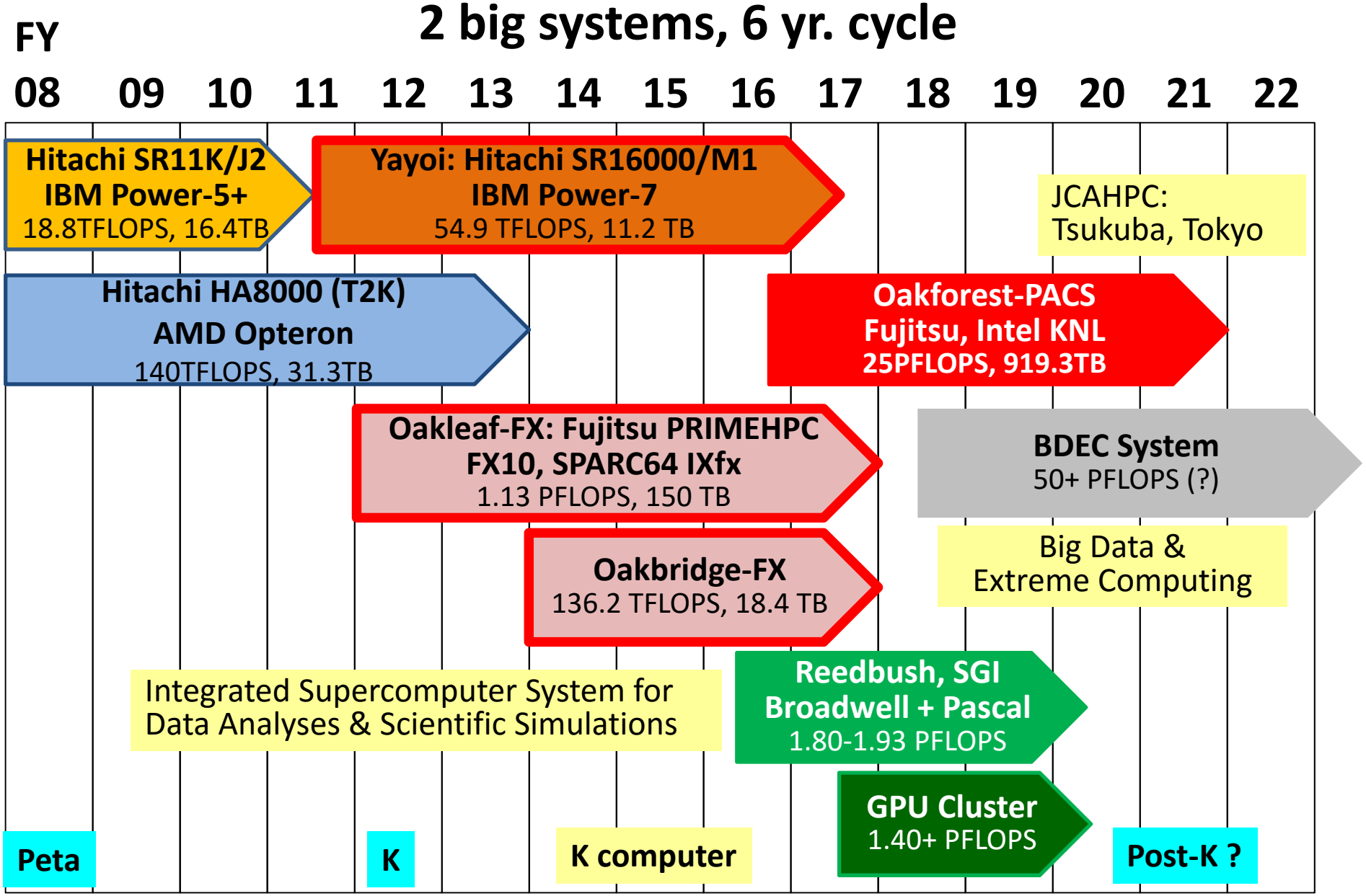
Supercomputers in ITC/U.Tokyo

2 big systems, 6 yr. cycle

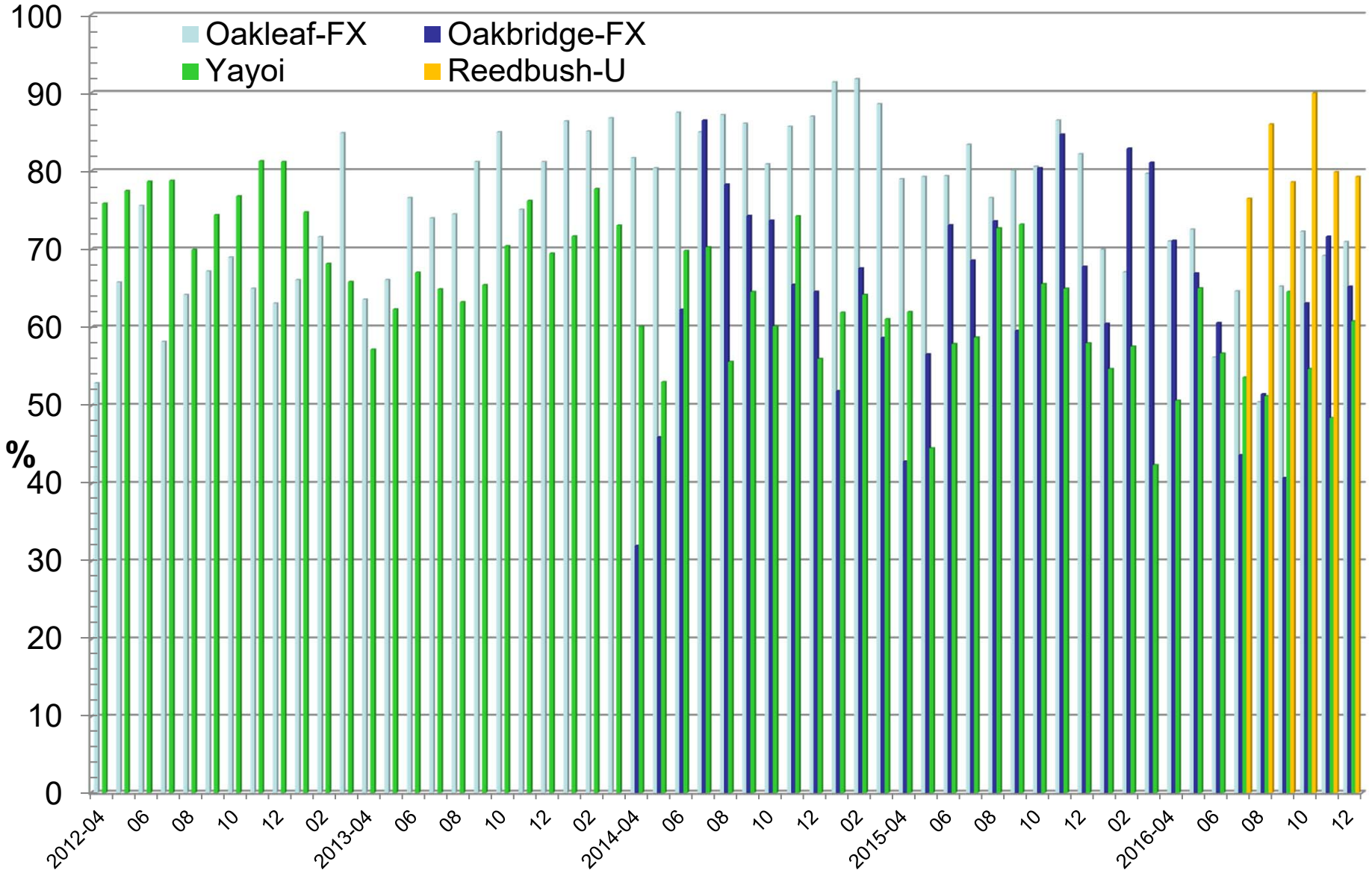


We are now operating 5 systems !!

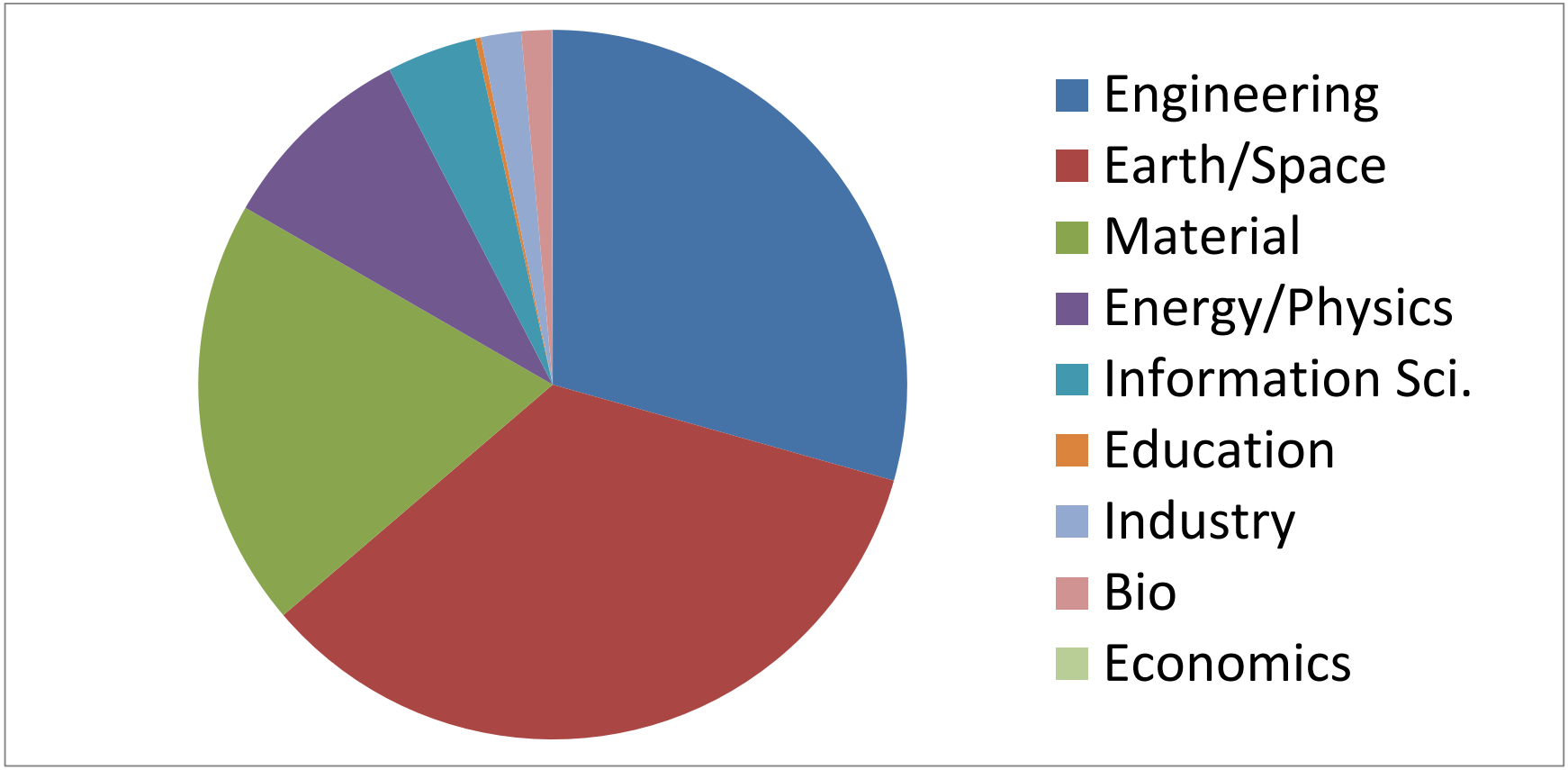
- Yayoi (Hitachi SR16000, IBM Power7)
  - 54.9 TF, Nov. 2011 – Oct. 2017
- Oakleaf-FX (Fujitsu PRIMEHPC FX10)
  - 1.135 PF, Commercial Version of K, Apr.2012 – Mar.2018
- Oakbridge-FX (Fujitsu PRIMEHPC FX10)
  - 136.2 TF, for long-time use (up to 168 hr), Apr.2014 – Mar.2018
- Reedbush (SGI, Intel BDW + NVIDIA P100 (Pascal))
  - Integrated Supercomputer System for Data Analyses & Scientific Simulations
  - 1.93 PF, Jul.2016-Jun.2020
  - Our first GPU System (Mar.2017), DDN IME (Burst Buffer)
- Oakforest-PACS (OFP) (Fujitsu, Intel Xeon Phi (KNL))
  - JCAHPC (U.Tsukuba & U.Tokyo)
  - 25 PF, #6 in 48th TOP 500 (Nov.2016) (#1 in Japan)
  - Omni-Path Architecture, DDN IME (Burst Buffer)

Work Ratio

80+% Average
Oakleaf-FX + Oakbridge-FX



Research Area based on CPU Hours FX10 in FY.2015 (2015.4~2016.3E)



Oakleaf-FX + Oakbridge-FX

- Yayoi (Hitachi SR16000, IBM Power7)
  - 54.9 TF, Nov. 2011 – Oct. 2017

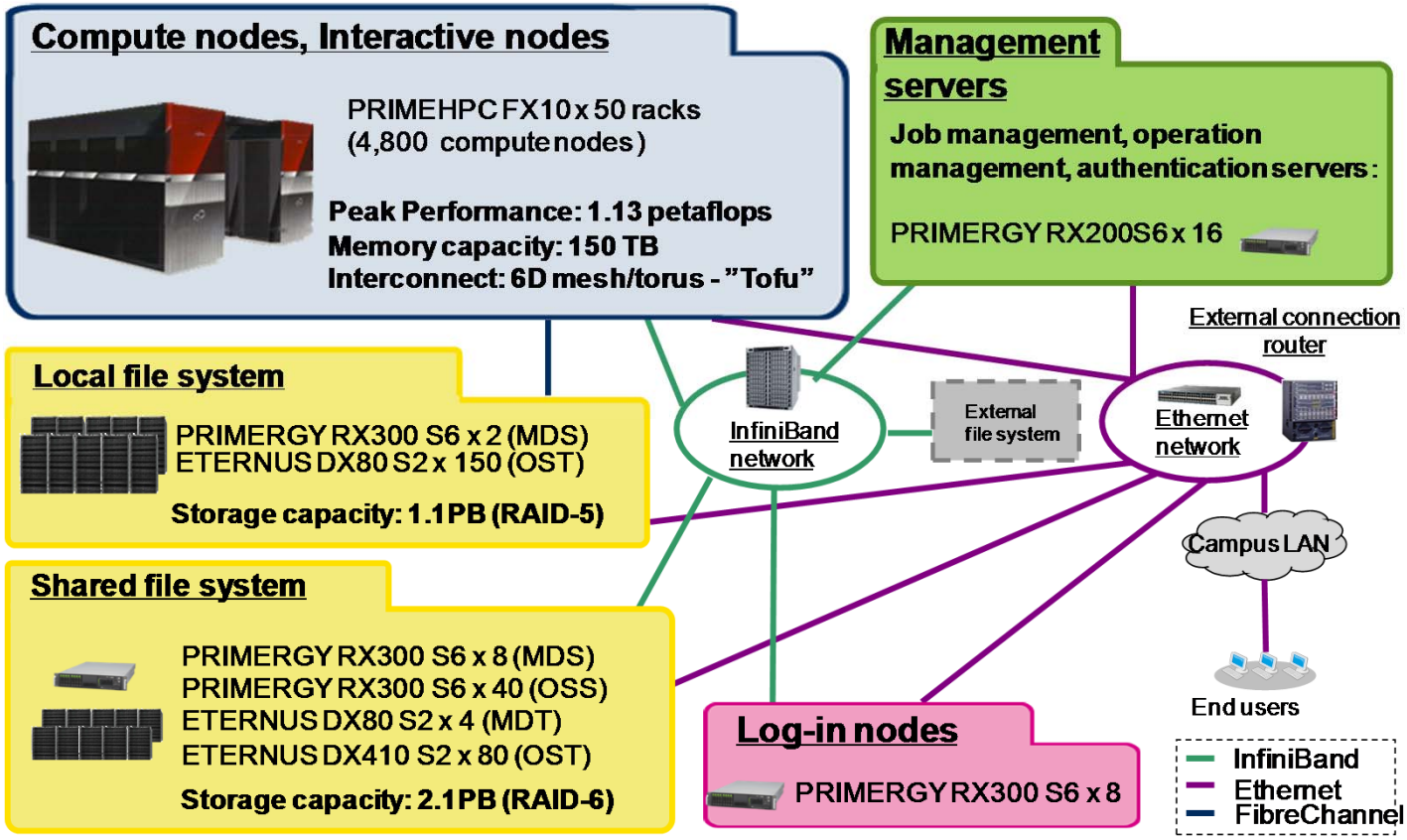
- **Oakleaf-FX (Fujitsu PRIMEHPC FX10)**
  - **1.135 PF, Commercial Version of K, Apr.2012 – Mar.2018**
- Oakbridge-FX (Fujitsu PRIMEHPC FX10)
  - 136.2 TF, for long-time use (up to 168 hr), Apr.2014 – Mar.2018
- Reedbush (SGI, Intel BDW + NVIDIA P100 (Pascal))
  - Integrated Supercomputer System for Data Analyses & Scientific Simulations
  - 1.93 PF, Jul.2016-Jun.2020
  - Our first GPU System (Mar.2017), DDN IME (Burst Buffer)

- Oakforest-PACS (OFP) (Fujitsu, Intel Xeon Phi (KNL))
  - JCAHPC (U.Tsukuba & U.Tokyo)
  - 25 PF, #6 in 48th TOP 500 (Nov.2016) (#1 in Japan)
  - Omni-Path Architecture, DDN IME (Burst Buffer)

Features of FX10 (Oakleaf-FX)

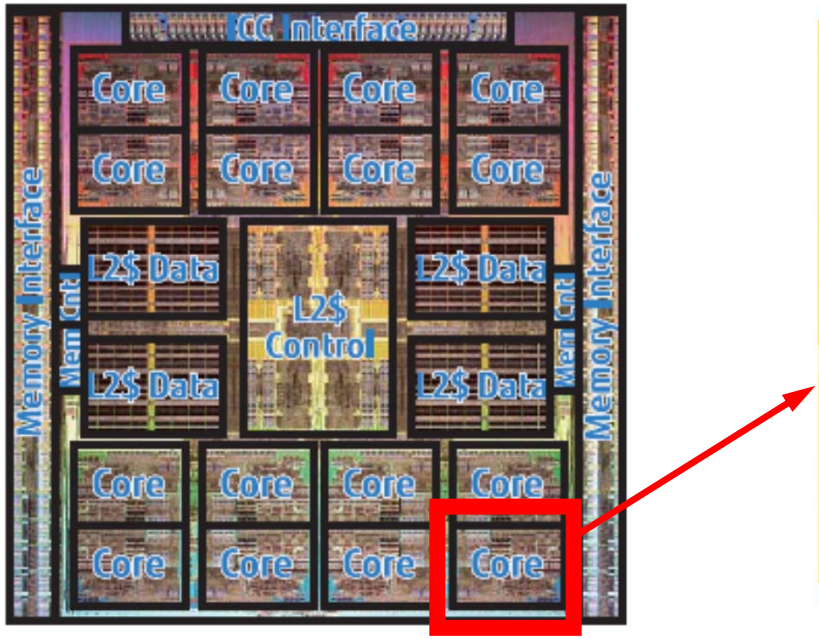
- Well-Balanced System
  - 1.13 PFLOPS for Peak Performance
  - Max. Power Consumption < 1.40 MW
    - < 2.00MW including A/C
- 6-Dim. Mesh/Torus Interconnect
  - Highly Scalable Tofu Interconnect
  - 5.0x2 GB/sec/link, 6 TB/sec for Bi-Section Bandwidth
- High-Performance File System
  - FEFS (Fujitsu Exabyte File System) based on Lustre
- Flexible Switching between Full/Partial Operation
- K compatible !
- Open-Source Libraries/Applications
- Highly Scalable for both of Flat MPI and Hybrid

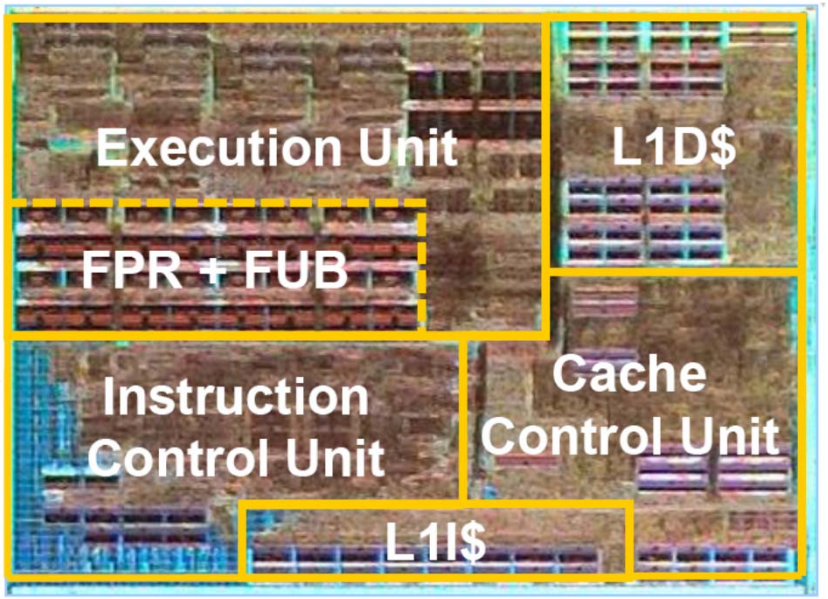
FX10 System (Oakleaf-FX)



- Aggregate memory bandwidth: 398 TB/sec.
- Local file system for staging with 1.1 PB of capacity and 131 GB/sec of aggregate I/O performance (for staging)
- Shared file system for storing data with 2.1 PB and 136 GB/sec.
- External file system: 3.6 PB

SPARC64™ IXfx

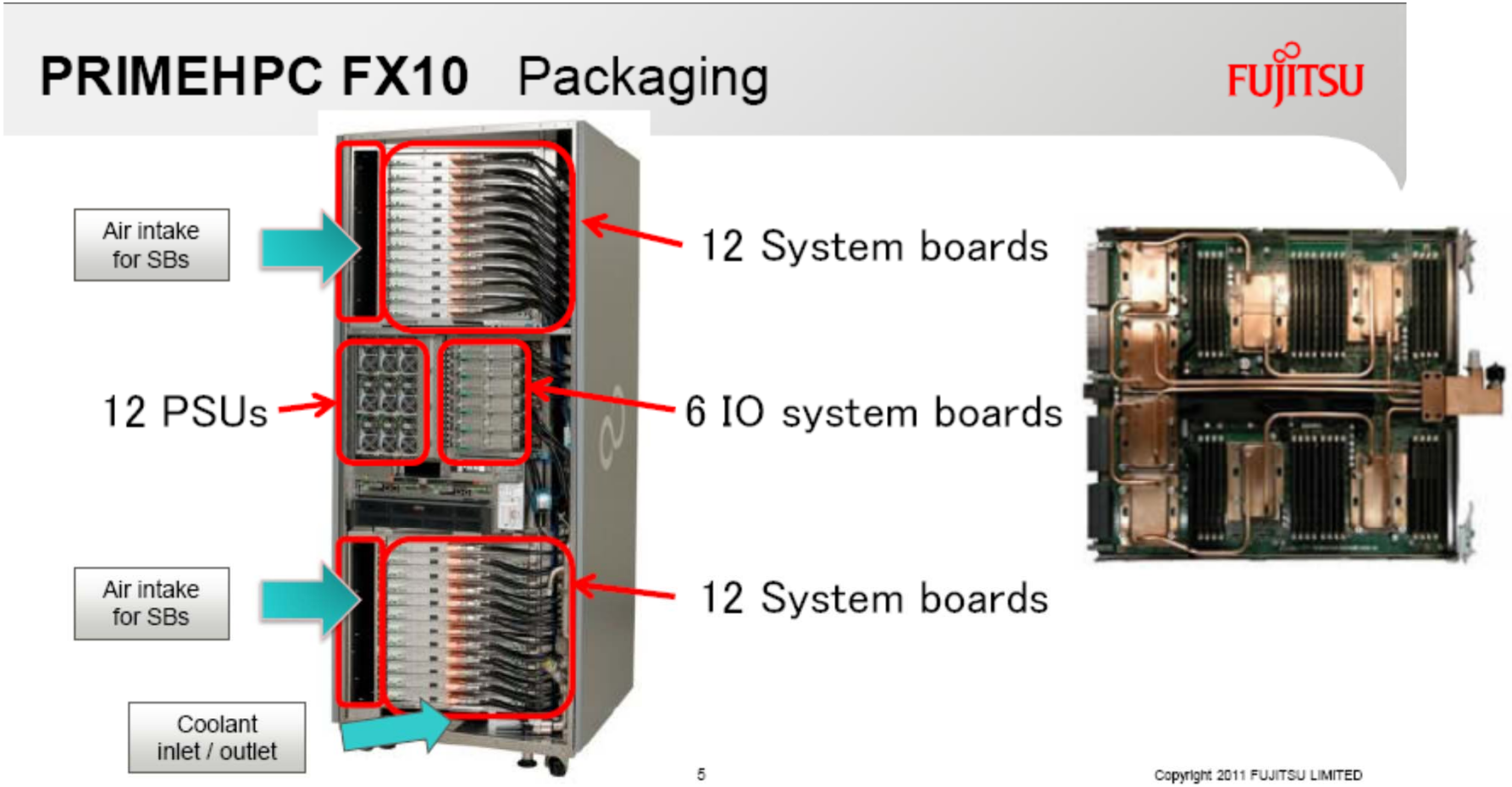




CPU	SPARC64™ IXfx 1.848 GHz	SPARC64™ VIIIfx 2.000 GHz
Number of Cores/Node	16	8
Size of L2 Cache/Node	12 MB	6 MB
Peak Performance/Node	236.5 GFLOPS	128.0 GFLOPS
Memory/Node	32 GB	16 GB
Memory Bandwidth/Node	85 GB/sec (DDR3-1333)	64 GB/sec (DDR3-1000)

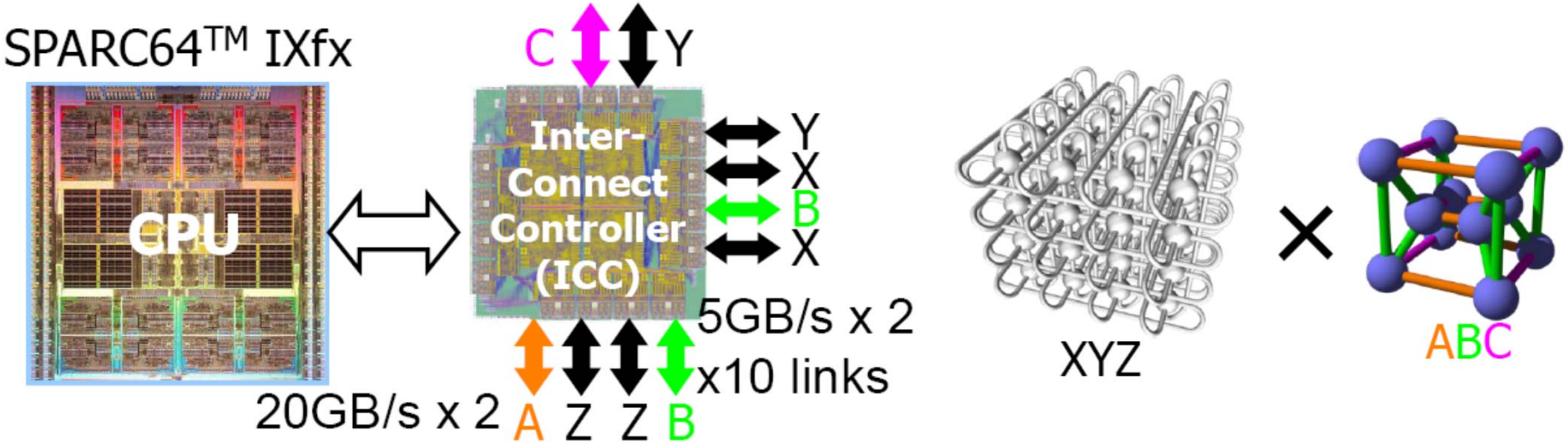
Racks

- A “System Board” with 4 nodes
- A “Rack” with 24 system boards (= 96 nodes)
- Full System with 50 Racks, 4,800 nodes



Tofu Interconnect

- Node Group
  - 12 nodes
  - A/C-axis: on system board, B-axis: 3 system boards
- 6D：(X,Y,Z,A,B,C)
  - ABC 3D Mesh: connects 12 nodes of each node group
  - XYZ 3D Mesh： connects “ABC 3D Mesh” group



Software of FX10

	Computing/Interactive Nodes	Login Nodes
OS	Special OS（XTCOS）	Red Hat Enterprise Linux
Compiler	Fujitsu Fortran 77/90 C／C++ GNU GCC，g95	Fujitsu (Cross Compiler) Fortran 77/90 C／C++ GNU (Cross Compiler) GCC，g95
Library	Fujitsu SSL II (Scientific Subroutine Library II)，C-SSL II，SSL II/MPI Open Source BLAS，LAPACK，ScaLAPACK，FFTW，SuperLU，PETSc，METIS， SuperLU_DIST，Parallel NetCDF	
Applications	OpenFOAM，ABINIT-MP，PHASE， FrontFlow/blue FrontSTR，REVOCAP	
File System	FEFS (based on Lustre)	
Free Software	bash, tcsh, zsh, emacs, autoconf, automake, bzip2, cvs, gawk, gmake, gzip, make, less, sed, tar, vim etc.	

NO ISV/Commercial Applications (e.g. NASTRAN, ABAQUS, ANSYS etc.)

- Yayoi (Hitachi SR16000, IBM Power7)
  - 54.9 TF, Nov. 2011 – Oct. 2017

- Oakleaf-FX (Fujitsu PRIMEHPC FX10)
  - 1.135 PF, Commercial Version of K, Apr.2012 – Mar.2018
- Oakbridge-FX (Fujitsu PRIMEHPC FX10)
  - 136.2 TF, for long-time use (up to 168 hr), Apr.2014 – Mar.2018
- **Reedbush (SGI, Intel BDW + NVIDIA P100 (Pascal))**
  - **Integrated Supercomputer System for Data Analyses & Scientific Simulations**
  - **1.93 PF, Jul.2016-Jun.2020**
  - **Our first GPU System (Mar.2017), DDN IME (Burst Buffer)**

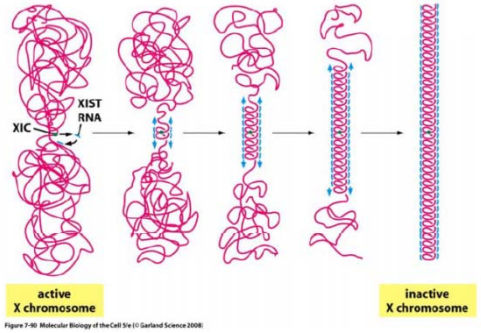
- Oakforest-PACS (OFP) (Fujitsu, Intel Xeon Phi (KNL))
  - JCAHPC (U.Tsukuba & U.Tokyo)
  - 25 PF, #6 in 48th TOP 500 (Nov.2016) (#1 in Japan)
  - Omni-Path Architecture, DDN IME (Burst Buffer)

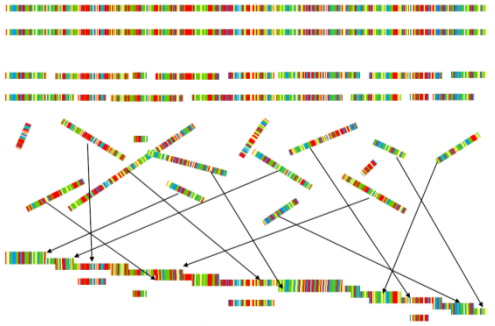
Reasons why we did not introduce systems with GPU's before …

- CUDA
- We have 2,000+ users
- Although we are proud that they are very smart and diligent …

Why have we decided to introduce a system with GPU's this time ?

- Experts in our division
  - Prof's Hanawa, Ohshima & Hoshino
- OpenACC
  - Much easier than CUDA
  - Performance has been improved recently
  - Efforts by Akira Naruse (NVIDIA)
- Data Science, Deep Learning
  - Development of new types of users other than traditional CSE (Computational Science & Engineering)
    - Research Organization for Genome Medical Science, U. Tokyo
    - U. Tokyo Hospital: Processing of Medical Images by Deep Learning

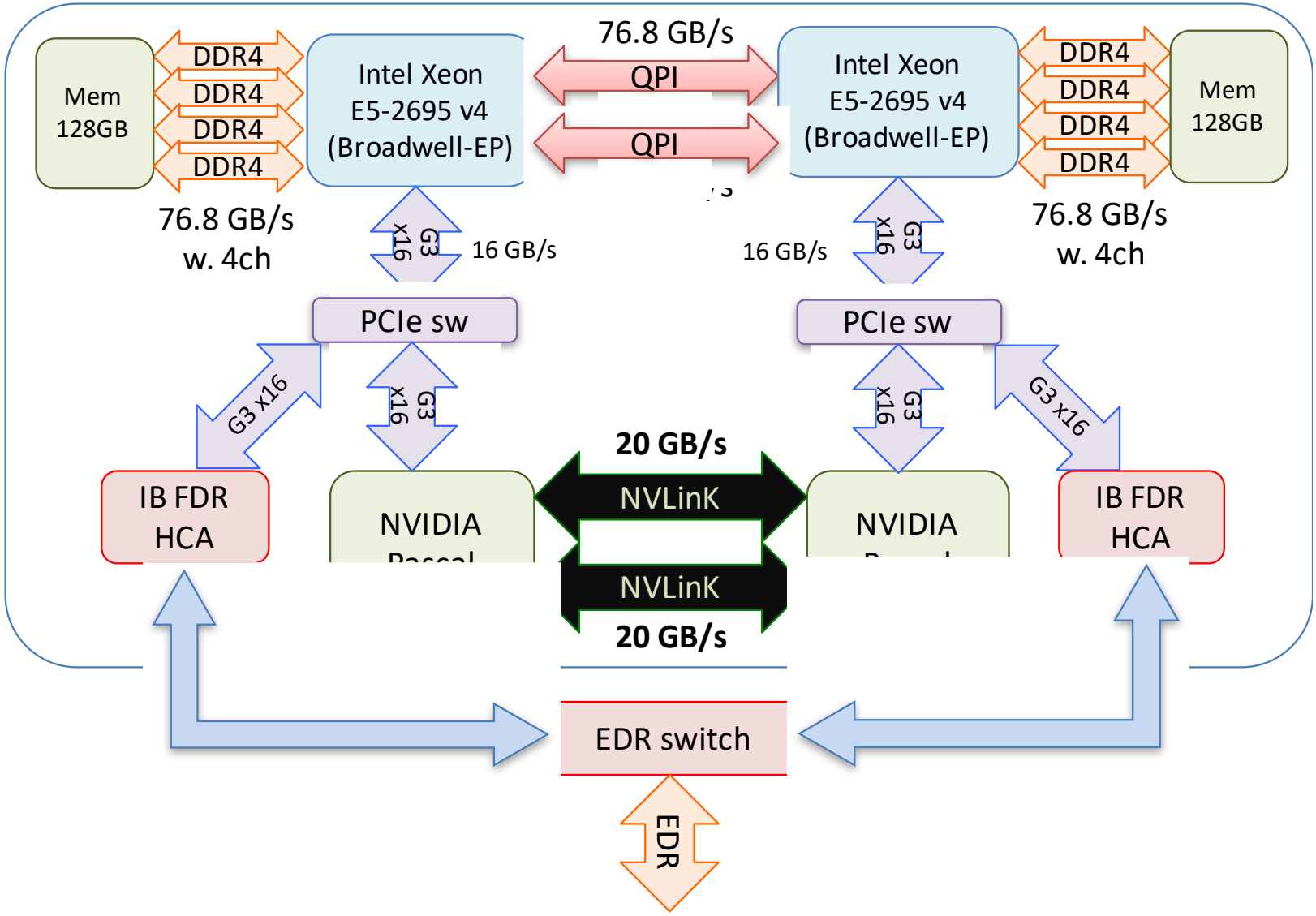


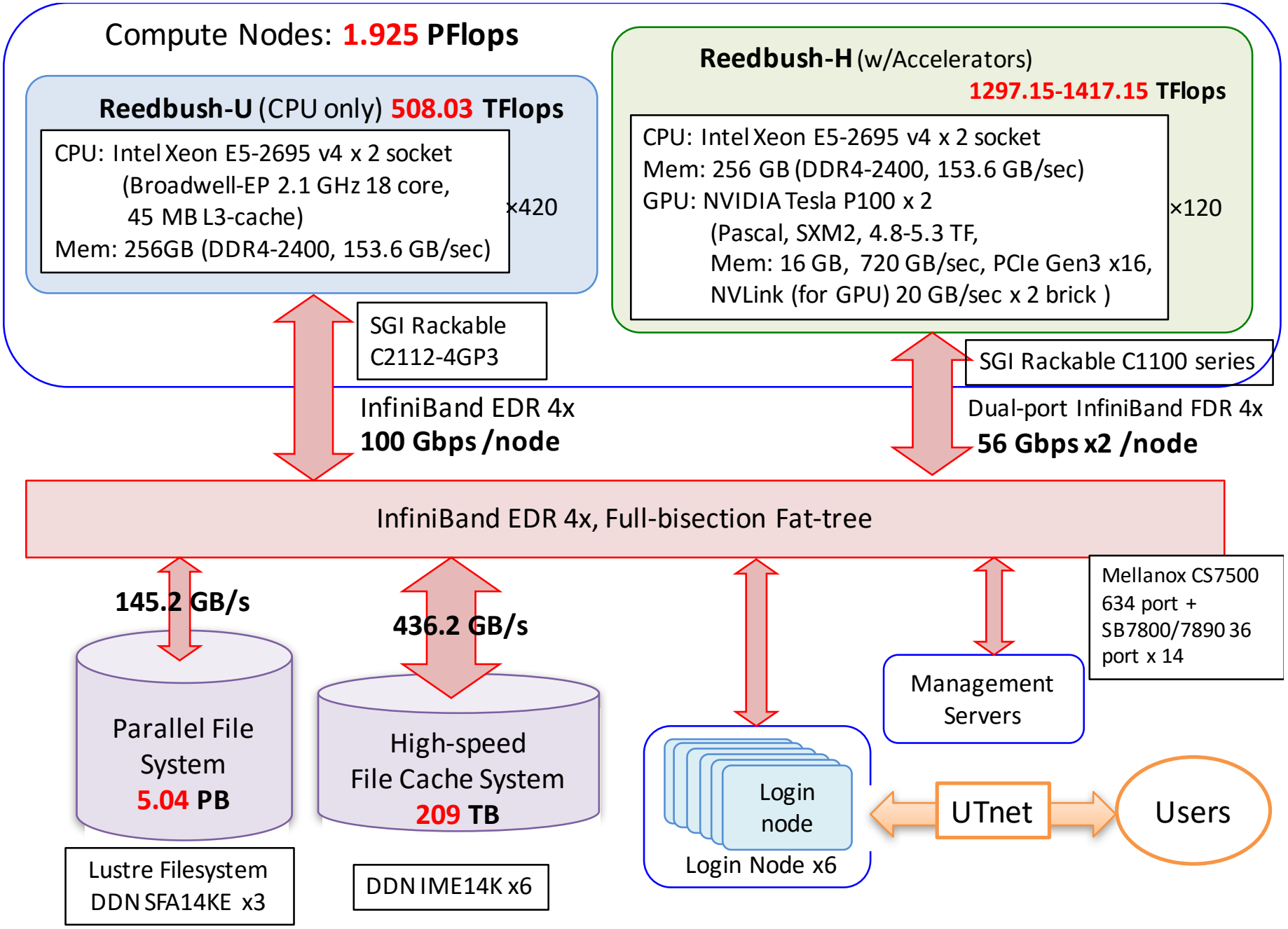


Reedbush (1/2)

- SGI was awarded (Mar. 22, 2016)
- **Compute Nodes (CPU only): Reedbush-U**
  - Intel Xeon E5-2695v4 (Broadwell-EP, 2.1GHz 18core) x 2socket (1.210 TF), 256 GiB (153.6GB/sec)
  - InfiniBand EDR, Full bisection Fat-tree
  - Total System: 420 nodes, 508.0 TF
- **Compute Nodes (with Accelerators): Reedbush-H**
  - Intel Xeon E5-2695v4 (Broadwell-EP, 2.1GHz 18core) x 2socket, 256 GiB (153.6GB/sec)
  - NVIDIA Pascal GPU (Tesla P100)
    - (4.8-5.3TF, 720GB/sec, 16GiB) x 2 / node
  - InfiniBand FDR x 2ch (for ea. GPU), Full bisection Fat-tree
  - 120 nodes, 145.2 TF(CPU)+ 1.15~1.27 PF(GPU)= 1.30~1.42 PF

Configuration of Each Compute Node of Reedbush-H



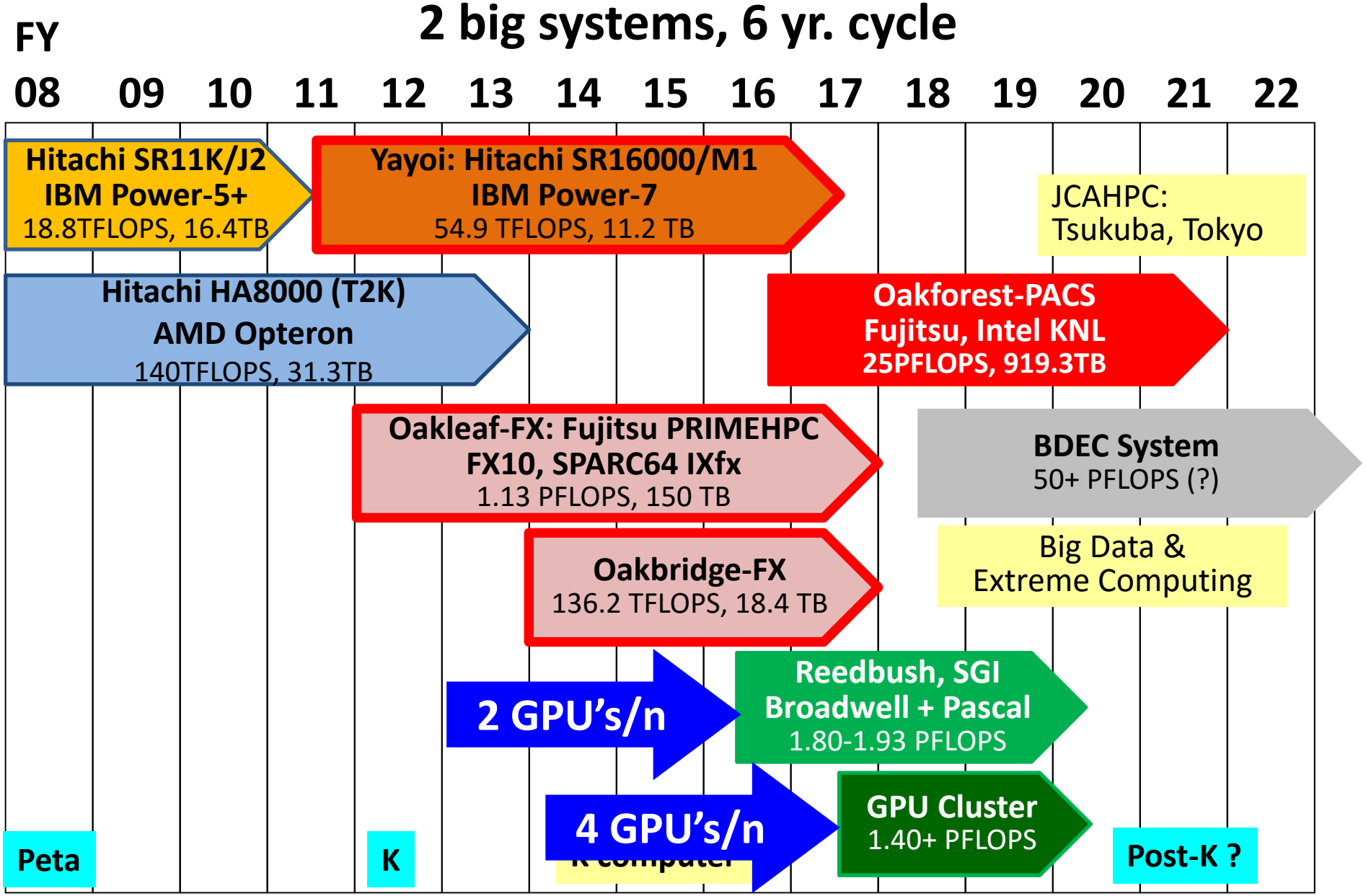

Compute Nodes: 1.925 PFlops
Reedbush-U (CPU only) 508.03 TFlops
CPU: Intel Xeon E5-2695 v4 x 2 socket
(Broadwell-EP 2.1 GHz 18 core,
45 MB L3-cache)
Mem: 256GB (DDR4-2400, 153.6 GB/sec)
×420
Reedbush-H (w/Accelerators)
1297.15-1417.15 TFlops
CPU: Intel Xeon E5-2695 v4 x 2 socket
Mem: 256 GB (DDR4-2400, 153.6 GB/sec)
GPU: NVIDIA Tesla P100 x 2
(Pascal, SXM2, 4.8-5.3 TF,
Mem: 16 GB, 720 GB/sec, PCIe Gen3 x16,
NVLink (for GPU) 20 GB/sec x 2 brick)
×120
SGI Rackable C2112-4GP3
SGI Rackable C1100 series
InfiniBand EDR 4x
100 Gbps /node
Dual-port InfiniBand FDR 4x
56 Gbps x2 /node
InfiniBand EDR 4x, Full-bisection Fat-tree
145.2 GB/s
436.2 GB/s
Mellanox CS7500
634 port +
SB7800/7890 36
port x 14
Parallel File
System
5.04 PB
High-speed
File Cache System
209 TB
Management
Servers
Login
node
Login Node x6
UTnet
Users
Lustre Filesystem
DDN SFA14KE x3
DDN IME14K x6

Reedbush (2/2)

- Storage/File Systems
  - Shared Parallel File-system (Lustre)
    - 5.04 PB, 145.2 GB/sec
  - Fast File Cache System: Burst Buffer (DDN IME (Infinite Memory Engine))
    - SSD: 209.5 TB, 450 GB/sec
- Power, Cooling, Space
  - Air cooling only, < 500 kVA (without A/C): 378 kVA, < 90 m^2
- Software & Toolkit for Data Analysis, Deep Learning …
  - OpenCV, Theano, Anaconda, ROOT, TensorFlow
  - Torch, Caffe, Cheiner, GEANT4
- **Pilot system towards BDEC system after Fall 2018**
  - **New Research Area: Data Analysis, Deep Learning etc.**

Supercomputers in ITC/U.Tokyo

2 big systems, 6 yr. cycle



- Yayoi (Hitachi SR16000, IBM Power7)
  - 54.9 TF, Nov. 2011 – Oct. 2017
- Oakleaf-FX (Fujitsu PRIMEHPC FX10)
  - 1.135 PF, Commercial Version of K, Apr.2012 – Mar.2018
- Oakbridge-FX (Fujitsu PRIMEHPC FX10)
  - 136.2 TF, for long-time use (up to 168 hr), Apr.2014 – Mar.2018
- Reedbush (SGI, Intel BDW + NVIDIA P100 (Pascal))
  - Integrated Supercomputer System for Data Analyses & Scientific Simulations
  - 1.93 PF, Jul.2016-Jun.2020
  - Our first GPU System (Mar.2017), DDN IME (Burst Buffer)
- **Oakforest-PACS (OFP) (Fujitsu, Intel Xeon Phi (KNL))**
  - **JCAHPC (U.Tsukuba & U.Tokyo)**
  - **25 PF, #6 in 48th TOP 500 (Nov.2016) (#1 in Japan)**
  - **Omni-Path Architecture, DDN IME (Burst Buffer)**

Oakforest-PACS

- Full Operation started on December 1, 2016
- 8,208 Intel Xeon/Phi (KNL), 25 PF Peak Performance
  - Fujitsu
- **TOP 500 #6 (#1 in Japan), HPCG #3 (#2), Green 500 #6 (#2) (November 2016)**
- **JCAHPC: Joint Center for Advanced High Performance Computing)**
  - University of Tsukuba
  - University of Tokyo
    - New system will installed in Kashiwa-no-Ha (Leaf of Oak) Campus/U.Tokyo, which is between Tokyo and Tsukuba
  - http://jcahpc.jp

JCAHPC

東京大学 THE UNIVERSITY OF TOKYO

筑波大学 University of Tsukuba

48th TOP500 List (November, 2016)

	Site	Computer/Year Vendor	Cores	R_{max} (TFLOPS)	R_{peak} (TFLOPS)	Power (kW)
1	National Supercomputing Center in Wuxi, China	**Sunway TaihuLight** , Sunway MPP, Sunway SW26010 260C 1.45GHz, 2016 NRCPC	10,649,600	93,015 (= 93.0 PF)	125,436	15,371
2	National Supercomputing Center in Tianjin, China	**Tianhe-2**, Intel Xeon E5-2692, TH Express-2, Xeon Phi, 2013 NUDT	3,120,000	33,863 (= 33.9 PF)	54,902	17,808
3	Oak Ridge National Laboratory, USA	**Titan** Cray XK7/NVIDIA K20x, 2012 Cray	560,640	17,590	27,113	8,209
4	Lawrence Livermore National Laboratory, USA	**Sequoia** BlueGene/Q, 2011 IBM	1,572,864	17,173	20,133	7,890
5	DOE/SC/LBNL/NERSC USA	**Cori**, Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Cray Aries, 2016 Cray	632,400	14,015	27,881	3,939
6	Joint Center for Advanced High Performance Computing, Japan	**Oakforest-PACS**, PRIMERGY CX600 M1, Intel Xeon Phi Processor 7250 68C 1.4GHz, Intel Omni-Path, 2016 Fujitsu	557,056	13,555	24,914	2,719
7	RIKEN AICS, Japan	**K computer**, SPARC64 VIIIfx , 2011 Fujitsu	705,024	10,510	11,280	12,660
8	Swiss Natl. Supercomputer Center, Switzerland	**Piz Daint** Cray XC30/NVIDIA P100, 2013 Cray	206,720	9,779	15,988	1,312
9	Argonne National Laboratory, USA	**Mira** BlueGene/Q, 2012 IBM	786,432	8,587	10,066	3,945
10	DOE/NNSA/LANL/SNL, USA	**Trinity**, Cray XC40, Xeon E5-2698v3 16C 2.3GHz, 2016 Cray	301,056	8,101	11,079	4,233

R_{max}: Performance of Linpack (TFLOPS)
R_{peak}: Peak Performance (TFLOPS), Power: kW

http://www.top500.org/

48th TOP500 List (November, 2016)

	Site	Computer/Year Vendor	Cores	R_{max} (TFLOPS)	R_{peak} (TFLOPS)	Power (kW)
1	National Supercomputing Center in Wuxi, China	**Sunway TaihuLight** , Sunway MPP, Sunway SW26010 260C 1.45GHz, 2016 NRCPC	10,649,600	93,015 (= 93.0 PF)	125,436	15,371
2	National Supercomputing Center in Tianjin, China	**Tianhe-2**, Intel Xeon E5-2692, TH Express-2, Xeon Phi, 2013 NUDT	3,120,000	33,863 (= 33.9 PF)	54,902	17,808
3	Oak Ridge National Laboratory, USA	**Titan** Cray XK7/NVIDIA K20x, 2012 Cray	560,640	17,590	27,113	8,209
4	Lawrence Livermore National Laboratory, USA	**Sequoia** BlueGene/Q, 2011 IBM	1,572,864	17,173	20,133	7,890
5	DOE/SC/LBNL/NERSC USA	**Cori**, Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Cray Aries, 2016 Cray	632,400	14,015	27,881	3,939
6	Joint Center for Advanced High Performance Computing, Japan	**Oakforest-PACS**, PRIMERGY CX600 M1, Intel Xeon Phi Processor 7250 68C 1.4GHz, Intel Omni-Path, 2016 Fujitsu	557,056	13,555	24,914	2,719
7	RIKEN AICS, Japan	**K computer**, SPARC64 VIIIfx , 2011 Fujitsu	705,024	10,510	11,280	12,660
8	Swiss Natl. Supercomputer Center, Switzerland	**Piz Daint** Cray XC30/NVIDIA P100, 2013 Cray	206,720	9,779	15,988	1,312
9	Argonne National Laboratory, USA	**Mira** BlueGene/Q, 2012 IBM	786,432	8,587	10,066	3,945
10	DOE/NNSA/LANL/SNL, USA	**Trinity**, Cray XC40, Xeon E5-2698v3 16C 2.3GHz, 2016 Cray	301,056	8,101	11,079	4,233
104	**ITC/U. Tokyo Japan**	**Oakleaf-FX SPARC64 IXfx, 2012 Fujitsu**	**76800**	**1043**	**1135**	**1177**

HPCG Ranking (SC16, November, 2016)

	Site	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	HPCG/ HPL (%)
1	RIKEN AICS, Japan	K computer	705,024	10.510	7	0.6027	5.73
2	NSCC / Guangzhou, China	Tianhe-2	3,120,000	33.863	2	0.5800	1.71
3	JCAHPC, Japan	Oakforest-PACS	557,056	13.555	6	0.3855	2.84
4	National Supercomputing Center in Wuxi, China	Sunway TaihuLight	10,649,600	93.015	1	0.3712	.399
5	DOE/SC/LBNL/NERSC USA	Cori	632,400	13.832	5	0.3554	2.57
6	DOE/NNSA/LLNL, USA	Sequoia	1,572,864	17.173	4	0.3304	1.92
7	DOE/SC/ Oak Ridge National Laboratory, USA	Titan	560,640	17.590	3	0.3223	1.83
8	DOE/NNSA/ LANL/SNL, USA	Trinity	301,056	8.101	10	0.1826	2.25
9	NASA / Mountain View, USA	Pleiades: SGI ICE X	243,008	5.952	13	0.1752	2.94
10	DOE/SC/ Argonne National Laboratory, USA	Mira: IBM BlueGene/Q,	786,432	8.587	9	0.1670	1.94

http://www.hpcg-benchmark.org/

Green 500 Ranking (SC16, November, 2016)

	Site	Computer	CPU	HPL Rmax (Pflop/s)	TOP500 Rank	Power (MW)	GFLOPS/W
1	NVIDIA Corporation	DGX SATURNV	NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100	3.307	28	0.350	9.462
2	Swiss National Supercomputing Centre (CSCS)	Piz Daint	Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100	9.779	8	1.312	7.454
3	RIKEN ACCS	Shoubu	ZettaScaler-1.6 etc.	1.001	116	0.150	6.674
4	National SC Center in Wuxi	Sunway TaihuLight	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	93.01	1	15.37	6.051
5	SFB/TR55 at Fujitsu Tech. Solutions GmbH	QPACE3	PRIMERGY CX1640 M1, Intel Xeon Phi 7210 64C 1.3GHz, Intel Omni-Path	0.447	375	0.077	5.806
6	JCAHPC	Oakforest-PACS	PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path	1.355	6	2.719	4.986
7	DOE/SC/Argonne National Lab.	Theta	Cray XC40, Intel Xeon Phi 7230 64C 1.3GHz, Aries interconnect	5.096	18	1.087	4.688
8	Stanford Research Computing Center	XStream	Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	0.781	162	0.190	4.112
9	ACCMS, Kyoto University	Camphor 2	Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect	3.057	33	0.748	4.087
10	Jefferson Natl. Accel. Facility	SciPhi XVI	KOI Cluster, Intel Xeon Phi 7230 64C 1.3GHz, Intel Omni-Path	0.426	397	0.111	3.837

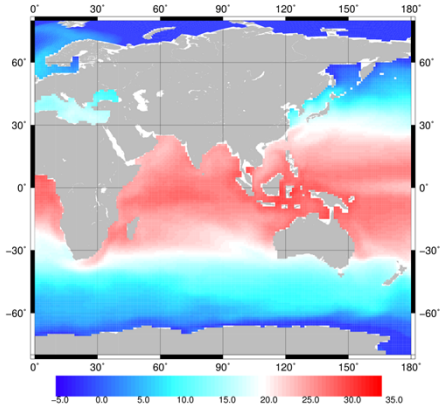
http://www.top500.org/

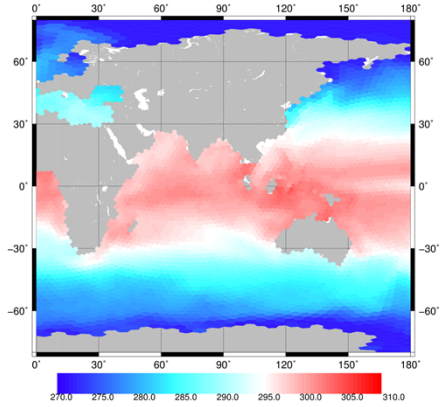
Software of Oakforest-PACS

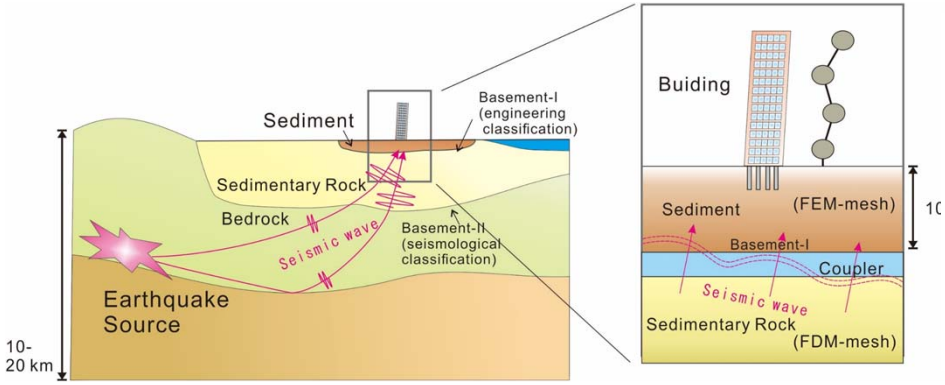
- OS: Red Hat Enterprise Linux (Login nodes), CentOS or McKernel (Compute nodes, dynamically switchable)
  - McKernel: OS for many-core CPU developed by RIKEN AICS
    - Ultra-lightweight OS compared with Linux, no background noise to user program
    - Expected to be installed to post-K computer
- Compiler: GCC, Intel Compiler, XcalableMP
  - XcalableMP: Parallel programming language developed by RIKEN AICS and University of Tsukuba
    - Easy to develop high-performance parallel application by adding directives to original code written by C or Fortran
- Application: Open-source softwares
  - : ppOpen-HPC, OpenFOAM, ABINIT-MP, PHASE system, FrontFlow/blue and so on

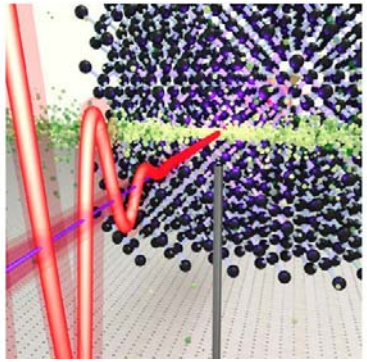
Applications on Oakforest-PACS

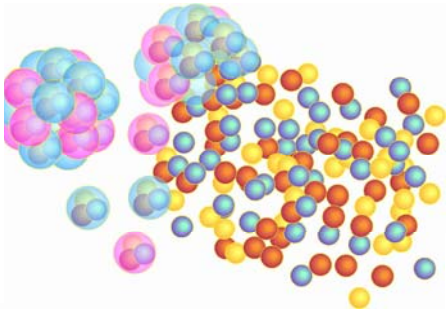
- ARTED: Ab-initio Real Time Electron Dynamics simulator
- Lattice QCD: Quantum Chrono Dynamics
- NICAM & COCO: Atmosphere & Ocean Coupling
- GAMERA/GOJIRA: Earthquake Simulations
- Seism3D: Seismic Wave Propagation







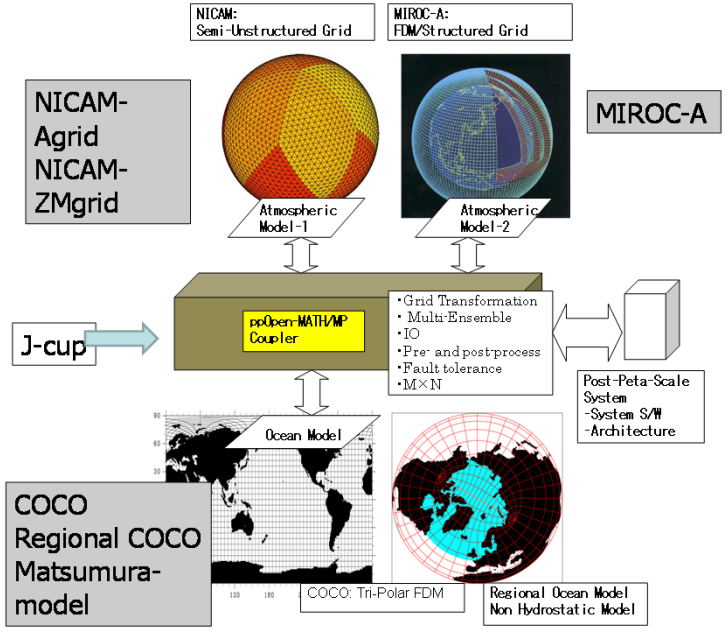


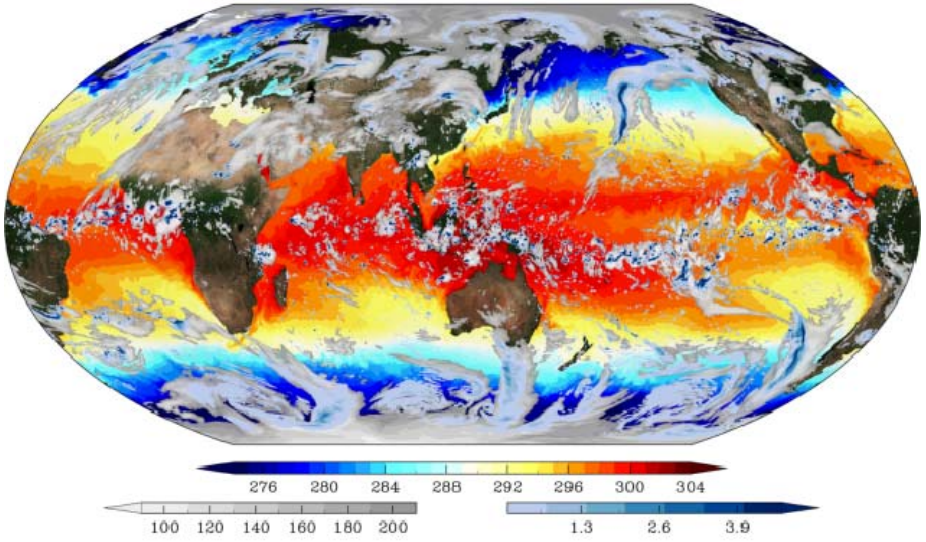


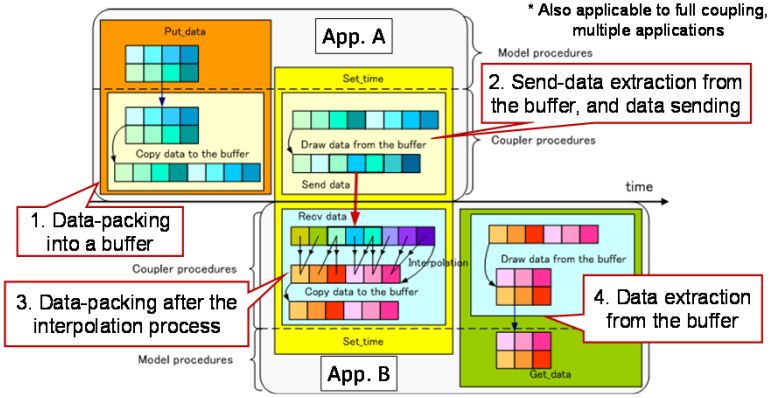
Atmosphere-Ocean Coupling

on OFP by NICAM/COCO/ppOpen-MATH/MP

- High-resolution global atmosphere-ocean coupled simulation by NICAM and COCO (Ocean Simulation) through ppOpen-MATH/MP on the K computer is achieved.
  - ppOpen-MATH/MP is a coupling software for the models employing various discretization method.
- An O(km)-mesh NICAM-COCO coupled simulation is planned on the Oakforest-PACS system.
  - A big challenge for optimization of the codes on new Intel Xeon Phi processor
  - New insights for understanding of global climate dynamics







[C/O M. Satoh (AORI/UTokyo)@SC16]