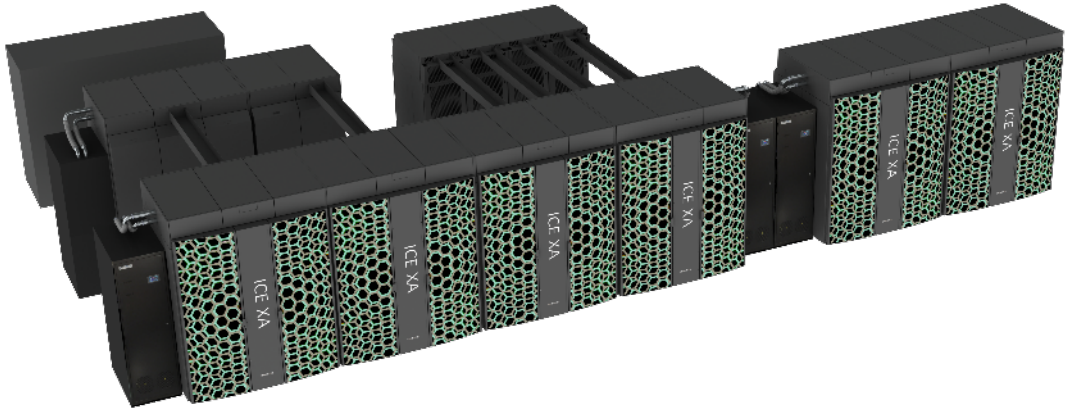
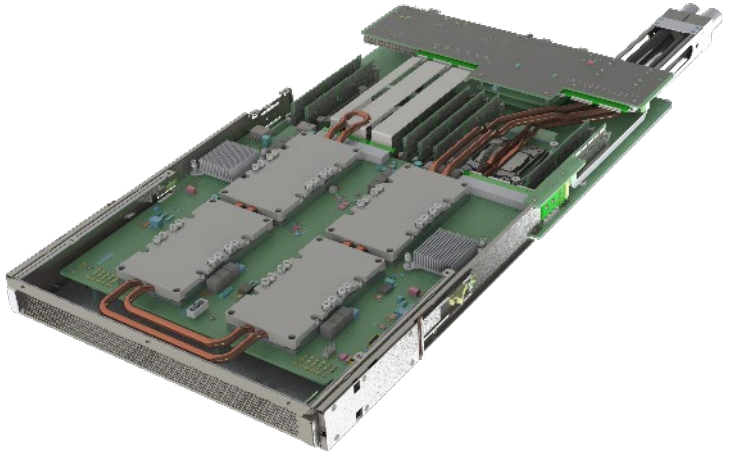
Activity Report from Tokyo Tech: Energy Efficiency of TSUBAME3.0

Toshio Endo

GSIC, Tokyo Institute of Technology





TSUBAME Supercomputer Series

2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019

TSUBAME1

TSUBAME1.0
Opteron+ClearSpeed, 85TF Peak

TSUBAME1.1
ClearSpeed/ Storage added

TSUBAME1.2
680 NVIDIA S1070 added

TSUBAME2.0
Westmere + 4224 M2050, 2.4PF Peak

TSUBAME2.5
Upgraded to K20X
5.7PF Peak

TSUBAME2

experiment

TSUBAME-KFC → KFC/DL

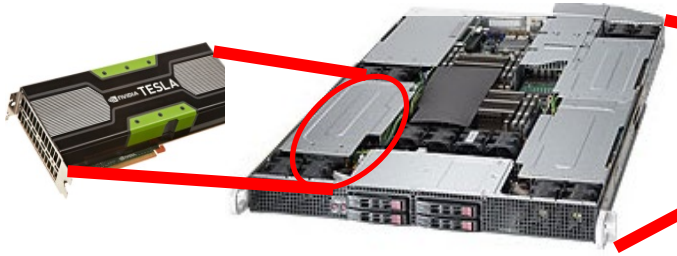
Testbed with oil-immersion cooling

TSUBAME3.0
Broadwell + 2160 P100, 12PF Peak

TSUBAME3

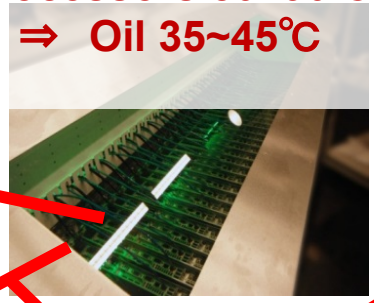
TSUBAME-KFC: Ultra-Green Supercomputer Testbed with Warm Liquid Cooling (2013—2019)

Compute Nodes with K20X → K80 Accelerators



Server with 4 accelerators x 40 are submerged into oil

GRC Oil-Submersion Rack
Processors 80~90℃
⇒ Oil 35~45℃



Heat Exchanger
Oil 35~45℃
⇒ Water 25~35℃

Heat Dissipation to Outside Air

Container
20 Feet Container (16m²)

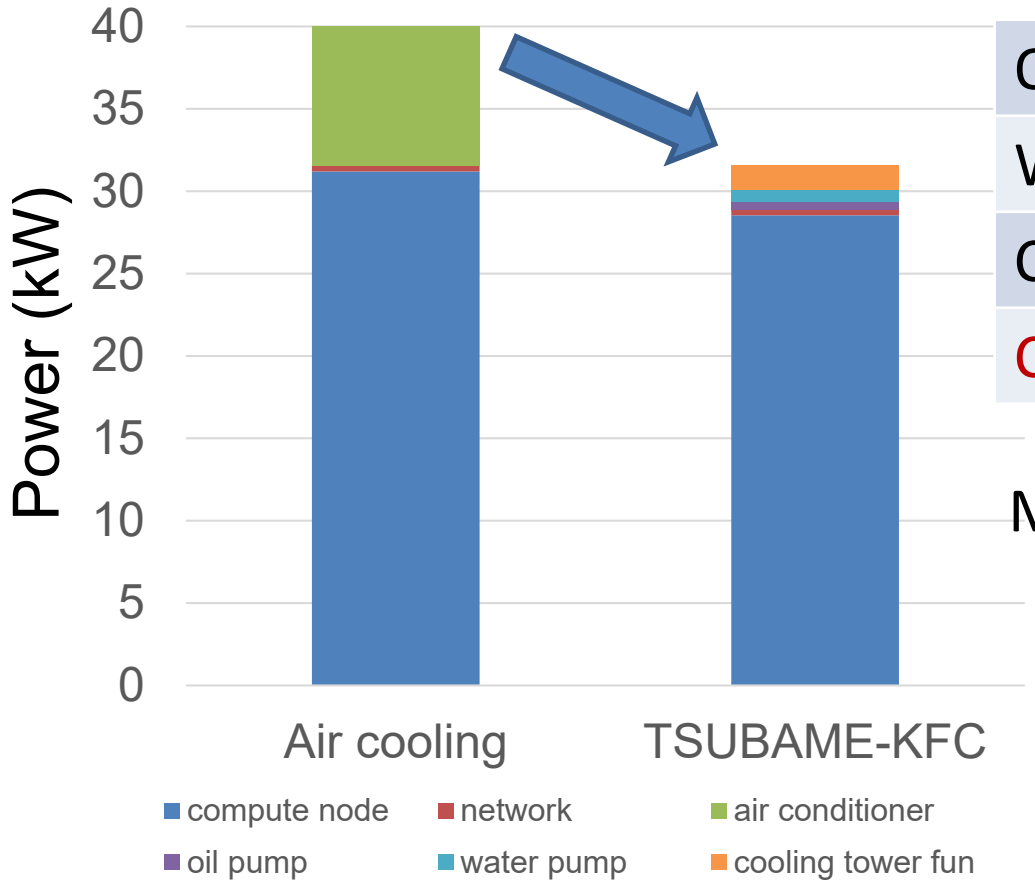
Cooling Tower:
Water 25~35℃
⇒ Outside

Peak performance
~200TFlops
➔ ~600TFlops

World No 1 in Nov 2013& Jun 2014 Green500

PUE (Power Usage Effectiveness) of TSUBAME-KFC

(= Total power / power for computer system)



PUE=1.3 in air cooling

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

Measurement during Linpack benchmark

PUE = 1.09!!

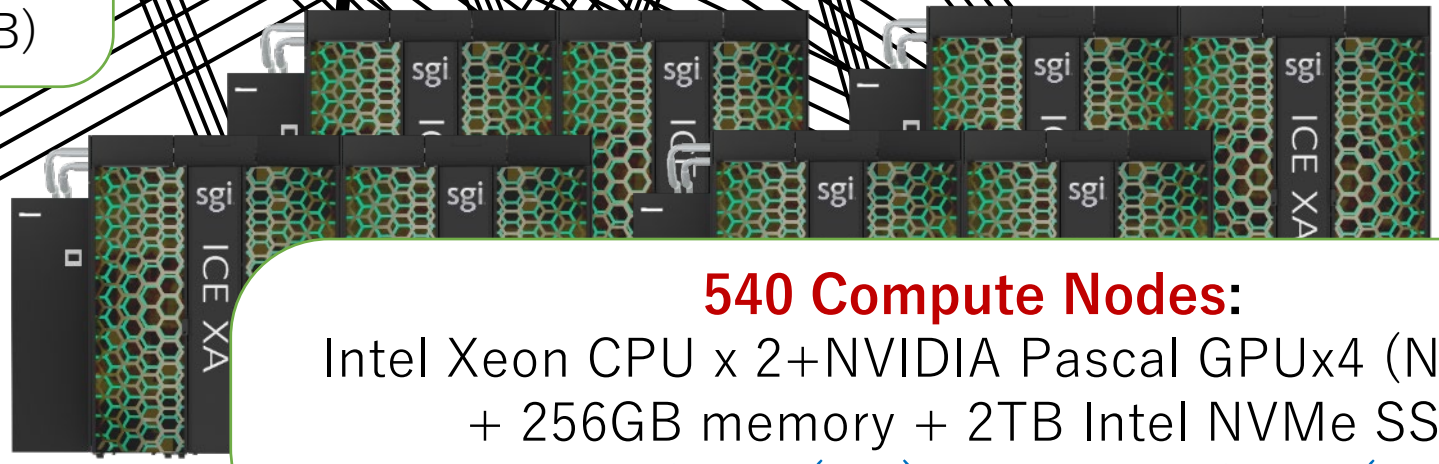
Overview of TSUBAME3.0
BYTES-centric, Scalable Architecture to all 2160 GPUs
(Aug 2017--)

World No 1 in
Jun 2017 Green500
14GFlops/Watt

Hewlett Packard Enterprise

Full Bisection Bandwidth
Intel Omni-Path Interconnect. 100Gx4/node
Full Bisection / 432 Terabits/s bidirectional

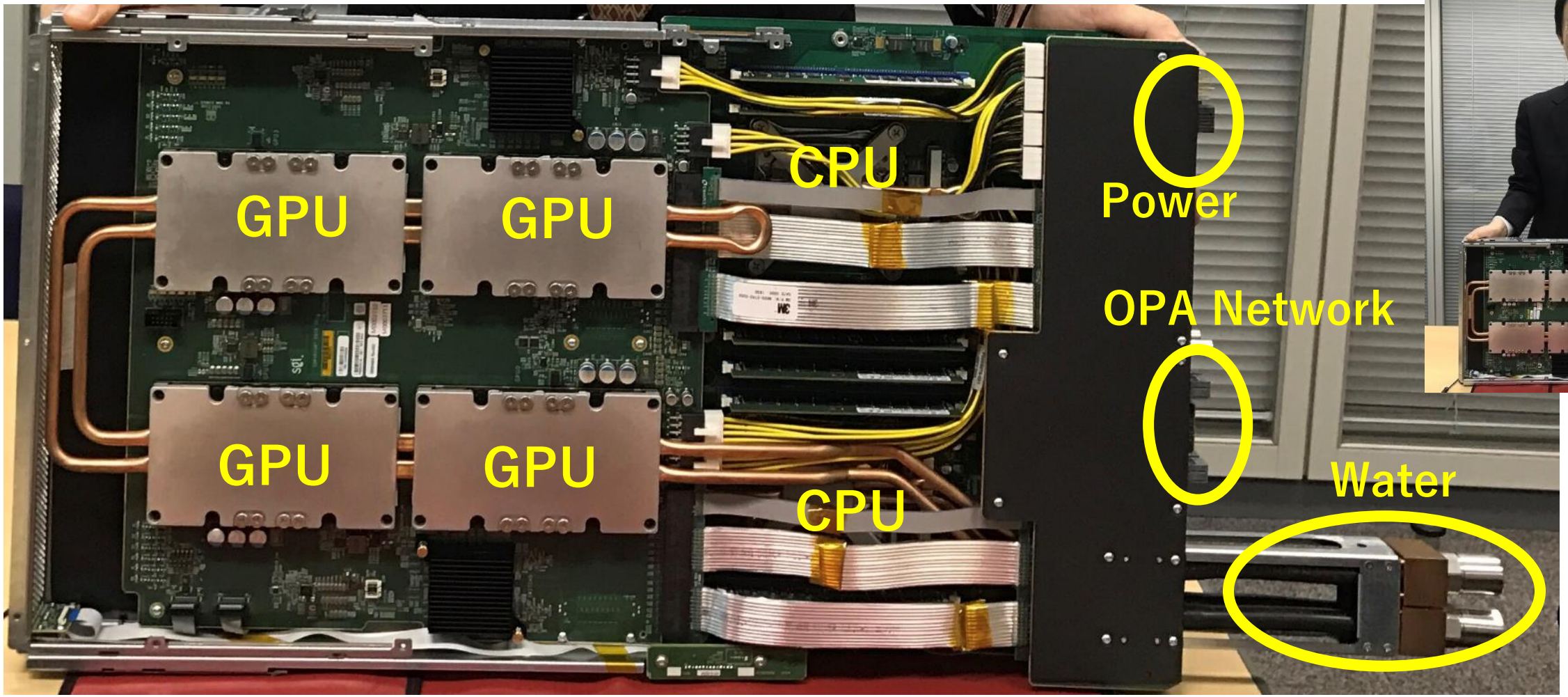
DDN Storage
(**Lustre FS 15.9PB**+Home 45TB)

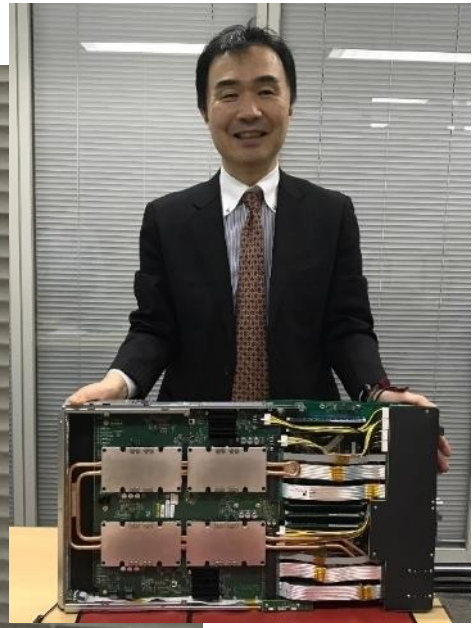


540 Compute Nodes:
Intel Xeon CPU x 2+NVIDIA Pascal GPUx4 (NV-Link)
+ 256GB memory + 2TB Intel NVMe SSD
12.1 Petaflops (DP) , 47.2 Petaflops (FP16)

TSUBAME3.0 Node

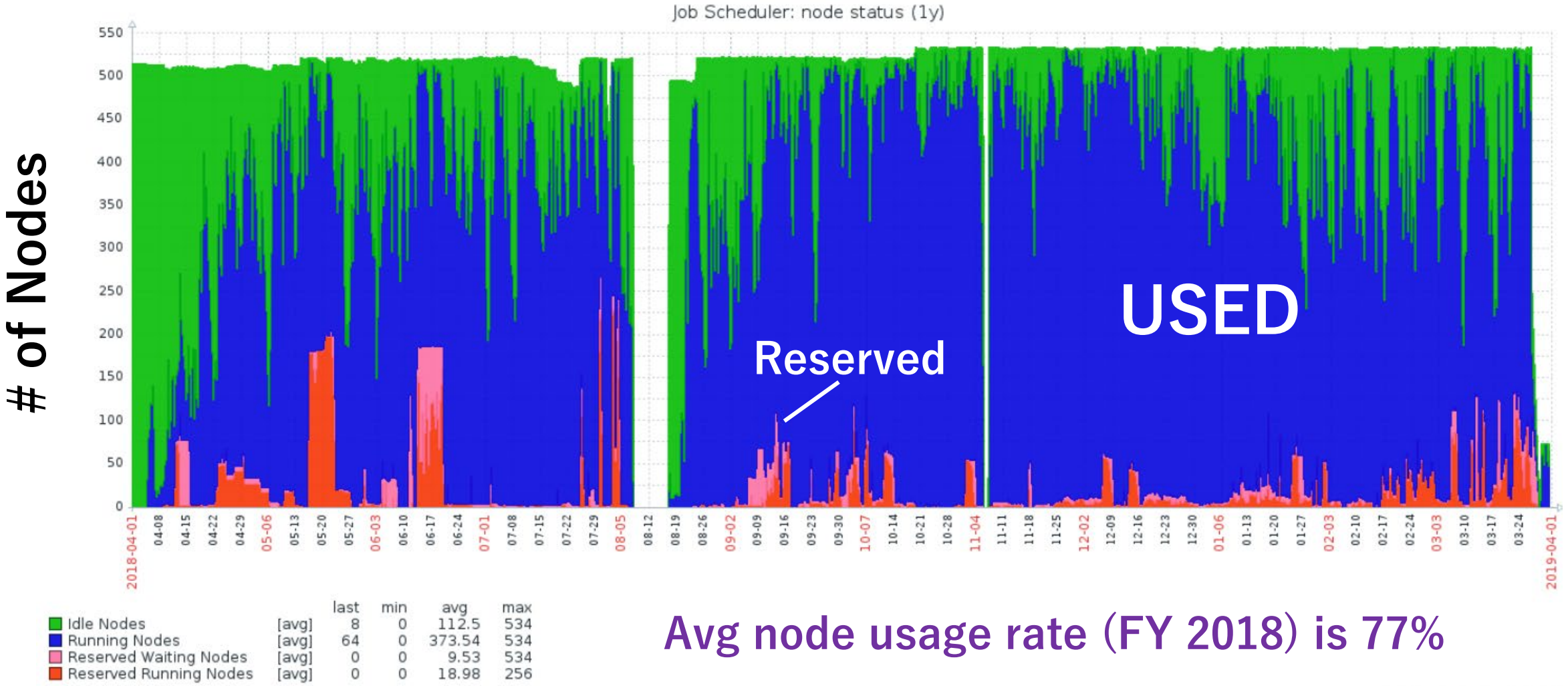
- Compact 2CPU+4GPU+4HCA Node
- CPUs/GPUs cooled by water, Others cooled by air





Heavily Crowded TSUBAME3 with 6000 Users

Apr 2018-Mar 2019



Avg node usage rate (FY 2018) is 77%

http://www.t3.gsic.titech.ac.jp/en/monitoring

Warm Water Cooling in TSUBAME3

Rooftop free cooling tower

Chillers on the ground
(Old and Not so efficient)

1MB Cooling Capacity

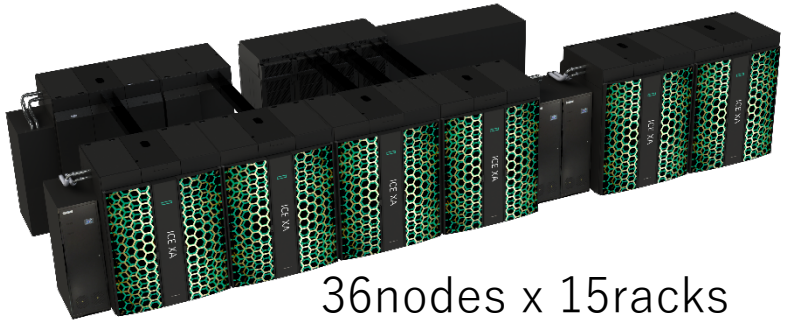
Return 40 degrees C

Outgoing 32 degrees C

2MW Cooling Capacity

Return 24 degrees C

Outgoing 17 degrees C



36nodes x 15racks

Compute Node
HPE SGI ICE XA

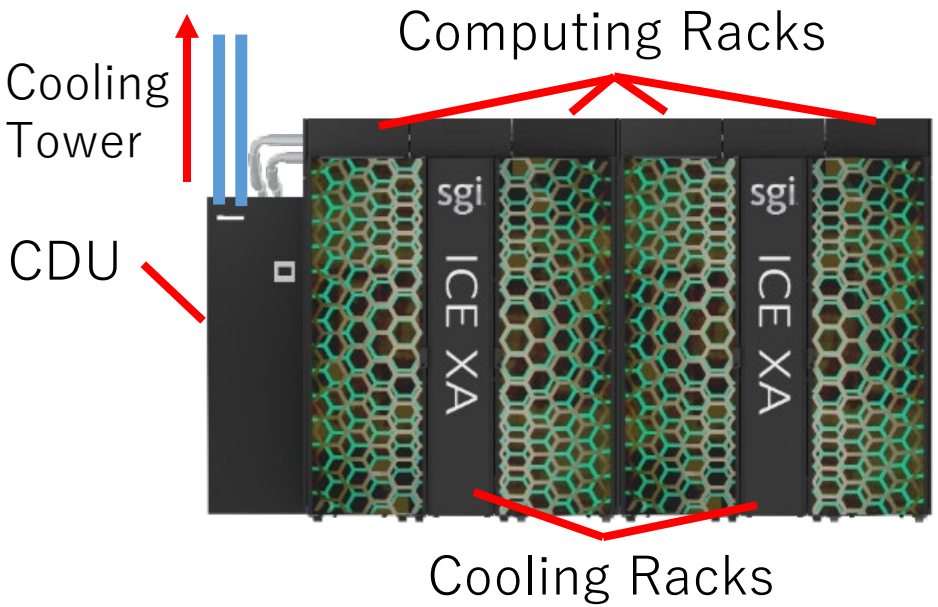
Max 900kW
Avg 350kW

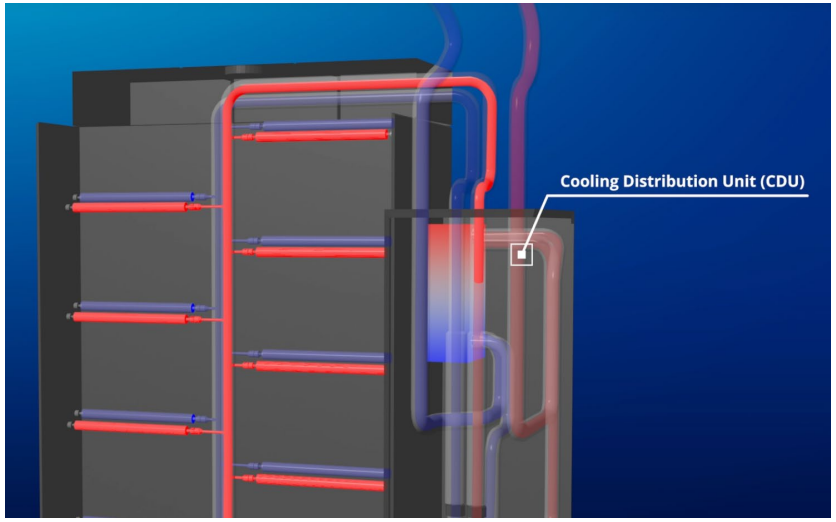
Backup Heat Exchanger

Storage
Interconnect SW

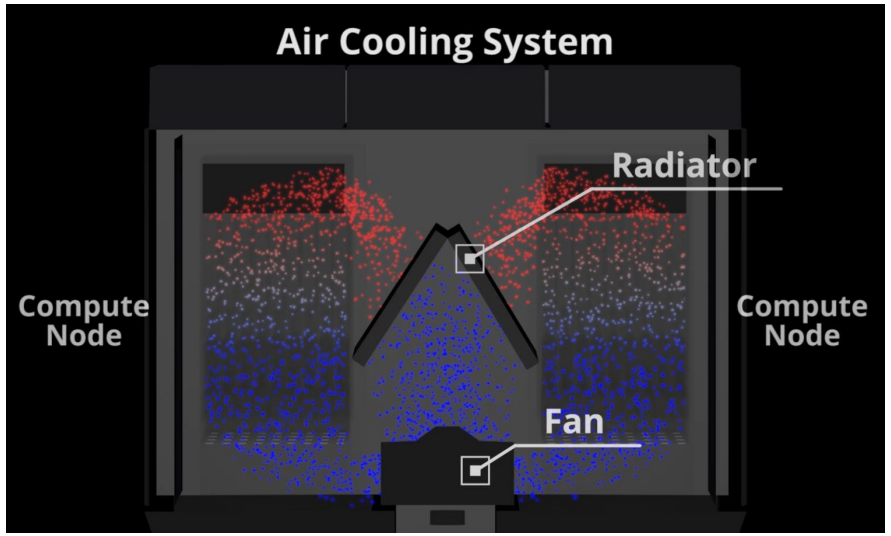
~70kW

HPE/SGI Cooling System

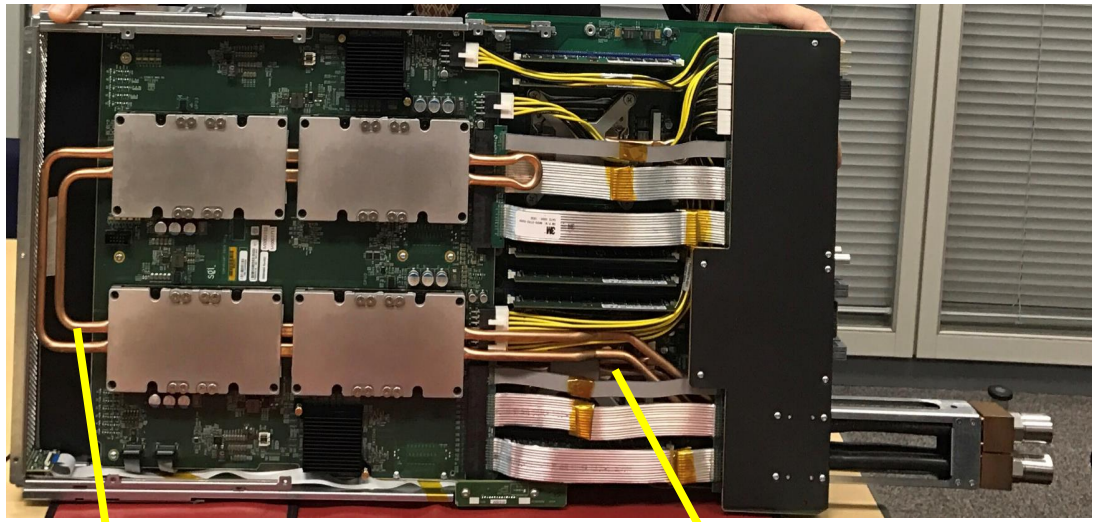




Pure water loop between CDU and Computing Racks



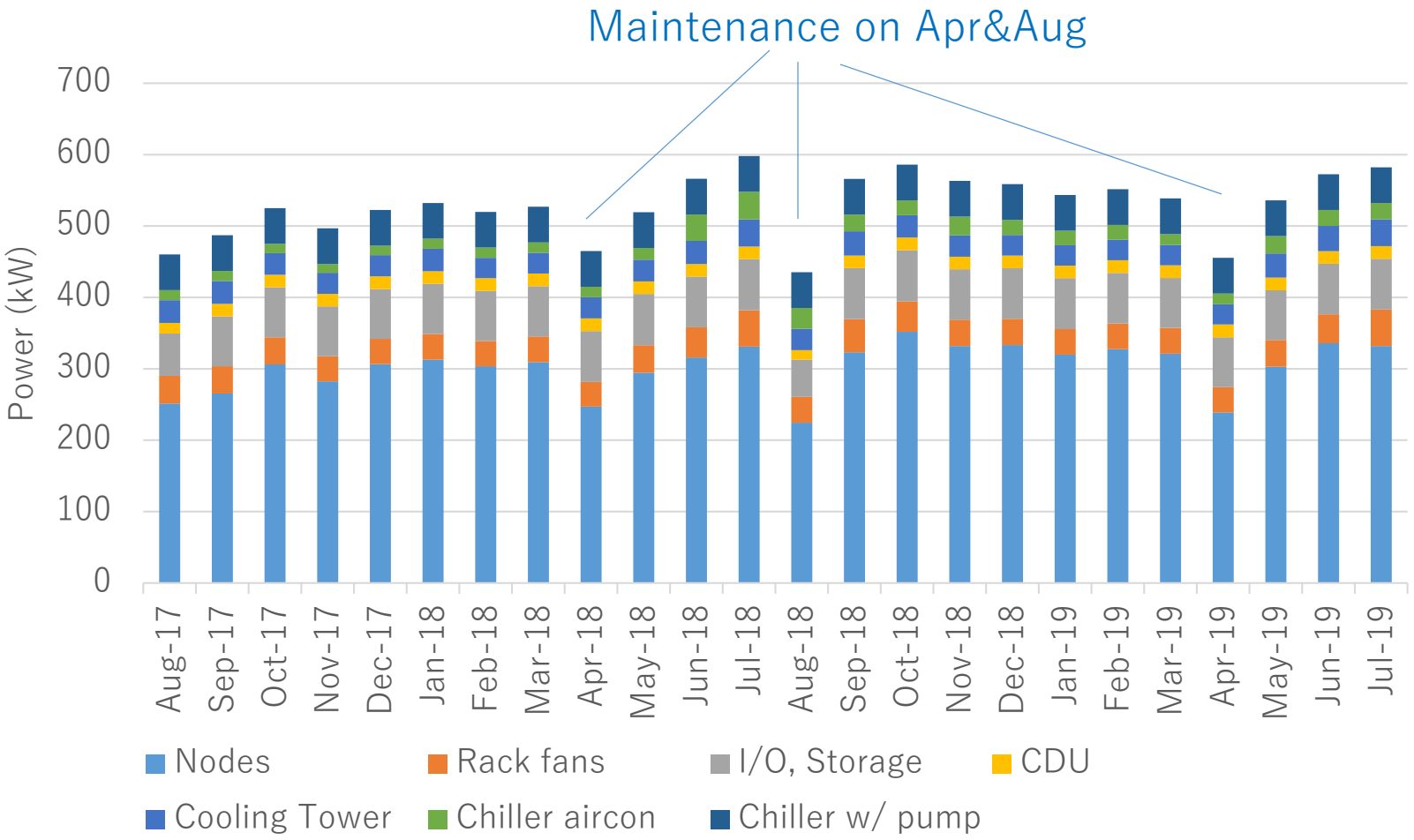
Water is also used to cool in-rack air (for memory, SSDs)

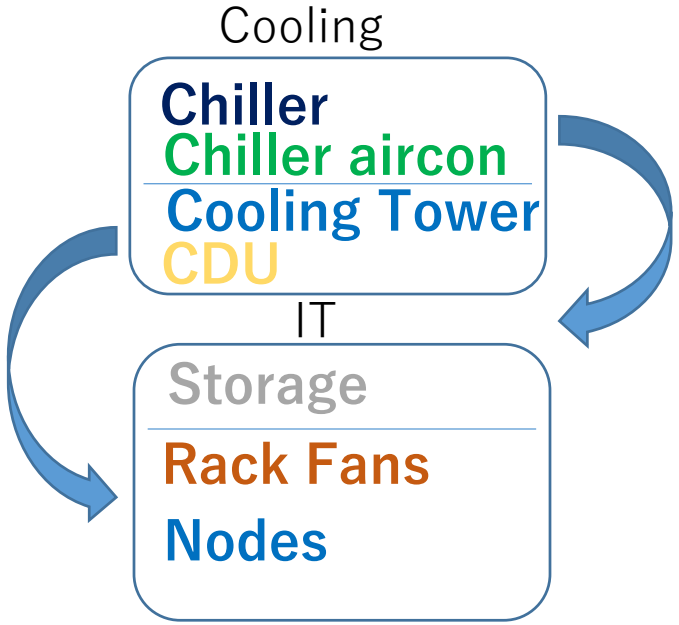


Pure water pipe For 4GPUs

Pure water pipe For 2CPUs

Power Usage of TSUBAME3.0 in 2Yr Operation

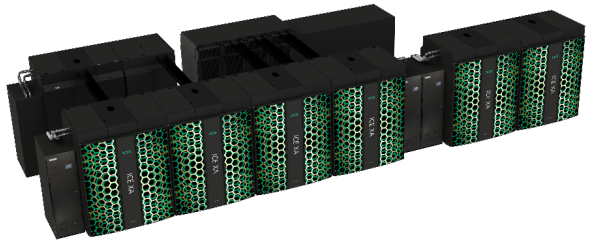




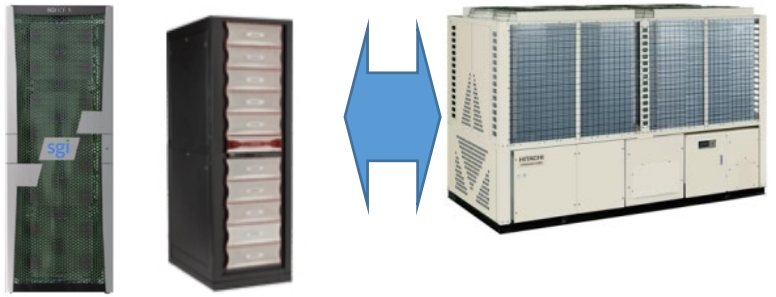
NOTE: Chillers have not been connected to power monitoring system, so chillers power values are based on estimation

Average PUE of TSUBAME3.0 in Operation

Cooling tower part



Chiller part



$$\frac{\text{Nodes}+\text{Rack Fans}+\text{CDU}+\text{Cooling Tower}}{\text{Nodes}+\text{Rack Fans}} = \frac{303+39+17+31}{303+39} = 1.14$$

$$\frac{\text{Storage}+\text{Chiller aircon}+\text{Chiller}}{\text{Storage}} = \frac{69+20+50}{69} = 2.00$$

Entire PUE: 1.29

Discussion on Current Efficiency

- The chiller part is inefficient and degrades entire PUE
  - Our current chillers were installed in 2010 for TSUBAME2
  - They (2MW capacity) are oversized for storage (70kW)
- PUE of cooling tower part (1.14) is not so good as that of KFC
  - Operational power usage (350kW) < Linpack power (800kW) < Theoretical peak (900kW)
  - We have room for tuning cooling parameters

Thank you for your attention

Toshio Endo

endo@is.titech.ac.jp