東京大学
THE UNIVERSITY OF TOKYO

Overview of Supercomputer Systems

Supercomputing Division
Information Technology Center
The University of Tokyo

Supercomputers at ITC, U. of Tokyo (April 2016)

Total Users > 2,000

Oakleaf-FX
(Fujitsu PRIMEHPC FX10)

Total Peak performance	: 1.13 PFLOPS
Total number of nodes	: 4800
Total memory	: 150 TB
Peak performance / node	: 236.5 GFLOPS
Main memory per node	: 32 GB
Disk capacity	: 1.1 PB + 2.1 PB

SPARC64 Ixfx 1.84GHz

since April 2012

Oakbridge-FX
(Fujitsu PRIMEHPC FX10)

Total Peak performance	: 136.2 TFLOPS
Total number of nodes	: 576
Total memory	: 18.4 TB
Peak performance / node	: 236.5 GFLOPS
Main memory per node	: 32 GB
Disk capacity	: 147TB + 295TB

SPARC64 Ixfx 1.84GHz

since April 2014
Special System for Long-Term Jobs up to 168 hours

Yayoi
(Hitachi SR16000/M1)

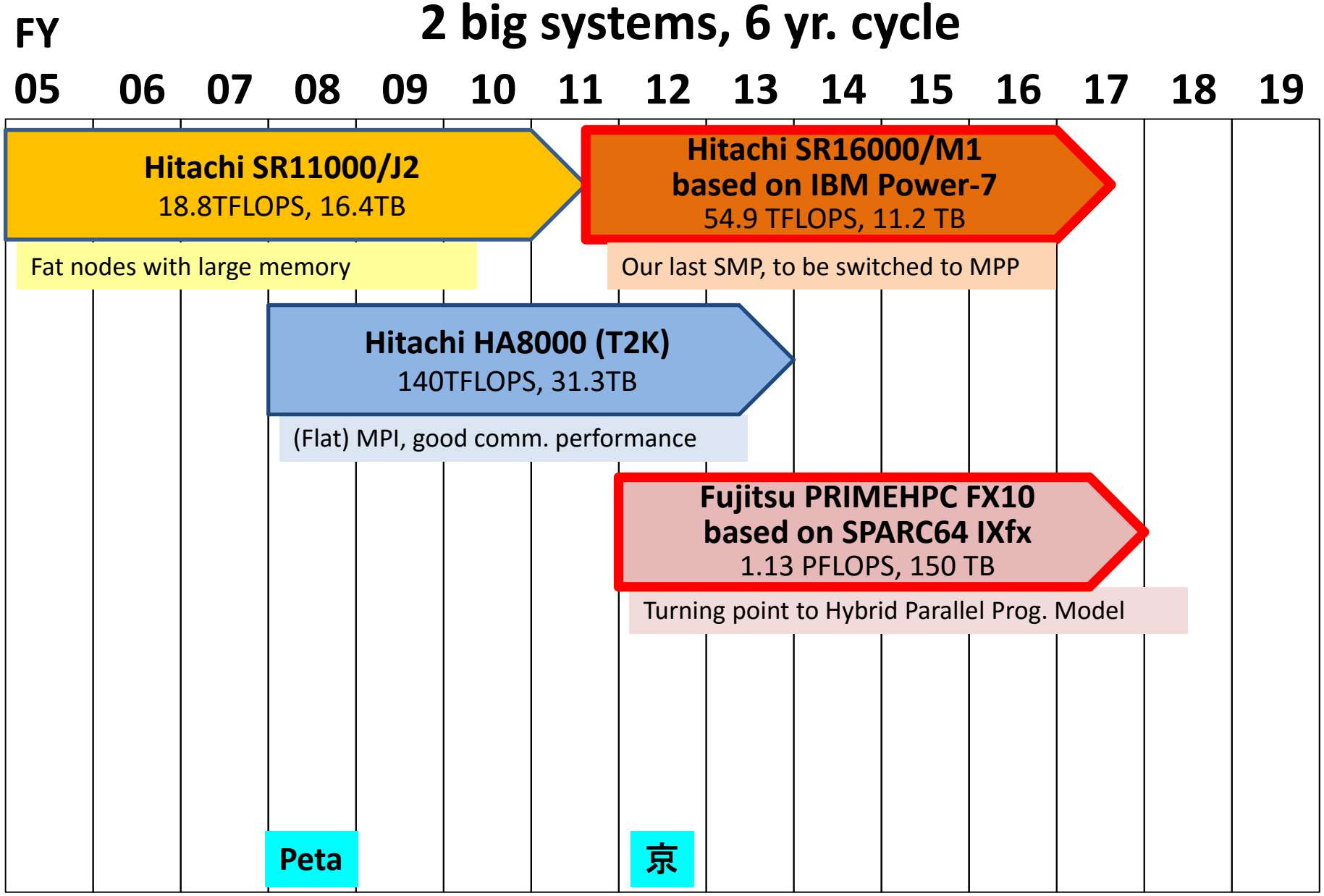
Total Peak performance	: 54.9 TFLOPS
Total number of nodes	: 56
Total memory	: 11200 GB
Peak performance / node	: 980.48 GFLOPS
Main memory per node	: 200 GB
Disk capacity	: 556 TB

IBM POWER 7 3.83GHz

since November 2011

Supercomputers in ITC/U.Tokyo

2 big systems, 6 yr. cycle



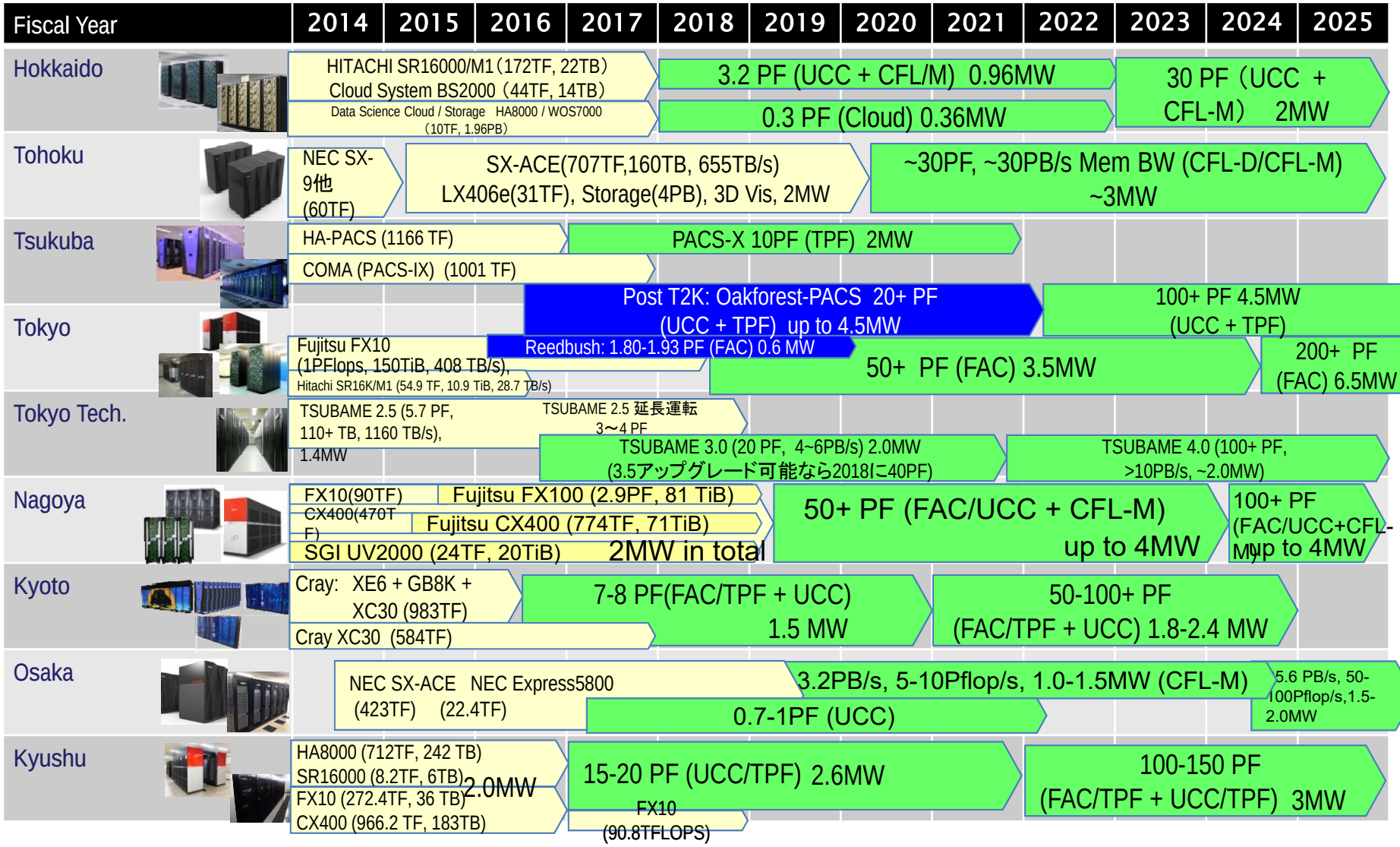
- HPCI
- Supercomputer Systems in SCD/ITC/UT
- Overview of Fujitsu FX10 (Oakleaf-FX)

- Post T2K System + Reedbush

Innovative High Performance Computing Infrastructure (HPCI)

- HPCI Consortium
  - Providing proposals/suggestions to the government and related organizations, operations of infrastructure
  - 38 organizations (Computer Centers, Users)
  - Operations started in Fall 2012
    - https://www.hpci-office.jp/
- Missions
  - Infrastructure (Supercomputers & Distributed Shared Storage System)
    - Seamless access to K, SC's (9 Univ's), & user's machines
  - Promotion of Computational Science
    - Strategic Programs for Innovative Research (SPIRE)
  - R&D for Future Systems (Post-peta/Exascale)

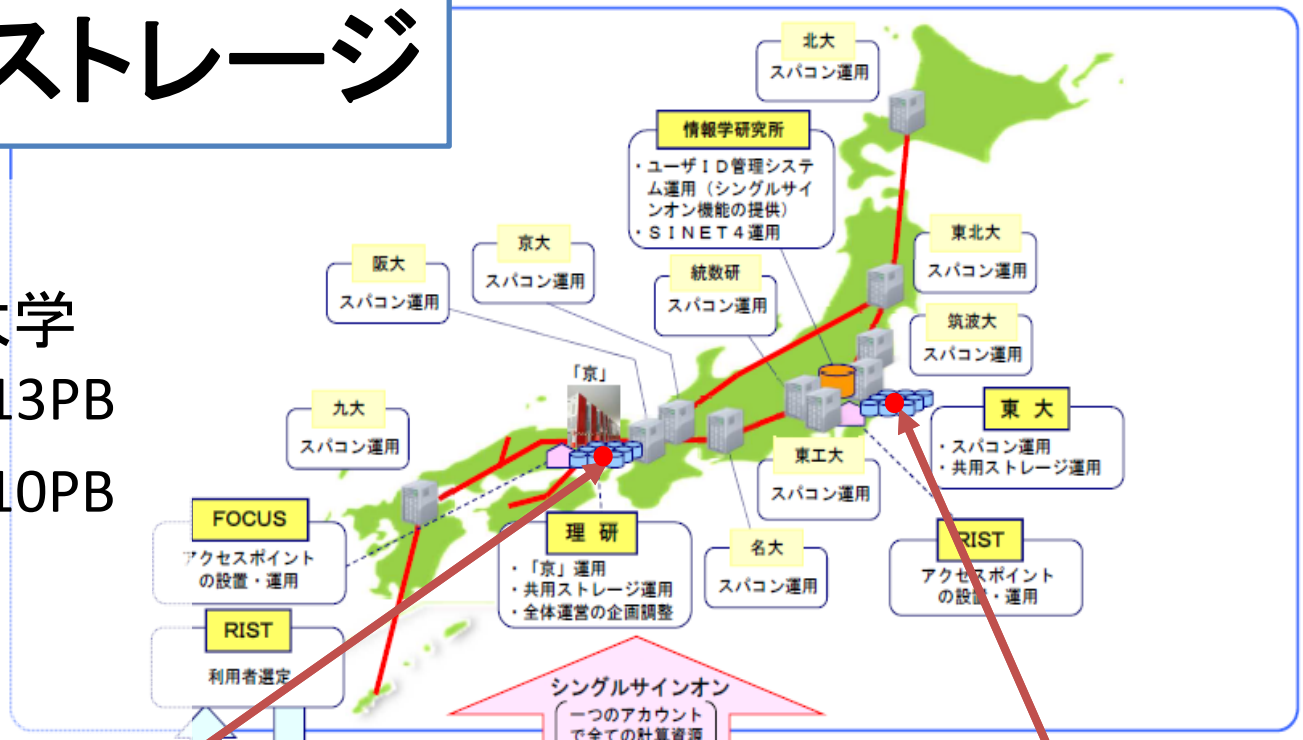
Current Status and Plan (April 2016): Supercomputer Systems in 9 National Universities in Japan



Maximum Power Consumption with A/C

HPCI共用ストレージ

- 文科省委託費
- 東拠点（東京大学柏キャンパス）13PB
- 西拠点（理研) 10PB



ストレージ

- W1 storage
  - Gfarm
  - DDN SFA10000(Total10PB)
  - メタデータサーバ2台
  - データサーバ16台
  - 10GbE ネットワーク
- 60 PB tape archive
- データ解析システム
  - 理論ピーク性能 12.37TFlops
  - 総主記憶容量8.4TB
  - 計算ノード88台 ログインノード2台

理研

- E1 storage
  - Gfarm
  - DDN SFA10000（Total 8PB)
  - データサーバ36台
  - 10GbE ネットワーク
- E2 storage
  - Gfarm
  - DDN SFA10000(Total 5.5PB)
  - メタデータサーバ2台
  - データサーバ8台
  - 10GbE ネットワーク
- 20 PB tape archive

東京大学情報基盤センター

SPIRE/HPCI

Strategic Programs for Innovative Research

- Objectives
  - Scientific results as soon as K computer starts its operation
  - Establishment of several core institutes for comp. science
- Overview
  - Selection of the five strategic research fields which will contribute to finding solutions to scientific and social Issues
    - Field 1: Life science/Drug manufacture
    - Field 2: New material/energy creation
    - Field 3: Global change prediction for disaster prevention/mitigation
    - Field 4: *Mono-zukuri* (Manufacturing technology)
    - Field 5: The origin of matters and the universe
  - A nation wide research group is formed by centering the core organization of each research area designated by MEXT.
  - The groups are to promote R&D using K computer and to construct research structures for their own area

HPCI戦略プログラム
Strategic Programs for Innovative Research

予測する生命科学・医療および創薬基盤

予測医療と革新的創薬

臓器レベルでの疾患を再現する階層統合シミュレーションを実現し、予測医療に貢献。また、標的タンパク質に強く結合する薬の候補化合物の設計を行い、創薬プロセスを加速。

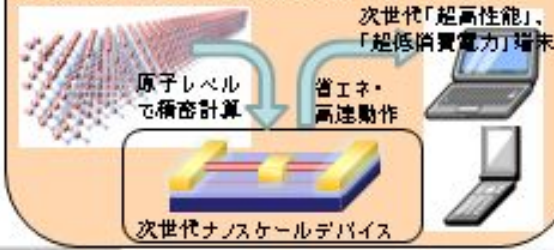
血栓成長による血管閉塞シミュレーション

薬候補のタンパク質への結合シミュレーション

新物質・エネルギー創成

世界に先駆けた次世代デバイスを提唱

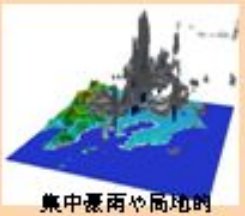
ナノスケールデバイスをまるごとシミュレーションし、機能・材料特性予測を実現することで、次世代デバイスの設計手法を提唱、超高性能・超低消費電力端末等の実現に貢献する。



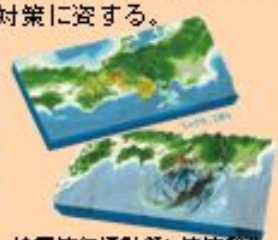
防災・減災に資する地球変動予測

集中豪雨や地震の予測

雲解像モデル、強震動モデル等を駆使して、集中豪雨の位置や地震の被害規模を高精度に予測し、防災・減災対策に資する。



集中豪雨や局地的大雨の予測

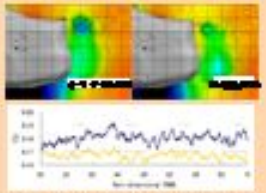


地震波伝播計算と津波発生伝播の連成シミュレーション

次世代ものづくり

設計プロセスの革新

熱流動の物理メカニズム理解に基づいた高度な設計制御技術を確立することで、環境(CO_2 NO_x)と製品性能のバランスを目指した将来の製品競争力強化に資する革新的ものづくりを実現



車体後部周りの超精緻解析による最適形状の究明

非定常空力・振動連成解析による、低空気抵抗、低振動車の開発

物質と宇宙の起源と構造

大質量星の超新星爆発の解明

磁場増幅、ニュートリノ輻射輸送などを考慮した3次元シミュレーションを、次世代スパコンを用いることで世界に先駆けて実行し、大質量星が重力崩壊から超新星爆発に至る過程を解明する。

- HPCI
- **Supercomputer Systems in SCD/ITC/UT**
- Overview of Fujitsu FX10 (Oakleaf-FX)

- Post T2K System + Reedbush

Current Supercomputer Systems University of Tokyo

- Total number of users ~ 2,000 (50% from outside of UT)
- Hitachi HA8000 Cluster System (T2K/Tokyo) (2008.6-2014.3)
  - Cluster based on AMD Quad-Core Opteron (Barcelona)
  - 140.1 TFLOPS
- Hitachi SR16000/M1 (Yayoi) (2011.10-)
  - Power 7 based SMP with 200 GB/node
  - 54.9 TFLOPS
- Fujitsu PRIMEHPC FX10 (Oakleaf-FX) (2012.04-)
  - SPARC64 IXfx
  - Commercial version of K computer
  - 1.13 PFLOPS (1.043 PFLOPS for LINPACK, 75th in Nov.2015)
  - Additional 576 Nodes with 136 TF (Oakbridge-FX, 2014.04-)

Supercomputers at ITC, U. of Tokyo (April 2016)

Total Users > 2,000

Oakleaf-FX
(Fujitsu PRIMEHPC FX10)

Total Peak performance	: 1.13 PFLOPS
Total number of nodes	: 4800
Total memory	: 150 TB
Peak performance / node	: 236.5 GFLOPS
Main memory per node	: 32 GB
Disk capacity	: 1.1 PB + 2.1 PB

SPARC64 Ixfx 1.84GHz

since April 2012

Oakbridge-FX
(Fujitsu PRIMEHPC FX10)

Total Peak performance	: 136.2 TFLOPS
Total number of nodes	: 576
Total memory	: 18.4 TB
Peak performance / node	: 236.5 GFLOPS
Main memory per node	: 32 GB
Disk capacity	: 147TB + 295TB

SPARC64 Ixfx 1.84GHz

since April 2014
Special System for Long-Term Jobs up to 168 hours

Yayoi
(Hitachi SR16000/M1)

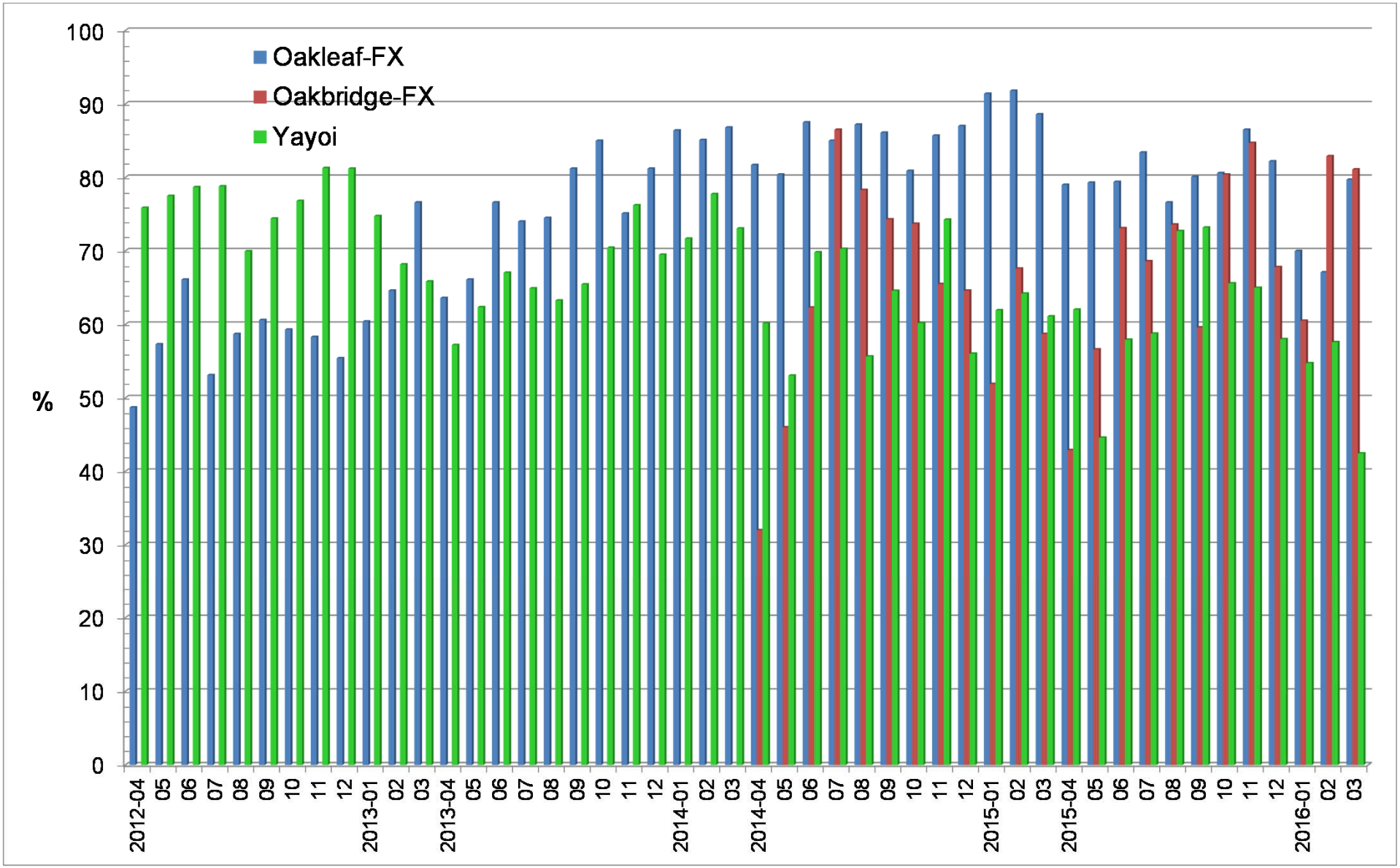
Total Peak performance	: 54.9 TFLOPS
Total number of nodes	: 56
Total memory	: 11200 GB
Peak performance / node	: 980.48 GFLOPS
Main memory per node	: 200 GB
Disk capacity	: 556 TB

IBM POWER 7 3.83GHz

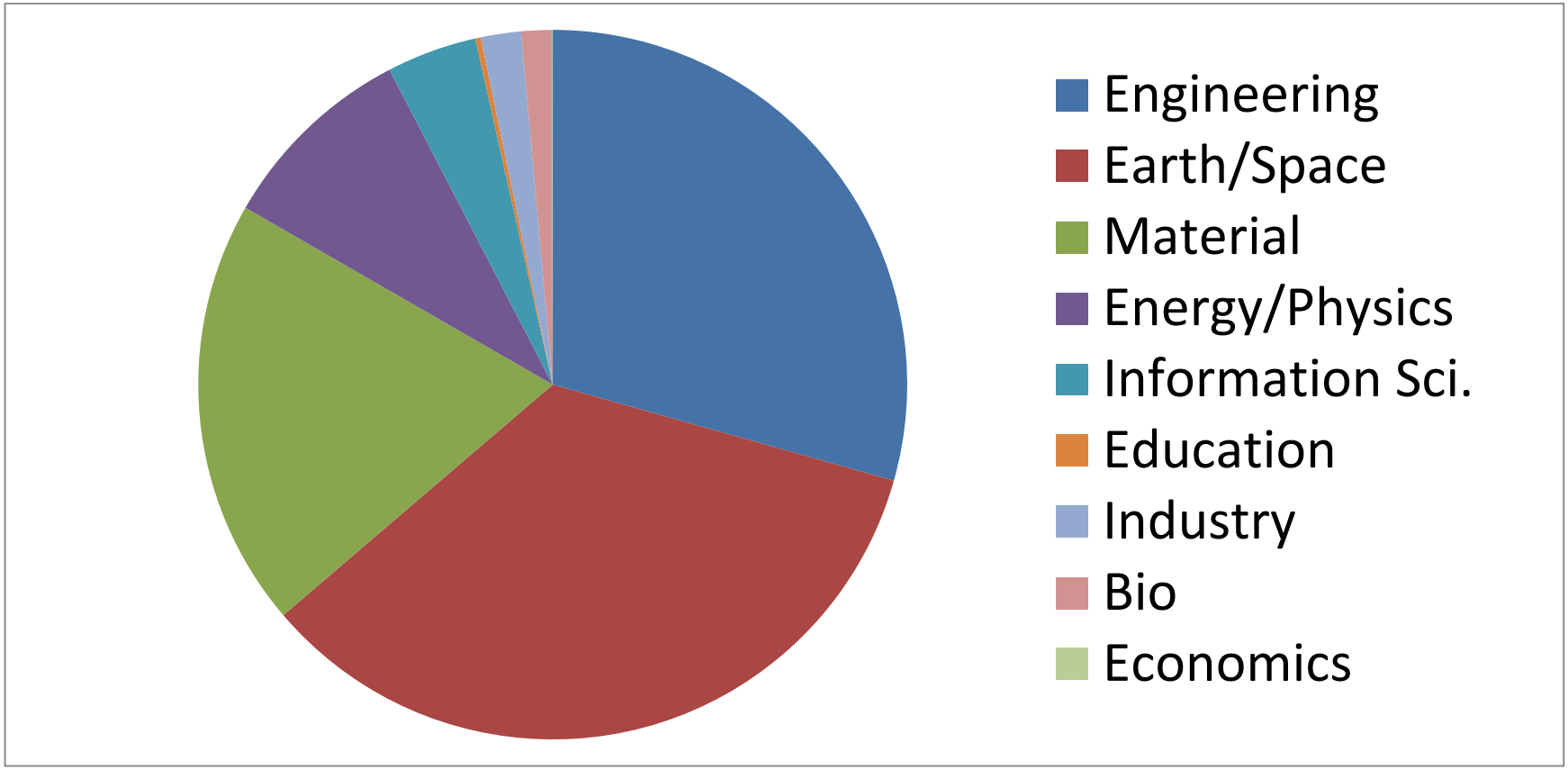
since November 2011

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 (FX10)

- 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
- Currently Oakleaf-FX is open for industry
  - 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 (FX10)

- 2-Day “Hands-on” Tutorials for Parallel Programming by Faculty Members of SCD/ITC (Free)
  - Fundamental MPI (3 times per year)
  - Advanced MPI (2 times per year)
  - OpenMP for Multicore Architectures (2 times per year)
  - 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
  - Proposal-based
  - Not limited to Classes of the University of Tokyo, 2-3 of 10
- RIKEN AICS Summer/Spring School (2011~)

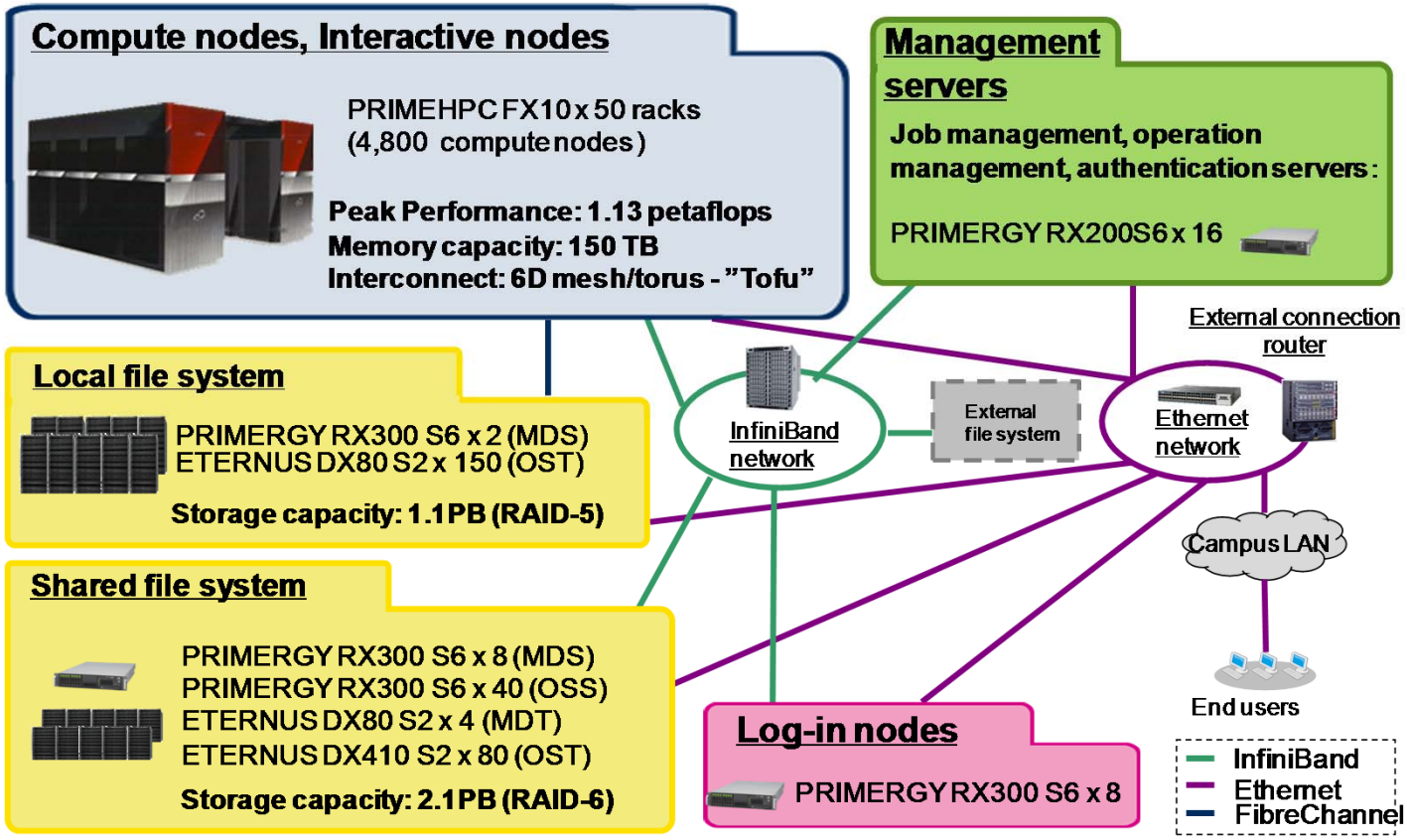
- HPCI
- Supercomputer Systems in SCD/ITC/UT
- **Overview of Fujitsu FX10 (Oakleaf-FX)**

- Post T2K System + Reedbush

Features of FX10 (Oakleaf-FX)

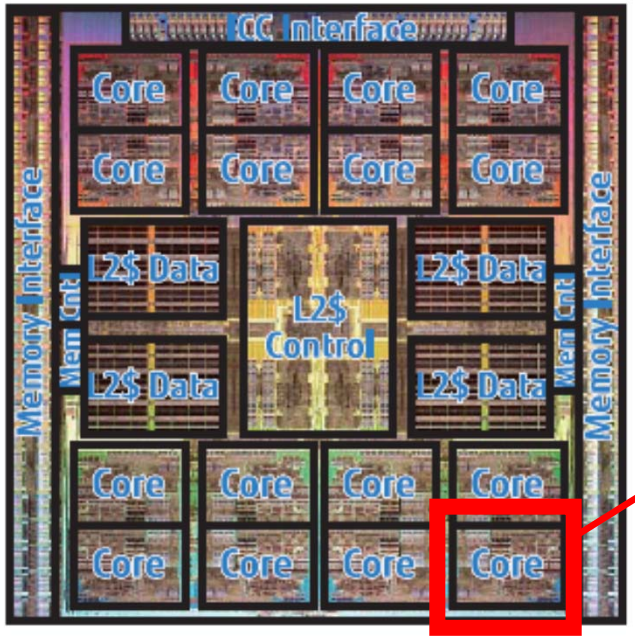
- Well-Balanced System
  - 1.13 PFLOPS for Peak Performance
  - Max. Power Consumption < 1.40 MW
    - < 2.00MW including A/C
- 6-Dim. Mesh/Torus Interconnect
  - Highly Scalable Tofu Interconnect
  - 5.0x2 GB/sec/link, 6 TB/sec for Bi-Section Bandwidth
- High-Performance File System
  - FEFS (Fujitsu 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

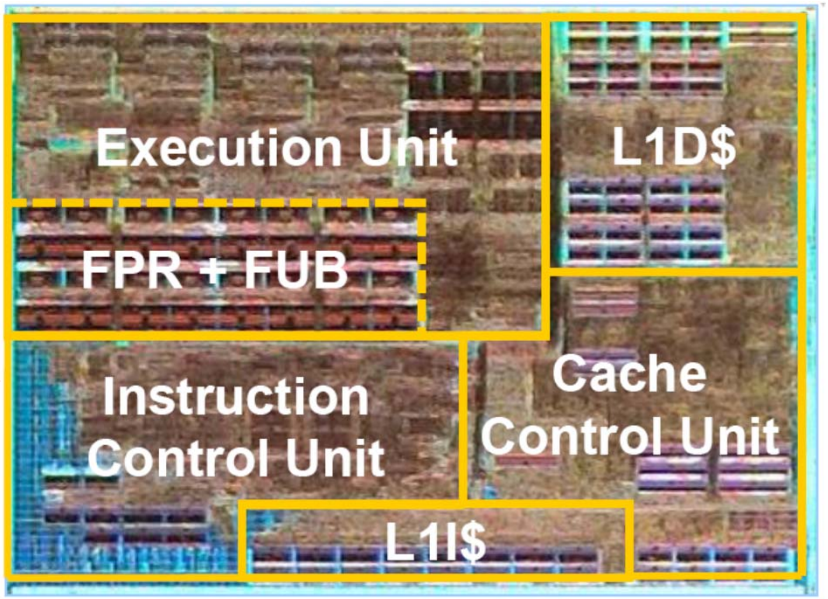
FX10 System (Oakleaf-FX)



- 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

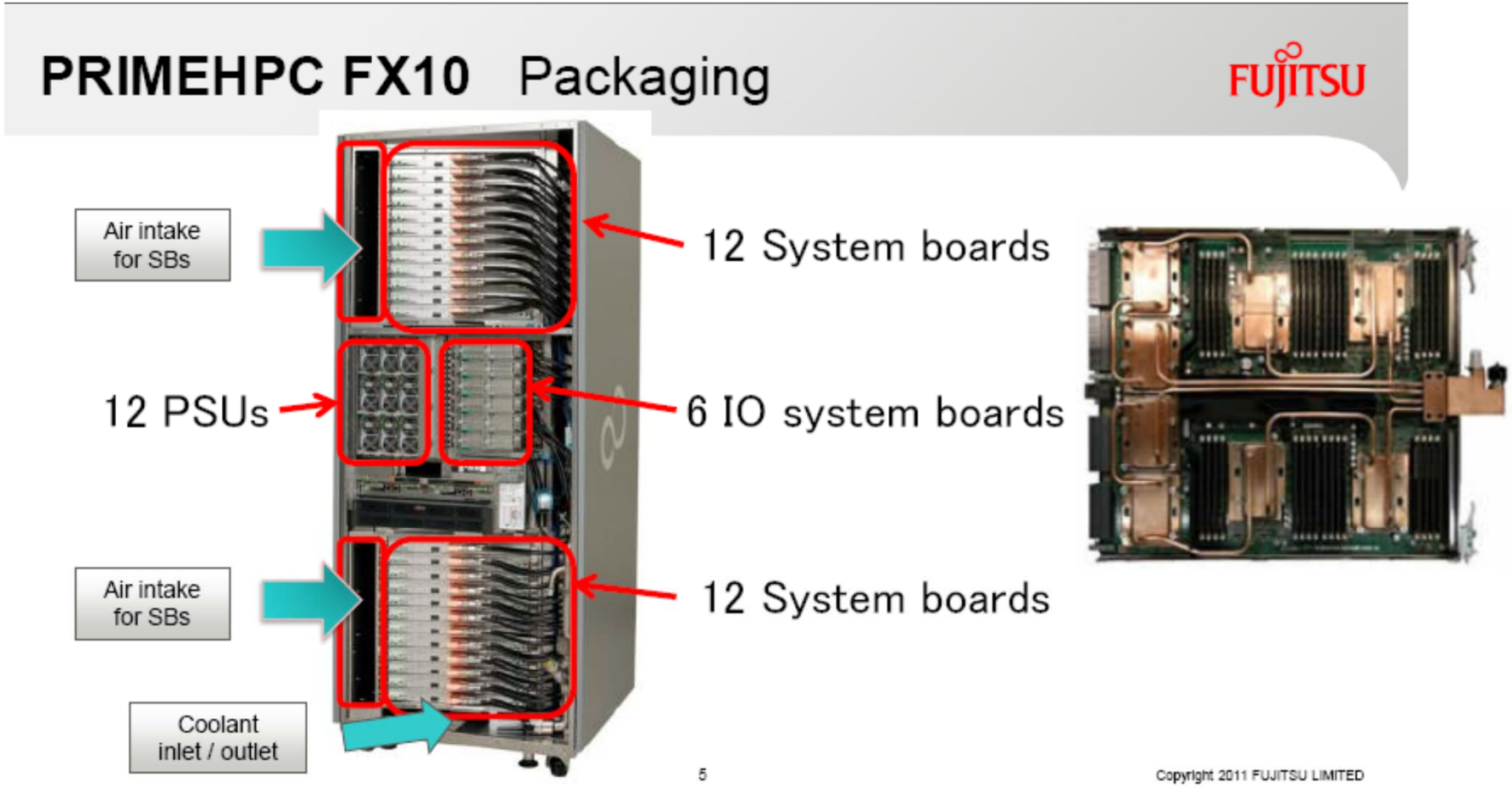




CPU	SPARC64™ IXfx 1.848 GHz	SPARC64™ VIIIfx 2.000 GHz
Number of Cores/Node	16	8
Size of L2 Cache/Node	12 MB	6 MB
Peak Performance/Node	236.5 GFLOPS	128.0 GFLOPS
Memory/Node	32 GB	16 GB
Memory Bandwidth/Node	85 GB/sec (DDR3-1333)	64 GB/sec (DDR3-1000)

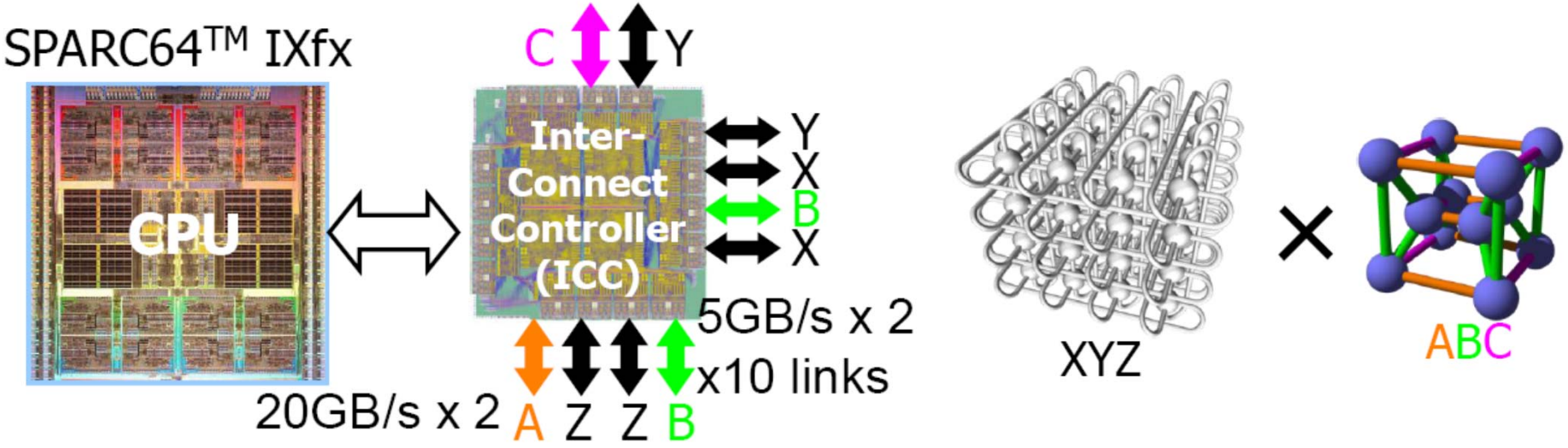
Racks

- 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.)

- HPCI
- Supercomputer Systems in SCD/ITC/UT
- Overview of Fujitsu FX10 (Oakleaf-FX)

- **Post T2K System + Reedbush**

Post T2K System: Oakforest-PACS

http://jcahpc.jp/pr/pr-en-20160510.html

- 25 PFLOPS, Fall 2016: by Fujitsu
- 8,208 Intel Xeon/Phi (KNL)
  - Full operation starts on December 1, 2016
- Joint Center for Advanced High Performance Computing (JCAHPC, http://jcahpc.jp/)
  - 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

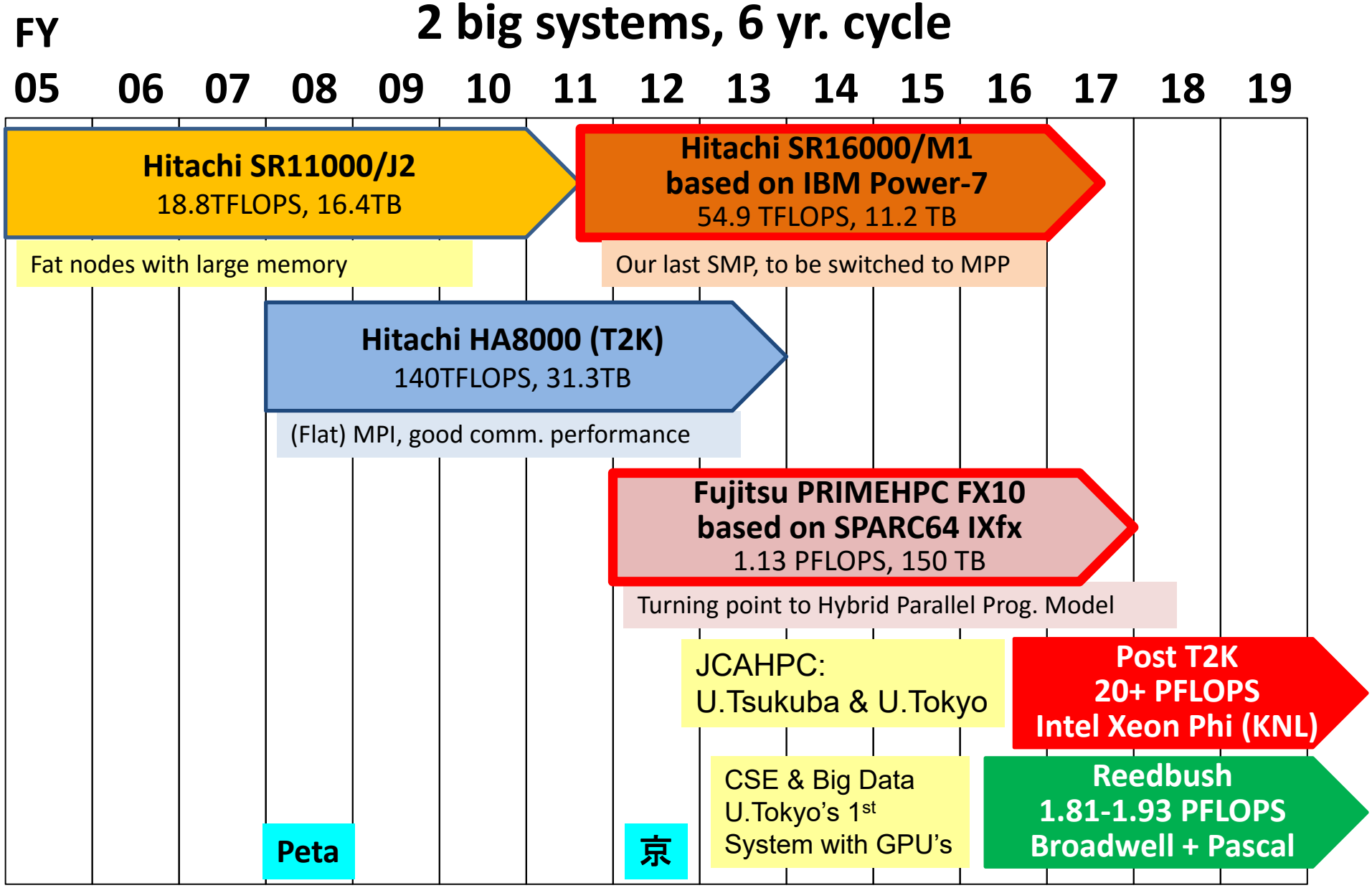
東京大学 THE UNIVERSITY OF TOKYO　JCAHPC　筑波大学 University of Tsukuba

Integrated Supercomputer System for Data Analyses & Scientific Simulations: MPT2K（Mini Post T2K）

- Two Types of Compute Nodes
- Compute Nodes (1) : CPU Only
  - Each Node: 1.2TF, 256GB, 150GB/sec, Total: 400TF+
- Compute Nodes (2)：CPU+GPU: Our First System with GPU
  - Architecture of CPU could be different that in (1)
  - GPU: 4TF, 16GB, 1TB/sec, Total: 960TF+
- File System
  - Shared File System: 4PB, 75GB/sec
  - High-Speed File Cache System: 150TB, 200GB/sec
- Air Cooling, 500kVA

Supercomputers in ITC/U.Tokyo

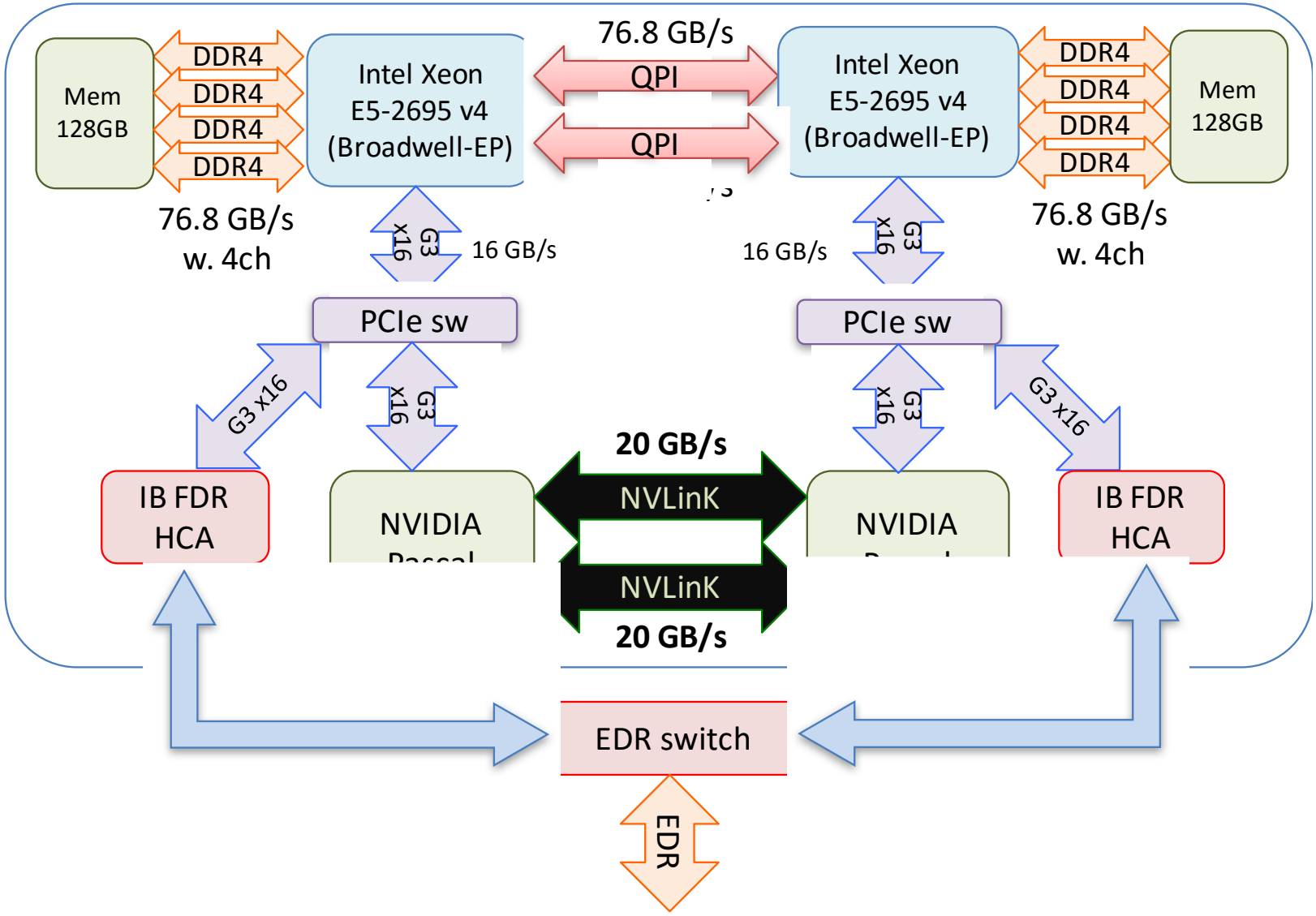
2 big systems, 6 yr. cycle



Reedbush (1/2): Actual Numbers

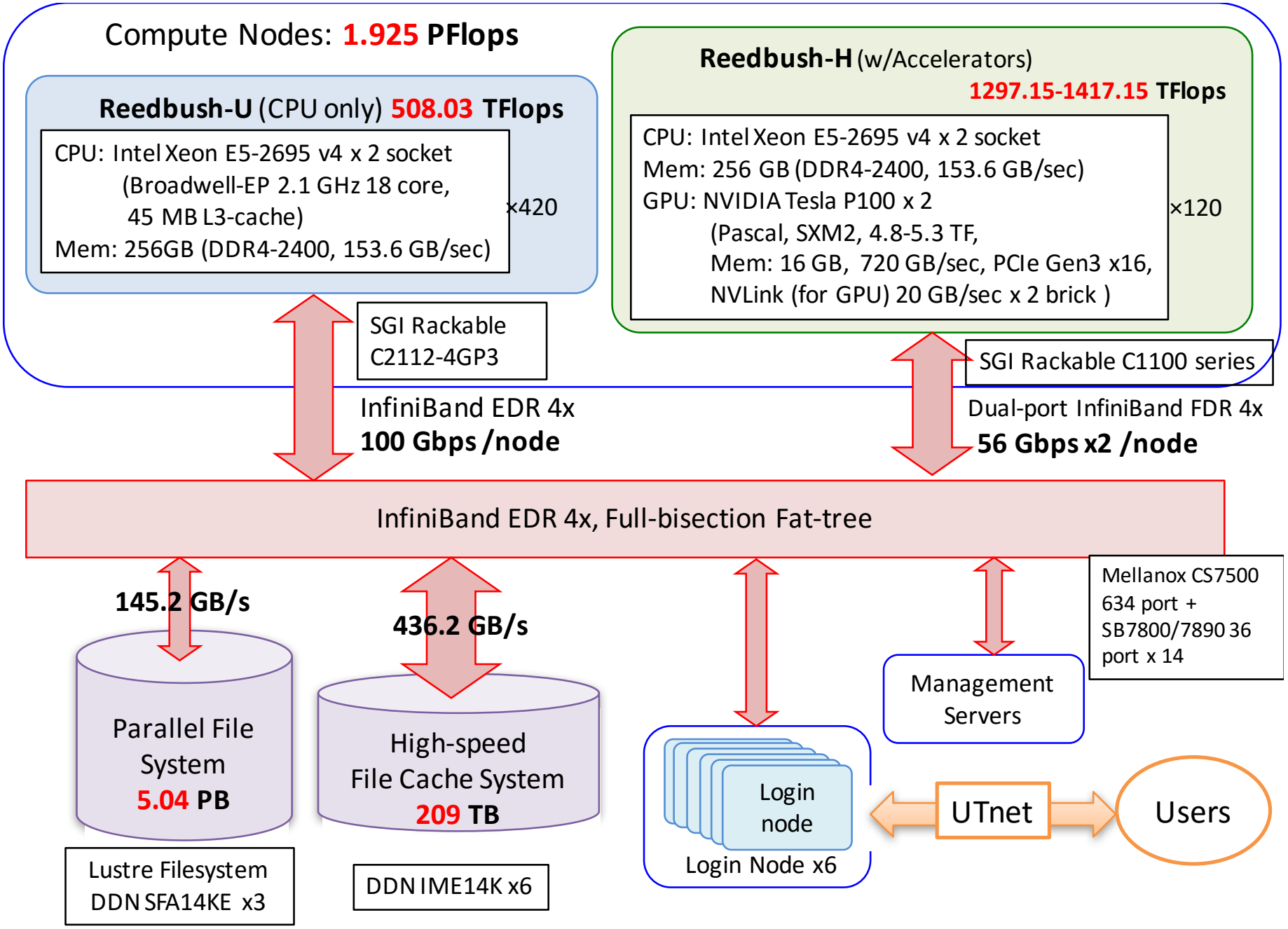
- SGI was awarded (Mar. 22, 2016)
- Compute Nodes (CPU only): Reedbush-U
  - Intel Xeon E5-2695v4 (Broadwell-EP, 2.1GHz 18core) x 2socket (1.210 TF), 256 GiB (153.6GB/sec)
  - InfiniBand EDR, Full bisection Fat-tree
  - Total System: 420 nodes, 508.0 TF
- Compute Nodes (with Accelerators): Reedbush-H
  - 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



Reedbush (Mini PostT2K) (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


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