ARTES project 0005-9
2000-05-22

Support for Real-Time 3-D Graphics on Future Mobile Terminals under Energy/Area Constraints

Research Proposal for the ARTES programme

Prof. Per Stenström

Department of Computer Engineering, Chalmers University of Technology,
SE-412 96 Gothenburg, Sweden

Phone: +46-31-772 1761, Email: pers@ce.chalmers.se

Dr. Fredrik Dahlgren

Technical Manager - Mobile Application Platforms

Research, R&D New Systems

Ericsson Mobile Communications AB, SE-221 83 Lund, Sweden

Phone: +46-46-194589, Email: fredrik.dahlgren@ecs.ericsson.se

Research plan Sept. 1, 2000 - Feb 28, 2003 (2.5 years)

Summary

One of the driving application domains for future mobile terminals is multimedia applications exemplified by video and high-quality audio processing as well as advanced games. These applications do not only require high peak performance levels, they also impose stringent real-time demands to provide a high user-perceived quality of service. At the same time, future mobile terminals must also accommodate other applications with a wide range of computational requirements that rule out a special-purpose computer infrastructure. What makes the design of the architecture of this infrastructure even more challenging is that it also must meet stringent constraints on area and power dissipation.

In this research project, our goal is to investigate design principles of a general-purpose single-chip processor/memory architecture that can provide high and predictable performance for multimedia applications under challenging constraints regarding chip area and power dissipation. It draws on the synergies with three ongoing projects at Chalmers in the areas of (1) worst-case timing analysis for high-performance architectural features (2) design principles for high-performance memory systems under power dissipation constraints, and (3) algorithms for 3-D computer graphics. These synergies together with the world-leading expertise of the industrial collaborator in the design of mobile terminals provide a solid base for advancing state-of-the-art in this area.

The expected results of the project are fundamental insights into what architectural support is needed to meet the challenging performance and predictability requirements of 3-D graphics applications for a simple processor/memory core built from embedded RAM technology. Besides the obvious potential industrial impact of this project, it advances state-of-the-art in the so far quite unexplored area of energy-efficient general-purpose architectures for the mobile market.

1. Problem Statement

A driving application domain for future mobile wireless terminals, e.g. high-end mobile phones, is multimedia applications including video and high-quality audio (e.g. digital music) processing as well as advanced games. First, these applications demand high-performance to provide a high user-perceived quality by having the processing rate conform to the sampling rate of audio and video. Second, since jitter in the audio and video streams will also contribute to a lower user-perceived quality, the computer infrastructure must conform to stringent real-time requirements. Finally, while this would call for a special-purpose architecture, e.g. vector architectures, future mobile terminals should also accommodate a wide range of applications whose computational needs are hard to anticipate which calls for a general-purpose computational infrastructure.

This general-purpose computational infrastructure for future mobile terminals is also a target for a set of other challenging requirements caused by constraints in the physical size and the limitations of battery technology. This makes area and energy consumption important constraints as well.

Multimedia support in general-purpose computer architectures is not appropriate for mobile terminals for the following reasons:

First, except for some simple standard functions, the execution of complex graphics code in most of today's mobile terminals is performed entirely by an embedded processor. This leads to poor performance, poor real-time responsiveness, and a high energy consumption. Second, the multimedia application area also drives high-end microprocessors for the desktops and has led to graphics acceleration in hardware tightly coupled to most processor cores. In addition, it has triggered the developments of special graphics cards in most PCs sold today, and high-bandwidth memory technologies such as Rambus. However, these solutions are for obvious reasons not viable for a power-sensitive, limited area, system-on-chip design aimed at mobile terminals. Finally, even though complex, real-time 3-D graphics applications for mobile terminals will not be as performance demanding as for desktops and home-entertainment systems, because of the limited display sizes, it is more challenging when factoring in the constraints imposed by the limited energy provided by current as well as future battery technology. This rules out traditional embedded solutions. An additional hurdle is that the computational infrastructure should also accommodate multi-programmed workloads that might include other timing sensitive applications such as audio.

In summary, the main issue to be considered in this project is how to design a single-chip general-purpose processor/memory architecture that delivers sufficient performance and predictability to meet user-perceived quality requirements under power dissipation and chip-area constraints for multimedia and other applications anticipated for future mobile terminals.

2. Main Ideas

Our approach to meet the application requirements under the technology constraints is to investigate the design space of a general-purpose processor/memory chip based on the Intelligent RAM (IRAM) concept[1]. Unlike the Berkeley IRAM approach, which is based on a vector architecture, we will base our initial design on the following considerations:

- A simple pipelined processor core that does not allow out-of-order execution
- A simple memory hierarchy that does selective data caching

The in-order pipelined processor core is motivated from recent observations that it provides a better energy-delay than an out-of-order processor core[2]. In addition, in our recent research we have found that the same

[1] R. Fromm et al., "The Energy Efficiency of IRAM Architectures", Proc. ISCA'97, pp. 327-337, 1997

[2] R. Gonzalez and M. Horowitz, "Energy Dissipation in General Purpose Microprocessors". *IEEE Journal of Solid-State Circuits*, 31(9):1277-1284, September 1996

design decisions also make it easier to achieve a high and predictable performance[3]. Moreover, in contrast to the vector architecture approach in the Berkeley IRAM project, our approach is not tied only to the computational requirements of multimedia applications, but can also accommodate applications with other computational requirements. The second design consideration to use instruction and selective data caching seems to be in contradiction to the predictability requirements of the applications under study. However, in our recent research we have noticed that it is possible to achieve a predictable and high performance by allowing data structures that are statically analyzable to be cached[4].

The goal of the project is to investigate how to best take advantage of the chip resources for this architectural framework to meet the application requirements (predictable performance) under the technology constraints (energy and area).

The scientific methodology is based on complete system simulation using SimICS to study performance and energy-efficiency issues of architectural design tradeoffs. In an on-going project, we have successfully used this methodology to study performance/energy tradeoffs for memory systems.[5] We will drive our models by a set of applications that we anticipate will be used for future mobile terminals such as 3-D graphics, handwriting recognition and video- and audio compression/decompression algorithms. This application suite is available already and used in on-going projects.

This is a collaborative project between Chalmers and Ericsson Mobile Communications, and we are seeking funding for one Ph.D. student at the Department of Computer Engineering at Chalmers. This Ph.D. student will be jointly advised by Prof. Per Stenström at Chalmers and Docent Fredrik Dahlgren at Ericsson. We are not seeking funding for this project elsewhere, and the project is dependent on external funding to be started. We currently have a candidate in mind and anticipate that the project can get started as soon as funding is granted.

3. Expected Results

The project will contribute with knowledge in the following main areas:

- Methods to predict performance and energy dissipation for multimedia applications under area constraints
- Design principles for processor/memory architectures for mobile terminals with challenging predictable performance requirements under energy dissipation and area constraints.

The industrial impact can be substantial. Looking at volumes of embedded 32-bit processors and DSPs, the mobile phones and terminals constitute a substantial part. These processors are developed to meet the expected needs of the vendors' customers. At the moment, we see a dramatic increase in the attempts to add DSP features into general-purpose processors to meet the performance demands on multimedia-related applications. Having a holistic approach with both the application-oriented (i.e. real-time requirements and memory behavior) related and hardware-oriented (i.e. silicon area, memory design, power dissipation) pieces is the only way to achieve a good and realistic solution. Moreover, all added knowledge into memory hierarchy design and efficiency for important emerging application will have a large and direct impact on the design of the ASICs.

4. Project Plan

The tasks for the first 2.5 years (assuming 80% activity level) are as follows.

Task 1 (Definition of application and system framework) This task aims at framing the set of applications that we anticipate will drive the development of future mobile terminals and developing a baseline processor/memory architecture for a single-chip system

---

[3] Thomas Lundqvist and Per Stenstrom: "Timing Anomalies in Dynamically Scheduled Processors" in *Proc. of the 1999 IEEE Int. Real-Time Systems Symposium,* pp. 12-21, Dec. 1999.

[4] Thomas Lundqvist and Per Stenstrom: "A Method to Improve the Estimated Worst-Case Performance of Data Caching," *in Proc. of 6th Int. Conf on Real-Time Computing Systems and Applications,* pp. 255-262., Dec 1999

[5] Jonas Jalminger and Per Stenstrom: "Boosting Energy-Efficiency of Off-Chip Caches using Selective Data Prefetching" *To appear in Proc.of the IEEE Workshop on Complexity-Effective Designs (in conjunction with ISCA),* June 2000.

Task 2 (Methodology development) is to develop an experimental infrastructure to estimate energy consumption and performance of a candidate processor (suitable for being embedded on a system-on-chip) as well as architectural acceleration methods. We will leverage on an available SimICS-based system simulation methodology already used at Chalmers.

Task 3 (Analysis) is to analyze the performance issues involved in the system architectural framework defined in Task 1 using the methodology developed in Task 2 and how it is affected by the predictability and power dissipation constraints imposed by the applications and technology, respectively. This provides a deep understanding of the problems of using standard, general- purpose processors and on-chip memory systems in this new setting.

Task 4 (Concept development) is to develop concepts in terms of architectural extensions for 3-D graphics and novel memory designs and analyze their impact on performance under predictability constraints. Since energy consumption and silicon area are also important metrics, the models of the various architectural extensions must be analyzed in detail. This provides a deep understanding of the cost-effectiveness of different architectural features to meet above goals.

Task 5 (Generalizations) makes use of the findings in Task 4 and generalize them in the context of extending processors with certain architectural support. The goal is to understand a broader design space.

The time-plan for the first 2.5 years is as follows.

- Task 1: 6 months
- Task 2: 6 months, one state-of-the-art report dealing with the design considerations behind the system and applications framework
- Task 3: 6 months, one state-of-the-art report dealing with the methodology and analysis of the system/applications framework and its major limitations.
- Task 4: 9 months, one state-of-the-art report dealing with architectural concepts to improve the performance under the various constraints
- Task 5. 3 months, Licentiate thesis

5. Preliminary Budget

We are seeking support for one Ph. D. student at 80% activity level (480 KSEK/year).

6. Related Work

The project that is mostly related to our project is the Berkeley Intelligent RAM project in which architectures for mobile computers are studied. They have chosen a vector architecture as the base for their study arguing that many multimedia applications need aggressive vector support. In addition, since caches have been considered hard to use for real-time applications, they exclude caches in their design. There are two major problems with their approach. First, mobile terminals must also meet other computational requirements which makes a vector architecture approach restrictive. Secondly, our as well as others recent work have cast doubts on the widely spread misconception that caches are useless in real-time systems. Both these observations have led to our focus on a general-purpose processor with a traditional, albeit restrictive cache architecture. Interestingly, Transmeta's Crusoe also takes a general-purpose approach and consider an in-order statically scheduled (VLIW) processor core.

7. Relation to ARTES and Industrial Relevance

The focus of this project proposal is to find and study solutions to a real-time application area: real-time 3-D graphics for mobile wireless terminals. Since we target *real* industry-related requirements, there is a number of other criteria that must be met for the solutions to be viable; power dissipation, silicon area, and performance. However, if these criteria are met while the real-time requirements are not, the solution will not be useful. This project emphasizes a holistic system view in that a solid understanding of hardware/software interaction is of paramount importance to success. In our on-going research, we have developed a productive methodology to study such interactions as was discussed in Section 3.

3-D computer graphics applications are one essential part of a multimedia terminal, while others include video and high-quality audio. As pointed out in the ARTES profile, these are good examples of real-time applications.

In addition, the proposed project is related to heterogeneous systems, since it deals with complex system designs based on mainstream processor cores as well as software which to a large degree is developed by others as building blocks. We are studying how tailor-made support can enhance the properties of the system.

The project is also related to the evaluation of early system designs, and methodologies for hardware-software design tradeoffs for real-time multi-program workloads in that we will further develop a methodology that can aid in tradeoffs between predictable (real-time) performance and energy consumption.

In summary, the project proposal matches the ARTES call-for-proposals well.

The market of mobile computers and mobile terminals is increasing at a tremendous rate. The industrial relevance should therefore be obvious for anyone familiar with the trends of mobile IT, wireless 3G network possibilities and services, the volumes and trends of current mobile phones and PDAs, and the volumes and aggressiveness of home entertainment products.

8. Relation to other SSF programs

This project fits well with the goals of the PCC program with the focus on multimedia computer infrastructures. However, no funding is available for new projects. It also relates to the goals of the INTELECT program although this program has a focus on electronics rather than on systems with interacting hardware and software. This proposal has been evaluated by INTELECT and it was pointed out to us that its focus on system-level issues did not match the goals of the program.

9. Context

For this project, we have an extremely strong set of people, cooperation, and infrastructure. The research is headed by Prof. Per Stenström at Chalmers, with Docent Fredrik Dahlgren from Ericsson Mobile Communication acting as co-advisor for the Ph.D. student. In addition, the research is carried out in collaboration with three other ongoing research projects at Chalmers: (1) worst-case execution time analysis of high-performance architectural features (Thomas Lundqvist), funded by TFR; (2) Real-time 3-D computer graphics (Dr. Tomas Möller), funded by PAMP, and (3) energy-efficient memory systems (Jonas Jalminger) funded by SSF.

Prof. Per Stenström is leading an internationally recognized research group in computer architecture at Chalmers. He has published more than 80 papers in international conferences and journals. His main areas of research are high-performance computer architectures and analysis methods. He is continuously supporting program committees for top conferences in the field such as ISCA, ASPLOS, and HPCA and is an associate editor of IEEE Trans. on Computers and Journal of Parallel and Distributed Computing. He is also the general chair of ISCA-2001 to be held in Göteborg.

Docent Fredrik Dahlgren has a Ph.D. and Docent degree in Computer System Architecture from Lund University and Chalmers University of Technology, respectively. He is the Technical Manager of the Mobile Application Platforms (MAP) group at the Research department at Ericsson Mobile Communications, Lund. MAP has collaborations with the ongoing designs and architecture work of mobile phones and terminals within Ericsson Mobile Communications around the world. He has published some 35 papers in scientific international journals and reviewed conferences.

Stenström and Dahlgren have a documented strong and long-term (>10 years) collaboration record!!!

Curriculum Vitae

Per Stenström
Professor, Ph.D., Docent, Senior Member of the IEEE

Affiliation

Chalmers University of Technology, Dept. of Computer Engineering, S-412 96 Göteborg, Sweden, Phone: +46 (31) 772 1761, Fax: +46 (31) 772 3663, E-mail: pers@ce.chalmers.se

Permanent positions and degrees:

- Prof. of computer eng. (chair in computer architecture), Chalmers (Sweden), since Nov. 1995. Since April 1999 Vice-dean of the School of Electrical and Computer Engineering.
- Assoc/assist. prof. of computer eng., Lund Univ. (Sweden), 1988-1995. Docent in 1993.
- Ph. D degree in computer eng., Lund Univ. (Sweden) in 1990, docent 1993.
- Master of Science degree in electrical eng., Lund Univ. (Sweden) in 1981.

Visiting positions:

- Visiting prof., EE dept., Univ. of Southern Calif. (USA), July/Aug. 1993.
- Visiting prof., Computer systems lab, Stanford Univ. (USA), June-Dec. 1991.
- Visiting scientist, CS dept., Carnegie-Mellon Univ. (USA), Aug. 1987- May 1988.

Main research interests

- Design principles for high-performance computer architectures
- Performance evaluation methodologies and tools
- Parallelization techniques for multiprocessors
- Energy-efficient computer architectures

Selected professional activities:

- Has authored more than 80 journal and conference publications in the computer architecture, performance analysis, and the compiler areas. Is author of two textbooks.
- Has supervised and graduated six Ph. Ds and three (non-overlapping) Licentiates
- Is associate editor for IEEE Trans. on Computers and the Journal of Parallel and Distributed Computing, guest editor of IEEE Computer, and Proceedings of the IEEE.
- Was vice chair of the program committees of the 14th IEEE Int. Conf. on Distributed Computing Systems, Poznan, Poland, 1994.
- Has been on the program committee of more than twenty computer architecture and parallel processing conferences.
- General Chair of ISCA-2001 to be held in Gothenburg July 2-4, 2001.

Journal publications: 1995--present (refereed)

[1] H. Grahn, P. Stenström, and M. Dubois: "Implementation and Evaluation of Update-Based Cache Protocols Under Relaxed Memory Consistency Models," in *Future Generation Computer Systems*, Vol. 11, No. 3, pp. 247-271, June 1995.

[2] F. Dahlgren and P. Stenström: "Using Write Caches to Improve Performance of Cache Coherence Protocols in Shared-Memory Multiprocessors," in *Journal of Parallel and Distributed Computing*, Vol 26. No 2, pp. 193- 210, April 1995.

[3] F. Dahlgren, M. Dubois, and P. Stenström: "Sequential Hardware Prefetching in Shared-Memory Multiprocessors," in *IEEE Trans. on Parallel and Distributed Systems*, Vol. 6 No 7, pp. 733-746, July 1995.

[4] M. Dubois, J. Skeppstedt, and P. Stenström: "Essential Misses and Memory Traffic in Coherence Protocols," in *Journal of Parallel and Distributed Computing*, Vol. 29, No 2, pp. 108-125, 1995.

[5] F. Dahlgren and P. Stenström "Evaluation of Stride and Sequential Hardware-based Prefetching in Shared- Memory Multiprocessors," in *Trans. on Parallel and Distributed Systems*, Vol. 7, No. 4, pp. 385-398, April 1996.

[6] M. Brorsson and P. Stenström: "Characterising and Modelling Shared-Memory Accesses in Multiprocessor Programs," in *Parallel Computing*, No 22, pp. 869-893, 1996.

[7] P. Stenström, M. Balldin, and J. Skeppstedt: "The Design of a Non-Blocking Load Processor Architecture," in *Microprocessors and Microsystems, No 20, pp. 111-123*, 1996.

[8] J. Skeppstedt and P. Stenström: "Using Dataflow Analysis to Reduce Overhead in Cache Coherence Protocols," in *Transactions on Programming Languages and Systems*, Vol 18, No 6, pp. 659-682, November 1996.

[9] H. Grahn and P. Stenström: "Evaluation of an Adaptive Update-Based Cache Protocol," in *Journal of Parallel and Distributed Computing*, 39(2):168-180, December 1996.

[10] P. Stenström, M. Brorsson, F. Dahlgren, H. Grahn, and M. Dubois: "Boosting Performance of Shared-Memory Multiprocessors," in *IEEE Computer*, pp. 63-70, July 1997.

[11] M. Karlsson and P. Stenström: "Effectiveness of Dynamic Prefetching in Multiple-Writer Distributed Virtual Shared Memory Systems," in *Journal of Parallel and Distributed Computing*, Vol. 43, No. 2, pp. 79-93, 1997.

[12] F. Dahlgren, M. Björkman and P. Stenström: "Reducing the Read Miss Penalty for Flat COMA Protocols, in *the Computer Journal*, Vol. 40, No. 4, pp. 208-219, 1997.

[13] P Stenström, Erik Hagersten, David Lilja, Margaret Martonosi, and Madan Venugopal: "Trends in Shared- Memory Multiprocessing," in *IEEE Computer* , Vol. 30, No. 12, pp. 44-50, December 1997.

[14] F. Dahlgren, J. Skeppstedt, and P. Stenström: "An Evaluation of Hardware-Based and Compiler-Controlled Snooping Cache Protocol Extensions," in *Journal of Future Generation Computer Systems*, No. 13, pp. 469- 487, 1998.

[15] F. Dahlgren, M. Dubois, and P. Stenström: "Performance Evaluation and Cost Analysis of Cache Protocol Extensions for Shared-Memory Multiprocessors," in *IEEE Transactions on Computers*, Vol. 47, No 10, pp. 1041-1055, Oct. 1998.

[16] J. Skeppstedt, F. Dahlgren, and P. Stenström: "Evaluation of Compiler-Controlled Updating to Reduce Coherence-Miss Penalties in Shared-Memory Multiprocessors," in *Journal of Parallel and Distributed Computing*, Vol. 56, No 2, pp. 122-153, 1999.

[17] H. Grahn and P. Stenström: "Comparative Evaluation of Latency-Tolerating and Reducing Techniques for Hardware-Only and Software-Only Directory Protocols", *Journal of Parallel and Distributed Computing*, to appear 1999.

[18] T. Lundqvist and P. Stenström: "An Integrated Path and Timing Analysis Method Based on Cycle-Level Symbolic Execution," Accepted for publication in *Journal of Real-Time Systems*, to appear Nov 1999.

[19] V. Milutinovic and P. Stenström “Opportunities and Challenges for Distributed Shared-Memory Multiprocessors. Guest Editors’ Introduction, *Proceedings of the IEEE*. Vol 87 No 3, pp 399-404, March 1999.

Conference Papers (refereed) 1995-present

[2] M. Björkman, F. Dahlgren, and P. Stenström: “Using Hints to Reduce Read Miss Penalties for Flat COMA Protocols, in *Proc. of 28th Hawaii International Conference on System Sciences,* pp. 242-251, January 1995.

[3] F. Dahlgren and P. Stenström “Effectiveness of Stride and Sequential Hardware-based Prefetching in Shared- Memory Multiprocessors, in *Proc. of First International Conference on High Performance Computer Architecture (HPCA-1),* pp. 68-77, January 1995.

[4] H. Grahn and P. Stenström: “Efficient Strategies for Software-Only Directory Protocols in Shared-Memory Multiprocessors,” in *Proc. of 22nd Annual International Symposium on Computer Architecture*, pp. 38-47, June 1995.

[5] J. Skeppstedt and P. Stenström: “A Compiler Algorithm that Reduces Read Latency in Ownership-Based Cache Coherence Protocols,” in *Proc. of Parallel Architectures and Compilation Techniques,* pp. 69-78, July 1995.

[6] F. Dahlgren, J. Skeppstedt, and P. Stenström: “Effectiveness of Hardware-Based and Compiler-Controlled Snooping Protocol Extensions,” in *Proc. of the International Conference on High Performance Computing, pages 87-92,* December 1995.

[7] M. Karlsson and P. Stenström: “Performance Evaluation of a Cluster-Based Multiprocessor Built from ATM- Switches and Bus-Based Multiprocessor Servers,” in *Proc. of Second International Conference on High Performance Computer Architecture*, pages 4-13, Jan. 1996.

[8] H. Grahn and P. Stenström: “Relative Performance of Software-Only and Hardware-Only Directory Protocols Under Latency Tolerating and Reducing Techniques,” in *Proceedings of the 11th International Parallel Processing Symposium*, pages 500-506, April 1997.

[9] J. Nilsson, F. Dahlgren, M. Karlsson, P. Magnusson, P. Stenström: “Computer System Evaluation with Commercial Workloads” in *Proc. of IASTED Conference on Modeling and Simulation*. pp. 293-297, May 1998.

[10] P Magnusson, F Dahlgren, H. Grahn, M. Karlsson, F. Larsson, A. Moestedt, J. Nilsson, P Stenström, and B. Werner: “SimICS/Sun4m: A Virtual Workstation. In *Proc. of USENIX98*,pp. 119-130, June 1998.

[11] T. Lundqvist and P. Stenström: “Timing Anomalies in Dynamically Scheduled Processors,” in *Proc. of 1999 IEEE Real-Time System Symposium (RTSS’99),* pp. 12-21 Dec. 1999.

[12] T. Lundqvist and P. Stenstrom. “A Method to Improve the Estimated Worst-Case Performance of Data Caching”. in *Proc. of 6th International Conference on Real-Time Computing Systems and Applications (RTCSA' 99),*pp. 255-262, Dec 1999.

[13] M. Karlsson, F. Dahlgren, and P. Stenström: “Prefetching Techniques for Irregular Accesses to Linked Data Structures,” *6th IEEE Int. Symp. on High-Performance Computer Architecture (HPCA-6)*, 2000.

[14] M. Karlsson, F. Dahlgren, and P. Stenström: “An Analytical Model for Working-Set Sizes in Decision Support Systems,” To appear in *SIGMETRICS*, 2000.

[15] A. Saulsbury, F. Dahlgren, and P. Stenström: “Recency-Based TLB Preloading” To appear in *27th IEEE Int. Symp. on Computer Architecture (ISCA-27)*, 2000.