EnergySaving Cluster Roll: Power Saving System for Clusters

HPC A
High Performance Computing and Architectures

Manuel F. Dolz, Juan C. Fernández, Rafael Mayo, Enrique S. Quintana-Ortí

High Performance Computing & Architectures (HPCA)
University Jaume I - Castellón (Spain)

Outline

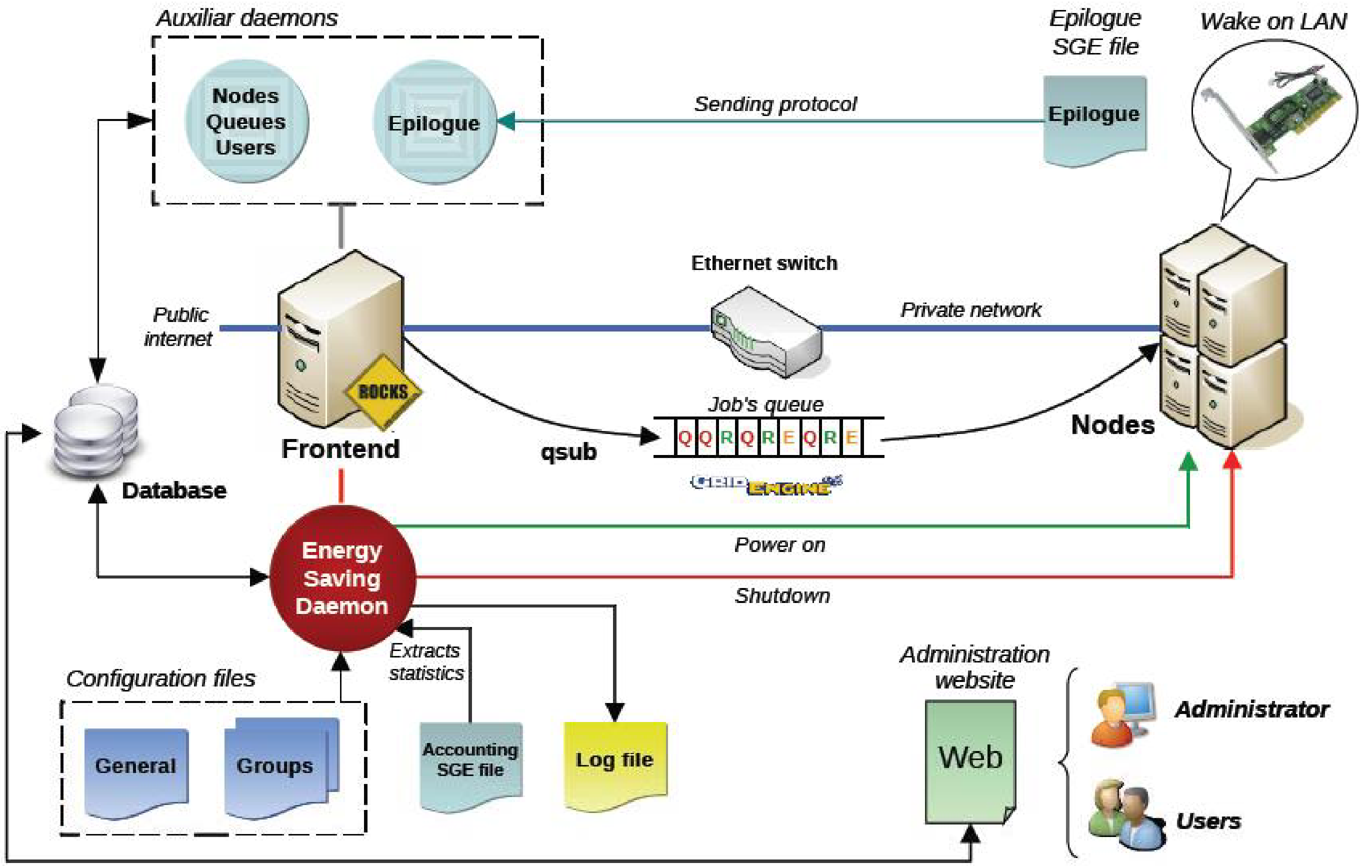
- Objectives
- Implementation of the Energy Saving Roll
- Experimental Results
- Conclusions

Objectives

- Development of a middleware that implements energy saving policies to turn on/off nodes of a cluster taking into consideration past and future users' requests

EnergySaving Roll

Implementation of Energy Saving Roll



Implementation of Energy Saving Roll

- The module includes the following components:

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

The three daemons

- Daemon for epilogue requests
  - To perform a series of updates in the energy saving database
- Daemon for the queues, users and nodes
  - To ensure that all information on users, nodes and queues is correctly reflected by the database
- Daemon for activation/deactivation actions and statistics
  - To activate/deactivate the nodes
  - To compare the threshold set by system administrator with the current values from the database to test if the activation/deactivation conditions are satisfied

Node activation conditions

- There are not enough appropriate active resources to run a job
- The average waiting time of a job in the queue exceeds a given threshold
- The number of jobs in the queue for a user exceeds the maximum value for its group

Options to select candidate nodes to turn on

- Ordered: By the name of the node
- Randomize: Randomly
- Balanced: Period that the nodes were active during the last *t* hours
- Prioritized: A priority assigned by the system administrator

Options to specify a strict threshold to power on nodes

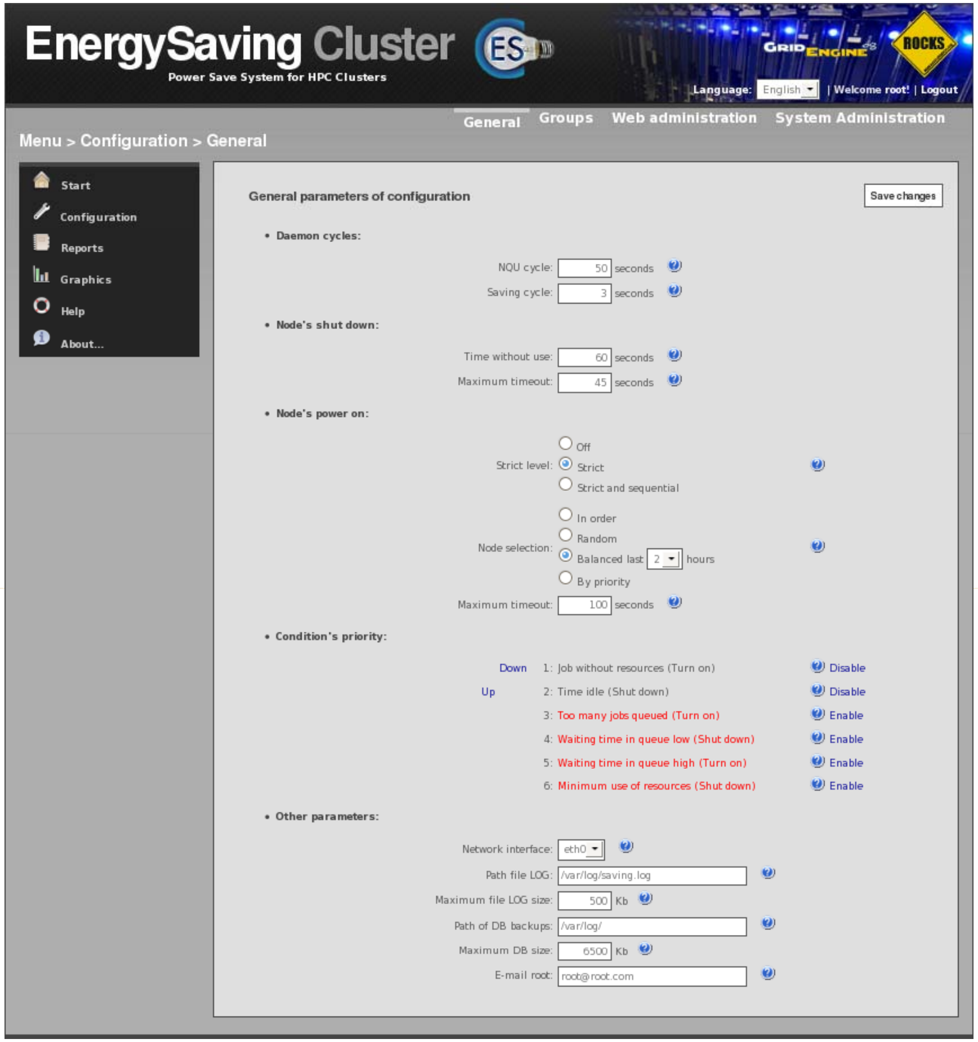
- No strict: Nodes are turned on to serve job request if there are not enough free slots on current active nodes
- Strict: Nodes are only turned on when the current active nodes do not provide enough slots (free or occupied) to serve requirements of the new job
- Strict and sequential: Nodes are only turned on to serve the job request when all current active nodes have their slots in free state

Node deactivation conditions

- The time that a node has been idle
- The average waiting time for user’s jobs is less than a threshold set by the administrator
- Current jobs can be served by a smaller number of active nodes

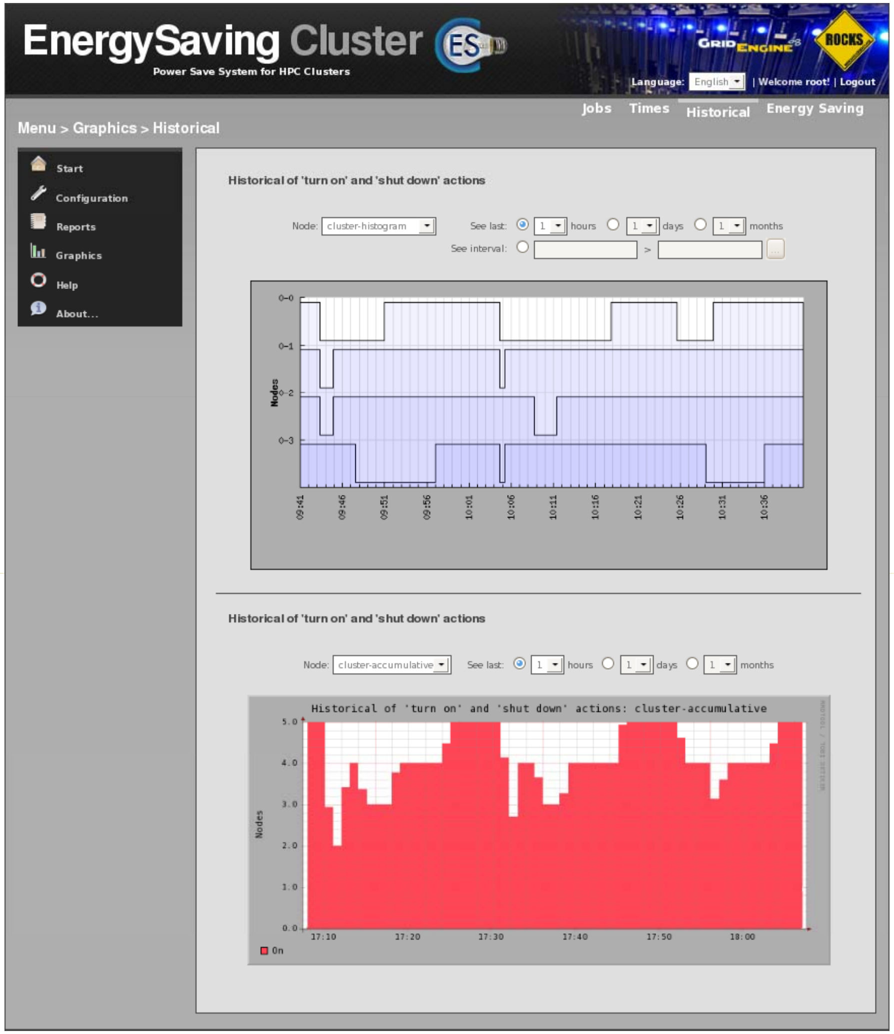
Website interface

Checking and modifing configuration parameters



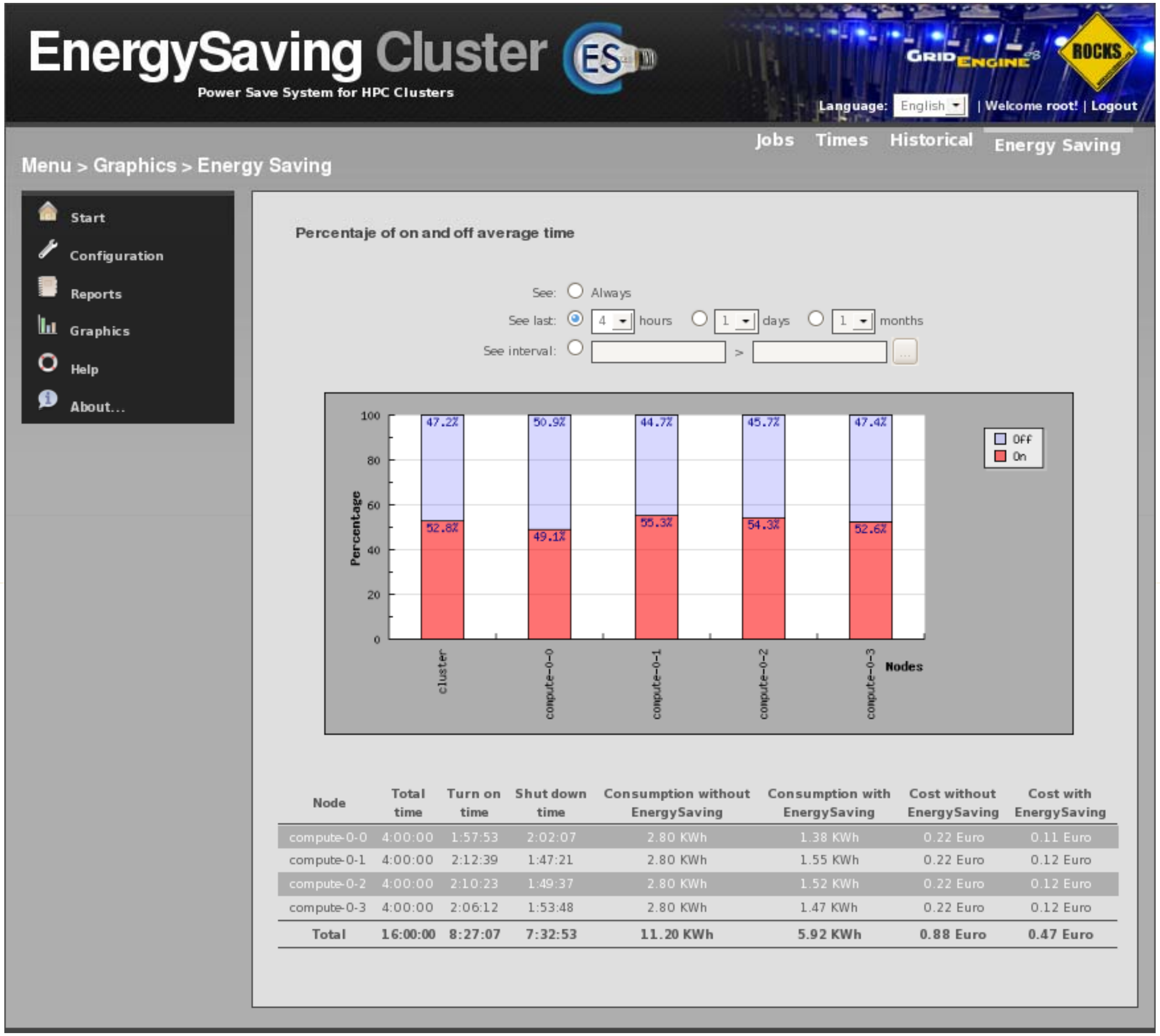
Website interface

Monitoring the operation of the cluster



Website interface

Monitoring the energy savings



Node	Total time	Turn on time	Shut down time	Consumption without EnergySaving	Consumption with EnergySaving	Cost without EnergySaving	Cost with EnergySaving
compute-0-0	4:00:00	1:57:53	2:02:07	2.80 KWh	1.38 KWh	0.22 Euro	0.11 Euro
compute-0-1	4:00:00	2:12:39	1:47:21	2.80 KWh	1.55 KWh	0.22 Euro	0.12 Euro
compute-0-2	4:00:00	2:10:23	1:49:37	2.80 KWh	1.52 KWh	0.22 Euro	0.12 Euro
compute-0-3	4:00:00	2:06:12	1:53:48	2.80 KWh	1.47 KWh	0.22 Euro	0.12 Euro
Total	**16:00:00**	**8:27:07**	**7:32:53**	**11.20 KWh**	**5.92 KWh**	**0.88 Euro**	**0.47 Euro**

Experimental Results

- To evaluate the benefits of the system we have developed a flexible simulator that provides information on the system and various platform configurations and under realistic workloads

EnergySaving-SIM

Experimental results

- We have configured the simulator to emulate the system of queues of the HPC computing service at the University Jaume I:
  - Front-end: HP Proliant DL360 G5 with 2 dual core Intel Xeon 5160 processors
  - Group 1: 26 nodes, Fujitsu Siemens RX200 with 2 Intel Xeon processors
  - Group 2: 27 nodes, HP Proliant DL360 G5 with 2 dual core Intel Xeon 5160 processors
  - Group 3: 11 nodes, HP Proliant BL460C with 2 quadcore Intel Xeon E5450 processors
  - Altix 3700 server with 48 Itanium2 processors

Experimental results

- The job benchmark was obtained from the real queue system logs of the computing facility at University Jaume I

- Composed by 10,415 jobs corresponding to the load submitted to the HPC during three months of 2009:
  - Number of processor required by the jobs: One processor (99.87%), 4 processors (0.12 %) and 8 processors (0.01%)
  - Jobs executed on: Group 1 (73,3%), group 2 (0%), group 3 (16.99%) and Altix server (9.7%)
  - The average execution time of the jobs is 1day, 2h, 53m

Experimental results

- We have evaluated the following policies:
  - No Policy (NP): Conventional cluster without energy saving
    - Nodes are permanently active
  - Policy 1 (P1):
    - Activation condition: job without resources
    - Deactivation condition: idle time of a node (60 sec.)
    - Node selection algorithm: ordered
    - Strict level: no strict
  - Policy 2 (P2): Same as P1, except strict level (strict)
  - Policy 3 (P3): Same as P1, except strict level (strict and sequential)

Experimental results

- Results are expressed by the following parameters:
  - Latency: Average time since jobs are submitted till their execution is completed (includes the time a job is enqueued as well as its execution time)
  - Power on time (%): Average fraction of the total time that the nodes remain turned on
  - Total time: Elapsed time since the first job is submitted till the last job completes its execution
  - Total consumption: In Mwatts-hour (we consider that a node consumes on average 250 Watts/hour)

Experimental results

- Results obtained with the simulator for different policies:

Policy	Latency	Power on time	Total time	Total consumption
NP	339 h, 44 m, 18 s	100.0%	4,022 h, 39 m, 50 s	65.37
P1	461 h, 54 m, 0 s	42.9%	4,022 h, 49 m, 15 s	29.51
P2	12,387 h, 56 m, 34 s	5.8%	29,962 h, 2 m, 41 s	46.50
P3	36,556 h, 28 m, 9 s	2.2%	86,712 h, 51 m, 31 s	85.73

Experimental results

- Results obtained with the simulator for different policies:

Policy	Latency	Power on time	Total time	Total consumption
NP	339 h, 44 m, 18 s	100.0%	4,022 h, 39 m, 50 s	65.37
P1	461 h, 54 m, 0 s	42.9%	4,022 h, 49 m, 15 s	29.51
P2	12,387 h, 56 m, 34 s	5.8%	29,962 h, 2 m, 41 s	46.50
P3	36,556 h, 28 m, 9 s	2.2%	86,712 h, 51 m, 31 s	85.73

- Conclusions of these results:
  - Policy P1 increases the job latency from NP, but the nodes are powered on only 42.9%

Experimental results

- Results obtained with the simulator for different policies:

Policy	Latency	Power on time	Total time	Total consumption
NP	339 h, 44 m, 18 s	100.0%	4,022 h, 39 m, 50 s	65.37
P1	461 h, 54 m, 0 s	42.9%	4,022 h, 49 m, 15 s	29.51
P2	12,387 h, 56 m, 34 s	5.8%	29,962 h, 2 m, 41 s	46.50
P3	36,556 h, 28 m, 9 s	2.2%	86,712 h, 51 m, 31 s	85.73

- Conclusions of these results:
  - Policy P2 produces worse results than P1
  - As most of the jobs of the benchmark require a single processor, policy P2 is not appropriate

Experimental results

- Results obtained with the simulator for different policies:

Policy	Latency	Power on time	Total time	Total consumption
NP	339 h, 44 m, 18 s	100.0%	4,022 h, 39 m, 50 s	65.37
P1	461 h, 54 m, 0 s	42.9%	4,022 h, 49 m, 15 s	29.51
P2	12,387 h, 56 m, 34 s	5.8%	29,962 h, 2 m, 41 s	46.50
P3	36,556 h, 28 m, 9 s	2.2%	86,712 h, 51 m, 31 s	85.73

- Conclusions of these results:
  - Policy P3 presents a long response time, for this particular benchmark is not appropriate
  - Policies P2 and P3 can deliver better best results in case the jobs request multiple processors

Experimental results

- More results for policy P1:

Measure	Total	Per node
Number of shutdowns	206	3.17
Maximum active nodes	37 of 65	-
Minimum active nodes	1 of 65	-
Active time	112,056 h, 24 m, 28 s	1,723 h, 56 m, 41 s
Inactive time	149,424 h, 46 m, 47 s	2,298 h, 50 m, 33 s
Active time with average of active intervals per node	25,462 h, 29 m, 0 s	391 h, 43 m, 49 s
Inactive time with average of inactive intervals per node	120,678 h, 57 m, 53 s	1,856 h, 35 m, 58 s

- Conclusions for policy P1 (in 4,000 hours):
  - A node was activated and deactivated slightly more than 3 times
  - Nodes are turned on basically 1,723 h or 42.9% of the time.

Experimental results

- More results for policy P1:

Measure	Total	Per node
Number of shutdowns	206	3.17
Maximum active nodes	37 of 65	-
Minimum active nodes	1 of 65	-
Active time	112,056 h, 24 m, 28 s	1,723 h, 56 m, 41 s
Inactive time	149,424 h, 46 m, 47 s	2,298 h, 50 m, 33 s
Active time with average of active intervals per node	25,462 h, 29 m, 0 s	391 h, 43 m, 49 s
Inactive time with average of inactive intervals per node	120,678 h, 57 m, 53 s	1,856 h, 35 m, 58 s

- Conclusions for policy P1:
  - Nodes were down a considerable time for this particular workload (1,856h). This value indicates that nodes have been deactivated for long periods of time » the decision of keeping them down is feasible

Experimental results

- More results for policy P1:

Measure	Total	Per node
Number of shutdowns	206	3.17
Maximum active nodes	37 of 65	-
Minimum active nodes	1 of 65	-
Active time	112,056 h, 24 m, 28 s	1,723 h, 56 m, 41 s
Inactive time	149,424 h, 46 m, 47 s	2,298 h, 50 m, 33 s
Active time with average of active intervals per node	25,462 h, 29 m, 0 s	391 h, 43 m, 49 s
Inactive time with average of inactive intervals per node	120,678 h, 57 m, 53 s	1,856 h, 35 m, 58 s

- Conclusions for policy P1:
  - For this particular workload is more convenient to turn off nodes than to keep them active because the time needed to reactivate a node is insignificant compared with the period of time they remain inactive

Summary and Conclusions

- EnergySaving Roll may yield substantial energy savings by turning on only those nodes that are actually needed at a given time during the execution of jobs
- This module is flexible:
  - There are three conditions to turn on the nodes and three conditions to turn off the nodes
  - There are also options to select candidate nodes to be powered on

Summary and Conclusions

- Choosing the best policy depends on the type of the jobs that are submitted to the system and the configuration of the cluster
- This module is currently in operation in the HPC clusters of the High Perfomance Computing & Architectures research group of University Jaume I

EnergySaving Cluster Roll: Power Saving System for Clusters

HPC A

High Performance Computing and Architectures

QUESTIONS?