USATLAS dCache system at BNL's RACF

Hironori Ito, **Zhenping (Jane) Liu**, Yingzi (Iris) Wu

RHIC/ATLAS Computing Facility, Physics Department

Brookhaven National Laboratory

10/15/2015 HEPIX Fall 2015 at BNL

USATLAS dCache system at BNL

- dCache: Distributed disk caching system software developed by DESY/FNAL/NDGF
- USATLAS T1 production storage since Y2004
- A very large scale disk storage system with tape back-end serving a geographically diverse, worldwide ATLAS scientific community.
  - Total size of name space: 17.9 PB. Among them, disk-only area data: 8.7 PB, and tape area data: 9.2 PB
  - Total space of disk pools: 14.2 PB, total used space of disk pools: 10.7 PB

Features

- dCache version:
  - Non-SRM head servers: 2.10.39; SRM: 2.10.41; Pools: 2.10.20
- ChimeraDB, SRMDB
  - Postgres v9.3.2
  - WAL based backup
- NFSv3
- NFSv4.1/pNFS
  - dCache mounted as any other NFS.
  - Direct I/O-operations
    - direct I/O access (cp, mv and etc) to space token area also works now

Features (Cont.)

- Replica Manager
  - RM is heavily used to manage replication of important DISK-only production data.
- Disk only data and Tape area data
  - Disk only data: disk copy only, probably multiple copies managed by RM;
  - Tape area data: all backed up in tape, most recently accessed files probably have disk copy.
- dCap transfer protocol
- SRM protocol
- GRIDFTP transfer protocol

Features (Cont.)

- Xrootd transfer protocol
- HTTPS/Webdav transfer protocol
- HPSS as Tape backend for tape area data
  - HPSS stage batch system for retrieval optimization
    - Batch system sorts out stage requests according to tape ID. Files on same tape retrieved in a batch.
- Storage pools are behind firewall
  - No direct access from WAN
- Space token
  - Enforces quota per spacetoken
  - Each user could get own space with own quota.
    - 5TB quota for a single USATLAS user in dCache name space.

dCache servers

- Head servers (all are RHEL 6.6)
  - Admin: 1 host
  - Chimera: 1 host
  - Chimera backup: 1 host
  - NFSV4.1 Export: 1 host
  - Replica Manager DB: 1 host
  - dCap: (4 VMs)
  - GridFTP/XROOTD doors: 18 hosts
    - dual network interface
  - SRM: 1 host
    - dual network interface
  - SRMDB: 1 host
  - SRMDB backup: 1 host
  - Billing/Monitoring : 1 host

dCache servers (cont.)

- Pool hosts:
  - 42 systems in production hosting about 600 pools:
    - 18 IBM x3650 M4
    - 20 IBM x3650 M2
    - 4 Dell R710
  - Reasons to choose them:
    - well known vendors
    - Stability
    - Reliability
    - compatibility with other systems in the facility

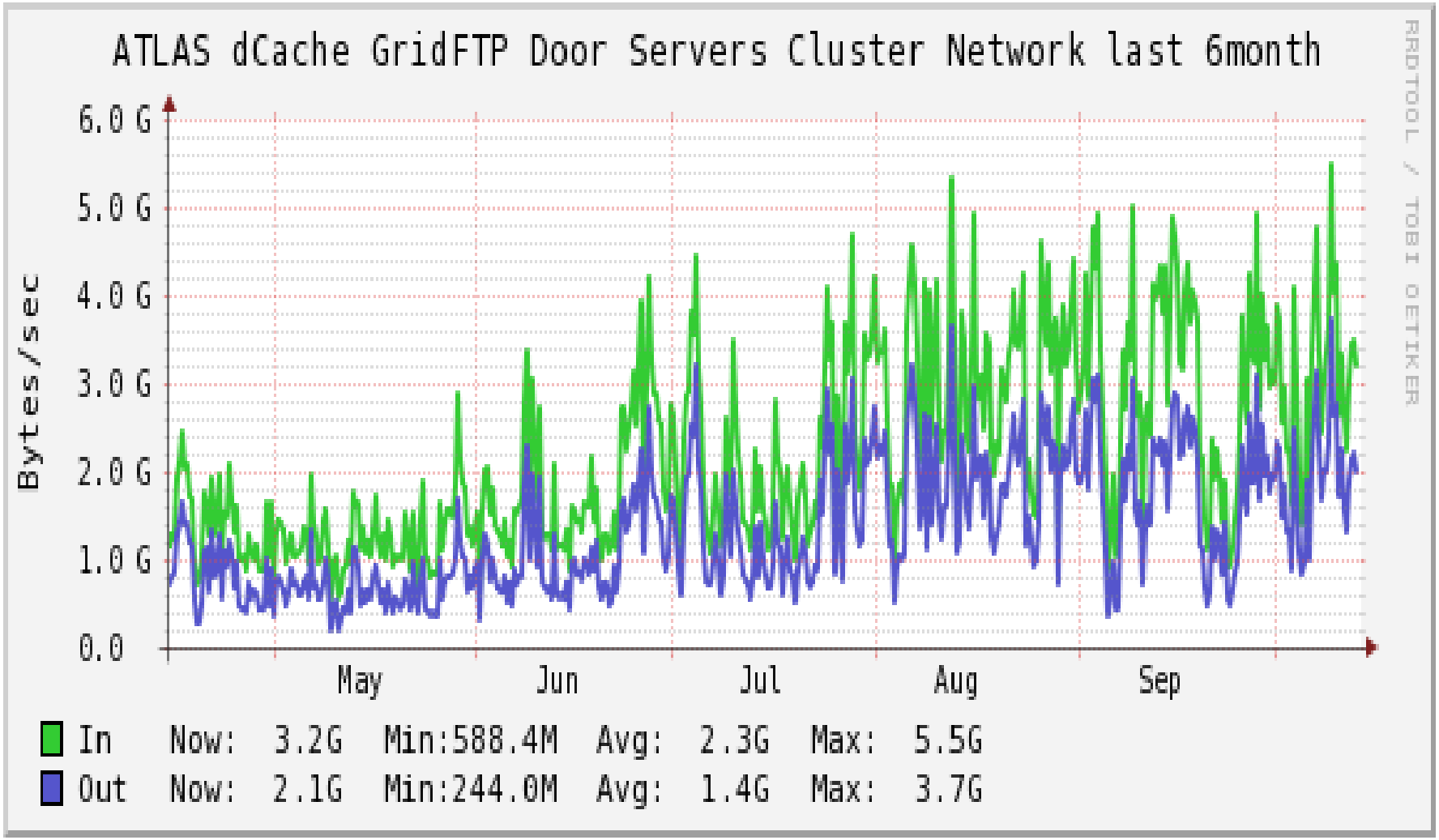
dCache servers (cont.)

- OS and file system
  - 28 hosts running Linux RHEL 6 with XFS
  - 14 hosts running Solaris 10 with ZFS
  - We are moving away from Solaris to Linux
    - license fees, better support in house for Linux
- Storage
  - Hitachi
  - JBOD
  - DDN
  - Nexsan
- Total disk space of pools: 14.2 PB; Total used disk space of pools: 10.7 PB

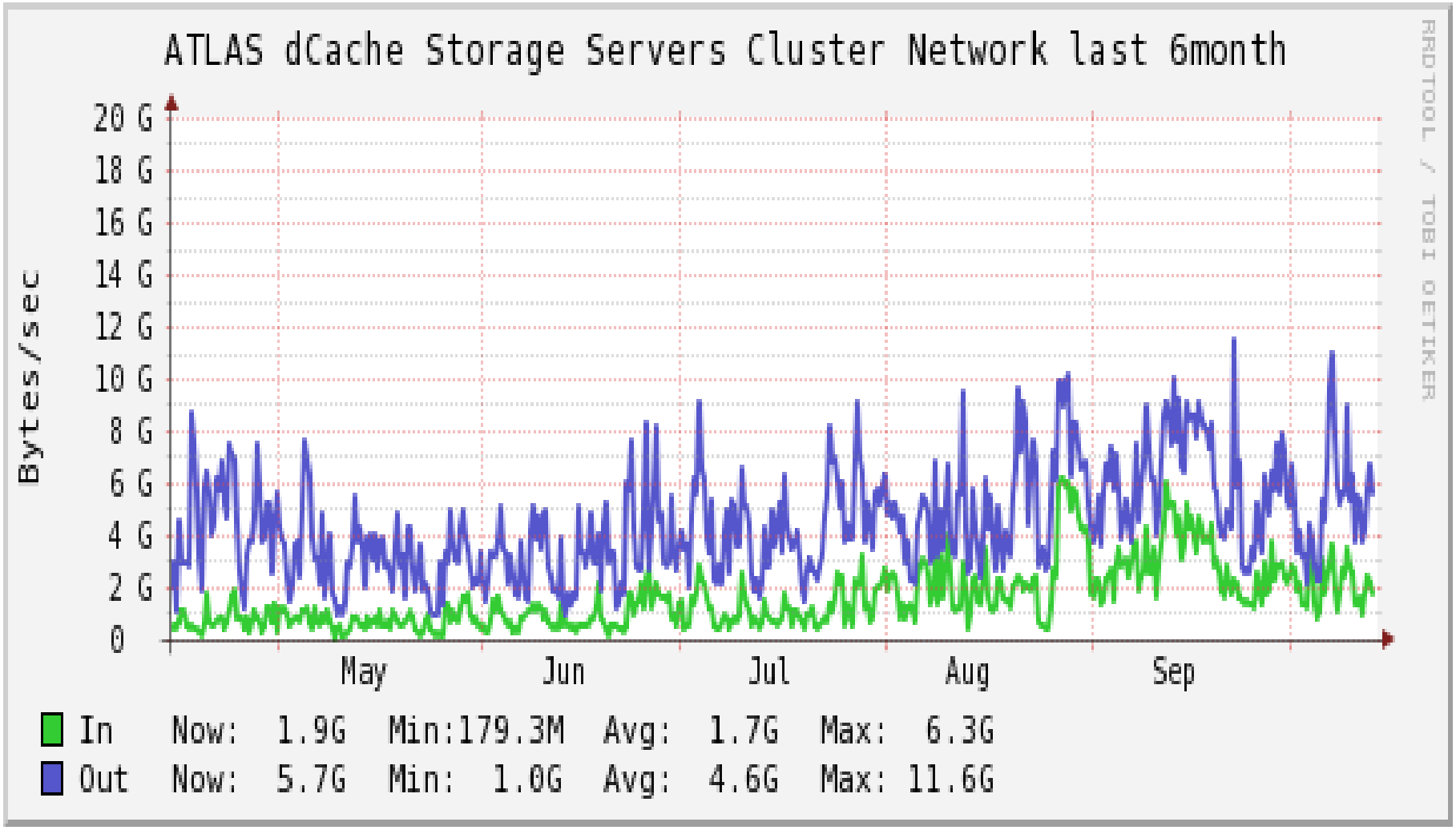
Network Configuration

- Only SRM and GRIDFTP servers are outside of BNL firewall for Atlas data transfer to ensure high performance
  - Each GridFTP door has two interfaces, one uses LHC-OPN (100Gbps) subnet IP (which is outside of BNL firewall), the other one uses 186 subnet (which is inside RCF/ACF firewall)
- All other dCache servers (including pools, admin, chimera, dcap, billing/monitoring, SRMDB and etc) are protected by two levels of firewalls, first BNL firewall then RCF/ACF firewall.

WAN Data Transfer (through GridFTP)



LAN Data Transfer (storage servers)



USATLAS dCache and FAX

- USATLAS dCache has its own Xrootd door, but only local Xrootd clients can access data.
  - Storage pools behind firewall, data not accessible from outside xrootd clients.
- dCache joins FAX (Federated ATLAS storage systems using XrootD)
  - FAX brings ATLAS Tier 1, Tier 2 and Tier 3 storage resources together into a common namespace, accessible from anywhere
  - List of FAX related services at BNL
    - dCache Storage
    - dCache's own xrootd services in pnfs namespace
    - FAX Reverse proxy redirector/servers
    - FAX (Forward) proxy redirector/servers
  - Software version
    - xrootd-server-atlas-n2n-plugin: v2.2
    - dcache-xrootd-n2n-plugin: v6.0.4

USATLAS dCache and FAX (Cont.)

- FAX Reverse proxy service
  - Allows the access to BNL data behind the firewall by clients located outside BNL.
  - FAX Reverse proxy understand ATLAS global namespace via its N2N plugin.
  - It is consisted of xRootd reverse proxy manager (redirector) and several xRootd reverse proxy servers (aka, data servers).
  - Reverse proxy servers (sharing hosts with GRIDFTP services) have dual network interfaces to access both external and internal networks. They proxies the data-transfer into/out-of the dCache storage servers without them going through site firewalls.

USATLAS dCache and FAX (Cont.)

- FAX (Forward) Proxy
  - Necessary since BNL's worker nodes are behind the firewall!
    - With reverse proxy, accessing BNL data from BNL worker nodes will not cause the network path to cross the firewall. But, the data will go through the proxy servers.
    - Accessing non-BNL data from BNL worker nodes will cause the network path to cross the firewall since the clients will directly communicate with remote xrootd services.
  - With FAX Forward Proxy
    - Only external data will be redirected to FAX Forwarding Proxy Servers; BNL worker nodes will talk to our dCache data servers directly.
    - BNL worker node clients can redirect to any other redirectors

Installation and Configuration Management

- Based on Puppet + Git + Cobbler + GLPI
- Cobbler: rapid setup & installation of systems through the network (for Linux type servers)
- Installation and configuration of dCache software and related software are managed by Puppets
  - Java, osg certificate, osg hostcert installation
  - dCache: software, configuration, cron, network related config, logrotate.
  - Major dCache version upgrade very quick
    - About several hours; Most upgrade time was actually spent on database conversion. E.g., 1.9 --> 2.6, 2.6 --> 2.10 upgrade
- Puppet scripts are stored in Git and pulled automatically to hosts
  - Various checks to prevent bad commit
- hosts and services specific parameters are stored in GLPI
  - Eg, open specific ports in firewall, disk partition, Ganglia group etc...

dCache Monitoring

- Nagios
  - Host up/down, dCache cell online/offline, pool space usage errors, disk space and etc.
  - Icinga2 in testing phase. Have plan to move to Icinga2.
- Ganglia
- Maintenance scripts
  - Various scripts to check logs to detect potential/ongoing problems on pools, gftp, srm, admin, Chimera, HPSS batch system and etc. Crons restart components (e.g., pools) if necessary.
  - New scripts developed after learning from problem debugging
  - Crons to check billing and do statistics on top errors to detect problematic pools/GFTPs/restore/store.
- ELK(Elasticsearch, Logstash, Kibana) framework used for the billing monitoring dashboard (Development phase)
  - See Carlos Gamboa's talk this afternoon
- Panda job pages
- DDM Dashboard

Recent Issues

- New storage pool servers
  - Frequent OutOfMemory problem, also invalid environment errors, process killed and so on
    - Cron to detect errors and do automatic restart
    - Will increase the RAM to 256G
- Replica Manager performance issue
  - Separate replica DB to a dedicated host
  - Wait for better software version to come out
- High load upon SRM restart sometimes due to busy PinManager or SRM component
  - Increased to 8G for PinManager component

Near-future plan of pool storage reconfiguration

- Currently symmetric replication
  - For important disk-only production data, RM replicate them to two copies on two groups of storage hosts.
  - Two replicas have same priority for clients to access.
  - Not much difference of storage qualities for two copies
  - Very expensive if both of replicas kept on good disk

Near-future plan of pool storage reconfiguration

- Near-future asymmetric replication (primary/secondary)
  - Primary copy on good disks
    - Hitachi Data Systems (HW RAID)
    - 4TB RAID JBOD
  - Secondary copy on bad disks
    - JBOD storage with 8TB Seagate SMR disks (6.336 PB usable space)
    - Other older HW RAID systems
  - All primary copies in CDCE and all secondary copies in Sigma 7
  - About 12.2PB of usable primary storage, 10PB of usable secondary storage

Data flow in asymmetric replication

- New files written into primary storage pools, then files get replicated from primary to secondary storage pools by Replica Manager
- Primary pools has higher read priority for client access than secondary pools
  - Most of the time, primary copy should be available. When client asks for a file, use primary storage pools to serve.
  - Occasionally when primary copy is not available due to maintenance or problems, use secondary storage pools to serve.
- Rather cost-effective but still ensures performance. At the same time it improves the file unavailability situation during storage maintenance window.