Technology | DOI:10.1145/2063176.2063183 Gary Anthes

Revamping Storage Performance

Great strides are being made in finding fast alternatives to the slow disks that dominate storage systems, but fast media are not nearly enough.

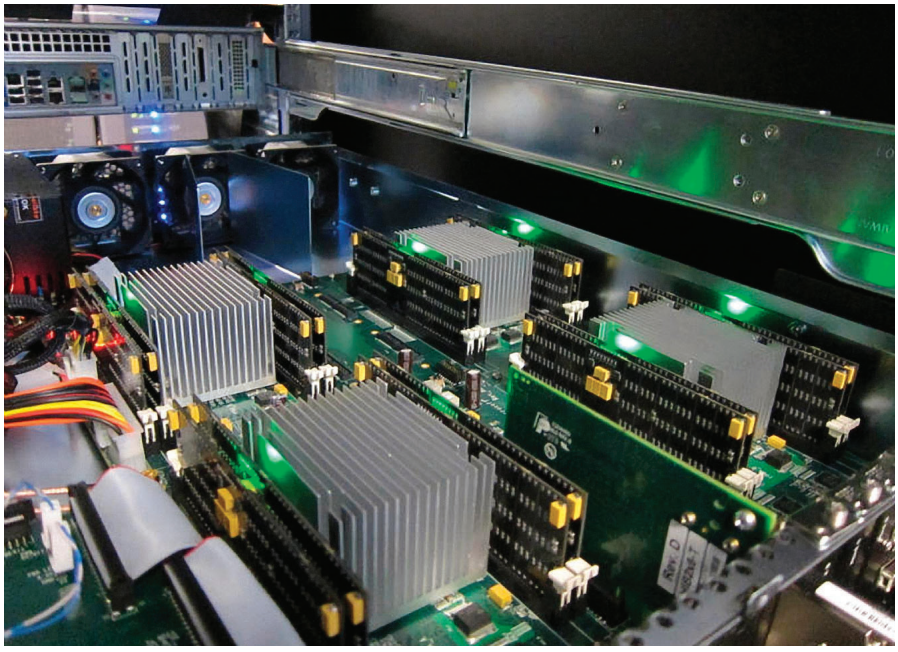
The growth in the power of computers, driven for decades by Moore's Law and more recently by increased core counts, coupled with huge improvements in network bandwidth, disk densities, and other metrics, is nothing short of astonishing. Yet practitioners at the high end say we are losing the battle against the data deluge.

"Our society is literally drowning in data," notes Allan Snavely, associate director of the San Diego Supercomputer Center at the University of California at San Diego (UCSD). "Data acquisition devices ranging from space telescopes to genomic sequencing machines, and the Internet itself, are producing data almost faster than it can be written to disk."

Indeed, disks are a major part of the problem. Disk-based storage systems are 10 to 100 times slower than a network and thousands of times slower than main memory in delivering data to an application, in part because the data comes from a relatively slow electromechanical device. In recent years, fast flash memories have begun replacing disks in some applications. And manufacturers have started showing promising prototypes of more exotic non-volatile storage devices such as phase-change memory (PCM).

While substituting fast media for slow disks can help, it is by no means the whole solution. In fact, computer scientists at UCSD argue that new technologies such as PCM will hardly be worth developing for storage systems unless the hidden bottlenecks and faulty optimizations inherent in storage systems are eliminated.

A team at UCSD led by Steven Swanson, assistant professor of computer science and engineering, is doing just that. A recent prototype, called Moneta, bypasses a number of functions in the operating system (OS) that typically slow the flow of data to and from storage. These functions were developed years ago to organize data on disk and manage input and output (I/O). The overhead introduced by them was so overshadowed by the inherent latency in a rotating disk that they seemed not to matter much. But with new technologies such as PCM, which are expected to approach dynamic random-access memory (DRAM) in speed, the delays stand in the way of the technologies' reaching their full potential. Linux, for example, takes 20,000 instructions to perform a simple I/O request.



A view of the internals of the Moneta storage array with phase change memory modules installed.

A phase-change memory prototype, called Moneta, bypasses a number of functions in the operating system that typically slow the flow of data to and from storage.

Introducing Moneta

Moneta is a prototype high-performance storage array. It uses field-programmable gate arrays to implement a scheduler and a distributed set of memory controllers attached to conventional DRAM emulating PCM. (A similar prototype from the same researchers, called Onyx, uses PCM.)

PHOTOGRAPH BY STEVEN SWANSON

The scheduler orchestrates Moneta's operations and is able to extract parallelism from some I/O requests. By redesigning the Linux I/O stack and by optimizing the hardware/software interface, researchers were able to reduce storage latency by 60% and increase bandwidth as much as 18 times.

The I/O scheduler in Linux performs various functions, such as assuring fair access to resources. Moneta bypasses the scheduler entirely, reducing overhead. Further gains come from removing all locks from the low-level driver, which block parallelism, by substituting more efficient mechanisms that do not.

"The OS gives you parallelism as an illusion," says Rajesh Gupta, a professor of computer science and engineering at UCSD. "It was designed to allow multiple users to support concurrency while actually doing shared access. But shared access and parallelism are two different things."

A further reduction in latency comes from bypassing an interrupt needed to wake up a thread that sleeps while waiting for completion of an I/O request. Instead, the thread does not sleep but spins in a busy loop.

In the hardware, I/O bandwidth is increased by providing separate queues for reads and for writes, and I/O is balanced by not processing big requests all at once while smaller ones wait.

As a result of these optimizations to software and hardware, Moneta performs I/O benchmarks 9.5 times faster than a RAID array of conventional disks, 2.8 times faster than a RAID array of flash-based solid-state drives (SSDs), and 2.2 times faster than Fusion-io's flash-based SSD, a high-end flash technology.

Actual speedups on large, data-intensive jobs could be even more dramatic, says Snavely. "For the right kind of random data access—such as examining a social network graph—flash is between 10 times and 100 times faster than spinning disk, and Moneta is 10 times faster than flash," he says. "That means some data-mining calculations that might take 24 hours today would only take 15 minutes or even less."

But, cautions Swanson, getting that kind of improvement will require rewriting or significantly reengineering the application software.

Moneta performs I/O benchmarks 9.5 times faster than a RAID array of conventional disks, 2.8 times faster than a RAID array of flash-based solid-state drives (SSDs), and 2.2 times faster than Fusion-io's high-end, flash-based SSD.

Sharing Principles

While Moneta is optimized for PCM, Swanson says many of Moneta's principles could be used with other fast non-volatile memories, such as spin-transfer torque random access memory. "The key takeaway is that when these new memories appear, we will shift to a place where the software really plays a critical role in the overall performance of the system," Swanson says. "In Moneta, we focused strongly on software for minimizing latency and maximizing concurrency."

Swanson's team is now moving parts of the Moneta software out of the OS and into the storage array itself, in a special library that can be accessed by applications. Additional enhancements to reduce latency will be possible, Swanson says. For example, a database typically uses the standard I/O calls, with their inherent overhead, which are provided by the OS. But software in the new library bypasses that overhead by taking over those calls and talking to the storage hardware directly. "I can now address non-volatile storage directly from my application, just like DRAM," Gupta says. "That's the broader vision—a future in which the memory system and the storage system are integrated into one."

Swanson says taking advantage of these capabilities will require extensive revamping of application software. "We have talked with companies like Oracle and Teradata, and they realize they will have to change a lot of

Overhauling Systems Operations

Onur Mutlu, assistant professor of electrical and computer engineering at Carnegie Mellon University, is considering the use of phase-change memory (PCM) not for storage systems, but for main memory, where conventional DRAM is nearing the limits of its scaling abilities. Like Steven Swanson, assistant professor of computer science and engineering at University of California, San Diego, Mutlu is not looking to improve PCM itself, but the systems operations that surround it.

PCM is thought to be more scalable and ultimately less costly than DRAM, but has access latencies several times slower than DRAM, requires very high write energies, and is much less durable. Mutlu and his collaborators have reorganized PCM buffers in a way that mitigates high-energy writes and hides latency. Their design performs partial writes so that only modified parts of cache are written, reducing wear. And it distributes wear more evenly by shifting the contents of a write along a row of memory cells and by periodically swapping memory segments of high and low write accesses. These techniques can extend the life of PCM, used as memory, from 171 days of continuous use to as much as 20 years, he says.

An advantage of PCM is that, unlike DRAM, it scales favorably. Mutlu says PCM is expected to scale down to 9nm, and it can also gain density by storing multiple bits in a cell. He says no one knows exactly how much more DRAM can scale, but that scaling it beyond 25nm will be extremely challenging.

"PCM definitely has a future in storage," Mutlu says. "Main memory is a lot harder because of very high bandwidth and low latency requirements, so the research we are doing is 10 years out or so." He and his colleagues are not trying to replace DRAM, but augment it in hybrid memory systems. "So you have a limited amount of DRAM in the system and a very large amount of PCM," he explains. "Hardware maintains counters about which pages should go to DRAM vs. PCM, with high-frequency pages going to DRAM."

the way their systems work," he says. A great deal of the complexity in database management systems lies in the buffer management and query optimization to minimize I/O, and much of that might be eliminated.

Also, storage systems will greatly benefit from changes in the way they access data over a network. Now, when a storage system accesses remote data on disk, it must navigate through the network stack, through file services software on both the local and remote machine, then do all that again coming back. "This change in storage performance is going to force us to look at all these different aspects of computer system design," Swanson says. "The reach of this is going to be surprisingly broad."

Swanson and colleagues at other universities are attempting to "catalyze these changes" by forming a consortium—not yet named—of storage system researchers. They include experts from the low levels of the OS, through the application layers, and on up to the data center and network architectures, he says. "The idea is to attack all these layers at once," says Sawnson, "and hopefully demonstrate that it's worth industry's time to make these changes to commercial systems."

Swanson's group is looking out five to eight years, he says. "The end point of this is you'll have non-volatile solid-state storage that's about as fast as DRAM," he notes. "That would be an increase of about 2,500 times improvement in latency and bandwidth. This is much faster than Moore's Law increases. I think it's the largest increase in any aspect of system performance, in the shortest time, ever. Fully exploiting these memories is going to require making changes throughout the system, but it's going to be very exciting time."

"I can now address non-volatile storage directly from my application, just like DRAM," Rajesh Gupta says. "That's the broader vision—a future in which the memory system and the storage system are integrated into one."

Further Reading

Caulfield, A., De, A., Coburn, J., Mollov, T., Gupta, R., and Swanson, S.
Moneta: A high-performance storage array architecture for next-generation, non-volatile memories, *Proceedings of the 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture*, Atlanta, GA, Dec. 4–8, 2010.

Akel, A., Caulfield, A., Mollov, T., Gupta, R., and Swanson, S.
Onyx: A prototype phase-change memory storage array, *Proceedings of the 3rd USENIX Conference on Hot Topics in Storage and File Systems*, Portland, OR, June 14, 2011.

Mollov, T., et al.
Understanding the impact of emerging non-volatile memories on high-performance, IO-intensive computing, *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, New Orleans, LA, Nov. 13–19, 2010.

Lee, B., et al.
Phase-change technology and the future of main memory, *IEEE Micro 3*, 1, Jan.–Feb. 2010.

Qureshi, K., Srinivasan, V., and Rivers, J.A.
Scalable high performance main memory system using phase-change memory technology, *ISCA '09 Proceedings of the 36th Annual International Symposium on Computer Architecture*, Austin, TX, June 20–24, 2009.

Gary Anthes is a technology writer and editor based in Arlington, VA.

Technology

Undergrads Seek Utility in Tech

A recent survey of technology preferences among college undergraduates offers results both predictable and surprising. As for the former, it should not raise eyebrows that the college computer lab is steadily fading in relevance on campuses, especially in this age of ubiquitous mobile devices.

As for the surprises? It defies convention that spending on physical textbooks, as opposed to digital versions, remains strong. They are the second-most commonly purchased item online (clothing is first). "The adoption of e-textbooks has progressed slowly," says Eric Weil, managing partner of Student Monitor LLC, which recently published the survey research in which 1,200 full-time students at 100 campuses participated. "They vastly prefer to buy used textbooks. Why is that? For the same reason they've bought used textbooks for decades: The highlighting is already done for them."

Other notable survey findings include:

- **Computer labs are not dead. More than six of 10 students still work on school-issued computers at least once a week. However, only 31% do so at least once a day.**
- **Old-school tech has not disappeared. Microsoft Word remains the most popular computer program, used by 72% of students compared to the second-most used, Skype, at 46%. Google Docs ranks far behind Word, accessed by just 18% of undergraduates.**
- **Smaller is better. Students like used physical textbooks because they are easy to tote around. And that factor is contributing to the increase of tablet computers. Among undergraduates who plan a computer purchase in the next year, 18% plan to buy a tablet, compared to 10% who did last year.**
- **Social media keeps surging. Nearly all undergraduates (98%) have joined Facebook and three of four use Twitter. Facebook commands nearly four hours of their time every week, and Twitter accounts for 101 minutes weekly. It is not all fun and games either. Even a site like LinkedIn is frequented by 65% of college students, who typically spend more than an hour a week there.**

"It isn't just the seniors going on there looking for a job," Weil says. "It's also younger students who are seeking internships or good summer employment. It's a grown-up site. But it's a very useful learning tool for undergraduates who want to get serious about their careers."

—Dennis McCafferty