MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Fault-tolerance

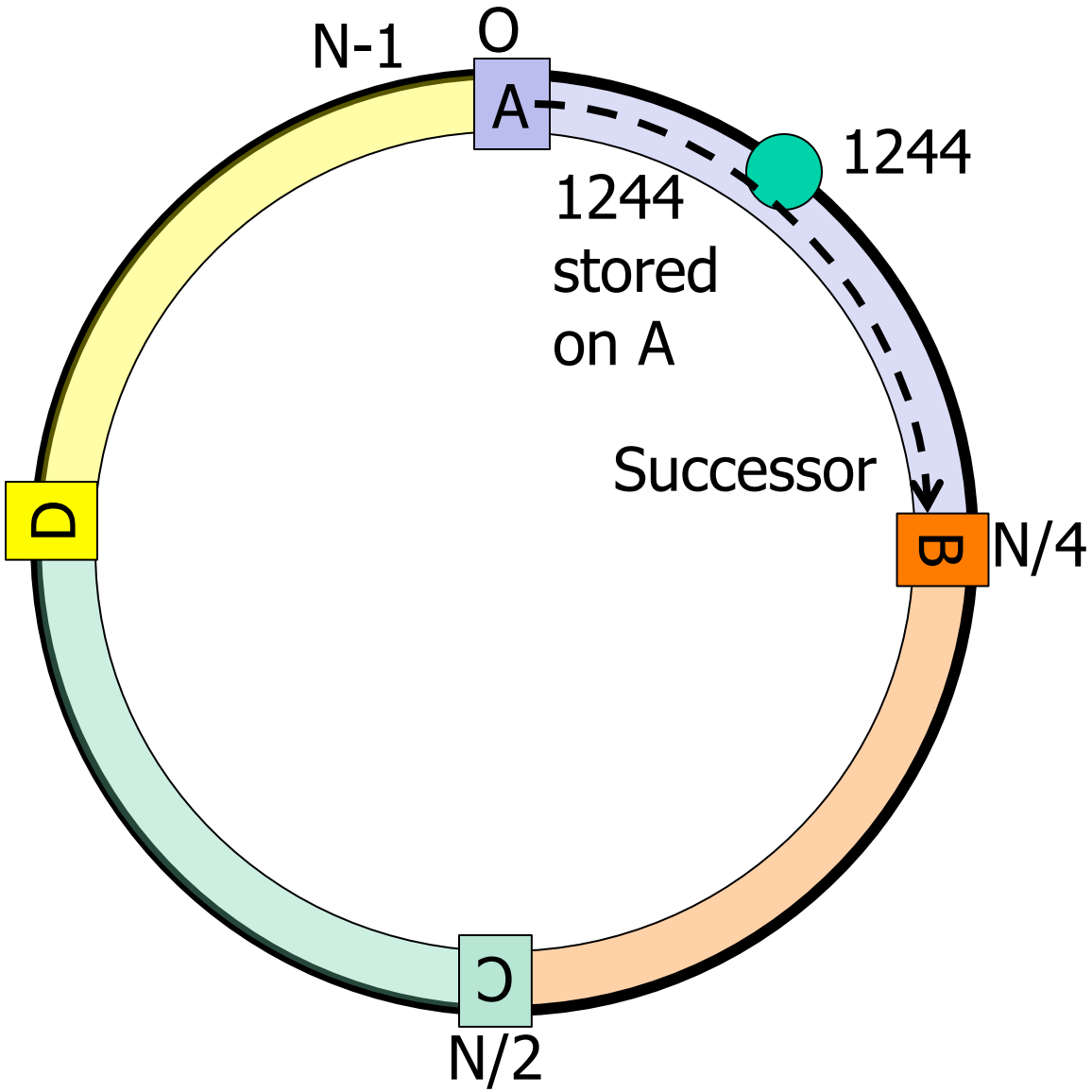
6.033 Lecture 15

Sam Madden

Today:

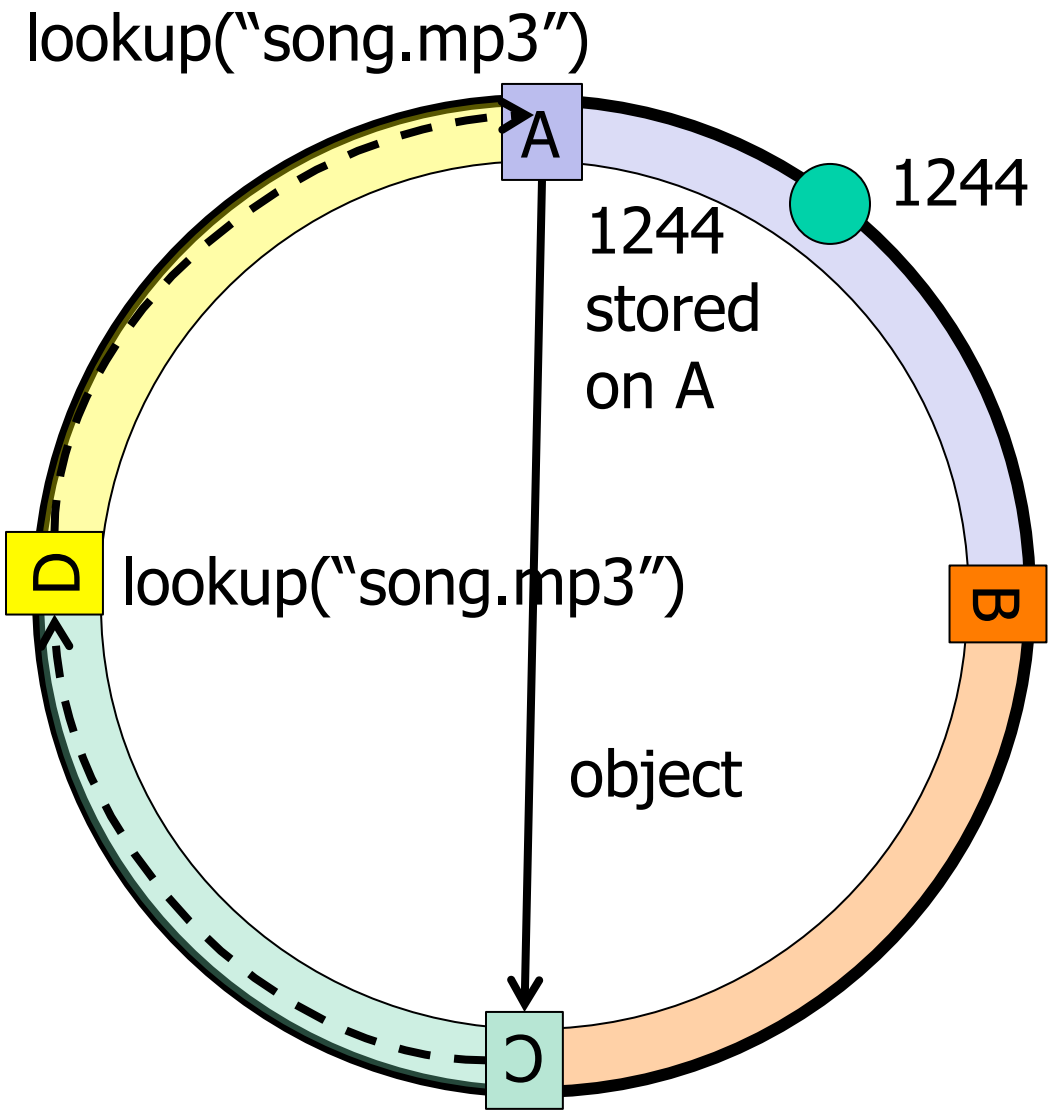
- Finish Chord
- Quantifying reliability
- General approach to fault tolerance
- Replication
- Study of disk failures

Chord Distributed Hash Table



- Objects assigned ids in range 0..N
  - Using hash function of object name
- Placed on ring
- Peers assigned range of ID space
  - Responsible for objects in range
- Successor pointers used to find objects

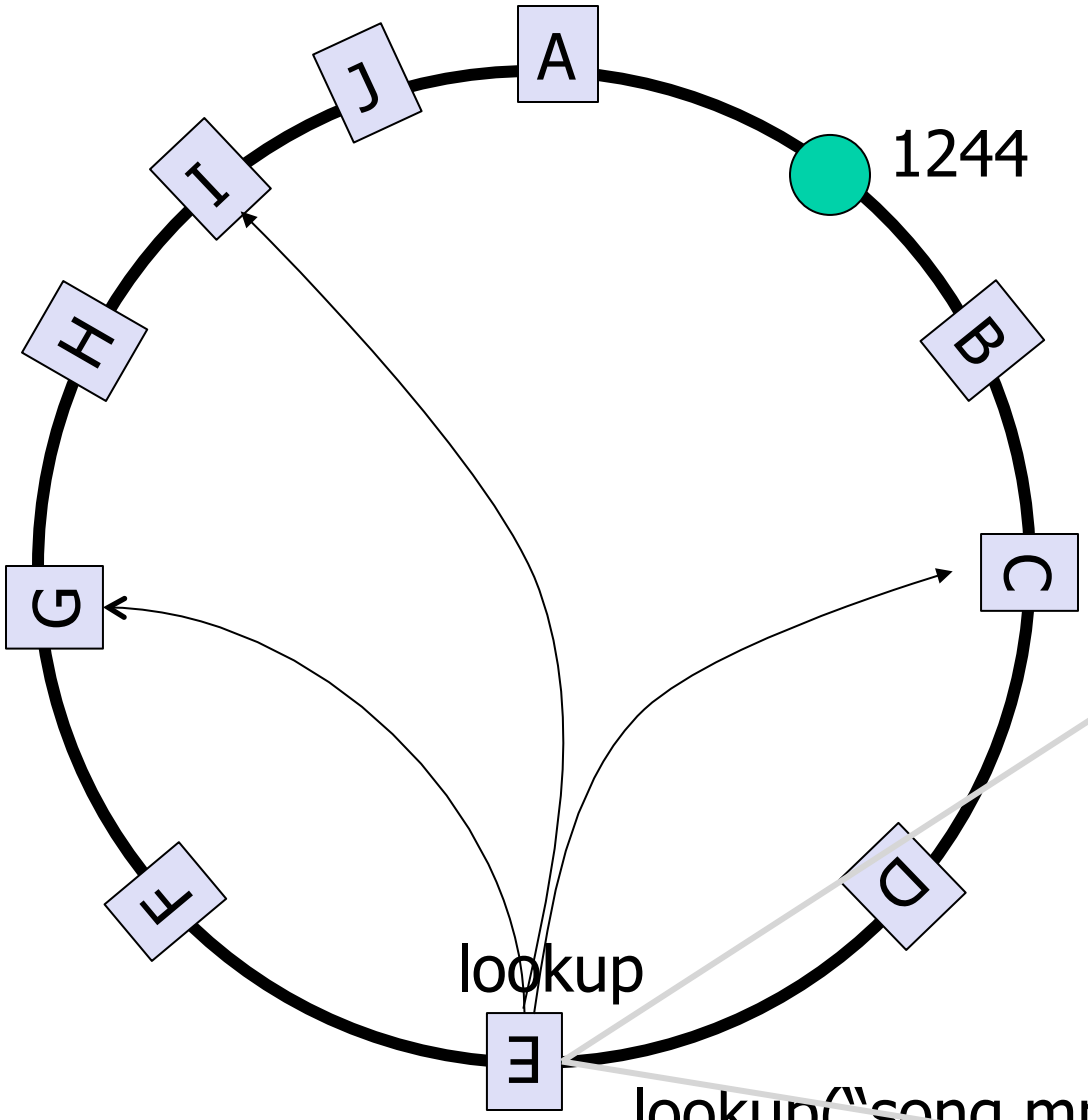
Object Lookup in Chord



Successor pointers used to find objects

```
lookup(name, dest):
  id = hash(name)
  if (id is local):
     send object to dest
  else
     successor.lookup(name, dest)
```

Optimizing lookup



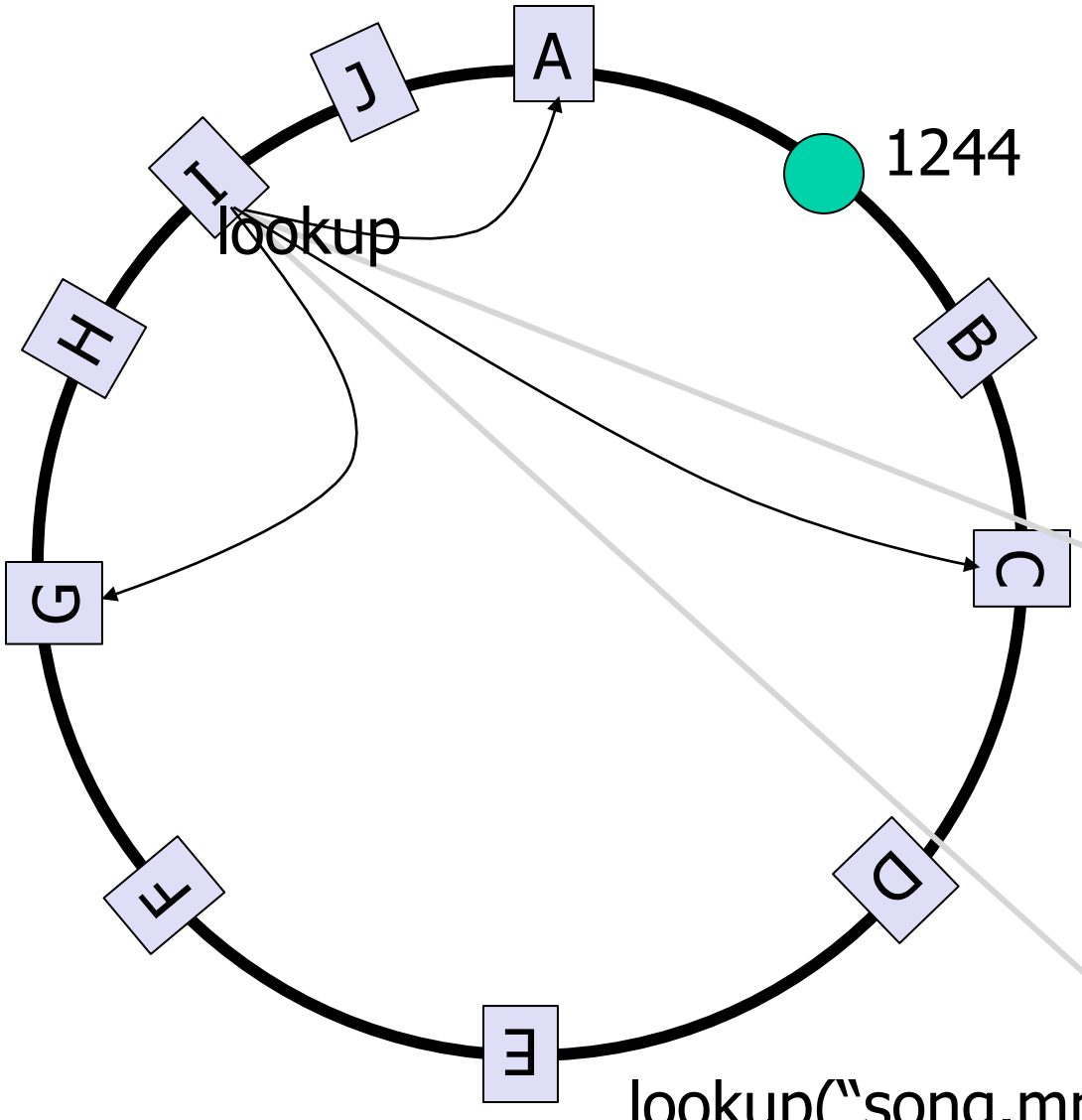
Problem: lookups require up to M messages, if M peers

Idea: store pointers to additional peers in "finger table"

Specifically to 2^{nd}, 4^{th}, 8^{th}, …, $2^{\log_2 M\ th}$ peers

Neigh	Range	IP
F	3500… 4000	5.6.7.8
G	4000… 6000	1.2.3.4
I	8000… 10000	2.3.4.5
C	2500… 3000	3.4.5.6

Optimizing lookup

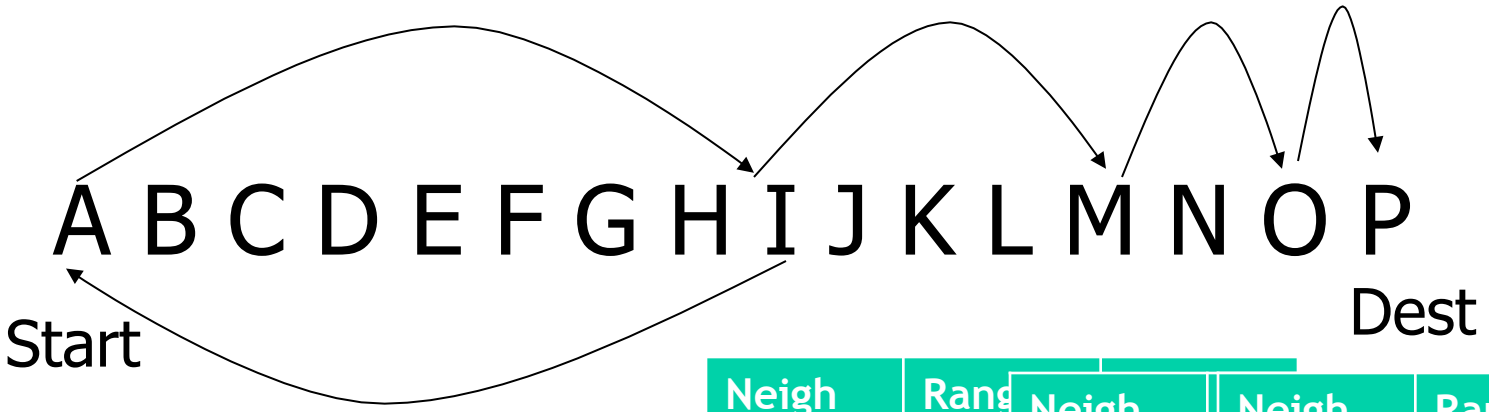


Log_2M hops

First step gets you at least 50% of the way, next step 25%, etc

Neigh	Range	IP
J	10000... 12000	4.5.6.7
A	0... 1500	1.2.2.3
C	2500... 3000	3.4.5.6
G	4000... 6000	1.2.3.4

Each hop uses next higher entry in finger table



Neigh	Range	IP
B	x	x
C	x	x
E	x	x
I	x	x

Neigh	Rang
J	x
K	x
M	x
A	x

Neigh	Neigh	Range	IP
N	P	x	x
O	A	x	x
A	E	x	x
E	I	x	x

Tables have Log_2M entries

Using Chord

- Objects can be whole files, blocks, or pointers to files
- Chord used in web-scale file systems, Akamai-like CDNs, and file-sharing applications

MASSACHUSETTS
INSTITUTE OF
TECHNOLOGY

Fault-tolerance

Concepts:

- Quantifying reliability
- General approach to fault tolerance
- Replication
- Study of disk failures

Relative frequency of hardware replacement

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

10,000 machines

Pr(failure in 1 year) ~.3

Schroeder and Gibson, FAST 2008


3107
A19

Data Sheet

Barracuda® 7200.10

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

80 GB to 750 GB • SATA 1.5Gb/s or 3Gb/s and PATA 100

Key Advantages

- First 3.5-inch drive to utilize capacity- and reliability-boosting perpendicular recording technology
- First drive to reach 750 GB—a full year ahead of competition—enabling new solutions for data-intensive applications.
- Industry's most proven and established desktop hard drive available today—more than 16 million shipped to date*
- "One-stop shopping" with a broad range of capacity, cache and interface options for all your computing needs
- Best-in-class environmental specifications and reliability features
- Adaptive Fly Height offers consistent read/write performance from the beginning to the end of your computing workload.
- Clean Sweep automatically calibrates your drive.
- Directed Offline Scan runs diagnostics when storage access is not needed.
- RoHS-compliant design assures an environmentally conscious product.
- Enhanced G-Force Protection™ defends against handling damage.
- Seagate® SoftSonic™ motor enables whisper-quiet operation.

Best-Fit Applications

Desktop and High-Performance PCs

- Gamer PCs
- Workstations
- High-end PCs
- Desktop RAID
- Mainstream PCs
- Point-of-sale devices/ATMs
- USB/FireWire/eSATA personal external storage

*16 million Barracuda 7200.10 drives shipped as of 4/16/07

Contact Start-Stops	50,000
Nonrecoverable Read Errors per Bits Read	1 per 10^{14}
Mean Time Between Failures (MTBF, hours)	700,000
Annualized Failure Rate (AFR)	0.34%

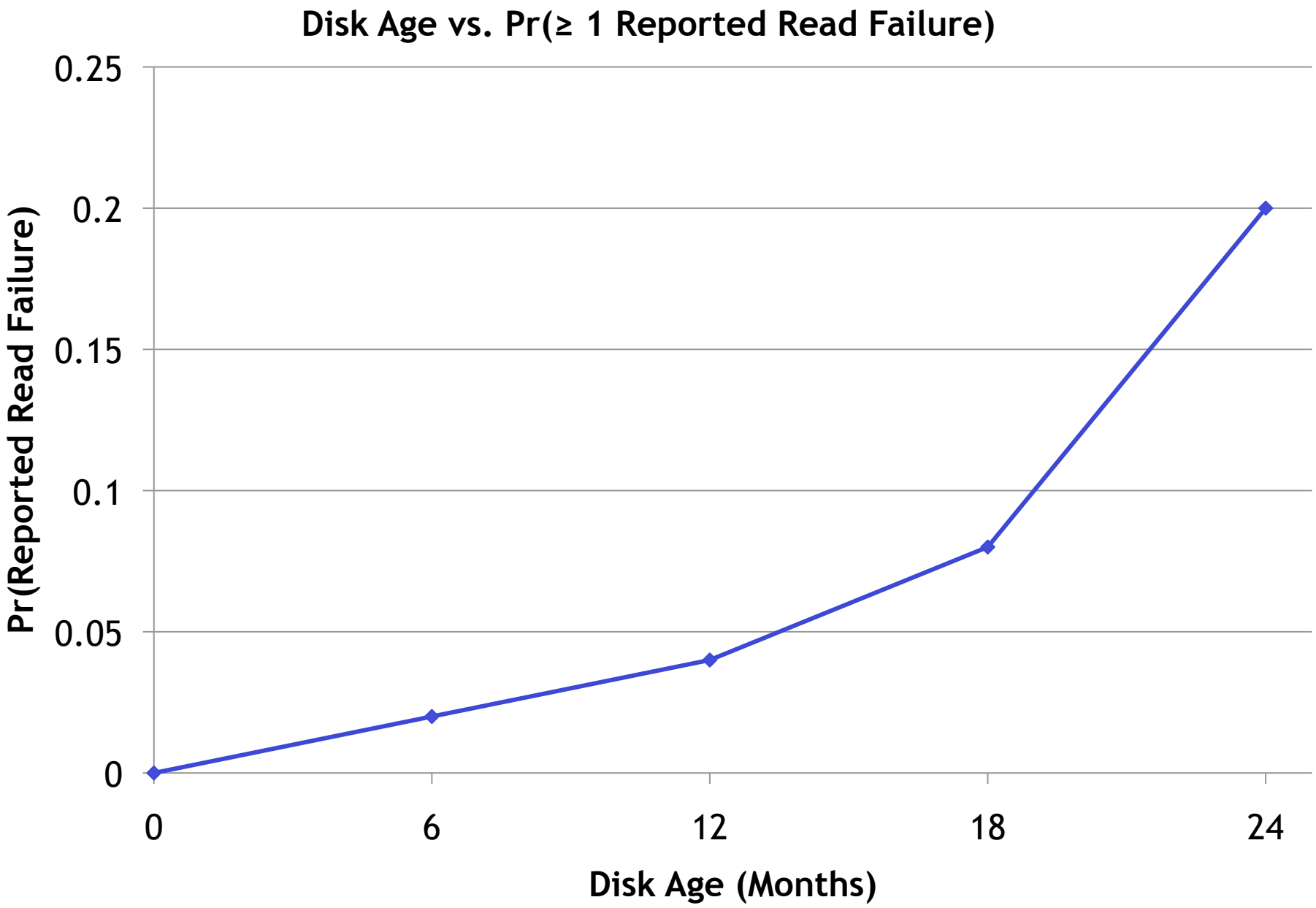
Data Sheet

Barracuda® ES.2

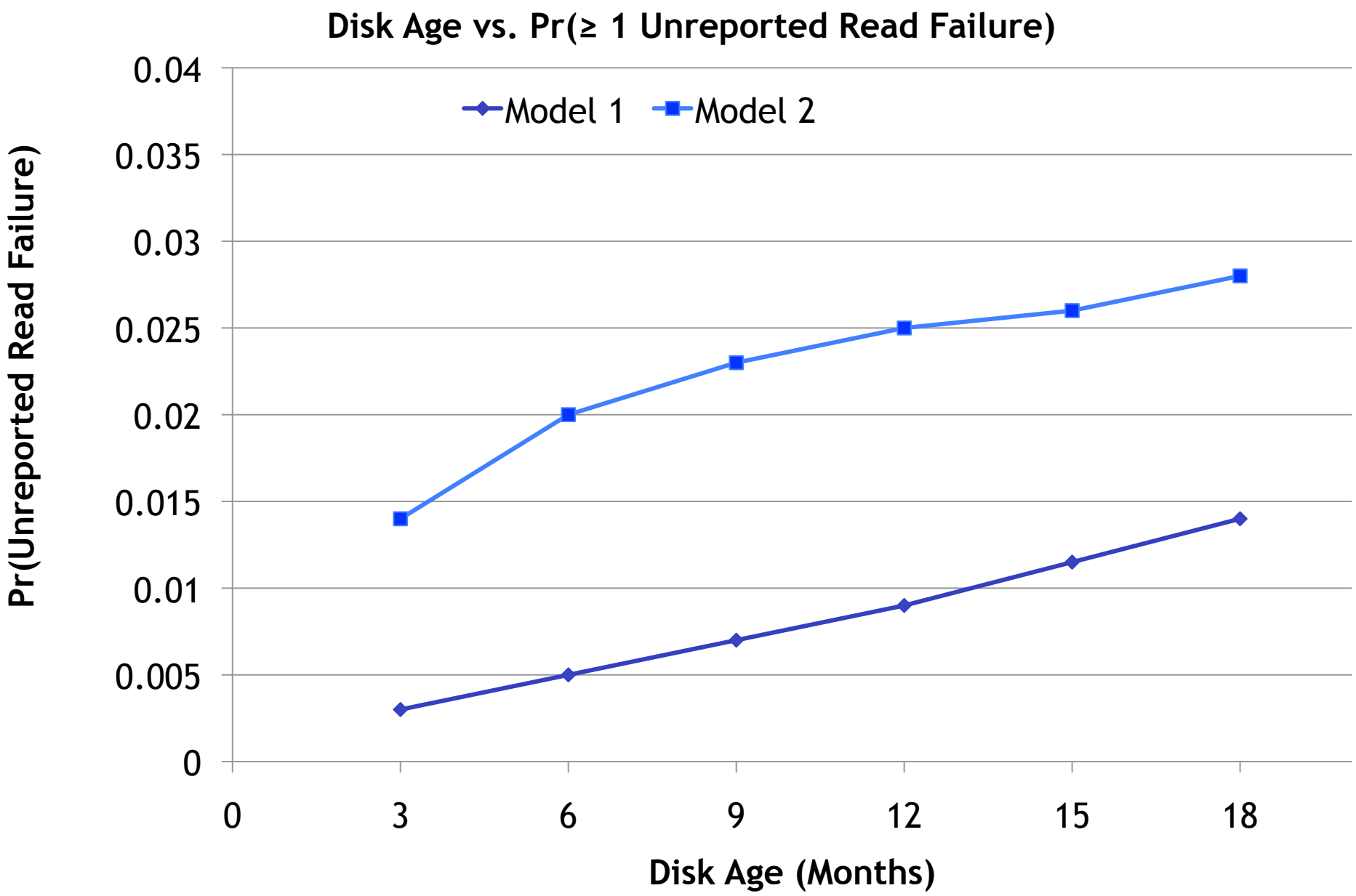
High-capacity, business-critical Tier 2 enterprise drives

1 TB, 750 GB, 500 GB and 250 GB • 7200 RPM • SATA 3Gb/s, SATA 1.5Gb/s and SAS 3Gb/s

Reliability/Data Integrity	
Mean Time Between Failures (MTBF, hours)	1.2 million
Reliability Rating at Full 24x7 Operation (AFR)	0.73%
Nonrecoverable Read Errors per Bits Read	1 sector per 10E15
Error Control/Correction (ECC)	10 bit
Interface Ports SATA SAS	 Single Dual



Bairavasundaram et al., SIGMETRICS 2007



Bairavasundaram et al., FAST 2008

Replicated Disks

```
write (sector, data):
    write(disk1, sector, data)
    write(disk2, sector, data)

read (sector, data):
    data = read (disk1, sector)
    if error
        data = read (disk2, sector)
        if error
            return error
    return data
```

Technical specifications

Processors	2–16 per node Intel Itanium processor 9100 series processors, 1.6 GHz single core processors
Cache	12 MB L3
RAM standard/maximum	Minimum: 4 GB Maximum: 16 GB (32 GB[2])
RAM type/speed	PC2100 ECC registered DDR266A/B
ServerNet I/O	Minimum: 10 Maximum: 60
I/O adapters supported	Fibre Channel, Gigabit Ethernet
Fibre Channel disk modules	14 disks per module
Disk drives supported	146 GB and 300 GB15K RPM Fibre Channel internal hard disk drives HP Disk Array family (e.g., XP24000, XP20000, XP12000, and XP10000 disk arrays)
Standard features	N + 1 power supplies N + 1 fans