中山大學
SUN YAT-SEN UNIVERSITY

国家超级计算广州中心
NATIONAL SUPERCOMPUTER CENTER IN GUANGZHOU

Towards Next Generation Chinese Supercomputing

Yutong Lu

Prof. School of computer science, SYSU & NUDT
Director, National Supercomputer Center in Guangzhou, China
yutong.lu@nscc-gz.cn

OUTLINE

国家超级计算广州中心

NATIONAL SUPERCOMPUTER CENTER
IN GUANGZHOU

Status of Chinese SC

Supercomputing Centers in China

CNGrid

NSCC-Guangzhou,2013
Tianhe-2

NSCC-Wuxi,2016
Shenwei-Taihu Light

NSCC-Changsha,2012
Tianhe-1A

NSCC-Jinan,2012
Shenwei-Bluelight

NSCC-Tianjin,2010
Tianhe-1A

NSCC-Shenzhen,2011
Dawning-6000

Status of Chinese SC

- **Increased and pervasive use of HPC capability**
  - Performance from 400GF to 125PF
  - Accelerators such as GPU or MIC， etc…
  - Easy use of HPC systems
- **Increased performance and coverage of networks**
  - From 1Gbps to 100Gbps
  - Network penetration increased dramatically
- **Advance in data storage and analysis technology**
  - A dramatic increase in data (scientific, social and application data)
  - Storage from TB to EB
  - Advance in hardware and software for data analytics
- **Advance in AI algorithms and technology**
  - Erupts of deep learning
  - Fast development of BigData-based AI algorithms
  - Hardware and software platforms supporting AI applications

Status of Chinese SC

Efforts of national R&D programs

	863 program	973 program	Key R&D program	R&D Mega-Program	NSFC
HPC	3 key projects	3 projects	1 Key project		1 key initiative
Cloud	1 key project		1 key project		
BD		6 projects	1 key project	Mega-project to be launched	10+ major projects
Net + Comm	3 key projects	8 projects		1 mega-project 15 years	1 key initiative
AI	30-year support	12+ on PR&MM		Mega-project to be launched	

Status of Chinese SC

HPC: MOST 863 Program & Key project

- 2001-2005：High Performance Computer and Core Software
- 2006-2010：High Productivity Computer and Grid Service Environment
- 2010-2016：High Productivity Computer and Application Service Environment
- New key project on HPC (2016-2020)
  - Strengthening R&D of kernel technologies and pursuing the leading position in high performance computer development
  - Promoting HPC applications
  - Building up an HPC infrastructure with service features and exploring the path to the HPC service industry
- Major tasks
  - Next generation supercomputer development
  - HPC applications development
  - CNGrid upgrading and transformation

Status of Chinese SC

HPC: NSFC Key Initiative

- **"Basic Algorithms and Computable Modeling for High Performance Scientific Computing"**
  - **8 years, 220 million yuan**
  - **Computable modeling methods**
  - **Innovative basic algorithms**
  - **Domain applications for demonstration**

- **NSFC/Guangdong province joint program on supercomputing**
  - **2years, 60 million**
  - **Supporting Tianhe-2 users in the form of CPU hours**
  - **600+ projects supported**

Status of Chinese SC

Cloud & BD: 863 Program

- **"China cloud (phase I and phase II) "863 key project**
  - **Cloud OS**
  - **Cloud kernel services (search, translation)**
  - **Cloud applications**
- **Cloud computing and Big Data, key project**
  - **Theory and methodology of software-defined cloud computing**
  - **Cloud-oriented network OS**
  - **Novel big data storage technologies and platforms**
  - **Domain-specific big data management system**
  - **Dataflow-based big data analysis system**
  - **Cloud computing and big data infrastructure**
    - New generation cloud servers / Data center key technology and equipment
  - **Cloud-based and data-driven software**
    - Theory, methodology and technology of evolutionary intelligent software
    - Integrated software platform for smart cities
  - **Big data analytics applications and human-like intelligence**
  - **…**

Status of Chinese SC

Cloud & BD: NSFC

- **NSFC/Guangdong province joint project on Big Data Science Center**
  - **5-year program, 300 million**
  - **Supporting about 10 projects for Smart city**

- **Big Data technology and application, Mege-project to be launched**

Status of Chinese SC

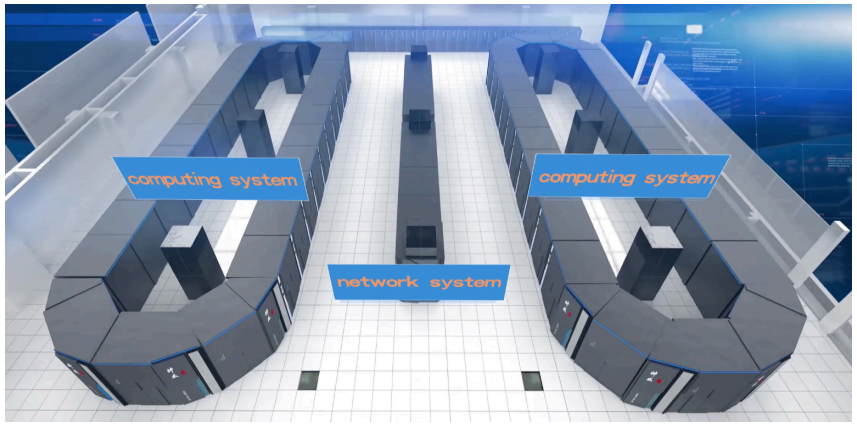
AI2.0—"New generation AI development plan" published by the state council

- **AI is raised to the level of national strategy**
  - Grasp the strategic opportunity of AI development
  - Establish the first-move advantage in AI development
  - Build up an innovative country with world level science and technology
- **Major tasks**
  - Establish an open collaborative system for AI research and innovation
  - Cultivate high-end, high efficient and intelligent economy
  - Construct secure convenient intelligent society
  - Civil-Military Integration in AI
  - Establish pervasive, secure and efficient intelligent infrastructure
  - Conduct new generation AI mega-project

Overview of Taihu-Light

Shenwei Taihu-Light

- SW processor
  - 1.5GHz, 260cores, 3.0TF/node
- 40,000nodes， 100mil cores
- Peak 125PF/s， Linpack 93PF/s
- 6GF/W

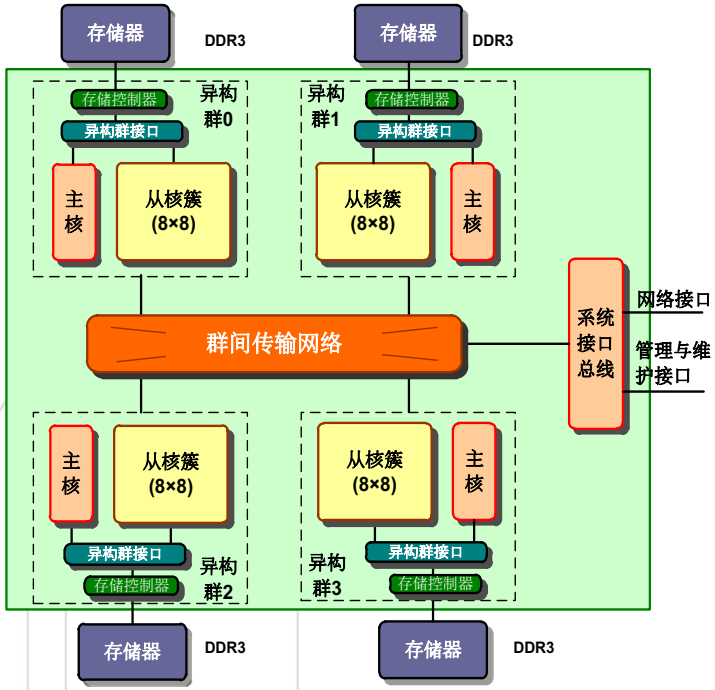


MPE 64-bit RISC core

- support both user and system modes
- 256-bit vector instructions
- 32 KB L1 instruction cache, and 32 KB L1 data cache
- 256 KB L2 cache， 8x8 CPE mesh

CPE 64-bit RISC core

- support only user mode
- 256-bit vector instructions
- 16 KB L1 instruction cache, and a Scratch Pad Memory (SPM)



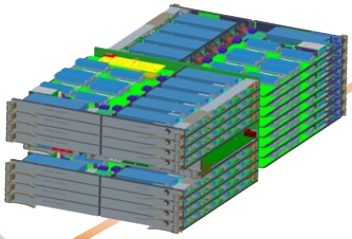
Overview of Tianhe-2

Perf	54.9PFlops / 33.86PFlops
System	16000nodes, 1.4PBmem
cabinet	125+8+13+24=170 (720m^2)
Power	17.8 MW (1902MFlops/W)
Cooling	Closet wind & water

No.1 on Top500 list from 2013 to 2015

TH-2 系统

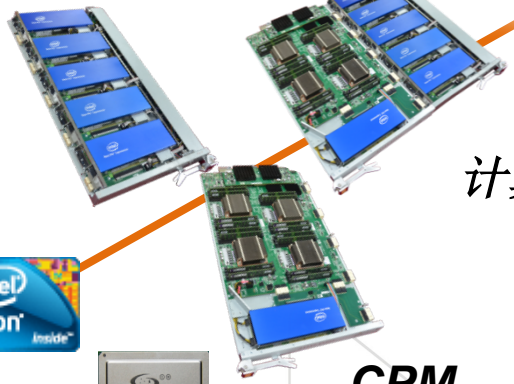
机柜
(8 x 机框)



机框
(16 x cn)

TH-Net

APM



计算主板

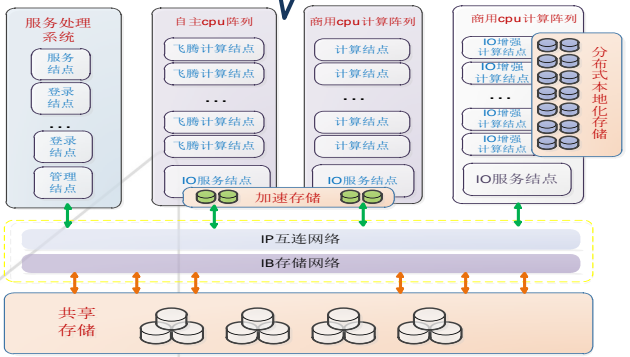
Phi
#48000

IVB
#32000

FT-1500
#4096

CPM

ION

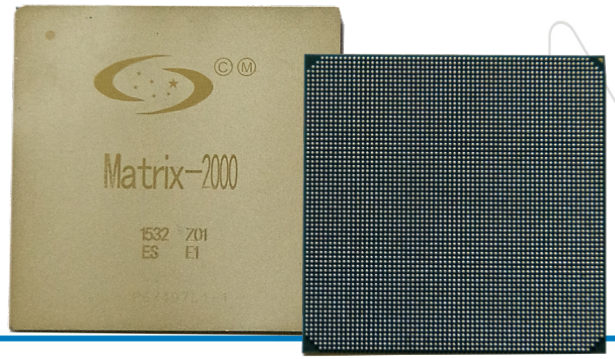


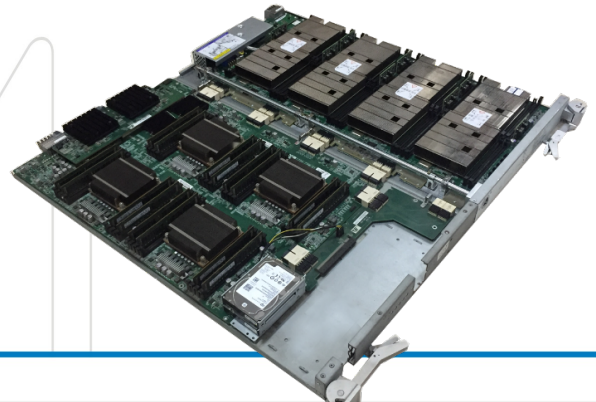
Large Scale Hybrid Storage System
12.4PB

Overview of Tianhe-2A

Tianhe-2A Supercomputer

- Designed **Proprietary accelerator Matrix-2000** to replace Intel KNC
- Developed **accelerator blade** with 4 Matrix-2000s
- **Customized software stack** for Matrix-2000
- Upgraded the proprietary interconnection chipset and network from 10G to **14G**
- System memory upgraded to 3.4PB from 1.4PB
- Expanded the I/O storage subsystem from 12.4PB to 20PB





Overview of Tianhe-2A

Comparison

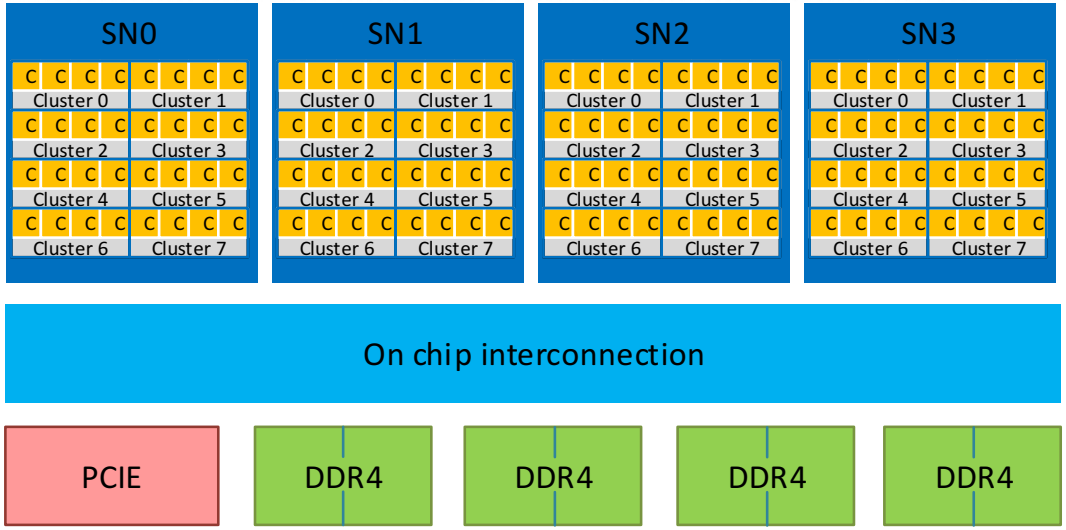
Items	Milkyway-2	**Milkyway-2A**
Nodes & Performance	16000 nodes with Intel CPU + KNC	**17792 nodes with Intel CPU + Matrix-2000**
	54.9Pflops	**94.97Pflops**
Interconnection	10Gbps, 1.57us	**14Gbps, 1us**
Memory	1.4PB	**3.4PB**
Storage	12.4PB, 512GB/s	**20PB, 1TB/s**
Energy Efficiency	17.8MW, 1.9Gflops/W	**About 18MW, >5Gflops/W**
Heterogeneous software	MPSS for Intel KNC	**OpenMP/OpenCL for Matrix-2000**

Matrix-2000 accelerator

Chip specification

- 128cores
  - 4 super-nodes (SN)
  - 8 clusters per SN
  - 4 cores per cluster
  - Core
    - Self-defined 256-bit vector ISA
    - 16 DP flops/cycle per core
- Peak performance: 2.4576Tflops@1.2GHz

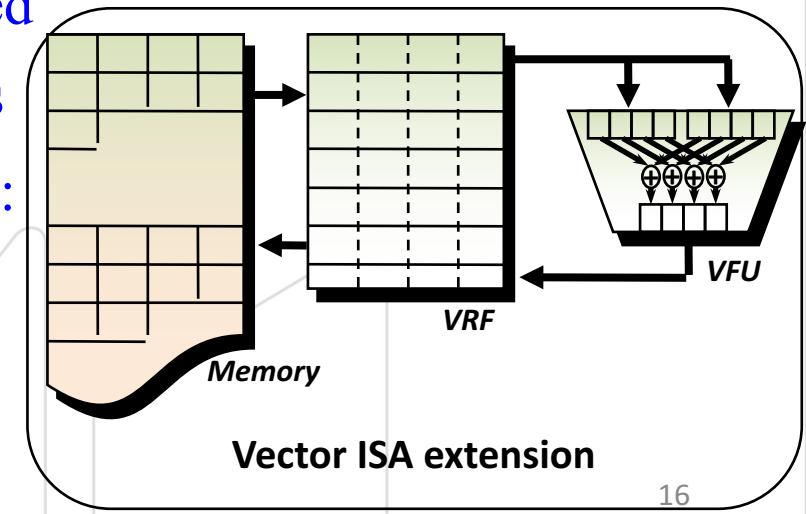
  4 SNs x 8 clusters x 4cores x 16 flops x 1.2 GHz = 2.4576 Tflops
- Peak power dissipation: ~240w
- Interface
  - 8 DDR4-2400 channels
  - X16 PCIE 3.0 EP Port



Matrix-2000 Vector ISA Extension

Self-defined 256-bit vector ISA extension

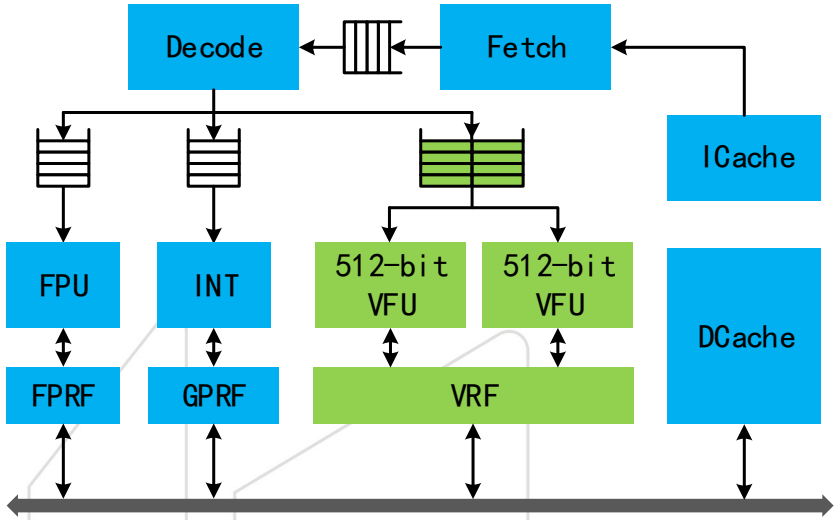
- General vector instructions
  - Integer/floating-point arithmetic, compare, bit manipulation, conversion, permutation and shuffling instructions, memory access operations
- Special features
  - Guarded(predicated) execution supported
  - Reduction-type execution supported
  - Pattern-aware prefetch instructions
  - 64-entry vector register file (VRF): for efficient and flexible programming



Matrix-2000 Core Structure

Core Structure

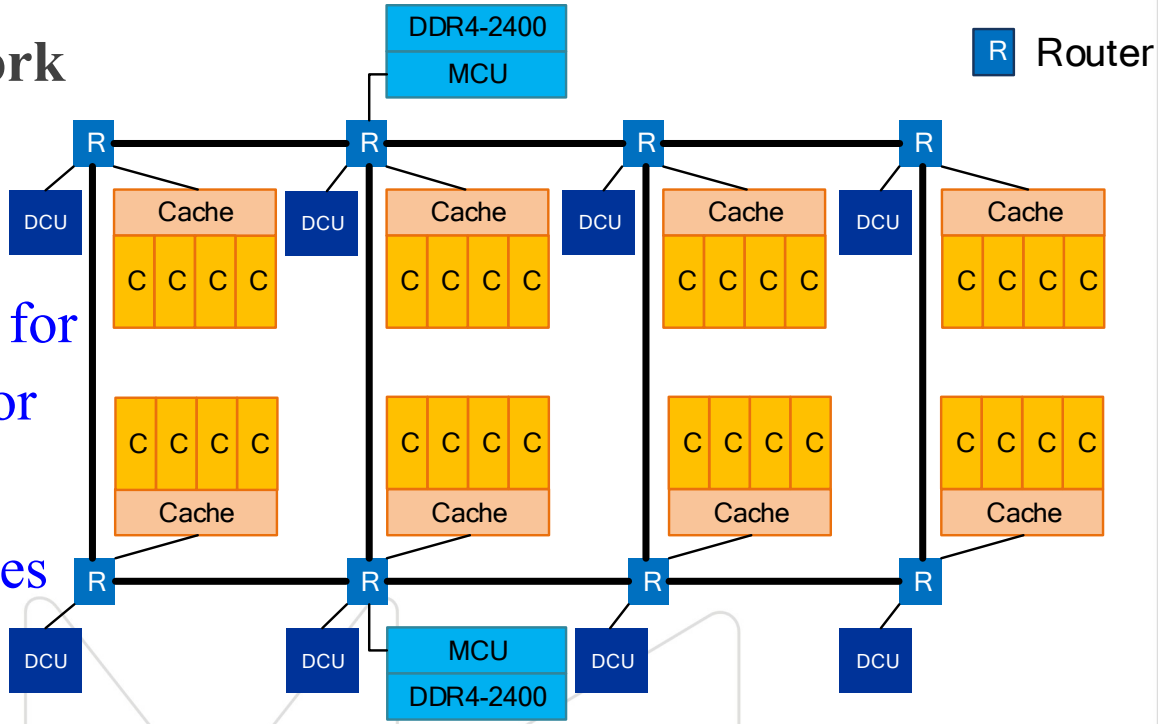
- In-order pipeline
  - 8~12 pipeline stages
- Two 256-bit vector functional units (VFU)
  - SP/DP MAC, multiply and add
  - Issue 2 vector instructions per cycle
  - 16 DP flops/cycle per core
- Memory access operations
  - 512-bit data access per cycle
  - Software guided prefetch



Matrix-2000 Interconnect in SN

Interconnect in SN

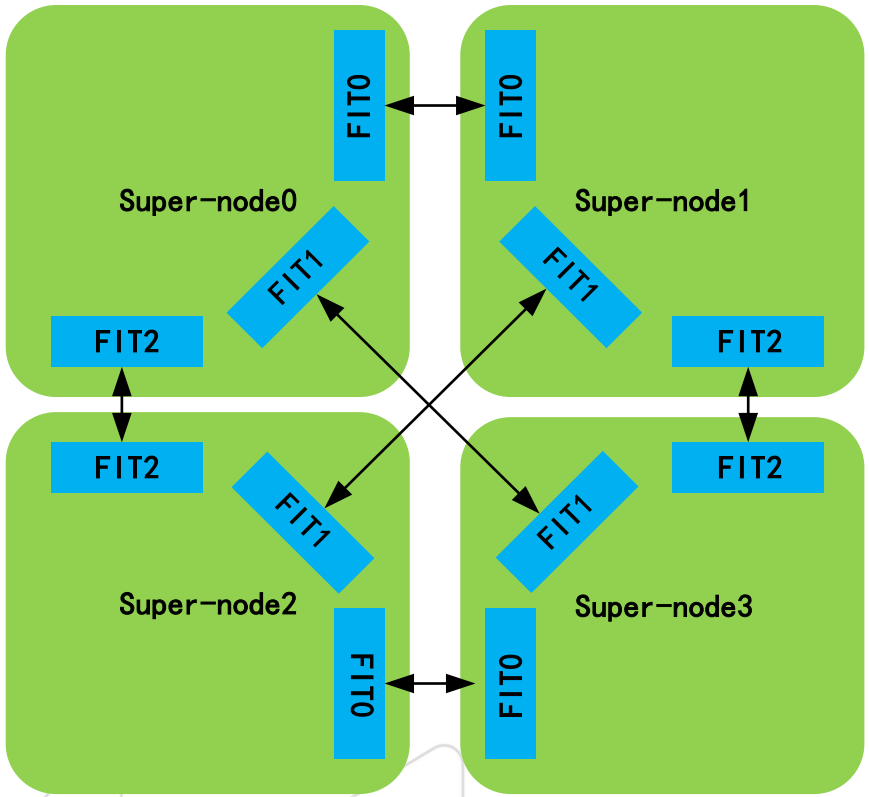
- Directory based cache coherence (CC) is supported by DCU
- 4x2 mesh based network
  - Router with 6 bidirectional ports
  - 4 physical channels for CC and 1 channel for debug
  - Low latency: 3 cycles for each hop
  - High bandwidth: 256GB/s per port



Matrix-2000 Interconnect in SN

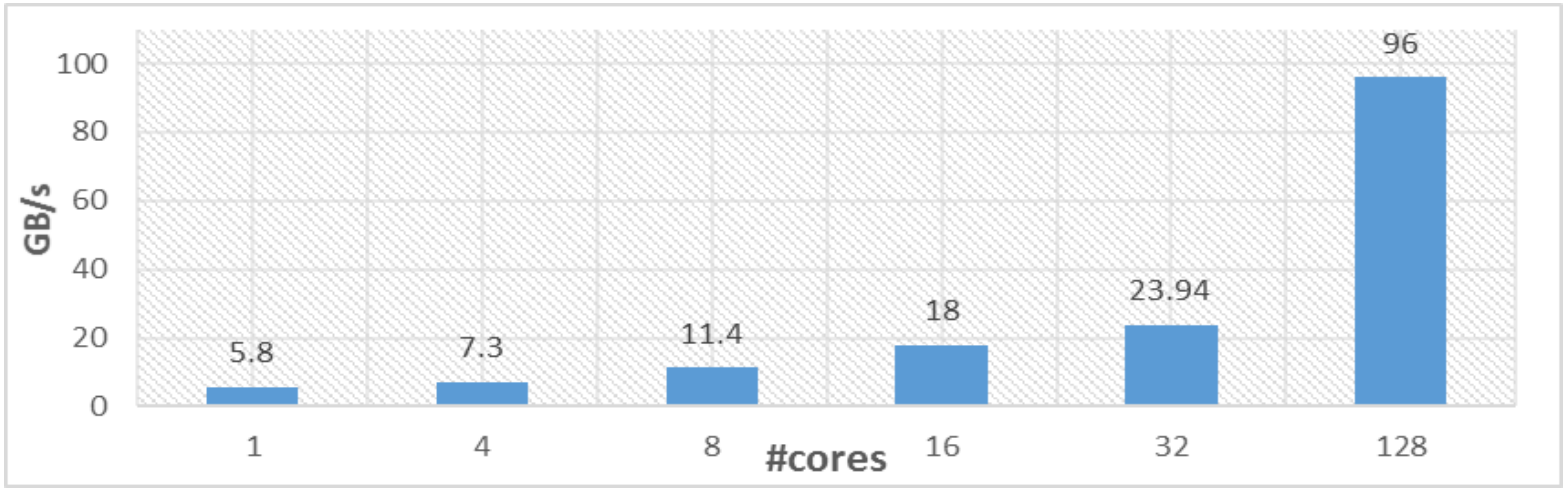
Interconnect between SNs

- Two SNs are connected to each other by Fast Interconnect Transports (FITs)
- DMA is supported for more efficient bandwidth usage
- Maximum bi-directional bandwidth of a FIT is 25.6GB/s
- Round-trip latency is about 20ns
- CRC data check and retransfer mechanism are supported to ensure the correctness of data

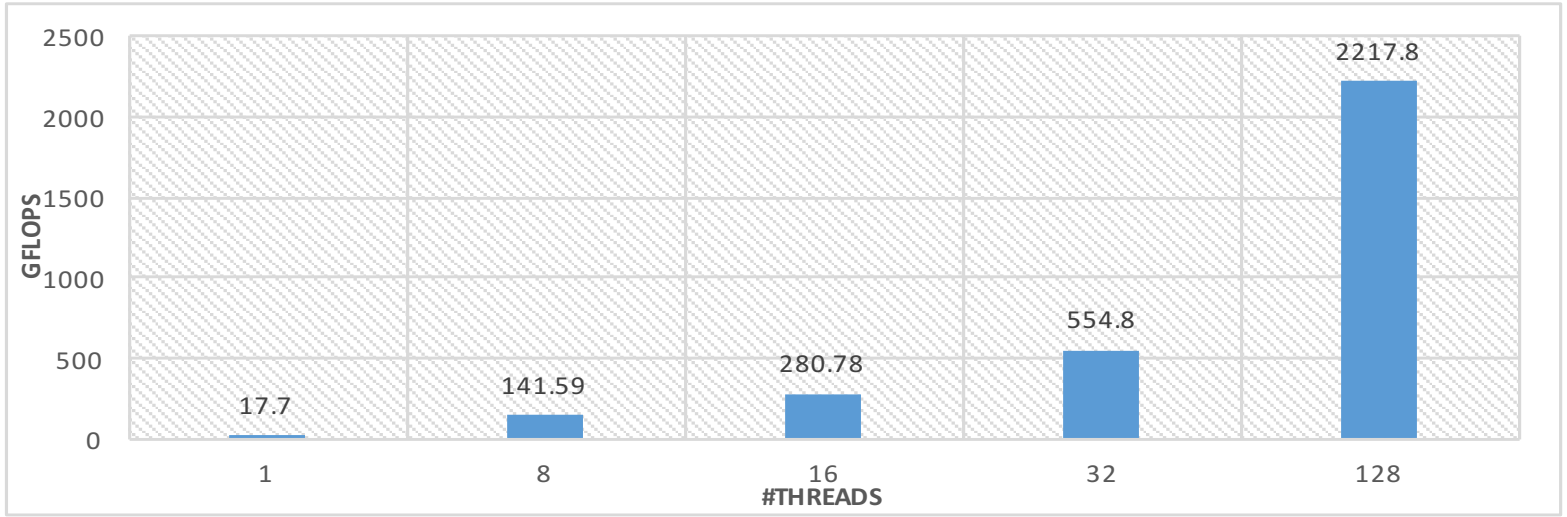


Matrix-2000 Evaluation

- Stream: 96GB/s with 128 cores (62.5%)



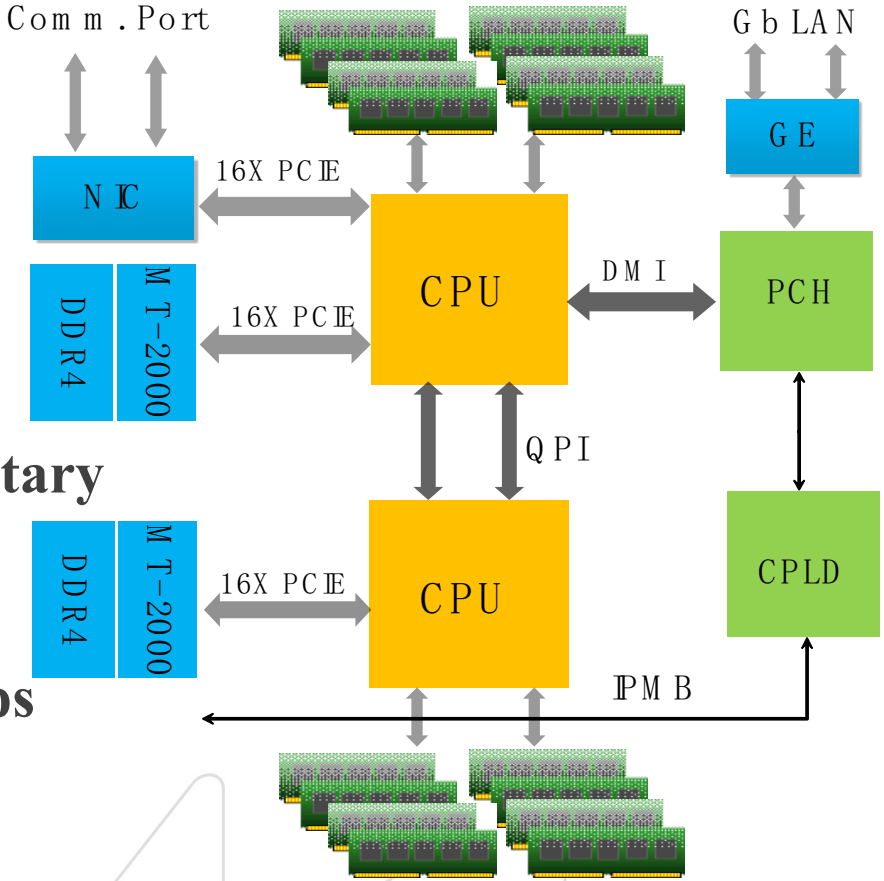
- DGEMM: 2.2Tflops with 128 threads(90.2%)



Compute nodes

Heterogeneous Compute Nodes

- Intel Xeon CPU x2
- Matrix-2000 x2
- Memory:192GB
- Interconnection:14G proprietary network
- Peak performance: 5.34Tflops

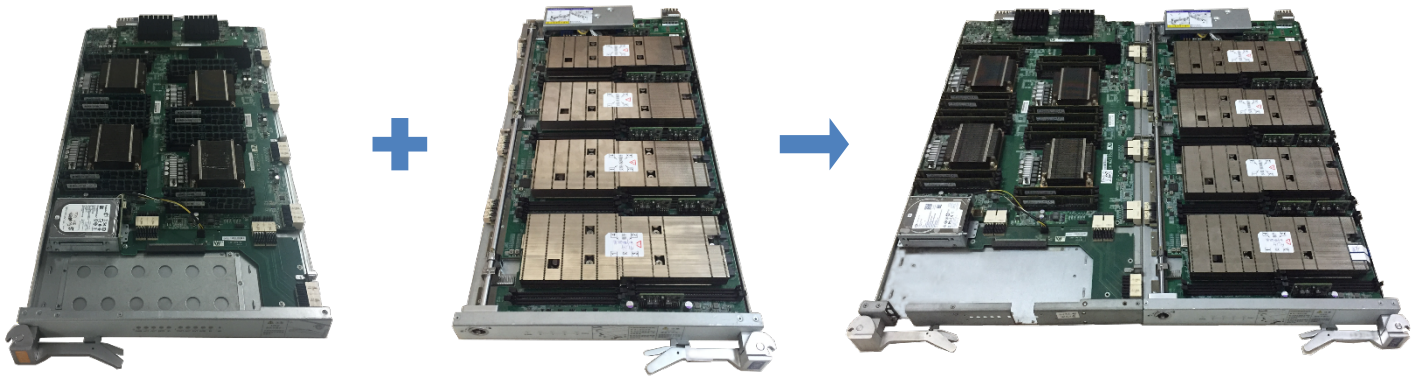


Compute nodes

Heterogeneous Compute Blades

- Compute blade = Xeon part + Matrix-2000 part

4 Intel Xeon CPUs **4 FT Matrix-2000** **2 Compute Nodes**



- Use the Matrix-2000 part to replace the KNC part

Interconnection network

High radix router ASIC: NRC

- Tile based switch architecture
- 24 network ports
- 8 lanes 14G SerDes per port
- Bidirectional bandwidth: 224Gbps
- Throughput of single NRC: 5.376Tbps

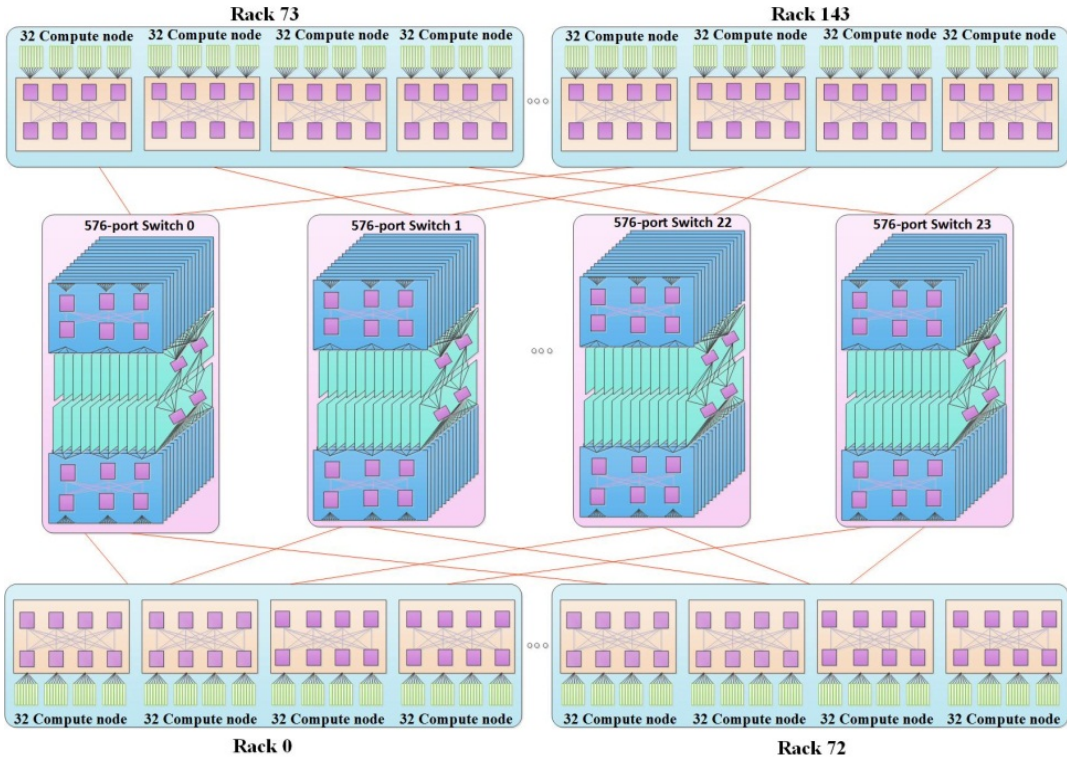
Network interface ASIC: NIC

- Proprietary MP/BLT communication and collective offload mechanisms
- 16x PCI-E Gen 3 interface
- 1 port with 8 lanes 14G SerDes
- Link rate:14Gbps

Interconnection network

Multi-level fat-tree topology

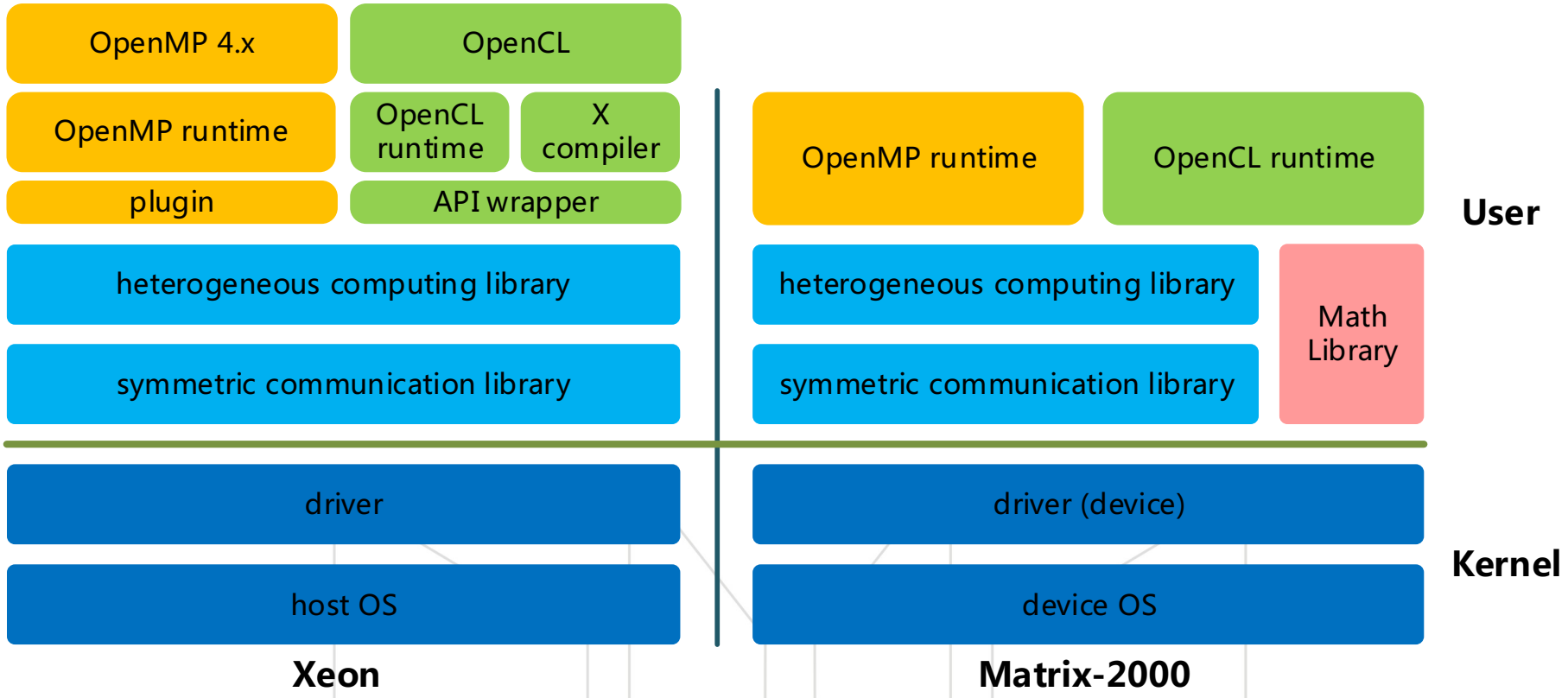
- 168 compute racks connected by 13 576-port switches
- Network diameter: 9
- Bi-section BW: 161TB/s
- Optical-electronic hybrid transport tech
- Proprietary network protocol



HPC Software Stack

Heterogeneous programming environment

- Supporting OpenMP 4.x and OpenCL for heterogeneous computing



HPC Software Stack

Heterogeneous programming environment

- Device operating system
  - A light-weight Linux-based OS embedded with the accelerator device driver
  - Providing a high-efficiency resource pool for thread scheduling
- Device driver & symmetric communication library
  - Connection: establishing a socket-like connections between processes on different nodes in the network
  - Messaging: exchange of short, latency-sensitive messages such as commands and synchronization operations
  - RMA: one-sided communication for the transfer of large , bandwidth-sensitive buffers

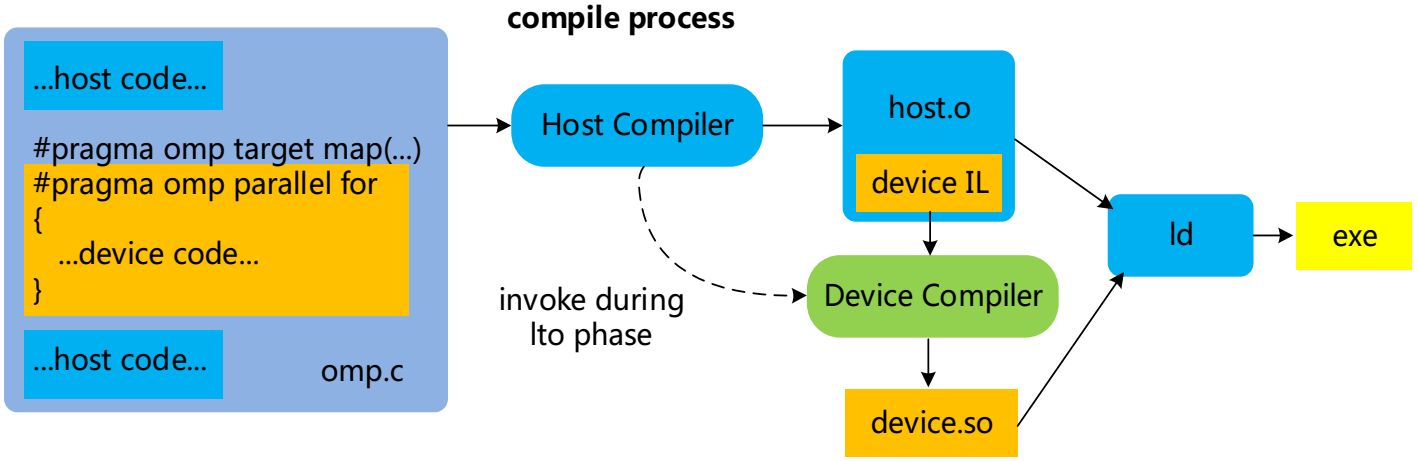
Heterogeneous programming environment

- Heterogeneous computing library
  - **Acting similar as COI in MPSS**
  - **Provides heterogeneous computing APIs such as**
    - **device manage: acquire device info, get device handle, clean up device resources,…**
    - **data manage: allocate/destroy buffers on device, map device buffers with local pointer,…**
    - **task manage: create/destroy process, run functions on device,…**
  - **Serves the runtime of high-level API implementations, e.g. OpenMP and OpenCL**

HPC Software Stack

Heterogeneous programming environment

- OpenMP
  - **OpenMP 4.x compiler based on GCC 7.1**
  - **Differ from implementation for MIC(KNL) since Xeon and Matrix-2000 are not binary-compatible**

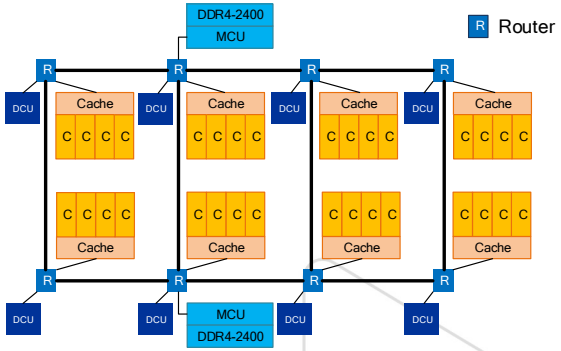


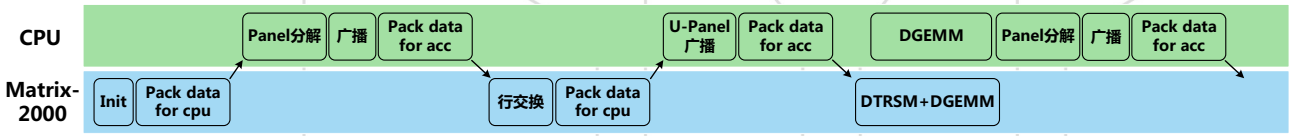
- OpenCL
  - **A thin wrapper to bridge the OpenCL runtime and the heterogeneous computing library**
  - **A cross compiler based on LLVM 3.9 for dynamic kernel compilation**

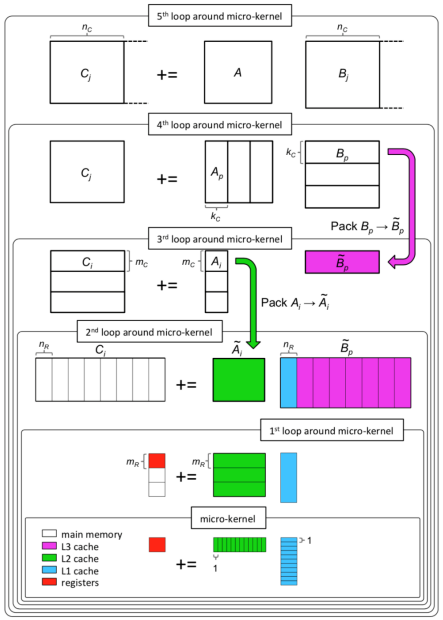
HPL testing and tuning

Implemented and tuned based on TH-2's version

- the whole array resides on accelerator's memory
- offloading trailing submatrix update (including pivoting)
- 17792 nodes, peak performance 94.97Pflops
- HPL performance 61.4434Pflops，efficiency **64.7%**
- N=9773000, P=256, Q=556, Time=10128.11s







Applications on Tianhe-2

- **Number of Users**
  - 2600+ scientific research institutions, universities and key enterprises
- **Number of Projects**
  - 800+ national projects and 100+ projects of Guangdong/Guangzhou
- **Classic Applications**
  - Big Computing: large scale of scientific and engineering applications
  - Big Data: Hadoop/Spark/Storm, peak IO 1TB/s of genetic data analysis
  - Transaction Processing: Guangdong Province Security e-Government Cloud

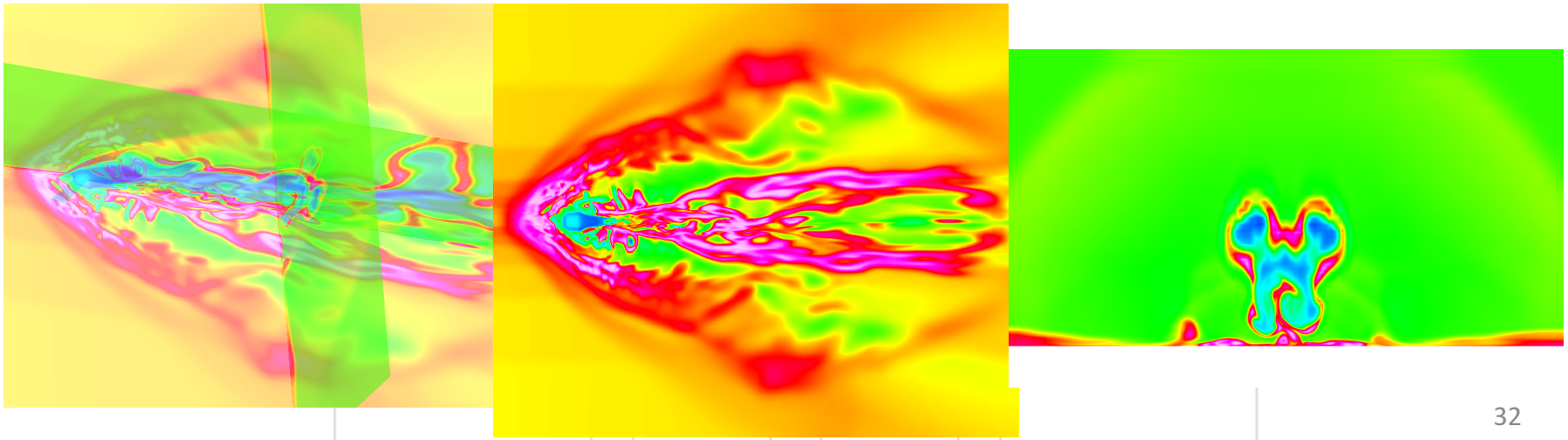
Applications on Tianhe-2A

Case study of Applications over million cores

- Supersonic Turbulent combustion simulation
- Neutron transport OpenSN
- Chinese domestic CFD software: PHengLEI
- Multi –depot Vehicle Routing Problem
- Electromagnetic scattering simulation
- High precision CFD simulation
- ……

Applications on Tianhe-2A

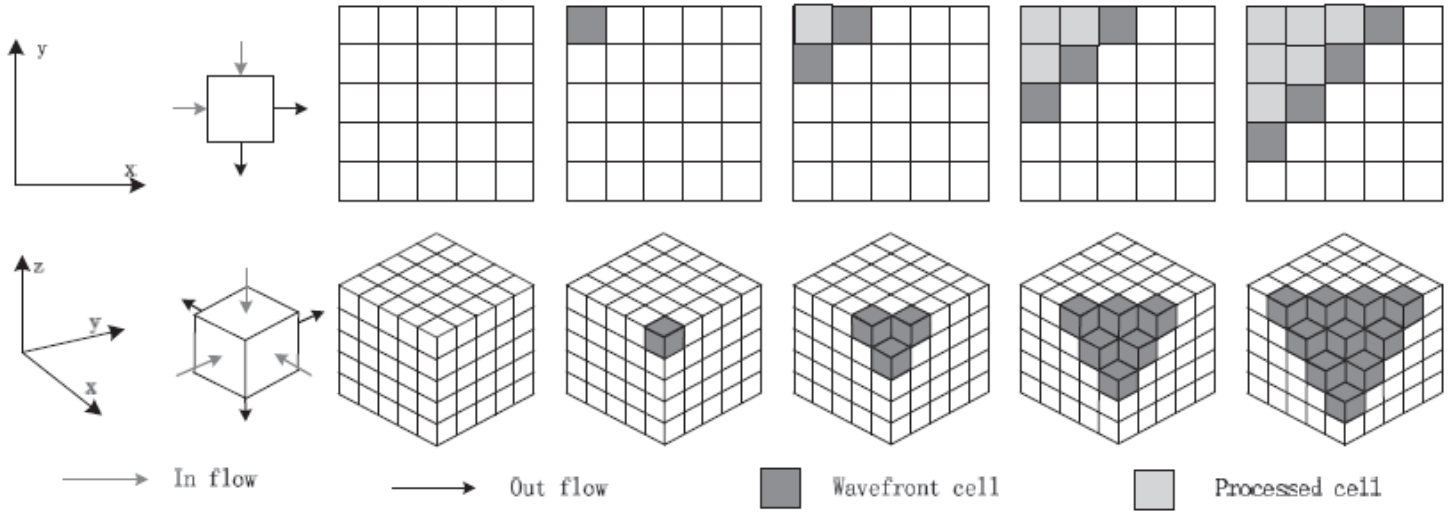
- **Heterogeneous Supersonic Turbulent combustion simulation for scramjet engine**
  - MPI+OpenMP4.5+SCIF
  - Optimization methods: single thread, OpenMP, heterogeneous, grid re-partition
  - The number of grid cells is 15.636 billions, 2.4M
  - CPU+Matrix-2000 is 5.28~8.92 times faster than CPU_Baseline
  - Scaling to 1,036,000 cores: 45.33%



Applications on Tianhe-2A

Neutron transport OpenSN

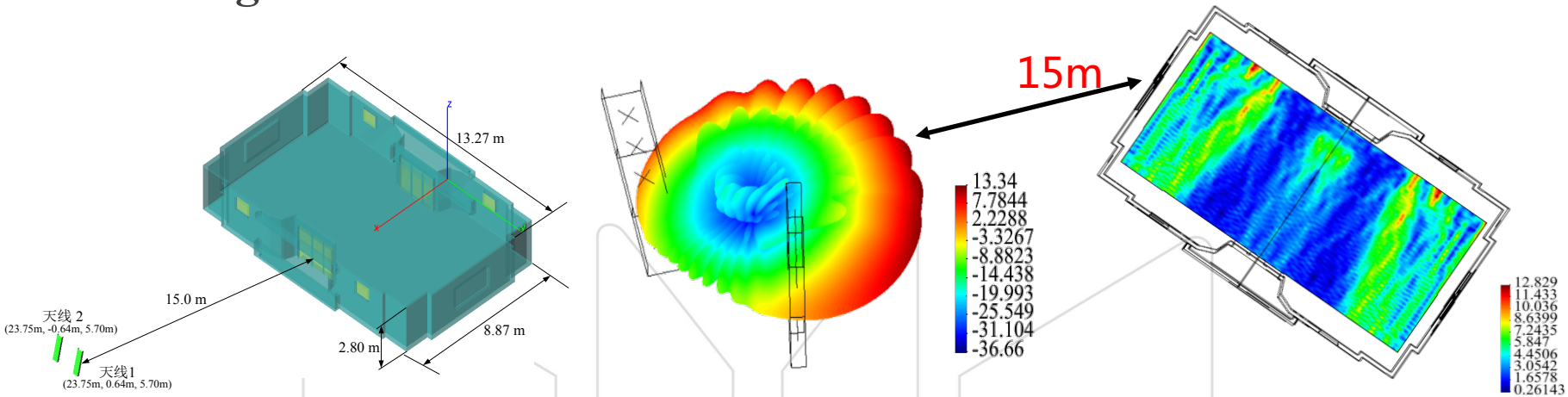
- Based on Sweep3D
- MPI+OpenMP+SCIF+COI
- CPU+Matrix-2000 is 6.42~10.32 times faster than CPU_Baseline
- Scaling to 1013760 cores: 52.03%



Applications on Tianhe-2A

Electromagnetic scattering simulation

- Large-Scale Parallel Method of Moments
- Full-Wave Analysis of Indoor Electromagnetic Pollution from Base-Station Antennas
- CPU+Matrix-2000 is 2.35~4.50 times faster than CPU_Baseline
- Scaling to 1152000 cores: 32%



Tianhe next

PreExa-Tianhe system

- FT2000+ CPU plus Chinese Accelerator ？
- ArmV8, 64bit

16nm	2.2-2.4GHz	615Gflops	64c	100W
工艺	主频	峰值浮点性能	核数	功耗

- 8*DDR4 2400
- 2 x16、1 x1 PCIE3.0

Tianhe next

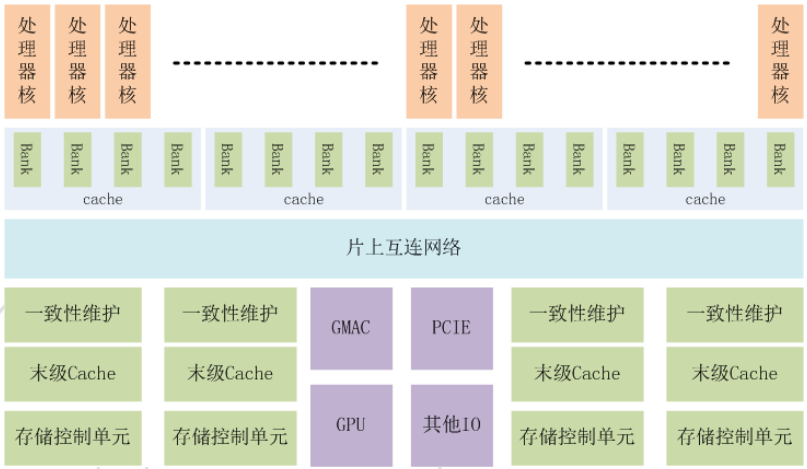
FT2000+

	SPEC 2006 Int	SPEC 2006 FP	Power
Intel X5670(6c, 2.93G,2010)	171	121	95W
Intel E5-2692v2 (12c , 2.2G , 2013)	456	329	115W
Intel E5-2695v3 (14c , 2.3G , 2014)	557	410	120W
Intel E5-2699v3 (18c , 2.3G , 2015)	693	460	145W
Intel E5-2699v4 (22c , 2.2G , 2016)	837	533	145W
Intel E7-8890v4 (24c , 2.2G , 2017)	880	590	165W
FT - 2000+ , 64 core 2.3GHz (2017)	**453**	**433**	**96W**

Tianhe next

FT2000+

- High efficient out-of-order superscalar processor core
- Hierarchical parallel storage structure on chip
- Multilevel heterogeneous network on chip
- Data affinity multi-core processor architecture
- Polymorphic parallel verification for full cycle verification management process
- High performance low power and high density CPU physical design
- High availability processor design technology
- Design technology of large size and high power packaging



Tianhe next

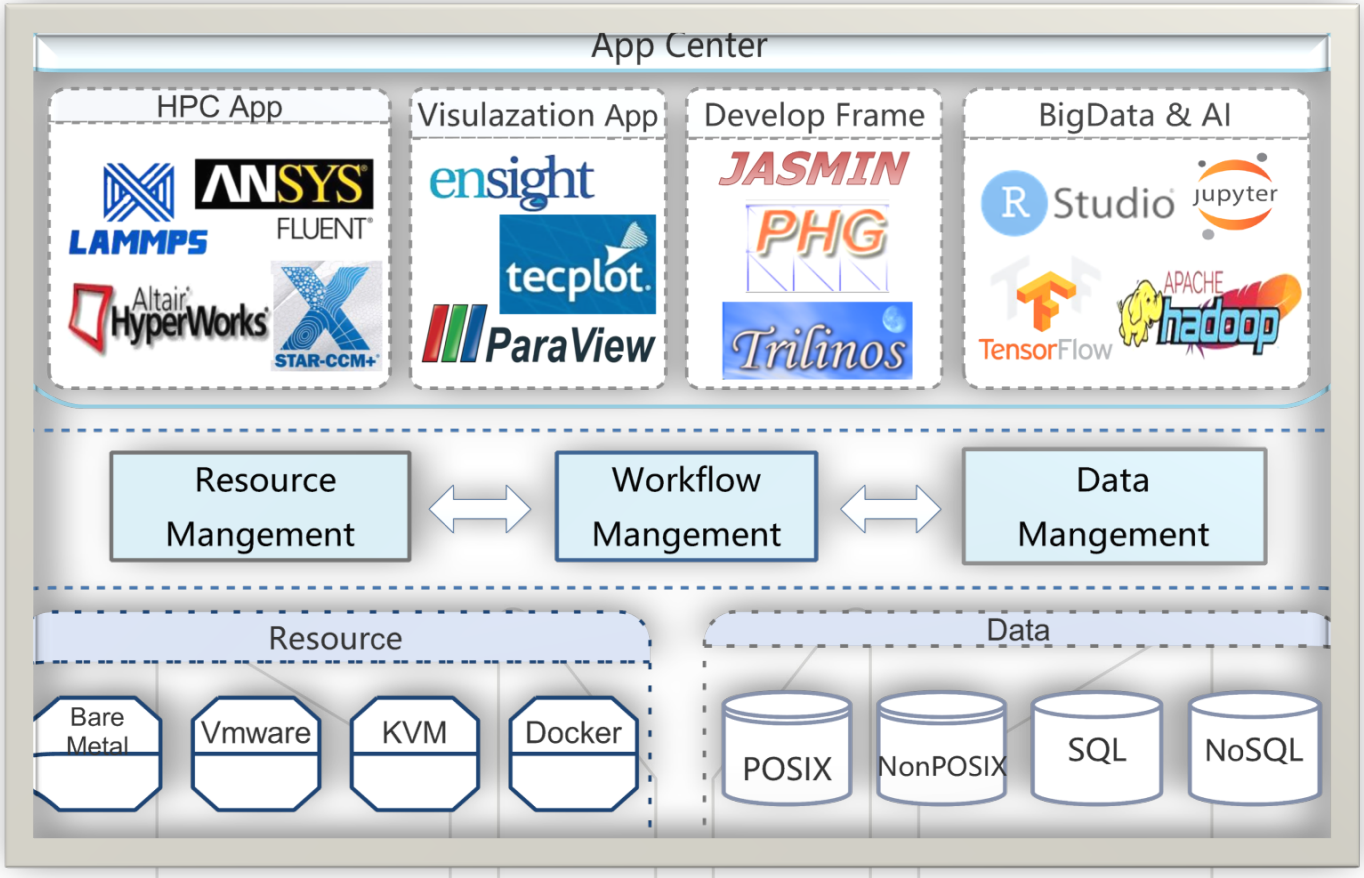
Leo System

- Prototype of PreExa Tianhe
- 17280 FT2000+
- 25Gbps interconnection
- >10PFlops

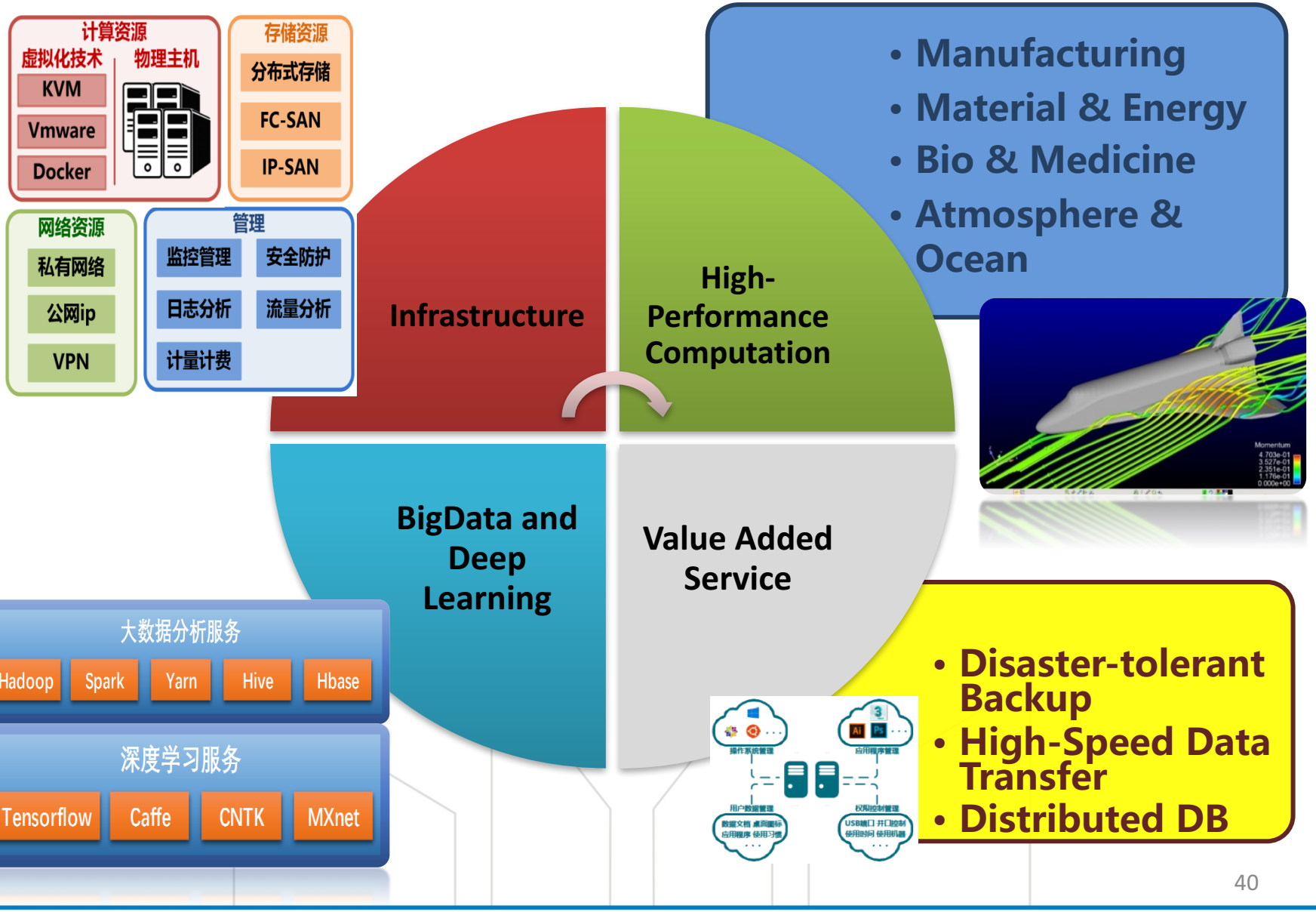
Eco-system

Tianhe-Starlight Platform

- Convergence of HPC，Cloud Computing，Bigdata
- Application-Centric Co-design Environment
- Support Cross-Software，Cross-domain Coupling Workflow

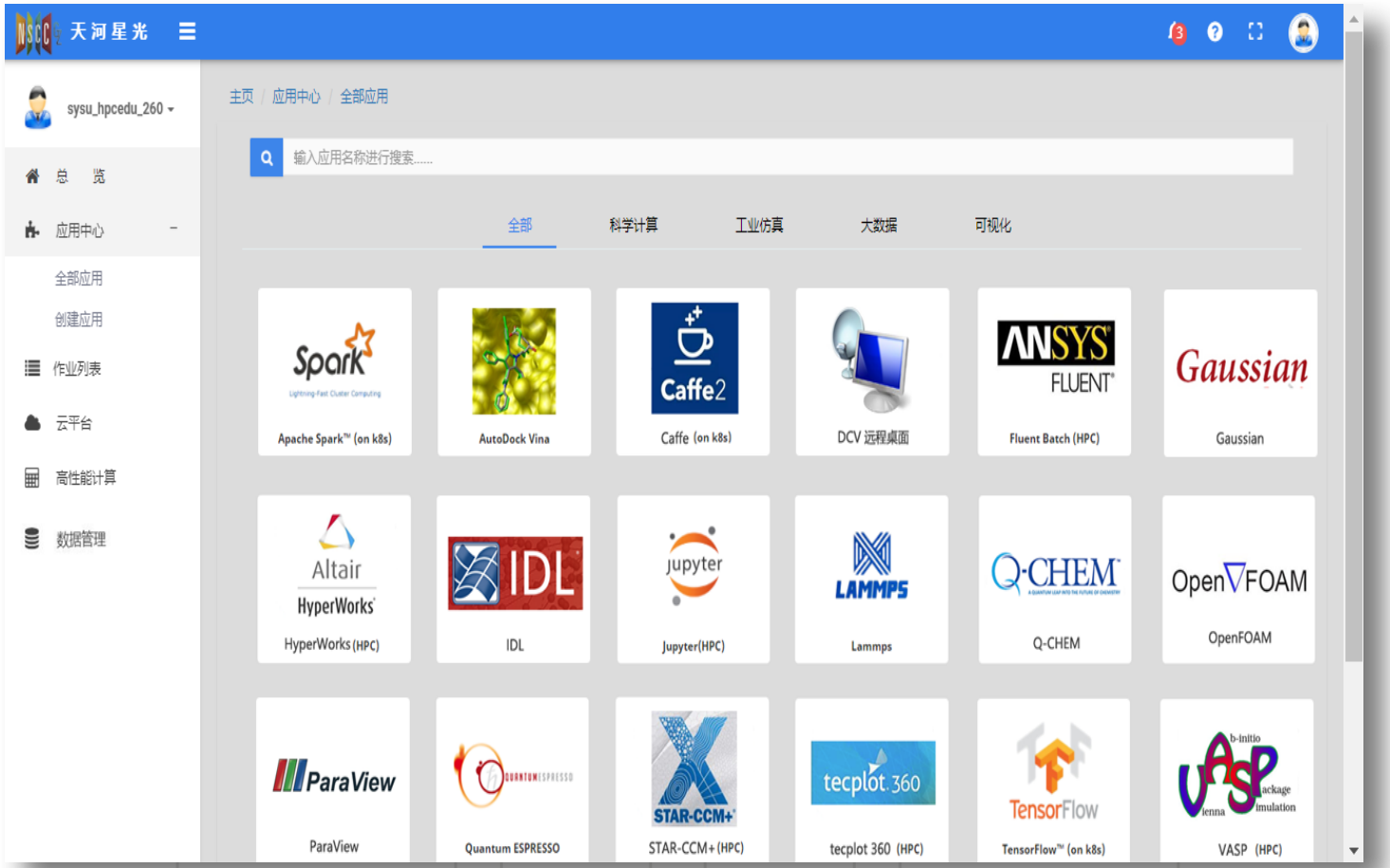


Tianhe-Starlight Platform



Application Center

- **Application Center: Users do not need to concern about resource management**



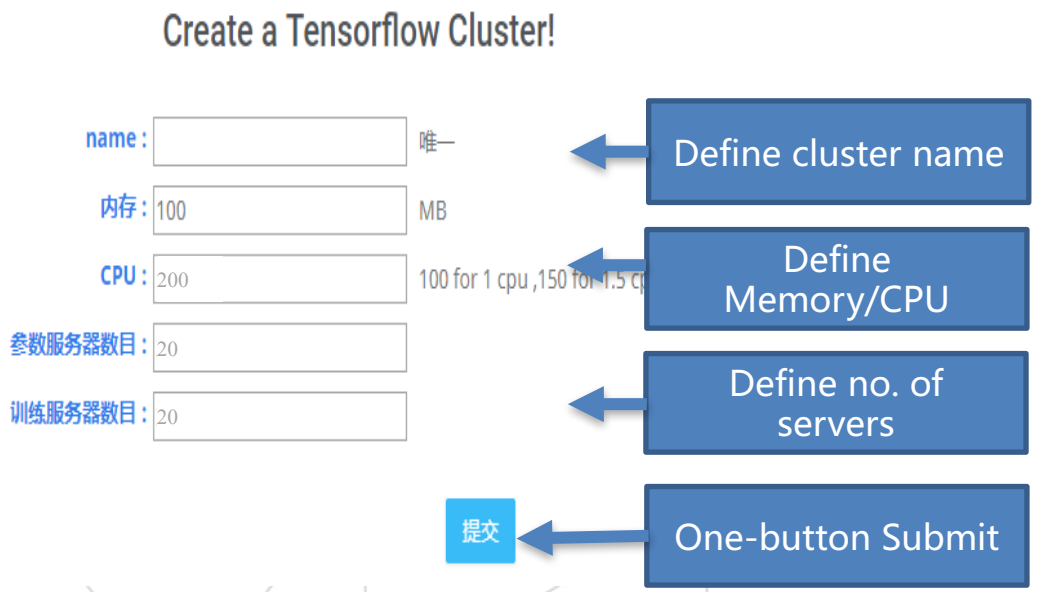
One-button Deployment

- **Support One-button Environment Deployment, which satisfy multiple requirement on both hardware and software**

– **HPC Environment**
  - SLURM Cluster
  - PBS Cluster

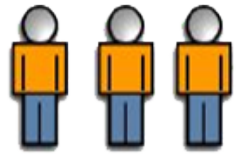
– **BigData Environment**
  - Hadoop Cluster
  - Spark Cluster
  - Storm Cluster

– **AI Environment**
  - Tensorflow Cluster
  - Caffe Cluster



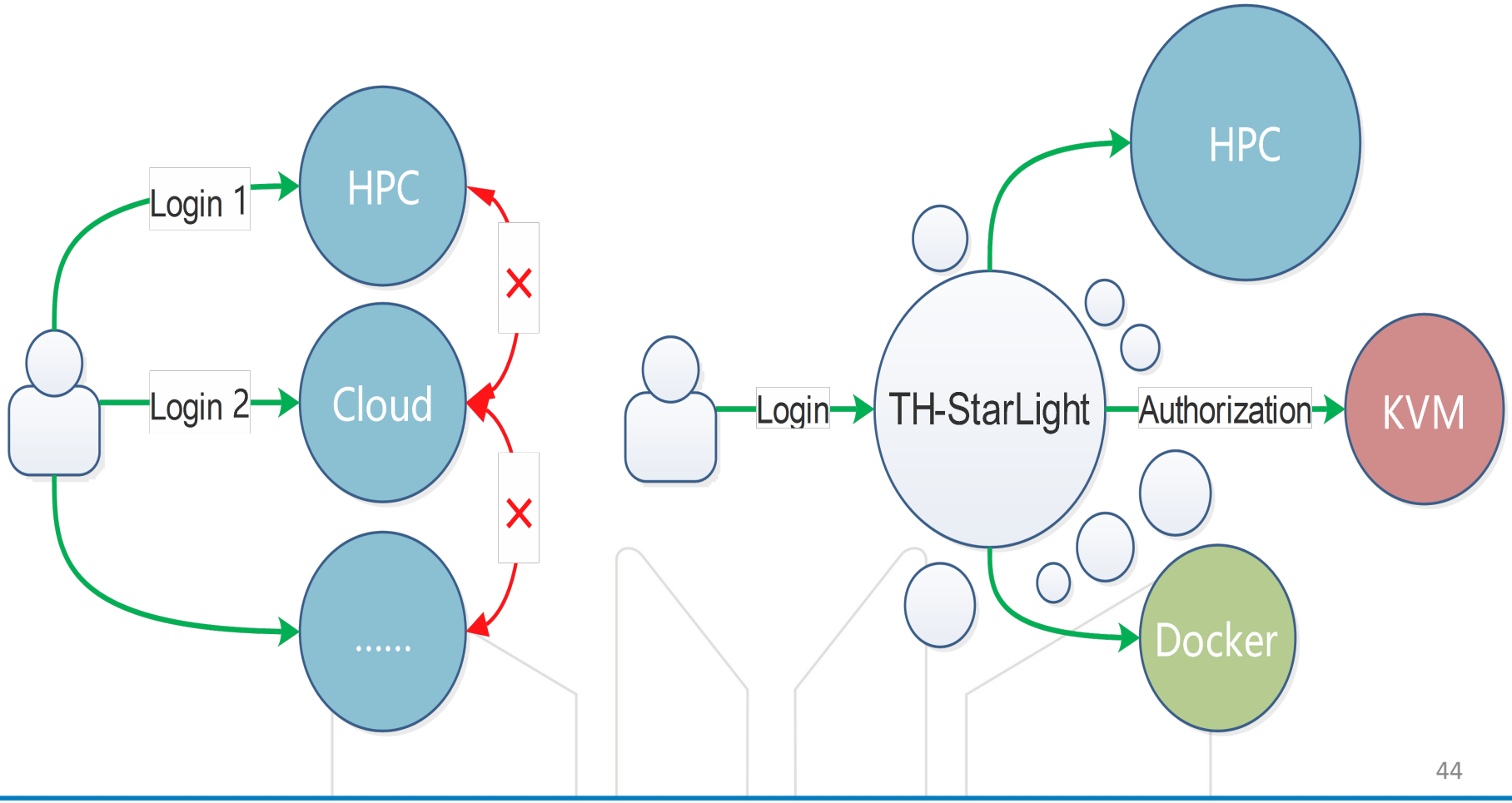
User-friendly Interface

- Unified Account Management
- Support Whole Workflow
- Convenient Data Interaction
- Online Application Development
- Application Store



Unified Account Management

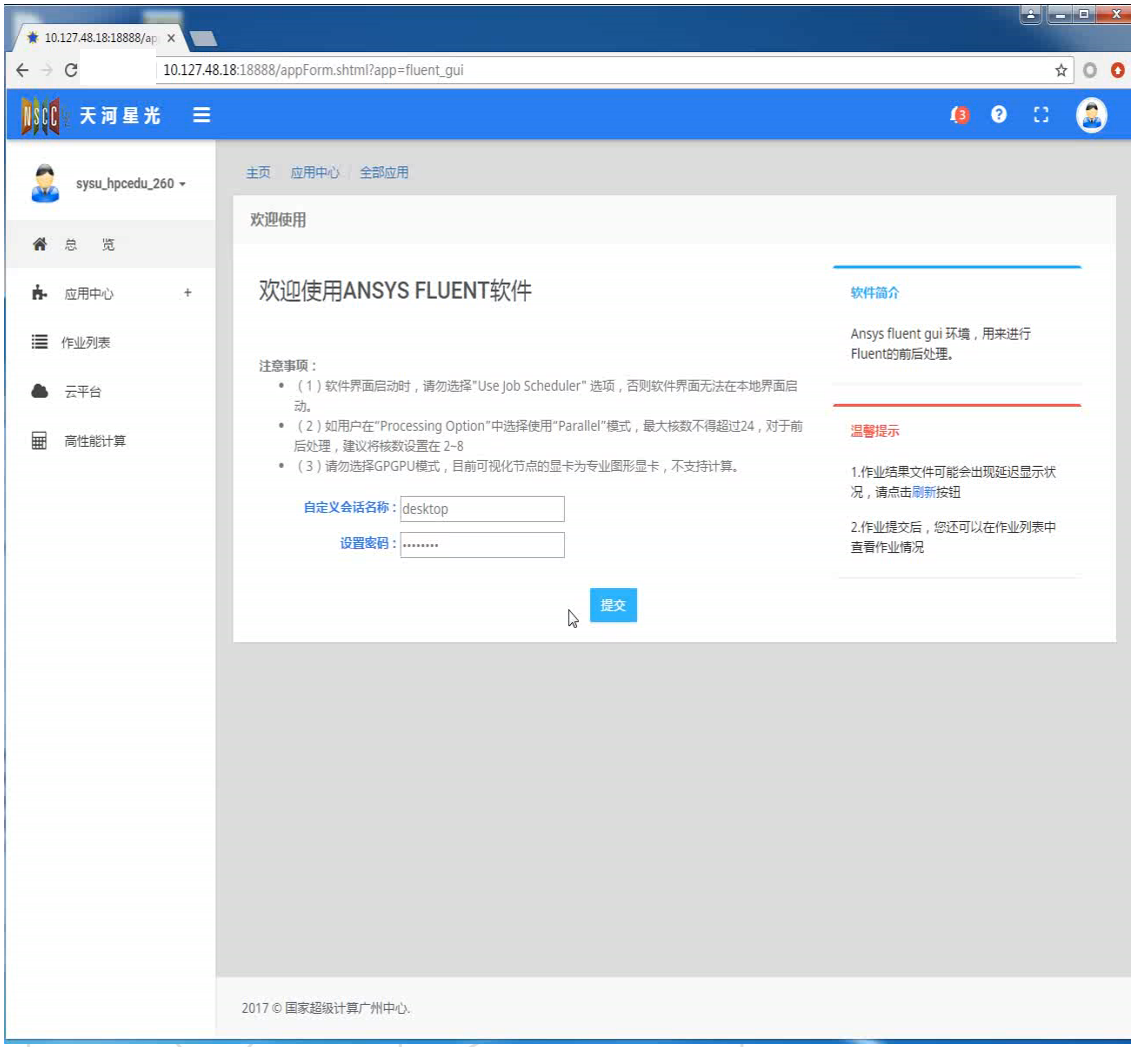
- One Account can access all the resource in NSCC-GZ



Support Whole Workflow

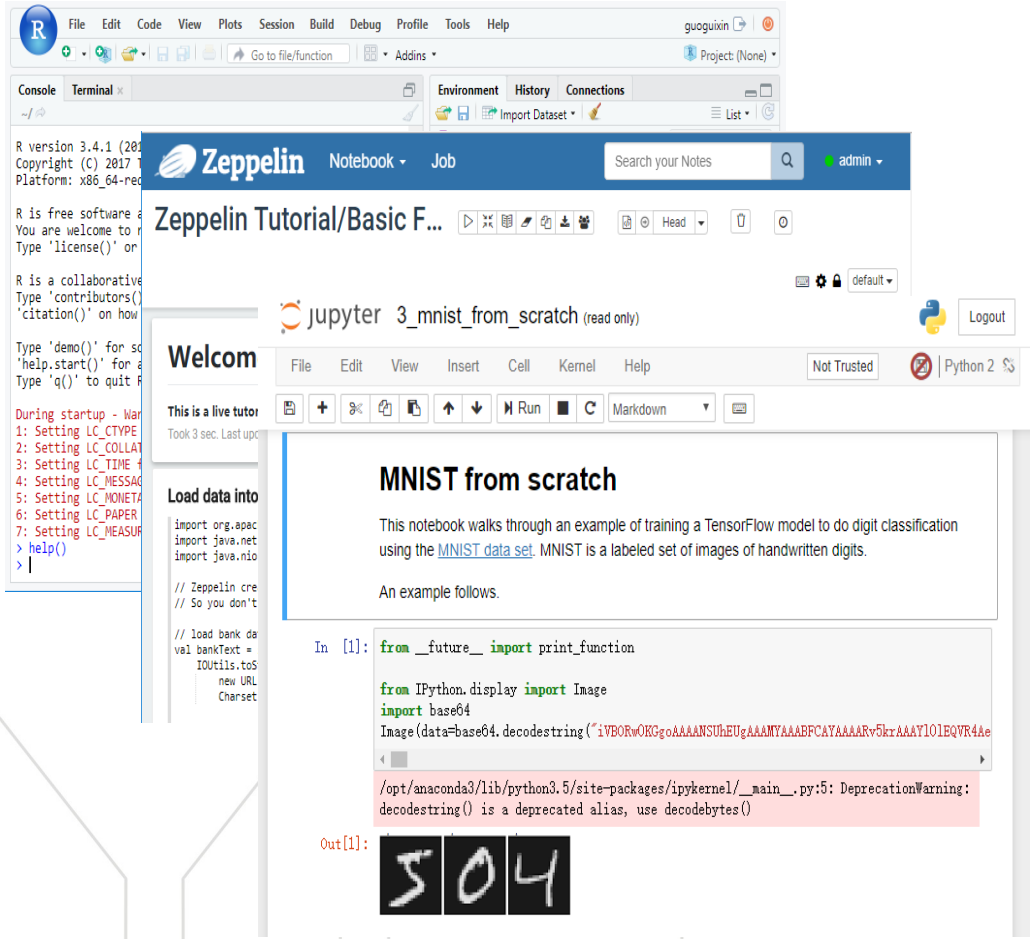
CAE Workflow

- Mesh Generation
- Model Configuration
- Job Submission and Status Monitor
- Result Analyze and Visualization



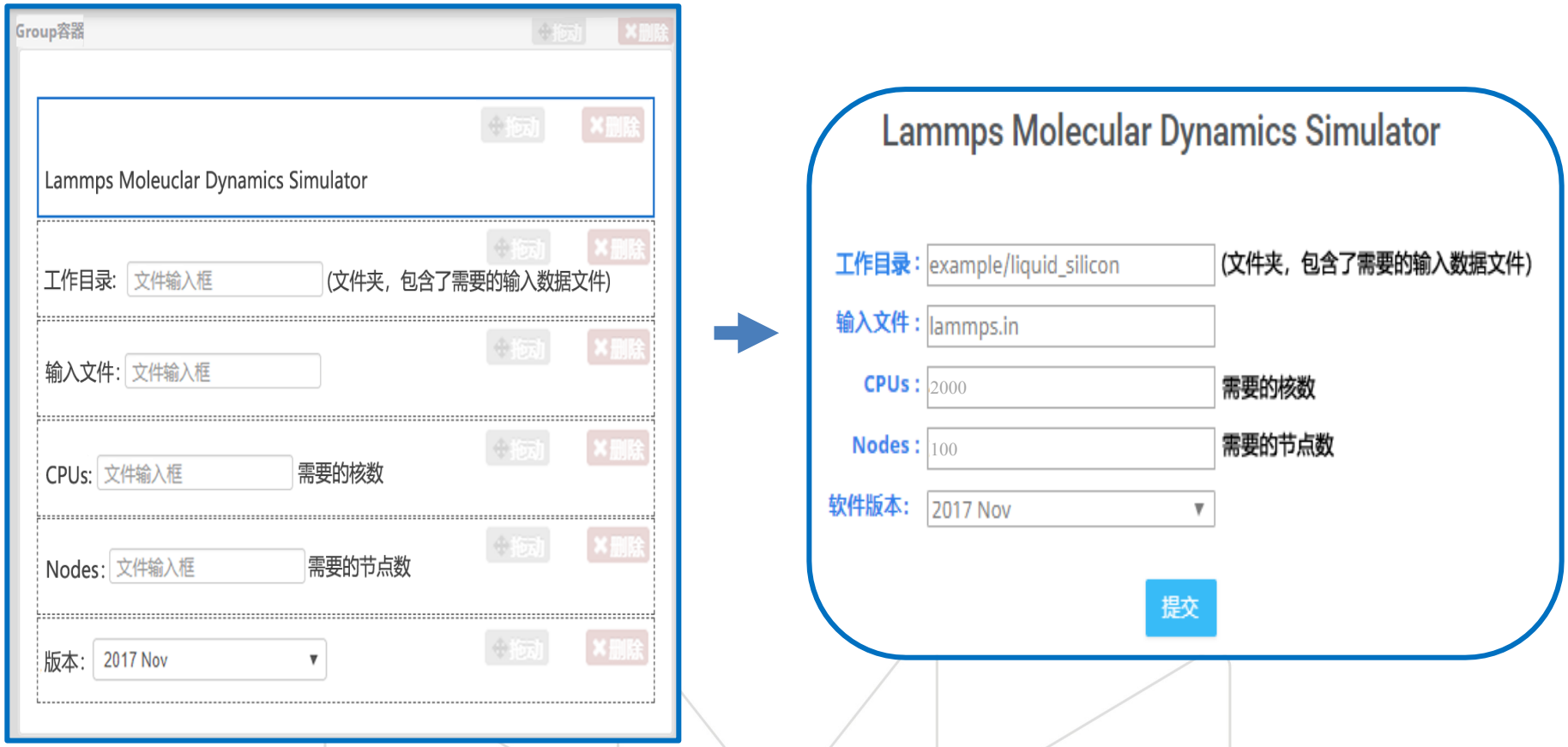
Online Application Development

- **Integrate with various Web IDE， which support online edit, compile and running of multiple languages**
  - **R**
  - **Python**
  - **C、C++**
  - **Fortran**
  - **MPI、OpenMP**



Application Store

- Allow users to publish their own application

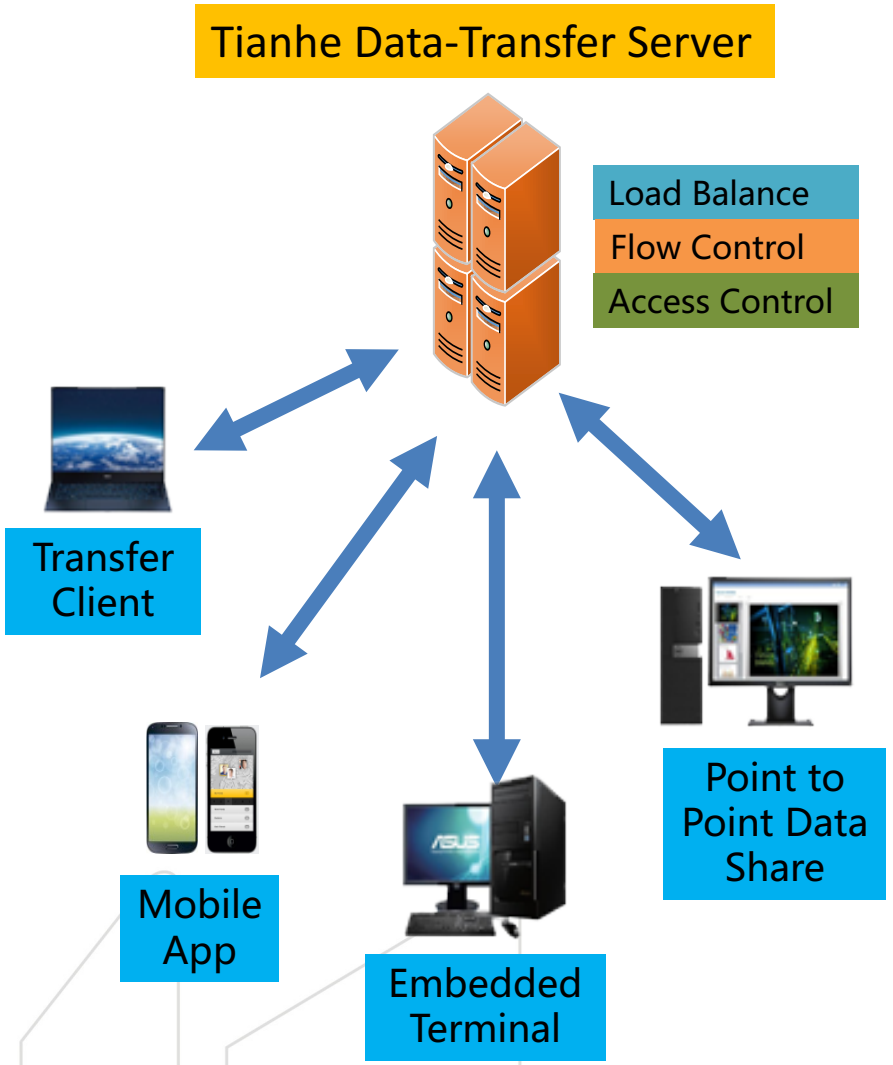


Convenient App editor

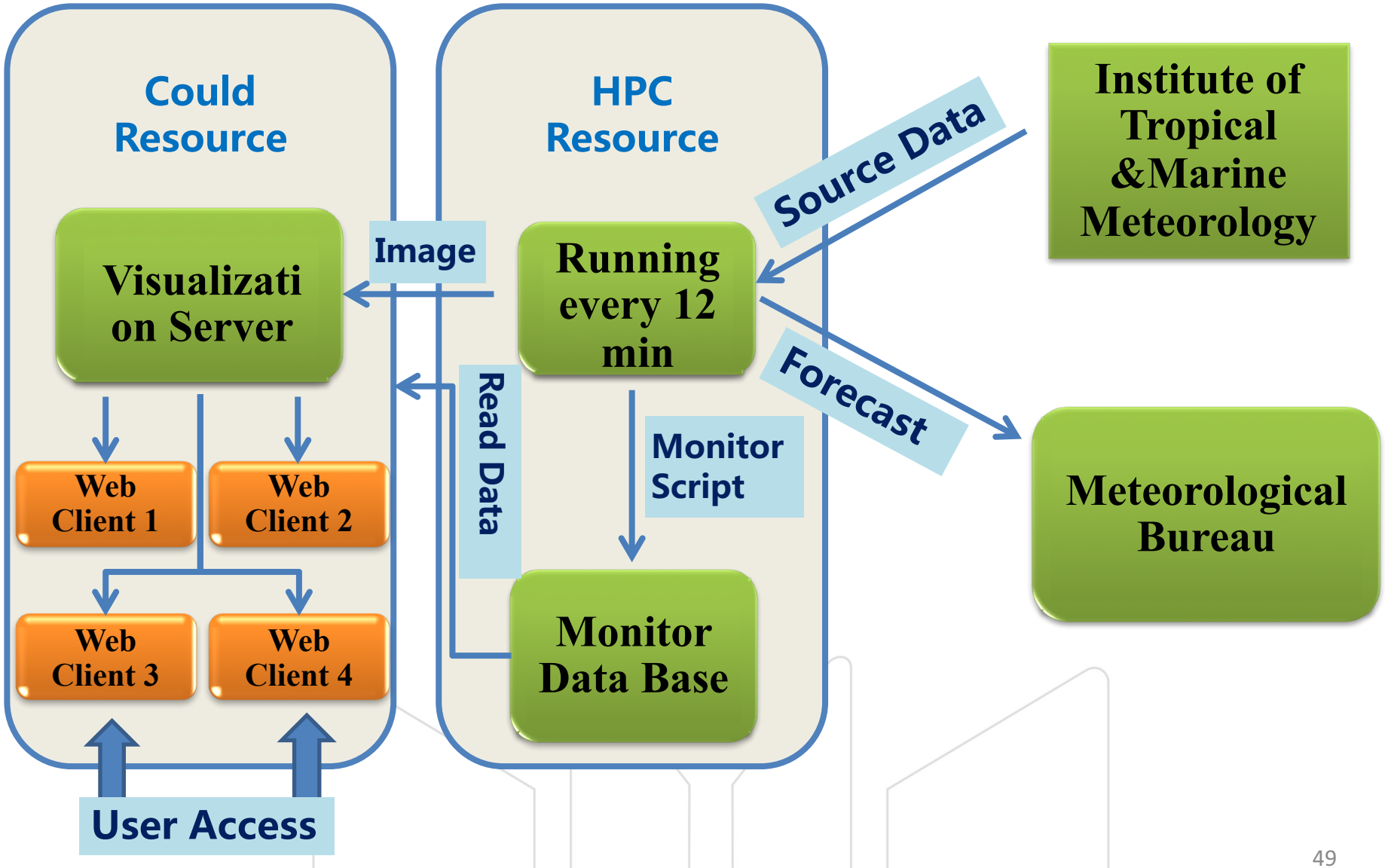
Convenient Data Interaction

- Based on Tianhe fast data transfer protocol, which is 10x faster than FTP or HTTP
- No affected by Size of file, Distance and Network Condition, average bandwidth Usage could reach 90%
- Support Various Terminal Access

  Web/App/Client



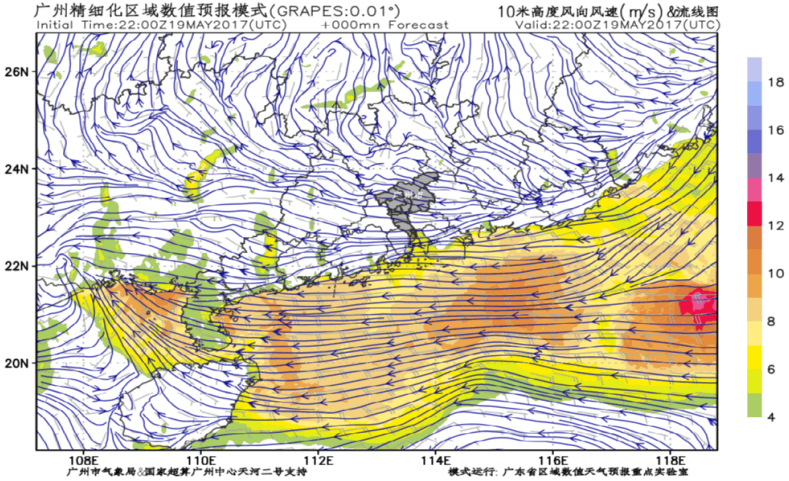
Refined Weather Forecast System

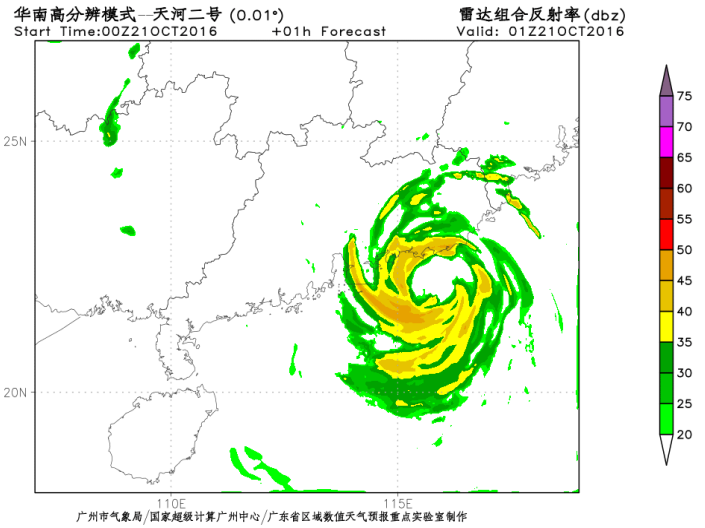


Refined Weather Forecast System

1km&3km Weather Forecast System

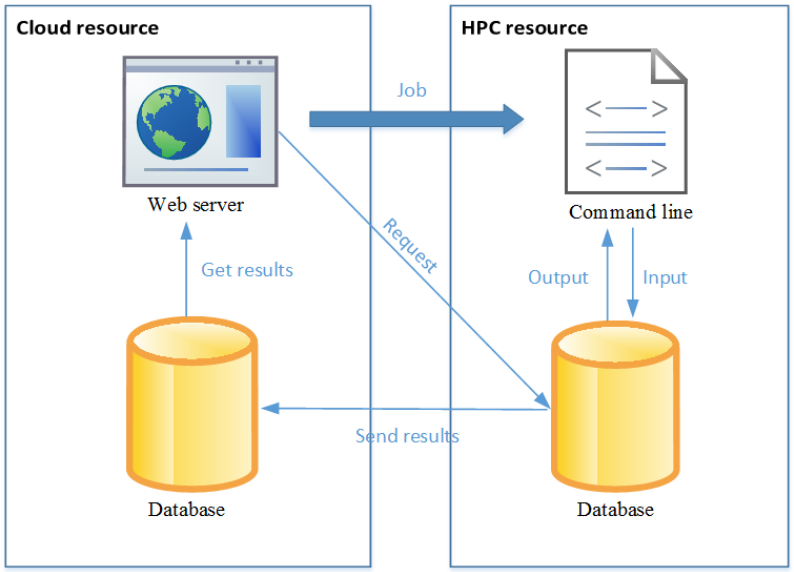
- First Supercomputer Center involved real weather forecast system in China
- 6 hour South China 1km weather forecast in 12 min



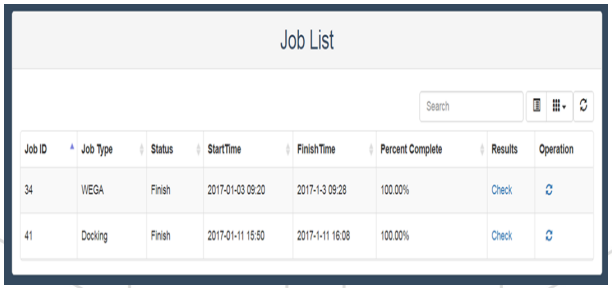


Virtual Drug Screening Platform

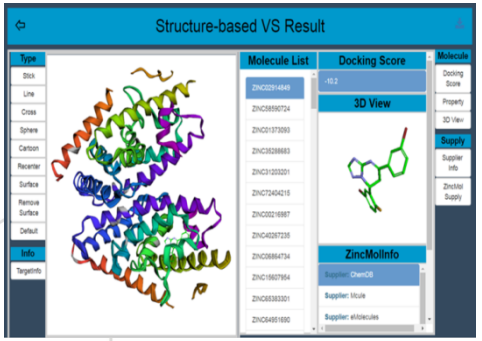
- 10 million molecular docking (ZINC) finished in 22.3 hour
- 2 billion molecular structural similarity calculation finished in 34 min
- Real time online analysis , high through-put visualization



Submit Job



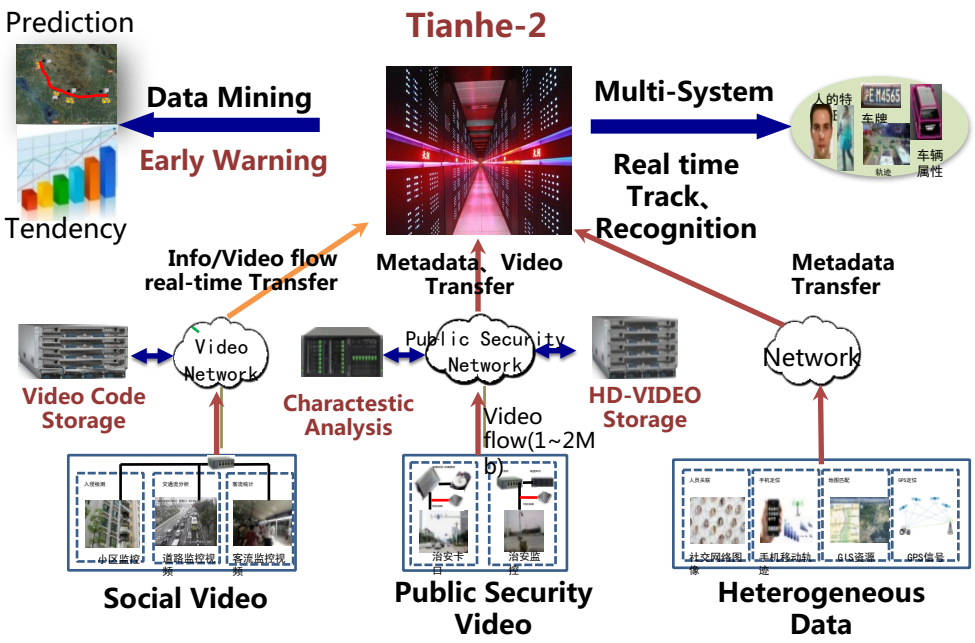
Job Monitor

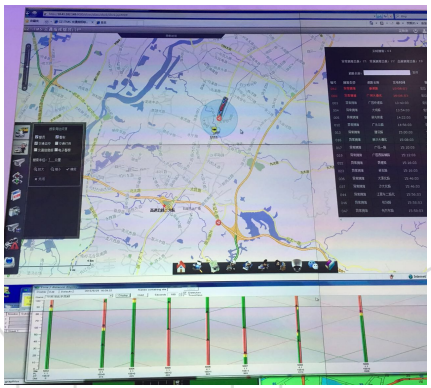


Online Visualization

Traffic Video Processing Platform

- Support 10 million vehicle image recognition , the degree of actuary is more than 95%
- Support 10 billion vehicle image parallel search, the efficiency is more than 3000 frame/sec
- Support 10 thousand channel video aggregation analysis on Tianhe-2
- Cover more than 5 cities in Guangdong Province





Video based on GIS

System Demo

HPC Education Platform

超算习堂
EasyHPC

gunicorn NGINX Flask

* Web Service Frame

MySQL VNC

* Virtual Lab
* Service Component

BIHU

* Provide Unified Resource Interface

openstack docker

slurm workload manager

kubernetes

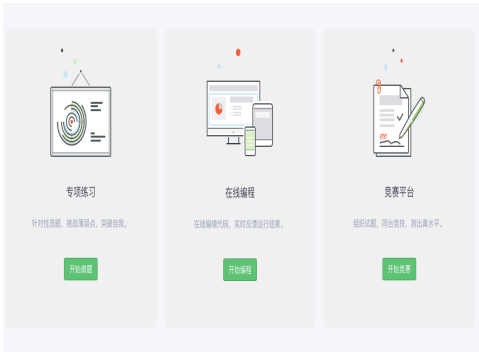
* User Behavior Learning, Course Recommendation …

HPC Education Platform

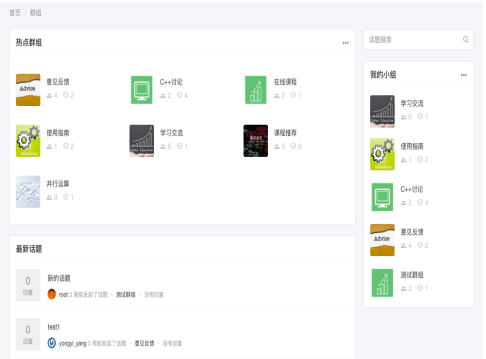
- Supported by National Key Research and Development Plan
- User 1800+
- Support Online Programming ，MPI、OpenMP, Python …
- Dynamic Environment Management
- Course Management
- User Community

Online Course

Online Programming



Online Test



Community

Converge HPC, BD, and AI

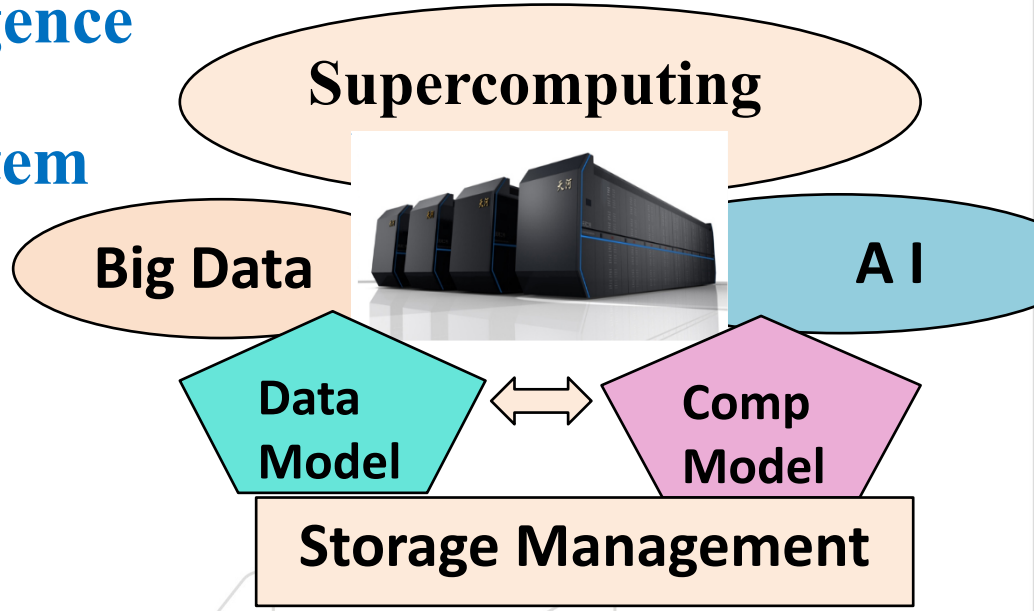
- HPC, BD, and AI are tightly connected and support each other
- HPC is the basic measure for big data analytics and deep learning-enabled AI technology and applications
- The existence of big data changes the way of AI research and applications
- Future supercomputer architecture and implementation will be deeply influenced by big data and AI
- Development of HPC, BD, and AI must be converged

Converge HPC, BD, and AI

- **Basic theory**
  - Common mathematical models, methods and basic algorithms
  - Novel computer architectures supporting BD and AI
  - Programming models and methodologies
- **Key technology and platform**
  - High efficient implementation of algorithms
  - Tools for performance and energy efficiency optimization
  - Runtime support
  - Platforms for fast generation and deployment of application platforms
- **Domain applications**
  - Exploring applications characterized by converged development of HPC, BD and AI
    - Smart Autonomous Systems, Smart health
  - Applications reformed by new methods

Summary

- Heterogeneous vs Homogeneous Architecture
- Adaptive system and software design
- HPC Bigdata AI Convergence
- Supercomputing Eco-system
- Collaborations



中山大學
SUN YAT-SEN UNIVERSITY

国家超级计算广州中心
NATIONAL SUPERCOMPUTER CENTER IN GUANGZHOU

Thanks