Mobile Supercomputers

Todd Austin, David Blaauw, Scott Mahlke, and Trevor Mudge, University of Michigan
Chaitali Chakrabarti, Arizona State University
Wayne Wolf, Princeton University

Moore's law has held sway over the past several decades, with the number of transistors per chip doubling every 18 months. As a result, a fairly inexpensive CPU can perform hundreds of millions of operations per second—performance that would have cost millions of dollars two decades ago.

We should be proud of our achievements and rest on our laurels, right? Unfortunately, no.

The human appetite for computation has grown even faster than the processing power that Moore's law predicted. We need even more powerful processors just to keep up with modern applications like interactive multimedia communications. To make matters more difficult, we need these powerful processors to use much less electrical energy than we have been accustomed to.

In other words, we need mobile supercomputers that provide massive computational performance from the power in a battery. These supercomputers will make our personal devices much easier to use. They will perform real-time speech recognition, video transmission and analysis, and high-bandwidth communication. And they will do so without us having to worry about where the next electrical outlet will be.

But to achieve this functionality, we must rethink the way we design computers. Rather than worrying solely about performance, with the occasional nod to power consumption and cost, we need to judge computers by their performance-power-cost product.

This new way of looking at processors will lead us to new computer architectures and new ways of thinking about computer system design.

Current trends in computer architecture and power cannot meet the demands of mobile supercomputing. Significant innovation is required.

A MOBILE COMPUTING WORLD

Untethered digital devices are already ubiquitous. The world has more than 1 billion active cell phones, each a sophisticated multiprocessor. With sales totaling about $400 million every year, the cell phone has arguably become the dominant computing platform, a candidate for replacing the personal computer.

We expect to see both the types and numbers of mobile digital devices increase in the near future. New devices will improve on the mobile phone by incorporating advanced functionality, such as always-on Internet access and human-centric interfaces that integrate voice recognition and videoconferencing. We also anticipate the emergence of relatively simple, disposable devices that support the pervasive computing infrastructure—for example, sensor network nodes.

The requirements of low-end devices are increasing exponentially, and computer architectures must adapt to keep up.

Some elements of high-end devices are already present in 3G cell phones from the major manufacturers. High-end PDAs also include an amazing range of features, such as networking and cameras.

SUPERCOMPUTING REQUIREMENTS

A mobile supercomputer will employ natural I/O interfaces to the mobile user. For example, input could come through a continuous real-time speech-processing component. Device output will include high-bandwidth graphics display, either as a semitransparent heads-up display or an ocular interface such as a retinal projector. An audio channel will support output for audio reception and sound cues. Finally, the device will include a high-bandwidth wireless interface for network and telecommunication access.

This platform will have to execute many computationally intensive applications: soft radio, cryptography, augmented reality, speech recognition, and mobile applications such as e-mail and word processing. We expect this platform to require about 16 times as much computing horsepower as a 2-GHz Intel Pentium 4 processor, for a total performance payload of 10,000 SPECInt benchmark units (www.specbench.org).

To remain mobile, the device must achieve this extremely high perfor-

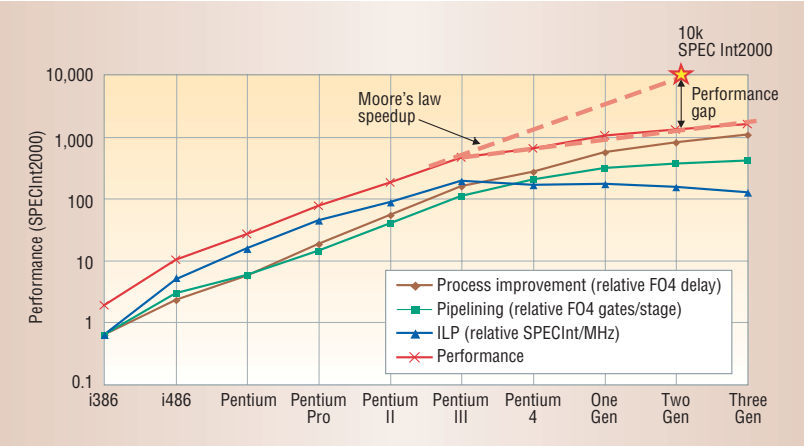

Figure 1. Performance trends for desktop processors. The star indicates the mobile supercomputer's requirement.

mance using only a small battery for power. Given the slow growth trend for batteries—5 percent capacity increase per year—we estimate that a mobile supercomputer (circa 2006) will require a 1,475 mA-hr battery weighing 4 oz. With a five-day battery lifetime under a 20 percent duty cycle (peak load versus standby), we estimate that the system's peak power requirement must not exceed 75 mW.

PERFORMANCE AND POWER TRENDS

Unfortunately, mobile supercomputing's requirements are in contrast to the trends we see in both computer architecture and power for future devices.

Figure 1 shows the trends in performance, measured in SPECInt, for a family of Intel x86 processors. Figure 2 shows the power consumption trends in the same processors. The graphs represent the published data for processors ranging from the 386 (in 1990) to the Pentium 4 (in 2002) in roughly two-year steps. The predicted trends through 2008 are derived from the 2003 edition of the *International Technology Roadmap for Semiconductors* (http://public.itrs.net/).

The star on each graph indicates our mobile supercomputer's performance and power requirements. Clearly, the trends will not meet mobile supercomputing demands anytime soon—and without significant innovation, perhaps they never will.

General-purpose limits

For more than three decades, architects have lavished attention on the design and optimization of general-purpose processors. As a result, current designs feature many advanced techniques such as superpipelining, superscalar execution, dynamic scheduling, multilevel memory caching, and aggressive speculation. Combined with fabrication technology improvements, these optimizations have resulted in a steady doubling of processor performance every 18 months.

But a growing body of evidence suggests that general-purpose processor optimizations are diminishing in value. A study examining the scalability of future general-purpose processor designs (V. Agarwal et al., "Clock Rate versus IPC: The End of the Road for Conventional Microarchitectures," *Proc. 27th Ann. Int'l. Symp. Computer Architecture* (ISCA 00), IEEE CS Press, 2000, pp. 248-259) identified two kinds of general-purpose processor optimizations:

- increased clock speed through pipelining, and
- higher instruction throughput via instruction-level parallelism.

The combined strength of these optimizations has led to the industry's impressive performance gains.

The study points out that clock-rate improvements from pipelining must soon diminish because current designs have little logic within pipe stages. As such, latch delay and clock skew will soon dominate the clock period. The pipeline curve in Figure 1 illustrates this leveling off. For example, Intel's Pentium 4 microprocessor has only 12 fanout-of-four (FO4) gate delays per stage, leaving little logic that can be bisected to produce higher clocked rates.

The negative trend of the instruction-level parallelism curve in Figure 1 suggests that increased instruction throughput cannot make up for anticipated clocking limits. The Pentium 4 microprocessor achieves only about 80 percent of its predecessor Pentium III's instruction throughput for some applications (measured in SPECInt/Mhz for the same technology).

As architectural optimizations reach their limits, they threaten a primary source of value in the computer industry, namely ongoing performance increases.

Nanometer impedences

Circuit-level effects in nanometer devices are also a leading barrier to continued performance scaling. Short-channel effects already prevent gate delay from scaling with feature size as originally expected. Figure 1 shows the technology curve flattening. Capacitive and inductive coupling and increased interconnect lengths pose a serious difficulty for fast signal transmission across the die.

Furthermore, as Figure 2 shows, the sharp rise in static leakage current in nanometer designs is impeding continued improvements in processor power consumption. The leakage current originates in a dramatic increase in both subthreshold current and gate-oxide

leakage current. In fact, static power consumption is now a primary issue in deep submicron design and is projected to account for as much as 50 percent of the total power dissipation for high-end processors in 90-nm technology.

REVOLUTIONARY CHANGES

In mobile applications, a device can be in standby mode a significant portion of the time. In this case, leakage power dominates total power dissipation and threatens the ability to meet the power requirements for high-performance mobile processors. It is becoming clear that incremental improvements within the architecture and circuit subdomains are not going to deliver the extra performance and power efficiency that high-end mobile applications will demand.

Furthermore, future generations of VLSI technology will not provide the reliable operation that we have so long assumed. The small size of future devices will make them vulnerable to radiation-induced upsets, circuit noise, and other factors that produce enough operational, transient failures to require architectural designs that can compensate for them. This means diverting a significant amount of the processor's computational effort to check the results. Thus, we must be even more clever about how we squeeze performance out of our machines, particularly since all that checking logic consumes energy that we can ill afford to lose.

JOINT OPTIMIZATIONS

To build practical mobile supercomputers, system architects need to jointly optimize across algorithms, architectures, and circuits. We don't have all the answers today about how to solve all the problems inherent in mobile supercomputing, but we believe that we have identified some useful approaches.

We can control tradeoffs in a vertically integrated manner:

- microarchitectures that can take advantage of advanced circuit features,
- programs that automatically exploit application-specific architectures, and
- software to glue the layers together and allow existing off-the-shelf applications to use the system efficiently.

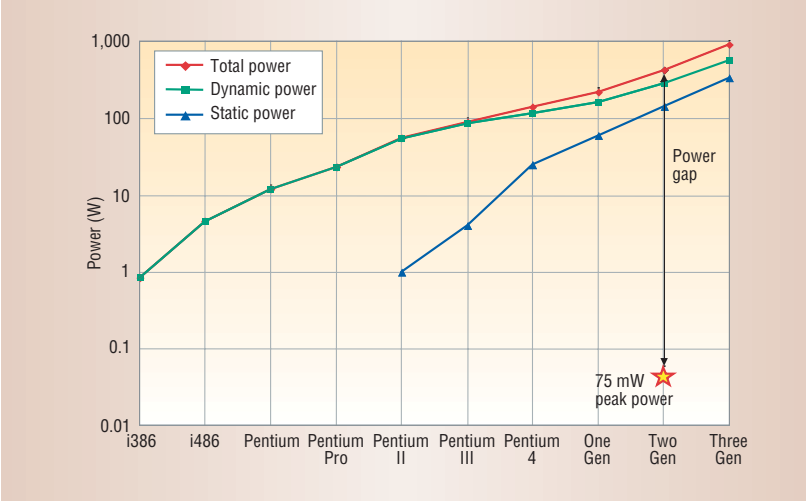

Figure 2. Power trends for desktop processors. The star indicates the mobile supercomputer's requirement.

We can also turn logic checking to our advantage by using it to run the system at its targeted performance level—even in the presence of errors caused by infrequently occurring "worst-case" scenarios. We can tune our designs for "better-than-worst-case" operation and remove the large safety margins necessary to insure against worst-case situations.

Of course, we must apply energy management aggressively at all abstraction levels to meet mobile supercomputing requirements. This means optimizing the hardware so that components can be turned on and off quickly. It also requires extracting program data during compilation for use in guiding energy management. Finally, we need to develop sophisticated user monitoring systems that can more accurately predict when to shut down parts of the system to save energy. ■

Todd Austin *is an associate professor of electrical engineering and computer science at the University of Michigan. Contact him at austin@umich.edu.*

David Blaauw *is an associate professor of electrical engineering and computer science at the University of Michigan. Contact him at blaauw@umich.edu.*

Scott Mahlke *is an assistant professor of electrical engineering and computer science at the University of Michigan. Contact him at mahlke@umich.edu.*

Trevor Mudge *is Bredt Family Chair of electrical engineering and computer science at the University of Michigan. Contact him at tnm@umich.edu.*

Chaitali Chakrabarti *is a professor of electrical engineering at Arizona State University. Contact her at chaitali@asu.edu.*

Wayne Wolf *is a professor of electrical engineering at Princeton University. Contact him at wolf@princeton.edu.*