Data Science

Data Life Cycle

Physical Storage Devices

- Computación Avanzada y e-Ciencia – IFCA-CSIC
- Ibán Cabrillo Bartolomé
- Santander Enero del 2020

UC UNIVERSIDAD DE CANTABRIA

IFCA Instituto de Física de Cantabria

CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Summary

- Physical Storage Devices
- Network Storage Devices
- Data Storage
- Data Management
- Backup

Schema I

- Physical Storage Devices
  - HDDs
    - About Interface
      - SATA
      - SAS
      - NL-SAS
      - SSD
      - FC
    - Physical Properties
  - Tapes
      - LTO
      - DDS
      - VXA
      - Compatibility table
  - Common RAID Levels
    - Raid 0 (Data Striping)
    - Raid 1 (Mirroring)
    - Raid 3 (Parallel Data Access)
    - Raid 5 (Striping with Parity)
    - Raid 6 (Striping wit dual parity)
    - Raid 1+0 (Raid 10), RAID 0+1 (RAID 01)
    - DDP

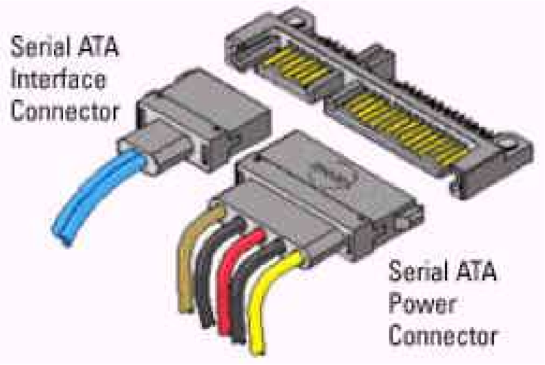
Schema II

- Disk Storage Systems
  - Arrays
  - LUN
  - Hot Spare
  - Storage Virtualization
- Examples

HDDs I

Interfaz Serial ATA (SATA)

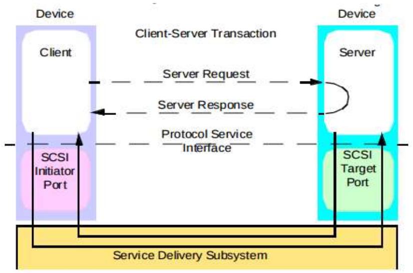
- The SATA specification defines three distinct protocol layers:
  - Physical layer define SATA's electrical and physical characteristics
  - link layer is responsible for transmission and reception of FISs over the SATA link
  - Transport layer has the responsibility of acting on the frames and for transmitting/receiving the frames in an appropriate sequence
- Parallel ATA (PATA) evolution
  - Data rate limited to 133MB/s
  - Two device per bus (Master & Slave)
- Direct connection (one device per bus)
  - We can use multiports
  - Three SATA levels
    - SATA I, 1500 MHZ, 150MB/s, year 2001
    - SATA II, 3000 MHZ, 300MB/s, year 2003
    - SATA III, 6000 Mhz, 600MB/s, year 2010
      - Native Command Queuing (NCQ) Management feature that helps optimize performance by enabling host processing
      - Multipath I/O
  - Wired to 2 m



HDDs II

Serial Attach SCSI (SAS) interface

- Evolution from common SCSI, year 2003
- Permit the fast connection/desconnection
- SAS allows up to 65,535 devices through the use of expanders, while Parallel SCSI has a limit of 8 or 16 devices on a single channel.
  - An *edge expander* allows for communication with up to 255 SAS addresses
  - A *fanout expander* can connect up to 255 sets of edge expanders
- SAS drives provide tagged command queuing.
- Three versions
  - SAS 1.0, 300 MB/s
  - SAS 2.0, 600 MB/s
    - From this version SATA device are allowed
  - SAS 3.0, 1200 MB/s 2013.
  - SAS 4.0, 22.5 Gbit/s Near Future



- Wired to 6 m
- Serial Attached SCSI system consists of the following basic components
  - the *Initiator*, is a device that originates device-service and task-management
  - The *Target*, is a device containing logical units and target ports that receives device service and task management requests.
  - The *Service Delivery Subsystem*, is the part of an I/O system that transmits information between an initiator and a target.
  - *The Expanders*, are devices that form part of a service delivery subsystem and facilitate communication between SAS devices (multiple SAS End devices to a single initiator port.

HDDs III

- The SAS specification defines distinct protocol layers:
  - Physical, defines electrical and physical characteristics
    - PHY layer, initialization, speed negotiation and reset sequences
  - Link, Establish and tear down native connections between SAS targets and initiators Establish and tear down tunnelled connections between SAS initiators and SATA targets connected to SAS expanders
    - Port, Combining multiple PHYs with the same addresses into wide ports
  - Transport layer Contains three transport protocols:
    - Serial SCSI Protocol (SSP): for command-level communication with SCSI devices
    - Serial ATA Tunnelled Protocol (STP): for command-level communication with SATA devices
    - Serial Management Protocol (SMP): for managing the SAS fabric
- NL-SAS
  - SATA drives with a SAS interface, head, media, and rotational speed of traditional enterprise-class SATA drives
    - Dual ports allowing redundant paths
    - Ability to connect a device to multiple computers
    - Full SCSI command set
    - Faster interface compared to SATA, up to 20%, no STP (Serial ATA Tunnelling Protocol) overhead

HDDs IV

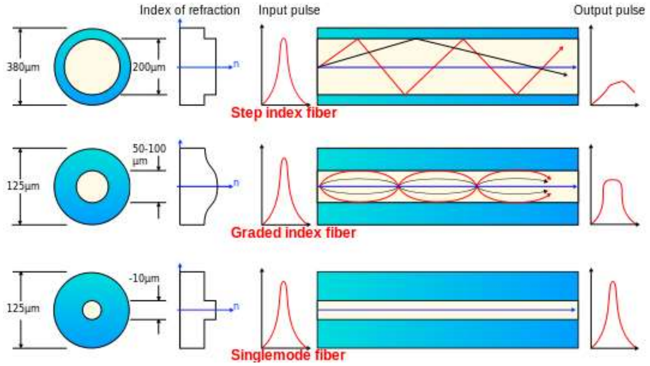
- Fibre Channel interface (FC)
  - Is a successor to parallel SCSI interface on enterprise market. It is a serial protocol Fibre Channel Protocol (FCP) is a transport protocol (similar to TCP used in IP networks) that predominantly transports SCSI commands over Fibre Channel networks
  - Not follow the OSI layer mode
    - FC0 – PHY includes cabling, connectors.
    - FC1 – Data link layer, which implements line coding of signals
    - FC2 – Network layer, defined by the FC-PI-2 standard, consists of the core of Fibre Channel, and defines the main protocols
    - FC3 – Common services layer, a thin layer that could eventually implement functions like encryption or RAID redundancy algorithms;
    - FC4 – Protocol-mapping layer, in which application protocols, such as SCSI or IP, are encapsulated into a PDU for delivery to FC2.
  - Topologies
    - Point-to-point (*FC-P2P)*. Two devices are connected directly to each other. This is the simplest topology, with limited connectivity (2 devices)

HDDs V

- **Arbitrated loop** (*FC-AL*). In this design, all devices are in a loop or ring, similar to token ring networking. Adding or removing a device from the loop causes all activity on the loop to be interrupted. The failure of one device causes a break in the ring. Fibre Channel hubs exist to connect multiple devices together and may bypass failed ports (128 devices).
  - In disk drives usually the Fibre Channel Arbitrated Loop (FC-AL) connection topology is used.

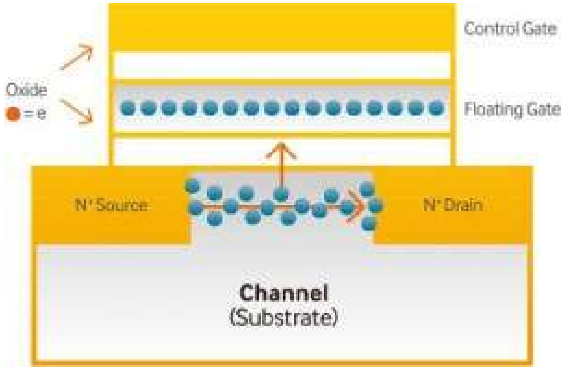
- **Switched fabric** (*FC-SW*). All devices or loops of devices are connected to Fibre Channel switches, similar conceptually to modern Ethernet implementations. ((2^{24} devices)

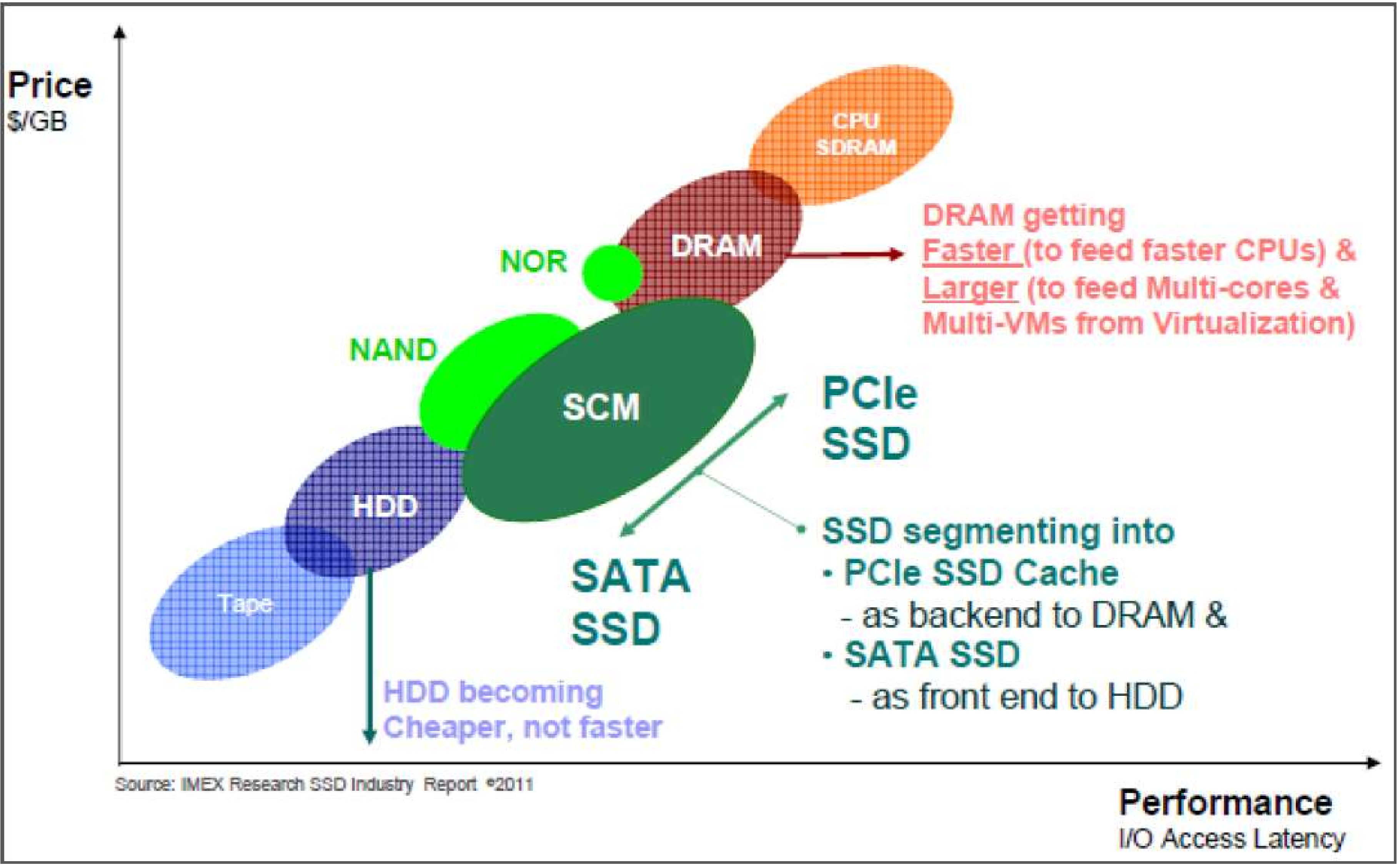
- high-speed network technology (commonly running at 2, 4, 8, 16, 32 and 128 Gbps) primarily used for storage networking
- Wired to 10 m using copper or 10Km with optical fibre.
- Has much broader usage than mere disk interfaces, and it is the cornerstone of storage area networks (SANs)
- Fiber Channel Host bus Adapter (HBA)has a unique World Wide Name (WWN).
  - **Node WWN** (**WWNN**), which can be shared by some or all ports of a device,
  - **Port WWN** (**WWPN**), which is necessarily unique to each port
- Fiber Types
  - Monomodo
    - Speeds 100-1600MB/s
    - Long distances
    - 1300 -1550 nm longwave light
  - Multimode
    - Sort distances
    - Speeds 100-1600MB/s
    - 850 nm shortwave light



HDDs VI

- SSDs offer significant performance and durability advantages over standard hard drives.
  - Have no moving parts
    - They are all semiconductor devices.
    - SSDs do not suffer from mechanical latencies like hard drives do
  - Can be subjected to much more shock and vibration than a typical mechanical hard drives.
  - Types
    - DRAM
      - Faster, but used for volatile puprposes, NOR gates
    - NAND ,Flash stores data in a large array of cells
      - NAND storage devices have a limited number of write cycles
        - Single level cell (SLC)
          - One bit for cell for SLC
          - SLC NAND would store a “0” or “1
        - Multi-Level Cell (MLC)
          - Two bits per cell for MLC
          - MLC NAND would store “00”, “01”, “10”, or “11”
        - Enterprise (grade) Multi-Level Cell (eMLC)
          - MLC with longer life.
        - Triple Level Cell (TLC)
          - TLC has higher power and error correction requirements.
          - TLC is targeted at environments with predominant read uses.
          - 000, 001, 010, 011, 100, 101, 110, 111
  - NAND is about 1000x faster than mechanical disks, but DRAM is 1000x faster than NAND





DRAM getting Faster (to feed faster CPUs) & Larger (to feed Multi-cores & Multi-VMs from Virtualization)

SSD segmenting into

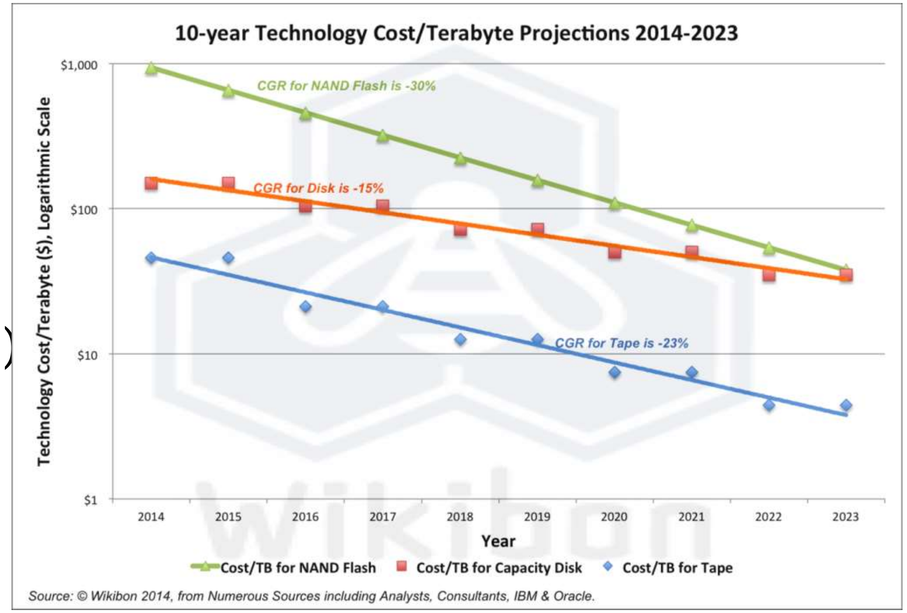
- PCIe SSD Cache
  - as backend to DRAM &
- SATA SSD
  - as front end to HDD

HDD becoming Cheaper, not faster

Source: IMEX Research SSD Industry Report ©2011

HDDs VII

- SAS and Fibre Channel Compared (Interesting reading)
- Physical properties



Source: © Wikibon 2014, from Numerous Sources including Analysts, Consultants, IBM & Oracle.

	RPMS	Seek (ms)	Read (MB/s)	Write (MB/s)	Errors	IOPS	€/GB	W	W (Idle)
SATA	5200-7200	9.5	150	120	10^{16}	~100	0.05	8	5
SAS	10000-15000	3.5	150-200	150-200	10^{14}	~300	0.2	7.5	4
SSD	NA	0.10	350	200	10^{16}	>3000~40000	0.25	3	4
FC	10000-15000	3.5	150-200	150-200	10^{16}	~300	0.33	7.5	4

$3398
10MB
THE HARD DISK YOU'VE BEEN WAITING FOR
XCOMP

XCOMP introduces a complete micro-size disk subsystem with more ...

- MORE STORAGE
- MORE SPEED
- MORE VALUE
- MORE SUPPORT

S100 users ... The XCOMP subsystem is now available with 10 megabytes of storage; 5 megabytes also available at $2,898.00. Compare the price and features of any other 5¼-inch — or even 8-inch system, and you'll agree that XCOMP's value is unbeatable.

OUTPERFORMS OTHER HARD DISKS
Floppy disk and larger, more expensive hard disks are no match for this powerful little system. More data is available on every seek: 64K on 10MB and 32K on 5MB. Faster seek time too — an average of 70MS. It provides solid performance anywhere with only 20 watts of power. Data is protected in the sealed enclosure, and the landing zone for heads provides another margin of safety. The optional power board plugs directly into the S100 bus and provides power for the drive.

FAST CONTROLLER
The XCOMP controller is the key to this system's high efficiency operation. Speed-up features include interleave without table lookup, block-deblock with controller buffer, and read lookahead. OEMs worldwide have already proven the outstanding performance of the XCOMP controller.

MORE SOFTWARE
Included with the system is software for testing, formatting, I/O drivers for CP/M*, plus an automatic CP/M driver attach program. Support software and drivers for MP/M* and Oasis* are also available. The sophisticated formatting program assigns alternate sectors for any weak sectors detected during formatting, assuring the lowest possible error rate — at least ten times better than floppies.

WARRANTY
The system has a full one-year warranty on parts and workmanship.

ALSO AVAILABLE FROM XCOMP

- General Purpose controllers (8 bit interface), with easy interface to microprocessor-based systems.
- GP controller adapter that plugs directly into most Z80 computers.
- ST/R GP controller for the 5MB and 10MB drive above, with ST506 type interface.
- SG/R GP controller for SA1000 interface.
- SM/R GP controller for storage module drives.
- ST/S, SG/S, and SM/S, same as above, for the S100 bus.

Quantity discounts available. Distributor, Dealer, and OEM inquiries invited.

See your local Dealer, or call

XCOMP, Inc.
7566 Trade Street
San Diego, CA 92121
Tel: (714) 271-8730
Telex: 182786

Circle 405 on inquiry card.

TAPES I

- Magnetic tape data storage technology, used mostly for backup or data tiering
- Slow speed access (If the tape is not at the drive, the autochanger has to load/unload the tape)
- Is a sequential device access
  - Linear Tape Open (LTO)
    - Developed in the late 1990s as an open standards alternative to the proprietary magnetic tape formats that were available at the time.

	LTO-1	LTO-2	LTO-3	LTO-4	LTO-5	LTO-6	LTO-7	LTO-8
release	2000	2003	2005	2007	2010	2012	2015	2017
Capacity (MB)	100	200	400	800	1.500	2.500	6000	12000
Speed (MB/s)	20	40	80	120	140	160	300	360
Compression	2:1	2:1	2:1	2:1	2:1	2.5:1	2.5:1	2:5:1
WORM	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Encryption	No	No	No	Yes	Yes	Yes	Yes	Yes

TAPES II

- Digital Data Storage
  - A format for storing computer data on a Digital Audio Tape (DAT).

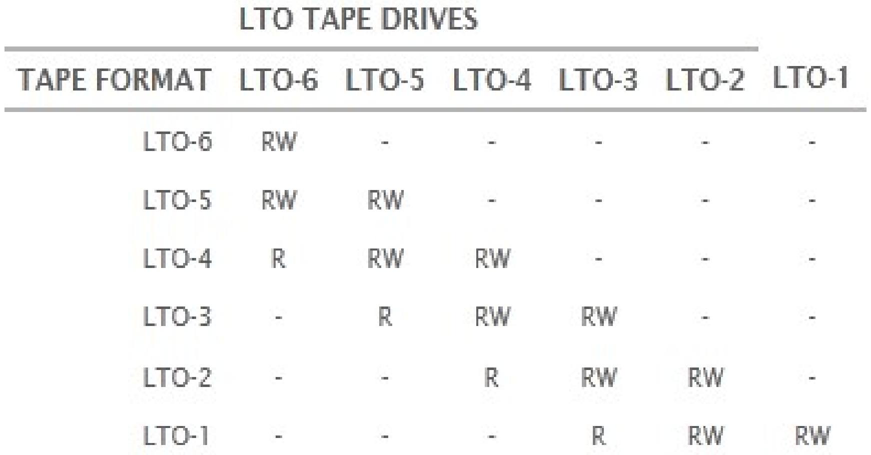
	DDS-1	DDS-2	DDS-3	DDS-4	DAT-72	DAT-160	DAT-320
release	1989	1993	1996	1999	2003	2007	2009
Capacity (MB)	2	4	12	20	36	80	160
Speed (MB/s)	0.18	0.6	1.1	3.2	3.2	6.9	12
Compression	2:1	2:1	2:1	2:1	2:1	2:1	2:1

- VXA
  - Tape backup format originally created by Ecrix and now owned by Tandberg Data
  - Based on helical scan technology
    - Data is written across the tape from side to side in helical strips.
    - ECC packet checksum

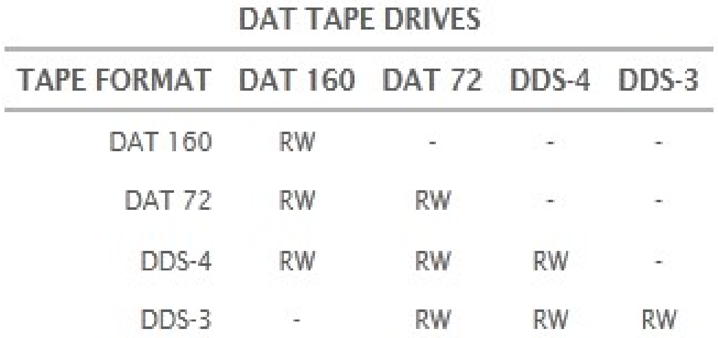
	VXA-1	VXA-2	VXA-3	VXA-4	VXA-5
release	1999	2002	2005	TBA	TBA
Capacity (MB)	22	80	160	320	640
Speed (MB/s)	3	6	20	24	48
Compression	2:1	2:1	2:1	2:1	2:1

TAPES IV

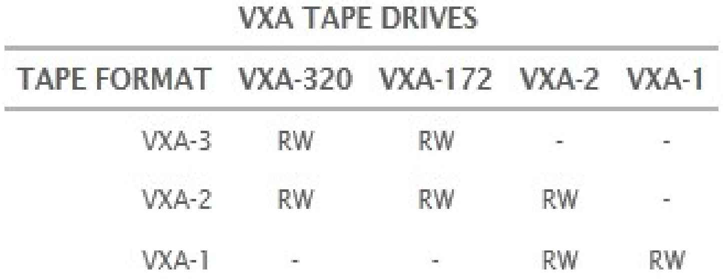
- Compatibility tape table



	LTO TAPE DRIVES					
TAPE FORMAT	LTO-6	LTO-5	LTO-4	LTO-3	LTO-2	LTO-1
LTO-6	RW	-	-	-	-	-
LTO-5	RW	RW	-	-	-	-
LTO-4	R	RW	RW	-	-	-
LTO-3	-	R	RW	RW	-	-
LTO-2	-	-	R	RW	RW	-
LTO-1	-	-	-	R	RW	RW



	DAT TAPE DRIVES			
TAPE FORMAT	DAT 160	DAT 72	DDS-4	DDS-3
DAT 160	RW	-	-	-
DAT 72	RW	RW	-	-
DDS-4	RW	RW	RW	-
DDS-3	-	RW	RW	RW



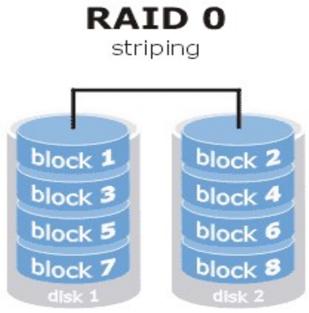
	VXA TAPE DRIVES			
TAPE FORMAT	VXA-320	VXA-172	VXA-2	VXA-1
VXA-3	RW	RW	-	-
VXA-2	RW	RW	RW	-
VXA-1	-	-	RW	RW

NOTE: LTO-8 brakes the compatibility table, can read LTO-7 but not LTO-6

RAIDs I

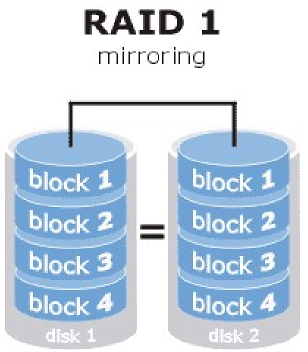
- Redundant Array of Independent Disks(RAID)
  - Makes reference a system that use multiple disk to distributed or replicate data

  - RAID 0 (Data Striping)
    - Splits data across two or more disks
    - No data redundancy
    - The array by each disk is limited to the size of the smallest disk
    - Reability decrease with the number of disk
    - the seek time of the array will be the same as that of a single drive for read and writes bigger than stripe size
    - The transfer speed of the array will be the transfer speed of all the disks added together, limited only by the speed of the RAID controller.

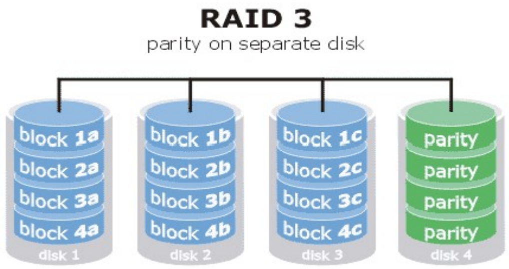


RAIDs II

- RAID 1 (Mirroring)
  - useful when read performance or reliability is more important than data storage capacity
  - can only be as big as the smallest member disk
  - not provide protection against data corruption due to viruses, accidental file changes or deletions, or any other data-specific changes
  - the read performance can go up roughly as a linear multiple of the number of copies, using independent raid controller for each disk

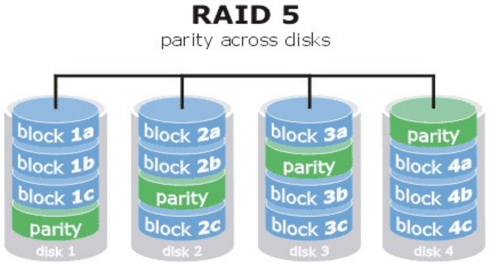


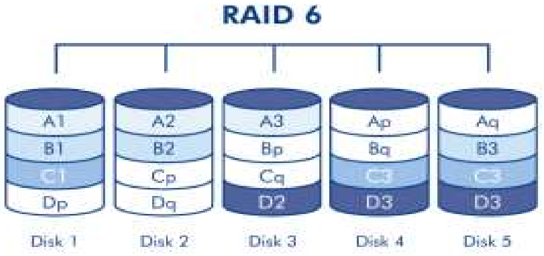
- RAID 3
  - byte-level striping with a dedicated parity disk
  - cannot service multiple requests simultaneously
  - The performance of the array is therefore identical to the performance of one disk in the array except for the transfer rate, which is multiplied by the number of data drives less the parity drives.



RAIDs III

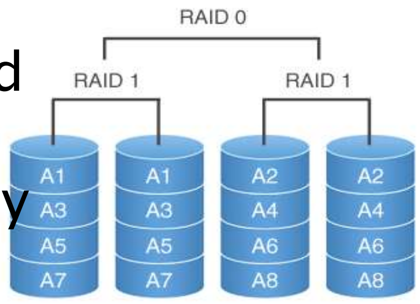
- RAID 5 (Striping with parity)
  - Block-level striping with parity data distributed across all member disks
  - Low cost of redundancy
  - Minimum of three disks is required for a complete RAID 5 configuration
  - Poor performance with writes smaller than the capacity of a single stripe
  - Modern controllers do not read the parity block
  - $\%Capacity = \frac{100*(N-1)}{N}$
  - No have a performance penalty for read operations.
  - Have a performance penalty on write operations because of the overhead associated with parity calculations
- RAID 6 (striping with dual parity)
  - extends RAID 5 by adding an additional parity block. Uses bloc-level striping with two parity blocks distributed across all member disks.
  - No have a performance penalty for read operations.
  - Have a performance penalty on write operations because of the overhead associated with parity calculations
  - $\%Capacity = \frac{100*(N-2)}{N}$



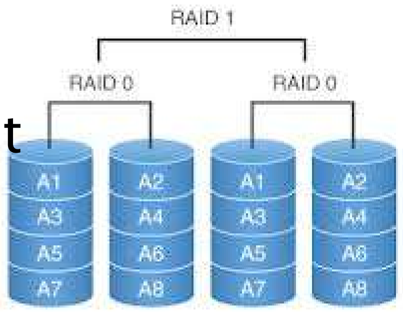


RAIDs IV

- RAID 1+0 (RAID 10) Strip of mirrors
  - top-level RAID-0 array (or *stripe set*) composed of two or more RAID-1 arrays
  - RAID 10 provides better throughput and latency than all other RAID levels except RAID 0
  - I/O-intensive applications such as database, email, and web servers



- RAID 0+1 (RAID 01) mirror of Stripes
  - top-level RAID-q array (mirror) composed of two or more RAID-0 arrays
  - 2 drive failure will cause the whole array to become, in essence, a RAID Level 0 array
  - for sites that need high performance but are not concerned with achieving maximum reliability



RAIDs V

- Dinamic Disk Pooling (DDP)
  - The Dynamic Disk Pool (DDP) feature dynamically distributes data, spare capacity, and protection information across a pool of disk drives.
  - DDP functions as effectively another RAID level offering in addition to the previously available RAID 0, 1, 10, 5, and 6 traditional RAID Disk Groups
  - DDP Rebuilding
    - Unlike RAID, there is no specific spare drive. All drives have spare space that is reserved.
    - When a drive fails, the remaining drives are read, the missing data is recomputed, and the result is written to multiple drives in their spare space. The result is parallel reads and parallel writes, which
  - significantly speeds up the rebuild time after a single drive failure.
    - It reduces rebuild times times by 5x for a single drive failure. (IE. 52hrs to 8hrs)
    - Reduces rebuild times by 10x when you have 2 drive failures. (100hrs to 12hrs)
    - Decreases the time required to add drives to the pool

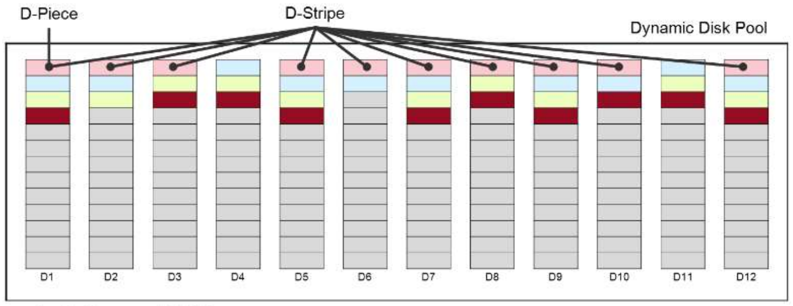


Figure 1 D-Piece and D-Stripe

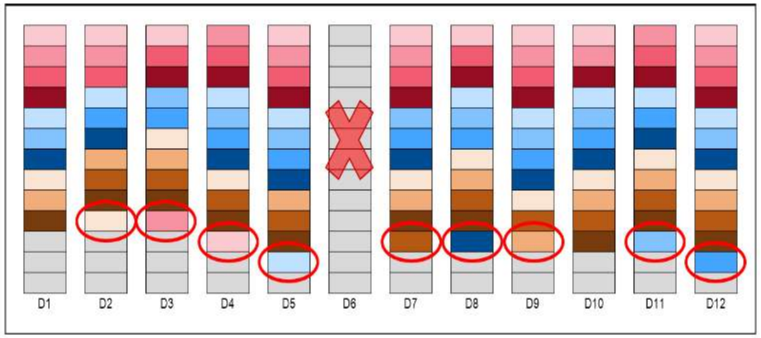


Figure 2 DDP Reconstruction

RAIDs VI

- The three pillars of RAID performance
  - Cache
    - Cache is simply RAM, or memory, placed in the data path in front of a disk or disk array.
    - Writing or reading cache is a lot faster than disk.
  - Striping
    - a virtual disk that the operating system sees, and spreading that virtual disk across several real, physical disks.
    - you now have the performance of several disks
  - Chunk size
    - Stripes go across disk drives
    - depends on your *average I/O request size*
      - big I/Os = small chunks (small files, DATBASES) sending each I/O to only one disk and spreading the I/Os evenly across the disks
      - small I/Os = big chunks (big files, sequential read/write) . Large chunk will mean that most I/Os get serviced by a single disk and more I/Os are available on the remaining disks.

RAIDs VII

	Min Drives	Read Perf	Write Perf	Degr Read Perf	Degr Write Perf	Data protection	Capacity	Application
RAID 0	2	High	High	N/A	N/A	No	100%	Transitory data, real time, rendering
RAID 1	2	High	Medium	Medium	High	Single-D	50%	Sos Databases
RAID 3	3	High	Low	Low	Low	Single-D	100(n-1)/n	---------
RAID 5	3	High	Low	Low	Low	Single-D	100(n-1)/n	Data, Archiving File serving
RAID 6	4	High	Low	Low	Low	Two-D	100(n-1)/n	Data, Archiving File serving High Avalability
RAID 1+0	4	High	Medium	High	High	Depend	100(n-2)/n	Fast Databases
RAID 0+1	4	High	High	High	High	Single-D	50%	---------
DDP	11	High	Low	Low	Low	Two-D	100(n-1)/n	Data, Archiving File serving High Avalability

Disk Storage

- Disk Storage Systems
  - From a simple external HDD (USB, Firewire, SAS) to the most sofisticated implementation (SAS,iSCSI, FC, FCoE)
  - A very big portfolio (IBM VDX, Fujitsu Eternus, DDN, EMC, NetApp, etc)
    - Dual Controller
      - Increasing the througput
      - Increasing the service availability
      - SAS, FC or FCoE external connection
      - Both can access to same enclosures
    - Expansion units (enclosure disk)
      - SAS or FC, redundan connection between them
  - Disk array
    - Disk storage system which contains multiple disk drives.
    - Cache memory and advanced functionality (RAID, virtualization)
  - Logical Unit number
    - Number used to identify a **logical unit**, which is a device addressed by the SCSI
    - A LUN may be used with any device which supports read/write operations often used to refer to a logical disk
  - Hot Spare Disk, a failover mechanism to provide reliability, when a disk fail, the hot spare is switched into operation
      - Storage Virtualizacion
        - LUNs are presented to the differents SOs like a real device

Examples

- Use cases.
  - Studied thus far, the best design storage implementation for the following cases. We must take into account economic factors, space and performance
    - Server SO consolidation
    - A WN local tmp big (1TB) area for an expensive I/O workload
    - Very big storage area, for video streaming (about 2PB)
    - A Critical Database (20000 IOPS)
    - Big Storage area, about 10 TB and a few disks (<20)
    - A “home” users area (about 1TB)