Review: Storage System Issues

- Historical Context of Storage I/O
- Storage I/O Performance Measures
- Secondary and Tertiary Storage Devices
- A Little Queuing Theory
- Processor Interface Issues
- I/O & Memory Buses
- Show and Tell
- RAID
- ABCs of UNIX File Systems
- I/O Benchmarks
- Comparing UNIX File System Performance

© Patterson 1994

Review: Buses

- Bus: a shared communication link between subsystems.
  - low cost: a single set of wires is shared multiple ways
  - versatility:Easy to add new devices & peripherals may even be ported between computers using common bus
- Disadvantage: a communication bottleneck, possibly limiting the maximum I/O throughput
- Bus speed is limited by physical factors:
  - the bus length
  - the number of devices (and, hence, bus loading).
  - these physical limits prevent arbitrary bus speedup.
- 2 generic types of buses
  - I/O buses: lengthy, many types of devices connected, wide range in the data bandwidth), and follow a bus standard
  - CPU–memory buses: high speed, matched to the memory system to maximize memory–CPU bandwidth, single device
  - To lower costs, low cost (older) systems combine together
- bus transaction: sending address & receiving or sending data

© Patterson 1994

Bus Options

Option	*High performance*	*Low cost*
Bus width	Separate address & data lines	Multiplex address & data lines
Data width	Wider is faster (e.g., 32 bits)	Narrower is cheaper (e.g., 8 bits)
Transfer size	Multiple words has less bus overhead	Single-word transfer is simpler
Bus masters	Multiple (requires arbitration)	Single master (no arbitration)
Split transaction?	Yes—separate Request and Reply packets gets higher bandwidth (needs multiple masters)	No—continuous connection is cheaper and has lower latency
Clocking	Synchronous	Asynchronous

© Patterson 1994

1990 Bus Survey

	VME	FurtureBus	Multibus II	IPI	SCSI
Signals	128	96	96	16	8
Addr/Data mux	no	yes	yes	n/a	n/a
Data width	16 - 32	32	32	16	8
Masters	multi	multi	multi	single	multi
Clocking	Async	Async	Sync	Async	either
MB/s (0ns, word)	25	37	20	25	1.5 (asyn) 5 (sync)
150ns word	12.9	15.5	10	=	=
0ns block	27.9	95.2	40	=	=
150ns block	13.6	20.8	13.3	=	=
Max devices	21	20	21	8	7
Max meters	0.5	0.5	0.5	50	25
Standard	IEEE 1014	IEEE 896.1	ANSI/IEEE 1296	ANSI X3.129	ANSI X3.131

© Patterson 1994

1993 I/O Bus Survey

Bus	SBus	TurboChannel	MicroChannel	PCI
Originator	Sun	DEC	IBM	Intel
Clock Rate (MHz)	16-25	12.5-25	async	33
Addressing	Virtual	Physical	Physical	Physical
Data Sizes (bits)	8,16,32	8,16,24,32	8,16,24,32,64	8,16,24,32,64
Master	Multi	Single	Multi	Multi
Arbitration	Central	Central	Central	Central
32 bit read (MB/s)	33	25	20	33
Peak (MB/s)	89	84	75	111 (222)
Max Power (W)	16	26	13	25

© Patterson 1994

1993 MP Server Memory Bus Survey

Bus	Summit	Challenge	XDBus
Originator	HP	SGI	Sun
Clock Rate (MHz)	60	48	66
Split transaction?	Yes	Yes	Yes?
Address lines	48	40	??
Data lines	128	256	144 (parity)
Data Sizes (bits)	512	1024	512
Clocks/transfer	4	5	4?
Peak (MB/s)	960	1200	1056
Master	Multi	Multi	Multi
Arbitration	Central	Central	Central
Addressing	Physical	Physical	Physical
Slots	16	9	10
Busses/system	1	1	2
Length	13 inches	12? inches	17 inches

© Patterson 1994

SCSI: Small Computer System Interface

- up to 8 devices to communicate on a bus or ``string'' at sustained speeds of 4-5 MBytes/sec
- SCSI-2 up to 20 MB/sec
- Devices can be slave (“target”) or master(“initiator”)
- SCSI protocol: a series of ``phases'', during which specific actions are taken by the controller and the SCSI disks
  - Bus Free: No device is currently accessing the bus
  - Arbitration: When the SCSI bus goes free, multiple devices may request (arbitrate for) the bus; fixed priority by address
  - Selection: informs the target that it will participate (Reselection if disconnected)
  - Command: the initiator reads the SCSI command bytes from host memory and sends them to the target
  - Data Transfer: data in or out, initiator: target
  - Message Phase: message in or out, initiator: target (identify, save/restore data pointer, disconnect, command complete)
  - Status Phase: target, just before command complete

© Patterson 1994

VME

- 3 96-pin connectors
- 128 defined as standard, rest customer defined
  - 32 address
  - 32 data
  - 64 command & power/ground lines

© Patterson 1994

Show and Tell

- In class demo of 10.5", 5.25", 3.5", and 3.5" disks
- In class demo SCSI chain
- In class demo SIMM cards
- In class demo of Sun 4, first commercially successful RISC machine: VME bus, rack, memory cards
- In class demo of SPARCstation-1, first "pizza box" workstation: SCSI, SIMM modules, S-bus

© Patterson 1994

Improving Bandwidth of Secondary Storage

- Processor performance growth phenomenal
- I/O?

"I/O certainly has been lagging in the last decade."
Seymour Cray, Public Lecture (1976)

"Also, I/O needs a lot of work."
David Kuck, Keynote Address, (1988)

© Patterson 1994

What have Magnetic Disk Designers been doing?

- Capacity: Improving 25% per year
- $/MB: Improving 25% per year
- Evolving to smaller physical size

14" ➡ 10" ➡ 8" ➡ 5.25" ➡
3.5" ➡ 2.5" ➡ 1.8" ➡ 1"?

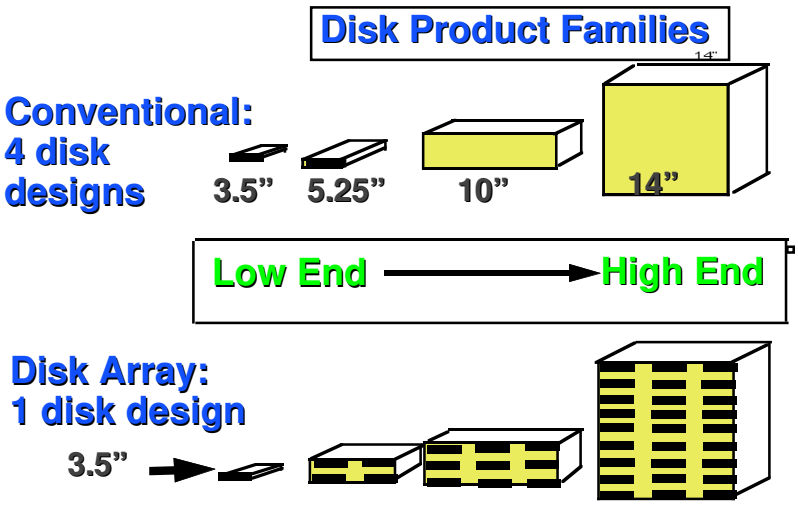
© Patterson 1994

Smaller and cheaper vs. Larger and faster

Can smaller disks be used to close gap in performance between disks and CPUs?

© Patterson 1994

Manufacturing advantages of disk arrays



© Patterson 1994

Arrays of Disks to Close the Performance Gap

Replace Small Number of Large Disks

With Large Number of Small Disks!

© Patterson 1994

	IBM 10.4" 3390	IBM 3.5" 0061	x 70
Data Capacity	23 GBytes	320 MBytes	23 GBytes
Volume	97 cu. ft.	0.1 cu. ft.	11 cu. ft.
Power	3 KW	11 W	1 KW
Data Rate	15 MB/s	1.5 MB/s	120 MB/s
I/O Rate	600 I/Os/s	55 I/Os/s	3900 IOs/s
Cost	≈$200K	≈$2K	≈$150K
MTTDL (hrs)	250K	50K	???

© Patterson 1994

Disk Arrays have potential for:

- large data and I/O rates
- high MB per cu. ft.
- high MB per KW

But is there a catch?

© Patterson 1994

Array Reliability

- Reliability of N disks =
  Reliability of 1 Disk ÷ N

  50,000 Hours ÷ 70 disks = 700 hours

  Disk system MTTF: Drops from
  6 years to 1 month!

- Arrays Too Unreliable to be useful!

© Patterson 1994

Improving Reliability with Redundancy

- Add redundant drives to handle failures
  Redundant
  Array of
  Inexpensive
  Disks
  RAID
- Redundancy offers 2 advantages:
  1) Data not lost:
     Reconstruct data onto new disk
  2) Continuous operation
     in the presence of failure

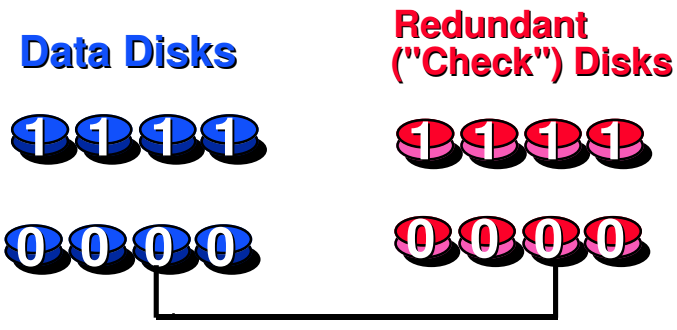
© Patterson 1994

Several RAID organizations

Mirroring/Shadowing	***(Level 1 RAID)***
ECC	***(Level 2 RAID)***
Parity	***(Level 3 RAID)***
Rotated Parity	***(Level 5 RAID)***

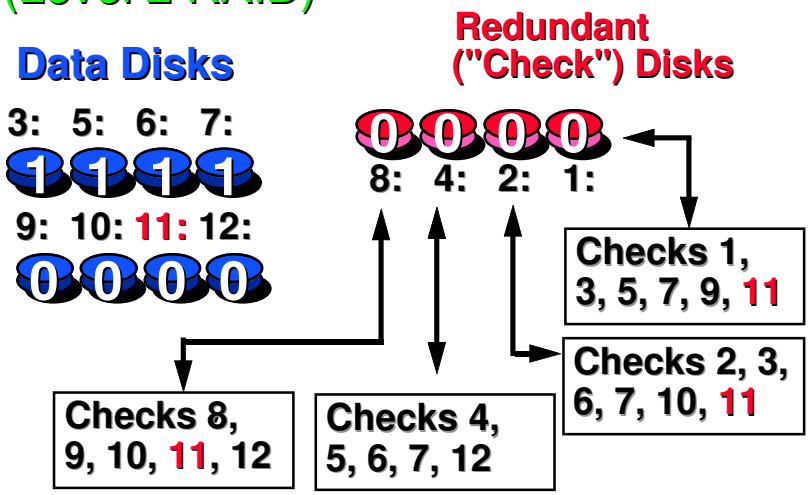
© Patterson 1994

Redundancy via Mirroring (Level 1 RAID)



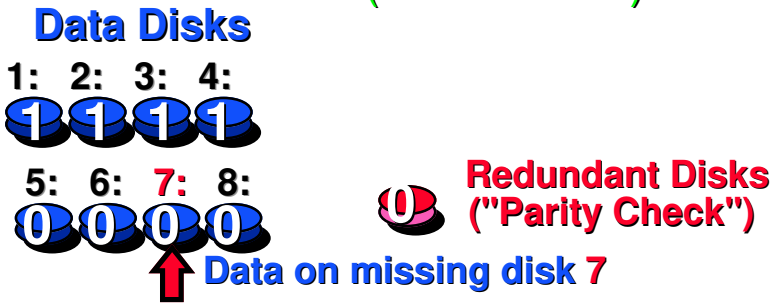
© Patterson 1994

Redundancy via ECC
(Level 2 RAID)



© Patterson 1994

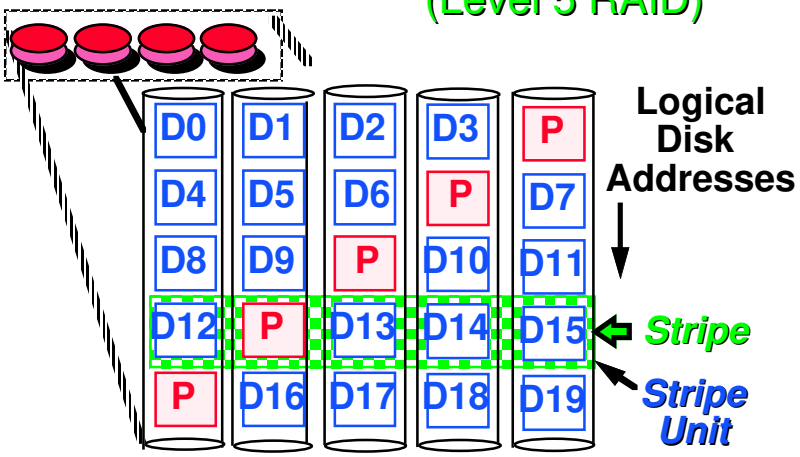
Redundancy via Parity
(Level 3 RAID)



Reed/Solomon code per sector identifies failed disk

© Patterson 1994

Redundancy via Rotated Parity
(Level 5 RAID)



© Patterson 1994