ABCI
AI Bridging Cloud Infrastructure

GREEN500 L2 Measurement on AI Bridging Cloud Infrastructure (ABCI)

Hirotaka Ogawa

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

Green 500 BoF, Dallas, TX, November 14th

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

東京大学 THE UNIVERSITY OF TOKYO AIST

Univ. Tokyo / Kashiwa II Campus

OPERATED SINCE AUGUST 1ST, 2018

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)

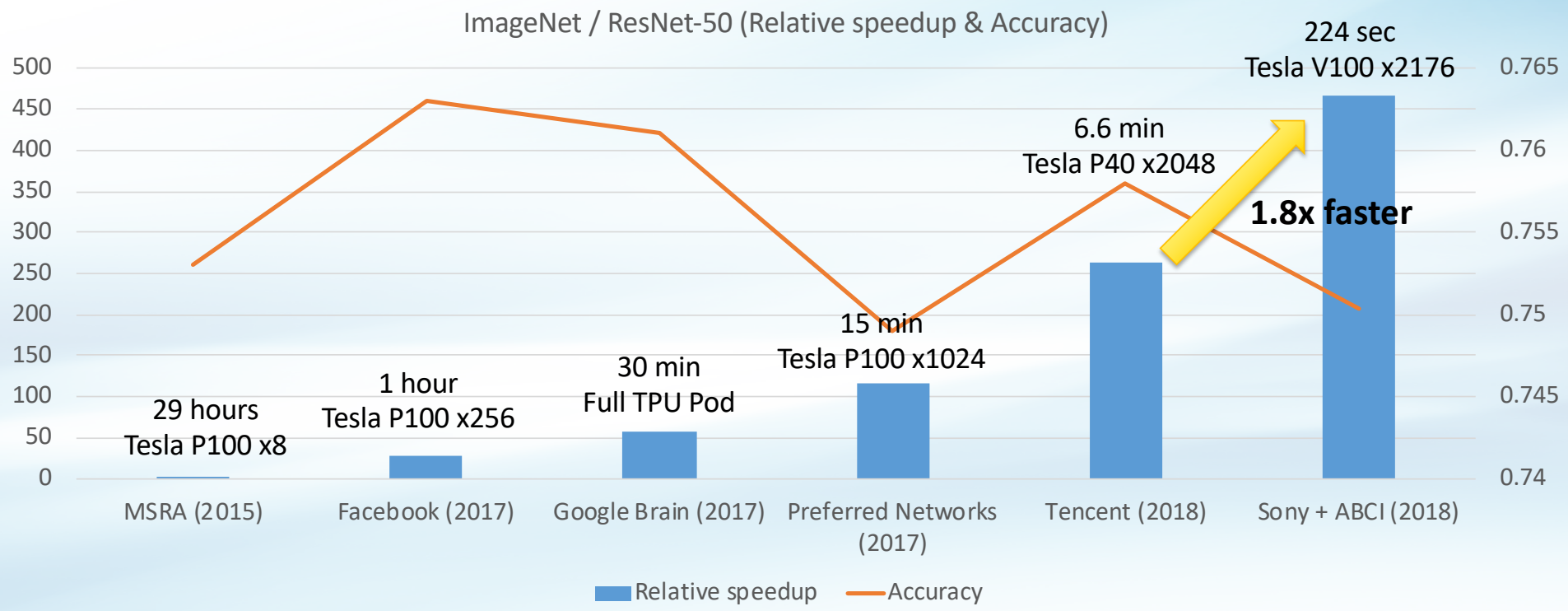
Effective Performance: (as of Nov 2018)
- 19.88 PFLOPS (#7 in TOP500)
- 14.423 GFLOPS/W (#4 in GREEN500)
- 508.85 TFLOPS (#5 in HPCG)

Power Usage: < 2.3 MW

Average PUE: < 1.1 (Estimated)

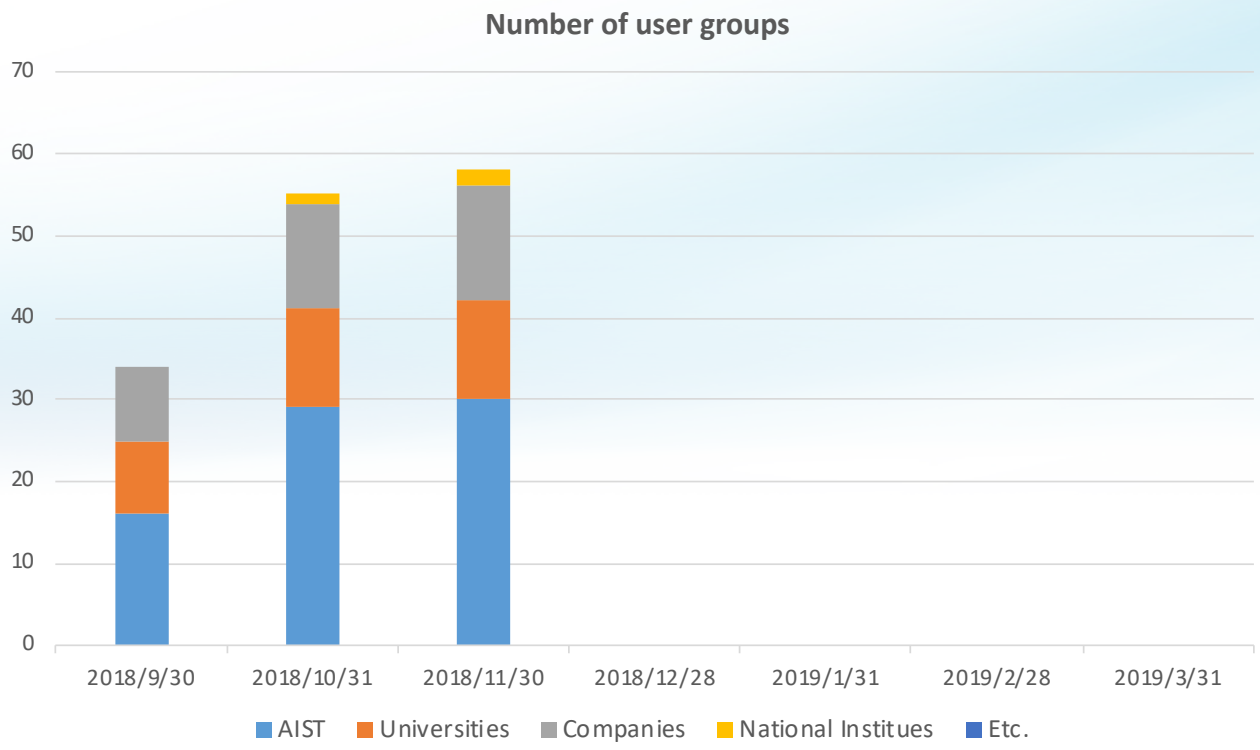
SC18 Updates

- News: Sony and AIST/ABCI achieve World's Fastest Deep Learning speeds through distributed learning (https://www.sony.net/SonyInfo/News/Press/201811/18-092E/index.html)



Fact

- ABCI operation began August 1st.
- Now serves 50+ research projects and 400+ users (we need 10x projects and users).



AIST ABCI Hardware Overview

High-Performance Computing System

550 PFlops(FP16), 37.2 PFlops(FP64)
476 TiB Memory, 1.74 PB NVMe SSD

Computing Nodes (w/ GPU) x 1088

GPU	NVIDIA Tesla V100 SXM2 x 4
CPU	Intel Xeon Gold 6148 (2.4GHz/20cores) x 2
Memory	384GiB
Local Storage	Intel SSD DC P4600 (NVMe) 1.6TB x 1
Interconnect	InfiniBand EDR x 2

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

Interactive Nodes x 4

Management and Gateway Nodes x 15

Large-scale Storage System

22 PB GPFS

DDN SFA14K (w/ SS8462 Enclosure x 10) x 3

- 12TB 7.2Krpm NL-SAS HDD x 2400
- 3.84TB SAS SSD x 216
- NSD Server x 12

Protocol Nodes, etc x 8

Interconnect (Infiniband EDR)

- Mellanox CS7500 x 2
- Mellanox SB7890 x 229

Service Network (10GbE)

Gateway and Firewall

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

100Gbps

SINET5

ABCI High-Performance Computing System

BOOTH #2409
PLEASE CHECK MOCK-UP OF OUR COMPUTING NODE

Chips (GPU, CPU)

Tesla V100

Xeon Skylake-SP

Compute Node (4GPUs, 2CPUs)

PRIMERGY CX2570 M4

Node Chassis (2 Compute Nodes)

PRIMERGY CX400 M4

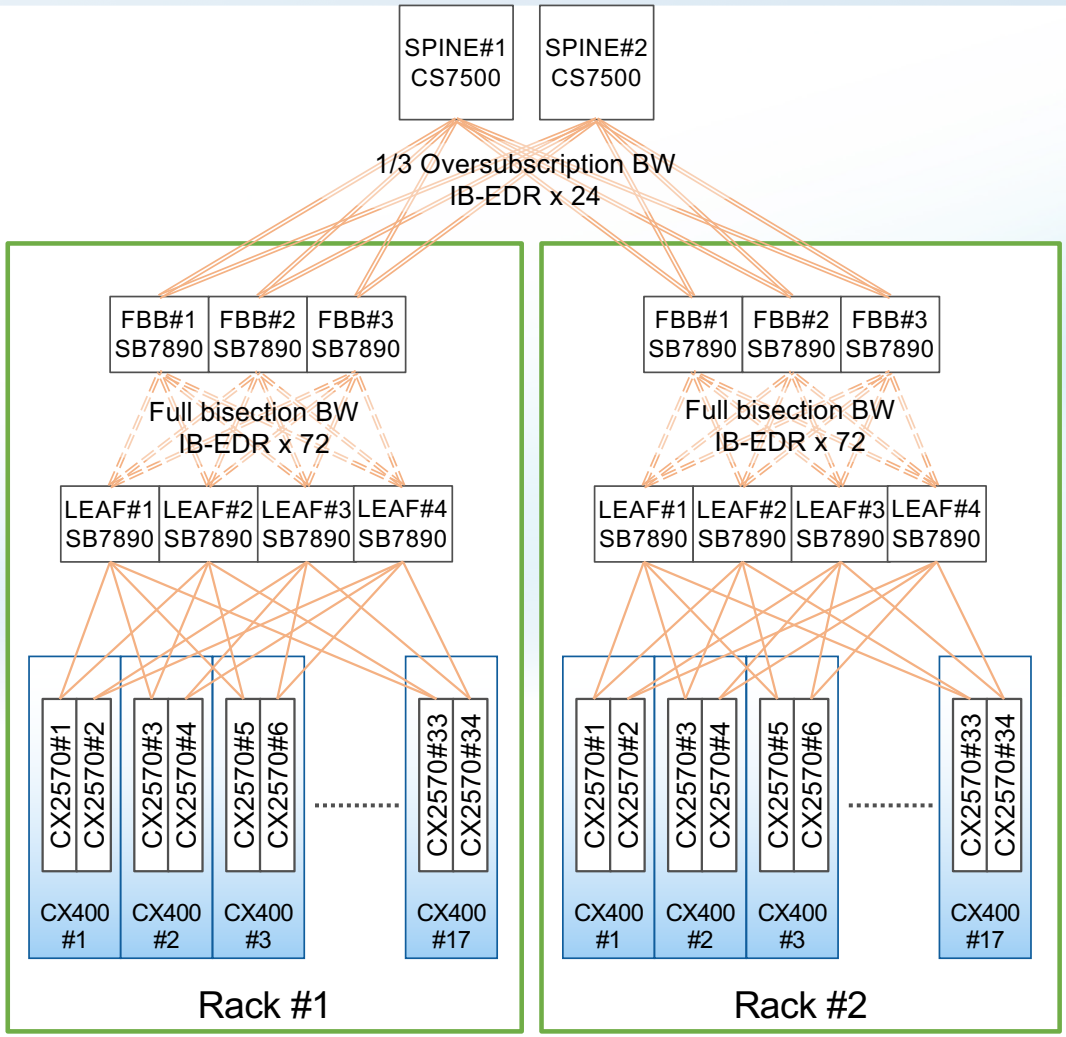
Rack (17 Chassis)

System (32 Racks)

1088 Compute Nodes
4352 GPUs

GPU:	CPU:				
7.8 TFlops(FP64)	**1.53 TFlops(FP64)**	**34.2 TFlops(FP64)**	**68.5 PFlops(FP64)**	**1.16 PFlops(FP64)**	**37.2 PFlops(FP64)**
125 TFlops(FP16)	**3.07 TFlops(FP32)**	**506 TFlops(FP16)**	**1.01 PFlops(FP16)**	**17.2 PFlops(FP16)**	**0.55 EFlops(FP16)**
		～3.72 TB/s MEM BW		～131TB/s MEM BW	～4.19 PB/s MEM BW
NVIDIA Tesla V100 (16GB SMX2)		384 GiB MEM 200 Gbps NW BW 1.6TB NVMe SSD		Full Bisection BW within Rack 70kW Max	1/3 of Oversubscription BW 2.3MW

ABCI Computing Rack / Interconnect

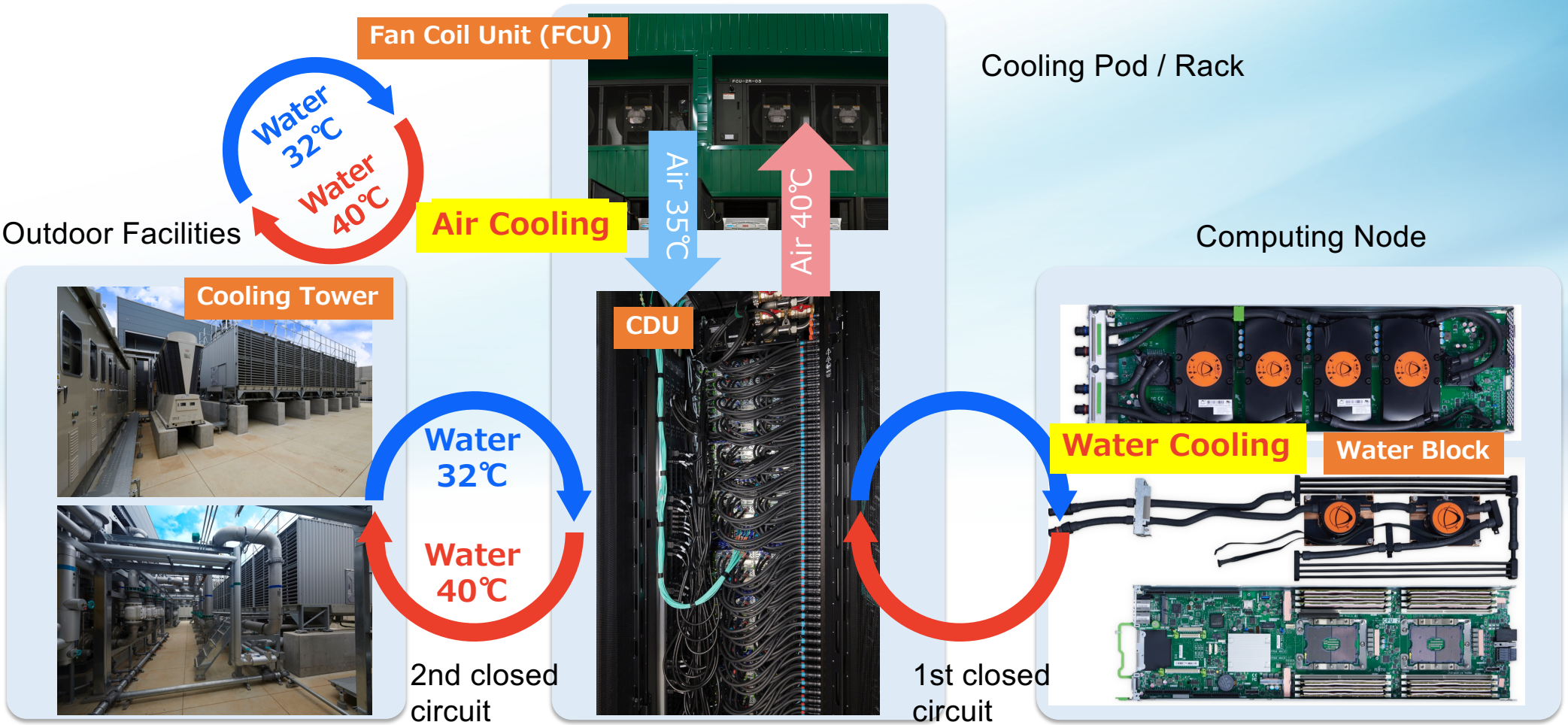


- **Dense-packaged rack: 34 nodes, 136 Tesla V100**
  - Theoretical peak performance per rack : 1.16 PFlops (FP64), 17 PFlops (FP16) c.f. Google TPU 3.0 Pod (>100PFlops/8racks)
  - Power consumption per rack: 67.33 kW
- **Interconnect**
  - Fat-tree topology
  - Intra-rack: full bisection BW
  - Inter-rack : 1/3 over-subscription (2400/6800)
  - Large-scale storage system: full bi-section BW

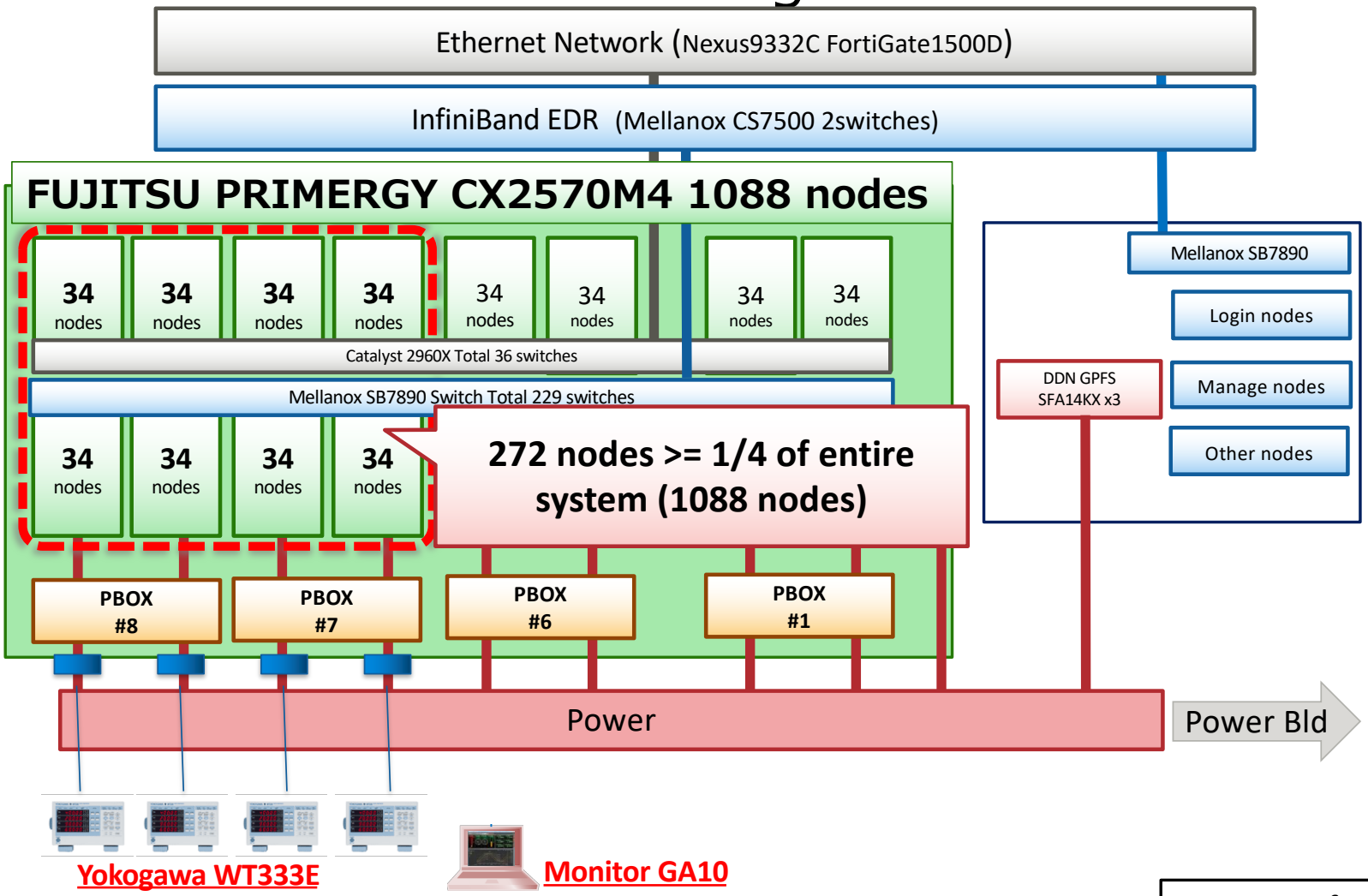
.....

InfiniBand EDR x4
InfiniBand EDR x6
InfiniBand EDR x1

Hybrid Water/Air Cooling System



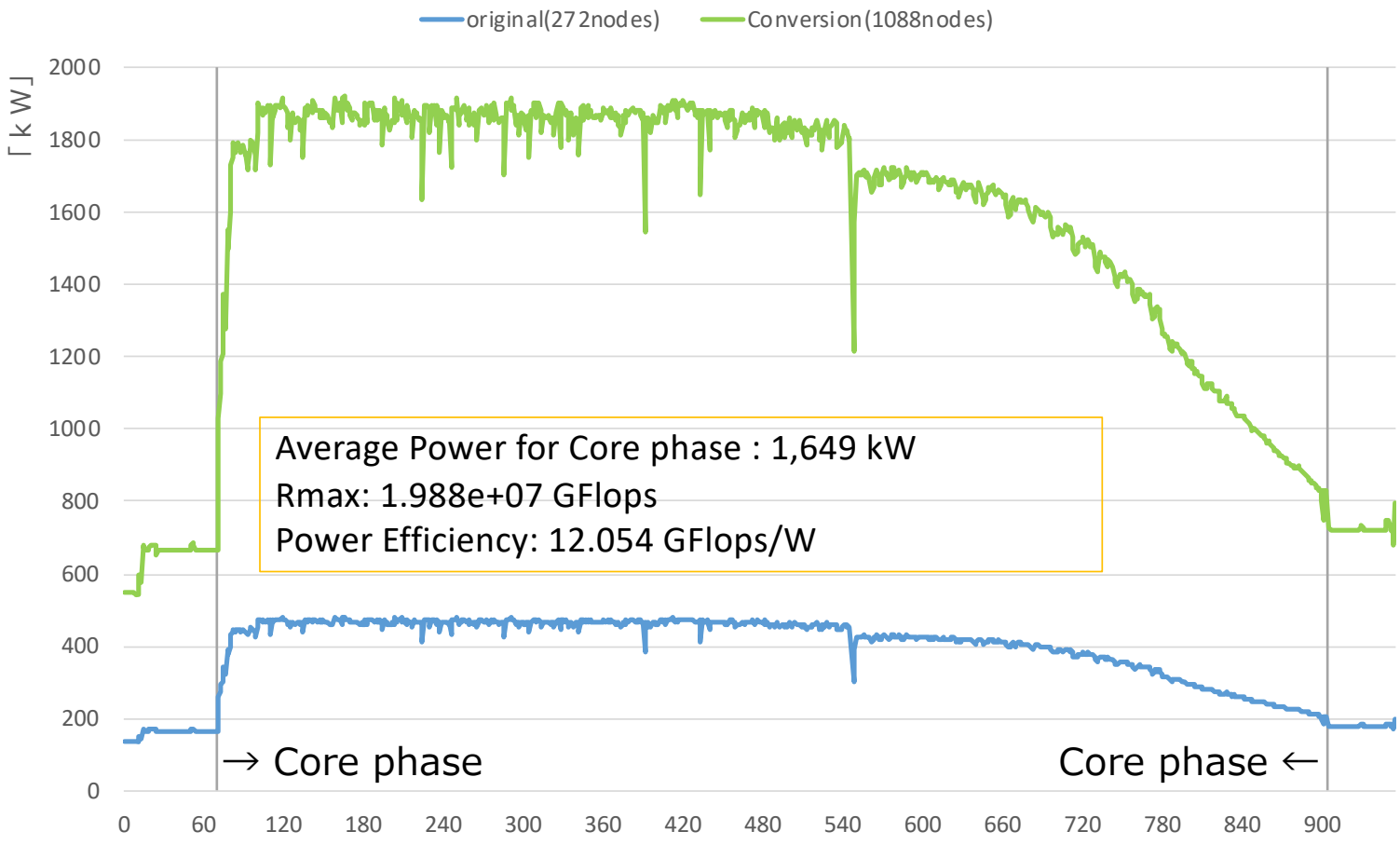
Power Monitoring on ABCI



Courtesy of FUJITSU LIMITED

Transition of Power Consumption on ABCI

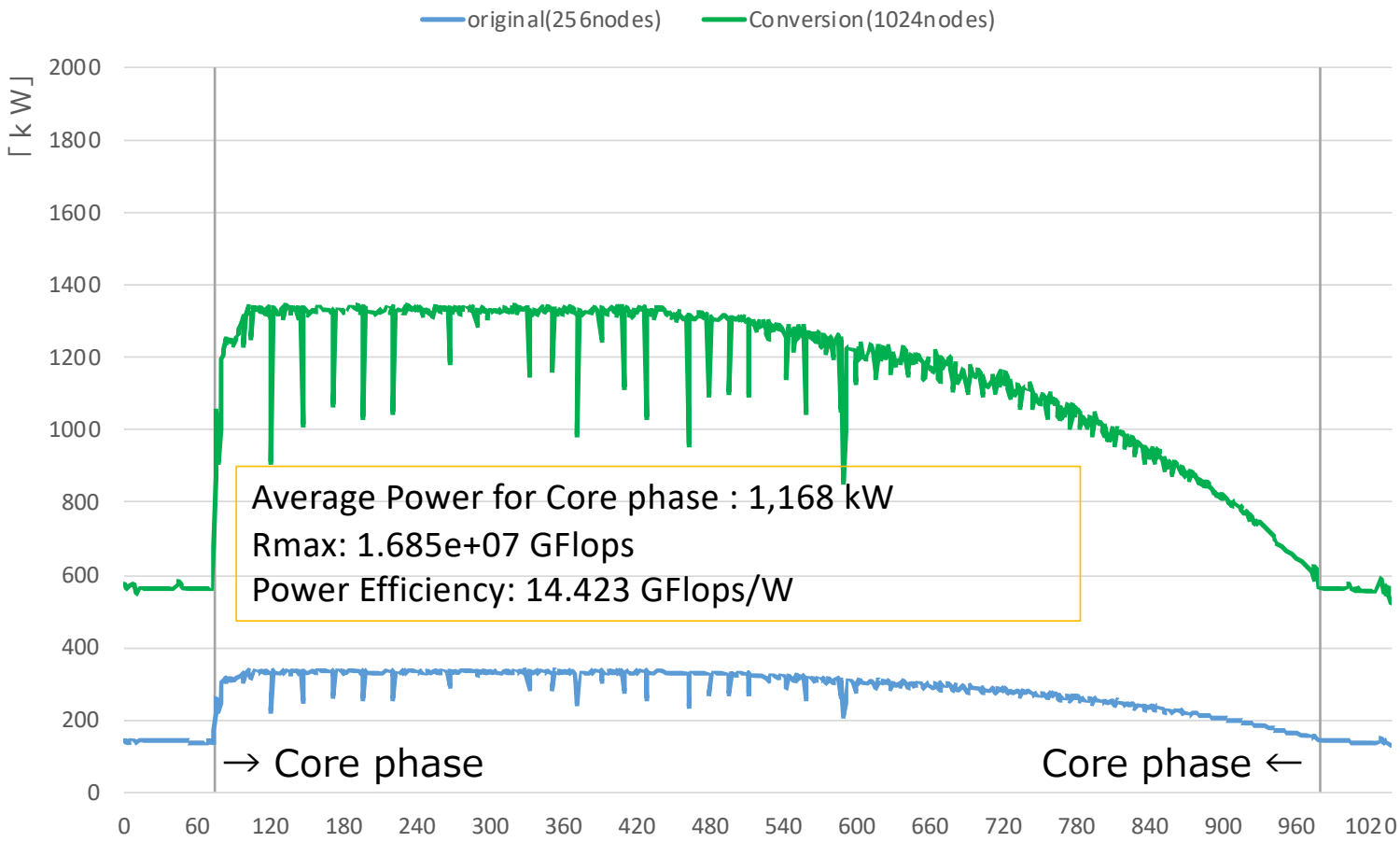
@ISC18 Measurement in June 2018



Courtesy of FUJITSU LIMITED

Transition of Power Consumption on ABCI

@SC18 Measurement in Sep 2018



Courtesy of FUJITSU LIMITED

Comments from Fujitsu

- Performance optimizations
  - Problem contained in small GPU memory
  - Communication optimizations are important to improve the performance for small memory HPL
  - GPU clock should be fixed on the suitable clock for avoiding load imbalance
- Preparation
  - HW screening is important to avoid performance degradation
- During measurement
  - Support systems are established for the rapid recovery
  - Power monitoring makes it easy to find out problems
  - Nature condition: The performance is better during night or early morning with green cooling system

ABCI

https://abci.ai/