



Confidence in a connected world.



# The Challenges of Storage System Growth

Denis Serenyi, May 14, 2007

Dartmouth Class of '96

# Disclaimer



- This presentation contains material from Symantec and from other sources openly published by others who have made that material publicly available on the world wide web. Our purpose for including portions of the publicly available material is to make readers aware of additional information available at the sites maintained by the copyright holders of those materials. All views expressed in this presentation are the views of Brian Witten and Denis Serenyi and not necessarily the views of Symantec Corporation or any Symantec affiliate. Please include this page in any copy of this presentation and in any copy of any portion of this presentation mentioning Symantec or including Symantec's logo.

# About today's talk



## Outline:

- Highlight scary numbers
- Introduce object storage & declustered RAID
- The execution challenge
  
- This talk is broad, looking forward to deep-diving in Q&A at the end
- Most slides have at least one Source URL, I also call out some good papers. The links are handy should you want to explore further

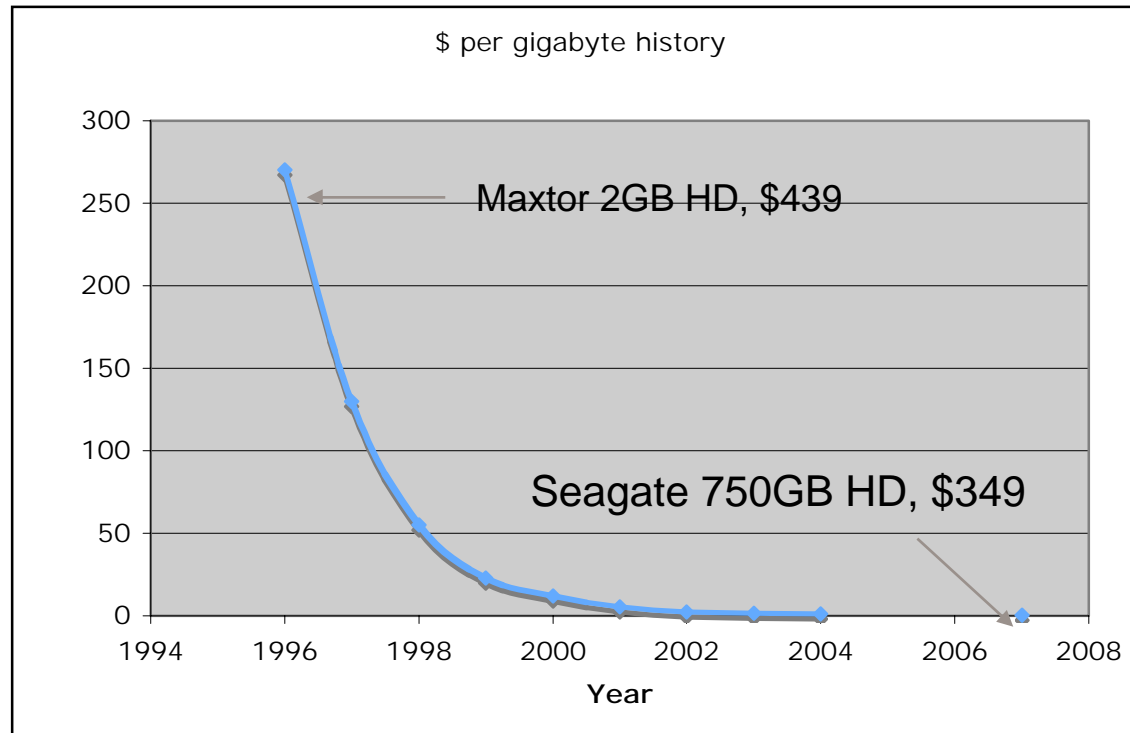
# Moore's Law of Storage



1996: Power Mac 6200  
1GB HD, \$2500

Disk storage cost halves every year

2007: Mac Pro  
750GB HD, \$2500



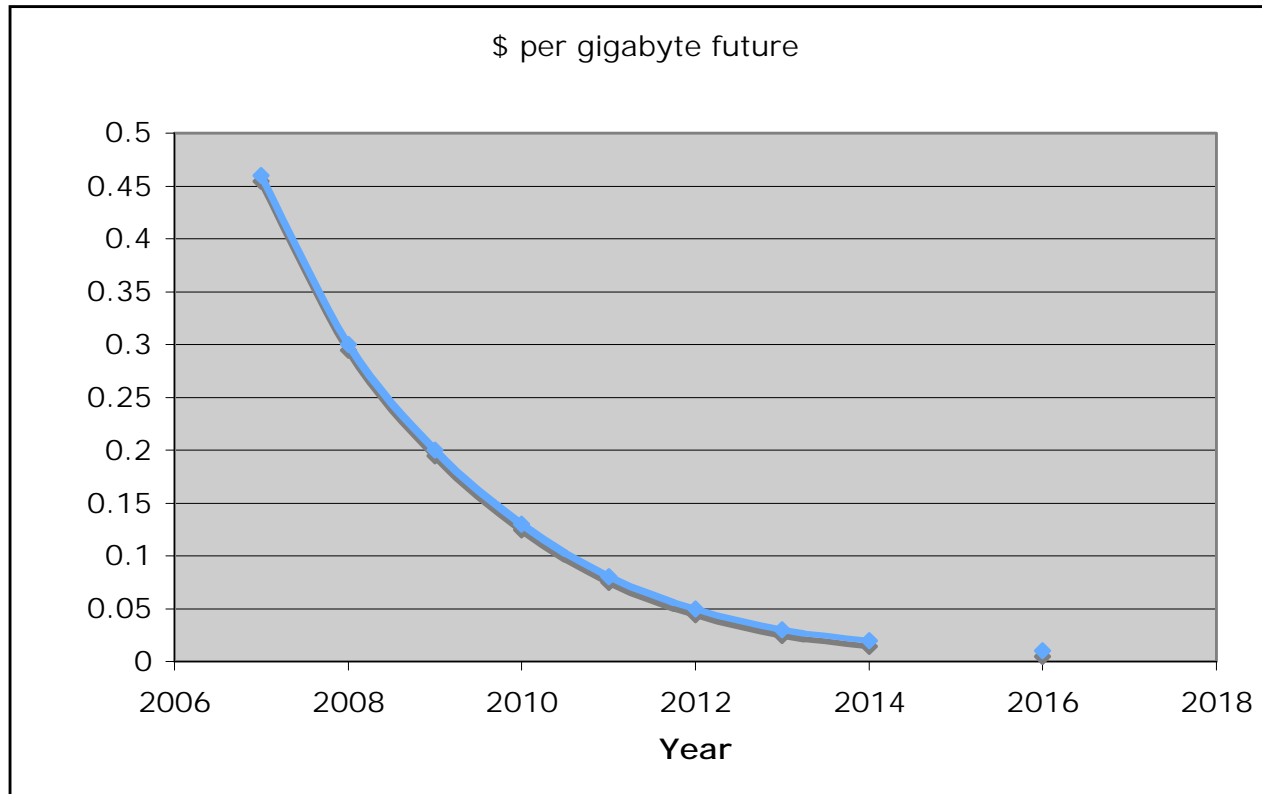
Source: <http://www.alts.net/ns1625/winchest.html>

# Extrapolating the trend



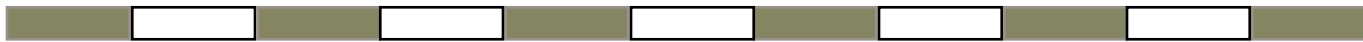
Today, 750 GB \$350

2017: 20 TB \$199?  
Physics may get in the way



# The Ever-Shrinking Bit

- Engineers fitting more and more bits/in<sup>2</sup> on platter surface
- 2005 drive densities are 150 Gbit/in<sup>2</sup>, an increase of 75 million times over when hard drives were first introduced in the '50s
- Perpendicular recording just being introduced



Traditional recording lays bits flat, interference results if too small



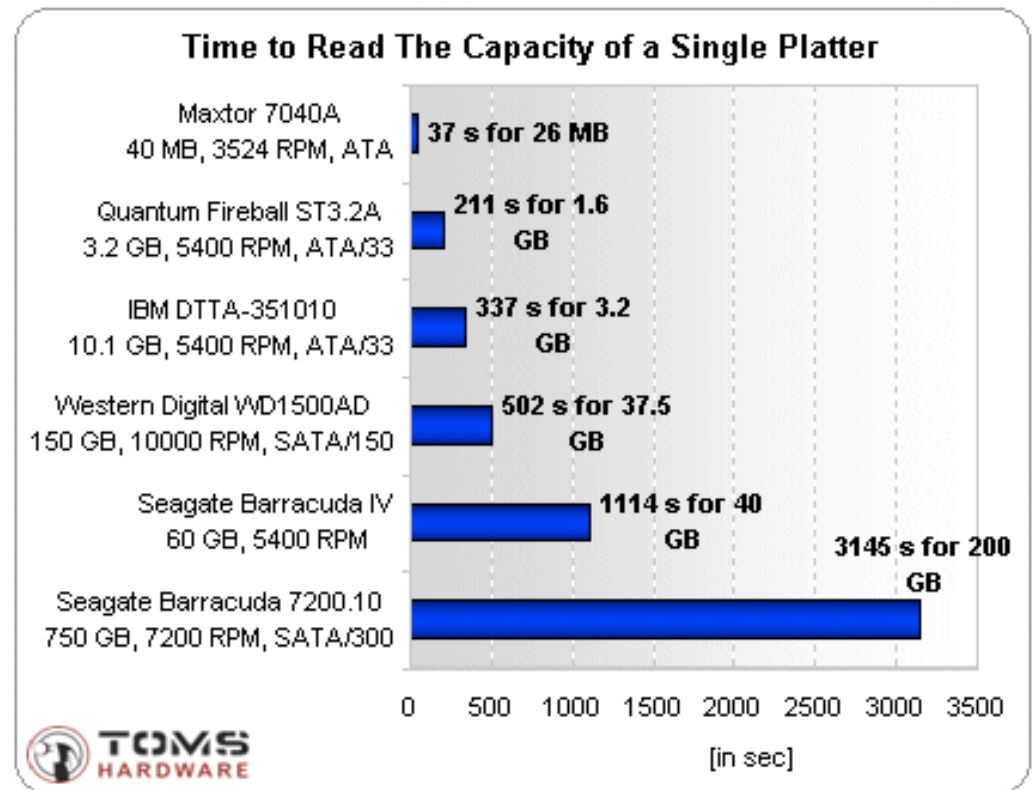
Perpendicular recording lays bits deep, counteracting the superparamagnetic effect, allowing smaller bits

- Perpendicular recording expected to top out at 1 Tb/in<sup>2</sup>
- More info on perpendicular recording:
  - [http://www.hitachigst.com/hdd/research/recording\\_head/pr/PerpendicularAnimation.html](http://www.hitachigst.com/hdd/research/recording_head/pr/PerpendicularAnimation.html)

Source: [http://en.wikipedia.org/wiki/Computer\\_storage\\_density](http://en.wikipedia.org/wiki/Computer_storage_density)

## But wait, it gets worse

- Per-spindle transfer speeds have not kept up with density growth. Also have not kept up with Moore's law.
- What has the “lunatic fringe” done? MORE spindles, so a super-exponential growth in aggregate size!
  - 1990: Supercomputers have disk farms of 50GB
  - 2002: ASCI Q: 700TB



Source: <http://www.dtc.umn.edu/resources/ruwart.ppt>  
<http://www.tomshardware.com/2006/11/27/15-years-of-hard-drive-history>

# Shhh... Disks Fail More Often Than Advertised

$$MTTF = \frac{\text{powered on hours per year}}{AFR}$$

AFR = annualized failure rate, or % of drives that fail in a given population during normal drive lifetime (5-10 years)

- Drive manufacturers claim
  - 1M - 1.5M MTTF, AFR of up to .88%
- Two recent papers provide substantial evidence that real-world AFR is around 3%, or an MTTF of 250,000 hours
  - Schroeder, Gibson (CMU): collected and analyzed failure data from real HPC sites
    - “Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?” (FAST ‘07)
  - Pinheiro, Weber, Barroso (Google): collected and analyzed failure data from Google’s data centers. Found that AFR for their 3 year old drives was 18%!!!
    - “Failure Trends in a Large Disk Drive Population” (FAST ‘07)

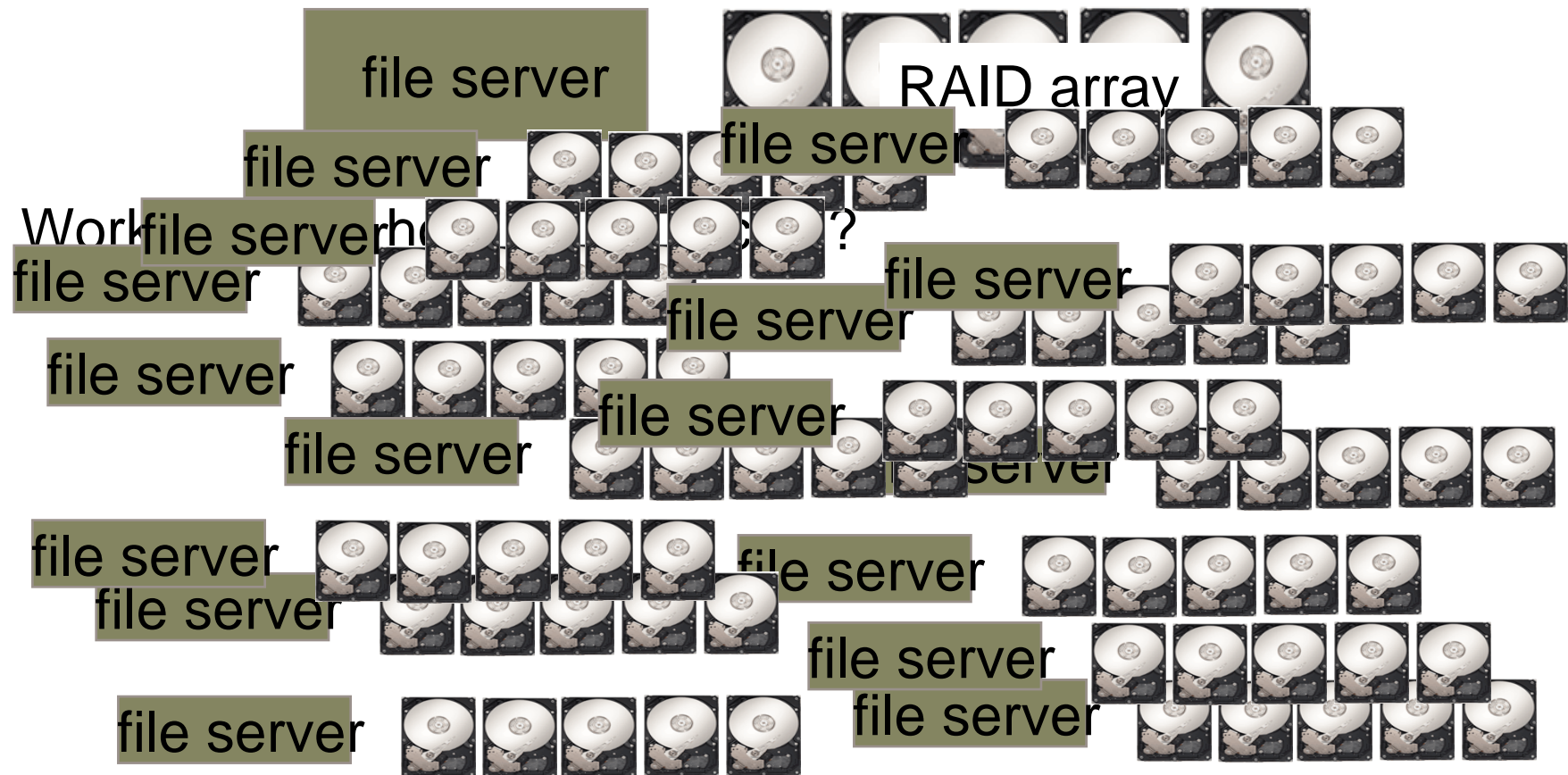
# The result: storage systems must tolerate *continuous* failure

These guys say it best:

- “Today's supercomputers, which perform trillions of calculations each second, suffer failures once or twice a day. Once supercomputers are built out to the scale of multiple petaflops, the failure rate could jump to once every few minutes. Petascale data storage systems will thus require robust designs that can tolerate many failures, mask the effects of those failures and continue to operate reliably.”
- “Imagine failures every minute or two in your PC and you'll have an idea of how a high performance computer might be crippled.”
  - Gary Grider, Los Alamos National Labs
- “With such a large number of components, it is a given that some component will be failing at all times”
  - Prof. Garth Gibson, CMU

Source: <http://www.hpcwire.com/hpc/876006.html>

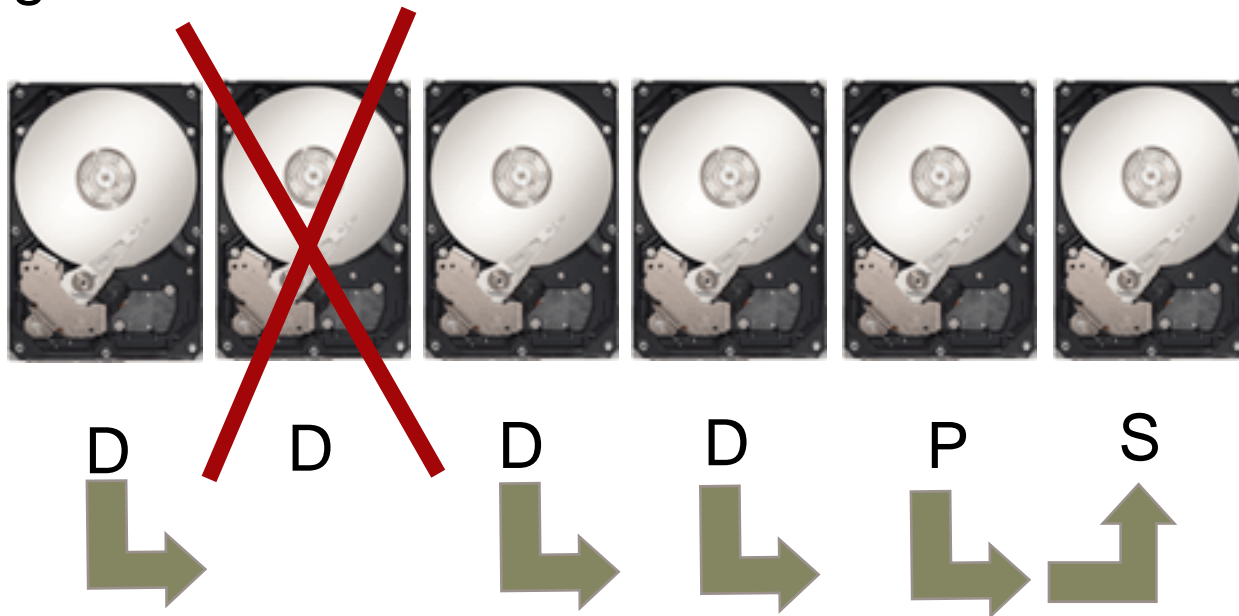
# Traditional Storage Systems



This can become a management nightmare.  
Can such a system tolerate failures gracefully?

# Traditional failure mode performs poorly in very large scale systems

RAID reconstruction: read all disks in RAID group, rebuild missing drive



Performance severely degraded during reconstruction

Reconstruction times are growing as bit densities increase:

RAID5 rebuild time for 500GB drives ~24hours

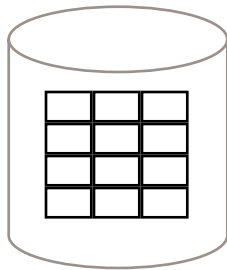
Source: [http://andirog.blogspot.com/2006\\_03\\_01\\_archive.html](http://andirog.blogspot.com/2006_03_01_archive.html)

# Object Storage



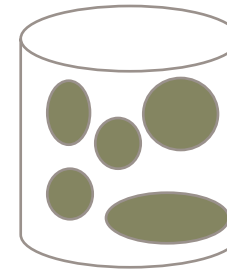
*A rethink of the storage system, starting with the drive itself*

The old way:  
Traditional block  
storage device



Operations: read block,  
write block  
Security: none

The spiffy new way:  
Object storage device



High level ops: create obj,  
read obj, write obj, del, etc  
Fine-grained access  
control with *capabilities*

# Object Storage System Architecture

- Storage subsystem moved to storage device
  - Storage addressed with compact metadata (object ID)
  - Media geometry aware placement
  - Data aware prefetching & caching
- File system can be managed with lightweight, scalable metadata manager
- Strong security enables direct client access to OSDs for most operations

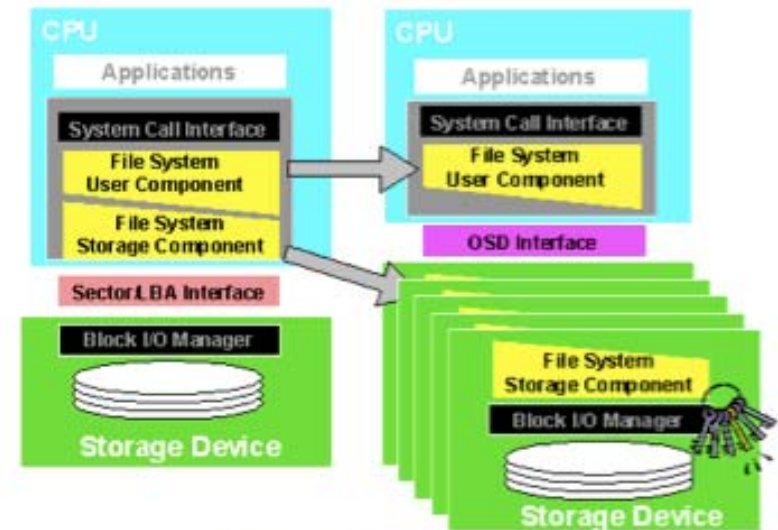
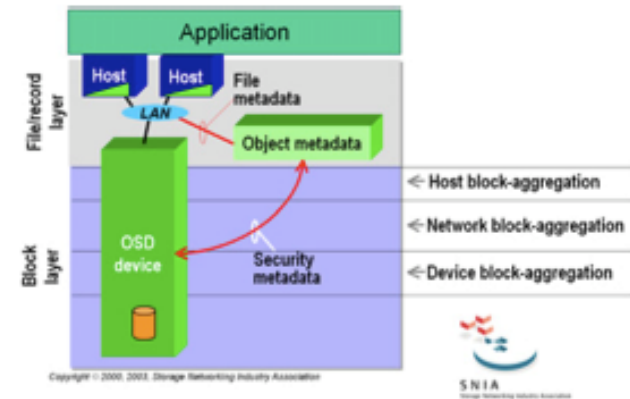


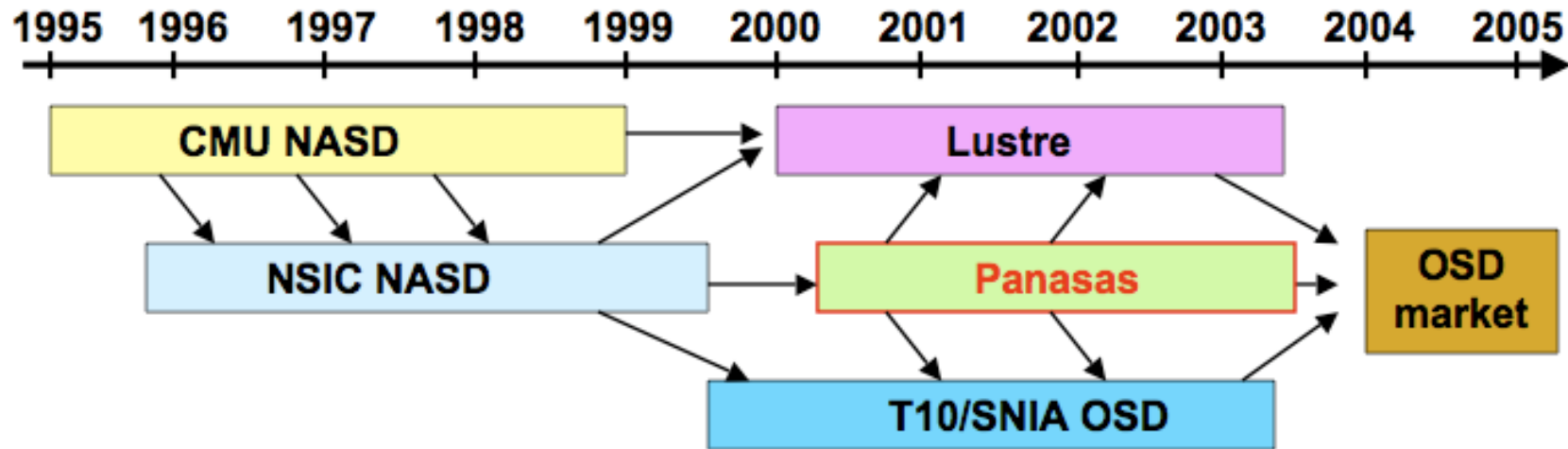
Figure 3. OSD in a typical software architecture

## Object-Based Storage Device (OSD), CMU NASD



Source: [http://www.seagate.com/docs/pdf/whitepaper/tp\\_536.pdf](http://www.seagate.com/docs/pdf/whitepaper/tp_536.pdf)

# Object Storage Adoption & Standardization



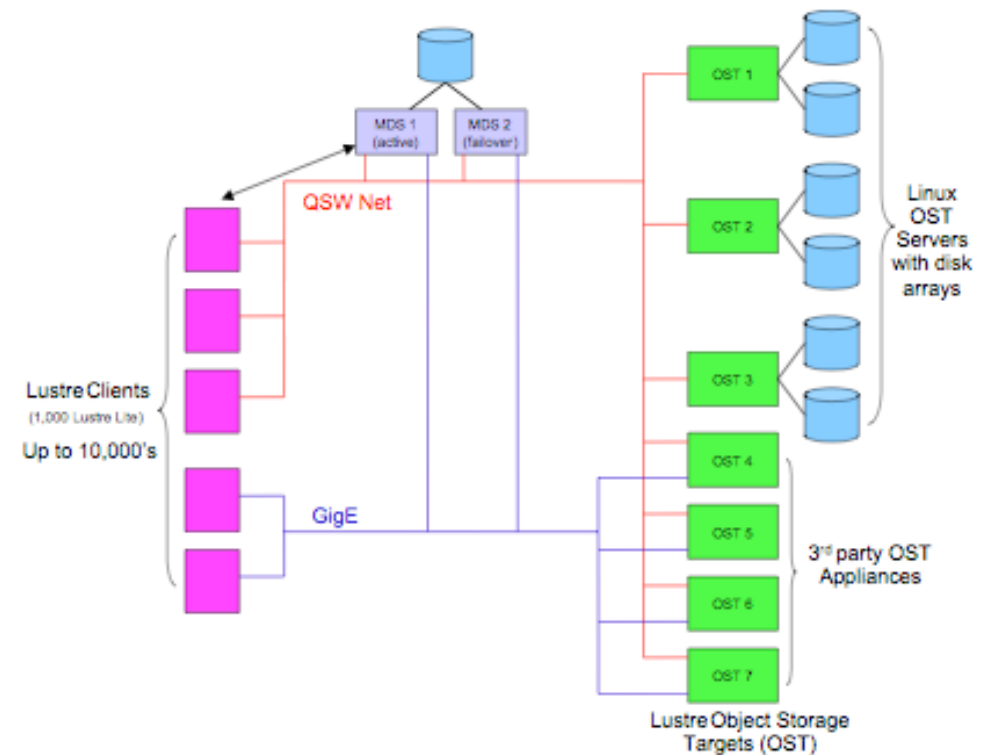
Source: <http://www.dtc.umn.edu/resources/welch.pdf>

- T10 OSD 1.0 released 2003, development of OSD 2.0 ongoing
- Broad industry support: products being developed, products released, or participation in the development of the standard
  - Symantec, IBM, HP, Seagate, EMC, Lustre, Panasas

# Case Study:

- HPC oriented object based filesystem
- Metadata servers, Linux VFS client, combined with “OST” OSDs combine to form (mostly) POSIX compliant filesystem
- OSTs use RAID internally (typically, not required)
- In use in some of the world’s largest supercomputers, including a 1.2 PB system at LLNL

Figure 1: Lustre Big Picture



Source: <http://www.lustre.org/docs/whitepaper.pdf>

# RAID Declustering: More Fault Friendly

- Basic idea: multiplex a RAID group of size  $N$  onto  $M$  disks, where  $M > N$
- Not every stripe needs to be read during reconstruction, so less work to reconstruct
- Results in faster recovery time and / or lighter load during recovery

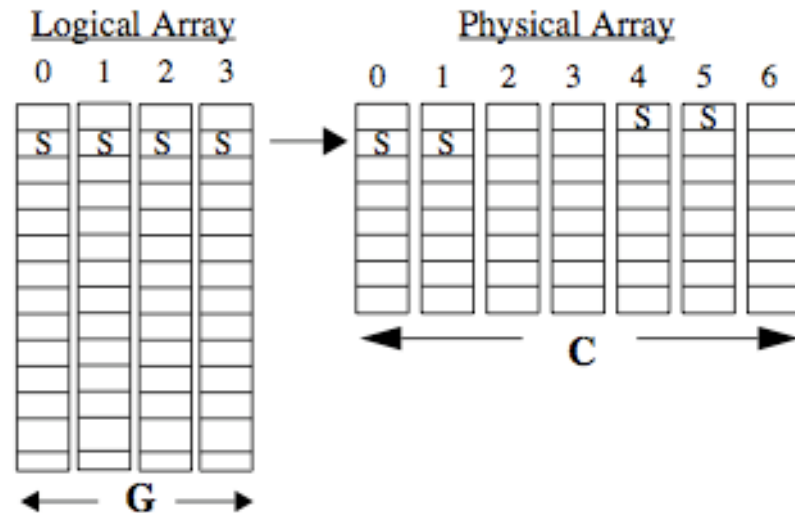


Figure 2-2: One possible declustering of a parity stripe of size four over an array of seven disks.

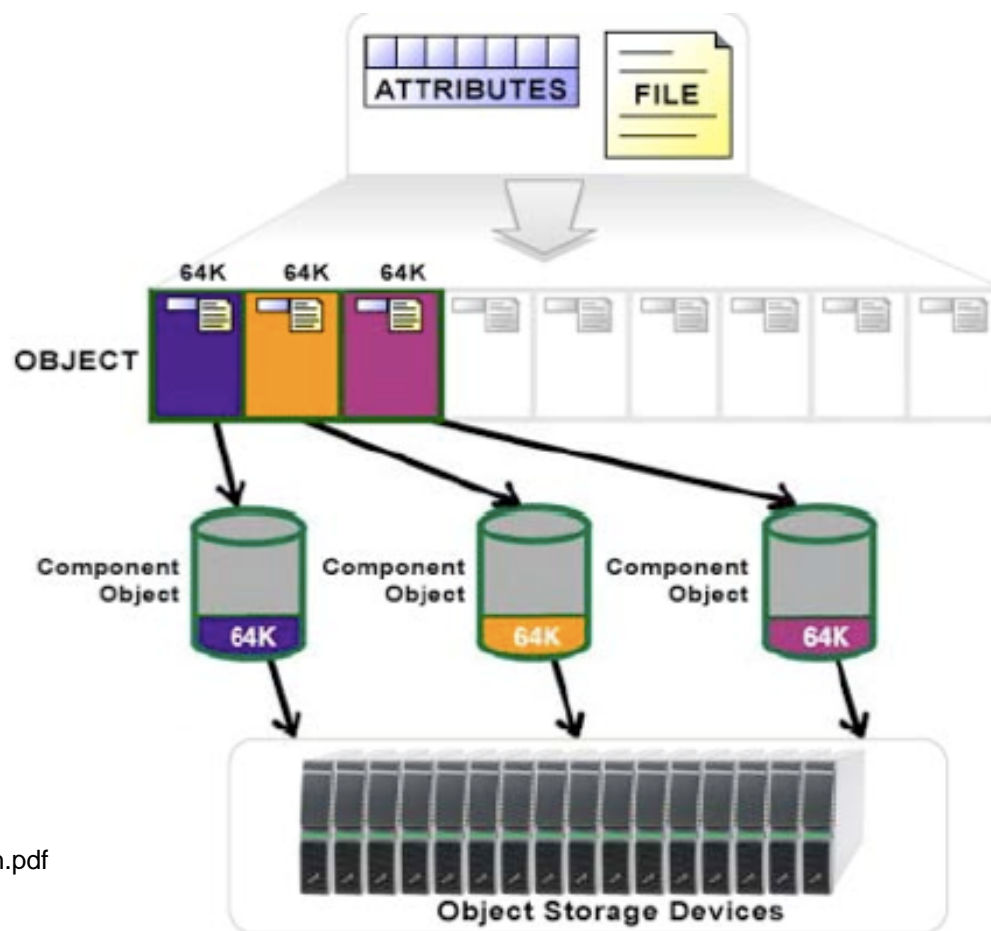
SOURCE: Gibson / Holland: "Parity Declustering for Continuous Operation in Redundant Disk Arrays"

## Case Study:

panasas



- Storage consists of arrays of OSDs and metadata managers
- Files stored as several *component objects*, which are distributed across OSDs
  - So, this is object storage with full declustering
  - Each file can have its own unique RAID layout



Source: <http://www.dtc.umn.edu/resources/welch.pdf>

# Case Study: Google™

Built its own distributed file system: Bigtable + GFS

- GFS files typically 3-way mirrored
- Bigtable files use double-parity erasure encoding
- GFS stores files in *chunks*, which are spread over a large number of *chunk servers*
  - Results in an “object like” file system, with full declustering
- Results in ability to tolerate up to 6 drives failing simultaneously with no data loss

Source: <http://labs.google.com/papers/bigtable.htm>  
<http://labs.google.com/papers/gfs.html>



# MTTDL Calculation



$$\text{MTTDL}_{(\text{RAID5})} = \frac{\text{MTTF}^2}{D * (D-1) * \text{MTTR}}$$

MTTR = mean time to repair (recon time)

D = size of RAID group

- Does declustering increase MTTDL?
  - No. D must increase because declustering increases the number of physical drives in the declustered RAID group. MTTR is lowered proportional to the increase in D, resulting in no change to the denominator
  - However, declustering is more *fault friendly*, lowering the performance impact of reconstruction by spreading the work

# The Challenge of Software Complexity

- The filesystem has one of the highest reliability bars in the industry
  - Users absolutely do not tolerate lost or corrupt data, system uptime must be at least 99.99%, 99.999% preferable
  - But as we've seen, hw continues to be *less* reliable, so failure codepaths executed as commonly as fault-free path.
- Not your daddy's filesystem
  - Modern, distributed file system has 1M+ lines of code
  - Lustre, Panasas both date back to 2000, After 7 years, still considered not ready for the enterprise
  - Peter Braam, Lustre creator: "It's not like backing your car out of the driveway. Installing Lustre is more like launching the space shuttle, with pieces of foam falling off."

# Tackling the Complexity Problem

- “Parallel File System Testing for the Lunatic Fringe: the care and feeding of restless I/O Power Users” (Hedges, Loewe, McLarty, Morrone)
  - LLNL’s experience evaluating and testing Lustre
- Takeaways
  - Highly parallelized, concurrent access demanded development of new test tools, which discovered subtle race conditions, problems at large scale, etc
  - Data corruption bugs found
  - Filesystem aging an issue: correctness and performance on day 1 does not assure correctness and performance on day 60

- Division of Symantec Research Labs – Darren Shou, Director
- *New SRL Graduate Research Fellowship*
  - Focus on innovative research with real-world value to Symantec's customers
  - Open to graduate students at all U.S. universities
  - Winners selected based on their scholarship and research proposals
    - Receive tuition, stipend, and a summer internship in SRL
    - Student fellows are paired with a research mentor
  - 2007-2008 SRL Graduate Research Fellows:
    - David Brumley, Carnegie Mellon University
    - Jack Lange, Northwestern University
    - Justin Ma, University of California San Diego
  - 2008-2009 Graduate Fellowship applications available Aug 2007
    - <http://www.symantec.com/about/careers/working/graduatefellowshippgms.jsp>

# Other Funding Opportunities



- University/Industry affiliate programs for directed research
- Requests for Proposals and collaborative research projects
  - UC Santa Cruz: Adaptive File Migration Policies project
  - UCLA: Image-Spam Detection project
  - Georgia Tech: Usable Security contest
- Technical speaker programs
  - Security researchers from CMU, Stanford, Johns Hopkins,...
  - 3-part series from storage systems researchers
- Product donations
- Support for conferences, lecture series, and student clubs



Confidence in a connected world.



**Thank You!!**  
**Go Big Green!**

# Symantec Research Labs



- “Ensuring Symantec’s long-term leadership through innovation, generation of new ideas, and development of next-generation technologies.”
- Experts in security and storage research
- Researchers nationwide/worldwide:
  - Major centers in Santa Monica and Mountain View, CA; Pune, India
  - Other researchers in Virginia, Chicago, Ireland...
- Collaboration with other researchers, government agencies, and universities:
  - National Science Foundation, Department of Homeland Security, European Commission
  - Carnegie Mellon, Stanford, George Mason, Georgia Tech, Purdue, and numerous other universities and research organizations