# *TSUBAME-KFC* : Ultra Green Supercomputing Testbed

Toshio Endo, Akira Nukada, Satoshi Matsuoka

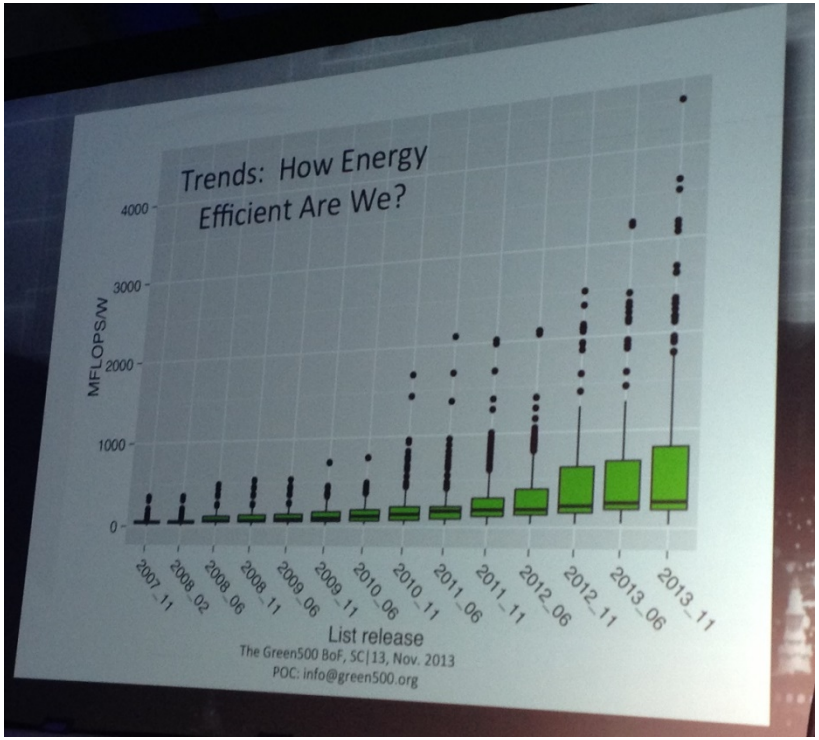
TSUBAME-KFC is developed by

**GSIC, Tokyo Institute of Technology**

**NEC, NVIDIA, Green Revolution Cooling,**

**SUPERMICRO, Mellanox**

# Performance/Watt is the Issue

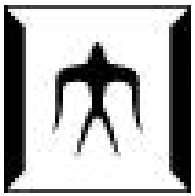



From Wu Feng's presentation
@Green500 SC13 BoF

- Realistic supercomputer centers are limited by power upper bound of 20MW
- In order to achieve Exaflops systems, technologies enabling 50GFlops/W is keys
- Around 2020

# 3 Years Ago

TSUBAME 2.0 achieved 0.96GFlops/W





- 2nd in Nov2010 Green500 (3rd in fact)
- Greenest Production Supercomputer award

Towards TSUBAME3.0 (2015 or 16),

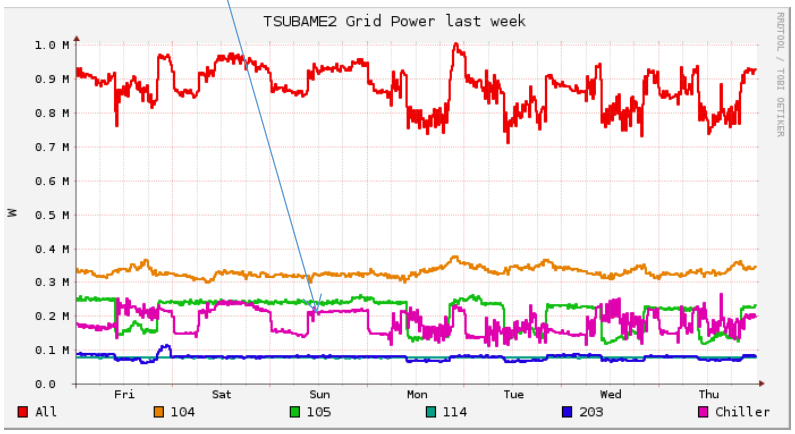
We should be Greener, Greener, Greener!!

# How Do We Make IT Green?

- Reducing computers power
  - Improvement of processors, process shrink
  - Node designs with richer many-core accelerators
  - System designs that reduces communication bottlenecks
  - Software technologies that efficiently utilize accelerators
- Reducing cooling power
  - Liquid cooling is keys due to higher heat capacity than air
  - We should avoid making chilled water

  → Fluid submersion cooling

In TSUBAME2, Chillers use ~25% power of the system



## ***TSUBAME-KFC***:
## Ultra-Green Supercomputer Testbed

TSUBAME-KFC

or Kepler Fluid Cooling

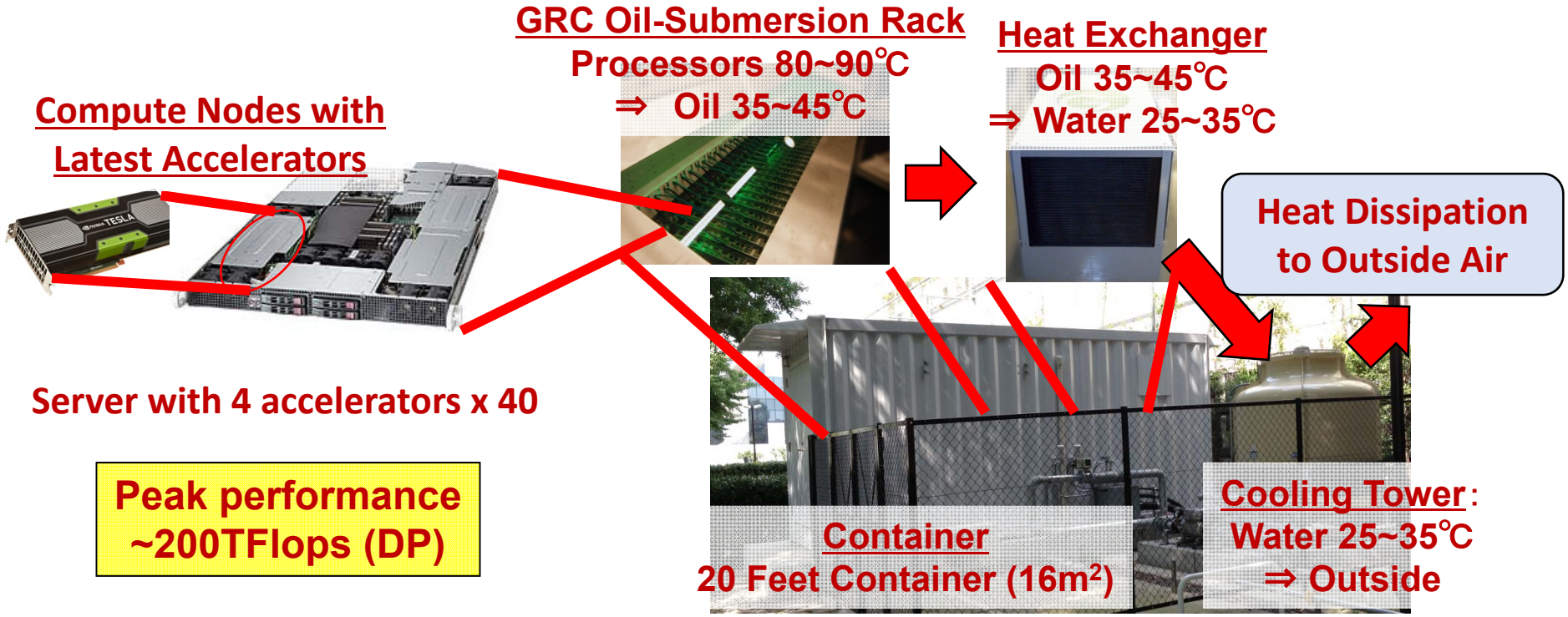
= (Hot Fluid Submersion Cooling

+ Outdoor Air Cooling

+ Highly Dense Accelerated Nodes)

in a 20-feet Container

# *TSUBAME-KFC*: Ultra-Green Supercomputer Testbed (as of planning)

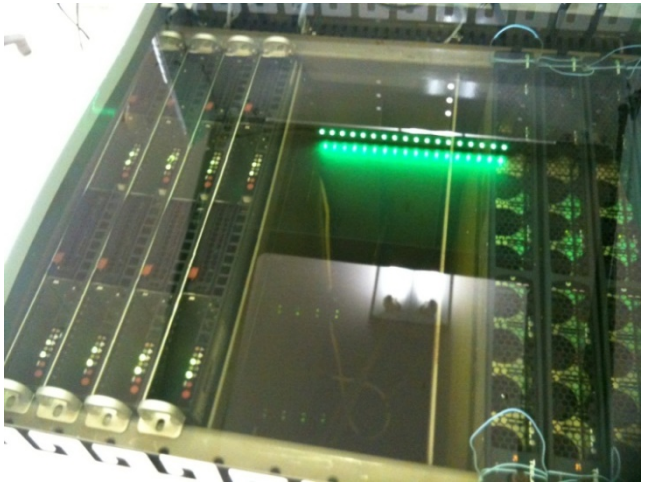



## Target

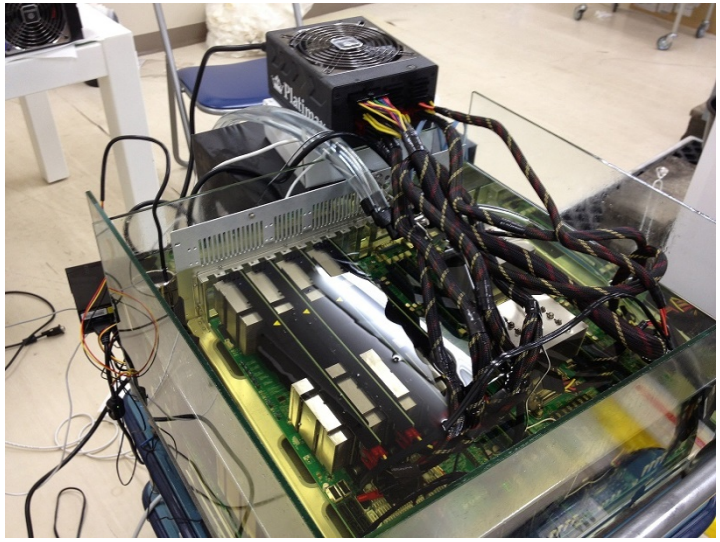
- Worlds' top class power efficiency, >3GFlops/W
- Avarage PUE of 1.05 (Cooling power is ~5% of system power)

**R&D Towards TSUBAME3.0. with >10GFlops/W!**

# We Started Small



Winter 2011: Green Revolution Cooling 13U evaluation kit

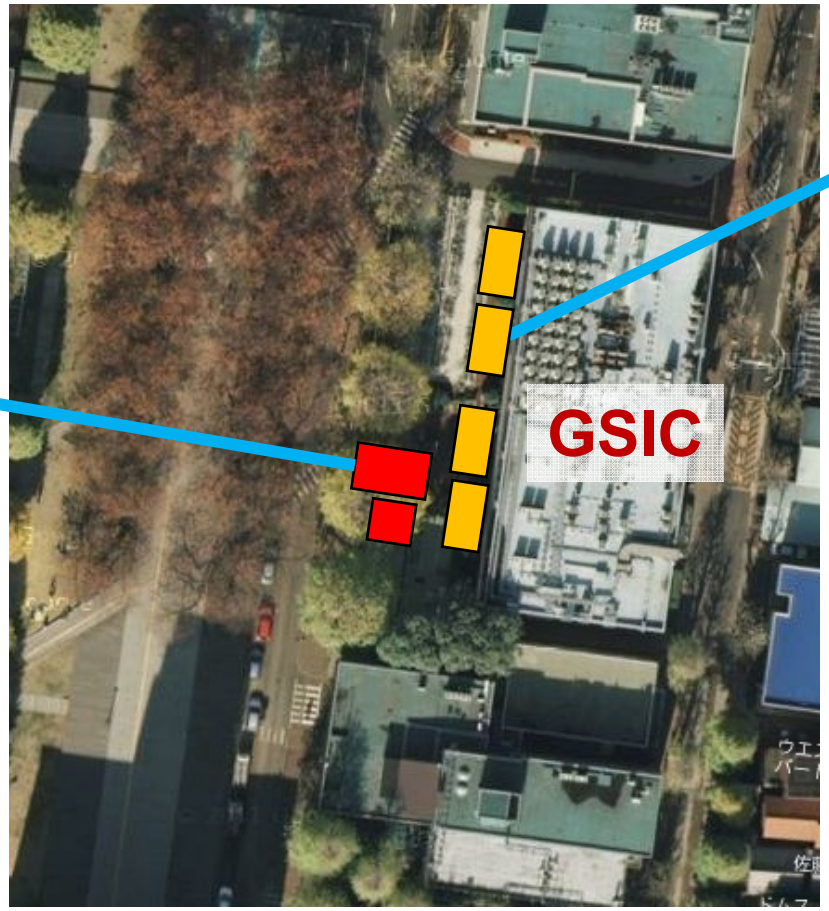


Summer 2012:
A self-made oil tank
with 4 K10 GPU machine

# Installation Site

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

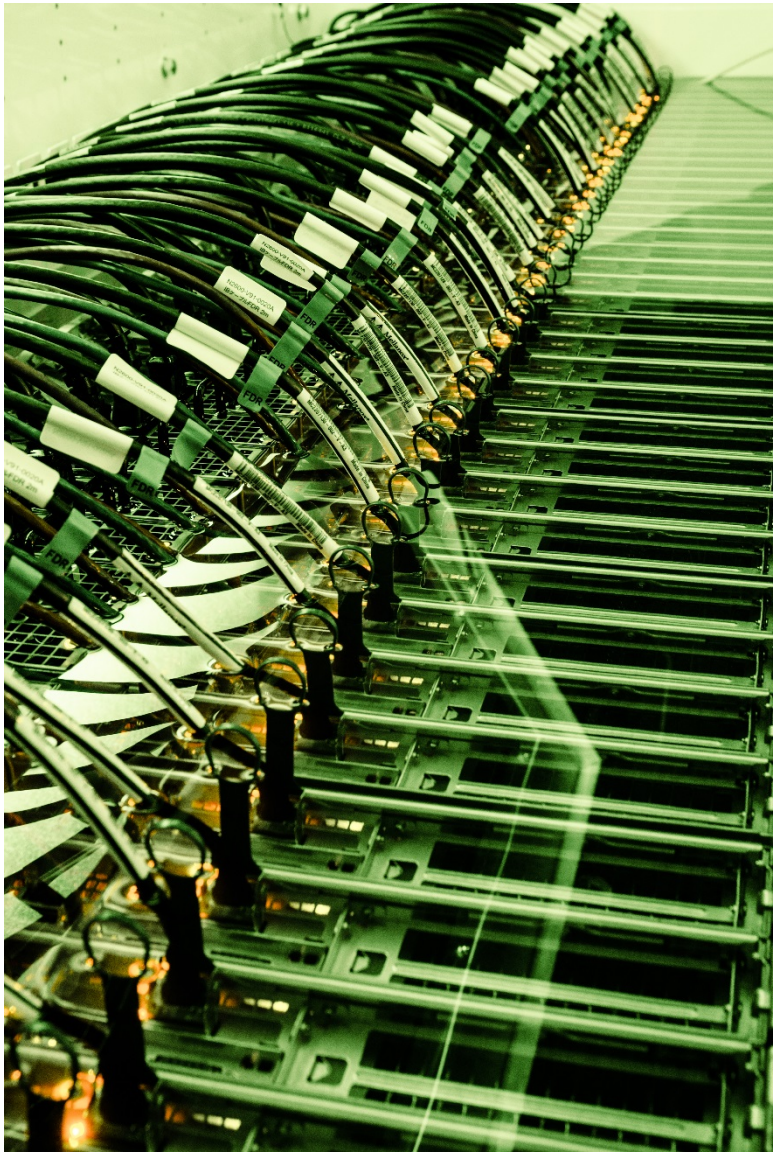
Flash point of oil must be >250℃,
Otherwise it is a hazardous material under the Fire Defense Law in Japan.

Still the officer at the fire station requested us to follow the safety regulations of hazardous material: sufficient clearance around the oil, etc.

# 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
- 1X 120GB SATA SSD

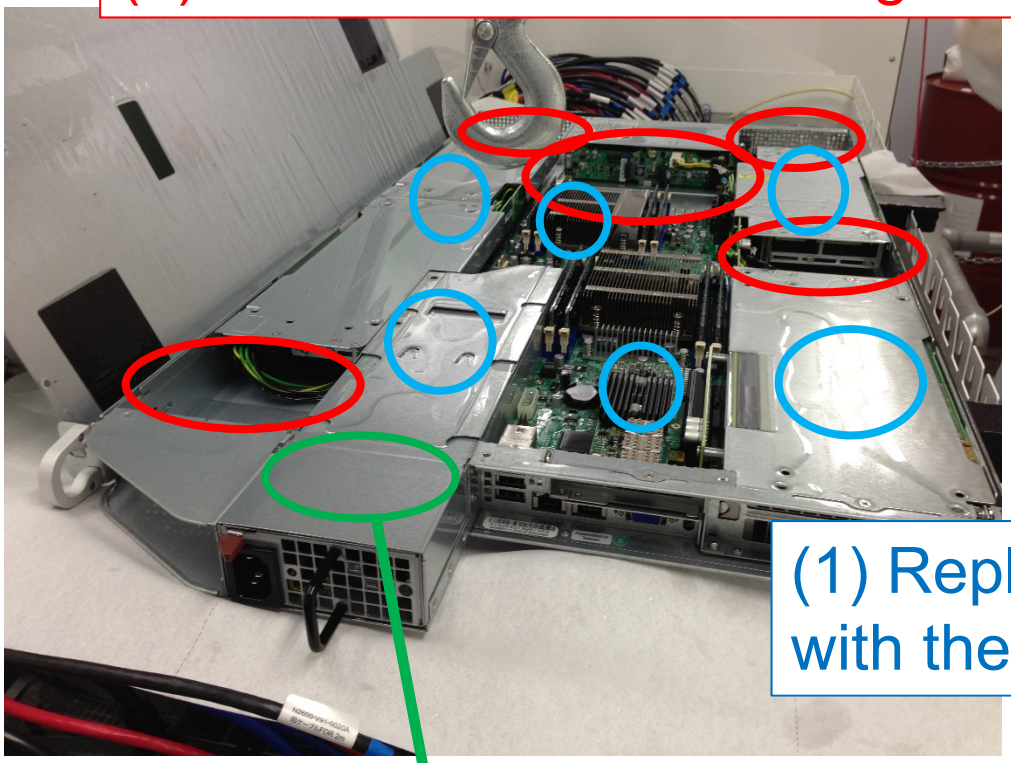
Peak Performance (DP)

| | |
|---|---|
| Single Node | 5.26 TFLOPS |
| System (40 nodes) | 210.61 TFLOPS |

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

# Modification to Compute Nodes

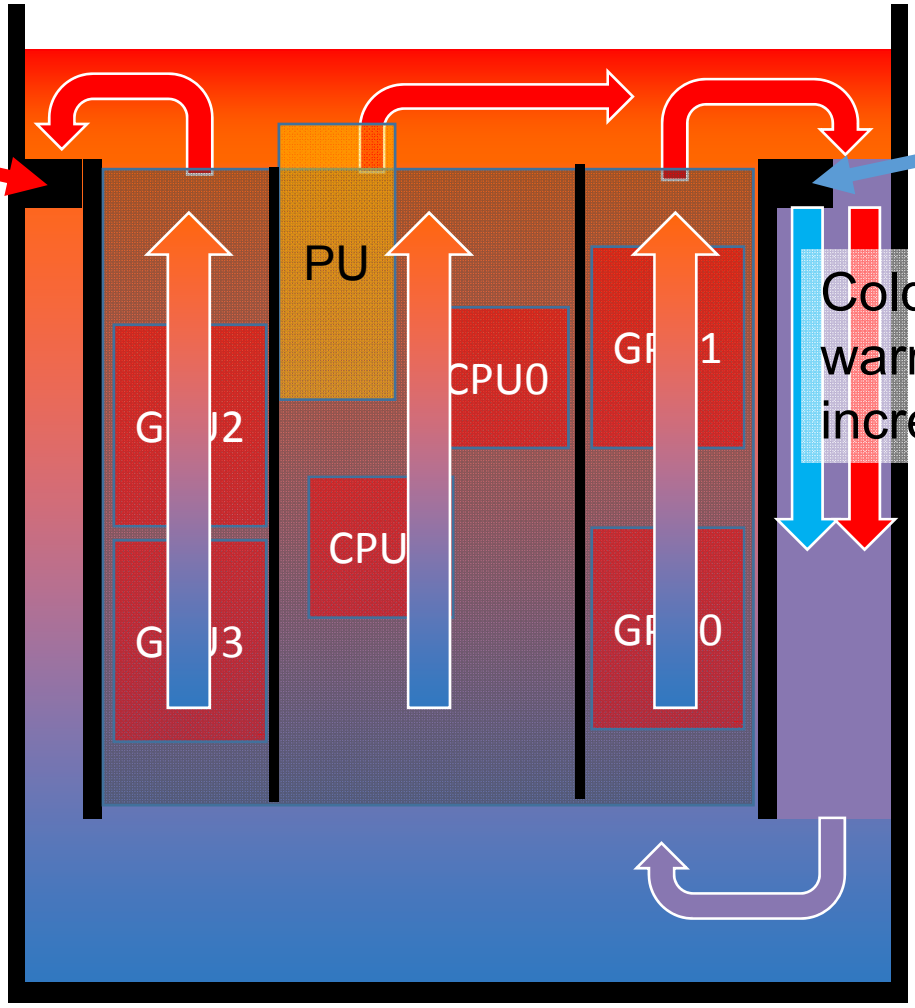
(2) Removed twelve cooling fans



(1) Replace thermal grease with thermal sheets

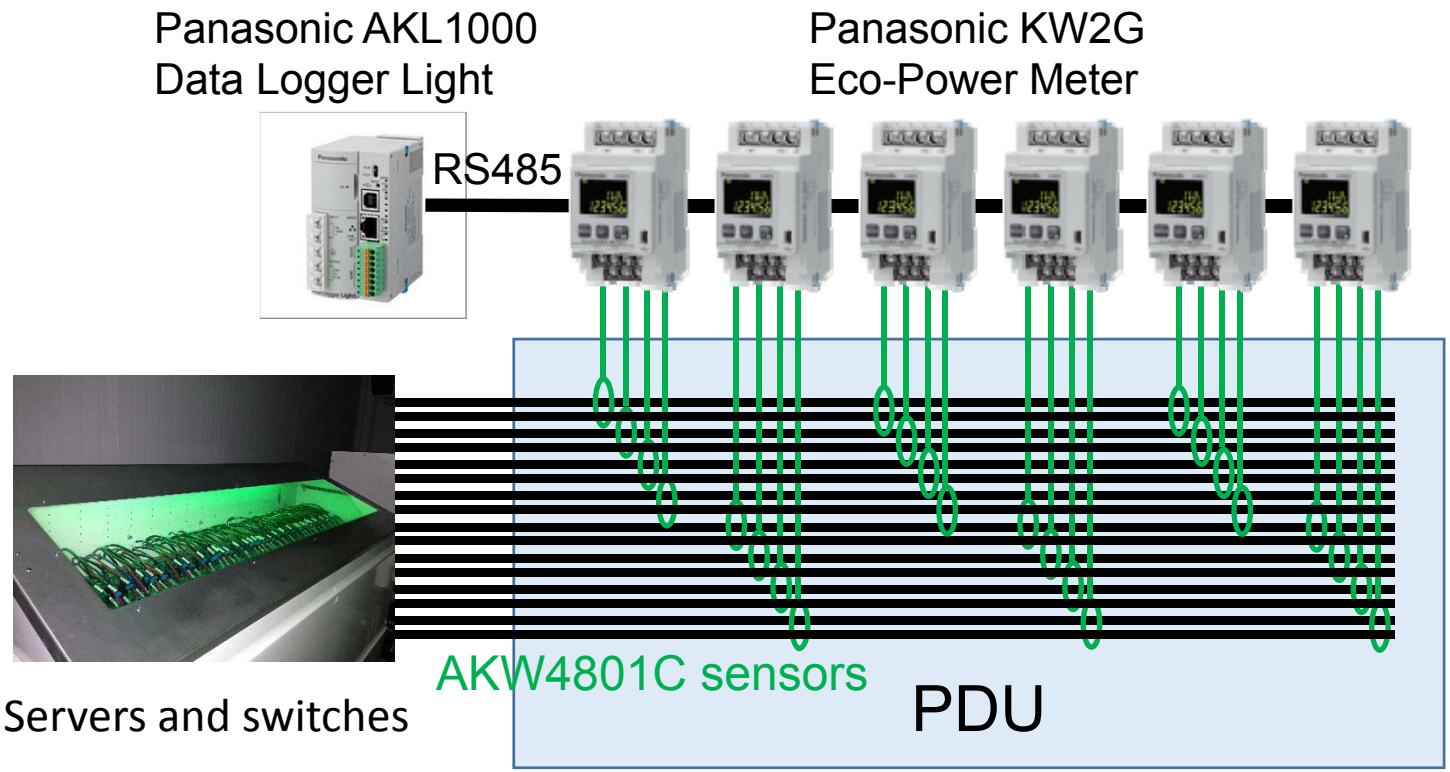
(3) Update firmware of power unit to operate with cooling fan stopped.

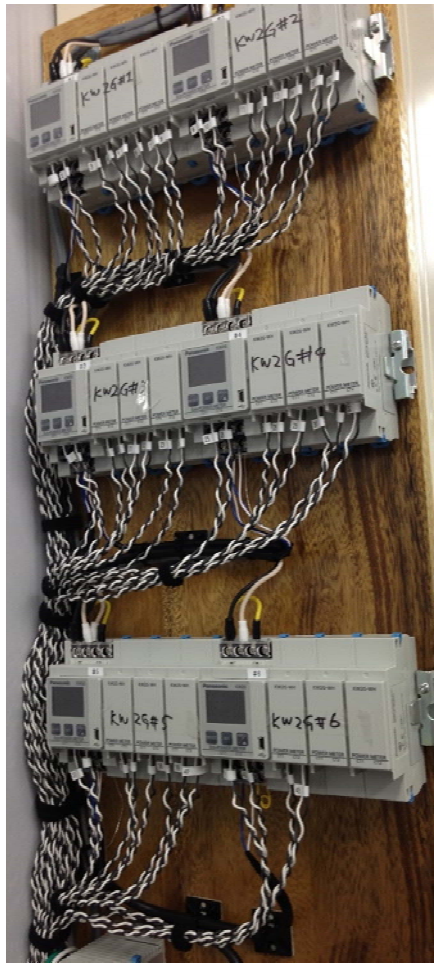
# GRC CarnotJet Fluid-Submersion Rack



# Power Measurement

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

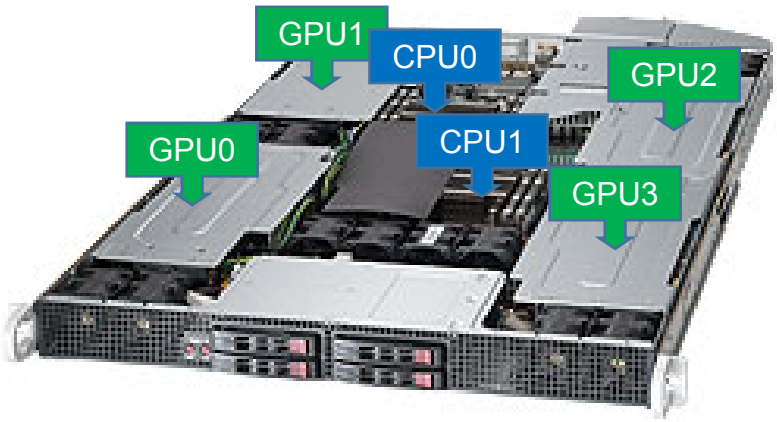


# Effects of Outdoor Environment

| | Rainy<br>Oct. 29th 17pm | Cloudy<br>Oct. 30th 17pm | Clear<br>Oct. 31th 17pm |
|---|---|---|---|
| Oil tank top | 25.7 + 28.0 C | 27.0 + 29.4 C | 25.4 + 27.4 C |
| Oil out | 24.2 C | 23.3 C | 23.5 C |
| Exchange in | 18.0 C | 19.3 C | 17.8 C |
| Exchange out | 18.9 C | 19.9 C | 18.5 C |
| Oil pump power | 572W | 566W | 555W |
| Outside air | 14.8 C | 19.7 C | 19.8 C |
| Outside air dew point | 15.2 CDP | 15.9 CDP | 11.7 CDP |
| Humidity | 99% | 75% | 56% |
| Water temp | 14.8 C | 16.8 C | 14.9 C |
| | | | |

# 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 )

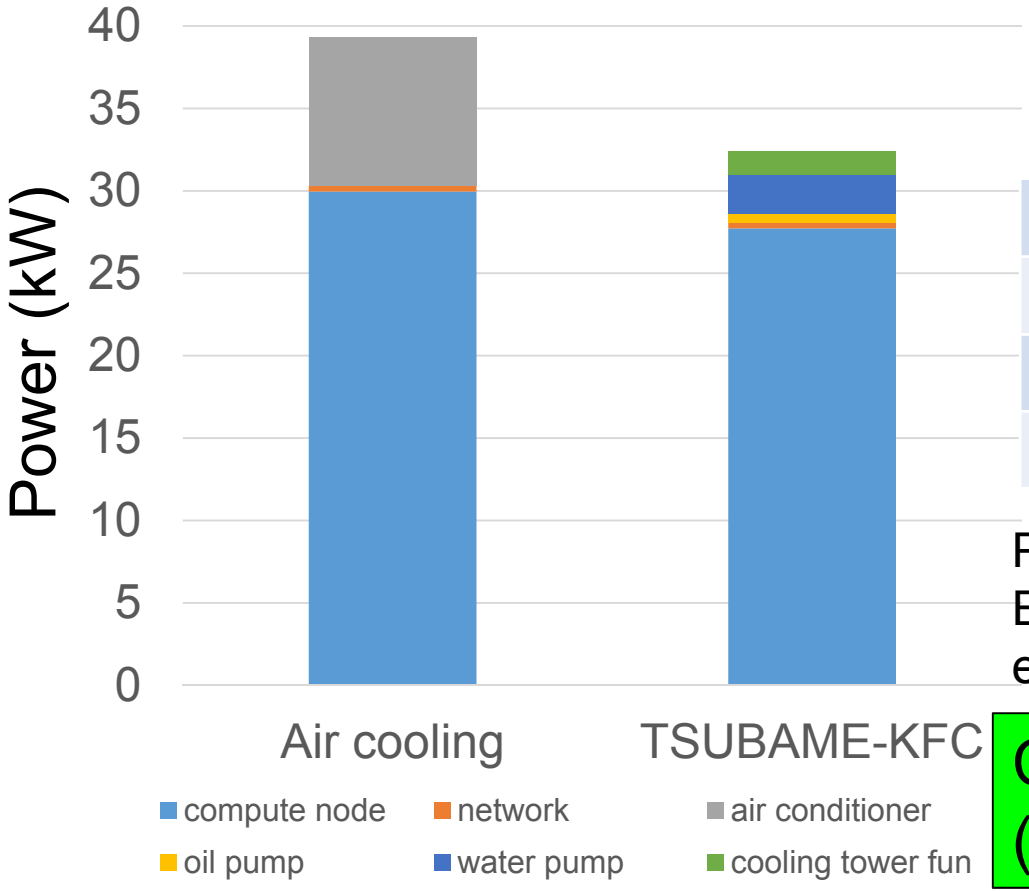
| | Air 26 deg. C | Oil 28 deg. C | Oil 19 deg. C |
|---|---|---|---|
| CPU0 | 50 (43) | 40 (36) | 31 (29) |
| CPU1 | | | 33 (28) |
| GPU0 | 52 (33) | 47 (29) | 42 (20) |
| GPU1 | 59 (35) | 46 (27) | 43 (18) |
| GPU2 | 57 (4 | 40 | 33 (18) |
| GPU3 | 4 (3 | (3 | 42 (18) |
| Node Power | 749W (228W) | 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 (60%) | 0.53 kW |
| Water Pump | 2.40 kW |
| Cooling Tower Fan | 1.40 kW |
| Total | 4.33 kW |

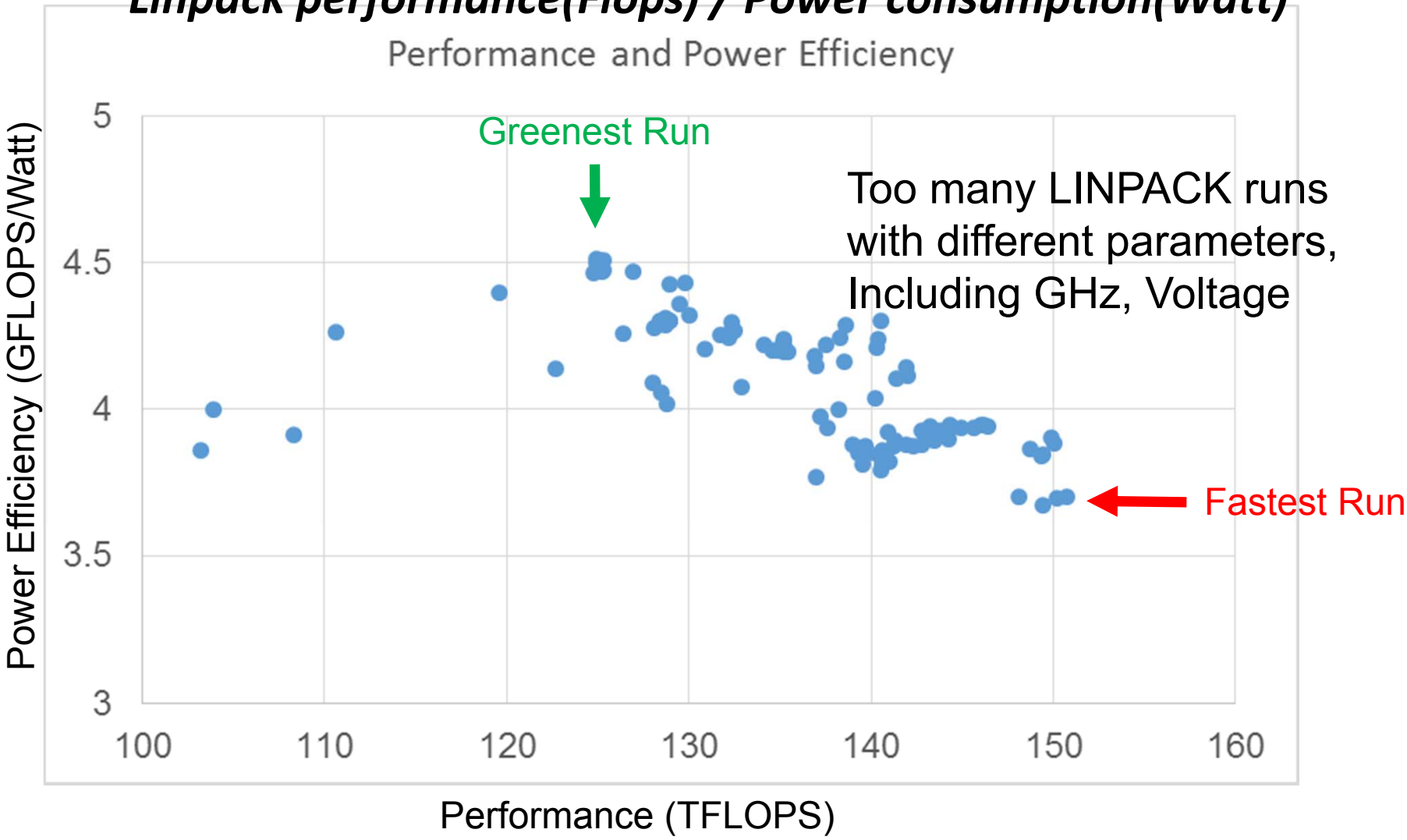
Power for cooling is basically constant. Especially water pump is higher than expected

Current PUE = 1.15
(1.068 based on air-cooling)

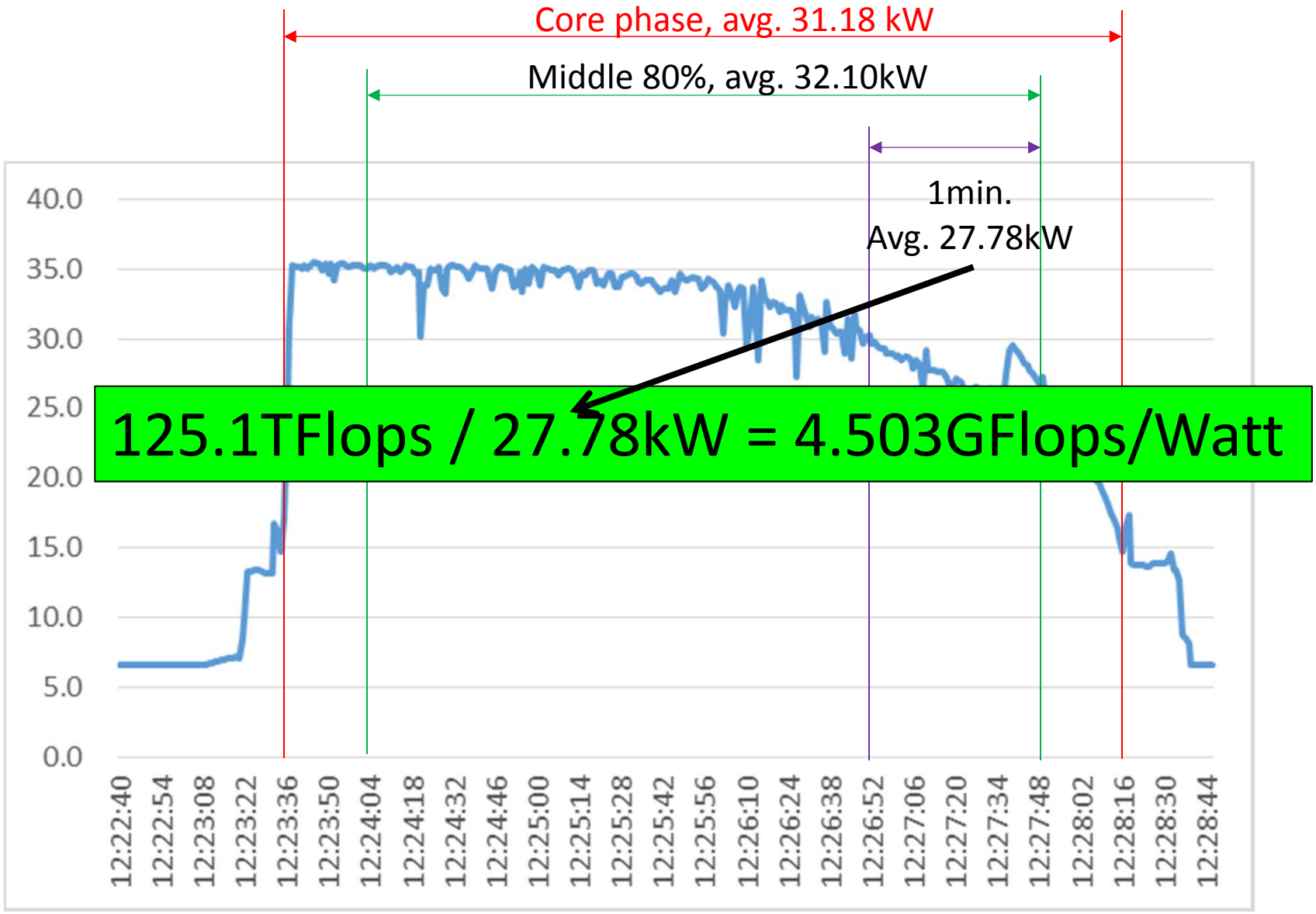
PUE=1.3 in air cooling

# Green500 submission

***Green500 ranking is determined by***
***Linpack performance(Flops) / Power consumption(Watt)***



# Power Profile during Linpack benchmark



# 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
  - Available GPU clocks (MHz):

    614 (best), 640, 666, 705, 732 (default), 758, 784

and 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.

# The Green500 List Nov 2013

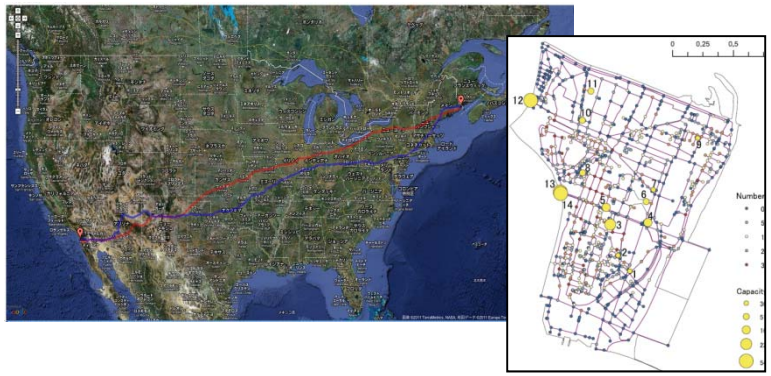
| Green500 Rank | MFLOPS/W | Site* | Computer* | Total Power (kW) |
|---|---|---|---|---|
| 1 | 4,503.17 | GSIC Center, Tokyo Institute of Technology | TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x | 27.78 |
| 2 | 3,631.86 | Cambridge University | 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<br>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 |

# Graph500 Benchmark

http://www.graph500.org

- **New Graph Search Based Benchmark for Ranking Supercomputers**
- BFS (Breadth First Search) from a single vertex on a static, undirected Kronecker graph with average vertex degree edgegactor (=16).
- Evaluation criteria: TEPS (Traversed Edges Per Second), and problem size that can be solved on a system, minimum execution time.

**US road network**
**24 million vertices & 58 million edges**



**Cyber-security**
**15 billion log entries / day**


**Neuronal network @ Human Brain Project**
**89 billion vertices & 100 trillion edges**

*Image: Illustration by Mirko Ilic*

# Green Graph500 list on Nov. 2013

- Measures power-efficient using **TEPS/W** ratio
- Results on various system such as **TSUBAME-KFC Cluster**
- **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** | Forschungszentrum Julich (FZJ) | JUQUEEN | 3 | 38 | 5848 | 16384 |
| 3 | **4.42** | Argonne National Laboratory | DOE/SC/ANL Mira | 2 | 40 | 14328 | 32768 |
| 4 | **4.35** | Tokyo Institute of Technology | EBD-RH5885v2 | 96 | 30 | 3.67 | 1 |
| 5 | **3.55** | Lawrence Livermore National Laboratory | DOE/NNSA/LLNL Sequoia | 1 | 40 | 15363 | 65536 |
| 6 | **1.89** | Research Center for Advanced Computing Infrastructure | altix | 50 | 30 | 37.66 | 1 |
| 7 | **0.73** | Mayo Clinic | grace | 68 | 31 | 10.32 | 64 |

# *KFC Got Double Crown!*