東京大学
THE UNIVERSITY OF TOKYO

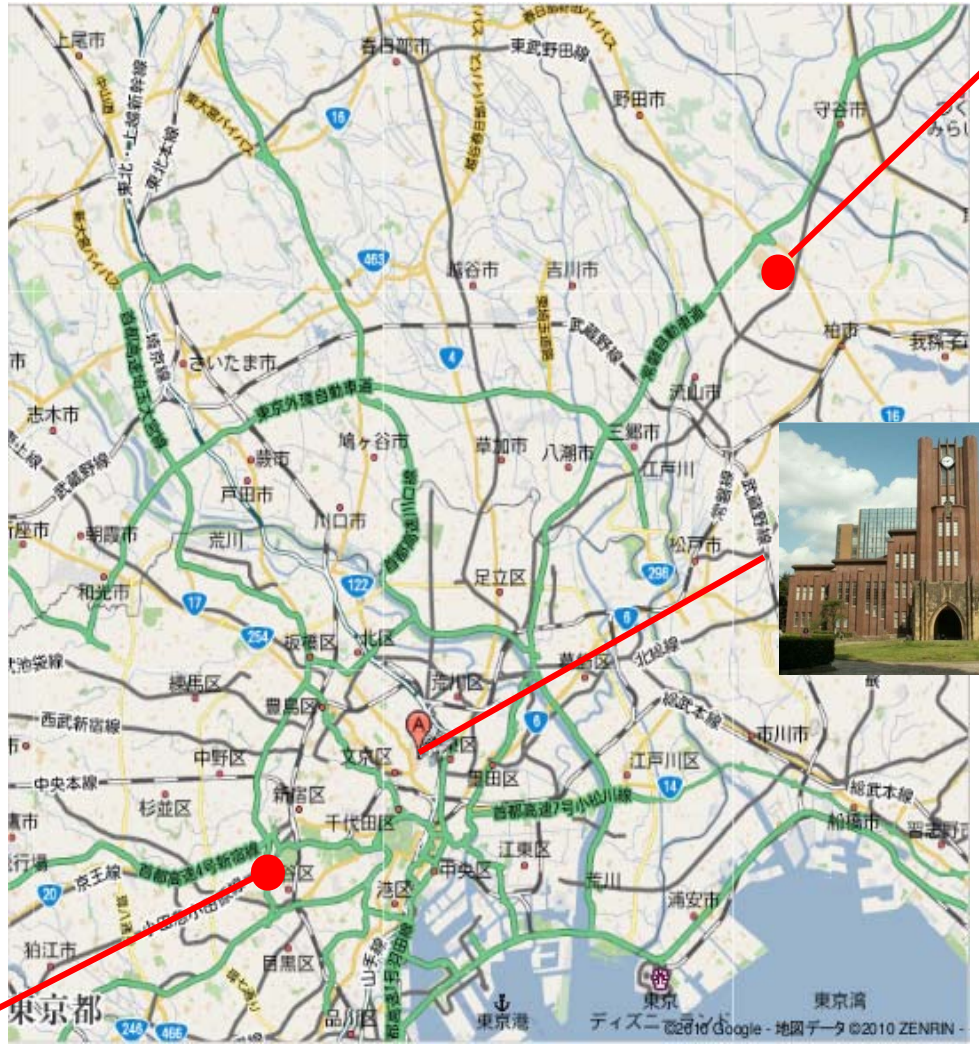
Accelerated Computing Activities at ITC/University of Tokyo

Kengo Nakajima
Information Technology Center (ITC)
The University of Tokyo

ADAC-3 Workshop
January 25-27 2017, Kashiwa, Japan

Three Major Campuses of University of Tokyo

東京大学
THE UNIVERSITY OF TOKYO



Kashiwa

Hongo

Komaba

Kashiwa（柏）: Oak

Kashiwa-no-Ha: Leaf of Oak

OAK RIDGE National Laboratory

“Kashi（樫）” is also called “Oak”

“Oak” covers many kinds of trees in Japan

- Overview of Information Technology Center (ITC), The University of Tokyo
  - Reedbush
    - Integrated Supercomputer System for Data Analyses & Scientific Simulations
    - Our First System with GPU's
  - Oakforest-PCAS
    - #1 System in Japan !
- Performance of the New Systems
  - 3D Poisson Solver by ICCG
  - Matrix Assembly for 3D FEM

Information Technology Center The University of Tokyo (ITC/U.Tokyo)

- Campus/Nation-wide Services on Infrastructure for Information, related Research & Education
- Established in 1999
  - Campus-wide Communication & Computation Division
  - Digital Library/Academic Information Science Division
  - Network Division
  - Supercomputing Division
- Core Institute of Nation-wide Infrastructure Services/Collaborative Research Projects
  - Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (JHPCN) (2010-)
  - HPCI (HPC Infrastructure)

Supercomputing Research Division of ITC/U.Tokyo (SCD/ITC/UT)

http://www.cc.u-tokyo.ac.jp

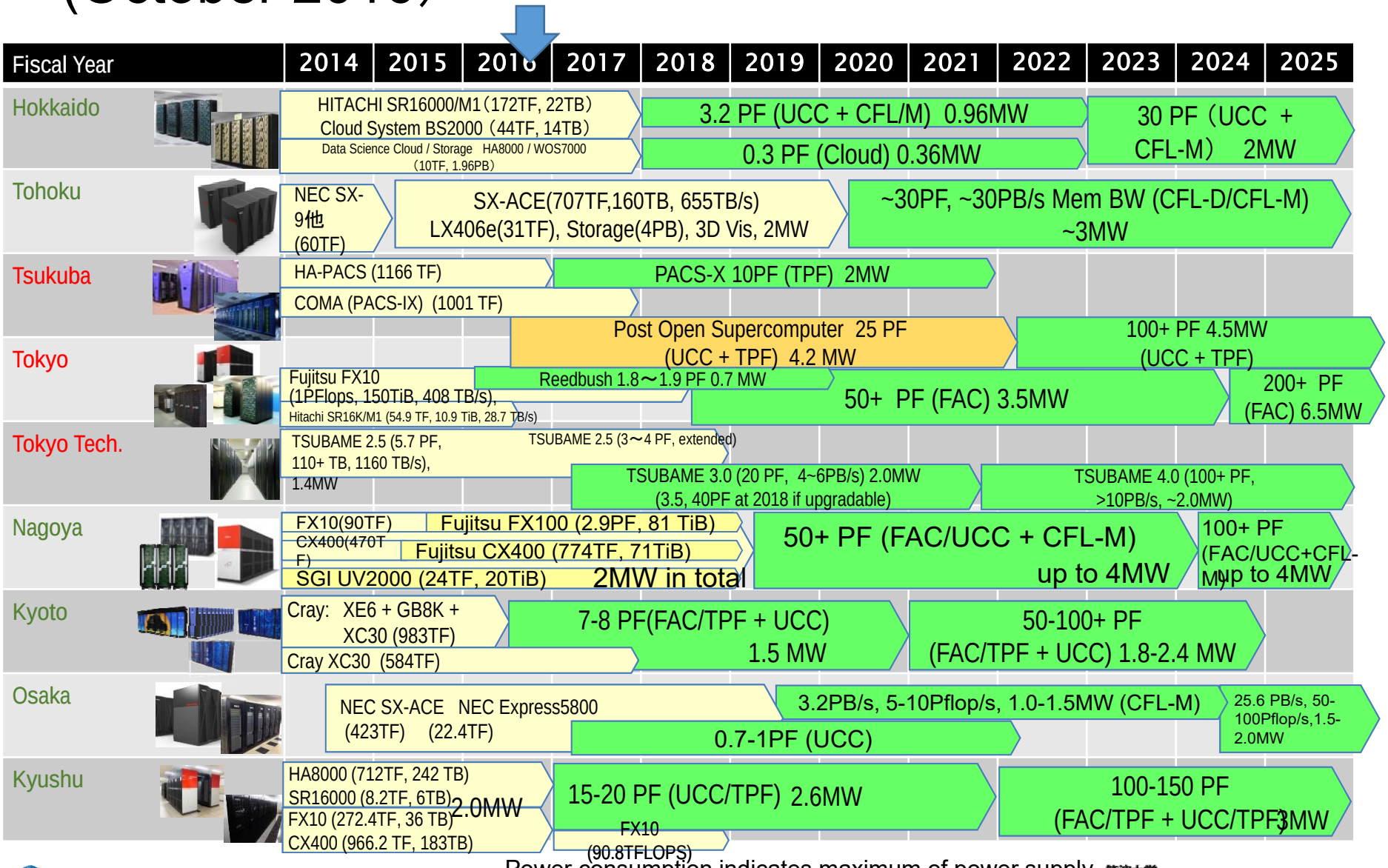
- Services & Operations of Supercomputer Systems, Research, Education
- History
  - Supercomputing Center, U.Tokyo (1965~1999)
    - Oldest Academic Supercomputer Center in Japan
    - Nation-Wide, Joint-Use Facility: Users are not limited to researchers and students of U.Tokyo
  - Information Technology Center (1999~) (4 divisions)
- 10+ Faculty Members (including part-time members)
  - Architecture, System S/W, Algorithms, Applications, GPU
- 8 Technical Staffs

Research Activities

- Collaboration with Users
  - Linear Solvers, Parallel Vis., Performance Tuning
- Research Projects
  - FP3C (collab. with French Institutes) (FY.2010-2013)
    - Tsukuba, Tokyo Tech, Kyoto
  - Feasibility Study of Advanced HPC in Japan (towards Japanese Exascale Project) (FY.2012-2013)
    - 1 of 4 Teams: General Purpose Processors, Latency Cores
  - ppOpen-HPC (FY.2011-2015, 2016-2018)
  - Post K with RIKEN AICS (FY.2014-)
  - ESSEX-II (FY.2016-2018): German-Japan Collaboration
- International Collaboration
  - Lawrence Berkeley National Laboratory (USA)
  - National Taiwan University (Taiwan)
  - National Central University (Taiwan)
  - Intel Parallel Computing Center
  - ESSEX-II/SPPEXA/DFG (Germany)

Plan of 9 Tier-2 Supercomputer Centers (October 2016)

JCAHPC



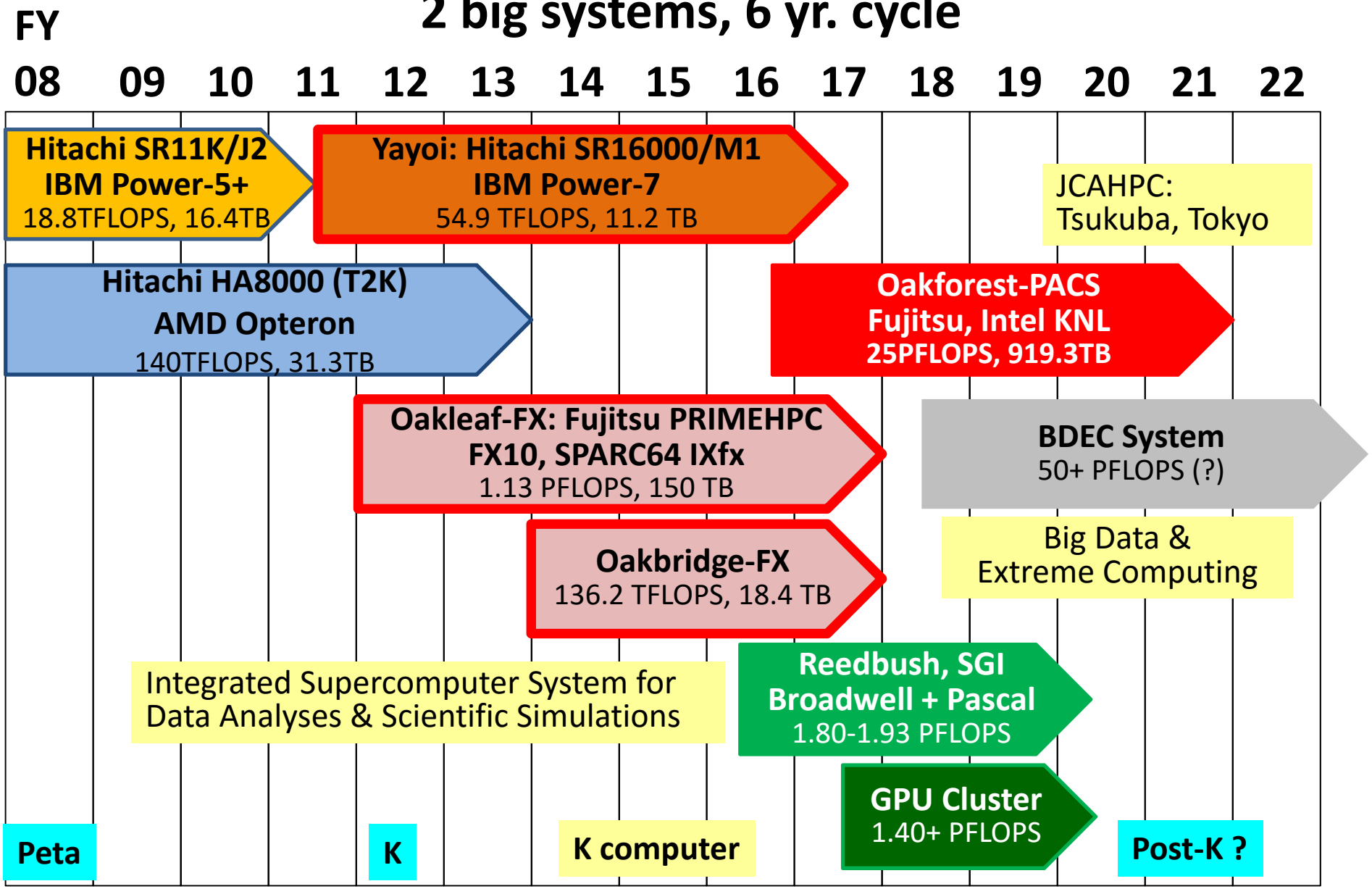
Power consumption indicates maximum of power supply (includes cooling facility)

東京大学情報基盤センター
INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

筑波大学 計算科学研究センター
Center for Computational Sciences

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)


Visiting Tour this Afternoon

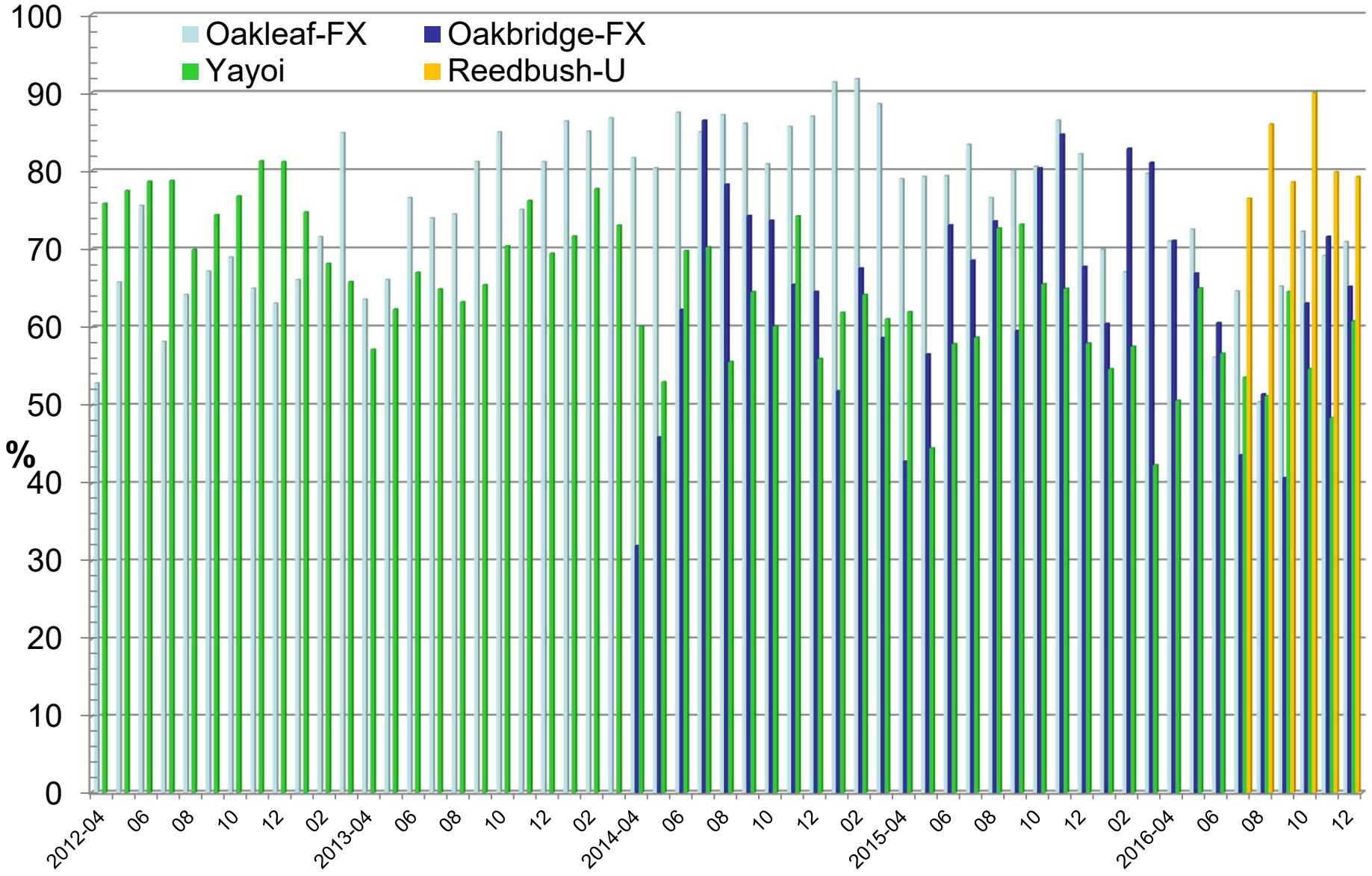
- Yayoi (Hitachi SR16000, IBM Power7)
  - 54.9 TF, Nov. 2011 – Oct. 2017

- **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)

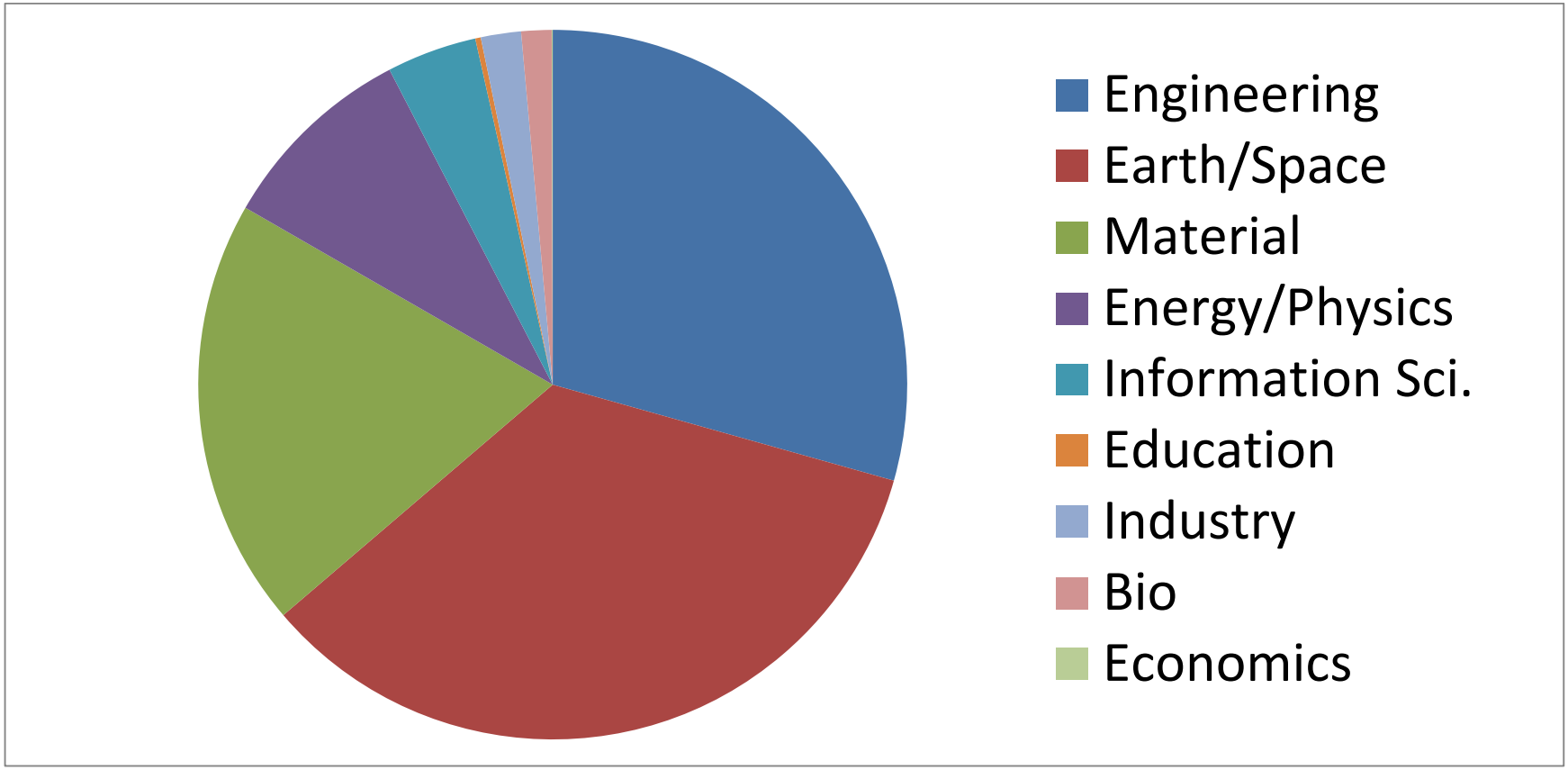
- **Oakforest-PACS (OFP) (Fujitsu, Intel Xeon Phi (KNL))**
  - **JCAHPC (U.Tsukuba & U.Tokyo)**
  - **25 PF, #6 in 48th TOP 500 (Nov.2016) (#1 in Japan)**
  - **Omni-Path Architecture, DDN IME (Burst Buffer)**

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

Services for Industry

- Originally, only academic users have been allowed to access our supercomputer systems.
- Since FY.2008, we started services for industry
  - supports to start large-scale computing for future business
  - not compete with private data centers, cloud services …
  - basically, results must be open to public
  - max 10% total comp. resource is open for usage by industry
  - special qualification processes/special (higher) fee for usage
- Various Types of Services
  - Normal usage (more expensive than academic users)
    - 3-4 groups per year, fundamental research
  - Trial usage with discount rate
  - Research collaboration with academic rate (e.g. Taisei)
  - Open-Source/In-House Codes (NO ISV/Commercial App.)

Training & Education

- 2-Day “Hands-on” Tutorials for Parallel Programming by Faculty Members of SCD/ITC (Free)
  - Participants from industry are accepted.
- Graduate/Undergraduate Classes with Supercomputer System (Free)
  - We encourage faculty members to introduce hands-on tutorial of supercomputer system into graduate/undergraduate classes.
  - Up to 12 nodes (192 cores) of Oakleaf-FX, 8 nodes (288 cores) of Reedbush-U
  - Proposal-based
  - Not limited to Classes of the University of Tokyo, 2-3 of 10
- RIKEN AICS Summer/Spring School (2011~)

- 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)**

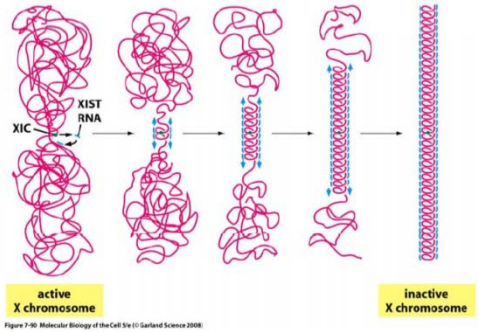
- Oakforest-PACS (OFP) (Fujitsu, Intel Xeon Phi (KNL))
  - JCAHPC (U.Tsukuba & U.Tokyo)
  - 25 PF, #6 in 48th TOP 500 (Nov.2016) (#1 in Japan)
  - Omni-Path Architecture, DDN IME (Burst Buffer)

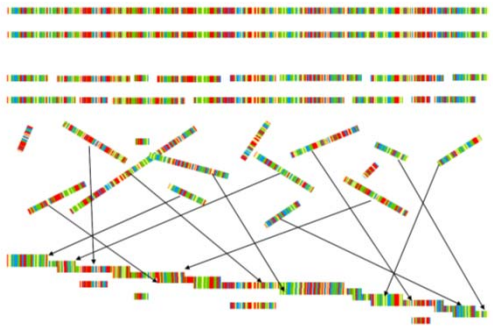
Reasons why we did not introduce systems with GPU's before …

- CUDA
- We have 2,000+ users
- Although we are proud that they are very smart and diligent …
- Therefore, we have decided to adopt Intel XXX for the Post T2K in Summer 2010.

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

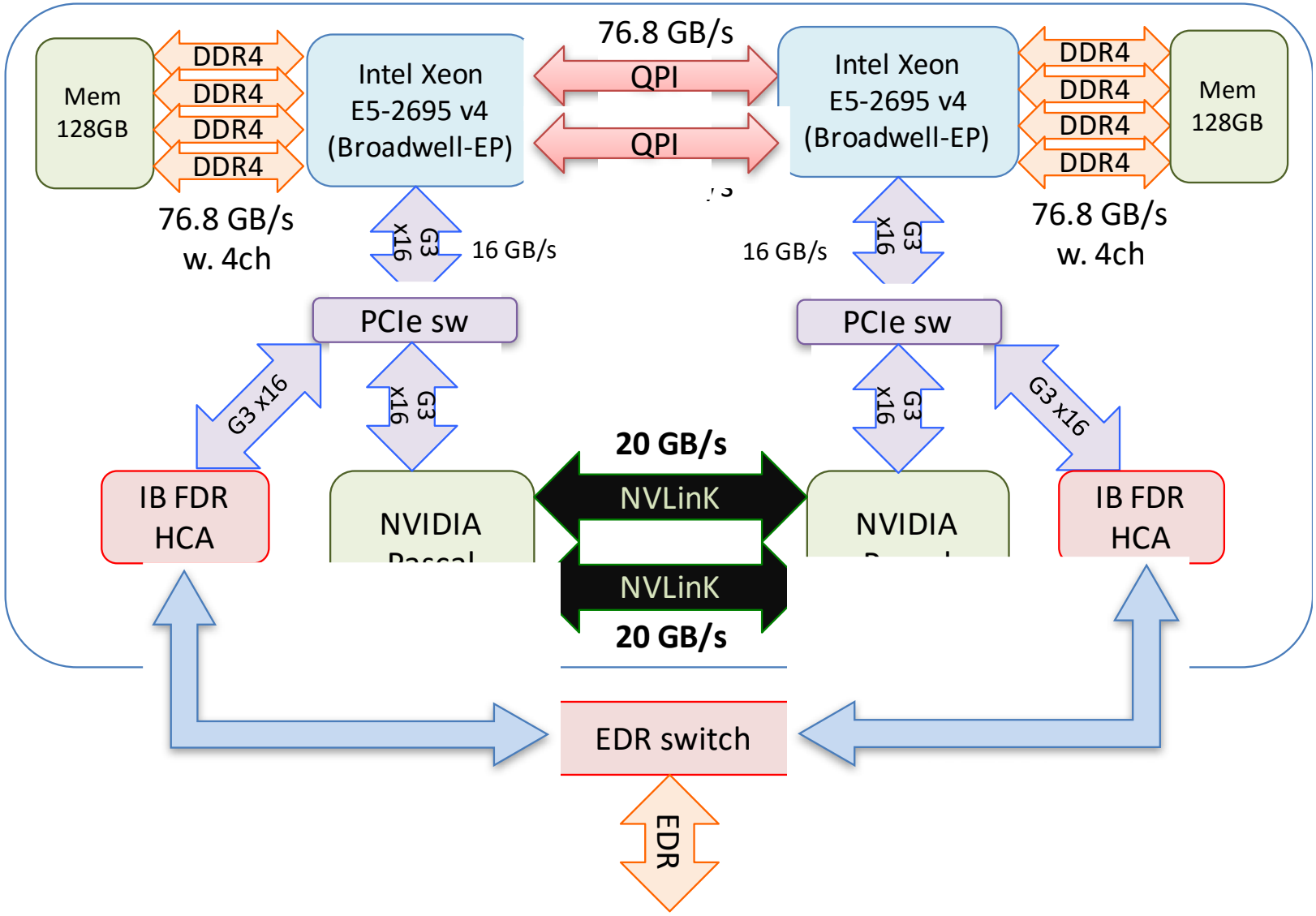


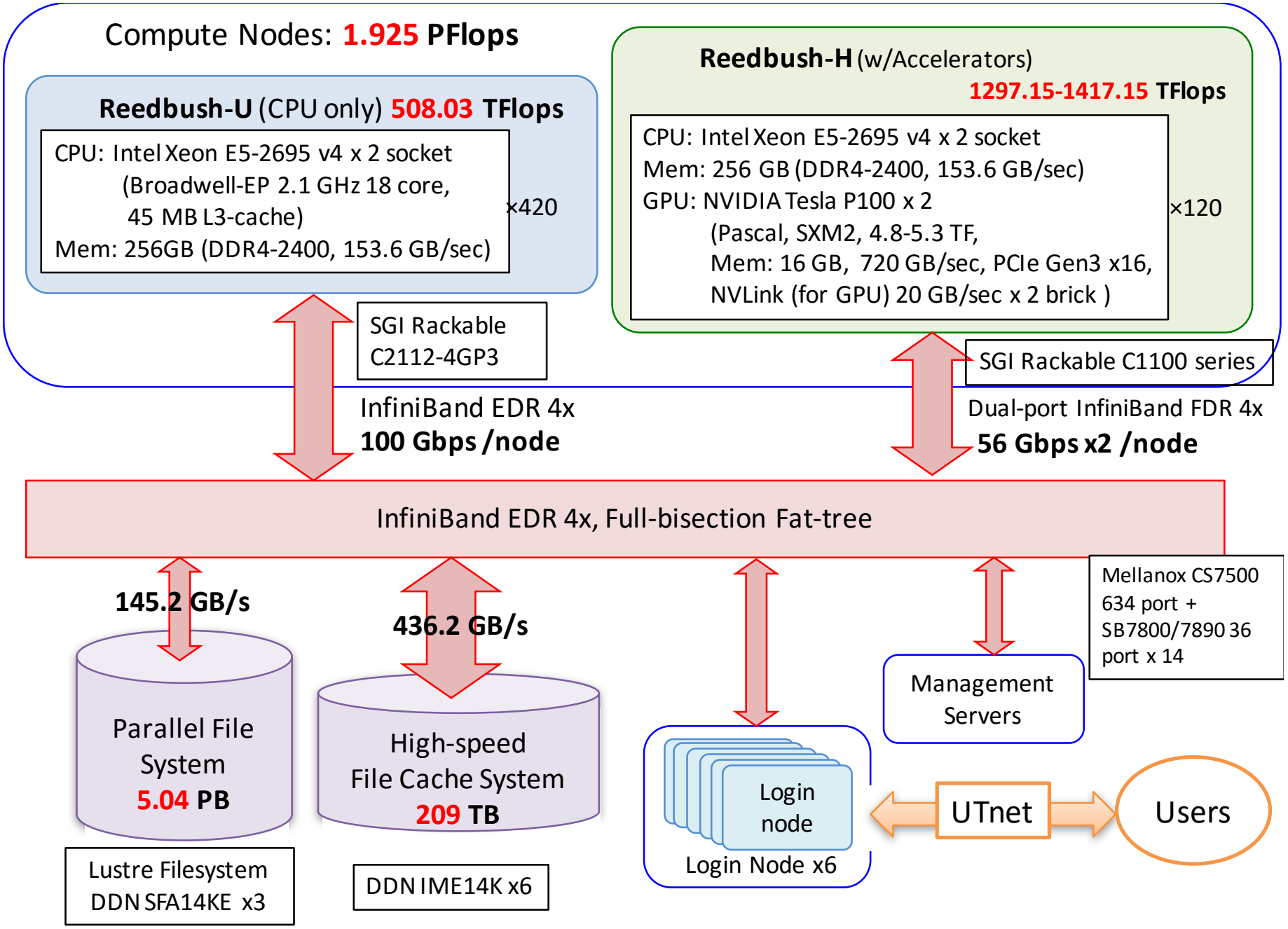


Reedbush (1/2)

- 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

Why “Reedbush” ?

Blaise Pascal
(1623-1662)

- L'homme est un roseau pensant.
- Man is a thinking reed.
- 人間は考える葦である

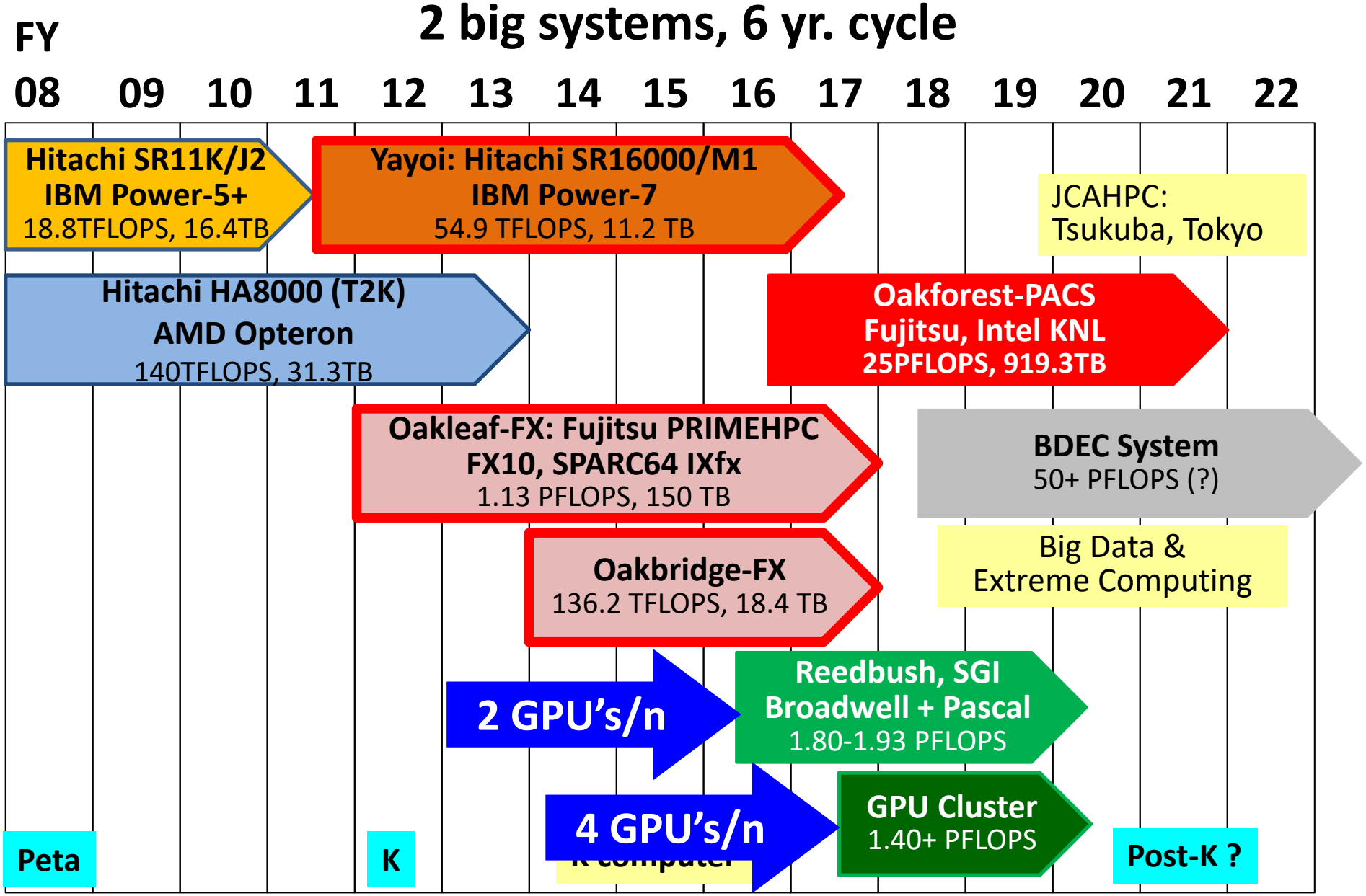
Pensées (Blaise Pascal)

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

JCAHPC: Joint Center for Advanced High Performance Computing

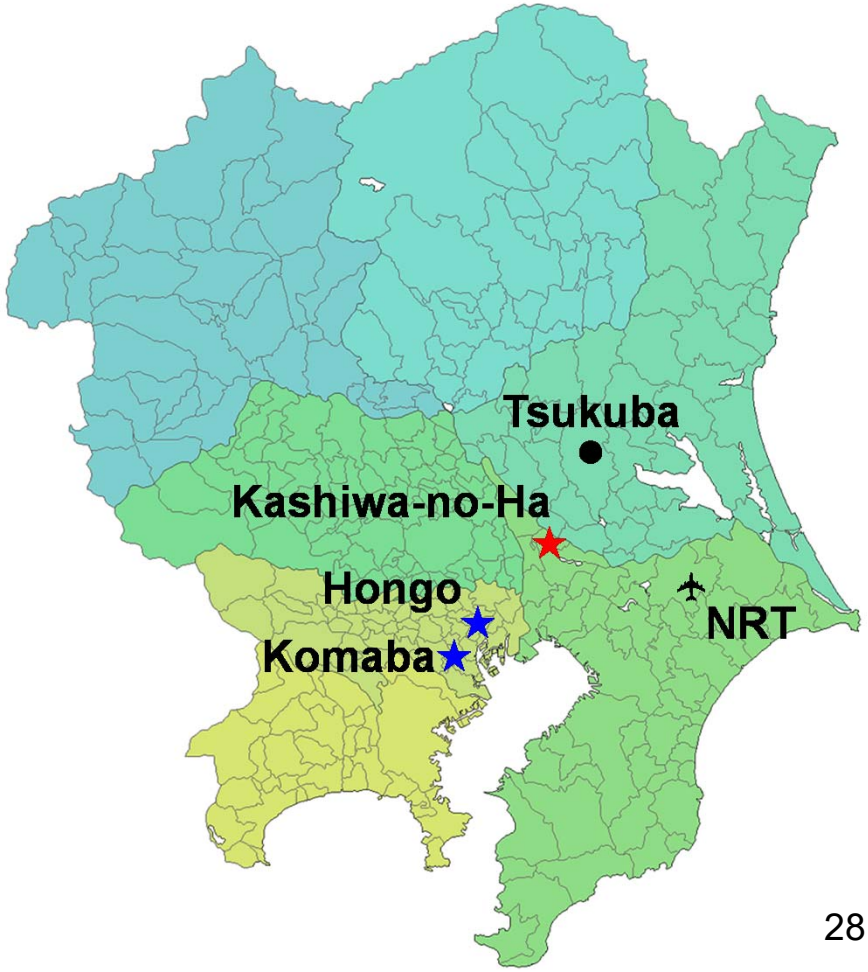
http://jcahpc.jp

- JCAHPC was established in 2013 under agreement between the University of Tsukuba & the University of Tokyo for collaborative promotion of HPC by:
  - Center for Computational Sciences (CCS) at University of Tsukuba, and
  - Information Technology Center (ITC) at the University of Tokyo.
- Design, operate and manage a next-generation supercomputer system by researchers belonging to two universities

Kashiwa is the ideal place for collaboration between U.Tokyo & U. Tsukuba

筑波大学
University of Tsukuba

東京大学
THE UNIVERSITY OF TOKYO



Features of Oakforest-PACS (1/2)

- Computing Node
  - 68 cores/node, 3 TFLOPS x 8,208= 25 PFLOPS
  - 2 Type of Memory
    - MCDRAM: High-Speed, 16GB
    - DDR4: Medium-Speed, 96GB
  - Variety of Selections for Memory-Mode/Cluster-Mode
- Node-to-Node Communication
  - Fat-Tree Network with Full Bi-Section Bandwidth
  - Intel's Omni-Path Architecture
    - High Efficiency for Applications with Full Nodes of the System
    - Flexible and Efficient Operations for Multiple Jobs

Specification of Oakforest-PACS system

Total peak performance			25 PFLOPS
Total number of compute nodes			8,208
Compute node	Product		Fujitsu PRIMERGY CX600 M1 (2U) + CX1640 M1 x 8node
	Processor		Intel® Xeon Phi™ 7250 (Code name: Knights Landing), 68 cores, 1.4 GHz
	Memory	**High BW**	**16 GB, 490 GB/sec (MCDRAM, effective rate)**
		Low BW	**96 GB, 115.2 GB/sec (peak rate)**
Inter-connect	Product		Intel® Omni-Path Architecture
	Link speed		100 Gbps
	Topology		Fat-tree with (completely) full-bisection bandwidth

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

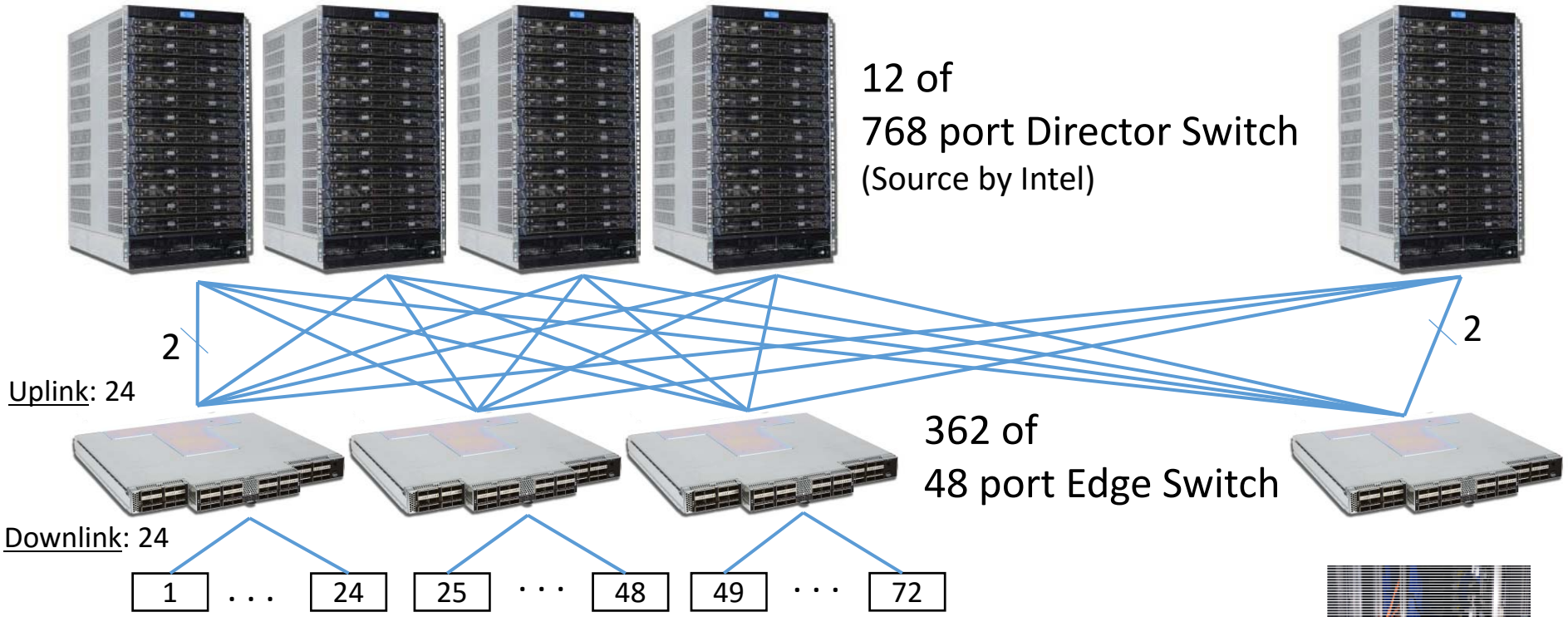
Chassis with 8 nodes, 2U size

15 Chassis with 120 nodes per Rack

東京大学情報基盤センター
INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

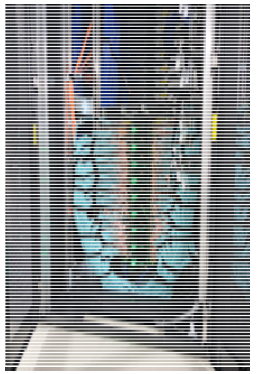
計算科学研究センター
Center for Computational Sciences

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 preferred for flexible job management.



Features of Oakforest-PACS (2/2)

- File I/O
  - Shared File System
    - Lustre, 26PB
  - File Cache System
    - (DDN IME, Burst Buffer): 1+TB/sec Performance for IOR Benchmarks
    - useful not only for computational sciences, but also for big-data analyses, machine learning etc.
- Power Consumption
  - #6 in Green 500
  - Linpack
    - 4,986 MFLOPS/W (OFP)
    - 830 MFLOPS/W（京）

Specification of Oakforest-PACS system (Cont'd)

Parallel File System	Type	Lustre File System
	Total Capacity	26.2 PB
	Product	DataDirect Networks SFA14KE
	Aggregate BW	500 GB/sec
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
	Aggregate BW	1,560 GB/sec
Power consumption		4.2 MW (including cooling)
# of racks		102

Benchmarks

- TOP 500 (Linpack, HPL(High Performance Linpack))
  - Direct Linear Solvers, FLOPS rate
  - Regular Dense Matrices, Continuous Memory Access
  - Computing Performance
- HPCG
  - Preconditioned Iterative Solvers, FLOPS rate
  - Irregular Sparse Matrices derived from FEM Applications with Many “0” Components
    - Irregular/Random Memory Access,
    - Closer to “Real” Applications than HPL
  - Performance of Memory, Communications
- Green 500
  - FLOPS/W rate for HPL (TOP500)

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/

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

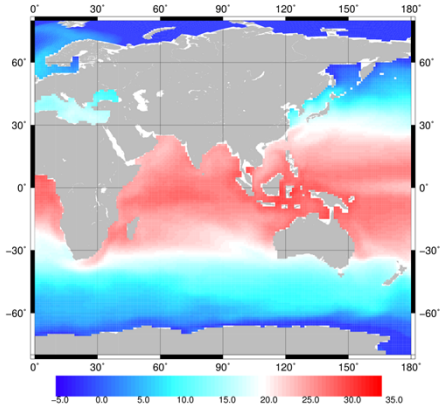
http://www.top500.org/

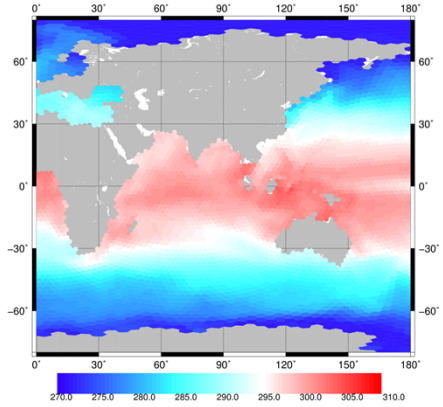
Software of Oakforest-PACS

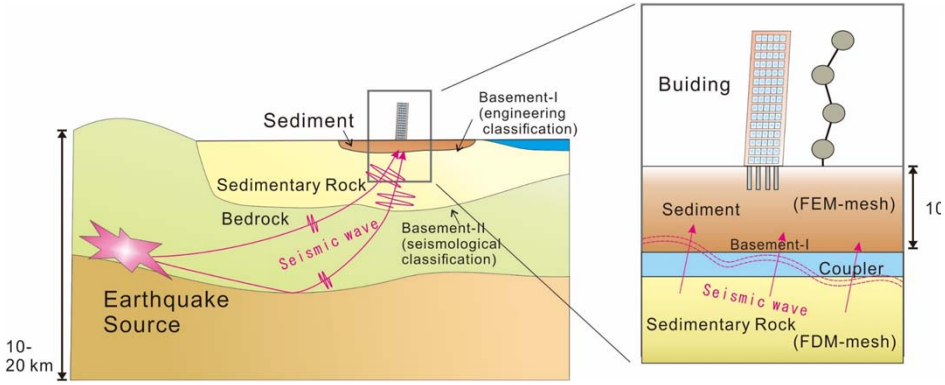
- OS: Red Hat Enterprise Linux (Login nodes), CentOS or McKernel (Compute nodes, dynamically switchable)
  - McKernel: OS for many-core CPU developed by RIKEN AICS
    - Ultra-lightweight OS compared with Linux, no background noise to user program
    - Expected to be installed to post-K computer
- Compiler: 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

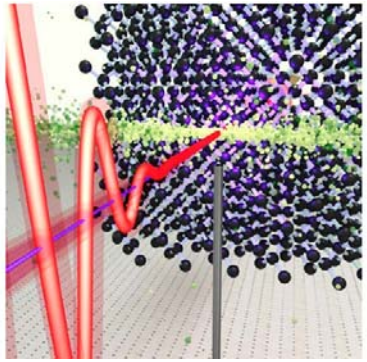
Applications on Oakforest-PACS

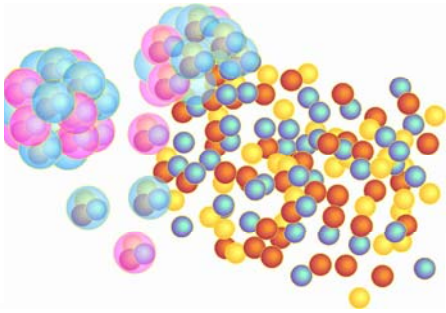
- ARTED: Ab-initio Real Time Electron Dynamics simulator
- Lattice QCD: Quantum Chrono Dynamics
- NICAM & COCO: Atmosphere & Ocean Coupling
- GAMERA/GOJIRA: Earthquake Simulations
- Seism3D: Seismic Wave Propagation







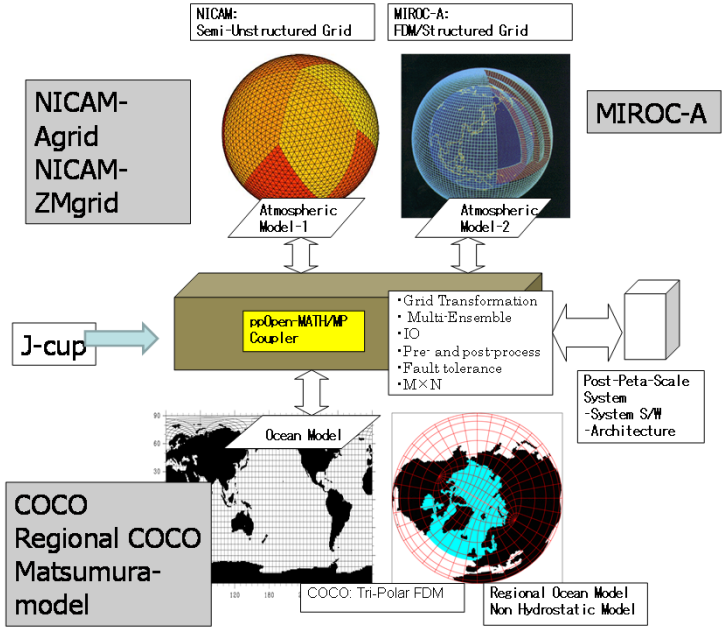


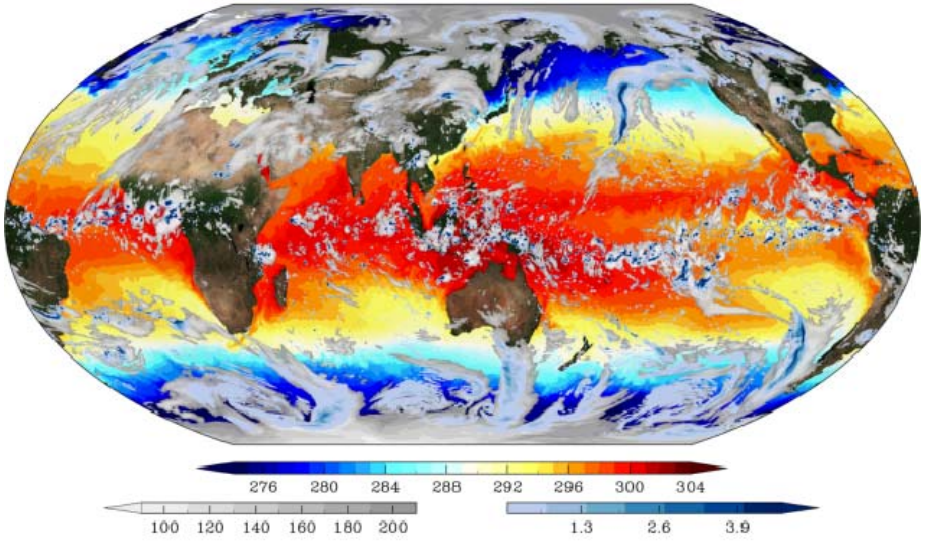


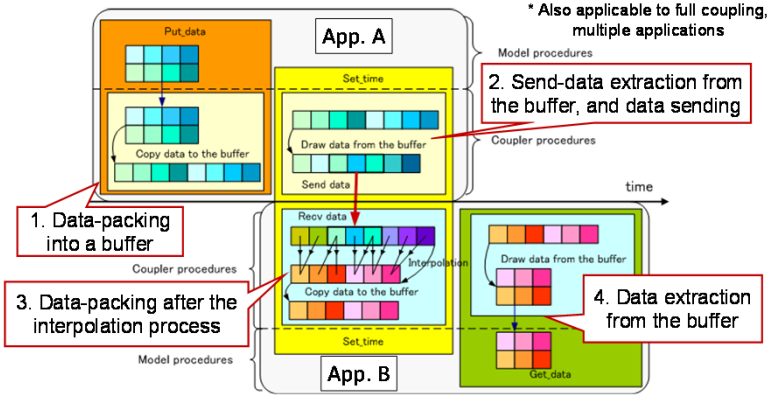
Atmosphere-Ocean Coupling

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]

Schedule for Operations

JCAHPC

- NO charge until the end of March 2017
- There are no explicit partition: All resources are shared
  - ✓in order to make full use of 8,208 nodes/25 PF
- Various Types of Program for Usage
  - ✓Each University's Own Program
    - Group Course (up to 2,048 nodes)
    - Personal Course (U.Tokyo only)
    - Industry Use (U.Tokyo only)
    - International Collaboration (Part of Group Course)
  - ✓HPCI
    - 20% of 8,208 node/yr. by "JCAHPC", including industry use
  - ✓JHPCN（学際大規模情報基盤共同利用共同研究拠点）
    - 5% of 8,208 node/yr.: Industry, International (U.Tokyo only)
  - ✓Education (Tutorials, Classes)
  - ✓Large-Scale HPC Challenge: Full Node, once per month for 24-hours, proposal-based

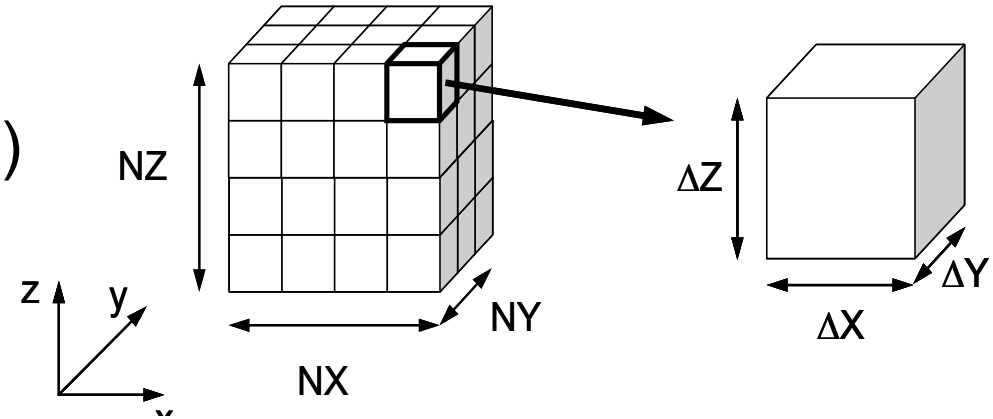
Greeting from Professor Olaf Schenk

- Institute of Computational Science, Universita della Svizzera italiana at Lugano, Switzerland
- Program Director, SIAM Activity Group in Supercomputing (2016-2017)
- **Developer of PARADISO library (direct solver in Intel MKL)**
- KFLOPS/Person for Each Country based on TOP 500 List

- Overview of Information Technology Center (ITC), The University of Tokyo
  - Reedbush
    - Integrated Supercomputer System for Data Analyses & Scientific Simulations
    - Our First System with GPU's
  - Oakforest-PCAS
    - #1 System in Japan !
- **Performance of the New Systems**
  - **3D Poisson Solver by ICCG**
  - Matrix Assembly for 3D FEM

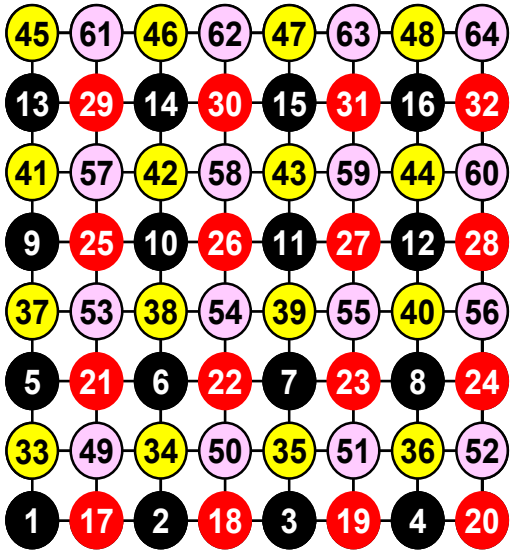
Target Application: Poisson3D-OMP

- Finite Volume Method, Poisson Equations (128^3 cells)
  - FDM-type mesh (7-pt. Stencil), Unstructured data structure
  - SPD matrix
  - ICCG: Data Dependency for Incomplete Cholesky Fact.
- Fortran 90 + OpenMP
- Thread Parallelization by OpenMP: Reordering needed
  - CM-RCM + Coalesced/Sequential
- Storage of Matrix
  - CRS, ELL (Ellpack-Itpack)
- Outer Loops
  - Row-Wise, Column-Wise

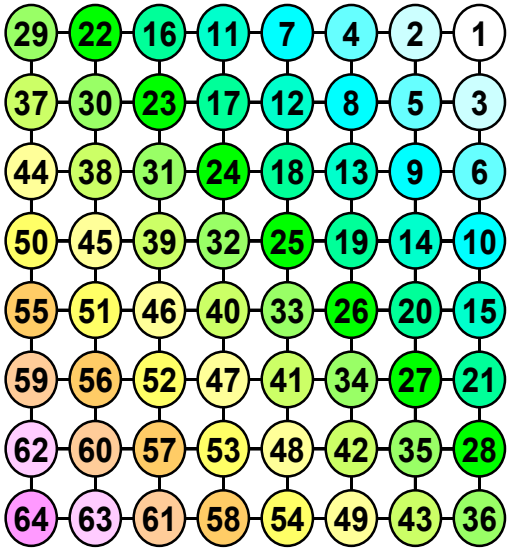


Elimination of Data Dependency in ICCG

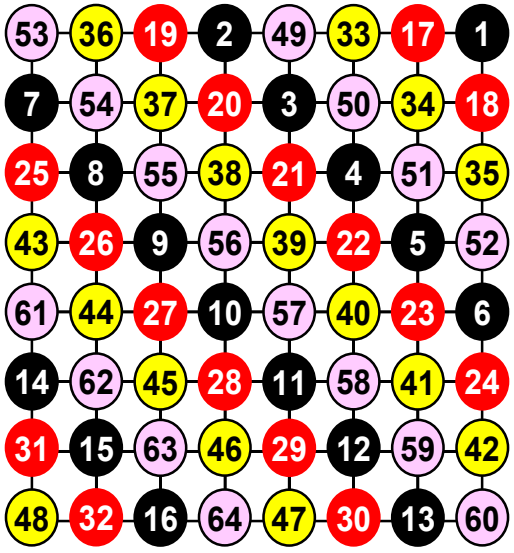
- Coloring + Reordering
- More colors, better convergence/larger sync. overhead
  - MC: good parallel performance, bad convergence
  - RCM: good convergence, many colors
  - **CM-RCM(*k*): combinations of MC + RCM**



MC (Color#=4)
Multicoloring



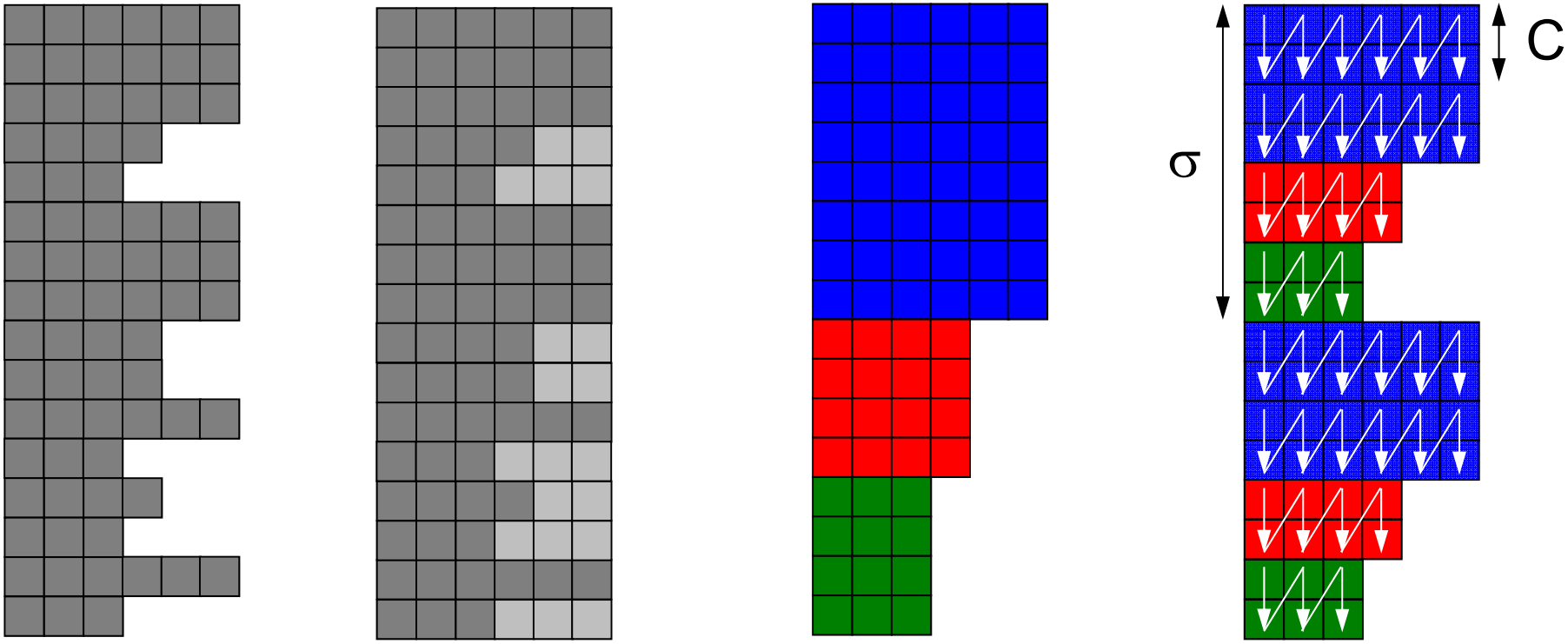
RCM
Reverse Cuthill-Mckee



CM-RCM (4)
Cyclic MC + RCM

Storing Formats for Sparse Matrices

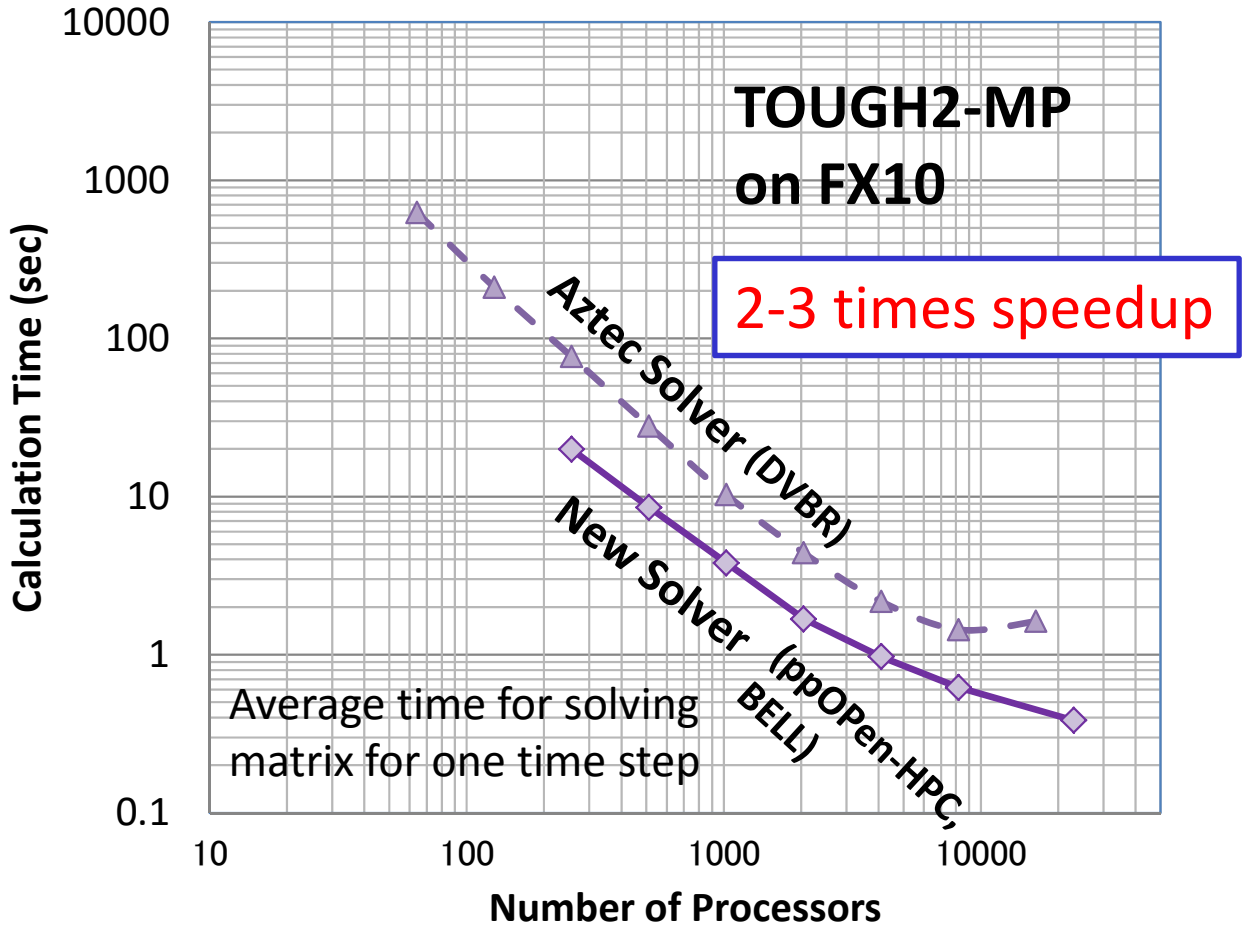
- ELL/Sliced ELL: Excellent Performance by Prefetching
- SELL-C-σ: Good for Vector/SIMD (by FAU/Erlangen)
  - This work is the first example where SELL-C-σ is applied to IC/ILU preconditioning.



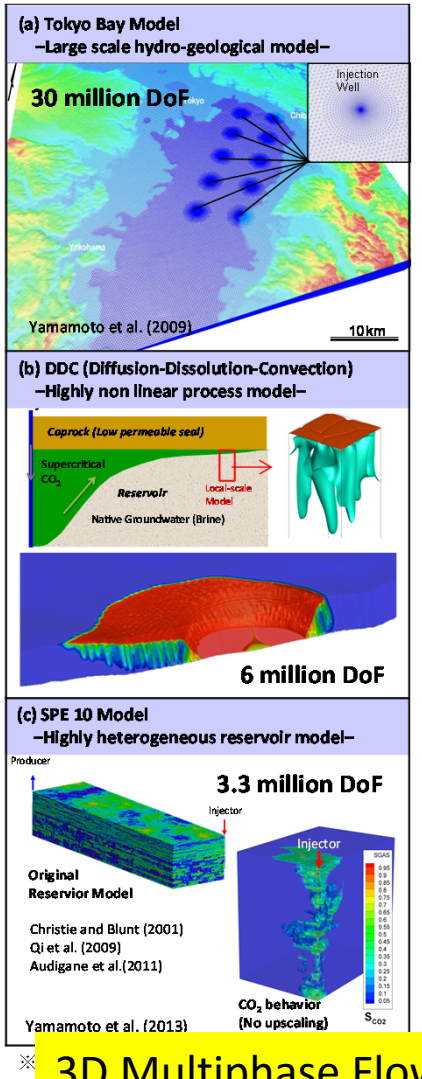
CRS ELL Sliced ELL SELL-C-σ (SELL-2-8)

Simulation of Geologic CO_2 Storage

30 million DoF (10 million grids × 3 DoF/grid node)



[Dr. Hajime Yamamoto, Taisei]



Fujitsu FX10（Oakleaf-FX），30M DOF: 2x-3x improvement

3D Multiphase Flow (Liquid/Gas) + 3D Mass Transfer

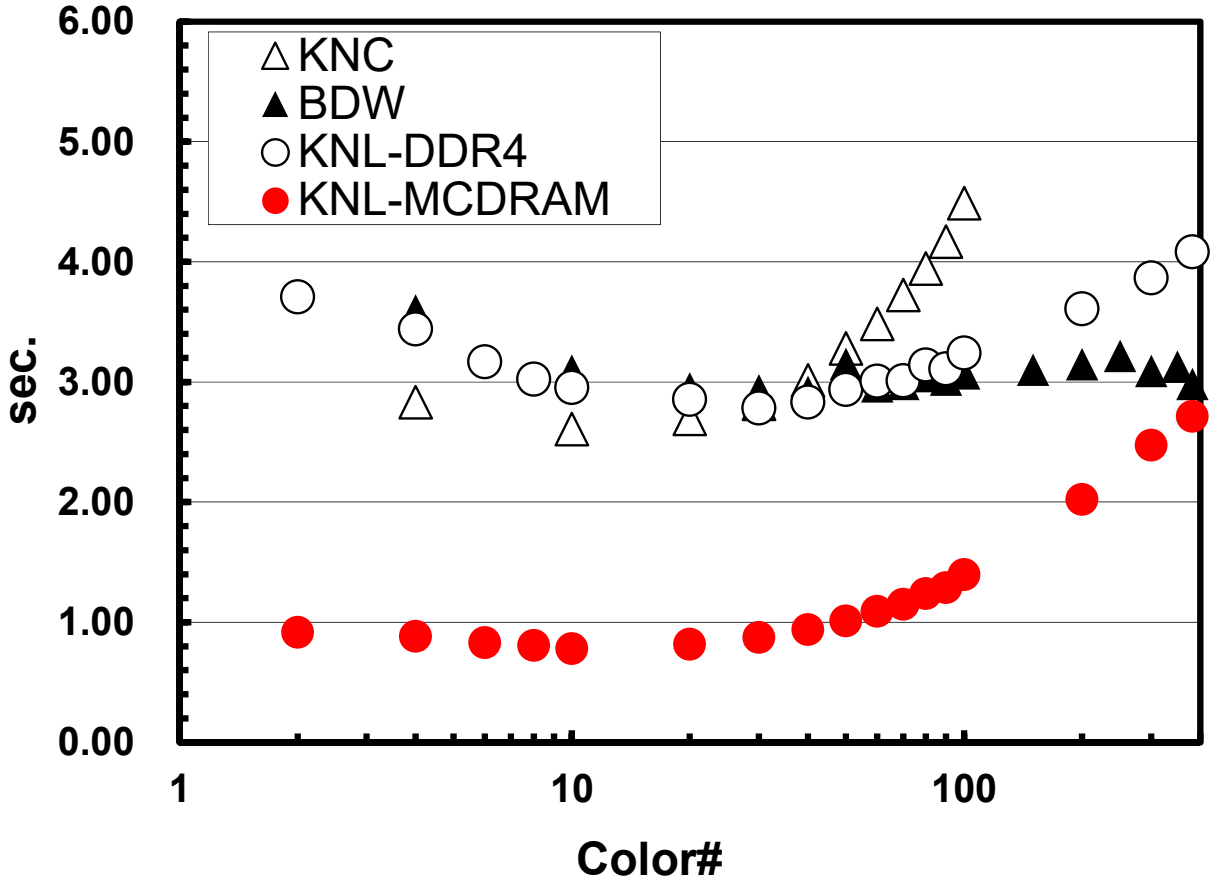
Code Name	KNC	KNL-0	BDW
Architecture	Intel Xeon Phi 5110P (Knights Corner)	Intel Xeon Phi 7210 (Knights Landing)	Intel Xeon E5-2695 v4 (Broadwell-EP)
Frequency (GHz)	1.053	1.30	2.10
Core # (Max Thread #)	60 (240)	64 (256)	18 (18)
Peak Performance (GFLOPS)	1,010.9	2,662.4	604.8
Memory (GB)	8	MCDRAM: 16 DDR4: 96	128
Memory Bandwidth(GB/sec., Stream Triad)	159	MCDRAM: 454 DDR4: 72.5	65.5
Out-of-Order	N	Y	Y
System		OFP-mini	Reedbush-U

Comp. Time for ICCG (Best ELL)

Effects of synchronization overhead are significant on KNL & KNC, if number of colors is larger

Generally, optimum number of color is 10 for KNL/KNC

Down is Good !!



Code Name	KNC	KNL-0	K40	P100
Architecture	Intel Xeon Phi 5110P (Knights Corner)	Intel Xeon Phi 7210 (Knights Landing)	NVIDIA Tesla K40	NVIDIA Tesla P100 (Pascal)
Frequency (GHz)	1.053	1.30	0.745	1.328
Core # (Max Thread #)	60 (240)	64 (256)	2,800	1,792
Peak Performance (GFLOPS)	1,010.9	2,662.4	1,430	4,759
Memory (GB)	8	MCDRAM: 16 DDR4: 96	12	16
Memory Bandwidth(GB/sec., Stream Triad)	159	MCDRAM: 454 DDR4: 72.5	218	530

Comp. Time for ICCG (Best ELL)

Xeon Phi (KNL) and P100 are competitive

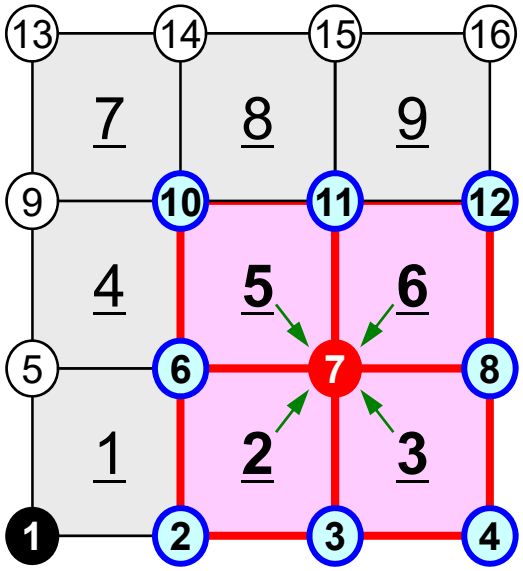
Down is Good !!

Please contact Kengo Nakajima (nakajima(at)cc.u-tokyo.ac.jp) if you need results by P100

- Overview of Information Technology Center (ITC), The University of Tokyo
  - Reedbush
    - Integrated Supercomputer System for Data Analyses & Scientific Simulations
    - Our First System with GPU's
  - Oakforest-PCAS
    - #1 System in Japan !
- **Performance of the New Systems**
  - 3D Poisson Solver by ICCG
  - **Matrix Assembly for 3D FEM**

Matrix Assembly in FEM

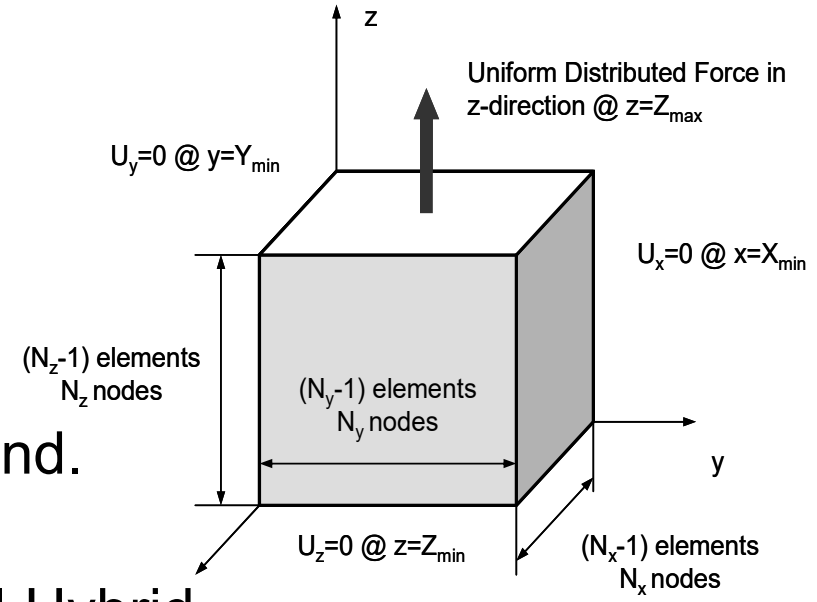
- Integration on each Element: Element Matrix (Dense)
- Accumulation of Elem. Mat: Global Matrix (Sparse)
- Integration/Matrix: Element
- Matrix Computation: Node
  - Components of each Node: Accumulation by Neighboring Elements.
  - Data dependency may happen
  - Coloring needed: Multicoloring
- Hardware
  - Intel Xeon Phi (KNC/KNL)
  - NVIDIA Tesla K40/P100
    - Atomic OP's supported by H/W



[KN, Naruse et al. HPC-152, 2015]

Target: GeoFEM/Cube

- Parallel FEM Code (& Benchmarks)
- 3D-Static-Elastic-Linear (Solid Mechanics)
- Performance of Parallel Preconditioned Iterative Solvers
  - 3D Tri-linear Hexa. Elements
  - SPD matrices: CG solver
  - Fortran90+MPI+OpenMP
  - Distributed Data Structure
  - Reordering for IC/SGS Precond.
    - MC, RCM, CM-RCM
  - MPI, OpenMP, OpenMP/MPI Hybrid
    - only OpenMP case



- Focusing on Matrix Assembling in this Study: 128^3 cubes

Code Name	KNC	KNL-1	K40	P100
Architecture	Intel Xeon Phi 5110P (Knights Corner)	Intel Xeon Phi 7250 (Knights Landing)	NVIDIA Tesla K40	NVIDIA Tesla P100 (Pascal)
Frequency (GHz)	1.053	1.40	0.745	1.328
Core # (Max Thread #)	60 (240)	68 (272)	2,800	1,792
Peak Performance (GFLOPS)	1,010.9	3,046.4	1,430	4,759
Memory (GB)	8	MCDRAM: 16 DDR4: 96	12	16
Memory Bandwidth(GB/sec., Stream Triad)	159	MCDRAM: 490 DDR4: 84.5	218	530
		Real OFP		

Results (Fastest Case)
sec. (GFLOPS)
We need further improvement on KNL

	Coloring + OpenMP/OpenACC	Atomic+ OpenACC	Atomic + CUDA
KNC	0.690 (94.0)		
KNL-1	0.304 (213.4)		
K40	0.675 (96.2)	0.507 (128.1)	0.362 (179.4)
P100			

Please contact Kengo Nakajima (nakajima(at)cc.u-tokyo.ac.jp) if you need results by P100

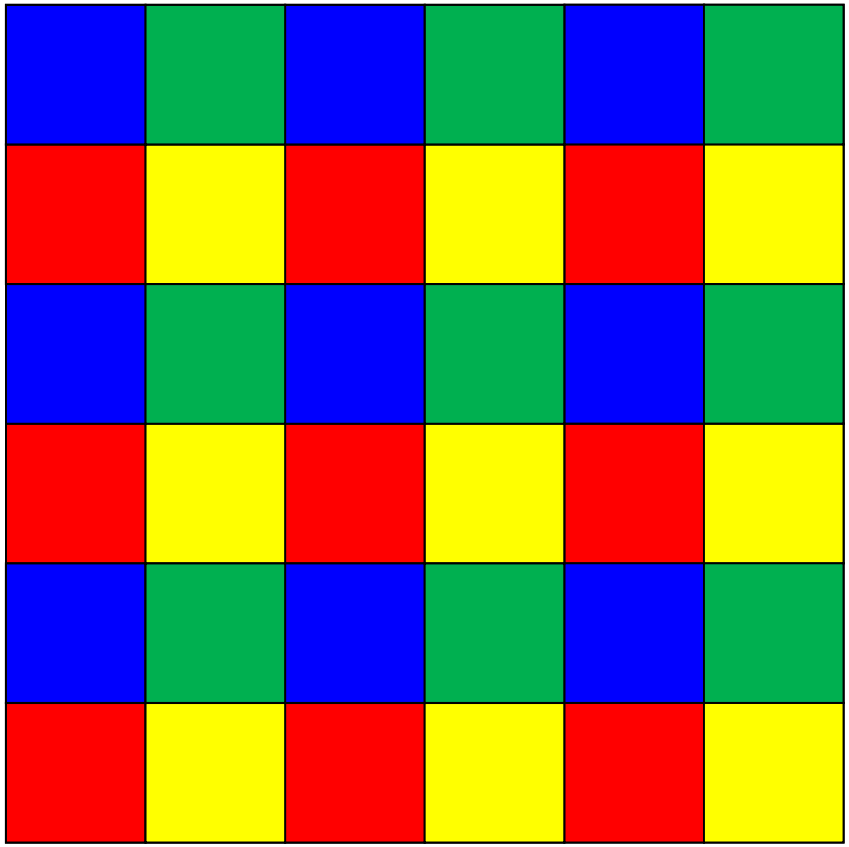
Why “Atomic” is faster ? [Naruse]

4-colors for 2D problem

Elements on same color are computed in parallel

NO cache effect: Each node is “touched” only once during operations for a certain color: this effect of “cache miss” is larger than overhead by Atomic operations

No problems on traditional vector processors



[KN, Naruse et al. HPC-152, 2015]

Summary

- Overview of Information Technology Center (ITC), The University of Tokyo
  - Reedbush
    - Integrated Supercomputer System for Data Analyses & Scientific Simulations, Our First System with GPU's
  - Oakforest-PCAS
    - #1 System in Japan !
- Performance of the New Systems
  - 3D Poisson Solver by ICCG
  - Matrix Assembly for 3D FEM

16th SIAM Conference on Parallel Processing for Scientific Computing (PP18)

March 7-10, 2018, Tokyo, Japan

- Venue
  - Nishiwaseda Campus, Waseda University (in Tokyo)
- Organizing Committee Co-Chairs's
  - Satoshi Matsuoka (Tokyo Institute of Technology, Japan)
  - Kengo Nakajima (The University of Tokyo, Japan)
  - Olaf Schenk (Universita della Svizzera italiana, Switzerland)
- Contact
  - Kengo Nakajima, nakajima(at)cc.u-tokyo.ac.jp
  - http://nkl.cc.u-tokyo.ac.jp/SIAMPP18/

Waseda University, Tokyo, Japan

Northwest of the City of Tokyo