Interactive European Grid

Accounting and Monitoring of Parallel & Interactive Jos in the GRID

Carlos Fernández Sánchez
et al.
(CESGA)

IBERGRID 08, Oporto, 12-14 May 2008

CESGA

Introduction

- As Grid Infrastructure Deployments involve the interaction of different geographically distributed communities. Need of accounting and monitoring functionalities.
- The key interested parts of interactive and parallel grid monitoring and accounting are discussed along with the new functionalities that have been implemented to monitor and account such infrastructure

Outline

- i2g project
- accounting
  - Parallel jobs
  - Waiting time
- monitoring
  - SAM tests
  - Daily report

int.eu.grid overview

- Interactive and parallel computing with message passing initiative procedures.
- Terabytes scale storage
- Grid and applications access with a graphical interface named Migrating Desktop with advanced visualization capabilities to show up the results of the simulations in real time
- Support for Virtual Organizations aiming to participate in the project

parallel and interactive job

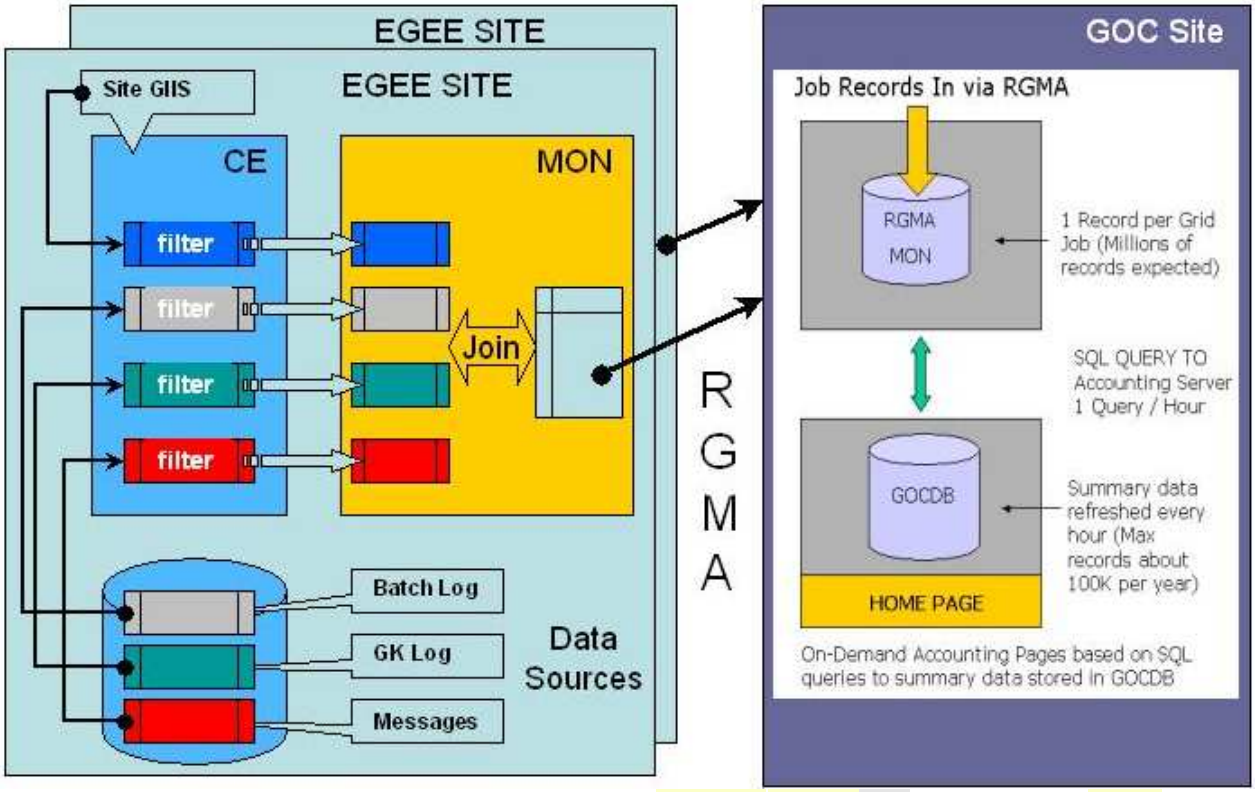
- Parallel based on OpenMPI and PACXMPI for intracluster & intercluster support
- mpi-start supports different schedulers and different MPI implementations (OpenMPI).
- Supports simple file distribution by using scripts to be inserted in the job definition language (JDL) file (hooks)
- Interactive jobs, based on the glogin agent
- Specific design for the project i2glogin which allow interactive parallel jobs.

Accounting

- What is and why is useful
- Based on the CESGA developments for EGEE.
- Some additional develps. had to be done to cover the project new features (parallelism & interactive)
- Its function is to provide quantitative information primarily statistical in nature, about users, sites, and VOs, that is intended to be useful and enrich our understanding of the utilisation of grid resources by the different agents.

APEL: Architecture

Information is gathered from each site job's accounting into a central repository R-GMA (Relational Grid Monitoring Architecture).



APEL: LcgRecords Table

The information is stored using the **schema proposed by the Global Grid Forum's Usage Record Working Group**.

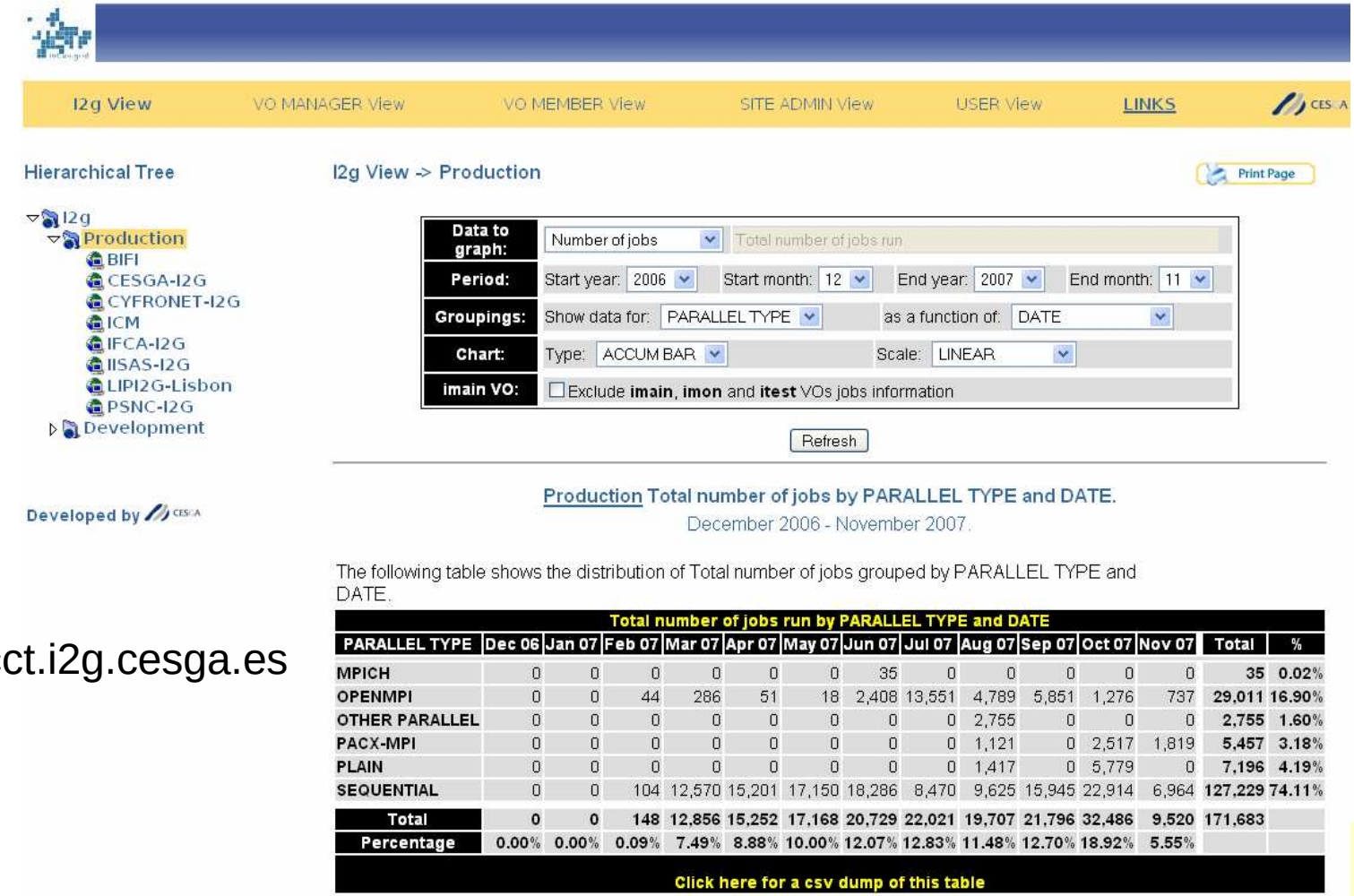
Column Name	Datatype	NOT NULL	AUTO INC	Flags	Default Value
RecordIdentity	VARCHAR(255)	✓		BINARY	
ExecutingSite	VARCHAR(50)			BINARY	NULL
LocalJobID	VARCHAR(50)			BINARY	NULL
LCGJobID	VARCHAR(255)			BINARY	NULL
LocalUserID	VARCHAR(50)			BINARY	NULL
LCGUserID	VARCHAR(255)			BINARY	NULL
LCGUserVO	VARCHAR(255)			BINARY	NULL
ElapsedTime	VARCHAR(30)			BINARY	NULL
BaseCpuTime	VARCHAR(30)			BINARY	NULL
ElapsedTimeSeconds	INTEGER			UNSIGNED ZEROFI	NULL
BaseCpuTimeSeconds	INTEGER			UNSIGNED ZEROFI	NULL
StartTime	VARCHAR(30)			BINARY	NULL
StopTime	VARCHAR(30)			BINARY	NULL
StartTimeUTC	VARCHAR(30)			BINARY	NULL
StopTimeUTC	VARCHAR(30)			BINARY	NULL
StartTimeEpoch	INTEGER			UNSIGNED ZEROFI	NULL
StopTimeEpoch	INTEGER			UNSIGNED ZEROFI	NULL
ExecutingCE	VARCHAR(50)			BINARY	NULL
MemoryReal	INTEGER			UNSIGNED ZEROFI	NULL
MemoryVirtual	INTEGER			UNSIGNED ZEROFI	NULL
SpecInt2000	INTEGER			UNSIGNED ZEROFI	NULL
SpecFloat2000	INTEGER			UNSIGNED ZEROFI	NULL
EventDate	DATE	✓			'0000-00-00'
EventTime	TIME	✓			'00:00:00'
MeasurementDate	DATE	✓			'0000-00-00'
MeasurementTime	TIME	✓			'00:00:00'

Monthly Summarization

- **Information is updated hourly** from R-GMA servers.
- Data is stored in **local database**.
- And data is **summarized by a monthly basis**.

Column Name	Datatype	NOT NULL	AUTO INC	Flags	Default Value
ExecutingSite	VARCHAR(50)	✓		BINARY	
LCGUserVO	VARCHAR(50)	✓		BINARY	
Njobs	INTEGER			UNSIGNED ZEROFILL	NULL
SumCPU	DECIMAL(10,0)			UNSIGNED ZEROFILL	NULL
NormSumCPU	DECIMAL(10,0)			UNSIGNED ZEROFILL	NULL
SumWCT	DECIMAL(10,0)			UNSIGNED ZEROFILL	NULL
NormSumWCT	DECIMAL(10,0)			UNSIGNED ZEROFILL	NULL
Month	INTEGER	✓		UNSIGNED ZEROFILL	0
Year	INTEGER	✓		UNSIGNED ZEROFILL	0
RecordStart	DATE				NULL
RecordEnd	DATE				NULL

New Parallel & Interactive job Accounting Portal. I



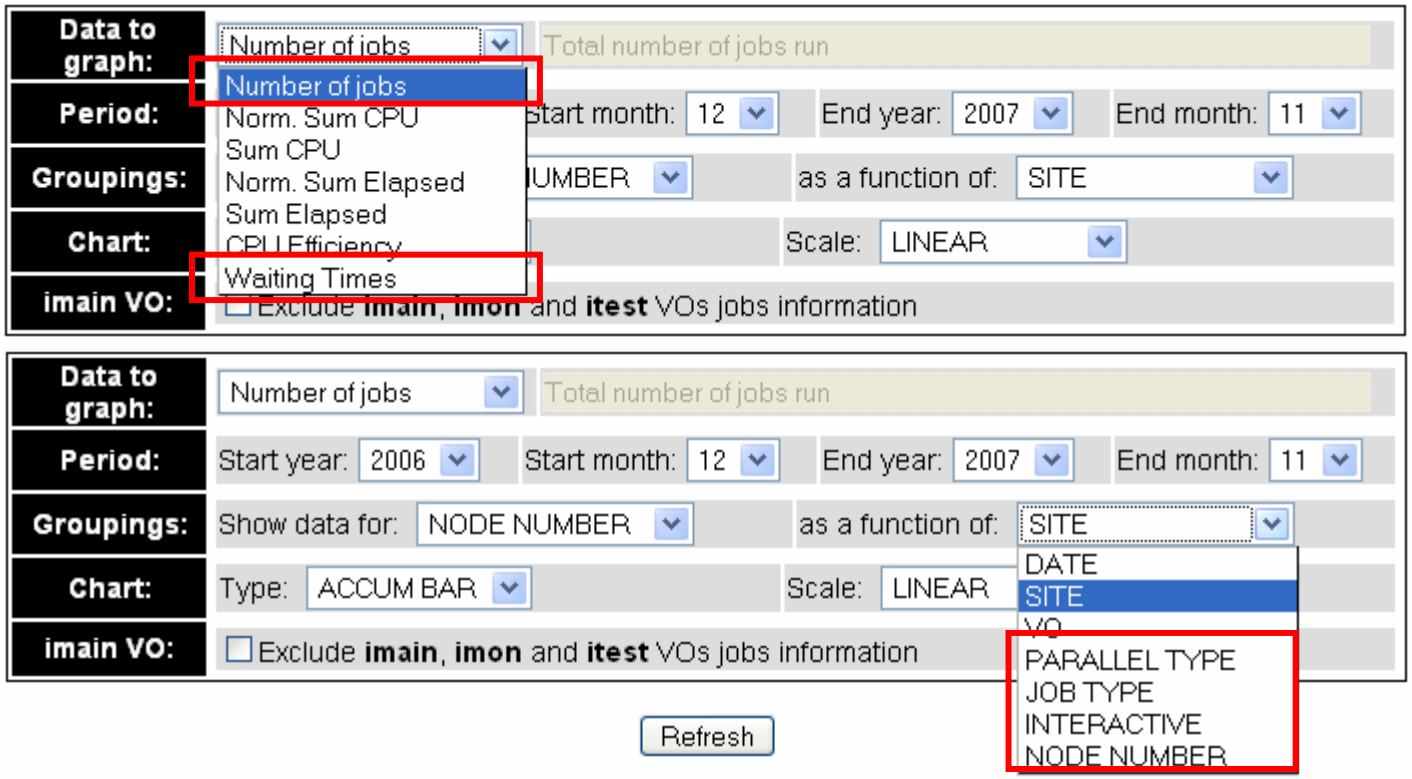
Total number of jobs run by PARALLEL TYPE and DATE

PARALLEL TYPE	Dec 06	Jan 07	Feb 07	Mar 07	Apr 07	May 07	Jun 07	Jul 07	Aug 07	Sep 07	Oct 07	Nov 07	Total	%
MPICH	0	0	0	0	0	0	35	0	0	0	0	0	35	0.02%
OPENMPI	0	0	44	286	51	18	2,408	13,551	4,789	5,851	1,276	737	29,011	16.90%
OTHER PARALLEL	0	0	0	0	0	0	0	0	2,755	0	0	0	2,755	1.60%
PACX-MPI	0	0	0	0	0	0	0	0	1,121	0	2,517	1,819	5,457	3.18%
PLAIN	0	0	0	0	0	0	0	0	1,417	0	5,779	0	7,196	4.19%
SEQUENTIAL	0	0	104	12,570	15,201	17,150	18,286	8,470	9,625	15,945	22,914	6,964	127,229	74.11%
Total	0	0	148	12,856	15,252	17,168	20,729	22,021	19,707	21,796	32,486	9,520	171,683	
Percentage	0.00%	0.00%	0.09%	7.49%	8.88%	10.00%	12.07%	12.83%	11.48%	12.70%	18.92%	5.55%		

Click here for a csv dump of this table

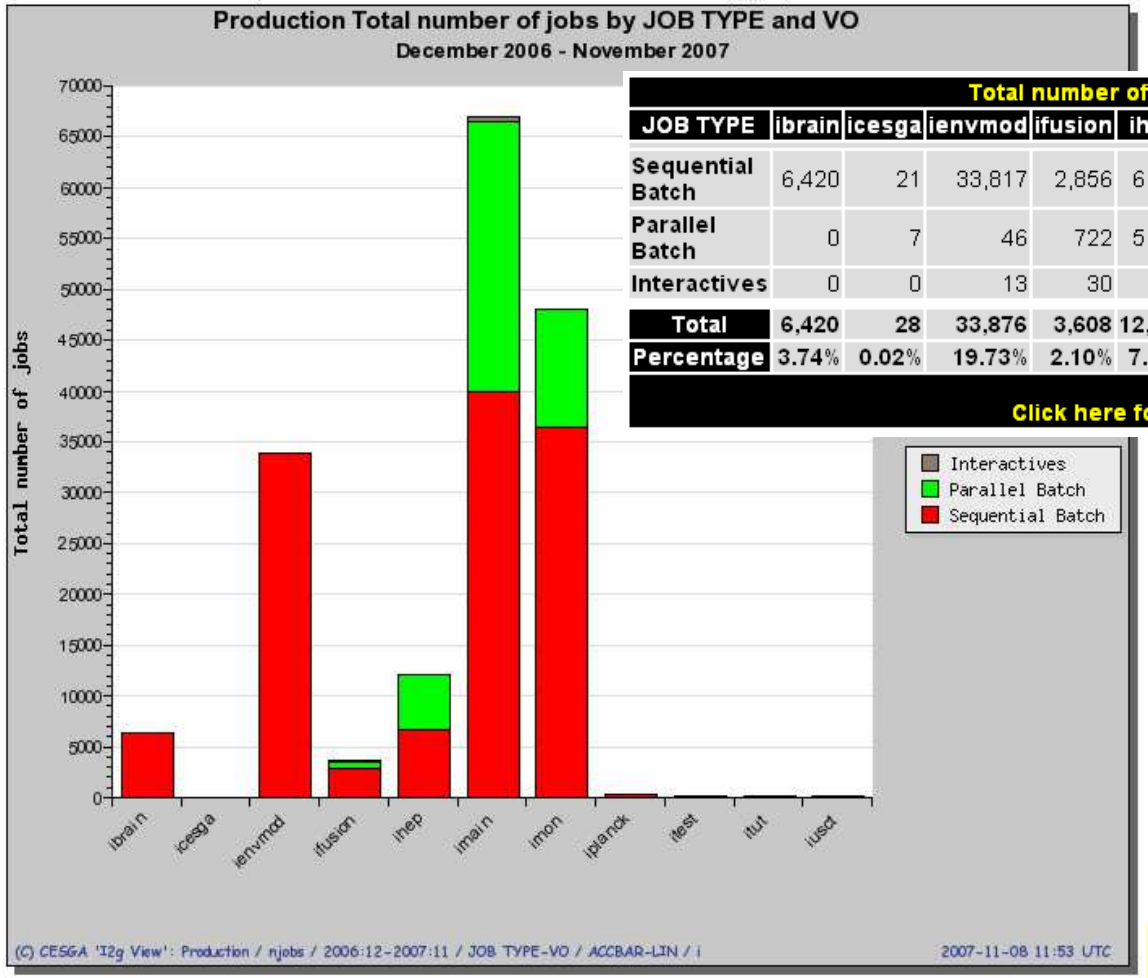
http://acct.i2g.cesga.es

New Parallel & Interactive job Accounting Portal. New user interface



- For "Waiting Times" and "Number of jobs"+(Parallel type, Job Type, Interactive, Node Number) **data is collected from RBs.**
- **All other cases** data comes **from R-GMA servers.**

New Parallel & Interactive job Accounting Portal. II



Total number of jobs run by JOB TYPE and VO

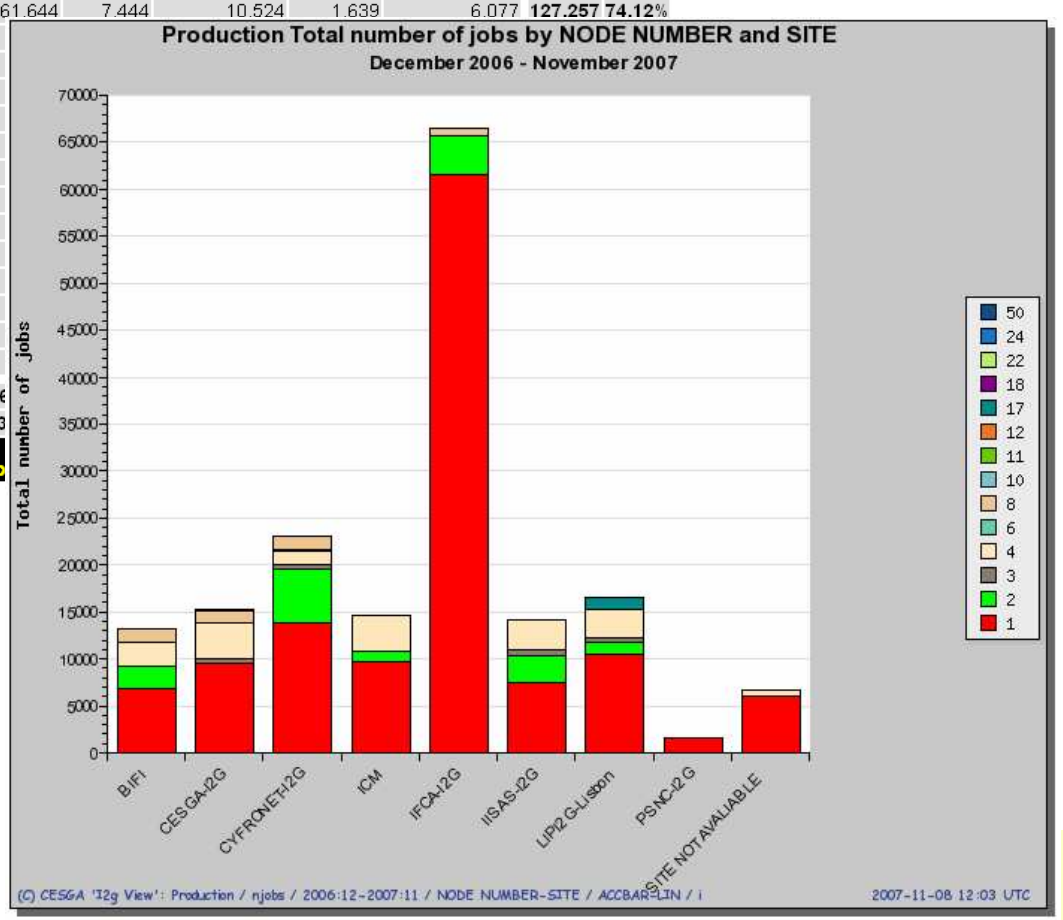
JOB TYPE	ibrain	icesga	ienvmod	ifusion	ihep	imain	imon	iplanck	itest	itut	iusct	Total	%
Sequential Batch	6,420	21	33,817	2,856	6,727	39,964	36,460	342	57	6	120	126,790	73.85%
Parallel Batch	0	7	46	722	5,330	26,579	11,615	6	4	97	33	44,439	25.88%
Interactives	0	0	13	30	0	374	0	0	37	0	0	454	0.26%
Total	6,420	28	33,876	3,608	12,057	66,917	48,075	348	98	103	153	171,683	
Percentage	3.74%	0.02%	19.73%	2.10%	7.02%	38.98%	28.00%	0.20%	0.06%	0.06%	0.09%		

Click here for a csv dump of this table

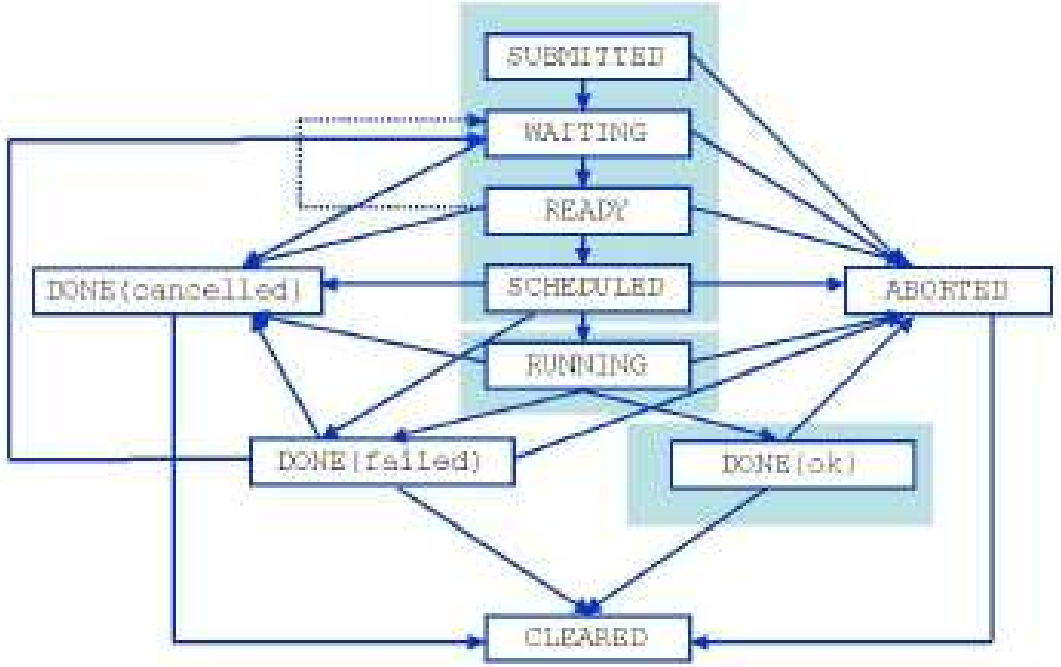
New Parallel & Interactive job Accounting Portal III

Total number of jobs run by NODE NUMBER and SITE											
NODE NUMBER	BIFI	CESGA-I2G	CYFRONET-I2G	ICM	IFCA-I2G	IISAS-I2G	LIPI2G-Lisbon	PSNC-I2G	SITE NOT AVALIABLE	Total	%
1	6,843	9,555	13,876	9,655	61,644	7,444	10,524	1,639	6,077	127,257	74.12%
2	2,461	7	5,753	1,093							
3	0	453	378	0							
4	2,474	3,890	1,547	3,947							
6	0	0	5	0							
8	1,355	1,288	1,472	1							
10	0	0	0	0							
11	31	0	0	0							
12	0	0	0	0							
17	0	0	0	0							
18	28	0	0	0							
22	0	1	0	0							
24	0	0	0	0							
50	0	0	0	0							
Total	13,192	15,194	23,031	14,696	6						
Percentage	7.68%	8.85%	13.41%	8.56%	3						

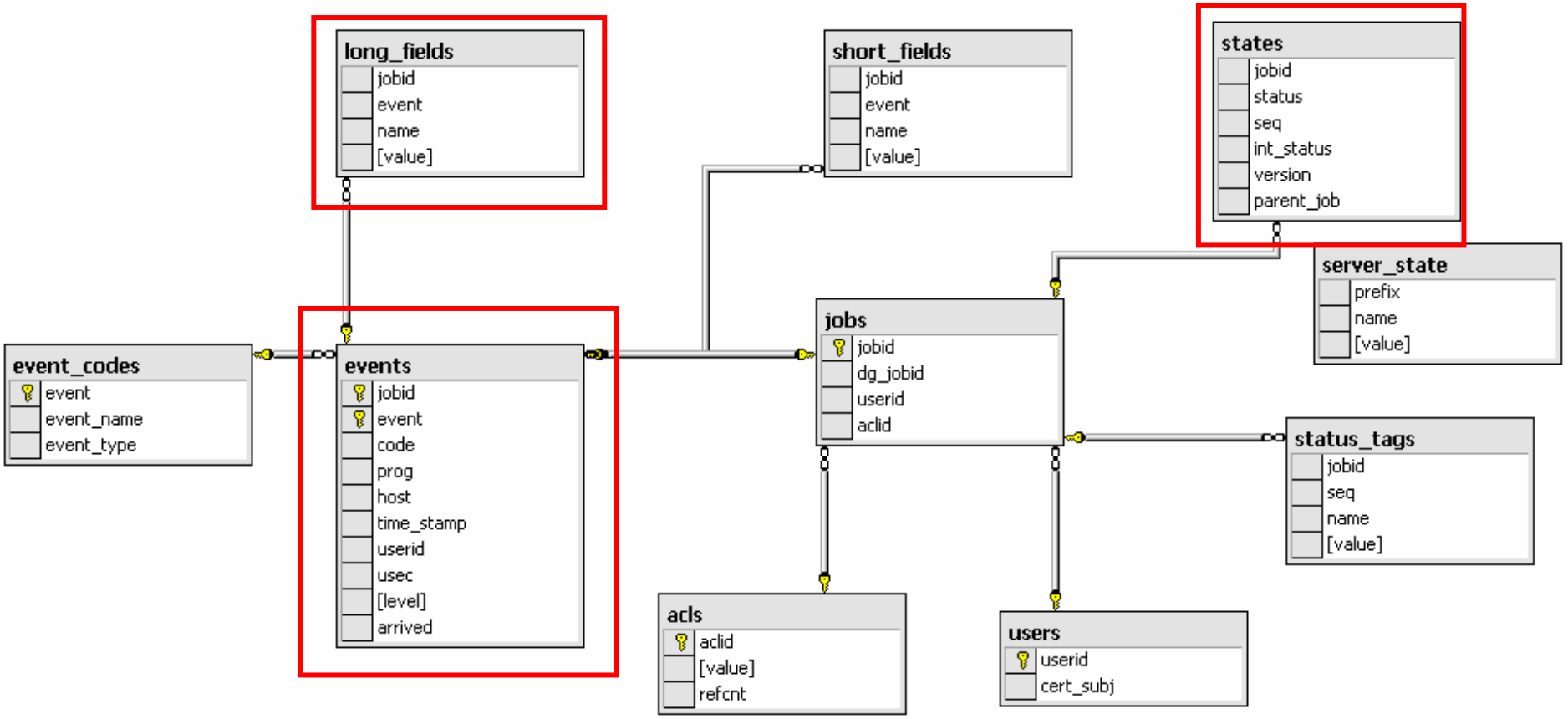
Click here fo



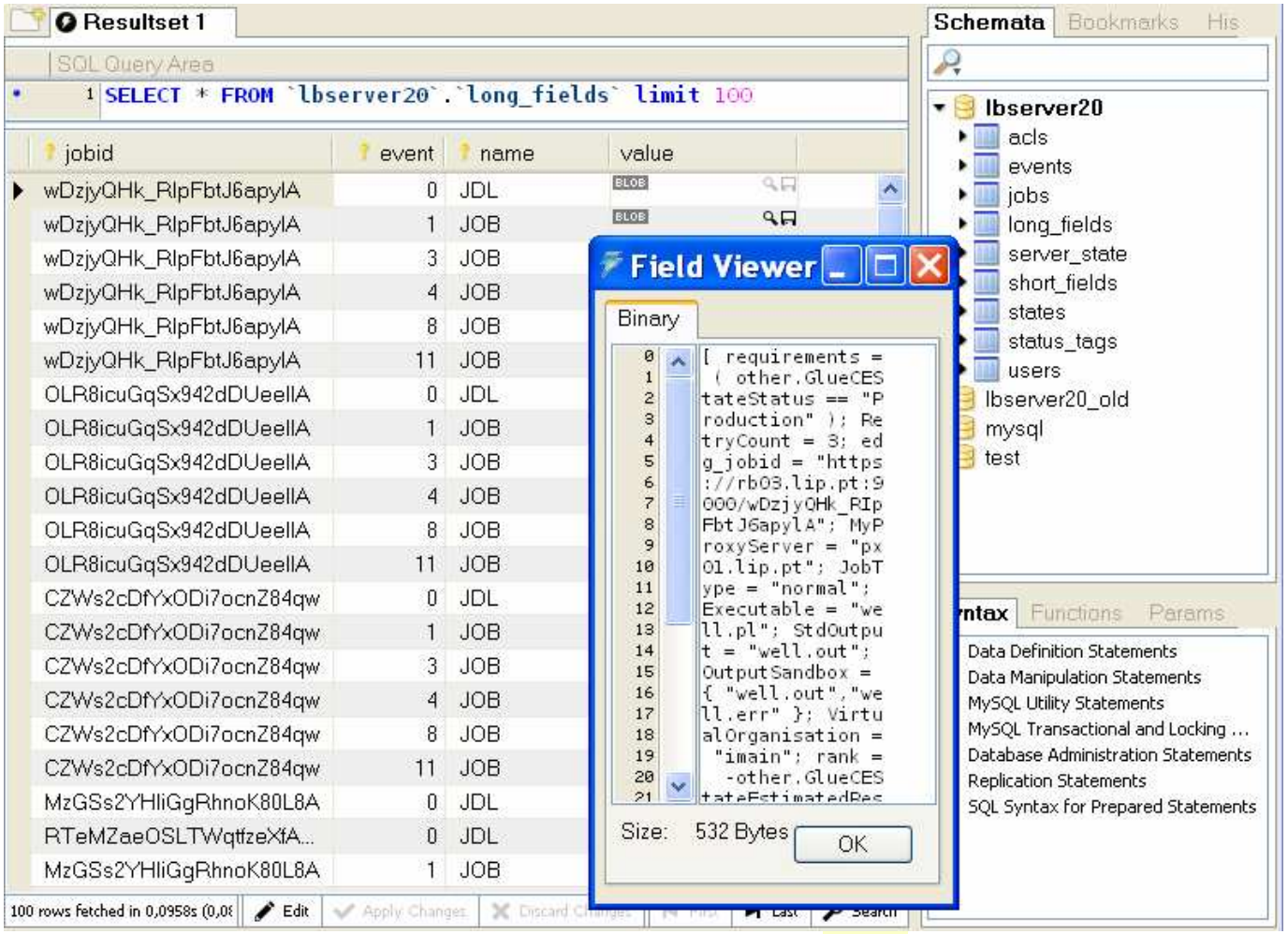
How Data is collected. Job's Life cicle



Resource Broker's Database



Resource Broker "long_fields" table



RB's data gathering process

- **Resource Broker** data query & transformation:
  - Waiting time.
- Job requirements **information extraction:**
  - JobType.
  - Interactive Agent.
  - Parallel Type.
  - NodeNumber.
  - Virtual Organization.
  - Executing Site.
- Insertion in **local database** "JobRequirements" table.
- **Monthly Summarization** in "SumNJobsRB " table.

"JobRequirements" table schema

- Data gathered from **all Resource Brokers**.
- **Job requirements are parsed** and transformed into a suitable database field format.
- Data is stored in **local mySQL database**.

Column Name	Datatype	NOT NULL	AUTO INC	Flags	Default Value
dg_jobid	VARCHAR(255)	✓		☐ BINARY	
VirtualOrganisation	VARCHAR(45)	✓		☐ BINARY	
ParallelType	VARCHAR(45)	✓		☐ BINARY	'SEQUENTIAL'
NodeNumber	INTEGER	✓		☑ UNSIGNED ☐ ZEROFILL	1
Processors	INTEGER	✓		☑ UNSIGNED ☐ ZEROFILL	1
ExecutingSite	VARCHAR(50)			☐ BINARY	NULL
time_stamp	DATETIME	✓			'0000-00-00 00:00:00'
InteractiveAgent	VARCHAR(45)			☐ BINARY	'BATCH'
SubmitingSite	VARCHAR(50)			☐ BINARY	NULL
EventDate	DATE	✓			'0000-00-00'
WaitingTime	INTEGER	✓		☑ UNSIGNED ☐ ZEROFILL	1

Gathering of RB's data

```
SELECT
    L.`jobid`,date (E1.`time_stamp`) AS EventDate,
    E1.`time_stamp` AS time_stamp,
    TIME_TO_SEC(TIMEDIFF(E2.`time_stamp`, E1.`time_stamp`)) AS WaitingTime,
    L.`value`
FROM
        lbserver20.`long_fields` AS L,
        lbserver20.`events`      AS E1
LEFT JOIN
    (SELECT
        E2.jobid, E2.time_stamp FROM lbserver20.`states` s ,
    (SELECT
        E2.jobid, MIN(E2.time_stamp) AS time_stamp
    FROM
        lbserver20.`events` E2 WHERE (E2.code='10') GROUP BY E2.jobid ) AS E2
    WHERE
        (s.jobid=E2.jobid) AND ((s.`status`=6) OR (s.`status`=7)) ) AS E2
ON
    L.jobid = E2.jobid
WHERE
    (E1.event=0) AND (L.event=0) AND (L.jobid= E1.jobid) AND (L.jobid = E2.jobid);
```

JobRequirements Gathered Data

dg_jobid	VirtualOrg...	ParallelType	NodeNumber	ExecutingSite	time_stamp	InteractiveAgent	EventDate	WaitingTime	Su
YmlQE6v6t2crta...	imon	SEQUENTIAL	1	LIPI2G-Lisbon	2007-08-07 12:10:31	BATCH	2007-08-07	289	
YmJHMT1Tfuz4...	imon	SEQUENTIAL	1	FZK-I2G	2007-08-13 07:10:28	BATCH	2007-08-13	364	
YmUnDSDQcA...	imain	OPENMPI	2	LIPI2G-Lisbon	2007-09-12 15:16:49	BATCH	2007-09-12	432	
YmXT5tAqbsP4...	imain	OPENMPI	2	CYFRONET-I2G	2007-10-20 10:16:40	BATCH	2007-10-20	197	
YmYh4RoX4Tg...	imain	SEQUENTIAL	1	CESGA-I2G	2007-11-05 11:30:55	I2GLOGIN	2007-11-05	133	
Ymau8Uw7RBT...	imon	OPENMPI	2	CYFRONET-I2G	2007-08-12 18:11:02	BATCH	2007-08-12	190	
Ymaygw_8dg5l...	imain	OPENMPI	2	CYFRONET-I2G	2007-09-05 01:16:46	BATCH	2007-09-05	171	
YmflSarVzPVC...	imon	OPENMPI	2	IFCA-I2G	2007-07-19 05:11:19	BATCH	2007-07-19	101	
YmitPnd9nVPv4...	imain	SEQUENTIAL	1	LIPI2G-Lisbon	2007-07-24 01:15:57	BATCH	2007-07-24	193	
YmlFDbl_JGRyi...	imon	OPENMPI	2	CYFRONET-I2G	2007-09-28 08:11:08	BATCH	2007-09-28	188	
Ymo2jgF2rcmB...	imon	SEQUENTIAL	1	ICM	2007-08-11 18:10:33	BATCH	2007-08-11	134	
Ymseq6wD5lOK...	imain	PACX-MPI	3	BIFI	2007-11-05 19:17:23	BATCH	2007-11-05	352	
YmwQzIDWRm...	ifusion	SEQUENTIAL	1	SITE NOT AVALIAB...	2007-10-31 13:00:12	BATCH	2007-10-31	193	
YmzBdSGxqb0...	imain	SEQUENTIAL	1	CYFRONET-I2G	2007-08-07 12:16:10	BATCH	2007-08-07	239	
Yn0Q5jnzlt8wAlz...	imon	SEQUENTIAL	1	CESGA-I2G	2007-09-06 23:10:21	BATCH	2007-09-06	175	
Yn236EVHc60L...	imain	SEQUENTIAL	1	LIPI2G-Lisbon	2007-08-29 21:16:01	BATCH	2007-08-29	432	
Yn5gc2MKqhFK...	imain	SEQUENTIAL	1	LIPI2G-Lisbon	2007-07-13 16:16:05	BATCH	2007-07-13	267	
Yn7ClKcnDF6yk...	imon	SEQUENTIAL	1	LIPI2G-Lisbon	2007-07-16 00:10:29	BATCH	2007-07-16	236	
YnFclblciCk2bX...	imain	OPENMPI	2	ICM	2007-08-09 04:17:05	BATCH	2007-08-09	72	
YnHpXRDIFZtQ...	imain	OPENMPI	2	ICM	2007-10-20 18:16:52	BATCH	2007-10-20	75	
YnKCMhGBbn7...	ibrain	SEQUENTIAL	1	CESGA-I2G	2007-11-07 08:31:00	BATCH	2007-11-07	87	
YnOWVA3pSyz...	imain	OPENMPI	2	BIFI	2007-10-01 09:17:00	BATCH	2007-10-01	139	
YnOc5eK7dx4A...	imon	OPENMPI	2	FZK-I2G	2007-08-27 03:11:10	BATCH	2007-08-27	196	
Yn_gYDFNosz...	imain	OPENMPI	2	BIFI	2007-06-27 22:17:10	BATCH	2007-06-27	158	
Yn_gYdBJ_nfc...	imain	OPENMPI	2	FZK-I2G	2007-11-03 04:17:02	BATCH	2007-11-03	240	

SumNJobsRB table

Column Name	Datatype	NOT NULL	AUTO INC	Flags	Default Value
ExecutingSite	VARCHAR(50)	✓		☐ BINARY	
LCGUserVO	VARCHAR(50)	✓		☐ BINARY	
Njobs	INTEGER			☐ UNSIGNED ☐ ZEROFILL	NULL
SumCPU	DECIMAL(10,0)			☐ UNSIGNED ☐ ZEROFILL	NULL
NormSumCPU	DECIMAL(10,0)			☐ UNSIGNED ☐ ZEROFILL	NULL
SumWCT	DECIMAL(10,0)			☐ UNSIGNED ☐ ZEROFILL	NULL
NormSumWCT	DECIMAL(10,0)			☐ UNSIGNED ☐ ZEROFILL	NULL
Month	INTEGER	✓		☐ UNSIGNED ☐ ZEROFILL	0
Year	INTEGER	✓		☐ UNSIGNED ☐ ZEROFILL	0
RecordStart	DATE				NULL
RecordEnd	DATE				NULL
ParallelType	VARCHAR(45)	✓		☐ BINARY	'Sequential'
NodeNumber	INTEGER	✓		☑ UNSIGNED ☐ ZEROFILL	1
Processors	INTEGER	✓		☑ UNSIGNED ☐ ZEROFILL	1
InteractiveAgent	VARCHAR(45)	✓		☐ BINARY	'Batch'
SubmitingSite	VARCHAR(50)	✓		☐ BINARY	
SumWaiting	DECIMAL(10,0)	✓		☐ UNSIGNED ☐ ZEROFILL	0

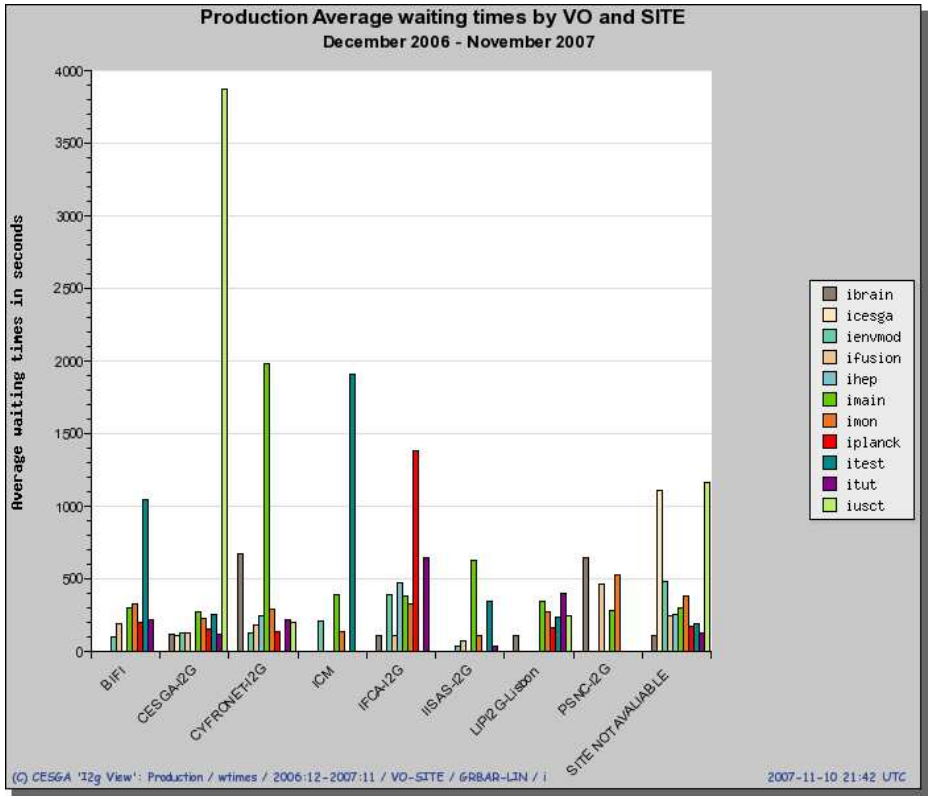
Waiting Time. Weighted average I

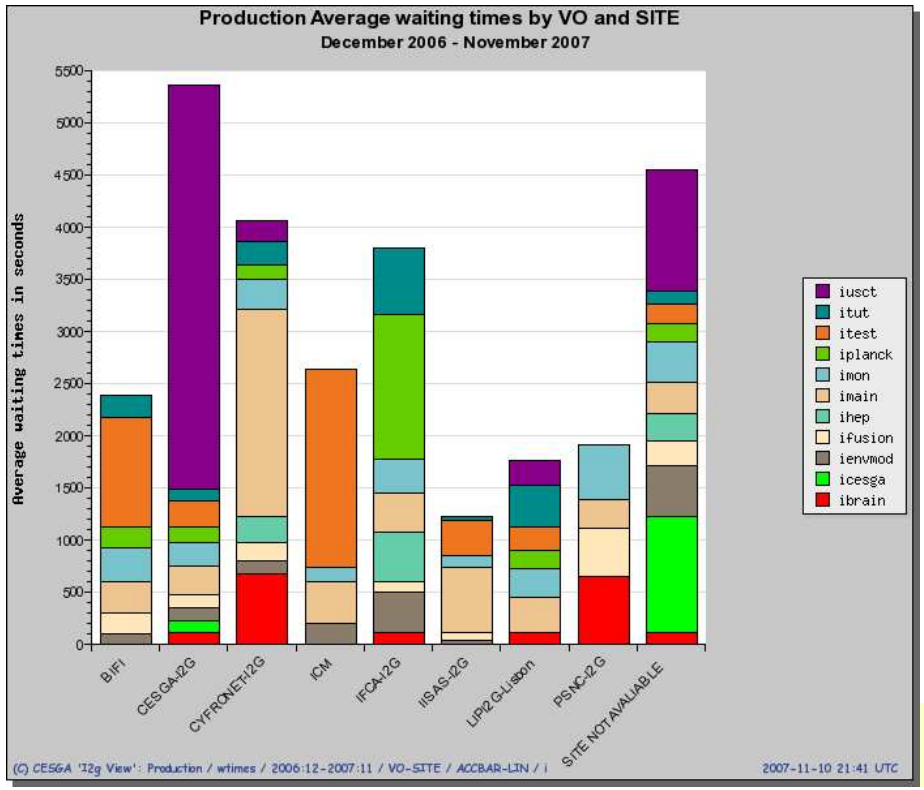
VO	BIFI	CESGA-I2G	CYFRONET-I2G	ICM	IFCA-I2G	IISAS-I2G	LIPI2G-Lisbon	PSNC-I2G	SITE NOT AVALIABLE	Weighted Time	%
ibrain	0	114	675	0	107	0	109	647	113	183	6.37%
icesga	0	111	0	0	0	0	0	0	1,109	503	4.40%
ienvmod	102	125	127	205	389	39	0	0	486	388	5.31%
ifusion	193	125	178	0	108	69	0	461	246	241	4.98%
ihep	0	0	249	0	471	0	0	0	259	337	3.53%
imain	304	271	1,984	391	379	631	342	280	303	562	17.62%
imon	327	226	289	138	326	107	277	529	383	241	9.39%
iplanck	201	154	136	0	1,378	0	168	0	172	1,099	7.97%
itest	1,044	251	0	1,906	0	341	233	0	192	362	14.31%
itut	219	115	219	0	646	33	398	0	123	374	6.32%
iusct	0	3,874	203	0	0	0	241	0	1,167	1,172	19.79%
Weighted Time	313	244	848	266	376	368	296	479	284	403	
Percentage	8.62%	19.36%	14.65%	9.52%	13.72%	4.40%	6.38%	6.92%	16.43%		
Click here for a csv dump of this table											

$$\overline{WT}(Site) = \frac{\sum_{VO} Njobs(VO,Site) \cdot WT(VO,Site)}{\sum_{VO} Njobs(VO,Site)} \qquad \overline{WT}(Site) = \frac{\overline{WT}(Site)}{\sum_{Site} \overline{WT}(Site)} \cdot 100$$

$$\overline{WT} = \frac{\sum_{Sites,VO} WT(VO,Site) \cdot Njobs(VO,Site)}{\sum_{Site,VO} Njobs(VO,Site)}$$

Waiting Time. Weighted average II



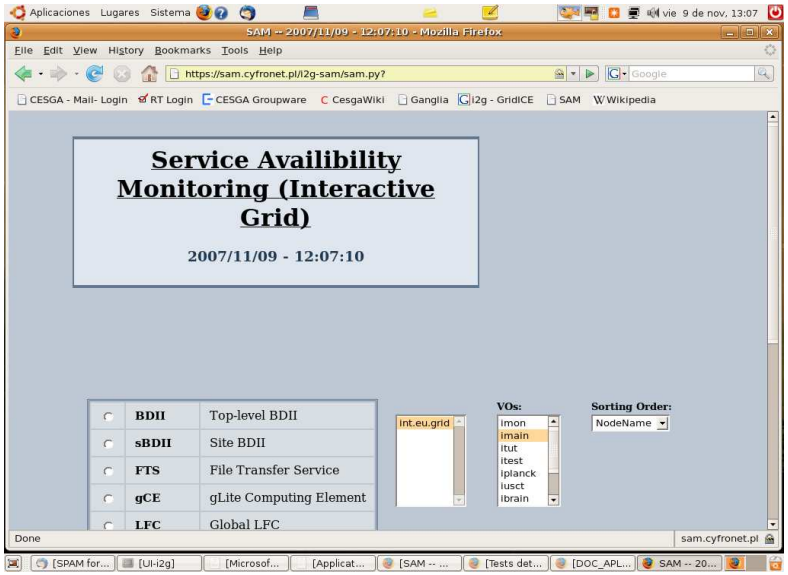


Monitoring

- Monitoring is an important proactive service that contributes to the detection and isolation of infrastructure problems. Site administrators need to know infrastructure status in order to solve any possible trouble. As the last chain point, users need reliability that is accomplished through proactive monitoring. Thus grid monitoring is required because all agents need to know about how the Grid is performing.
- Service Available Monitoring (SAM) Environment .
- GridIce servers (Production and Development).
- Central RGMA server.

SAM Monitoring

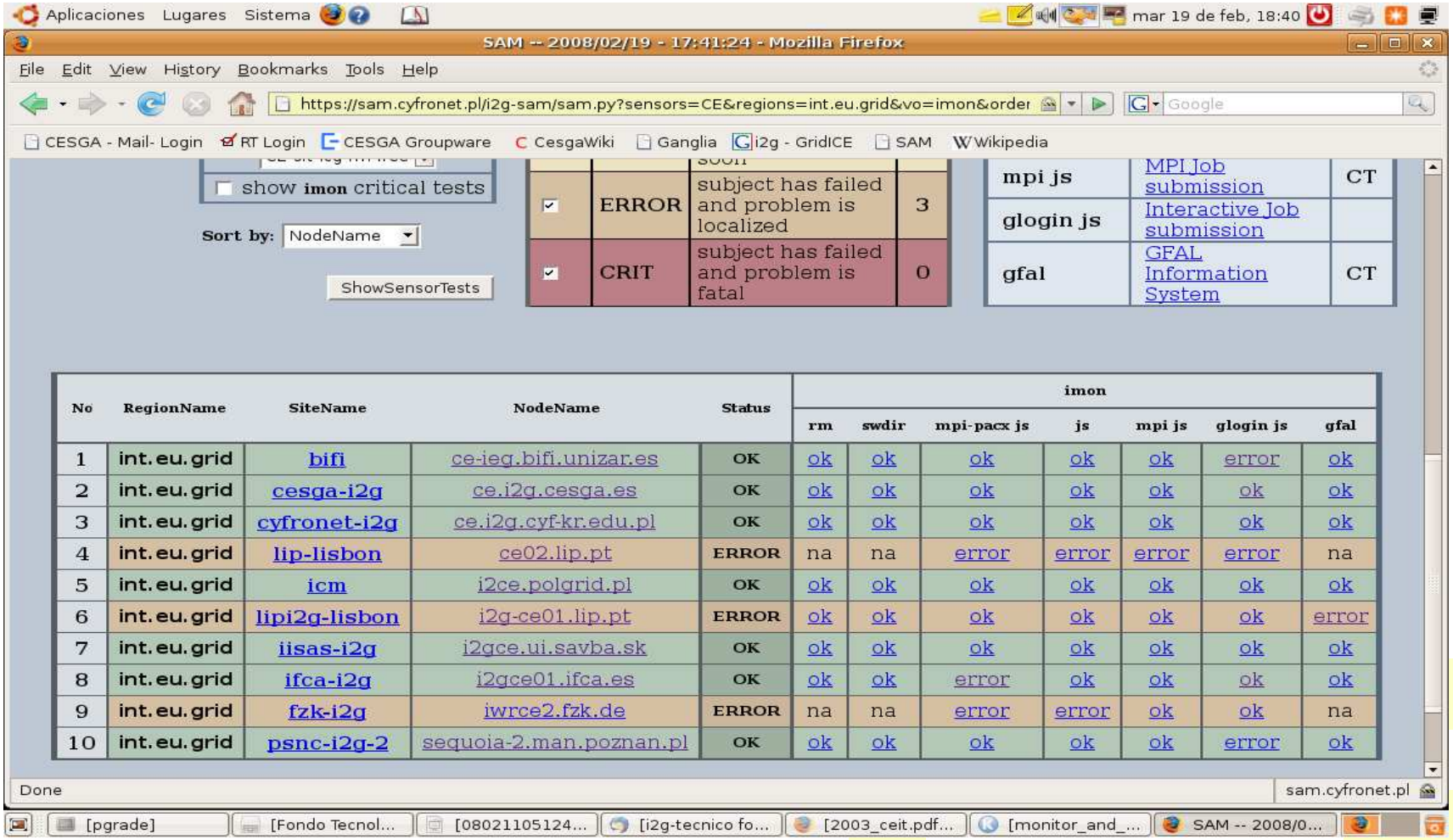
- The Infrastructure is monitored with SAME framework.
- https://sam.cyfronet.pl/i2g-sam/sam.py?
  - Three VOs:
    - imon
    - imain
    - ihep
  - Three sensors:
    - CE
    - SE
    - RB
  - Tests are executed hourly.



New SAM Tests

- New tests have been developed
  - MPI job submission.
    - ask for 2 nodes, computes pi.
  - PACX MPI job submission.
    - ask for 3 nodes,.
  - Interactive jobs with i2glogin.
- A daily version executed at 2:30 ask for 8 nodes (openmpi and pacx-mpi).

New SAM test



No	RegionName	SiteName	NodeName	Status	imon						
					rm	swdir	mpi-pacx js	js	mpi js	glogin js	gfal
1	int.eu.grid	bifi	ce-ieg.bifi.unizar.es	OK	ok	ok	ok	ok	ok	error	ok
2	int.eu.grid	cesga-i2g	ce.i2g.cesga.es	OK	ok	ok	ok	ok	ok	ok	ok
3	int.eu.grid	cyfronet-i2g	ce.i2g.cyf-kr.edu.pl	OK	ok	ok	ok	ok	ok	ok	ok
4	int.eu.grid	lip-lisbon	ce02.lip.pt	ERROR	na	na	error	error	error	error	na
5	int.eu.grid	icm	i2ce.polgrid.pl	OK	ok	ok	ok	ok	ok	ok	ok
6	int.eu.grid	lipi2g-lisbon	i2g-ce01.lip.pt	ERROR	ok	ok	ok	ok	ok	ok	error
7	int.eu.grid	iisas-i2g	i2gce.ui.savba.sk	OK	ok	ok	ok	ok	ok	ok	ok
8	int.eu.grid	ifca-i2g	i2gce01.ifca.es	OK	ok	ok	error	ok	ok	ok	ok
9	int.eu.grid	fzk-i2g	iwrce2.fzk.de	ERROR	na	na	error	error	ok	ok	na
10	int.eu.grid	psnc-i2g-2	sequoia-2.man.poznan.pl	OK	ok	ok	ok	ok	ok	error	ok

OpenMPI test example

Generating JDL file:

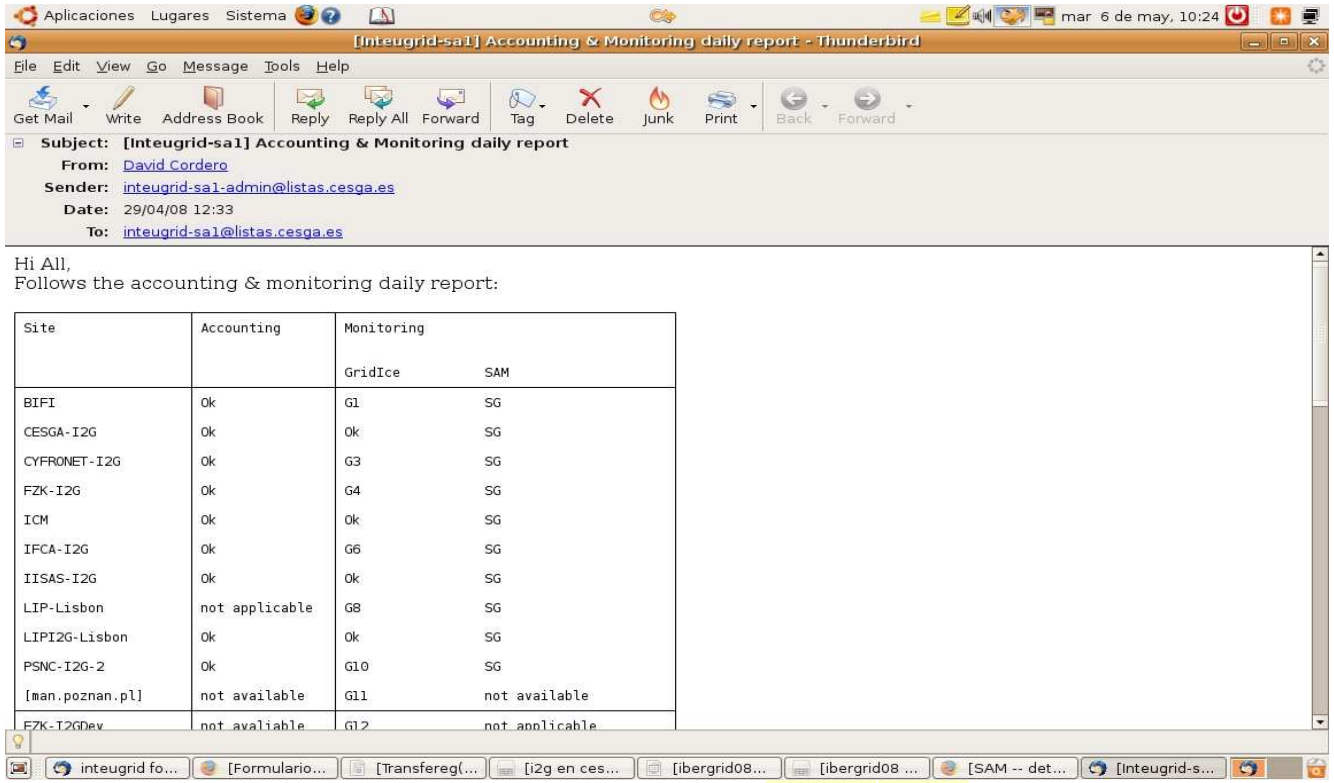
```
JobType             = "openmpi";
NodeNumber          = 2;
VirtualOrganisation = "imain";
Executable          = "cpi";
StdOutput           = "mpi-testjob.out";
StdError            = "mpi-testjob.err";
InputSandbox        = {"cpi"};
OutputSandbox       = {"mpi-testjob.out","mpi-testjob.err"};
Environment = {"I2G_MPI_START=/opt/i2g/bin/mpi-start","I2G_MPI_START_DEBUG=1","I2G_MPI_START_VERBOSE=1"};
Requirements    = (other.GlueCEInfoHostName == "ce.i2g.cesga.es");
```

Submitting a job

```
+ i2g-job-submit --vo imon -o mpi-testjob.jid mpi-testjob.jdl
Selected Virtual Organisation name (from --vo option): imon
Connecting to host i2g-rb01.lip.pt, port 7772
Logging to host i2g-rb01.lip.pt, port 9002
============================ i2g-job-submit Success ============================
 The job has been successfully submitted to the Network Server.
 Use i2g-job-status command to check job current status. Your job
 identifier (edg_jobId) is:
 - https://i2g-rb01.lip.pt:9000/6gjtF64qEjPHAG662S9JqA
 The edg_jobId has been saved in the following file:
 /home2/people/ymskital-imon/.same/CE/nodes/ce.i2g.cesga.es/mpi-testjob.jid
==============================================================================
```

New Services

- A dayly report sent notifiying all sites monitoring and accounting status



Aplicaciones Lugares Sistema — mar 6 de may, 10:24

[Inteugrid-sa1] Accounting & Monitoring daily report - Thunderbird

File Edit View Go Message Tools Help

Get Mail | Write | Address Book | Reply | Reply All | Forward | Tag | Delete | Junk | Print | Back | Forward

Subject: [Inteugrid-sa1] Accounting & Monitoring daily report
From: David Cordero
Sender: inteugrid-sa1-admin@listas.cesga.es
Date: 29/04/08 12:33
To: inteugrid-sa1@listas.cesga.es

Hi All,
Follows the accounting & monitoring daily report:

Site	Accounting	Monitoring	
		GridIce	SAM
BIFI	Ok	G1	SG
CESGA-I2G	Ok	Ok	SG
CYFRONET-I2G	Ok	G3	SG
FZK-I2G	Ok	G4	SG
ICM	Ok	Ok	SG
IFCA-I2G	Ok	G6	SG
IISAS-I2G	Ok	Ok	SG
LIP-Lisbon	not applicable	G8	SG
LIPI2G-Lisbon	Ok	Ok	SG
PSNC-I2G-2	Ok	G10	SG
[man.poznan.pl]	not available	G11	not available
FZK-I2GDev	not available	G12	not applicable

inteugrid fo... | [Formulario... | [Transfereg(... | [i2g en ces... | [ibergrid08... | [ibergrid08 ... | [SAM -- det... | [Inteugrid-s...

Questions?

- carlosf@cesga.es
- egee-admin@cesga.es