

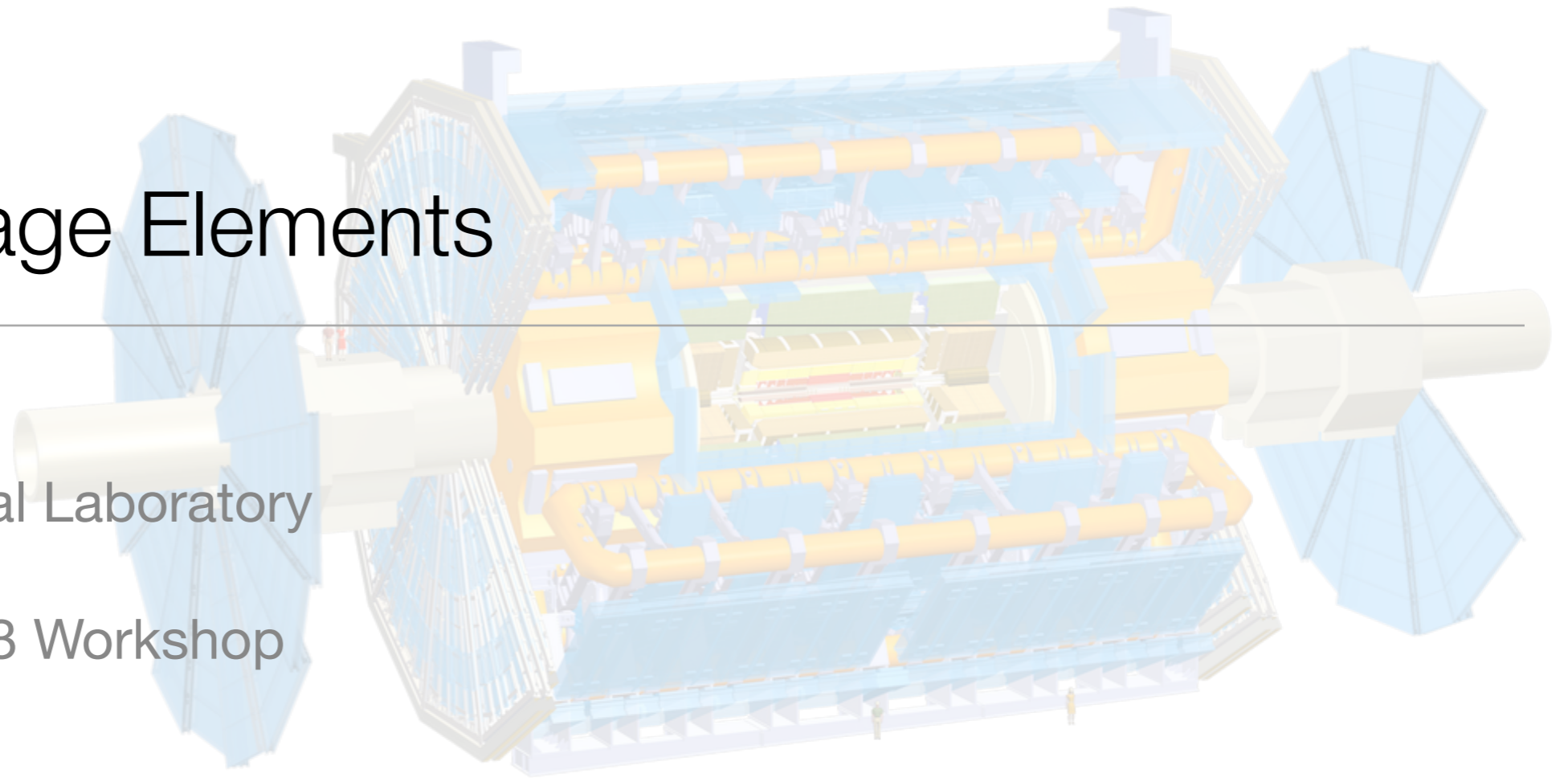
# Fabric Storage Elements

---

Robert Petkus  
Brookhaven National Laboratory

US ATLAS Tier 2 & 3 Workshop  
SLAC

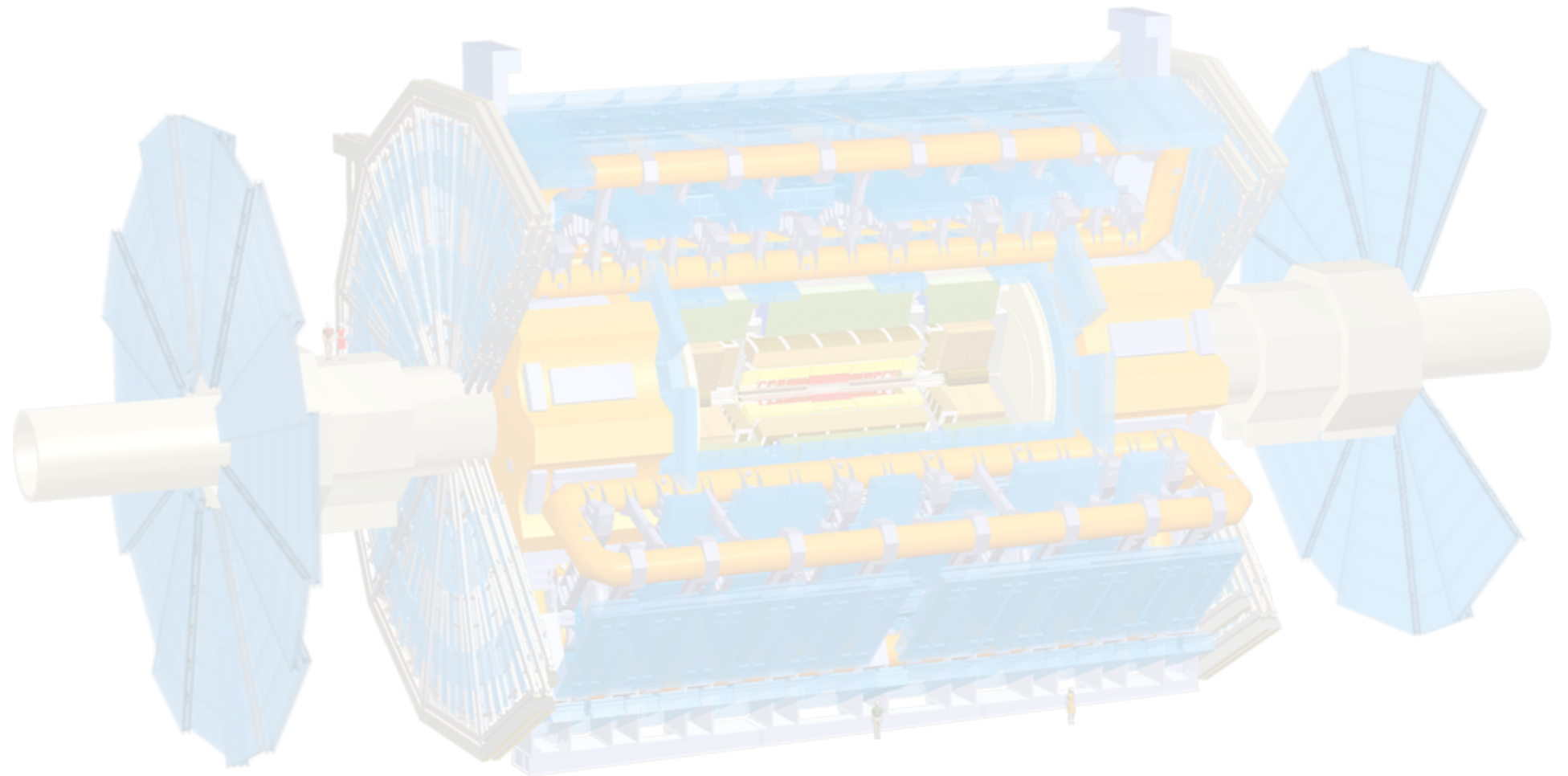
29 November, 2007



# Overview

---

- Storage Systems
  - 2007
  - 2008
  - Future
  - Tier 3

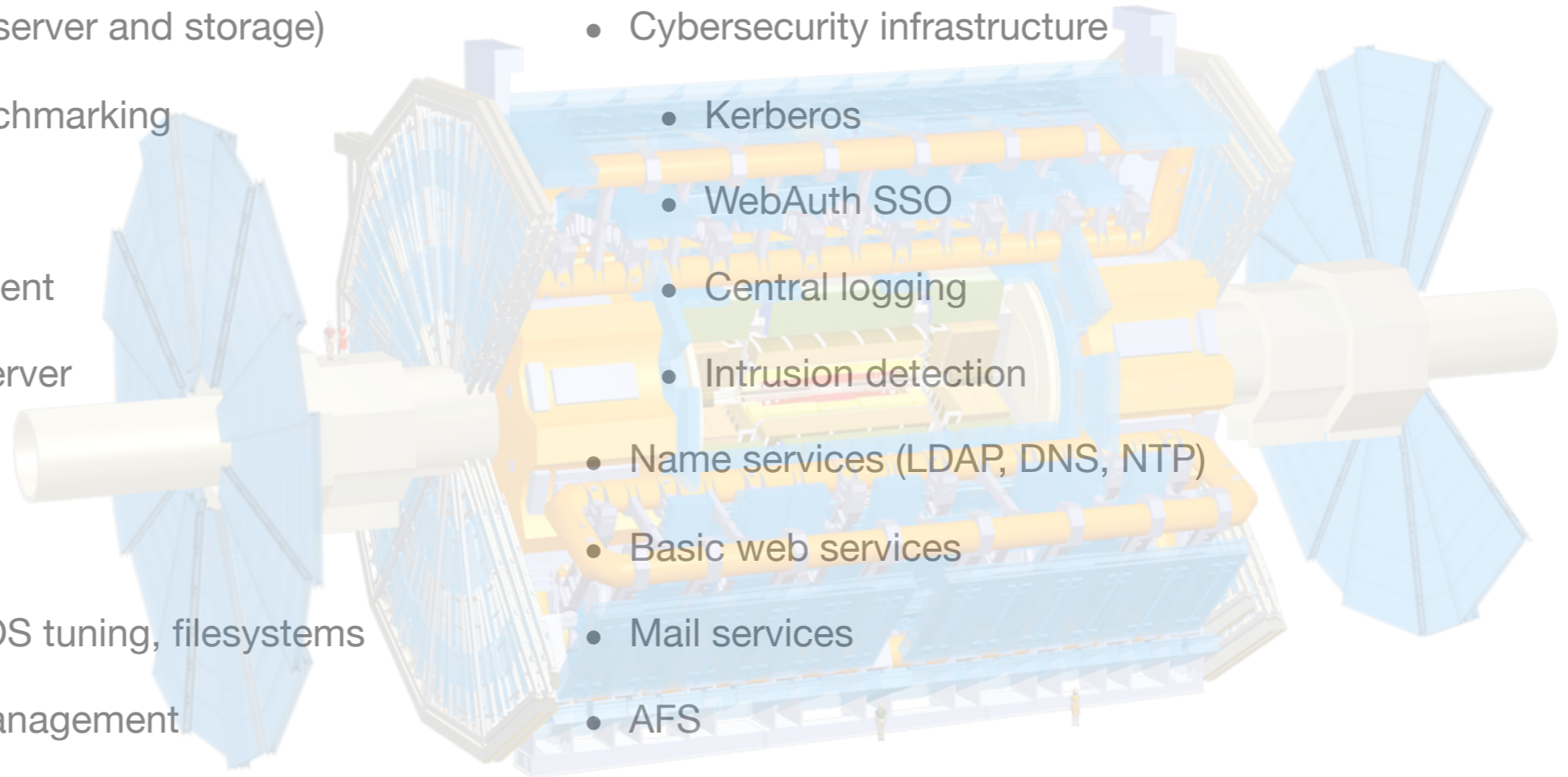


# General Computing Environment

## Infrastructure Management

- Hardware provisioning (server and storage)
  - Evaluation and benchmarking
  - System builds
- Configuration management
  - Red Hat Satellite Server
  - Cfengine
- System administration
  - Linux and Solaris, OS tuning, filesystems
  - Package / patch management
  - Remote access
  - Monitoring (Nagios, Ganglia, Cacti, snmp)
- Network operations and infrastructure

- Cybersecurity infrastructure
  - Kerberos
  - WebAuth SSO
  - Central logging
  - Intrusion detection
- Name services (LDAP, DNS, NTP)
- Basic web services
- Mail services
- AFS
- User life-cycle management

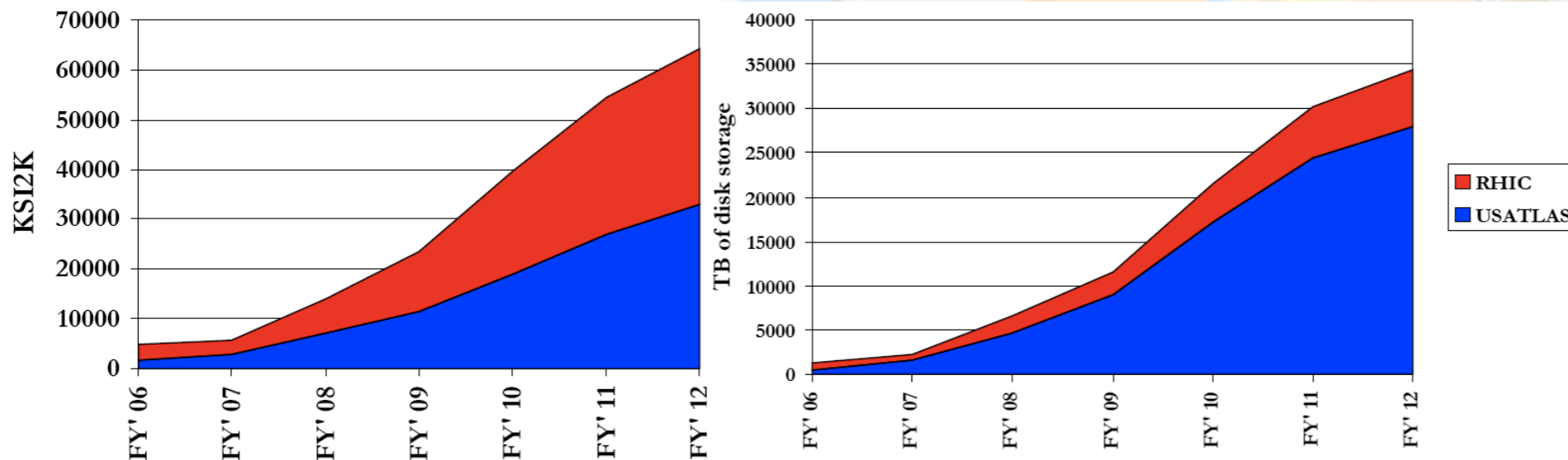




# Storage Systems

We are in the midst of an unprecedented disk growth spurt:

- 3PB coming in the door in 2008 & >30PB in 2012
- Storage elements are moving from the processor farm to dedicated file servers. Why?
  - Storage requirements outpace computational requirements
  - Management and scalability concerns
  - Limited data center real estate, power, and cooling



- There are (2) classes of disk storage at RACF: **distributed** (dCache, xrootd) & **centralized** (NFS)

# Tier 1/2 Distributed Storage System Options

## Primary considerations for candidate systems

- Preference for high density vertical disk arrays:
  - SunFire x4500 (integrated server + storage) and Nexsan SATAbeast (storage only) are 5U, 48-bay stand-alone systems. 500GB and 1TB disks (240-480TB/rack)
    - ★ *Within the realm of reason for a Tier 3 facility*
  - Xyratex disk arrays with DDN S2A controllers (also re-branded by IBM) with 48 (480TB/rack) and 60 bay (600TB/rack) 4U disk enclosures. 1TB SATA.
  - 3PAR S400/800 controllers and disk arrays with 40 bay disk enclosures (10x4 disk magazines) for 150TB/rack. 750 GB SATA

★ Note that many data center floors will not accommodate the weight of a fully populated

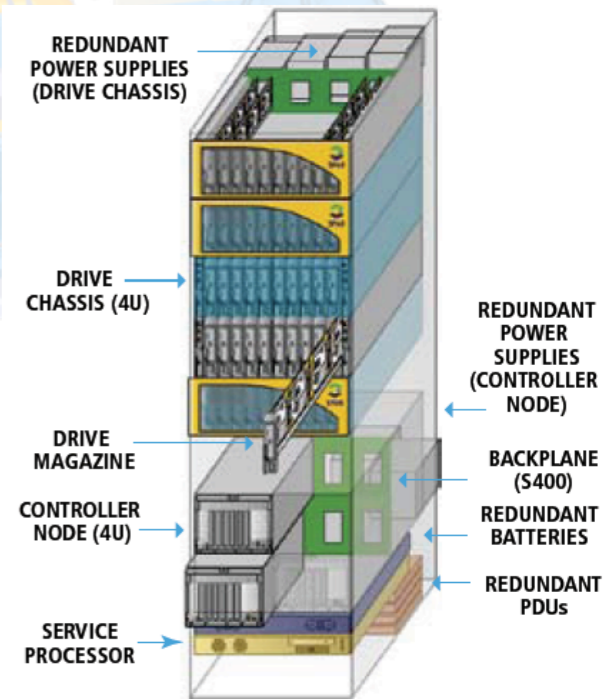
DDN StorageScaler

SunFire x4500

Nexsan SATAbeast

DDN S2A 9550

3PAR S800





# Distributed Storage System Considerations

## Primary considerations for candidate systems continued

- Hardware/software compatibility
- Infrastructure connectivity
  - Front end: Channel-bonded 1GE, 10GE, iSCSI, IB
  - Back end: IB, FC, SAS, SATA
- Supported RAID levels. Do the RAID controllers support RAID-6? They had better with 1TB SATA disks.
- Management
  - Simple set-up, flexible configuration options
  - Built-in monitoring tools (snmp, e-mail alerts, etc.)
  - Remote access (ILOM, RSC)
- Incremental growth profile; e.g., is it cheap and easy to add several TB?
- Fault tolerant: dual power supplies, dual controllers
- Service and support. Ideally 3-year support on all components or disks at a minimum. Next-day replacement



# 2007 Distributed Storage Deployment

---

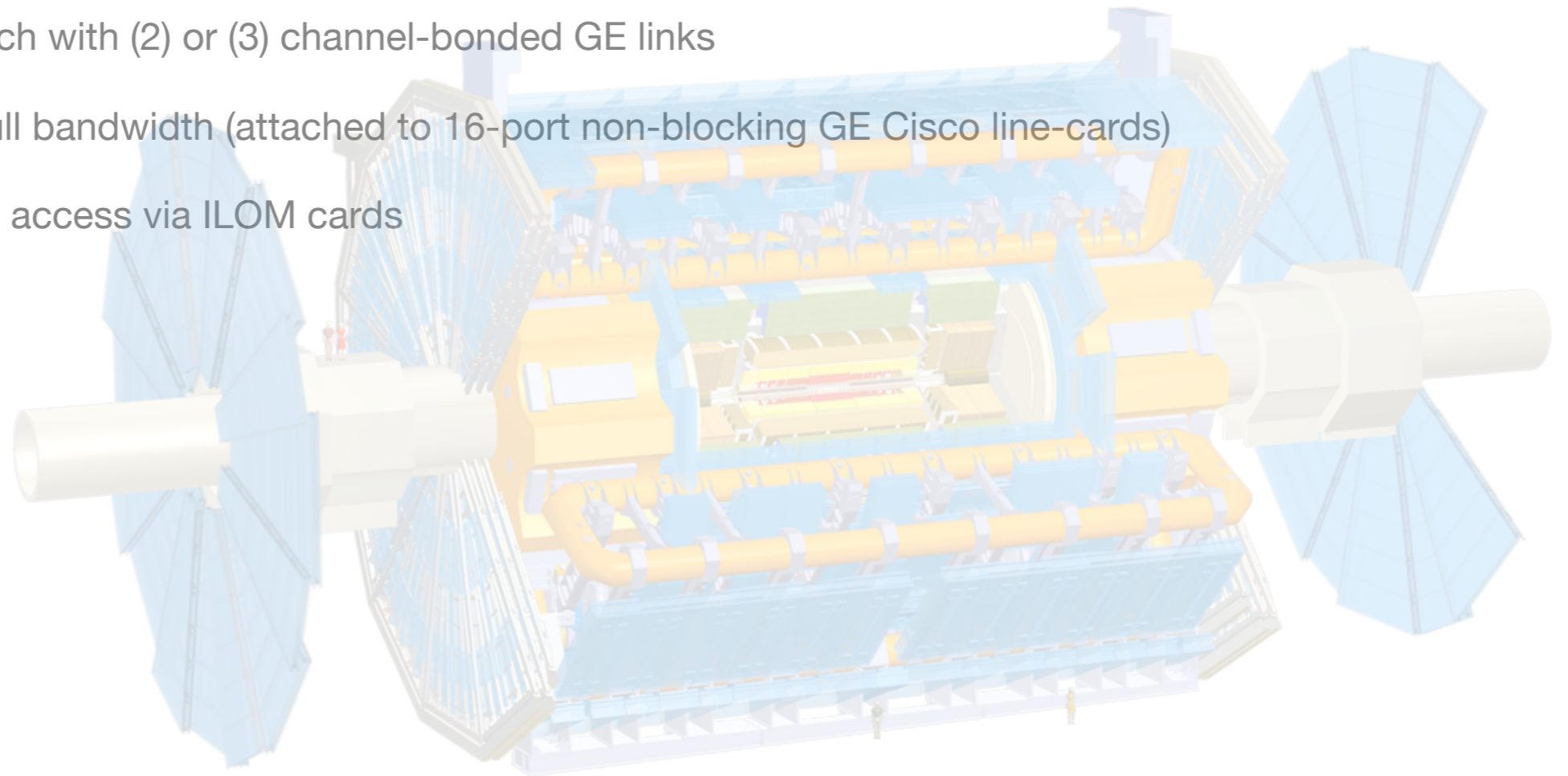
- (28) SunFire x4500 storage servers in (4) APC racks each with (3) metered 2 Phase 30AMP PDU
- System specs:
  - (2) dual-core AMD Opteron 285 processors and (48) Hitachi Deskstar 500 GB SATA II drives
  - (6) paired SATA controllers connected to HyperTransport PCI-X 2.0 tunnels
  - (2) dual-Gb ethernet controllers
  - 16 GB RAM
- Solaris 10 update 3 (11/06)
- Java 1.6 64-bit
- Disk configuration as follows:
  - (2) SVM mirrored system disks, UFS file system
  - A single ZFS RAID 60 storage pool and file system. (4) striped RAIDz2 sets in (11+2) and (12+2) bundles
  - (1) hot spare



# 2007 Distributed Storage Deployment

---

- (5) systems configured as write pool nodes, each with (4) LACP channel-bonded GE links
- (23) read pool nodes each with (2) or (3) channel-bonded GE links
- All connections are at full bandwidth (attached to 16-port non-blocking GE Cisco line-cards)
- Full poweron/off remote access via ILOM cards

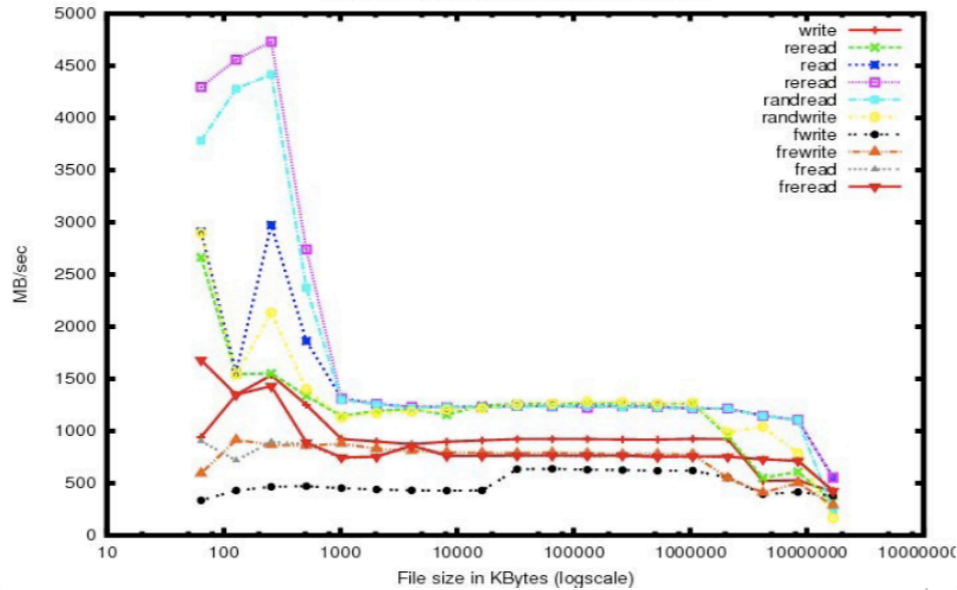




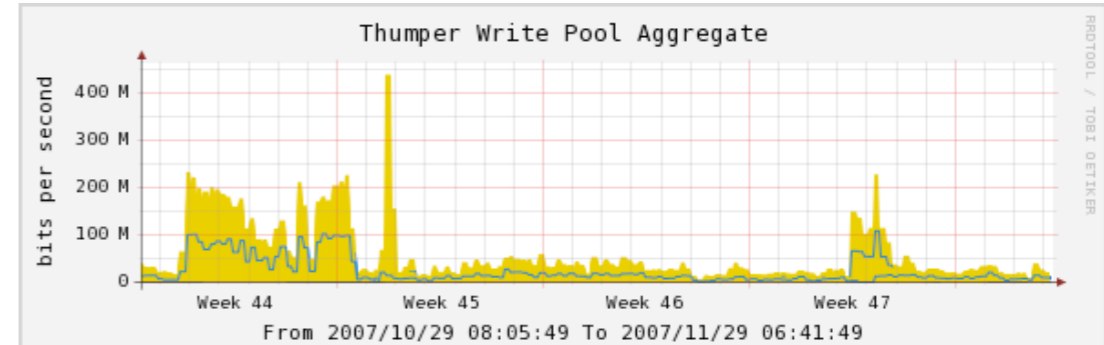
# x4500 Performance Profile

- Local IO tests: >400MB/sec R/W with RAID6
- Controlled dccp test (150 clients each R or W 3x1.5GB files) 200-300MB/sec
- So far, real-world production has not taxed the systems
- There is room to grow!

Aggregate Iozone Performance: ThumperZFS



Zooming Graph 'Thumper Write Pool Aggregate'



sw22 Inbound  
 sw22 Current: 15.03 M  
 sw22 Average: 42.83 M  
 sw22 Maximum: 210.20 M

sw22 Outbound  
 sw22 Current: 11.52 M  
 sw22 Average: 21.37 M  
 sw22 Maximum: 101.11 M

sw22 Inbound  
 sw22 Current: 2.64 M  
 sw22 Average: 29.65 M  
 sw22 Maximum: 436.09 M

sw22 Outbound  
 sw22 Current: 1.12 M  
 sw22 Average: 14.33 M  
 sw22 Maximum: 119.08 M

sw22 Inbound  
 sw22 Current: 18.09 M  
 sw22 Average: 43.95 M  
 sw22 Maximum: 223.65 M

sw22 Outbound  
 sw22 Current: 7.93 M  
 sw22 Average: 21.38 M  
 sw22 Maximum: 112.51 M

sw22 Inbound  
 sw22 Current: 15.29 M  
 sw22 Average: 43.90 M  
 sw22 Maximum: 207.61 M

sw22 Outbound  
 sw22 Current: 6.87 M  
 sw22 Average: 22.21 M  
 sw22 Maximum: 112.76 M

sw22 Inbound  
 sw22 Current: 16.43 M  
 sw22 Average: 43.49 M  
 sw22 Maximum: 230.12 M

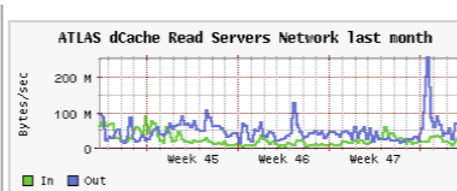
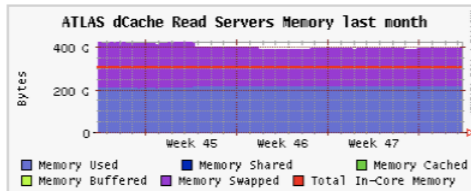
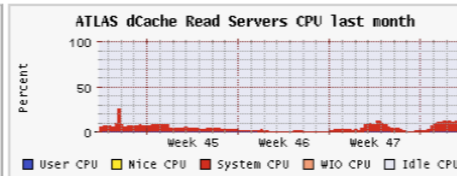
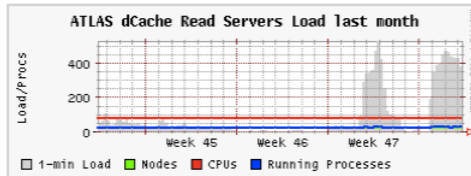
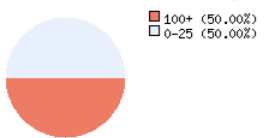
sw22 Outbound  
 sw22 Current: 9.50 M  
 sw22 Average: 21.83 M  
 sw22 Maximum: 102.14 M

Monthly (2 Hour Average)

CPU's Total: 80  
 Hosts up: 20  
 Hosts down: 0

Avg Load (15, 5, 1m):  
 104%, 104%, 104%  
 Localtime:  
 2007-11-29 06:15

Cluster Load Percentages



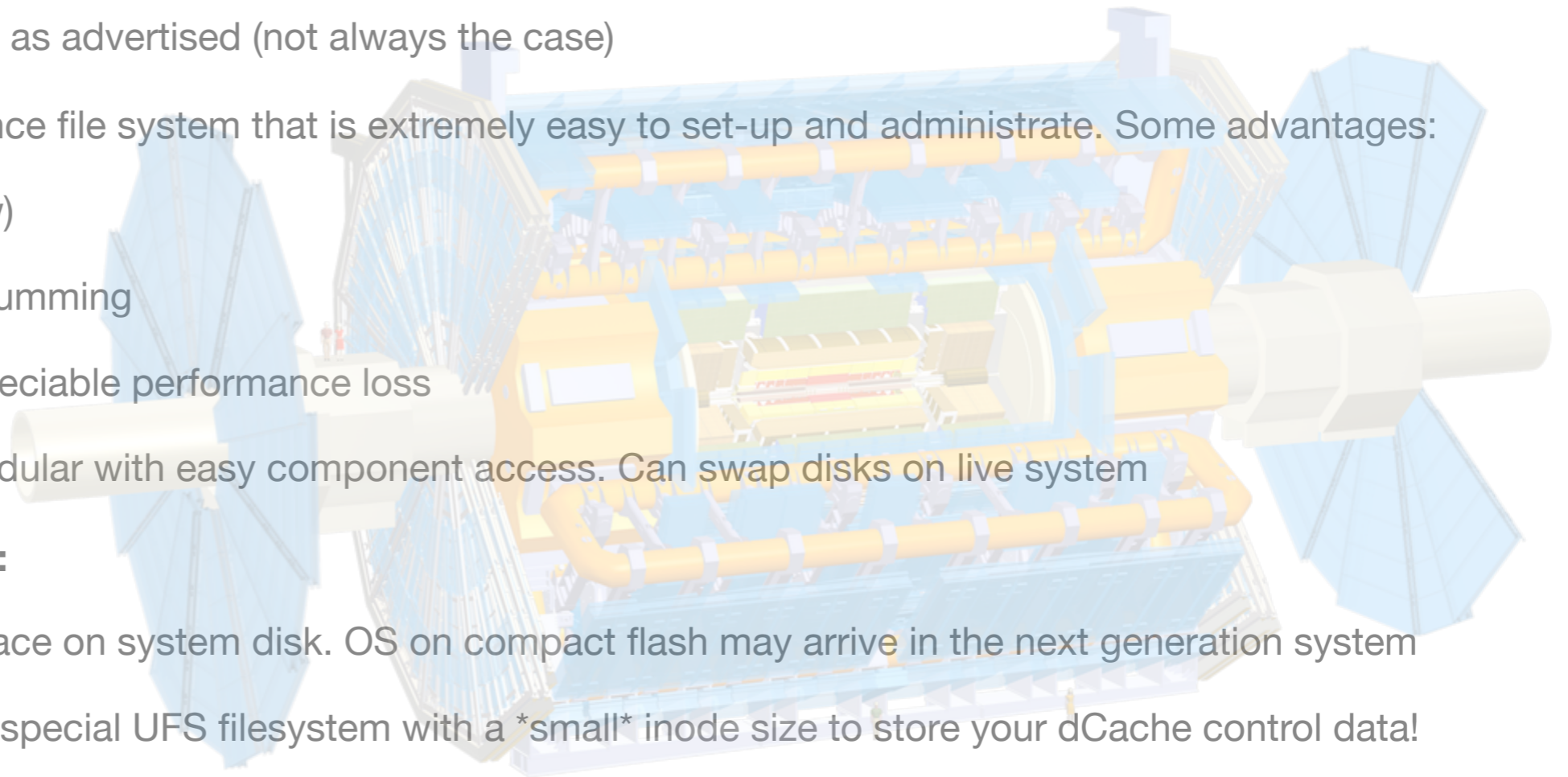
# x4500 Experience

---

- Generally, this is a nice system that is performing well for us
- Channel-bonding works as advertised (not always the case)
- ZFS is a high-performance file system that is extremely easy to set-up and administrate. Some advantages:
  - Copy-on-write (cow)
  - End-to-end checksumming
  - RAID6 with no appreciable performance loss
- The entire system is modular with easy component access. Can swap disks on live system

## Some caveats, though:

- Waste 1-2TB of disk space on system disk. OS on compact flash may arrive in the next generation system
- Make sure you create a special UFS filesystem with a \*small\* inode size to store your dCache control data!
- No way to know if there has been a disk failure unless you set-up the fmadm snmp framework
- No automatic insertion of hot spare into faulty RAID set (!)
- Some infant mortality -- bad memory, CPU, motherboard, disks



# Future of x4500 at RACF

- We were eagerly awaiting the next generation x4500 which would offer

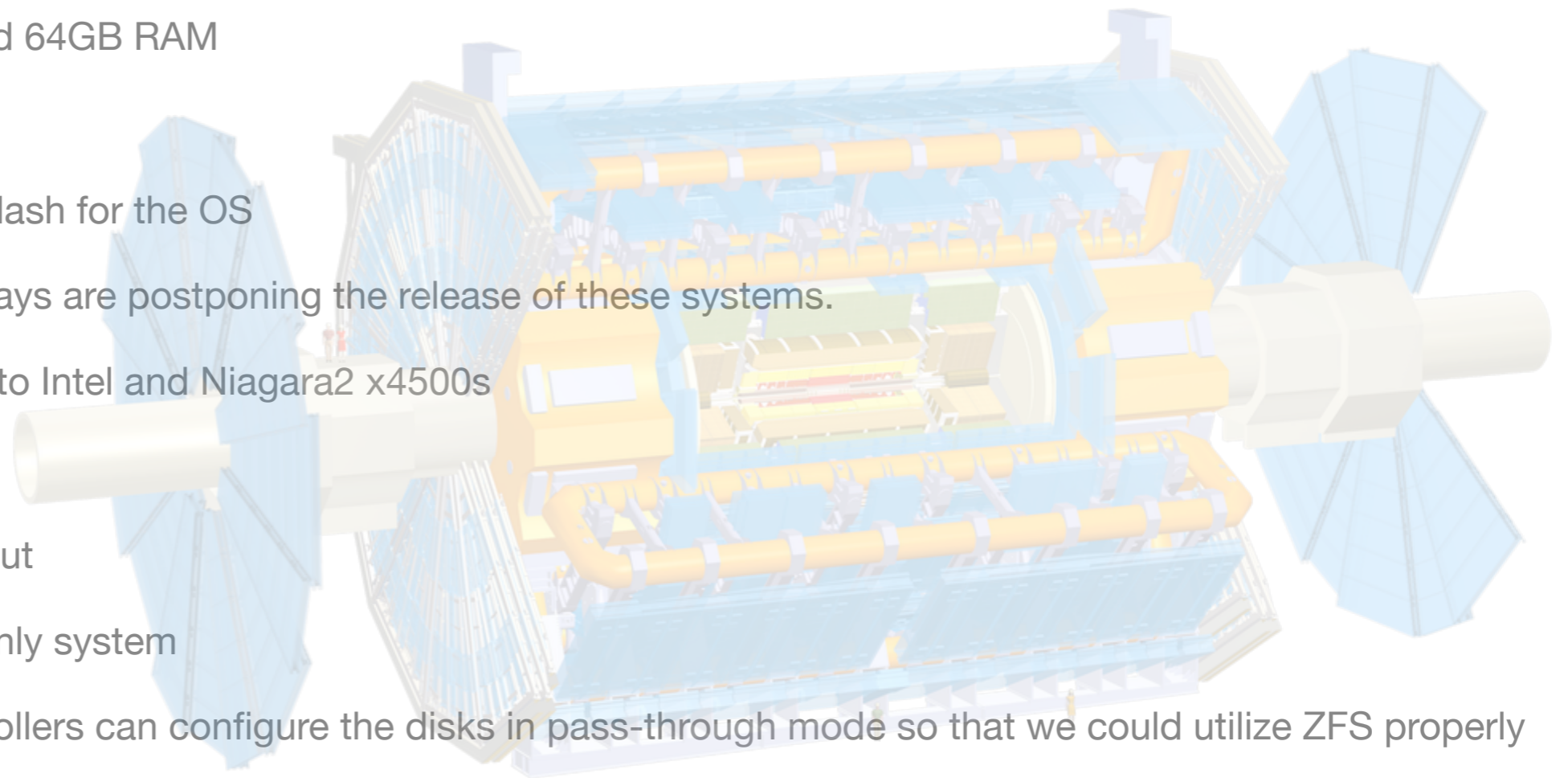
- Dual Quad-core and 64GB RAM
- 1 TB disks
- Possible compact flash for the OS

- But AMD Barcelona delays are postponing the release of these systems.

- SUN should look into Intel and Niagara2 x4500s

## Other options

- Nexsan -- ultra-dense but
  - Stand-alone disk-only system
  - Unclear if the controllers can configure the disks in pass-through mode so that we could utilize ZFS properly
- DDN -- Fast HW RAID6 with no overhead but
  - Incremental upgrades may be expensive
- Actively looking at new technology. May need to be creative.





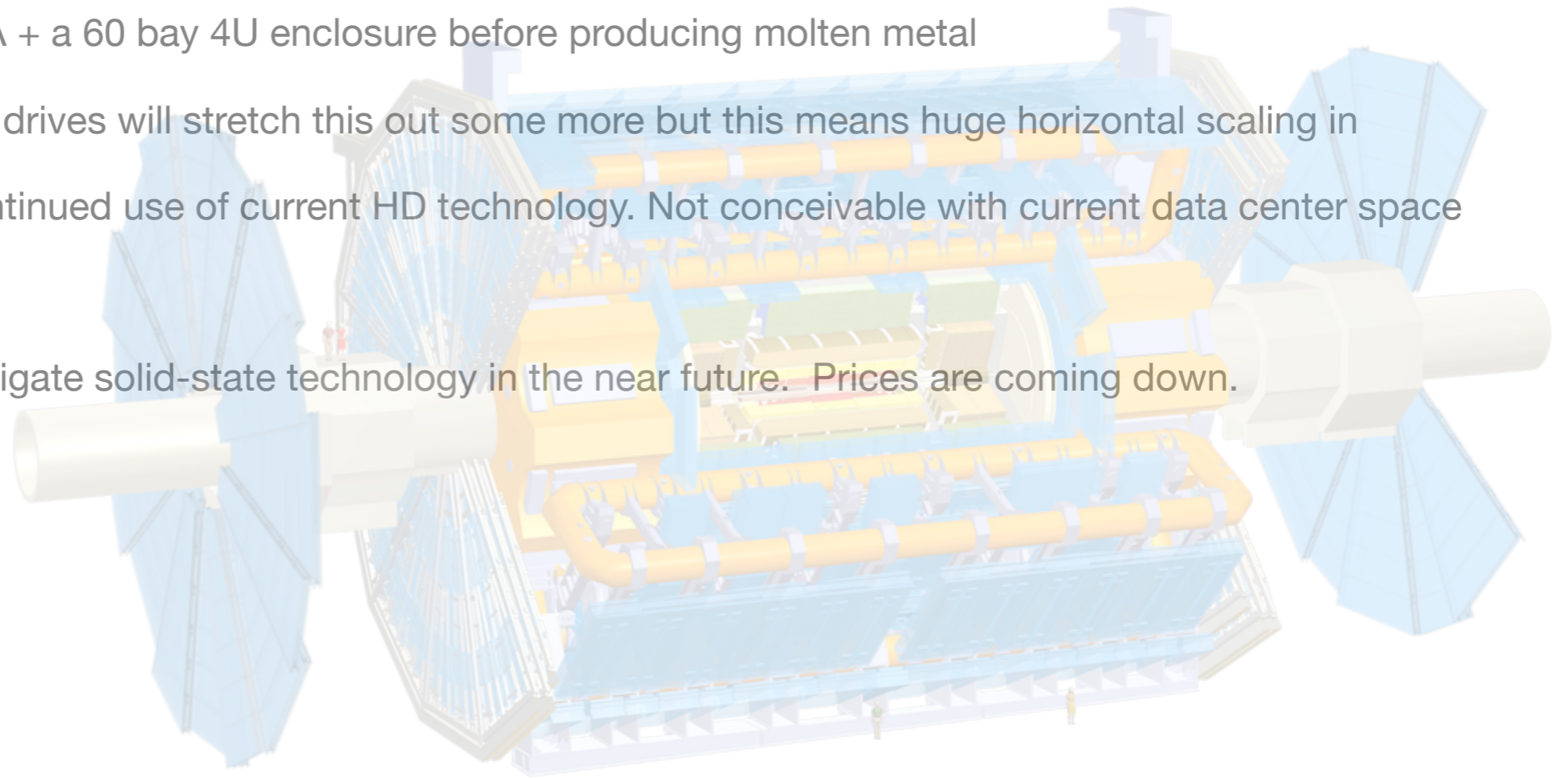
# 5 - 10 Years Down the Road

---

- Disk capacity may cap at 2-3TB per disk so the max amount of storage per rack is probably 1.8PB/rack ??

This assumes 3.5" SATA + a 60 bay 4U enclosure before producing molten metal

- High capacity SAS 2.5" drives will stretch this out some more but this means huge horizontal scaling in the future assuming continued use of current HD technology. Not conceivable with current data center space issues
- It behooves us to investigate solid-state technology in the near future. Prices are coming down.



# Tier 3 Storage Recommendations

---

- Disk-heavy compute nodes are viable options. Recent processor farm purchases yielded <\$6k systems with
  - Dual quad core
  - 16GB RAM
  - 4.5 TB disk (6 TB now available)
- Ideal xrootd / Proof node
- If using Linux, recommend HW RAID 5
- Invest in the right network hardware. Aim for low-latency, high bandwidth
- If space is not an issue, inexpensive and fairly reliable JBODs and RAID arrays can be purchased from a myriad of vendors like Aberdeen and RAIDinc.
- JBODs paired with Solaris/ZFS can offer decent integrity and performance for little \$\$\$
- Tier 3s need to be investigated on a case by case basis. They may already have decent pre-existing infrastructure to work with

