



d-Cache and Grid Enabled Analysis

lan Fisk June 23, 2003

Introduction



What dCache is?

US CMS

- D-cache is a disk caching system developed at DESY as a front end for Mass Storage Systems
 - It now has significant developer support from FNAL and is used in several running experiments
- We are using it as a way to utilize disk space on the worker nodes and efficiently supply data in intense applications like simulation with pileup.
 - Applications access the data in d-cache space over a POSIX compliant interface. The d-cache directory (/pnfs) from the user perspective looks like any other cross mounted file system
 - Since this was designed as a front-end to MSS, once closed, files cannot be appended
- Very promising set of features for load balancing and error recovery
 - d-cache can replicate data between servers if the load is too high
 - if a server fails, d-cache can create a new pool and the application can wait until data is available.





Michael Ernst and I demonstrated it could be used for CMS very high data rate applications as a replacement for some of the Objectivity AMS functionality

- CMS event building with pile-up is the most intense application
 - 70MB of pile-up accessed at random per event simulated
 - Pile-up sample is large and needs to be spread across many servers
- dCache can handle the loss of a data server without killing the application
- dCache is capable of making multiple copies of data to balance the load across data servers







Data Delivered to processes during 12 digi jobs 12 systems test



Fairly flat and stable across the test

lan M. Fisk Fermilab

CPT Week



Load Balancing



C I	Micrp	ncoft Dowe	Doint - Fon may]				- 🕑 Micr	peoft DowerDoint - Fon m	ا بد						
	dCache ONLINE - Microsoft Internet Explorer						- Die Eile	🚰 dCache ONLINE - Microsoft Internet Explorer							
	Fi	Hie Edit View Pavorites Tools Heip ← Back ← → ← ③ ② ③ ① ③ ③ Search ③ Favorites ④ Media ③ □ □ → → → ○ ③ ② ① △ □ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○					- -	File Edit View Favorites Tools Help							_
ſ	4						<u>د</u>	→ Back → → → · · · · · · · · · · · · · · · ·							
1	Ad						7	Address 🕲 http://cms-dcache-serv.sdsc.edu:2288/queueInfo							
- (Lin		compute-1-10_1	compute-1-10Domain		100		Links 🕘 Customize Links	Eree Hotmail	🕘 Windows 🛛 👗 Free AOL	& Unlimited I	internet	RealPlayer	🕘 Wind	0
²۱	Γ		. 1 11 1			100	8	compute-1	-10_1 c	ompute-1-10Domain	25	100	0	0	
٦ſ			compute-1-11_1	compute-1-11Domain	20	100		compute-1	-11_1 c	ompute-1-11Domain	21	100	0	0	
ૈ			compute-1-12_1	compute-1-12Domain	71	100	9	compute-1	-12_1 c	ompute-1-12Domain	32	100	0	0	Ī
4			compute-1-13_1	compute-1-13Domain	60	100		compute-1	-13_1 c	ompute-1-13Domain	31	100	0	0	l
ļ			compute-1-14_1	compute-1-14Domain	80	100	10	compute-1	-14_1 c	ompute-1-14Domain	31	100	0	0	Î
5	Įτ		compute-1-15_1	compute-1-15Domain	90	100	11	compute-1	-15_1 c	ompute-1-15Domain	31	100	0	0	l
ء			compute-1-16_1	compute-1-16Domain	80	100		compute-1	-16_1 c	ompute-1-16Domain	31	100	0	0	Ī
ૌ			compute-1-17_1	compute-1-17Domain	57	100	12	compute-1	-17_1 o	ompute-1-17Domain	31	100	0	0	l
7			compute-1-18_1	compute-1-18Domain	52	100		compute-1	-18_1 c	ompute-1-18Domain	35	100	0	0	Î
c c			compute 1 1 1	compute 1 1Domain	0	100	13	compute-1	-19_1 c	ompute-1-19Domain	33	100	0	0	ĺ
8			compute-1-2_1	compute-1-1Domain	0	100	14	compute-	1-1_1 c	:ompute-1-1Domain	23	100	0	0	Î
9			compute-1-3 1	compute-1-3Domain	0	100	1 ¹⁴	compute-3	1-2_1 c	compute-1-2Domain	20	100	0	0	ĺ
٦ ر			compute-1-4 1	compute-1-4Domain	0	100	15	compute-	1-3_1 o	:ompute-1-3Domain	29	100	0	0	Ĩ
10			compute-1-5_1	- compute-1-5Domain	0	100		compute-1	1-4_1 o	compute-1-4Domain	25	100	0	0	Ī
			compute-1-6_1	compute-1-6Domain	0	100	16	compute-	1-5_1 o	compute-1-5Domain	16	100	0	0	Ĩ
11			compute-1-7_1	compute-1-7Domain	0	100] ₁₇	compute-	1-6_1 c	compute-1-6Domain	18	100	0	0	ĺ
12			compute-1-8_1	compute-1-8Domain	0	100	1 <i>'</i>	compute-	1-7_1 c	compute-1-7Domain	15	100	0	0	Ĩ
			compute-1-9_1	compute-1-9Domain	0	100	18	compute-	1-8_1 o	compute-1-8Domain	23	100	0	0	Ì
.▲ Di	aw -			Total	520	2010		compute-	1-9_1 o	compute-1-9Domain	32	100	0	0	Ĩ
		Done					•		Tot	al	520	2010	0	0	l
😹 Start 🛛 🥭 🦃 🕨 🖃 📾 🗗 🗗 🖗 🖉 🖉 🖉 🖉 😓 🗧															
Ian M. Fisk Fermilab												i si s	a		ĺ





What dCache isn't?

- By itself dCache doesn't have any distributed data handling capabilities
 - It was designed to improve access to MSS data on a local site
 - Access to data is handled by UNIX file permissions
 - Local File Catalogue is NFS file system
- dCache has a lot of the elements that you want for a distributed analysis system
 - Fault Tolerance
 - Load Balancing
- Without addition of wide area transfer mechanism and a distributed data catalogue it doesn't have anything more to do with Grid Enabled Analysis than NFS does





Services are needed to build upon local data distribution tools to provide distributed and largely transparent data access

- Need implementations of RLS and RLI services to identify and distribute the data files available at sites
- Need tools for wide area replication with strong authentication
 - Needs to be linked to local data providing tools
- Need functionality to provide a data gateway for transfers in and out of centers
 - Much like the globus gateway currently used for batch requests





There has been development on an SRM interface directly to dCache

- This provides a robust and efficient wide area transfer tool
- Provides a uniform interface to a variety of MSS

The proposal by members of the LCG and US-CMS is to create a Storage Element (SE) based on RLS and RLI data catalogue, SRM and a number of local data distribution tools

- dCache is promising because it is POSIX compliant and doesn't require changing the application
- Other local tools being investigated are RFIO, ROOTD, and simple NFS file systems





A lot of the SE development has been planned with reconstruction users in mind

- Infrastructure planned is advanced, but it still expects a level of predictability in terms of the data transferred.
- At some point resource brokers will attempt to make intelligent scheduling decisions based on data location

The grid enabled analysis user doesn't currently have the infrastructure to know if a request is reasonable or possible

The production and reconstruction user has a much more organized and predictable environment to work in.





As the Storage Element Development proceeds it would be interesting work on developed the applications that help the user estimate the feasibility of an analysis requests

- User interfaces to RLS and RLI catalogues
- Estimation Tools
 - Time for transfer estimation tools based on bandwidth measurements
 - Amount of data to access