET based software development, as well as soft-computing especially bacterial algorithms.

Tibor Cinkler has received M.Sc.('94) and Ph.D.('99) degrees from the Budapest University of Technology and Economics (BUTE), Hungary, where he is currently associate professor at the Department of Telecommunications and Media Informatics. His research interests focus on optimisation of routing, traffic engineering, design, configuration, dimensioning and resilience of IP, Ethernet, MPLS, ngSDH, OTN and particularly of heterogeneous GMPLS-controlled WDM-based multilayer networks. He is author of over 180 refereed scientific publications and of 4 patents. He has been involved in numerous related European and Hungarian projects including ACTS METON and DEMON; COST 266, 291, 293; IP NOBEL I and II and MUSE; NoE e-Photon/ONe, NoE e-Photon/ONe+ and NoE BONE; CELTIC PROMISE and CELTIC TIGER 2; NKFP, GVOP, ETIK; and he is member of ONDM, DRCN, BroadNets, AccessNets, IEEE ICC and Globecom, EUNICE, CHINACOM, Networks, WynSys, ICTON, etc. Scientific and Program Committees. He has been guest editor of a Feature Topic of the IEEE ComMag and reviewer for many conferences and journals.

Incentive scheme for
voluntary and autonomous cooperation in distributed networks

LÁSZLÓ HARRI NÉMETH, RÓBERT SZABÓ

Budapest University of Technology and Economics,
Department of Telecommunication and Media Informatics
{nemethl, szabo}@tmit.bme.hu

Reviewed

Keywords: ambient networks, voluntary cooperation, game theory, peer-to-peer, distributed networks, promise theory

Today's communication networks are becoming increasingly dynamic in the sense that they do not have fixed infrastructure, or the configuration of infrastructure-based networks continuously changes. Examples include distributed access networks using WLAN technology, ad-hoc networks, ambient intelligence networks [1,2] or sensor networks. These networks have considerable independence and autonomy and they might frequently act in a selfish manner. Autonomy means that such networks have no central administrative or management principles that would determine their operation.

1. Introduction

In this kind of environment a distributed architecture becomes necessary for the voluntary cooperation of autonomous networks, which controls the cooperations [3]. No central confidence of infrastructure is to be assumed.

Promise theory is a graph theoretical framework, which simplifies the understanding of complex relationships in a network environment that requires compliance with diverse restrictions [3], [4]. According to the basic idea, fully autonomous nodes connect with each other through promises. The cooperative nodes organize groups. Every single promise implies a restriction on the behavior of the promising node.

In large scale distributed networks the components of the network share their services and network-management functions with each other. However, it is not a good choice for the nodes to share all their services with the others.

Each network node needs services from other nodes. If a node only requires services, but does not serve the requests of the other nodes, that means that this node behaves in a selfish way. In order to terminate such behavior in the network and motivate the nodes to cooperate, one may use several kinds of techniques. The principle of these solutions is that one rewards the generous nodes and punishes the selfish ones. If a node receives a reward, it is more likely that its requests will be served by other nodes. If a node receives a punishment, it will be less likely that such node is served. The game theory approach is the most suitable way to model the above described method. The most fitting game for this model is the general prisoner's dilemma. In order to make a decision whether or not to serve a certain service request, the nodes must store some kind of information about the behavior of the other nodes to make the system work.

Behavioral information and history can be stored basically in two ways: by shared history or by private histories [5]. The two storage methods have different drawbacks, in case of storage in a commonly used area a node may send false recommendation related to another node, that is to say it lies about another node and this can ruin the cooperation. To store information in a common field a distributed data-storage method is also required, e.g., by way of distributed hash tables. In case of a large number of nodes individually stored history results in infeasible memory requirements, so the above mentioned method can be used only to a limited extent.

Description of resource sharing by game theory models is a widely researched field, especially since the P2P file-distributed networks have become popular. Several approaches have been developed to motivate the participants of the network to share their resources. In these reputation-based incentive systems the nodes have a utility value, which they want to increase and maximize during their operation. The calculation of the utility value is based on the resource sharing level of the node and the extent of the utilization of other nodes. One of the most comprehensive studies in this field was conducted by Ion Stoica and his team [5], but many other valuable publications were made on this topic. These researches differ in several ways, e.g., the type of the game theory used to analyze the system. Ion Stoica and his team used an asymmetric model with two participants, while for example Philippe Golle conducted his analysis with a multi-agent reinforcement learning model [6].

Existing game theoretic descriptions are based on P2P principle, i.e., any participant may contact any other participant to request or to perform a service. The solution, described in this paper, differs from these approaches in the fact that a topological network is used to deliver service interactions as the chain of physical, node-to-node interactions. As an example, in ambient networks [1] the nodes have only a limited coverage area, so they can communicate directly only with their neighbours. Consequently, routing is required in the network, and a service request goes through several nodes. Therefore, upon a service request three different kinds of nodes participating in the process can be distinguished: an initiator node, which requests the service, a target node, from which the service is requested and optionally some transport nodes, which transmit the requests and the answers. Naturally, a node may request service from its direct neighbor. In this case the transport nodes are left out.

2. System Model

Game theory is a branch of mathematics trying to answer the question: which behavior is reasonable in a situation when the results and effects of a participant's decisions are also affected by other participants' decisions. The description of a game basically requires the specification of three elements: the players, the strategies and the payments, or in other word, payoffs.

Players are the participants of the game, who want to maximize their payoffs. By strategy we mean the behavior of the players, namely, the kind of decisions the players may come to. By payoff we mean the player's utility diagram, the value, which may be recorded to the player's credit at the end of the game. This value depends on the strategy the player has chosen and the strategies of other players. Since the player is rational, he wishes this utility value to be as high as possible. To reach this, the player has to consider the other players' decisions or decision options, as well as his own payoffs in relation to the above. There are several kinds and classifications of the games, e.g., normal form or extensive form games, symmetrical or asymmetrical, zero sum or non-zero sum games. The easiest way to specify a normal form game is the payoff matrix. This matrix shows the players, the strategies and the payoffs.

In order to understand the operation of the system first we should discuss the prisoner's dilemma. There are many versions of this game. The basic idea is that two prisoners, suspected of a crime are imprisoned in separate cells. They have the same options: if a prisoner testifies against the other he will be released and the other is punished to 10 years' imprisonment. If neither of them testifies, they receive 6 months each, if both of them testify, they get 6 years each. The prisoners must not communicate with each other hence they are unable to cooperate (non-cooperative game). Thus the duration of the punishment may be considered as a kind of negative utility we wish to minimize. The payoff matrix of the above described game is illustrated in *Table 1* (in a cell the first number is the payoff of the Player 1 /utility/ and the second number belongs to the Player 2).

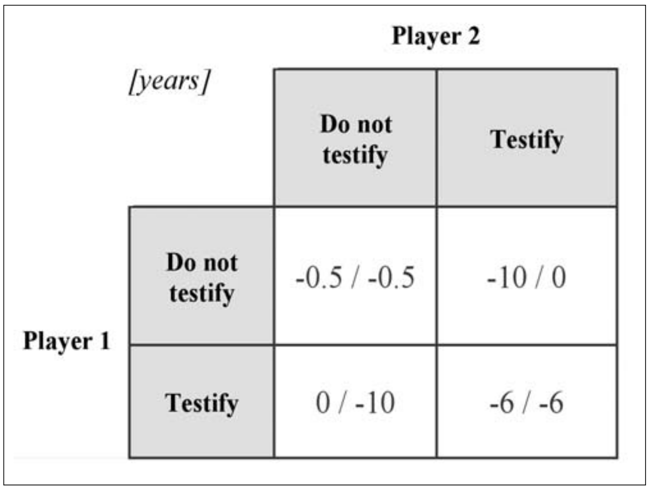
The difference between the original and the generalized game is that several restrictions and rules were defined for the payoff values. Based on the above various prisoner's dilemma games may be described which fulfill these rules. We do not discuss these in details.

For asymmetric games, like a client-server interaction, the classical prisoner-dilemma game can be extended as shown in *Table 2*. The numbers in Table 2 indicate the utility and payoffs of certain players. This game is played many times by the participants of the network and the scores are cumulated. In the very case of *Table 2*, when a node requests a service from another node, two events can occur: the node either serves the client node's request, in which case the server node receives -1 point and the client receives 7 points, or the server rejects the request, so each of them receive 0 points.

The players may have 3 different strategies: always cooperate, always defeat (never cooperative) and to be reciprocative. The first strategy means that the node fulfils every inbound request unconditionally. The second strategy is the opposite of the first: the node never fulfils any request.

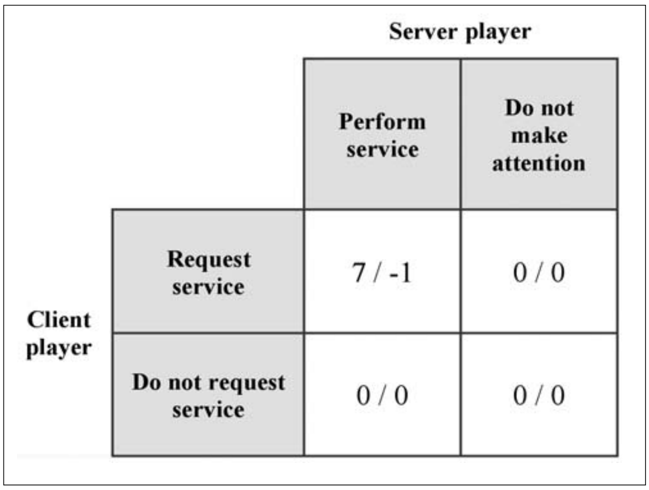
The information about the behavior of the requesting nodes stored by the nodes becomes relevant in the reciprocative strategy. Using this strategy, the decision of a node whether to serve the requesting node or not, is based on some stored information. During the pro-

Table 1.
Payoff matrix - Classical prisoner's dilemma game



		Player 2	
[years]		Do not testify	Testify
Player 1	Do not testify	-0.5 / -0.5	-10 / 0
	Testify	0 / -10	-6 / -6

Table 2.
Payoff matrix for the game played by nodes



		Server player	
		Perform service	Do not make attention
Client player	Request service	7 / -1	0 / 0
	Do not request service	0 / 0	0 / 0

cedure the nodes collect their scores (or loose them) game by game. Each node compiles statistics about which strategy has been the most profitable for them. If a node considers that another strategy would be more profitable than the one it currently uses, it changes strategy. In this case the identifier of the node also changes, so the information about this node stored by the others loses its relevance. (A traitorous node is an exception to this rule, since it keeps its identifier even if it changes strategy. This issue will be discussed later.)

A node may increase its utility not only by serving, but also transferring requests. The value of transferring requests is identical to the value of serving a request. For the requesting node, it is practically transparent who provides the service. The transport of the services is implemented in a way of a routing mechanism. The nodes are aware of the routes through which they can reach other nodes, thus they know which of their neighbors they have to turn to first if they request service from a specific node.

The following question may arise: why would a server node perform services upon a client's request if this results in a negative score for such a node? The answer lies behind the previously described private history stored by the nodes. If a node does not perform services to the other nodes, then sooner or later its requests will be declined as well, so it would be unable to collect scores. This means that in the long term it would not profit from such operation. Performing or not performing services also depends on the relationship of the serving node with other nodes, since as it will be subsequently shown, in certain cases a node may prefer the non-cooperative strategy to the other strategies.

Additionally, a traitorous type of node has also been introduced into the system with the following operation: When this kind of node changes strategy its identifier remains unchanged and the information stored by the other nodes about it also remains valid. Theoretically, a node like this may cooperate with every other node in the first part of the operation, while it refuses to serve any requests in the second part, since due to the high score collected in the first part its requests will most likely be served by the other nodes, which conduct reciprocative strategy. We have examined the operation of the system also in the presence of such of nodes.

During the operation of the system the nodes also store information on the nodes they had previous connections with. The nodes "remember" the clients which had requested services from them. They use this memory when they act as client nodes and they are more likely to request services from those nodes which had already requested services from them. Thus, a node can return a service by performing a request for the other node.

Due to this principle, during the simulation the behavior of the network converges to a relatively stable condition, and although some strategy changes may occur at the last stages of the simulation, no drastic u-turns take place, thus the system becomes stable.

3. Numerical Results

The examination of the above described system was conducted by way of simulation. The simulation was divided into cycles and every node requested service from another node in each cycle, that is, they played the above described game. The game goes through the entire service path, that is, the path on which the performance of services takes place between the client and server nodes. Each simulation contains 1,000 cycles. The examination of the operation of the system was conducted with respect to several cases.

The storing method of histories stored about the nodes was examined both from short-term and long-term respect. If we store such information only for a short-term, this means, that a node may quickly "whitewash" itself, so the system is forgiving, however this behavior might be disadvantageous for the other nodes subsequently. However the storage of long-term history requires extra memory and for satisfactory operation an efficient search must be implemented as well. These two cases we examined in relation to private and shared history.

During the simulation we examined the operation of a network containing 100 nodes. The nodes were randomly positioned, so the topology developed in this way is also random. We examined which strategy is the most profitable for a certain node. The use of a certain strategy depends on several circumstances, e.g., on the position of the node in the network (whether it has a few or a lot of neighbours) or the strategies its neighbours use. At the beginning of the simulation the strategies were randomly distributed between the nodes in the same proportion, thus 1/3 of the nodes were cooperative, 1/3 were defective (non-cooperative) and 1/3 played the reciprocative strategy. In general it can be established that in most cases the cooperative and the reciprocative strategies were the most profitable ones. However, in certain cases, in some parts of the network the non-cooperative behavior became more popular. The system acted differently if the presence of traitorous nodes were also allowed, the proportion of which was set to 25%.

During the simulation the network approached to a stable state. This means that the majority of the nodes were not interested in strategy change and the frequency of strategy changes decreased in the entire network. The diagrams show the number of nodes that use a certain strategy in a certain simulation cycle, but it does not indicate which specific nodes use such strategy, so we cannot find out if the strategy was used by the same nodes or some others. To demonstrate the aforementioned characteristics, we prepared a network topology in each simulation cycle, which indicates each strategy by different color.

By examining these topologies, we came to the conclusion that the bulk strategy changes took place at the beginning of the simulation and at further stages no substantial changes happened. The examination of this process provides an opportunity to focus on the distribution of the strategies depending on position within the topology.

Fig. 1. shows the topology reached by the end of the simulation in case of various simulation scenarios. It is obvious that in the case of the presence of traitorous nodes, the number of the non-cooperative nodes is larger than the number of defective nodes if only normal operating nodes are present in the network. It is worth noticing when the short-term history is used and some traitorous nodes are present every node behaved in a non-cooperative way towards the others shown at extension of the right-hand side portion of the graph. Thus, this effect spread over in that part of the network and such behavior could be observed at the presence of the traitorous nodes. In those parts of the network where the nodes are relatively densely positioned the behavior of the nodes is more or less the same, however, there are some areas, where, because of the presence of the traitorous nodes, the nodes become less cooperative.

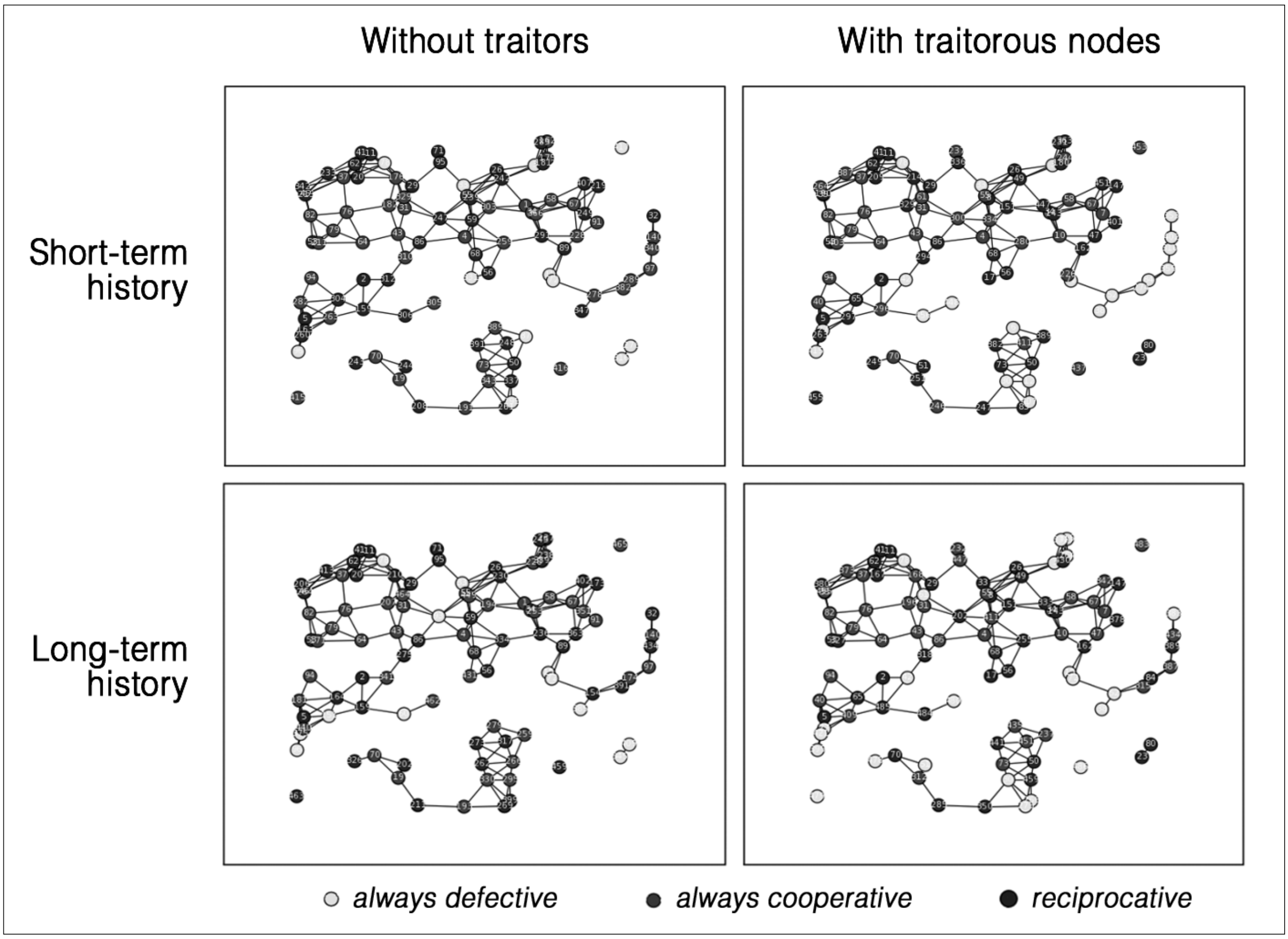
Fig. 2. shows the distribution of the nodes using specific strategies. It can be seen that, if traitorous nodes are present, the distribution of the nodes is more unsteady, the nodes more frequently change strategies. This effect can be clearly seen also when comparing the solutions using short-term and long-term history. In accordance with previous diagram it can be observed that how many nodes followed the various strategies by the end of the simulation. At the presence of the traitorous nodes the difference is clearly noticeable, by the end of the simulation more nodes used the strategy of never cooperating with the others.

4. Summary

In summary, we may establish that the proposed incentive system is able to motivate the nodes to voluntary cooperation. In some cases this cooperation is high-level and the number of the non-cooperative nodes is insignificant, while in other cases some parts of the network form non-cooperative groups.

The study of the system may be continued different ways, e.g., we might examine a specific situation when the nodes are not steady, but change their positions. In this case we certainly must provide effective routing for the nodes to be able to find each other in a quickly changing network. Several studies were made to this effect, however small but frequent changes the network topology had a significant effect on the nodes' strategy selection.

Figure 1. Distribution of nodes by strategies in the topology graph



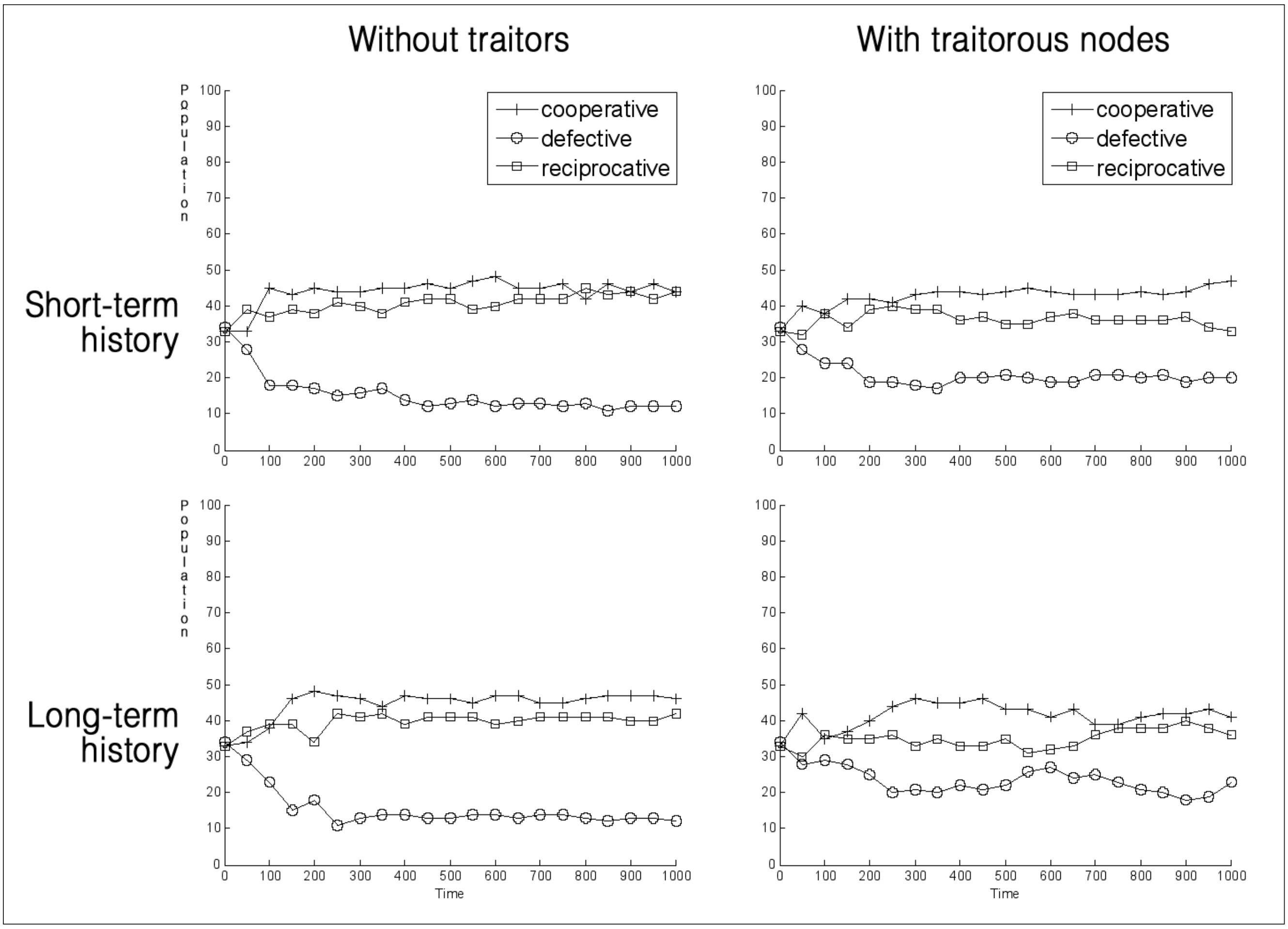


Figure 2. The numbers of the nodes using the different strategies during the simulation

This means, that we may not draw many conclusion from describing diagrams like the above ones. The study of such case constitutes the subject of further research.

References

[1] N. Niebert, H. Flinck, R. Hancock, H. Karl, C. Prehofer, Ambient Networks – Research for Communication Networks Beyond 3G, 2004.

[2] Kovács Balázs, Simon Csaba, "Ambient" hálózatok, 2005.

[3] Mark Burgess, An Approach to Understanding Policy Based on Autonomy and Voluntary Cooperation, Lecture Notes on Computer Science, 2005.

[4] Mark Burgess and Siri Fagernes, Pervasive Computer Management: A Model of Network Policy with Local Autonomy, IEEE Transactions on Networking, 1999.

[5] Michalel Feldman, Kevin Lai, Ion Stoica, John Chuang, Robust incentive techniques for peer-to-peer networks, ACM Conference on Electronic Commerce, June 2004.

[6] Philippe Golle, Kevin Leyton-Brown, Ilya Mironov, Incentives for sharing in peer-to-peer networks, 3rd ACM conference on Electronic Commerce, Tampa, Florida, USA, 2001.

Authors

László Harri Németh obtained his M.Sc. in Computer Engineering, graduated in Budapest University of Technology and Economics in 2006. Currently he is a Ph.D. student of Budapest University of Technology and Economics, Department of Telecommunication and Mediainformatics. His research interests are peer-to-peer and ambient networks and Wi-Fi based positioning techniques for presence and location based services. He participated in the development of positioning algorithms for WLANpos, a Wi-Fi based indoor local positioning system.

Róbert Szabó is an associate professor at the Department of Telecommunication and Media Informatics, Budapest University of Technology (BME). He is the head of the High Speed Networks Laboratory (HSNLab) at BME; and is the President of the Telecommunications Section of the Scientific Association for Infocommunications, Hungary. His main research interests are architectures, protocols and performance of communication networks.

Security API analysis with the spi-calculus

LEVENTE BUTTYÁN, TA VINH THONG

Budapest University of Technology and Economics, Department of Telecommunications
Laboratory of Cryptography and Systems Security
{buttyan, thong}@crysys.hu

Reviewed

Keywords: hardware security module, formal methods, process algebra, security, confidentiality

API level vulnerabilities of hardware security modules represent a serious threat, thus, discovering and patching security holes in APIs are important. In this paper, we argue and illustrate that the application of formal verification methods is a promising approach for API analysis. In particular, we propose an API verification method based on process algebra. The proposed method seems to be extremely well-suited for API analysis as it allows for the straightforward modelling of the API, the precise definition of the security requirements, and the rigorous verification of the security properties offered by the API.

1. Introduction

Hardware Security Modules (HSM) are indispensable in many applications, such as ATM (Automatic Teller Machine) networks, public key infrastructures, electronic ticketing in public transportation, electronic payment systems, and electronic commerce, in general. A HSM is a hardware device (including the firmware and software components) which has some tamper resistance properties, and it is used to store cryptographic keys and to perform various security-critical cryptographic operations (e.g., generation of digital signatures and PIN codes).

HSMs appeared in civilian applications starting from the late 1960s. At that time, driven by the explosion of the number of banking card forgery attacks, IBM (the main supplier of the computer systems of the banks) developed a system where the customer's PIN was computed from the account number placed on the card by encrypting it using a key called the *PIN derivation key*.

Therefore, the protection of the PIN derivation key against both the bank employees and outside attackers became an importnat requirement. This led to the development of the IBM 3848 co-processor, which represents the first generation of HSMs that were widely used in ATM networks later. Today, the application of HSMs is expanded, and besides the banking sector, they became widely used also in Public Key Infrastructures, in Automated Fare Collection systems, and generally in electronic commerce.

The primary goal of attacking a HSM is to extract the secret data stored in it. The long list of potential attacks [2] starts with invasive attacks where the attacker physically penetrates the HSM and gains access to its internal parts, and it continues with non-invasive side channel attacks where the operational environment of the HSM (e.g., its timing and power consumption) is observed or manipulated. These attacks can be very effective, but at the same time, they often require expensive equipments. Finally, HSMs can also be attacked through their APIs by exploiting some design weaknesses in the API's logic. Being fully software based, this kind of attacks is much less expensive than physical and side-channel attacks, and depending on the weaknesses that are exploited, it may have devastating effects. This means that attacking HSMs through their APIs has a potentially high risk. Many API attacks have been found against several widely-used, commercially available HSMs, which otherwise provide very strong physical protection [3-7,10,11]. Thus, discovering and patching security holes in APIs are required, ideally, still before the large-scale deployment of the HSMs. At the same time, APIs used in practice are complex, containing hundreds of functions, which renders their analysis difficult.

One promising approach of API analysis is to apply some formal verification method used in software engineering [8,9,11,12,14,16]. In this paper, we follow this approach, and propose an API verification method based on process algebra that seems to be extremely well-suited for the formal modelling of security APIs, the precise definition of the security requirements, and the rigorous analysis of the provided security properties. In particular, the method introduced here is based on the spi-calculus [1], which was originally designed for analysing key exchange protocols. To the best of our knowledge, we are the first who use the spi-calculus for analysing security APIs.

In the rest of the paper, we first introduce API attacks against the Visa Security Modul in Section 2 for illustration and motivation purposes. Similar attacks also work against other HSMs. The subtlety of these attacks motivate the formal API analysis method introduced in Section 4. Our method is based on the spi-calculus, which is briefly reviewed in Section 3.

2. An API attack against the VISA Security Module

The primary function of the VISA Security Module (VSM) is to protect PINs transmitted over the ATM networks. VISA's goal in promoting this technology was to persuade member banks to connect their ATMs to VISA's network, so that a customer of one member bank could get cash from an ATM operated by another member bank. VISA wanted to minimize the loss that could be caused by dishonest or negligent employees at member banks. The goal was to ensure that no single employee of any bank in the network can learn the clear value of any customer's PIN. This means that PIN numbers should not simply be managed in the software running on the mainframes of the bank. Instead, PIN numbers are managed in a physically protected, tamper-resistant environment implemented by the VSM.

Due to the limitations of its internal memory size, the VSM only stores the most important master keys inside the module; other keys are stored outside secured under the master keys. The key storage method of the VSM follows a hierarchical structure [3] illustrated in *Figure 1.*, which has the advantage of efficient key sharing. However, if a key at a top layer is compromised, every key below it in the hierarchy will be also compromised. The VSM uses five different master keys to encrypt other keys according to their relevancy and roles. The VSM supports nine key types to distinguish roles. As we can see, master keys are placed at the top layer of the hierarchy, and are illustrated as circles, and the nine key/data types are illustrated as rectangles at the lower layers. The keys that belong to a given layer and a given type are secured with the corresponding keys at the upper layers, except the master keys.

The master key *ZCMK* (Zone Control Master Key) is used to encrypt *ZCK* (Zone Control Key) keys. *ZCK*s are keys to be shared with other banking networks, used to protect the exchange of working keys. Working Keys (*WK*s) are used to protect trial PINs that customers have entered while they travel through the network on the way to the bank for verification, and are not used for intra-bank communications. Working keys are stored outside encrypted with the Working Master Key (*WMK*). Terminal Communications keys (*TCK*s) are for protecting control information going to and from ATMs, compute MACs of messages exchanged between *VSM*s, and are secured with the Terminal Communication Master Key (*TCMK*). The Terminal Master Key (*TMK*) and the PIN generation key (*P*) are very important keys and are considered as keys with the same relevancy. Thus, they are both encrypted under master key *MK*, in other words, they are treated as the same key type. The *TMK* keys are shared between ATMs and used to protect all keys sent to an ATM. The PIN generation key is used to generate customer PINs, as we know.

Finally, at the lowest layer we can find *user data* that are encrypted with the operational keys according to their type, where *X{ }* means that the user data is encrypted with a key of type *X*.

Before putting a new ATM in operation, the bank has to supply the ATM with every necessary key. To do this, first, a fresh *TMK* key is shared with the new ATM. All other keys are protected with this *TMK* during transmission to the ATM.

The generation of the key *TMK* is as follow: Function *GenerateKeyShares* of the VSM API is called by the Host:

$$Host \rightarrow VSM : \text{"}GenerateKeyShares\text{"}$$

The VSM generates a key part *TMK*$_i$, and at the same time, it prints the key part to a secure printer to which only authorized persons have access:

$$VSM \rightarrow SecurePrinter : TMK_i$$

Then, it returns the key part encrypted under the master key *MK* to the host.

$$VSM \rightarrow Host : \{TMK_i\}_{MK}$$

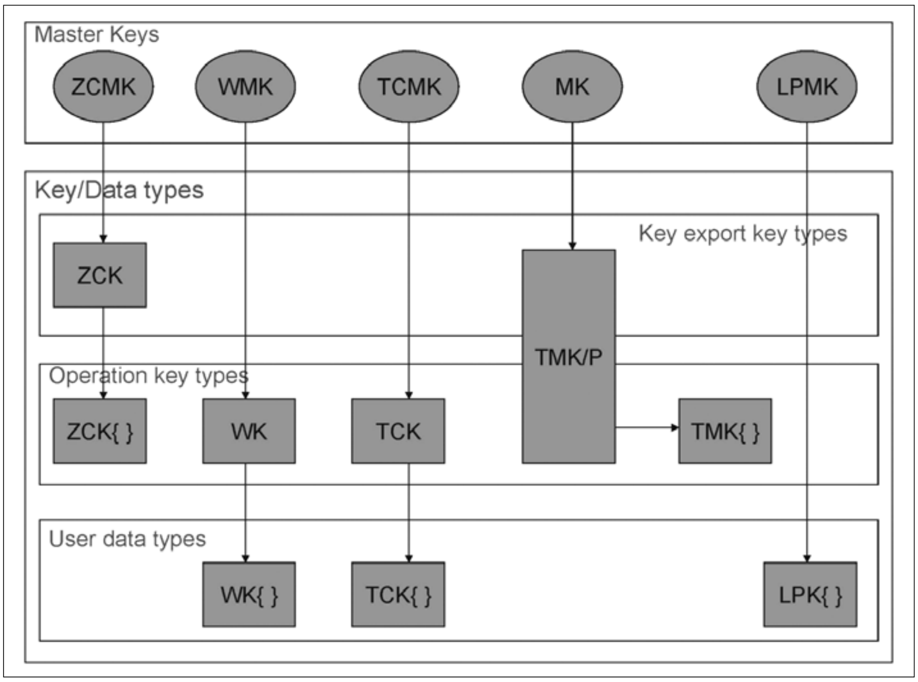
We assume that two key parts are required to construct the *TMK*. The key parts *TMK*$_1$ and *TMK*$_2$ printed by the secure printer are given to separate authorized couriers, who carry it to the new ATM and load it in. After receiving both parts of the key, the new ATM computes the *TMK* key with XORing the two key parts, $TMK = TMK_1 \oplus TMK_2$.

The same *TMK* key is produced at the bank with the *CombineKeyShares* command:

$$Host \rightarrow VSM : \text{"}CombineKeyShares\text{"}, \{TMK_1\}_{MK}, \{TMK_2\}_{MK}$$

$$VSM \rightarrow Host : \{TMK_1 \oplus TMK_2\}_{MK} = \{TMK\}_{MK}$$

Figure 1.
The key hierarchy of the VISA Security Module (VSM)



There exists an API attack that exploits the Terminal Master Key generation function above.

Namely, instead of inputting $\{TMK_1\}_{MK}$ and $\{TMK_2\}_{MK}$, the host (or a programmer at the host) calls *Combine KeyShares* with inputting twice the same key token $\{TMK_1\}_{MK}$ (or $\{TMK_2\}_{MK}$).

$$Host \rightarrow VSM : \text{"}CombineKeyShares\text{"}, \{TMK_1\}_{MK}, \{TMK_1\}_{MK}$$
$$VSM \rightarrow Host : \{TMK_1 \oplus TMK_1\}_{MK} = \{0\}_{MK}$$

Thus, the programmer can achieve that the all zero key becomes the *TMK*. He can then exploit this to produce customer PINs, since the PIN derivation key (*P*) is protected with the *TMK* key during transmission to the ATM for PIN verification. In other words, the programmer can now easily decrypt $\{P\}_0$ with the key 0, and obtains *P* in clear. With the key *P*, he can generate the PIN of any customer.

There is an another attack that uses the *Encrypt CommsKey* function of the API, which inputs a clear *TCK* key and returns the encrypted version under the master key *TCMK*. This key token is stored in an external storage.

$$Host \rightarrow VSM : \text{"}EncryptCommsKey\text{"}, TCK$$
$$VSM \rightarrow Host : \{TCK\}_{TCMK}$$

As mentioned above, every key, including the *TCK* key, must be transferred to a new ATM. The transmission of the key *TCK* is also protected with the master key *TMK*: $\{TCK\}_{TMK}$. The function *TranslateCommsKey toTMK* ensures the generation of this key token:

$$Host \rightarrow VSM : \text{"}TranslateCommsKeytoTMK\text{"}, \{TCK\}_{TCMK}, \{TMK\}_{MK}$$
$$VSM \rightarrow Host : \{TCK\}_{TMK}$$

The attack exploits that *TMK* and *P* are treated as having the same type. The malicious programmer calls *EncryptCommsKey*, but instead of inputting *TCK*, he inputs the customer's account number *PAN*:

$$Host \rightarrow VSM : \text{"}EncryptCommsKey\text{"}, PAN$$
$$VSM \rightarrow Host : \{PAN\}_{TCMK}$$

Next, he calls *TranslateCommsKeytoTMK*, but instead of inputting $\{TCK\}_{TCMK}$, he inputs the resulted key token $\{PAN\}_{TCMK}$ of the previous step. Besides this, he inputs $\{P\}_{MK}$ instead of $\{TMK\}_{MK}$.

$$Host \rightarrow VSM : \text{"}TranslateCommsKeytoTMK\text{"}, \{PAN\}_{TCMK}, \{P\}_{MK}$$
$$VSM \rightarrow Host : \{PAN\}_{P} = PIN$$

The returned value is the account number *PAN* encrypted under the PIN derivation key, which is exactly the PIN number of the account holder.

3. Overview of the spi-calculus

In this section, we give a brief overview of the spi-calculus [1], an extension of the π-calculus [13] with cryptographic primitives. Similarly to the π-calculus, the spi-calculus can be seen as a programming language. Hence, the spi-calculus seems to be well-suited for modeling security APIs.

3.1. Syntax of the spi-calculus

In the spi-calculus, communication channels are represented with names. We assume an infinite set of names. In addition, we assume an infinite set of variables that is important at initialization. Let *x, y,* and *z* range over variables, and let *m, n,* and *c* range over names. We distinguish *terms* and *processes*. Terms (messages, channel identification, keys, etc.) represent data, while processes describe behaviour. A term can be an atom, such as a constant or a variable, or it can be a complex term.

The set of terms is defined by the following grammar:

$L, M, N ::=$	*terms*
n	*name*
(M, N)	*pair*
0	*zero*
$suc(M)$	*successor*
x	*variable*
$\{M_1, M_2, \ldots, M_k\}_N$	*shared - key encryption*

As we can see, a term can be a name, a pair of terms, a constant zero, the successor of a given term, or a variable. We emphasize the term $\{M_1, M_2, \ldots, M_k\}_N$, which represents shared-key encryption, where *N* represents the key, and $M_1, M_2, \ldots, M_k$ terms represent the fields of the plaintext message.

The set of processes is defined by the following grammar:

$P, R, Q ::=$	*process*
$\overline{M}\langle N_1, N_2, \ldots, N_k\rangle.P$	*ouput* $(k \geq 0)$
$M(x_1, x_2, \ldots, x_k).P$	*input* $(k \geq 0)$
$P \mid Q$	*(parallel) composition*
$(\nu n)P$	*restriction*
$!P$	*replication*
$[M \text{ is } N]P$	*match*
0	*nil process*
$let\ (x, y) = M\ in\ P$	*pair splitting*
$case\ M\ of\ 0 : P\ suc(x) : Q$	*integer case*
$case\ L\ of\ \{x_1, x_2, \ldots, x_k\}_N\ in\ P$	*shared - key decryption* $(k \geq 0)$

The above constructions of the spi-calculus have the following intuitive meanings:

- **Output**

Here, the term *M* represents a channel. This process is ready to output terms $N_1, N_2, \ldots, N_k$ on channel *M*. If a *reaction step* (see below) can occur, then terms $N_1, N_2, \ldots, N_k$ are sent on channel *M* and then process *P* runs.

- **Input**

This process is the pair of the ouput process. In a reaction step, an output process sends terms $N_1, N_2,$

...,N_k as a message on channel *M*, and an input process inputs these terms from the same channel, and then process $P[N_1/x_1,N_2/x_2,...,N_k/x_k]$ runs, where *N/x* represents the binding of variable *x* to term *N*. More precisely, variables are substituted with the inputted terms in process *P*.

- **Composition** (P|Q)

This conctruction represents the parallel execution of processes *P* and *Q*. They can interact with each other via channels known to both, or they can interact with the outside world independently of each other.

- **Restriction** (*vn*)*P*

The process *P* creates a new local name *n*. This name cannot appear in other processes unless it has been sent explicitly during some communications. With this construction, we can model the generation of a new secret key.

- **Replication** (!*P*)

This construction represents an infinite number of copies of process *P* running in parallel.

- **Match** ([*M is N*]*P*)

This process behaves as *P* provided that terms *N* and *M* are the same; otherwise it is stuck, meaning that it does nothing.

- **Nil process** (0)

The nil process does nothing.

- **Pair splitting** (*let* (*x,y*)*=M in P*)

If *M*=(*N,L*) holds, then process *P*[*N/x*][*L/y*] will execute, otherwise the process will stuck.

- **Integer case** (*case M of* 0:*P suc* (*x*):*Q*)

This process behaves as *P* if term *M* is 0, and as *Q*[*N/x*], if *M=suc(N)*. Otherwise, the process is stuck.

- **Shared-key decryption**

Process *case L of* $\{x_1, x_2, ..., x_k\}_N$ *in P* attempts to decrypt the term *L* with the key *N*. If *L* is a ciphertext of the form $\{M_1,M_2, ...,M_k\}_N$, then the process will behave as $P[M_1/x_1,M_2/x_2,...,M_k/x_k]$. Otherwise, the process is stuck.

As usual, there are some important assumptions made about cryptography and messages:

– The only way to decrypt an encrypted packet is to know the corresponding key.

– An encrypted packet does not reveal the key that was used to encrypt it.

– There is sufficient redundancy in messages so that the decryption algorithm can detect whether a ciphertext was encrypted with the expected key.

– The attacker cannot find out or generate any secret data of the protocol.

3.2. Modeling secrecy property in the spi-calculus

In the spi-calculus the attacker is an arbitrary *R* process about which we assume only that at the beginning it does not have any secret data. The attacker process runs in parallel with the process that models the system, and they can interact (communicate) via public channels. The attacker attempts to obtain some secret data using only the information that he gets during the interaction.

Secrecy, which is a basic security property in the spi-calculus, is based on the *indistinguishability* of processes. Namely, the system *P* keeps data *M* secret, if for arbitrary data *M'*, the attacker process *R* cannot distinguish *P(M)* and *P(M')*.

A formal definition of indistinguishability in the spi-calculus is given by using the notion of *testing equivalence*. To make this clear, first we introduce some additional notions:

- **Free and bound variables**

Variable *x* is bound in process *P* if process *P* contains an input subprocess *m*(*x*) (for arbitrary *m*). Variable *x* is free in process *P* if process *P* does not contain an input subprocess *m*(*x*). Let *fv*(*P*) denote the set of free variables in *P*.

- **Closed term/process**

We say that a term or process is closed if it has no free variables. In the spi-calculus, we assumed that the attacker process is closed.

- **Reaction step**

A reaction step arises from the interaction of an input process *m*(*x*).*Q* and an output process $\overline{m}\langle M\rangle.P$. During the interaction the output process sends term *M* via channel *m*, while the input process receives it on channel *m*, and binds variable *x* to the received term. Then process *Q* runs with this term. Formally,

$$\overline{m}\langle M\rangle.P \mid m(x).Q \rightarrow P \mid Q[M/x]$$

- **Barb exhibiting**

Exhibiting a barb means that a process uses a given channel to send or receive messages. Barb exhibition is denoted by ↓. Exhibiting a barb is entirely independent from the content of the output or input messages. Barb exhibition is defined by the two axioms:

– **Barb In**: If a process *immediately* uses channel *m* to receive data, then it exhibits the barb *m*, namely, $m(x).P{\downarrow}m$.

– **Barb Out**: If a process *immediately* uses channel *m* to send data, then it exhibits the barb $\overline{m}$, namely, $\overline{m}\langle M\rangle.P{\downarrow}\overline{m}$.

- **Convergence**

Convergence intuitively means that a process does not definitely use a given channel immediately, but only after some reaction steps. Convergence is denoted by ⇓, and there are two related axioms:

– If a process exhibits a barb β, then it will converge to β.

– If a process *P* transforms to process *Q,* that exhibits barb β, then process *P* will converge to barb β.

Next, after introducing the required notions, we give a formal definition of *testing equivalence*:

Definition (Testing equivalence)

A test is a pair (R,β), where *R* is an arbitrary closed process and β is a barb (*m* or $\overline{m}$). Testing equivalence holds between *P* and *Q*, written as $P \approx Q$, if and only if $P \sqsubseteq Q$ and $Q \sqsubseteq P$ holds, where $P \sqsubseteq Q$ holds if and only if $(P|R){\Downarrow}\beta$ implies $(Q|R){\Downarrow}\beta$ for any test (R,β).

Intuitively, $P \approx Q$ means that the behaviors of the processes P and Q are *indistinguishable* for any external observer R. More precisely, P and Q may have different internal structure, but a third process R cannot distinguish running in parallel with P from running in parallel with Q.

4. Modeling security APIs in the spi-calculus

Although the spi-calculus is designed for modelling key exchange protocols, we argue that it is also well-suited for modeling the interaction with a HSM via its API. This is because the interaction can be thought of as a set of two-party protocols, each describing an exchange of messages between the HSM and the user. We can model the entire API as the parallel composition of the replication of the processes that represent individual API function calls. We show an example in this section.

For this purpose, we first define a simplified security API. We assume that the security module has a master key, denoted by *MK*, which is stored inside the module. In addition, we distinguish two types of keys: data encryption keys (denoted by K_i), and key encryption keys (denoted by KEK_j), to which we link the type indicator constants *DataKey* and *KEKKey*, respectively. Key tokens that contain a data encryption key K_i will carry a type indicator *DataKey*. Similarly, key tokens containing key encryption keys KEK_j will carry *KEKKey* as a type indicator.

We also tag encrypted data with the type indicator *TData*. In addition, we assume that the modul does not store K_i and KEK_j inside, instead it exports them in encrypted forms $\{DataKey, K_i\}_{MK}$ and $\{KEKKey, KEK_j\}_{MK}$ under the master key *MK*.

Our example API consist of four functions:

- **Data-encryption**

This function inputs some data *Data* and some key token $\{DataKey, K_i\}_{MK}$. Then, it decrypts $\{DataKey, K_i\}_{MK}$ with the internally stored master key *MK*, and checks its type. If the type is *DataKey*, then it uses K_i to encrypt *Data*. Finally, it outputs the cipher $\{TData, Data\}_{K_i}$.

- **Data-decryption**

This function inputs some encrypted data $\{TData, Data\}_{K_i}$ and some key token $\{DataKey, K_i\}_{MK}$. Then, it decrypts $\{DataKey, K_i\}_{MK}$ with the internally stored master key *MK*, and checks its type. If the type is *DataKey* then it uses K_i to decrypt the cipher $\{TData, Data\}_{K_i}$. Finally, it checks if the type is *TData*, and if so, then it outputs *Data*.

- **Data-key export**

This function takes two key tokens, $\{DataKey, K_i\}_{MK}$ and $\{KEKKey, KEK_j\}_{MK}$ as inputs. It decrypts both of them with the master key, and checks their types. If the types are *DataKey* and *KEKKey* respectively, then it encrypts K_i with KEK_j. It then outputs the key token $\{Data Key, K_i\}_{KEK_j}$. This token will be sent to another modul that may import key K_i.

- **Data-key import**

This function takes two key tokens; $\{DataKey, K_i\}_{KEK_j}$, $\{KEKKey, KEK_j\}_{MK}$ as inputs. It first decrypts $\{KEKKey, KEK_j\}_{MK}$ with the master key *MK*, and checks its type. Then it decrypts $\{DataKey, K_i\}_{KEK_j}$ with KEK_j, and checks its type. Finally, if the types are correct, it encrypts K_i with the master key, and outputs the key token $\{DataKey, K_i\}_{MK}$.

We can model the API defined above with the spi-calculus as follow:

Let $MODULE^{ENC}$, $MODULE^{DEC}$, $MODULE^{EXP}$, $MODULE^{IMP}$ denote the data-encryption, data-decryption, data-key export and data-key import processes. Each process receives data (e.g., input arguments) via channels. The names c_{enc}, c_{dec}, c_{exp}, c_{imp} denote the communication channels through which the processes can receive data. Moreover, we define a channel $\overline{c_{user}}$ through which the processes output data to the environment.

The formal definition of these processes is the following:

1. $MODULE^{ENC}(MK)$

$$c_{enc}(x_{data}, x_{token0}).case\ x_{token0}\ of\ \{x_{typeK}, x_{K_i}\}_{MK}\ in\ [x_{typeK}\ is\ DataKey]$$
$$\overline{c_{user}} < \{TData, x_{Data}\}_{x_{K_i}} >$$

2. $MODULE^{DEC}(MK)$

$$c_{dec}(x_{token1}, x_{token2}).case\ x_{token2}\ of\ \{x_{typeK}, x_{K_i}\}_{MK}, x_{token1}$$
$$of\ \{x_{typeData}, x_{Data}\}_{x_{K_i}}\ in\ [x_{typeK}\ is\ DataKey]\ [x_{typeData}\ is\ TData]$$
$$\overline{c_{user}} < x_{Data} >$$

3. $MODULE^{EXP}(MK)$

$$c_{\exp}(x_{token3}, x_{token4}).case\ x_{token3}\ of\ \{x_{typeK}, x_{K_i}\}_{MK}, x_{token4}$$
$$of\ \{x_{typeKEK}, x_{KEK}\}_{MK}\ in\ [x_{typeK}\ is\ DataKey]\ [x_{typeKEK}\ is\ KEKKey]$$
$$\overline{c_{user}} < \{DataKey, x_{K_i}\}_{x_{KEK}} >$$

4. $MODULE^{IMP}(MK)$

$$c_{imp}(x_{token5}, x_{token6}).case\ x_{token6}\ of\ \{x_{typeKEK}, x_{KEK}\}_{MK}, x_{token5}$$
$$of\ \{x_{typeK}, x_{K_i}\}_{x_{KEK}}\ in\ [x_{typeK}\ is\ DataKey]\ [x_{typeKEK}\ is\ KEKKey]$$
$$\overline{c_{user}} < \{DataKey, x_{K_i}\}_{MK} >$$

Then, the API can be represented as the parallel composition of the replication of the above processes with an initial output of some key tokens. These key tokens are stored outside of the HSM, and thus, they are available to everyone (including the attacker).

$$Sys_{API}(K_i, KEK_j)$$
$$(\nu MK)\begin{pmatrix} \overline{c_{user}} < \{DataKey, K_i\}_{MK}, \{KEKKey, KEK_j\}_{MK} >. \\ (!MODULE^{ENC}(MK) \mid !MODULE^{DEC}(MK) \mid !MODULE^{EXP}(MK) \mid !MODULE^{IMP}(MK)) \end{pmatrix}$$

It is possible to prove formally that this simplified API never leaks out keys in clear. In the formal proof we have to prove the following testing equivalences:

$Sys_{API}(K_i,\ KEK_j) \approx$
$Sys_{API}(K_i',\ KEK_j)$ and $Sys_{API}(K_i,\ KEK_j) \approx$
$Sys_{API}(K_i,\ KEK_j')$ for every $K_i,\ K_i',\ KEK_j,\ KEK_j'$.

The proof of this is based on induction. We assume that at first, the attacker process R does not has any key, that is, the system is in safe state. Then, we prove that if the system is in safe state, it will remain in safe state after any reaction step between process R and the system. This means that the attacker cannot extract any key from the system via its API. We omit further details of the proof here due to space limitations; the interested reader, however, can find the entire proof in [15].

5. Conclusion

API attacks on hardware secutiy modules represent a serious risk. In this paper, we proposed a formal method for analysing security APIs. This method enables us to prove that an external attacker cannot extract any key from the modul via its API (given that indeed this is the case). A failed proof does not directly gives us an attack scenario, however, it often reveals the weak points of the API. The proposed method is based on the spi-calculus, which was originally designed for analysing key exchange protocols. In this paper, we showed that it can also be successfully used to analyise security APIs. Our experience shows that the spi-calculus is well-suited for this kind of analysis.

Acknowledgements

The work presented in this paper has been supported
by the SeVeCom Project that receives funding from
the European Commission withint
the context of the 6th Framework Programme.

References

[1] M. Abadi, A. Gordon,
A calculus for cryptographic protocols: the Spi calculus.
Technical Report SRC RR 149, Digital Equipment Co., Systems Research Center, January 1998.

[2] R. Anderson, M. Bond, J. Clulow, S. Skorobogatov,
Cryptographic processors – a survey.
Techn. Report UCAM-CL-TR-641, Univ. of Cambridge, Computer Laboratory, August 2005.

[3] M. Bond,
Attacks on cryptoprocessor transaction sets.
In Proceedings of the CHES 2001 Workshop, Springer LNCS 2162, 2001.

[4] M. Bond,
Understanding security APIs. PhD thesis,
University of Cambridge, 2004.

[5] M. Bond, R. Anderson,
API level attacks on embedded systems.
IEEE Computer Magazine, October 2001.

[6] M. Bond, J. Clulow,
Encrypted? Randomised? Compromised?
(when cryptographically secured data is not secure).
In Proceedings of the Workshop on
Cryptographic Algortihms and their Uses, 2004.

[7] M. Bond, P. Zielinski,
Decimalisation table attacks for PIN cracking.
Technical Report UCAM-CL-TR-560, University of Cambridge, Computer Laboratory, January 2003.

[8] E.M. Clarke, A. Biere, R. Raimi, Y. Zhu,
Bounded model checking using satisfiability solving.
Formal Methods in System Design, 19 July 2001.

[9] J. Clulow,
The design and analysis of cryptographic APIs.
MSc thesis, University of Natal, South Africa, 2003.

[10] J. Clulow,
On the security of PKCS#11. In Proceedings of the CHES 2003 Workshop. Springer LNCS 2779, 2003.

[11] V. Ganapathy, S.A. Seshia, S. Jha,
T.W. Reps, R.E. Bryant,
Automatic discovery of API-level vulnerabilities.
In Proceedings of the ACM/IEEE Conference on Software Engineering (ICSE), 2005.

[12] A. H. Lin,
Automated analysis of security APIs. MSc Thesis, Massachusetts Institute of Technology, May 2005.

[13] R. Milner, J. Parrow, D. Walker,
A calculus of mobile processes, parts I and II.
Information and Computation, September 1992.

[14] M. Moskewicz, C. Madigan, Y. Zhao, L. Zhang, S. Malik,
Engineering an efficient SAT solver.
In Proc. of the 38th Design Automation Conference (DAC), June 2001.

[15] Ta Vinh Thong,
Security API analysis with the Spi calculus.
Student Scientific Conference, BME, Budapest, 2007.

[16] P. Youn,
The analysis of cryptographic APIs using
the theorem prover otter. MSc Thesis,
Massachusetts Institute of Technology, May 2004.

Authors

Levente Buttyán received the MSc degree in Computer Science from the Budapest University of Technology and Economics (BME) in 1995, and the Ph.D. degree from the Ecole Polytechnique Fédérale de Lausanne (EPFL) in 2002. In 2003, he joined the Department of Telecommunications at BME, where he currently holds a position as an Associate Professor and works in the Laboratory of Cryptography and Systems Security (www.crysys.hu). His research interests are in the design and analysis of security protocols for wired and wireless networks, with a special emphasis on wireless networked embedded systems, including wireless sensor networks, vehicular communications, and RFID systems. Dr Buttyán served on the Technical Program Committee of several conferences and workshops in this field. He was member of the Steering Committee of the European Workshop on Security and Privacy in Ad hoc and Sensor Networks (ESAS), and he was co-chair of the TPC of ESAS in 2006. Currently, he is a Steering Committee member of the ACM Conference on Wireless Network Security (WiSec). He was a Guest Editor of the IEEE Journal on Selected Areas in Communications, Special Issue on Non-Cooperative Behavior in Networking.

Ta Vinh Thong received his MSc degree in Computer Science from the Budapest University of Technology and Economics (BME) in 2008. Since 2007 he has been working in the Laboratory of Cryptography and Systems Security. His research interests are in the analysis of security protocols, his current research activities are security API analysis and formal languages.

Tactile sensing arrays – design and processing

GÁBOR VÁSÁRHELYI, MÁRIA ÁDÁM, CSABA DÜCSŐ, ISTVÁN BÁRSONY

Research Institute for Technical Physics and Materials Science, Hungarian Academy of Sciences
{vasarhelyi, adam, ducso, barsony}@mfa.kfki.hu

ATTILA KIS

Péter Pázmány Catholic University, Faculty of Information Technology
kis@itk.ppke.hu

Reviewed

Keywords: tactile sensor, elastic cover, MEMS, three-axial force sensing

Tactile sensors are commonly used in industrial, medical or virtual-reality applications, but the majority of commercial tactile systems are capable to detect pressure maps only. In this article we present a novel tactile sensing array that processes all three components (normal and shear) of the tactile information at every sensory element (taxel – tactile pixel). We describe the processing technology of the integrated micro-sensors, write about the information coding behaviour of its elastic cover and, finally, we show a robotic application example, where the three-component force measurements play a fundamental role.

1.Introduction

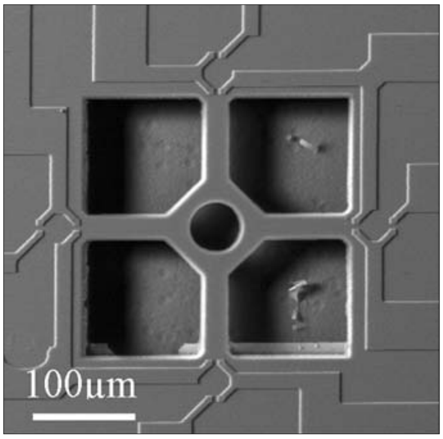
Tactile sensing is probably the second most important, most complex perception of the human body after the vision. The human skin is filled with tiny mechanoreceptors that separate different components of tactile input (static pressure, motion, vibration, etc.) into parallel channels, and send them to the central processing unit, the brain.

Our final goal is to mimic the operation of this sensing-processing system with artificial tactile sensors that could be integrated into robotic hands, medical diagnostic devices or even into e.g. arm prostheses to improve their efficiency.

The core element of our research is a three-axial force sensing array, developed by the Research Institute for Technical Physics and Materials Science (MFA). This tiny MEMS (Micro-Electro-Mechanical System) device is a single-crystalline Si-based sensor that – unlike the commercial tactile sensory arrays – measures and processes not only the normal, but also shear components of the force vectors at its surface.

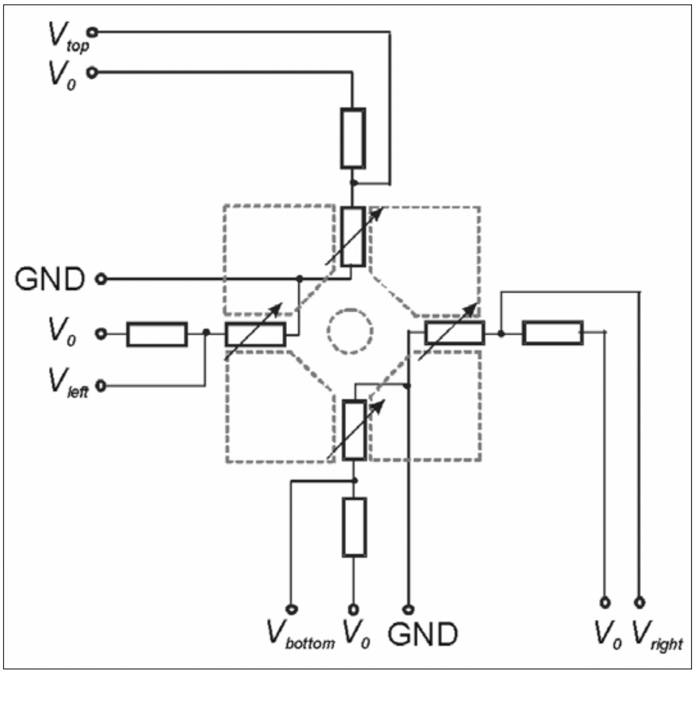
The elastic cover is an indispensable key component of the tactile sensors. Besides offering a certain amount of physical protection, it also plays a fundamental role in the overall procedure of sensation as a mechanical information-coding layer between the sensors and the environment (let us just think about the increased tactile sensitivity around an abrasion, or our thickening sole during summer holidays). The elastic cover can be treated as the *first spatial-temporal, dynamic information-coding layer* of the sensory structure, therefore, its behaviour must be taken into account throughout the design of every tactile device. Our goal is on the one hand

Figure 1.



SEM view of a piezoresistive sensing element (taxel).
Characteristic dimensions:
suspension beams: 80x32x10 µm³,
central reinforced membrane: 100x100x10 µm³,
hole diameter: 50 µm,
cavity etch depth: ~35 µm.
The piezoresistor pairs (deforming and non-stressed reference) are symmetrically formed around the joints of each beam.

Figure 2.



The layout of a taxel indicating the placement and the circuit connection of the four piezoresistor pairs.

to better understand human tactile sensing. On the other hand, we would like to copy ideas from nature to improve the quality of our artificial heptic sensory devices.

The general description of our system is followed by an application example through a *proactive robotic grasping task*.

This article is the English version of a review in Hungarian review that summarizes our previous international publications listed in the references [1-7].

2. Sensors by MEMS technology

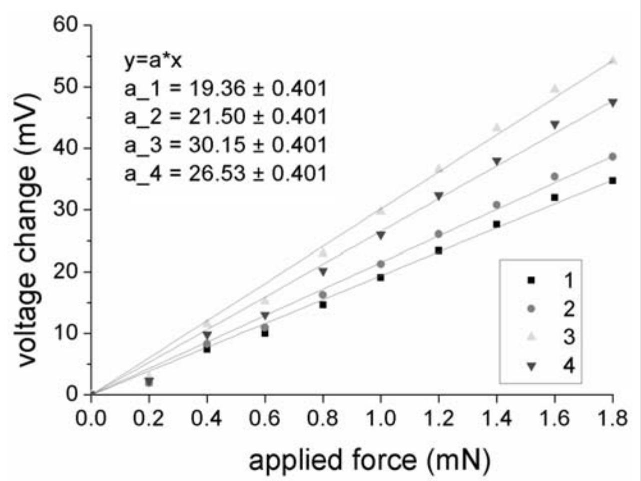
The monolithic tactile sensor arrays are formed in single-crystalline Si (c-Si) by the well-known IC processing technology complemented with an appropriate bulk micromachining technique in order to form the 3D deforming structure. The extraordinary mechanical properties of single crystalline Si combined with the above fabrication techniques enable us to produce intelligent smart sensors of various functions.

All the tactile elements (taxels) of the sensor array are suspended, perforated c-Si membranes with perfectly positioned, embedded piezoresistors *(Fig.1)*. The change of the resistance of each piezoresistor is proportional to the emerging stress in the deforming membrane during loading.

Single-side porous Si bulk micromachining technique was used for releasing the n-type c-Si membranes. The location and direction of the ion-implanted p+ piezoresistors was determined by finite element model calculations in order to select the most sensitive area, i.e. where the load-generated mechanical stress reaches its maximum.

All the piezoresistors are coupled with a serially connected non-deforming reference element placed in the Si bulk maintaining constant resistance even when loading the taxel. The two resistors form a simple voltage divider or a half Wheatstone bridge, therefore, the readout is an analogue DC signal which is proportional to the generated mechanical stress *(Fig. 2)*.

Figure 3.



Linear responses of the four sensing elements exposed to normal load. The sensitivity difference is due to the geometric mispositioning of the attacking needle of the test device.

All the taxels consist of four independent piezoresistor pairs in order to resolve the vectorial components of the attacking force. The linear relationship between the voltage changes and the attacking force in the centre of the sensor element can be described by the following equations:

$$F_x = \frac{1}{V_0 \alpha_{ls} \pi_{44}} \left(\Delta V_{right} - \Delta V_{left} \right),$$
$$F_y = \frac{1}{V_0 \alpha_{ls} \pi_{44}} \left(\Delta V_{top} - \Delta V_{bottom} \right), \qquad (1)$$
$$F_z = \frac{1}{V_0 \alpha_{ln} \pi_{44}} \frac{\left(\Delta V_{left} + \Delta V_{right} + \Delta V_{top} + \Delta V_{bottom} \right.}{2}$$

where F_i are the three components of the attacking force (*z:* normal, *x* and *y:* tangential) V_0 is the common voltage, ΔV is the measured voltage change, π_{44} is the dominant piezoresistive coefficient, α_{ln} and α_{ls} are the linear normal and shear coefficients in the given geometric arrangement.

The measured sensitivity and the linear characteristics of the sensors correspond well to the preliminary finite element model calculations (4-6 mV/mN/V) *(Fig. 3)*.

3. Integrated sensor arrays

Real tactile applications often require sensor arrays of different size and density. Therefore, two array chips were developed. A four element array (2x2) can easily be formed with simple multiplication of the taxels *(Fig. 4)*.

When aiming at further increasing the number of taxels, one faces the contact wiring problem. The large number of wires requires considerable floor space, multilevel metallization and an equal number of bonding pads, which further complicates the inherently critical assembling process.

The metallization problem of large arrays can easily be circumvented by integrating decoders or multiplex-

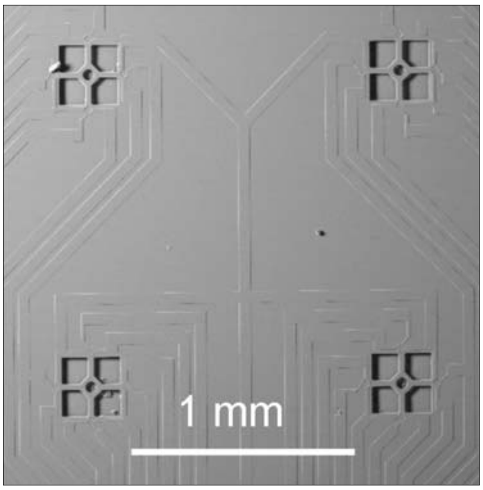


Figure 4. A 2x2 element sensor array. Taxel-size: 0.3x0.3 mm² Spacing: 1.5 mm

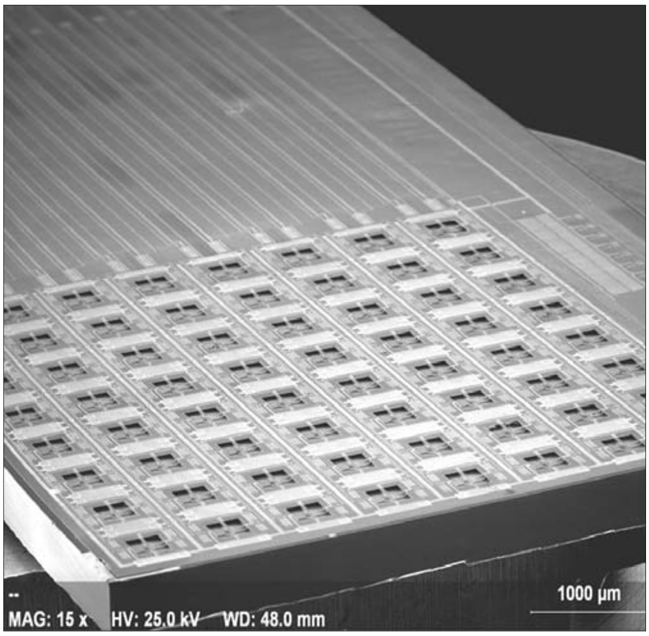


Figure 5.
The 8x8 element tactile sensor array fabricated with CMOS compatible micromachining processes (patent pending)

ers in conventional chips. Nevertheless, in most of MEMS sequences the 3D micromachining as well as the fragile suspended membranes formed make the integration of the required circuitry quite difficult.

Therefore, we developed a proprietary CMOS compatible process sequence, which enables the integration of driver circuitry with porous Si micromachined sensors. Using this patented process we fabricated an 8x8 element tactile sensor chip with on-chip current generators and decoders *(Fig. 5)*.

4. Effects of the elastic cover

As mentioned in the introduction, the elastic cover is an indispensable and fundamental part of every tactile sensor or organ. The elastic layer transfers the surface forces to the sensors in the form of distributed mechanical stress/strain/deformation, no matter which system receives them – mechanoreceptors in the deep skin or artificial tactile sensors – receives them.

Continuum-mechanics is the key word for the mathematical description of the elastic cover of the sensors. In the first run, the elastic material can be treated as a homogeneous, isotropic, infinite *half-space* that obeys Hooke's law. The input forces act only on the open surface of the half-space, and create a complex stress profile inside the elastomer. Since the stress is mostly concentrated around the indentation and decays rapidly with distance, we can fairly approximate the behavior of the real, finite rubber with the infinite half-space at a depth corresponding to the real elastomer thickness.

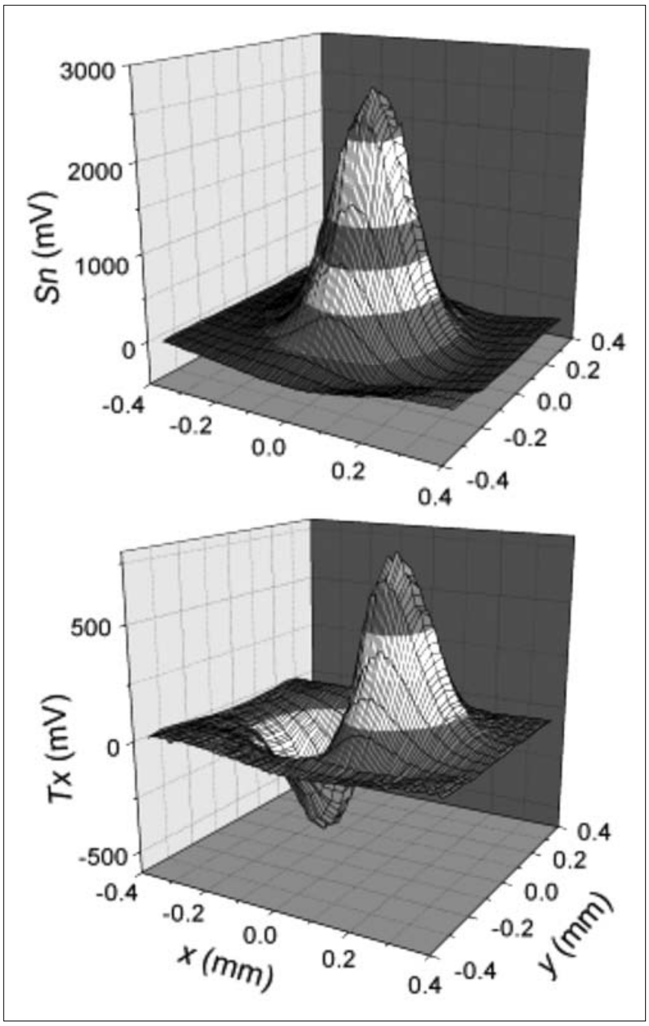
The first task is to solve the equilibrium equations of the rubber for a given surface-indentation profile, and to find the stress/strain/deformation distribution at that specific depth *(Fig. 6)*. A much more important practical task is the solution of the inverse problem, i.e. the reconstruction of the surface indentation profile from discrete number of strain measurements under the rubber.

Although first solutions to the direct problem of the elastic half-space were elaborated long time ago, towards the end of the 19th century; by that time the elastic theory had nothing to do with tactile sensors. It was only in the mid-eighties of the last century when the model became the primary mathematical description of the skin and the artificial cover of pressure sensors. With the appearance of three degree-of-freedom tactile sensors, the theory called for enhancements again.

One of our results is that we changed the flat surface of the cover to a certain, defined shape, mimicking the human skin with finger ridges or other sophisticated tactile organs developed by the evolution *(Fig. 7)*.

Consequently, the half-space model could not be used any more in the original form. Therefore, as an extension of the elastic half-space, we created a *finite-element model* to describe the mechanical behaviour of the cover.

Figure 6.
Two components of the stress profile inside the elastic cover at a given depth, generated by the simplest point load on the surface. The measurement results correlate well with the theoretical predictions.



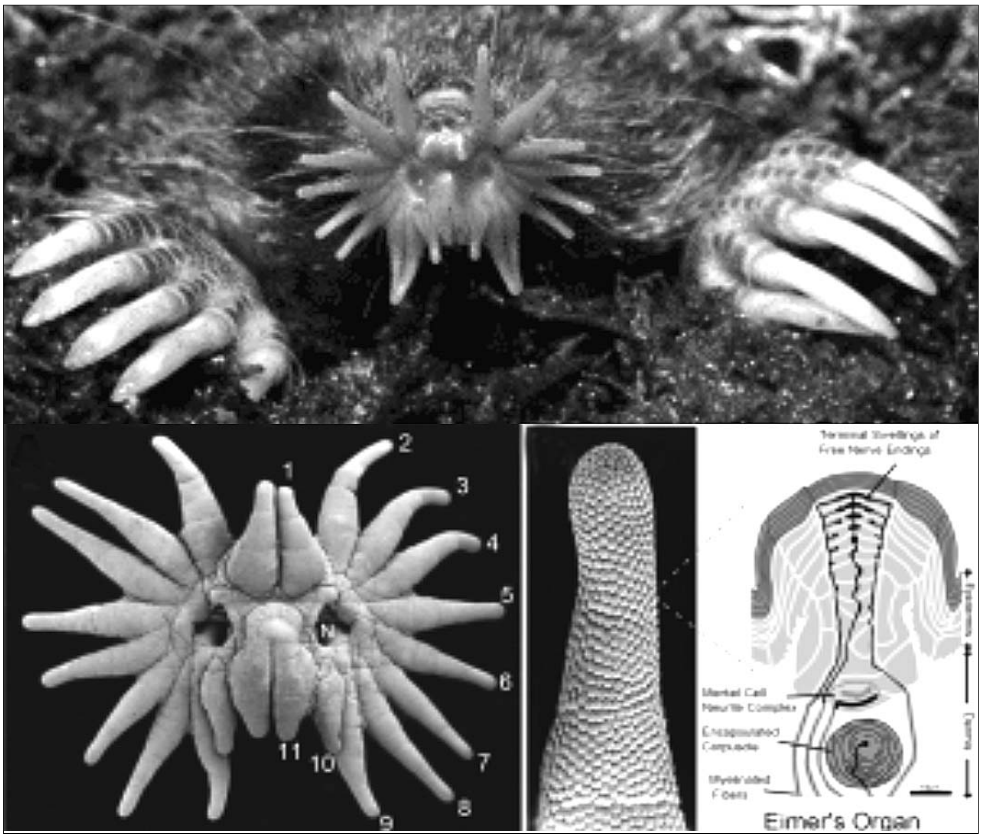

1
2
3
4
5
6
7
8
9
10
11
N
Terminal Swellings of Free Nerve Endings
Merkel Cell Neurite Complex
Encapsulated Corpuscle
Myelinated Fibers
Eimer's Organ

Figure 7.
The world's most developed tactile organ of the star-nosed mole (Condylura cristata), consisting of elastic hemispheres for mechanically filtering and amplifying the tactile signals.

The neuromorphic cover „stolen" from the star-nosed mole basically consists of elastic hemispheres over the cover surface with the following intriguing properties:

– The hemispheres convert the spatially-continuous input force distribution into a discrete one by localizing the forces to their tip.

Figure 8.
Finite-element model of the elastic hemispheres under loads with different directions (above).
The linear and independent stress components arising at the location of the sensors (below).

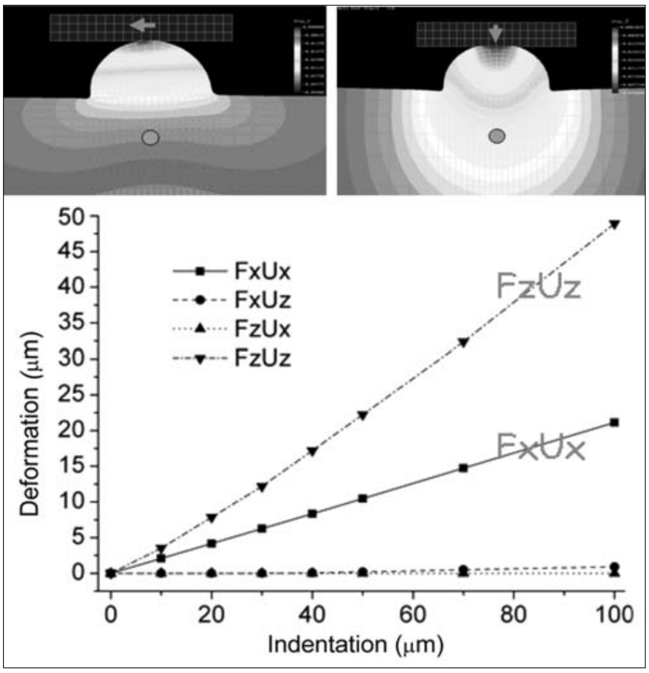
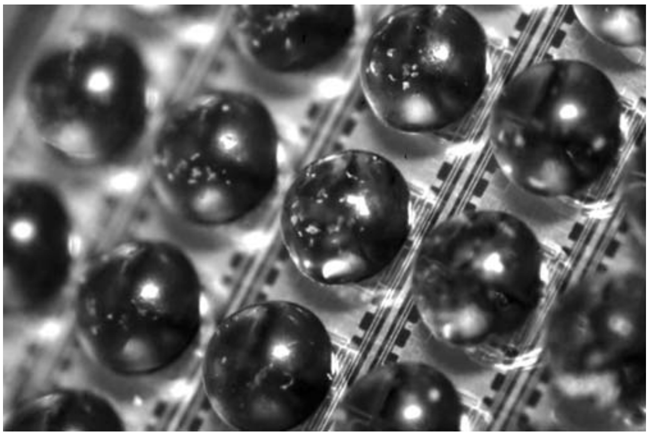


Figure 9.
8x8 sized sensor array with elastic silicone hemispheres on top

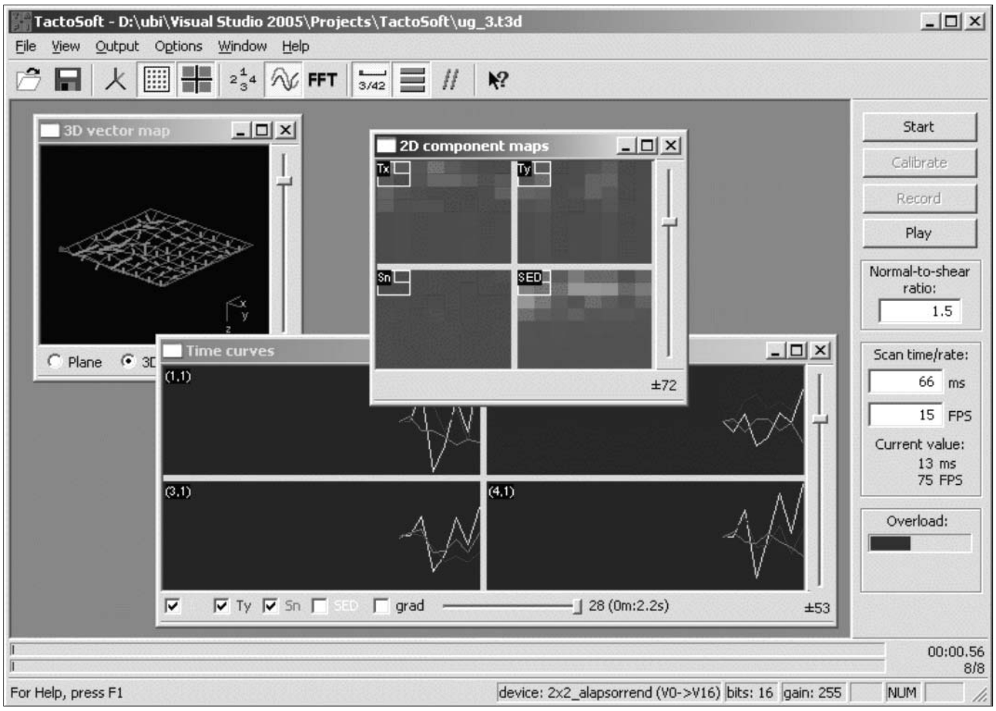


– The hemispheric structure modifies the information coding behaviour of the whole cover in such a way that the three components of the local input forces can be measured linearly and independently with sensors located under the structure *(Fig. 8)*.

We equipped the sensor arrays with geometric elastic covers designed according to the finite-element simulations *(Fig. 9)*, and improved thereby the shear sensitivity of the system. We could also verify experimentally the role of these geometric mechanical structures in biological systems.

5. The tactile system

The signals are pre-processed and transferred to a PC by a read-out circuitry. This read-out board filters and amplifies the analog signals and also calibrates the sen-



Normal-to-shear ratio: 1.5

Scan time/rate: 66 ms 15 FPS

Current value: 13 ms 75 FPS

device: 2x2_alapsorrend (V0->V16) bits: 16 gain: 255

Figure 10.
The graphical user interface of the tactile software

sor array. After A/D conversion the signals are sent to the PC through RS232 or USB communication line.

The data are stored, processed (on-line or off-line) by PCs running under WinXP, by a software developed for analyzing tactile events *(Fig. 10)*.

6. Proactive-adaptive robotic arm control-slippage detection

In order to illustrate the use of the three axial tactile sensors, an experimental system was constructed. The main component of the system is a two fingered robotic arm that can hold small and medium sized objects *(Fig. 11)*.

For handling fragile or slippery objects with unknown parameters, a continuous tactile feedback is indispensable in order to prevent slippage. Let us consider the case when a two-fingered robotic arm holds an empty glass and we start to fill it with water: the glass changes its weight with time. In that case the grasping force has to be adapted, too, proportionally to the increasing weight of the object. If the holding force is too small, the object can slip out of the fingers. This must be detected by the system in due time to give an adequate response and prevent slippage, namely, to increase the grasping forces.

A great advantage of three-axial tactile sensors is the capability of measuring shear forces. Thereby knowing the slippage threshold, the robot's control can be alerted before the actual slip, preventing any relative motion between the object and the robot fingers *(Fig. 12)*.

Figure 11.
The two-fingered Katana robot grasping an object between the fingertips equipped with tactile sensors

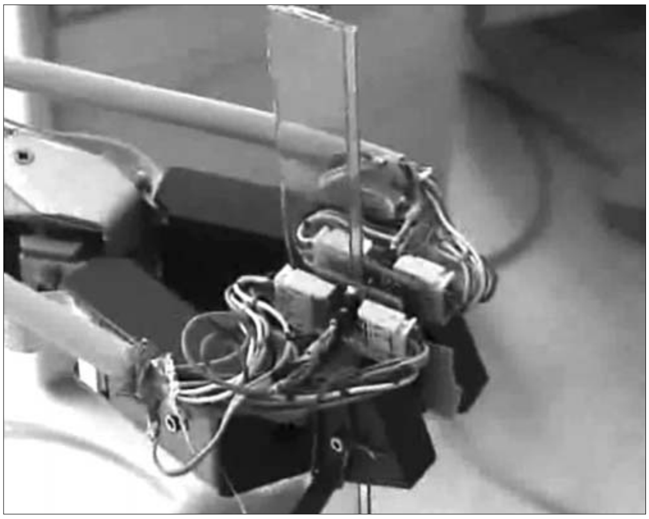
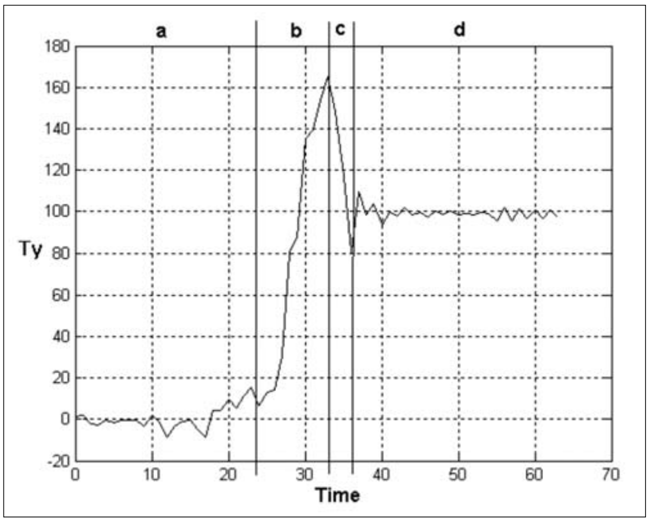


Figure 12.
The time evolution of 3D grasping through tactile forces: a) nothing is grasped; b) shear force increasing proportionally to the growing weight; c) object starts to slip out, force decreases; d) constant motion with constant force determined by the kinetic friction coefficient.



6. Summary, applications

An introductory overview to the design and processing issues of integrated tactile sensor arrays was presented. The main achievement of the novel system is the capability of 3D resolution of the attacking force in every taxel. Since shear forces appear in every grasping task, the capability of three-axial measurements is a must in tactile sensing.

In order to exploit the results presented here in a nutshell, and to expand our capabilities, MFA, PPKE and RG Co. established a spin-off company, TactoLogic Ltd., Budapest. Besides commercializing complex systems dedicated for education and research purposes, the company is going to introduce the tactile devices in medical application. Endoscopes, catheters, autonomous microrobots equipped with three-axial taxels could provide tactile information from remote places, where human hand could never touch before. Moreover, accurate physical diagnostics can also be targeted in long term.

By integrating this system with tactile displays, tactile tele-sensing will also be achievable. In the long run, tactile sensors could be also exploited in any prosthetic device.

Acknowledgements

This tactile research was initiated by the TELESENSE project (NKFP 035/02/2001) of the Széchenyi Program. Part of the work was supported by OTKA via T047002 project.

References

[1] G. Vásárhelyi, M. Ádám, É. Vázsonyi, Zs. Vízváry, A. Kis, I. Bársony, Cs. Dücső,
"Characterization of an Integrable Single-Crystalline 3D Tactile Sensor,"
IEEE Sensors Journal, August 2006, Vol. 6, No. 4, pp.928–934.

[2] G. Vásárhelyi, M. Ádám, É. Vázsonyi, I. Bársony, Cs. Dücső,
"Effects of the Elastic Cover on Tactile-Sensor Arrays,"
Sens. Actuators A, 2006, Vol. 132, pp.245–251.

[3] G. Vásárhelyi, B. Fodor, T. Roska,
"Tactile Sensing-Processing: Interface Cover Geometry & Inverse Elastic Problem,"
Sens. Actuators A, 2007, Vol. 140, pp.8–18.

[4] Zs. Vízváry, P. Fürjes, M. Ádám, Cs. Dücső, I. Bársony,
"Mechanical Modelling of an Integrable 3D Force Sensor by Silicon Micromachining,"
National Institute for Research and Development in Microtechnologies (Bucharest) (ed.)
Special issue featuring selected papers from the 13th European Micromechanics workshop (MME'02), Bristol, Institute of Physics Publishing, 2003, pp.165–168.

[5] É. Vázsonyi, M. Ádám, Cs. Dücső, Zs. Vízváry, A.L. Tóth, I. Bársony,
"Three-dimensional Force Sensor by Novel Alkaline Etching Technique,"
Sens. Actuators A, Vol. 123-124, No. 23, Sep. 2005, pp.620–626.

[6] M. Ádám, T. Mohácsy, P. Jónás, Cs. Dücső, É. Vázsonyi, I. Bársony,
"CMOS Integrated Tactile Sensor Array by Porous Si Bulk Micromachining",
Sens. Actuators A,
In Press, Corrected Proof available online 6 Aug. 2007.

[7] A. Kis, F. Kovács, P. Szolgay,
"3D Tactile Sensor Array Processed by CNN-UM: A Fast Method for Detecting and Identifying Slippage and Twisting Motion,"
Int. Journal on Circuit Theory and Application (CTA), 2006, No. 34, pp.517–531.

Authors

Gábor Vásárhelyi received his M.Sc. degree in engineering-physics from the Technical University of Budapest, Hungary, in 2003, and his Ph.D. degree in technical sciences (infobionics) from the Péter Pázmány Catholic University, Faculty of Information Technology, Hungary, in 2007. He is currently a research fellow at the Institute for Technical Physics and Materials Sciences, and the chief technology officer of TactoLogic Ltd. His main research fields are connected to artificial tactile sensory systems.

Mária Ádám received the MSc degree in electrical engineering from the Technical University of Budapest, in 1973. She was R&D engineer at TUNGSRAM Ltd., at the Department of Semiconductors, from 1973 to 1982. From 1982-1985 she spent three years with the Microelectronics Co. Currently she is a research engineer at the MEMS Laboratory of the Research Institute for Technical Physics and Materials Science (MFA). Her research interests include design, development and processing of silicon-based mechanical and gas microsensor structures. She is co-author of about 25 scientific papers and one patent.

Csaba Dücső was born in Hungary, in 1959. He was graduated as a chemist in 1983 and received the Ph.D. degree in chemistry from the Eötvös Lóránd University, Budapest, Hungary. He was the head of the MEMS laboratory of MFA, Budapest between 1992 and 2007. His research interests include Si based MEMS with special emphasis on integrated gas and mechanical sensors as well as the development of related processing technology. At present he is in charge of R&D of thin film solar cells at BudaSolar Technologies Ltd., Budapest. He is co-author of over 70 papers published in periodicals or conference proceedings and holds 4 patents pending.

István Bársony graduated in electrical engineering from the Technical University of Ilmenau, Germany in 1971. He holds a C.Sc from the Hungarian Academy of Sciences (1978) and a PhD from the Technical University of Budapest (1996) and a D.Sc from the HAS (2001). During his professional career he was working on research assignments in Hungary mainly in silicon technology, in Japan (1983–1986) on imaging application of the Static Induction Transistor, in the Netherlands (1988–1993) at the MESA Research Institute of the University of Twente on Rapid Thermal Multi-Processing. Since 1993 he has been with the Research Institute for Technical Physics and Materials Science of the Hungarian Academy of Sciences, MFA Budapest, from 2004. He is the director of MFA. He has led several international projects a/o on solar cell and microsystems research. He holds 12 patents, published over 80 scientific papers and is a professor of nanotechnology at the University of Veszprém.

Attila Kis received his MSc degree in electrical engineering from the Technical University of Tirgu Mures, Romania, in 2001. From 2001 to 2002 he was co-researcher at the Computer and Automation Research Institute of the Hungarian Academy of Sciences, Analogical and Neural Computing Systems Laboratory. He received his PhD degree at Péter Pázmány Catholic University, Multidisciplinary Technical Sciences Doctoral School, Budapest, Hungary, in 2007. His research interests are biologically inspired real-time information processing on tactile sensor arrays, wave computing, CNN-UM, autonomous robots, integrated tactile and vision systems and smart sensors.

Makyoh topography:
a simple and powerful method for the flatness characterisation of semiconductor wafers

FERENC RIESZ, ISTVÁN ENDRE LUKÁCS, JÁNOS SZABÓ, JÁNOS P. MAKAI, ISTVÁN RÉTI, BÉLA SZENTPÁLI, IMRE EÖRDÖGH, BÁLINT PŐDÖR*, ZSOLT JOHN LACZIK#

Research Institute for Technical Physics and Materials Science, Hungarian Academy of Sciences
riesz@mfa.kfki.hu

**Budapest Tech, Kandó Kálmán Faculty of Electrical Engineering*

#Department of Engineering Science, University of Oxford

Reviewed

Keywords: semiconductor technology, surface flatness, optical metrology

The paper presents our research in the field of Makyoh topography, a method based on an ancient principle. The method's application is the qualitative and quantitative study of semiconductor wafers and other mirror-like surfaces.

1. Introduction

Semiconductor technology requires perfectly flat and defect-free single crystal wafers as starting material. Any deviation from the ideal flatness can hinder processing steps or degrade process quality. Therefore, a high interest exists both from wafer manufacturers and users towards contactless, highly accurate, clean and fast characterisation tools which can be used to screen or characterise the wafers with regard to geometrical or topographical defects.

Numerous methods exist for flatness characterisation. The surface topography can be measured with high accuracy using surface stylus methods, but they are slow, require mechanical movement and the stylus can scratch the surface. The requirements of non-contact operation are fulfilled by the optical methods [1], such as laser scanning or other, mainly interferometric, techniques. However, the realisation of these methods for large-size samples is difficult.

As an adaptation of the ancient Japanese magic mirror [2], a new alternative tool, *Makyoh topography* appeared in the late 1970s [3,4] (Makyoh means 'magic mirror' in Japanese). The principle of the method is the following: the surface under study is illuminated by a homogeneous, collimated light beam, and the reflected beam is intersected by a screen placed some distance away. Because of the surface's microdeformations, a non-uniform intensity distribution characteristic to the surface topography appears on the screen *(Fig. 1)*.

In practice, optical set-ups containing a CCD camera and other optical elements, equivalent to the original one, are used. A number of researchers and manufacturers were inspired to apply the technique by its exceptional simplicity. The method has been applied more widely from the '90s mainly for wafer screening and the assessment of lapping-polishing technology [4-6]. The construction of the set-up is simple, the scaling for large-area studies is straightforward, and the method is able to detect surface defects in real time and with high sensitivity. However, a drawback of the technique in its original form was its inability to perform quantitative studies.

The aim of the present paper is a short, concise description of the basic and applied research in the field of Makyoh topography conducted at MFKI (Műszaki Fizikai Kutatóintézet) and at the legal successor MFA.

2. Initial steps

Research on the method commenced at MFKI in the beginning of the 1990s. *Fig. 2.* shows the scheme of the first set-up constructed at MFKI [7]. The light source is a pigtailed LED (wavelength, 820 nm) providing a 50-μm-diameter light spot. A 500-mm focal lenght lens, positioned above the sample, serves as a collimator for the light source and as a 'magnifier' for the camera. The

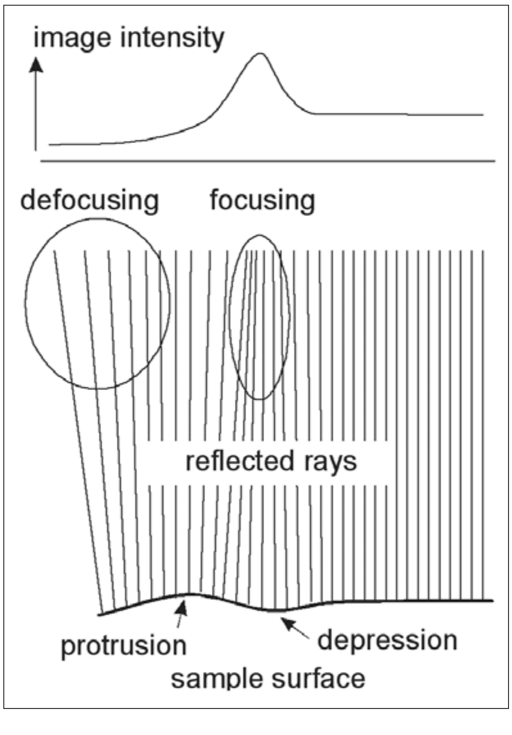


Figure 1. Scheme of Makyoh-topography imaging

maximum sample diameter is limited to 75 mm by the lens aperture. The images obtained by the camera were recorded and stored on a video tape recorder.

The research and manufacturing of GaAs devices at the Department of Microwave Devices of MFKI as well as our international contacts provided a large number of samples for characterisation [8,9]. These studies – in accordance with the state of the art at the time – remained within the framework of qualitative interpretation. Some characteristic examples are shown in *Fig. 3.* The pattern of closely parallel arcs indicates saw marks, the periodic, dark spots indicate a rough surface morphology, while large-size features with the slowly varying contrast are due to large-scale wafer deformation.

3. Fundamentals of image formation

As measurements providing numerical results are basic requirement of modern technology assessment, our further research aimed at the understanding of the image formation of Makyoh topography. Since for Makyoh topography set-ups an optically equivalent system can be found, which, disregarding a magnification factor, consists only of a collimated illumination and a distant screen, the imaging can be characterised by a single parameter, the equivalent screen-sample distance (L) [10,11].

The geometrical optics model of Makyoh imaging was described in Ref. [12] in detail. Here we present only the final result. The screen position $\mathbf{f}(\mathbf{r})$ of a light beam reflected from a given $\mathbf{r}$ point of the surface is given by the following formula (for small incidence angles, that is, for a relatively smooth surface):

$$\mathbf{f}(\mathbf{r}) = \mathbf{r} - 2L\,\mathrm{grad}\,h(\mathbf{r}). \qquad (1)$$

This equation follows trivially from the law of reflection: the shift of the reflected beam's position relative to a flat sample surface is proportional to the surface gradient at the given point. The $I(\mathbf{f})$ intensity of the $\mathbf{f}(\mathbf{r})$ point (relative to that of a flat surface with unity reflectivity) is described by the following formula:

$$I(\mathbf{f}) = \frac{\rho(\mathbf{r})}{\left|(1-2LC_{\min})(1-2LC_{\max})\right|}, \qquad (2)$$

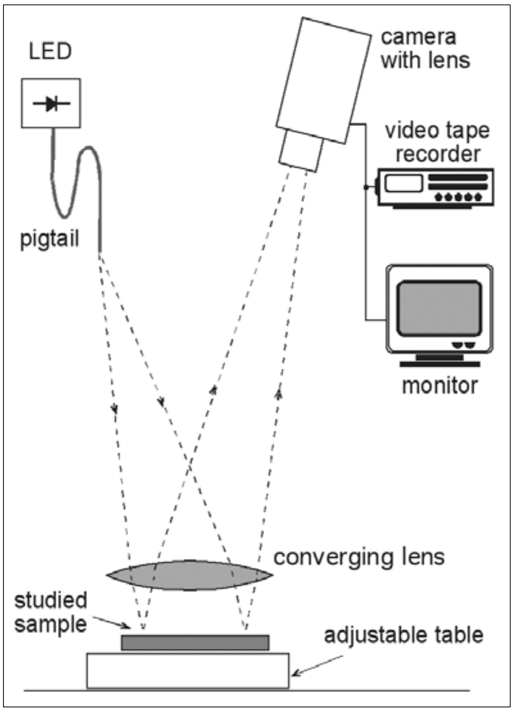


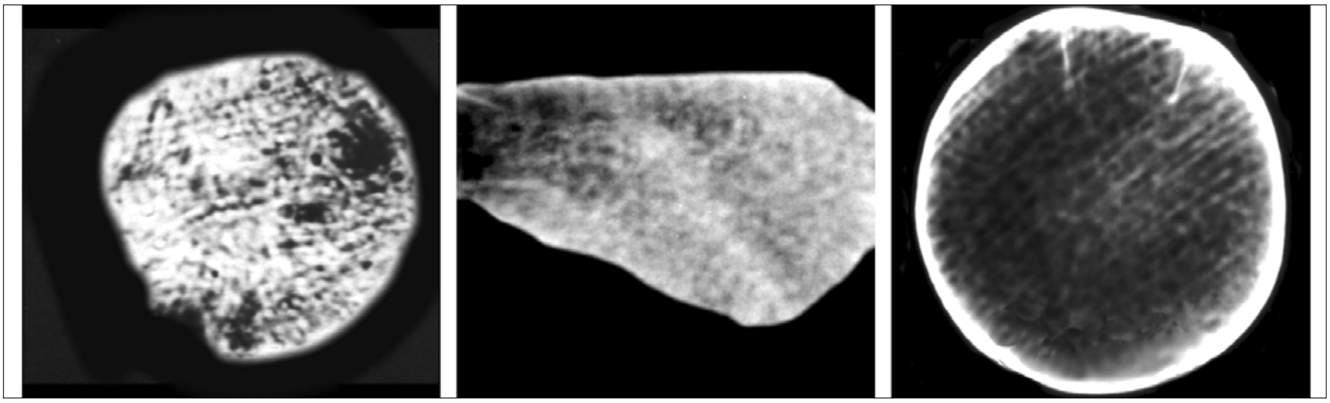
Figure 2. The first Makyoh-topography set-up built at MFKI

where $\rho(\mathbf{r})$ is the local reflectivity of the surface, and $C_{\min}$ and $C_{\max}$ are the local maximum and minimum curvatures of the surface. That is, the intensity of the reflected beam is determined by the second-order properties of the surface.

It follows from the above equation, that at small $|L|$ values, a given surface point and its image are close to each other (assuming the sample and screen are effectively in the same plane), and the main component of the image contrast is resulted from the inhomogeneities of the surface reflectivity. Increasing $|L|$ increases the contrast and the distance between the point and its image, suppressing the component due to reflectivity variations. The optimum setting is therefore in that medium region of $|L|$ which gives high enough contrast for the safe observation of the image while retaining the integrity of the surface topology in the image.

Although the geometrial optics description is approximate, is gives satisfactory description for most cases. Significant diffraction effects are encountered only if the sample has sharp surface topography features, e.g. at

Figure 3. Typical Makyoh images of semiconductor wafers



the edges of openings. The geometrical optics approach may also not be sufficiently accurate if many beams meet at a certain point in the image. In practice however, the wafers studied using the technique have a uniform reflectivity, their surface topography is mainly smooth and varies slowly, and the optimum imaging regime is just when focussing effects are not encountered.

4. Quantitative Makyoh topography

Although the imaging laws of Makyoh topography are simple, the equations describing the imaging are not invertable, thus the analytic determination of the surface topography from the Makyoh image is not possible in the general case [13]. However, if the homogenous illumination is structured by some mask, certain points of the surface are 'labelled'. Thus, based on Equation (1), the surface gradient can be determined at the labelled points if the positions pertaining to the ideal flat surface are known. Equation (2) thus becomes superfluous. The most expedient way of structuring is a square grid. The $h(x, y)$ height topography in the grid points can then be approximated by the following sum [13]:

$$h(x,y) = \frac{1}{2L}\sum_i \left[\Delta x(x_i - f_{xi}) + \Delta y(y_i - f_{yi})\right] \quad (3)$$

Here Δx and Δy are the grid cell sizes, and (f_x, f_y) are the co-ordinates of the (x, y) grid point; (x_i, y_i) denotes the co-ordinates pertaining to the ideal flat surface, which can be determined by the measurement of a flat reference mirror. The summing starts at a point whose height can be chosen arbitrarily. In principle, the summation path can be chosen arbitrarily since all paths with the same starting and end points should give the same sum. In practice, however, because of the finite resolution of the grid, the error of the integral sum depends on the path and its value is not predictable. The accuracy of the method can significantly be increased if the sums of all paths (or, more precisely, the paths contained in a rectangle spanned by the starting and end points) are calculated and averaged. This procedure, however, can take a long time even for a small grid. We therefore developed a recursive algorithm which gives the same result but is much faster [14,15].

We have also developed an algorithm for the location of the grid points. The algorithm runs a cross-like weight function over the Makyoh image and their correlation is determined. By finding the local maxima of the correlation function, the coordinates of the grid points can be determined with subpixel accuracy.

The described method can easily be automated, it allows simple and fast (quasi real time for a 50x50 grid) quantitative studies. It is important to note that, provided that the grid lines are significantly thinner than the grid pitch, the Makyoh image remains visible showing the contrast caused by small-size surface defects. This property corresponds to the requirements of the semiconductor industry, since the topography of the wafers, in general, is a result of the superposition of a slowly varying deformation (curvature, warp) and localised defects (polishing, lapping marks etc.). It is advantageous if the image of the grid is nearly focussed. This can be achieved by the arrangement described in Section 2.

The path-dependent error component of the integral sum can be eliminated by an iterative (so called relaxation) procedure [15,16]. This method is more accurate than the direct integration, but it is slower, rendering it unsuitable for real-time measurements.

5. Applications

5.1. Studies of deformations induced by wafer reclaim

Large-diameter semiconductor wafers are expensive, and wafers that were rejected by one the technology line may still be suitable for certain other purposes. Wafer reclaim is therefore a dynamically growing branch of semiconductor industry. The same considerations apply for the novel, costly compound semiconductor materials, such as SiC. Wafer reclaim includes the removal of the device layers and subsequent polishing of the wafer. A model experiment was carried out in our institute [17] with an aim to study the deformations induced by the wafer reclaim steps and to explore their possible causes. In the course of this experiment, the deformations of two-inch-diameter processed Si wafers were examined after each layer removal.

We have shown that the removal of the functional layers (oxide or metallisation) induces a uniform change of the wafer curvature, while the final polishing step causes a non-uniform deformation, which depends on the amount of the original deformation and the parameters of the polishing process. Comparing the samples after final polishing, we established that the originally flat or uniformly curved wafers remained flat or uniformly curved upon polishing. Our interpretation is that upon polishing, the attachment of an originally curved wafer to the supporting plate makes it flat, and, upon releasing it after polishing, it re-acquires its original shape. In contrast, the topography of the wafers having irregular topography before polishing changed upon polishing showing no obvious correlation with the original shape. These deformations are presumably related to imperfections of the polishing process.

5.2. Studies of the deformation of MEMS elements

Although the chief application of Makyoh topography is the study of the large-area surfaces, and the strongly structured MEMS samples cause strong diffraction patterns, in simple cases, the method can still be used efficiently. We studied the deformation of 4-10 mm side length Si/SiN$_x$ square membranes [18]. We compared the measured deflection of the centre point of the prepared membranes with the deflection values obtained by finite element thermo-mechanical simulations in order to determine the thermal expansion coefficient of SiN$_x$. We obtained good agreement if we set 2.62x

$10^{-6}K^{-1}$ for the thermal expansion coefficient of SiN_x. We emphasize that the height of the membranes' centre points were below 0.1 μm, which shows the high sensitivity of the method.

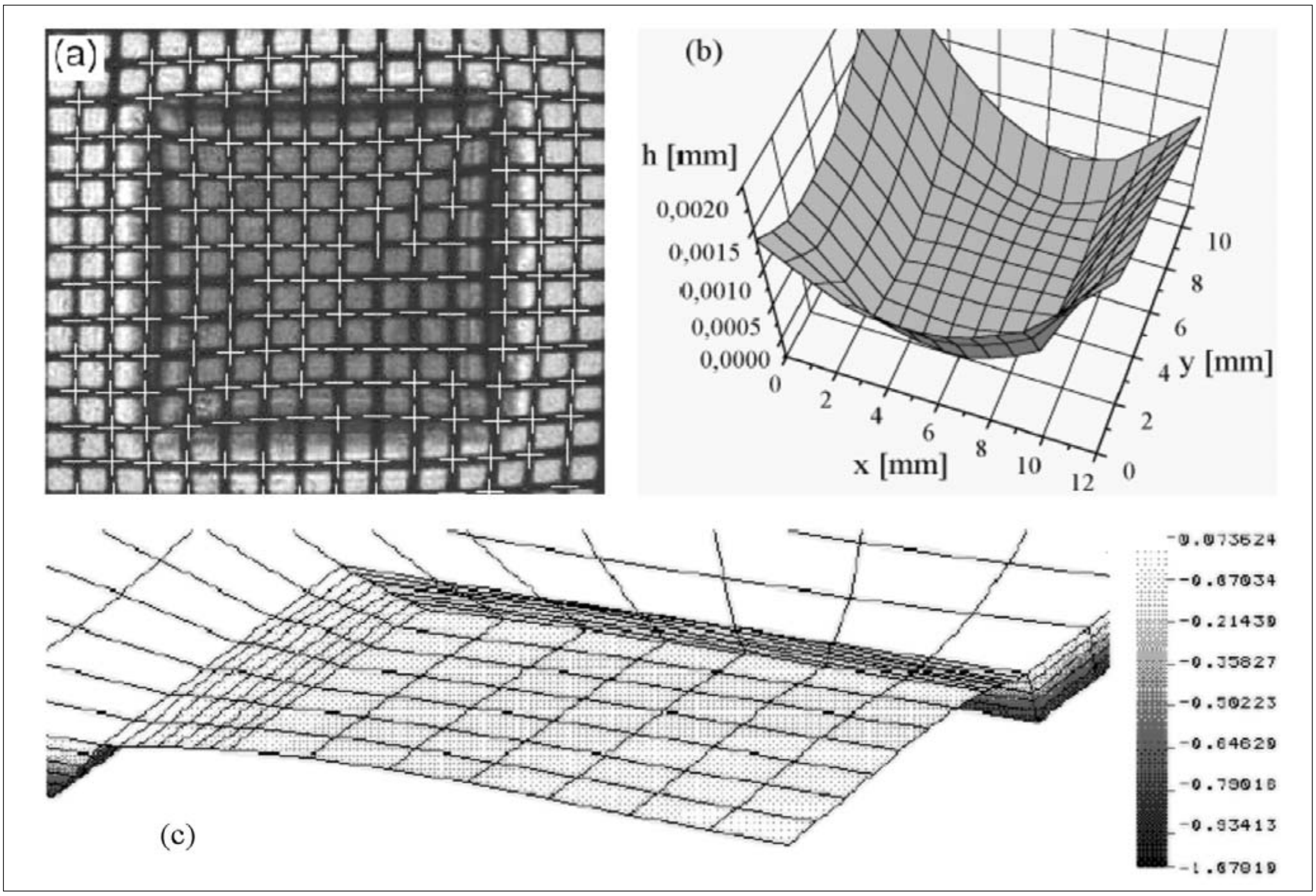
Fig. 4. shows a Makyoh image and the corresponding calculated height map as well as a characteristic simulation result. According to the measured Makyoh topogram, a shallow (≈ 0.01 μm deep) depression is situated in the centre of the otherwise convex membrane, well reproduced also by the simulations. The deformation of the substrate around the membrane area is visible on the Makyoh topogram as well as on the simulated topography.

6. Novel experimental set-ups: a mirror-based arrangement and the Digital Micromirror Device

The greatest disadvantage of the arrangement described in Section 2 is the inability of studying large-diameter samples, because a large-diameter collimator/magnifier lens without significant aberrations it is extremely costly. To circumvent this problem, we constructed a mirror-based system *(Fig. 5)* [19-22]. The parabolic mirror applied in off-axis arrangement eliminates spherical aberration, and the beam splitter makes the imaging free of parallax errors. The diameter of the parabolic mirror is 300 mm, its focal length is 1524 mm. The value of *L* can be varied between approx. 0 and 5500 mm by changing the distance setting of the camera lens. With this equipment we have a modern, sensitive, high dynamic range, widely applicable tool. The greatest advantage of the device is its scalability: off-axis parabolic mirrors of 450 mm diameter and λ/20 surface quality are available on the market. As an alternative arrangement, a set-up employing separate spherical mirrors in the illuminating and detecting paths was built and its operation was demonstrated [20]. The advantage of this arrangement over the one based on the parabolic mirror is its smaller cost.

The greatest disadvantage of the grid version is the bad lateral resolution: a grid must be sparse enough in order to allow the safe detection of the grid points. The lateral resolution can be increased by applying a shifted grid and sequential image recording; the grid is shifted by a fraction of the grid period between each exposure, thus we obtain a "supergrid" with a period equalling the shift distance. Real-time measurements cannot be realised. The most expedient way of the realisation of this concept is a programmable mirror matrix (Digital Micromirror Device, DMD). The DMD consists of a matrix array of individually addressable micromirrors that can be tilted around their diagonal. (Such devices are used e.g. in DLP projectors.)

Figure 4.
The Makyoh image of a 10x10 mm SiN_x membrane (a) with the localised grid points, (b) the calculated height profile and (c) the profile as simulated by the finite element method (height data are in μm)



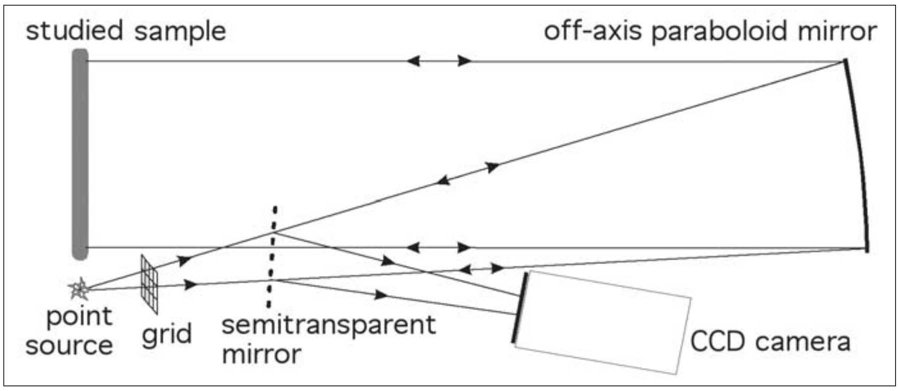


*Figure 5.
The scheme of a mirror-based Makyoh-topography set-up*

Because the pattern is finer than that of the traditional fixed grid, it is more important to ensure that the image of the mask be sharp on the Makyoh image. Because it is more difficult to achieve this with the mirror-based system, we positioned the DMD in a telescopic system consisting of two converging lens. The first (lens-based) version of the set-up has been built in the University of Oxford [16]; 0.7 mm lateral resolution was achieved, and a maximum 10% difference projected to the total height span of 7 µm was demonstrated comparing the results to interferometry.

The application of DMDs opens new perspectives in Makyoh topography [21]. In addition to the described shifted-grid set-up, a grid with any period (or even any arbitrary pattern) can be realised, thus the trade-off between measurement speed and lateral resolution can be optimised.

7. Summary

The research we described contributed to the basic understanding of an already known research tool: the “magic” phenomenon became an understood, widely applicable method that knocks on the door of industrial applications. Regarding further applications and other aspects we kindly refer the reader to the literature [23-25] and to the home page of the research:
http://www.mfa.kfki.hu/~riesz/makyoh/.

Acknowledgements

The authors gratefully acknowledge the contributions of their colleagues: István Bársony, János Chalupa, Csaba Dücső, Magdolna Erős, Péter Fürjes, János Gubics, József Karányi, Margit Németh-Sallay, Károly Somogyi, Károly Szász, Erika Tunyogi and Zsolt Vízváry.
The work was partly supported by the Hungarian Scientific Research Fund (OTKA, Grants F 25224, T 037711, K 68534 and M 041735), Bolyai János Postdoctoral Research Fellowship (F. Riesz), KPI (contract No. GVOP-3.2.1.-2004-04-0337/3.0.), European Union (projects MEMSWAVE, INCO-Copernicus [cont. No. 977131], ANNA [cont. 026134 (RII3)], HYPHEN [cont. FP6-027455]) and The Royal Society of London.

References

[1] Muller T., Kumpe R., Gerber H.A., Schmolke R., Passek F., Wagner P.:
Techniques for analysing nanotopography on polished silicon wafers,
Microel. Engin. 56. (2001), p.23.

[2] Riesz F.:
A 2000 years old principle in high technology – the Japanese magic mirror,
Élet és Tudomány 55 (2000), p.41. (in Hungarian)

[3] Kugimiya K:
Makyoh: The 2000 year old technology still alive,
Journal Crystal Growth 103. (1990), p.420.

[4] Blaustein P., Hahn S.:
Realtime inspection of wafer surfaces,
Solid State Technol. 32. (1989), p.27.

[5] Tokura S., Fujino N., Ninomiya M., Masuda K.:
Characterization of mirror-polished silicon wafers by Makyoh method,
Journal Crystal Growth 103, (1990), p.437.

[6] Pei Z.J., Xin X.J., Liu W.:
Finite element analysis for grinding of wire-sawn silicon wafers: a designed experiment,
Int. Journal Machine Tools Manufact. 43 (2003), p.7.

[7] Szabó J., Makai J.:
Investigation of mirror-like surfaces using the Makyoh method,
Elektronikai technológia, mikrotechnika 32 (1993), p.15. (in Hungarian).

[8] Németh-Sallay M., Minchev G.M., Pődör B., Pramatarova L.D., Szabó J., Szentpáli B.:
Investigation of the surface preparation of GaAs substrates for MBE and VPE with whole sample optical reflection,
Journal Cryst. Growth 126. (1993), p.70.

[9] Minchev G.M., Pramatarova L.D., Pődör B., Szabó J.:
Experimental confirmation of the peculiar behavior of the coherent-type twin boundaries in sphalerite crystals,
Crystal Research and Technology 29. (1994), p.1131.

[10] Riesz F.:
Camera length and field of view in Makyoh-topography instruments,
Review of Scientific Instruments 72. (2001), p.1591.

[11] Szabó J., Riesz F., Szentpáli B.:
Makyoh topography: curvature measurements and implications for the image formation,
Jpn. Journal Applied Physics 35. (1996), L258.

[12] Riesz F.:
Geometrical optical model of the image formation in Makyoh (magic-mirror) topography,
Journal Physics D: Applied Physics 33 (2000), p.3033.

[13] Riesz F., Lukács I.E.:
Possibilities of quantitative Makyoh topography,
Proc. of 3rd International EuroConf. Advanced Semiconductor Devices and Microsys.,
16-18 October 2000, Smolenice,
Editors: Osvald J., Hascík S., Kuzmík J., Breza J.,
IEEE, Piscataway (2000), p.215.

[14] Lukács I. E., Riesz F.:
Error analysis of Makyoh-topography surface height profile measurements,
Eur. Phys. Journal – Appl. Phys. 27. (2004), p.385.

[15] Riesz F., Lukács I.E.:
Sensitivity and measurement errors of Makyoh topography,
Physica Status Solidi (A) 202. (2005), p.584.

[16] Lukács I.E., Riesz F., Laczik Z.J.:
High spatial resolution Makyoh topography using shifted grid illumination,
Physica Status Solidi (A) 195. (2003), p.271.

[17] Lukács I.E., Riesz F.:
Makyoh-topography assessment of etch and polish removal of processed circuits for substrate re-use,
Microel. Engin. 65. (2003), p.380.

[18] Lukács I.E., Vízváry Zs., Fürjes P., Riesz F., Dücső Cs., Bársony I.:
Determination of deformation induced by thin film residual stress in structures of millimetre size,
Adv. Eng. Mater. 4. (2002), p.625.

[19] I.E. Lukács, J.P. Makai, F. Riesz, I. Eördögh, B. Szentpáli, I. Bársony, I. Réti, A. Nutsch,
Wafer flatness measurement by Makyoh (magic-mirror) topography for in-line process control,
Proc. 5th European Advanced Equipment Control / Advanced Process Control (AEC/APC) Conference,
14-16 April 2004, Dresden, Germany, p.514.

[20] Makai J.P., Riesz F., Lukács I.E.:
Practical realizations of the Makyoh arrangement for the investigation of large area mirror-like surfaces,
3rd International Conference on Metrology [CD-ROM],
14-16 November 2006, Tel Aviv, Israel.

[21] Riesz F., Lukács I.E., Makai J.P.:
Realisation of quantitative Makyoh topography using a Digital Micromirror Device,
SPIE Europe Optical Metrology, 17-21 June 2007,
Munich; Proc. of SPIE, Vol. 6616., Paper 66160L.

[22] Lukács I.E., Makai J.P., Pfitzner L., Riesz F., Szentpáli B.:
Apparatus and measurement procedure for the fast, quantitative, non-contact topographic investigation of semiconductor wafers and other mirror like surfaces,
European Patent EP 1 434 981 B1, 5 July 2006.
US Patent 7,133,140 B2, 7 November 2006.

[23] Riesz F.:
Makyoh topography for the morphological study of compound semiconductor wafers and structures,
Mater. Science and Engineering B 80 (2001), p.220.

[24] Lukács I.E., Riesz F.:
A simple algorithm for the reconstruction of surface topography from Makyoh-topography images,
Crystal Res. Technol. 36. (2001), p.1059.

[25] Lukács I.E., Fürjes P., Dücső Cs., Riesz F., Bársony I.:
Process monitoring of MEMS technology by Makyoh topography,
Proc. of 13th Micromechanics Europe Workshop (MME'2002), 6-8 October 2002,
Sinaia, Romania, p.283.

Authors

Ferenc Riesz PhD, Senior Research Fellow, graduated from the Technical University of Budapest in 1989. Since that, he has been with the Research Institute for Technical Physics and Materials Science (formerly Research Institute for Technical Physics). He obtained his PhD degree in 1994; his thesis was devoted to the structural characterisation of lattice mismatched semiconductor heterostructures. His main research interest is Makyoh topography and related optical metrology topics as well as their application in semiconductor technology and MEMS characterisation. He has published over 60 papers in international scientific journals, three invited contributions and holds an international patent (with co-inventors).

János P. Makai earned his M.Sc.EE. at the Technical University Budapest (BME), 1977. First he joined the Development Division, Works of Mechanical Measuring Equipments Co., (MMG-AM) Budapest then in 1979 the Research Institute for Technical Physics (MTA MFKI) as research fellow and in 1986 he became head of the Optics Department. In 1991 he was invited as guest scientist to the Optical Technology Division of NIST, USA for a three year period. As part time research fellow he worked in 1995-96 for the Coherent Optics Group of the Physics Institute at the BME. In 1997 he earned his Ph.D. in ME from the Department of Optics of the Mechanical Engineering Faculty at the BME. In the same year he left for the Central Bureau of the International Commission on Illumination (CIE) Vienna, Austria and became the Technical Manager till present while he is a part-time researcher at the Research Institute for Technical Physics and Material Sciences (MTA MFA). Field of research: radiometric physics, fiber optic communication, holographic and speckle pattern interferometry, spectrometry, optical methods for material testing. He published more than 70 journal and conference papers and owns several patents.

János Szabó was born in 1920. After receiving his doctoral degree in chemical engineering from the Pázmány Péter University of Sciences, he joined Tungsram Research Labs, where he carried out research on fluorescent light sources. He had pivoting role in the world success of Tungsram products: among others, he had international patents concerning a high-efficiency fluorescent powder and a novel starter for fluorescent lamps. Dr. Szabó retired in 1980. Then, he joined Research Institute for Tehnical Physics, where he was involved in various tasks of semiconductor research. He initiated the research on Makyoh topography. He passed away in 2001.

Béla Szentpáli received the Dipl. Physicist degree from the Eötvös Loránd University, Budapest in 1967, and the CS. Degree in electronics from the Scientific Qualification Board of the Hungarian Academy of Sciences in 1980. After graduation he joined to the Research Institute for Technical Physics, he still works for the successor of that institute. He lead the development and pilot line production of the microwave Schottky-diodes. He dwelt at length on the different applications of microwave techniques and their innovation in the industry. His research interests are sensors, microwave field measurement and low-frequency noise.

Solar Cell Technology Innovation Center at MTA MFA

Ágoston Németh, Zoltán Lábadi, Vilmos Rakovics, István Bársony

Research Institute for Technical Physics and Materials Science, Hungarian Academy of Sciences
nemeth@mfa.kfki.hu

István Krafcsik

Energosolar S.A.

Reviewed

Keywords: ***solar cells, Cu(InGa)Se$_2$, vacuum technologies***

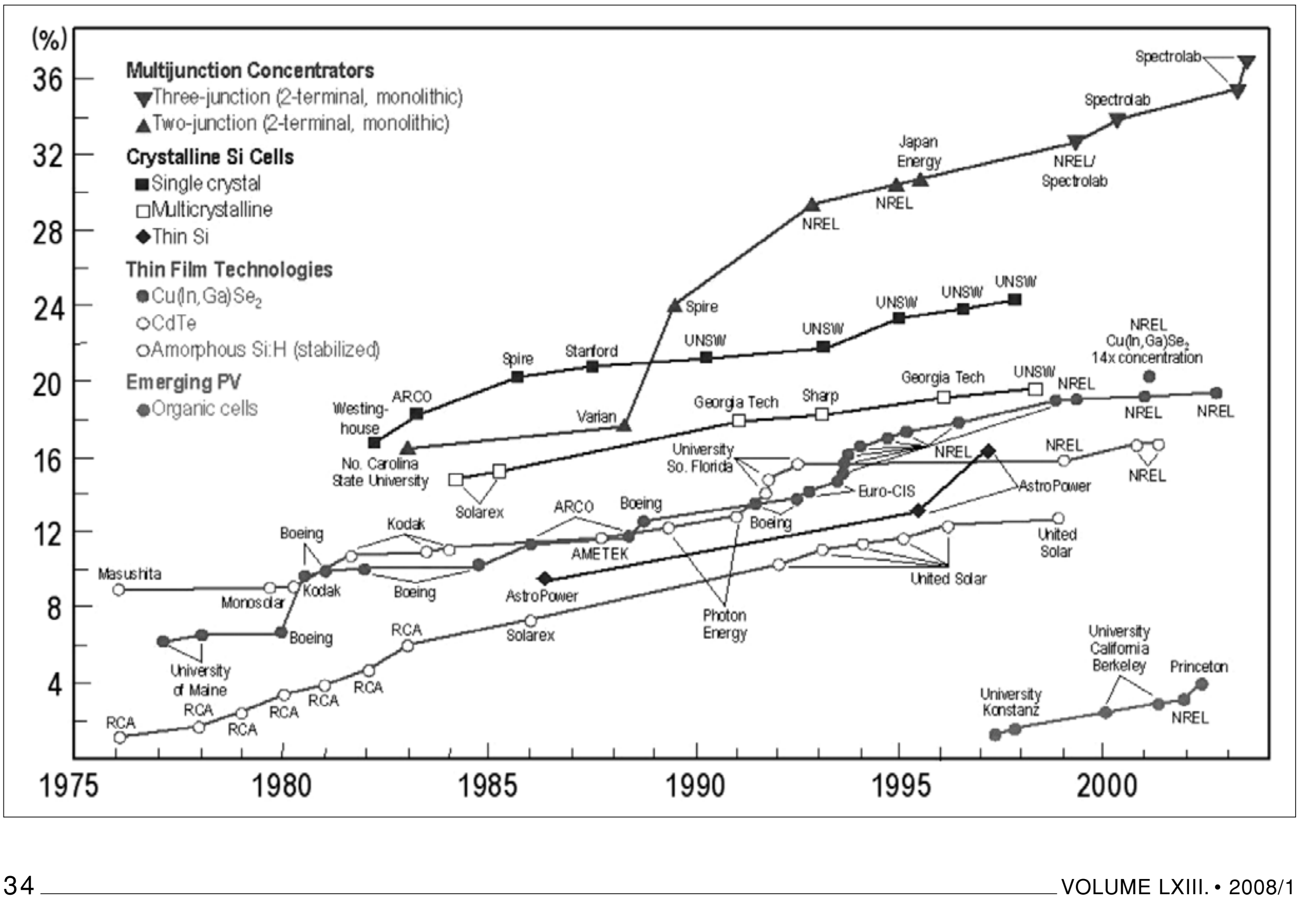
This paper introduces to the reader one of the largest facilities of the solar cell research and development in Hungary – the Solar Cell Technology Innovation Center. The R&D equipment is an integrated vacuum system designed and built for the preparation of thin film Copper Indium Gallium diSelenide (CIGS) solar cell layer structures. The facility was built on the premises of the Hungarian Academy of Sciences by the Energosolar Co. in the frame of a main project funded by the Hungarian National Office for Research and Technology. This paper reviews the layout of the solar cell structure and the equipment for its preparation, introduces the main materials science issues raising in the CIGS system and presenting challenges for the research.

1. Introduction

The worldwide market of renewable energy sources (and especially the photovoltaic cell market) is currently in the phase of dynamic extensive growth. This is due to political factors (increasing concerns about global warming, the Kyoto and Rio protocols) as well as to rapid technological development. The production of photovoltaic (PV) cells and modules increased by 35% over the last decade and reached the 1 GW per year level in 2004. The largest segment of the production is based on crystalline silicon (c-Si) technologies.

In the same time the PV industry has to face the limited feedstock of crystalline and polycrystalline silicon and this problem became a bottleneck for the production. Although silicon is one of the most abundant ele-

Figure 1.
Cell efficiencies of different type solar cells versus production year (Source: www.nrel.gov)



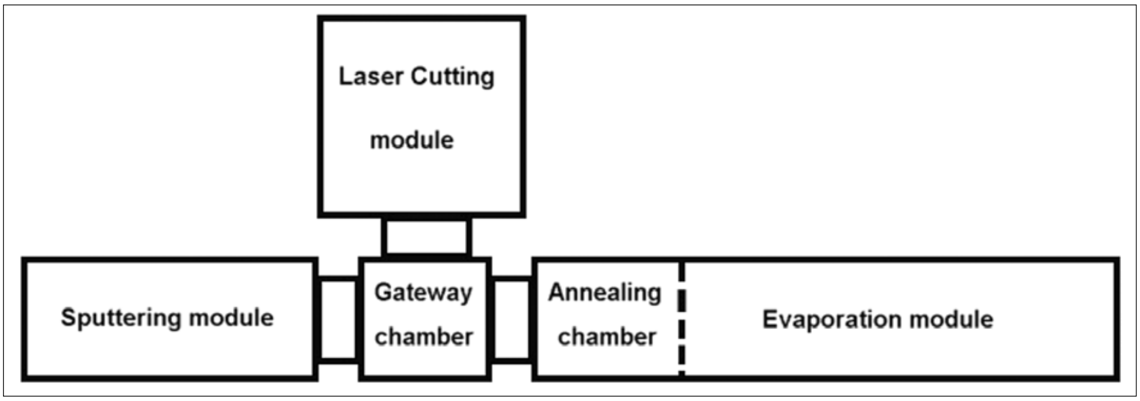


Figure 2. Schematic layout of the integrated vacuum system

ments available in the Earth's crust, the production of solar grade (crystalline) c-Si is an expensive and energy-consuming process. According to reliable market studies this will lead to the saturation of the global solar cell production at the ca. 3-4 GW per year level within the next decade [1-3].

These factors gave impetus to the development of *non-silicon based thin film* solar cells. The most promising alternative to silicon is the copper-indium-gallium-diselenide (CIGS) based thin film PV cell. CIGS already emerged as an ideal choice for PV material in the 80's and the research and development of this material gained momentum in the recent years.

The main advantages of $CuInGaSe_2$ for PV application are the followings:

- a stable chalcopyrite structure,
- p-type conductivity easily achieved by Cu-poor growth processes, and
- very good feasible cell efficiency (current value of laboratory record is 19% while commercially available products have 11%).

This value is very promising compared to the 12,7-13,5% typical efficiencies of c-Si modules, moreover, laboratory research shows the possibility of further improvement. *Fig. 1.* summarizes the trends of cell efficiencies of different solar cell types (based on the data of the US National Renewable Energy Laboratory).

The largest Hungarian R&D project of this promising field started in 2001. The aim of the project was to build an integrated vacuum technology system suitable for the deposition of a CIGS solar cell layer structure, for the development of the complex technology and for the education and training of professionals. The project was financed by the Hungarian National Office for Research and Technology (NKTH). A leading enterprise of the Hungarian vacuum technology industry that time – Kraft Rt. – played the role of initiator in the project. This company was also the first leader of the R&D consortium, but for various reasons handed over this role in 2004 to the Research Institute for Technical Physics and Materials Science of the Hungarian Academy of Sciences (MTA MFA).

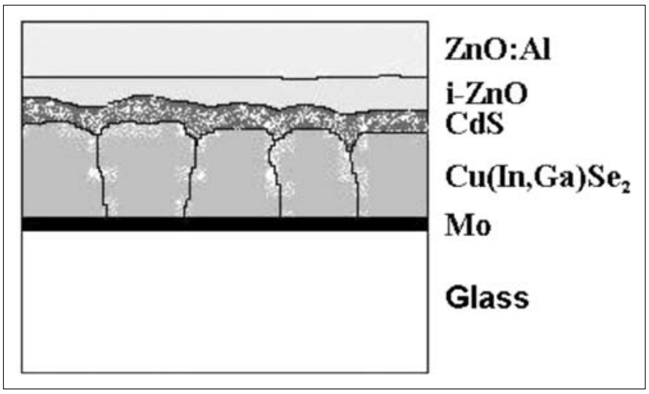


Figure 3. Schematic cross-section of a CuInGaSe2 solar cell structure

Another industrial partner in the consortium was Electrical Drives and Vehicle Electronics Ltd. (VHJ Kft.), while further academic partners were the Institute for Nuclear Research of the HAS (MTA ATOMKI) in Debrecen, Department of Optics and Quantum Electronics of the Science University of Szeged, and the Department of Electron Devices of the Budapest University of Technology and Economics (BME EET).

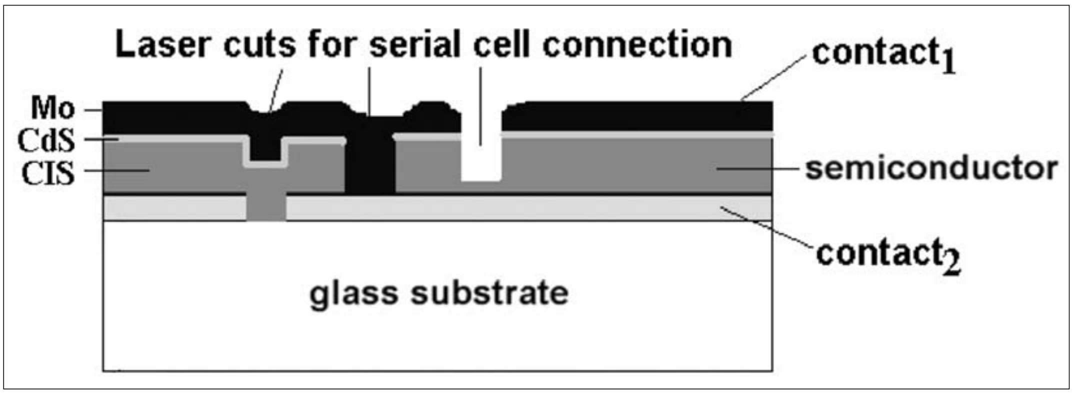


Figure 4. Laser cut in the layer structure and cells connected in series

2. Structure of the integrated vacuum system

In the frame of the project mentioned above an integrated vacuum system was built at the premises of the MTA MFA (completed in 2007) which is suitable for the deposition of CIGS solar cell layer structure on a 30x30 cm^2 glass substrate. The system was designed and built by Energosolar Co. *Fig. 2.* represents the schematic layout of the equipment, while *Fig. 3.* shows the cross-section of the layer structure to be deposited in the system.

In order to form the solar cell structure the CIGS semiconductor layer has to be inserted between two contact layers (in this case between a Mo back contact and a ZnO top window layer) and the whole layer sequence has to be deposited onto the surface of a glass substrate. In order to achieve this goal an integrated vacuum system consisting of four main modules was built:

- Deposition of the contact layers is by magnetron sputtering while the deposition of the CIGS layer is carried out by vacuum evaporation. Therefore the main layer growth units in the system are the sputter- and the evaporation chambers.
- In order to obtain a solar module with the proper terminal voltage the deposited layers have to be segmented and the individual cells have to be connected in series electrically. Therefore, proper grooves have to be formed in every deposited layer according to *Fig. 4.* Formation of these cuts is made by focused laser beam, this technological function is located in the laser cutting chamber (Fig. 2.)
- The fourth processing unit is the gateway chamber situated in the middle of the system. This module ensures the bidirectional movement of the glass substrate between the technological units.

A 10^{-6} mbar end-vacuum can be achieved in the large chambers by using oil diffusion pumps. The chambers are separated by pneumatic latches. The valves, latches and the elements of the transport mechanics are controlled by a purpose-made software integrated into the system.

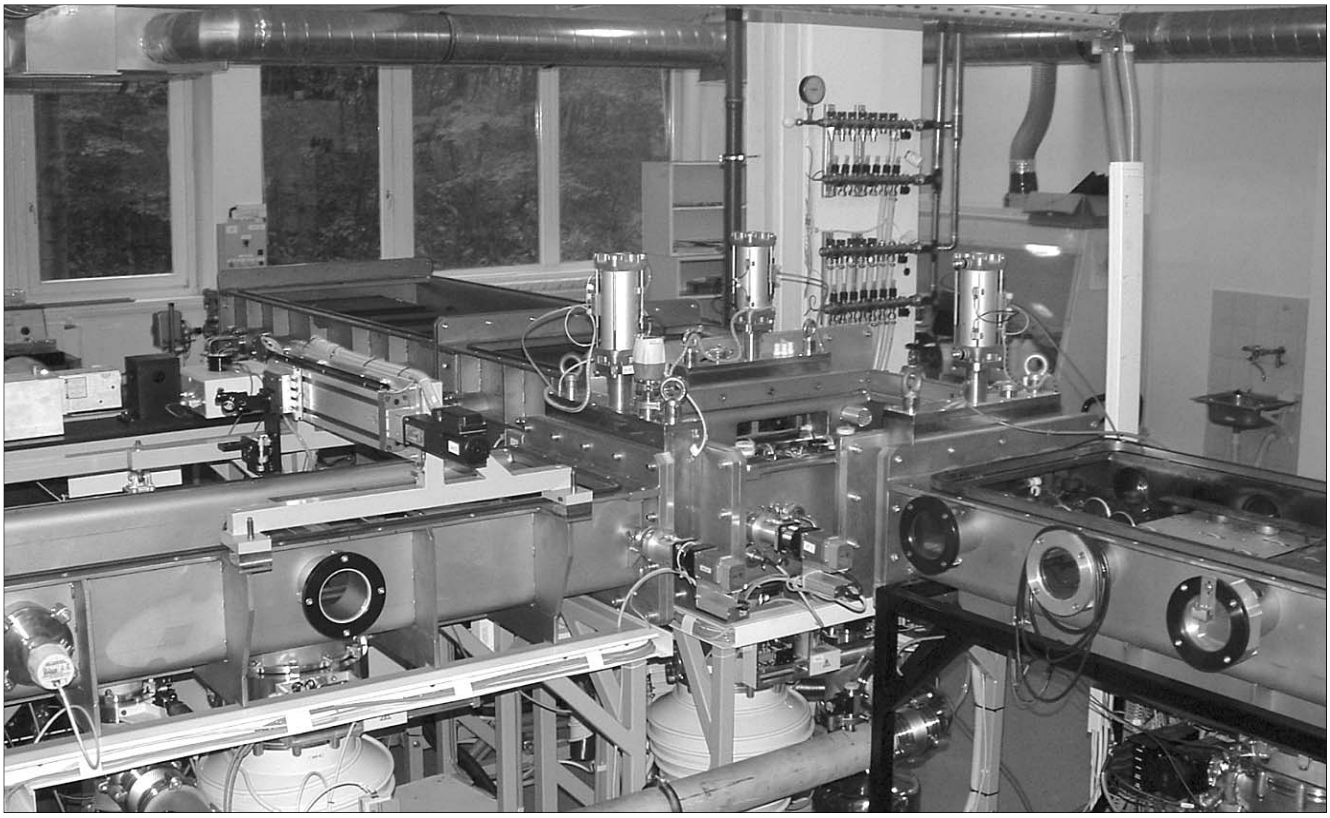
Deposition of the transparent conductive contact layer (ZnO window layer) is performed by reactive sputtering. The sputtered target is a metallic Al-Zn alloy while the reactive deposition takes place under argon-oxygen plasma. The aluminium is incorporated into the material as an n-type dopant and thereby provides the proper conductivity for the transparent contact layer.

The most sophisticated and most critical parts of the system are the evaporating sources containing graphite distributor pipes. The line sources consist of four point sources and their proper dimensioning and arrangement ensures the thickness uniformity of the evaporated layers.

Deposition is carried out by using the co-evaporation method. Individual sources evaporate the elemental source materials (Cu, In, Ga, Se) while the final crystalline structure and morphology is determined by an appropriate thermal annealing programme in the preheating/cooling chamber (Fig. 2.)

Fig. 5. shows a photograph of the vacuum system.

Figure 5.
The integrated vacuum system with the laser cutting chamber in the foreground



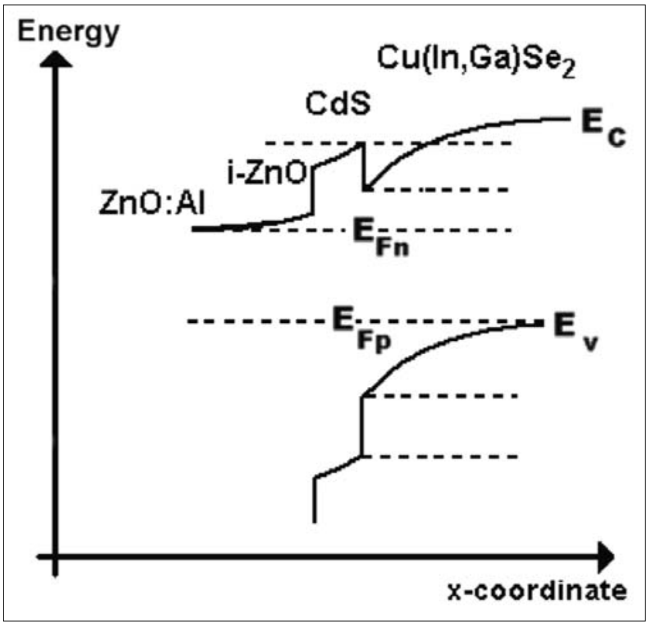


Figure 6.
Band structure of a CIGS solar cell layer sequence [4]

3. Materials science issues related to the CuInGaSe$_2$ material system

The high optical absorption of the direct semiconductor chalcopyrite makes solar cells based on very thin absorbers feasible. However, it also means that the incident sunlight is absorbed close to the surface. Assuming, it would be possible to dope chalcopyrite in a well controlled manner it would still be challenging to reach high efficiency with homojunction solar cell since the major part of carriers generated between the surface and the pn-junction would be lost by the surface recombination.

This problem is avoided by introducing the heterojunction concept with window (transparent conductive ZnO)-layer and absorber layer. Due to the wide band gap of the window layer, the absorption is shifted away from the surface to the internal hetero-interface The most effective approach to lowering recombination lies in minimizing the density of electrons or holes at the interface, which requires appropriating doping band line-up (matching) and interface charge.

The structure should contain an n-window – p-absorber heterojunction *(Fig. 6)*, where the Fermi-level at the interface is close to the conduction band and where the Fermi-level intersects the midgap energy at a short distance from the interface in the absorber. The interface charge should be positive to assist in establishing the structure.

Deposition of a CIGS layer structure with optimal properties therefore necessitates the study of the following five materials science issues:

– characterization of the shallow acceptor levels formed under Cu-poor growth conditions (these make possible the p-type autodoping of the material);
– formation of the optimal band gap by changing the In/Ga ratio in the layer, and formation of a graded band gap CIGS layer;
– study of the effect of the grain size distribution on the layer properties;
– study of the Na outdiffusion from the substrate glass;
– deposition of a buffer layer between the CIGS and transparent conductive ZnO contact by vacuum-technology compatible means (Fig. 6.).

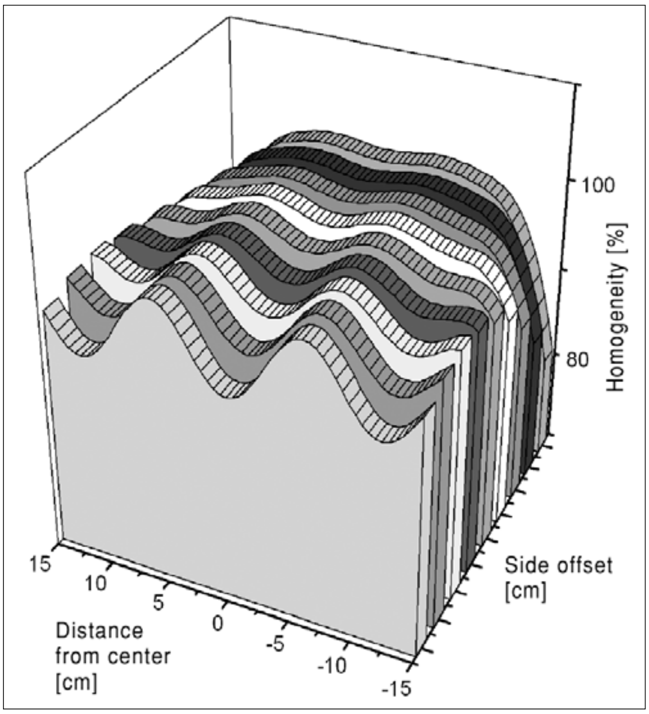
The results of the experimental work already carried out in the Solar Cell Innovation center can be summarized as follows [5-14]:

• The effect of the deposition parameters on the quality of Mo contact layer and the ZnO:Al window layer were studied in detail in the sputtering module. An optimal technology was elaborated together with the necessary conditions for reproducibility at room temperature. The obtained layer has the specific resistance of 1.7x 10^{-4} Ωcm, which is compatible with the best results published in the literature albeit with heated substrate.

• From our processing experience we determined that deviation from the optimal composition in the ZnO layer (towards the metallic as well as ceramic direction) can be monitored by spectroscopic ellipsometry. This allows the in-line integration of an efficient measurement technique into the system.

• The technology for selective laser cutting of ZnO and Mo layers was successfully elaborated in the laser module in cooperation with the researchers from the Department of Optics and Quantum Electronics of the Science University of Szeged.

Figure 7.
Dependence of lateral thickness homogeneity of evaporated CIGS layers as a function of the source geometry



• A process for wet chemical deposition of CdS buffer layer between the CIGS and transparent conductive ZnO contact (Fig. 6.)

• A computer model for the thickness uniformity of layers deposited from line sources was developed on the basis of evaporation experiments from an individual source. This model served as a basis for the design and construction of the evaporation chamber.

The materials analysis support required by the Solar Cell Innovation Center is provided by MFA and the other academic institutes in the consortium. The complex system of analyses and characterizations includes the following items:

1. Morphology study by Scanning Electron Microscopy SEM–FESEM (MFA)
2. Elemental composition analysis by Electron Dispersive Spectra (EDS) (MFA)
3. Elemental composition and phase analysis by X-ray diffraction (MFA)
4. Photoluminescence analysis (MFA)
5. Ellipsometric layer thickness and composition analysis (MFA)
6. Electron spectroscopy (XPS) and Secondary Ion Mass Spectrometry (SIMS) (ATOMKI)
7. Surface potential measurement (Kelvin probe method) and open circuit voltage measurements (BME EET)

5. Summary

This paper is an overview of the Solar Cell Technology Innovation Center which was built by a Hungarian R&D consortium at MTA MFA. This unique facility in Hungary is suitable for the development of process sequence for CIGS solar cells. It consists of a closed cycle vacuum pilot production-line equipped with laser cutting facility and in-line measurement techniques, an is applicable for:

– processing R&D purposes;
– professional training and education;
– support of the marketing activity of the industrial partner;
– pilot production of 300x300 mm^2 CIGS photovoltaic modules with an efficiency of ca. 12%.

The results of the project are also documented at http://www.mfa.kfki.hu/Napelem-CIS/.

Acknowledgement

The authors thank for the for the financial support of the Hungarian National Office for Research and Technology (NKTH) project No. 3/025/2001, and the participants of the R&D consortium for their contribution and Energosolar S. A. for the construction of the designed vacuum system.

References

[1] Dhere, N.G., Toward GW/year of CIGS production within the next decade, Solar Energy Materials & Solar Cells 91. (2007), pp.1376–1382.

[2] Thin Film Solar Cells, Fabrication, Characterization and Applications (ed. J. Poortmans and V. Arkhipov), Wiley Series in Mat. for Electr. & Optoelectr. Appl., John Wiley & Sons, 2006.

[3] Dhere, N.G., Present status and future prospects of CIGS thin film solar cells, Solar Energy Materials & Solar Cells 90. (2006), pp.2181–2190.

[4] Rau, U., Schock, H.W., Electronic properties of Cu(In,Ga)Se_2 heterojunction solar cells-recent achievements, current understanding and future challenges, Applied Physics A 69. (1999), pp.131–147.

[5] E. Horváth, A. Németh, A.A. Koós, A. L. Tóth, L.P. Biró, J. Gyulai, Focused Ion Beam based sputtering yield measurements on ZnO and Mo thin films, Superlattices and Microstructures, In Press, available online 8 June 2007.

[6] Á. Németh, Cs. Major, M. Fried, Z. Lábadi, I. Bársony, Characterisation of transparent conductive ZnO layers by spectroscopic ellipsometry, Submitted to Thin Solid Films.

[7] A. Buzás, Zs. Geretovszky, Patterning ZnO layers with frequency doubled and quadrupled Nd:YAG laser for PV application, Thin Solid Films, In Press, Corrected proof available online 16 April 2007 (doi:10.1016/j.tsf.2007.04.026).

[8] Á. Németh, E. Horváth, Z. Lábadi, L. Fedák, I. Bársony, Single step deposition of different morphology ZnO gas sensing films, Sensors and Actuators B, accepted for publication.

[9] Rakovics V., Chemical bath deposition of CdS and CdPbS nanocrystalline thin films and investigation of their photoconductivity, 2005 MRS Fall Meeting, 27 November - 2 December, Boston, MRS Symposium Proceedings 900, pp.87–91.

[10] V. Rakovics, Zs.J. Horváth, Z.E. Horváth, I. Bársony, C. Frigeri, T. Besagni, Investigation of CdS/InP heterojunction prepared by chemical bath deposition, 8th Expert Evaluation and Control of Compound Semiconductor Materials and Technologies, EXMATEC'06, 14-17 May 2006, Cádiz, Spain, Physica Status Solidi C 4. (2007), pp.1490–1494.

[11] V. Rakovics, Zs.J. Horváth, K.T. Eppich, B. Pődör,

Electrical and photoelectrical behaviour of nanocrystalline CdS/InP heterojunction p-n diodes, XXXV Int. School on the Physics of Semiconducting Compounds, 17-23 June 2006, Jaszowiec, Poland, Abstract Booklet, p.44.

[12] Zs.J. Horváth, V. Rakovics, Z.E. Horváth, Electrical properties of nanocrystalline CdS/InP heterojunction p-n diodes prepared by chemical bath deposition, International Workshop on Nanostructured Materials, NANOMAT 2006, 21-23 June 2006, Antalya, Turkey, Book of Abstracts, p.69.

[13] V. Rakovics, Zs.J. Horváth, B. Pődör, Electrical and optical behaviour of nanocrystalline CdS/InP heterojunction p-n diodes, 6th Int. Conference Advanced Semiconductor Devices and Microsystems, ASDAM'06, 16-18 October 2006, Smolenice, Slovakia, p.155.

[14] Á. Németh, V.Rakovics, E.B. Kuthi, Z. Lábadi, Á. Nemcsics, S. Püspöki, A.L. Tóth, I. Bársony, Study the properties of sulphide buffer layers as a function of deposition parameters and annealing, Proc. of the 21st European Photovoltaic Solar Energy Conference, 4-8 September 2006, Dresden, Germany pp.1986–1989.

Authors

Ágoston Németh PhD student, graduated in 2002 as an electrical engineer in the Budapest University of Technology and Economics. He is currently working on his PhD thesis. His interest is thin film solar cell technologies and related materials.

Zoltán Lábadi received his PhD in Budapest University of Technology and Economics in 2001. He was working on mid-infrared compound semiconductor devices at the Physics Department of Lancaster University int he UK (1997-2001). He is currently a research fellow at the Research Institute for Technical Physics and Materials Science (Budapest). His current interest is thin film solar cell materials and characterization.

Vilmos Rakovics graduated in chemistry from the Eötvös Loránd University in 1979. He holds a C.Sc from the Hungarian Academy of Sciences (1995) and a PhD from the Technical University of Budapest (1996). He is currently a senior research fellow at the Research Institute for Technical Physics and Materials Science. His current interest is preparation and characterization of optoelectronic devices.

István Bársony graduated in electrical engineering from the Technical University of Ilmenau, Germany in 1971. He holds a C.Sc from the Hungarian Academy of Sciences (1978) and a PhD from the Technical University of Budapest (1996) and a D.Sc from the HAS (2001). During his professional career he was working on research assignments in Hungary mainly in silicon technology, in Japan (1983-1986) on imaging application of the Static Induction Transistor, in the Netherlands (1988-1993) at the MESA Research Institute of the University of Twente on Rapid Thermal Multi-Processing. Since 1993 he has been with the Research Institute for Technical Physics and Materials Science of the Hungarian Academy of Sciences (MFA) Budapest, from 2004. He is the director of MFA. He has led several international projects a/o on solar cell and microsystems research. He holds 12 patents, published over 80 scientific papers and is a professor of nanotechnology at the University of Veszprém.

István Krafcsik was graduated in electrical engineering in Kiev in 1973. He holds Ph. D. degree in electrical engineering (1978). Until 1991 he worked for the Research Institute for Microelectronics, Budapest as research fellow and head of department. He was the the founder of Kraft Electronics Ltd., and served as technical director between 1991 and 2001, and CEO of Kraft Inc. until 2005. Since 2005 he has been working as CTO of EnergoSolar Hungary Ltd. He is author of 17 research papers with over a hundred independent citations and holds 7 patents. Main field of interest: vacuum equipment for thin film solar cell production, coordination of research and development.

On the scaling characteristics of MMORPG traffic

SÁNDOR MOLNÁR, GÉZA SZABÓ

Budapest University of Technology and Economics,
Department of Telecommunication and Media Informatics
{molnar, szabog}@tmit.bme.hu

Reviewed

Keywords: gaming traffic, MMORPG, scaling analysis, traffic modelling

In this paper a comprehensive scaling analysis of the traffic of the four most popular Massively Multiplayer On-line Role Playing Games (MMORPG) is presented. The examined games are World of Warcraft, Guild Wars, Eve Online and Star Wars Galaxies. Both server and client generated traffic are analyzed in details. Our study reveals the basic statistical properties of the investigated games focusing on the correlation and scaling behavior. Although the examined games are all from the same genre and basic statistics such as the mean packet rate, variation of the packet rate, skewness of the packet rate distributions fall into the same magnitude, the games exhibit diverse traffic characteristics. We have found that in spite of the fact that some similarities can be found among the scaling characteristics of these games they show versatile scaling properties and the games can not be treated with one common model.

1. Introduction

Today's Internet usage tends to serve the expansion of the entertainment industry. Besides the content-delivery traffic (e.g. web, P2P), significant traffic appeared which is generated by online games. *Massive Multiplayer Online Role Playing Games* (MMORPG) attract the most users who play simultaneously in virtual worlds over the Internet.

Earlier studies focused on games that were popular at that time. These games include the popular first person shooters, e.g. Counterstrike which was analyzed in [1]. Today most of the gaming traffic is generated by massively multiplayer online games thus such works dealing with the new type of traffic have recently appeared. Chen et al. analyzed ShenZhou Online, a mid-scale, commercial MMORPG in Taiwan [2]. They extended their work in [3] where they performed scaling analysis on the measurements. They explained the scaling results with the fact that an ON-OFF model can be constructed based on the results of the analysis where ON and OFF periods are in connection with the players' active and idle times indirectly. In [4] authors analyzed Lineage II which was one of the world's largest MMORPGs in terms of the number of concurrent users at that time. In [5], authors took Ragnarok Online, and studied the traffic generated by mainstream game bots and human players. In [6], authors used CrossFire, an open source MMOG to evaluate their performance model. All of these works used packet level network traces and statistical methods for traffic characteristics analysis.

However, situation has recently changed. According to [7] the top game having the most active subscribers is World of Warcraft. The number of active subscribers is four times higher than in Lineage II. We decided to analyze the following games from the charts of [7]: *World of Warcraft, Eve Online, Star Wars Galaxies* and *Guild Wars*. There are several reasons behind this decision. All of these games are commercial, and it was only recently possible to access the games via Internet and play with them free during a trial period. In addition, the target market of the games used in previous analysis was definitely the Asian market. However, we can hardly come across with any of the traffic of those games in a European or American network.

The motivation of our work was to understand the traffic characteristics and, especially, the scaling behavior of the traffic generated by the selected games. Although the traffic rates generated by the clients are low comparing to other applications, their aggregation on the server side can become significant due to the large population of players. The scaling characteristics of the internet traffic, with special attention to the growing gaming traffic, can have significant impact on network performance and engineering.

2. Measurements

The measurements took place on a client machine connected to a campus network with Internet access via a 100 Mbps FDDI. The network parameters of this connection is far above the capabilities of a network for which these games are designed for, thus we assumed that we did not have to deal with any parameter change in the game traffic due to the network inadequacy. The advantage of the measurement configuration is that we can observe the client network traffic practically without loss of packets and network delay. The measurements were conducted during the 19-20 hour periods on weekdays in January, 2007.

We have measured both the downstream traffic from the server to the client (we will call it server traffic throughout the paper) and the upstream traffic from the client to the server which will be called client traffic. The network traffic of the client machine running the games was captured by Wireshark with microsecond accuracy.

The traffic of the different games can be seen in *Figures 1-4.* As the statistical methods which were applied to the measurements presumed the stationarity property of the examined data series, the selected intervals for examination are shown in the figures.

3. Basic traffic characteristics

Observing the probability density function (PDF) of the interarrival times of the packets derived from the clients to the server, there are characteristic values for some specific packet inter-arrival time values. These are the effects of the internal working mechanism of the game client application as the measurement setup does not

Figure 1. World of Warcraft traffic intensity (packets/sec), selected interval: 1100-2000

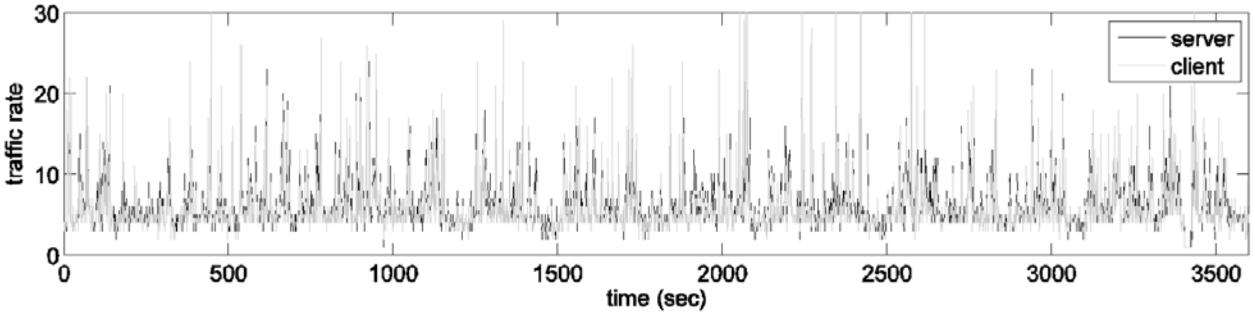


Figure 2. Guild Wars measured traffic intensity (packets/sec), selected interval: 1600-2800

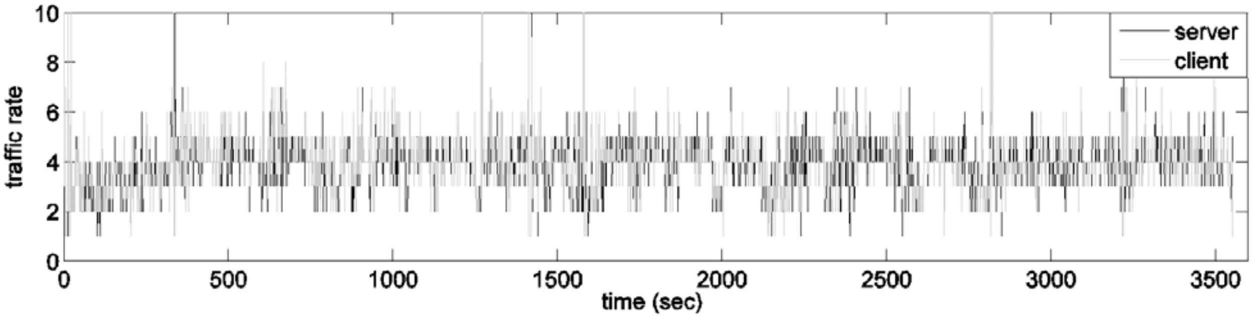


Figure 3. Eve Online measured traffic intensity (packets/10 sec), selected interval: 50-450

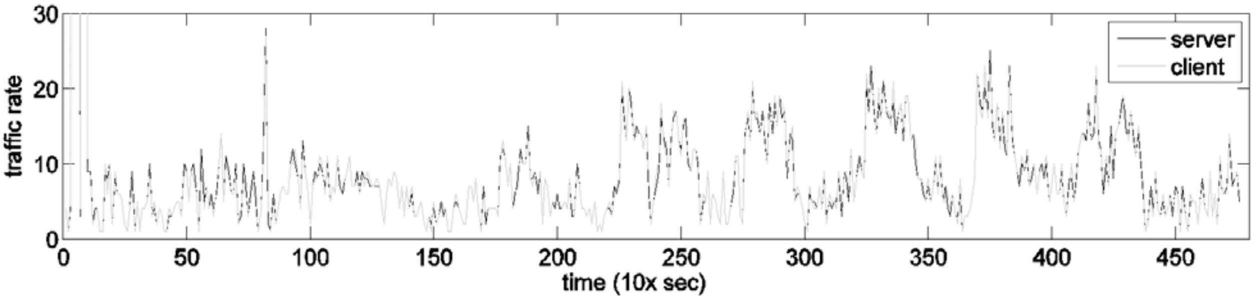
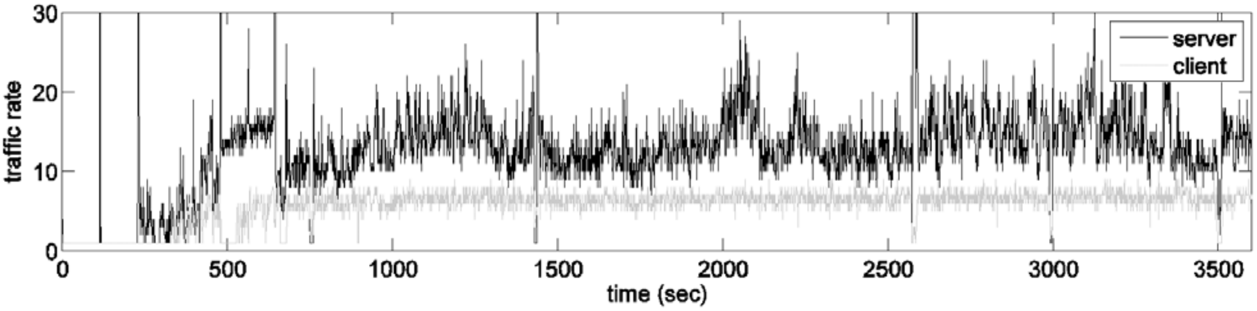


Figure 4. Star Wars Galaxies measured traffic intensity (packets/sec), selected interval: 1500-2000



add any delay to the captured packets derived from the clients.

All the games have a high probability value about the 200 msec packet inter-arrival time. This value was a reasonable design decision, as MMOGs are designed to run smoothly even with 1250 msec latency in game play thus with the 200 msec periodicity even a retransmission fits into this interval length. World of Warcraft and Guild Wars have peaks at their PDF at about 300 msec and Star Wars Galaxies has a high peak at 140 msec. This lower packet inter-arrival time can be explained by the situation that Star Wars Galaxies uses UDP protocol with plenty of small packets, thus the communication model is different from the other analyzed games. Eve Online generates packets much rarely than the other games thus the probability of high packet inter-arrival time values decreases slower.

In case of the server packets the very low packet inter-arrival time values are due to the fragmentation of packets when a data burst is transmitted towards the client. The identification of packet inter-arrival time values can be effectively used during traffic classification.

Investigating the probability density function of the packet payload sizes, it can be experienced that the zero and few-byte payloads occur frequently both at the client and at the server side. One reason for this is that at least the TCP packets have to be acknowledged even if the party itself does not want to send data. Another reason is that the game protocol is constructed as an overlay protocol on TCP. As an example, we can check the general structure of the World of Warcraft (WoW) packets, where we can see that the TCP data carries a 4 byte WoW packet header if it is a server packet and 6 byte if it is a client packet. This header contains a WoW packet type field which is necessary for parsing the rest of the packet accordingly. The WoW packet header is encrypted. If either the client or the server sends a packet apart from the TCP acknowledgements, these packets have at least 6 or 4 bytes length even if they do not carry any game data.

We can confirm earlier works which found that comparing the client and server packet size distributions the client packets are smaller as they contain the commands of one player, while server packets convey nearby the actions of nearby players and monsters as well as system messages.

Comparing the probability density function of the server and the client packet rates we can find that those games which applies TCP for communication has similar PDF, while Star Wars Galaxies which uses UDP for communication has very distinct PDF characteristics as the probability of high packet rate on the server side is higher than on the client side. Other basic statistical descriptors are shown in *Table 1*.

4. Long-range dependence analysis

The Long-Range Dependent (LRD) property of a traffic flow is revealed in the power law decay of the autocorrelation function at large lags, i.e.

$$r(k) \sim c|k|^{2H-2}, k \to \infty, H \in (0.5,1)$$

and *c* is constant.

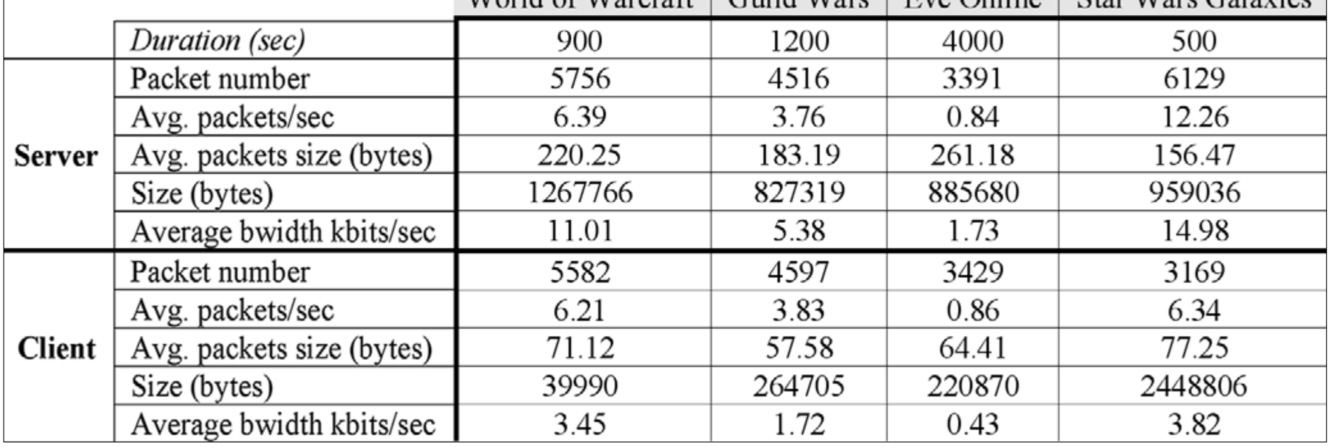
The degree of this slow decay is determined by the Hurst parameter (*H*). Intuitively, long-range dependence measures the memory of a process. For LRD data the ACF decays very slowly (power-law decay). On the contrary, Short-Range Dependence (SRD) is characterized by quickly (exponential-like) decaying correlations.

Among the several statistical methods of LRD testing [10] we choose periodogram analysis, R/S analysis, variance of residuals, variance-time plot, and the Whittle estimator and use the logscale diagram based on the wavelet transform [8] to verify the results.

The results of our LRD analysis can be found in *Table 1*. We can see that World of Warcraft traffic is strongly long-range dependent for the server traffic. However, the LRD tests results have not confirmed the same for the client traffic due to the statistical inaccuracy.

In case of Guild Wars, the client traffic shows LRD property, but in case of the server traffic the test can not be performed due to the lack of data in higher time

Table 1. Basic data of the selected traffic trace segments



		World of Warcraft	Guild Wars	Eve Online	Star Wars Galaxies
	Duration (sec)	900	1200	4000	500
Server	Packet number	5756	4516	3391	6129
	Avg. packets/sec	6.39	3.76	0.84	12.26
	Avg. packets size (bytes)	220.25	183.19	261.18	156.47
	Size (bytes)	1267766	827319	885680	959036
	Average bwidth kbits/sec	11.01	5.38	1.73	14.98
Client	Packet number	5582	4597	3429	3169
	Avg. packets/sec	6.21	3.83	0.86	6.34
	Avg. packets size (bytes)	71.12	57.58	64.41	77.25
	Size (bytes)	39990	264705	220870	2448806
	Average bwidth kbits/sec	3.45	1.72	0.43	3.82

scales. Star Wars Galaxies' server traffic shows LRD property with parameter *H=0.75*. The client traffic can not be estimated due to similar reasons as in the case of Guild Wars server traffic. In case of Eve Online server traffic the higher ranges can not be used for LRD parameter estimation due to the lack of data in that ranges. The same statements are true for the client traffic of Eve Online.

The summary of the results of the long range analysis can be found in *Table 2*.

5. Scaling analysis

Scaling properties of traffic can be efficiently investigated by multifractal analysis via wavelet-based methods [8]. The discrete wavelet transform represents a data series *X* of size *n* at a scaling level *j* by a set of wavelet coefficients $d_X(j,k), k=1,2,...n_j$, where $n_j=2^{-j}n$. Define the q^{th} order Logscale Diagram (q-LD) by the log-linear graph of the estimated q^{th} moment

$$\mu_j(q) = \frac{1}{n_j}\sum_{k=1}^{n_j}\left|d_X(j,k)\right|^q$$

against the octave *j*.

Linearity of the LDs at different moment order *q* indicates the scaling property of the series, i.e. $\log_2 \mu_j q = j\alpha(q)+c_2(q)$, where $\alpha(q)$ is the scaling exponent and $c_2(q)$ is a constant. In our test results we plot $y_j=\log_2 \mu_j(q)$-t for $q=2$ which is called the second-order logscale diagram (LD). The plot of $\alpha(q)$ against *q* can reveal the type of scaling [9].

In case of *monofractal* scaling $\alpha(q)$ varies linearly with *q* while for *multifractals* the variation is non-linear. For testing this behavior the Linear Multiscale Diagram (LMD) can efficiently be used which is defined as $h_q=\alpha(q)/q-1/2$.

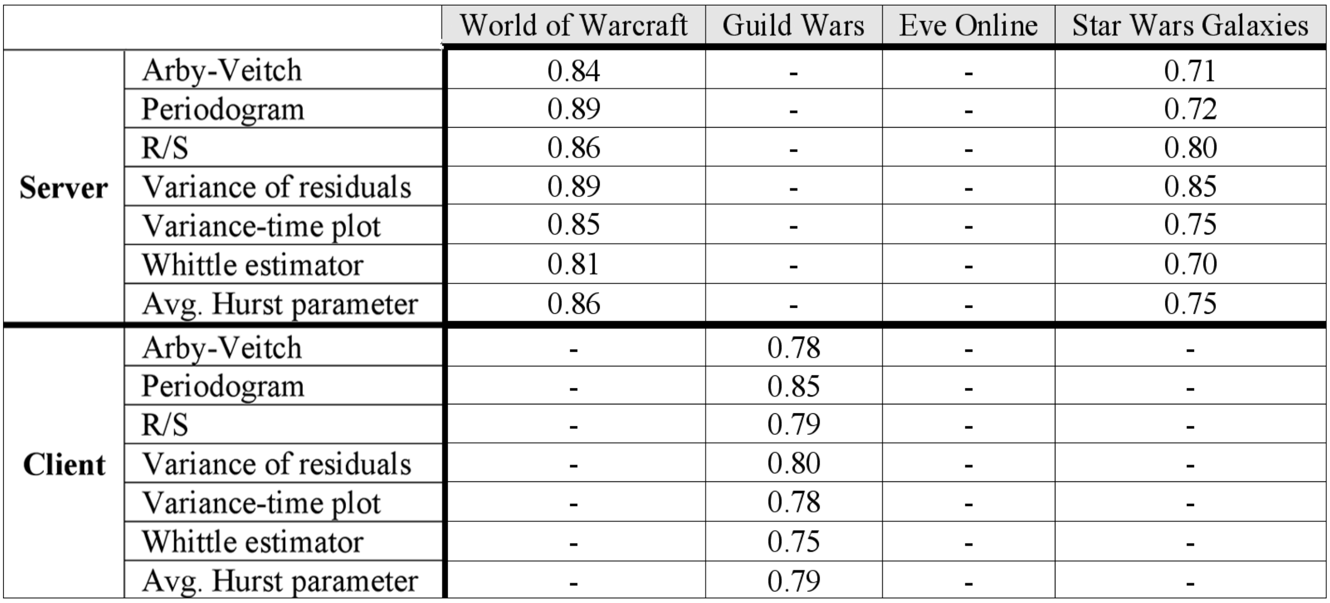
World of Warcraft

It can be seen that the logscale diagram of the WoW server traffic is approximately linear *(Fig. 5.)* for the whole range and supports the LRD property suggested by the LRD tests. Since the linearity holds for the whole investigated range it also suggests possible statistical self-similarity over these time scales. The linear multiscale diagram depicted in *Fig. 13.* confirms this observation. The LMD of World of Warcraft soon takes up a stabilized value around $h_q=-0.16$ which gives an estimate of H=0.84 since $H=h_q+1$ for all *q* in case of self-similar traffic.

The estimated value is in accordance with the values calculated by the LRD tests (H=0.86). We can conclude that World of Warcraft server traffic is not only LRD but the statistical *self-similarity* is a good model for this type of traffic in these time scales. The range of the time scales selected for the analysis based on the fact that there is no reasonable rate function below the 1 sec time intervals, thus the low packet rate of the traffic imposes a lower bound for the analyzed time scale. On the higher time scales we selected the longest stationary parts of the measurements but even with this method it was not possible gain enough samples from higher time scales.

A different behavior can be observed for the World of Warcraft client traffic. Examining the logscale diagram in *Fig. 6.* we can only find scaling region in the range between j=1 and j=4 (1 sec-16 sec). The multiscale diagram *(Fig. 14.)* reveals the scaling type in the range between j=1 and j=4 (1 sec-16 sec): the non-linear LMD plot shows *multifractal* behavior. The multifractal behavior frequently found together with the non-Gaussian like marginals of the rate distribution. This property holds for this case too. The kurtosis (13.53) and skewness (2.89) are also far from the Gaussian-like distributions. (A Gaussian distribution has kurtosis and skewness metrics 3

Table 2.
Summary of the long-range dependence analysis (n.a.=statistical results are not reliable due to insufficient sample size)



		World of Warcraft	Guild Wars	Eve Online	Star Wars Galaxies
Server	Arby-Veitch	0.84	-	-	0.71
	Periodogram	0.89	-	-	0.72
	R/S	0.86	-	-	0.80
	Variance of residuals	0.89	-	-	0.85
	Variance-time plot	0.85	-	-	0.75
	Whittle estimator	0.81	-	-	0.70
	Avg. Hurst parameter	0.86	-	-	0.75
Client	Arby-Veitch	-	0.78	-	-
	Periodogram	-	0.85	-	-
	R/S	-	0.79	-	-
	Variance of residuals	-	0.80	-	-
	Variance-time plot	-	0.78	-	-
	Whittle estimator	-	0.75	-	-
	Avg. Hurst parameter	-	0.79	-	-

and 0, respectively.) For the upper time scales (above 16 sec) no scaling property can be found.

It is important to note that self-similarity is a characteristic property for time scales higher than 50-100 msec, e.g. in the case of the round trip time of a TCP packet. Below this limit the fractional property can be found, but in our case the multifractal property of the client traffic can be observed for as large time scales as 1-16 sec.

Guild Wars

The logscale diagram of Guild Wars server traffic *(Fig. 7.)* can be divided into two ranges: *j*=1–4 (1 sec-16 sec) and *j*=4–6 (16 sec-1 min) where scaling region can only be detected in the lower ranges. Depicting the LMD of ranges 1-4 in *Fig. 15.*, it can be seen that it has the same value over all the investigated moments. Thus it can be concluded that Guild Wars server traffic can be modeled with a *monofractal* model with *h*=0.63 scaling parameter in these time scales.

Examining *Fig. 8.* we can see that the logscale diagram of the Guild Wars client traffic is approximately linear which suggests a self-similar scaling over all the investigated time scales. The LMD in *Fig. 16.* shows that the Guild Wars client traffic indeed has *self-similar* scaling. The estimated *H*=0.78 from the LD diagram is in good accordance with the estimated *H*=0.79 obtained by the LRD tests.

Because of the self-similar scaling we can expect a Gaussian-like rate distribution. Both the shape of the rate distribution and also the estimated kurtosis (3.09) and skewness (0.04) metrics confirms that our expectation is true.

Eve Online

The logscale diagram of Eve Online server traffic plotted in *Fig. 9.* can be divided into two ranges where the scaling property can be examined: 1-3 (10 sec-80 sec) and 3-5 (80 sec-over 5 min). The range 3-5 contains very few data thus the estimators are very inaccurate in this range. Analyzing the range 1-3 by the multiscale diagram (in *Fig. 17.*) it can be seen that the calculated scaling parameter is around 0.54 which suggests a non-scaling *noise-like* behavior. Thus we can conclude that there is *no scaling* property of Eve Online server traffic for the whole range.

Similar statements are true for the client traffic as well: the scaling parameter between 1-3 (10 sec-80 sec) is *h*=0.52, and the range between 3-5 (80 sec-5 min) contains few data *(Fig. 10. and 18.)*, thus we can conclude that there is *no scaling* property of Eve Online client traffic for the whole range.

Star Wars Galaxies

Investigating the server traffic of Star Wars Galaxies it can be seen on the logscale diagram in *Fig. 11.* that it is also approximately linear for the whole range and in *Fig. 19.* the LMD gives values around h_q=0.29. Thus Star Wars Galaxies server traffic can also be modeled with a statistical *self-similar* process with *H*=0.71 estimated by the LD plot. This estimation matches the *H*=0.75 obtained by the LRD tests. The self-similar property also comes together with the Gaussian-like marginals as could be seen in the rate distribution curves and also from the estimated kurtosis (3.23) and skewness (0.45) metrics.

Looking at the logscale diagram in *Fig. 12.* of Star Wars Galaxies client traffic we can divide two ranges where the scaling property can be examined: 1-3 (1 sec-8 sec) and 3-5 (8 sec-1 min). The range 3-5 consists of too few data so that the estimators are very inaccurate in this range. Examining the range 1-3 by the multiscale diagram (in *Fig. 20.*) it can be seen that the calculated scaling parameter is around 0.5 which suggests a non-scaling *noise-like* behavior. Thus we can conclude that there is *no scaling* property of Star wars Galaxies client traffic for the whole range.

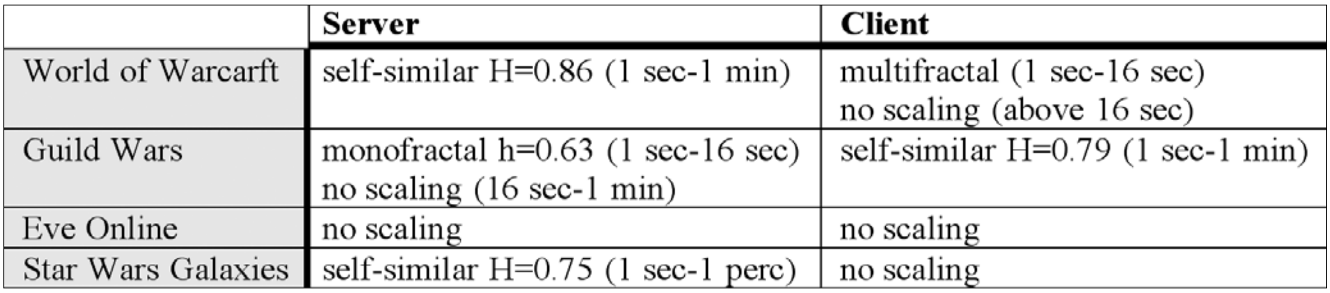
In *Table 3.* the summary of the scaling analysis can be found.

6. Conclusions

In this paper we have analyzed four popular games traffic in both server and client directions. We have presented the important statistical characteristics of these games and we have carried out a comprehensive scaling analysis including long-range dependence analysis with several tests and a detailed scaling analysis by a wavelet-based multifractal analysis.

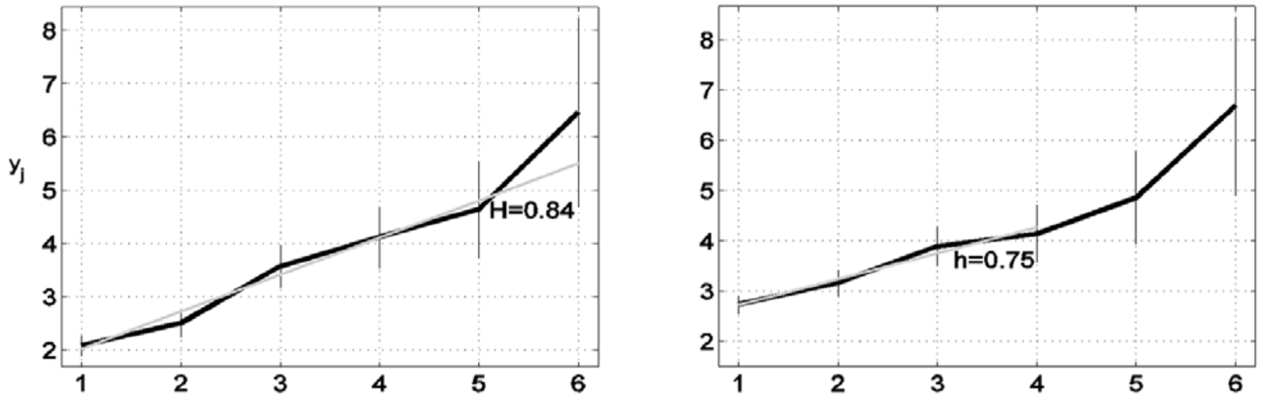
We have found different scaling properties of the investigated MMORPG traffic types. The server traffic of World of Warcraft is statistically self-similar with Hurst parameter around 0.86. However, the client traffic of World of Warcraft is multifractal below 16 sec time scales. The Guild Wars client traffic is statistically self-similar with

Table 3. Summary of the scaling analysis

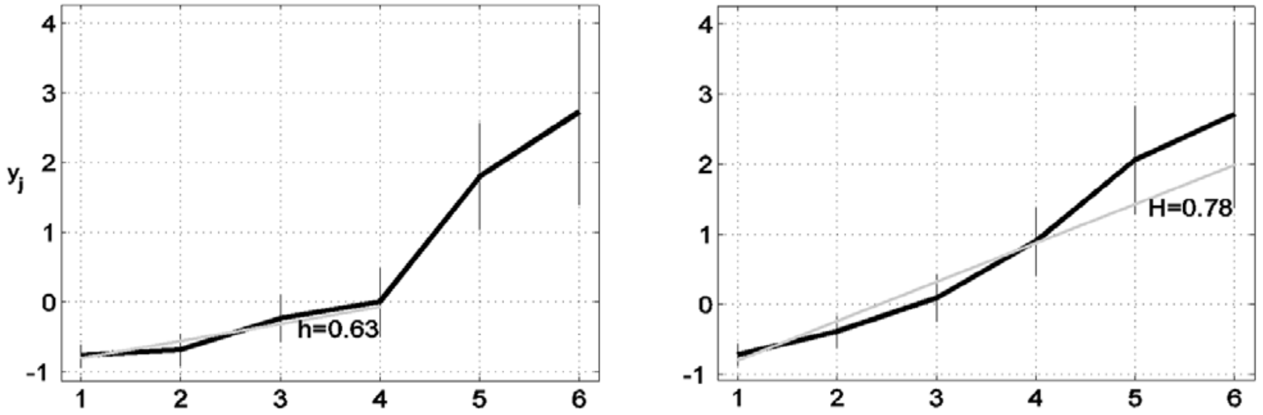


	Server	Client
World of Warcarft	self-similar H=0.86 (1 sec-1 min)	multifractal (1 sec-16 sec) no scaling (above 16 sec)
Guild Wars	monofractal h=0.63 (1 sec-16 sec) no scaling (16 sec-1 min)	self-similar H=0.79 (1 sec-1 min)
Eve Online	no scaling	no scaling
Star Wars Galaxies	self-similar H=0.75 (1 sec-1 perc)	no scaling

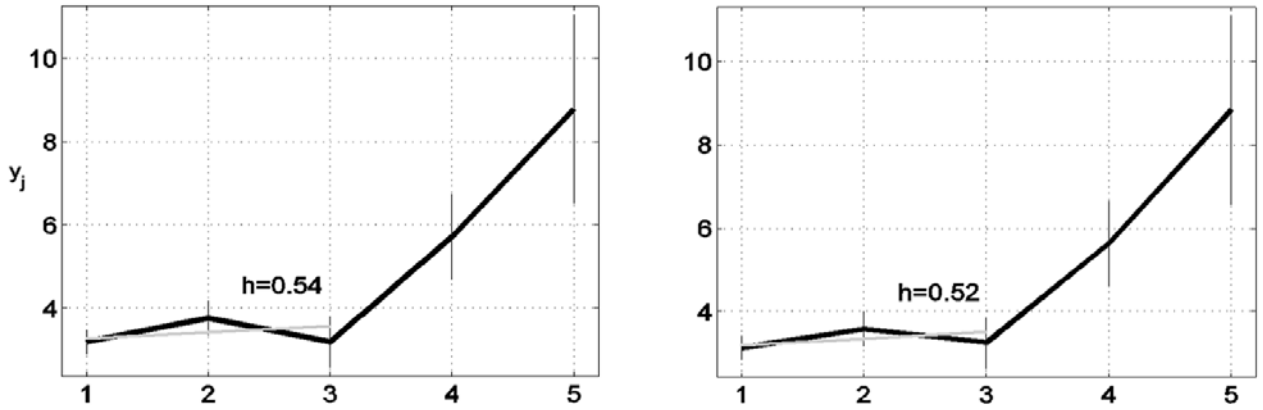
Figures 5-6. World of Warcraft server and client logscale diagram covering timescales from 1 sec to 1 min



Figures 7-8. Guild Wars server and client logscale diagram covering timescales from 1 sec to 1 min



Figures 9-10. Eve Online server and client logscale diagram covering timescales from 10 sec to 1 min



Figures 11-12. Star Wars Galaxies server and client logscale diagram covering timescales from 1 sec to 32 sec

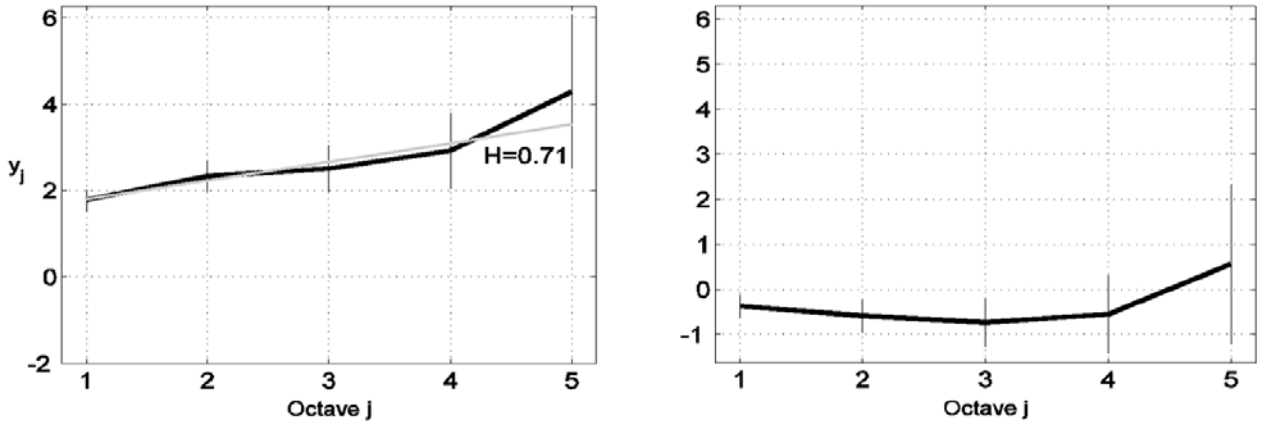


Figure 13. World of Warcraft server multiscale diagram depicted on the time scales between 1 sec and 1 min

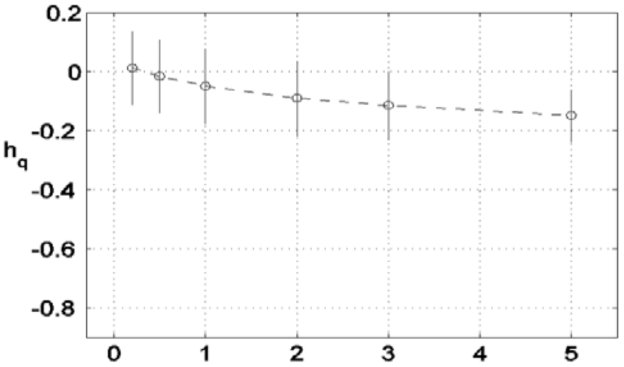


Figure 14. World of Warcraft client multiscale diagram depicted on the time scales between 1 sec and 16 sec

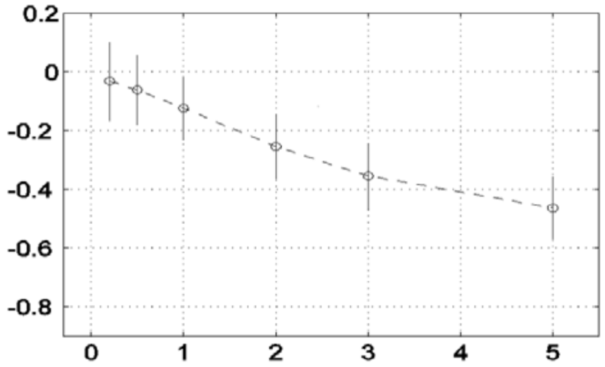


Figure 15. Guild Wars server multiscale diagram depicted on the time scales between 1 sec and 16 sec

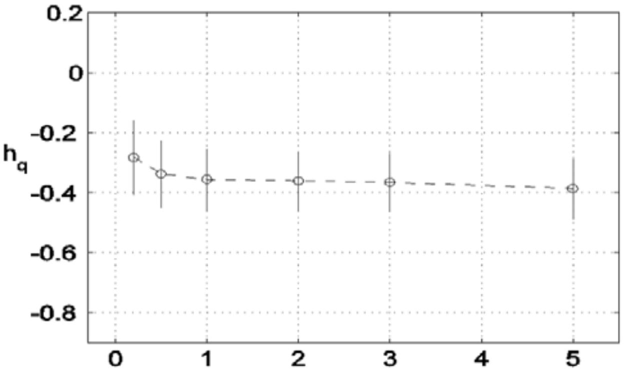


Figure 16. Guild Wars client multiscale diagram depicted on the time scales between 1 sec and 1 min

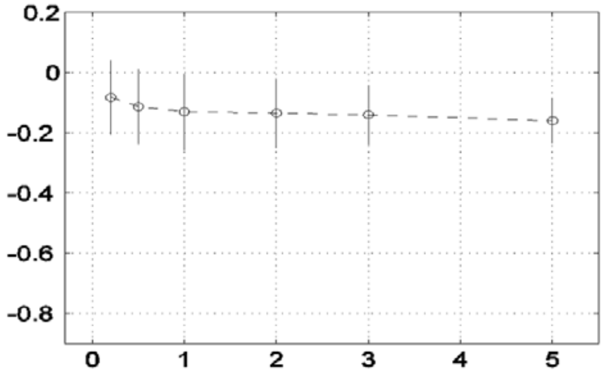


Figure 17. Eve Online server multiscale diagram depicted on the time scales between 10 sec and 1 min

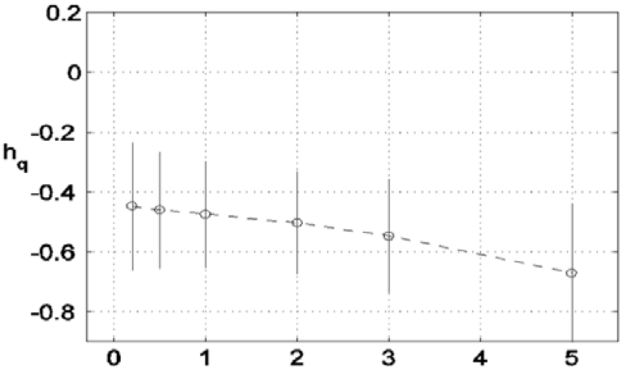


Figure 18. Eve Online client multiscale diagram depicted on the time scales between 10 sec and 1 min

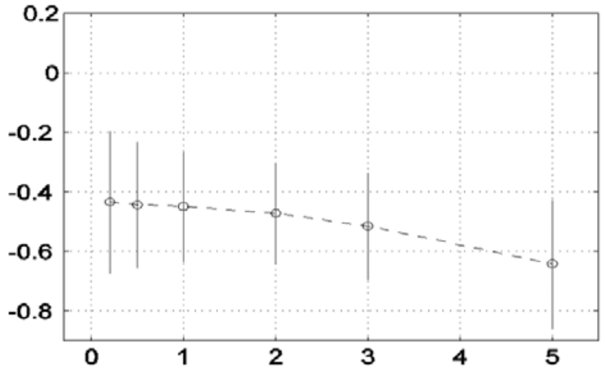


Figure 19. Star Wars Galaxies server multiscale diagram depicted on the time scales between 1 sec and 32 sec

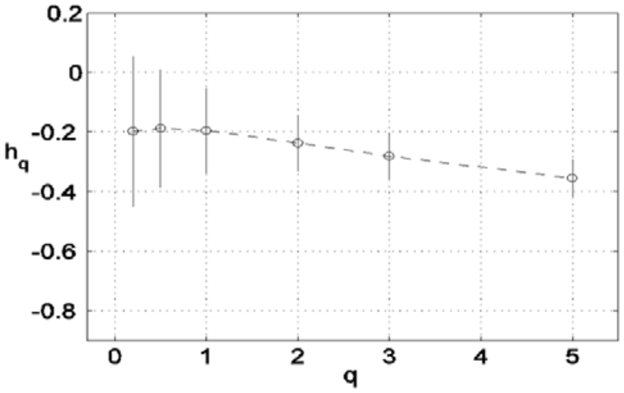
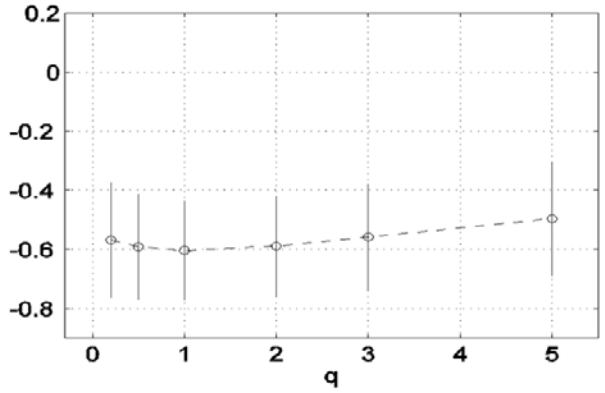


Figure 20. Star Wars Galaxies client multiscale diagram depicted on the time scales between 1 sec and 8 sec



Hurst parameter around 0.79. The server traffic in this case also shows scaling behavior over small time scales, namely, it has monofractal scaling. Star War Galaxies' server traffic has self-similar scaling with Hurst parameter 0.75. However, this game traffic does not have this scaling characteristics from the other direction. Finally, both server and client traffic of Eve Online have no scaling behavior.

As a conclusion we have found that in spite of the fact that some similarities can be found among the scaling characteristics of these games they show versatile scaling properties. From these results we conjecture that the emerging network traffic in the Internet cannot be classified by a typical gaming traffic behavior but rather will depend on the characteristics of the actual dominant gaming application.

Our future work will address the analysis of the network game traffic aggregates and the modeling of these traffic types. Furthermore, we would like study the network performance implications of these game traffic characteristics.

References

[1] W. Feng, F. Chang, W. Feng, J. Walpole,
Provisioning on-line games:
A traffic analysis of a busy Counter-strike server.
In SIGCOMM Internet, Measurement Workshop,
Marseille, France, 2002.

[2] K. Chen, P. Huang, C. Huang, C. Lei,
Game traffic analysis: an MMORPG perspective.
In NOSSDAV'05, New York, USA, 2005.

[3] K.-T. Chen, P. Huang, C.-L. Lei,
Game traffic analysis: An MMORPG perspective.
Computer Networks, 51(3), 2007.

[4] J. Kim, J. Choi, D. Chang, T. Kwon, Y. Choi, E. Yuk,
Traffic characteristics of a massively multi-player
online role playing game.
In NetGames'05, New York, USA, 2005.

[5] K. Chen, J. Jiang, P. Huang, H. Chu, C. Lei, W. Chen,
Identifying MMORPG bots: A traffic analysis approach.
In ACM SIGCHI ACE'06, Los Angeles, Jun 2006.

[6] M. Ye, L. Cheng,
System-performance modeling for
massively multiplayer online role-playing games.
IBM Syst. Journal, 45(1), pp.45–58., 2006.

[7] MMOGChart.com (http://www.mmogchart.com)

[8] P. Abry, D. Veitch,
Wavelet analysis of long-range-dependent traffic.
IEEE Transactions on Information Theory,
44(1), pp.2–15., 1998.

[9] P. Abry, P. Flandrin, M. Taqqu, D. Veitch,
Wavelets for the analysis, estimation and synthesis
of scaling data. Self Similar Network Traffic Analysis
and Performance Evaluation,
K. Park and W. Willinger (Eds.), 1999.

[10] J. Beran,
Statistics for long-memory processes.
Chapman and Hall, One Penn Plaza, 1995.

Authors

Sándor Molnár received his M.Sc. and Ph.D. in electrical engineering from the Budapest University of Technology and Economics (BME), Hungary, in 1991 and 1996, respectively. In 1995 he joined the Department of Telecommunications and Media Informatics, BME. He is now an Associate Professor and the principal investigator of the teletraffic research program of the High Speed Networks Laboratory. Dr. Molnár has been participated in several European projects and he is a member of the IFIP TC6 WG 6.3. He is also a member of the Editorial Board of the Springer Telecommunication Systems journal and recently works as a General Chair of SIMUTOOLS 2008. Dr. Molnár has more than 120 publications in international journals and conferences. His main interests include teletraffic analysis and performance evaluation of modern communication networks.

Géza Szabó received the degree of Master in Computer Science in 2006 from the Budapest University of Technology and Economics, in Budapest, Hungary. The author's main interests include internet traffic classification and modelling. He currently pursues a PhD degree in the High Speed Networks Laboratory of the Budapest Univeristy of Technology and Economics.

ENUM in everyday practice – is it a dream or an opportunity?

ISTVÁN TÉTÉNYI, GYULA SZABÓ, ANDRÁS KISS, ANDRÁS TÓTH

MTA SZTAKI, Center for Internet Applications and Technology
tetenyi@sztaki.hu

Reviewed

Keywords: next generation network, telecommunication, Internet, ENUM

The aim of this paper is to elaborate on Electronic NUmber Mapping (ENUM) technology. In the introduction, the role of ENUM is presented. Afterwards an ENUM measuring method is introduced, and several determining parameters are identified and it is shown how these parameters influence the performance of ENUM. The closing part shows overall ENUM and DNS performance parameters, apart from the DNS server raw performance. Finally, as a sanity check, the Hungarian voice communication profile is compared with the measured ENUM performance in order to have sizing guidelines for ENUM related services.

1. Introduction

ENUM is an IETF standard that makes it possible to assign phone numbers (in E.164 format) and standard domain names used on the Internet. ENUM is also a tool that allows to access Internet based services via phone numbers. Consequently, ENUM is one of the relatively new technologies that pave the way towards convergent networks, the so called Next Generation Network.

The IETF working group was established in 1999, its core standard in RFC 3761 was published in 2004 which is an update of RFC 2916 from 2000. Current development activity in the ENUM working group is focusing on:

– continuously broadening services based on DNS,
– definition of new services in the DNS that use ENUM,
– separation of user ENUM and infrastructure ENUM,
– interworking issues with other IETF working groups like SPEERMINT.

An E.164 phone number provides universal access for phones, and using this number several value added services can be provided, like SMS or MMS. The Universal Resource Identifier (URI) supports communication between computers that are connected to the Internet. The assumption is that telephony services and Internet services will coexist for a long time, so in order to establish synergy there is a need for a standardised gateway between the traditional telephone services and Internet services. In this respect ENUM is one of the mechanisms that ties together the two worlds of communication systems, as convergent services are provided by the applications. It is well known that the first *Voice* over the Internet was realised in the mid-nineties. The SIP IETF signalling standard is dating back to 1999 (RFC 2546, RFC 3261).

By 2008 the term convergence became commonplace as service platforms that are merging continuously. However, there are warnings that make people and businesses very cautious. An announcement of Nominum [1] in March 2005 assured the public that Nominum's DNS solution for ENUM were more than satisfactory. A study [2] of several US ISP's DNS service concluded that the SLA of DNS services needs significant improvement.

In this article we will assess if ENUM as a technology is mature enough to be deployed and the concerns over ENUM's performance are substantiated.

2. ENUM measurement scheme

Applications that need ENUM make DNS requests and interpret the responses.

The overall set-up is very simple: there is a DNS server that answers the ENUM requests *(Fig. 1)*. DELL 1855 blade servers were used, with a blade of 2 CPU-s. There were two blades that participated in the measurements; it provided very good physical coupling, and compact arrangement. The blades were dually connected to the network but this does not have any impact as the level

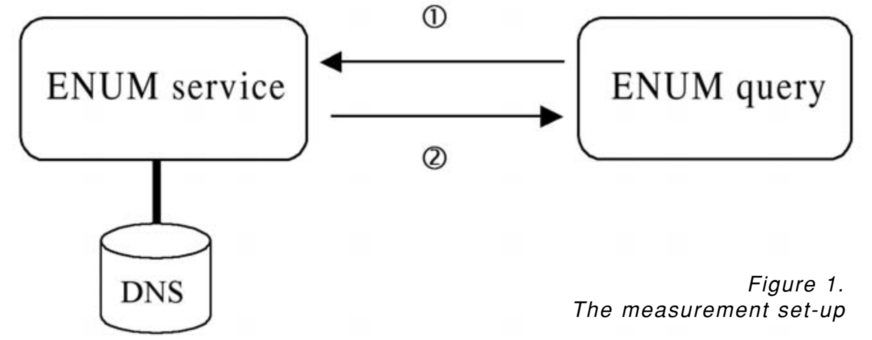


Figure 1. The measurement set-up

of network traffic was much less than the 2*1 Gbps capacity of the network interface card. The blades each have 2 Gbyte of RAM and two Intel 3.2 GHz hyper threading Xeon processors, the operating systems were Linux with 2.6 kernel.

Our obvious choice as DNS server was BIND 9 [3]. The software that was used for issuing the requests was Nominum's *dnsperf* [4]. The same software was used as in Nominum's case, so the differences caused by different software requester packages were eliminated.

Common elements of the measurements are that different DNS zone files were generated by a home made software. The zone files in the measurements represented different complexities in terms of ENUM responses.

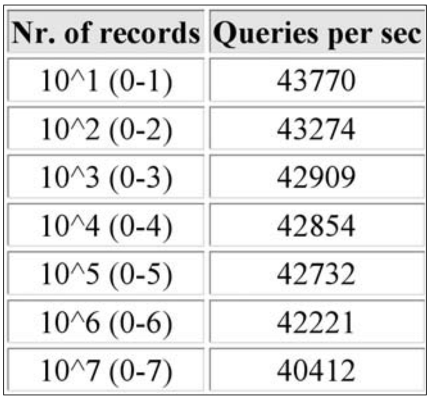
Our aims with the measurements were to determine:
– the number of responses per second of DNS servers that are loaded with zone files being different in ENUM resolving complexities
– what other parameters affect the number of served DNS requests per second

3. ENUM performance measurements

3.1. DNS responses by number of records

This is a baseline performance measurement. The DNS zone files did not contain any ENUM specific values. The result was that a simple DNS server without any tuning could respond more than forty thousand requests per second *(Table 1)*. Compared with the original Nominum press release and article [2] the expected throughput was much lower.

Table 1. DNS performance differences by number of records

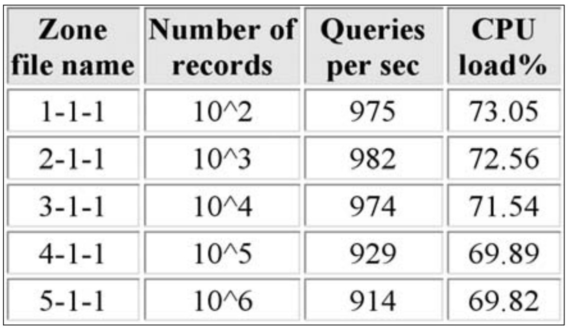


Nr. of records	Queries per sec
10^1 (0-1)	43770
10^2 (0-2)	43274
10^3 (0-3)	42909
10^4 (0-4)	42854
10^5 (0-5)	42732
10^6 (0-6)	42221
10^7 (0-7)	40412

3.2. DNS server performance and CPU capacity

For this test the BIND DNS was tested running on an old Intel P3 machine.

Table 2. DNS performance with low capacity CPU



Zone file name	Number of records	Queries per sec	CPU load%
1-1-1	10^2	975	73.05
2-1-1	10^3	982	72.56
3-1-1	10^4	974	71.54
4-1-1	10^5	929	69.89
5-1-1	10^6	914	69.82

Due to memory limitations only one million records could be loaded to the server. The sixth test reassures us that ENUM needs cannot be met by very low performance CPU-s *(Table 2)*. It also indicates why DNS could also have been a bottleneck in the early stages of the development of the Hungarian Internet.

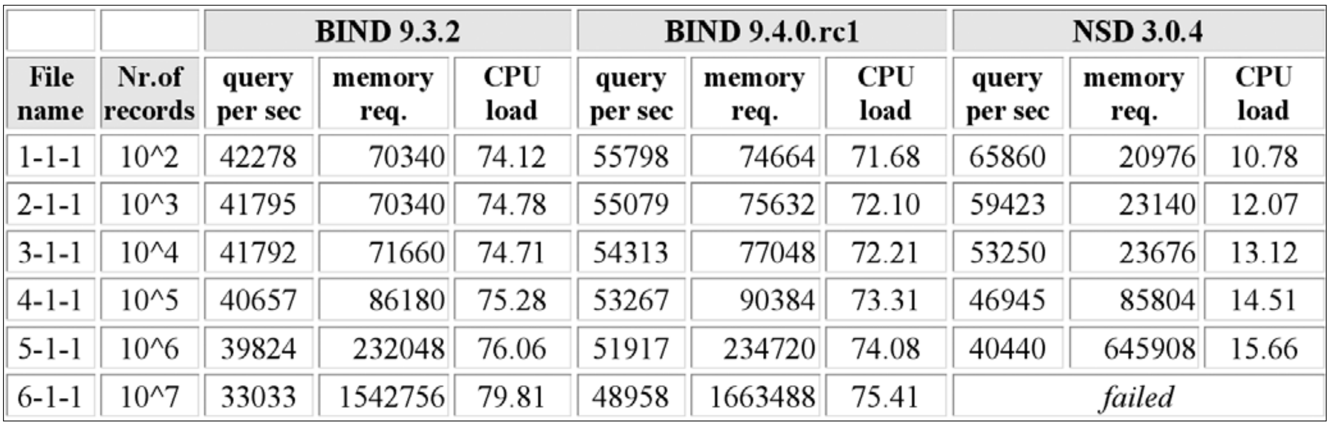
3.3. Resolving ENUM records by different type of DNS servers

The performance difference *(Table 3.)* of the two BIND versions is attributed to different software versions, and whether local optimization of the code was allowed or the pre-packaged version was used. NSD [5] is the open source version of the root name servers.

We conclude something very trivial: if the number of NAPTR records is growing than the DNS performance is slightly decreasing. NSD and BIND 9.4.0 are roughly equivalent in performance apart from the problem of the inability of NSD of loading ten million records.

The message of this test is that a simple DNS server providing ENUM records could surpass the 40000 resolutions per second. In our case the DNS servers utilised the modern Linux kernel with SMP and multithreaded functionality. It is good to keep in mind that an authoritive DNS server that is responsible for lots of ENUM records needs more memory, in our case the usable memory was 2 Gbyte.

Table 3. DNS servers' performance resolving simple ENUM records



		BIND 9.3.2			BIND 9.4.0.rc1			NSD 3.0.4		
File name	Nr.of records	query per sec	memory req.	CPU load	query per sec	memory req.	CPU load	query per sec	memory req.	CPU load
1-1-1	10^2	42278	70340	74.12	55798	74664	71.68	65860	20976	10.78
2-1-1	10^3	41795	70340	74.78	55079	75632	72.10	59423	23140	12.07
3-1-1	10^4	41792	71660	74.71	54313	77048	72.21	53250	23676	13.12
4-1-1	10^5	40657	86180	75.28	53267	90384	73.31	46945	85804	14.51
5-1-1	10^6	39824	232048	76.06	51917	234720	74.08	40440	645908	15.66
6-1-1	10^7	33033	1542756	79.81	48958	1663488	75.41	*failed*		

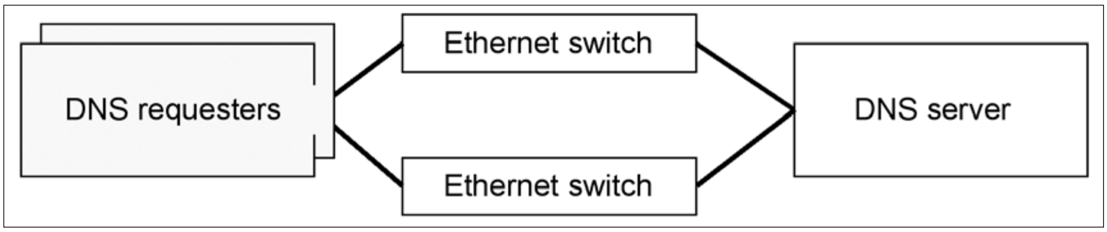
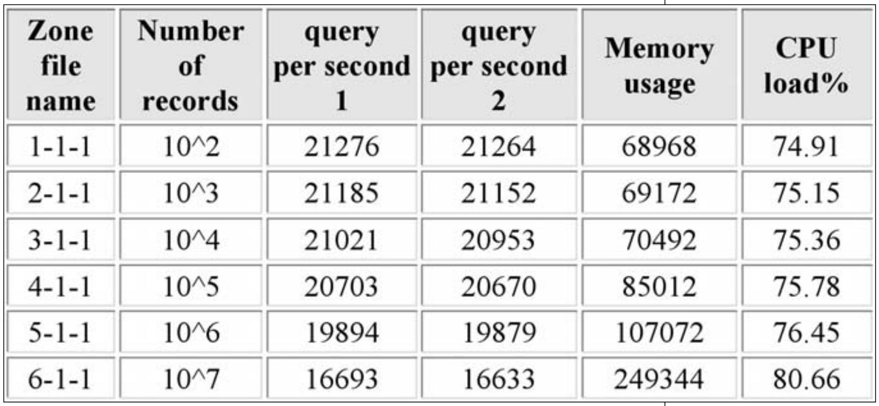


Figure 2. Test measurement arrangement

3.4. The effect of parallel requests on DNS server performance

The performance of the DNS server was measured with requests from two computers at the same time *(Fig. 2)*. The NAPTR (ENUM) structures were the same as in the previous measurement. From the results we conclude that the server handles the requests independently and the aggregated performance is the same as in the previous test *(Table 4)*.



Zone file name	Number of records	query per second 1	query per second 2	Memory usage	CPU load%
1-1-1	10^2	21276	21264	68968	74.91
2-1-1	10^3	21185	21152	69172	75.15
3-1-1	10^4	21021	20953	70492	75.36
4-1-1	10^5	20703	20670	85012	75.78
5-1-1	10^6	19894	19879	107072	76.45
6-1-1	10^7	16693	16633	249344	80.66

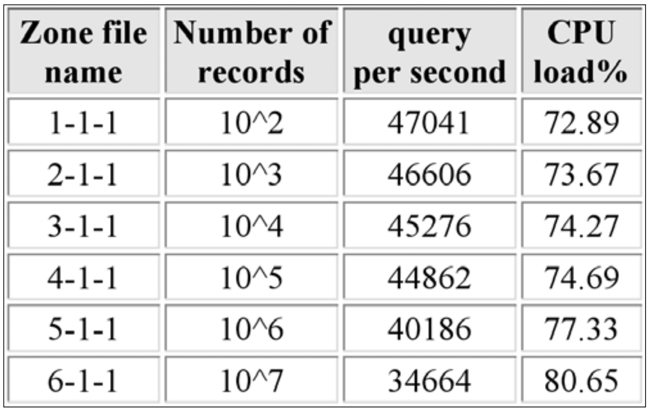
Table 4. DNS performance with dual load

3.5. Serving non-existing DNS records

In this measurement the performance of the DNS server was tested against non-existing records.

For BIND 9.3 implementation there was only a slight degradation of the performance for serving non-existent records *(Table 5)*.

Table 5. Performance data serving non-existing DNS records



Zone file name	Number of records	query per second	CPU load%
1-1-1	10^2	47041	72.89
2-1-1	10^3	46606	73.67
3-1-1	10^4	45276	74.27
4-1-1	10^5	44862	74.69
5-1-1	10^6	40186	77.33
6-1-1	10^7	34664	80.65

3.6. The effect of EDNS0 on BIND performance

The original DNS used UDP protocol with a maximum of 512 bytes of payload. When it turned out that the DNS responses might grow over 512 bytes, two solutions were introduced. One of the solutions is the DNS over TCP protocol but its drawback is the performance penalty, which results in slow responses. The other option is to allow longer responses up to 4096 bytes. When a DNS client is able to handle longer responses, it is indicated with an OPT element, so the DNS server can respond with longer UDP records.

In this measurements the effect of longer responses, with EDNS0 are assessed. In this particular test the structure of the ENUM record becomes gradually more complex in respect of the *response size*. During the measurement the number of records and the size of the NAPTR records are increased. It is a real life test in terms of usage, as it is equivalent the assignment of several `sip://`, `mailto:`, `IM`, etc. records to the same phone number. This measurement scenario represents a big provider with user ENUM enabled. The standard also allows the truncation of DNS response to 512 bytes provided the server was not specifically asked to respond with long records if it was needed.

The tests were carried out with and also without EDNS0 support.

As our measurements show the BIND DNS server performance depends on EDNS0. There is a slight decrease in performance, provided the responses are bigger. Although the whole range of possible response sizes in this test were not measured, the results show that with moderate long size DNS responses the performance is realistic *(Table 6)*.

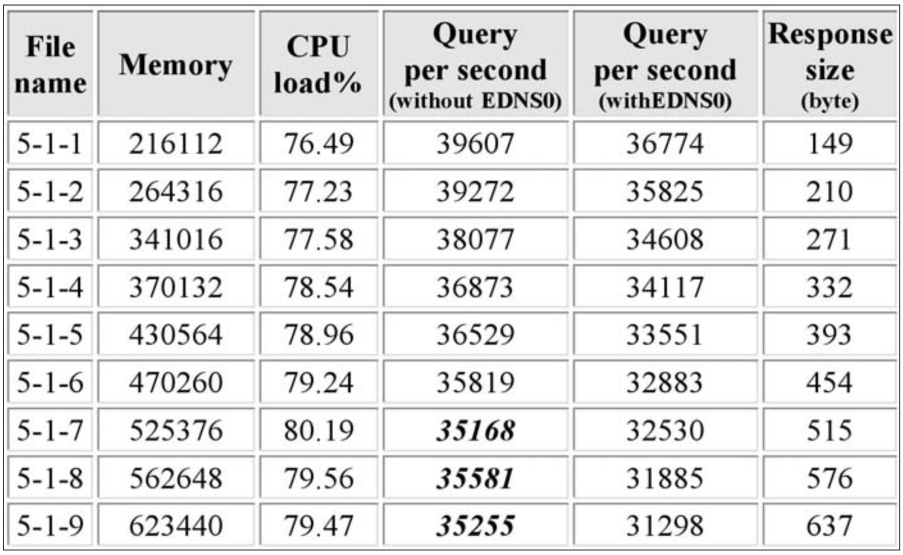
3.7. How the response size affects DNS performance

In this measurement the DNS response size is increased for two data sets, and the DNS performance is analysed.

According to the DNS Response Size Issues internet draft the DNS response size can be larger then the original 512 byte limit maximum.

Our aim is to find out how the response size affects ENDS0 operations.

In one of the data sets there were 100 records, with different sizes up to 4096, in the other data set there were 100,000 records and the response size changed accordingly *(Table 7).* The results of this test show that the response size affects the performance *heavily:* with growing response size the DNS performance becomes significantly less.



File name	Memory	CPU load%	Query per second (without EDNS0)	Query per second (withEDNS0)	Response size (byte)
5-1-1	216112	76.49	39607	36774	149
5-1-2	264316	77.23	39272	35825	210
5-1-3	341016	77.58	38077	34608	271
5-1-4	370132	78.54	36873	34117	332
5-1-5	430564	78.96	36529	33551	393
5-1-6	470260	79.24	35819	32883	454
5-1-7	525376	80.19	***35168***	32530	515
5-1-8	562648	79.56	***35581***	31885	576
5-1-9	623440	79.47	***35255***	31298	637

Table 6. The effect of EDNS0 on BIND performance

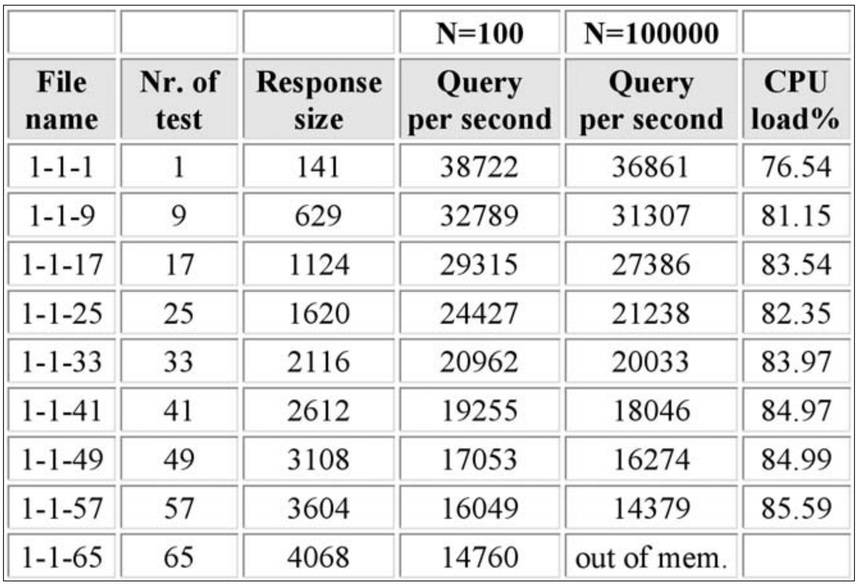
This concludes to the following design rule: for bigger DNS performance requirements one needs higher speed CPU-s, and higher speed memory access. The need for longer NAPTR records is expected in the future, as user ENUM registration is becoming a service.

3.8. DNS update performance and ENUM

The purpose of the measurement is to get performance details of DNS server update capability. The update/second is the scale of the measurements *(see Table 8)*.

The update performance of BIND DNS for very low record numbers is relatively high. It is assumed that this is due to some kind of internal cache mechanism. For bigger ENUM sets the DNS performance gets relatively low, and the performance is almost independent from the number record within the DNS zone files. The measurements show that the results are almost identical with newer BIND DNS. The increasing rate of IO WAIT-s show that the upgrade limit is Linux kernel related. The tuning of IO WAIT-s is a possible follow-up of these measurements.

We conclude that DNS update operations is not one of the strong points of BIND 9. Provided we assume the slow change of ENUM records, this update rate allows 1.7 million changes per day/server. This is substantial, although most probably it is not enough for very large customer base and for applications with very heavy change rate. If we stick to the original ENUM idea, the measured rate is definitely enough, as ENUM data is *static* like data on a business card.

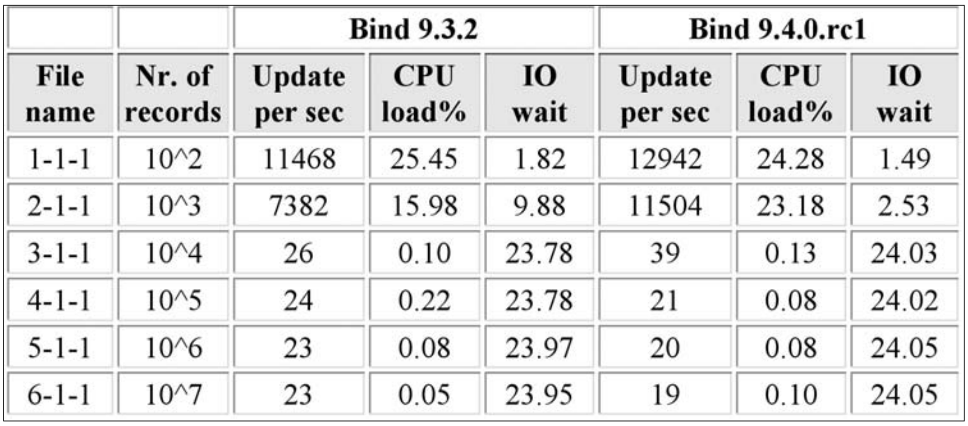


			N=100	N=100000	
File name	Nr. of test	Response size	Query per second	Query per second	CPU load%
1-1-1	1	141	38722	36861	76.54
1-1-9	9	629	32789	31307	81.15
1-1-17	17	1124	29315	27386	83.54
1-1-25	25	1620	24427	21238	82.35
1-1-33	33	2116	20962	20033	83.97
1-1-41	41	2612	19255	18046	84.97
1-1-49	49	3108	17053	16274	84.99
1-1-57	57	3604	16049	14379	85.59
1-1-65	65	4068	14760	out of mem.	

*Table 7.
The DNS performance change against the number of records and length of response*

3.9. Comparison of the measured results with other sources

1. NLnetlabs published its BIND 9 measurements [6] in October 2005. Their results are comparable to ours. NLnetlabs conclusion is that BIND requires modern 2.6 Linux kernel for higher performance operations.



		Bind 9.3.2			Bind 9.4.0.rc1		
File name	Nr. of records	Update per sec	CPU load%	IO wait	Update per sec	CPU load%	IO wait
1-1-1	10^2	11468	25.45	1.82	12942	24.28	1.49
2-1-1	10^3	7382	15.98	9.88	11504	23.18	2.53
3-1-1	10^4	26	0.10	23.78	39	0.13	24.03
4-1-1	10^5	24	0.22	23.78	21	0.08	24.02
5-1-1	10^6	23	0.08	23.97	20	0.08	24.05
6-1-1	10^7	23	0.05	23.95	19	0.10	24.05

*Table 8.
The BIND DNS server update performance*

2. Several publications were published during the summer of 2006, with the key message that the core of the problem of Internet applications responsiveness is the slow DNS answers. Our view is that Nominum started a media campaign in 2006, and journalists got the wrong message, or only half of the picture. An Australian Internet forum clearly attributes the wrong message to Nominum [7].

3. BIND DNS has been with us in the last 15-20 years, it has been updated successfully, and its scalability and reliability are its most advantageous points. Our view is the BIND DNS is capable to serve well mid-ranged ISP-s, till the customer base reaches 10 million. So BIND DNS – and not just alone – is a real, free alternative of Nominum DNS server.

4. Finally, have a look at the original Nominum press release [8]:

> *Running on commodity hardware*, Nominum's Foundation Authoritative Name Server (ANS) answered to 45,000 queries per second against 200M NAPTR records with an average latency of 2 milliseconds. Nominum's ANS outperformed by as much as four times all the other tested softwares. The company is also hosting a demonstration of its ENUM solution and benchmarks during the VON Conference in San Jose, California.*
>
> ** DNS servers were running*
> *on the following configuration:*
> *Red Hat Enterprise Linux 3.0,*
> *Intel Pentium XEON 2.4 GHz, 2 GB RAM,*
> *160 GB Raid 5 Disk array,*
> *Gigabit Ethernet Interface.*

This announcement clearly indicates the advantages of Nominum. The reason for BIND showing so low performance is the 2.4 Linux kernel. Our measurements show, that by early 2007 in a new environment BIND performs almost equally to Nominum.

5. If we concentrate on the update performance [9] of Nominum DNS server, this is also in the same range as that of BIND – 30 updates/sec vs. 24 updates/ sec.

> *For example, Nominum tested the Navitas server with a load representative of production carrier environments: 200 millions records, 30 updates/ second, serving simultaneous queries.*

6. Obviously, there are several advantages of Nominum's DNS server:

- It needs much less physical memory, which is an advantage for huge zone files.
- It support DNS EPP protocol.
- There are several extensions that makes Nominum attractive for VoIP providers.
- Service providers very often have more trust in a commercial product than an open-source solution without official support.

Our conclusion is that Nominum's DNS server advantage for ENUM services is not purely in the given performance, as BIND DNS can reach that ENUM performance level too.

4. Deployment considerations of DNS servers supporting ENUM

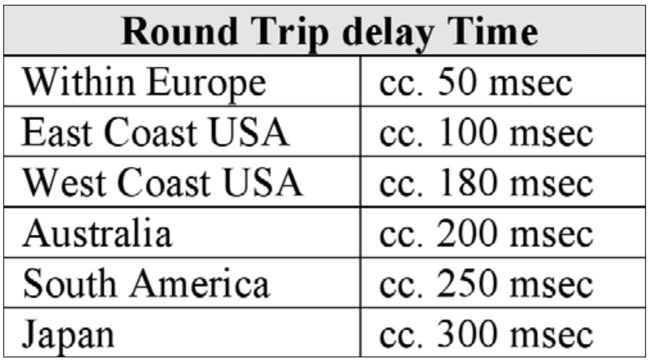
The primary aim of an ENUM DNS is to serve session setup with proper information that is in a DNS server and can be used within a certain time limit. ENUM is built upon DNS, the delay of the name resolution process has to be minimised.

DNS resolution time depends on:

a) The time the requester needs to issue the request.
b) The time the DNS request travels till it reaches the server that provides authoritive data.
c) The time the authoritive server needs to respond.
d) The DNS response transit time.
e) The processing time of the response at the requester .

In the previous sections ENUM performance data was presented that corresponds to point "c". Points "a" and "e" fully depend on the end user environment in case of user ENUM, therefore a telecommunication service provider cannot affect these time parameters. Due to the modern environment it can be safely assumed that "a"+"e" is smaller than 2-5 msec.

The transit times are constrained by the global IP networks, it cannot be significantly improved at the moment *(Table 9)*.



Round Trip delay Time	
Within Europe	cc. 50 msec
East Coast USA	cc. 100 msec
West Coast USA	cc. 180 msec
Australia	cc. 200 msec
South America	cc. 250 msec
Japan	cc. 300 msec

Table 9. RTT as a lower estimate of "b"+"d"

Assume that an average DNS server could resolve 20000 ENUM requests/second, it is equivalent to 0.05 msec. This processing time is negligible compared to the request travelling times within the network.

One consequence of the above mentioned observation is that the geographical distance is the crucial factor between the ENUM DNS server and its user. There are potentially 3 billion phone numbers, so international voice traffic based on simple ENUM requests might have serious problems with call set-up times due to geographic dispersity.

4.1. The effect of DNS service modernization

There has been a significant modernization in the DNS root name servers responsiveness. The aim was that although the number of root name servers is limited to thirteen still a sort of geographically dispersed DNS service should be available for the end users.

The solution is characterized by the so-called Anycast groups [10]. The DNS servers that participate in the Anycast service have the same IP address, and with the help of the properly configured BGP routing protocol this solution allows to find the nearest member of the Anycast group. There are several studies that summarize the effectiveness of the deployment of Anycast services [11].

The consequence of this modernization is that it is possible to deploy DNS servers around the world with phone numbers of the e164.arpa or ie164.arpa domains. Independently of the global Anycast service, the DNS cache servers and secondary servers allow resilient and quick reach of ENUM records for smaller communities.

If and when the Hungarian Regulatory Authority (HRA) starts the Hungarian ENUM trial, its central reference database would prove an important source of information for all the registered phone numbers in Hungary. This central reference database should be used for infrastructure ENUM. It will be HRA, or another organization selected and authorized by NHH that will play the role of the central ENUM registry. This registry will build upon a service that utilizes the Anycast DNS for the e164.arpa and ie164.arpa zone. Hungarian ISP-s, VoIP providers and telecommunication companies could build a very effective service on this basis of ENUM DNS.

4.2. What is the right sized DNS server for ENUM in Hungary?

The purpose of this section is to find out the performance requirements of the future Hungarian ENUM service based on publicly available data. If our aim is to access each Hungarian phone via ENUM, one needs a properly sized ENUM DNS service.

The basic question is: what is the call/second value that corresponds to the Hungarian voice traffic? Our estimation is based upon the Hungarian Central Statistical Office Quarterly Press Release on voice traffic [12].

In this quarter:

– there were 640 million PSTN calls,
– there were 1724 million mobile calls,
– the total number of initiated calls were:
 2364 million,
– the length of the quarter was 92 days,
– the average number of calls per day was:
 25.7 million,
– the average number of new calls were:
 297.4 per second.

The statistics does not give detailed background about the distribution of the calls. It is assumed that the Poisson distribution is applicable, as it is often used in telecommunications.

The Poisson distribution is:

- x = [0;23]
- Λ = 13,7

 (with this value the call distributions gives back
 99.915% of the total voice calls,
 with peek hours between 10-15.)

The diagram in *Fig. 3.* corresponds to the daily average call distribution of the Hungarian voice calls.

The peak hour is at 11 a.m., that corresponds 3 million calls/hour that is equal to *833.33 calls/second*.

Obviously the demand is distributed unevenly on workdays. It is assumed that on a very busy day the peak hour might take three times higher load than the average. That is equivalent of 2500 calls/second. To be on the safe side let us estimate the topmost DNS resolution requirement is 8000 calls/second. As this is an

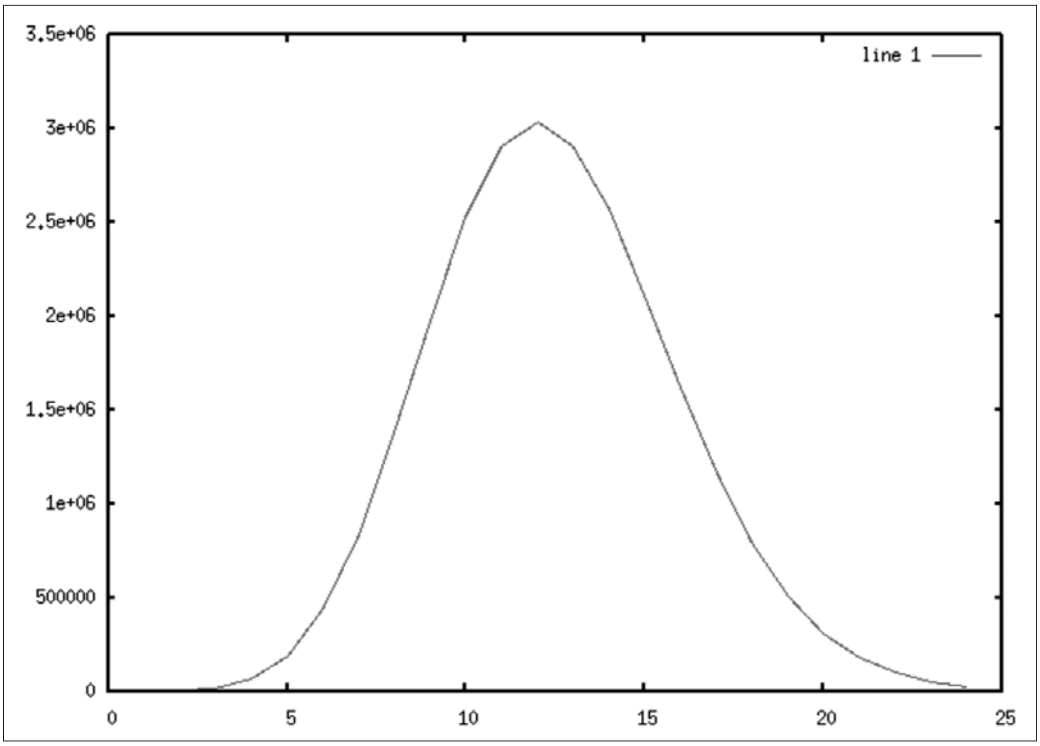


Figure 3. Estimated call initiation per hour in the Hungary

aggregate value, it is distributed back to telecommunication service providers, to smaller demand values.

Our final conclusion is that for voice call establishment, the ENUM (DNS) requirements for the Hungarian population in a converged network could easily be met by current server computer hardware and software. The technical problems that were shown in the introduction are not relevant in Hungary, so ENUM related services could be introduced and be part of everyday practice. The Hungarian telecommunication industry can firmly build on ENUM technology and utilize the advantages of this new opportunity.

5. Summary

The article is a summary of performance measurement results that were obtained during testing different DNS servers loading with predefined ENUM structures. Several parameters were identified that affect the performance of domain name resolution between E.164 phone numbers and records in e164.arpa domain.

Our conclusions of the measurements are the following: high quality ENUM resolution can be provided for a Hungarian sized population with PC category servers and open-source software packages; ENUM resolution primarily depends on the geographical distance between the caller and the recipient; higher ENUM DNS resolution demand can be addressed with clustering for bigger populations.

The technical conditions and practical experiences to introduce ENUM are available, the new convergent services will appear soon.

References

[1] http://www.nominum.com/popupPressRelease.php?id=338 (retrieved 26 July 2007.)

[2] http://www.lionbridge.com/competitive_analysis/reports/nominum/Nominum_2006_03_DNS_Survey_v3.1.pdf (retrieved 26 July 2007.)

[3] http://www.isc.org/index.pl?/sw/bind/

[4] http://www.nominum.com/testing_tools.php (retrieved 26 July 2007.)

[5] http://www.nlnetlabs.nl/nsd/

[6] http://www.nlnetlabs.nl/downloads/bind9-measure.pdf

[7] http://www.nik.com.au/archives/2006/08/19/346/

[8] http://www.nominum.com/popupPressRelease.php?id=338

[9] http://www.nominum.com/getFile.php?file=nominum_wp_enum.pdf

[10] http://root-servers.org/ and http://en.wikipedia.org/wiki/DNS_root_zone

[11] ftp://ftp.ripe.net/ripe/docs/ripe-393.pdf

[12] http://portal.ksh.hu/pls/ksh/docs/hun/xftp/gyor/tav/tav20609.pdf (retrieved 23 December 2007.)

Authors

István Tétényi graduated at the Technical University of Budapest, at the Measurements and Instrumentation Department in 1977, got his "Dr. Univ." degree at 1986, has been working for the Computer and Automation Institute since 1977 as a researcher and as a head of department since 1991. Before the embargo was lifted, he participated in several equipment development activities for telecommunication purposes that mainly addressed X.25 and IBM networking. Till end of 2006 he participated in the development and management of the Hungarian Academic and Research Network as the Head of the Technical Steering Group and also in several international research networking projects. His recent activities focus on mobile communications. He is co-author of the "Internet világa" book that was published in 1998.

Gyula Szabó graduated in 1999 at University of Pécs at the Pollack Mihály Faculty of Engineering. He has been working for the Computer and Automation Institute since 2001. He participated in several maintenance and development projects that focus on network monitoring, web applications. His recent activities belongs to authentication and authorization based on LDAP technology and federated identity management.

The evolution of Grid Brokers: union for interoperability

ATTILA KERTÉSZ

Institute of Informatics, University of Szeged
MTA SZTAKI Computer and Automation Research Institute
keratt@inf.u-szeged.hu

Reviewed

***Keywords:* Grid Brokers, grid interoperability, Meta-brokering, matchmaking**

Grid resource management is probably the research field most affected by user demands. Though well-designed, evaluated and widely used resource brokers, meta-schedulers have been developed, new capabilities are required, such as agreement and interoperability support. Existing solutions cannot cross the border of current middleware systems that are lacking the support of these requirements. In this paper we examine and compare different research directions followed by researchers in the field of Grid Resource Management, in order to establish Grid Interoperability. We propose a meta-brokering approach, which means a higher level resource management by enabling communication among existing Grid Brokers and utilizing them.

1. Introduction

Grid Computing has become a detached research field in the '90s and since then it has been targeted by many world-wide projects. Several years ago users and companies having computation and data intensive applications looked skeptically at the forerunners of grid solutions, who promised less execution times and easy-to-use application development environments by creating a new high performance network system of interconnected computers from all around the world. Research groups were forming around specific middleware components and different research branches have grown out of the trunk.

Many user groups from various research fields (biology, chemistry, physics, etc.) put their trust in grids, and today's usage statistics and research results show that they were undoubtedly right. Grid Computing is in the spotlight, several international projects aim at establishing sustainable grids (CoreGRID [1], LA Grid [2], Globus [3], etc.).

Nowadays research efforts are focusing on user needs: more efficient utilization and interoperability play the key roles. Grid resource management is probably the research field most affected by user demands. Though well-designed, evaluated and widely used resource brokers have been developed, new capabilities are required, such as agreement (Service Level Agreements, WS-Agreements [4]) and interoperability support. These two directions also depend on other grid middleware capabilities and services, and since they can hardly cross the border of these middleware solutions, they need revolutionary changes affecting the whole system. Solving these problems is crucial for the next generation of grids, which should rise up from the academic to the business world.

To achieve this, capabilities such as advance reservation and co-allocation need to become reality, but the currently used grid middleware solutions do not provide these services. Therefore, usually estimations and predictions are used in the scheduling process of the resource managers to overcome these lacking features and provide a more efficient schedule. (For example, Lőrincz et. al. monitor runtime information to determine the behavior of the job and use these additional data in scheduling [14]). Trying to enlarge the limitation borders, in this paper we are focusing on interoperability approaches in the field of Grid Resource Management.

The current solutions of grid resource management will not be able to fulfill the high demands of future generation grid systems, though several grid resource brokers [5] have been developed supporting different grid systems. Their main problem is that most of them cannot cross the borders of current grid middleware solutions, therefore the newly arisen problems need to be treated with novel research approaches. Nowadays grid systems have their own researchers and user groups. This means borders not only for the development but also for the interoperable utilization.

Figure 1. illustrates the current grid utilization: grid resources can be accessed either directly by using grid middleware components or through grid brokers that help finding a proper execution environment or through grid portals that provide a convenient user interface to grid services.

The need for interoperability among different grid systems has raised several questions and directions. The advance of grids seems to follow the way assigned by the Next Generation Grids Expert Group, which has been established by the European Commission. In their latest publication [6] they have pointed out that grid and web services are converging and envisaged a new hybrid architecture called SOKU (Service Oriented Knowledge Utility), which enables more flexibility, adaptability and advanced interfaces, therefore interoperability is evident and congenital in these systems.

Following these expert guidelines and the latest requirements of grid user groups, in this paper we propose a grid resource management solution that does not require major changes of the whole grid middleware and still provides interoperability.

2. Resource management and matchmaking in grids

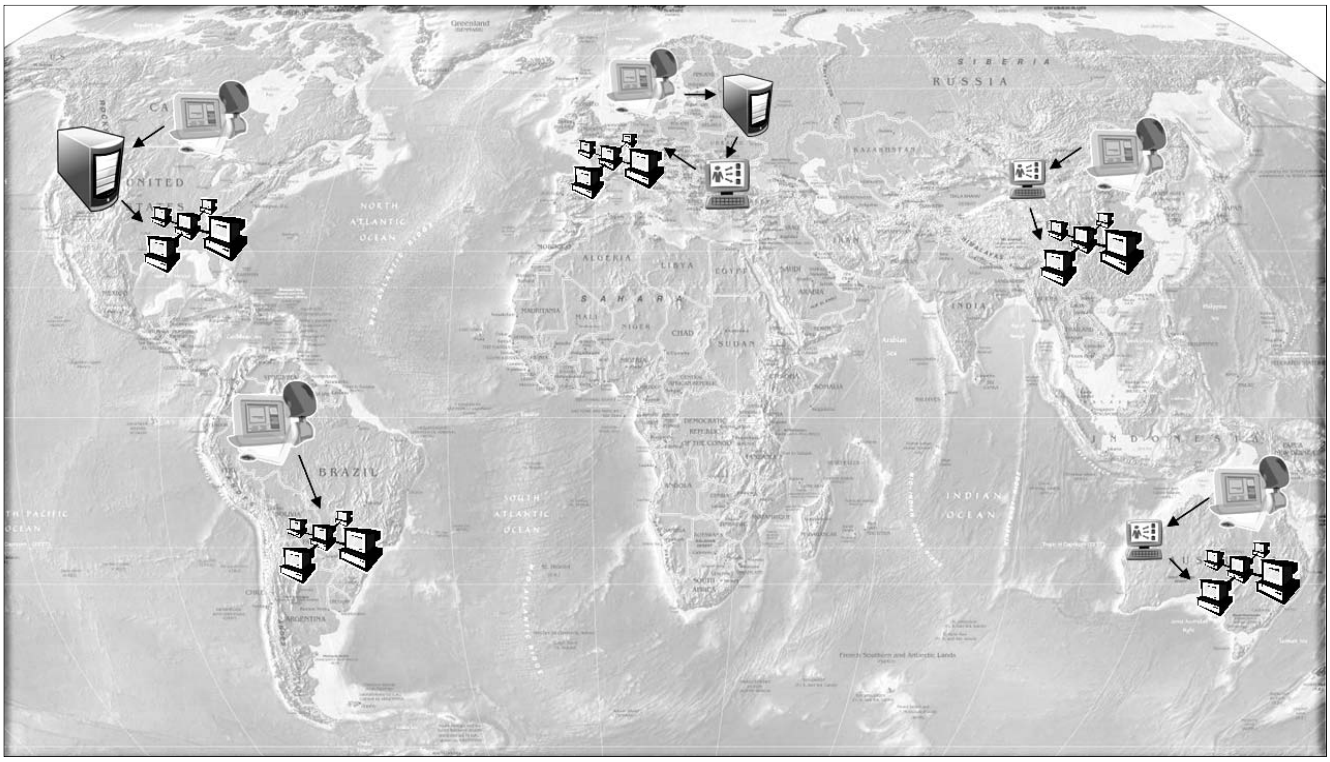
When grids were born, resource management components of the middleware provided various interfaces to submit jobs, transfer files, query resource information, track job states and retrieve execution results. As grids and the number of users were growing, the dynamicity and heavy load made users unable to cope with manual resource selection. Automatic matchmaking between user requests and available resources came into view and resource brokers were born. This additional component contacts the Information System of the grid and schedules user requests to a proper execution environment, computing resource (most of the time *proper* means likely the fastest execution). In addition to contacting the resources, transferring the jobs, tracking the states and staging back the result files are also the tasks to be performed by the brokers. Describing the job requirements is done with a middleware-specific language (in general job description language).

This document needs to be submitted to the broker with the necessary input files and executable. This is the first time where interoperability problems appear. If users wanted to use different grid solutions, they need to use different description forms for the same requirements. Furthermore grids use different protocols to store resource information, transfer files, access resources, etc., though they implement the same methodology. Knowing these facts it is not surprising that users and developers have started to form separate user groups and developer communities around various grid solutions. Because of the same reasons resource brokers generally support one grid middleware and its job description language, therefore they are tightly coupled to that middleware.

Up to now most of the broker developers identified this separation and have started to redesign and extend their solutions for multi-middleware and multi-language support to provide a basic level of interoperability. Though carrying out these extensions take much time and still in progress nowadays, several solutions are ready to serve different user communities of different grids. The additional components for understanding other language descriptions and using other protocols make the extended brokers more and more robust and unmanageable. These redesigns are usually done for similar description forms and protocols, or for middleware solutions having common components. These observations show that broker extensions cannot be done for all the available middleware solutions, and the more grids an extended broker supports, the more failures can occur during its utilization.

Another possibility to enhance interoperability is the use of grid portals. These tools provide easy to use graphical interfaces to utilize various grid components. There are general purpose and specialized ones for supporting specific applications. An instance of the second approach is the Conflet framework (CONFigurable

Figure 1. The utilization of production Grids



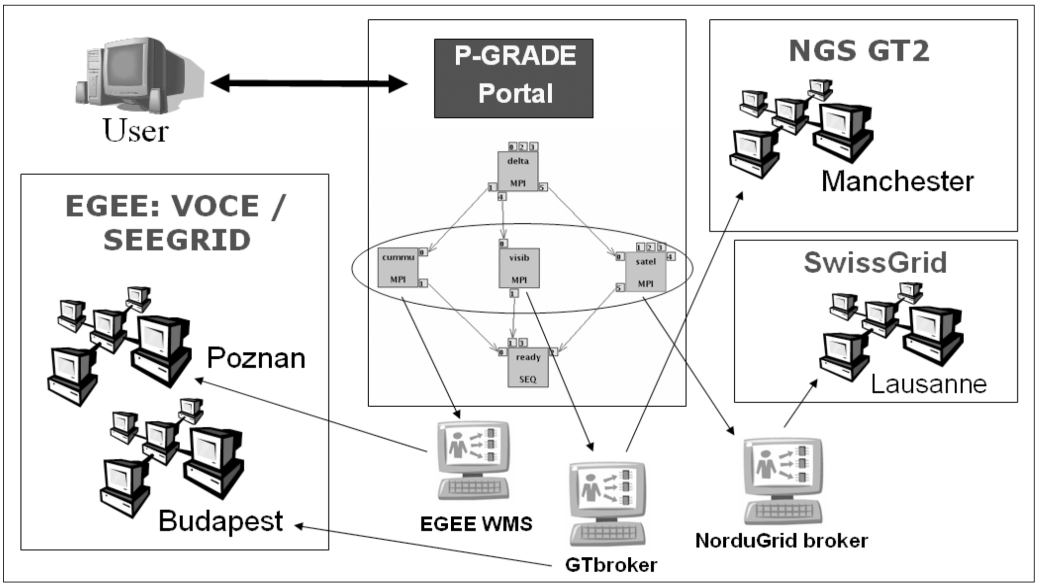


Figure 2. Multi-grid and multi-broker support in the P-GRADE Portal

portLET [13]), which can be used to create specific portlets to one's application. Interfacing different brokers to portals is another option to extend interoperability and support more middleware solutions. Nevertheless these portals also attract other user communities and provide more computational power. In *Figure 2.*, we can see how the P-GRADE Portal [7] supports various production grids by interfacing different resource brokers.

The P-GRADE Portal is a workflow-oriented grid portal with the main goal to support all stages of grid workflow development and execution processes. It enables a graphical design of workflows created from various types of executable components, executing these workflows in Globus-based computational grids [3] relying on user credentials, and finally analyzing the monitored trace-data by the built-in visualization facilities. In the Portal box field of Fig. 2., the bigger boxes represent the executable files, jobs (delta, cummu, etc.), the smaller numbered boxes (ports) on their sides represent input and output files. Connecting these ports the user can create an application of dependent jobs, which together form a workflow. In the last step of the workflow edition the user can select brokers or resources to the jobs. During the execution these brokers take care of the execution of the jobs, or they are directly submitted to the manually selected resource. The disadvantage of this solution is the same as in the previous case: interfacing additional brokers requires modifications to the system.

3. The evolutionary step: unifying Grid Brokers

Facing with the problems stated in the previous section, several research groups turned their attentions to new solutions to establish interoperability. It has become obvious that keeping the same architecture would not bring interoperability in the near future; they need to wait for revolutionary changes in the whole middleware to enable a world-wide interoperable grid, which would take long years. The only way to achieve a higher level of interoperability in reasonable time is to unify brokers by enabling communication and data-flow among them.

One of the biggest grid research organizations is the OGF (Open Grid Forum), which has many research groups to share innovative ideas and standardize solutions in various fields of Grid Computing. The GSA-RG (Grid Scheduling Architecture Research Group [8]) is currently working on a project enabling grid scheduler interaction. They try to define common protocol and interface among schedulers enabling inter-grid usage. Implementing such an interface and using it by all the brokers would enable sharing different user jobs, workloads. Agreeing on a common interface and implementing it to the brokers definitely takes a long time.

The other similar approach enables communication among the same broker instances. Since in this case no negotiation is needed with other researchers and solutions, it is easy to agree on an interface and the implementation needs to be done only for their own solution. (Note that in this case different protocols will be created and used by different developers, again.) This approach is followed by the following projects: Koala [9], LA Grid [2] and Gridway [10].

Comparing these approaches we can see that all of them use a new method to expand current grid resource management boundaries. The interconnected domains are being examined as a whole, and they delegate resource information among broker instances managing different domains. Usually the local domain has preference and when it is overloaded, some jobs are passed to somewhere else – in this case the results should be passed back to the initial domain. Though this is a novel approach and all of them proved that they achieved better load balancing, the interoperability problem among different systems is still not solved.

The final solution lies in meta-brokering. This approach means a higher level brokering, which uses the existing resource brokers to reach different grids. Unlike existing brokers it uses metadata about the available broker capabilities and maps user requests to brokers not to resources. In order to achieve this we need to store and consume metadata about user jobs and resource brokers. The OGF has already developed a standard language for describing jobs - this is the JSDL (Job Submission Description Language [11]).

Regarding broker capabilities we designed a BPDL (Broker Property Description Language [12]) description format together with researchers from the Barcelona Supercomputing Center. Scheduling at the metabrokering level requires additional metadata about the scheduling requirements of the users and scheduling properties of the brokers. Since the JSDL is lacking these attributes and the BPDL incorporates some of them, we decided to create a separate language called MBSDL (Meta-Broker Scheduling Description Language). Here we mention that the OGF-GSA-RG [8] has started to define an SDL (Scheduling Description Language) for enabling the aforementioned inter-broker communication, but they have not got too far, yet. We believe that MBSDL can be regarded as a contribution to their work. Once SDL becomes a standard, our system will be ready to use it.

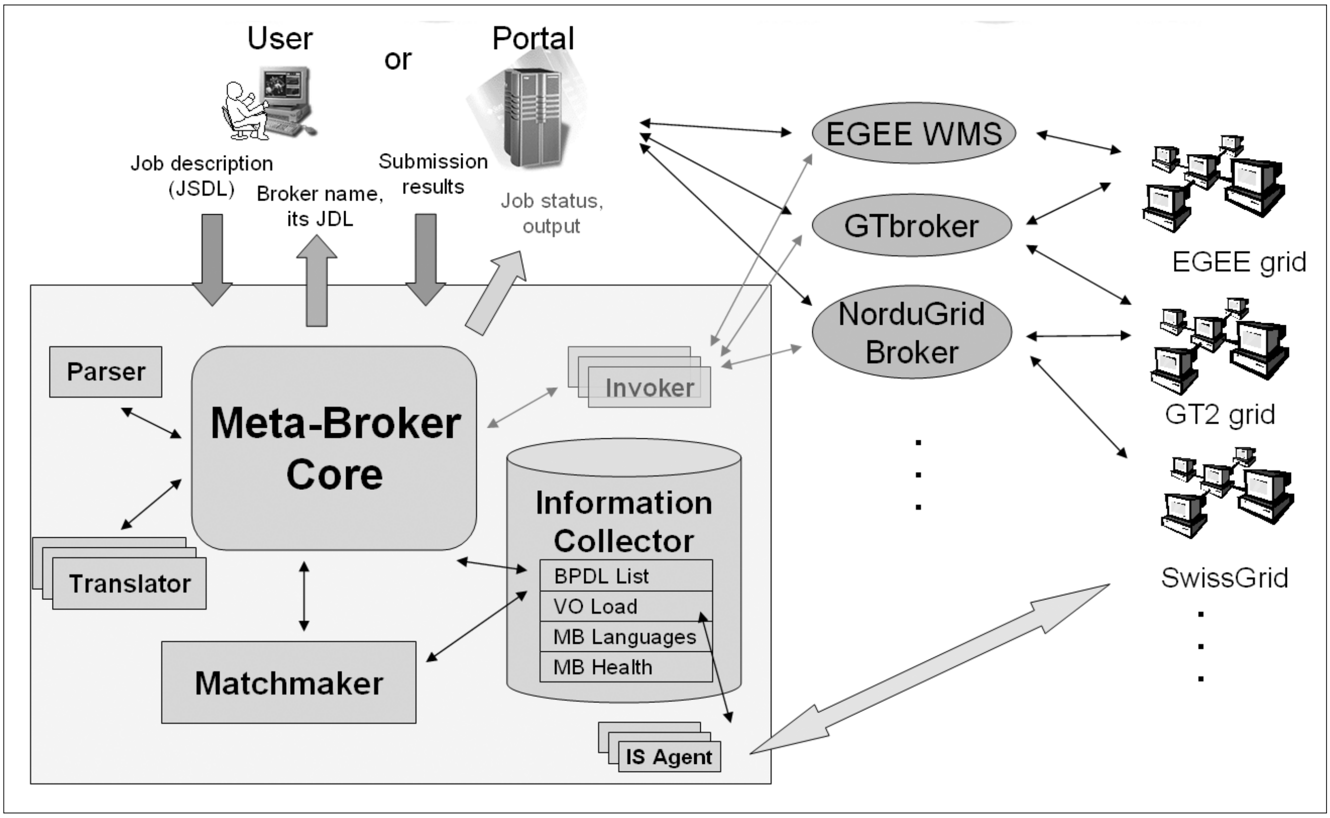
Using these tools we developed a meta-brokering service as a general web-service and named it as Grid Meta-Broker (shown in *Figure 3*). Together with BPDL and MBSDL metadata can be stored about resource brokers in the Information Collector (IC) component of the system. The users can specify their requirements with JSDL and MBSDL. Consuming these documents the MatchMaker component executes a scheduling algorithm to select a broker and a grid for the actual user job. So-called IS Agents are used to provide up-to-date information about the background grid load to help the MatchMaker skip grids with overloaded or unavailable resources. The next step is to translate the user request to the language of the selected broker and let the Invoker submit the job contacting the underlying broker. This component is responsible for tracking job states through the broker and retrieving the result files and logging information. The final step is to provide the results to the user and update the IC with broker performance data.

Another scenario can also be done, when instead of the Invoker the user or a portal contacts the selected broker and does the actual job submission. In this case they need to report the submission results themselves to the Meta-Broker.

4. Conclusions

We have learned several reasons why existing resource management systems cannot fulfill the newly arisen requirements of grid users. Providing bigger computation power and serving business-oriented investments requires a novel, higher level approach in grid resource management: we need to unify the separated grid islands and manage them together. Extending the current resource brokers with multi-middleware support or

Figure 3. The architecture of the Grid Meta-Broker



interfacing them by widely used grid portals can be a good starting point, but in the long run they become unmanageable and vulnerable.

More successful solutions have been developed by enabling inter-broker communication among specific broker instances operating in different domains of the same grid. Though it brings some level of interoperability, these brokering systems still cannot work together, do not have common interfaces.

The final solution for grid interoperability is the Grid Meta-Broker, which has been built on the latest standards of web and grid technologies taking into account the guidelines of NGG. The goals of meta-brokering are to use the widespread resource brokers to manage their own grids and to provide an intelligent way to unify these brokers and offer it to the users as a transparent multi-grid service.

In a unified, world-wide grid (WWG [15]) these Meta-Brokers will bridge the yet separated islands of grids and serve the whole user community in a fully interoperable manner.

Acknowledgement

This research work is carried out under the FP6 Network of Excellence CoreGRID funded by the European Commission (Contract IST-2002-004265).

References

[1] http://www.coregrid.net
[2] http://www.latinamericangrid.org/
[3] http://www.globus.org/
[4] http://www.ogf.org/documents/GFD.107.pdf
[5] A. Kertész, P. Kacsuk:
A Taxonomy of Grid Resource Brokers, 6th Austrian-Hungarian Workshop on Distributed and Parallel Systems (DAPSYS 2006) in conjunction with the Austrian Grid Symposium 2006, Innsbruck, pp. 201-210, Austria, September 21-23, 2006
[6] Next Generation Grids Report:
Future for European Grids: GRIDs and Service Oriented Knowledge Utilities –
Vision and Research Directions 2010 and Beyond, December 2006 (NGG3).
[7] A. Kertész, G. Sipos, P. Kacsuk:
Multi-Grid Brokering with the P-GRADE Portal,
In post-proc. of the Austrian Grid Symp. (AGS'06), OCG Verlag, Austria, 2007.
[8] https://forge.gridforum.org/sf/projects/gsa-rg
[9] A. Iosup, D. H.J. Epema, T. Tannenbaum, M. Farrellee, M. Livny:
Inter-Operating Grids through Delegated MatchMaking, In proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'07), Reno, Nevada, November 2007.
[10] T. Vazquez, E. Huedo, R. S. Montero, I. M. Llorente:
Evaluation of a Utility Computing Model Based on the Federation of Grid Infrastructures,
pp.372–381, Euro-Par 2007, 28 August 2007.
[11] http://www.ogf.org/documents/GFD.56.pdf
[12] A. Kertész, I. Rodero, F. Guim:
Data Model for Describing Grid Resource Broker Capabilities, CoreGRID Workshop on Grid Middleware in conjunction with ISC'07 Conference, Dresden, Germany, 25-26 June 2007.
[13] D. Pasztuhov, I. Szeberényi:
A new architecture of the Conflet system,
Networkshop 2007, (in Hungarian).
https://nws.niif.hu/ncd2007/docs/ehu/036.pdf
[14] L. Cs. Lőrincz, A. Ulbert, Z. Horváth, T. Kozsik:
Towards an Agent Integrated Speculative Scheduling Service,
6th Austrian-Hungarian Workshop on Distributed and Parallel Systems (DAPSYS'2006),
pp.211–222, Innsbruck, Austria, 21-23 Sept. 2006.
[15] P. Kacsuk, A. Kertész, T. Kiss:
Can We Connect Existing Production Grids into a World Wide Grid?,
8th International Meeting High Performance Computing for Computational Science (VECPAR'08), Toulouse, France, 24-27 June 2008. (Submitted)

Author

Attila Kertész is a PhD student at the University of Szeged, Hungary and also a researcher at the Laboratory of Parallel and Distributed Systems of MTA SZTAKI Computer and Automation Research Institute, Hungary. He graduated as a program-designer mathematician, his research interests include grid brokering, scheduling and web services. He is participating two leading European projects S-Cube and CoreGRID Network of Excellence, and also a member of the CoreGRID Institute on Resource Management and Scheduling.

Saleve: toolkit for developing parallel Grid applications

PÉTER DÓBÉ, RICHÁRD KÁPOLNAI, IMRE SZEBERÉNYI

Budapest University of Technology and Economics, Centre of Information Technology
{dobe,kapolnai,szebi}@iit.bme.hu

Reviewed

Keywords: Saleve, Grid application development, EGEE, parameter study

We present the Saleve tool, which helps the migration of existing parameter study applications into grid environment. Programs linked against the Saleve library can be integrated into grids using different middleware systems, so the application developer need not deal with the technical details of the middleware.

1. Introduction

In our days we have numerous resources available for running scientific computations, yet in many cases their usage is not encouraged by an adequate support. Although a large amount of development has been carried out, a researcher has many challenges to face in order to benefit from a distributed, parallel computational system.

Our intention is to ease these difficulties by presenting the *Saleve framework* which provides a virtual layer over the underlying infrastructures for the developer of a parallel application. Saleve focuses on implementing a specific type of parallel algorithms called *parameter study* programs. The parallel applications developed using Saleve can be executed on several different distributed infrastructures without any modification or recompilation.

In the next section we give a quick overview of the parameter study problems and of the EGEE which is the most important Grid project in the EU. Then we introduce the motivation and objective of Saleve, and we continue by outlining some details of the operation of the Saleve system. Finally we give a summary and present some future plans.

2. Utilizing the Grid for parameter study tasks

2.1. Parameter study tasks

In practice, there is frequently a need for an algorithm to be run with hundreds or even thousands of different input parameter values: such tasks are called parameter studies or parameter scans. In certain cases the requested result is the set of outputs obtained using each parameter, but the end result is often acquired via a final aggregating step. For instance, in an exhaustive optimization this last step is to seek one specific parameter value. Another simple example is the numeric integration of a non-analytic function over a given domain. We can split the domain into non-overlapping subdomains which will be used as the input parameter for the integrating routine, and as the final step we add up all the integral parts.

The problem of PS emerges in several experimental sciences, especially in physical simulations but also in other areas such as high energy physics, astrophysics, genetics, biomedical research and seismology. Like parameter studies, it is easy to split into subtasks all the engineering problems that can be described with ordinary differential equations. Such problems include statics tasks, like the research project at BME that involves the planning of reinforced concrete bridge beams using a parallelizable algorithm similar to PS [1].

The large number of executions, each with different parameters, would however take very long time when done sequentially. On the other hand, we should note that there is no causal relation among the runnings, so the chronological order of these is arbitrary, and they can even be done parallel according to the Single Program, Multiple Data (SPMD) model [2]. Therefore the presence of multiple CPUs can be taken advantage of, either in the form of a multiprocessing system or a cluster of several nodes – and in the best case, we can even utilize a grid infrastructure for this purpose.

2.2. The current state of Grids

In order to satisfy the rapidly increasing demand for computing and data storage, the plan for a geographically distributed network of resources called the grid [3] emerged more than a decade ago. Since then, several initiatives all over the world have been launched to implement it.

Among these, *Enabling Grids for E-sciencE* (EGEE) [4] is the largest in Europe. It was initially built to process data from the sensors of the Large Hadron Collider (LHC) at CERN, Switzerland, but now it has a wide range of scientific applications, for example in astrophysics, bioinformatics and geophysics. The project brings together more than 240 institutions from 45 countries including Hungary. Currently the grid consists of approximately 41,000 CPUs, can store up to 5 petabytes of data and can process 100,000 concurrent jobs.

BME participates in EGEE, too, by network activities and also by providing resources. In our own grid site called BMEGrid, there are currently 8 quad-core server machines which execute the jobs submitted into the grid. Connected to the site, we have a high capacity, efficient, parallel accessible storage system, the Scalable File Share (SFS), which is capable of storing nearly 3 terabytes of data. Our resources are mostly used by members of the Atlas HEP project and biomedical researchers.

Grid projects focus on facilitating the development of new applications, thus recruiting more grid users. For this purpose, portals [1,10], sometimes extended by development and workflow management tools [11], or other environments with complex functionalities [12] can provide a solution. The Saleve concept differs from these: using it, parallel applications can be created that are capable of transferring themselves to the runtime environment without separate tools, and besides staying lightweight, and can be run on a personal computer as well. The system, just like the aforementioned environments, is not specific to the application area.

3. Overview of Saleve

3.1. Motivation

Most of the researchers and engineers have programming skills, especially in C and Fortran languages, therefore creating sequential programs for performing scientific calculations means no obstacle to them. However, the development of distributed programs running in parallel requires more advanced knowledge and experience in programming.

The situation is made even more difficult by the large number of diverse, rapidly evolving technologies in use. These include different batch systems like PBS, LSF and Condor [5]. Although the final goal of grid development is a worldwide service that is accessible in a standardized way, its implementation is expected to delay several years yet. Currently each grid system has its own middleware system that is incompatible with the others. Getting acquainted with all these technologies and learning to use them in order to solve a general problem, for example PS, would take away effort from the real task of research. In addition, it is quite common that one would like to use an existing program without radical redesign in case the underlying computing infrastructure changes. Such change is for example switching from a single computer to a local cluster, or switching from the local cluster to the execution on a grid system.

3.2. A Proposed Solution

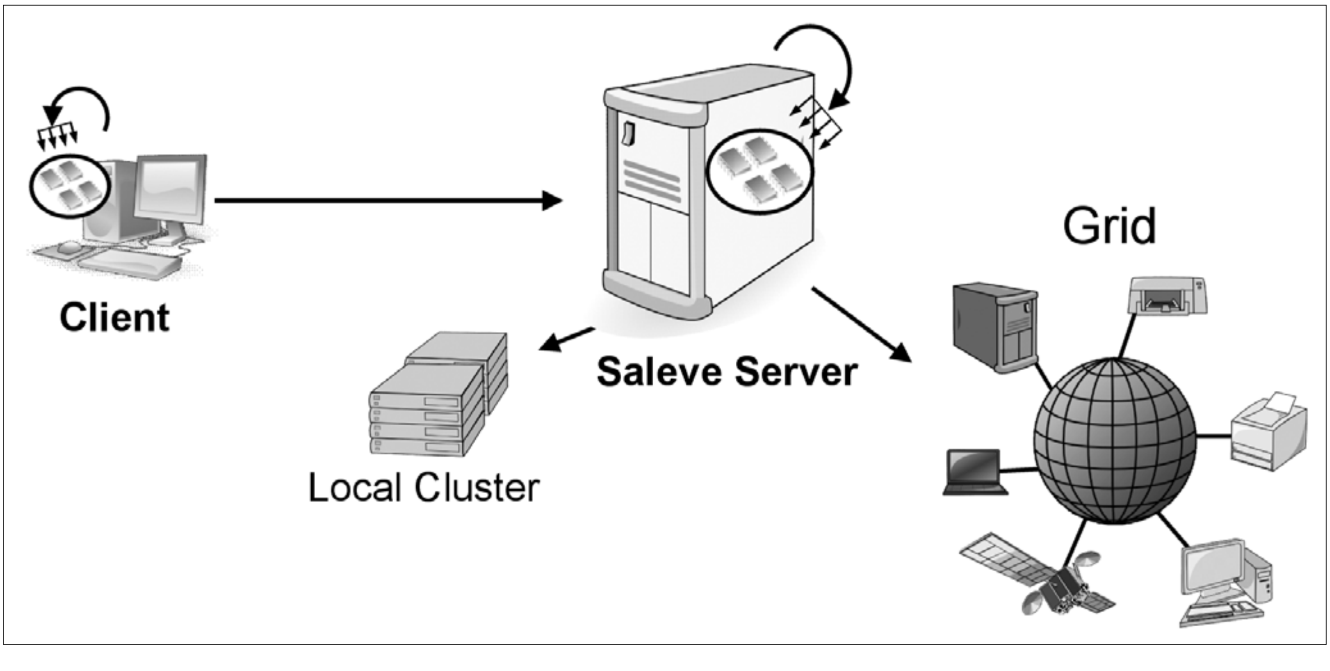
Handling these difficulties is the aim of the Saleve system, an open source tool to aid the development of C programs that are capable of running in parallel. Saleve can be used either to make new programs or to upgrade existing sequentially running ones. The main goal is to hide the details of the distinct computing technologies, and to provide an invariant, easy-to-learn methodology to create PS applications that, in addition to the simple sequential execution, are also able to take advantage of parallel computing systems.

4. Design of Saleve

4.1. Client-Server Architecture

To understand the operation of Saleve we begin with the derivation of the Saleve client. For that purpose, let us consider a traditional, sequential PS application writ-

Figure 1. The features of Saleve



ten in C. From the user's point of view, the only duty is that the original program has to be gently transformed into the Saleve client. First, a more exact structure of the program should be defined i.e. the following modules have to separated: the partition of the parameter space, the calculation over a subdomain and the summary of the subresults. Finally the resulting source code has to be linked against the Saleve programming library.

The client can be launched to compute the subresults over the subdomains and summarize them. By default, the running is similar to the original program: the subresults are computed consecutively, but the client can be configured to launch parallel jobs locally to finish faster in a multi-CPU machine. However, the most important feature of the client is the ability of transmitting itself and the input to a given Saleve server.

After receiving the executable client binary and its input data, the Saleve server forwards a new job for each subdomain to a Grid or to a cluster or simply executes it instead of dispatching. We emphasize that the user is not aware at all which distributed system the jobs were forwarded to, therefore Saleve provides full transparency *(Fig.1)*.

The server monitors the submitted jobs, resubmits them on failure and stores the temporary subresults. Under this phase the client may disconnect from the server and later another instance may resume the session from a different machine.

The server continuously returns the available subresults to connected client. As soon as each subresult has been returned, the client computes the final result from the subresults as the user defined.

Fig. 2. illustrates the course of events during the lifecycle of a task.

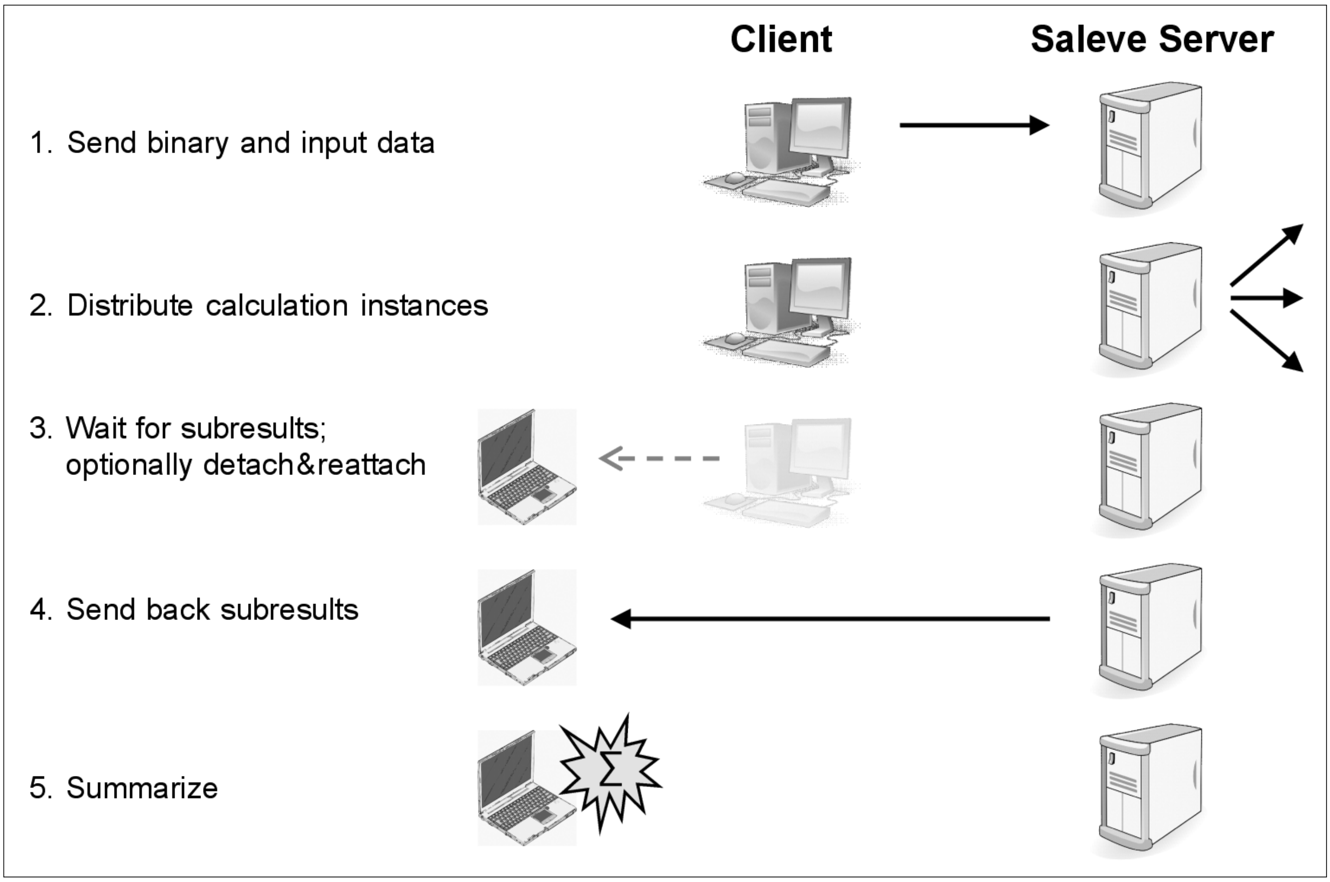
4.2. The Architecture of the Server

To meet our main requirements, the server should support the most popular distributed computing environments and, moreover, it should be easy to extend to operate with a new scheduler or grid middleware. This approach has led to the split-up of the server components into two groups: the components of the first group serve general purposes which are independent of any specific computing environment, the second group contains the components named the *plugins*.

One of the generic components implements the communication interface based on SOAP. The server provides web services towards the clients to upload a task and the input data and to download the subresults. The web services of Saleve are built on the gSoap [6] implementation. The generic group includes more components such as the one responsible for user management or job management which are discussed in detail in [7,8].

The plugins make it possible to the server to cooperate with several different infrastructures, in addition, adapting the server to a new infrastructure would not

Figure 2. Execution of a task in Saleve


Client
Saleve Server
1. Send binary and input data
2. Distribute calculation instances
3. Wait for subresults;
optionally detach&reattach
4. Send back subresults
5. Summarize

involve any change in the generic components. For this reason, a dedicated plugin for every distributed system handles the envorinment-specific communication *(Fig. 3)*.

Up to present the following plugins have been made:

– executing jobs parallel on the server host,

– submitting jobs to the Condor scheduler [5] which is widely spread on cluster sites,

– forwarding jobs to the EGEE grid infrastructure through the gLite middleware.

4.3. Interoperation between the EGEE infrastructure and Saleve

Saleve supports submitting tasks to the EGEE grid with the help of the gLite plugin mentioned above. Developing a new Saleve plugin principally requires knowledge of the interface of the corresponding middleware or scheduler, does not presume deep experience in Saleve internals.

To create a new plugin, one has to implement an interface where the data management is aided by the Saleve development library. The major challenge is dealing with the authentication issues towards the grid and the job management rather than using the Saleve library. Looking after the grid jobs is essential due to the instability of the current infrastructures: some jobs may abort and must be submitted again.

The gLite middleware which is the software engine of the EGEE infrastructure has adopted certificate-based authentication and resource-allocation methods where the permissions of a user are determined by his or her memberships in virtual organisations (VOs). When a user wishes to access to a resource e.g. by submitting a task, a temporary, short range proxy certificate has to be generated using the long range certificate, has to be attached to the submitted task and periodically renewed. This procedure is useful to protect the long range certificate if the proxy certification was compromised.

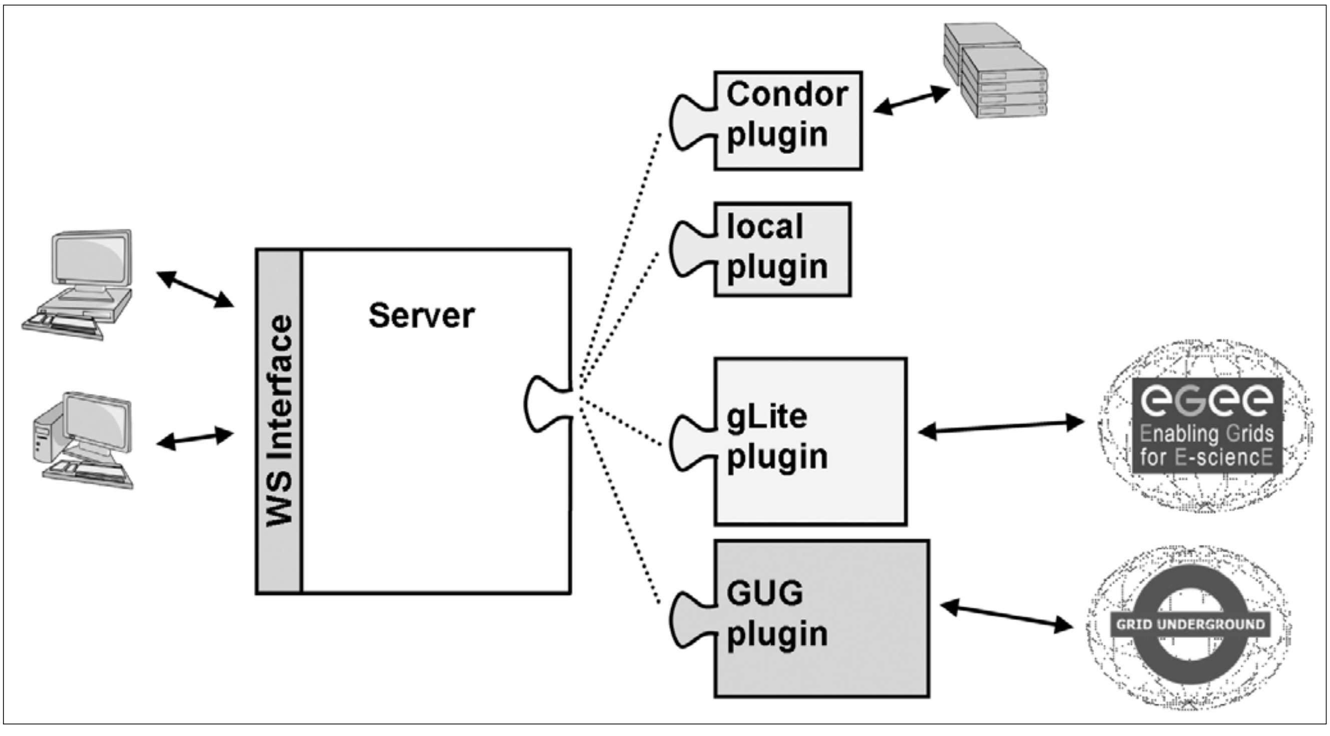
In our current implementation the Saleve server keeps an own certificate for accessing grid resources directly thus the server is a member of some virtual organisation. In this manner the proxy generation and renewal is completely hidden from the user who cannot tell whether the task has been executed in the EGEE grid or in a local cluster.

5. Summary

The presented Saleve system aids the development of parameter study type parallel applications by forming a transparent abstraction layer above the middleware or batch systems of the different distributed environments. Its main advantage is that a slightly modified version of the legacy application called the Saleve client can be run without change in several types of runtime environment and even on the local machine, so it is easy to develop applications possibly for EGEE, the largest infrastructure of Europe, without knowing the technical details.

Regarding the possible upgrades of the system, we are going to focus on implementing the dynamic loading and unloading of plugins and a better management of jobs submitted into the grid. Our plans also include improving the flexibility of client-server communication with the help of webstreams. For more information on the current status of the Saleve project, please visit the web page [9].

Figure 3. The architecture of Saleve



Acknowledgements

Part of this work was funded by the Péter Pázmány program (RET-06/2005) of the National Office for Research and Technology, and the authors would like to thank EGEE project (EU INFSO-RI-031688) and NKFP MEGA (2_009_04) project.

References

[1] D. Pasztuhov, A. Sipos and I. Szeberényi: Calculating Spatial Deformations of Reinforced Concrete Bars Using Grid Systems, MIPRO 2007 – Hypermedia and Grid Systems, Opatija, Croatia, 2007. pp.189–194.

[2] P. E. Black: Algorithms and Theory of Computation Handbook, CRC Press LLC, U.S. National Institute of Standards and Technology, 1999.

[3] I. Foster and C. Kesselman: The Grid 2: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers Inc., 2003.

[4] EGEE-II Information Sheet, 2007. http://www.eu-egee.org/sheets/uk/egee-ii.pdf

[5] D. Thain, T. Tannenbaum and M. Livny: Distributed Computing in Practice: the Condor Experience, Concurrency and Computation: Practice and Experience, 2005. pp.323–356.

[6] R. A. van Engelen and K. Gallivan: The gSOAP Toolkit for Web Services and Peer-To-Peer Computing Networks, In the Proc. of the 2nd IEEE Int. Symposium on Cluster Computing and the Grid (CCGrid'02), Berlin, Germany, 2002. pp.128–135.

[7] Zs. Molnár and I. Szeberényi: Saleve: simple web-services based environment for parameter study applications. In Proc. of the 6th IEEE/ACM International Workshop on Grid Computing, 2005. pp.292–295.

[8] P. Dóbé, R. Kápolnai and I. Szeberényi: Simple grid access for parameter study applications. In 6th International Conference on Large-Scale Scientific Computations, Sozopol, 2007. (in press)

[9] Saleve project, http://egee.ik.bme.hu/saleve/

[10] P. Kacsuk, Z. Farkas and G. Hermann: Workflow-level parameter study support for production Grids, Proc. of ICCSA'2007, Kuala Lumpur, Springer LNCS 4707, pp.872–885.

[11] P. Kacsuk, G. Dózsa, J. Kovács, R. Lovas, N. Podhorszki, Z. Balaton and G. Gombás: P-GRADE: A Grid Programming Environment. Journal of Grid Computing, Vol. 1, No. 2, 2004. pp.171–197.

[12] T. Fahringer, A. Jugravu, S. Pllana, R. Prodan, C. Seragiotto Jr. and H.-L. Truong: Askalon: a tool set for cluster and Grid computing, Concurrency and Computation: Practice and Experience, Vol. 17, 2005. pp.143–169.

Author

Richárd Kápolnai received his M.Sc. degree in technical informatics from Budapest University of Technology and Economics (BME), Hungary, in 2006. His M.Sc. thesis is about designing networks for selfish users. Currently he is pursuing a Ph.D. degree at the Centre of Information Technology, BME, in grid systems. He helped develop the Saleve system, and participated in the 2nd phase in the project EGEE. His research interests include supporting grid application development, and mechanism design.

Summaries • of the selected papers published in this issue

Periodic reconfiguration of groomed multicast trees in WDM networks

In a typical multicast scenario the tree members (users attached to the tree) change all the time. New users join the tree, while some existing users leave it. Here we consider these dynamically changing multicast trees in two-layer, grooming-capable, optical networks. The continuous changing of the tree members (users) causes the degradation of the tree. Therefore a huge amount of network resources can be spared by periodically repeated reconfiguration. In this paper the benefits of reconfiguration are investigated for different multi-cast routing algorithms and reconfiguration periods.

(In: 2007/8, pp.14–23.)

Incentive scheme for voluntary and autonomous cooperation in distributed networks

Today's communication networks are becoming increasingly dynamic in the sense that they do not have fixed infrastructure, or the configuration of infrastructure-based networks continuously changes. Examples include distributed access networks using WLAN technology, ad-hoc networks, ambient intelligence networks or sensor networks. These networks have considerable independence and autonomy and they might frequently act in a selfish manner. Autonomy means that such networks have no central administrative or management principles that would determine their operation.

(In: 2007/8, pp.38–42.)

Security API analysis with the spi-calculus

API level vulnerabilities of hardware security modules represent a serious threat, thus, discovering and patching security holes in APIs are important. In this paper, we argue and illustrate that the application of formal verification methods is a promising approach for API analysis. In particular, we propose an API verification method based on process algebra. The proposed method seems to be extremely well-suited for API analysis as it allows for the straightforward modelling of the API, the precise definition of the security requirements, and the rigorous verification of the security properties offered by the API.

(In: 2007/8, pp.43–48.)

Tactile sensing arrays – design and processing

Tactile sensors are commonly used in industrial, medical or virtual-reality applications, but the majority of commercial tactile systems are capable to detect pressure maps only. In this article we present a novel tactile sensing array that processes all three components (normal and shear) of the tactile information at every sensory element (taxel – tactile pixel). We describe the processing technology of the integrated microsensors, write about the information coding behaviour of its elastic cover and, finally, we show a robotic application example, where the three-component force measurements play a fundamental role.

(In: 2007/10, pp.47–52.)

Makyoh topography: a simple and powerful method for the flatness characterisation of semiconductor wafers

The paper presents our research in the field of Makyoh topography, a method based on an ancient principle. The method's application is the qualitative and quantitative study of semiconductor wafers and other mirror-like surfaces.

(In: 2007/10, pp.19–24.)

Solar Cell Technology Innovation Center at MTA MFA

This paper introduces to the reader one of the largest facilities of the solar cell research and development in Hungary – the Solar Cell Technology Innovation Center. The R&D equipment is an integrated vacuum system designed and built for the preparation of thin film Copper Indium Gallium diSelenide (CIGS) solar cell layer structures. The facility was built on the premises of the Hungarian Academy of Sciences by the Energosolar Co. in the frame of a main project funded by the Hungarian National Office for Research and Technology. This paper reviews the layout of the solar cell structure and the equipment for its preparation, introduces the main materials science issues raising in the CIGS system and presenting challenges for the research.

(In: 2007/10, pp.30–34.)

On the scaling characteristics of MMORPG traffic

In this paper a comprehensive scaling analysis of the traffic of the four most popular Massively Multiplayer On-line Role Playing Games (MMORPG) is presented. The examined games are World of Warcraft, Guild Wars, Eve Online and Star Wars Galaxies. Both server and client generated traffic are analyzed in details. Our study reveals the basic statistical properties of the investigated games focusing on the correlation and scaling behavior. Although the examined games are all from the same genre and basic statistics such as the mean packet rate, variation of the packet rate, skewness of the packet rate distributions fall into the same magnitude, the games exhibit diverse traffic characteristics. We have found that in spite of the fact that some similarities can be found among the scaling characteristics of these games they show versatile scaling properties and the games can not be treated with one common model.

(In: 2007/11, pp.20–26.)

ENUM in everyday practice – is it a dream or an opportunity?

The aim of this paper is to elaborate on Electronic NUmber Mapping technology. In the introduction, the role of ENUM is presented. Afterwards an ENUM measuring method is introduced, and several determining parameters are identified and it is shown how these parameters influence the performance of ENUM. The closing part shows overall ENUM and DNS performance parameters, apart from the DNS server raw performance. Finally, as a sanity check, the Hungarian voice communication profile is compared with the measured ENUM performance in order to have sizing guidelines for ENUM related services.

(In: 2007/11, pp.13–19.)

The evolution of Grid Brokers: union for interoperability

Grid resource management is probably the research field most affected by user demands. Though well-designed, evaluated and widely used resource brokers, meta-schedulers have been developed, new capabilities are required, such as agreement and interoperability support. Existing solutions cannot cross the border of current middleware systems that are lacking the support of these requirements. In this paper we examine and compare different research directions followed by researchers in the field of Grid Resource Management, in order to establish Grid Interoperability. We propose a meta-brokering approach, which means a higher level resource management by enabling communication among existing Grid Brokers and utilizing them.

(In: 2007/12, pp.21–25.)

Saleve: toolkit for developing parallel Grid applications

We present the Saleve tool, which helps the migration of existing parameter study applications into grid environment. Programs linked against the Saleve library can be integrated into grids using different middleware systems, so the application developer need not deal with the technical details of the middleware.

(In: 2007/12, pp.32–36.)

Networks 2008 will focus on the challenges of planning networks to deliver on the promise of Convergence and NGN. The challenges are many – how to build high performance networks for converged services where every step is cost justified and drives profitable growth, where difficult issues of scalability, end-to-end network performance, network management, network and service control, reliability, security and interoperability are planned and then realized, and where flexibility is maintained to allow experimentation with new applications that can foster new and compelling revenue streams for operators. Unleashing the core value of convergence and the reduction in the number of network platforms requires innovation in network planning methods, scalable architectures, new optimization algorithms, and understanding the tradeoffs between different technology choices and migration paths.

We are living in very interesting times right now. The revolution we had been hoping for is finally taking place. Mobile and fixed operators are moving to next generation networks. Hundreds of trials are underway. Many operators have made recent commitments, very significant investments towards their future vision, often reducing spending drastically on extensions to current infrastructures to allow spending on new platforms. At such times, industry professionals are on a high learning growth curve and there is an urgent need for forums where true learning can occur thru the exploration of diverse perspectives and opposing viewpoints, the presentation of leading-edge results in planning specific networks, and state-of-the art methods and tools.

Networks 2008 provides such an opportunity. It is the 13th of such symposia, held every two years, and attracting participants from all over the world – from network operators, to systems and software companies, to researchers from universities and industry, to systems integrators. At **Networks 2008**, we will continue the tradition of state-of-the art papers, invited presentations, and panel sessions, as international experts present their latest findings and share experiences in network strategy, planning, operations, management, control and design.

May we invite you to participate in Networks 2008!

THE NETWORKS 2008 WILL CONSIST OF:

Invited Talks and Plenary Panel Sessions on Hottest Topics

The plenary sessions of the symposium will be distinguished by presentations and debates of some of the foremost figures in the telecommunications sector.

Technical Contributed Sessions

The symposium will feature technical presentations on network strategy, planning, management, control and design.
They will be selected from submitted papers through peer review. Paper submission deadline: February 29, 2008.

Tutorials

Tutorial sessions on Monday, September 29 will provide education on various topics in network strategy, planning, management control and design. Experts from industry and science are invited to propose relevant contributions for the tutorial programme.
Tutorial proposals due: April 11, 2008.

Exhibition, Tools Demonstration

A tools demonstration is planned to give companies and universities the opportunity to present themselves and their services as well as latest developments in hardware and software. For more details, please, contact the Conference Secretariat.

NETWORKS 2008 KEY THEMES:

1 Convergence at Different Domains

a) Business strategy for convergence in competition
b) Economies of scale at network, services, access, terminals and operation domains
c) Triple play and Multiple play. Architectures and solutions
d) Fixed-Mobile-Nomadic convergence
e) Broadband Mobile-Broadcast/Multicast convergence
f) IMS architecture and applications solutions
g) Addressing and numbering issues
h) Example cases on convergence

2 Migration to NGN and Mobile Broadband

a) NGN and IMS architectures and solutions
b) Migration steps to NGN based on economical evaluation and security
c) New services driving evolution to NGN
d) Migrating to multimedia/multiservice environment
e) VoIP solutions: service, performance, cost and revenues impacts
f) IPTV demand and design
g) Migration alternatives from 2G to 3G and economical evaluation, Services driving 3G business
h) Interoperation across multiple domains
i) Regulation and interconnection issues in NGN, Network certification process

3 Routing, Traffic Flows and Optimization

a) Multiservice traffic measurement, analysis, characterization and simulation at network level
b) Multiservice flow matrix and aggregation methods
c) New signaling and control in multimedia/multisystems
d) P2P (peer-to-peer) traffic flows and implications on network demand
e) Impact of GRID services
f) Intra-Domain and Inter-Domain routing, TE (Traffic Engineering) and resilience
g) Optimization process with technical and economic criteria
h) Using Game Theory, Meta-heuristics and LP/ILP

4 Network Design and Planning Methods

a) Network design methods for multimedia services
b) Network topology design and optimization at different layers: physical, optical, media and control
c) Quality of Service, Quality of Resilience, performance and SLA
d) VPN design
e) Ad-hoc networking
f) Sensor networks and autonomic networks
g) End-to-End service performance evaluation
h) Network security
i) Network resilience: survivability, protection, restoration and availability
j) Cost modelling and pricing

5 Role of New Technologies, Developments and Standards

a) Optical switching
b) New generation SDH, OTN
c) New generation internet and IPv6
d) Carrier Grade/Class Ethernet: 100GEth, VPLS, PBB,
e) GMPLS, PCE
f) Mobile 3.5 and 4G, LTE (Long Term Evolution)
g) WiMax
h) xDSL and FTTx, Power Line Communication (PLC)
i) Broadcast/multicast systems. DVB-T/H

6 Network Planning Support Processes and Tools

a) Multilayer planning process
b) Business planning methods and tools
c) Optical network design tools
d) Access planning methods and tools
e) NGN planning methods and tools
f) 3G planning methods and tools
g) OSS and BSS processes and tools
h) Network Management support tools

AUTHOR GUIDELINES:

Each paper should clearly and concisely state the problem addressed, the methodology used, and the central conclusions. The title of the paper should convey the essence of the work. Papers must be unpublished. All papers will be reviewed by Technical Program Committee members and other experts active in the field to ensure high quality and relevance. Authors of accepted papers must attend the conference to present their contributions. Papers must be submitted electronically through the EDAS system (details will be available at the www.networks2008.org). Each submission must be accompanied by the following information: a short abstract (up to 300 words), a complete list of authors and their affiliations, a contact person for correspondence, postal and e-mail addresses, phone and fax numbers.

Please, indicate preferred topic areas, using the list of main topics (e.g. 3f). The total length of papers should be between 6 and 10 pages.

Accepted papers will be published by the HTE and the IEEE. HTE will own the copyright transferred to them by the authors. These papers will also be available electronically from IEEE Xplore (ieeexplore.ieee.org), the digital library of the IEEE.

Format requirements:
Guidelines for formatting the paper can be obtained from http://www.ieee.org/web/publications/authors/transjnl/

IEEE style files for Latex and templates for MS Word are also available:
Unix LaTeX2e Transactions Style File (TAR.GZ, 204 KB), Unix Bibliography File (TAR.GZ, 204 KB)
Template and Instructions on How to Create Your Paper (MS Word, 92 KB)

Best paper award and Journal publications:
The top 10% papers receiving highest reviewer score will be invited to related Journals. Based on scores and the oral presentation up to 3 papers will be chosen for the best paper award.

Important Dates:

Electronic Paper Submission	February 29, 2008
Tutorial Submission	April 11, 2008
Acceptance Notification for papers	May 9, 2008
Final Version of Accepted papers and copyright	August 1, 2008

NETWORKS 2008 IS ORGANIZED BY:

Department of Telecommunications and Media Informatics (TMIT)
Budapest University of Technology and Economics (BME) *(www.tmit.bme.hu)*

Scientific Association for Infocommunications (HTE) *(www.hte.hu)*

Information Technology Society, Association for Electrical, Electronic & Information Technologies (ITG-VDE) *(www.vde.com)*

Committees:

International Management and Scientific Committee (IMSC)

IMSC Chair: Gyula Sallai, *BME-TMIT and HTE, Hungary (sallai@tmit.bme.hu)*

IMSC Members:

Alberto Ciarniello, *Telecom Italia Mobile, Italy*
Oscar Gonzalez-Soto, *ITU Consultant, Spain*
Joachim Gross, *Arcor, Germany*
Wolfgang Gross, *Telekom, Germany*
Hideaki Yoshino, *NTT, Japan*
Bernard Jarry-Lacombe, *France Telecom, France*
Sang-Baeg Kim, *Korea Telecom, Korea*
Hussein T. Mouftah, *University of Ottawa, Canada*
Lawrence Paratz, *Telstra Corporation, Australia*
Michal Pióro, *Warsaw University of Technology, Poland*
Rati C. Thanawala, *Lucent Technologies, USA*
Andy Valdar, *University College London, UK*

International Programme Committee (IPC)
Chair:
Tibor Cinkler, *BME-TMIT, Hungary (cinkler@tmit.bme.hu)*

Conference Secretariat
HTE
Address: Kossuth Lajos tér 6-8, H-1055 Budapest, Hungary
Phone: +36 1 353 1027 *E-mail:* info@hte.hu

National Advisory Board (NAB)
Chair:
László Pap, *BME-HIT and HTE, Hungary (pap@hit.bme.hu)*

Local Organizing Committee (LOC)
Chair:
Péter Nagy, *HTE, Hungary (nagy.peter@hte.hu)*

Technical co-sponsors:

CALL FOR PAPERS AND PARTICIPATION

NETWORKS 2008

13th International Telecommunications Network Strategy and Planning Symposium

„Convergence in Progress"

DANUBIUS HEALTH SPA RESORT MARGITSZIGET
BUDAPEST, HUNGARY
September 28 – October 2, 2008

www.networks2008.org