Arsenal: Exploiting Massive Heterogeneity for Power-Limited, High-Performance Processors

Michael B. Taylor Steven Swanson
Department of Computer Science & Engineering
University of California, San Diego
mbtaylor@cs.ucsd.edu swanson@cs.ucsd.edu
http://www.cs.ucsd.edu/users/{mbtaylor,swanson}

10 October 2006

1 Introduction

As transistor density continues to scale with each process generation, the power density of CMOS circuits is increasing. At the same time, cooling and energy costs, battery capacity and acceptable noise levels make increases in processor operating power undesirable. These two trends, increasing power density and fixed power budgets, will force designers onto one of two paths: 1) Build smaller designs so that power consumption remains constant or 2) Design processors so that only a small fraction of the design is operating at one time. The first scenario is likely the effective demise of Moore's law, since the benefits of increased integration will decline. To avoid this unfortunate outcome and continue to enjoy performance returns from Moore's Law, we pursue the second scenario.

We propose a new style of processor design called *Arsenal*. Arsenals comprise a diverse collection of specialized computing and communication circuits, of which a small subset (determined by the power budget) will be active at any point. The applications use the circuits best suited to it subject to an energy budget and performance target. Arsenal designs employ customized transistors and interconnect to implement heterogeneous, *specialized processing elements* (SPEs), interconnection networks, and memories. SPEs differ along many axes, such as functional units (e.g., specialized units for graphics, signal processing, physics, database, general-purpose), pipeline structures (to connect functional units), control-structures (e.g, macro- and micro- architectures), arithmetic structures, circuit styles, and voltages. An Arsenal's on-chip networks are specialized for traffic-type (e.g., single-word static or multi-word dynamic) and provide a selection of bandwidths, latencies, energy tradeoffs and topologies. Memories can be organized according to access style, aliasing possibilities, banking, volatility, density and read/write power. Each resource provides different levels of performance and power consumption.

Arsenal systems run programs on the set of resources that allow it to meet its performance target while consuming the least power possible. Alternately, the system can select components to provide the maximum performance within a fixed power budget.

To succeed, Arsenal systems must accomplish two objectives. First, they will need to match the diversity that exists within workloads with diversity in the hardware. For instance, general-purpose programs vary in power demands, performance targets, instruction mix, and control behavior. Arsenal systems could match this diversity by providing SPEs that vary in transistor characteristics, clock and voltage scaling, instruction set, and pipeline depth. Arsenal systems will include static and dynamic mechanisms for varying the performance and power characteristics of SPEs, on-chip memories, and on-chip interconnects.

Second, Arsenal systems must be able to map applications onto a set of diverse processing, communication, and memory resources the hardware provides. Arsenal systems will use models of application