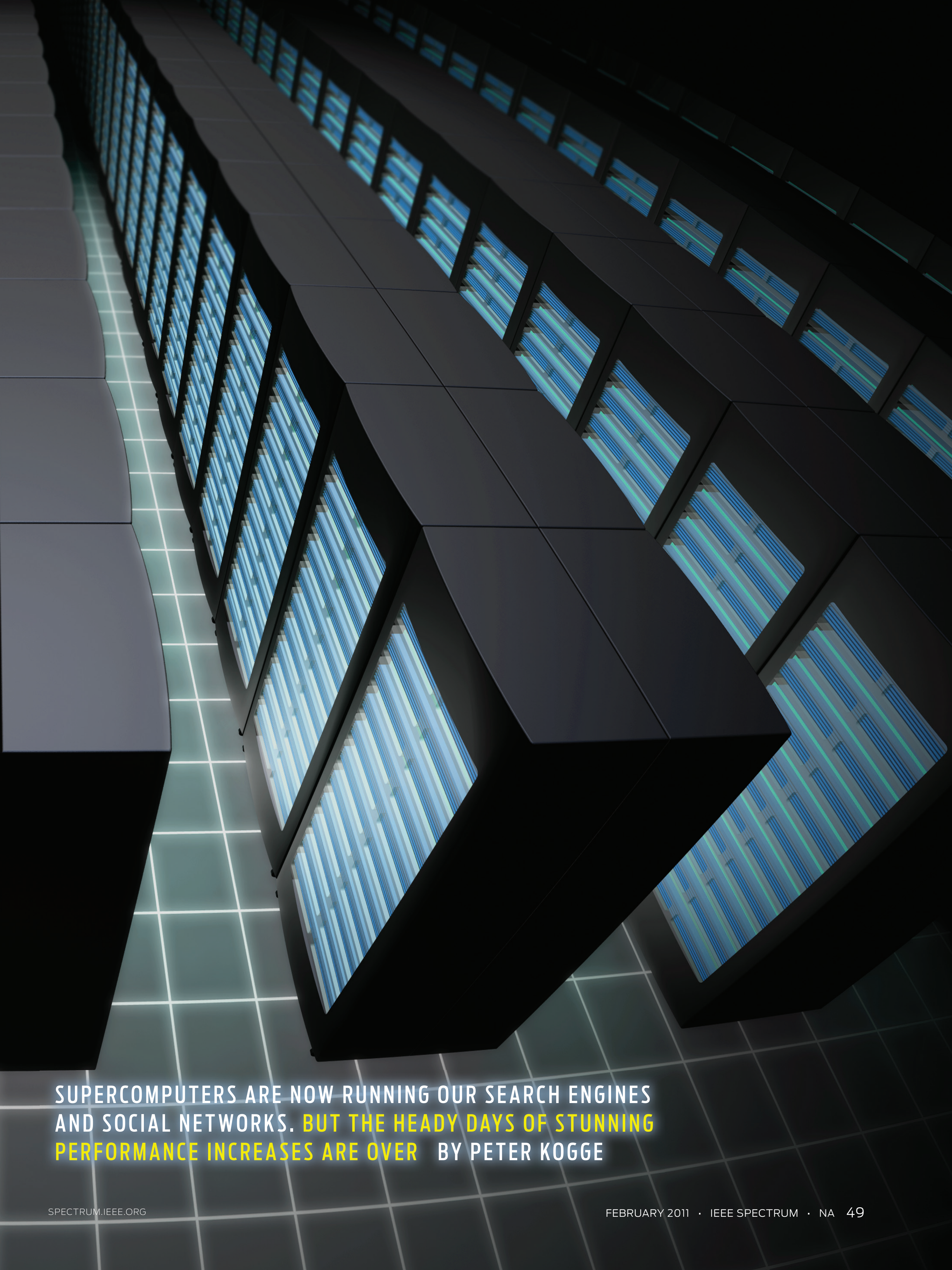
THE TOPS IN FLOPS



SUPERCOMPUTERS ARE NOW RUNNING OUR SEARCH ENGINES AND SOCIAL NETWORKS. BUT THE HEADY DAYS OF STUNNING PERFORMANCE INCREASES ARE OVER BY PETER KOGGE

Supercomputers are the crowning achievement of the digital age. Yes, it's true that yesterday's supercomputer is today's game console, as far as performance goes. But there is no doubt that during the past half-century these machines have driven some fascinating if esoteric pursuits: breaking codes, predicting the weather, modeling automobile crashes, simulating nuclear explosions, and designing new drugs—to name just a few. And in recent years, supercomputers have shaped our daily lives more directly. We now rely on them every time we do a Google search or try to find an old high school chum on Facebook, for example. And you can scarcely watch a big-budget movie without seeing supercomputer-generated special effects.

So with these machines more ingrained than ever into our institutions and even our social fabric, it's an excellent time to wonder about the future. Will the next decade see the same kind of spectacular progress as the last two did?

Alas, no.

Modern supercomputers are based on groups of tightly interconnected microprocessors. For decades, successive generations of those microprocessors have gotten ever faster as their individual transistors got smaller—the familiar Moore's Law paradigm. About five years ago, however, the top speed for most microprocessors peaked when their clocks hit about 3 gigahertz. The problem is not that the individual transistors themselves can't be pushed to run faster; they can. But doing so for the many millions of them found on a typical microprocessor would require that chip to dissipate impractical amounts of heat. Computer engineers call this the power wall. Given that obstacle, it's clear that all kinds of computers, including supercomputers, are not going to advance at nearly the rates they have in the past.

So just what can we expect? That's a question with no easy answer. Even so, in 2007 the U.S. Defense Advanced Research Projects Agency (DARPA) decided to ask an even harder one: What sort of technologies would engineers need by 2015 to build a supercomputer capable of executing a quintillion (10^{18}) mathematical operations per second? (The technical term is floating-point operations per second, or flops. A quintillion of them per second is an exaflops.)

DARPA didn't just casually pose the question. The agency asked me to form a study group to find out whether exaflops-scale computing would be feasible within this interval—half the time it took to make the last thousandfold advance, from teraflops to petaflops—and to determine in detail what the key challenges would likely be. So I assembled a panel of world-renowned experts who met about a dozen times over the following year. Many of us had worked on today's petaflops supercomputers, so we had a pretty good idea how hard it was going to be to build something with 1000 times as much computing clout.

We consulted with scores of other engineers on particular new technologies, we made dozens of presentations to our DARPA sponsors, and in the end we hammered out a 278-page report, which had lots of surprises, even for us. The bottom line, though, was rather glum. The practical exaflops-class supercomputer DARPA was hoping for just wasn't going to be attainable by 2015. In fact, it might not be possible anytime in the foreseeable future. Think of it this way: The party isn't exactly over, but the police have arrived, and the music has been turned way down.

This was a sobering conclusion for anyone working at the leading edge of high-performance computing. But it was worrisome for many others, too, because the same issues come up whether you're aiming to construct an exaflops-class supercomputer that occupies a large building or a petaflops-class one that fits in a couple of refrigerator-size racks—something lots of engineers and scientists would dearly like to have at their disposal. Our panel's conclusion was that to put together such "exascale" computers—ones with DARPA's requested density of computational might, be they building-size supercomputers or blazingly fast rack-size units—would require engineers to rethink entirely how they construct number crunchers in the future.

How far away is an exaflops machine? A decent supercomputer of the 1980s could carry out about a billion floating-point operations per second. Today's supercomputers exceed that by a factor of a million. The reigning champion today is China's Tianhe-1A supercomputer, which late last year achieved a world-record 2.57 petaflops—that's 2.57 quadrillion (2.57×10^{15}) flops—in benchmark testing. Still, to get to exaflops, we have a factor of almost 400 to go.

The biggest obstacle to that by far is power. A modern supercomputer usually consumes between 4 and 6 megawatts—enough electricity to supply something like 5000 homes. Researchers at the University of Illinois at Urbana-Champaign's National Center for Supercomputing Applications, IBM, and the Great Lakes Consortium for Petascale Computation are now constructing a supercomputer called Blue Waters. In operation, this machine is going to consume 15 MW—more actually, if you figure in what's needed for the cooling system. And all that's for 10 petaflops—two orders of magnitude less than DARPA's exaflops goal.

If you tried to achieve an exaflops-class supercomputer by simply scaling Blue Waters up 100 times, it would take 1.5 gigawatts of power to run it, more than 0.1 percent of the total U.S. power grid. You'd need a good-size nuclear power plant next door. That would be absurd, of course, which is why DARPA asked our study group to figure out how to limit the appetite of such a computer to a measly 20 MW and its size to 500 conventional server racks.

To judge whether that is at all feasible, consider the energy expended per flop. At the time we did the study, computation circuitry required about 70 picojoules for each operation, a picojoule being one millionth of one millionth of a joule. (A joule of energy can run a 1-watt load for one second.)

The good news is that over the next decade, engineers should be able to get the energy requirements of a flop down to about 5 to 10 pJ. The bad news is that even if we do that, it won't really help. The reason is that the energy to perform an arithmetic operation is trivial in comparison with the energy needed to shuffle the data around, from one chip to another, from one board to another, and even from rack to rack. A typical floating-point operation takes two 64-bit numbers as input and produces a 64-bit result. That's almost 200 bits in all that need to be moved into and out of

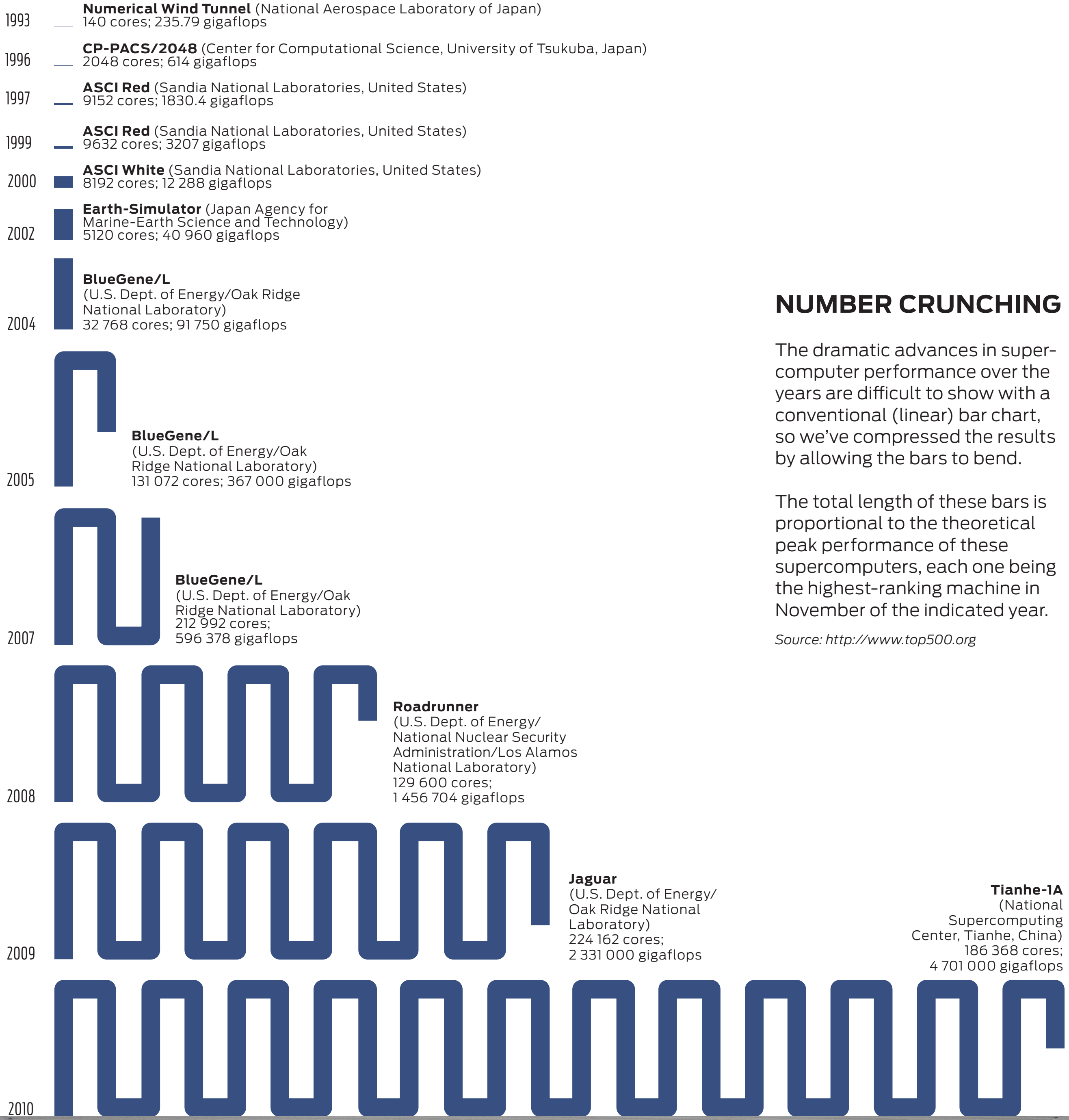
PRECEDING PAGE: ILLUSTRATION: GEORGE RETSECK

NUMBER CRUNCHING

The dramatic advances in supercomputer performance over the years are difficult to show with a conventional (linear) bar chart, so we've compressed the results by allowing the bars to bend.

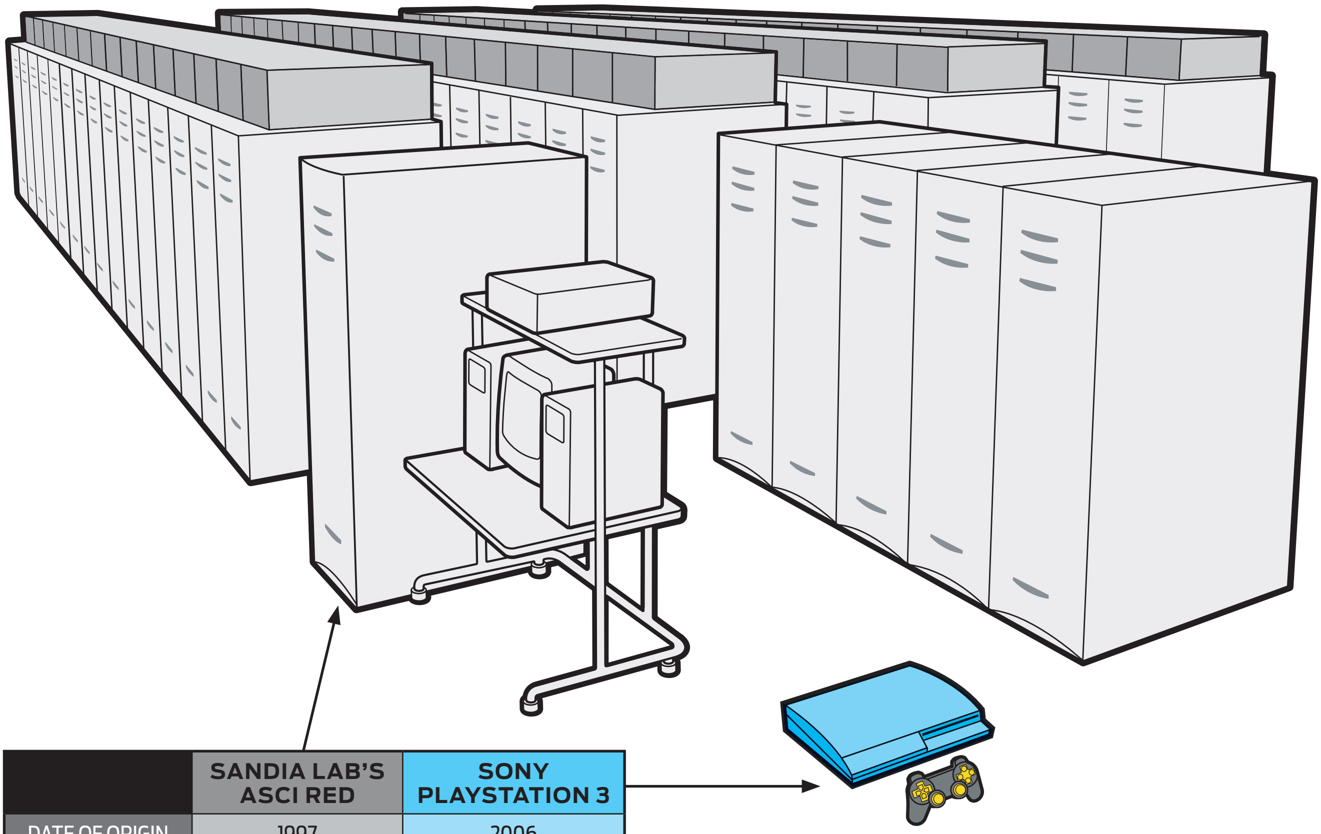
The total length of these bars is proportional to the theoretical peak performance of these supercomputers, each one being the highest-ranking machine in November of the indicated year.

Source: http://www.top500.org



The Tianhe-1A supercomputer in Yianjin, China PHOTO: CHINAFOTOPRESS/GETTY IMAGES

TALE OF THE TAPE: SUPERCOMPUTER VS. GAME CONSOLE



GEORGE RETSECK

	SANDIA LAB'S ASCI RED	SONY PLAYSTATION 3
DATE OF ORIGIN	1997	2006
PEAK PERFORMANCE	1.8 teraflops	1.8 teraflops*
PHYSICAL SIZE	150 square meters	0.08 square meter
POWER CONSUMPTION	800 000 watts	<200 watts

* For GPU; CPU adds another 0.2 teraflops

some sort of memory, likely multiple times, for each operation. Taking all that overhead into account, the best we could reasonably hope for in an exaflops-class machine by 2015 if we used conventional architecture was somewhere between 1000 and 10 000 pJ per flop.

Once the panel members realized that, we stopped thinking about how to tweak today's computing technology for better power efficiency. We'd have to start with a completely clean slate.

TO GET a handle on how best to minimize power consumption, we had to work out a fairly detailed design for the fundamental building block that would go into making up our hypothetical future supercomputer. For this, we assumed that the microprocessors used would be fabricated from silicon, as they are now, but using a process that would support chip voltages lower than the 1 volt or so that predominates today. We picked 0.5 V, because it represented the best projection for what industry-standard silicon-based logic circuitry would be able to offer by 2015. Lowering the operating voltage involves a trade-off: You get much lower power consumption, because power is proportional to the square of voltage, but you also reduce the speed of the chip and make circuits more prone to transient malfunctions.

Bill Dally (then at Stanford and now chief scientist of Nvidia Corp.), working largely on his own, hammered out the outlines of such a design on paper. The basic module he came up with consists of a chip with 742 separate microprocessor cores running at 1.5 GHz. Each core includes four floating-point units and a small amount of nearby memory, called a cache, for fast data access. Pairs of such cores share a somewhat slower second-level cache, and all such pairs can access each other's second-level (and even third-level) memory caches. In a novel twist, Dally's design has 16 dynamic RAM chips directly attached to each processor. Each processor chip also has ports for connections to up to 12 separate routers for fast off-chip data transfers.

One of these processor-memory modules by itself should be able to perform almost 5 teraflops. We figured that 12 of them could be packaged on a single board and that 32 of these boards would fit in a rack, which would then provide close to 2 petaflops, assuming the machine was running at peak performance. An exaflops-class supercomputer would require at least 583 such racks, which misses DARPA's target of 500 racks but is nevertheless a reasonable number for a world-class computing facility.

The rub is that such a system would use 67 MW, more than three times the 20 MW that DARPA had set as a limit. And that's not even the worst problem. If you do the arithmetic, you'll see that our 583-rack computer includes more than 160 million microprocessor cores. It would be tough to keep even a small fraction of those processors busy at the same time.

Realistic applications running on today's supercomputers typically use only 5 to 10 percent of the machine's peak processing power at any given moment. Most of the other processor cores are just treading water, perhaps waiting for data they need to perform their next calculation. It has proved impossible for programmers to keep a larger fraction of the processors working on calculations that are directly relevant to the application. And as the number of processor cores skyrockets, the fraction you can keep busy at any given moment can be expected to plummet. So if we use lots of processors with relatively slow clock rates to build a supercomputer that can perform 1000 times the flops of the current generation, we'll probably end up with just 10 to 100 times today's computational oomph. That is, we might meet DARPA's targets on paper, but the reality would be disappointing indeed.

The concerns we had with this approach did not end there. Accessing memory proved especially vexing. For example, in analyzing how much power our hypothetical design would use, we assumed that only 1 out of every 4 floating-point operations would be able to get data from a nearby memory cache, that only 1 out of every 12 memory fetches would come from a separate memory chip attached to the microprocessor chip, and only 1 out of every 40 of them would come from the memory mounted on another module. Real-world numbers for these things are invariably larger. So even our sobering 67-MW power estimate was overly optimistic. A later study indicated the actual power would be more like 500 MW.

For this design we added only the amount of memory that we thought we could afford without the power requirements of connecting it all together becoming too much of an issue. The resultant amount of memory, about 3.6 petabytes in all, seems large at first blush, but it provides far less memory than the 1 byte per flops that is the supercomputer designer's holy grail. So unless memory technologies emerge that have greater densities at the same or lower power levels than we assumed, any exaflops-capable supercomputer that we sketch out now will be memory starved.

And we're not even done with the seemingly insurmountable obstacles! Supercomputers need long-term storage that's dense enough and fast enough to hold what are called checkpoint files. These are copies of main memory made periodically so that if a fault is discovered, a long-running application need not be started over again from the beginning. The panel came to the conclusion that writing checkpoint files for exaflops-size systems may very well require a new kind of memory entirely, something between DRAM and rotating disks. And we saw very limited promise in any variation of today's flash memory or in emerging nanotechnology memories, such as carbon nanotubes or holographic memory.

As if the problems we identified with excessive power draw and memory inadequacies weren't enough, the panel also found that lowering the operating voltage, as we presumed was necessary, would make the transistors prone to new and more-frequent faults, especially temperature-induced transient glitches. When you add this tendency to the very large number of components projected—more than 4 million chips with almost a billion chip contacts—you have to worry about the resiliency of such systems. There are ways to address such concerns, but most solutions require additional hardware, which increases power consumption even further.

So are exaflop computers forever out of reach? I don't think so. Meeting DARPA's ambitious goals, however, will require more than the few short years we have left before 2015. Success in assembling such a machine will demand a coordinated cross-disciplinary effort carried out over a decade or more, during which time device engineers and computer designers will have to work together to find the right combination of processing circuitry, memory structures, and communications conduits—something that can beat what are normally voracious power requirements down to manageable levels.

Also, computer architects will have to figure out how to put the right kinds of memory at the right places to allow applications to run on these systems efficiently and without having to be restarted constantly because of transient glitches. And hardware and software specialists will have to collaborate closely to find ways to ensure that the code running on tomorrow's supercomputers uses a far greater proportion of the available computing cores than is typical for supercomputers today.

That's a tall order, which is why I and the other DARPA panelists came away from the study rather humbled. But we also found a greater understanding of the hurdles, which will shape our research for many years to come. I, for example, am now exploring how new memory technologies can reduce the energy needed to fetch data and how architectures might be rearranged to move computation to the data rather than having to repeatedly drag copies of that data all around the system.

Perhaps more important, government funding agencies now realize the difficulties involved and are working hard to jump-start this kind of research. DARPA has just begun a program called Ubiquitous High Performance Computing. The idea is to support the research needed to get both very compact high-performance computers and rack-size supercomputers built, even if bringing a warehouse full of them together to form a single exaflops-class machine proves to be prohibitive. The hope is to be able to pack something equivalent to today's biggest supercomputers into a single truck, for example. The U.S. Department of Energy and the National Science Foundation are funding similar investigations, aimed at creating supercomputers for solving basic science problems.

So don't expect to see a supercomputer capable of a quintillion operations per second appear anytime soon. But don't give up hope, either. If rack-size high-performance computers do indeed become as ubiquitous as DARPA's new program name implies they will, a widely distributed set of these machines could perhaps be made to work in concert. As long as the problem at hand can be split up into separate parts that can be solved independently, a colossal amount of computing power could be assembled—similar to how cloud computing works now. Such a strategy could allow a virtual exaflops supercomputer to emerge. It wouldn't be what DARPA asked for in 2007, but for some tasks, it could serve just fine. ❑

TELL US WHAT YOU THINK at http://spectrum.ieee.org/exaflops0211.