HPCL State Key Laboratory of High Performance Computing

Overview of Tianhe-2 (MilkyWay-2) Supercomputer

Yutong Lu

School of Computer Science, National University of Defense Technology;
State Key Laboratory of High Performance Computing, China
ytlu@nudt.edu.cn

国防科学技术大学
National University of Defense Technology

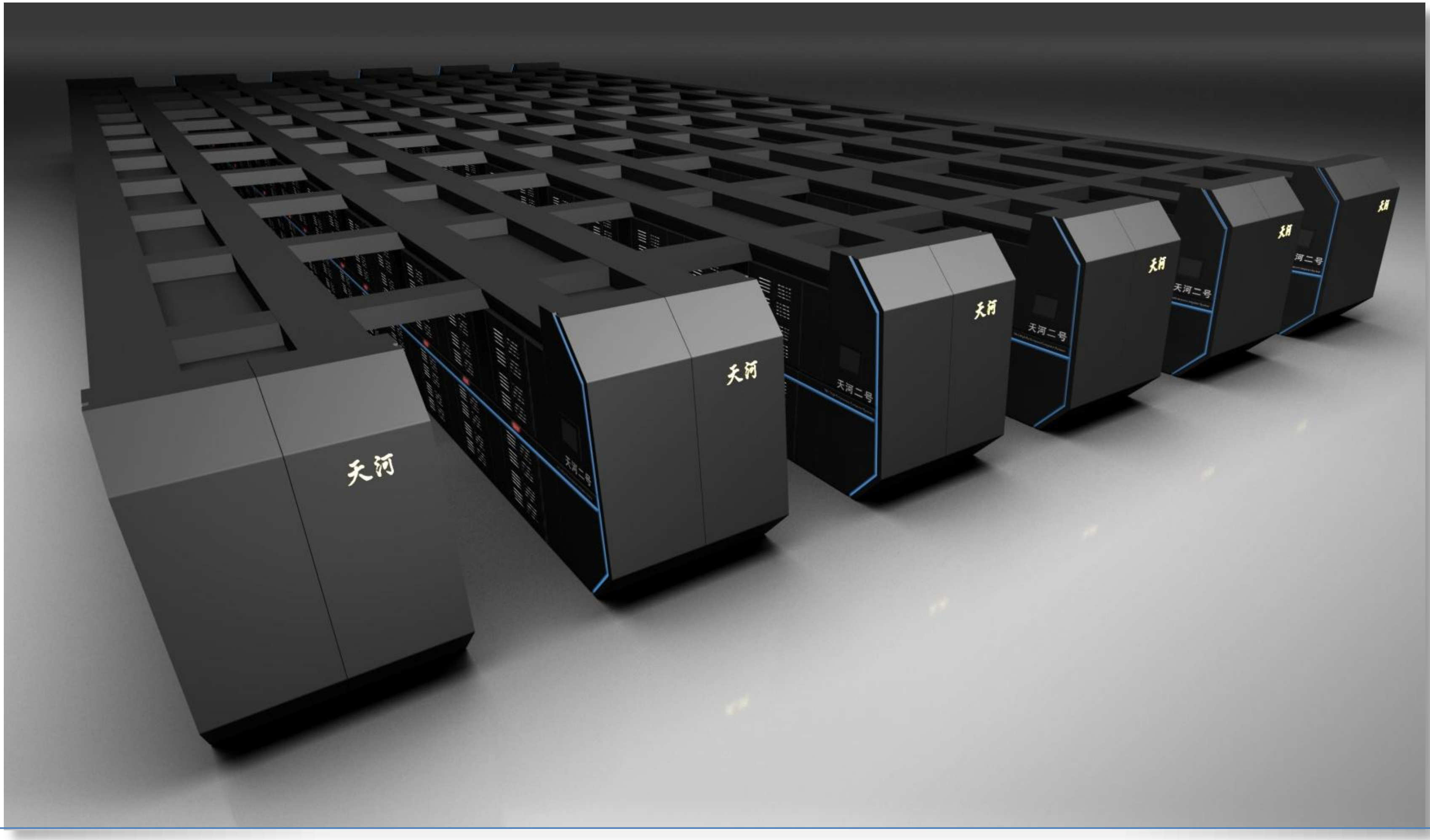
Outline

- Motivation
- Specification
- Hardware & Software
- Applications

Motivation

- ~100 petaflops system
  - ◆ 863 High tech. Program of Chinese Government
  - ◆ Government of Guangdong province and Government of Guangzhou city
- NSCC-GZ
  - ◆ Open platform for research and education
  - ◆ Public information infrastructure
- Goal
  - ◆ Scalability
  - ◆ Power consumption
  - ◆ Resilience
  - ◆ Usability

Motivation



Tianhe-2 (Milkyway-2) Supercomputer

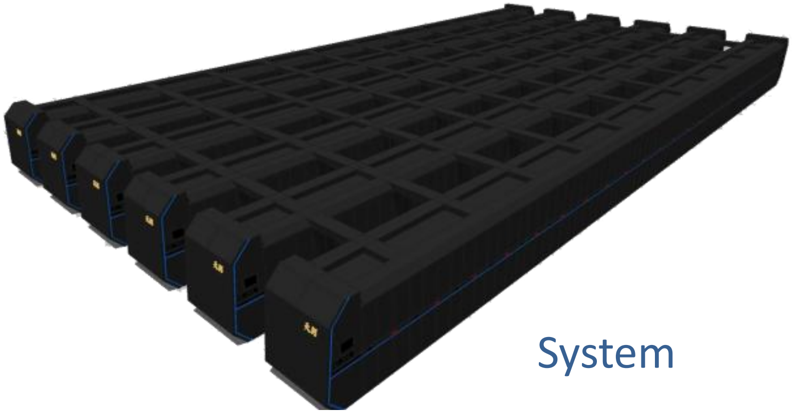
Specification

- **Hybrid Architecture**
  - ◆ Xeon CPU & Xeon Phi

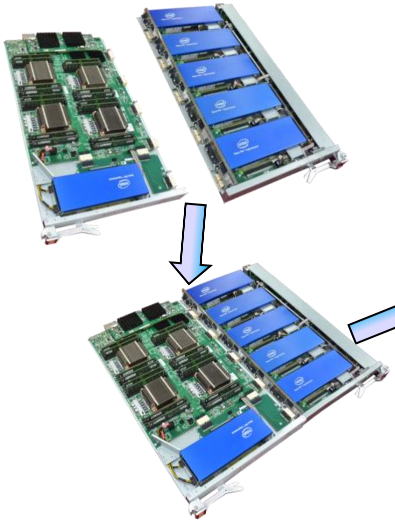
Items	Configuration
Processors	32000 Intel Xeon CPUs + 48000 Xeon Phis + 4096 FT CPUs Peak performance is 54.9PFlops, HPL
Interconnect	Proprietary high-speed interconnection network TH Express-2
Memory	1.4PB in total
Storage	Global shared parallel storage system, 12.4PB
Cabinets	125+13+24=162 compute/communication/storage Cabinets
Power	17.8 MW (1902MFlops/W)
Cooling	Closed Air cooling system

From Chips to Entire System

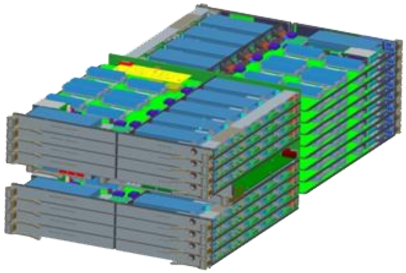
- ◆16000 compute nodes in total
- ◆Frame: 32 compute Nodes
- ◆Rack: 4 Compute Frames
- ◆Whole System: 125 Racks



System



Compute Node

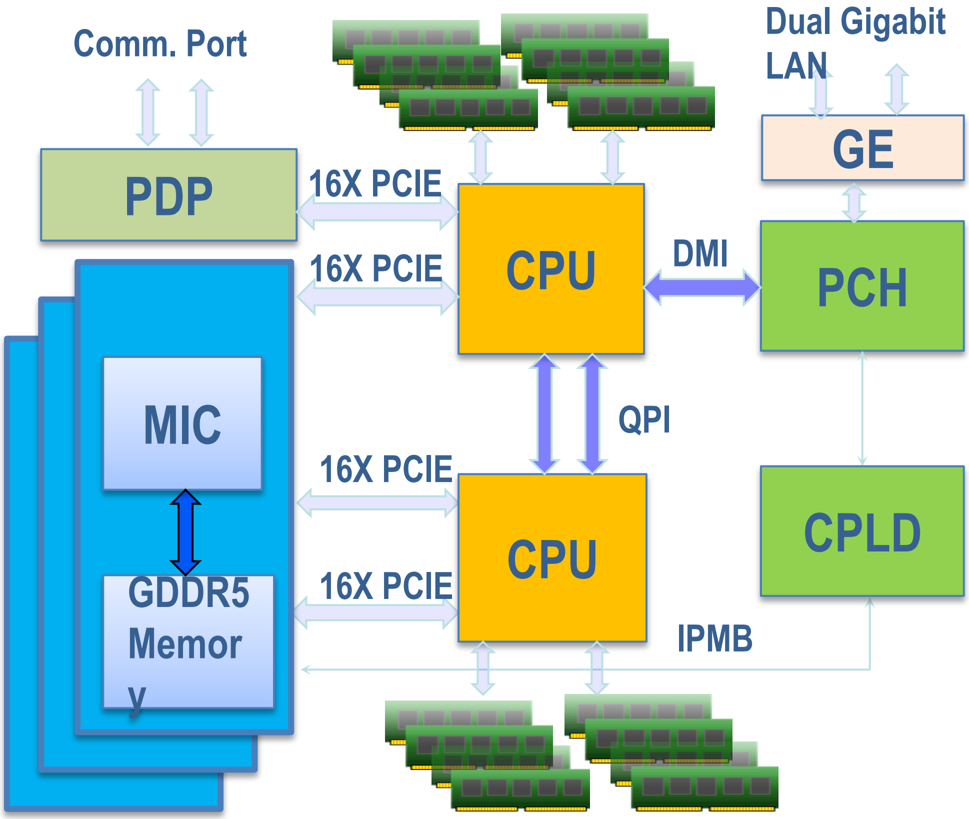


Compute Frame

Compute Rack

Compute Node

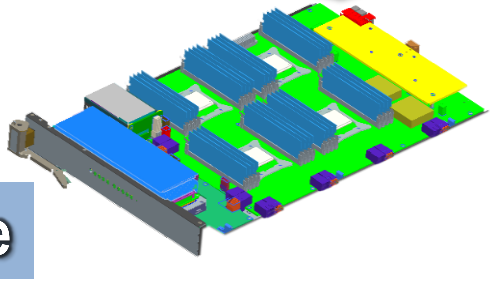
- Neo-Heterogeneous Compute Node
  - ◆ Similar ISA, different ALU
  - ◆ 2 Intel Ivy Bridge CPU + 3 Intel Xeon Phi
  - ◆ 16 Registered ECC DDR3 DIMMs, 64GB
  - ◆ 3 PCI-E 3.0 with 16 lanes
  - ◆ PDP Comm. Port
  - ◆ Dual Gigabit LAN
  - ◆ Peak Perf. : 3.432Tflops



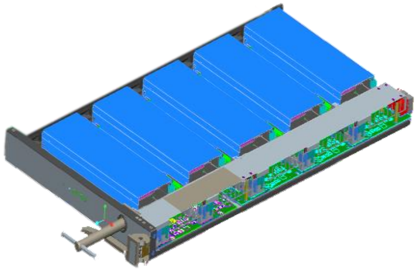
Compute Node

- Compute Blade = CPM Module + APU Module

4CPUs and 1 Intel Xeon Phi



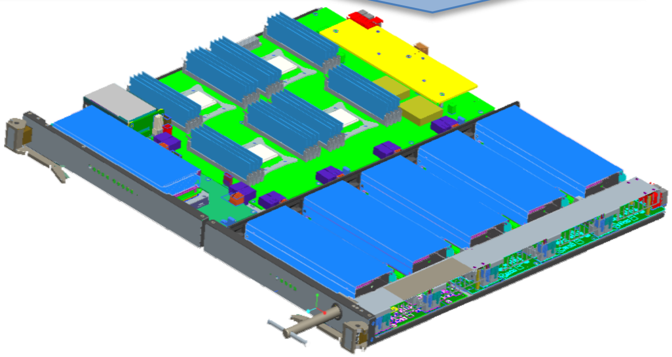
CPM module



APU module

5 Intel Xeon Phis

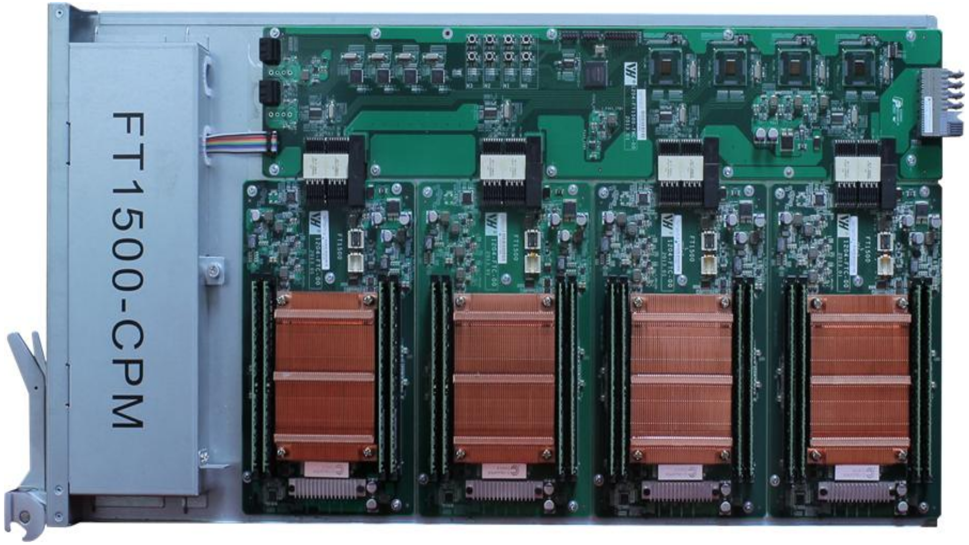
2 Compute Nodes with 128G memory and two comm. ports



Compute Blade

Operate Node

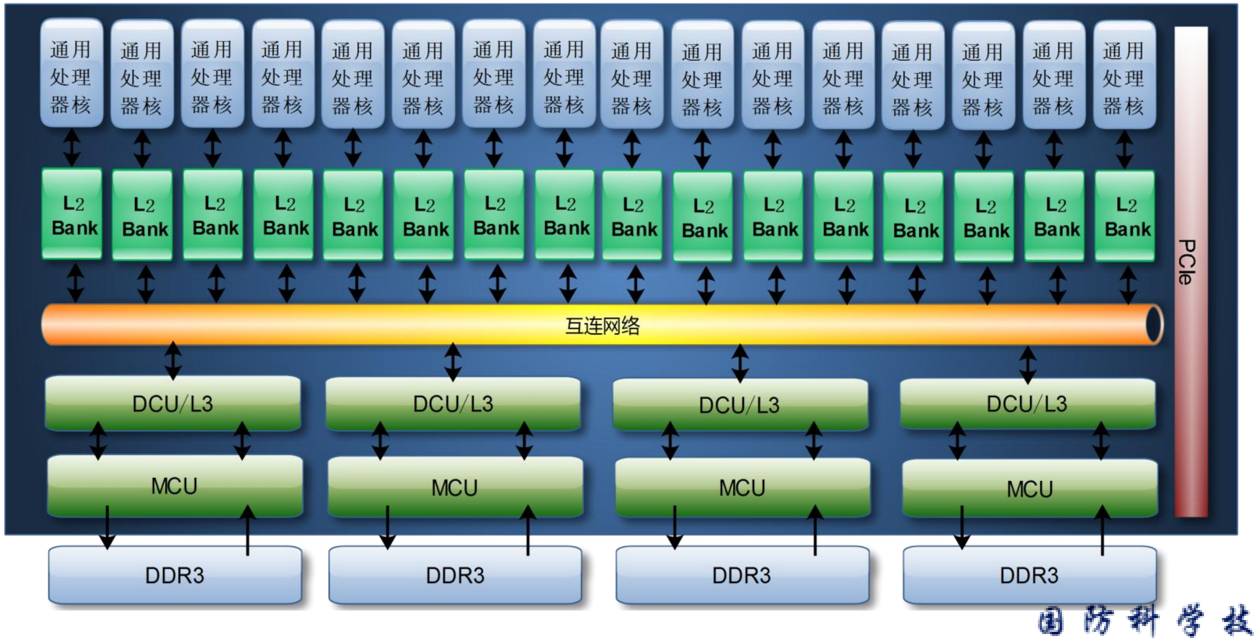
- 4096 FT-1500 processor based operation nodes
  - ◆Performance 144GFlops
  - ◆Four DDR3 channels
  - ◆One 16x PCIE 2.0



Blade of FT-1500

FT-1500 CPU

- 4096 FT-1500 processor based operation nodes
  - SparcV9， 16 cores， 4 SIMD
  - 40nm, 1.8GHz
  - Performance： 144GFlops
  - Typical power: ~65W



I/O node

- **Storage system**
  - **256 I/O nodes and 64 storage servers with total capacity of 12.4PB**
- **I/O node**
  - **2TB SSD storage**
  - **Burst I/O bandwidth: 5GB/s**
  - **PDP Comm. Port**
  - **IB QDR storage network Port**

IO Module

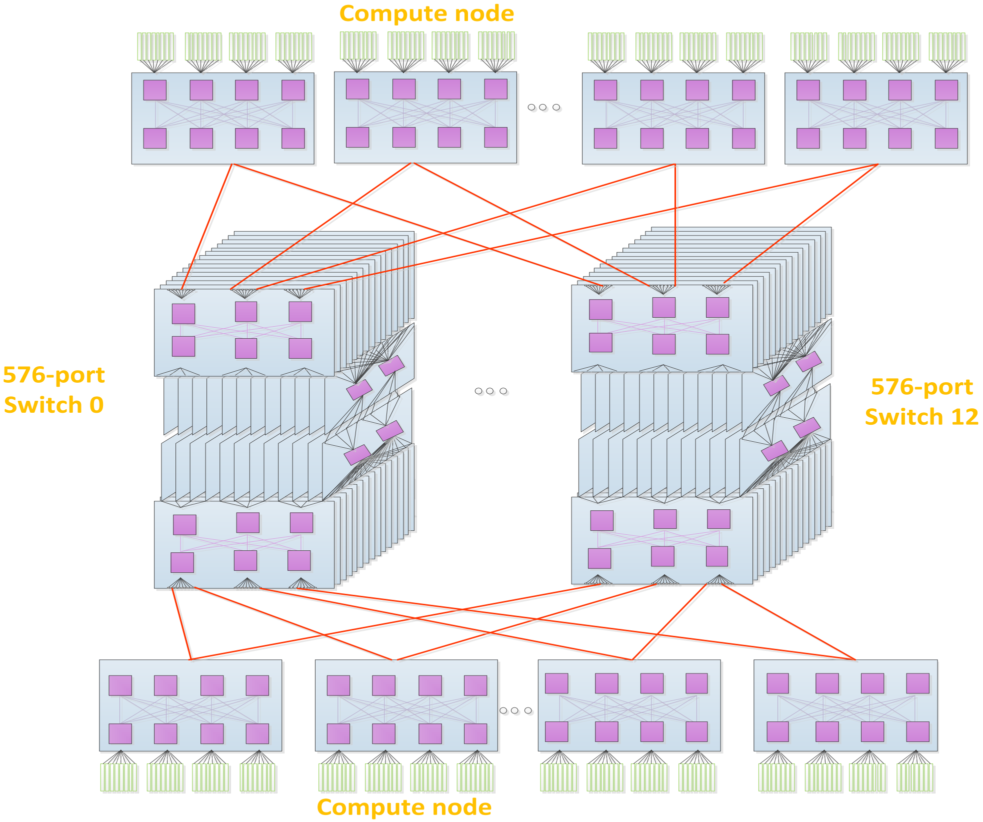
SSD Module

ION Blade(2I/O nodes)

Interconnection network

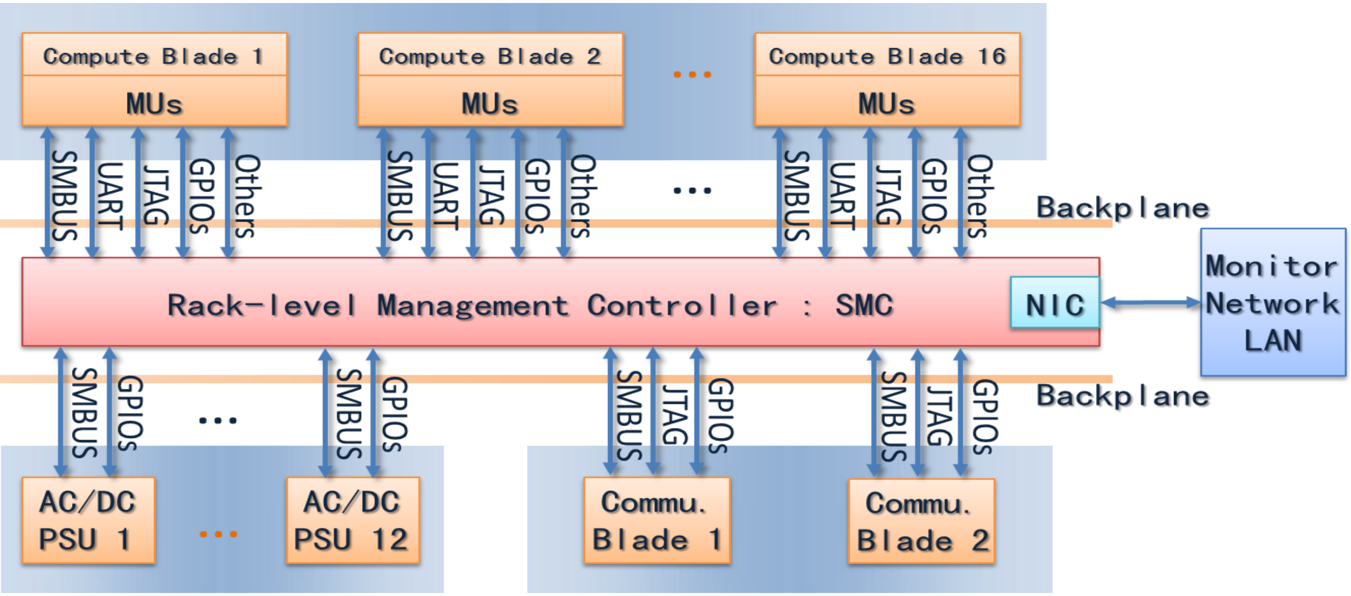
■ TH Express-2 interconnection network

- ◆Fat-tree topology using 13 576-port top level switches
- ◆Opto-electronic hybrid transport tech.
- ◆Proprietary network protocol
- ◆NRC +NIC



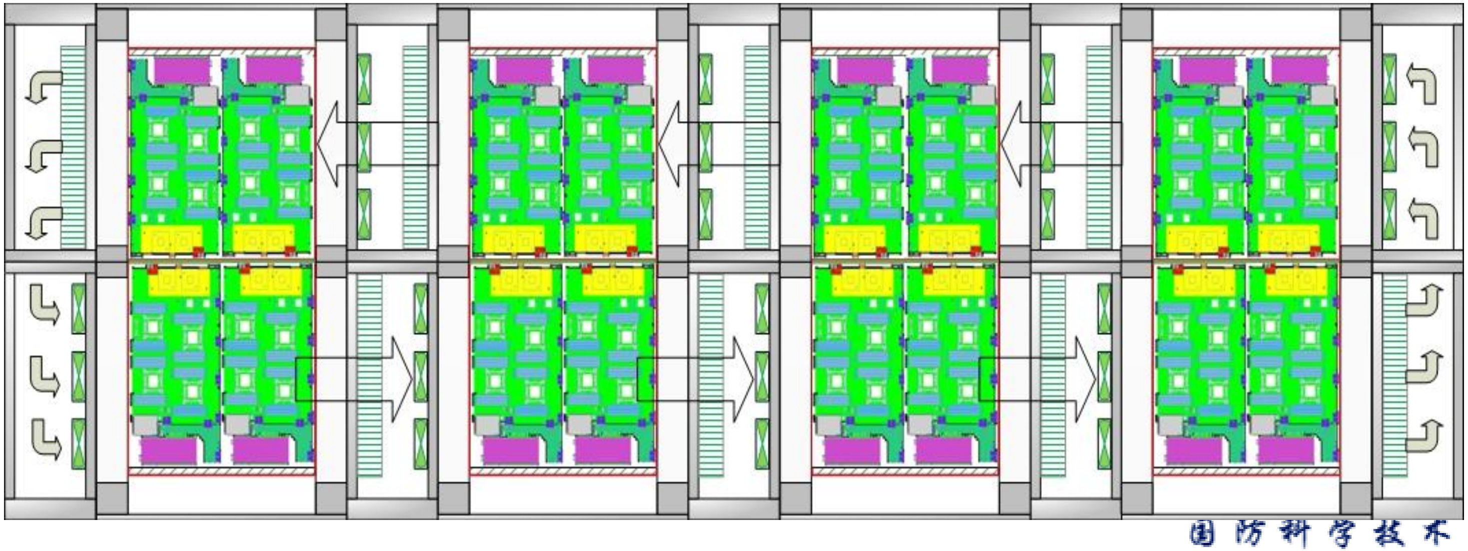
Monitor System

- Three levels of monitor entities
  - System-level : MCC
  - Rack-level : SMC
  - Board-level : MU
- Gigabit Ethernet for monitoring

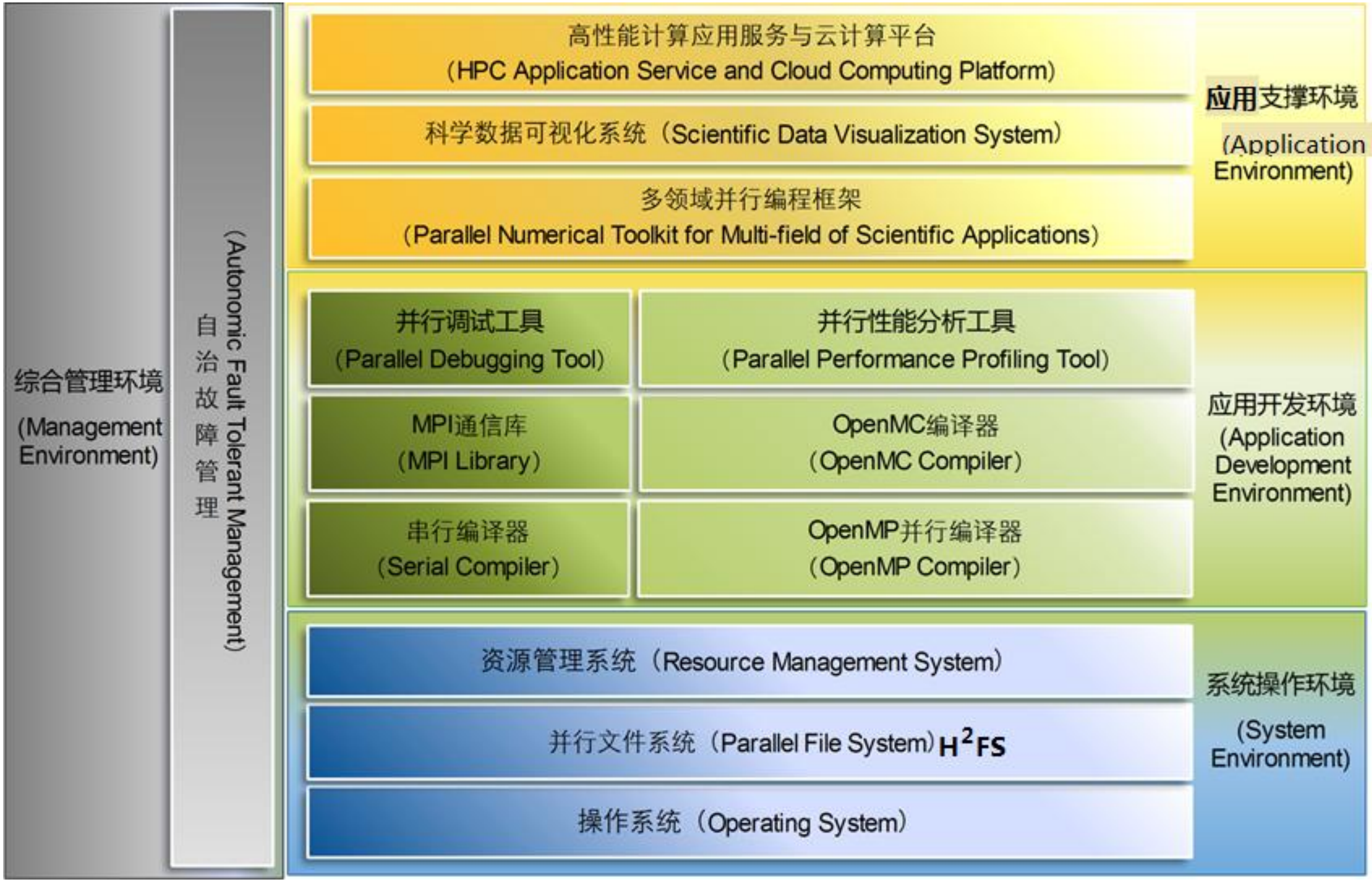


Cooling System

- Cooling Type
  - ◆ Close-coupled chilled water cooling
- Customized Liquid Cooling Unit
  - ◆ High Cooling Capacity: 80kW
- NSCC-GZ will use city cooling system to supply cool water to LCUs



HPC Software stack



综合管理环境
(Management Environment)

自治故障管理
(Autonomic Fault Tolerant Management)

高性能计算应用服务与云计算平台
（HPC Application Service and Cloud Computing Platform）

科学数据可视化系统（Scientific Data Visualization System）

多领域并行编程框架
（Parallel Numerical Toolkit for Multi-field of Scientific Applications）

应用支撑环境
(Application Environment)

并行调试工具
（Parallel Debugging Tool）

并行性能分析工具
（Parallel Performance Profiling Tool）

MPI通信库
（MPI Library）

OpenMC编译器
（OpenMC Compiler）

串行编译器
（Serial Compiler）

OpenMP并行编译器
（OpenMP Compiler）

应用开发环境
(Application Development Environment)

资源管理系统（Resource Management System）

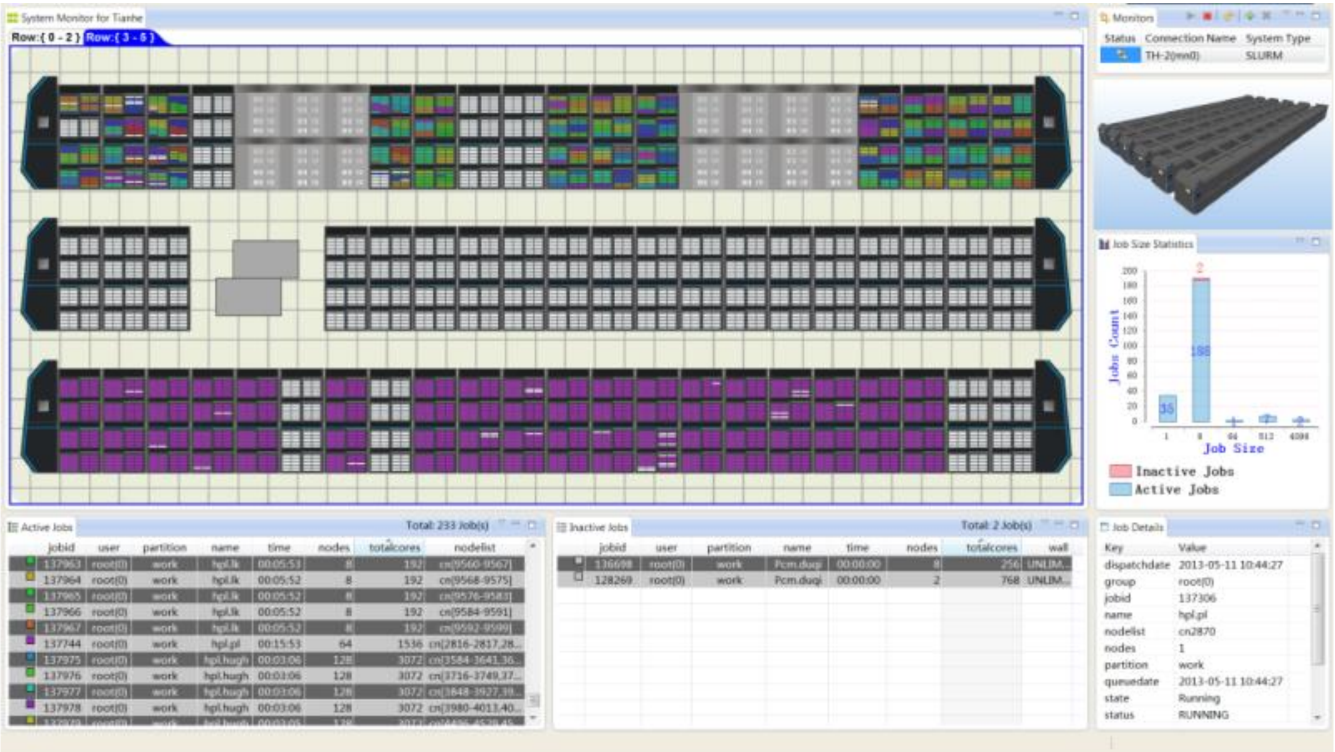
并行文件系统（Parallel File System）H^2FS

操作系统（Operating System）

系统操作环境
(System Environment)

OS & RMS

- **Operating System**
  - Kylin Linux
- **Resource manage system**
  - Power-aware resource allocation
  - Multiple custom schedule policies



Status	Connection Name	System Type
	TH-2(mn0)	SLURM

Active Jobs Total: 233 Job(s)

jobid	user	partition	name	time	nodes	totalcores	nodelist
137963	root(0)	work	hpl.lk	00:05:53	8	192	cn[9560-9567]
137964	root(0)	work	hpl.lk	00:05:52	8	192	cn[9568-9575]
137965	root(0)	work	hpl.lk	00:05:52	8	192	cn[9576-9583]
137966	root(0)	work	hpl.lk	00:05:52	8	192	cn[9584-9591]
137967	root(0)	work	hpl.lk	00:05:52	8	192	cn[9592-9599]
137744	root(0)	work	hpl.pl	00:15:53	64	1536	cn[2816-2817,28...
137975	root(0)	work	hpl.hugh	00:03:06	128	3072	cn[3584-3641,36...
137976	root(0)	work	hpl.hugh	00:03:06	128	3072	cn[3716-3749,37...
137977	root(0)	work	hpl.hugh	00:03:06	128	3072	cn[3848-3927,39...
137978	root(0)	work	hpl.hugh	00:03:06	128	3072	cn[3980-4013,40...

Inactive Jobs Total: 2 Job(s)

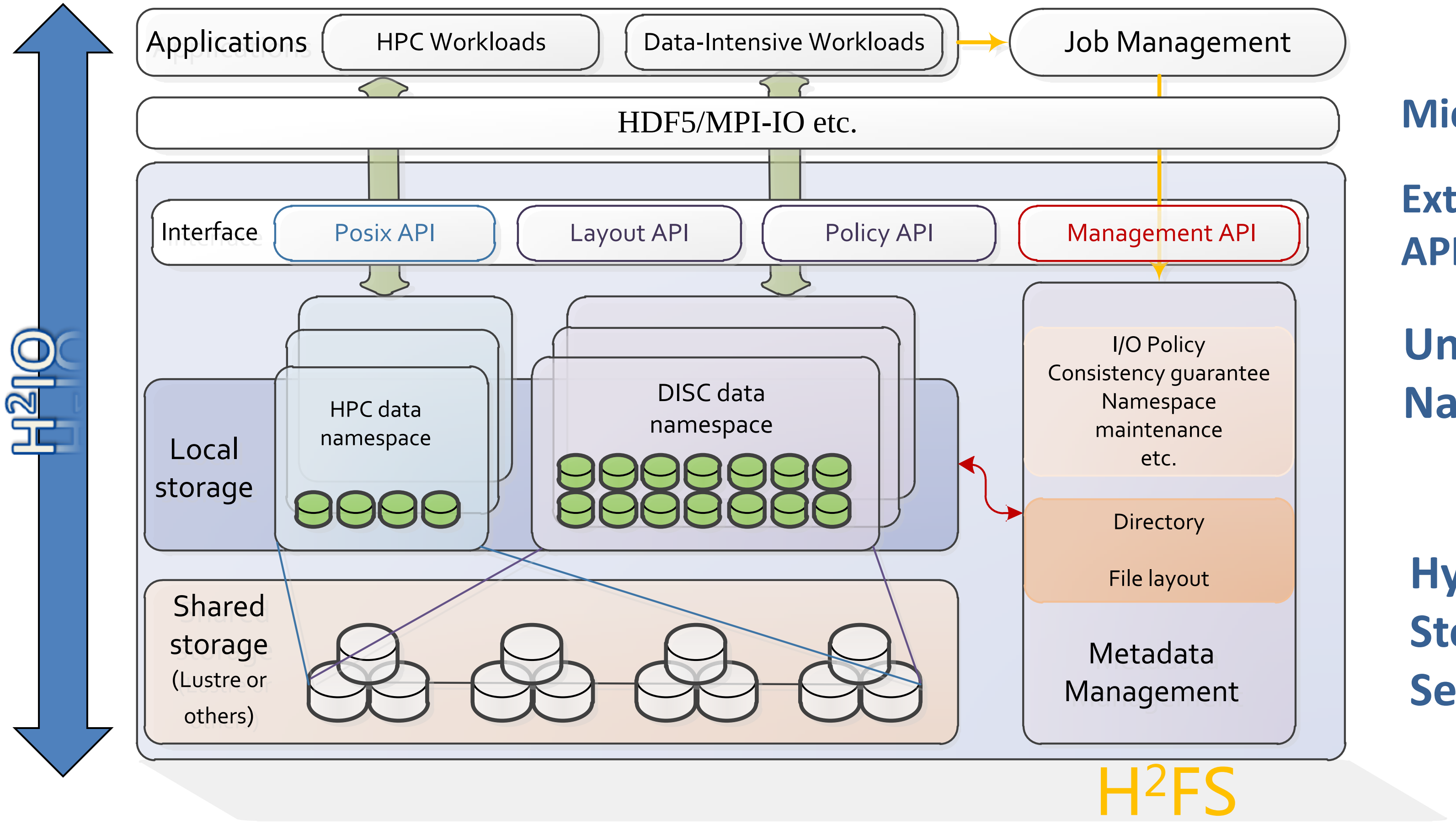
jobid	user	partition	name	time	nodes	totalcores	wall
136698	root(0)	work	Pcm.duqi	00:00:00	8	256	UNLIM...
128269	root(0)	work	Pcm.duqi	00:00:00	2	768	UNLIM...

Job Details

Key	Value
dispatchdate	2013-05-11 10:44:27
group	root(0)
jobid	137306
name	hpl.pl
nodelist	cn2870
nodes	1
partition	work
queuedate	2013-05-11 10:44:27
state	Running
status	RUNNING

Parallel Filesystem-H^{2}FS

- **H2IO: Hybrid and Hierarchy I/O stack**



Middleware

Extended I/O API

Unified Namespace

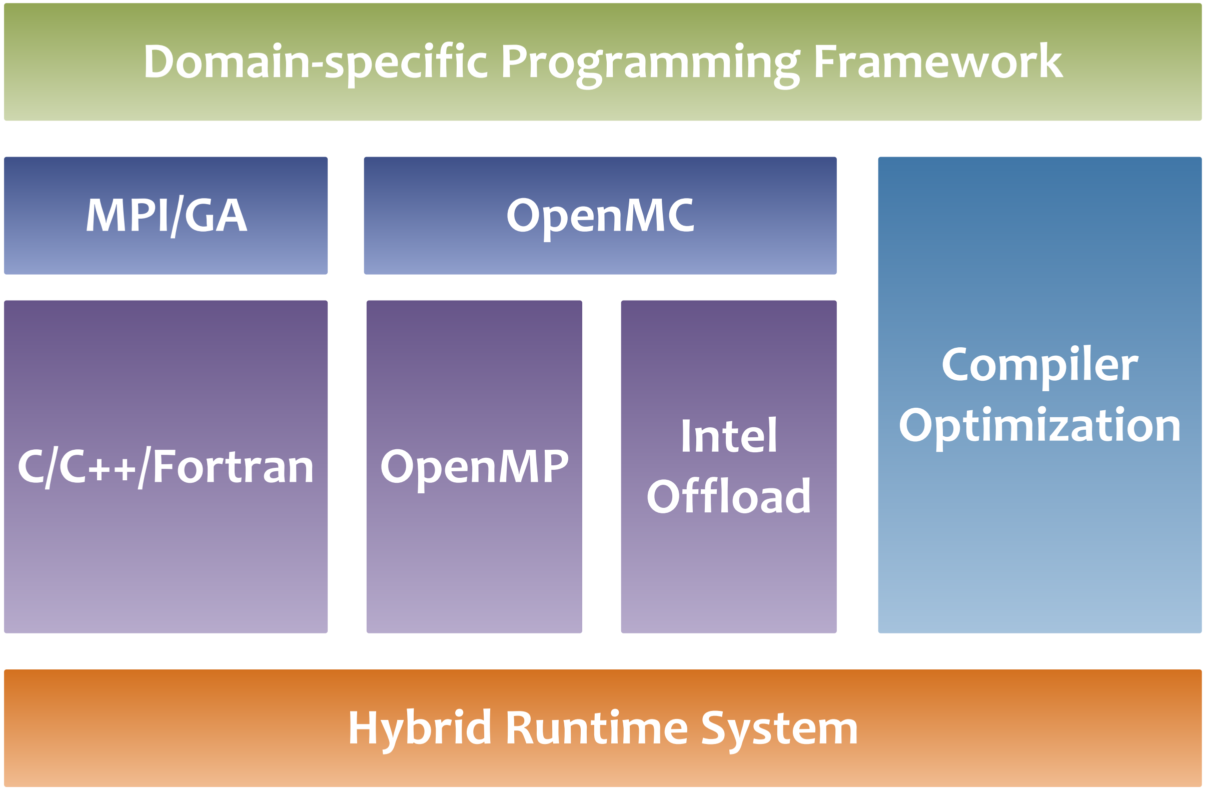
Hybrid Storage Service

Hybrid and locality aware file layout

国防科学技术大学
National University of Defense Technology

Compile system

- **Programming Languages**
  - C/C++/Fortran
  - OpenMP
  - OpenMC
  - MPI/GA
  - Intel Offload

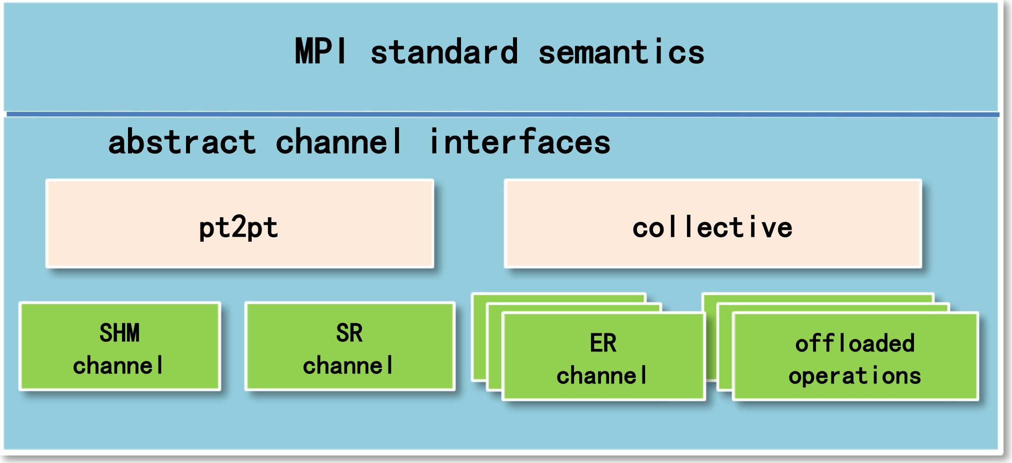


OpenMC

- A directive-based heterogeneous programming model
  - a substitution for existing intra-node OpenMP+X model
  - higher abstraction level than CUDA/OpenCL
- New abstraction for hardware and software
  - provides a unified logical layer above all computing cores, including CPU cores and MIC cores
  - all computation tasks are inherently asynchronous
  - can better orchestrate multiple tasks across multiple devices than OpenACC and Offload on TianHe-2 system

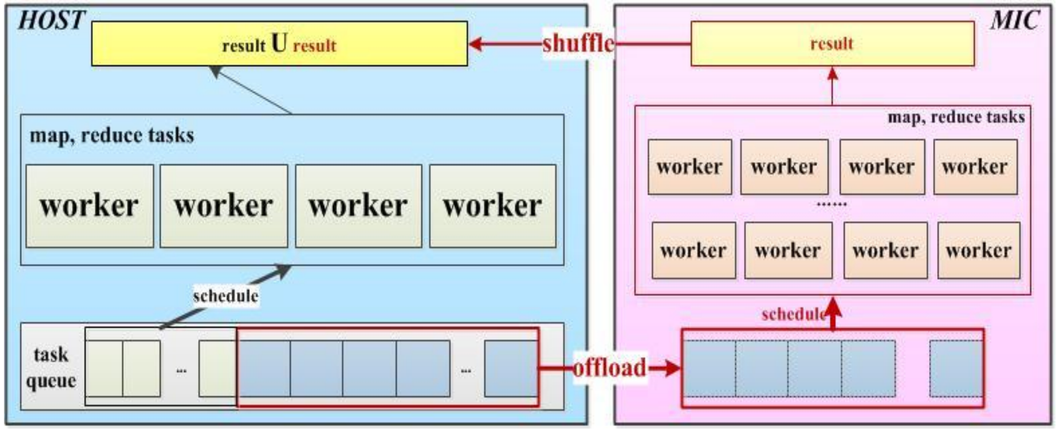
Customized MPI

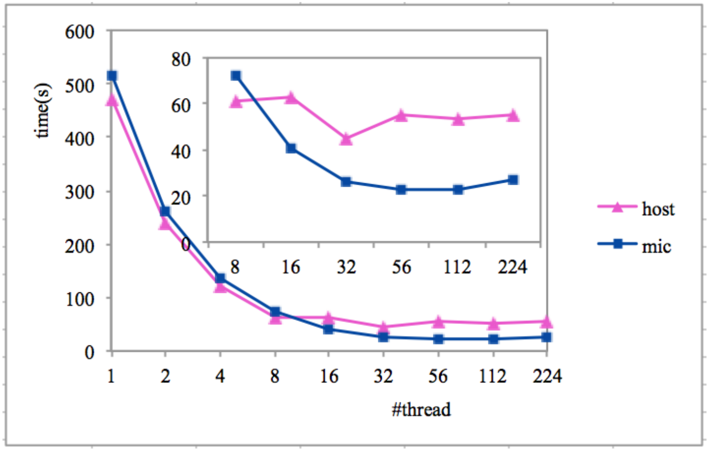
- MPI 3.0 standard compliance
- high-performance RDMA data transferring protocol
- scalability-oriented optimization
  - ◆multi-channel message data transferring
  - ◆dynamic flow control communication protocol
  - ◆offloaded collective operations



MicMR

- MicMR
  - Extend Map/Reduce framework on CPU/MIC heterogeneous architecture for big data processing
    - optimizes data transfer scheme between host CPUs and MIC
    - designs an efficient SIMD parallel optimization strategy for big data applications

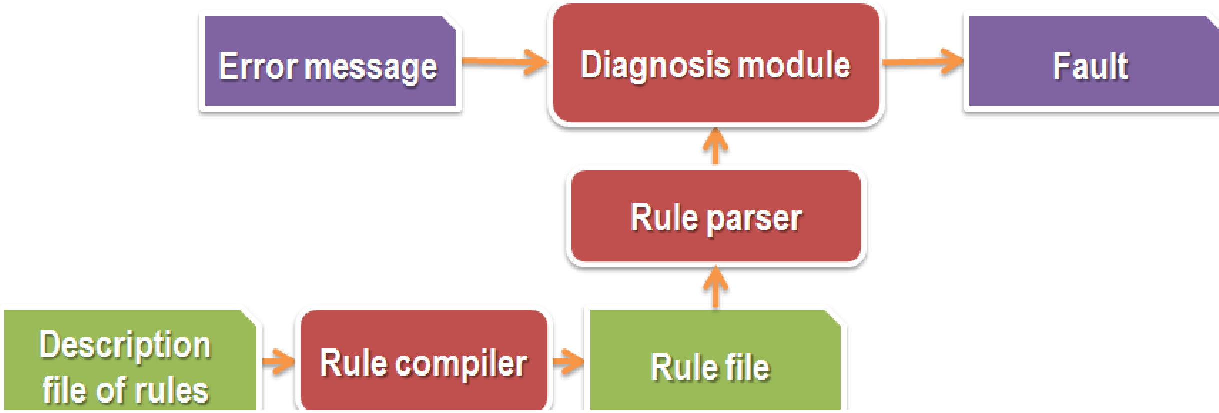




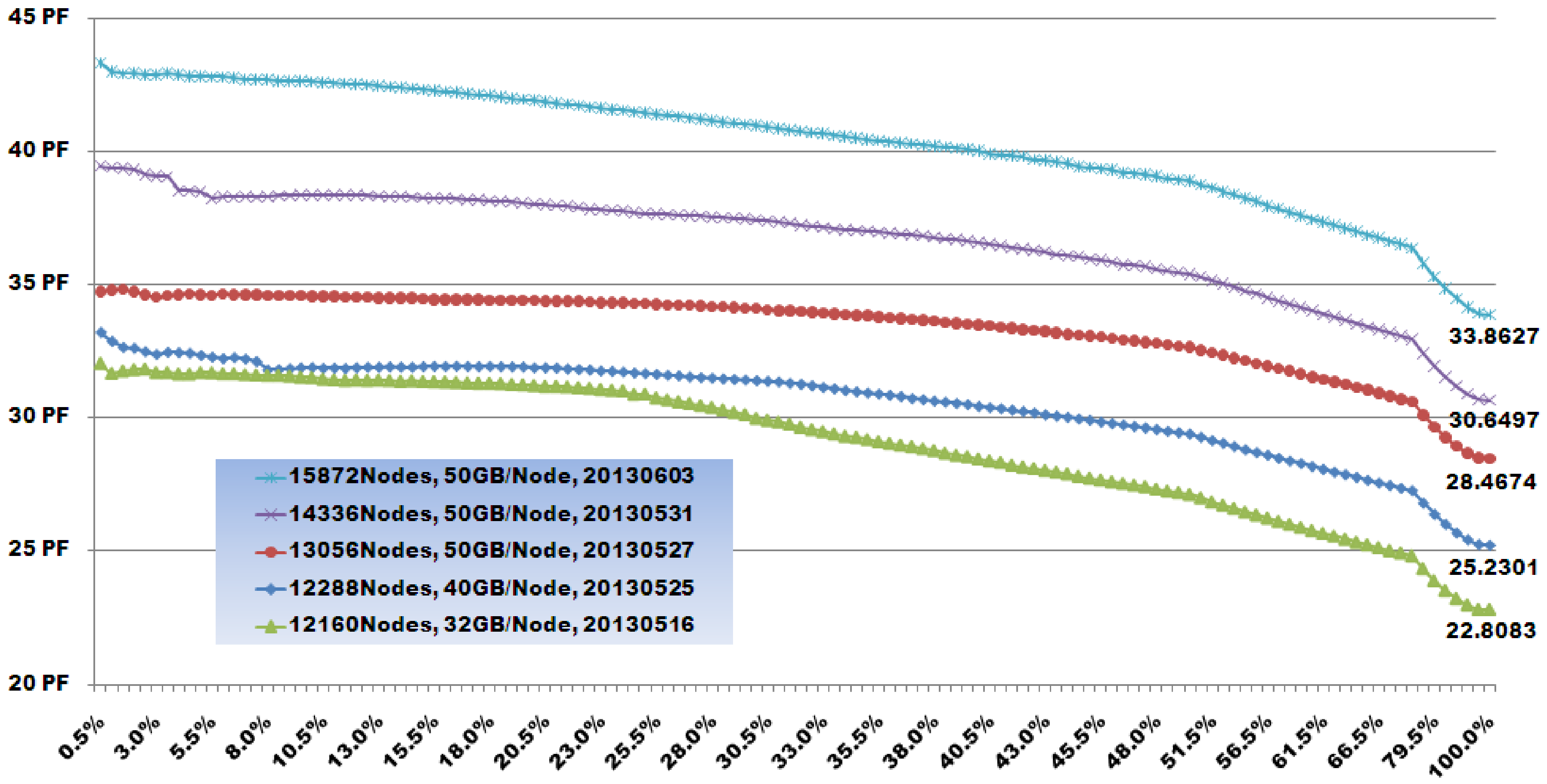
of OK-Means performance ne MIC v.s 2 CPUs with hyper-threads

Autonomic fault Tolerant Management

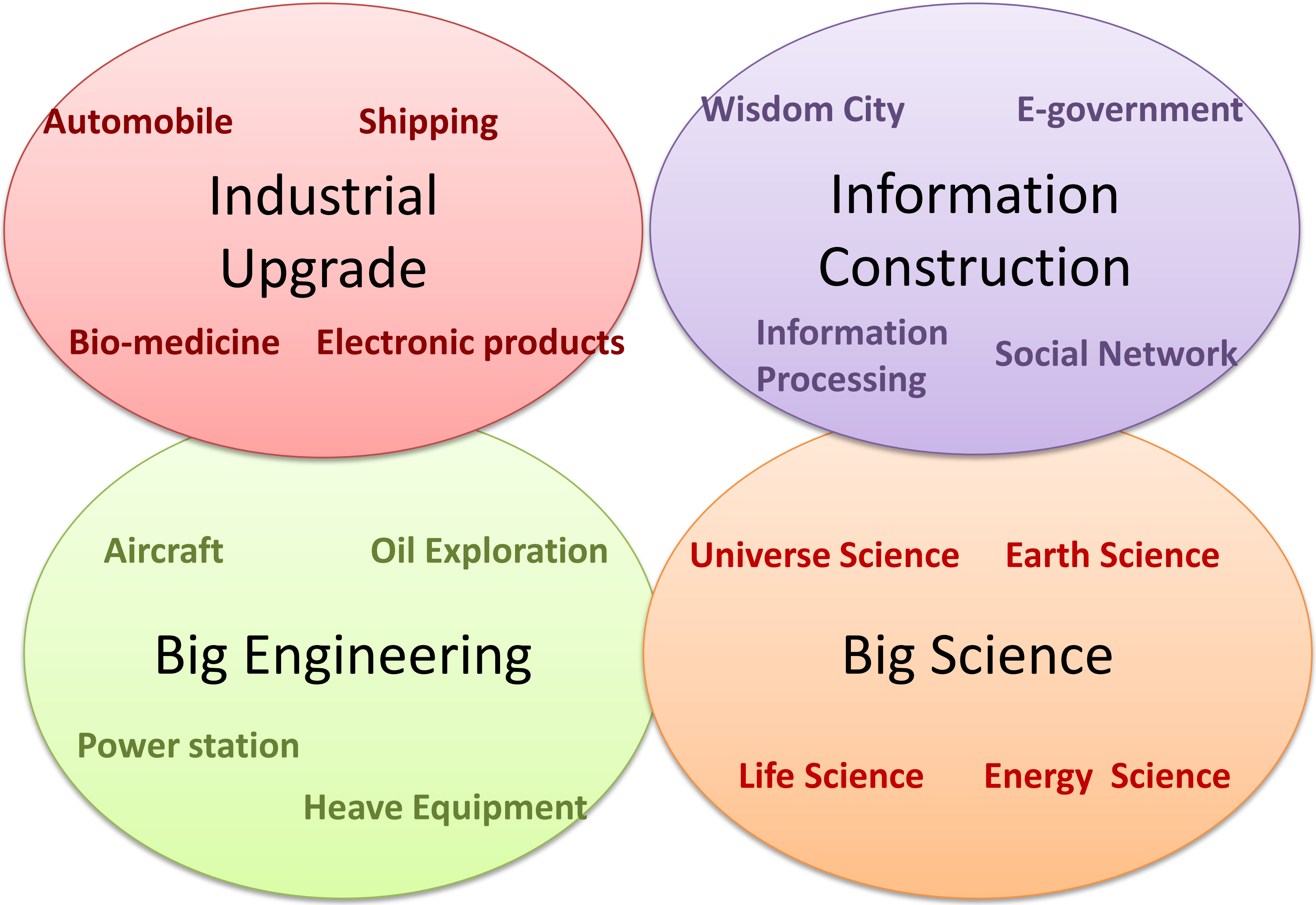
- Fault diagnosis
  - Diagnosis and Rules are independent
  - Diagnosis module supports 1:1 & N:M diagnose functions
  - Rules are added by system manager with time going
  - Analyze infected jobs and parts
- Predict the fault and monitor the healthy of the system
- Lightweight probing



HPL Testing and Tuning



Applications



Automobile

Shipping

Industrial Upgrade

Bio-medicine

Electronic products

Wisdom City

E-government

Information Construction

Information Processing

Social Network

Aircraft

Oil Exploration

Big Engineering

Power station

Heave Equipment

Universe Science

Earth Science

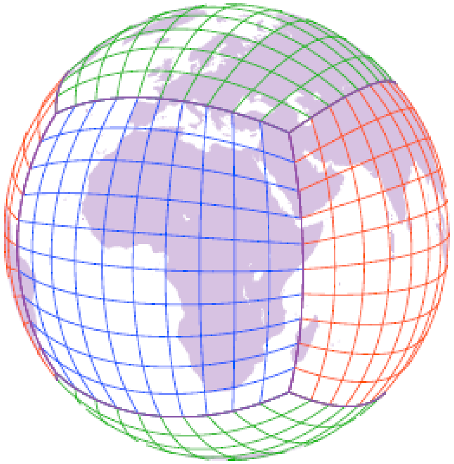
Big Science

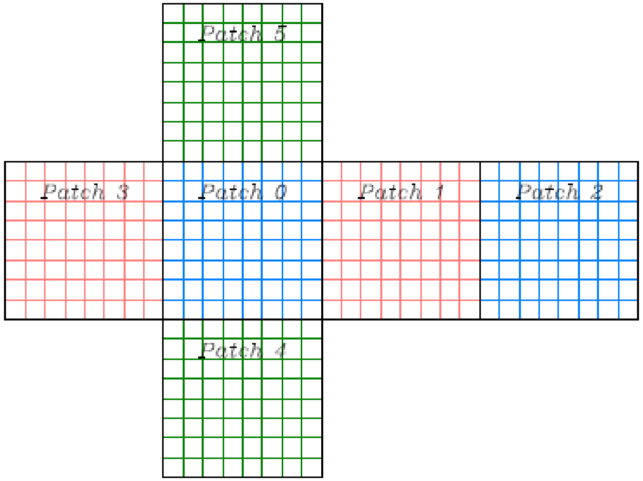
Life Science

Energy Science

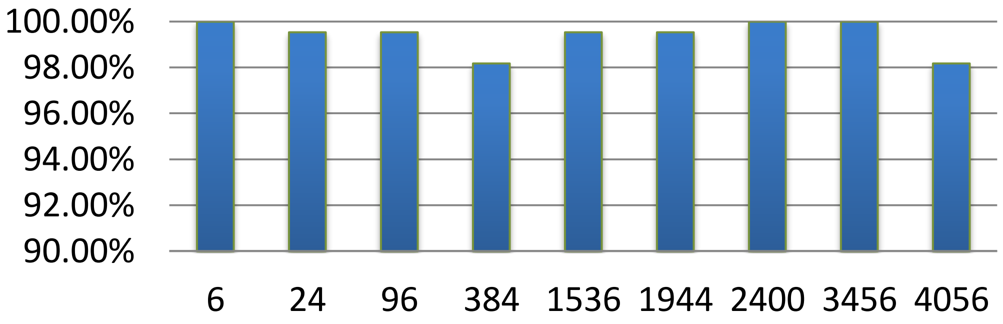
Application

- **Application of a global shallow water model: algorithms**
  - **Hierarchical data partition & communication on cubed-sphere**
  - **Balanced partition between CPU/MIC inside each node**
  - **Communication hiding algorithm based on "Pipe-flow" scheme**





- **Nearly ideal weak scaling on the Tianhe-2**
  - **Using up to 4,056 nodes (97,344 CPU cores + 693,576 MIC cores)**
  - **# of unknowns for the largest run: 200 billion**



Slides provided by Chao Yang from Institute of Software, Chinese Academy of Sciences

国防科学技术大学
National University of Defense Technology

Summary

- Challenge issues
  - New Parallel Model & Algorithm
    - Scalable
    - Power aware
    - Resilience
  - Domain-specific Application Framework
- Broad International Collaboration

……

Thanks