6.033 Spring 2017

Lecture #14

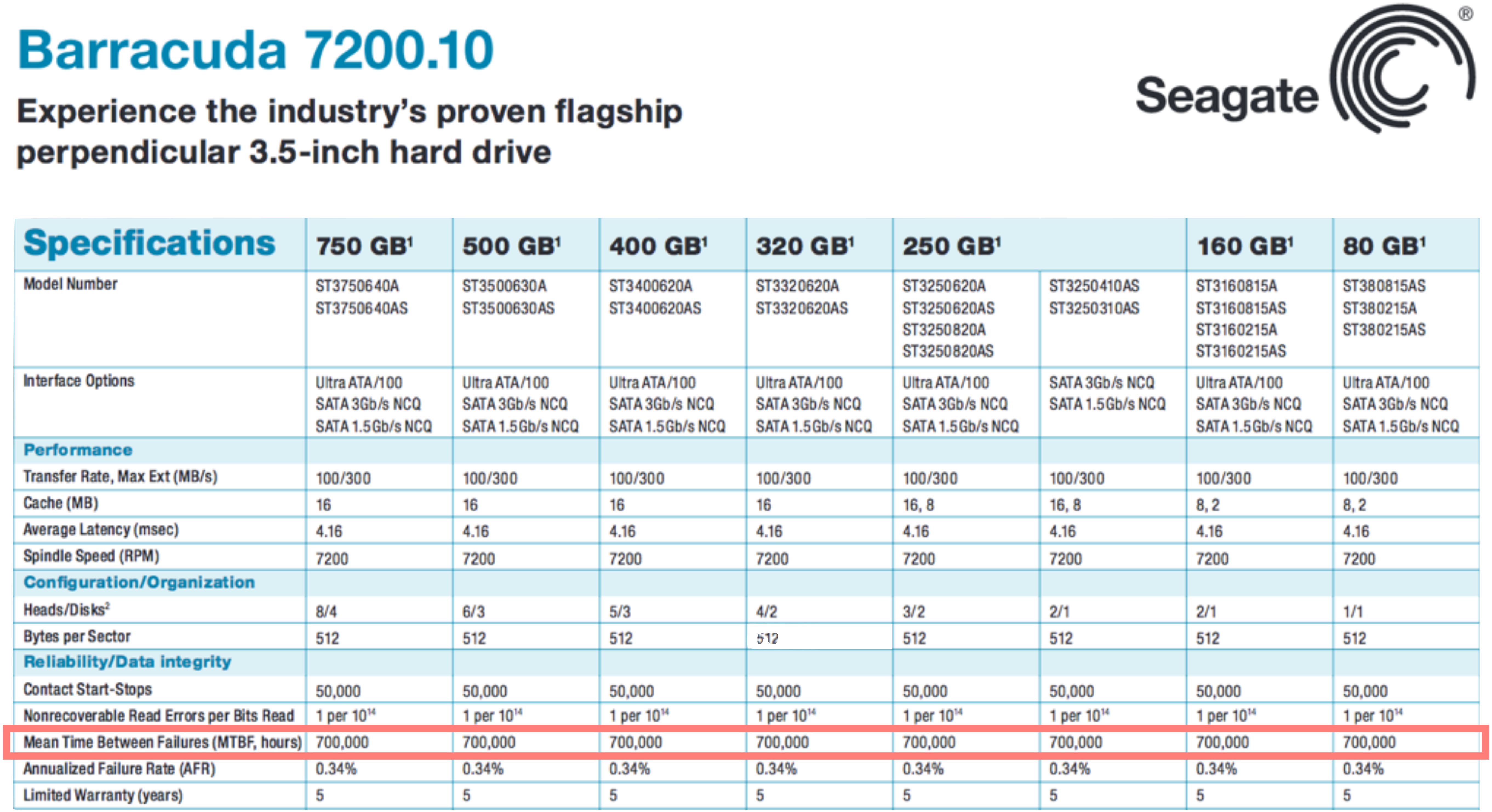
- **Reliability via Replication**
  - **General approach to building fault-tolerance systems**
  - **Single-disk failures: RAID**

How to Design Fault-tolerant Systems in Three Easy Steps

1. identify all possible faults
2. detect and contain the faults
3. handle the fault

quantifying reliability

dealing with disk failures



Barracuda 7200.10

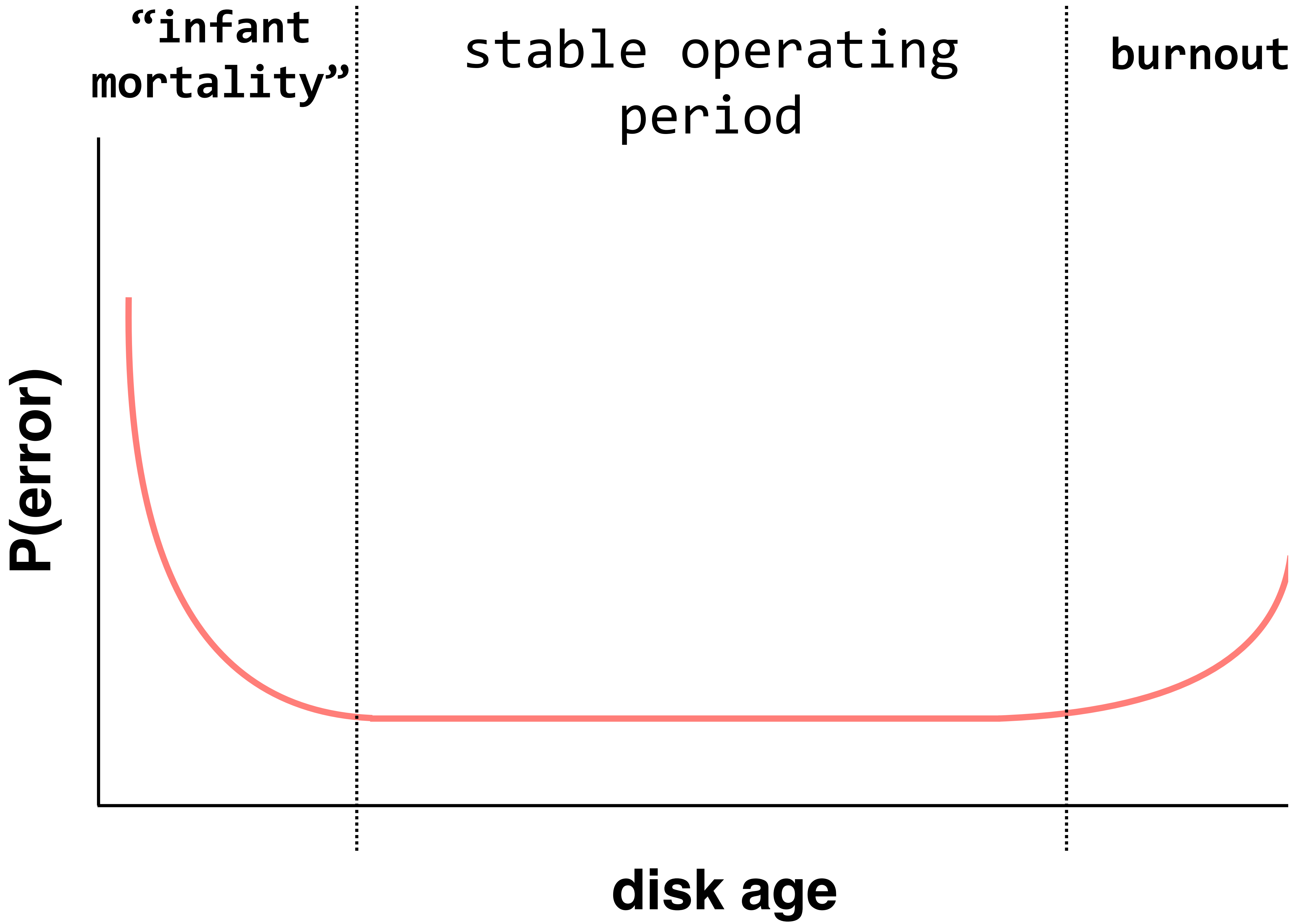
Experience the industry's proven flagship perpendicular 3.5-inch hard drive

Seagate®

Specifications	750 GB[1]	500 GB[1]	400 GB[1]	320 GB[1]	250 GB[1]		160 GB[1]	80 GB[1]
Model Number	ST3750640A ST3750640AS	ST3500630A ST3500630AS	ST3400620A ST3400620AS	ST3320620A ST3320620AS	ST3250620A ST3250620AS ST3250820A ST3250820AS	ST3250410AS ST3250310AS	ST3160815A ST3160815AS ST3160215A ST3160215AS	ST380815AS ST380215A ST380215AS
Interface Options	Ultra ATA/100 SATA 3Gb/s NCQ SATA 1.5Gb/s NCQ	Ultra ATA/100 SATA 3Gb/s NCQ SATA 1.5Gb/s NCQ	Ultra ATA/100 SATA 3Gb/s NCQ SATA 1.5Gb/s NCQ	Ultra ATA/100 SATA 3Gb/s NCQ SATA 1.5Gb/s NCQ	Ultra ATA/100 SATA 3Gb/s NCQ SATA 1.5Gb/s NCQ	SATA 3Gb/s NCQ SATA 1.5Gb/s NCQ	Ultra ATA/100 SATA 3Gb/s NCQ SATA 1.5Gb/s NCQ	Ultra ATA/100 SATA 3Gb/s NCQ SATA 1.5Gb/s NCQ
Performance								
Transfer Rate, Max Ext (MB/s)	100/300	100/300	100/300	100/300	100/300	100/300	100/300	100/300
Cache (MB)	16	16	16	16	16, 8	16, 8	8, 2	8, 2
Average Latency (msec)	4.16	4.16	4.16	4.16	4.16	4.16	4.16	4.16
Spindle Speed (RPM)	7200	7200	7200	7200	7200	7200	7200	7200
Configuration/Organization								
Heads/Disks[2]	8/4	6/3	5/3	4/2	3/2	2/1	2/1	1/1
Bytes per Sector	512	512	512	512	512	512	512	512
Reliability/Data integrity								
Contact Start-Stops	50,000	50,000	50,000	50,000	50,000	50,000	50,000	50,000
Nonrecoverable Read Errors per Bits Read	1 per 10^{14}	1 per 10^{14}	1 per 10^{14}	1 per 10^{14}	1 per 10^{14}	1 per 10^{14}	1 per 10^{14}	1 per 10^{14}
Mean Time Between Failures (MTBF, hours)	700,000	700,000	700,000	700,000	700,000	700,000	700,000	700,000
Annualized Failure Rate (AFR)	0.34%	0.34%	0.34%	0.34%	0.34%	0.34%	0.34%	0.34%
Limited Warranty (years)	5	5	5	5	5	5	5	5

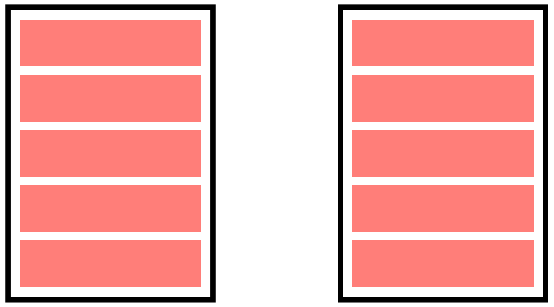
700,000 hours ≈ 80 years

(which seems.. suspicious)


“infant mortality”
stable operating period
burnout
P(error)
disk age

dealing with disk failures

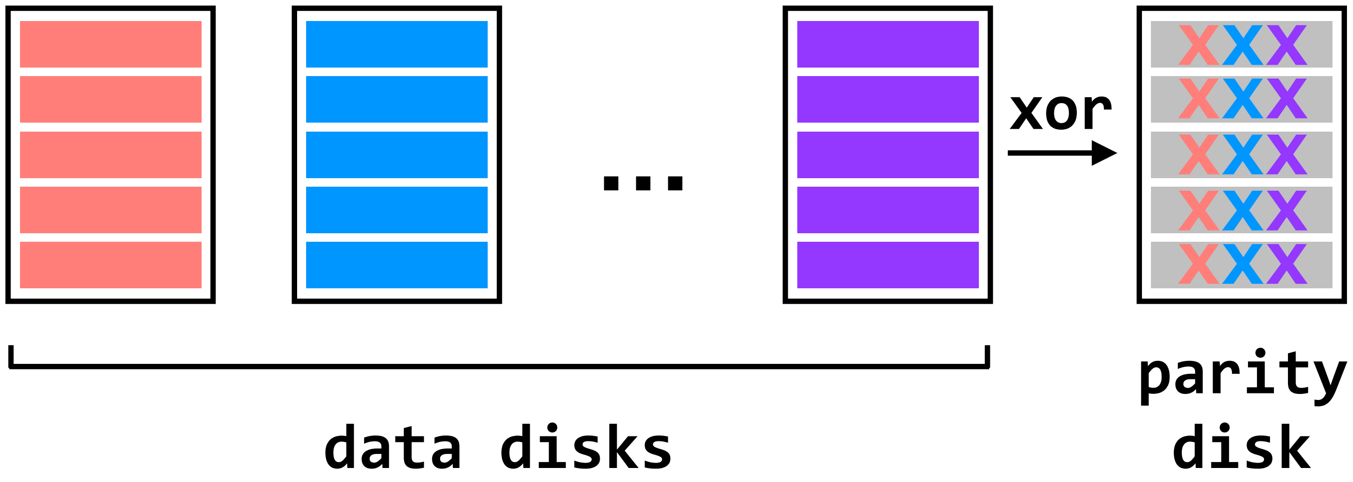
RAID 1 (mirroring)



+ can recover from single-disk failure

- requires 2N disks

RAID 4 (dedicated parity disk)



sector i of the parity disk is the xor of sector i from all data disks

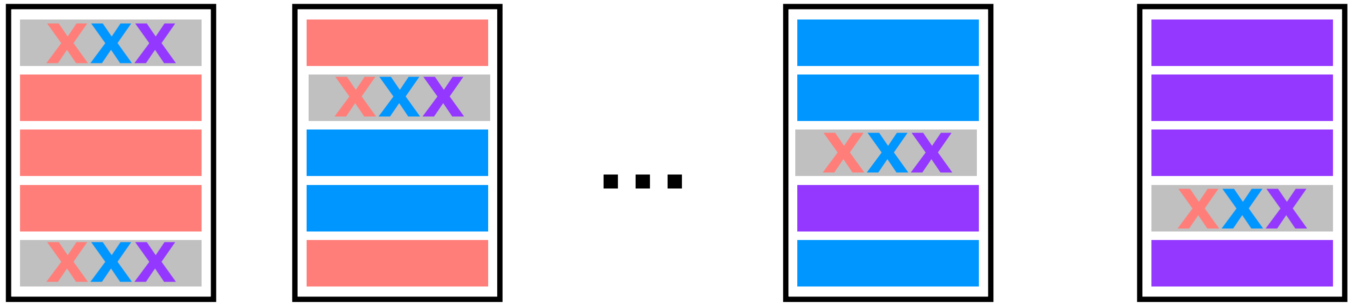
+ can recover from single-disk failure

+ requires N+1 disks (not 2N)

+ performance benefits if you stripe a single file across multiple data disks

- all writes hit the parity disk

RAID 5 (spread out the parity)



- + can recover from single-disk failure
- + requires N+1 disks (not 2N)
- + performance benefits if you stripe a single file across multiple data disks
- + writes are spread across disks

- Systems have faults. We have to take them into account and build reliable, **fault-tolerant systems**. Reliability always comes at a cost — there are tradeoffs between reliability and monetary cost, reliability and simplicity, etc.

- Our main tool for improving reliability is **redundancy**. One form of redundancy is **replication**, which can be used to combat many things including disk failures (important, because disk failures mean lost data).

- **RAID** replicates data across disks in a smart way: RAID 5 protects against single-disk failures while maintaining good performance.