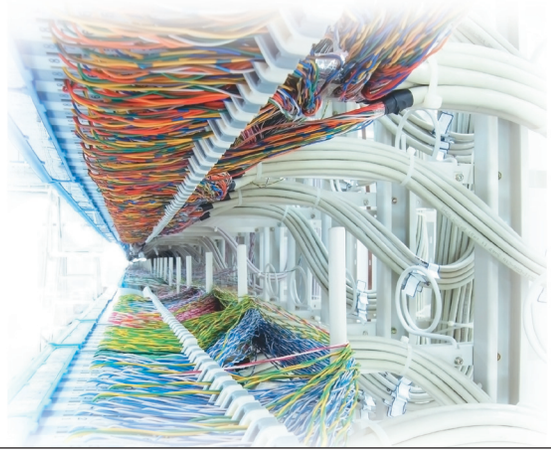
Will Power Problems Curtail Processor Progress?

Neal Leavitt



The use of ever-increasing numbers of ever-shrinking transistors has raised processors' demand for power, which threatens the steady performance increases that chips have experienced for decades.

In a famous 1965 paper, Intel cofounder Gordon E. Moore predicted that the number of transistors that could be placed inexpensively on an integrated circuit would roughly double every year. In 1975, he changed his prediction, saying the doubling would occur every two years. Since then, processors by and large have followed Moore's law.

"The performance of processors has continued to improve each generation," said Stanford University professor William Dally, who is also chief scientist and senior vice president of research at GPU maker Nvidia. "Previously, this scaling was due to improvements in process technology. Today, the scaling is due to improvements in architecture and circuit design that make processors more efficient."

However, a new concern threatens these many years of progress.

"Moore's law continues to pack twice as many transistors onto a chip, but this is leading to a doubling of power consumption and is creating extreme overheating problems," stated University of California, San Diego professor Michael Taylor.

In fact, many chips have to run with some transistors turned off, reducing their overall performance. "Modern chips have become power-constrained, and this is getting worse with every generation," said Dally.

Computer scientists from Microsoft Research, the University of Washington, the University of Wisconsin-Madison, and the University of Texas at Austin studied a broad range of microprocessors, as well as transistor-industry roadmaps, and projected chip performance and power for the next 10 years.

The study found that even multicore processors will be constrained by power consumption and thus won't be able to deliver successive performance improvements every few years.

Figure 1 illustrates this trend, which could slow the ongoing development of new technologies and the release of products with major new capabilities.

The recent chip performance and power study has generated considerable discussion within the computing industry.

"There are many viable approaches to power reduction right now," said Tony King-Smith, vice president of marketing for Imagination Technologies, which designs and licenses multimedia and communications semiconductor cores.

"Gone are the days of packing more transistors to deliver higher frequency for performance," added Intel Labs Fellow Shekhar Borkar. He said the intelligent design and integration of system components have yielded efficient chips that offer better performance while consuming less power.

University of Wisconsin-Madison professor Karthikeyan Sankaralingam expressed concern about the power-related issues that today's processor designers face. "This is a serious threat to Moore's law," he said, "but there are research opportunities and many ideas are being explored in the community. Unfortunately, these are disruptive changes and must occur in a short span, which is the key challenge here."

SCALING ROADBLOCKS

The dominant chip-design trend through the 1990s was to increase clock frequency. A problem, though, is that power consumption increases along with frequency.

In the late 1990s and early 2000s, designers began relying on chips performing more work in a single clock cycle—by simultaneously handling multiple instructions—rather than by increasing clock frequency. Another trend was placing

0018-9162/12/$31.00 © 2012 IEEE Published by the IEEE Computer Society

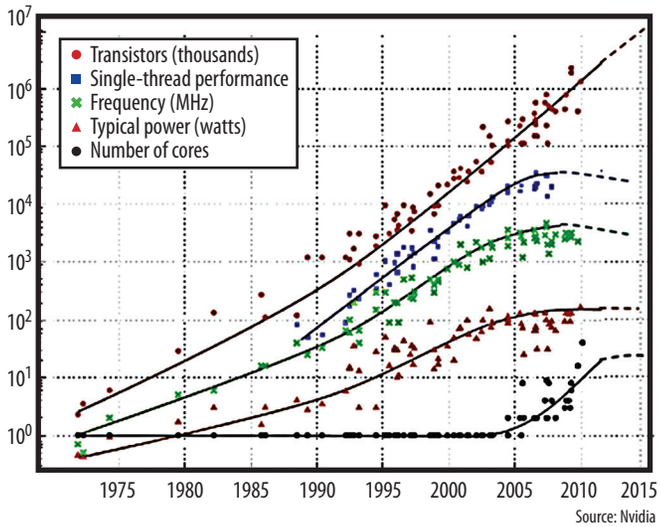


Figure 1. Vendors have increased the number of transistors they pack on chips and the number of processing cores they put on chipsets. However, this has caused power demands to increase, which could slow performance improvement.

multiple processing cores—which increase performance by conducting various parts of a task simultaneously in parallel—on chips.

However, power requirements of today's smaller transistors have caused a phenomenon known as *Dennard scaling* to stop.

In 1974, electrical engineer, inventor, and IBM Fellow Robert Dennard wrote a landmark paper that explored different ways to scale CMOS devices. The findings were that the devices' relative power demands would remain constant even as the number and switching speed of transistors on a chip increased, as long as voltages were reduced in proportion to the decrease in transistor size.

A key reason Dennard scaling is slowing down is that as transistors shrink, leakage current increases. This limits the ability to continue reducing voltage.

Increased power generates more heat on a chip, which can affect performance and reliability, as well as damage circuitry. Practical constraints on the ability to cool chips limit the amount of power that can be used, explained Nvidia's Dally.

Power issues even affect the use of multicore chips, said Tom Hackenberg, semiconductors research manager for market-analysis firm IMS Research.

These chips distribute potential heating problems over multiple cores and can either run some cores at lower speeds or not at all to improve power efficiency. Up to a point, Hackenberg noted, chips using this approach can increase performance by adding cores, rather than solely increasing the number of transistors on a processor.

But, he added, this process is limited because adding cores increases a chipset's size. Moreover, power-related issues still affect efforts to increase the performance of individual cores.

STUDYING PERFORMANCE

The recent chip performance and power study used analytical models to review performance and power-consumption issues for various multicore processors.

Multicore scaling

The study suggested that increasing the number of cores on a chip alone won't sustain the performance improvements that industry and consumers expect.

Multicore processors are limited in their ability to utilize parallelism with many applications, noted the University of Wisconsin's Sankaralingam. Similarly, he added, vendors are constrained because adding cores increases a chipset's power consumption.

Dark silicon

Dark silicon—transistors turned off because the power they demand would create problems for the processor as a whole, a condition that many advanced processors experience—could contribute to decreasing chip-performance improvements, noted Sankaralingam.

According to the recent study, by early 2013, 21 percent of the transistors in advanced chips will need to go dark at any one time; and in seven years, more than one-half will have to be turned off. The researchers predicted this would enable multicore-chip performance to increase only threefold in the next 10 years.

"By comparison, we saw about a 40× [chip] performance improvement every 10 years from 1988 to 2004," noted University of California, Berkeley, professor Kurt Keutzer.

On the other hand

Numerous industry experts said that while the study raised important issues, the problem is not quite as dire as the researchers claim.

For example, noted Dally, current chip architectures consume more energy per instruction or operation than they need to, so there's a lot of room for improvement.

Added Keutzer, "We've had dark silicon for some time, so I don't find this notion at the processor-cores level a troubling prospect."

"The rate of chip performance will not lag—not for a decade—so the innovations will continue," said Intel's Borkar. "We'll be able to increase the number of transistors by making them efficient with process

technology and by being innovative in how to use them."

ANSWERING THE CHALLENGE

According to Sankaralingam, the most pessimistic outlook is that solving chip technology's power-related challenges is too difficult and processor performance growth will decline, causing problems for the IT and IT-dependent economic sectors.

However, these issues might also trigger and accelerate innovations that complement and surpass performance as chip design's primary goal.

For example, instead of large-scale, all-purpose devices that need powerful chips, there's already an increase in special-purpose devices that don't need high-performance processors.

According to Nathan Brookwood, Research Fellow with market-research firm Insight 64, designers will have to optimize the use of the transistors that are already on chips.

Researchers could pursue optimization techniques—perhaps including energy-efficient architectures, specialized hardware, or improved parallelization—along with increasing multicore implementation.

Industry and academic experts are already using or researching other energy-efficient techniques, including

- new digital-logic approaches such as the ability to perform less-precise calculations when absolute accuracy isn't critical;
- new processing paradigms;
- fundamental microarchitecture innovations such as extreme parallelism; and
- utilizing otherwise dark transistors to provide specialized functionality that takes some of the processing load off the main processor core, thereby reducing energy consumption.

Sankaralingam is studying processing techniques that eliminate many power-hungry elements.

His team is also looking at the Dynamically Synthesized Execution Resource (DySER) approach, which improves programmable processors' energy efficiency and performance via techniques that dynamically change the specialized functions they perform.

Dally has worked on architectures that would either eliminate or revise the design of many hardware structures—like caches and instruction-fetch mechanisms—to reduce energy consumption.

Other projects add specialized processing blocks, such as video decoders and math hardware, to a chip to offload work from the main processor that the new components could handle more efficiently. Such specialization is already used commonly in cellular-phone chips.

UC San Diego researchers are looking at a technique to build application-specific blocks into chips. Also, Taylor's group has proposed multicore chips that contain hundreds of specialized *conservation cores*, each designed to execute common fragments of applications with maximum energy efficiency.

His research group is building a prototype chip called GreenDroid, which automatically generates conservation cores from dark silicon to reduce smartphones' energy consumption.

Imagination Technologies' King-Smith said too many chip designers are trying to force older processor designs to reduce power consumption even though they were never designed to do so.

"It's time for engineers to realize that they need to rip up those designs and start thinking about the interaction between the application, OS, and hardware, and utilize new, highly scalable architectures," he said.

The increasing movement of computing to mobile devices—which use lower-performing processors—will delay the impact of many of the industry's power-related challenges, said UC Berkeley's Keutzer.

The solutions ultimately found to address these challenges will shape the post-CMOS generation of processor technology and carry the industry forward.

For example, employing new architectural approaches—such as chips with different types of cores and using new materials and ultralow-voltage techniques—will enable performance increases for "at least the next decade," stated Linley Gwennap, president of market-research firm The Linley Group.

"Facing the challenges of power forces us to be far more efficient in everything about our designs," said analyst Tom Starnes of market research firm Objective Analysis.

"Given the severity of [the power-consumption] problem, both industry and academia face a race against time. To address these challenges, the industry must be open to significant, disruptive design changes," said Sankaralingam.

Nonetheless, added Starnes, the chip industry has always found ways to meet its serious challenges and will do so now. "There's so much business riding on continued advancements in semiconductors," he said, "that the problems will be resolved."

Neal Leavitt *is president of Leavitt Communications (www.leavcom.com), a Fallbrook, California-based international marketing communications company with affiliate offices in Brazil, France, Germany, Hong Kong, India, and the UK. He writes frequently on technology topics and can be reached at neal@leavcom.com.*

Editor: Lee Garber, *Computer*; l.garber@computer.org