Power Management for Energy Efficient HPC systems

Architect of an Open World™

Teratec – June 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

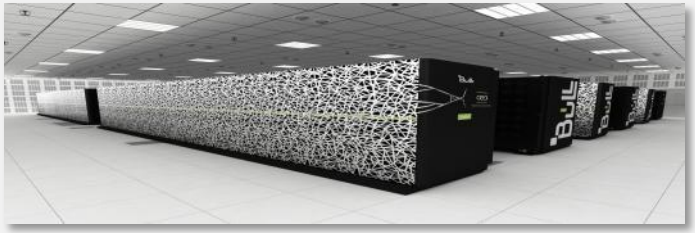
Software

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

bullx supercomputer suite

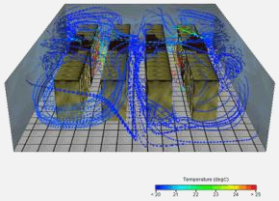
Servers

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



Infrastructure

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



Leading HPC technology with Bull

TERA100 – 2010

1st European PetaFlop-scale System

Rank #6

CURIE – 2011

1st PRACE PetaFlop-scale System

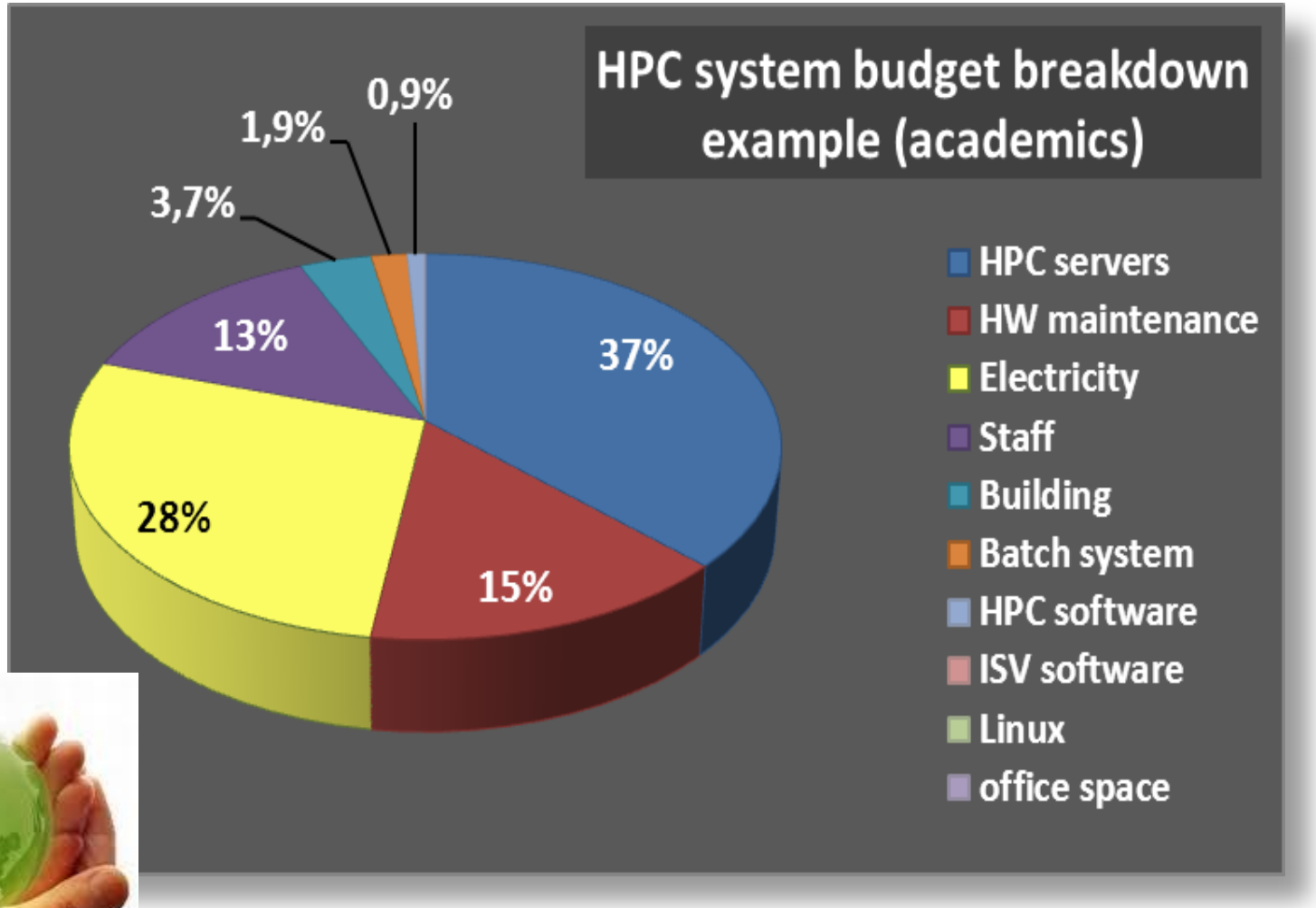
Rank #9

BEAUFIX – 2013

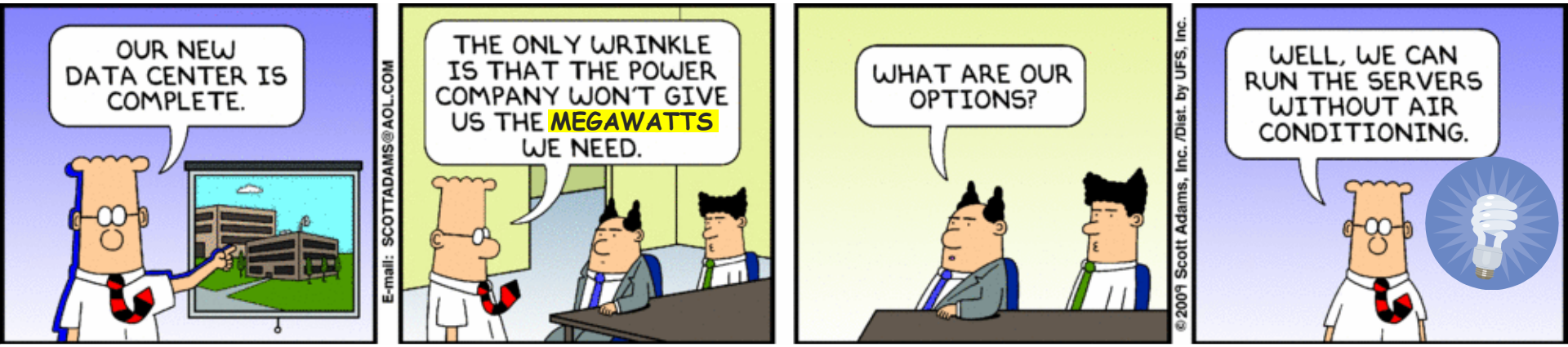
1st Intel Xeon E5-2600 v2 System

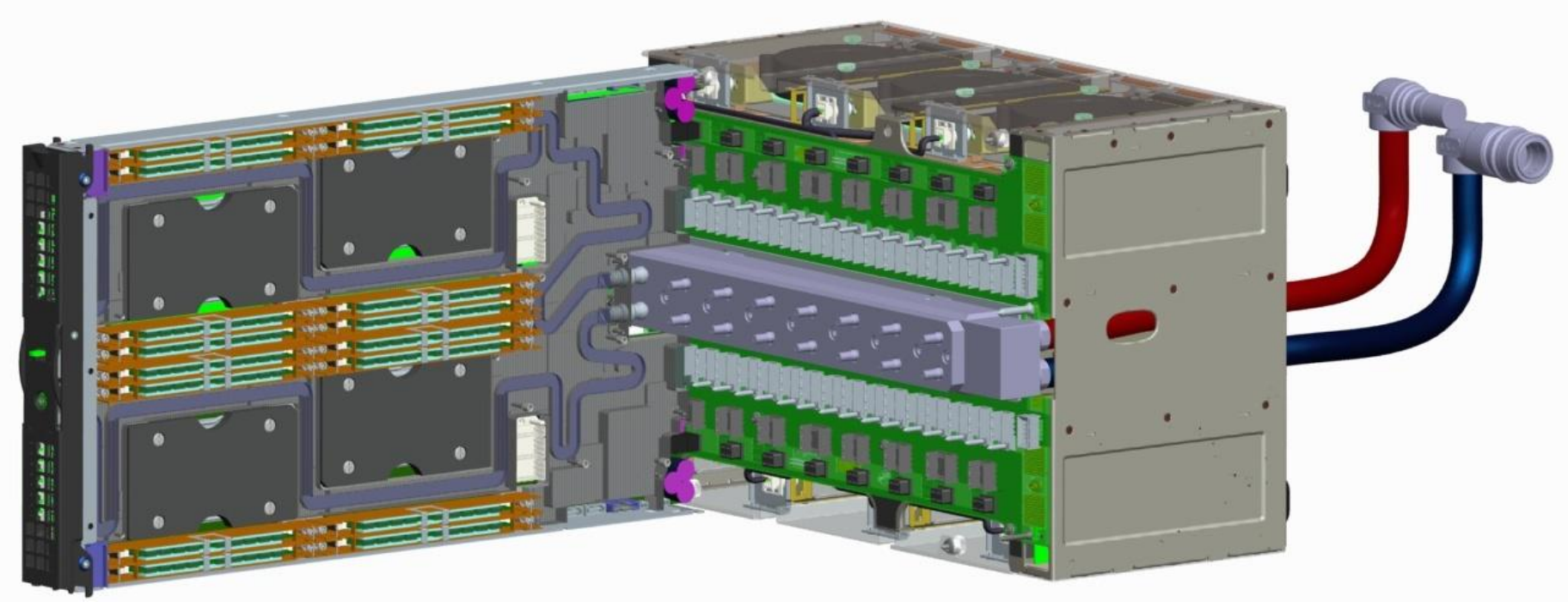
Direct Liquid Cooling Technology

Electricity, a significant part of HPC budget

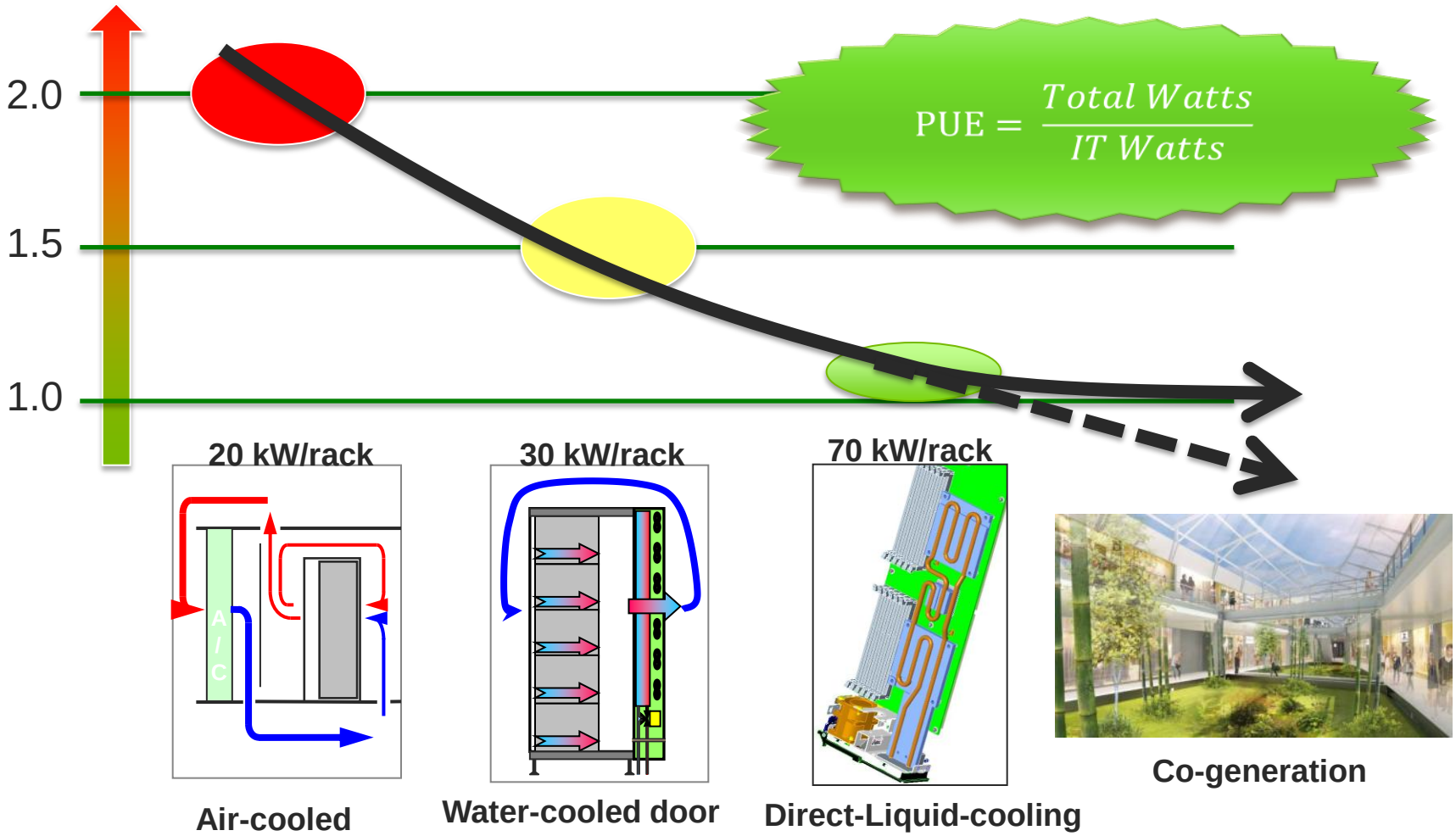


Power to the datacenter



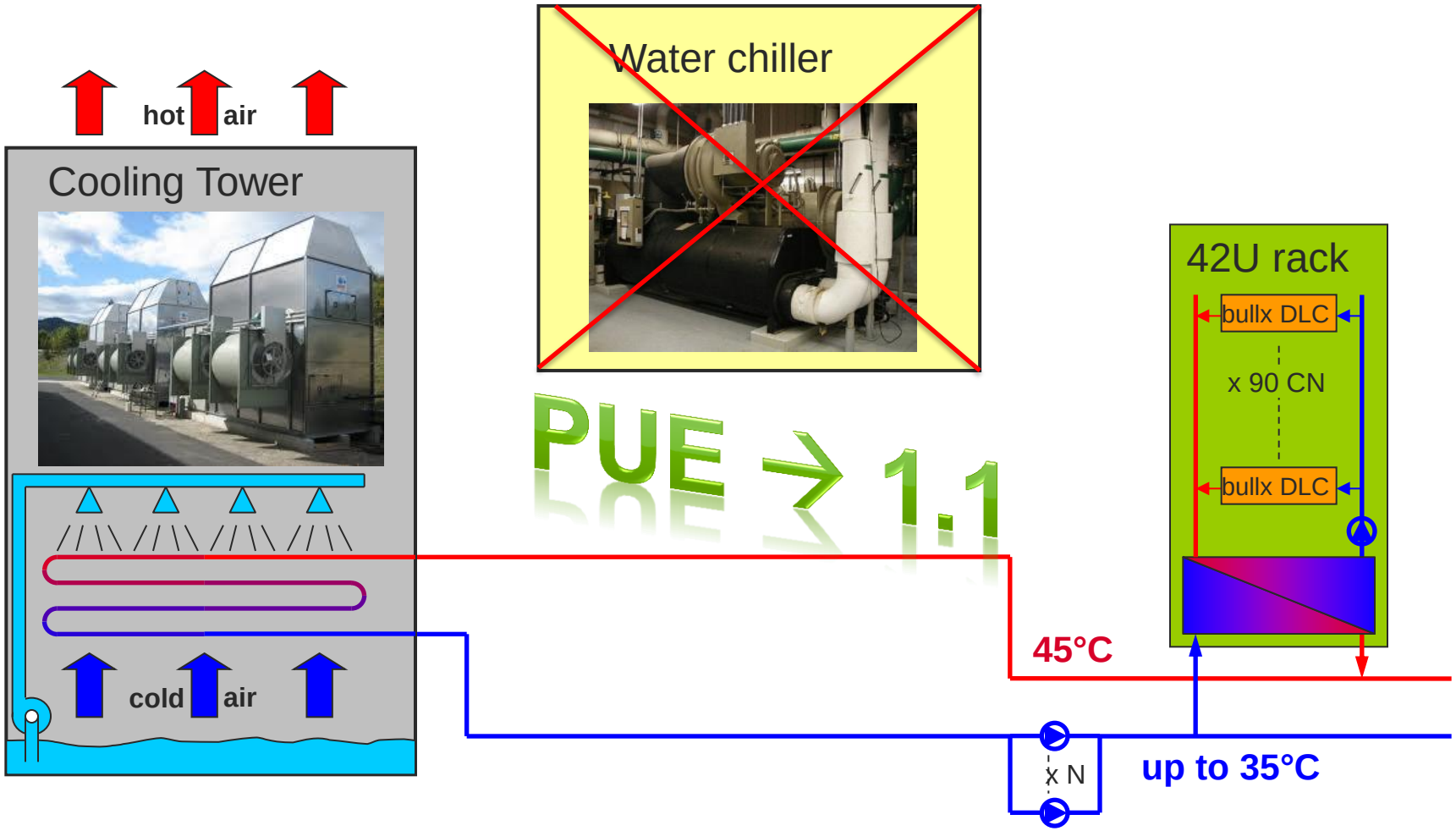


Cooling & Power Usage Effectiveness (PUE)

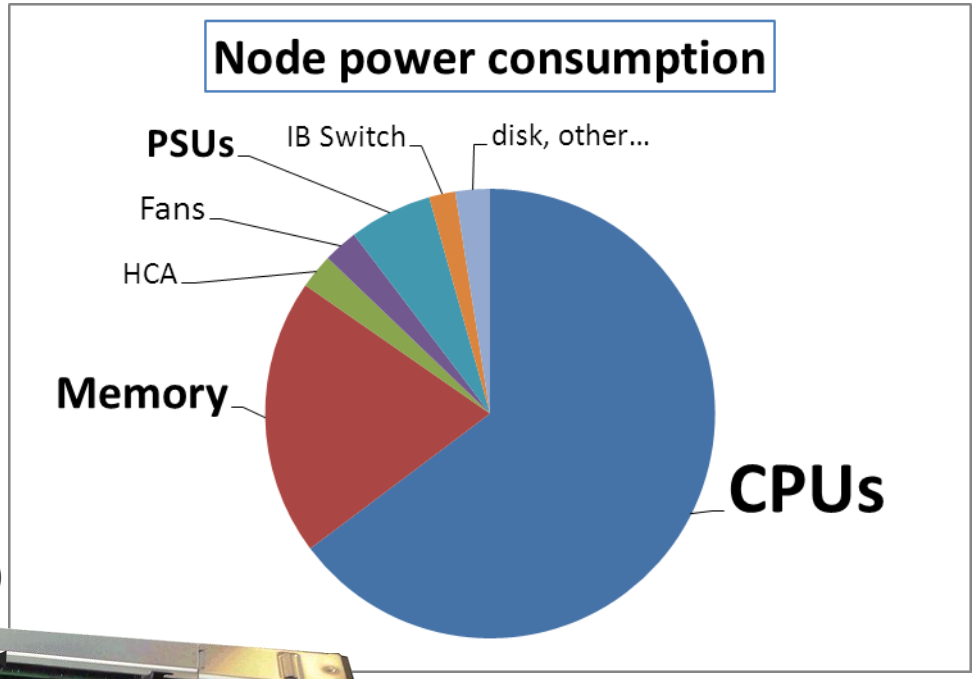


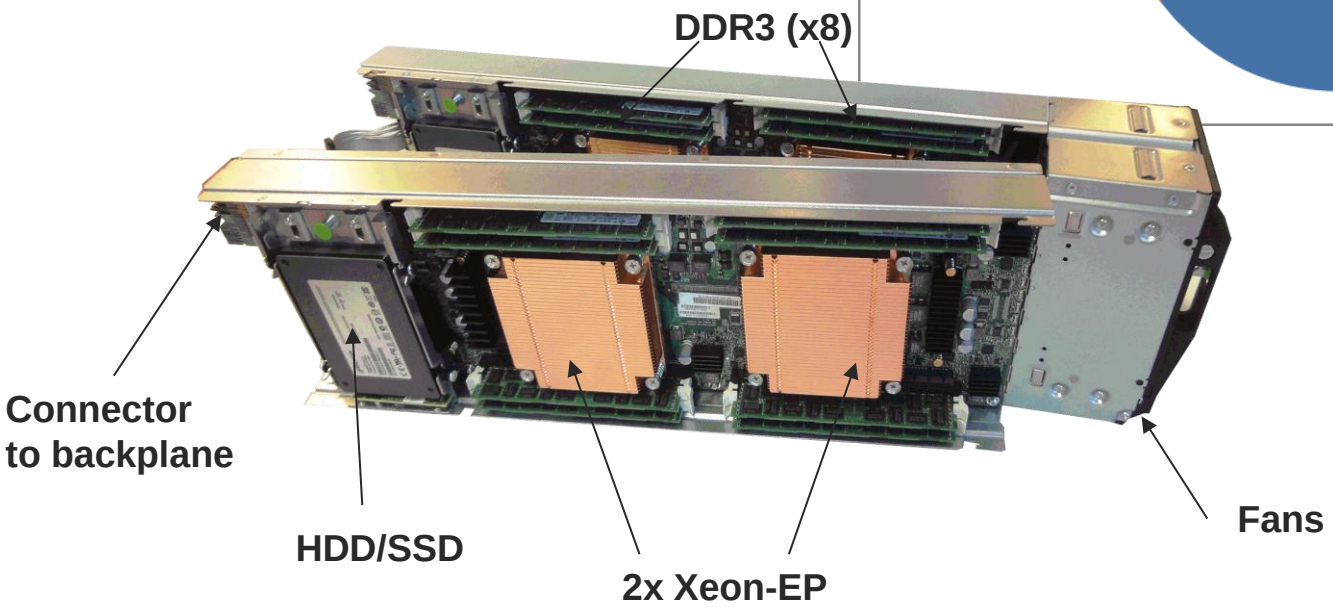
Direct Liquid Cooling Infrastructure

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

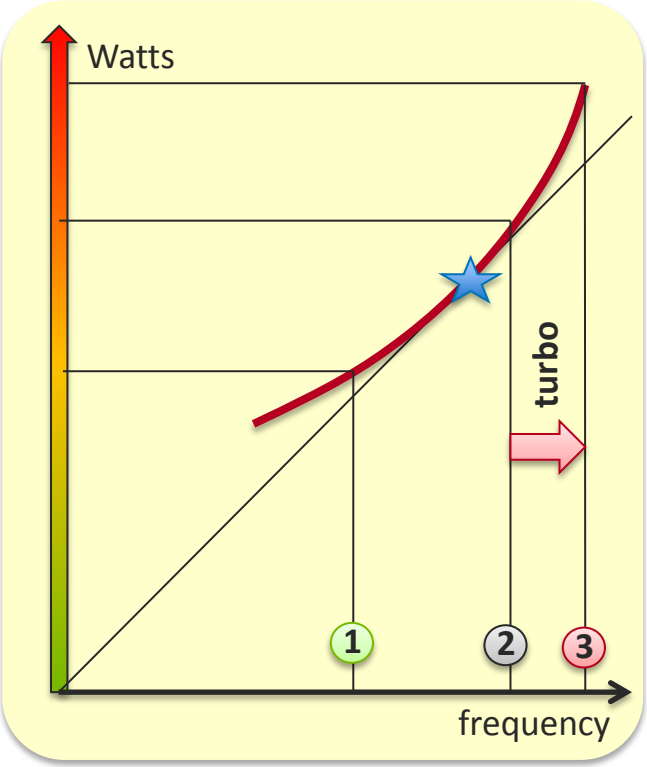


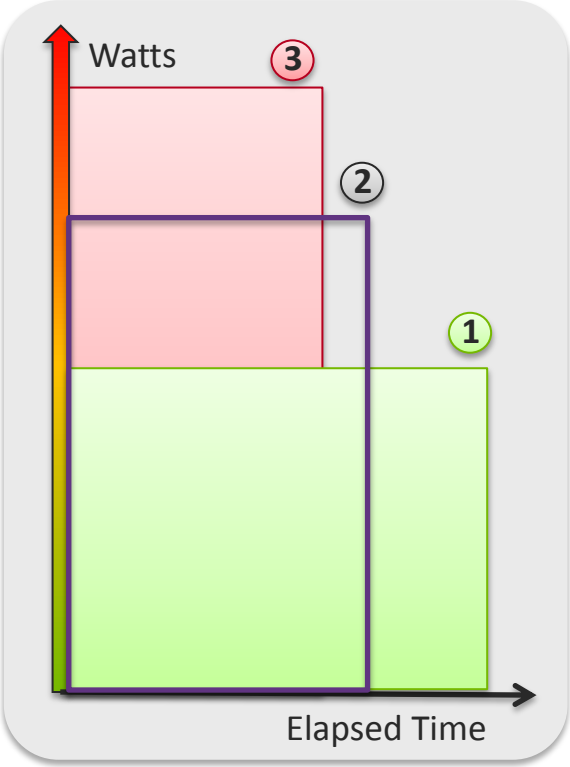
Where do all these Watts go ?

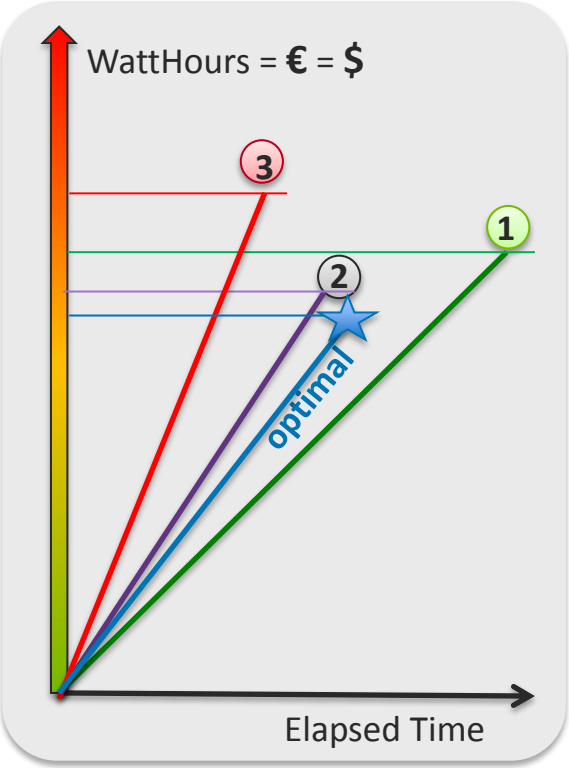




CPU frequency vs System energy consumption

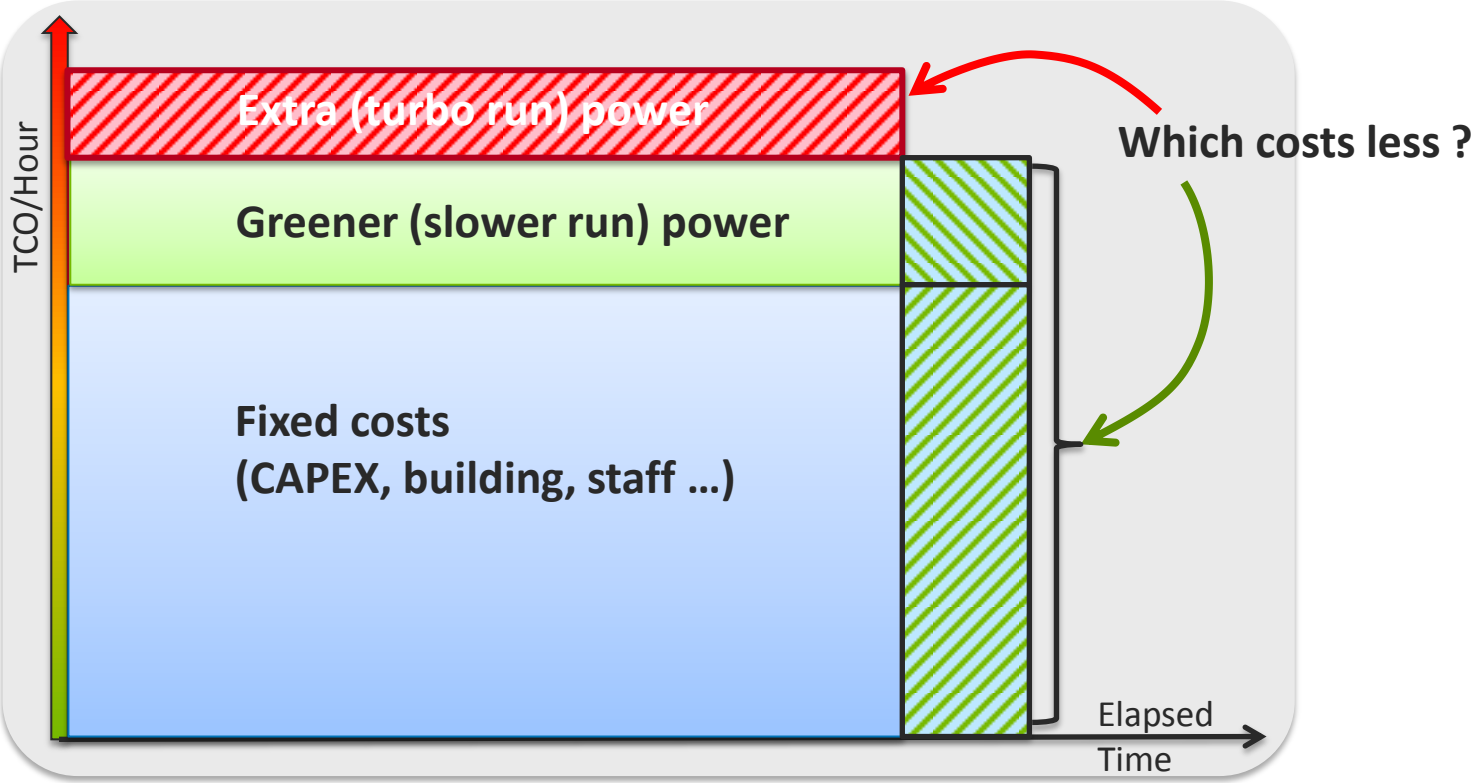






- The faster the CPU, the more power
- The slower the CPU, the less power
- But minimal energy consumption is achieved for intermediate (lower than nominal) frequency

Total Cost of Ownership (TCO): the CFO view

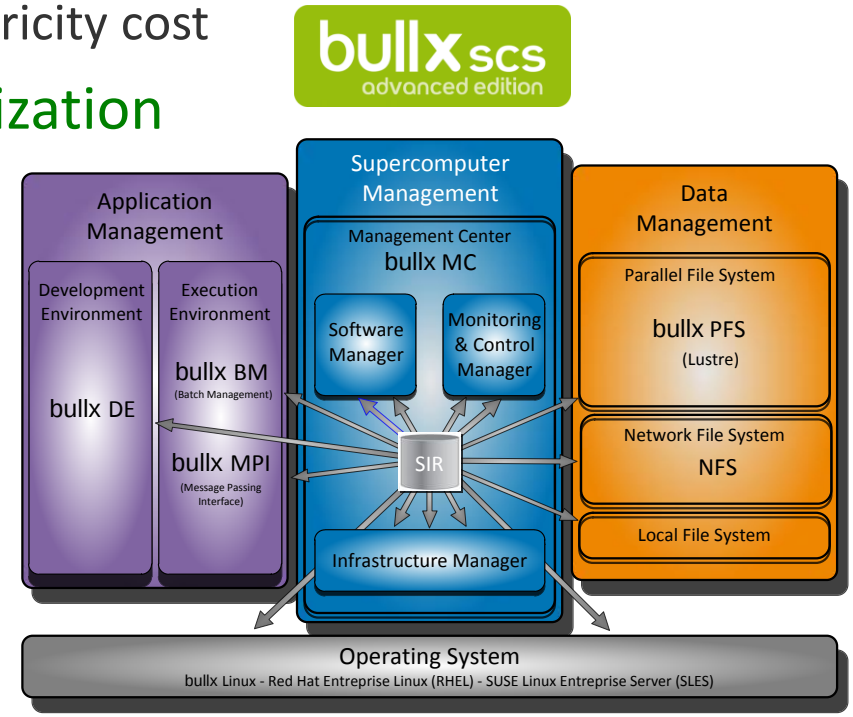


- Energy (electricity cost) is only a portion (25-30%) of the TCO
- When taking into account fixed expenses … slower runs are more expensive
- Greener might not mean Cheaper TCO

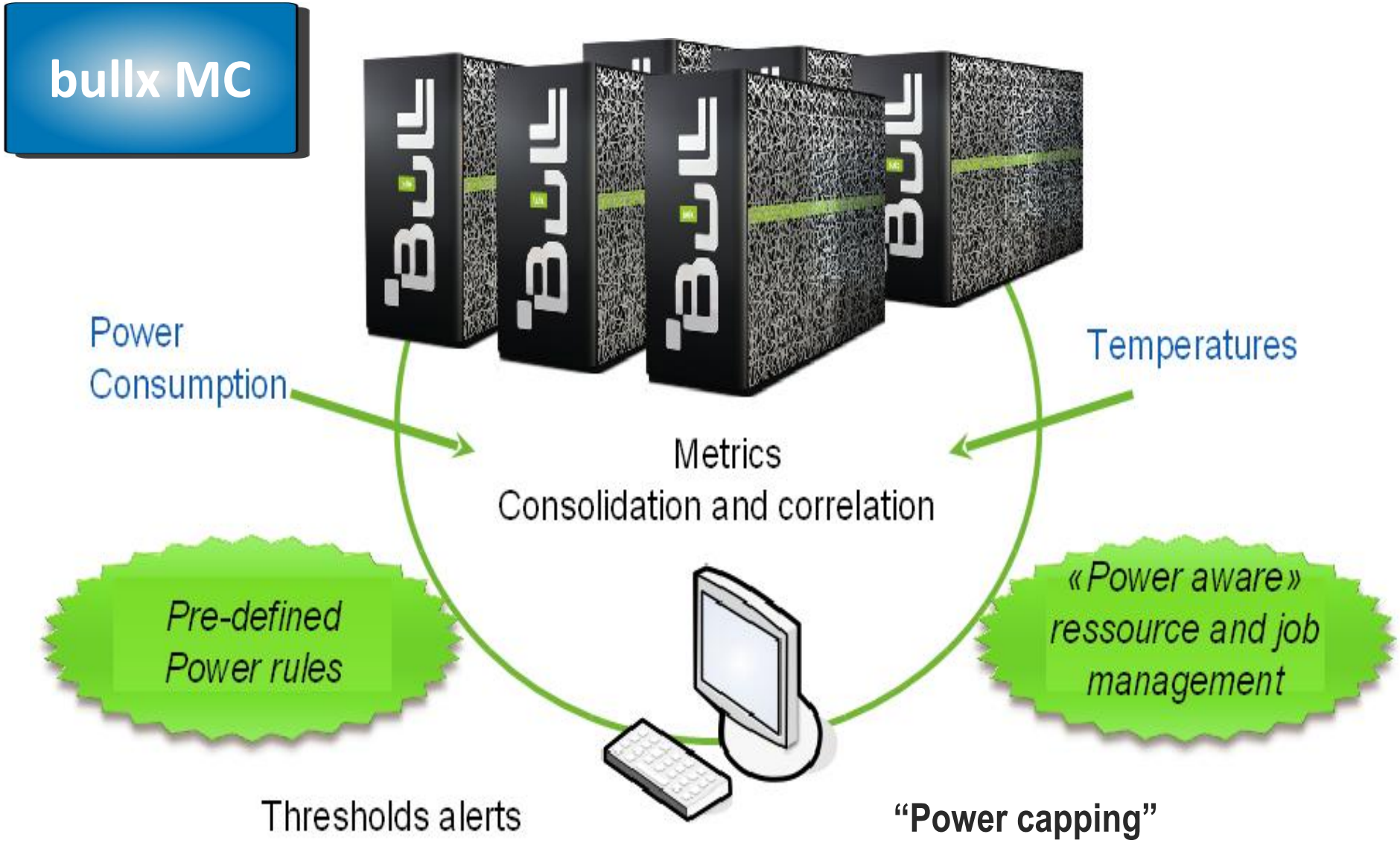
Power Management

- 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

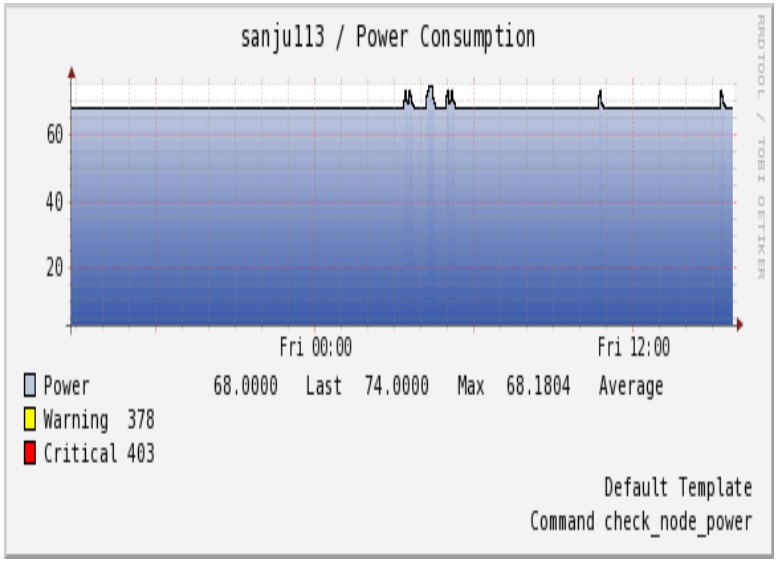


Power Control scenario



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

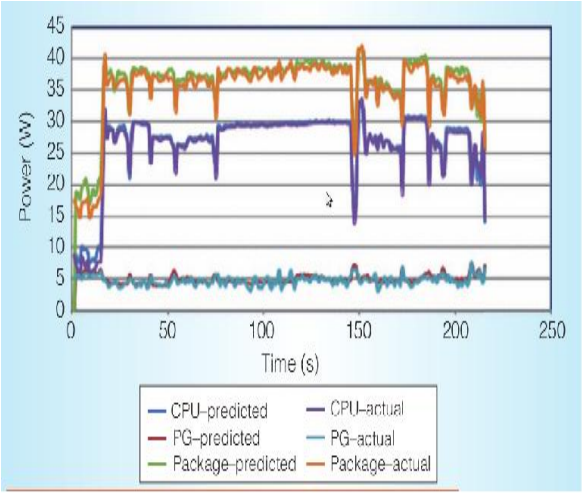


- Power capping
  - Automatic action to decrease power level
  - Automatic information for system monitoring
  - Open framework, based on SEC (Simple Event Correlator)
  - Allow new rules creation
  - But slow reaction time (minutes)

What do consume your applications?

bullx BM

- 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

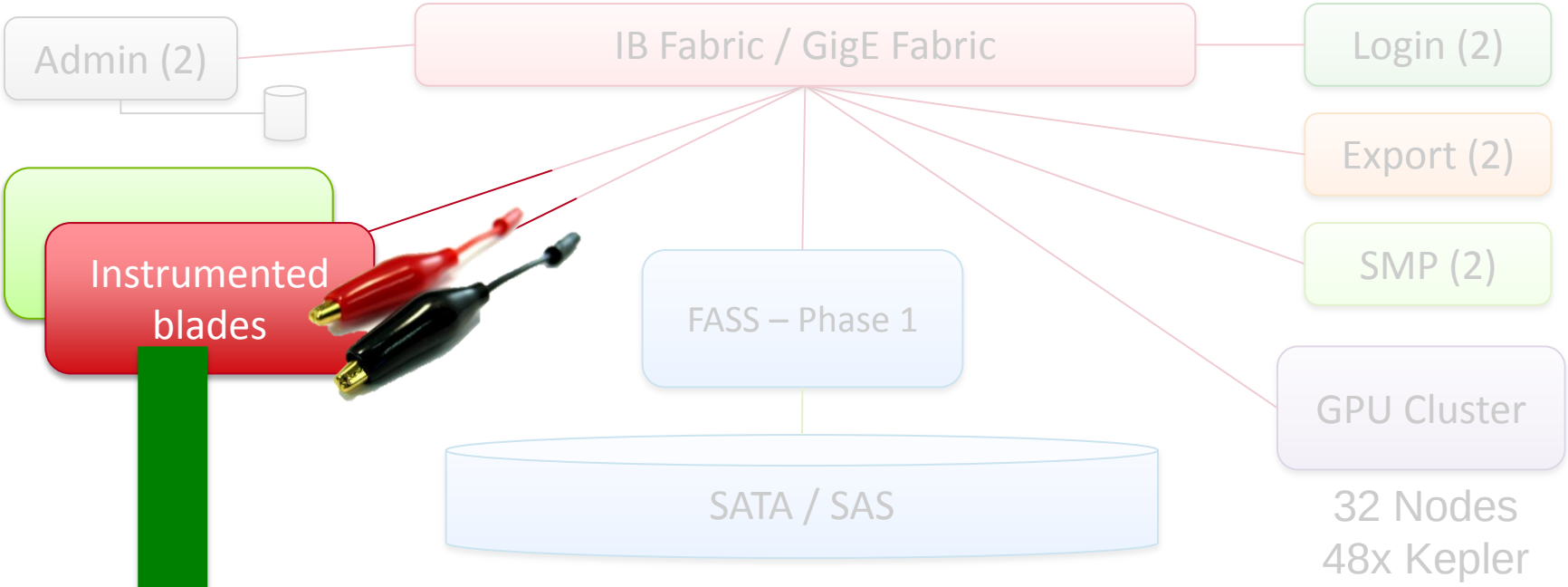
MIDDLEWARE
- New modules in VAMPIR
- Scalable High Definition Power Monitoring API

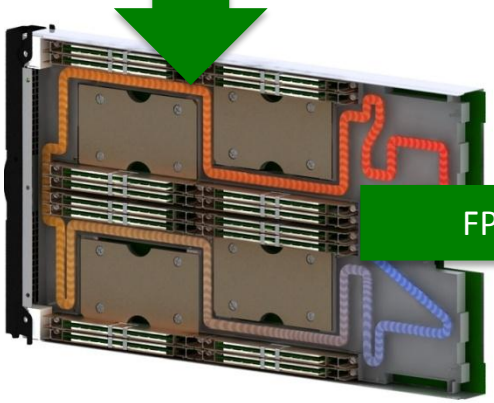
APPLICATION
- Development of new optimization methodologies
- Demonstration of energy efficiency improvement

OPENSOURCE

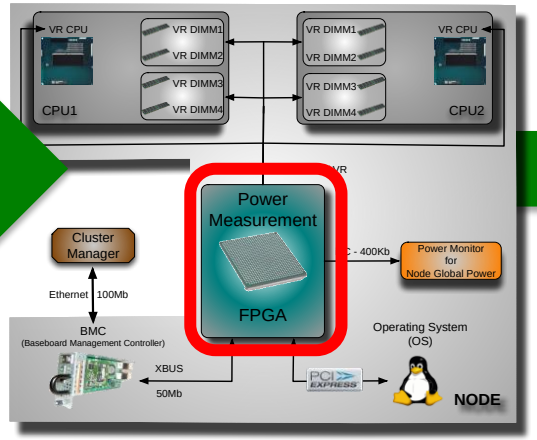
- ✓ Energy efficient operation
  - ✓ CPU states
  - ✓ Turn off devices
  - ✓ Interface with batch scheduler
  - ✓ Mesurement environment
  - ✓ Vampir integration
- ✓ Energy accounting
- ✓ FPGA integration
- ✓ Measuring system Accuracy
- ✓ Energy Efficiency research at application level

Bull TU Dresden high frequency monitoring

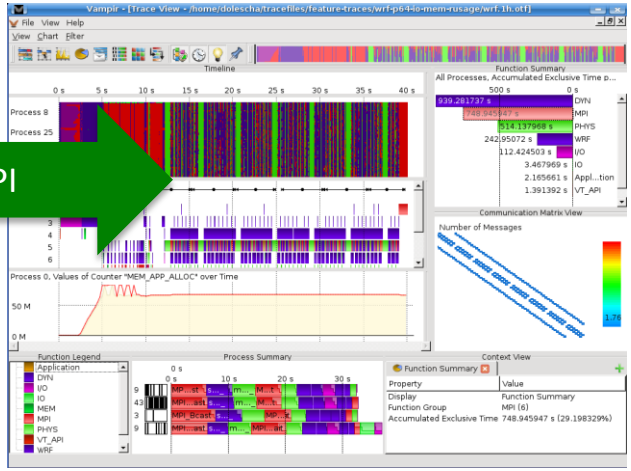




FPGA



API



Freq: 500Hz - Error: 2%

Power Management …

- Interest driven by energy cost and green attitude
- Minimal TCO might not agree with Green
- Non-intrusive power monitoring at low frequency (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 to be defined

Architect of an Open World™