INVENTEURS DU MONDE NUMÉRIQUE

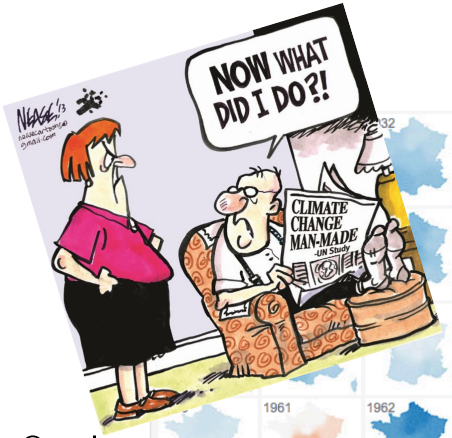
Efficacité énergétique.. L'aventure Green 5000
A la recherche de la proportionalité énergétique…

Laurent Lefèvre
Inria Avalon / LIP Lab / Ecole Normale Supérieure de Lyon
laurent.lefevre@inria.fr

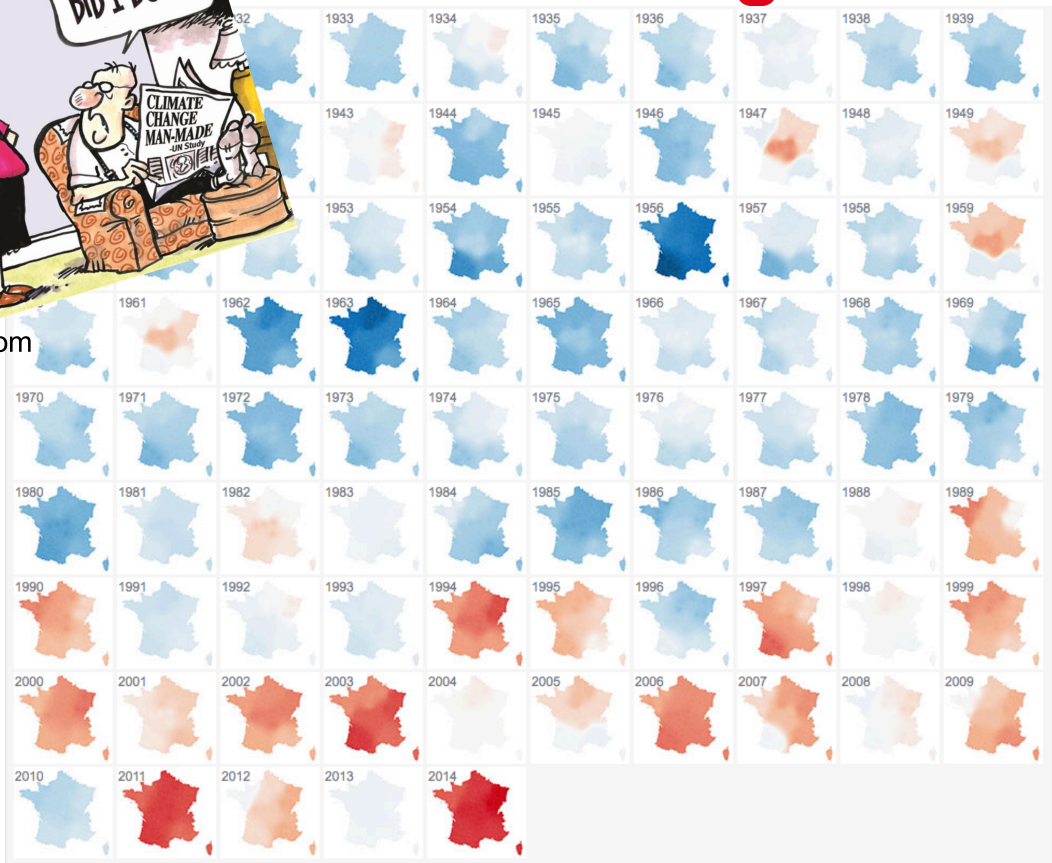
SUCCES 2015

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

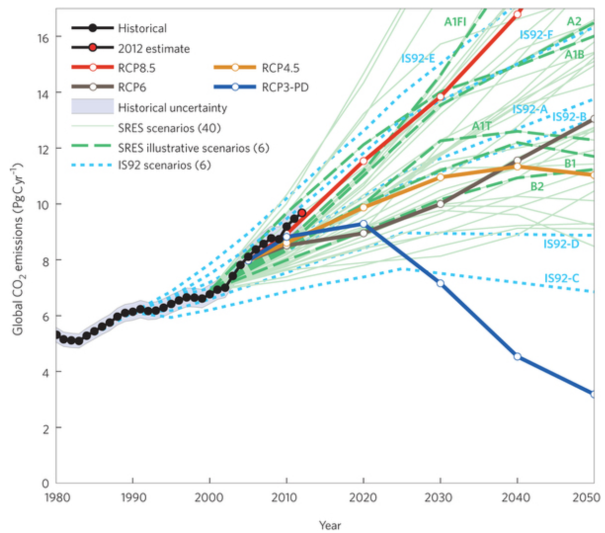
An urgent issue !



©cagle.com



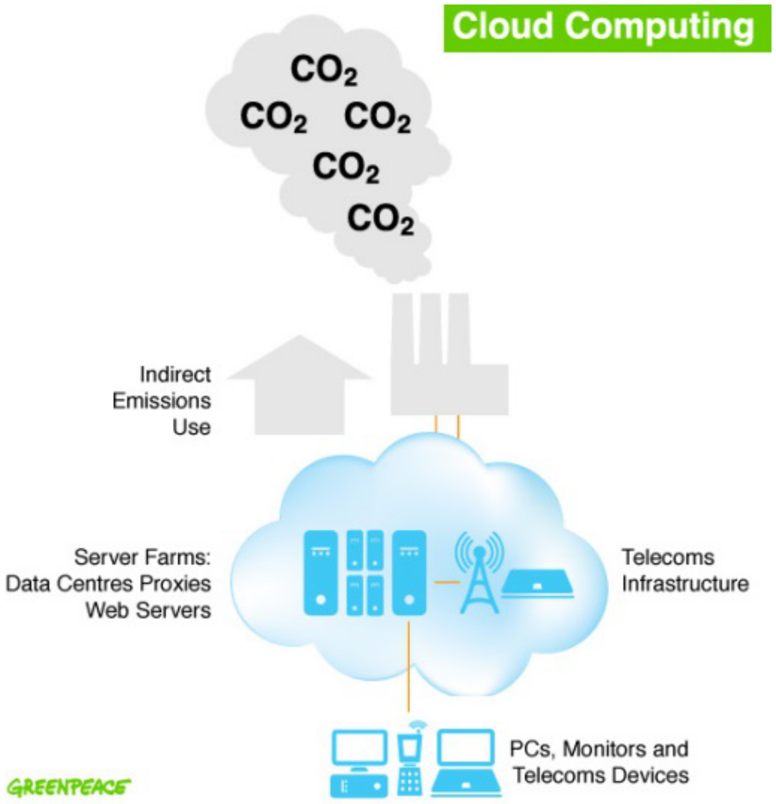
COP21 · CMP11
PARIS 2015
UN CLIMATE CHANGE CONFERENCE

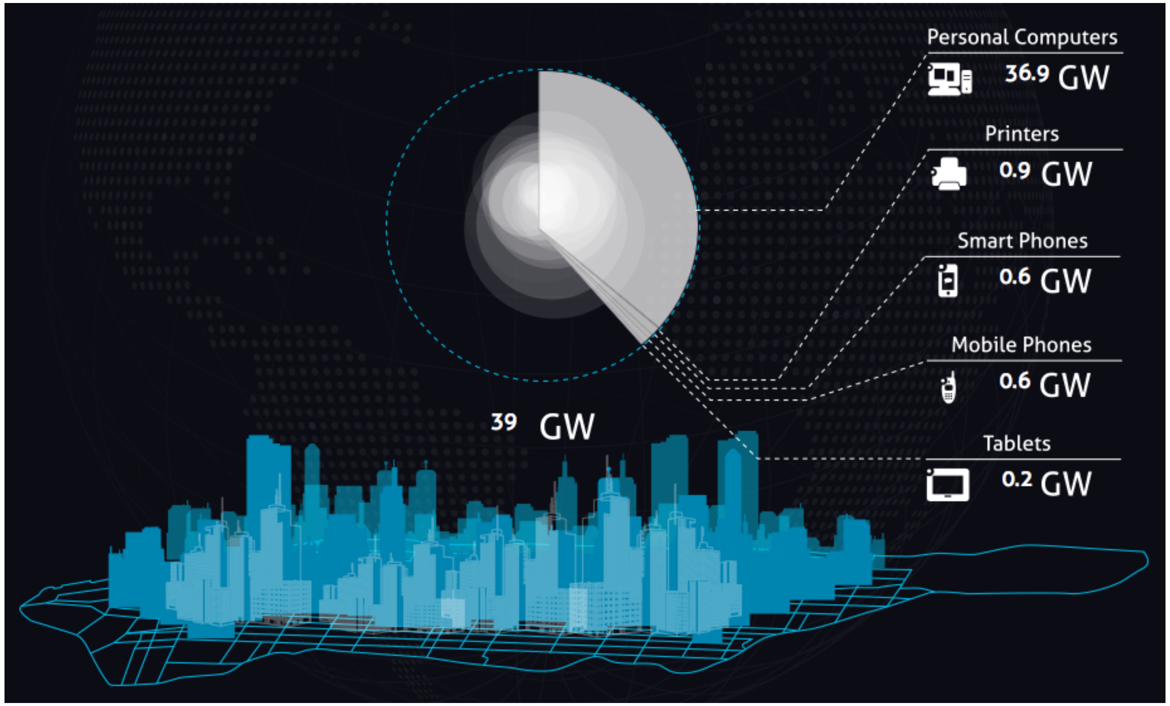


Links with ICT ?

Electrical consumption of ICT….

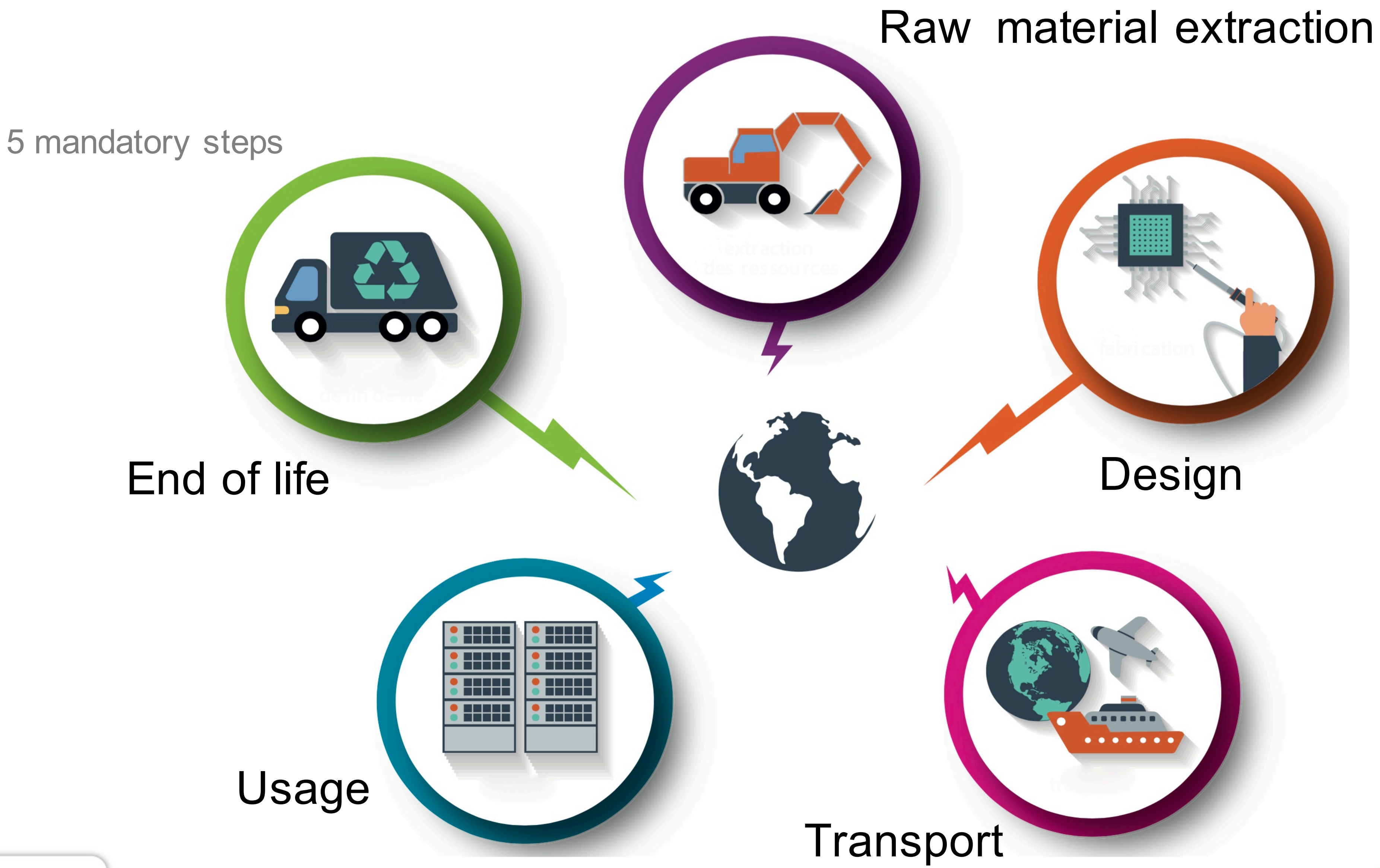
2013… gwatt.net





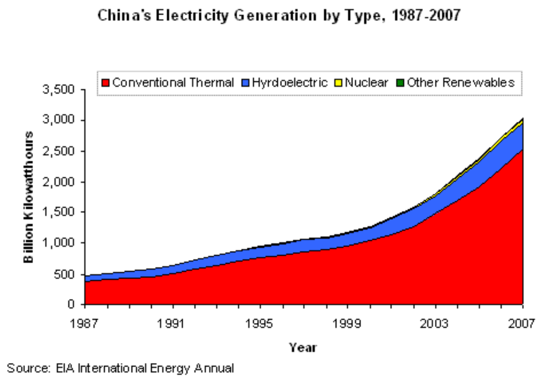
EcoInfo
http://ecoinfo.cnrs.fr

Lifecycle of ICT

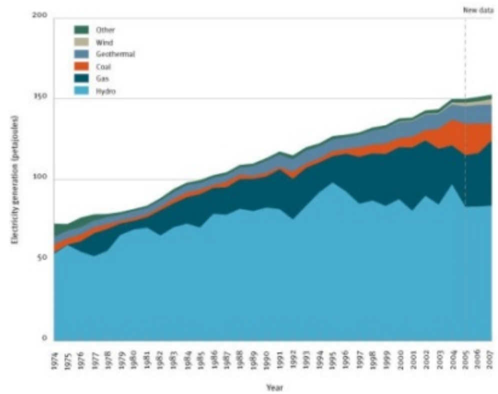


Energy : 1st limiting factor for large scale systems ((hpc)datacenter, clouds, internet)?

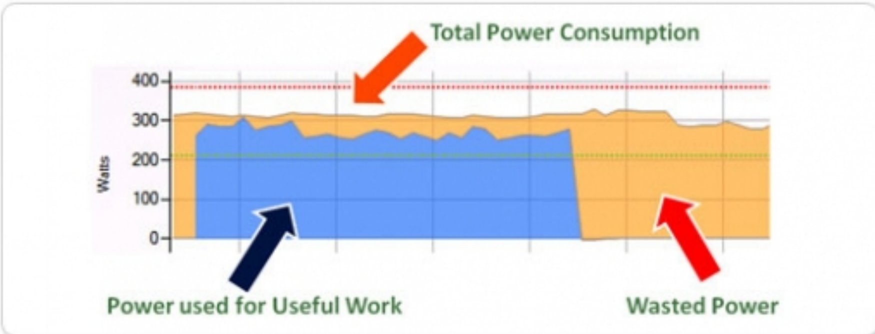
- ICT – 5% of CO2 emissions
- Green IT ➔ reducing electrical consumption of IT equipment - CO2 impact depends on countries
- But, IT by definition is not clean - Bad lifecycle : planned obsolescence, rebound effect, design (rare minerals), recycling...



- Future exascale/datacenters platforms -> systems from 20 100MW



- How to build such systems and make them (more) energy sustainable/responsible ? - Multi dimension approaches : hardware, software, usage



top500.org

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	2,325
7	King Abdullah University of Science and Technology Saudi Arabia	**Shaheen II** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	196,608	5,537.0	7,235.2	2,834
8	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462,462	5,168.1	8,520.1	4,510
9	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458,752	5,008.9	5,872.0	2,301
10	DOE/NNSA/LLNL United States	**Vulcan** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393,216	4,293.3	5,033.2	1,972

Exascale race !

green500.org

Listed below are the June 2015 The Green500's energy-efficient supercomputers ranked from 1 to 10.

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	7,031.58	RIKEN	Shoubu - ExaScaler-1.4 80Brick, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SC	50.32
2	6,842.31	High Energy Accelerator Research Organization /KEK	Suiren Blue - ExaScaler-1.4 16Brick, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband, PEZY-SC	28.25
3	6,217.04	High Energy Accelerator Research Organization /KEK	Suiren - ExaScaler 32U256SC Cluster, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, PEZY-SC	32.59
4	5,271.81	GSI Helmholtz Center	ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150	57.15
5	4,257.88	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	39.83
6	4,112.11	Stanford Research Computing Center	XStream - Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	190.00
7	3,962.73	Cray Inc.	Storm1 - Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40m	44.54
8	3,631.70	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
9	3,614.71	TU Dresden, ZIH	Taurus GPUs - Bull bullx R400, Xeon E5-2680v3 12C 2.5GHz, Infiniband FDR, Nvidia K80	58.01
10	3,543.32	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	54.60

* Performance data obtained from publicly available sources including TOP500

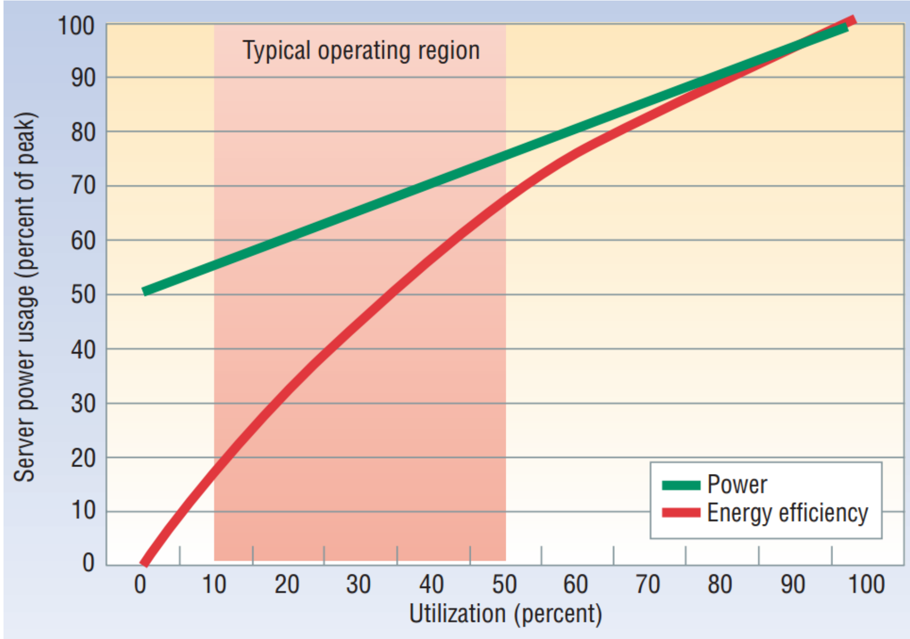
Exascale race ? Sustainable ?

Energy Proportionality

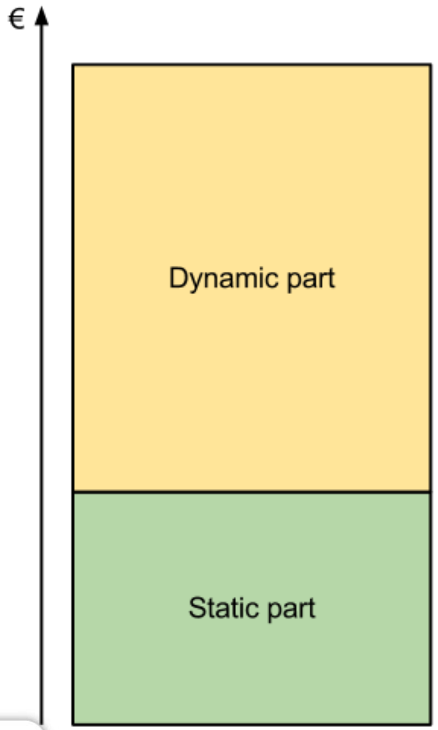
Luiz André Barroso and Urs Hölzle, « The case for Energy-Proportional Computing », IEEE Computer, 2007

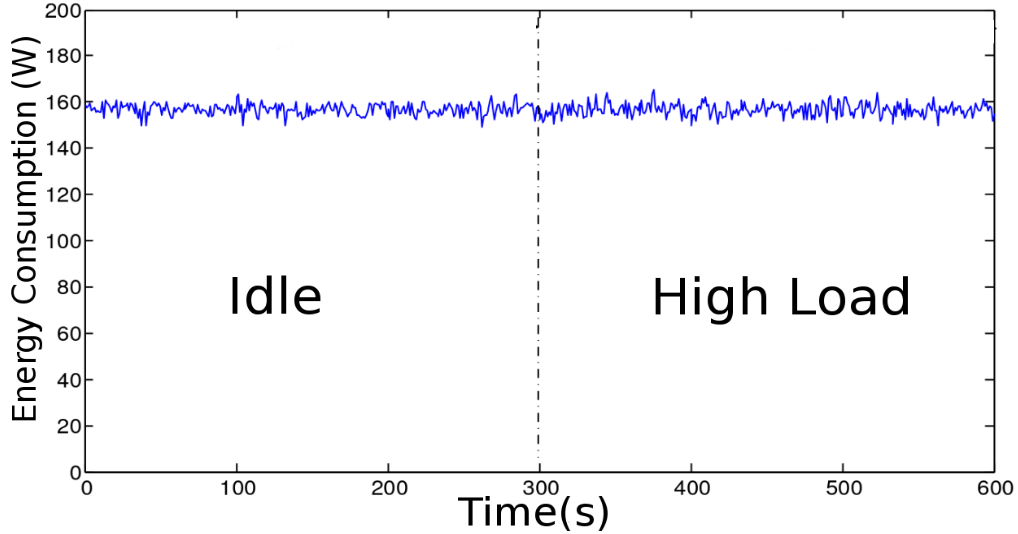
Servers : *idle* consumption loosely proportional

Network equipment : consumption unaware of usage

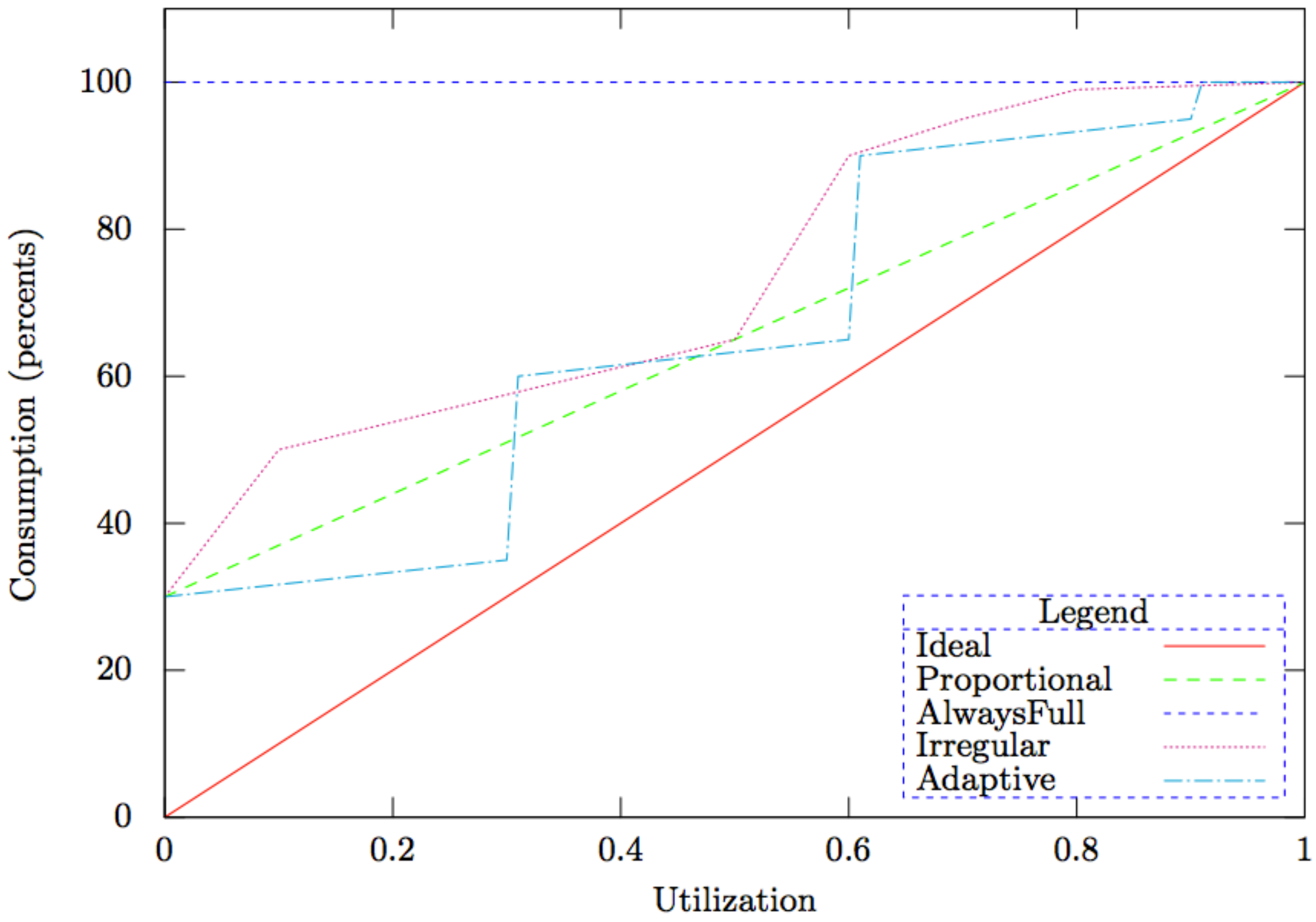


Energy consumption and energy efficiency of a server according to its load





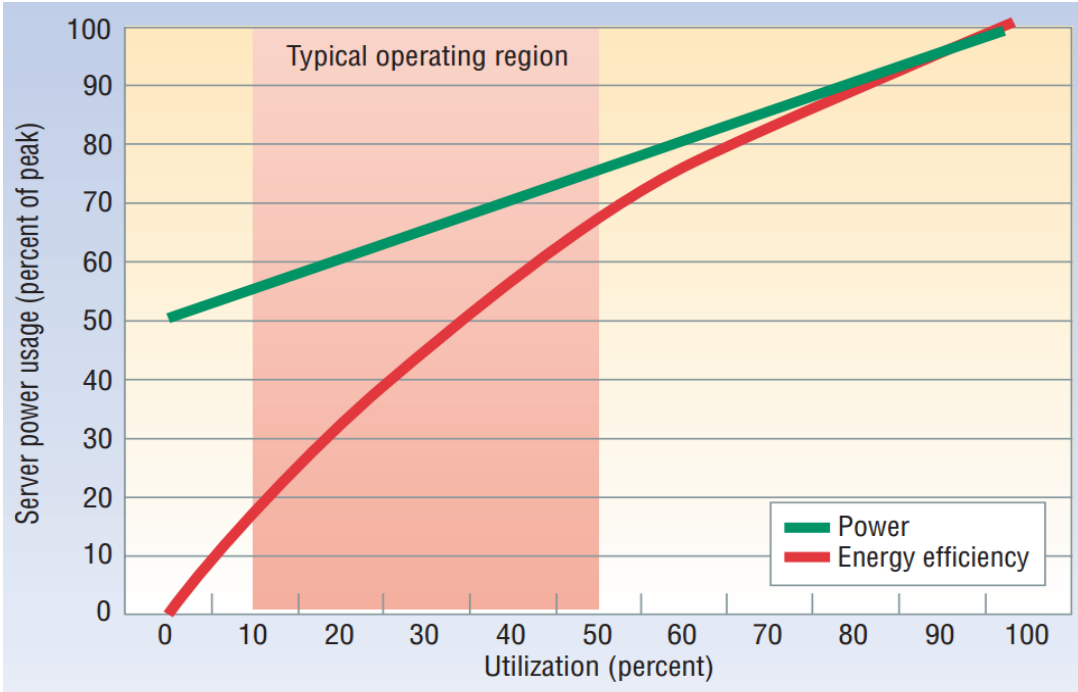
EP in reality



Green capabilities !

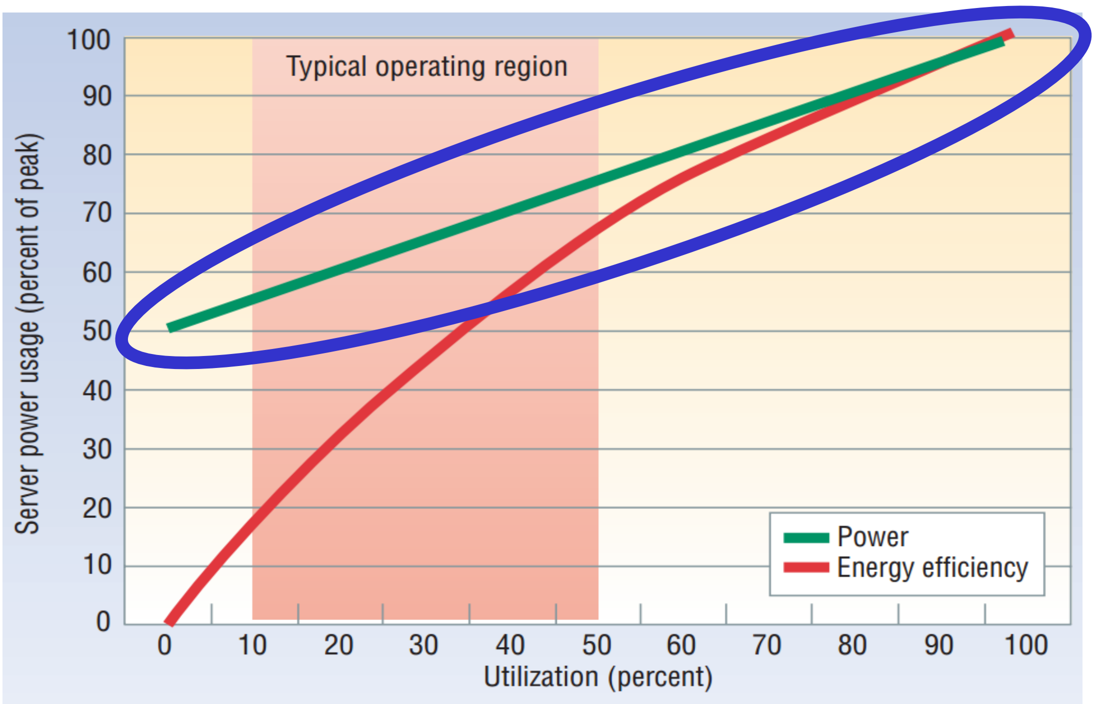
- **Shutdown** : to reduce the number of useless plugged resources
- **Slowdown** : to adapt performances (and energy consumption) to the real needs of applications, services and protocols
- **Optimizing** : to modify application sand services in order to malke them greener
- **Consolidating / aggregating** : relocate services and applications on smaller number of physical resources

But where to focus on the curve ?



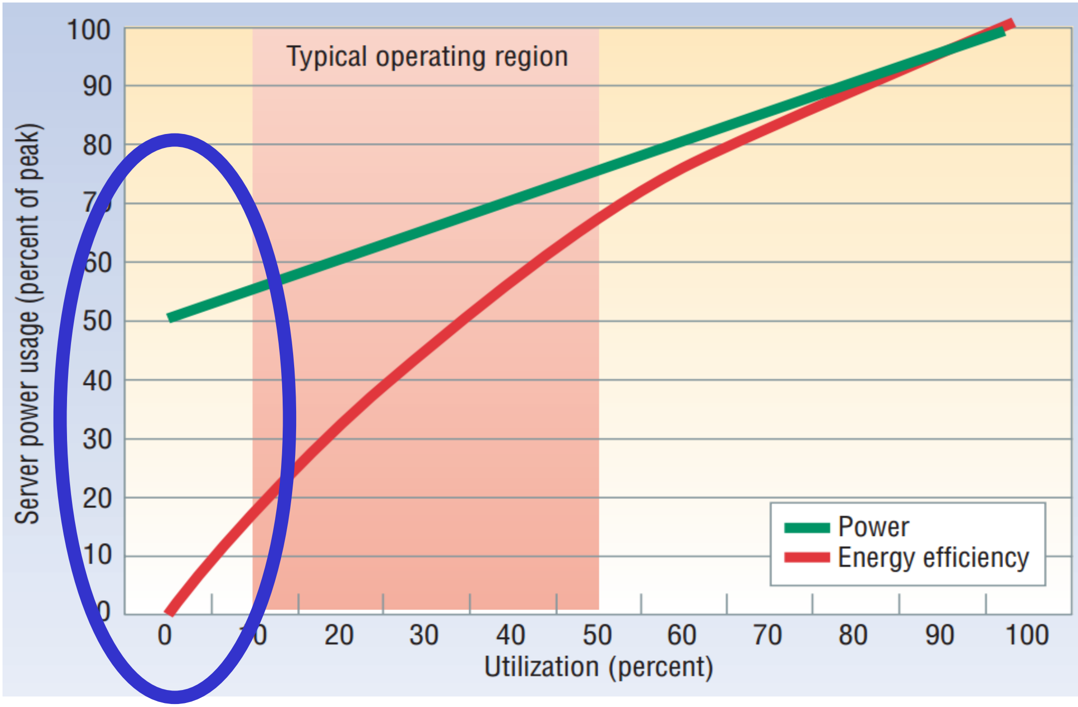
But where to focus on the curve ?

- Energy Profiling and monitoring



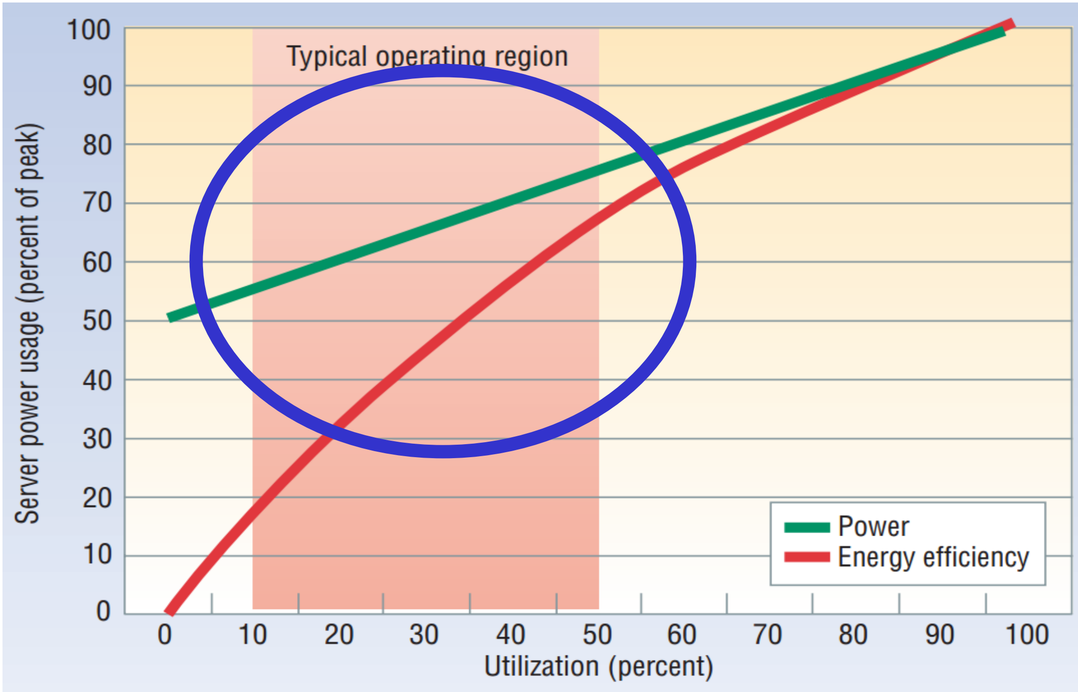
But where to focus on the curve ?

- Energy Profiling and monitoring
- On/Off



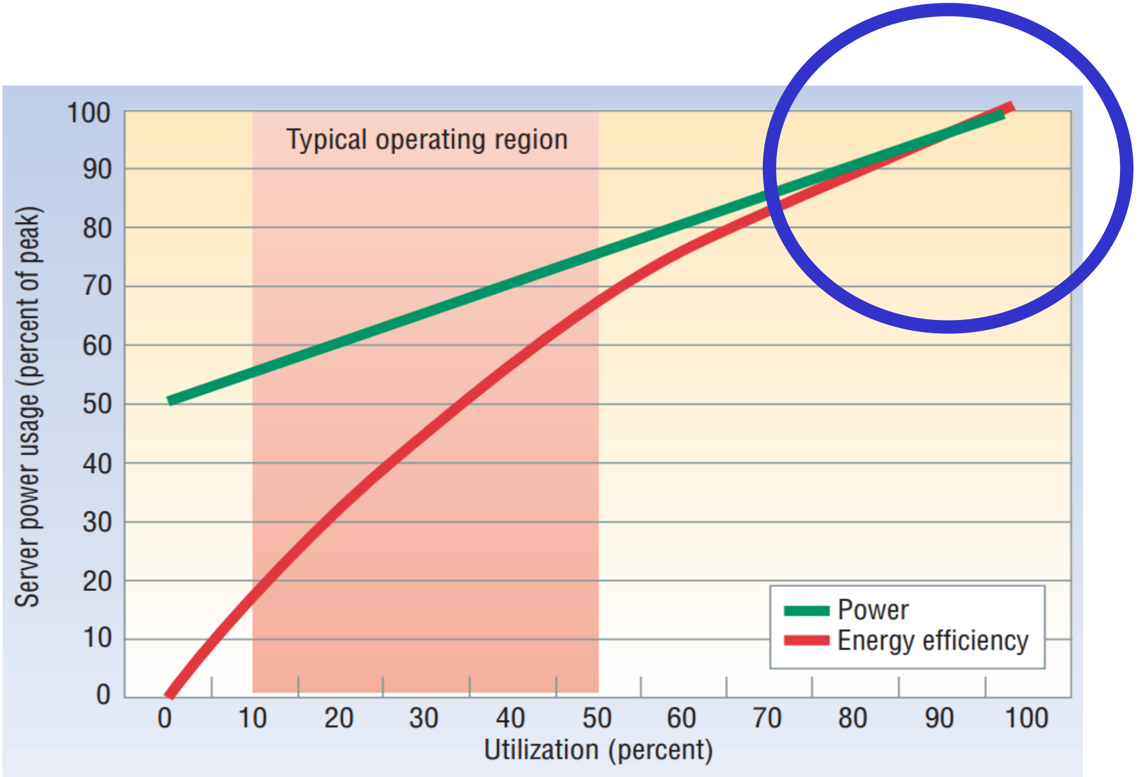
But where to focus on the curve ?

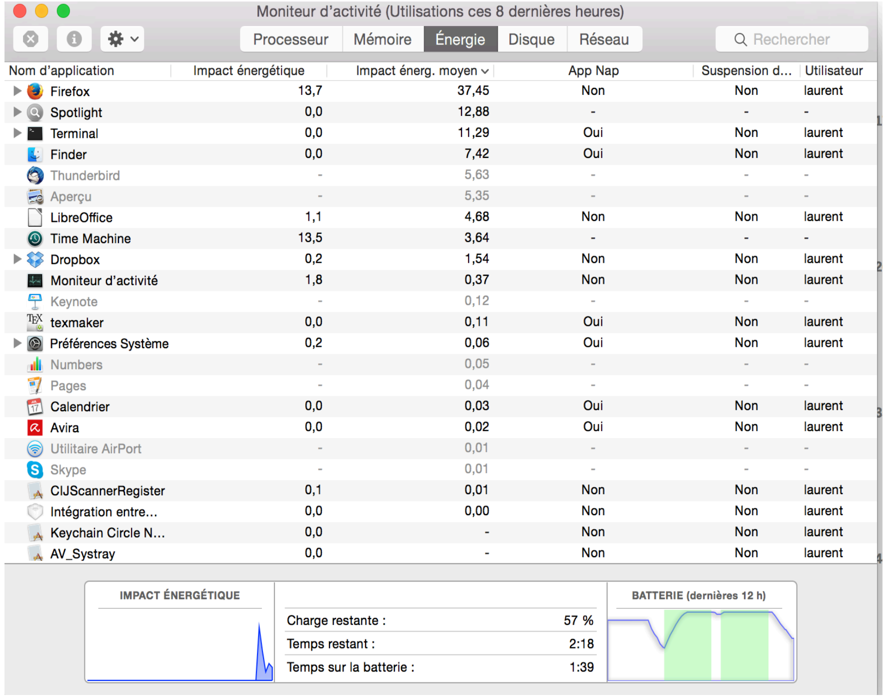
- Energy Profiling and monitoring
- On/Off
- Resizing



But where to focus on the curve ?

- Energy Profiling and monitoring
- On/Off
- Resizing
- Consolidating





Moniteur d'activité (Utilisations ces 8 dernières heures)

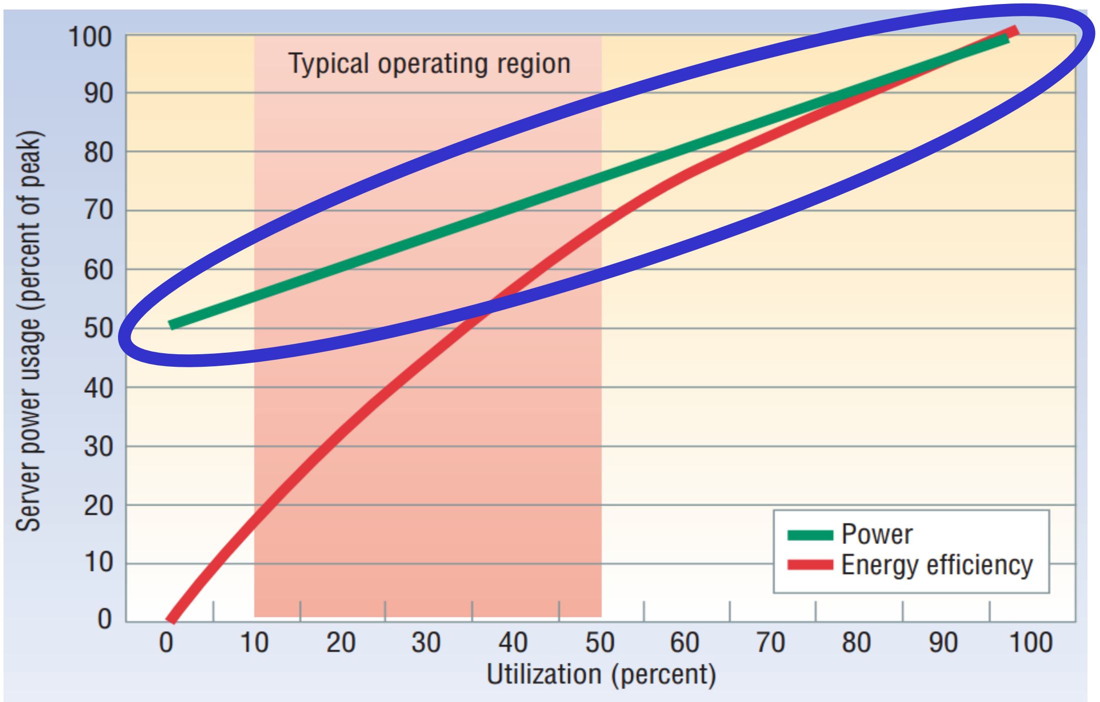
Processeur | Mémoire | Énergie | Disque | Réseau

Nom d'application	Impact énergétique	Impact énerg. moyen	App Nap	Suspension d...	Utilisateur
Firefox	13,7	37,45	Non	Non	laurent
Spotlight	0,0	12,88	-	-	-
Terminal	0,0	11,29	Oui	Non	laurent
Finder	0,0	7,42	Oui	Non	laurent
Thunderbird	-	5,63	-	-	-
Aperçu	-	5,35	-	-	-
LibreOffice	1,1	4,68	Non	Non	laurent
Time Machine	13,5	3,64	-	-	-
Dropbox	0,2	1,54	Non	Non	laurent
Moniteur d'activité	1,8	0,37	Non	Non	laurent
Keynote	-	0,12	-	-	-
texmaker	0,0	0,11	Oui	Non	laurent
Préférences Système	0,2	0,06	Oui	Non	laurent
Numbers	-	0,05	-	-	-
Pages	-	0,04	-	-	-
Calendrier	0,0	0,03	Oui	Non	laurent
Avira	0,0	0,02	Oui	Non	laurent
Utilitaire AirPort	-	0,01	-	-	-
Skype	-	0,01	-	-	-
CIJScannerRegister	0,1	0,01	Non	Non	laurent
Intégration entre...	0,0	0,00	Non	Non	laurent
Keychain Circle N...	0,0	-	Non	Non	laurent
AV_Systray	0,0	-	Non	Non	laurent

IMPACT ÉNERGÉTIQUE

Charge restante :	57 %
Temps restant :	2:18
Temps sur la batterie :	1:39

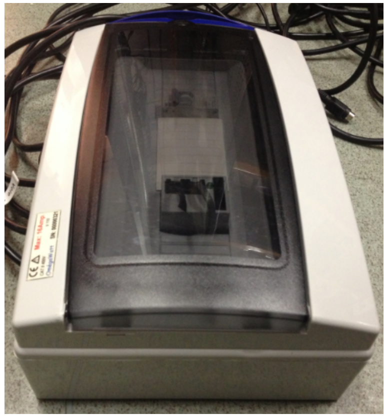
BATTERIE (dernières 12 h)

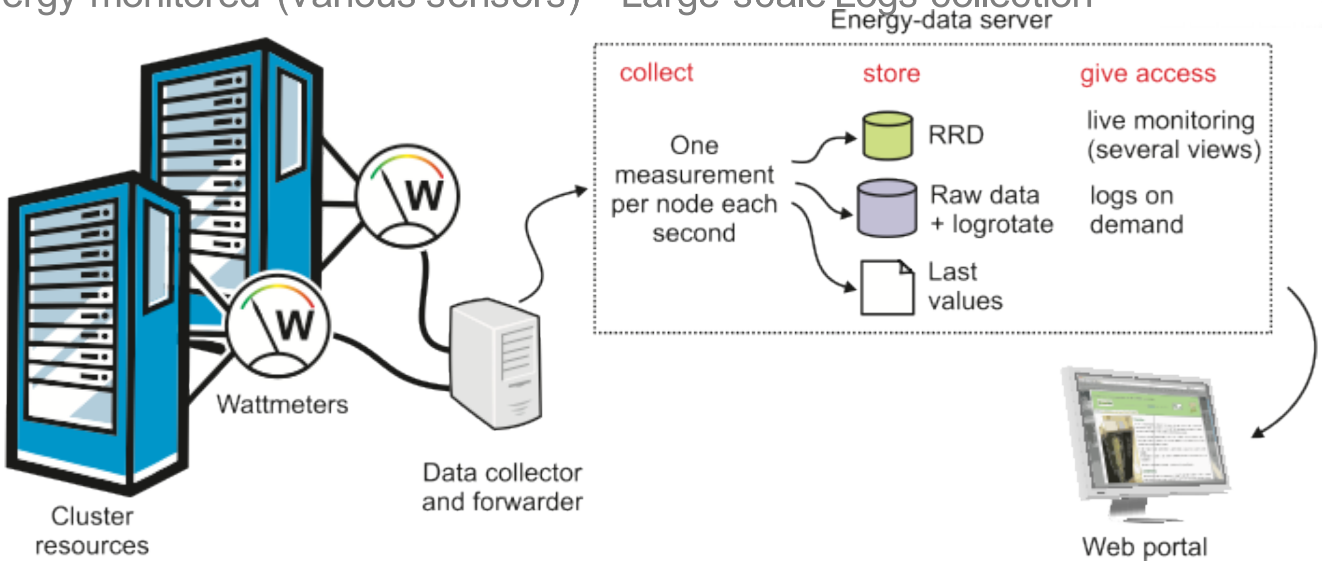


Energy Profiling !!!

A (1slide) short history of Green Grid 5000

- 2007 : first energy efficiency considerations
- 2008 : Launch of research activities on Green : Inria support (ARC Green-Net project : "GREEN-NET : Power aware software frameworks for high performance data transport and computing in large scale distributed systems" with Lyon, Toulouse, Grenoble and Virginia Tech) - 1st challenge : finding a First (real) powermeter SME (Omegawatt)
- 2009 : Deployment of 3 powermeters : Lyon, Toulouse, Grenoble - Design of the Green Grid5000
- 2010 : Increasing scalability : Lyon Grid5000 site - a fully energy monitored site > 150 powermeters in big boxes.. Hemera Launch
- 2015 : All G5K sites are energy monitored (various sensors) – Large scale Logs collection in production (kwapi)





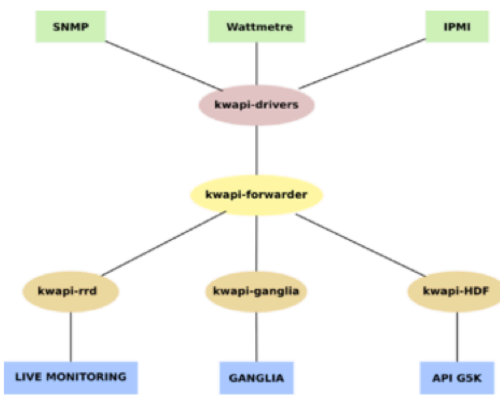
Marcos Dias de Assuncao, Jean-Patrick Gelas, Laurent Lefèvre, and Anne-Cécile Orgerie. "The Green Grid5000: Instrumenting a Grid with Energy Sensors", 5th International Workshop on Distributed Cooperative Laboratories: Instrumenting the Grid (INGRID 2010), Poznan, Poland, May 2010

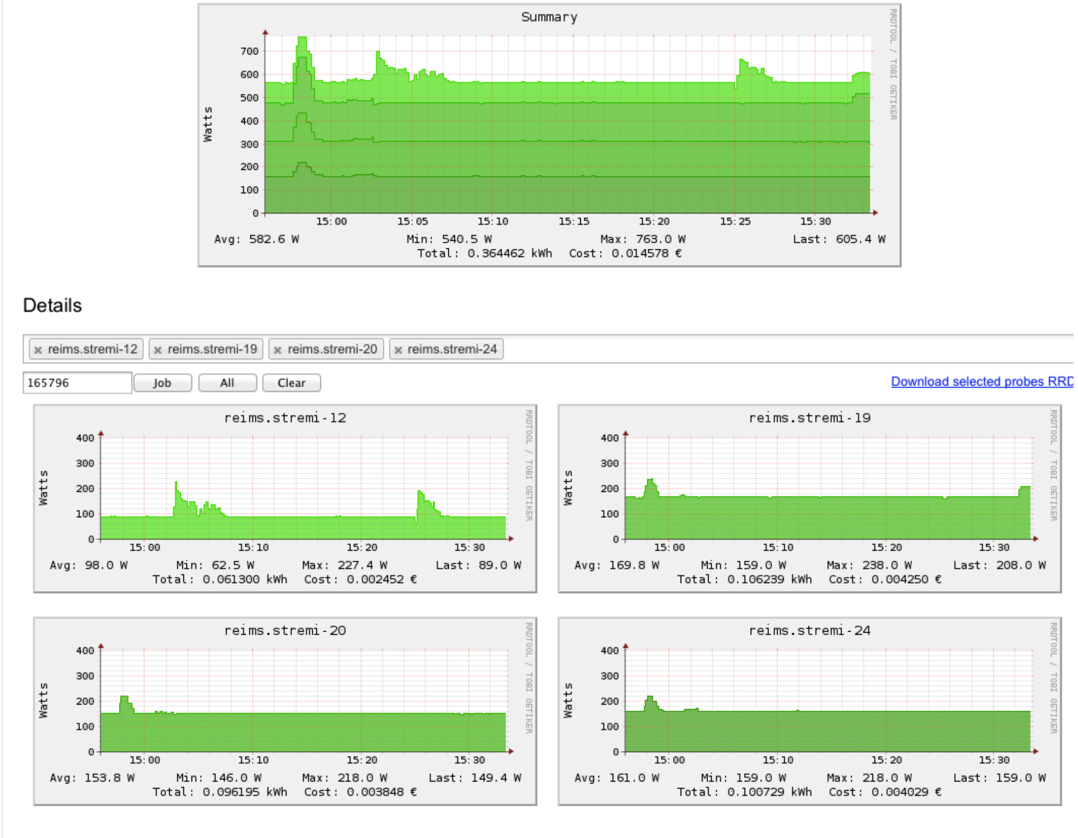
New Kwapi Grid5000 API for Energy monitoring

Unified way to access energy logs for energy profiling

- Energy consumption per resource
- Energy consumption per user jobs

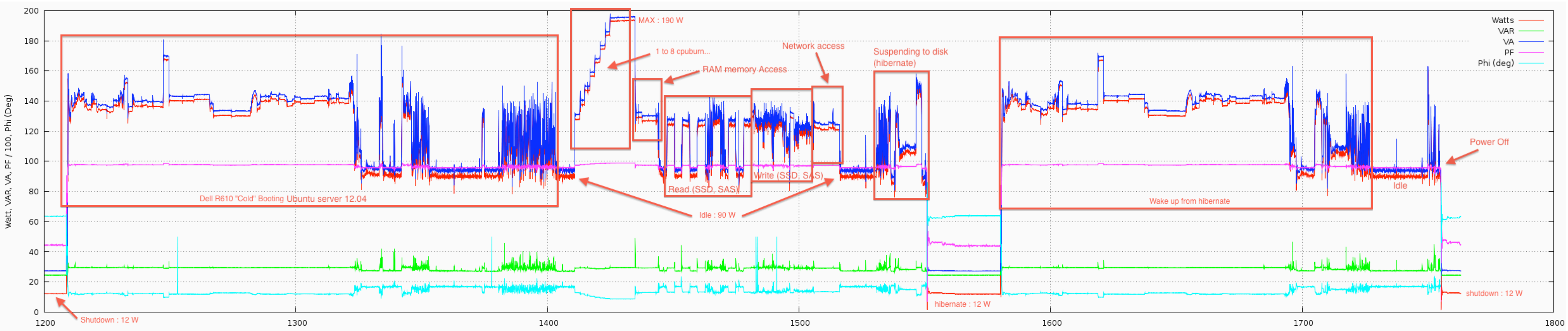
Energy measurement injection in user application and services



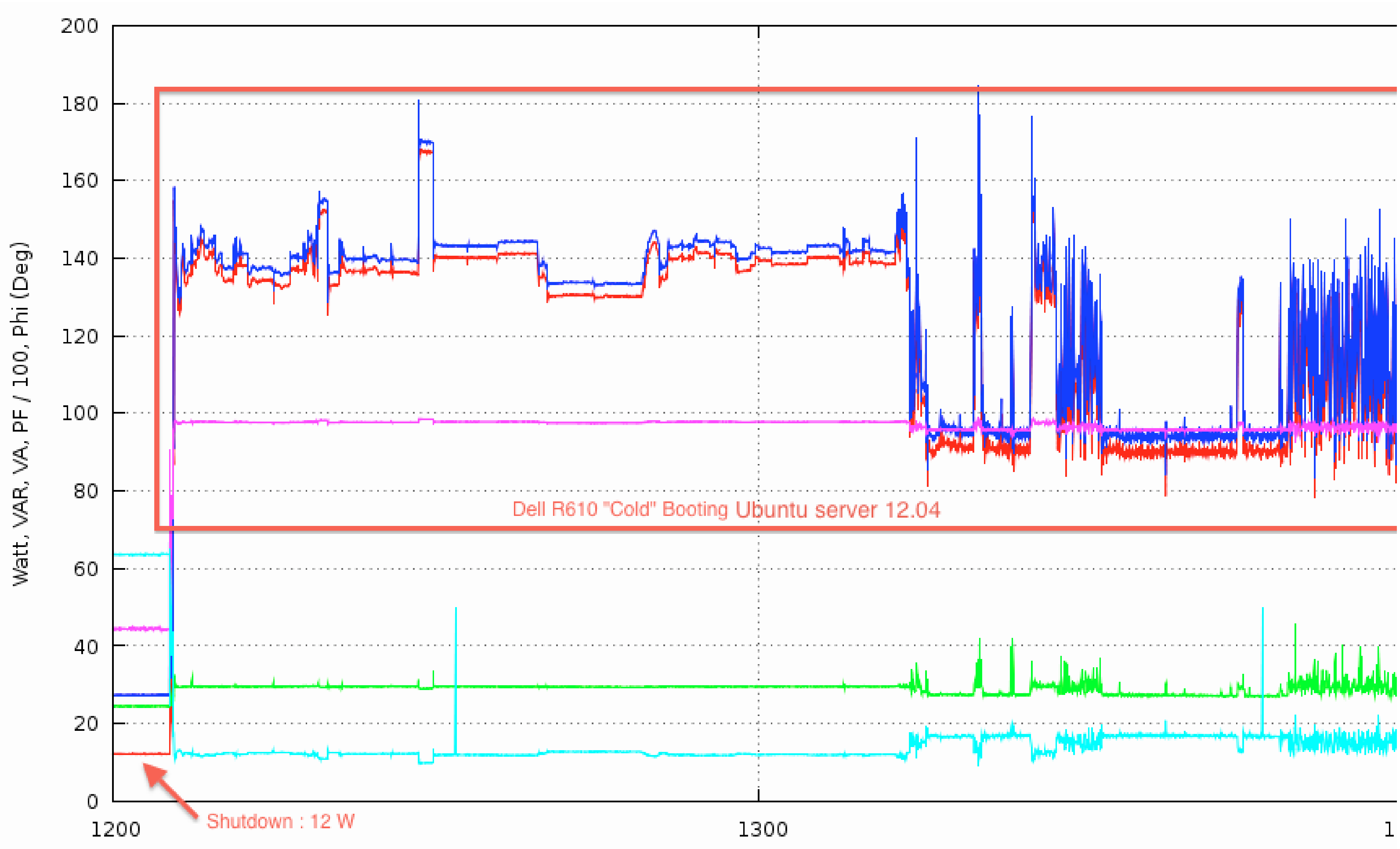


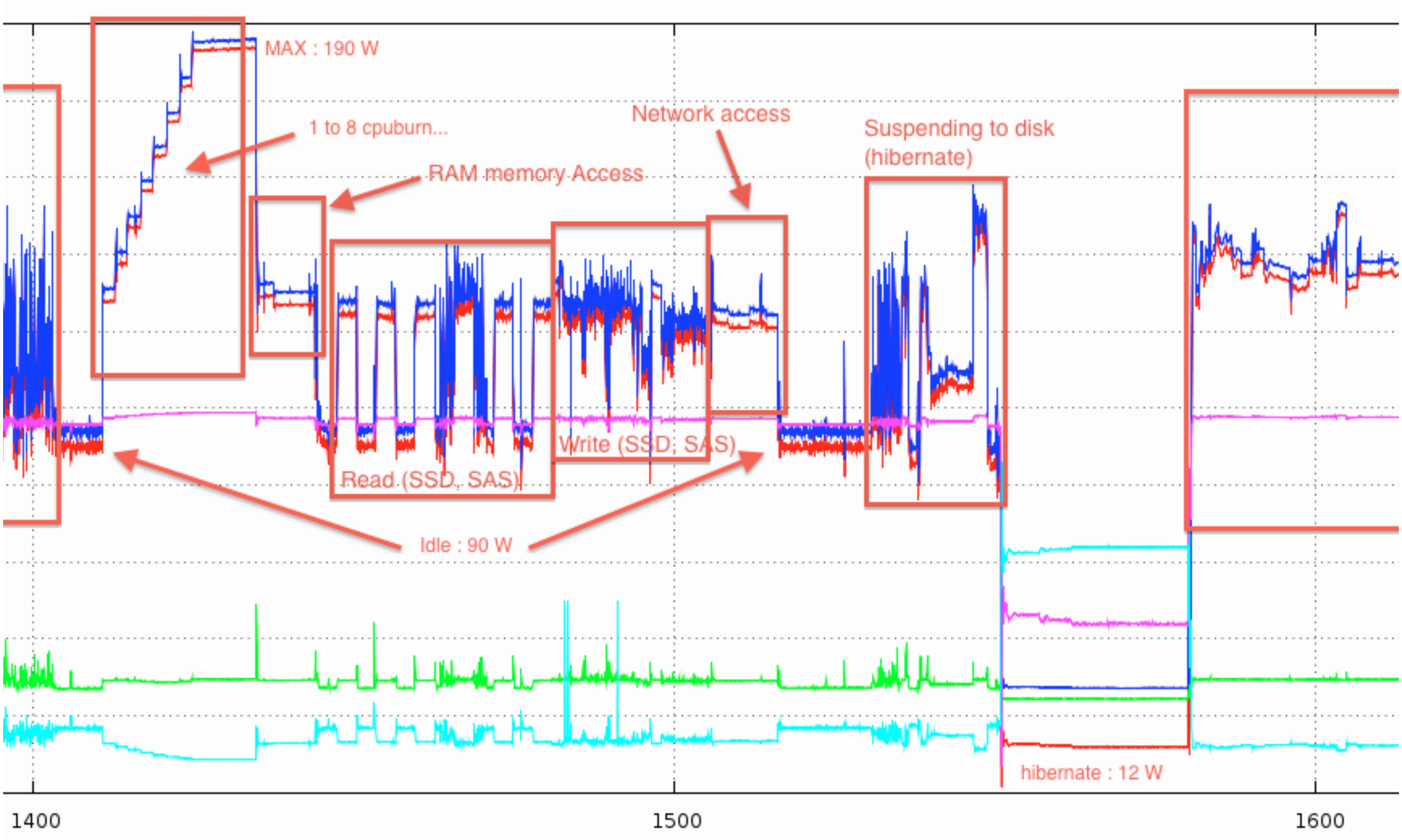
Sites	Monitoring	Via	Hosts	VM	Installation	API	Configuration	Live	LongTerm	Ganglia	Nodes alive
Grenoble	✔	IPMI-SNMP-Wattmetre	ALL ?	✘							103
Lille	✘										54
Luxembourg	⟳	SNMP	ALL ?	✔	✔	⟳	✘	✘	✘	✘	36
Lyon	✔	Wattmetre	ALL	✔	✔	✔	✔	✔	⟳	✘	82
Nancy	✔	SNMP	graphene[105-144]	✔	✔	✔	✔	✔	⟳	✘	218
Nantes	✘										?
Reims	✔	SNMP	ALL	✔	✔	✔	✔	✔	⟳	✘	42
Rennes	✔	SNMP	parapluie	✔	✔	✔	✔	✔	⟳	✘	87
Sophia	✘										105
Toulouse	✘										37

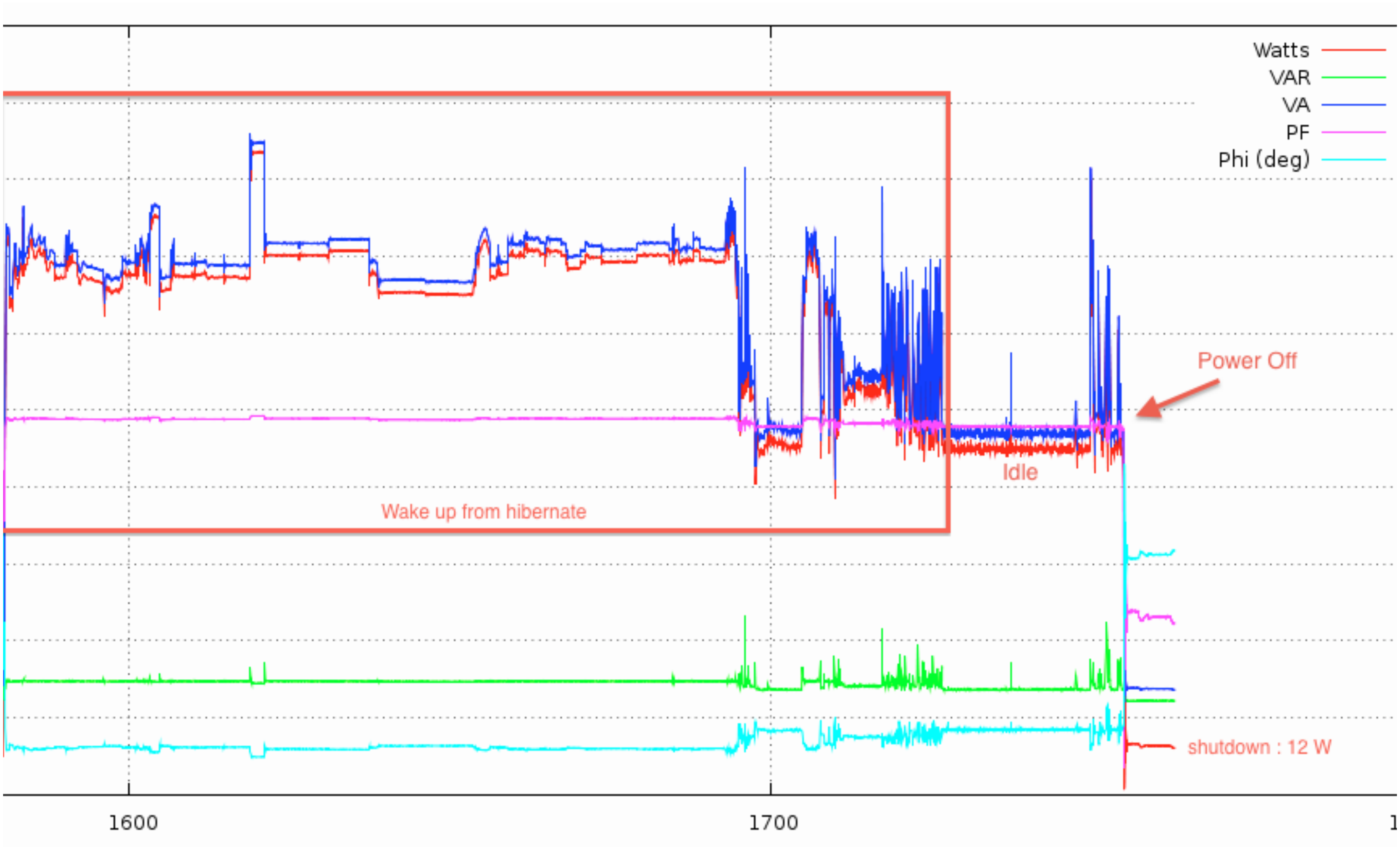
What to observe : Profiling applications on a server



Dell R610 - Zimmer LMG450

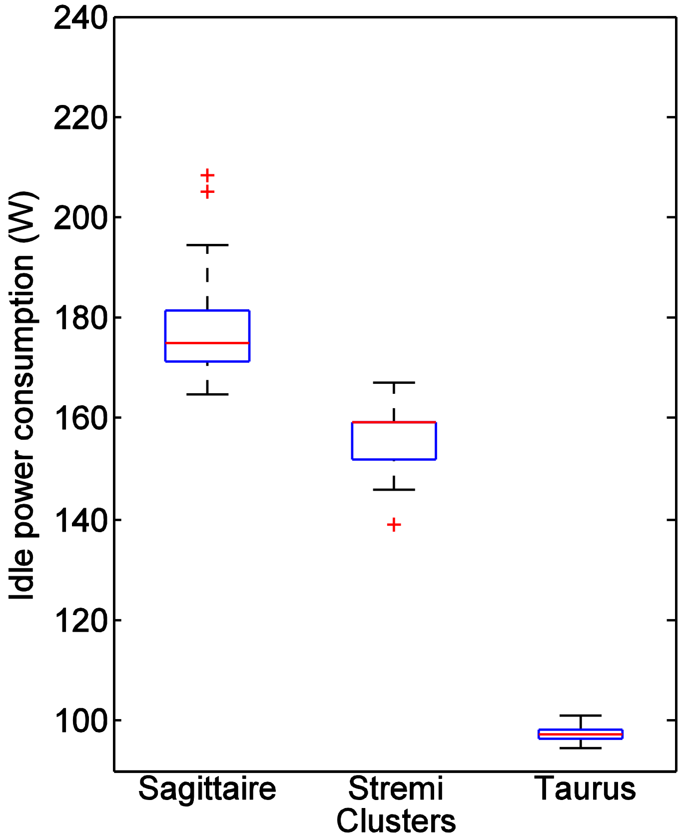

200
180
160
140
120
100
80
60
40
20
0
Watt, VAR, VA, PF / 100, Phi (Deg)
1200
1300
1
Dell R610 "Cold" Booting Ubuntu server 12.04
Shutdown : 12 W

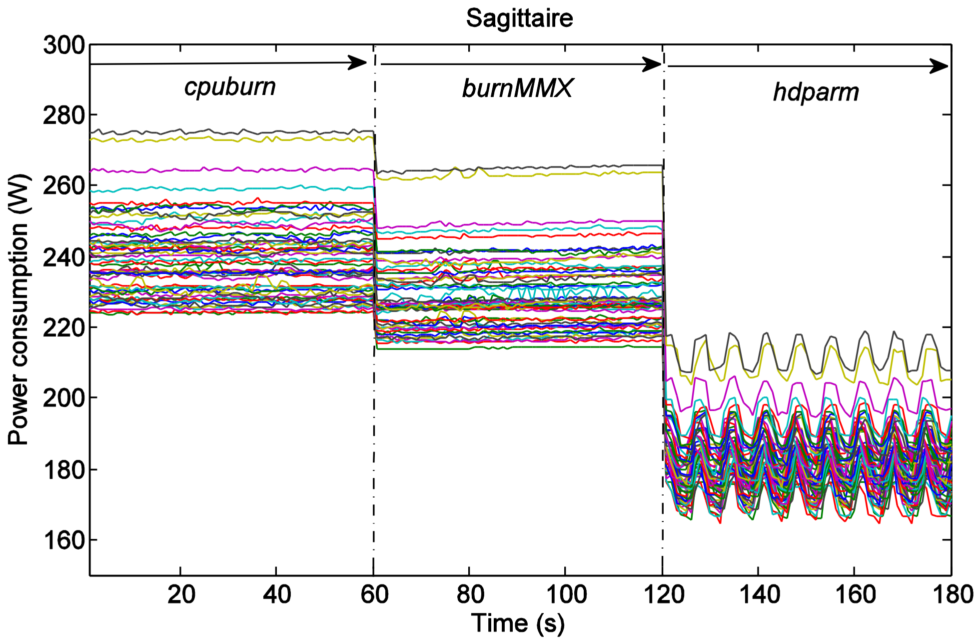

MAX : 190 W
1 to 8 cpuburn...
RAM memory Access
Network access
Suspending to disk
(hibernate)
Write (SSD, SAS)
Read (SSD, SAS)
Idle : 90 W
hibernate : 12 W
1400
1500
1600


Watts
VAR
VA
PF
Phi (deg)
Power Off
Idle
Wake up from hibernate
shutdown : 12 W
1600
1700

Homogeneous electrical consumption ? Does it exist ?

- Depends on technology
- Same flops but not same flops per watt
- Idle / static cost
- CPU : main responsible
- Green scheduler designers must incorporate this issue !

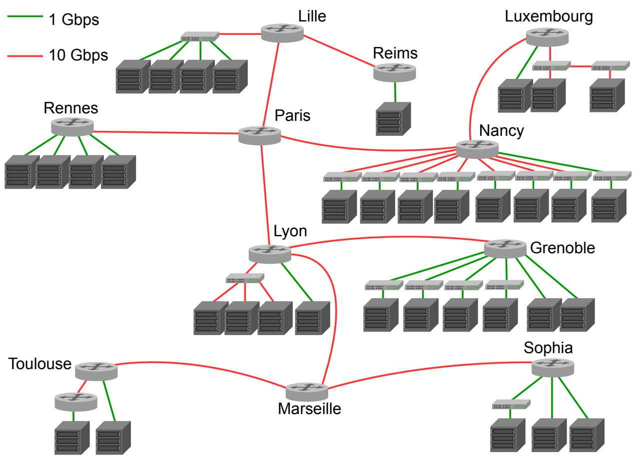


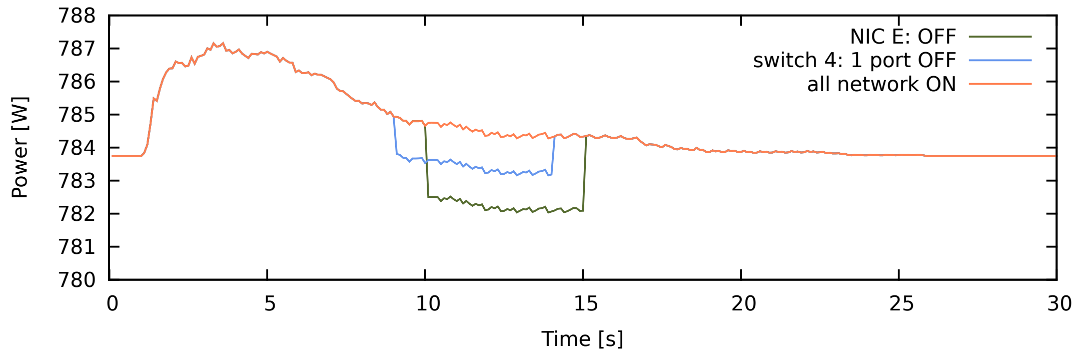


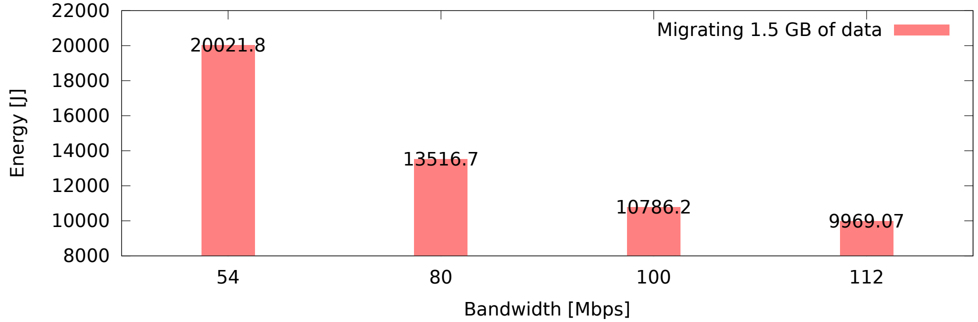
Mohammed el Mehdi Diouri, Olivier Gluck, Laurent Lefevre and Jean-Christophe Mignot. **"Your Cluster is not Power Homogeneous: Take Care when Designing Green Schedulers!"**, *IGCC2013 : International Green Computing Conference*, Arlington, USA, June 27-29,

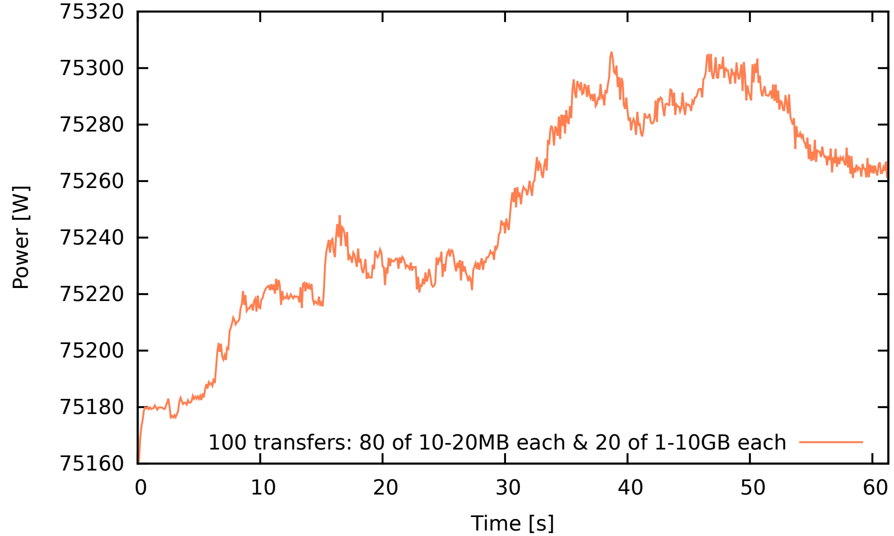
Profiling End2end energy consumption of networks ?

- Need of End2end measurement
- How to evaluate impact of equipment, protocols and services in front of usage/traffic ?
- How to evaluate application of green capabilities ?
- Simulation mandatory ? Yes but with calibrated values ! i.e. Ecofen

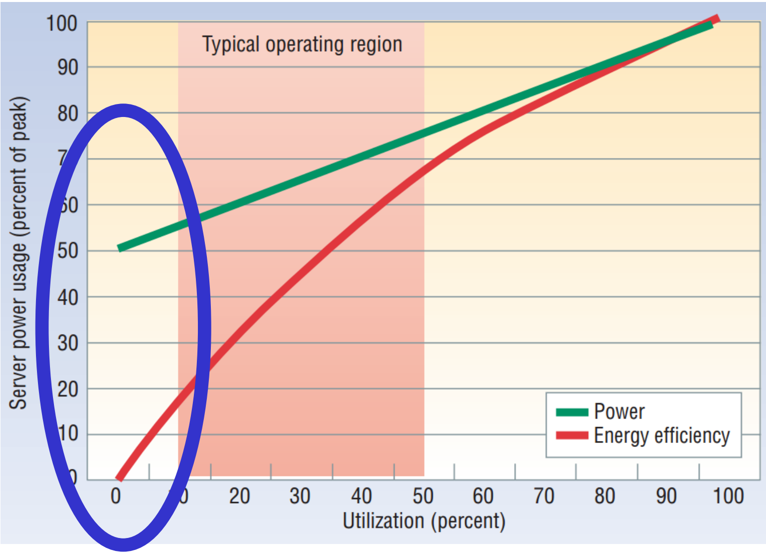








[CloudNet14] : Cornea, Orgerie, Lefevre

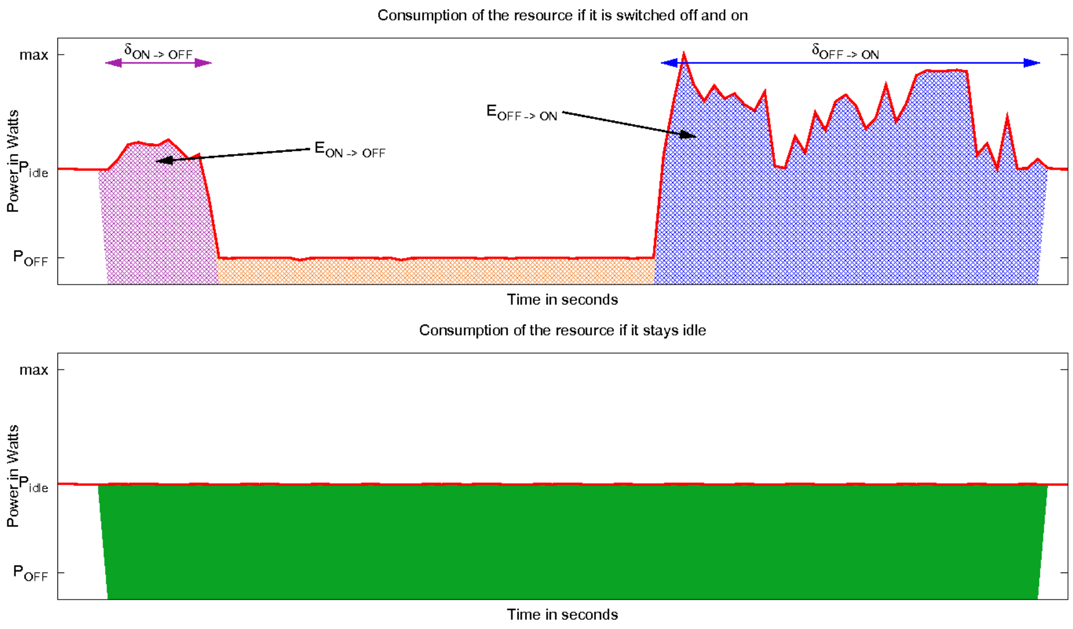


The shutdown approach : Switching ON & OFF for energy proportionality

On/Off Models

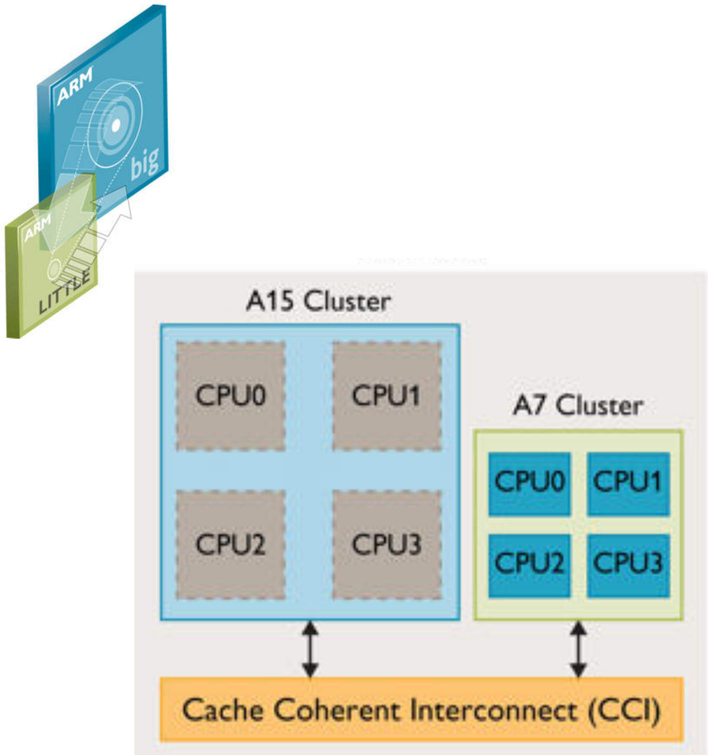
- Switching off useless resources
  - Coarse grain : servers, disks, network equipment
  - Middle grain : network ports, fibers, computing cores, co-processors
  - Small grain : memory banks, electronic components

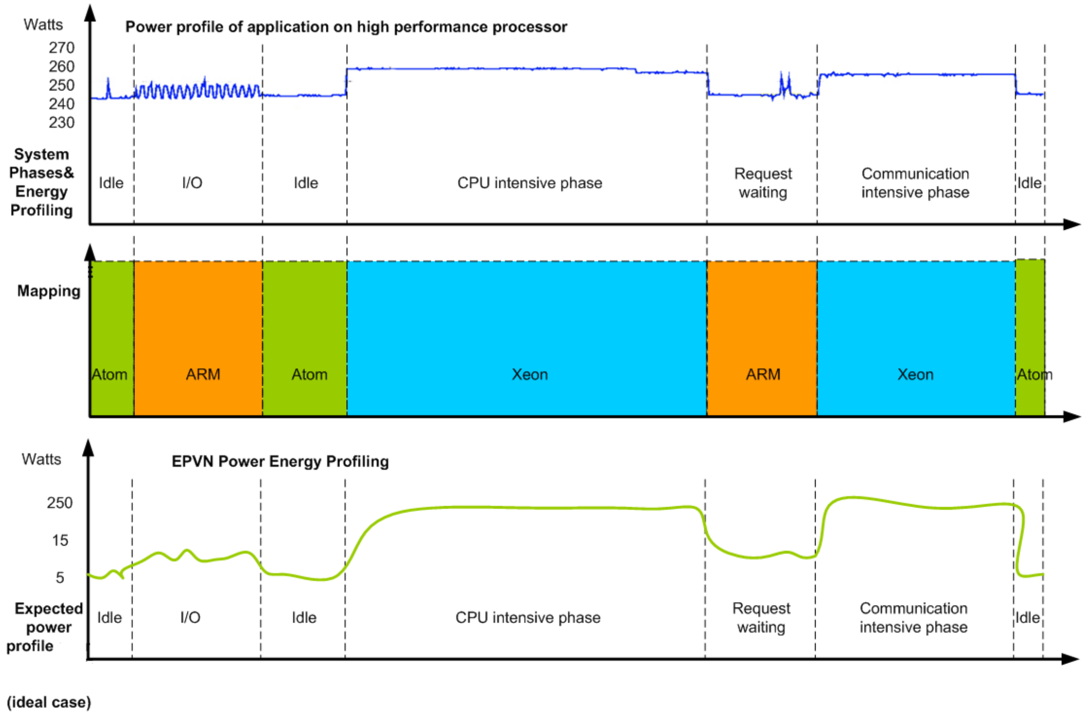
- First approach and widely explored
- Allow the development of *sleep states* and fast *wake up*
- Directly fights the static electrical consumption issue



Energy proportionality on hybrid architectures

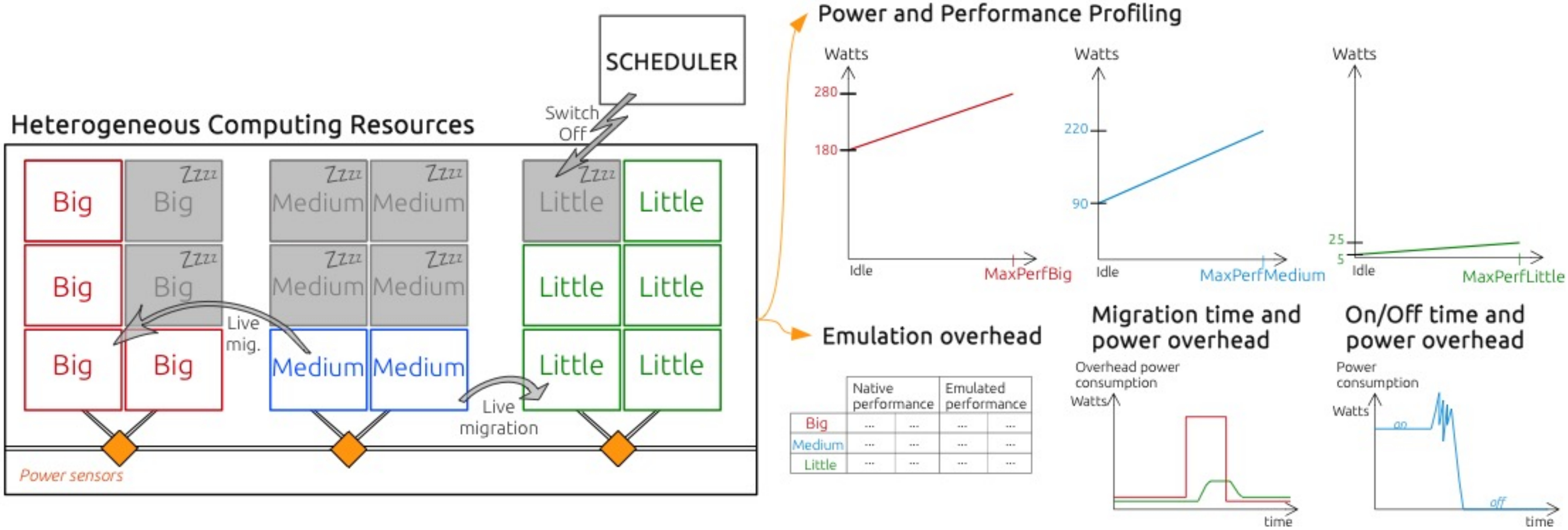
- Hybrid computing architectures : low power processors, co processors, GPUs…
- Supporting a “Big, Medium, Little” approach : the right processor at the right time



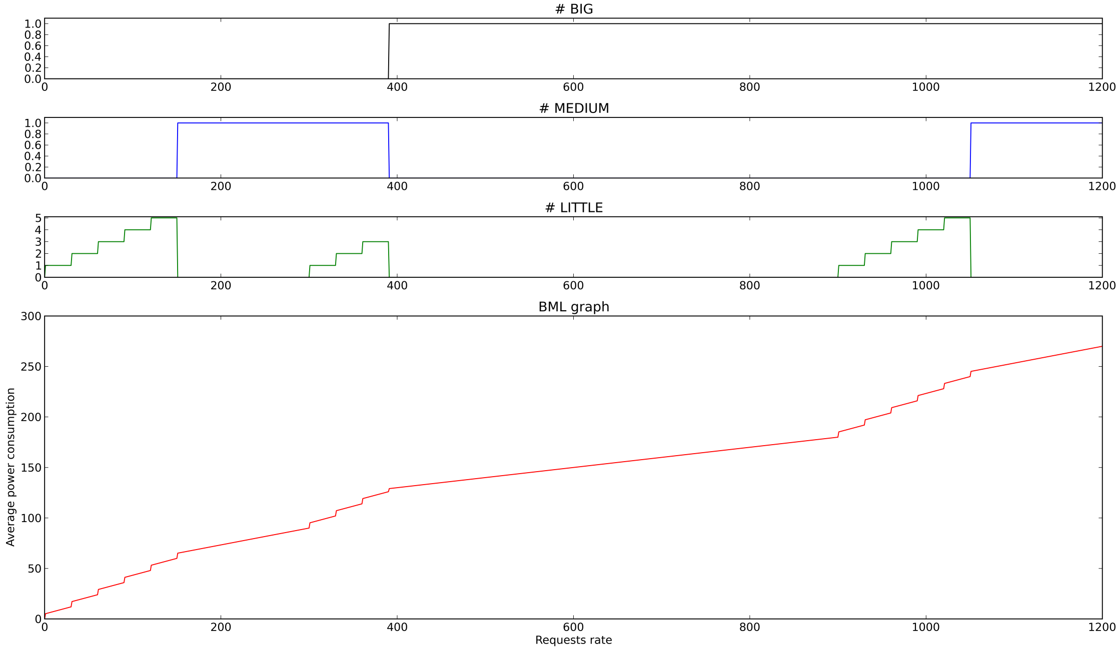


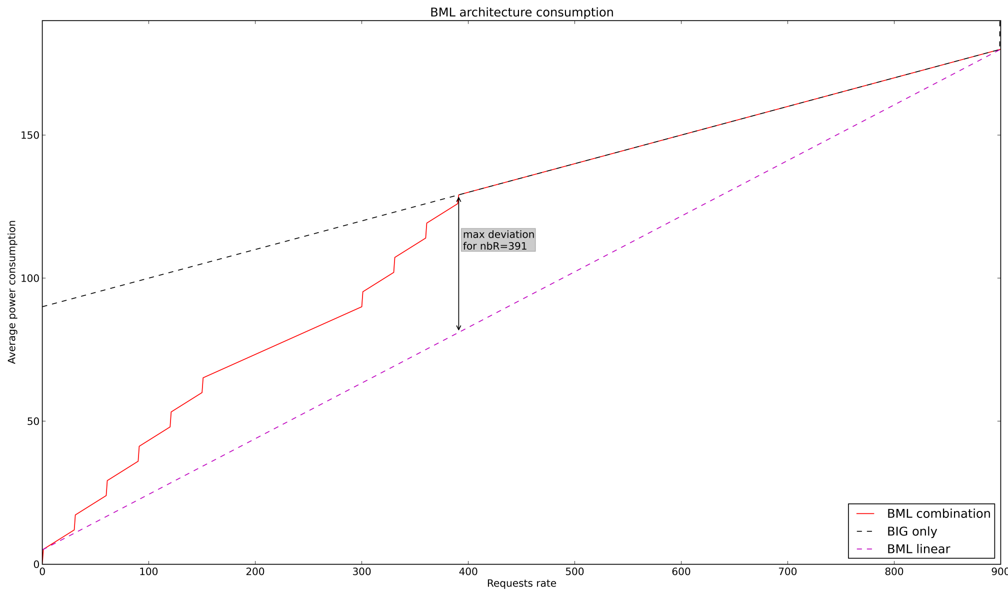
Energy proportionality on hybrid architectures

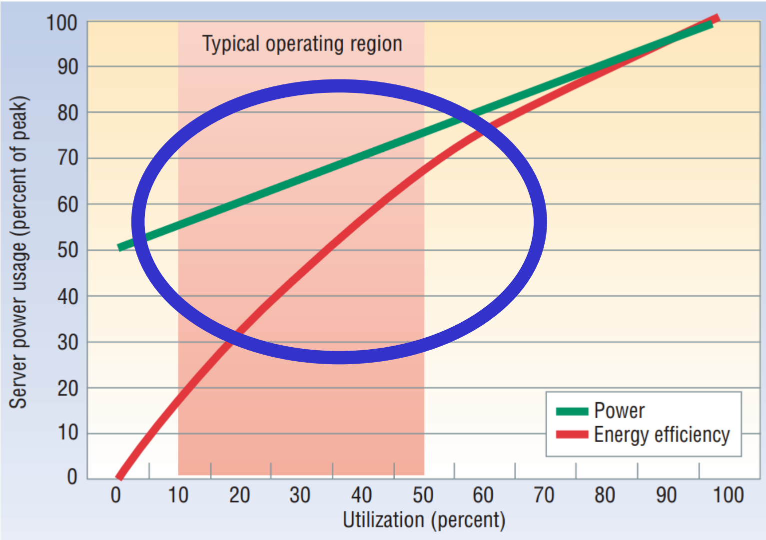
- Need of pertinent resource profiling
- Mapping choices and migration costs must be included
- Switching on/off useful/useless resources



Energy proportionality on hybrid architectures







The slowdown approach : efficient resizing

EE&EP Without knowledge of applications and services ?

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 ?

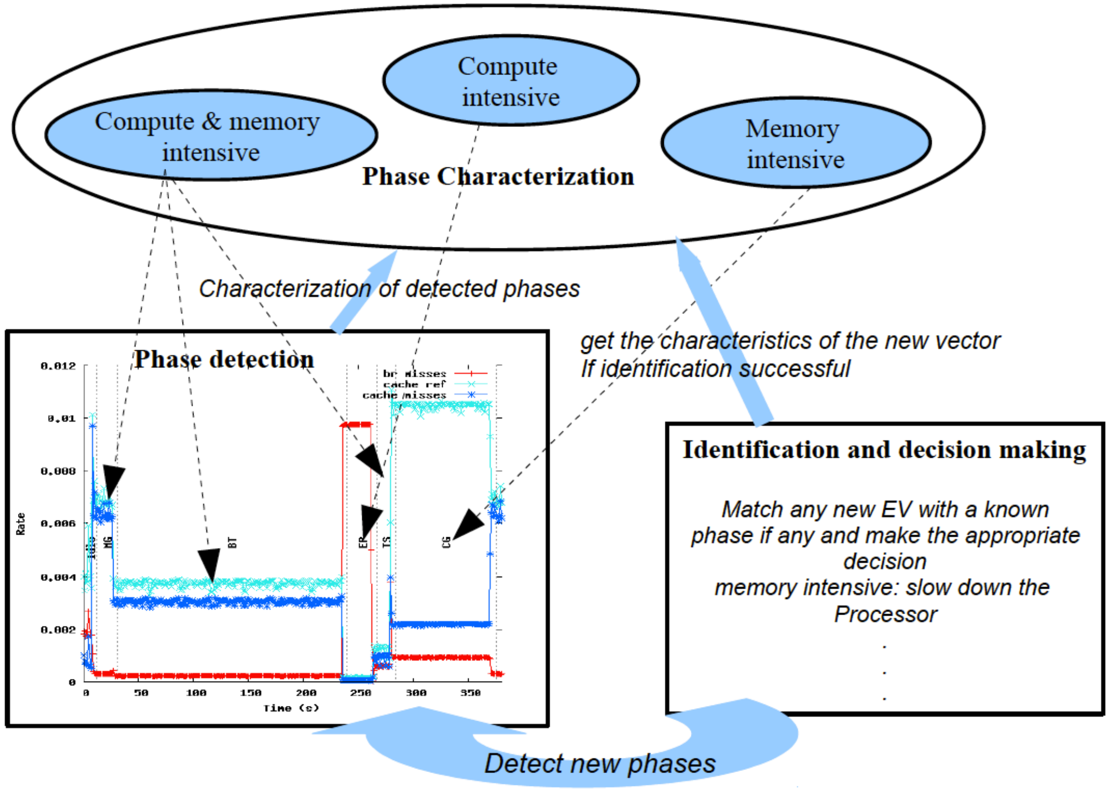
Applications can share the same infrastructure : Optimizations made for saving energy considering some applications are likely to impact the performance of others

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

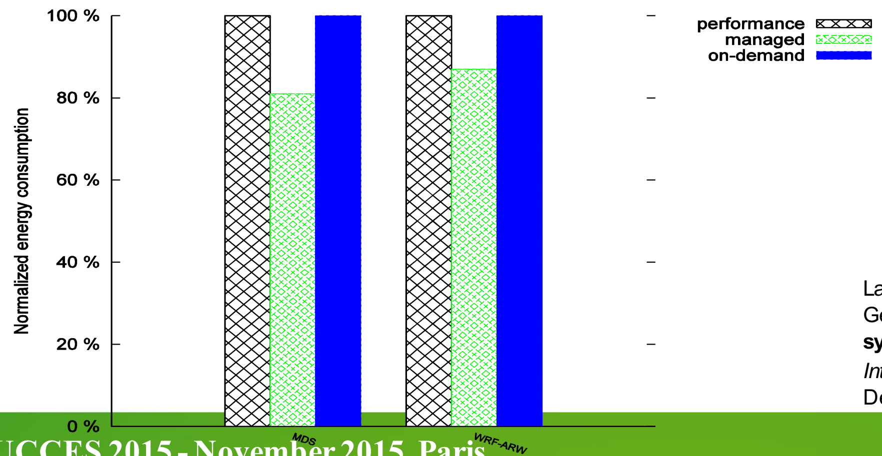
- Detect and characterize system's runtime behaviours/phases
- Optimize each subsystem (storage, memory, interconnect, CPU) accordingly

Without knowledge on applications

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



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)

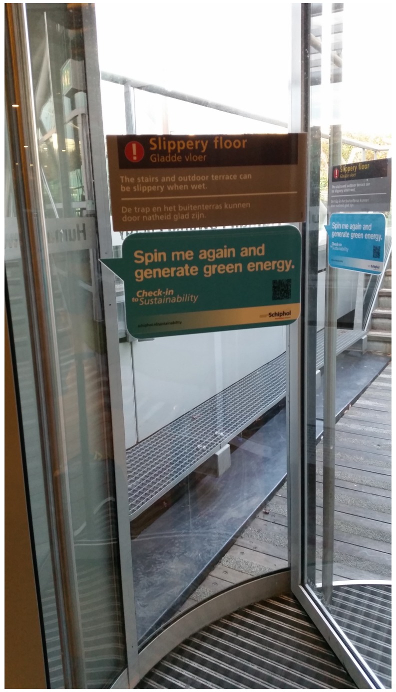


Landry Tsafack, Laurent Lefevre, Jean-Marc Pierson, Patricia Stolf, and Georges Da Costa. **"A runtime framework for energy efficient HPC systems without a priori knowledge of applications"**, *ICPADS 2012 : 18th International Conference on Parallel and Distributed Systems*, Singapore, December 2012

Conclusions

- On the road to EP...
- Energy is driving the way we design applications, services, protocols, equipment…
- Grid 5000 is a needed experimental environment for experimenting energy awareness in distributed systems
- Based on realistic measurements, Energy Efficient applications and services can be built
- Green Grid5000 eco-system has reached a first « production level » which allows easy reporting (for green programming) and injection (for energy aware software design) – New sensors, new metrics ?

- Must take care of the Full lifecycle loop and include (human) users choices



Venez verdoyer avec nous sur Grid5000 !

Thanks to : D. Balouek, F. Cappello, E. Caron, R. Carpa, G. DaCosta, M. Dias de Asuncao, M. Diouri, J.-P. Gelas, O. Gluck, J.C. Mignot, A.-C. Orgerie, G. Tsafack, J.-M. Pierson, L. Pouilloux, F. Rossigneux, P. Stolf, V. Villebonnet

Questions ?

INVENTEURS DU MONDE NUMÉRIQUE

Laurent.lefevre@inria.fr

INVENTEURS DU MONDE NUMÉRIQUE