ABCI
AI Bridging Cloud Infrastructure

ABCI: An Open Innovation Platform for Advancing AI Research and Deployment

Ryousei Takano

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

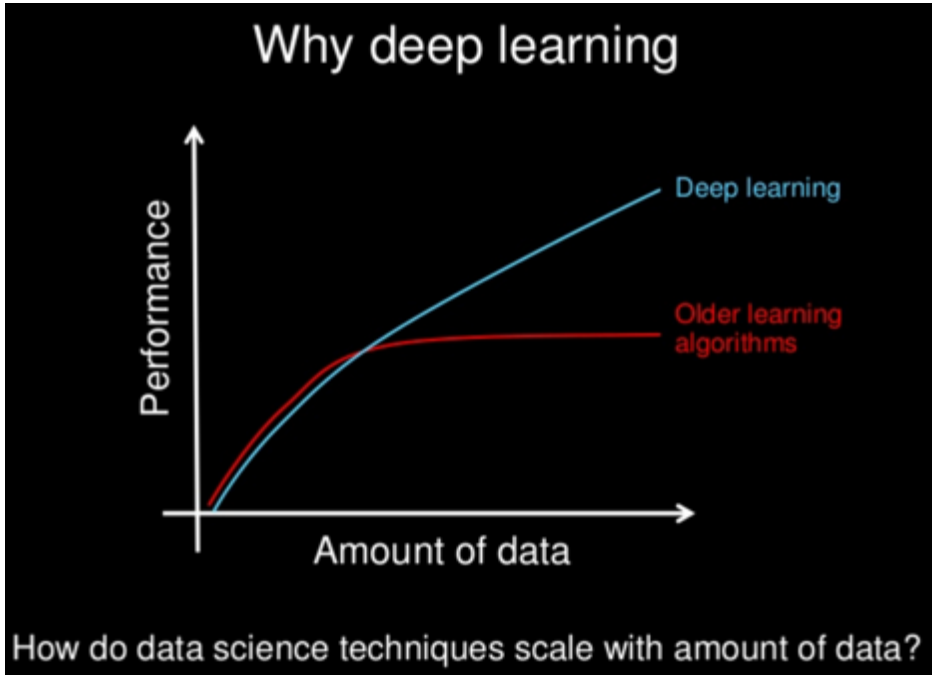
PRAGMA Workshop 35, Penang, October 4th

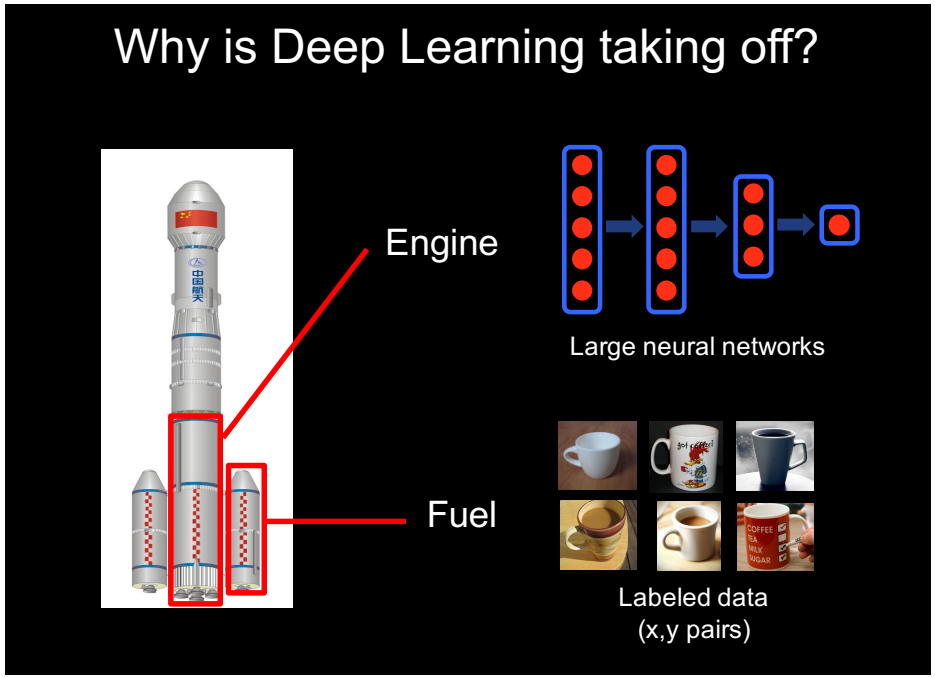
Outline

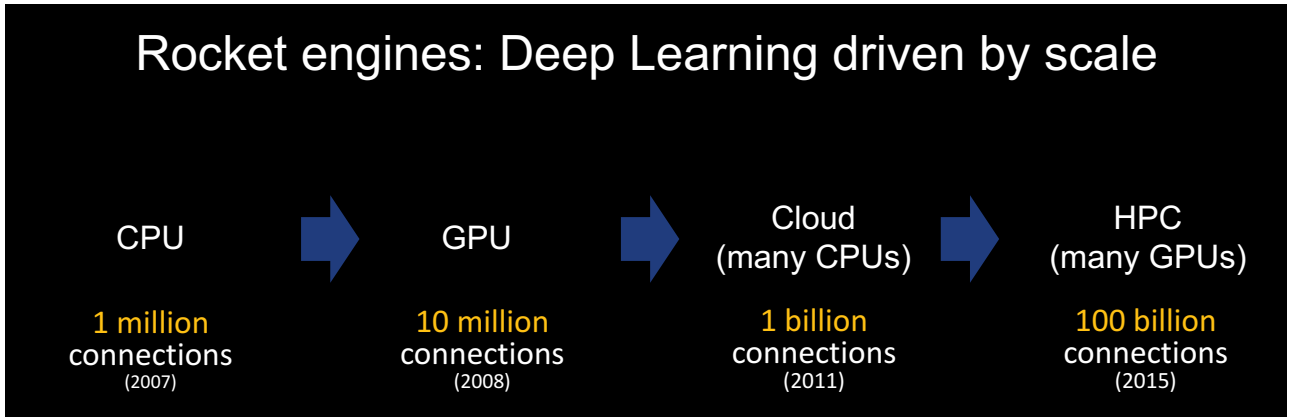
- Open Innovation Platform for advancing AI R&D
- ABCI: AI Bridging Cloud Infrastructure
  - Hardware Architecture
  - Software Stack
  - Services
  - Data Center Facilities
- International Collaboration

Deep Learning

- "Deep Learning is scalable"
- "Performance just gets better if you feed in more data"







Andrew Ng (Baidu) "What Data Scientists Should Know about Deep Learning"

Open Innovation Platform for advancing AI R&D

Common Public Platform

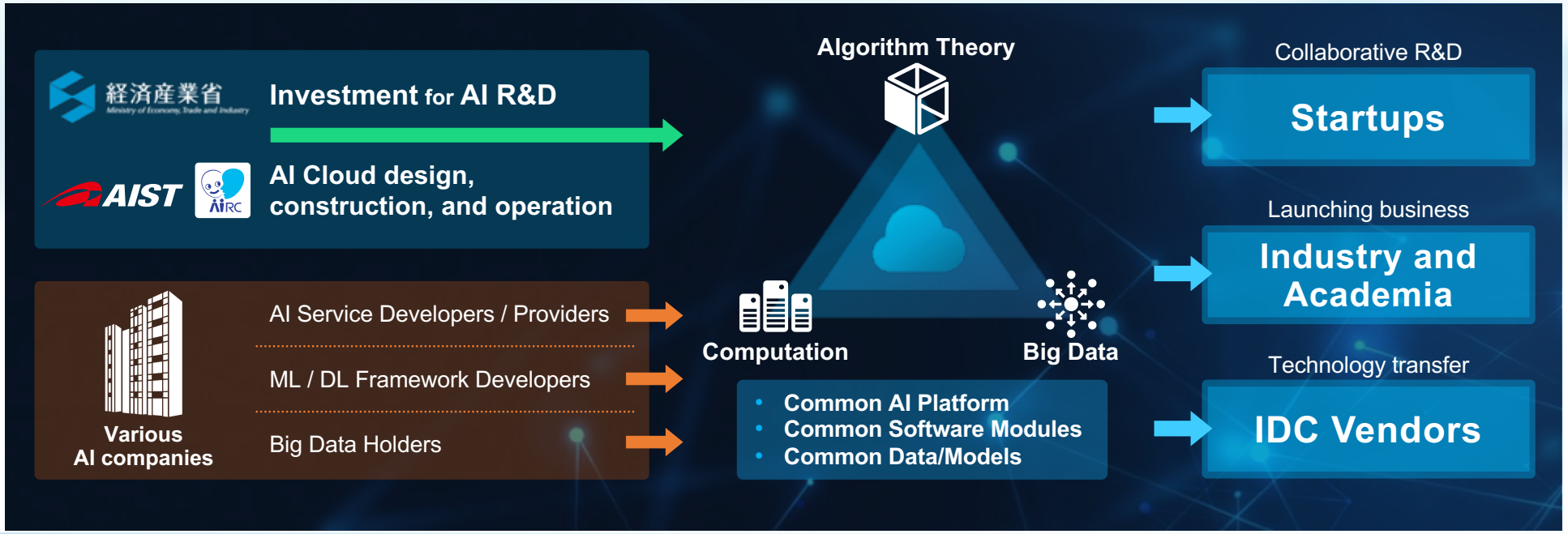
- For AI apps, services, and infrastructure designs
- Aiming fast tech transfer through industry and academia collaboration

Open Hardware and Software Architecture

- w/ AI acceleration support based on commodity devices

Multi-PB-class Sharable Big Data Storage

- For AI R&D collaboration



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

東京大学 THE UNIVERSITY OF TOKYO AIST

Univ. Tokyo / Kashiwa II Campus

ABCI AI Bridging Cloud Infrastructure

- World Top-Level compute and data process capability
- **Open, Public, and Dedicated** infrastructure for Al & 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 (#5 in TOP500)
12.054 GFlops/W (#8 in GREEN500)
Power Usage: < 2.3 MW
Average PUE: < 1.1 (Estimated)

TOP500 List (June 2018)

#	System	Architecture	Country	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	Summit, ORNL	IBM, POWER9 + GPU(GV100)	USA	122,300.0	187,659.3	8,806
2	TaihuLight, NSCW	Sunway, SW26010	China	93,014.6	125,435.9	15,371
3	Sierra, LLNL	IBM POWER9 + GPU(GV100)	USA	71,610.0	119,193.6	-
4	Tianhe-2, NSCG	NUDT, CPU + Matrix-2000	China	61,444.5	100,678.7	18,482
5	ABCI, AIST	Fujitsu, CPU + GPU(V100)	Japan	19,880.0	32,576.6	1,649
6	Piz Daint, CSCS	Cray, CPU + GPU(P100)	Switzerland	19,590.0	25,326.3	2,272
7	Titan, ORNL	Cray, CPU + GPU(K20x)	USA	17,590.0	27,112.5	8,209
8	Sequoia, LLNL	IBM, BlueGene/Q	USA	17,173.2	20,132.7	7,890
9	Trinity, LANL/SNL	Cray, MIC(KNL)	USA	14,137.3	43,902.6	3,844
10	Cori, NERSC-LBNL	Cray, MIC(KNL)	USA	14,014.7	27,880.7	3,939

AI Infrastructure for Everyone (Democratization AI)

Expert

- ABCI Grand Challenge: Demonstration of highly challenging academic and/or industrial themes using the whole ABCI resources for 24 hours
- ABCI Data Challenge: Competition of accelerating open science using open data

Advanced & Intermediate

- Up to 512-node computing resource is available for everyone
- Software, datasets, and pre-trained models are ready for use
- High computing capability enables to accelerate AI R&D and promote social implementation

Beginner

- User friendly WebUI based IDE for supporting beginners of deep learning
- PoC platform of B2B2C

Building ecosystem of AI R&D from beginners to state-of-the-art researches

A reference design
Technology transfer

IDC

Promoting the use by over 100 institutions and over 1000 researchers/engineers

AI Infrastructure by co-Design

High Density Datacenter Design

- Cost-effective lightweight “warehouse” building and cooling pod
- **x20 thermal density of standard IDC**

Ultra Green

- Free cooling, and high-efficiency power supplies, etc.
- **Commoditizing supercomputer cooling technologies to Clouds (70KW/rack)**

De facto and Commodity Architecture

- Wide-ranging Big Data and HPC standard software stacks
- AI accelerator support based on commodity software

Container-optimized Software Ecosystem

- Advanced cloud-based operation
- Convergence of HPC and AI/Big Data software stack

Secure Data Utilization

- Promoting industrial R&D using private data sets

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)

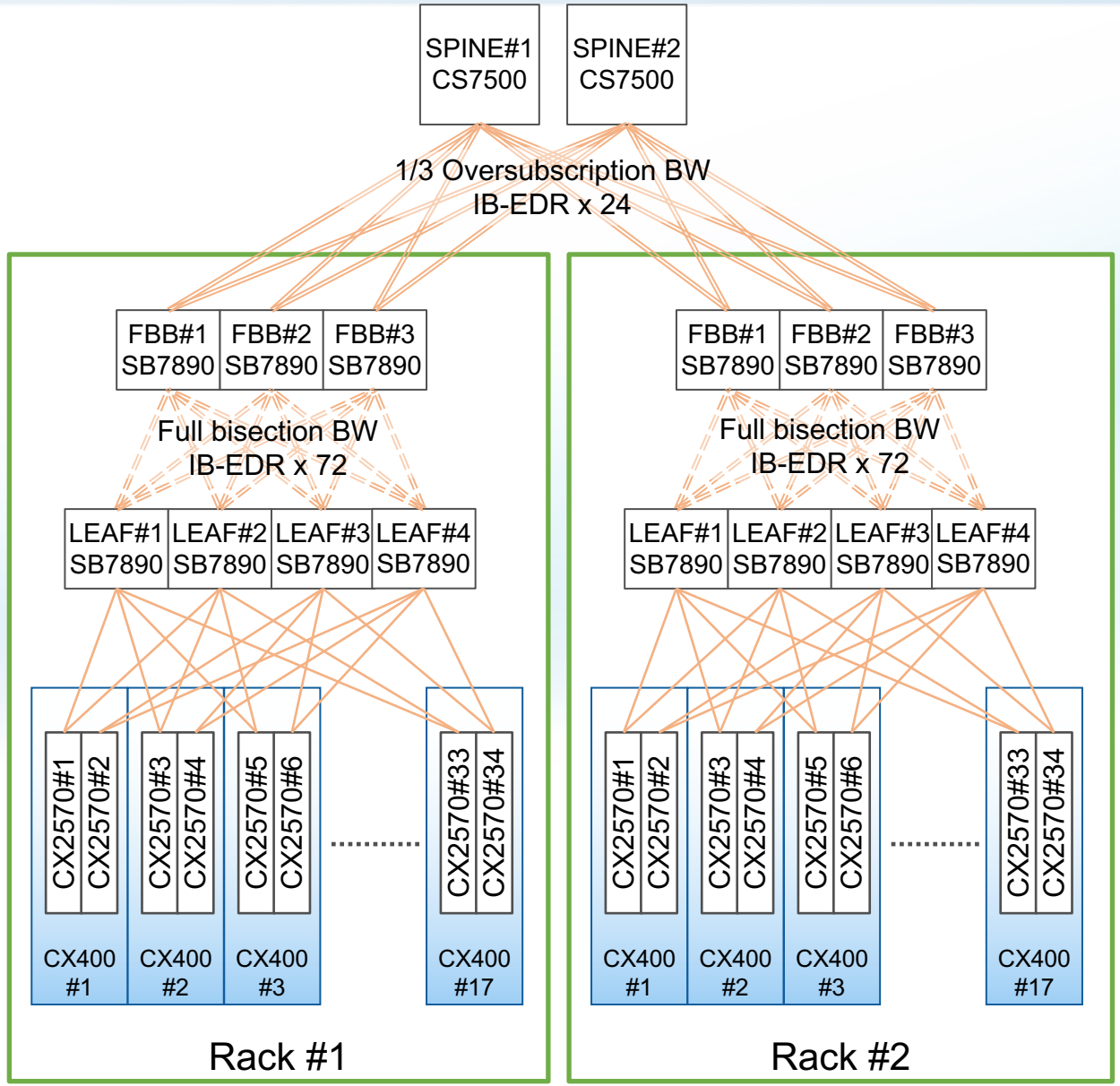
Gateway and Firewall

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

100Gbps

SINET5

ABCI Computing Rack / Interconnect



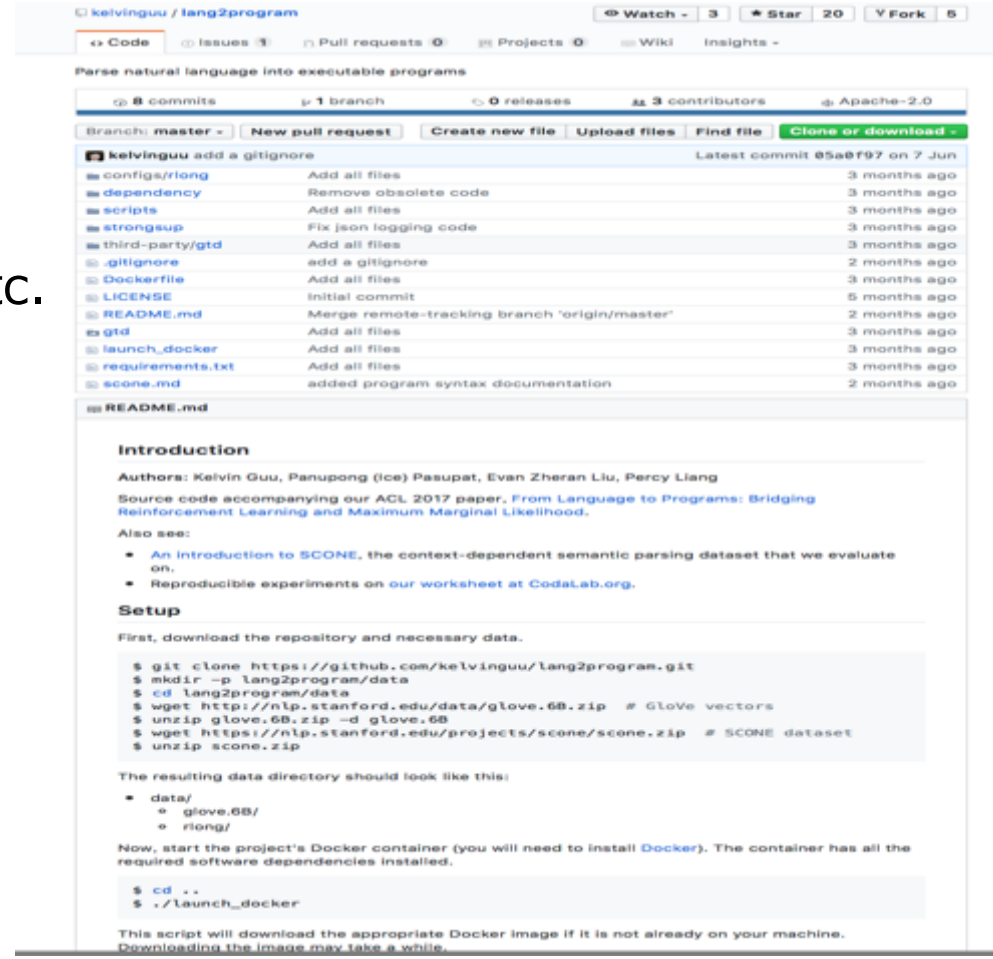
- **Dense-packaged rack: 34 nodes, 136 Tesla V100**
  - Theoretical peak performance per rack : 1.16 PFlops (FP64), 17 PFlops (FP16) c.f. Google TPU 3.0 Pod (>100PFlops)
  - Power consumption per rack: 67.33 kW
- **Interconnect**
  - Fat-tree topology
  - Intra-rack: full bisection BW
  - Inter-rack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

.....

InfiniBand EDR x4
InfiniBand EDR x6
InfiniBand EDR x1

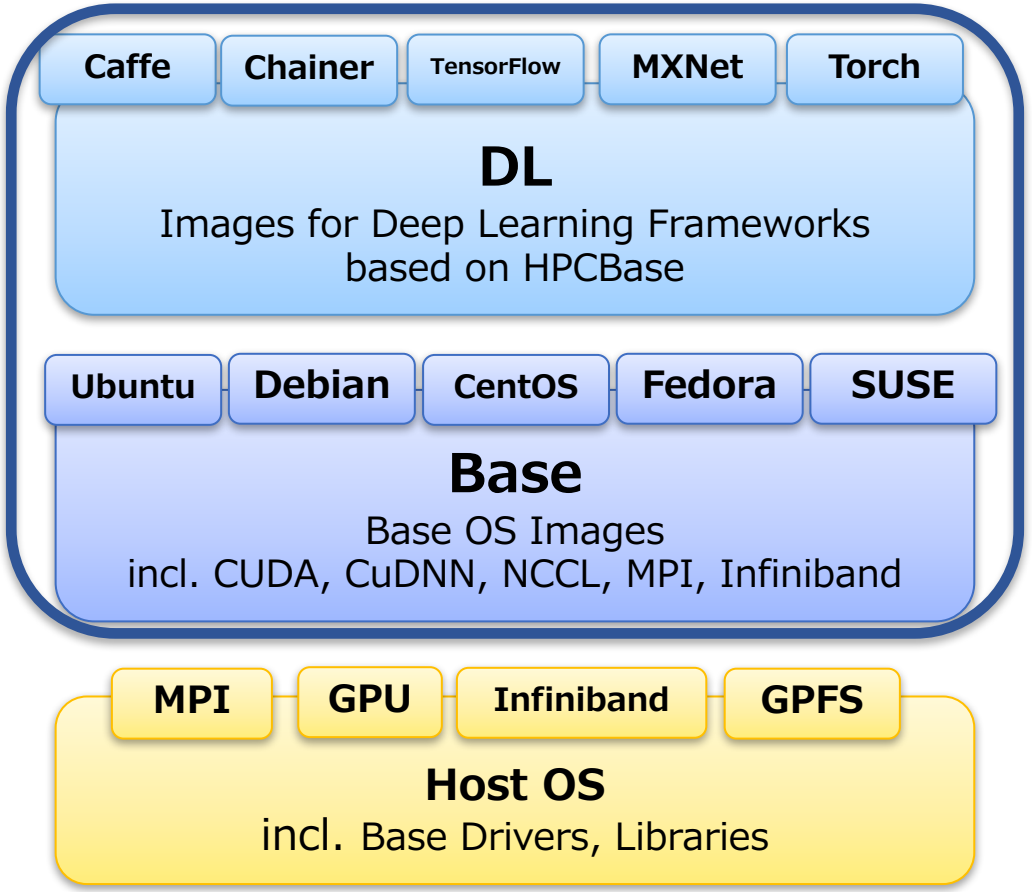
Container is critical for AI R&D

- lang2program (referred in ACL2017)
  - https://github.com/kelvinguu/lang2program
  - Provided as a Dockerfile
  - Bunch of software needed to be run
    - Tensorflow, PostgresQL, Python Pip Packages, etc.
- It is not possible to run it on traditional large-scale HPC systems
  - Arbitrary/Voluntary installation of software => chaos
  - Docker => security reason
- We're developing an easy-to-manage and flexible-to-use platform for deploying AI apps as "modules"



AI Frameworks & Modules

- Host OS only provides the minimum set of software including:
  - Base drivers, libraries
- “Base” modules provide customized OS images including:
  - CUDA, cuDNN, NCCL, MPI, etc.
- “DL” modules provide deep learning frameworks and apps, which extend “Base” images.
- We mainly employ Singularity as the basis of our AI platform.



ABCI Software Stack

Software

Operating System	CentOS, RHEL
Job Scheduler	Univa Grid Engine
Container Engine	Docker, Singularity
MPI	OpenMPI, MVAPICH2, Intel MPI
Development tools	Intel Parallel Studio XE Cluster Edition, PGI Professional Edition, NVIDIA CUDA SDK, GCC, Python, Ruby, R, Java, Scala, Perl
Deep Learning	Caffe, Caffe2, TensorFlow, Theano, Torch, PyTorch, CNTK, MXnet, Chainer, Keras, etc.
Big Data Processing	Hadoop, Spark

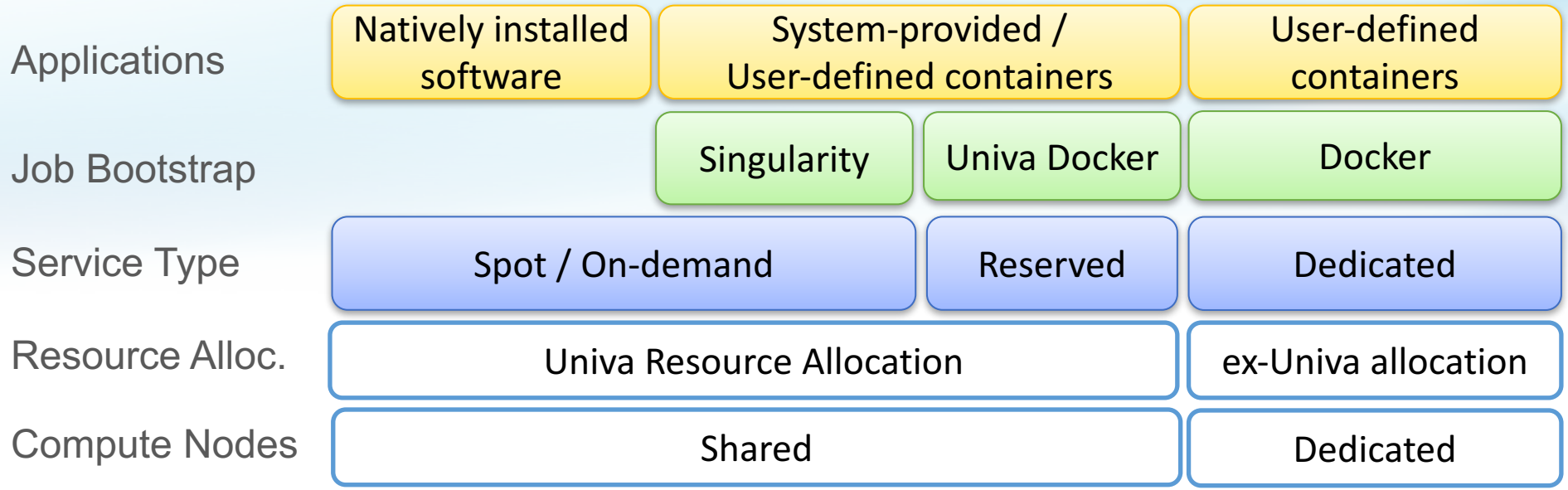
Container support

- Containers enable users to instantly try the state-of-the-art software developed in AI community
- ABCI supports two container technologies
  - **Docker**, having a large user community
  - **Singularity**, recently accepted HPC community
- ABCI provides various single-node/distributed deep learning framework container images optimized to achieve high performance on ABCI

ABCI Cloud Services

Service Types

Service Type	Description	#Nodes (Min./Max.)
Spot	Batch job service	1 / 512
On-demand	Interactive job service	1 / 32
Reserved	Advanced reservation service	1 / 32
Group Storage	Shared storage service	N/A



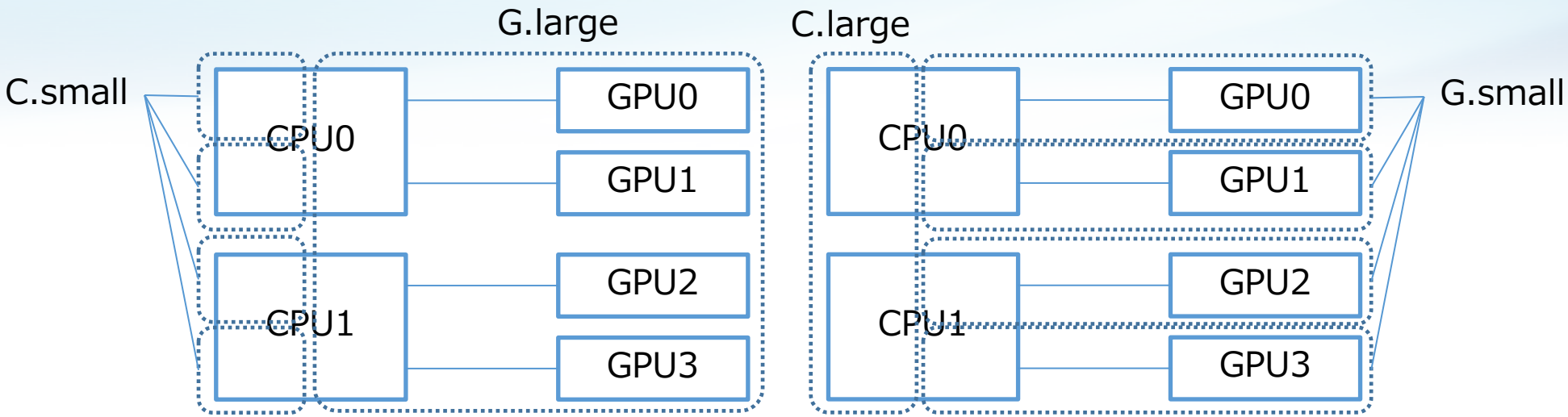
※ABCI provides batch and interactive-type job execution service for maximizing throughput, advanced reservation service for dedicated use of nodes, IDE, and storage services.

ABCI Cloud Services

Resource Types

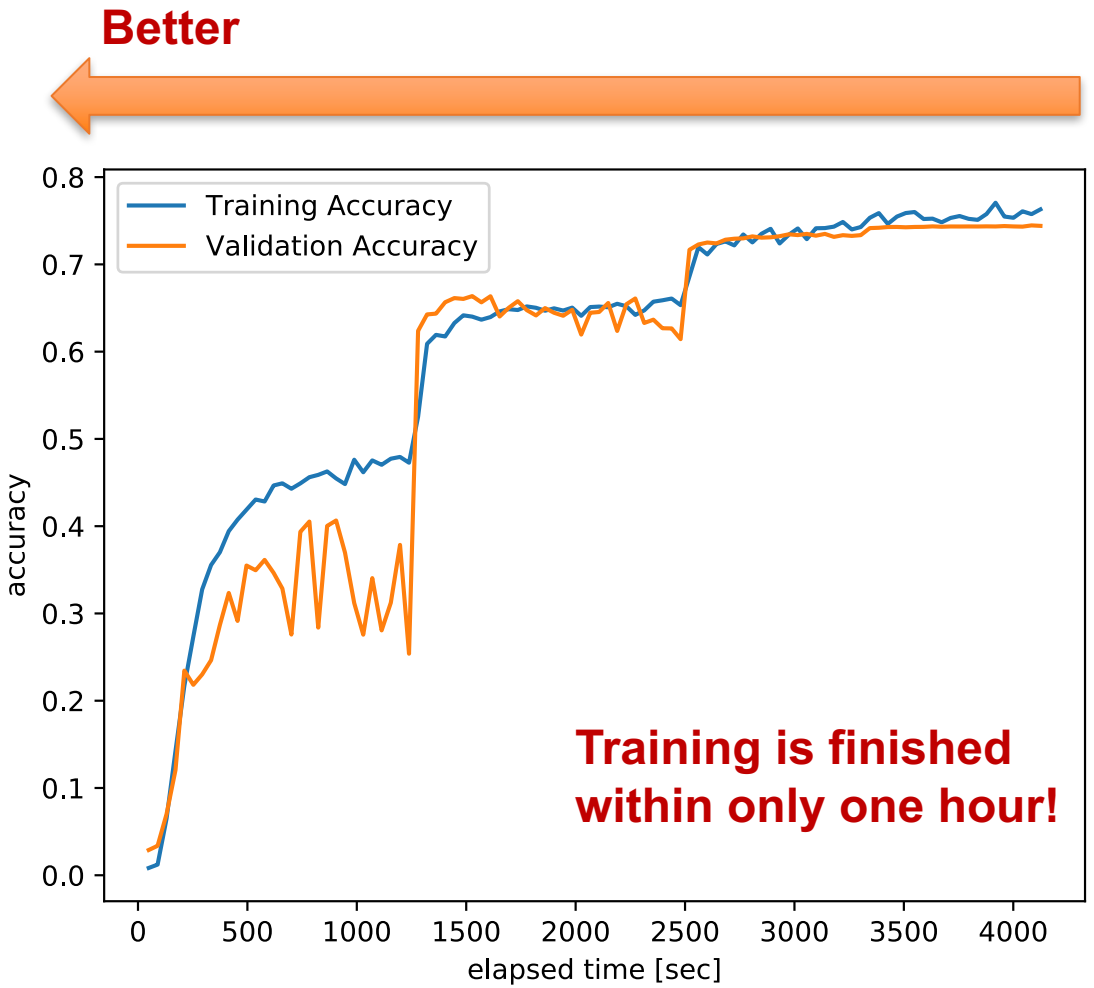
※Users can choose the most suitable computing instance from five resource types.

Types	#CPU core Assign / Total	#GPU Assign / Total	Memory (GB) Assign / Total	Storage (TB) Assign / Total
F (Full Node)	40 / 40	4 / 4	360 / 384	1.4 / 1.6
G.large	20 / 40	4 / 4	240 / 384	0.7 / 1.6
G.small	5 / 40	1 / 4	60 / 384	0.175 / 1.6
C.large	20 / 40	0 / 4	120 / 384	0.7 / 1.6
C.small	5 / 40	0 / 4	30 / 384	0.175 / 1.6

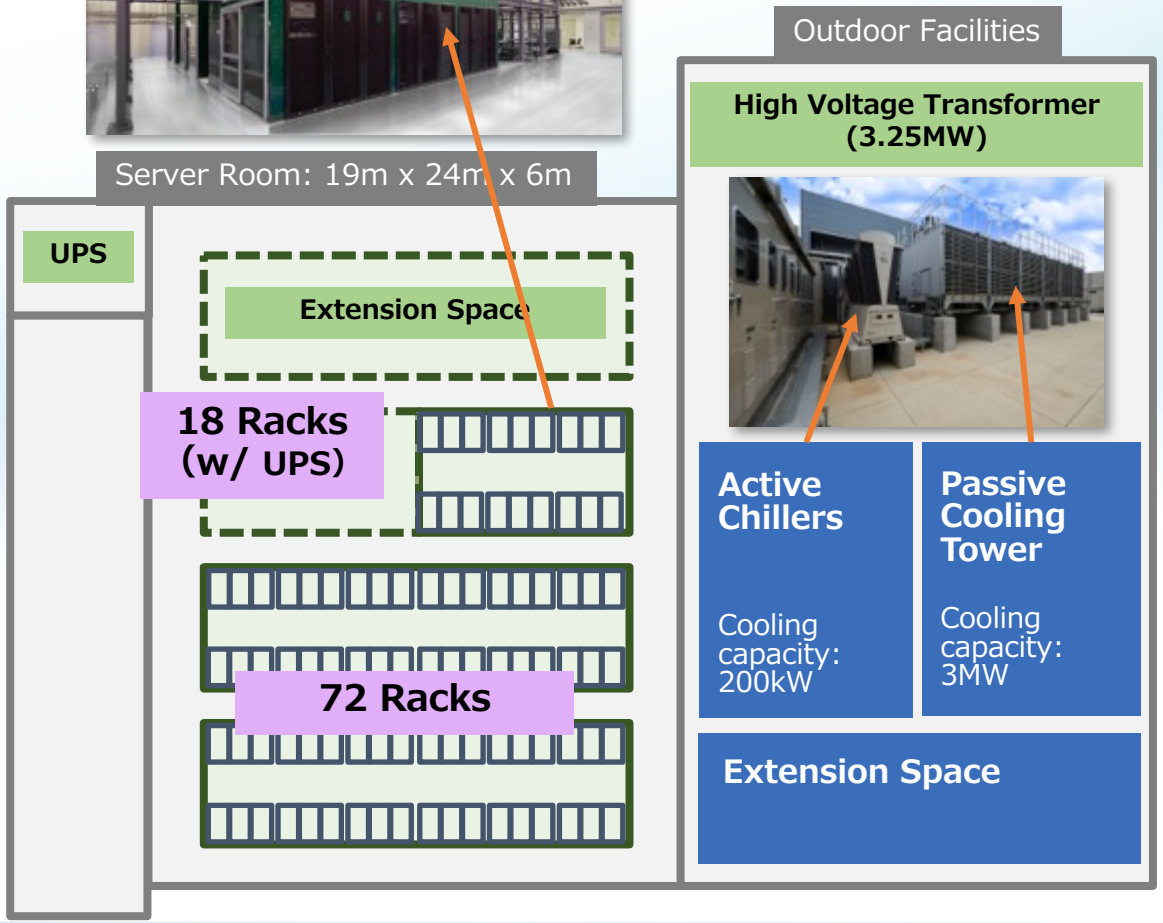


ABCI Use Case: Distributed Deep Learning

- 64 Nodes (256 GPUs) in 1 hour
- DL Framework: ChainerMN v1.3.0
- Data Set: ImageNet-1K
- Model: ResNet-50
  - Batch size: 32 x 256
  - Learning Rate: starting 0.1 and x0.1 at 30, 60, 90 epoch w/ warm up scheduling
  - Momentum SGD (momentum=0.9)
  - Weight decay: 0.0001
  - Training Epoch: 100



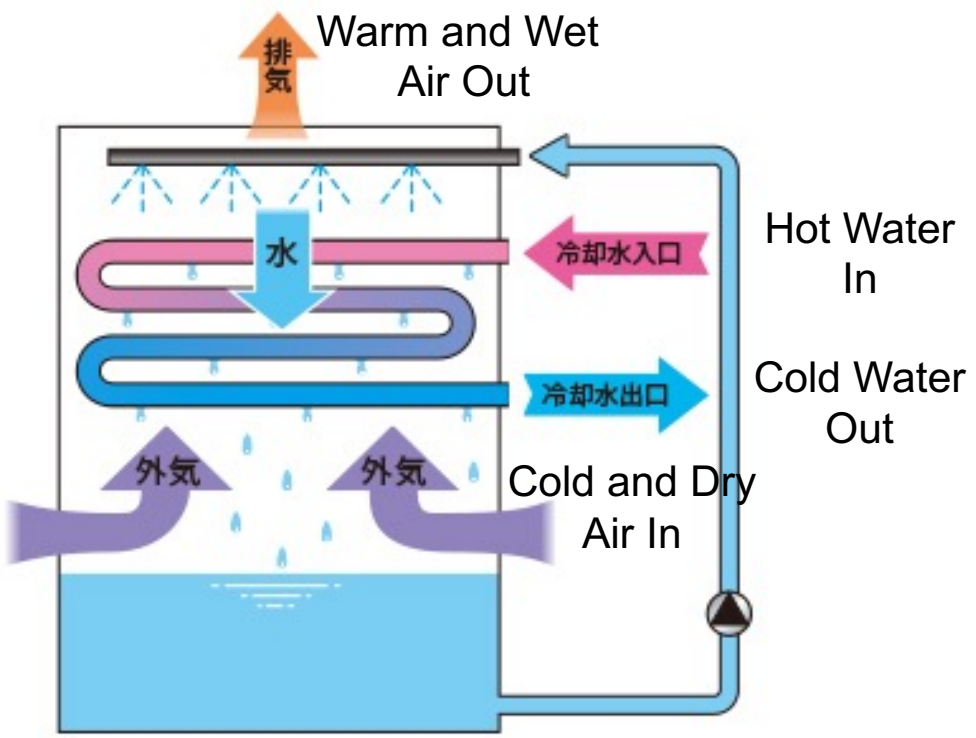
AI Datacenter
“Commoditizing supercomputer cooling technologies to Cloud (70kW/rack)”



- **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

100% Free Cooling over the Entire Year

- Free cooling using a passive cooling tower
  - ⭕ Low OPEX. No active chiller (mechanical refrigeration)
  - ❌ Generated water temperature is depended on the external weather (temp. and humidity)
- Hybrid water/air cooling with high-temperature cooling water (32℃)



https://www.sinko.co.jp/product/technical-column/09/

Hybrid Water/Air Cooling System

Fan Coil Unit (FCU)

Cooling Pod / Rack

Water 32℃

Water 40℃

Outdoor Facilities

Air Cooling

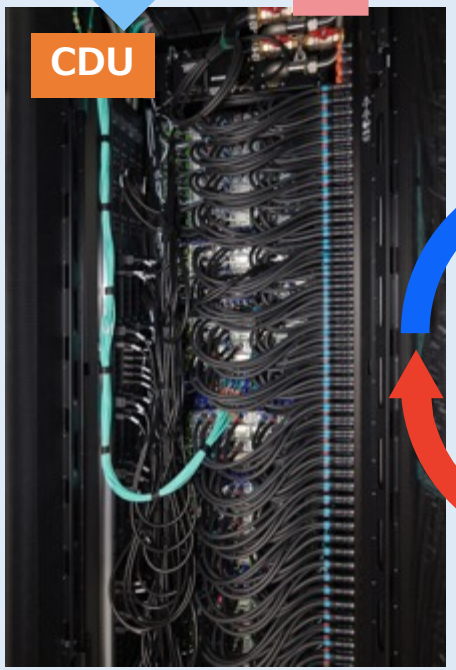
Air 35℃

Air 40℃

Computing Node

Cooling Tower

CDU



Water 32℃

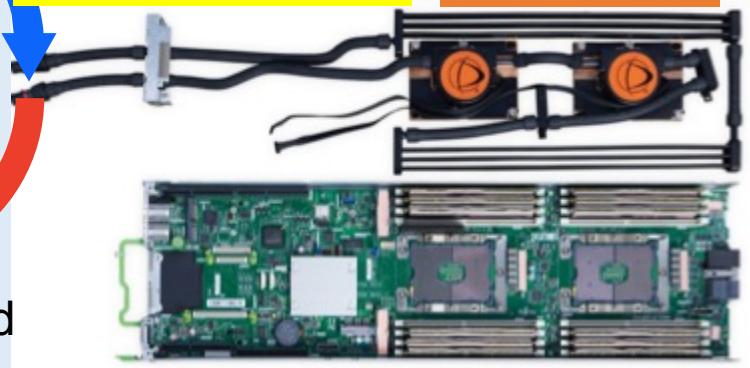
Water 40℃

2nd closed circuit

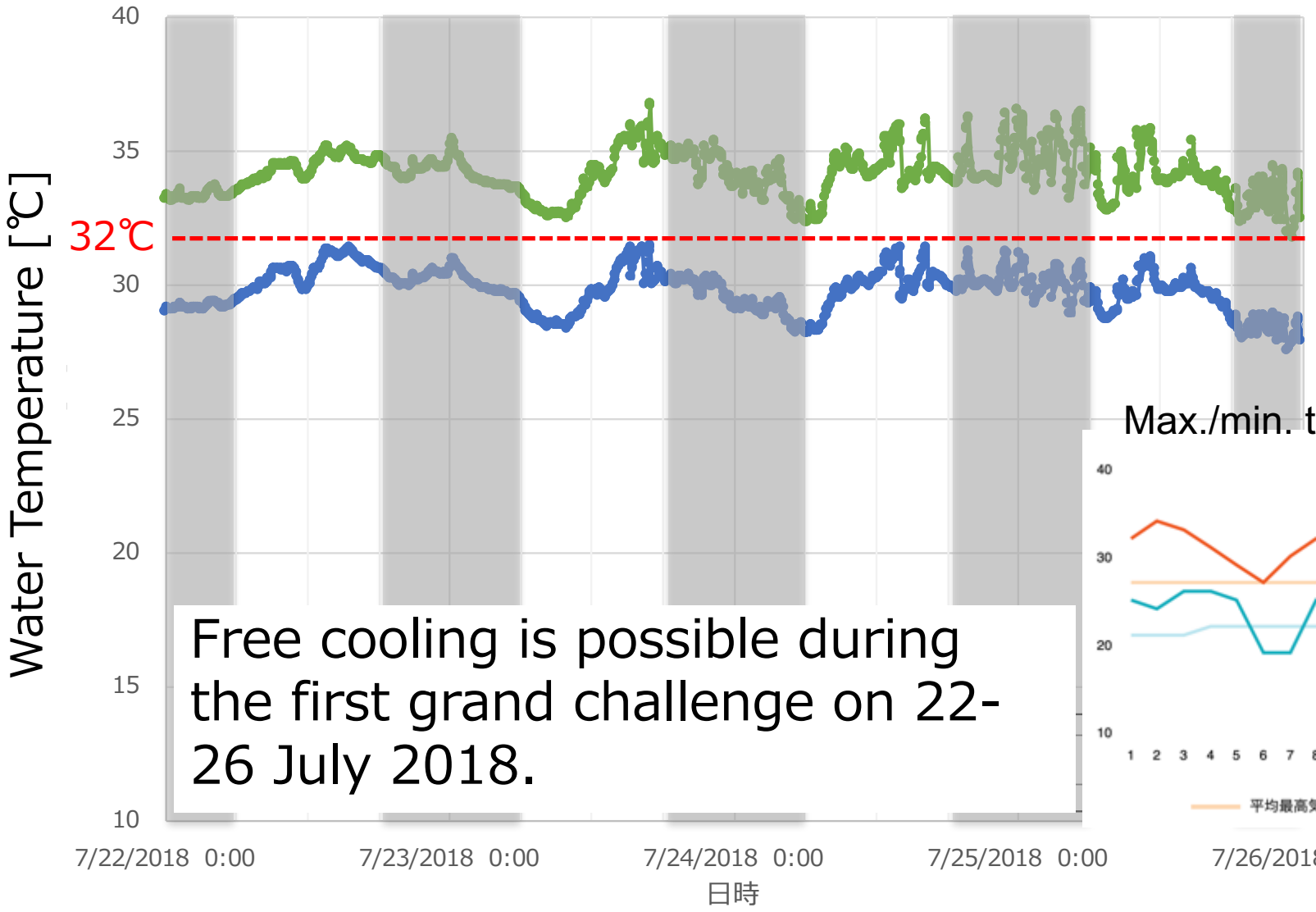
1st closed circuit

Water Cooling

Water Block



Survived the extreme heat this summer!

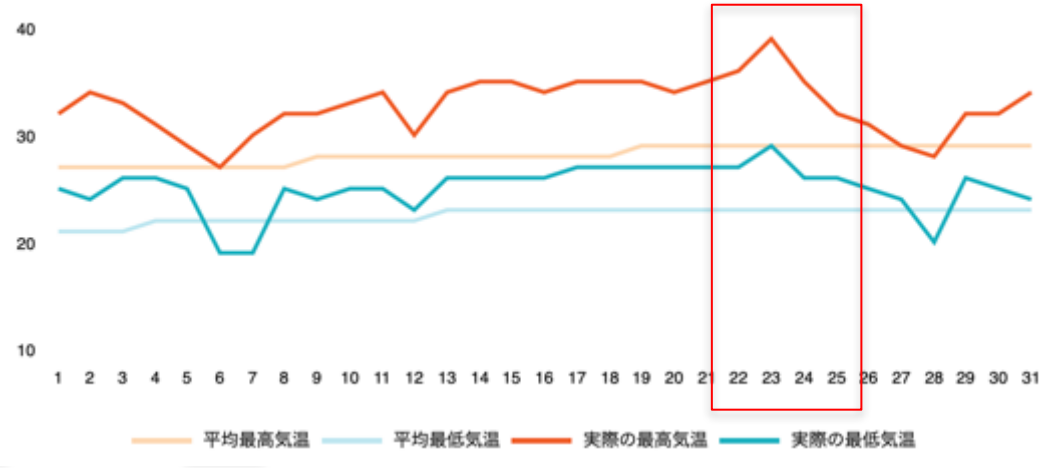


Free cooling is possible during the first grand challenge on 22-26 July 2018.

Cooling Water (From ABCI): <40℃

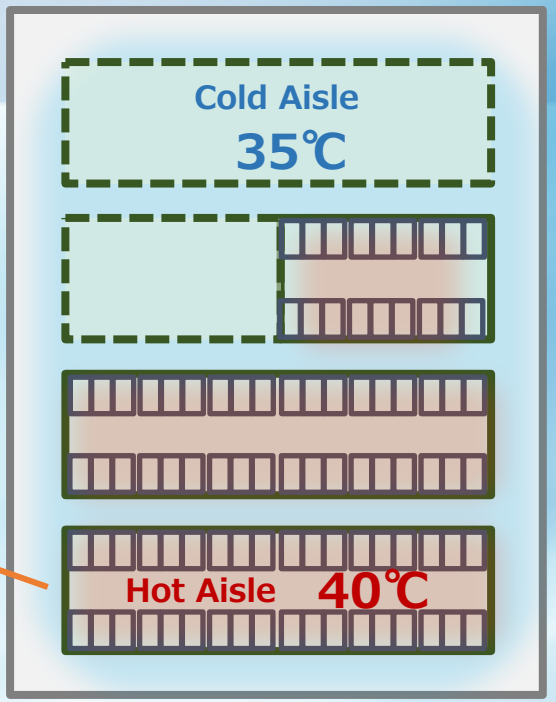
Cooling Water (To ABCI): <32℃

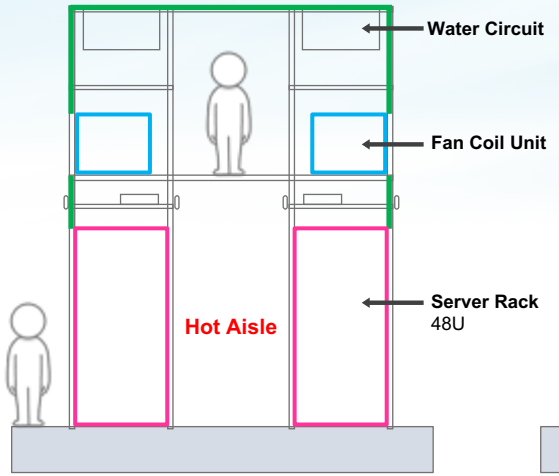
Max./min. temperature in Kashiwa in July 2018



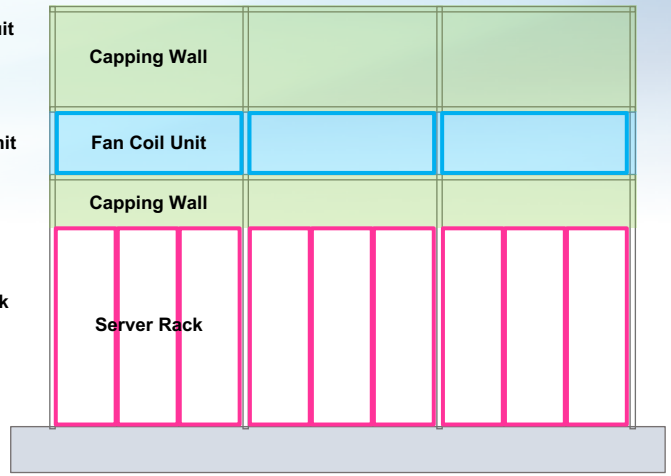
Cooling Pod

- Modularization of DC facilities: rack space, water/air cooling and power equipment
- No raised floor and skeleton frame built on concrete slab
- Effective air cooling by hot aisle containment
- Ease of maintenance



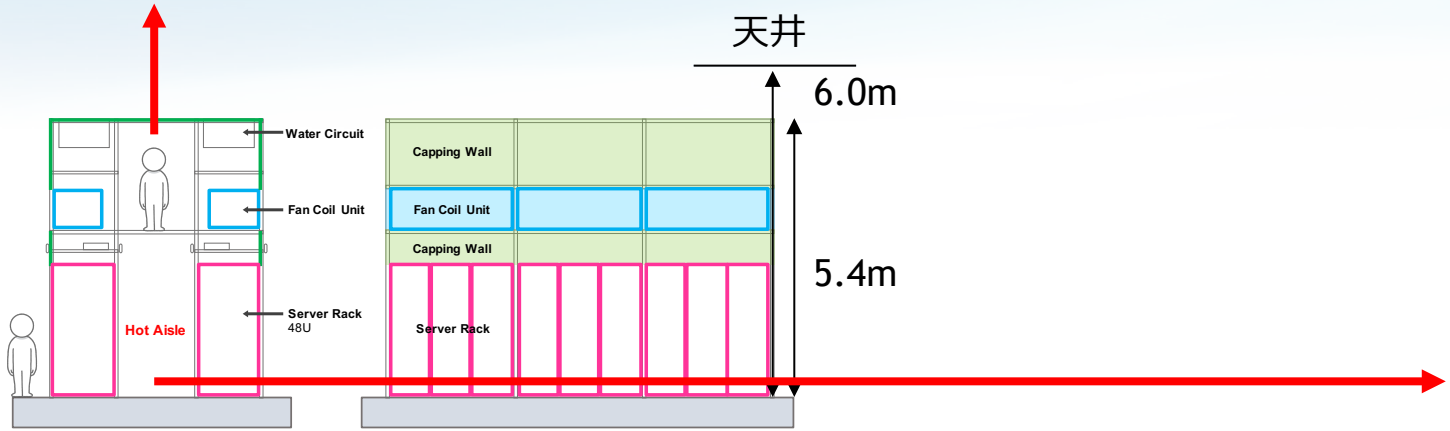


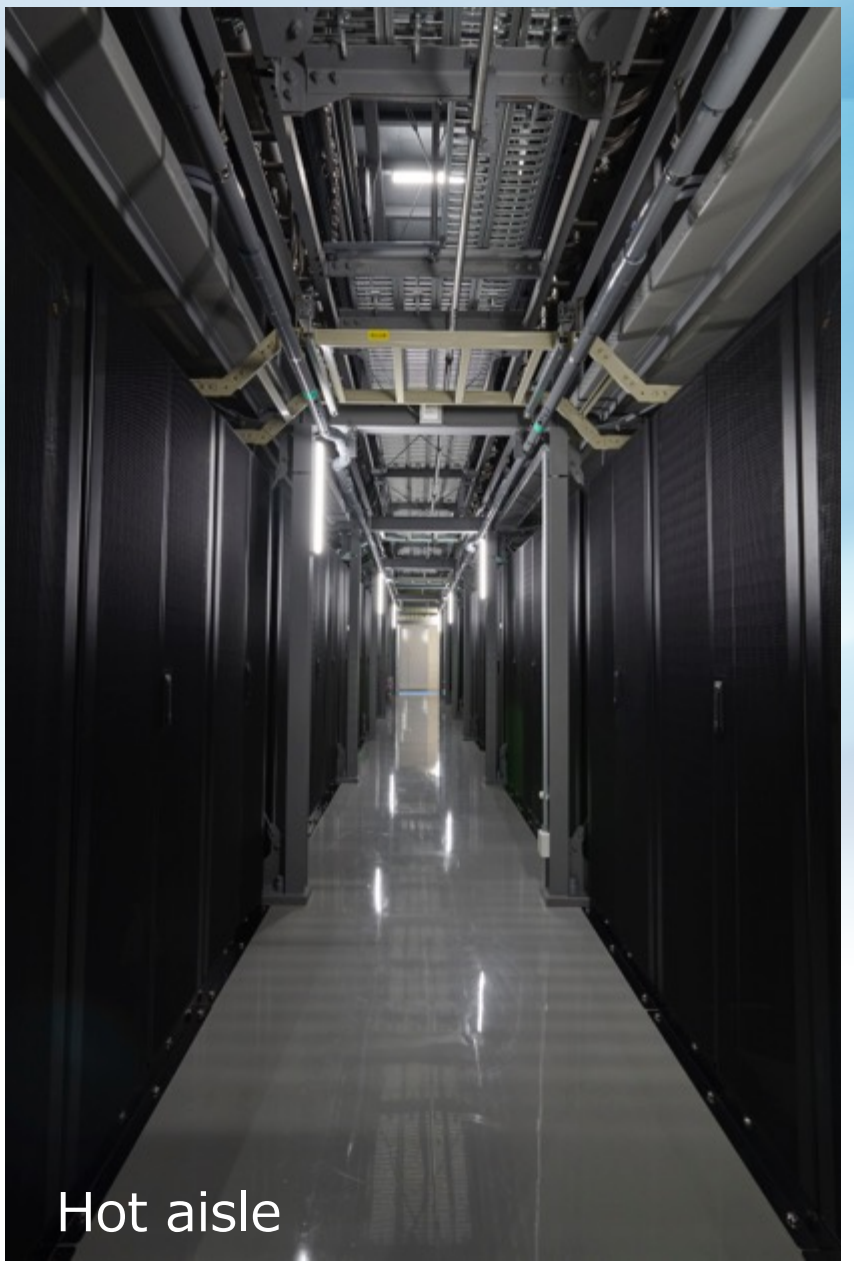
Front view



Side view

Inside Cooling Pod



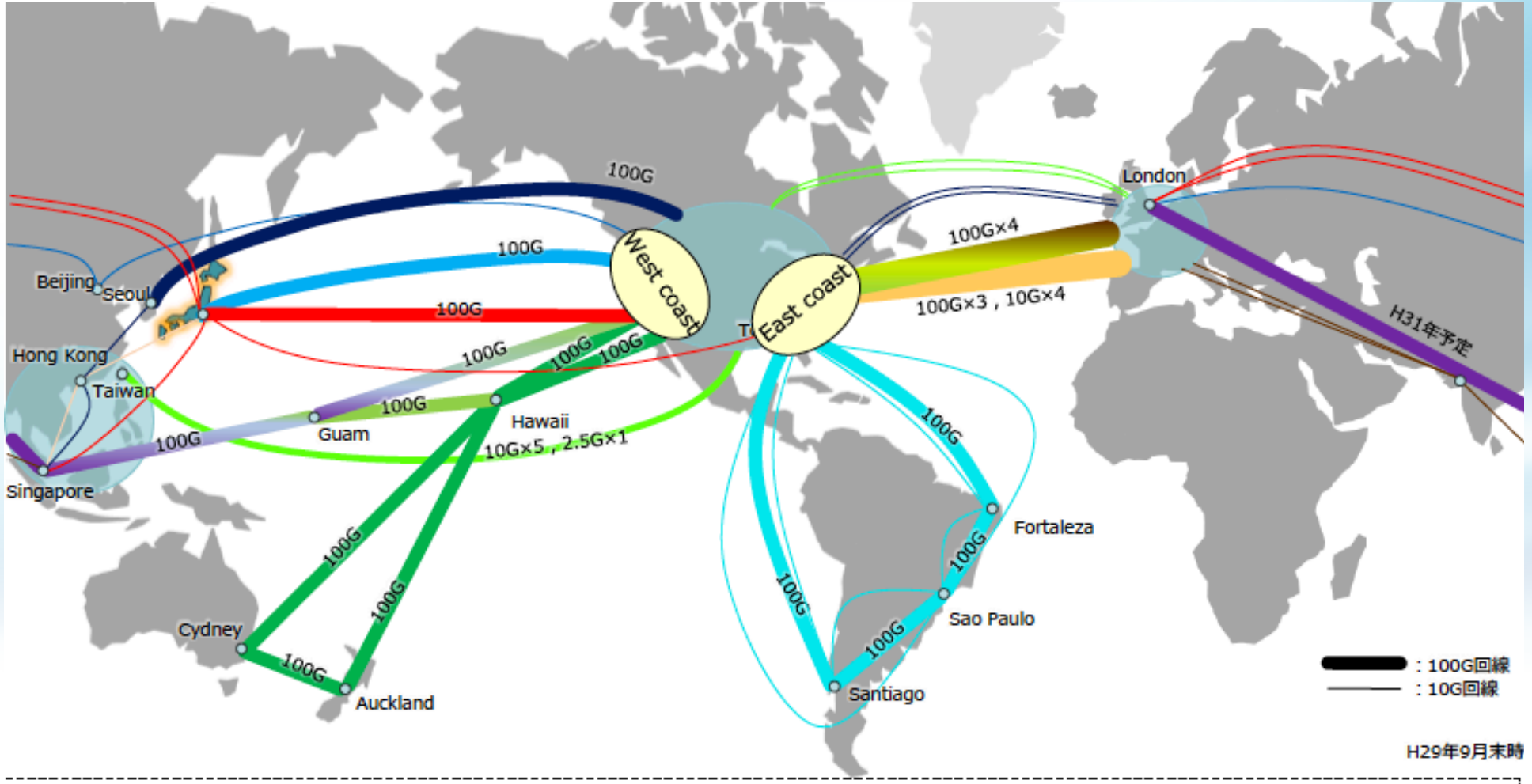


Construction of Cooling Pods

February 6, 2018

March 21, 2018

ABCI with SINET-5 Connectivity



ABCI-Pacific Research Platform (PRP)

- Objective: Establish a close working relationship with respect to the research, education, and application of scientific knowledge in AI and more broadly in data intensive science and robotics

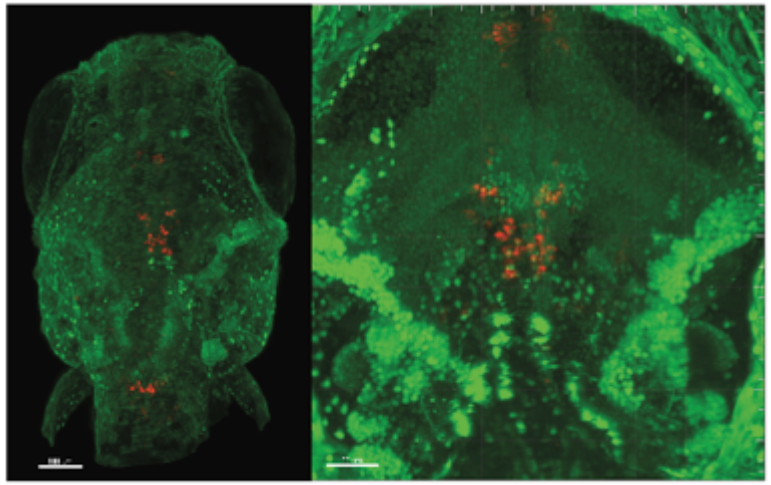
- New 5-year MoU completed between AIST and UCSD
- Activities include:
  - Organization of workshops in the fields of mutual interest both in the U.S. and Japan
  - Exchange of UCSD's faculty and post-doctorates and AIST's researchers
  - UCSD students participate in research projects at AIST in Japan
  - **Collaborative infrastructure projects between UC's PRP and ABCI**
  - **UCSD and AIST researchers use ABCI for collaborative projects**

ABCI-PRP: Grand Challenge Project

- Gerald Pao, Salk Institute; George Sugihara, SIO
- Creation of neuromorphic deep learning architectures by large-scale dynamic modeling of transparent fish brains on ABCI

ABCI is ideally suited to perform the convergent cross mapping (CCM) to interrogate the relationships between the ~120,000 neurons of the larval zebrafish brain during exposure to various stimuli

Check Wassapon's Poster (PP25)!



From Pao, 2018. Green are neurons, red are active neurons responding to hypoxic environment.

ABCI Grand Challenge

- "ABCI Grand Challenge" program encourages researchers and students to leverage ABCI's enormous compute capability (1,088 nodes, 4,352 GPUs) for up to 24 hours to support highly challenging themes in the field of AI.

■ Important Date

	Application Deadline	Notification	Grand Challenge Week
#1	April 30 2018	May 30 2018	Last week in Jul. 2018
#2	August 31 2018	September 30 2018	Last week in Oct. 2018
#3	November 30 2018	December 22 2018	Last week in Jan. 2019

Thank you for your attention!

More Information is available!
http://abci.ai/

ABCI