Maurice Askinazi
Ofer Rind
Tony Wong

HEPIX @ Cornell
Nov. 2, 2010

Storage at BNL

Traditional Storage

NFS servers

SAN switches

- Dedicated compute nodes and NFS SAN storage
- Simple and effective, but SAN storage became very expensive as facility grew
- Not a scalable solution

Fibre Channel Disk Storage

A New Model

- Having thousands of compute nodes in our facility provided an opportunity to install large disks internally. This is an inexpensive solution to provide data storage capacity. The task was to figure out how to use this storage for production. STAR collaboration accesses this storage with Xrootd and PHENIX use dCache.

- Having reduced the facility demand for network storage, its use became more focused, and spending a little more to improve its performance became possible.

- NFS use: Home directories (user environment), high value files which are backed up, work group scratch space

Linux Farm

<u>RHIC</u>
1200 (+279) machines
7500 (+2232) cores
2.25 (+1.6) PB storage

<u>ATLAS</u>
900 machines
5800 cores

RHIC production uses local farm storage:
  STAR accesses this storage with Xrootd, PHENIX uses dCache
ATLAS dCache uses dedicated storage servers

Centralized NFS Storage

BlueArc (hardware NFS)

As the large demand for production storage was redirected to distributed local storage, the need for other NFS filesystems, smaller in size and more targeted in use, became evident.

Requirements of any new NFS solution were, the ability to have many concurrent connections and to backup large amounts of data to tape.

Our first purchase of two Titan 2200 servers was in 2007, which was quickly upgraded to three Titan 3200s the next year

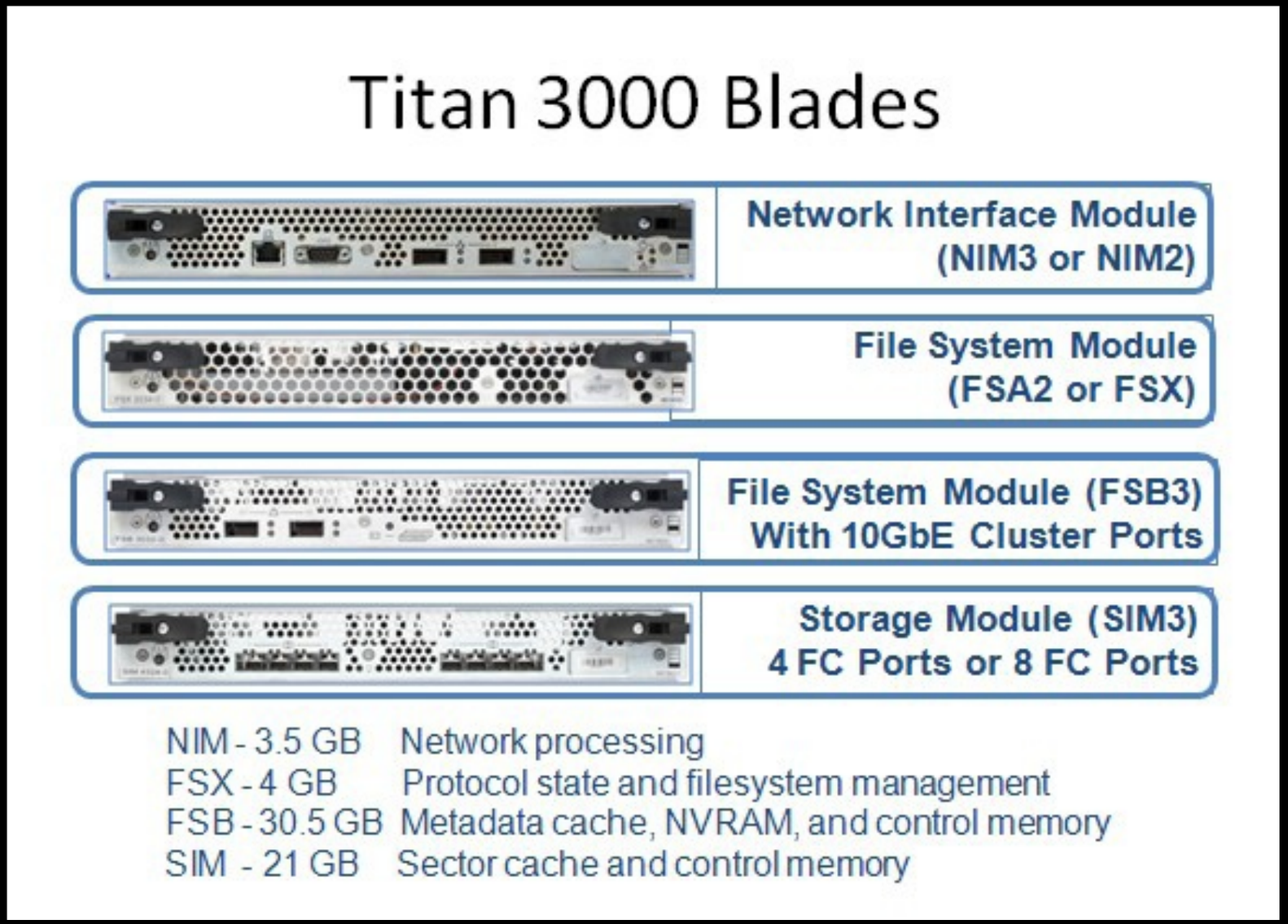
Titan 3000 Series

High Performance Network Storage

	3100	3200
IOPS	100,000	200,000
Throughput	Up to 10Gbps	Up to 20Gbps
Scalability	Up to 2 PB in a single namespace	Up to 4PB in a single namespace
Ethernet Ports	2 10GbE or Six GbE	2 10GbE or Six GbE
Fibre Channel Ports	Four 4Gb	Eight 4Gb
Clustering Ports	Two 10GbE	Two 10GbE

Comparison of Titan Network Storage Performance

Titan 3000 Blades



The BlueArc server model passes data through boards with specialized processors, rather than expecting a server cpu to do it all.

All The Performance Money Can Buy?

The Titan was working well, but could we get by with a smaller server?

<u>BlueArc list price</u>:

Titan 3100 is $97,500
Titan 3210 is $150,000

Mercury 50 is $45,000
Mercury 100 is $65,000

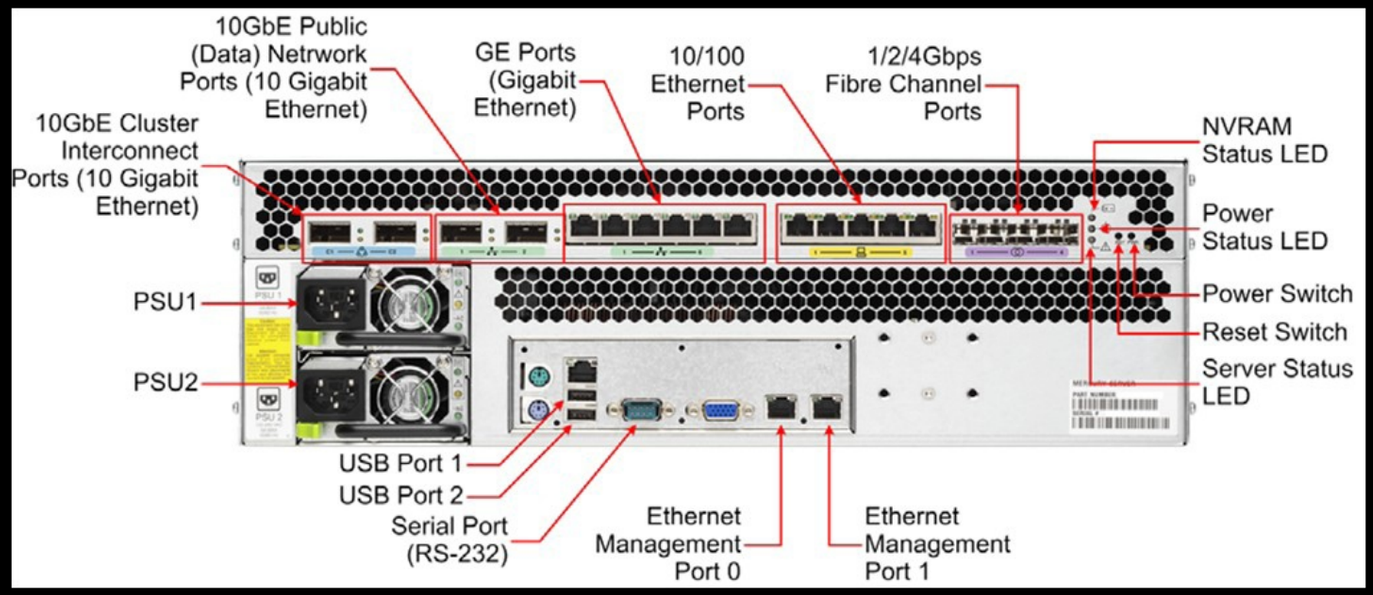
BlueArc Platforms Comparison

	Mercury 50	Mercury 100	Titan 3100	Titan 3200
Product Class	Lower Mid-range	Mid-range	Mid-range	High End
Cluster Nodes	2	Up to 8	Up to 8	Up to 8
Max Storage Capacity	1 PB	2 PB	2 PB	4 PB
Max Filesystem Size	256 TB	256 TB	256 TB	256 TB
NFS Throughput	700 MB/s	1100 MB/s	1200 MB/s	1900 MB/s
Performance (Ops/Sec)	60,000	100,000	100,000	200,000
Storage Options	All BlueArc storage array options are available with each platform			
Software / File Services	All software and file system options (NFS, CIFS, iSCSI) available			

Mercury Architecture

Main Mother Board (MMB)

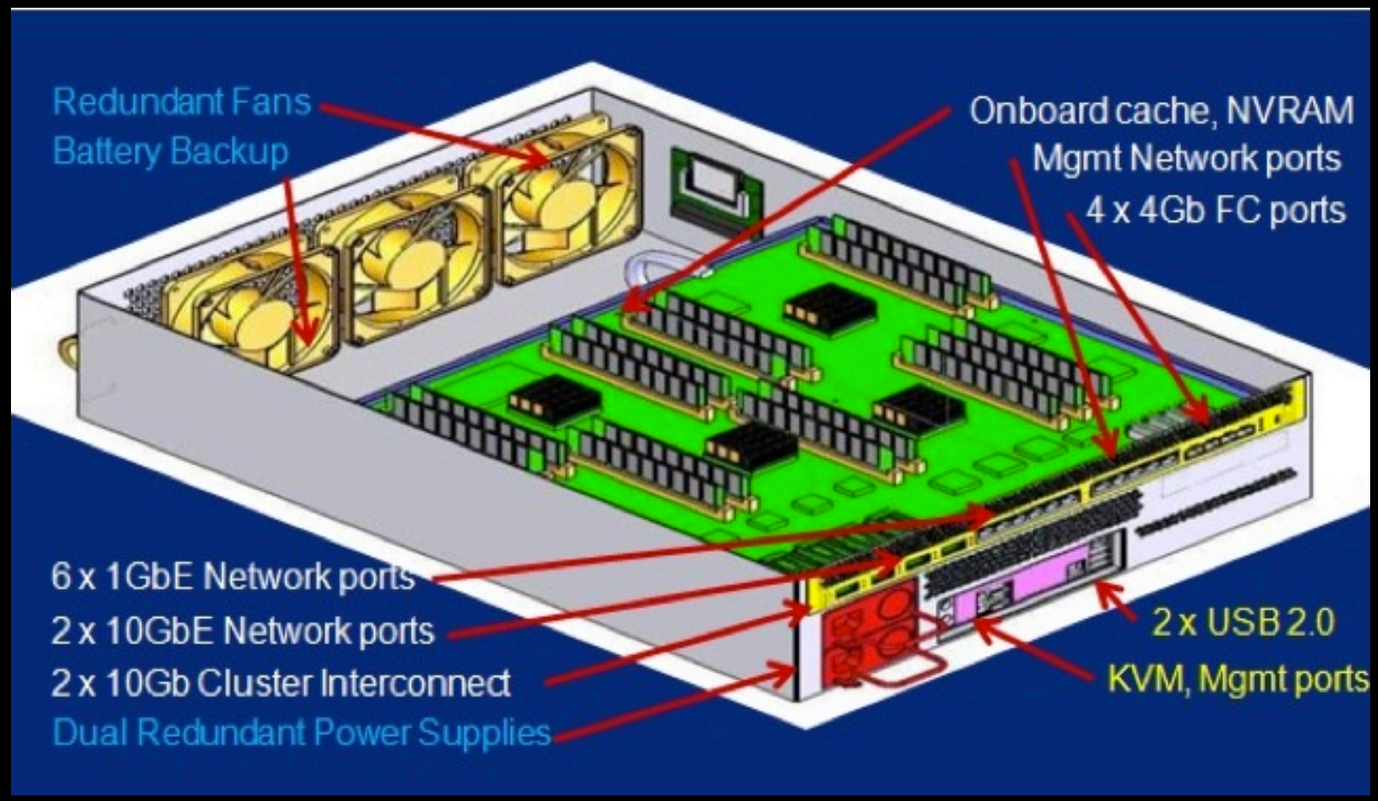
- Standard X86, Dual-Core 3Gz
- 64 bit Linux
- 2 x 2.5” HDD (SW RAID 1)
- 10/100/1000 Ethernet
- 8G mem (Mercury 100)



Main FPGA Board (MFB)

- Single custom PCB, similar in size to MMB
- PCIe (4 lane) connection to MMB
- 6 FPGAs (replacing 13 in Titan)
- 24G mem (Mercury 100)

Hybrid solution with HW acceleration for filesystem data movement and metadata processing



Current Usage

RHIC:

2 (3 x Titan 3210)* Clusters – 20 LSI storage arrays
- 17 x RC16TB contollers with 300GB FC disk (RAID5)
- 1 x RC16SA controller with 1TB SATA disk
- 2 x RC12 controllers with 2TB NLSAS disk

937 raw TB

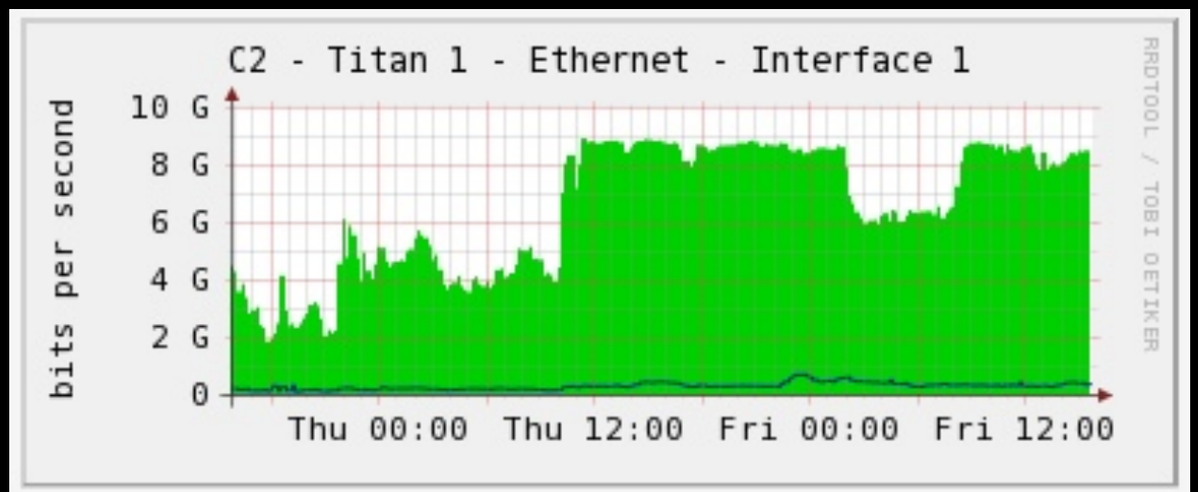
ATLAS:

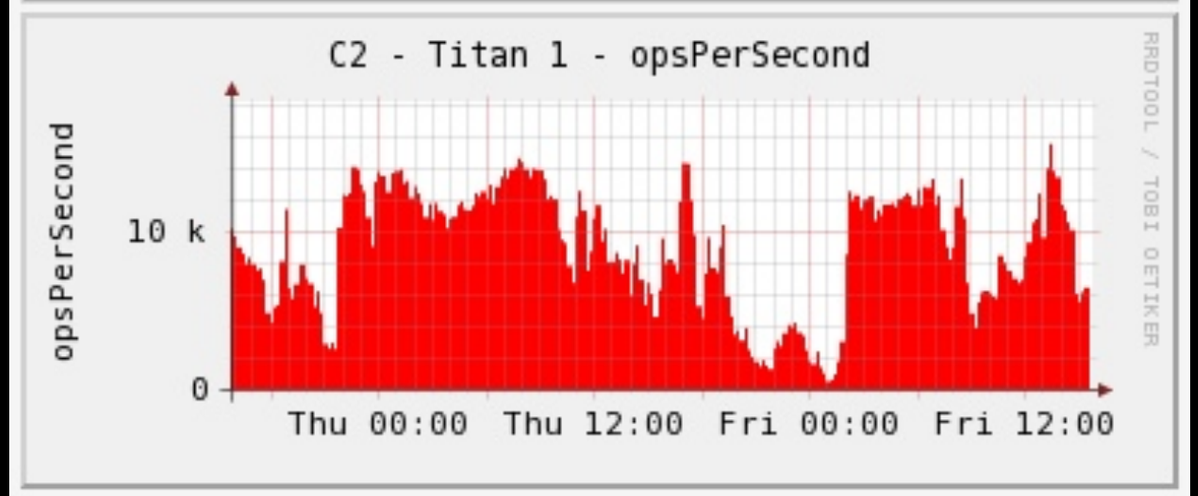
1 (2 x Mercury 100) Cluster – 4 LSI storage arrays
- 2 x RC12 controllers with 450GB SAS disk
- 2 x RC12 controllers with 1TB NLSAS disk

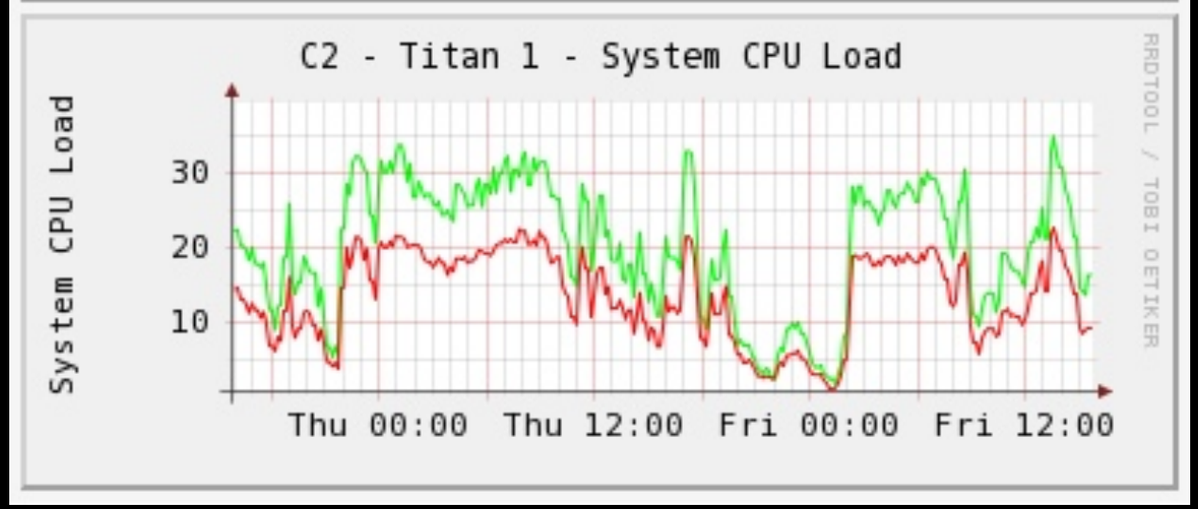
128 raw TB

* Second cluster necessitated by 128 LUN limit

Titan Performance at RACF

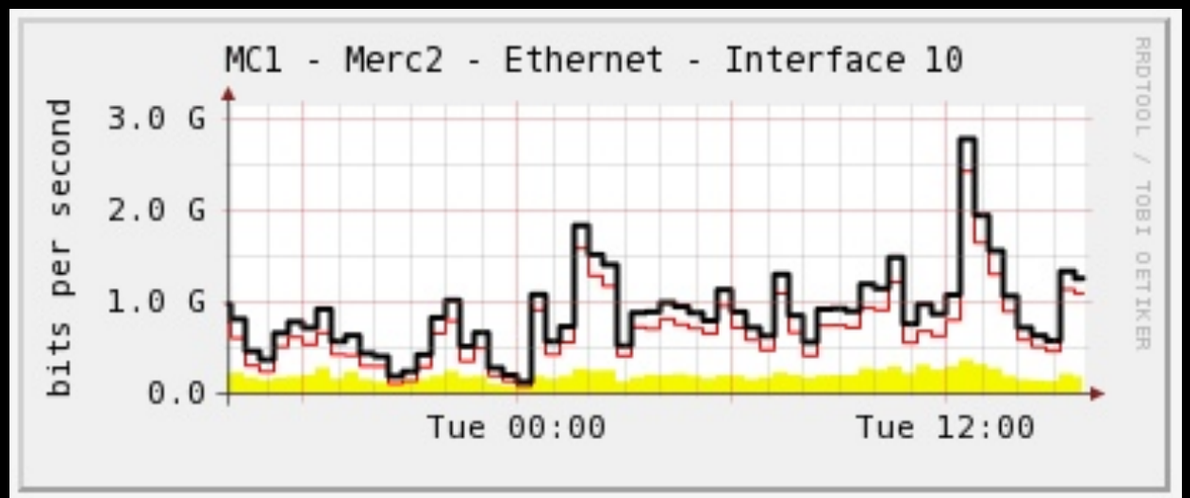


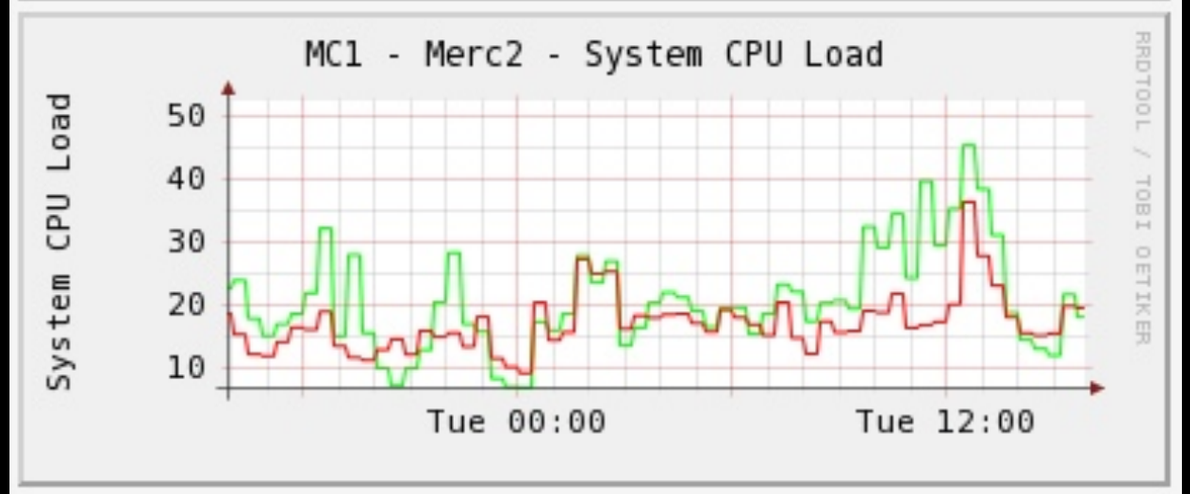


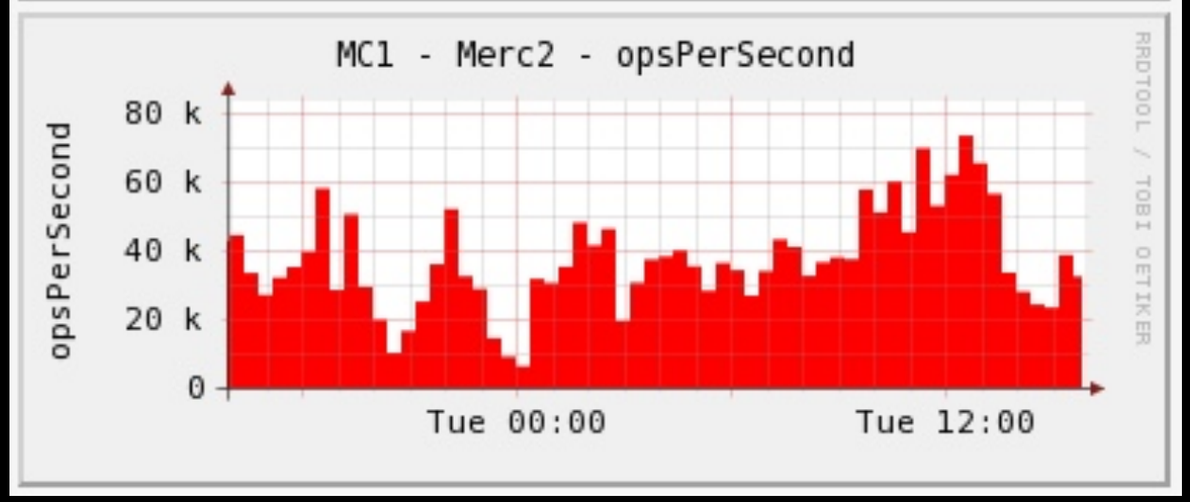


As an example of performance, these SNMP graphs show a Titan server sustaining 9 Gb/s bandwidth, 12,000 ops/sec while system load < 30%

Mercury Performance at RACF

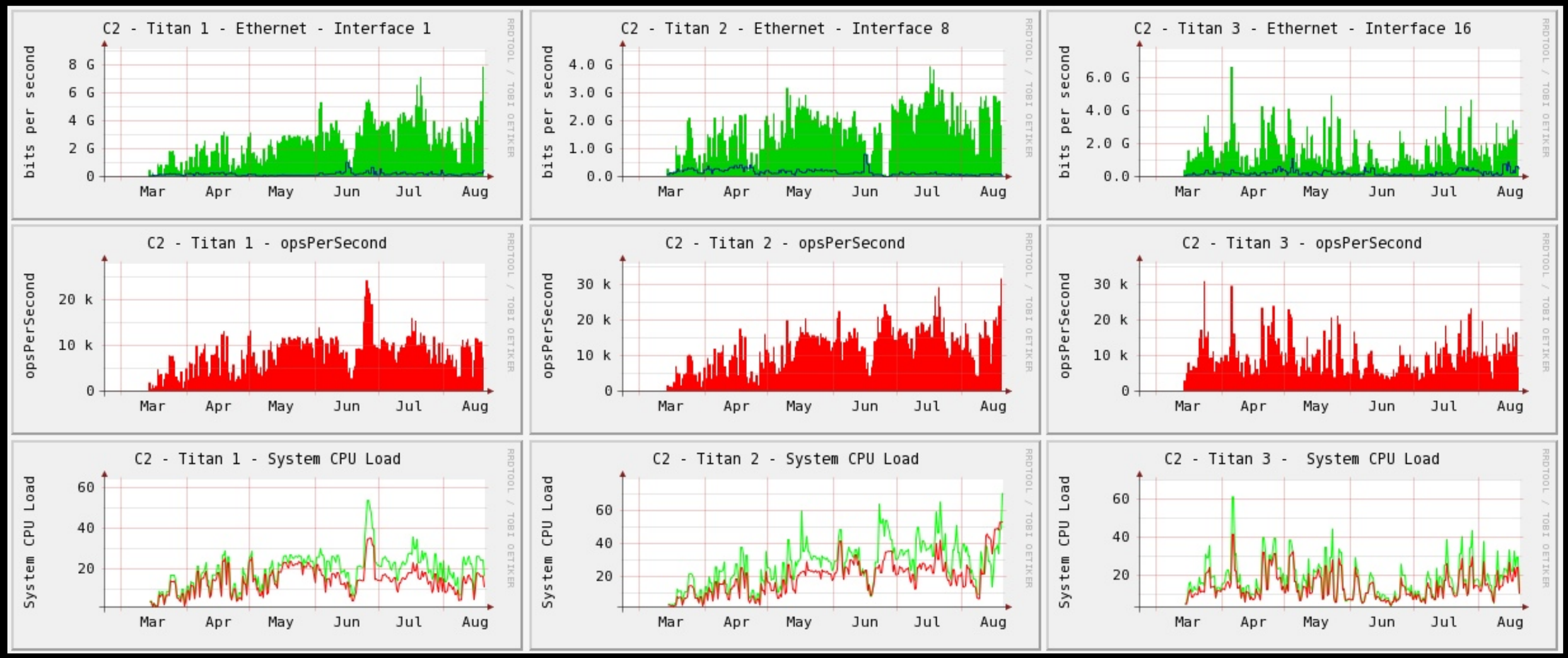






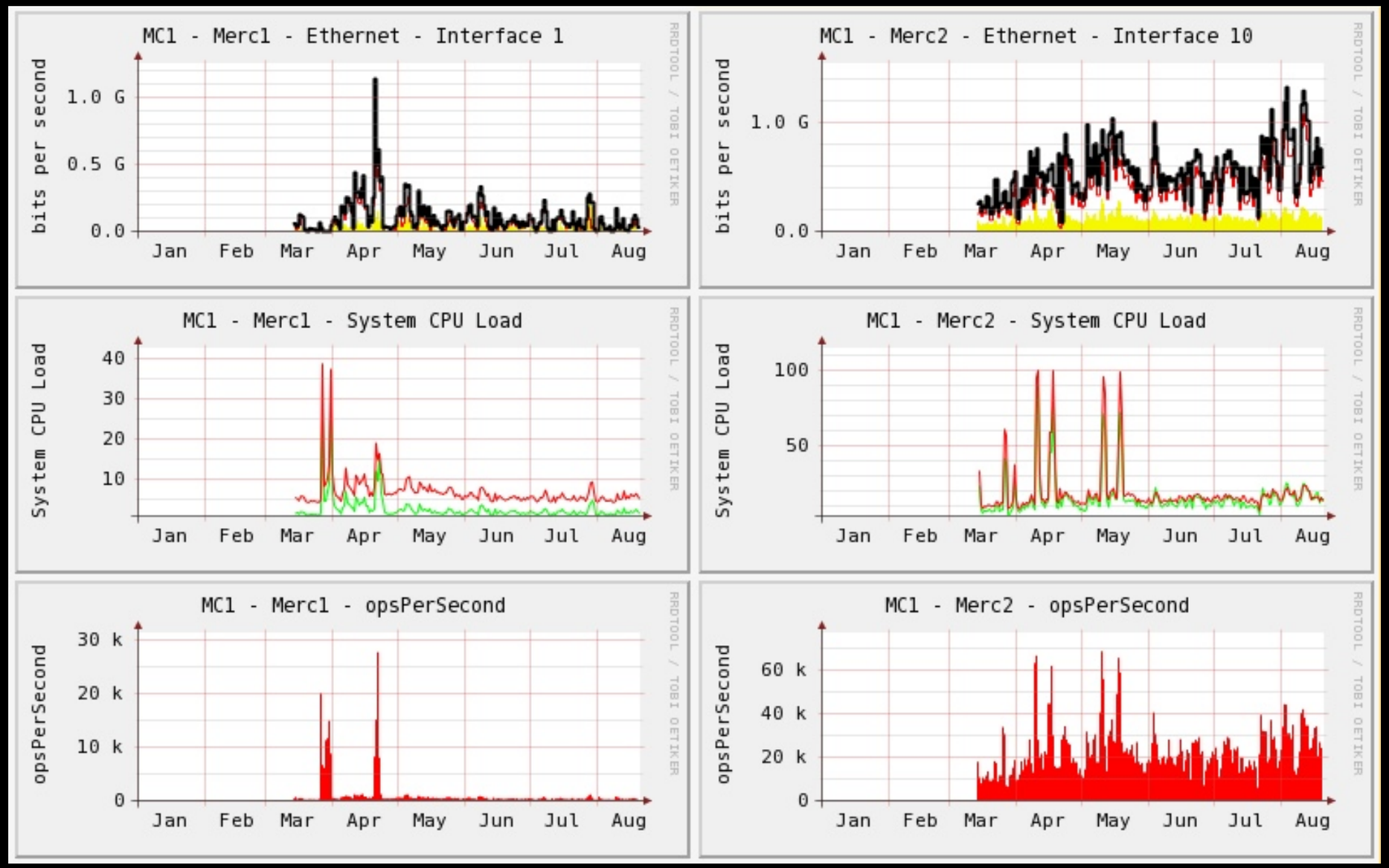
As an example of performance, these SNMP graphs show a Mercury server sustaining 2Gb/s bandwidth, 60,000 ops/sec while system load < 40%

Titan cluster uptime



Our 3 server Titan Cluster. It has provided uninterrupted service since installation. Performance graphs go back to March 2009 when our SNMP database was reset.

Mercury cluster uptime



Our 2 server Mercury Cluster. It has had uninterrupted service for the 6 months since installation.

Future

Though the Titan server is made to deliver greater performance, the current Bluearc operating system limits the amount of storage the cluster can mount to 128 LUNs.

Without being able to add more storage to the cluster, some of the server performance may go unutilized.

If the Mercury performance meets expectations for serving the desired amount of storage, then it is a better buy.

BlueArc will be providing 512 LUN support in November. That will expand the capacity of the Mercury 100 from 2PB to 8PB, and the Titan 3210 from 4PB to 16PB.

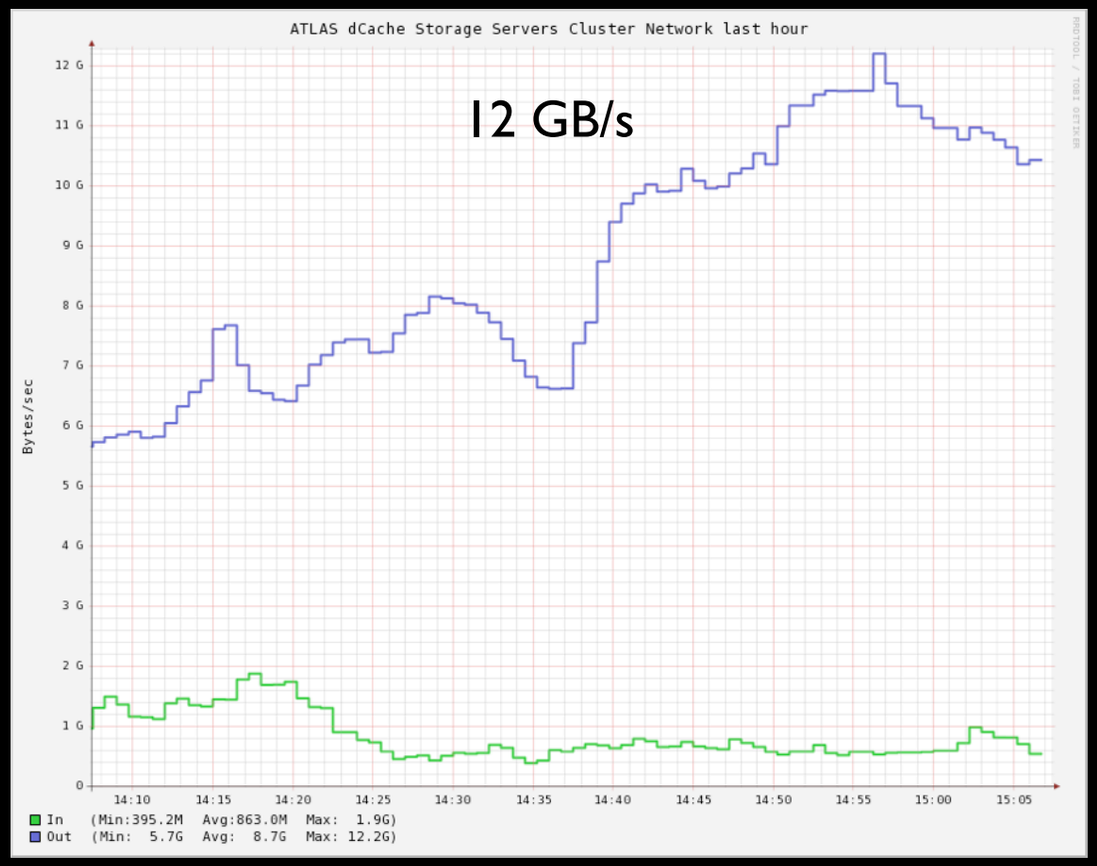
Distributed Data Storage

Distributed Storage: Overview

- US ATLAS dCache Storage element
  - 4.5 PB on ~100 Sun/Nexsan Servers (up to 120 TB behind one server)
  - 2 PB DDN served by 4 Dell R710
  - Recently added 1.3 PB on 19 IBM x3650 M2 servers w/Nexsan storage (Solaris 10 + ZFS)
    - Choice driven by cost and some hardware concerns
- PHENIX dCache
  - 1.1 PB local storage (up to 8 TB/server) + some dedicated storage servers (Aberdeen, etc.). Soon to be 2 PB
- STAR Xrootd
  - 1 PB local storage. Moving toward 2 PB soon.
- Daya Bay, LBNE, ATLAS analysis
  - Small scale Xrootd deployments on local storage

DDN Storage

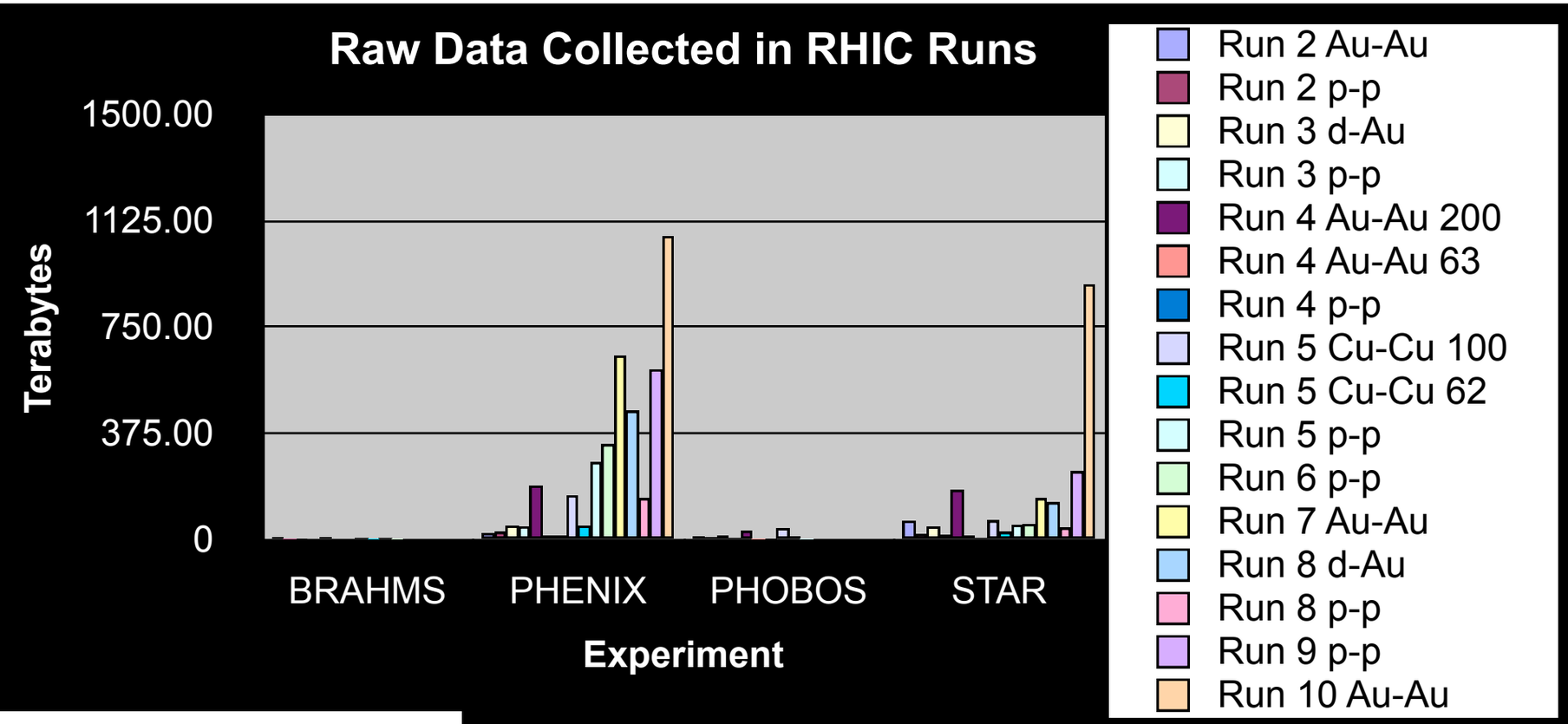
- 1200 2TB SATA (7200 RPM) drives
  - RAID6 (8+2); 120 x 14.5 TB luns
- Dual S2A9900 controllers
- 4 Dell R710 Servers
  - 48G RAM; RH5.5 + XFS
  - Mirrored SSD for journal log
  - Myricom 10G NIC
- 5 X 88 TB dCache pools per server
- Up to 1.5 GBps measured throughput per server
- ReadOnly (data migrated)

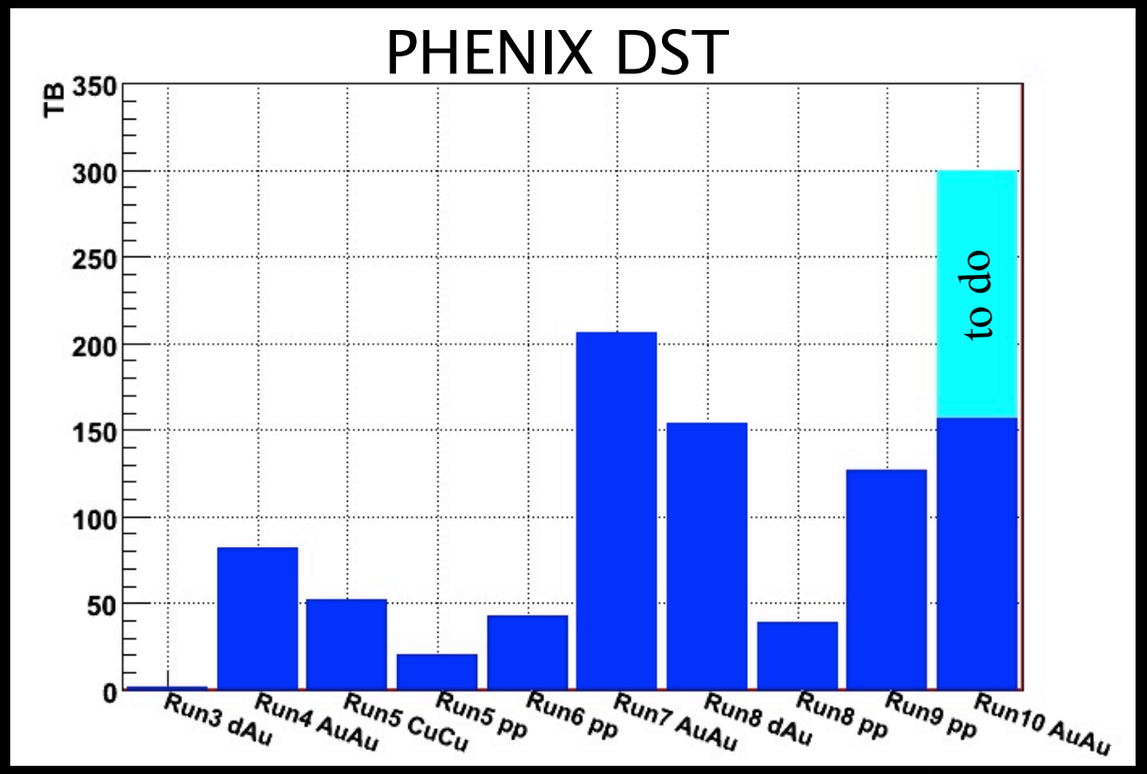


Example of dCache throughput after addition of DDN storage

RHIC Data Volumes

PB sized raw data sets in Run 10 – should now be the norm





C. Pinkenburg

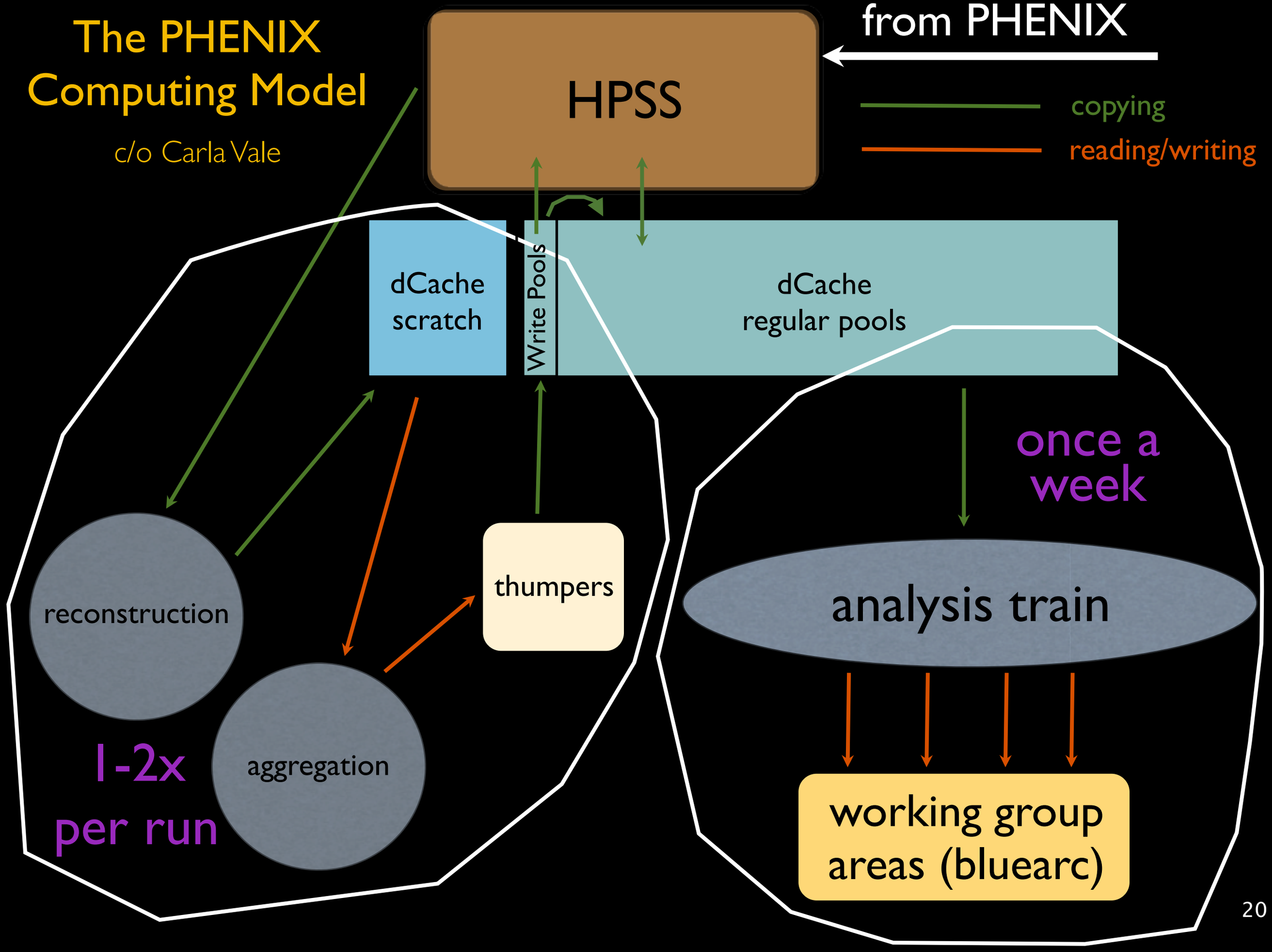
Reconstructed data: 700+ TB

File aggregation – now 9 GB file sizes

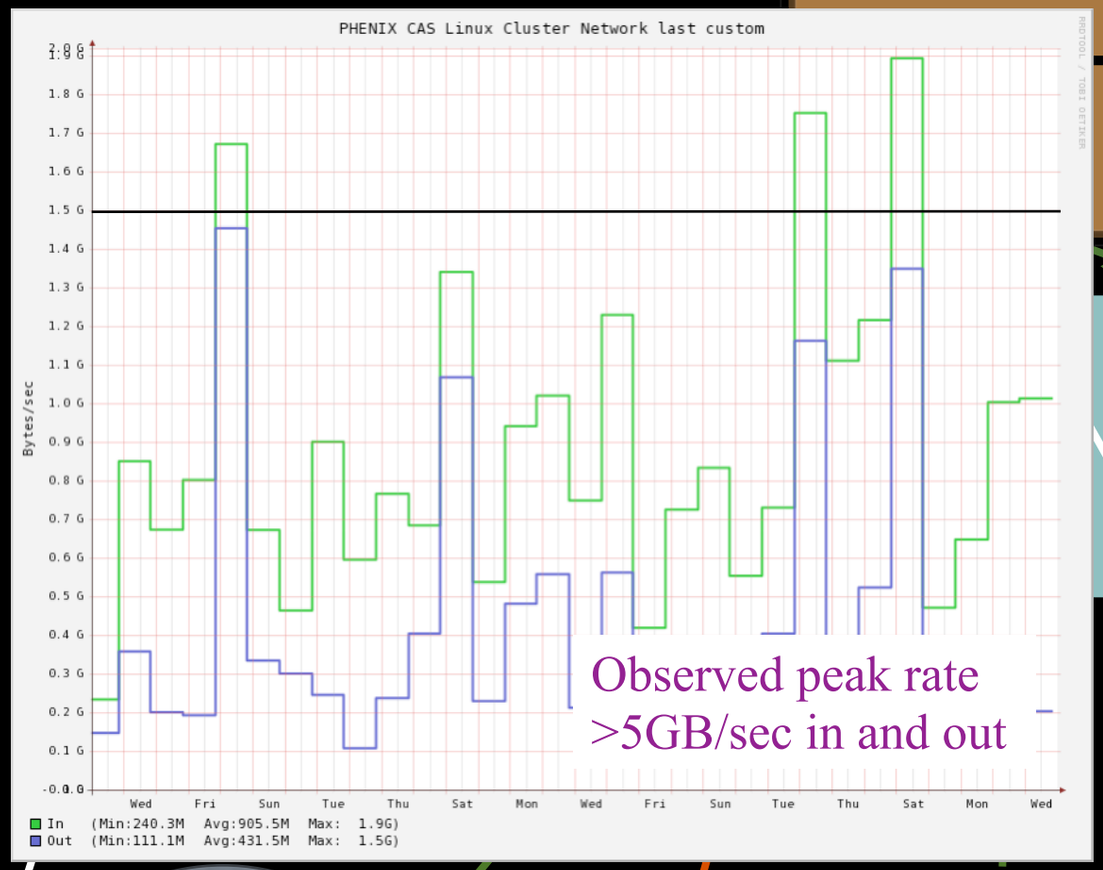
Weekly passes over full data sets – "Analysis Train" model (jobs split and grouped by data subset for local I/O)

The PHENIX Computing Model

c/o Carla Vale



The PHENIX



Observed peak rate
>5GB/sec in and out

C. Pinkenburg

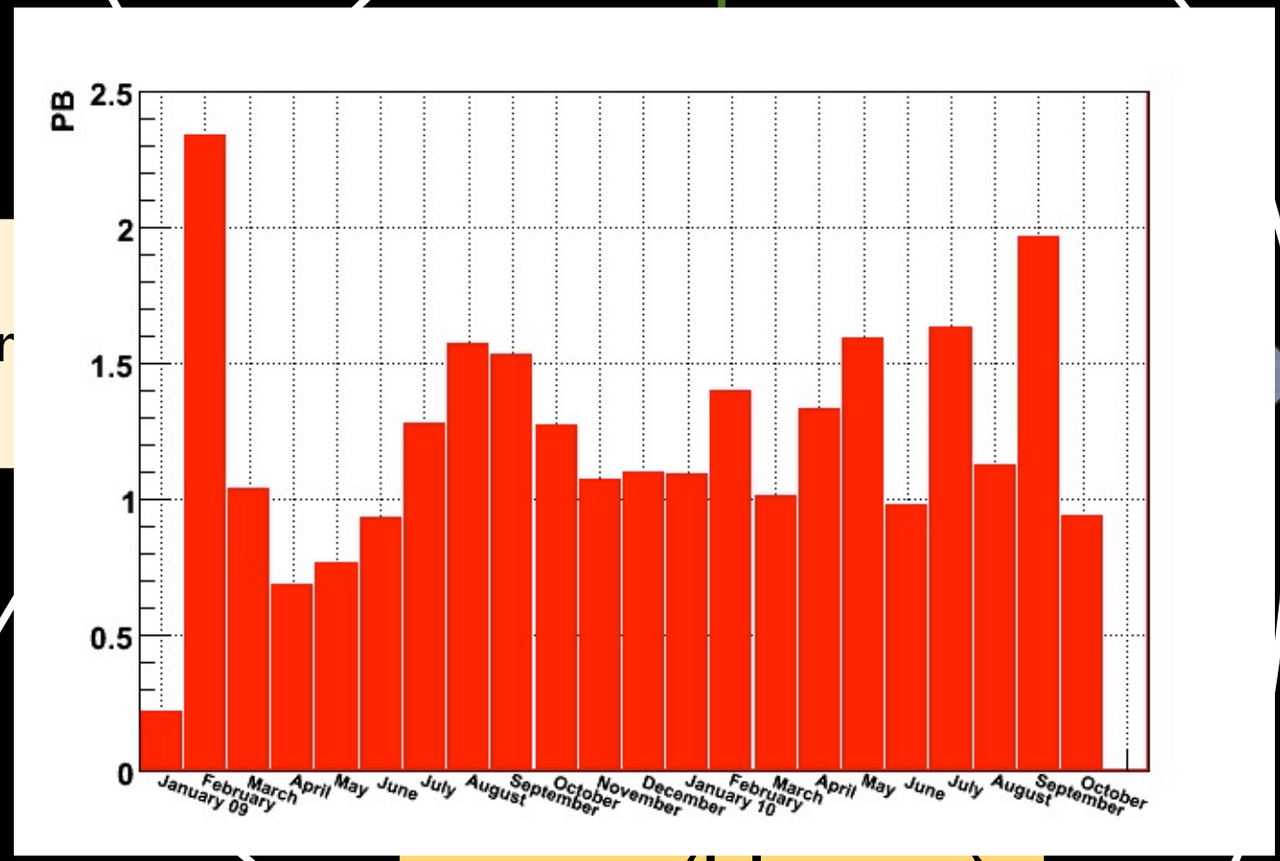
HPSS

from PHENIX

copying

reading/writing

dCache
regular pools



C. Pinkenburg

areas (bluearc)

thumper

reconstruction

aggregation

1-2x
per run

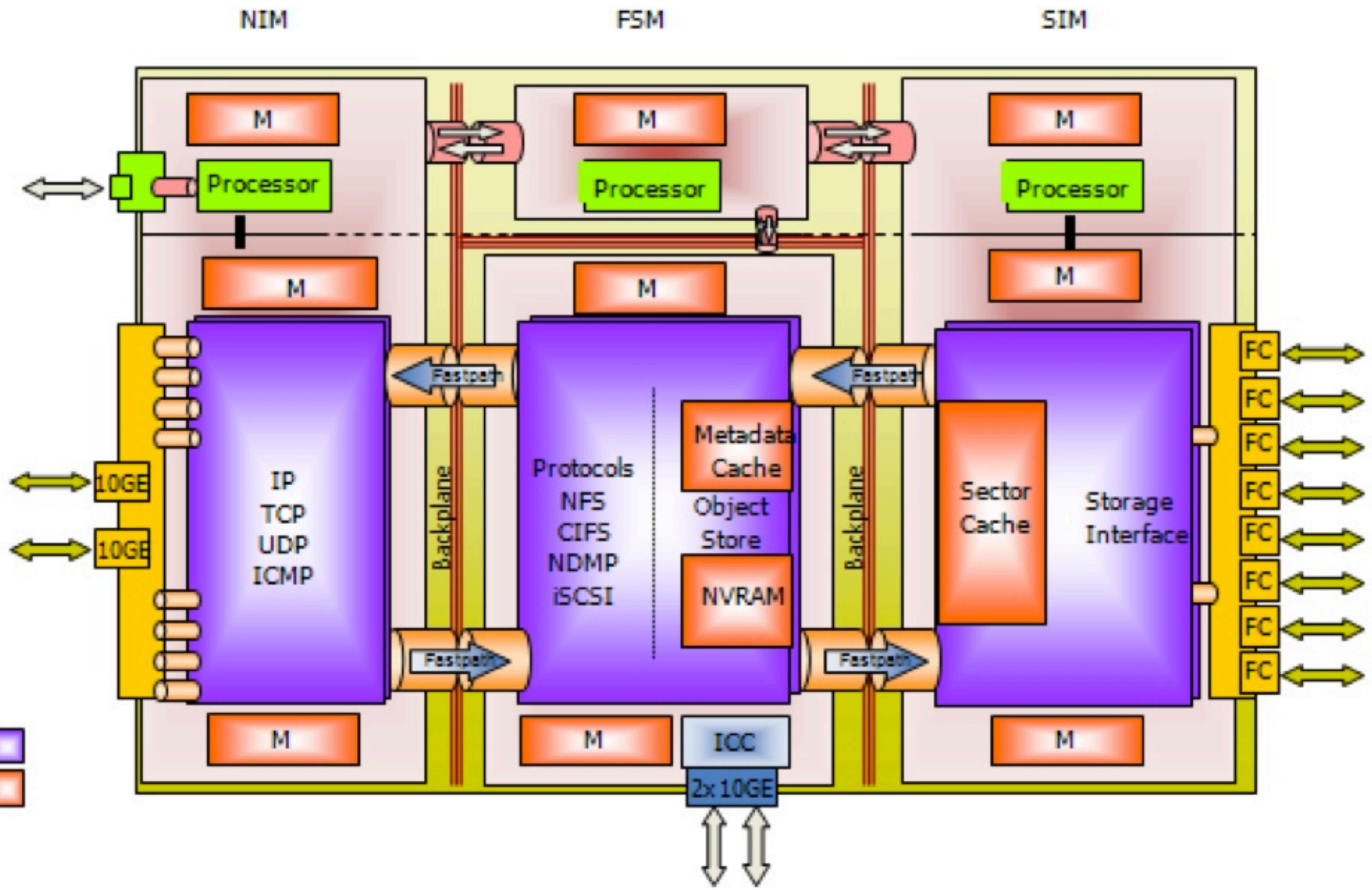
Conclusions and Plans

- Distributed storage performing well within current computing models
- Deploying alternatives to Sun/Oracle storage solutions for ATLAS
- PHENIX and STAR expanding within the farm hybrid model
  - At what point does increasing core density break the scaling?
  - When does local storage become the bottleneck?
- Increasing interactive end-user analysis
  - Growing use of PROOF farm for ATLAS
- Expanding storage solutions for other experiments

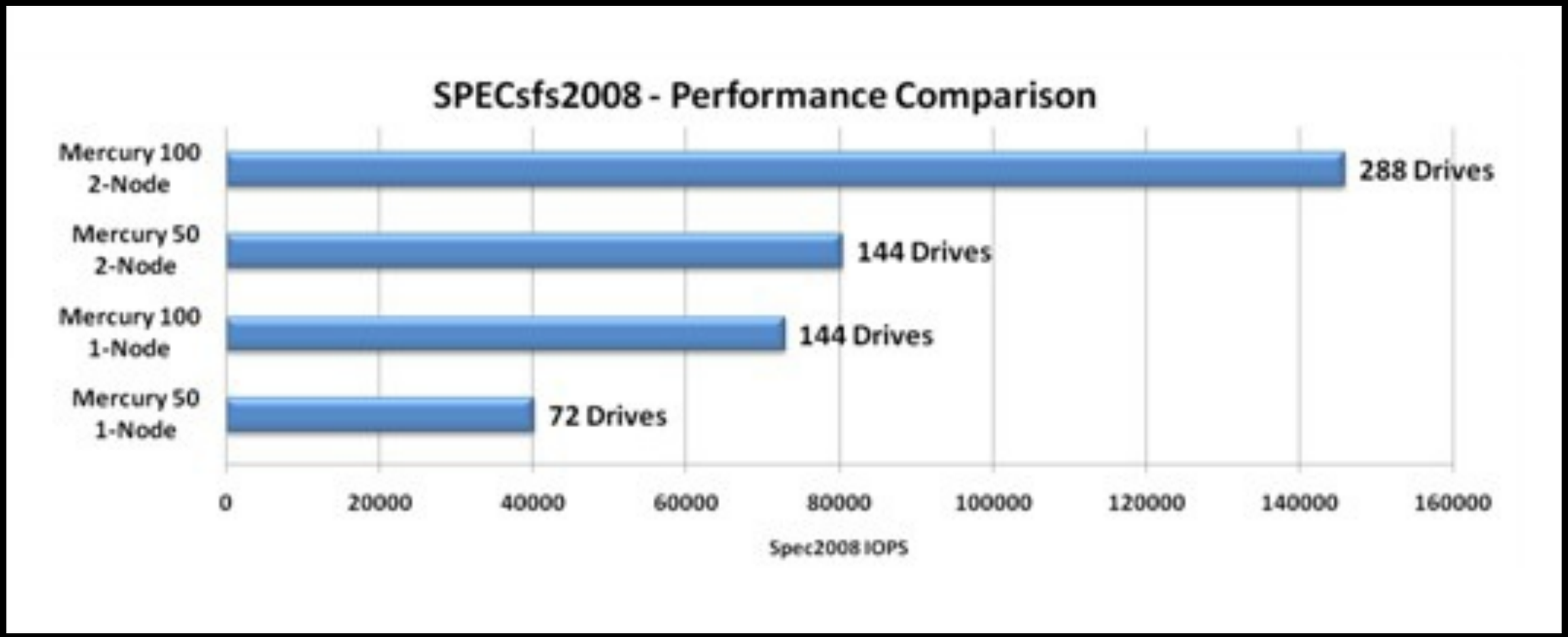
Questions?

Extra Slides

Titan System Architecture



Mercury Benchmarking Data



ATLAS dCache Architecture