Use of glide-ins in CMS for production and analysis



On behalf of the CMS Offline and Computing Projects

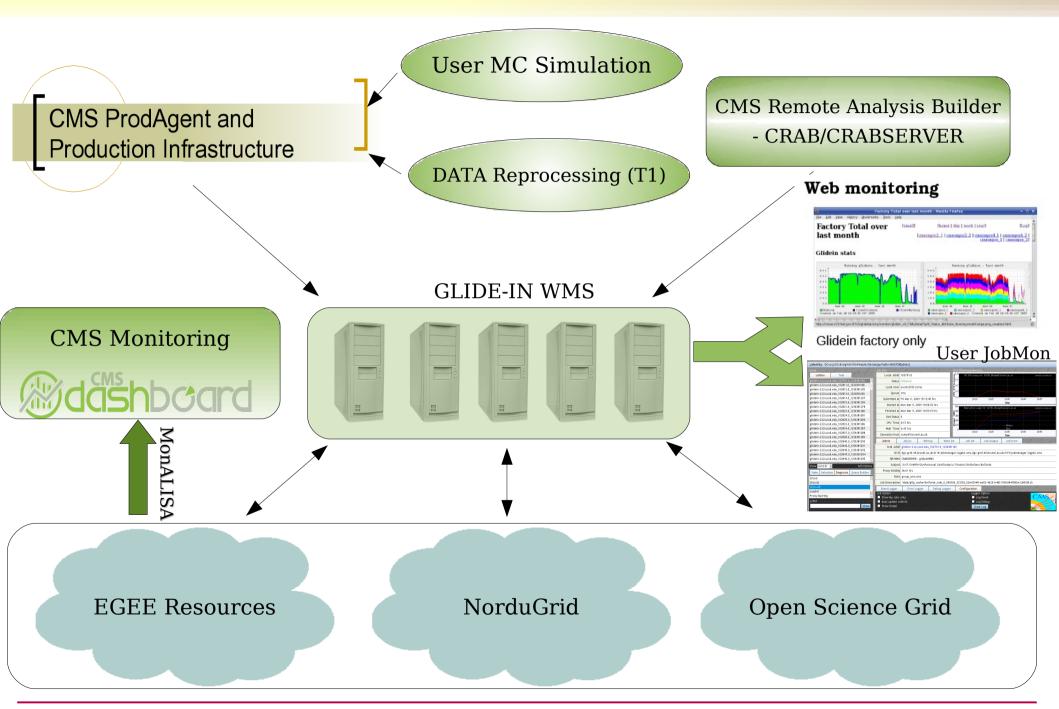
Sanjay Padhi
University of California, San Diego (UCSD)
With

Oliver Gutsche (FNAL), Kristian Hahn (MIT), Burt Holzman (FNAL), Dave Mason (FNAL), Haifeng Pi (UCSD), Igor Sfiligoi (FNAL), Eric Vaandering (FNAL), Frank Wuerthwein (UCSD)

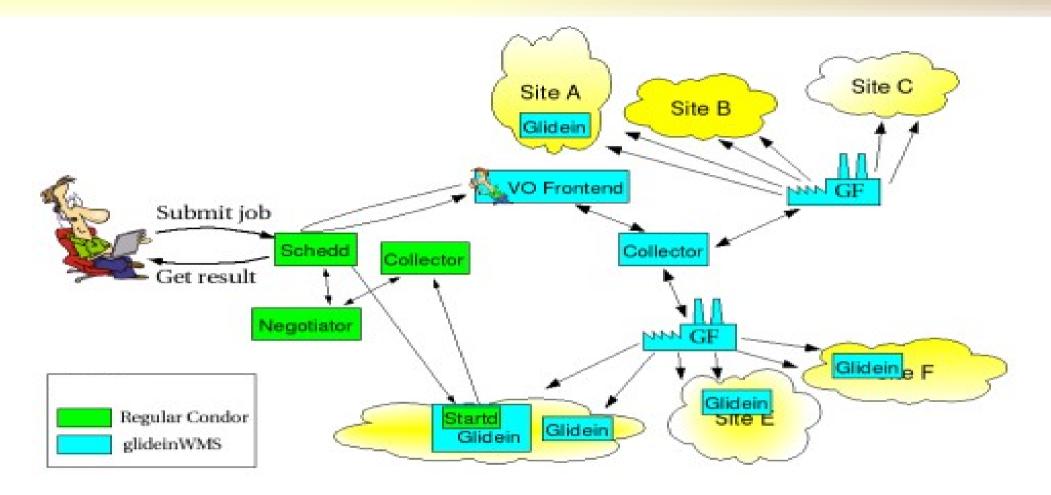
Outline

- Components using glide-in based WMS
- GlideinWMS
- Scalability of the system
- Use of glideinWMS for Data Reprocessing
 - Data reprocessing at the Tier 1s.
 - Events Merging
- Use of glideinWMS for Data analysis @ T2s
 - Analysis experience during CCRC-08
 - Integration of glideinWMS with Crabserver
 - JobRobots using glideinWMS via Crabserver
 - Recent results
 - Use for Monte Carlo Simulations
- Coherent Monitoring interface to the system
 - JobMon and Pseudo-Interactive tools
- Experience using next generation CREAM CE

Components using glide-in based WMS



Glide-in based WMS



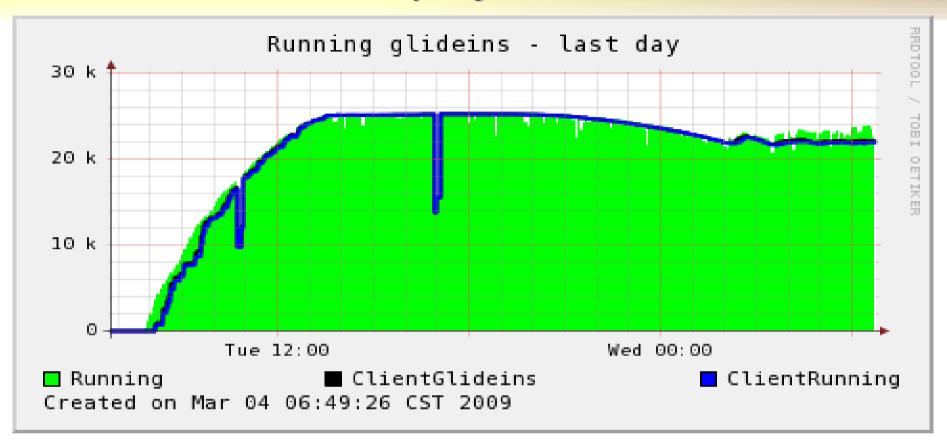
Glide-ins are Grid jobs that start regular Condor daemons

i.e are pilots implemented with Condor

Glide-in WMS is an autonomous glide-in submission system

Uses GSI authentication for all interactions with Condor daemons over the network

Scalability of glideinWMS

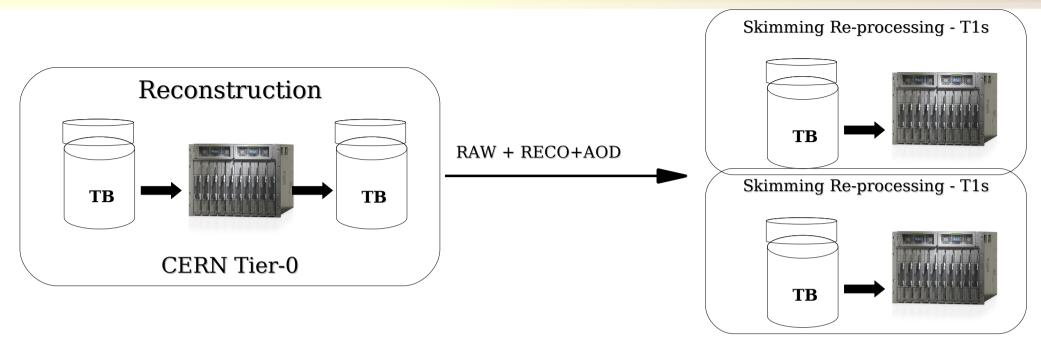


Using a test setup the glide-ins along with clients (user jobs) were tested intensively For details:

- "Interoperability and Scalability within glideinWMS" by D. Bradley, ID-218, CHEP 09.
- Scaled up to running 22,000 25,000 clients/jobs in parallel
- Submitted more than 500K jobs, running on average of 3 hour duration

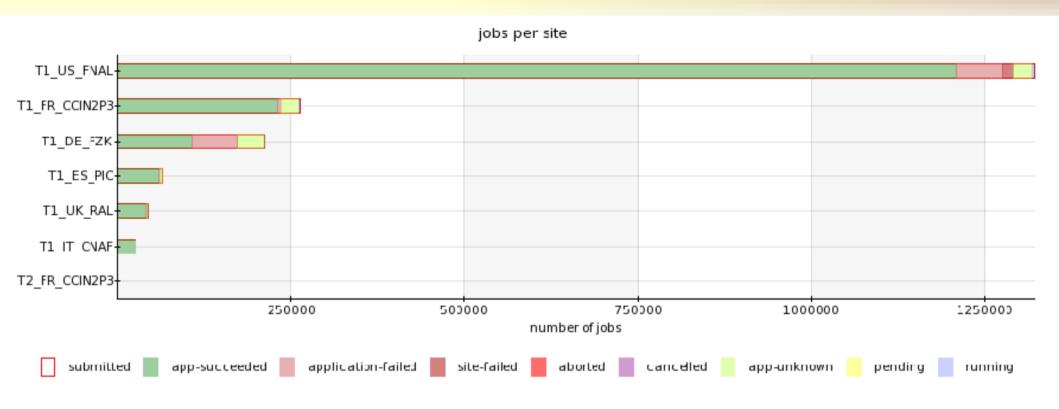


Use of glideinWMS for DATA Reprocessing



- ◆ Tier-1s receive a continuous stream of RAW and reconstructed data
- Data skimming and re-reconstructions are performed at the Tier1s.
- Data transfers are managed by PhEDEx
- Description of Data files organized in blocks are kept in global DBS
- ◆ ProdAgent, is used to Skim and re-process → uses glideinWMS
- Late binding features in glide-ins provides a homogeneous pool of T1 resources
 - Ensures proper CMSSW ENV, publishes installed Versions at the site/WNs
 - Provides reliability in reprocessing jobs/workflows (Shields from faulty WNs)

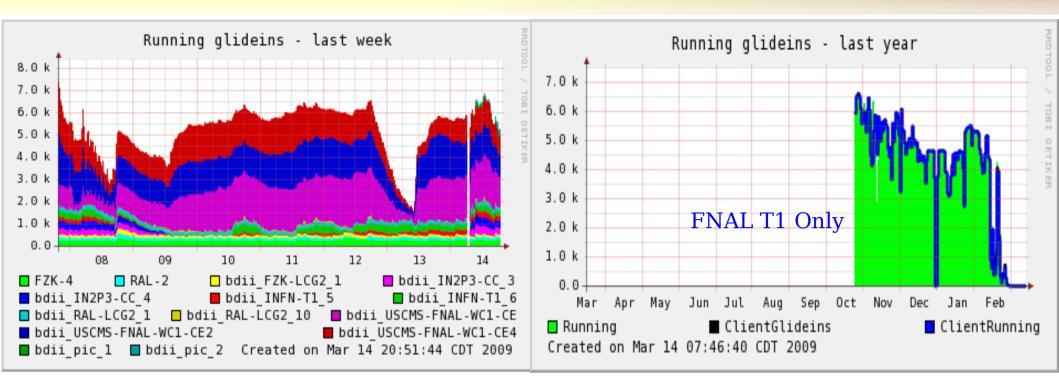
DATA Reprocessing @ Tier1s



Results shown starting Jan 2009: Overall very good efficiency !!!

		current status					grid exit status						appli	/		status		
site	Info	Sub	Pend	🛚 Run 🔼	Term	Done	'∕ Canc Y	Aport	V Unk V	Grid%	Succ ~	AppFail	SiteFail	AllFail	Unk >	App%	Site%	Overal\%
T1_DE_FZK	Link	211988	<u>C</u>	<u>70</u>	211918	<u>0</u>	<u>0</u>	<u>0</u>	211918	0	109982	63669	3	63672	38264	63.33	100	51.9
T1_ES_PIC	<u>Link</u>	65615	<u>C</u>	<u>0</u>	65615	<u>0</u>	<u>0</u>	<u>0</u>	65615	0	51048	1205	<u>0</u>	1205	3362	98.06	100	93.04
T1_FR_CCIN2P3	<u>Link</u>	263885	<u>C</u>	<u>655</u>	263230	<u>0</u>	<u>0</u>	<u>0</u>	263230	0	232748	5397	<u>0</u>	<u>5397</u>	25085	97.73	100	88.42
T1_IT_CNAF	<u>Link</u>	27912	<u>C</u>	233	27679	<u>0</u>	<u>0</u>	<u>0</u>	27679	0	27482	192	<u>0</u>	192	<u>5</u>	99.31	100	99.29
T1_UK_RAL	<u>Link</u>	45604	<u>C</u>	<u>187</u>	<u>45417</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>45417</u>	0	<u>42540</u>	845	<u>0</u>	<u>845</u>	2032	98.05	100	93.67
T1_US_FNAL	<u>Link</u>	1323278	3 <u>C</u>	<u>3353</u>	131992	<u>5 0</u>	<u>0</u>	<u>0</u>	1319925	0	1208873	66580	16437	83017	28035	93.57	98.75	91.59
T2_FR_CCIN2P3	<u>Link</u>	524	<u>C</u>	<u>0</u>	<u>524</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>524</u>	0	<u>511</u>	<u>11</u>	<u>0</u>	<u>11</u>	2	97.89	100	97.52

DATA Reprocessing @ Tier1s



Data Reprocessing, skimming, merging etc. requires a large number of resources Many of the task workflows are short jobs.

CEs can afford a rate ~ 0.1Hz submissions (I. Sfiligoi, ID 216, CHEP 07)

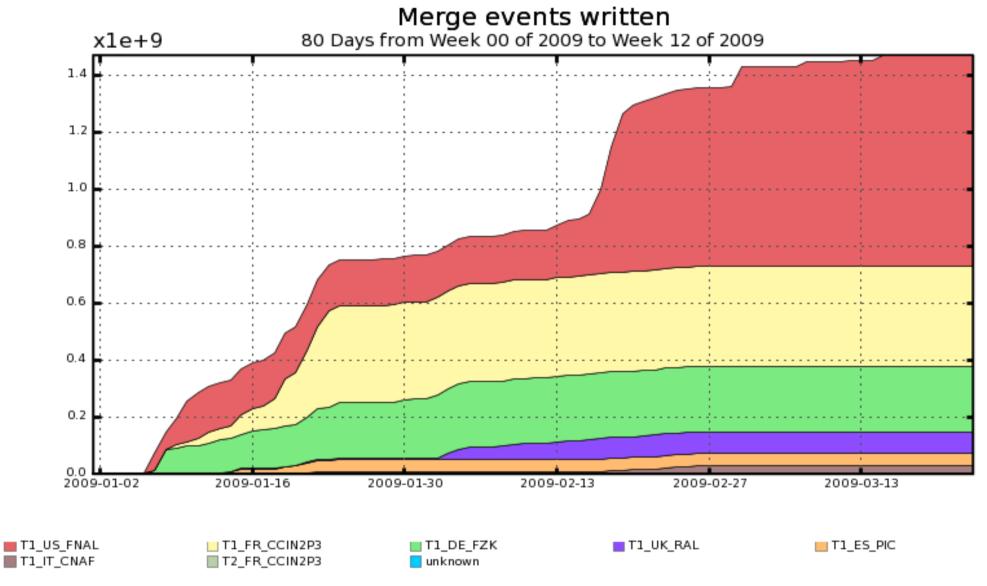
- In order to fill 6000 CPUs will require 16.67 hours.

Allow "long" lived glide-ins

- each can pull jobs sequentially until expiry of their lifetime/proxy
- Scalable solution required at FNAL in order to "fill" the resources

Prodagent can use glideinWMS in a production mode for all reprocessing activities

Events processed using glideinWMS

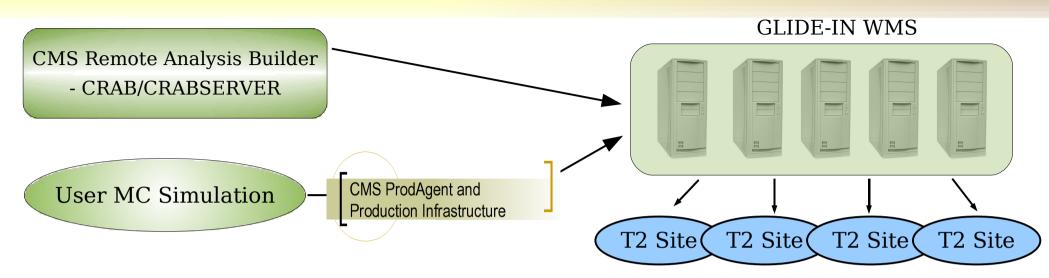


Total: 24,961,648 , Average Rate: 3.57 /s

Results shown starting Jan 2009 (80 days) - Very successful workflow

Use of glideinWMS for DATA analysis

Use of glideinWMS for DATA analysis



- CMS Remote Analysis Builder (CRAB) uses both late & early binding WMS
 - On early binding See talk by Giuseppe CODISPOTI, ID-139, CHEP 09
- Interacts with data discovery services (DBS and DLS) to split task into smaller jobs
- Provides input to the WMS the SE names
- SE to CE mapping is then performed using BDII
- Jobs with "DESIRED Gatekeepers" are then matched via glideinWMS
- Gfactory sends glide-ins to the desired sites.
- Once the glide-ins ensure the reliability of the node along with requested CMSSW
 - Jobs directly flow to the Worker Node
- BOSSLite is used to interact:
 - with the glideinWMS scheduler to provide instantaneous updates

Analysis experience during CCRC-08

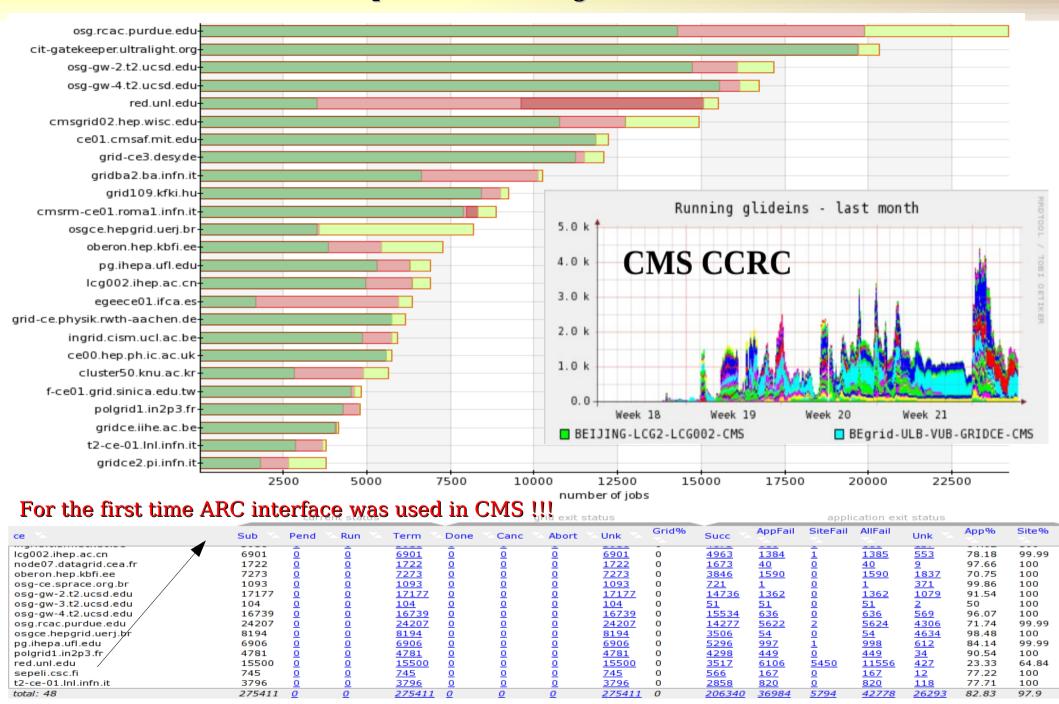
Various analysis exercises were performed during CCRC-08

Goal was to gain overall understanding (See - A. FANFANI, ID 213, CHEP 09)

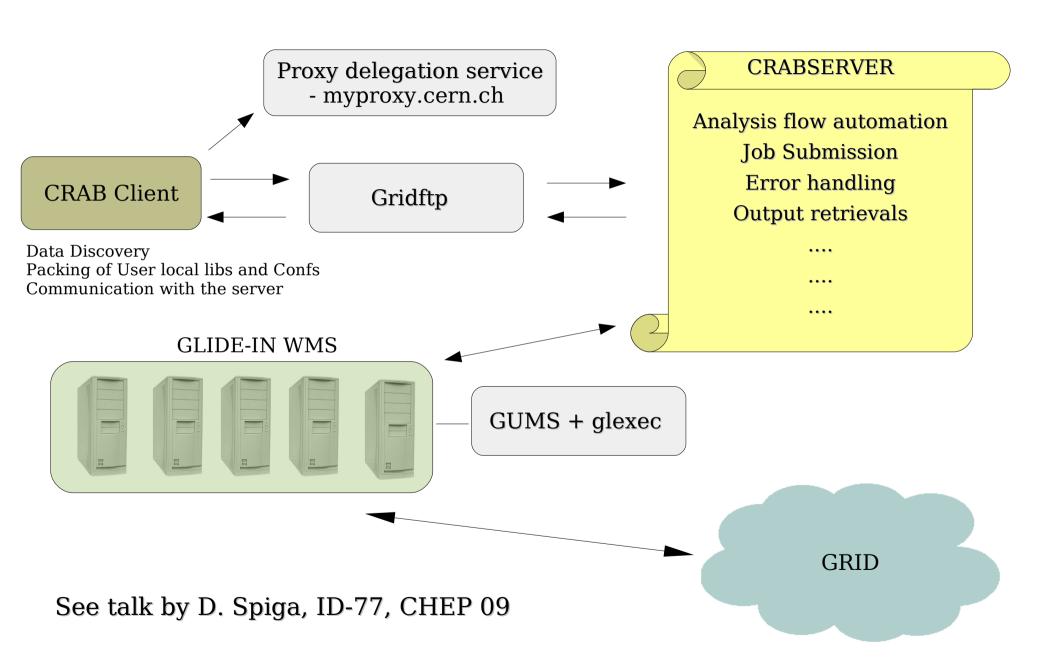
- of the performance and readiness of the CMS Tier-2 sites for data analysis
 CRAB client using glideinWMS and gliteWMS were used
 Site performance was evaluated based on increasing in complexity of the jobs
 - Long running CPU intensive jobs with moderate I/O
 - Long running I/O intensive jobs
 - Short-running jobs with local stage out of O(10MB) files
- Over 40 sites across EGEE, OSG and Nordugrid were involved
- Mix of errors at a few sites due to catastrophic storage failures

Overall success rate (without SE issues) ranged from 92-99% (Based: more than 100k submitted jobs)

Experience during CCRC-08



Integration of glideinWMS with Crabserver



JobRobots using Crabserver based glideinWMS

JOB DETAILED VIEW

total: 25987 (seeing 0 to 50)

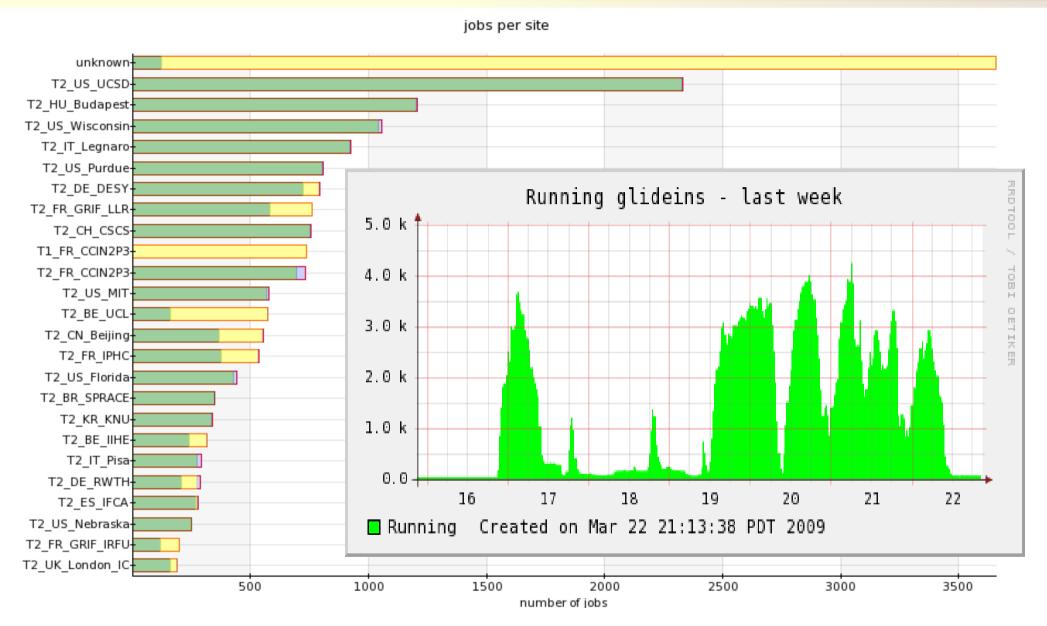
prev 50 next 50

num	JobId		Grid Status	ExitCode	IdInTask	NEvReq	NEvProc	Submitted	Started	Finished	Task	IP	Target
1	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/1	T2_BE_IIHE	DONE	0	1	642		00:10:14	00:40:50	00:30:33	<u>belforte_crab_0_090223_010949</u>	193.190.247.140	44_Selected_SE
2	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/10	T2_BE_IIHE	DONE	8001	10	0	0			2009-02-23 00:42:50	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
3	<u>https://glidein</u> /00d2173d96cb6bc45c1271ee10ca54017f665126/100	T2_BE_IIHE	DONE	0	100	642	16/17			2009-02-23 01:13:48	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
4	/00d21/3d96cb6bc45c12/1ee10ca5401/1665126/11	T2_BE_IIHE	DONE	0	11	642	16/12			2009-02-23 00:53:43	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
5	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/12	T2_BE_IIHE	DONE	8001	12	0	10			2009-02-23 00:42:53	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
6	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/13	T2_BE_IIHE	DONE	0	13	642	6/12			2009-02-23 00:53:43	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
7	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/14	T2_BE_IIHE	DONE	0	14	642	6/17	00:10:15	00:43:40	00:55:11	<u>belforte_crab_0_090223_010949</u>	193.190.247.140	44_Selected_SE
8	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/15	T2_BE_IIHE	DONE	8001	15	0	0			2009-02-23 00:43:49	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
9	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/16	T2_BE_IIHE	DONE	0	16	642	6/19			2009-02-23 00:57:06	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
10	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/17	T2_BE_IIHE	DONE	8001	17	0	0			2009-02-23 00:46:46	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
11	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/18	T2_BE_IIHE	DONE	0	18	642				2009-02-23 00:58:27	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
12	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/19	T2_BE_IIHE	DONE	8001	19	0	In			2009-02-23 00:46:28	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
13	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/2	T2_BE_IIHE	DONE	0	2	642	6/12			2009-02-23 00:51:00	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
14	https://glidein /00d2173d96cb6bc45c1271ee10ca54017f665126/20	T2_BE_IIHE	DONE	0	20	642				2009-02-23 00:58:06	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE
15	https://glidoin	T2_BE_IIHE	DONE	0	21	642	6/12			2009-02-23 00:59:01	belforte_crab_0_090223_010949	193.190.247.140	44_Selected_SE

Entertaining about ~25k JobRobot jobs (span of days) !!!

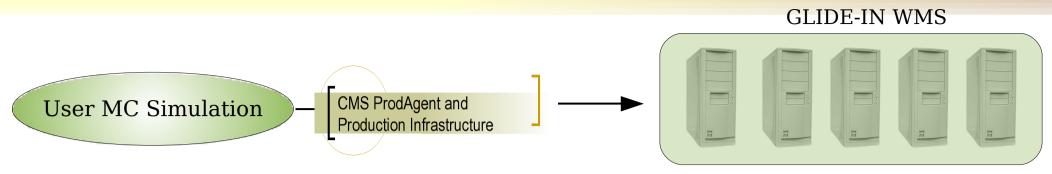
[The Job Robot is a tool to automatically create/submit analysis tasks to the grid.]

Recent results using Crabserver and glideinWMS

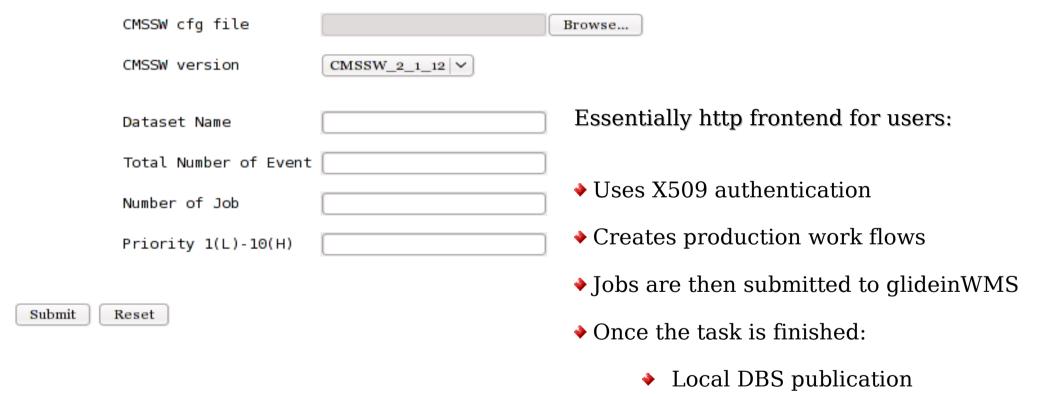


Results from last week:
Running more than 3.0 K CPUs over 38 CMS T2 sites using Crabserver

User Monte Carlo Simulation

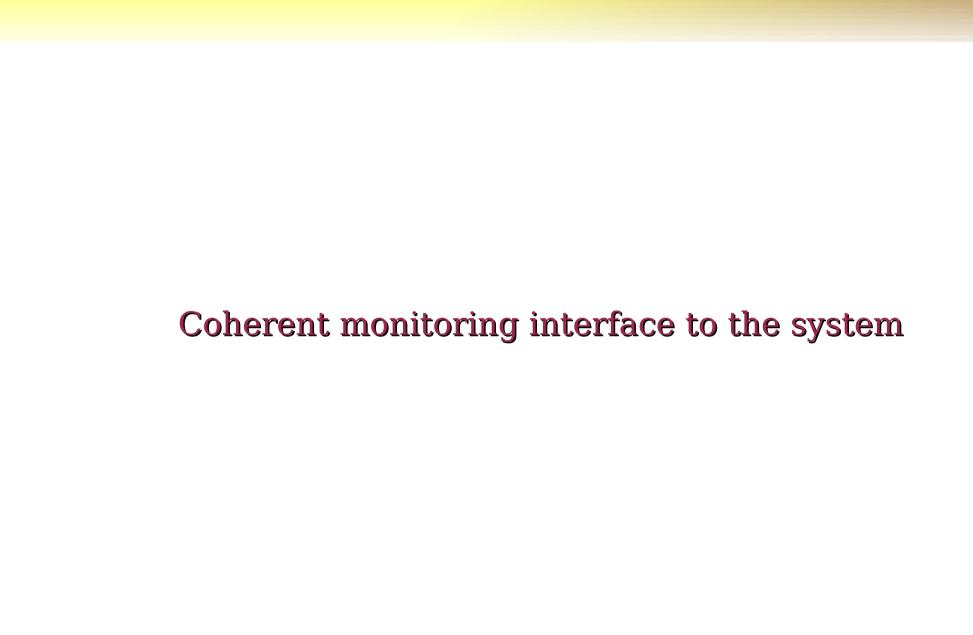


Production Request Form



User Registration

Registration



User JobMon using glideinWMS

The JobMon (See: S. Sarkar, ID 259, CHEP 09) enables users to track jobs in details. The input to the monitoring framework is taken from ClassAds of the glideinWMS.

- Uses X509 authentication
- CPU, Memory usage as a function of time
- Provides summary information

Queries are based on various machine/job attributes:

- CEs, Submission time, CPU efficiencies are supported

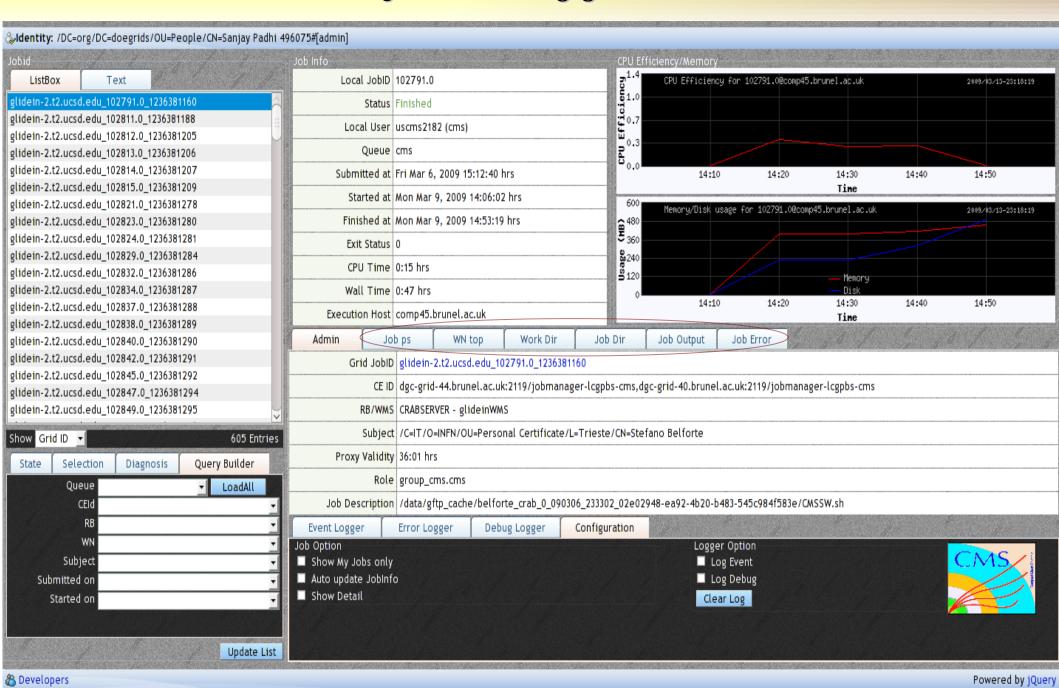
Supports (after proper authentication) the users to be able to use:

- "ls, cat, top" etc on the real-time running jobs at the WNs
- can use the pseudo-interactive tools in order to debug applications.

Similar developments from dashboard (A user-centric monitoring view)

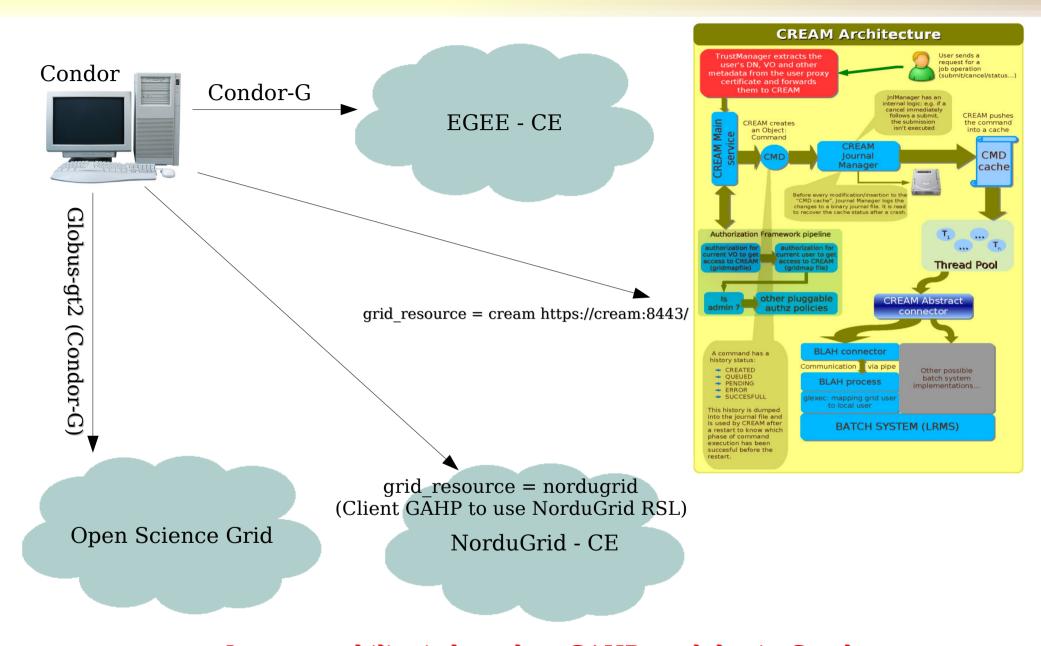
- See: Edward Karavakis, ID 121, CHEP 09

User JobMon using glideinWMS



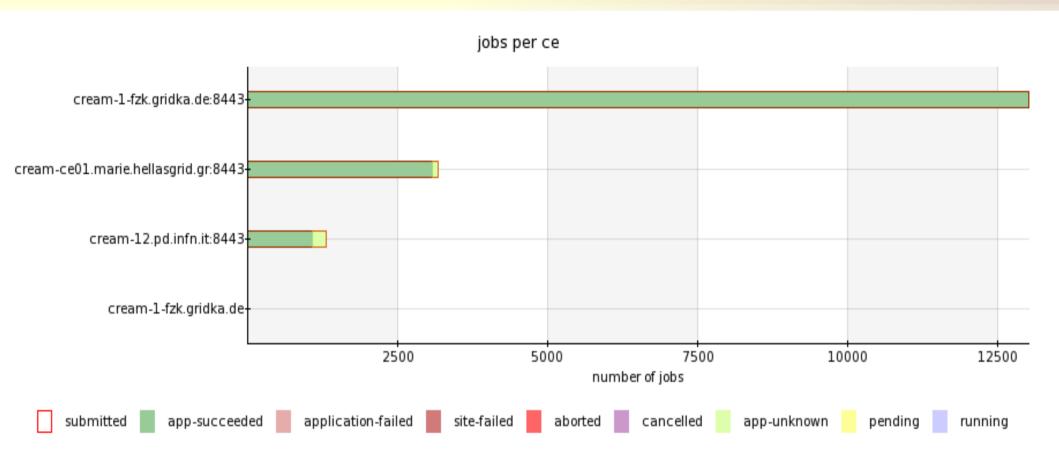
Experience using next generation CREAM CE

Interoperability with Condor



Interoperability is based on GAHP modules in Condor

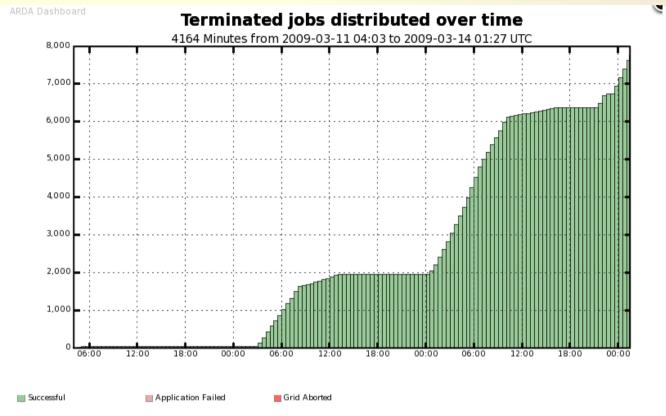
Condor tests with CREAM CE (Successful Jobs)



Results from last Week:

		curr	ent statu:	S	_		grid exit s		application exit status							
ce 🔀	Sub 🔼	Pend	∑ Run	△ Term >	Done	✓ Canc	™ Abort	¼Unk ¼	Grid%	Succ	AppFail	SiteFail	AllFail	Unk 🔼	App%	Site%
cream-1-fzk.gridka.de cream-1-fzk.gridka.de:8443	17 13037	0	0	<u>17</u> 13037	0	0	0	<u>17</u> 13037	0	<u>17</u> 13037	0	0	0	0	100 100	100 100
cream-12.pd.infn.it:8443	1314	<u>0</u>	<u>0</u>	1314	0	<u>0</u>	<u>0</u>	1314	0	1088	<u>0</u> 0	<u>0</u>	<u>0</u>	<u>226</u>	100	100
cream-ce01.marie.hellasgrid.gr:8443	3184	<u>0</u>	<u>0</u>	<u>3184</u>	<u>0</u>	<u>0</u>	<u>0</u>	<u>3184</u>	0	<u>3077</u>	0	<u>0</u>	<u>0</u>	<u>107</u>	100	100

Condor tests with CREAM CE



Maximum: 7,766, Minimum: 0.00, Average: 2,546, Current: 7,766

385 LastHoldReason = "CREAM error: CREAM_Job_Cancel Error: job status does not match"

100 LastHoldReason = "CREAM error: CREAM_Job_Purge Error: EOF detected during communication"

234 LastHoldReason = "CREAM error: CREAM_Job_Register Error: EOF detected during communication."

1 LastHoldReason = "CREAM error: CREAM_Job_Status Error: EOF detected during communication."

2000 LastHoldReason = "Failed to create proxy delegation"

165 LastHoldReason = "Unspecified gridmanager error"

About 25% of the jobs failed - Major source of errors are due to proxy renewal/delegation Engaged with the developers - Work in progress to address these issues

Summary and Conclusions

GlideinWMS is widely used in CMS for central data reprocessing & skimming @T1s

- provides the capability to fill a very large number of resources quickly and sustaining the usage with short jobs ($\sim 1/2$ hour)

Provides a homogeneous pool of resources over heterogeneous grid environments.

Crab-based user analysis efficiency benefits significantly from late binding methods

Job monitoring tools provide CMS users detailed information for their jobs:

- provides tools for users to access jobs interactively for debugging.

Interface to the next generation CE has been created using Condor client

Recent successful results using CREAM CE looks promising.

Appendix- Condor GAHP and CREAM CE

Condor-GAHP (Translates the states and maps to CREAM IDs)

CREAM Job States

GM_SUBMIT REGISTERED

GM PENDING PENDING

GM PENDING IDLE

GM RUNNING RUNNING

GM RUNNING REALLY-RUNNING

GM HELD HELD

GM CANCEL CANCELLED

GM DONE DONE-OK

GM FAILED DONE-FAILED

GM ABORTED ABORTED

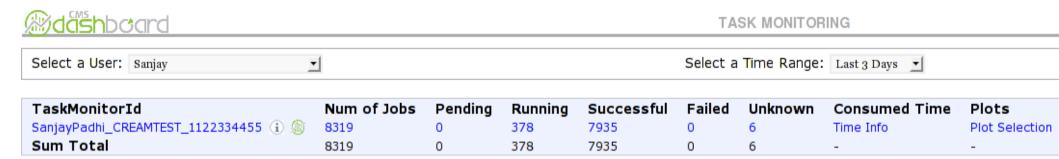
GM ERROR UNKNOWN

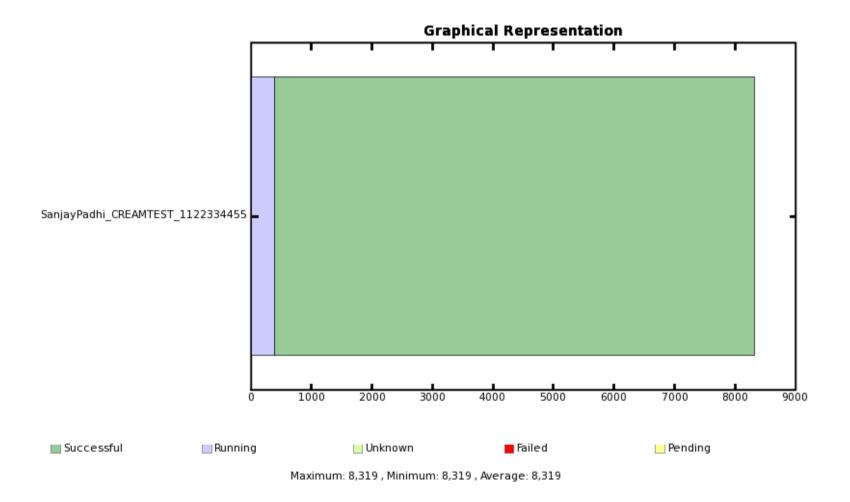
GAHP[10522] -> '32655' 'NULL' '1' 'CREAM500503337' 'DONE-OK' '0' 'NULL' GAHP[10522] -> '32656' 'NULL' '1' 'CREAM650881586' 'IDLE' 'NULL' 'NULL' GAHP[10522] -> '32657' 'NULL' '1' 'CREAM323014883' 'DONE-OK' '0' 'NULL' GAHP[10522] -> '32658' 'NULL' '1' 'CREAM897757278' 'DONE-OK' '0' 'NULL' GAHP[10522] -> '32659' 'NULL' '1' 'CREAM901297268' 'DONE-OK' '0' 'NULL' GAHP[10522] -> '32660' 'NULL' '1' 'CREAM101916692' 'DONE-OK' '0' 'NULL' GAHP[10522] -> '32662' 'NULL' '1' 'CREAM854095215' 'IDLE' 'NULL' 'NULL' GAHP[10522] -> '32661' 'NULL' '1' 'CREAM108996394' 'IDLE' 'NULL' 'NULL' GAHP[10522] -> '32663' 'NULL' '1' 'CREAM327701067' 'IDLE' 'NULL' 'NULL' GAHP[10522] -> '32664' 'NULL' '1' 'CREAM156148436' 'IDLE' 'NULL' 'NULL'

GAHP[10522] -> '32665' 'NULL' '1' 'CREAM758758977' 'IDLE' 'NULL' 'NULL' GAHP[10522] -> '32666' 'NULL' '1' 'CREAM041113197' 'IDLE' 'NULL' 'NULL'

Jo bRegister Registered Jo bStatt Job submission unsuccesfull Pending. LRMS assigns LRMS ID JobSuspend Held Idle JobReame LRMS executes JW JobSuspend Held Ruhhing LRMS executes the job JobSuspend 🚅 Really Running Held Job Resume Job term instes Job terminates JobCancel. successful by Abouted. Cancelled Done-OK Done-Failed

Appendix - Condor tests with CREAM CE (Successful Jobs)





March 24th 2009, Sanjay Padhi