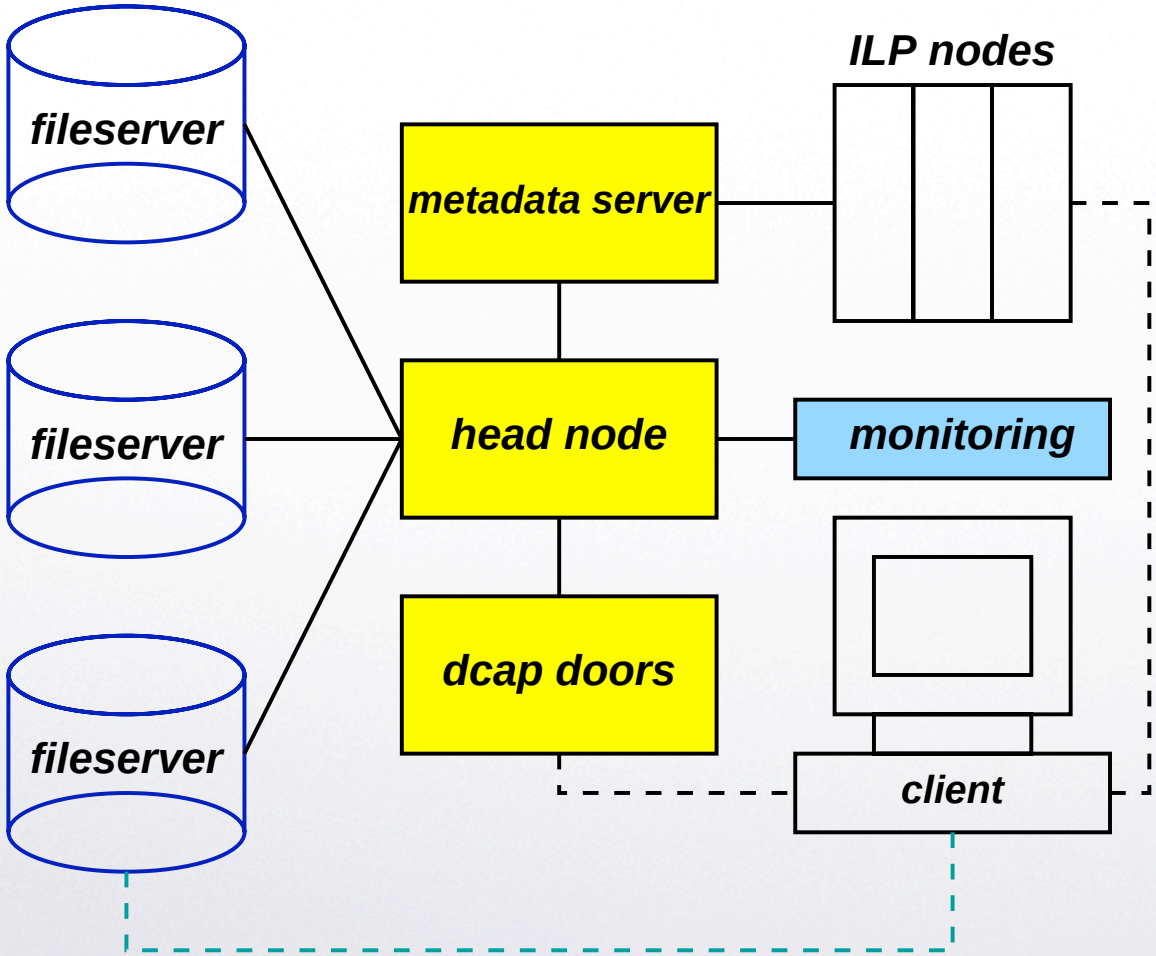
Diskpool Support and Operation

Alexei Varganov, Dave Gerdes
University of Michigan

Outline

- Prototype setup
- Diskpool support in a nutshell
  - server operation
  - client care
- Analyzing diskpool activity
- Summary

Prototype Architecture



- *fcdfrdc[2, 3, 4]* used for metadata, door, and head hosts respectively
- *fcfdrdc5* is monitoring
- 50 TB in 8 RAID servers *fcdfdata[034, 039, 040, 064, 119, 126, 140, 141]*
- All nodes use SL 3, dCache v1.6.5

Disk Space Organization

Group	Allocated
stn	23.5 TB
ewk	6.9 TB
top	6.7 TB
qcd	5.2 TB
bnt	3.3 TB
exo	1.7 TB

- The diskpool name space is grouped into several areas

  /pnfs/diskpool/[ewk, top, ... , exo]

- Each physics group area is *linked* to its dedicated pool group

- If a new file server is added its disk partitions a.k.a *pools* are assigned to various pool groups without any interruptions!

dCache ONLINE

Disk Space Usage

CellName	DomainName	Total Space/MB	Free Space/MB	Precious Space/MB	Layout (precious/used free)
data1-fcdfdata064	fcdfdata064Domain	1004342	743157	261184	
data1-fcdfdata119	fcdfdata119Domain	1115638	96729	1018908	
data1-fcdfdata126	fcdfdata126Domain	2044595	134468	1910127	
data1-fcdfdata140	fcdfdata140Domain	1813140	88767	1724372	
data1-fcdfdata141	fcdfdata141Domain	1813140	346420	1466719	
data2-fcdfdata064	fcdfdata064Domain	1004342	669473	334869	
data2-fcdfdata119	fcdfdata119Domain	1115638	0	1115637	
data2-fcdfdata126	fcdfdata126Domain	558609	20286	538323	
data2-fcdfdata140	fcdfdata140Domain	1813140	31625	1781515	
data2-fcdfdata141	fcdfdata141Domain	1813140	504651	1308488	
data3-fcdfdata119	fcdfdata119Domain	1115638	94470	1021167	
data3-fcdfdata126	fcdfdata126Domain	2044595	130265	1914329	
data3-fcdfdata140	fcdfdata140Domain	1813140	52892	1760247	
data3-fcdfdata141	fcdfdata141Domain	1813140	937459	875681	
data4-fcdfdata119	fcdfdata119Domain	1115638	0	1115637	
data4-fcdfdata126	fcdfdata126Domain	558609	19554	539055	
data4-fcdfdata140	fcdfdata140Domain	1813140	49319	1763820	
data4-fcdfdata141	fcdfdata141Domain	1813140	181919	1631220	
data5-fcdfdata119	fcdfdata119Domain	1115638	98637	1016369	

Support Cases

Support level structure

- *basic level* -- support by power users and offline
- *intermediate level* -- by the University of Michigan
- *expert level* -- support by FNAL computing division

What does *intermediate level* support include

- System maintenance and expansion
- Data relocation
- Data access support
- Privileges management

Diskpool Node Support

- Software configuration: on boot each node automatically starts all dCache services, identifies itself as a valid system component, and joins the diskpool.

- Details specified in the file /ets/init.d/dcache

- dCache processes run by

  root on two admin nodes in area */opt*

  non-root user *cdfdata* on other nodes in home area

- For diskpool diagnostics, archived log files are stored for last 30 days

Diskpool Expansion

- Adding a new file server is easy!
- Home area ~*cdfdata* from an active server gets replicated to a new machine.
- New raid partitions declared as pools.
- New pools are linked to the physics groups areas using dCache admin interface.
- It takes less than an hour to do the dCache-specific part of the disk server configuration!

Account Management

- The head node *fcdfrdc4* hosts diskpool-wide authentication process
- User Kerberos principals are mapped to *read* and *write* accounts in file config/gss.kpwd
- Similarly, user X.509 certificate info mapped in config/gsi.kpwd
- Users with namespace editing privileges listed in /pnfs/diskpool/.k5login
- All diskpool data owned by *cdfdata* account, which is used for collaborative management of the physics group data.
- Physics group conveners are able to manage the disk space efficiently

Maintenance: from ILP

- CDF has interactive login pool of linux nodes (ILP)
- ILP has diskpool mounted via NFS in */pnfs/diskpool*
- Power users access the area */pnfs/diskpool* to create directories and remove diskpool data
- The diskpool file path on ILP looks the same as used in dcap URL
- Power users can get diskpool report for a physics group pool

  /pnfs/diskpool/space ewk

DCAP Library

- Analysis jobs access data directly from pool using network based *dcap* protocol

  myChain.Add("dcap://fcdfrdc3:22125/pnfs/diskpool/test/out.root")

- *dcap* has been a *cdfsoft* product for a while
- For data copying *dccp* command is used

  dccp /tmp/pong dcap://fcdfrdc3:22125/pnfs/diskpool/test/ping

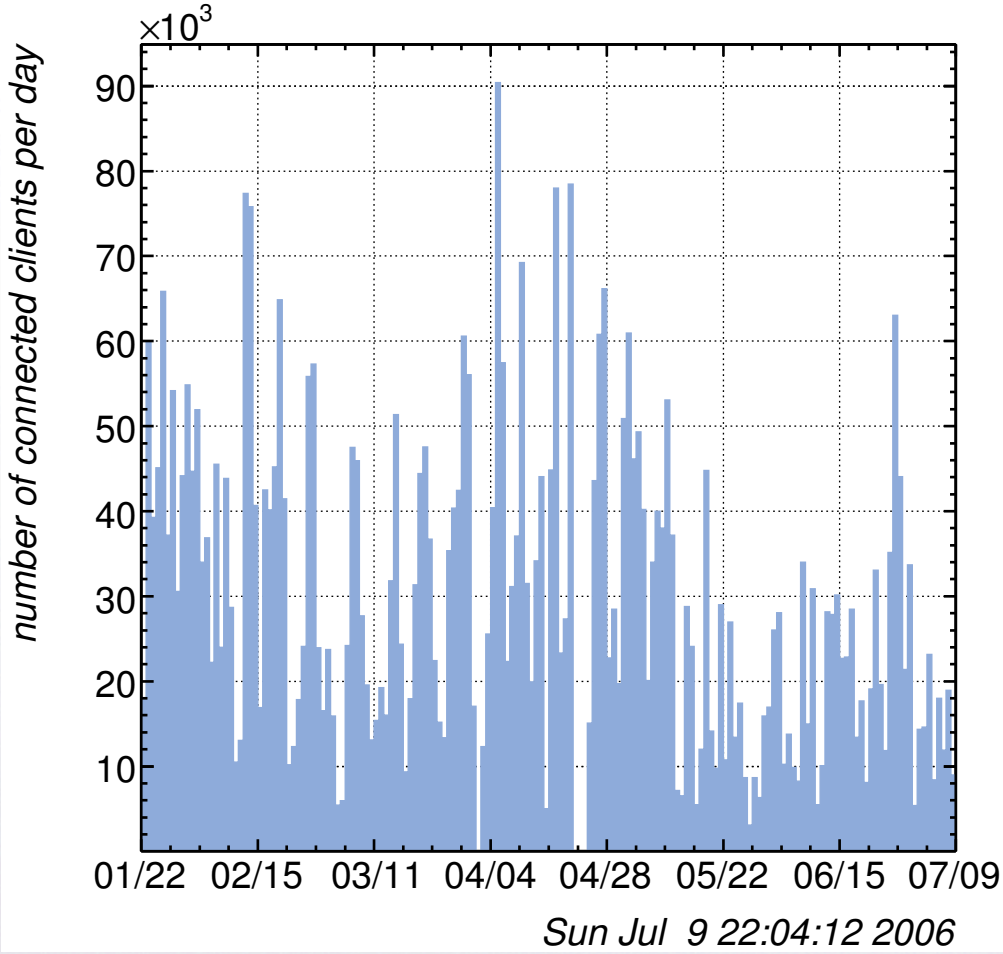
User Resources

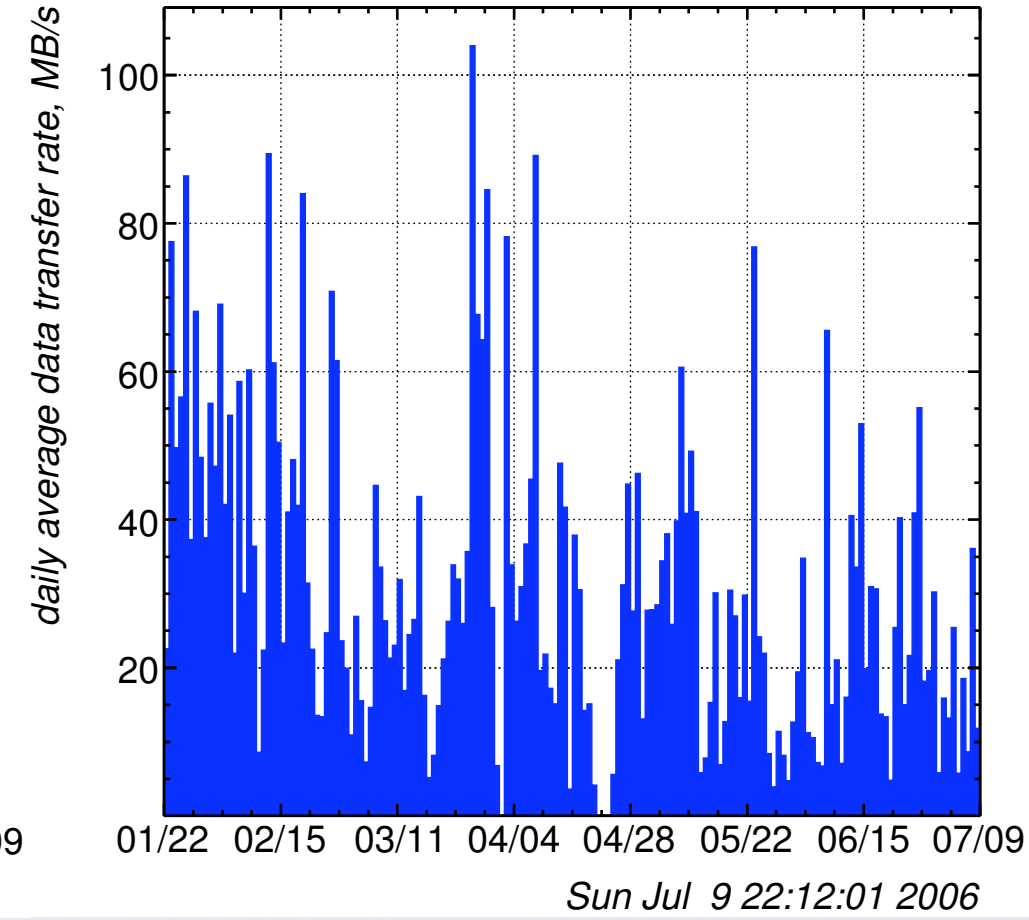
- Analysis Diskpool link on Fast Navigator
- Archived e-mail lists
  cdf_scalable_diskpool@fnal.gov
  cdf_diskpool_admin@fnal.gov
- Web based blog and forum

Operational Experience

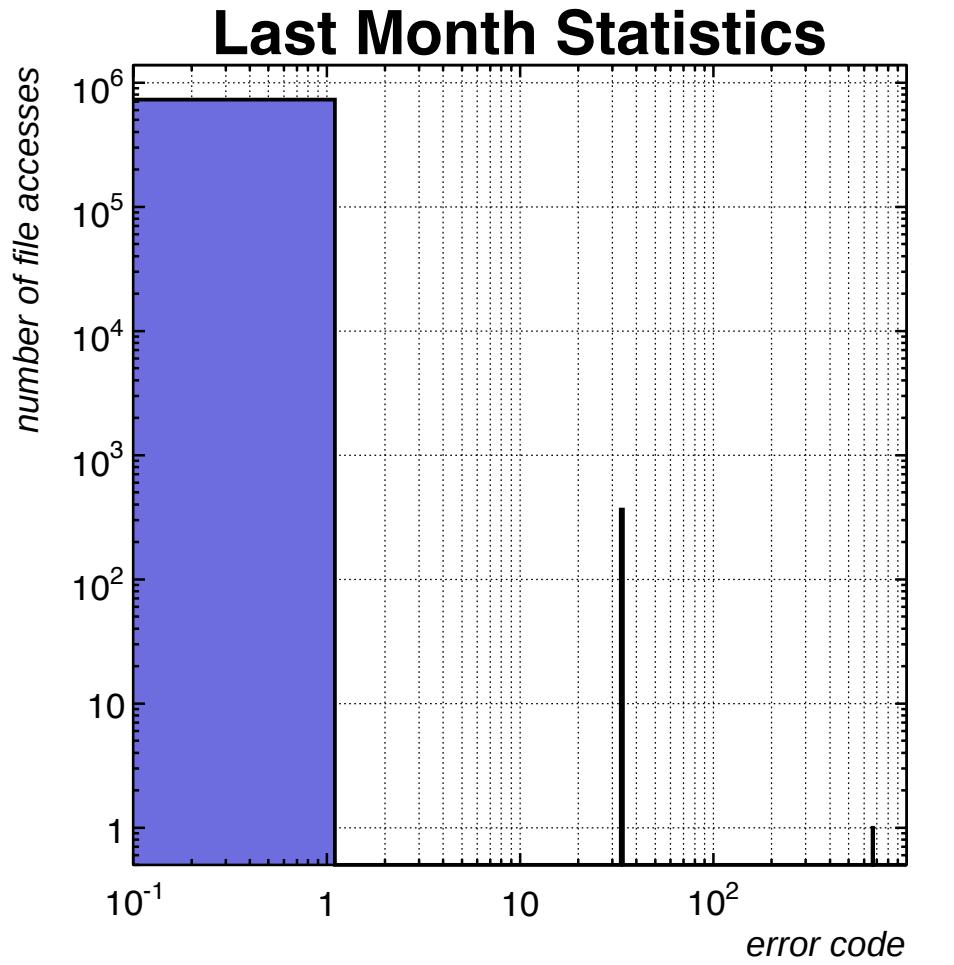
- Kerberos door hang-ups under high loads. Solved by implementing KDC multiplexor in November 2005.
- Misconfigured door or PNFS node would not boot up after restart: 3 'weekend downtimes'. Fixed by the REX sysadmin group
- Partition fill-up rendering nodes unresponsive. All pools readjusted to take less than 97.5% of its disk
- Out of warranty raid array failure in August'06: only 7 files out of 2 TB of relocated data (< 0.4%) were lost
- Data integrity check. We now use *adler32* check sum

Six Months Activity



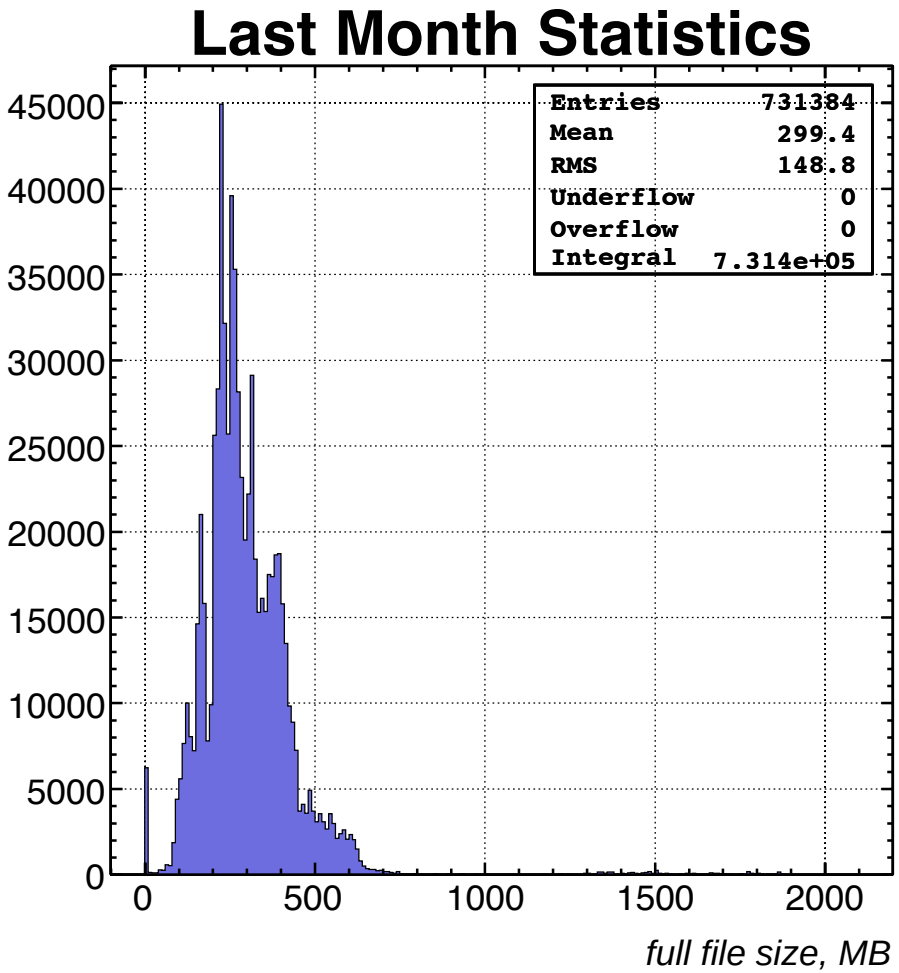


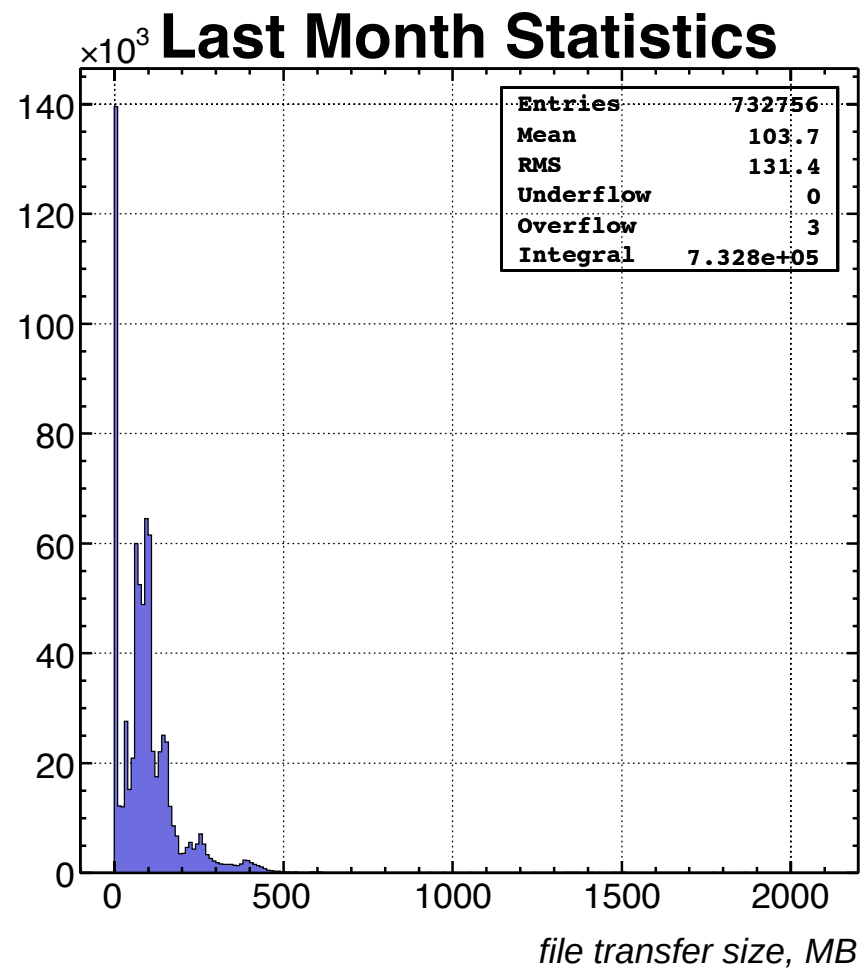
Last Month Activity



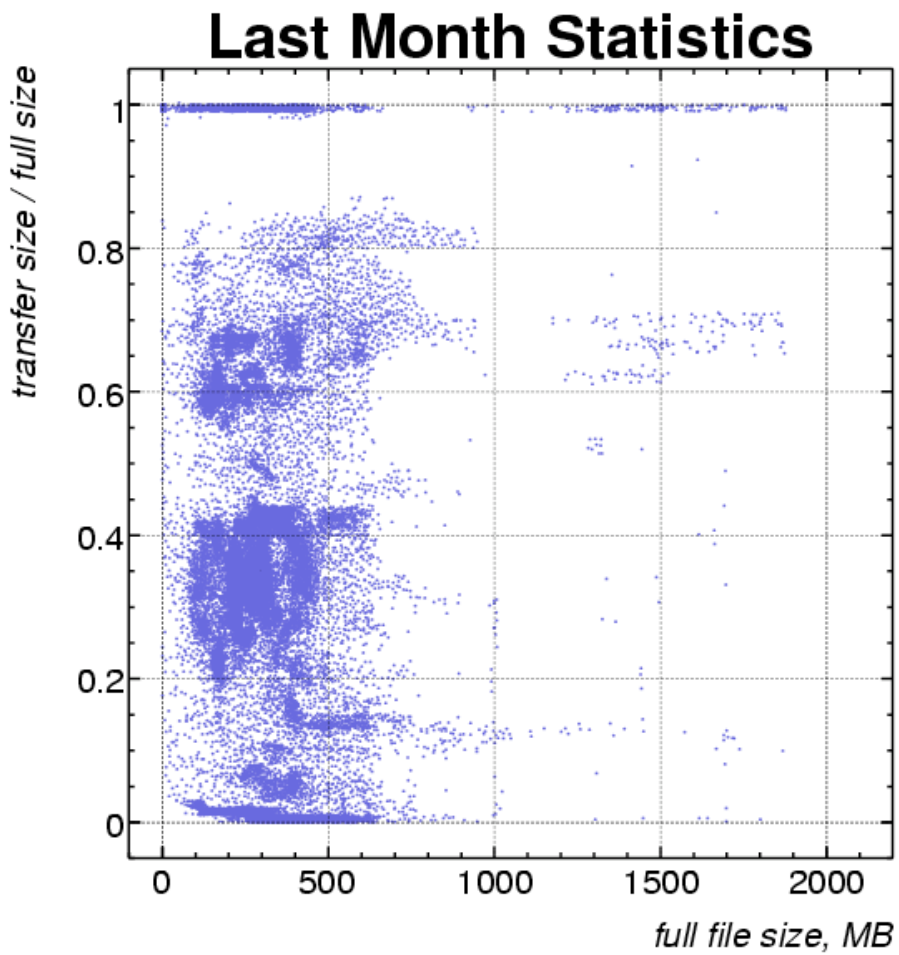
- All displayed information comes from billing database hosted on *fcdfrdc5*
- Entries with errors:
  - understood
  - bulk coming from the attempts to access one erroneously created file with 0 length

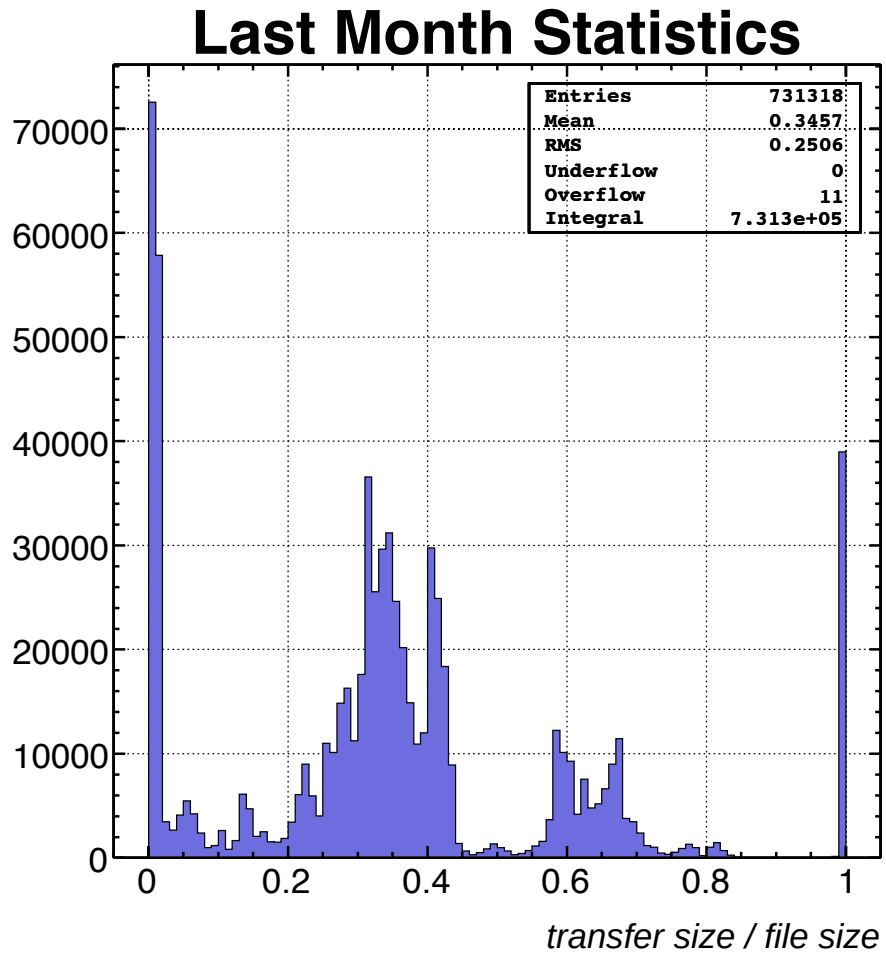
File and Transfer Size





35% Transfer Size Ratio



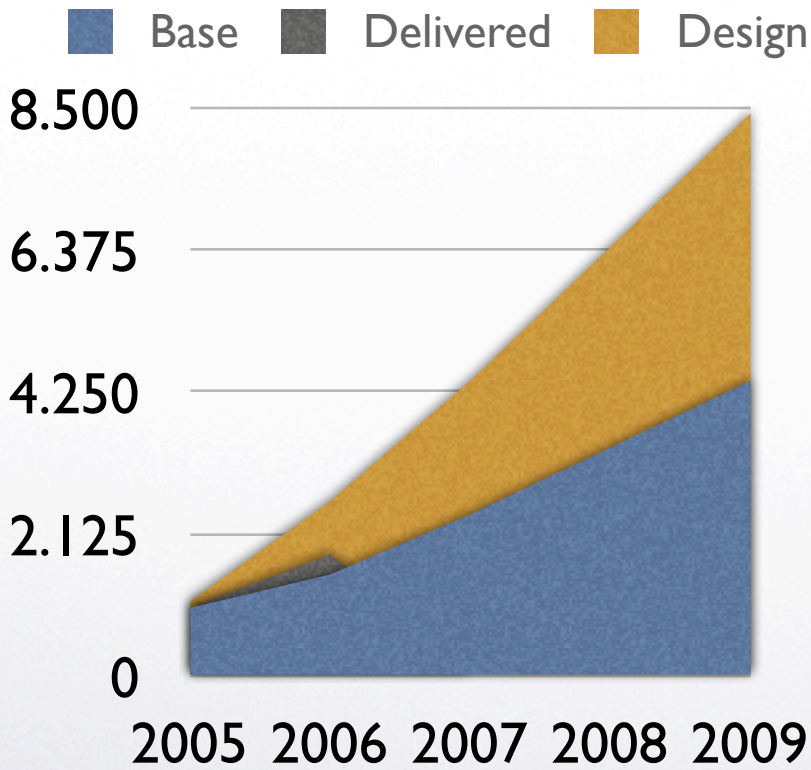


Summary

- dCache has matured and became well-documented
- The product fits very well within CDF computing environment.
- Building and validating the prototype from scratch took less than a week!
- Most of the installation and configuration steps learned myself.
- Diskpool is easy to run and monitor.
- Low maintenance level. Estimated support effort about 10 hr/week!

Backup Slides

Motivation



- Current array of isolated servers is difficult to maintain
- Need to organize disk space into centrally managed resource

March 2005 Workshop

- Several candidates for distributed file system
  - open source Parallel Virtual File System
  - General Parallel File System by IBM
  - Lustre by CFS
  - Global File System by Red Hat
  - dCache DESY/FNAL
  - a few other commercial products
- The final solution was constrained to the boundary conditions

Support Cases

On the server side

- startup and restart
- user privileges management
- space management
  - data relocation
  - adding new space

On the client side

- namespace export and operations
- *dcap* library support
  - *dccp* command
  - user applications

dCache Solution

- Supported by FNAL CD
- Adopted widely by HEP community

  extended shared knowledge base
- Mature and well documented product
- Very easily integrated with CDF hardware and software

Other Usage Patterns

- Getting the disk space report

```
/pnfs/diskpool/space ewk
```

- Finding files

```
find /pnfs/diskpool/stn -name “dm02f953.*”
```

- Calculating disk usage

```
du -sh /pnfs/diskpool/stn
```

Handy PNFS Data

- Each files in dCache identified by its ID

```
cat "/pnfs/diskpool/test/.(id)(ping)"
```

- Retrieving file name is easy with its ID

```
cat "/pnfs/diskpool/.(nameof)(ID)"
```

- Even more info: original *adler32* checksum

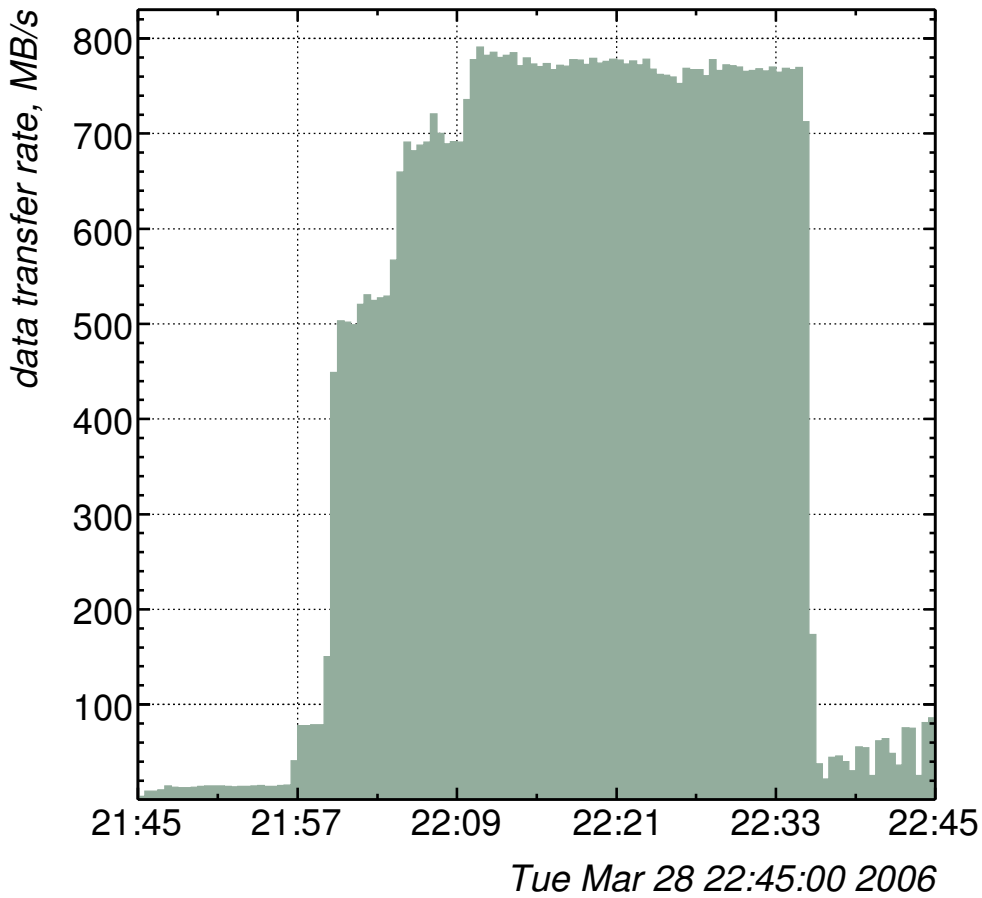
```
cat "/pnfs/diskpool/test/.(use)(2)(ping)"
```

Benchmarking Tests

To build confidence in diskpool performance we carried out this benchmarking tests

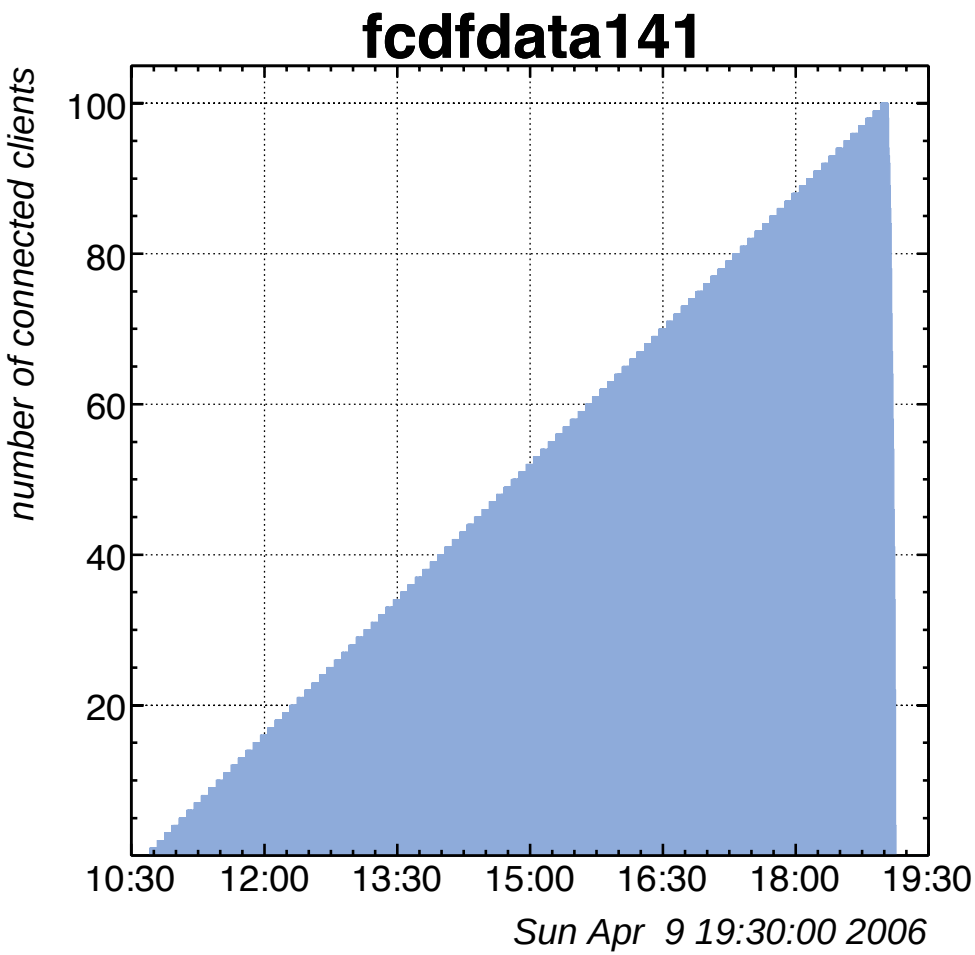
- aggregate throughput test using *dccp*
- single server *dccp* test under high load
- single server *root* test under high load

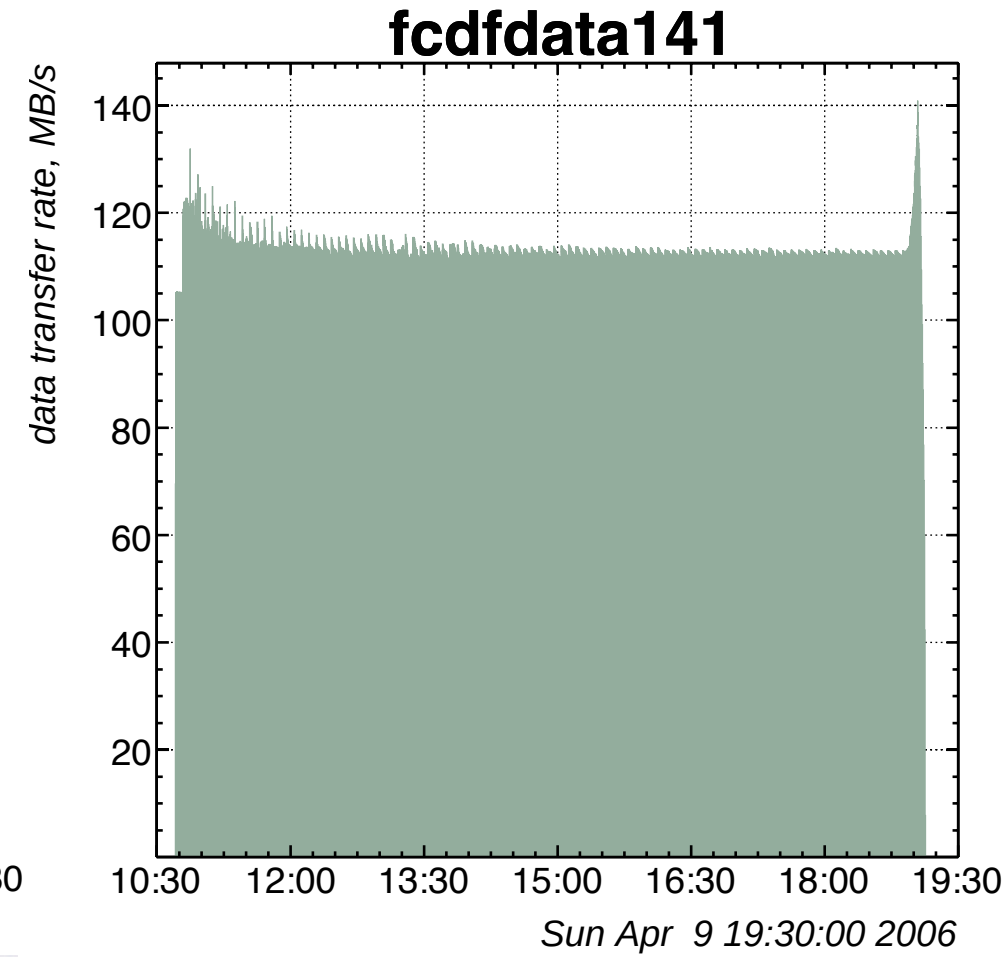
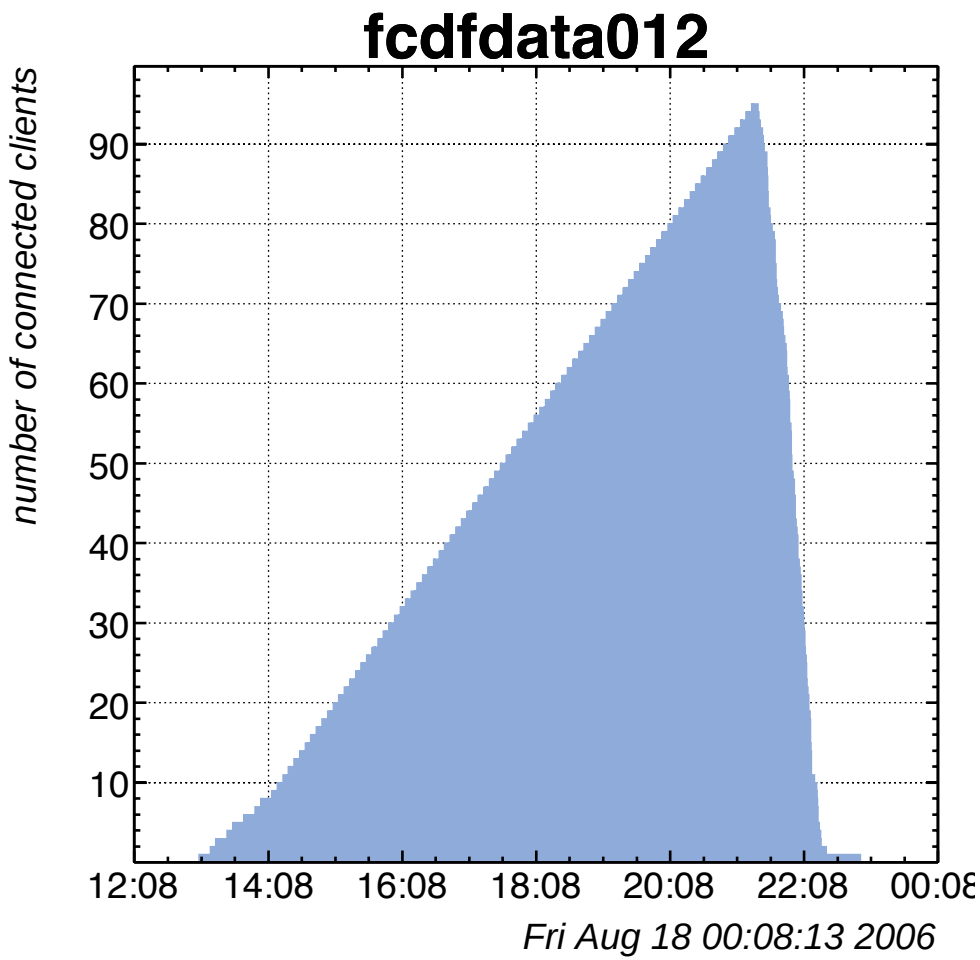
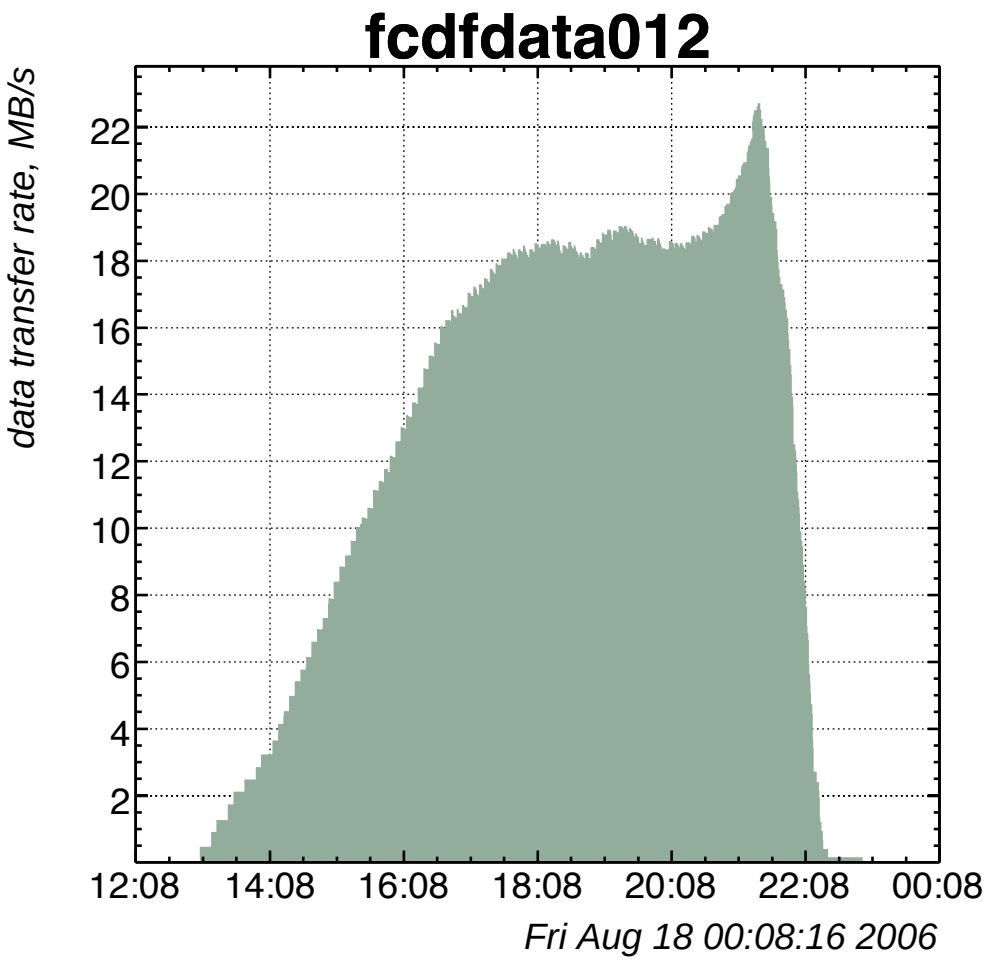
Max Throughput Test



- Concurrent *dccp* clients ramped up from CAF to reach disk I/O saturation
- Aggregate I/O throughput consistent with 100MB/s per node

DCCP I/O per Server





ROOT I/O per Server