First Experiences With Validating and Using The Cray Power Management Data Base Tool

Gilles Fourestey, Ben Cumming, Ladina Gilly and Thomas C. Schulthess, CSCS

HPC Performance Metric

Classic HPC metric: Time To Solution (TTS)

How do we minimize TTS?

For CPU-bound applications it means:

maximizing the flops count.

HPL (top500) is a directed reflection of this fact.

HPC Performance Metric

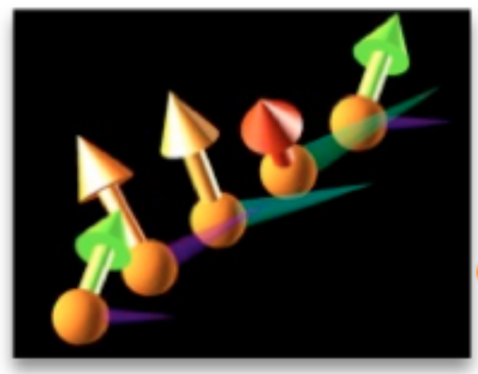
Application performance seems to keep up with supercomputing systems performance (!)

~100 Kilowatts → ← ~5 Megawatts → ← 20-30 MW →

~1 Exaflop/s

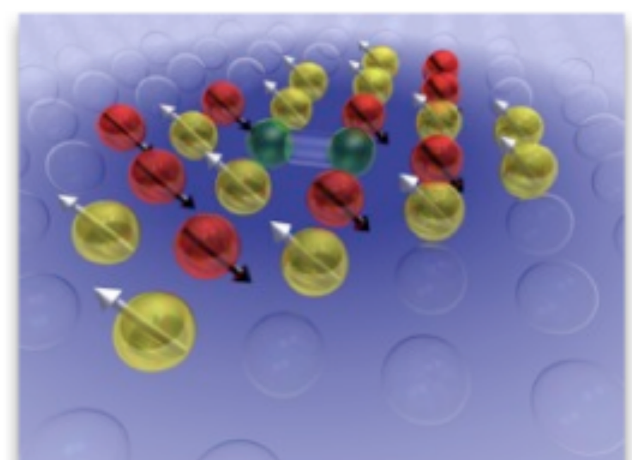
100 million or billion processing cores (!)

1 Gigaflop/s
Cray YMP
8 processors



1.02 Teraflop/s
Cray T3E
1'500 processors

1.35 Petaflop/s
Cray XT5
150'000 processors



1988	1998	2008	2018
First sustained GFlop/s Gordon Bell Prize 1988	First sustained TFlop/s Grondon Bell Prize 1998	First sustained PFlop/s Gordon Bell Prize 2008	Another 1,000x in sustained performance increase

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS
Swiss National Supercomputing Centre

HPC Performance Metric

We need to not only consider:

Time To Solution

but also:

Energy To Solution

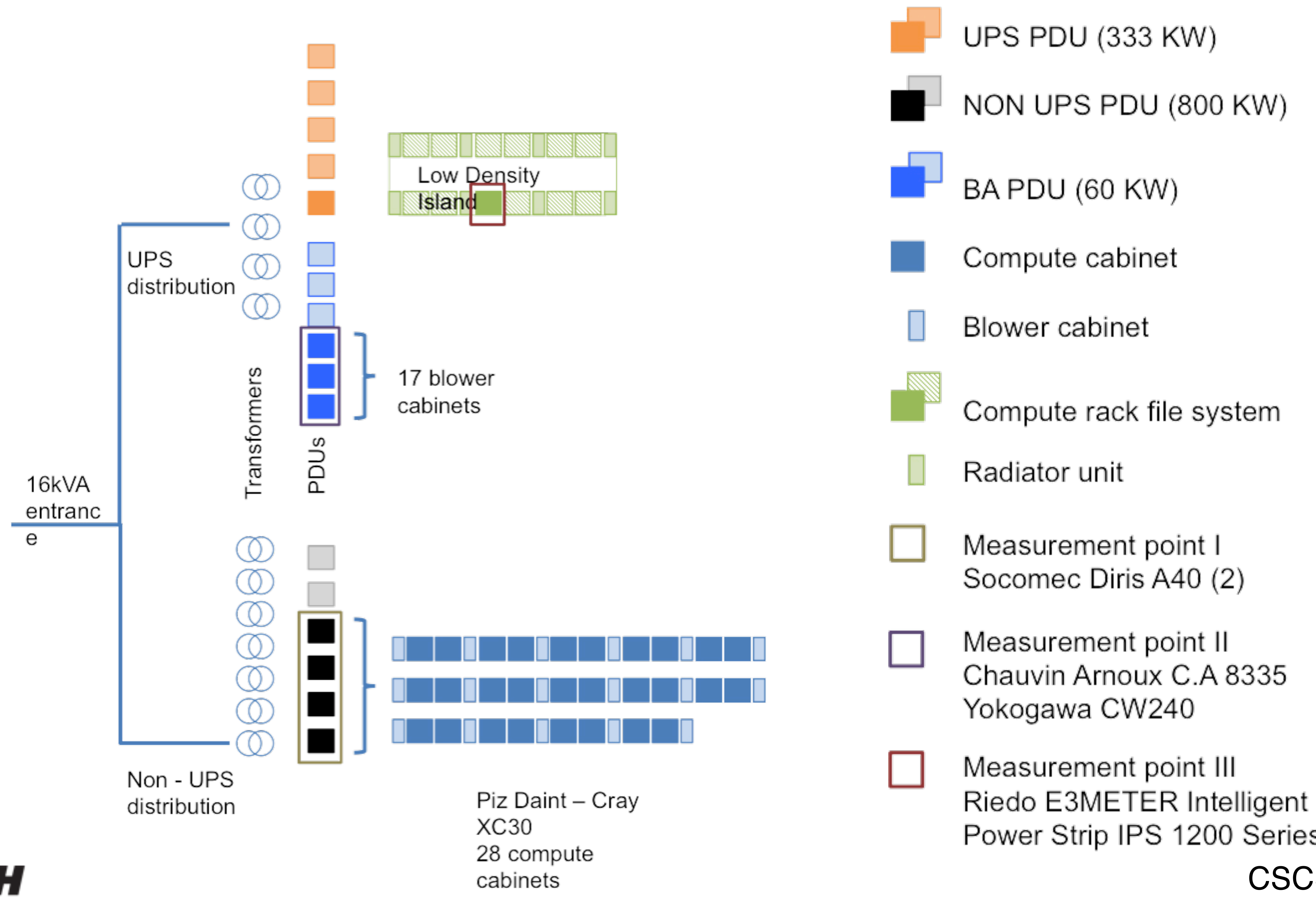
How do we measure Energy To Solution?

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS
Swiss National Supercomputing Centre

External Power Meters

Level 3 capable measurements at CSCS



Power Management DataBase

Integrated Power/Energy measurement for the nodes, GPUs, blades, racks, network and blowers stored by time or APID.

- PMDB: direct access to the database (node, blade and cabinet level, NOT in user space)
- RUR: Resource Utilization Reporting (node level, in user user space)
- PM counters: files on each node storing power/energy data (node level, in user space)

PMDB example:

```
  APID    |   Joules    |         KWh
---------+------------+----------------------
 2134412 | 2928139984 | 813.372217777777778
```

Aries chips and blowers power consumption are missing.

Real power (node-level) = (Joules/TTS + (# nodes)/4*Aries + Blowers)/0.95
Real power (cab-level) = (Joules/TTS + Blowers)/0.95

0.95: AC/DC conversion rate Aries: 100W (static) Blowers: 4440~5300W

Resource Utilization Report

RUR is the simplest way to get energy consumed by a job.

In /scratch/daint/RUR/rur-<date> (in Joules):

uid: 21553, apid: 2380700, jobid: 289724, cmdname: ./hpcg energy ['energy_used', 11240718]

- Per APID/jobID: so full job, not parts of the code
- Node energy, i.e CPU + GPU + RAM, no network, no blowers: the node-level PMDB formula applies!

PM Counters

Sysfs files on each nodes that are updated approximately every 0.1 seconds with power and energy for the node and the accelerator.

Polling those files will trigger an interrupt in the system so don't poll too often.

Can be used to measure energy/power consumption for regions of a code.

PM Counters

```
int get_acc_{energy,power}(){
    int value;
    char buff[16];
    FILE *fid;

    fid =  fopen("/sys/cray/pm_counters/accel_{energy,power}", "r");
    fscanf(fid, "%d %s", &value, buff);
    fclose(fid);
    return value;
}
```

```
int get_{energy,power}(){
    int value;
    char buff[16];
    FILE *fid;

    fid =  fopen("/sys/cray/pm_counters/{energy,power}", "r");
    fscanf(fid, "%d %s", &value, buff);
    fclose(fid);
    return value;
}
```

Example:

```
$> aprun -n 1 ./energy.sh
62 W
33404642 J
33404707 J
```

energy.sh:

```
#!/bin/bash
cat /sys/cray/pm_counters/power
cat /sys/cray/pm_counters/energy
sleep 1
cat /sys/cray/pm_counters/energy
```

PM Counters

Cublas DGEMM, 15000x15000x15000

-------------- Idle Power

idle node_power = 60 W

idle acc_power = 20 W

-------------- Data Transfer Power

xfer rate = 5.692945 GB/s (0.883399 s.)

xfer node_power = 113 W

xfer acc_power = 52 W

PMDB: 113 W

xfer nvml_power = 46 W

-------------- DGEMM Power

kernel perf = 1167.82 Gflops (5.786003 s.)

kernel node_power = 268 W, (1549 J)

kernel acc_power = 210 W, (1215 J)

Kernel node_energy = 1545 J

Kernel acc_energy = 1211 J

PMDB: 267W

kernel nvml_power = 190 W, (1099 J)

Real Life Applications

Idle power consumption, in kW (Clogin at Chippewa Falls, 3 racks):

C0 (cab-PMDB)	C1 (cab-PMDB)	C2 (cab-PMDB)	Sum	Corrected	Ext. PM
16.333	16.043	16.452	48.829	65.420	66.067

Blowers at rest, in kW (Piz Daint at CSCS, 17 blowers):

PMDB	Corrected	Full System	Ext. PM
4440	4673.7	79.452	79.448

DCA+

- Dynamic Cluster Approximation (DCA) models of high-temperature superconductors
- Continuous time quantum Monte-Carlo solver with delayed updates which allows to use an efficient algorithm based on BLAS level 3 operations (CPU + GPU)
- Each test has a different Temperature (from high temperature to below Tc) and time to solution

	#0	#1	#2	#3	#4	#5	#6	#7
TTS (s)	3787	2725	922	605	329	182	75	23
Cab PMDB (kW)	58.5	57.2	55.9	53.6	53.2	49.3	46.8	43.0
External PM (kW)	58.6	57.1	53.9	52.9	52.2	47.0	42.7	30.5

Comparison is very good for large jobs

Small jobs: the external power meter had a 0.1Hz smapling frequency, e.g for test #7 we only had 4 samples.

COSMO Application

- COSMO is an atmospheric simulation code
  - Used for both weather forecasting and climate modeling

- Fully ported run on both multi-core and GPUs.
  - Currently production climate simulations are run on GPUs on Piz Daint.
  - Ideal for comparing both time to solution and energy to solution on different architectures.

- We use COSMO-2 to compare Cray systems (XE6, XK7, XC30 & hybrid XC30)
  - COSMO-2 is 2-km model of the Alps currently used for daily weather forecasting by MeteoSwiss.
  - Use ensemble configuration with 9 nodes per member
    - enough ensemble members to fill an entire cabinet of each system
    - 10 members on XE6 and XK7 systems
    - 20-21 members on XC30 and hybrid XC30 systems

COSMO Validation

- First we validated the PMDB measurements on XC30 with an external power meter
  - External meter measured entire system: 3 cabinets + 3 blowers
  - PMDB cabinet level measurements for each cabinet
    - We add 3*4440 W (unadjusted) for blowers
  - 62 ensemble members fill 3 cabinets on system and we perform simultaneous external and PMDB measurement

	PMDB (kWh)	external meter (kWh)	estimated efficiency
Run 1	53.63	56.45	95.0%
Run 2	53.47	56.27	95.0%

- Results are consistent between runs : 0.3% difference between run 1 and run 2
- The estimated efficiency of 95% for AC-DC conversion is valid for COSMO

COSMO Comparison

- Fill a cabinet on each test system with ensemble members
  - Include blowers on XC30 (XE6 and XK7 systems have integrated blowers)
  - New systems improve time and energy to solution (XE6 vs XC30 and XK7 vs hybrid XC30)
  - GPU has better time to solution and energy to solution than CPU implementation
  - Energy to solution improvements are bigger than time to solution
- Measuring energy and power was much easier with PMDB

System	Rosa	Todi	Daint	Clogin
Type	XE6	XK7	XC30	Hybrid XC30
Ensemble members	10	10	20	21
Time to solution (s)	3683	2579	2083	1539
Mean cabinet power (kW)	40.22	62.07	28.27	41.6
Energy to solution (kWh)	41.14	44.47	16.34	17.77
Energy per member (kWh)	4.11	2.22	1.64	0.85
TTS scaling	1.0	1.4	1.8	2.4
ETS scaling	1.0	1.9	2.5	4.8

Green500

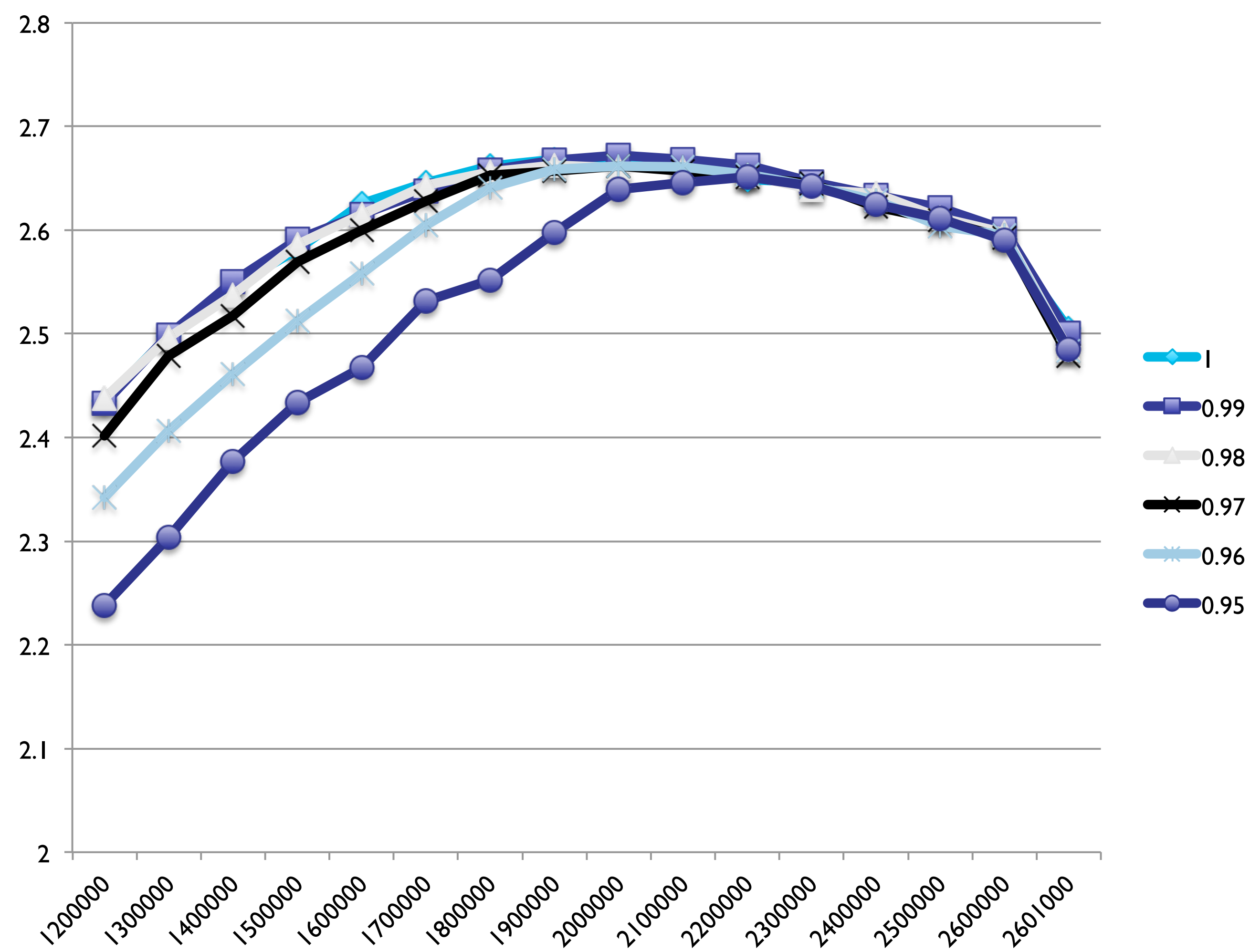
Green500: maximize energy efficiency (Gflops/W)

- Maximize Gflops
- Minimize power consumption

Energy efficiency per component:
- CPU energy efficiency is ~1.28 Gflops/W
- GPU energy efficiency is ~5.95 Gflops/W

We have to maximize GPU work, minimize CPU

HPL parameters tuning for green500:
- GPU/CPU split (between 90 and 100%)
- CPU throttling (16 p-states)

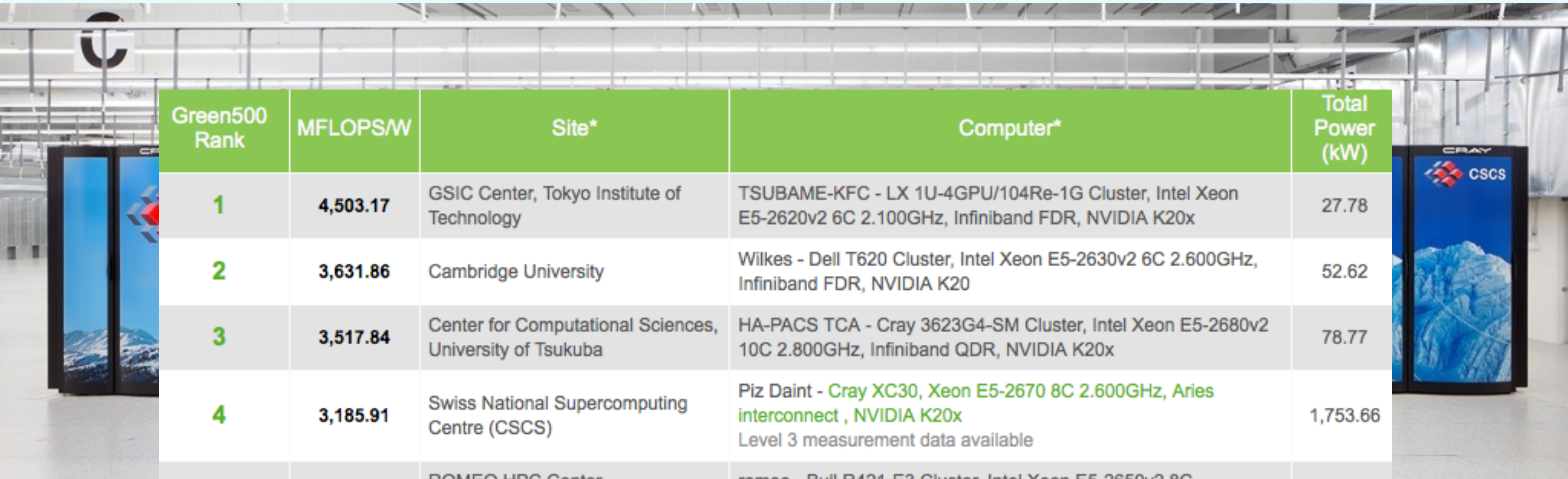


CPU freq.	RUR	PMDB (node)	PMDB (cab)	Facility
1.9	1526	1536	1600	1635

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS
Swiss National Supercomputing Centre

Green500



Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,503.17	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	27.78
2	3,631.86	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
5	3,130.95	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
6	3,068.71	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.930GHz, Infiniband QDR, NVIDIA K20x	922.54

Most energy efficient Petaflop system:
- 1.753 MW
- 3'185 Mflops/W

Top500 was:
- 2.3 MW
- 2'69 Mflops/W

Thanks To:

- Cray for assistance in Chippewa Falls, in particular Steve Martin and Ron Rongstad.

- Nina Suvanphim (Cray) at CSCS for her assistance.

- Tiziano Belotti, Rolando Summermatter and Luca Bacchetta from the Facility Management Group at CSCS.

- Massimiliano Fatica (nVidia) for the hybrid HPL code.

Thank You!

Questions?...

gilles.fourestey@cscs.ch

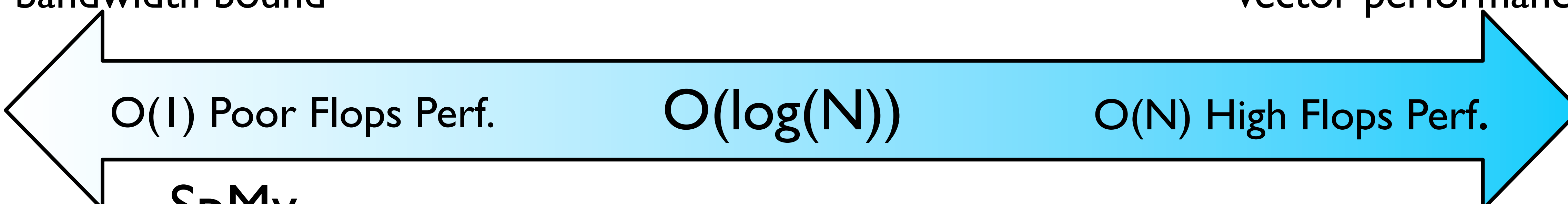
HPC Performance Metric

Classic HPC metric: Time To Solution (TTS)

How do we minimize TTS? More flops.

Arithmetic Intensity AI := flop/DRAM accesses

Memory-bandwidth bound | Vector performance bound



SpMv, Blas 1 & 2 | Stencil | FFT | Blas 3, HPL

- Long stride memory access
- Little reuse of accessed memory
- Flops < DRAM access

- Unit-stride memory access
- Reuse of accessed memory
- Flops > DRAM accesses

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS
Swiss National Supercomputing Centre