HPC基盤の現状と将来

石川裕
東京大学情報理工学系研究科／情報基盤センター
理化学研究所　計算科学研究機構

Outline of This Talk

- Introduction of Key Organizations/Programs/Activities
  - HPCI
    - Innovative High Performance Computing Infrastructure
      - Seamless access to K computer, supercomputers, and user's machines
  - SPIRE
    - Strategic Programs for Innovative Research
- Feasibility Study for future HPC in Japan
  - Background
  - Introduction of “Feasibility study on advanced and efficient latency core-based architecture for future HPCI R&D”
- Post T2K

What is HPCI (Innovative High Performance Computing Infrastructure)

HPCI Consortium

Institutional/University computer centers

Computational Science communities

RIKEN AICS (Advanced Institute for Computational Science)

Providing proposals/suggestions to the government and related organizations

- ✓ Plan and operation of HPCI system
- ✓ Promotion of computational sciences
- ✓ Future supercomputing

proposals
suggestions

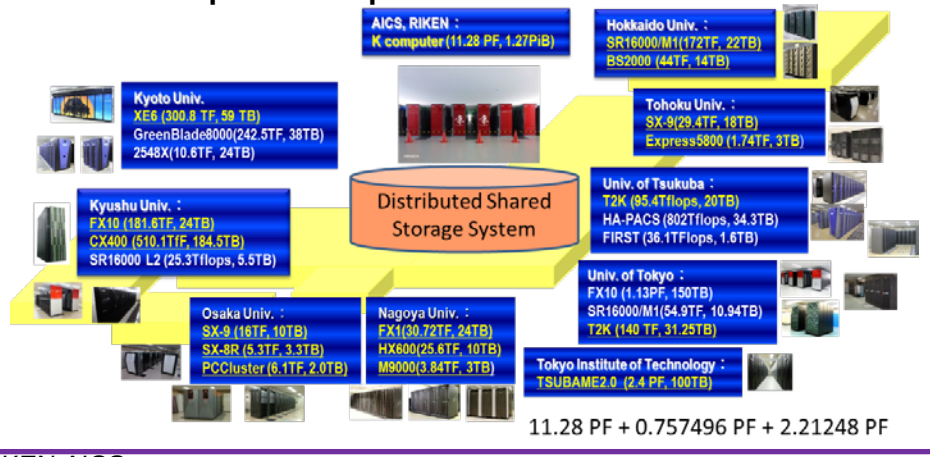
Japanese Government

contract

Operated by RIST, Riken, NII, and U. of Tokyo with 9 supercomputer centers' collaboration

HPCI system

- ✓ HPCI Operation
  - ✓ Seamless access to K computer, supercomputers, and user's machines
  - ✓Distributed shared storage system
- ✓ Joint selection of proposals for K and other supercomputers



- Preparation of HPCI Consortiam starts, October 2010
- Establishment of HPCI Consortium as legal entity 2nd April 2012, 1st General Assembly 6th June 2012 (39 members)
- Submission deadline of proposals using HPCI, 15th June 2012
- Start of HPCI operation, the end of September 2012

RIST: Research Organization for Information Science and Technology
NII: National Institute of Informatics

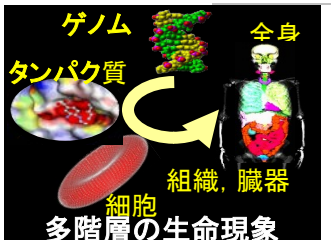
SPIRE (Strategic Programs for Innovative Research)

- Objectives
  - Scientific results as soon as K computer starts its operation
  - Establishment of several core institutes for computational 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

Five strategic groups of SPIRE

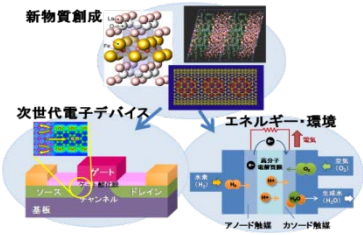
1. Computational Life Science and Application Drug Discovery and Medical Development

   Led by Toshio Yanagida, RIKEN



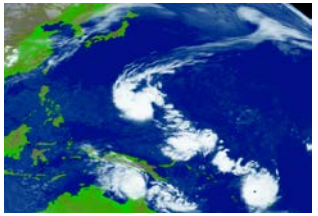
2. Computational Materials Science Initiative

   Led by Shinji Tsuneyuki, University of Tokyo



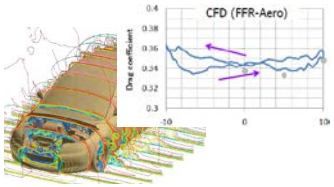
3. Projection of Planet Earth Variations for Mitigating Natural Disasters

   Led by Shiro Imawaki, JAMSTEC



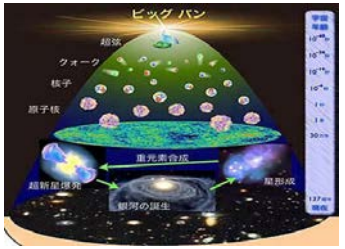
4. Industrial Innovation

   Led by Chisachi Kato, University of Tokyo



5. The origin of matters and the universe

   Led by Shinya Aoki, University of Tsukuba



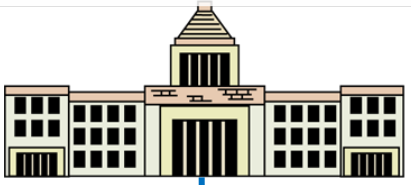
Outline of This Talk

- Introduction of Key Organizations/Programs/Activities
  - HPCI
    - Innovative High Performance Computing Infrastructure
      - Seamless access to K computer, supercomputers, and user's machines
  - SPIRE
    - Strategic Programs for Innovative Research
- Feasibility Study for future HPC in Japan
  - Background
  - Introduction of “Feasibility study on advanced and efficient latency core-based architecture for future HPCI R&D”
- Post T2K

What happened in FY2011

http://www.open-supercomputer.org/workshop/sdhpc/

Japanese Government



MEXT: Ministry of Education, Culture, Sports, Science, and Technology

HPCI Consortium

Providing proposals/suggestions to the government and related organizations
- ✓ Plan and operation of HPCI system
- ✓ Promotion of computational sciences
- ✓ Future supercomputing

Institutional/University computer centers

Computational Science communities

Advanced Institute for Computational Science (AICS), RIKEN

proposals suggestions

contract

report/advise

Council on HPCI Plan and Promotion

Discussion on basic concept and policy

Science roadmap, draft machine requirements to achieve the roadmap, and discussion of challenges to meet requirements

report

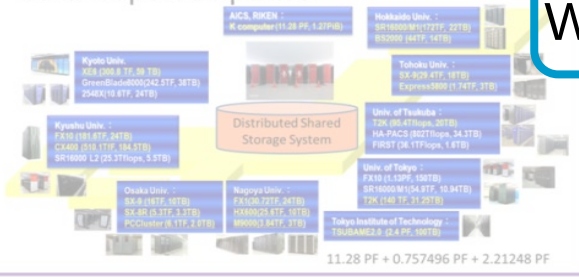
WG for applications

WG for systems

HPCI system

Operated by RIST, Riken, NII, and U. of Tokyo with 9 supercomputer centers' collaboration

- ✓ HPCI Operation
  - ✓ Seamless access to K computer, supercomputers, and user's machines
  - ✓ Distributed shared storage system
- ✓ Joint selection of proposals for K and other supercomputers

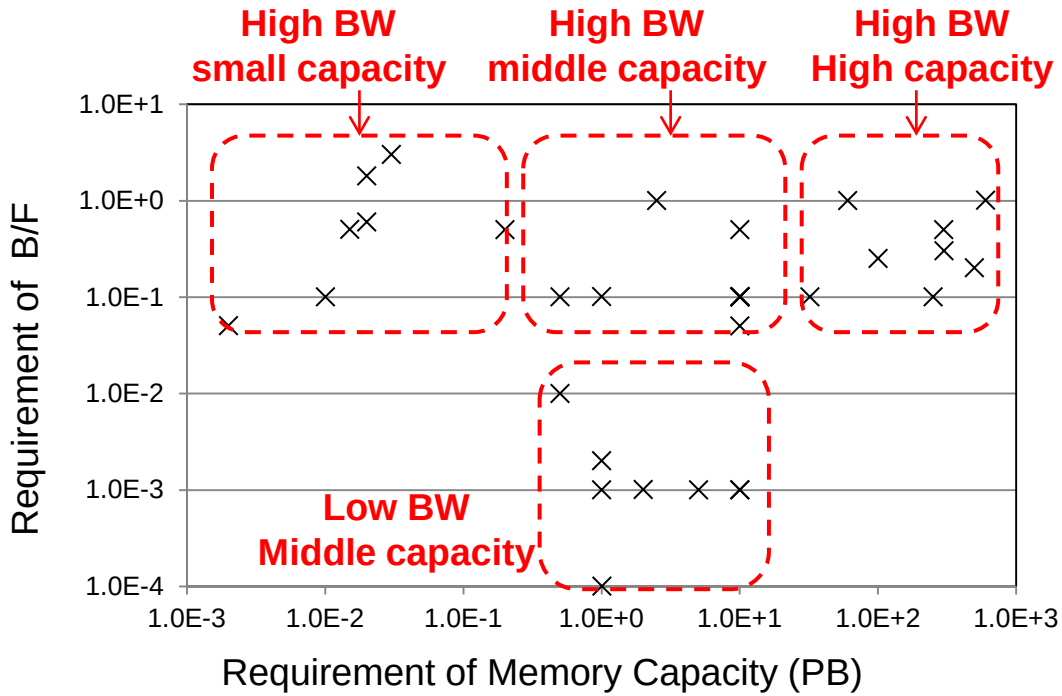


RIST: Research Organization for Information Science and Technology
NII: National Institute of Informatics

White paper for strategic direction/development of HPC in JAPAN, Published in March 2012

System Requirement for Target Sciences by 2020

- System performance
  - FLOPS: 800 – 2500PFLOPS
  - Memory capacity: 10TB – 500PB
  - Memory bandwidth: 0.001 – 1.0 B/F
  - Example applications
    - Small capacity requirement
      - MD, Climate, Space physics, …
    - Small BW requirement
      - Quantum chemistry, …
    - High capacity/BW requirement
      - Incompressibility fluid dynamics, …

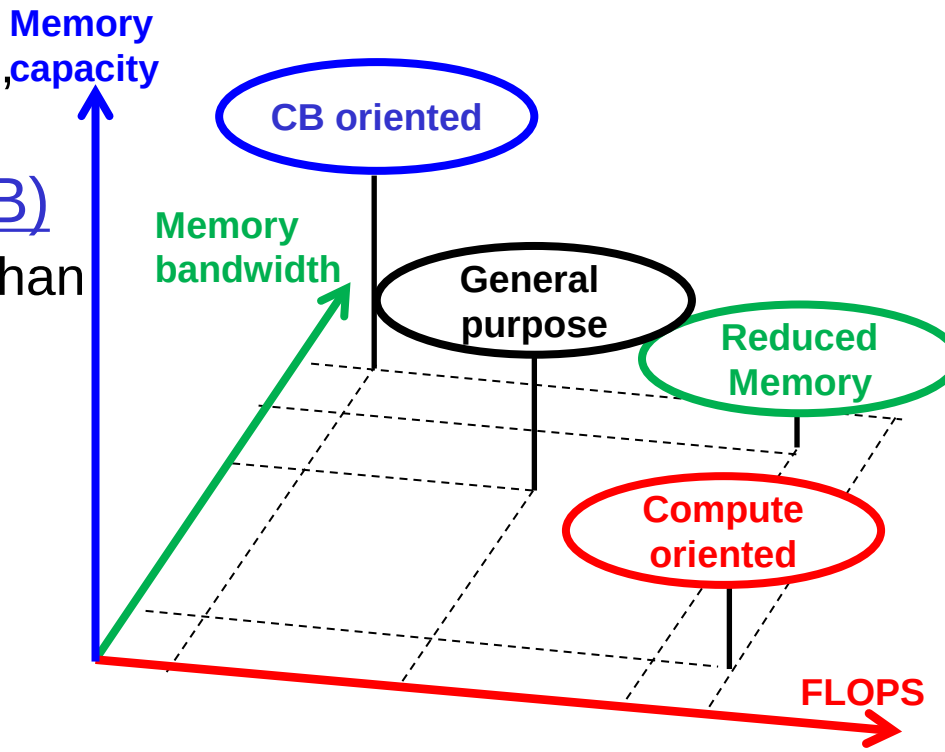


- Interconnection Network
  - Not enough analysis has been carried out
  - Some applications need >1us latency and large bisection BW
- Storage
  - There is not so big demand

Source: Masaaki Kondo's presentation at IESP Kobe meeting, 2012

Candidate of the Post Peta-scale Architectures

- Four types of architectures are considered
  - General Purpose (GP)
    - Ordinary CPU-based MPPs
    - e.g.) K-Computer, GPU, Blue Gene, x86-based PC-clusters
  - Capacity-Bandwidth oriented (CB)
    - With expensive memory-I/F rather than computing capability
    - e.g.) Vector machines
  - Reduced Memory (RM)
    - With embedded (main) memory
    - e.g.) SoC, MD-GRAPE4, Anton
  - Compute Oriented (CO)
    - Many processing units
    - e.g.) ClearSpeed, GRAPE-DR

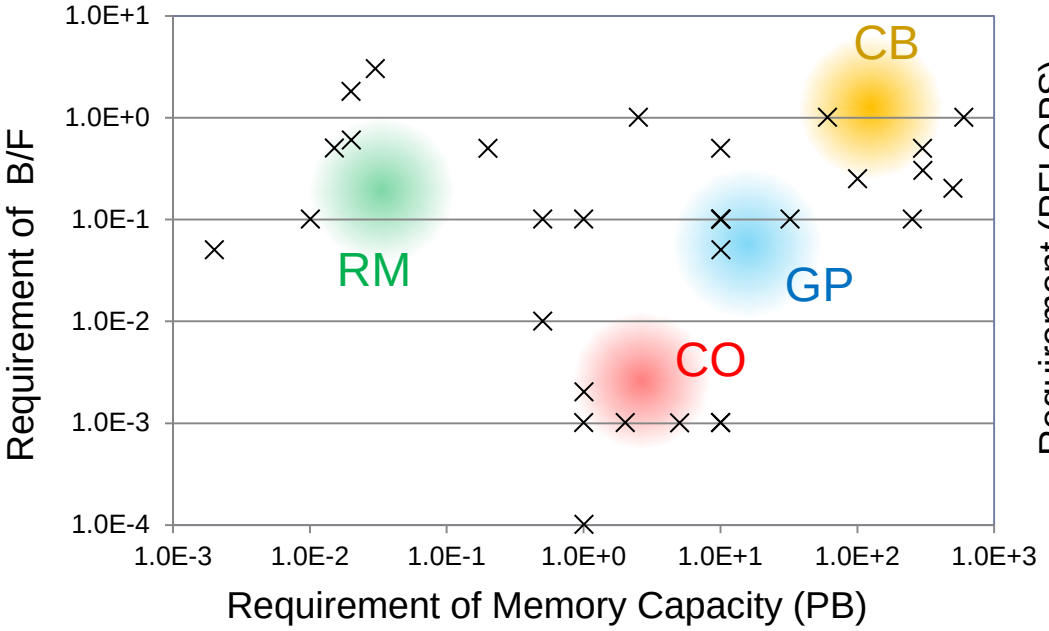


Source: Masaaki Kondo's presentation at IESP Kobe meeting, 2012

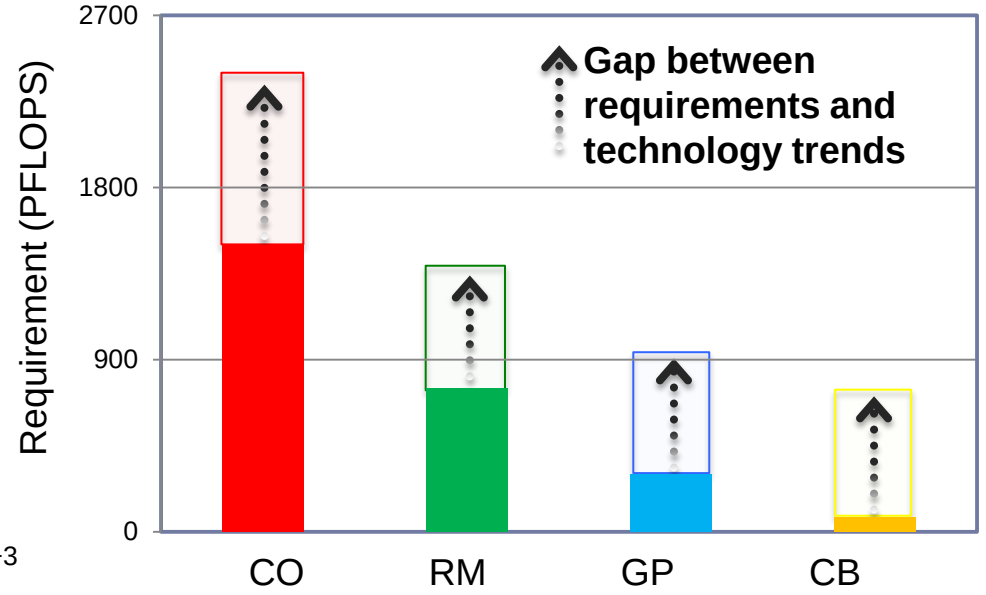
Gap Between Requirement and Technology Trends

- Mapping four architectures onto science requirement
- Projected performance vs. science requirement
  - Big gap between projected and required performance

Mapping of Architectures



Projected vs. Required Perf.

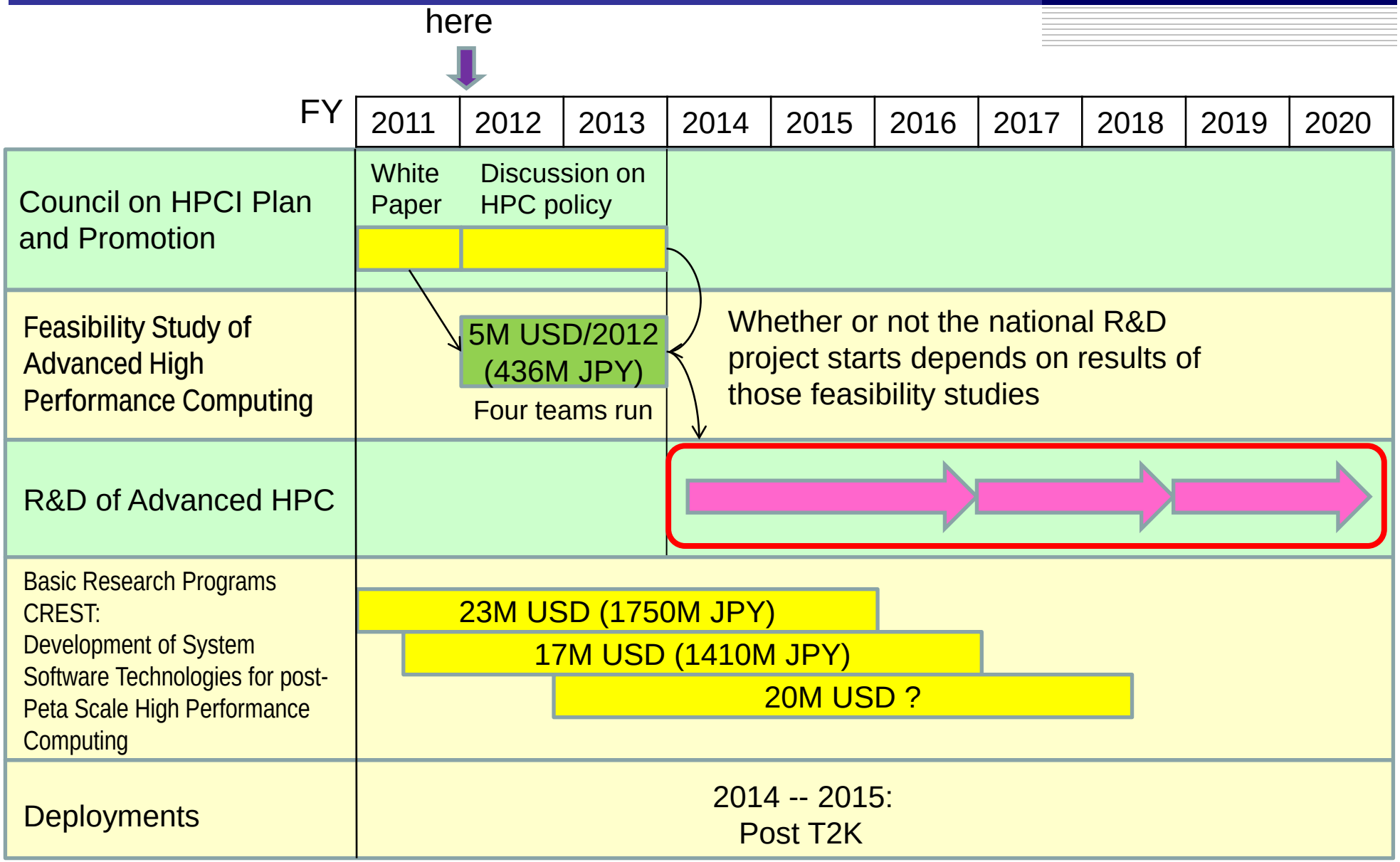


Needs national research project for science-driven HPC systems

GP (General Purpose), Capacity-Bandwidth oriented (CB), Reduced Memory (RM, Compute Oriented (CO)

Source: Masaaki Kondo's presentation at IESP Kobe meeting, 2012

Plans


here
FY
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Council on HPCI Plan and Promotion
White Paper
Discussion on HPC policy
Feasibility Study of Advanced High Performance Computing
5M USD/2012 (436M JPY)
Four teams run
Whether or not the national R&D project starts depends on results of those feasibility studies
R&D of Advanced HPC
Basic Research Programs
CREST:
Development of System Software Technologies for post-Peta Scale High Performance Computing
23M USD (1750M JPY)
17M USD (1410M JPY)
20M USD ?
Deployments
2014 -- 2015:
Post T2K

Feasibility Study on Future HPC R&D in Japan

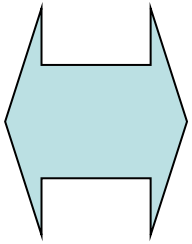
Program promotion board

Member: The head of each team and other specialists
Role: To check the progress of the each team and to coordinate the collaboration among the teams

1 application study team

RIKEN AICS and TITECH
Collaboration with application filelds

- Identification of scientific and social issues to be solve in the future
- Drawing Science road map until 2020
- Selection of the applications that plays key roles in the roadmap
- Review of the architectures using those applications



3 system study teams

Tohoku Univ. and NEC

U. of Tsukuba, Titech, and Hitachi

U. of Tokyo, Kyushu U., Fujitsu, Hitachi, and NEC

- Design of computer systems solving scientific and social issues
- Identification of R&D issues to realize the systems
- Review of the system using the application codes
- Estimation of the system's cost

Outline of This Talk

- Introduction of Key Organizations/Programs/Activities
  - HPCI
    - Innovative High Performance Computing Infrastructure
      - Seamless access to K computer, supercomputers, and user's machines
  - SPIRE
    - Strategic Programs for Innovative Research
- Feasibility Study for future HPC in Japan
  - Background
  - Introduction of “Feasibility study on advanced and efficient latency core-based architecture for future HPCI R&D”
- Post T2K

Towards Next-generation General Purpose Supercomputer

Approach:
- ✓Material and Climate Sciences are the first target applications
- ✓Approach from evolution of the K architecture
- ✓System Software Stack is designed for both the proposed machine and commodity-based machines

PI: Yutaka Ishikawa, U. of Tokyo
- Organization
- System Software Stack
- Performance Prediction and Tuning

Co-PI: Yuichi Nakamura, NEC
- System Software Stack

Applications

System Software Stack
(MPI, parallel file I/O, PGAS, Batch Job Scheduler, Debugging and Tuning Tools)

Commodity-based Supercomputer

Next-Gen General Purpose Supercomputer

Co-PI: Tsuneo Iida, Hitachi
- Storage Architecture and System Software Stack

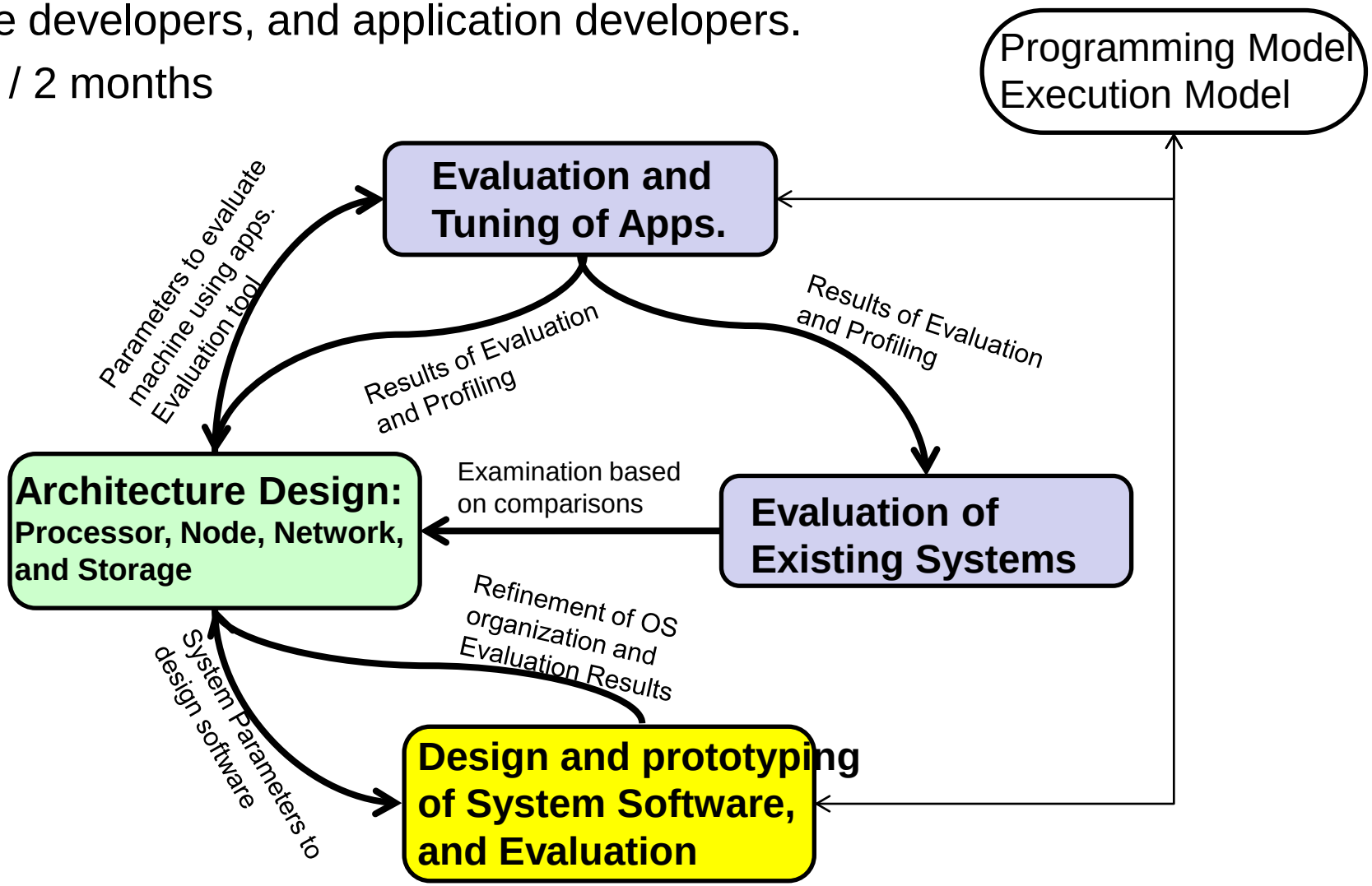
Co-PI: Kei Hiraki, U. of Tokyo
- Architecture Evaluation, Compiler, and Low power technologies

Co-PI: Mutsu Aoyagi, Kyushu U.
- Network Evaluation Environment

Co-PI: Naoki Shinjo, Fujitsu
- Processor, Node, Interconnect Architecture and System Software Stack

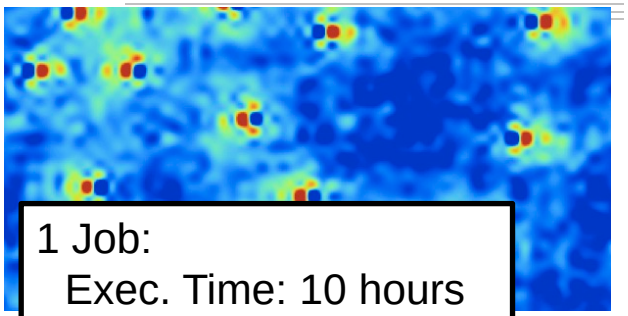
Co-design

- Tightly coupled design of architecture by architects, software developers, and application developers.
- 1 Cycle / 2 months

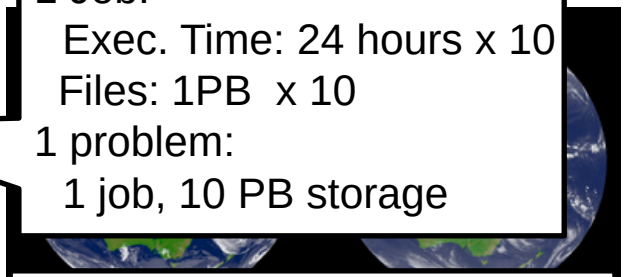


Part of Target Applications in FY2012

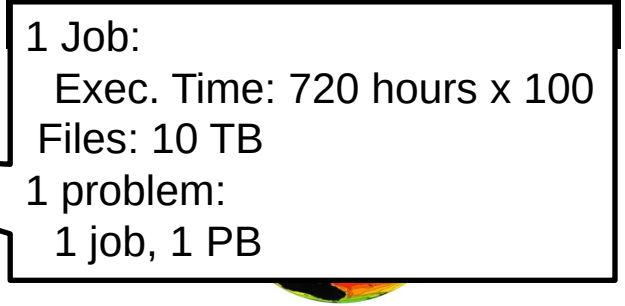
- ALPS(Algorithms and Libraries for Physics Simulations)
  - Providing high-end simulation codes for strongly correlated quantum mechanical systems
  - **Total Memory: 10～100PB, low latency and high radix network**
- RSDFT (Real-Space Density Functional Theory)
  - A DFT(Density Functional Theory) code with real space discretized wave functions and densities for **molecular dynamics** simulations using the Car-Parrinello type approach
  - **Total Memory: 1PB, Performance: 1EFLOPS(B/F 0.1)**
- NICAM (Nonhydrostatic ICosahedral Atmospheric Model)
  - A Global Cloud Resolving Model (GCRM)
  - **Total Memory: 140 TB, Memory Bandwidth: 300 PB/sec, Performance: 700 PFLOPS(B/F = 0.4)**
- COCO (CCSR Ocean Component Model)
  - ocean general circulation model developed at Center for Climate System Research (CCSR), the University of Tokyo
  - **Total Memory: 320 TB, Memory Bandwidth: 150 PB/sec, Performance: 50 PFLOPS (B/F = 3)**



1 Job:
Exec. Time: 10 hours
Files: 500TB
1 problem:
10 jobs, 5 PB storage



1 Job:
Exec. Time: 24 hours x 10
Files: 1PB x 10
1 problem:
1 job, 10 PB storage

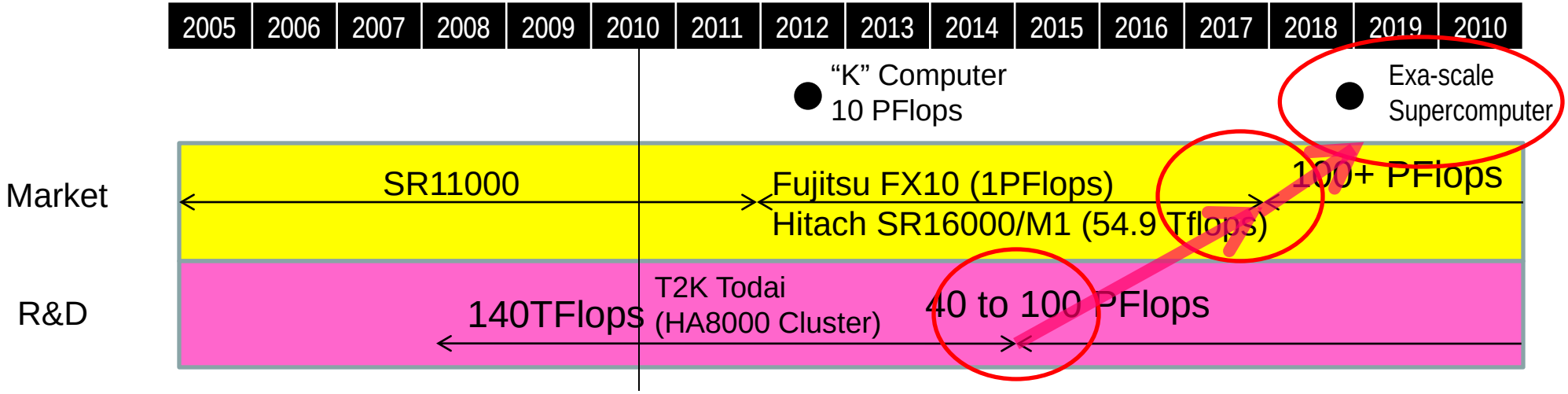


1 Job:
Exec. Time: 720 hours x 100
Files: 10 TB
1 problem:
1 job, 1 PB

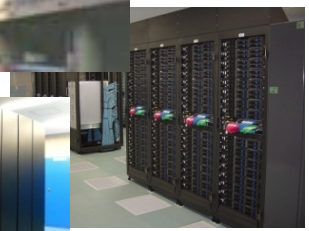
Outline of This Talk

- Introduction of Key Organizations/Programs/Activities
  - HPCI
    - Innovative High Performance Computing Infrastructure
      - Seamless access to K computer, supercomputers, and user's machines
  - SPIRE
    - Strategic Programs for Innovative Research
- Feasibility Study for future HPC in Japan
  - Background
  - Introduction of “Feasibility study on advanced and efficient latency core-based architecture for future HPCI R&D”
- Post T2K

Post T2K



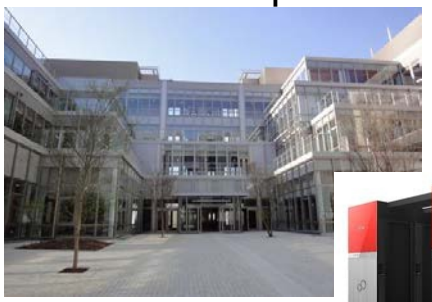
Hongo Campus



SR16000/M1

HA8000

Kashiwa Campus

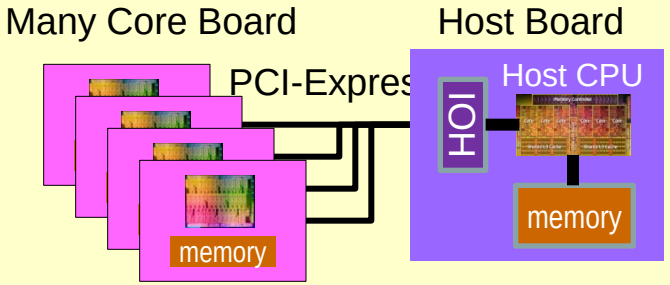


FX10

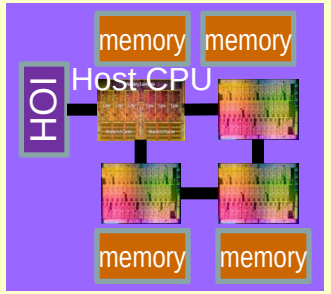
- PRIMEHPC FX10
  - 4800 Node (16 core/node)
  - 1.13 PFlops
  - 150 TB main memory
- Hitachi SR16000/M1
  - 56 Node (32 Core/Node)
  - 54.9 TFlops
  - 11200 GB main memory

Variations of Many-core based machines

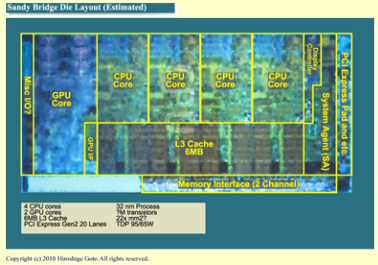
Many-core board connected to PCI-Express
e.g., Intel Knights Ferry, Knights Corner



Many-core chip connected to system bus
Not existing so far

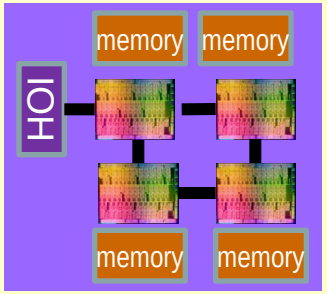


Many-core inside CPU die
c.f., Intel Sandy Bridge with GPU



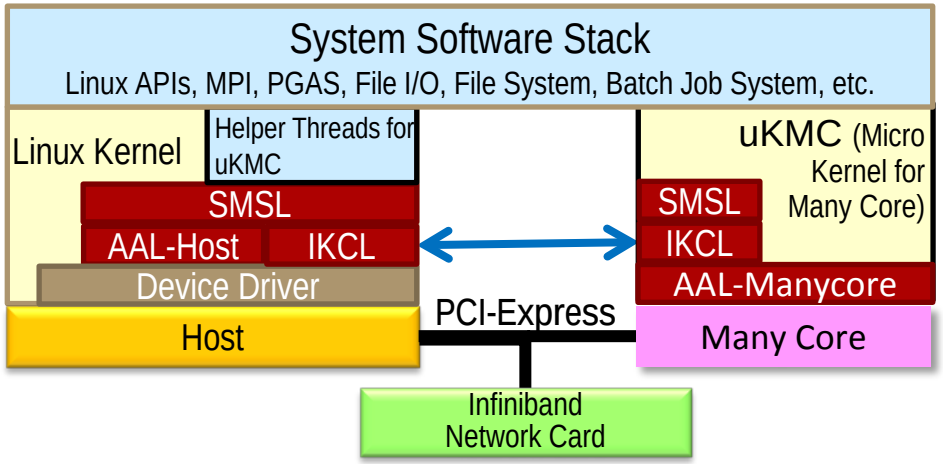
http://pc.watch.impress.co.jp/docs/column/kaigai/20100412_360173.html

Many-core only
Not existing so far

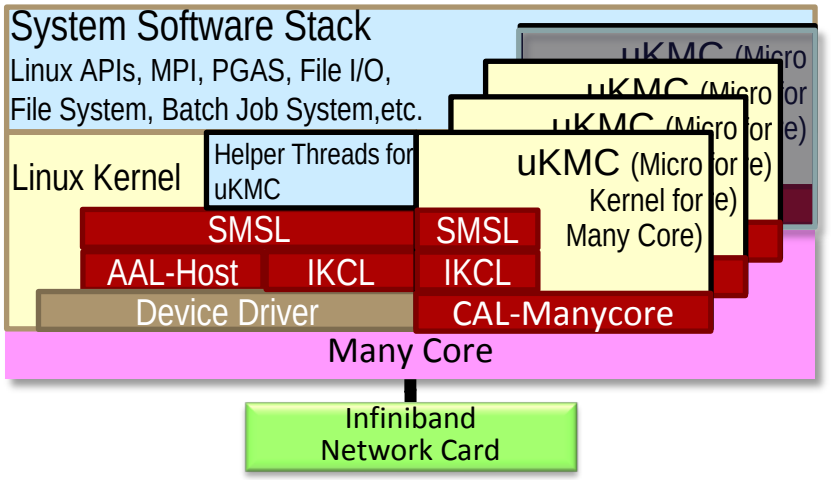


System Software Stack

In case of Non-Bootable Many Core



In case of Bootable Many Core



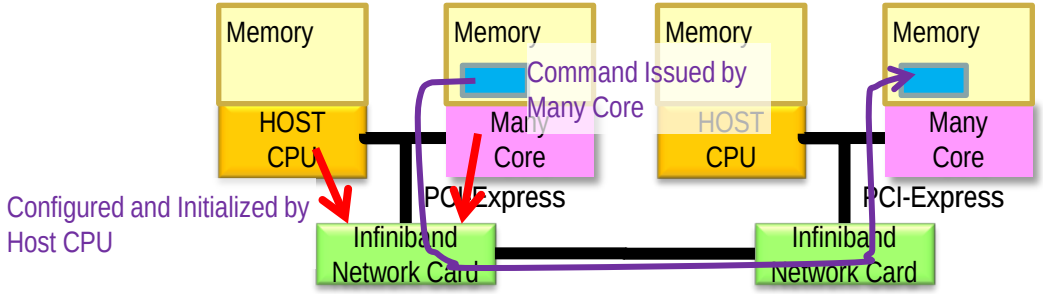
Design Criteria

- Cache-aware system software stack
- Scalability
- Minimum overhead of communication facility
- Portability

- AAL (Accelarator Abstraction Layer)
  - Provides low-level accelerator interface
  - Enhances portability of the micro kernel
- IKCL (Inter-Kernel Communication Layer)
  - Provides generic-purpose communication and data transfer mechanisms
- SMSL (System Service Layer)
  - Provides basic system services

DCFA: Direct Communication Facility for Accelerator

- Limitations of a PCI-Express device
  - cannot configure another device such as a communication device, and thus it does not know the other device address.
  - Cannot receive interrupts from other devices.
- DCFA
  - The host configures and initializes an Infiniband HCA, and informs the HCA address to an MIC device so that it may issue commands to that device
  - The MIC device directly accesses the Infiniband HCA registers

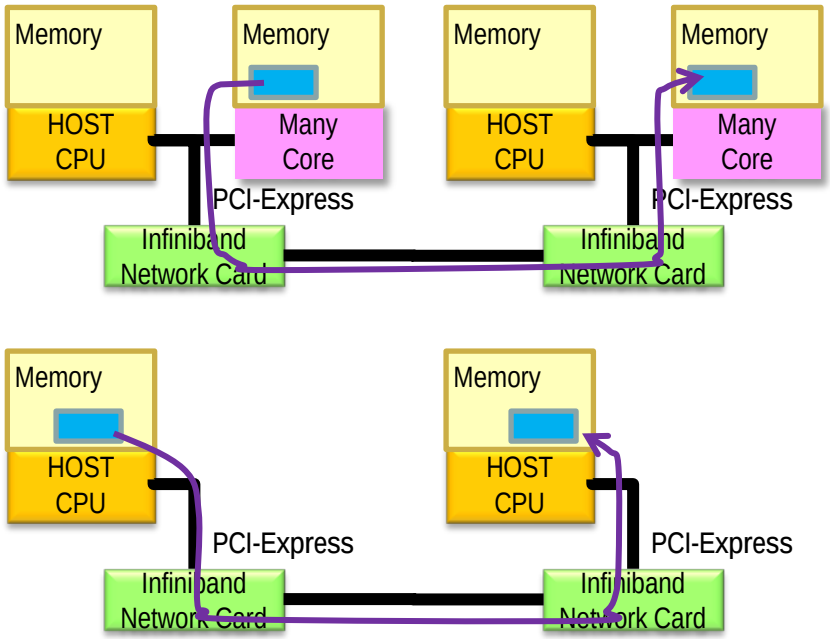


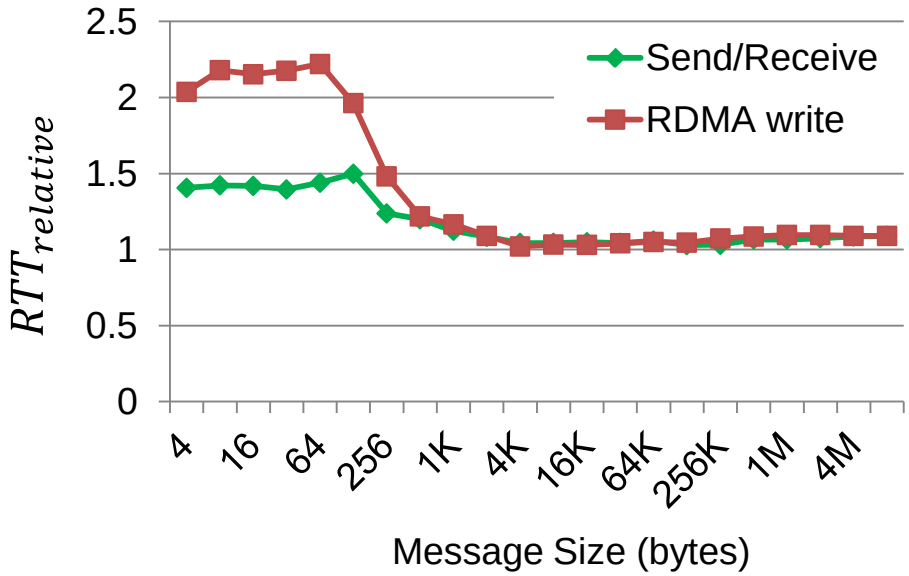
Min Si and Yutaka Ishikawa, "Design of Direct Communication Facility for Manycore-based Accelerators, " to appear at CASS2012 in conjunction with IPDPS2012.

Latency

- The same performance as that of host to host data transfer for large message size

$$RTT_{relative} = {}^{RTT_{DCFA}}/_{RTT_{HOST}}$$

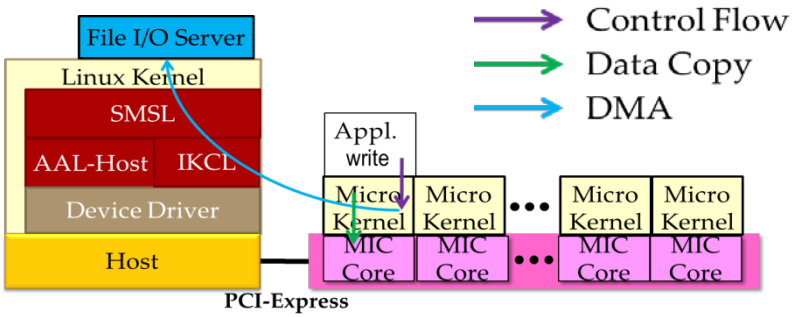




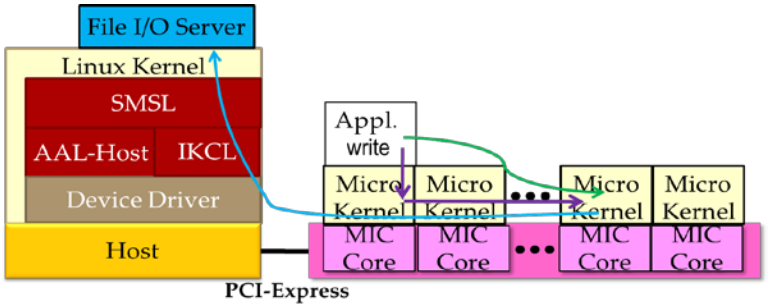
Intel Xeon X5680 3.33GHz x 2
Mellanox MT26428
Knights Ferry

Design Considerations of File I/O System

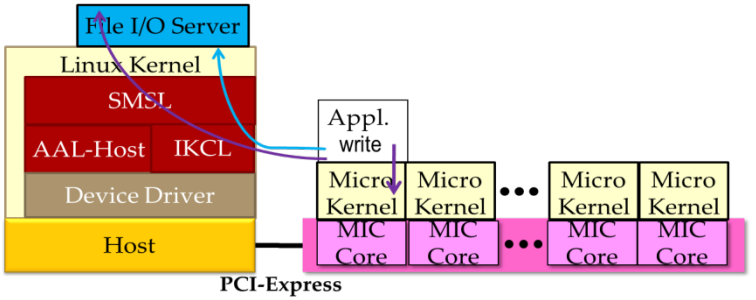
- File I/O functions run on computing core in MIC



- File I/Os are delegated to the OS-dedicated core in MIC



- File I/Os are delegated to the host OS



Yuki Matsuo, Taku Shimosawa, and Yutaka Ishikawa, "A File I/O System for Many-Core Based Clusters," in conjunction with ICS2012, 2012.

Performance Differences

Iterative

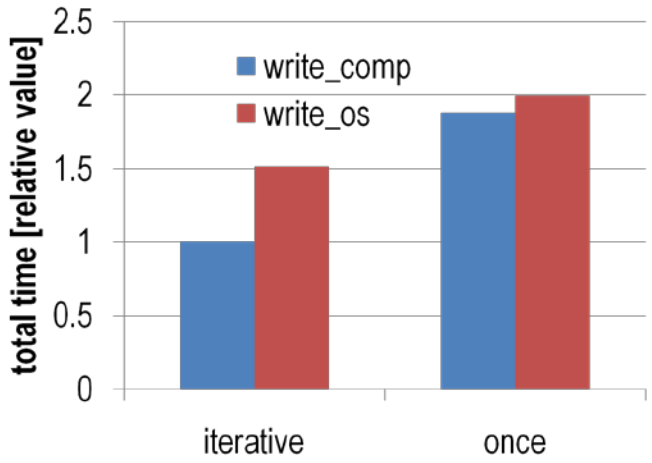
```
size = 64KB;
for(n = 0;  n < DIVISOR;  n++) {
   for(i = 0;  i < size/4;  i++)  buf[i] = n;
   write(fd, buf, size);
}
```

Once

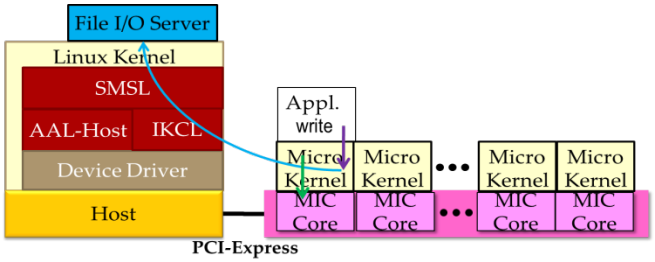
```
size = 64KB; j = 0;
for(n = 0;  n < DIVISOR;  n++) {
   for(i = 0;  i < size/4;  i++) buf[j++] = n;
}
write(fd, buf, size*DIVISOR);
```

In the iterative simple benchmark, the write system call is issued during the user data is located on L2 cache.

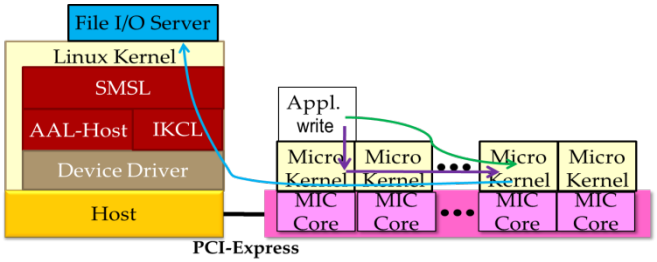
Relative Total Execution Time



write_comp



write_os



Summary

- Building Nation-wide infrastructure and collaboration structure
  - HPCI: Innovative High Performance Computing Infrastructure
  - SPIRE: Strategic Programs for Innovative Research
- Starting Feasibility Study for future HPC in Japan
  - 1 application and 3 architecture teams have been selected
- Studying OS mechanisms for Post T2K
  - A manycore-based cluster has been considered and been studied