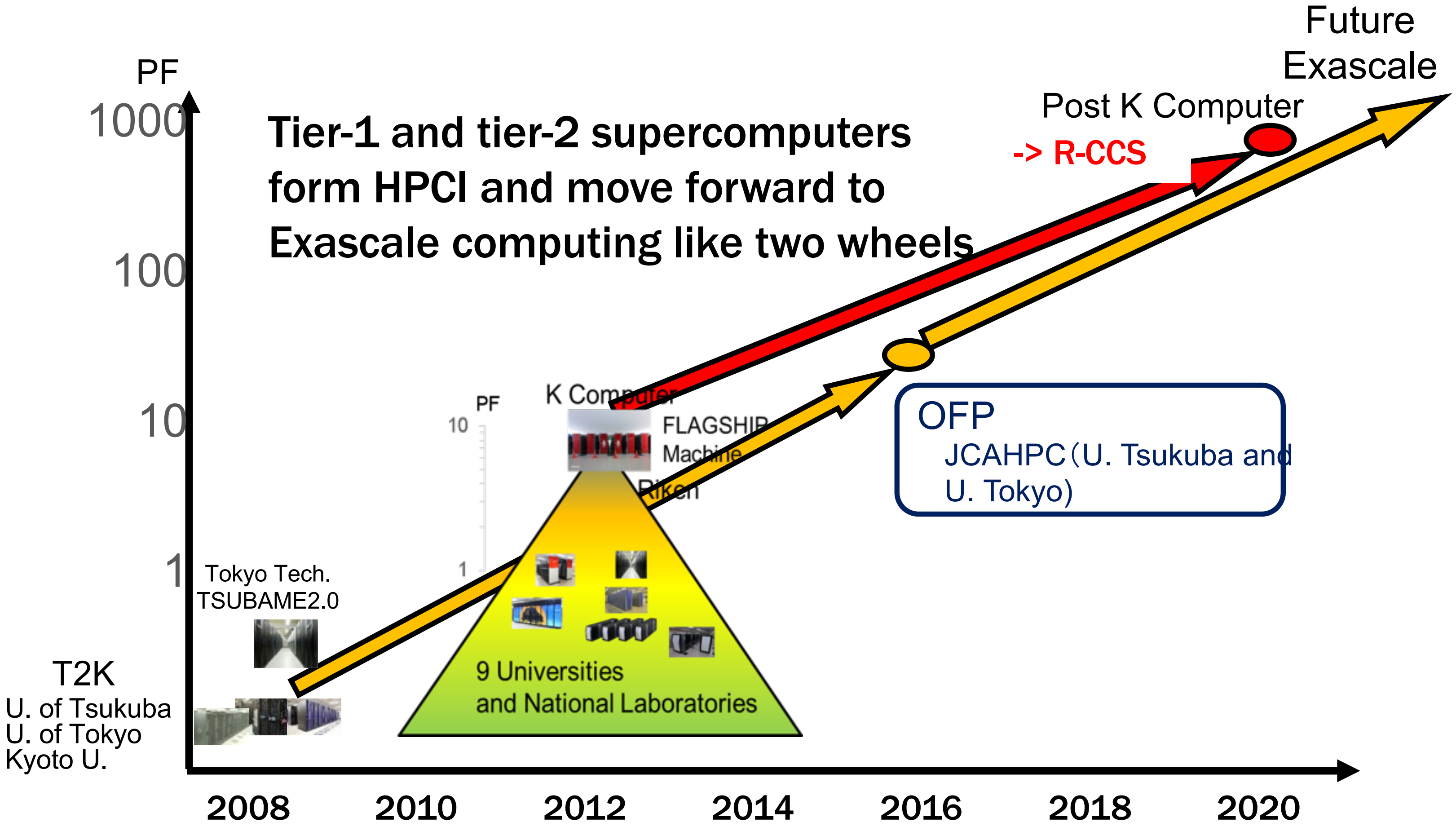
Site Update for Oakforest-PACS at JCAHPC

Taisuke Boku

Vice Director, JCAHPC

University of Tsukuba

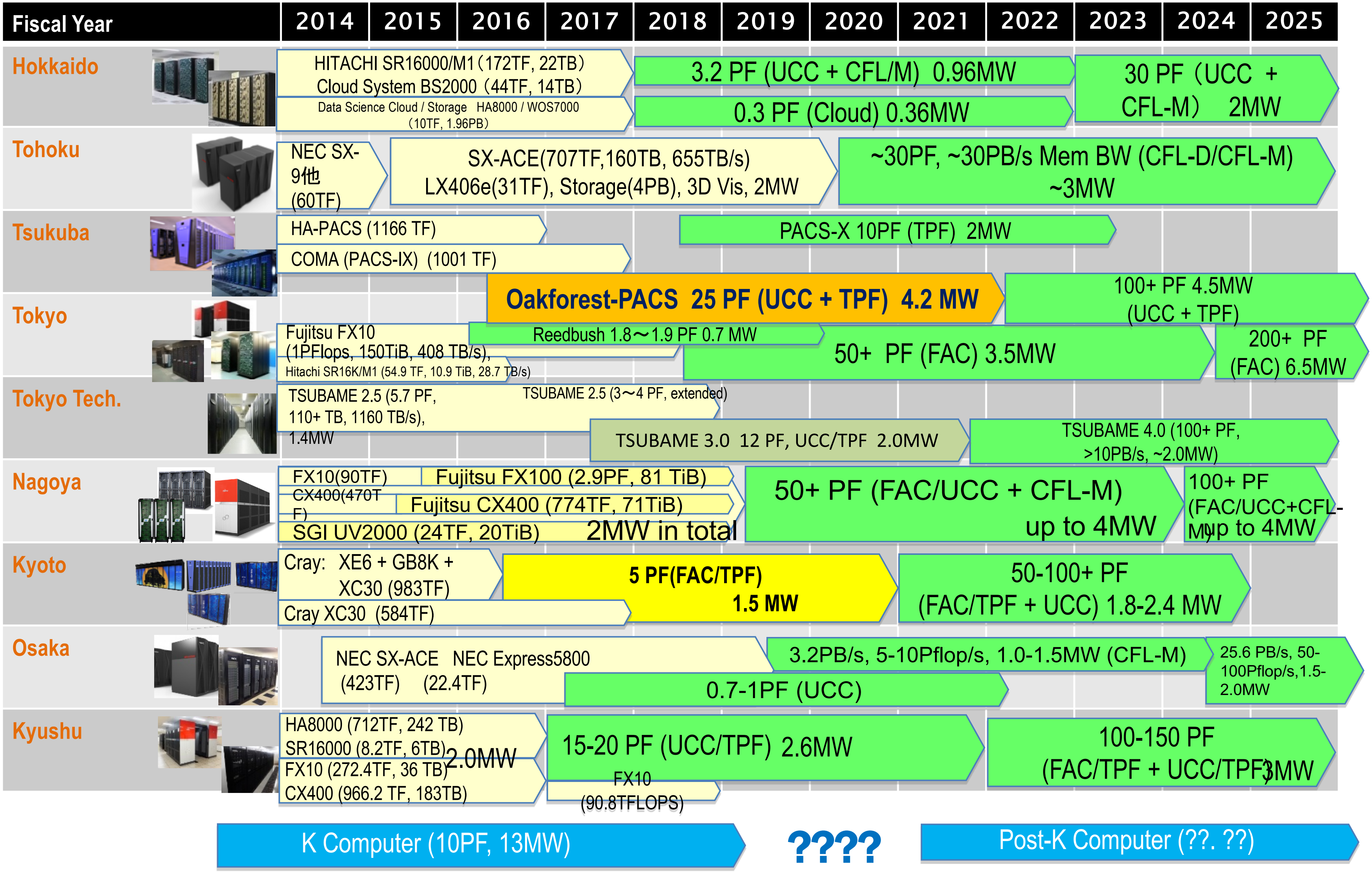
Towards Exascale Computing



Tier-1 and tier-2 supercomputers form HPCI and move forward to Exascale computing like two wheels

Deployment plan of 9 supercomputing center（Feb. 2017）

Power consumption indicates maximum of power supply (includes cooling facility)



Fiscal Year: 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025

Hokkaido

HITACHI SR16000/M1（172TF, 22TB）
Cloud System BS2000（44TF, 14TB）
Data Science Cloud / Storage HA8000 / WOS7000（10TF, 1.96PB）

3.2 PF (UCC + CFL/M) 0.96MW

0.3 PF (Cloud) 0.36MW

30 PF（UCC + CFL-M） 2MW

Tohoku

NEC SX-9他 (60TF)

SX-ACE(707TF,160TB, 655TB/s)
LX406e(31TF), Storage(4PB), 3D Vis, 2MW

~30PF, ~30PB/s Mem BW (CFL-D/CFL-M) ~3MW

Tsukuba

HA-PACS (1166 TF)

COMA (PACS-IX) (1001 TF)

PACS-X 10PF (TPF) 2MW

Tokyo

Oakforest-PACS 25 PF (UCC + TPF) 4.2 MW

100+ PF 4.5MW (UCC + TPF)

Fujitsu FX10 (1PFlops, 150TiB, 408 TB/s),

Reedbush 1.8〜1.9 PF 0.7 MW

Hitachi SR16K/M1 (54.9 TF, 10.9 TiB, 28.7 TB/s)

50+ PF (FAC) 3.5MW

200+ PF (FAC) 6.5MW

Tokyo Tech.

TSUBAME 2.5 (5.7 PF, 110+ TB, 1160 TB/s), 1.4MW

TSUBAME 2.5 (3〜4 PF, extended)

TSUBAME 3.0 12 PF, UCC/TPF 2.0MW

TSUBAME 4.0 (100+ PF, >10PB/s, ~2.0MW)

Nagoya

FX10(90TF)

Fujitsu FX100 (2.9PF, 81 TiB)

CX400(470TF)

Fujitsu CX400 (774TF, 71TiB)

SGI UV2000 (24TF, 20TiB)

2MW in total

50+ PF (FAC/UCC + CFL-M) up to 4MW

100+ PF (FAC/UCC+CFL-M) up to 4MW

Kyoto

Cray: XE6 + GB8K + XC30 (983TF)

Cray XC30 (584TF)

5 PF(FAC/TPF) 1.5 MW

50-100+ PF (FAC/TPF + UCC) 1.8-2.4 MW

Osaka

NEC SX-ACE (423TF)

NEC Express5800 (22.4TF)

3.2PB/s, 5-10Pflop/s, 1.0-1.5MW (CFL-M)

0.7-1PF (UCC)

25.6 PB/s, 50-100Pflop/s,1.5-2.0MW

Kyushu

HA8000 (712TF, 242 TB)
SR16000 (8.2TF, 6TB)
FX10 (272.4TF, 36 TB)
CX400 (966.2 TF, 183TB)
2.0MW

15-20 PF (UCC/TPF) 2.6MW

FX10 (90.8TFLOPS)

100-150 PF (FAC/TPF + UCC/TPF) 3MW

K Computer (10PF, 13MW)

????

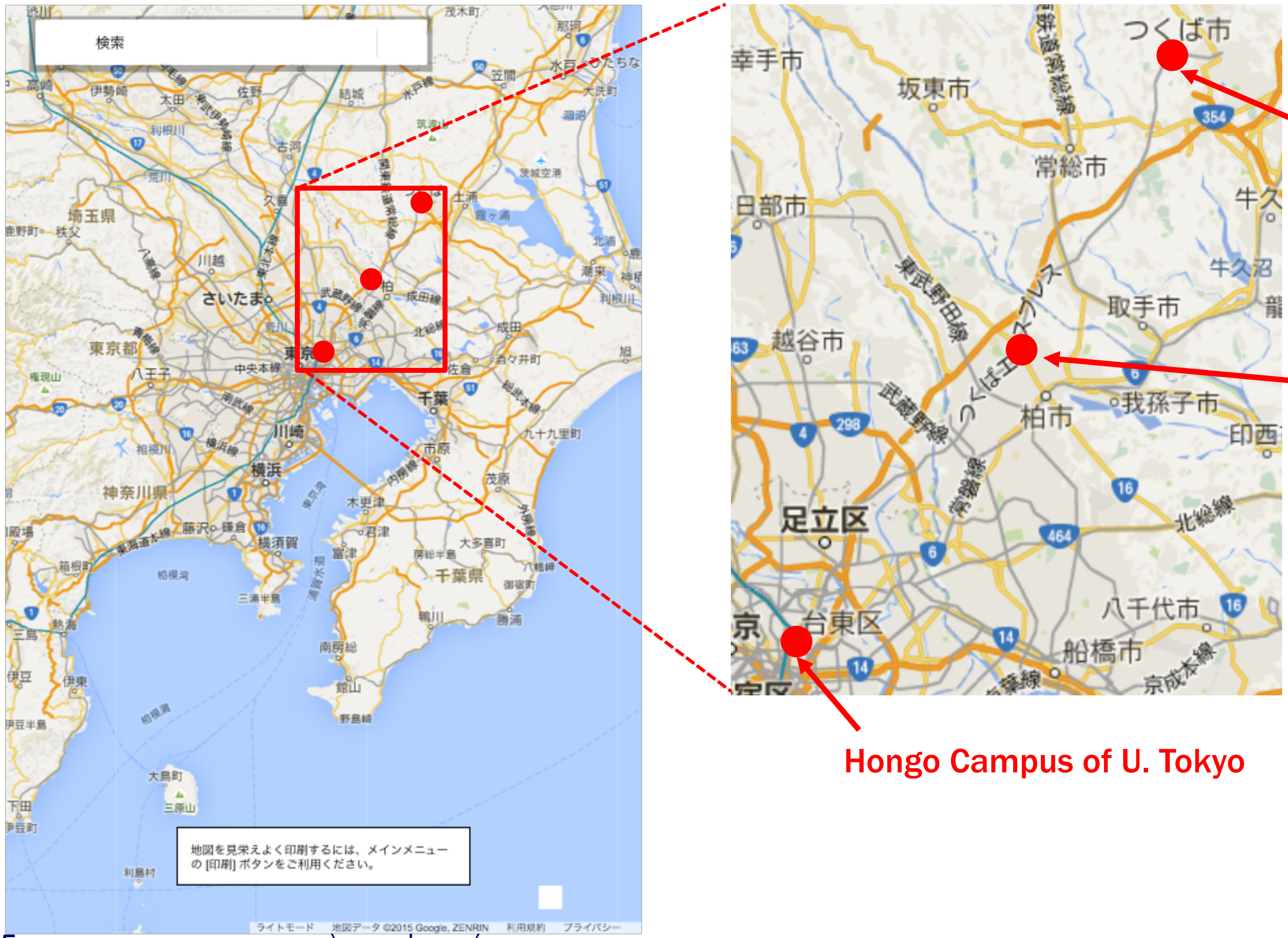
Post-K Computer (??. ??)

JCAHPC

- Joint Center for Advanced High Performance Computing (http://jcahpc.jp)

- Very tight collaboration for “post-T2K” with two universities
  - For main supercomputer resources, *uniform specification* to ***single shared system***
  - Each university is financially responsible to introduce the machine and its operation
    -> unified procurement toward single system with ***largest scale in Japan***
  - To manage everything smoothly, a joint organization was established
    -> JCAHPC

Machine location: Kashiwa Campus of U. Tokyo



地図を見栄えよく印刷するには、メインメニューの [印刷] ボタンをご利用ください。

ライトモード　地図データ ©2015 Google, ZENRIN　利用規約　プライバシー

U. Tsukuba

Kashiwa Campus of U. Tokyo

Hongo Campus of U. Tokyo

Oakforest-PACS (OFP)

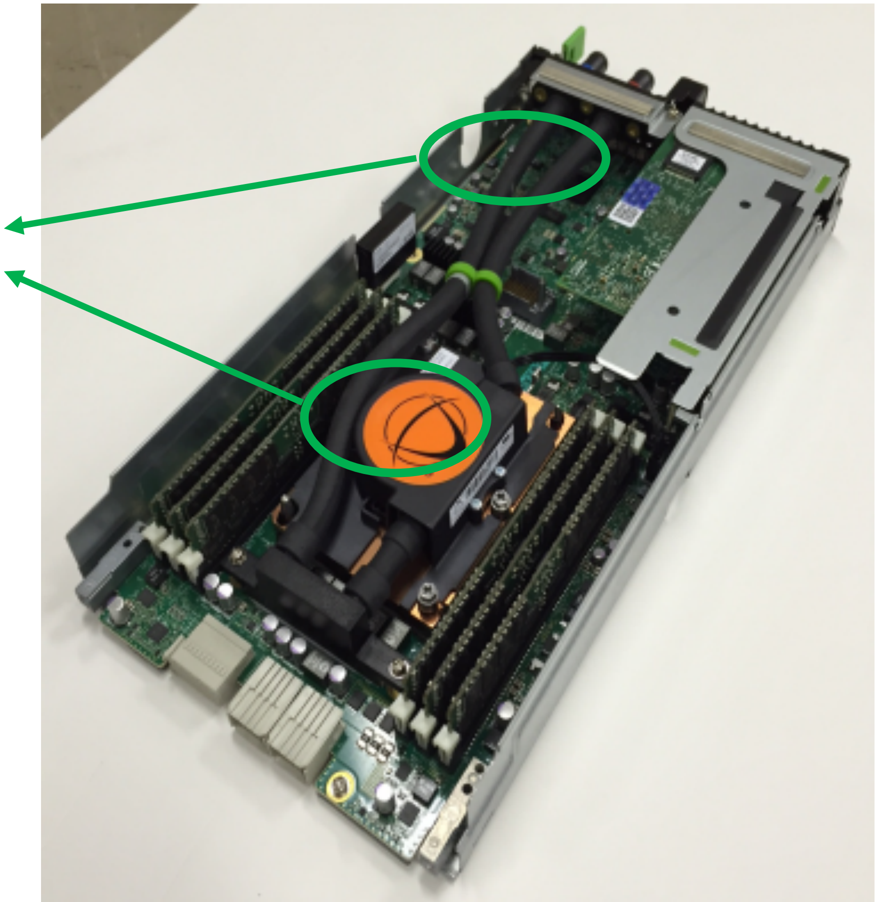
U. Tokyo convention U. Tsukuba convention

⇒ Don’t call it just “Oakforest” !
“OFP” is much better

- 25 PFLOPS peak
- 8208 KNL CPUs
- FBB Fat-Tree by OmniPath
- HPL 13.55 PFLOPS
  #1 in Japan
  #6→#7
- HPCG #3→#5
- Green500 #6→#21

- Full operation started Dec. 2016
- Official Program started on April 2017

Computation node & chassis

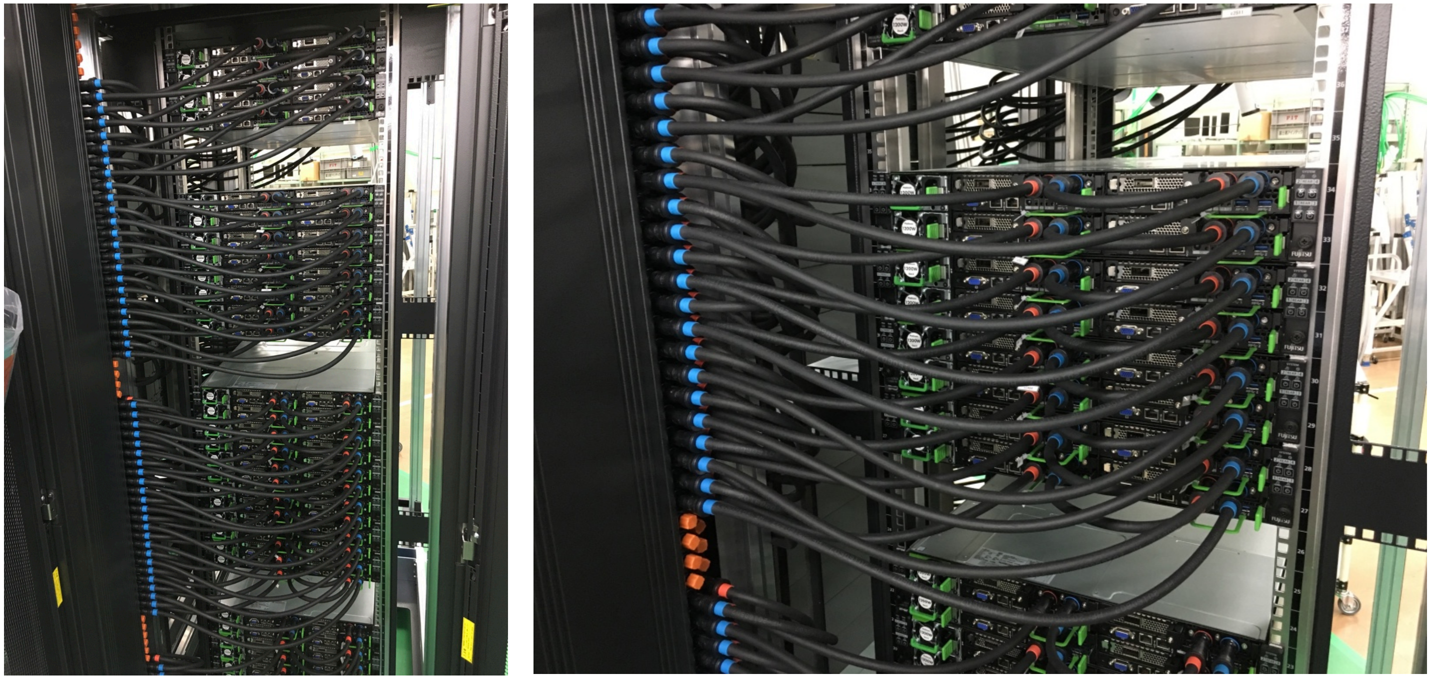


Water cooling wheel & pipe

Computation node (Fujitsu next generation PRIMERGY) with single chip Intel Xeon Phi (Knights Landing, 3+TFLOPS) and Intel Omni-Path Architecture card (100Gbps)

Chassis with 8 nodes, 2U size

Water cooling pipes and rear panel radiator



Direct water cooling pipe for CPU

Rear-panel indirect water cooling for others

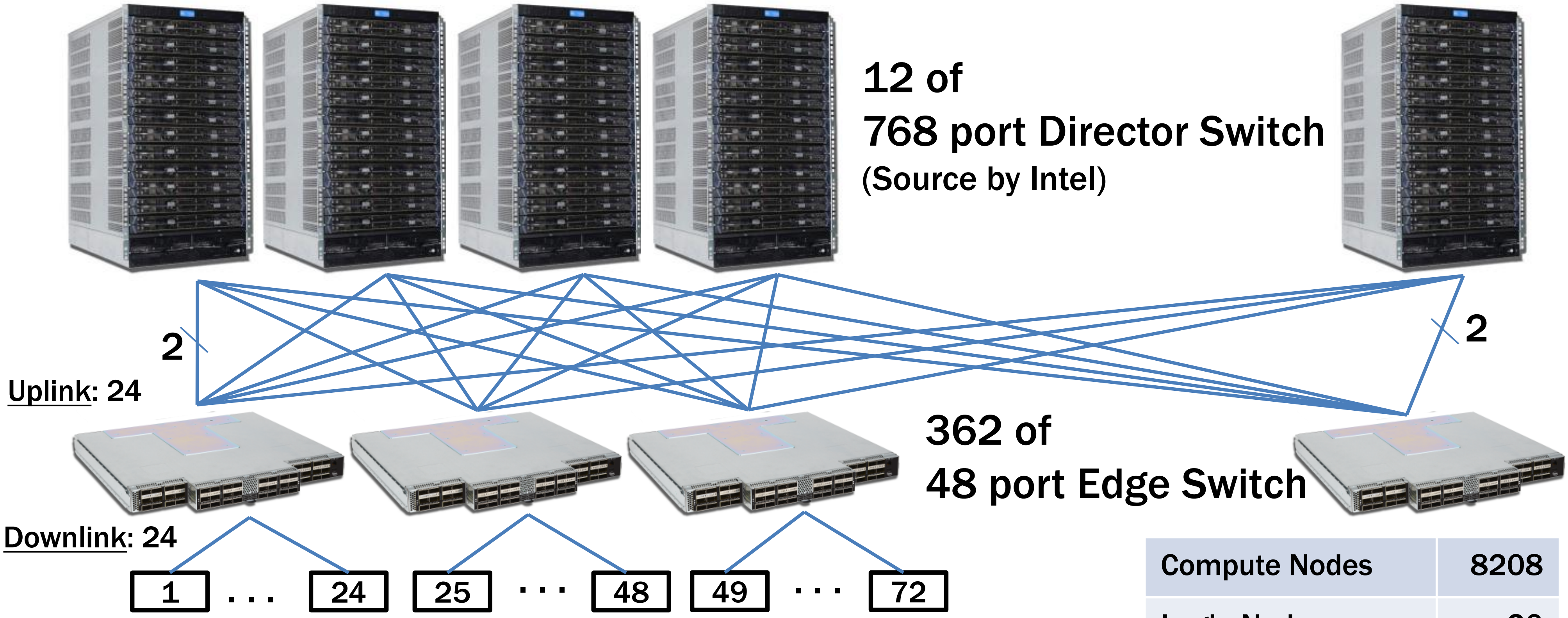
Specification of Oakforest-PACS

Total peak performance			25 PFLOPS
Total number of compute nodes			8,208
Compute node	Product		Fujitsu Next-generation PRIMERGY server for HPC (under development)
	Processor		Intel® Xeon Phi™ （Knights Landing） Xeon Phi 7250 (1.4GHz TDP) with 68 cores
	Memory	High BW	16 GB, > 400 GB/sec (MCDRAM, effective rate)
		Low BW	96 GB, 115.2 GB/sec (DDR4-2400 x 6ch, peak rate)
Inter-connect	Product		Intel® Omni-Path Architecture
	Link speed		100 Gbps
	Topology		Fat-tree with full-bisection bandwidth
Login node	Product		Fujitsu PRIMERGY RX2530 M2 server
	# of servers		20
	Processor		Intel Xeon E5-2690v4 (2.6 GHz 14 core x 2 socket)
	Memory		256 GB, 153 GB/sec (DDR4-2400 x 4ch x 2 socket)

Specification of Oakforest-PACS (I/O)

Parallel File System	Type		Lustre File System
	Total Capacity		26.2 PB
	Meta data	Product	DataDirect Networks MDS server + SFA7700X
		# of MDS	4 servers x 3 set
		MDT	7.7 TB (SAS SSD) x 3 set
	Object storage	Product	DataDirect Networks SFA14KE
		# of OSS (Nodes)	10 (20)
		Aggregate BW	~500 GB/sec
Fast File Cache System	Type		Burst Buffer, Infinite Memory Engine (by DDN)
	Total capacity		940 TB (NVMe SSD, including parity data by erasure coding)
	Product		DataDirect Networks IME14K
	# of servers (Nodes)		25 (50)
	Aggregate BW		~1,560 GB/sec

Full bisection bandwidth Fat-tree by Intel® Omni-Path Architecture



12 of
768 port Director Switch
(Source by Intel)

362 of
48 port Edge Switch

Uplink: 24

Downlink: 24

Firstly, to reduce switches&cables, we considered :

- All the nodes into subgroups are connected with FBB Fat-tree
- Subgroups are connected with each other with >20% of FBB

But, HW quantity is not so different from globally FBB, and globally FBB is preferredfor flexible job management.

Compute Nodes	8208
Login Nodes	20
Parallel FS	64
IME	300
Mgmt, etc.	8
Total	**8600**

Facility of Oakforest-PACS system

Power consumption			4.2 MW (including cooling) → actually around 3.0 MW
# of racks			102
Cooling system	Compute Node	Type	Warm-water cooling Direct cooling (CPU) Rear door cooling (except CPU)
		Facility	Cooling tower & Chiller
	Others	Type	Air cooling
		Facility	PAC

Software of Oakforest-PACS

	Compute node	Login node
OS	CentOS 7, McKernel	Red Hat Enterprise Linux 7
Compiler	gcc, Intel compiler (C, C++, Fortran)	
MPI	Intel MPI, MVAPICH2	
Library	Intel MKL LAPACK, FFTW, SuperLU, PETSc, METIS, Scotch, ScaLAPACK, GNU Scientific Library, NetCDF, Parallel netCDF, Xabclib, ppOpen-HPC, ppOpen-AT, MassiveThreads	
Application	mpijava, XcalableMP, OpenFOAM, ABINIT-MP, PHASE system, FrontFlow/blue, FrontISTR, REVOCAP, OpenMX, xTAPP, AkaiKKR, MODYLAS, ALPS, feram, GROMACS, BLAST, R packages, Bioconductor, BioPerl, BioRuby	
Distributed FS		Globus Toolkit, Gfarm
Job Scheduler	Fujitsu Technical Computing Suite	
Debugger	Allinea DDT	
Profiler	Intel VTune Amplifier, Trace Analyzer & Collector	

TOP500 list on Nov. 2017 (#50)

#	Machine	Architecture	Country	Rmax (TFLOPS)	Rpeak (TFLOPS)	MFLOPS/W
1	TaihuLight, NSCW	MPP (Sunway, SW26010)	China	93,014.6	125,435.9	6051.3
2	Tianhe-2 (MilkyWay-2), NSCG	Cluster (NUDT, CPU + KNC)	China	33,862.7	54,902.4	1901.5
3	Piz Daint, CSCS	MPP (Cray, XC50: CPU + GPU)	Switzerland	19,590.0	25,326.3	10398.0
4	Gyoukou, JAMSTEC	MPP (Exascaler, PEZY-SC2)	Japan	19,125.8	28,192.0	14167.3
5	Titan, ORNL	MPP (Cray, XK7: CPU + GPU)	United States	17,590.0	27,112.5	2142.8
6	Sequoia, LLNL	MPP (IBM, BlueGene/Q)	United States	17,173.2	20,132.7	2176.6
7	Trinity, NNSA/ LABNL/SNL	MPP (Cray, XC40: MIC)	United States	14,137.3	43,902.6	3667.8
8	Cori, NERSC-LBNL	MPP (Cray, XC40: KNL)	United States	14,014.7	27,880.7	3556.7
9	Oakforest-PACS, JCAHPC	Cluster (Fujitsu, KNL)	Japan	13,554.6	25,004.9	4985.1
10	K Computer, RIKEN AICS	MPP (Fujitsu)	Japan	10,510.0	11,280.4	830.2

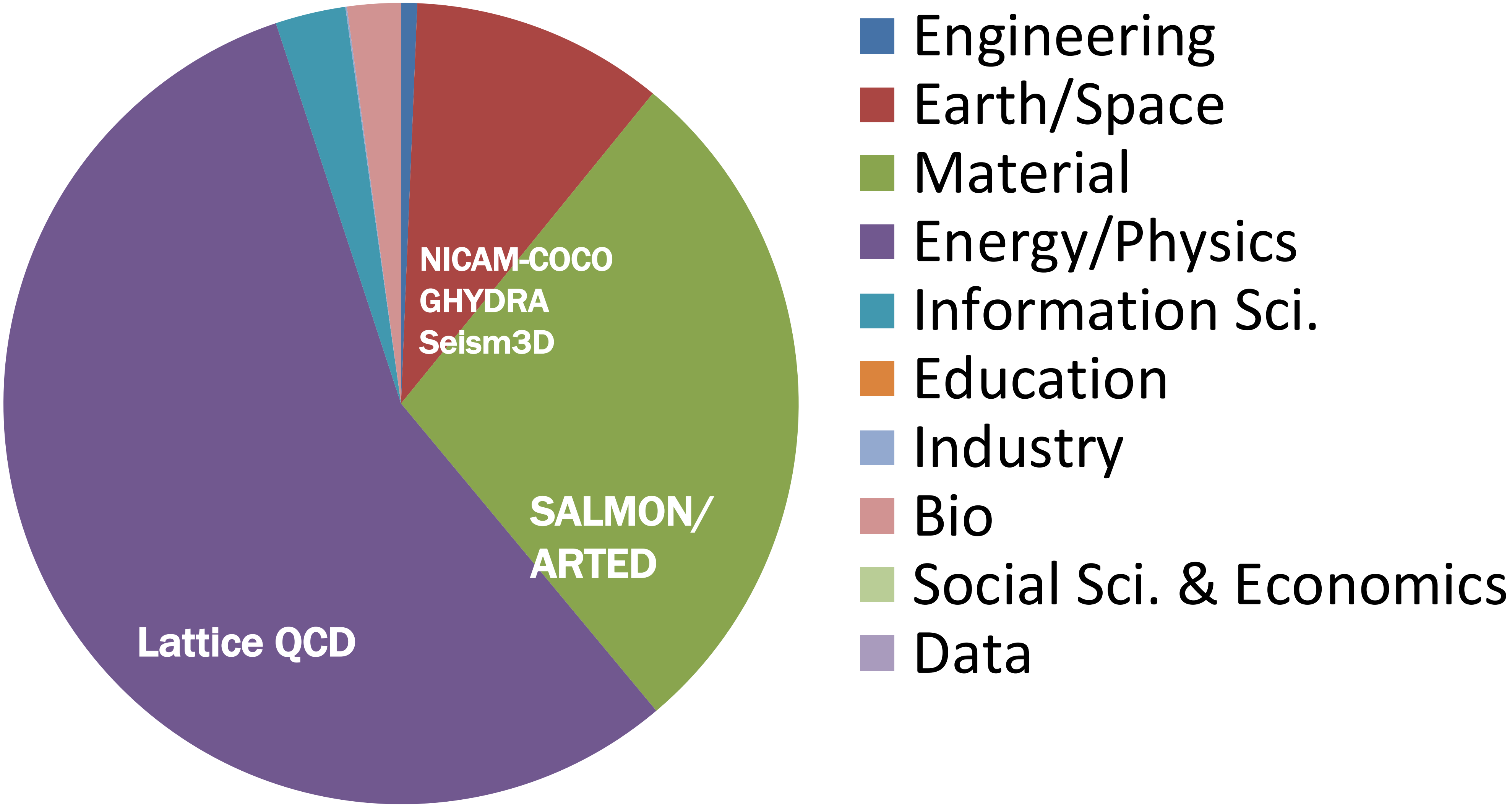
Post-K Computer and OFP

- OFP fills gap between K Computer and Post-K Computer
  - Post-K Computer is planned to install 2020-2021 time frame
  - K Computer will be shutdown around 2018-2019 ??
- Two system software developed in AICS RIKEN for Post-K Computer
  - McKernel
    - OS for Many-core era, for a number of thin-cores without OS jitter and core binding
    - Primary OS (based on Linux) on Post-K, and application development goes ahead
  - XcalableMP (XMP) (in collaboration with U. Tsukuba)
    - Parallel programming language for directive-base easy coding on distributed memory system
    - Not like explicit message passing with MPI

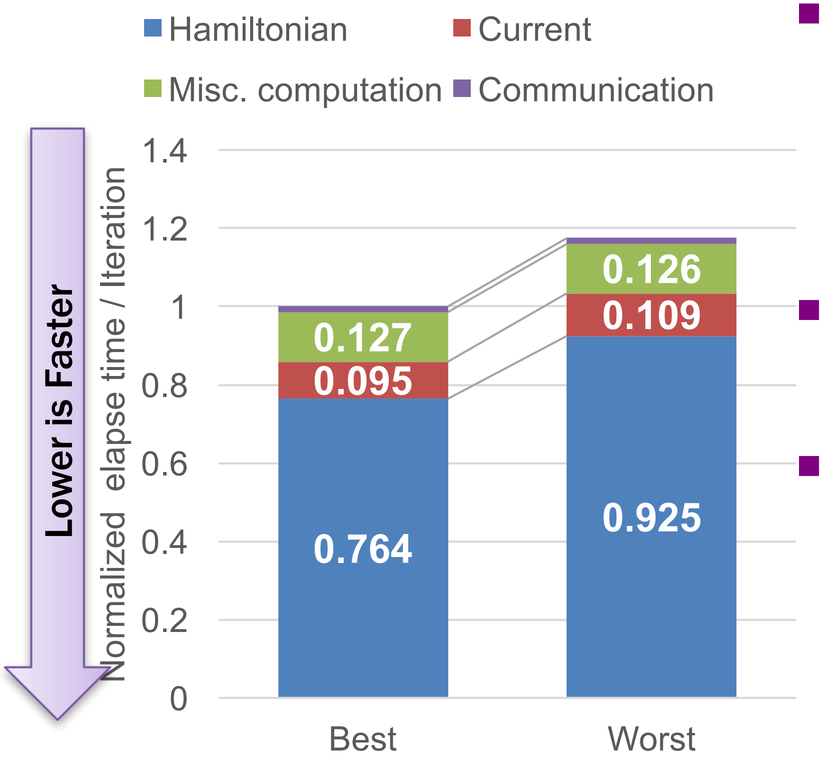
OFP resource sharing program (nation-wide)

- JCAHPC (20%)
  - HPCI – HPC Infrastructure program in Japan to share all supercomputers (free!)
  - Big challenge special use (full system size) – opportunity to use entire 8208 CPUs by just one project for 24 hours, every end of month
- U. Tsukuba (23.5%)
  - Interdisciplinary Academic Program (free!)
  - Large scale general use
- U. Tokyo (56.5%)
  - General use
  - Industrial trial use
  - Educational use
  - Young & Female special use
- Ordinary job can use up to 2048 nodes/job

Research Area based on CPU Hours Oakforest-PACS in FY.2017 (TENTATIVE: 2017.4~2017.9)



Performance variant between nodes



normalized to best case

- most of time is consumed for Hamiltonian calculation
  - not including communication time
  - domain size is equal for all nodes
- root cause of strong scaling saturation
  - performance gap exists on any materials
- <u>Non-algorithmic load-imbalancing</u>
  - dynamic clock adjustment (DVFS) on turbo boost is applied individually on all processors
  - it is observed on under same condition of nodes
  - on KNL, more sensitive than Xeon
  - serious performance degradation on synchronized large scale system

Operation summary

- Memory model
  - basically 50:50 for cache:flat modes
  - started to watch the queue condition for “gently” changing the ratio ~ ±15%
  - planning to introduce “dynamic on-demand switching” in job by job manner
- KNL CPU
  - almost good and failure rate is enough under estimation by Fujitsu
  - enough stability to support up to 2048 node job
- OPA network
  - at first there was a problem at booting up time, but now it’s fixed almost
    -> it was the main reason against to the dynamic memory mode change
  - hundreds of links have been changed by initial failure, but now stable
- Special operation
  - every month, 24hours operation for just one project to occupy entire system

New machine planned at CCS, U. Tsukuba “PACS-X” with GPU+FPGA

CCS at University of Tsukuba

- Center for Computational Sciences
- Established in 1992
  - 12 years as Center for Computational Physics
  - Reorganized as Center for Computational Sciences in 2004
- Daily collaborative researches with two kinds of faculty researchers (about 35 in total)
  - Computational Scientists
    who have NEEDS (applications)
  - Computer Scientists
    who have SEEDS (system & solution)

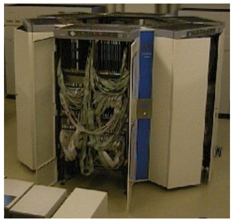
PAX (PACS) series history in U. Tsukuba

- Started in 1977 (by Hoshino and Kawai)
- 1st generation PACS in 1978 with 9 CPUs
- 6th generation CP-PACS awarded #1 in TOP500

1978
1st PACS-9

1980
2nd PAX-32

1989
5th QCDPAX



1996
6th CP-PACS
#1 in the world

2006
PACS-CS (7th)
first PC cluster solution

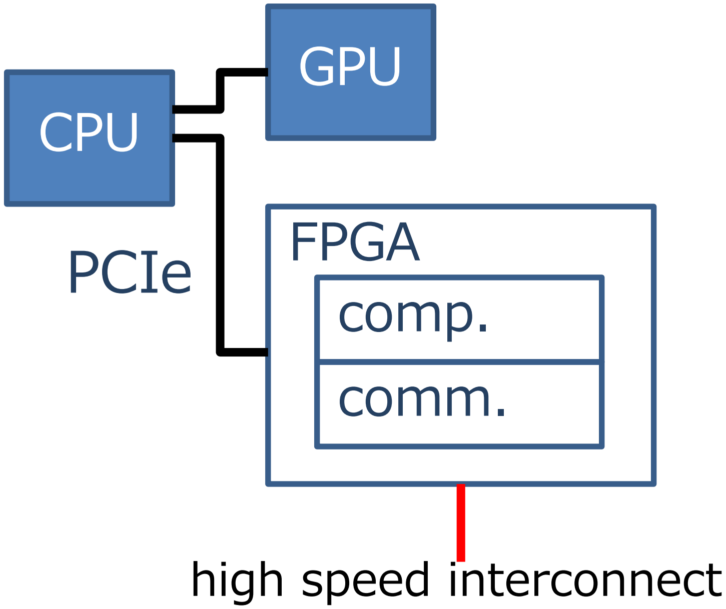
2012~2013
HA-PACS (8th) introducing
GPU/FPGA

Year	Name	Performance
1978年	PACS-9	7 KFLOPS
1980年	PACS-32	500 KFLOPS
1983年	PAX-128	4 MFLOPS
1984年	PAX-32J	3 MFLOPS
1989年	QCDPAX	14 GFLOPS
1996年	CP-PACS	614 GFLOPS
2006年	PACS-CS	14.3 TFLOPS
2012~13年	HA-PACS (PACS-VIII)	1.166 PFLOPS
2014年	COMA (PACS-IX)	1.001 PFLOPS

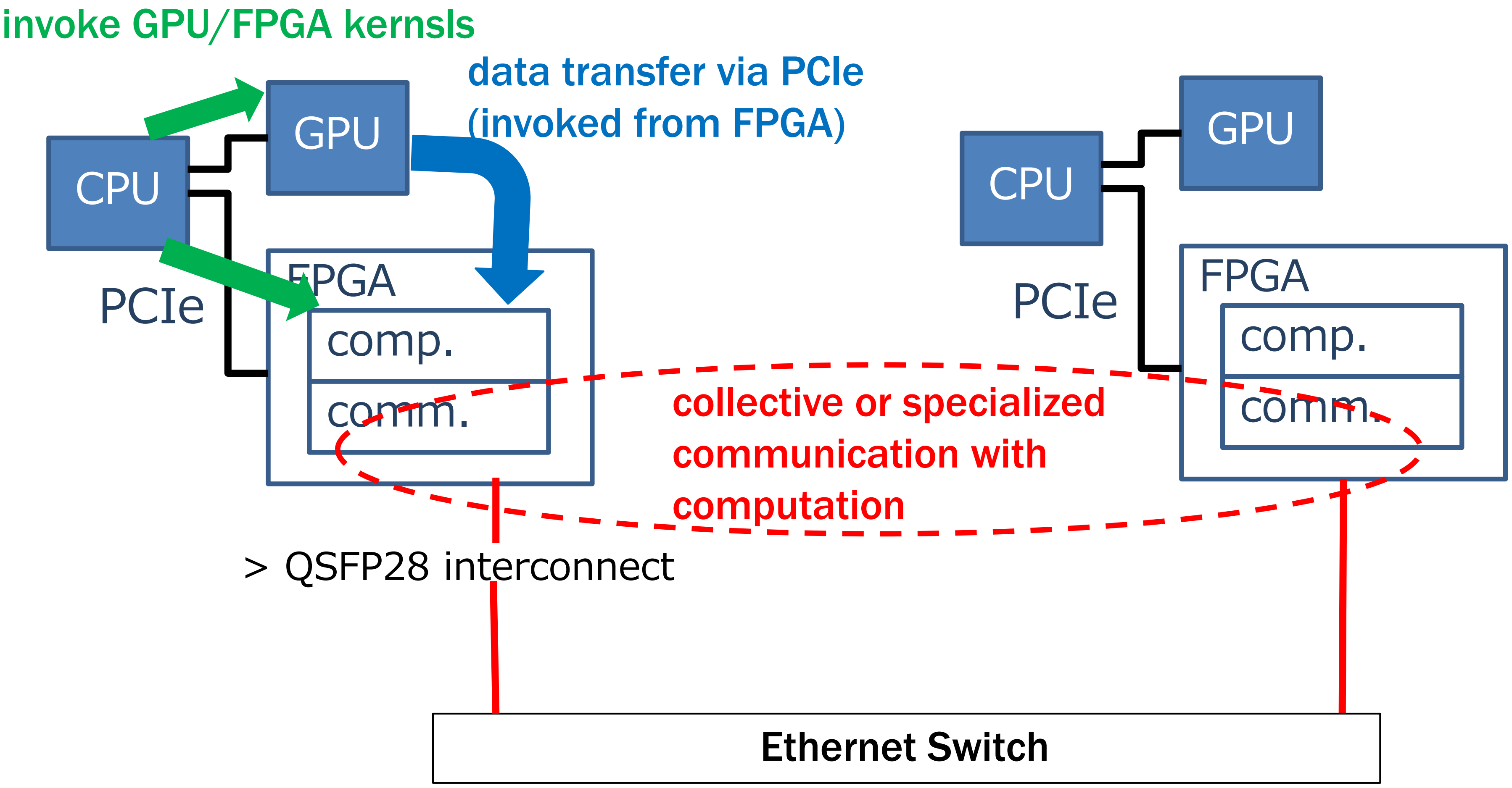
- *co-design* by computational scientists and computer scientists
- *Application-driven* development
- *Accumulation of experiences* by continuous development

AiS

- AiS: Accelerator in Swtich
  - Using FPGA not only for computation offloading but also for communication
  - Combining computation offloading and communication among FPGAs for ultra-low latency on FPGA computing
  - Especially effective on communication-related small/medium computation (such as collective communication)
  - Covering GPU non-suited computation by FPGA
  - OpenCL-enable programming for application users



AiS computation model



invoke GPU/FPGA kernsls

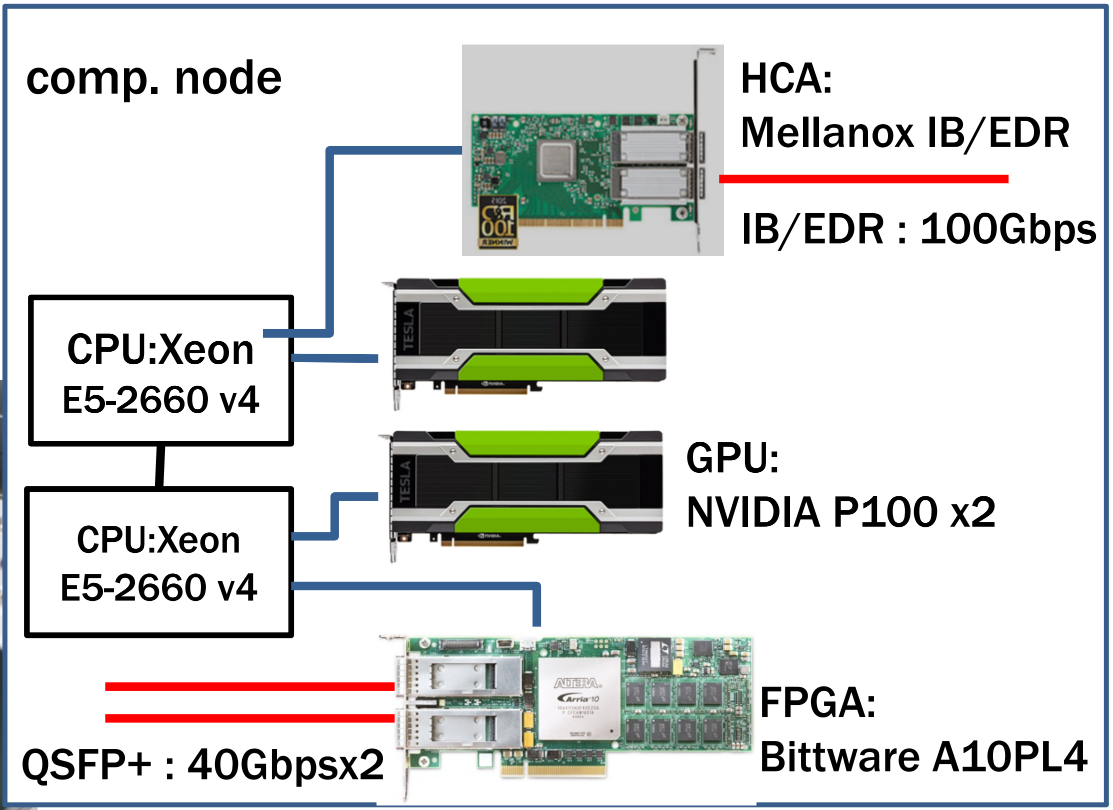
data transfer via PCIe
(invoked from FPGA)

collective or specialized communication with computation

> QSFP28 interconnect

Evaluation test-bed

- Pre-PACS-X (PPX)
  - CCS, U. Tsukuba
  - PACS-X prototype



comp. node

HCA:
Mellanox IB/EDR

IB/EDR : 100Gbps

CPU:Xeon
E5-2660 v4

CPU:Xeon
E5-2660 v4

GPU:
NVIDIA P100 x2

FPGA:
Bittware A10PL4

QSFP+ : 40Gbpsx2

Host OS	CentOS 7.3
Host Compiler	gcc 4.8.5
FPGA Compiler	Intel FPGA SDK for OpenCL, Intel Quartus Prime Pro Version 17.0.0 Build 289

Time Line

- Feb. 2018: Request for Information
- Apr. 2018: Request for Comment (followings are just requirement)
  - basic specification: AiS-based large cluster with up to 256 nodes
  - V100 class of GPU x2
  - Stratix10 or UltraScale class of FPGA x1 (25% of total count of nodes)
  - OPA x2 or InfiniBand HDR class interconnection
- Aug. 2018: Request for Proposal
  - Bidding closed on begin of Sep. 2018
- Mar. 2019: Deployment
- Apr. 2019: Starting official operation