



Enabling Grids for E-science

SGE

J. Lopez, A. Simon, E. Freire, G. Borges, K. M. Sephton

All Hands Meeting

Barcelona, Spain

23 May 2007



Information Society



■ Introduction

■ LRMS comparison

- LSF, Torque/Maui, Condor and Sun Grid Engine

■ Sun Grid Engine gLite integration (for the Icg-CE)

- JobManager
- Accounting Information
- Information plug-in
- YAIM Integration

■ Conclusions and Future Work

- **948 registered users**
- **HP Superdome:**
 - 2 nodes SMP Cluster
 - 128 Itanium2
 - HP-UX
- **SVG:**
 - 423 cores
 - PIV and Xeon
 - Linux
- **Compaq HPC320**
 - 8 SMP nodes
 - 32 Alpha processors
 - Tru64

LRMS	Pros	Cons
<p>LSF</p>	<ul style="list-style-type: none"> ■ Flexible Job Scheduling Policies ■ Advance Resource Management <ul style="list-style-type: none"> ○ Checkpointing & Job Migration, Load Balacing ■ Good Graphical Interfaces to monitor Cluster functionalities 	<ul style="list-style-type: none"> ■ Expensive comercial product
<p>Torque/ Maui</p>	<ul style="list-style-type: none"> ■ Very well known because it comes from PBS: Torque=PBS+bug fixes ☺ ■ Good integration of parallel libraries ■ Flexible Job Scheduling Policies <ul style="list-style-type: none"> ○ Fair Share Policies, Backfilling, Resource Reservations ■ Very good support in gLite 	<ul style="list-style-type: none"> ■ Two separate products -> Two separate configurations ■ No user friendly GUI to configuration and management ■ Software development uncertain ■ Bad documentation
<p>Condor</p>	<ul style="list-style-type: none"> ■ CPU harvesting ■ Special ClassAds language ■ Dynamic check-pointing and migration ■ Mechanisms for Globus Interface 	<ul style="list-style-type: none"> ■ Not optimal to parallel applications ■ Complex configuration


- **Grid Engine, an open source job management system developed by Sun**
 - Queues are located in server nodes and have attributes which characterize the properties of the different servers
 - A user may request at submission time certain **execution features**
 - *Memory, execution speed, available software licences, etc*
 - Submitted jobs wait in a holding area where its requirements/priorities are determined
 - *It only runs if there are queues (servers) matching the job requests*
- **N1 Grid Engine, commercial version including support from Sun**
- **Some Important Features**
 - Extensive **operating system** support
 - **Flexible Scheduling Policies:** Priority; Urgency; Ticket-based; Share-Based, Functional, Override
 - Supports **Subordinate Queues**
 - Supports **Array Jobs**
 - Supports **Interactive Jobs** (qlogin)
 - **Complex Resource Attributes**
 - **Shadow Master Hosts** (high availability)
 - Accounting and Reporting Console (**ARCo**)
 - **Tight integration of parallel libraries**
 - Implements **Calendars** for Fluctuating Resources
 - Supports **Check-pointing and Migration**
 - Supports **DRMAA 1.0**
 - **Transfer-queue Over Globus (TOG)**
 - **Intuitive Graphic Interface**
 - Used by users to manage jobs and by admins to configure and monitor their cluster
 - **Good Documentation:** Administrator's Guide, User's Guide, mailing lists, wiki, blogs
 - Enterprise-grade scalability: 10,000 nodes per one master (promised 😊)



- CLI: qconf

- GUI interface: qmon




Job Control

Pending Jobs
Running Jobs
Finished Jobs

JobId	Priority	JobName	Owner	Status	Queue
1316661	3.49928	mpijob.sh	usceljlc	qw	*pending*
1316649	1.50024	wrfprep_in	uscfmifd	qw	*pending*
1316651	1.49988	geogrid.sh	uscfmifd	qw	*pending*
1316656	1.49910	cola.sh	uscFMLMR	qw	*pending*
1316644	1.49258	PdR1BrLLR2	uviqoar1	qw	*pending*
1316396	1.37689	exe8_t.sh	uscelrvf	qw	*pending*
1316646	1.22073	pru5.sh	cseezjoc	qw	*pending*
1316590	1.16650	run2gs1dd.	uscfaeci	qw	*pending*
1316592	1.16650	run2gs1du1	uscfaeci	qw	*pending*
1316594	1.16650	run2gs1du2	uscfaeci	qw	*pending*
1316596	1.16650	run2gs1uu.	uscfaeci	qw	*pending*
1316616	1.16647	run2gs1uu.	uscFambf	qw	*pending*
1316474	1.09792	sas7Cioba	ulcqivoc	qw	*pending*
1316478	1.09779	tasbNio2a	ulcqimrp	qw	*pending*
1316480	1.09777	taabCioa	ulcqimrp	qw	*pending*
1316658	1.06583	vai	uscqfecl	qw	*pending*
1316663	1.02215	pru1.sh	cseeznhh	qw	*pending*

Refresh

Submit

Tickets

Force

Suspend

Resume

Delete

Reschedule

Select All

Why ?

Hold

Priority


Qalter

Clear Error

Customize

Done

Help


Complex Configuration

Attributes

Name	Shortcut	Type	Relation	Requestable	Consumable	Default	Urgency
<input type="text"/>	<input type="text"/>	INT	==	NO	NO	0	0

Name	Shortcut	Type	Relation	Requestable	Consumable	Default	Urgency
s_core	s_core	MEMORY	<=	YES	NO	0	0
s_cpu	s_cpu	TIME	<=	YES	NO	0:0:0	0
s_data	s_data	MEMORY	<=	YES	NO	0	0
s_fsize	s_fsize	MEMORY	<=	YES	NO	0	0
s_rss	s_rss	MEMORY	<=	YES	NO	0	0
s_rt	s_rt	TIME	<=	FORCED	NO	0:0:0	-10
s_stack	s_stack	MEMORY	<=	YES	NO	0	0
seq_no	seq	INT	==	NO	NO	0	0
slots	s	INT	<=	YES	YES	1	1000
swap_free	sf	MEMORY	<=	YES	NO	0	0
swap_rate	sr	MEMORY	>=	YES	NO	0	0
swap_rsvd	srsv	MEMORY	>=	YES	NO	0	0
swap_total	st	MEMORY	<=	YES	NO	0	0
swap_used	su	MEMORY	>=	YES	NO	0	0
tmpdir	tmp	RESTRING	==	NO	NO	NONE	0
virtual_free	vf	MEMORY	<=	YES	NO	0	0
virtual_total	vt	MEMORY	<=	YES	NO	0	0
virtual_used	vu	MEMORY	>=	YES	NO	0	0
x86	x86	RESTRING	==	YES	NO	NONE	0
s_vmem	s_vmem	MEMORY	<=	FORCED	YES	0	0

■ The JM is the core service of the **Globus GRAM Service**

- Submits jobs to SGE based on Globus requests and through a **jobwrapper** script
- Intermediary to query the status of jobs and to cancel them

■ SGE command client tools (qstat, qsub, qdel) have to be **available in the CE**

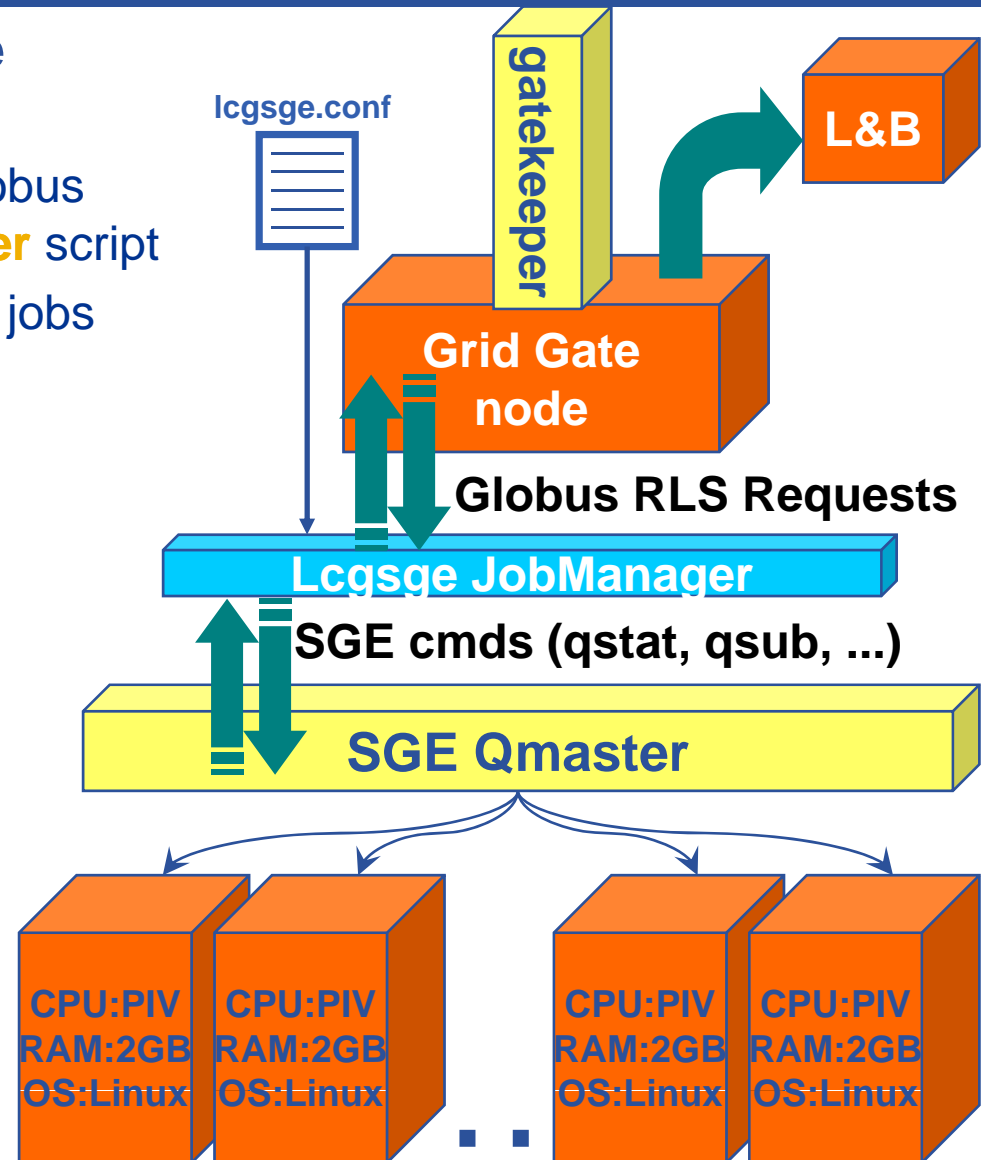
- Even if the Qmaster machine is installed in another machine

■ **Doesn't require shared homes**

- But home dirs must have the same path on the CE and WNs

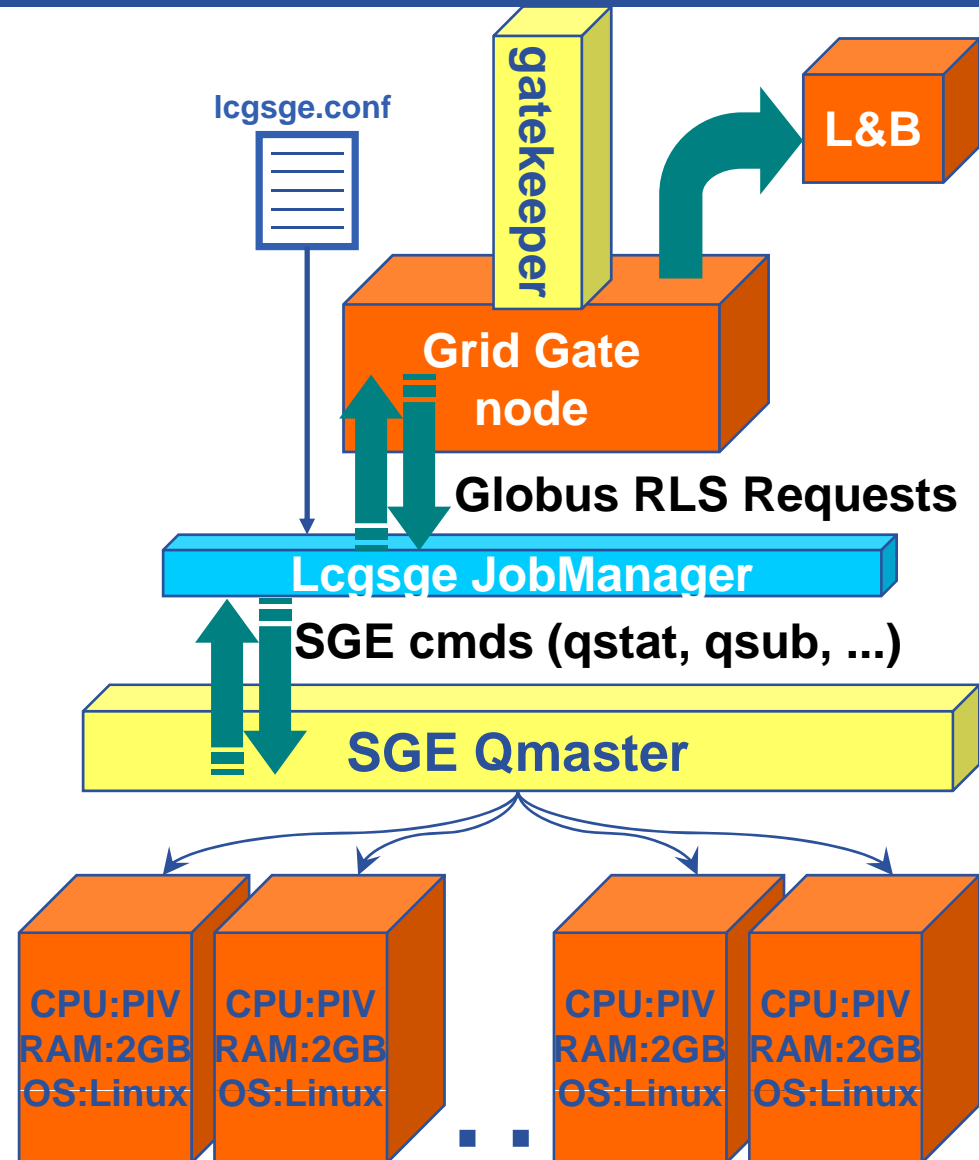
■ The SGE JM is based on the **LCGPBS JM**

- Requires XML::Simple.pm



■ SGE JM re-implements the following functions:

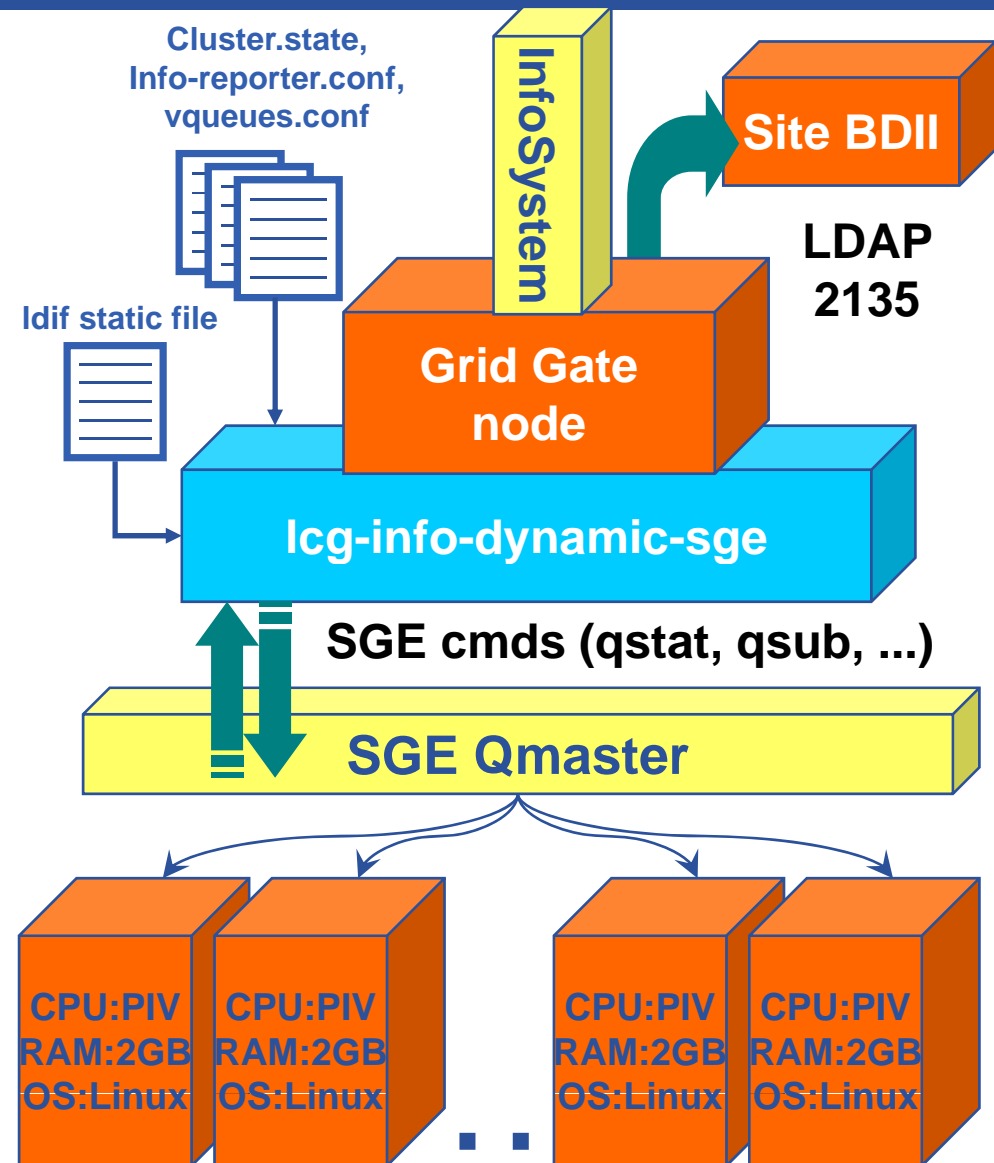
- **Submit:** Checks Globus RSL arguments returning a Globus error if the arguments are not valid or if there are no resources
- **Submit_to_batch_system:** Submits jobs to SGE, after building the **jobwrapper** script, by getting the necessary information from the RSL variables
- **Poll:** Links the present status of jobs running in SGE with the Globus appropriate message
- **Poll_batch_system:** Allows to know the status of running jobs parsing the **qstat** SGE output.
- **Cancel_in_batch_system:** Cancels jobs running in SGE using **qdel**



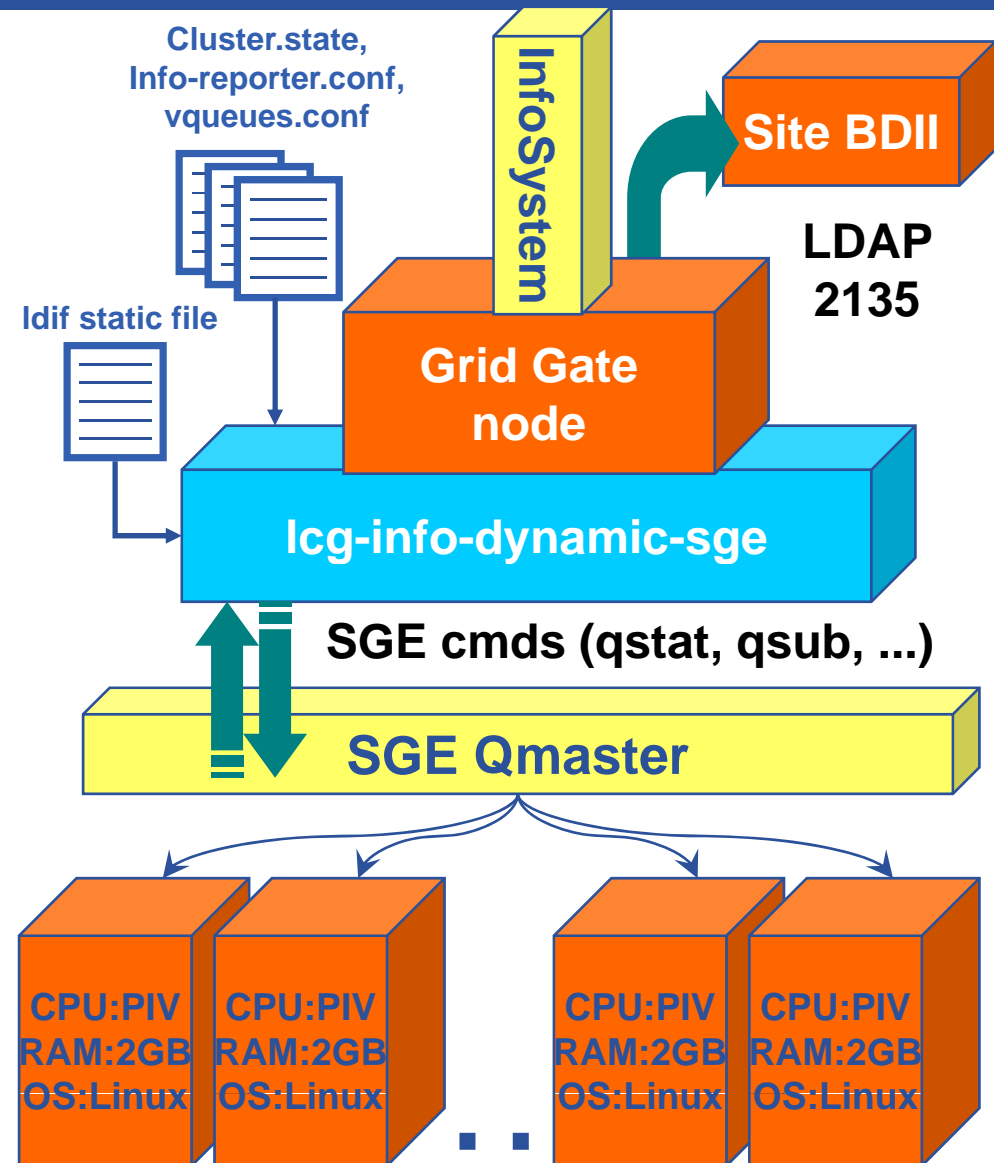
- The solution implemented for SGE does not currently use the generic EGEE scripts
 - **lcg-info-dynamic-sge**
 - A standalone Information plugin script that examines SGE queuing system state

- Information expected to be reported is based on queues
 - SGE does not assign a job to a queue until execution time.
 - **virtual queues** are used

- The info reporter reads...
 - A copy of a static Idif file with details of all "virtual queues"
 - **Config files** specifying how virtual queues map into a list of resource requirements



- The dynamic information
 - single call to SGE's "qstat"
- The system determines which virtual queues the job should be associated with
- Each virtual queue is considered to count up
 - Nb of job slots, Nb of pending/running jobs
 - Total amount of runtime left on all of the jobs assuming that they will run for their max duration
- The state of the batch queues can change quite fast ...
 - Option to capture a copy of all information provider input data, which can be replayed to the information provider

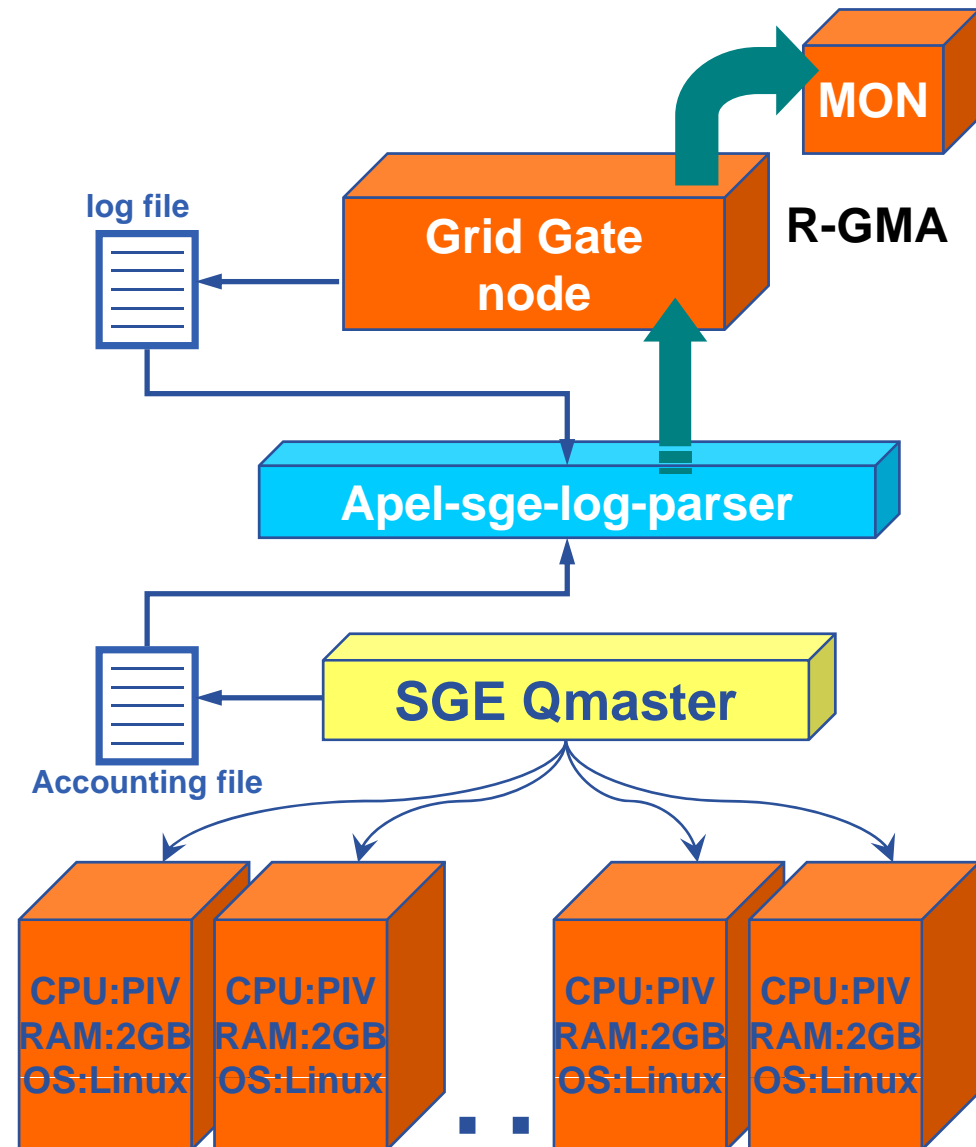


- Information provider produced which allows for a greater variety of SGE configurations. This is continuing to be improved based on feedback from other sites.
- Work started on IP for gliteCE (*blah port*). Problems getting reference gliteCE installed.
- Latest IP release contains a patch to work on a gLiteCE
- Again, aiming to be as flexible as possible to fit with existing SGE configurations.

- **APEL SGE plug-in is a log processing application**
 - Used to produce CPU job accounting records
 - Interprets gatekeeper & batch system logs

- **Requires the JM to add ``gridinfo'' records in the log file (not anymore)**
 - Standard Globus JMs do not log them but LCG JMs do it

- **apel-sge-log-parser parses the SGE accounting log file**
 - This information, together with the gridinfo mappings from the JobManager are joined together to form accounting records
 - Published using R-GMA to an accounting database.



■ Development of **two integration rpms**

- lcgCE-yaimtosge-0.0.0-2.i386.rpm
- gliteWN-yaimtosge-0.0.0-2.i386.rpm
- Requirements
 - SGE installed (we presently made SGE rpms to install it)
 - lcg-CE and glite-WN
 - glite-yaim ($\geq 3.0.0-34$), perl-XML-Simple ($\geq 2.14-2.2$), openmotif ($\geq 2.2.3-5$) and xorg-x11-xauth ($\geq 6.8.2-1$)

- **\$SGE_ROOT** software dir must be set to **/usr/local/sge/pro**
 - May be changed by the site admin in a future release

- **The SGE Qmaster can only be installed in the CE**
 - May be installed in another machine in a future release

- **Three new variables must be set in the **site-info.def****
 - **SGE_QMASTER, DEFAULT_DOMAIN, ADMIN_EMAIL**

- **The integration rpms do...**
 - Change the **node-info.def** file to include two new node types
 - CE_sge and WN_sge
 - Run the same functions as the CE and WN nodes, plus at the end
 - *Config_sge_server and Config_sge_client*

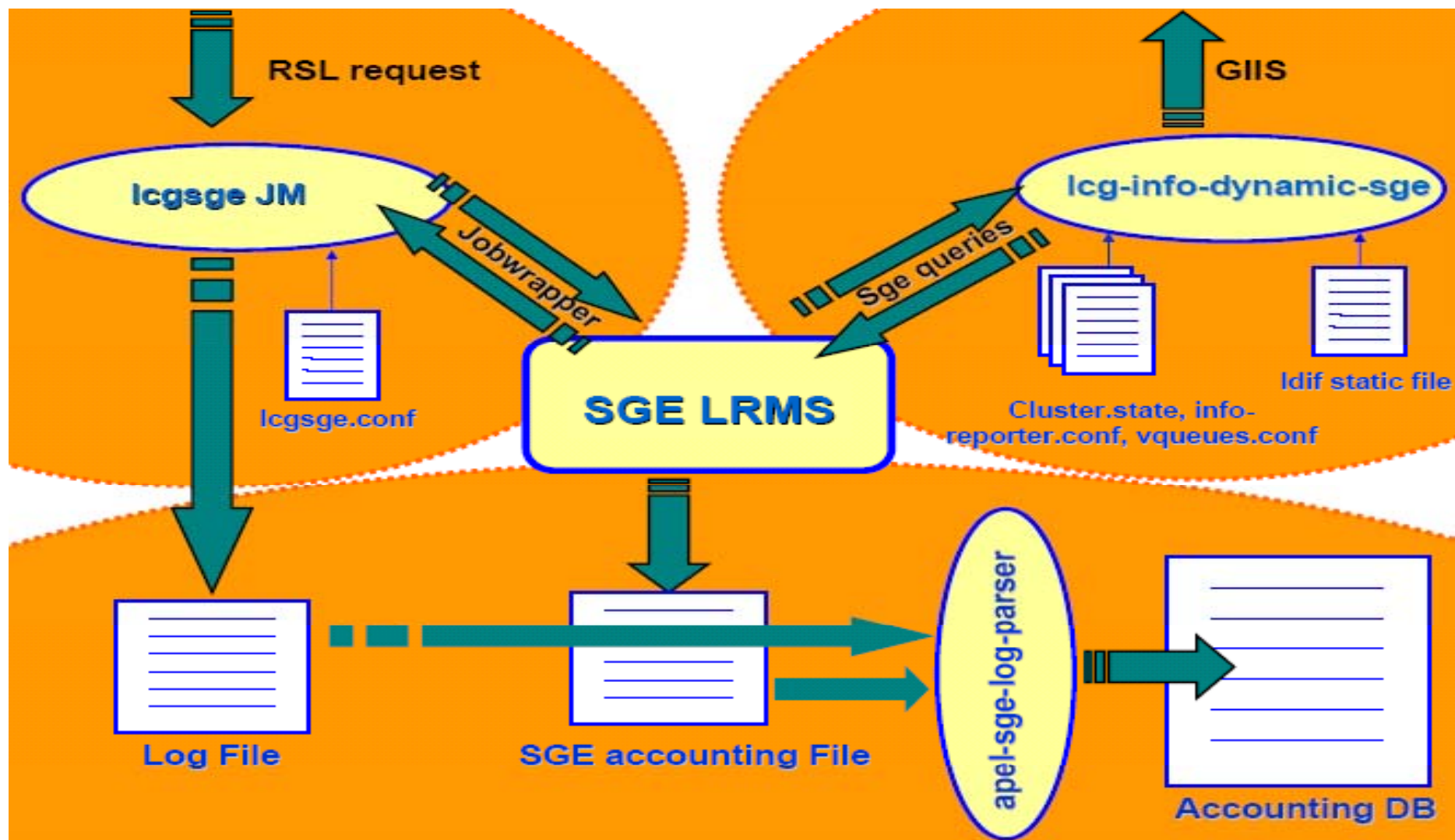
■ The **Config_sge_server**

- Uses an auxiliary perl script (**configure_sge_server.pm**)
 - Builds all the default SGE directory structure
 - Configures environment setting files, sets the global SGE configuration file, the SGE scheduler configuration file and SGE complex attributes
- Defines one cluster queue for each VO
- Deploys the **lcgsgc JM** and builds its configuration files
- Deploys **SGE Information plug-in** and builds its configuration files
- Accounting is not properly integrated but will be soon...

■ The **Config_sge_client**

- Uses an auxiliary perl scrip (**configure_sge_client.pm**)
 - Builds all the default SGE directory structure in the client

`/opt/glite/yaim/bin/yaim -c -s site-info.def -n CE_sge`



- **SGE is working on a lcg-CE although additional work is required**
 - **YAIM SGE integration**
 - More flexible allowing site admins to dynamically set a broader range of options
 - Separate Qmaster from the CE
 - Fully integrate the SGE Accounting
 - **SGE Information Provider** needs to improve its flexibility and take into account overlapping cluster queues / virtual queues definitions. Some bugs have been detected and they are being solved.

- **Started on integrating support for BLAH, running on glite-CE**
 - Work started on blah port. Problems getting reference gliteCE installed. Expect initial release next month. Again, aiming to be as flexible as possible to fit with existing SGE configurations.
 - Will be used within glite-CE and CREAM to interface with the LRMS
 - Expected to share the configuration files and concept of virtual queues with the information provider.
 - Other local middleware elements (GIIS, YAIM) basically remain unchanged for this glite-CE flavour.

- **Still missing**
 - GridICE sensors for SGE

■ Grid Engine

- <http://gridengine.sunsource.net/>

■ N1 Grid Engine

- <http://www.sun.com/software/gridware/index.xml>

■ SGE Wiki Page

- <https://twiki.cern.ch/twiki/bin/view/LCG/ImplementationOfSGE>

Thank you!

