Towards energy proportional HPC and Clouds

Laurent Lefèvre
laurent.lefevre@inria.fr

CCDSC2014, Dareizé, September 5, 2014

Thanks to Jack and Bernard !

INRIA **AVALON / LIP**
Ecole Normale
Supérieure de Lyon

Some messages from our planet

Ice 500 Gtons 2011-2014 : Groenland 375 Gt /Antartic 125 Gt : *2/*3 compared to average between 03-09

Rising > 1 m (2100)

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

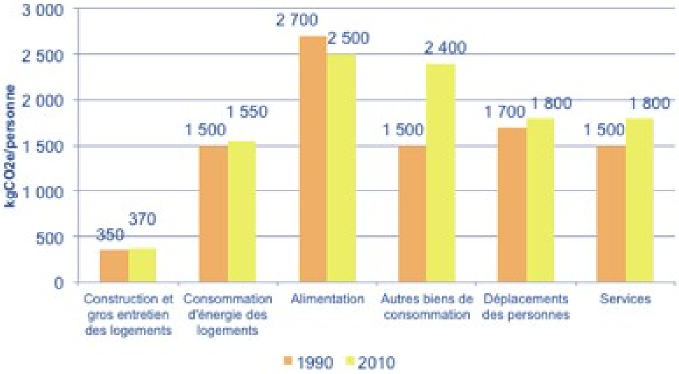
No more petrol in 50 years …

IT -> electricity -> CO2 -> impact

So we should change our way to use energy with IT - Chasing watts / chasing overprovisioning / unuseful services…

But IT/HPC/Cloud are good for the planet : IT4Green

- Problem : IT 4 Green is not yet proven (at least in France)
  - France : total increase (co2 emission) : 25 % between 1990 and 2010
  - +11 % population = +13 % per person increase
- Cloud/visio do not avoid travels



I am sure that the fox wants to…

- Avoid resources wasting

- Avoid sea rising increase

- Avoid global warming

- Avoid biodiversity loss

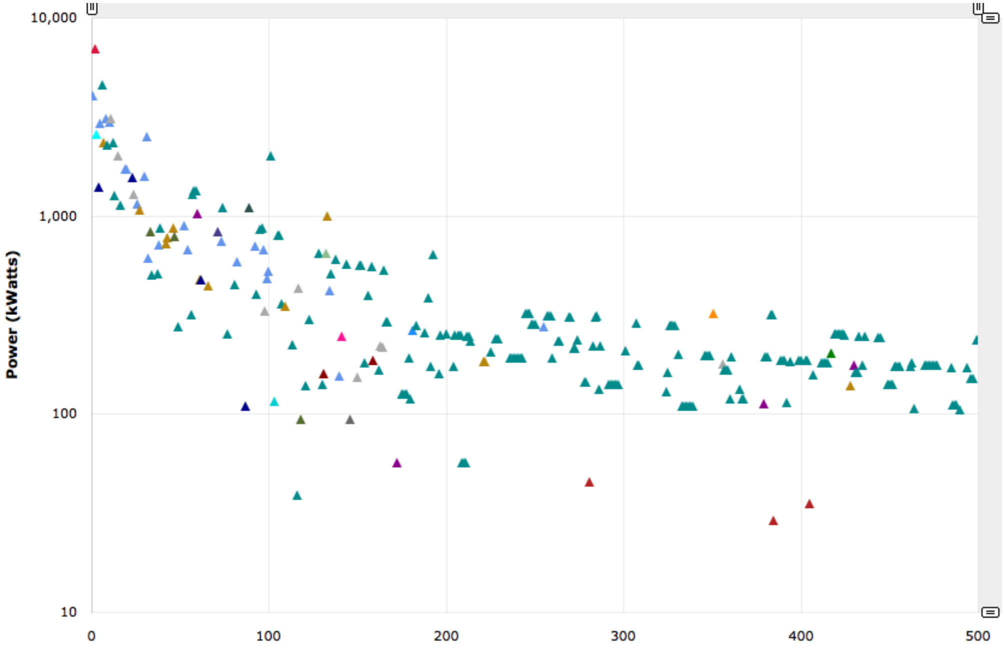
So my interpretation/assumption : « the fox wants to promote energy efficient infrastructures »

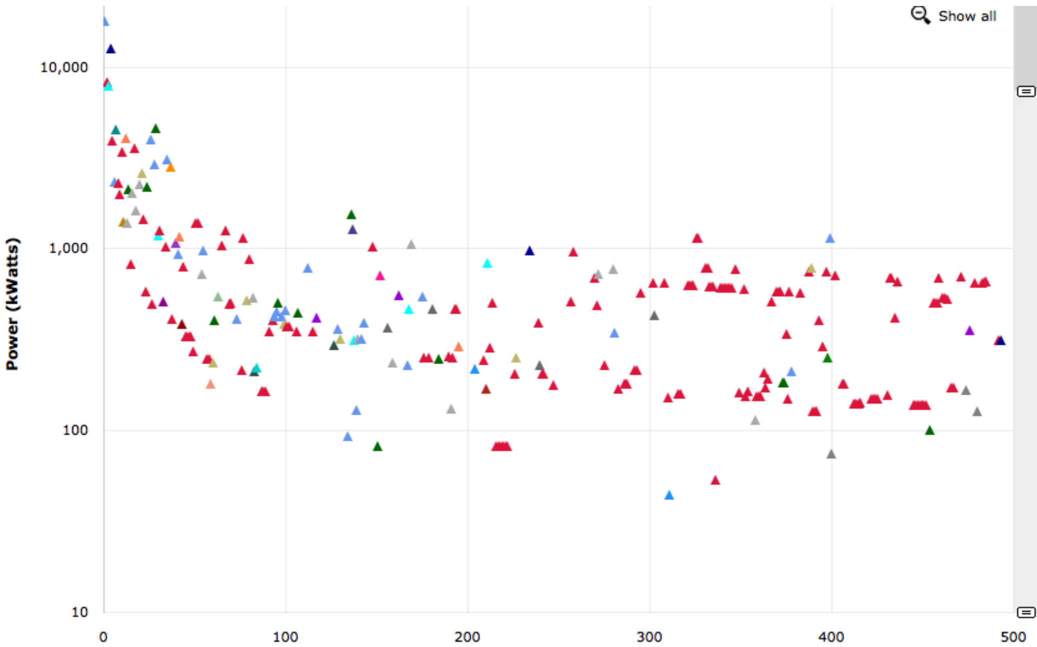
Energy : 1st limiting factor for large scale systems ((hpc)datacenter, clouds, internet)?

- Energy consumption is growing : Top500 : Nov 2010 : 127 MW – Nov 2013 : 205 MW (not all referenced) - Green500 : 550 MW (Nov. 13 – all referenced)

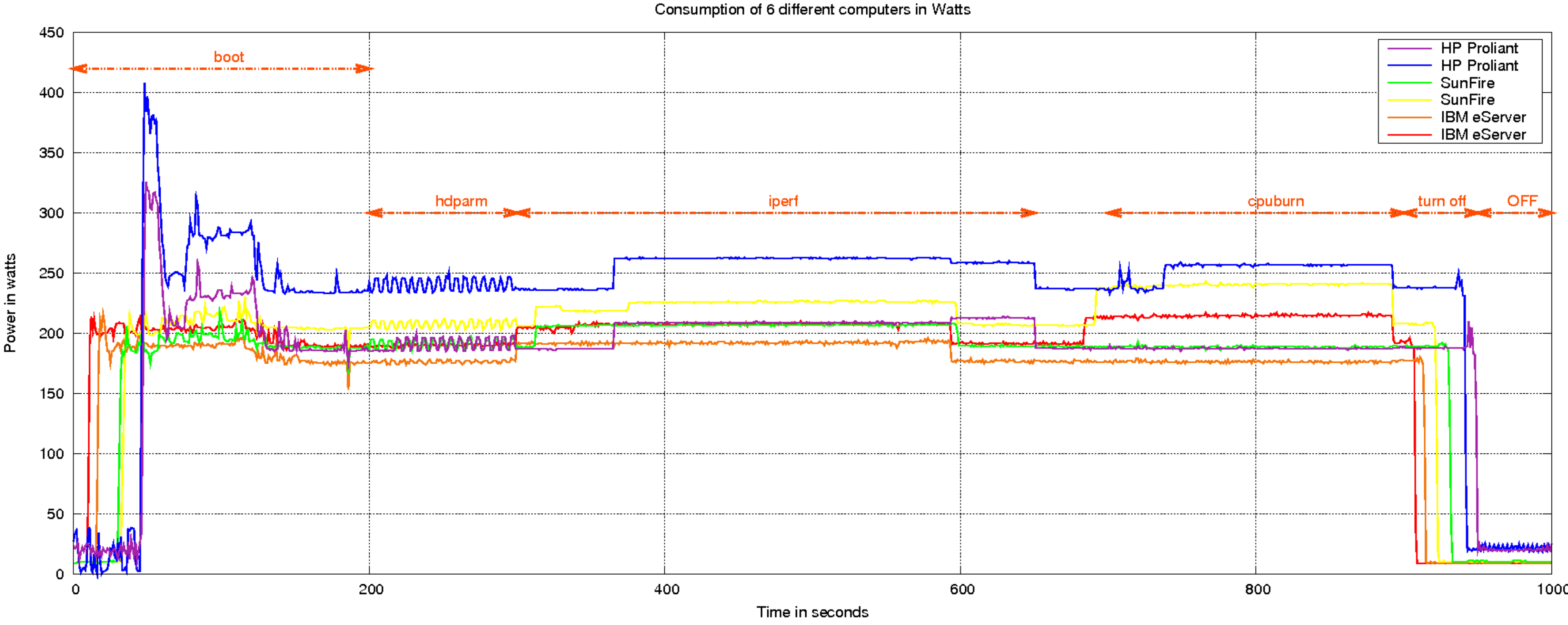
- Only usage ! not the full life cycle which is bad : planned obsolescence, rebound effect, design (rare minerals), difficult recycling...

- How to build future exascale/datacenters platforms and make them (more) energy sustainable/responsible ? - Multi dimension approaches : hardware, software, usage



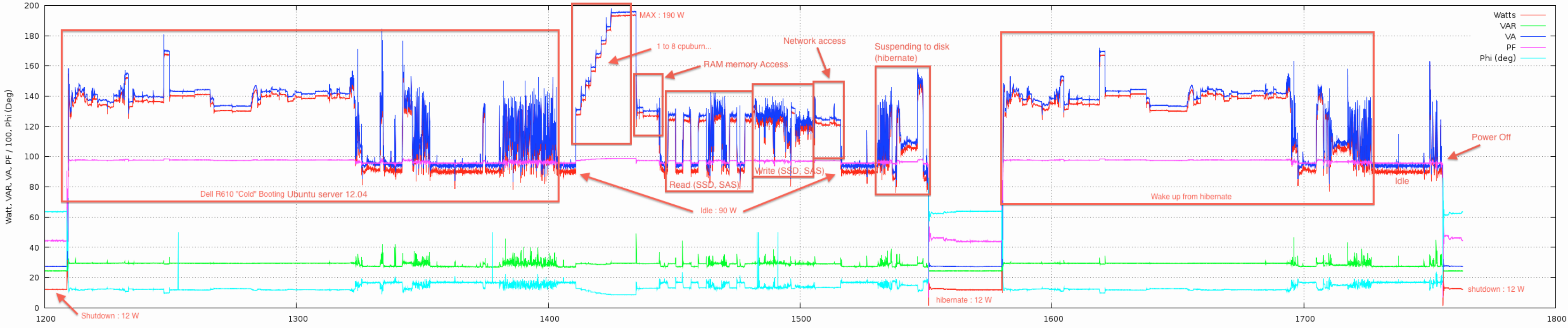


Power profiling : some good old servers (2009)

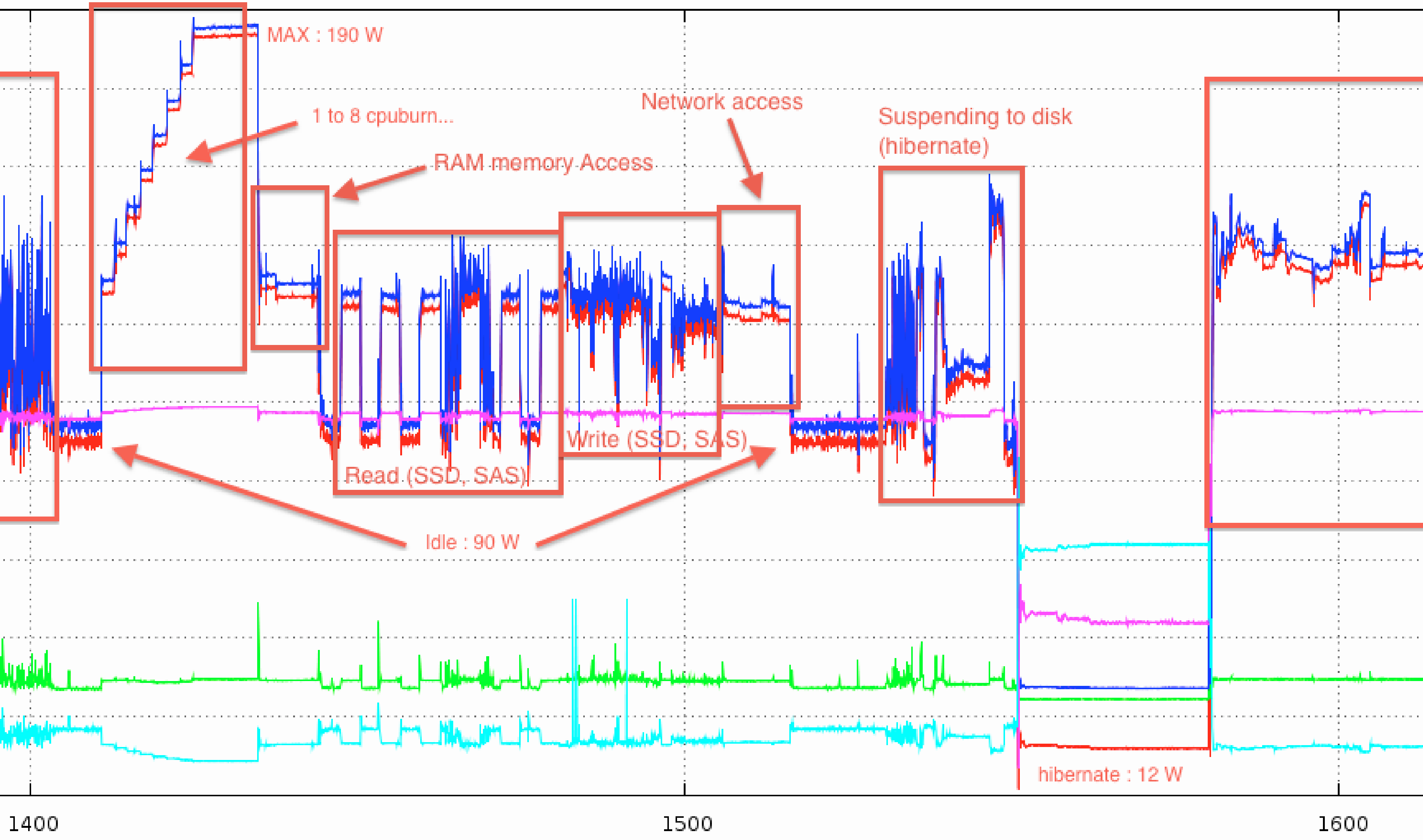


Easy to analyze, easy to understand, no cores only CPUs…

Power profiling of a more recent server



Dell R610 - Zimmer LMG450 - watts in red


MAX : 190 W
1 to 8 cpuburn...
RAM memory Access
Network access
Suspending to disk
(hibernate)
Read (SSD, SAS)
Write (SSD, SAS)
Idle : 90 W
hibernate : 12 W
1400
1500
1600

Energy proportionality

*Luiz André Barroso and Urs Hölzle, « **The case for Energy-Proportional Computing** », IEEE Computer, 2007*

<u>At servers level :</u>

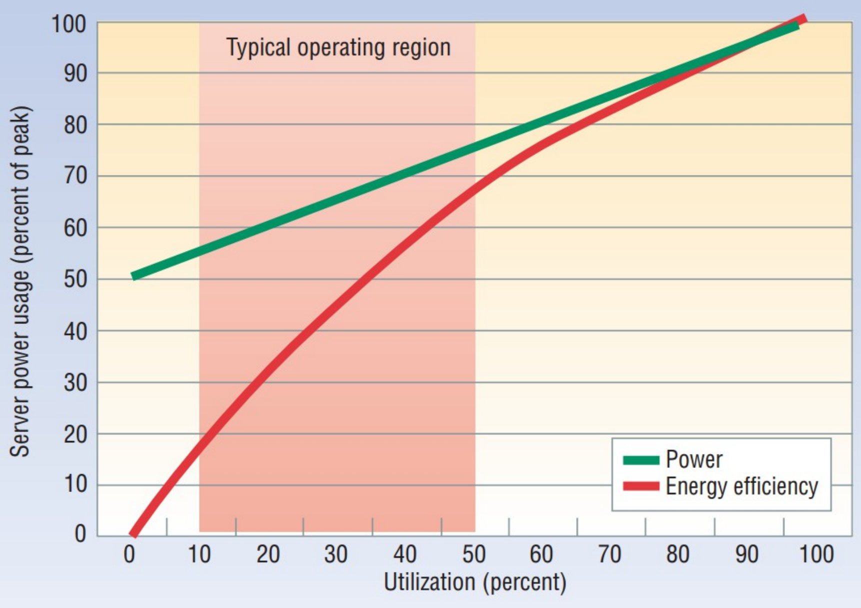
Idle power consumption

Inefficient region depending on load

<u>At network level :</u>

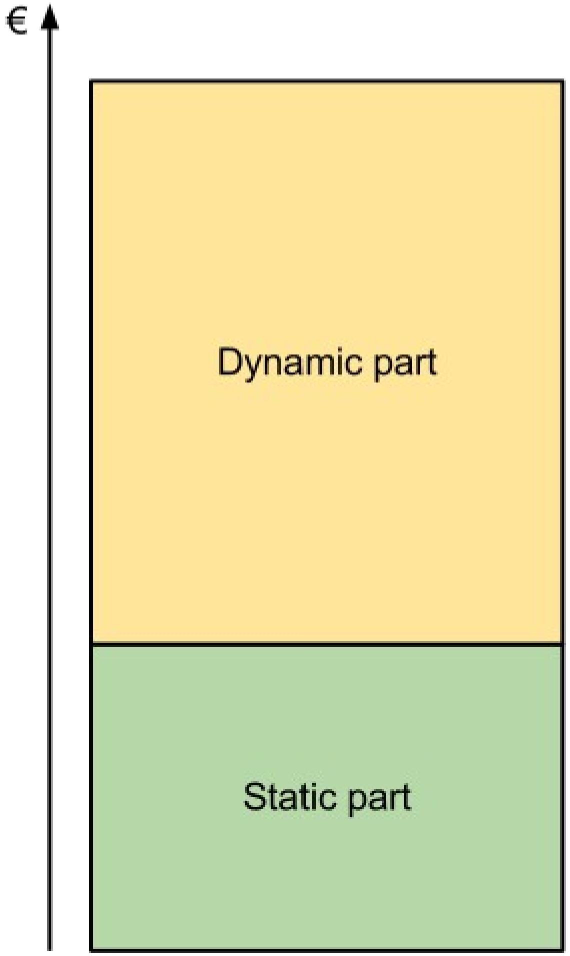
Even less proportional

Switches energy consumption almost consta



Energy consumption and energy efficiency of a server according to its load

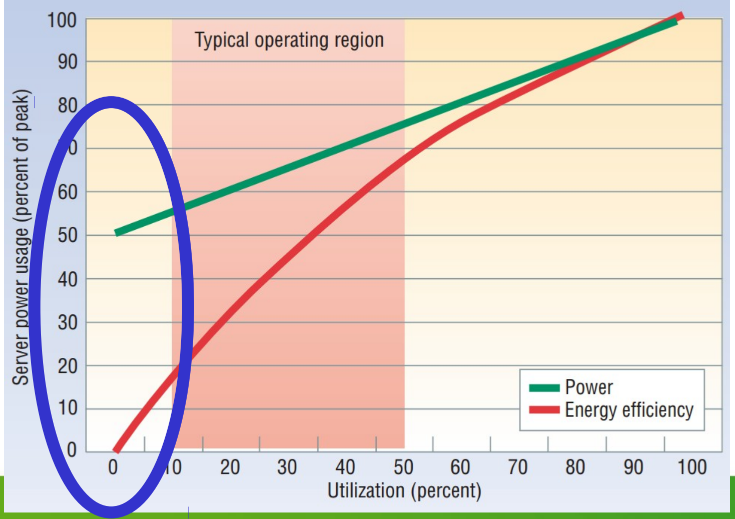
Static / dynamic part of power



Even reducing a lot some static part can remain important

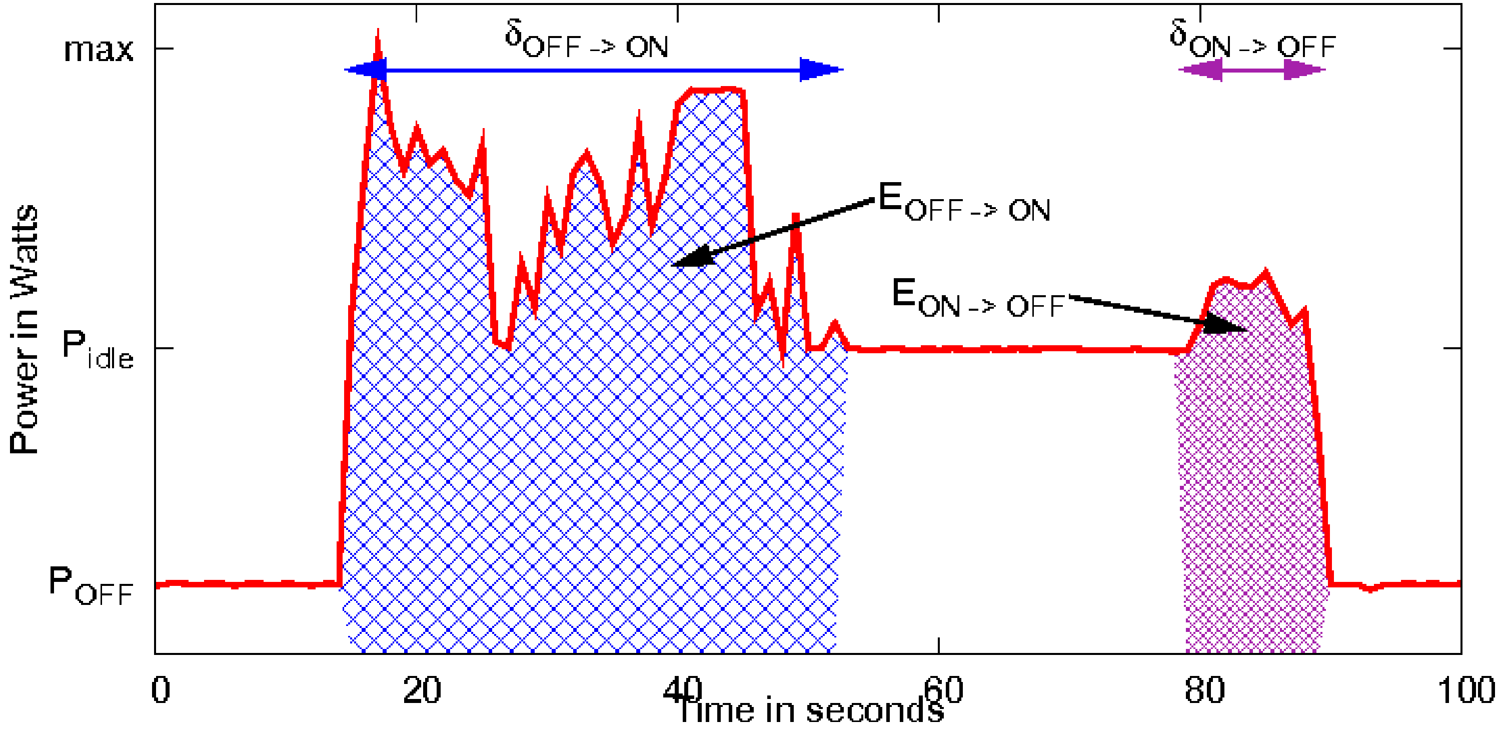
from GOS : 240 W / 260 W (92%) to recent one 90W / 190W (47%)

First LHF : switch off unused resources : delete the static part !



Aggressive ON/OFF is not always the best solution

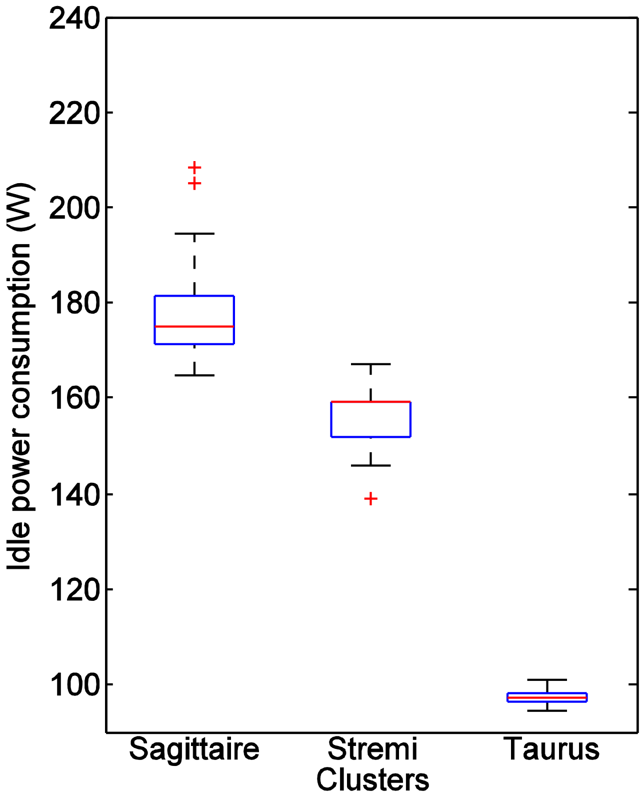
- Exploiting the gaps between activities to reduce unused plugged ressources number
- But only switiching off → if potential energy saving
- ON -> OFF can be really long (at large scale)

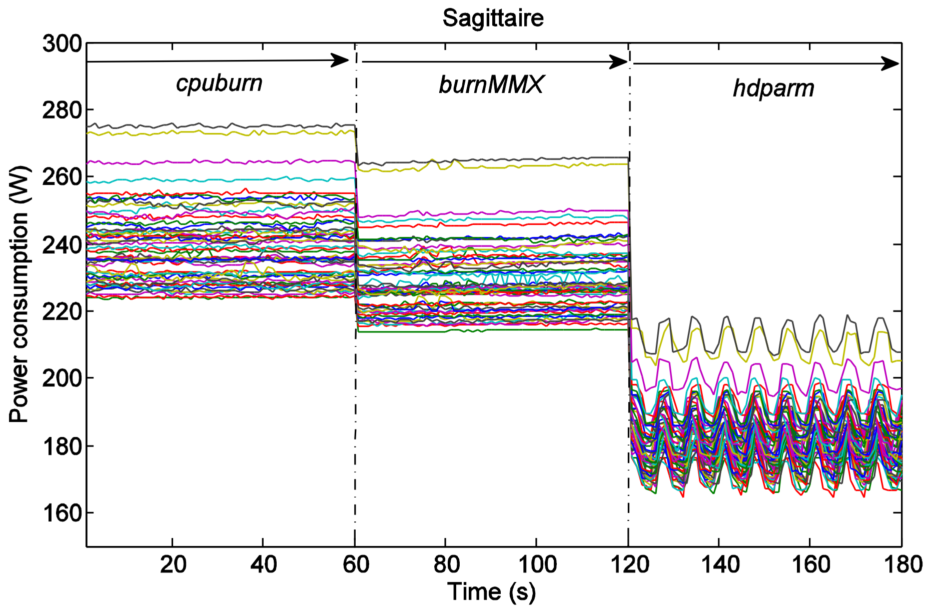


Anne-Cecile Orgerie, Laurent Lefevre, and Jean-Patrick Gelas. "Save Watts in your Grid: Green Strategies for Energy-Aware Framework in Large Scale Distributed Systems", ICPADS 2008 : The 14th IEEE International Conference on Parallel and Distributed Systems, Melbourne, Australia, December 2008

Other difficulty : homogeneity (in energy consumption) does not exist ! Must switch off/on the right resource

- Depends on technology
- Same flops but not same flops per watt
- Idle / static cost
- CPU : main responsible





Mohammed el Mehdi Diouri, Olivier Gluck, Laurent Lefevre and Jean-Christophe Mignot. **"Your Cluster is not Power Homogeneous: Take Care when Designing Green Schedulers!"**, *IGCC2013 : International Green Computing Conference*, Arlington, USA, June 27-29,

Reservation based Openstack Clouds

Switching off and on is difficult and complex at large scale without good prediction

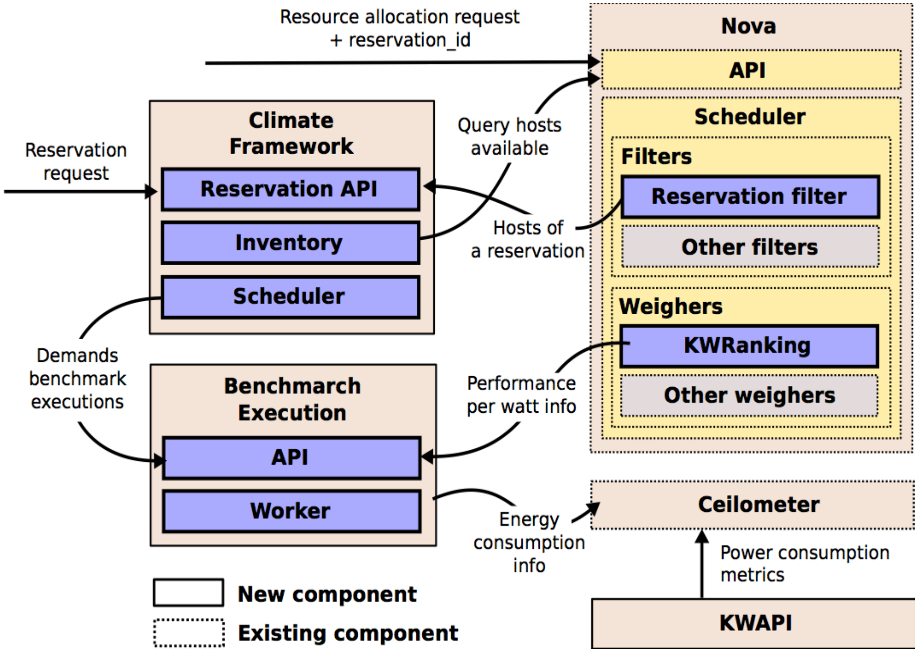
Avoiding on-demand & overpovisioning

Needs of scheduling and planification -> need of reservation based systems

FSN XLCLOUD Project (2012-2015)

Partners : Bull SAS, Serviware, Institut Telecom, HPC-Project, CEA List, EISTI, ATEME, OW2, Inria

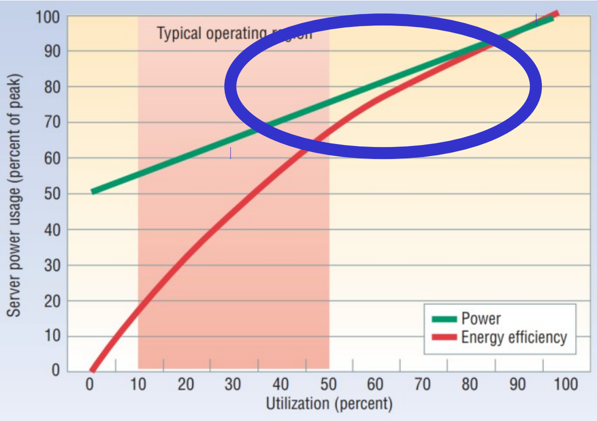
Target : HPC as a service : supporting HPC applications with interactive remote visualization in energy efficient Cloud : GPUs, Infiniband… etc…



http://xlcloud.org/

Climate / Blazar project : capacity leasing in Openstack (Inria, Bull, Mirantis)

Address the dynamic part with green levers : adapt resources to the need of applications



HPC applications keep growing in complexity : too many bugs in HPC applications already present, adding energy management and considerations won't help :=)

Are HPC programmers ready for eco design of applications ?

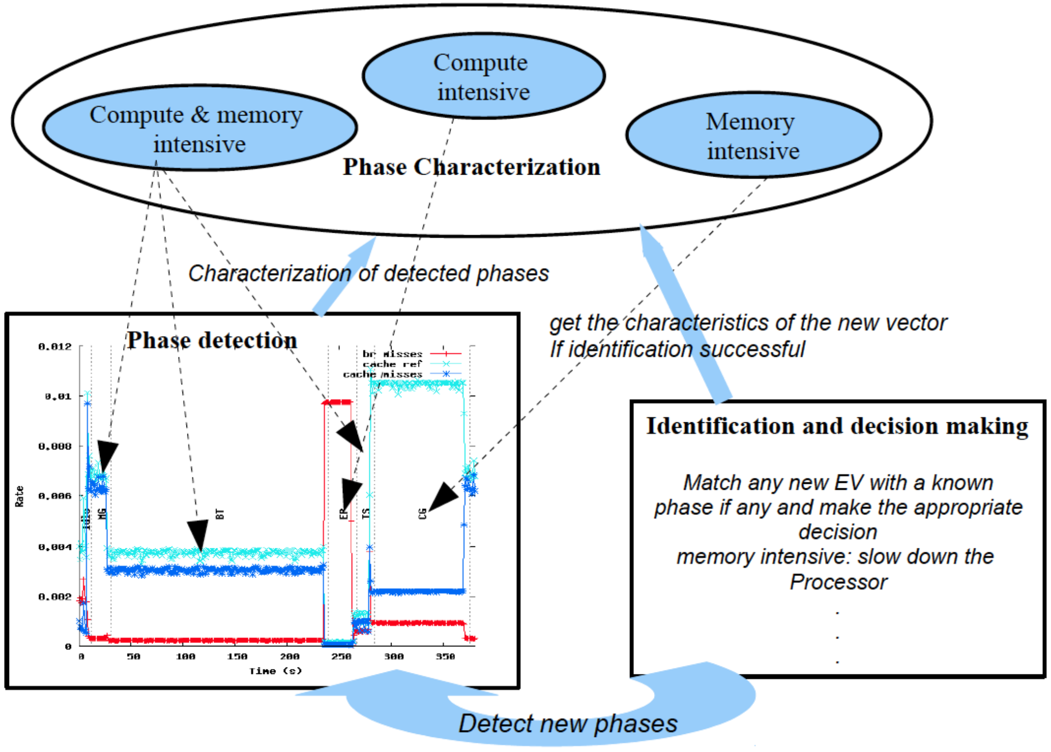
Applications can share the same infrastructure : Optimizations made for saving energy considering some applications are likely to impact the performance of others

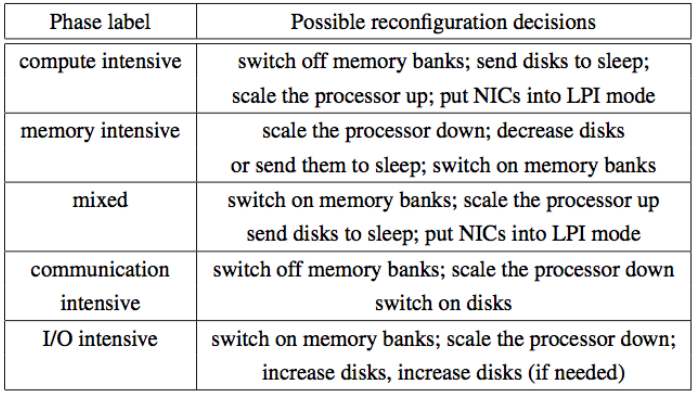
Instead of looking at applications and service => Focusing on the infrastructure

- Detect and characterize system's runtime behaviours/phases
- Optimize each subsystem (storage, memory, interconnect, CPU) accordingly

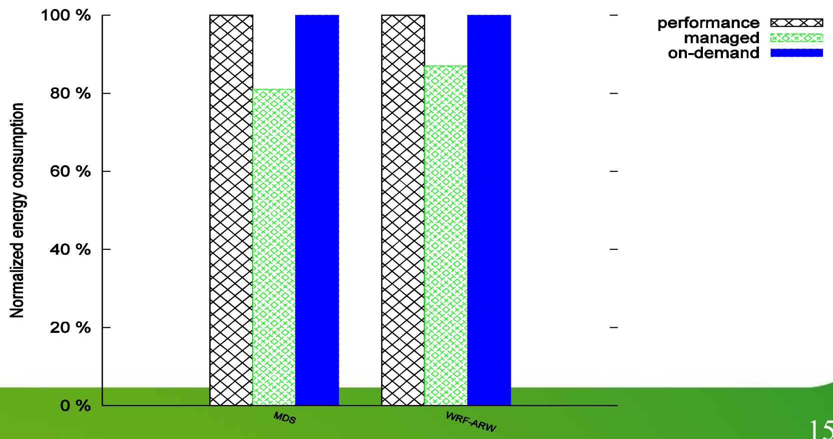
Online analysis without knowledge on applications

- Irregular usage of resources
- Phase detection, characterisation
- Power saving modes deployment
- MREEF framework


Compute intensive
Compute & memory intensive
Memory intensive
Phase Characterization
Characterization of detected phases
get the characteristics of the new vector
If identification successful
Phase detection
0.012
0.01
0.008
0.006
0.004
0.002
0
Rate
0
50
100
150
200
250
300
350
Time (s)
br misses
cache ref
cache misses
Idle
MG
BT
EP
IS
CG
Identification and decision making
Match any new EV with a known phase if any and make the appropriate decision
memory intensive: slow down the Processor
.
.
.
Detect new phases



Phase label	Possible reconfiguration decisions
compute intensive	switch off memory banks; send disks to sleep; scale the processor up; put NICs into LPI mode
memory intensive	scale the processor down; decrease disks or send them to sleep; switch on memory banks
mixed	switch on memory banks; scale the processor up send disks to sleep; put NICs into LPI mode
communication intensive	switch off memory banks; scale the processor down switch on disks
I/O intensive	switch on memory banks; scale the processor down; increase disks, increase disks (if needed)

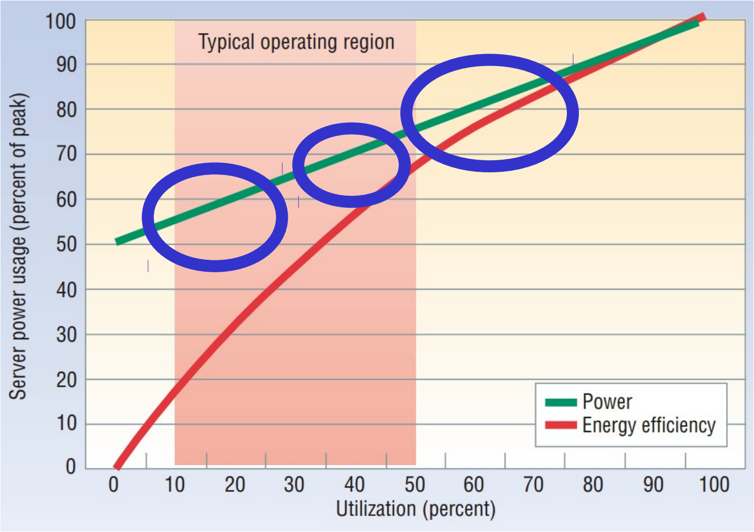


Landry Tsafack, Laurent Lefevre, Jean-Marc Pierson, Patricia Stolf, and Georges Da Costa. **"A runtime framework for energy efficient HPC systems without a priori knowledge of applications"**, *ICPADS 2012 : 18th International Conference on Parallel and Distributed Systems*, Singapore, December 2012

What about missing parts of the curve ?

- Specific conditions of workload
- Gaps between bursts

- Exploiting heterogeneity of processors (flops, watts, flops per watt) to fill the missing parts



Heterogeneous multicore processors

ARM big.LITTLE

2 processors (4 cores each) :

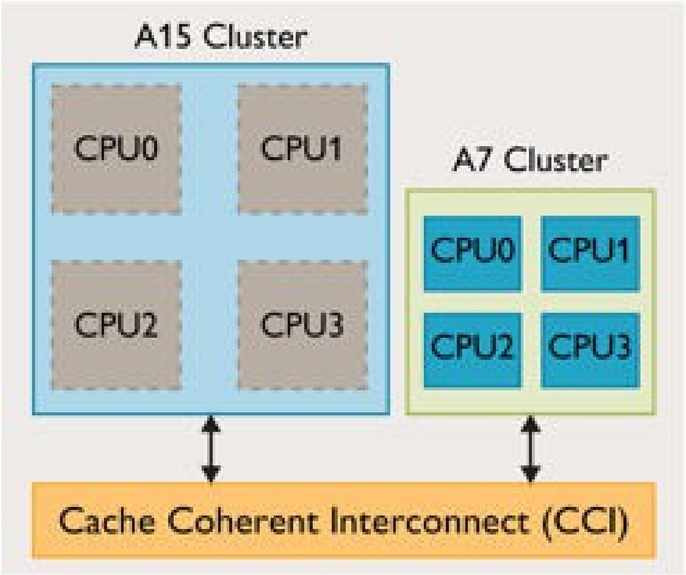
- **LITTLE** (Cortex A7)
- **big** (Cortex A15)

Interconnected by a Cache Coherence system

GOAL→ Extend battery life time of mobile devices which are idle most of the time

Some utilization modes :

- Cluster migration (4 / 4)
- Global Task Scheduling (8)



big.LITTLE « Cluster migration »

Heterogeneous architectures

A the scale of a datacenter → ARM may be not enough
We could need real performance to absorb load peaks

Exploring a combination of :

Low-power processors for low load
and
high performance processors for heavy load

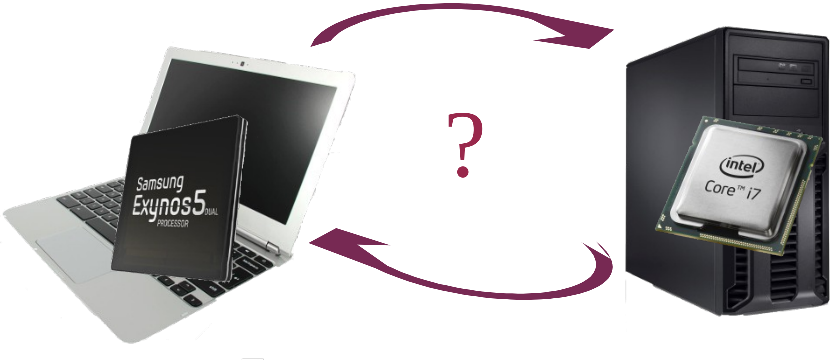
→ reduces static costs

→ use classical servers only at their most energy efficient load level

✚ other classical levers : DVFS, switch off/on,… to improve consumption proportionality

Technical challenges

- Different architectures : ARM and x86



How to combien them and be able to go from one architecture to another ?

- live migration without impact on the moving application

- migration fastest as possible

→ First idea Classical cloud approach : Virtual machines

2 physical architectures → 2 choices for virtual machine architecture

When the VM is not on the right physical architecture, we use emulation with QEMU

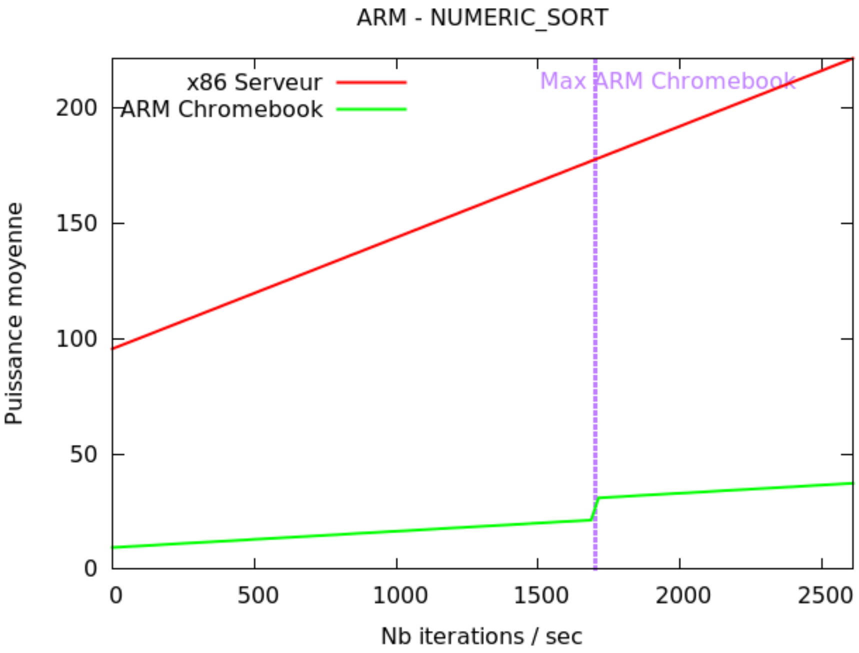
→ What is the cost of emulation ?

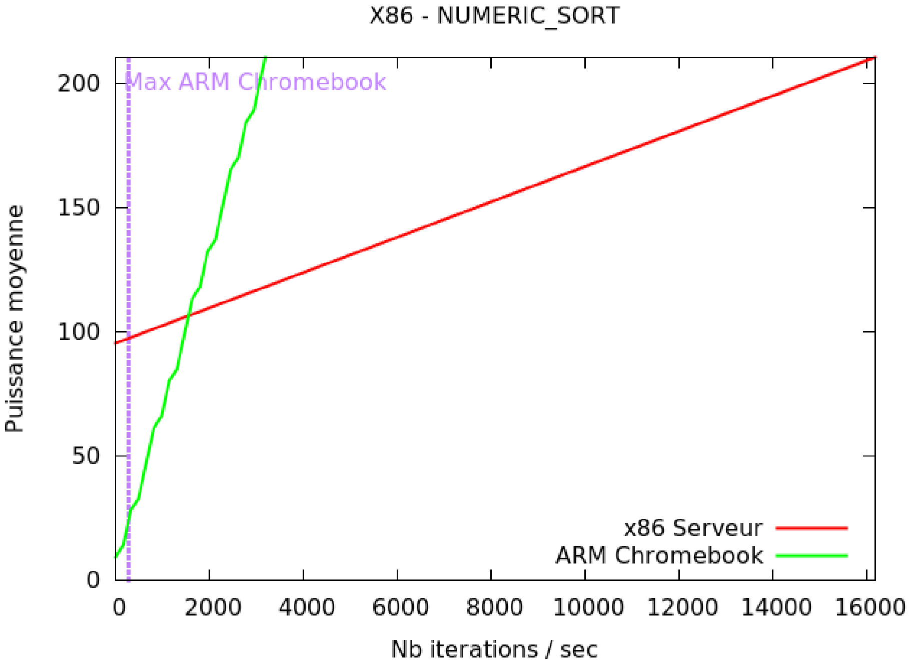
→ Which architecture to choose for the VM ?

Comparison of VM architecture – First results

- ARM VM:
  Native on ARM processor
  Emulated on x86 processor

- X86 VM :
  Native on x86 processor
  Emulated on ARM processor





ARM : Samsung Chromebook (2 processors ARM Cortex-A15)

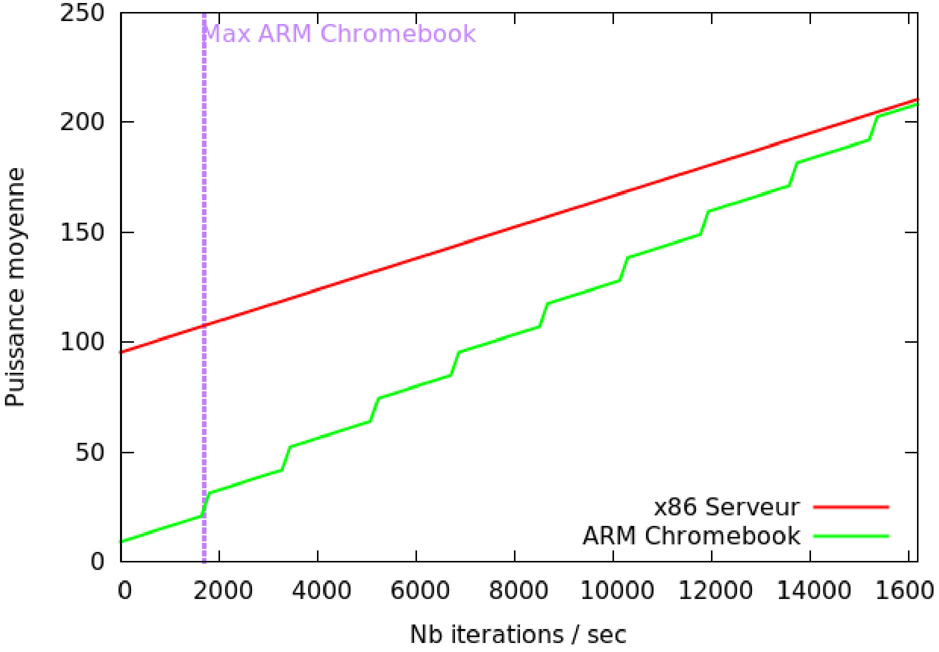
x86 : Dell PowerEdge R720 (2 processors Intel Xeon 6 cores)

Benchmark nbench : integer/float

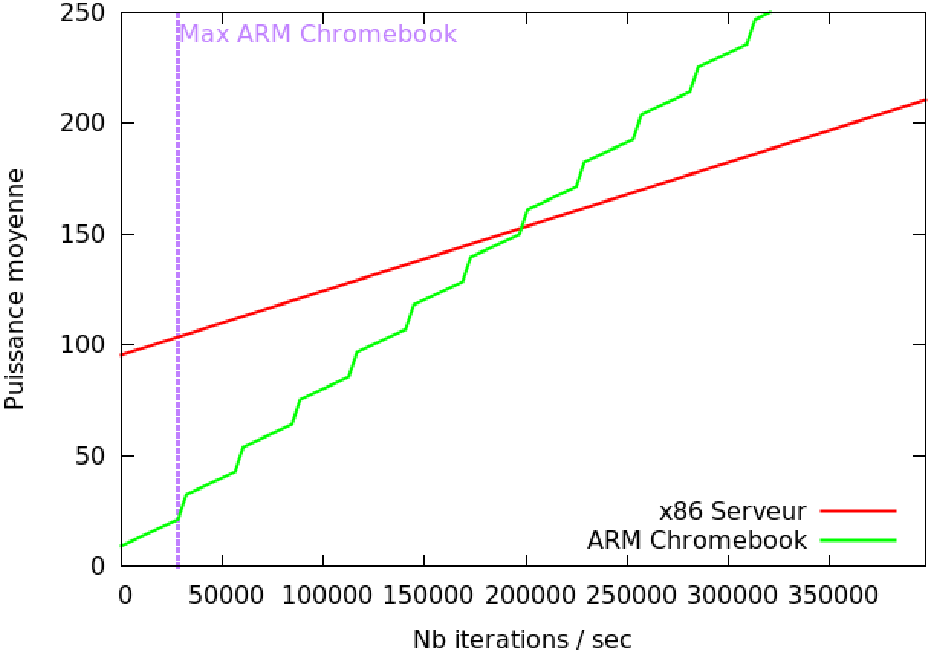
Comparison of native performances – First results

- If we can benefit from native performances of each architecture, what is the impact on proportionality

Native - NUMERIC_SORT *(Integer)*

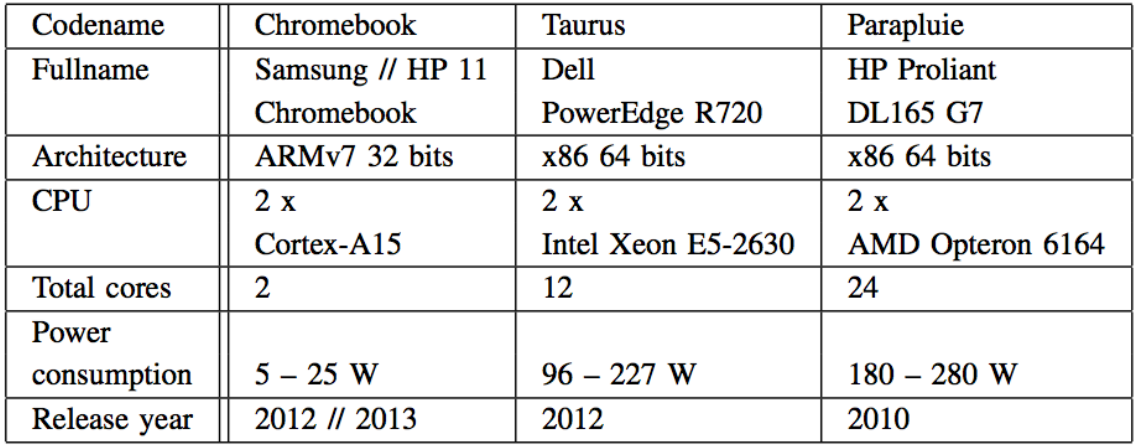


Native - FOURIER *(Float)*

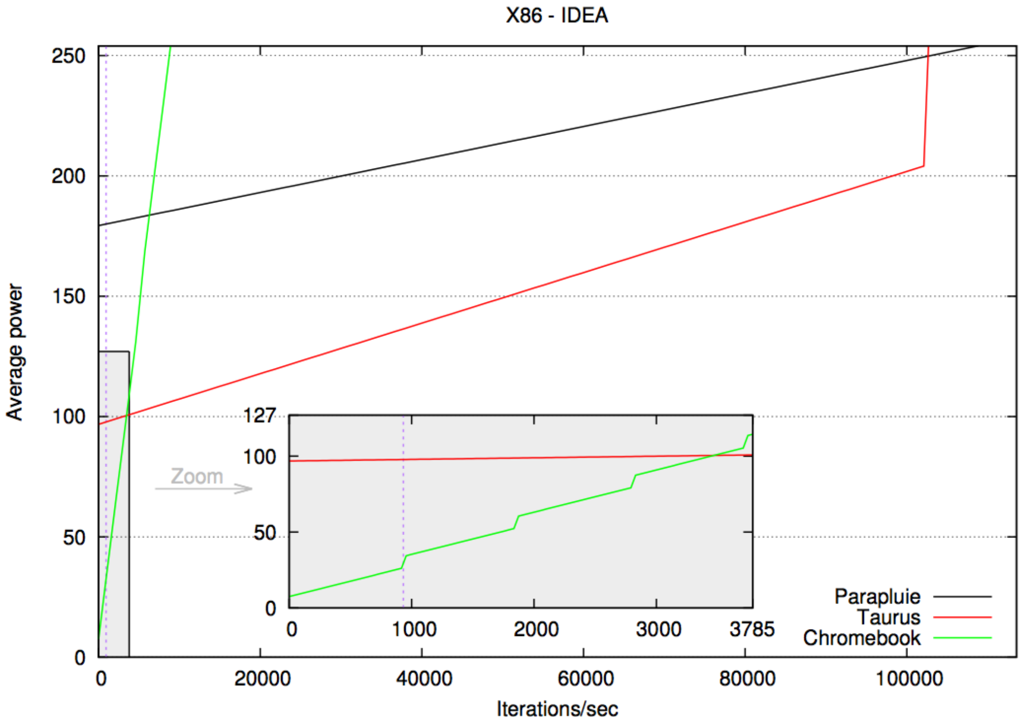


Number of Chromebook before reaching x86 server performances 10 7

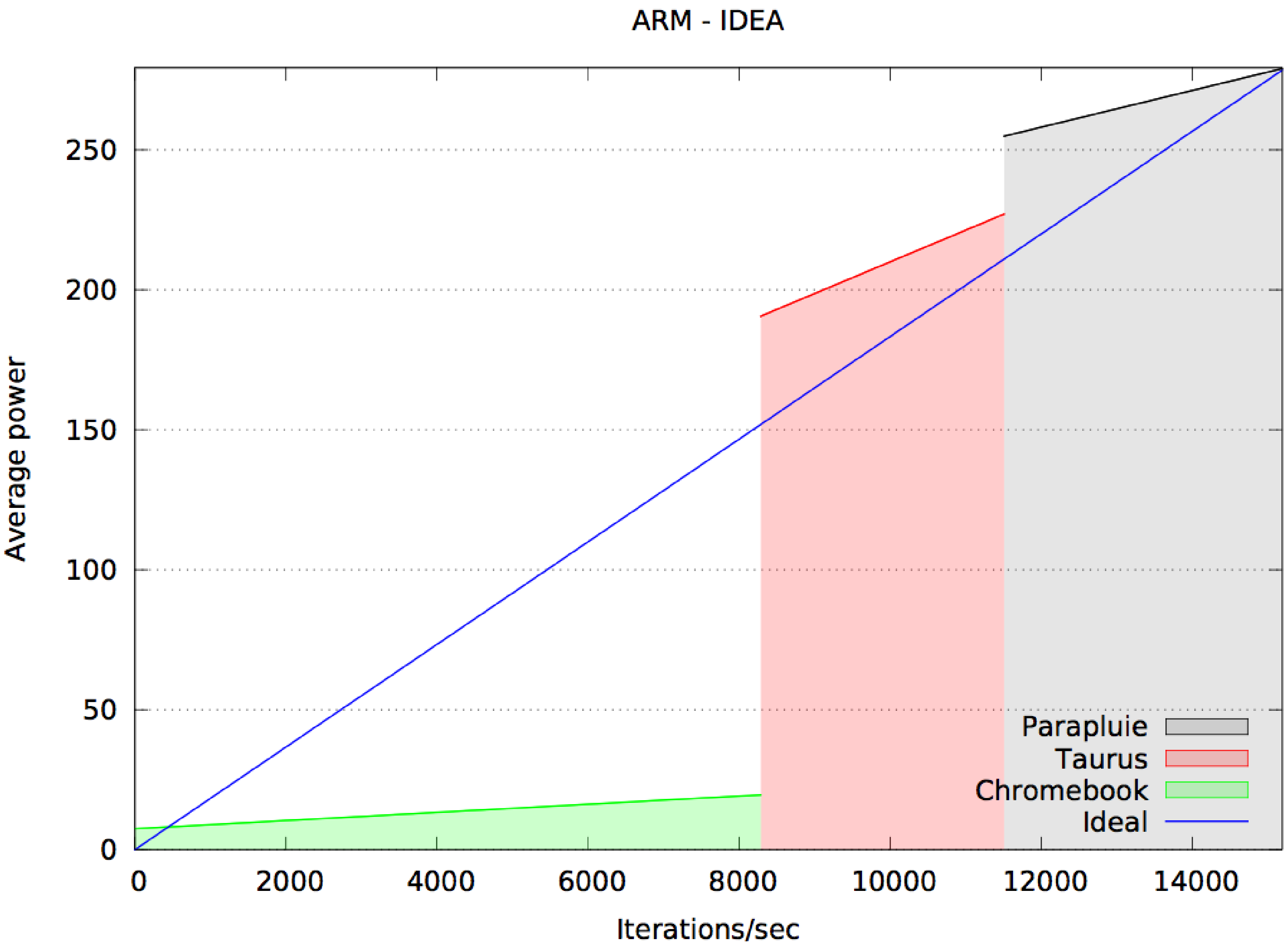
Comparison of VM performances – First results



Codename	Chromebook	Taurus	Parapluie
Fullname	Samsung // HP 11 Chromebook	Dell PowerEdge R720	HP Proliant DL165 G7
Architecture	ARMv7 32 bits	x86 64 bits	x86 64 bits
CPU	2 x Cortex-A15	2 x Intel Xeon E5-2630	2 x AMD Opteron 6164
Total cores	2	12	24
Power consumption	5 – 25 W	96 – 227 W	180 – 280 W
Release year	2012 // 2013	2012	2010



Comparison of native performances – First results



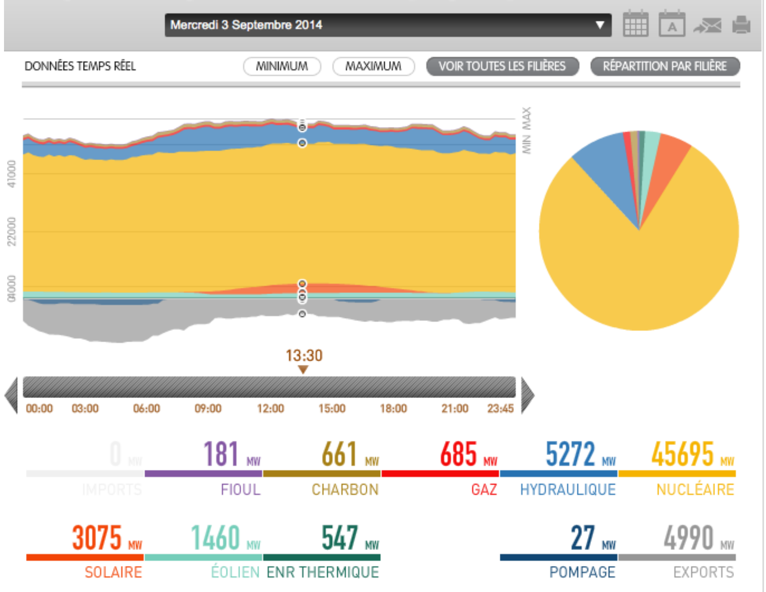
Still some work to do to reach a nice energy proportional curve

Don’t say !

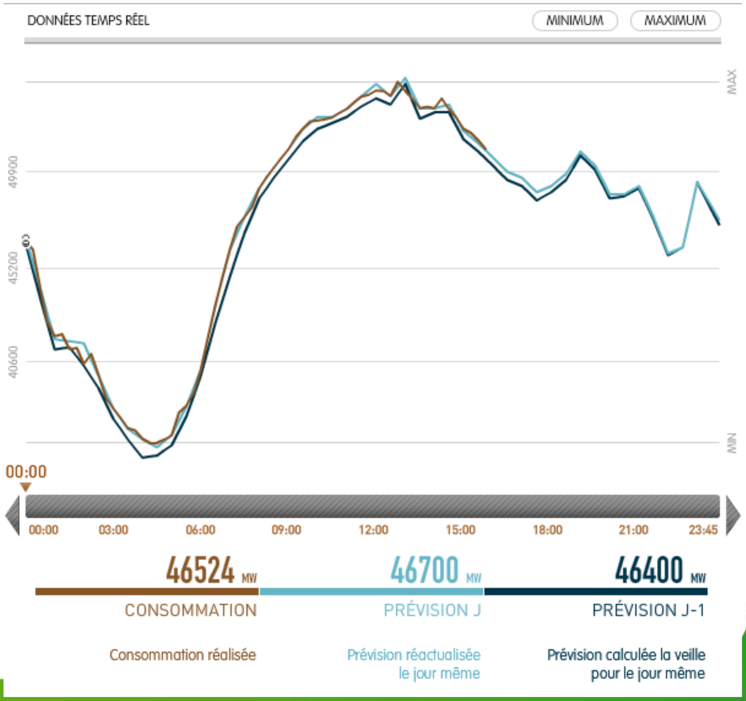
- Not possible « I need tu use a constant power »
Ex : power usage in France yesterday
-> negociate with your provider – combine reservation/prediction

-Not possible « my DC needs to consume a minimum amount of power » -> renegociate your contract

-Not possible, when my machines (re)boot I face too much risks -> negociate with your system designer, add resilience solutions (see Yves for that)



0 MW IMPORTS
181 MW FIOUL
661 MW CHARBON
685 MW GAZ
5272 MW HYDRAULIQUE
45695 MW NUCLÉAIRE
3075 MW SOLAIRE
1460 MW ÉOLIEN
547 MW ENR THERMIQUE
27 MW POMPAGE
4990 MW EXPORTS



46524 MW CONSOMMATION
Consommation réalisée

46700 MW PRÉVISION J
Prévision réactualisée le jour même

46400 MW PRÉVISION J-1
Prévision calculée la veille pour le jour même

Current Challenges

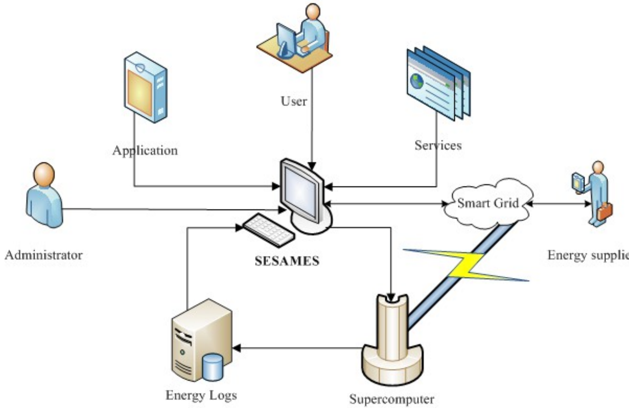
- Large scale frequent energy monitoring remains a challenge
  - Data deluge of energy info
  - Energy sensors : less interest for external monitoring (too much cores) - relying on internal sensors (quality, intrusiveness…)

- Possible supported scenario :
  - Cloud with workload variations
  - HPC with batch jobs

- Large scale energy variations : need live exchange with energy provider

- Need to adapt software and infrastructures to support computing power jitter and resilience to boot failures

- Full lifecycle of EP IT : from design, transport, deployment, usage, destroying, recycling



M. Diouri, O. Gluck, and L. Lefevre. "Towards a novel Smart and Energy-Aware Service-Oriented Manager for Extreme-Scale applications, First Workshop for Power Grid-Friendly Computing (PGFC'12), San Jose, USA, June 2012

INVENTEURS DU MONDE NUMÉRIQUE