Lecture 23: I/O—Redundant Arrays of Inexpensive Disks

Professor Randy H. Katz

Computer Science 252

Spring 1996

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
- RAID
- ABCs of UNIX File Systems
- I/O Benchmarks
- Comparing UNIX File System Performance

Review: Busses

- **Bus: a shared communication link between subsystems.**
- **Disadvantage: a communication bottleneck, possibly limiting the maximum I/O throughput**
- **Bus speed is limited by physical factors**
- **Two generic types of busses: I/O and CPU**
- **Bus transaction: sending address & receiving or sending data**

Review: 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

Review: 1990 Bus Survey

	VME	FutureBus	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

Review: 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

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

Review: 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)

Network Attached Storage

Decreasing Disk Diameters

14" » 10" » 8" » 5.25" » 3.5" » 2.5" » 1.8" » 1.3" » . . .
high bandwidth disk systems based on arrays of disks

Network provides
well defined physical
and logical interfaces:
separate CPU and
storage system!

High Performance
Storage Service
on a High Speed
Network

Network File Services

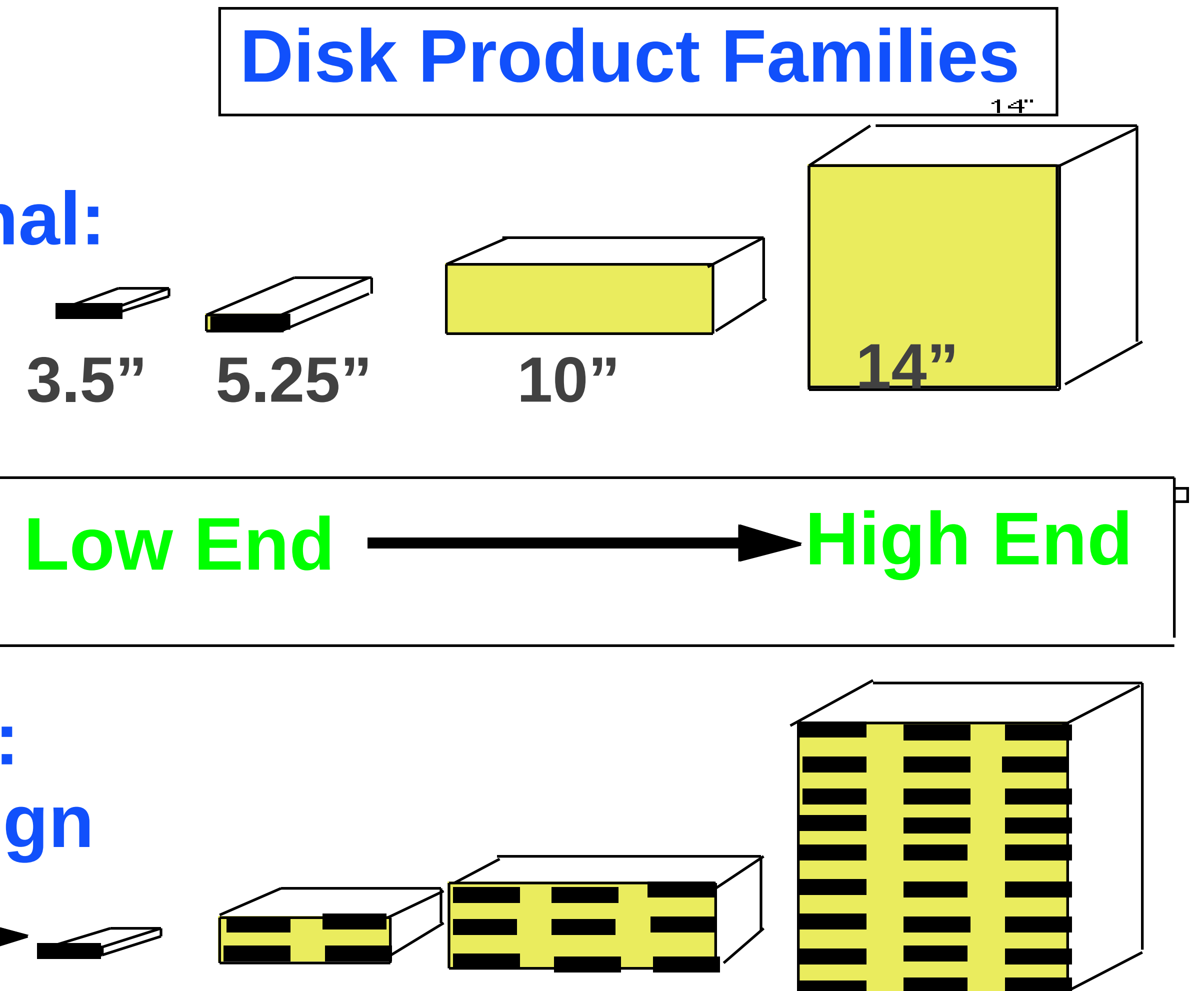
OS structures
supporting remote
file access

3 Mb/s » 10Mb/s » 50 Mb/s » 100 Mb/s » 1 Gb/s » *10 Gb/s*
networks capable of sustaining high bandwidth transfers

Increasing Network Bandwidth

Manufacturing Advantages of Disk Arrays

Conventional:
4 disk designs



Disk Array:
1 disk design

3.5”

Replace Small # of Large Disks with Large # of Small Disks!

	IBM 3390 (K)	IBM 3.5" 0061	*x70*
Data Capacity	20 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*
MTTF	250 KHrs	50 KHrs	*??? Hrs*
Cost	$250K	$2K	*$150K*

Disk Arrays have potential for

- **large data and I/O rates**
- **high MB per cu. ft., high MB per KW**
- **awful reliability**

Redundant Arrays of Disks

- **Files are "striped" across multiple spindles**
- **Redundancy yields high data availability**

Disks will fail

Contents reconstructed from data redundantly stored in the array

→ **Capacity penalty to store it**

→ **Bandwidth penalty to update**

Techniques:

- **Mirroring/Shadowing (high capacity cost)**
- **Horizontal Hamming Codes (overkill)**
- **Parity & Reed-Solomon Codes**
- **Failure Prediction (no capacity overhead!)**
  VaxSimPlus — Technique is controversial

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 without redundancy too unreliable to be useful!

Hot spares support reconstruction in parallel with access: very high media availability can be achieved

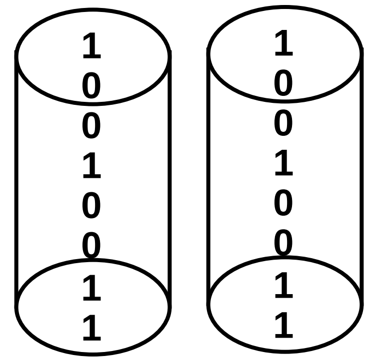
Redundant Arrays of Disks (RAID) Techniques

- ***Disk Mirroring, Shadowing***

  Each disk is fully duplicated onto its "shadow"

  Logical write = two physical writes

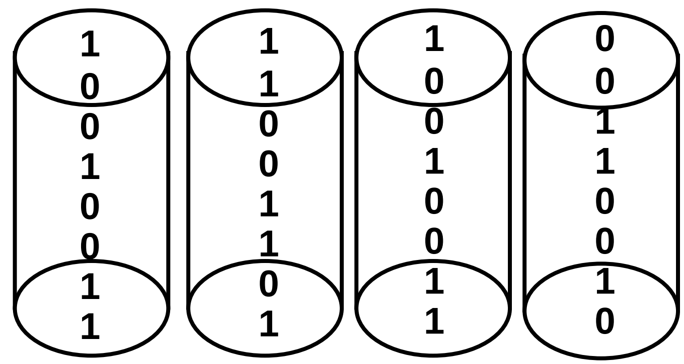
  100% capacity overhead



- ***Parity Data Bandwidth Array***

  Parity computed horizontally

  Logically a single high data bw disk



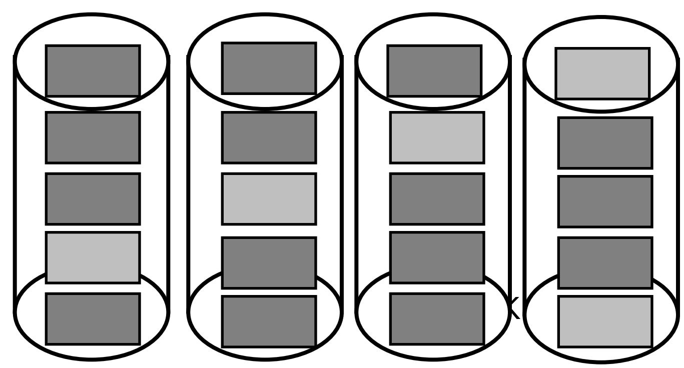
- ***High I/O Rate Parity Array***

  Interleaved parity blocks

  Independent reads and writes

  Logical write = 2 reads + 2 writes

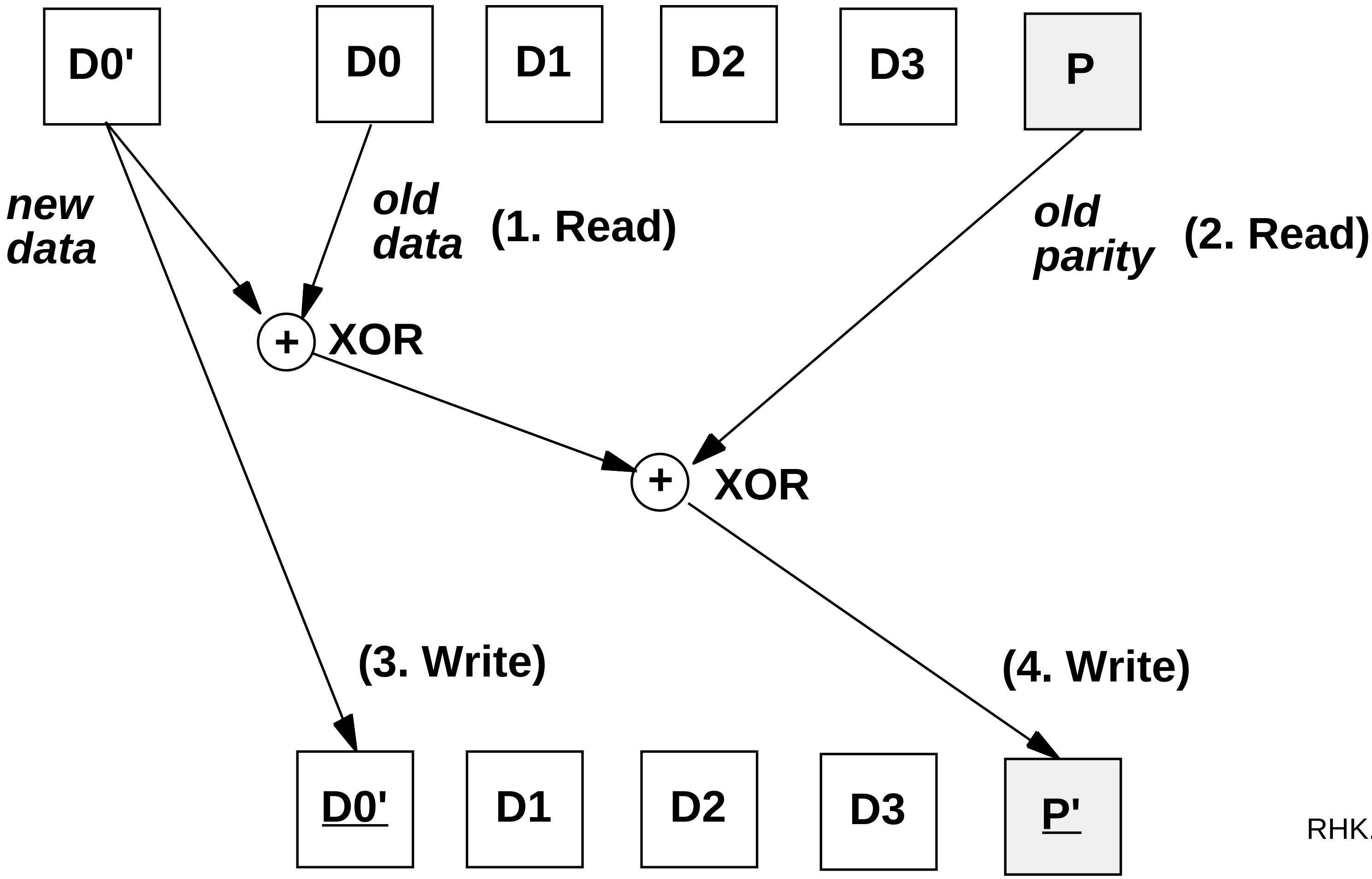
  Parity + Reed-Solomon codes



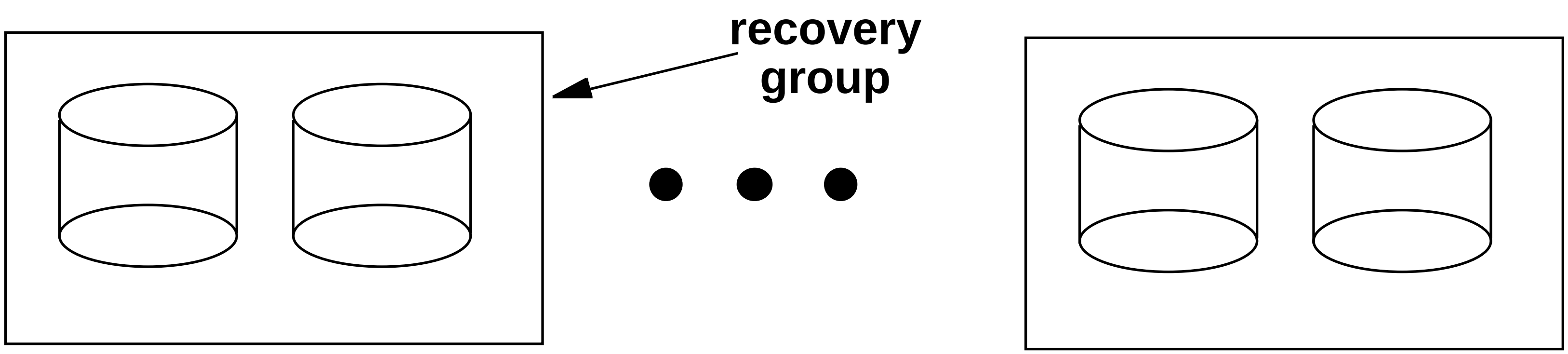
Problems of Disk Arrays: Small Writes

RAID-5: Small Write Algorithm

1 Logical Write = 2 Physical Reads + 2 Physical Writes



Redundant Arrays of Disks
RAID 1: Disk Mirroring/Shadowing



- **Each disk is fully duplicated onto its "shadow"**
  Very high availability can be achieved

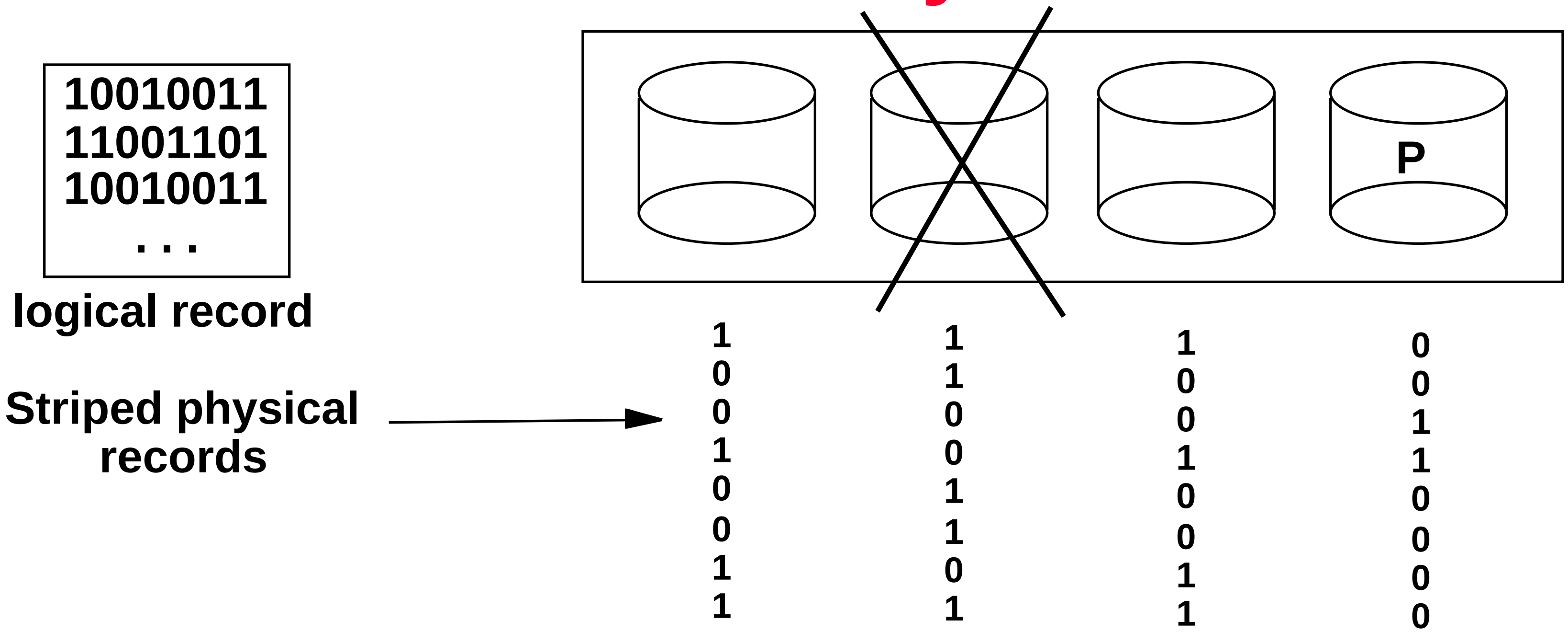
- **Bandwidth sacrifice on write:**
  Logical write = two physical writes

- **Reads may be optimized**

- **Most expensive solution: 100% capacity overhead**

Targeted for high I/O rate , high availability environments

Redundant Arrays of Disks
RAID 3: Parity Disk



- **Parity computed across recovery group to protect against hard disk failures**
  - **33% capacity cost for parity in this configuration**
  - **wider arrays reduce capacity costs, decrease expected availability, increase reconstruction time**

- **Arms logically synchronized, spindles rotationally synchronized**
  - **logically a single high capacity, high transfer rate disk**

Targeted for high bandwidth applications: Scientific, Image Processing

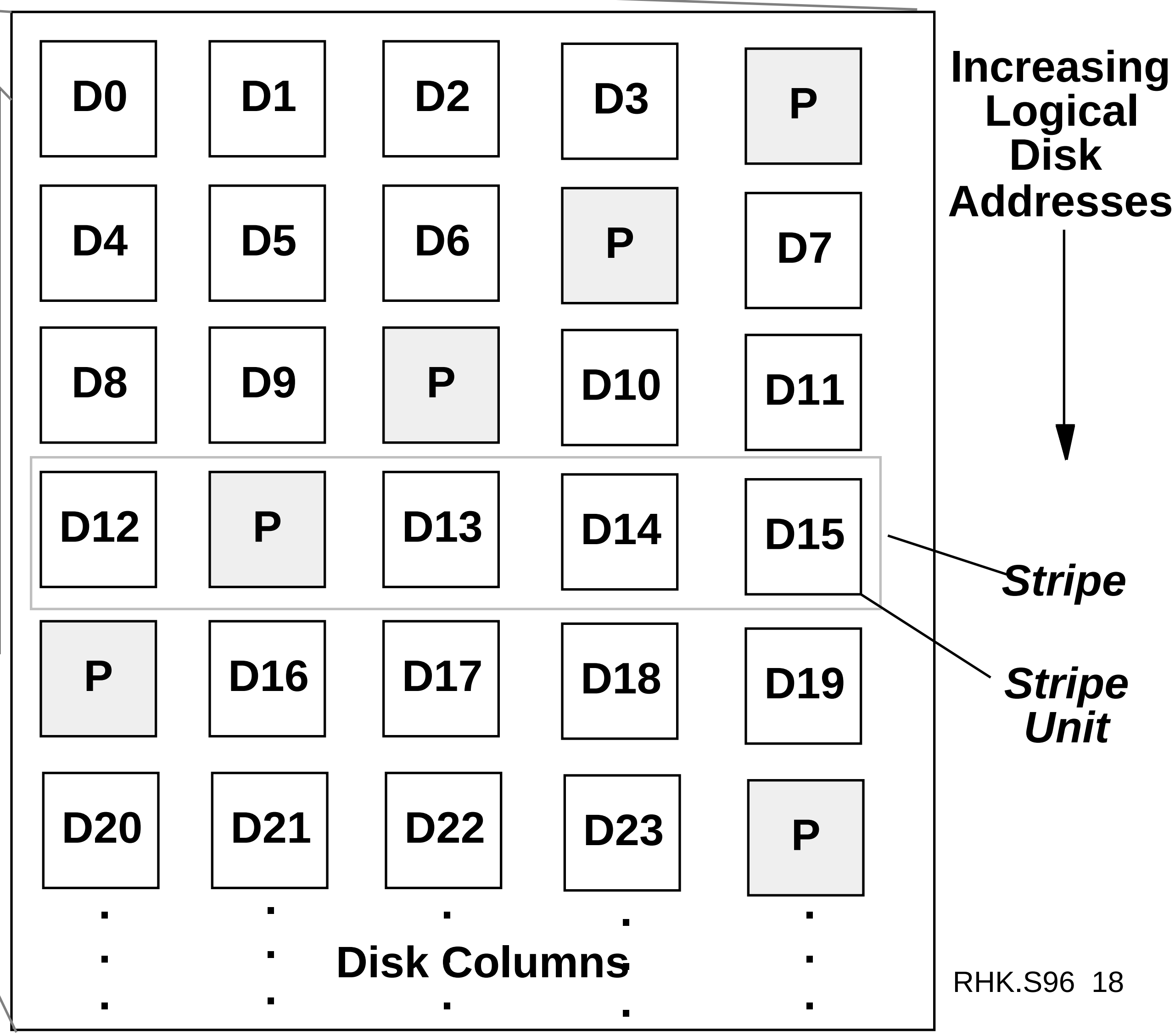
Redundant Arrays of Disks RAID 5+: High I/O Rate Parity

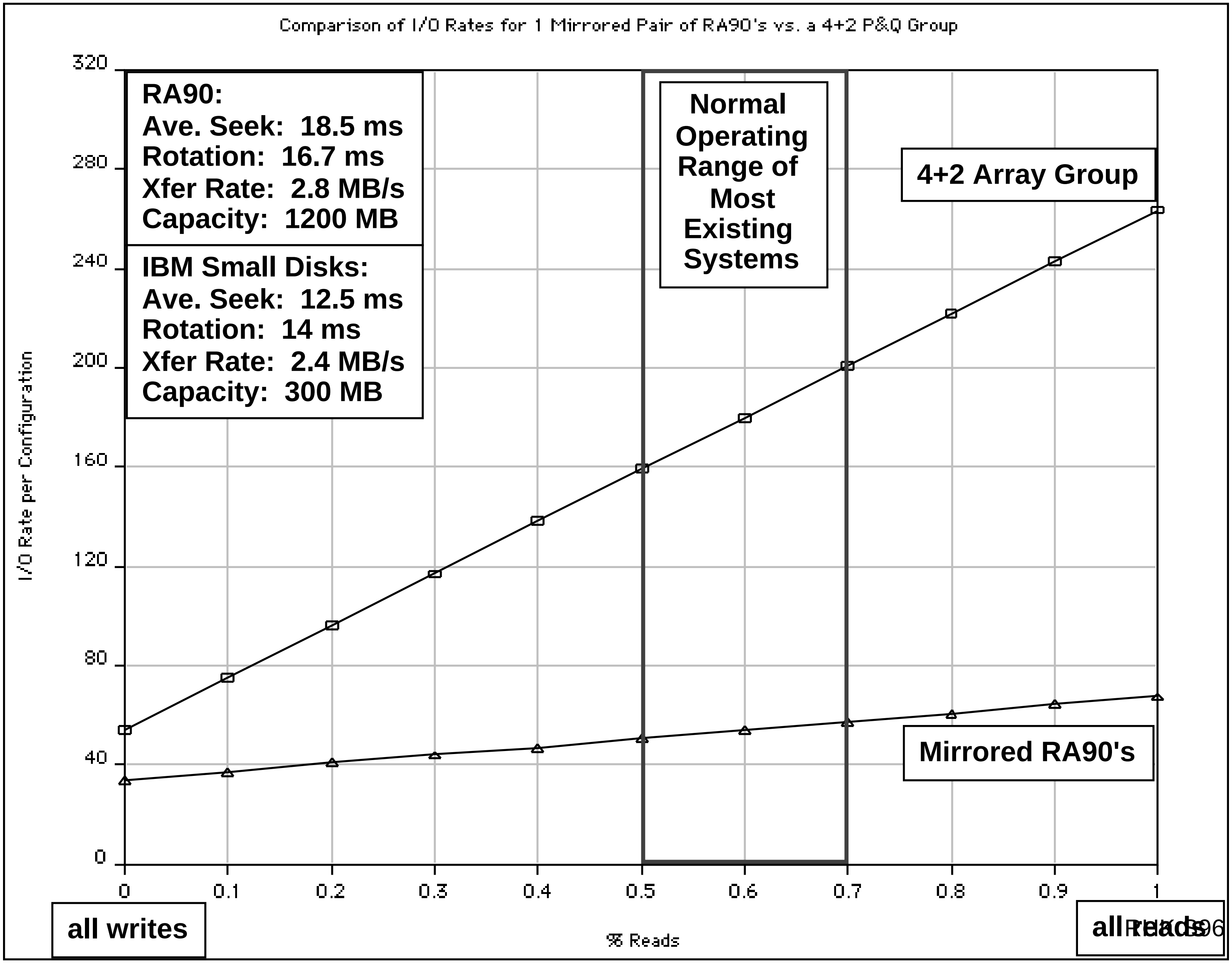
A logical write becomes four physical I/Os

Independent writes possible because of interleaved parity

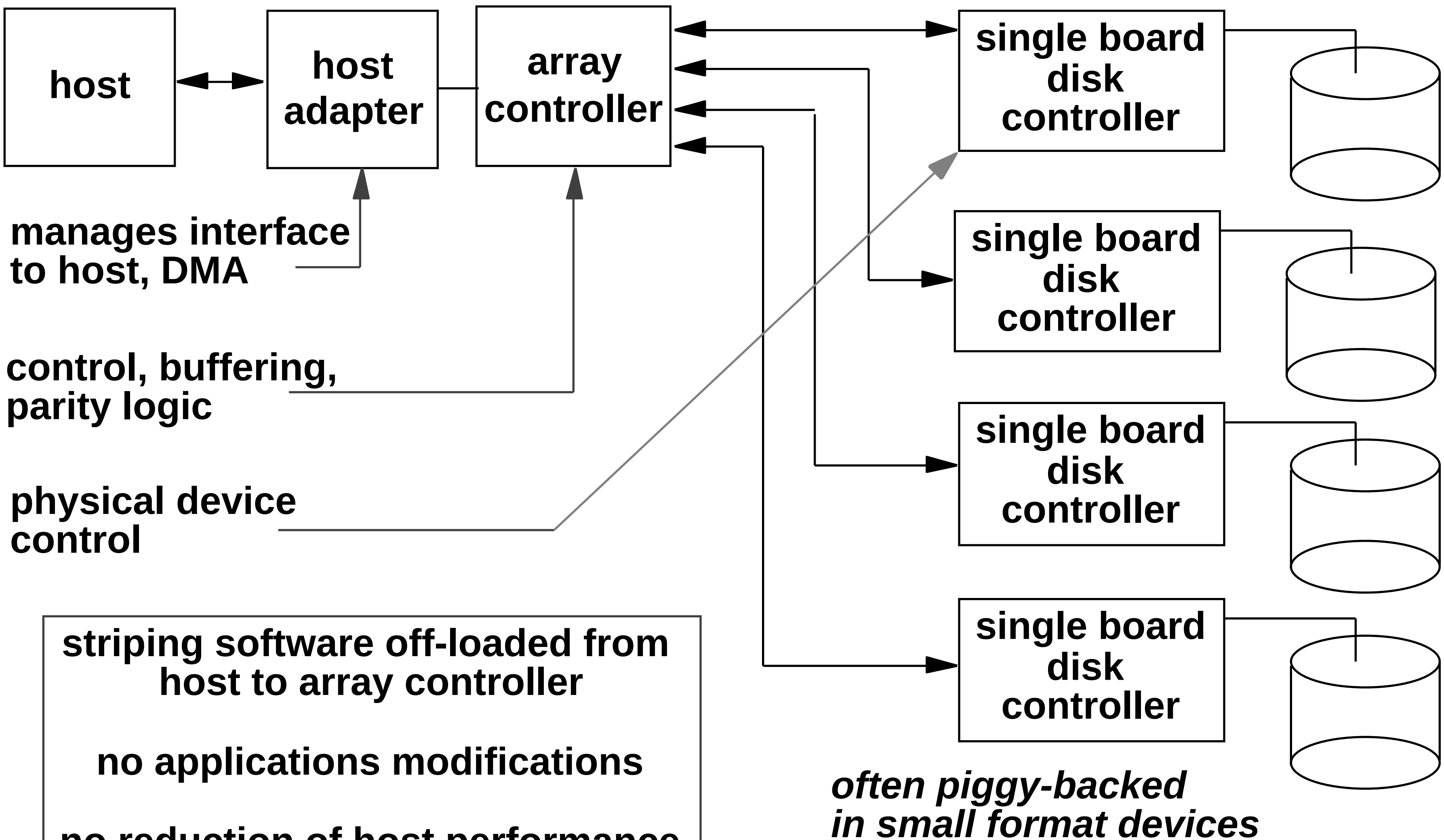
Reed-Solomon Codes ("Q") for protection during reconstruction

Targeted for mixed applications




Comparison of I/O Rates for 1 Mirrored Pair of RA90's vs. a 4+2 P&Q Group
RA90:
Ave. Seek: 18.5 ms
Rotation: 16.7 ms
Xfer Rate: 2.8 MB/s
Capacity: 1200 MB
IBM Small Disks:
Ave. Seek: 12.5 ms
Rotation: 14 ms
Xfer Rate: 2.4 MB/s
Capacity: 300 MB
Normal Operating Range of Most Existing Systems
4+2 Array Group
Mirrored RA90's
I/O Rate per Configuration
320
280
240
200
160
120
80
40
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
% Reads
all writes
all reads

Subsystem Organization

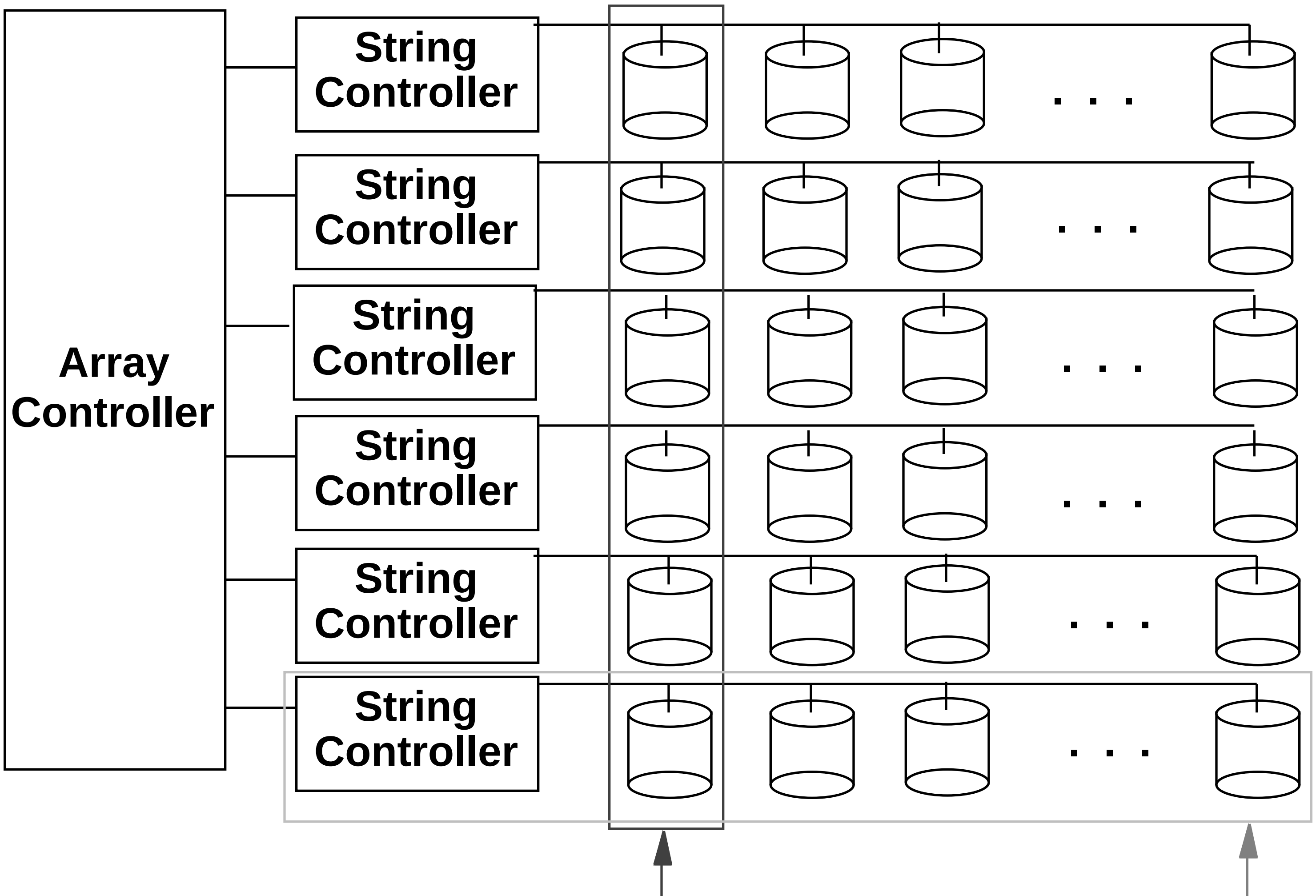


striping software off-loaded from host to array controller

no applications modifications

no reduction of host performance

System Availability: Orthogonal RAIDs

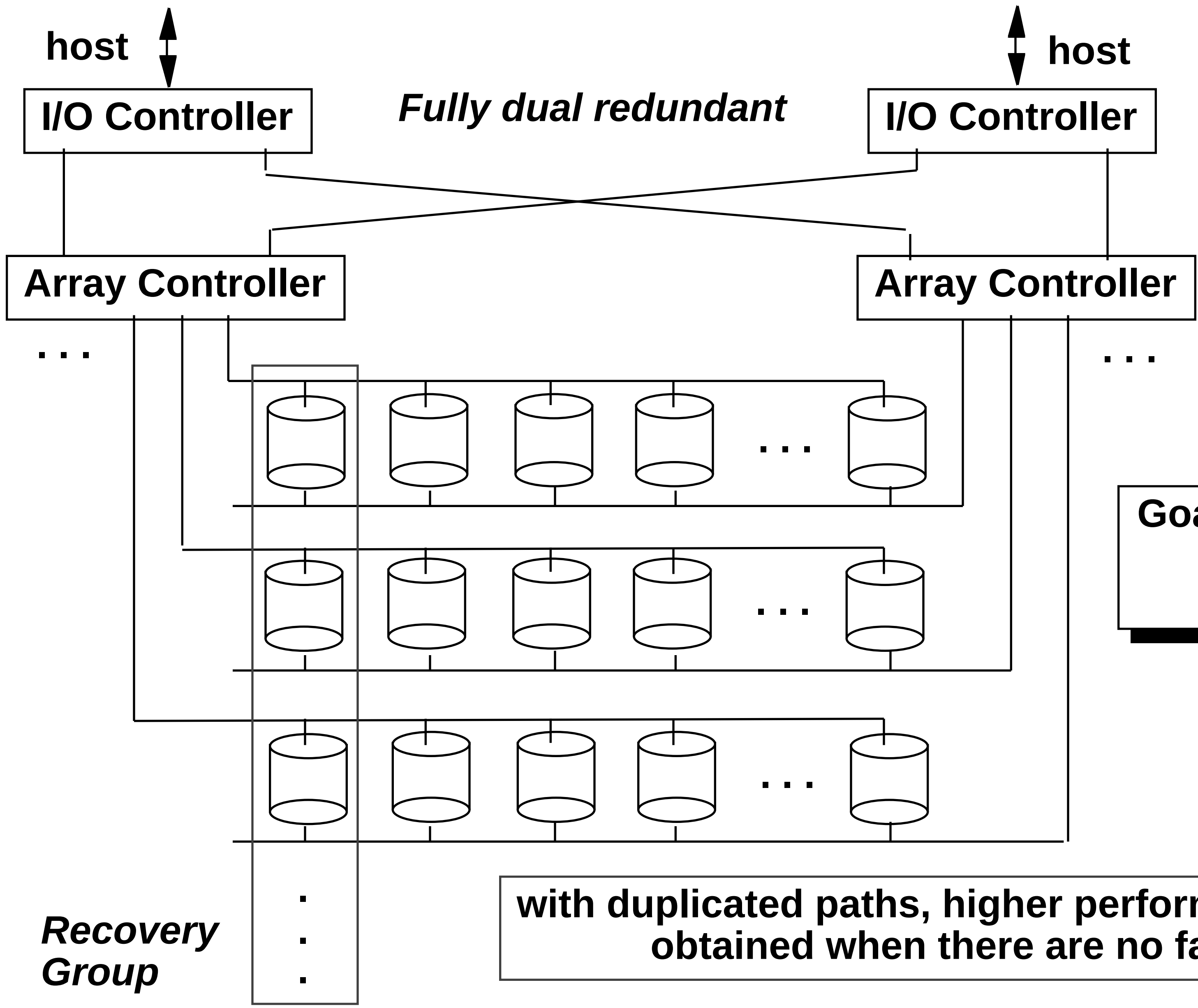


Data Recovery Group: **unit of data redundancy**

Redundant Support Components: **fans, power supplies, controller, cables**

End to End Data Integrity: **internal parity protected data paths**

System-Level Availability



Goal: No Single Points of Failure

with duplicated paths, higher performance can be obtained when there are no failures