

RESEARCH STATEMENT

Memory Management for Real-Time (Embedded) Multidimensional Signal Processing Systems

Summary: Many signal processing systems, particularly in the multimedia and telecommunication domains, are synthesized to execute data-dominant applications. The system behavior is described in a high-level programming language; the code is typically organized in sequences of loop nests and the data structures are multi-dimensional arrays. In these (embedded) VLSI systems, data transfer and storage have a significant impact on both the system performance and the major design cost parameters – power consumption and chip area. This is why, during the system development process, the designer must often focus on the exploration of the memory subsystem in order to achieve a cost optimized end-product.

Starting from the algorithmic specification and taking into account the designer's constraints, this research work (funded by the National Science Foundation under the CAREER Award), has the general goal of synthesizing a (hierarchical) memory subsystem, optimized for chip area and/or power consumption, subject to performance constraints. Note that this topic is considered by the Semiconductor Research Corporation (SRC) one of the top synthesis problems still unsolved.

Overview: In the last three years, the main focus of my research has steadily shifted in the direction of devising novel techniques based on data-flow analysis in the memory management of real-time multi-dimensional signal processing systems. The guiding idea of this research is to consistently use data-flow analysis as the basic exploration mechanism of the solution space since it offers both exploration freedom and generality in comparison with the traditional scheduling-based investigation. This strategy is realistic since memory management tasks usually need only *relative* (rather than *exact*) information on the signal lifetimes. In addition, this strategy enables the study of the memory management tasks at the desired level of granularity – between whole array and the scalar level – trading-off computational effort and solution optimality.

1. Part of this work has investigated non-scalar techniques for computing *exactly* the minimum data memory in real-time multimedia algorithms. This is a clear advance in the field since all the previous approaches achieved only a memory size *estimation* rather than an *exact computation*. Our exact computation approach has many potential benefits: for instance, it can be used to evaluate the impact of different code (and, in particular, loop) transformations on the data storage. Also, the approach can be used to assess the different models of mapping the arrays from a specification to the physical memory locations. This research uses both algebraic techniques specific to the data-flow analysis used in modern compilers and, also, more recent advances in the theory of n -dimensional polyhedra. This exact approach could be extended to deal with algorithmic specifications of high-throughput applications, where the code contains explicit parallelism, which will be a novelty as well.

2. We have recently developed a formal model for data reuse analysis based on lattices. Due to the manipulation of large sets of data in real-time communication and multimedia processing applications, a multi-layer memory hierarchy can enhance the system performance and, also, can reduce the energy consumption. Savings of dynamic energy can be obtained by accessing frequently used data from smaller memories rather than from large background memories. The optimization of the hierarchical memory architecture implies the addition of layers of smaller memories to which heavily used data can be copied. This optimization must trade-off the reduction of power consumption by accessing the data from smaller memories, and the increased energy demands caused by the additional transfers between memory layers.

Our data reuse model formally identifies those parts of the arrays heavily accessed which duplication in the

memory layer(s) closer to the processor is likely to produce the largest saving in the dynamic energy. The intention is to firstly test this data reuse model on a 2-level memory hierarchy (scratch-pad and off-chip memories) and to extend it, afterwards, to an arbitrary number of memory layers.

3. In a later phase, we plan to extend the hierarchical memory allocation model to save also leakage energy. Different from dynamic energy which increases only when a memory access occurs, leakage energy is spent as long memory is powered on. Since leakage becomes the dominant part of energy consumption for 0.10 μm (and finer) technologies, one of the future developments we are considering is to take leakage into account as well. In principle, savings of leakage energy can be obtained improving the performance of the VLSI system (in terms of number of clock cycles).

4. Chip area represents another important component in the cost of the memory architecture of multimedia processing systems. Since data is organized in several memory layers, the optimization must trade-off the reduction of power consumption due to memory fragmentation, and the increase in area and interconnect cost due to the additional memory necessary to store the copies of data, and also because of the additional area overhead (like addressing logic) due to the memory fragmentation.

5. Another problem currently studied is the mapping of the typically large multi-dimensional arrays into the physical memory. Based on the analysis of the lifetimes of signals, different signal-to-memory mapping models are studied and implemented. These formal models aim to obtain minimum bounding windows for the array elements simultaneously alive.

6. A project in an earlier phase of development refers to the memory management of configurable architectures. Typical configurable computing systems consist of arrays of reconfigurable logic blocks and reprogrammable interconnect. In order to offer greater computing capabilities, high-performance commercial configurable architectures have integrated, besides configurable logic blocks (CLB), fixed components like DSP and microprocessor cores, custom hardware, distributed block random-access memory (RAM) modules. For instance, the Xilinx Virtex II Pro field programmable gate array (FPGA) series provides up to 125K logic cells, up to 4 PowerPC processor cores, and distributed, embedded RAM blocks having the same capacity. In homogeneous architectures, like Xilinx Virtex II FPGAs, the block RAM modules of a same capacity are evenly distributed on the chip. Heterogeneous architectures contain a variety of block RAM modules with different capacities. For example, the on-chip memory on an Altera Stratix II FPGA chip consists of three types of RAM modules. These configurable architectures exhibit superior computing possibilities, storage capacities, and flexibility over traditional FPGAs. In addition, the research work on the memory subsystem for configurable architectures will offer memory management solutions to be used within the more recent design methodologies for *dynamically reconfigurable systems*.

Two Ph.D. students – Hongwei Zhu and Ilie I. Luican – are currently involved in this project. Hongwei defended his preliminary thesis in April 2006 and is expected to finish his Ph.D. by May 2007. Ilie will probably have his preliminary defense in Fall 2007. A master program based on building part of the software infrastructure was also completed (Karthik Chandramouli) in the Fall of 2003.

Our research work within this project gained the interest of several European academic centers. For instance, we have recently initiated a co-operation with the groups of prof. Francky Catthoor from the Interuniversity Microelectronics Center (IMEC), Leuven, Belgium and, also, with the group of prof. Per Gunnar Kjaldsberg from the Norwegian University of Science and Technology, Trondheim, Norway (see the articles *J12* and *C'20* in the list of publications). In addition, the group of prof. Donatella Sciuto from Politecnico di Milano, Italy, expressed interest of using our work within their framework for dynamically reconfigurable systems.