IPDPS 2014 PHOENIX

Scalability-Centric System Design for Large Scale Computing

Prof. Yutong Lu

School of computer science, NUDT &

State Key Laboratory of High Performance Computing

ytlu@nudt.edu.cn

国防科学技术大学

National University of Defense Technology

Overview

National university of defense technology

~2,000 Teachers

~15,000 Students

… Others

国防科学技术大学

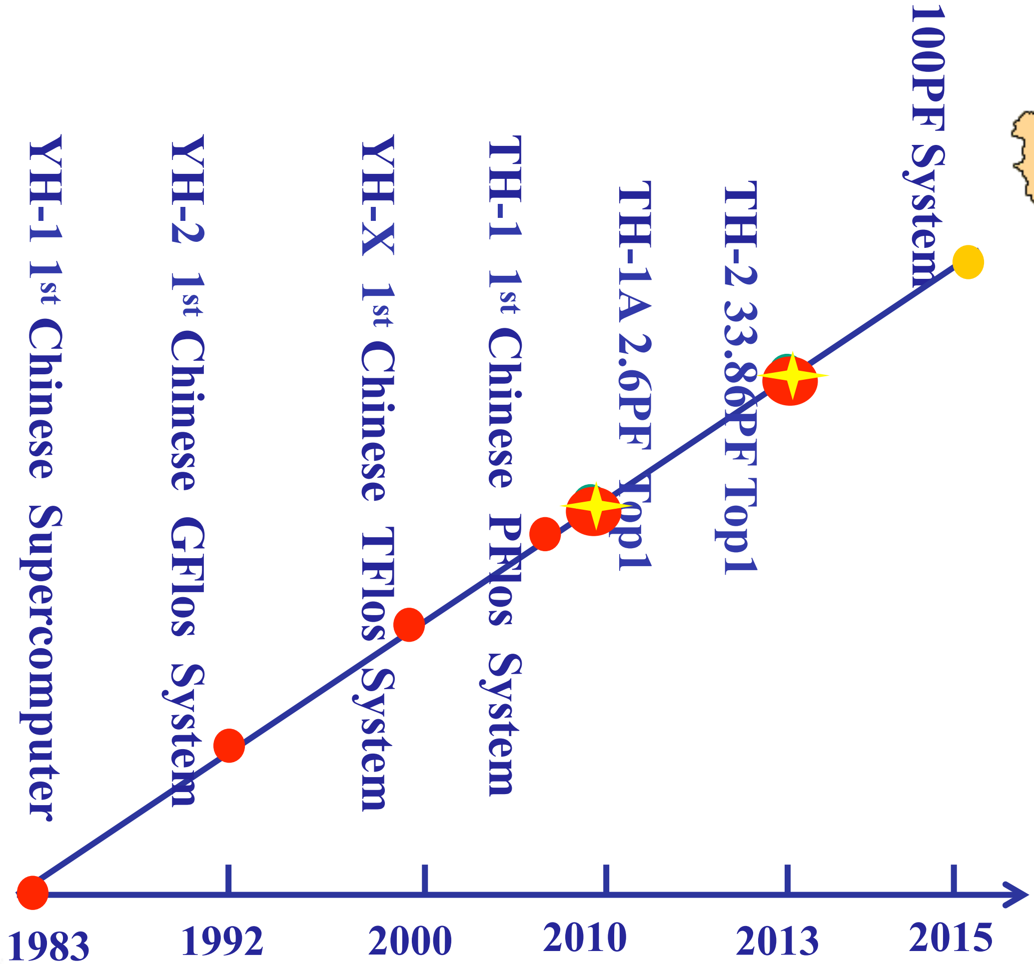
National University of Defense Technology

HPCL

Tianhe 天河

Overview

Supercomputers in NUDT, Changsha, China



NSCC-Tianjin,2010

NSCC-Guangzhou,2013

NSCC-Changsha,2012

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

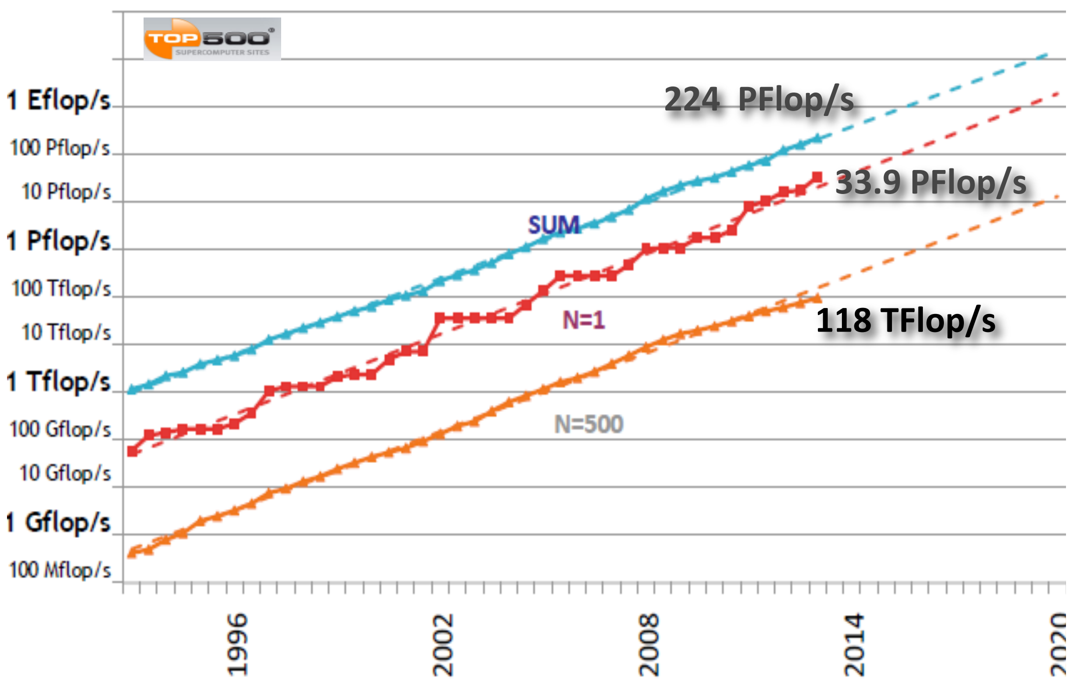
Outline

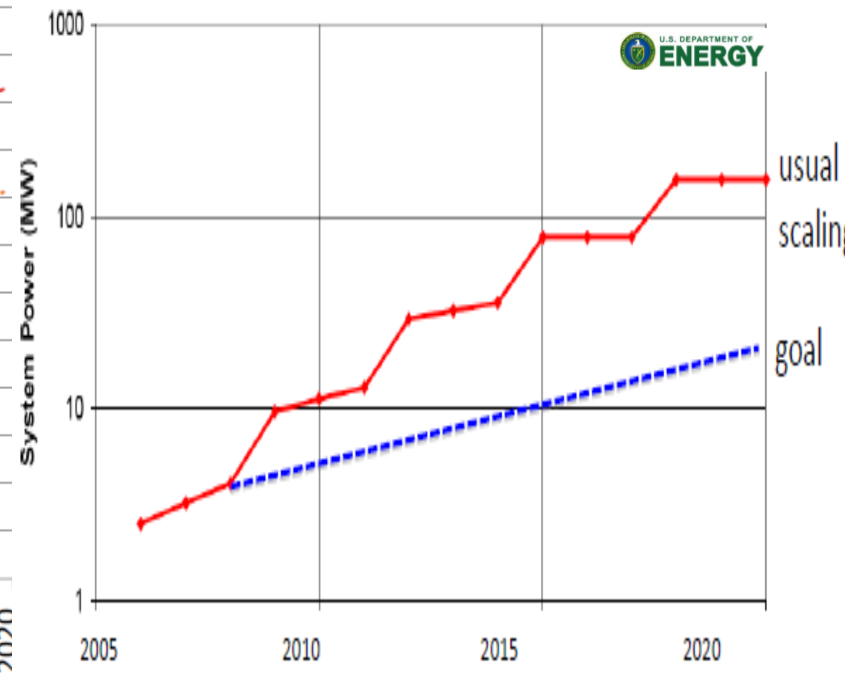
- Trend of HPC Architecture
- Scalable System Software Design
- Applications

Challenges

PSPR

- ❒ Performance
- ❒ Scalability
- ❒ Power consumption
- ❒ Reliability





Trend of Architecture

- Tree carriages of Performance
  - Frequency
  - ILP
  - Parallelism
- Performance = Parallelism
  - ……
  - Year 2010：TH-1A，4.7Pflops，7168Nodes，186,368 Cores
  - Year 2013：TH-2， 54.9Pflops， 16000Nodes， 3,120,000 Cores
  - ……
- Exploit parallelism
  - Longitude (100,000nodes)
  - Latitude（multi/many cores, SIMD、ILP）

Trend of Architecture

❒ Heterogeneous architecture

- ➢ Some of top-level supercomputers
  - ◆ Tiahhe-1A
    - – NVIDA M2050 GPU
  - ◆ Tianhe-2
    - – Intel Xeon Phi
  - ◆ Titan
    - – NVIDIA K20X GPU
- ➢ Heterogeneous systems on latest Top500 list
  - ◆ #53 /Top500, #24 /Top100，#4 /Top10

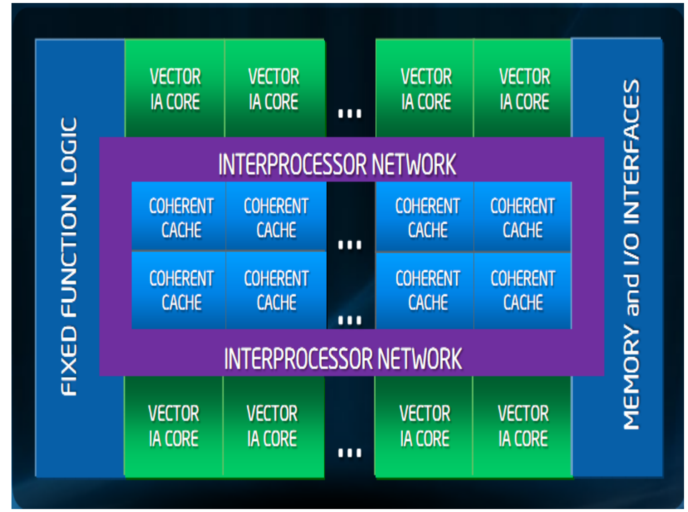
❒ Compute Efficiency

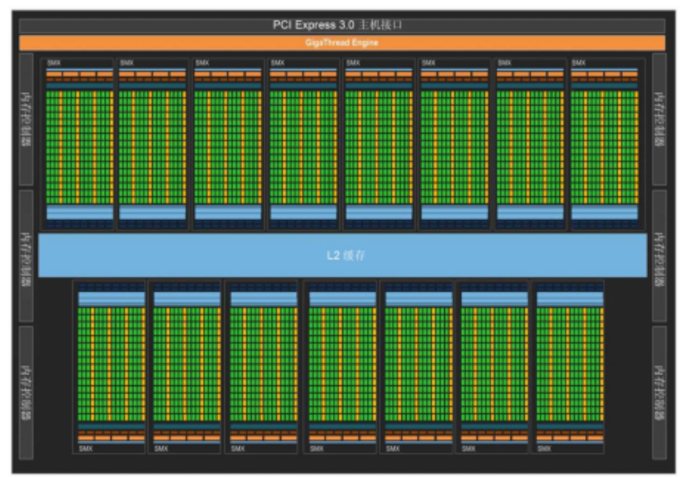
- ➢ More computations per joule
- ➢ More computations per transistor

Trend of Architecture

Tianhe 天河

- ❐ **Many core processor**
  - ➢ **Intel MIC**
    - ◆ >60cores, >200threads
    - ◆ 1.15GHz
    - ◆ > 1TFlops performance
    - ◆ 512b SIMD
  - ➢ **GPU, NVIDIA Kapler**
    - ◆ 2688 cores
    - ◆ 732MHz
    - ◆ 1.31TFlops





HPCL

Trend of Architecture

Tianhe♪ 天河

Tianhe supercomputers

TH-1A

GPU vs

- Data Parallel
- Simple instruction
  - Limited scheduling
- GPU Direct available
  - ~40% ↑ MPI communication on Tianhe-1A
  - 5% ↑ Linpack
- Steep learning curve
- Supporting
  - Cuda
  - Open CL
  - …
- 2CPU + 2GPU Linpack ~71%
- Whole system Linpack 56.5%

TH-2

MIC

- Multi threads & SIMD
- Flexible modes
  - Native, Offload, Symmetric, Shared
- SIMD available
  - ~ 4.5 times speedup on Tianhe-2
- Relatively easy to get started
- Intel Supporting
- 2CPU + 3MIC Linpak ~76.5%
- Whole system Linpack 61.6%

HPCL

Trend of Architecture

Tianhe 天河

GPU

- Computational Chemistry and Biology
- Numerical Analytics
- Physics
- Manufacturing: CAD and CAE
- Oil and Gas
- Defense and Intelligence
- Computational Finance
- Media and Entertainment

MIC

- Computational Chemistry and Biology
- Electronic Structure
- Physics
- Computational Fluid Dynamic
- Astrophysics
- Environment
- Oil and Gas
- Computational Finance

Trend of Architecture

Memory Hierarchy

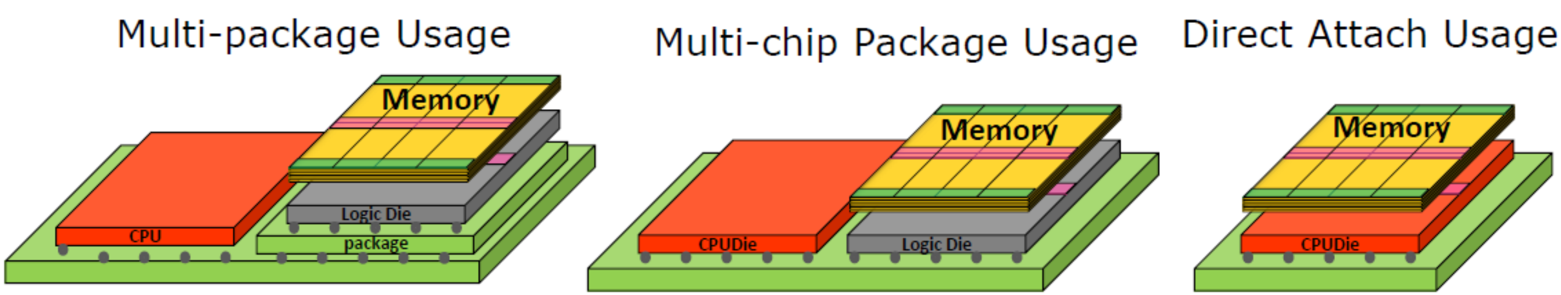
- Performance of CPU ↑59%, Perf of MEM ↑ 26%

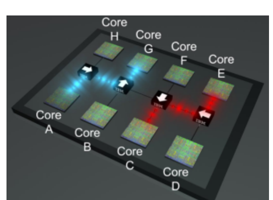
Register	1 circle
L1 Cache	3 circle
L2 Cache	10 circle
L3 Cache	30 circle
Local Mem	150 circle
Non-local MEM	>1500 circle

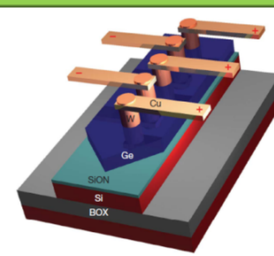
- Exploit Data Locality, reduce communication and memory accessing

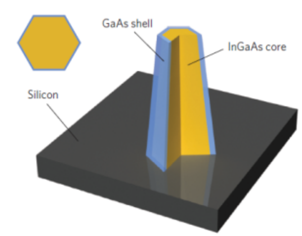
Trend of Architecture

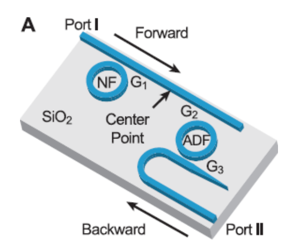
- **Memory architecture will be benefited from multiple technologies**
  - **Deeper memory hierarchy**
  - **Advanced package technology**
    - **3D stack、MCM**
  - **Optical connection btw chips**











Trend of Architecture

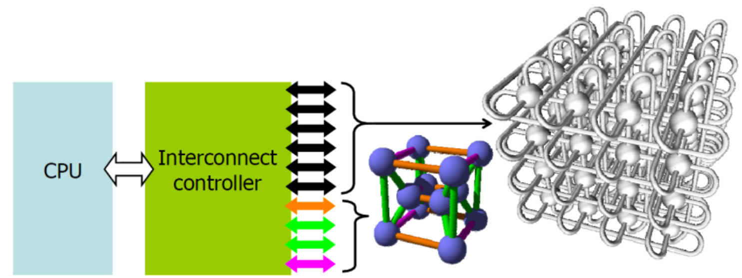
Power Consumption

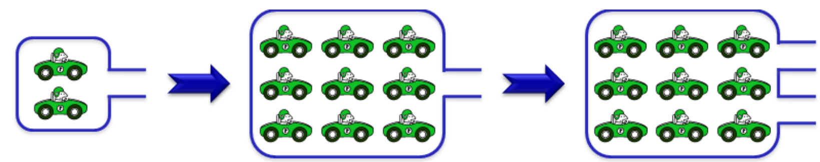
- PW for data moving / 48X PW for data computing
  - MLA inside core: 100PJ
  - Read inside CPU: 4800PJ
  - Data moving btw cores: 7500PJ
  - Data moving btw nodes: 9000PJ
- DTF, reduce 20% power consumption, with 5% performance losing
- Power control applications, power aware, minimum data moving

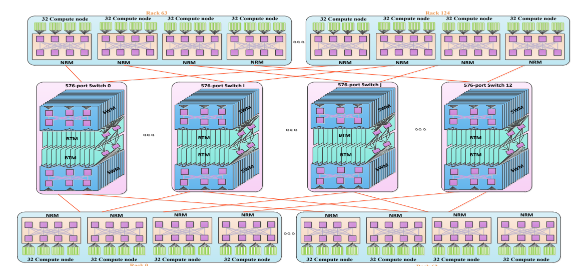
Trend of Architecture

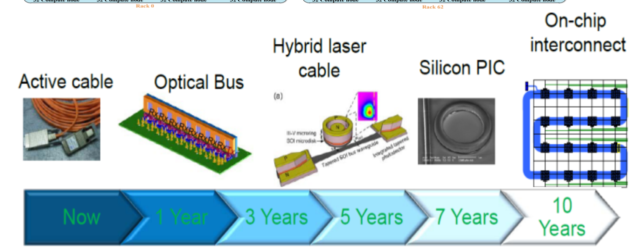
Interconnection network

- ❐ **NIC**
  - ➢ High Bandwidth
  - ➢ Multiple Lanes
- ❐ **Router**
  - ➢ High radix Vs. Low radix
- ❐ **Topology**
  - ➢ N-D Torus Vs. Fat Tree
  - ➢ N Dimension Tree
- ❐ **Optical**
  - ➢ High BW, Low Latency, EMC
- ❐ **Cost**
- ❐ **Topo-aware software**

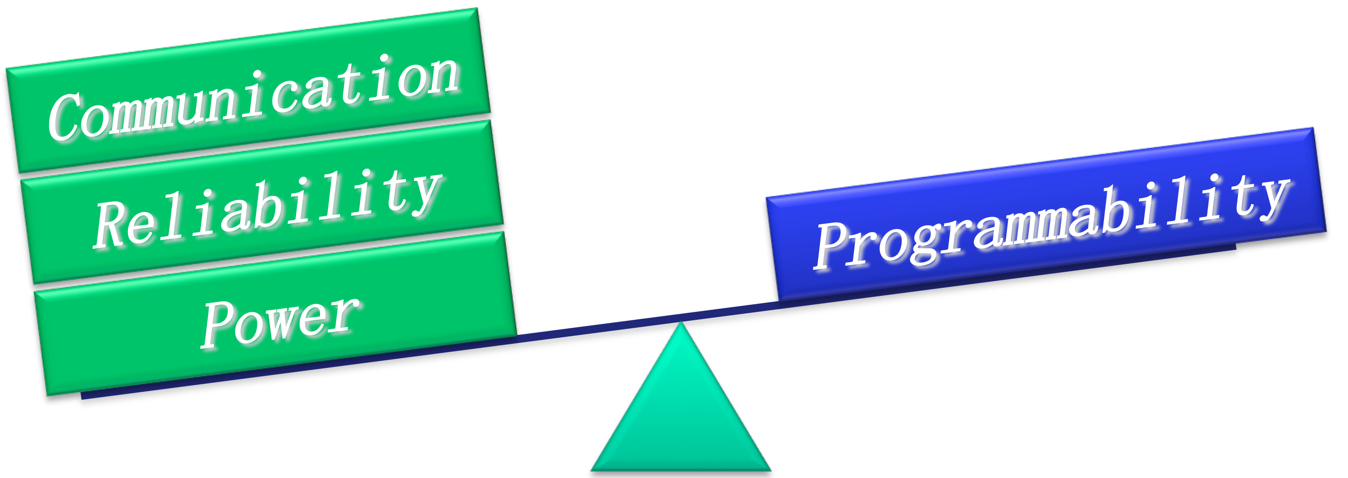








Trend of Architecture



Heavy the burden of Software

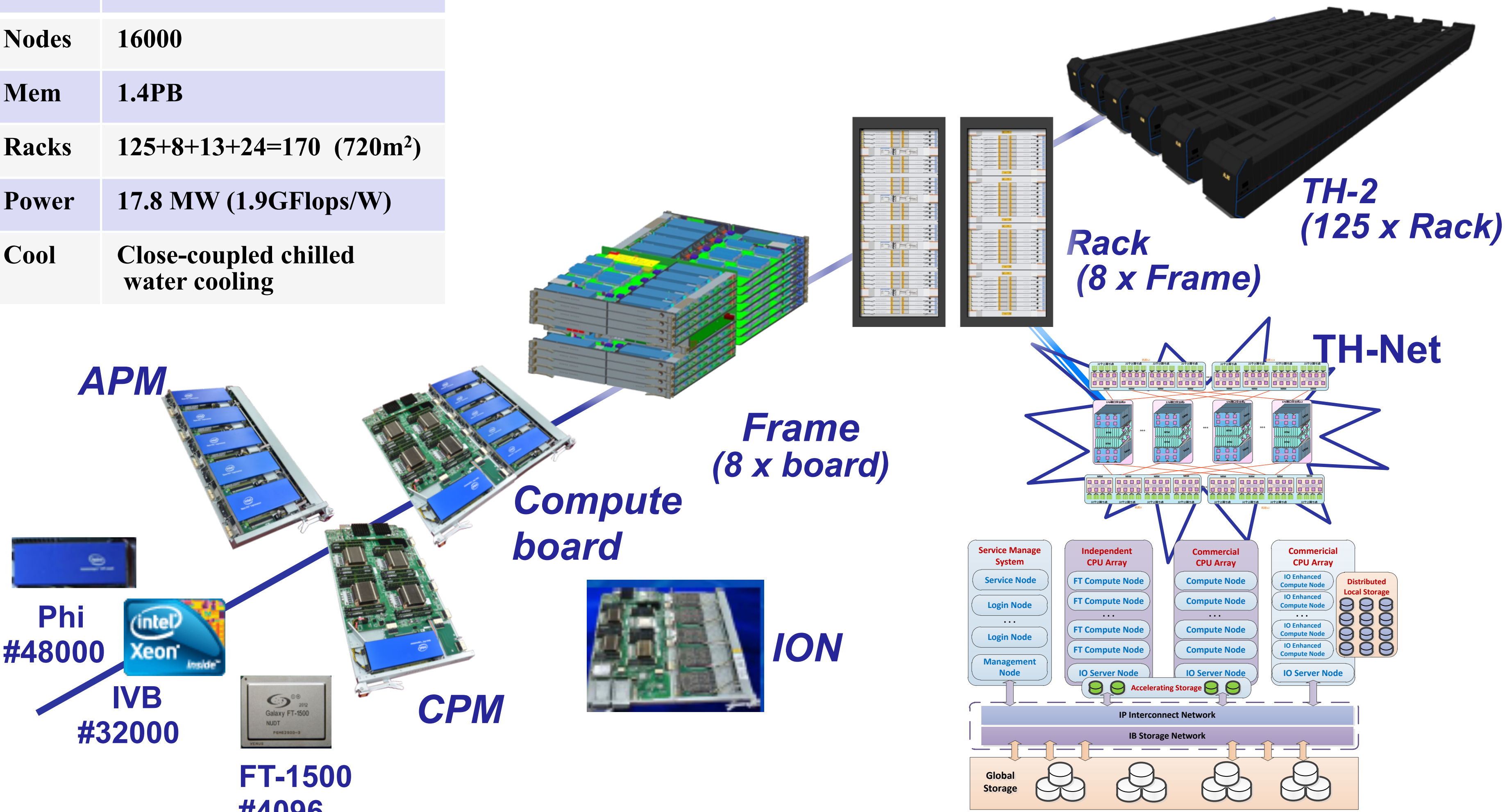
Software issues

- Scalability
  - How to use the exist systems better
  - How to explore the next generation systems
- Resilience
  - Reduce the CR overhead
  - Lightweight resilience method
- Power Control
- Programmability
- HPC vs Big data
  - Data management and filesystem

Highlights of Tianhe-2

Tianhe 天河

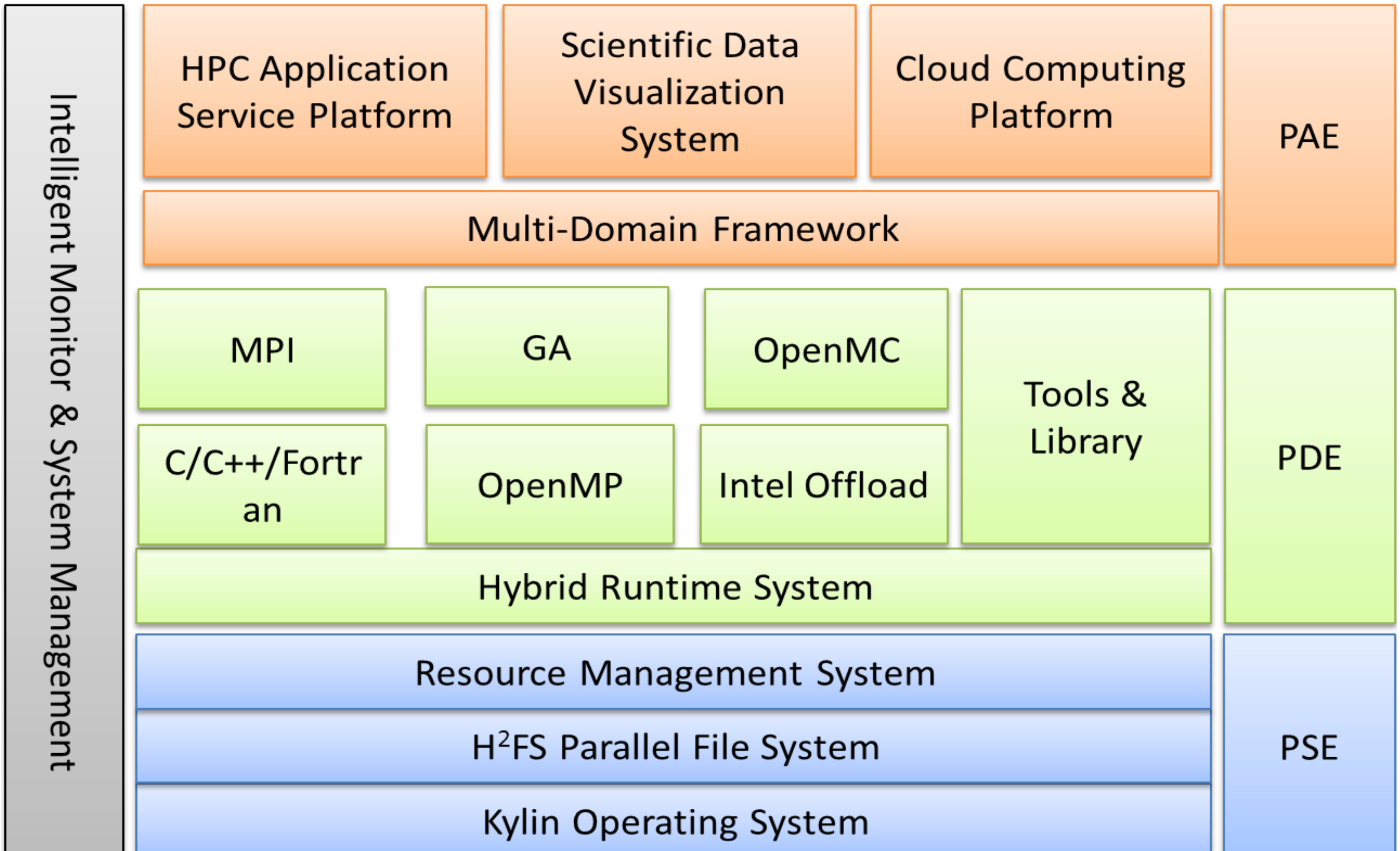
Perf	54.9PFlops / 33.86PFlops
Nodes	16000
Mem	1.4PB
Racks	125+8+13+24=170 (720m²)
Power	17.8 MW (1.9GFlops/W)
Cool	Close-coupled chilled water cooling



Hybrid Hierarchy shared storage System
12.4PB

Highlights of Tianhe-2

❒Software Stack



Programming model

❑ Trend of programming model

- ➢ Whole system
  - ◆ MPI
  - ◆ New Data-driven model
- ➢ Intra node
  - ◆ Various
    - – OpenMP, Cuda/OpenCL, OpenACC
- ➢ Others
  - ◆ PGAS (Global Array)

Portability
Performance
Simplicity and Symmetry
Modularity
Compatibility
Completeness Distributed memory

Scalable MPI

- **Performance**
  - P2P: Bandwidth/Latency
  - Collective communication
  - Communicator/Group operations
  - MPI-Init
- **Resource consumption**
  - Memory
  - Network connection
- **Measurement?**

Scalable MPI

- Mem consumption for MPI implementation
  - p: System Scale(#rank)
  - $M \propto O(p^2)$ --conventional implementation based table
  - $P=10^3$, $M=4B*10^6= 4MB$ ☺
  - *$P=10^6$, $M=4B*10^{12}= 4TB$ 😐
  - $P=10^8$, $M=4B*10^{16}=$ 40PB ☹
  - *$P=10^9$, $M=4B*10^{18}=$ 4ZB
- Data structures should be redesigned
  - Communicator, RMA window, protocol buffer…

Scalable MPI

- TH-Express2 & TH-Express2^{+}
  - Network Interface Chip: NIC
    - 10Gbps X 8lane
    - 14Gbps X 8lane(plus)
  - Network Router Chip: NRC
    - 16 ports, more(plus)
  - Optic and electronic hybrid network
  - Topology: Fat tree → N Dimension Tree
  - Design for extension to 100PFlops

Scalable MPI

Message Passing services over TH-Express

- Galaxy Express (GLEX)
  - Basic message passing infrastructure on network interface
  - User level communication technology
  - User and kernel API
- MPICH-GLEX Design Consideration
  - Protocol: different communication mechanisms exhibit different performance and resource usage
  - Application characteristic: communication mode, such as nearest-neighbor communication
  - Scalability: balance between performance and resource usage

Scalable MPI

- Message passing protocols
- Various protocols in low level with TH-Net
  - Eager Protocol
    - Exclusive RDMA Channel
      - Performance oriented
    - Shared RDMA Channel
      - Scalability oriented
    - Hybrid channels
      - Combine application model
  - Rendezvous protocol
    - Zero-copy data transfer based on RDMA Get
- Performance benefit from the neighborhood communication in a number of applications

Scalable MPI

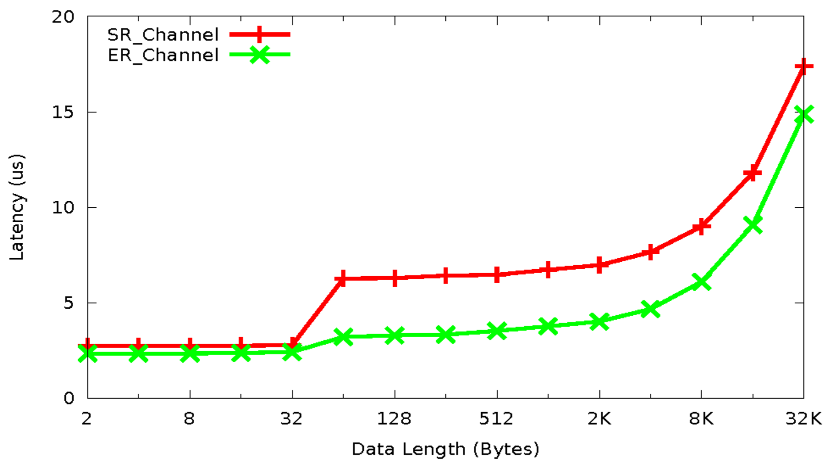
❑ P2P Performance

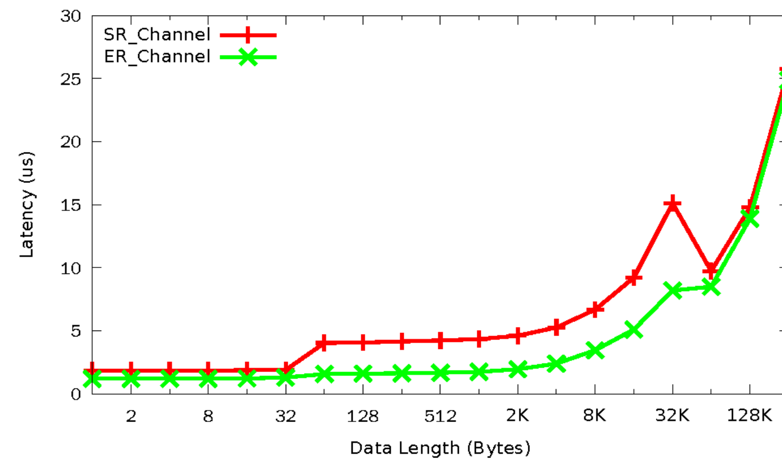
TH-Express2

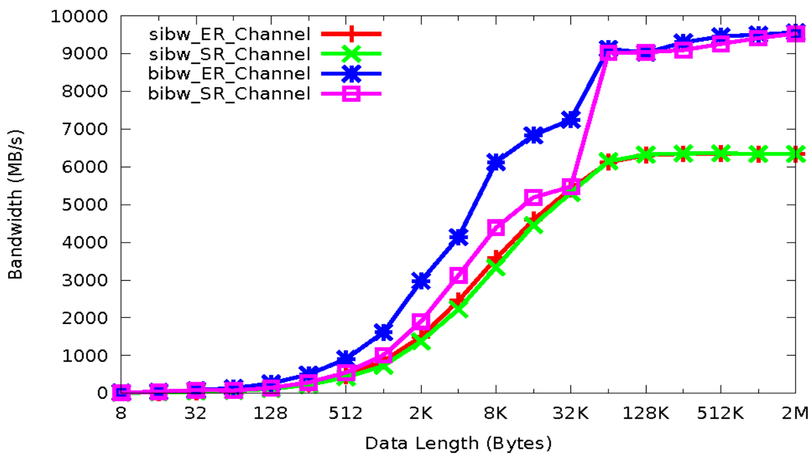
- MPI P2P Bandwidth 6.3GB/s
- Latency ~2us

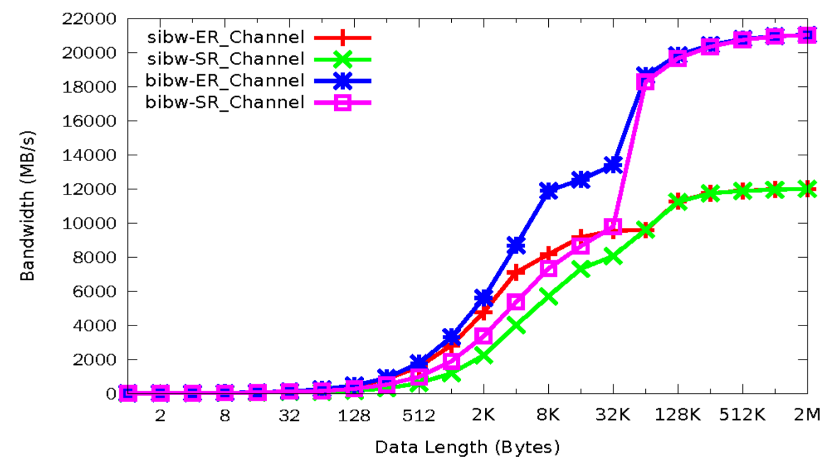
TH-Express2+

- MPI P2P Bandwidth 12GB/s
- Latency ~1us



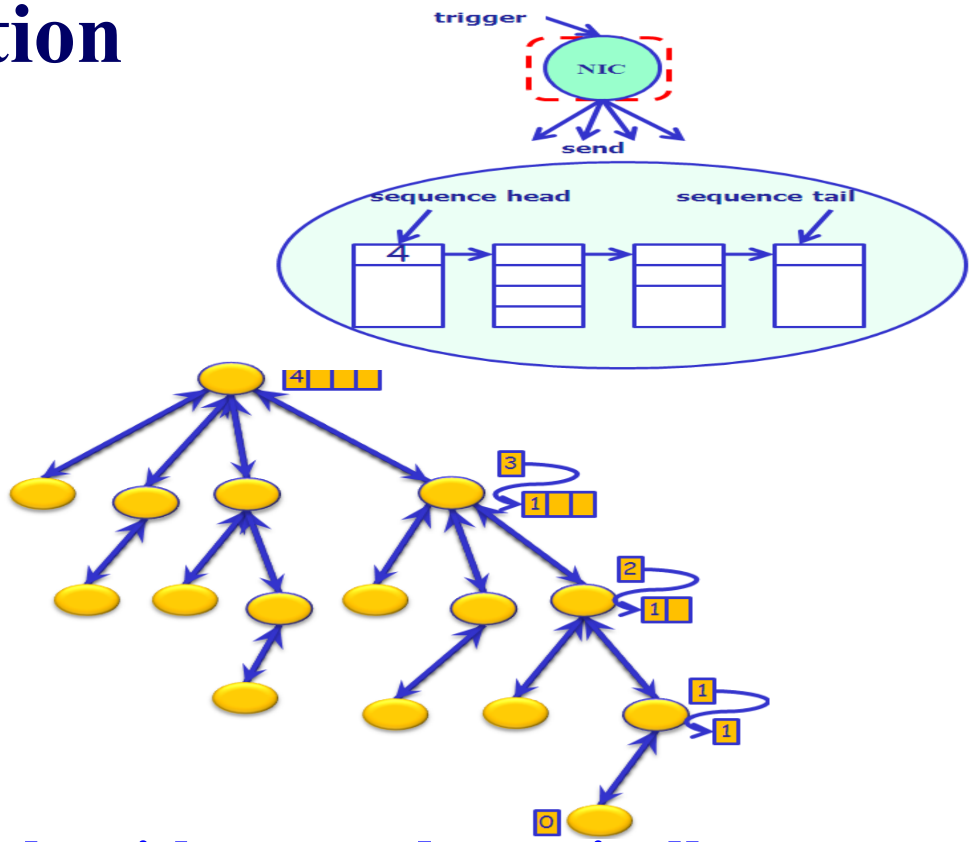






Scalable MPI

- **Collective communication**
  - **MPI interface level**
    - **NonBlock collective**
    - **Alltoallv/AllGetherV**
    - **Group-split**
  - **Implementation level**
    - **Scalable algorithm**
    - **Topology aware**
    - **Hardware offload**



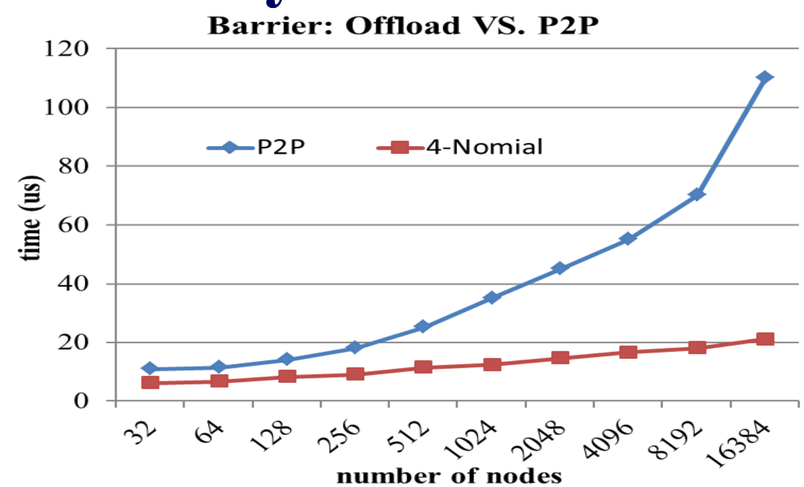
- **Collective offload**
  - **Construct topology-aware algorithm tree dynamically**
  - **Message pass automatically based on the trigger of NIC**
  - **Bypass effect of OS noise**

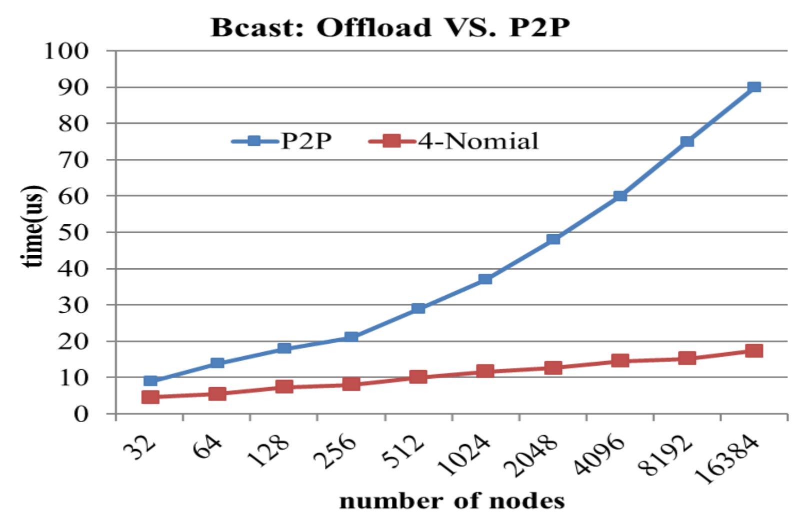
Scalable MPI

Tianhe 天河

❐ Collective Optimization for Scalability

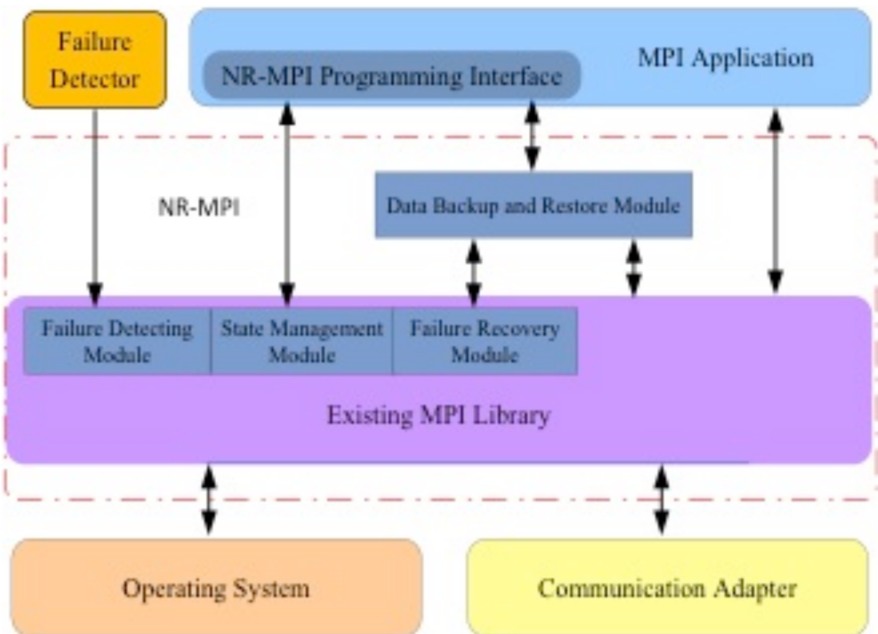
- Two-level Collective Operations
  - Intra-node: shared-memory
  - Inter-node: network
- Adaptive tree structure
  - K-nominal
  - K-ary
  - K is a variable value
- Optimization based on topology
  - Mapping processes to nodes

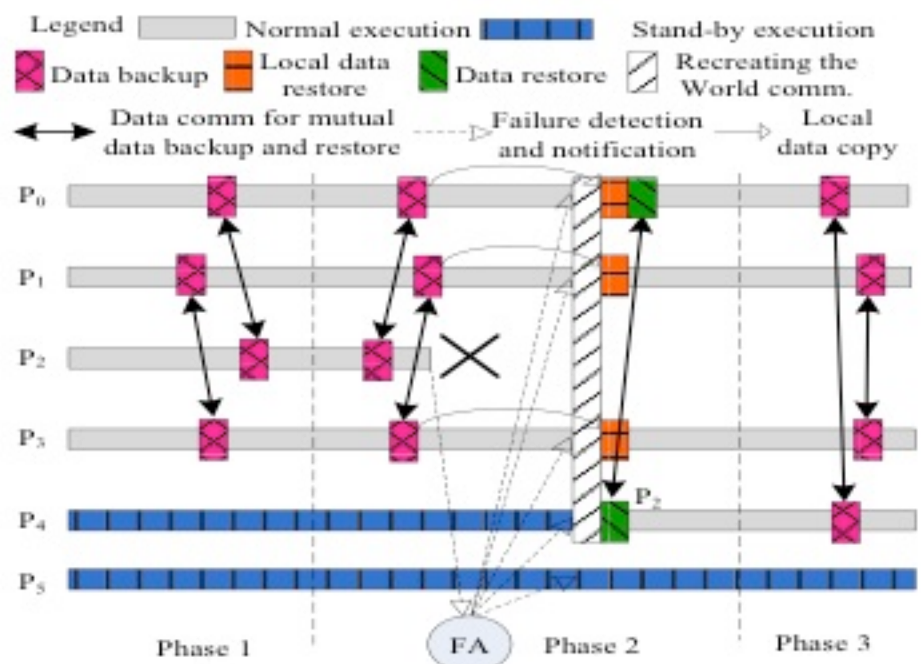




Scalable MPI

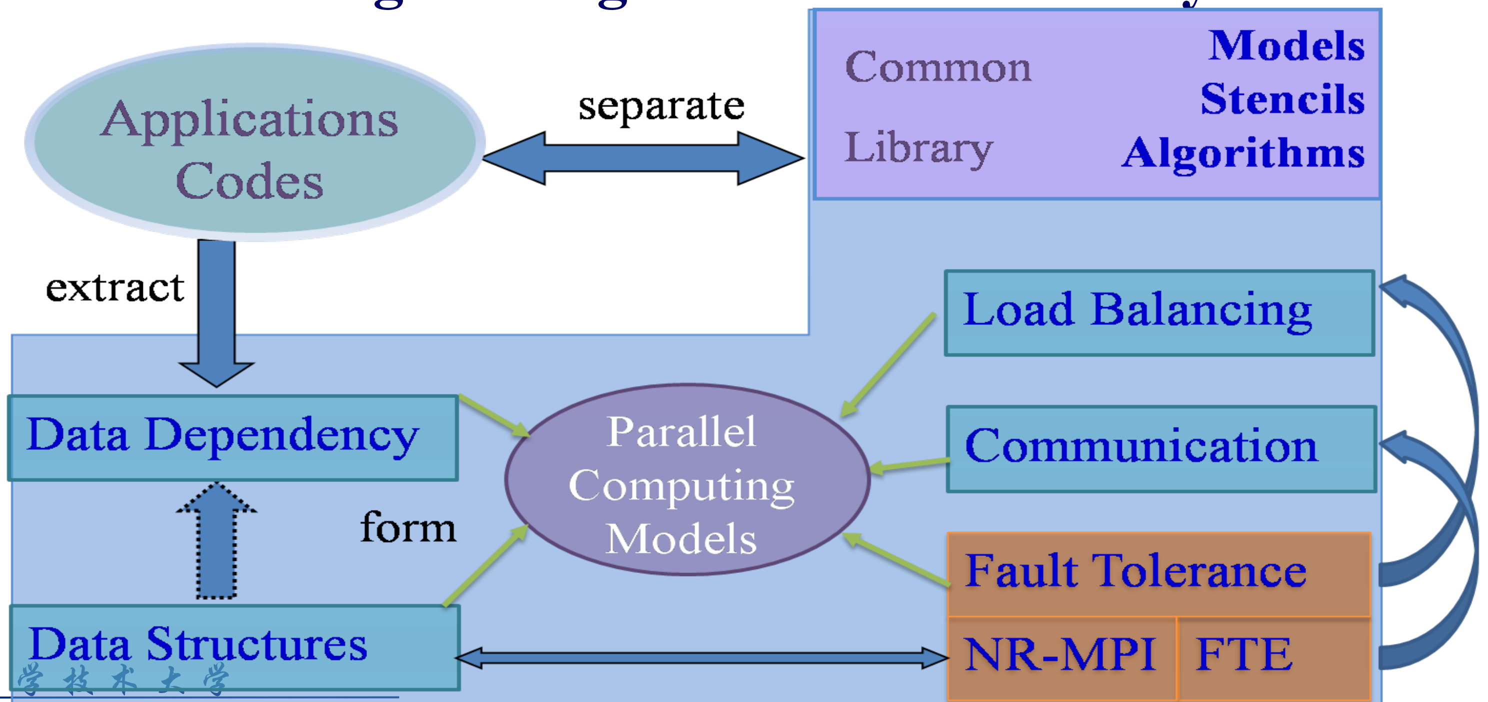
- Non-stop and fault Resilient MPI (NR-MPI)
  - Application continue execution without being relaunched
  - Failure detection and MPI state recovery done by runtime
  - Data-backup by application-level diskless C/R
  - Reconstruct of MPI communicator and channel





Domain Framework

- Hides parallel programming complexity and the hierarchy of parallel computers
- Integrates the efficient implementations of parallel fast algorithms
- Provides efficient data structures and solver libraries
- Supports software engineering for code extensibility

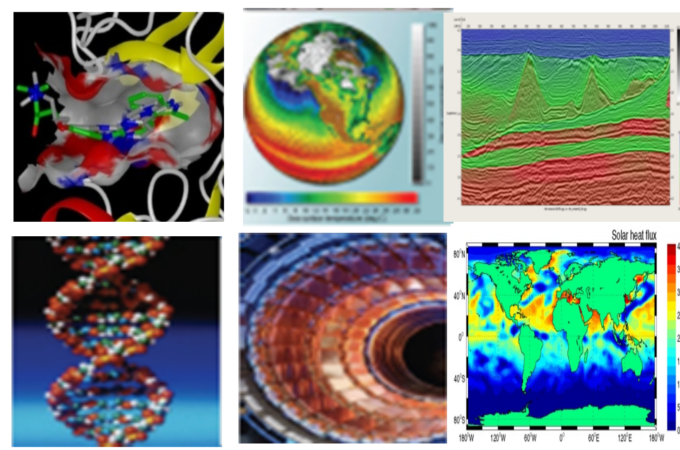


Dynamic Software

- Application Complexity： Multidisciplinary, Multi-physics, Multi-scale, Multi-method
- Legacy applications: Long term for developing, Expensive, Difficult
- Autotuning the performance
- Dynamic resources requirement and providing
- Topo-aware and Latency hiding
- Resource sharing & Hybrid runtime
- Fault tolerant and Resilience
- Rethink & Redesign the software

Scientific Discovery

- ❑ **Creative Computing Technology**
  - ➢ **Hardware, system software, algorithm, applications**
- ❑ **Creative Data Processing Technology**
  - ➢ **Data management, Analysis, Visualization**
- ❑ **Big Data come from**
  - ➢ **Experiment**
  - ➢ **Observation**
  - ➢ **Sensor network**
  - ➢ **Simulation**



- ❑ **Challenge of computing/throughput**

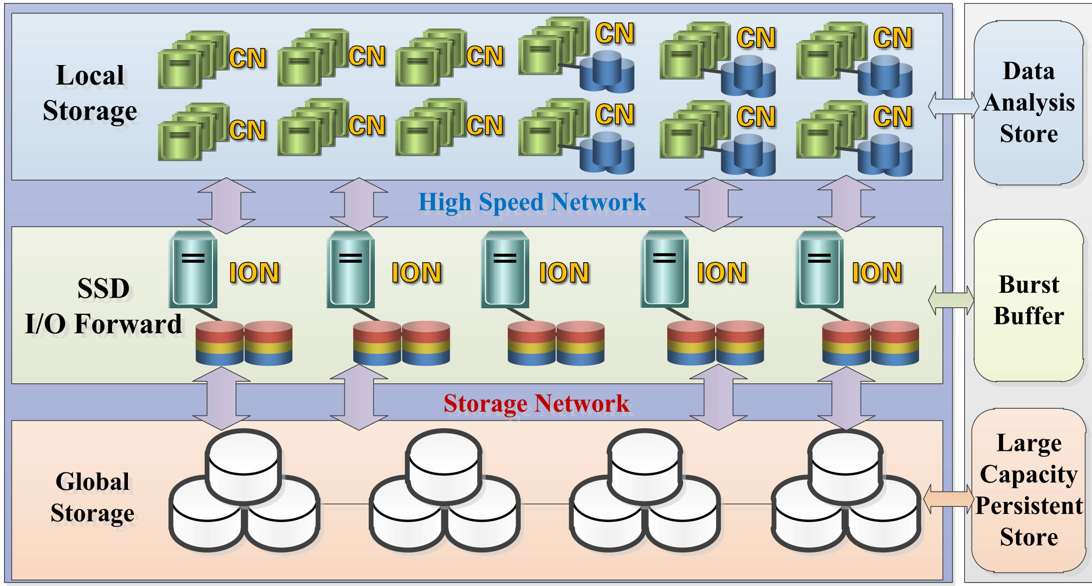
HPC Vs Big Data

- Increasing I/O requirements
  - Large scale Pre/Post data sets
  - Visualization and Analysis
  - Big science with Big data
  - Expected data volume per simulation from ~GB to ~PB, typically ~100 TB
- I/O Bottleneck
  - Scalability, Efficiency, Performance, Economic and durability
- What's needed for Parallel IO interface
  - More hints could be expressed
  - More patterns could be supported
  - Interface to application IO library

Scalable IO Structure

❐ IO Architecture on Tianhe-2

- ➢ Multiple Layers & Hybrid Storages
  - ◆ Local Disk
  - ◆ PCI-E SSD
  - ◆ Disk Array
- ➢ 6400 local Disks
  - ◆ Bus attached
- ➢ 256 IO nodes
  - ◆ Burst: above 1TB/s
  - ◆ TH-Express and IB QDR port
- ➢ 64 Storage Servers
  - ◆ Sustained： about 100GB/s



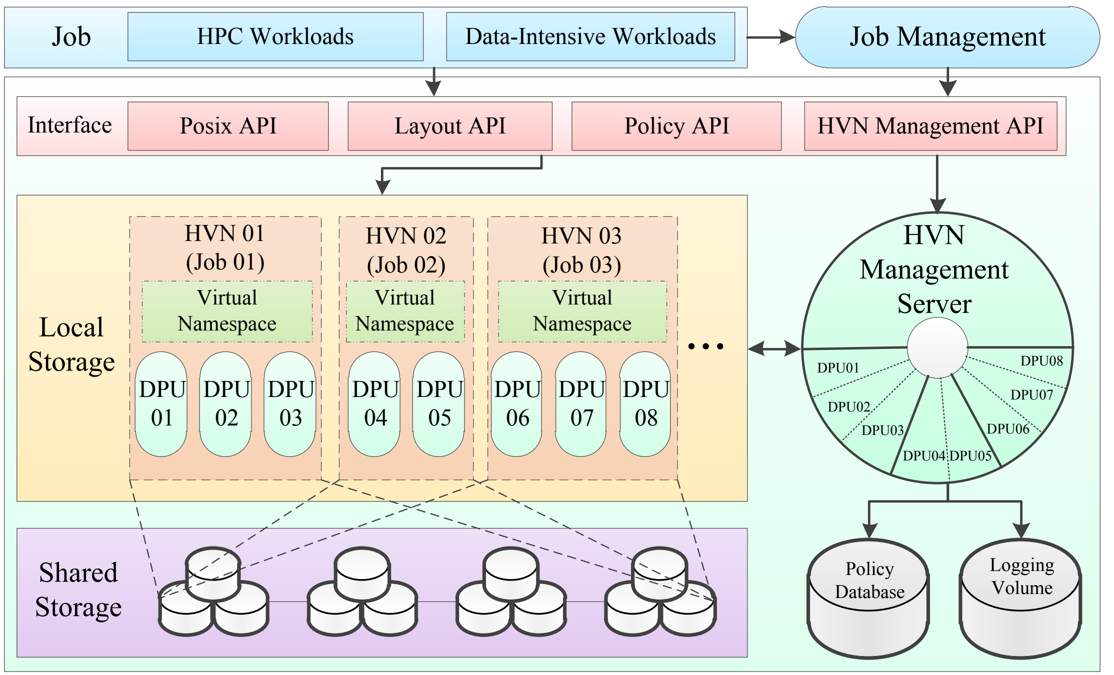
Scalable IO Structure

❐ H^2FS: Hybrid Hierarchy File System

- DPU, A fundamental unit for data processing, tightly couples a compute node with its local storage
- HVN, Hybrid, Unified and Isolated dynamic namespace maintained by centralized servers
- Layered and enriched metadata, I/O hints as high level metadata

❐ I/O API

- POSIX
- MPI-IO
- Extended API, layout and policy guide
- HDF5 over POSIX and extended API
- Object API(todo)

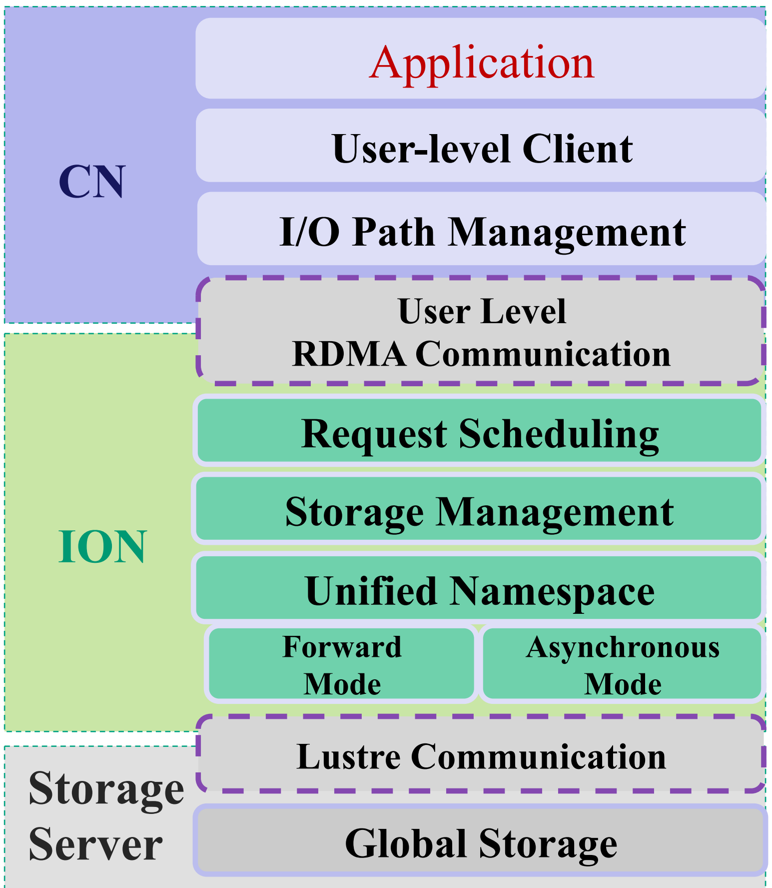


Scalable IO Structure

- Multi Modes supported in Customized HVN
  - Forward Mode
    - local storage bypassed, forward & aggregate requests
  - Burst Buffer Mode
    - Local storage attached as independent buffer for draining burst I/O, transparent data movement
  - Local Cooperation Mode
    - Local storage unified with individualize layout, DHT for unique-file, partitioned layout for shared-file, with minimum global storage involved
  - Fusion Mode
    - Local Cooperation + Global, single unification namespace of H^2IO storage, customized data moving policy

Scalable IO Structure

❐ Contributions of components in H²FS



	Performance	Scalability	Ease-to-Use	Reliability
User-level Client		√	√	
I/O Path Management				√
RDMA Communication	√	√		
Request Scheduling	√	√		
Storage Management			√	√
Unified Namespace		√	√	
Forward mode		√	√	
Asynchronous mode	√	√		

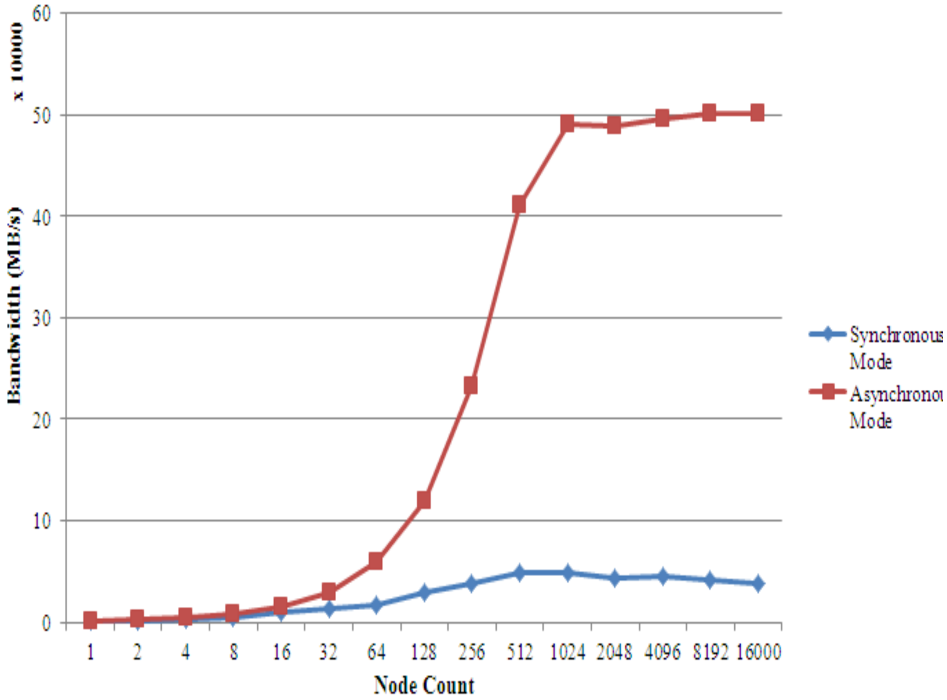
HPCL

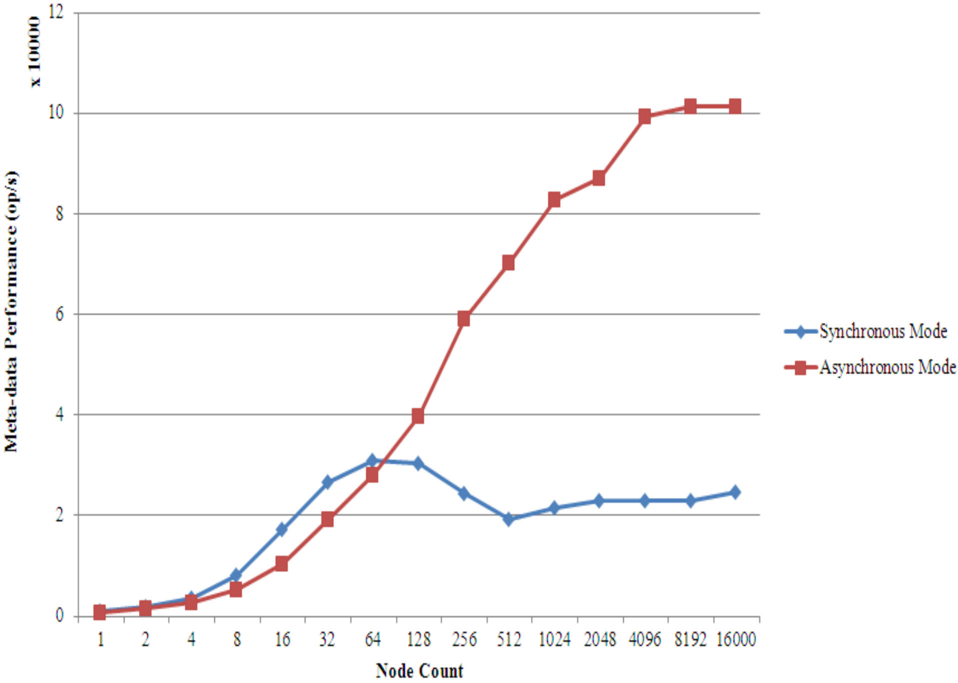
Scalable IO Structure

Tianhe 天河

❐Scalable I/O operation

- Aggregate burst Bw>500GB/s, IOR benchmark
- Aggregate metadata throughput > 100,000 op/s, mdtest



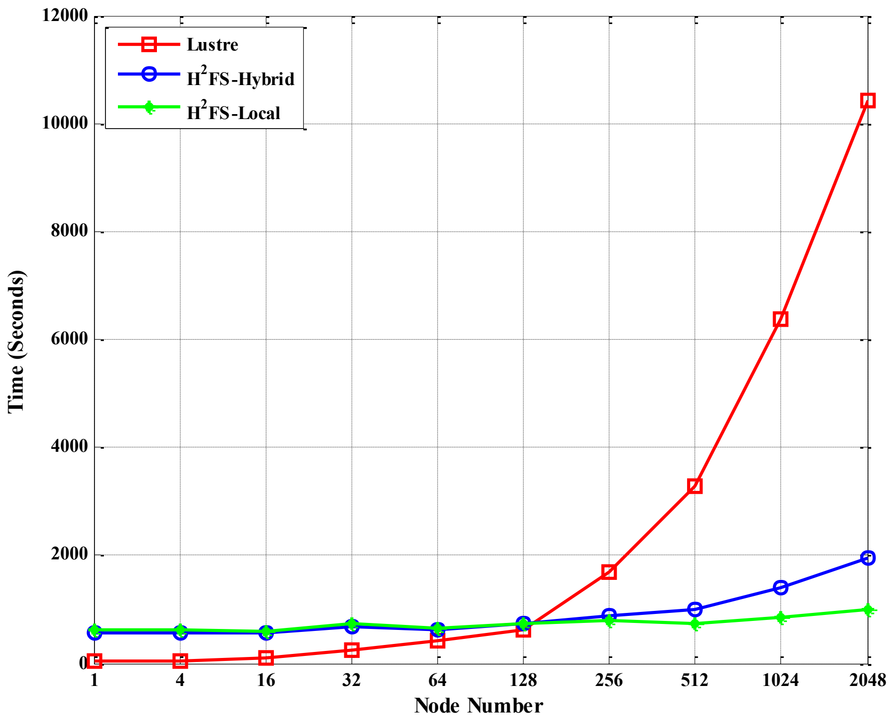


Local cooperation HVN

Scalable IO Structure

❑ Evaluation on typical HPC application

- Geoeast, seismic data processing software
- MEASTRO, MADBench2, S3D



	Non-HVN (GB/s)	HVN (GB/s)	
MEASTRO (Unique file)	7.28	37.06	5X
MADBench2 (Unique file)	30.43	94.92	3.1X
S3D (shared file)	5.33	14.82	2.78X

HVN is flexible, more work todo

Scalable IO Structure

❐HPC benefits

- ➢ Scalable burst BW for typical HPC application
- ➢ Isolated HVN makes data intensive application individualize their optimization
- ➢ Reduced requirements for costly shared storage
- ➢ Scalability, Efficiency, Economic and Ease of use

❐Data processing benefits

- ➢ Maximum locality, DPU provides opportunity to schedule tasks close to data
- ➢ Single namespace make post-processing easy
- ➢ Reduction of data movement, better support for in-situ data analysis and data in-transit analysis

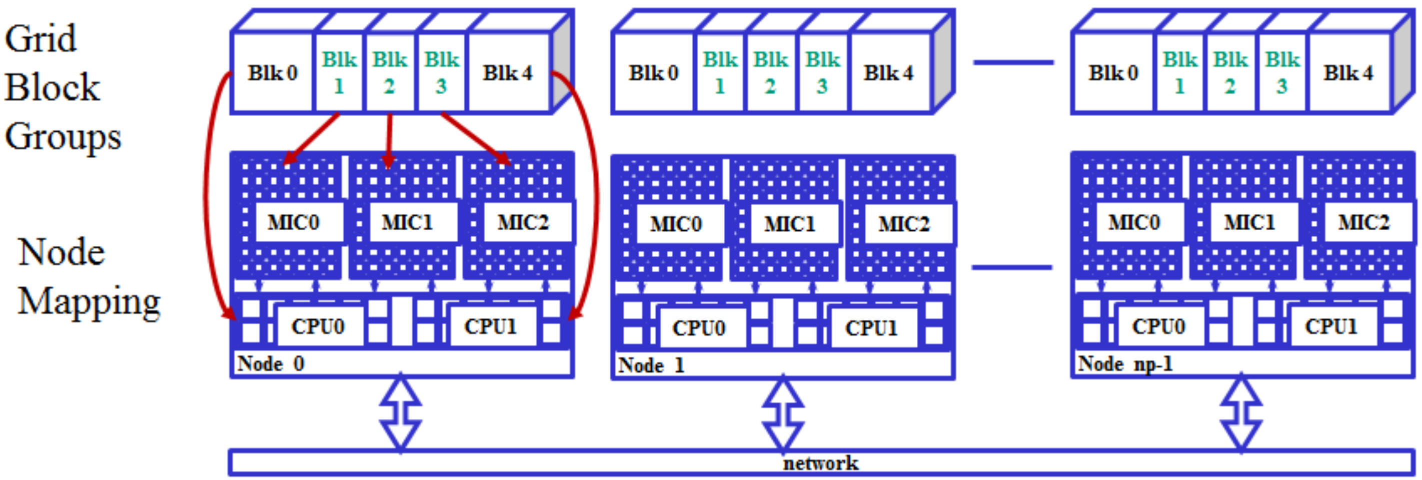
Different Levels of Performance

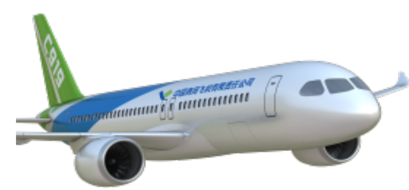
- Peak performance
- LINPACK performance
  - Avg. 80%
- Gordon Bell Prize performance
  - ~30%
- Application sustained performance
  - <5%~10%
- HPCG Benchmark
  - ~1%

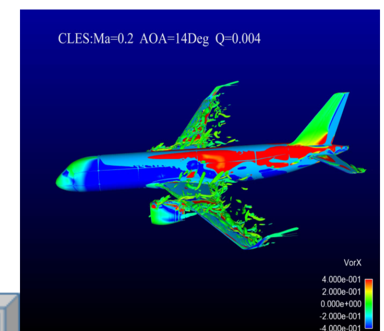
Scalable Applications

HCFD: High-Order SimulaTor of Aerodynamics

- **WCNS- Weighted Compact Nonlinear Scheme**
- **Explicit Runge-Kutta**





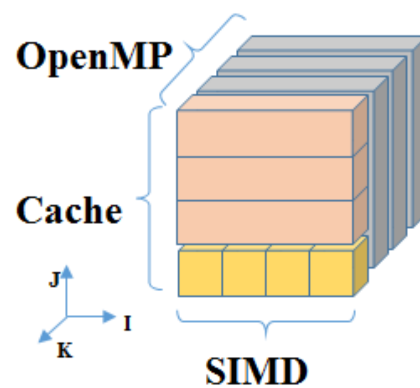


MPI: Grid Block groups

Offload: In a grid block group（CPU+MIC）

OpenMP: In one grid block

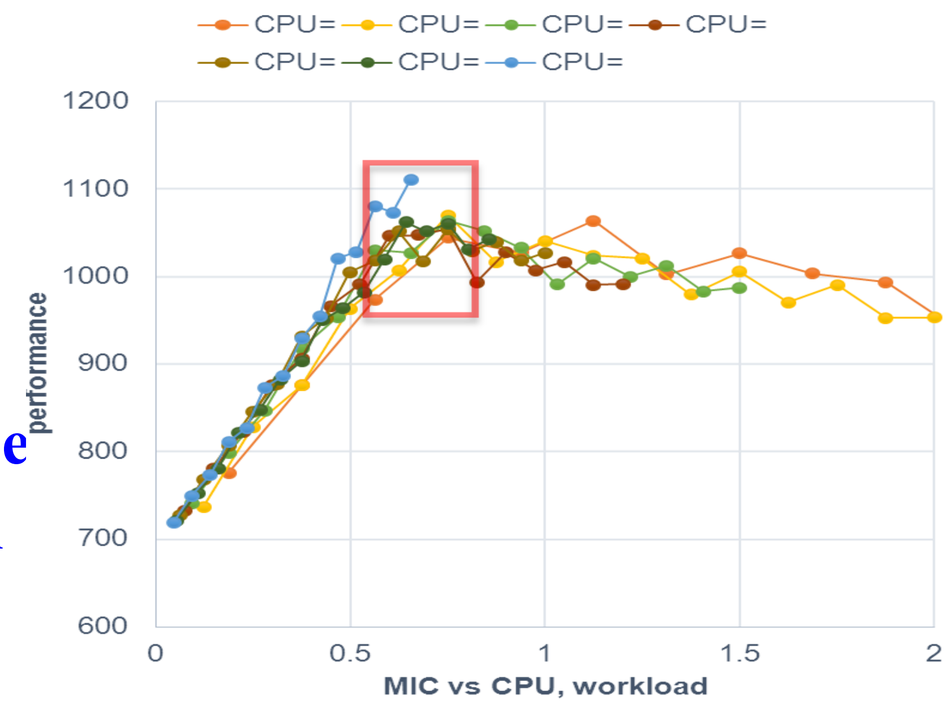
SIMD: slice of one block



Scalable Applications

HCFD: High-Order SimulaTor of Aerodynamics

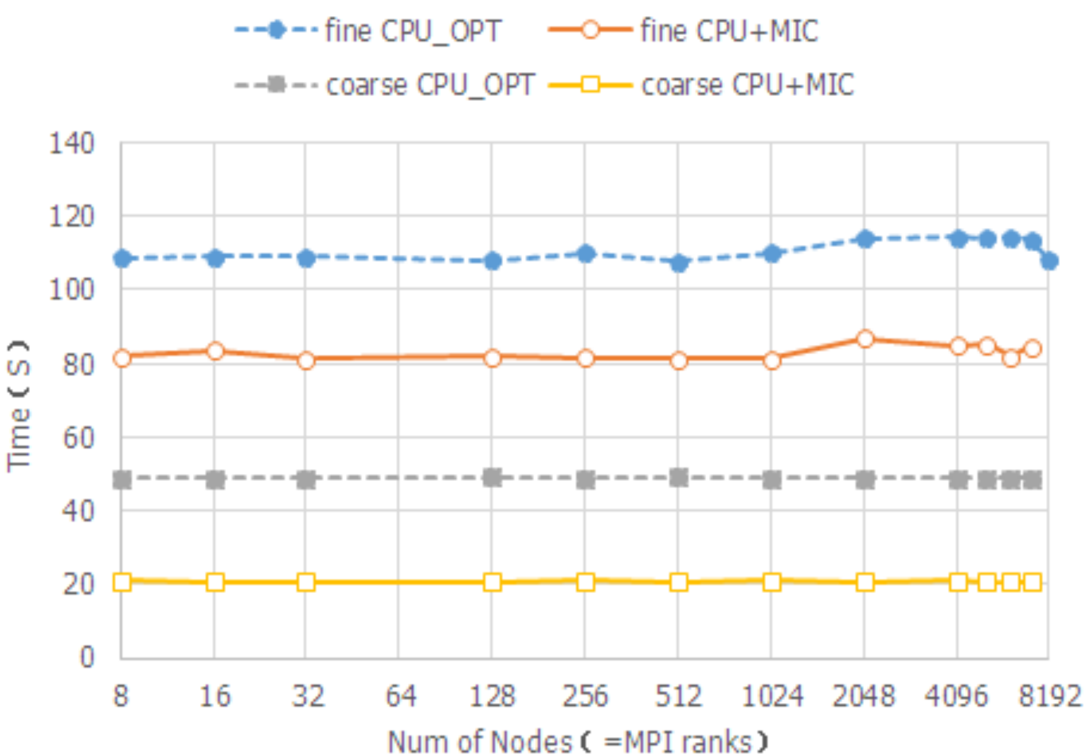
- Balanced partition between CPU/MIC inside each node
  - MIC: CPU 0.6~0.8
- Hierarchical data partition & communication
- Overlap the communication and computation using pipeline
- Memory & cache optimization
- Offload mem reuse
- Exploit SIMD

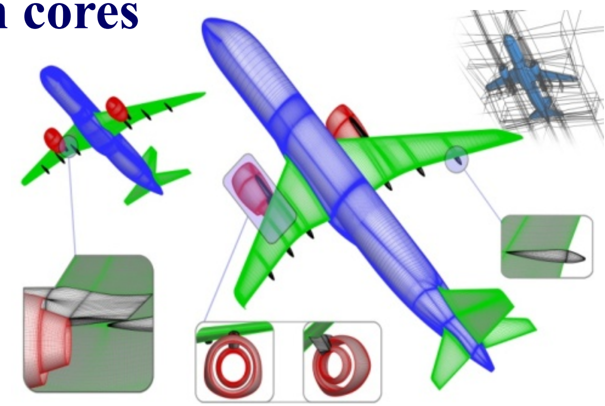


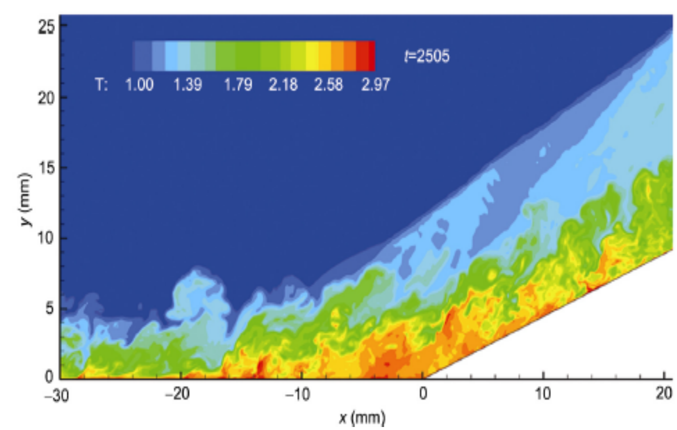
Scalable Applications

HCFD: High-Order SimulaTor of Aerodynamics

- **CPU+MIC**
  - **7168nodes with 3mics/node，1.376million cores**
  - **Gird 682.4 Billion**







Applications

Tianhe
天河

CARDC 中国空气动力研究与发展中心 China Aerodynamics Research and Development Center

ISCAS

UNIVERSITAS AMOIENSIS

HUNAN UNIVERSITY 976·1926

SOUTH CHINA UNIVERSITY OF TECHNOLOGY 1934

HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY HUST

SIAT CHINESE ACADEMY OF SCIENCES

HHEE

COMAC 中国商飞

The University of Hong Kong

PEKING UNIVERSITY 1898

SHANGHAI JIAO TONG UNIVERSITY 1896

TSINGHUA UNIVERSITY 1911

NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY 1953

SIMM

UNIVERSITY OF SCIENCE AND TECHNOLOGY BEIJING 1952

AVIC

BEIHANG UNIVERSITY 1952

SUN YAT-SEN UNIVERSITY 1924

XIDIAN UNIVERSITY 1931

COMMUNICATION UNIVERSITY OF CHINA

Institute of Mechanics CAS ~1956~

NORTHWEST UNIVERSITY XI'AN CHINA 1902

BEIJING NORMAL UNIVERSITY 1902

SHANDONG UNIVERSITY 1901

CENTRAL SOUTH UNIVERSITY

西北大学

inspur 浪潮

华大基因 BGI

HUAZHONG AGRICULTURAL UNIVERSITY 1898

SCHOOL OF PHARMACEUTICAL SCIENCES

视骏科技 STRONGENE

CSSC 广州船舶及海洋工程设计研究院

中国科学院 CHINESE ACADEMY OF SCIENCES

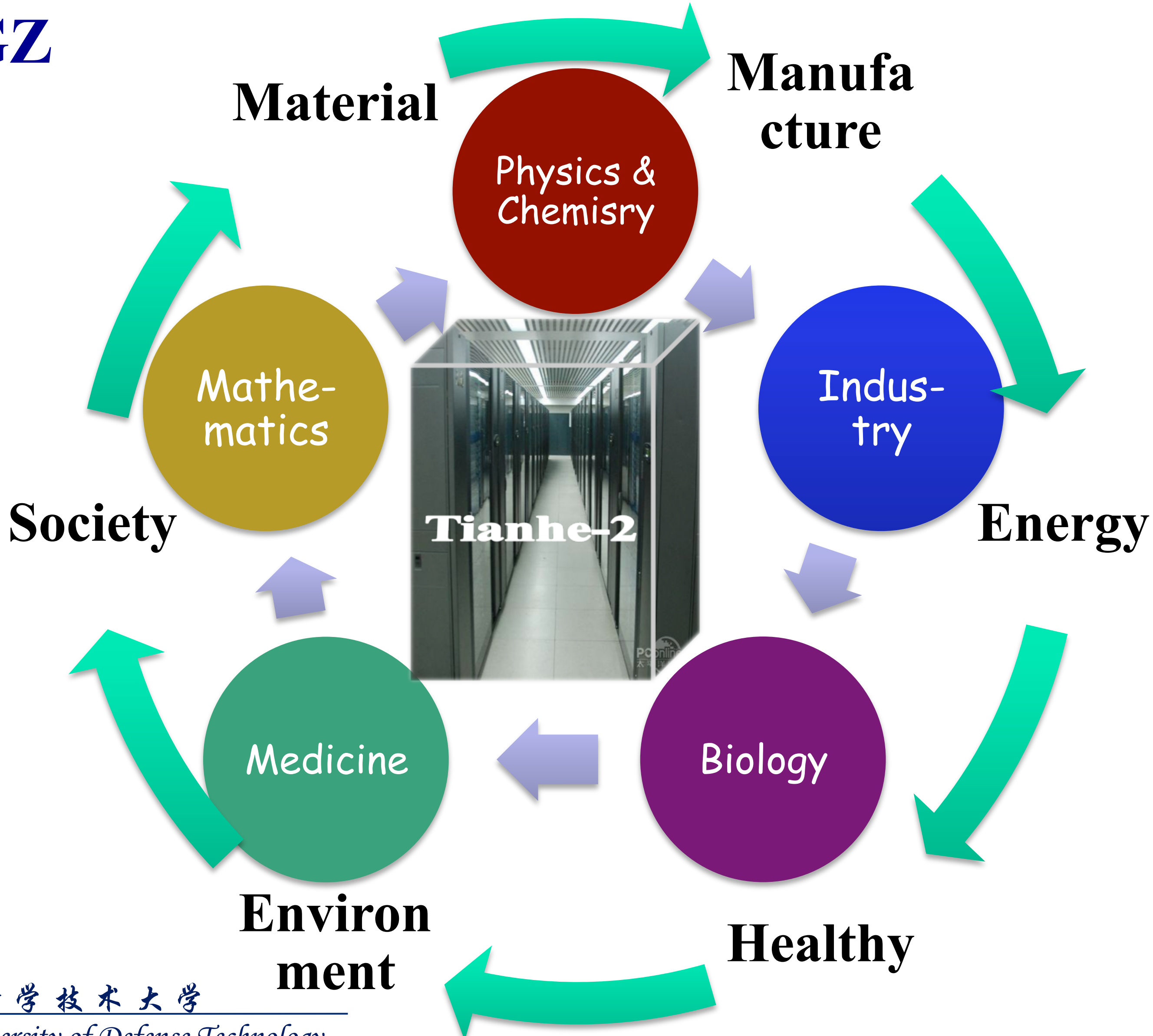
清华大学 Tsinghua University

北京大学 PEKING UNIVERSITY

Applications

Tianhe
天河

NSCC-GZ



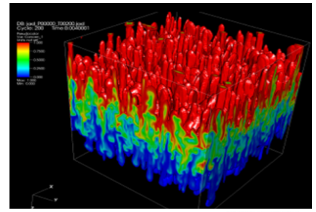
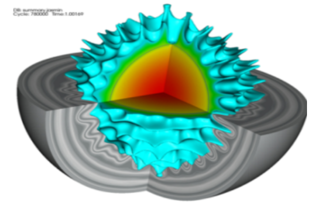
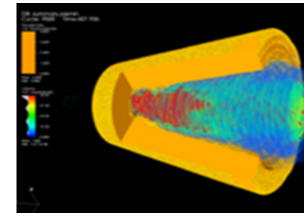
国防科学技术大学
National University of Defense Technology

HPCL

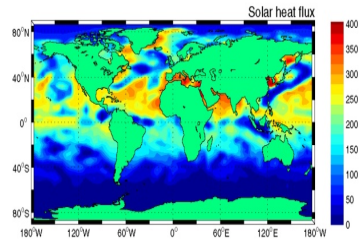
Tianhe
天河

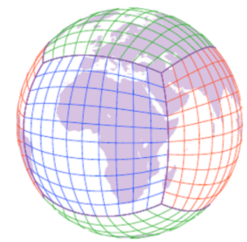
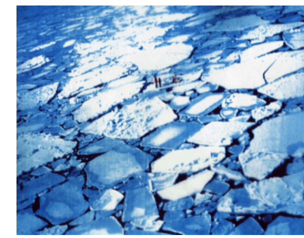
Applications

- High Energy Density Physics

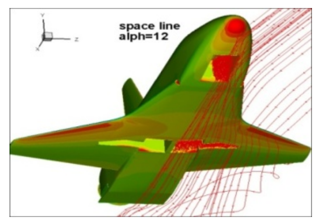




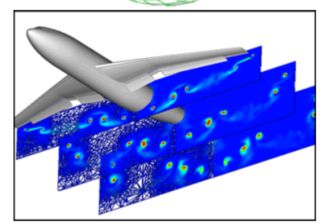
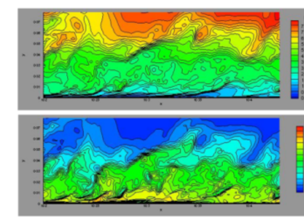
- Weather & Climate



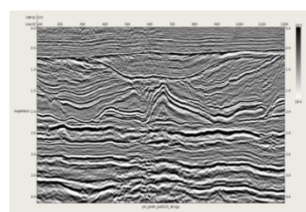
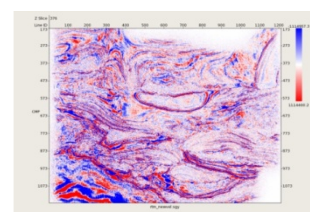
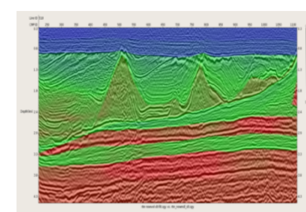



- CFD

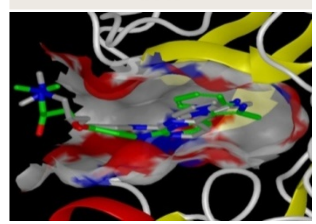
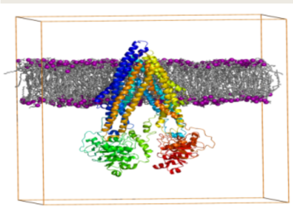
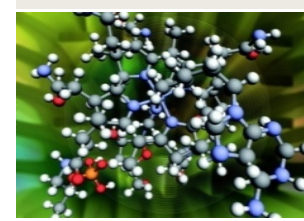


- Seismic data processing

- Bio-information

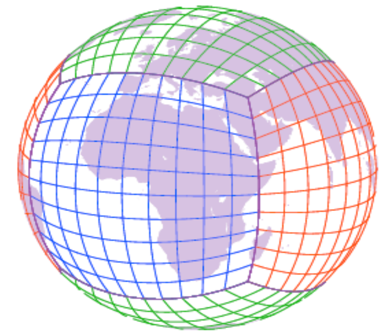




- E-Gov & Service

Applications

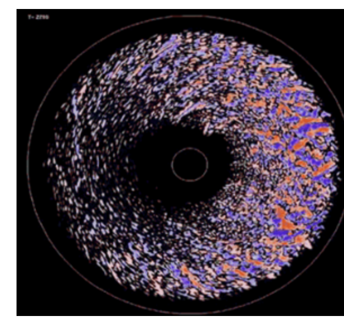
❐ Climate

- Global shallow water model， #8664， ~1.7million cores, 77%



❐ Physics

- Gyrokinetic Toroidal Code GTC， #2048， ~160,000 cores

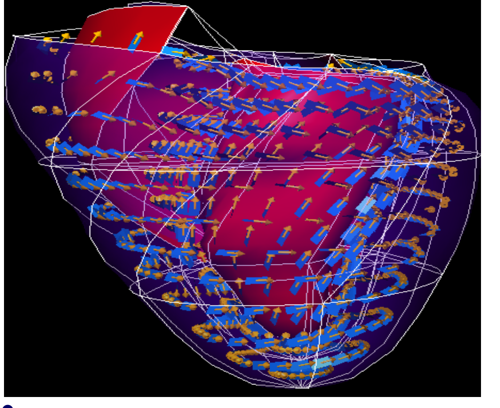


❐ Business Opinion Analysis

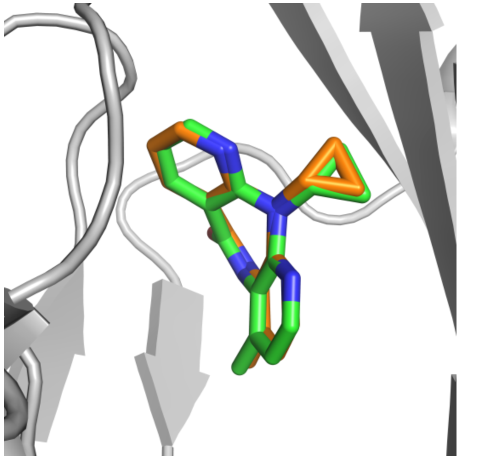
- 600TB structured/non structured data with micMR (Hadoop over MIC)， #1024， 100Million Rec/day

Applications

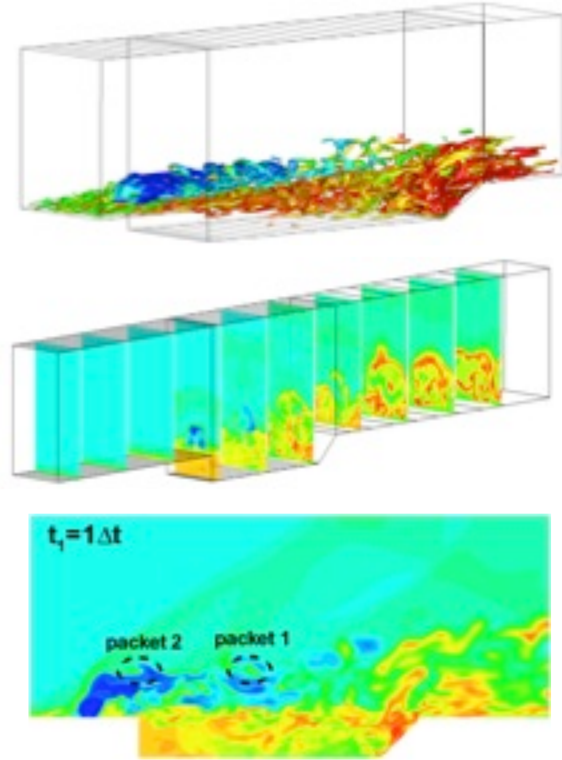
- **Cardiac subcellular level nanoscale calcium-spread mechanical simulation**
  - **Explore the pathogenesis of heart disease**
  - **4096nodes with mic, 1.27PF**



- **Virtual drug screening - molecular docking calculations**
  - **DOCK6.5**
  - **303,826 compounds conformation(specs)**
  - **1,100 drug target（pdtd）**
  - **Over 334 million docking calculation**

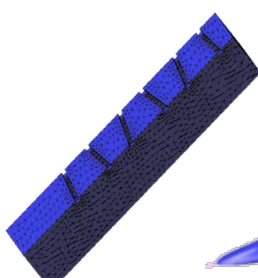


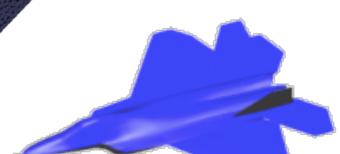
Applications

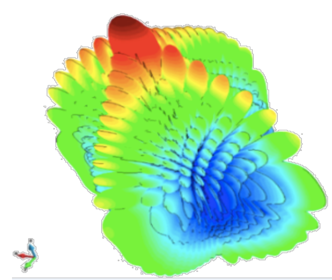


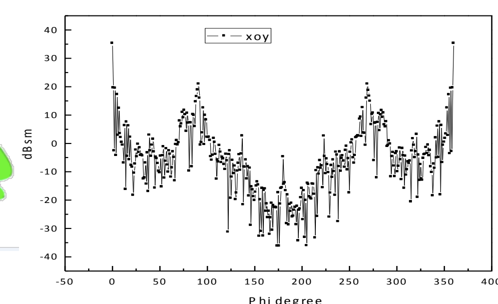
- ❒ **Combustion flow in the turbulent**
  - ➢ Stability and flame propagation mechanism, combustion oscillation mechanism

- ❒ **Fast simulation of complex electromagnetic environment**
  - ➢ FDTD
  - ➢ MOM
  - ➢ PO



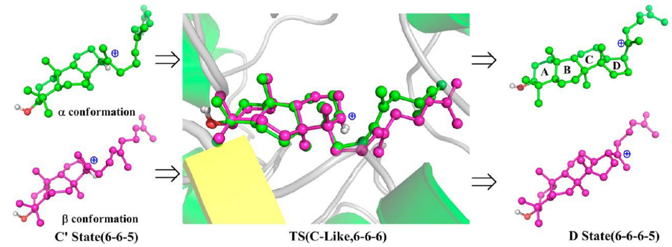




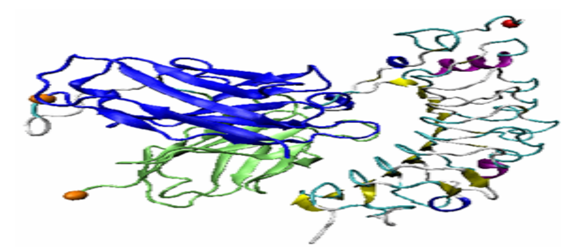


Applications

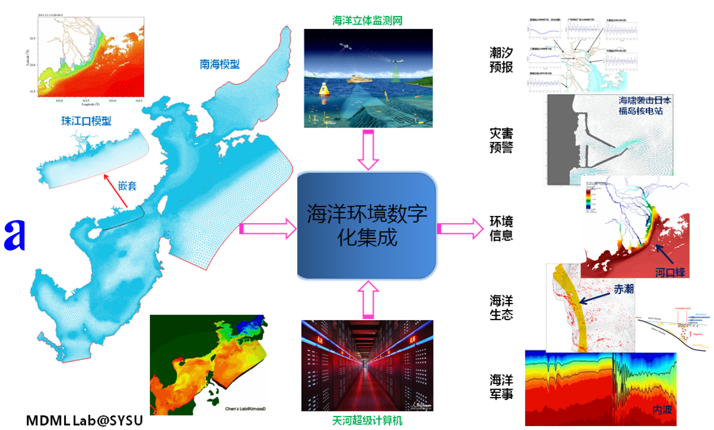
- **The Catalytic Mechanism of Human Oxidosqualene Cyclase**
  - **QM/MM MD simulation (Qchem-Tinker)**



- **Study the pathogenesis of Flavobacterium**
  - **Research and product development of the key technology in freshwater fish immune disease prevention and control**

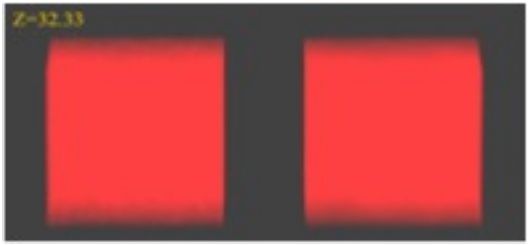


- **Regional Marine digitizing system**
  - **Pearl River Estuary South China Sea**



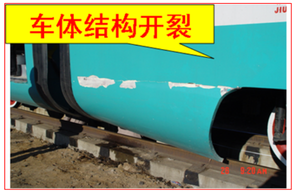
Applications

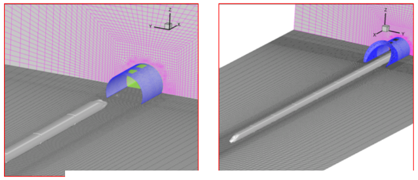
- ❑ **Neutrino Mass Measurement**
  - ➢ Simulate 13.7-billion-years cosmic evolution



Massless Neutrino   Massive Neutrino

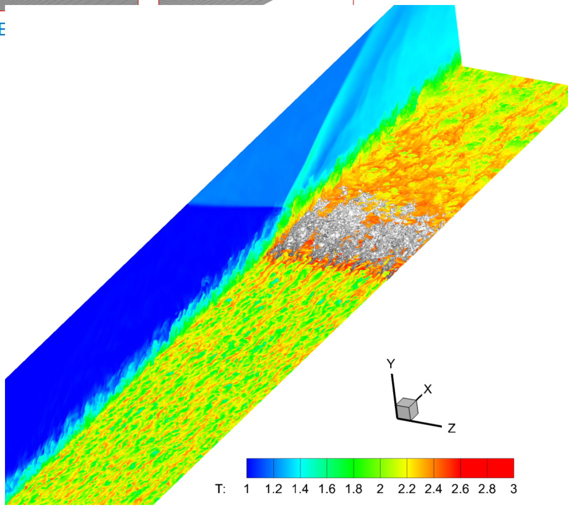
- ❑ **High-speed rail tunnel aerodynamic effects**
  - ➢





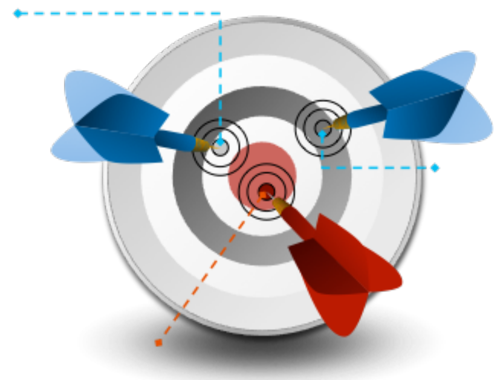
网格数 > 1500万

- ❑ **Shock Wave/Turbulent Boundary Layer Interaction**
  - ➢ Structural safety of the high-speed aircraft

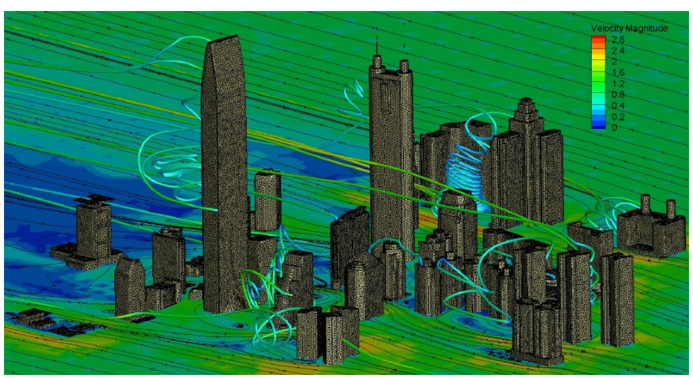


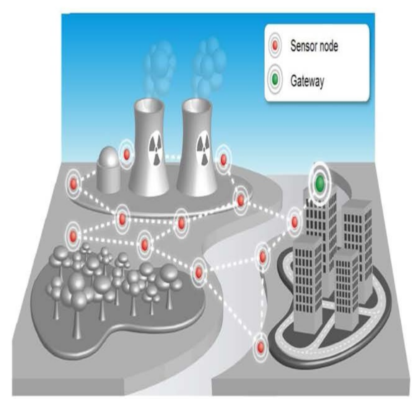
Applications

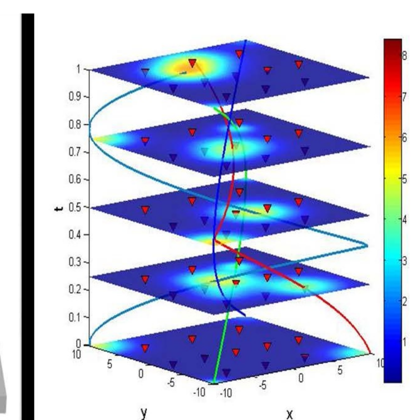
- **Real-time financial market risk quantification computing**



- **Sources of air pollution in city**
  - **Pollutant concentration distribution and temporal trace**







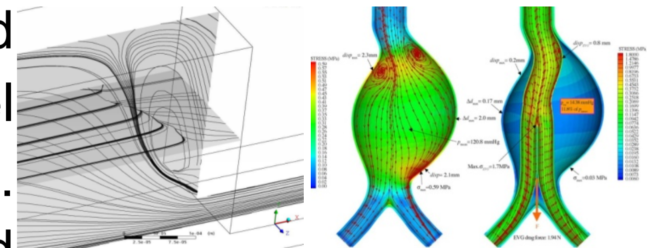
Applications

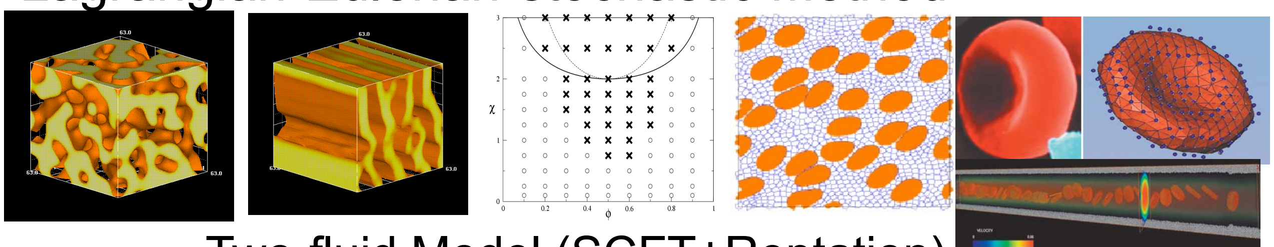
Multi-Scale numerical simulation framework

Length

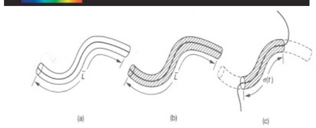
10^{-3} m

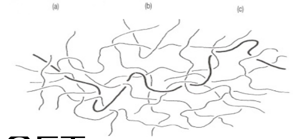
Immersed Boundary Method
Two-fluid Model
Microscopic kinetic models + LBM.
Lagrangian-Eulerian-stochastic Method



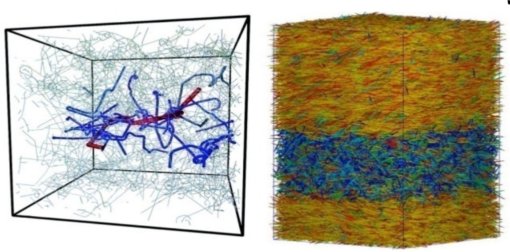


Two-fluid Model (SCFT+Reptation)
Stochastic Entanglement Dynamics
Lattice Boltzmann Methods (LBM)
Smooth Particle Hydrodynamics (SPH)





10^{-9} m



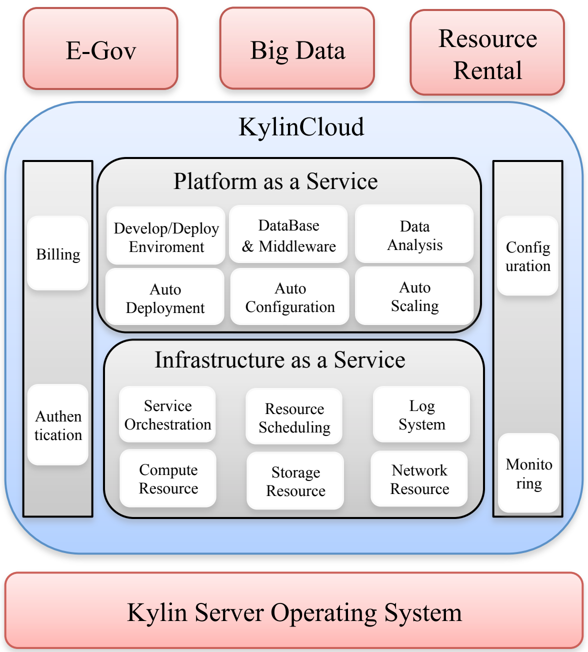
Theoretical approach: SCFT
Kinetics of signalling and metabolic pathways
Coarse-grained Monte Carlo (MC) and Molecular Dynamics (MD) and Non-equilibrium MD
Car-Parrinello MD, Quantum MC

10^{-9} sec — Time — 10^{3} sec

Applications

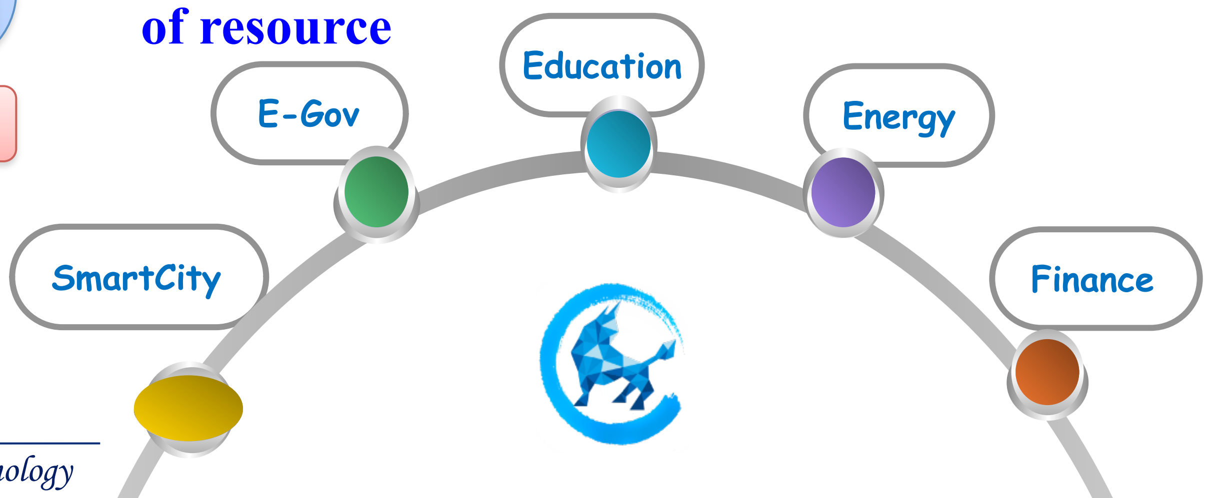
KylinCloud Cloud Platform

❒ Architecture



❒ Features

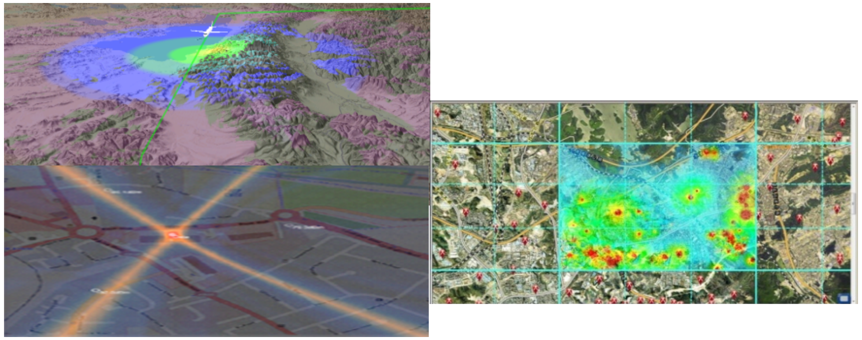
- ➢ **Customized according to the need of various applications and the arch. of TH-2**
- ➢ **Provide IaaS and PaaS services to applications with efficient resource management and scheduling mechanisms**
- ➢ **Provide multiple-level user management and quota management to tenants**
- ➢ **Provide friendly self-service portal and the statistics, reporting and displaying of the usage of resource**



Applications

❑ Applications

- ➢ E-Gov
- ➢ RenderCloud
- ➢ micMR
- ➢ Video Processing
- ➢ Electromagnetic Spectrum Management



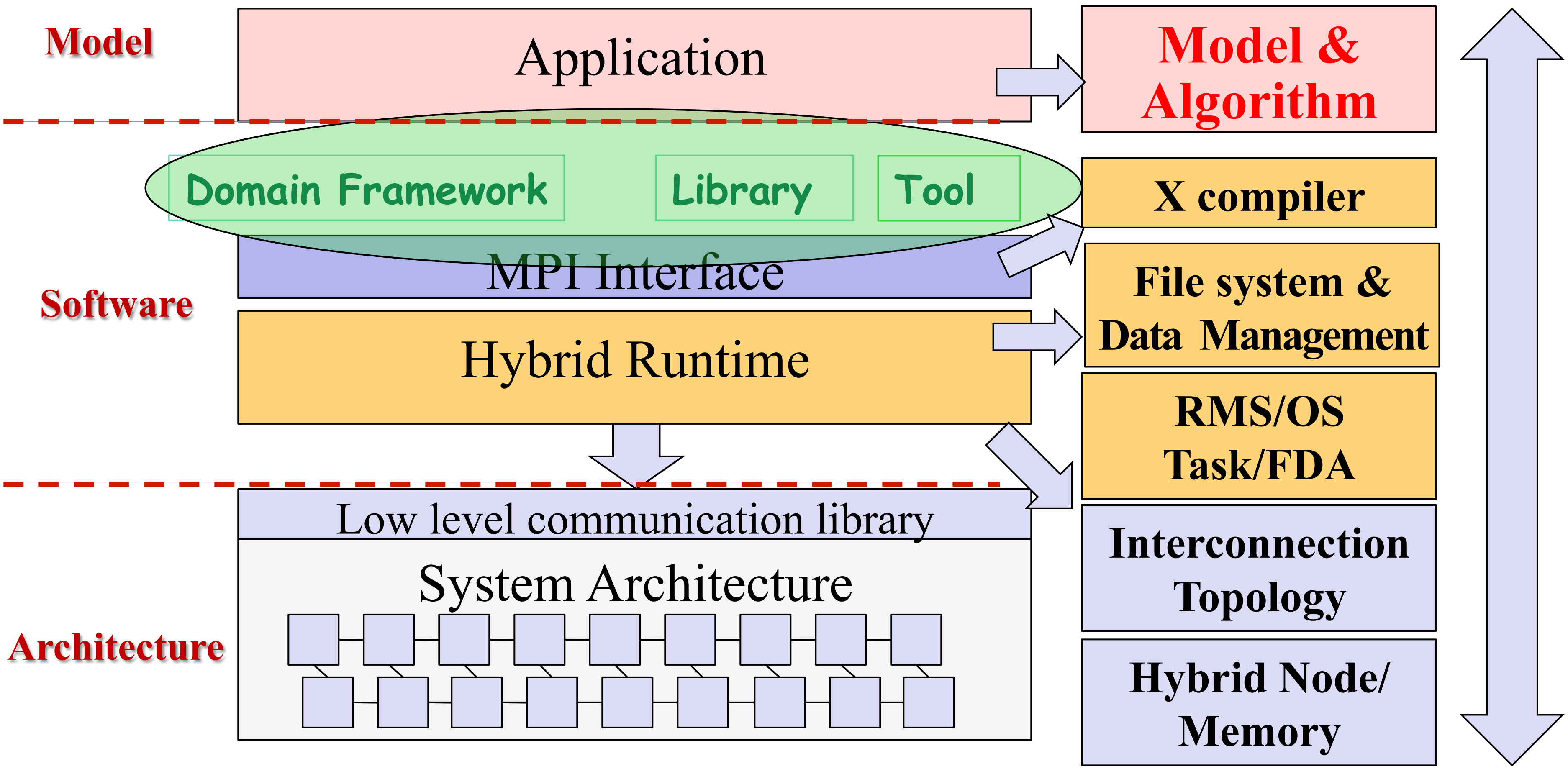
Applications

- Need custom hybrid algorithms
  - Performance-oriented programming
  - Communication reduction
  - Architecture aware algorithm
  - Dynamic management of resources at all levels
  - Fault Resilient and Oblivious
  - Rethinking heterogeneous new algorithms at the physics model to maximize the performance
- Application Code
  - Scalability， Maintainable
  - Portable, Productivity

HPCL

Co-design for Scalable System

Tianhe 天河



HPCL

Summary

Tianhe 天河

- Use the existing systems better
- Many-core will be the main trend for next generation system
- Interconnection communication is critical
- Hybrid hierarchy IO structure
- System designers and application designers should share the burden of Scalability
- Domain-specific application framework may be helpful
- International collaboration is important

HPCL

Thanks

Tianhe
天河