High Performance Computing Technology Toward Exascale Simulation

Taisuke Boku

Center for Computational Sciences

University of Tsukuba

Outline

- The K Computer
- System overview
- HPCI (HPC Infrastructure)
- SDHPC (Strategic Development of HPC systems) and cooperation with application groups
- Feasibility Study toward Exascale
- Activity in U. Tsukuba
  - HA-PACS
  - TCA

The K Computer

World Fastest computer “K” (Jun. 2011)

K computer

Full operation for public utilization starts in 2012

- Latest spec. of K Computer (as on Nov. 2011)
  - Computation nodes (CPUs): 88,128
  - Total core#: 705,024
  - Peak perforance: 11.28 PFLOPS Linpack: 10.51 PFLOPS
  - Memory capacity: 1.41 PB (16GB/node)

Nicknamed the "K computer"

Kei (京) represents the numerical unit of 10 Peta (10^{16}) in the Japanese language, representing the system's performance goal of 10 Petaflops. The Chinese character 京 can also be used to mean " a large gateway" so it could also be associated with the concept of a new gateway to computational science.

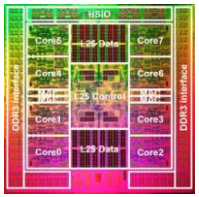
一、	十、	百、	千、	万、	億、	兆、	京、	垓、	杼、	穰、	溝、	澗、	正、	載、	極、
10^{0}	10^{1}	10^{2}	10^{3}	10^{4}	10^{8}	10^{12}	10^{16}	10^{20}	10^{24}	10^{28}	10^{32}	10^{36}	10^{40}	10^{44}	10^{48}

恒河沙、	阿僧祇、	那由他、	不可思議、	無量大数
10^{52}	10^{56}	10^{60}	10^{64}	10^{68}

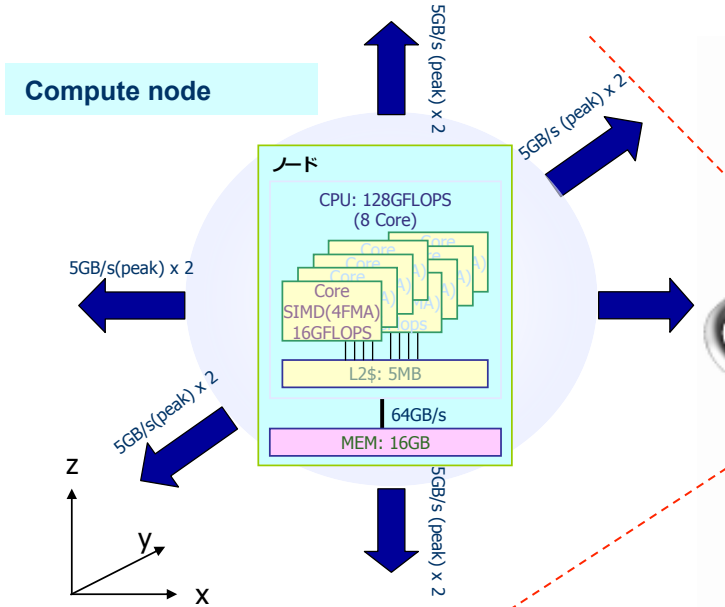
K computer: compute nodes and network

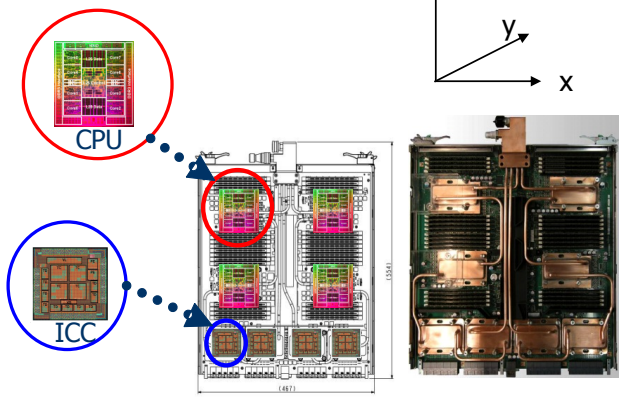
- Computation nodes (CPUs): 88,128
- Total core#: 705,024
- Peak perforance: 11.28 PFLOPS
  Linpack: 10.51 PFLOPS
- Memory capacity: 1.41 PB (16GB/ node)

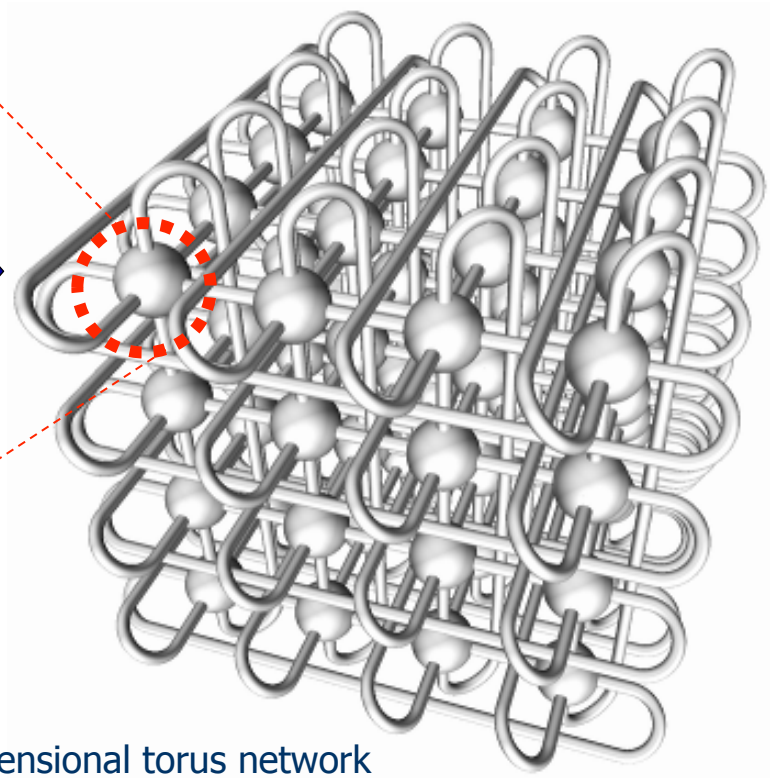
- Logical 3-dimensional torus network
- Peak bandwidth: 5GB/s x 2 for each direction of logical 3-dimensional torus network
- bi-section bandwidth: > 30TB/s



SPARC64™ VIIIfx







Logical 3-dimensional torus network for programming

Courtesy of RIKEN/FUJITSU Ltd.

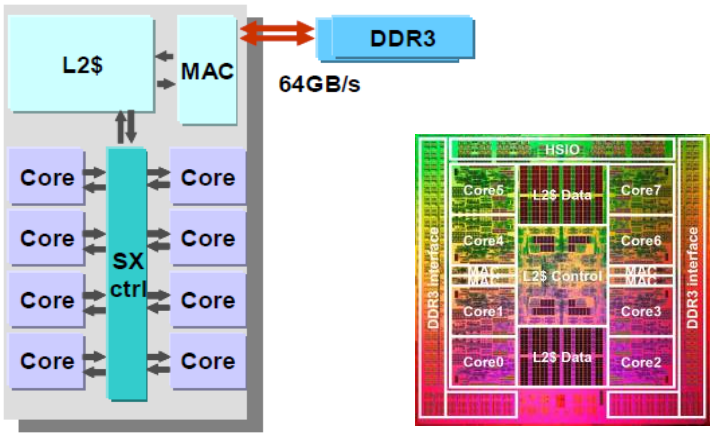
CPU Features (Fujitsu SPARC64™ VIIIfx)

- 8 cores
- 2 SIMD operation circuit
  - 2 Multiply & add floating-point operations (SP or DP) are executed in one SIMD instruction

16GF/core(2*4*2G)

- 256 FP registers (double precision)

- Shared 5MB L2 Cache (10way)
  - Hardware barrier
  - Prefetch instruction
  - Software controllable cache
    - Sectored cache

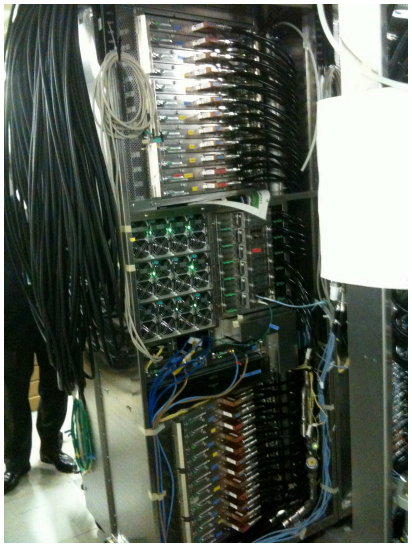


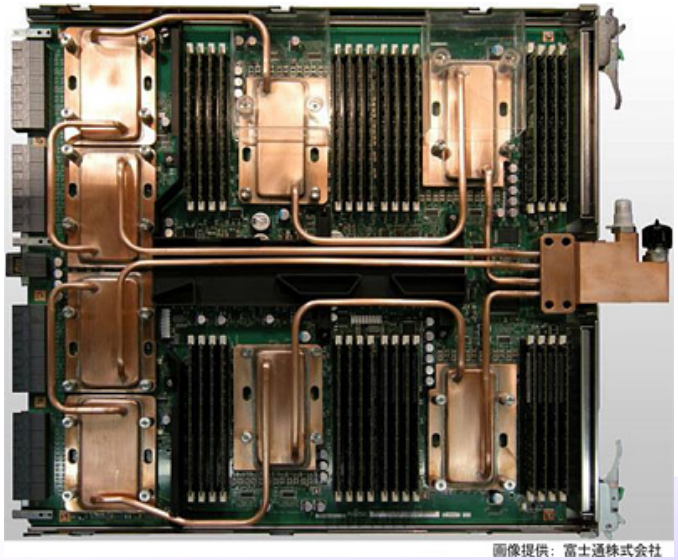
- Performance
  - 16GFLOPS/core, 128GFLOPS/CPU

45nm CMOS process, 2GHz
22.7mm x 22.6mm
760 M transisters
58W（at 30℃ by water cooling)

Reference: SPARC64™ VIIIfx Extensions
http://img.jp.fujitsu.com/downloads/jp/jhpc/sparc64viiifx-extensions.pdf

The K Computer





HPCI
~ Nation-wide High Performance Computing Infrastructure ~

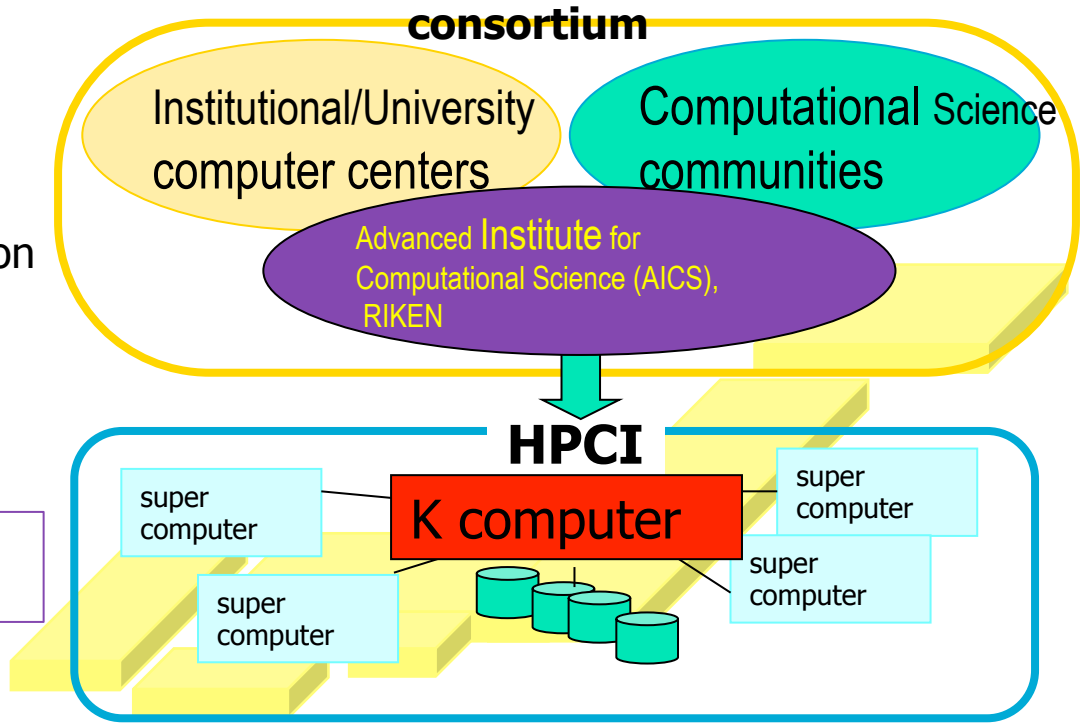
What is HPCI ?

- Nation-wide High Performance Computing Infrastructure in Japan
- Gathering main supercomputer resources in Japan including “K” and all universities’ supercomputer centers
- Grid technology to enable “single sign-on” utilization of any supercomputers in Japan
- Ultra large scale distributed file system in two sites (West and East) as data pool to be accessed from any supercomputer resources under HPCI
- Science-driven project base proposals to utilize HPCI resources are submitted, then free CPU budget is decided

Formation of HPCI

- Background:
  - After re-evaluation of the project at "government party change" in 2011, the NGS project was restarted as "Creation of the Innovative High Performance Computing Infra-structure (HPCI)".
- Building HPCI: High-Performance Computing Infrastructure
  - To establish a hierarchical organization of supercomputers linked with the K computer and other supercomputers at universities and institutes
  - To set up a large-scale storage system for the K computer and other supercomputers
- Organizing HPCI Consortium
  - To play a role as the main body to design and operate HPCI.
  - To organize computational science communities from several application fields and institutional/university supercomputer centers.
    - Including Kobe Center

The preparatory consortium has been organized



(slide is courtesy by M. Sato, U. Tsukuba)

HPCI Preparatory Consortium Members

User Communities (13)

- RIKEN
- Computational Materials Science Initiative
- Japan Agency for Marine-Earth Science and Technology
- Institute of Industrial Science at University of Tokyo
- Joint Institute for Computational Fundamental Science
- Industrial Committee for Super Computing Promotion
- Foundation for Computational Science
- BioGrid Center Kansai
- Japan Aerospace Exploration Agency
- Center for Computational Science & e-Systems, Japan Atomic Energy Agency
- National Institute for Fusion Science
- Solar-Terrestrial Environment Laboratory, Nagoya University
- Kobe University

Resource Providers (25)

- Information Initiative Center, Hokkaido University
- Cyberscience Center, Tohoku University
- Information Technology Center, University of Tokyo
- Information Technology Center, Nagoya University
- Academic Center for Computing and Media Studies, Kyoto University.
- Cybermedia Center, Osaka University
- Research Institute for Information Technology, Kyushu University
- Center for Computational Sciences, University of Tsukuba
- Global Scientific Information and Computing Center, Tokyo Institute of Technology
- Institute for Materials Research, Tohoku University
- Institute for Solid State Physics, University of Tokyo
- Yukawa Institute for Theoretical Physics, Kyoto University
- Research Center for Nuclear Physics, Osaka University
- Computing Research Center, KEK(High Energy Accelerator Research Organization)
- National Astronomical Observatory of Japan
- Research Center for Computational Science, Institute for Molecular Sciencce
- The Institute of Statistical Mathematics
- JAXA's Engineering Digital Innovation Center
- The Earth Simulator Center
- Information Technology Research Institute, AIST
- Center for Computational Science & e-Systems, Japan Atomic Energy Agency
- Advanced Center for Computing and Communication, RIKEN
- Advanced Institute for Computational Science
- National Institute of Informatics
- Research Organization for Information Science & Technology

(slide is courtesy by M. Sato, U. Tsukuba)

AICS and Supercomputer Centers in Japanese Universities

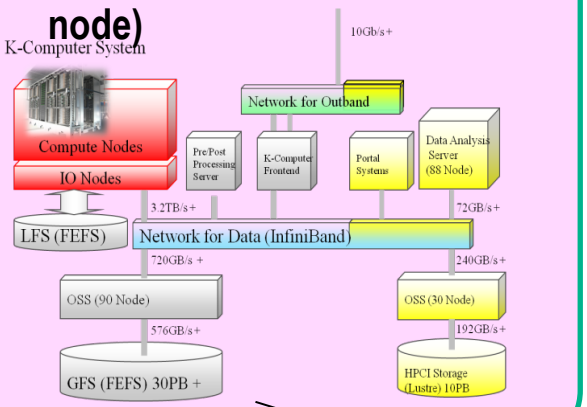
(slide is courtesy by M. Sato, U. Tsukuba)

Storage System in first phase for HPCI

HPCI WEST HUB

AICS, RIKEN

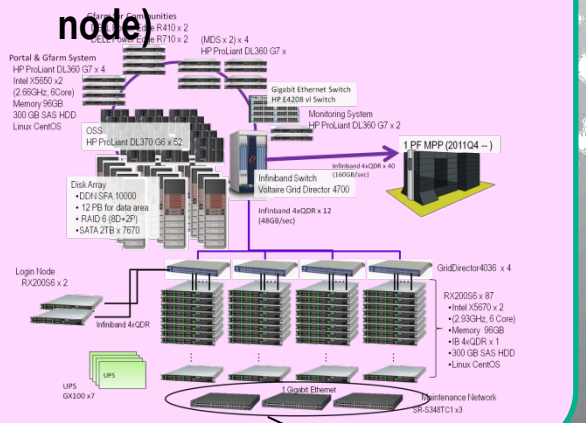
- 10 PB storage (30 OSS)
- Cluster for data analysis (88 node)



HPCI EAST HUB

University of Tokyo

- 12 PB storage (52 OSS)
- Cluster for data analysis (87 node)



Gfarm2 is used as the global shared file system

(slide is courtesy by M. Sato, U. Tsukuba)

HPCI operation schedule

- Originally scheduled to start the operation from autumn of 2012
- Basic system construction and experiments started from April 2011
- Currently, partial operation starting date is under consideration with a half year earlier than original plan
  ⇒ Sept. 2012

Strategic Study toward Exascale Computing in Japan

Governmental activities on HPC

- MEXT organized a working group, discussing future development of HPC technologies, under the “Council for HPCI Plan and Promotion”

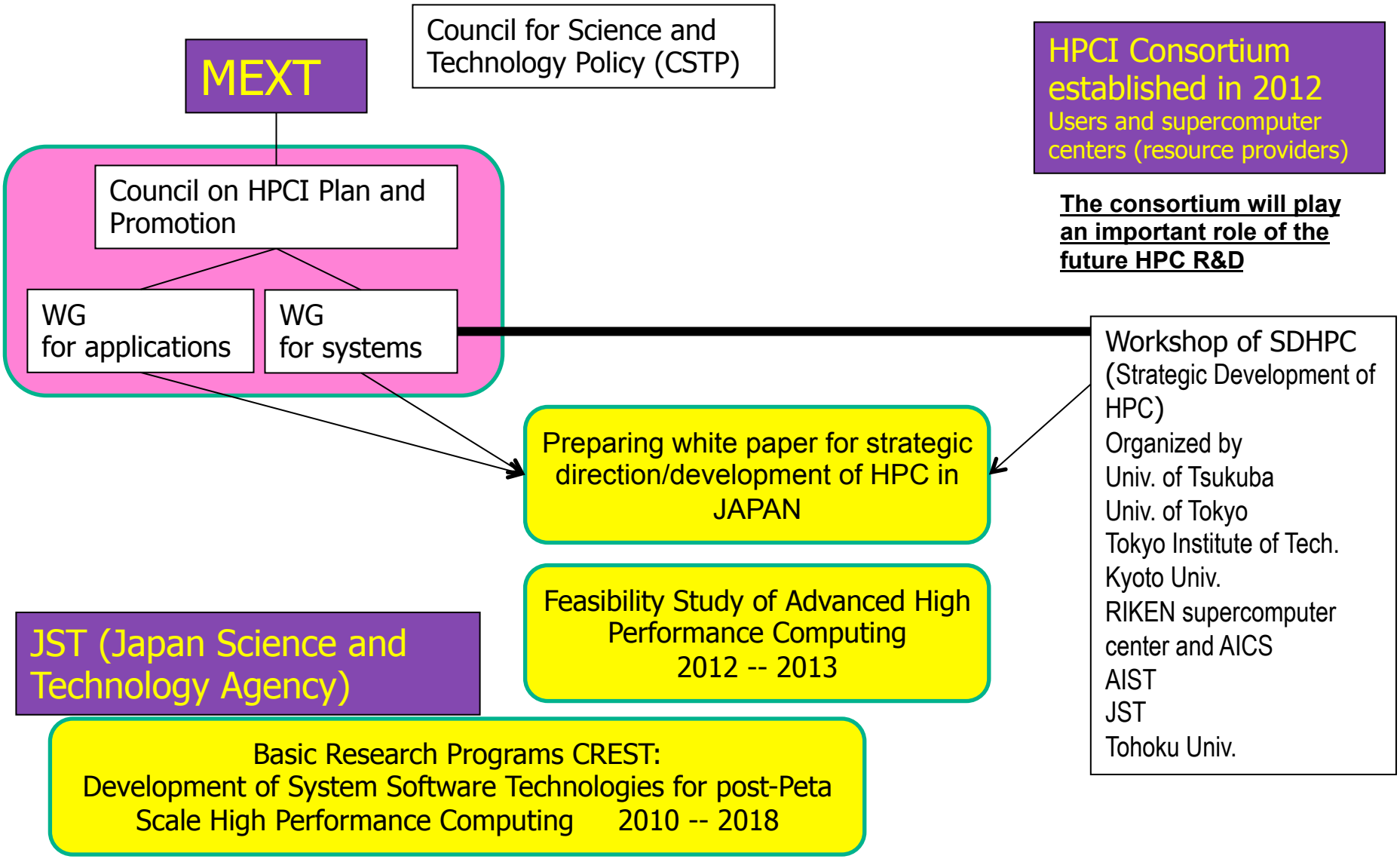
In other words, demonstrating scientific and technological values in exa-scale computing is very important for tax payers

The WG advised

- The first step should be discussion on how HPC technologies are applied to science and technology fields
- The system should be co-designed and developed with applications, computer architectures, and system software
- Several application-oriented computer architectures should be considered
- Prior to developing an exa-scale machine, tens to hundreds peta-scale machines should be deployed
- Strategic development, which pieces are developed in Japan and which pieces are internationally collaborated, should be determined
- The development of novel computational models and algorithms is also important
- The new application fields should be investigated

(slide is courtesy by Y. Ishikawa, U. Tokyo)

Players and Projects in Japan



(slide is courtesy by Y. Ishikawa, U. Tokyo)

SDHPC

- White paper for Strategic Direction/Development of HPC in JAPAN is now being written by young Japanese researchers with advisers (seniors)
- The white paper will be approved by the Council for HPCI Plan and Promotion by the end of FY 2011.
- Contents (draft)
  - Expected scientific and technological values in exa-scale computing
  - Challenges towards exa-scale computing
  - Current research efforts in Japan and other countries
  - Approaches
  - Competitive vs Collaboration
  - Plan of R&D and organization
- The white paper will be used for call for proposals in “Feasibility Study of Advanced High Performance Computing”

(slide is courtesy by Y. Ishikawa, U. Tokyo)

SDHPC working teams

- Categories for white paper work
  - Architecture
  - System Software
  - Language, Compiler & Run-Time System
  - Math Libraries
- Work is on-going…
  - (discussion for white paper completion)

Feasibility Study of Advanced High Performance Computing

- MEXT (Ministry of Education, Culture, Sports, Science, and Technology) has just proposed two-year project for feasibility study of advanced HPC which will start in FY2012 (April of 2012)
- Objectives
  - The high performance computing technology is an important national infrastructure
  - Keeping development of top-level HPC technologies is one of Japanese international competitiveness and contributes national security and safety
  - This two-year project is to study feasibilities of such development that Japanese community should focus on
- Budget requested
  - Approximately US$ 10M / year ⇒ US$9M

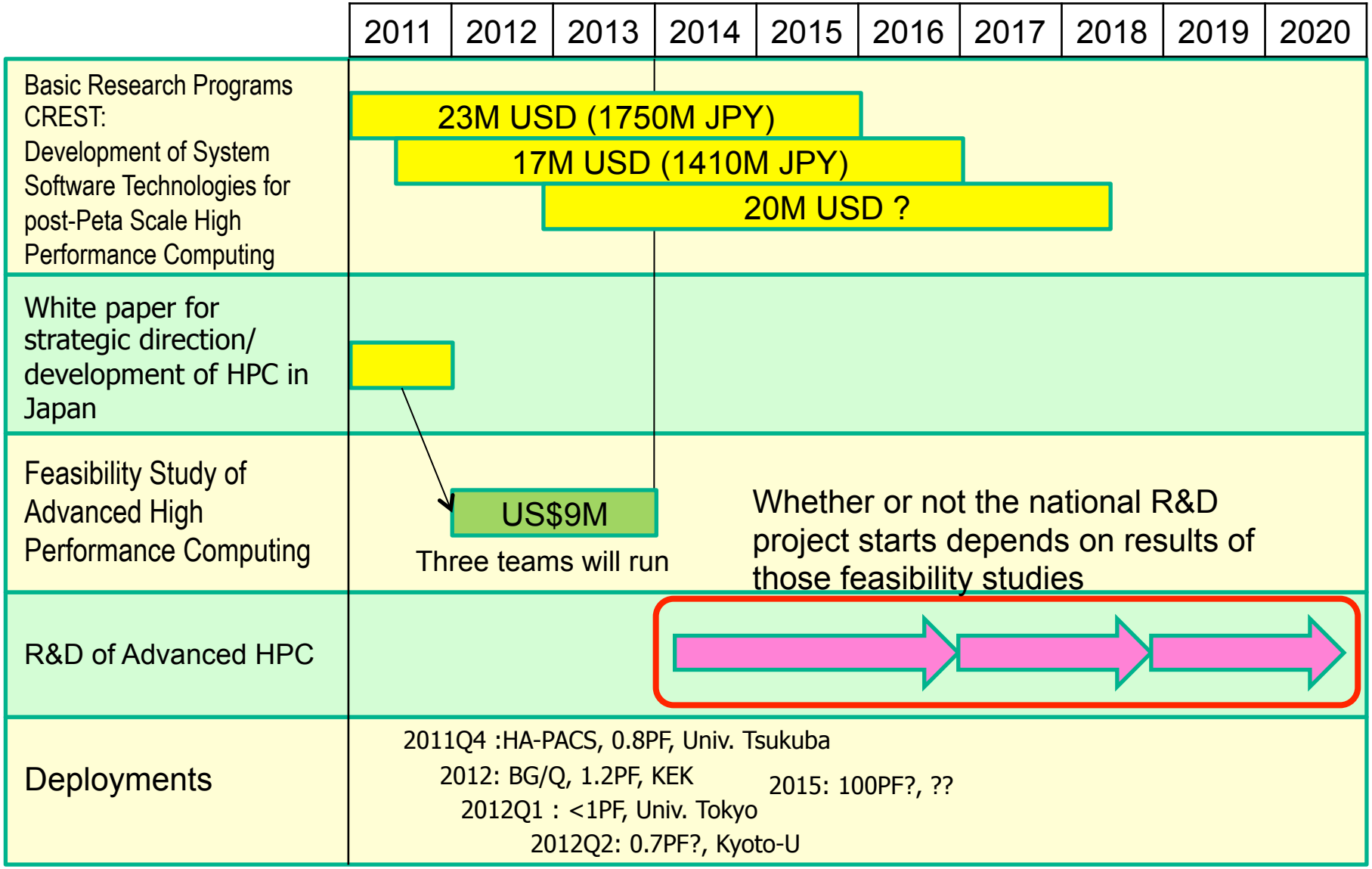
(slide is courtesy by Y. Ishikawa, U. Tokyo)

Feasibility Study of Advanced High Performance Computing

- Plan
  - Several computer architectures will be selected based on their potential advantages in solving respective key socio-scientific problems in Japan Japan
  - For each system
    - Hardware trends are investigated
    - The system architecture and system software are designed (with applications and, its prototype system is developed)
  - As a result of this study, further R&D for Japanese HPC funded by the government is decided
- Organizations
  - Several R&D teams (maybe three) will be selected.

(slide is courtesy by Y. Ishikawa, U. Tokyo)

Plans



(slide is courtesy by Y. Ishikawa, U. Tokyo)

Keyword: *Co-Design*

- HPC and application development so far
  - "Best hardware & software is made by HPC people"
    ⇒ Applications are developed/ported to the system after completion
  - In Exascale era, there is no room to depend with each other
    ⇒ Hardware design parameters are limited (quantitatively & qualitatively)
    ⇒ Specified by application area with "science driven approach"
- Co-design is the word for "application"
  - Application researcher's role is very important
  - They are not "guests" any more
  - Applications real request must be reflected to all the elements of the system

The hardest challeng: POWER

- Exascale system target power consumption
  = 25MW: limit which can be supported by the largest class of supercomputer center
- Today...
  - K computer: 10PFLOPS with 13MW ⇒ 0.77GF/W
  - BG/Q: 2GF/W
  - GPU cluster (HA-PACS): 1GF/W ?? (peak=2GF/W)
  - TARGET: 40GF/W (!!)
- Possible technology
  - Accelerated Computing (in any form)
  - Ultra low power process and communication devices
  - But just accelerators are not enough...
    Tightly coupled accelerators for massively parallel processing is required

Activity in University of Tsukuba

“HA-PACS” Project

HA-PACS Project in Univ. of Tsukuba

- HA-PACS (Highly Accelerated Parallel Advanced system for Computational Sciences)
  - 2011/4 – 2014/3 3years project
  - Base Cluster Unit
    - Large scale GPU cluster with latest standard CPUs and latest standard GPUs, especially employing advanced I/O bus technology to connect multiple GPUs with full speed
    - A platform for developing large scale application using latest GPUs and an system for production run of the applications
  - TCA (Tightly Coupled Accelerator)
    - Feasibility study for direct communication between GPUs
    - Develop a new network board (PEACH2) directly using PCIe Bus
    - Reduce the latency between GPUs

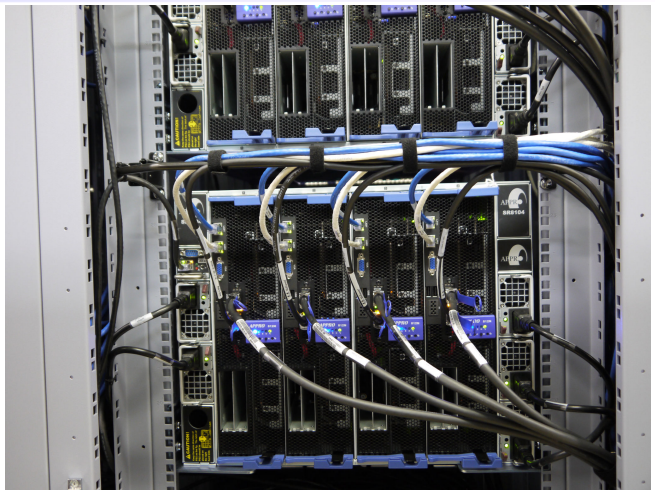
HA-PACS base cluster (Feb. 2012)

HA-PACS base cluster

Front view

Side view

HA-PACS base cluster

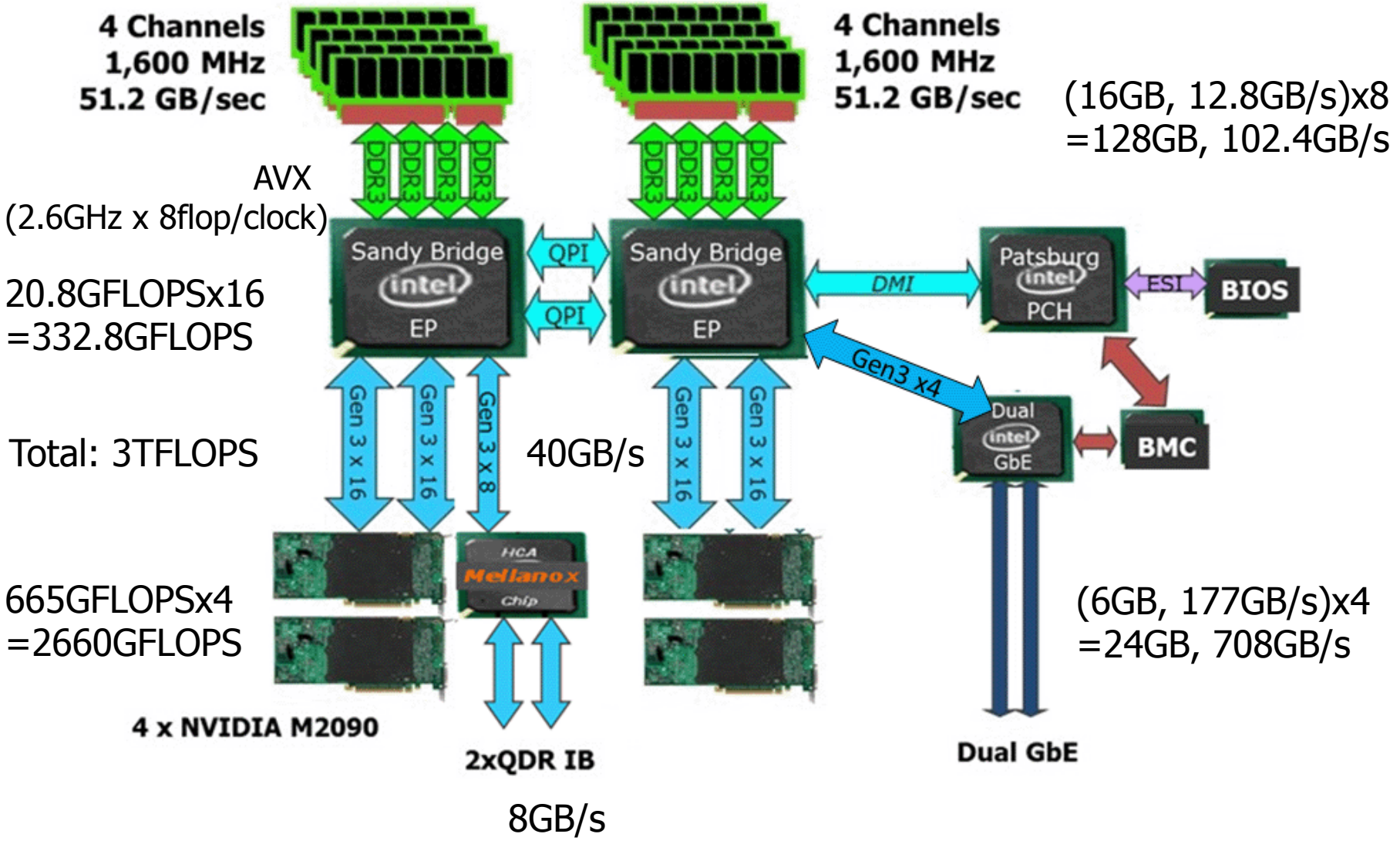


Rear view of one blade chassis with 4 blades

Front view of 3 blade chassis

Rear view of Infiniband switch and cables
(yellow=fibre, black=copper)

HA-PACS: Base Cluster Unit (computation node)



HA-PACS: Base Cluster Unit（Total）

- 268 nodes are connected by 2 IB switches
- CPU: 89TFLOPS + GPU: 713TFLOPS = total 802TFLOPS
- CPU: Memory size 34TByte, Bandwidth 27TByte/sec,
  GPU: Memory size 6.4TByte, Bandwidth 190TByte/sec
- Bisection bandwidth 2.1TByte/sec
- Storage User Area 504TByte
- Power Consumption 408kW (monitoring available）
  ⇒ 2 GF/W
- 26 racks (5.5m x 10m including maintenance area)
- Delivery: End of January 2012

HA-PACS: TCA (Tightly Coupled Accelerator)

- TCA: Tightly Coupled Accelerator
  - Direct connection between accelerators (GPUs)
  - Using PCIe as a communication device between accelerator
    - Most acceleration device and other I/O device are connected by PCIe as PCIe end-point (slave device)
    - An intelligent PCIe device logically enables an end-point device to directly communicate with other end-point devices
- PEARL: PCI Express Adaptive and Reliable Link
  - We already developed such PCIe device (PEACH, PCI Express Adaptive Communication Hub) on JST-CREST project “low power and dependable network for embedded system”
  - It enables direct connection between nodes by PCIe Gen2 x4 link

⇒ Improving PEACH for HPC to realize TCA

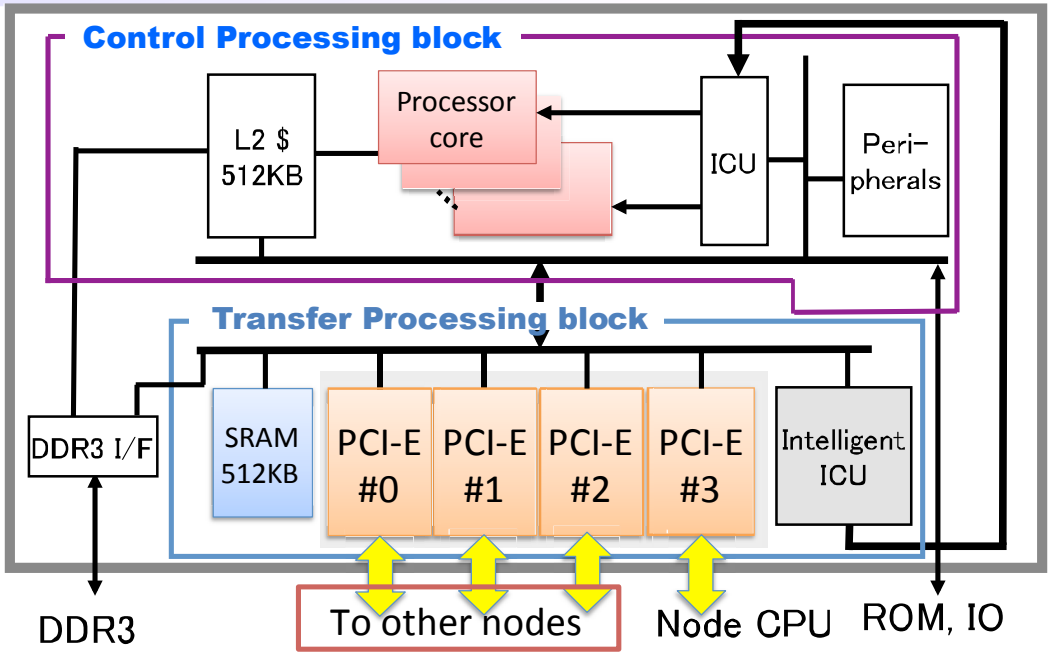
HA-PACS/TCA

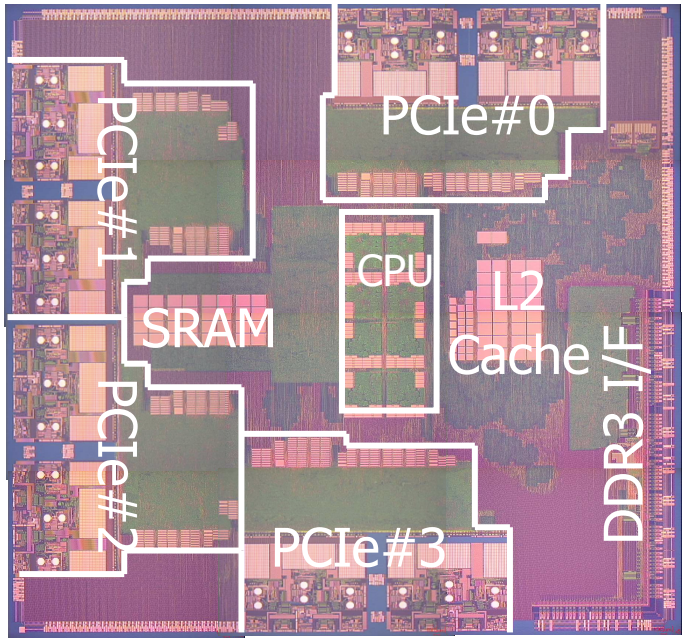
- Elementary technology research for new technology for advanced accelerated computing system
- Avoid the bottleneck on PCI-E connection between CPU & GPU, and interconnection network bottleneck between nodes
- PEARL : PCI-Express Adaptive and Reliable Link
  - Research supported by JST-CREST “DEOS” (dependable OS)
  - Applying PCIe link as “a direct link between nodes”
  - Short range communication (= high density) and make direct GPU-GPU connection to reduce the latency of GPU-GPU communication
  - 1st generation hardware prototype for PEARL link named “PEACH” was implemented by ASIC
  - PEACH includes 4-core M32R embedded CPU, Linux controlled, for 4-port PCI-E switch, and one of 4 ports will be a communication link with its host CPU

PEACH

- PEACH: PCI-Express Adaptive Communication Hub
- An intelligent PCI-Express communication switch to use PCIe link directly for node-to-node interconnection
- Edge of PEACH PCIe link can be connected to any peripheral devices, including GPU
- Prototype PEACH chip
  - 4-port PCI-E gen.2 with x4 lane / port
  - PCI-E link edge control feature: “root complex” and “end points” are automatically switched (flipped) according to the connection handling
  - Other fault-tolerant (reliability) function is implemented:
    “flip network link” to allow single link fault
- in HA-PACS/TCA prototype development, we will enhance current PEACH chip ⇒ PEACH2

PEACH chip (previous) [Otani et al., ISSCC2011]





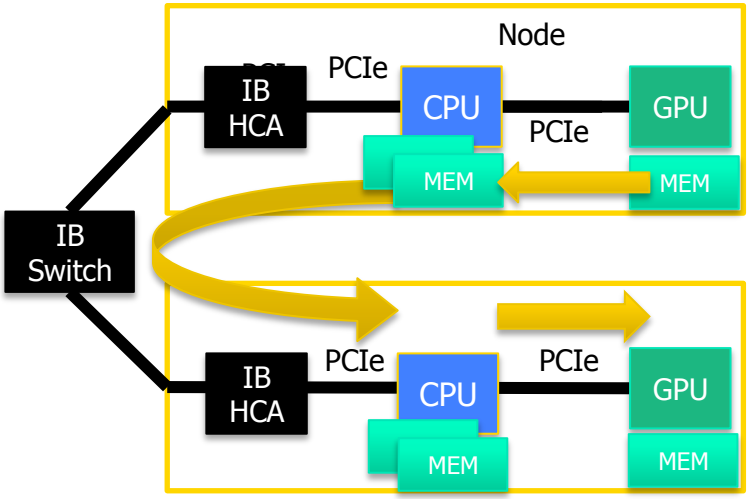
- **CPU: Renesas M32R 4core SMP (max. 400MHz)**
  - **small core size, low power**
  - **SMP**
  - **Controlling PCIe 4 port**
  - **health check for the node, link**
  - **communication link management**
  - **route reconfiguration**

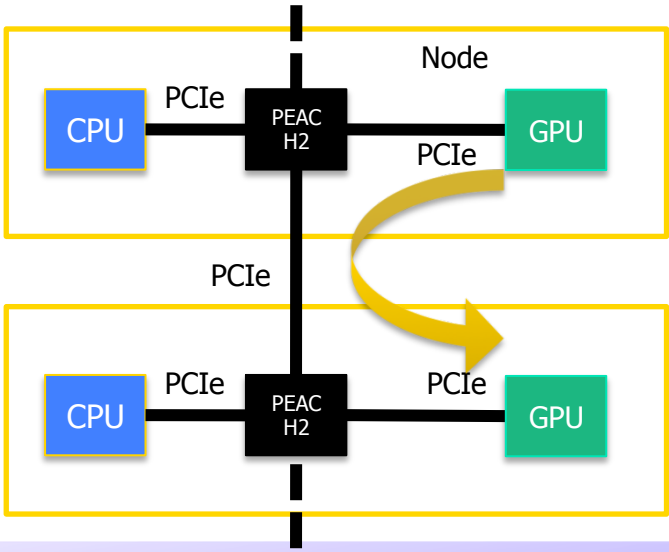
- comm. link:
  PCI Express Gen2 x4 lanes (20Gbps)
  x4 port
- SuperHyway, DMAC
- Low power
  - Controlling #lanes for each port
  - gen1/gen2 switching
  - core frequency control

HA-PACS/TCA (Tightly Coupled Accelerator)

- **True GPU-direct**
  - current GPU clusters require 3-hop communication (3-5 times memory copy)
  - For strong scaling, Inter-GPU direct communication protocol is needed for lower latency and higher throughput
- **Enhanced version of PEACH**

  ⇒ ***PEACH2***
  - **x4 lanes -> x8 lanes**
  - **hardwired on main data path and PCIe interface fabric**



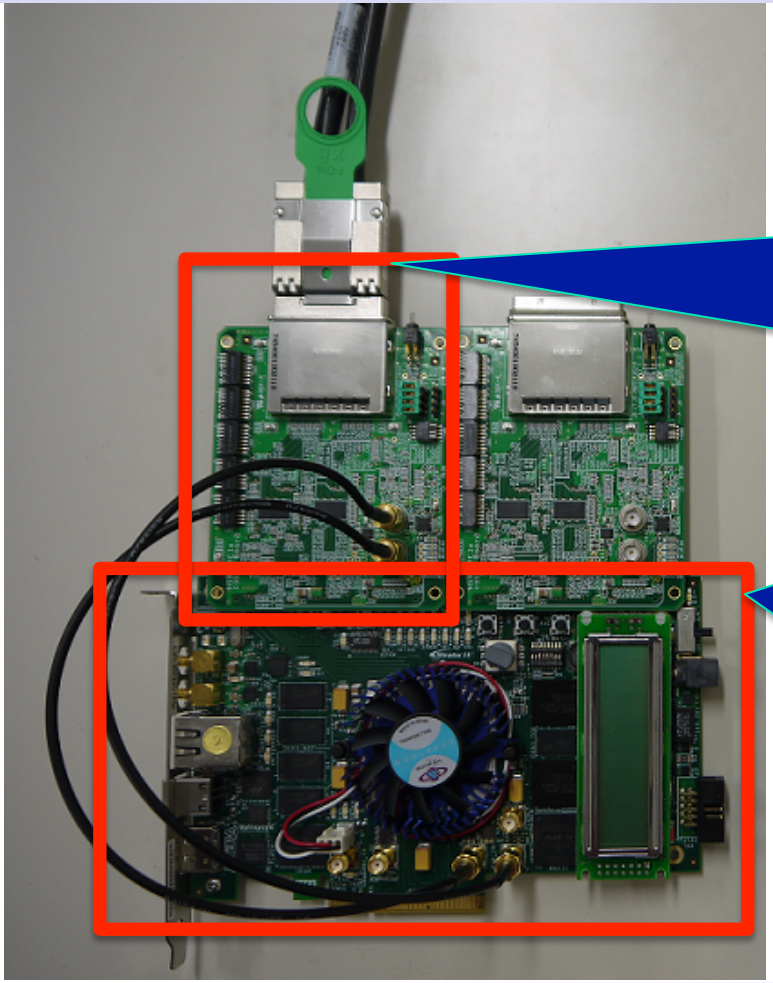


Implementation of PEACH2: ASIC⇒FPGA

- FPGA based implementation
  - today's advanced FPGA allows to use PCIe hub with multiple ports
  - currently gen2 x 8 lanes x 4 ports are available
    ⇒ soon gen3 will be available (?)
  - easy modification and enhancement
  - fits to standard (full-size) PCIe board
  - internal multi-core general purpose CPU with programmability is available
    ⇒ easily split hardwired/firmware partitioning on certain level on control layer
- Controlling PEACH2 for GPU communication protocol
  - collaboration with NVIDIA for information sharing and discussion
  - based on CUDA4.0 device to device direct memory copy protocol

PEACH2 FPGA test bed (~ Mar. 2012)

PCI-ExpressハードIPの性能や機能を確認

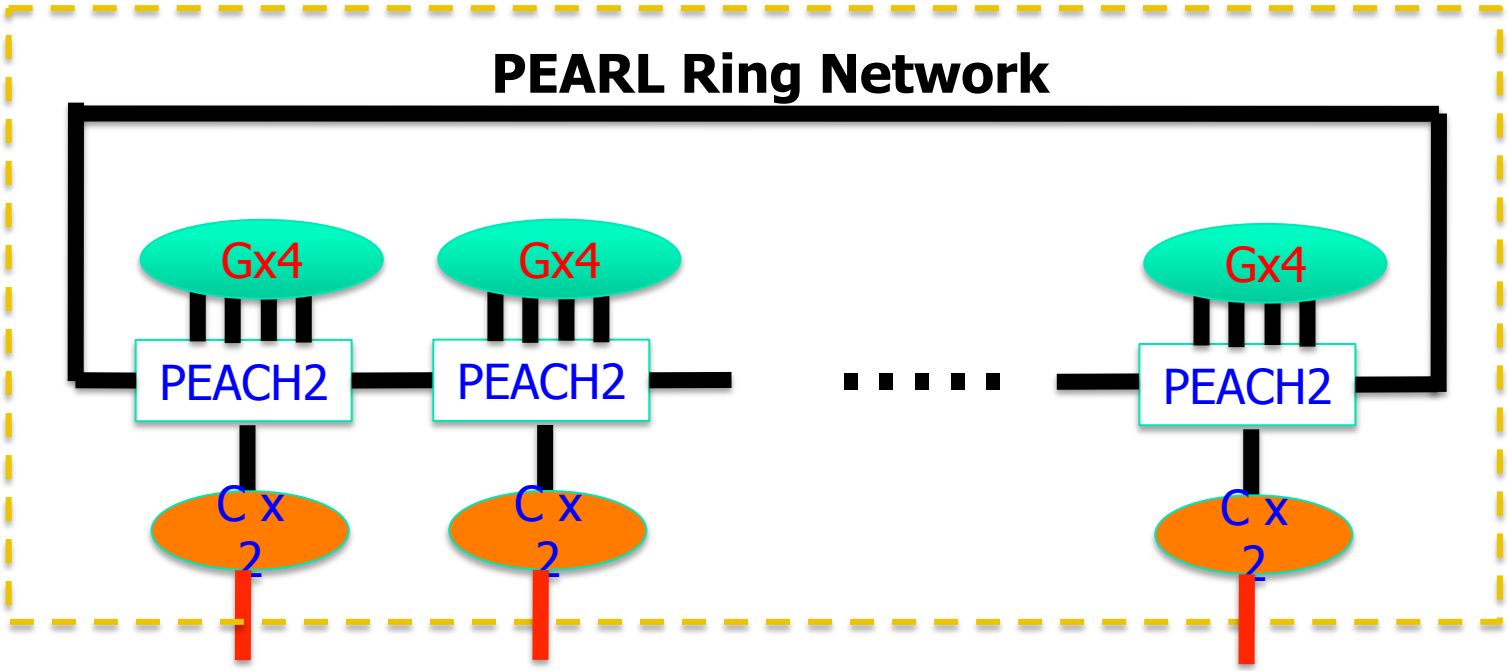


HSMC-PCIe converter board (newly developed)
- HSMC (High-Speed Mezzanine Card) standard
- PCIe x8 cable connector
- RootComplex / Endpoint switchable
- both x4 / x8 available

generic FPGA evaluation board (DEV-4SGX503N)
- PCIe x8 endpoint
- HSMC support PCIe x 8 Gen2
- HSMC support PCIe x 4 Gen2
- Stratix IV GX530KH40 (1517pin, 531K LE, 20Mbit, 4 PCIe IP, 24 8.5Gbps Transceiver)

HA-PACS/TCA

Node Cluster = NC

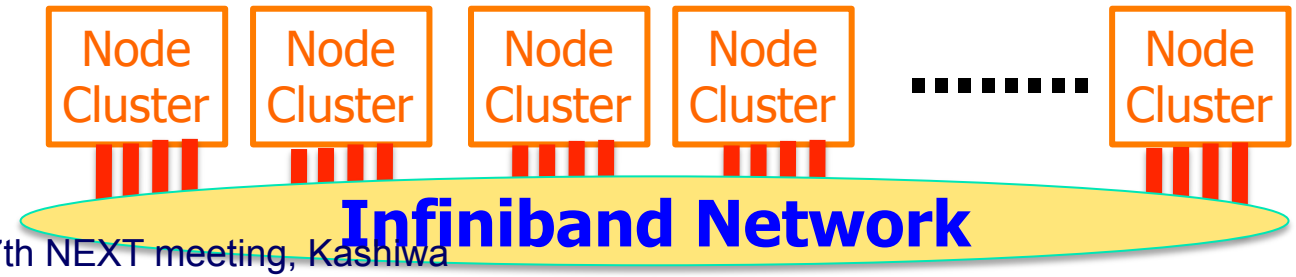


Node Cluster with 16 nodes
- GPUx64 (G)
- CPUx32 (C)
- GPU comm with PCIe
- IB link / node
- CPU: Xeon E5
- GPU: Kepler

Infiniband Link

- High speed GPU-GPU comm. by PEACH within NC (PCI-E gen2x8 = 5GB/s/link)
- Infiniband QDR (x2) for NC-NC comm. (4GB/s/link)

4 NC with 16 nodes, or 8 NC with 8 nodes = 360 TFLOPS extension to base cluster



Toward Exa-scale: what's the next step?

- Target=40GF/W
  - x20 more perf./power efficiency than HA-PACS
- Is GPGPU forever ?
  - maybe "NO", because GPU is based on GPU business solution
  - x10 performance is really required ?
  - x10 memory capacity comes ?
  - catch up to PCIe gen.4 ?
- Many-core solution will be technologically merged
  - Intel MIC (IA32 based many core, Knights Ferry, Knights Corner and Knights Landing)
  - Mixture core of general purpose (small # of complex cores) and accelerator (large # of simple cores)

Toward Exa-scale Computing

- Exa-scale target year: 2018-2019
- In all aspects, memory/network bandwidth is limited and severe (esp. for memory)
  - Maximize the locality of data access
  - Multi-layered parallelization (hybrid)
- Accelerated computing technology is necessary
  - I/O bandwidth of accelerators is critical
  - Latency hiding/solution is important, not just bandwidth
- Network and accelerators must be much closer as well as CPU and accelerators

"Free lunch is over!"

THANK YOU