Energy efficiency in reservation-based large scale distributed systems

Laurent Lefèvre, Anne-Cécile Orgerie
INRIA RESO Project-Team
Ecole Normale Supérieure of Lyon - Université de Lyon

laurent.lefevre@inria.fr

HPC Workshop, Cetraro, June 27, 2011

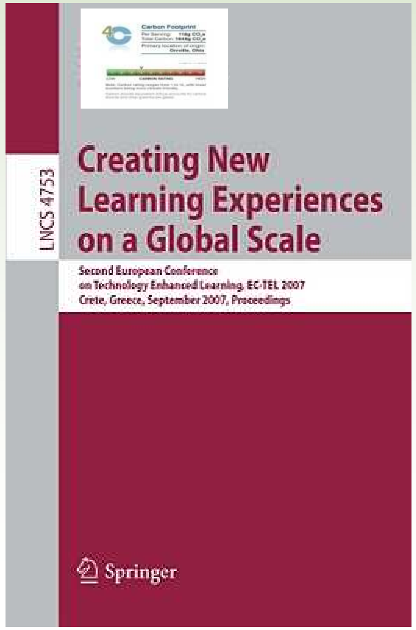
ENS DE LYON

INRIA RHÔNE-ALPES

Lyon 1

Energy : 1st challenge for large scale systems (datacenter, grids, clouds, internet)?

- Future exascale platforms -> systems from 20 to 100MW (current 4-6 MW – 10 MW for Kei)
- How to build such systems and make them energy sustainable/responsible ?
  - Hardware can help (component by component)
  - Software must be adapted to be scalable but also more energy efficient
  - Usage must be energy aware



Between incentive and reality :

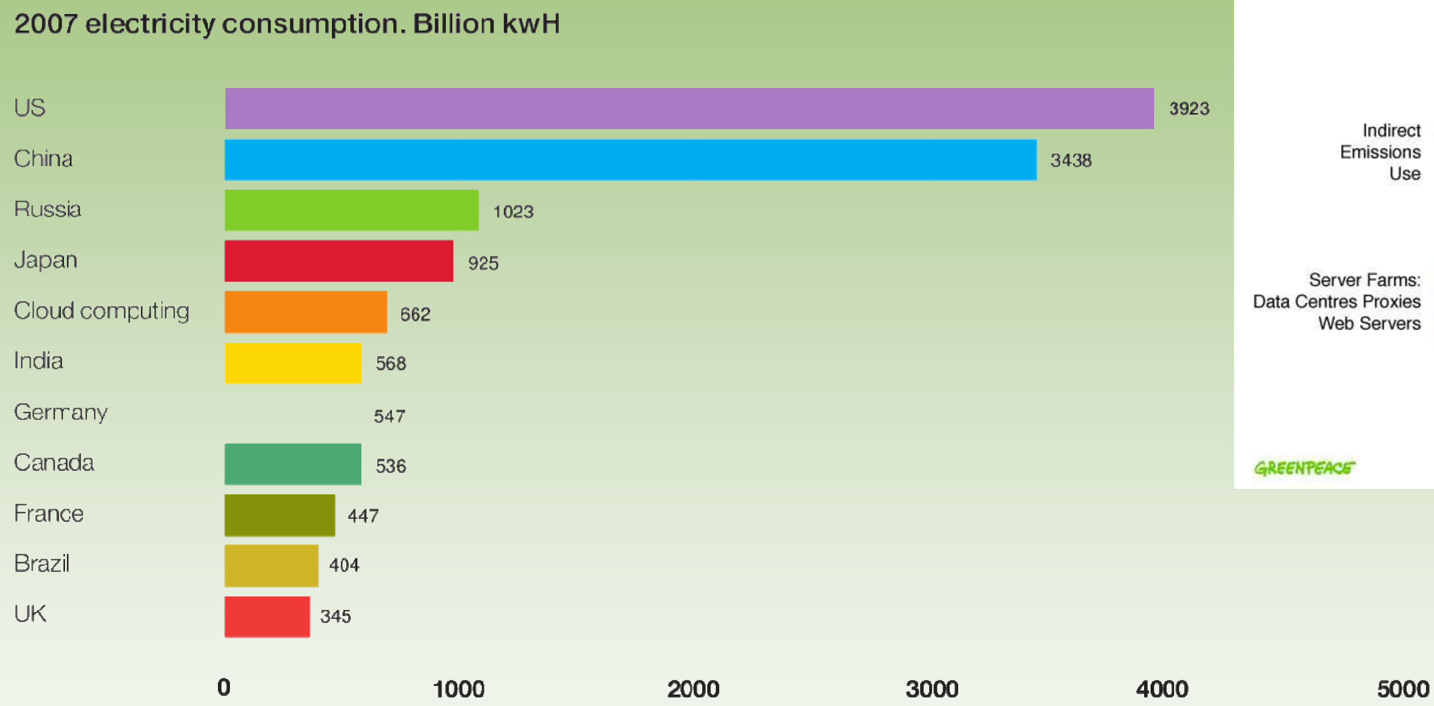
2010 : record year in CO_2 emission : 30.6 Gigatons (+5% previous record in 2008) IEA

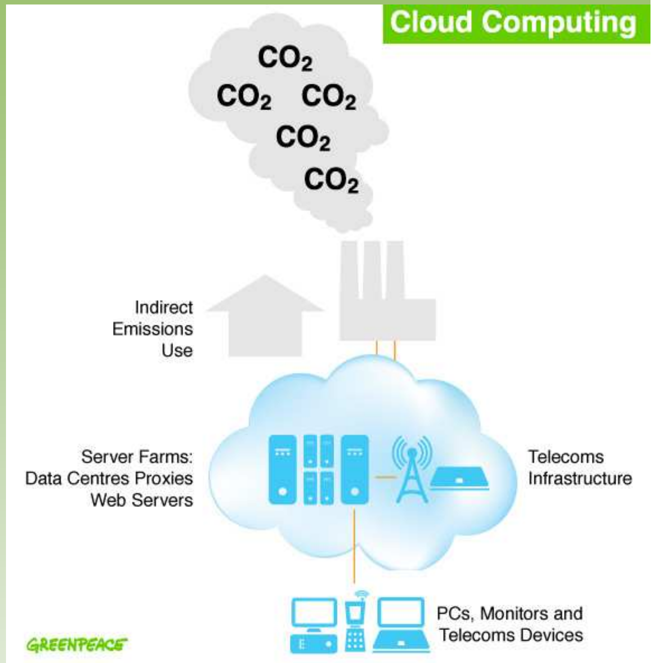
44% coal – 36% petrol – 20% natural gas

Temperature increasing (2°C – 2100) -> 4°C (50% chance – 2100)

We are part of climate changing !

- Or at least of enormous electricity usage
- As IT users/designers





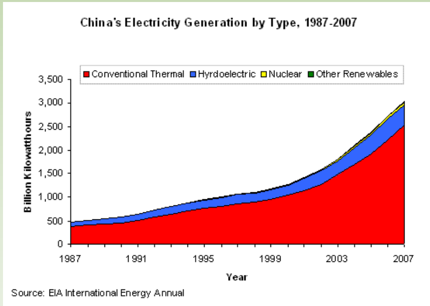
- « Greenpeace reports 2010-2011 »

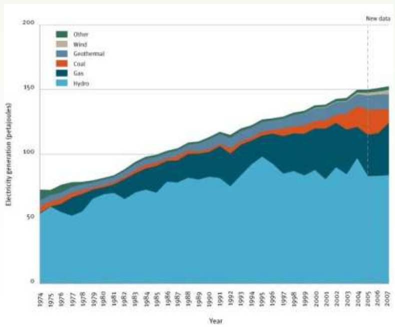
Because it is (intensively) used !

- OK results maybe not very accurate, but estimations…
- Seems exact :
  - 200 M emails / minute
  - 700 000 google request per minute
  - 48 hours of youtube video per minute

Power demand and Green IT

- IT – 2-5% of CO_2 emissions
- Green It ➔ reducing electrical consumption of IT equipments - CO_2 impact depends on countries
- Focus on usage : fighting un-used/over-provisioned plugged resources





The 4 loops issue

- **Usage loop** : hardware , applications & services energy usage. What is the impact of my application to energy consumption ?
- **Infrastructure Loop** : embedding into the models the infrastructure energy cost (air/water/free/un cooling). What is the impact of my environment to infrastructure consumption ?
- **Life cycle loop** : from production, usage to destruction (recycling) : What is the energy cost of my IT during its whole life?
- **Human loop** : add the human factor around the precedent loop. How to keep the world « happy » while reducing energy usage? ☺
- The holy Grail : a model and frameworks with the 4 loops inside

- This apply to all kind of products : food (local apple vs foreign bananas), cars (prius vs hummer)… etc…
- But with IT it should be easier/quantifiable… hum.. not yet…

Towards Energy Aware Large Scale Systems

How to decrease the energy consumption without impacting the performances?

- How to understand and to analyze the usage and energy consumption of large scale platforms?
- How to monitor lively such usage from pico to large scale views?
- How to design energy aware software frameworks ?
- How to help users to express theirs Green concerns and to express tradeoffs between performance and energy efficiency ?

Green-IT Leverages

- **Shutdown** : reducing the amount of powered unused resources
- **Slowdown** : adapting the speed of resources to real usage
- **Optimizing** : improving hardware and software for energy reduction purpose (i.e. energy aware libraries). Adapt software to green hardware.
- **Coordinating :** using large scale approaches to enhance green leverages

Explosion of (uncoordinated) initiatives

For each domain

- **Data centers/HPC : Green500, EU CoC, The Green Grid**
- **Storage : SNIA**
- **Networks : Green Touch / EEE**

Our Methodology

- Proposing a generic model able to be derivated onto different scenario (Grids, Clouds, Networks)
- Designing software solutions for infrastructures
- Simulating and Validating at medium and large scale

Reservation-based systems

Every usage is based on a reservation (resources, duration, deadline):

- Reserving cpu in HPC and Grids
- Reserving Virtual machines time in Clouds
- Reserving Bandwidth in large transport of data

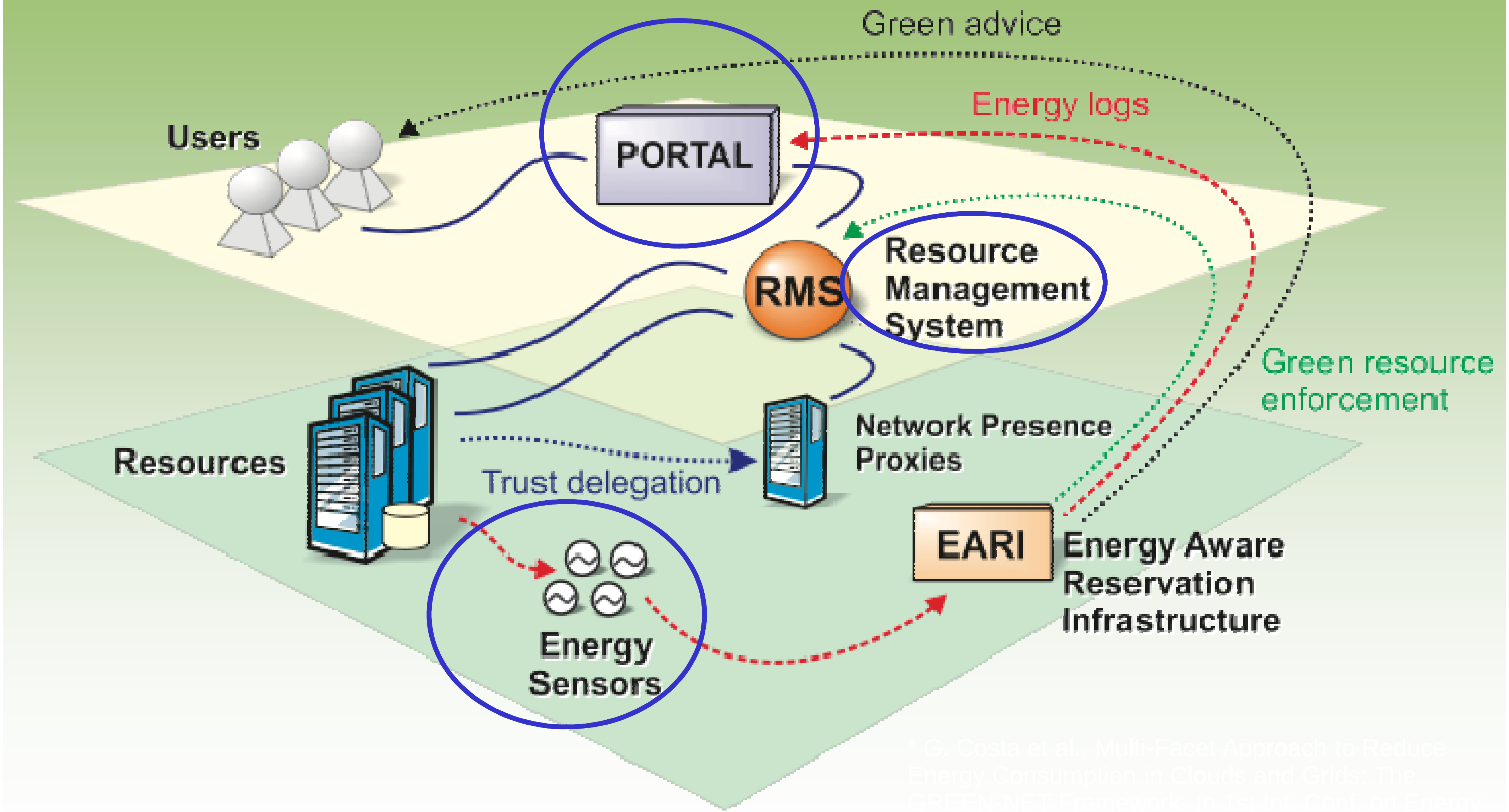
- Leverages:
  - Finding and powering the optimal number of resources in front of needs of applications
    - HPC and Grids : switching on/off physical components
    - Clouds : switching on/off VMs
    - Networks : lighting or switching off paths, nics, links, routers, LPI
  - Adapting « speed » (and consumption) to the need of applications/users
    - HPC, Grids : dvfs
    - Clouds : tuning, capping
    - Networks : adaptive link rate

The ERIDIS model

Energy-efficient Reservation Infrastructure for large-scale DIstributed Systems

- Collecting and exposing : usage, energy profiling of applications and infrastructures
- Predicting usage of infrastructures
- Expressing and Proposing : to deal with tradeoffs between perf and energy, Green Policies
- Agregating resources reservations and usage in time and space
- Enforcing Green leverages : switch on/off or adapt performancs

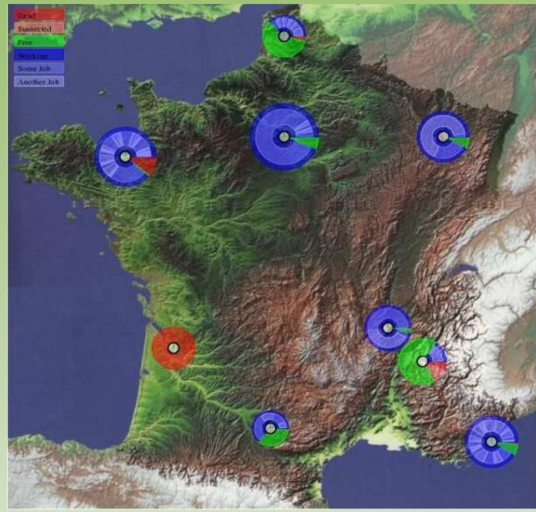
The ERIDIS Framework

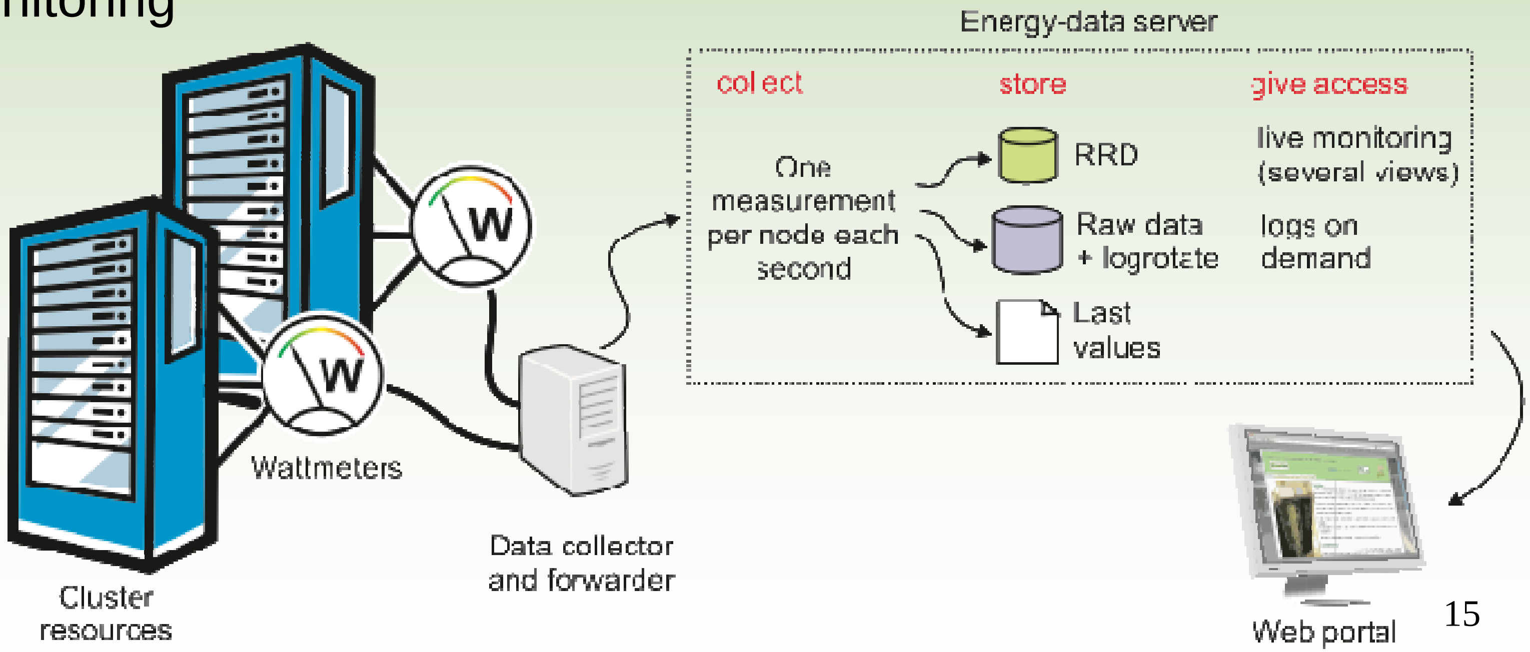


* G. Costa et al., Multi-Facet Approach to Reduce Energy Consumption in Clouds and Grids: The GREEN-NET Framework. In 1st Int. Conf. on Energy Efficient Computing and Networking, Passau, 2010.

Collecting and exposing

- (Green) Grid'5000
  - French experimental testbed
  - 7400 cores
  - 10 sites
  - External energy sensors
  - Full site monitoring



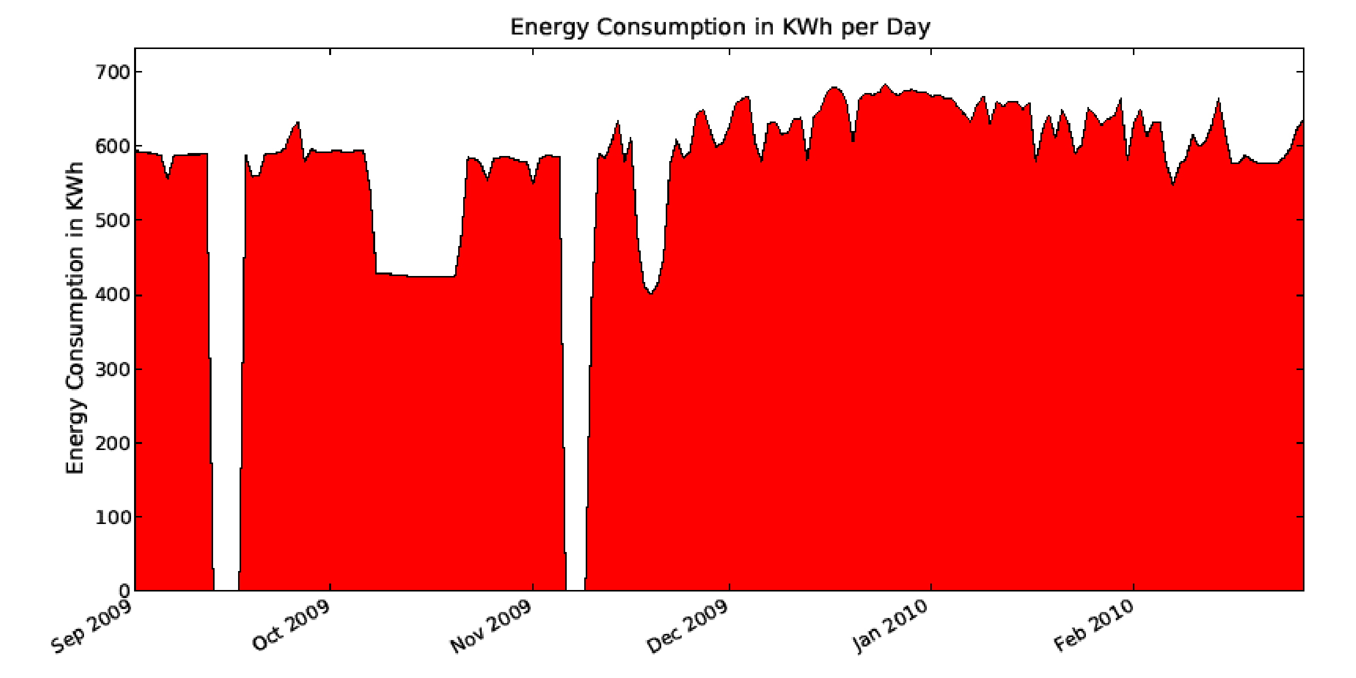


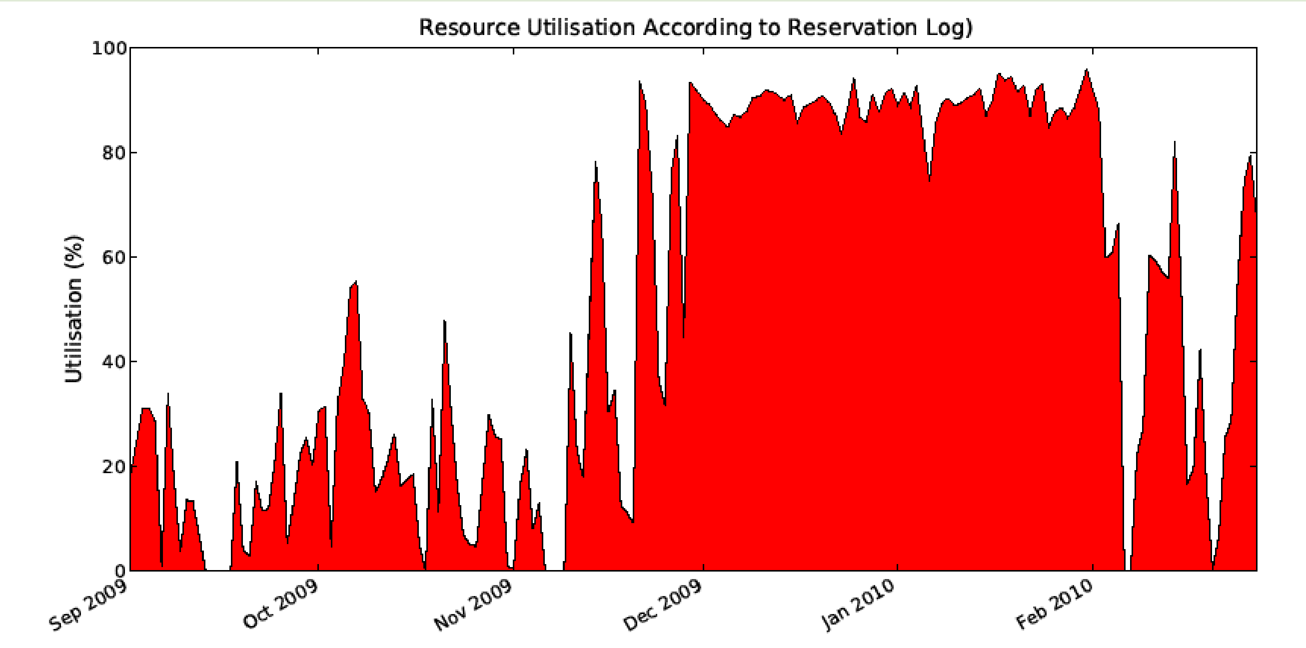
Electrical consumption / Usage

Periodicity of **energy measurements**:

One measurement per **second** for each equipment

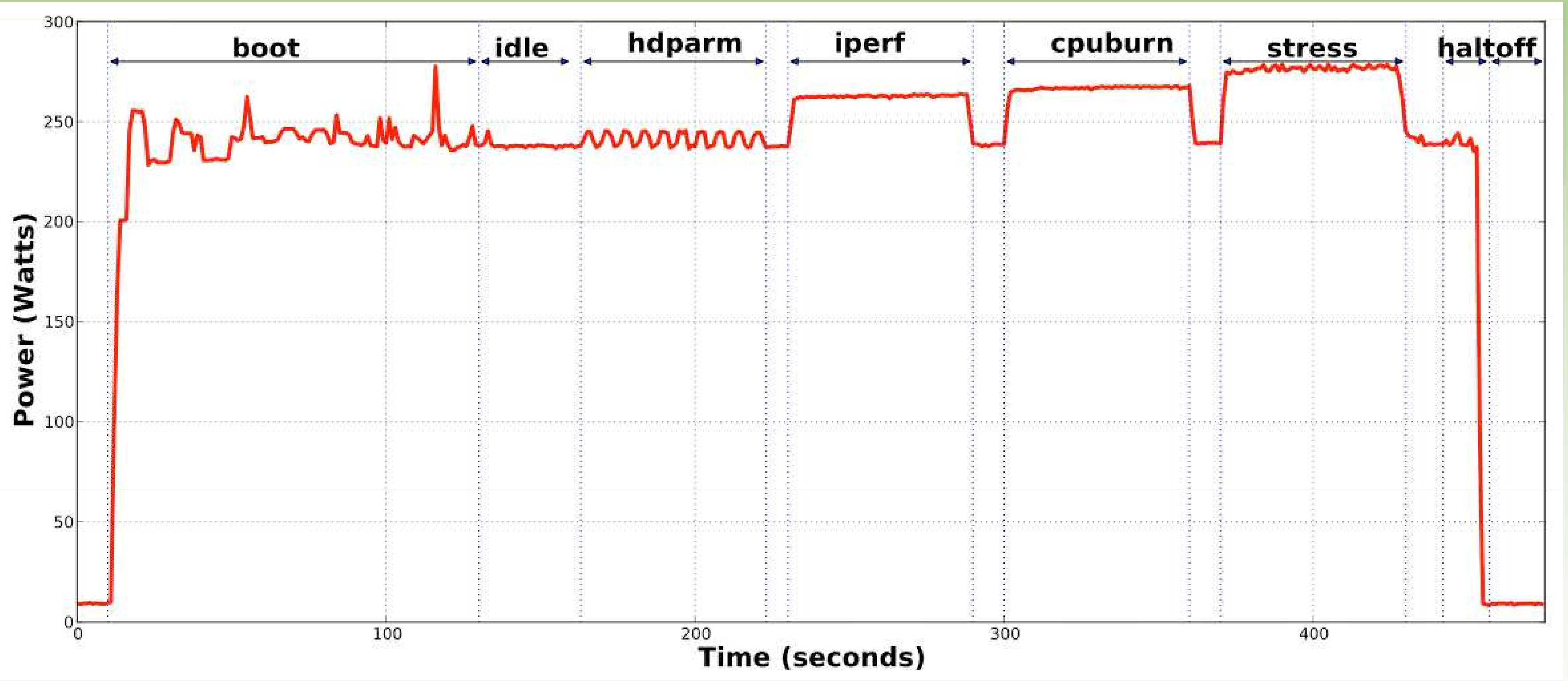
*





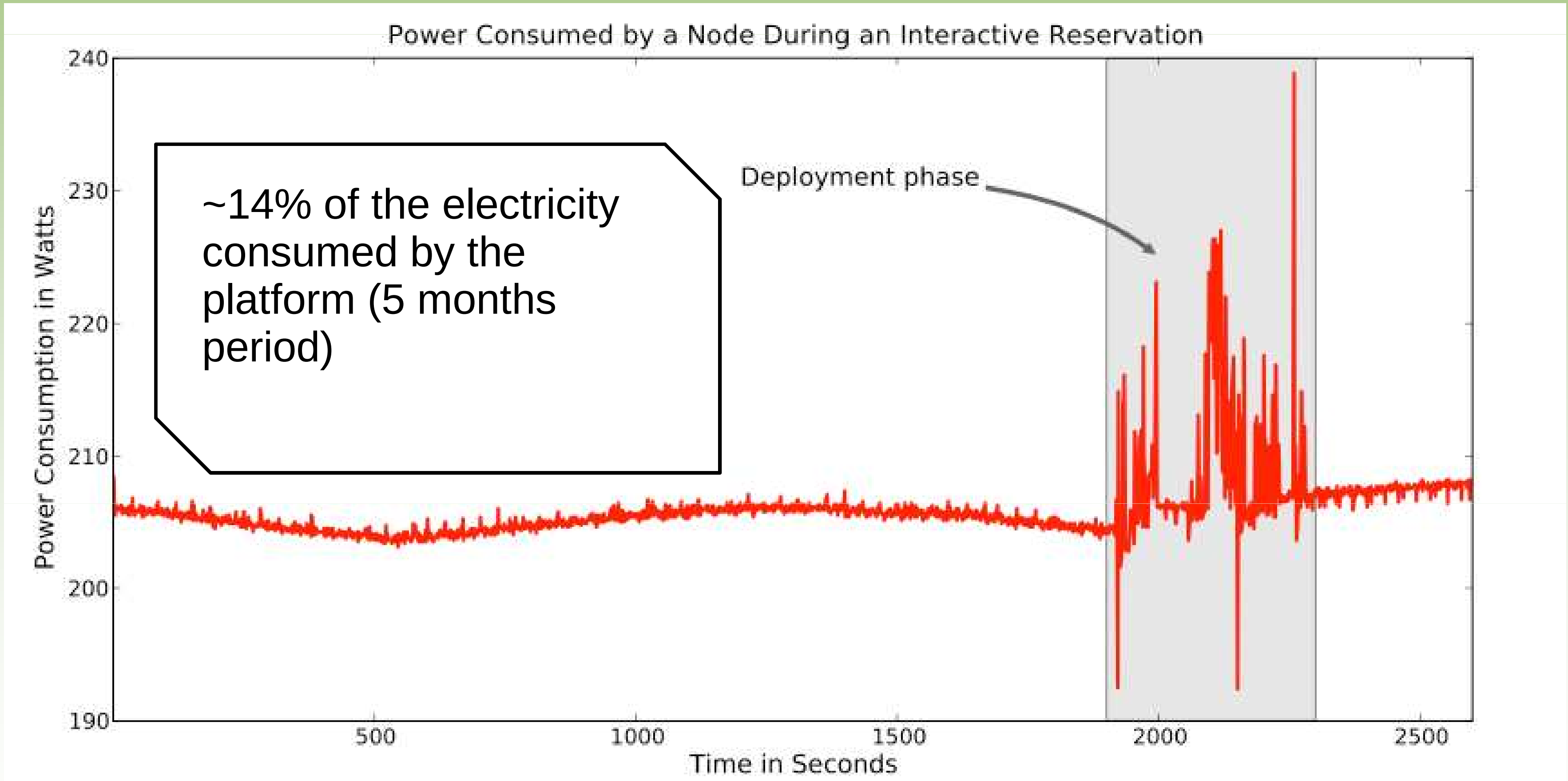
Example I : Profiling applications

Profiling the energy consumption of applications

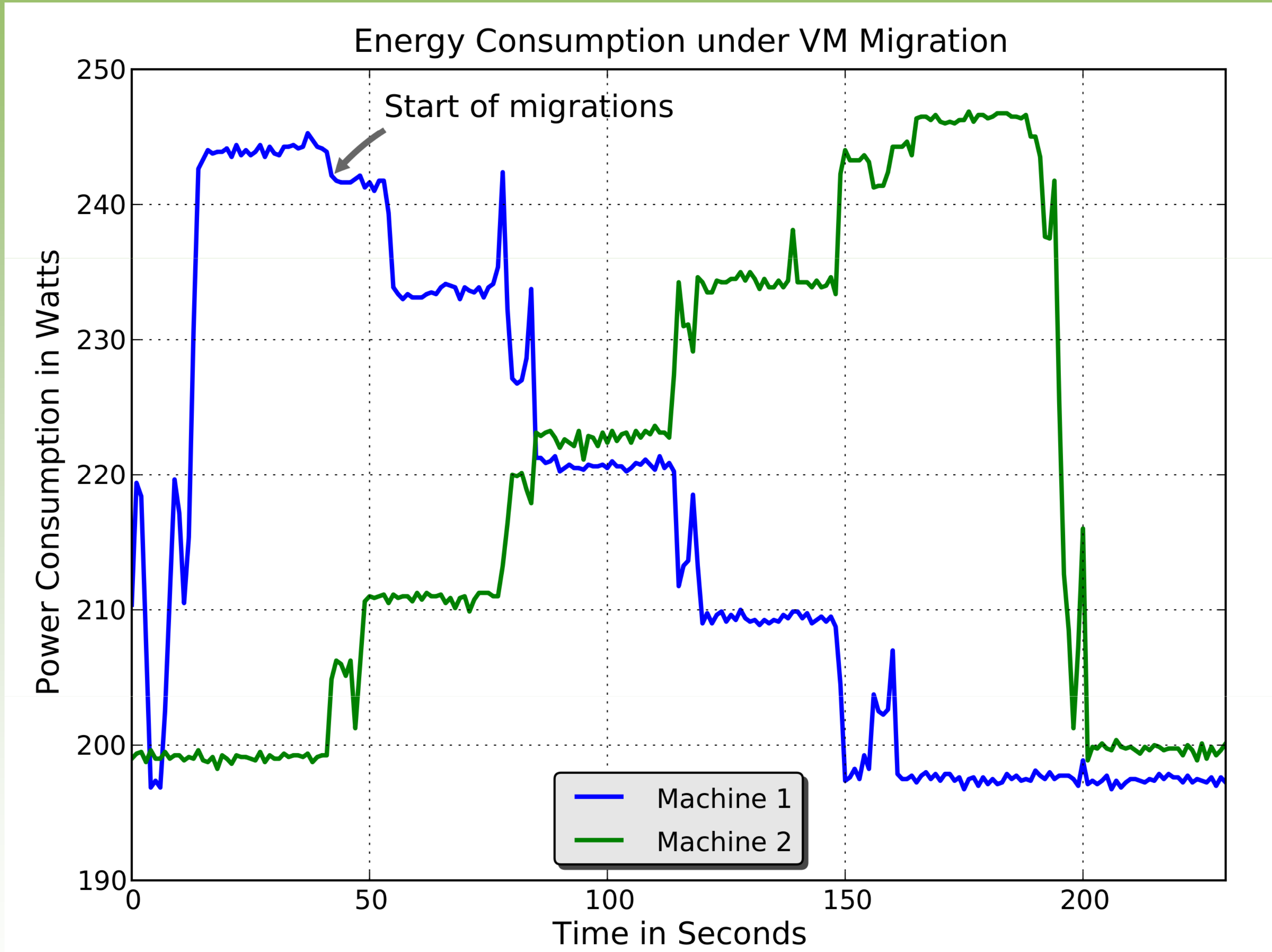


Example II : detecting anomalies

Improving frameworks/middleware and policies

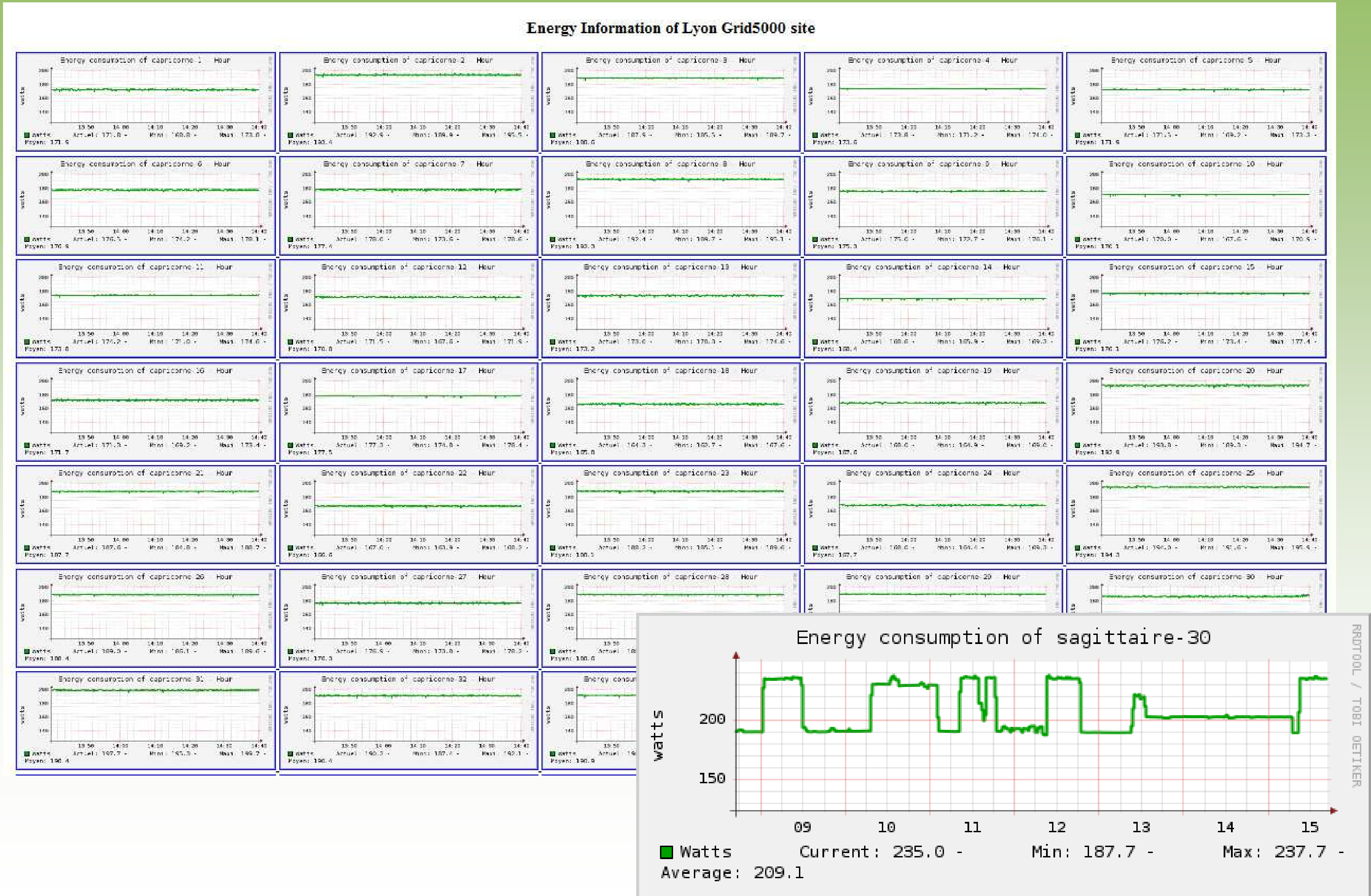


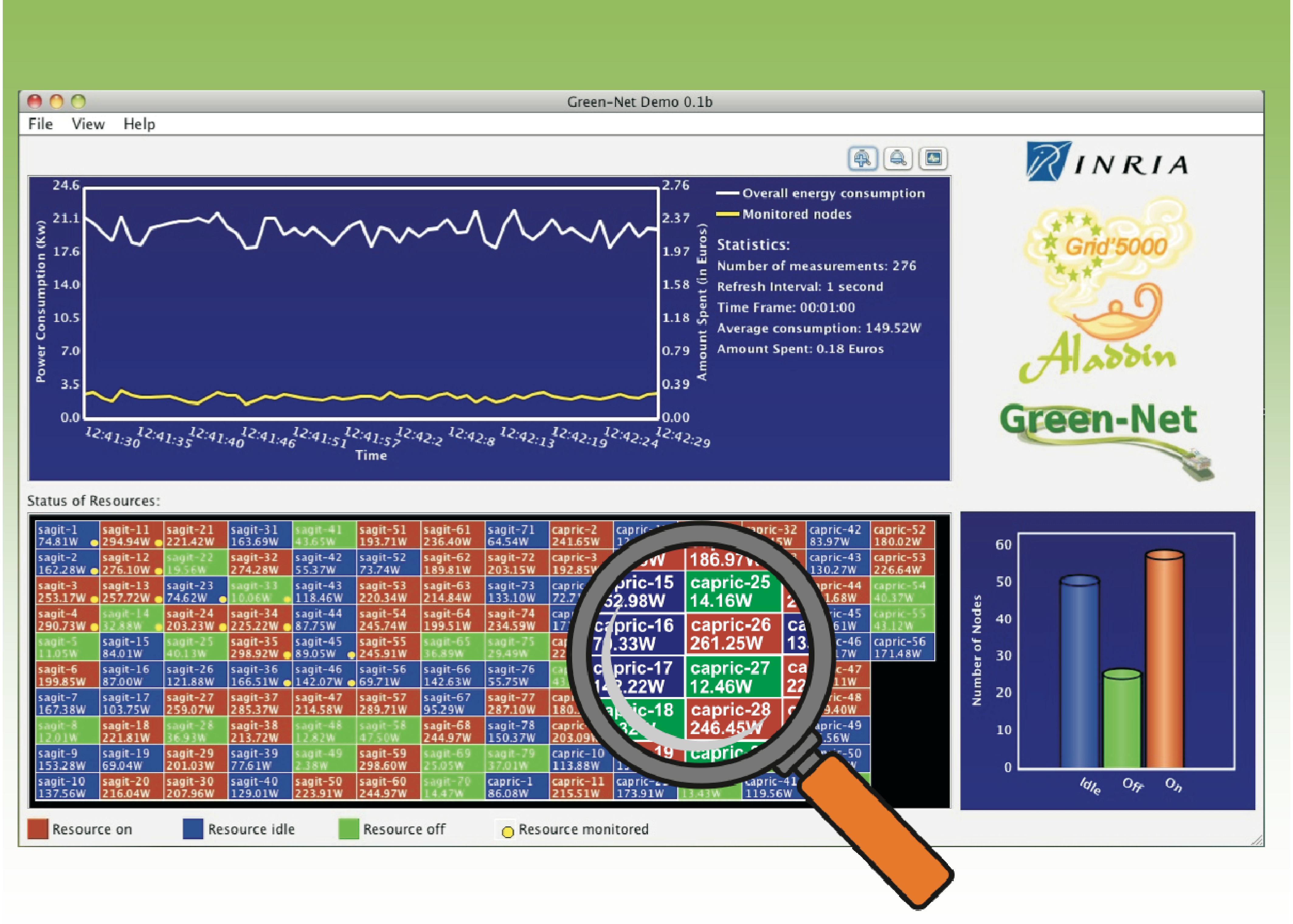
Migration



- Bad moment in energy during the migration

Large scale energy exposing

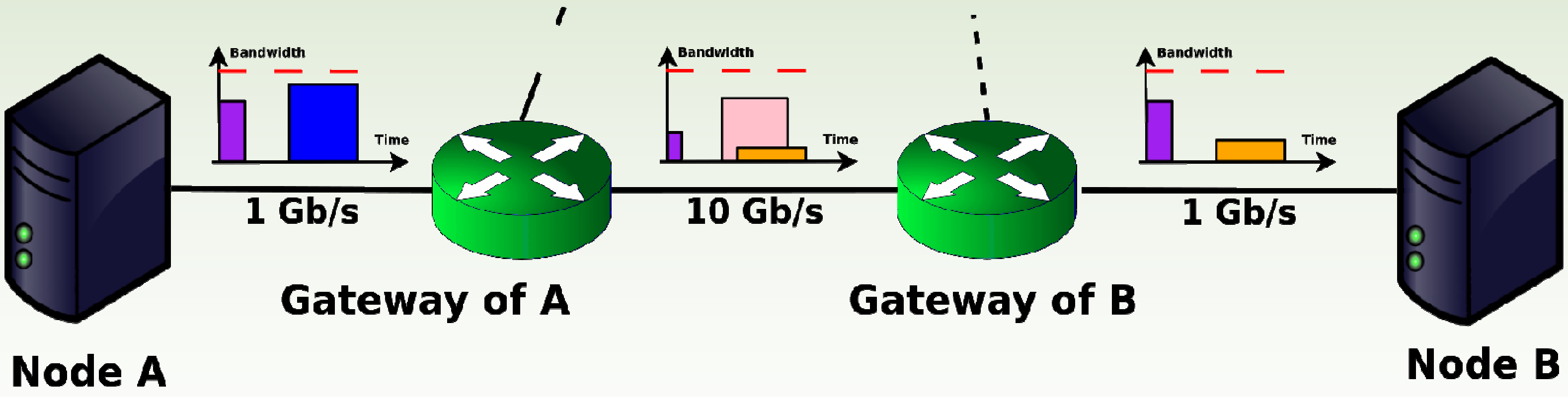




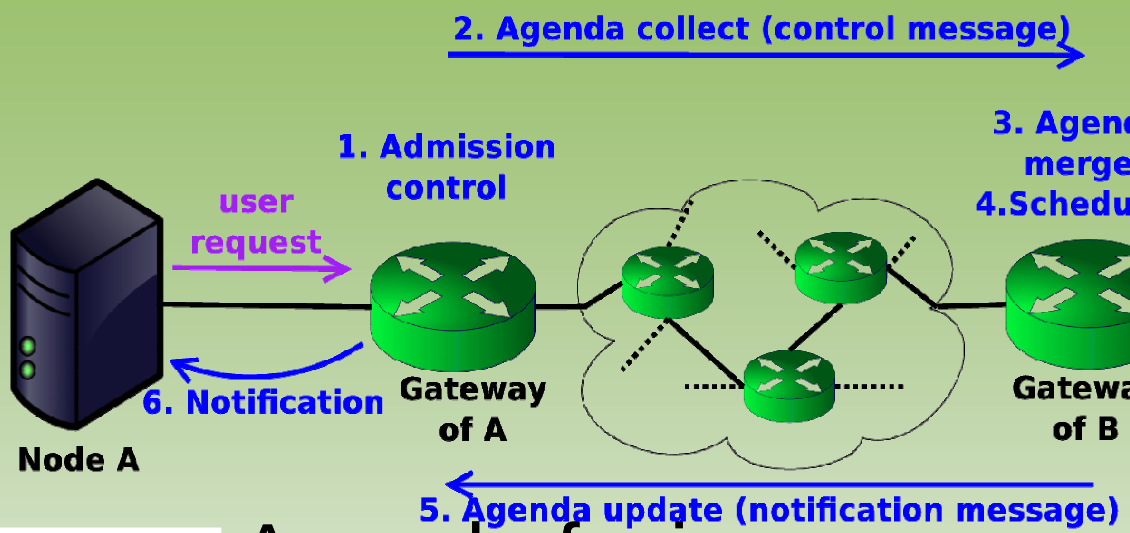
Green-Net Demo 0.1b
File
View
Help
Power Consumption (Kw)
Amount Spent (in Euros)
Time
Overall energy consumption
Monitored nodes
Statistics:
Number of measurements: 276
Refresh Interval: 1 second
Time Frame: 00:01:00
Average consumption: 149.52W
Amount Spent: 0.18 Euros
INRIA
Grid'5000
Aladdin
Green-Net
Status of Resources:
Number of Nodes
Idle
Off
On
Resource on
Resource idle
Resource off
Resource monitored

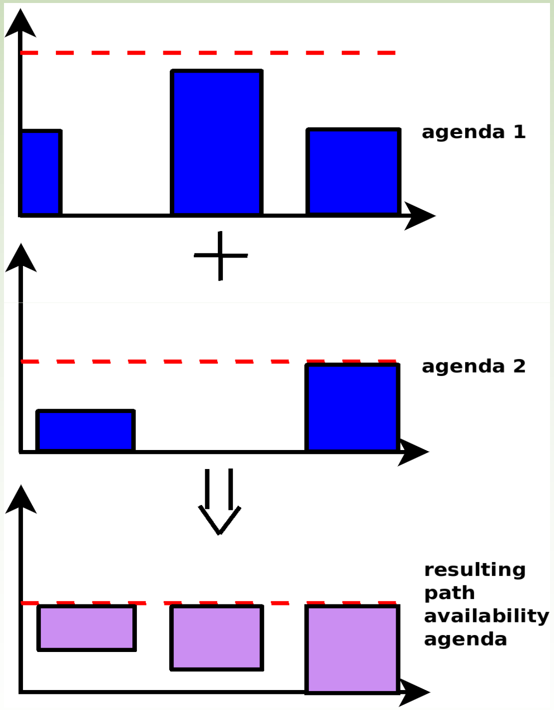
HERMES : High-level Energy-awaRe Model for bandwidth reservation in End-to-end networkS

- Switching off unused parts of the network : NIC, routers, links
- Distributed network management
- Energy-efficient scheduling with reservation aggregation
- Usage prediction to avoid on/off cycles
- Minimization of the management messages
- Usage of DTN (Disruptive-Tolerant Network) for network management purpose



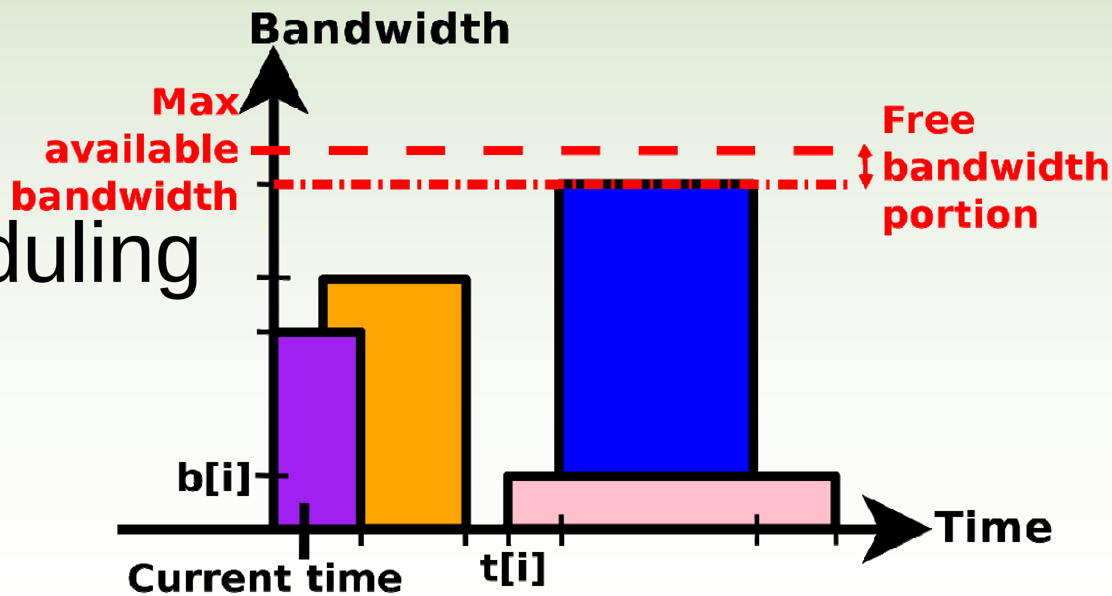
Hermes





Agenda fusion

ABR scheduling



Hermes results

- Network simulated: 500 nodes, 2 462 links.
- Random Network (Molloy & Reed method)
- All the nodes can be sources and destinations.
- Time to boot: 30 s.; time to shutdown: 1 s.
- 1 Gbps per port routers

Component	State	Power
Chassis	ON	150 W
	OFF	10 W
Port	1 Gbps	5 W
	100 Mbps	3 W
	idle, 10 Mbps	1 W

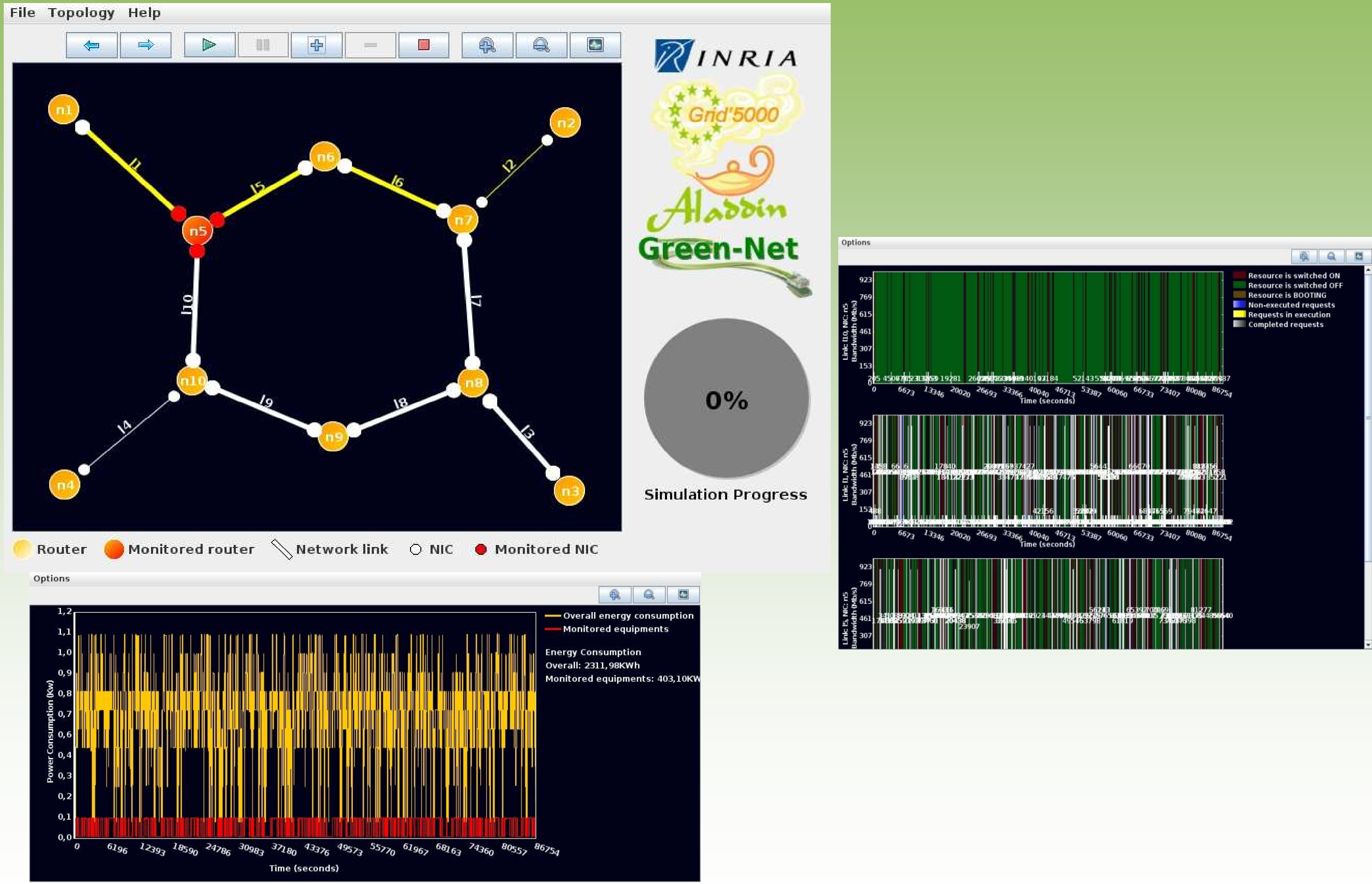
- 31% workload : Energy consumption in Wh

Scheduling	No off	First	First green	Last	Last green	Green
Average	412 306	205 270	203 844	204 949	196 260	203 342
Standard deviation	2 685	2 477	1 938	2 375	2 695	2 145
Accepted volume (Tb)	2 148	2 148	2 128	2 014	1 853	2 149
Cost in Wh per Tb	191.92	95.55	95.78	101.74	105.92	94.60

- Cost in Wh per Tb
- Compared to current case (no-off), HERMES could save 51%, 46% and 43% of the energy consumed depending on the workload

Workload	No off	First	First green	Last	Last green	Green
31%	191.92	95.55	95.78	101.74	105.92	94.60
46%	149.84	81.61	81.95	87.74	92.40	80.63
61%	130.45	74.73	74.91	80.09	84.63	73.79

Replayer



- 2010 SuperComputing demo, Marcos Dias de Assunção

Conclusions

- Big role for IT: Green IT and IT for Green
- Challenge : design energy proportional equipments and frameworks (computing, memory or network usage)
- Need to take out energy efficient models from the lab and put them in operationnal conditions
- Adress all levels from hardware to software – production and usage (4 loops)
- Explore EE best effort environments

Thanks to Anne-Cécile Orgerie

Questions?

Laurent Lefèvre
Laurent.lefevre@inria.fr

ENS DE LYON

INRIA RHÔNE-ALPES

Lyon 1