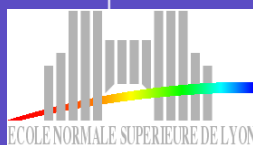


# Towards Energy Aware Resource Infrastructure for Large Scale Distributed Systems (GREEN-\*)

CCGSC 2008 – 16/9/2008 – Flat Rock, NC



INSTITUT NATIONAL  
DE RECHERCHE  
EN INFORMATIQUE  
ET EN AUTOMATIQUE



**Green-Net**

*Laurent Lefèvre, Anne-Cécile Orgerie, Jean-Patrick Gelas*

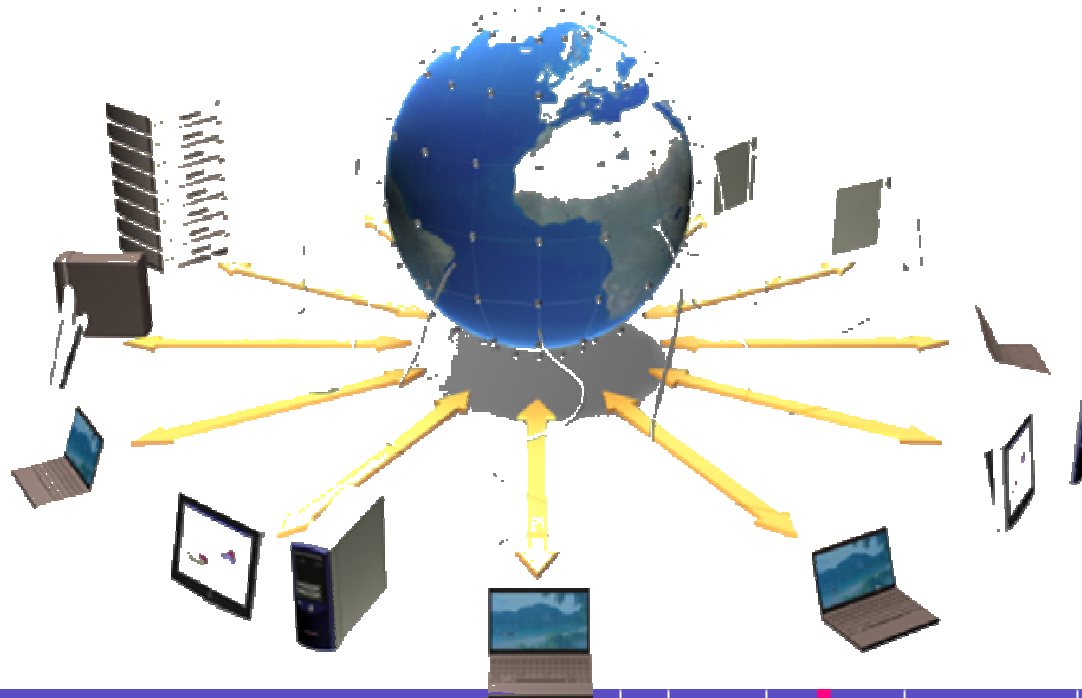
*INRIA RESO – University of Lyon - LIP (UMR CNRS, ENS, INRIA,  
UCB), France*

*[laurent.lefevre@inria.fr](mailto:laurent.lefevre@inria.fr)*

# Challenge : general estimations

Currently in the world:

- 1.5 billion of computers
- between 5 and 10% of the world electric consumption



# Example of Servers

- Electricity used for servers doubled over the period 2000 to 2005 both in the U.S. and worldwide
- Total power used by servers represented about 0.6% of total U.S. electricity consumption in 2005.
- When cooling and auxiliary infrastructure are included, that number grows to 1.2%, an amount comparable to that for color televisions.
- The total power demand in 2005 (including associated infrastructure) == five 1000 MW power plants for the U.S. and 14 such plants for the world.
- The total electricity bill for operating those servers and associated infrastructure in 2005 was about \$2.7 B and \$7.2 B for the U.S. and the world, respectively.

*“ESTIMATING TOTAL POWER CONSUMPTION BY SERVERS IN THE U.S. AND THE WORLD” Jonathan G. Koomey, Lawrence Berkeley National Laboratory and Consulting Professor, Stanford University February 15, 2007*

# Problem :

Bad usage (some French statistics ADEME / ENERTECH) about machines usage in companies :

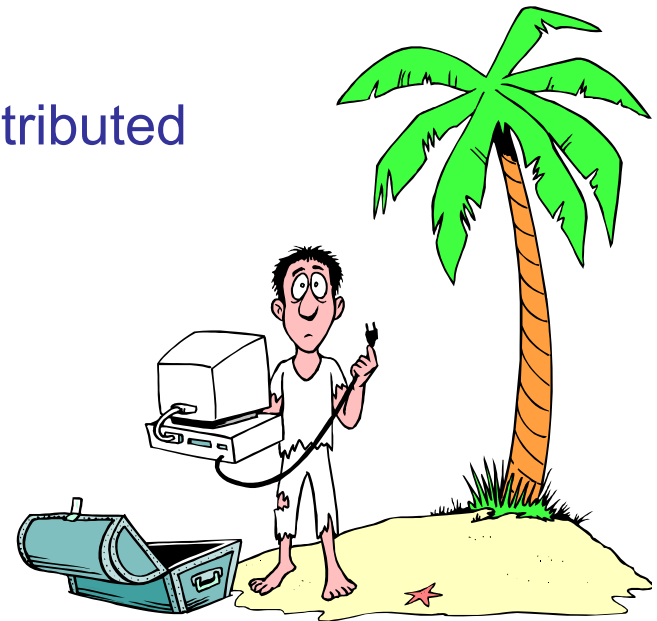
- CPU : 4004 h of running per year = 17.8 h per working day
- Screen : 2510 h per year = 11,2 h per working day

But users effectively use their machines :686 h per year = 3 hours a day

Computers are up and running  $\frac{3}{4}$  time for nothing !

# Roadmap

- Green-\* problems and contributions
  - End hosts
  - Clusters/Data centers
  - Grids
  - Internet and networks
- How to understand usage of large scale distributed system ?
- Proposition of an energy aware reservation infrastructure



# GREEN-\* approach

Derived from “Autonomic Computing” (IBM)

Like Self-\* patterns :

- self-managing
- self-configuring
- self-optimizing
- self-protecting
- self-healing/repairing
- ...

Green-\* : “when Power aware / Green intervention is possible”

Proposing set of alternatives solutions to address the energy/power consumption issues in distributed systems (plugged)



# Not covered by this talk

- Urgent / real-time applications and requirements

- Ex : SPRUCE : Special PRIORITY and Urgent Computing Environment

**Urgent Computing:  
I Need it Now!**

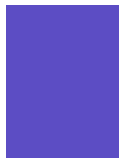
- Applications with dynamic data and **result deadlines** are being deployed
- Late results are useless
  - Wildfire path prediction
  - Storm/Flood prediction
  - Influenza modeling
- Some jobs need priority access  
**"Right-of-Way Token"**



Argonne Nat'l Lab/U Chicago SPRUCE Urgent Computing Flat Rock, North Carolina, 2006  
<http://www.mcs.anl.gov/~beckman>

Green, green, green : Services / Servers Hosting

green ISP

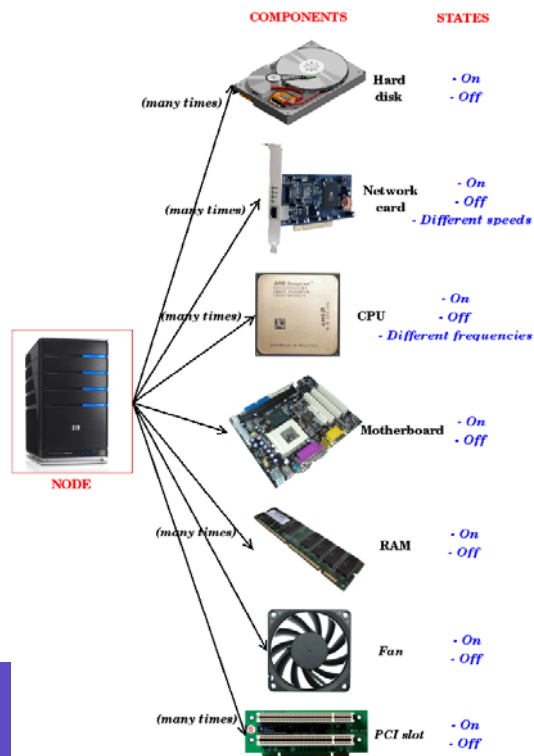
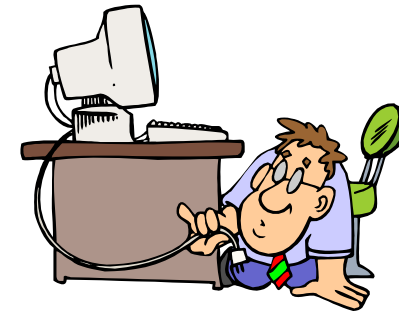


# What you do on your machine -> how much energy you use

Power usage per device, per process, per service

Tools like **powertop** for Linux machines

<http://www.lesswatts.org/projects/powertop/>



```
File Edit View Terminal Go Help
PowerTOP version 1.8 (C) 2007 Intel Corporation

Cn      Avg residency      P-states (frequencies)
C0 (cpu running)  (12.9%)            1.71 Ghz    9.8%
C1      0.0ms ( 0.0%)      1200 Mhz   0.3%
C2      10.7ms (87.1%)     800 Mhz    0.5%
C3      0.0ms ( 0.0%)      600 Mhz    89.4%
C4      0.0ms ( 0.0%)

Wakeups-from-idle per second : 81.2      interval: 15.0s
Power usage (ACPI estimate): 14.1W (6.6 hours) (long term: 136.4W,/0.7h)

Top causes for wakeups:
34.4% ( 31.9)  <interrupt> : ipw2200, Intel 82801DB-ICH4, Intel 82801DB-
19.4% ( 18.0)  firefox-bin : futex_wait (hrtimer_wakeup)
15.5% ( 14.4)  X : do_setitimer (it_real_fn)
11.5% ( 10.7)  evolution : schedule_timeout (process_timeout)
4.3% ( 4.0)   <kernel module> : usb_hcd_poll_rh_status (rh_timer_func)
3.9% ( 3.6)   <interrupt> : libata
1.8% ( 1.7)   <kernel core> : sk_reset_timer (tcp_delack_timer)
1.2% ( 1.1)   X : schedule_timeout (process_timeout)
1.1% ( 1.0)   Terminal : schedule_timeout (process_timeout)
1.1% ( 1.0)   xfce4-panel : schedule_timeout (process_timeout)
0.6% ( 0.5)   <kernel module> : neigh_table_init_no_netlink (neigh_periodic
0.5% ( 0.5)   spamd : schedule_timeout (process_timeout)
0.5% ( 0.5)   events/0 : ipw_gather_stats (delayed_work_timer_fn)
0.4% ( 0.3)   xfdesktop : schedule_timeout (process_timeout)
0.4% ( 0.3)   firefox-bin : sk_reset_timer (tcp_write_timer)
0.3% ( 0.3)   nscd : futex_wait (hrtimer_wakeup)
0.2% ( 0.2)   xscreensaver : schedule_timeout (process_timeout)
0.2% ( 0.2)   ksnapshot : schedule_timeout (process_timeout)

Suggestion: Disable the unused bluetooth interface with the following command:
hciconfig hci0 down ; rmmod hci_usb
Bluetooth is a radio and consumes quite some power, and keeps USB busy as well.
Q - Quit | R - Refresh | B - Turn Bluetooth off
```



# For cluster / data centers : Green 500

Green 500 : the energy aware TOP 500

Supported by Virginia Tech (US)

<http://green500.org>



Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)	TOP500 Rank*
1	357.23	Science and Technology Facilities Council - Daresbury Laboratory	Blue Gene/P Solution	31.10	121
2	352.25	Max-Planck-Gesellschaft MPI/IPP	Blue Gene/P Solution	62.20	40
3	346.95	IBM - Rochester	Blue Gene/P Solution	124.40	24
4	336.21	Forschungszentrum Juelich (FZJ)	Blue Gene/P Solution	497.60	2
5	310.93	Oak Ridge National Laboratory	Blue Gene/P Solution	70.47	41
6	210.56	Harvard University	eServer Blue Gene Solution	44.80	170
7	210.56	High Energy Accelerator Research Organization /KEK	eServer Blue Gene Solution	44.80	171
8	210.56	IBM - Almaden Research Center	eServer Blue Gene Solution	44.80	172
9	210.56	IBM Research	eServer Blue Gene Solution	44.80	173
10	210.56	IBM Thomas J. Watson Research Center	eServer Blue Gene Solution	44.80	174
11	210.56	Renaissance Computing Institute (RENCI)	eServer Blue Gene Solution	44.80	175
12	210.56	University of Canterbury	eServer Blue Gene Solution	44.80	176
13	208.31	Forschungszentrum Juelich (FZJ)	eServer Blue Gene Solution	179.20	28
14	208.31	Computational Biology Research Center, AIST	eServer Blue Gene Solution	89.60	52
15	208.31	EDF R&D	eServer Blue Gene Solution	89.60	53
16	208.31	Ecole Polytechnique Federale de Lausanne	eServer Blue Gene Solution	89.60	54
17	208.31	High Energy Accelerator Research	eServer Blue Gene Solution	89.60	55

# Green 500 : what about the most using ones ?

Green500	Site	Computer	Country	Mflops Per Wa	Total Power	TOP500
494	The Earth Simulator Center	Earth-Simulator	Japan	5,6	6400	30
498	SARA (Stichting Academisch Rekencentrum)	eServer pSeries 575, p5+ 1.9 GHz, Infiniband	Netherlands	2,3	4992	120
496	Lawrence Livermore National Laboratory	Intel Itanium2 Tiger4 1.4GHz - Quadrics	United States	4,06	4915,2	47
499	Aerospace Company (E)	eServer pSeries 575, p5+ 1.9 GHz, Infiniband	France	2,3	3827,2	210
497	Los Alamos National Laboratory	ASCI Q - AlphaServer SC45, 1.25 GHz	United States	3,65	3800	91
21	DOE/NNSA/LLNL	eServer Blue Gene Solution	United States	166,6	2870,4	1
500	Aerospace Company (E)	eServer pSeries 575, p5+ 1.9 GHz, Infiniband	France	2,3	2704	450
390	DOE/NNSA/LLNL	eServer pSeries p5 575 1.9 GHz	United States	28,64	2645,07	11
451	University of Edinburgh	Cray XT4, 2.8 GHz	United Kingdom	21,01	2600,82	17
444	Leibniz Rechenzentrum	Altix 4700 1.6 GHz	Germany	22	2569,19	15
448	NNSA/Sandia National Laboratories	PowerEdge 1850, 3.6 GHz, Infiniband	United States	21,36	2481,6	18
450	Wright-Patterson Air Force Base/DoD ASC	Altix 4700 1.6 GHz	United States	21,13	2433,97	21
247	Oak Ridge National Laboratory	Cray XT4/XT3	United States	43,42	2341,99	7
274	NERSC/LBNL	Cray XT4, 2.6 GHz	United States	37,2	2295	9
238	NNSA/Sandia National Laboratories	Sandia/ Cray Red Storm, Opteron 2.4 GHz du	United States	45,63	2239,83	6
407	NASA/Ames Research Center/NAS	SGI Altix 1.5 GHz, Voltaire Infiniband	United States	26,59	1950,4	20
413	Commissariat a l'Energie Atomique (CEA)	Novascale 3045, Itanium2 1.6 GHz, Infiniband	France	24,93	1689,6	26
492	HWW/Universitaet Stuttgart	SX8/576M72	Germany	6,2	1440	202
495	NCSA	TeraGrid, Itanium2 1.3/1.5 GHz, Myrinet	United States	5,08	1419,2	351
78	Computational Research Laboratories, TATA SONS	Cluster Platform 3000 BL460c, Xeon 53xx 3G	India	83,94	1404,53	4
84	Government Agency	Cluster Platform 3000 BL460c, Xeon 53xx 2.6	Sweden	75,92	1354,03	5
58	SGI/New Mexico Computing Applications Center (NMCAC)	SGI Altix ICE 8200, Xeon quad core 3.0 GHz	United States	99,61	1274	3
239	GSIC Center, Tokyo Institute of Technology	Sun Fire x4600 Cluster, Opteron 2.4/2.6 GHz	Japan	45,53	1239,3	16
491	Pacific Northwest National Laboratory	Cluster Platform 6000 rx2600 Itanium2 1.5 GHz	United States	7,43	1161,6	222
357	Lawrence Livermore National Laboratory	Appro Xtreme Server - Quad Opteron Dual Co	United States	31,79	1152	29
493	Joint Supercomputer Center	MVS-15000BM, eServer BladeCenter JS20 (P	Russia	5,79	1148	408
477	CINECA	eServer 326 Cluster, Opteron Dual Core 2.6	Italy	14,98	1052,16	71
248	National Supercomputer Centre (NSC)	Cluster Platform 3000 DL140 Cluster, Xeon 2.	Sweden	42,75	1040	23
485	Cray Inc.	Cray XT4, 1.8 GHz	United States	13,01	988,16	102
489	Financial Institution (M)	Cluster Platform 4000 BL685c, Dual Core Opt	United States	10,51	975,58	145
467	Los Alamos National Laboratory	xSeries x3755 Cluster Opteron dual core 2.2	United States	16,28	864	86
360	Lawrence Livermore National Laboratory	Appro Xtreme Server - Quad Opteron Dual Co	United States	31,69	864	38

# For Grids : Green Grid



Big consortium mixing industrials and academics

Mission :

- defining meaningful, user-centric models and metrics;
- developing standards, measurement methods, processes and new technologies to improve data center performance against the defined metrics;
- promoting the adoption of energy efficient standards, processes, measurements and technologies.

<http://thegreengrid.org>

Geography / location aware load balancing :

- Inside the cluster : distributing the heat in the machine room
- On the grid : balancing the load

# Green-\* in networks ?

- In Networks:

- Suspending / Launching network interfaces
- « DVS on interfaces »



- Impact on Routers and LAN switches

- Green Internet (Univ. of Florida) : computational costs of network protocols (like TCP)

- Energy Efficient Ethernet Study Group (Nov. 2006) : IEEE 802.3az task force. Standardization of Adaptive Link Rate (ALR) protocols

- Goal : reduce power consumption during periods of low link usage (10M/100M/1G/10G)

# Towards energy aware platforms : open questions

## How to reduce energy without compromising QoE : Quality of Experiment ?

- How to understand and to analyze the usage of large scale platforms?
- How to apply energy usage models on this experimental usage ?
- How to monitor lively such usage (multiple views (Grids, datacenters, clusters, nodes, services, processes, threads)) ?
- How to design energy aware solutions ?

# Our current approach in one slide

- To analyze and understand how Grids are used
- To analyze and understand the life of a Grid node
- Collect energy usage information
- Expose information to users/middleware
- Inject energy information in distributed information systems
- Adapt schedulers to benefit of energy collection
- Propose alternative usage models
  - ON/OFF models / DVS models
- Assume net presence / virtualization / trust
- Validate on real platforms :
  - Grid5000 (high performance distributed system : 4K nodes)
  - DSLLAB (low performance « home » system : 45 nodes)

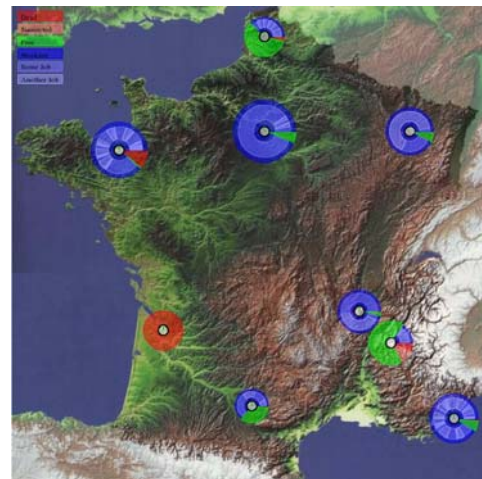
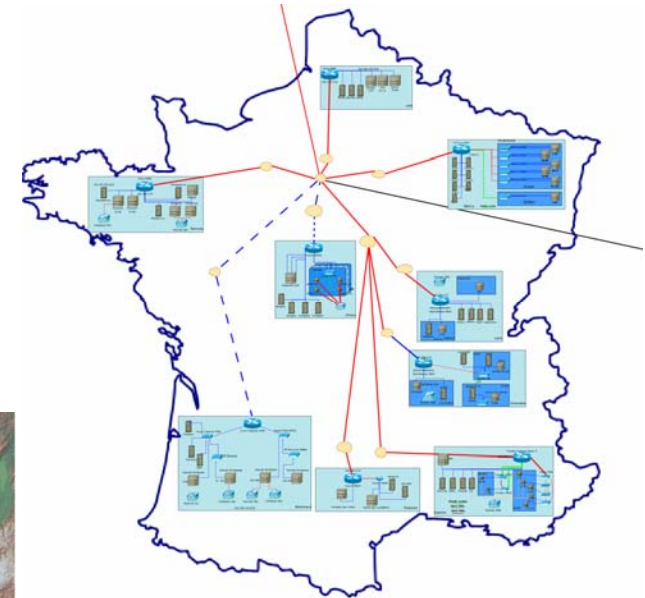
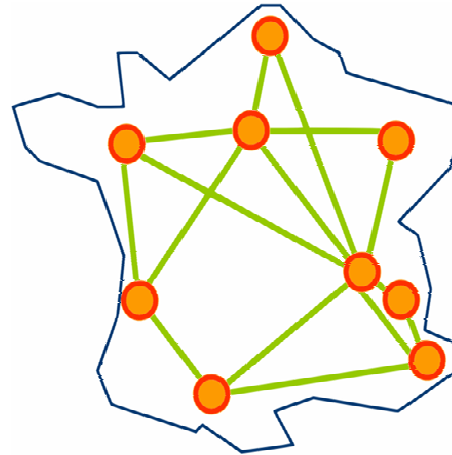
# How an experimental platform is used : the french Grid5000 case

Experimental testbed for research

- 9 sites geographically distributed in France

- 4000 processors

- Usage : Nodes reservation, image deployment, node reboot, exclusive usage of reserved nodes



# Analysis of global usage

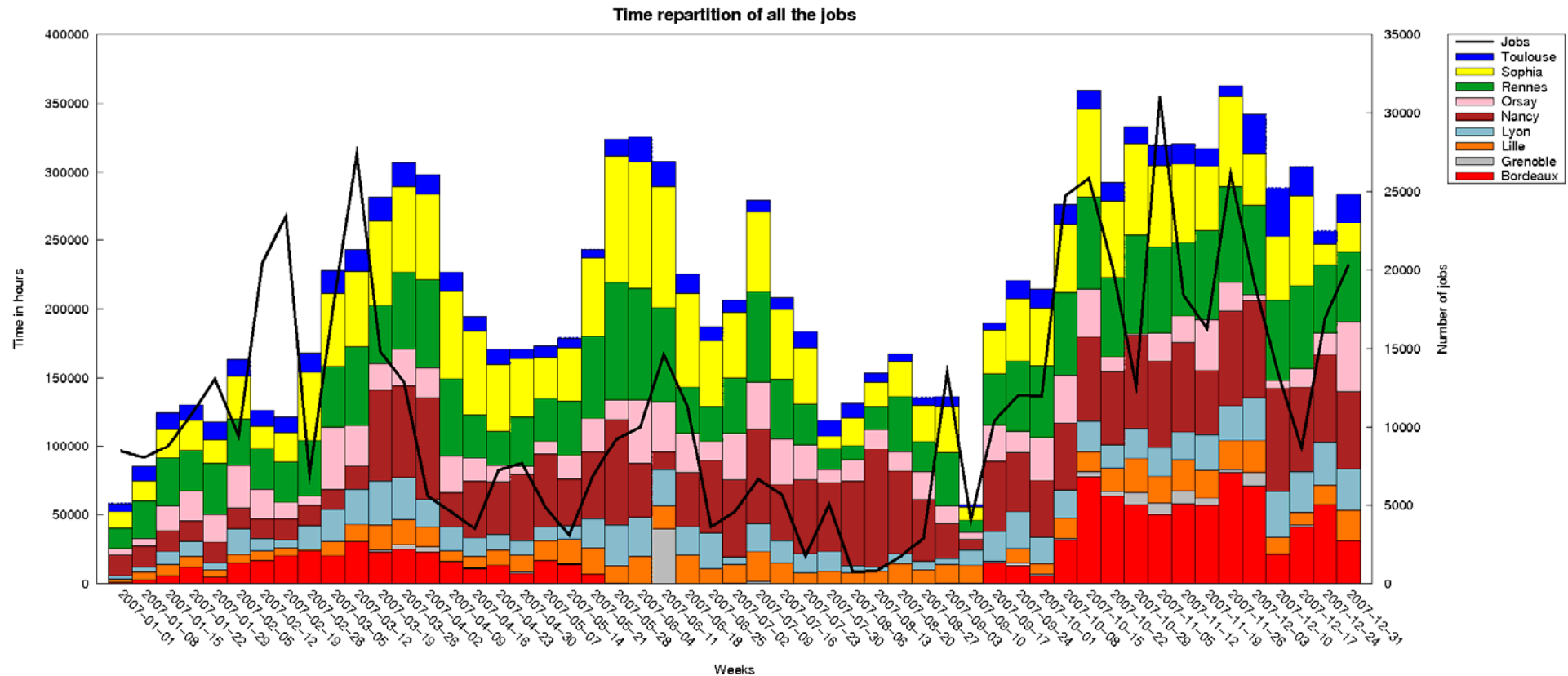
- Using batch scheduler reservation system (OAR)
- 1.2 Gbytes of reservation logs collected for 12 months period (2007) - 600 000 reservation events

Site	nb of reservations	nb of cores	nb of core per reservation	mean length of a reservation	real work
Bordeaux	45775	650	55.50	5224.59 s.	47.80%
Lille	330694	250	4.81	1446.13 s.	36.44%
Lyon	33315	322	41.64	3246.15 s.	46.38%
Nancy	63435	574	22.46	19480.49 s.	56.41%
Orsay	26448	684	47.45	4322.54 s.	18.88%
Rennes	36433	714	54.85	7973.39 s.	49.87%
Sophia	35179	568	57.93	4890.28 s.	51.43%
Toulouse	20832	434	12.89	7420.07 s.	50.57%

Some low average value but burst support - In operational Grids :  
60% to 70% average usage



# Usage over one year period

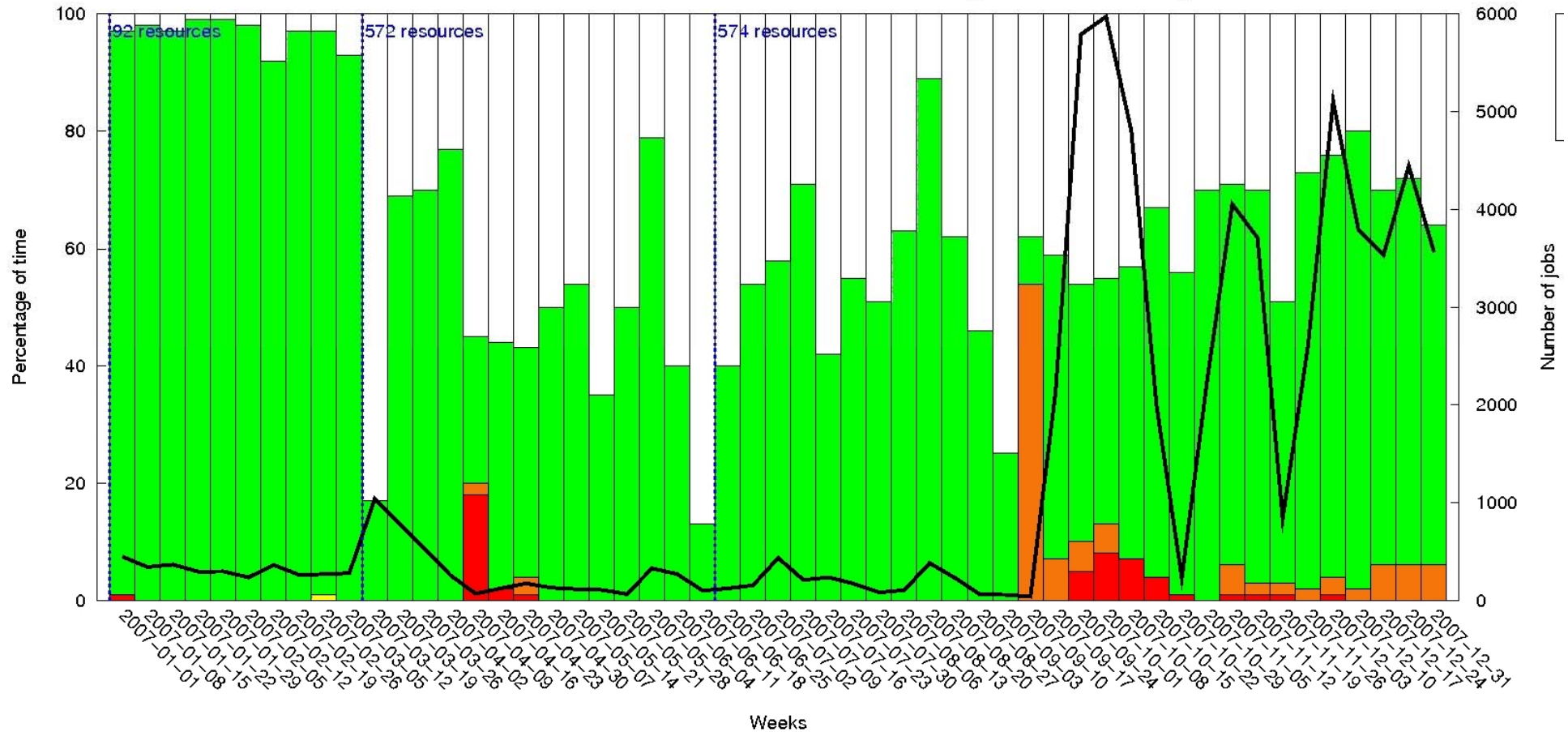


Global usage is not enough -> need more precise views

# Zoom on a site: Nancy



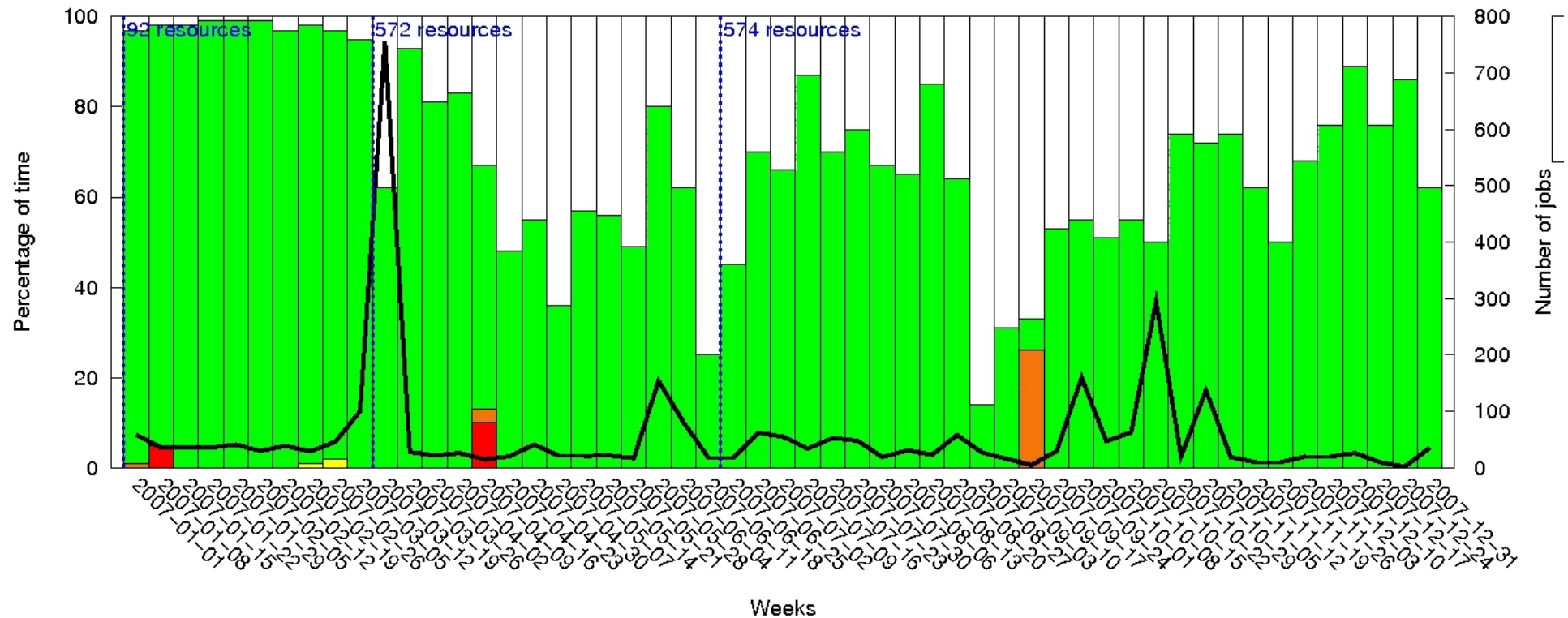
Distribution in time of the different resource's states per week for Nancy



# Zoom on a node : maximal resource



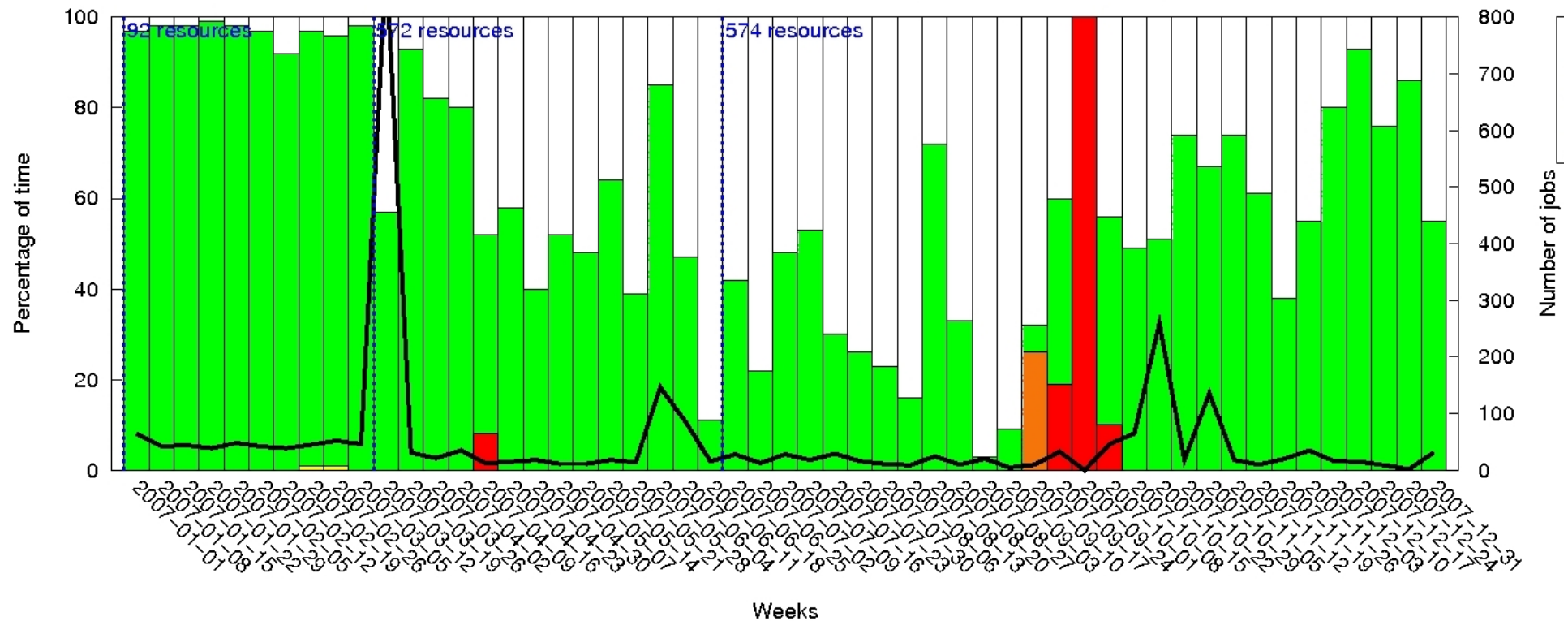
Distribution in time of the maximal resource's states per week for Nancy – number 67



# Zoom on a node : median resource



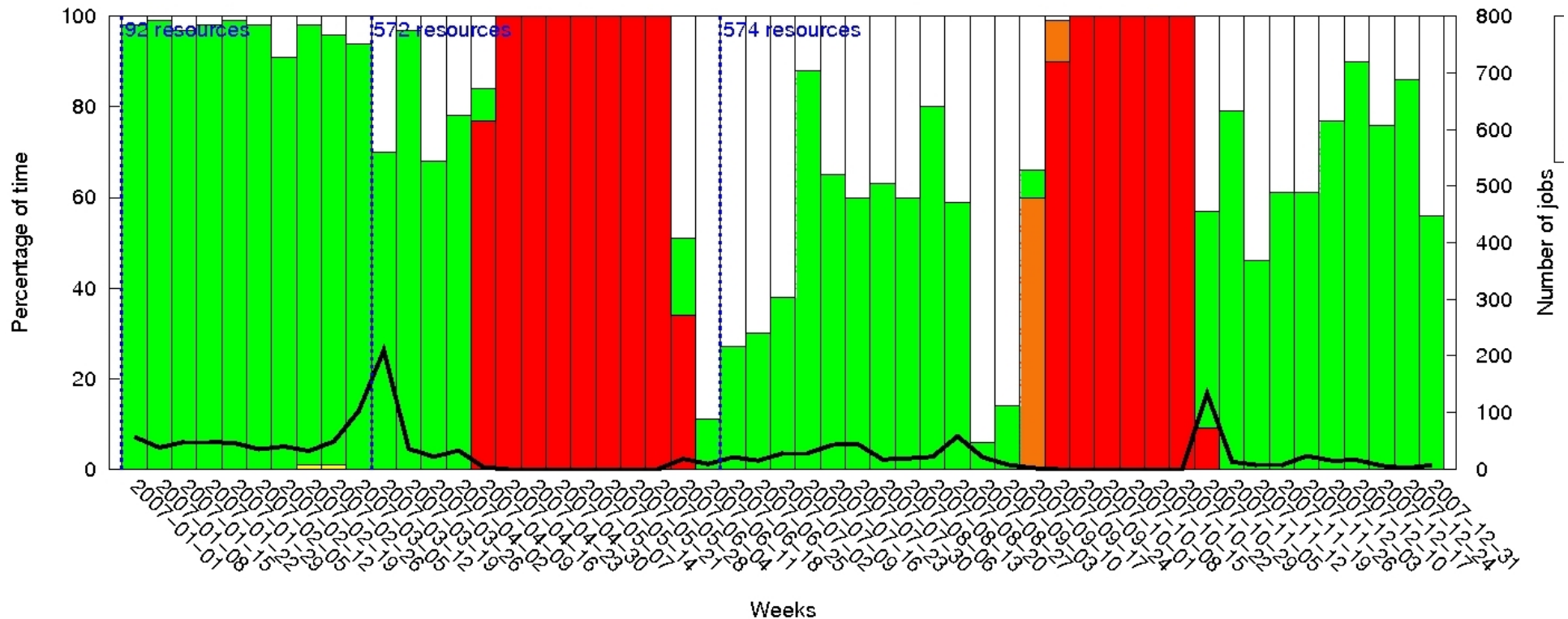
Distribution in time of the median resource's states per week for Nancy – number 29



# Zoom on a node : minimal resource



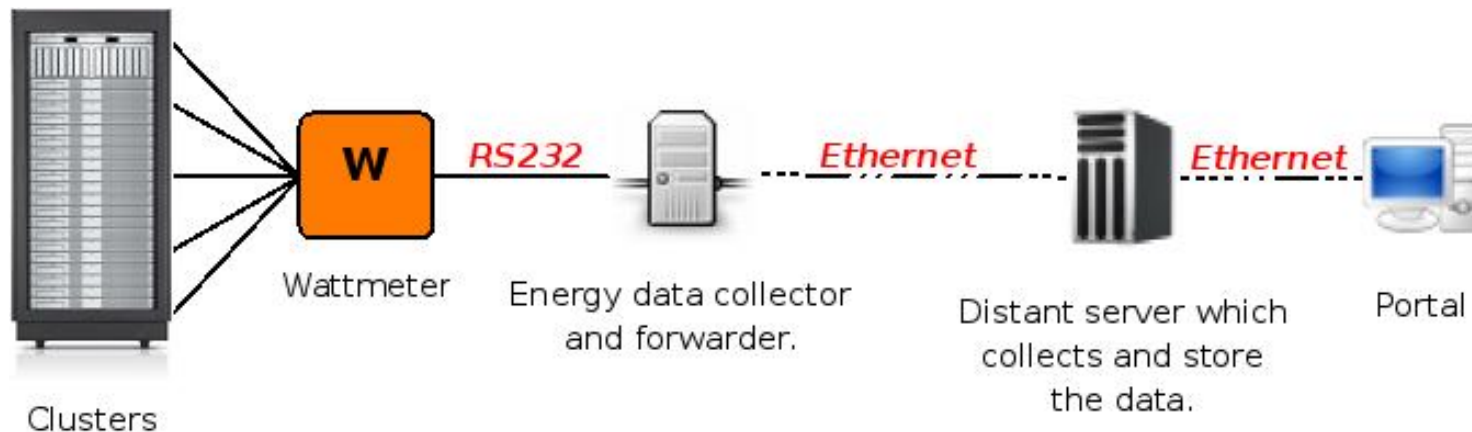
Distribution in time of the minimal resource's states per week for Nancy – number 46



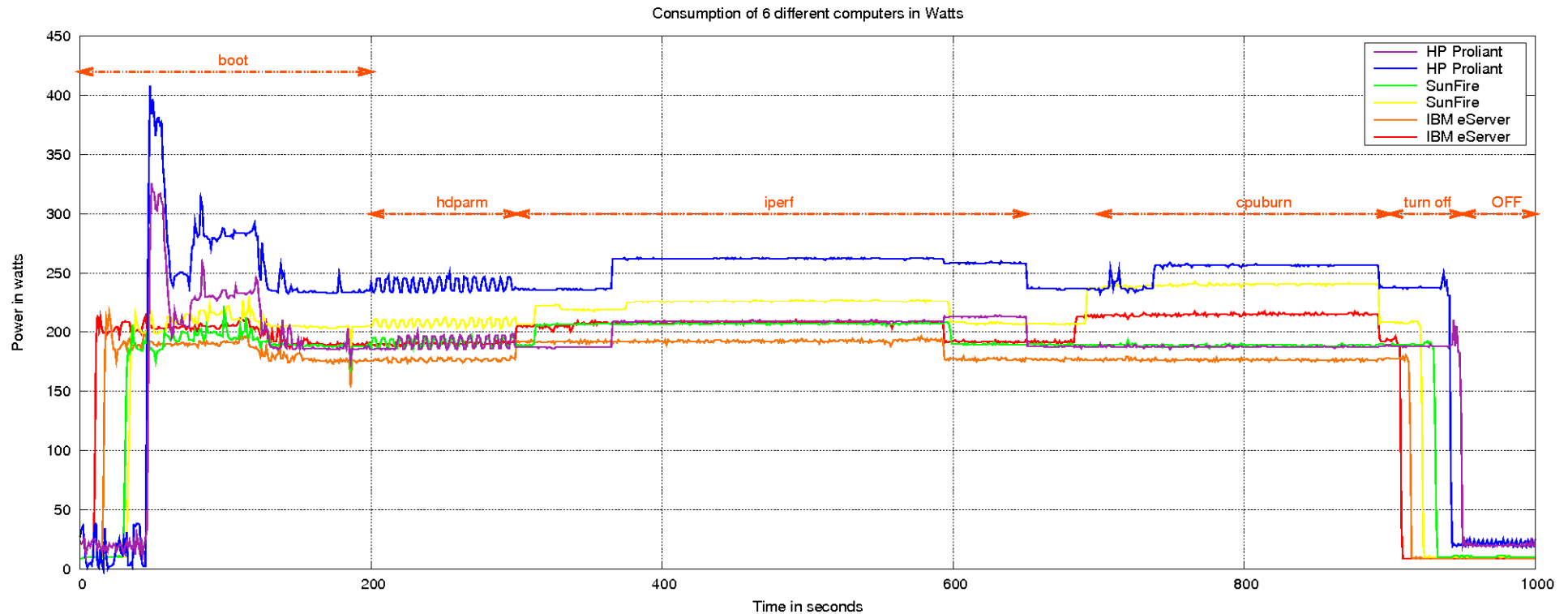
More precise view : Monitoring energy usage of a node

# Consumption measurement and collection

- Dynamic and autonomic
  - Controllable and external
- Energy sensor

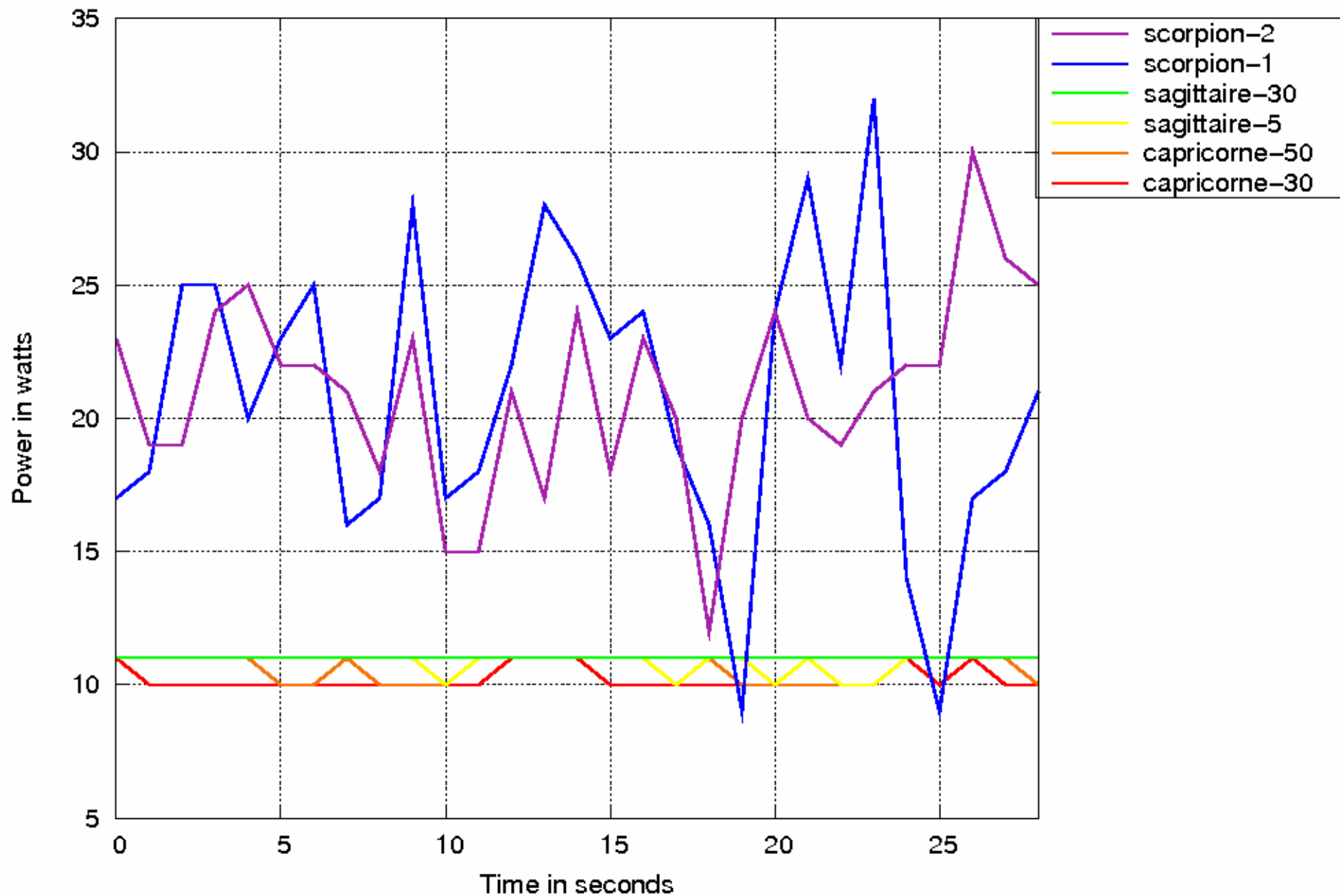


# Energy usage of Lyon nodes



- IBM eServer 325 (2.0GHz, 2 CPU)
- Sun Fire v20z (2.4GHz, 2 CPU)
- HP Proliant 385 G2 (2.2GHz, 2 CPU bicore)

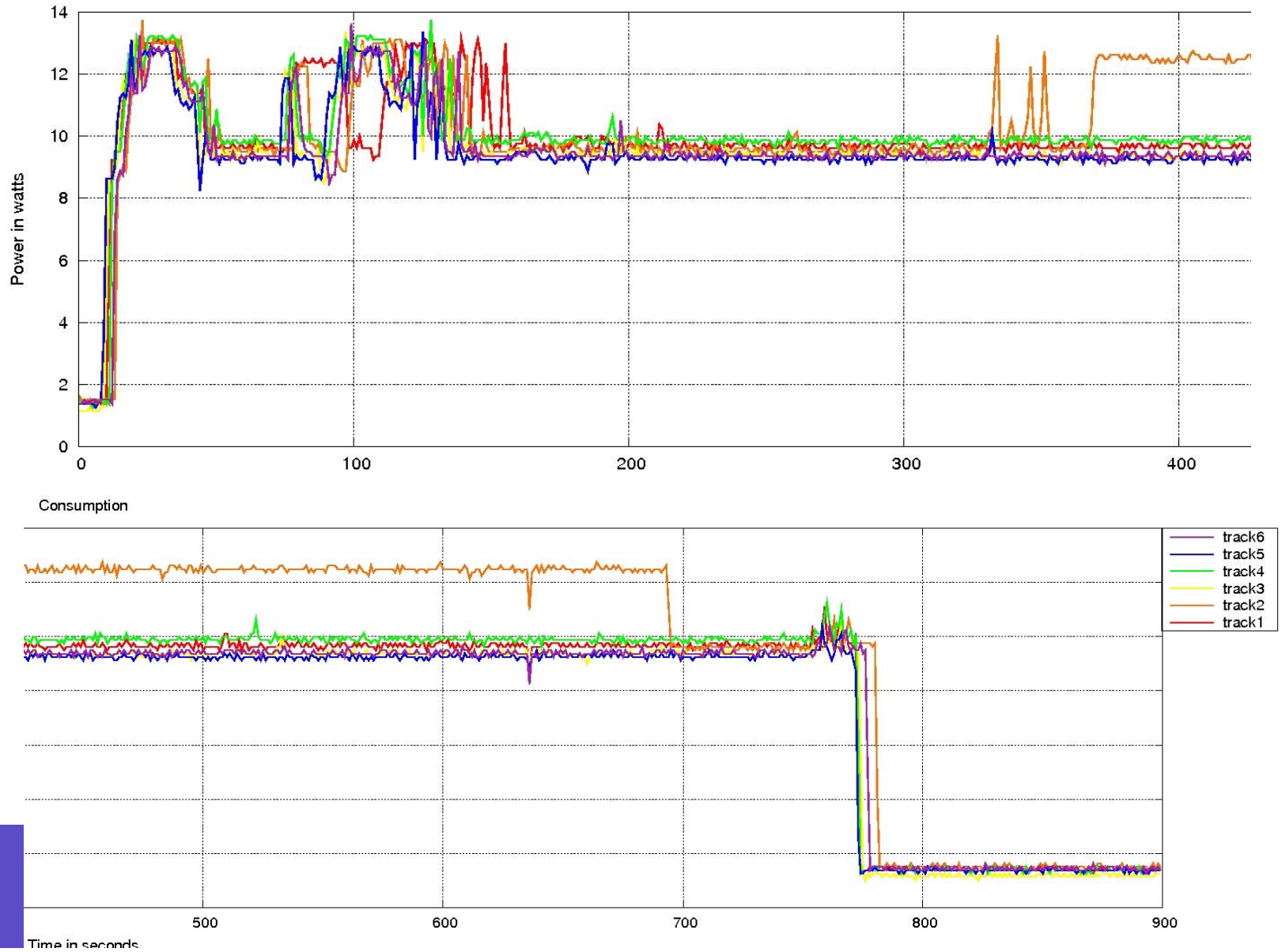
Consumption when PCs sleep







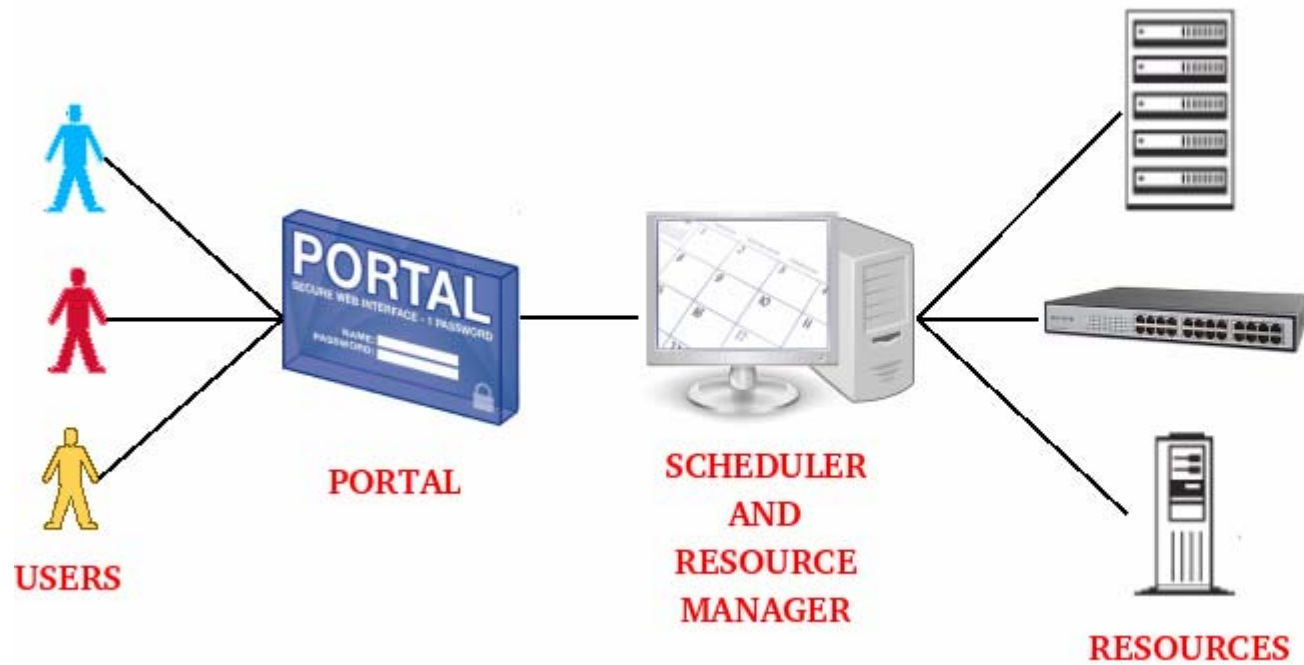
## Comparison with low power nodes (DSL-LAB platform)



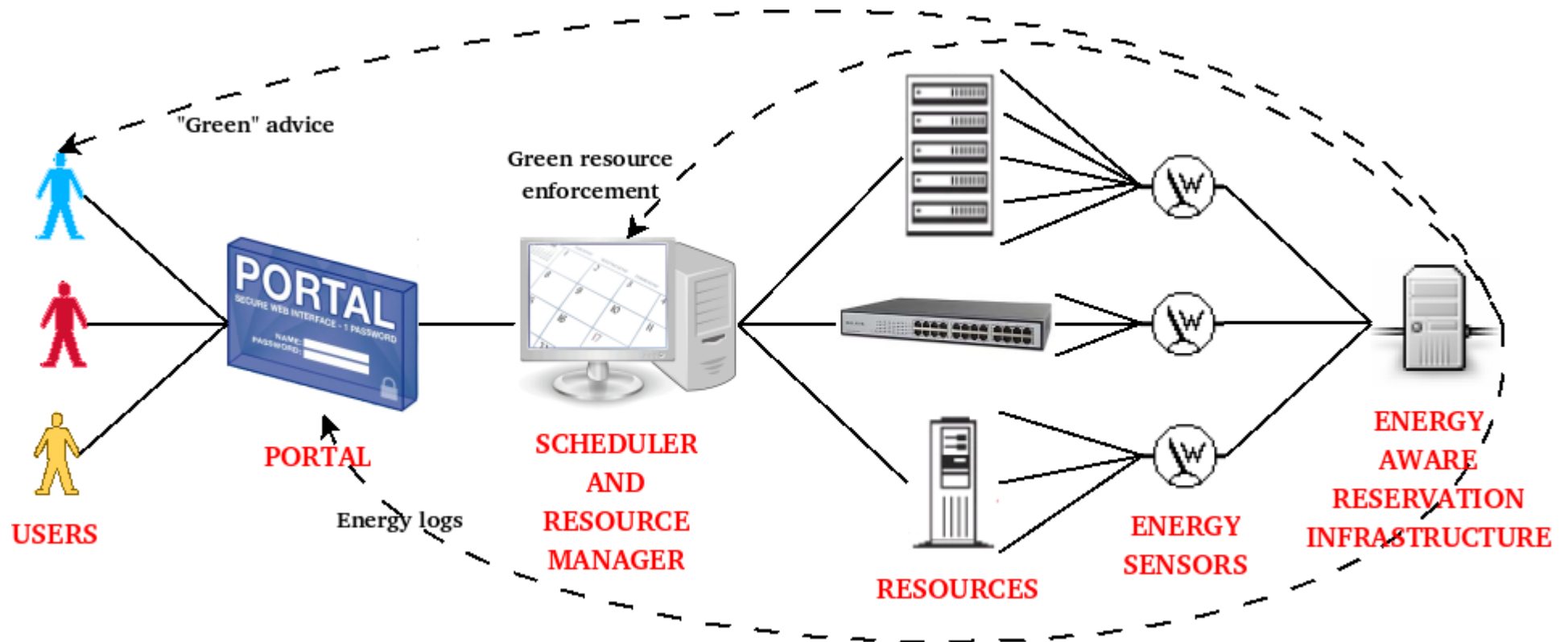
# Lessons learned for usage analysis

- Significant bursts -> significant gaps!
- A lot of small reservations (experimental platform)
- Significant energy usage while nodes are idle
- Benefit from gaps to propose energy savings models

# Architecture of an energy aware platform

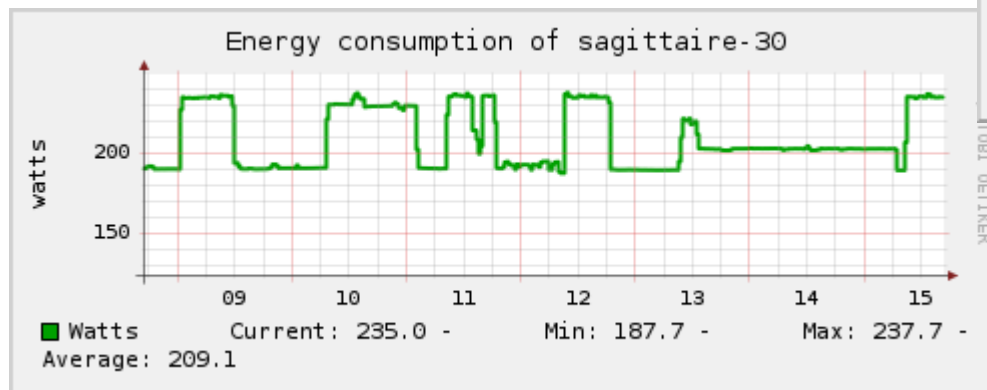
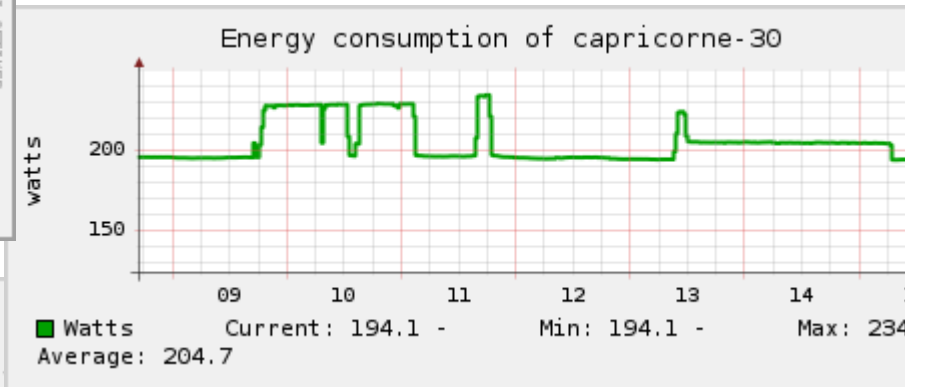
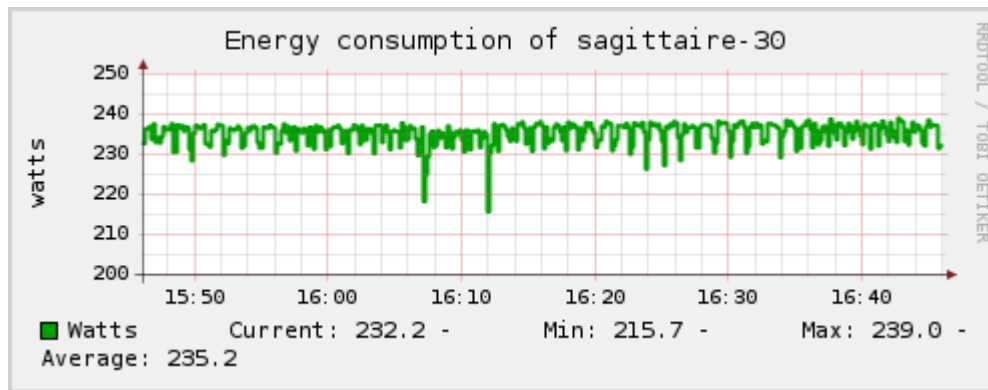


# Architecture of an energy aware platform



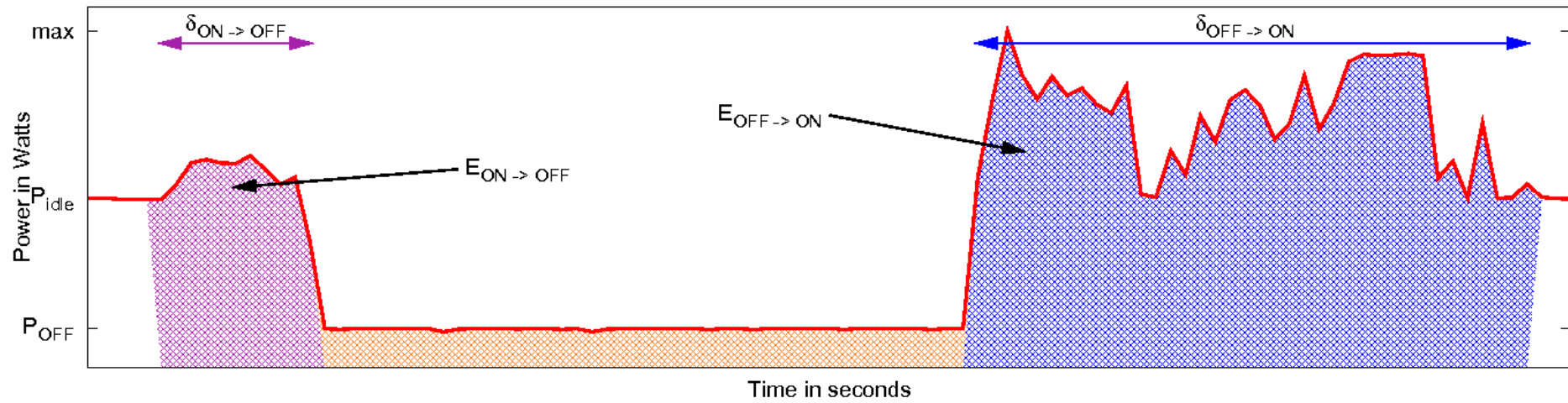
# Live Energy Monitoring

Feedback for users (days, weeks, month) for the reserved nodes

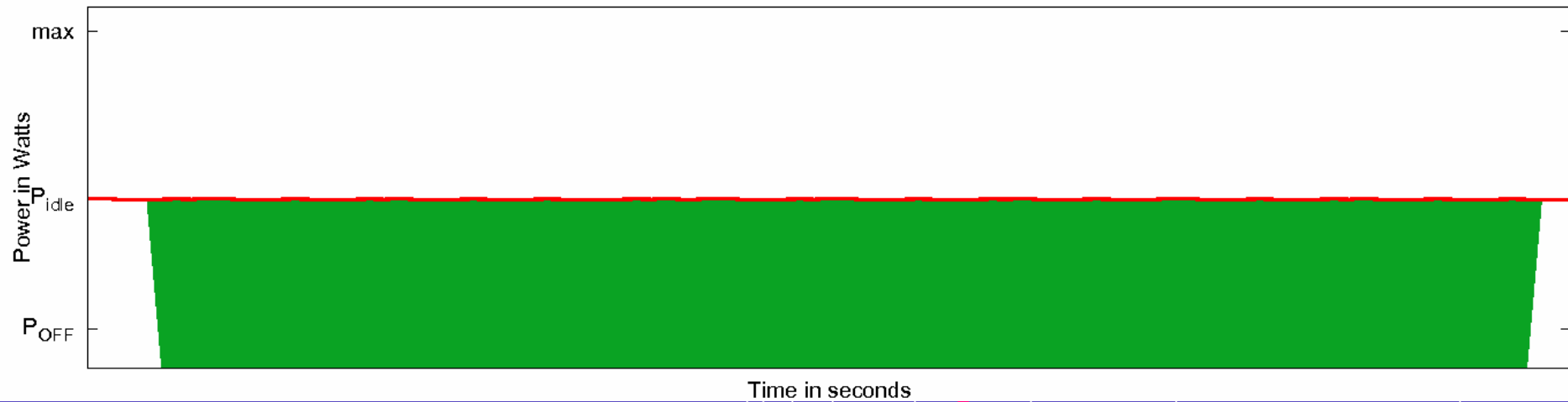


# On/Off model? → optimization

Consumption of the resource if it is switched off and on



Consumption of the resource if it stays idle



# On/Off model? → prediction

## Idea of the Algorithm:

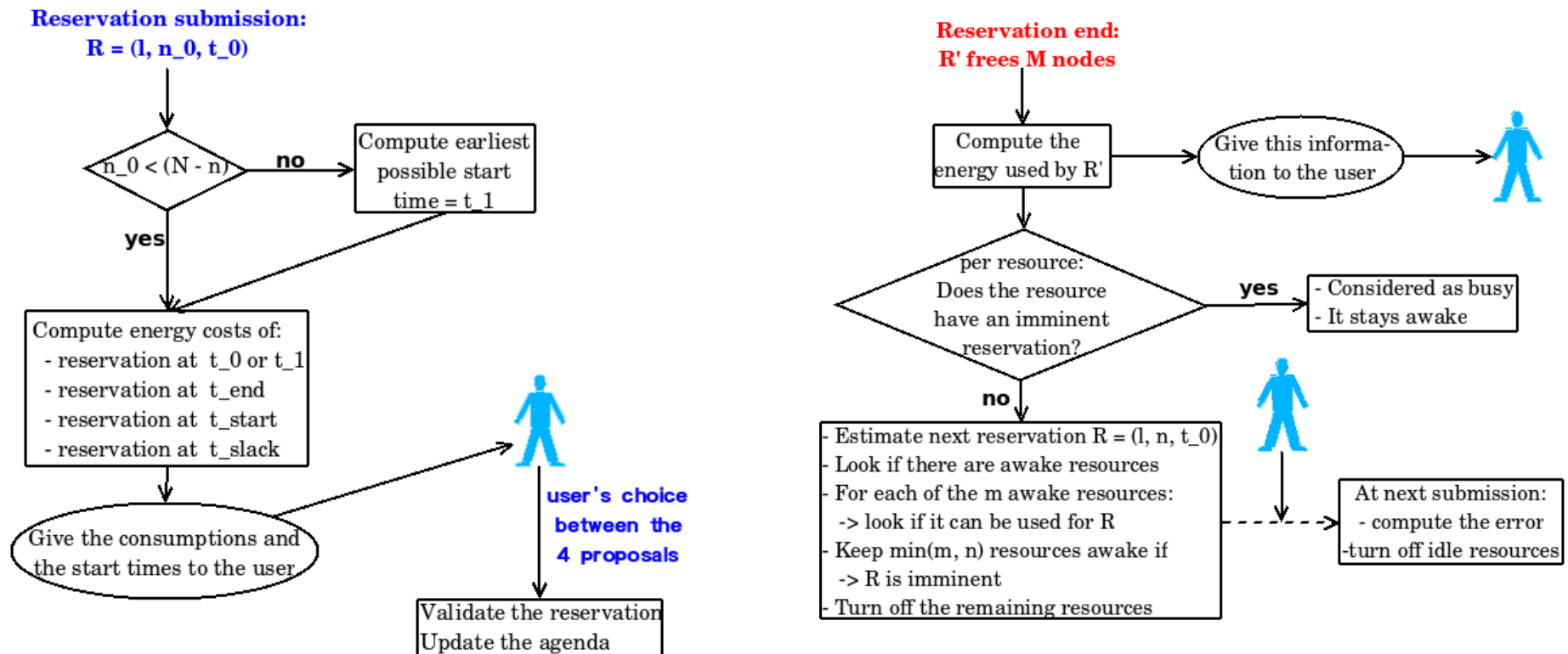
- At the end of a reservation, we want to know if it is better to shut down the nodes or not.
- We should then predict the next reservation.
- If the next predicted reservation is “*sufficiently near*”, we keep the nodes awake; otherwise we turn them off.
- Compute  $T_s$  such as:  $T_s - (\delta_{ON \rightarrow OFF} + \delta_{OFF \rightarrow ON}) \geq 0$ ;

and :

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

# Prediction algorithm ! Lightweight approach

After a reservation submission / at the end of a reservation





# Prediction

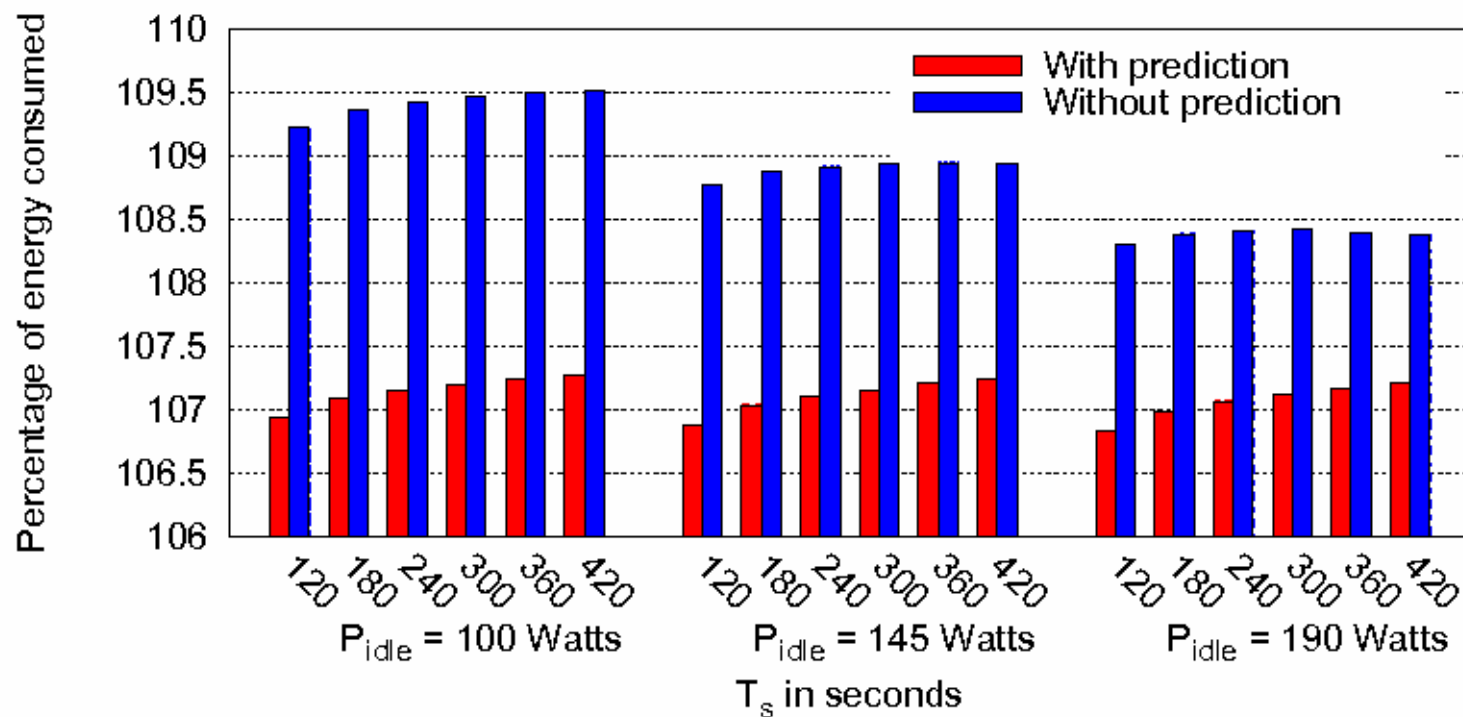
- What :
  - Next reservation (size, duration, start time)
  - Next empty period
  - Energy consumption of a reservation
- With :
  - Recent history (last reservation) + feedback
  - Recent reservations days + feedback
  - User history + resources
- How to validate the prediction ? How to validate the green policies ?

# Prediction evaluation based on replay

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

100 % : theoretical case (future perfectly known)

Currently (always on) : 185 % energy



# Green policies

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

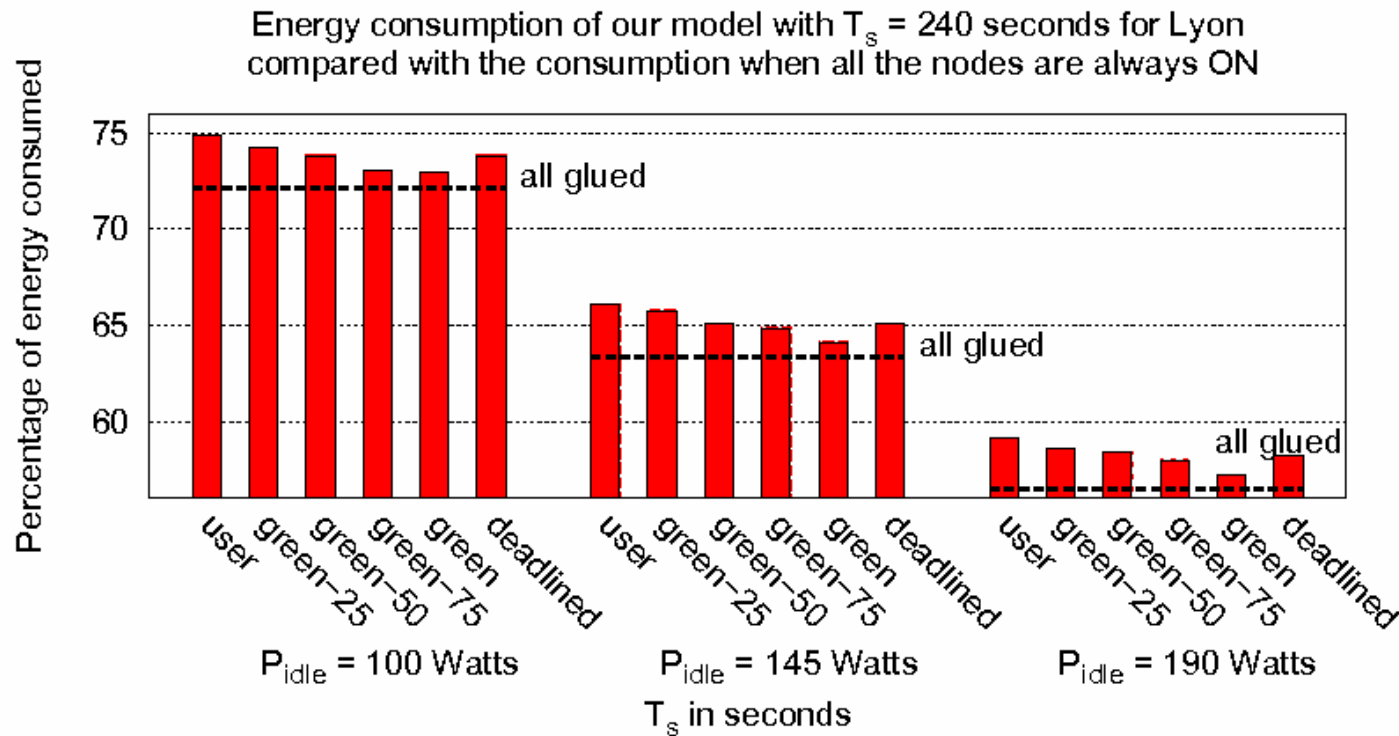
# Green policies evaluation

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

Current situation : always on nodes(100 %)

All glued : unreachable theoretical limit

Whatever P<sub>idle</sub> => same energy profile



# Local and global energy savings : now and in the future

- For Lyon site : current hardware :  $P_{idle} = 190$  W,  $P_{off} = 10$ W,  $P_{work} = 216$ W

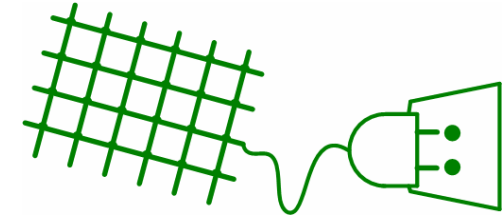
$P_{idle}$	present state	user	50% of green	fully green	all glued
100	135500	101500	100000	98300	97300
145	154000	101700	100300	98500	97300
190	172500	102000	100800	98700	97300

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

-1209159 kWh for the full Grid5000 platform (without aircooling and network equipments) on a 12 month periods

# Green-Net

## Running project



ARC GREEN-NET : (Action de Recherche Coopérative supported by INRIA) : 2008-2009

- Partners tams :

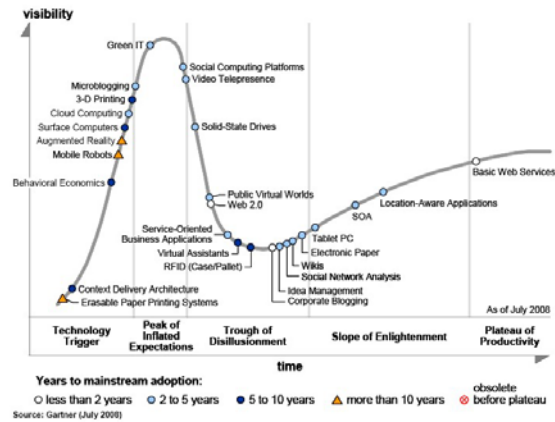
- IRIT (Toulouse)
- INRIA MESCAL (Grenoble)
- INRIA RESO (Lyon)
- Virginia Tech (USA)

- Goals :

- Energy usage analysis
- Modelization of electrical cost/process/services
- Adaptating a scheduling framework
- Large scale trust delegation : with constant equipment, with proxies
- Validation between partners sites

*<http://www.ens-lyon.fr/LI/RESO/Projects/green-net/>*

Figure 1. Hype Cycle for Emerging Technologies, 2008



# Conclusions



Good to be on top of the wave (Gartner Hype cycle) !

How to be energy efficient → how to reduce the watts on a large scale ?

The energy aspects will change the way we design software, protocols and services

The frameworks and middleware can/must help

Solutions for large scale / worldwide Grids

Human factor : are we ready to sacrifice some QoS/performance ?

# Future works

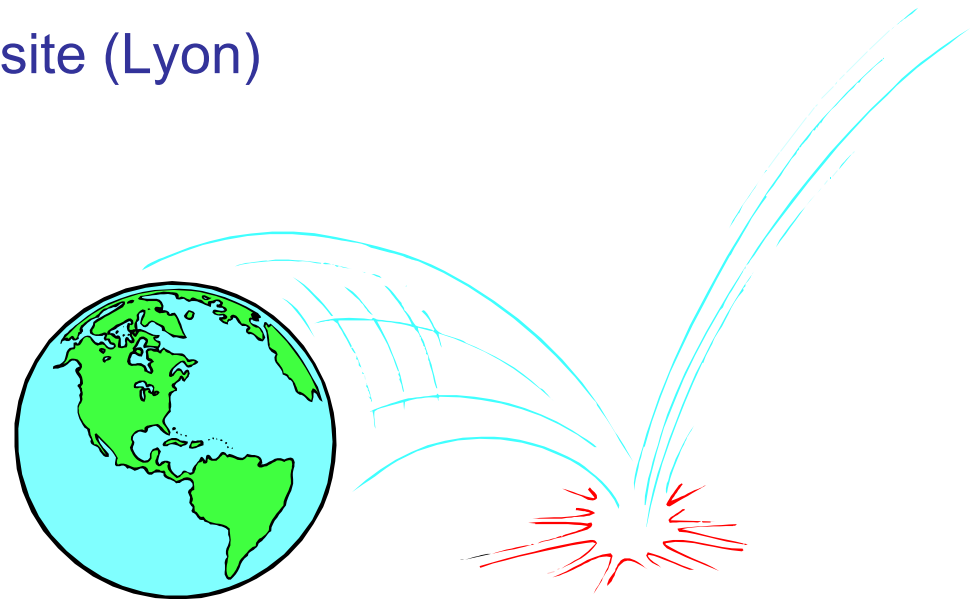
A lot !

We need new energy aware resources and equipments !

Need of energy aware benchmarks

Install energy sensors on a full site (Lyon)

We must save the world 😊





# Questions ?

*laurent.lefevre@inria.fr, annececile.orgerie@ens-lyon.fr, jpgelas@ens-lyon.fr*

*Anne-Cecile Orgerie, Laurent Lefèvre, and Jean-Patrick Gelas. « Chasing Gaps between Bursts : Towards Energy Efficient Large Scale Experimental Grids » In PDCAT 2008 : The Ninth International Conference on Parallel and Distributed Computing, Applications and Technologies, Dunedin, New Zealand, Dec 2008*

*Anne-Cecile Orgerie, Laurent Lefèvre, and Jean-Patrick Gelas. « Save Watts in your Grid: Green Strategies for Energy-Aware Framework in Large Scale Distributed Systems » In ICPADS 2008 : The 14th IEEE International Conference on Parallel and Distributed Systems, Melbourne, Australia, Dec 2008*