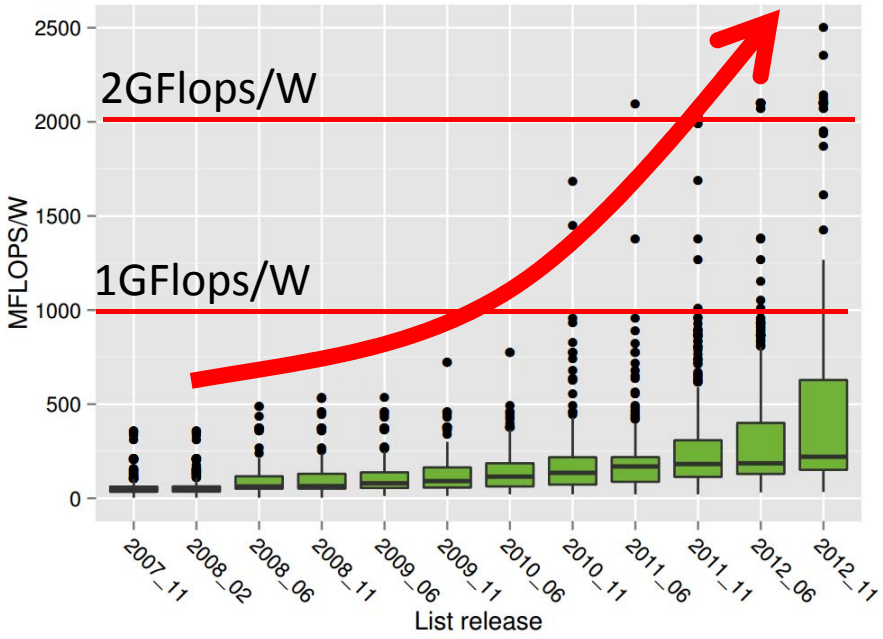
TSUBAME-KFC : a Modern Liquid Submersion Cooling Prototype Towards Exascale

Toshio Endo, Akira Nukada, Satoshi Matsuoka

GSIC, Tokyo Institute of Technology (東京工業大学)

Performance/Watt is the Issue

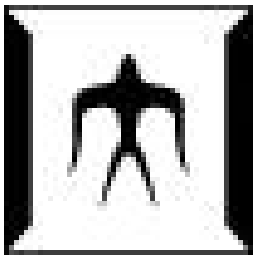


- Development of supercomputers are capped by power budget
  - Realistic supercomputers, data centers are limited by ~20MW
- In order to achieve Exascale systems, we will require technologies enabling 50GFlops/W
  - Around 2020~2022

Development of supercomputers' power efficiency.
From Wu Feng's presentation@Green500 SC13 BoF

Achievement 4 Years Ago

TSUBAME 2.0 supercomputer achieved
~1GFlops/W (1.2PFlops Linpack, 1.2MW)



- World's 3rd in Nov2010 Green500 ranking
- Greenest Production Supercomputer award

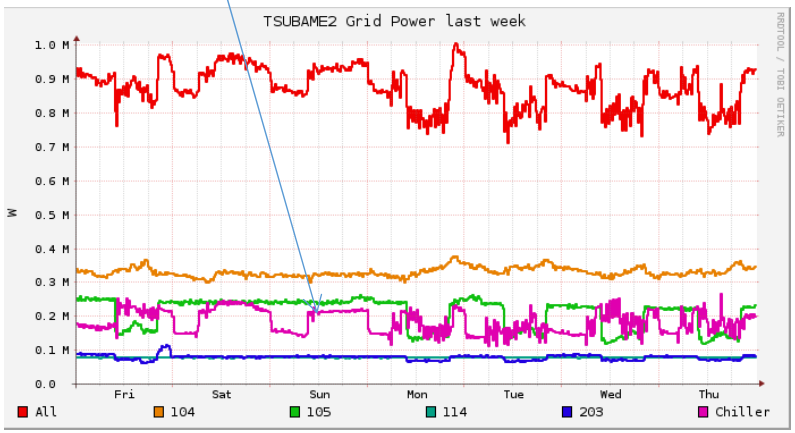
Towards TSUBAME3.0 (2016),
We should be much more power efficient!!

How Do We Make IT Green?

- Reducing computers power
  - Improvement of processors, process shrink
  - Utilization of power-efficient many-core accelerators
  - Software technologies that efficiently utilize accelerators
  - Management technologies for power control

Today's focus

In TSUBAME2, cooling system consumes >25% power of the system



- Reducing cooling power
  - Liquid has higher heap capacity than air

→Liquid cooling is preferable

  - We should avoid making chilled water

→Warm/hot liquid cooling

  - Designing water pipe / plate is expensive
  - Control of coolant speed is more difficult

→ Fluid submersion cooling

TSUBAME-KFC:
Ultra-Green Supercomputer Testbed

TSUBAME-KFC

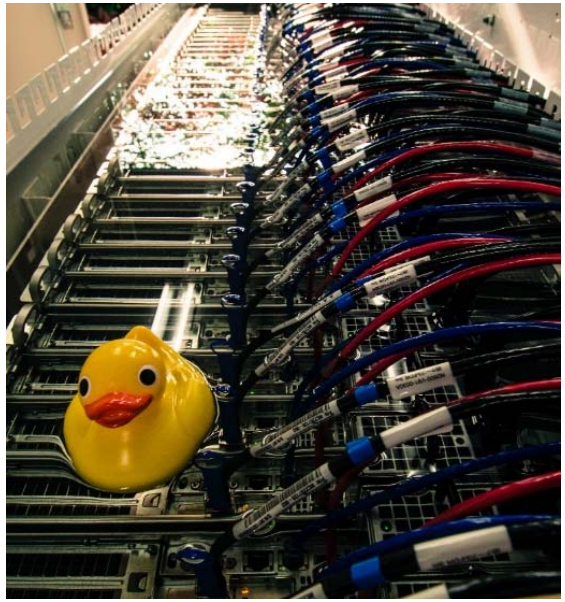
or **K**epler **F**luid **C**ooling

= (Hot Fluid Submersion Cooling

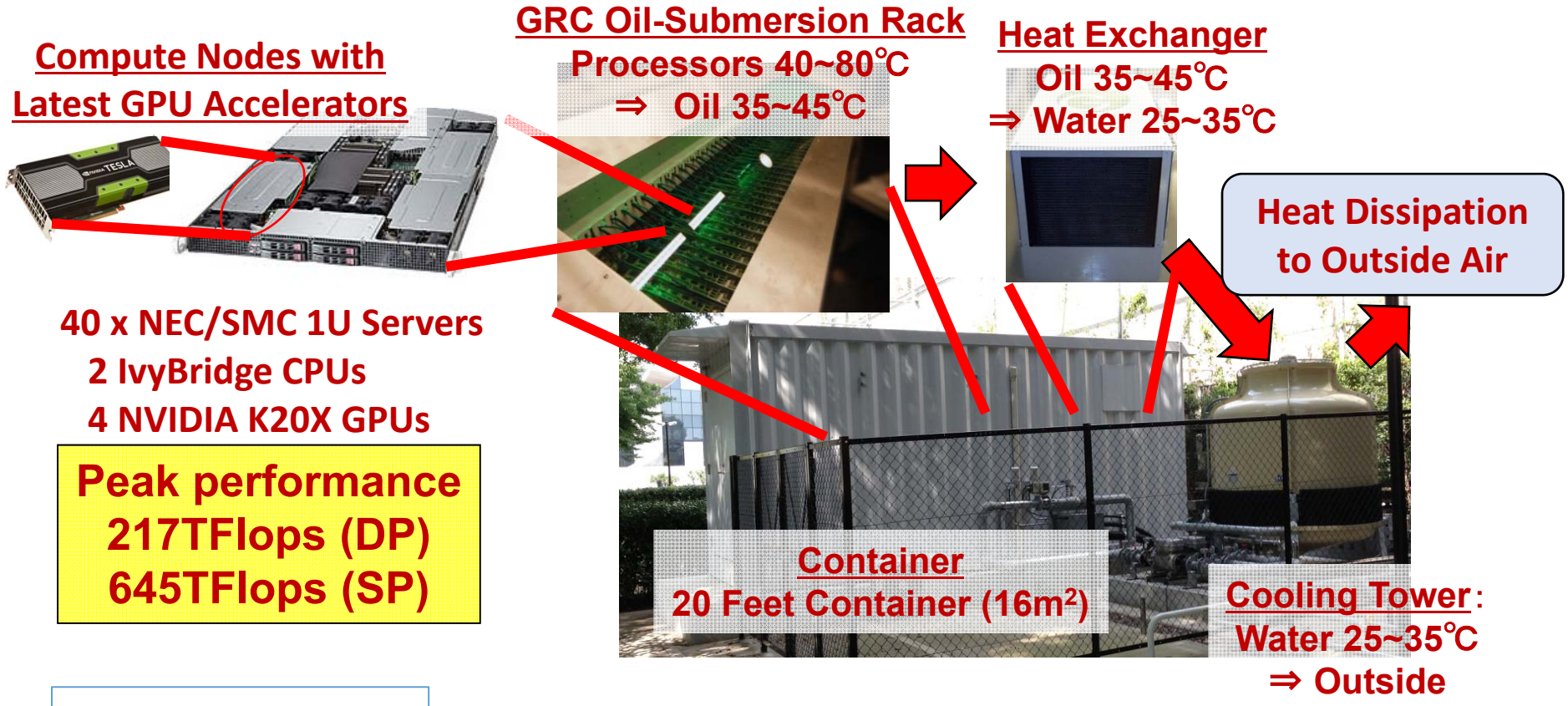
+ Outdoor Air Cooling

+ Highly Dense Accelerated Nodes)

in a 20-feet Container



TSUBAME-KFC: Ultra-Green Supercomputer Testbed



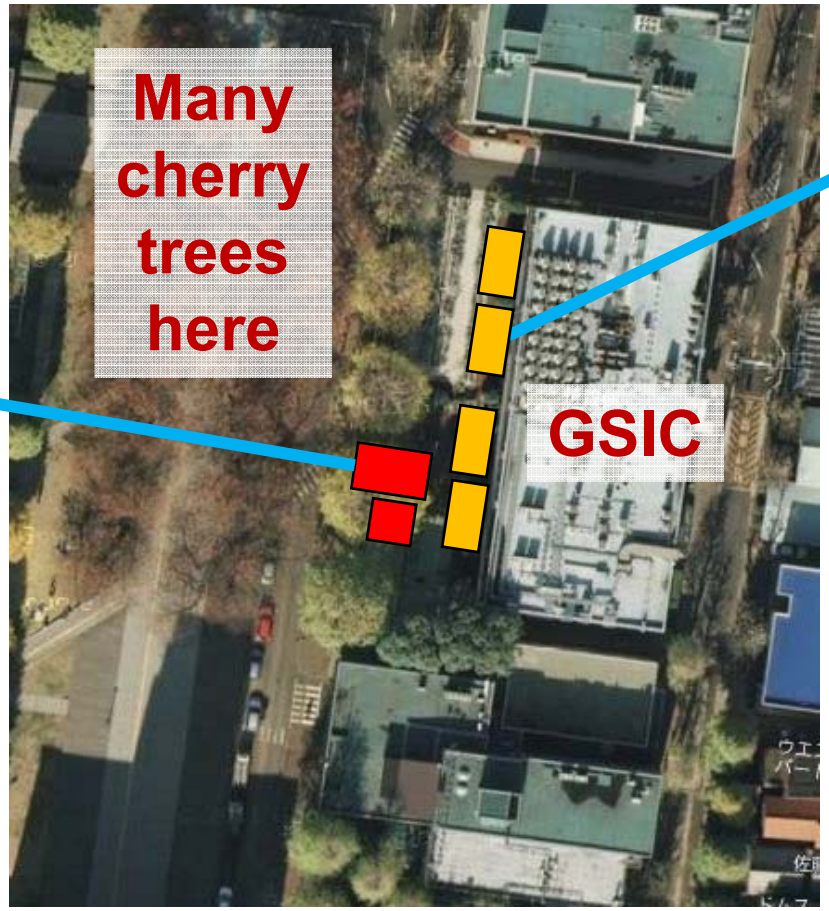
Achievement

- Worlds' top power efficiency, 4.5GFlops/W
- Avarage PUE <1.1 (Cooling power is ~10% of system power)

Installation Site of TSUBAME-KFC

Neighbor space of GSIC, O-okayama campus of Tokyo Institute of Technology

- Originally a parking lot for bicycles



Chillers for TSUBAME2

KFC Container & Cooling tower

Coolant Oil Configuration

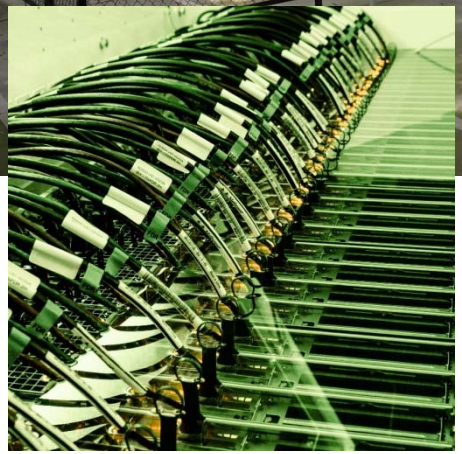
ExxonMobil SpectraSyn Polyalphaolefins (PAO)

	4	6	8
Kinematic Viscosity@40C	19 cSt	31 cSt	48 cSt
Specific Gravity@15.6C	0.820	0.827	0.833
Flash point (Open Cup)	220 C	246 C	260 C
Pour point	-66 C	-57 C	-48 C

Fire Station at Den-en Chofu

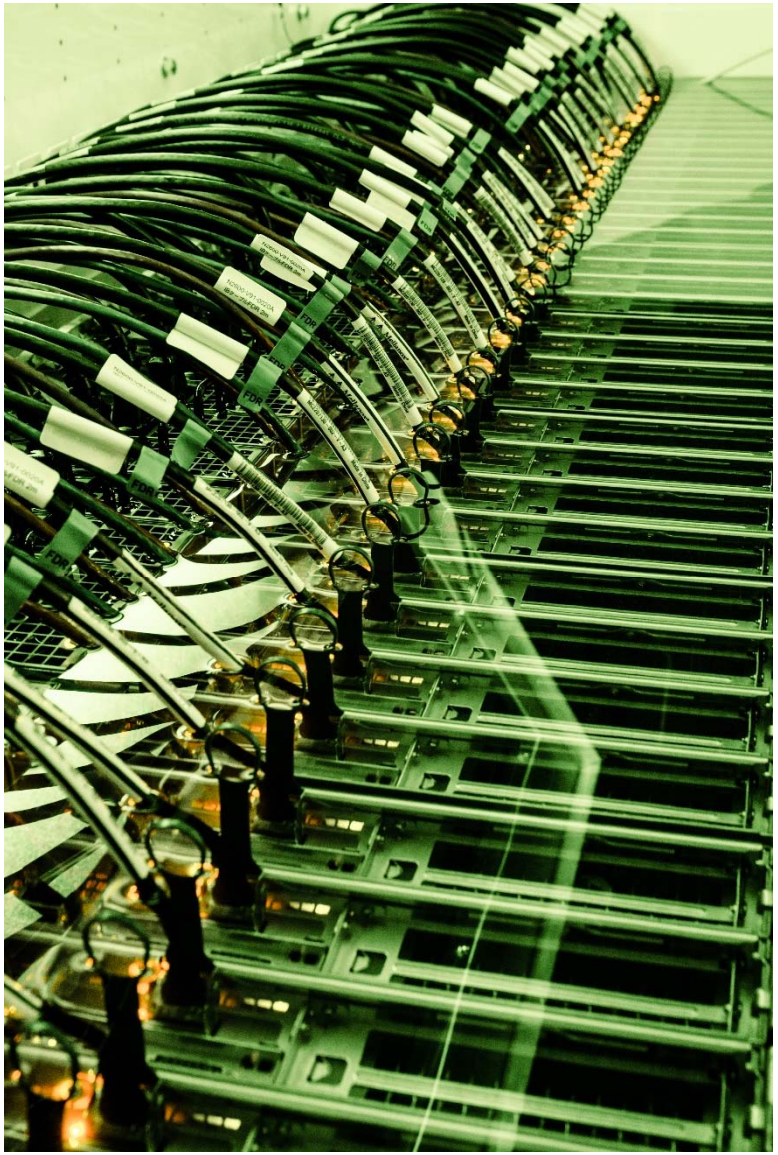
We are using ~1,200 liters oil
Flash point of oil must be >250℃,
Otherwise it is a hazardous material under the Fire Defense Law in Japan.

Installation



Installation completed in Sep 2013

40 KFC Compute Nodes



NEC LX 1U-4GPU Server, 104Re-1G

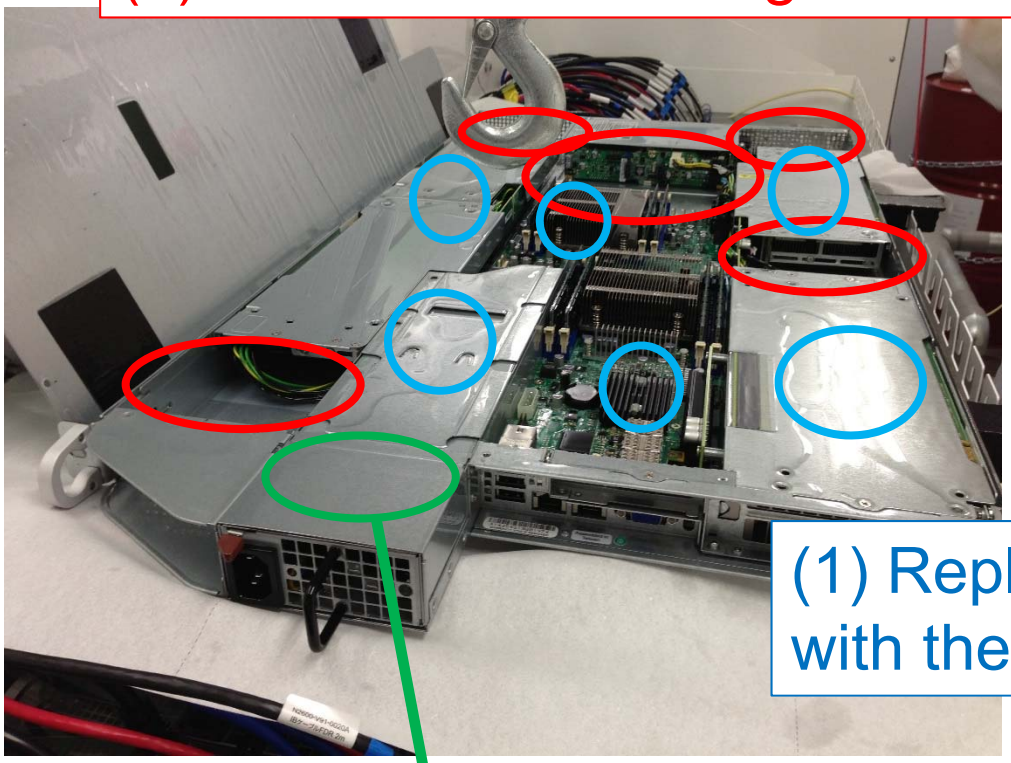
(SUPERMICRO OEM)

- 2X Intel Xeon E5-2620 v2 Processor (Ivy Bridge EP, 2.1GHz, 6 core)
- **4X** NVIDIA Tesla K20X **GPU**
- 1X Mellanox FDR InfiniBand HCA
- 1.1TB SATA SSD (120+480+480)

CentOS 6.4 64bit Linux
Intel Compiler, GCC
CUDA 5.5
OpenMPI 1.7.2

Modification to Compute Nodes

(2) Removed 12 cooling fans

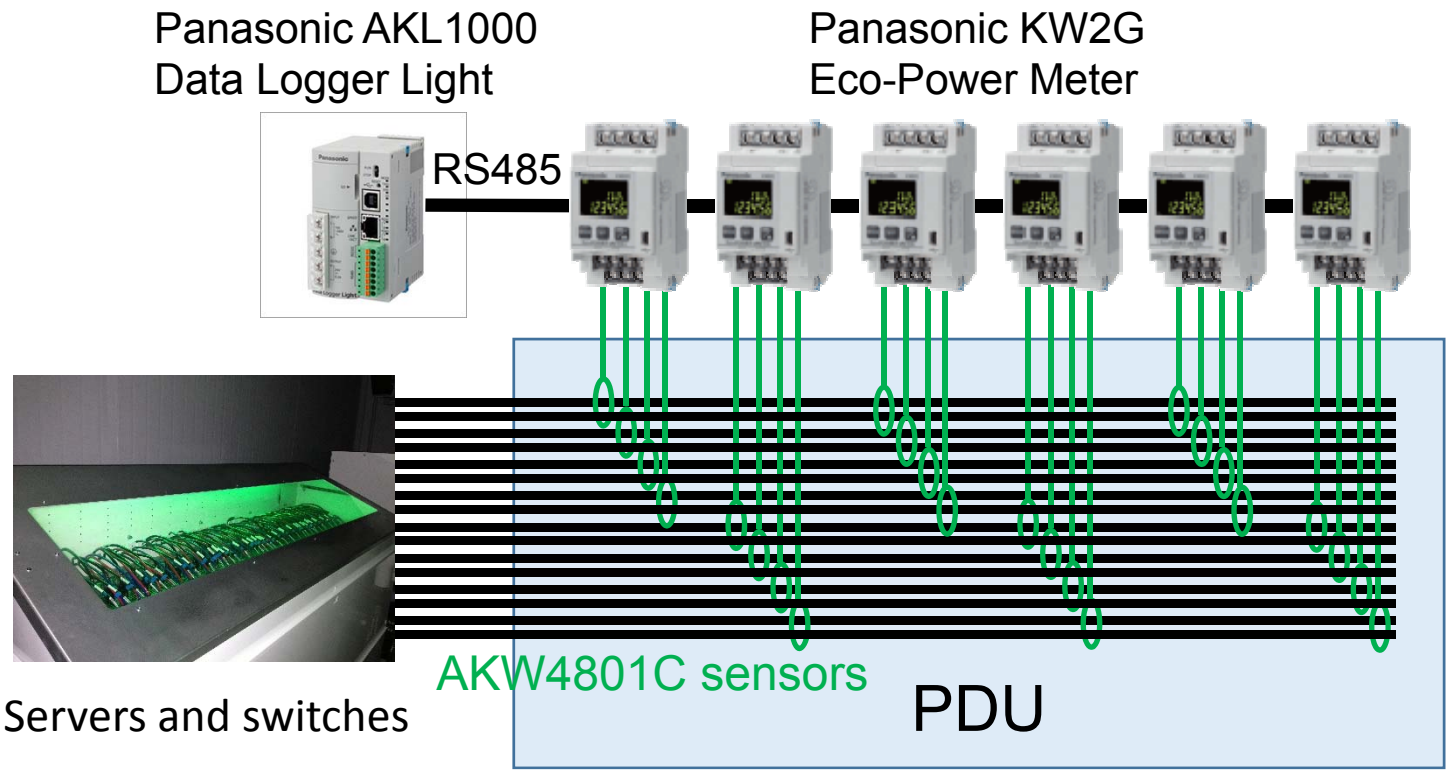


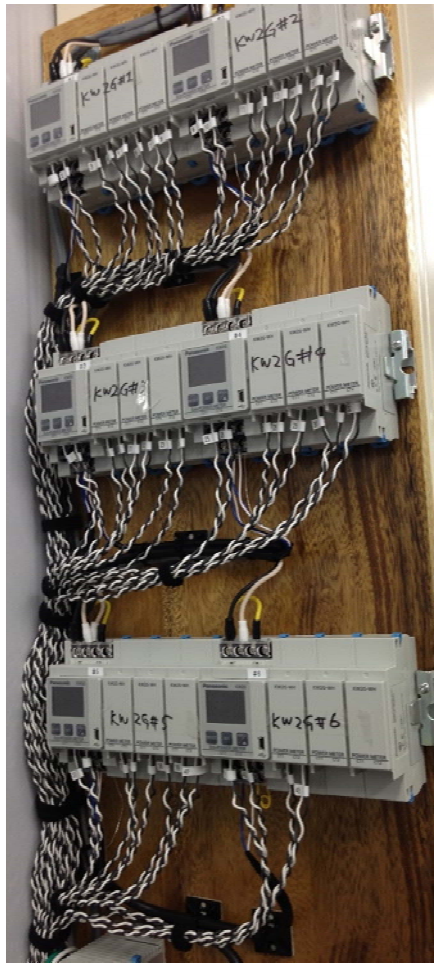
(1) Replace thermal grease with thermal sheets

(3) Update firmware of power unit to operate with cooling fan stopped.

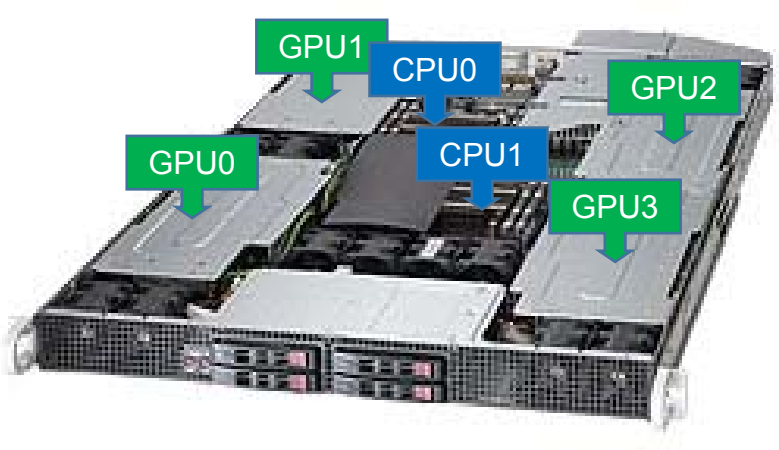
Power Measurement

In TSUBAME-KFC, we are recording power consumption of each compute node and each network switch, in one sample per second.





Node Temperature and Power



Using IPMI to fetch Temp. data.

Lower oil temp results in lower chip temp.
But no further power reduction achieved.

Upper: Running DGEMM on GPU
Lower: (IDLE)

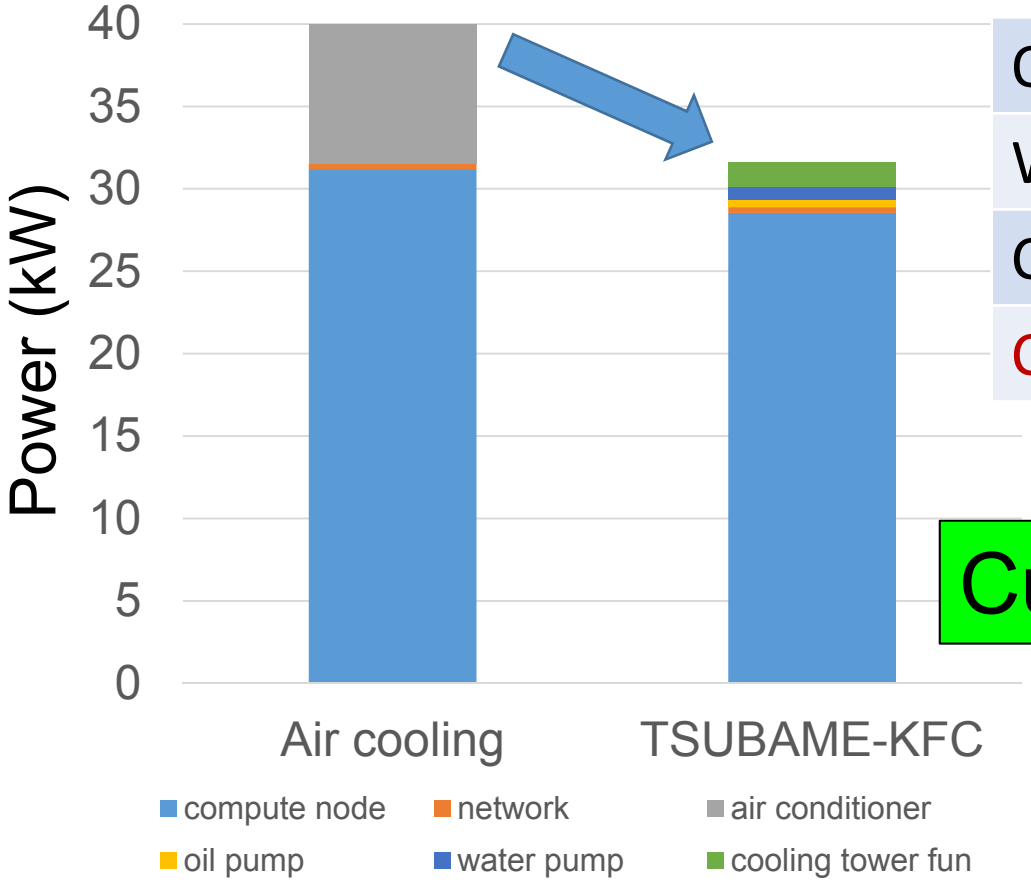
	Air 26 deg. C	Oil 28 deg. C	Oil 19 deg. C
CPU0	50 (43)	40 (36)	31 (29)
CPU1			33 (28)
GPU0	52 (33)	47 (29)	42 (20)
GPU1	59 (35)	46 (27)	43 (18)
GPU2	57 (4	40	33 (18)
GPU3	4 (3	(3	42 (18)
Node Power	749W (228W)	693W (160W)	691W (160W)

26℃ Oil is “cooler” than 28℃ Air !

~8% power reduction!

PUE (Power Usage Effectiveness)

(= Total power / power for computer system)



Oil Pump	0.46 kW
Water Pump	0.75 kW
Cooling Tower Fan	1.48 kW
Cooling Total	2.69 kW

Current PUE = 1.09!!

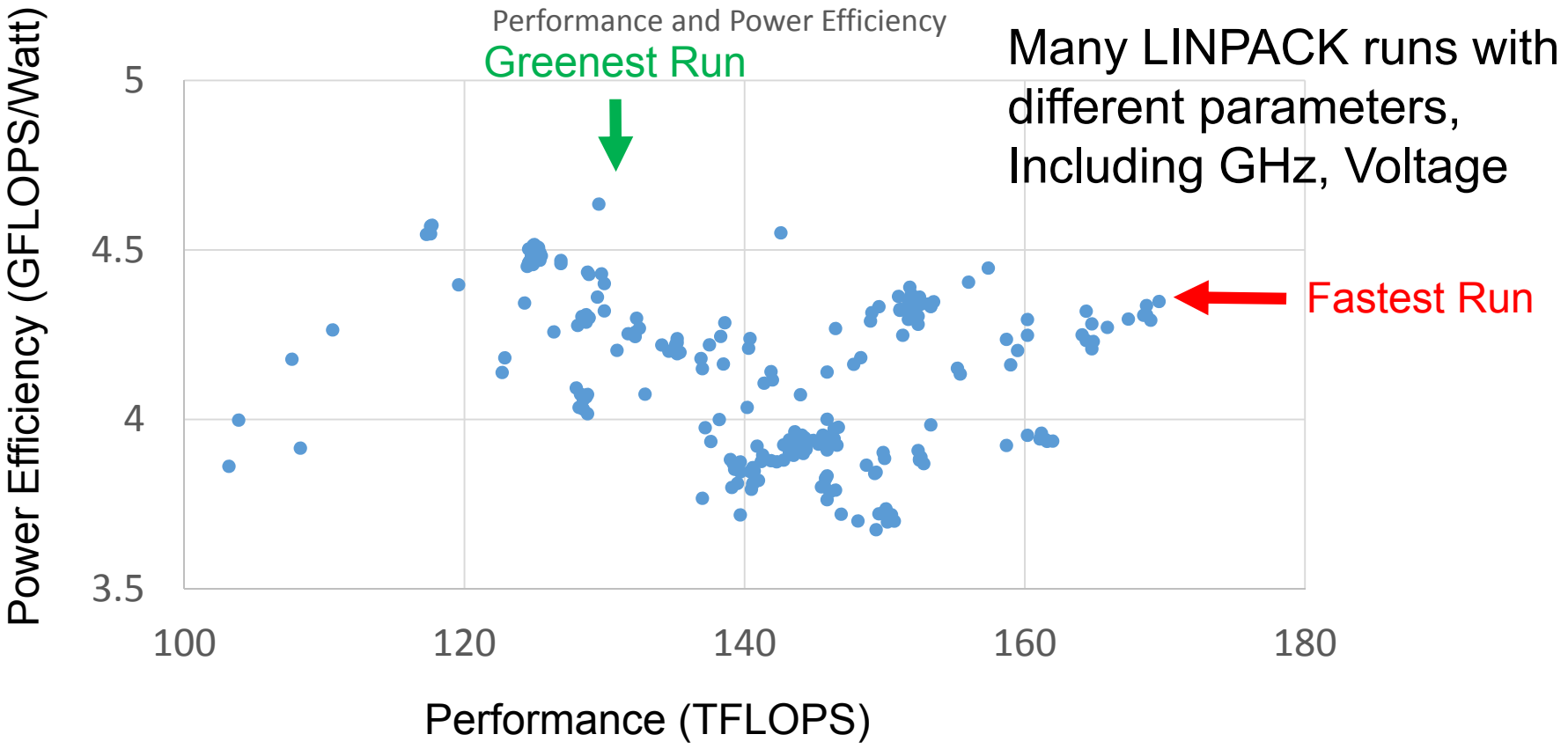
PUE=1.3 in air cooling

Green500 submission

Green500 ranking is determined by

Linpack performance(Flops) / Power consumption(Watt)

- Linpack: Dense matrix benchmark used in Top500
- In current rule, cooling power is NOT included

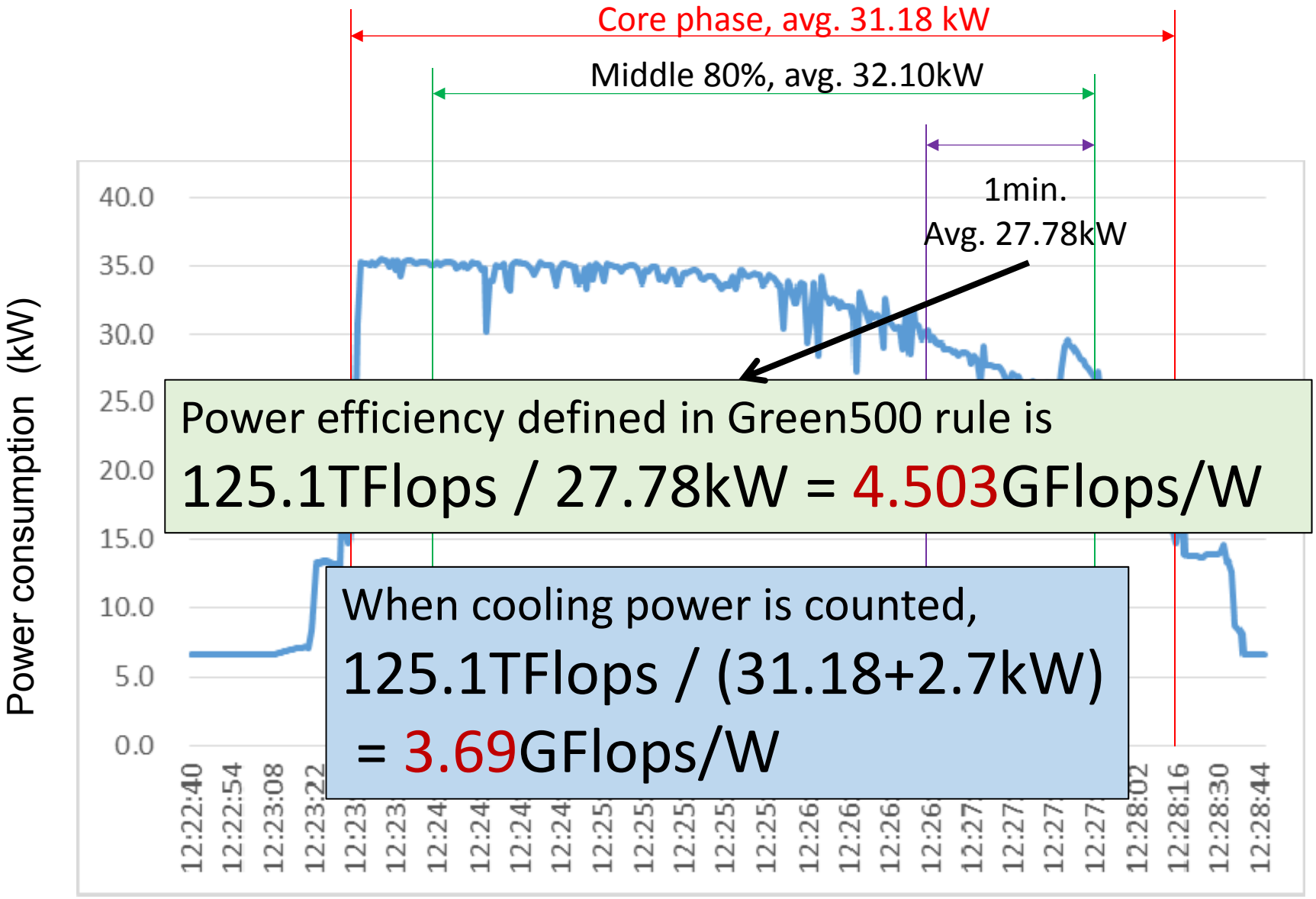


Optimizations for Higher Flops/W

‘Lower’ speed performance leads higher efficiency

- Tuning for HPL parameters
  - Especially, block size (NB), and process grid (P&Q)
- Adjusting GPU clock and voltage
  - The lowest voltage and the highest clock.
- Advantages of hardware configuration
  - GPU:CPU ratio = 2:1
  - Low power Ivy Bridge CPU (this also lower the perf.)
  - Cooling system. No cooling fans. Low temperature.

Power Profile during Linpack benchmark in Nov 13



The Green500 List Nov 2013

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
7	2,702.16	University of Arizona	iDataPlex DX360M4, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR14, NVIDIA K20x	53.62
8	2,629.10	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94
9	2,629.10	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	55.62
10	2,358.69	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2.000GHz, Infiniband FDR, Nvidia K20m	71.01

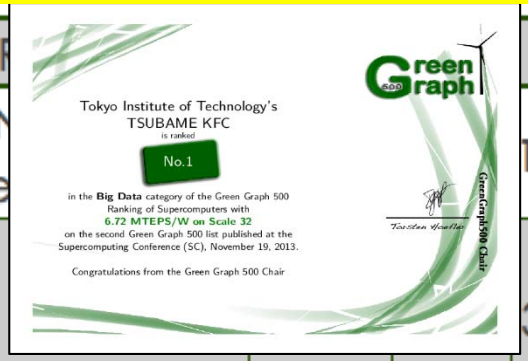
Green Graph500 list on Nov. 2013

- Ranking of power efficiency in Big Data benchmark
- Measures power-efficiency using **TEPS/W** ratio
  - TEPS: Traversed Edges Per Second
- **http://green.graph500.org**

In the **Big Data** category:

Rank	MTEPS/W	Site	Machine	G500 rank	Scale	GTEPS	Nodes
1	6.72	Tokyo Institute of Technology	TSUBAME KFC	47	32	44.01	32
2	5.41	Forsc				348	16384
3	4.42	Argo				328	32768
4	4.35	Tokyo				3.67	1
5	3.55	Lawren				15363	65536
6	1.89	Re Adv Infrastructure				37.66	1
7	0.73	Mayo Clinic	grace	68	31	10.32	64

KFC Got Double Crown in Nov 13!



KFC in Green500

Nov 13	Jun 14	Nov 14
No.1 4.504GF/W	No.1 4.389GF/W	No. 3 4.447GF/W

Latest winners announced in SC14
No.1: L-CSC (GSI Helmholtz center)
No.2: Suiren PEZY-SC (KEK)

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	5,271.81	GSI Helmholtz Center	L-CSC - ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150 Level 1 measurement data available	57.15
2	4,945.63	High Energy Accelerator Research Organization /KEK	Suiren - ExaScaler 32U256SC Cluster, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, PEZY-SC	37.83
3	4,447.58	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	35.39
4	3,962.73	Cray Inc.	Storm1 - Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40m Level 3 measurement data available	44.54
			Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband	

How about Maintenance Cost?

- After operation starts in Sep 13, we added two SSDs into each node in Mar 14
  - To enhance big-data experiments
- We needed to pull up each node from oil!

Details of SSD Installation Time

Procedure	Approx. time
Removing external cables	50s
Pulling up the node	2m10s
Opening the node cover	1m10s
Removing GPUs	1m20s
Installing SATA cables	1m50s
Restoring GPUs	2m20s
Closing the node cover	2m40s
Inserting SSDs into drive bay	1m10s
Submerging the node	2m30s
Installing external cables	40s
Total	16m40s

- Procedures directly related to oil-submersion occupies 4m40s (~28% of time)

Summary

- TSUBAME-KFC: A Ultra Green Supercomputer testbed has been installed
  - Fluid submersion cooling for improving power efficiency
- Further development is required towards 50GF/W
  - TSUBAME1.0(2006): ~0.05GF/W
  - TSUBAME2.0(2010): 1GF/W
  - TSUBAME-KFC(2013): 4.5GF/W (3.7GF/W including cooling)