
# EnergySaving Cluster experience in CETA-CIEMAT

Manuel F. Dolz, Juan C. Fernández, Sergio Iserte, Rafael Mayo,
Enrique S. Quintana, Manuel E. Cotallo, and Guillermo Díaz

June 8–10, 2011, Santander (Spain)

# Motivation

High Performance Computing Clusters in the Grid Infrastructure:

- Normally composed by a high number of nodes
- Multi-processors/multi-cores nodes at high frequencies
- Infrastructure requires big cooling systems

- High power consumption
- Environmental impact and high economic cost

- Power-aware techniques and tools to reduce negative effects

# Outline

# Objectives

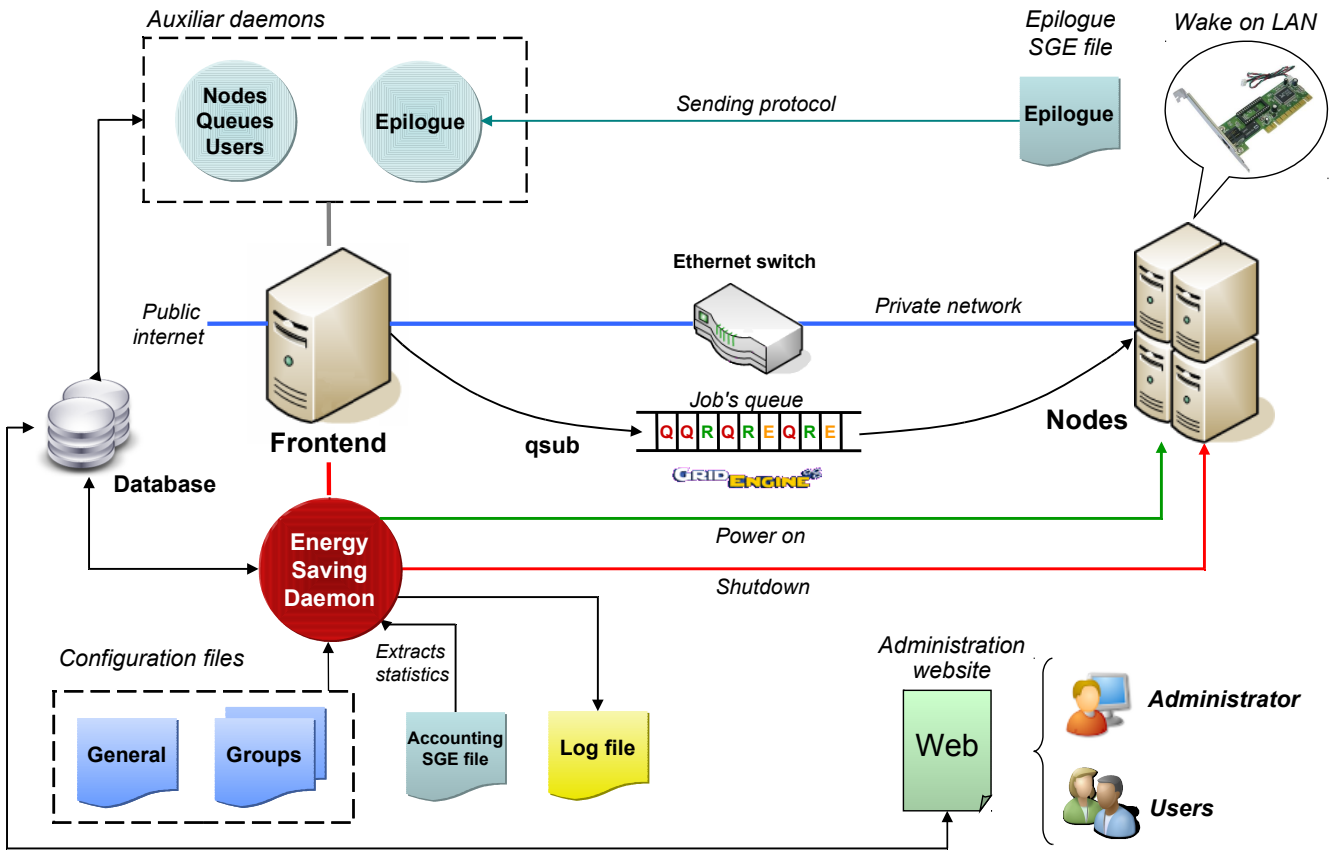
- Development of a middleware that implements energy saving policies to turn on/off nodes of a clusters taking into consideration past and future computational load

Find a solution!

EnergySaving Cluster

- Evaluate the performance of this middleware to the CETA-CIEMAT Grid Computing Center

# Middleware architecture

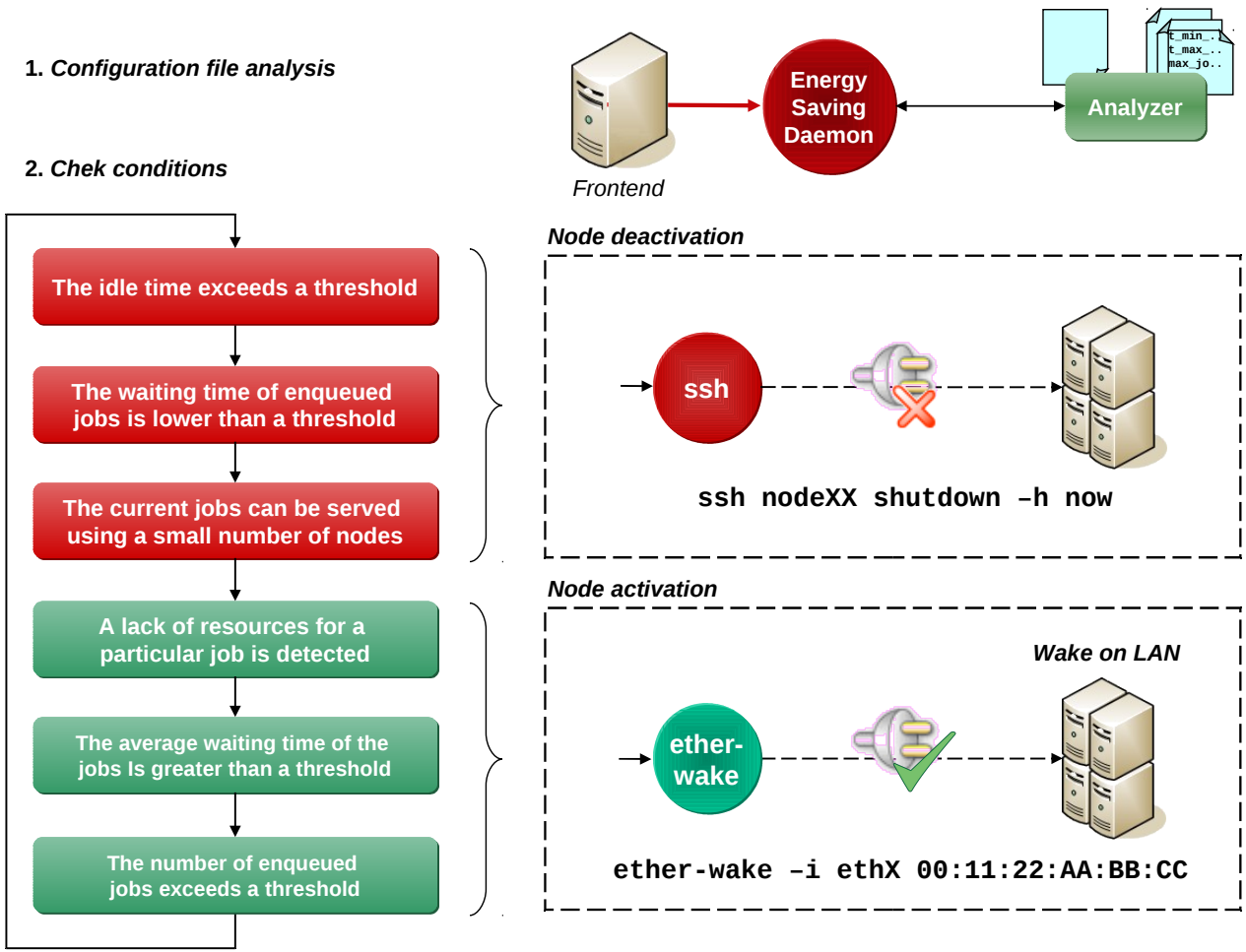


# System architecture

The module includes the following components:

- Three daemons to manage the database, collect statistics and execute the commands that power on/off the nodes
- The database stores all necessary information to make decisions
- The web interface to configure and administer users' groups and set the threshold to define the power saving policy

# Daemon for activation/deactivation actions



# Installation and compatibility issues

CETA-CIEMAT is the first grid-computing center where EnergySaving Cluster (ESC) middleware has been installad and tested by Jaume I University developers

Installation and tuning issues:

- Compatibility:
  - Computing node hardware issues
    - WOL capability is required in computing nodes
  - Network concerns
    - Nodes must be contained in the same layer 2 subnetwork, as well as, node hosting daemons

# Installation and compatibility issues

- Adaptation for datacenter architecture and SGE's own configuration:
  - Web frontend, database and daemons running in the same node
    - Use TCP sockets instead of UNIX sockets to host modules in different machines
  - ESC daemons run in the same SGE master node
    - Adapt daemons to connect remotely to SGE for issuing q-commands
    - Adapt system to use remotely SGE's accounting logs
  - There is not a notion of isolated cluster queues with dedicated computing resources
    - ESC involves the whole SGE system, and, currently, do not work with a grouped resources infrastructure

# Testing environment

In order to verify that a correct ESC deployment was made in CETA-CIEMAT, the following testing environment was set as follows:

- Web frontend/ESC daemons machine/MySQL machine:
  - Hosted in the same VMWare virtual machine, 1 processor, 1GB RAM, CentOS 5.3
- Subclusters "A" and "B":
  - 5 machines each, Bull Novascale, 2 Intel 5230 quad-core processors, 16GB RAM, Scientific Linux 5.3

→ ESC database was modified to collect data about power of Bull's Novascale chassis

# User tests

Established test plan in CETA-CIEMAT's environment

- Minimal functional tests
  - Loop simulating arrival of sequential/parallel jobs with no processing (sleeping 30 sec.)
- Stress tests
  - Bursts of ultra-short jobs (1 s), CPU intensive (99 %)
  - Bursts of short jobs (1 h), CPU intensive (99 %), with a period of 1 hour between bursts
- Performance tests
  - Some performance data were gathered during strees tests to taken into account for simulation purposes

# Integration within gLite (I)

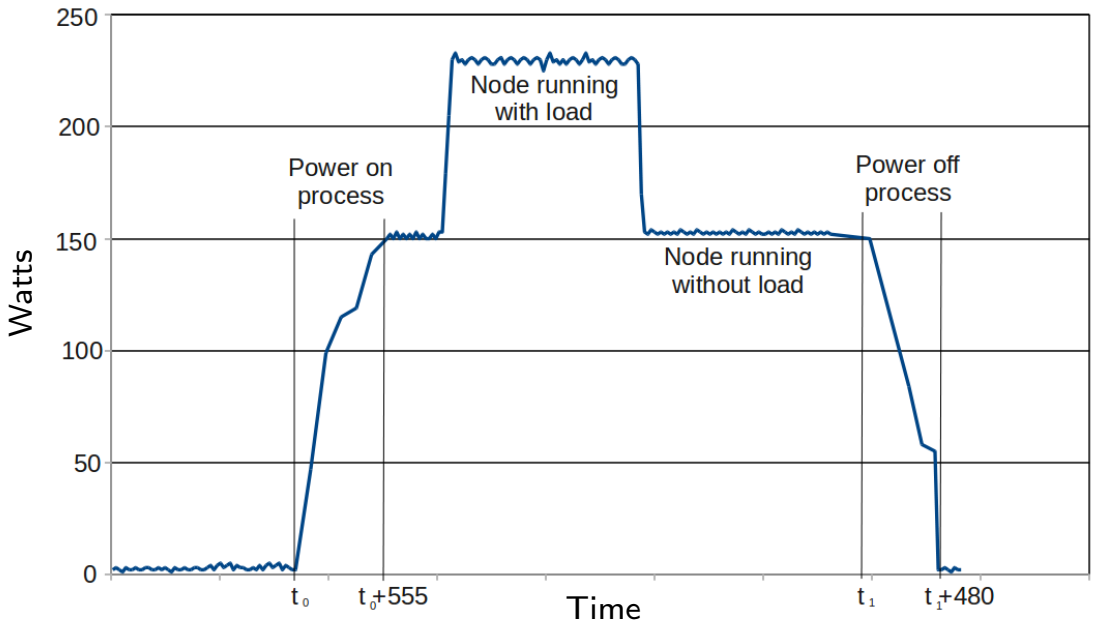
- Tight integration is not possible right now. Why?
  - How would information systems show CPU resources when "asleep"? $\rightarrow$ As not available.
  - Statistics of availability and reliability of affected sites
- Future: Information System's schemas needs a change to reflect different:
  - CPU states (available, offline, asleep)
  - QoS of resources (quick-online, slower-asleep)

# Integration within gLite (II)

- Batch Queue information providers for information systems need to be modified accordingly
  - Sun Grid Engine can reflect asleep nodes? $\rightarrow$ No, but, maybe "a" state of node queues can be used.
- Should GLUE Schema for Information Systems be changed?
  - Sure, not just due to ESC, but for any power saving schema that needs to stop nodes.

# Configuration

- We have configured a simulator of ESC with power consumption parameters of nodes int the CETA-CIEMAT:
  - 16 nodes with 8 cores per node
  - Power real data



# Benchmark and experimental results

- We have used the following of synthetic workloads form the Paraellel Workloads Archive:
  - OSC: OSC Linux Cluster, a workload composed of 80,714 jobs
  - NASA: NASA Ames iPSC/860 is a set of 42,264 jobs
- From our simulation we have obtained the following table wich displays the energy savings:

| **Workload** | **Time (days, hours, minutes, seconds)** | **Energy (MWh)** |
|---|---|---|
| OSC without ESC | 677 d, 2 h, 55 m, 51 s | 40.42 MWh |
| OSC with ESC | 868 d, 20 h, 50 m, 39 s | 12.87 MWh |
| NASA without ESC | 92 d, 0 h, 3 m, 43 s | 6.72 MWh |
| NASA with ESC | 92 d, 0 h, 12 m, 59 s | 4.79 MWh |

# Experimental results

- From our simulation we have obtained the following table wich displays the energy savings:

| Workload | Time (days, hours, minutes, seconds) | Energy (MWh) |
|---|---|---|
| OSC without ESC | 677 d, 2 h, 55 m, 51 s | 40.42 MWh |
| OSC with ESC | 868 d, 20 h, 50 m, 39 s | 12.87 MWh |
| NASA without ESC | 92 d, 0 h, 3 m, 43 s | 6.72 MWh |
| NASA with ESC | 92 d, 0 h, 12 m, 59 s | 4.79 MWh |

Conclusions of these results:

- It is possible to obtain an important level on energy savings with ESC.
- Depending on the load, the throughtput can be lowered (e.g. OSC load).

# Experimental results

- From our simulation we have obtained the following table wich displays the energy savings:

| Workload | Time (days, hours, minutes, seconds) | Energy (MWh) |
|---|---|---|
| OSC without ESC | 677 d, 2 h, 55 m, 51 s | 40.42 MWh |
| OSC with ESC | 868 d, 20 h, 50 m, 39 s | 12.87 MWh |
| NASA without ESC | 92 d, 0 h, 3 m, 43 s | 6.72 MWh |
| NASA with ESC | 92 d, 0 h, 12 m, 59 s | 4.79 MWh |

Conclusions for the OSC load:

- The time to process all the jobs is increased by a factor of 28 %
- Energy consumption with ESC is reduced by a factor of 68 %

# Experimental results

- From our simulation we have obtained the following table wich displays the energy savings:

| Workload | Time (days, hours, minutes, seconds) | Energy (MWh) |
|---|---|---|
| OSC without ESC | 677 d, 2 h, 55 m, 51 s | 40.42 MWh |
| OSC with ESC | 868 d, 20 h, 50 m, 39 s | 12.87 MWh |
| NASA without ESC | 92 d, 0 h, 3 m, 43 s | 6.72 MWh |
| NASA with ESC | 92 d, 0 h, 12 m, 59 s | 4.79 MWh |

Conclusions for the NASA load:

- The time to process all the jobs is increased by a factor of 0.000069 %
- Energy consumption with ESC is reduced by a factor of 29 %.

# Summary and conclusions

Conclusions:

- EnergySaving Cluster middleware implements a power-on/power-off policy so that, at any moment only the necessary computational resources are actuve, and those that are not needed remain powered off
- Modular design: enables integration with different queue systems, e.g. Sun Grid Engine, Portable Bath System/Torque or SLURM
- We have developed a **simulator** in order to evaluate the energy savings produced by our middleware in a production environment:
  - Usefulness to evaluate how affects the productivity and performance on the system
  - Predict the potential energy savings
- We have also discussed how to integrate the middleware into gLite environment and SGE queue system

# Thanks for your attention!

*Questions?*