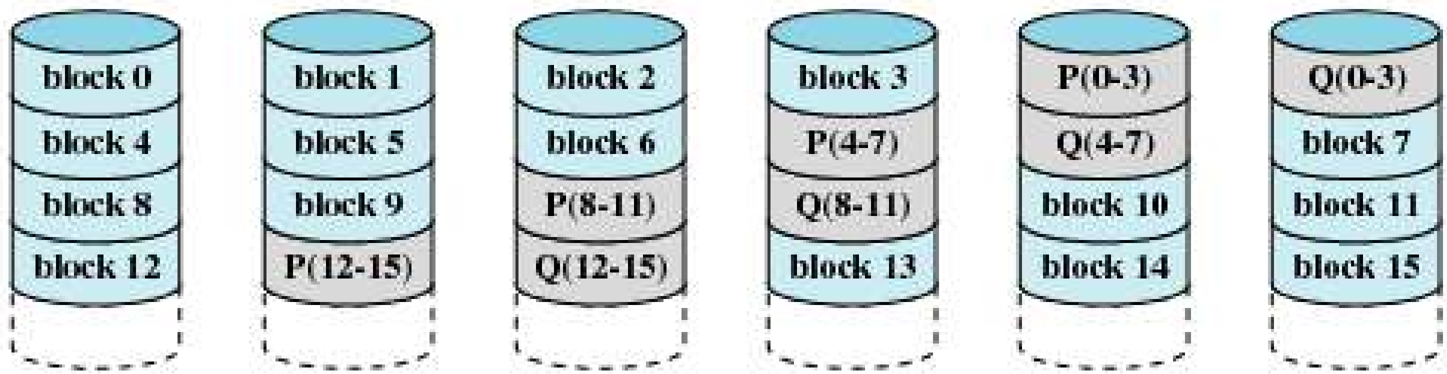
DATA STORAGE TECHNOLOGIES & NETWORKS (CS C446, CS F446 & IS C446)

LECTURE 21– STORAGE

RAID – RAID 6 or RAID PQ

- All the RAID schemes (1 to 5) are targeted at correcting a single self-identifying failures
  - What about multiple disk failures?
  - What about a read error while attempting to correct a disk failure (by reading all the other disks in an array)?
    - Typical "uncorrectable bit error" rate: 1 in 10^{14} as advertised by disk manufacturers
      - i.e. about 1 in 25 billion sectors
    - Typical errors – occur at write time or due to magnetic decay.

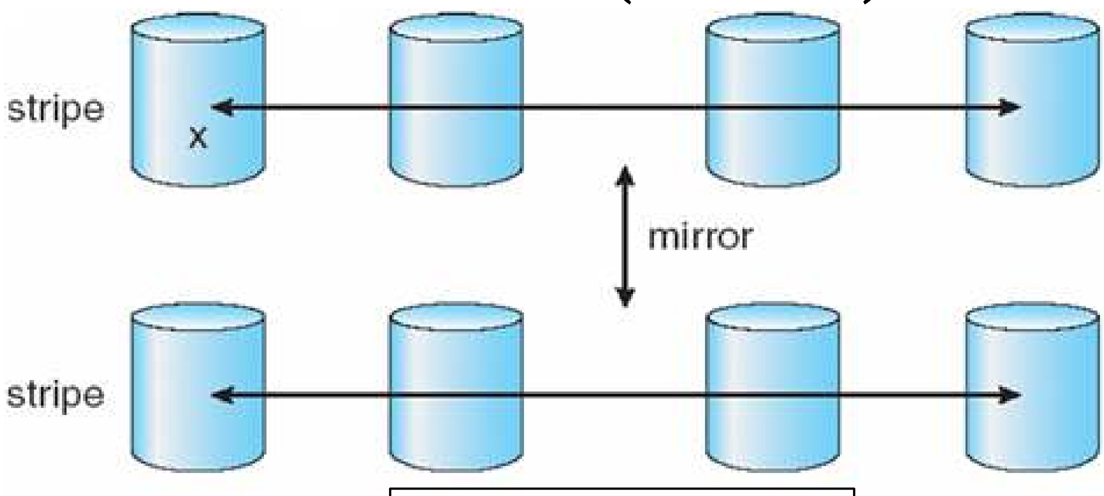
RAID 6 (P+ Q dual redundancy)

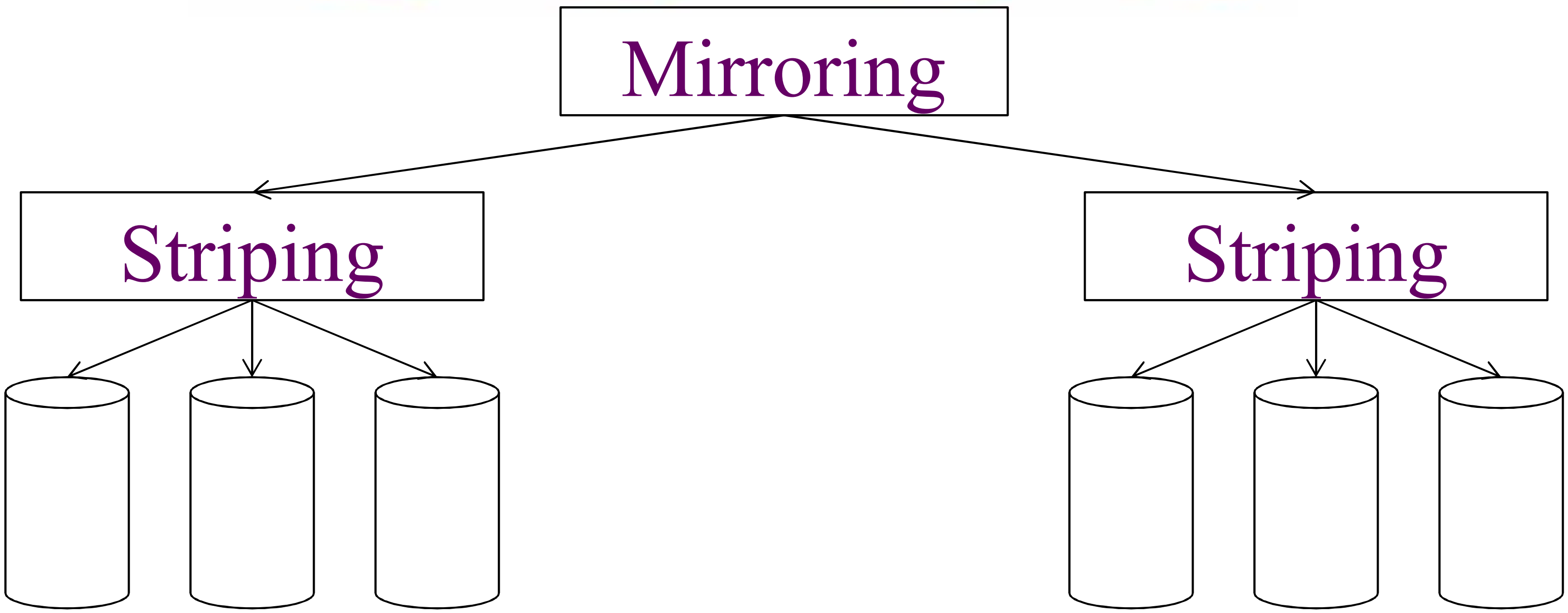


(g) RAID 6 (dual redundancy)

- Block inter leaved P+Q redundancy scheme organization
  - Stores extra redundant information to guard against multiple disk failures
  - Instead of parity, error correction codes such as Reed – Solomon codes are used
  - Can handle 2 disk failures
  - Store D data blocks and 2 parity blocks in a stripe Block interleaved Distributed Parity Storage

RAID (0 + 1)

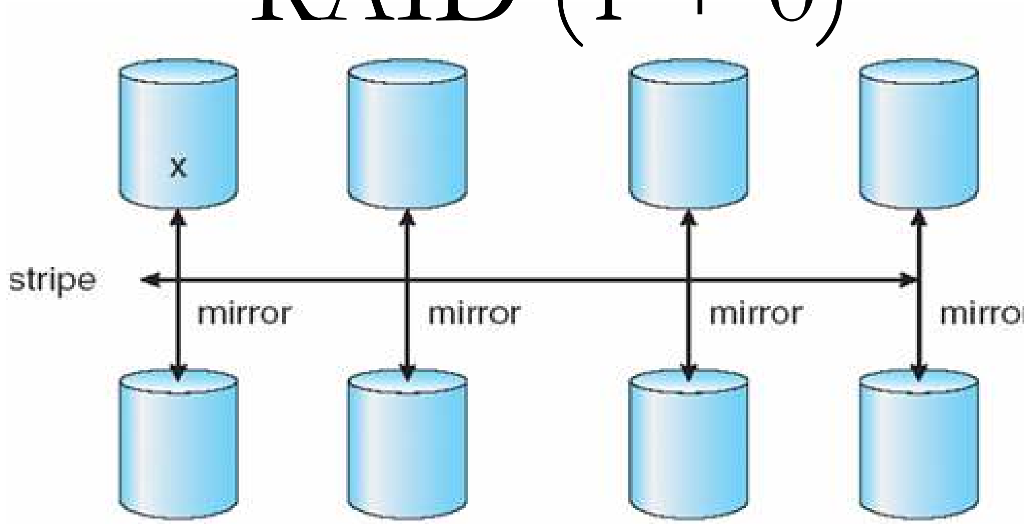


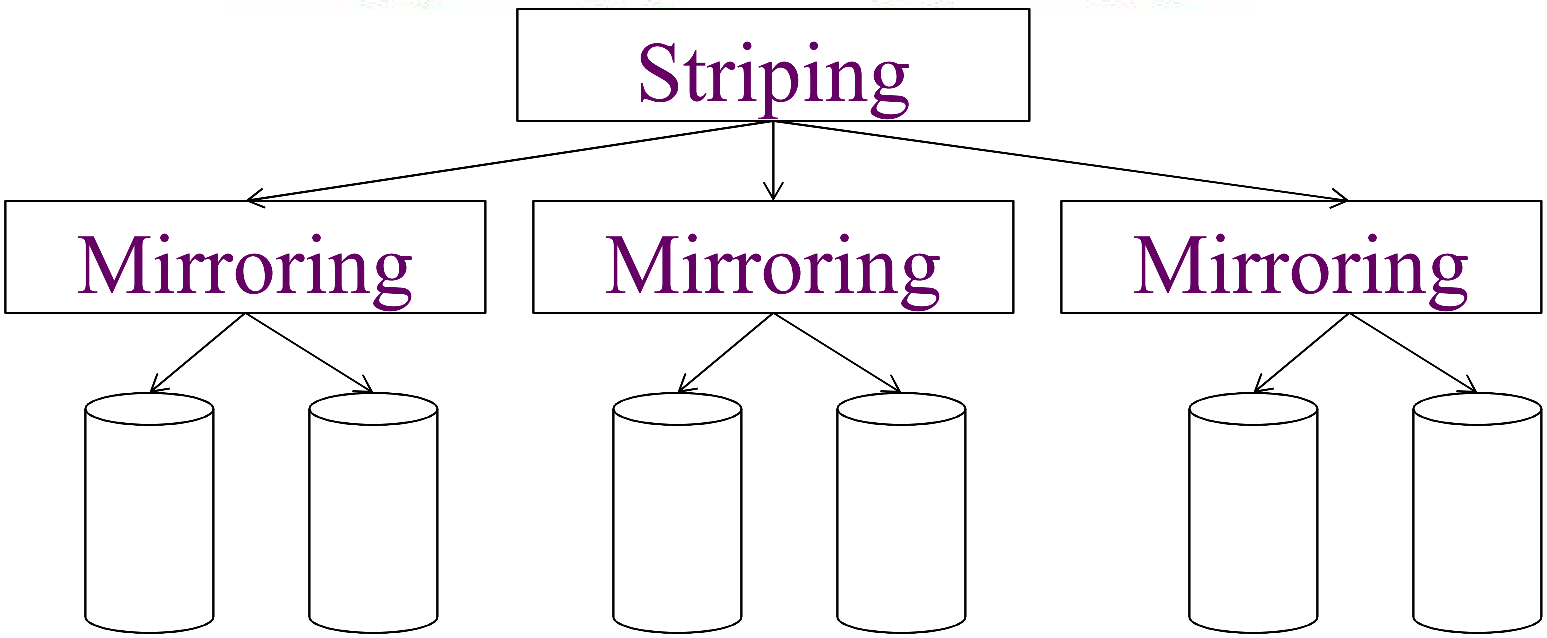


NESTED RAID

- RAID 0 + 1 [also known as RAID 01 or RAID 0/1]
  - Striped sets in a mirrored set [mirrored stripe]
  - Require even number of disks [min # disks =4]
  - If one drive fails, the entire stripe is faulted. Rebuild by copying entire stripe from each healthy disk to its corresponding failed
  - Causes increased and unnecessary disk I/O load on the surviving disks and makes the RAID set more vulnerable to a second disk failure.

RAID (1 + 0)


x
stripe
mirror
mirror
mirror
mirroı


Striping
Mirroring
Mirroring
Mirroring

- RAID 1 + 0 [also known as RAID 10 or RAID 1/0]
  - Mirrored sets in a striped set [called striped mirror]
  - Require even number of disks [min # disks =4]
  - Performs well for workloads that use small, random write intensive I/O
  - Applications
    - High transaction rate Online Transaction Processing (OLTP)
    - Large messaging installations
    - Database applications that require high I/O rate, random access and high availability

- RAID 5 + 1
  - Mirrored striped set with distributed parity

RAID TYPE	MIN. DISKS	STORAGE EFFICIENCY %	COST	READ PERFORMANCE	WRITE PERFORMANCE	WRITE PENALT
RAID 0	2	100	LOW	VERY GOOD FOR BOTH RANDOM AND SEQUENTIAL READ	VERY GOOD	NO
RAID 1	2	50	HIGH	GOOD. BETTER THAN A SINGLE DISK	GOOD. SLOWER THAN SINGLE DISK, AS EVERY WRITE MUST BE COMMITTED TO ALL DISKS	MODER ATE
RAID 3	3	(N-1)*100/N WHERE N=NUMBER OF DISKS	MODERATE	GOOD FOR RANDOM READS AND VERY GOOD FOR SEQUENTIAL READS	POOR TO FAIR FOR SMALL RANDOM WRITES. GOOD FOR LARGE, SEQUENTIAL WRITES	HIGH
RAID 4	3	(N-1)*100/N WHERE N=NUMBER OF DISKS	MODERATE	VERY GOOD FOR RANDOM READS. GOOD FOR SEQUENTIAL WRITES	POOR TO FAIR FOR SMALL RANDOM WRITES. FAIR TO GOOD FOR SEQUENTIAL WRITES	HIGH
RAID 5	3	(N-1)*100/N WHERE N=NUMBER OF DISKS	MODERATE	VERY GOOD FOR RANDOM READS. GOOD FOR SEQUENTIAL READS	FAIR FOR RANDOM WRITES. SLOWER DUE TO PARITY OVERHEAD. FAIR TO GOOD FOR SEQUENTIAL WRITES	HIGH
RAID 6	4	(N-2)*100/N WHERE N=NUMBER OF DISKS	MODERATE BUT MORE THAN RAID 5	VERY GOOD FOR RANDOM READS. GOOD FOR SEQUENTIAL READS	GOOD FOR SMALL. RANDOM WRITES (HAS WRITE PENALTY)	VERY HIGH
RAID 1+0 & 0+1	4	50	HIGH	VERY GOOD	GOOD	MODER ATE

Hot Spares

- Spare HDD in RAID array
  - Temporarily replaces a failed HDD of a RAID set
  - Data reconstructed [from parity if parity RAID is used or from mirror if mirroring is used] on to the hot spare
- When a new HDD replaces the old HDD
  - New HDD gets data from hot spare
  - Hot spare returns to idle state – ready to replace the next failed HDD
- Hot spare can be automatic or user initiated

Energy Efficiency in RAID

In RAID - 1

- Policies to dispatch a read request to disks at the RAID 1 controller to obtain high-performance
  - Send all read requests to a single primary replica
  - random selection
  - round-robin
  - shortest-seek first and shortest-queue first
  - selecting the replica with the shortest request queue on disk drive and having ties broken by random selection

- Energy efficient strategies (eRAID, EERAID)
- Better model: send all requests to one group such as primary dispatch
  - other group should always be idle.
  - This may result in spending more energy for the intensive I/O workloads because the aggregate access time for all requests is substantially stretched

- Design features of eRAID and EERAID
  - even small increases in the request interval length of inactive disks can result in significant energy savings
  - make sure the performance is not compromised after applying new schemes