Drive technology overview

technology brief, 2nd edition

Abstract

Disk drive capacity has increased at rates similar to those of microprocessor performance. New and improved interconnect technologies allow the rapid transfer of large amounts of data to and from the disk. New and more cost-effective applications are made possible by increased storage capacity and reduced cost per bit. Through innovation and engineering expertise, HP develops industry-leading disk technologies that optimize overall system capacity, performance, reliability, and value.

This paper reviews the classes of disk drives, the key factors determining capacity, performance and reliability in single drives, the options available to connect the drives to the system, and the use of multiple drives to further increase performance and reliability.

Introduction

Disk drives provide the primary mechanism for storing and retrieving permanent, or non-volatile, data in almost all servers, desktop computers, and notebook computers. Disk drives are also increasingly common in portable electronic devices such as music players and automobile navigation systems.

The key performance differences between main memory (semiconductor RAM) and primary storage (typically, magnetic disk drives) are speed of access and capacity. Accessing primary storage is typically approximately 1000 times slower than accessing main memory. Primary storage is typically at least 100 times larger than main memory. Most engineering in disk drive and interconnect technologies is driven by a desire to reduce this difference in access speed while simultaneously increasing disk drive capacity and reliability. Innovative strategies in disk and disk controller design continue to deliver dramatic increases in disk capacity, performance, and availability.

Flash memory technology, which has previously been used as a low-performance, lower capacity storage medium in consumer devices, is being adapted for use as primary storage in computers. This technology has the promise of delivering enterprise class storage with low latencies and performance approaching that of RAM. The cost per bit for flash memory is between that of RAM and traditional disk drives. Flash-based solid state drives that can meet both the performance and the heavy duty cycle requirements of server storage are being introduced.

Categories of server disk drives

HP has refined and expanded its drive family to offer three distinct classes of server disk drives—Entry, Midline, and Enterprise drives. Each drive category has a different set of performance, reliability, and cost/capacity characteristics designed to meet the needs of different usage environments. Meeting the requirements of each environment heavily influences both the design and the component selections for the drives.

HP Entry drives have the lowest unit cost and provide a basic level of reliability and performance. They are best suited for use in non-mission-critical environments where I/O workloads are 40 percent or less. Typical intended applications for Entry drives are internal/archival storage or as boot disks for entry-level servers.

HP Enterprise drives provide maximum reliability, highest performance, scalability, and error management under the most demanding conditions. They are the only class of drives designed for use at unconstrained I/O workloads and are intended for use in mission-critical applications such as large databases, e-mail servers, and CRM.

HP Midline drives bridge the gap between Entry and Enterprise class by providing larger capacity and greater reliability than Entry drives. HP Midline drives have improved resistance to rotational and operational vibration, so they are better suited than Entry drives for use in multi-drive configurations. For maximum flexibility, Midline drives are available with both Serial ATA (SATA) and Serial Attached SCSI (SAS) interfaces.

HP Midline drives are designed for use in high-capacity applications such as external storage that may require increased reliability. Like Entry drives, however, Midline drives are designed for use in moderate workload environments and should not be considered for mission-critical applications.

Table 1 provides a side-by-side comparison of the three classes of HP server drives.

Table 1. Categories of HP server disk drives

	Entry drives (ETY)	**Midline drives (MDL)**		**Enterprise drives (ENT)**
General description	Performance and reliability intended for entry-level servers, lowest unit cost	High capacity, lowest cost per gigabyte		Maximum reliability and performance using state of the art design
Usage environments	Low I/O, non-mission critical usages - Boot disk - Entry server storage	- External storage - Backups/archival - Redundancy		- Mission critical - High I/O - Large database - e-mail/messaging
Workload	Designed for workloads < 40%	Designed for workloads < 40%,		Designed for unconstrained workloads
Reliability		2 times Entry drive reliability		3.5 times Entry drive reliability
Interface	SATA 1.5 and 3 Gb/s	SATA 3 Gb/s	SAS 3 Gb/s	SAS 3 Gb
Connectivity	Single port	Single port	Dual port	Single and dual port
RPM	5400 and 7200	7200	7200	10,000 and 15,000
Warranty	1 year	1 year	1 year	3 year

Characteristics of disk drives

This section identifies basic characteristics of industry standard disk drives and factors that affect them.

Form factor

HP disk drives for servers are available in both 2.5-inch and 3.5-inch form factors. In general, 2.5-inch drives are used when power savings and space savings are considered important. The smaller 2.5-inch drives can require as little as half the power and generate significantly less heat than 3.5-inch drives. On the other hand, 3.5-inch drives are better suited for uses that require large single drive capacities and lower cost per gigabyte.

HP provides two lines of Universal Carrier for disk drives, one for 2.5-inch form factor drives and one for 3.5-inch form factor drives. These carriers allow any hot-pluggable drive from a family to fit mechanically and electrically with HP ProLiant servers or storage products. This mechanical commonality extends to most SAS-based StorageWorks and HP Integrity server products as well. Thus, mixed HP Enterprise solutions can be supported by a common family of hard drives.

Capacity

The capacity of a drive, measured in gigabytes, is set at manufacturing, and today's drives are capable of storing hundreds of gigabytes. The drive's capacity is determined by the number of platters it contains, the surface area of each platter, and the number of bits that can be stored per unit area (called areal density). Areal density is determined by the number of tracks-per-inch of disk radius multiplied by the number of bits--per-inch of track.

A common source of confusion regarding disk drive capacity is the definition of a gigabyte. In a disk drive, a gigabyte is exactly 1,000,000,000 bytes, but operating systems often use the binary-based approximation of 2^{30}, or 1,073,741,824 bytes, per gigabyte. Thus, the operating system may report that a disk drive with 100 actual gigabytes of storage has only 93 gigabytes.

Performance

Several factors determine the performance of a disk drive. They include the rotational speed of the platters, seek performance, mechanical latency, read/write bandwidth, queuing strategies, and interface technologies.

When preparing to read data from the disk, the drive head must move to the position above the correct track and then wait for the target segment to pass under the head. This mechanical delay—the time to move the head to the correct track and then wait for the target segment—is called the latency or seek time.

Latency, which is fundamental to disk system performance, is measured in milliseconds (ms). Typical values are 4 to 10 ms. A number of strategies have been developed to directly or indirectly avoid or reduce this mechanical latency (Table 2). For example, doubling the rotation rate of the disk platter can reduce the time spent waiting for the target segment to pass under the head.

Disk drive performance is usually characterized under one of two data transfer scenarios—continuous data transfer rate of the media and random Input/Output operations per second (IOPs).

Continuous data transfer occurs when reading or writing relatively large blocks of data to sequential disk sectors. It sets the upper boundary of performance for the drive. It should be noted, however, that the maximum continuous data rate is valid only for the outermost tracks on the drive, and that this rate can be up to 50 percent lower on the inner tracks.

Random access occurs when reading or writing relatively small blocks of data to sectors that may be scattered across the disk. The speed of the actuator and the spindle determine performance in this scenario and set the lower boundary of performance for the drive.

The performance of disk drives deployed in actual computing environments is heavily dependent on the nature of the application; for example, whether it is dealing with large blocks of sequential data (for example, video files) or small blocks of unrelated data (for example, customer records in an e-commerce database). As a disk drive fills up, large blocks of data may have to be written to non-sequential segments or non-adjacent tracks. This scattering of data across the disk, called fragmentation, can significantly degrade performance.

Table 2. Strategies to improve single disk capacity, performance, and reliability

Mechanical	Magnetic	Disk I/O
Increase platter rotation rate	Increase bit density per unit of track	Write cache buffer data to be written to disk
Increase areal density of data		Queue read operations
Reduce platter size		Queue write operations
Decrease seek times		Reorder read and write operations to execute the next operation physically available on drive

Mechanical design strategies are used to reduce the physical distance that the read/write heads must travel to reach the target segment:

- smaller diameter platters
- multiple platters per drive
- increased speed of platter rotation
- increased areal density of data
- decreased seek time per track

Magnetic storage strategies are used to increase the amount of data in each track. Increasing the bit density per unit length of track reduces the need to move among tracks and allows immediate read-after-write verification.

Disk I/O strategies are used to reduce the time that a logical read/write spends waiting for the physical read/write operation. These strategies seek to effectively decouple the logical and physical operations of the disk. Increasingly sophisticated approaches become practical as the embedded processing power and memory incorporated into the drive increase. These approaches include:

- buffering the data to be written to disk (write caching)
- queuing read operations
- read-ahead caching
- queuing write operations
- write caching

Reliability

Disk drive reliability is measured in terms of Annual Failure Rates (AFR). The AFR is the percentage of disk drive failures occurring in a large population of drives in operation for one year. For example, a population of 100,000 drives with an AFR of 1.5 percent would experience approximately 1,500 failures per year. An AFR calculated from a small number of drives is subject to large statistical variations that render it inaccurate.

Major factors in determining drive reliability are the duty cycle and the I/O workload to which the drives are subject. Duty cycle is simply *power-on time*, which is calculated as “the number of hours that the disk drive is powered on” divided by “the number of calendar hours.” I/O workload is *disk working time*, which is calculated as “the number of hours that the disk drive is aggressively reading and writing data” divided by “the number of calendar hours.”

Enterprise drives are designed for unlimited I/O workloads, that is, for continuous I/O activity. Midline and Entry drives are designed for constrained I/O workloads of 40 percent or less. If any doubt exists about the expected workload, and if reliability is a priority, then Enterprise drives should be used.

Drives are subject to mechanical problems created by shock, vibration, environmental extremes, and thermal effects. These problems may degrade performance or reliability (for example by displacing the read head from the data track). They may cause data loss, or even cause catastrophic failure of the drive. Of the three categories of HP drives, Enterprise drives are the most resistant to vibration effects. Midline drives have a lower tolerance to vibration than Enterprise drives. Both HP Enterprise and Midline drives have internal sensors that detect operational/rotational vibration and reduce the performance impact from system, drive-to-drive, and environmental vibrations. Entry class drives will exhibit degraded performance in high vibration environments.

Multiple drives in a single enclosure may interact to create coupled vibration problems. This problem can occur when using Entry or Midline drives with Enterprise class I/O workloads. In fact, Entry drives are not supported in some external storage solutions due to the higher levels of system and drive-to-drive vibrations.

Temperature is a major factor influencing reliability and is usually best managed by controlling the operating environment. HP Enterprise and Midline drives include workload management instrumentation designed to protect the drives if operating temperatures exceed pre-determined limits. Humidity, corrosive environments, and static electricity may also degrade disk performance.

Drive qualification process

HP employs best-in-class qualification and quality control processes to ensure that the disk drives it ships remain reliable, meet customer requirements, and integrate seamlessly into HP server and storage systems. The HP processes also ensure continuous improvement in both current and future products and processes. The qualification process consists of four specific steps:

1. Selection evaluation
2. Development verification
3. Supplier production qualification
4. Continuos improvement/performance monitoring

HP development engineers work closely with disk drive suppliers to execute a comprehensive set of approximately 50 different procedures and specifications that determine the testing and metrics that a candidate drive design must satisfy. Approximately 1000 unique hard disk drives are typically used to evaluate a product family during the selection evaluation and development verification steps, and approximately 2 million drive test hours transpire.

The supplier production qualification phase includes a thorough analysis of the supplier's capabilities, focused on validating supplier process capability and process controls, and on measured product quality. The analysis includes extensive review of the supplier's process controls, closed-loop corrective action processes, and overall quality control system. The final stage of the supplier production qualification includes a comprehensive analysis of the product's quality performance via the HP configuration pilot.

Disk drive products that pass the extensive HP qualification process proceed into HP's continuous improvement/disk drive performance monitoring phase during volume production. This phase includes three main areas of focus:

- Validate that volume production is in process control
- Measure, analyze, and react to product quality data
- Deliver continuous product improvements

HP and the disk drive suppliers work as a team during the volume production phase of a product. The team monitors the performance of each product through quality control methods at the supplier's factory and at HP option kitting configuration sites. Product quality data is reviewed on a daily, weekly, and monthly basis.

Interconnect technology

Various interconnect technologies are used to connect one or more disk drives to a computer system. During the past five years technology has transitioned from parallel bus data interfaces (ATA, IDE, and the original SCSI interface) to SATA and SAS serial interfaces in which each drive has its own high-speed serial communication channel to the disk controller. Table 3 lists basic characteristics of SATA and SAS interfaces. The parallel interfaces are considered end of life in all current HP servers and are not discussed in this paper.

Certain capabilities have traditionally been inherent in SAS or SATA, but this is changing. The benefits and constraints of these two interfaces may become blurred over the next year or two.

Interconnect bandwidths are now exceeding the bandwidth available from the physical drives. With SAS moving to 6 Gb/s and SATA to 3 Gb/s, disk drives will not be able to fully use these bandwidths until well beyond 2012.

Table 3. Comparison of SAS and SATA interfaces for industry-standard servers

	SATA 1.5e	**SAS**
Architecture	point-to-point serial bus	point-to-point full duplex serial bus
Maximum throughput	150 MB/s (1.5 Gb/s)	300 MB/s (3.0 Gb/s)
Cable length	1 m	8 m
Number of devices supported	15	16,256
Command set	ATA	SCSI +
Hot swap support	yes	yes
Drive ID		worldwide unique ID set at factory
Drives supported	SATA only	SAS or SATA, can be mixed
Typical Use	Entry and Midline low IO workload	Enterprise mission critical high IO workload best performance
Seek time, typical drive	9 – 12 ms	3.5 – 4.0 ms

Serial Attached SCSI

SAS is generally considered the most cost-effective solution for mission critical, high I/O workload applications, such as business critical databases. SAS disk drives are dual ported and provide two active-active paths to each device. SAS inherits the proven SCSI command set but uses a point-to-point serial interface, with each device connecting directly to a SAS point. The serial interface makes the complete bandwidth of the link available to each device, which greatly increases performance and scalability. In addition, the links are full duplex and can be grouped to further increase bandwidth. First generation SAS supports a link speed of 3 Gb/s. The second generation supports a link speed of up to 6 Gb/s.

SAS is built upon the SATA physical characteristics. This means that SATA drives can be used with SAS controllers. In fact, SATA and SAS drives can be mixed in a single enclosure. However, SAS devices cannot be used with SATA controllers.

HP was instrumental in developing the SAS standard. For a more detailed discussion of SAS, see the HP technology brief entitled “Serial Attached SCSI technology” available at http://h20000.www2.hp.com/bc/docs/support/SupportManual/c00302340/c00302340.pdf.

Serial ATA

SATA is usually considered the best solution for price-sensitive, low I/O workload server applications, such as entry level and external storage. SATA is also expected to dominate the desktop market due to its low cost.

SATA introduces a serial communication interface that operates in simplex mode, increases data transfer rate, requires a small-diameter cable, supports additional disks, and supports hot swapping. SATA uses a seven-pin, small-diameter cable. The controller may be built into the mother board or provided as an adaptor card.

The SATA specification was initially released in three variants:

- 1.5 Gb/s
- 1.5 Gb/s with extensions
- 3.0 Gb/s

The initial SATA 1.5 Gb/s variant was targeted at replacing ATA in the desktop and consumer markets. It introduced a serial interface that supports one drive per controller port.

SATA 1.5 Gb/s with extensions is targeted to workstations and low-end servers. It adds native command queuing.

SATA 3.0 Gb/s is targeted to workstations and low-end servers. It increases the data transfer rate.

The SATA roadmap calls for SATA to reach speeds of 6 Gb/s.

For a more detailed discussion of SATA, see the technology brief entitled “Serial ATA technology” at http://h20000.www2.hp.com/bc/docs/support/SupportManual/c00301688/c00301688.pdf

Excess bandwidth

Certain capabilities have traditionally been inherent in SAS or SATA, but this is changing. The benefits and constraints of these two interfaces may become blurred over the next year or two.

Interconnect bandwidths are now exceeding the bandwidth available from the physical drives. With SAS moving to 6 Gb/s and SATA to 3 Gb/s, disk drives will not be able to fully use these bandwidths until well beyond 2012.

Improved performance and reliability with RAID

Drives do fail; therefore, storing data on a single disk drive creates a risk of data loss. HP recommends always using some form of fault-tolerant RAID (Redundant Array of Independent Disks) across multiple drives.

RAID strategies can be characterized by how they achieve data reliability (how parity or other error correction data is distributed across the array), the minimum number of drives required, and data storage efficiency. The performance of multiple drives is better than the performance of a single drive.

The choice of RAID strategy and how it is implemented affect read performance, write performance, and robustness in the face of hardware failures:

- Raid 0 – striping to two or more disks; no redundancy, performance improvement only
- Raid 1 – mirroring; duplicates same data on two disks; redundancy and potential performance improvements
- Raid 1 + 0 – mirroring and striping; redundancy and performance improvement
- Raid 5 – block striping with distributed parity; three or more drives; fault tolerance
- Raid 6 – block level striping with dual distributed parity; three or more drives, increased fault tolerance

Both read performance and write performance also vary with the workload; that is, whether I/O (many small data units) or bandwidth (fewer, large data units) predominates.

Advanced controllers

Advanced controllers, such as the HP Smart Array, decouple the logical disks seen by applications from the physical devices used to implement the disk subsystem. These controllers include both hardware and software. A single logical disk (as seen by an application) may be mapped onto an array of multiple physical disks. This approach provides greatly enhanced flexibility, expandability, maintainability, and performance. Smart Array controllers are available for SAS, SATA, and SCSI interfaces.

For further details, see the technology brief entitled "HP Smart Array Controller Technology," at http://h20000.www2.hp.com/bc/docs/support/SupportManual/c00687518/c00687518.pdf.

Emergence of solid state drives for servers

Solid state drives represent a new class of storage technology that is entering the market. Unlike traditional disk drives, SSDs use flash memory—the same basic technology used in devices such as USB drives—to store and retrieve data. Until recently, storage devices based on flash memory were too slow to be considered for use in server environments. Advances in flash memory, including the use of faster and more reliable Single-level Cell (SLC) technology and advanced controller technologies, have led to the creation of solid state drives that can meet the performance and reliability requirements for use in server environments.

Although specifics may change slightly as actual solid state drive products are introduced, Table 4 provides a good starting point for comparing today's SATA and SAS server disk drives with the expected first generation server SSDs.

Table 4. Comparison of SATA/SAS drives with first generation SSDs for servers

	Typical Small Form Factor SATA/SAS drive	**1st generation server SSD (preliminary)**
Maximum capacity	Up to 250 Gb	32 and 64 Gb
Interconnects and form factors	Defined industry standards	SATA interface Low profile form factor
Write performance sustained throughput	85 Mb/s	50 Mb/s
Write performance random IOP/s	160	100
Read performance sustained throughput	90 Mb/s	100 Mb/s
Read performance random IOP/s	180	4300
Typical power	4 – 9 watts	2 watts
Operating environment	60 degrees C	Up to 70 degrees C
Reliability	Entry – suitable for non-mission-critical environments Midline – 2 times Entry class Enterprise – 3.5 times Entry class	Comparable to Midline drives

Performance of solid state drives

As Table 4 indicates, server SSDs will not be particularly fast in terms of write performance. The underlying slow write process for flash memory cannot be completely compensated for by write cache and other technologies used in SSD controllers. Similarly, sustained read performance is only comparable to that of SATA and SAS drives. The one area where SSDs truly excel is in random reads, where performance is over 20 times that of traditional rotating media. This makes SSDs good candidates for use in applications that make extensive use of random read operations.

Reliability and operational environment of solid state drives

As shown in Table 4, first generation solid state drives are expected to have an overall reliability, in terms of annualized failure rates, that is roughly equivalent to that of midline server disk drives. However, SSDs can maintain this level of reliability in environments that are unsuitable for traditional disk drives:

- Higher temperature environments
- Limited airflow environments
- Environments subject to higher shock and vibration
- Environments requiring drives with lower power consumption

Conclusion

Disk technology is evolving rapidly. Hardware should be reviewed periodically to ensure that the most cost-effective solutions are deployed. Enterprise class drives with SAS interconnects and HP SAS Smart Array Controllers provide the highest performance and most reliable storage for industry-standard servers. Other technologies may be appropriate for non-mission-critical applications.

Solid state drives are a new and rapidly evolving class of storage technology that can operate reliably in more extreme environments while providing random I/O performance that is superior to traditional disk drives.

For more information

For additional information, refer to the resources listed below.

Resource description	Web address
Serial ATA Technology technology brief	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c00301688/c00301688.pdf
Serial Attached SCSI Technology technology brief	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c00302340/c00302340.pdf
HP Serial ATA Drives overview	http://h18004.www1.hp.com/products/servers/proliantstorage/serial/sata/index.html
HP Disk Drives for ProLiant Servers	http://h18004.www1.hp.com/products/servers/proliantstorage/drives-enclosures/index.html

Call to action

Send comments about this paper to TechCom@HP.com.

TC080906TB, September 2008

invent