HPC and Energy Efficiency: an oxymoron?

Jean-Marc Pierson

University of Toulouse

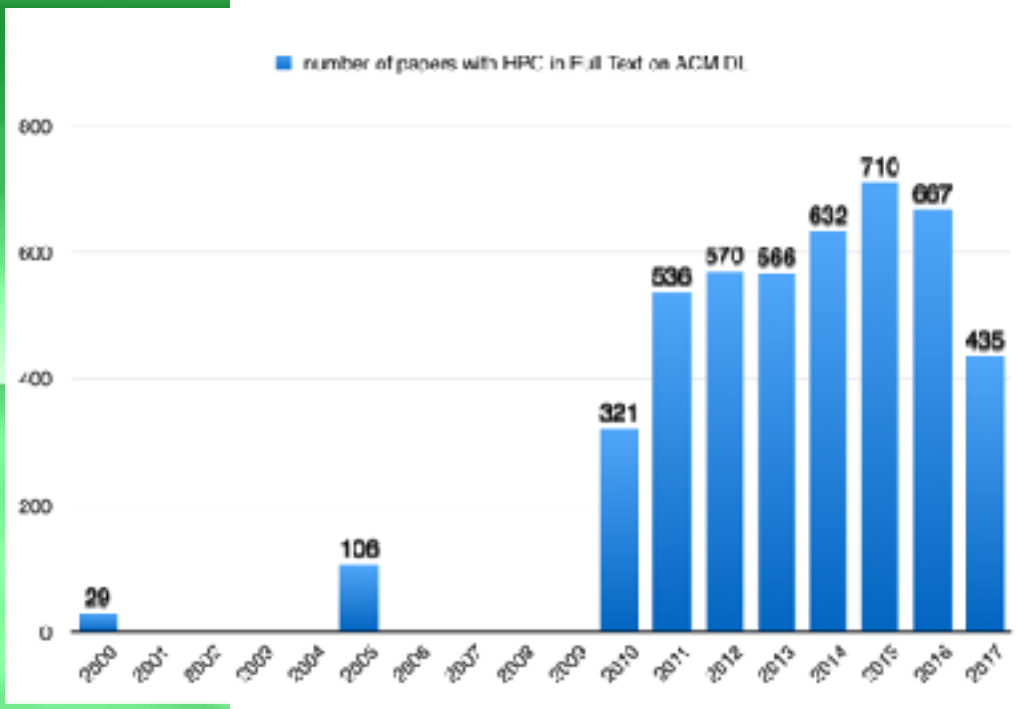
FRANCE

IRIT

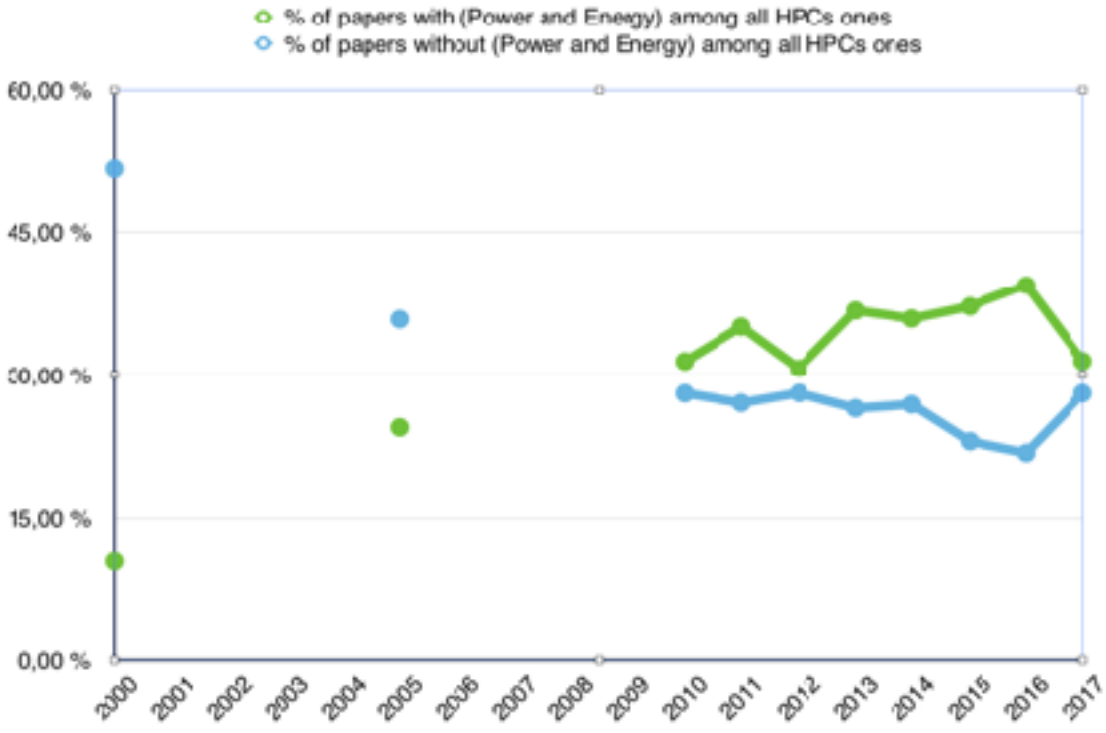
September 2017

pierson@irit.fr

Increasing interests in EE in HPC



in September 2017



pierson@irit.fr

Agenda

- The problem: Joules, Watt, … and performances !
- Hardware view point
- Knowing the system, measuring, modelling
- Actions for energy savings in HPC
- and exascale…
- Conclusion

pierson@irit.fr

What is energy?

Energy = Power * Time

Power in watts
Time in seconds, hours,…

Energy in Wh

Or in Joule: 1 J = 1 Ws

James Prescott Joule

pierson@irit.fr

What is it all about?

UNIVAC I : (UNIVersal AutomaticComputer) machine in the 1950' was consuming 125 kW for 1905 operations per second.

Sunway (NRCPC-China, 1st June 2016 at Top500): 93 Petaflops (48 billions times more) at a cost of 15371 kW (122 times more)

pierson@irit.fr

TOP 500

Rank	System	Cores	(TFlop/s)	(TFlop/s)	(kW)
1	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , **NRCPC** National Supercomputing Center in Wuxi **China**	10,649,600	93,014.6	125,435.9	15,371
2	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P , **NUDT** National Super Computer Center in Guangzhou **China**	3,120,000	33,862.7	54,902.4	17,808
3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , **Cray Inc.** Swiss National Supercomputing Centre (CSCS) **Switzerland**	361,760	19,590.0	25,326.3	2,272
4	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , **Cray Inc.** DOE/SC/Oak Ridge National Laboratory **United States**	560,640	17,590.0	27,112.5	8,209
5	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom , **IBM** DOE/NNSA/LLNL **United States**	1,572,864	17,173.2	20,132.7	7,890
6	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , **Cray Inc.** DOE/SC/LBNL/NERSC **United States**	622,336	14,014.7	27,880.7	3,939
7	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path , **Fujitsu** Joint Center for Advanced High Performance Computing **Japan**	556,104	13,554.6	24,913.5	2,719
8	**K computer**, SPARC64 VIIIfx 2.0GHz, Tofu interconnect , **Fujitsu** RIKEN Advanced Institute for Computational Science (AICS) **Japan**	705,024	10,510.0	11,280.4	12,660
9	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom , **IBM** DOE/SC/Argonne National Laboratory **United States**	786,432	8,586.6	10,066.3	3,945
10	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect , **Cray Inc.** DOE/NNSA/LANL/SNL **United States**	301,056	8,100.9	11,078.9	4,233

Src: www.top500.org

pierson@irit.fr

TOP 500

Rank	System	Cores	(TFlop/s)	(TFlop/s)	(kW)
1	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , **NRCPC** National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
2	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P , **NUDT** National Super Computer Center in Guangzhou China	3,120,000	33,862.7	54,902.4	17,808
3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , **Cray Inc.** Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	2,272
4	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , **Cray Inc.** DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	8,209
5	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom , **IBM** DOE/NNSA/LLNL United States	1,572,864	17,173.2	20,132.7	7,890
6	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , **Cray Inc.** DOE/SC/LBNL/NERSC ed States	622,336	14,014.7	27,880.7	3,939
7	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, l Omni-Path , **Fujitsu** t Center for Advanced High Performance Computing Japan	556,104	13,554.6	24,913.5	2,719
8	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect , **Fujitsu** RIKEN Advanced Institute for Computational Science (AICS) Japan	705,024	10,510.0	11,280.4	12,660
9	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom , **IBM** DOE/SC/Argonne National Laboratory United States	786,432	8,586.6	10,066.3	3,945
10	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect , **Cray Inc.** DOE/NNSA/LANL/SNL United States	301,056	8,100.9	11,078.9	4,233

6051 MFlops / watt
= 6 GFlops / watt

Src: www.top500.org

Energy Efficiency in HPC

June 2017

Rank	TOP500 Rank	System	Cores	Rmax [TFlop/s]	Power [kW]	Efficiency [GFlops/watts]
1	61	**TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 , HPE GSIC Center, Tokyo Institute of Technology **Japan**	36,288	1,998.0	142	14.110
2	465	**kukai** - ZettaScaler-1.6 GPGPU system, Xeon E5-2650Lv4 14C 1.7GHz, Infiniband EDR, NVIDIA Tesla P100 , **ExaScalar** Yahoo Japan Corporation **Japan**	10,080	460.7	33	14.046
3	148	**AIST AI Cloud** - NEC 4U-8GPU Server, Xeon E5-2630Lv4 10C 1.8GHz, Infiniband EDR, NVIDIA Tesla P100 SXM2 , **NEC** National Institute of Advanced Industrial Science and Technology **Japan**	23,400	961.0	76	12.681
4	305	**RAIDEN GPU subsystem** - NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , **Fujitsu** Center for Advanced Intelligence Project, RIKEN **Japan**	11,712	635.1	60	10.603
5	100	**Wilkes-2** - Dell C4130, Xeon E5-2650v4 12C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , **Dell** University of Cambridge **United Kingdom**	21,240	1,193.0	114	10.428
6	3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , **Cray Inc.** Swiss National Supercomputing Centre (CSCS) **Switzerland**	361,760	19,590.0	2,272	10.398
7	69	**Gyoukou** - ZettaScaler-2.0 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , **ExaScalar** Japan Agency for Marine-Earth Science and Technology **Japan**	3,176,000	1,677.1	164	10.226
8	220	**Research Computation Facility for GOSAT-2 (RCF2)** - SGI Rackable C1104-GP1, Xeon E5-2650v4 12C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , **NSSOL/HPE** National Institute for Environmental Studies **Japan**	16,320	770.4	79	9.797
9	31	NVIDIA DGX-1/Penguin Relion 2904GT, Xeon E5-2698v4 20C 2.2GHz/ E5-2650v4, Mellanox Infiniband EDR, NVIDIA Tesla	60,512	3,307.0	350	9.462

Src: www.green500.org

pierson@irit.fr

June 2017

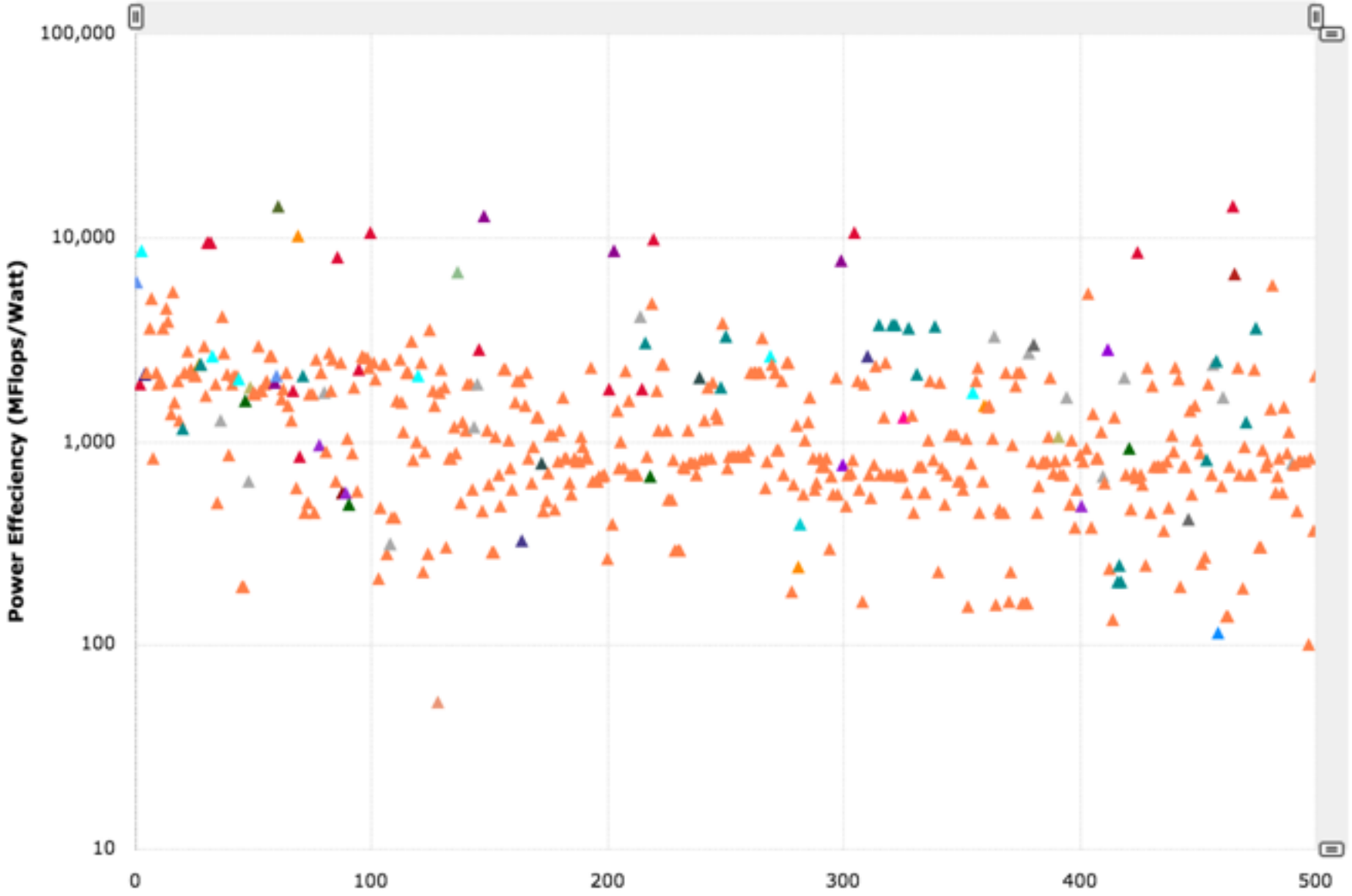
Rank	TOP500 Rank	System	Cores	Rmax [TFlop/s]	Power [kW]	Efficiency [GFlops/watts]
1	61	**TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 , HPE GSIC Center, Tokyo Institute of Technology **Japan**	36,288	1,998.0	142	14.110
2	465	**kukai** - ZettaScaler-1.6 GPGPU system, Xeon E5-2650Lv4 14C 1.7GHz, Infiniband EDR, NVIDIA Tesla P100 , **ExaScalar** Yahoo Japan Corporation **Japan**	10,080	460.7	33	14.046
3	148	**AIST AI Cloud** - NEC 4U-8GPU Server, Xeon E5-2630Lv4 10C 1.8GHz, Infiniband EDR, NVIDIA Tesla P100 SXM2 , **NEC** National Institute of Advanced Industrial Science and Technology **Japan**	23,400	961.0	76	12.681
4	305	**RAIDEN GPU subsystem** - NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , **Fujitsu** Center for Advanced Intelligence Project, RIKEN **Japan**	11,712	635.1	60	10.603
5	100	**Wilkes-2** - Dell C4130, Xeon E5-2650v4 12C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , **Dell** University of Cambridge **United Kingdom**	21,240	1,193.0	114	10.428
6	3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , **Cray Inc.** Swiss National Supercomputing Centre (CSCS) **Switzerland**	361,760	19,590.0	2,272	10.398
7	69	**Gyoukou** - ZettaScaler-2.0 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , **ExaScalar** Japan Agency for Marine-Earth Science and Technology **Japan**	3,176,000	1,677.1	164	10.226
8	220	**Research Computation Facility for GOSAT-2 (RCF2)** - SGI Rackable C1104-GP1, Xeon E5-2650v4 12C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , **NSSOL/HPE** National Institute for Environmental Studies **Japan**	16,320	770.4	79	9.797
9	31	NVIDIA DGX-1/Penguin Relion 2904GT, Xeon E5-2698v4 20C 2.2GHz/ E5-2650v4, Mellanox Infiniband EDR, NVIDIA Tesla	60,512	3,307.0	350	9.462

All hybrid architectures ! (almost all Tesla P100)

Src: www.green500.org

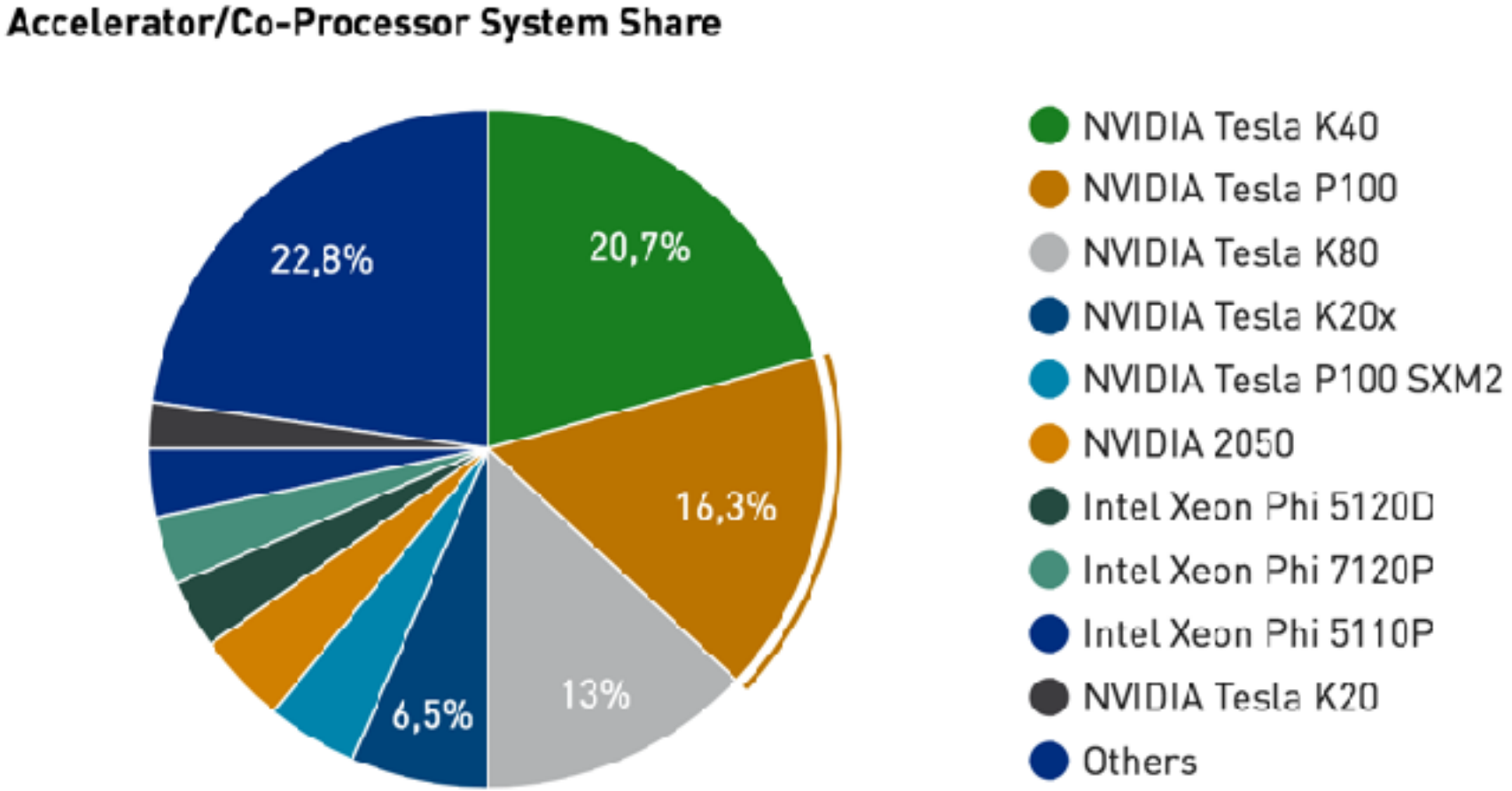
pierson@irit.fr

Power Efficiency of the Top500 (Jun 2017)



pierson@irit.fr

Accelerators / Coprocessors :
(only) 90 / 500 have accelerators in top500 list

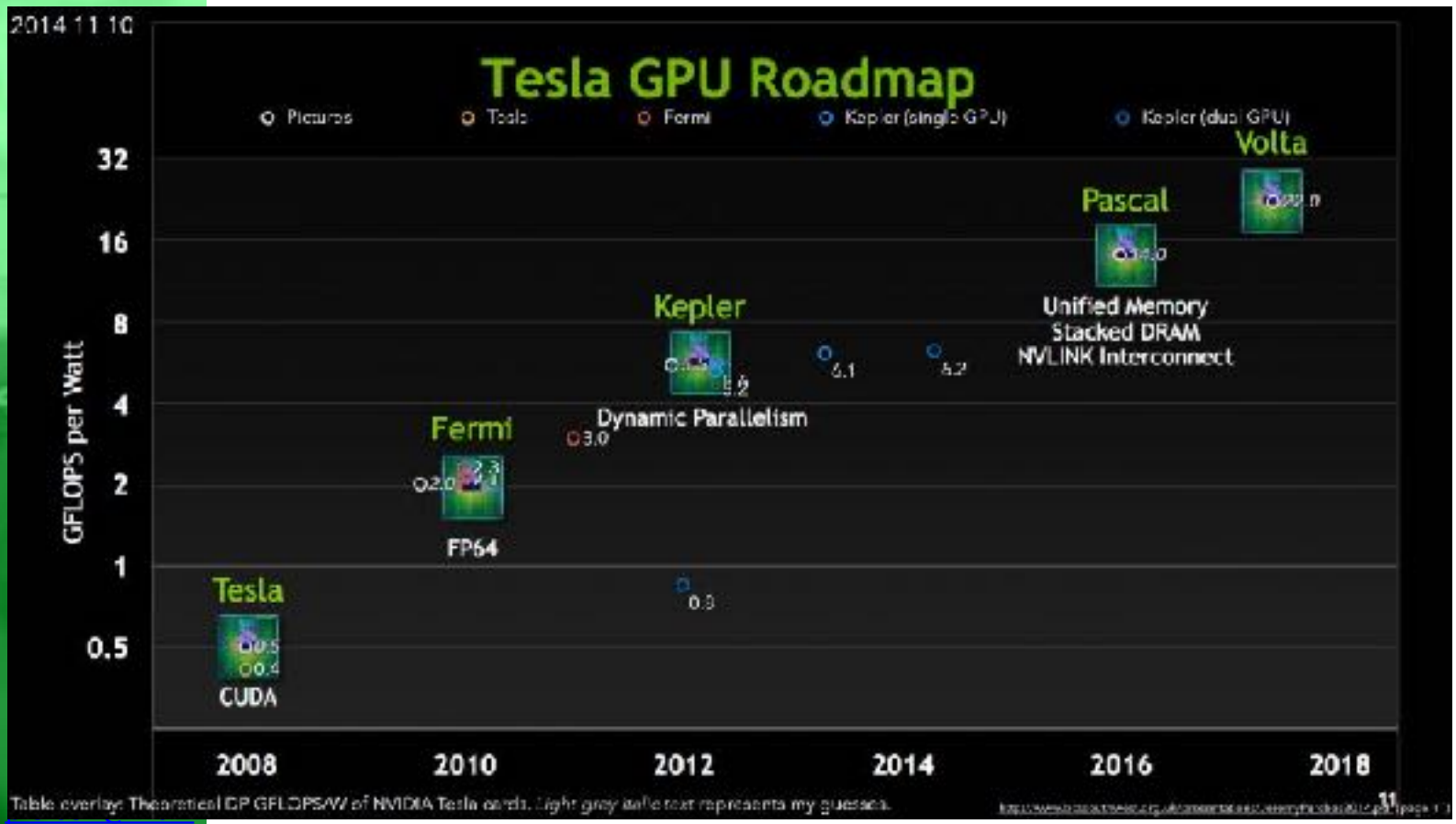


pierson@irit.fr

Evolution of energy efficiency for NVIDIA GPUs

» More integration (FinFET manufacturing process at 16nm)

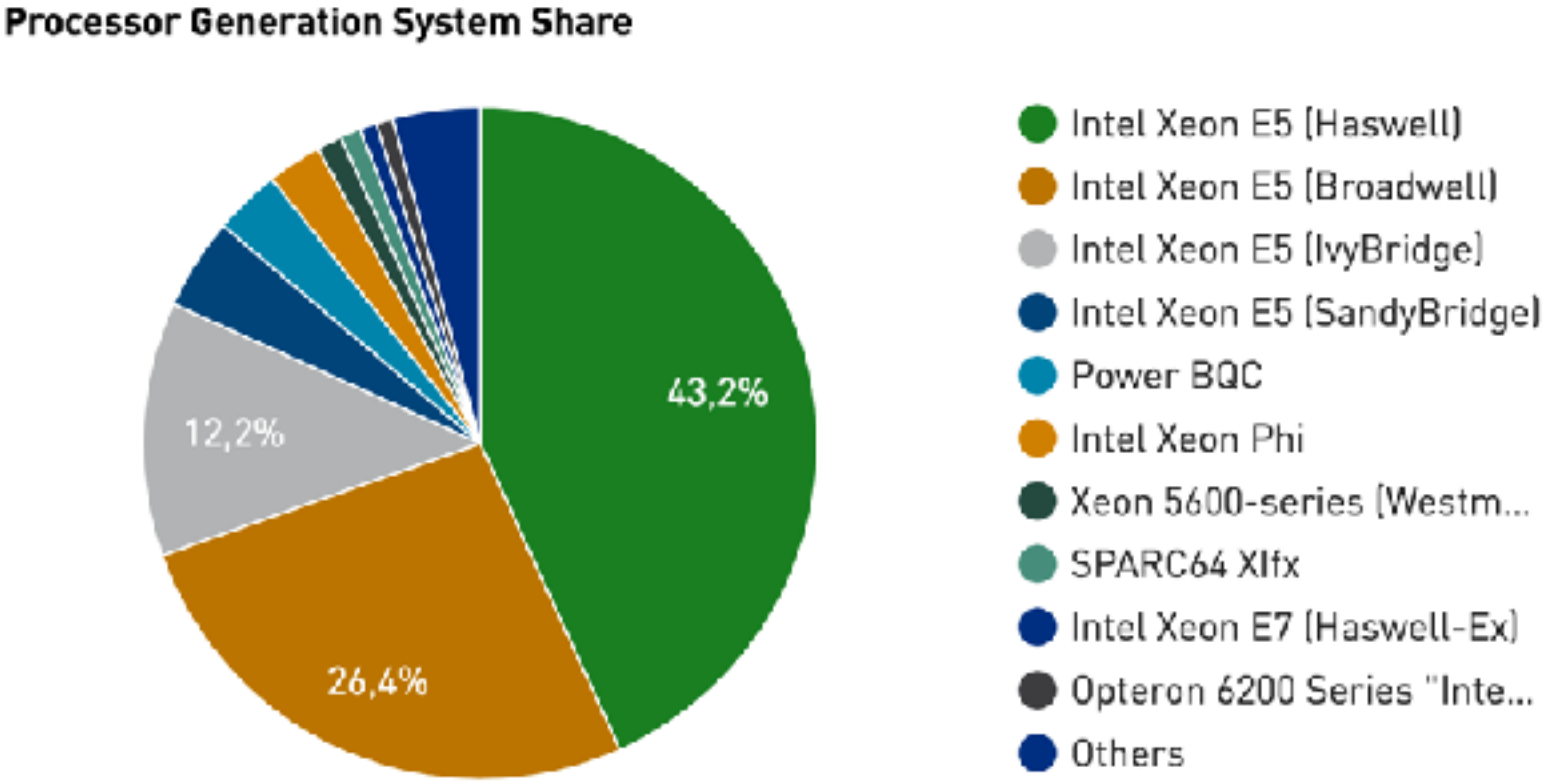
» More cores : in Tesla P100 (Pascal), each Streaming Multiprocessor has 64 CUDA cores. A GPU can host more than 3000 CUDA cores.



[nvidia.com]

Impact of architectures

Architectures of HPC ?



>90% Intel, then IBM, SPARC, IBM, AMD, …

top500.org

pierson@irit.fr

ARM-based HPC

Designed originally for mobile phones

but improving…

For instance, the EU Mont-Blanc project at BSC.

(EUROPEAN APPROACH TOWARDS ENERGY EFFICIENT HIGH PERFORMANCE)

2160 CPUs and 1080 GPUs.

Originally based on Exynos 5 SoC Dual

ARM Cortex-A15 (ARMv7, 32 bits)

Integrated ARM Mali-T604 GPU.

Performance :

34.7 teraflops with 24 kilowatts : 1445 Mflops / W

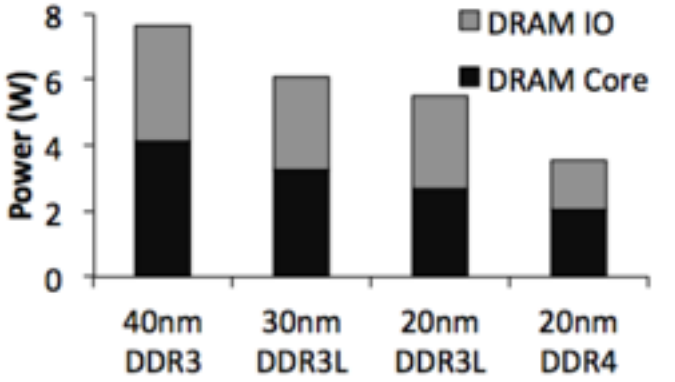
Soon (2017-): Dibona- BULL Sequana, ARMv8, 64bits with Double Precision Floating Point (ThunderX2 ARM). 48 Nodes. 3000 cores

pierson@irit.fr

Memory energy

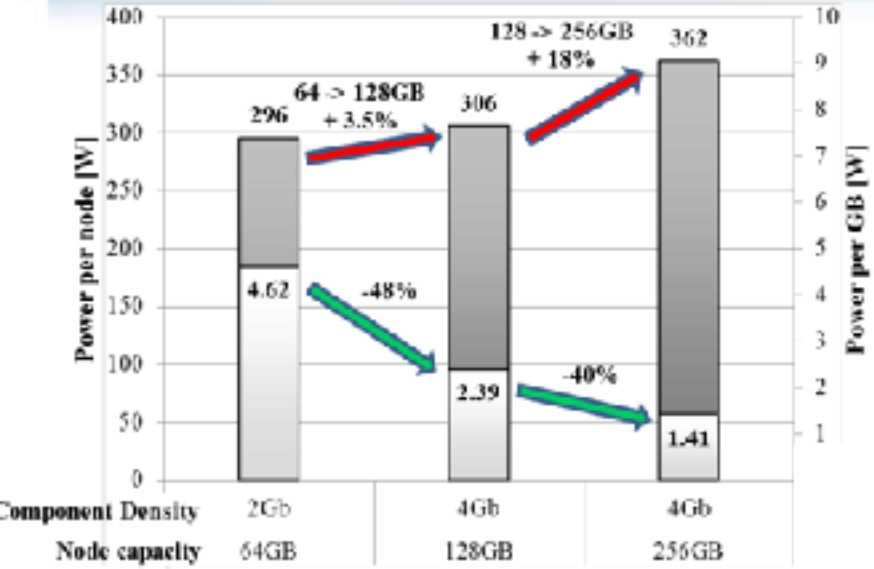
Need for FAST and Power efficient uniform access to memory DRAM and GFX memory together: Memory bandwidth should not become the bottleneck

Decrease of power thanks to low voltage and higher density



Samsung

pierson@irit.fr



Memory capacity and component density:

- Higher memory density per node consumes more power: +21.5% between 64GB and 256GB
- the power consumption per GB of capacity decreases due to power efficiency of 4Gb component vs. 2Gb.

Within 100 KW power limit:

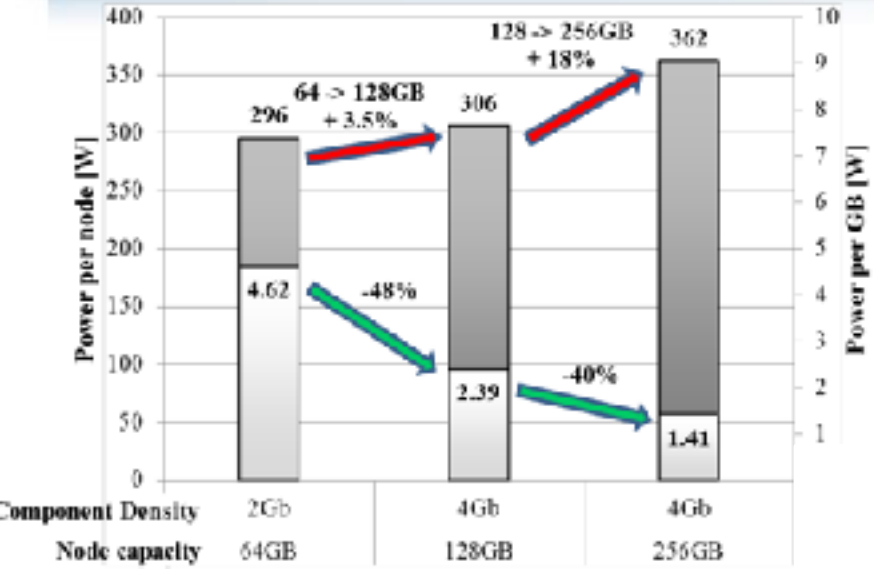
- **276** nodes with 256GB@1600 Mbps, vs.
- **352** nodes with 64GB

i.e. **18% less nodes** with high density modules will provide **3.1x more total** memory in the cluster

The most energy efficient configuration:

- **64GB** capacity per node
- **30nm DRAM process technology**
- running at **low voltage (1.35V)** and **1600Mbps**

pierson@irit.fr



Memory capacity and component density:

- Higher memory density per node consumes more power: +21.5% between 64GB and 256GB
- the power consumption per GB of capacity decreases due to power efficiency of 4Gb component vs. 2Gb.

Within 100 KW power limit:

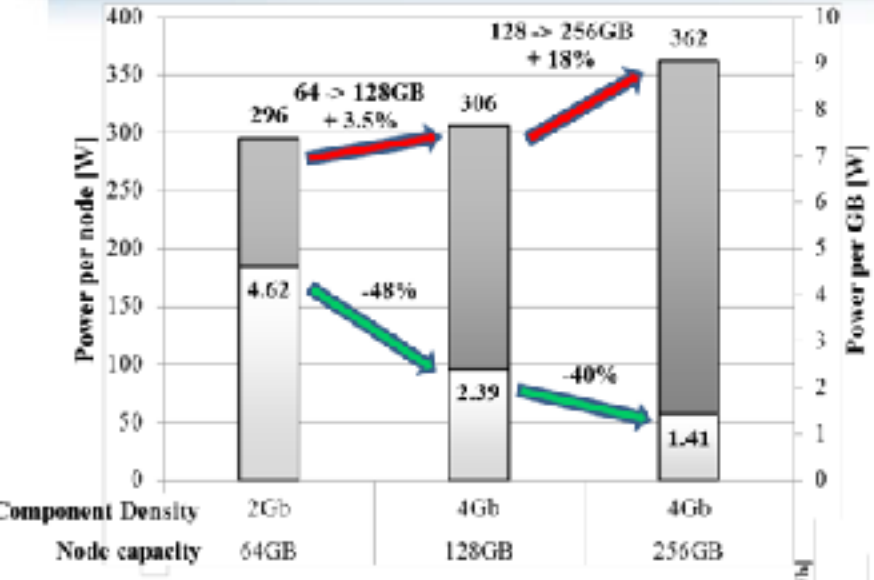
- **276** nodes with 256GB@1600 Mbps, vs.
- **352** nodes with 64GB

i.e. **18% less nodes** with high density modules will provide **3.1x more total** memory in the cluster

The most energy efficient configuration:

- **64GB** capacity per node
- **30nm DRAM process technology**
- running at **low voltage (1.35V)** and **1600Mbps**

pierson@irit.fr



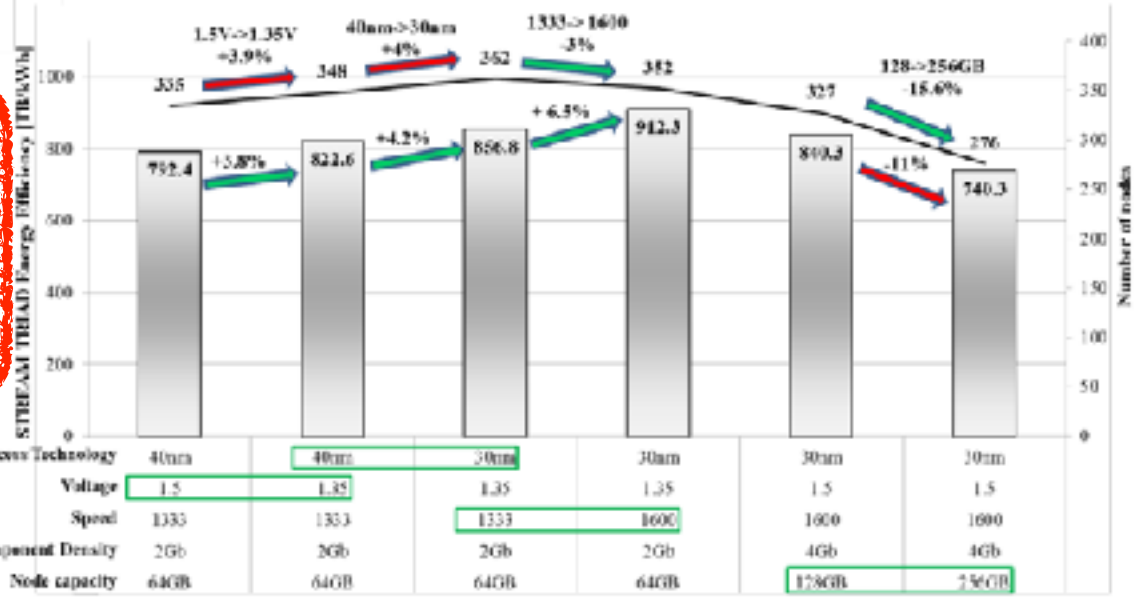
Memory capacity and component density:

- Higher memory density per node consumes more power: +21.5% between 64GB and 256GB
- the power consumption per GB of capacity decreases due to power efficiency of 4Gb component vs. 2Gb.

Within 100 KW power limit:

- **276** nodes with 256GB@1600 Mbps, vs.
- **352** nodes with 64GB

i.e. **18% less nodes** with high density modules will provide **3.1x more total** memory in the cluster

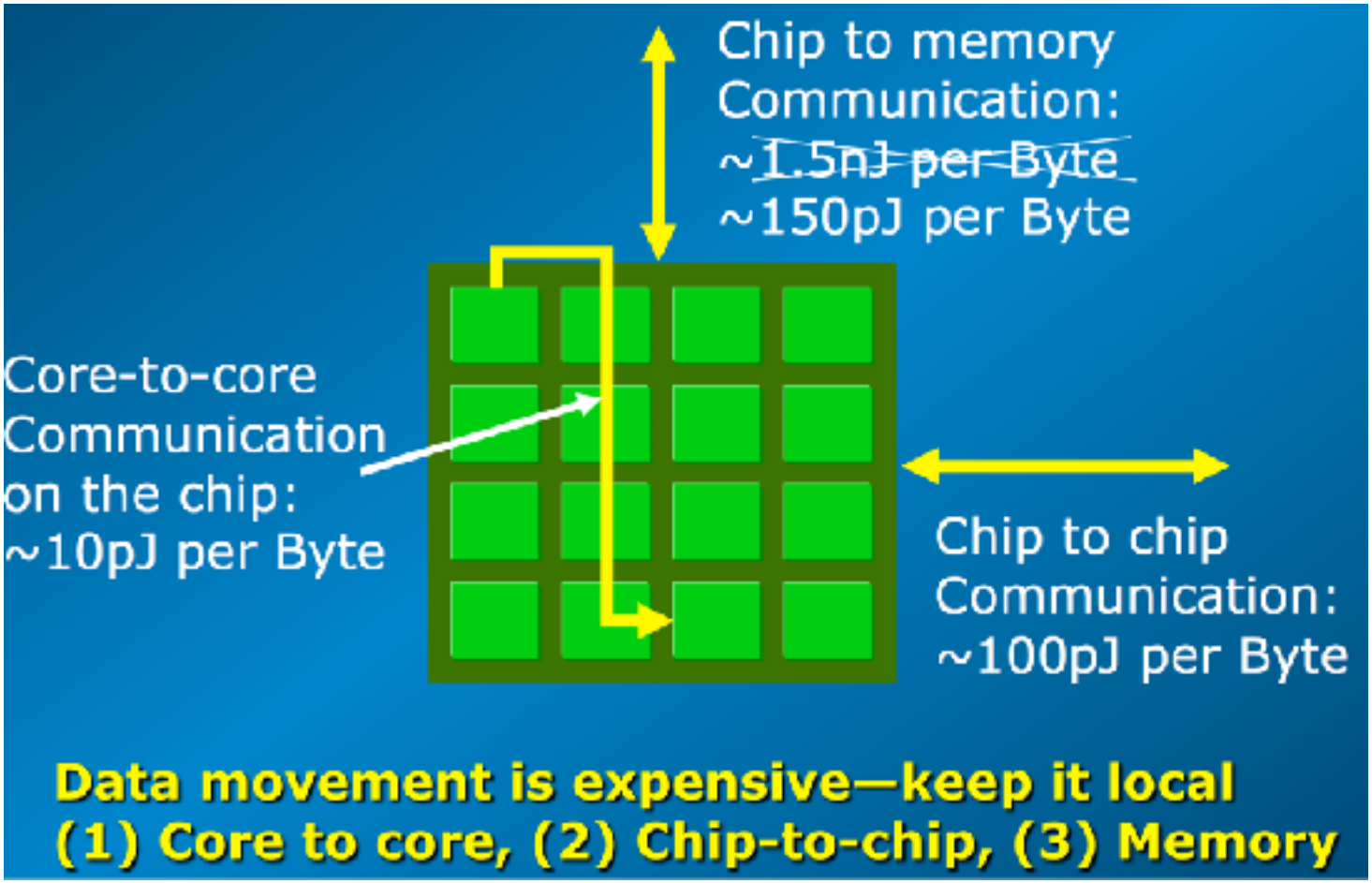


The most energy efficient configuration:

- **64GB** capacity per node
- **30nm DRAM process technology**
- running at **low voltage (1.35V)** and **1600Mbps**

pierson@irit.fr

Communication energy



pierson@irit.fr

A lot of energy is spent in off-chip Data Movement (40-50%!) [Lefurgy, IEEE Computer'03]

It takes time and energy to load/store data

and even more time than computing…

Functionality	Energy	Performance	Capacity (FP)
Core/FP Unit	50 pJ	125 ps	-
Register Bank	10 pJ	250 ps	100
Cache/SRAM	100 pJ	2 ns	100.000
Memory/DRAM	1000 pJ	100 ns	100.000.000

Source: Dally, 2011

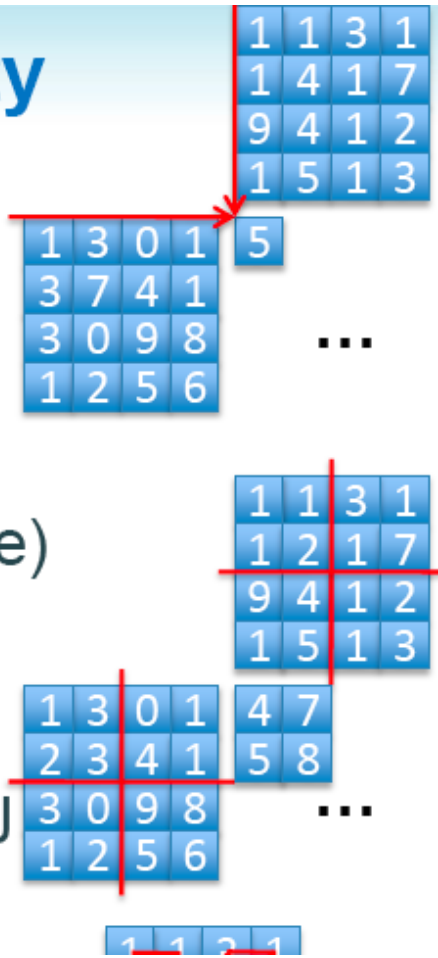
Memory hierarchy impact, per byte

pierson@irit.fr

For a matrix-multiplication

I/O Complexity and Power Complexity

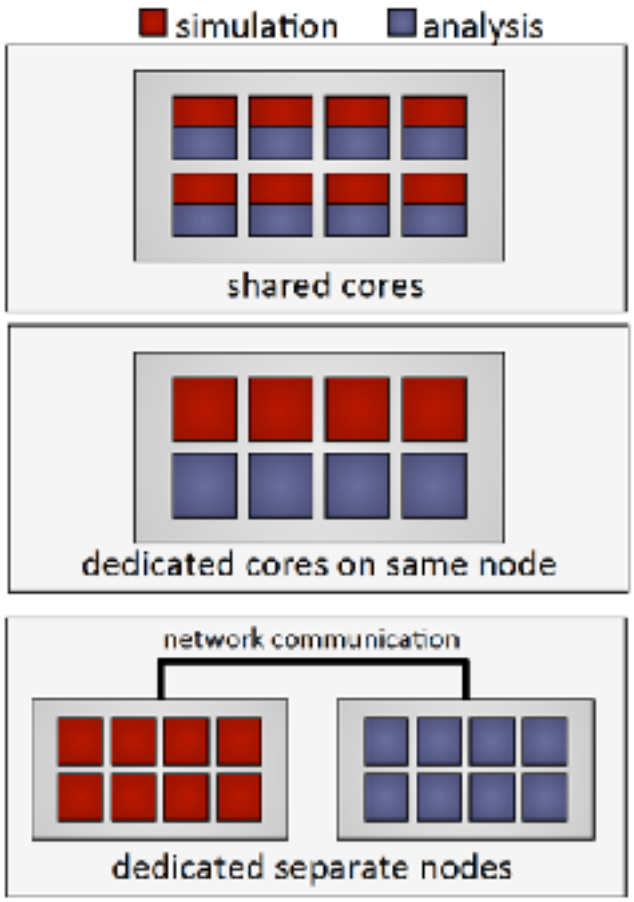
- Trivial algorithm (no reuse, N>50k):
  - $E(N) = (2N^3 + N^2) * 1\ nJ$
  - $E(55k) = 332.75\ kJ$
  - $FP(55k) = 55.000^3 * 50\ pJ = 8.32\ kJ$
- Block algorithm ($B=(N/C)^2$ CxC blocks fit in cache)
  - DRAM ops: $B(2N/C + C^2)$
  - Cache ops: $B(2C^3 + C^2)$
  - E(N,C) = [DRAM ops]*1nJ+[Cache ops]*0.1nJ
  - E(55k,35) = 10.78 kJ + 21.48 kJ = 32.26 kJ

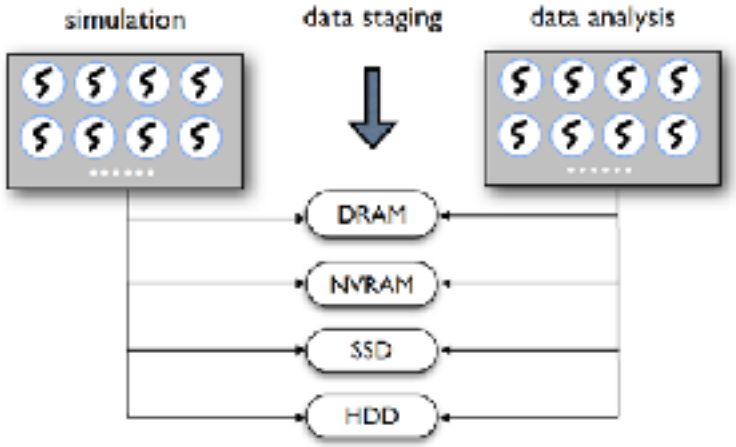


from T. Hoefler, EnA-HPC 2011

pierson@irit.fr

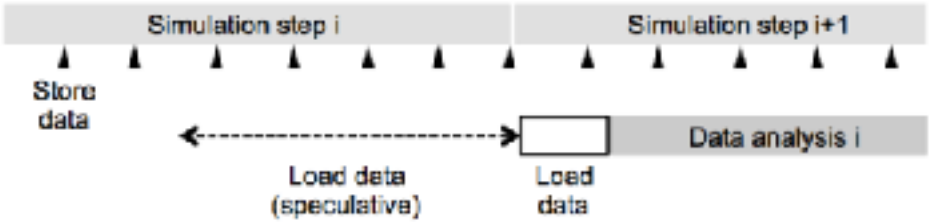
Towards In-situ Data Analysis





Data staging both vertically across multi-level memory hierarchy and horizontally across compute nodes

Speculative data movement



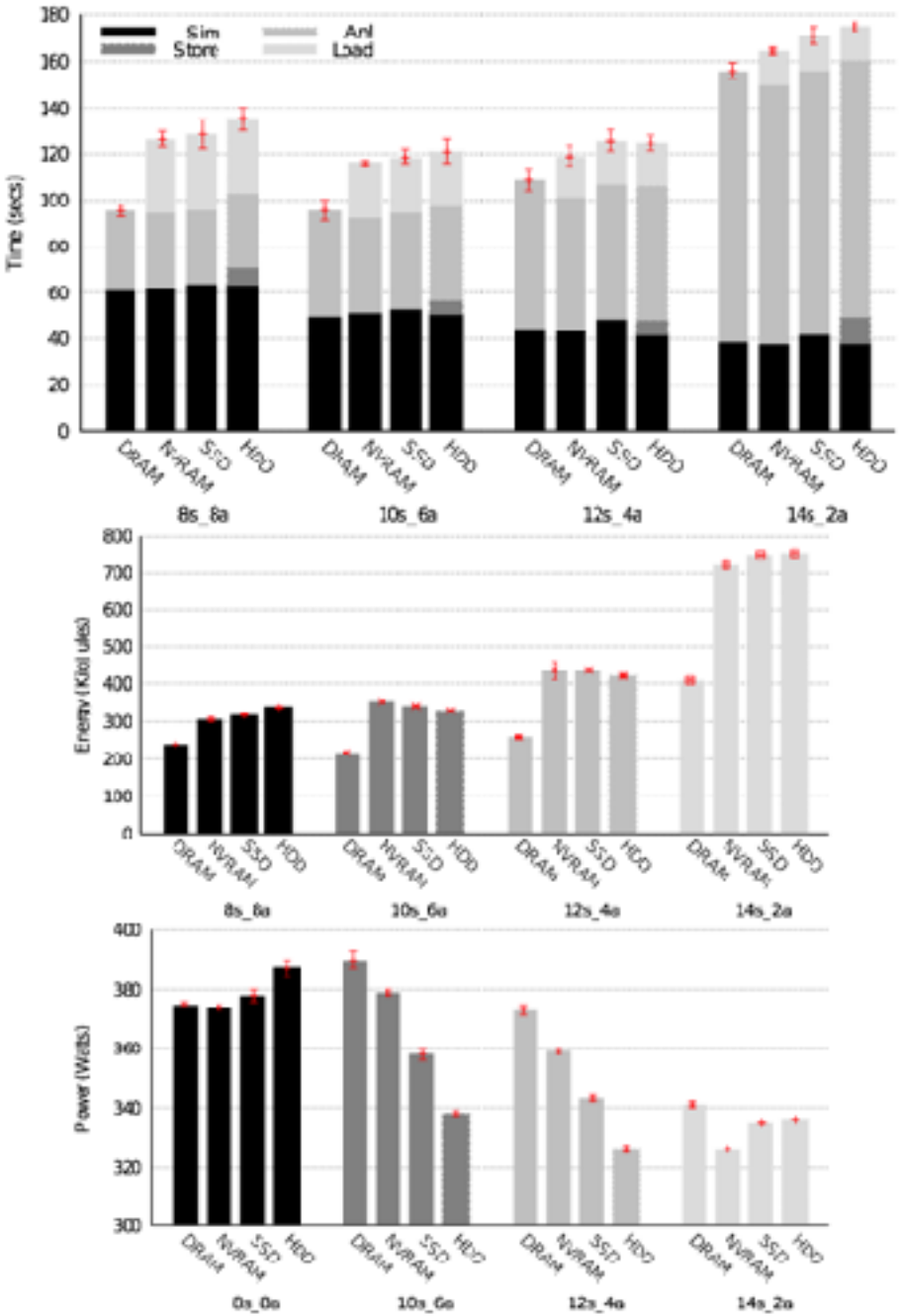
Rodero et al. 2014

pierson@irit.fr

Data Staging Impact

Key findings:

- the best is 12 cores for simulation and 4 for analysis
- using only 2 cores for analysis delays simulation tasks
- higher power for power demanding memory (DRAM)

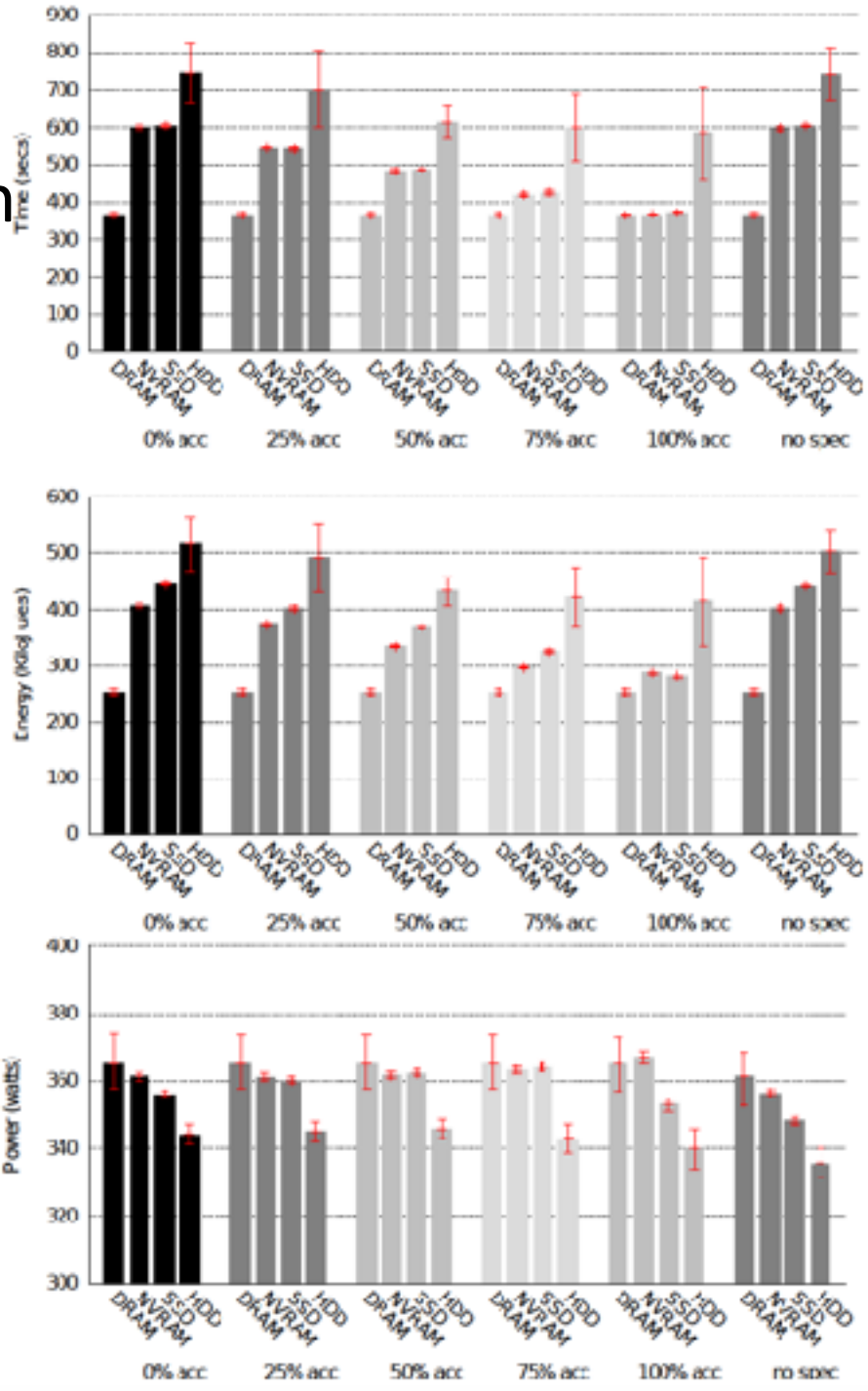


pierson@irit.fr

Impact of data movement speculation

Key findings:

- Almost no impact in terms of energy and time
- but average power increases with speculation! due to the data movement done in parallel to the simulation and analysis



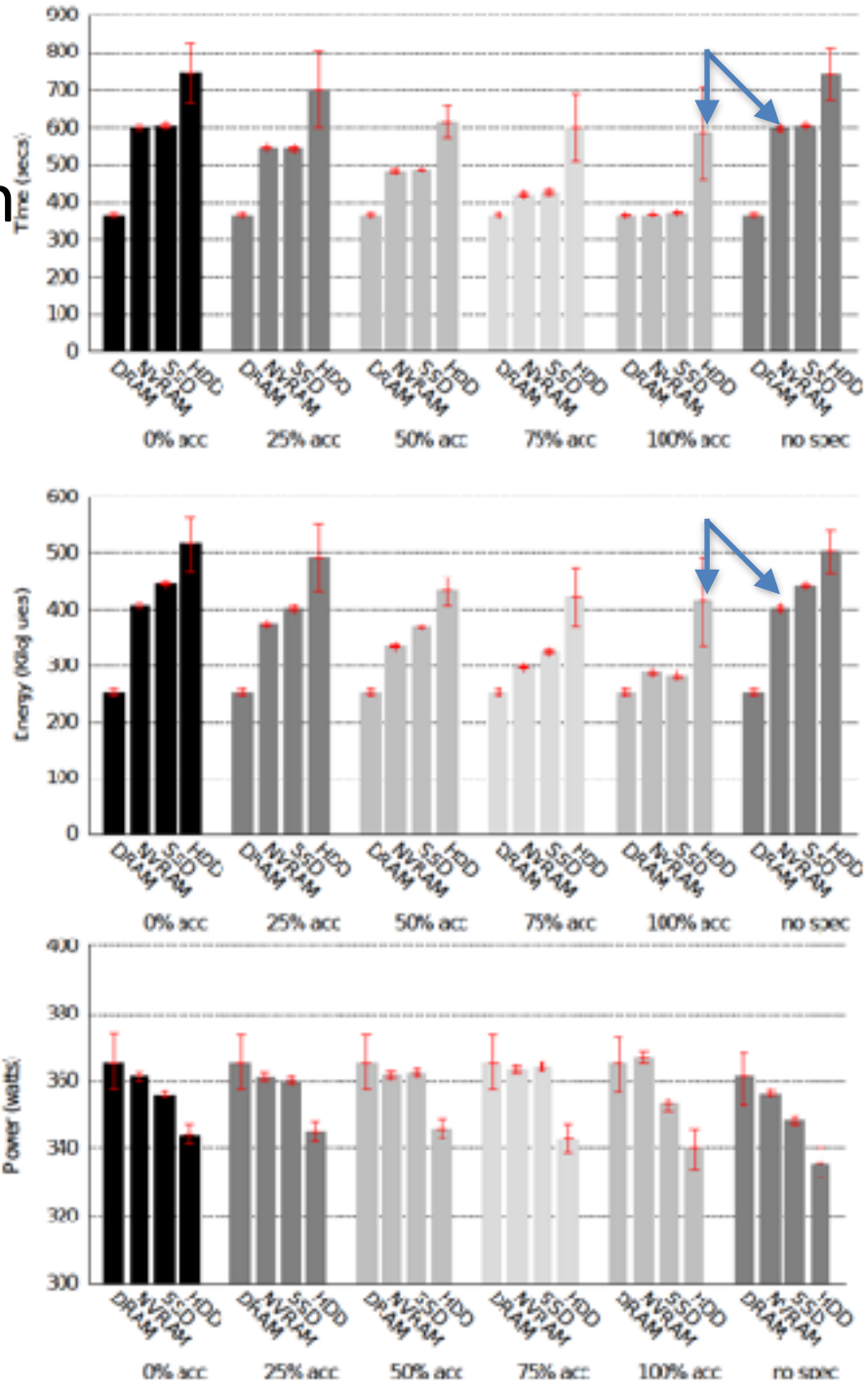
pierson@irit.fr

Impact of data movement speculation

Key findings:

- Almost no impact in terms of energy and time
- but average power increases with speculation! due to the data movement done in parallel to the simulation and analysis

And HDD with 100% accurate speculation gives better results than NVRAM both for energy and time



pierson@irit.fr

Reducing Energy…

Classical methods to save energy in large scale IT

- Switch off / remove useless resources:
  - Avoid redundancies in system (servers, networks) and data
  - Consolidation and virtualisation

pierson@irit.fr

Classical methods to save energy in large scale IT

- Switch off / remove useless resources:
  - Avoid redundancies in system (servers, networks) and data
  - Consolidation and virtualisation

NOT YET FOR HPC

pierson@irit.fr

Advanced technics

- **Adjust** automatically the system to the demand
- Take into account the electricty production means and energy market (smartgrids)
- And life cycle impact

pierson@irit.fr

Advanced technics

- **Adjust** automatically the system to the demand
- Take into account the electricty production means and energy market (smartgrids)
- And life cycle impact

NOT YET FOR HPC!!!

pierson@irit.fr

Measuring and Modelling Power Consumption

- hardware level
- external devices
- software models

Different sensors

- Hardware sensors: RAPL (Intel), NVML (NVidia)
- Intra-resource sensors: (to measure voltage or power at component level): PowerMon, Linux Energy Attribution and Accounting Platform (LEA2P)
- External devices: ePDU, Watt's Up, PowerPack, pmlib, KWAPI, PowerScope, …
- Software interfaces:
  - PAPI, Mummi, eclib, EML, for estimating power consumption

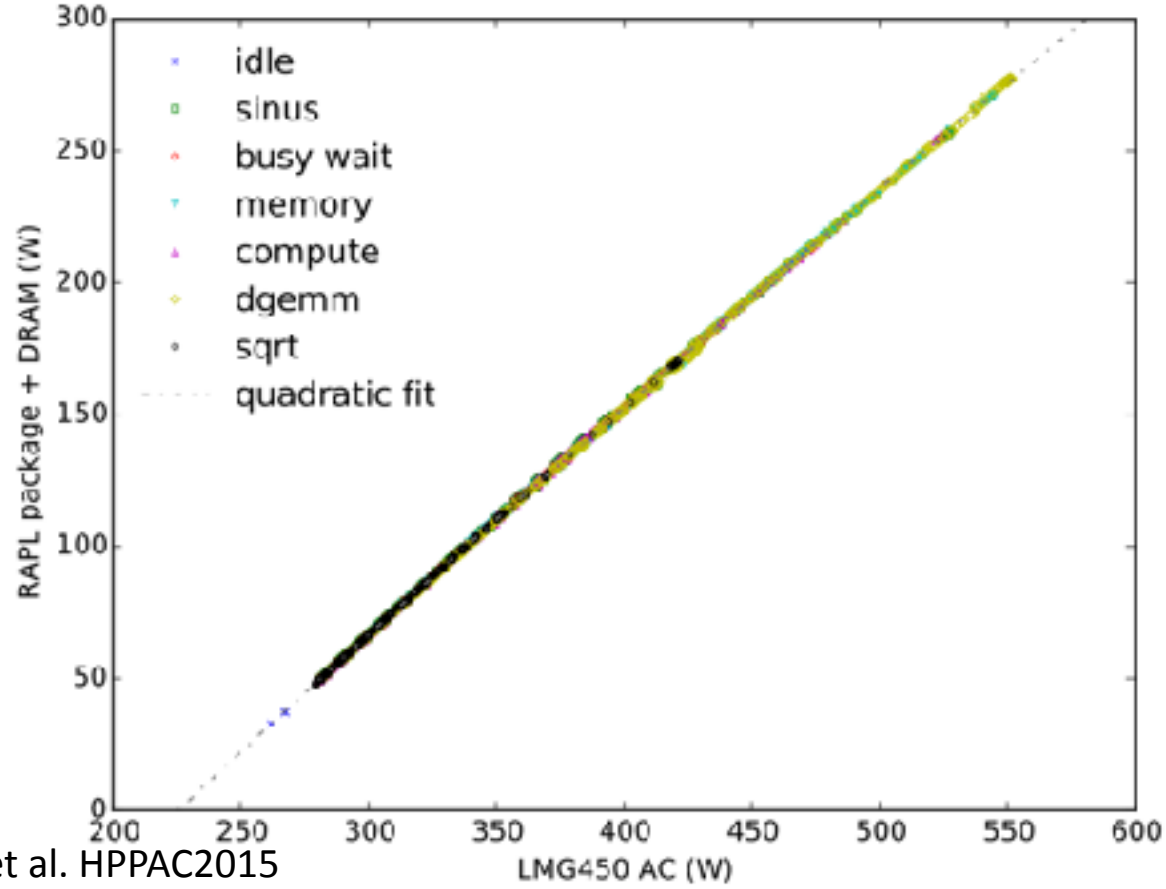
pierson@irit.fr

Hardware sensors
Measuring with RAPL
Running Average Power Limiting

Originally designed for power capping

Since Intel Haswell, RAPL values based on real measurements

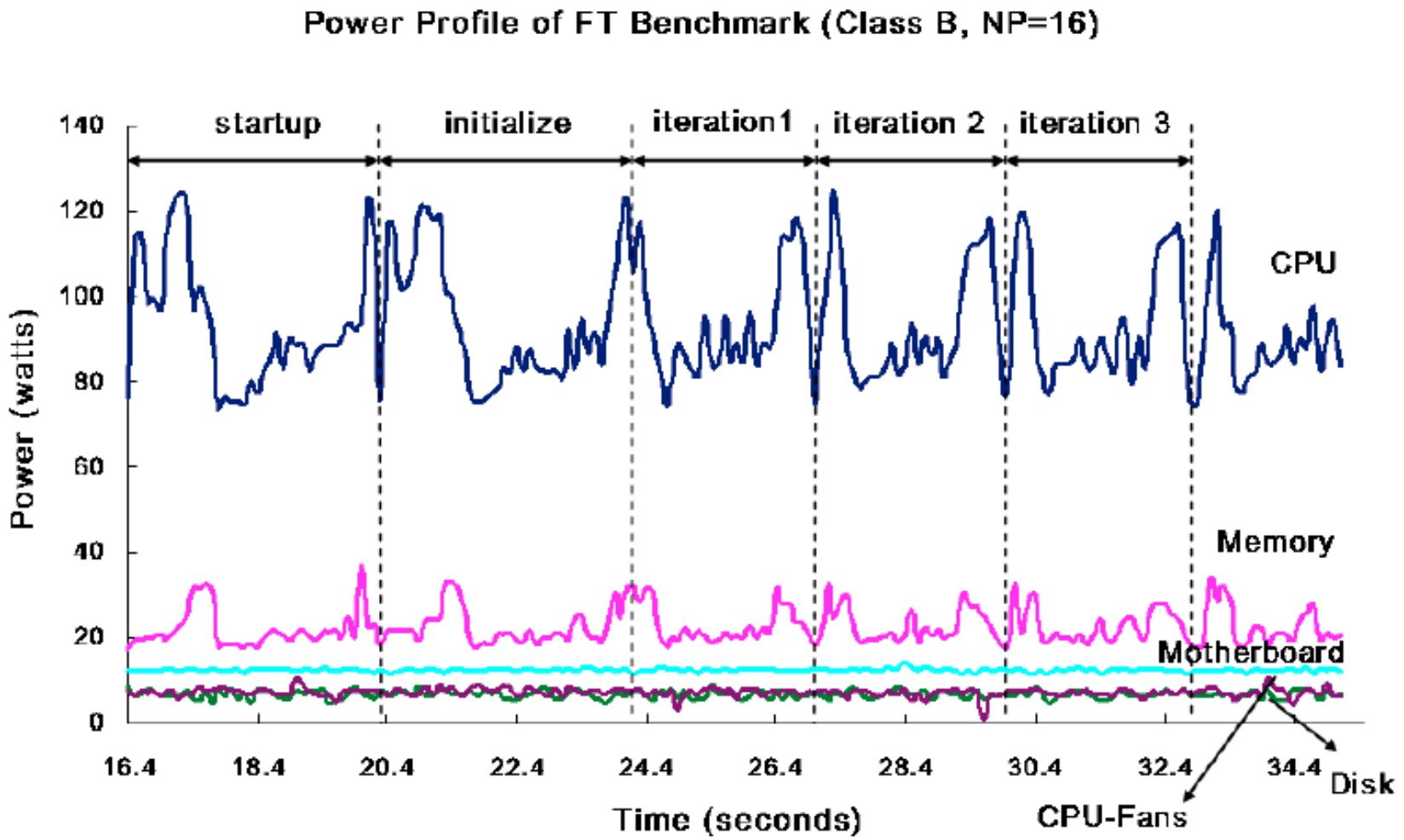
Correlation between RAPL reported data and direct AC measurements



Hackenberg et al. HPPAC2015

pierson@irit.fr

Power profile over time



From K. Cameron, with Powerpack

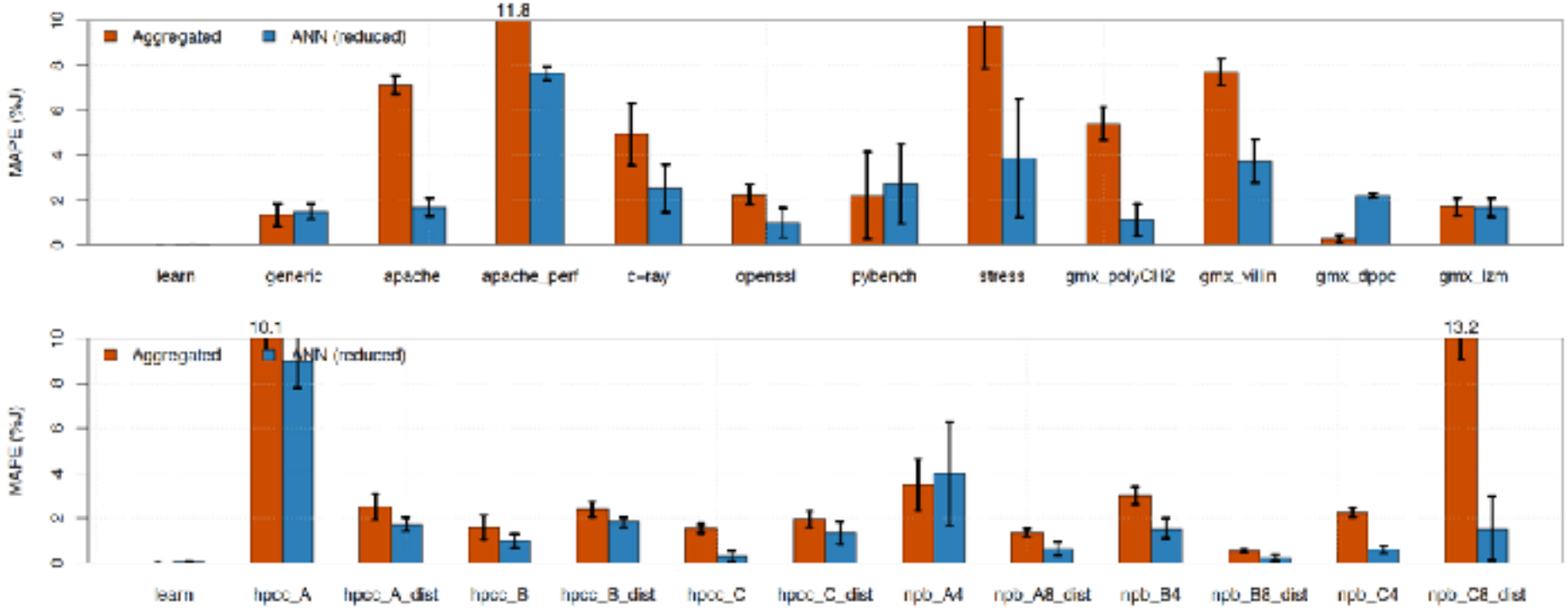
pierson@irit.fr

Estimating power, software approach

Predict power consumption using machine learning

- multivariate linear regression on HW Performance Counter preselected, or selected from PCA analysis (e.g. MuMMi framework)
- artificial neural network using HW Performance Counter and OS collected data without a priori selection

Using Artificial Neural Networks



The quality of prediction is higher than a priori approach

pierson@irit.fr

Cupertino, Da-Costa, Pierson, under review

Actions at the OS level, without knowing applications

CMOS Power

$P = C * V^2 * f$

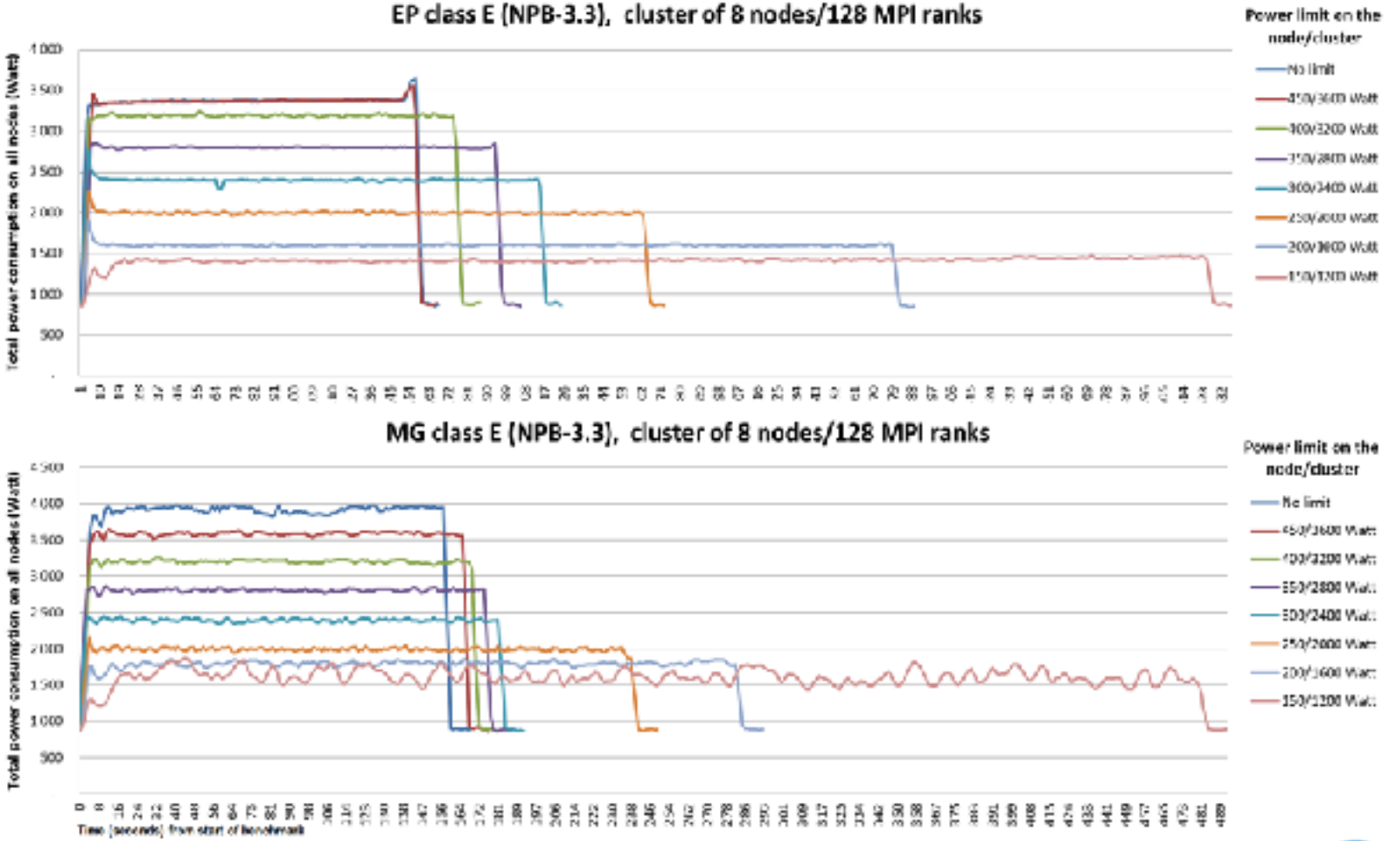
where

- C = capacitance of the circuit
- V = tension (voltage)
- f = frequency

To decrease Power -> decrease voltage or frequency (DVFS: Dynamic Voltage and Frequency Scaling)
OR limit the allowed power

pierson@irit.fr

Power Capping Effects (1/2)



Power Capping Effects (2/2)

NPB Test	Energy (kWh)		Gain	Most energy efficient power envelope per node (Watt)	Performance/Watt (Mops/Watt)		Gain	Best power envelope per node (Watt) for power/performance
	Total at no power limit	Min energy			at no power limit	Best Perf./Watt		
CG	1.63	1.24	1.31x	**300**	5.83	7.70	1.31x	**300**
MG	0.17	0.12	1.41x	**300**	42.10	61.86	1.47x	**300**
LU	3.87	2.62	1.47x	**300**	46.02	67.87	1.47x	**300**
BT	3.28	2.66	1.23x	**300**	79.16	96.56	1.22x	**300**
SP	4.79	3.2	1.49x	**250**	27.42	40.50	1.44x	**250**
EP	0.145	0.143	1.01x	**350**	4.21	4.49	1.06x	**300**

- Amount of consumed energy varies from application to application and depends on the imposed power limit on the node
- The most **"power efficient"** power limit won't necessarily be the most **"energy efficient"** one!

Right choice of power envelope for application can result in significant energy savings

pierson@irit.fr

Adapting the OS at runtime

Idea: Adapt the governor according to the activity

Our approach: NetSched, using network activity indicator.
Simple Rule based:
Every 100ms :

- If Currently in Slowest frequency and Network Activity < Threshold --> Change frequency to Fastest
- If Currently in Fastest frequency and Network Activity > Threshold --> Change frequency to Slowest

Results: (compared to performance governor)

- makespan: +8% on IS, -5% on LU, stable for others
- energy: down to -25% on FT

From Da Costa, PDP 2015.

Same kind of idea with REST (Intel / INRIA)
- If the program is memory bound, lower the frequency
- If the program is CPU bound, heighten the frequency

pierson@irit.fr

Still at system level: Use phases of applications

Approach:

- detect and characterize the system's runtime behaviours/ phases
- partial phase recognition for phase identification
- systems adaptation (storage, memory, interconnect, CPU) for each phase .

Table: Translation of phase characteristics into system adaptation (IO related sensors include network and disk activities).

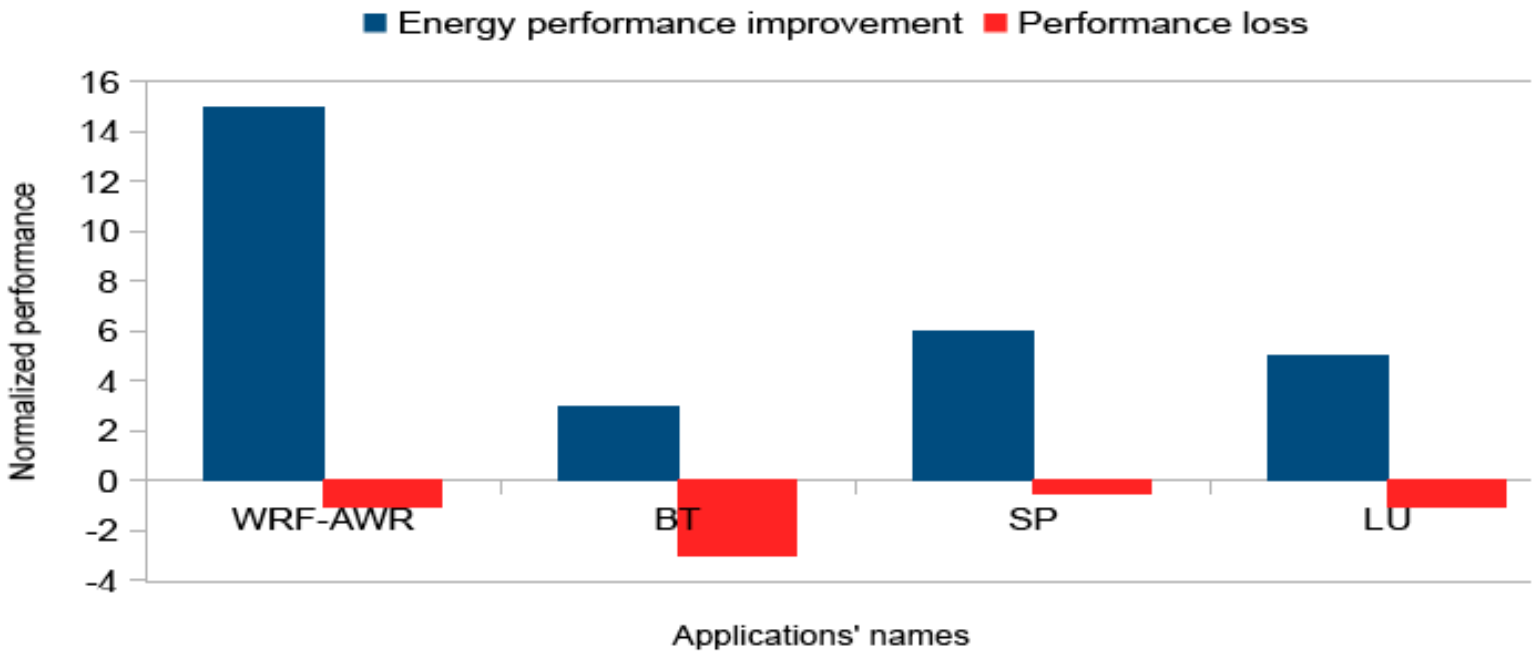
Sensors selected from PCA for phase characterization	Decisions
cache_references & cache_misses & IO related sensors	CPU frequency set to its maximum spin down the disk network speed scaled down
no IO related sensors	CPU frequency set to its lowest network speed scaled up
instructions & last level cache misses (llc)	CPU frequency set to its minimum network speed scaled up
instructions or llc & IO related sensors	CPU frequency set to its average value network speed scaled down spin down the disk
IO related sensors	CPU frequency set to its maximum spin down the disk network speed scaled down

From Ghislain Landry Tsafack Chetsa. ICPADS 2012

pierson@irit.fr

Saving energy at the small price of performances

Comparison with baseline execution (Linux's on-demand governor)

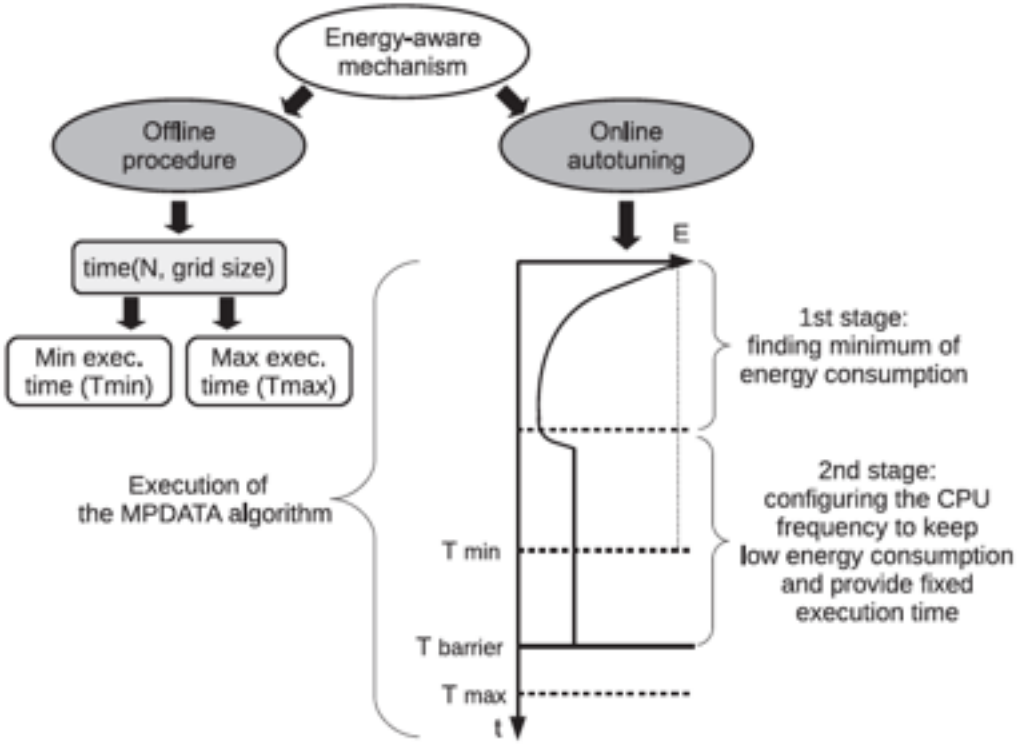


From the application runtime

- Observe the iterative application running for a couple of iterations, learn the behaviour
- Use DVFS to adapt frequency to avoid idle time

pierson@irit.fr

Example with Stencil applications



Stage 1: starting from the max frequency, find the most energy efficient one for each stencil at runtime, so that estimated time constraints is not exceeded

Stage 2: reconsider frequency to not exceed time constraints

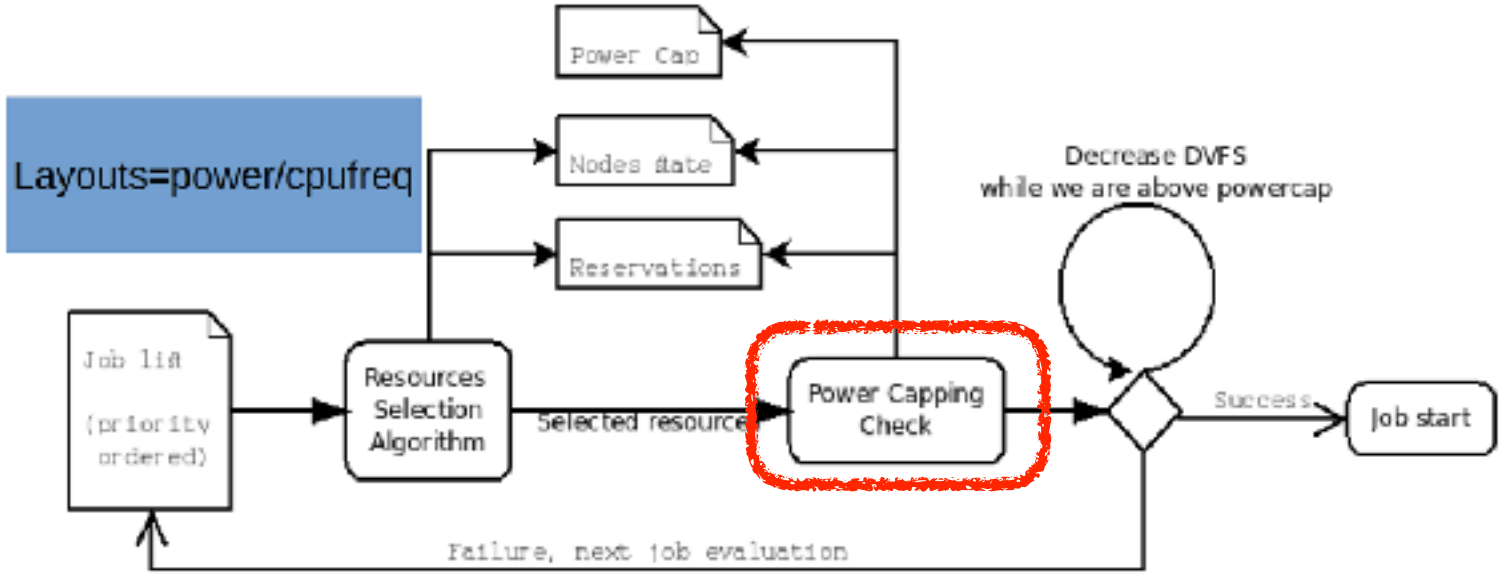
Also, adapted when another application share at one point the CPU…

K. Rojek, A. Ilic, R. Wyrzykowski, L. Sousa, 2016

pierson@irit.fr

Actions at the middleware level
(batch scheduler)
- Power Capping
- Energy Capping
- Scheduling

In SLURM (15.08)



Calculate the power consumption of the cluster if the job is executed

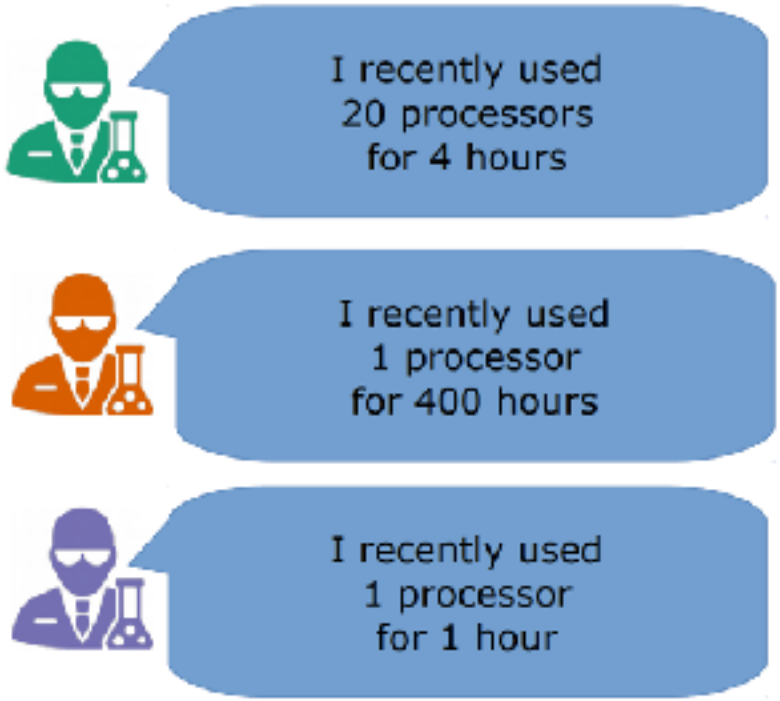
- If higher than the allowed power budget, check if DVFS is allowed for the job

Reductions through DVFS, idle and shut-down nodes

Extension also to Energy capping

pierson@irit.fr

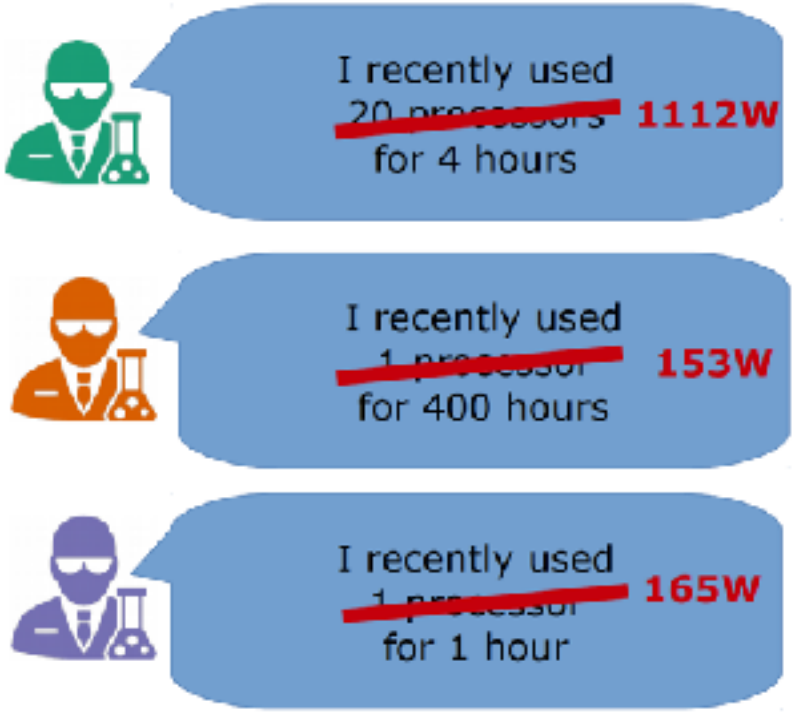
Fair Sharing



Fairshare counters	Users weights	Score
20*4=80	1	80/1=80
1*400=400	4	400/4=100
1*1=1	1	1/1=1

pierson@irit.fr

Energy Fair Sharing



Fairshare counters	Users weights	Score
1112*4= 4448	1	4448/1= **4448**
143*400= 57200	4	57200/4= **14300**
165*1=165	1	165/1=**165**

pierson@irit.fr

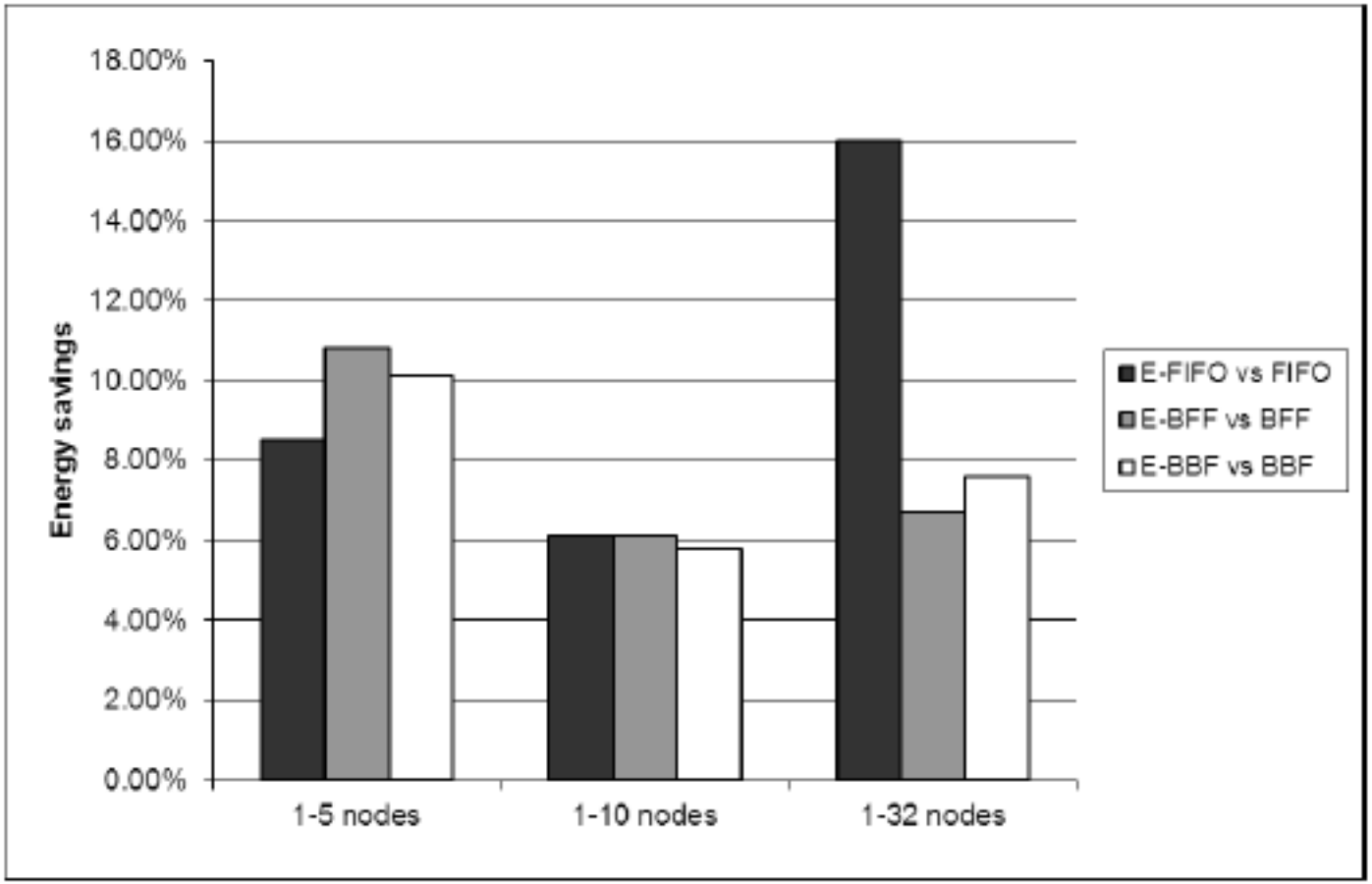
Glesser 2015

Energy Efficient Job scheduling in HPC

Several possibilities exist for energy efficient or energy-aware scheduling (Mammela et al.):
-E-FIFO: switch-off unused nodes if first job in the queue can not be run before T seconds
-E-BBF and E-BFF: backfilling (best fit, first fit) + switch-off servers like E-FIFO

AND / OR, use DVFS:
- Adjust frequency of the processor when a job start based on the number of jobs in the queue, and the Utilization level of the HPC System: low, medium, high (Etinski et al.)
- Adjust frequency based on the application DAG. (Dolz et al.)

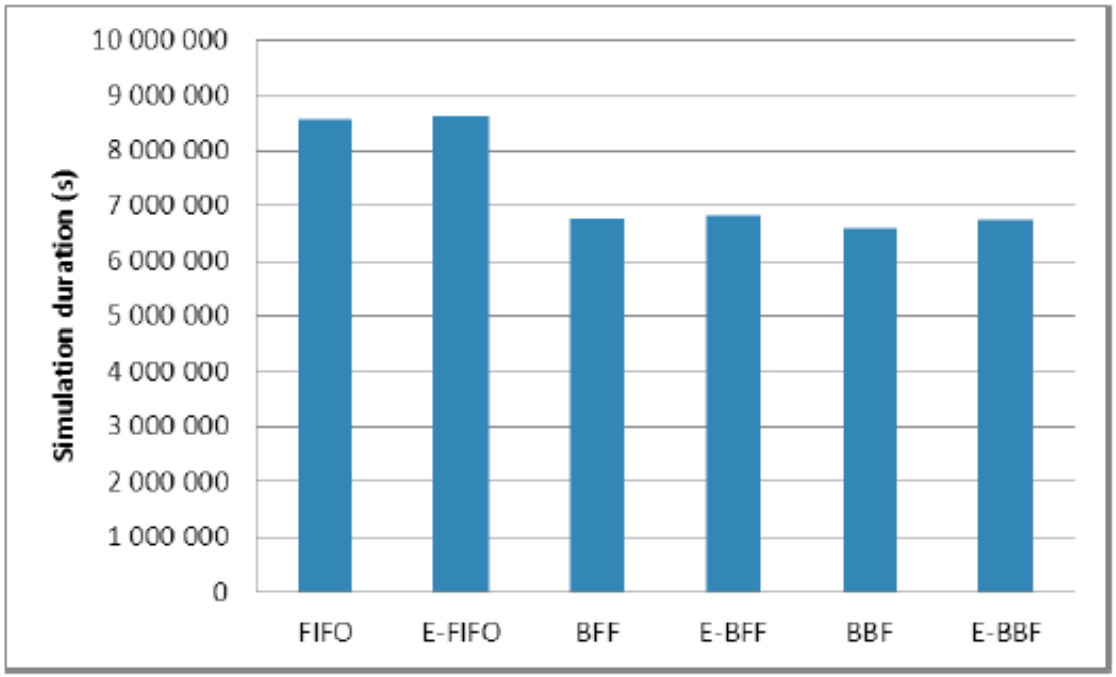
pierson@irit.fr



Results: potential for energy savings, depends on workload using energy-aware versions

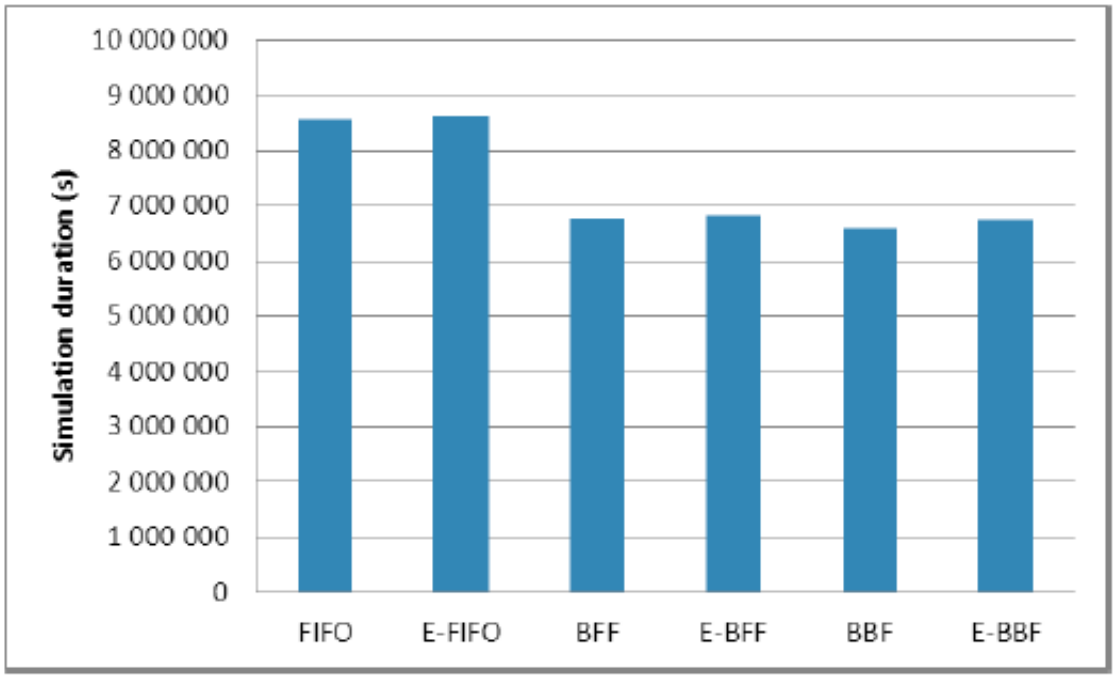
from O. Mammela, EnA–HPC 2011

pierson@irit.fr



Maximum +2.37% on makespan

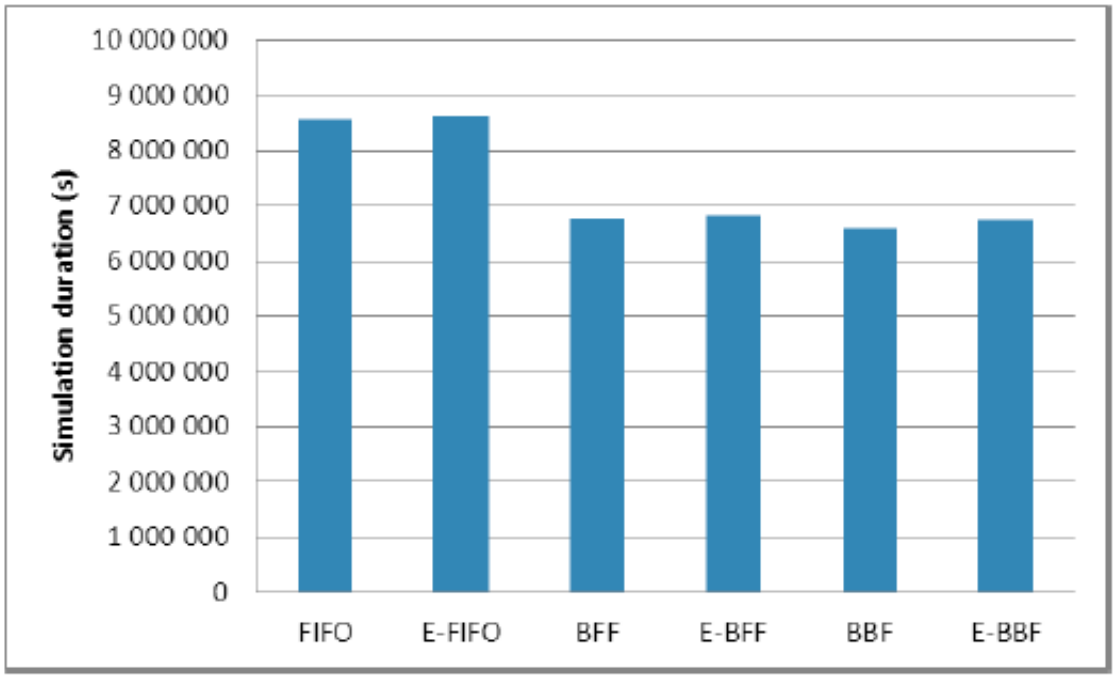
pierson@irit.fr



Maximum +2.37% on makespan

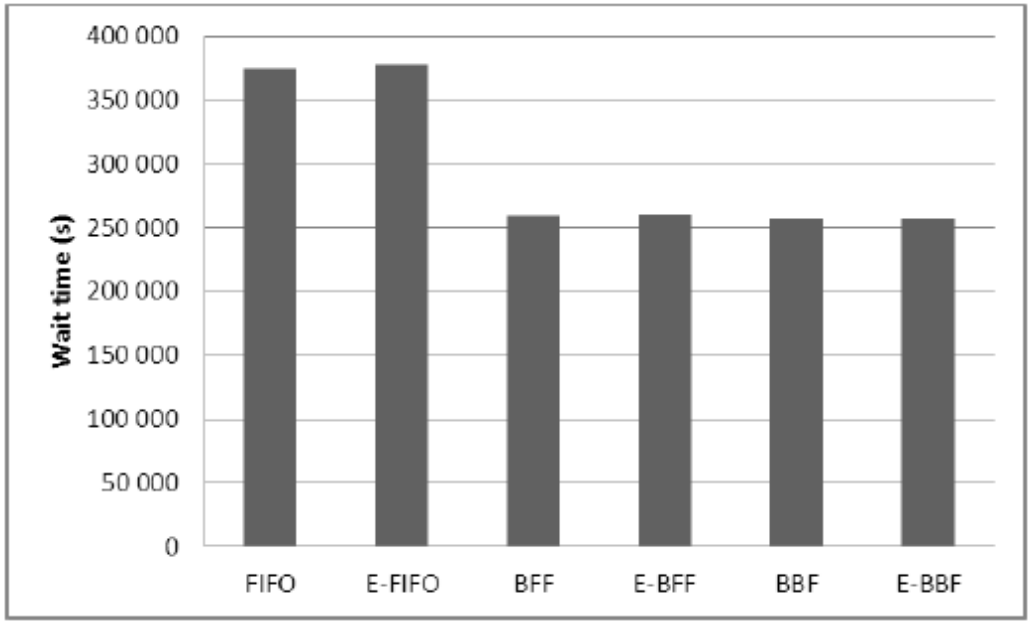
Maximum +0.81% on waiting time

pierson@irit.fr



Maximum +2.37% on makespan

Maximum +0.81% on waiting time



pierson@irit.fr

Using the right speed for fork/join tasks

- Allocate tasks with a greedy algorithm based on execution time, then use a scaling factor on each task/node to adjust frequency
- It gives the same assignment than a greedy based on energy consumption
- the scaling factor can be analytically expressed

T. Rauber, G. Rünger, CCPE2015

pierson@irit.fr

Reducing slack and adjusting DVFS, or Run and Rest

Two strategies:

1. SRA: Slack Reduction Algorithm:
   Reduce frequency of cores/processors that execute non-critical tasks to decrease idle times without sacrifying total performance of the algorithm: find critical path on DAG and slow down tasks not of the critical path
   Result: Higher execution time, more energy consumption
2. RIA: Race to Idle Algorithm
   Execute all tasks at highest frequency to enjoy longer inactive periods
   Result: Stable execution time, reduced energy consumption

from M. Dolz, Univ. Jaume, EnA-HPC 2011

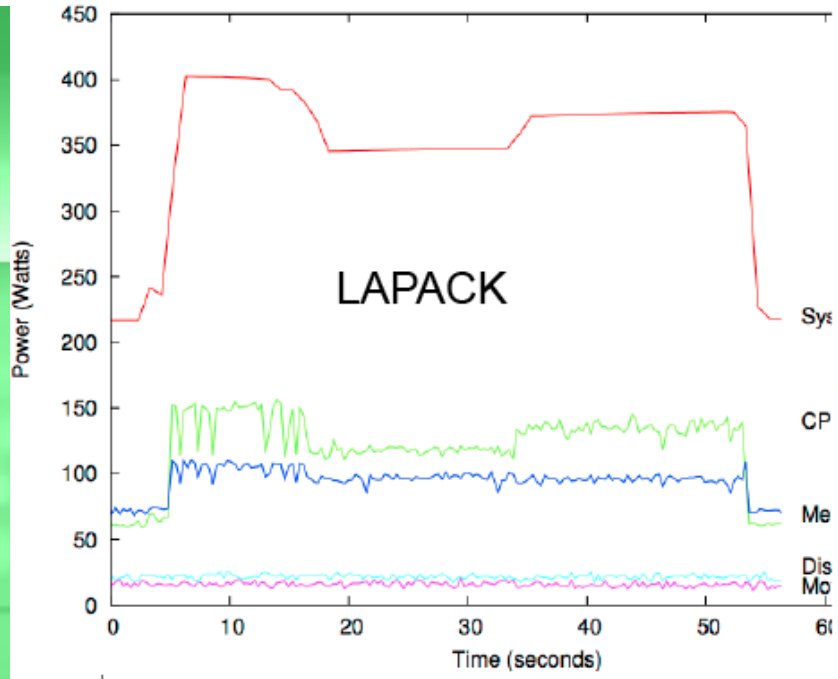
Hybrid computing

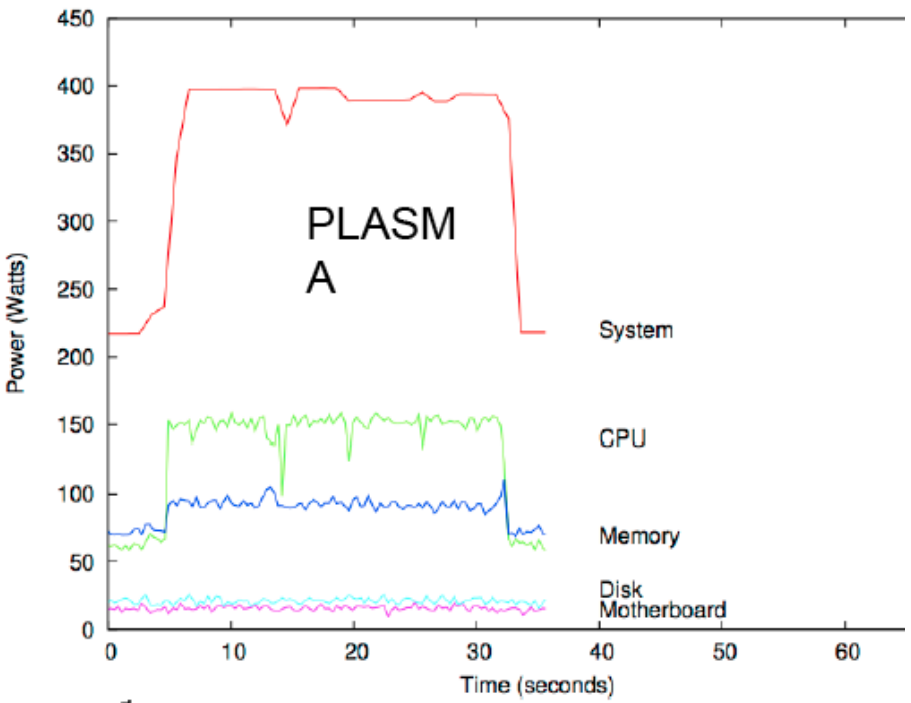
- CPU / GPU processing
- GPU are VERY efficient for SOME algorithms (regular, matrix based)
- CPU for others
- Need communications between both, i.e. between memories and processors
- Synchronous barriers between GPU and CPU
- Which part of the code on which processor family?
  - At RunTime: for instance StarPU algorithm
  - At placement phase: need to be able to model in advance, e.g. Hydrasim (developed by G. Da Costa, IRIT, http://hydrasim.sourceforge.net/)

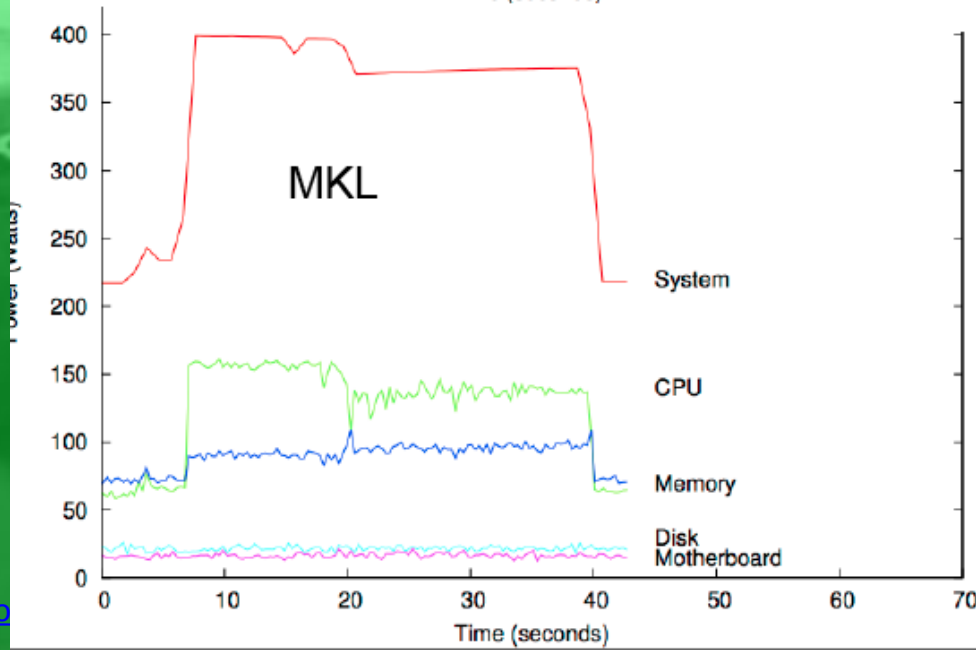
pierson@irit.fr

Impact of software developments

Impact of the programming language and the library used







Power consumption over time

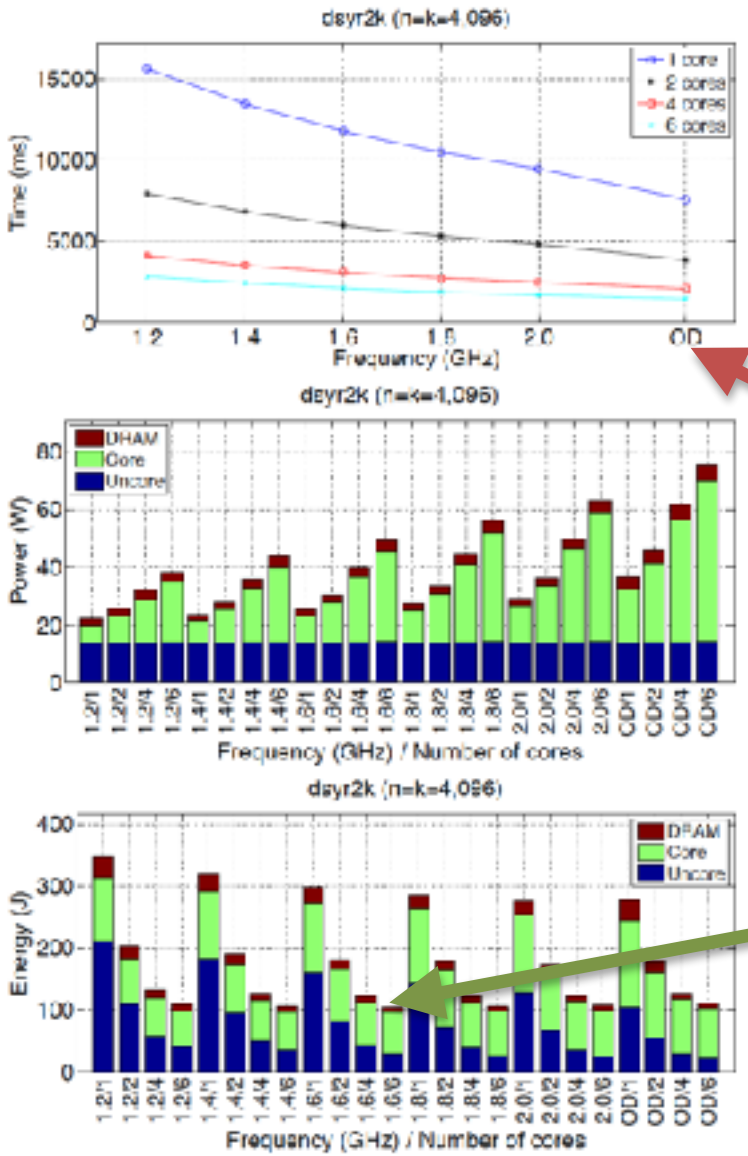
Matrix inverse

Comes from Powerpack

Sources:

Piotr Luszczek Hatem Ltaief

Characterising HPC kernels



BLAS3, 1st experiment

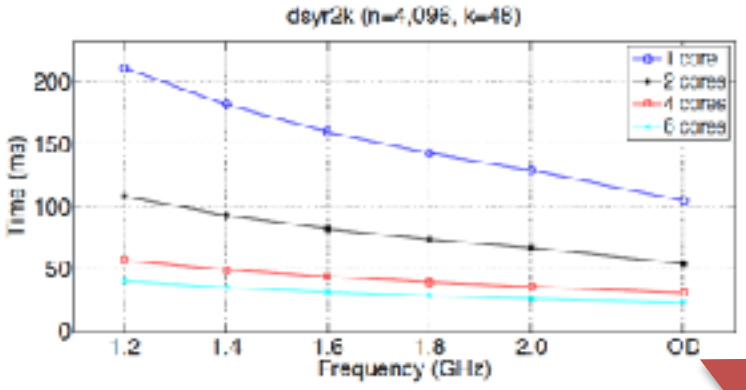
dsyr2k
(n=k=4,096)

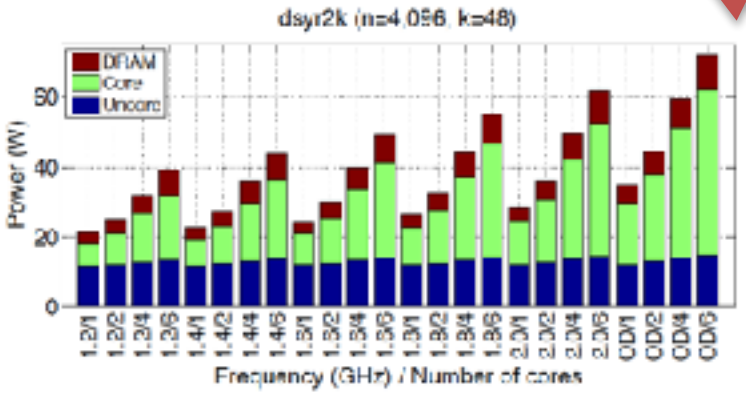
- **Execution time**
  - Decreases when the number of cores and CPU frequency increases
  - Best option: OD/6 cores!
- **Average power**
  - Increases with the number of cores and CPU frequency
  - Almost only Core power changes!
- **Energy consumption**
  - The greenest is 1.6 GHz/6 cores
  - The fastest is OD/6 cores
  - Performance: 1.43×; Energy efficiency: 0.93×

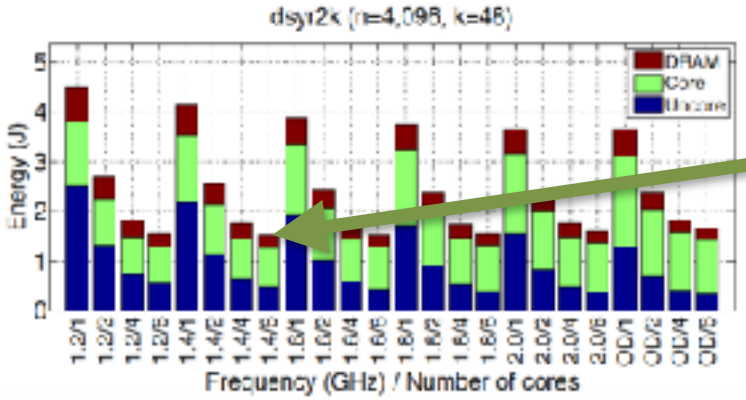
Dolz et al. 2014

pierson@irit.fr

Characterising HPC kernels







BLAS3, 2nd experiment: change a parameter

dsyr2k
(n=4,096, k=48)

- **Execution time**
  - Decreases when the number of cores and CPU frequency increases
  - Best option: OD/6 cores!
- **Average power**
  - Increases with the number of cores and CPU frequency
  - Core and DRAM power changes!
- **Energy consumption**
  - The greenest is 1.4 GHz/6 cores
  - The fastest is OD/6 cores
  - Performance: 1.52×; Energy efficiency: 0.92×

Results depend on the **arithmetic intensity of operations**: **ratio of FLOPS over memory operations** (depends on 2k)

pierson@irit.fr

Source instrumentation

- Access to the source code
- Add inside the code some timestamps when using some methods
- Examples: Energy Checker Intel, ADIOS/CIAO, EML

Very precise

Intrusive, time consuming

pierson@irit.fr

Source instrumentation
EML : Energy Monitoring Library

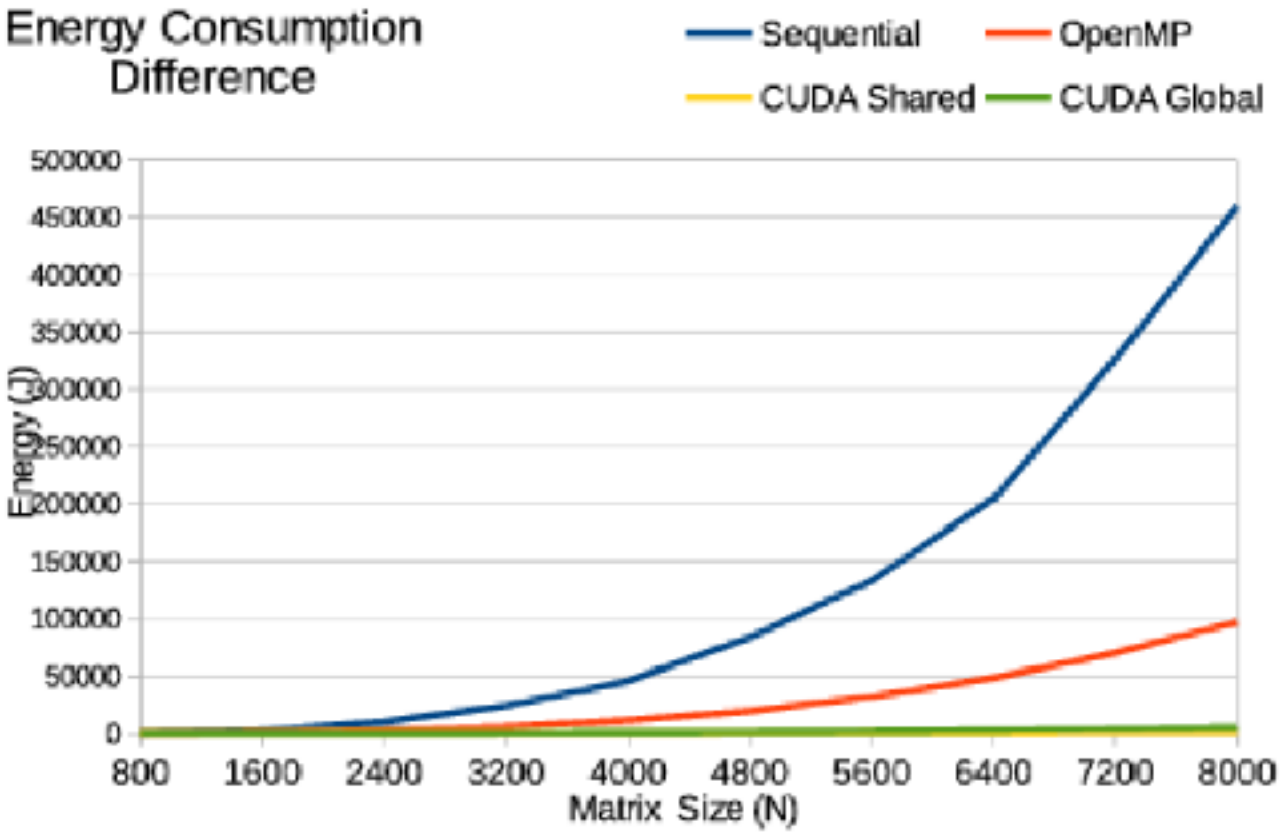
```
#include <eml.h>
#include <stdlib.h>

int main() {
  emlInit();
  //get total device count and allocate result handles
  size_t count;
  emlDeviceGetCount(&count);
  emlData_t* data[count];

  emlStart();
  //...do work...
  emlStop(data);
  //...use data...
  emlShutdown();
}
```

Used for instance to compute communication energy or to compare computation energy on Sandy Bridge and GPU (here a matrix multiplication)

Cabrera et al. 2014

pierson@irit.fr

Source instrumentation EML, Comparing



CUDA versions consume much less than Sandy Bridge ones

pierson@irit.fr

Near Threshold Computing

Near-Threshold Voltage Computing (1/3)

- Set the voltage of processors under the minimum recommended voltage
- Leads to “some” marginal errors (bit flips) —> decreases performances by 5-10 times
- Power savings potential of 10-50 times
- Energy reductions from 2 to 5 times
- Only usable when computation with errors possible (iterative solvers, signal processing) or redundancy

pierson@irit.fr

Near-Threshold Voltage Computing (2/3)

» Notion of Code Significance: measure the susceptibility of code to errors and effects on end results.

» Experiments with different location of bit errors

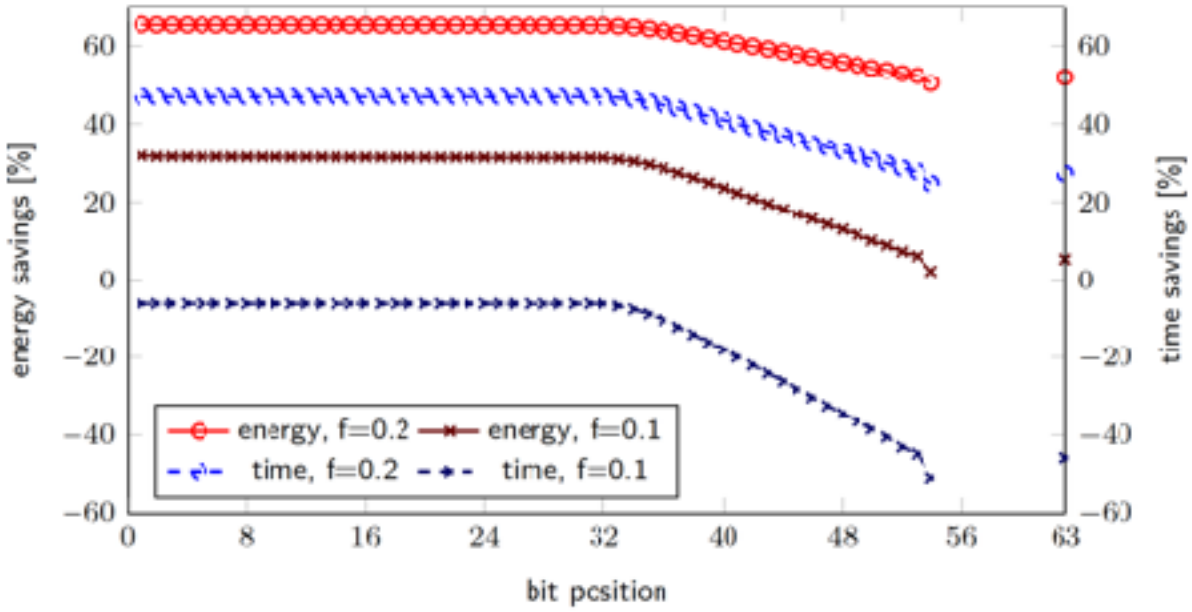


Figure: Energy and time savings over correct, sequential Jacobi for 16 unreliable cores.

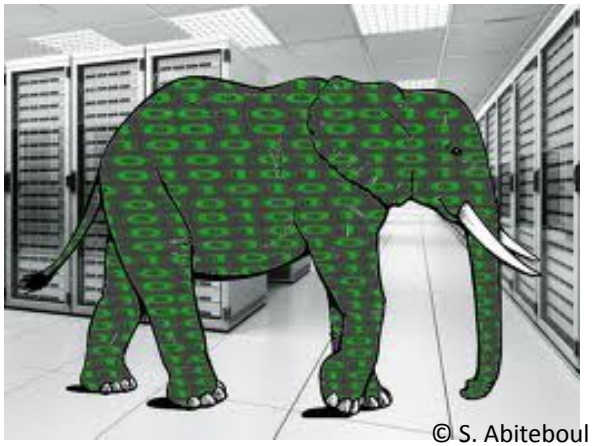
Gschwandtner 2014

pierson@irit.fr

Future of HPC and EE

From Compute intensive to Data intensive applications

- More and more data,
- Big Data


© S. Abiteboul

- More to do with:
  - Loading data in the system : IO, disks and networks
  - Processing huge amount of data: memory!
  - Communicating data among nodes:
    - reduce data movement, favour locality

pierson@irit.fr

Towards Exascale

Today's best: ~93 petaflops
Exaflops and beyond: > x10

Sunway energy consumption: ~15 MW
Exaflop machine ?? --> 150 MW

From Green500 list from jun 2017, an exascale system with the best level of energy efficiency would draw already 70 MW.

pierson@irit.fr

The question is not anymore
Can I save energy when doing HPC?
There is no choice!
but
How to do energy efficiency successfully? »