Report on int.eu.grid MPI Challenge

Pablo Orviz Fernández[1]

[1] Instituto de Física de Cantabria (CSIC-UC)

Second MPI Workshop - Bologna, March 2008

Outline

Objetives of the Challenge

Question

Which are the purposes of this Challenge?

Well...

- Reliability of int.eu.grid infrastructure
- Test *mpi-start* and *Crossbroker* int.eu.grid middleware components
- Behaviour of MPI parallel applications in the grid
- Raise awareness of MPI viability for user needs
- Force cluster limits

Challenge Approach

About the Challenge...

- Celebrated on March 3rd 2008
- Representation of every site belonging to int.eu.grid project
- Workload distributed to cover the whole infrastructure
- Challenge participant reported incidents via form

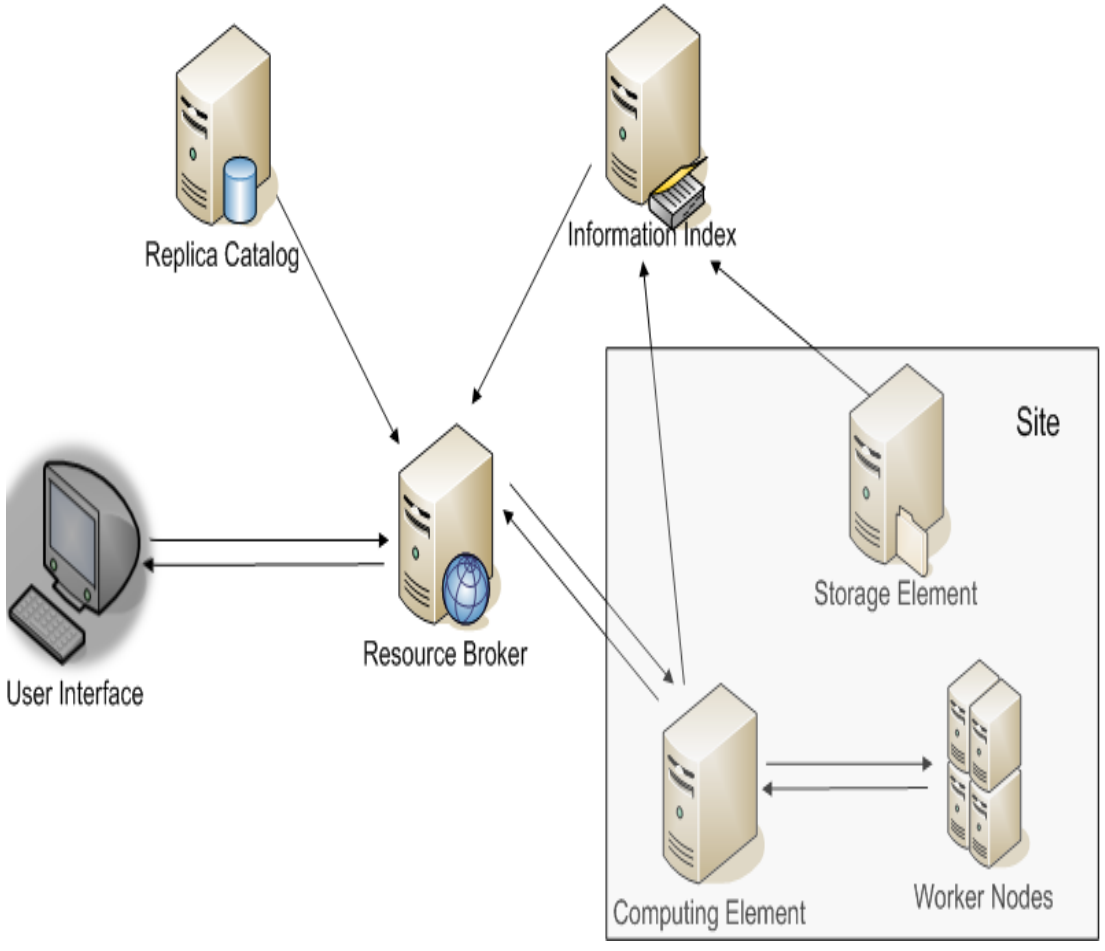
What conditions must be fulfilled?

- OpenMPI based jobs
- 24-hour intensive
- Several processor request
  - Varied from 4 to 50 CPUs
- Tested with different VOs

I2G TestBed

Which resources have been used?

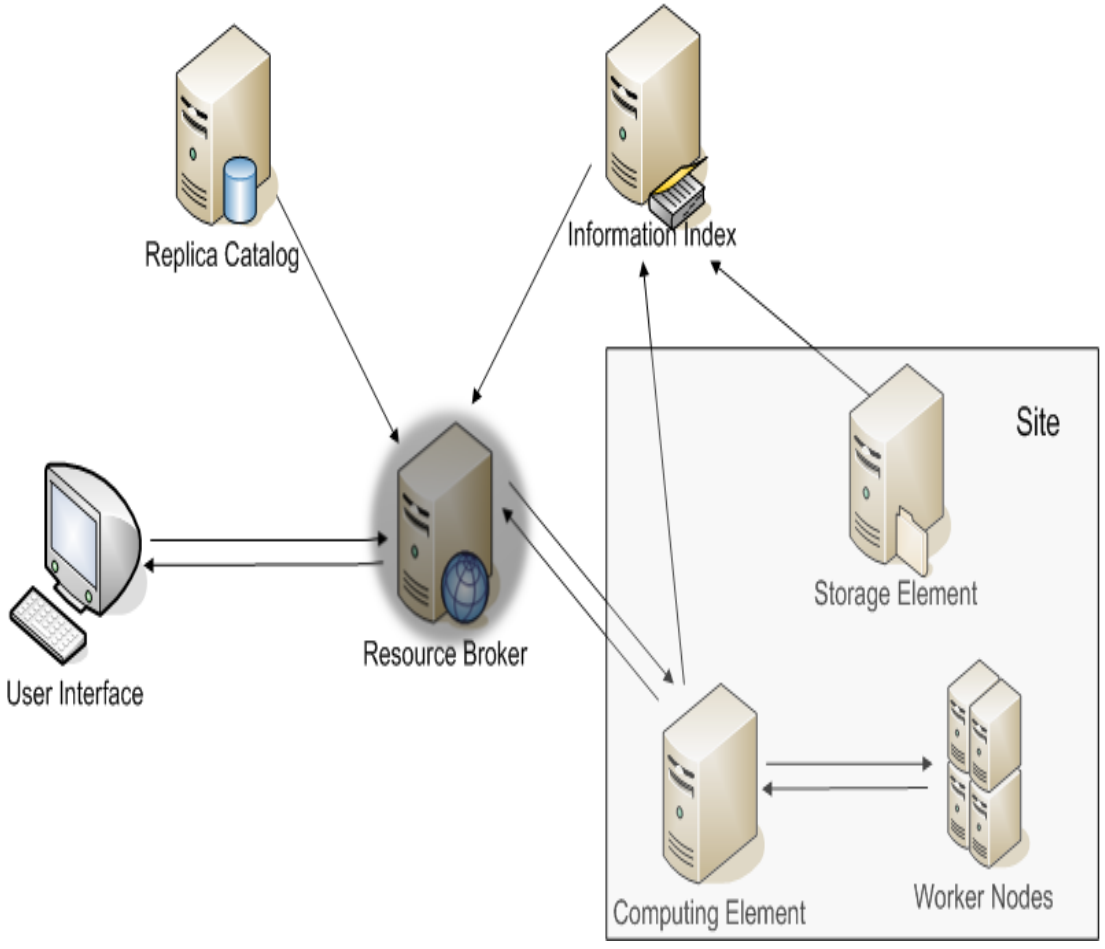
I2G TestBed



Jobs submitted both from

gLite-based UI
Migrating Desktop

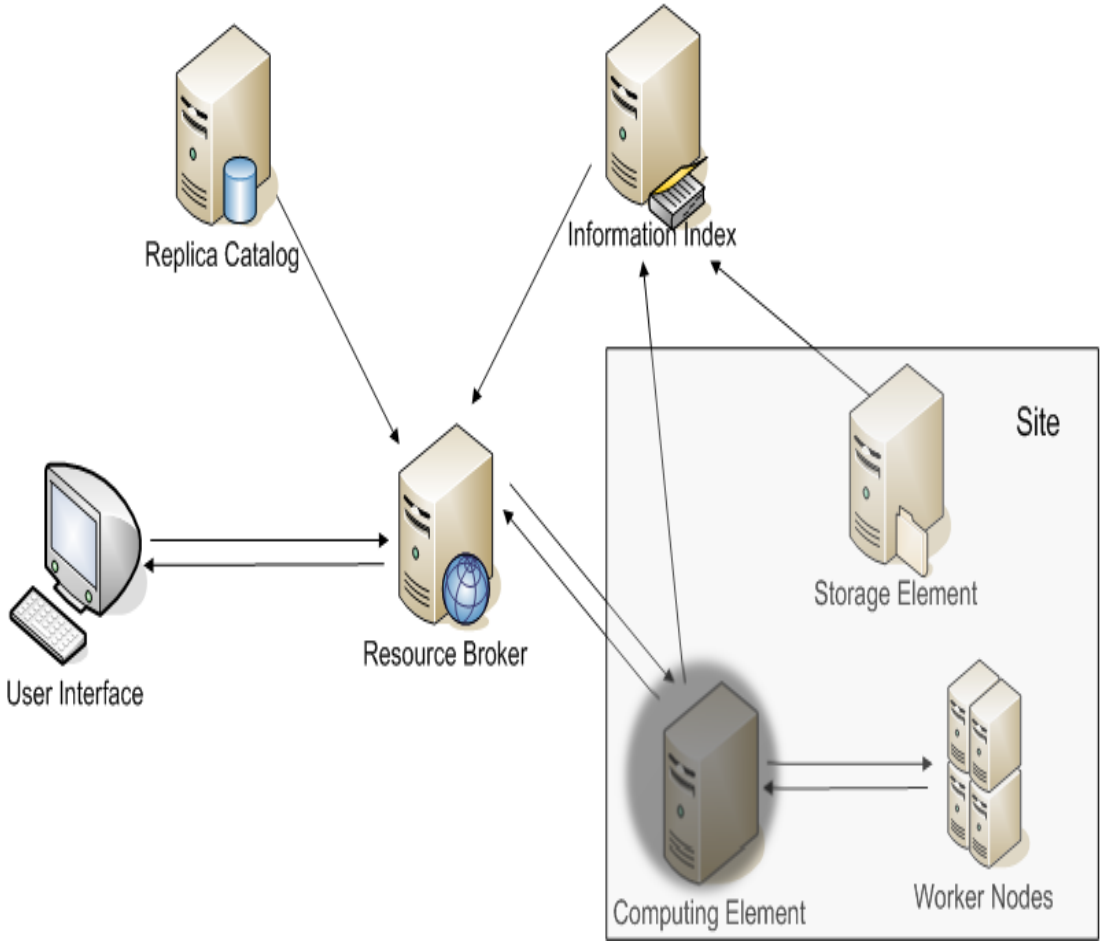
I2G TestBed



RBs
i2g-rb01.lip.pt
i2g-rb02.lip.pt
i2grb01.ifca.es

Infrastructure specification

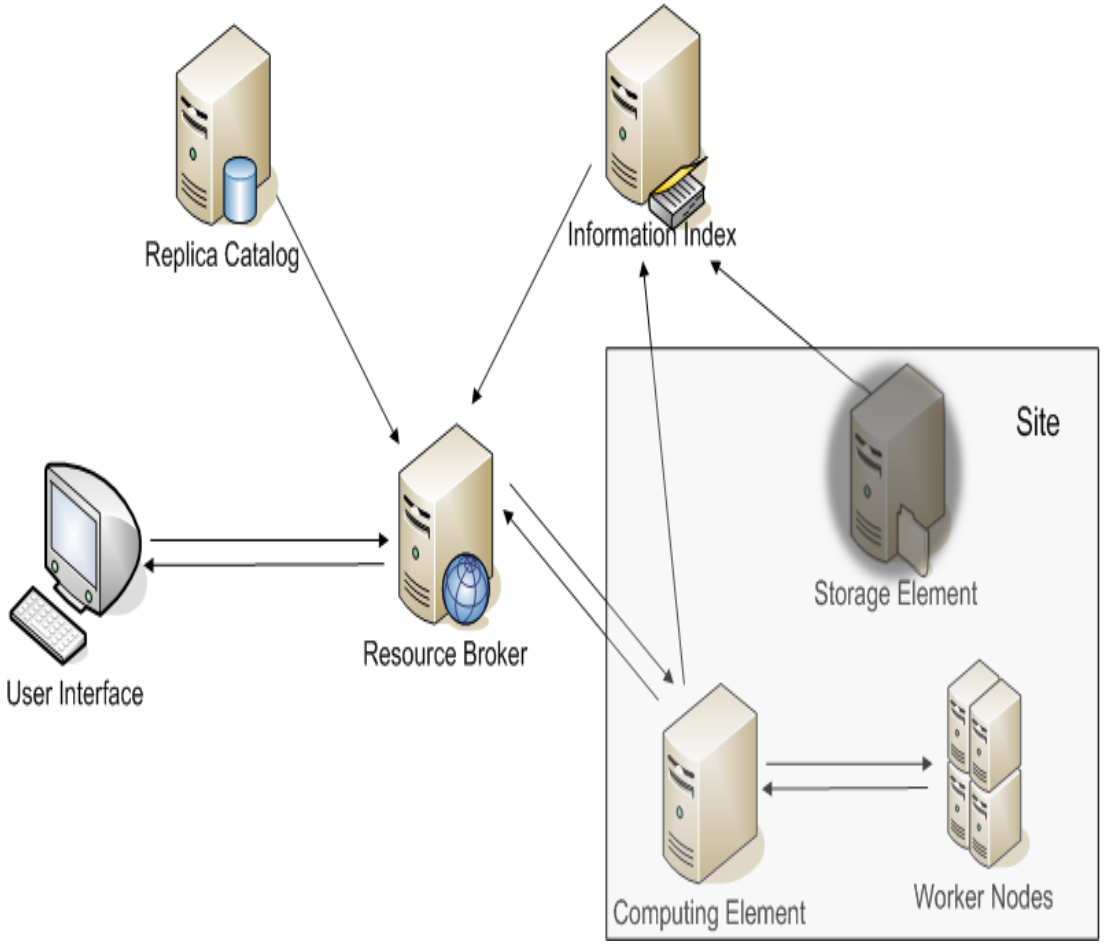
I2G TestBed



CPUs	CEs
28	ce-ieg.bifi.unizar.es
17	ce.i2g.cesga.es
20	ce.i2g.cyf-kr.edu.pl
92	iwrce2.fzk.de
32	i2ce.polgrid.pl
288	i2gce01.ifca.es
32	i2gce.ui.savba.sk
51	i2g-ce01.lip.pt
80	sequoia-2.man.poznan.pl

Infrastructure specification

I2G TestBed



Avail. Space	Used Space	SEs
101444.6	8807.6	se-ieg.bifi.unizar.es
3263.1	7258.5	se.i2g.cesga.es
12390000	4158000	dpm.cyf-kr.edu.pl
1840000	153.8	iwrse2.fzk.de
1368604.3	296.3	i2se.polgrid.pl
1117479	14099	i2gse01.ifca.es
491980	78.1	i2gse.ui.savba.sk
500278.3	1071679.7	dcache01.lip.pt
72587.5	428	i2g-se01.lip.pt
3150000	39.6	se1.reef.man.poznan.pl

Global results

Some statistics...

Global int.eu.grid MPI Challenge results

Overall job statistics

Total of jobs submitted during the int.eu.grid MPI Challenge.

Submitted	Failed	Success	Efficiency
330	104	226	68.49 %

Job features

- Requesting different CPUs (4, 10, 20 and 50)
- With several duration (15 minutes, 1 hour, 4 hours)

A more detailed approach...

Jobs per VO

Jobs were run with different VOs.

VOs	Submitted	Failed	Success	Efficiency
ienvmod	40	10	30	75 %
ifusion	4	4	0	0 %
imain	120	31	89	74.17 %
iplanck	166	59	107	64.46 %

Comment on results

- VO sucess depends on queue priority

A more detailed approach...

Jobs per CE

Job submission was planned to cover int.eu.grid infrastructure.

Site	CEs	CPUs	Submitted	Failed	Success	Efficiency
BIFI	ce-ieg.bifi.unizar.es	28	34	12	22	64.71 %
CESGA	ce.i2g.cesga.es	17	24	9	15	62.5 %
CYFRONET	ce.i2g.cyf-kr.edu.pl	20	69	28	41	59.42 %
FZK	iwrce2.fzk.de	92	16	6	10	62.5 %
POLGRID	i2ce.polgrid.pl	32	14	1	13	92.86 %
IFCA	i2gce01.ifca.es	288	110	33	77	70 %
LIP	i2g-ce01.lip.pt	51	19	3	16	84.21 %
IISAS	i2gce.ui.savba.sk	32	42	10	32	76.19 %
PSNC	sequoia-2.man.poznan.pl	80	2	2	0	0 %

Comment on results

- Better results on sites with lower workload
- Some sites do not publish their real CPU resources

A more detailed approach...

Jobs per CPU request

Job asked for different CPU number

Number of CPUs	Submitted	Failed	Success	Efficiency
4 CPUs	136	28	108	79.41 %
10 CPUs	79	24	55	69.62 %
20 CPUs	82	34	48	58.54 %
50 CPUs	33	18	15	45.46 %

Comment on results

- Better performance with less CPU request

Troubleshooting

Troubleshooting

Someone can be thinking about low efficiency...

MPI non-related issues

- Standard middleware failures
- YAIM utilization affects MPI configuration files already tunned (like `JobManager`)
  - Needs of a systematic testing
- VO site configuration problems (like `ifusion`)
- Hardware problems (power cuts, bad filesystem mountings, ...)

Someone can be thinking about low efficiency...

MPI related issues

- Scheduler configuration
  - Queue limits
    - Submitting more jobs than queue allows
  - Allocation policies
  - Missconfiguration of MPI `submit-filter` configuration file (happened at IFCA Site)
  - Maui `backfilling` policy does not properly work due to predefined 3-day job lifetime
    - A job requesting 50 CPUs can block other smaller jobs (20 CPUs) although there were 30 free slots
    - We need to tell Maui to ignore this 3-day lifetime at the level of Crossbroker or through a JDL parameter

Someone can be thinking about low efficiency...

MPI related issues (Continuation)

- Proxy certificate
  - We got `Proxy Expired` aborted jobs although it was not expired
  - Seemed to be a Condor issue on the Crossbroker

Something to take into account...

Problem

During the Challenge we noticed a very low CPU time consumed by jobs

Diagnosis

- The accounting of CPU time is handled by *pls – tm* libraries
- These libraries hand the monitoring control to PBS which then does the accounting of resources being used

Solution

- *pls – tm* libraries need to be compiled for OpenMPI and installed in every Worker Node