Storage-class memory: The next storage system technology

R. F. Freitas
W. W. Wilcke

The dream of replacing rotating mechanical storage, the disk drive, with solid-state, nonvolatile RAM may become a reality in the near future. Approximately ten new technologies—collectively called ***storage-class memory*** *(SCM)—are currently under development and promise to be fast, inexpensive, and power efficient. Using SCM as a disk drive replacement, storage system products will have random and sequential I/O performance that is orders of magnitude better than that of comparable disk-based systems and require much less space and power in the data center. In this paper, we extrapolate disk and SCM technology trends to 2020 and analyze the impact on storage systems. The result is a 100- to 1,000-fold advantage for SCM in terms of the data center space and power required.*

Introduction

Maintaining the performance growth rate of the 30-year-old system memory and storage hierarchy, primarily based on DRAM (dynamic RAM) and disks, has become a major challenge in the design of large-scale, high-performance computer systems. This challenge manifests itself in many ways. For example, the gap [1] between the performance (measured as latency) of disks and the rest of the system—which is already five orders of magnitude—continues to widen rapidly. In addition, the energy consumption, space usage, and cost of the memory and storage hierarchy are major obstacles to the development of exascale systems capable of 10^{18} operations per second.

To overcome these obstacles and maintain the historic growth rate in the capabilities of large-scale, high-performance computers, either significant advances in disk drives must be made or entirely new approaches to storage must be developed. Research and development efforts are underway worldwide on several nonvolatile memory technologies that not only complement the existing memory and storage hierarchy but also reduce the distinctions between memory (fast, expensive, evanescent) and storage (slow, inexpensive, permanent). An overview of these technologies appears as a companion paper [2] in this issue of the *IBM Journal of Research and Development*. One or more of these technologies may eventually replace disks and perhaps even DRAM. We call this newer group of technologies *storage-class memory* (SCM). Flash memory can be considered as an early version of SCM, and it is slowly being adopted for certain enterprise niche uses. However, its cost, write performance, and write endurance (number of times a flash bit can be written) will limit the extent to which it will replace disks on a large scale.

However, future SCM technologies will overcome these limitations and thereby compete effectively with disk drives and potentially replace them by 2020. SCM promises random and sequential I/O performance many times that of comparable disk-based systems, as well as a major reduction in space and power for the data center. The realization of SCM should give rise to a major new industry that will be on a scale similar to that of disk drives or DRAM. In 2007, the worldwide memory chip industry—including SRAM (static RAM), DRAM, and flash memory—had annual revenues exceeding $60 billion, with the NAND flash memory sector showing the largest growth [3].

The purpose of this paper is to provide an early view of the competitive relationship of SCM to disk technology in

0018-8646/08/$5.00 © 2008 IBM

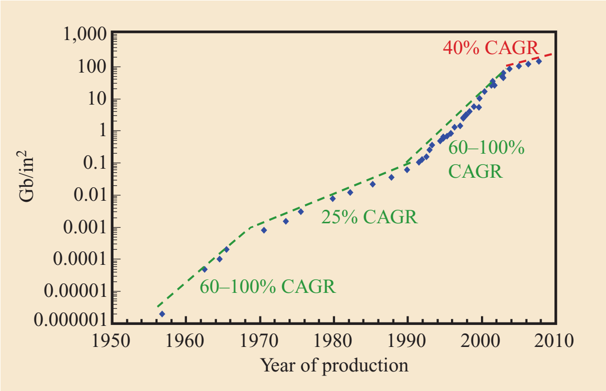

Figure 1

History of disk areal density. See References [4] and [5] for additional background.

the arena of storage systems. In this paper, we describe the trends in disk drive technology, provide a brief introduction to SCM technologies, describe SCM-based systems, extrapolate both sets of technology trends to 2020, and then compare storage systems targeted at two important application areas.

Disk technology trends

The disk has played a pivotal role in the advancement of computer systems. Its steady growth in areal density and corresponding reduction in cost per byte have enabled computers to handle increasingly ambitious applications. The aspect of disk technology of interest here is areal density and its impact on cost, capacity, performance, and power use.

Areal density, cost, and capacity

The growth in disk areal density [4] was striking in the late 1990s and early 2000s, but then it slowed. **Figure 1** shows that the compound annual growth rate (CAGR) for disk drive areal density started to drop in 2004 from approximately 100% to 40% (see red line) [5]. It will likely stay at that rate for the foreseeable future. Areal density is important, as it drives both cost and performance (bandwidth) of disks.

We believe that the cost per gigabyte (GB) of disk drives will continue to decrease, but at a reduced rate. Instead of drive capacity doubling every 12–18 months, it will now double in 24–36 months. The reduction in the initial price of a new drive—in a family of drives differing only in capacity—will likely remain at 3–5% CAGR. The 2007 high volume price estimates are \$1.00–2.00/GB for enterprise disks and \$0.30–0.60/GB for consumer disks. The recent trend, which is expected to continue indefinitely, is for the cost per gigabyte to decline at approximately 40% per year (CAGR).

The form factor of disk drives is also evolving. Currently, a transition is taking place from 3.5-inch to 2.5-inch drives. This transition is driven by a number of factors, including the proliferation of laptops and the need of the enterprise system for higher storage performance in a smaller space. Many enterprise applications are limited by the storage system performance, measured in storage I/O operations per second (iops). This number scales with the number of disk drives but does not depend on disk drive capacity. Therefore, in order to increase the storage performance, more disk drives are needed. Thus, there is a strong trend toward using physically smaller disk drives. It is likely that the next form factor transition will take place after 2015. This transition will be from 2.5-inch to 1.8-inch disks. If current areal density and packaging trends continue, then in 2020, the likely capacity of disk drives will be about 20 terabytes (TB) for a 3.5-inch drive, 10 TB for a 2.5-inch drive, and 5 TB for a 1.8-inch drive.

Performance

Disk performance is measured by bandwidth and access time, and it has been improving at a much lower rate than the areal density. Patterson [6] observes that latency improvements are lagging bandwidth improvements throughout the computing industry. Disks are no exception, which is a major problem for those applications that are more dependent on improved latency than on improved bandwidth.

Bandwidth (external data rate)

The factors that caused a decrease in the growth of the disk areal density are also causing a decrease in the growth of maximum sustainable disk bandwidth, which is proportional to the product of the head velocity and the linear density. The linear density is related to the areal density through the equation *areal density* = *linear density* × *track density*. Therefore, the growth rate of areal density is related to the growth of linear density and track density. Historically, they have both grown at nearly the same rate, with the track density usually growing slightly faster than the linear density [7]. If this trend continues and areal density growth remains at about 40% CAGR, then the linear density growth and, therefore, the annual bandwidth growth should be roughly 15%. The dotted line in **Figure 2** represents a 40% CAGR, while the shaded region depicts a range of 10–25%.

Access time (latency plus seek time)

The average access time for a disk drive is equal to the average rotational latency plus the average seek time. The average rotational latency is half the rotation period. In 2007, drives used in large systems spin at 7,200, 10,000, or 15,000 revolutions per minute (rpm). There is little expectation that drives spinning much faster than 15,000 rpm will become common within the next 10 years [8]. Thus, the rotational latency of disks will stay at or above 2 ms. Similarly, the other component of access time, seek time, is not expected to improve much anytime soon. Historically, its improvement has been less than 5% CAGR, and there are no expectations that the growth will be better in the future.

Power

The power P supplied to a disk drive is given by $P = I + M + S$. In this equation, I is the power supplied to the interface and control logic of the disk. M is the power supplied to the motor to spin the disk, overcoming the friction of spinning the disk in air. $M \sim d^{4.6} \times r^{2.8}$, where d is the diameter of the disk, and r is the rotational speed [4]. S is the power supplied when the heads are moved to a new track. It is dissipated *only* when the heads are moving. Only S varies during normal system operation.

A common approximation rule suggests that each of these draws about one third of the total disk power. Because disk motors and actuators already operate close to their theoretical efficiency limits, there is little room for managing the power of disks other than transitioning to a smaller disk form factor or shutting them down completely when not in use. However, because it takes such a long time to power up a disk drive (~20 seconds), the latter is practical only for disk-based archival systems [9]. As power becomes the central issue for data centers, this power constraint will become a significant drawback for using disks as the storage medium. A commercially viable variable-speed disk drive [10] could play a role in data centers where there are a significant number of underutilized disk drives.

Storage-class memory

The goal of SCM development is to create compact, robust storage (and memory) systems with greatly improved cost/performance ratios relative to other technologies. The defining requirements for all SCM technologies are nonvolatility, solid-state implementation (no moving parts), very low latencies (tens to hundreds of nanoseconds), low cost per bit, and physical durability during practical use. Bandwidth is not listed here as a differentiating requirement because all candidate SCM technologies offer good device-level bandwidth and the external bandwidth is mostly determined by packaging cost constraints.

Numerous materials exhibit bistable hysteretic transitions between two easily distinguishable, stable states. Such materials can be sandwiched between two orthogonal planes of parallel conductors, providing an X (wordline) and Y (bitline) addressing scheme [11]. A data cell is formed at each of the intersections of a wordline and a bitline. By activating a wordline and a bitline in a technology-specific manner, a bit in this cross-point memory is selected for either reading or writing.

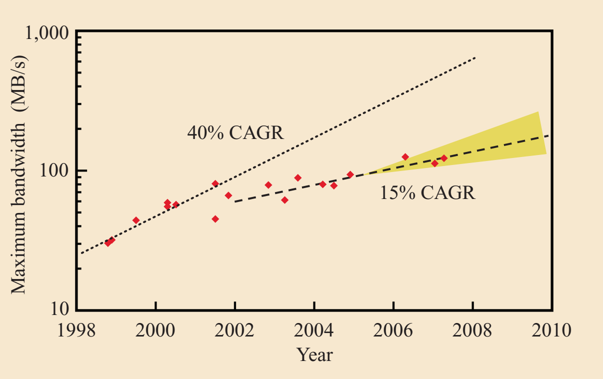

Figure 2

History of maximum bandwidth for disks.

An array of such data cells forms a memory bank, and an SCM chip contains several such memory banks. A collection of SCM chips, combined with appropriate control logic, forms a memory or storage system. Since all of the technologies under consideration operate at semiconductor speeds, the access time of an SCM chip is three to five orders of magnitude faster than a disk drive. The usable system bandwidth will be determined by how the components are packaged. SCM will enable much better tradeoffs between performance, space, and power than disk-based systems.

However, write endurance is a significant challenge for most SCM technologies. Each act of writing a bit may slightly damage a cell. A flash bit cell can be written 10^4–10^5 times, depending on the flash technology, before it becomes unusable. In comparison, DRAM chips will survive being written 10^{15} times, and a disk drive may survive being written 10^{12} times. This corresponds to being continuously written for the lifetime of the devices. SCM developers are striving for a write endurance of 10^8–10^{12} writes per cell. If writes were performed at the maximum possible write rate, an SCM cell—for most of the SCM technologies discussed here—would be worn out within a few minutes. This problem can be ameliorated by system wear-leveling schemes, as it is done for flash technology today. The notion is that by distributing the writes across the entire address space, write hotspots are eliminated. Strong wear-leveling algorithms will be important for effective SCM storage systems. Note that the write endurance problem is more

Table 1 Projected 2020 characteristics of storage-class memory devices. (SIO: start I/O.)

Capacity	1 TB
Read or write access time	100 ns
Data rate	>1 GB/s
Sustained I/O rate [1/(0.1 μs + 4 KB/1 GB/s) = 1/4.1 μs]	238,000 SIO/s
Sustained bandwidth (4 KB/4.1 μs = 975 MB/s)	975 MB/s
Write endurance	10^{12} writes

severe if SCM devices are used as memory, that is, with more direct communication with the CPU, rather than as storage elements.

The second major challenge is to scale cross-point memories down to a size that makes them cost competitive with disk drives. The industry is currently studying three techniques that are expected to yield such memories: multilayers [12], multibit cells [13], and sublithographic addressing and patterning [14]. In their article in this journal issue, Burr et al. [2] provide an overview of SCM candidate device technologies and then compare them in terms of their potential for scaling to ultrahigh areal density [2]. Of the many SCM technologies described in their paper, the one that seems to be in the best position to replace the current flash technology and serve as SCM in the next decade is phase-change memory (PCM) [15, 16]. For the remainder of this paper, all numerical estimates are based on the use of PCM.

Currently, at least 18 companies are working on PCM. The key concept involves the use of certain chalcogenide alloys (typically based on germanium, antimony, and tellurium) as the bistable storage material. These alloys exist in two stable solid phases. One phase (RESET) is amorphous and exhibits low electrical conductivity (and low optical reflectivity). The second phase (SET) is polycrystalline and exhibits two to three orders of magnitude higher conductivity (and reflectivity). For writing or reading a bit, an electrical current is addressed via the cross-point conductors to a small amount of phase-change material located at a specific cross-point intersection. A small current can be used to measure the resistance (i.e., reading a bit). A larger current, via ohmic heating, is used to write a bit. By controlling the temperature and the duration of the heating, either phase can be obtained. In either case, the phase can be changed in a few tens of nanoseconds.

PCM has three major advantages over flash technology. First, PCM has much better size scaling; flash data retention times decrease more than exponentially when gate-oxide thickness [17] is decreased. Second, PCM is write-in-place, while flash is not. With PCM, a cell can be written repeatedly without any intervening operations. Flash requires that a cell be erased before it can be written again. For random write operations, this contributes to a major performance decrease. Flash random write times are measured in milliseconds, whereas PCM write times are on the order of 100 ns. Third, the write endurance of PCM is several orders of magnitude better than that for flash.

Initially, the cost per bit for all SCM technologies will be much higher than that for disks and comparable to that for flash. Thus, SCM will be used in systems in which size (e.g., in mobile devices), performance, and/or reliability in harsh environments are paramount. Over time, the cost per bit for SCM will decrease dramatically and may—by the middle of the next decade—approach that of enterprise disks of that time. This rapid price drop is plausible because SCM density can be increased with the various methods discussed earlier. We expect the introduction of PCM to follow three phases: 1) availability of a system-usable PCM chip in 2008 or 2009, 2) availability of PCM-based enterprise appliances and systems by 2011 or 2012, and 3) complete replacement of disks in most systems by 2020. The exceptions are systems that need ultralow-cost storage, such as archival and consumer video applications

A PCM-based SCM is expected to have roughly the specifications shown in **Table 1** by 2020. For the system comparison section below, we use these parameters. We have assumed that the data block size is 4 KB and defined the sustained I/O rate as the reciprocal of the sum of the access time and the transfer time for a data block, the block transfer time as the block size divided by the module transfer rate, and the sustained bandwidth as the block size divided by the sum of the access time and the block transfer time.

SCM system impact

SCM-based high-performance systems

SCM reduces the boundaries between storage and memory and could be used for many tasks in the memory and storage hierarchy. However, the finite write endurance of SCM (and flash) requires careful attention [18] to how these devices are used. The higher performance of SCM makes it possible to wear out SCM very quickly, unless the system architecture guards against such calamities.

Table 2 depicts a memory and storage hierarchy, with associated access times. Note that SCM is associated with a large gap between DRAM and disks. As shown, if access time is measured in CPU clock cycles, then the L1, L2, and L3 caches and the DRAM-based primary

Table 2 Hierarchy of latencies (access times) in a computer system.

CPU cycles	*Device*	*Comment*
10^7–10^8	Disk	Nonvolatile, slow, and inexpensive
	—Gap in access time—	
10^3	SCM	Nonvolatile, fast, and inexpensive
10^2	DRAM	Volatile, fast, and expensive
10–100	L2 and L3 cache	Volatile, fast, and expensive
1	L1 cache	Volatile, fast, and expensive

memory are at a distance between 1 and 100 cycles from the CPU, while the disk is at a distance between 10^7 and 10^8 cycles. SCM is at 10^3–10^4 cycles with performance near DRAM, but cost and persistence near disks. Tape is at 10^{10}–10^{12} cycles from the CPU and is not considered here.

A taxonomy of SCM uses is given in **Table 3**. Note that SCM can be used as a new cache layer (e.g., L4 cache), positioned just beyond the main (DRAM-based) memory. The primary justification is the expected low price per bit, compared to DRAM. However, this stack position exposes SCM to the potential for high wear. SCM can also be used as part of main memory, with user- or compiler-controlled addressing. As shown in **Figure 3**, a distinction is made between using SCM logical addressing or SCM physical addressing. These distinctions are of great practical importance because they determine which part of the computer industry (e.g., CPU vendors, independent chip houses, or software vendors) will control the system use of SCM. Three major SCM options may be considered. First, a separate hardware controller may be used for SCM addressing and for wear leveling (e.g., distributing write operations to increase write endurance). Second, wear level and existing virtual memory (VM) translation engines may be combined into one controller. Third, the runtime kernel software may be modified to handle wear leveling, using mechanisms similar to VM management [19]. This is simple and flexible but introduces latency. In all cases, a smart compiler, which is aware of the distinction between DRAM and SCM space and possibly supplied with compiler hints from a user, could allocate variables to either type of memory.

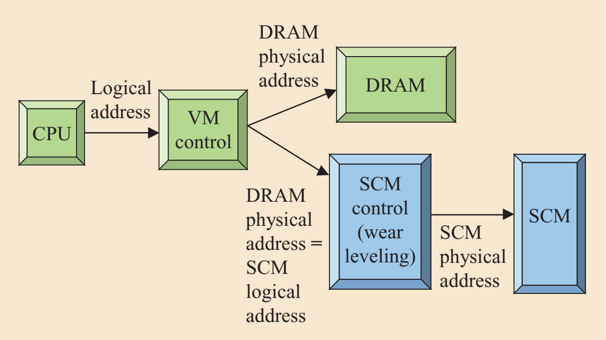

Figure 3

Translation of addresses in a virtual memory system that utilizes both DRAM (dynamic RAM) and storage-class memory (SCM). The virtual memory manager translates some of the CPU logical address space into physical addresses for the DRAM. Addresses destined for SCM undergo a second translation, which takes wear leveling into account. (VM: virtual memory.)

As depicted in Table 3, SCM can be used like traditional storage, that is, addressed as a block device and part of a (file) namespace. It may use legacy I/O protocols such as SAS [Serial-Attached SCSI (Small Computer System Interface)], SATA (Serial Advanced Technology Attachment), or Fibre Channel protocols. This approach is simple but will not fully exploit the performance of SCM.

SCM can be used as a new storage device connected via new interfaces. This application may be optimized for use as a memory-mapped device, which is an important paradigm for emerging data-centric applications. Additionally, SCM would make an excellent paging device. The high performance of SCM may restore the usefulness of the VM concept to high-performance computing. (Paging based on the use of disks is so slow that performance-critical applications avoid the use of VMs entirely.) Finally, SCM could be built as a fast cache inside a storage controller for disks. Because SCM is nonvolatile, it can safely buffer write operations. All of these options will make the address space much flatter, compared to current systems. (Here, the concept of a “flat address space” implies a memory and storage system in which all bytes can be addressed in the same manner and in the same access time.)

As with disks and DRAM, because SCM will have occasional runtime errors or manufacturing defects, mechanisms for dealing with these errors must be invented. These mechanisms may range from device-internal correction to controller-based ECC-type circuits to table-based, software-controlled mapping mechanisms, linked with the wear-level control engines. The optimal choice will depend on the use case.

Finally, the fact that read accesses do not wear out SCM, whereas writes do, suggests making read/write access to SCM asymmetrical. The unit of read access should be a single word.

Table 3 Taxonomy of SCM system uses. (NVRAM: nonvolatile RAM; VM: virtual memory.)

Access mode	*Use mode*	*Comments*
Address oriented (memory)	Cache (e.g., Level 4)	Wear level is critical
	Main memory, version 1	Separate SCM controller
	Main memory, version 2	Integrated RAM/SCM controller
	Main memory, version 3	SCM wear level managed by software and VM manager
Block oriented (storage)	Via legacy I/O buses	Wastes SCM performance
	Via new interfaces	Good for memory mapping
	Paging device	Very promising use
	I/O cache for a disk controller	Act as NVRAM

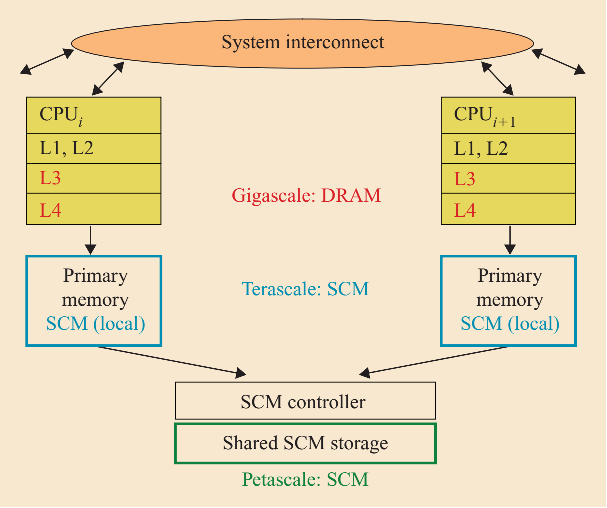

Figure 4

Possible SCM-based high-performance system architecture with gigabytes in the caches, terabytes in the primary memory, and petabytes in the shared storage. The SCM controller provides dual-port access (for fault-tolerance reasons) to a petascale shared SCM storage unit. The controller manages wear leveling and controls physical access to the SCM modules. Logical sharing is managed by the CPUs and not by the SCM controller.

Making the optimal choice of the block size for keeping track of wear level is difficult. Large blocks keep wear-leveling tables small but needlessly label bits worn even if they have not been actually written. Note that the actual wear state of each bit cell could be occasionally measured and used in the wear-level algorithms. This will be a rich area for future inventions.

Figure 4 shows an example of a system built with SCM technology. DRAM is used only for giga-scale caches (gigabytes). SCM is used both as the main memory (terabytes) and as permanent storage (petabytes). The latter is shared across several nodes for fault tolerance and programming-model reasons. All sharing occurs via the high-speed interconnect linking the node CPUs. The SCM for the main memory and the storage applications may be separate implementations, differing in materials with respect to write speed and volatility materials tradeoffs and how errors are handled. Countless variations of this basic architecture are conceivable.

The realization of very large, flattened memory address spaces and very fast I/O devices will greatly improve speeds on practical applications, presumably greatly reducing the gap between peak and sustained performance. Also, the simple predictable behavior of SCM, compared to disks, will simplify performance tuning. It will eliminate many of the unexpected interactions seen today when optimizing the memory and storage hierarchy.

We have described both disk and SCM technology and forecasted their attributes in 2020. Now, we discuss how these technologies will have an impact on future systems and applications. We have chosen power and floor space in the data center as key metrics for comparing disks and SCM-based storage systems under several representative workloads.

Workloads

The storage needs of computer systems are typically measured by three metrics: I/O rate, bandwidth, and capacity. The I/O rate measures how many I/O requests per second are issued by the system and is independent of how many bytes are transmitted per request. Bandwidth can be viewed as the product of I/O rate and average request size. Applications for high-performance systems fall into one of two broad classes: compute-centric and data-centric workloads.

Compute-centric workloads

Historically, researchers exploring large compute-centric problems have used the most powerful computer systems

Table 4 Storage system projections.

2007		*2020*			
Performance requirement	*CAGR*	*Performance requirement*		*Disk*	*SCM*
Compute-centric: 0.4 TB/s of sustained storage bandwidth	70%	0.4 PB/s	*Devices*	1.3 million disks	406,000 modules
			Space	6,192 sq ft	85 sq ft
			Power	6 MW	41 kW
	90%	1.7 PB/s	*Devices*	5.6 million disks	1.7 million modules
			Space	26,292 sq ft	300 sq ft
			Power	25 MW	173 kW
Data-centric: 2 million SIO/s	70%	2.0 Giga-SIO/s	*Devices*	5 million disks	8,000 modules
			Space	23,220 sq ft	12 sq ft
			Power	22 MW	1 kW
	90%	8.4 Giga-SIO/s	*Devices*	21 million disks	35,000 modules
			Space	98,568 sq ft	12 sq ft
			Power	93 MW	4 kW

in the world. These researchers seek to solve problems [20] involving weapons design, biotechnology, climate and weather forecasting [21], structural engineering, oil exploration, automotive accident studies, medicine, and many other topics. Most of these applications focus on solving coupled partial differential equations, and their defining characteristics are that CPU and memory performance is critical, whereas I/O performance is often less critical.

As a typical system used for compute-centric workloads, consider the ASC Purple [22] system of the U.S. government. It has more than 12,000 processors, 50 TB of primary DRAM memory, and 2 petabytes (PB) of storage. Typical computation times are very long (e.g., days), and a major use of storage is to occasionally create a checkpoint of the system state to protect against failures. Even in this ancillary role, the demands on the storage system are significant. For example, the requirement could be to save a checkpoint (write) of a significant fraction of the 50 TB of primary memory to disks every 5 hours and require no more than 5 minutes for this activity. Thus, 1/60 of wall-clock time is sacrificed for checkpointing. This translates into a storage system bandwidth requirement of 122 GB/s. This entails writing data to 8,000 data disks (10,000 total disks, when including redundancy data) at a sustained bandwidth of more than 15 MB/s per disk. Although this bandwidth may seem low for a disk, if one considers the disk seek time, latency, and other relevant factors, then it is actually at the limit of the capabilities of the drive.

Data-centric workloads

Data-centric problems involve analyzing or creating large amounts of data. Examples of such problems include intelligence and surveillance work, Google-type searches, reconnaissance, cryptographic analysis, analysis of data from large scientific experiments (such as the Large Synoptic Survey Telescope [23] and the CERN Large Hadron Collider [24]), analysis of extremely large-scale graphs [25] (e.g., graphs with hundreds of billions of edges that occur in the search for social or terrorist networks [26]), image and video applications, and multilingual text analytics. Such applications have become increasingly important. Unlike most compute-centric applications, data-centric applications demand very high I/O rates. Additionally, these applications may require large numbers of small reads and writes randomly throughout the storage space, and their performance is governed by access times to I/O devices. Given that disks are slow, this is a serious problem. Throughout the industry, a great deal of effort is being directed toward finding algorithms and techniques to compensate for the very long access time of disks by multi-threading, by including more memory in the system, and by caching frequently accessed items. In 2007, state-of-the-art examples of such systems must perform at least 2,000,000 storage operations per second. Providing this I/O rate with disks requires a heroic effort.

Storage requirement scaling

Over the years, the computational requirements for compute-centric applications have continually grown. This growth is not tied directly to the growth of the underlying technology but reflects the *needs* of the user of the system. An example of this can be found in Grider [27], which shows a growth in computation of approximately 70% per year and that is expected to continue at this rate. Data-centric applications require even more system performance growth. We next discuss the system implications of this scaling.

Analysis

The results of our analysis of power and space requirements are shown in **Table 4**. The table includes two classes of systems. One is optimized for compute-centric (storage-bandwidth-limited) applications, and the other is optimized for data-centric (I/O-rate-limited) applications. The first column specifies the system and the requirement for 2007. The next column gives CAGR values for 2007. Two CAGR values for each class of system are shown. Historically, the growth rate has been at the upper end of the range shown. The columns labeled Disk and SCM show results for systems based on disks and on SCM.

In performing the analysis, the following additional assumptions have been made for the 2020 timeframe. An enterprise disk drive will use a disk with a diameter of 1.8 inches and sustain a bandwidth of 300 MB/s and an I/O rate of 400 start I/O operations per second (SIO/s). Furthermore, the disk drive will average 4 W of power, and 256 drives will be packaged in a standard 4U (7-inch-high) rack drawer. Ten such 4U drawers will be packaged in a standard 19-inch rack.

The packaging assumptions for SCM are as follows. The SCM chip will be packaged in a 15-mm × 15-mm × 5-mm module. Its performance characteristics are shown in Table 1. An active module will dissipate 100 mW, 3,200 SCM modules will be packaged in a standard 2U drawer, and 21 of those drawers will be packaged in a standard rack.

Our assumption for floor space of a standard rack is 12 square feet [2-ft wide × (3 ft for rack depth plus 3 ft of clearance for maintenance)]. The bandwidth of a large-scale compute-centric system is 400 GB/s, and the I/O rate of a large data-centric system is 2,000,000 SIO/s. These represent the starting conditions for the analysis in 2007. Both are evaluated at growth rates of 70% and 90%.

Results

An easy way to understand the dramatic advantage of SCM technology is to calculate the amount of floor space required for storage. Using the above requirements for I/O rate and other relevant factors, the floor space needed for SCM-based systems could be 1/1,000th of that needed for a disk drive-based system with the same I/O performance. While NAND flash technology will be better suited for many storage applications than disk technology, its performance will not match that of SCM technology. This is particularly true for data-centric applications with high write rates.

Conclusion

This paper presents an analysis of the storage requirements for data center applications through the year 2020. Given the foreseeable trends in the development of disks, we have reached the conclusion that disks cannot be used as the storage medium for reasons of insufficient performance and excessive power and data center floor space use. Disks must be replaced by technologies that have the proper combination of performance, cost, power, and reliability in order to fulfill the needs of future high-end applications. SCM technology (based on phase-change memory) as described in this paper promises to provide an effective storage system that is 50–1,000 times the I/O performance of disk drive-based systems.

*Trademark, service mark, or registered trademark of International Business Machines Corporation in the United States, other countries, or both.

**Trademark, service mark, or registered trademark of SPARC International in the United States, other countries, or both.

References

1. J. Gray and A. Reuter, *Transaction Processing: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA, 1993.
2. G. W. Burr, B. N. Kurdi, J. C. Scott, C. H. Lam, K. Gopalakrishnan, and R. S. Shenoy, "Overview of Candidate Device Technologies for Storage-Class Memory," *IBM J. Res. & Dev.* **52**, No. 4/5, 449–464 (2008, this issue).
3. C. Hirst, R. Gordon, J. Unsworth, and A. Norwood, "Forecast: Memory, Worldwide, 2001–2011 (3Q07 Update)," Gartner Group Report, 2007; see *http://www.gartner.com/DisplayDocument?id=522309*.
4. E. Grochowski and R. D. Halem, "Technological Impact of Magnetic Hard Disk Drives on Storage Systems," *IBM Syst. J.* **42**, No. 2, 338–346 (2003).
5. E. Grochowski, "HDD Data Systems Report," Hitachi Global Storage Technologies; see *http://www.hgst.com/hdd/hddpdf/tech/hdd_technology2003.pdf*.
6. D. A. Patterson, "Latency Lags Bandwidth," *Commun. ACM* **47**, No. 10, 71–75 (2004).
7. G. J. Tarnopolsky, "Perfect Devices: The Amazing Endurance of Evolving Hard Disk Drives," *PARC Forum*, Palo Alto, CA, July 15, 2004; see *http://www.parc.com/cms/get_article.php?id=332*.
8. Fujitsu, "Analyzing the Trends in the Enterprise Hard Disk Drive Industry," white paper (2006); see *http://www.fujitsu.com/downloads/COMP/fcpa/hdd/enterprise-hdd-single_wp.pdf*.
9. D. Colarelli, D. Grunwald, and M. Neufeld, "The Case for Massive Arrays of Idle Disks (MAID)," *USENIX Conference*

on File and Storage Technologies, Monterey, CA, January 28–30, 2002; see *http://www.usenix.org/events/fast02/wips/colarelli.pdf*.

10. S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke, "Reducing Disk Power Consumption in Servers with DRPM," *IEEE Computer* **36**, No. 12, 59–66 (2003).
11. P. Arnett and J. Chang, "Non-volatile Diode Cross Point Memory Array," U.S. Patent No. 3838405, 1973.
12. "3D Integration," *IEEE Design & Test of Computers* **22**, No. 6 (2005, entire issue).
13. B. Eitan, R. Kazerounian, A. Roy, G. Crisenza, P. Cappelletti, and A. Modelli, "Multilevel Flash Cells and Their Trade-offs," *International Electron Devices Meeting, IEDM Technical Digest*, December 8–11, 1996, pp. 169–172.
14. K. Gopalakrishnan, R. S. Shenoy, C. T. Rettner, R. S. King, Y. Zhang, B. Kurdi, L. D. Bozano, et al., "The Micro to Nano Addressing Block (MNAB)," *IEEE International Electron Devices Meeting, IEDM Technical Digest*, December 5–7, 2005, pp. 471–474.
15. S. Raoux, G. W. Burr, M. J. Breitwisch, C. T. Rettner, Y.-C. Chen, R. M. Shelby, M. Salinga, et al., "Phase-Change Random Access Memory: A Scalable Technology," *IBM J. Res. & Dev*. **52**, No. 4/5, 465–479 (2008, this issue).
16. S. Hudgens and B. Johnson, "Overview of Phase-Change Chalcogenide Nonvolatile Memory Technology," *MRS Bull.* **29**, No. 11, 829–832 (2004).
17. L. C. Tran, "Challenges of DRAM and Flash Scaling – Potentials in Advanced Emerging Memory Devices," *Proceedings of the Seventh International Conference on Solid-State and Integrated Circuits Technology*, October 18–24, 2004, pp. 668–672.
18. J. D. Aasheim and Y. Yang, "System and Method for Achieving Uniform Wear Levels in a Flash Memory Device," European Patent Application No. EP20030000541, 2003.
19. D. Bovet and M. Cesati, *Understanding the Linux Kernel*, Third Edition, O'Reilly and Associates, Sebastopol, CA, 2001, pp. 45–63.
20. S. L. Graham, M. Snir, and C. A. Patterson, *Getting Up to Speed: The Future of Supercomputing*, National Academic Press, Washington, DC, 2005.
21. U.K. Meteorological Office Report, "The Unified Model," 2007; see *http://www.metoffice.gov.uk/research/nwp/numerical/unified_model/*.
22. Advanced Simulation and Computing Purple, 2005; see *https://asc.llnl.gov/computing_resources/purple/*.
23. R. F. Green, W. N. Brandt, D. E. Vanden Berk, D. P. Schneider, and P. S. Osmer, "AGN Science with the Large Synoptic Survey Telescope," *Astronomical Society of the Pacific Conference Series*, Volume 373, *The Central Engine of Active Galactic Nuclei*, 2007, pp. 707–714.
24. A. Barr, "The Large Hadron Collider"; see *http://www.so.stfc.ac.uk/lhcresources/BarrPressOfficers3.ppt*.
25. B. Hayes, "Graph Theory in Practice," *Am. Sci.* **88**, No. 2, 104–109 (2000).
26. V. E. Krebs, "Uncloaking Terrorist Networks," *FirstMonday* **7**, No. 4 (2002); see *http://www.firstmonday.org/issues/issue7_4/krebs/*.
27. G. Grider, "The ASCI/DOD Scalable I/O History and Strategy," *Los Alamos National Laboratory*, Report No. LAUR 042787, May 2004; see *http://ardra.hpcl.cis.uab.edu/sfast04/presentations/Grider.pdf*.

Received September 21, 2007; accepted for publication October 18, 2007; Internet publication July 23, 2008

Richard F. Freitas *IBM Almaden Research Center, 650 Harry Road, San Jose, California 95120 (freitas@almaden.ibm.com)*. Dr. Freitas is an IBM Research Staff Member at the IBM Almaden Research Center. He received his Ph.D. degree in EECS from the University of California at Berkeley in 1976. He then joined the IBM RISC computing group at the IBM Thomas J. Watson Research Center, where he worked on the IBM 801 project. He has held various management and research positions in architecture and design for storage systems, servers, workstations, and speech recognition hardware at the IBM Almaden Research Center and the IBM T. J. Watson Research Center. His current interests include exploring the use of emerging nonvolatile solid-state memory technology in storage systems for commercial and scientific computing.

Winfried W. Wilcke *IBM Almaden Research Center, 650 Harry Road, San Jose, California 95120 (winfriedwilcke@us.ibm.com)*. Dr. Wilcke is Program Director at the IBM Almaden Research Center. He received a Ph.D. degree in nuclear physics in 1976 from the Johann Wolfgang Goethe Universitaet, Frankfurt, Germany, and worked at the University of Rochester, Lawrence Berkeley Laboratory and Los Alamos on heavy-ion and muon-induced reactions. In 1983, he joined the IBM T. J. Watson Research Center in New York, where he managed the first two MIMD message-passing supercomputer projects of IBM Research (Victor and Vulcan), which were the precursors of the very successful IBM SP* supercomputers. In 1991, he joined HaL Computer Systems, initially as Director of Architecture and later as CTO. With Sun Microsystems, his team created the 64-bit SPARC** architecture. Later, he rejoined IBM Research in San Jose, California, where he launched the IBM IceCube project, which became the first funded spinoff venture of IBM Research. Recently, Dr. Wilcke became engaged in research on storage-class memories and future systems based on such memories. In addition to his industrial work, he has published more than 100 papers, has coauthored numerous patents, and is active in aviation.