Decentralized Services

- We want to take a look at decentralization as a way towards survivability
- The case study is: Survivable Storage Systems
  - What is Survivable Storage
  - Have we seen "flavors" of such concept before?
    - RAID technology can be considered survivable
    - However, malicious concepts were not considered
  - We want to look at the PASIS project
    - basis for the discussion on Survivable Storage are the PASIS papers
      - http://www.pdl.cmu.edu/Pasis/

Decentralization

- Before discussing Survivable Storage, I would like to briefly discuss the concept of RAIDs and how it plays into "thinking survivable"
  - The basis for the RAID discussion is the 1988 paper by Patterson
    - Patterson, D.A., et. al., "A Case for Redundant Arrays of Inexpensive Disks (RAID)", *ACM SIGMOD Records, International Conference on Management of Data*, Vol.~17, No.~3, pp.~109-116, June~1988.
  - The following material is probably to detailed for our discussion. I will only outline the basic concepts of RAID, as they will help to get a feeling for the performance issues associated with survivable storage
  - Note that the Patterson paper is very dated, yet, there are very interesting issues that are still valid!

RAID

- RAID Redundant Arrays of Inexpensive Disks
- Motivation
  - single chip computers improved in performance by 40% per year
  - RAM capacity quadrupled capacity every 2-3 years
  - Disks (magnetic technology)
    - » capacity doubled every 3 years
    - » price cut in half every 3 years
    - » raw seek time improved 7% every year
  - Note: values presented in Pattersons' paper are dated!
  - Note: paper discusses "pure" RAID, not smarter implementations, e.g. caching.

RAID

- Amdahl's Law:

$$S = \frac{1}{(1-f) + f/k}$$

Effective Speedup

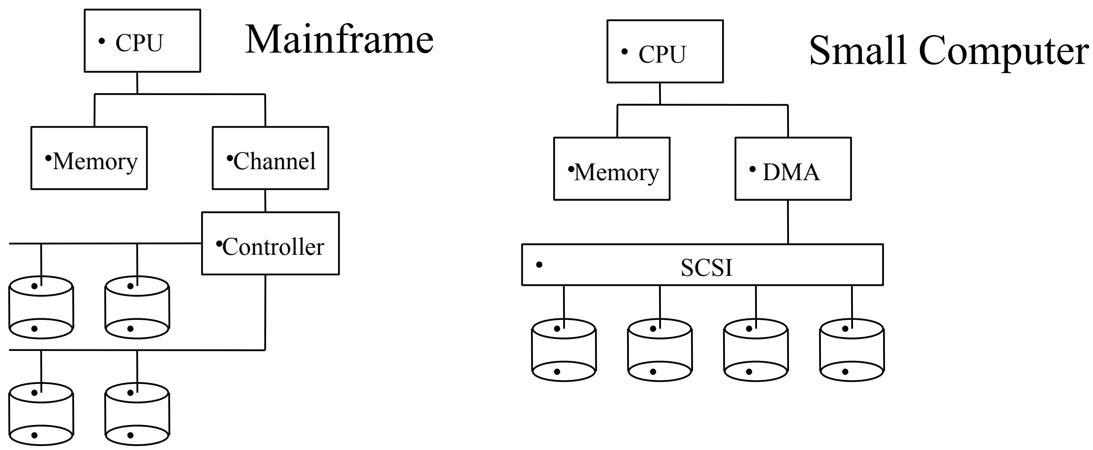
- » f = fraction of work in fast mode
- » k = speedup while in fast mode

Example:

- » assume 10% I/O operation
- » if CPU 10x => effective speedup is 5
- » if CPU 100x => effective speedup is 10
  - ■ 90 % of potential speedup is wasted

RAID

- Motivation
  - compare “mainframe mentality” with “today's” possibilities, e.g. cost, configuration



RAID

- Reliability

$$MTTF_{Array} = \frac{MTTF_{single}}{\#\ disks}$$

Bad news!

- e.g. $MTTF_{disk} = 30{,}000$ h

  $MTTF_{100} = 300$ h (< 2 weeks)

  $MTTF_{1000} = 30$ h
- Note, that these numbers are very dated. Todays drives are much better. MTBF > 300,000 to 800,000 hours.
- Is that really true though?????
- even if we assume higher MTTF of individual disks, the problem stays.

RAID

- RAID Reliability
  - partition disks into reliability groups and check disks
    - D = total number of data disks
    - G = # data disks in group
    - C = # check disks in group

$$MTTF_{\text{RAID group}} = \frac{MTTF_{\text{disk}}}{G+C} \times \frac{1}{\text{Prob. of failure during repair}}$$

$$\text{Prob. of failure during repair} = \frac{MTTR}{MTTF_{\text{disk}} / (G+C-1)}$$

$$MTTF_{\text{RAID}} = \frac{MTTF_{\text{RAID group}}}{\#\ \text{groups}}$$

RAID

- Target Systems
  - Different RAID solutions will benefit different target system configurations.
  - Supercomputers
    - larger blocks of data, i.e. high data rate
  - Transaction processing
    - small blocks of data
    - high I/O rate
    - read-modify-write sequences

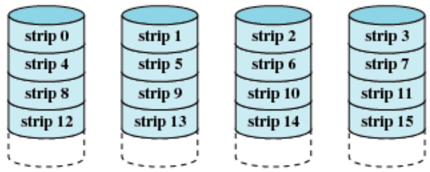
RAID

- 5 RAID levels
  - RAID 1: mirrored disks
  - RAID 2: hamming code for ECC
  - RAID 3: single check disk per group
  - RAID 4: independent read/writes
  - RAID 5: no single check disk

RAID

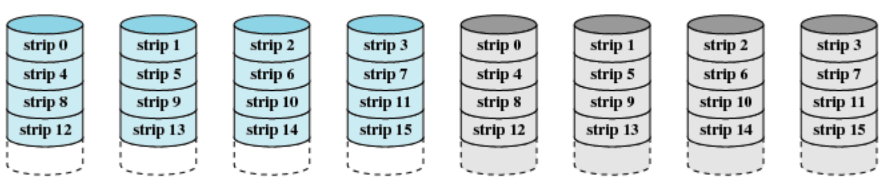
- other RAIDs (derived after the paper)
  - RAID 0
    - » employs striping with no redundancy at all
    - » claim of fame is speed alone
    - » has best write performance, but not the best read performance
      - why? (other RAIDs can schedule requests on the disk with the shortest expected seek and rotational delay)
  - RAID 6 (P + Q Redundancy)
    - » uses Reed-Solomon code to protect against up to 2 disk failures using the bare minimum of 2 redundant disks.

RAID 0 (non-redundant)



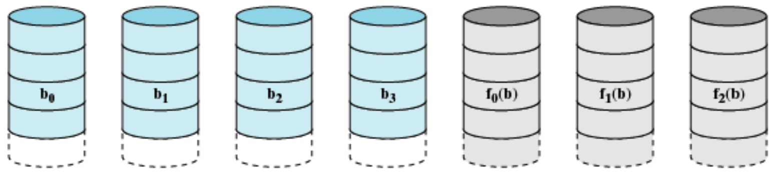
(a) RAID 0 (non-redundant)

RAID 1 (mirrored)



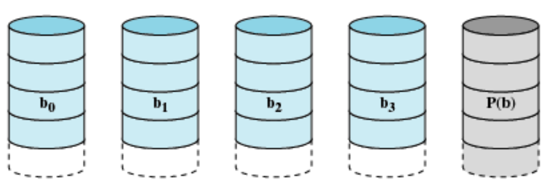
(b) RAID 1 (mirrored)

RAID 2 (redundancy through Hamming code)



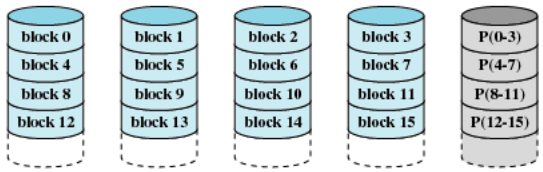
(c) RAID 2 (redundancy through Hamming code)

RAID 3 (bit-interleaved parity)



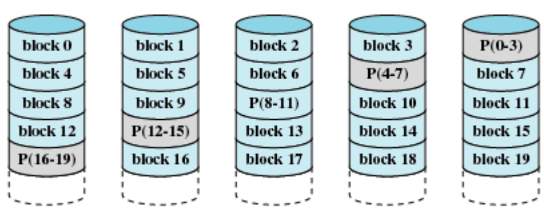
(d) RAID 3 (bit-interleaved parity)

RAID 4 (block-level parity)



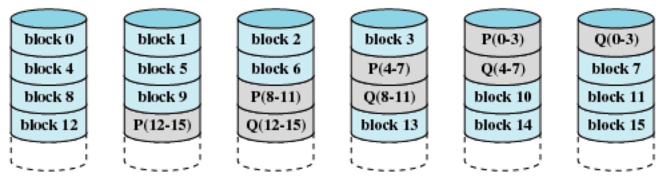
(e) RAID 4 (block-level parity)

RAID 5 (block-level distributed parity)



(f) RAID 5 (block-level distributed parity)

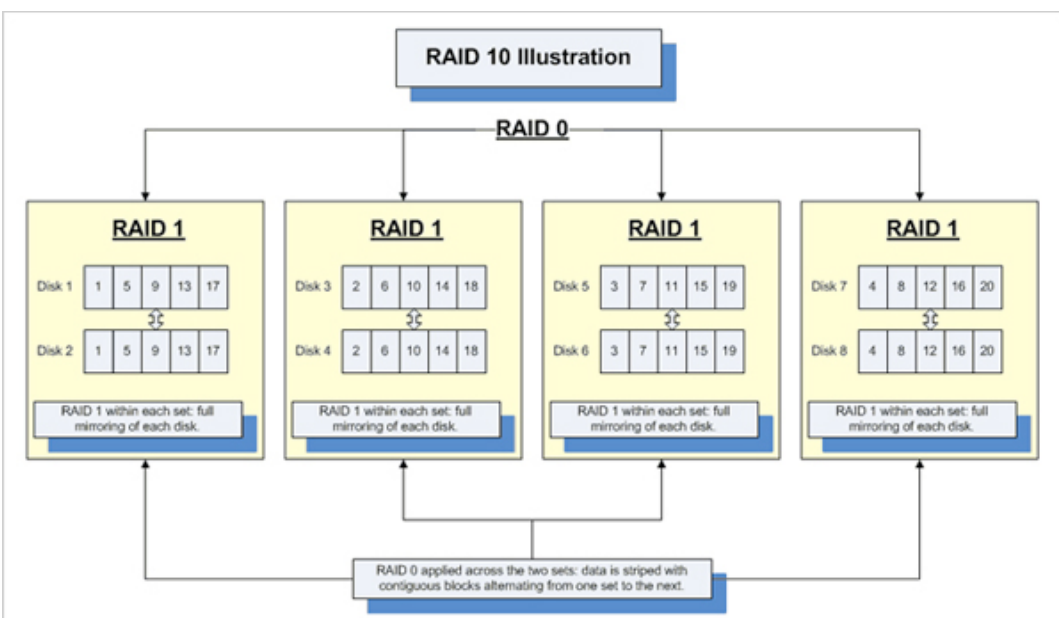
RAID 6 (dual redundancy)



(g) RAID 6 (dual redundancy)

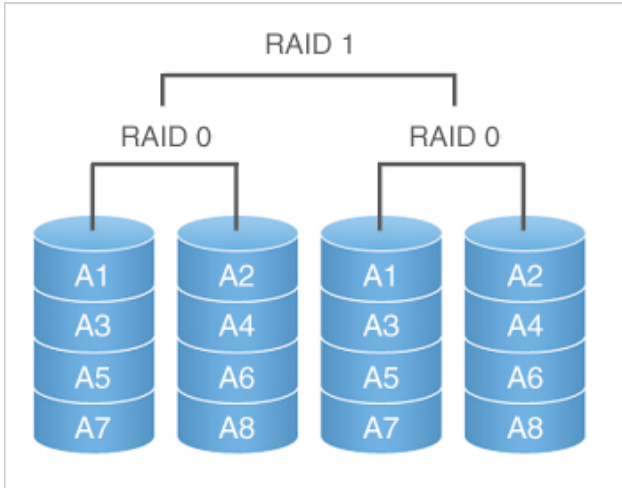
RAID 10

RAID 10 is sometimes also called RAID 1+0



source: http://www.illinoisdataservices.com/raid-10-data-recovery.html

RAID 0+1



source: http://www.illinoisdataservices.com/raid-10-data-recovery.html

RAID

- RAID level 1: Mirrored Disks
  - Most expensive option
  - Tandem doubles controllers too
  - Write to both disks
  - Read from one disk
  - Characteristics:
    - S = slowdown. In synchronous disks spindles are synchronized so that the corresponding sectors of a group of disks can be accessed simultaneously. For synchr. disks S = 1.
    - Reads= 2D/S, i.e. concurrent read possible
    - Write = D/S, i.e. no overhead for concurrent write of same data
    - R-Modify-Write = 4D/(3S)
    - Pat88 Table II (pg. 112)

RAID Pat88 Table II

MTTF	Exceeds Useful Product Lifetime (4,500,000 hrs or > 500 years)	
Total Number of Disks	2D	
Overhead Cost	100%	
Useable Storage Capacity	50%	
Events/Sec vs Single Disk	*Full RAID*	*Efficiency Per Disk*
Large (or Grouped) Reads	2D/S	1 00/S
Large (or Grouped) Writes	D/S	50/S
Large (or Grouped) R-M-W	4D/3S	67/S
Small (or Individual) Reads	2D	1 00
Small (or Individual) Writes	D	50
Small (or Individual) R-M-W	4D/3	67

Table II. *Characteristics of Level 1 RAID Here we assume that writes are not slowed by waiting for the second write to complete because the slowdown for writing 2 disks is minor compared to the slowdown S for writing a whole group of 10 to 25 disks Unlike a "pure" mirrored scheme with extra disks that are invisible to the software, we assume an optimized scheme with twice as many controllers allowing parallel reads to all disks, giving full disk bandwidth for large reads and allowing the reads of read-modify-writes to occur in parallel*

RAID

- RAID level 2: Hamming Code
  - DRAM => problem with α-particles
    - » Solution, e.g. parity for SED, Hamming code for SEC
  - Recall Hamming Code
  - Same idea using one disk drive per bit
  - Smallest accessible unit per disk is one sector
    - » access G sectors, where G = # data disks in a group
  - If operation on a portion of a group is needed:
    1) read all data
    2) modify desired position
    3) write full group including check info

Recall Hamming Code

12	11	10	9	8	7	6	5	4	3	2	1	Bit Position
0	1	0	1	0	1	0	1	0	1	0	1	
0	1	1	0	0	1	1	0	0	1	1	0	
1	0	0	0	0	1	1	1	1	0	0	0	
1	1	1	1	1	0	0	0	0	0	0	0	
				c4				c3		c2	c1	Check Bit
d8	d7	d6	d5		d4	d3	d2		d1			Data Bit

$$2^k > m + k + 1$$

m = data bits
k = parity bits

Compute Check

$$c1 = d1 \oplus d2 \oplus d4 \oplus d5 \oplus d7$$
$$c2 = d1 \oplus d3 \oplus d4 \oplus d6 \oplus d7$$
$$c3 = d2 \oplus d3 \oplus d4 \oplus d8$$
$$c4 = d5 \oplus d6 \oplus d7 \oplus d8$$

RAID

- Allows soft errors to be corrected "on the fly".
- Useful for supercomputers, not useful for transaction processing

  e.g. used in Thinking Machine (Connection Machine) "Data Vault" with G = 32, C = 8.
- Characteristics:
  - » Pat88 Table III (pg 112)

RAID Pat88 Table III

MTTF		Exceeds Useful Lifetime			
		G=10 (494,500 hrs or >50 years)		G=25 (103,500 hrs or 12 years)	
Total Number of Disks		1 40D		1.20D	
Overhead Cost		40%		20%	
Useable Storage Capacity		71%		83%	
Events/Sec (vs Single Disk)	Full RAID	Efficiency Per Disk L2	L2/L1	Efficiency Per Disk L2	L2/L1
Large Reads	D/S	71/S	71%	86/S	86%
Large Writes	D/S	71/S	143%	86/S	172%
Large R-M-W	D/S	71/S	107%	86/S	129%
Small Reads	D/SG	07/S	6%	03/S	3%
Small Writes	D/2SG	04/S	6%	02/S	3%
Small R-M-W	D/SG	07/S	9%	03/S	4%

Table III *Characteristics of a Level 2 RAID The L2/L1 column gives the % performance of level 2 in terms of level 1 (>100% means L2 is faster) As long as the transfer unit is large enough to spread over all the data disks of a group, the large I/Os get the full bandwidth of each disk, divided by S to allow all disks in a group to complete Level 1 large reads are faster because data is duplicated and so the redundancy disks can also do independent accesses Small I/Os still require accessing all the disks in a group, so only D/G small I/Os can happen at a time, again divided by S to allow a group of disks to finish Small Level 2 writes are like small R-M-W because full sectors must be read before new data can be written onto part of each sector*

RAID

- RAID level 3: Single Check Disk per Group
  - Parity is SED not SEC!
  - However, often controller can detect if a disk has failed
    - » information of failed disk can be reconstructed
    - » extra redundancy on disk, i.e. extra info on sectors etc.
  - If check disk fails
    - » read data disks to restore replacement
  - If data disk fails
    - » compute parity and compare with check disk
    - » if parity bits are equal => data bit = 0
    - » otherwise => data bit = 1

RAID

- Since less overhead, i.e. one check disk only

  => Effective performance increases
- Reduction in disks over L2 decreases maintenance
- Performance same as L2, however, effective performance per disk increases due to smaller number of check disks
- Better for supercomputers, not good for transaction proc.
- Maxtor, Micropolis introduced first RAID-3 in 1988
- Characteristics:
  - » Pat88 Table IV (pg 113)

RAID Pat88 Table IV (pg 113)

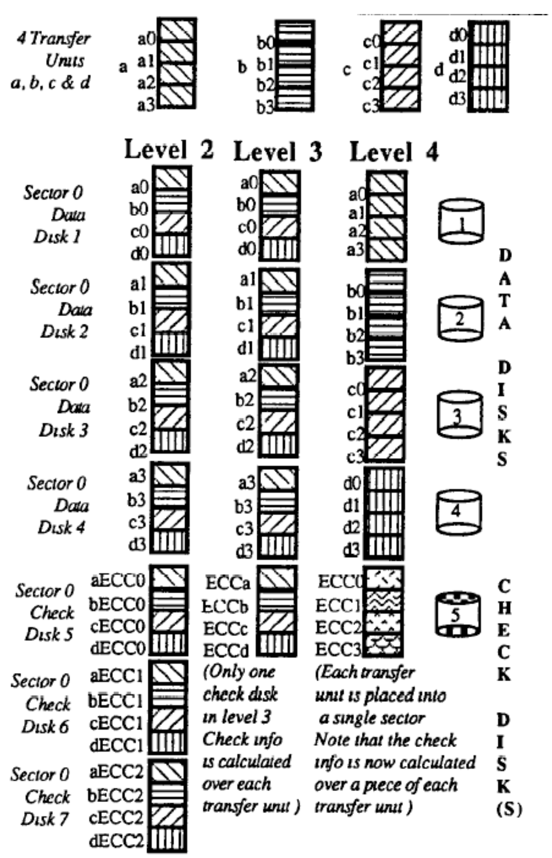
MTTF		Exceeds Useful Lifetime	
		G=10 (820,000 hrs or >90 years)	G=25 (346,000 hrs or 40 years)
Total Number of Disks		1 10D	1 04D
Overhead Cost		10%	4%
Useable Storage Capacity		91%	96%

Events/Sec (vs Single Disk)	Full RAID	Efficiency Per Disk L3	L3/L2	L3/L1	Efficiency Per Disk L3	L3/L2	L3/L1
Large Reads	D/S	91/S	127%	91%	96/S	112%	96%
Large Writes	D/S	91/S	127%	182%	96/S	112%	192%
Large R-M-W	D/S	91/S	127%	136%	96/S	112%	142%
Small Reads	D/SG	09/S	127%	8%	04/S	112%	3%
Small Writes	D/2SG	05/S	127%	8%	02/S	112%	3%
Small R-M-W	D/SG	09/S	127%	11%	04/S	112%	5%

Table IV *Characteristics of a Level 3 RAID The L3/L2 column gives the % performance of L3 in terms of L2 and the L3/L1 column gives it in terms of L1 (>100% means L3 is faster) The performance for the full systems is the same in RAID levels 2 and 3, but since there are fewer check disks the performance per disk improves*

RAID

- RAID level 4: Independent Reads/Writes
  - Pat88 fig 3 pg. 113 compares data locations



RAID

- RAID level 4: Independent Reads/Writes
  - Disk interleaving has advantages and disadvantages
  - Advantage of previous levels:
    - large transfer bandwidth
  - Disadvantages of previous levels:
    - all disks in a group are accessed on each operation (R,W)
    - spindle synchronization
      - if none => probably close to worse case average seek times, access times (tracking + rotation)
  - Interleave data on disks at sector level
  - Uses one parity disk

RAID

- for small accesses
  - need only access to 2 disks, i.e. 1 data & parity
  - new parity can be computed from old parity + old/new data
  - compute: $P_{new} = data_{old}$ XOR $data_{new}$ XOR P_{old}
- e.g. small write
  1) read old data + parity
  2) write new data + parity

  in parallel
- Bottleneck is parity disk
- e.g. small read
  - only read one drive (data)
- Characteristics:
  - Pat88 Table V (pg 114)

RAID

Pat88 Table V (pg 114)

MTTF	Exceeds Useful Lifetime	
	G=10 (820,000 hrs or >90 years)	G=25 (346,000 hrs or 40 years)
Total Number of Disks	1 10D	1 04D
Overhead Cost	10%	4%
Useable Storage Capacity	91%	96%

Events/Sec (vs Single Disk)	*Full RAID*	*Efficiency Per Disk* L4	L4/L3	L4/L1	*Efficiency Per Disk* L4	L4/L3	L4/L1
Large Reads	D/S	91/S	100%	91%	96/S	100%	96%
Large Writes	D/S	91/S	100%	182%	96/S	100%	192%
Large R-M-W	D/S	91/S	100%	136%	96/S	100%	146%
Small Reads	D	91	1200%	91%	96	3000%	96%
Small Writes	D/2G	05	120%	9%	02	120%	4%
Small R-M-W	D/G	09	120%	14%	04	120%	6%

Table V. *Characteristics of a Level 4 RAID The L4/L3 column gives the % performance of L4 in terms of L3 and the L4/L1 column gives it in terms of L1 (>100% means L4 is faster) Small reads improve because they no longer tie up a whole group at a time Small writes and R-M-Ws improve some because we make the same assumptions as we made in Table II the slowdown for two related I/Os can be ignored because only two disks are involved*

RAID

- RAID level 5: No Single Check Disk
  - Distributes data and check info across all disks, i.e. there are no dedicated check disks.
  - Supports multiple individual writes per group
  - Best of 2 worlds
    - » small Read-Modify-Write
    - » large transfer performance
    - » 1 more disk in group => increases read performance
  - Characteristics:
    - » Pat88 Table VI (pg 114)

RAID

- Pat88 Table VI (pg 114)

MTTF		Exceeds Useful Lifetime	
		G=10 (820,000 hrs or >90 years)	G=25 (346,000 hrs or 40 years)
Total Number of Disks		1 10D	1 04D
Overhead Cost		10%	4%
Useable Storage Capacity		91%	96%

Events/Sec (vs Single Disk)	*Full RAID*	*Efficiency Per Disk* L5	L5/L4	L5/L1	*Efficiency Per Disk* L5	L5/L4	L5/L1
Large Reads	D/S	91/S	100%	91%	96/S	100%	96%
Large Writes	D/S	91/S	100%	182%	96/S	100%	192%
Large R-M-W	D/S	91/S	100%	136%	96/S	100%	144%
Small Reads	(1+C/G)D	1 00	110%	100%	1 00	104%	100%
Small Writes	(1+C/G)D/4	25	550%	50%	25	1300%	50%
Small R-M-W	(1+C/G)D/2	50	550%	75%	50	1300%	75%

Table VI *Characteristics of a Level 5 RAID The L5/L4 column gives the % performance of L5 in terms of L4 and the L5/L1 column gives it in terms of L1 (>100% means L5 is faster) Because reads can be spread over all disks, including what were check disks in level 4, all small I/Os improve by a factor of 1+C/G Small writes and R-M-Ws improve because they are no longer constrained by group size, getting the full disk bandwidth for the 4 I/O's associated with these accesses We again make the same assumptions as we made in Tables II and V the slowdown for two related I/Os can be ignored because only two disks are involved*

RAID

- Patterson Paper
  - discusses all levels on pure hardware problem
  - refers to software solutions and alternatives, e.g. disk buffering
  - with transfer buffer the size of a track, spindle synchronization of groups not necessary
  - improving MTTR by using spares
  - low power consumption allows use of UPS
  - relative performance shown in Pat88 fig. 5 pg. 115

RAID

- relative performance shown in Pat88 fig. 5 pg. 115

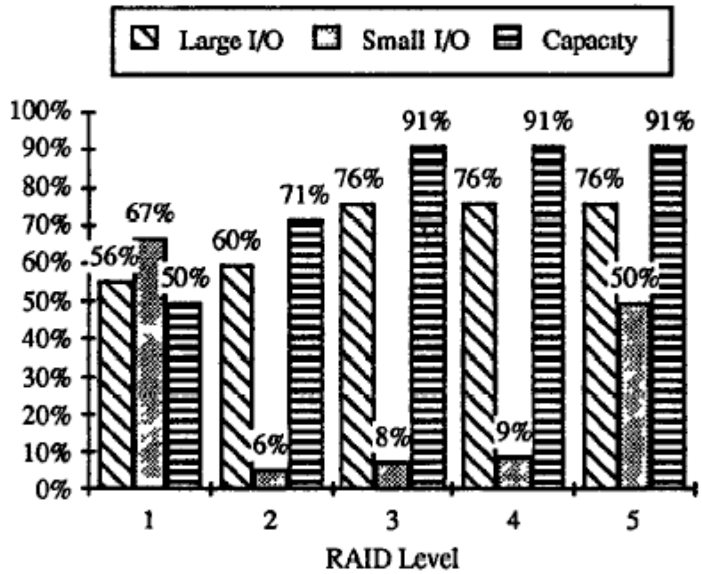


Figure 5 *Plot of Large (Grouped) and Small (Individual) Read-Modify-Writes per second per disk and useable storage capacity for all five levels of RAID (D=100, G=10) We assume a single S factor uniformly for all levels with S=1 3 where it is needed*

RAID

- Summary
  - Data Striping for improved performance
    - distributes data transparently over multiple disks to make them appear as a single fast, large disk
    - improves aggregate I/O performance by allowing multiple I/Os to be serviced in parallel
      - independent requests can be serviced in parallel by separate disks
      - single multiple-block block requests can be serviced by multiple disks acting in coordination
  - Redundancy for improved reliability
    - large number of disks lowers overall reliability of disk array
    - thus redundancy is necessary to tolerate disk failures and allow continuous operation without data loss

RAID

String management

Source:

RAID: high-performance, reliable secondary storage, Peter M. Chen, Edward K. Lee, Garth A. Gibson, Randy H. Katz, David A. Patterson, Journal ACM Computing Surveys (CSUR), Volume 26 Issue 2, June 1994, Pages 145 - 185

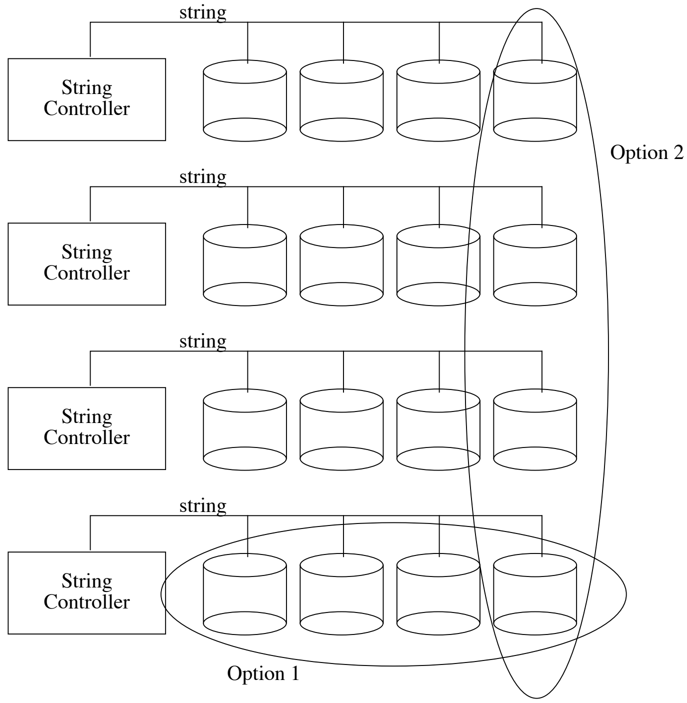


Figure 7: Orthogonal RAID. This figure present two options of how to organize error-correction groups in the presence of shared resources, such as a string controller. Option 1 groups four disks on the same string into an error-correction group; Option 2 groups one disk from each string into a group. Option 2 is preferred over Option 1 because the failure of a string controller will only render one disk from each error inaccessible.

RAID

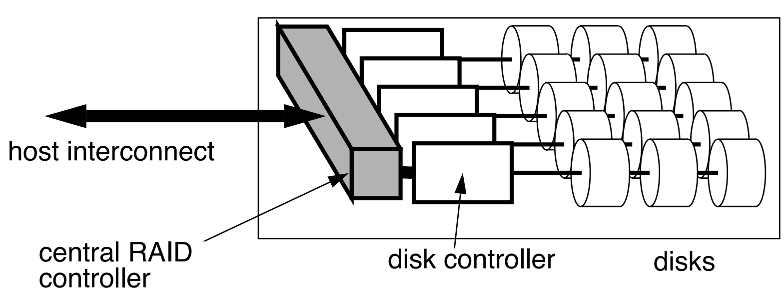
- Just to give you an idea about issues in commercial system
- These are old examples and serve only to give a historic perspective of key issues

RAID

- Case Studies
  - very early
  - Thinking Machines Corp.: TMC ScaleArray
    - » RAID level 3 for CM-5 massively parallel processor (MPP)
    - » high bandwidth for large files
    - » OS provides file system that can deliver data from a single file to multiple processors from multiple disks
    - » uses 4 SCSI-2 strings with 2 disks each (= 8 disks)
    - » these 4 strings are attached to an 8MB disk buffer
    - » 3 of these units are attached to the backbone (=> 3x8=24 disks)

RAID

- Case Studies
  - HP: TickerTAIP/DataMesh
    - » material shown is from "The TickerTAIP Parallel RAID Architecture", Cao et.al., ACM Trans. on Computer Systems, Vol.12, No.3, August 1994, pp.236-269.
    - » traditional RAID architecture
      - host interface
        - bottleneck
        - single point of failure

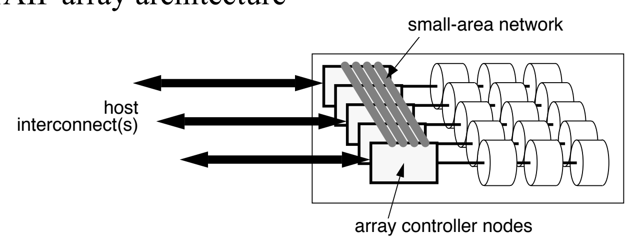


RAID

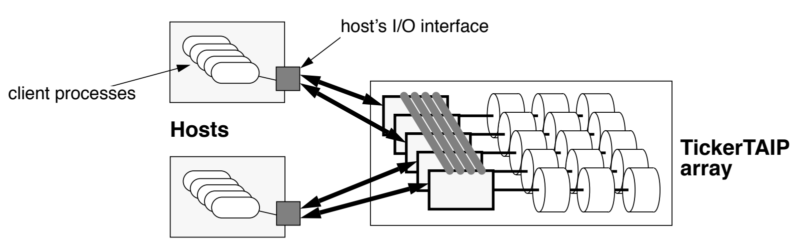
- Case Studies cont.
  - TickerTAIP/DataMesh Issues
    - » getting away from centralized architecture
    - » different algorithms for computing RAID parity
    - » techniques for establishing request atomicity, sequencing, and recovery
    - » disk-level request-scheduling algorithms inside the array

RAID

- Case Studies
  - HP: TickerTAIP/DataMesh
    - » TickerTAIP array architecture



TickerTAIP system environment



RAID

- Case Studies
  - HP: AutoRAID
    - provide a RAID that will provide excellent performance and storage efficiency in the presence of dynamically changing workloads
    - provides both level 1 and level 5 RAID
    - dynamically shift data to the “appropriate” level
    - dynamically shift data to level 5 if approaching maximum array capacity
    - parity logging
    - hot pluggable disks, spare controller, dynamically adapts to added capacity
    - Wilkes, J. et. al. “The HP AutoRAID hierarchical storage system”, ACM Trans. on Computer systems, 14, 1 (Feb.), 108-136, 1996.

RAID

- Case Studies
  - StorageTek: Iceberg 9200 Disk Array Subsystem
    - using 5.25-inch disks to look like traditional IBM mainframe disks
    - implements an extended RAID level 5 and level 6 disk array
    - array consists of 13 data drives, P and Q drives, and a hot spare
    - data, parity and Reed-Solomon coding are striped across the 15 active drives