東京大学
THE UNIVERSITY OF TOKYO

# Challenges towards Post-Peta/Exascale Computing

## Information Technology Center
## The University of Tokyo


**Kengo Nakajima**

Information Technology Center
The University of Tokyo



53rd HPC User Forum, RIKEN AICS, Kobe, Japan
July 16, 2014

# 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)

# 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)

# ＞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

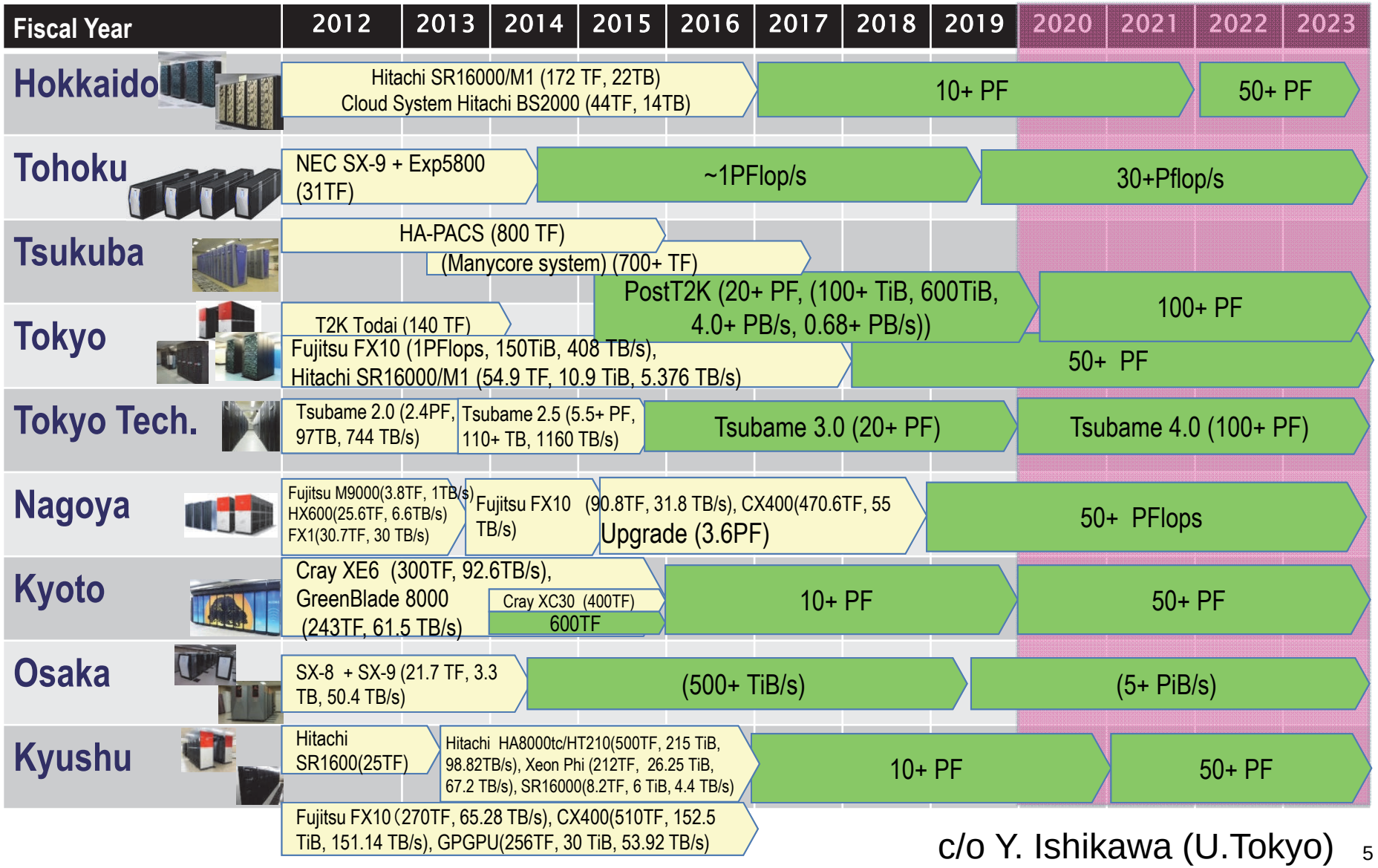
Kyushu Univ.：
FX10 (181.6TF, 24TB)
CX400 (811.9TF, 184.5TB)
HA8000-tc (712.5TF, 24.7TB)
SR16000 VM1 (8.19TF, 16TB)

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

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

# 9 supercomputer centers located at universities in Japan

**The next national flagship machine**




c/o Y. Ishikawa (U.Tokyo)

# Supercomputing 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)
- 12 Faculty Members
  - System Software, Numerical Library, Applications, GPU

# 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**
    - **Fujitsu**
  - **ppOpen-HPC (FY.2011-)**
- **International Collaboration**
  - **Lawrence Berkeley National Laboratory (USA)**
  - **National Taiwan University (Taiwan)**
  - **Intel Parallel Computing Center**

# ICT/U.Tokyo joined IPCC in Dec.2013

officially announced in June 2014

focusing on optimization of FEM/ICCG solver on Xeon/Phi

- **Supercomputer Systems in SCD/ITC/UT**
- Post T2K System, ppOpen-HPC

# 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, 36$^{th}$ in 43$^{rd}$ TOP500)
  - Additional 576 Nodes with 136 TF (Oakbridge-FX, 2014.04-)

# 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** | |

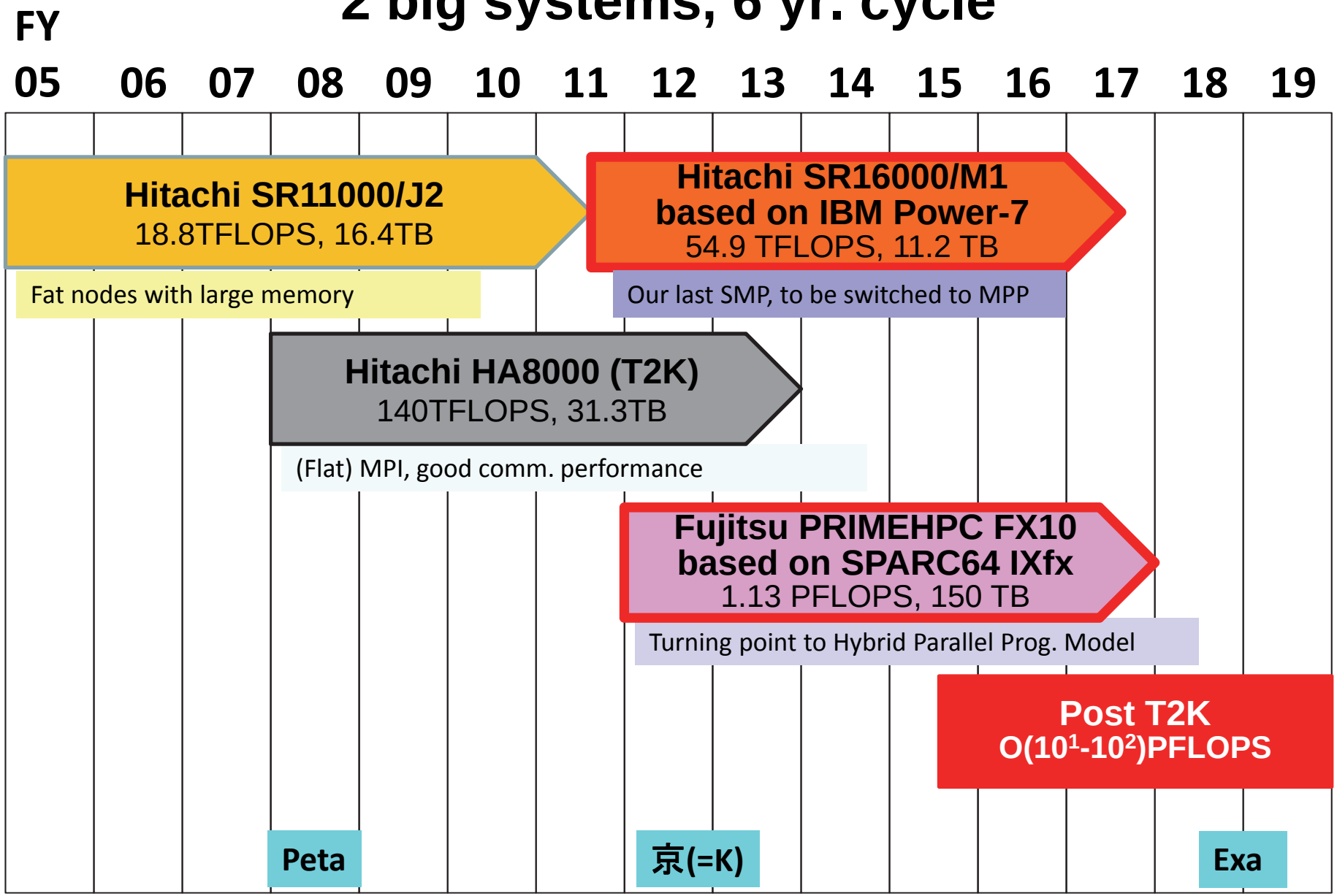
| Yayoi (Hitachi SR16000/M1) | |
|---|---|
| Total Peak performance | : 54.9 TFLOPS |
| Total number of nodes | : 56 |
| Total memory | : 11200 GB |
| Peak performance / node | : 980.48 GFLOPS |
| Main memory per node | : 200 GB |
| Disk capacity | : 556 TB |
| **IBM POWER 7 3.83GHz** | |

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

**Total Users > 2,000**

# Supercomputers in U.Tokyo

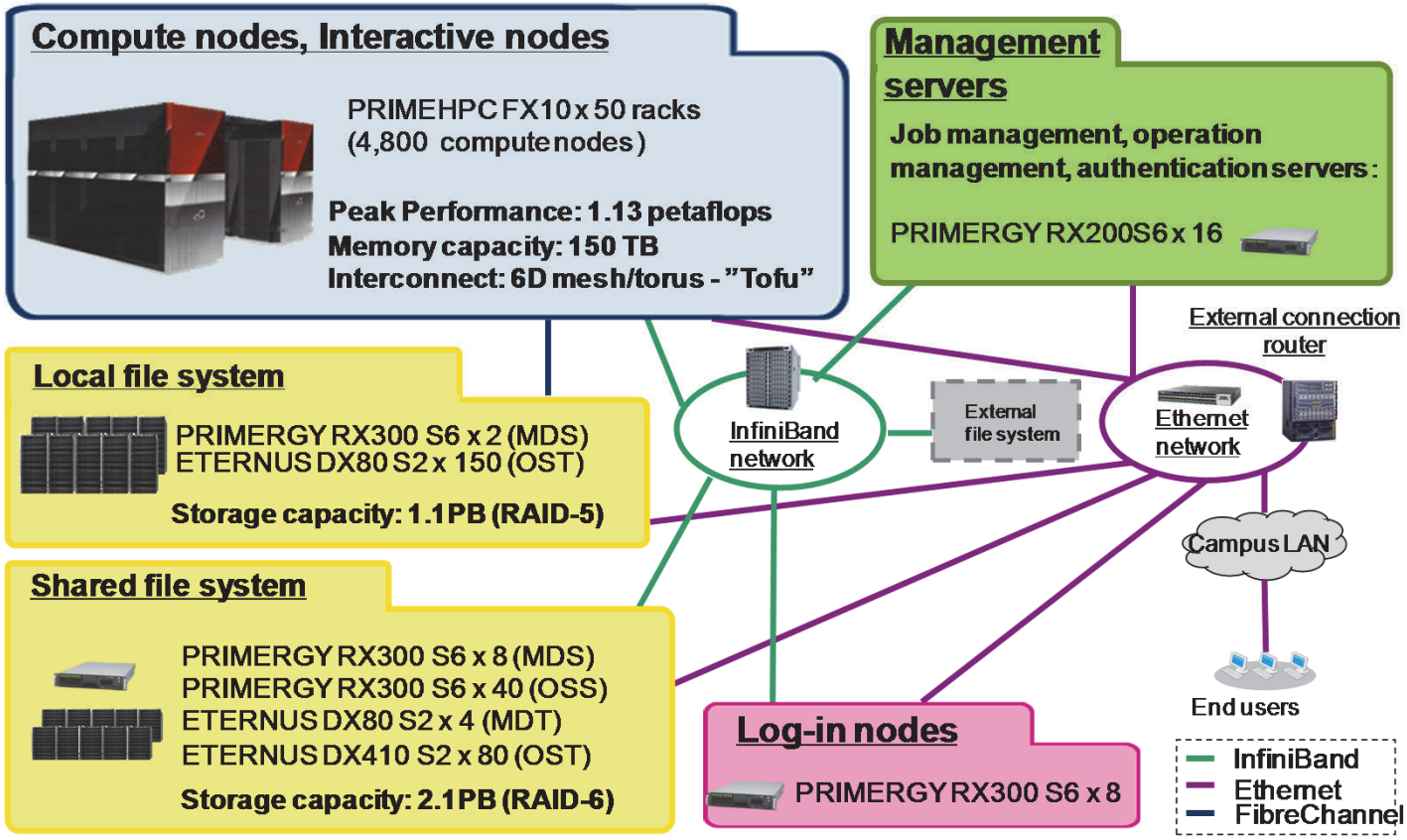
## 2 big systems, 6 yr. cycle



# Features of FX10 (Oakleaf-FX)

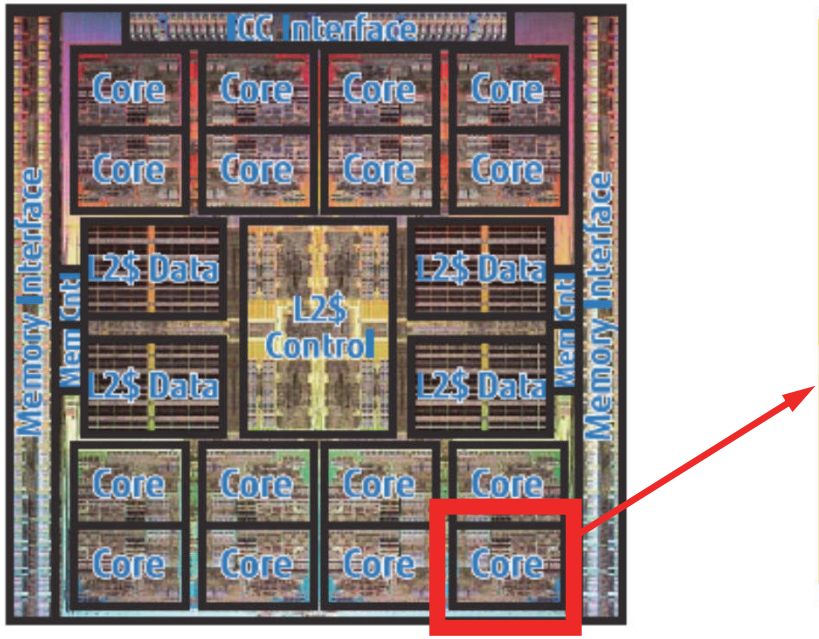
- Well-Balanced System
  - Peak Performance: 1.13 PFLOPS, 398 TB/sec
  - Max. Power Consumption < 1.40 MW (<2.00MW with A/C)
    - Strict Requirement after March 11, 2011
    - 1.043 PFLOPS for Linpack with 1.177 MW (excluding A/C)
- 6-Dim. Mesh/Torus Interconnect
  - Highly Scalable Tofu Interconnect
  - 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)

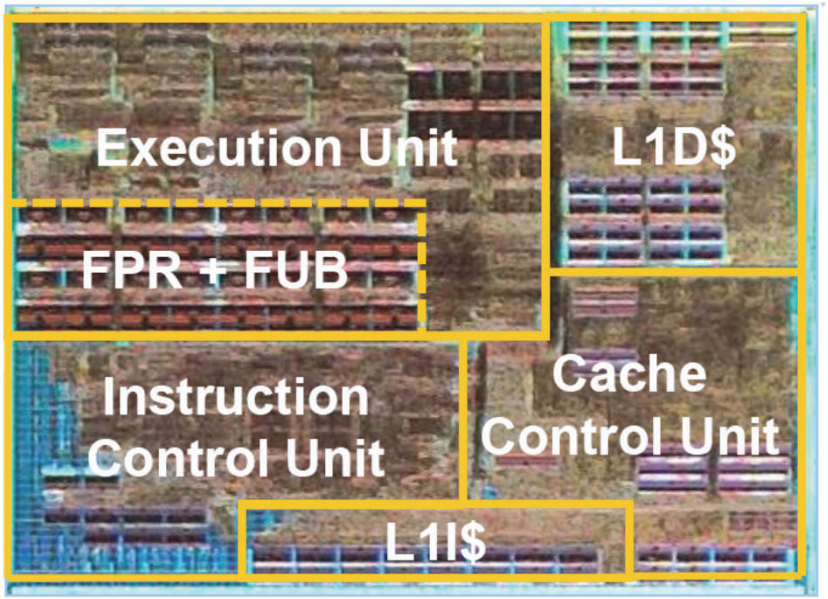
# FX10 System (Oakleaf-FX)




- 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

# SPARC64™ IXfx





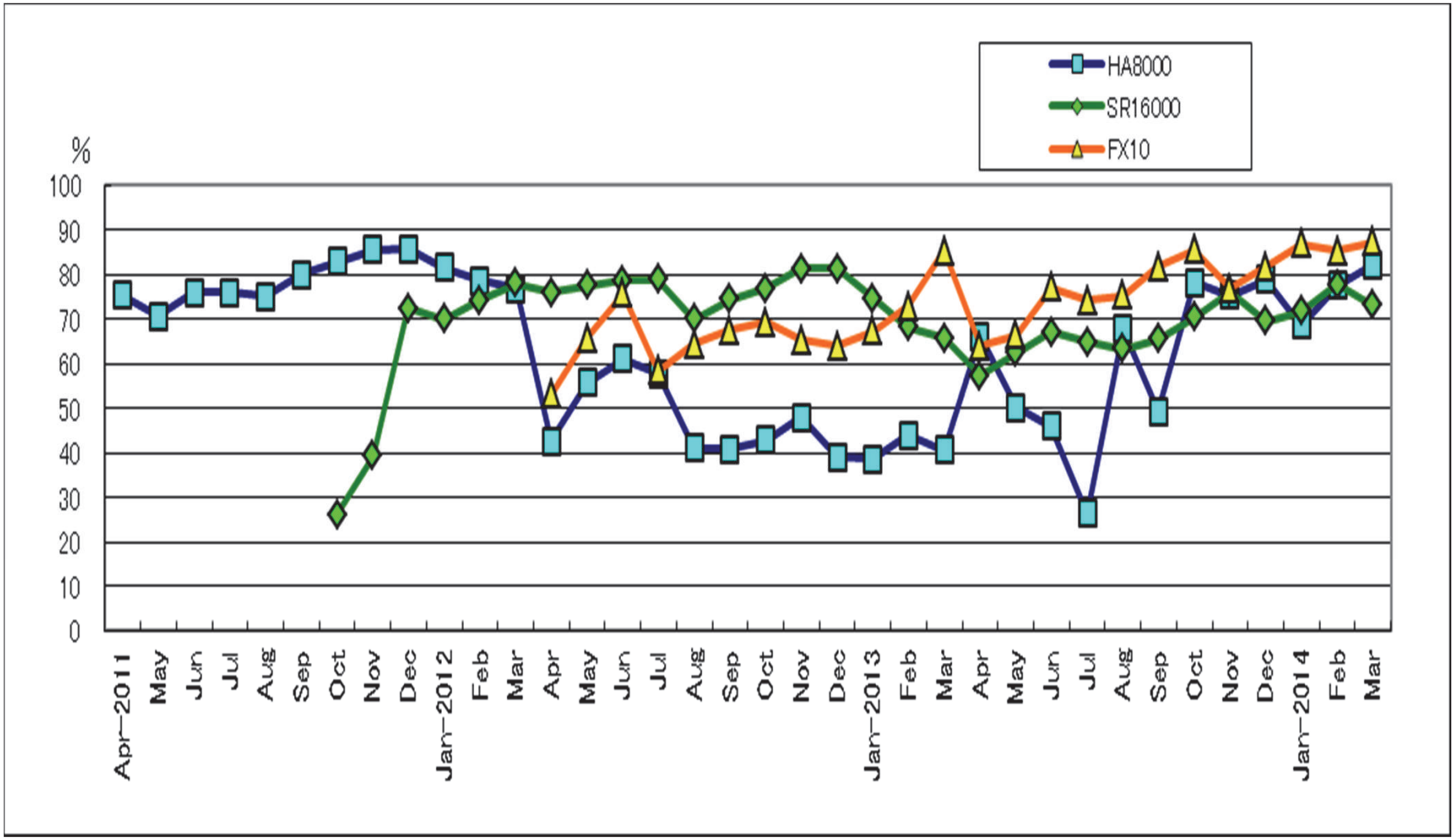

| CPU | SPARC64™ IXfx 1.848 GHz | SPARC64™ VIIIfx 2.000 GHz |
|---|---|---|
| Number of Cores/Node | 16 | 8 |
| Size of L2 Cache/Node | 12 MB | 6 MB |
| Peak Performance/Node | 236.5 GFLOPS | 128.0 GFLOPS |
| Memory/Node | 32 GB | 16 GB |
| Memory Bandwidth/Node | 85 GB/sec (DDR3-1333) | 64 GB/sec (DDR3-1000)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

# Software of FX10

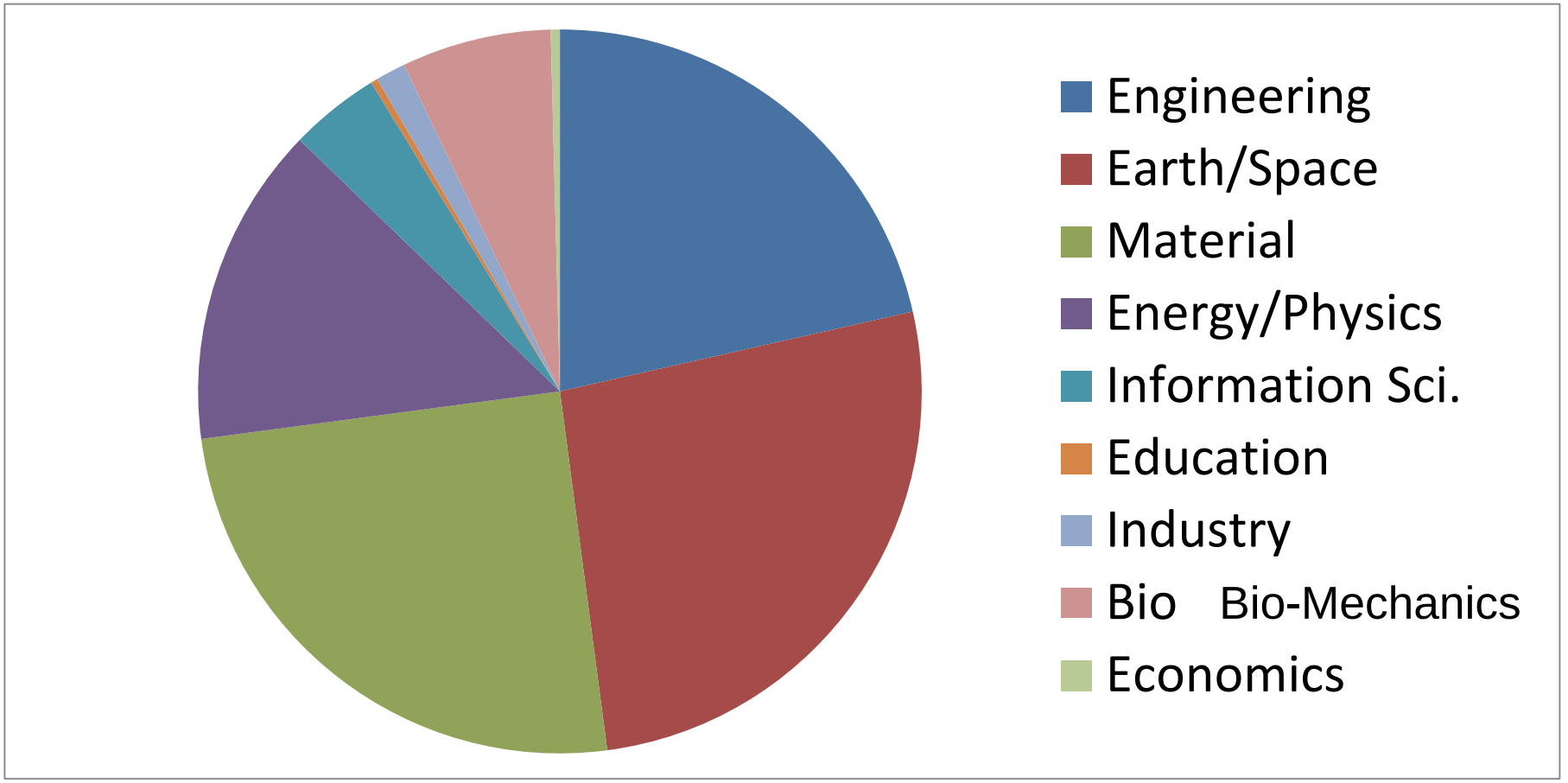
| | Computing/Interactive Nodes | Login Nodes |
|---|---|---|
| OS | Special OS（XTCOS） | Red Hat Enterprise Linux |
| Compiler | Fujitsu<br>Fortran 77/90<br>C／C++<br>GNU<br>GCC，g95 | Fujitsu (Cross Compiler)<br>Fortran 77/90<br>C／C++<br>GNU (Cross Compiler)<br>GCC，g95 |
| Library | Fujitsu<br>SSL II (Scientific Subroutine Library II)，C-SSL II，SSL II/MPI<br>Open Source<br>BLAS，LAPACK，ScaLAPACK，FFTW，SuperLU，PETSc，METIS，<br>SuperLU_DIST，Parallel NetCDF | |
| Applications | OpenFOAM，ABINIT-MP，PHASE，<br>FrontFlow/blue<br>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.)**

# History of Work Ratio

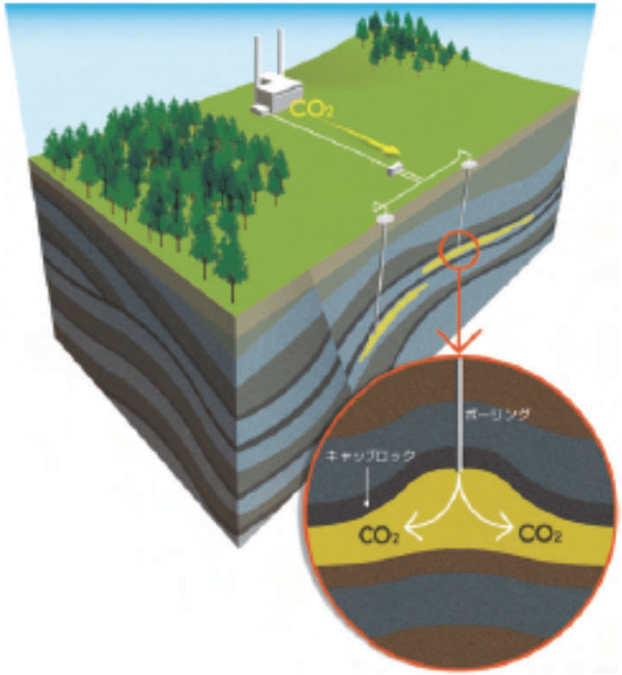


# Research Area based on CPU Hours FX10 in FY.2013 (2013.4~2014.3E)




Applications

# Simulation of Geologic $CO_2$ Storage



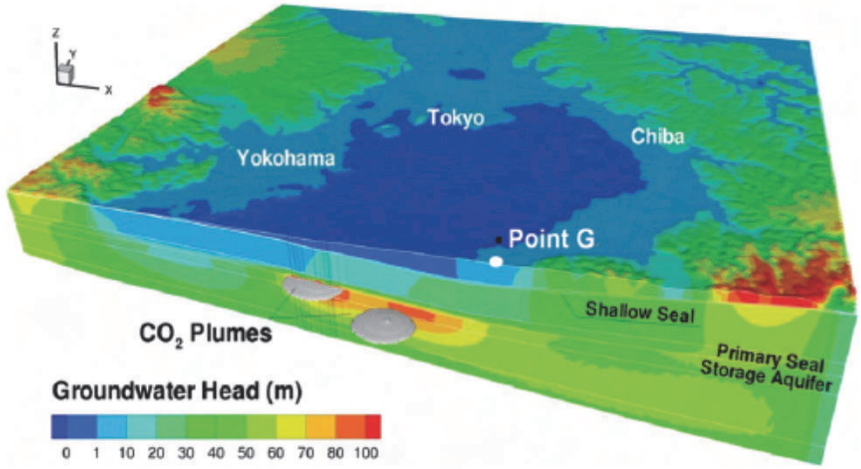
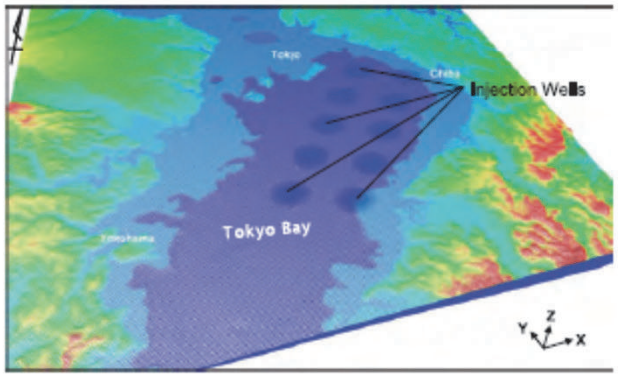
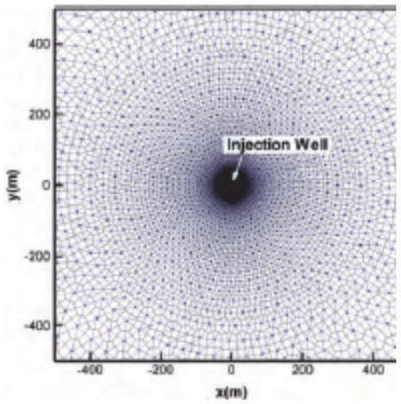
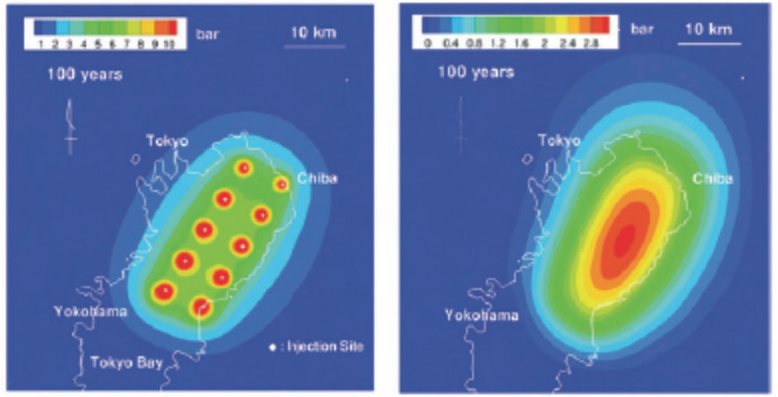




図-4　$CO_2$ 圧入後の地下水圧（全水頭換算）の分布（100 年後）



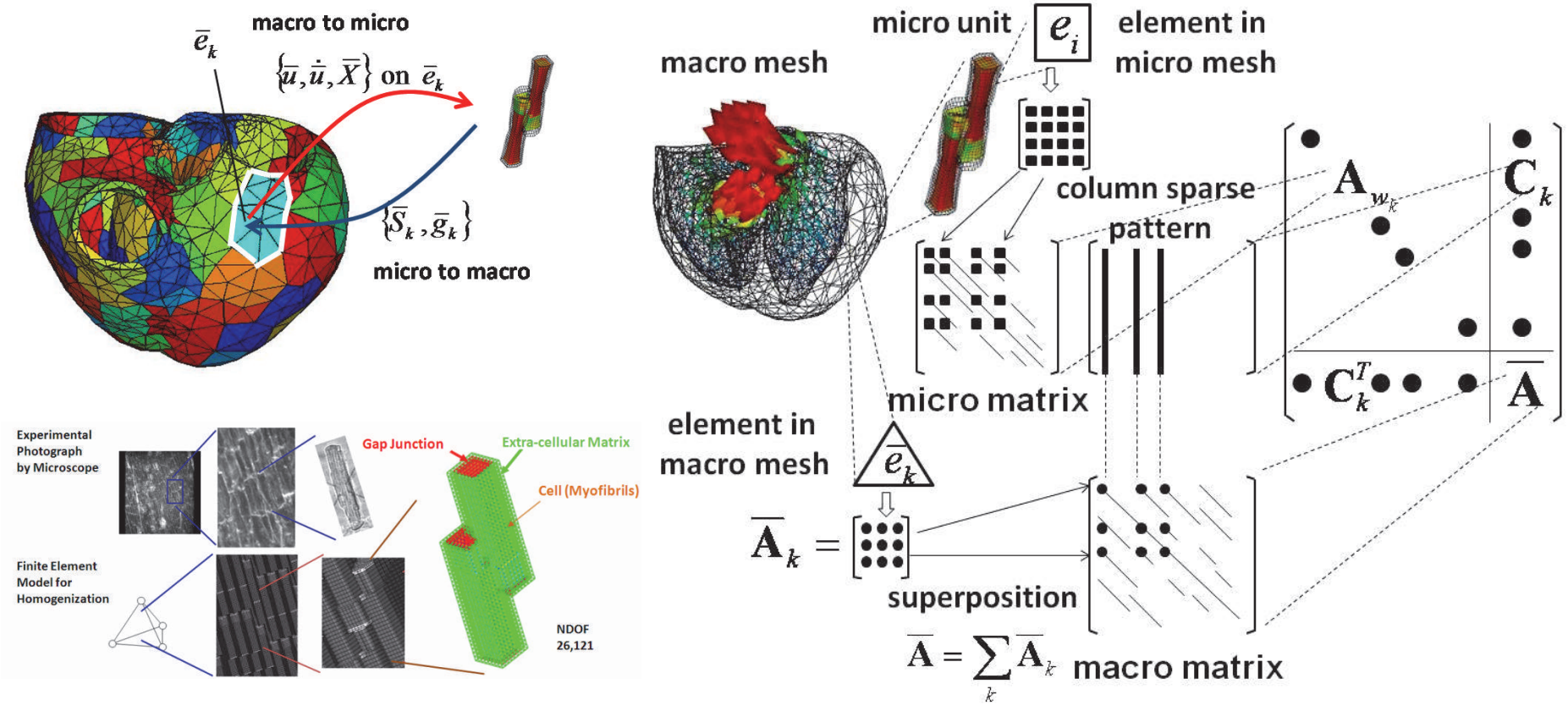

(a) 深部遮蔽層下面　　(b) 浅部遮蔽層下面

図-5　圧力上昇量の平面分布（初期状態からの増分、圧入開始から 100 年後）

[Dr. Hajime Yamamoto, Taisei]

# Multi-Scale/Physics Heart Simulator (UT Heart)

Prof. T. Hisada (U.Tokyo) et al., SC10




[Hisada-Sugiura Lab., U.Tokyo]

# 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.)

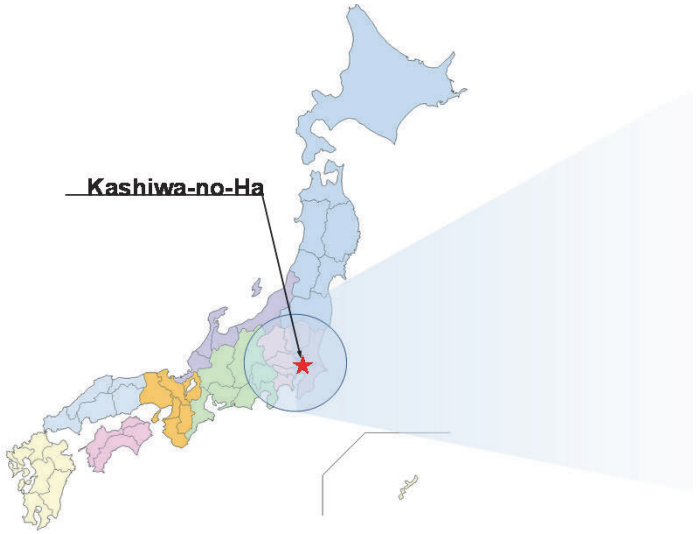
# Training & Education (FX10)

- 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~)

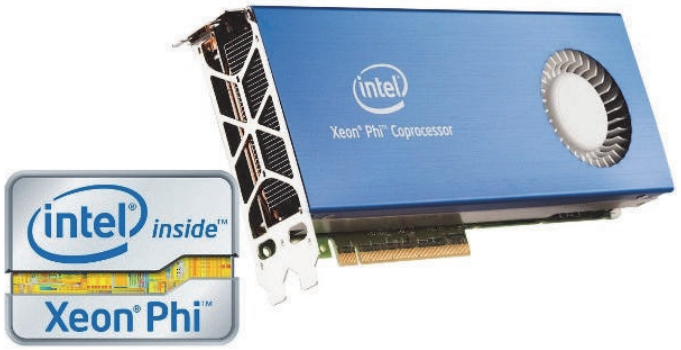
- Supercomputer Systems in SCD/ITC/UT
- **Post T2K System, ppOpen-HPC**

# Post T2K System

- 20-30 PFLOPS, FY.2015
- 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

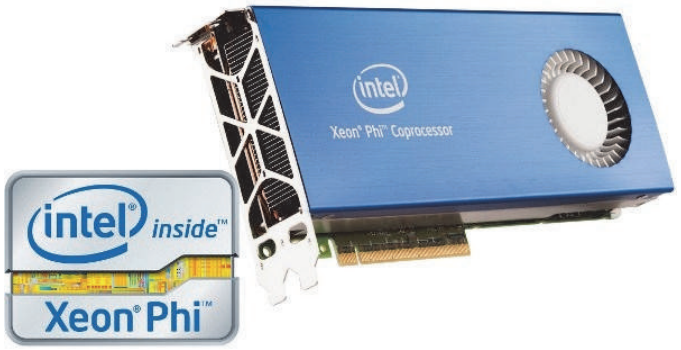






# Post T2K System

- 20-30 PFLOPS, FY.2015
- 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



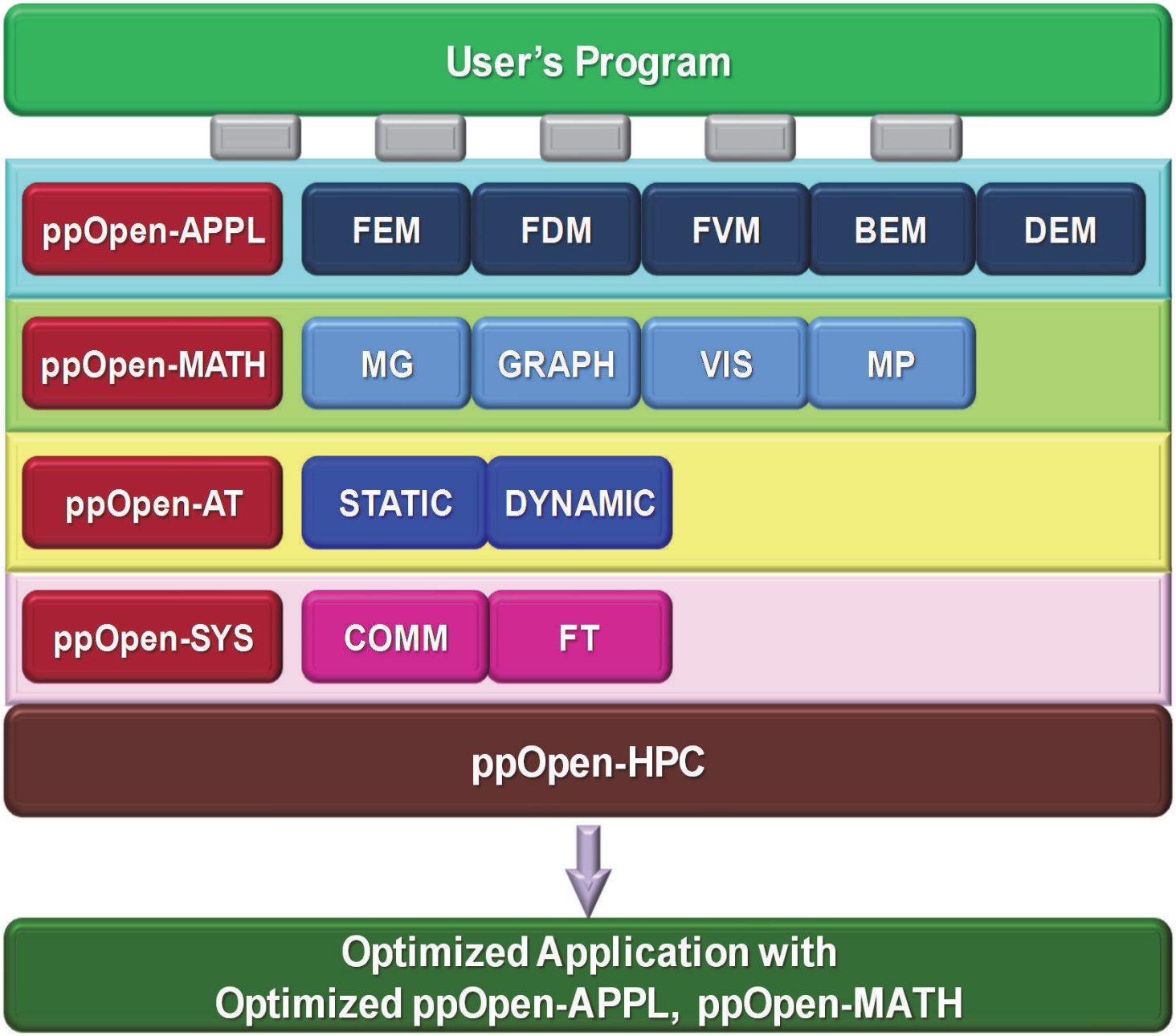
# Key-Issues for Appl's/Algorithms towards Post-Peta & Exa Computing

Jack Dongarra (ORNL/U. Tennessee) at ISC 2013

- Heterogeneous/Hybrid Architecture
- Communication/Synchronization Reducing Algorithms
- Mixed Precision Computation
- Auto-Tuning/Self-Adapting
- Fault Resilient Algorithms
- Reproducibility of Results

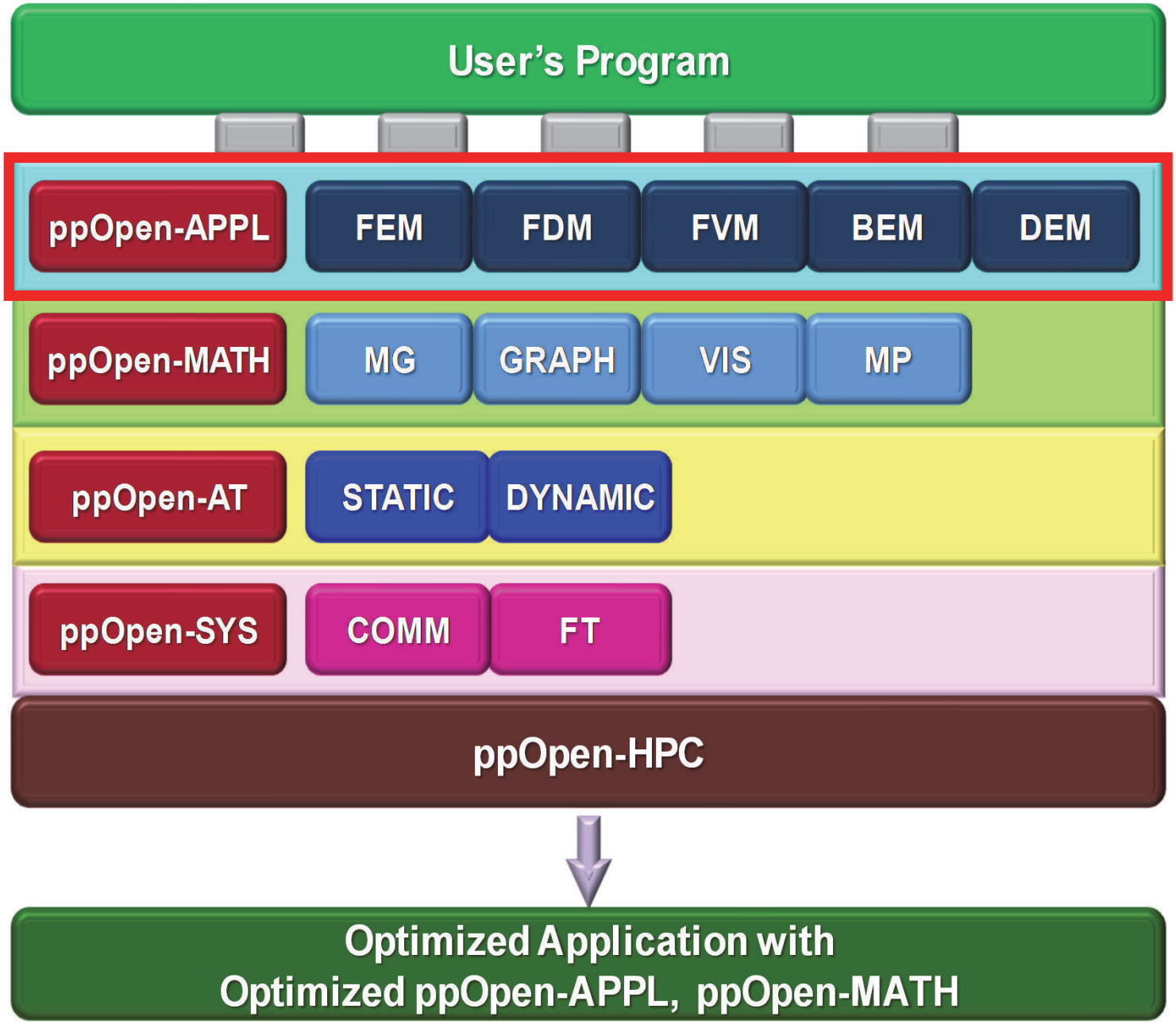
# ppOpen-HPC (1/3)

- Open Source Infrastructure for development and execution of large-scale scientific applications on post-peta-scale supercomputers with automatic tuning (AT)
  - “pp” : post-peta-scale
- Five-year project (FY.2011-2015) (since April 2011)
  - P.I.: Kengo Nakajima (ITC, The University of Tokyo)
  - Part of “Development of System Software Technologies for Post-Peta Scale High Performance Computing” funded by JST/CREST (Japan Science and Technology Agency, Core Research for Evolutional Science and Technology)
    - Supervisor: Prof. Akinori Yonezawa (Co-Director, RIKEN AICS)
  - 4.5 M$ for 5 yr.
- Team with 7 institutes, >30 people (5 PDs) from various fields: Co-Desigin
  - U.Tokyo (4 divisions), Kyoto U., Hokkaido U., JAMSTEC

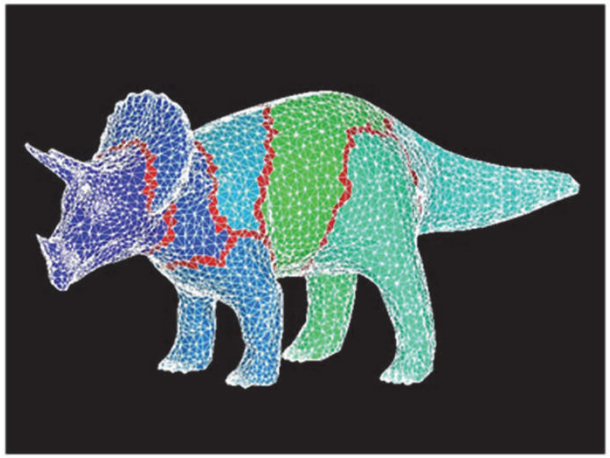

User's Program
ppOpen-APPL
FEM
FDM
FVM
BEM
DEM
ppOpen-MATH
MG
GRAPH
VIS
MP
ppOpen-AT
STATIC
DYNAMIC
ppOpen-SYS
COMM
FT
ppOpen-HPC
Optimized Application with
Optimized ppOpen-APPL, ppOpen-MATH

# ppOpen-HPC (2/3)

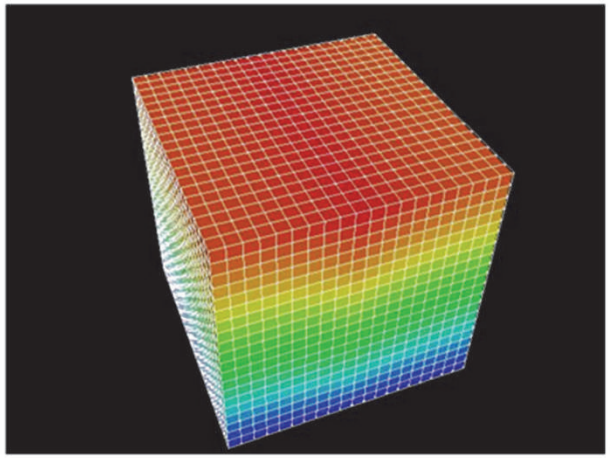
- ppOpen-HPC consists of various types of *optimized* libraries, which covers various types of procedures for scientific computations.
  - ppOpen-APPL/FEM, FDM, FVM, BEM, DEM
- Source code developed on a PC with a single processor is linked with these libraries, and generated parallel code is optimized for post-peta scale system.
- Users don’t have to worry about optimization tuning, parallelization etc.
  - Part of MPI, OpenMP
  - (OpenACC)


User's Program
ppOpen-APPL
FEM
FDM
FVM
BEM
DEM
ppOpen-MATH
MG
GRAPH
VIS
MP
ppOpen-AT
STATIC
DYNAMIC
ppOpen-SYS
COMM
FT
ppOpen-HPC
Optimized Application with
Optimized ppOpen-APPL, ppOpen-MATH

# ppOpen-HPC covers …

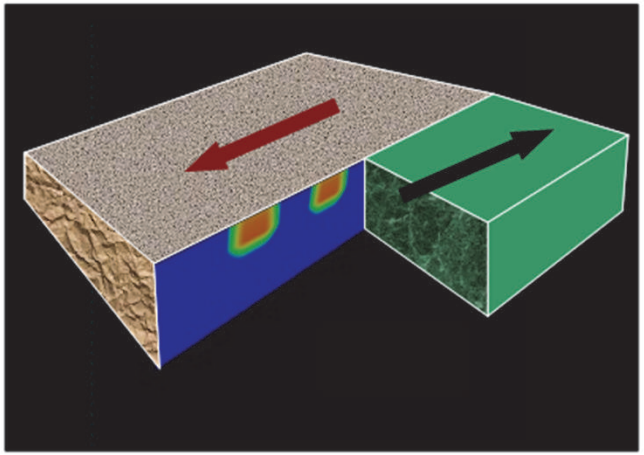


**FEM**
Finite Element Method

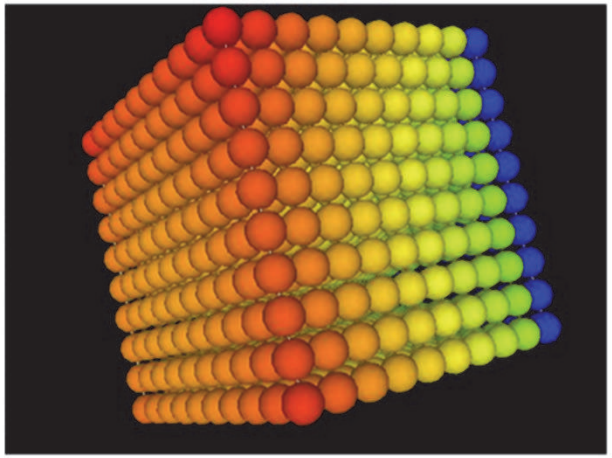


**FDM**
Finite Difference Method

**FVM**
Finite Volume Method



**BEM**
Boundary Element Method



**DEM**
Discrete Element Method

# ppOpen-APPL

- A set of libraries corresponding to each of the five methods noted above (FEM, FDM, FVM, BEM, DEM), providing:
  - I/O
    - netCDF-based Interface
  - Domain-to-Domain Communications
  - Optimized Linear Solvers (Preconditioned Iterative Solvers)
    - Optimized for each discretization method
  - H-Matrix Solvers in ppOpen-APPL/BEM
  - Matrix Assembling
  - AMR and Dynamic Load Balancing
- Most of components are extracted from existing codes developed by members

# FEM Code on ppOpen-HPC

## Optimization/parallelization could be hidden from application developers

```
Program My_pFEM
use ppOpenFEM_util
use ppOpenFEM_solver

call ppOpenFEM_init
call ppOpenFEM_cntl
call ppOpenFEM_mesh
call ppOpenFEM_mat_init

do
  call Users_FEM_mat_ass
  call Users_FEM_mat_bc
  call ppOpenFEM_solve
  call ppOPenFEM_vis
  Time= Time + DT
enddo

call ppOpenFEM_finalize
stop
end
```

# ppOpen-HPC (2/3)

- ppOpen-HPC consists of various types of *optimized* libraries, which covers various types of procedures for scientific computations.
  - ppOpen-APPL/FEM, FDM, FVM, BEM, DEM
- Source code developed on a PC with a single processor is linked with these libraries, and generated parallel code is optimized for post-peta scale system.
- Users don’t have to worry about optimization tuning, parallelization etc.
  - Part of MPI, OpenMP
  - (OpenACC)

# ppOpen-HPC (3/3)

- Capability of automatic tuning (AT) enables development of optimized codes and libraries on emerging architecture based on results by existing architectures and machine parameters.
  - Mem. Access, Host/Co-Proc. Balance, Comp./Comm. Overlapping
  - Solvers & Libraries in ppOpen-HPC
  - OpenFOAM, PETSc etc.
- Target system is Post T2K system
  - 20-30 PFLOPS, FY.2015
  - Many-core based (e.g. Intel MIC/Xeon Phi)
  - **ppOpen-HPC helps smooth transition of users to new system**

# Schedule of Public Release

(with English Documents, MIT License)

We are now focusing on MIC/Xeon Phi

- 4Q 2012 (Ver.0.1.0)
  - ppOpen-HPC for Multicore Cluster (Cray, K etc.)
  - Preliminary version of ppOpen-AT/STATIC
- 4Q 2013 (Ver.0.2.0)
  - ppOpen-HPC for Multicore Cluster & Xeon Phi (& GPU)
  - available in SC'13
- 4Q 2014
  - Prototype of ppOpen-HPC for Post-Peta Scale System
- 4Q 2015
  - Final version of ppOpen-HPC for Post-Peta Scale System
  - Further optimization on the target system

# ppOpen-HPC v.0.1.0

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

- Released at SC12 (or can be downloaded)
- Multicore cluster version (Flat MPI, OpenMP/MPI Hybrid) with documents in English
- Collaborations with scientists

| Component | Archive | Flat MPI | OpenMP/MPI | C | F |
|---|---|---|---|---|---|
| ppOpen-APPL/FDM | ppohFDM_0.1.0 | ○ | | | ○ |
| ppOpen-APPL/FVM | ppohFVM_0.1.0 | ○ | ○ | | ○ |
| ppOpen-APPL/FEM | ppohFEM_0.1.0 | ○ | ○ | ○ | ○ |
| ppOpen-APPL/BEM | ppohBEM_0.1.0 | ○ | ○ | | ○ |
| ppOpen-APPL/DEM | ppohDEM_0.1.0 | ○ | ○ | | ○ |
| ppOpen-MATH/VIS | ppohVIS_FDM3D_0.1.0 | ○ | | ○ | ○ |
| ppOpen-AT/STATIC | ppohAT_0.1.0 | - | - | ○ | ○ |

# What is new in Ver.0.2.0 ?

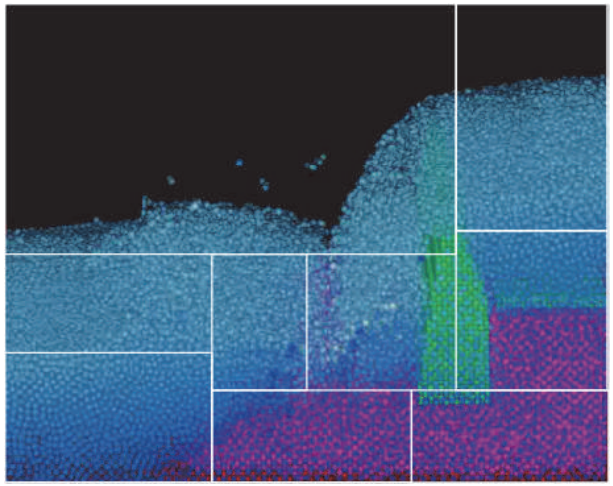
http://ppopenhpc.cc.u-tokyo.ac.jp/

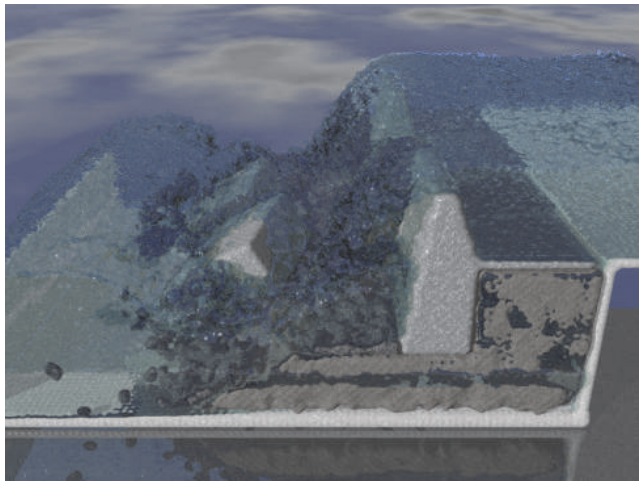
- Available in SC13 (or can be downloaded)

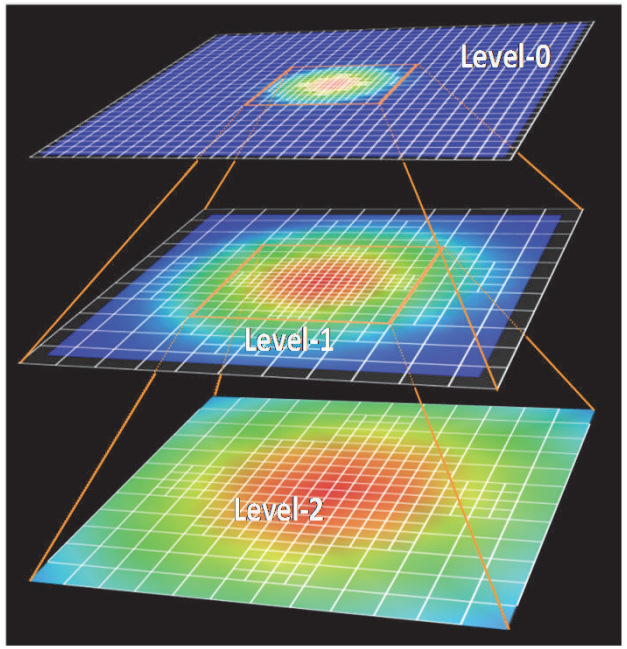
| Component | New Development |
|---|---|
| **ppOpen-APPL/FDM** | • OpenMP/MPI Hybrid Parallel Programming Model<br>• Intel Xeon/Phi Version<br>• Interface for ppOpen-MATH/VIS-FDM3D |
| **ppOpen-APPL/FVM** | • Optimized Communication |
| **ppOpen-APPL/FEM** | • Sample Implementations for Dynamic Solid Mechanics |
| **ppOpen-MATH/MP-PP** | • Tool for Generation of Remapping Table in ppOpen-MATH/MP |
| **ppOpen-MATH/VIS** | • Optimized ppOpen-MATH/VIS-FDM3D |
| **ppOpen-AT/STATIC** | • Sequence of Statements, Loop Splitting (Optimized)<br>• ppOpen-APPL/FVM<br>• ppOpen-APPL/FDM・BEM |

# Collaborations, Outreaching

- Collaborations
  - International Collaborations
    - Lawrence Berkeley National Lab., National Taiwan University
  - We are happy to do any types of research collaborations
- Outreaching, Applications
  - Large-Scale Simulations: Most Important for Demonstrations of the Potential of ppOpen-HPC
    - JHPCN (Joint Usage/Research Ctr. for Interdisciplinary Large-scale Information Infrastructures )
    - ppOpen-AT, ppOpen-MATH/VIS, ppOpen-MATH/MP, Linear Solvers
  - SPNS Workshop (2012, 2013)
  - Tutorials, Classes







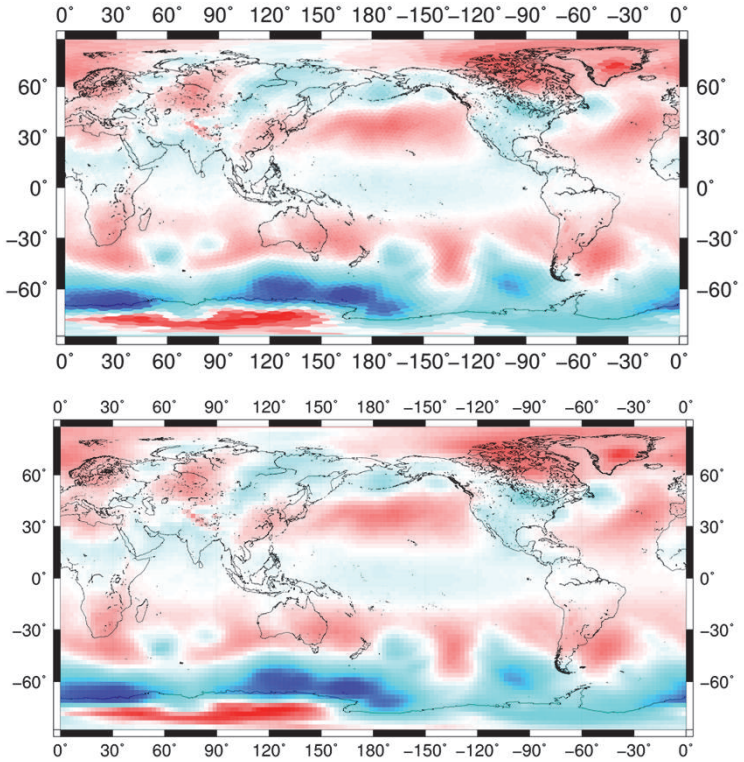
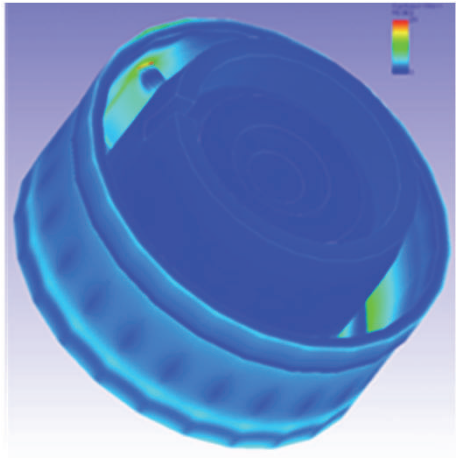
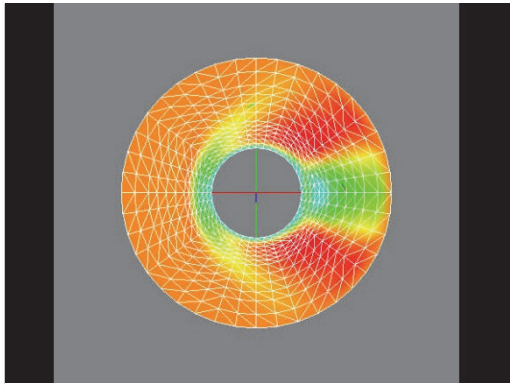
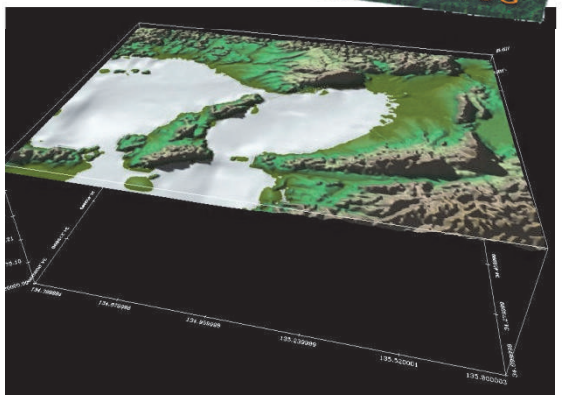




図 1　NICAM と IO モジュールの海面気圧



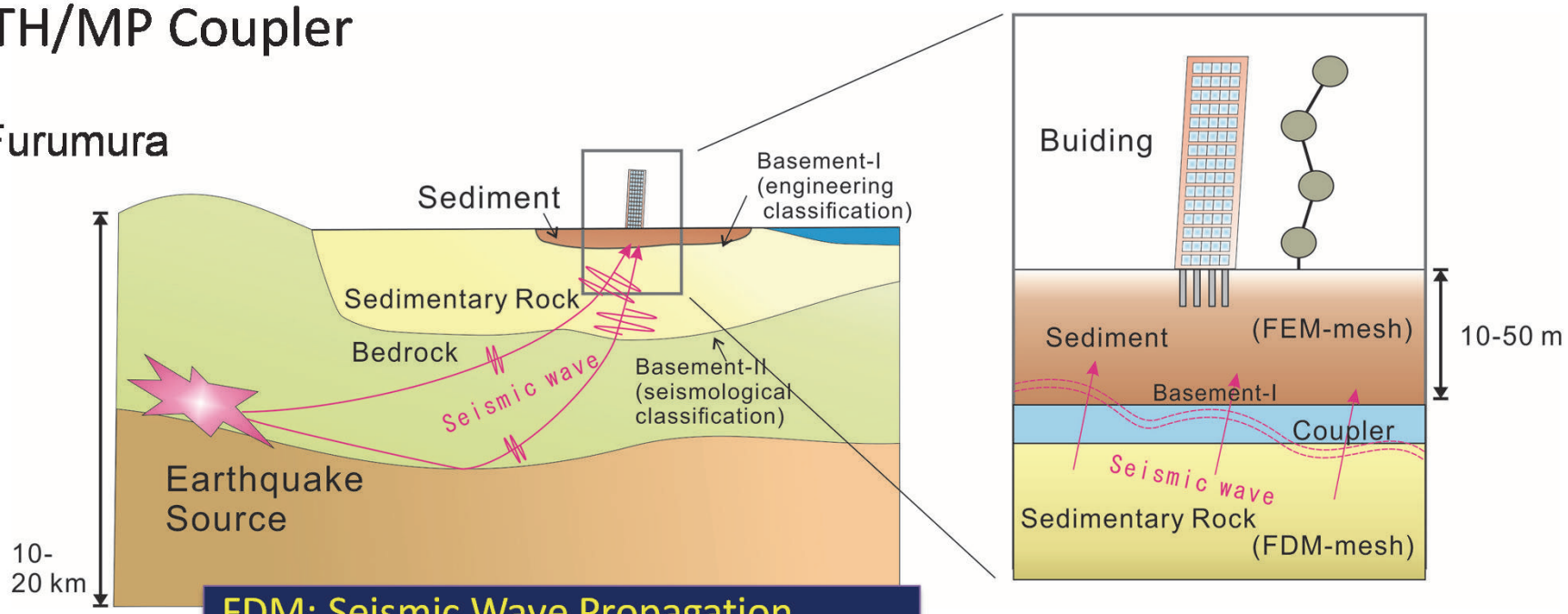



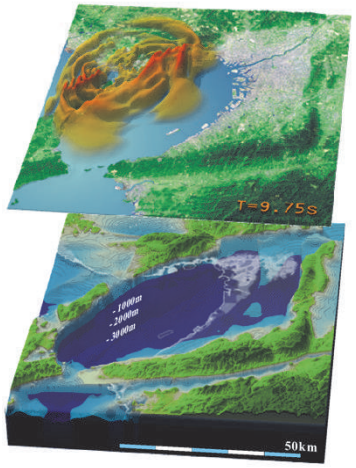


# Large-Scale Coupled Simulations in FY.2013

Challenge (FY2013) : A test of a coupling simulation of FDM (regular grid) and FEM (unconstructed grid) using newly developed ppOpen-MATH/MP Coupler

c/o T.Furumura




**FDM: Seismic Wave Propagation**
Model size: 80x80x400 km
Time : 240 s
Resolution (space) : 0.1 km (regular)
Resolution (time) : 5 ms
(effective freq.<1 Hz)

**FEM: Building Response**
Model size : 400x400x200 m
Time : 60 s
Resolution (space) : 1 m
Resolution (time) : 1 ms

**ppOpen-MATH/MP:** Space-temporal interpolation, Mapping between FDM and FEM mesh, etc.

# from Post-Peta to Exascale

- Currently, we are focusing on Post-T2K system by manycore architectures (Intel Xeon/Phi)
- Outline of the Exascale Systems is much clearer than which was in 2011 (when we started this project).
  - Feasibility Study in Japan towards Exascale System
    - More complex, and huge system
    - More difficult to extract performance of applications
  - Frameworks like ppOpen-HPC are really needed
  - Smooth transition from post-peta to exa will be possible through continuous development and improvement of ppOpen-HPC (We need funding for that !)
- **Research Topics in Exascale Era**
  - **Power-Aware Algorithms/AT**
  - **Communication/Synchronization Reducing Algorithms**