# COMA (PACS-IX) Project

Taisuke Boku Deputy Directory, HPC Division Center for Computational Sciences University of Tsukuba



1 CCS Ext. Review 2014 2014/02/20

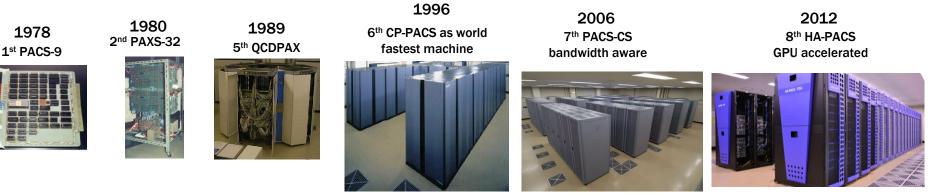
## Two Streams of Supercomputers at CCS

- 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



# Hisotry of PAX (PACS) MPP series

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



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



CCS Ext. Review 2014

2014/02/18

# PACS (PAX) Series

- MPP system R&D continued at U. Tsukuba for more than 30 years
- Naming history
  - (original) PACS : Processor Array for Continuum Simulation
  - PAX : Parallel Array eXperiment
  - (recent) PACS : Parallel Advanced system for Computational Sciences
- 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
  - $\rightarrow$  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



#### Next PACS with another accelerating devices

- HA-PACS : large scale highly dense GPU cluster
- Another accelerating device today many-core arch.
- Intel Xeon Phi (MIC: Many Integrated Core)
  - a number of simple cores
  - currently delivered as an accelerating device attached to CPU through PCIe bus (similar to GPU)
  - each core is available to run ordinary Linux on x86 ISA
- CPU core of Xeon Phi (KNC: Knights Corner generation)
  - similar to Pentium4 class x86, approx. 1GHz of frequency
     each core is relatively weak, but 512bit AVX SIMD instruction enhances floating point performance (FLOPS)
    - ☆ "throughput core" (vs "latency core")
  - 60 (or 61) cores + GDR5 memory to provide wide-bit high bandwidth memory access



## Toward JCAHPC system procurement

- Joint system procurement and operation with U. Tokyo on 2015 at Kashiwa Campus of U. Tokyo
- Currently targeting Many-Core processor for CPU
  - Xeon Phi will be available as "main CPU", not as "accelerator board"
  - Large performance improvement on each core, increasing # of cores, higher frequency, ...
  - We need much of experience to utilize throughput-core system for wide variety of HPC applications
  - Code tuning, new algorithm, new library, ...
  - Operating system for throughput-core (U. Tokyo)



We need a test-bed for these purposes



## What is COMA ?

- Cluster Of Many-core Architecture processor
- COMA
  - a famous "cluster of galaxies"
  - galaxy = cluster of stars (= many core)
  - cluster of galaxies = cluster of many-cores
- We will continue the name of PACS as a simple name of series of machines with "index code" (not unique name)
   COMA is PACS-IX



## **Basic specification of COMA**

- 9<sup>th</sup> generation of PACS (PACS-IX)
   (operated with HA-PACS in parallel)
- operation starts after shutdown of T2K-Tsukuba
  - deployment at end of March 2014
  - operation starts at 15<sup>th</sup> April
- System configuration
  - computation node : general CPU + MIC
  - node construction
    - CPU x 2: Intel Xeon E5-2670v2
    - MIC x 2: Intel Xeon Phi 7110P
    - Memory: CPU=64GB MIC=16GB (8GB x 2)
    - Network: IB FDR Full-bisection b/w Fat Tree
  - # of nodes: 383+10=393
  - peak performance: CPU=157.2 TFlops MIC=843.8 TFlops TOTAL: 1001 TFlops = 1.001 PFLOPS
- System delivery: Cray Inc.



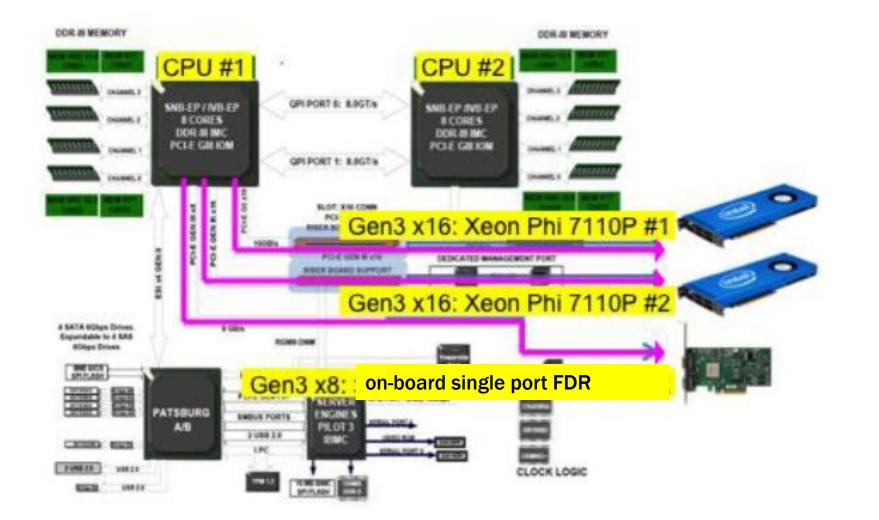
#### **Computation node**

- CPU (x2): Intel Xeon E5-2670v2 (Ivy Bridge)
  - 10 core/CPU, 2.5GHz
  - 200GFLOPS x 2 = 400GFLOPS
  - memory: 64GB
     DDR3 1866MHz x 4chan x 2CPU = 119.4 GB/s
- MIC (x2): Intel Xeon Phi 7110P
  - 61 core/MIC, 1.1GHz
  - 1.0736 TFLOPS x 2 = 2.1472 TFLOPS
  - memory: 8GB x 2 = 16GB GDR5 352GB/s x 2 = 704 GB/s
- Interconnection: InfiniBand FDR (on-board)
  - Mellanox Connect-X3
- Local HDD: 1TB x 2 (RAID-1)





## Block diagram of node

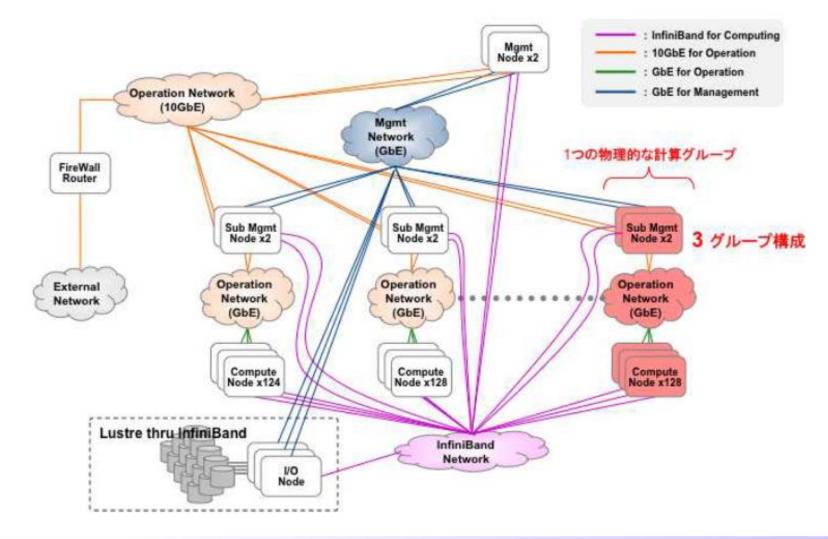


## System performance summary

- # of nodes: 393
  - peak performance
    - CPU: 157.2 TFLOPS
    - MIC: 843.8 TFLOPS
    - TOTAL: 1001 TFLOPS = 1.001 PFLOPS
  - memory capacity
    - CPU: 25.1 TB
    - MIC: 6.3 TB
- Interconnection: Fat-Tree full-bisection b/w
  - Bisection bandwidth: 2.75 TB/s

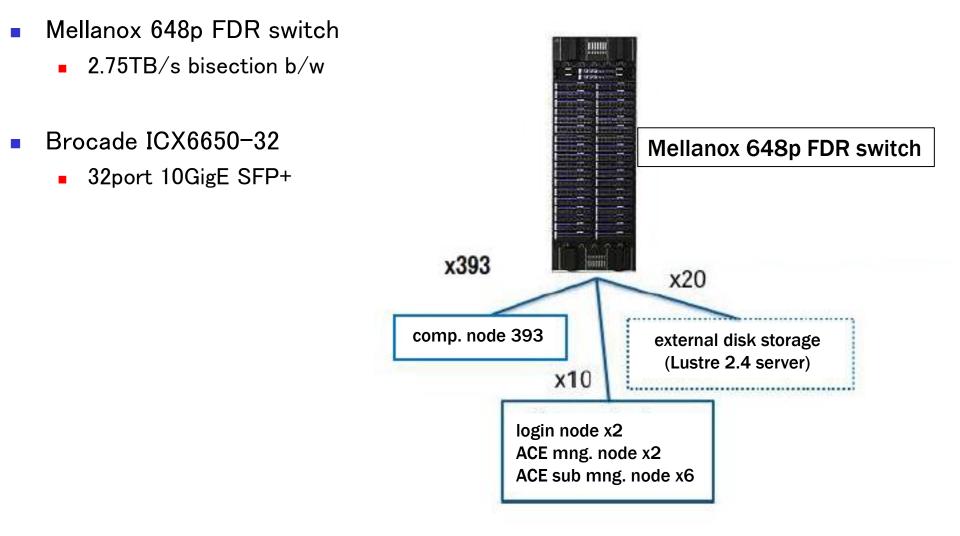


#### **Network construction**





## **Interconnection Network for Computtion**





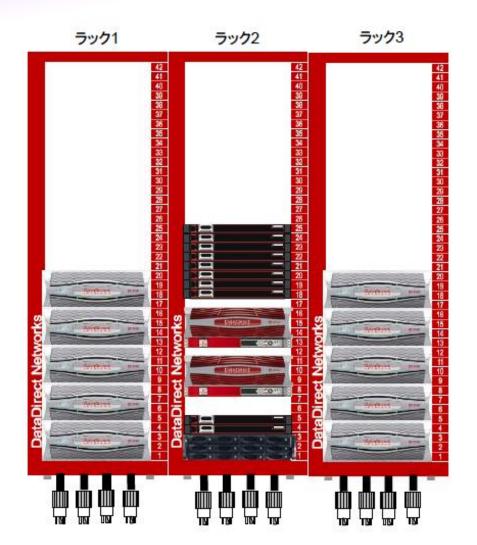
## System Image

Connection Result	Consule Radi	Contraction Reaction	Canada Rada	Companie Reals	Contrado Redi	Care Balais	Computer Name	Consult Radi	Contrade Resk	Canada Red	Companie Reals	Computer Reals	Consult Radi
	1	1	4				7					<b>0</b>	13
APC AREAS	APC AROOD	APC ARCES	APC MEDO	APC/AR000	APC MODE	APC AROUND	APC AROUND	APC AREAD	APC ARCES	APC AREOD	APC AREAD	ARC AROUND	APC AREAD
All Galled	At Castel	At Cashed	All Cashed	Al Coded	All Cardinal	As Called	As Galled	At Galed	An Control	All Casted	Al Calei	As Galled	All Cashed
8.08	B.MB.	49-041(04)	47 24 (10)	60 Opt (Opt)	47-34 (34)		40* 367 (34)	47 0 (194)	497 (0)(1(0)4)	4P 3/ 010	8.44	8.05	BLARE
8.05	BARK.	8.06	BARK .	8.44	47 3 g (April		4P 3 (F 34p-0	4702.04	B.OK.	BARK .	8.46	8.05	BLAR
8405	848		-	848			-		ALC: N	848	8/8	8405	
-	8.08		845	848			8.05	8.08	105	848	848		BAR .
102709-197	10000	100.00	COLUMN TW	UCTION THE	10704 10		10734 10	LOUGH BY	10734 107	12100.10	ECTIV TH	10709-107	12753 50
102708-78	102706 797	12706-797	102109-199	127106-792	100709-787		102709-792	102704 707	12734-74	102109-09	12708-79	102739-79	102204 297
100709-100	ILLING BUT	1298.90	12109.100	12710.70	100709-30		100709-100	122504 50	12754 50	1210.79	12104.70	102709-102	122708-90
102709-792	102708-797	12708-797	KOTON THE	127106-797	100709-707		1027109-779	102704-797	12704-757	100000	127109-797	100708-78	102704-797
102704-107	RECEIPT OF	12734.50	Restore The	10000	ICCOME THE		100704-00	100004 90	12554.50	12102.00	10000	100708-107	ICCIDE TO
102708-787	102704-397	10708-00	ICCTOR THP	107106-797	100708-767		1027104-107	102704-797	10704-79	1022 CORE TRAF	107106-797	100708-787	102708-797
100104-00	10004 99	12708.90	100.10	1000 10	100108-00		10704 10	10004.90	10,0140	1207.10	1000	10704 10	102104 90
100708-199	102504 797	10706-97	ICTOR THE	10708 79	100708-767		1027394 789	10008-99	10734797	12104.19	10108-19	100709-799	102704-797
102739-397	ICOMA TOP	12708-727	COLOR THE	107109 79	100709-76*	Melanos 648 Port IB	102739-79	102304 797	12734787	1210.79	100 M TV	102739-79	102704-707
102708-19	10000 77	12706-397	102109-119	10708-79	102736-767	Managed Over Policity	102708-78	10,00 77	10734 77	10106.19	10708 79	102739-79	ICCION TOP
	10000 70	12738-50	1010.10	10000	100704 10		107104 107	10000 00	101810	1210.19	ECON IN	10708 10	122504 500
102739-797	102706 797	12720-227	102109-79	127106-79	100709-755		102739-78	102204-202	10724-72	102109-09	12709-79	102720-707	102204 297
102709-799	102708-707	12708-99	102109-178	127106-797	100709170		1027109-170	102704-797	12734-757	100108-09	127106-797	1007094 799	102704 797
102709-199	102704 797	10708-99	ICCION THE	102108-79	100709-787		1007109-170	102704-797	10704-757	100108-09	12708-79	100708-78	102704-797
ICCOM TO	RECEIPT OF	10000.00	NAME OF TAXABLE PARTY.	10000	ICCOME NO		100704-00	102103-109	12534.50	Contraction of the	600 M 10	100708-107	102108-90
100709-199	102708-797	12739-797	ICCION THE	102108-79	100709-782		1027109-170	102708-797	10704-76	102109-09	12709-79	100708-78	102708-797
102739-78	102706-797	12709-70	102108-79	102108-79	100709-767		102739-79	102704-70	12734-757	100109-09	127.05.79	102739-78	102704-70
100708-78	ICCIDIN THE	12734-50	ICCOM THE	10709-79	100709-76*		100709-199	10230A 99	10754-97	12108.19	10710A TW	1027394 789	ICCIDIA TRP
10736 10	102504 907	1278.50	10000	10708-19	10734 10		102734 10	ICON IV	12734 30	1210.10	10000 70	10734 10	ICCOR TO
102709-797	102304-397	12726.727	SCOOL OF	10706.79	100709-707		102708-70	ICOM IN	12734.79	1210.79	12708.79	100709-107	102708-707
102709-797	10000 70	1270.32	102109-179	127106-79	100709-787		102739-79	10000	12734.72	12108.79	12708.79	107106-797	10000 20
102708-78	102706-797	12706-797	KOTON THE	102109-79	100709-787		100709-79	102304-397	12724-797	102108-07	122108-292	102739-79	102708-797
100704-10	102504 90	12734.50	10104.10	10000	100709-107	100 0008 00 540	100704-10	100004-909	12734-90	12104.70	12104 10	100704-00	102304 90
100709-197	102704-797	12706-707	ICCOM THP	107106-797	100708-767		100709-789	102304-99	10704-07	102106-09	12106-79	100708-78	102708-797
102708-78	102306-707	12708-767	102208-789	102109-79	1027291707	191.000	102739-78	102306-397	10734/72	102308-005	127106-79	102739-78	102304-397
100709-199	102704 797	10708-99	KOTON THP	10709 79	1027391787	173.000	100709-79	102704-797	10704797	10106.09	10109.79	100709-199	102704-797
10754.78	10004 10	12708-50	COOK NO	1010110	100709-90		10704 10	102504 90	10,04.00	12104.10	1010110	10794 10	102504 90
8.05	BACK		8498	242	1027291707	3.05	102739-787	102704-707	Real Procession	8498	848	8.05	
-	8.06	8.04	BANK .	RAK.	8.08	8.08	8.05	B.VE	8.06	BAR	8.48	8.05	BLUE.
Ad Heaptoni	8.08	Rate Management	845	Ref Management	Lage L	ACE Management	Lingto	Rd Management	LUX LUX	The Management	848	Ad Management	Red Management
8.08	8.48	Read and a r	84.9	848	8.05		8.05	BAR.	844	848	848	8.00	BAR .
845	848		845	845	1 100	ACT Management	-	-	-	848	848	-	
8.08	8.46	8.05	BANK .	848	8.05	8.05	8.05	B.VK	B.L.B.	84.6	848	8.05	8.48
8.08	8.46	8.0%	BASE .	846	8.05	8.05	8.05	B.ME	846	BASE .	846	8.05	8.48
													ز <b>السب</b>
<u>н</u> р			ц р		<b>D</b>	н Г		14 1	1 P	ц — р			/H – P



#### Storage

- Shared file server
  - RAID6 + Luster
  - OST: DDN SS8460 x 10
    - 4TB x 550基
  - OSS x 8
  - Controller: DDN SFA12K-40
  - InfiniBand FDR x8 (OSS)
  - User space: 1.5PB
  - Flat access from all computation nodes
- NFS server
  - 12TB for /home, etc.





# Software (OS, programming)

- OS: Red Hat Enterprise Linux (login server)
- OS: CentOS (compute node)
- Cluster management: ACE
- Job scheduler: SLURM
- Programming environment:
  - Intel Cluster Studio XE2013 (Composer/XE)
  - Fortran95/C/C++
  - Intel MPI



## Programmin on MIC

- Linux is permanently running on each MIC (KNC)
- Two programming model
  - Native Mode: Direct execution of code on each MIC
    - Multi-threaded code (OpenMP) can be executed
    - Max 240 threads with hardware thread control
    - No I/O (HDD) HDD on host CPU is mounted through NFS
    - MPI is possible for MIC-to-MIC communication through host InfiniBand
  - Offload Mode: "offloaded" part of host CPU code runs on MIC
    - Composer XE (Intel extended compiler for MIC)
    - Code parts for "device" (MIC) is explicitly described
    - Writing OpenMP in offloaded part, it can be executed in thread parallel on manycores



## Example of offload programming

```
#include <stdio.h>
#include <omp.h>
#define SIZE 1000
int main()
```

```
int inarray[SIZE], sum, validsum;
int i;
int nth;
```

```
validsum=sum=0;
for(i=0; i<SIZE; i++){
    inarray[i]=i;
    validsum+=i;
}
nth=0; // for checking
```

```
#pragma offload target(mic:0) in(inarray:
   alloc_if(1) free_if(0)) out(sum) out(nth)
  int lsum = 0;
  int i:
#pragma omp parallel for default(none) ¥
       shared(inarray) reduction (+:lsum)
  for(i = 0; i < SIZE; i++)
    Isum += inarray[i];
   sum = lsum:
#pragma omp parallel
#pragma omp master
   nth = omp get num threads();
 printf("sum = \%d validsum = \%d + n", sum, validsum);
 printf("num thread = %d¥n", nth);
```



#### Job control on MIC

- Three types of partitions (job queue)
  - CPU partition
    - using 16 cores out of 20 on general CPUs on each node
    - ordinary programming for multi-core (OpenMP + MPI)
  - MIC partition
    - dedicating 4 cores out of 20 on general CPUs to control two MICs
    - (maybe) running with offload model
  - Mixed partition
    - all CPU and MIC resources is available for the job
    - hybrid program, work sharing, MIC native mode execution

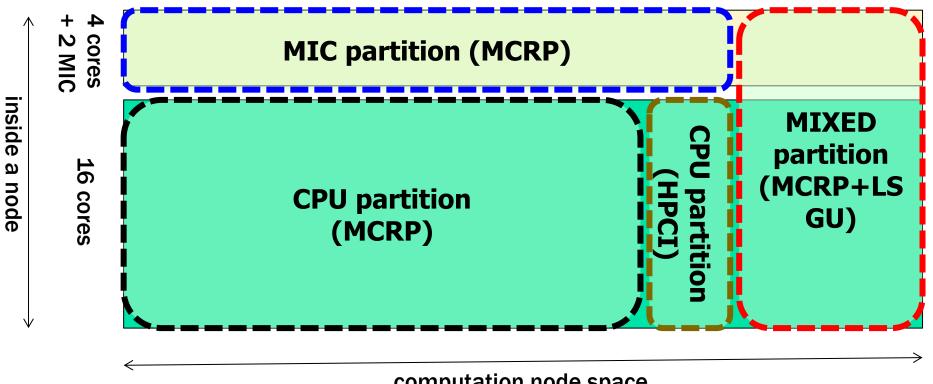


## **Operation program of COMA**

- Multi-disciplinary Collaborative Research Program
  - supporting advanced computational science/engineering
  - free of charge (application + ext. review)
  - CPU, MIC, Mixed
- HPCI
  - networking nation-wide all supercomputers under MEXT by single-sign-on system
  - free of charge (application + HPCI review committee)
  - CPU
- Large Scale General Use
  - assigning job to node-to-node manner
  - charge CPU utilization cost
  - Mixed
- Official operation starts on 15<sup>th</sup> of April, 2014



## System partitioning for job queue



computation node space



#### User support and education (planned)

- Tutorial for many-core system utilization
  - by support of Intel and Cray
  - HPC Seminar (Summer + Winter) by CCS
  - International HPC School
- Programming support
  - currently, the policy is the same as HA-PACS

⇒ we basically hope users & projects to write and tune their code with their own effort

▷ MCRP strongly recommends to make a team of computational scientists and computer scientists



#### Summary

- COMA (PACS-IX) is a new cluster in CCS deployed in the end of March 2014, based on Intel Xeon Phi (MIC) accelerating devices
- 393 nodes, 786 MICs (Intel Xeon Phi 7110P) for 1 PFLOPS of peak performance
- With advanced many-core architecture, accelerated computing with high-density, high-performance and low-power is available
- Three types of operation model: CPU, MIC, mixed
- Program: CCS Multidisciplinary Collaborative Research, general use (not free)
- Operation starts at 15<sup>th</sup> April, 2014

