Energy-Efficiency in Computer Grids

Anne-Cécile Orgerie

ENS de LYON, FRANCE

annececile.orgerie@ens-lyon.fr

www.cost804.org

Action IC0804

9th March 2011, COST Training School, Poiana Brasov, Romania

ENS DE LYON

ENS DE LYON

Where do I come from?

- ENS de Lyon http://www.ens-lyon.eu
- More than 2 000 students (400 doctoral students)
- 350 researchers and 270 professors
- 31 nationalities represented on campus
- 245 international partnerships

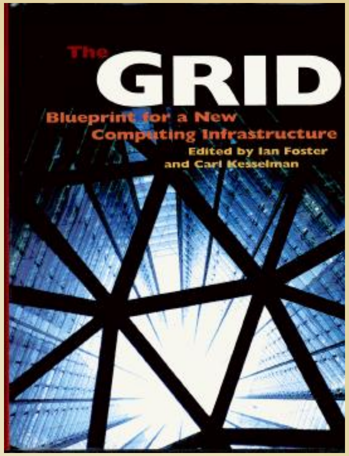
- LIP laboratory (47 permanent faculties and researchers and 48 PhD students)
- RESO team working on networks and distributed systems

Outline

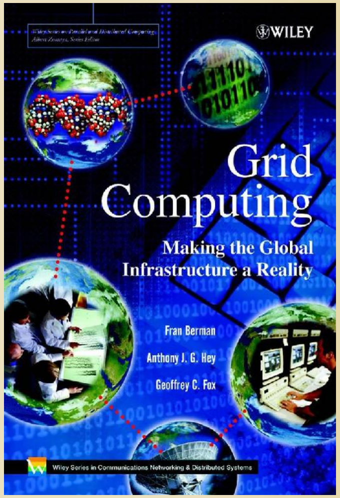
- Definitions and examples
- Issues
- Energy consumption reductions
- Grid usage
- Energy cost of Grid servers
- EARI
- Conclusions

Definitions and Examples

Grid Computing



- Born in the early 1990s
- "A service for sharing computer power and data storage capacity over the Internet."
- "Computing resources are not administered centrally.
- Open standards are used.
- Nontrivial quality of service is achieved."



GridCafé (CERN) http://www.gridcafe.org

"*What is the Grid? A Three Point Checklist*" Ian Foster

Grid examples

- Desktop Grids
- Volunteer Grids
- Enterprise Grids
- E-science Grids
- Production Grids
- Sensor Grids

Grid Architecture

- **Network layer**: connects the resources
- **Resource layer**: computers, storage systems, sensors, … often heterogeneous and geographically distributed
- **Middleware layer**: provides the tools to use the resources.
- **Application layer**: where the users can interact with the Grid; provides various services

Grid Middleware

- Batch scheduler
- Data access
- Remote job initiation
- Resource discovery
- Resource co-allocation
- Location management

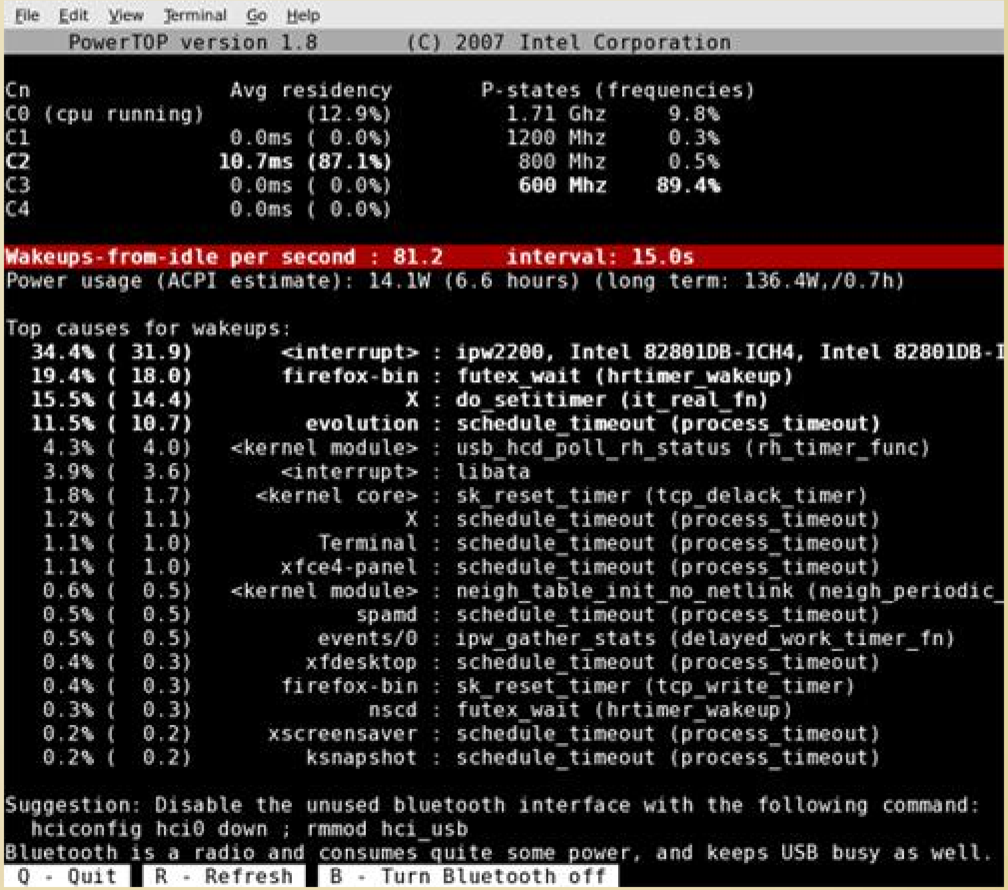
Energy Issues

How to measure energy consumption in Grids ?

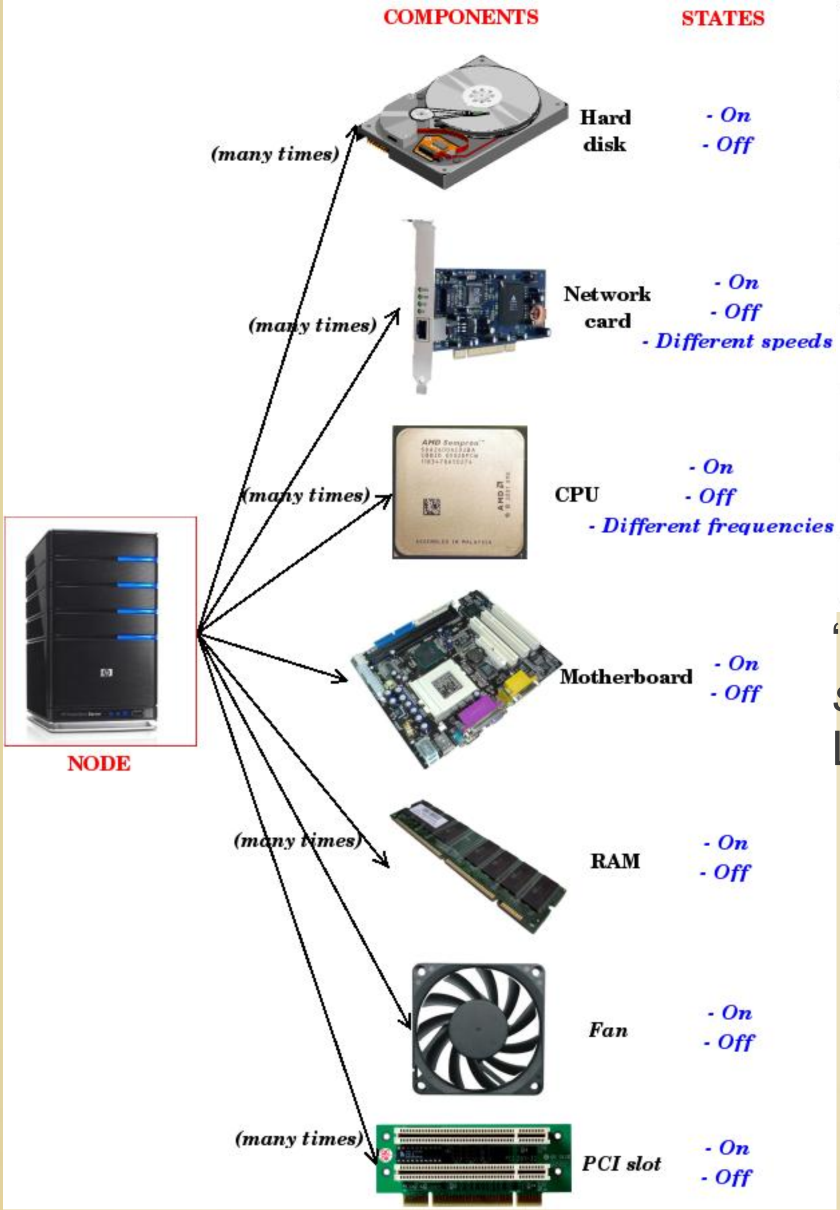
- Power usage per device, per process, per service, per rack?
- New tools: **powertop**
- Event counters
- Sensors

http://lesswatts.org/projects/powertop/

"*Balancing power consumption in multiprocessor systems*", A. Merkel and F. Bellosa, SIGOPS Oper. Syst. Rev., 2006.

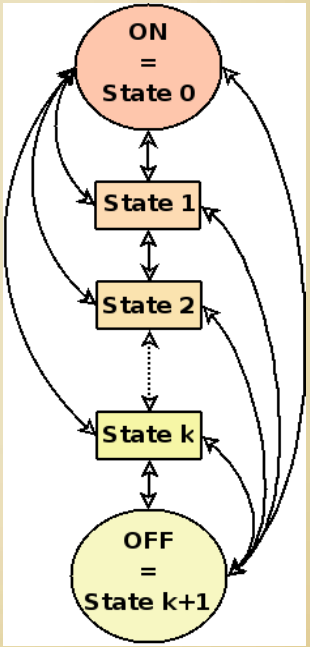


Understanding nodes



Component	Peak Power	Count	Total
CPU [16]	40 W	2	80 W
Memory [18]	9 W	4	36 W
Disk [24]	12 W	1	12 W
PCI slots [22]	25 W	2	50 W
Motherboard	25W	1	25 W
Fan	10 W	1	10 W
System Total			213 W

"*Power Provisioning for a Warehouse-sized Computer* ", X. Fan, W. Weber and L. Barroso, ISCA, 2007.



Issues

- Bad usage of machines in companies:

  - CPU: 4 004 h of running per year (17.8 h per working day)

  - Screen: 2 510 h per year (11.2 h per working day)

- But users effectively use their machines 686 h per year (3 h per working day)
- Computers are switch on ¾ time for nothing!

Energy consumption: some figures

- Energy consumption of data centers worldwide doubled between 2000 and 2006.
- Incremental US demand for data center energy between 2008 and 2010 is the equivalent of 10 new power plants.
- Power loss of 0.01% per meter of cable above 6 meters.

EcoInfo

"*Revolutionizing Data Centre Energy Efficiency*"
McKinsey & Company, April 2008.

How to measure energy efficiency?

- Benchmarks and metrics
- PUE: Power usage effectiveness

$$PUE = \frac{Total\ Facility\ Power}{IT\ Equipment\ Power}$$

- DCIE: Data Center Infrastructure Efficiency

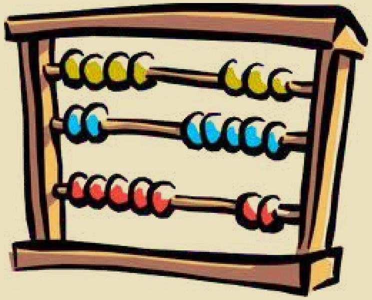
$$DCIE = \frac{1}{PUE} = \frac{IT\ Equipment\ Power}{Total\ Facility\ Power} \times 100$$

the green grid

ENERGY STAR

“*Green Grid Data Center Power Efficiency Metrics: PUE and DCIE*”, Green Grid White Paper, 2008.
“*Models and Metrics to Enable Energy-Efficiency Optimizations*”, S. Rivoire et al., IEEE Computer, 2007.

Techniques to reduce energy consumption in data centers and grids

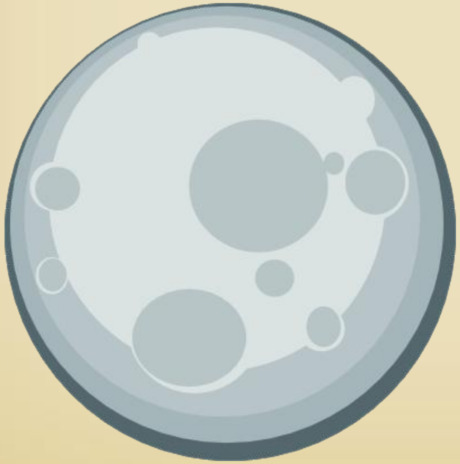
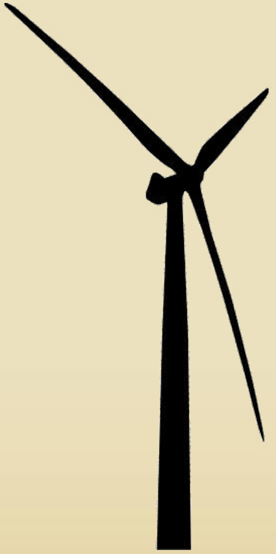


Different levels

- Grid level
- Data center level
- Node level
- Component level

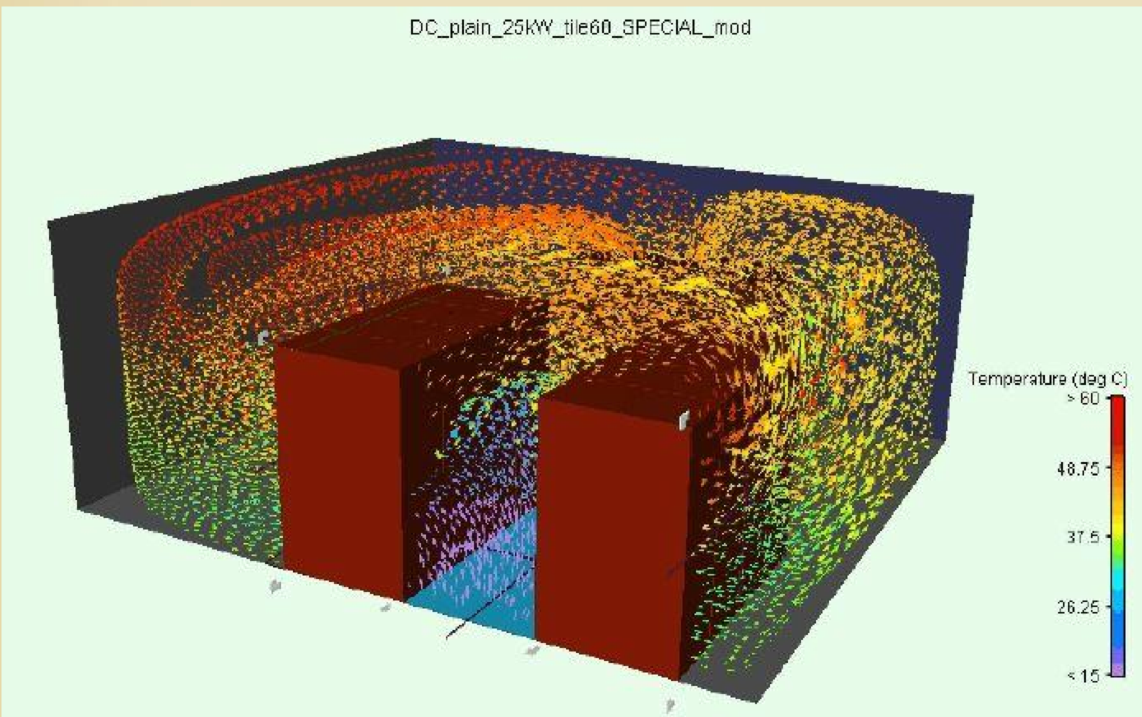
Follow-the-* approaches

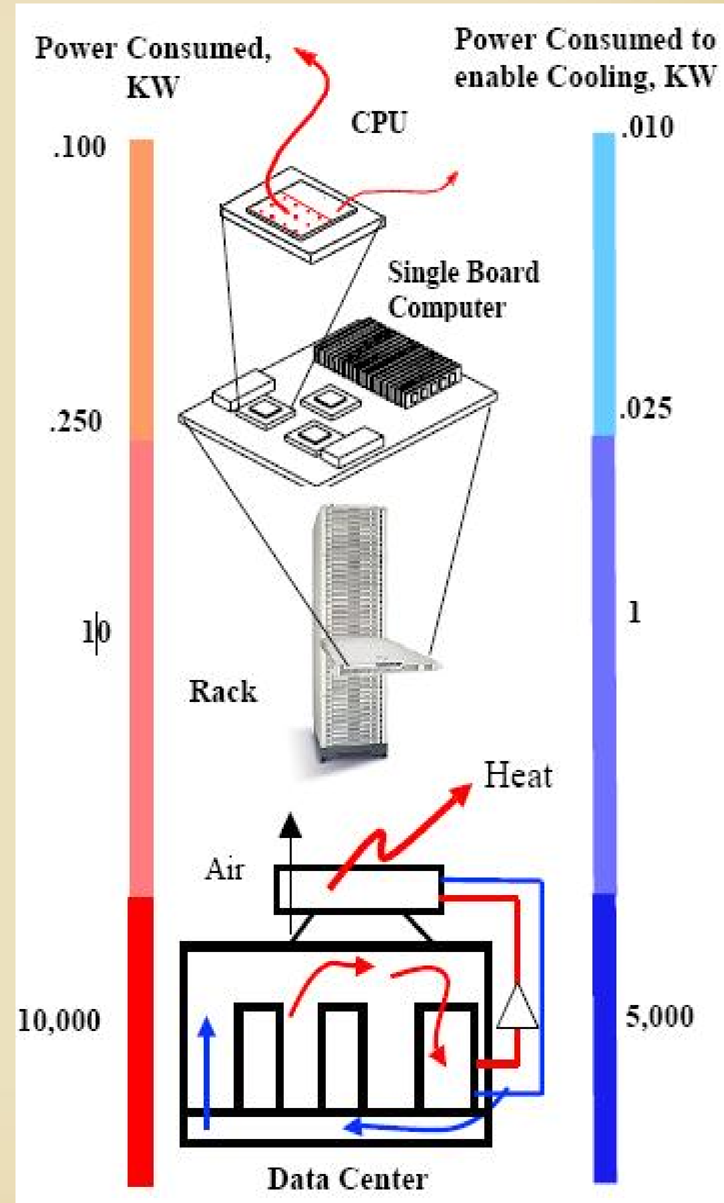
- Follow the moon: free cooling with air from outside during cool days, and on hot weather days, computing load is shifted to other data centers
- Follow the sun/wind: use renewable energy sources


Thermal management

- Cooling costs a lot, so reduce heat production to reduce energy consumption





"*Energy Aware Grid: Global Workload Placement based on Energy Efficiency* ", HP Technical Report, 2002.

"*Power Provisioning for a Warehouse-sized Computer* ", X. Fan, W. Weber and L. Barroso, ISCA, 2007.

On/off techniques

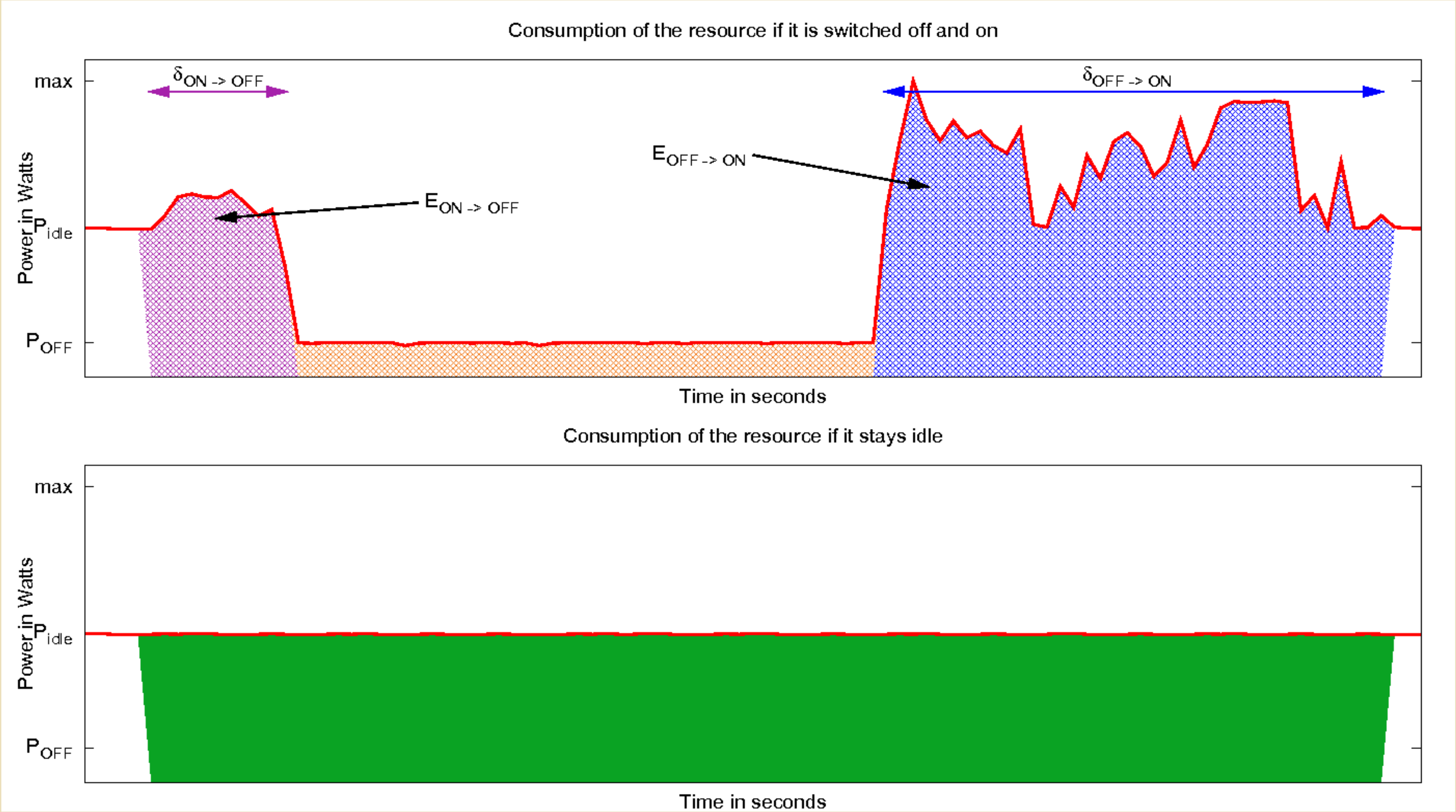
- Switch off unused resources
- Minimize set of active resources, adapt to the load

Issues:

- does it reduce the life time of resources?
- does it lower the reactivity of the platform?
- does the middleware consider the resources as dead?

“*Managing Energy and Server Resources in Hosting Centers* ”, J. Chase et al., SOPS, 2001.
“*Balance of power: Dynamic thermal management for Internet data centers*”, R. Sharma et al., IEEEInternet Computing, 2005.

When can we switch off ?



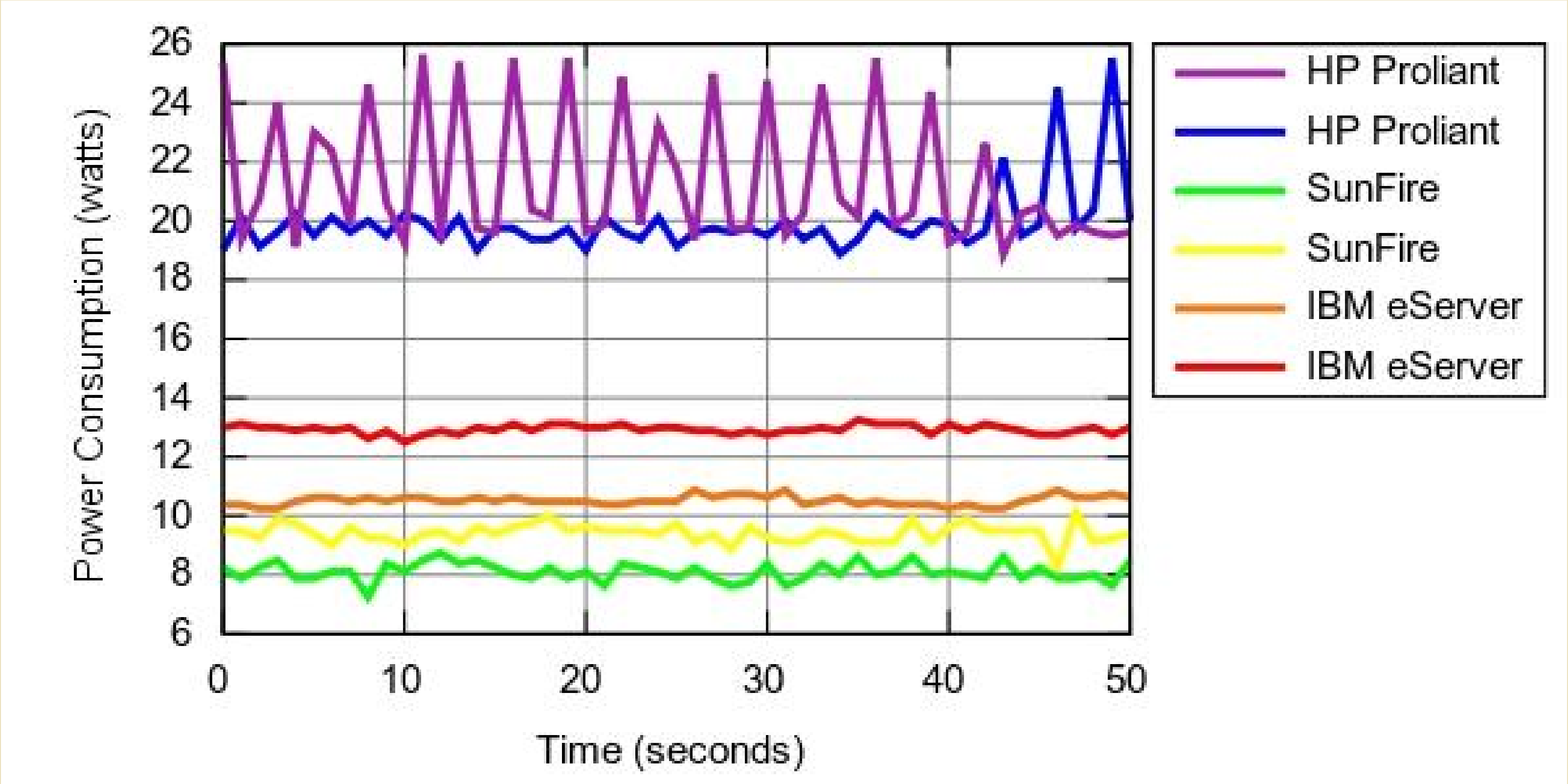
$$T_s = \frac{E_s - P_{OFF}(\delta_{ON\to OFF} + \delta_{OFF\to ON}) + E_{ON\to OFF} + E_{OFF\to ON}}{P_I - P_{OFF}} + T_r$$

"*Save Watts in your Grid: Green Strategies for Energy-Aware Framework in Large-Scale Distributed Systems* ", A.-C. Orgerie et al., ICPADS, 2008.

On/off tools

- Suspend to disk (hibernation)
- Suspend to RAM (standby or sleep)
- Wake on LAN
- IPMI (Intelligent Platform Management Interface)
- Presence proxies
- Smart PDU (Power Distribution Unit)

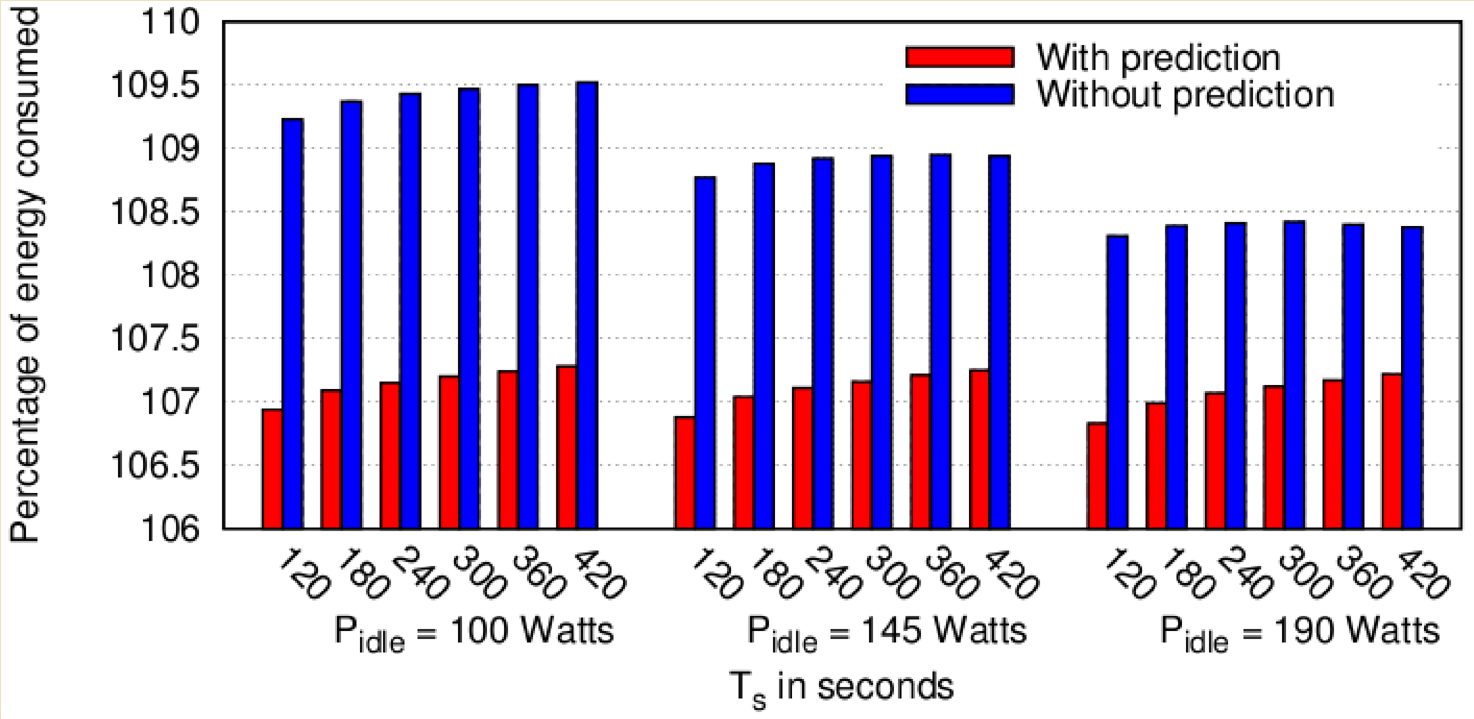
Does a switched off node consume energy?



"*Demystifying Energy Consumption in Grids and Clouds*
", A.-C. Orgerie et al., WIPGC, 2010.

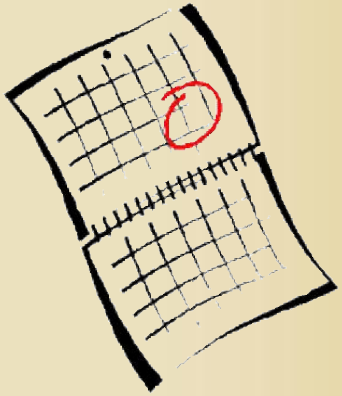
Is intensive and systematic on/off the perfect approach for saving energy?

Must be done cleverly and coupled with usage prediction



"*Chasing Gaps between Bursts: Towards Energy-Efficient Large-Scale Experimental Grids* ", A.-C. Orgerie et al., PDCAT, 2008.

Energy-Aware Task Scheduling



- Exploiting performances and energy consumptions of heterogeneous nodes
- Use first the most energy-efficient nodes

Issues:

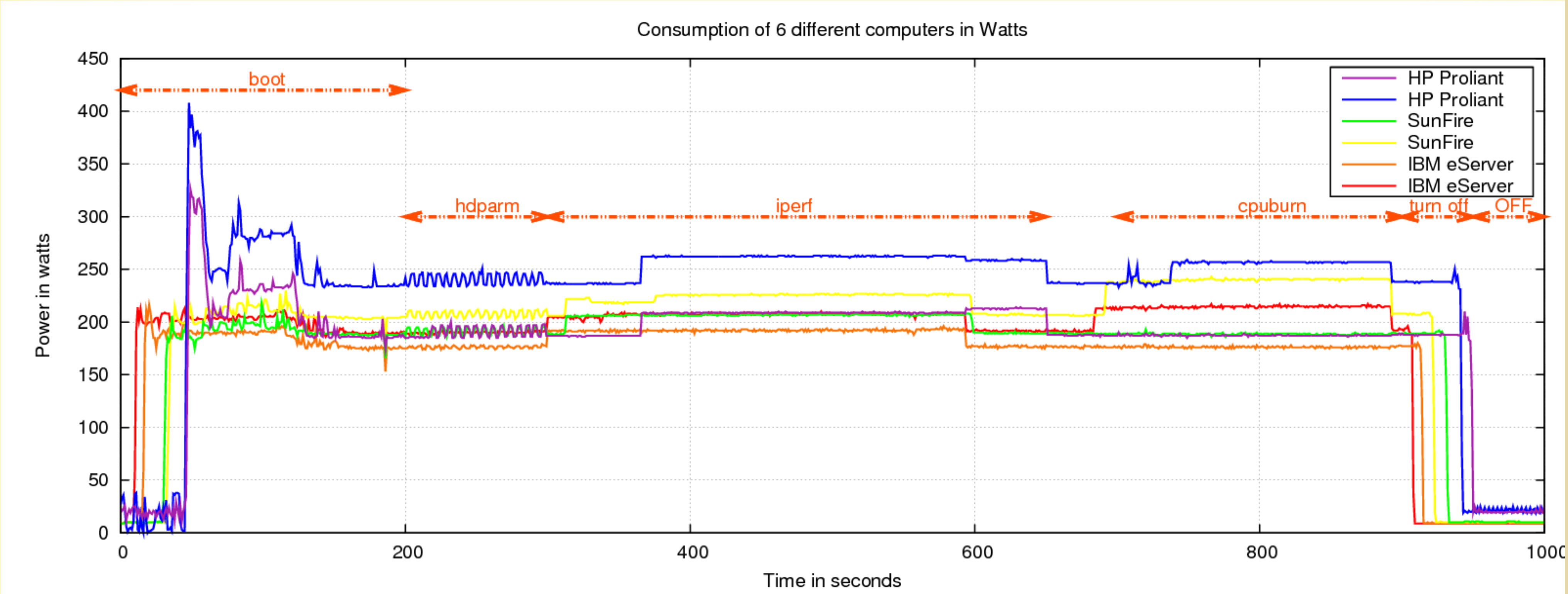
- hot spots

- need first a measurement campain of all the nodes

- require quick and efficient scheduling algorithms

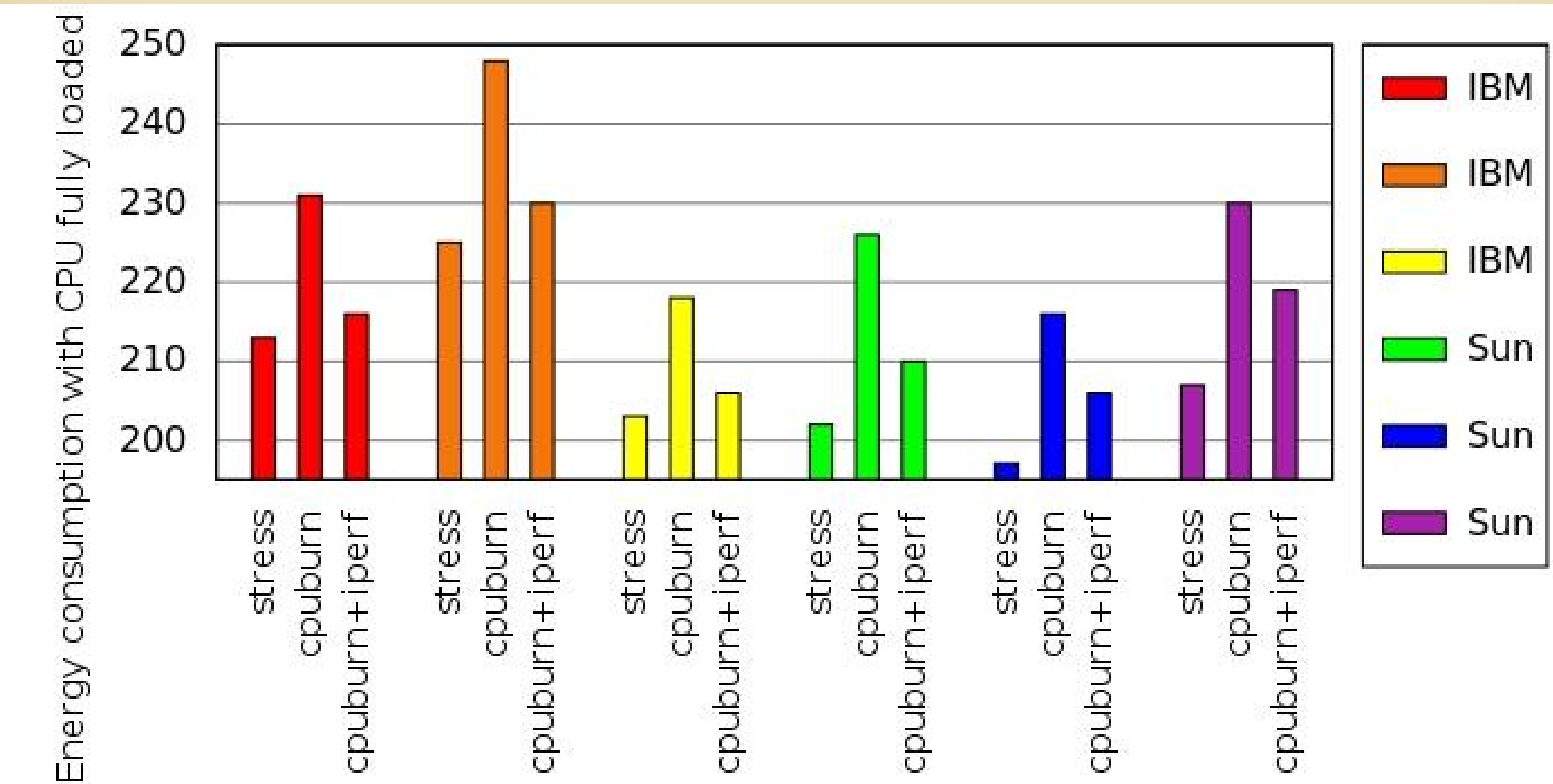
"*Energy-Aware Task Scheduling with Task Synchronization for embedded real-time systems* ", R. Jejurikar and R. Gupta, IEEE Trans. On Computer-Aided Design Of Integrated Circuits and Systems, 2006.

Does homogeneous nodes have the same energy consumption?



"*Demystifying Energy Consumption in Grids and Clouds* ",
A.-C. Orgerie et al., WIPGC, 2010.

Is the relation between CPU load and energy consumption linear ?



"*Demystifying Energy Consumption in Grids and Clouds* ",
A.-C. Orgerie et al., WIPGC, 2010.

Virtualization techniques

- Consolidation and aggregation → reduce number of servers
- Isolation → application workload characterization
- Live Migration → live adaptation

Issues:

- Time overhead, resource sharing
- Hypervisor needs resources → reduce capacities of the nodes

"*Using Virtualization to Improve Data Center Efficiency* ",
Green Grid White Paper, 2009.

Does virtualization increase energy consumption?

- No paper to either deny or confirm this assumption.
- Experiments with Xen 3.2 and Xen Server 5.0

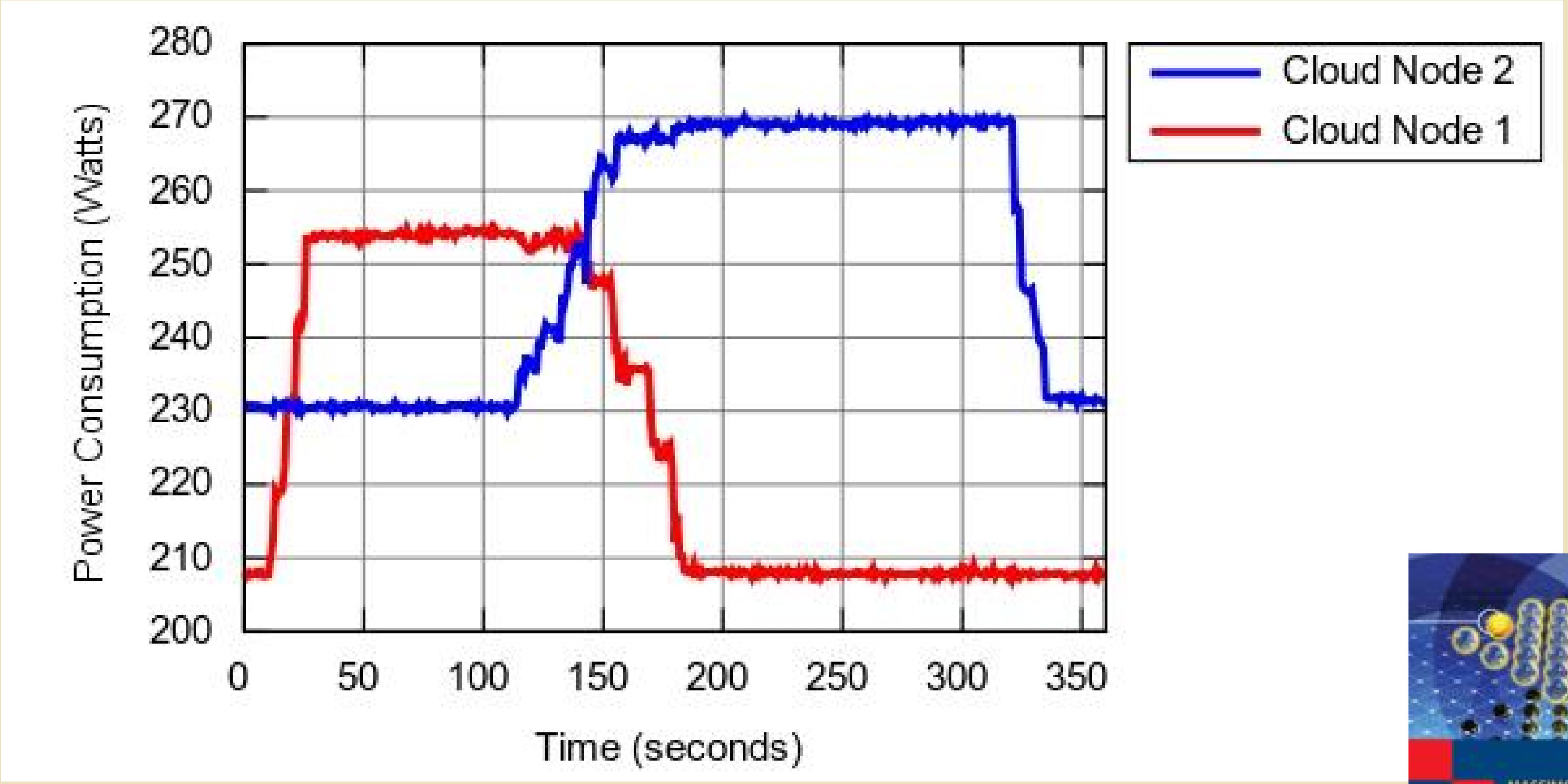
  - idle consumption identical

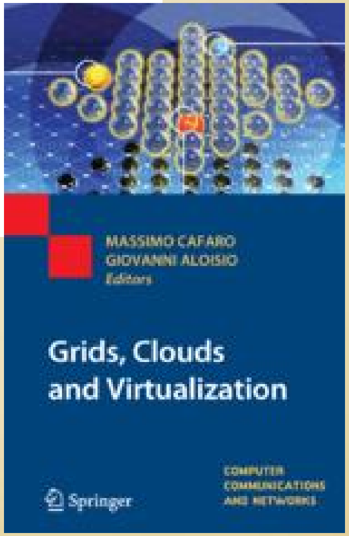
  - cpuburn experiments show a difference of less than 1 Watt

  - iperf experiments show a difference of 7 Watts (due to a poor management of network allocation in Xen 3.2)

"*Demystifying Energy Consumption in Grids and Clouds*",
A.-C. Orgerie et al., WIPGC, 2010.

Is live migration of VM free?





"*Designing and Evaluating an Energy-Efficient Cloud* ",
A.-C. Orgerie et al., Journal of Supercomputing, 2010.

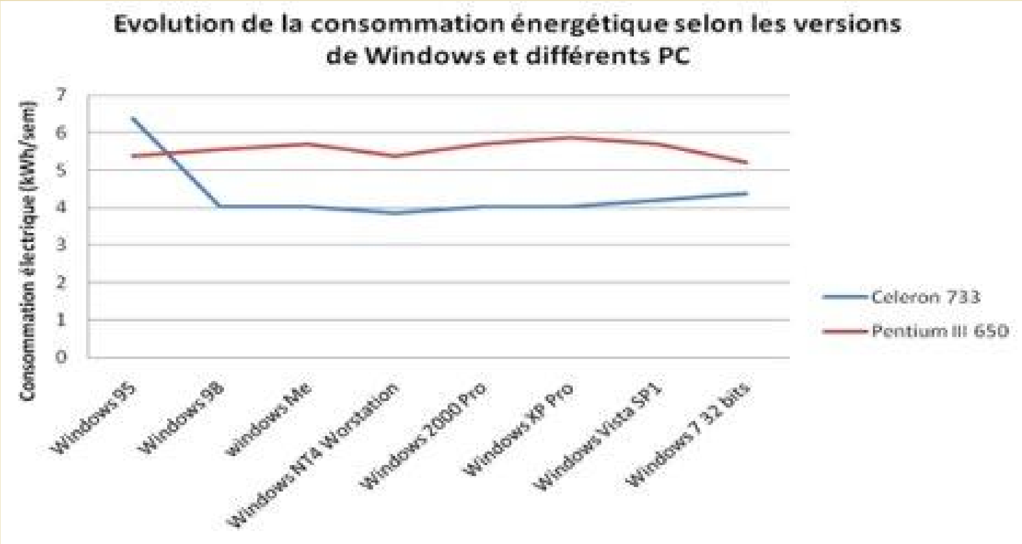
Dynamic Voltage/Frequency Scaling

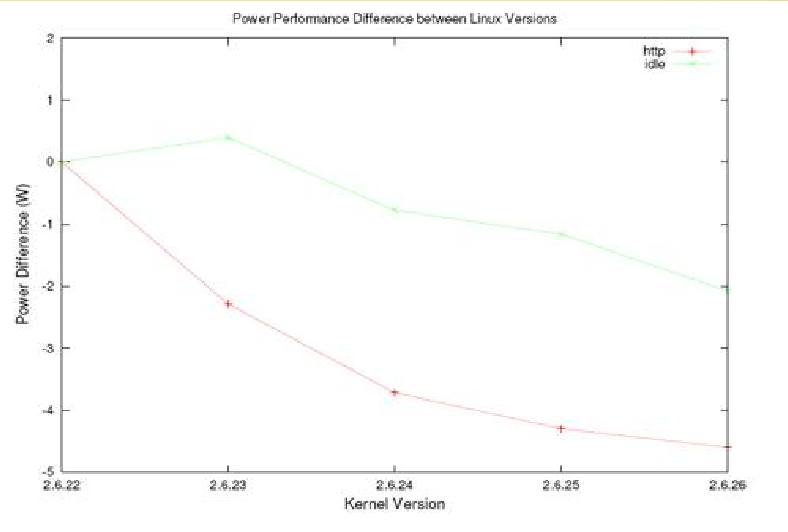
- CPU-throttling: frequency of a microprocessor can be automatically adjusted "on-the-fly" (Cool'n'Quiet for AMD, SpeedStep for Intel)
- Different voltages/frequencies are available on the CPU
- Trade-off between time and energy
- Need to predict application profile in terms of CPU usage

“*Critical Power Slope: Understanding the Runtime Effects of Frequency Scaling*”, A. Miyoshi et al., ICS, 2002.
“*Energy-Efficient Real-Time Heterogeneous Server Clusters*”, C. Rusu et al., RTAS, 2006

Code improvements

- OS code
- Green programing
- Automatically turning off unused components (USB ports, Bluetooth, Wifi, kernel modules)





http://lesswatts.org/projects/powertop/

http://www.ecoinfo.cnrs.fr/spip.php?article211

Hardware improvements

- Improve energy-efficiency of the components (power supply, fans, water cooling…)
- Energy-efficient architectures and infrastructures
- Allow components to be put into sleep modes separately (memory benchs, cores, …)
- Allow different levels of usage (voltages, frequencies)
- Ecodesign and recycling

Standards, Consortiums, Projects

- Energy Star
- Green Grid
- Efficient Servers
- The green 500
- ...

the green grid

ENERGY STAR

EFFICIENT-SERVERS

EPEAT

EcoInfo

THE GREEN 500

cost

www.cost804.org

Action IC0804

Experimental Scenario

Objective and Methodology

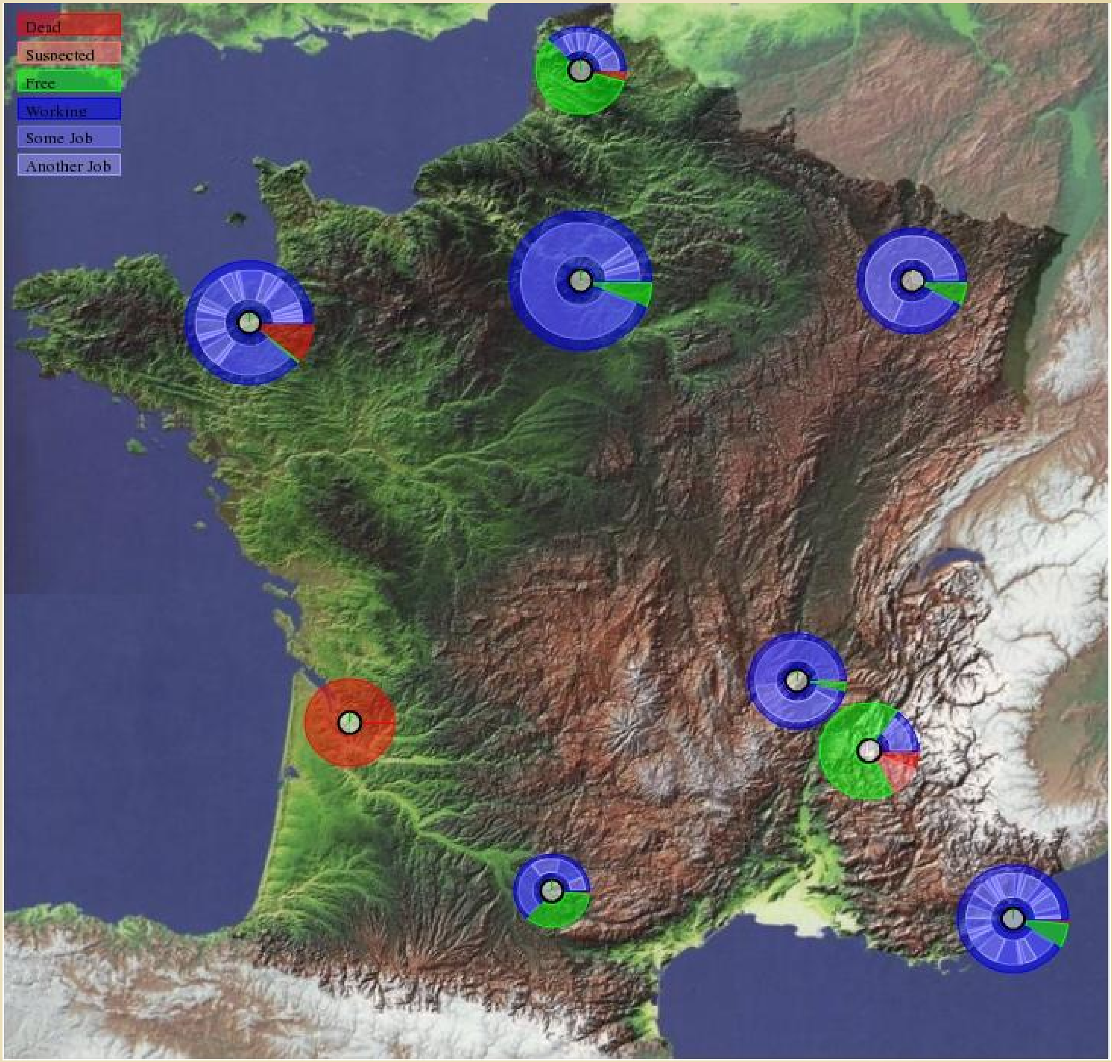
How to decrease the consumption without impacting the performances?

- Switch off the unused resources;
- Predict the next use;
- Aggregate the reservations.

- Study the existing solutions for energy savings;
- Analyze a platform usage and its power consumption;
- Propose an energy-aware infrastructure for resource reservation;
- Validate experimentally this infrastructure.

Grid'5000

- French experimental testbed
- 5000 cores
- 9 sites
- Dedicated Gb network
- Designed for research on large-scale parallel and distributed systems



Lyon: a Monitored Site

- 135 nodes
- One power measurement per node and per second

“The Green Grid'5000: Instrumenting a Grid with Energy Sensors”

M. Dias de Assunção, J.-P. Gelas, L. Lefèvre and A.-C. Orgerie, *INGRID 2010*, May 2010.

Resource Reservation: OAR

- Open source batch scheduler
- Resource management system of Grid'5000
- Advance reservations
- Exclusive reservation of resources (super-user)
- Image deployment with Kadeploy

  → resource reservation logs + deployment logs

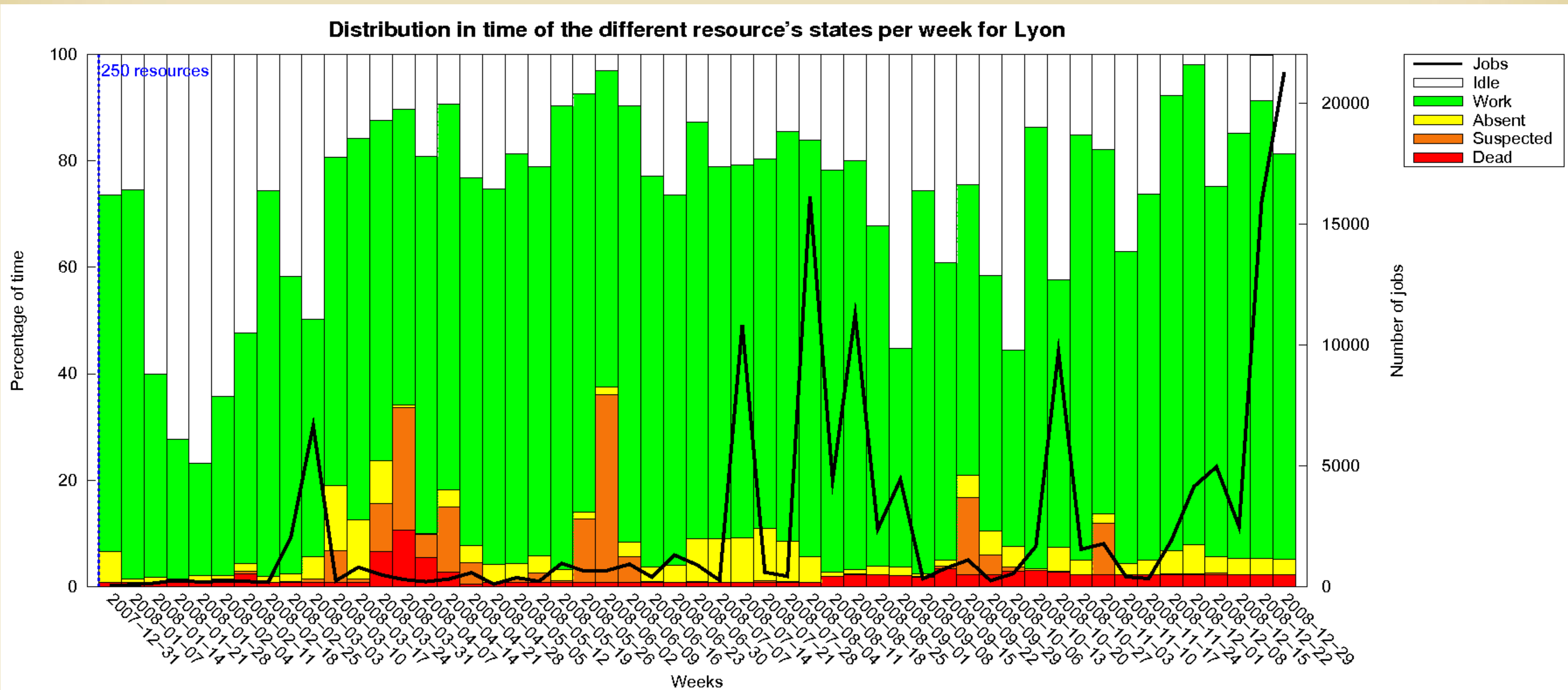
Grid view

OAR (batch scheduler and resource manager)
Logs of 2008 year (2 GB; 1,290,000 reservations)

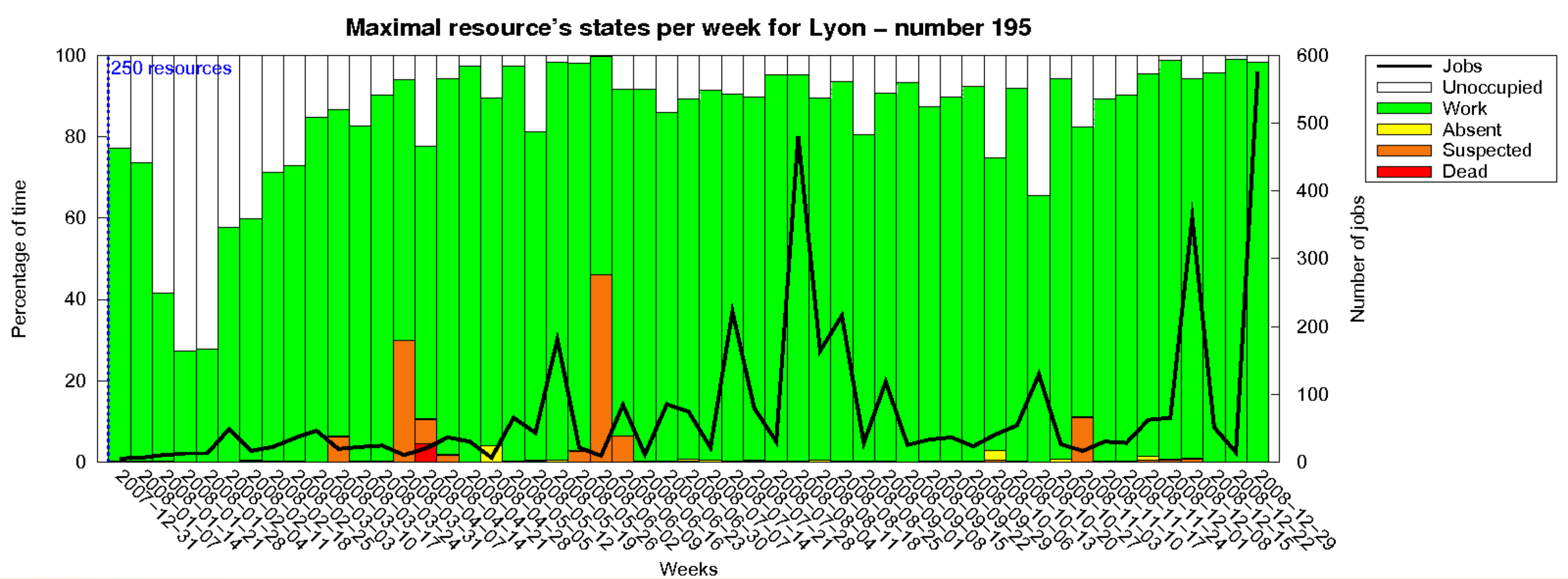
Site	number of cores	number of reservations	Mean number of resources per job	Mean duration of a job (mn)	Percentage of activity
Bordeaux	650	356,222	7.44	41	53.20 %
Lille	618	344,538	8.11	53	72.89 %
Lyon	322	138,217	4.39	62	69.27 %
Nancy	574	74,592	14.63	149	60.08 %
Orsay	684	92,862	14.58	104	57.82 %
Rennes	714	58,843	27.32	118	64.58 %
Sophia	568	58,142	22.14	146	81.51 %
Toulouse	434	166,191	6.29	37	61.67 %

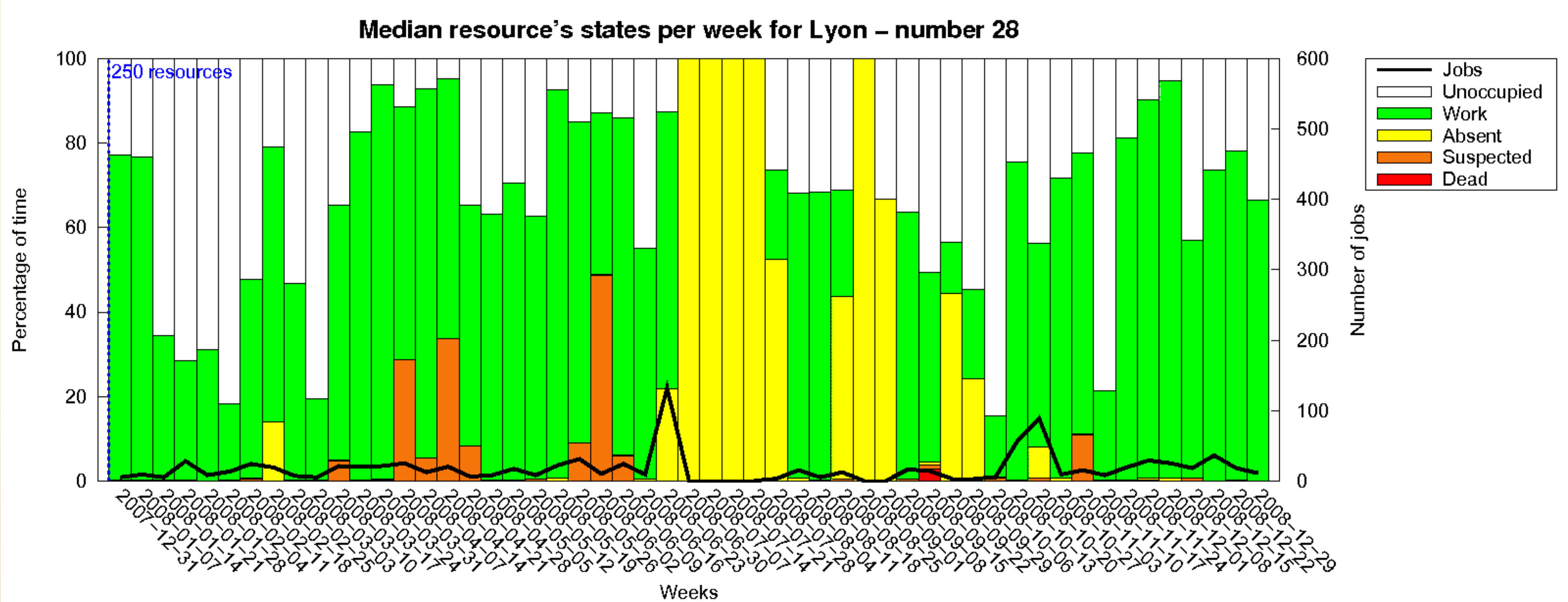
- Lots of variations from one site to another.
- Operational Grid: between 60 and 70% of usage.
  ”*How are Real Grids Used? The Analysis of Four Grid Traces and Its Implications*” A. Iosup et al., Grid Computing, 2007.

Zoom on a site: Lyon site

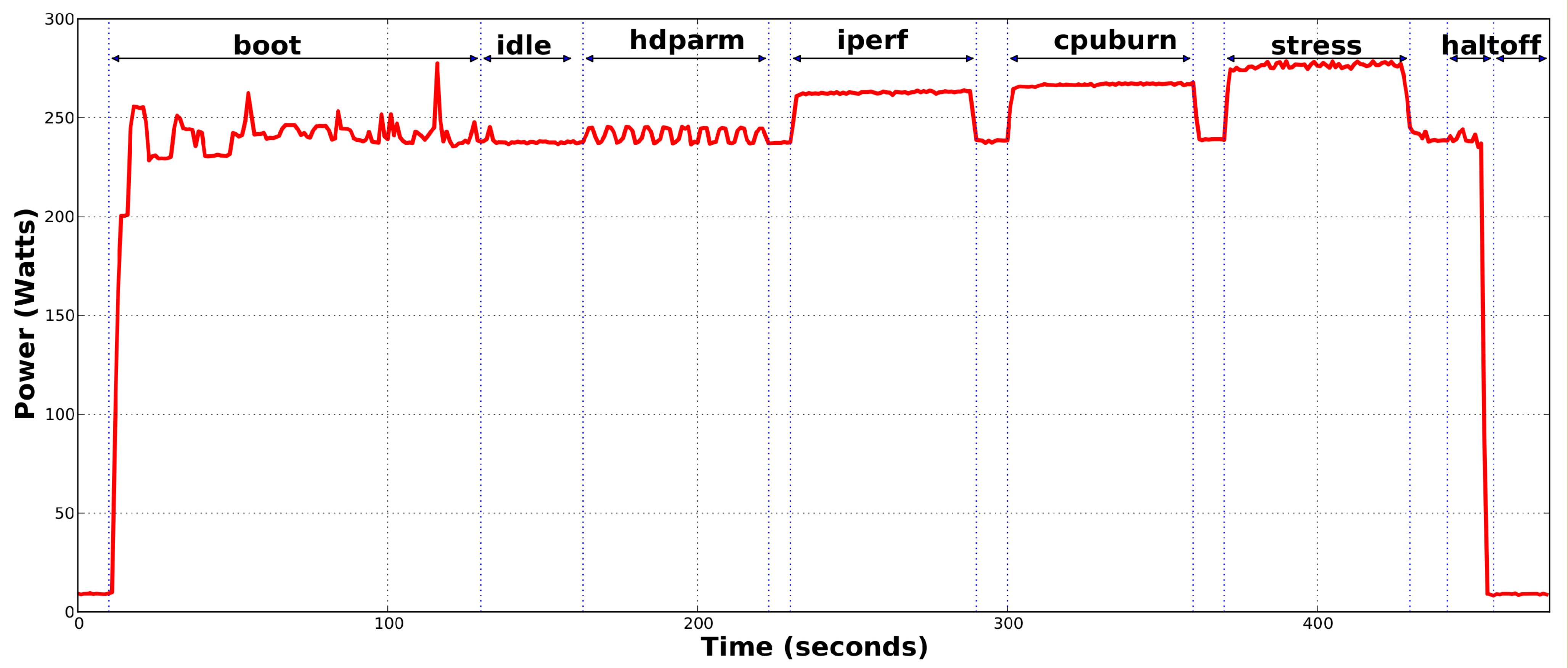


Heterogeneous Resources





Impact of Usage on Power Consumption



Idle consumption really high, no prediction possible

→ Huge energy savings are possible

Power Consumption Logs: Failures

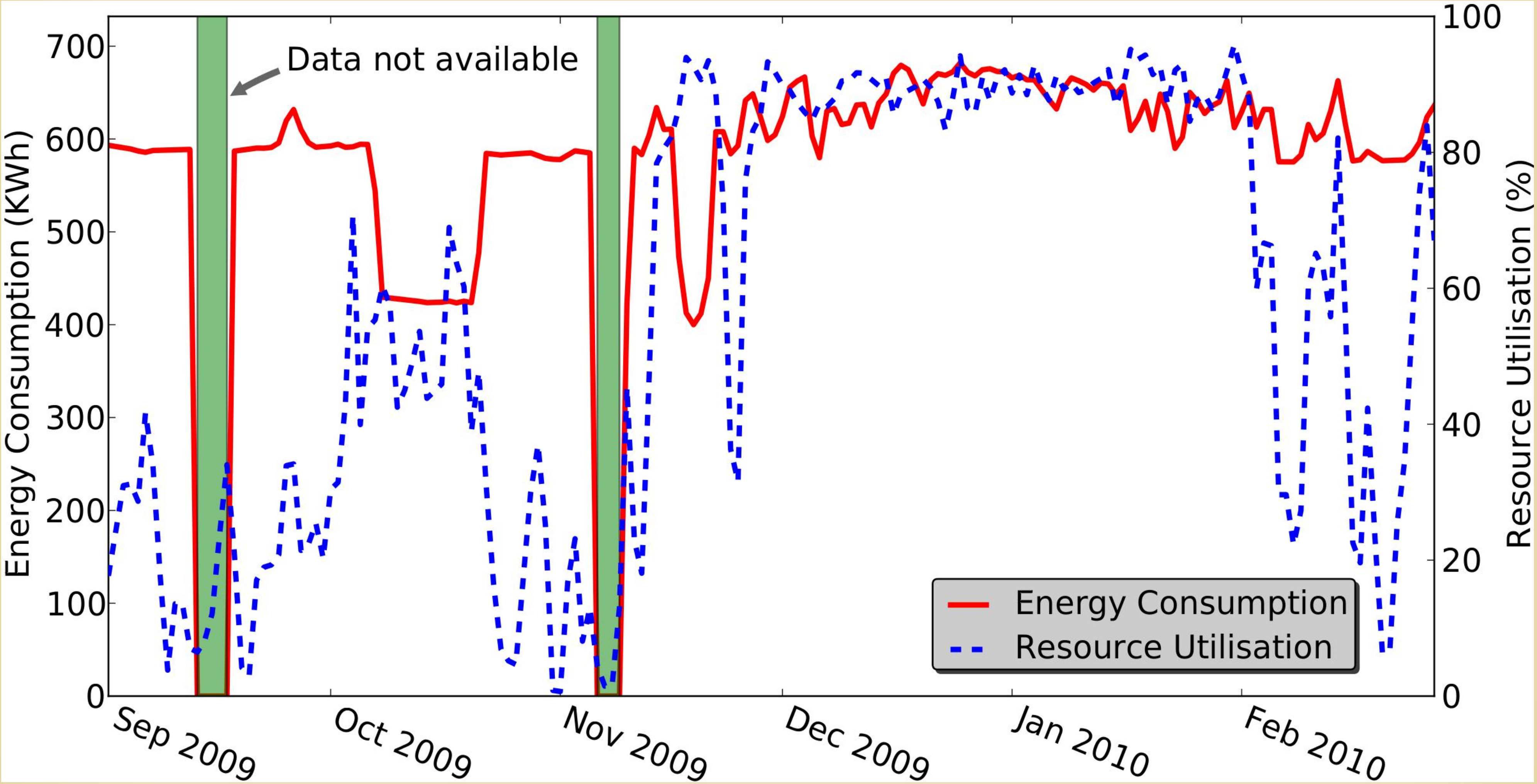
- 6 months logs:
from 1st Sep. 2009
to 27th Feb. 2010

- 135 2-CPU nodes
56 IBM eServer 326
79 Sun Fire V20z

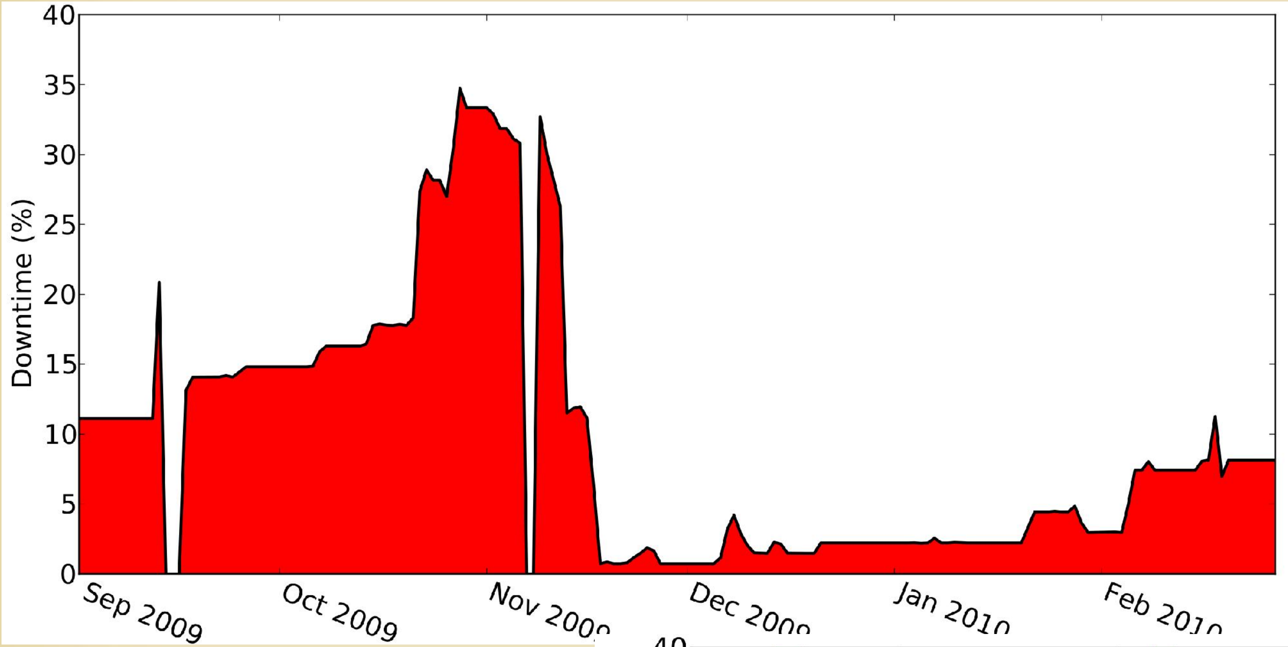
Start Time	Duration	Description
06-09-2009 04:25:41	01:11:34	Failure in the wattmeters.
13-09-2009 19:13:10	90:13:40	Platform downtime.
17-09-2009 13:30:56	00:02:40	
19-09-2009 08:31:26	01:11:34	Failure in the equipments responsible for measuring the power consumption.
20-09-2009 04:30:25	01:11:34	
27-09-2009 04:25:02	01:12:33	
24-10-2009 13:33:52	00:17:51	
25-10-2009 05:34:31	01:11:34	
01-11-2009 05:25:16	01:12:05	
06-11-2009 16:11:19	67:57:15	Platform downtime.
10-11-2009 11:47:32	01:28:38	Failure in the equipments responsible for measuring the power consumption.
15-11-2009 05:25:16	01:11:34	
22-11-2009 05:25:05	01:12:56	
13-12-2009 05:25:08	01:13:40	
20-12-2009 05:25:46	01:11:34	
10-01-2010 05:25:13	01:13:11	
17-01-2010 05:25:13	01:11:34	
31-01-2010 05:25:33	01:13:19	
07-02-2010 05:25:37	01:11:34	

Global Energy Consumption

Global Energy Consumption

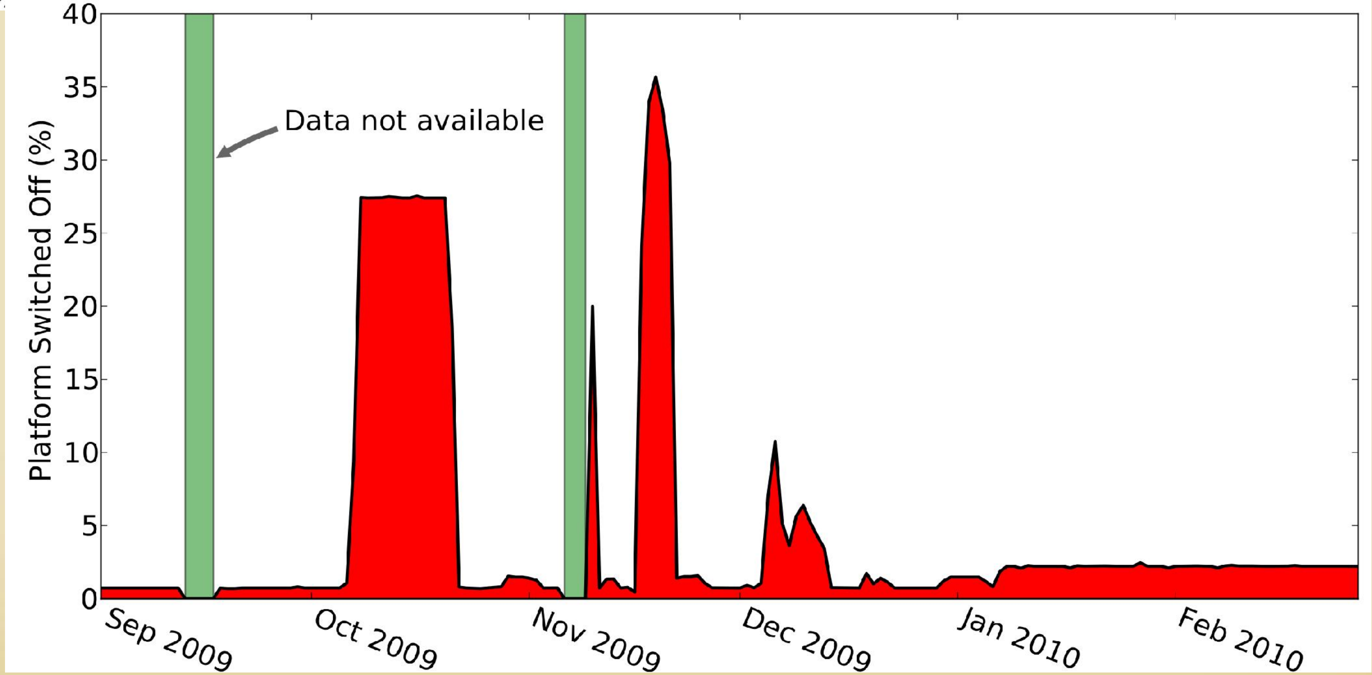


Platform downtime and switch off

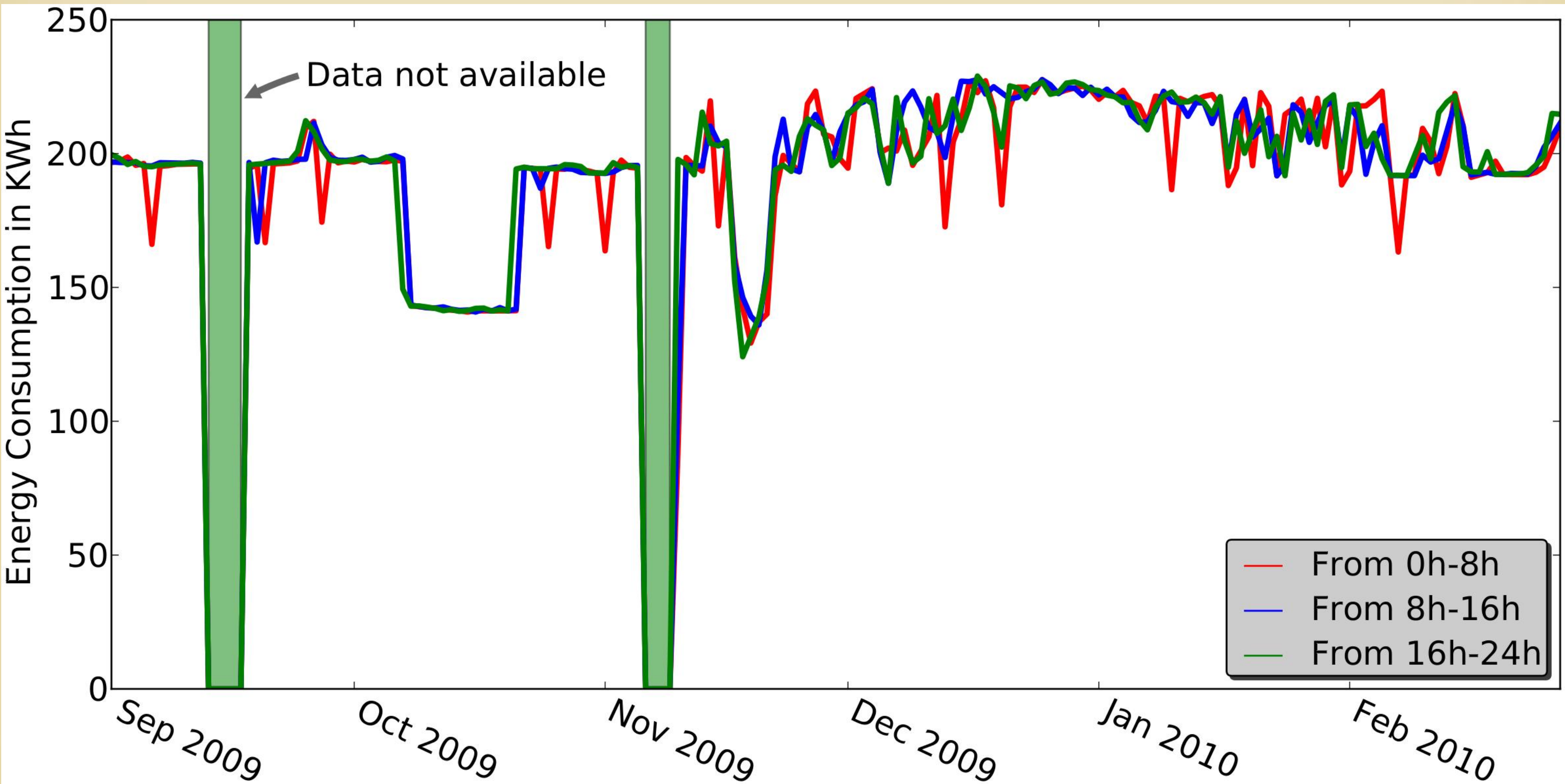


Downtime

Switch off



Energy consumption during the day



Resource Reservations

Energy consumed by reservation categories

Number of nodes used	Number of reservations	Overall consumption in KWh	Average Watts per Node
1 to 2	3844	1529.77	197.23
3 to 5	681	1203.71	192.36
6 to 10	611	7384.92	200.44
11 to 30	7408	35371.17	216.49
31 to 70	205	9821.79	178.78
71 to 100	45	1918.33	185.91
101 to 135	50	5447.23	185.85

→ reservations with up to 30 nodes tend to consume more energy

Energy consumed by reservation groups

Duration	Number of nodes	
	≤ 10 nodes	> 10 nodes
≤ 1 hour	SN (27.18%)	SW (47.26%)
> 1 hour	LN (12.81%)	LW (12.75%)

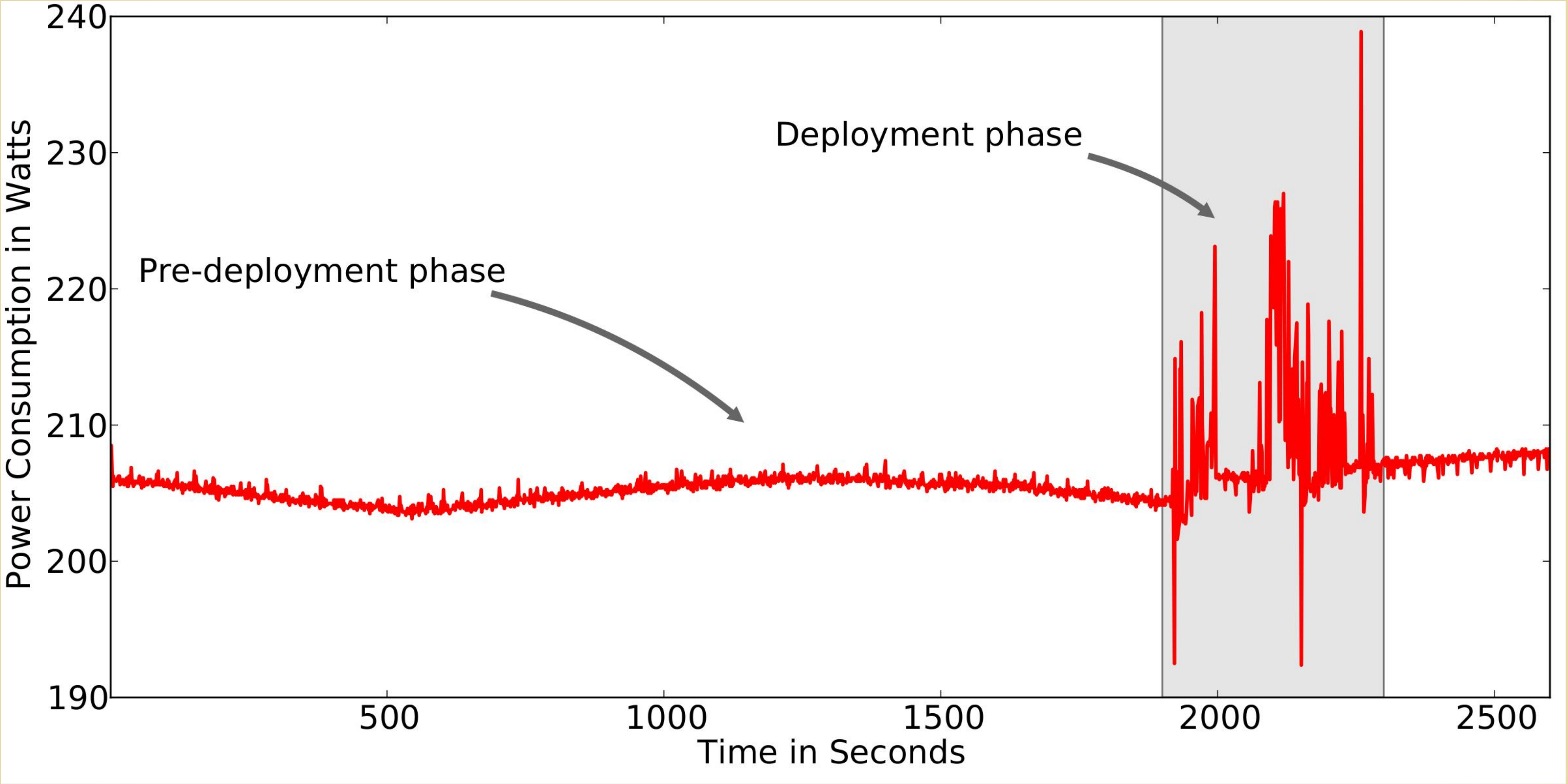
Average Weighted Power Consumption

$$AWPC = \frac{\sum_{j \in \tau_k} p_j \cdot m_j \cdot wn_j}{\sum_{j \in \tau_k} p_j \cdot m_j}$$

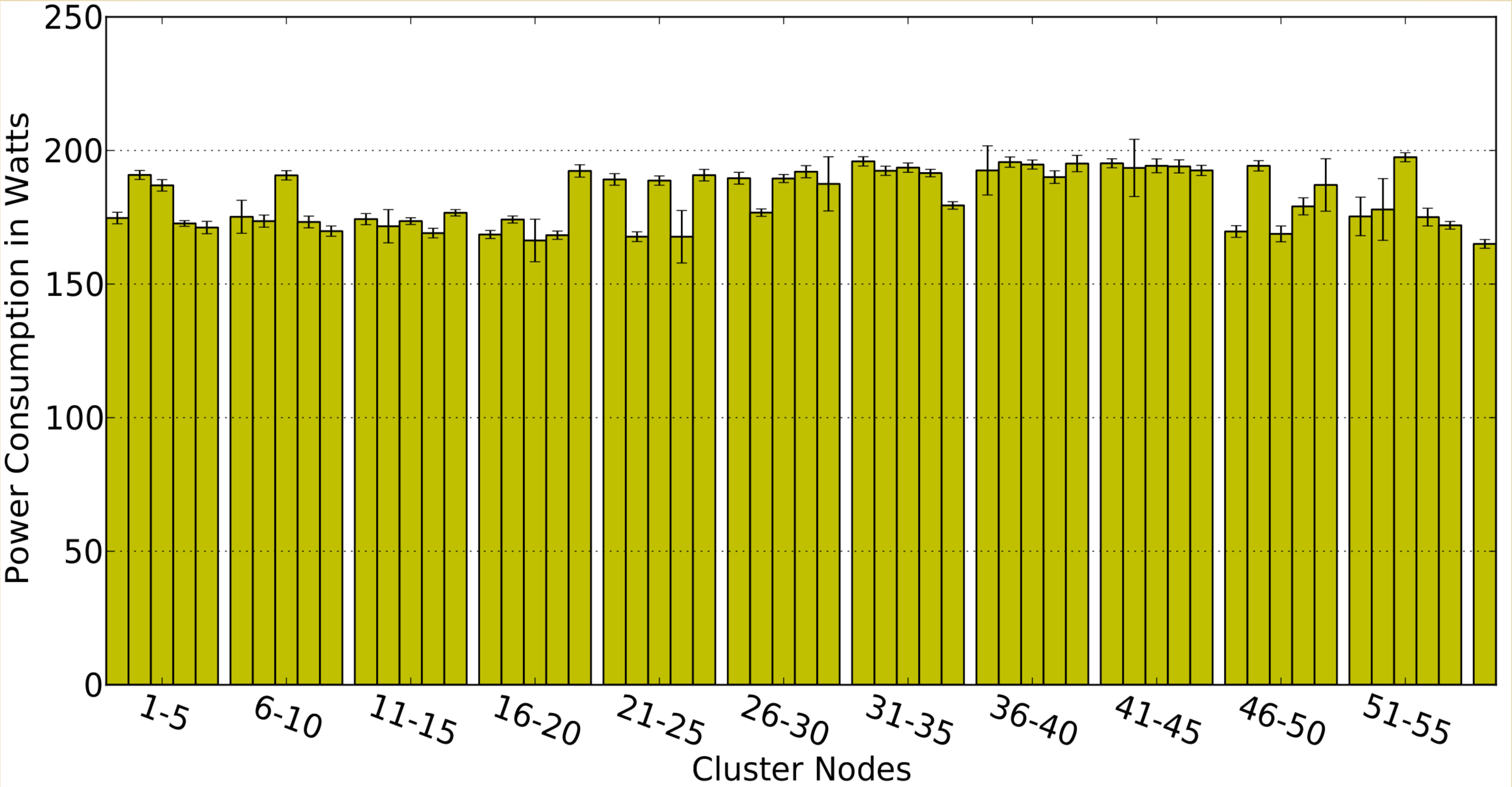
Reservation Group	Overall Consumption in KWh	Average Watts per Node	AWPC
SN	278.29	198.78	191.46
SW	6406.21	217.72	207.09
LN	9840.11	193.10	197.84
LW	46152.31	205.42	184.23

Energy Cost of Resource Under-Utilisation

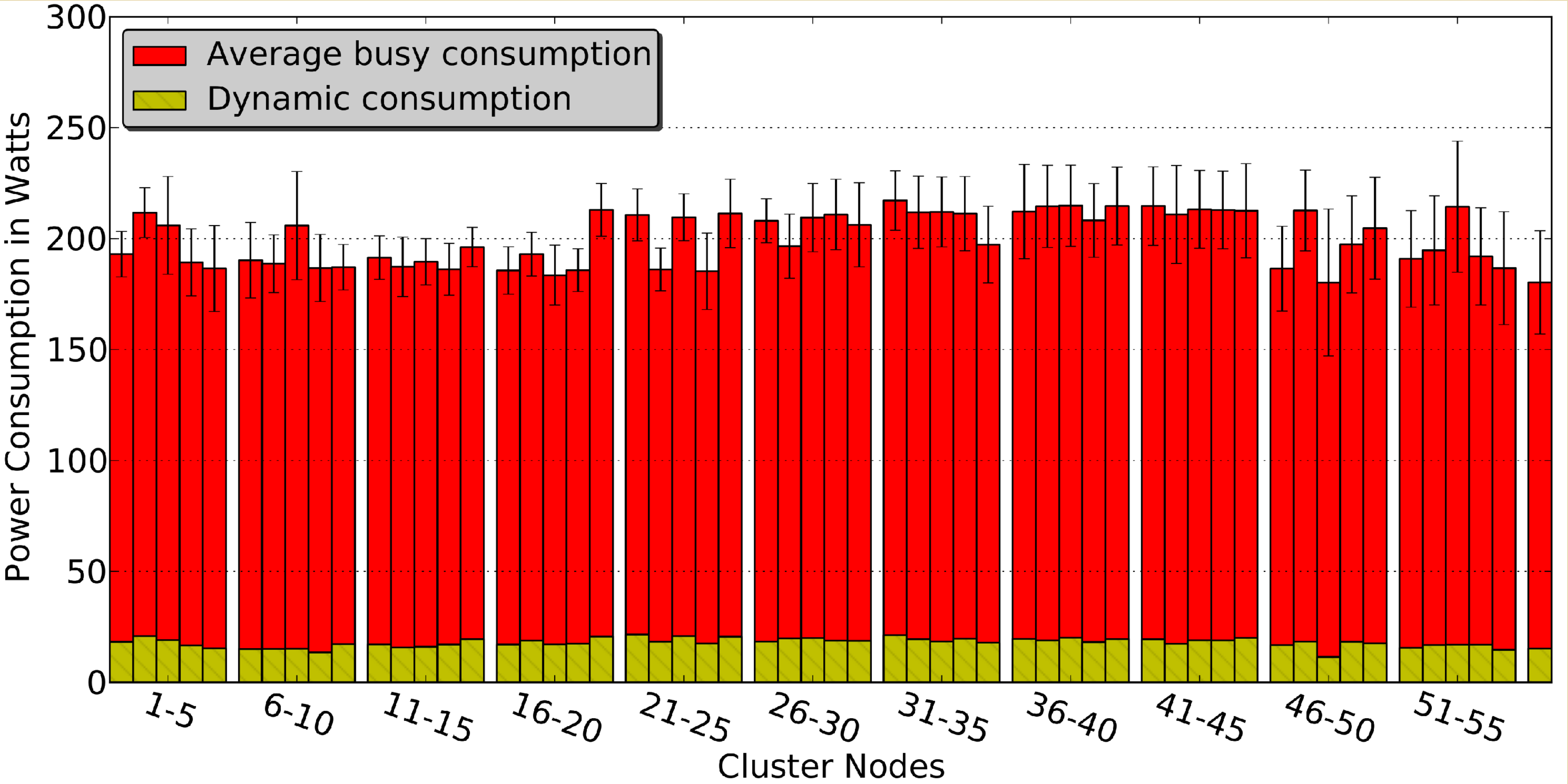
Consumption during late deployment



Idle Power Consumption



Busy Power Consumption



Lessons Learnt

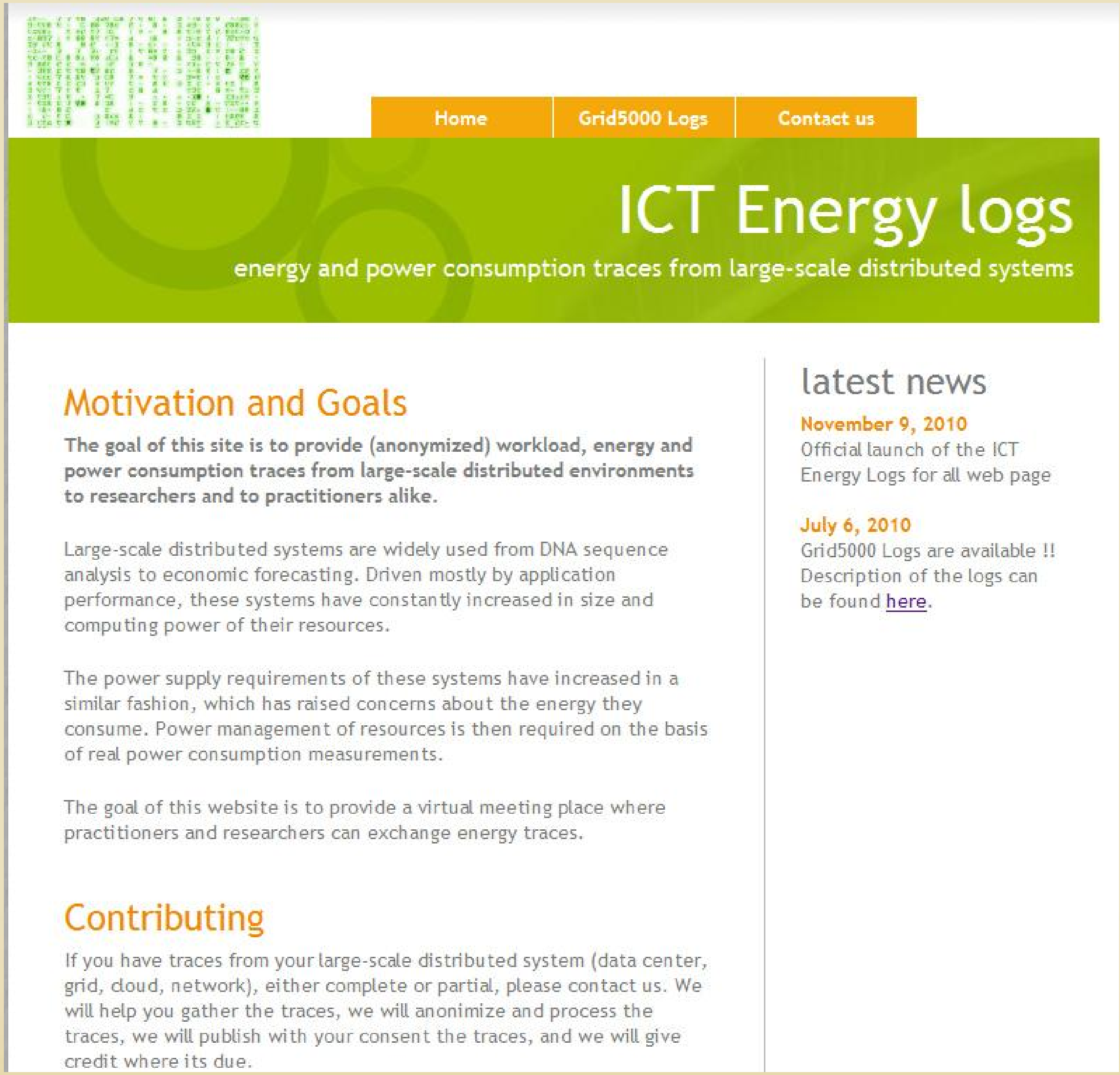
Lessons learnt

- Dynamic energy consumption corresponds to **3.05%** of the overall energy consumed during reservations. → maximum that CPU throttling can save currently
- Idle power consumption is really high.
- Energy consumed during pre-deployment phase corresponds to **14.41%** of the overall energy.
- Small reservations are more resource intensive.

→ Users behavior should be taken into account.

ICT Energy logs

http://www.ens-lyon.fr/LIP/RESO/ict-energy-logs/



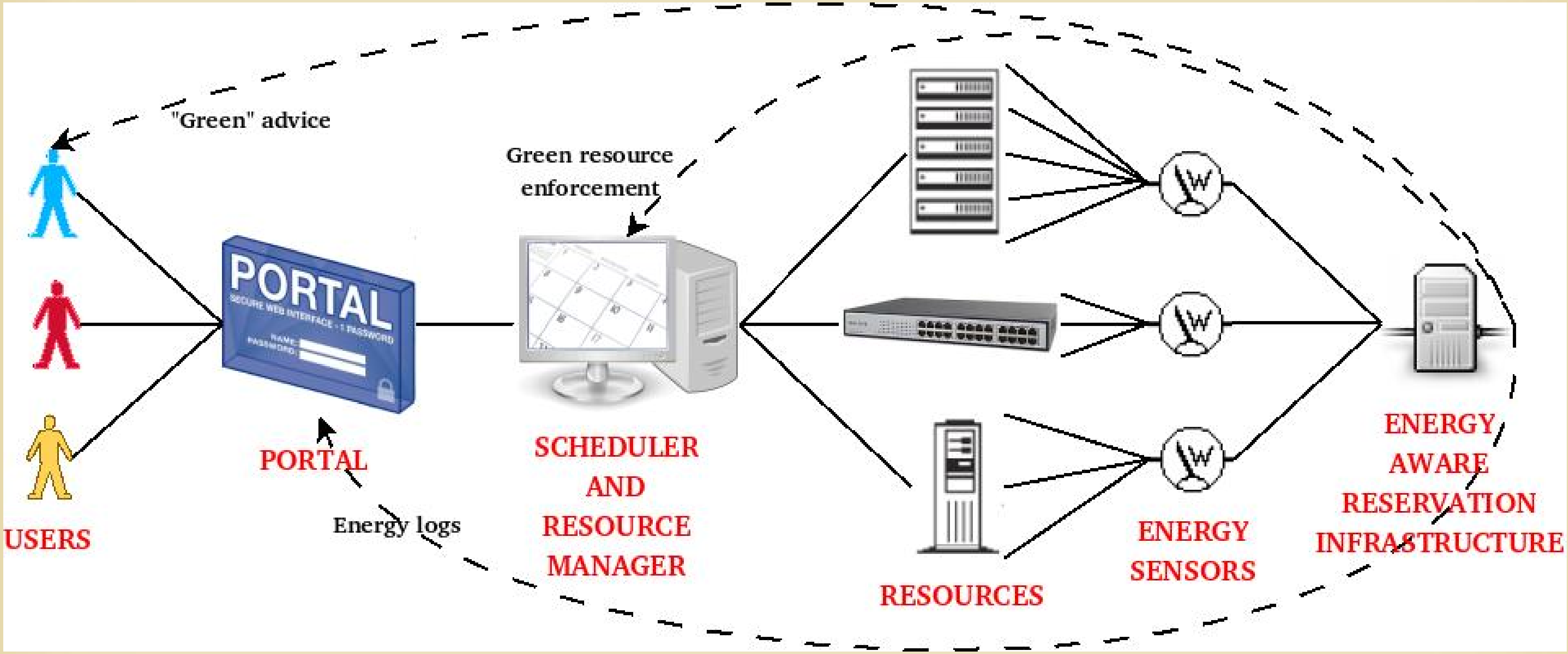
Energy-Aware Reservation Infrastructure

Energy-Aware Reservation Infrastructure (EARI)

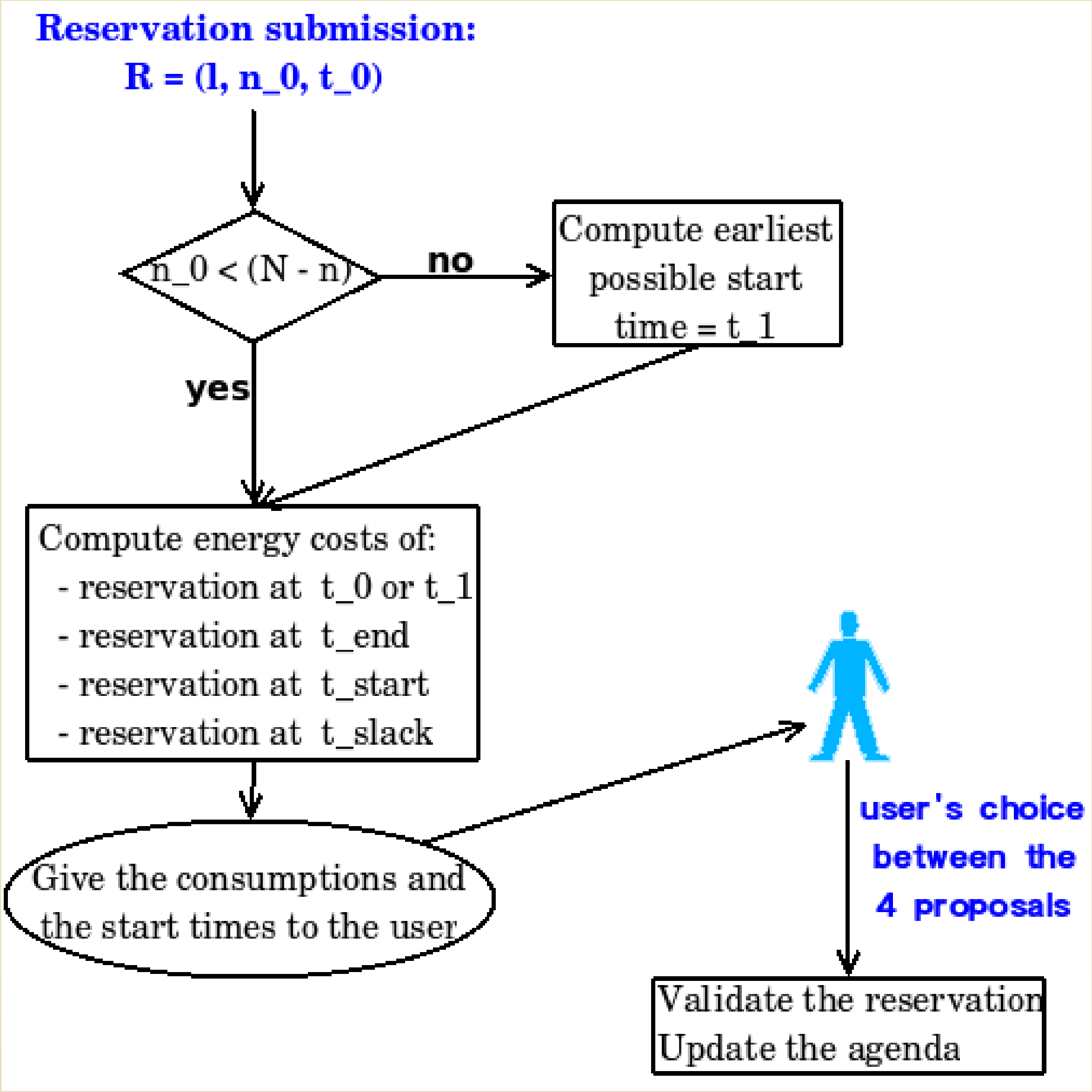
The main features are:

- **Switch off** unused computing resources;
- **Predict** next use;
- **Aggregate** the reservations by giving green advice to the users.

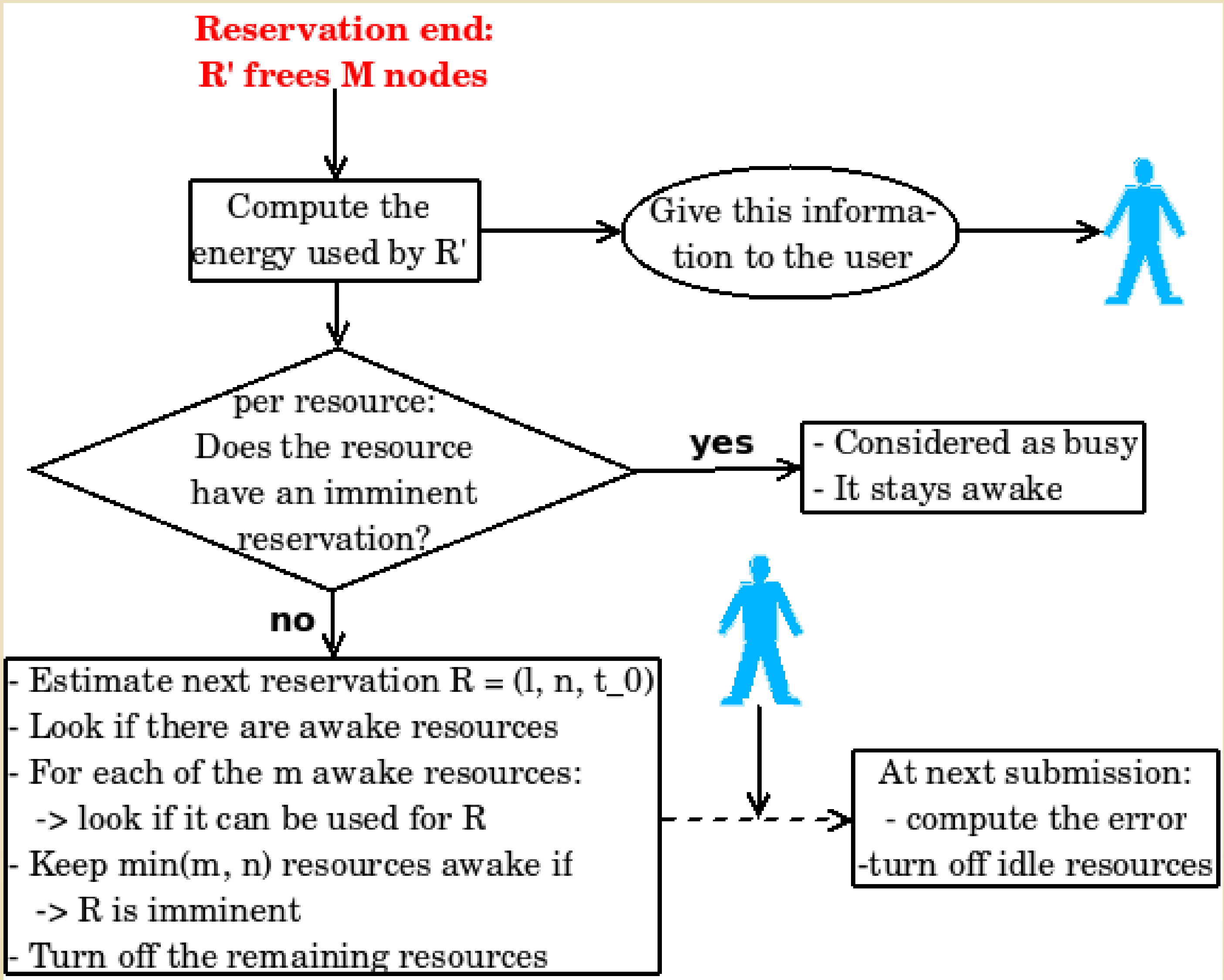
Architecture



After a reservation request



At the end of a reservation



Predictions

What :

- Next reservation (size, duration, start time)
- Next empty period
- Energy consumption of a reservation

With :

- Recent history (last reservation) + feedback
- Recent reservations days + feedback
- User history + resources

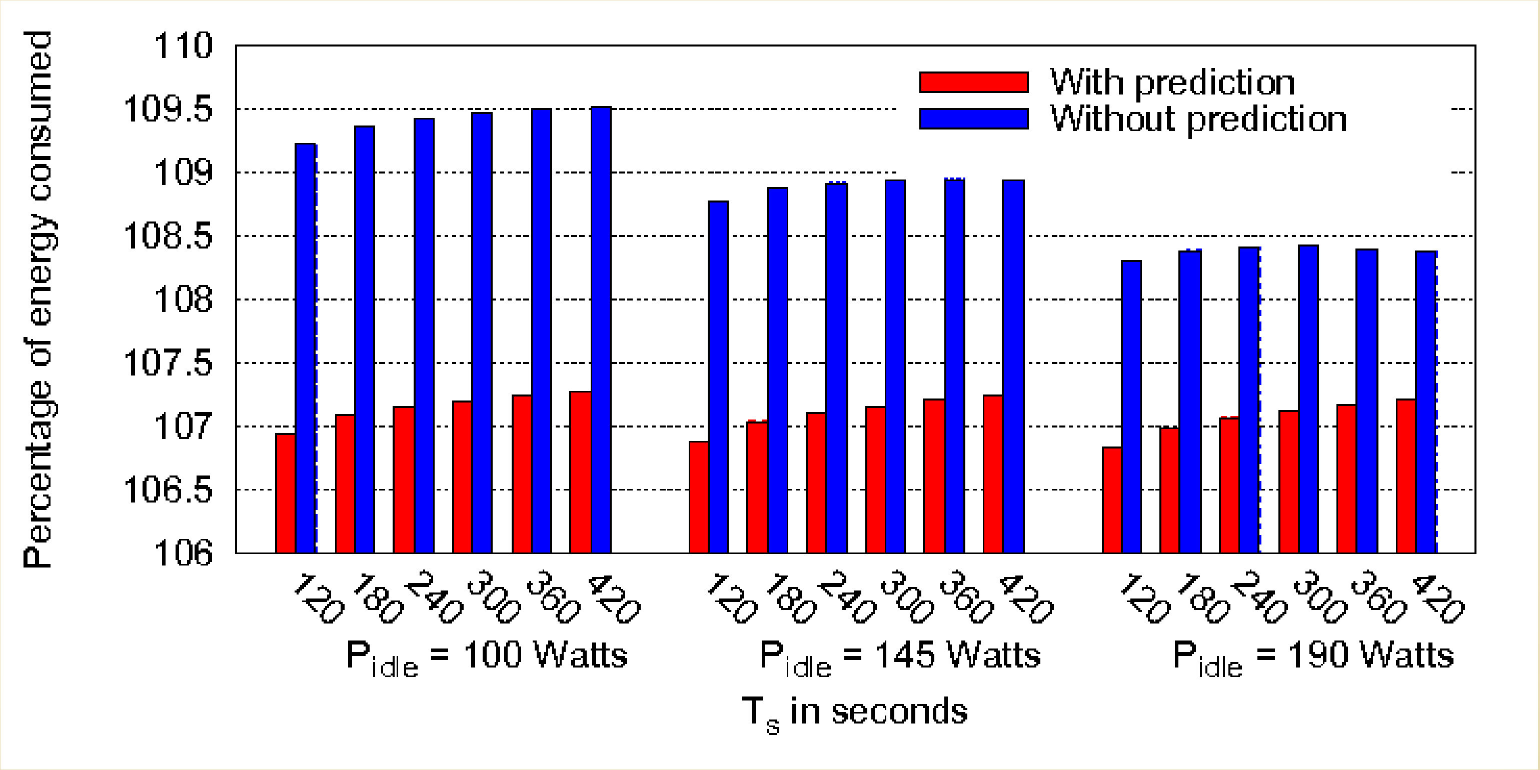
How to validate the predictions?

Prediction evaluation based on replay

Example: Bordeaux site (650 cores, 45K reservations, 45% usage)

100 % : theoritical case (future perfectly known)

Currently (always on) : 185 % energy



Green Policies

- **user**: requested date
- **25% green**: 25% of jobs follow Green advices – the rest follows user request
- **50% green**: 50% of jobs follow Green advices – the rest follows user request
- **75% green**: 75% of jobs follow Green advices – the rest follows user request
- **fully green**: solution with uses the minimal amount of energy and follows Green advices
- **deadlined**: fully green for 24h – after: user

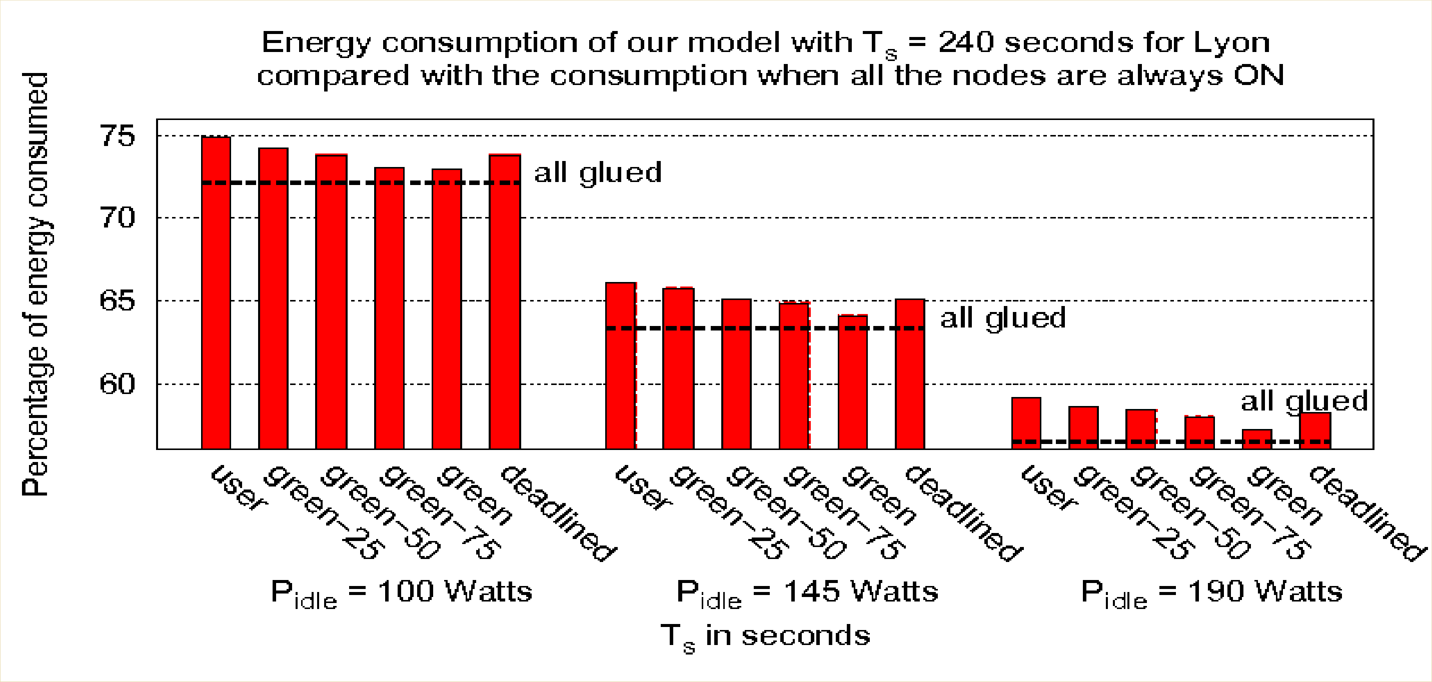
Evaluation on Lyon example

Example of Lyon site (322 cores, 33K reservations, 46% usage)

Current situation: always ON nodes (100 %)

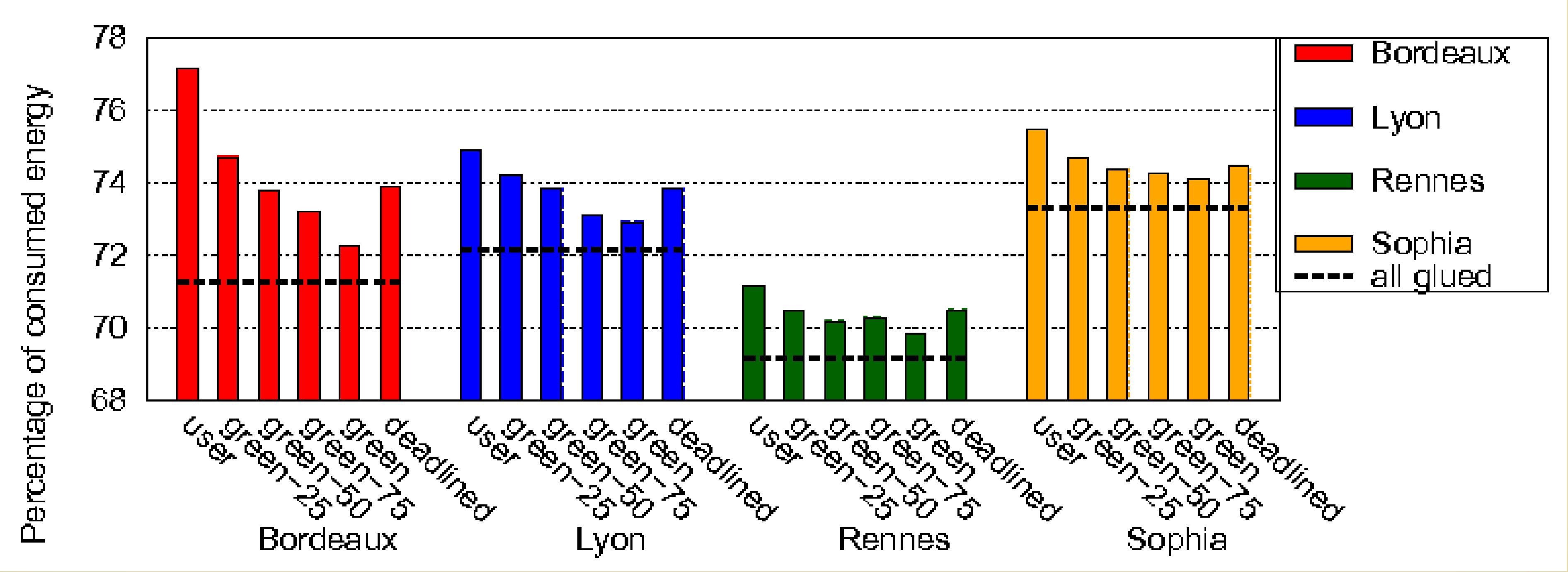
All glued: unreachable theoretical limit

For Lyon site: saving of 73,800 kwh for 2007 period



Experimental validation of EARI

- Real traces of an experimental Grid: Grid'5000
- 4 different sites, one year period



Extrapolation to the whole Grid

209,159 kWh for the full Grid'5000 platform (without aircooling and network equipments) on a 12 month periods (2007)

It represents the consumption of a french village of 600 inhabitants.

So roughly, a village of 1200 inhabitants for the whole infrastructure.

Conclusions

- Proposition of an energy-aware infrastructure for resource reservation

  - simple and quick in terms of computing time

  - including heuristics

  - proposing energy saving solutions to the users without forcing them and impacting performances

  - leading to important energy savings.

Publications

Laurent Lefèvre and Anne-Cécile Orgerie.
"Towards Energy Aware Reservation Infrastructure for Large-Scale Experimental Distributed Systems"
Parallel Processing Letters, 19(3):419-433, Sept. 2009.

Georges Da-Costa, Jean-Patrick Gelas, Yiannis Georgiou, Kamal Sharma, Laurent Lefèvre, Anne-Cécile Orgerie, Jean-Marc Pierson and Olivier Richard.
"The GREEN-NET Framework: Energy Efficiency in Large Scale Distributed Systems",
HPPAC 2009, Workshop in conjunction with IPDPS 2009 , May 2009.

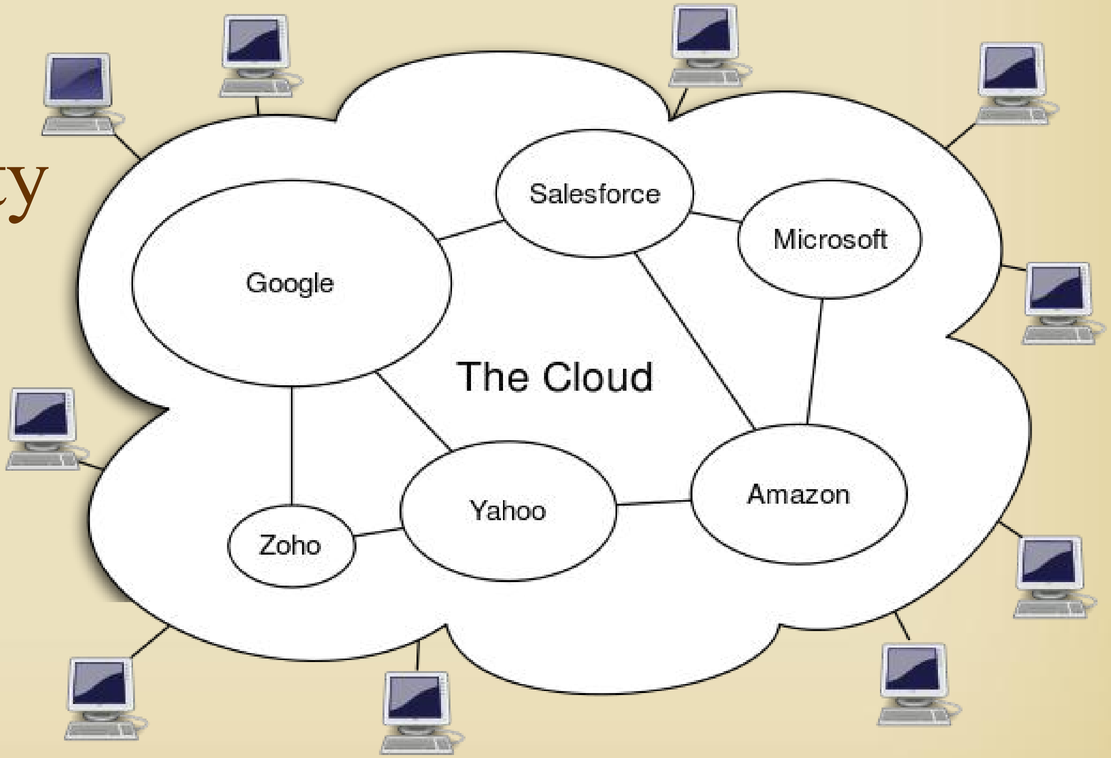
Anne-Cécile Orgerie, Laurent Lefèvre and Jean-Patrick Gelas.
"Chasing Gaps between Bursts : Towards Energy Efficient Large Scale Experimental Grids",
PDCAT 2008, pages 381-389, Dec. 2008.

Anne-Cécile Orgerie, Laurent Lefèvre and Jean-Patrick Gelas.
"Save Watts in your Grid: Green Strategies for Energy-Aware Framework in Large Scale Distributed Systems"
ICPADS 2008, pages 171-178, Dec. 2008.

Green Open Cloud

Cloud Features

- Virtualization
- Resources as a service
- Accounting, pricing
- Scalability, reliability
- Security



Differences between Grids and Clouds (that interest us)

- Agenda (reservations in advance)
- Virtualization
- Possibility to use live migration
- Usage predictions
- Resource management

Energy Cost of VMs

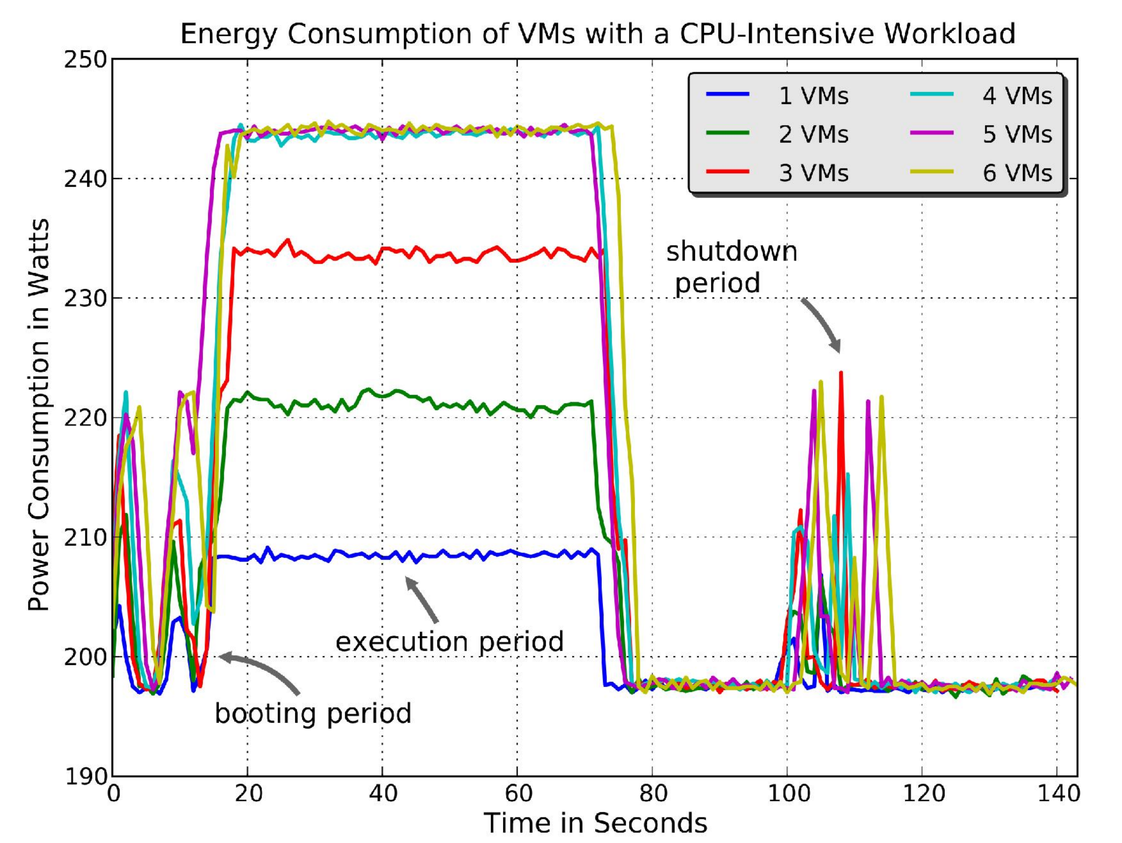
- **Energy analysis** of virtual machines using **energy sensors** plugged on the cloud nodes during typical and advanced **cloud operations** (boot of a VM, run, halt, migration)

- Case of computing jobs (worst case for the energy consumption)

Experimental Conditions

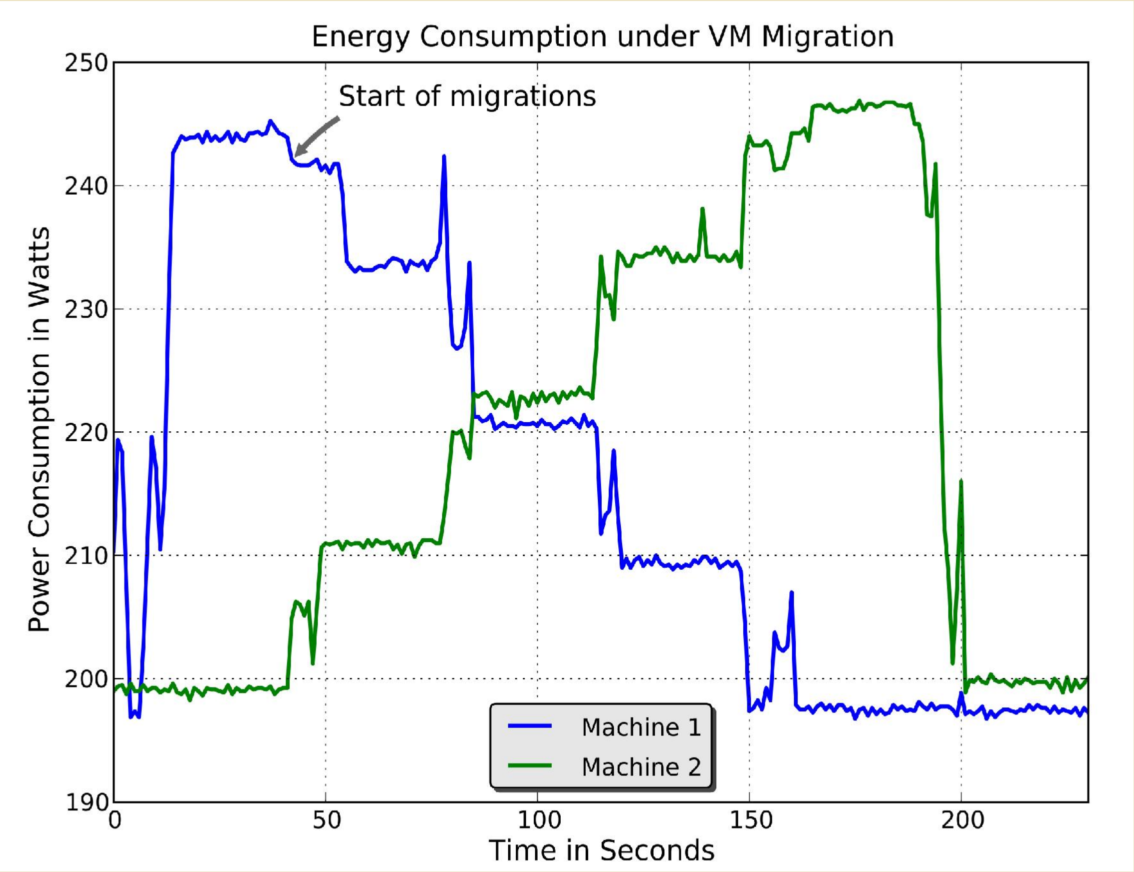
- 2 HP Proliant 85 G2 Servers (2.2 GHz, 2 dual core CPUs per node)
- XenServer 5.0
- VMs : - Debian Etch 4.0
  - 1 virtual CPU + 512 Mo
  - 6 VMs maximum per node

Boot, Run and Halt



6% increase of energy with 1 VM running

Migration

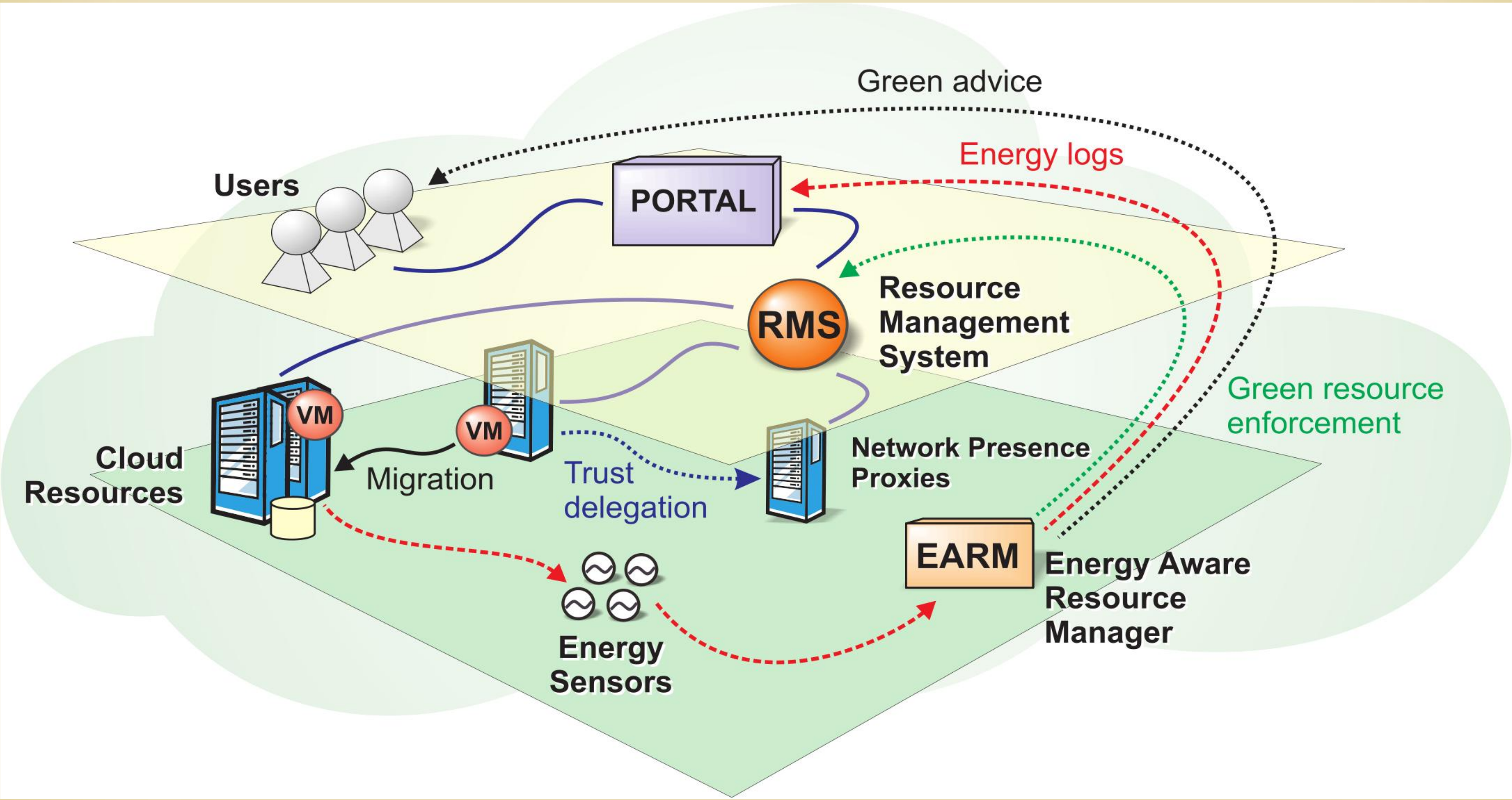


Bad moment in energy during the migration

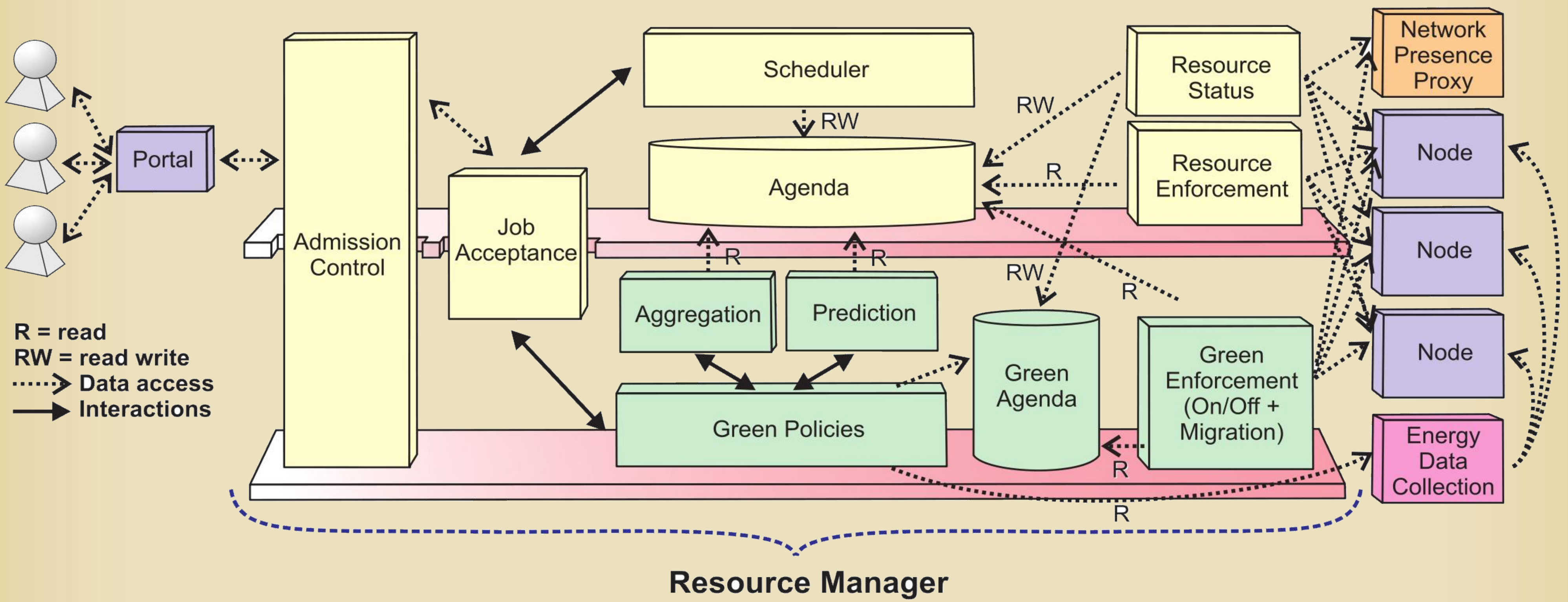
Green Open Cloud Features

- Switch off unused resources
- Predict usage
- Aggregate reservations (live migration)
- Green policies for the users
- Network presence proxy

GOC Architecture



GOC Resource Manager



- Smooth integration in Cloud infrastructure

Algorithm for the end of a reservation

Algorithm At the end of each job

ForEach totally or partially free resource M **Do**
  Predict the next use of M.
  If M will be used in less than T_s **Then**
    Nothing to do for M.
  Else
    If M is totally free **Then**
      Switch off M.
    Else
      If the job(s) of M will end in more than T_m **Then**
        Try to migrate the remaining VMs of M on other partially free resources to minimize the number of used resources.
      Else
        Nothing to do for M.

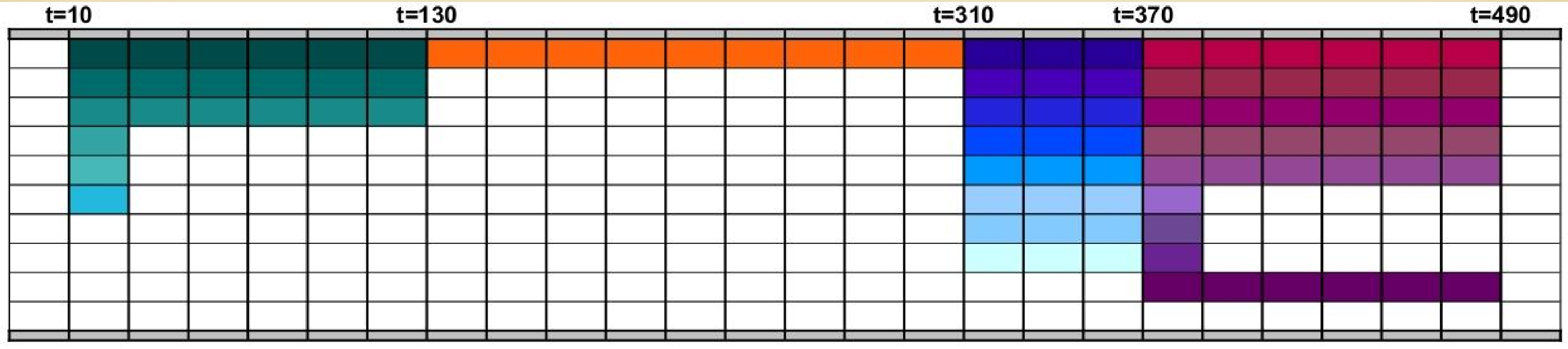
Experimental Methodology

<u>Cloud job arrival example:</u>

- t = 10: 3 jobs of 120 s. + 3 jobs of 20 s.
- t = 130: 1 job of 180 s.
- t = 310: 8 jobs of 60 s.
- t = 370: 5 jobs of 120 s. + 3 jobs of 20 s. + 1 job of 120 s.

→ limited time experiment

→ identical nodes



Experimental Methodology

- Two different simple **schedulings**: *round-robin* and *unbalanced.*
- Four **scenarios**:

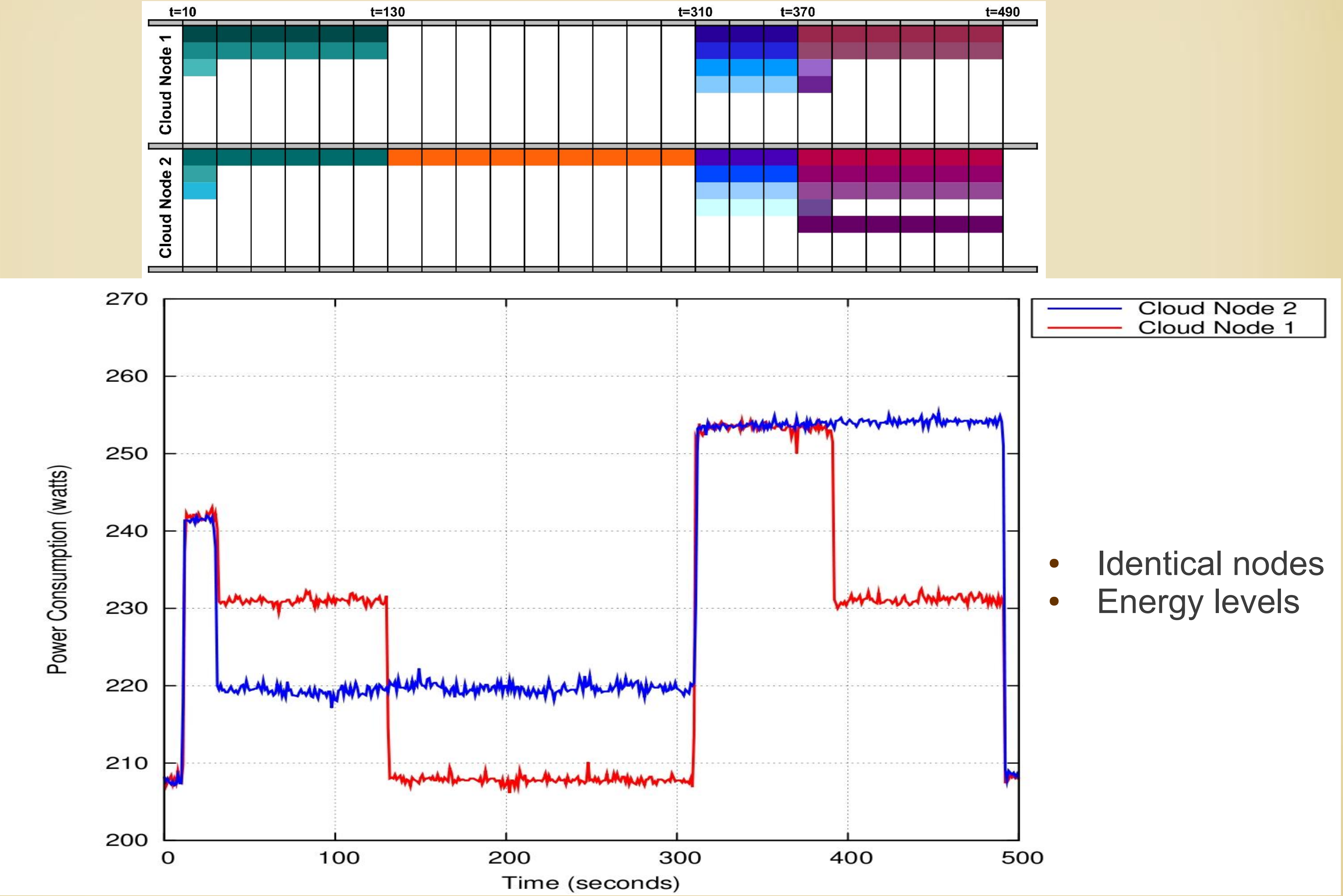
  - *basic*: nothing to do;

  - *balancing*: use migration to balance the load;

  - *on/off:* switch off unused nodes;

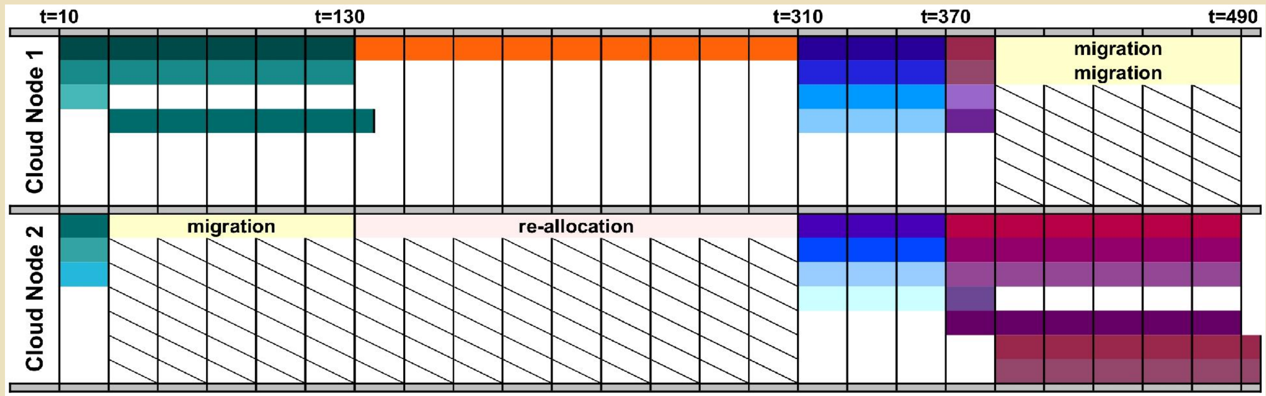
  - *green*: switch off unused nodes and use migration to unbalance the load.

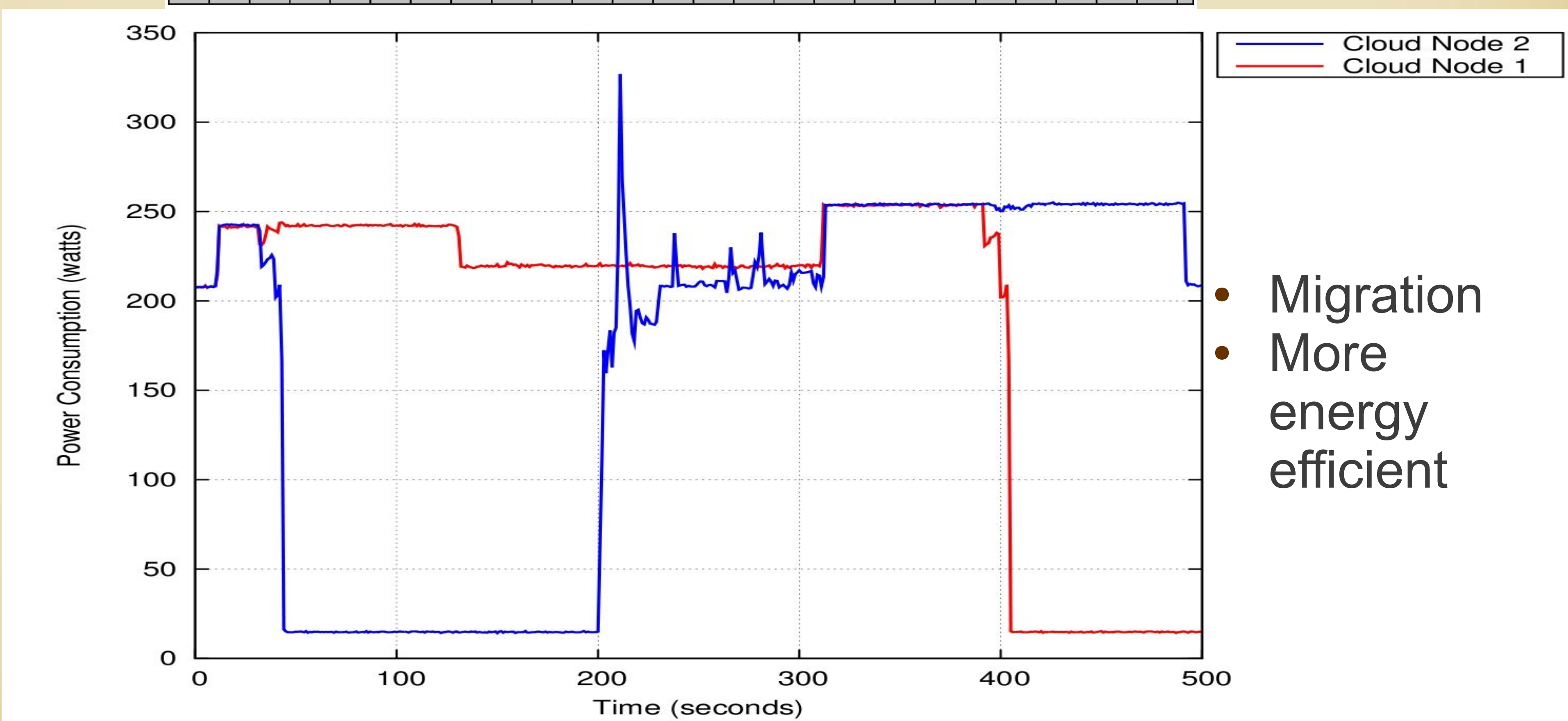
Round-Robin with Basic Scenario



- Identical nodes
- Energy levels

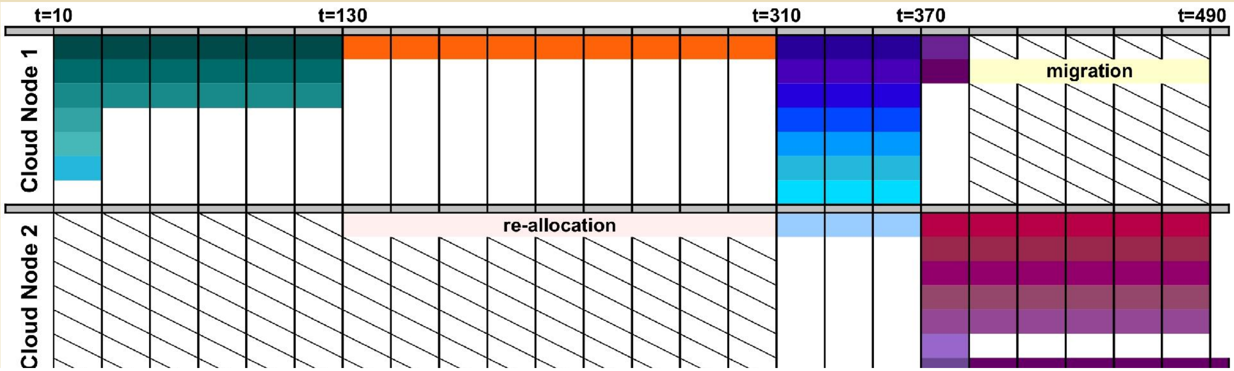
Round-Robin with Green Scenario

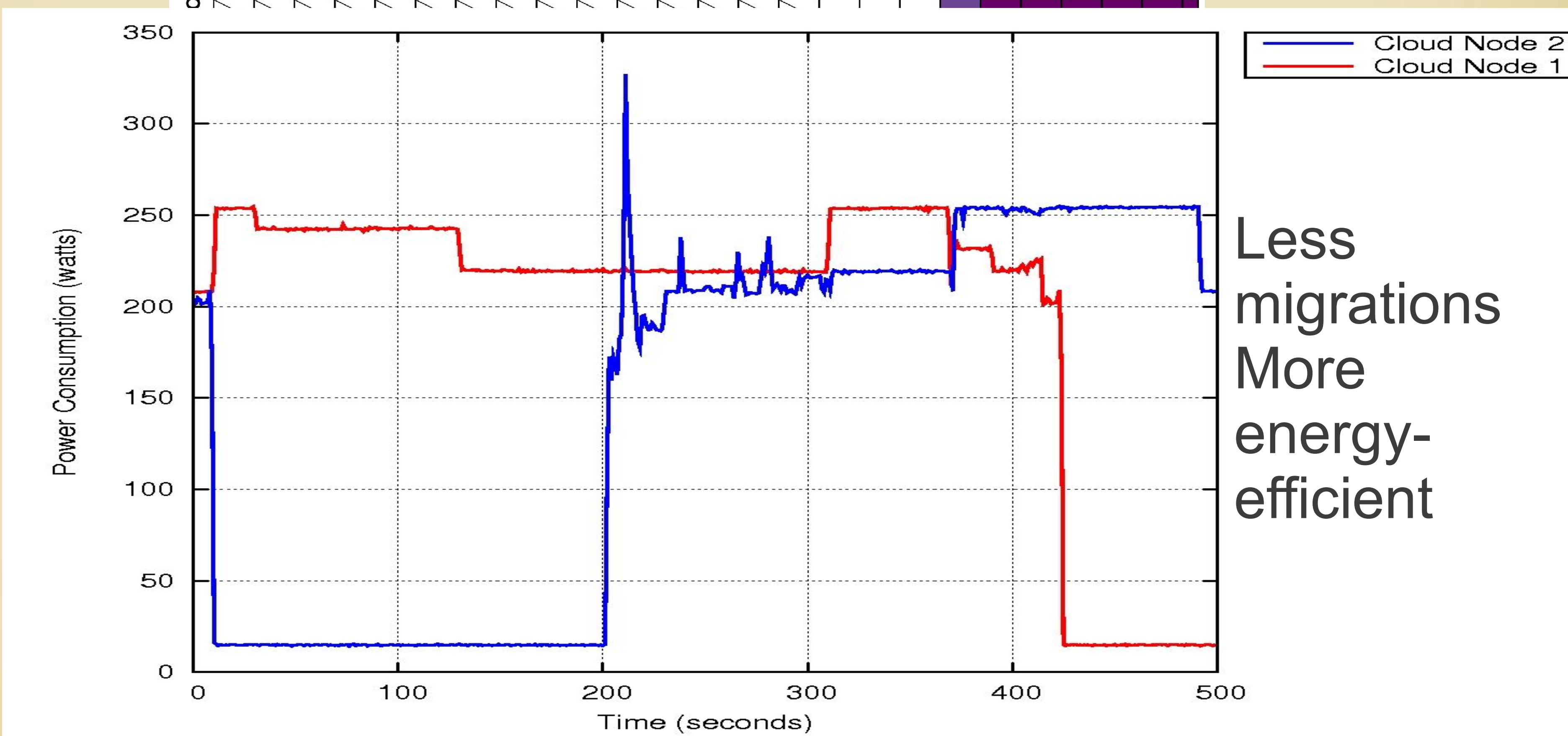




- Migration
- More energy efficient

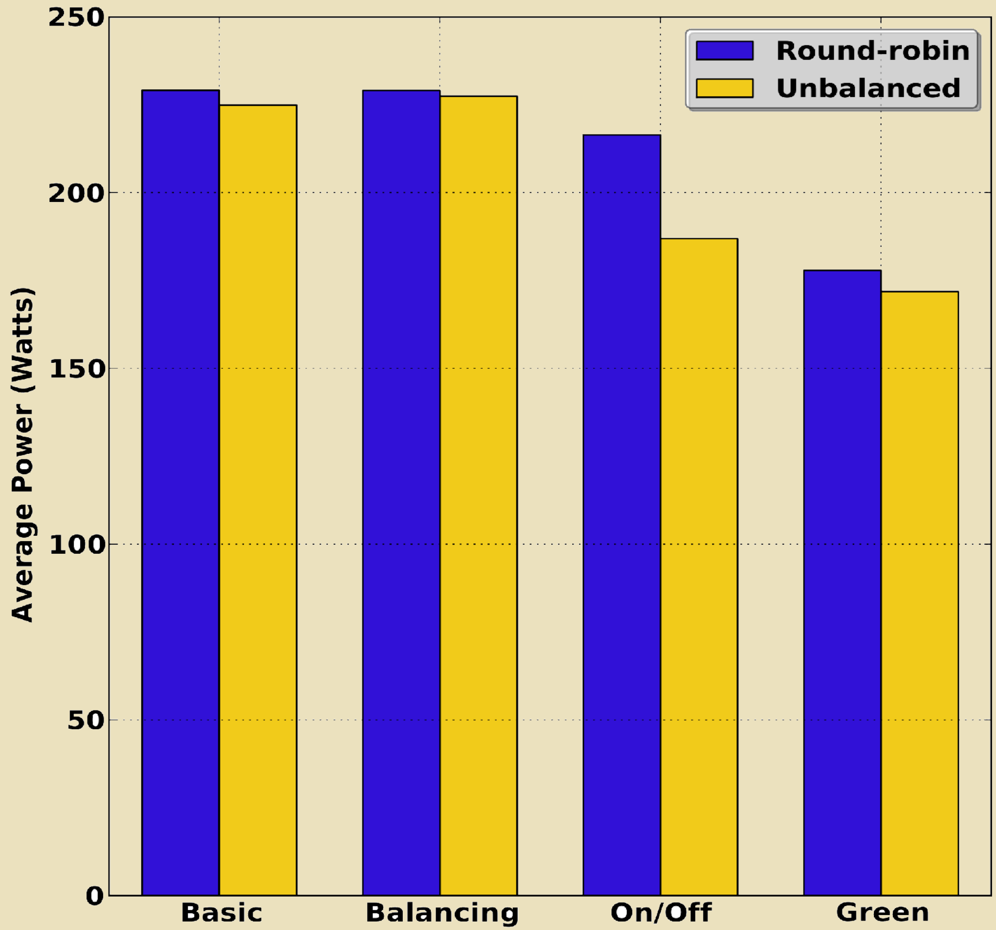
Unbalanced with Green Scenario





Less migrations
More energy-efficient

Comparison between the scenarios



Same execution time for all the experiments

Conclusions

- Analysis of energy usage of VM in Cloud context
- Test on real nodes leads to 25% of energy saved with GOC
- Significant energy savings are achievable.
- GOC can be integer in current and future Cloud infrastructures (with reservation, accounting, ...)

Publications

Anne-Cécile Orgerie, Marcos Dias de Assunção and Laurent Lefèvre.
"Energy Aware Clouds"
chapter in *Grids, Clouds and Virtualization*, pages 145-170, Springer Book, to appear in 2010.

Laurent Lefèvre and Anne-Cécile Orgerie.
"Designing and Evaluating an Energy Efficient Cloud"
Journal of Supercomputing, 51:352-373, March. 2010.

G. Da-Costa, M. Dias de Assunção, J.-P. Gelas, Y. Georgiou, L. Lefèvre, A.-C. Orgerie, J.-M. Pierson, O. Richard and A. Sayah.
"**Multi-Facet Approach to Reduce Energy Consumption in Clouds and Grids: the GREEN-NET Framework".**
E-Energy 2010, pages 95-104, April 2010.

Laurent Lefèvre and Anne-Cécile Orgerie.
"**When Clouds Become Green: the Green Open Cloud Architecture**",
ParCo 2009, pages 228-237, Sept. 2009.

Conclusions

Conclusions

- Issues about energy efficiency in Grids
- State of the art on energy saving techniques used in Grid environments
- Analysis on 6 months energy and reservation logs of a Grid'5000 site (experimental Grid).
- Proposition of EARI and validation on traces with measured parameters
- Proposition of GOC and validation on real nodes

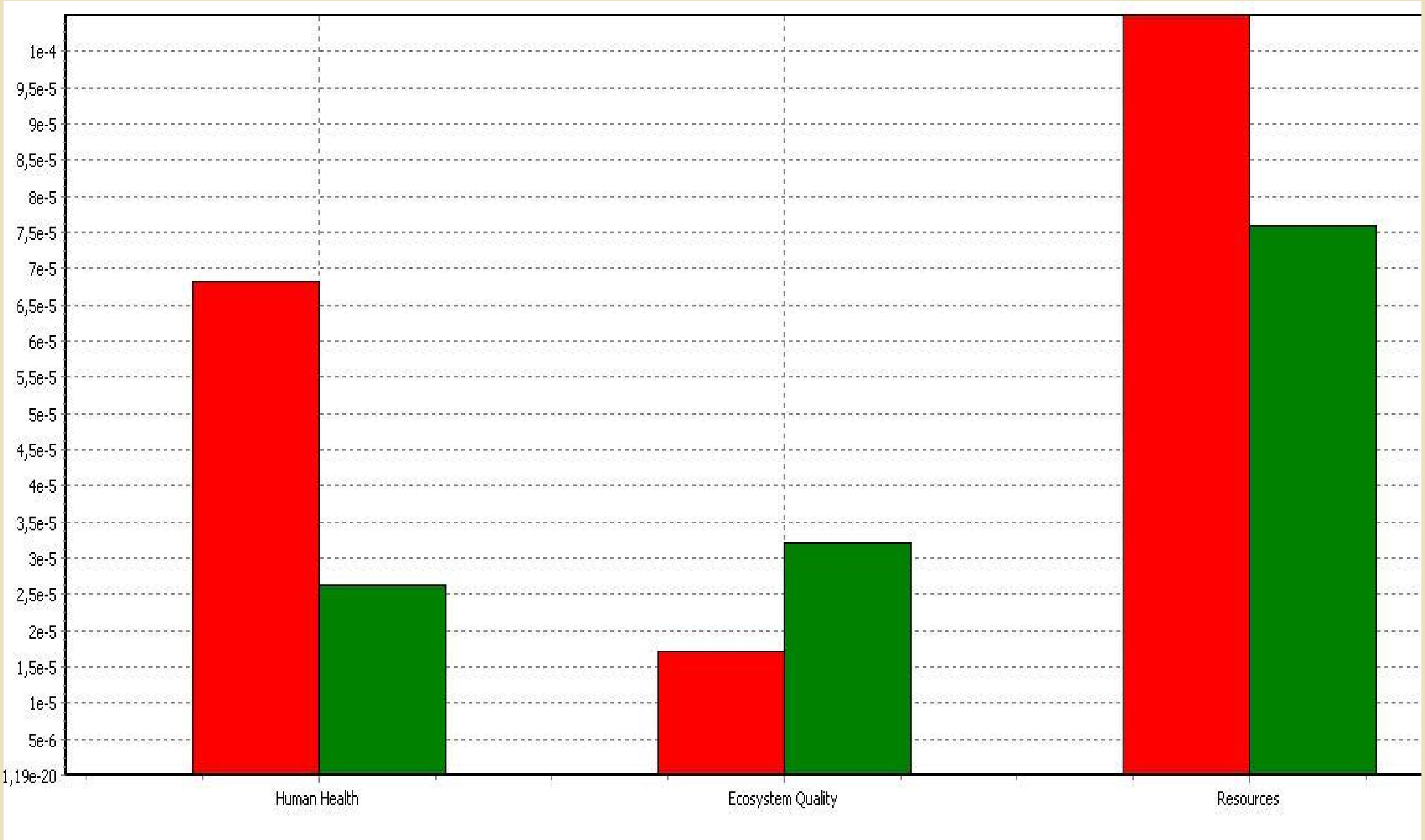
Bad good ideas

- Give 50 pages to read
- Two solutions:

  - usb key + 2 hours of computer and screen working

  - print on a color laser printer (front and back) on regular paper

Bad good ideas



Human health　　Ecosystem quality　　Resources

Bad good ideas (2)

- New scenario:

  500 pages

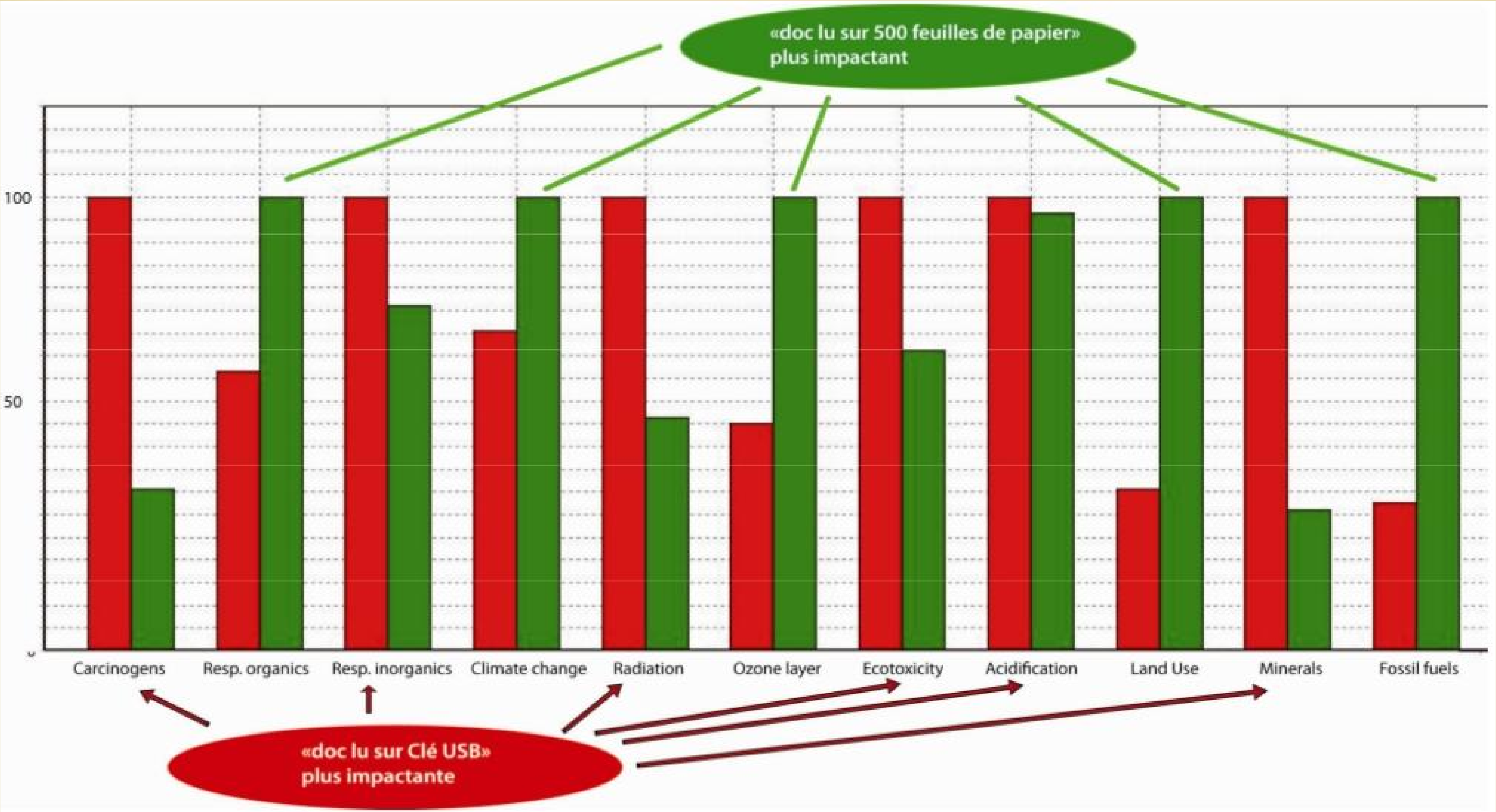
  usb key with 20h of computer working

EcoInfo

ADEME

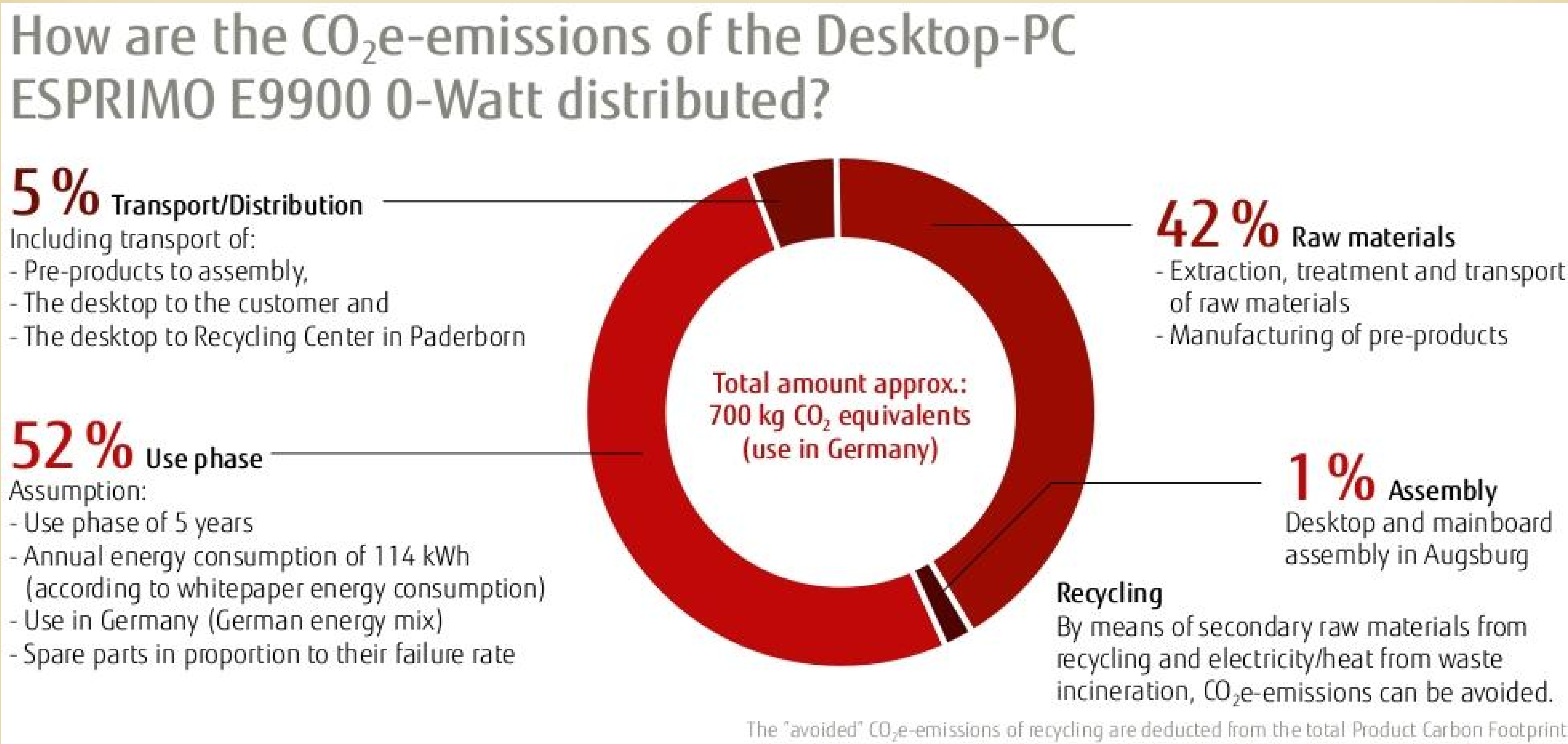
Agence de l'Environnement
et de la Maîtrise de l'Energie

Bad good ideas (2)



For 20h working, the impact of the computer is greater than the impact of the usb key.

Top of the iceberg



→ 48 years in France, 5 years in Germany without screen

Fujitsu, Fact Sheet: CO2-Footprint of a Desktop-PC ESPRIMO E9900

Top of the iceberg

- Production + Transport often dominates the carbon footprint
- Lots of other criteria to take into account
- Need to take the whole loop into account

  → Life Cycle Assessment
- Standards required

Thank you for your attention!

Questions?

annececile.orgerie@ens-lyon.fr

ENS DE LYON RESO INRIA RHÔNE-ALPES LIP UCBL CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE UNIVERSITÉ DE LYON