

## Design Challenges for Single-Chip Servers

Vikas Agarwal, Karthikeyan Sankaralingam, Doug Burger, Steve Keckler  
*Department of Computer Science, The University of Texas at Austin*

In collaboration with Chuck Moore and Alex Mericas  
*IBM Enterprise Systems Group*

With the future transistor budgets that will soon be available, it will be possible to build chips that each have dozens of processors which can handle hundreds of tasks at once: single-chip servers. We are building the infrastructure to evaluate the design space of these high-throughput devices. The challenges to building these chips are four-fold:

- Permitting reliable operation in an environment where radiation and inductive effects can cause transistors to switch randomly,
- Incorporating dynamic power management that prevents the chip from overheating, plus minimizes energy consumption over time,
- Designing each processing core on the chip so that it can exploit intra-task, fine-grain parallelism as well as run at multi-GHz clock speeds,
- Incorporating mechanisms and policies that perform on-line resource management: balancing single-thread performance, on-chip memory consumption, off-chip bandwidth utilization, and overall throughput of tens to hundreds of threads running concurrently. The management of resources should not be limited to expensive, infrequent protection domain crossings.

To evaluate these future systems, we will need full system simulation capability, a detailed single-chip server simulator, and technology models that permit incorporation of the correct parameters into the simulator and design. We are working on all three, and describe the first and third below.

In the machine simulation component of this project, we are developing a full system simulator that accurately and efficiently emulates both the microprocessor and the surrounding components such as memory and disk I/O. We are starting with IBM ARL's SimOS-PPC simulator that has accurate models for the caches and disk subsystems, but uses a simple instruction emulator to model the core microprocessor. In this part of the project we will integrate both the microprocessor core from the SimpleScalar toolset and GP-Timer, the IBM Power4 microarchitecture simulator, into SimOS. With this infrastructure, we will be able to measure the performance of real server applications on accurate microprocessor models and ultimately evaluate new architectures and architectural features for improved server performance.

Looking forward to future architectures, we are developing a methodology for incorporating technology into the design of a high performance, scalable microarchitecture. We have developed technology-based models of memory-oriented structures to determine the relationship between structure sizes and access time across a range of technology generations. We are using these models and the SimpleScalar simulator, to evaluate tradeoffs between structure capacity and latency in high clock rate processors. We will also use these models to study efficient on-chip memory hierarchies and their impact on performance. We are currently using similar techniques to evaluate the scalability of memory arrays in commercial IBM server microprocessors.