Power Capping Scheduling on TSUBAME2.5 and Upgrade of TSUBAME-KFC

Toshio Endo, Akira Nukada, Satoshi Matsuoka

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

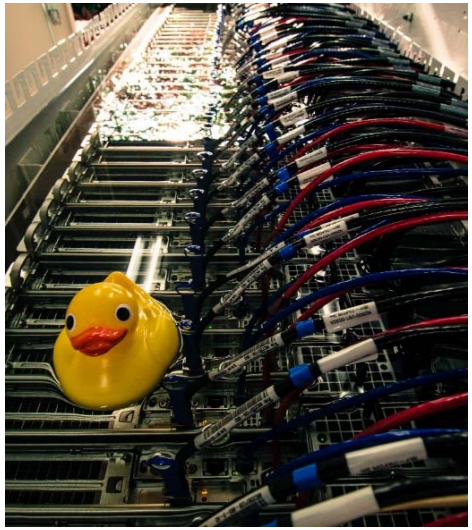
Two Topics

1) Power Capping Scheduling on TSUBAME2.5

- TSUBAME2.5: 5.7PFlops heterogeneous supercomputer
  - Upgraded from TSUBAME2.0 in 2013
- Capping V1 (2011~2013): Static
- Capping V2 (2014~): Dynamic

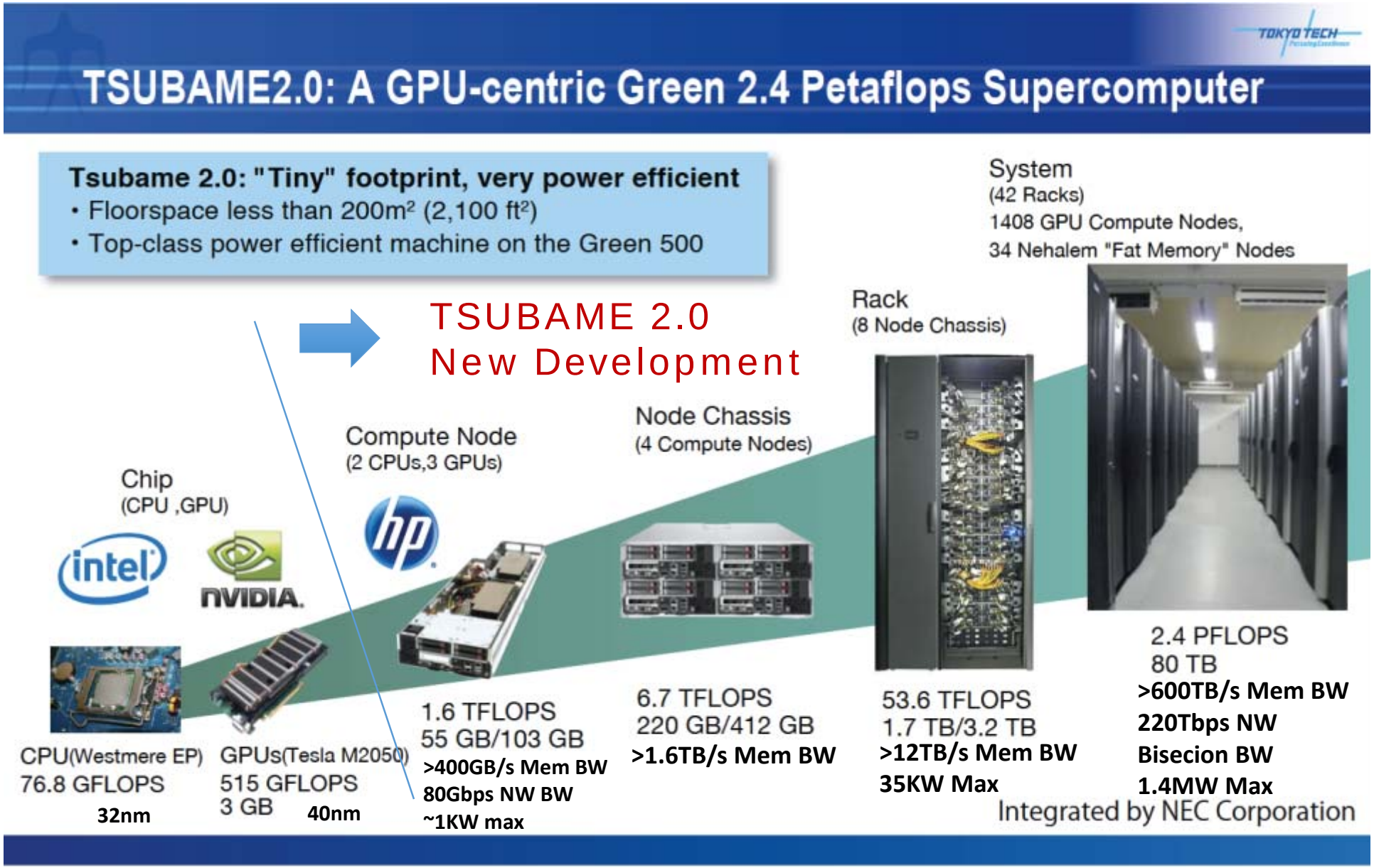
2) Upgrade of TSUBAME-KFC

- TSUBAME-KFC: a prototype of "ultra-green" supercomputer
  - Green No.1 in Nov 2013 & Jun 2014
- Upgraded Tesla K20X GPUs➔ K80
  - The resultant system: TSUBAME-KFC/DL



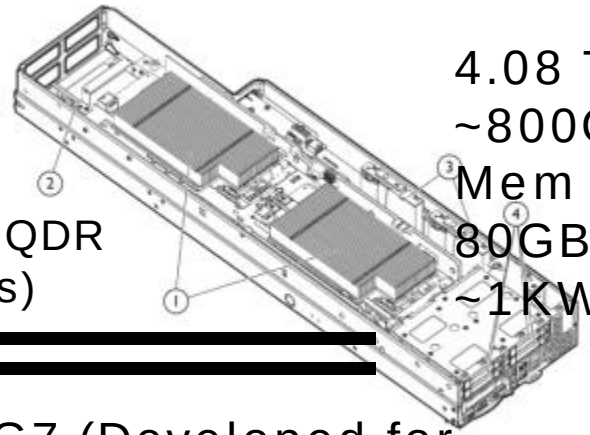
TSUBAME2.0 Nov. 1, 2010

"The Greenest Production Supercomputer in the World"



TSUBAME2.0⇒2.5 Thin Node Upgrade in 2013

Thin Node



Peak Perf.

4.08 Tflops
~800GB/s
Mem BW
80GBps NW
~1KW max

Infiniband QDR
x2 (80Gbps)

HP SL390G7 (Developed for TSUBAME 2.0, Modified for 2.5)
GPU: NVIDIA Kepler K20X x 3
1310GFlops, 6GByte Mem(per GPU)
CPU: Intel Westmere-EP 2.93GHz x2

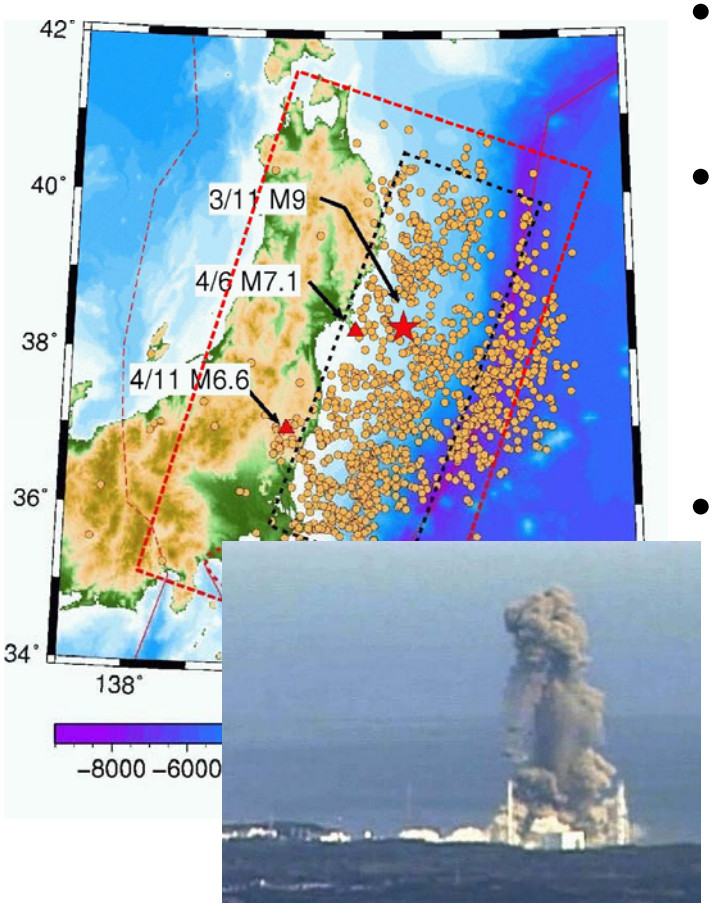
NVIDIA Fermi
M2050
1039/515
GFlops

NVIDIA Kepler
K20X
3950/1310
GFlops

System total: 2.4PFlops → 5.7PFlops

Power Capping Scheduling on T2.5: Motivation

- We suffer from Mar 11, 2011 earthquake (yet)



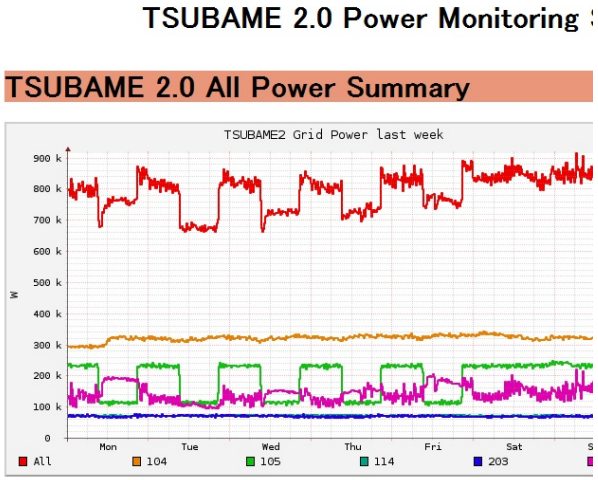
- Power plant accident caused power crisis in the capital area
- Although TSUBAME2 is “green”, it consumes ~10% (~1MW) power of the campus!
- University requested us to reduce power consumption of TSUBAME2 every summer
  → < 800kW in daytime

Power Capping Scheduling V1: Static (2011—2013)

- It was allowable to increase running nodes in nighttime and weekend

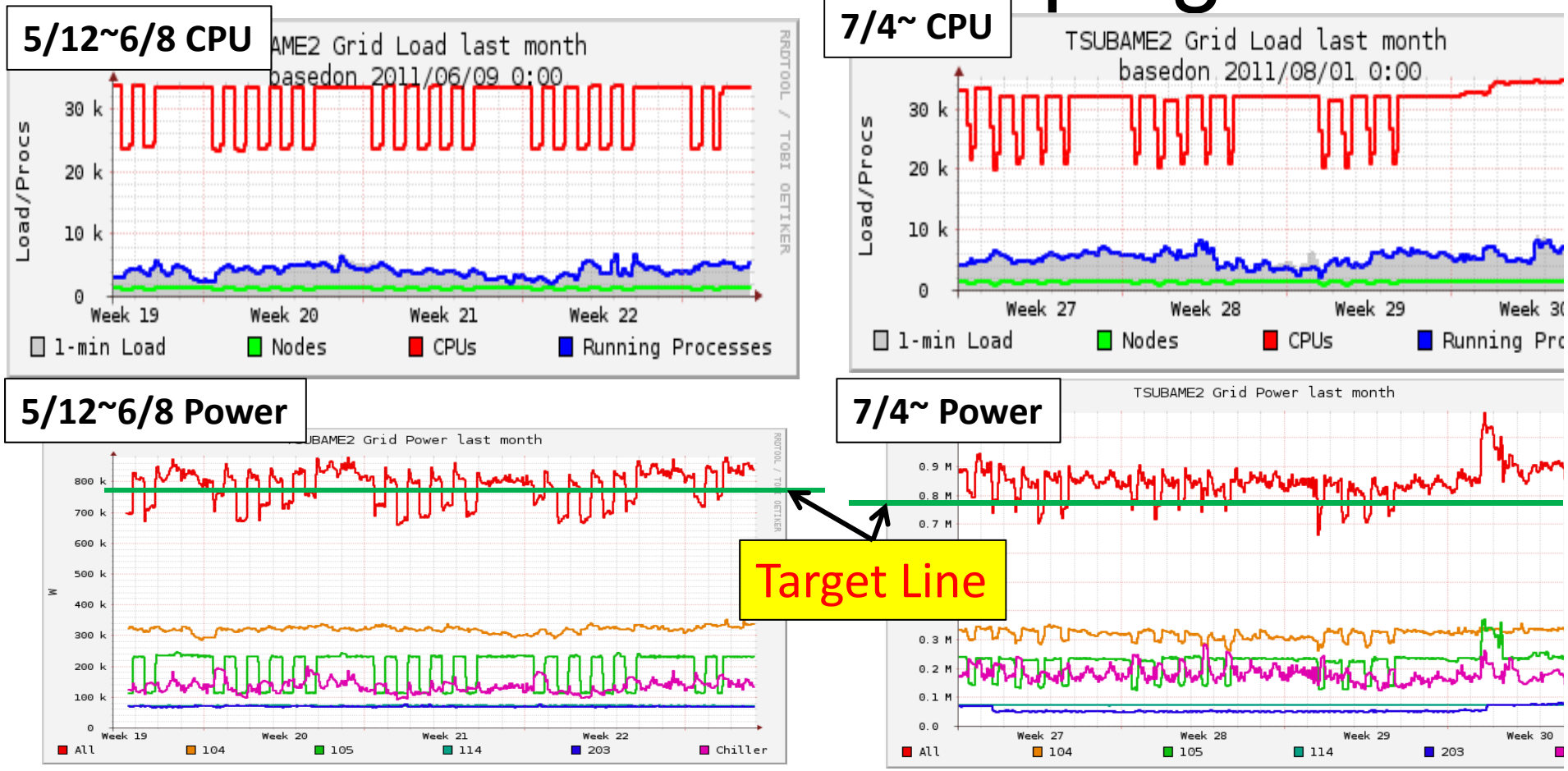
65% operation in daytime

100% operation in nighttime/weekend



- Tokyo-Tech and NEC developed tools to shutdown/ wake up nodes automatically everyday
  - 35% nodes are turned off around 9:00 and on around 19:00
- The rate of “35%” is pre-determined by measurement of power consumption in typical operation
  - The rate is static

Results of Power Capping V1



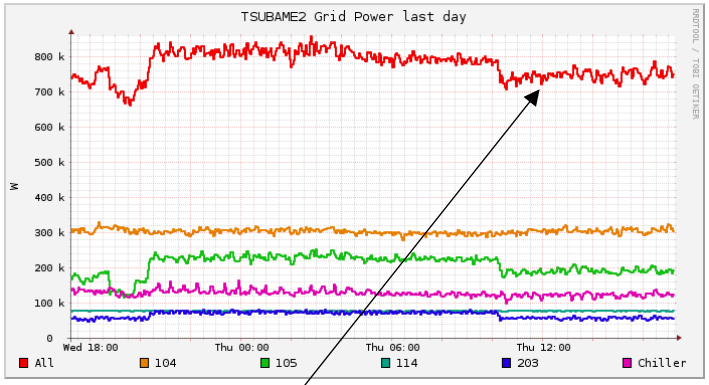
- Successfully capped the power below the target
- But, it is sometimes too conservative

→ Users largely suffer from shortage of computing resources

Power Capping Scheduling V2: Dynamic (2014—)

During daytime,

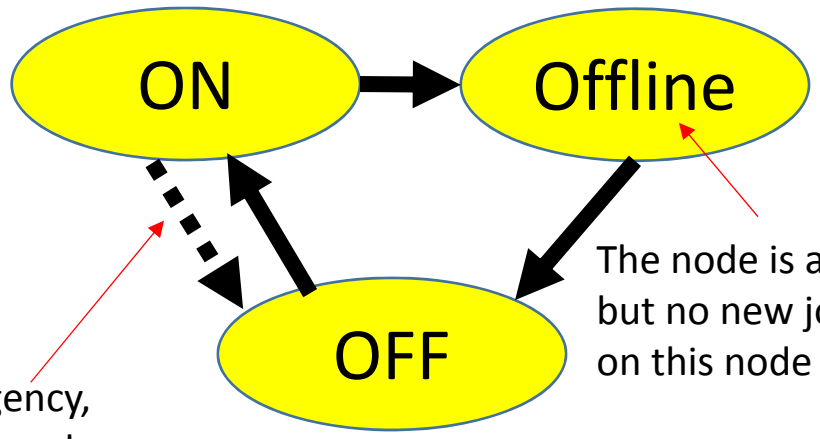
(1) Monitor system power consumption



Is current power sufficiently lower than the specified cap?
Or are we in a critical zone?

(2) Control the number of running nodes dynamically & automatically

Transition of modes of each node
(simplified)

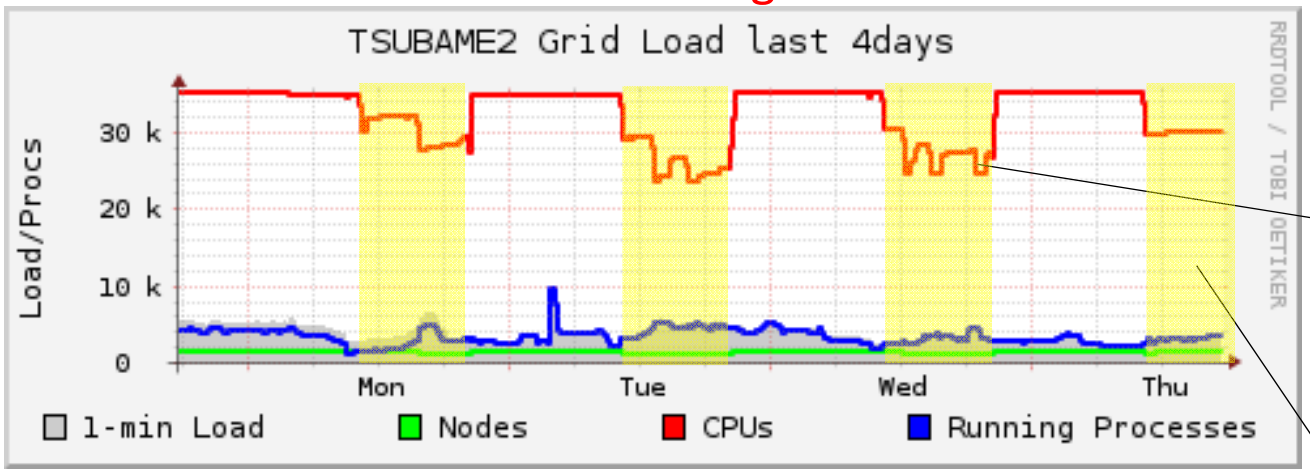


The node is awaken, but no new job starts on this node

Only in emergency, we stop jobs&nodes

Results of Power Capping V2

The number of working CPU cores

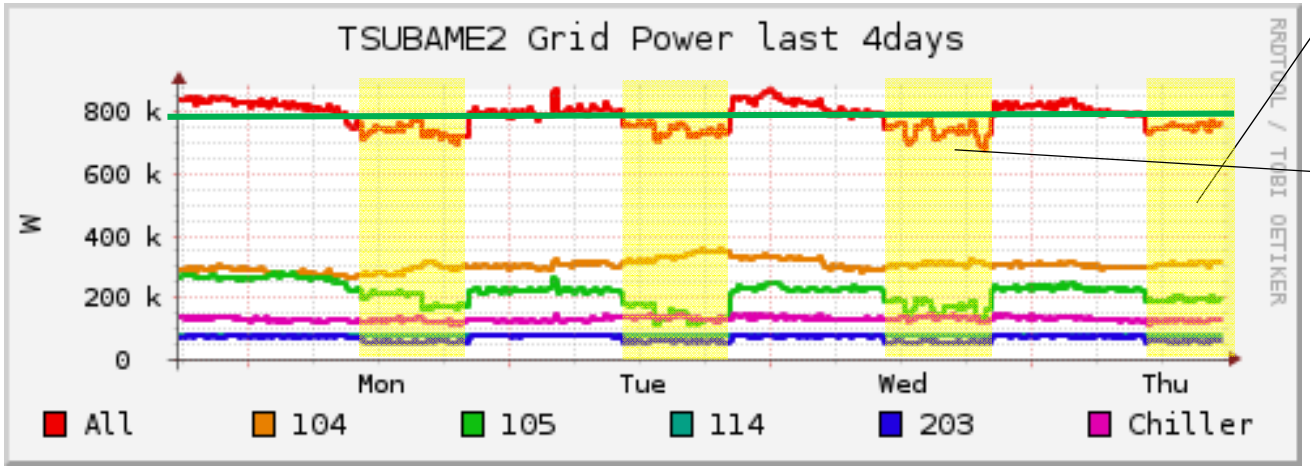


The number of nodes are controlled dynamically

# of available nodes are better than fixed 65%

Daytime

Power Consumption



The power is capped by 800 kW successfully

Real-time graphs are at *http://mon.g.gsic.titech.ac.jp/*

TSUBAME-KFC and its Upgrade

TSUBAME-KFC: Ultra-Green Supercomputer Testbed

TSUBAME-KFC

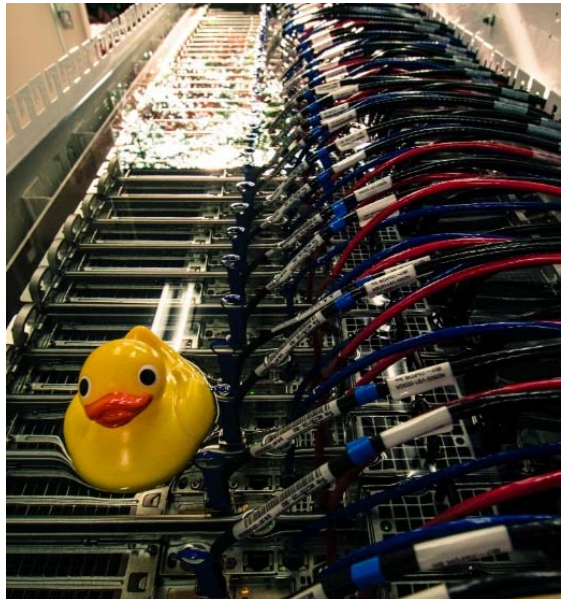
or **K**epler **F**luid **C**ooling

= (Hot Fluid Submersion Cooling

+ Outdoor Air Cooling

+ Highly Dense Accelerated Nodes)

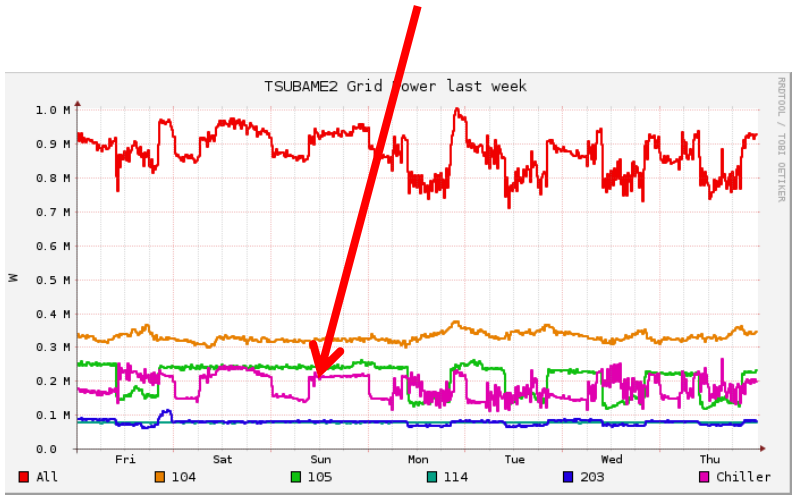
in a 20-feet Container



Motivation of TSUBAME-KFC

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

- Making 9℃ water is expensive

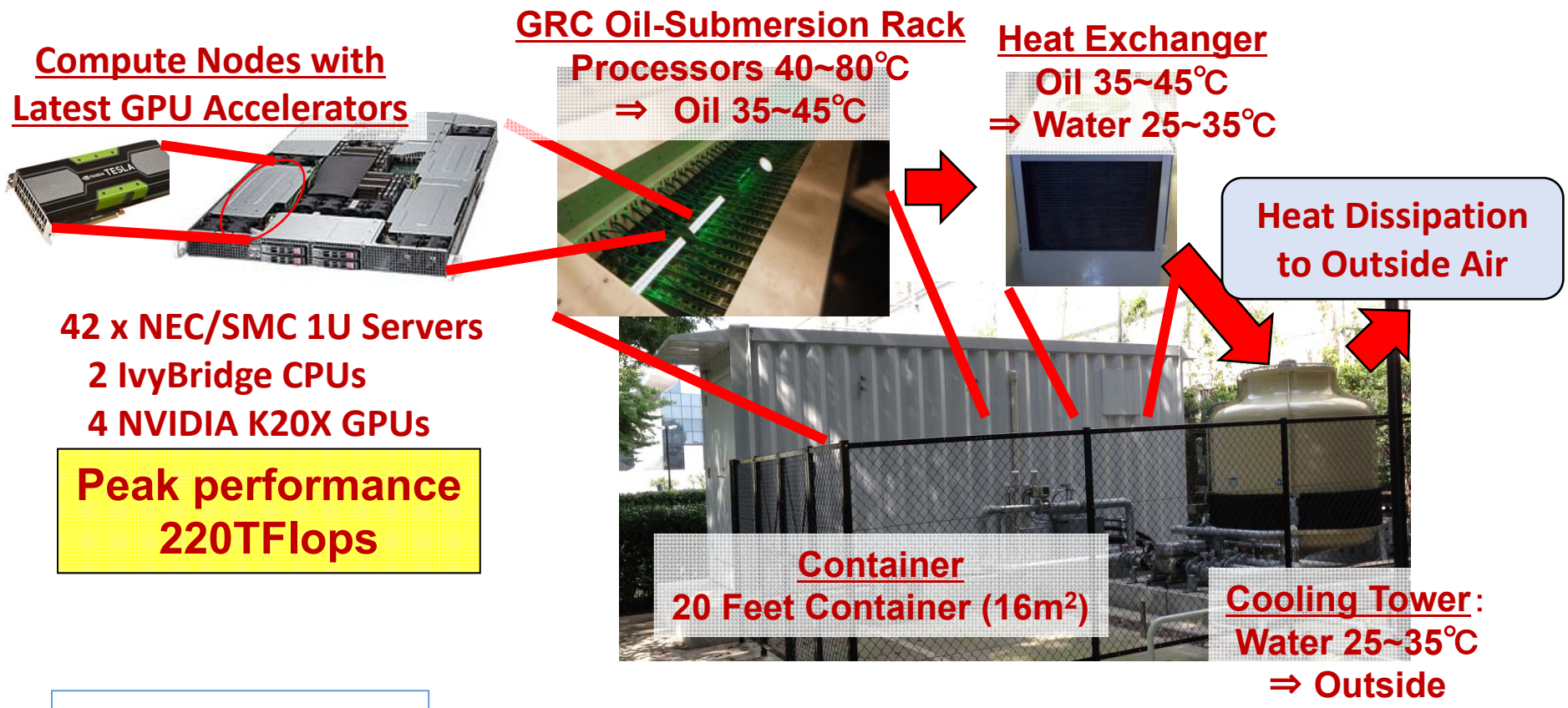


- To reduce cooling power
  - We should avoid making chilled water

  →Warm/hot liquid cooling
  - Designing water pipe / plate is expensive
  - Control of coolant speed is more difficult

  → Liquid submersion cooling

TSUBAME-KFC: Ultra-Green Supercomputer Testbed as of 2013



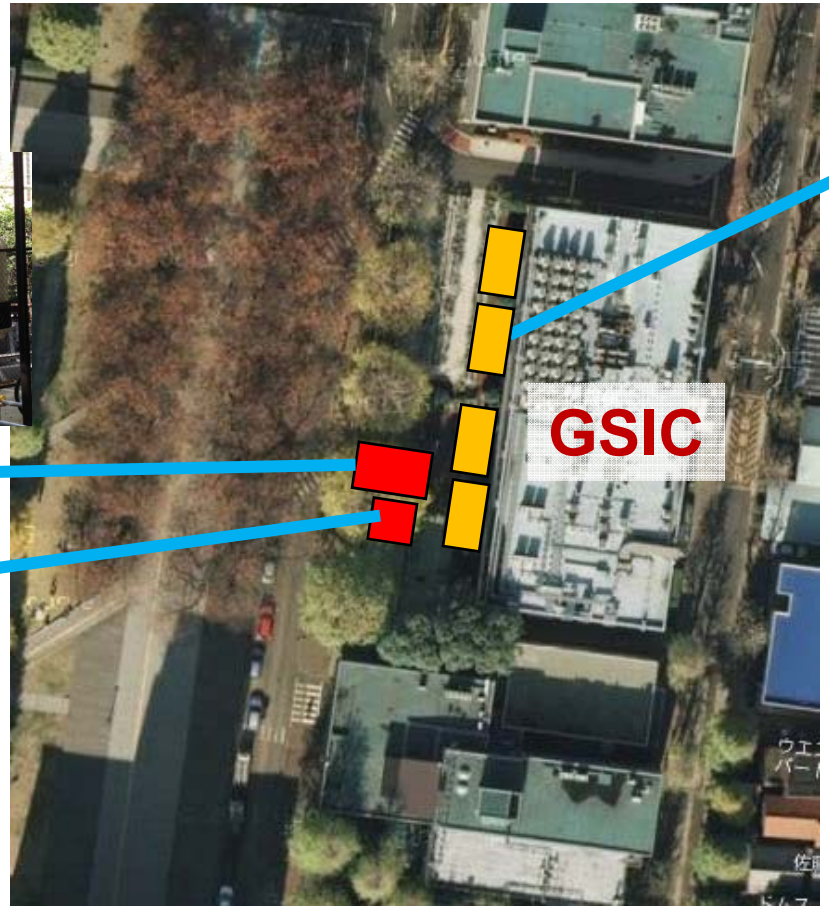
Achievement

- 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



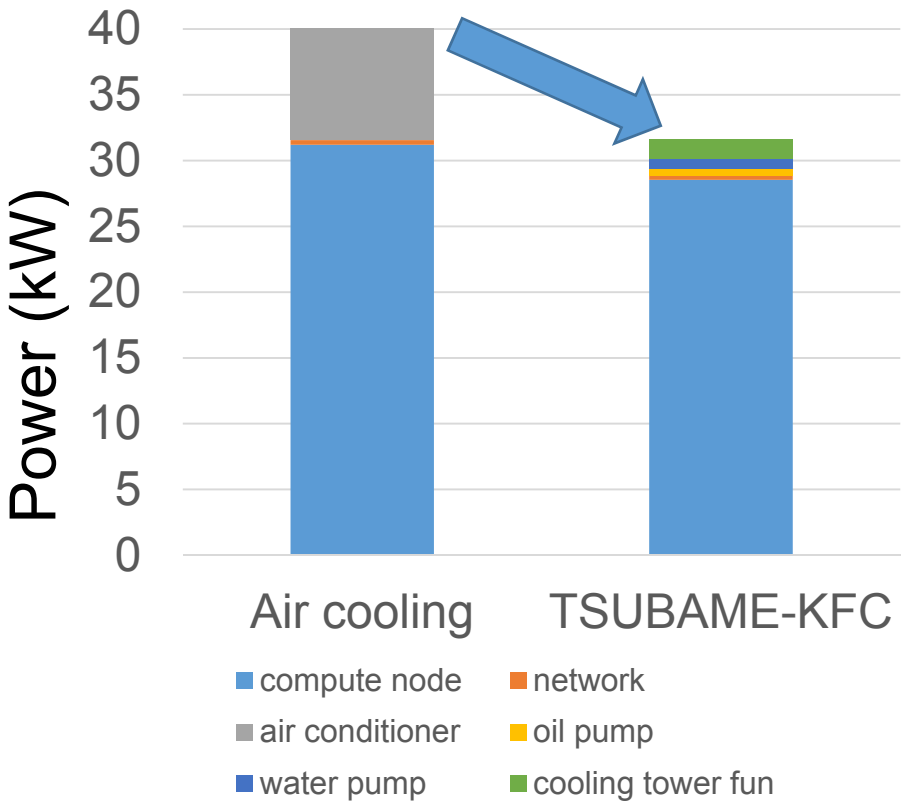
KFC Container

Cooling tower

Chillers for TSUBAME2

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

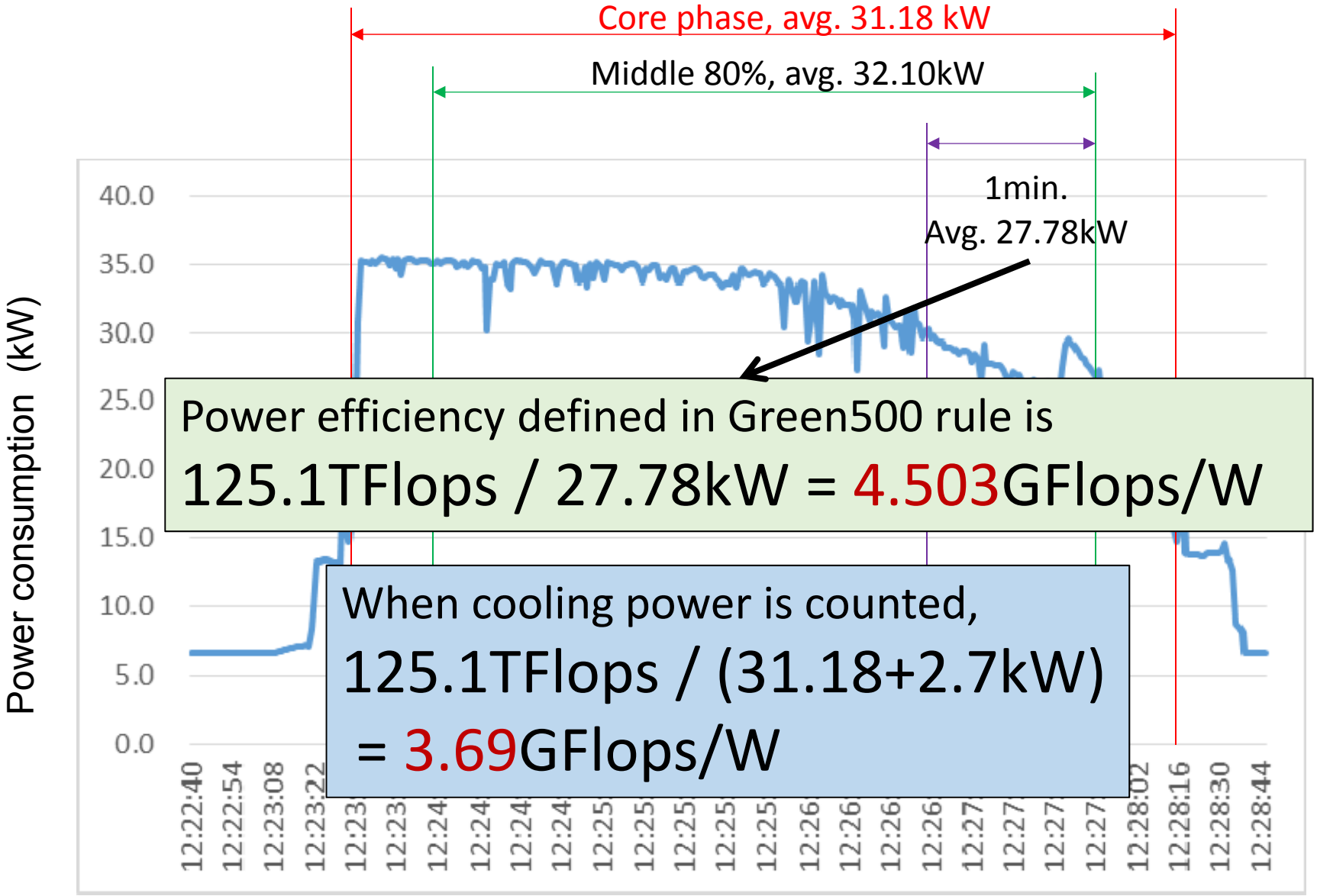
Not only cooling power, nodes power is reduced for

- Removing cooling fans inside
- Lower chip temperature

PUE = 1.09!!

Air cooling case is based on a guess with assumption of PUE=1.3

Power Profile during Linpack benchmark in Nov 2013



Power efficiency defined in Green500 rule is

125.1TFlops / 27.78kW = 4.503GFlops/W

When cooling power is counted,

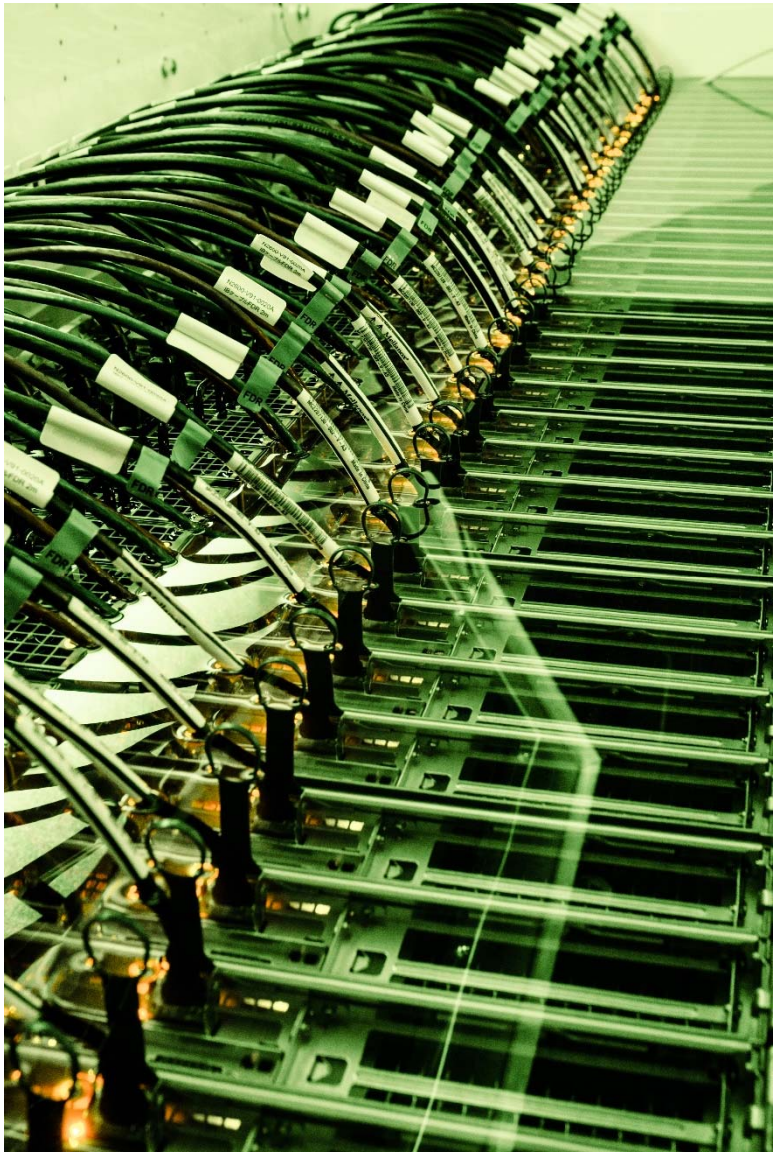
125.1TFlops / (31.18+2.7kW)

= 3.69GFlops/W

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	Xeon E5-2630v2 6C 2.600GHz, Infiniband	52.62
3	3,517.84	Center for Computational University of Tsukuba	M Cluster, Intel Xeon E5-2680v2 10C IA K20x	78.77
4	3,185.91	Swiss National Supercor Centre (CSCS)	2670 8C 2.600GHz, Aries interconnect , ble	1,753.66
5	3,130.95	ROMEO HPC Center - C Ardenne	tel Xeon E5-2650v2 8C 2.600GHz,	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

42 TSUBAME-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)
- 64GB Main Memory

System Total Performance

SP	DP
669 TF	224 TF

Recent Upgrade:
42 TSUBAME-KFC/DL 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~~
- **4X** NVIDIA Tesla K80 **GPU**

HCA

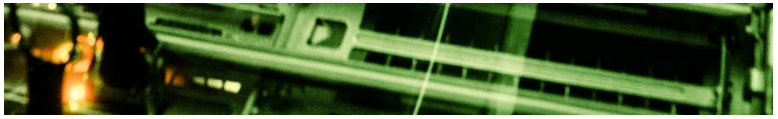
-480)

Please watch results in Top500 and Green500 announced in SC!

SP	DP
669 TF	224 TF

	SP	DP
Base clock (560MHz)	951 TF	318 TF
Boosted clock (875MHz)	1476 TF	493 TF

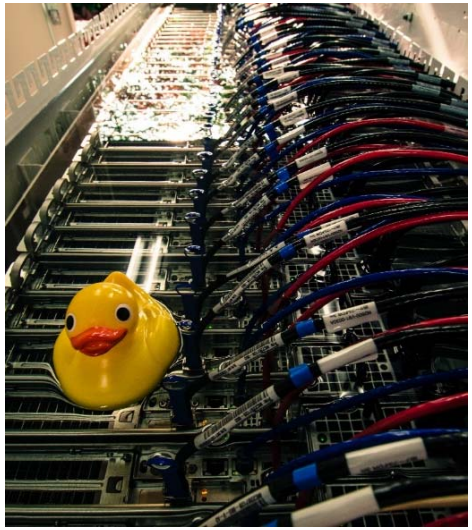
With support of NVIDIA, NEC



Two Topics I Talked

1) Power Capping Scheduling on TSUBAME2.5

- TSUBAME2.5: 5.7PFlops heterogeneous supercomputer
  - Upgraded from TSUBAME2.0 in 2013
- Capping V1 (2011~2013): Static
- Capping V2 (2014~): Dynamic

2) Upgrade of TSUBAME-KFC

- TSUBAME-KFC: a prototype of "ultra-green" supercomputer
  - Green No.1 in Nov 2013 & Jun 2014
- Upgraded Tesla K20X GPUs➔ K80
  - The resultant system: TSUBAME-KFC/DL