Mixed adaptation and fixed-reservation QoS for improving picture quality and resource usage of multimedia (NoC) chips

Milan Pastrnak, Peter H.N. de With, *Senior Member*, IEEE, Calin Ciordas
Jef van Meerbergen, *Senior Member*, IEEE, Kees Goossens, *Member* IEEE

Abstract - *Advanced video systems running multiple applications require an efficient distribution of system resources. An adaptable computing system can be created with a reconfigurable Network-on-Chip (NoC). Execution of multiple multimedia processing tasks implicitly ask for a reservation-based Quality-of-Service (QoS) control. However, pure reservation-based systems have rigid rules for assigning resource budgets, leading to a slow reaction time on activity change or a long reconfiguration time. In this paper, we present an application-specific QoS solution that combines a reservation-based approach with a run-time adaptation facility. We demonstrate this framework by mapping an MPEG-4 arbitrary-shaped video decoder on an NoC of eight ARM cores with specific monitoring features in the network (e.g. Æthereal NoC). First, we have found that our advanced QoS can save up to 32% of communication resources. Second, we have obtained experimental results showing the absolute PSNR of approximately 35 dB with a quality improvement of 1–5 dB, using the "Stefan" tennis sequence, as compared to the implementation with reservation-based approach only.*[1].

Index Terms - QoS, multiprocessor architecture, prediction based systems, NoC monitoring, best effort computation.

I. Introduction

The increasing complexity of systems and continuous miniaturization lead to increasingly complex systems, requiring a modular approach in building complete systems-on-chip. The strong cost constraint in consumer electronics demands integration of a broad functionality into a single chip. In our previous work, we addressed Multiprocessor Systems-on-Chip (MP-SoC) as a promising solution for recent multimedia applications combining various types of computing. The inherited modularity of the application nicely matches with the modularity of a multiprocessor setup. However, the natural limitations of consumer multimedia systems with respect to the resource consumption and the overall system cost, do not allow to worst-case design approach for the system realization. An alternative is Quality-of-Service (QoS) based system design, where the total system resources are shared and distributed at run-time among various active applications.

Several QoS approaches have been reported in the literature. For example, economical reservation-based Quality-of-Service (QoS) solutions are presented by Bormans *et al.* [1] and by Bril *et al.* [2]. We have presented our hierarchical QoS proposal in [3]. However, more recent experimental results showed that pure reservation-based QoS control of the system yields an average efficiency of about 70%. For this reason, we focus on further maximizing possible output quality by using the reservation-based technique in combination with a best-effort run-time adaptation of the computation.

In this contribution, we show that this combination indeed improves the picture quality. We present a QoS management model using a NoC run-time monitoring system [4] for run-time adaptation of the computation graph. In the majority of the processed pictures, a switch to a video processing at higher quality level is obtained.

The paper is divided as follows. Section II illustrates the limitations of a reservation-based QoS. Section III outlines the NoC model with monitoring features. The combined reservation and adaptation QoS framework is given in Section IV. Section V presents the details of our experimental setup for the MPEG-4 arbitrary-shaped video decoder. Section VI concludes the paper.

II. Limitations of Reservation-based QoS

Pure reservation-based QoS control systems have several limitations which are discussed in this section. The QoS control selects the best combination of quality settings of active applications and reserve resources for those applications. Due to the long reconfiguration time and the overhead connected with the reconfiguration after each processed picture, the reservation and reconfiguration takes place at

[1] Milan Pastrnak is with LogicaCMG Nederland, P.O.Box 7089, NL-5605 JB Eindhoven, The Netherlands (e-mail: m.pastrnak@tue.nl).
Calin Ciordas is with TU Eindhoven, NL-5600 MB Eindhoven, The Netherlands.
Peter H.N. de With is with TU Eindhoven / LogicaCMG Nederland, NL-5600 MB Eindhoven, The Netherlands.
Jef van Meerbergen is with TU Eindhoven / Philips Research Eindhoven, NL-5600 MB Eindhoven, The Netherlands.
Kees Goossens is with Philips Research Eindhoven, NL-5656 MB Eindhoven, The Netherlands.

the end of a group of pictures (Video Object Planes). Let us now discuss two disadvantages of this approach.

A. *Long-time resource reservation*

In the decoding of arbitrary-shaped MPEG-4 video objects [5], the reservation of resources for the whole Group of Video Object Planes (GOV) requires that the system has sufficient resources for decoding each Video Object Plane (VOP). However, the MPEG-4 GOV length is not known in advance and is fixed by the actual encoder. Therefore, the QoS control of the decoder has to decide on the reservation of resources for the decoding application for the whole length of a GOV. This GOV set has a variable length (the authors observed sequences of several hundreds of VOPs in one GOV). In the worst case, the decoder QoS control has to decide only on a fragment of the GOV size. Consequently, this approach can lead to a QoS decision for a lower quality level for a long sequence of VOPs. This lower quality level already occurs when only one VOP cannot be decoded within available resources.

B. *Increase of available resources*

We have observed that the reservation-based QoS is also sloth in covering the increase of available resources. The time for the reallocation of resources and increase of the guaranteed quality level for an application is only possible at the end of a GOV. When the quality changes or when a termination of other applications occurs, these resources cannot be directly used for the subsequent VOP decoding. The decoding at a higher quality level starts at the first frame of the next GOV. In the case that such increase of resources occurs at the beginning of the GOV, the response time of the system might be too long for the system user.

These two limitations motivated us to supplement the reservation-based model with the run-time QoS adaptation.

III. NoC with Monitoring Features

The observation of ongoing computations in a system has received a lot of attention in literature. NoC monitoring systems have been proposed in order to cope with observing the communication at run-time. This work was mainly driven by testing and debugging aspects [6]. Passive hardware monitors make use of the so-called SPY feature on twelve chip pins. The internal signals are grouped in sets of twelve signals and are hierarchically multiplexed on twelve pins.

In this paper, we present our solution based on the Æthereal [7] NoC that offers run-time monitoring features. Up till now, the monitoring was used mainly for debugging purposes. The role of monitoring becomes more valuable when coupling it to advanced QoS management that can explore the monitoring information for better distribution of resources. The mechanism of using the monitoring information is highlighted in Section IV.

The NoC monitoring service, as illustrated in Figure 1, consists of configurable monitoring probes (P) attached to NoC components, i.e. router (R) or network interface (NI), their associated programming model, and a monitoring traffic management strategy.

The *monitoring probes* are responsible for collecting the required information from the NoC components. The probes capture the monitored information in the form of events. Multiple classes of events can be generated by each probe, based on a predefined instance of an event model. Monitoring probes are not necessarily attached to all NoC components. The placement of probes is a design-time choice and is related to the cost versus observability tradeoff.

The *traffic management* regulates the traffic from the Monitoring Service Access point (MSA) to the probes, which is required to configure the probes, and the traffic from the probes to the MSA is used to obtain the monitoring information from the NoC. Already available NoC communication services, e.g. guaranteed throughput (GT) or best-effort (BE) connections, or even dedicated solutions can be used for the traffic information for monitoring.

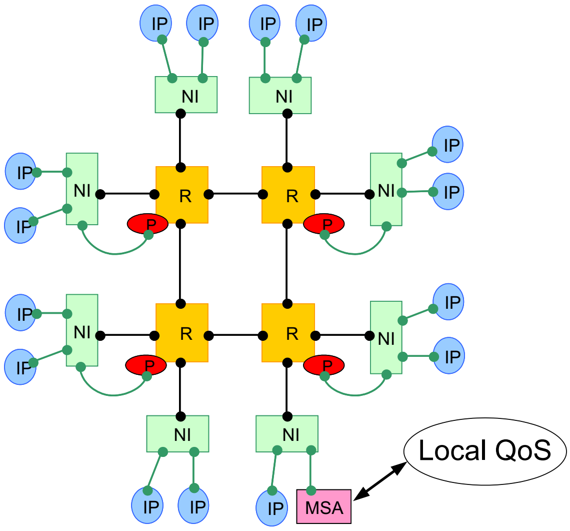

Fig. 1. The NoC architecture view with the MSA connected to the Local QoS control.

The above framework is integrated in our experiments in the following way. The presented NoC with communication-monitoring features offers the combination of mixed GT and BE connections. GT connections support the principles of reservation-based QoS control. BE connections fit to our adaptation technique, which will be presented in the next section. The monitoring mechanism is needed to avoid non-optimal communication of data between tasks that will be completed after their deadline.

IV. Combining Best-effort and Guaranteed Services for QoS management

We have presented our view on two-layer hierarchical QoS management in [3]. The QoS should bridge the application domain with the hardware-platform domain. In Section II, we have identified the limitations of the resource-based QoS model. In our opinion, a pure reservation-based QoS control using guaranteed services has to avoid possible dead-locks resulting from special structures in resource occupation. As a consequence, the QoS control has to reserve extra capacity to circumvent those deadlocks. However, our experiments show that the *combination* of reservation-based control and the best-effort usage of resources is highly increasing the performance of the final system. This combination is outlined in the following paragraphs.

The *Global QoS manager* assigns the resources and appropriate quality level to each application. The important part of our mapping approach is the *resource estimator*. The estimator calculates for each application the amount of required resources for processing the set of new data (in MPEG-4 video decoding, we decided to lock it to the size of the GOV). The resource request is then evaluated with the available resources and the Global QoS manager sets the highest quality that just fits to the platform resources. These resources are reserved for the application until the end of the GOV or any exceptional situation occurs.

The *Local QoS manager* is responsible for monitoring the prediction model and real resource consumption. The monitoring of execution is done at fine granularity, in our case at the VOP level. The Local QoS sets the parameters for scalable communication connections and scalable tasks at run-time, based on the resource availability. The details about connecting the Global QoS manager with the Local QoS manager and run-time monitors are depicted in Figure 2.

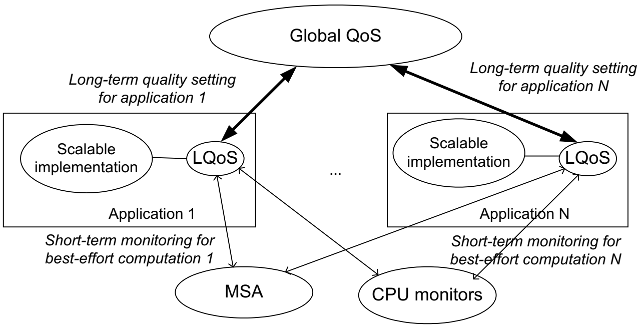
Global QoS
Long-term quality setting for application 1
Long-term quality setting for application N
Scalable implementation
LQoS
...
Scalable implementation
LQoS
Application 1
Application N
Short-term monitoring for best-effort computation 1
Short-term monitoring for best-effort computation N
MSA
CPU monitors

Fig. 2. The application view on the combined QoS management.

The Global QoS knows all applications and estimates resources for them, and depending on the platform status, it sets the long-term settings for each application. The Local QoS controls an individual application and translates the settings into a scalable setting of the tasks. The Local QoS connects to run-time monitors of resources (MSA of communication, CPU monitors) to check the availability of resources for a best-effort computation. If processing at a higher quality level is possible, the Local QOS manager modifies the functional task execution of an application to a higher level.

The observability of the platform at run-time enables us to employ so-called *best-effort* principles to obtain a higher quality level for a short time (fragment of the GOV). This level is higher than the quality level that was assigned by the Global QoS manager. Statistical experiments showed that a worst-case approach is not necessary in 80% of the execution time. This effectively means that the Global QoS manager is overestimating the required resources. These reserved resources are actually used only 20% of the time and during the remaining time can be used by other applications. If we would have only a reservation-based solution, the system would have to wait until the next suitable time for changing the quality level and the corresponding reconfiguration (end of GOV). In the new solution, the Local QoS calculates for each VOP the resource requirements of the succeeding VOP and compares it with the run-time information from MSA monitor and CPU's monitor. If the MSA monitor reports sufficient available resources for the communication, the Local QoS allows the computation at higher quality level. The Local QoS temporary sets parameters for scalable tasks to a higher quality level for decoding of the next VOP within the actual GOV.

V. Experiments with MPEG-4 Arbirary Shaped VOP Decoding

A. MPEG-4 scalable decoding application

Object-based video processing offers a complex application with dynamic resource requirements. Scalable algorithms are important for consumer electronic devices using MPEG video coding, because they offer a trade-off between picture quality and the embedded available computational performance [8]. In our work, we focus on arbitrary-shaped MPEG-4 video objects (VO) decoding. We have presented our proposal for scalability of an arbitrary-shaped VO MPEG-4 decoder in [3]. Figure 3 portrays the details of the computation with the indicated scalability of communication (GT and BE) and tasks-to-processors assignment.

In order to deploy the combination of QoS techniques, we initiate the system with the worst-case mapping from the communication point of view, where we map each task to a different processor. After each GOV, this mapping is reconsidered. Since our experiments are very recent and still ongoing, the actual execution was done without the mapping reconsideration, but was converging to a stable outcome of the actual resource usage that allows broader conclusions.

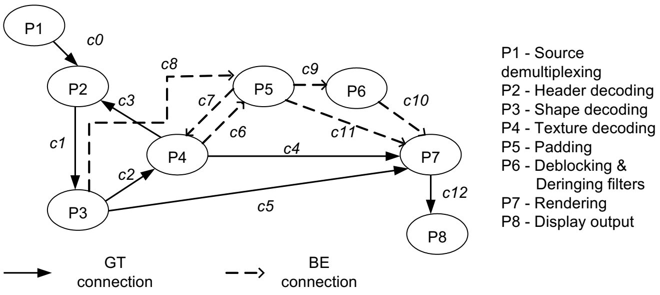

Fig. 3. The scalable mapping of AS VO decoder.

We have defined three quality levels of our experimental AS VO MPEG-4 decoding:

- Level 0 - basic quality, the shape is fully decoded; the basic quality of texture after IDCT is communicated to the *Rendering* task.
- Level 1 - medium quality, the MPEG-4 padding [5] of the texture data is activated; there are no artifacts on edges.
- Level 2 - highest quality, the complete chain with post-processing of deblocking and deringing filters is executed.

B. Architecture for experimental setup

Our system architecture employs a 2x4 mesh Æthereal NoC with eight ARM processing cores. The ARM cores are one-to-one mapped to Network Interfaces (NI). We have implemented a centralized performance monitoring service. Each router was probed with performance monitors able to monitor link utilization. Each monitor communicates performance data to the MSA by means of a low-bandwidth GT connection through the closest located NI, which received an extra NI port for this. The single MSA connects to NI7 (see Fig. 4) by means of an extra NI port. The complete overview is given in Figure 4.

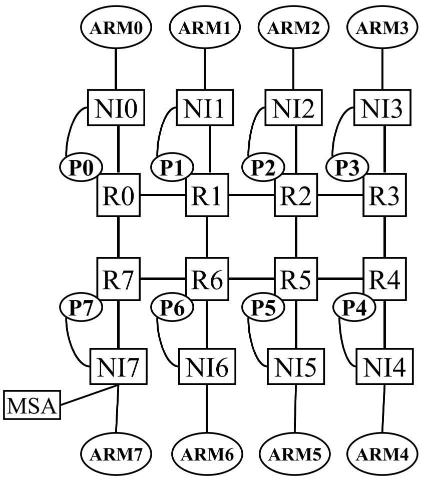

Fig. 4. Experimental system architecture with 2x4 mesh Æthereal NoC connected with eight ARM cores and MSA.

We have chosen the Advanced Coding Efficiency (ACE) Profile, Level 3, at CCIR-601 resolution from the MPEG-4 standard. Figure 5 illustrates the varying communication requirements of the “Stefan” tennis sequence that was segmented from the original resolution of 688×464 pixels. The bold line indicates the other applications running in parallel within the system.

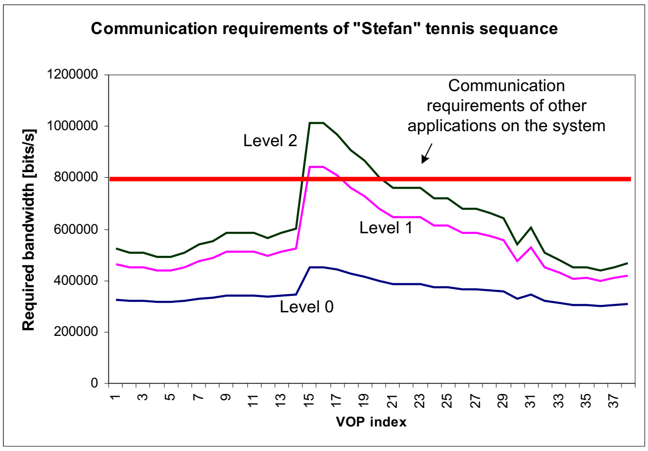

Fig. 5. Communication requirements of “Stefan” tennis sequence. The bold line represents the communication requirements of other applications also executed within the system.

Reservation of resources. At the start of the GOV, the *Estimator* calculates the computation and communication resource requirements at all three quality levels. Next, the Global QoS selects the quality level at which all VOPs can be decoded. In our example (Fig. 5), the quality *Level 0* will be selected because of the requirements of the VOPs with indexes from 13 to 22.

Best-effort computation. Our proposed solution is based on exploring the best-effort communication where it is possible. When compared to the GT connections, the BE connections do not have guarantees on the timing of data delivery. With the option to monitor the NoC connections, the Local QoS can verify at finer granularity (frame level) if there are available resources for BE communication. If the Local QoS received a positive response for all BE connections at a higher quality level, the extended computation at higher quality level is activated.

In our setup we integrated *alien traffic generators* that program the system to a minimum level of communication activity. We assigned the following connections from Figure 3 to the corresponding quality levels:

- Level 0 : c0–c5, c12
- Level 1 : all connections at Level 0 + c6, c7, c8, c11
- Level 2 : all connections at Level 1 + c9, c10

The Local QoS has to monitor the connections c6–c11, as they are of BE type. As is depicted in Figure 5, the initial quality is at quality Level 0. Prior to starting the next VOP decoding, the Local QoS checks the status of the connections

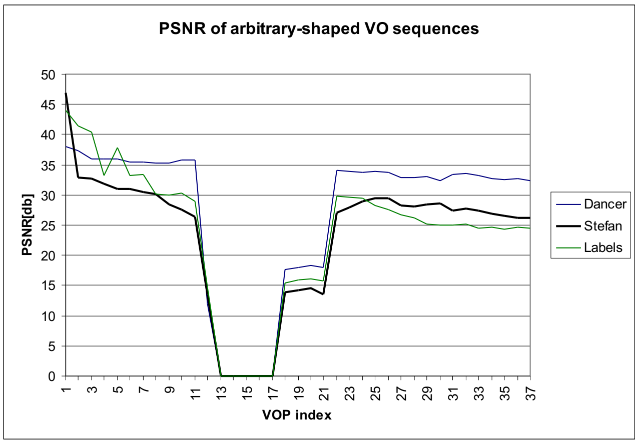

Fig. 6. PSNR of arbitrary-shaped video sequences. "Stefan" tennis sequence at 688×464 pixel resolution while the other two sequences are at CIF resolution.

and if the estimated communication resources are available, then it activates the scalable tasks at the highest possible level.

C. Experimental Results

With the novel mixed approach, the system is decoding only VOPs 14–18 at quality Level 0. The quality Level 1 is achieved for VOPs 13 and 19–22 and the rest of the GOV is decided to be decoded at the highest quality level. The obtained improvement of quality is depicted in Figure 6. As it can be noticed, there is no improvement for frames that were decoded at Level 0. However, there is significant improvement (to the absolute PSNR of approximately 35 dB) on the remainder of the VOPs.

The communication monitoring typically introduces overhead that is orders of magnitude lower than the required bandwidth of the experimental multimedia application. It can be concluded that the NoC monitoring allows the run-time adaptation of the decoding process to a higher quality. It should be noted that the obtained time fraction of 76% where the quality levels are increased, is highly dependent on the video input data and the run-time status of the platform.

VI. Conclusions

We have studied the mapping of object-oriented MPEG-4 decoding on a multiprocessor NoC. Since the amount of objects is unknown in advance and the decoding characteristics are highly variable in resource usage, the guaranteed execution of all decoding tasks cannot be ensured. We have employed the combined solution for reservation-based QoS management with run-time resource adaptation. This adaptation was implemented by using best-effort communication connections instead of the initialized guaranteed-throughput connections, where it was possible. The monitoring features in the network were formed by Monitoring Service Access (MSA) probes at network interfaces. The complete system was experimentally verified with a network of eight ARM processor cores, executing an MPEG-4 Video Object decoder at ACE profile and CCIR-601 and CIF resolution. The proposed framework showed that the adaptation at finer granularity, e.g. VOP level within a GOV, can improve the image quality significantly (experimental results show the absolute PSNR of approximately 35 dB with a quality improvement of 1–5 dB). Furthermore, it can be concluded that the monitoring of resources shortens the reaction time of the system to the system change due to video input changes or application changes. We conclude that the implementation of advanced monitoring systems as described in this paper is indispensable for the realization of multimedia consumer electronics systems designed for high-efficiency usage of the resources.

References

[1] J. Bormans, N.P. Ngoc, G. Deconinck, and G. Lafruit, *"Terminal QoS: advanced resource management for cost-effective multimedia appliances in dynamic contexts" in Ambient intelligence: impact on embedded system design*, Kluwer Academic Publ., NL., 2003, pp. 183-201.

[2] R. J. Bril, Ch. Hentschel, E. F. M. Steffens, M. Gabrani, G. C. van Loo, and J. H. A. Gelissen, "Multimedia QoS in consumer terminals," in *IEEE Workshop on Signal Proc. Systems (SIPS)*, 2001, pp. 332–344.

[3] M. Pastrnak, P. Poplavko, P. H. N. de With, and J. van Meerbergen, "Novel QoS model for mapping of MPEG-4 coding onto MP-NoC," in *9th IEEE International Symposium on Consumer Electronics (ISCE)*, 2005, pp. 93–98.

[4] C. Ciordas, T. Basten, A. Radulescu, K. Goossens, and Jef van Meerbergen, "An event-based monitoring service for Network-on-Chip," in *ACM Transactions on Design Automation of Eletronic Systems*, Vol. 10, No. 4, October 2005, pp. 702–723.

[5] ISO/IEC 14496-2:199/ Amd 1:2000, *"Coding of Audio-Visual Objects - Part 2:Visual, Amendement 1: Visual Extensions"*, Maui, December 1999.

[6] B. Vermeulen, S. Oostdijk, and F. Bouwman, "Test and debug strategy of the PNX8525 nexperia digital video platform system chip," in *IEEE International Test Conference (ITC)*, 2001, pp. 121–131.

[7] Kees Goossens, John Dielissen, and Andrei Rădulescu, "The Æthereal network on chip: Concepts, architectures, and implementations," *IEEE Design and Test of Computers*, vol. 22, no. 5, Sept-Oct 2005, pp. 21-31.

[8] S. Mietens, P. H. N. de With, and Ch. Hentschel, "Computational-complexity scalable motion estimation for mobile MPEG encoding," in *IEEE Transactions on Consumer Electronics*, Vol. 50, No. 1, February 2004, pp. 281–291.

Milan Pastrnak received the M.S. degree in information systems from Zilina University, Slovak Republic, in 1999. In 2002, he completed the post-graduation designers course in software technology of the Eindhoven University of Technology. He received PDEng degree in September 2002 on the project "Distributed Visualization and Simulation with Object-Oriented Networks".

Currently, he is a Junior Researcher at the Information & Distribution Technology, LogicaCMG Nederland B.V., The Nehterlands. He is a guest at the Video Coding and Architectures group, University of Technology, Eindhoven and he closely cooperates with Philips Research Laboratories, Eindhoven, The Netherlands. He is focusing on the application SW-HW co-design, heterogenous multiprocessor systems, quality-of-service for multimedia systems and system-level design.

Peter H.N. de With graduated in electrical engineering from the University of Technology in Eindhoven. In 1992, he received his Ph.D. degree from the University of Technology Delft, The Netherlands. He joined Philips Research Labs Eindhoven in 1984, where he became a member of the Magnetic Recording Systems Department. From 1985 to 1993, he was involved in several European projects on SDTV and HDTV recording. In this period, he contributed as a principal coding expert to the DV standardization for digital camcording. In 1994, he became a member of the TV Systems group at Philips Research Eindhoven, where he was leading the design of advanced programmable video architectures. In 1996, he became senior TV systems architect and in 1997, he was appointed as full professor at the University of Mannheim, Germany, at the faculty of Technical Computer Science. Since 2000, he is with LogicaCMG as a principal consultant and he is professor at the University of Technology Eindhoven, at the faculty of Electrical Engineering. He has written and co-authored numerous papers on video coding, architectures and their realization. In 1995, 2000 and 2004, he coauthored papers that received the IEEE CES Transactions Paper Award and SPIE paper awards. Dr. de With is a senior member of the IEEE, program committee member of the IEEE CES and ICIP, chairman of the Benelux community for Information Theory, scientific advisor of the Dutch Imaging school ASCII, IEEE ISCE and board member of various working groups.

Calin Ciordas is a junior researcher at the Eindhoven University of Technology (TU/e), Information and Communication Systems department. Currently, he is a guest at Philips Research Laboratories, Eindhoven, The Netherlands. He obtained an M.Sc. in computer science from Technical University of Cluj-Napoca, Romania, and a PDEng degree from Eindhoven University of Technology, Netherlands. His current research interest includes system-level design, on-chip multiprocessor systems and networks on chip.

Jef Van Meerbergen (M'87-SM'92) received the electrical engineering and the Ph.D. degrees from the Katholieke Universsiteit Leuven, Belgium, in 1975 and 1980, respectively.

In 1979, he joined the Philips Research Laboratories, Eindhoven, The Netherlands. He was engaged in the design of MOS digital circuits, domain-specific processors, and general-purpose digital signal processors. In 1985, he started working on application-driven high-level synthesis. Initially, this work was targeted towards audio and telecom DSP applications. Later, the application domain shifted towards high-throughput applications. His current interests are in system-level design methods, heterogenous multiprocessor systems, and reconfigurable architectures. He is the Associate Editor of *Design Automation for Embedded Systems*. He is a part-time Professor at the Eindhoven University of Technology, Eindhoven.

Dr. van Meerbergen is a Philips Research Fellow. His Phideo paper received the Best Paper Award at the 1997 ED&TC conference.

Kees Goossens received his BSc in computer science from the University of Wales in 1988, and obtained his PhD from the University of Edinburgh in 1993. In his thesis he investigated the formal verification of hardware, in particular by using semi-automated proof systems in conjunction with formal semantics of hardware description languages such as ELLA and VHDL. He continued this work at several other universities before joining Philips Research in the Netherlands in 1995. At Philips he worked on behavioural synthesis for high-throughput video processing, then on on-chip communication protocols and memory management. Since 2000, he has worked on networks on chip for consumer electronics systems, where real-time predictability (QoS) and costs are major constraints.