Monitoring GreenClouds

Evaluating the trade-off between Performance and Energy Consumption in DAS-4

System and Networking Engineering

Renato Fontana | Katerina Mparmpopoulou

Monitoring GreenClouds

Presentation Flow

- Green concepts
- Project objective
- Experimental environment
- Metrics and Workload
- Experiment Results
- Conclusions
- Future work

Green Concepts

- What does it mean to be green?
- Refers to environmentally sustainable
- Energy becomes a key challenge in large-scale distributed systems
- IT requires more and more power

Known techniques

- Event-monitoring counters
  - Deducing energy consumption
- On/off algorithms
  - Switch on/off nodes in long idle state
- Load balancing
  - Distribute workload amongst multiple nodes
- Task scheduling
  - Slowdown factors
- Thermal management
  - Monitoring heat generation

Research Question

- How to evaluate the trade-off between energy and performance in DAS-4?
- How to correlate performance and energy consumption in Cloud Computing Systems?

Approach

- Compare workload with power-monitoring tools
- Estimate energy consumption in nodes
- Correlate main components (CPU, memory)
- CPU load and energy consumed

Monitoring GreenClouds

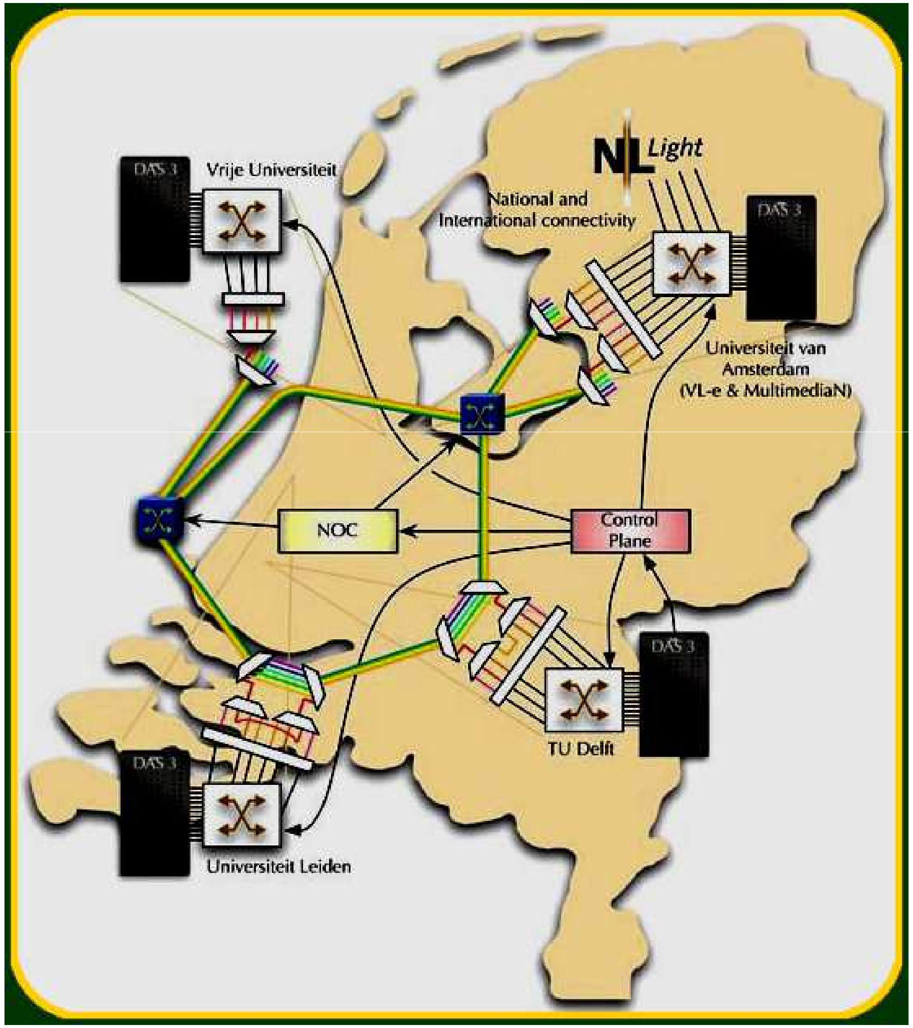
Experimental environment

DAS-4 (The Distributed ASCI Supercomputer 4)

- Six-cluster wide-area distributed system
  - UvA and VU nodes (PDU enable)
- Grid Computing
  - DAS-4 mainly composed by cluster nodes
- Cloud Computing
  - OpenNebula

Monitoring GreenClouds

Topology

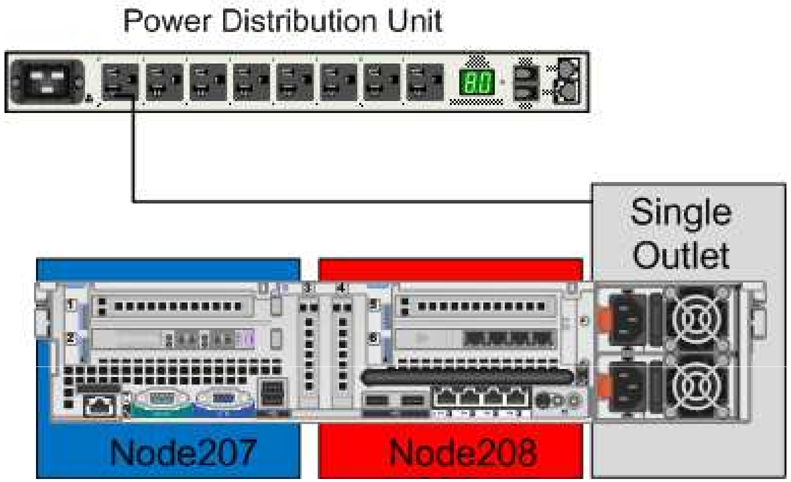


Cluster	Head node	Compute nodes
VU	fs0.das4.cs.vu.nl	001-075
LU	fs1.das4.liacs.nl	101-116
UvA	fs2.das4.science.uva.nl	201-218
TUD	fs3.das4.tudelft.nl	301-332
UvA-MN	fs4.das4.science.uva.nl	401-436
ASTRON	fs5.das4.astron.nl	501-523

Current Setup

Cluster environment

- 2U Twin Server
- Single outlet for the entire server

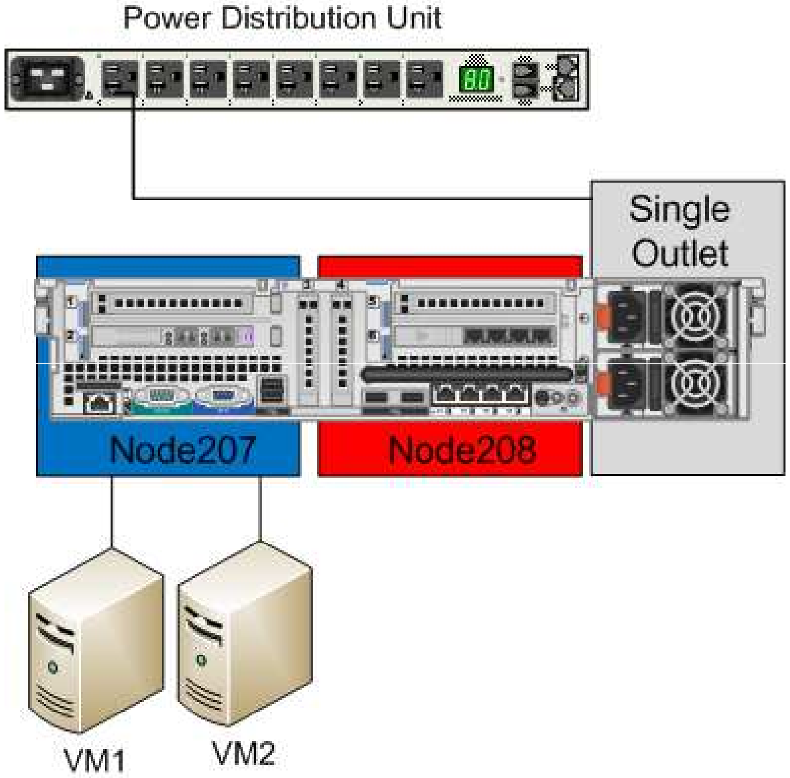


Rear View

Environment Approximation

Cloud environment

- Single node with two VMs
- Only one energy source for both VMs
- Why?
  - No monitoring tools;
  - Concurrent resource share;



Monitoring GreenClouds

Metrics and Workload

Workload measurement

- Bright Cluster Manager

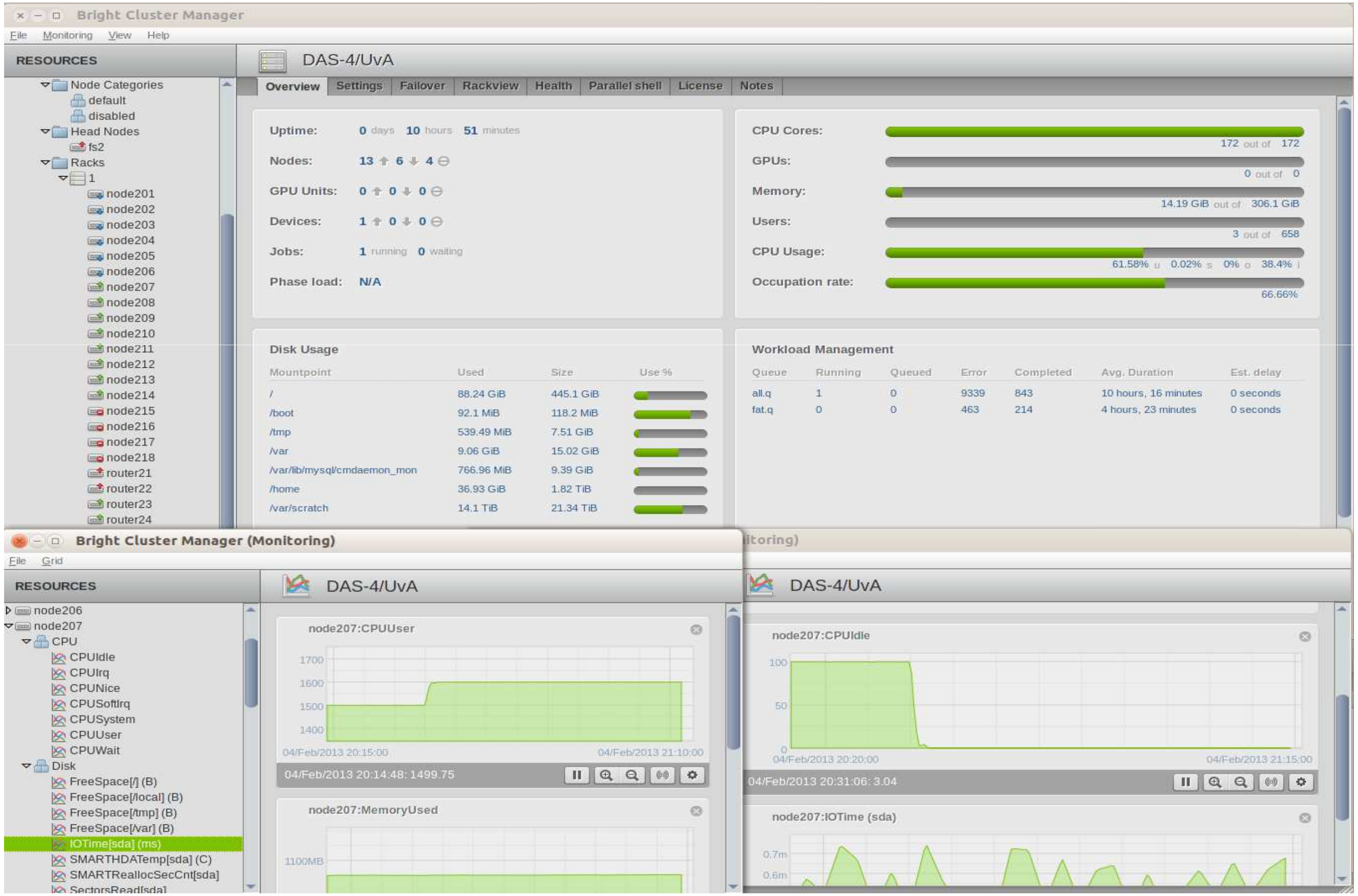
Power management

- Racktivity PDUs

Correlation of the two systems

- Workload and energy

Bright Cluster Manager



Uptime: 0 days 10 hours 51 minutes

Nodes: 13 ↑ 6 ↓ 4 ⊖

GPU Units: 0 ↑ 0 ↓ 0 ⊖

Devices: 1 ↑ 0 ↓ 0 ⊖

Jobs: 1 running 0 waiting

Phase load: N/A

CPU Cores: 172 out of 172

GPUs: 0 out of 0

Memory: 14.19 GiB out of 306.1 GiB

Users: 3 out of 658

CPU Usage: 61.58% u 0.02% s 0% o 38.4% i

Occupation rate: 66.66%

Disk Usage

Mountpoint	Used	Size	Use %
/	88.24 GiB	445.1 GiB	
/boot	92.1 MiB	118.2 MiB	
/tmp	539.49 MiB	7.51 GiB	
/var	9.06 GiB	15.02 GiB	
/var/lib/mysql/cmdaemon_mon	766.96 MiB	9.39 GiB	
/home	36.93 GiB	1.82 TiB	
/var/scratch	14.1 TiB	21.34 TiB	

Workload Management

Queue	Running	Queued	Error	Completed	Avg. Duration	Est. delay
all.q	1	0	9339	843	10 hours, 16 minutes	0 seconds
fat.q	0	0	463	214	4 hours, 23 minutes	0 seconds

Metrics

Metric	Extraction Method	Source
Execution time	As reported by the Job	Job
Power Consumption	Python Script	PDU
Energy Consumption	Python Script	PDU
CPU Load	Python script	Bright Cluster Manager

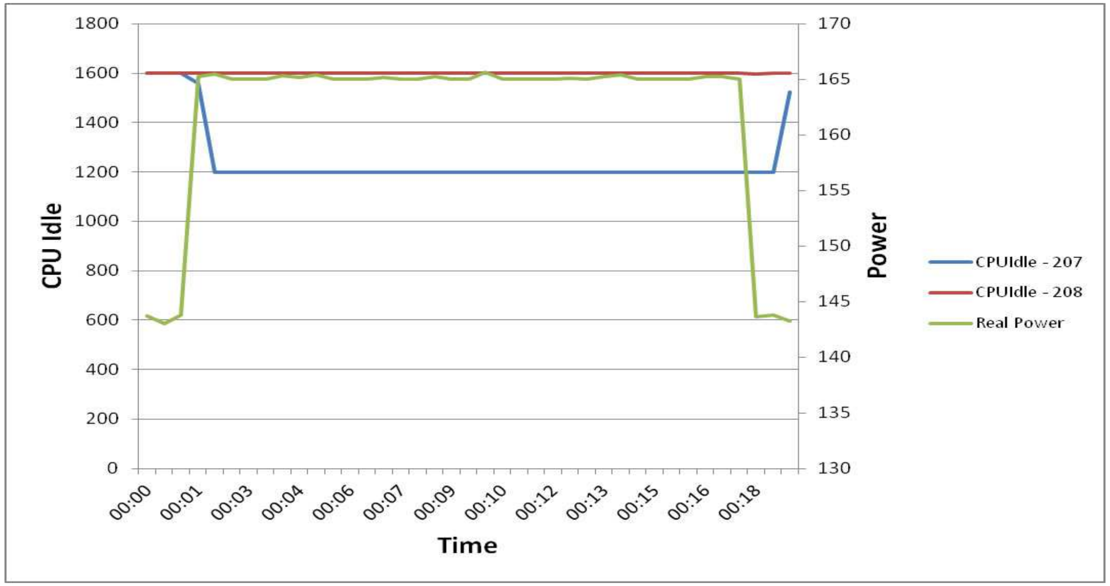
Monitoring GreenClouds

Linpack Vs Polyphase Filter

- Linpack lacks the configuration option to control the amount of resources that it uses
- Polyphase filter is configurable, as regards the number of its runs and the used threads
- We define two different jobs; job1 and job2, so that job1 causes the double workload of job2
- We treat every single job as a unit and measure the power produced by each of them under various rates of CPU utilization

Monitoring GreenClouds

Polyphase Filter – 25% workload

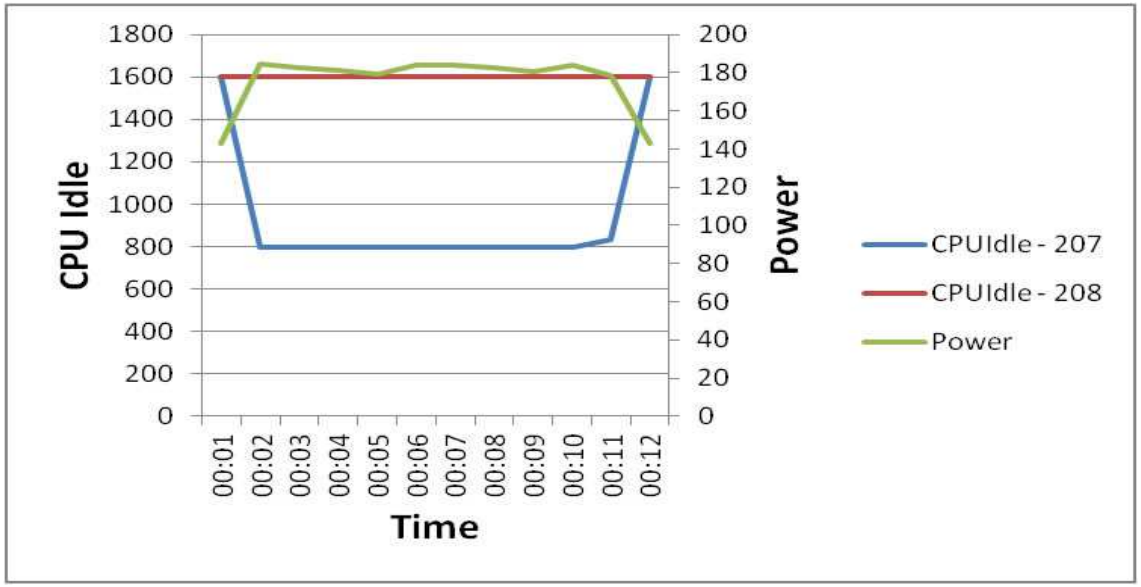


- job 1 is running on node-207 and the adjacent node-208 is idle

CPU Load Node-207	CPU Load Node-208	Peak of Power Consumption	Max Execution Time
25%	0%	165,4 W	1028 sec

Monitoring GreenClouds

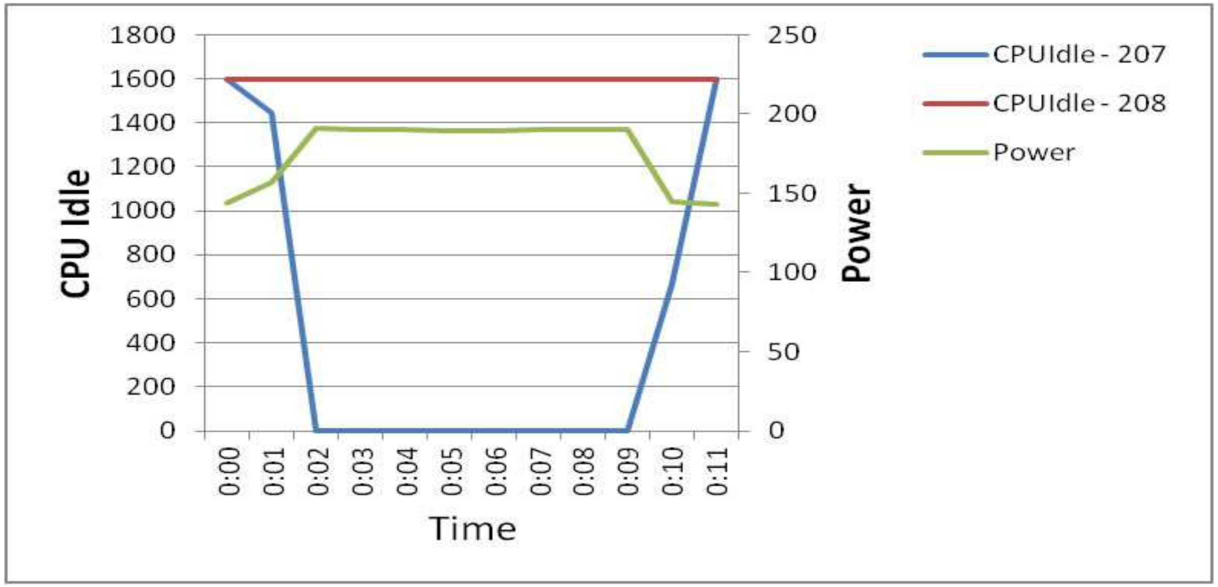
Polyphase Filter – 50% workload



job 1 is running on node-207 and the adjacent node-208 is idle

CPU Load Node-207	CPU Load Node-208	Peak of Power Consumption	Max Execution Time
50%	0%	184 W	587,6 sec

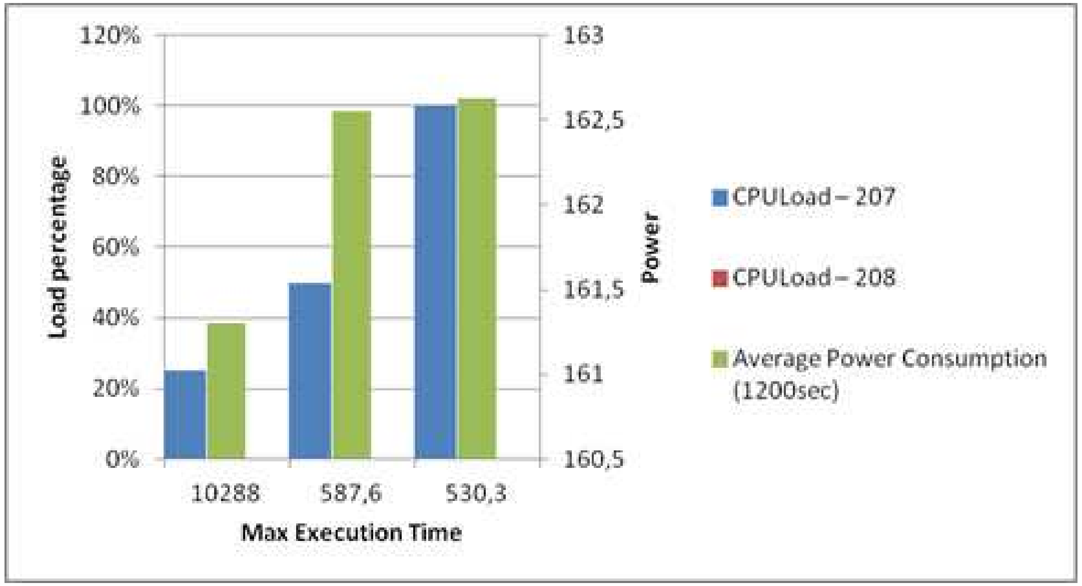
Polyphase Filter – 100% workload



job 1 is running on node-207 and the adjacent node-208 is idle

CPU Load Node-207	CPU Load Node-208	Peak of Power Consumption	Max Execution Time
100%	0%	190 W	530,3 sec

Results evaluation

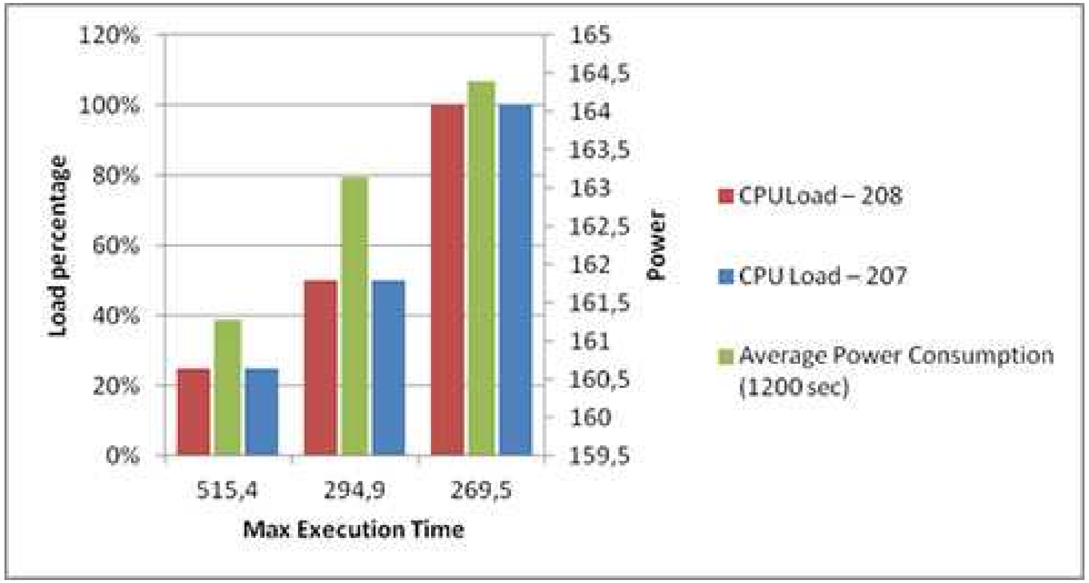


To evaluate the trade-off between power consumption and performance for all the above cases, we built a **coupled in time** environment of 1200 sec

CPU Load Node-207	CPU Load Node-208	Average Power Consumption In time interval equal to 1200 sec	Max Execution Time
25%	0%	161,30 W	1028 sec
50%	0%	162,54 W	587,6 sec
100%	0%	162,62 W	530,3 sec

Monitoring GreenClouds

Results evaluation



Finally in a short time interval, approximately equal to the longer execution time, **gains in power saving are almost negligible**.

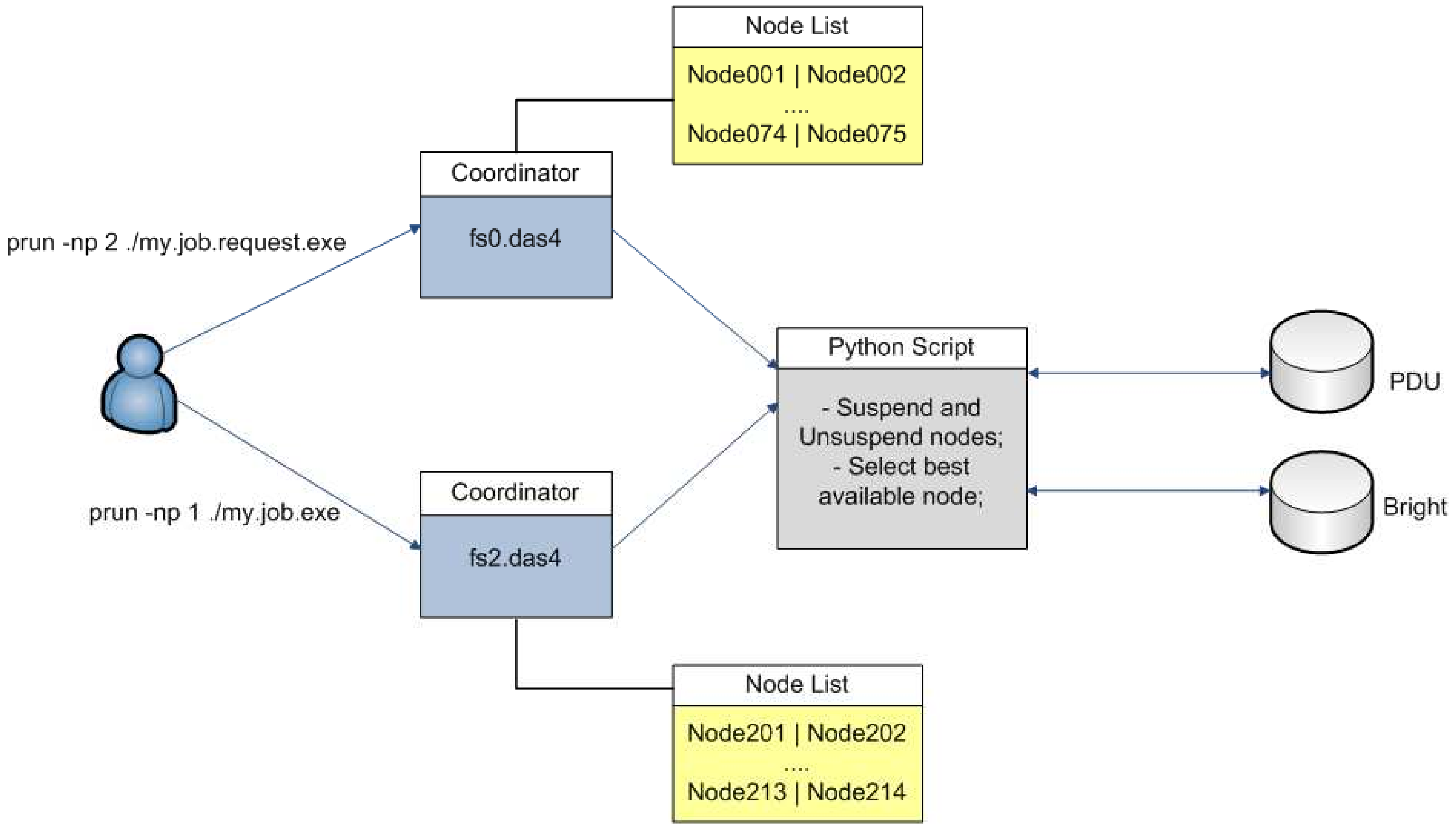
job2 = ½ job1

CPU Load Node-207	CPU Load Node-208	Average Power Consumption In time interval equal to 1200 sec	Max Execution Time
25%	25%	161,27 W	515,4 sec
50%	50%	163,14 W	294.9 sec
100%	100%	164,39 W	269,5 sec

Conclusions

- Definite execution time job
  - Better performance using roughly the same amout of power
  - Grant execution in available nodes which share the same physical server
- In the current cluster implementation, it is impossible to execute more then one job at a time
  - Queue system

Future work



Node List
Node001 | Node002
....
Node074 | Node075
Coordinator
fs0.das4
prun -np 2 ./my.job.request.exe
Python Script
- Suspend and Unsuspend nodes;
- Select best available node;
PDU
Bright
prun -np 1 ./my.job.exe
Coordinator
fs2.das4
Node List
Node201 | Node202
....
Node213 | Node214

Questions?