Oakforest-PACS (OFP)

Taisuke Boku

Deputy Director, Center for Computational Sciences

University of Tsukuba

(with courtesy of JCAHPC members)

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

Oakforest-PACS (OFP)

U. Tokyo convention　U. Tsukuba convention

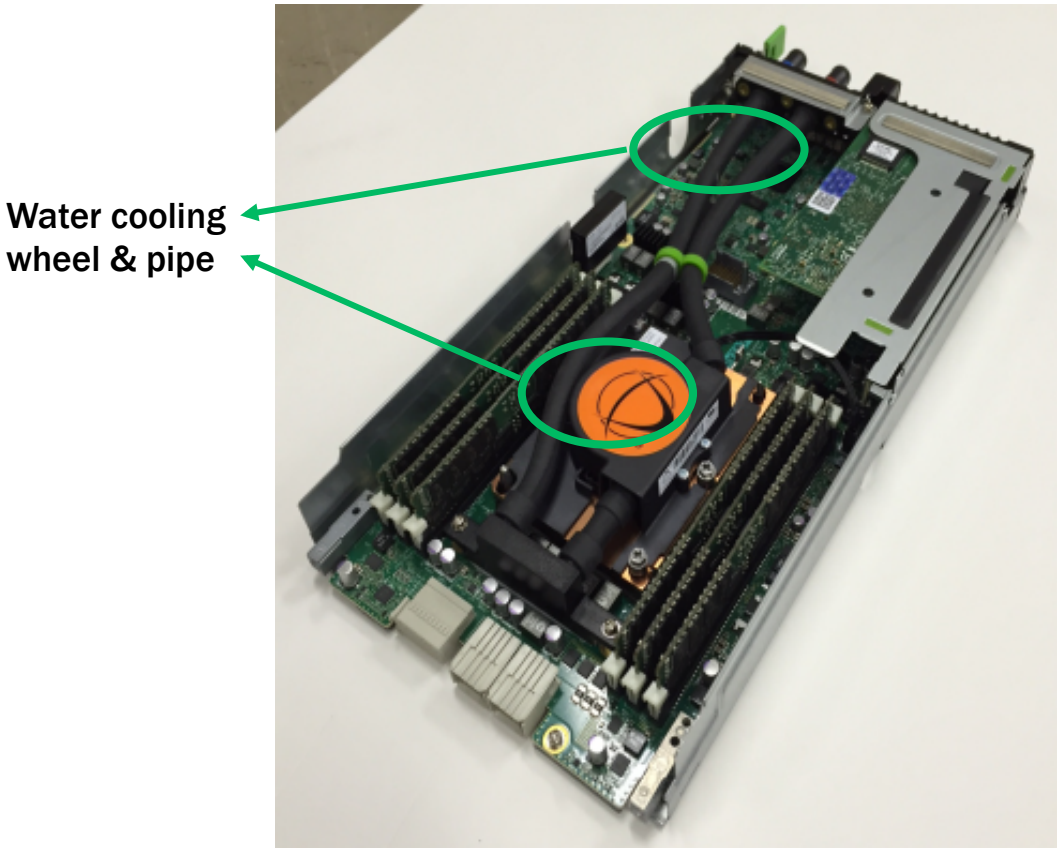
⇒ Don't call it just "Oakforest" !
"OFP" is much better

- 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

JCAHPC



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

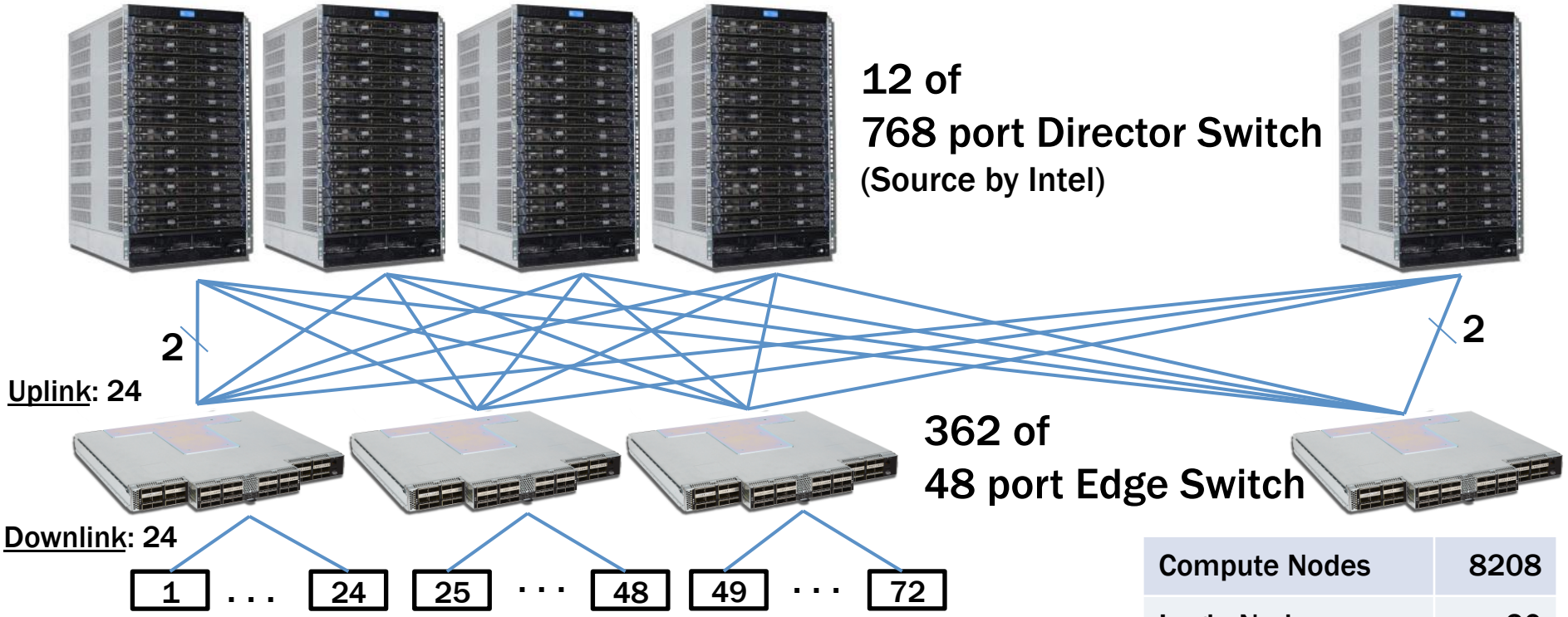
Specification of Oakforest-PACS

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



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

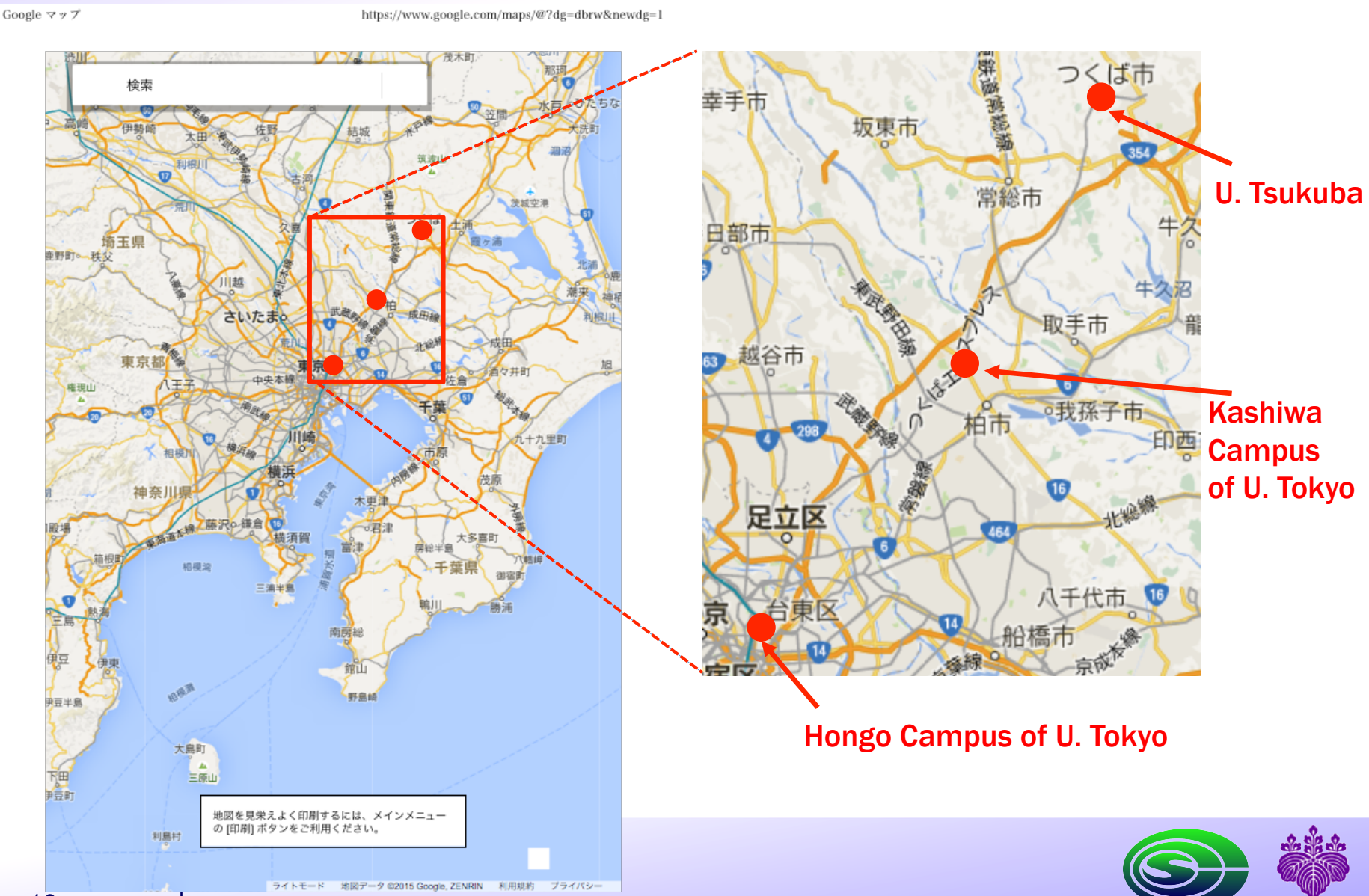
JCAHPC

	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	

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

Machine location: Kashiwa Campus of U. Tokyo



Large Scal Applications on Oakforest-PACS

- **ARTED (SALMON)**
  - Electron Dynamics
- **Lattice QCD**
  - Quantum Chrono Dynamics
- **NICAM & COCO**
  - Atmosphere & Ocean Coupling
- **GHYDRA**
  - Earthquake Simulations
- **Seism3D**
  - Seismic Wave Propagation