# T2K & HA-PACS Projects

## *Supercomputers at CCS*

**Taisuke Boku**

**Deputy Director, HPC Division**

**Center for Computational Sciences**

**University of Tsukuba**

# Two Streams of Supercomputers at CCS

- **Service oriented general purpose machine with regular budget**
  **→ T2K-Tsukuba & follow-up**
  - Supercomputer rental budget to support 4-5 years period as national shared supercomputer resource (including HPCI)
  - High performance, commodity base and easy to program
  - Large scale general purpose parallel processing
  - Latest system: T2K-Tsukuba

- **Research & mission oriented project machine with special budget**
  **→ PAX/PACS series**
  - Supercomputer development for specific application area
  - Peak performance centric
  - “Highly skilled” high performance system for high-end computing
  - Latest system: HA-PACS

# T2K-Tsukuba

# History and Background

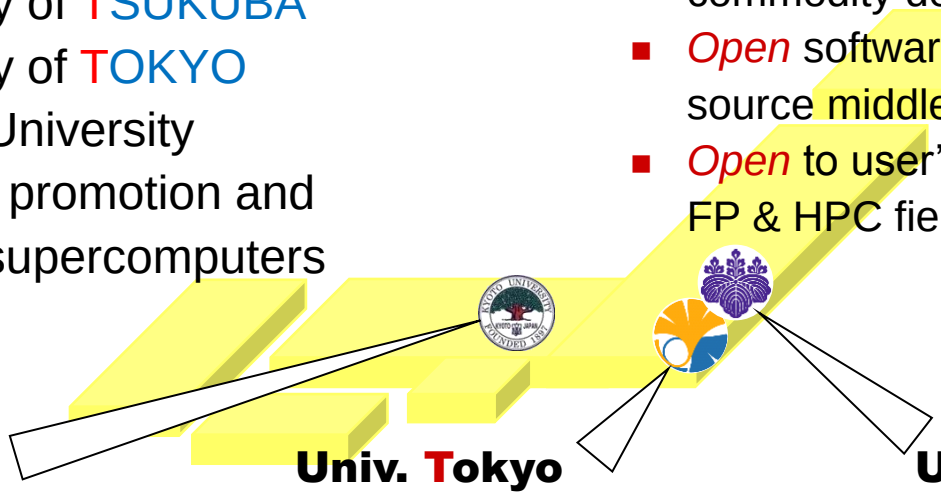
- Service oriented supercomputer operation had been originally a work in Media, Education and Network Center in U. Tsukuba which was operating Fujitsu Vector machines (VPP500, VPP5000)
- In 2007, U. Tsukuba decided to shift the procurement and operation of these supercomputers to CCS
- At the procurement of new machine, CCS decided to shift the architecture from Vector to Scalar (PC Cluster) in the supercomputer trends on those days
- At the same time, U. Tokyo and Kyoto U. were also going to start the procurement on their machines
- We agreed to make a “coordinated” procurement based on the same system architecture and performance target on each node to strongly proceed the collaborated researches and portability of application codes

  → “T2K” (Tsukuba, Tokyo and Kyoto) Alliance

# What is “T2K” ?

- Alliance between three Japanese national universities
  - University of TSUKUBA
  - University of TOKYO
  - KYOTO University
- Procurement, promotion and education of supercomputers

- *T2K Open Supercomputer Alliance*
- *Open* hardware architecture with commodity devices & technologies.
- *Open* software stack with open-source middleware & tools.
- *Open* to user’s needs not only in FP & HPC field but also INT world.



**Kyoto Univ.**
416 nodes (61.2TF) / 13TB
Linpack Result:
Rpeak= 61.2TF (416 nodes)
Rmax = 50.5TF

**Univ. Tokyo**
952 nodes (140.1TF) / 31TB
Linpack Result:
Rpeak= 113.1TF (512+256 nodes)
Rmax = 83.0TF

**Univ. Tsukuba**
648 nodes (95.4TF) / 20TB
Linpack Result:
Rpeak= 92.0TF (625 nodes)
Rmax = 76.5TF

# Common spec. of T2K Supercomputers

Open Supercomputer T2K
Open Supercomputer Alliance

- Quad-core / Quad-socket multi-core fat node
  - AMD quad-core Opteron (Barcelona) 2.3GHz
  - 32GB memory (partially 128GB in U. Tokyo)
  - 147GFLOPS / node
- Quad-rail high performance SAN
  - Infiniband 4xDDR (U. Tsukuba & Kyoto U.)
  - Myrinet10G (U. Tokyo)
- Software stacks
  - RedHat Linux
  - C, C++, Fortran90
  - multi-rail ready MPI

Sharing the basic hardware and software
⇒ sharing applications with performance portability

# T2K-Tsukuba (by Appro Int. & Cray Japan Inc.)

Open Supercomputer T2K
Open Supercomputer Alliance

#20 at TOP500 on June 2008 (95 TFLOPS peak, 76.46 TFLOPS Linpack)

# Computation Nodes and File Servers

Open Supercomputer T2K
Open Supercomputer Alliance

Computation node（70racks）
(Appro XtremeServer-X3)

648 node (quad-core x 4socket / node)
Opteron "Barcelona" 8356 CPU
2.3GHz x 4FLOP/c x 4core x 4socket
= 147.2 GFLOPS / node
= 95.3 TFLOPS / system
20.8 TB memory / system

800 TB (physical 1PB) RAID-6
Luster cluster file system
Infiniband x 2
Dual MDS and OSS config.
⇒ high reliability

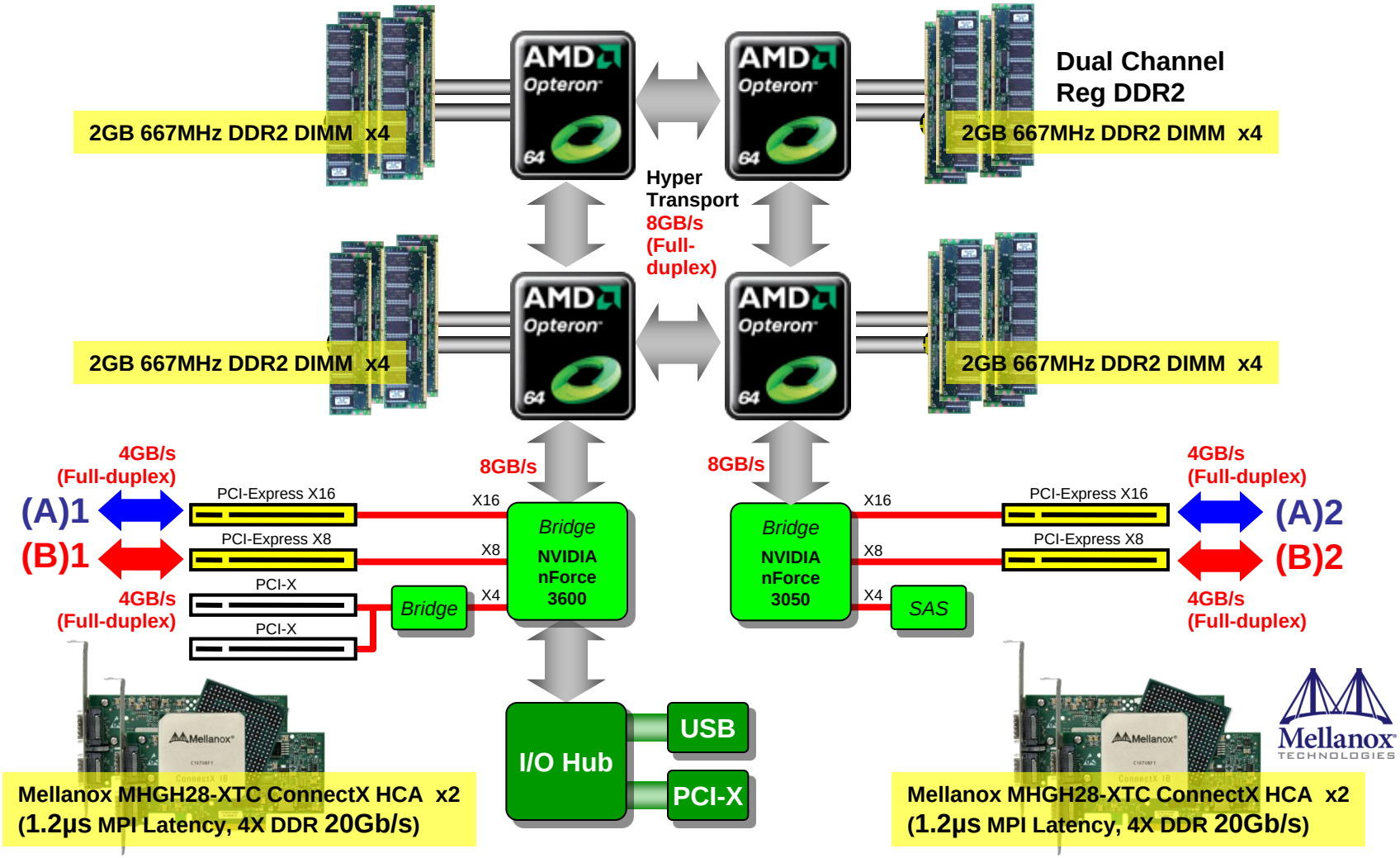

File server (disk array only)
(DDN S2A9550)

# Performance Overview of T2K-Tsukuba System

- System Specification
  - # of computation nodes: 648 (Appro XtremeServer-X3)
  - Peak performance: 95.39 TFLOPS
  - Total communication bandwidth: 5.18 Tbyte/sec
  - Total file system capacity: 800 Tbyte (RAID6, available user space)
  - Network: Full bisection-bandwidth Fat Tree with Quad-Rail
  - Shared file system bandwidth: 12 Gbyte/sec
- "Open Architecture" based on commodity processor network
- Fat node with multi-core/multi-socket architecture (4x4 = 16 cores)
- Shared file system provides a complete "flat-view" for all computation node
- Easy node allocation and scheduling for large scale jobs

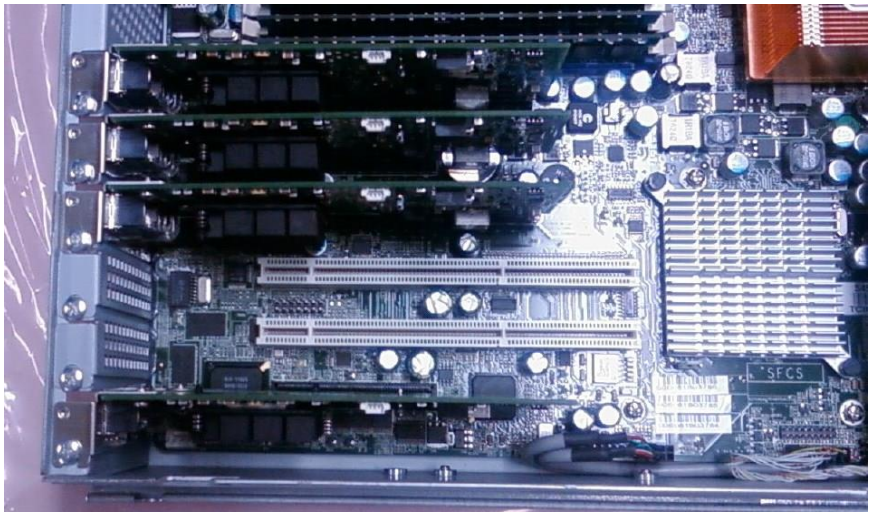
# Block diagram of T2K-Tsukuba computation node

## 2.3GHz quad-core Opteron (8356, Barcelona) x 4



# Inside the chassis

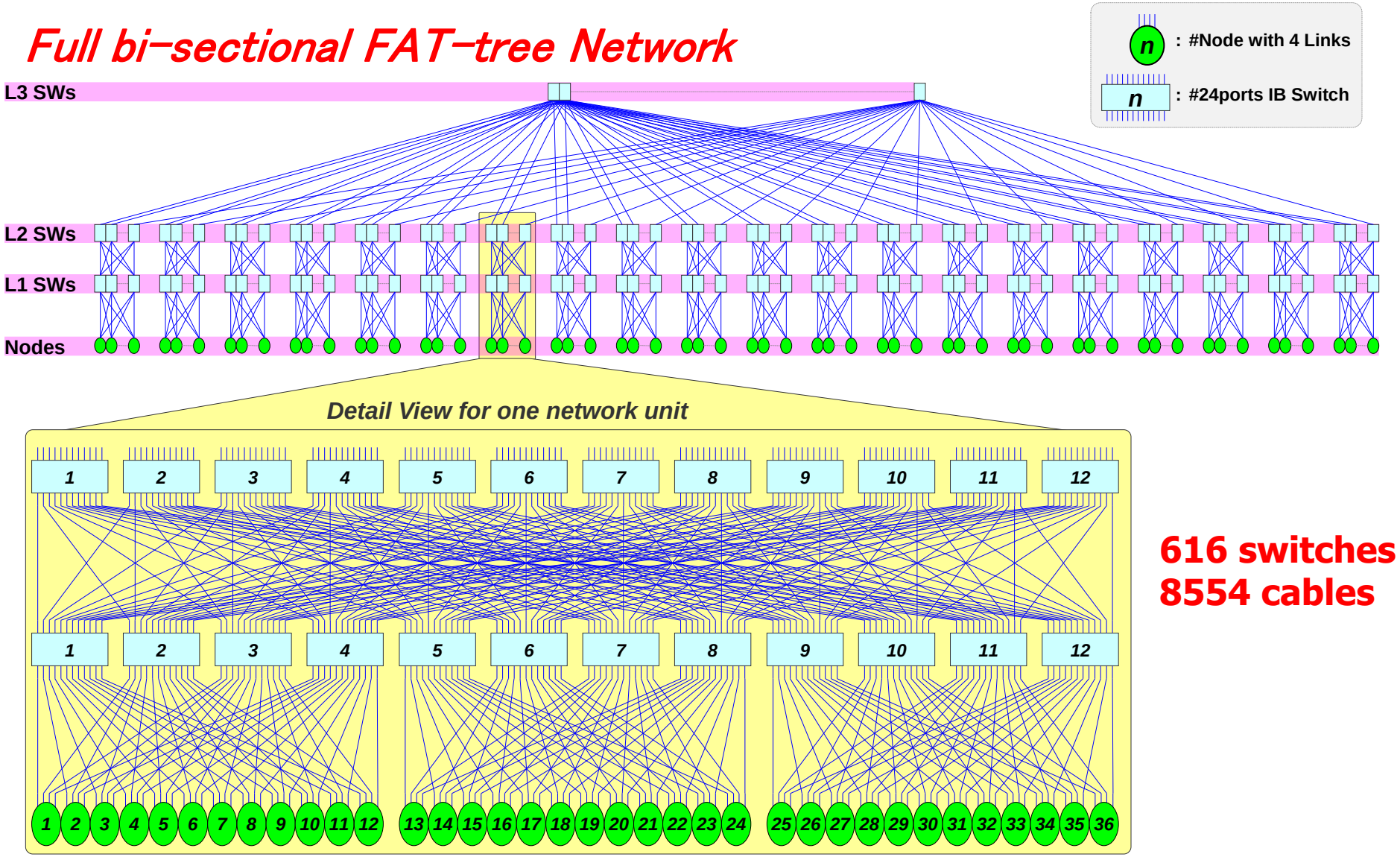
Inside the enclosure



Mellanox DDR IB ConnectX x 4
(PCI-Express gen.1 x 8lane, each)

# Infiniband Fat Tree Network

## *Full bi-sectional FAT-tree Network*



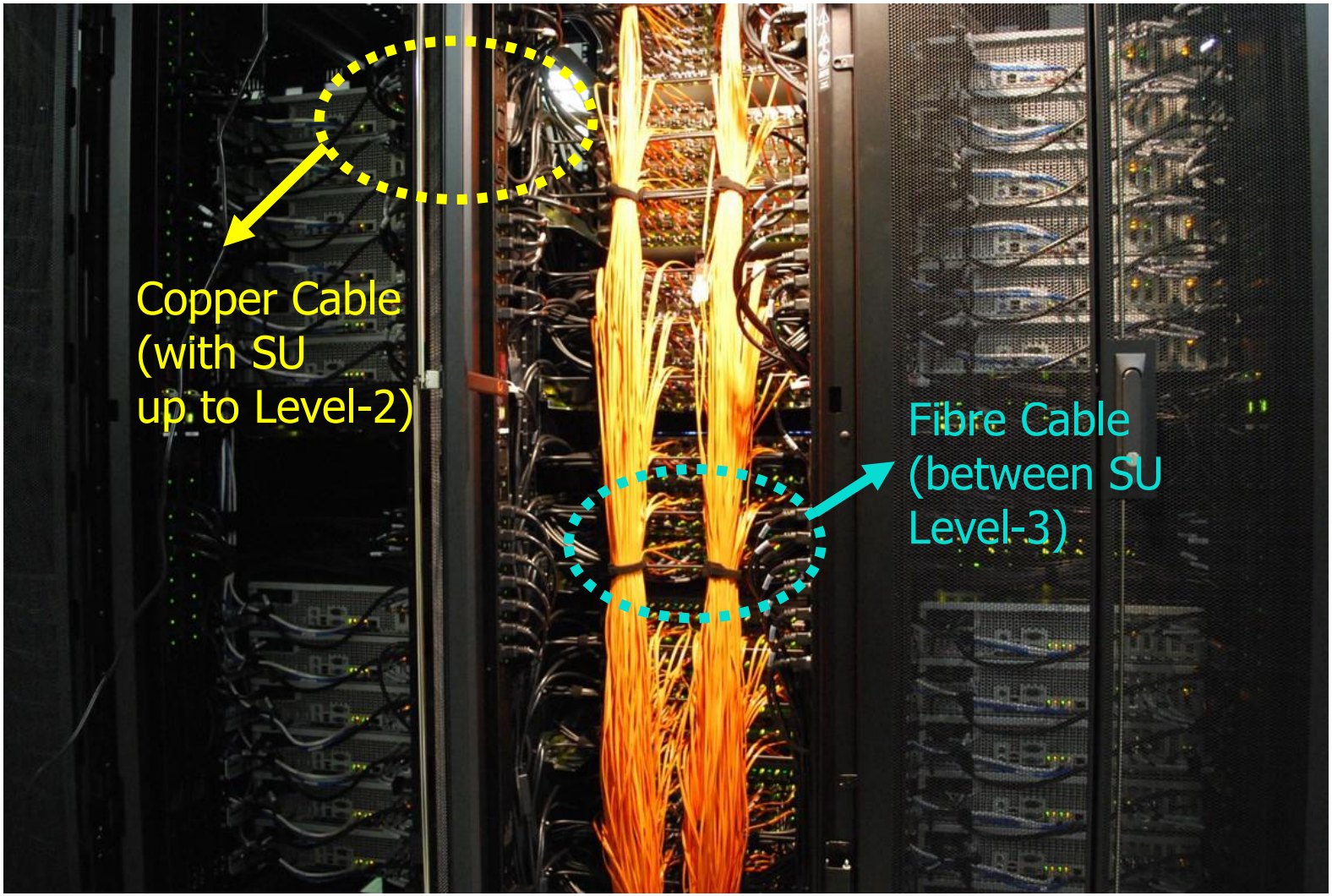

**616 switches**
**8554 cables**

*x 20 network units*

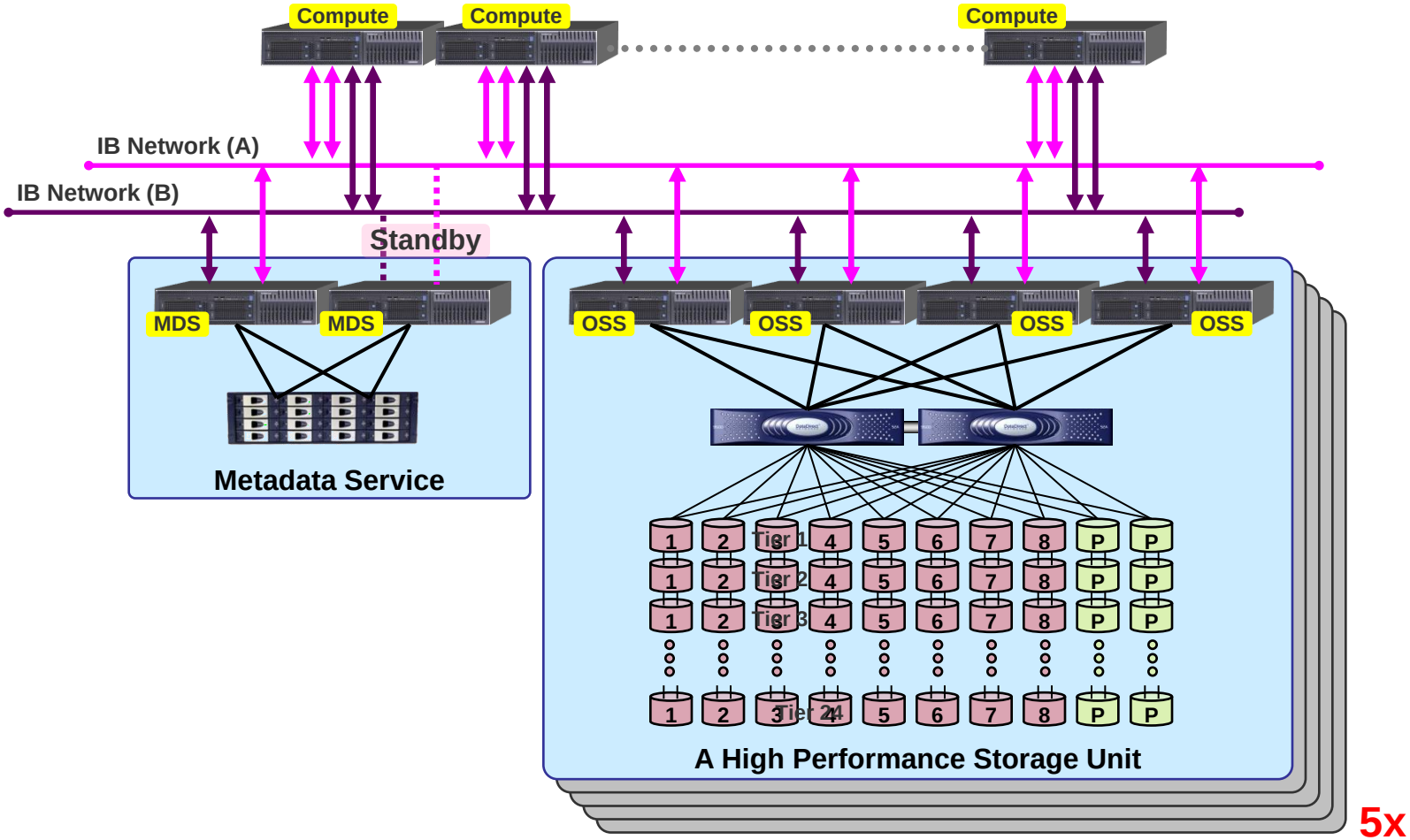
# Floor plan and Infiniband network link

CCS U. Tsukuba, 2nd Supercomputer Building

# Infiniband switch rack and cables



# Lustre shared file system



# Software Stack

- **Software stack**
  - OS: Red Hat Enterprise Linux v.5 WS (Linux kernel 2.6)
  - Programming languages: F90, C, C++, Java
  - Compilers: PGI, Intel, GNU
  - MPI library: MVAPICH2 (OSU)
  - Numerical library: IMSL, ACML, SCALAPACK
  - Performance monitor: PGPROFR, PAPI

- **Programming**
  - Flat MPI
  - Hybrid with MPI + OpenMP
  - Multithread compiler + MPI


# T2K-Tsukuba operation program

- Special Promotion Program for High-end Computational Sciences (approx. 50%)
  - Proposal base CPU time assignment by external review
  - Machine time is free, but limited according to the review
  - The program is performed for promotion of large scale computational sciences toward Peta-scale computation
- HPCI: High Performance Computing Infrastructure, national program (approx. 25%)
  - Proposal base CPU time assignment by HPCI program committee
  - Machine times is free, but limited according to the committee's review
  - Under HPCI concept, Gfarm shared file system is applied
- General use for Japanese universities and institutes (approx. 25%)
  - Fixed size of nodes for each project is assigned, and CPU time fare charged according to the node size
  - Any application is welcome from scientific and engineering fields

# Applications on T2K-Tsukuba (by CCS)

- **Large scale parallel applications**
  - MPP: QCD (particle phys.), RSDFT (material science), 3—DRISM (nano),...
  - Vector -> Scalar: NICAM (global atomospheric)
  - Others: Phantom-GRAPE (Astrophys.) , WRF (climate), MD (nano-bio),...

- **World-class large simulation**
  - QCD: toward physical point (48x48x48x96 lattice)
  - RS-DFT: 10,000 atom
  - Astro: first object simulation
  - Climate: 7km-mesh for entire globe simulation at “cloud-level”

**ported to K Computer**

# Latest result: computation of "π"

- Massively large digits computation based on FFT
- 2,576,980,370,000 digits: WORLD RECORD (submitted to Guinness)
  (Previous record: 1,2411,000,000 digits)
- 73 hours with 10,240 cores on T2K-Tsukuba (～full system)
  - Two time computation in different alrgorithm
- Purpose
  - Stability check of T2K-Tsukuba (confirmation of MTBF)
  - Records so far were made by large scale vector
    -> 12% of running time of previous record

# Operation Status of T2K-Tsukuba

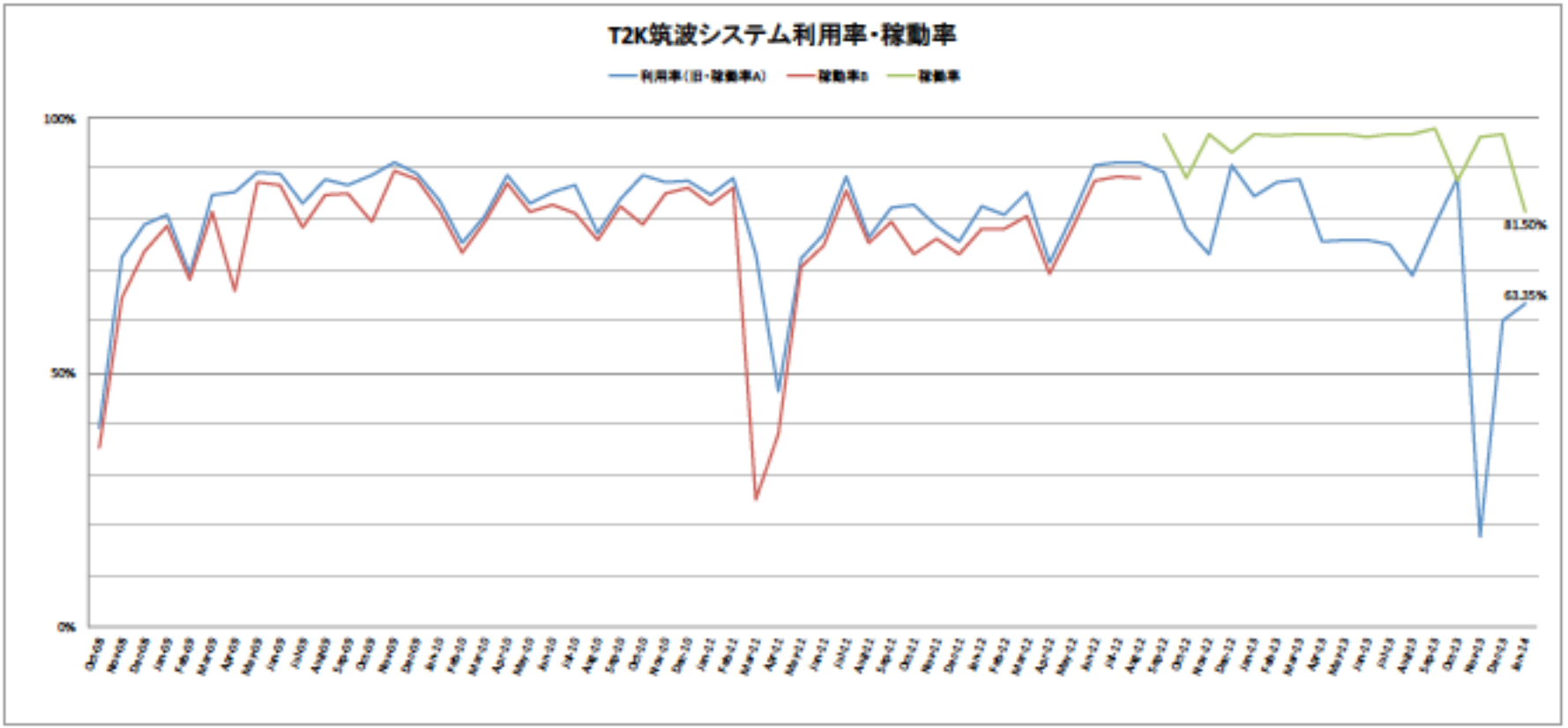
- Originally planned to operate from Jun. 2008 to May 2013, but we decided to extend it until Feb. 2014 according to our new plan for JCAHPC (joint system procurement and operation with U. Tokyo)
  → T2K-Tsukuba will be shut down in the end of this month

- Average utilization ratio for more than 5 years is more than 70% thanks to wide variety of program for nation wide users, freedom of node scheduling and high availability of the system itself

- On 11$^{th}$ of March 2011, the system was temporarily shut down due to the heavy earthquake followed by the power shortage in Japan, but recovered on May and continuously operated
  → almost no damage on hardware

- Incident: the system was intruded by some user's private key leak and insufficient security patch on Linux kernel, on Oct. 2013 and Jan. 2014

# T2K-Tsukuba Operation and Utilization Ratio



# HA-PACS

# Hisotry of PAX (PACS) MPP series

- Launched in 1977 (Prof. Hoshino and Prof. Kawai)
- First machine was completed in 1979
- 6th generation machine CP-PACS was ranked #1 in TOP500 in Nov. 1996

1978
1st PACS-9

1980
2nd PAXS-32

1989
5th QCDPAX

1996
6th CP-PACS as world fastest machine

2006
7th PACS-CS bandwidth aware

2012
8th HA-PACS GPU accelerated

| Year | Name | Performance |
|---|---|---|
| 1978年 | PACS-9 | 7 KFLOPS |
| 1980年 | PACS-32 | 500 KFLOPS |
| 1983年 | PAX-128 | 4 MFLOPS |
| 1984年 | PAX-32J | 3 MFLOPS |
| 1989年 | QCDPAX | 14 GFLOPS |
| 1996年 | CP-PACS | 614 GFLOPS |
| 2006年 | PACS-CS | 14.3 TFLOPS |
| 2012年 | HA-PACS | 802 TFLOPS |

- High end supercomputer based on MPP architecture towards “practical machine” under collaboration with computational scientists and computer scientists
- Development in Application-driven
- Continuation of R & D by an organization

# PAX (PACS) Series

- MPP system R&D continued at U. Tsukuba for more than 30 years
- Coupling of need from applications and seeds from the latest HPC technology, the machines have been developed and operated with the effort by application users on programming
  → a sort of application oriented machine
  (not for a single application)
- HA-PACS is the first system in the series to introduce accelerating devices (GPUs)
- CCS has been focusing on the accelerating devices for ultra high performance to provide to “high-end” users who require extreme computing facilities

# CCS and Accelerated Computing

- FIRST (a hybrid cluster with GRAPE6 engine for astrophysics) was the first large scale system with accelerators towards very high peak performance for applications
- HA-PACS Project proceeds the porting of major applications in CCS to GPU accelerated systems
- We will proceed the accelerated computing research both on system and application development
- Our Feasibility Study towards Exascale Computing System basic design under MEXT is also based on an idea of extremely accelerated chip
- Wide variety of system architecture is our target to port and execute our applications which require extremely high performance and large scale computation

# Project plan of HA-PACS

- **HA-PACS (Highly Accelerated Parallel Advanced system for Computational Sciences)**
- Accelerating critical problems on various scientific fields in Center for Computational Sciences, University of Tsukuba
  - The target application fields will be partially limited
  - Current target: QCD, Astro, QM/MM (quantum mechanics / molecular mechanics, for life science)
- Two parts
  - HA-PACS *base cluster*:
    - for development of GPU-accelerated code for target fields, and performing product-run of them
  - HA-PACS/*TCA*: (TCA = Tightly Coupled Accelerators)
    - for elementary research on new technology for accelerated computing
    - Our original communication system based on PCI-Express named "PEARL", and a prototype communication chip named "PEACH2"

# GPU Computing: current trend of HPC

- **Major GPU clusters in TOP500 on Nov. 2011**
  - #5 天河 Tienha-1A (Rmax=2.57 PFLOPS)
  - #10 星雲 Nebulae (Rpeak=1.27 PFLOPS)
  - #14 TSUBAME2.0 (Rpeak=1.19 PFLOPS)
  - (1st Sequoia Rpeak=16.32 PFLOPS)
- **Features**
  - high peak performance / cost ratio
  - high peak performance / power ratio
  - large scale applications with GPU acceleration don't run yet in production on GPU cluster

  ⇒ **Our First target is to develop large scale applications accelerated by GPU in real computational sciences**

# Issues of GPU Cluster

- **Problems of GPGPU for HPC**
  - Data I/O performance limitation
    - Ex) GPGPU: PCIe gen2 x16
    - Peak Performance： 8GB/s (I/O) ⇔ 665 GFLOPS (NVIDIA M2090)
  - Memory size limitation
    - Ex) M2090: 6GByte vs CPU: 4 – 128 GByte
  - Communication between accelerators: no direct path (external)
    ⇒ communication latency via CPU becomes large
    - Ex) GPGPU:
      GPU mem ⇒ CPU mem ⇒ (MPI) ⇒ CPU mem ⇒ GPU mem
- **Researches for direct communication between GPUs are required**

**Our another target is developing a direct communication system between external GPUs for a feasibility study for future accelerated computing**

# Project Formation

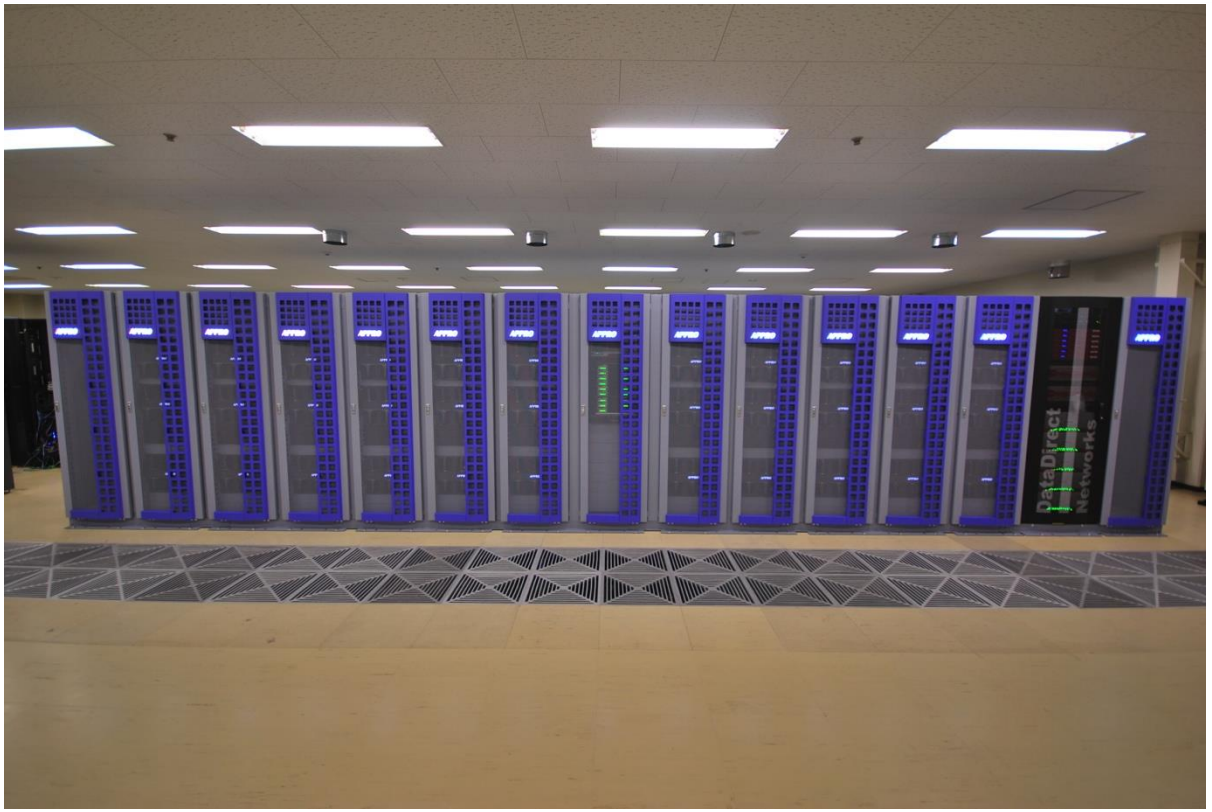
- **HA-PACS (Highly Accelerated Parallel Advanced system for Computational Sciences)**
  - **Apr. 2011 – Mar. 2014, 3-year project (the system will be maintain until Mar. 2016), lead by Prof. M. Sato**
    "Advanced research and education on computational sciences driven by exascale computing technology", $4.5M supported by MEXT
  - **Project Office for Exascale Computational Sciences (Leader: Prof. M. Umemura)**
    - Develop large scale GPU applications : 14 members

    Elementary Particle Physics, Astrophysics, Bioscience, Nuclear Physics, Quantum Matter Physics, Global Environmental Science, Computational Informatics, High Performance Computing Systems
  - **Project Office for Exascale Computing System Development (Leader: Prof. T. Boku)**
    - Develop two types of GPU cluster systems: 15 members

# HA-PACS base cluster (Feb. 2012)

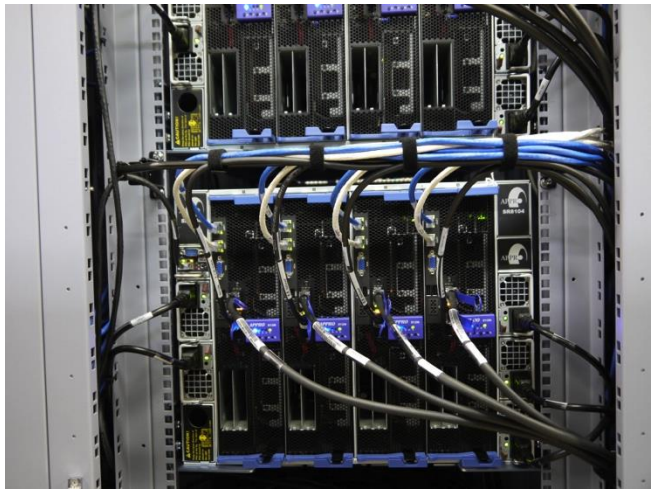
# HA-PACS base cluster


Front view




Side view

# HA-PACS base cluster

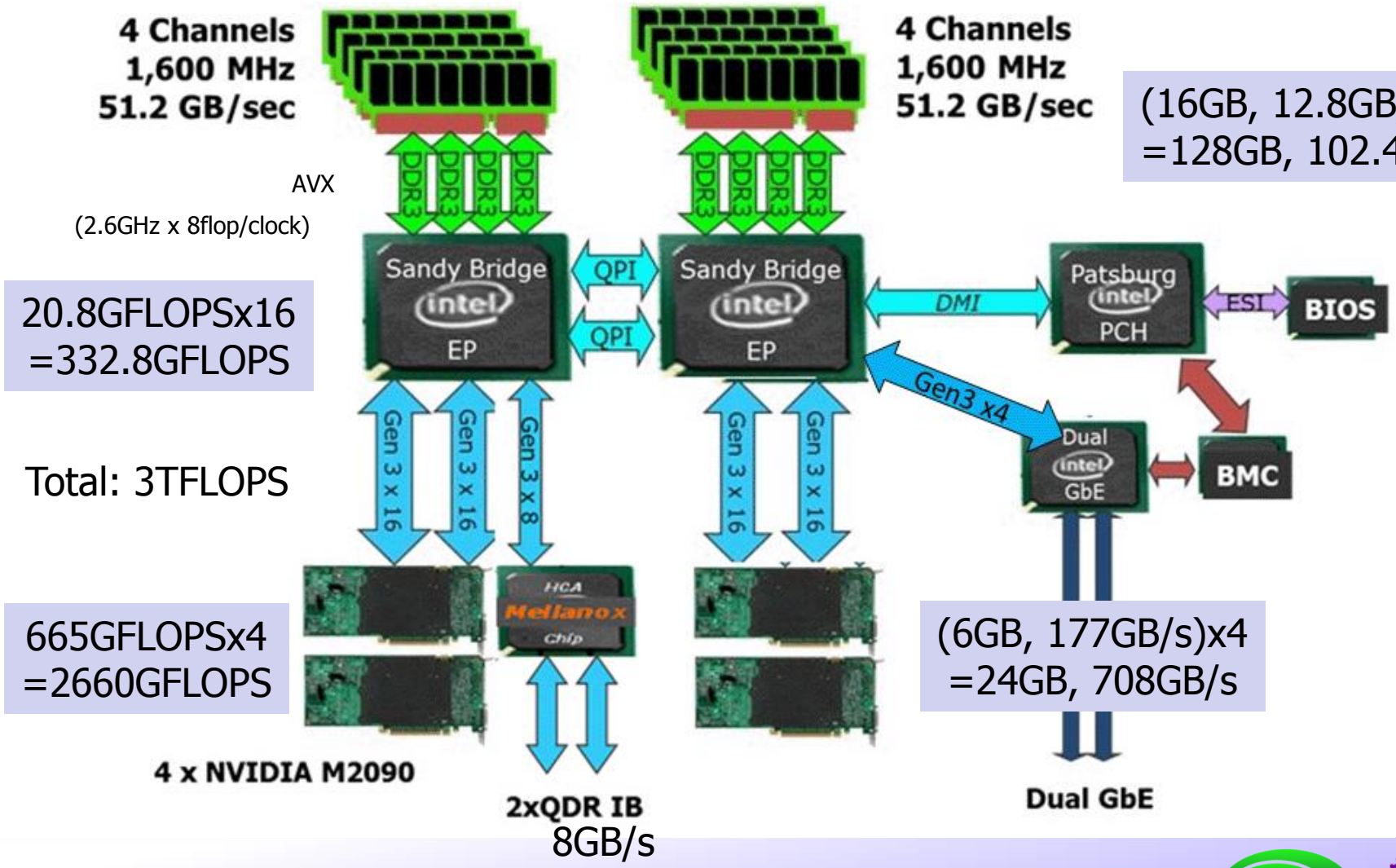


Rear view of one blade chassis with 4 blades

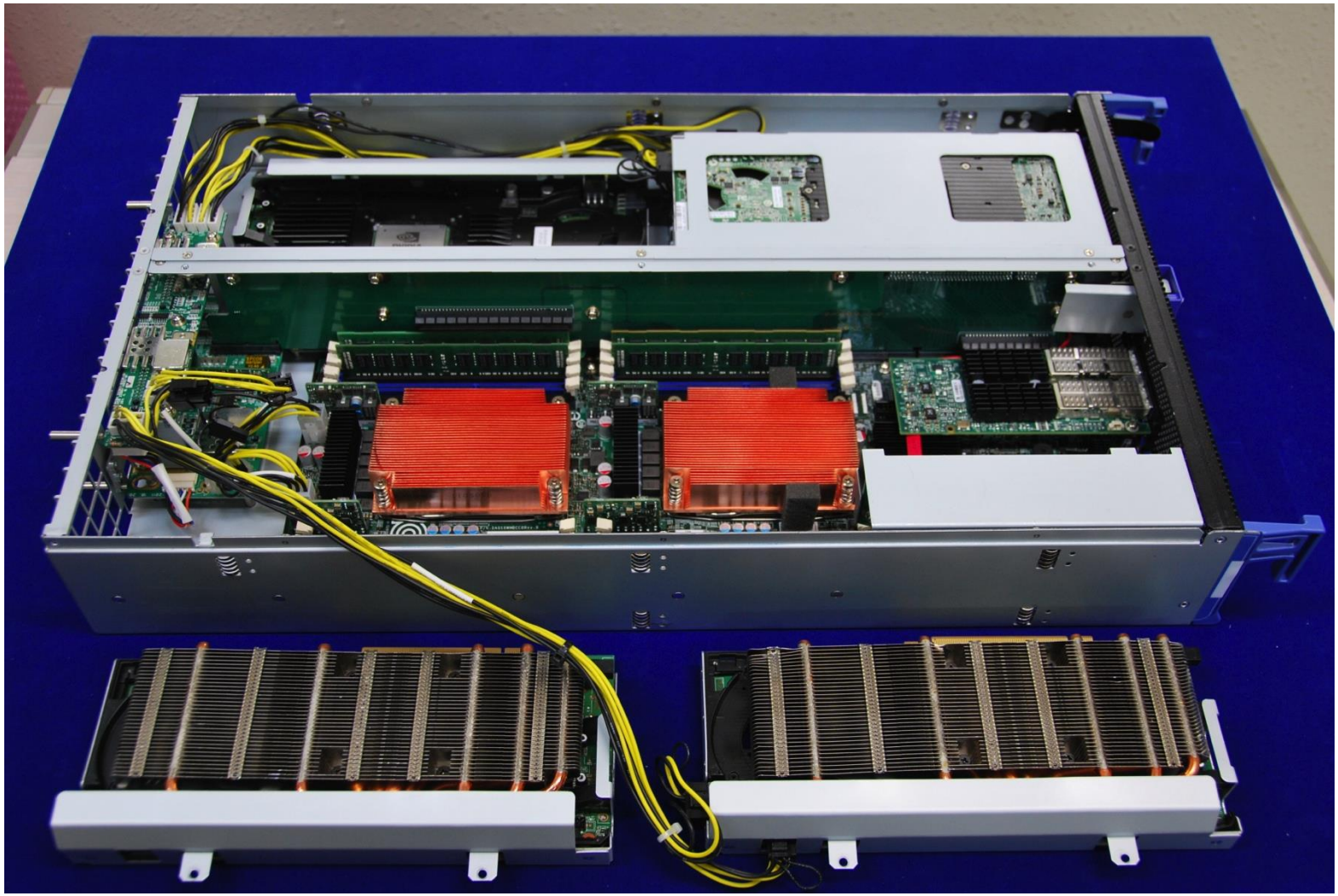
Front view of 3 blade chassis


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

# HA-PACS: base cluster (computation node)



# Computation node of base cluster



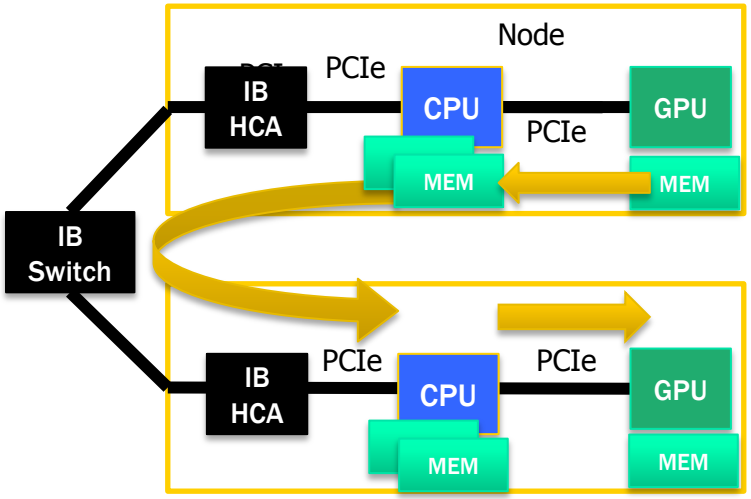
# HA-PACS: TCA

- **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
- **PEACH: PCI Express Adaptive Communication Hub**
  - We already developed such PCIe device on JST-CREST project "low power and dependable network for embedded system"
  - It enables direct connection between nodes by PCIe Gen2 x4 link

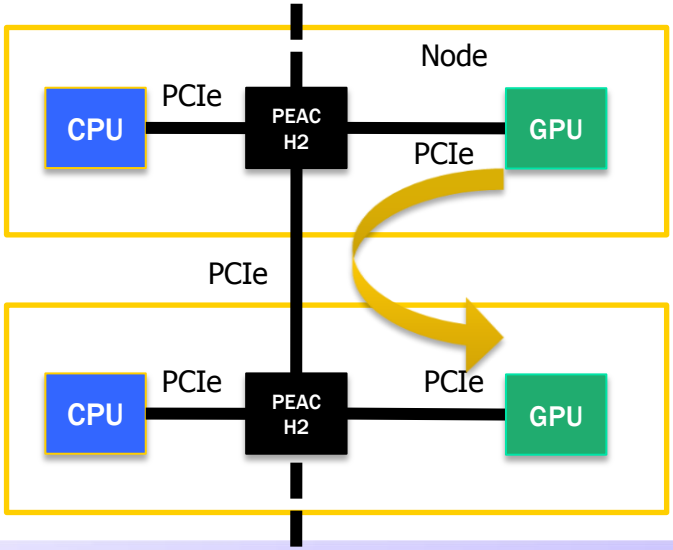
⇒ **Improving PEACH for HPC to realize TCA**

# HA-PACS/TCA (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



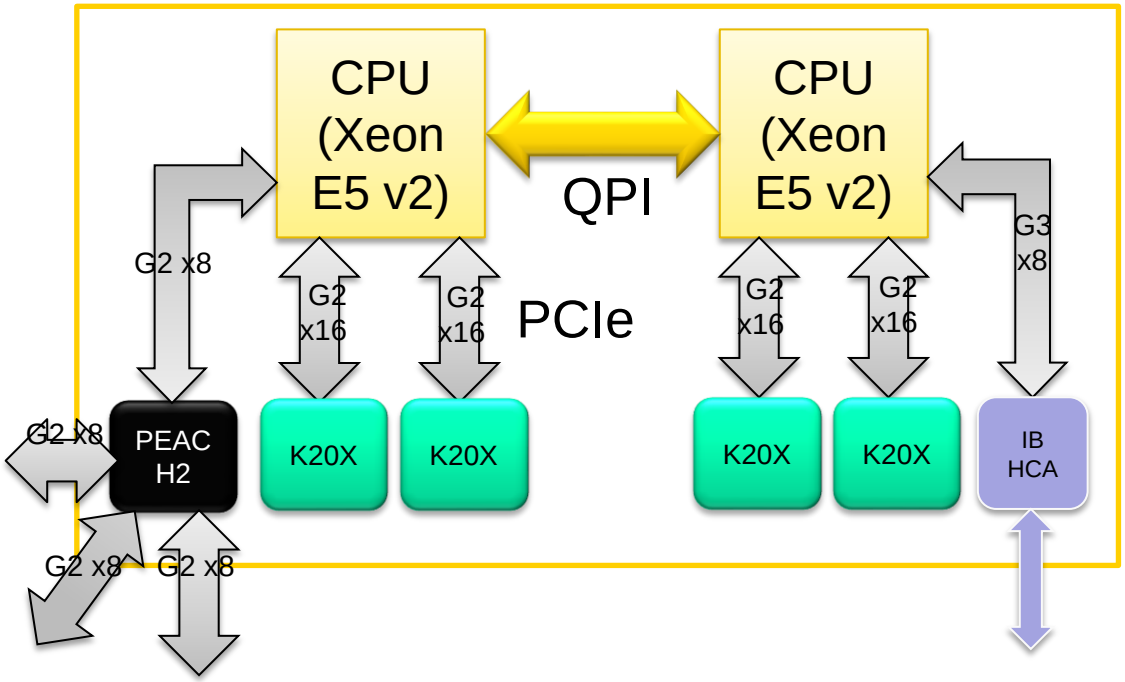
# Implementation of PEACH2: FPGA solution

- **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 and CUDA5.0 PCIe RDMA feature

# TCA node structure

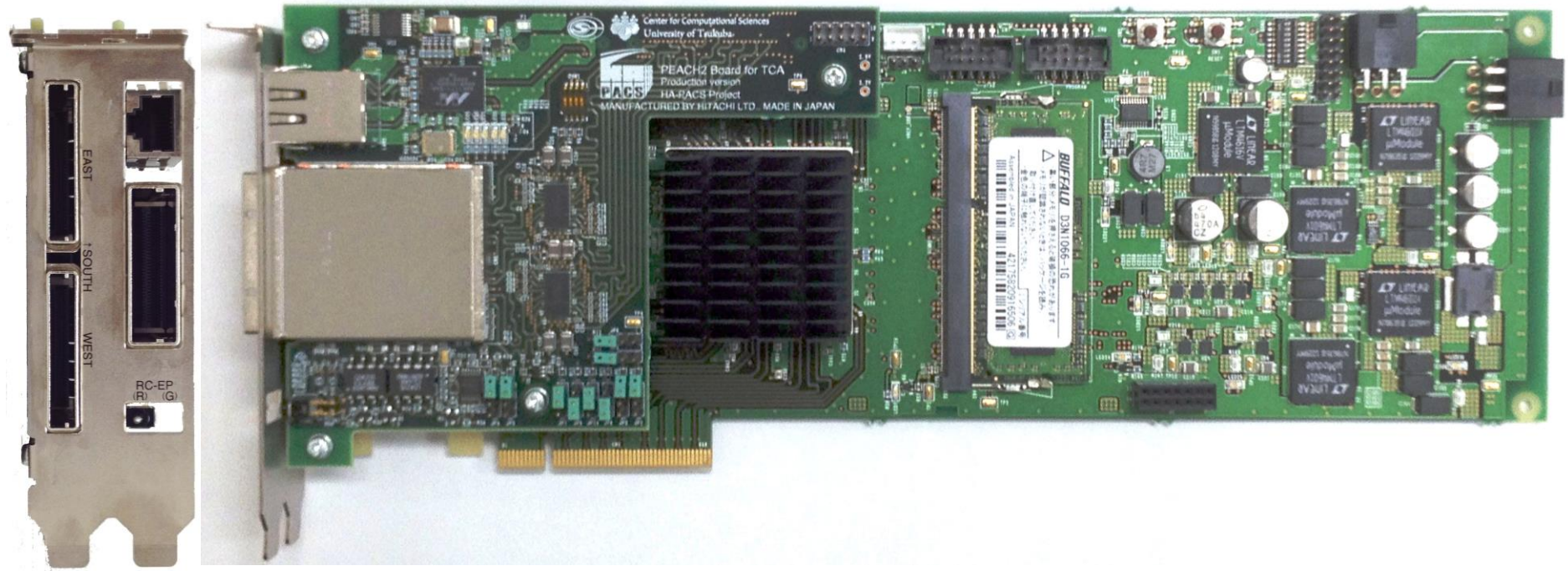
- CPU can uniformly access to GPUs.
- PEACH2 can access every GPUs
  - Kepler architecture + CUDA 5.0 “GPUDirect Support for RDMA”
  - Performance over QPI is quite bad.
    => support only for two GPUs on the same socket
- Connect among 3 nodes

- This configuration is similar to HA-PACS base cluster except PEACH2.
  - All the PCIe lanes (80 lanes) embedded in CPUs are used.



# PEACH2 board

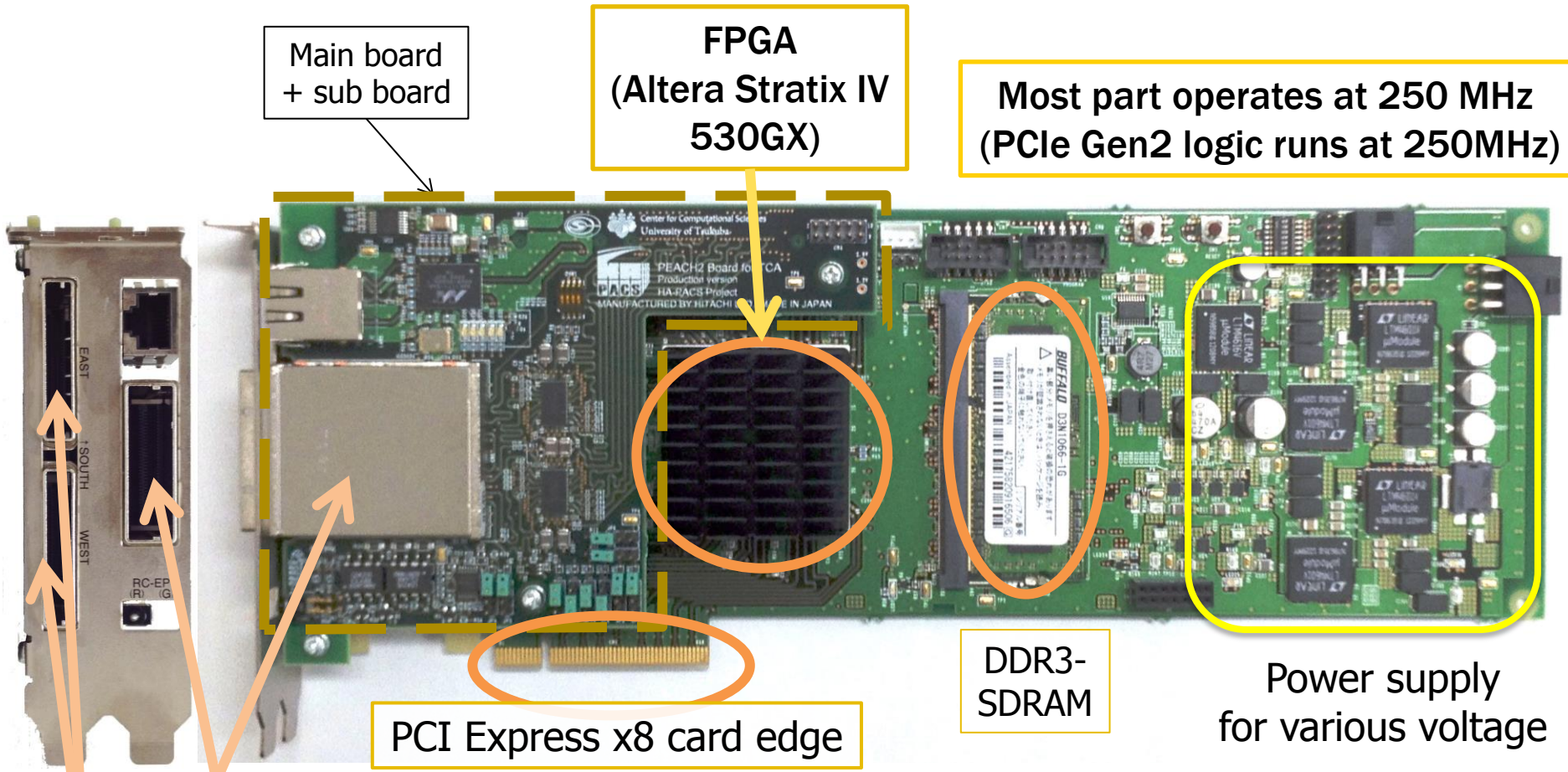
- **PCI Express Gen2 x8 peripheral board**
  - **Compatible with PCIe Spec.**



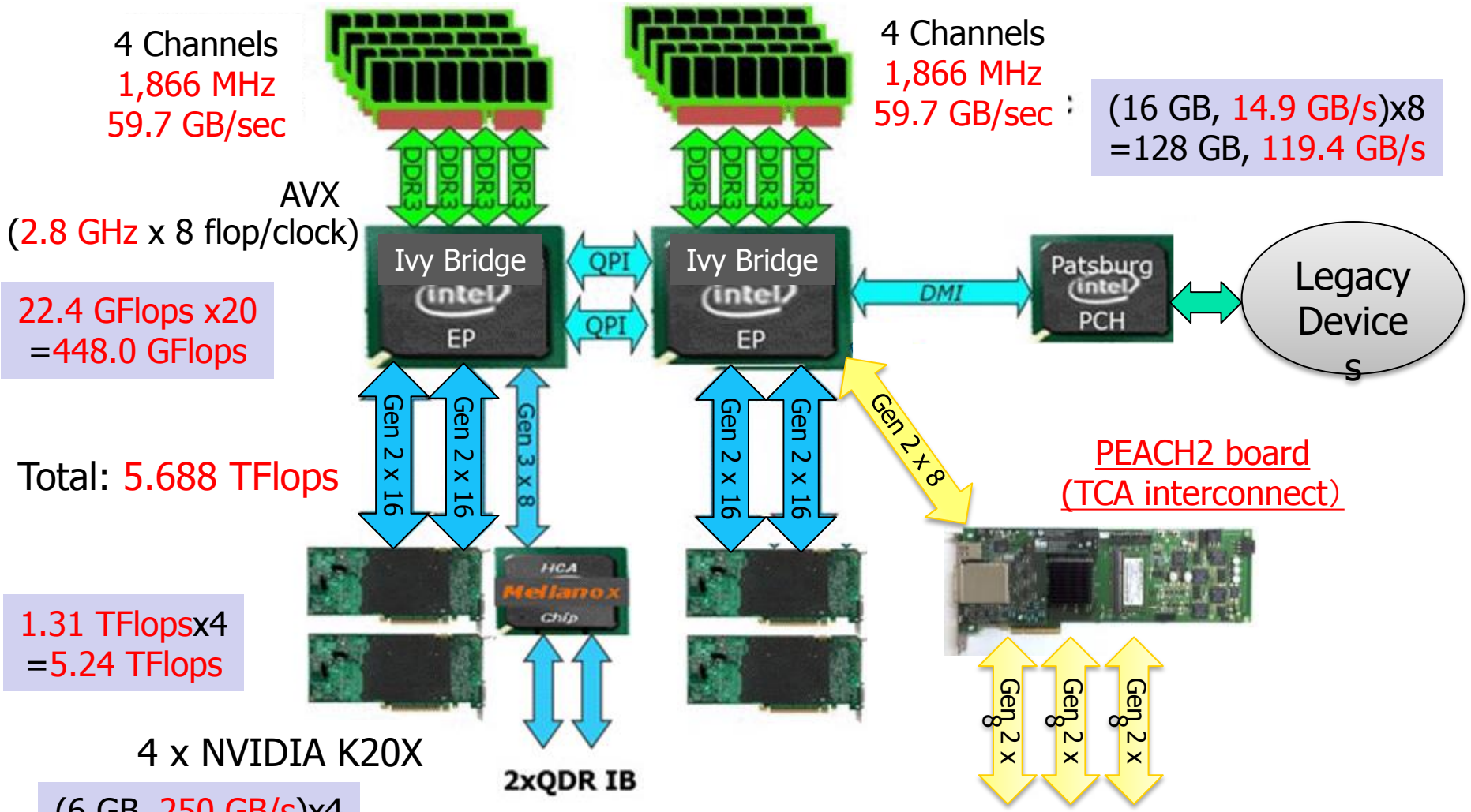

Side View

Top View

# PEACH2 board



# HA-PACS/TCA (computation node)



# HA-PACS Base Cluster + TCA
## (TCA part starts operation on Nov. 1st 2013)


- HA-PACS Base Cluster = 2.99 TFlops x 268 node = 802 TFlops
  - HA-PACS/TCA = 5.69 TFlops x 64 node = 364 TFlops
    - TOTAL: 1.166 PFlops

# HA-PACS/TCA computation node inside

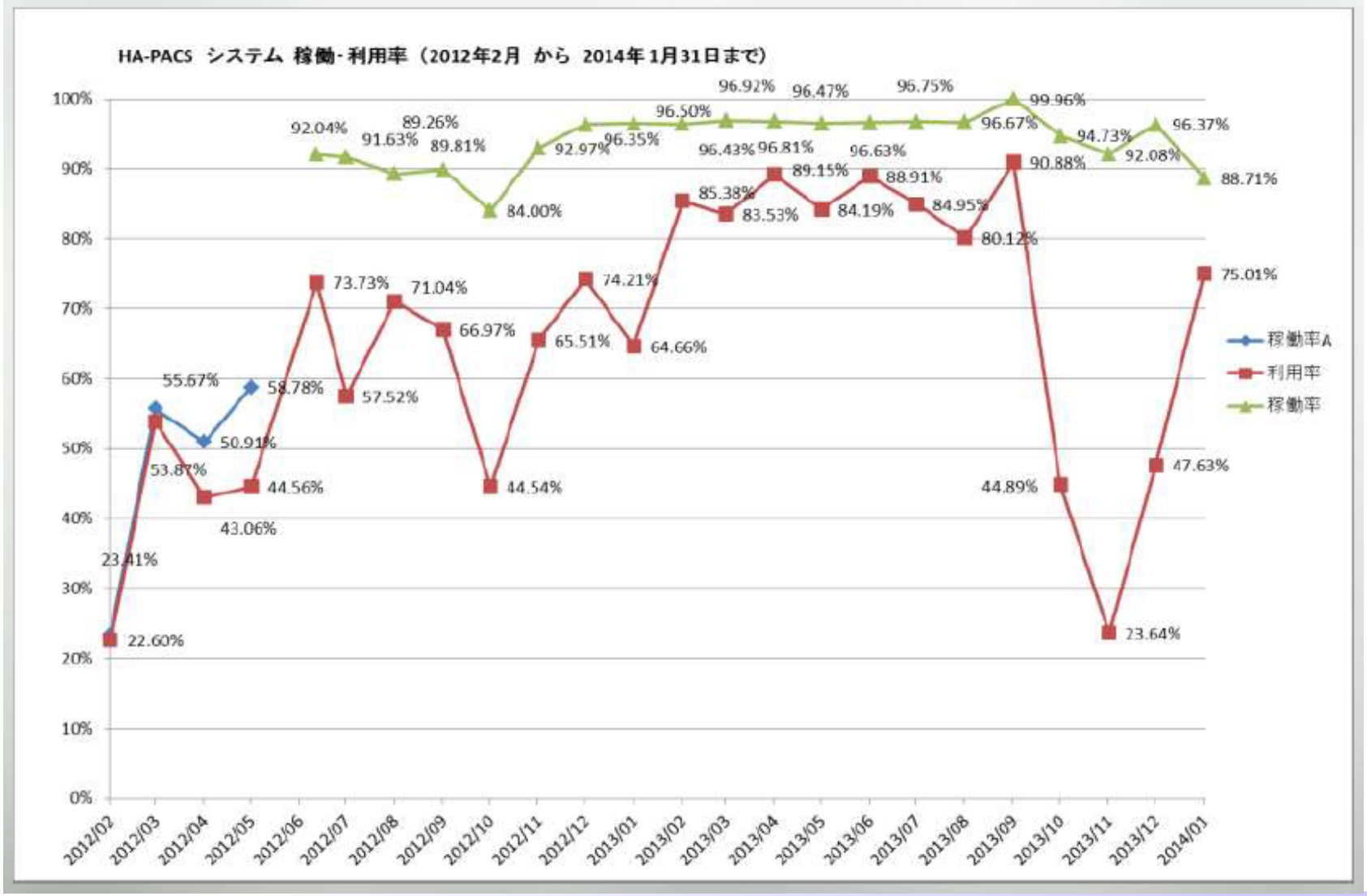
# TOP500 and Green500

- TOP500 (HPL) on Base Cluster
  - 421.6 TFLOPS (ranked #41 in TOP500, June 2012)
  - #7 as GPU cluster
  - Computing efficiency: 54.2% of theoretical peak
- Green500 on Base Cluster
  - 1151.91 MFLOPS/W (ranked #24 in Green500, June 2012)
  - #3 as GPU cluster
  - #1 as “large scale” (within TOP50) GPU cluster
- Green500 on TCA Part (without TCA feature)
  - 3518 MFLOPS/W (ranked #3 in Green500, November 2013)
  - 76% of HPL efficiency : quite high as GPU cluster
  - ranked #134 (277 TFLOPS) in TOP500 Nov. 2013

# Operation and Utilization Program of HA-PACS

- HA-PACS was deployed on Feb. 2012, and used privately within CCS for shake-down and concentrated development of several major applications (QCD, Astrophysics, Material Science, Life Science)
- From Oct. 2012, 100% of resource has been dedicated for Multidisciplinary Cooperative Research Program of CCS
  → Proposal base & external review, providing GPU computation resources to wide variety of nation wide researchers
  (free of charge)
- System condition has been kept stable almost, except relatively frequent failure of GPU (4-6 GPU devices/month out of 1072 devices)
- Average utilization ratio after Oct. 2012 is approximately 65%

# HA-PACS Operation and Utilization Raio



# Summary

- CCS has been operating several big machines in two streams:
  - PAX/PACS series for peak performance and high-end users on specific applications
  - T2K-Tsukuba (and follow-up system) for general purpose use including HPCI resource sharing program
- PAX/PACS series with 8 generations have been contributing to high-end computational sciences under the collaboration with computational domain scientists and computer scientists
- T2K-Tsukuba and follow-up system has been contributing to wide area of computational science/engineering through Multidisciplinary Cooperative Research Program and HPCI Resource Sharing Program
- For the challenge on PAX/PACS, we are strongly promoting the accelerated computing based on GPU, FPGA and Many-core technology