The Use and Integration of Distributed and Object-Based File-Systems at Brookhaven National Laboratory

CHEP 2006
Tata Institute of Fundamental Research
Mumbai, India

Robert Petkus, Maurice Askinazi, David Free, Bruce Gibbard, Jerome Lauret, Zhenping Liu, Ofer Rind,Thomas Throwe, Yingzi Wu

RHIC/USATLAS Computing Facility
Brookhaven National Laboratory

RCF Fileserver Overview

RHIC Computing Facility

- 2000+ dual-CPU analysis/reconstruction Linux farm
- 680 TB local disk storage (SCSI, SATA, and PATA)
  - 132 TB xrootd storage on 650 nodes utilized by STAR
  - 25 TB dCache storage on 128 nodes (pool servers) utilized by PHENIX
- 220 TB FC RAID5 NFS storage on 37 Sparc/Solaris servers, Veritas 4.0 (VxVM, VxFS)
- 100 TB object-based storage on 20 Panasas shelves
- 6 TB FC AFS storage on 3 Solaris servers

ACF Fileserver Overview

USATLAS Computing Facility

- 330+ dual-CPU Linux farm
  - 146 TB local SATA disk storage dedicated to dCache
  - 322 internal/external read pool nodes
  - 8 internal/external write pool nodes
- 20+ TB FC RAID5 NFS storage on 4 Sparc/Solaris servers, Veritas 4.0 (VxVM, VxFS)
- 1 TB FC AFS storage on 3 Solaris servers

Fileserver Criteria

- Fast, scalable, reliable, and fault tolerant
- Load balancing
- Security and centralized management
- Single, global namespace
- Transparent, uniform, "POSIX-like" data access

Different Implementation Philosophies:

- Hardware vs. software
- Central vs. distributed
- Existing protocol vs. new protocol
- Proprietary vs. open

NFS Centralized Storage

NFS, Solaris, Veritas

- NFS consistently performs reliably on Solaris 9
- Network data transfer rate 70-80 MB/sec with single 1 Gb NIC
- Ubiquitous, compatible, and mature

However,

- No load balancing, poor scalability, and insecure
- Veritas is expensive and unpredictable in the midst of a host of more able, free competitors; e.g., XFS

Usage at RHIC and USATLAS

- Home directories
- Container for reconstructed data, job output
- Predominantly used for read access or scratch space

AFS Centralized Storage

OpenAFS

- Secure, reliable distributed filesystem well suited as a repository for static data and WAN access
- Replication creates redundancy and increases read performance
- Bottleneck potential during writes

Usage at RHIC and USATLAS

- Central software repositories
- Home directories
- Web content
- Container for finished data

Panasas - Clustered Centralized Storage

- Integrated hardware-software solution
- Fast: direct and parallel data access
- Global namespace
- Security and centralized management
- Distributed metadata

Front

Components of each “shelf” at the RCF:

- 1 Director blade and 10 storage blades
- Up to 8TB raw storage
- 1 Gb of internal connectivity per blade (iSCSI)
- 4 Gb of external connectivity (etherchannel)

Panasas Architecture

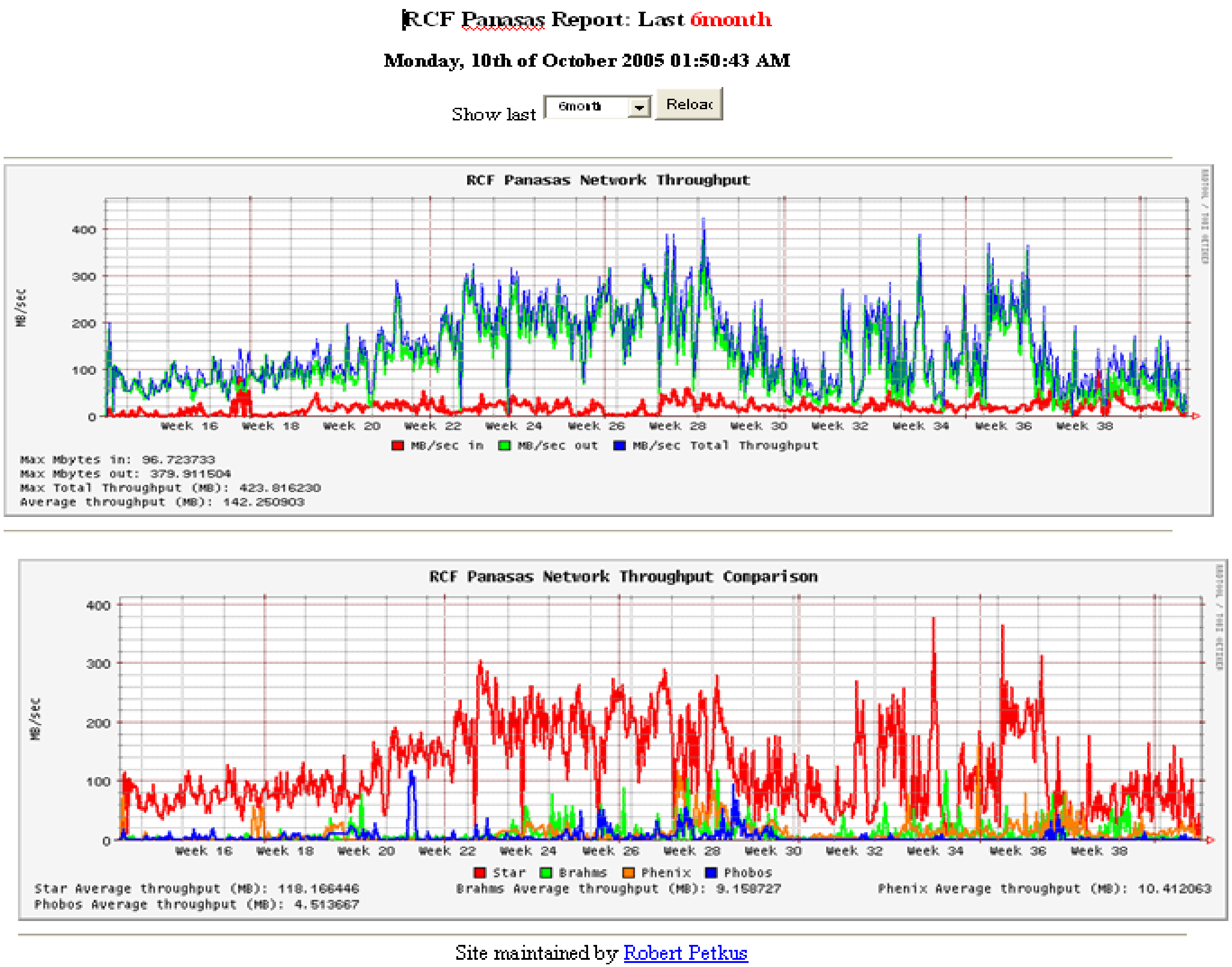
ActiveScale Operating System

- Runs on Director Blade
- Divides files into data objects, which are arbitrary in size, and stripes them across storage blades
- Dynamically distributes workload across storage blades
- Each storage blade is filled to 90% capacity. The remaining 10% is reserved for rebuilding parity

Direct Flow Software

- Installed on the Linux compute node
- Direct data path from client to storage blades
- Optimizes data layout, caching and prefetching
- File is reconstructed at the compute node

Panasas performance snapshot



RCF Panasas Report: Last 6month

Monday, 10th of October 2005 01:50:43 AM

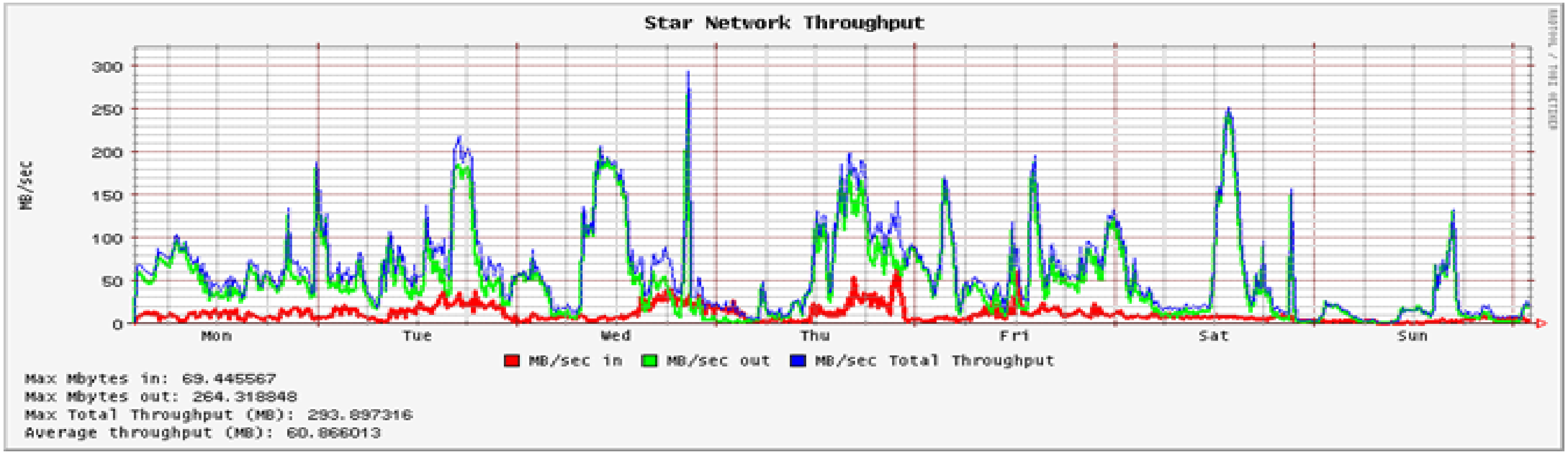
Show last 6month Reload

```
Max Mbytes in: 96.723733
Max Mbytes out: 379.911504
Max Total Throughput (MB): 423.816230
Average throughput (MB): 142.250903
```

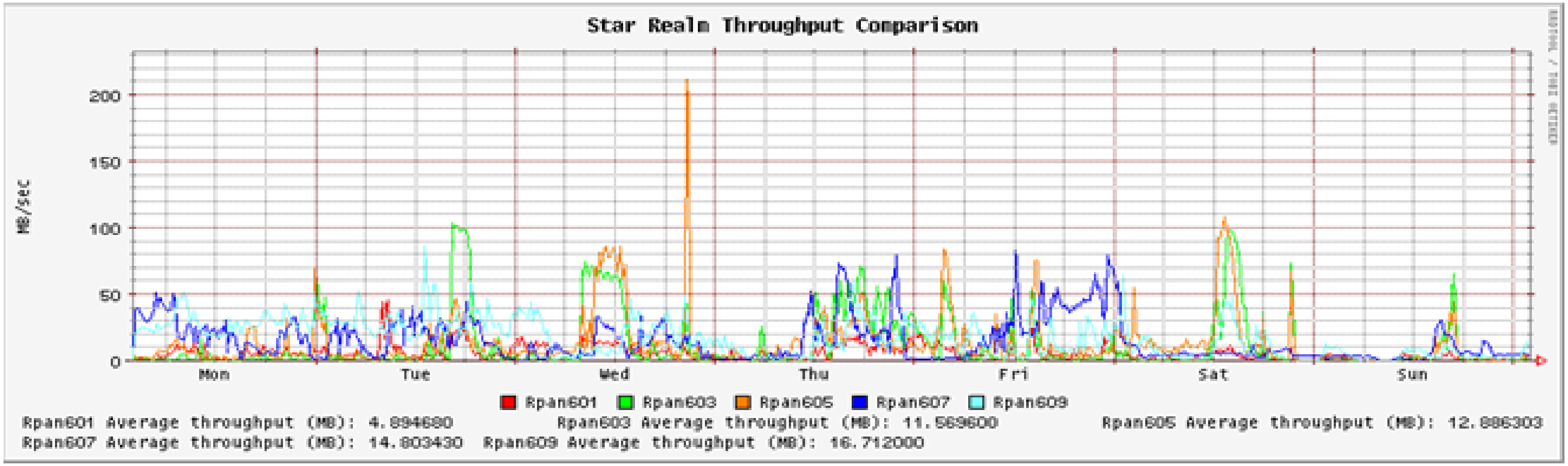
```
Star Average throughput (MB): 118.166446        Brahms Average throughput (MB): 9.158727        Phenix Average throughput (MB): 10.412063
Phobos Average throughput (MB): 4.513667
```

Site maintained by Robert Petkus

Panasas performance snapshot



```
Max Mbytes in: 69.445567
Max Mbytes out: 264.318848
Max Total Throughput (MB): 293.897316
Average throughput (MB): 60.866013
```



```
Rpan601 Average throughput (MB): 4.894680          Rpan603 Average throughput (MB): 11.569600          Rpan605 Average throughput (MB): 12.886303
Rpan607 Average throughput (MB): 14.803430  Rpan609 Average throughput (MB): 16.712000
```

Panasas in Practice

Sure it's fast and the technology is innovative and promising,

But...

- Proven to be unreliable and finicky
- Expensive, but utilizing low-rent hardware – you wouldn't or couldn't recycle this for something else
- Maxed-out volumes are unstable requiring the use of hard quotas which further diminish usable capacity
- A storage blade failure frequently equates to lengthy <filesystem offline> reconstruction and recovery
- The Linux clients are buggy kernel modules that often crash the node
- No LDAP support.

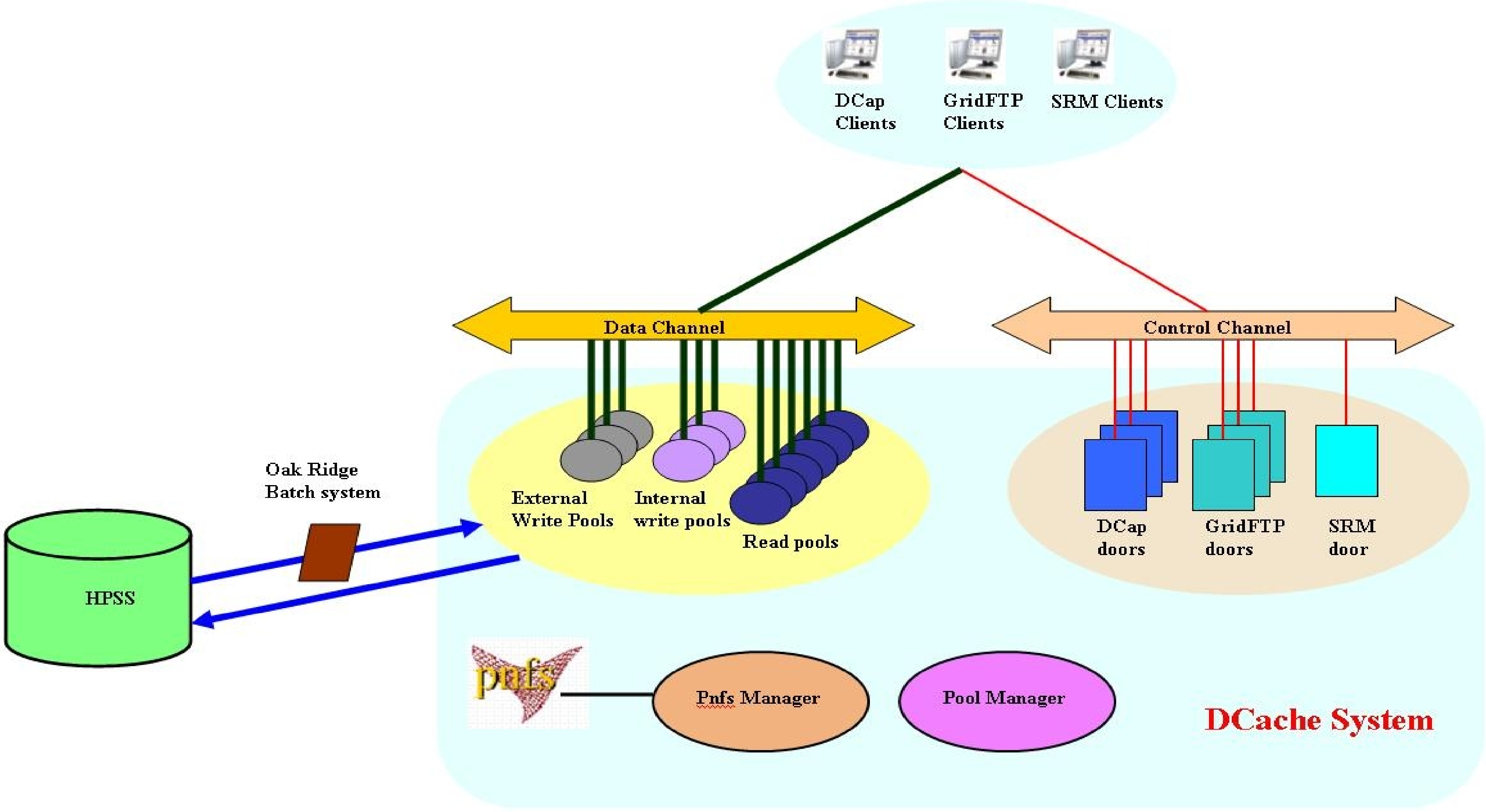
Usage at RHIC

- A replacement for NFS

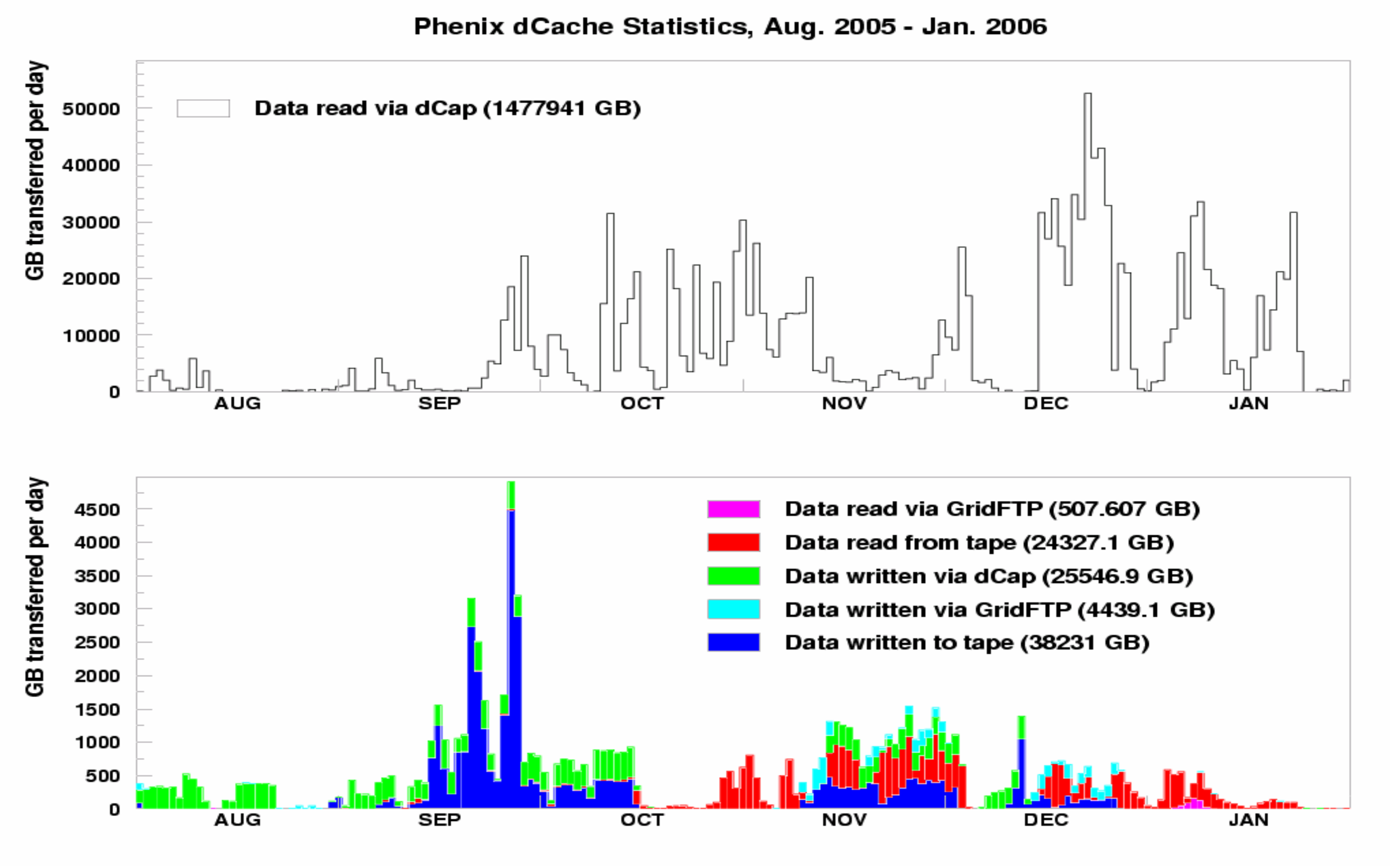
dCache Distributed Disk Caching System

- Heterogeneous data access – data is stored in disk pools or in HPSS.
- Global namespace
- Secure (GSI, Kerberos), scalable, redundant, and fault-tolerant
- Load balancing: Detection of a "hot spot" will dynamically replicate files using cost metric algorithms; e.g., lowest latency, least load, disk space, internodal data transfer vs. HPSS refetch
- Transparent access to data in disk pools or HPSS
- Multiple data access methods (doors):
  - Local POSIX-like access via DCAP
  - SRM: Protocol negotiation, checksumming, space allocation
  - GsiFTP: secure WAN data transport
- General UNIX filesystem access via PNFS, a database and exportable filesystem made available via NFS that serves and stores metadata. *PNFS does not store the actual files
- Improved administration and monitoring

USATLAS dCache Architecture



PHENIX dCache Usage



dCache Issues

- Write pool nodes:
  - Need dedicated write pool nodes – frequent crashing when sharing computing resources
  - Better performance using the XFS filesystem
  - Need reliable disks
- PNFS database is single point of failure and potential bottleneck. Need multiple, fault tolerant metadata dbs
- If the PNFS server crashes on transfer (very rare), files written into the system might be improperly registered and/or invalid requiring a manual clean-up of residual data
- Read or write access only (not RW)
- No mechanism or policy for user/group based quotas on transfer limits

dCache at the RACF

USATLAS Usage

- Hybrid model: The majority of compute nodes are also read pool nodes thus maximizing the potential of each system
- Data analysis: write raw data into HPSS, analyse on farm, write back to HPSS
- Grid production: On-site, dCap is the only source and destination for data. Off-site, data is accessible via GridFTP and SRM clients
- Oak Ridge Batch System for backend tape prestage

RHIC Usage

- Hybrid model for a subset of compute nodes in the PHENIX experiment
- dCache metadata (PNFS) exported/imported directly to/from PHENIX file catalogue
- Data retrieval from HPSS via Carousel (developed at RHIC) which allows for policy-based request throttling

Xrootd: Accessing and Managing Distributed Data

- High performance, multiplexed data access per client
- Global namespace
- Secure (Kerberos and GSI plugins)
- P2P architecture with distinct data/control flow akin to GnuTella
- Low CPU overhead
- Load balancing (open load balancers (OLB)) – conceptually similar to dCache "hot spot" detection wherein data is dynamically replicated on demand via definable cost metrics
- No single POF – each element in the system is completely redundant and fault-tolerant
  - Data clients (also data servers as configured in the STAR experiment) have transparent access to data
  - Data Servers – A master copy of the data in HPSS will be retrieved to another server in the event of a failure
  - Redirectors – traffic managers that direct clients to data
  - OLB Managers (monitors load and availability of data) and Servers (maps data location)

Xrootd continued

Issues

- Limited data access methods – however, the merging of xrootd with SRM is in development to provide a complete grid solution
- No access via a general UNIX filesystem-like interface – psychological convenience?
- Read or write access only (not RW)
- No mechanism or policy for user/group based quotas on transfer limits

Usage at RHIC

- The STAR experiment with 650 xrootd dataservers in deployment is the largest in the world
- Hybrid deployment – compute nodes double as fileservers

Other Solutions

- GPFS: data block striping across disks. Parallel reads/writes. Fail-over, replication, distributed metadata. Extremely fast and reliable but cost prohibitive
- BlueArc: ASICs dedicated to NFS, network, and filesystem. Feature-rich, fast and robust NFS solution – but still NFS…
- Ibrix: Meta servers assigned to segments in a disk pool. Traditional NFS or "Fusion", a proprietary protocol
- Lustre: Object-based, distributed storage solution. Software only. Active development. Older versions are free
- NFSv4.1: One protocol for file access, locking, and mounting. Kerberos integration, client caching and delegation policies. When??
- ZFS: Combined volume manager and 128-bit filesystem. Transactional writes, 64-bit data checksums. Proprietary, Solaris-only and still not ready for general consumption

Feature Comparison of RHIC/USATLAS “Fileservers”

	NFS	AFS	Panasas	dCache	xrootd
Low Cost	✓	✓		✓	✓
Reliability	✓	✓		✓	✓
Speed			✓	✓	✓
Scalability		✓	✓	✓	✓
Fault Tolerance		✓[1]	✓	✓[2]	✓
Load Balancing		✓[1]	✓	✓	✓
Security		✓	✓	✓	✓
Centralized Mngmt		✓	✓	✓	✓
Global Namespace		✓	✓	✓	✓
POSIX Access	✓	✓	✓	✓[3]	✓
Open Source	✓	✓			✓
User Policy Mngmt	✓[4]	✓	✓[5]		

1: When using replication

2: Single PNFS DB is a POF

3: POSIX-like access

4: Dependent on the underlying filesystem being exported

5: No group-based quotas

Conclusion

- No single implementation is a panacea providing all solutions
- New fileservers are constantly being developed/perfected and will continue to be tested
- Heterogeneous solutions will remain in place but the following trends are clear:
  - A move from centralized toward distributed storage
  - USATLAS and PHENIX will continue to deploy as well as expand and enhance their utilization of dCache
  - STAR, with the largest xrootd deployment in the world, is focused on xrootd, specifically a grid implementation using SRM
- Panasas robustness and reliability is ultimately disappointing
- NFS will hobble along for some time functioning as a home directory and scratch storage server
- AFS, popular at CERN and BNL, will remain the repository of choice for static software