

CDF Analysis Disk Pool

Ashutosh Kotwal

(for the CDF Offline Group)

- Motivation
- Requirements
- Prototype Status
- Next Steps
- Summary

Motivation

- Broadly speaking, physicists use large amounts of disk during analysis in two ways:
 - General dCache (backed up by Enstore tape system) for accessing official production datasets (e.g. reconstruction farm output, Monte Carlo)
 - Static disk (without automatic tape backup) allocated to physics groups for
 - Small, analysis-specific or group-specific data splits
 - Analysis-specific or group-specific Monte Carlo, e.g. For studying systematics, SUSY parameter scans etc.
 - Private or semi-private ntuples
 -
- Latter use case (topic of this talk) is driven by physics groups, allocation to users is driven by their analysis initiatives: this disk provides the users the flexibility they need to do research and publish.

Physics Group Static Disk

- Currently consists of many file servers (some owned by universities), with privileges assigned to specific users/groups
- Accessed via Rootd
- Size ~ 120 TB (including university file servers)
- Access rate ~ 25 TB / day (read + write)
- Issues with this system that need to be improved upon:
 - Difficult to do accounting and impose accountability
 - Difficult to reallocate disk resources from one user to another
 - Performance of Rootd not scalable
- Anticipated need for better solution last year – a common, shared disk-pool

Analysis Disk Pool

- The vision:
 - All file servers allocated for physics group static usage will become part of the analysis disk pool
 - Create 'quotas' on the disk pool for each of 5 physics groups, to be internally allocated and managed by physics group conveners / disk coordinators
 - Create additional 'quotas' for temporary storage (for validation and concatenation) of experiment-wide 'production ntuples' which will be treated as official datasets
- Discussions / brainstorming occurred over the last year, including workshop in spring 2005

Dcache-based Analysis Disk Pool Prototype

- Few months ago, we decided to prototype disk pool using dCache
 - Fermilab-supported product, with local expertise and good experience, made dCache a natural choice
 - Alexei Varganov (U. Michigan postdoc) volunteered to do the work, with help from Doug, Krzysztof and Pasha
 - Guidance from CD dCache experts much appreciated
- CDF “1-pass processing” made 2005 collider data rapidly available to physicists (in anticipation of 1/fb Challenge for Winter 2006 conferences)
- We made analysis disk pool prototype available for staging “production ntuples” for expediting analysis of 2005 data
- Prototype is currently 23 TB, being used to stage & validate production ntuples (& giving us feedback on performance and stability)

Dcache-based Analysis Disk Pool Prototype

- Benefits already apparent: open access to information, browsable, public accountability, reallocation of resources on-the-fly

Disk Space Usage

CellName	DomainName	Total Space/MB	Free Space/MB	Precious Space/MB	Layout (precious/used/free)
data1-fcdfdata119	fcdfdata119Domain	1024000	32861	991138	
data1-fcdfdata126	fcdfdata126Domain	1945600	756158	1189441	
data2-fcdfdata119	fcdfdata119Domain	1024000	528070	495929	
data2-fcdfdata126	fcdfdata126Domain	563200	554485	8714	
data3-fcdfdata119	fcdfdata119Domain	1024000	33148	990851	
data3-fcdfdata126	fcdfdata126Domain	1945600	756461	1188859	
data4-fcdfdata119	fcdfdata119Domain	1024000	527652	496347	
data4-fcdfdata126	fcdfdata126Domain	563200	554513	8686	
data5-fcdfdata119	fcdfdata119Domain	1024000	33106	990893	
data5-fcdfdata126	fcdfdata126Domain	1945600	756393	1188909	
data6-fcdfdata119	fcdfdata119Domain	1024000	528271	495728	
data6-fcdfdata126	fcdfdata126Domain	563200	554487	8712	
stn-fcdfdata012	fcdfdata012Domain	1740800	756417	983984	
stn-fcdfdata029	fcdfdata029Domain	1740800	759675	981124	
stn-fcdfdata034	fcdfdata034Domain	1433600	33033	1400566	
stn-fcdfdata039	fcdfdata039Domain	1433600	46014	1387585	
stn-fcdfdata040	fcdfdata040Domain	1433600	32964	1400635	
test-fcdfdata034	fcdfdata034Domain	51200	31652	19547	
test-fcdfdata039	fcdfdata039Domain	51200	32466	18733	
test-fcdfdata040	fcdfdata040Domain	51200	32564	18635	
top-fcdfdata034	fcdfdata034Domain	716800	333484	383315	
top-fcdfdata039	fcdfdata039Domain	716800	333022	383777	
top-fcdfdata040	fcdfdata040Domain	716800	332782	384017	

Dcache-based Analysis Disk Pool Prototype

- So far so good....next steps:
 - Demonstrate production-level system reliability
 - Demonstrate scalability to O(100 TB) size and O(25 TB/day) bandwidth
 - Measure the various user analysis patterns in more detail to specify requirements
 - Add physics group 'quotas' in addition to existing 'production ntuple' quotas to complete the prototype use cases
- Performance / stress testing is in progress, results will be analysed and discussed at future GDM meeting
- Consultation with dCache/CMS experts on optimal configuration may be beneficial at this stage
 - As our questions become better-defined

Summary

- Concept of shared, common analysis disk pool as a way of providing physics group disk space seems viable
- Benefits of disk pool over scattered file servers (served by Rootd) are apparent
 - Physics group disk becomes a managed resource, while still maintaining the flexibility to be driven by physics priorities
 - Total size and access rate more scalable
- Anticipate need for O(100 TB) disk pool before winter conferences (expect requests from physics groups soon)
- Working to bring to production-level: consultation with dCache/CMS experts requested
- If/when project succeeds, we would like to discuss long-term support (i.e. After U.Michigan) with CD Central Services