KURT CHAN

a comparison of disk drives for enterprise computing

Kurt Chan is a Technical Director at Network Appliance, responsible for storage subsystems.

kurtc@netapp.com

FOR END USERS, THE FIVE MOST externally visible characteristics of a disk drive are capacity, price, interface type (e.g., SCSI, ATA, Fibre Channel, SATA), performance (e.g., access time, I/Os per second, sustained transfer rate), and reliability (e.g., MTBF or unrecoverable read error rate). When evaluating a drive for a particular application, these attributes carry varying weight. We'll examine how these attributes are related in real disk drive implementations, what applications are best suited to specific drive types, and what the future holds for disk storage in the enterprise.

Disk Drive Economics

The disk drive business has undergone heavy consolidation over the past decade, and even the survivors operate on relatively thin margins compared with those who integrate drives into enterprise systems. Here's a chart of some disk drive manufacturer gross margins for 2005, along with some major storage integrators [1]:

Disk Drive Manufacturer	*Gross Margin*
Maxtor	11.1%
WD	18.4%
Seagate	25.1%

Disk Drive Integrator	*Gross Margin*
Dell	17.8%
EMC	53.7%
NetApp	61.3%

Disk Drive Integrator	*Units (2004)*
Dell	16.1%
EMC	0.6%
NetApp	0.5%

Source: *IDC Worldwide Disk Storage Systems Market Forecast and Analysis*, 2002-9

Note that although EMC and NetApp have superior gross margins, Dell accounted for almost 15 times the unit shipments of both companies put together—16.1% market share versus 1.1%. This is because the volumes of the consumer and desktop markets dwarf the volume associated with the enterprise storage market. Furthermore, overall enterprise HDD revenue has remained relatively flat over the past 3–4 years, and cost/GB enterprise disk pricing has dropped about fourfold in the past four years. This means that, to maintain revenue, drive vendors must offer higher and higher capacity drives for about the same unit cost, which explains the speed at which we learn new Greek prefixes. (Terabyte disks will be commonplace by the end of the decade, and petabyte configurations are now possible.) These economic factors will be important in understanding the target designs of various drive types.

Classifying Disk Drives by Application

While a growing number of disk drives are finding their way into mobile and consumer appliances (e.g., notebooks, music and video recorders, personal electronics), disk drives for the computing industry are segmented into enterprise and desktop applications. Also arising is a new segment called "nearline enterprise" that combines some of the attributes of the classic desktop and enterprise markets.

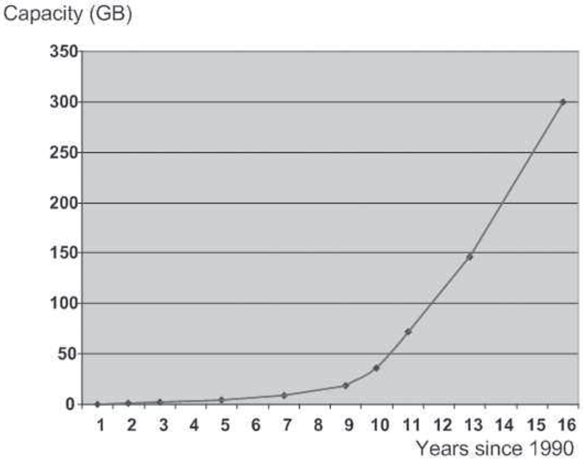
Application Attribute	*High-Performance Enterprise*	*Nearline Enterprise*	*Typical 2006 Desktop*
Rotational speed (rpm)	15,000	7,200	5,400–7,200
Interface	FC, SAS	SATA	SATA
Avg Power:			
operating	18–20 W	10–13 W	8–12 W
idle	12–14 W	7–9 W	6–9 W
Nonrecoverable read errors per bits read	1 sector per 10^{15}–10^{16}	1 sector per 10^{14}–10^{15}	1 sector per 10^{14}
Serial link rate (Gb/s)	2–4 FC, 3.0 SAS	3.0 SATA	1.5–3.0 SATA
Noise (ISO 7779, bels)			
idle	3.5–3.8	2.8–3.4	2.5
performance seek	4.3–5.9	3.5–3.9	3.1–3.7
Capacities (2006)	37–174 GB	320–500 GB	160–320 GB
Performance:			
sustained transfer	58–98 MB/s	35–65 MB/s	32–58 MB/s
average seek	3–4 ms	8–9 ms	8–10 ms
Relative price per GB	5–10x	1.5x	1x

Notable niches include 300 GB, 10k rpm FC, and 150 GB; 10k rpm SATA drives exist, but are not as broadly sourced among vendors.

Capacity

Although the capacities of each drive category will change over time, the lowest capacities are found in the enterprise markets, where performance is more important than capacity. The highest capacities are found in the nearline market, where disks are sometimes used for secondary storage, replacing tape for disk-to-disk backup applications or for storing less frequently used data that still require online access. The desktop market, where cost/GB is the lowest, focuses on the capacities—these typically lie somewhere between performance and nearline enterprise capacities, and strong discounting takes place as inventory is purged from one capacity generation to the next.

Even though SCSI/FC disk drive capacity has been growing exponentially at a compound annual growth rate of 53.2% over the past fifteen years, it has slowed dramatically over the past five.[2] Whereas capacity would normally double every 18–19 months given trends from the early 1990s, the last five years of data indicate we are doubling capacity only every 29–30 months. One of the reasons for this change is the need to balance reliability with capacity. As a product generation matures, the various electromechanical margins are eroded as capacities and performance increase. Decreasing head fly heights and increasing spindle speed and platter count all make it more difficult to maintain MTBF and unrecoverable error rate (UER) specifications. The ceilings encountered in recent years are partly related to maintaining the same or better reliability with disk drives spinning 50–100% faster, thus generating more heat and mechanical stresses. This is another reason why the highest capacity drives are not found in the performance enterprise, but, rather, in the desktop and nearline categories. This year, perpendicular recording will provide a new generation of drives with more margin, and capacity growth should improve as a result.



Year	*Capacity (GB)*	*Rate of Increase*	*Annual Rate of Increase*
1990	0.5		
1991	1	100.0%	100.0%
1992	2	100.0%	100.0%
1994	4	100.0%	41.4%

Source: "Why Tape Won't Die," *Enterprise Storage Forum*, June 16, 2005

Power

Power is another area of differentiation and generally increases in proportion to performance. Lower-speed drives consume less power, make less noise, and generate less heat, placing less demand on air conditioning. But they also provide lower sustained transfer rates and I/Os per second compared to performance enterprise drives. However, for many applications that do not demand high I/O per second rates, SATA drives are often a better choice. Archived email, digital photographs, or archived customer records do not require high transaction rates, and using high-performance enterprise drives for such bulk information can be wasteful. Although power differences may not seem significant, if a large disk user such as Google or Yahoo had 1,000 drives running 24/7, the difference in electricity costs between performance and nearline disk drives could amount to more than a quarter of a million dollars a year in electricity for power and cooling.

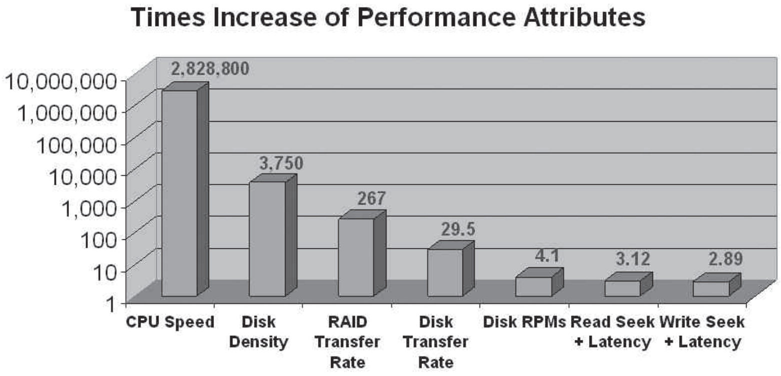
Reliability

A UER on SATA of 1 in 10^{14} bits read means a read failure every 12.5 terabytes. A 500 GB drive has 0.04E14 bits, so in the worst case rebuilding that drive in a five-drive RAID-5 group means transferring 0.20E14 bits. This means there is a 20% probability of an unrecoverable error during the rebuild [3]. Performing the same calculation for a 174 GB enterprise drive with a UER of 1 in 10^{15}, we get a 1.2% probability of data loss. Although SATA is expected to reach a UER of 10^{-15} by 2007, and enterprise drives 10^{-16} in the same timeframe, corresponding to 2% and 0.1%, respectively, this is still unacceptably high for many enterprise applications.

This phenomenon is not going away—as drives get larger, the problem becomes worse because there are more bits to move in a rebuild. Furthermore, product reliability can vary greatly among vendors as well as among product families from the same vendor. Instead of relying on advertised average failure rate (AFR), MTBF, and UER numbers from vendors, storage integrators tend to use their own empirical information to assess overall product quality and determine the necessary data protection methods. Since all drive integrators work with basically the same disk drives, what storage integrators are looking for is a means of making customer data availability more immune to drive reliability. Whereas reliability continues to be an important metric to control support and maintenance costs, measures such as double-parity RAID, full mirroring, rebuilding only used capacity, end-end checksums, and background media scans can help make the differences in reliability among drive families less important when it comes to ensuring overall customer data availability.

Performance

Disk drive performance in general has been relatively static compared to CPU clock speed and areal density growth, but it remains a meaningful differentiator between FC/SAS and SATA drives.



All 15k rpm drives and almost all 10k rpm drives available today have only FC, SCSI, or serial attached SCSI (SAS) interfaces. Enterprise drive suppliers in general have been reluctant to rush toward providing 10k speeds in a SATA drive, to avoid cannibalization of their high-margin markets as well as to keep costs low for their volume markets. Because many OLTP enterprise applications are limited in performance by IOP rates, by putting low-speed drive assemblies behind SATA interfaces the drive industry will remain segmented, barring any new designs that fill the gap between low-cost SATA and high-performance SAS drives. It might be possible to construct a high-performance OLTP system with "half-speed" drives, but the infrastructure and connectivity costs would make the solution impractical at the high end. However, important issues for the disk industry involve the amount of enterprise data being created that does not demand high-performance FC/SAS storage, and whether or not end users will begin matching their data to storage attributes using Information Lifecycle Management (ILM) and other tools to help lower their disk hardware costs. The recent growth of nearline SATA storage is evidence that users are becoming more aware of these options.

Rotational vibration plays a role in performance as well. Mechanical interferences caused by vibration patterns increase seek time, since it takes longer for heads to settle on track in the presence of severe vibration. Also, if the actuator vibrates off track, this can result in read retries and aborted writes. Since device driver timeouts can be lengthy, even a small number of retries can prove costly to performance. It's not unusual to see desktop drives drop to 50% of their nominal peak performance in the presence of 10 rad/s^2 of vibration, whereas enterprise drives might see no drop-off until around 15 rad/s^2 and might hit 50% of their nominal peak performance at 40 rad/s^2. Rotational vibration is also exacerbated by random operation and bursty workloads—the kind often found in enterprise high-OLTP traffic applications. More stringent rotational vibration specifications may be needed for SATA cabinets to ensure that performance remains at expected levels.

Finally, tests have shown that drives designed for desktop workloads can fail more frequently when exposed to heavier workloads. When Seagate performed accelerated life testing on three groups of 300 desktop drives while exposing them to high-duty-cycle sequential workloads, these drives failed twice as often as when they were exposed to normal desktop workloads. And, when exposed to random server workloads, they failed four times as often [4]. If nearline systems are deployed in the wrong workload environments without the proper data protection precautions, loss of data availability could result.

Interfaces

Over the past five years there has been a rapid adoption of serial disk interfaces over their parallel counterparts. Virtually no new computer designs are incorporating parallel SCSI or ATA today, and disk drive manufacturers will ramp down their production of parallel interfaces as demand lowers for legacy applications. The move to serial interfaces has been motivated by several factors: the inability of scaling parallel cables in both speed and distance, the cost and bulk of parallel cables and connectors in embedded desktop applications, the larger number of devices supported by serial protocols, and the ability to support more than one disk type over the same wire protocol.

FIBRE CHANNEL

The most broadly networked disk protocol is Fibre Channel. At the high end, 256-port nonblocking switches are available from multiple vendors, with 4 Gb/s and multi-kilometer distances supported through various copper and fiber-optic cabling options. At the low end, Fibre Channel Arbitrated Loop switches are available for interconnecting disk drives within RAID or disk enclosures over high-speed backplanes. Although the Fibre Channel architecture makes it convenient for connecting drives directly to initiators without protocol conversion, Fibre Channel as a storage system interface carries more momentum than as a disk drive interface. Part of the reason is the realization that the disk drive doesn't need to have as much network intelligence as is required by the Fibre Channel standards. Furthermore, bridging and RAID technologies are becoming more prevalent, allowing Fibre Channel to be used where its distance and multi-initiator capabilities are best leveraged—at the server interface—while allowing the disk drive interface to be chosen independently.

Fibre Channel as a disk drive interface is expected to level off in volume owing to the rise of both SAS (at the high end) and SATA (in nearline) beginning in 2007. One reason is that although only a few vendors are committed to producing Fibre Channel drives, almost every drive vendor is offering both SAS and SATA, making for increased competition. Also, SAS will offer the same performance characteristics as Fibre Channel, with the option of tunneling SATA protocols over the same physical and link layers.

SATA

One of the motivating factors for SATA was bandwidth. The maximum theoretical limit for parallel IDE interfaces was 133 MB/s. The 1.5, 3.0, and 6.0 Gb/s interfaces defined for SATA correspond to 150, 300, and 600 MB/s, offering a growth path that parallel interfaces could not match. SATA was also looked upon as an opportunity for nonenterprise drive vendors to gain a toehold in the enterprise space. A number of features were added to enable this:

- Native Command Queuing (NCQ) with scatter/gather features to improve random I/O performance
- 32-bit CRC checking for data and commands
- Hot-plug, blind-mate connectors for active sparing in RAID environments
- Point-to-point cabling versus "daisy-chaining," and SAS physical layer support

- The definition of port multipliers, allowing the connection of up to 15 disks to the same port
- Active–passive port selectors and active–active port multiplexors that provide dual-initiator options for higher availability

SAS

SAS and SATA are unique in that although SATA can be used to connect initiator ports directly to target ports in a point-to-point fashion for embedded desktop applications, the SAS protocol was defined to support both SAS and SATA drives over the same interconnect network. The same underlying physical and link-layer protocols support both interfaces, which presents a unique and compelling value proposition for many storage integrators. For the first time, both performance-oriented SAS and value-oriented SATA drives can be supported using the same cable plant.

Three transport protocols are supported over the SAS physical and link layers:

- Serial SCSI Protocol (SSP), which defines the mapping of SCSI commands over the link layer. Frame formats are based on Fibre Channel Protocol.
- Serial ATA Tunneling Protocol (STP), which defines connection delimiters, frames, and flow control unique to SATA devices.
- Serial Management Protocol (SMP), which adds management functions for the SAS expanders (circuit switches that distribute SAS traffic) using simple request-response functions related to discovery, status, and low-level hardware control.

SCSI Application	*SATA Application*	*Management Application*	*Architecture Defines*
SSP Transport	STP Transport	SMP Transport	Framing and information units
SSP Link	STP Link SAS Link	SMP Link	Encoding, primitives, flow control, connection management
	SAS Phy		Cables, connectors, electrical

FC VERSUS SAS DISKS IN THE ENTERPRISE

SAS is growing at the expense of SCSI, which was a premeditated outcome for early industry supporters of SAS. What perhaps was not expected was the rate at which SAS would gain in popularity at the expense of FC. Although this has not happened yet, both IDC and Seagate market research expect that within the next 12–18 months, storage suppliers will be shipping more SAS+SATA drives than FC+SCSI to enterprise customers, and within a year after that, two-thirds of enterprise drive shipments will be SAS+SATA. Considering how new these interfaces are, that adoption rate is unprecedented. Four reasons that may explain this trend are as follows:

1. There is a great deal of competition. Many of the silicon and HDD vendors that missed the FC bandwagon in the mid-1990s are attacking the SAS market with a vengeance to make sure they don't get left behind again in the lucrative enterprise market. The increased competition and combined marketing forces of these suppliers, along with price advantages, advanced feature sets, and greater motivation for interoperability compared to FC, are making SAS more attractive from a developer's perspective.
2. The new breed of high-density 1–2U and blade-based servers has increased demand for small-form-factor drives. The 2.5" drive interface of choice is SAS for this market, which has grown more quickly than many expected and is expected to accelerate the adoption of SAS in general.

3. With SATA support available using the same expander complex as SAS, and with SAS drives promising performance identical to that of FC, many developers are looking at SAS infrastructure as a means of getting two products for the development cost of one. FC–SATA bridging and tunneling solutions are either proprietary or late to the game, have fewer vendors supporting them, and have given SAS-SATA a lengthy head start.

4. SAS is leveraging many of the lessons learned from implementing high-speed serial interfaces. The SAS link and physical layers from FCP to 8b/10b encoding borrow from Fibre Channel. Also, the first SAS implementations are coming in the form of expanders for direct disk attachment, and cascaded expanders allow dozens of disk drives to be directly connected to host bus adapters without the need for external retiming hubs or switches. It wasn't until the later stages of adoption that commercially available loop switches provided options for native disk attachment, forcing early adopters to use external hubs and switches or to restrict themselves to modest configurations using primitive loop bypass circuits. Early switch interoperability issues combined with limited vendor selection also slowed adoption.

Fibre Channel still has its advantages. One is maturity: Fibre Channel is in its tenth year of multivendor implementation, whereas SAS is in its second, and there are bound to be early implementation glitches in any new technology. In addition, what started out as a relatively straightforward drive interface definition is sliding down the slippery slope of complexity that has somewhat plagued Fibre Channel as a disk interface. Zoning, security, and other "network" features threaten to delay standards and add complexity, and the SAS community must avoid the temptation to be all things to all developers. Fibre Channel is a better system network interface, provides distances up to multiple kilometers using fiber-optic options, and finally has multiple vendors providing interoperable switch solutions at both the high end and the low end. Attempts to compete with FC in this arena may slow the interoperability of storage subsystem components, cause a ripple effect back to the drive interface itself by adding complexity, and inadvertently slow the overall adoption of SAS if architecture, design, or interoperability problems result.

The bottom line is that 4 Gb and 8 Gb Fibre Channel will continue be the dominant storage system interconnect in the enterprise for the foreseeable future, but we'll see SAS begin to take significant Fibre Channel market share in 2007 as a disk interface, and before the end of the decade more SAS drives will be shipped than FC and SCSI put together.

SAS VERSUS SATA DISKS IN THE ENTERPRISE

Historically, the overriding priority for SATA drive design has been cost/GB, and this tradeoff shows up in the following areas when comparing SATA to SAS (or FC) drives in the enterprise [4]:

Attribute	*SAS/FC Feature Differentiators*
Mechanical	Larger magnets, stiffer covers, air control devices, faster seeks, low rotational vibration susceptibility
Head stack	More heads, low mass/high rigidity, higher-cost designs
Motor	Higher rpm, less runout, more expensive
Electronics	Dual processors, multi-host, dual-port, twice the firmware, high rpm control and rotational position sensing, superior error correction, smart servo algorithms, more sophisticated performance optimization and command scheduling, deeper queues, larger caches, and more sophisticated data integrity checks
Disks	More platters, smaller diameter, full media certification, and fully characterized
Format	Variable sector sizes (e.g., SATA is moving to large, fixed 4096-byte sectors)

Workloads that are optimal for nearline storage are sequential reads, compliance data, archived email, and other record archives with low duty cycles and low IOP requirements. Workloads optimal for performance storage are random reads and writes, high IOP rates, and high-duty-cycle traffic. Real-time OLTP workloads are an example.

The new features in SATA described previously will put pressure on the normally simple differentiation between the classic desktop and the classic enterprise drive. The cost advantage of SATA, particularly for nearline workloads, is compelling enough for drive integrators to be willing to spend a little more on data protection and enclosure features to accommodate these drives. While end users will want the best of all worlds, drive vendors will continue to prefer to withhold performance and reliability features from SATA drives to maintain their margins in their performance markets as well as to use the same drives to fight for market share on the desktop. This is why performance SAS and nearline SATA drives will continue to coexist in the enterprise for the foreseeable future.

However, systems are now being introduced that can accept both SATA and SAS drives coexisting in the same enclosure. This means that, for the first time, the choice of SATA versus SAS can become a post-purchase decision for customers. It is only fitting that, after 30 years of evolution, storage technology has finally allowed the consumer to more directly dictate the ultimate winner.

REFERENCES

[1] IDC Worldwide Disk Storage Systems Market Forecast and Analysis, 2002-9: http://www.itresearch.com/getdoc.jsp?containerId=33477.

[2] Enterprise Storage Forum, June 16, 2005, “Why Tape Won’t Die”: http://www.enterprisestorageforum.com/continuity/features/article.php/3513406.

[3] WinHEC 2005, “SATA in the Enterprise,” and Seagate Market Research: http://download.microsoft.com/download/9/8/f/98f3fe47-dfc3-4e74-92a3-088782200fe7/TWST05005_WinHEC05.ppt.

[4] Enterprise Storage Forum, Dec. 15, 2005, “Storage Headed for Trouble”: http://www.enterprisestorageforum.com/technology/features/article.php/3564426.