ACM/IFIP/USENIX
12th International Middleware Conference

2nd International Workshop on
Green Computing Middleware

A Flexible Simulator to Evaluate a Power Saving System for HPC Clusters

Manuel F. Dolz, Juan C. Fernández, Sergio Iserte,
Rafael Mayo, Enrique S. Quintana

Universitat Jaume I

December 12, 2011, Lisbon (Portugal)

Motivation

High Performance Computing Clusters:

- 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

Motivation

High Performance Computing Clusters:

- 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

Motivation

High Performance Computing Clusters:

- 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

Simulator

- Evaluate the performance of the ESC middleware within different kind of workloads by using our the ESC simulator.

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

Simulator

- Evaluate the performance of the ESC middleware within different kind of workloads by using our the ESC simulator.

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

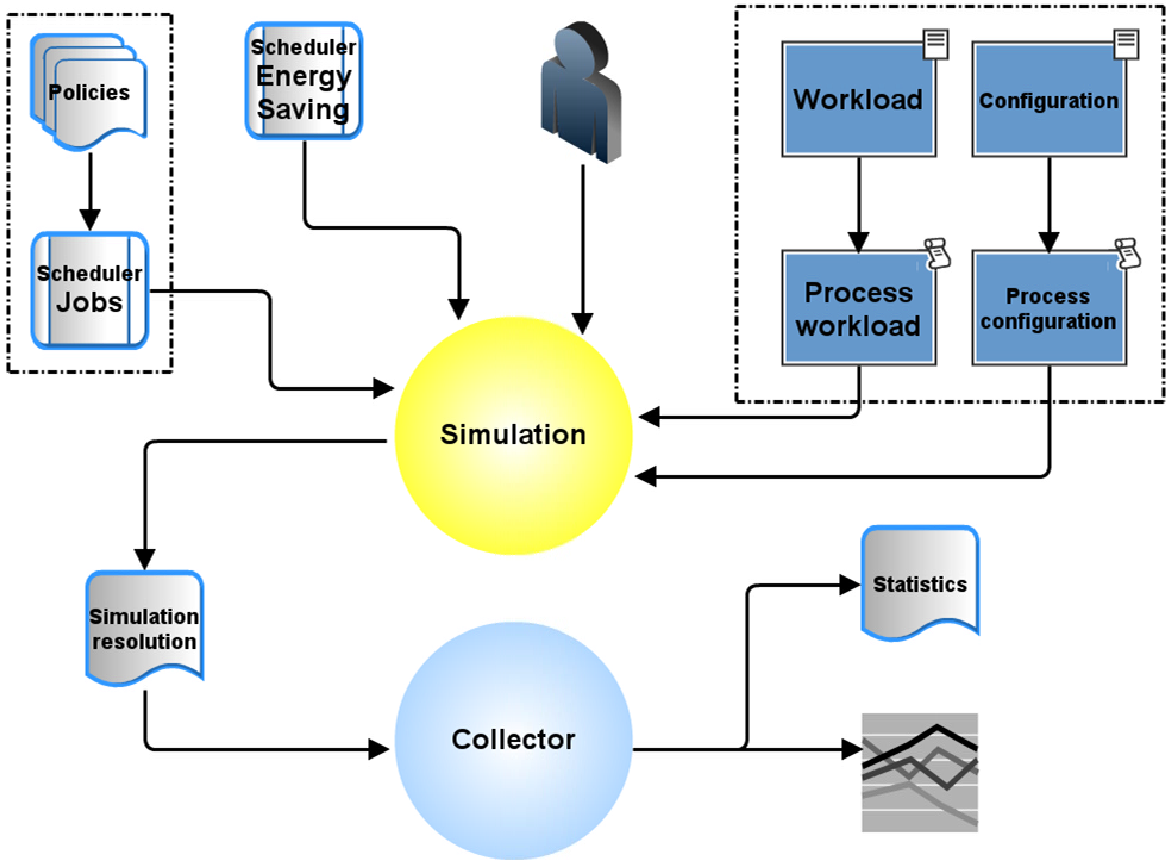
Find a solution!

EnergySaving Cluster

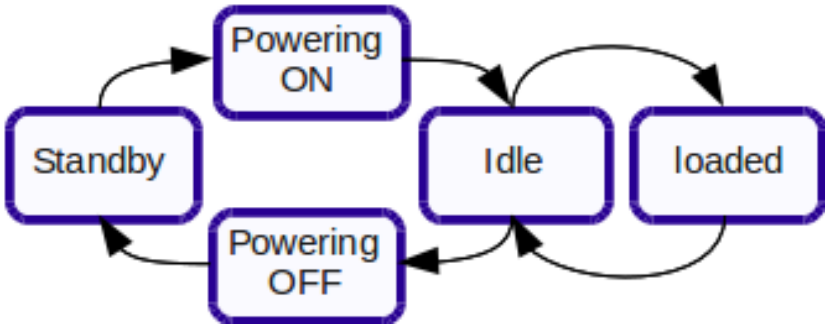
Simulator

- Evaluate the performance of the ESC middleware within different kind of workloads by using our the ESC simulator.

General schema



Model of the node energy consumption



Node states:

- *Standby*: Node still consumes a residual energy.
- *Powering on*: Consumption and time needed to power on
- *Powering off* : Consumption and time needed to power off
- *Idle*: Node is waiting for jobs, but it still consumes.
- *Loaded*: Node is executing a job, it employs the 100 % of computational power.

Workload file loader

Standart workload format:

- *First 4 lines*: global aspects, number of jobs, start/finish dates, nodes, processors, queues.
- *Remaining lines*: jobs running and informacion about the jobs: identifier, submission time, user, queue, used processors, duration.

Loader module:

1. Receives the workload file with the *Standard Workload Format*.
2. Builds a B-Tree structure with information of all jobs in chronological order.

The B-Tree contain events of type *a new job is submitted to the system*.

System configuration

The module uses a standard configuration file with the following information:

- Users of the system, Groups they belong to, and configuration queues for groups
- Nodes in the cluster and parameters of each froup of nodes in cluster
- General operations of the simulator:
  - Parameters defining the policies applied to job exections.
  - Energy saving policies.
  - Duration of events ocurring during simulations.

Queueing system/Energy Saving scheduler

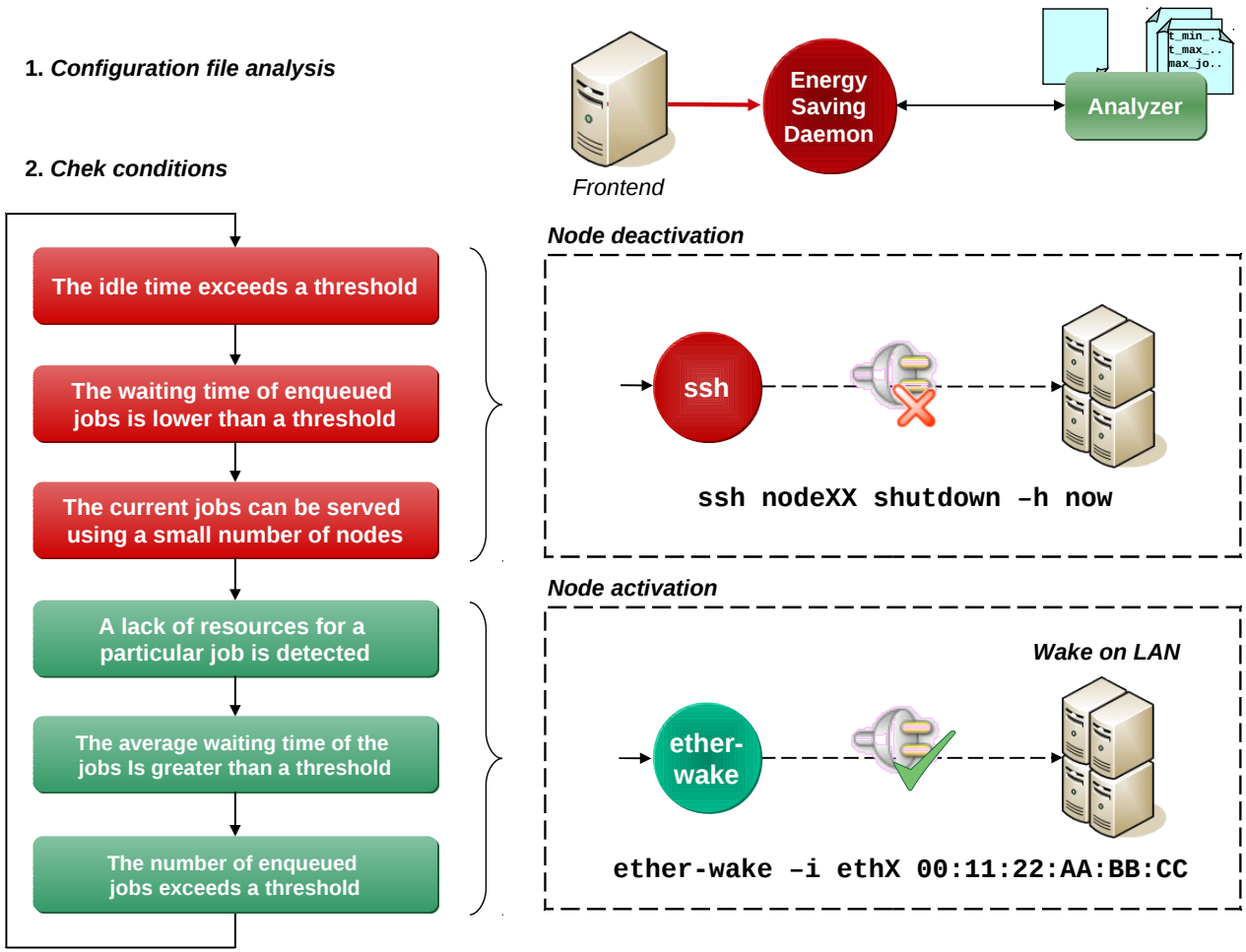
Queueing system scheduler: The simulatior employs a scheduler similar to the Sun Grid Engine:

- Is encharged to handle the execution of jobs.
- For each queue, the FIFO policy is applied.
- Due the modular structure of the simulator, adding new policies is easy.

Energy Saving scheduler:

- The simulator employs the Energy Saving system adapted to employ the interfaces provided by the queuing system scheduler module.
- This module provides the activation/deactivation policies provided by the Energy Saving Cluster tool.

Activation/deactivation actions



Simulation module

- How? The simulation looks up the B-Tree for the next event in the time, analizes and process it.
- During simulation module inserts events in the B-Tree.
- There are 11 events that may appear during execution of the simulation:
  - *Node turn-on starts / ends*
  - *Node turn-off starts / ends*
  - *New job is submitted to the system*
  - *Energy saving scheduler starts / ends*
  - *Queue system scheduler starts / ends*
  - *Job execution starts / ends*

Simulation and statistics module

- For each simulation a trace file is produced
- For each event the module saves a line
  - Timestamp
  - Elements involved
  - Results of any decisions taken.

$$\text{Simulation} \rightarrow \text{Trace file} \rightarrow \text{Statistics} \rightarrow \text{Results}$$

- Statistics:
  - Maximum number of active nodes
  - Number of shitdowns during the smulation period
  - Average queue/user waiting/execution time
  - Average node active/execution/idle time
- Finally, the statistics module elaborates graphs and tables to ease the visualization of results.

Web interface

Provides a full control of the simulator:

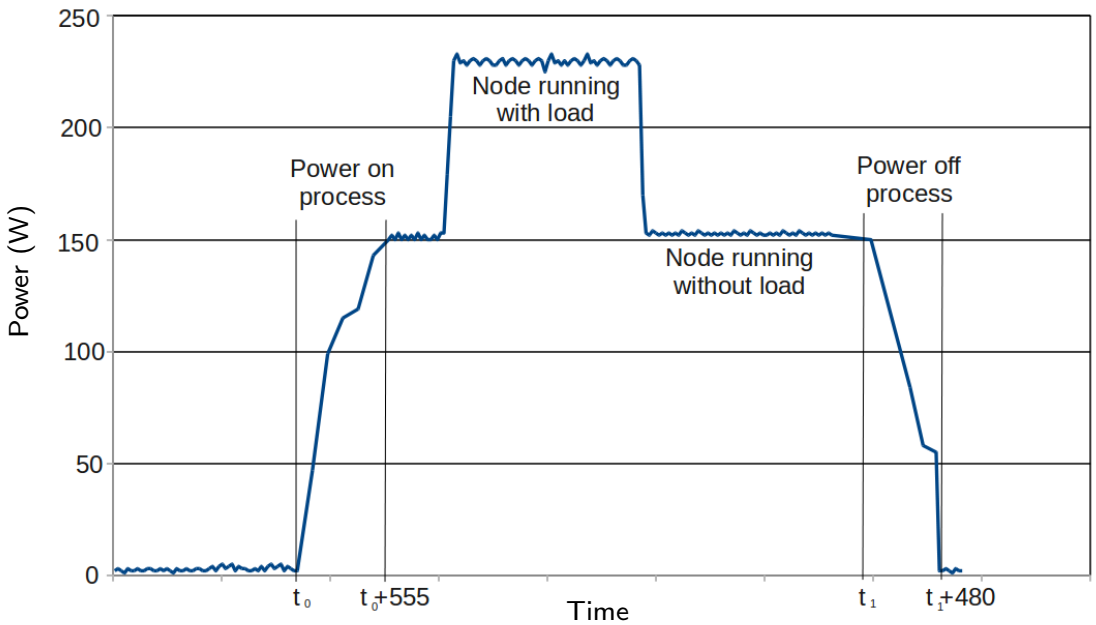
- Set parameters of simulation
- Import configuration files to apply them to simulations
- Import workload files for simulation
- Run and check simulations
- Manage simulations (abort, clear results, view traces, etc.)
- View results (graphs and tables).

Configuration parameters for evaluation

- We have used two workloads form the Parallel Workload Archive:

Workload	Number of jobs	Platform
OSC Linux Cluster	80,714	16-node cluster
NASA Ames iPSC/860	42,264	57-node cluster

- To obtain power consumption statistics of simulations, we have supposed that clusters are composed of Intel Xeon E5230 with 16 GB of RAM:



Benchmark and 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, 14 m, 7 s	40.42 MWh
OSC with ESC	997 d, 0 h, 25 m, 37 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	4.67 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, 14 m, 7 s	40.42 MWh
OSC with ESC	997 d, 0 h, 25 m, 37 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	4.67 MWh

Conclusions of 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, 14 m, 7 s	40.42 MWh
OSC with ESC	997 d, 0 h, 25 m, 37 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	4.67 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, 14 m, 7 s	40.42 MWh
OSC with ESC	997 d, 0 h, 25 m, 37 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	4.67 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 30 %.

Experimental results

- Detailed results for the OSC workload with ESC:

Measure	**Total**	**Per node**
Number of shutdowns	817	14.33
Maximum active nodes	33 of 57	-
Active time	24,009d, 1h, 33m, 53s	421d, 54h, 4m, 48s
Inactive time	32,820d, 3h, 14m, 7s	575d, 19h, 20m, 48s
Active time with average of active intervals per node	31d, 13h, 11m, 4s	13h, 17m, 2s
Inactive time with average of inactive intervals per node	43d, 3h, 5m, 20s	18h, 9m, 34s

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 active, and those that are not needed to remain powered off
- 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
- Highly efficient: simulation of months in a real cluster can be reduced to minutes which accelerates the analysis of the data
- Modular design: enhaces its flexibility, so that, adding new features is relatively easy.

Thanks for your attention!

Questions?