OPTIMIZATION OF HEP ANALYSIS ACTIVITIES USING A TIER2 INFRASTRUCTURE

G.BAGLIESI[A], T.BOCCALI[A], S.COSCETTI[B], E.MAZZONI[A], S.SARKAR[C], S.TANEJA[D]

A: INFN SEZIONE DI PISA, PISA, ITALY, B: SCUOLA NORMALE SUPERIORE, PISA, ITALY, C: SAHA INSTITUTE, KOLKOTA, INDIA, D: S.TANEJA, DIPARTIMENTO DI INFORMATICA, PISA, ITALY

INTRODUCTION

The GRID Data Center at INFN Pisa hosts a big Tier2 for the CMS experiment, together with local usage from other HEP related/not related activities

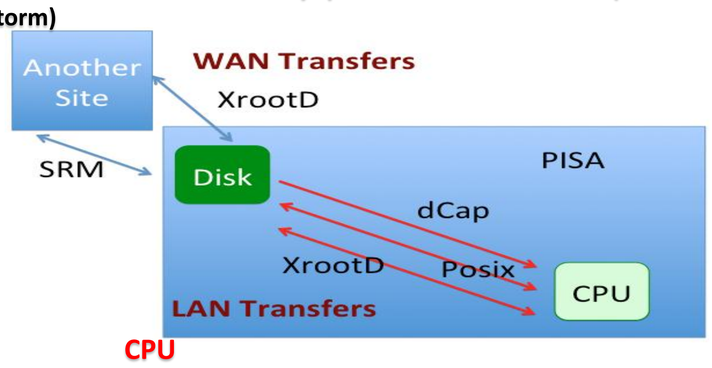
The Tier2 has to fulfill on one side a series of activities mandatory in the Memorandum of Understanding of CMS, on the other side support as much as possible Italian Physicists doing research

Typical activities
- MC prod (MoU): ~1000 cores, 50 TB temporary space
- Analysis via GRID: ~1000 cores, 800 TB
- Local analysis: ~50 users LSF ~ 500 cores, ~ 100 TB local space
- Interactive work: ~50 users, few cores each (mostly multi threaded applications)

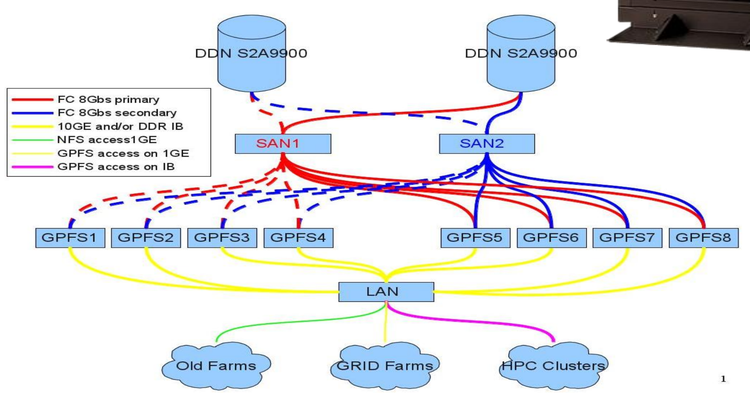
There are three ingredients to a successful large scale data center: storage, CPUs, infrastructure

Storage

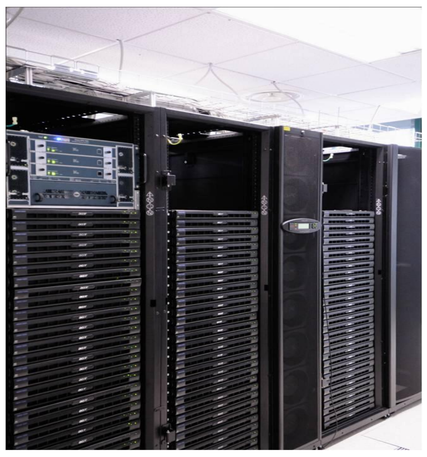
A center used for diverse activities must offer a solid storage infrastructure, accessible through all the protocols needed by the use cases
- SRM for WAN organized transfers (we use Storm, an italian SRM implementation, and dCache for legacy files)
- At least one protocol for locally running GRID jobs (we offer dCap, POSIX, XrootD)
- At least one protocol for WAN direct (streaming) access (we offer XrootD both from dCache and Storm)
- Interactive users in the end always prefer POSIX direct access (we offer that on Storm)

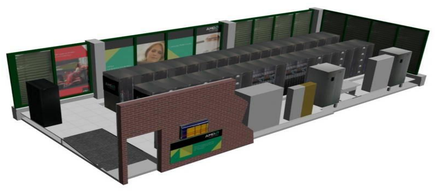


- 2 DataDirect DDN9900 systems, with 300 disks each (>1 PB raw storage)
- Total of 4x8 8 Gbit/s FC connections, served to GPFS NSD via a FC Switch
- 8 GPFS NSD with MultiPath enabled, 10 Gbits connectivity
- GPFS version 3.4.0, serving more than 1 PB
- DDN EF3015 storage dedicated to Metadata handling
- DataDirect 12k on arrival
- Throughput from storage measured in excess of 32 Gbit/s



The GRID data center in Pisa (~5000 cores) is engineered to make sure that the biggest number of tasks is running at any time, which means
- No jobs are guaranteed to run instantaneously
- Every user/group of users can in principle saturate the whole farm
- We are happy to host GRID jobs from users from experiments not present in Pisa
- No restriction is imposed to jobs when the farm is not full
- In case of multiple queued jobs, a fairshare is implemented via LSF which matches the share of resources
- Only case in which some resources are left unused is a small part of the cluster (~500 cores) which can be preempted by parallel jobs



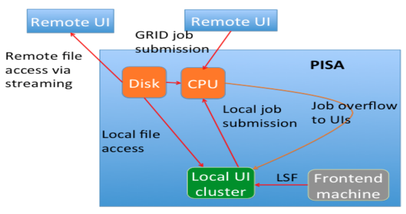


Force10 E1200

APC InRow

Users need specific machines to carry on their research work
- Possibility to submit to the GRID
- Possibility to use the local farm
- Possibility to run interactively a process on a single/multi cores with guarantees on RAM availability

We decided to have a number of machines available, using the batch system as a load balancer.
The batch system directs you to a machine where you have the # of cores and memory you asked for, and makes sure no other task is interfering.
When the interactive load is low, the same machines can be used as standard GRID processing nodes



Request 4 interactive cores in the UI cluster

```
bsub -Is -n 4
  -R "span[hosts=1]"
  -q fai /bin/bash -l
```

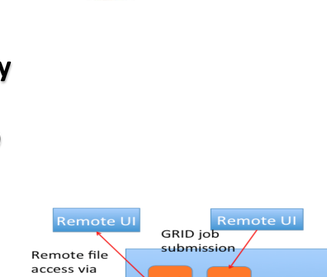
Infrastructure

The Pisa Data Center largely exceeds the typical dimension of a Department's computing center. As such, it is equipped with enterprise level infrastructure
- Rack space: the computing room can host up to 34 full height standard racks.
- Networking: we use a flat switching matrix, implemented via a Force10 E1200 fabric. It allows for a point-to-point 1Gbit/s (or 10Gbit/s) connection between all the machines. We use 1Gbit/s on all the Computing nodes, and 10Gbit/s on all the storage nodes.
- Electrical distribution: we are served with a 300 KVA connection. All the services, and part of the Computing nodes, are served by UPS.
- Cooling: the computing room is served by 12 APC InRow air sources, cooled by 3 Emerson Chillers on the roof (300 kW refrigerating capacity)

We had to develop a number of Monitoring tools to make sure the site is operational. These complement the standard CMS and WLCG existing tools.

Jobs completed during the last year

VO/Group	Total Jobs	Succ Jobs	Succ Rate(%)	Walltime (sec)	CPU Time (sec)	CPU Eff(%)	Walltime Share(%)	Avg Wait (sec)
cms	1748850	1701717	97.30	35150328316	30696944466	87.33	40.62	5531
theophys	104772	85958	82.04	11491148942	10716358951	93.26	13.28	15734
theodip	119116	108838	91.37	9240664958	8311394603	89.94	10.68	12046
theoinfn	82109	70677	86.08	6250310792	5089695297	81.43	7.22	11116
cmsprt	109676	97102	88.54	5378418064	4960889464	92.24	6.21	8710
compchem	148984	145047	97.36	4033296845	3724179321	92.34	4.66	8401
lhcb	170745	167793	98.27	4021745987	3649911712	90.75	4.65	4376
cdf	51003	50020	98.07	3266905235	2516073323	77.02	3.77	1131
glast	43050	39363	91.44	1690453135	1585595668	93.80	1.95	20886
thpi12	50930	49475	97.14	1554803800	1409094314	90.63	1.80	9369
thpi11	72596	71199	98.08	1124985450	1052938178	93.60	1.30	1928
biomed	110580	107308	97.04	709152367	489742590	69.06	0.82	2133
atlas	15286	14000	91.59	513245498	396208565	77.20	0.59	41797
thto61	6511	5289	81.23	407803387	398918891	97.82	0.47	4910
gridit	33607	32505	96.72	228735477	206572232	90.31	0.26	7982



JobFlow: (Submitted | Dispatched | Completed) Last Hour

CPU Slots

Maximum Seen	Current Max	Available	Occupied	Free
5477	4889	4885	4414	471

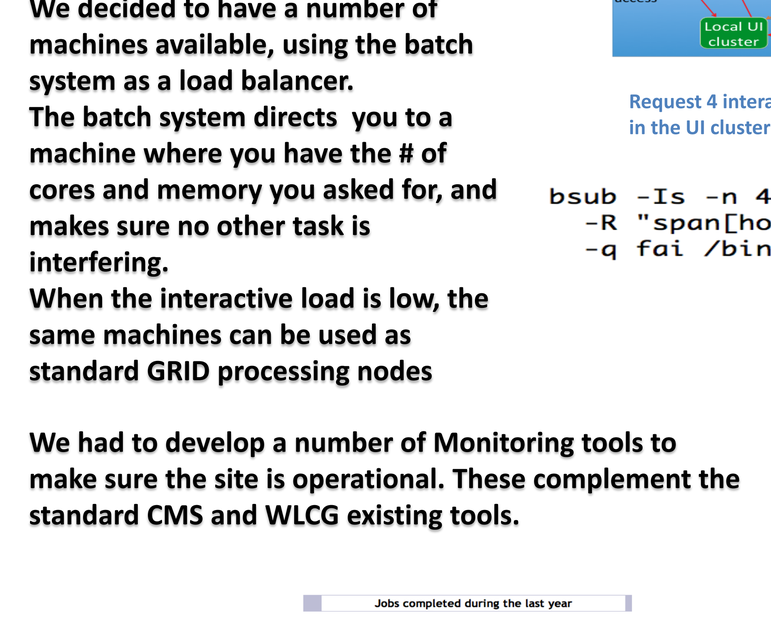
Jobs

Jobs	Running	Pending	Held	CPU Eff(%)	Jobs(Eff<10%)	JobFlow
4291	3495	796	0	51.45	53	151 | 156 | 232

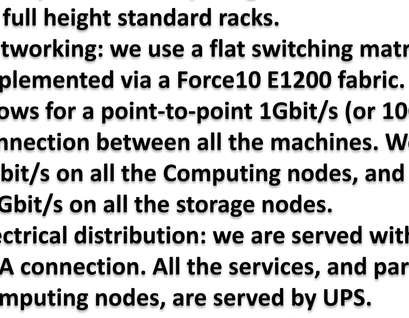
VOs

VO/Group	Jobs	Running	Pending	Held	CPU Eff(%)	Jobs(Eff<10%)	Walltime Share(%)	JobFlow
cms	2554	2110	444	0	42.53	49	81.92	142 | 4 | 158
theophys	726	726	0	0	95.46	1	11.22	0 | 0 | 33
theodip	633	336	297	0	83.01	1	0.65	0 | 149 | 1
cdf	149	144	5	0	67.02	1	2.32	1 | 1 | 0
theoinfn	99	99	0	0	98.34	0	2.14	0 | 0 | 1

LSFMon



Infrastructure status



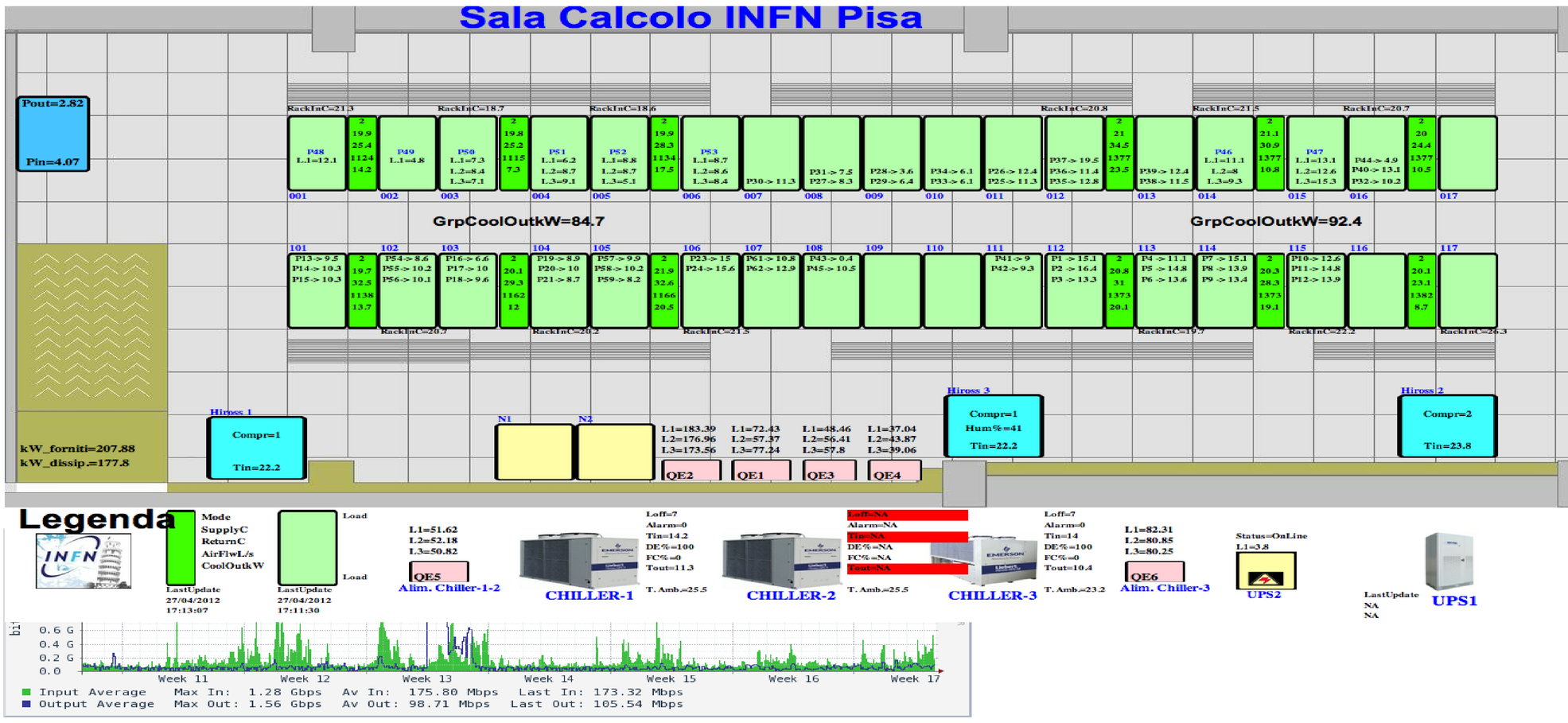
Netvibes widgets to monitor dCache, LSF

Partially developed under PRIN 2008MHENNA MIUR Project

Optimization of HEP Analysis activities using a Tier2 Infrastructure

authors

PRIN statement



T2_IT_Pisa

Site Readiness Status:	R	R	R	R	R	R	R	R	R	W	R	R	R	R	R							
Daily Metric:	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	E	O	O	O	O	O	
Maintenance (Topology):	Up	Up	Up	Up	Up	Up	Up	UD	UD	UD	UD	Up	Up	UD	Up	Up	Up	UD	Up	Up	Up	
Job Robot:	100%	98%	89%	99%	94%	94%	94%	100%	97%	92%	97%	91%	100%	100%	99%	96%	93%	94%	99%	100%	96%	
SAM Availability:	100%	100%	100%	100%	92%	100%	100%	100%	100%	100%	100%	100%	100%	91%	96%	100%	96%	100%	100%	96%	n/a	
Good T2 links from T1s:	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8	1/5	8/8	8/8	7/8	8/8	n/a	
Good T2 links to T1s:	6/7	7/7	7/7	7/7	7/7	7/7	6/7	7/7	7/7	7/7	7/7	7/7	7/7	7/7	6/7	6/7	6/7	6/7	7/8	7/8	7/7	
Active T2 links from T1s:	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	
Active T2 links to T1s:	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	01	02
	Feb																				Mar	

Report made on 2012-03-02 03:30:02 (UTC)

JobFlow: (Submitted|Dispatched|Completed) Last Hour

CPU Slots

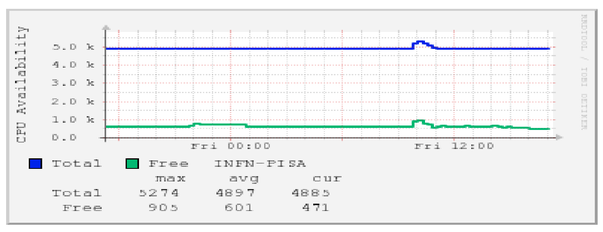
Maximum Seen	Current Max	Available	Occupied	Free
5477	4889	4885	4414	471

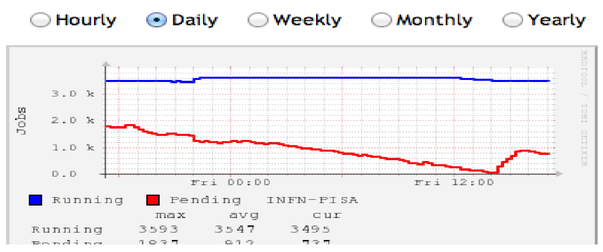
Jobs

Jobs	Running	Pending	Held	CPU Eff(%)	Jobs(Eff<10%)	JobFlow
4291	3495	796	0	51.45	53	151|156|232

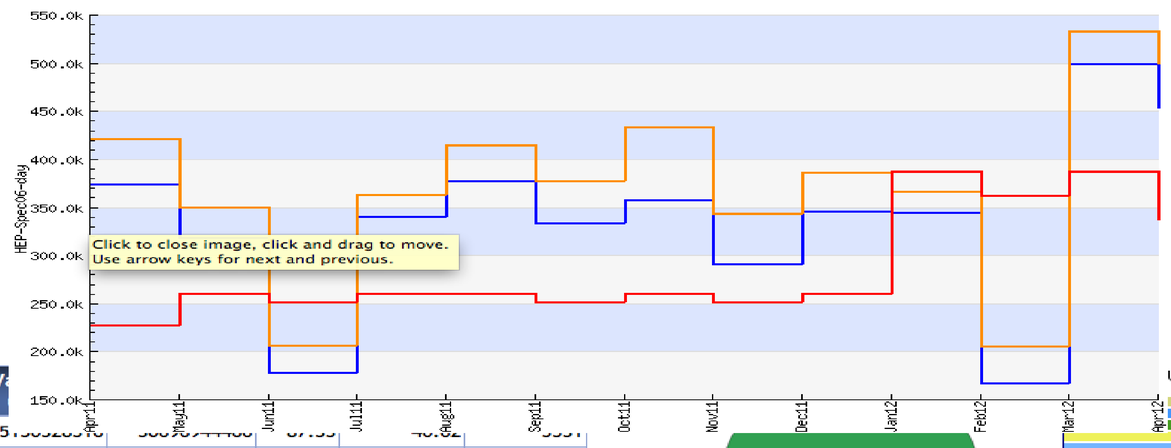
VOs

VO/Group	Jobs	Running	Pending	Held	CPU Eff(%)	Jobs(Eff<10%)	Walltime Share(%)	JobFlow
cms	2554	2110	444	0	42.53	49	81.92	142|4|158
theophys	726	726	0	0	95.46	1	11.22	0|0|33
theodip	633	336	297	0	83.01	1	0.65	0|149|1
cdf	149	144	5	0	67.02	1	2.32	1|1|0
theoinfn	99	99	0	0	98.34	0	2.14	0|0|1

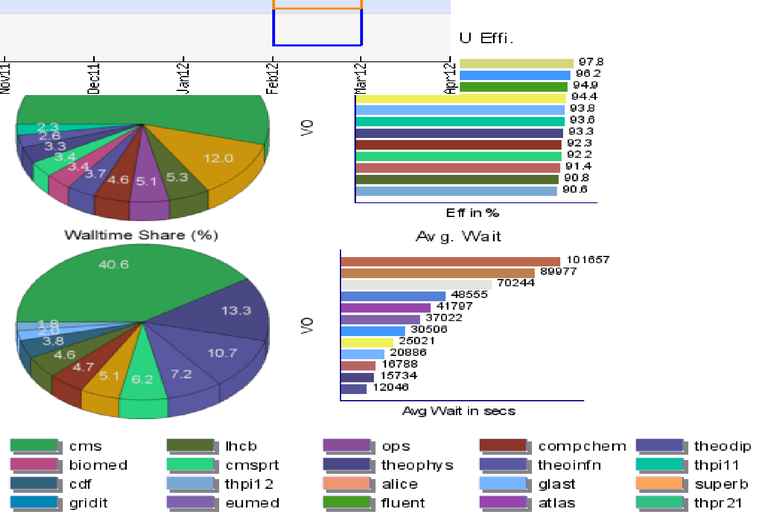




HEP-Spec06-day LHC CPT/WCT per month



VO/Group	Total Jobs	Succ Jobs	Succ Rate(%)	W				
cms	1748850	1701717	97.30	35				
theophys	104772	85958	82.04	11491148942	10716358951	93.26	13.28	15734
theodip	119116	108838	91.37	9240664958	8311394603	89.94	10.68	12046
theoinfn	82109	70677	86.08	6250310792	5089695297	81.43	7.22	11116
cmsprt	109676	97102	88.54	5378418064	4960889464	92.24	6.21	8710
compchem	148984	145047	97.36	4033296845	3724179321	92.34	4.66	8401
lhcb	170745	167793	98.27	4021745987	3649911712	90.75	4.65	4376
cdf	51003	50020	98.07	3266905235	2516073323	77.02	3.77	1131
glast	43050	39363	91.44	1690453135	1585595668	93.80	1.95	20886
thpi12	50930	49475	97.14	1554803800	1409094314	90.63	1.80	9369
thpi11	72596	71199	98.08	1124985450	1052938178	93.60	1.30	1928
biomed	110580	107308	97.04	709152367	489742590	69.06	0.82	2133
atlas	15286	14000	91.59	513245498	396208565	77.20	0.59	41797
thto61	6511	5289	81.23	407803387	398918891	97.82	0.47	4910
gridit	33607	32505	96.72	228735477	206572232	90.31	0.26	7982



Remote UI

Remote UI

GRID job submission

Remote file access via streaming

PISA

Disk

CPU

Local job submission

Job overflow to UIs

Local file access

Local UI cluster

LSF

Frontend machine

Another Site

WAN Transfers

XrootD

SRM

PISA

Disk

dCap

XrootD

Posix

CPU

LAN Transfers