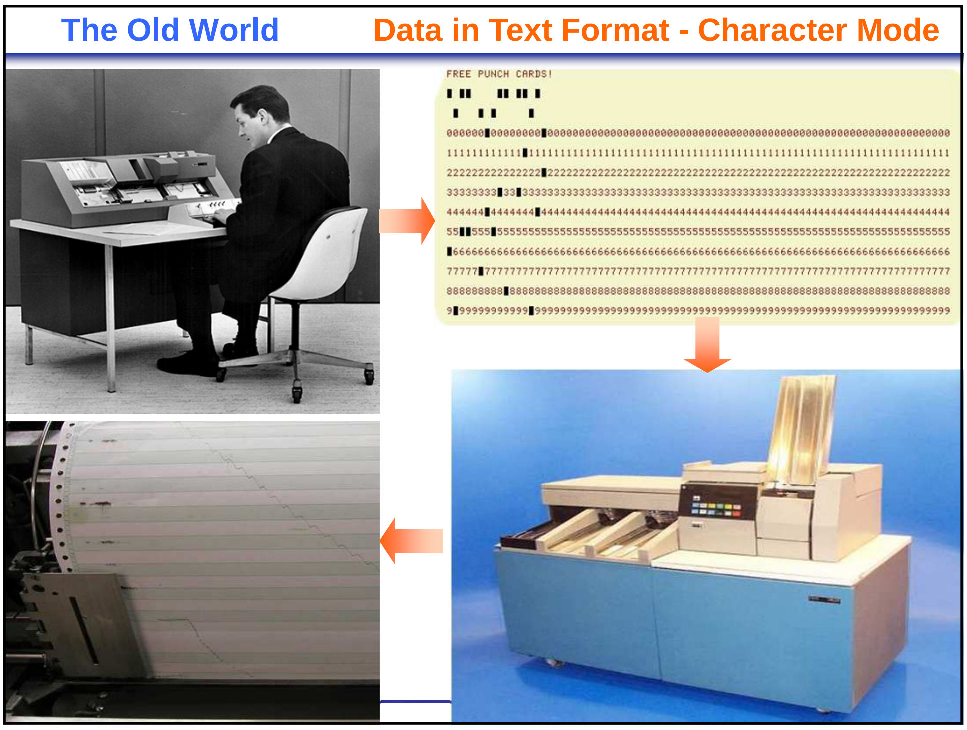
Storage Overview – part 1

As a database administrator, you need to understand how storage works for best administration, performance, and disaster-recovery requirements.

When creating a database, it is important to understand how database stores data so that you can calculate and specify the amount of disk space to allocate for the data files and log files.

- Storage History
- Storage Overview
- RAID, DAS, NAS, SAN

The Old World   Data in Text Format - Character Mode

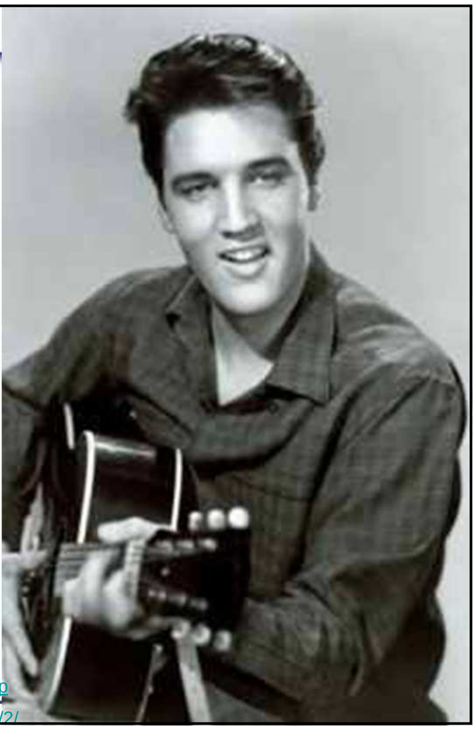


The King of Rock and Roll

http://www.Elvis.com

When did *Elvis* start his singing carrier?

1956



http://www.elvis.com/elvisology/bio/elvis_1935_1957_5.asp

http://cwflyris.computerworld.com/t/851617/224813/33674/2/

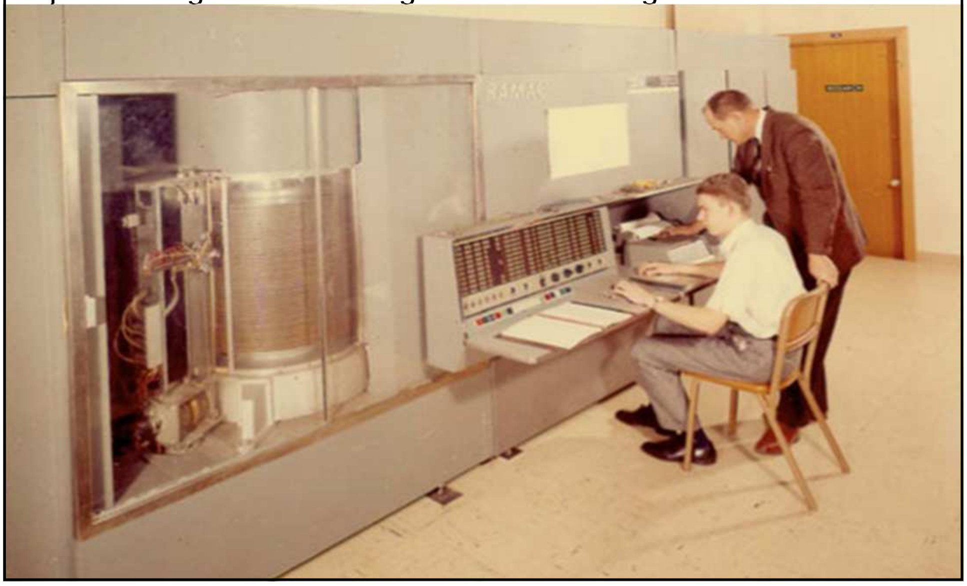
Hard Drive History – 50+ Years in The Making

- When IBM delivered its first hard drive on September 13, 1956, few could have imagined the impact it would have on our everyday lives.
- The RAMAC (also known as '*Random Access Method of Accounting and Control*') was the size of two refrigerators and literally weighed a ton.
- It required a separate air compressor to protect the heads, had pizza-sized platters and was able to store a then whopping 5 MB of data.
- The **IBM 350 Disk Storage Unit** offered unprecedented performance for its time by allowing random access to any of the million characters distributed over both sides of 50 two-foot-diameter disks containing 50,000 sectors -- each of which held 100 alphanumeric characters. The purchase price was about $10,000 per MB, the equivalent of $70,000 in today's dollars.

http://www.youtube.com/user/ibmdisk

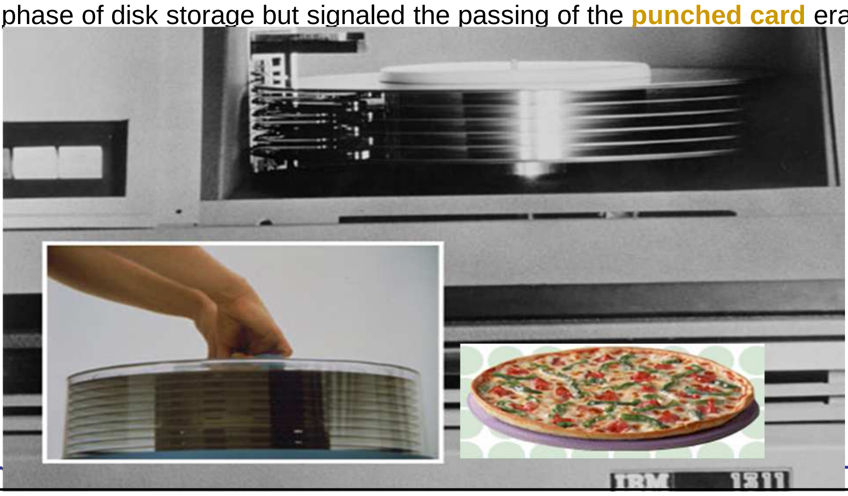
IBM 305A RAMAC SYSTEM

- Santa Clara University - IBM 305 System offered transaction processing based on Magnetic Disk Storage.



IBM 1311 Disk Storage Drive

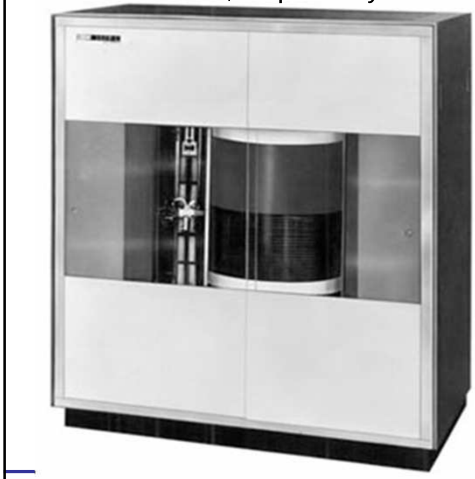
- The IBM 1311 Disk Storage Drive provided storage for two million characters. Debuting in 1962, the 1311 used the IBM Disk Pack -- an interchangeable package containing six 14-inch-diameter disks in a four-inch stack, weighing 10 pounds.
- Some analysts believe that the removable disk pack not only led to a new phase of disk storage but signaled the passing of the punched card era.



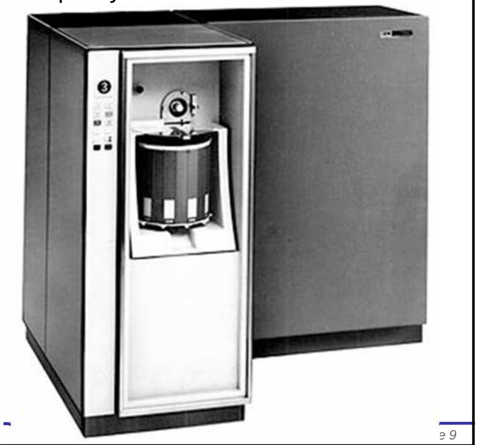
IBM 1311 disk storage drive

IBM Disk Storage

- The IBM 1405 Disk Storage was offered by IBM throughout the 1960s in 25-disk and 50-disk models, for a storage capacity of 10 MB and 20 MB, respectively.



- The IBM 2321 Data Cell Drive stores up to **400 MB** on magnetic strips mounted vertically around a rotating cylinder.
- Announced in 1964, the 2321 could be linked in multiple drives to the IBM System/360 to provide a storage capacity of **3.2 GB** of information.



On July 20, 1969, N. Armstrong successfully landed the Apollo 11 on the moon. The project was supported by IBM System/360.

IBM hard disk drive on PC and Laptop

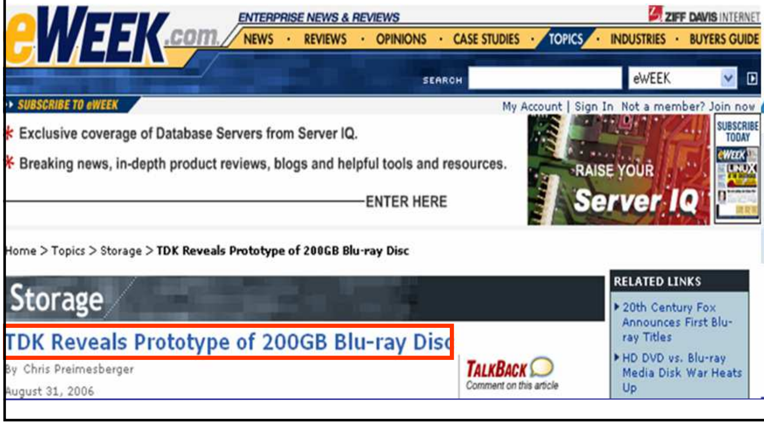
PC (1980s)

ThinkPad 770 (1997)

Blu-ray Disc

http://www.sonypictures.com/homevideo/bluray/

- A high-density optical disc format for the storage of digital media, including high-definition video.
- The name Blu-ray Disc is derived from the blue-violet laser used to read and write this type of disc. Because of its shorter wavelength (405 nm), substantially more data can be stored on a Blu-ray Disc than on the DVD format, which uses a red, 650 nm laser.
- A Blu-ray Disc can store 25 GB on each layer, as opposed to a DVD's 4.7 GB.
- About 9 hours of high-definition (HD) video can be stored on a 50 GB disc.
- About 23 hours of standard-definition (SD) video can be stored on a 50 GB disc.



eWEEK.com
ENTERPRISE NEWS & REVIEWS
ZIFF DAVIS INTERNET
NEWS · REVIEWS · OPINIONS · CASE STUDIES · TOPICS · INDUSTRIES · BUYERS GUIDE
SEARCH
eWEEK
SUBSCRIBE TO eWEEK
My Account | Sign In Not a member? Join now
RAISE YOUR Server IQ
SUBSCRIBE TODAY

Exclusive coverage of Database Servers from Server IQ.

Breaking news, in-depth product reviews, blogs and helpful tools and resources.

ENTER HERE

Home > Topics > Storage > TDK Reveals Prototype of 200GB Blu-ray Disc

Storage

TDK Reveals Prototype of 200GB Blu-ray Disc

By Chris Preimesberger

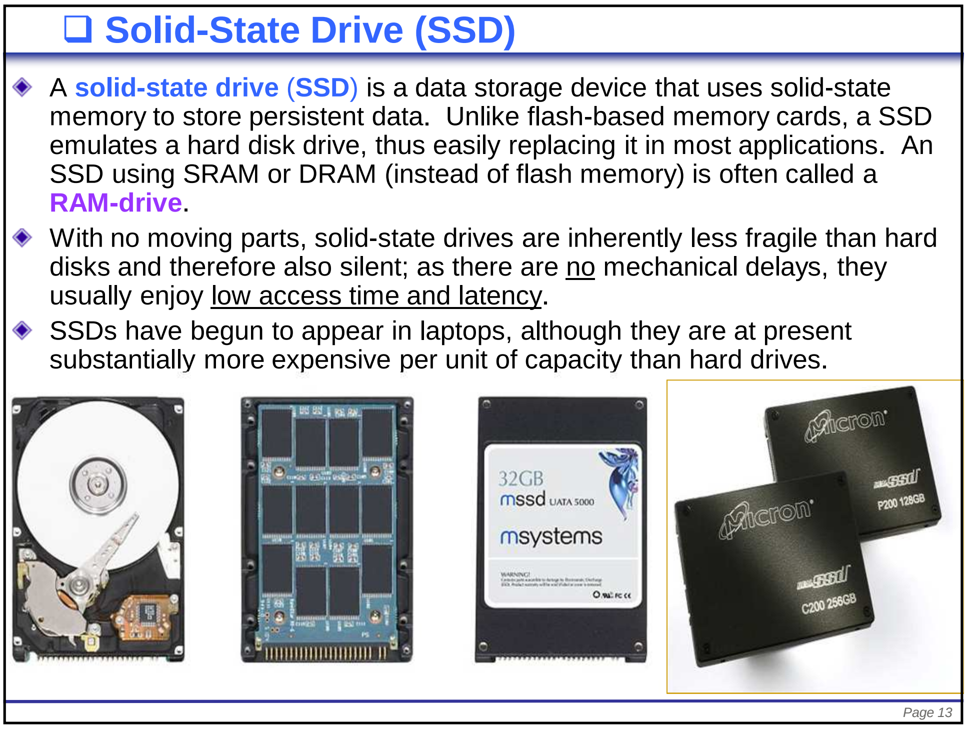
August 31, 2006

TalkBack Comment on this article

RELATED LINKS

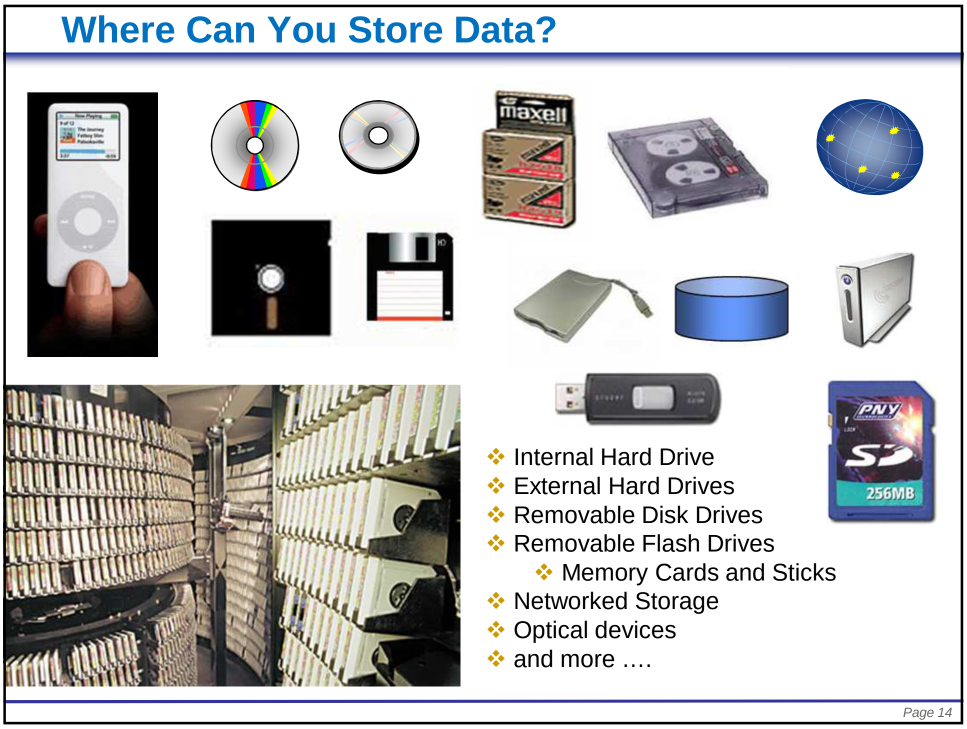
- 20th Century Fox Announces First Blu-ray Titles
- HD DVD vs. Blu-ray Media Disk War Heats Up

Blu-ray Disc TM



❑ Solid-State Drive (SSD)

- A **solid-state drive (SSD)** is a data storage device that uses solid-state memory to store persistent data. Unlike flash-based memory cards, a SSD emulates a hard disk drive, thus easily replacing it in most applications. An SSD using SRAM or DRAM (instead of flash memory) is often called a **RAM-drive**.
- With no moving parts, solid-state drives are inherently less fragile than hard disks and therefore also silent; as there are no mechanical delays, they usually enjoy low access time and latency.
- SSDs have begun to appear in laptops, although they are at present substantially more expensive per unit of capacity than hard drives.



Where Can You Store Data?

- Internal Hard Drive
- External Hard Drives
- Removable Disk Drives
- Removable Flash Drives
  - Memory Cards and Sticks
- Networked Storage
- Optical devices
- and more ….

maxell
PNY
SD
256MB

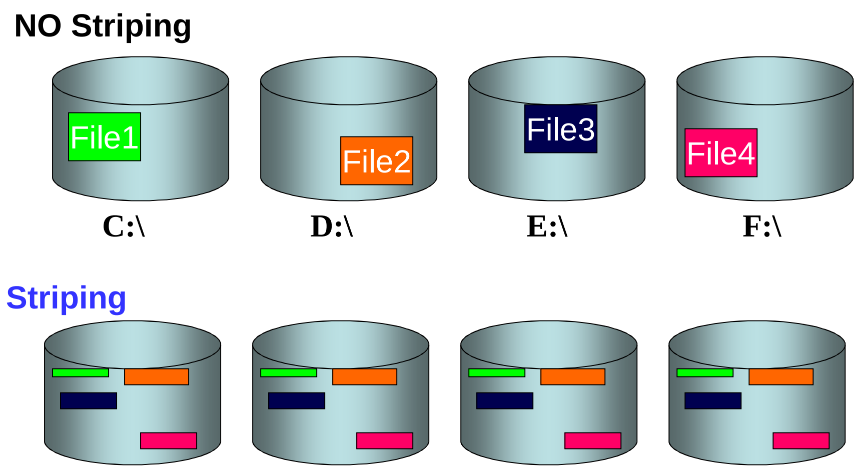
❑ RAID

- **What is RAID?** *(Inexpensive)*
  - Redundant Array of Independent Disks is a common system for high-volume data storage at the server level. **RAID** systems use many small-capacity disk drives, which appear as a single logical unit.
- **Levels of RAID**
  - 16 different RAID levels
  - Only 4 (0, 1, 5, 6) are commonly used in the marketplace
    - **RAID - 0 - stripping**
    - **RAID - 1 - mirroring**
    - **RAID - 0+1 - stripping and mirroring** (Most OLTP systems are RAID 0+1)
    - **RAID - 5 - stripping and parity** (can recover from a single failure without doubling the space)
    - **RAID - 6 - stripping and double parity** (can recover from a single failure with redundant parity bit)
  - Different options to implement fault tolerance
- **Benefits of Using RAID**
  - Performance
  - Capacity
  - Reliability
  - Simplicity
- *Note: **RAID does not protect your system from failures related to disk bus, host adaptor, interface, disk controller, cooling system, power system, host, application and human beings.***

The Evolution Of RAID

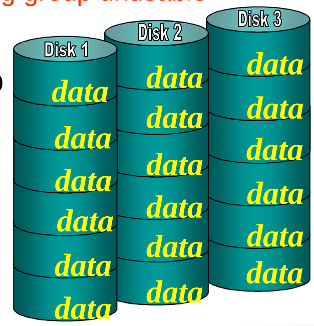
1988	1989	1991	1994	1998	2002	2008
"A Case For Redundant Arrays Of Inexpensive Disks (RAID)" is published	Compaq SystemPro, the first RAID controller for LAN servers, arrives	EMC (NYSE: EMC) introduces Symmetrix RAID for the Mainframe	ANSI Fibre Channel Standard opens the SAN era	Xiotech's Magnitude offering virtualizes RAID	RAID-6 (double parity) goes mainstream	Self-healing arrays from Xiotech and Atrato come to market

Striping - RAID 0



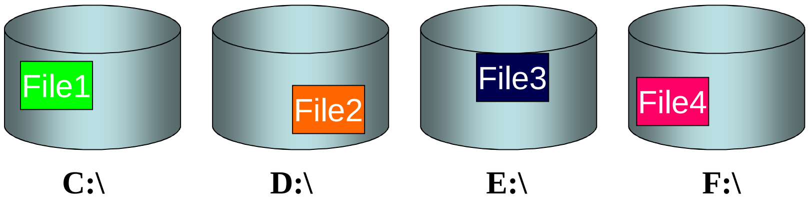
RAID Level 0 – Disk Stripping

- Requires two or more drives
- Data is read simultaneously from each drive
- Writes evenly distributed across drives
- Full capacity of the installed drives can be used
- Excellent performance
- Cannot tolerate the loss of one disk
  - Loss of a single drive makes the entire stripping group unusable
- Not fault tolerance
- Often, RAID 0 is combined with RAID 1 to create a highly available and efficient drive array

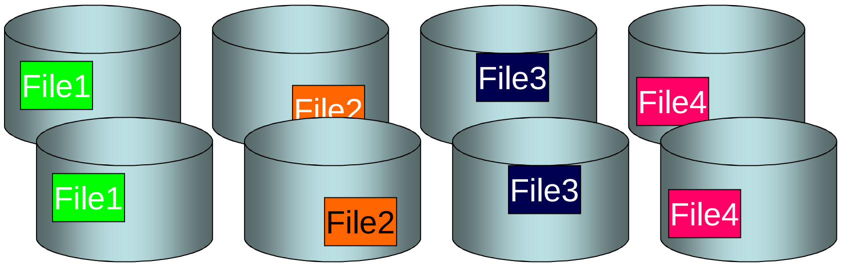


Mirroring - RAID 1

NO Mirroring

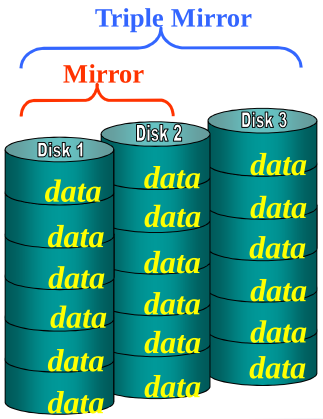


Mirroring



RAID Level 1 – Disk Mirroring

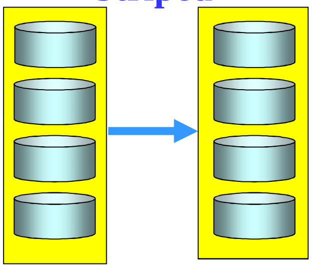
- Additional drives are exact images
- The RAID 1 Controller divides the hard drives in the array into two (mirror) or three (triple mirror) groups
  - Writes identical data onto drives in each group
  - Data is read out from either group
- Excellent Performance
- 100% Redundancy
- Only use 50% (or 1/3) of disk space
- Triple mirroring will increase availability
- Remote mirroring provides site recovery



RAID Level 0+1 – Mirroring of Striped Disks

- Minimum of 4 drives
- Striped array of disks, which are then mirrored to another identical set of striped disks
- Can tolerate loss of 1 drive only per stripe set
- Excellent performance through striping
- Fault tolerance through mirroring
- Next Best Availability/Fault Tolerance (0+1 Not = 1+0, 1+0 > 0+1)

Striped

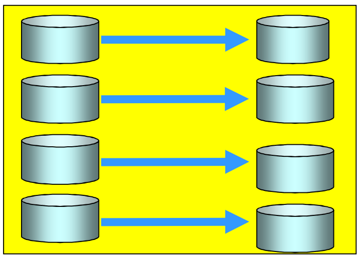


Mirrored

RAID Level 1+0 (RAID 10) – Striping of Mirrored Disks

- Minimum of 4 drives
- Use multiple mirror set, which are then striped into a logical disk
- Can tolerate loss of more than 1 drive as long as no two within a mirrored set are lost
- Performance gains through striping; less than RAID 0 + 1
- Fault tolerance through mirroring
- Best Availability/Fault Tolerance *(if lose 1 drive, it doesn't lose the whole strip)*

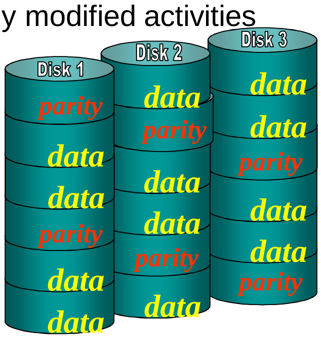
Mirrored



Striped

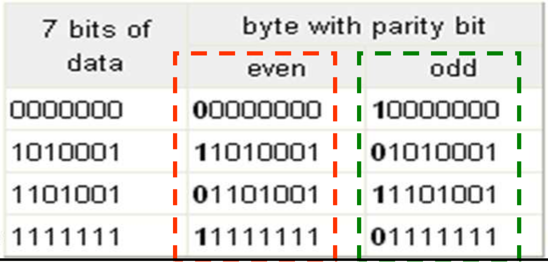
RAID Level 5 – Disk Striping with Parity

- Minimum of 3 drives
- Can tolerate loss of 1 drive only
- Performance penalty on write
- Performance degradation for calculation of parity when a stripe member is missing
- N-1 drives for data; 1 drive for fault tolerance
  - Parity bits are striped across all disks
- Not the best choice of performance on heavily modified activities
- Offers redundancy at the lower cost



Parity bit

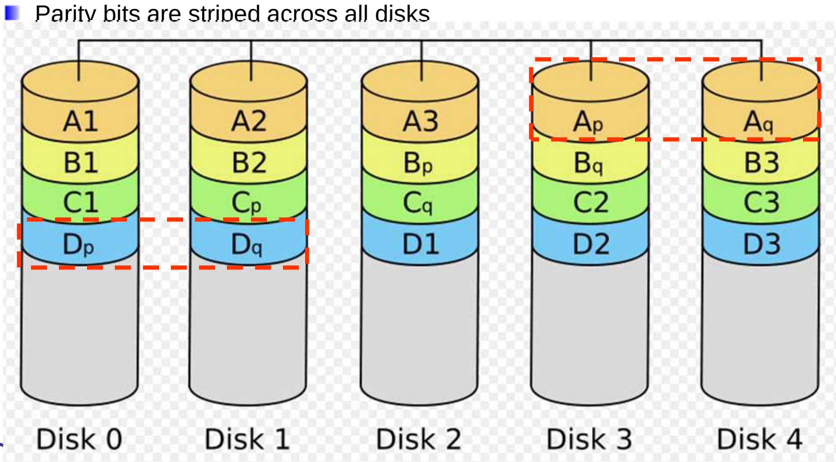
- A **parity bit** is a binary digit that indicates whether the number of bits with value of one in a given set of bits is even or odd.
- Parity bits are used as the simplest error detecting code.
- There are two types of parity bits:
  - **even parity bit -** An even parity bit is set to 1 if the number of ones in a given set of bits is **odd** (making the total number of ones even).
  - **odd parity bit -** An odd parity bit is set to 1 if the number of ones in a given set of bits is **even** (making the total number of ones odd).
- It's in the same way as a checksum to detect accidental alteration of data during transmission or storage.



7 bits of data	byte with parity bit	
	even	odd
0000000	**0**0000000	**1**0000000
1010001	**1**1010001	**0**1010001
1101001	**0**1101001	**1**1101001
1111111	**1**1111111	**0**1111111

RAID Level 6 – Disk Striping with Double Parity

- **RAID 6** extends RAID 5 by adding an additional parity block, thus it uses block-level striping with two parity blocks distributed across all member disks.
- **RAID 6** is designed for tolerating two simultaneous HDD failures by storing two sets of distributed parities.
- N-2 drives for data; 2 drives for fault tolerance
  - Parity bits are striped across all disks



Why is RAID-DP Needed?

Protection

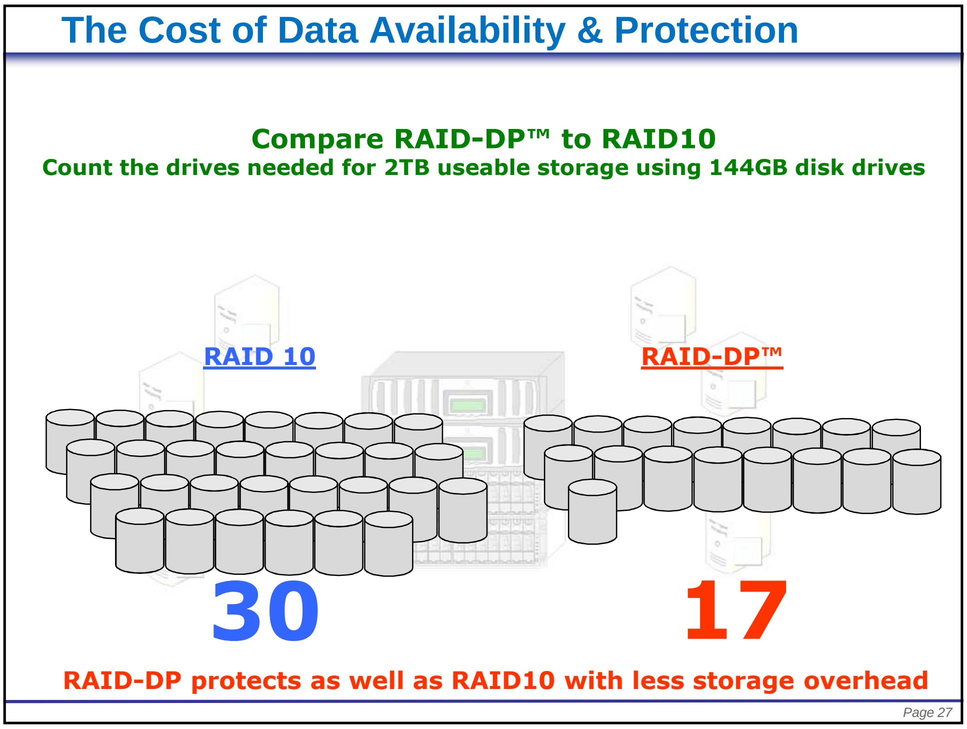
- ‘Traditional’ single-parity-drive RAID group no longer provides enough protection
  - Reasonably-sized RAID groups (e.g. 8 drives) are exposed to data loss during reconstruction
- RAID-DP’s double disk-failure protection does what RAID5 and RAID1/0 cannot:
  - Reduces RISK: limits exposure to same RAID group second disk failure or non-recoverable media error

Cost

- RAID 1 is too costly for widespread use
  - Mirroring doubles the cost of storage
  - Not affordable for all data
- RAID-DP exceeds RAID1/0 protection levels without the associated doubling of capacity and cost

Performance

- Optimized for Performance
- Reduces RAID group rebuild time

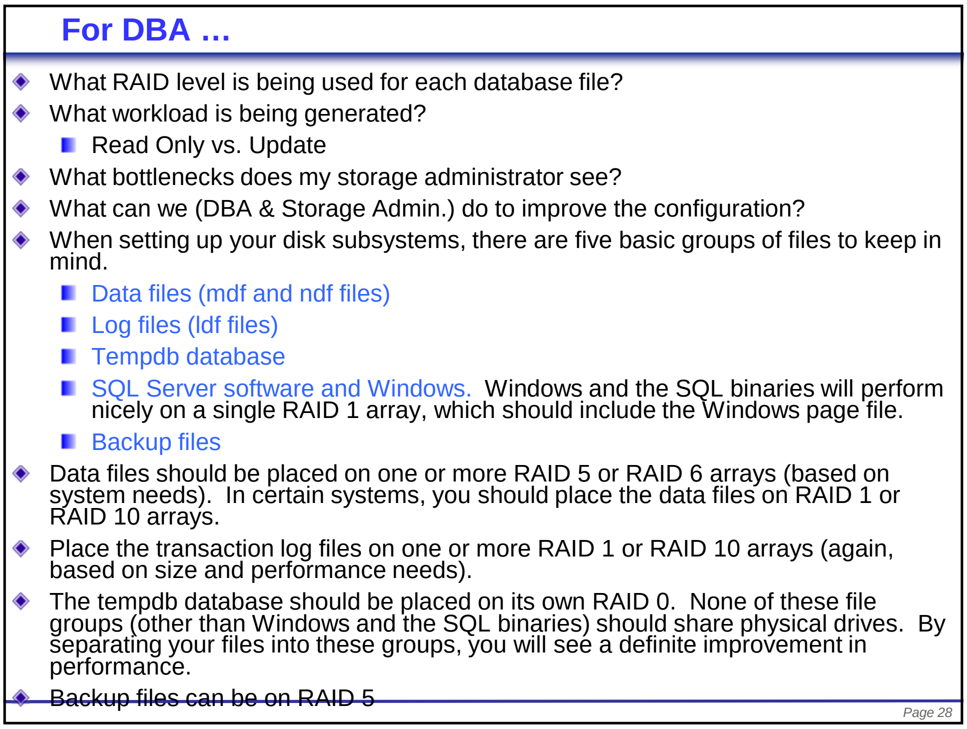


The Cost of Data Availability & Protection

Compare RAID-DP™ to RAID10

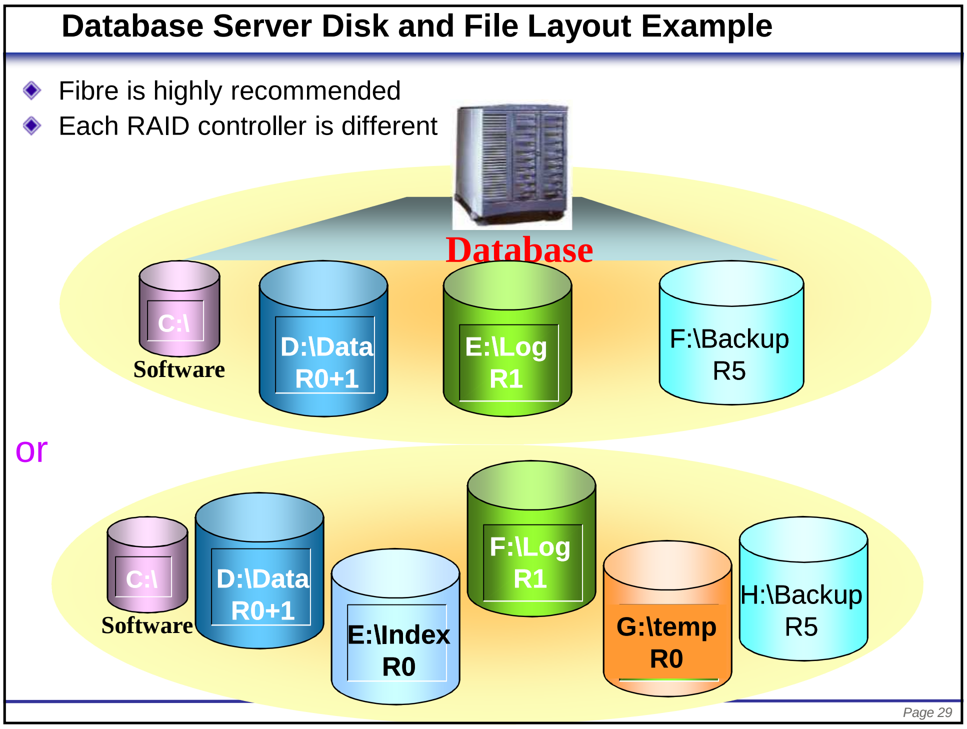
Count the drives needed for 2TB useable storage using 144GB disk drives

RAID-DP protects as well as RAID10 with less storage overhead



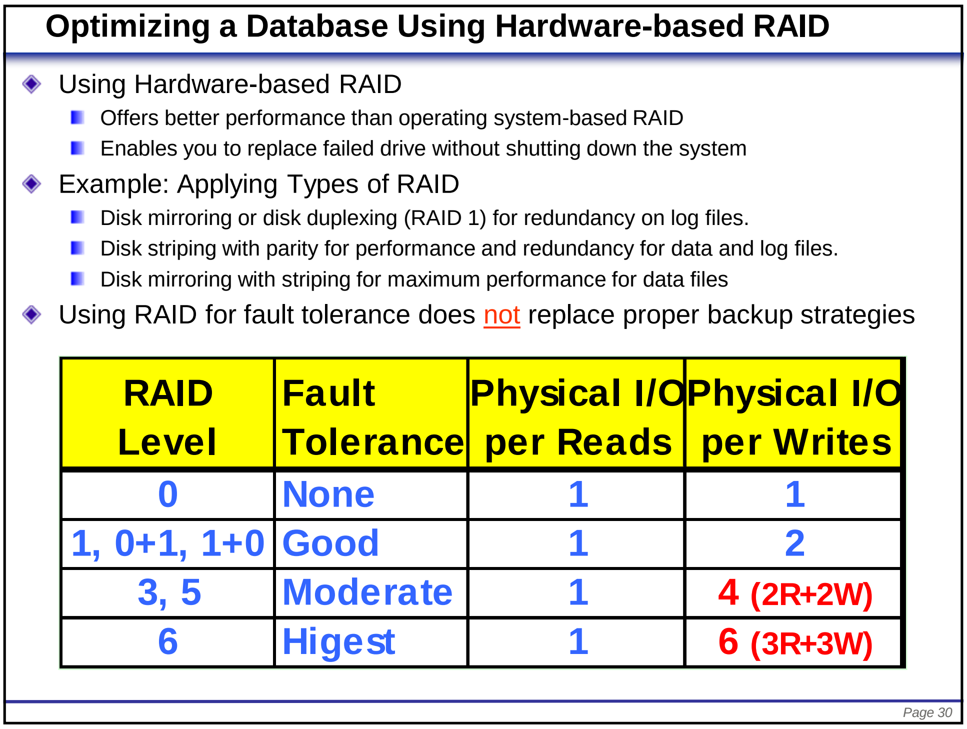
For DBA …

- What RAID level is being used for each database file?
- What workload is being generated?
  - Read Only vs. Update
- What bottlenecks does my storage administrator see?
- What can we (DBA & Storage Admin.) do to improve the configuration?
- When setting up your disk subsystems, there are five basic groups of files to keep in mind.
  - Data files (mdf and ndf files)
  - Log files (ldf files)
  - Tempdb database
  - SQL Server software and Windows. Windows and the SQL binaries will perform nicely on a single RAID 1 array, which should include the Windows page file.
  - Backup files
- Data files should be placed on one or more RAID 5 or RAID 6 arrays (based on system needs). In certain systems, you should place the data files on RAID 1 or RAID 10 arrays.
- Place the transaction log files on one or more RAID 1 or RAID 10 arrays (again, based on size and performance needs).
- The tempdb database should be placed on its own RAID 0. None of these file groups (other than Windows and the SQL binaries) should share physical drives. By separating your files into these groups, you will see a definite improvement in performance.
- Backup files can be on RAID 5



Database Server Disk and File Layout Example

- Fibre is highly recommended
- Each RAID controller is different

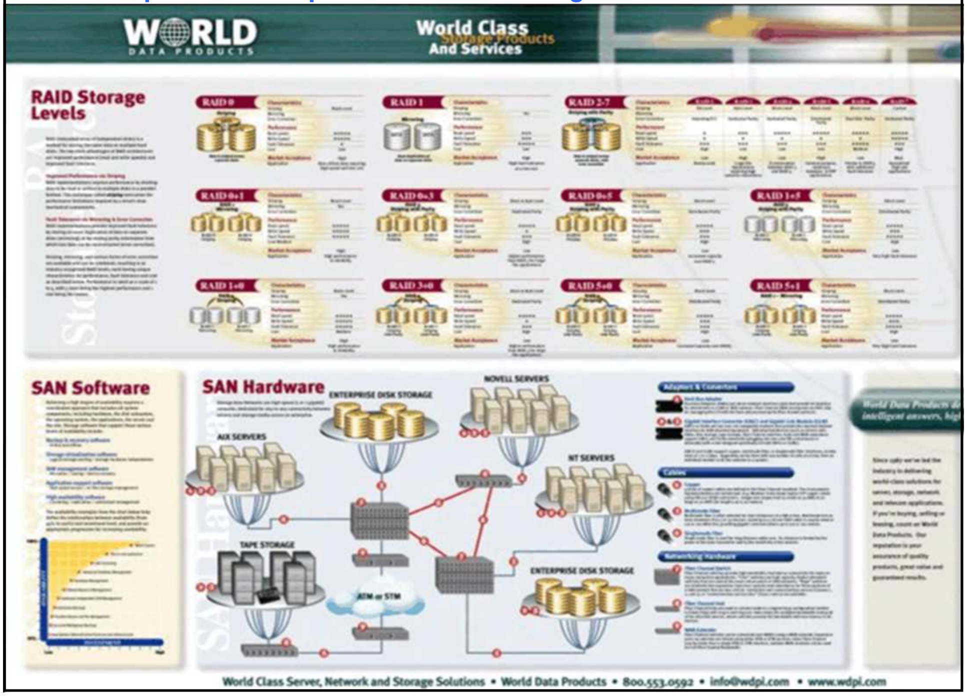


Optimizing a Database Using Hardware-based RAID

- Using Hardware-based RAID
  - Offers better performance than operating system-based RAID
  - Enables you to replace failed drive without shutting down the system
- Example: Applying Types of RAID
  - Disk mirroring or disk duplexing (RAID 1) for redundancy on log files.
  - Disk striping with parity for performance and redundancy for data and log files.
  - Disk mirroring with striping for maximum performance for data files
- Using RAID for fault tolerance does not replace proper backup strategies

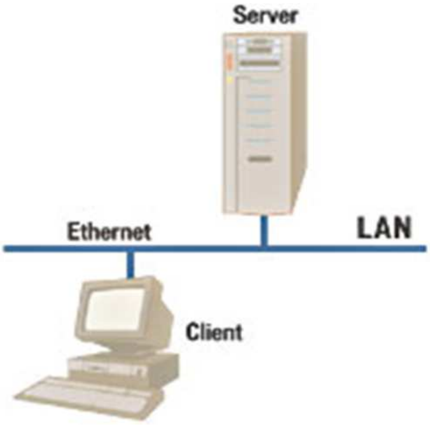
RAID Level	Fault Tolerance	Physical I/O per Reads	Physical I/O per Writes
0	None	1	1
1, 0+1, 1+0	Good	1	2
3, 5	Moderate	1	4 (2R+2W)
6	Higest	1	6 (3R+3W)

http://www.computerworld.com/ads/go2/7056165.html



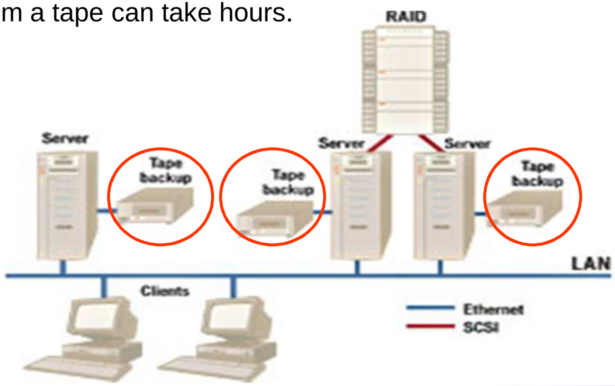
Client/Server Storage at Its Simplest

- In the simplest network storage configuration, a user saves data either to his own PC or over a LAN to a server.
- This provides the user with backup of important files but is inadequate in all but the smallest of environments, such as a small office or home network.



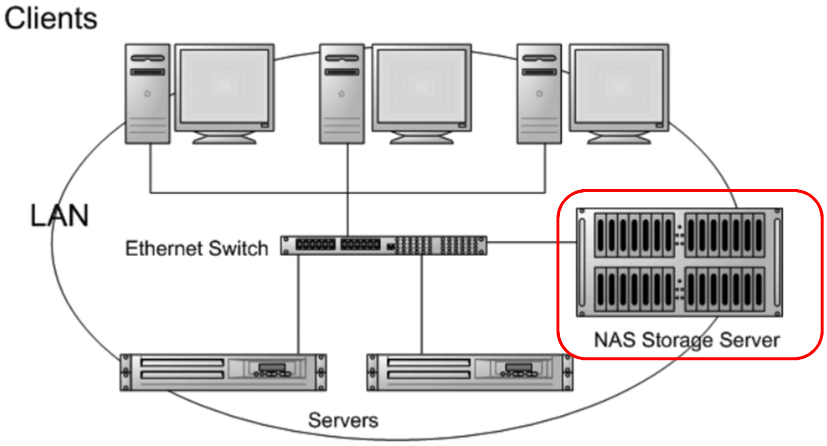
❑ Direct-Attached Storage (DAS)

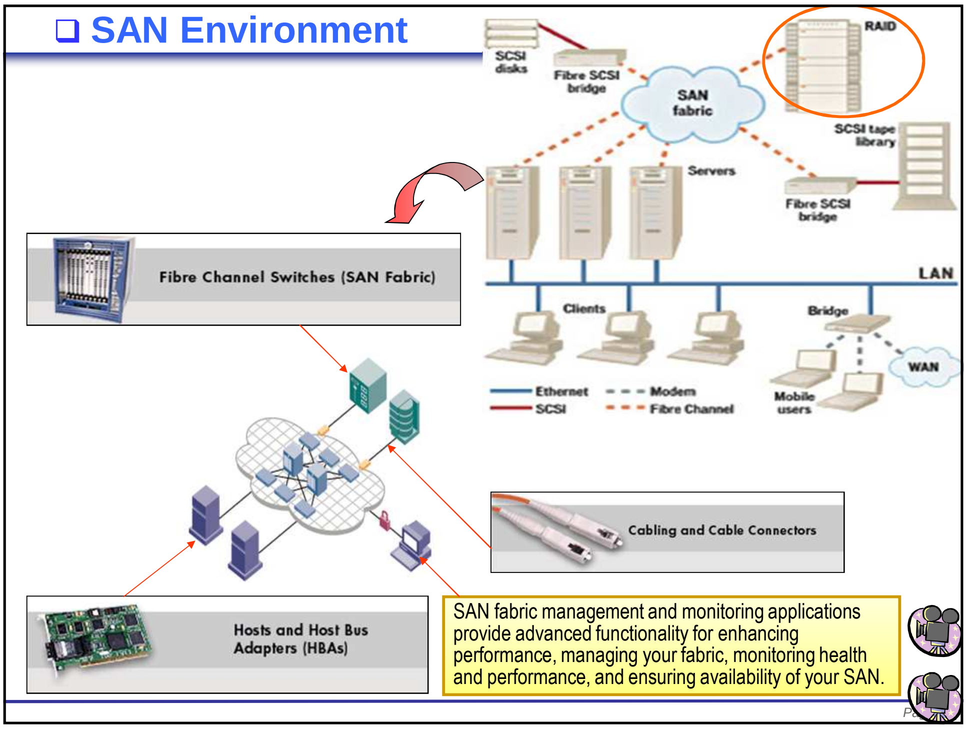
- As a network grows, it typically starts with direct-attached storage, in which a server is regularly backed up to tape or a RAID array. There may be multiple servers, such as file servers, database servers and Web servers, each with its own backup tape drive.
- But tapes must be manually swapped out, and if someone forgets to insert a fresh tape, or if a tape or drive goes bad, the backup doesn't happen.
- Backup schedules typically vary—a production database should be backed up more frequently than an ordinary file server. Tasks such as retrieving single files from a tape can take hours.



❑ Network-Attached Storage (NAS)

- By attaching high-speed RAID storage to the network, servers can get data from the RAID array instead of individual servers.
- The RAID array increases data reliability and redundancy, but it doesn't replace the need for regular backups.



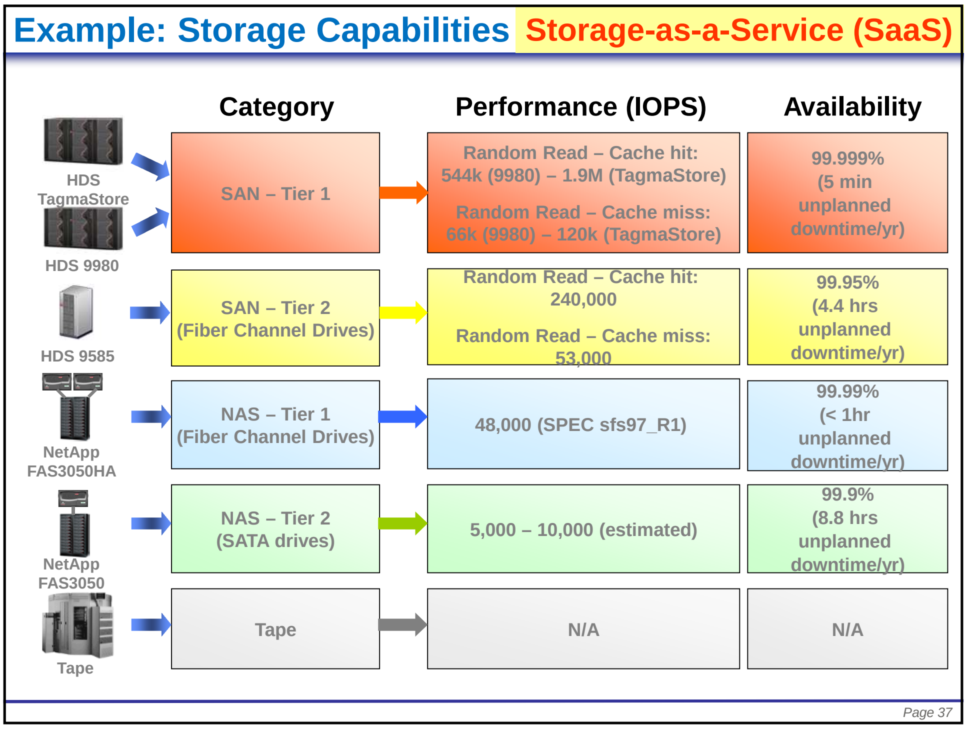


❑ SAN Environment

SAN fabric management and monitoring applications provide advanced functionality for enhancing performance, managing your fabric, monitoring health and performance, and ensuring availability of your SAN.

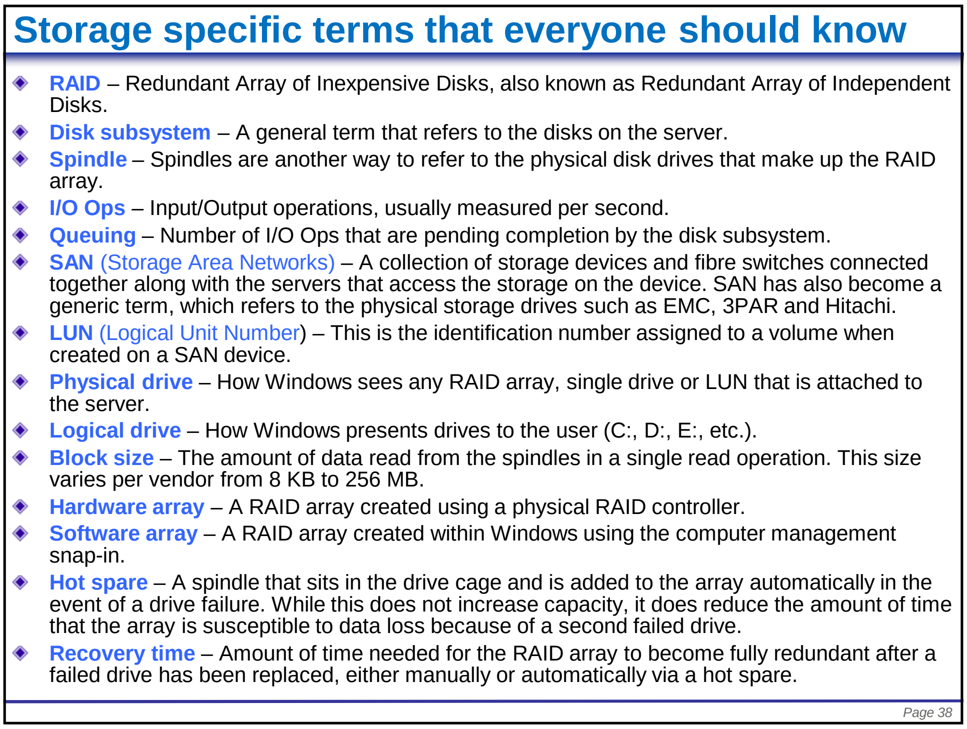
Logical Unit Number (LUN) of SAN

- **A LUN is a logical entity that converts raw physical disk space into logical storage space that a host server's operating system can access and use.**
  - LUNs differentiate between different chunks of disk space.
  - A LUN is part of the address of the storage that you're presenting to a [host] server.
- **LUNs are often referred to as logical "volumes," reflecting the traditional use of "drive volume letters," such as volume C: or volume F: on your computer.**
  - The 'volume' is a piece of a volume group, and the volume group is composed of multiple LUNs.
- **LUNs are created as a fundamental part of the storage provisioning process using software tools that typically accompany the particular storage platform.**
  - There is not a 1-to-1 ratio between drives and LUNs. Numerous LUNs can easily be carved out of a single disk drive.
  - A 500 GB drive can be partitioned into one 200 GB LUN and one 300 GB LUN, which would appear as two unique drives to the host server.
- **Storage administrators can employ Logical Volume Manager software to combine multiple LUNs into a larger volume.**
  - Veritas Volume Manager from Symantec Corp. is just one example of this software.
  - Disks are first gathered into a RAID group for larger capacity and redundancy (e.g., RAID-5), and then LUNs are carved from that RAID group.
- **Once created, LUNs can also be shared between multiple servers. For example, a LUN might be shared between an active and standby server. If the active server fails, the standby server can immediately take over.**



Example: Storage Capabilities Storage-as-a-Service (SaaS)

	Category	Performance (IOPS)	Availability
HDS TagmaStore HDS 9980	SAN – Tier 1	Random Read – Cache hit: 544k (9980) – 1.9M (TagmaStore) Random Read – Cache miss: 66k (9980) – 120k (TagmaStore)	99.999% (5 min unplanned downtime/yr)
HDS 9585	SAN – Tier 2 (Fiber Channel Drives)	Random Read – Cache hit: 240,000 Random Read – Cache miss: 53,000	99.95% (4.4 hrs unplanned downtime/yr)
NetApp FAS3050HA	NAS – Tier 1 (Fiber Channel Drives)	48,000 (SPEC sfs97_R1)	99.99% (< 1hr unplanned downtime/yr)
NetApp FAS3050	NAS – Tier 2 (SATA drives)	5,000 – 10,000 (estimated)	99.9% (8.8 hrs unplanned downtime/yr)
Tape	Tape	N/A	N/A



Storage specific terms that everyone should know

- **RAID** – Redundant Array of Inexpensive Disks, also known as Redundant Array of Independent Disks.
- **Disk subsystem** – A general term that refers to the disks on the server.
- **Spindle** – Spindles are another way to refer to the physical disk drives that make up the RAID array.
- **I/O Ops** – Input/Output operations, usually measured per second.
- **Queuing** – Number of I/O Ops that are pending completion by the disk subsystem.
- **SAN** (Storage Area Networks) – A collection of storage devices and fibre switches connected together along with the servers that access the storage on the device. SAN has also become a generic term, which refers to the physical storage drives such as EMC, 3PAR and Hitachi.
- **LUN** (Logical Unit Number) – This is the identification number assigned to a volume when created on a SAN device.
- **Physical drive** – How Windows sees any RAID array, single drive or LUN that is attached to the server.
- **Logical drive** – How Windows presents drives to the user (C:, D:, E:, etc.).
- **Block size** – The amount of data read from the spindles in a single read operation. This size varies per vendor from 8 KB to 256 MB.
- **Hardware array** – A RAID array created using a physical RAID controller.
- **Software array** – A RAID array created within Windows using the computer management snap-in.
- **Hot spare** – A spindle that sits in the drive cage and is added to the array automatically in the event of a drive failure. While this does not increase capacity, it does reduce the amount of time that the array is susceptible to data loss because of a second failed drive.
- **Recovery time** – Amount of time needed for the RAID array to become fully redundant after a failed drive has been replaced, either manually or automatically via a hot spare.