Storage Systems

Storage Systems

- Internet services such as search engines has enhanced the importance of I/O for computers
- shift in focus from computation to communication and storage of information
- Improving the capacity of disk – areal density

$$\text{Areal density} = \frac{\text{Tracks}}{\text{Inch}} \text{ on a disk surface} \times \frac{\text{Bits}}{\text{Inch}} \text{ on a track}$$

- *access time gap* between disks and DRAM
- DRAM latency is about 100,000 times less than disk, and that performance advantage costs 30 to 150 times more per gigabyte for DRAM

Understanding SAS, SATA, SCSI and ATA

- parallel interface has been widely used in storage systems
- parallel interface is a channel capable of transferring date in parallel mode
- Almost all personal computers come with at least one parallel interface
- parallel interfaces include SCSI and ATA

SCSI and ATA

- SCSI
  - small computer system interface
  - parallel interface standard systems for attaching peripheral devices to computers
  - provide for data transmission rates up to 80 megabytes per second
  - commonly used in servers, and more in industrial applications than home uses
- ATA
  - Advanced Technology Attachment
  - common interface used in many personal computers
  - used to connect hard disk drives, CD-ROM drives and similar peripherals
  - supports 8/16-bit interface that transfer up to 8.3MB/s for ATA-2 and up to 100MB/s for ATA-6

SATA

- SATA
  - Serial ATA
  - connects mass storage devices such as hard disk drives, optical drives, and solid-state drives.
  - Serial ATA succeeded the older Parallel ATA (PATA) standard,[a] offering several advantages over the older interface: reduced cable size and cost (seven conductors instead of 40 or 80), native hot swapping, faster data transfer through higher signaling rates, and more efficient transfer
- ATA
  - Advanced Technology Attachment
  - common interface used in many personal computers
  - used to connect hard disk drives, CD-ROM drives and similar peripherals
  - supports 8/16-bit interface that transfer up to 8.3MB/s for ATA-2 and up to 100MB/s for ATA-6

Disk Arrays

- Improves both dependability and performance of storage systems
- Potential throughput can be increased by having many disk drives and, hence, many disk arms, rather than fewer large drives
- Spreading data over multiple disks, called *striping*, automatically forces accesses to several disks if the data files are large
- With more devices, dependability decreases: *N* devices generally have 1/*N* the reliability of a single device
- Although disk array would have more faults than a smaller number of larger disks when each disk has the same reliability, dependability is improved by adding redundant disks to the array to tolerate faults

Disk Arrays

- If a single disk fails, the lost information is reconstructed form redundant information
- The only danger is in having another disk fail during the mean time to repair (MTTR)
- Since the mean time to failure (MTTF) of disks is tens of years and the MTTR is measured in hours, redundancy can make the measured reliability of many disks much higher than that of a single disk
- This redundant disk arrays are known by, RAID – Redundant Array of Inexpensive (Independent) Disks
- The ability to recover from failure plus the higher throughput, either measured as megabyte per second or as I/Os per second, makes RAID attractive
- When combined with the advantage of smaller size and lower power of smaller diameter drives, RAIDs now dominate large scale storage systems

Disk Arrays

- Although disk array would have more faults than a smaller number of larger disks when each disk has the same reliability, dependability is improved by adding redundant disks to the array to tolerate faults
- If a single disk fails, the lost information is reconstructed form redundant information
- The only danger is in having another disk fail during the mean time to repair (MTTR)
- Since the mean time to failure (MTTF) of disks is tens of years and the MTTR is measured in hours, redundancy can make the measured reliability of many disks much higher than that of a single disk
- This redundant disk arrays are known by, RAID – Redundant Array of Inexpensive (Independent) Disks

RAID Levels

- RAID 0: it has no redundancy- just a bunch of disks (JBOD). This level is generally included to act as a measuring stick for the other RAID levels in terms of cost, performance and dependability

- RAID 1: also called mirroring or shadowing. There are two copies of every piece of data. It is the simplest and oldest disk redundancy scheme, but has highest cost. Takes more time for the mirrored writes to complete

- RAID 2: it was inspired by applying memory style error correcting codes to disks. Not common

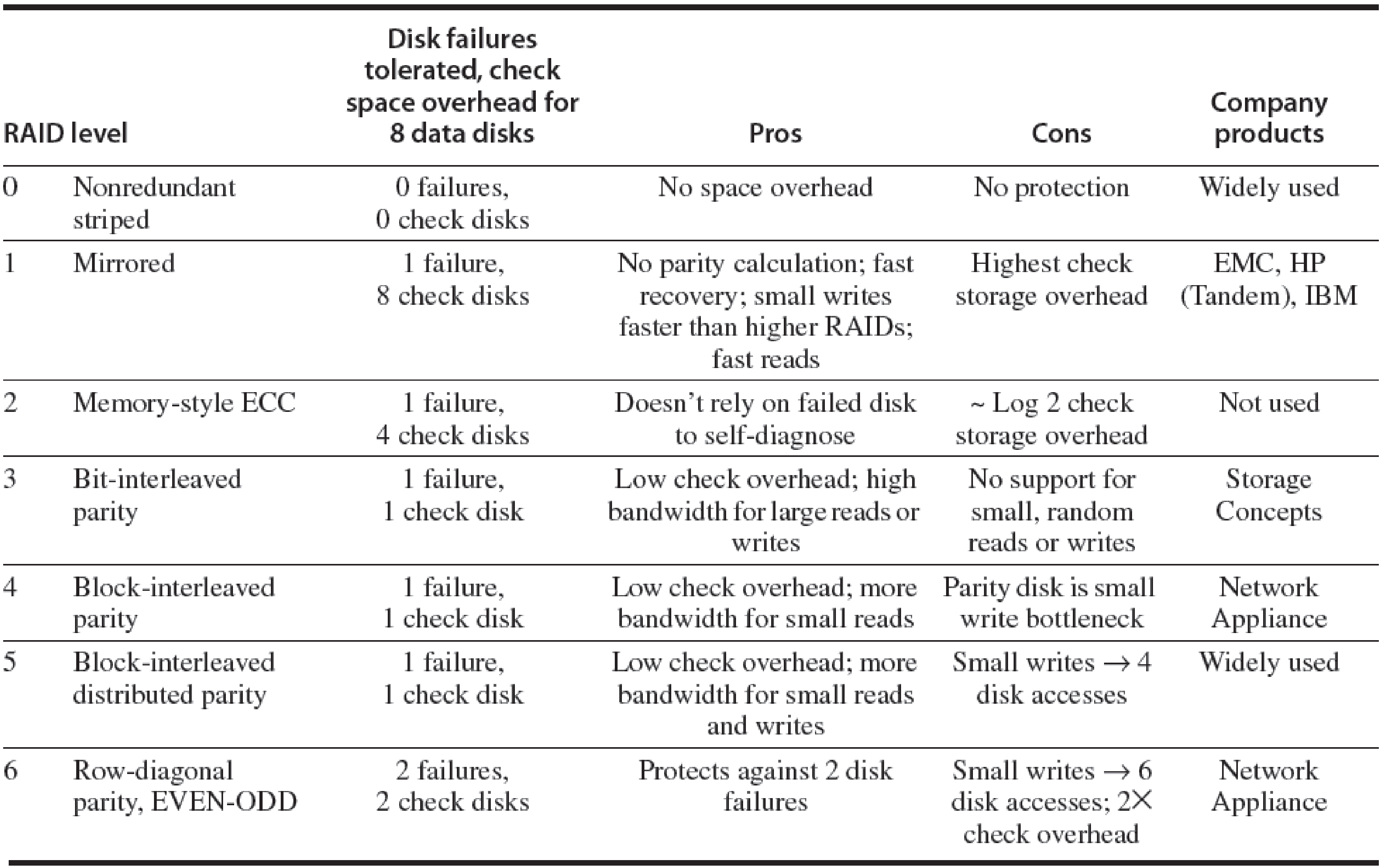
RAID Levels

- RAID 3: designers have realized that if one extra disk contains the parity of the information in the data disks, a single disk allows recovery from a disk failure. The data is organized in stripes with N data blocks and one parity block. When a failure occurs, you just subtract the good data from the good blocks, and what remains is the missing data

- RAID 4: many applications are dominated by small accesses. We can increase the number of reads per second by allowing each disk to perform independent reads. Writes would still be slower. To increase the number of writes per second, some alternative approaches are done in RAID 4.

- RAID 5: a performance flaw for small writes in RAID 4 is that they all must read and write the same check disk, so it is a performance bottleneck. RAID 5 distributes the parity information across all disks in the array, thereby removing the bottleneck. It requires the most sophisticated controller of the classic RAID levels

RAID Levels, Fault tolerance



RAID level		Disk failures tolerated, check space overhead for 8 data disks	Pros	Cons	Company products
0	Nonredundant striped	0 failures, 0 check disks	No space overhead	No protection	Widely used
1	Mirrored	1 failure, 8 check disks	No parity calculation; fast recovery; small writes faster than higher RAIDs; fast reads	Highest check storage overhead	EMC, HP (Tandem), IBM
2	Memory-style ECC	1 failure, 4 check disks	Doesn't rely on failed disk to self-diagnose	~ Log 2 check storage overhead	Not used
3	Bit-interleaved parity	1 failure, 1 check disk	Low check overhead; high bandwidth for large reads or writes	No support for small, random reads or writes	Storage Concepts
4	Block-interleaved parity	1 failure, 1 check disk	Low check overhead; more bandwidth for small reads	Parity disk is small write bottleneck	Network Appliance
5	Block-interleaved distributed parity	1 failure, 1 check disk	Low check overhead; more bandwidth for small reads and writes	Small writes → 4 disk accesses	Widely used
6	Row-diagonal parity, EVEN-ODD	2 failures, 2 check disks	Protects against 2 disk failures	Small writes → 6 disk accesses; 2× check overhead	Network Appliance

Warehouse-Scale Computers

Warehouse-scale Computers

- – Provides Internet services
  - Search, social networking, online maps, video sharing, online shopping, email, cloud computing, etc.

- – Differences with HPC “clusters”:
  - Clusters have higher performance processors and network
  - Clusters emphasize thread-level parallelism, WSCs emphasize request-level parallelism

- – Differences with datacenters:
  - Datacenters consolidate different machines and software into one location
  - Datacenters emphasize virtual machines and hardware heterogeneity in order to serve varied customers

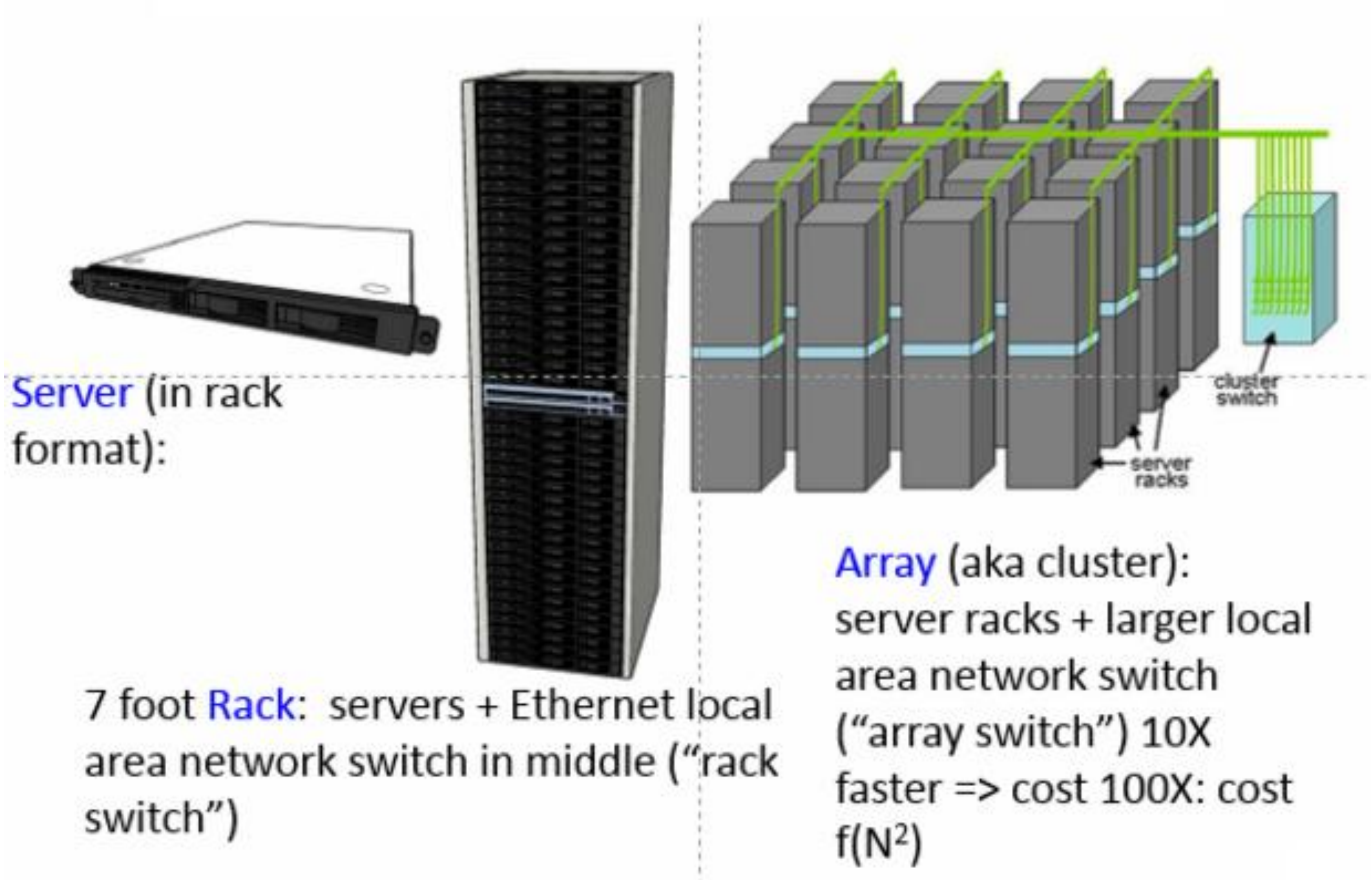
WSC

- Important design factors for WSC:
  - Cost-performance
    - Small savings add up
  - Energy efficiency
    - Affects power distribution and cooling
    - Work per joule
  - Dependability via redundancy
  - Network I/O
  - Interactive and batch processing workloads
  - Ample computational parallelism is not important
    - Most jobs are totally independent
    - “Request-level parallelism”
  - Operational costs count
    - Power consumption is a primary, not secondary, constraint when designing system
  - Scale and its opportunities and problems
    - Can afford to build customized systems since WSC require volume purchase

Computer Architecture of WSC

- WSC often use a hierarchy of networks for interconnection
- Each 19” rack holds 48 1U servers connected to a rack switch
- Rack switches are uplinked to switch higher in hierarchy
  - Uplink has 48 / n times lower bandwidth, where n = # of uplink ports
    - “Oversubscription”
  - Goal is to maximize locality of communication relative to the rack

Computer Architecture of WSC



Server (in rack format):

7 foot Rack: servers + Ethernet local area network switch in middle (“rack switch”)

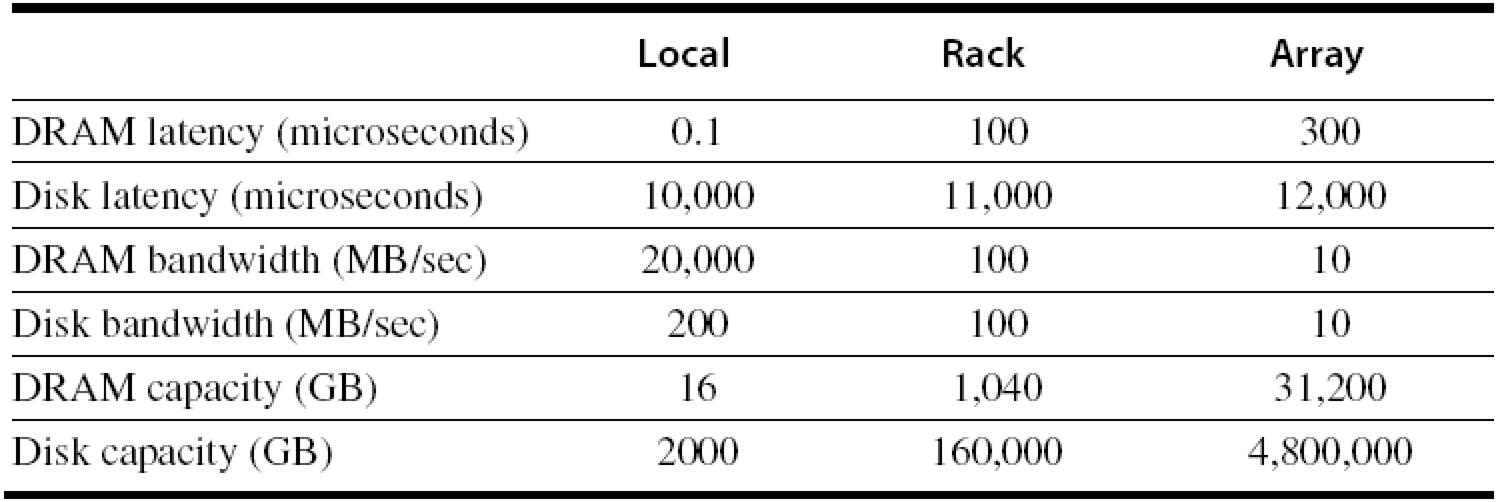
Array (aka cluster): server racks + larger local area network switch (“array switch”) 10X faster => cost 100X: cost f(N^2)

Computer Architecture of WSC

- Storage
  - Fill rack with servers
  - Collection of racks Array
- Switch that connects an array of racks
  - Array switch should have 10 X the bisection bandwidth of rack switch
  - Cost of *n*-port switch grows as n^2
  - Often utilize content addressible memory chips and FPGAs

Computer Architecture of WSC

- WSC Memory Hierarchy
  - **Servers can access DRAM and disks on other servers**



	Local	Rack	Array
DRAM latency (microseconds)	0.1	100	300
Disk latency (microseconds)	10,000	11,000	12,000
DRAM bandwidth (MB/sec)	20,000	100	10
Disk bandwidth (MB/sec)	200	100	10
DRAM capacity (GB)	16	1,040	31,200
Disk capacity (GB)	2000	160,000	4,800,000