CNRS - INPT - UPS - UT1 - UT2J

Mutli-scale optimization of datacenters Energy and Performance

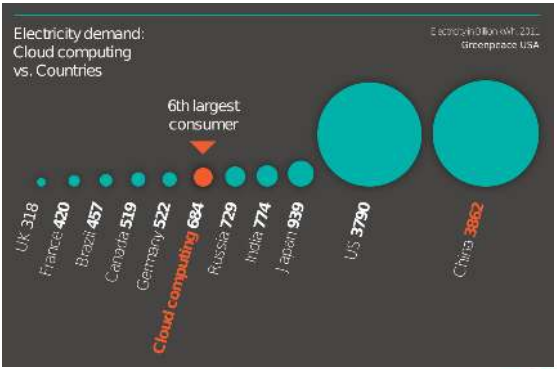
Georges Da Costa

ICICafé

SEPIA Team

IT impact on electricity

- Recent datacenters: 40000 servers, 500000 services (virtual machines). Google, Facebook > 1 million servers
- One major power consumer
  - 2000 : 70 TWh
  - 2007 : 330 TWh, 2% of CO_2 world production
  - 2011 : 6^{th} electricity consumer in the world
  - 2020 : 1000 TWh
- Rising
  - 2014 to 2016: 90% of datacenters were expected to need hardware upgrades



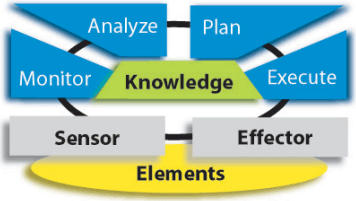
Sustainable datacenters

- Action can be done at several different levels
  - Hardware level: changing servers or cooling system
    - If entropy is constant, theoretical energy consumtion is 0 !
  - Application level: rewrite applications while changing paradigm*or library
  - Middleware level: manages servers and services/applications
- Middleware: minimal cost, maximal impact
  - OpenStack: 30% of market share in 2014
  - OpenSource solutions: 43% (+72% in 2 years)

* Georges et al. *Exascale machines require new programming paradigms and runtimes*, SFI journal, 2015

Middleware systems

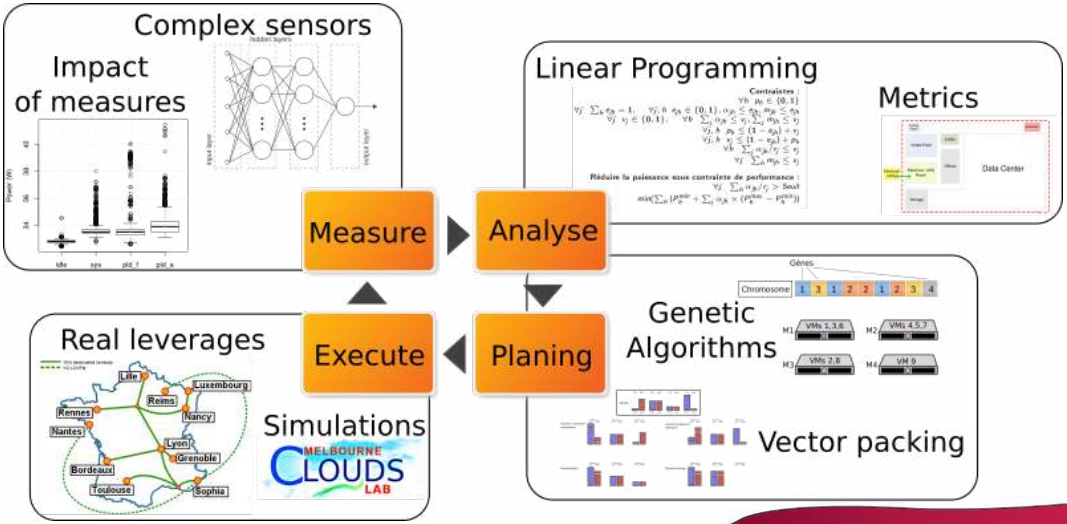
- Two goals:
  - Managing (needs, errors, faults, overheating)
  - Optimizing (Energy, performance)
- Leverages
  - Switching on/off, DVFS
  - Migration (x86/ARM)*, reduction of allocated resources, suspend



MAPE-K loop ©IBM

* Violaine et al., *Big, Medium, Little : Reaching Energy Proportionality with Heterogeneous Computing Scheduler*, Parallel Processing Letters, journal, 2015.

Autonomic loop



Node optimization

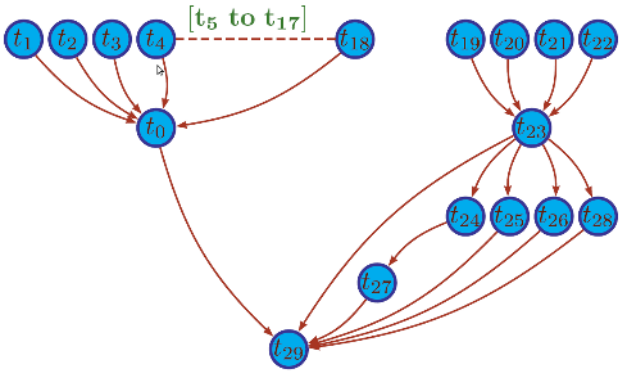
- Three temporalities
  - Large-grained (minute) : Optimal frequency in function of the task graph*
    - 13% of energy savings
  - Medium-grained (second) : Phase detection†
    - 20% of energy savings, 3% of time increase
  - Fined-grained (1/10s) : Frequency policy at the kernel level‡
    - 25% of energy savings, 1% of time decrease
- No coordination between the three temporality, no objectives

* Tom et al., *Energy-aware simulation with DVFS*, SMPT journal, 2013 †Landry et al., *Exploiting performance counters to predict and improve energy performance of HPC systems*, SUSCOM journal, 2014 ‡Georges et al., *DVFS governor for HPC: Higher, Faster, Greener*, PDP conference, 2015

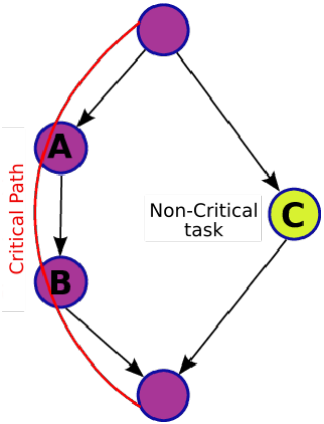
Plan

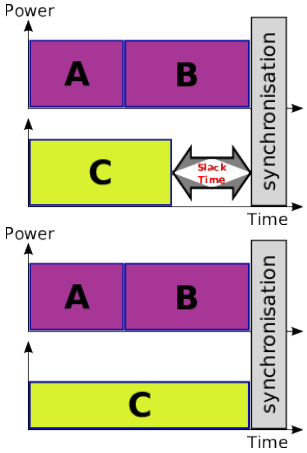
At the scale of a node: Large-grained

- Use of contextual external information
- Example at the scheduler level: Task DAG

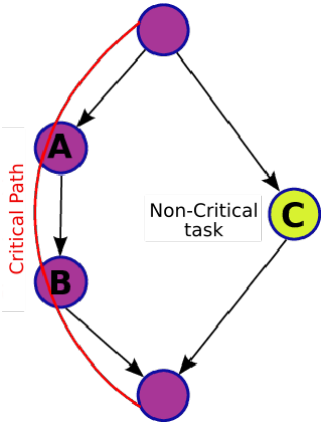


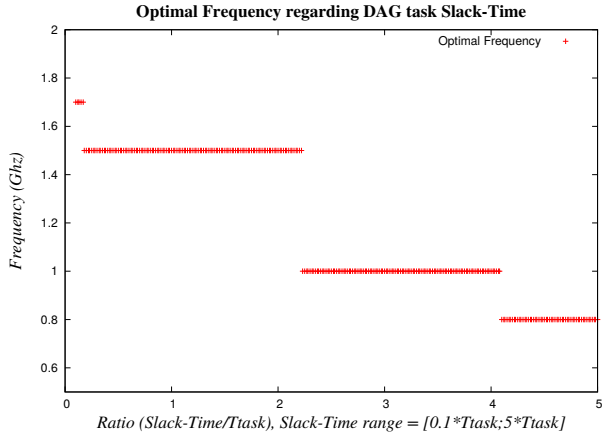
Coordination of node speeds





Coordination of node speeds



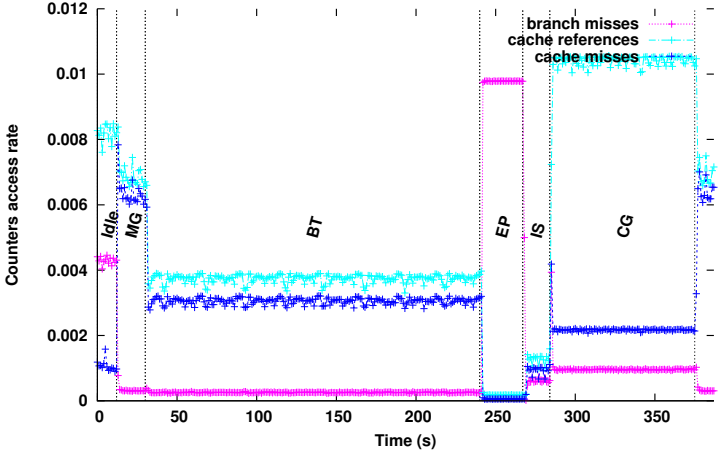


Plan

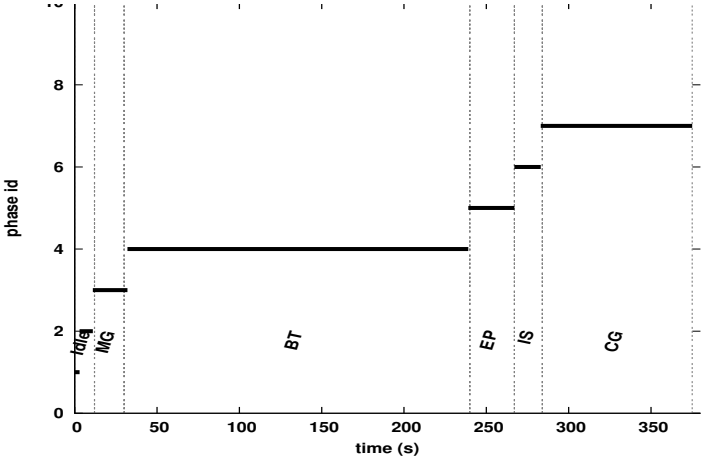
At the level of a node: Medium-grained

- React at medium latency at the level of the node
  - Change the processor frequency
  - Change the hard drive mode
  - Reconfigure the network card
- Detection of the current phase
- React in function of this profile
- Light impact on the infrastructure

Resource consumption of a complex application



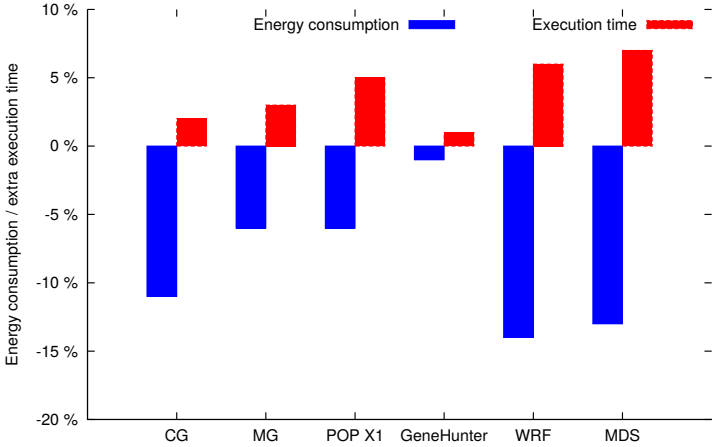
Phases where resource consumption are constant



Decision method

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
IO-intensive	switch on memory banks; scale the processor down; increase disks, increase disks (if needed)

Energy and performance, 28 node



Plan

Fine-grained = DVFS ?

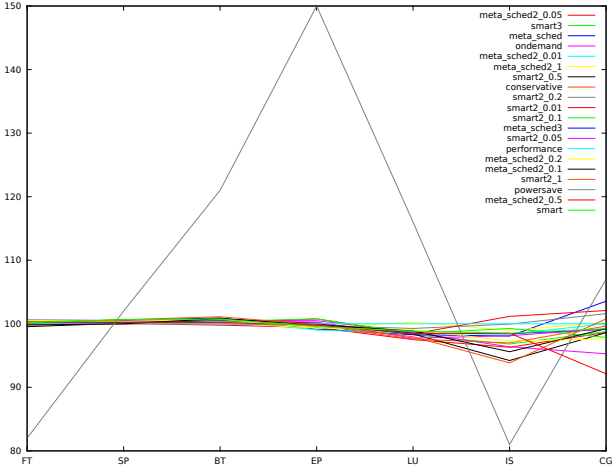
Relative values between *performance* and *ondemand* governors

Benchmark	FT	SP	BT	EP	LU	IS	CG
Time increase (%)	0	-3	-1	1	-2	2	0
Energy increase (%)	0	-3	-1	-1	-2	-1	-1

- HPC applications are rarely in Idle... Surprise !
- MPI libraries are spinning

Classical HPC benchmarks from NPB (Nas Parallel Benchmark)

DVFS = function of load

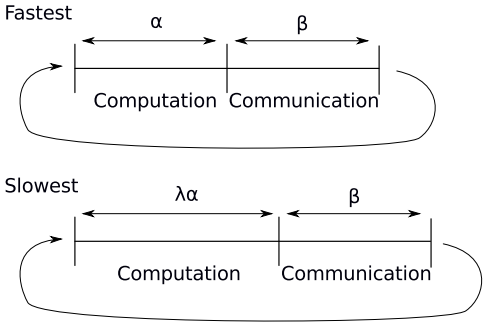


HPC Hypothesis

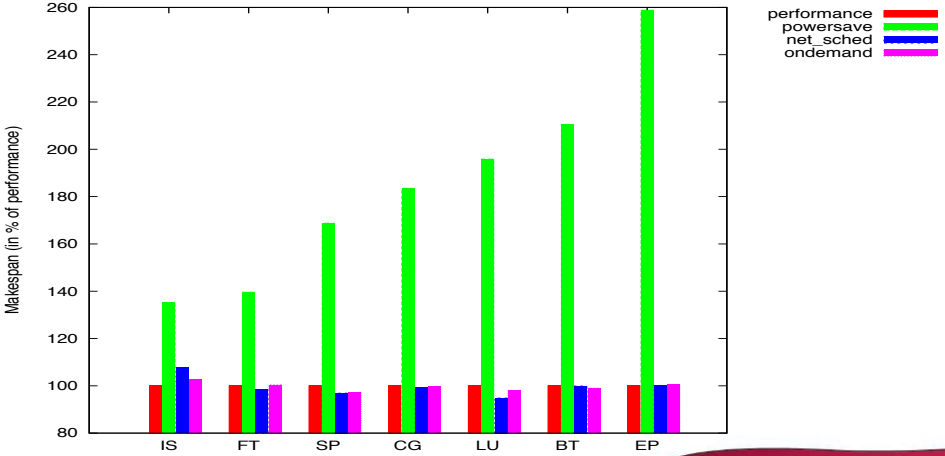
- State of applications
  - Computing
  - Communications
  - Disk I/O
  - Idle

HPC Hypothesis

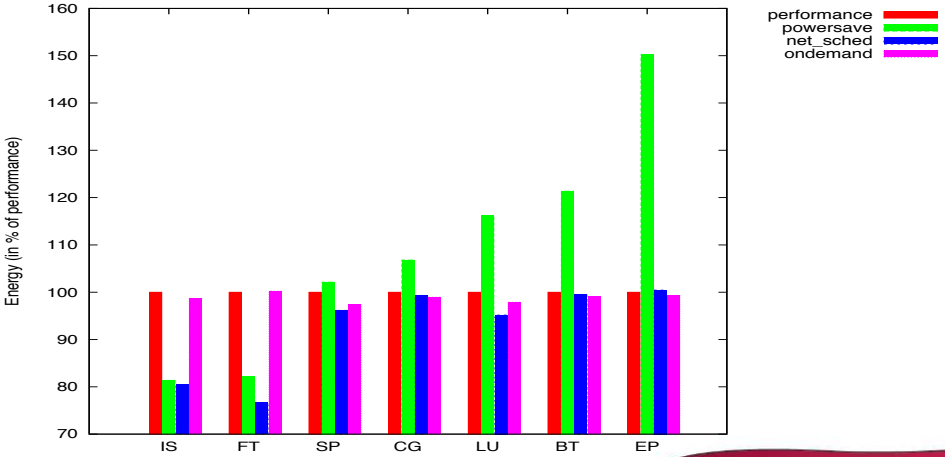
- State of applications
  - Computing
  - Communications



Results: Makespan



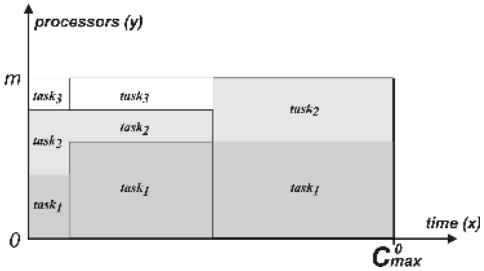
Results: Energy-to-solution



Plan

Malleable HPC applications: Energumen project

1. Interaction between batch scheduler and applications
2. Dynamic redimensioning of parallel jobs.
3. Optimize performance and energy
   - Speed-up models
   - Reconfiguration models



Open research questions

- Programming paradigms
  - Ability to describe parallelism intuitively
  - Remove the burden from developer
- Runtimes
  - Capability to adapt to particular profiles and their interactions
  - Ability to change kernels in function of context
- Communication between these two levels
- Improvement of RJMS (Resources and Job Management Systems)
  - Spatial management
  - Temporal management