HA-PACS Project

Yuetsu Kodama
University of Tsukuba
Faculty of Engineering , Information and Systems /
Center for Computational Sciences

History of parallel computer PAX(PACS) in U-Tsukuba

1978
1: PACS-9
Started by Prof. Hoshino and Kawai

1980
2: PAXS-32

1989
5: QCDPAX

1996
6: CP-PACS

Top1 in Top500 List

2006
7: PACS-CS

Service out in last Sep.

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

- Cooperation with Computational Scientists and Computer Engineers
- Target performance driven by application
- Continuous development with experience accumulation

Next PACS system toward Exa-scale era

- **HPC system toward post-Peta/Exa Scale era**
  - Should introduce some accelerator
  - Develop applications that are really accelerated by accelerator including new algorithm development
  - General purpose accelerator technology
    - GPGPU
    - Many Integrated Core(Ex. Intel MIC)
    - Cell Broadband Engine
    - FPGA
    - GRAPE-DR
  - Current most high performance/cost ratio accelerator ⇒ GPGPU

GPU Computing: current trend of HPC

- **GPU clusters in TOP500 2011/6**
  - 2nd 天河Tienha-1A (Rpeak=4.7PFLOPS)
  - 4th 星雲Nebulae (Rpeak=3PFLOPS)
  - 5th TSUBAME2.0 (Rpeak=2.3PFLOPS)
  - (1st K Computer Rpeak=8.8PFLOPS)
- **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 developing large scale applications accelerated by GPU in real computational sciences**

Problems 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
    ⇒ 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 GPUs for a feasibility study

Project Formation

- **HA-PACS (Highly Accelerated Parallel Advanced system for Computational Sciences)**
  - 2011/4 – 2014/3 3yeas project (the system will be maintain until 2016/3)
  - Project Office for Exascale Computational Sciences (Leader: Prof. Umemura)
    - Develop large scale GPU applications : 14 members

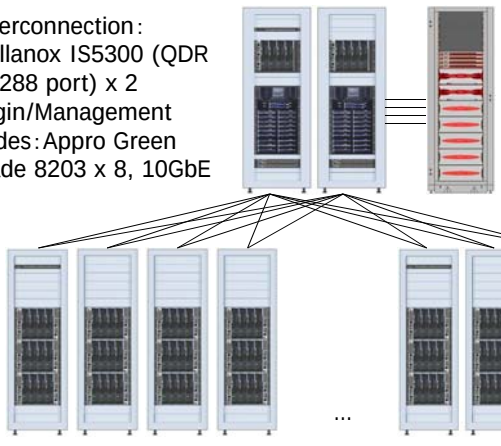
    Elementary Particle Physics, Astrophysics, Bioscience, Nuclear Physics, Quantum Matter Physics, Global Environmental Science, Computational Informatics, High Performance Computing Systems
  - Project Office for Exascale Computing System Development (Leader: Prof. Boku)
    - Develop two types of GPU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

HA-PACS

- HA-PACS system
  - Base Cluster Unit
    - Large scale GPU cluster with latest standard CPUs and latest standard GPUs
    - Using 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) Unit
    - 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 Unit

Interconnection: Mellanox IS5300 (QDR IB 288 port) x 2 Login/Management nodes: Appro Green Blade 8203 x 8, 10GbE I/F

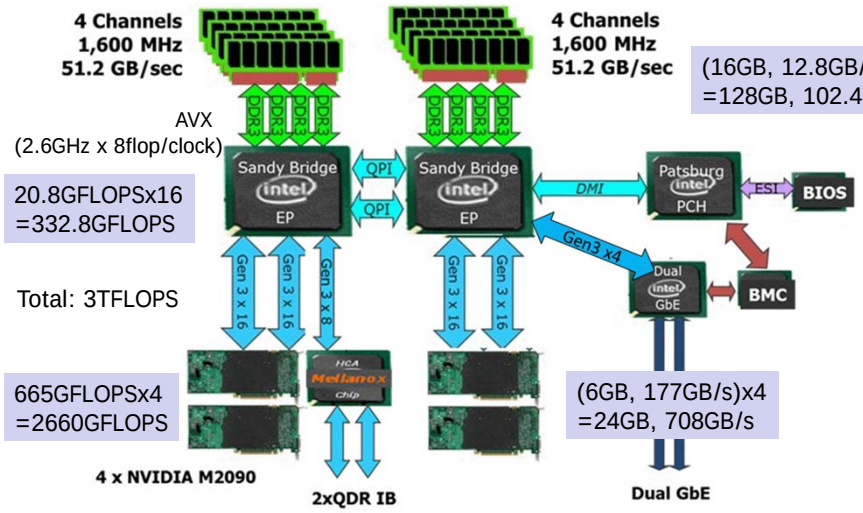


Storage: DDN SFA10000, connecting QDR IB, Luster File system, User area: 504TB

Computation nodes: Appro Green Blade 8204 (8U enc. 4 node) 268 node (67 enc./23 rack), 800TFLOPS

Total 26 racks

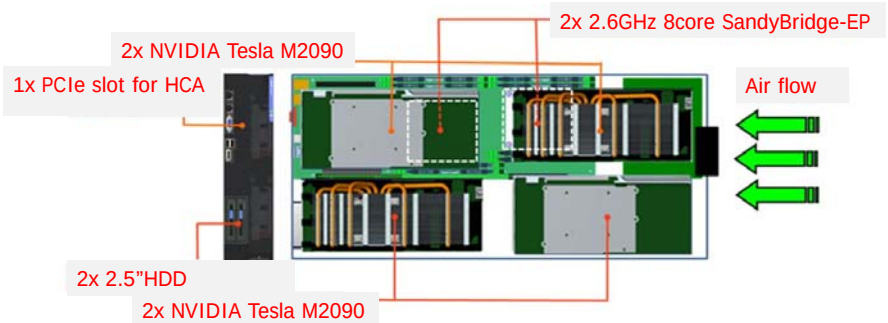
HA-PACS: Base Cluster Unit (computation node)

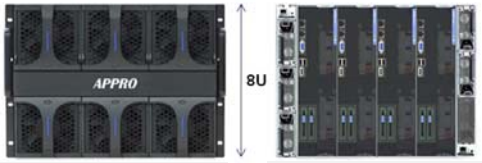


HA-PACS: Base Cluster Unit（GPU）

- **NVIDIA M2090**
  - Number of processor core: 512
  - Processor core clock: 1.3 GHz
  - Double 665 GFLOPS, Float 1331GFLOPS
  - PCI Express Gen2 ×16 system interface
  - Board power dissipation: < = 225 W
  - Memory clock: 1.85 GHz, size: 6GB with ECC, 177GB/s
  - Shared/L1 Cache: 64KB, L2 Cache: 768KB

HA-PACS: Base Cluster Unit（Blade system）





Front view　　Rear view

Power Supply Unit and Fan
- 8U enclosure
- 4 nodes
- 3 PSU(Hot Swappable)
- 6 Fans(Hot Swappable)

HA-PACS: Base Cluster Unit（Total）

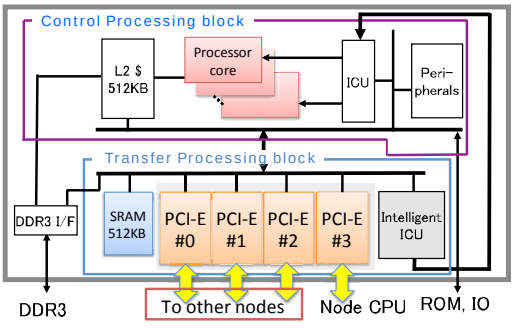
- 268 nodes are connected by 2 IB switches
- CPU: 89TFLOPS + GPU: 713TFLOPS = total 802TFLOPS
- CPU: Memory size 34TByte, Bandwidth 27TByte/sec,
  GPU: Memory size 6.4TByte, Bandwidth 190TByte/sec
- Bisection bandwidth 2.1TByte/sec
- Storage User Area 504TByte
- Power Consumption 408kW (monitoring available)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

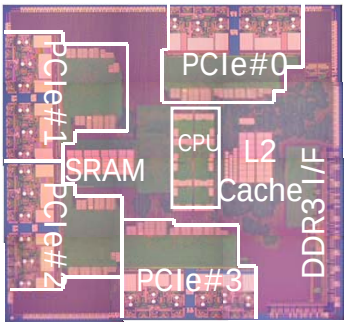
HA-PACS: TCA (Tightly Coupled Accelerator)

- **TCA: Tightly Coupled Accelerator**
  - Direct connection between accelerator (GPU)
  - 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**

PEACH Chip[Otani et al., ISSCC2011]





- **CPU: Renesas M32R 4core SMP (max. 400MHz)**
- **Communication Link: PCI Express Gen2 x4 lane (20Gbps) *4 port**

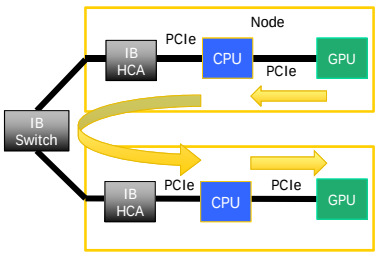
Improvement PEACH ⇒ PEACH2

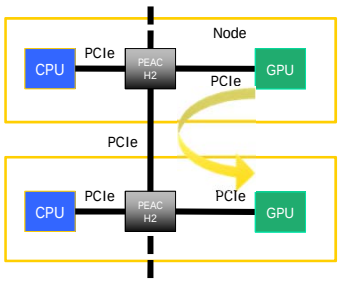
- **Bandwidth**
  - GPU: Gen2 x16, but PEACH: Gen2 x4
    => Gen2 x8 (Gen3 x8 will be desirable, but …)
  - Performance improvement of DMA controller
    => Chained DMA, Scatter/Gather
- **Latency**
  - Handling by embedded processor
    => Hardwired logic by FPGA
- **Using latest FPGA with PCIe Gen2 Hard IP**

Altera FPGA (Stratix IV GX)

HA-PACS/TCA with PEACH2

- **For HPC**
  - High bandwidth and low latency
  - Connecting to GPU
- **True GPU-direct**
  - Direct communication protocol between GPU
    ⇒ cooperate with NVIDIA





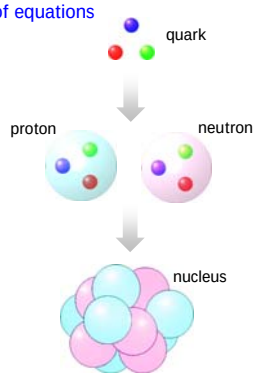
Schedule for TCA

- 2011/Middle: Considering about schemes for communication between GPUs on commercial FPGA evaluation board (Altera DK-DEV-4SGX530N) with developed daughter-card for PCIe cable.
- 2012/Begin: Develop evaluation board of PEACH2 with 4 port of PCIe Gen2 x 8.
- 2012/Middle: Produce several tens board of PEACH2 and Develop TCA system.
- 2013/Begin: Complete TCA system for application programming. Total performance of HA-PACS will be 1PFLOPS

HA-PACS Application (1) : Elementary Particle Physics

Multi-scale physics

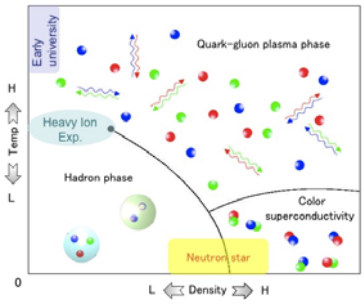
Investigate hierarchical properties via direct construction of nuclei in lattice QCD
GPU to solve large sparse linear systems of equations



Finite temperature and density

Phase analysis of QCD at finite temperature and density
GPU to perform matrix-matrix product of dense matrices

Expected QCD phase diagram



HA-PACS Applications (2) : Astrophysics

(A) Collisional N-body Simulation **(B) Radiation Transfer**

Globular Clusters

- Formation of the most primordial objects formed more than 10 giga years.
- Fossil object as a clue to investigate the primordial universe

Massive Black Holes in Galaxies

- Understanding of the formation of massive black holes in galaxies
- Numerical simulations of complicated gravitational interactions between stars and multiple black holes in galaxy centers.

➢ Direct (brute force) calculations of acceleration and jerks are required to achieve the required numerical accuracy

➢ Computations of the accelerations of particles and their time derivatives (jerks) are time consuming.

➢ Accelerations and jerks are computed on GPU

First Stars and Re-ionization of the Universe

- Understanding of the formation of the first stars in the universe and the succeeded re-ionization of the universe.

Accretion Disks around Black Holes

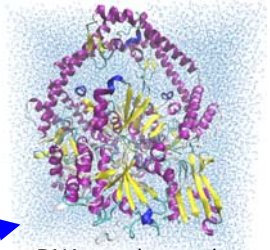
- Study of the high temperature regions around black holes

➢ Calculation of the physical effects of photons emitted by stars and galaxies onto the surrounding matter.

➢ So far, poorly investigated due to its huge amount of computational cost, though it is of critical importance in the formation of stars and galaxies.

➢ Computations of the radiation intensity and the resulting chemical reactions based on the ray-tracing methods can be highly accelerated with GPUs owing to its high concurrency.

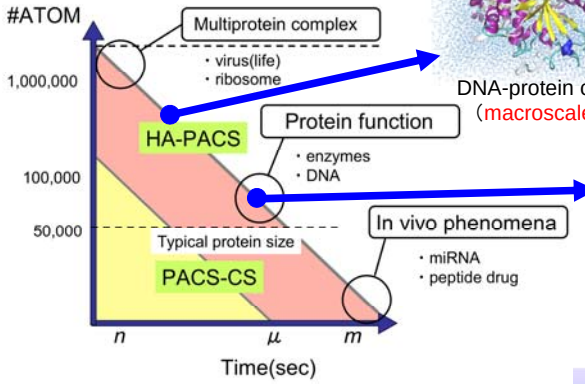
Application of HA-PACS(3) : bioscience



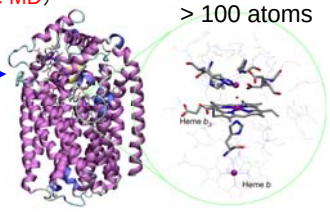
GPU acceleration

- Direct coulmb (Gromacs, NAMD, Amber)
- -2 electron integral

DNA-protein complex (macroscale MD)



QM region
> 100 atoms



Reaction mechanisms (QM/MM-MD)

Conclusion

- **HA-PACS: Next Generation GPU Cluster in Univ. Tsukuba**
  - Using Base Cluster unit, we develop large scale parallel application using GPU, and highly accelerated computational application will be produced.
  - We will develop TCA as direct connection between accelerators and study the feasibility of the next-generation accelerated computing.
- **Base Cluster Unit will be delivered in Jan. 2012. TCA will be developed in 2013**
- **Total 1PFLOPS system will be developed and production run of large scale computational application for GPU on it.**