15-712:
Advanced Operating Systems & Distributed Systems

A Case for Redundant Arrays of Inexpensive Disks (RAID)

Prof. Phillip Gibbons

Spring 2023, Lecture 10

Course Projects

- **Project Topics**
  - Prior project list on 15-712 webpage
  - Look to research you are already doing, if on topic for course
  - I can also suggest some projects – come talk to me
- **Friday 2/17 Deadline to form project teams (teams of 3)**
- **Friday 2/24: Day of meetings to discuss project ideas**
  - Each team comes with 2-3 ideas for project proposals
- **Wednesday 3/1: Project proposals due 11:59 pm**

Today's Papers

"A Case for Redundant Arrays of Inexpensive Disks (RAID)"

David A. Patterson, Garth Gibson, Randy H. Katz 1988

Optional Further Reading:

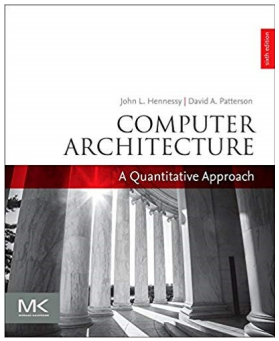
"Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?"

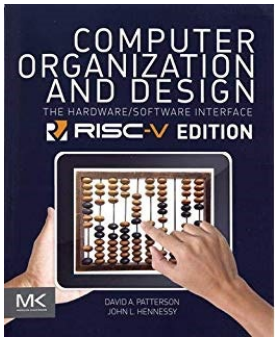
Bianca Schroeder, Garth A. Gibson 2007

"A Case for Redundant Arrays of Inexpensive Disks (RAID)"

David A. Patterson, Garth Gibson, Randy H. Katz 1988

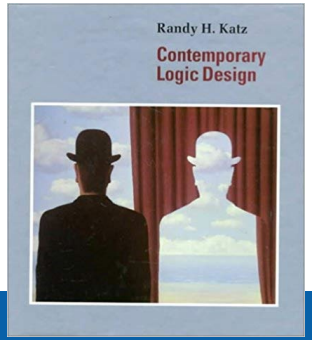
- **Dave Patterson** (UC Berkeley)
  - ACM Turing Award
  - Eckert-Mauchly Award
  - NAE, NAS, AAAS





- **Garth Gibson** (UC Berkeley PhD, CMU, Vector Institute)
  - Founder/CTO panasas, President/CEO Vector Inst.
  - IEEE Info. Storage Award, ACM Fellow, IEEE Fellow

- **Randy Katz** (UC Berkeley)
  - IEEE Info. Storage Award, NAE, AAAS
  - ACM Outstanding Educator Award



“A Case for Redundant Arrays of Inexpensive Disks (RAID)”

David A. Patterson, Garth Gibson, Randy H. Katz 1988

SigOps HoF citation (2011):

The paper shows how to achieve efficient, fault tolerant and highly available storage using cheap, unreliable components.

Birth of RAID

- CPUs are going along nicely
- But **Amdahl's Law** says CPU cycles wasted if disk doesn't keep up

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

S = speedup
f = frac work faster
k = how much faster

- e.g., if you make 30% of the system run 9x faster:

$$S = \frac{1}{(1-0.3)+0.3/9}$$ = speedup of only 1.36x (yikes!)

General Problem!

Balancing performance of components in a computer system == eternal challenge

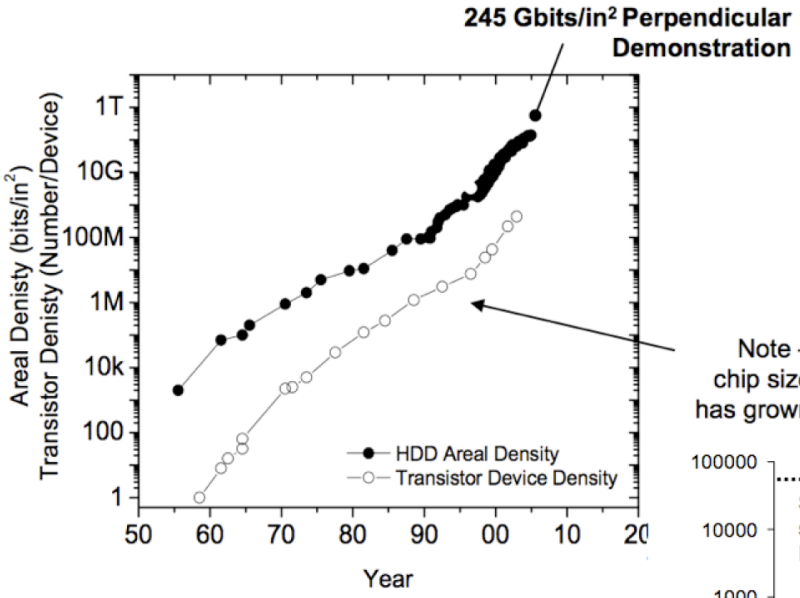
- CPU speed
- Memory cache speed (L1, L2, L3, ...)
- Bus speed
- Disk throughput
- Disk IO operations / sec
- Network throughput
- Network latency

Trying to substitute one for another == great fun, popular

- Transistors for memory speed: prediction...
- Spend local disk instead of network BW: Caching
- Spend network BW instead of local disk: RDMA

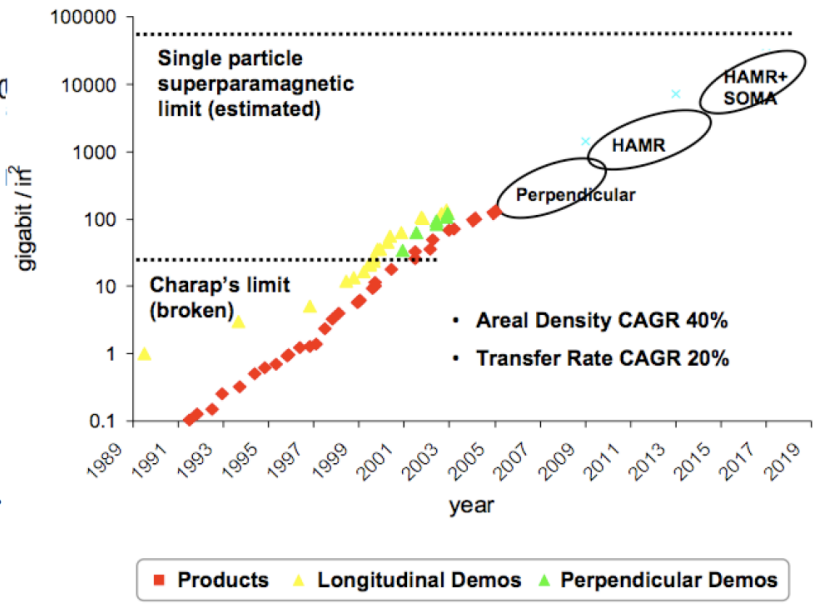
The Pending I/O Crisis

Areal Density vs. Moore's Law



Technology Trends

- Disk density tracks Moore's Law & no stalls in sight
- Transfer rate ~SQRT
- Random accesses ~5%/yr



Comparison of Disks (circa 1988)

Characteristics	IBM 3380	Fujitsu M2361A	Conners CP3100	3380 v. 3100	2361 v. 3100
	Mainframe	Minicomputer	PC	(>1 means 3100 better)	
Disk diameter (in)	14	10.5	3.5	4	3
Formatted Data Capacity (MB)	7500	600	100	.01	.2
Price/MB (cntl incl)	$18-$10	$20-$17	$10-$7	1-2.5	1.7-3
MTTF Rated (hrs)	30,000	20,000	30,000	1	1.5
MTTF in practice (hrs)	100,000	?	?	?	?
No. Actuators	4	1	1	2	1
Max IOs/s/Actuator	50	40	30	.6	.8
Typ IOs/s/Actuator	30	24	20	.7	.8
Max IOs/s/box	200	40	30	.2	.8
Typ IOs/s/box	120	24	20	.2	.8
Transfer Rate (MB/s)	3	2.5	1	3	4
Power/box (W)	6,600	640	10	660	64
Volume (cu ft)	24	3.4	.03	800	110

Disk Array of 75 PC disks has 12x I/O BW of IBM 3380, same capacity, lower power consumption, lower cost!

1988

Characteristics	IBM 3380	Fujitsu M2361A	Conners CP3100	3380 v. 3100 (>1 means 3100 better)	2361 v. 3100
Disk diameter (in)	14	10.5	3.5	4	
Formatted Data Capacity (MB)	7500	600	100	.01	
Price/MB (cntl incl)	$18-$10	$20-$17	$10-$7	1-2.5	
MTTF Rated (hrs)	30,000	20,000	30,000	1	
MTTF in practice (hrs)	100,000	?	?	?	
No. Actuators	4	1	1	2	
Max IOs/s/Actuator	50	40	30	.6	
Typ IOs/s/Actuator	30	24	20	.7	
Max IOs/s/box	200	40	30	.2	
Typ IOs/s/box	120	24	20	.2	
Transfer Rate (MB/s)	3	2.5	1	3	4
Power/box (W)	6,600	640	10	660	64
Volume (cu ft)	24	3.4	.03	800	110

2013 vs. 1988

- Capacity up 1000X
- Transfer rate up 100X
- MTTF up 30X
- IO/sec up 5X

Disk Specs (2005, 2009, 2013)

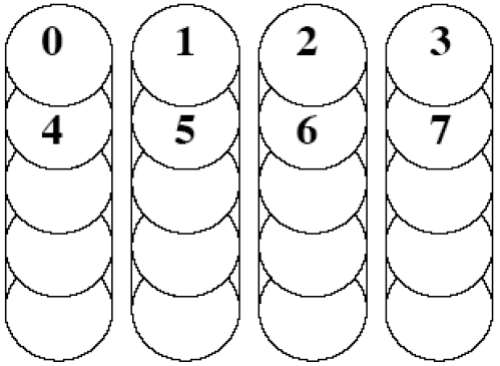
3.5 inch Nearline	2005 (Longitudinal)	2009 (Perpendicular)	2013 (HAMR)
Drive Capacity (GB)	500	2,000	8,000
Number of Discs	3	3	3
Capacity (GB/disc)	168	670	2,670
Product Areal Density (Gbpsi)	120	500	1,800
Transfer Rate (Mb/sec)	995	2,000	5,000
RPM	7,200	7,200	10,000
Read Seek Time (ms)	8	7.2	6.5
2.5 inch Enterprise	**2005**	**2009**	**2013**
Drive Capacity (GB)	75	300	1,000
Number of Discs	2	2	2
Capacity (GB/disc)	40	150	500
Product Areal Density (Gbpsi)	70	300	1,000
Transfer Rate (Mb/sec)	750	2,000	4,000
RPM	10,000	15,000	15,000
Read Seek Time (ms)	4.7	3.8	3.1

Striping for Read Throughput

Goals:

- **Load balance high-concurrency, small accesses across disks**
- **Enable parallel transfers for low-concurrency large reads**

Striping to the rescue!



- **Uniform load for small reads**
  - If striping unit contains the whole object (e.g., small read is contained on one disk)
- **Parallelism for large reads**
  - Stripe unit small enough to spread read across many disks

And Now The Bad News

MTTF of Disk Array = MTTF of Single Disk / Number of Disks in Array

100 CP 3100 disks has MTTF = 300 hours
vs. 3 years for IBM 3380: This is dismal

1000 CP 3100 disks has MTTF = 30 hours,
requiring an adjective worse than dismal

Adding the R to RAID

D = total number of data disks

G = number of data disks in a group

C = number of check disks in a group

n_G = D/G = number of groups

$$MTTF_{Group} = \frac{MTTF_{Disk}}{G + C} \times \frac{MTTF_{Disk}}{(G + C - 1)MTTR}$$

$$MTTF_{RAID} = MTTF_{Group}/n_G$$

RAID Level 1: Mirrored Disks

D = total number of data disks

G = number of data disks in a group = 1

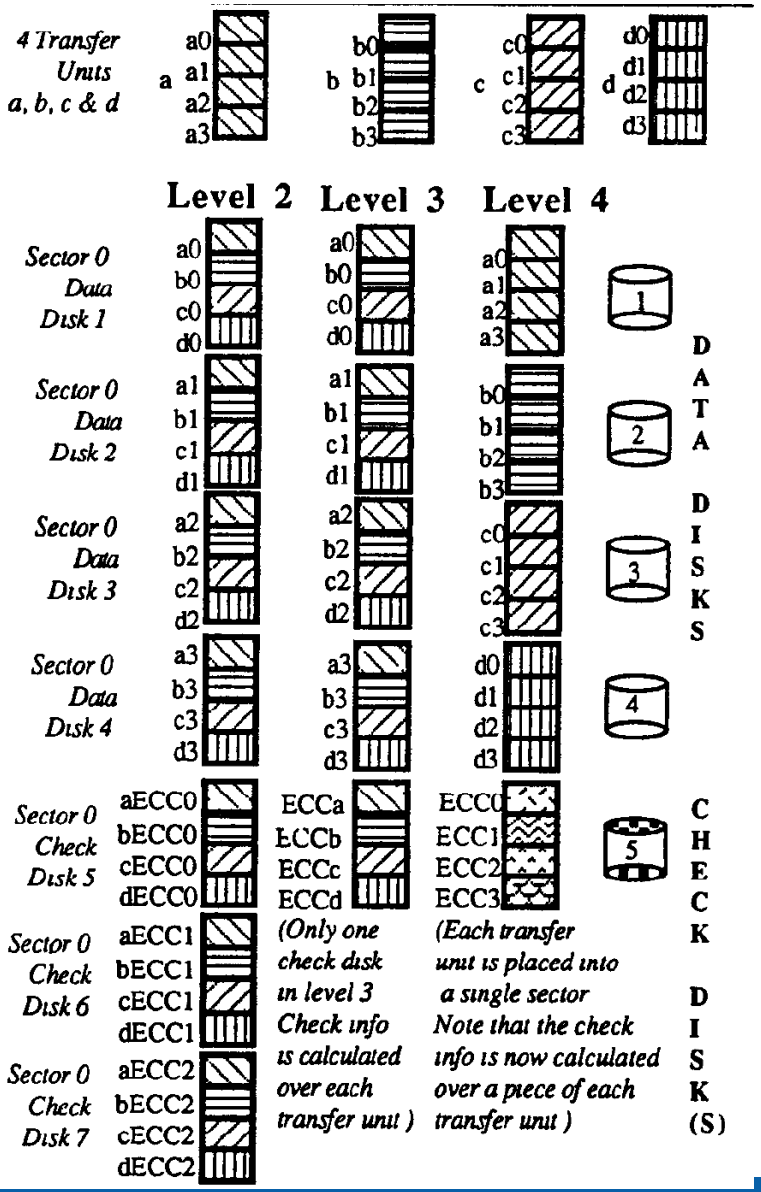
C = number of check disks in a group = 1

n_G = D/G = number of groups

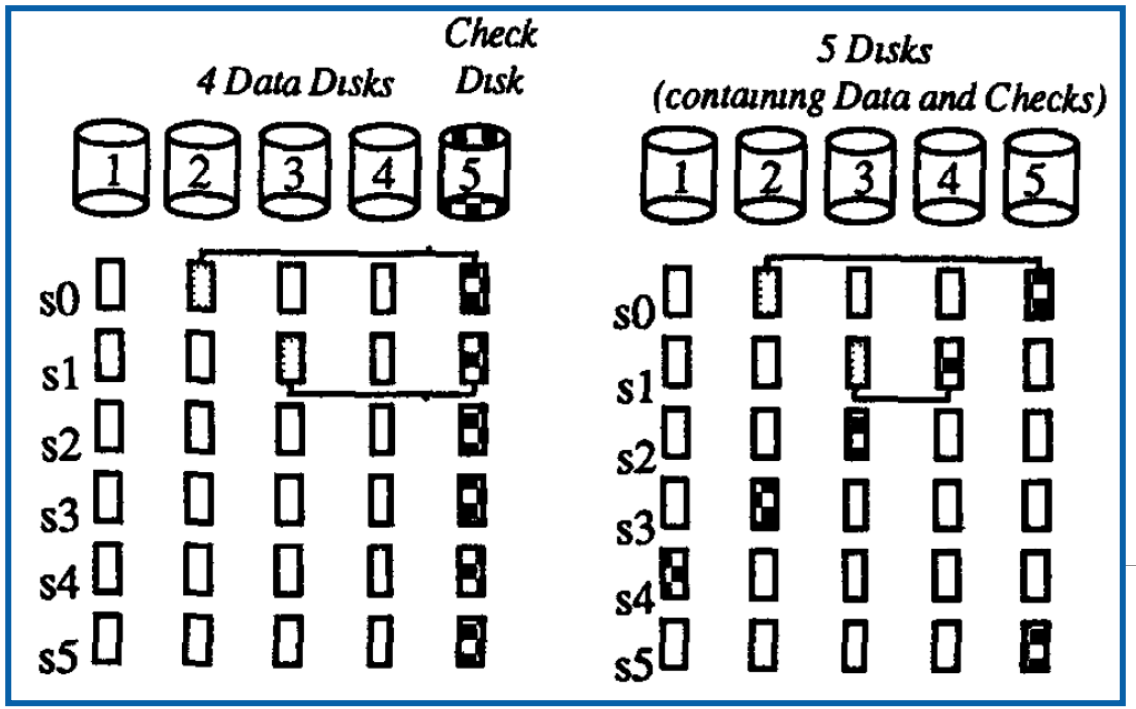
$$MTTF_{Group} = \frac{MTTF_{Disk}}{G + C} \times \frac{MTTF_{Disk}}{(G + C - 1)MTTR}$$

$$MTTF_{RAID} = MTTF_{Group}/n_G$$

RAID Levels 2-4



RAID Level 5



	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*			*Efficiency Per Disk*		
		L5	*L5/L4*	*L5/L1*	*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%

D = data disks

G = data disks/group

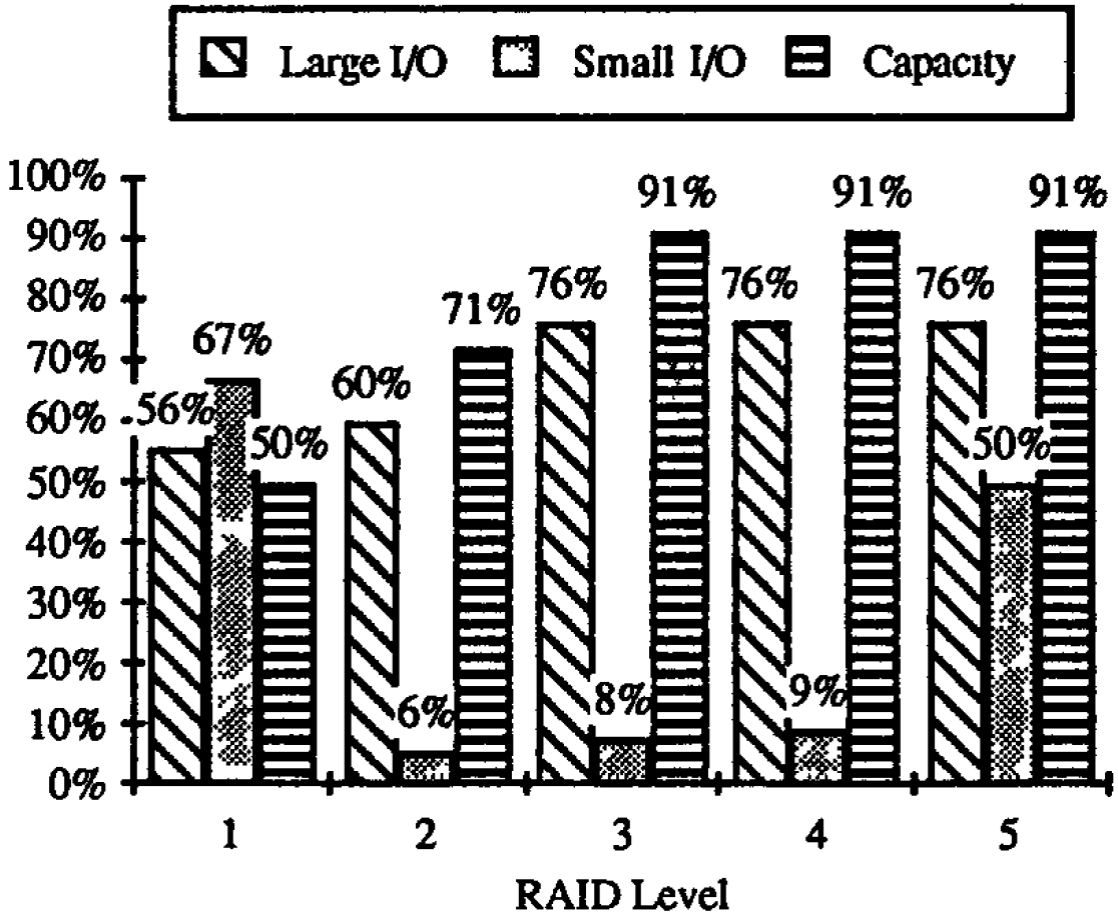
C = check disks/group

Discussion: Summary Question #1

- **State the 3 most important things the paper says.** These could be some combination of their motivations, observations, interesting parts of the design, or clever parts of their implementation.

RAID Level 1-5 Comparison

RMW Efficiency/Disk & Capacity



D=100

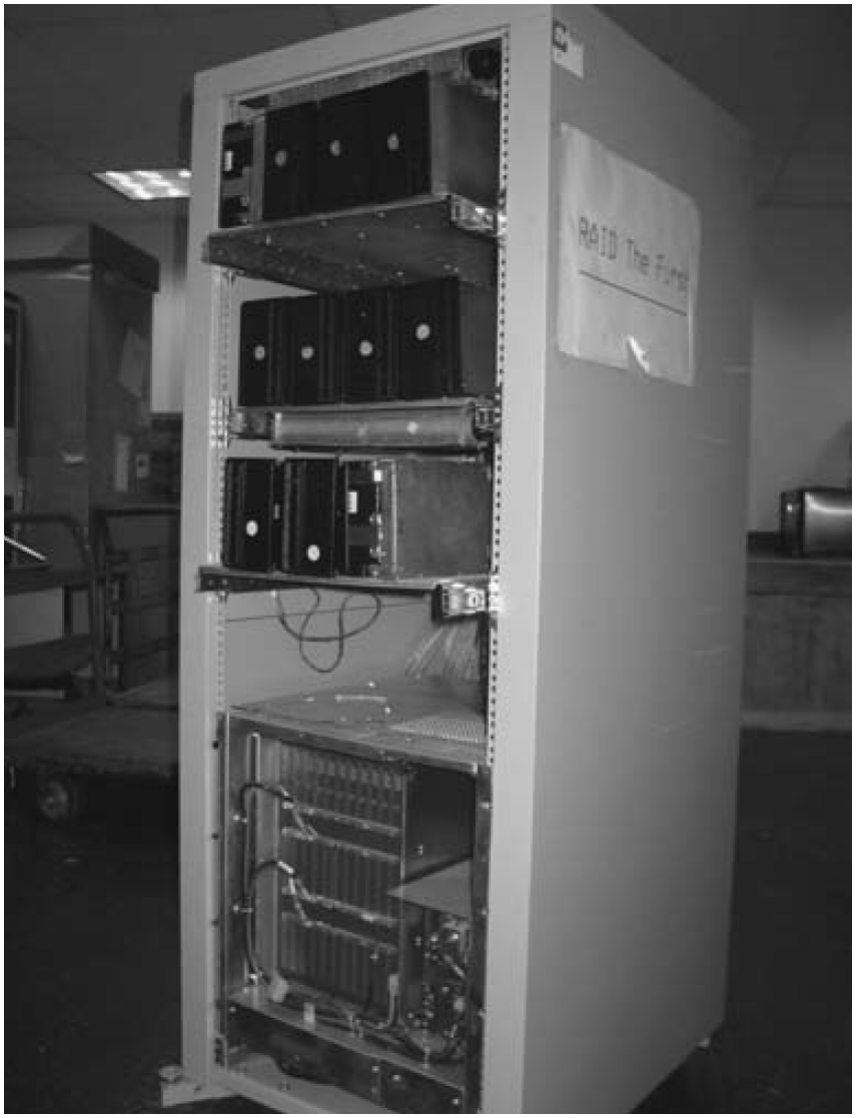
G=10

S=1.3

RAID 5 vs. IBM 3380

Characteristics	*RAID 5L (100 10) (CP3100)*	*SLED (IBM 3380)*	*RAID v SLED (>1 better for RAID)*
Formatted Data Capacity (MB)	10,000	7,500	1.33
Price/MB (controller incl.)	$11-$8	$18-$10	2.2-.9
Rated MTTF (hours)	820,000	30,000	27.3
MTTF in practice (hours)	?	100,000	?
No. Actuators	110	4	22.5
Max I/O's/Actuator	30	50	.6
Max Grouped RMW/box	1250	100	12.5
Max Individual RMW/box	825	100	8.2
Typ I/O's/Actuator	20	30	.7
Typ Grouped RMW/box	833	60	13.9
Typ Individual RMW/box	550	60	9.2
Volume/Box (cubic feet)	10	24	2.4
Power/box (W)	1100	6,600	6.0
Min Expansion Size (MB)	100-1000	7,500	7.5-75

First & Second RAID Prototypes



Discussion: Summary Question #2

- **Describe the paper's single most glaring deficiency.** Every paper has some fault. Perhaps an experiment was poorly designed or the main idea had a narrow scope or applicability.

"Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?"

Bianca Schroeder, Garth A. Gibson 2007

- **Bianca Schroeder** (CMU PhD/post-doc, Toronto)
  - Outstanding Young Canadian Comp. Sci. Prize
  - PC Chair for Systor 2020, Usenix FAST'14, ACM Sigmetrics'14, IEEE NAS'11

- **Garth Gibson** (CMU, Vector Institute)

FAST'07 Best Paper

USENIX Test of Time Award 2019

Seven Disk Failure Data Sets

Data set	Type of cluster	Duration	#Disk events	# Servers	Disk Count	Disk Parameters	MTTF (Mhours)	Date of first Deploym.	ARR (%)
HPC1	HPC	08/01 - 05/06	474	765	2,318	18GB 10K SCSI	1.2	08/01	4.0
			124	64	1,088	36GB 10K SCSI	1.2		2.2
HPC2	HPC	01/04 - 07/06	14	256	520	36GB 10K SCSI	1.2	12/01	1.1
HPC3	HPC	12/05 - 11/06	103	1,532	3,064	146GB 15K SCSI	1.5	08/05	3.7
	HPC	12/05 - 11/06	4	N/A	144	73GB 15K SCSI	1.5		3.0
	HPC	12/05 - 08/06	253	N/A	11,000	250GB 7.2K SATA	1.0		3.3
HPC4	Various	09/03 - 08/06	269	N/A	8,430	250GB SATA	1.0	09/03	2.2
	HPC	11/05 - 08/06	7	N/A	2,030	500GB SATA	1.0	11/05	0.5
	clusters	09/05 - 08/06	9	N/A	3,158	400GB SATA	1.0	09/05	0.8
COM1	Int. serv.	May 2006	84	N/A	26,734	10K SCSI	1.0	2001	2.8
COM2	Int. serv.	09/04 - 04/06	506	9,232	39,039	15K SCSI	1.2	2004	3.1
COM3	Int. serv.	01/05 - 12/05	2	N/A	56	10K FC	1.2	N/A	3.6
			132	N/A	2,450	10K FC	1.2	N/A	5.4
			108	N/A	796	10K FC	1.2	N/A	13.6
			104	N/A	432	10K FC	1.2	1998	24.1

From disk replacement logs in production systems

ARR = Annual replacement rate

Caveat: 43% of returned disks have no problems [Seagate]

Disks vs. Other HW Components

HPC1	
Component	%
CPU	44
Memory	29
Hard drive	**16**
PCI motherboard	9
Power supply	2

Outages attributed to HW

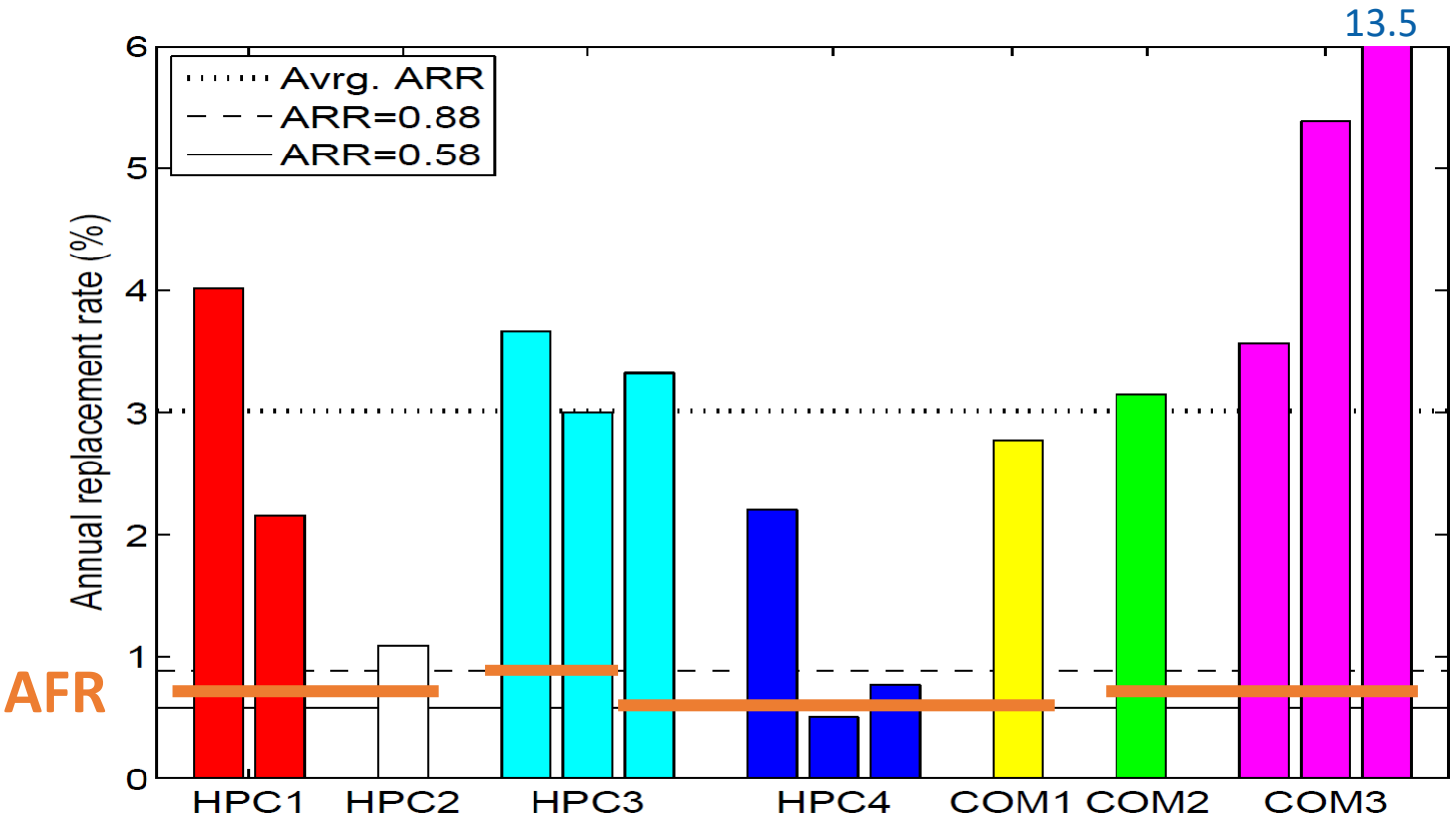
Relative frequency of HW component replacement

HPC1	
Component	%
Hard drive	**30.6**
Memory	28.5
Misc/Unk	14.4
CPU	12.4
PCI motherboard	4.9
Controller	2.9
QSW	1.7
Power supply	1.6
MLB	1.0
SCSI BP	0.3

COM1	
Component	%
Power supply	34.8
Memory	20.1
Hard drive	**18.1**
Case	11.4
Fan	8.0
CPU	2.0
SCSI Board	0.6
NIC Card	1.2
LV Power Board	0.6
CPU heatsink	0.6

COM2	
Component	%
Hard drive	**49.1**
Motherboard	23.4
Power supply	10.1
RAID card	4.1
Memory	3.4
SCSI cable	2.2
Fan	2.2
CPU	2.2
CD-ROM	0.6
Raid Controller	0.6

Datasheet AFRs vs. Observed ARRs



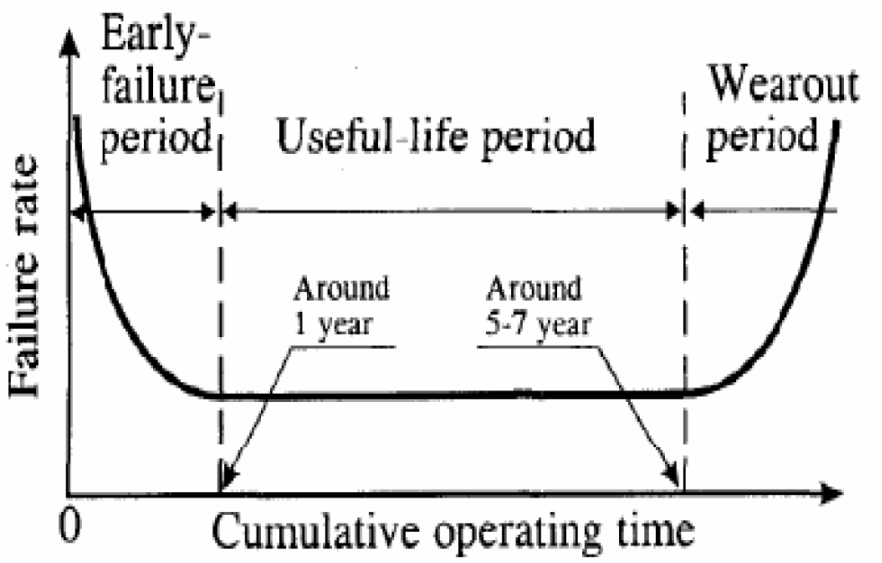
Obs 1: Avg. ARRs > 3.4x higher than datasheet MTTFs

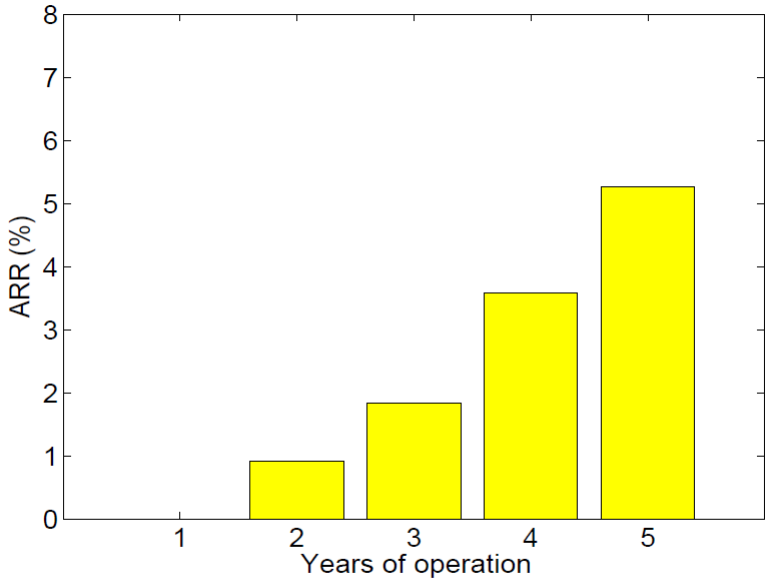
Obs 2: ARRs for age 5-8 upto 30x higher than datasheet MTTFs

Obs 3: ARRs for age < 3 upto 6x higher than datasheet MTTFs

Obs 4: ARRs for SATA disks not worse than SCSI/FC disks

Age-dependent Replacement Rates



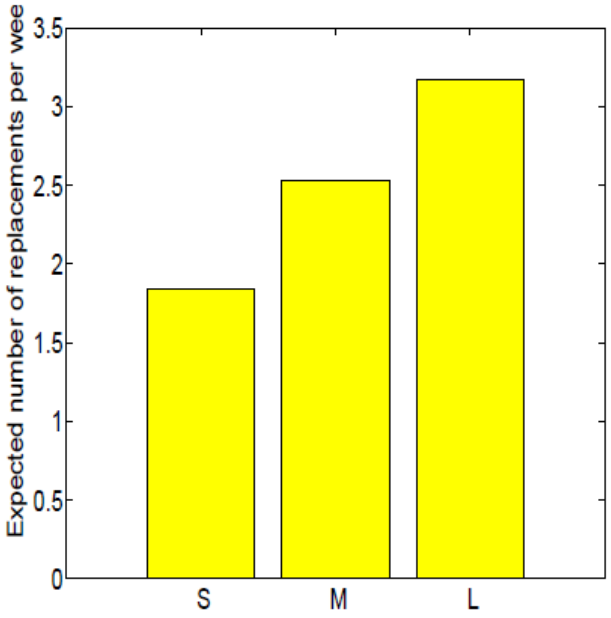


HPC1 (filesystem nodes)

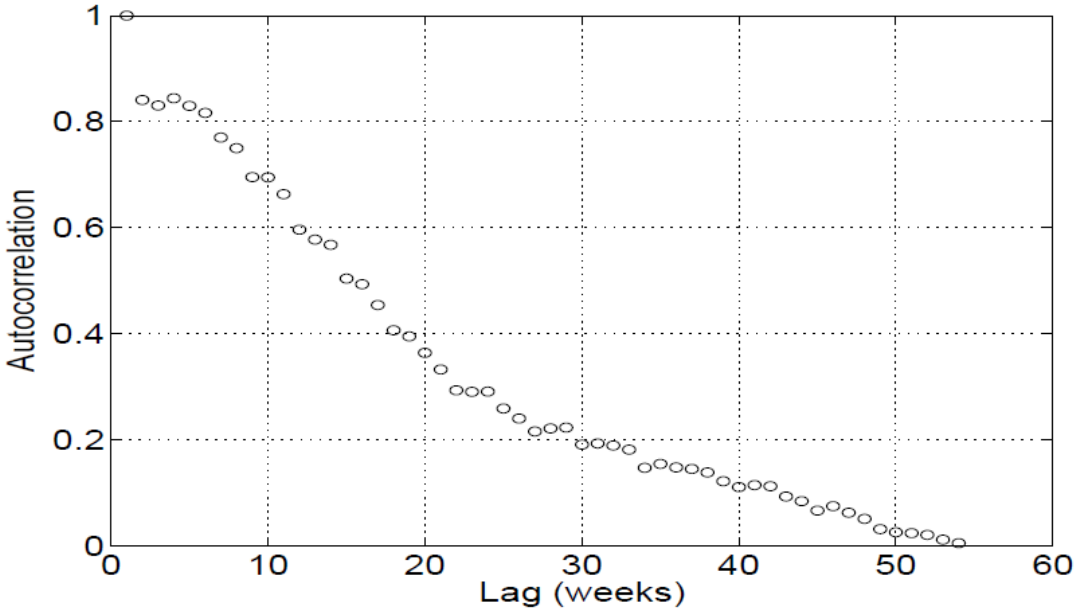
Obs 5: Contrary to conventional wisdom, replacement rates steadily increase over time (no bottom of the bathtub)

Obs 6: Early onset of wear-out is significant, infant mortality is not

Correlation of Disk Replacements



Year 3



Year 3

Obs 7 & 8: Disk replacement counts exhibit significant levels of autocorrelation & long-range dependence

Distribution of Time Between Failures

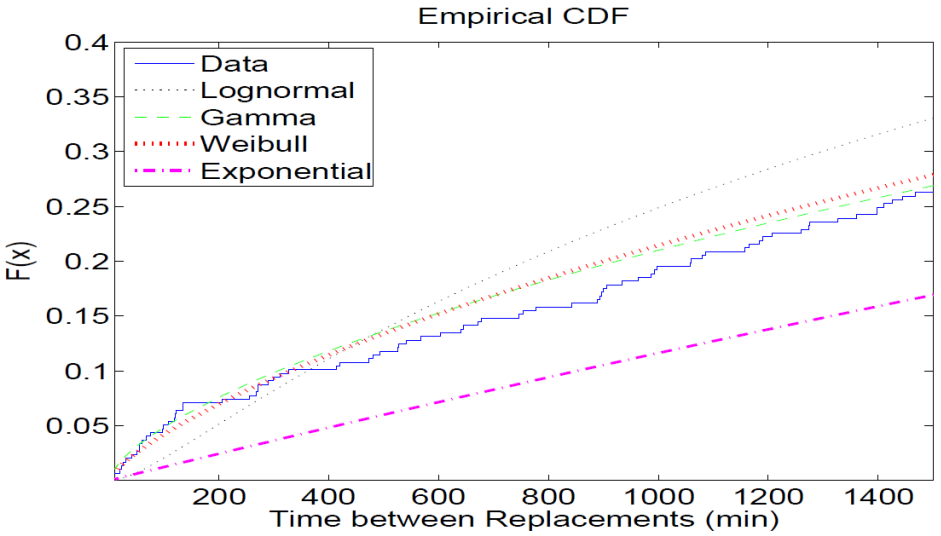


Figure 8: *Distribution of time*

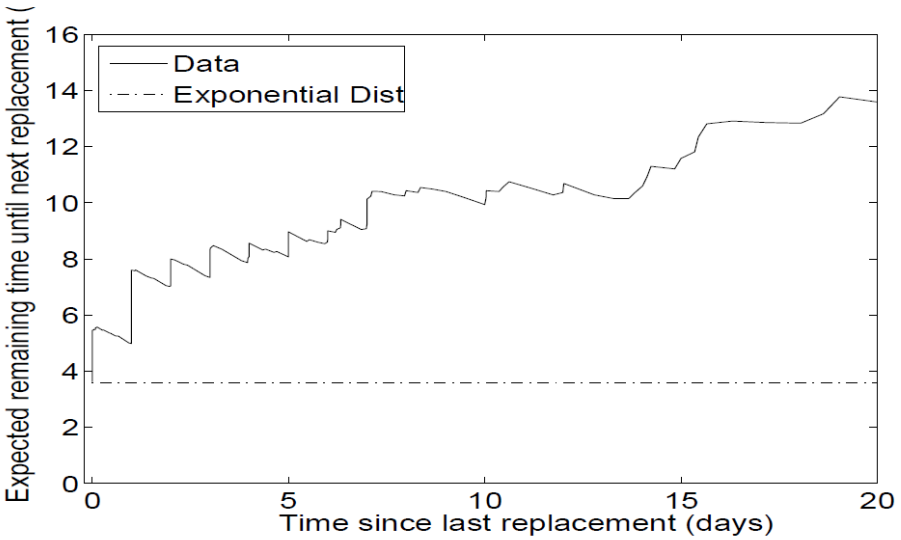


Figure 10: *Illustration of decreasing hazard rates*

Obs 9 & 10: Time between disk replacements does NOT follow an exponential distribution & it has a higher variability

Obs 11: Expected remaining time until the next disk was replaced grows with the time since the last replacement

Discussion: Summary Question #3

For Both Papers

- **Describe what conclusion you draw from the paper as to how to build systems in the future.** Most of the assigned papers are significant to the systems community and have had some lasting impact on the area.

Monday's Paper

Transactions and Databases (I)

"On Optimistic Methods for Concurrency Control"

H. T. Kung, John T. Robinson 1981