ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS
Swiss National Supercomputing Centre

Measuring performance and energy efficiency in the High-Performance Linpack benchmark

- Gilles Fourestey and Thomas C. Schulthess

Generally, what we need to optimize in HPC

Time to solution

&

Energy to solution

How do we minimize time to solution?

It depends on the application!

Specifically for High-Performance LINPACK:

(1) high arithmetic density: $$\frac{\#\text{ of flop}}{\#\text{ of load-stores}} \propto O(N)$$

(2) total work measured in number of retired floating point operations ($totflop$) can be easily computed

(3) normalized time to solution: $$\frac{\Delta t_c}{totflop}$$ or performance $$\frac{totflop}{\Delta t_c}\left[\frac{\text{flop}}{\text{sec.}} = \text{flops}\right]$$

Maximizing floating point performance is equivalent to minimizing time to solution for high-performance LINPACK

How do we minimize energy to solution?

It depends on the application!

Specifically for High-Performance LINPACK:

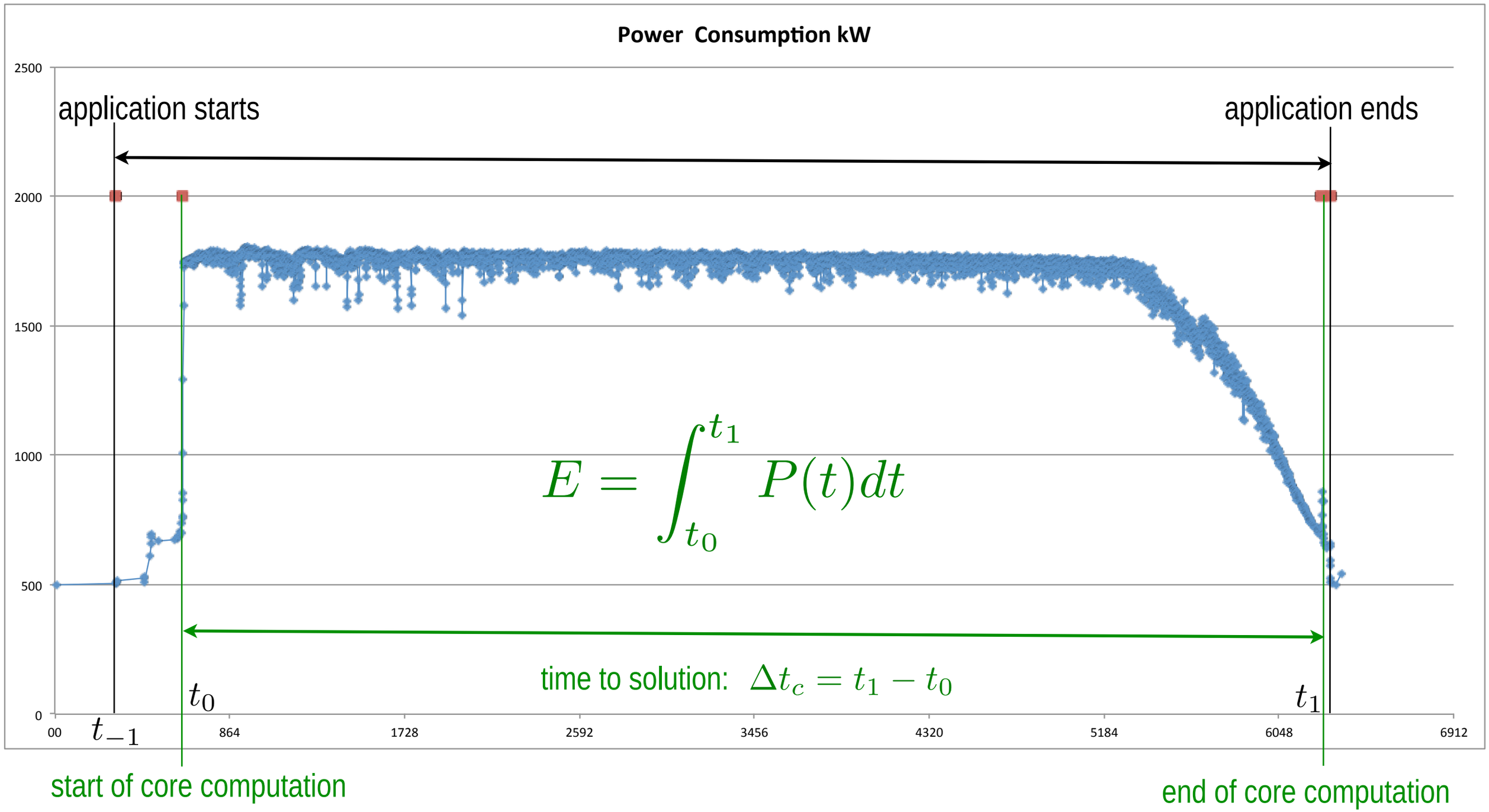
(1) normalized energy to solution E by simple measure of work $\frac{E}{totflop}$

(2) minimizing energy to solution is equivalent to maximizing $\frac{totflop}{E}\ \left[\frac{\text{flop}}{\text{Joule}}\right]$

(3) ... and of course $\left[\frac{\text{flop}}{\text{Joule}}\right] = \left[\frac{\frac{\text{flop}}{\text{sec.}}}{\text{Watt}}\right]$

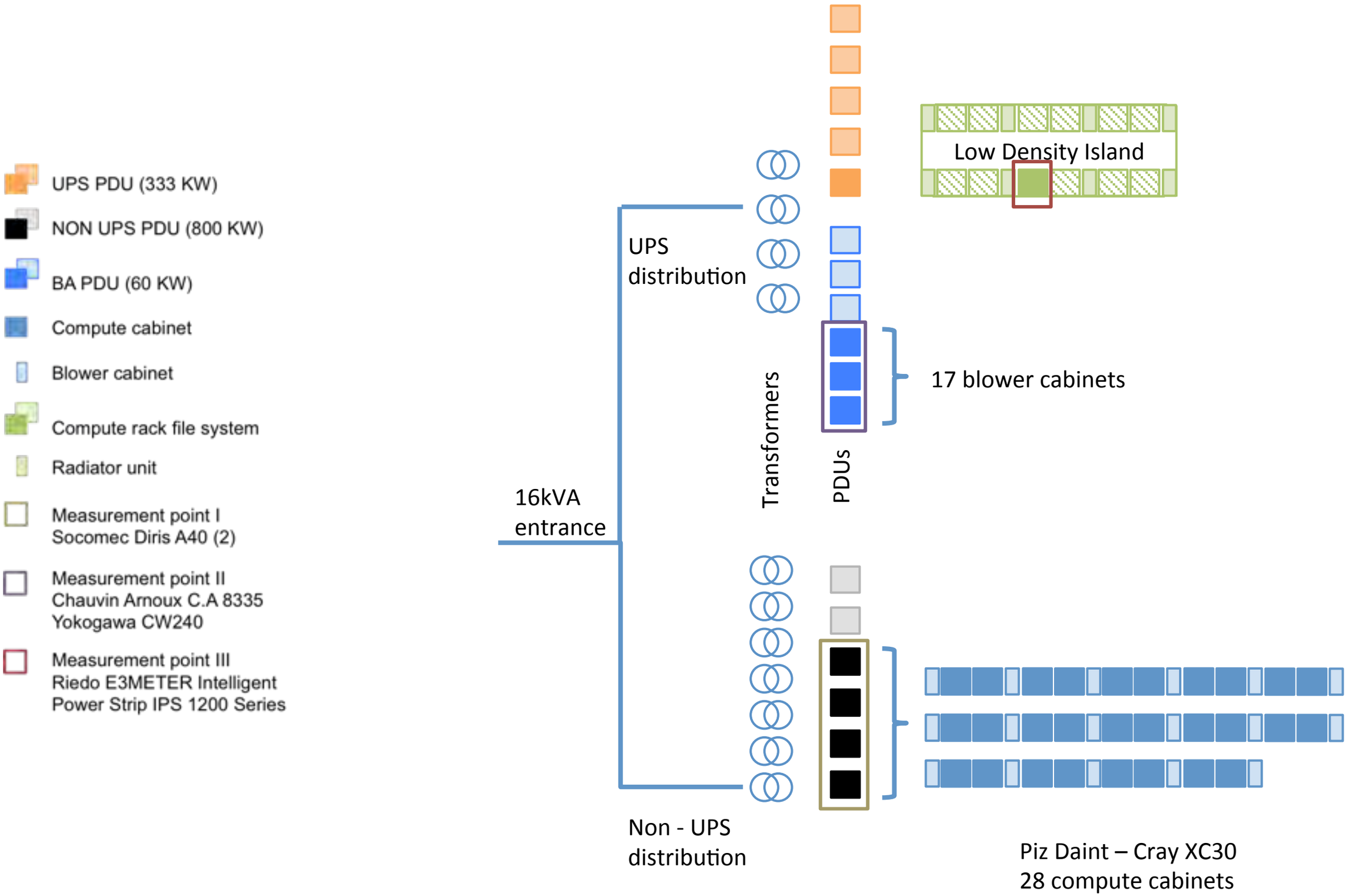
Minimizing energy to solution is equivalent to maximizing [flop/Joule] or [flop/sec./Watt]

How do we measure energy to solution?



Power has to be measured on entire system or at least a representative cross section

Level 3 measurement on “Piz Daint” at CSCS



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS
Swiss National Supercomputing Centre

Results for “Piz Daint”

Level 3 (*) measurement on run optimized for power efficiency:

Execution time core phase 5625.76 s
Avg. power for core phase 1753.66 kW
Measured HPL performacne 5.587 Pflops

3.186 Gflops/W

(*) this is compatible with all I have said so far

Piz Daint cheat sheet
- Cray XC30, 28 cabinets
- 5272 compute nodes
  - Intel Xeon E5-2670 (SandyBridge)
  - NVIDIA K20x GPU (Kepler)
  - 32GB DDR3-1600 memory
  - 6GB GDDR5 memory
- Aries network with 1344 router chips
  - 3276 of 3360 optical ports used
  - Bisection bandwidth 33075 GB/s
  - Global bandwidth 11.6 GB/s per node
- File system has 2.5 PB capacity
  - 138 GB/s max bandwidth
- Floor space: 106 m^2
- Inlet water temperature: 16C

The (Level 1) rules of Green500 would have allowed us to
- use system internal power monitoring tool, using a 5% correction to allow for power conversion loss and measure only compute nodes (ignoring network -- sounds familiar?)
- take measurement on compute nodes only for the full core phase (we could do only 20% of core phase)

Execution time core phase and performance measurements is same as above
Avg. power for core phase 1446 kW

3.864 Gflops/W, but this is wrong!

What if we can't measure the entire system power?

Nevertheless integrate the power for the entire core phase!

Titan @ ORNL: measured power consumption on 1/4 of 200 cabinets

2.143 Gflops / Watt

Tödi @ CSCS: measured power consumption on all 3 cabinets

2.112±0.001* Gflops / Watt

(*) error is standard deviation taken from five independent runs; there might be an additional systematic error, however, comparing to the number measured at ORNL, this error seems to be of order or smaller than 1%; the difference to ORNL measurement is more likely due to fluctuations in cabinet power consumption and summing over fewer cabinets at CSCS.

Independent but complete measurements on either side of the Atlantic yield the same result!

THANK YOU!