Gabriele Compostella University of Trento and INFN compostella@tn.infn.it

on behalf of CDF Computing Group

CDF Experience using the Grid

Workshop 2007 su Calcolo e Reti dell'INFN Rimini

8 Maggio 2007

CDF Experiment Luminosity



CDF Experiment Data Volumes



CDF Data Production





Data Handling model relies on SAM (Sequential data Access via Metadata), a centralized system that manages all official data

SAM

- File Catalog for data on tape and on disk:
 - Files based, with metadata attached
 - Files are organized in datasets, users can create their own
- Manage File Transfers: tape-to-cache, cache-to-disk, disk-to-tape
- For jobs accessing data:
 - copy necessary files from the closest location to local cache
 - provide files in the optimal order
 - keep track of processed and unprocessed files
 - Failed sections can be automatically recovered
- Each site serving data needs to have a SAM Station

Users Analysis and MC Production

- CAF Central Analysis Facility: mostly User analysis FNAL hosts data -> farm used to analyze them
- Decentralized CAF (dCAF): mostly Monte Carlo production Remote sites produce Monte Carlo, some (ex. CNAF) have also data Produced files are then uploaded to SAM
- CAF and dCAF are CDF dedicated farms

CAF Model



The CAF Headnode



Job Monitoring

,		FNAL C	AF Web Moni	itor - Mozill	a Firefox					_ •	×		_				9
ile <u>M</u> odifica <u>V</u> isualizza V <u>a</u> i S <u>e</u> gnalibri <u>S</u> tri	menti <u>G</u> uida									a m a	1						1
🔹 - 😓 🛽 🚯 🚺 http://cdfcaf.fnal.gov/groupcaf/									<u>2241</u>			<u>2242</u>		2243		2244	
CAF Web Monitor [History Analyze]									Running Ready:Apr 10 22:34 Started:Apr 10 22:42 Condor ID: 306503			Running Ready.Apr 10 22:34 Startad.Apr 10 22:43 Condor ID: 3065041		Running Ready.Apr 10 22:34 Startad:Apr 10 22:42 Condor ID: 3065044	Completed RetCode: 0 (Real: 10h 38; CPU: 9h 37') ReadyApr 10 22:35 Stattad:Apr 10 22:42 Ended:Apr 11 09:21 Condor ID: 3005047		
<u>site Overview</u> CAF C <u>System status</u> System	CAF overview _[History] System Info										2:35 22:43 3065050	2246 Runnin Ready:Apr Started:Apr Condor I	1g 10 22:35 r 10 22:43 ID: 3065053	2247 Running Ready:Apr 10 22:35 Started:Apr 10 22:43 Condor 1D: 3065057	2248 Ru Rea Star Cor	i nning .dy:Apr 10 22:36 .ted:Apr 10 22:43 ndor ID: 3065059	
	atal VMa IV		imed I e	ad-O I a		nai an a d	Enco V/	I Lood I		<u>2249</u>		2250	2	<u>2251</u>		<u>2252</u>	
bs status					163	28	171	1052 1	Comp	leted RetCo	de: 0	Dunning		Dunning	Derest		
y accounting group:	2000 2	2490 2	2297 1	008	183	38	35	History	(Real: 7 Ready:Ap	7h 46', CPU: 6 or 10 22:36	h 52')	Kunning	<u></u>	Kunning	Running		
• common	Avg. 2972 2477 2403 228 183 38 35 <u>Abb.</u>							<u>1110001 j</u>	Started:A Ended:A	pr 11 01:43 pr 11 09:30	User:	andrew Length: long			Load on vm2@fcdfca	f979.fnal.gov	
• group_dqm Total C/	Total CAFMarks Claimed CAFMarks Free CAFMarks VM Load CAFMarks Avg. CAF								Condor	Condor ID: 3065062		nting group_MCprod.andrew				·	
• group_florida 270	2700.0k 2511.5k 188.5k 2078.6k 10						107	9.0		User:	t						
• group_mai • group_geneva										-	Source: none				CPU		
group_highprio/h	/hour Started sections Finished sections Submitted jobs Terminated jobs										Status:	Running	Load: 0	.99	0.0 - 22:00 00:00 02:00 04:00 06:00		
group_italy Last 10 m	Last 10 mins 318.00 1014.00 6.00 18.00									Submitted: Apr 06 11:41 Ready: Apr 07 20:32			pr 07 20:32	■ Load Created on Wed Apr 11 09:46:11 2007			
group physmon	Last 60 mins 344.90 533.85 8.00 14.00									Started: Apr 10 20:37			01-	Memory usage on vm2@fcdfcaf979.fnal.gov			
• group_pitt			~								Usea u VM:	wm2@fcdfcaf	P79 fnal dov	211	40		
• <u>group prd</u> • group rochester	Sections by Accounting Group(ordered by AcctGroup)										Site: FermiGridCDF Entry: fnalcdf1 5 002			nalcdf1 5 002	20 × 20		
• group sam											Condor	ID: 94162	Schedd: S	chedd_1@fcdfhead5.fnal.gov	£		
• <u>group_uk</u> <u>Accoun</u>	ting Quot	a <u>Running</u>	Assigned	Idle	Wait	Held	Completed	l <u>Removed</u>	Total	Jobs					0 22:00	00:00 02:00	04:00 06:00
/ user:	700*	1262	0	514	3264	0	5870	7	10917	84 Histo					MemUsed Creat	ted on Wed Apr 11 09:46:1	4 2007
aaltonen groun da	n 50	0	0	7	0	0	0	0	7	7 Histo	Dunan						
• andrew				_		0	10		50	4 I Ii-+-	Proces	es [<u>Hide]</u>					
mpletato						A	dblock 🔏	FoxyProxy: Mod	delli í í 🗐 A	Apri blocco note	PID STAF 5052 20:34	FED %CPU VSZ CMD 53 0.0 2324 /bin/ba	ash /grid/home/c	df/.qlobus/.gass_cache/local/md5/ff/7	/e28732a0dcdf79887	7ele24d6ec44/md5/54/2fcc	83232a5bd52266e07945i
Moh hac		lor	ito	r. i	nfa	rm	sti/	h	1919.		5301 20:34 5366 20:34	57 0.0 2316 /bin/ba	ash ./condor_sta	rtup.sh glidein_config _dvn _f			
	eu r		IILU		IIIC	7111	all										

on running jobs for all users and jobs/users history

Interactive job Monitor: Available commands:

Latmon list	Catmon kill
CafMon dir	CafMon tail
CafMon ps	CafMon top

ș Carmon	Jobs						
Analysis Job	Farm:	lcgcaf Group	Host: From	pcdf11 To	L.pd Sta	.infn.it atus	
1208		short	Tota	1: 5	50		
1208		short	1	11	L Per	nding	
1208		short	7	15	5 Ru	nning	
1208		short	16	50) Su	ccess	
1208		short	Succ	ess:	42	Pending:	8
1208		short	Succ	ess:	84%	Pending:	16%
\$							
	\$ CarMon Analysis Job 1208 1208 1208 1208 1208 1208 1208 \$	\$ carmon jobs Analysis Farm: Job 1208 1208 1208 1208 1208 1208 1208 5	Analysis Farm: lcgcaf Job Group 1208 short 1208 short 1208 short 1208 short 1208 short 1208 short 1208 short 1208 short	Analysis Farm: lcgcaf Host: Job Group Prom 1208 short Tots 1208 short 1 1208 short 1 1208 short 1 1208 short 16 1208 short 5ucc 5	Analysis Farm: lcgcaf Host: pcdfl: Job Group From To 1208 short Total: 5 1208 short 1 11 1208 short 7 15 1208 short 16 5 1208 short Success: 1208 short Success: 5	<pre>Analysis Farm: lcgcaf Host: pcdfll.pd Job Group From To St: 1208 short Total: 50 1208 short 1 11 Pei 1208 short 7 15 Ru 1208 short 16 50 Su 1208 short Success: 42 1208 short Success: 84%</pre>	Analysis Farm: lcgcaf Host: pcdfll.pd.infn.it Job Group From To Status 1208 short Total: 50 1208 short 1 11 Pending 1208 short 7 15 Running 1208 short 16 50 Success 1208 short Success: 42 Pending: 1208 short Success: 84% Pending: 5

CDF batch computing in numbers

Up to 4000 batch slots (last 3 months) Almost 800 registered users, Approx. 100 active at any given time

5.0 k

4.5 |

4.0 k

3.5 k

3.0 k

2.5

2.0 1

1.5 k

1.0 k

0.5 k

ASCAF

Week

FermiGrid

Nr. of VMS



on Mon Apr 9 11\:11\:37 2007

Created

10

Moving to the Grid...



- Need to expand resources, luminosity expect to increase by factor ~4 until CDF stops taking data.
- Resources were in dedicated pools: limited expansion and need to be maintained by CDF personnel
- As a first step, we can move MC jobs to the Grid (no data access...)
- Keep on using CAF (and CNAF) for data analysis.
- CDF developed 2 portals to the GRID:
 - NamCAF for OSG and LcgCAF for LCG

NamCAF Condor Glidein based CAF

More info at: http://cdfcaf.fnal.gov http://cdfcaf.fnal.gov/namcaf

CDF use of Condor glide-ins



CDF usage of **OSG** resources per site



- ...Up to 1500 batch slots
- Self contained tarballs
- No specific support on Grid sites

http://cdfcaf.fnal.gov/namcaf/

LCGCAF CDF Submission Portal to LCG

More info at: http://www.ts.infn.it/cdf-italia/public/offline/lcgcaf.html

LcgCAF: General Architecture



Job Submission and Execution

kx509: transforms user's CDF kerberos ticket to valid Grid x509 proxy certificate

Kdispenser service: sends kerberos ticket to worker node, needed to rcp output to CDF data servers

Submitter, completely rewritten for lcg tools:

- internally enqueue user job, then send to WMS
- create job wrapper to run the job on GRID environment
- make available user tarball

Job wrapper:

- discover site and choose correct configuration/proxy
- copy user tarball from http server to WN
- run user job and fork monitoring processes
- copy whatever is left at the end of user code to user specified location

Job Monitor



Web based Monitor: information on running jobs for all users and jobs/users history are read from LcgCAF Information System running jobs information are read from LcgCAF Information System and sent to user desktop. Available commands: CafMon list CafMon kill CafMon dir CafMon tail CafMon ps CafMon top

CDF Code Distribution



Parrot is used as virtual file system to distribute CDF code:

- It's just a wrapper: when a user job needs a file/library not available to the worker node, parrot puts the job on hold and downloads the required file from our code server
- No specific software requirement on site
- Only the portions of CDF code that are really needed by the job are transferred to the worker node
- No need to choose in advance what to transfer

To have good performances with Parrot, a local Squid web proxy cache is also required

Frontier

Run Condition DB distribution for data Analysis and MC production



Prevent central DB overloads

Improvements up to 60% in certain type of jobs

LcgCAF Usage

- LcgCAF opened to all users and officially supported by CDF since November 2006
- European CDF portal to LCG resources
 Currently an average of 10 users running, peaks of more than 2500 running sections



Sites accessed trough LcgCAF

INFN-Padova	Italy
INFN-Catania	Italy
INFN-Bari	Italy
INFN-Legnaro	Italy
INFN-Roma1	Italy
INFN-Roma2	Italy
INFN-Pisa	Italy
FZK-LCG2	Germany
IFAE	Spain
PIC	Spain
IN2P3-CC	France
UKI-LT2-UCL-HE	UK
Liverpool	UK

Week 1/

Performances:

efficiency ~95% using wms 3.0 (discarding temporary sites inefficiencies)

Issues:

- wms 3.0 bugs and stability problems
- problems with matchmaking, logging and bookkeeping
- misconfigured sites, ex. clock problem: if date and time are not correct, kerberos authentication doesn't work, so it's impossible to move job output to finale destination

Now testing WMS 3.1

CDF Usage of LCG resources



(European sites not included in this plot, only italian sites)

Conclusions

...GRID can serve also a running experiment!

CDF proved to have an adaptable, expandable and succesfull computing model

Great effort has been put on exploiting GRID:

- Used standard and common tools (http + squid)
- No specific software requirement on sites
- Resources are used oportunistically
- Nothing changes from the user point of view
- MC production is completely moving towards distributed systems

Future Developments

- Plan to add more resources into NamCAF and LcgCAF
- Plan to implement a mechanism of policy management for LcgCAF on top of GRID policy tools:
 - now testing GPBox for policies enforcement at CE level
 - waiting for DGAS: we need good accounting to be able set priorities dinamically
- Developed SAM-SRM2 interface to allow:
 - output file declaration into catalogue directly from LcgCAF
 - automatic transfer of output files from local CDF storage to FNAL tape
 - waiting for a SRM2 compliant storage element
- Explore the possibility of running on data moving jobs towards sites that can serve the desired files

Backup

GlideCAF: Condor Binaries Transfer



NamCAF: Working with remote Grid sites



NamCAF: How GCB works



CDF usage of **OSG** resources

Integrated CPU time consumed per VO



http://grid02.uits.indiana.edu:8080/show?page=index.html

Overview of Parrot



Parrot Performances with Http in a CDF-like job



LcgCAF Output Storage



- User job output is transferred to CDF Storage Element (SE) via gridftp or to a CDF-Storage location via rcp.
- From CDF-SE or CDF-storage output copied using gridftp to FNAL fileservers.
- Automatic procedure copies files from fileservers to tape and declares them to the CDF catalogue.
- Planning to use a SRM-SE interface for big sites.

LcgCAF Performances

CDF B Monte Carlo production by Power Users in July 2006: ~5000 jobs in ~1 month

Samples	Events on tape	3	ϵ with 1 Recovery
$\mathbf{B}_{s} \rightarrow \mathbf{D}_{s} \pi \mathbf{D}_{s} \rightarrow \phi \pi,$	~6,2 · 10 ⁶	~94%	~100%
$D_s \rightarrow K^*K, D_s \rightarrow K_sK$			NO TO ALL

Possible Failures:

• GRID: site miss-configuration, temporary authentication problems, WMS overload, LB "loose" informations

• LcgCAF: Parrot/Squid cache stacked (Parrot cache coherency problems are under investigation)

Automatic Retry mechanism:

- GRID failures: WMS retry shallow retry
- LcgCAF failures: "home-made" retry (log files parsing) only for certain type of jobs