Energy Efficiency Considerations for HPC Procurements

Energy Efficiency HPC Working Group

SC19, Denver, Tuesday 19 November 2019

Agenda

- On-going work in the EE HPC WG **procurement team**, Gert Svensson
- Three procurements with novel technology:
  - Using modularized **containers to achieve extreme energy savings**, William Thigpen, NASA Ames Research Center
  - Energy Efficiency Considerations for HPC **Procurements @ Cineca**, Carlo Cavazzoni, CINECA, Italy
  - **ABCI** and the **Energy Efficient Datacenter Operation**, Ryousei Takano, AIST, Japan
- Q/A - **discussion**

KTH ROYAL INSTITUTE OF TECHNOLOGY

On-going Work in the EE HPC WG Procurement Team

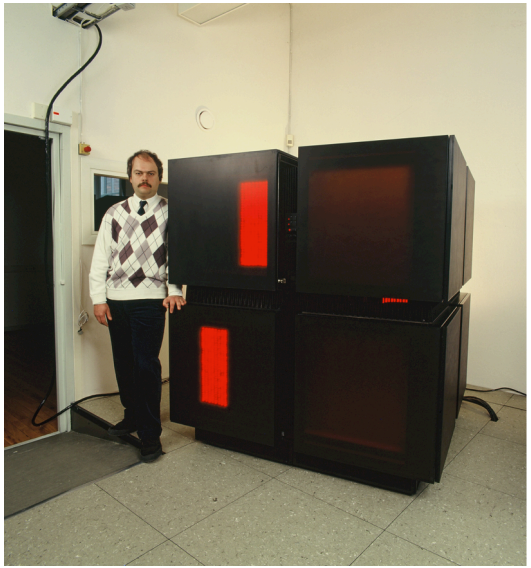
Gert Svensson, KTH, Stockholm, Sweden

Supported by the PRACE project

SC19, Denver, Tuesday 19 November 2019

Who am I?

- Have run procurements for HPC during 30 years at the KTH, PDC Center for High Performance Computing



. . .

How did I get involved in the EE HPC WG?

- Attended several EE HPC WG Workshops at SC
- Attended an EE HPC WG BOF about procurements
- The main topic was:
  - The frequency of energy measurements in different parts of the system
- The procurement document started with the same topic
- Talked to Natalie Bates, leader of the EE HPC WG
- Been an active member in the procurement team since 2017

Highly active members of the procurement team

- Natalie Bates, EE HPC WG coordinator
- Jason Hick, Los Alamos National Laboratory
- James H. Laros, Sandia National Laboratories
- Gert Svensson, KTH Royal Institute of Technology

The goal of the procurement team

- To disseminate **best practices** about how to deal with **energy efficiency** in the **procurement** process for HPC systems via
  - This **BOF**
  - The document ”**Energy Efficiency Considerations for HPC Procurements**”
    - > *First version 2013*
    - > *We now release a draft of the 2019 year’s version:*
    - > *https://eehpcwg.llnl.gov/compsys_pro.html*

Analyzed procurement documents from major HPC sites for energy-related requirements

- Thanks to all the sites that provided documents!

  - ORAL
  - CTS1
  - Cineca
  - Crossroads
  - GENCI
  - JCAHPC
  - Jülich
  - KTH
  - LRZ
  - Trinity & NERSC-8
  - Vanguard

Examples from the procurement document- Cooling section: Starts with some questions

- What type of cooling **exists** in the building today? Is it sufficient to cool the new HPC system and economical to use?
- What is the **local climate** like? Can the heat from computers be disposed of in open air without using compressors?
- Would **free cooling be sufficient** for the warmest periods of the year or will other methods be needed for some periods?
- Do **lakes** with **cool water** that can be used for cooling of the new computer exists in the area near the building?
- Can the **heat from the computer be used for other purposes**, like heating buildings, heating a pool or heating greenhouses?

Cooling section: Selected requirements

Information to the vendor	Include a **description of the facility cooling** systems showing component detail how heat is transferred from the compute platform to the facility environment.
Basic requirement	The tenderers should **describe** the computing equipment and its **method(s) of cooling** (e.g. direct liquid cooling, indirect liquid cooling, air-cooling, rear-door heat exchangers, hybrid systems, etc.).
Basic requirement	The tenderers or contractor should participate in a **site survey**.

Liquid cooling section: Selected requirements

Advanced requirement	For IT equipment located in facilities with inlet water temperature not exceeding 35°C, at least **90% of the energy** consumed by the system should be **removed by water**.
Innovative requirement	For inlet water temperatures in the range between 35°C and 45°C and a room temperature below 35°C, **>98% of the energy** consumed by the system should be removed by water.
Advanced requirement	For a procurement of a system using a heat re-use the vendor should **guarantee the outgoing water** temperature for a typical **defined workload** of the system.

The structure of the procurement document

- Ideas, things to think of regarding energy efficiency
- Expressed as NEW STRUCTURE
  - background information & questions to the author of the RFP
  - request for information
  - requirements
    - > *Basic*
    - > *Advanced*
    - > *Innovative*

NEW SECTIONS:
Cooling
Interfacing with facilities

  - Note: We present ideas, not a copy-paste document for RFP:s

Thanks!

gert@kth.se

natalie.jean.bates@gmail.com

Agenda

- On-going work in the EE HPC WG **procurement team**, Gert Svensson
- Three procurements with novel technology:
  - Using modularized **containers to achieve extreme energy savings**, William Thigpen, NASA Ames Research Center
  - Energy Efficiency Considerations for HPC **Procurements @ Cineca**, Carlo Cavazzoni, CINECA, Italy
  - **ABCI** and the **Energy Efficient Datacenter Operation**, Ryousei Takano, AIST, Japan
- Q/A - **discussion**

Energy Efficiency Considerations for HPC Procurements

William W Thigpen
Manager, High-End Computing Capability Project
william.w.thigpen@nasa.gov

Energy-Efficient Driven Procurement

Prototype module deployment

- NASA developed a series of prototypes beginning in 2016 to prove the concept of utilizing outside air to cool supercomputers and to grow the system in multiple modules. The prototype resulted in a 96% reduction in water and 90% reduction in electricity to cool the systems.

- In 2017 we produced a new RFI followed by an industry day and later RFP targeted at building a new modular facility capable of housing a computer with at least 1,000,000 cores.
  - Vendors were responsible for building the pad, module, cooling and electrical infrastructures, as well as the computer and their plan had to show how they would grow from the initial deployment to the final system.
  - Vendors were scored on how energy efficient their total system was including compute, cooling and electrical distribution

Artist rendition of full site provided during industry day

Energy-Efficient Deployment

Initial module

- HPE and Schneider Electric teamed to develop the site and deploy the initial module. The site is prepared to house compute modules and 3 data modules. The electricity used for cooling is less than 5% of the power to run the computer.
- The system is cooled using a warm water (90° F) closed cooling loop. The processors and network ASICS are cooled using cold sinks. The rest of the heat is transferred to the water loop in a closed cell using suction to pull the air across the nodes. Adiabatic cooling systems use outside air to cool the closed loop.

- Each module and its support infrastructure will be deployed as required by NASA and can take advantage of future improvements in module and cooling technologies leading to further energy saving.

Artist rendition of full site duplicating initial module

Agenda

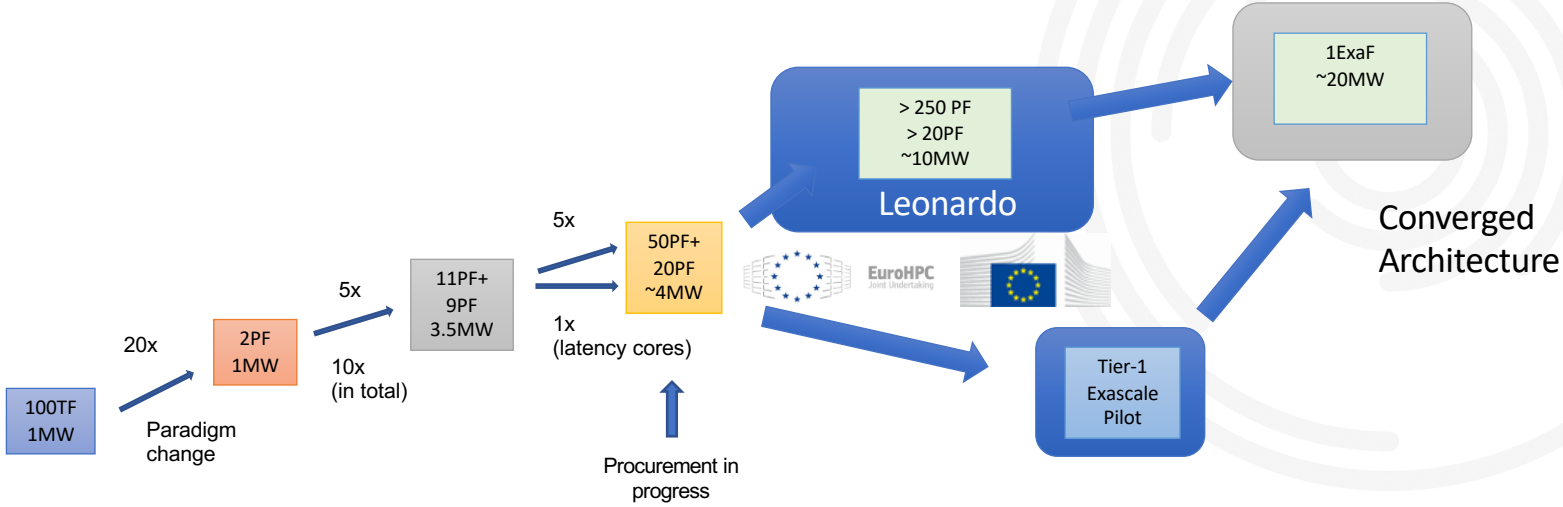
- On-going work in the EE HPC WG **procurement team**, Gert Svensson
- Three procurements with novel technology:
  - Using modularized **containers to achieve extreme energy savings**, William Thigpen, NASA Ames Research Center
  - Energy Efficiency Considerations for HPC **Procurements @ Cineca**, Carlo Cavazzoni, CINECA, Italy
  - **ABCI** and the **Energy Efficient Datacenter Operation**, Ryousei Takano, AIST, Japan
- Q/A - **discussion**

Energy Efficiency Considerations for HPC Procurements @ Cineca

Carlo Cavazzoni – c.cavazzoni@cineca.it
Business Unit HPC High Performance Computing

CINECA

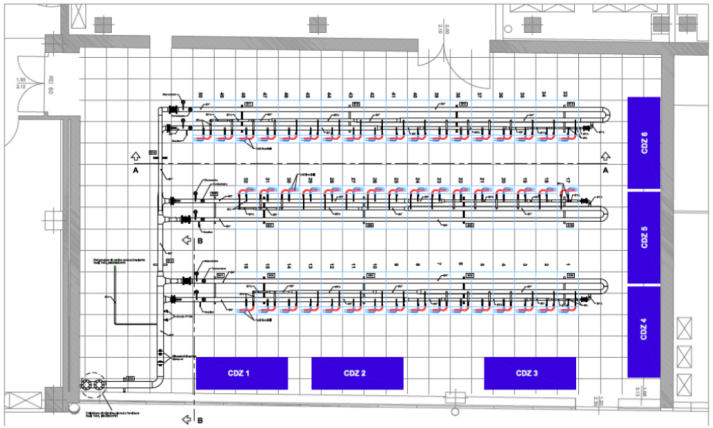
Italian exascale RoadMap



2009	2012	2016	2019/2020	2021	2023-2025	2025-2027
IBM SP6 Power6	**Fermi IBM BGQ PowerA2**	**Marconi Lenovo Xeon+KNL**	**Marconi PPI4HPC ICEI - PPIHBP**	**Leonardo With EuroHPC contribution**	**Post-Marconi Exascale pilot technology**	**Post-Exa**

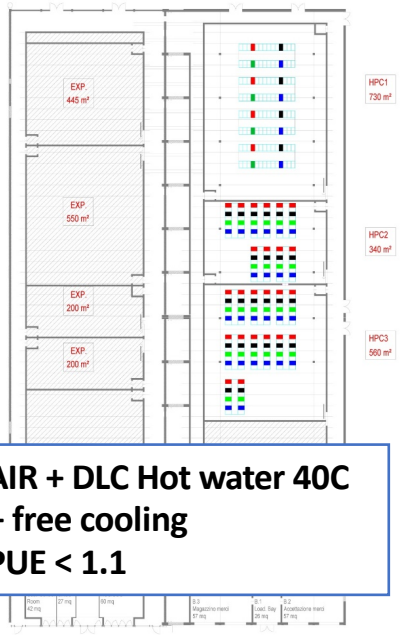
Two different datacentres

Marconi DC



AIR+ Cool water 15-18C
+ free cooling
PUE ~1.4

Leonardo DC



DLC room

AIR room

AIR + DLC Hot water 40C
+ free cooling
PUE < 1.1

DATA ROOM STAGE 1: 1600 sqm
DATA ROOM STAGE 2: 2600 sqm
ANCILLARY SPACES: 900 sqm

CINECA

Procurement RFP

- Cooling technology (*proposed for Leonardo*)
- TCO approach based on application sustained performance (*required for Marconi*)
- Monitoring & Power Capping (*required for Marconi*)

CINECA

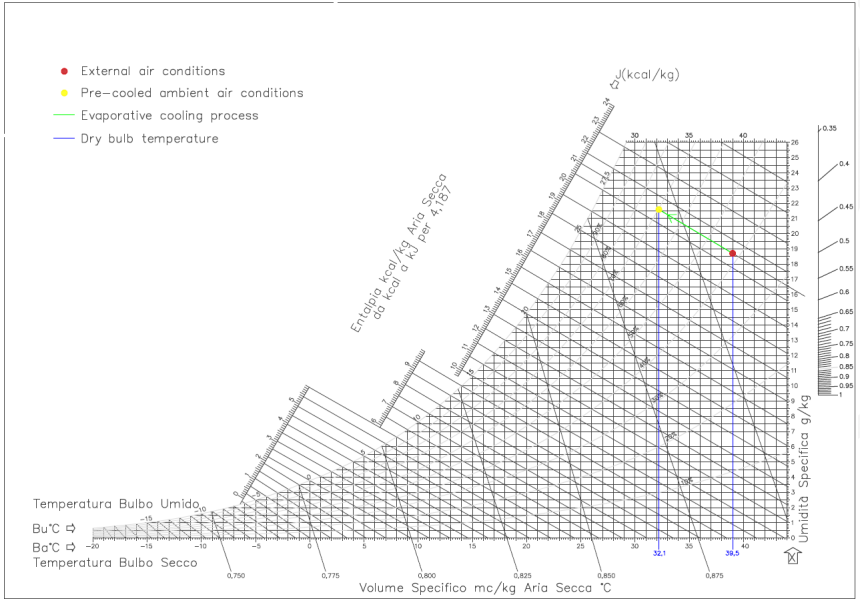
Cooling Technology

- External data as per ASHRAE Handbook – Fundamentals (2017)
- Location: Bologna
- Extreme conditions at n = 20 years

o Dry-bulb Temperature +39,5°C
o Wet-bulb Temperature +27,5°C

The proposed design strategy uses adiabatic dry coolers to provide warm water to the HPC racks at 40°C throughout the year. Dry coolers lower down warm water from 50°C to 40°C and can operate in two ways:

- Dry operation: water is cooled only via external air
- Wet operation: external air is pre-cooled via adiabatic system, which lowers ambient air dry-bulb temperature at constant wet-bulb temperature; water is then cooled via the pre-cooled ambient air



Adiabatic cooling working principles, which allow the procured system to provide water operating temperature at 40°C even at extreme external design conditions

Extreme Annual Design Conditions

Extreme Annual WS				Extreme Annual Temperature				n-Year Return Period Values of Extreme Temperature							
				Mean		Standard Deviation		n=5 years		n=10 years		n=20 years		n=50 years	
1%	2.5%	5%		Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
(n)	(o)	(p)		(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)
7.3	6.2	5.4	DB	-7.0	36.5	2.7	1.6	-8.9	37.7	-10.5	38.6	-12.0	39.5	-13.9	40.7
			WB	-7.6	25.4	2.6	1.1	-9.5	26.2	-11.0	26.8	-12.4	27.5	-14.3	28.3

TCO approach

TCO relevance: 50 points out of 100!

TCO = CAPEX + OPEX

CAPEX in our tender is fixed, for the last procurement was 25M€

OPEX -> workloads (production) in the system lifetime

Benchmark made up of 4 user app + HPC + HPCG

$$Energy.LifeTime.App = [x * PUEDLC + (1 - x) * PUEAC] * n.App * ETS.App$$

Take weighted average over the 6 benchmarks, usually 1/6

Define a score -> $HPL'_C = HPL_C(1 + p * Eff_C)$

CINECA

Monitoring and Power Capping

For the last procurement

The combined system + cooling power requirements were above the datacentre capacity in the worst case scenario:

Linpack during heat waves

In the RFP we ask the HPC system to have a energy monitoring system and a power capping solution to gracefully reduce the power consumption during heat waves.

monitoring requirement: 95% accuracy @ 1% performance overhead, & job level granularity

CINECA

Thank you!

CINECA

Agenda

- On-going work in the EE HPC WG **procurement team**, Gert Svensson
- Three procurements with novel technology:
  - Using modularized **containers to achieve extreme energy savings**, William Thigpen, NASA Ames Research Center
  - Energy Efficiency Considerations for HPC **Procurements @ Cineca**, Carlo Cavazzoni, CINECA, Italy
  - **ABCI** and the **Energy Efficient Datacenter Operation**, Ryousei Takano, AIST, Japan
- Q/A - **discussion**

ABCI and the Energy Efficient Datacenter Operation

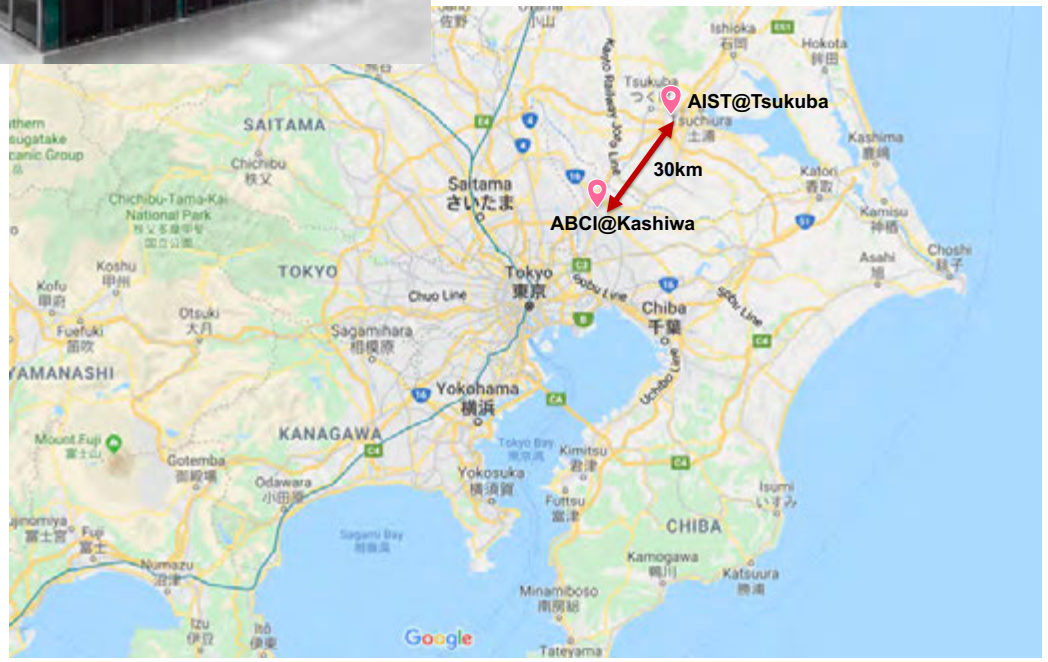
Ryousei Takano

National Institute of Advanced Industrial Science and Technology (AIST), Japan

Energy Efficiency Considerations for HPC Procurements BoF, 19 November 2019

ABCI: The World's First Large-Scale Open AI Infrastructure

Univ. Tokyo / Kashiwa II Campus



ABCI AI Bridging Cloud Infrastructure

- World Top-Level compute and data process capability
- **Open, Public, and Dedicated** infrastructure for AI & Big Data Algorithms, Software, and Applications
- **Open Innovation Platform** to accelerate joint academic-industry R&D for AI

Peak Performance:
550 PFlops (FP16)
37 PFlops (FP64)
Effective Performance:
19.88 PFlops (#8 in TOP500)
14.423 GFlops/W (#3 in GREEN500)
509 GFlops (#5 in HPCG)

Operation: October 2018~

ABCI Status
Nov 13, 2019 13:29:01 to Nov 13, 2019 19:54:32
PUE
Current PUE
1.097
PUE Changes
Utilization
Current Utilization
93.25%
Utilization
Disabled Current: 15
Reservation Current: 8
Batch Current: 993
Idle Current: 72
Power Consumption
Current Power Consumption
1.101 MW
Power Consumption
kWatt
Total
Servers
Cooling System

ABCI: Optimized for large-scale DL with commodity hardware

High-Performance Computing System

550 PFlops (FP16), 37.2 PFlops (FP64)

476 TiB memory, 1.74 PB NVMe SSD

Compute Node (w/ GPU) x 1,088

GPU	NVIDIA Tesla V100 SXM2 x 4 (Total 4,352)
CPU	Intel Xeon Gold 6148 x 2 (Total 2,176)
Memory	384GiB
Local Storage	Intel SSD DC P4600 (NVMe) 1.6TB x 1
Interconnect	InfiniBand EDR x 2

Multi Platform Node (w/o GPU) x 10
- Intel Xeon Gold 6132 (2.6GHz/14cores) x 2
- 768GiB Memory, 3.8TB NVMe SSD

Interactive Node x 4 | Management and Gateway Nodes x 15

Large Volume Storage System

22 PB GPFS + 1 PB All-Flash Storage + 17 PB Cloud Storage

DDN SFA14K (w/ SS8462 Enclosure x 10) x 3
- 12TB 7.2Krpm NL-SAS HDD x 2400
- 3.84TB SAS SSD x 216
- NSD S

DDN S
- 7.68TI
- MDS S

HPE A
- 12TB SATA HDD x 1440
- 3.2TB SSD x 24

Interconnect Network (Infiniband EDR)

- Mellanox CS7500 x 2
- Mellanox SB7890 x 229

Service Network (10GbE)

Gat
- Nexsus 3232C x2
- FortiGate 1500D x2
- FortiAnalyzer 400E x1

SINET5

■ Point 1: Large-scale and State-of-the-art GPU farm

Total 4352 GPUs. Users can run a large-scale AI/deep learning jobs with friction-less.

■ Point 2: Supercomputer-based network technology

All GPUs and CPUs can communicate with each other through ultra low latency (less than 2.5 us) and wide bandwidth (200Gbps) network.

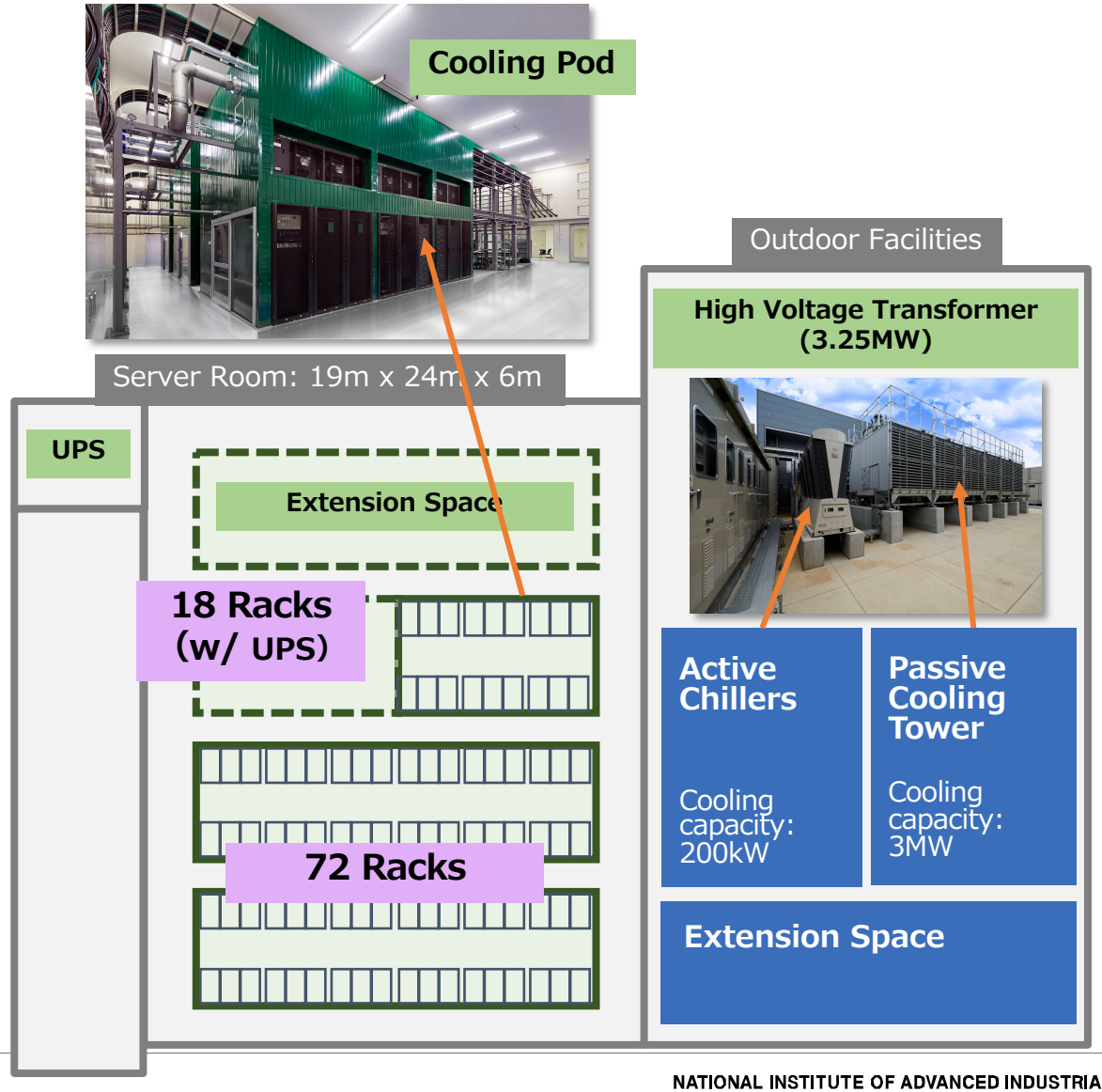
ABCI achieves ultra-dense packaged rack

*June 2019

	TSUBAME3.0	ABCI	Summit	TPU 3.0 Pod
Organization	Tokyo Tech	AIST	ORNL	Google
Start of operation	2017	2018	2018	2018
Number of nodes	540	1088	4608	unknown
Throughput Processor (TP)	NVIDIA Tesla P100	NVIDIA Tesla V100	NVIDIA Tesla V100	TPU 3.0
Number of TP	2160	4352	27648	unknown
Theoretical Perf. (FP64)	12.2 PF	37.2 PF	200 PF	unknown
Theoretical Perf. (DL)	47.2 PF	550 PF	3.3 EF	100 PF / Pod
TOP500*	25	8	1	unknown
GREEN500*	5	3	2	unknown
#Nodes / Rack	36	34	16	unknown
#GPU / Rack	144	136	96	unknown
kW / Rack	64.8 kW	67.33 kW	45-55 kW (est.)	unknown
DL Perf. / Rack	3.1 PF	17 PF	12 PF	12.5 PF (100 PF / 8 rack)

ABCI Datacenter

"Commoditizing supercomputer cooling technologies to Cloud (70kW/rack)"

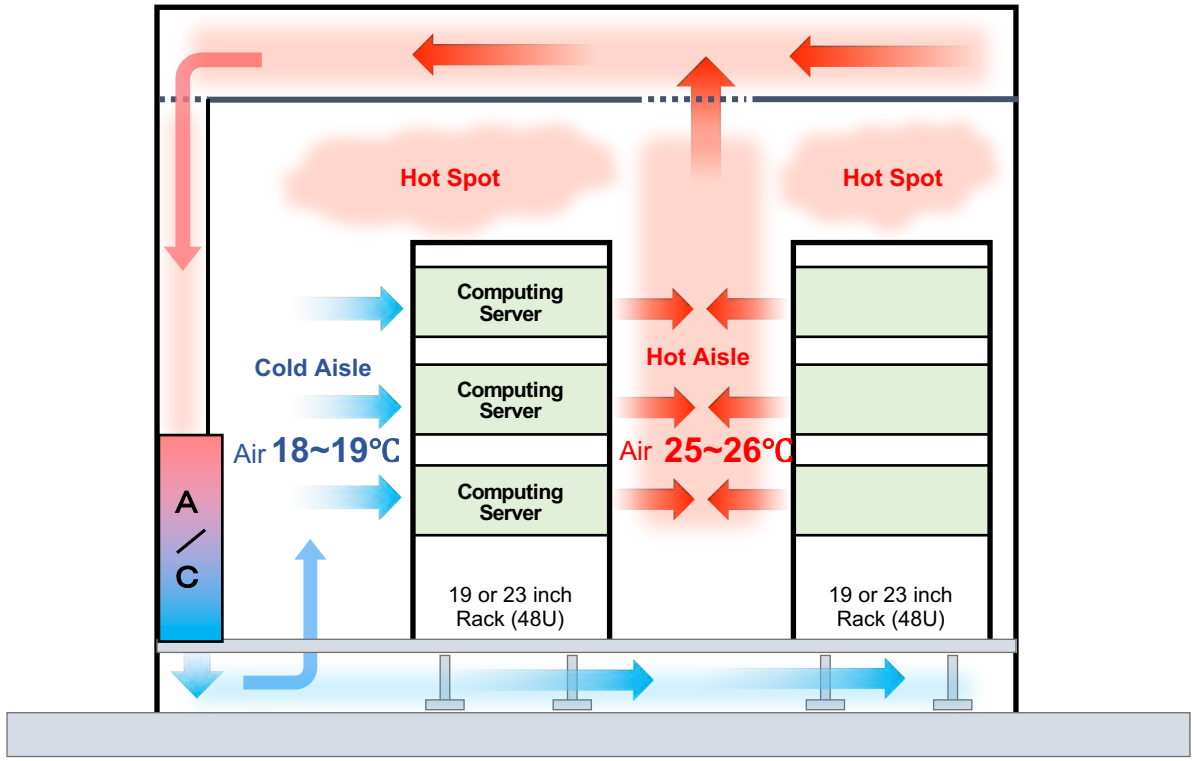


- **Single floor, cost effective building**
- **Hard concrete floor 2t/m² weight tolerance for racks and cooling pods**
- **Number of Racks**
  - Initial: 90 (ABCI uses 41 racks)
  - Max: 144
- **Power capacity: 3.25 MW**
  - ABCI uses 2.3MW max
- **Cooling capacity: 3.2MW**
  - 70kW/rack: 60kW water + 10kW air

Free cooling with water/air hybrid cooling system

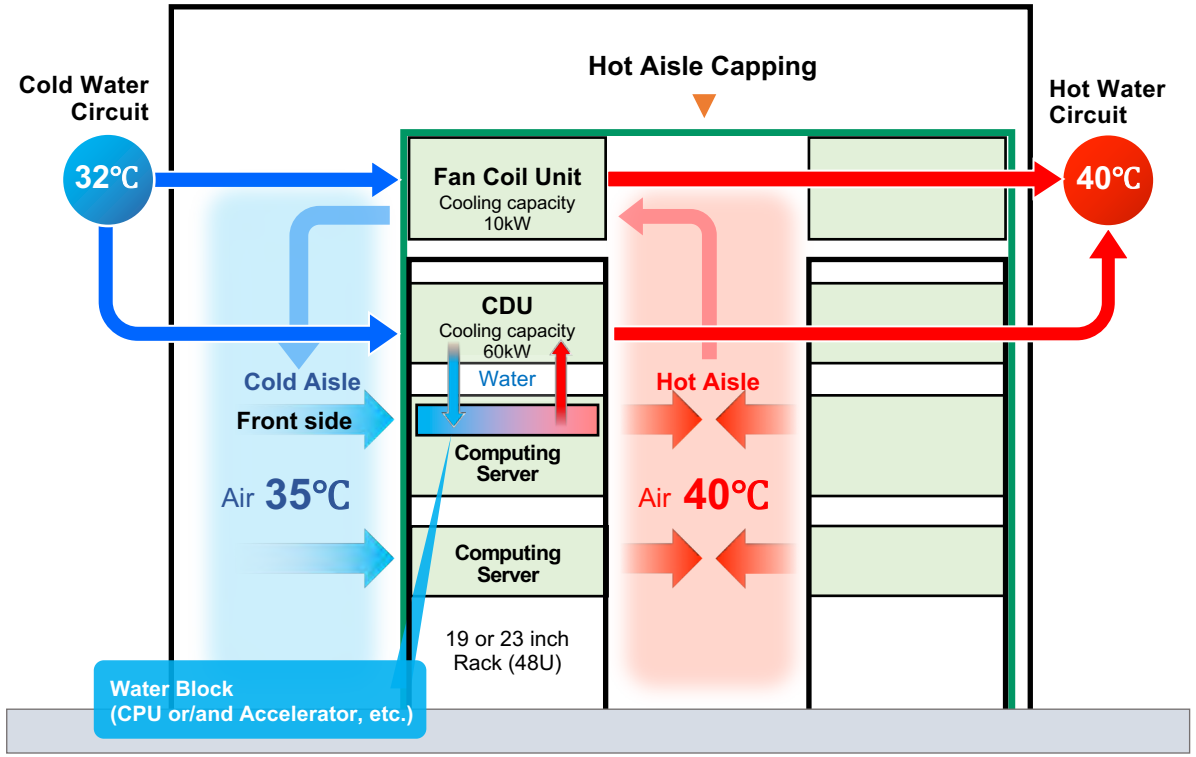
■ Conventional Datacenter Design

Average PUE : ≈ 2.0



■ ABCI Datacenter Design

Average PUE : ~ 1.1

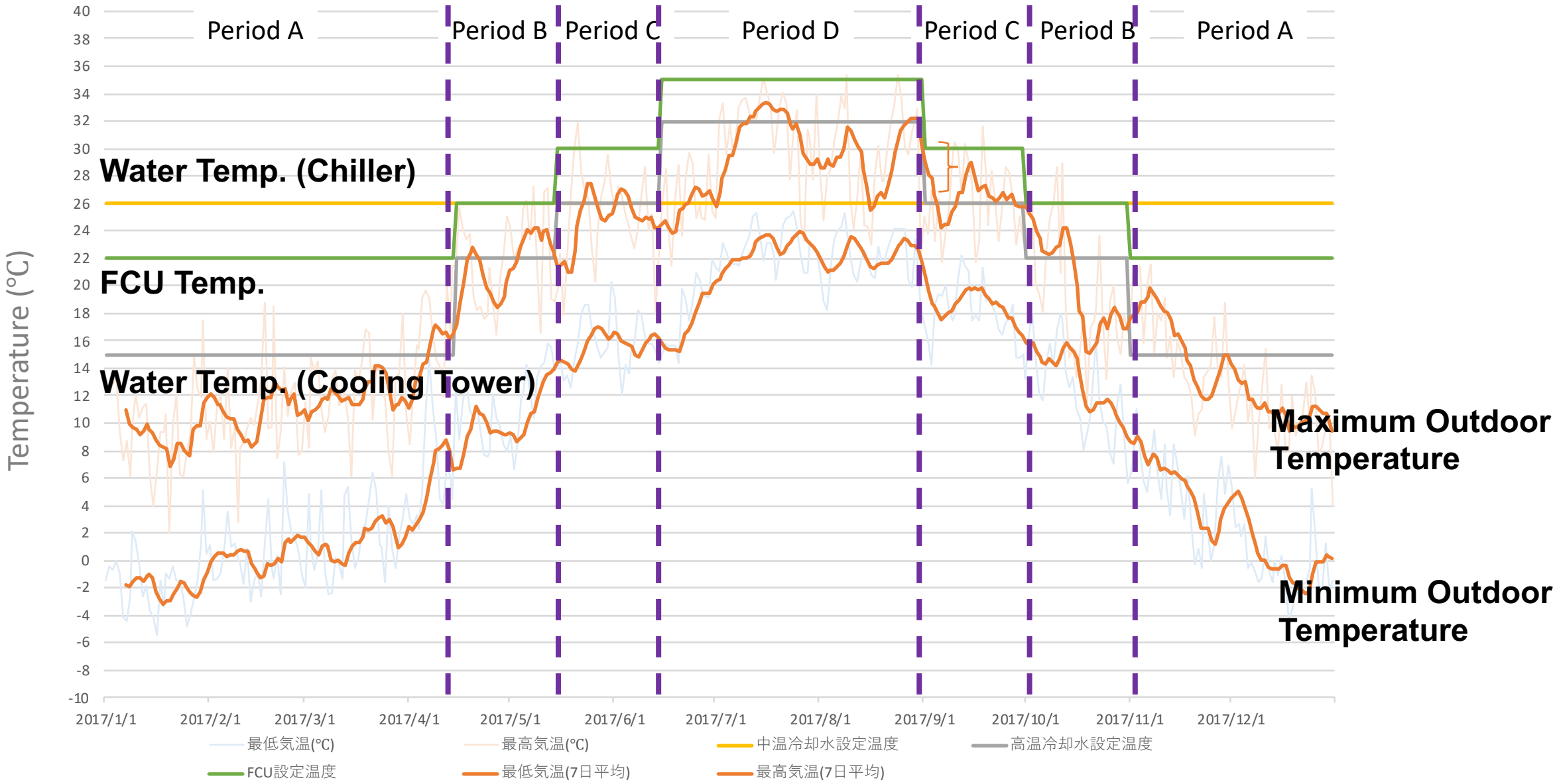


Configuration in Summer

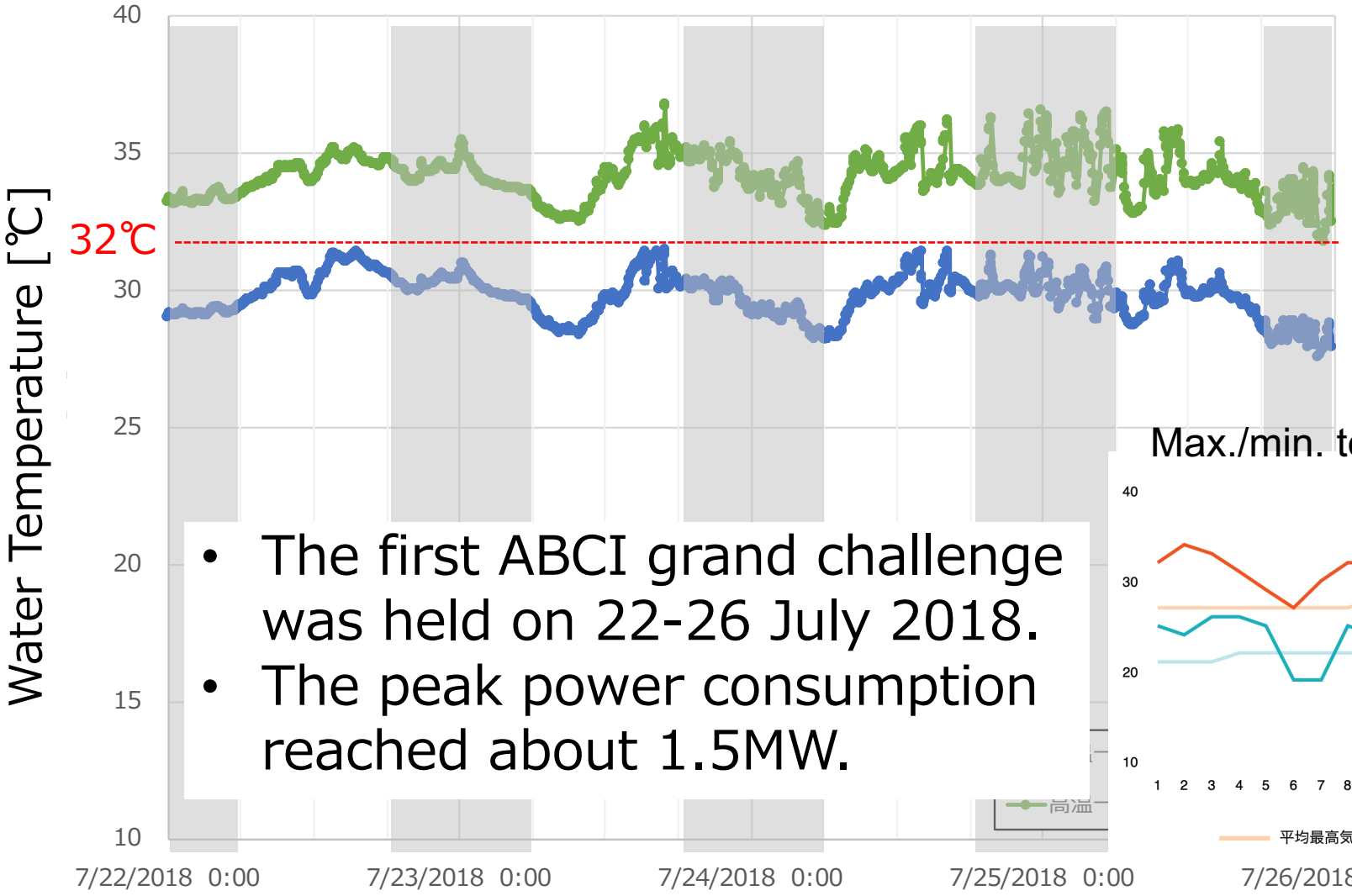
$$\text{Power Usage Effectiveness (PUE)} = \frac{\text{Total Facility Energy}}{\text{IT Equipment Energy}}$$

(Lower the better. 2.0 Standard, 1.4 Good, 1.0 Best)

Cooling System Parameters



100% free cooling is possible in high summer

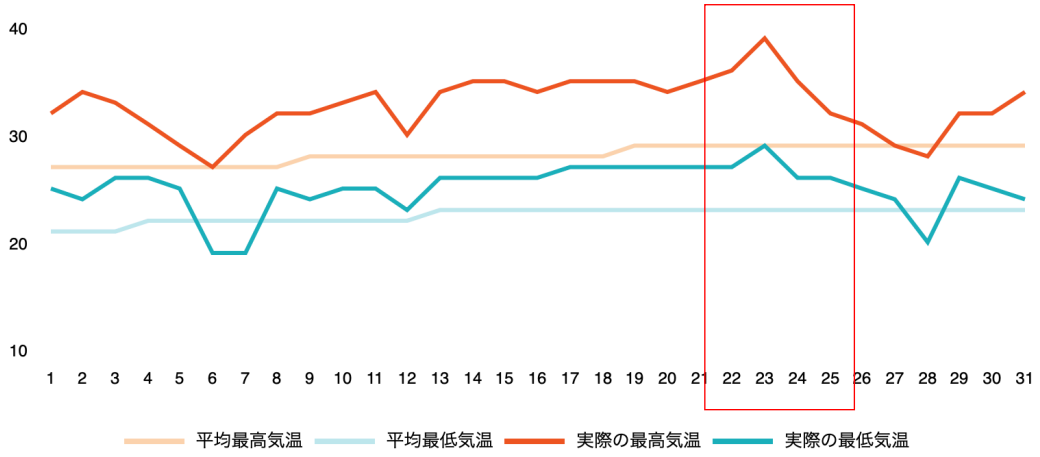


Cooling Water (From ABCI): <40℃

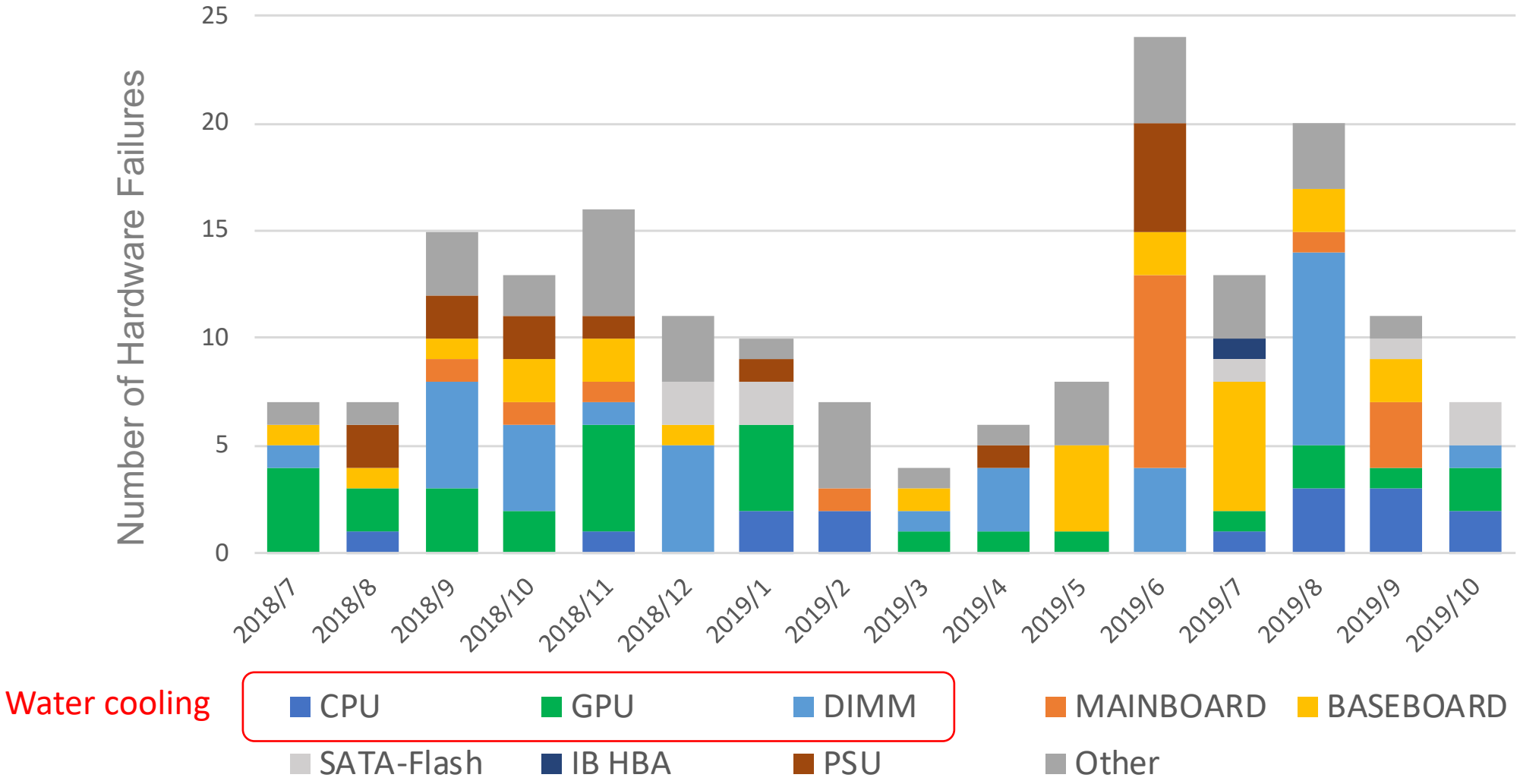
Cooling Water (To ABCI): <32℃

- The first ABCI grand challenge was held on 22-26 July 2018.
- The peak power consumption reached about 1.5MW.

Max./min. temperature in Kashiwa in July 2018



Hardware failure in computing nodes



Weak correlation between hardware failures and rack space layout

The total number of hardware failures in each rack since July 2018

Cooling Pod #1 (Computing node)

14	4	3	9	7	6	9	6		
6	3	11	11	2	5	5	8		

Cooling Pod #2 (Computing node)

6	6	9	3	4	2	1	3		
5	6	1	3	2	8	5	8		

Univa Grid Engine generally assigns jobs in a rack number order (from top left to bottom right).

ABCI Procurement Timeline

AI Datacenter

2016.11 2017.1 2017.8 2018.1

Proposal & Contract → Design → Construction

Datacenter Facility

2017.5 .7 .12 2018.3

Proposal & Contract → Design → Installation

ABCI HW/SW

2016.8 .10 2017.8 .10 2018.3 .6

Conceptual Design → Spec & Benchmark Definition → Proposal & Contract → Detailed Design → Installation & Benchmark

Production

End of 2018.7

Test Op. → Operation

January 16, 2017

October 30, 2017

January 23, 2018

2016.11
2017.1
2017.8
2018.1
Proposal & Contract
Design
Construction
2017.5
.7
.12
2018.3
Proposal & Contract
Design
Installation
.10
2017.8
.10
2018.3
.6
Conceptual Design
Spec & Benchmark Definition
Proposal & Contract
Detailed Design
Installation & Benchmark
End of 2018.7
Production
Test Op.
Operation

March 31, 2018

FCU-2R-01
2016.11
2017.1
2017.8
2018.1
Proposal & Contract
Design
Construction
2017.5
.7
.12
2018.3
Proposal & Contract
Design
Installation
.10
2017.8
.10
2018.3
.6
Conceptual Design
Spec & Benchmark Definition
Proposal & Contract
Detailed Design
Installation & Benchmark
End of 2018.7
Production
Test Op.
Operation

Summary

- We have successfully design and launched an energy efficiency supercomputer dedicated to AI/DL workloads for only two years from scratch.
- Free cooling and warm water cooling are keys for high energy efficiency, and they are working perfectly in Japan.
- We achieve 70kW / rack cooling capacity and the average PUE 1.1 using commodity (cost effective) cooling technologies.
- Automation of facility operation can be critical in the next generation AI/HPC datacenter.
  -- “AI data center by self-adapting ML/DL technologies”

Thank you for your attention!

More Information is available from https://abci.ai/

ABCI