Energy Efficient HPC systems

Architect of an Open World™

EnA-HPC– Sept. 2013

Jean-Pierre Panziera
Chief Technology Director

Bull: from Supercomputers to Cloud Computing

Expertise & services

- HPC Systems Architecture
- Applications & Performance
- Energy Efficiency
- Data Management
- HPC Cloud

extreme factory
stay lean: compute smart

center for excellence in parallel programming

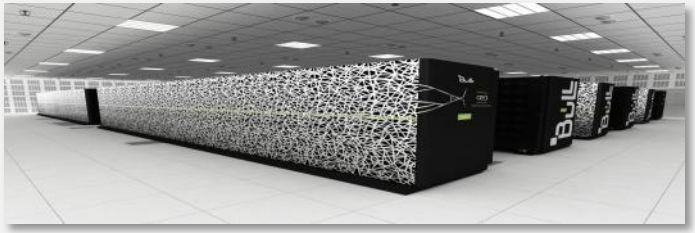
Software

- Open, scalable, reliable SW
- Development Environment
- Linux, OpenMPI, Lustre, Slurm
- Administration & monitoring

bullx supercomputer suite

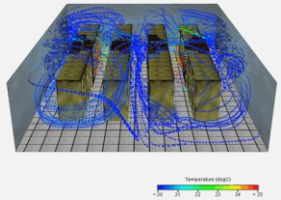
Servers

- Full range development from ASICs to boards, blades, racks
- Support for accelerators



Infrastructure

- Data Center design
- Mobile Data Center
- Water-Cooling



Bull Architect of an Open World
ServiWare A Bull Group Company
sc science + computing A Bull Group Company

Leading HPC technology with Bull

TERA100 – 2010

1st European PetaFlop-scale System

Rank #6 TOP500 SUPERCOMPUTER SITES

CURIE – 2011

1st PRACE PetaFlop-scale System

Rank #9 TOP500 SUPERCOMPUTER SITES

BEAUFIX – 2013

1st Intel Xeon E5-2600 v2 System

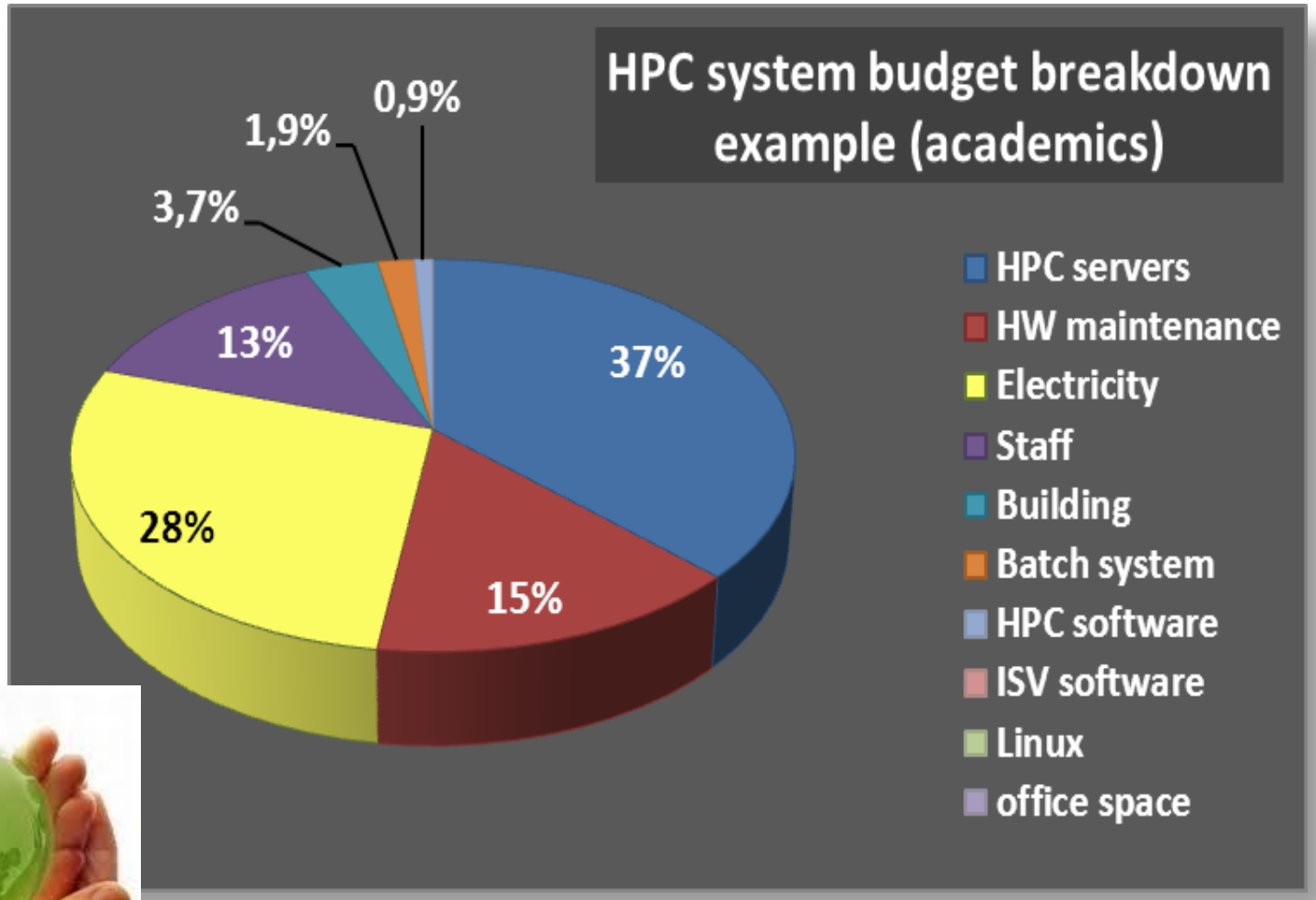
Direct Liquid Cooling Technology

DE LA RECHERCHE À L'INDUSTRIE
cea

GENCI
GRAND EQUIPEMENT NATIONAL POUR LE CALCUL INTENSIF

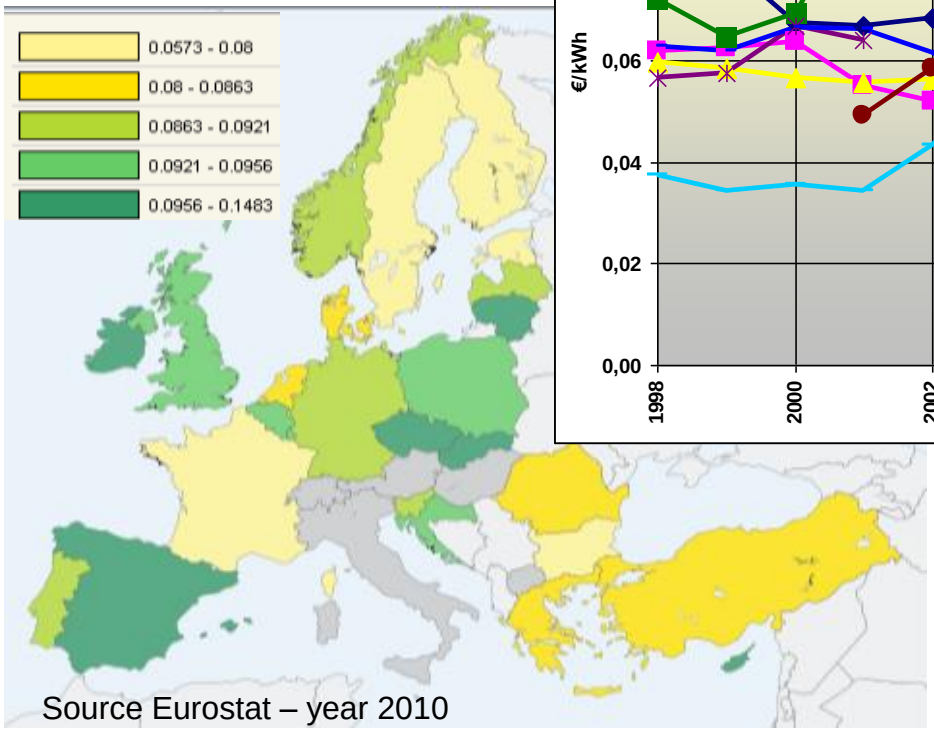
METEO FRANCE
Toujours un temps d'avance

Energy (Electricity): a significant part of HPC budget

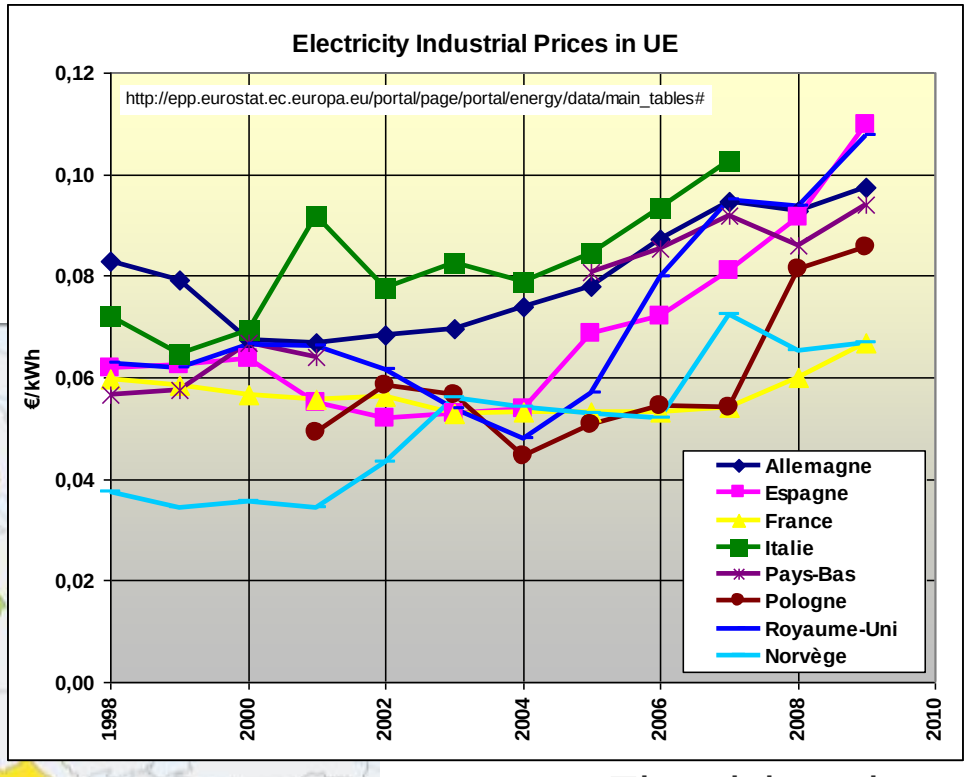


Industrial Electricity Prices in Europe

Electricity prices highly variable across Europe
Avg 0.11€/kWh

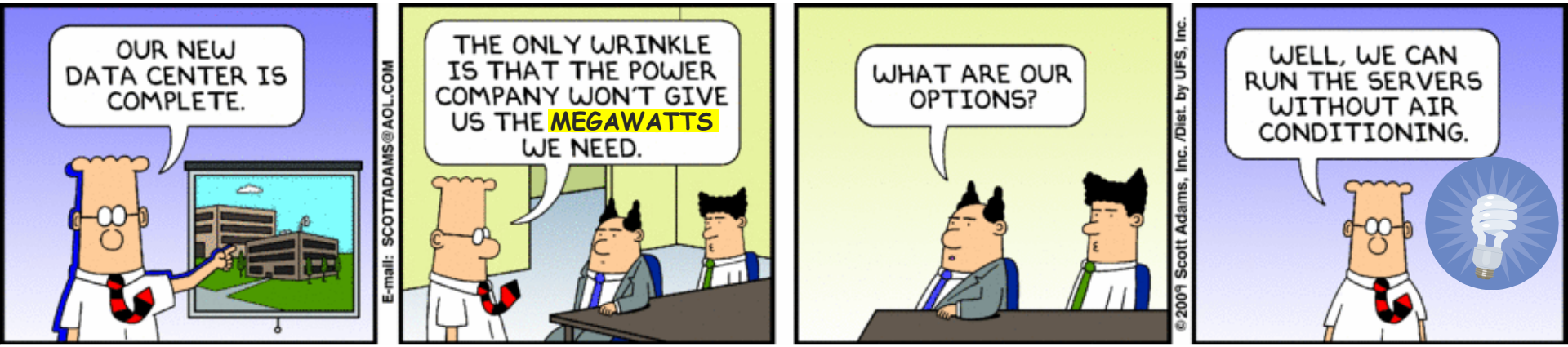


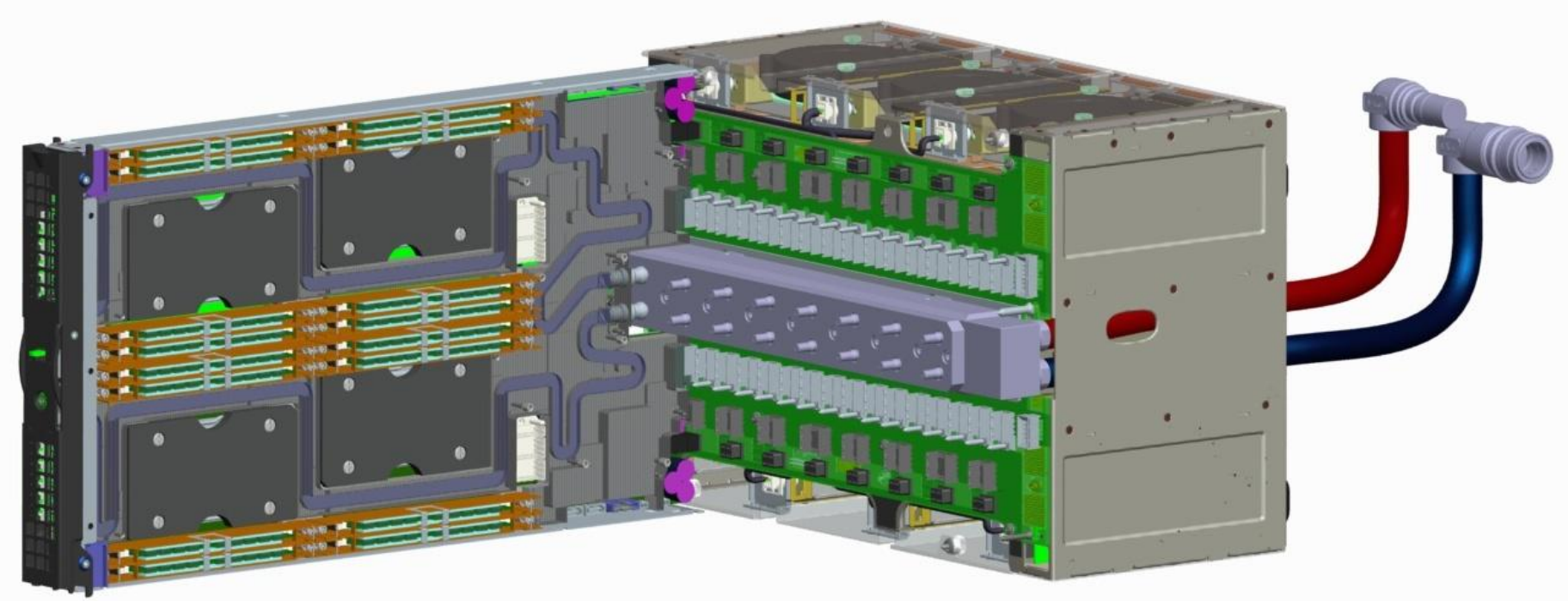
Source Eurostat – year 2010



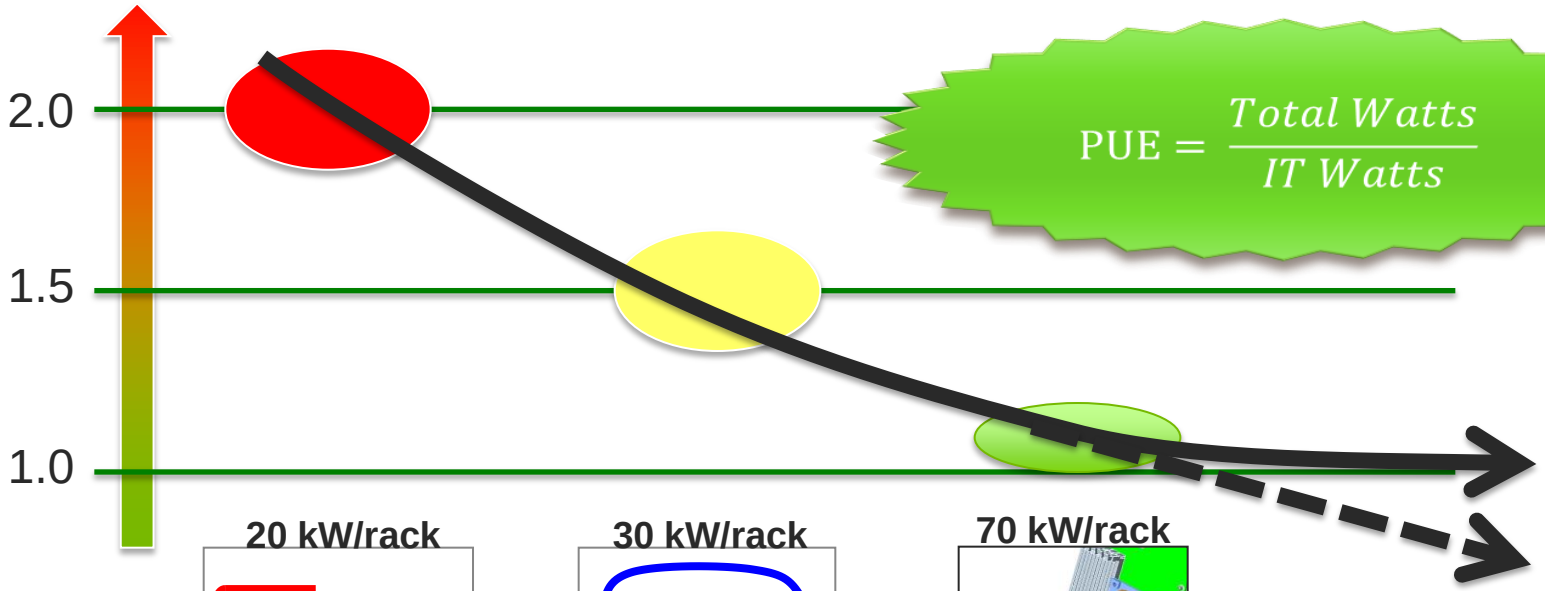
Electricity prices
Rising steadily
CAGR 12%

Power to the datacenter

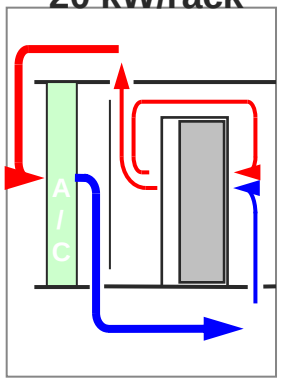




Cooling & Power Usage Effectiveness (PUE)

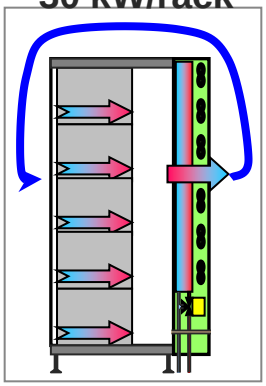


20 kW/rack



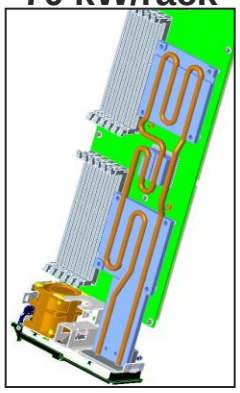
Air-cooled

30 kW/rack



Water-cooled door

70 kW/rack

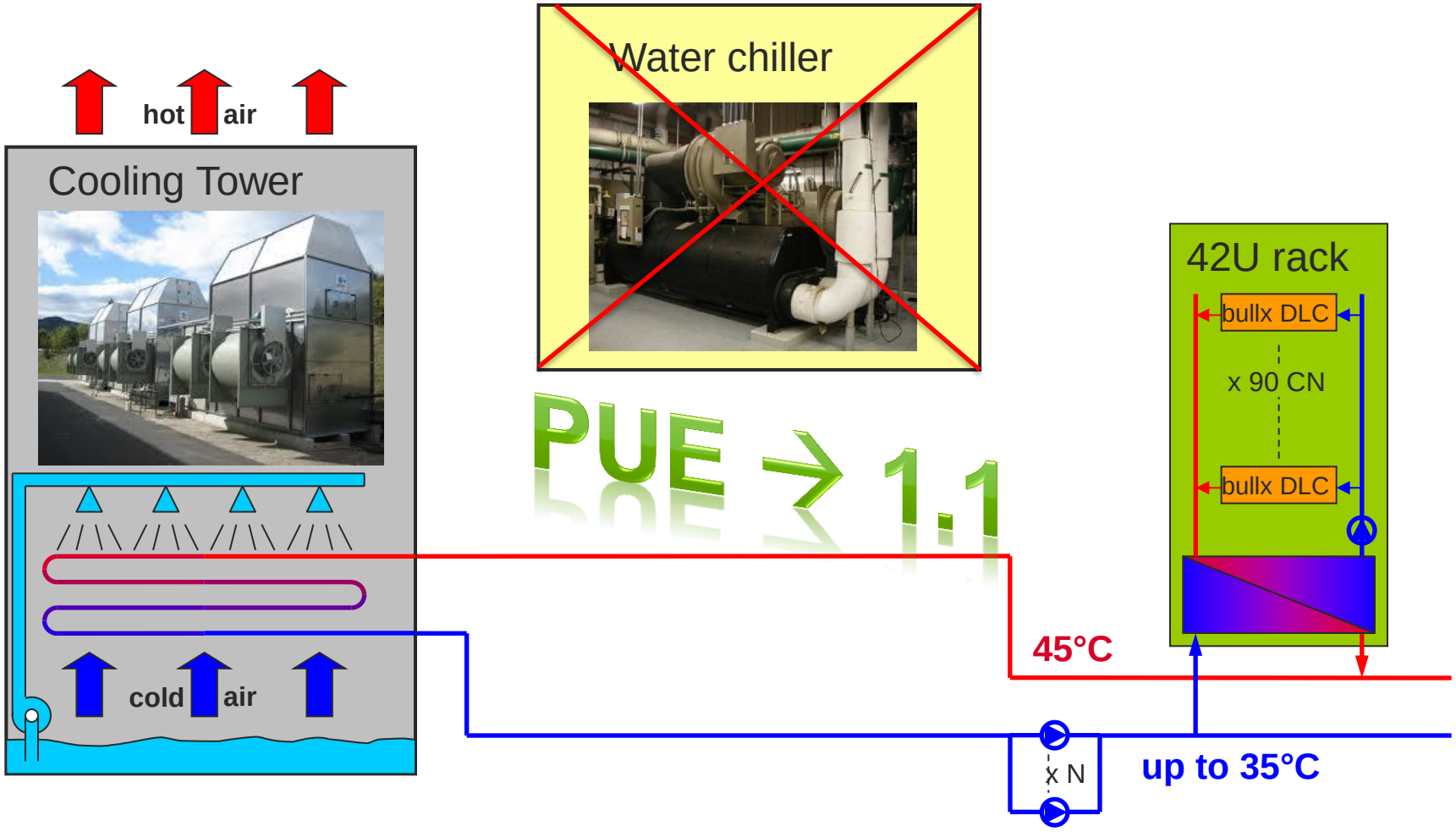


Direct-Liquid-cooling

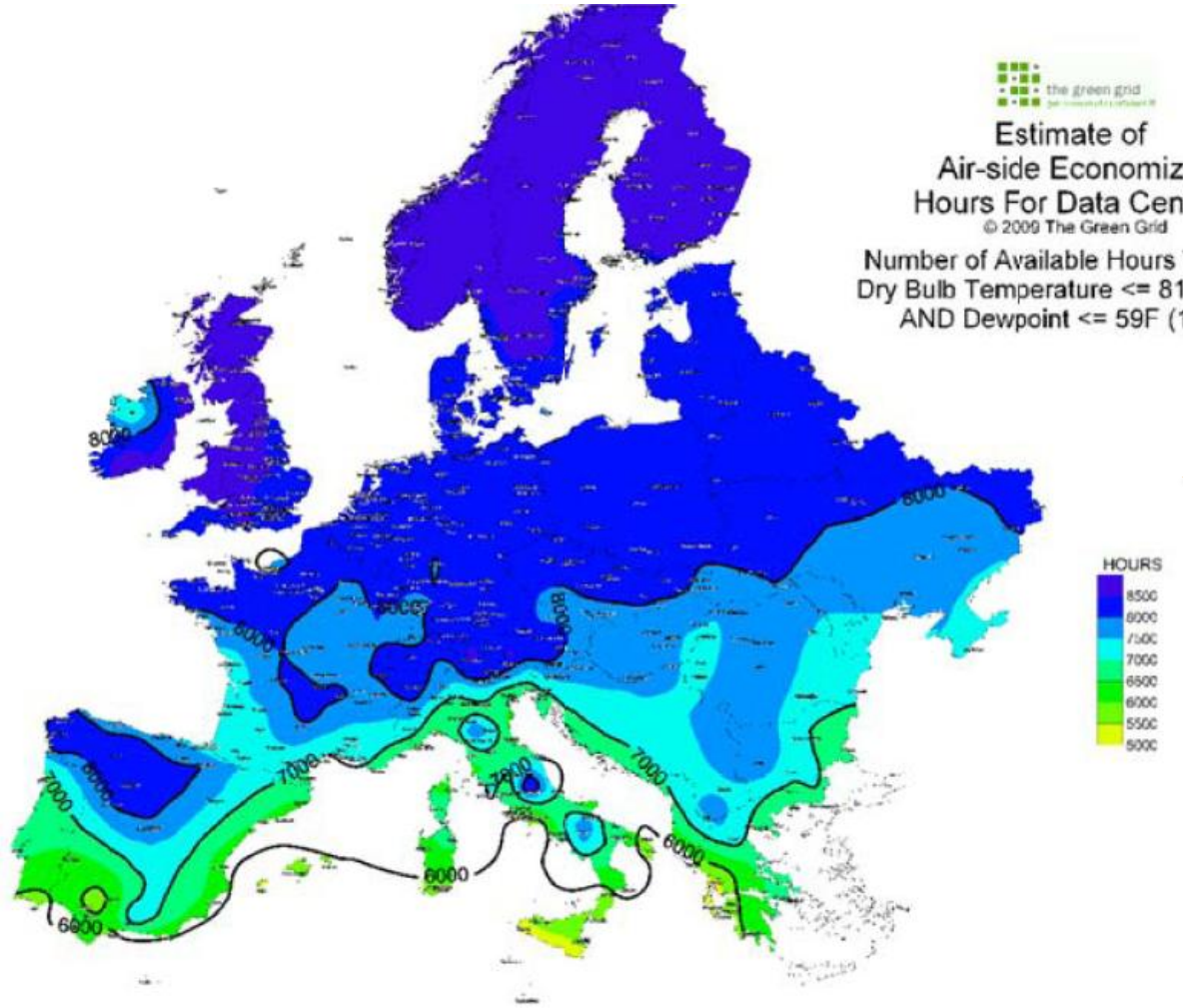
Co-generation

Direct Liquid Cooling Infrastructure

- With hot water cooled servers, water chillers are not required anymore



Fresh Air for (almost) free-cooling



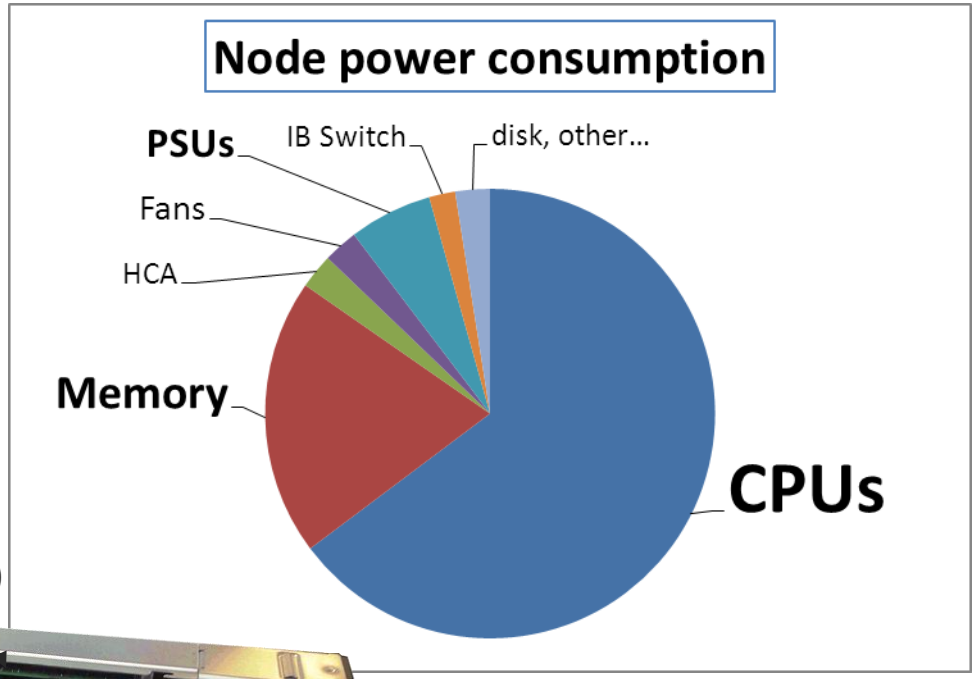
When system density does not matter

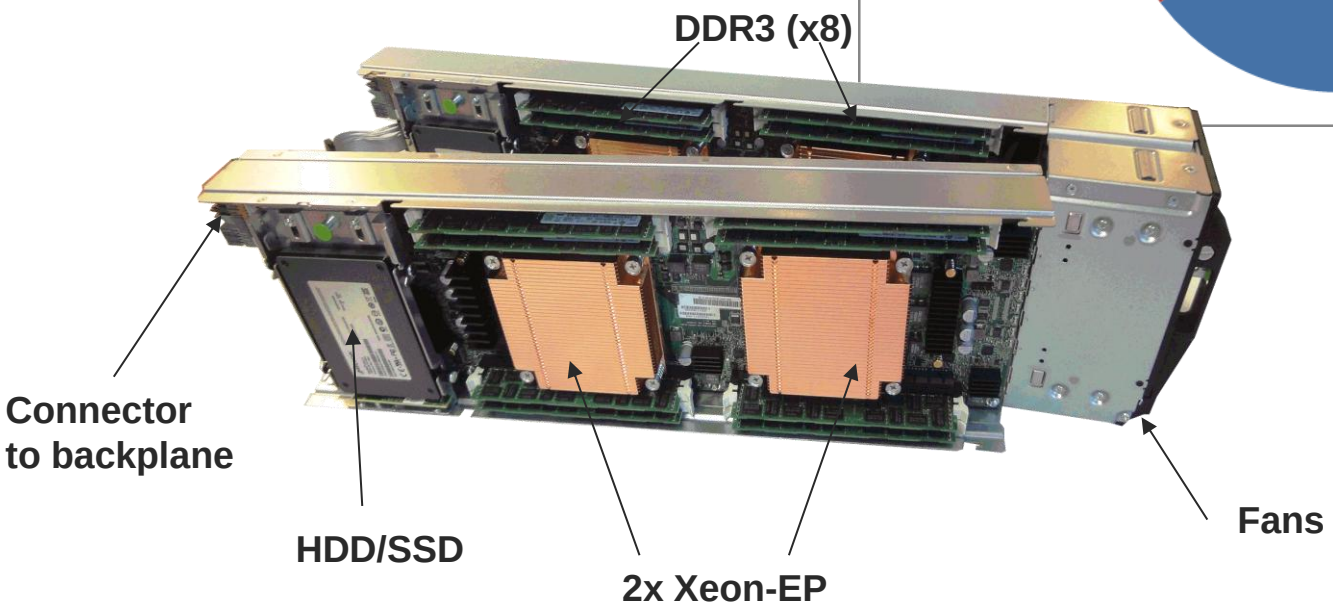
Plenty of cool air available

8760 hours/year … 8000h/y is > 90% of time

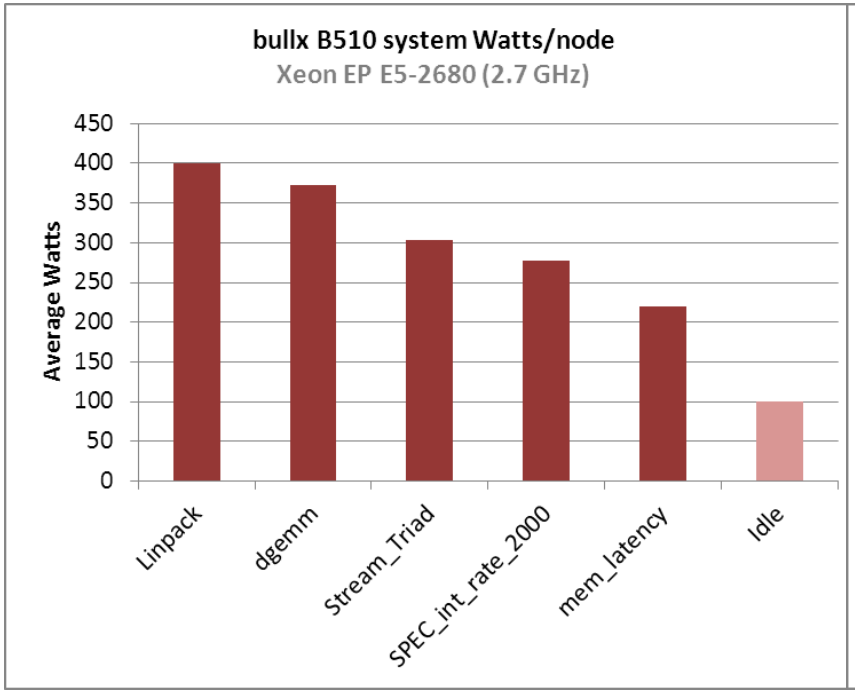
Let's have another look at air-cooling

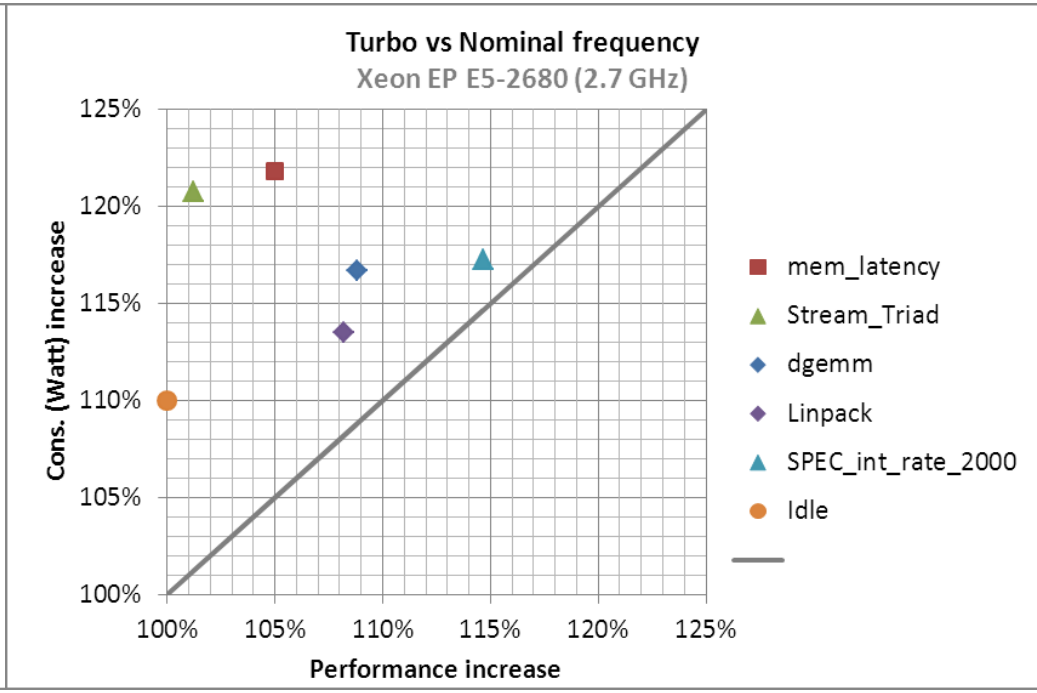
Where do all these Watts go ?





Node consumption varies with workload

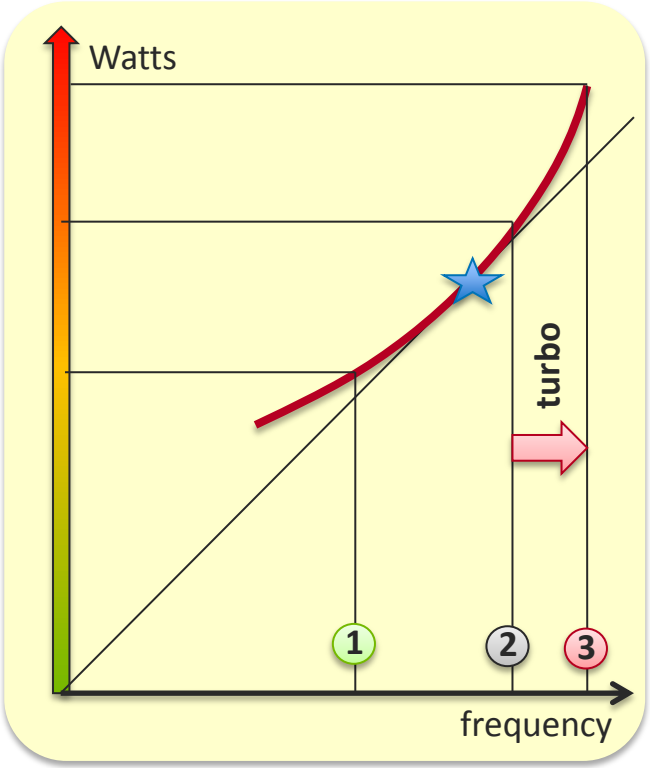


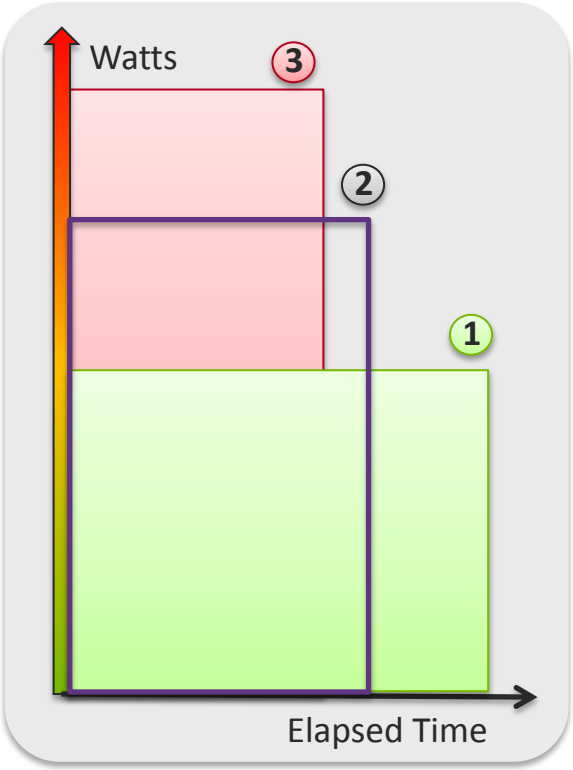


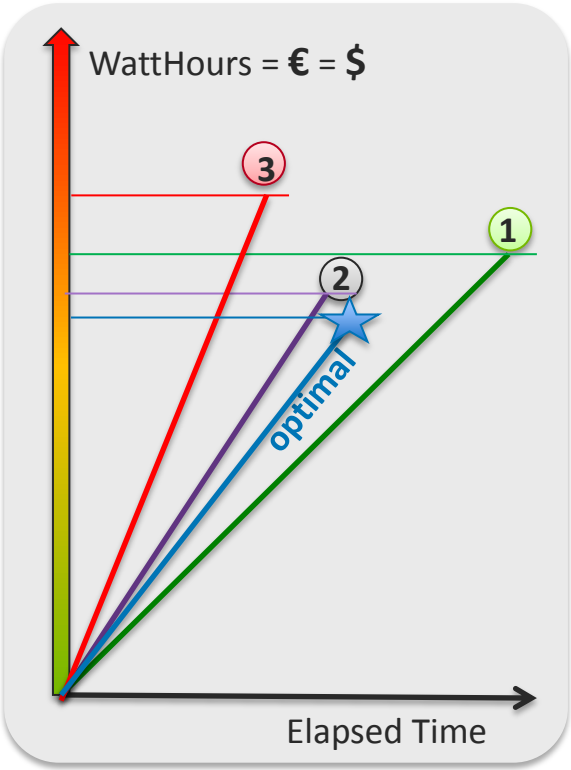
Wrt Linpack (max -> 100%)
Memory streaming 75%
Irregular memory access 55%
Iddle 25%

Using turbo is never energy efficient

CPU frequency vs System energy consumption

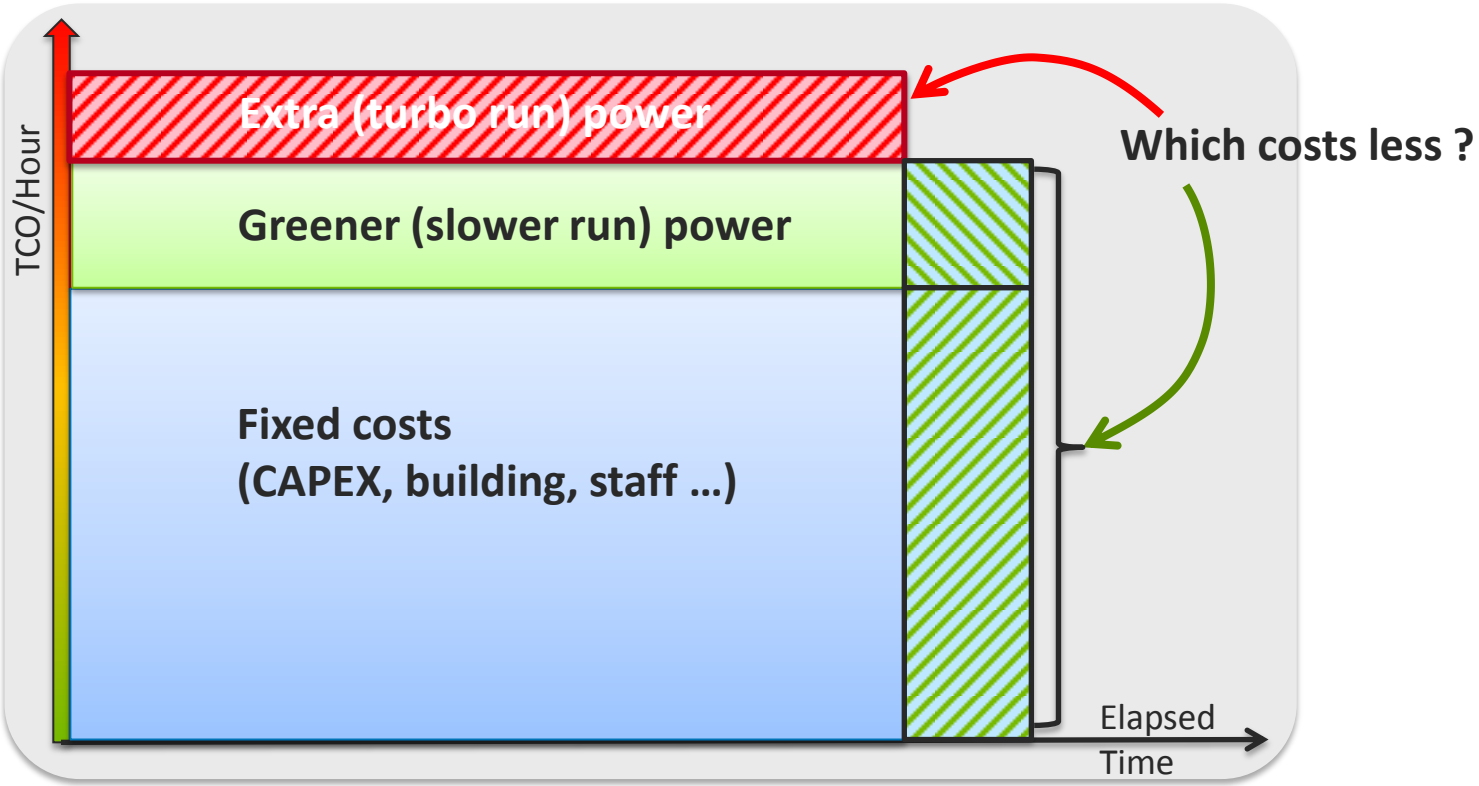






- The CPU frequency is the HPC system throttle
- The faster the CPU, the more power
- The slower the CPU, the less power
- Minimal energy consumption is achieved for intermediate (<nominal) frequency

Total Cost of Ownership (TCO): the CFO view



- Energy (electricity cost) is only a portion (25-30%) of the TCO
- When taking into account fixed expenses … slower runs are more expensive
- Greener might not mean Cheaper TCO

Global optimization for Parallel HPC Applications

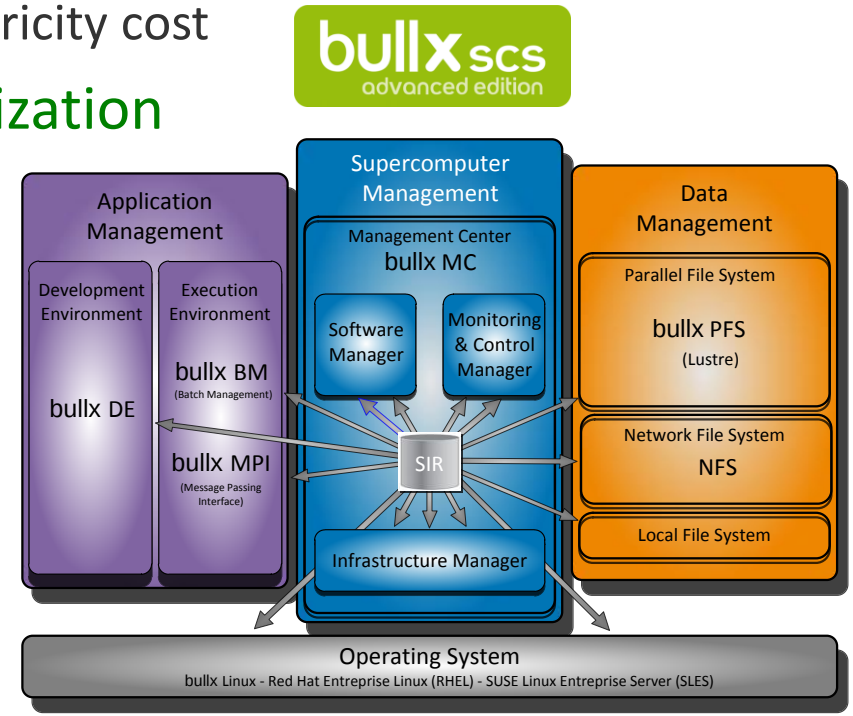
Adding many more parameters to the equation:

- HPC applications == highly parallel
- Interference with other jobs on system creates variability
  - Job placement
  - Interconnect
  - Storage
- Good load balancing is hard to achieve
  - Everyone waiting for the slower thread, let's speed it up.
  - →non uniform parameters for different tasks/nodes.
- Complex optimization requires detailed understand / precise measurements

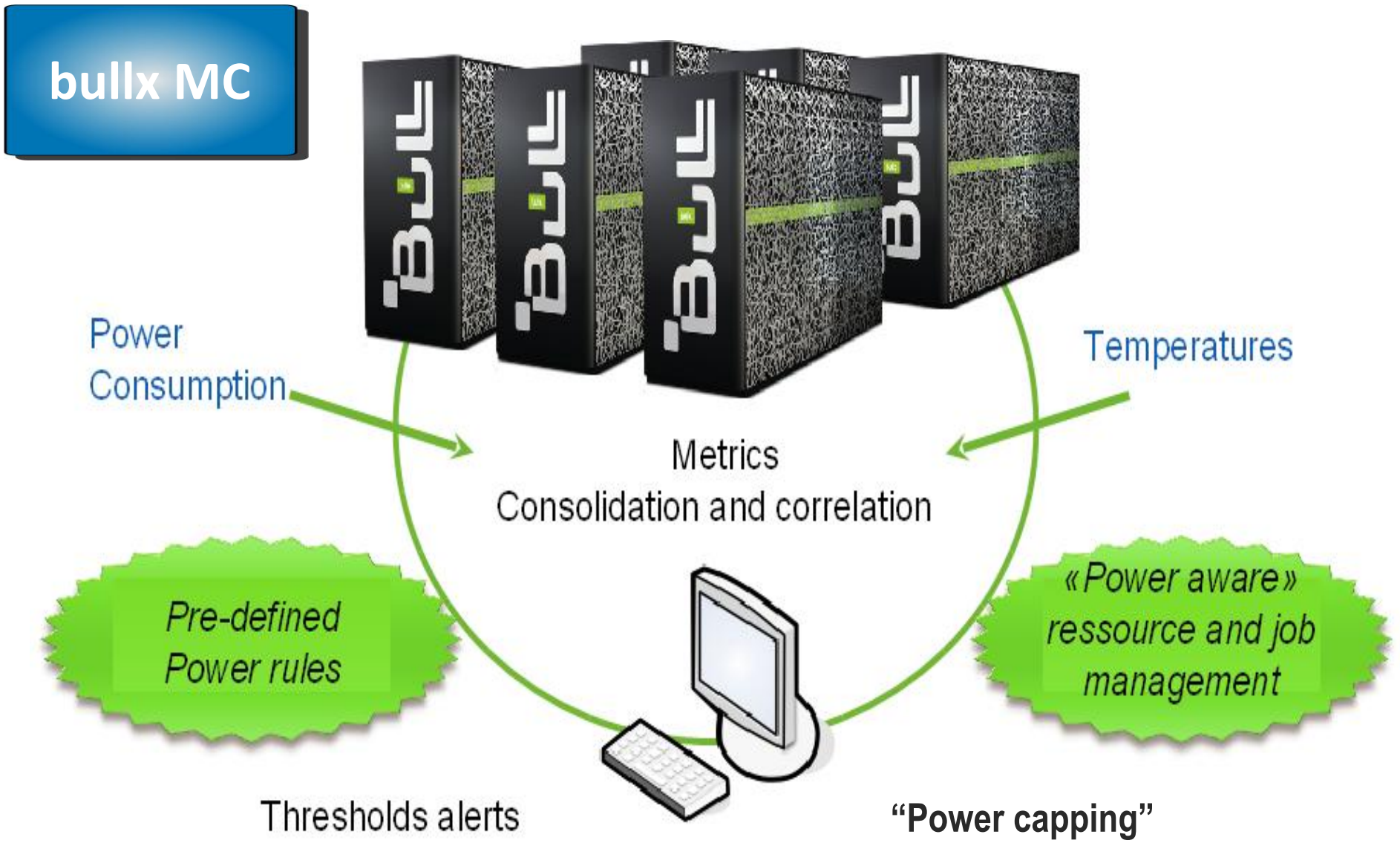
Power Management

- Accounting
  - Users billed separately for CPU, IO, … and Energy
  - Keep compute center electricity bill within budget
- Control power
  - Avoid running over capacity
  - Allow for priority jobs
  - Adjust power consumption with electricity cost
- Energy consumption / cost optimization
  - Fine & precise power monitoring
  - Power data analysis
  - Control all system resources power

… enter software



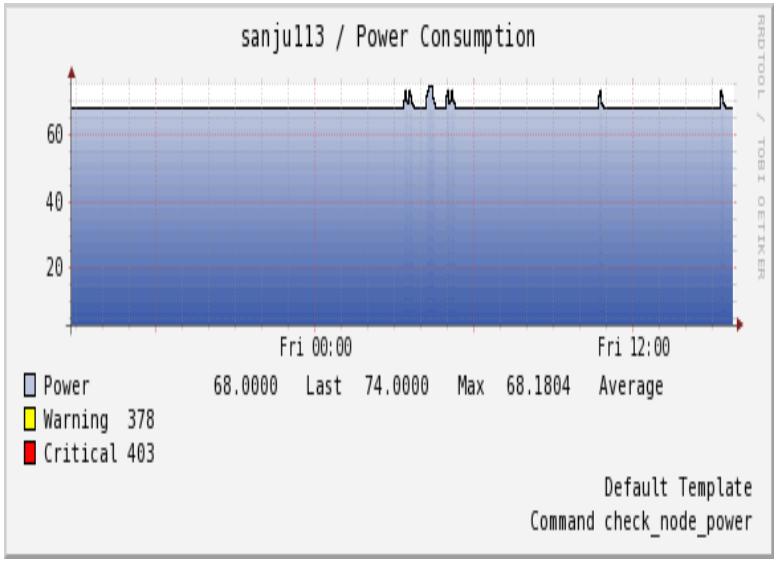
Power Control scenario



Bullx MC Power Manager

Monitoring

- All HW with available power sensors
- Consolidation every 10 mn
- Store info in database
- Graphical web interface
- Out-of-band queries



Power capping

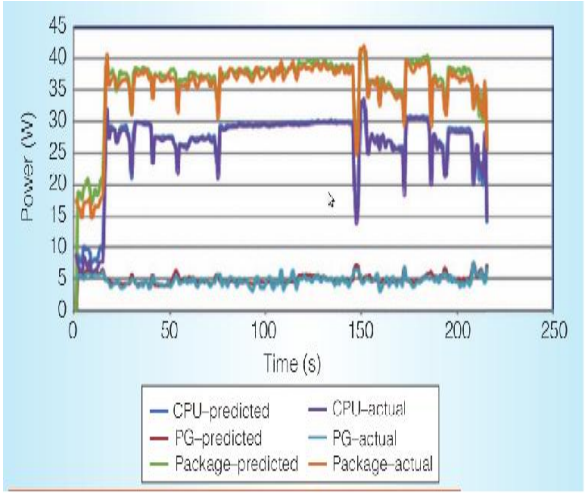
- Automatic action to decrease power level
- Automatic information for system monitoring
- Open framework, based on SEC (Simple Event Correlator)
- Allow new rules creation
- But slow reaction time (minutes)

What do consume your applications?

bullx BM

slurm workload manager

- Pluggins: RAPL, IPMI (OS) and RRD
- Per job (global value & time slice)
- Per node
- Per user
- New srun parameter to allow CPU frequency scaling for job execution



Bull TU Dresden high frequency monitoring

HARDWARE

- Regular B700 blades + innovative power mesurement tools

SOFTWARE

- API (Opensource)

PROJECT

- Project Management
- IP Management
- Contract Management

MIDDLEWARE

- New modules in VAMPIR
- Scalable High Definition Power Monitoring API

APPLICATION

- Development of new optimization methodologies
- Demonstration of energy efficiency improvement

OPENSOURCE

- ✓ Energy efficient operation
  - ✓ CPU states
  - ✓ Turn off devices
  - ✓ Interface with batch scheduler
  - ✓ Mesurement environment
  - ✓ Vampir integration
- ✓ Energy accounting
- ✓ FPGA integration
- ✓ Measuring system Accuracy
- ✓ Energy Efficiency research at application level

Bull TU Dresden high frequency monitoring

Admin (2)

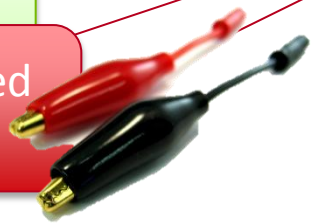
IB Fabric / GigE Fabric

Login (2)

Export (2)

SMP (2)

Instrumented blades

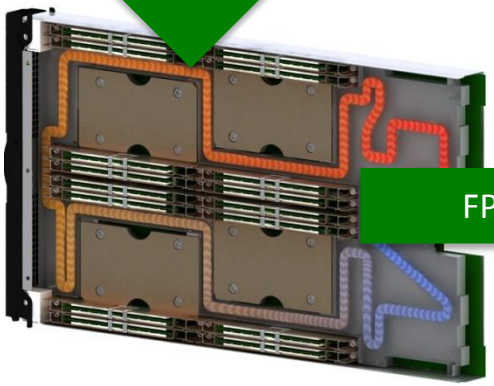


FASS – Phase 1

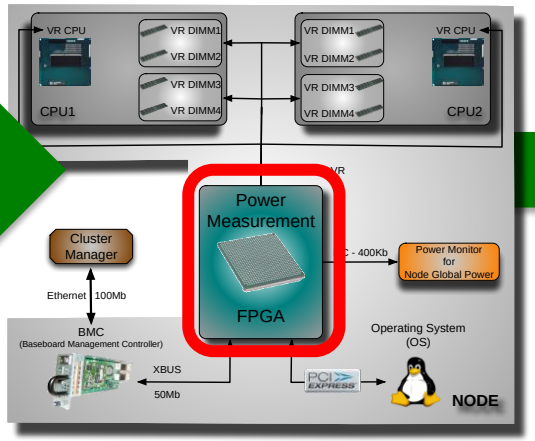
GPU Cluster

SATA / SAS

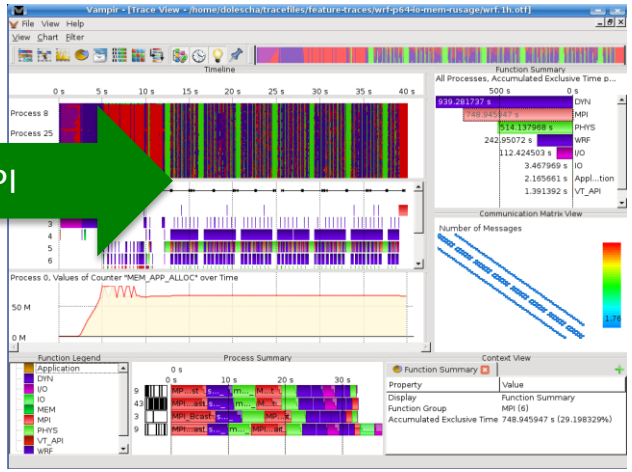
32 Nodes
48x Kepler



FPGA



API



Freq: 500Hz - Error: 2%

Energy efficient HPC systems …

Green systems

- Interest driven by energy cost and green attitude
- Green systems start with Green components (CPUs, Memory, PSUs, Interconnect…)
- Free-Cooling either with Liquid or fresh-air (save on CAPEX & OPEX)
- Optimize runtime parameters for best overall system performance (incl. Power)

Power Monitoring

- Non-intrusive power monitoring at low frequency (seconds, minutes)
- Accounting – Energy billing separately from CPU time
- Fine grain monitoring (seconds) possible but slightly intrusive (RAPL and OS IMPI)
- For high rate power sampling, HW instrumentation required
- Complete power management framework is still under development

Architect of an Open World™