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

(retired, March 2014)

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	

T2K-Todai （Hitachi HA8000-tc/RS425 ）	
Total Peak performance	: 140 TFLOPS
Total number of nodes	: 952
Total memory	: 32000 GB
Peak performance / node	: 147.2 GFLOPS
Main memory per node	: 32 GB, 128 GB
Disk capacity	: 1 PB
AMD Quad Core Opteron 2.3GHz	

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	

“Oakbridge-fx” with 576 nodes installed in April 2014 (separated) (136TF)

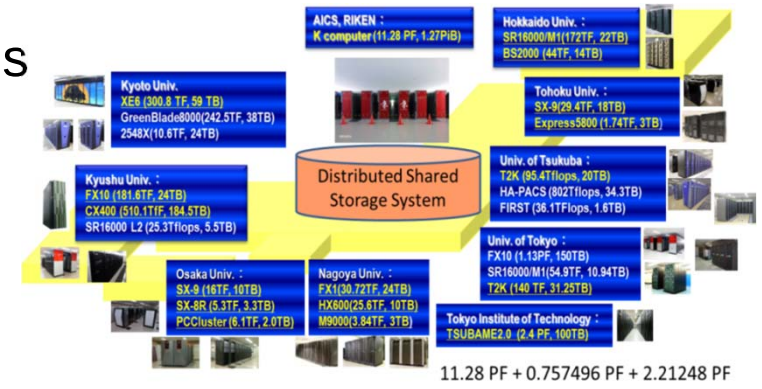
Total Users > 2,000

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

- Post T2K System

Innovative High Performance Computing Infrastructure (HPCI)

- HPCI
  - Seamless access to K computer, supercomputers, and user's machines
  - Distributed shared storage system
- HPCI Consortium
  - Providing proposals/suggestions to the government and related organizations
    - Plan and operation of HPCI system
    - Promotion of computational sciences
    - Future supercomputing
  - 38 organizations
  - Operations started in Fall 2012
    - https://www.hpci-office.jp/



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。

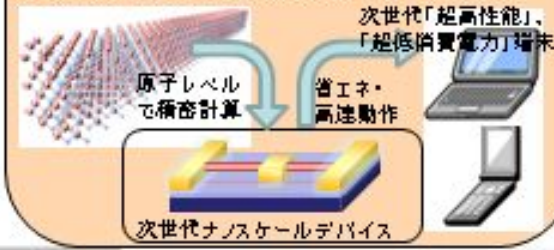
血栓成長による血管閉塞シミュレーション

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

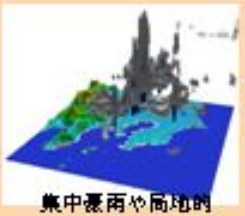
ナノスケールデバイスをまるごとシミュレーションし、機能・材料特性予測を実現することで、次世代デバイスの設計手法を提唱、超高性能・超低消費電力端末等の実現に貢献する。



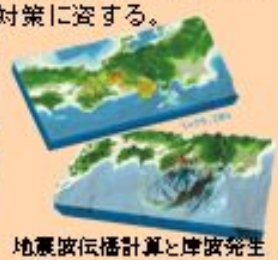
防災・減災に資する地球変動予測

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

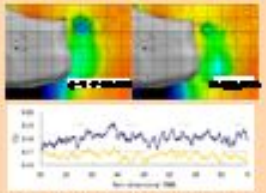


地震波伝播計算と津波発生伝播の連成シミュレーション

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。

＞22 PFLOPS

November 2013

AICS, RIKEN：
K computer (11.28 PF, 1.27PiB)

Hokkaido Univ.：
SR16000/M1(172TF, 22TB)
BS2000 (44TF, 14TB)

Tohoku Univ.：
SX-9(29.4TF, 18TB)
Express5800 (1.74TF),
3TB)

Kyoto Univ.
XE6 (300.8 TF, 59 TB)
GreenBlade8000(242.5TF, 38TB)
2548X(10.6TF, 24TB)

Univ. of Tsukuba：
T2K (95.4Tflops, 20TB)
HA-PACS (802Tflops,
34.3TB)
FIRST (36.1TFlops, 1.6TB)

Osaka Univ.：
SX-9 (16TF, 10TB)
SX-8R (5.3TF, 3.3TB)
PCCluster (22.7 TF, 4.6TB)

Univ. of Tokyo：
FX10 (1.13PF, 150TB)
SR16000/M1(54.9TF, 10.94TB)
T2K (75.36TF,16TB/140 TF, 31.25TB)
EastHubPCCluster(10TF,5.71TB/13TF,8.1
GPU Cluster(CPU 4.5TF, GPU 16.48TF,1.5
WestHubPCCluster(12.37TF,8.25TB)
RENKEI-VPE:VM Hosting

Kyushu Univ.：
FX10 (181.6TF, 24TB)
CX400 (811.9TF, 184.5TB)
HA8000-tc (712.5TF, 24.7TB)
SR16000 VM1 (8.19TF, 16TB)

Nagoya Univ.：
FX10 (90.8TF, 12TB)
CX400 (471TF, 43TB)

Tokyo Institute of Technology
：
TSUBAME2.5 (5.79 PF, 150TB)

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

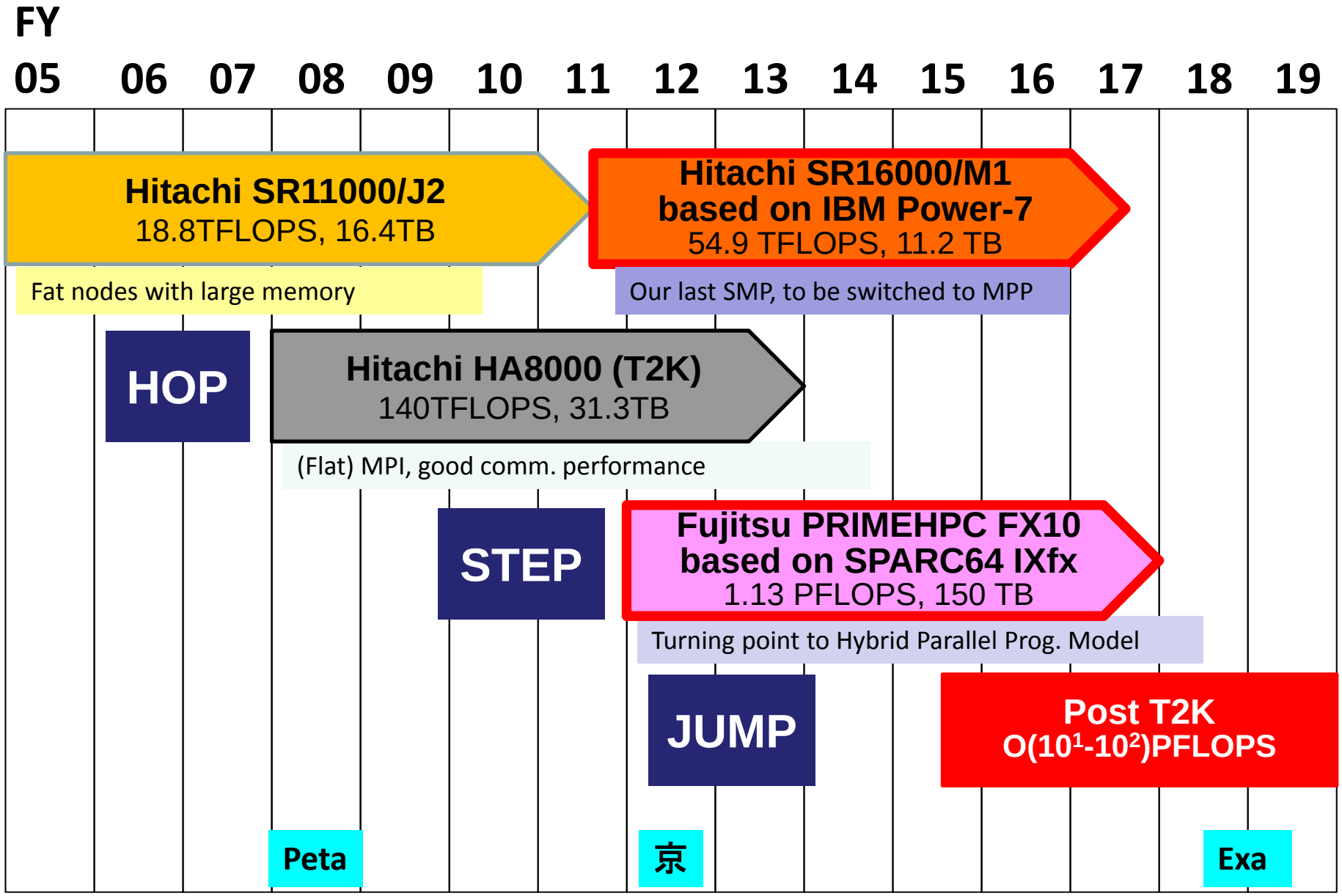
- Post T2K System

Current Supercomputer Systems University of Tokyo

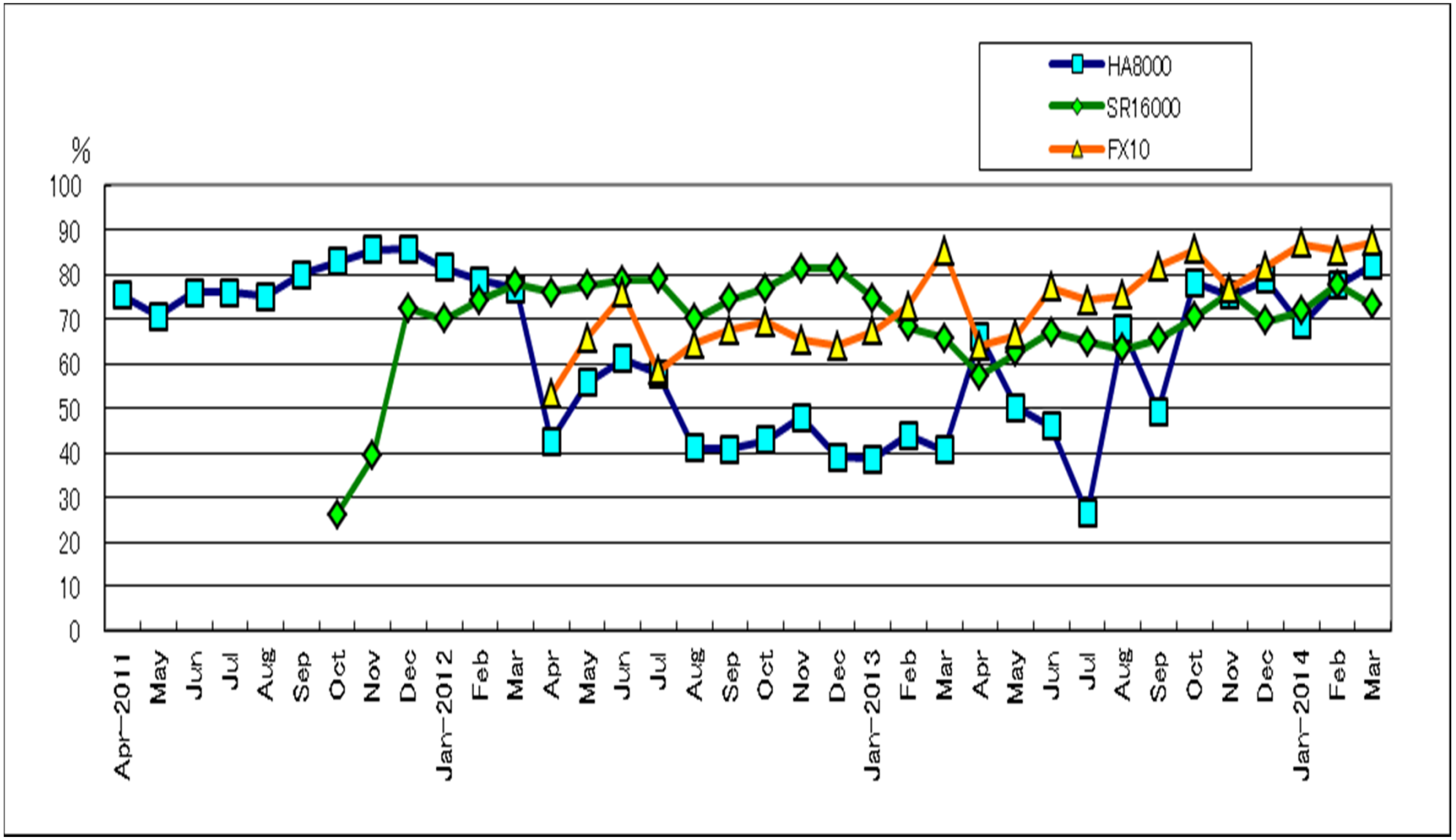
- Total number of users ~ 2,000 (50% from outside of UT)
- Hitachi HA8000 Cluster System (T2K/Tokyo) (2008.6-2014.3)
  - Cluster based on AMD Quad-Core Opteron (Barcelona)
  - 140.1 TFLOPS
- Hitachi SR16000/M1 (Yayoi) (2011.10-)
  - Power 7 based SMP with 200 GB/node
  - 54.9 TFLOPS
- Fujitsu PRIMEHPC FX10 (Oakleaf-FX) (2012.04-)
  - SPARC64 IXfx
  - Commercial version of K computer
  - 1.13 PFLOPS (1.043 PFLOPS for LINPACK, 36th in 43rd TOP500)
  - Additional 576 Nodes with 136 TF (Oakbridge-FX, 2014.04-)

	HA8000 (T2K)	SMP (Yayoi) SR16000/M1	FX10 (Oakleaf-FX) PRIMEHPC FX10
CPU	AMD Quad Core Opteron 2.3GHz	IBM Power7 3.83GHz	FUJITSU SPARC64IXfx 1.8GHz
Total # of core	15232	1792	76800
Total Peak FLOPS	140 TFLOPS	54.9 TFLOPS	1.13 PFLOPS
Total # of nodes	952	56	4800
Total Memory	32 TB	11200 GB	150 TB
# of core / node	16	32	16
Perk FLOPS / node	147.2 GFLOPS	980.5 GFLOPS	236.5 GFLOPS
Memory / node	32 GB, 128 GB	200 GB	32 GB
Network	Myrinet 10G Full-bisection	Hierarchical Full-bisection	Tofu 6D Mesh/Torus
Storage	1 PB	556 TB	1.1PB + 2.1 PB

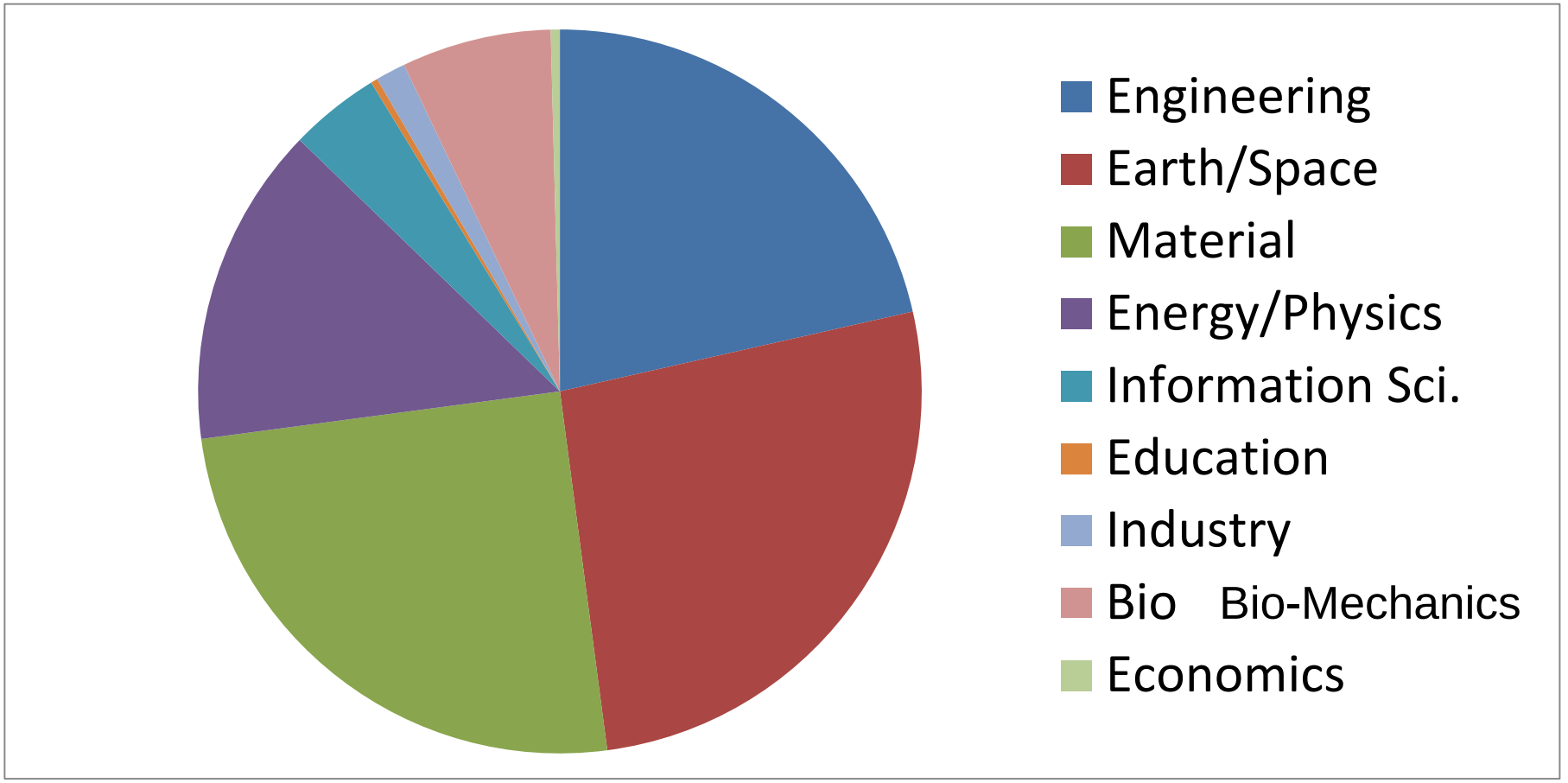
Supercomputers in U.Tokyo



History of Work Ratio



Research Area based on CPU Hours FX10 in FY.2013 (2013.4~2014.3E)



Applications

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)
    - 4 groups (FY.2014) (1 IT, 3 manufacturing), 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~)

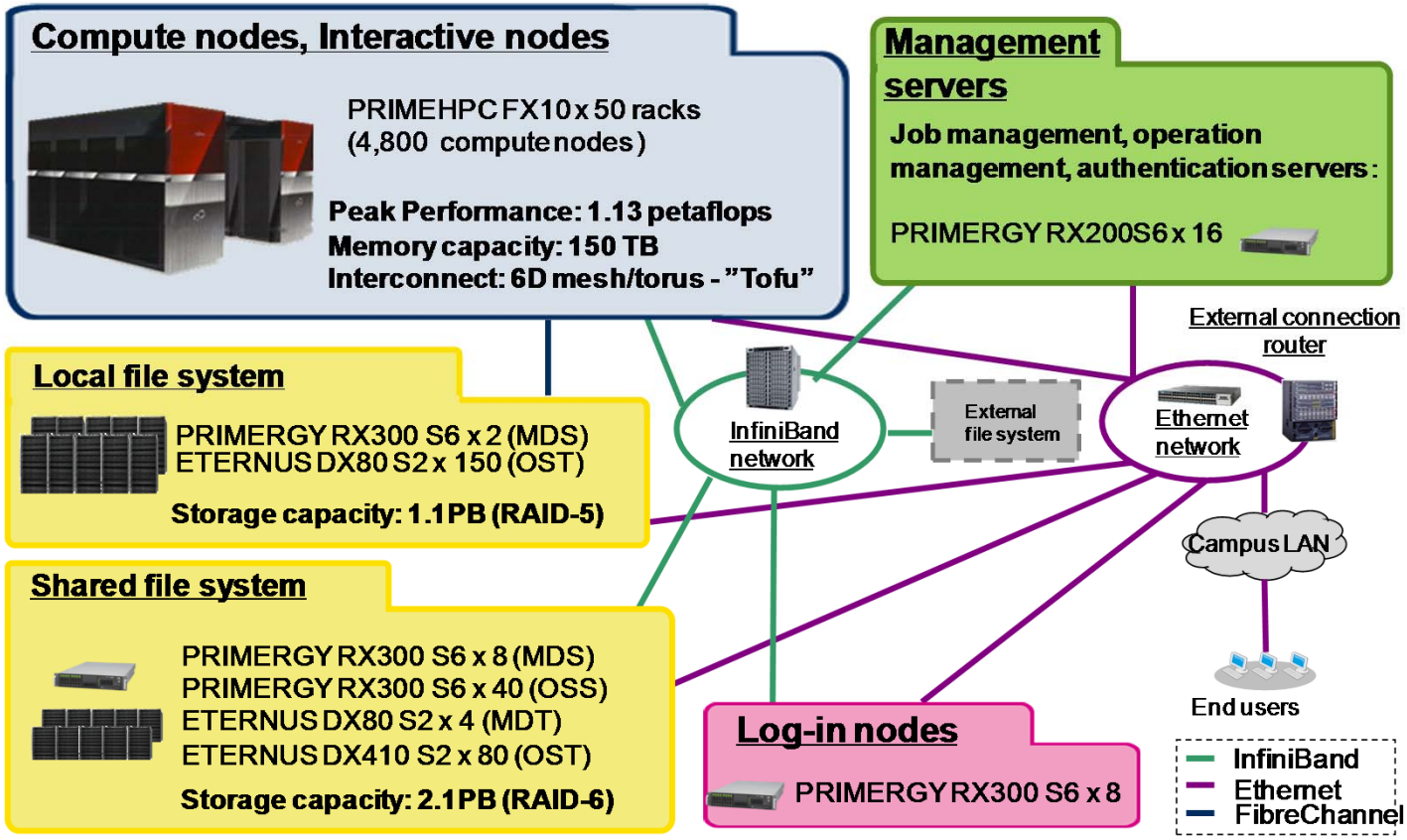
- HPCI
- Supercomputer Systems in SCD/ITC/UT
- **Overview of Fujitsu FX10 (Oakleaf-FX)**

- Post T2K System

Features of FX10 (Oakleaf-FX)

- Well-Balanced System
  - Peak Performance: 1.13 PFLOPS, 398 TB/sec
  - Max. Power Consumption < 1.40 MW (<2.00MW with A/C)
    - Strict Requirement after March 11, 2011
    - 1.043 PFLOPS for Linpack with 1.177 MW (excluding 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 (16 cores/node, K: 8 cores/node) !
- Open-Source Libraries/Applications
- Highly Scalable for both of Flat MPI and Hybrid (OpenMP + MPI)

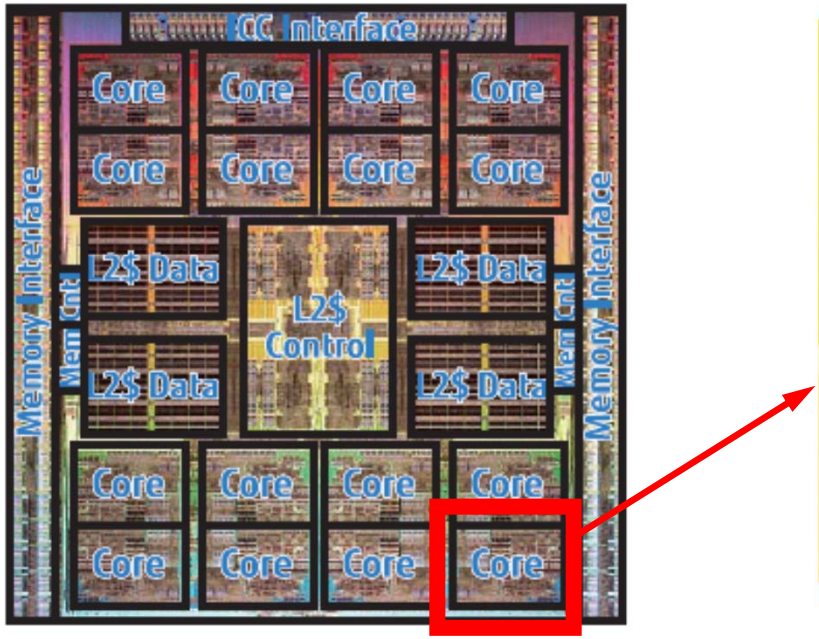
FX10 System (Oakleaf-FX)

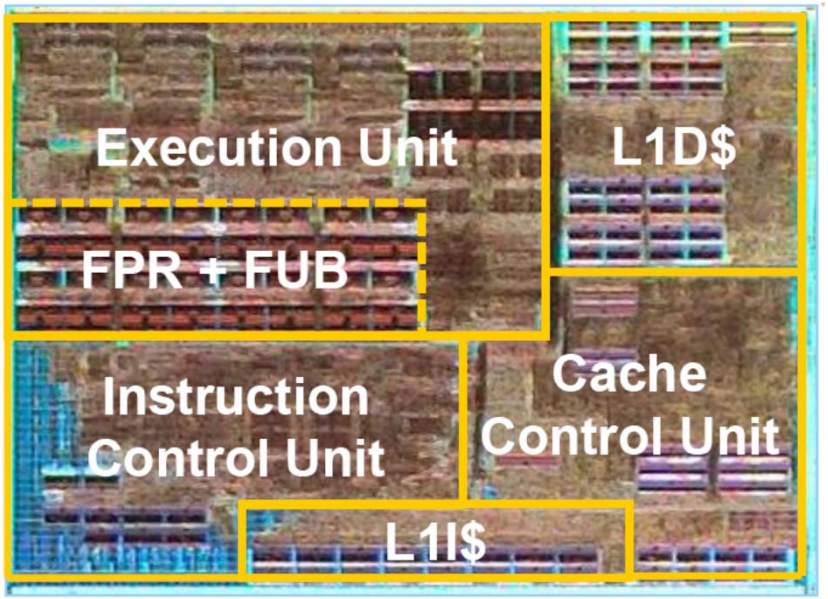


- Aggregate memory bandwidth: 398 TB/sec.
- Local file system for staging with 1.1 PB of capacity and 131 GB/sec of aggregate I/O performance (for staging)
- Shared file system for storing data with 2.1 PB and 136 GB/sec.
- External file system: 3.6 PB

Copyright 2011 FUJITSU LIMITED

SPARC64™ IXfx



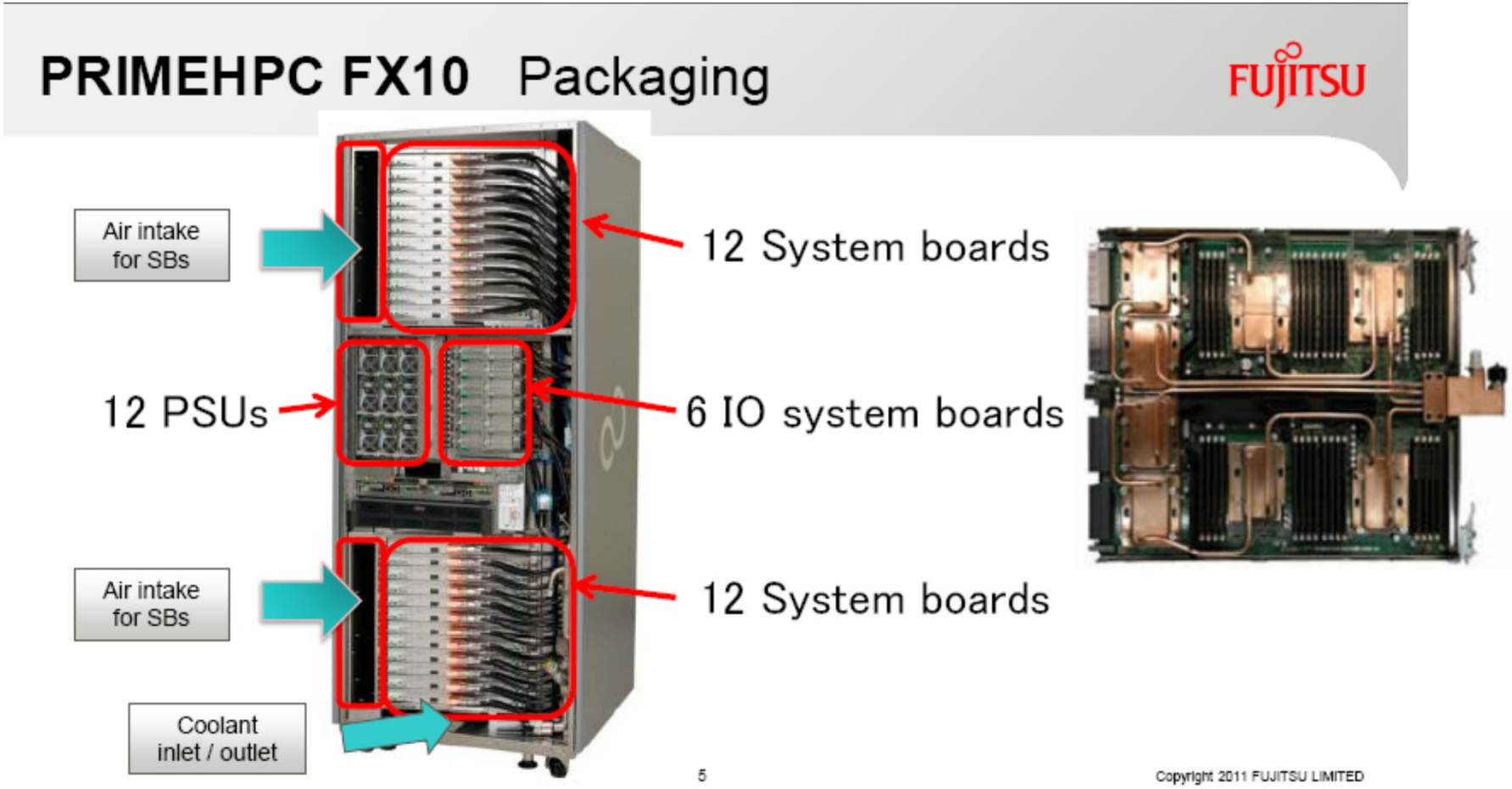


Copyright 2011 FUJITSU LIMITED

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)

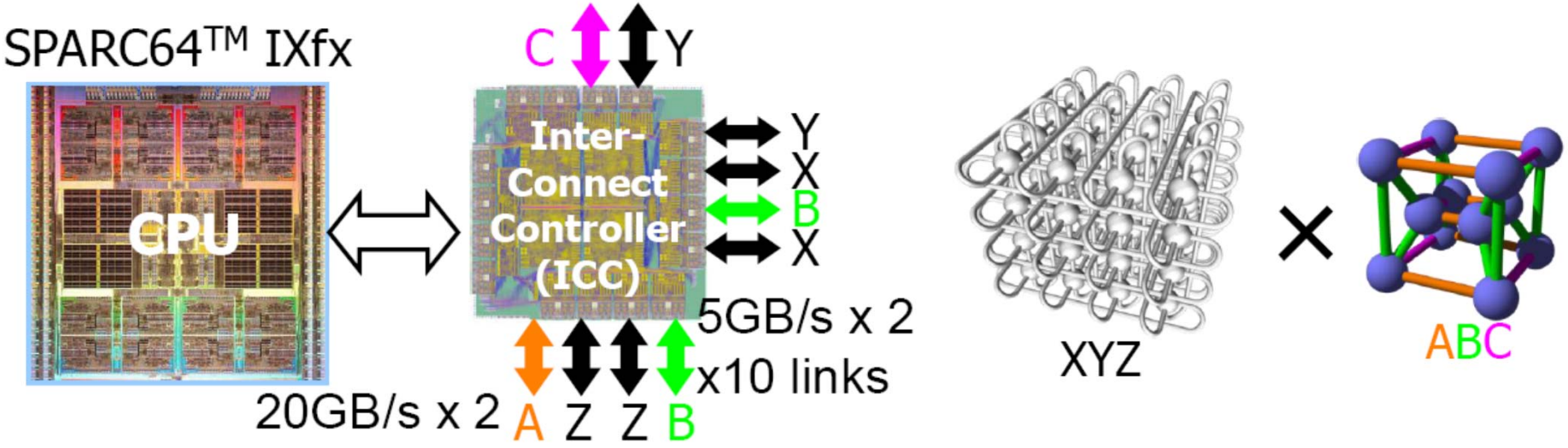
Racks

- 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



Copyright 2011 FUJITSU LIMITED

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.	

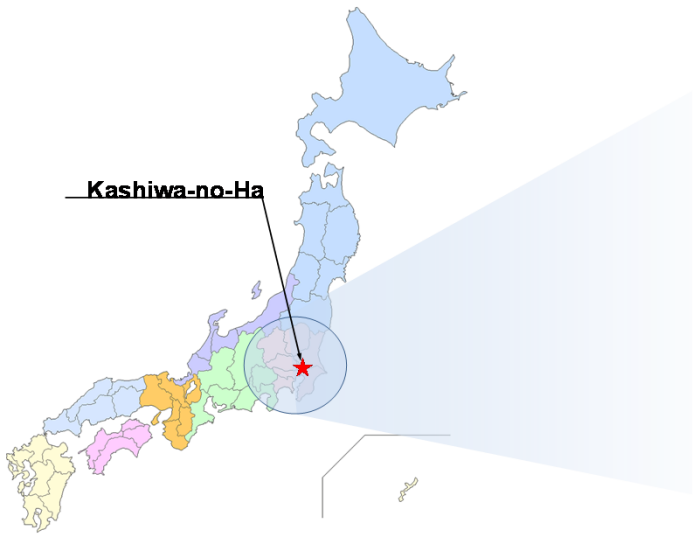
NO ISV/Commercial Applications (e.g. NASTRAN, ABAQUS, STAR-CD etc.)

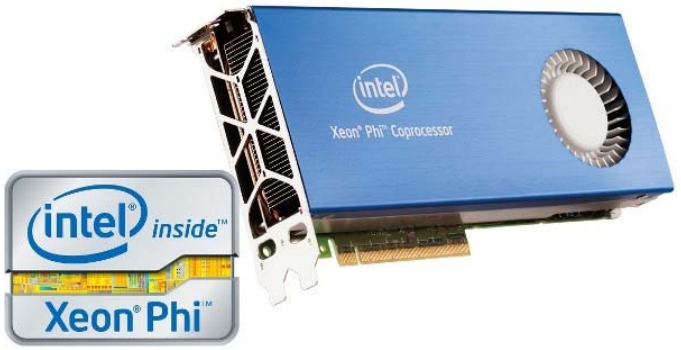
- HPCI
- Supercomputer Systems in SCD/ITC/UT
- Overview of Fujitsu FX10 (Oakleaf-FX)

- **Post T2K System**

Post T2K System

- 20-30 PFLOPS, FY.2015-2016
- Many-core based (e.g. (only) Intel MIC/Xeon Phi)
- 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





Post T2K System

- 20-30 PFLOPS, FY.2015-2016
- Many-core based (e.g. (only) Intel MIC/Xeon Phi)
- 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
- Programming is still difficult, although Intel compiler works.
  - (MPI + OpenMP) + Y
  - Tuning for performance (e.g. prefetching) is essential
  - Some framework for helping users needed