Russian Supercomputing Days

Moscow, september 28-29

Russian Supercomputing Days is the first united international supercomputing conference carried on the best traditions of Russian and international supercomputing events.

Overview of Tianhe2 System and Application

Yutong Lu

Professor

School of computer science, NUDT &

State Key Laboratory of High Performance Computing, China

国防科学技术大学

National University of Defense Technology

HPCL

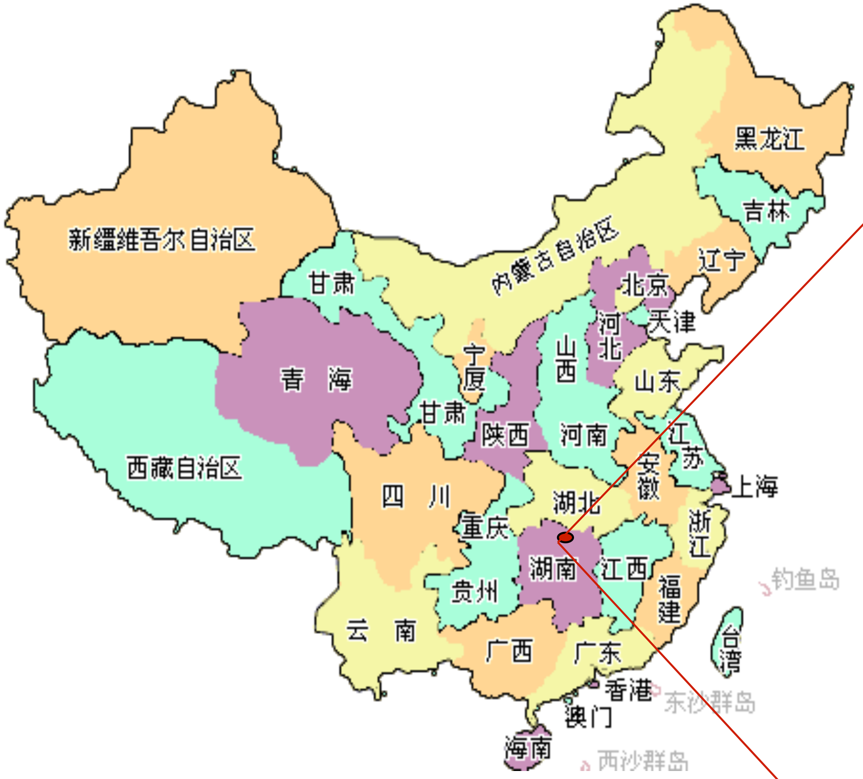
Outline

- □NUDT HPC Background
- □Design and APP of Tianhe2 System
- □Prospect of Tianhe-2A
- □Summary

Overview

天河

National university of defense technology



~2,000 Teachers

~15,000 Students

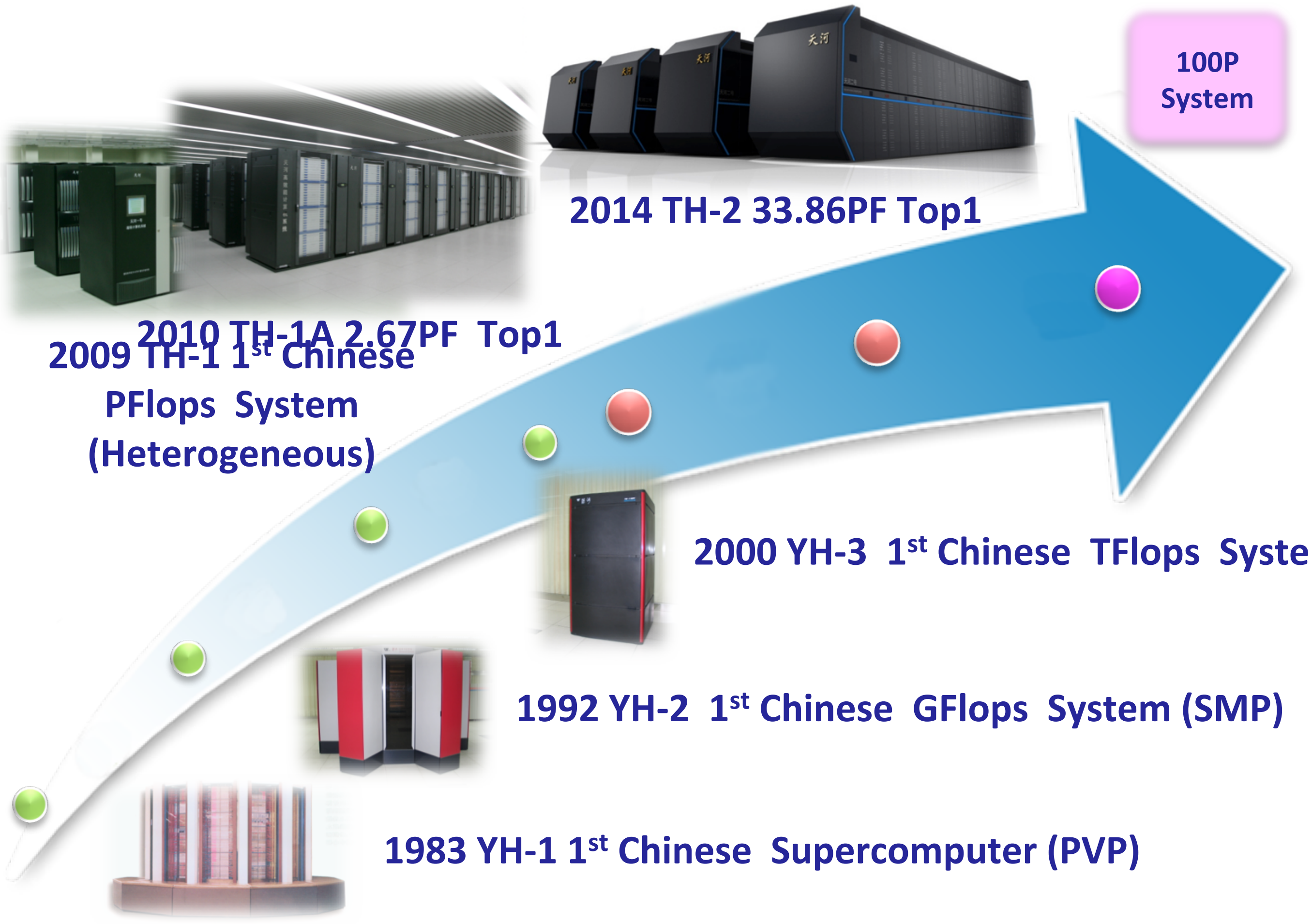
… Others

国 防 科 学 技 术 大 学

National University of Defense Technology

Overview

Supercomputers in NUDT, Changsha, China



100P System

2014 TH-2 33.86PF Top1

2010 TH-1A 2.67PF Top1

2009 TH-1 1st Chinese PFlops System (Heterogeneous)

2000 YH-3 1st Chinese TFlops System (MPP)

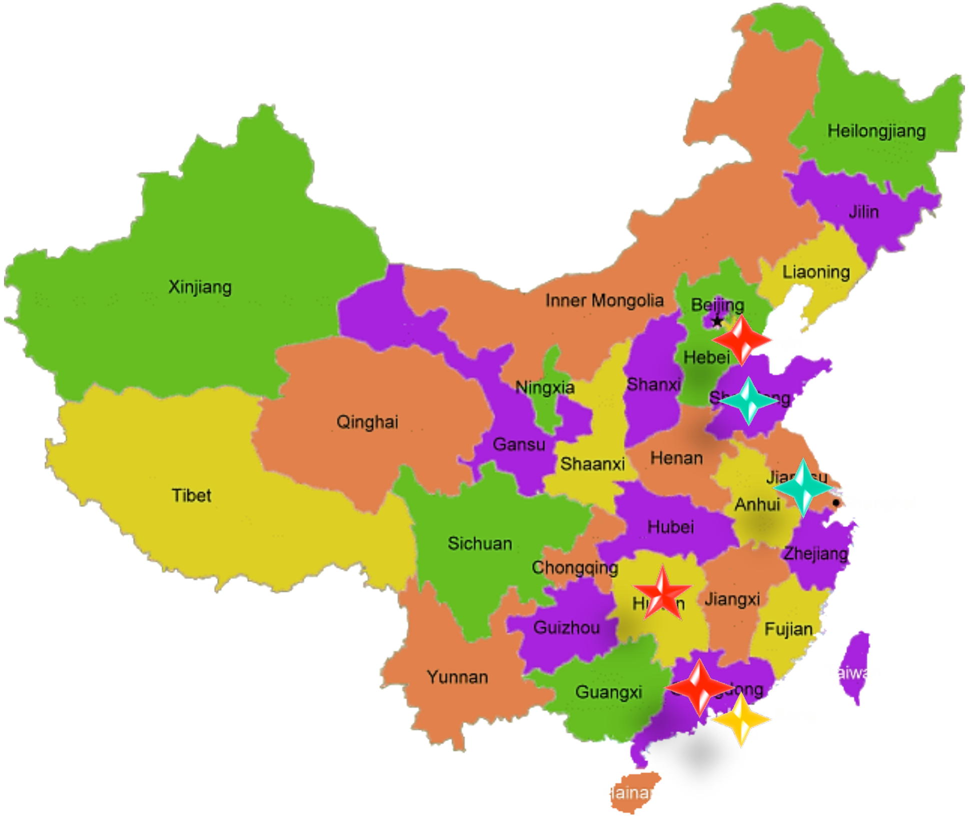
1992 YH-2 1st Chinese GFlops System (SMP)

1983 YH-1 1st Chinese Supercomputer (PVP)

Overview

天河

Supercomputing Centers in China



NSCC-Guangzhou,2013
Tianhe-2

NSCC-Wuxi
Shenwei-NG

NSCC-Changsha,2012
Tianhe-1A

NSCC-Jinan,2012
Shenwei-Bluelight

NSCC-Tianjin,2010
Tianhe-1A

NSCC-Shenzhen,2011
Dawning-6000

Overview

天河

Challenges(PSPRD)

- ➢ Performance
- ➢ Scalability
- ➢ Power consumption
- ➢ Reliability
- ➢ big Data

❒ Heterogeneous Trend

- ➢ Some of top-level supercomputers
- ➢ Tianhe-1/2, Titan …

❒ Compute Efficiency

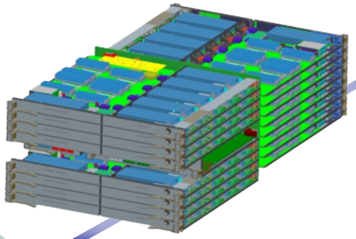
- ➢ More computations per transistor
- ➢ More computations per joule

Highlights of Tianhe-2

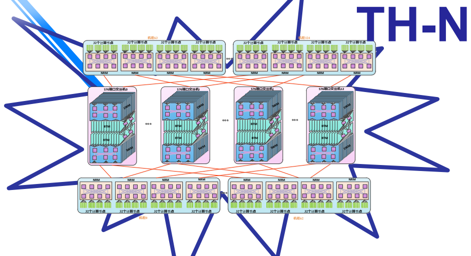
Perf	54.9PFlops / 33.86PFlops
Nodes	16000
Mem	1.4PB
Racks	125+8+13+24=170 (720m^2)
Power	17.8 MW (1902MFlops/W)
Cool	Close-coupled chilled water cooling

TH-2
(125 x Rack)

Rack
(8 x Frame)

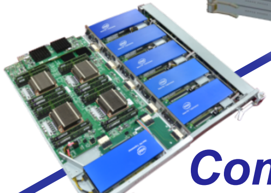


Frame
(8 x board)

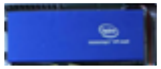


TH-Net

APM

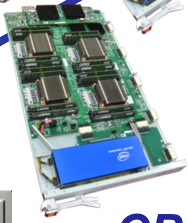


Compute board



Phi
#48000

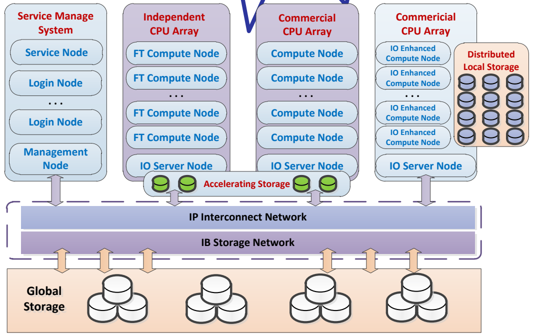
IVB
#32000



CPM

FT-1500
#4096

ION

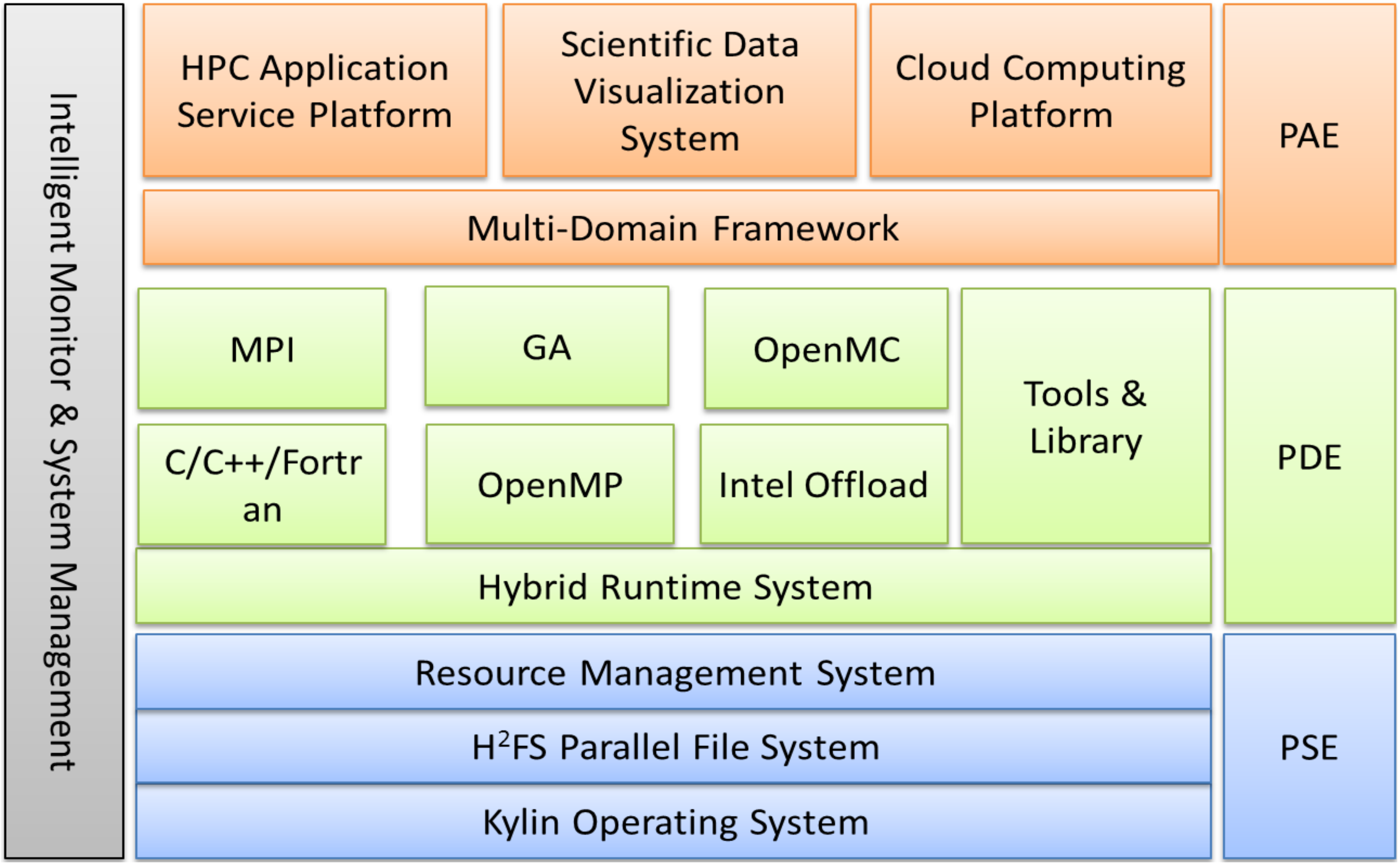


Hybrid Hierarchy shared storage System
12.4PB

Highlights of Tianhe-2

天河

❐Software Stack


Intelligent Monitor & System Management
HPC Application Service Platform
Scientific Data Visualization System
Cloud Computing Platform
PAE
Multi-Domain Framework
MPI
GA
OpenMC
Tools & Library
PDE
C/C++/Fortr an
OpenMP
Intel Offload
Hybrid Runtime System
Resource Management System
H²FS Parallel File System
Kylin Operating System
PSE

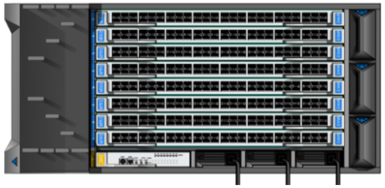
Adaptive Interconnect

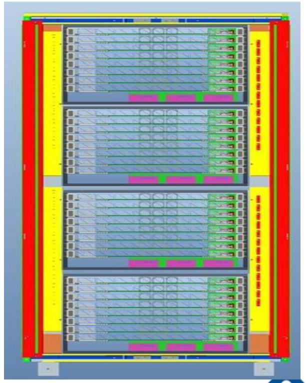
TH-Express 2+

- ❐ Network Interface Chip (NIC)
  - Bandwidth
    - ◆ 14Gbps x 8lane
    - ◆ 12GB/s
  - Function extension
    - ◆ Low latency and high rate operations
    - ◆ Collective comm offload
    - ◆ Parallel processing using Multiple DMA engines
    - ◆ Enhanced reliability

- ❐ Network Router Chip (NRC)
  - 24 ports
    - ◆ Up to 5.376 Tb/s of switching capacity
  - Table-based routing
    - ◆ Multi-path adaptive routing
    - ◆ Oblivious routing

Extending to >100PFlops system

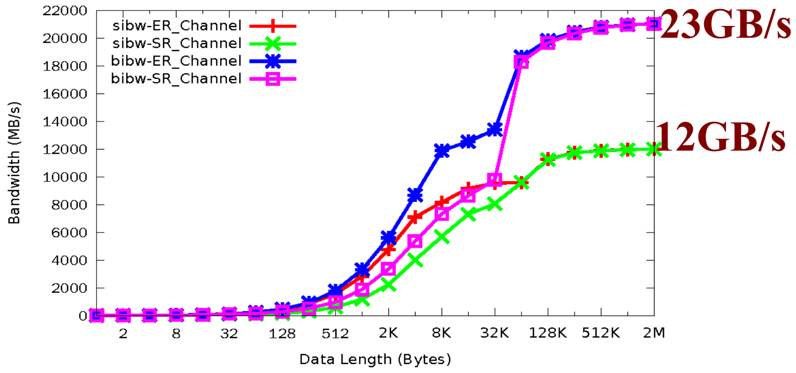


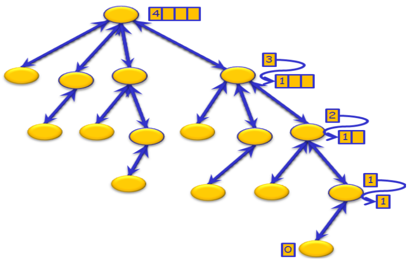


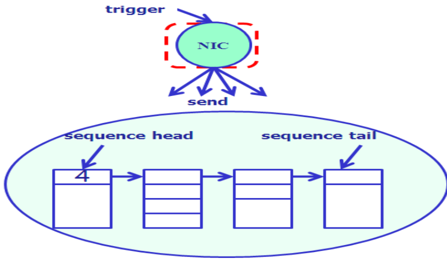
Scalable MPI

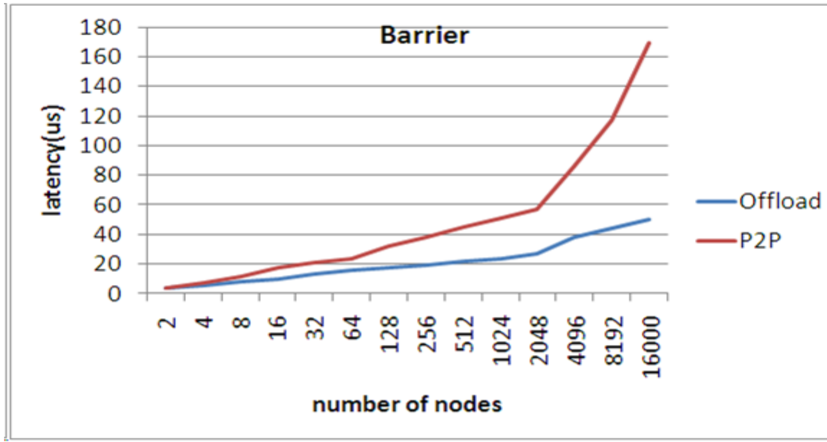
Message passing system based on TH-Express

- ❐ Multiple Communication Protocols
  - ➢ Performance oriented
  - ➢ Scalability oriented
  - ➢ Combine application model
  - ➢ Zero-copy data transfer
- ❐ Collective Operation Offload
  - ➢ Construct topology-aware algorithm tree dynamically
  - ➢ Message pass automatically based on the trigger of NIC
  - ➢ Bypass effect of OS noise, Reduce Latency of large scale



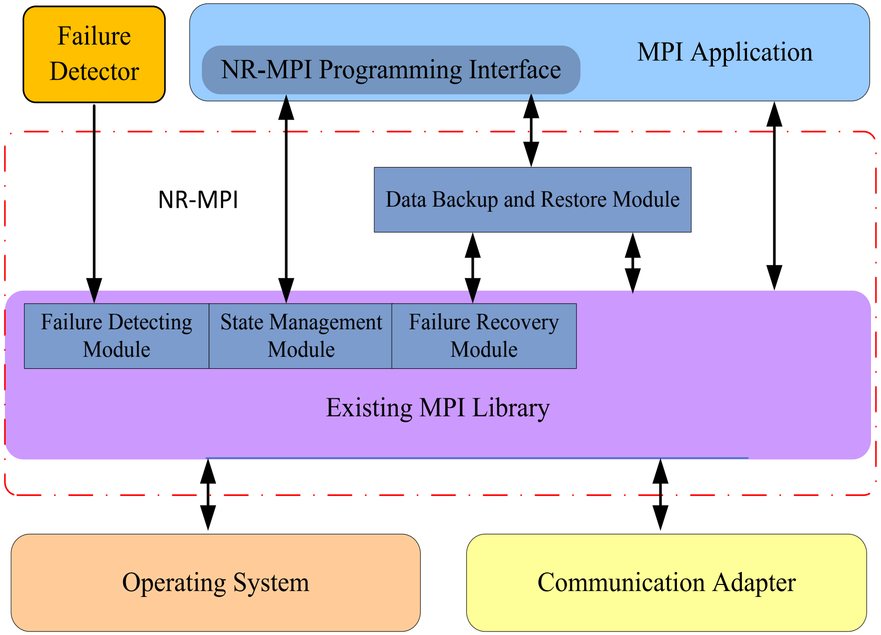


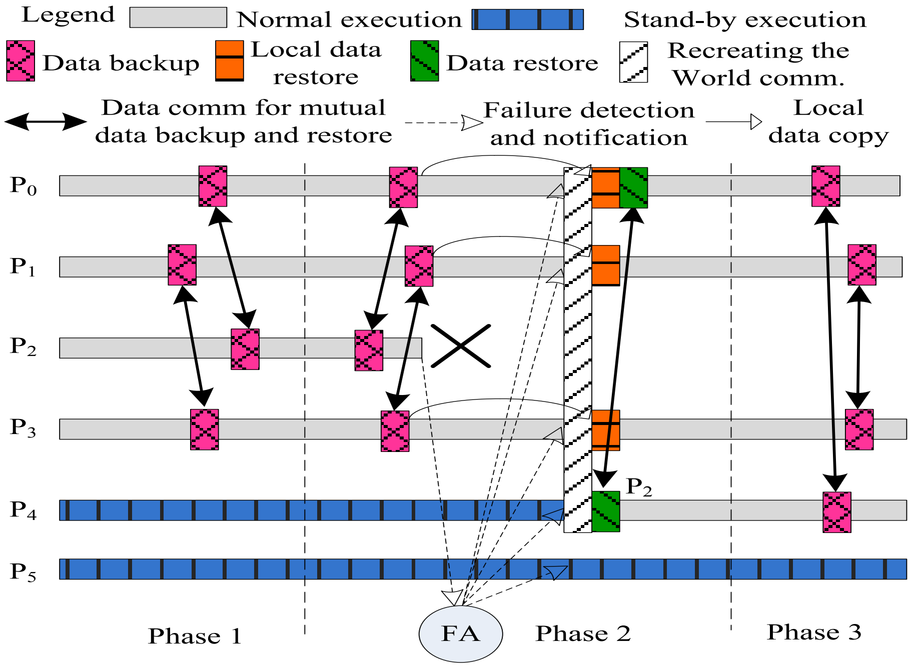




Scalable MPI

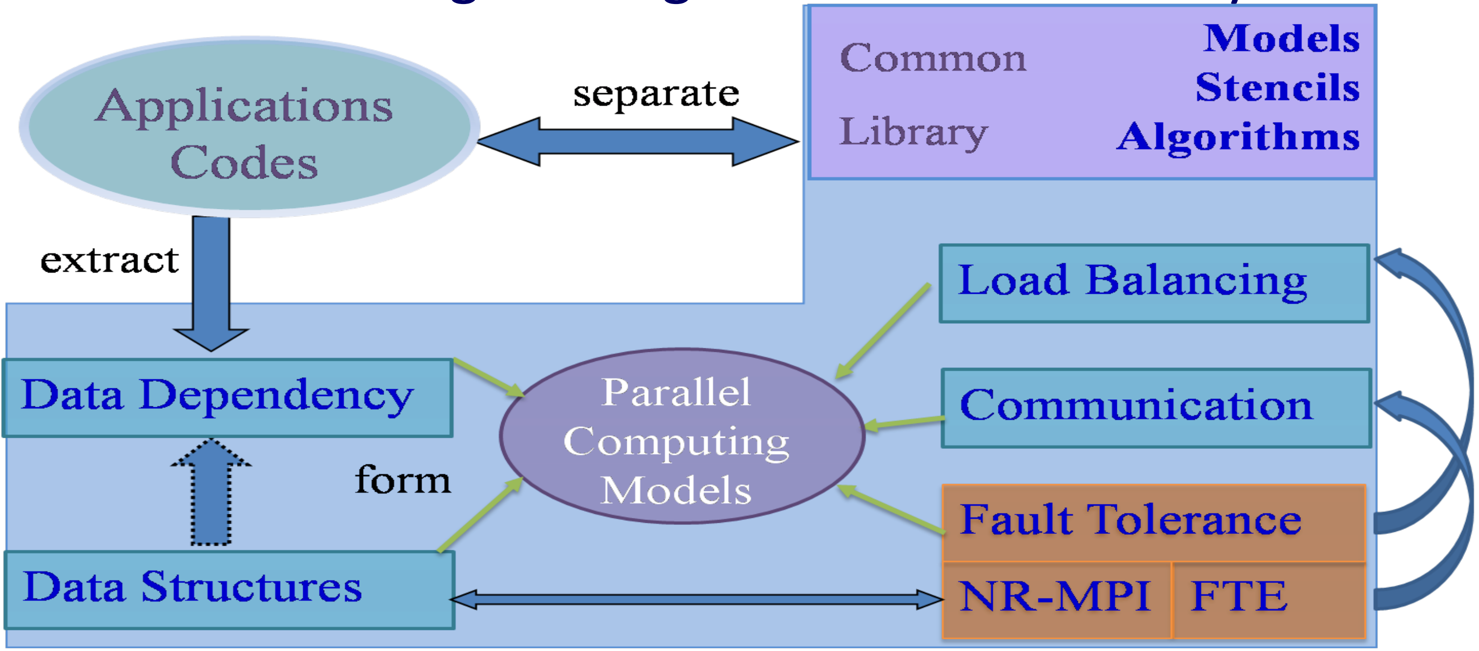
- Non-stop and fault Resilient MPI (NR-MPI)
  - Application continue execution without being relaunched
  - Failure detection and MPI state recovery done by runtime
  - Data-backup by application-level diskless C/R
  - Reconstruct of MPI communicator and channel





Domain Specific Framework

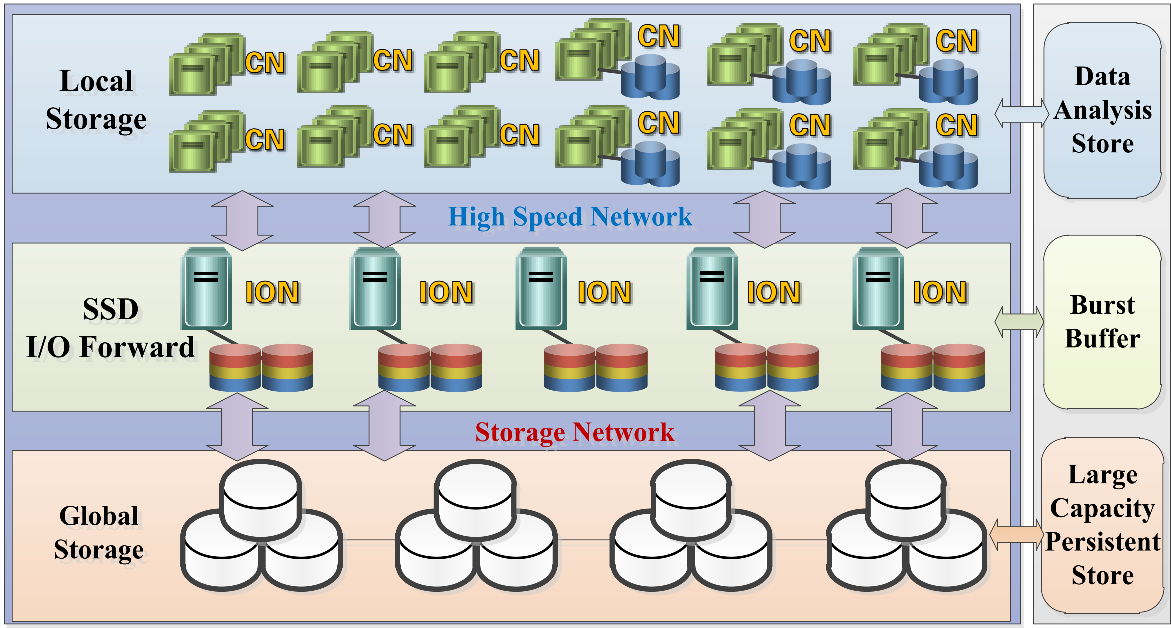
- Hide parallel programming complexity using millions of cores and the hierarchy of parallel computers
- Integrates the efficient implementations of parallel fast algorithms
- Provides efficient data structures and solver libraries
- Supports software engineering for code extensibility



Scalable IO Structure

天河

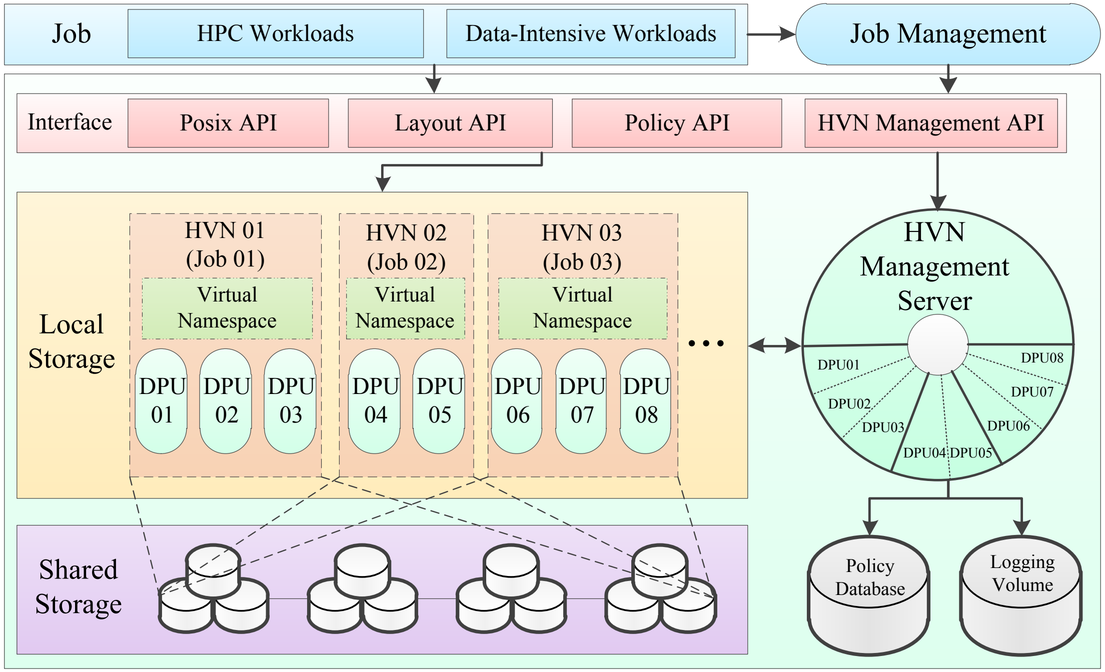
- ❑ IO Architecture on Tianhe-2
  - ➢ Multiple Layers & Hybrid Storages
    - ◆ Local Disk
    - ◆ PCI-E SSD
    - ◆ Disk Array
  - ➢ 6400 local Disks
    - ◆ Bus attached
  - ➢ 256 IO nodes
    - ◆ Burst: above 1TB/s
    - ◆ TH-Express and IB QDR port
  - ➢ 64 Storage Servers
    - ◆ Sustained： about 100GB/s



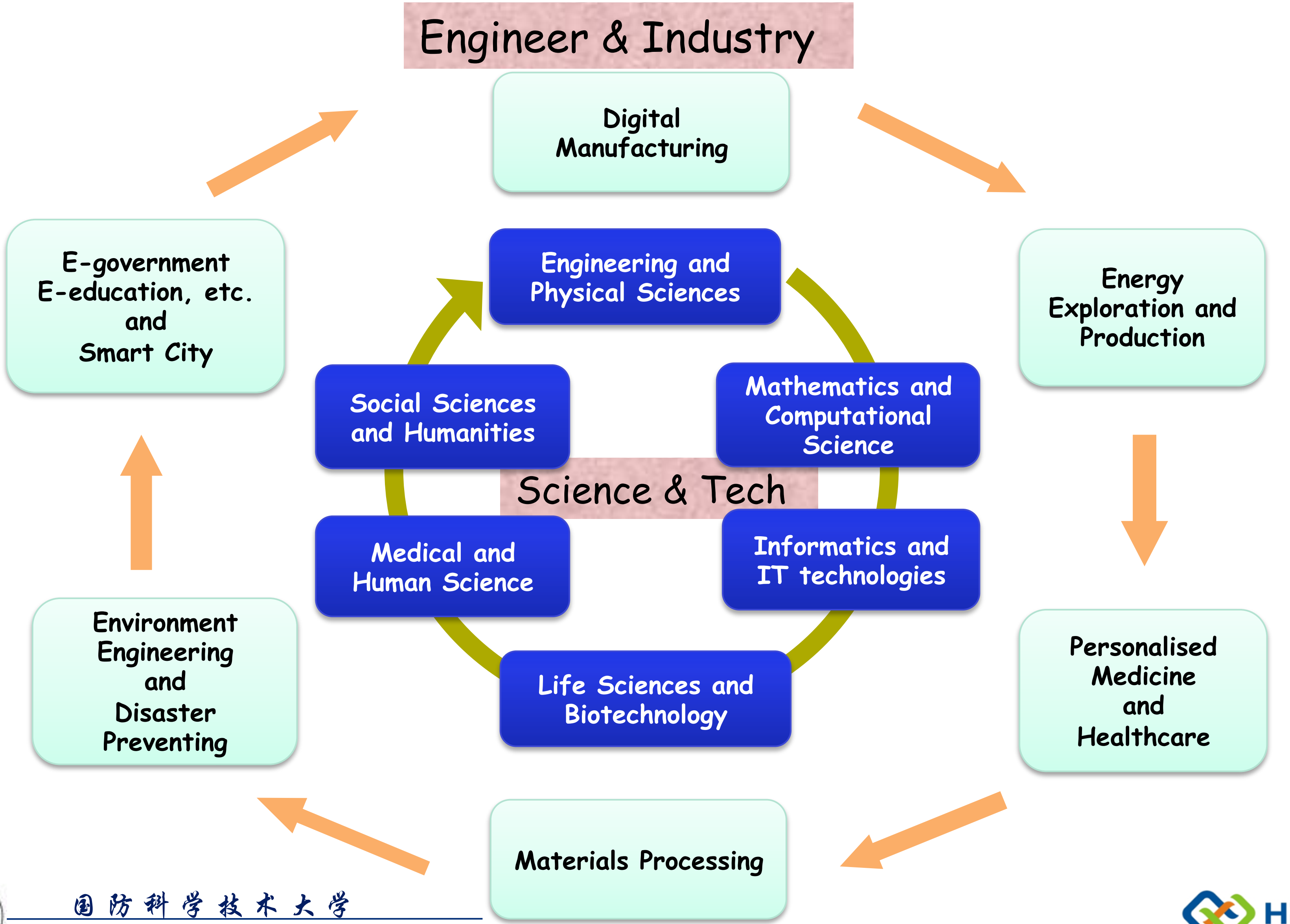
Scalable IO Structure

天河

- ❒ 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



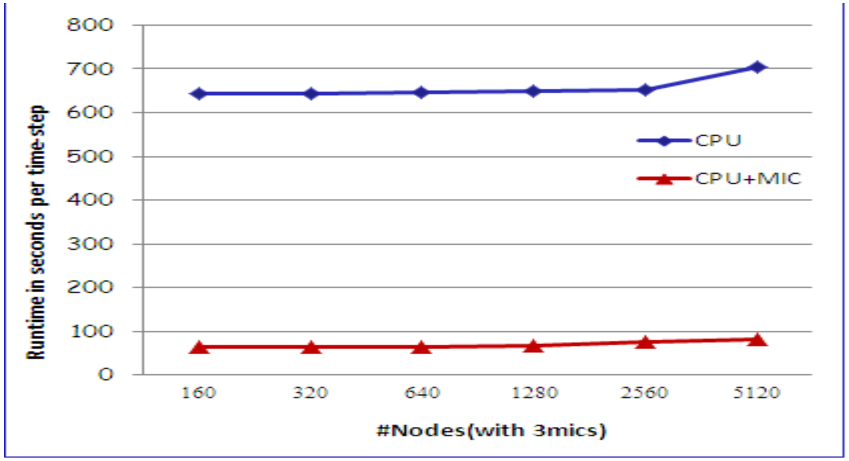
Scalable Application

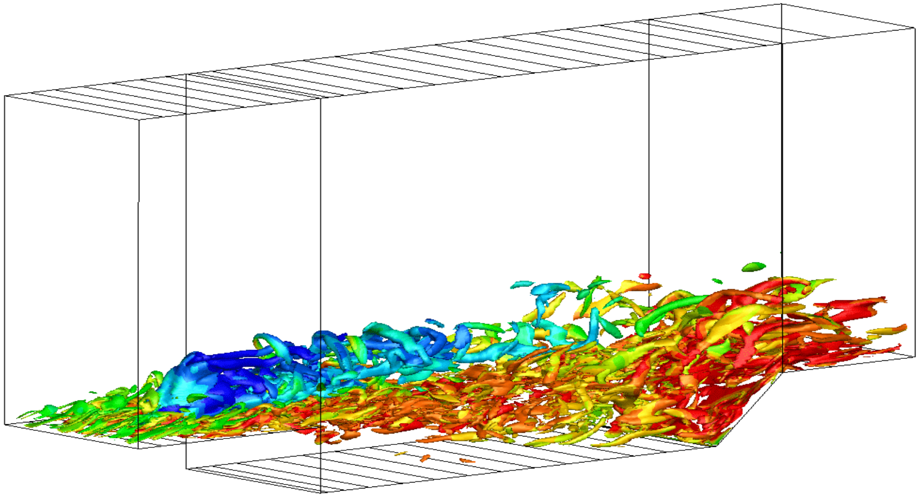

Engineer & Industry
Digital Manufacturing
Energy Exploration and Production
Personalised Medicine and Healthcare
Materials Processing
Environment Engineering and Disaster Preventing
E-government E-education, etc. and Smart City
Engineering and Physical Sciences
Mathematics and Computational Science
Informatics and IT technologies
Life Sciences and Biotechnology
Medical and Human Science
Social Sciences and Humanities
Science & Tech

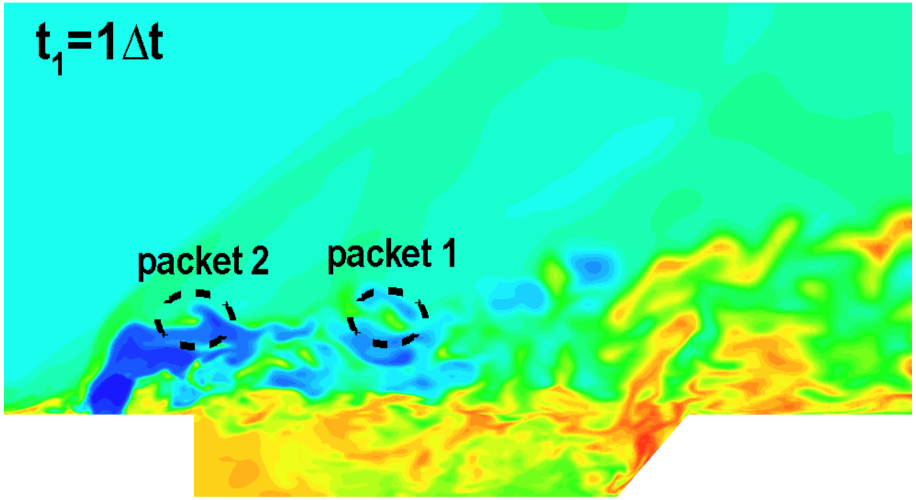
Application Case -- CFD

天河

- Hybrid RANS/LES simulation of scramjet combustion
  - CPU+MIC version is 8.63X to 13.86X faster than the original CPU version
  - The largest scramjet combustion simulation up to now: totally 998,400 cores on 26,880 million cells, parallel efficiency 79%.







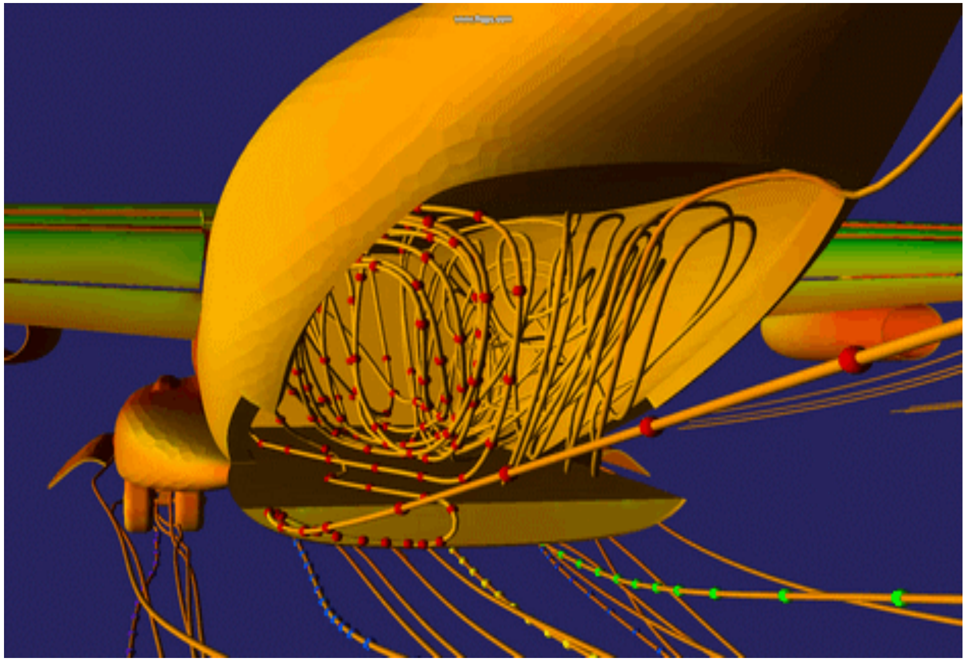
国防科学技术大学

National University of Defense Technology

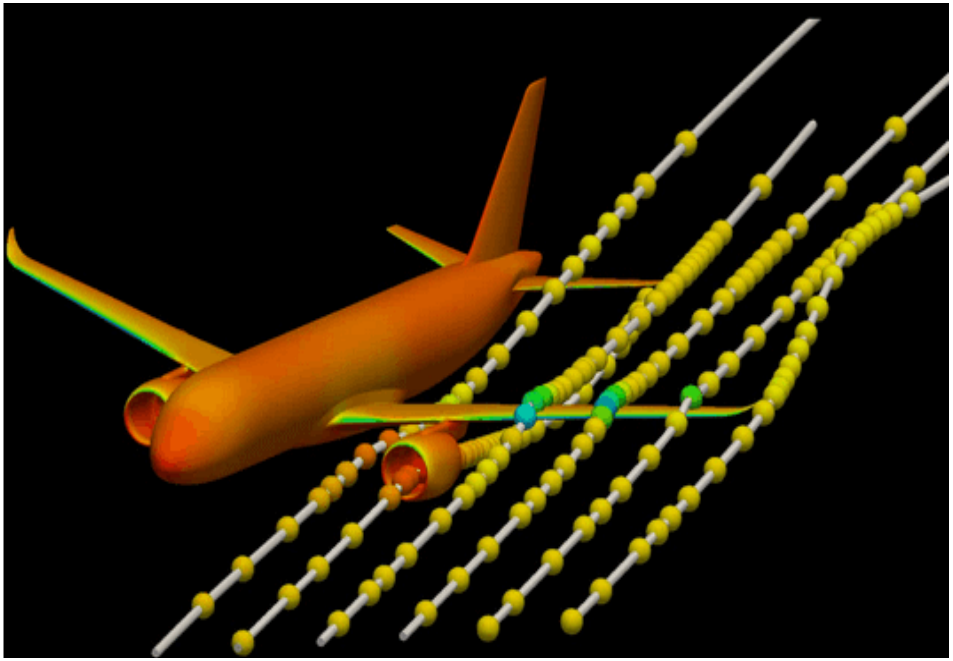
HPCL

Application Case -- CFD

- ❐ Large transport aircraft

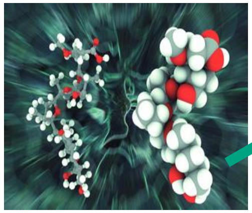


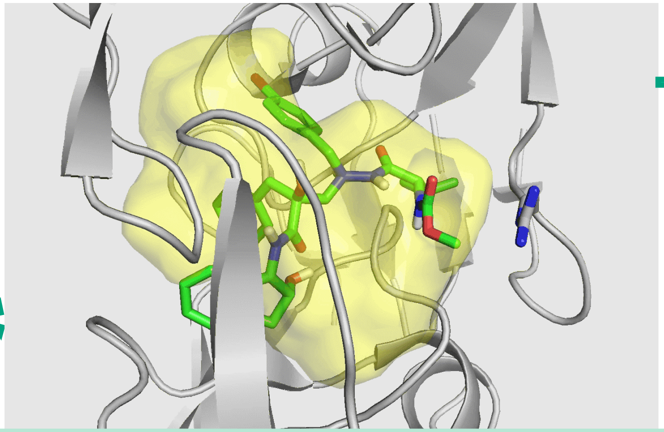
- ❐ Large passenger aircraft



Application Case – Bio-Medicine

- High Throughput Virtual Screening
- Applications: Computer Aided Drug Design, Molecular Docking, and Virtual screening
- mD^{3}DOCKxb, Lamarckian Genetic Algorithm
- Data Scale, 40 millions molecules, 800TB
- Bottleneck, IO BW, Comm BW





Find out the best 100 molecules;

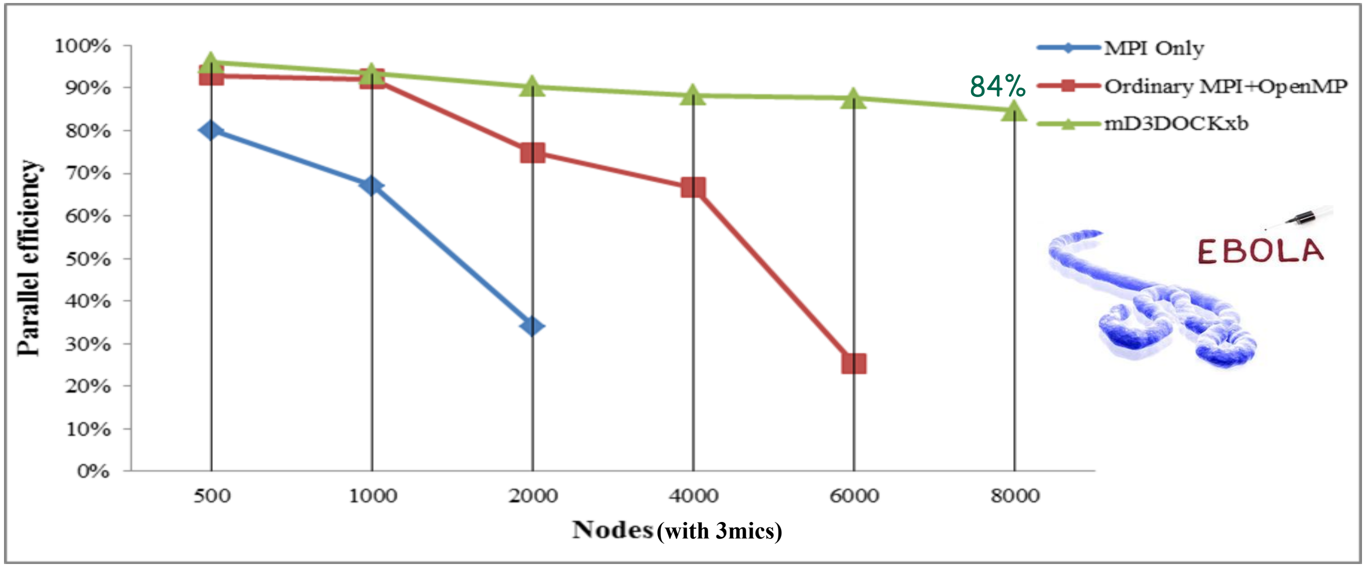
Do experiment;

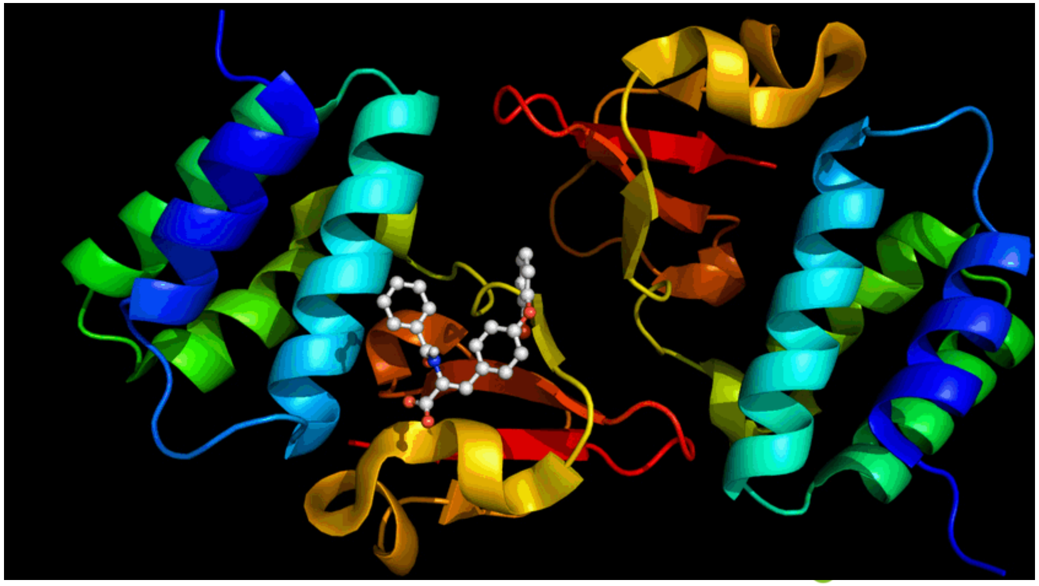
Clinical

Aim: Finish docking of all the drug molecules on earth within one day.

Application Case – Bio-Medicine

- Against Ebora virus 40 millions real drug molecules docking in 20hours ON TH-2

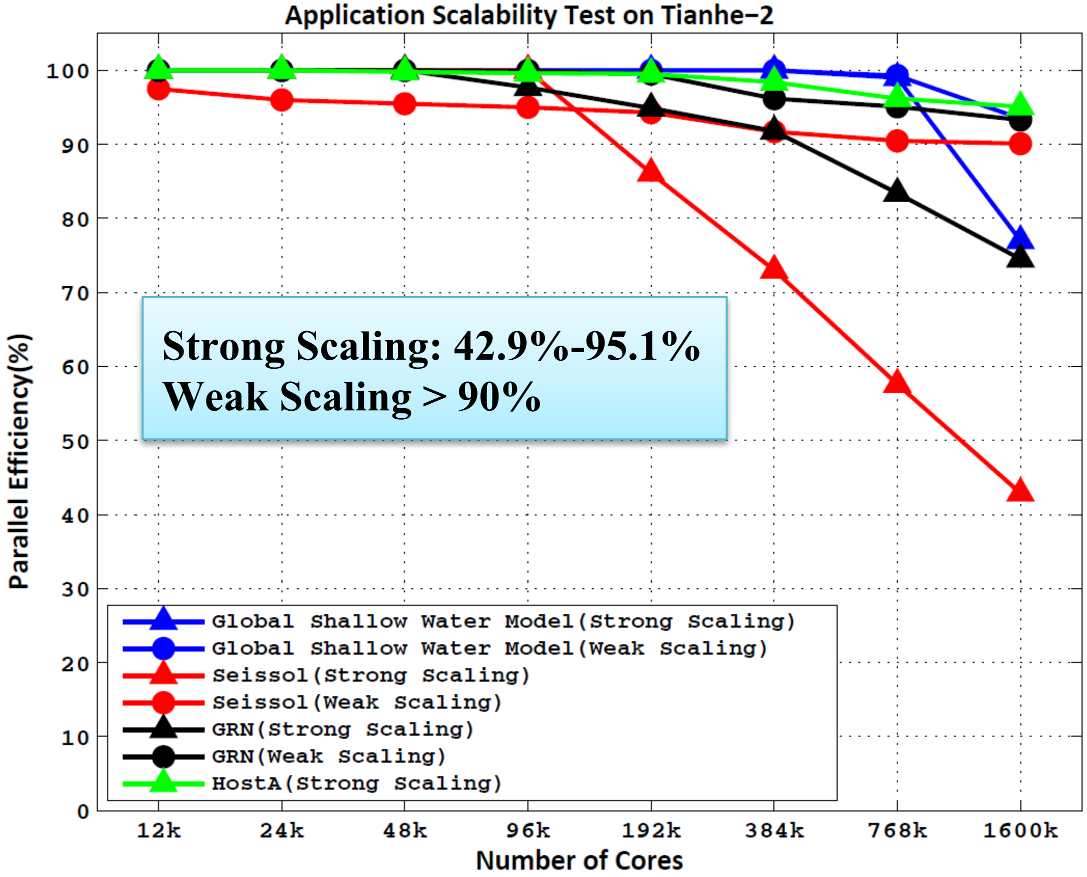




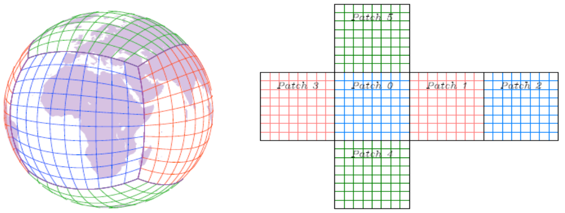
Other Applications

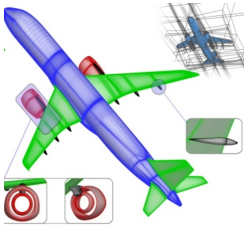
天河

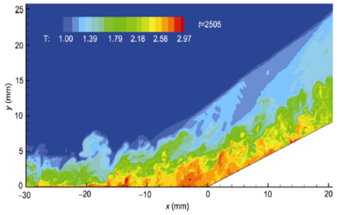
❒Other Applications

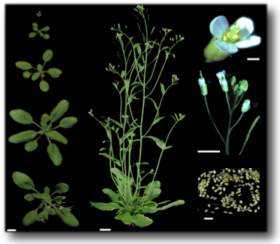


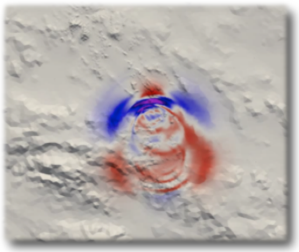
Strong Scaling: 42.9%-95.1%
Weak Scaling > 90%

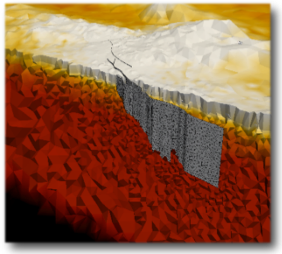












Cloud Computing & Bigdata

天河

- ❐ Additional Applications
  - ➢ E-Gov
  - ➢ micMR
  - ➢ RenderCloud
  - ➢ Health care
  - ➢ Smart city
  - ➢ Video Processing
  - ➢ Education
  - ➢ Social Network Analysis

广州市电子政务数据中心管理系统

用户：t_dc | 我的配置 |

管理 系统管理 统计库

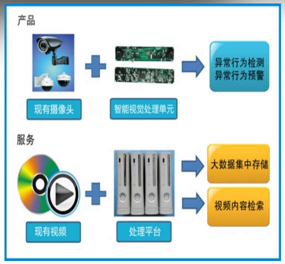
数据中心

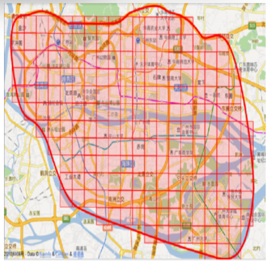
个 | 日均交换：条

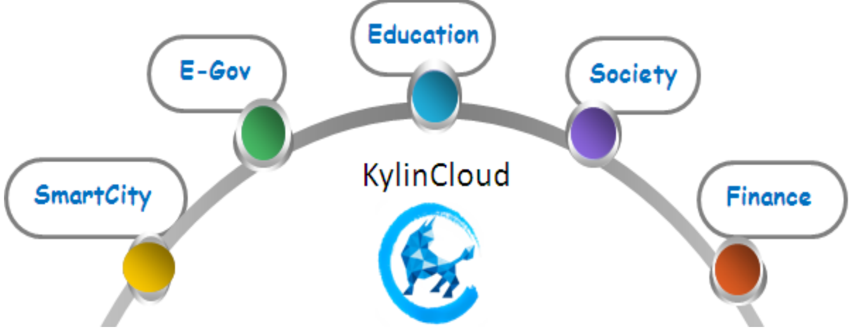
供数据 0 获取数据0

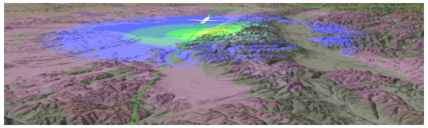
当日获取数据：0 条

最后获取时间：--









创意云

首页 | 需求发布 | 在线创作 | 渲染农场 | 创意商城 | 游戏大厅

Hello!

01 需求

02 创作

03 渲染

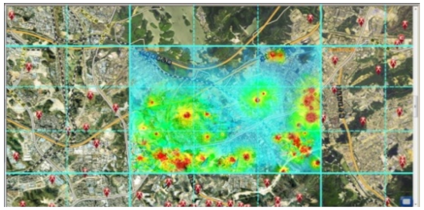
04 商城

05 游戏

登录创意云・云端体验

登 录

需求 创作 渲染 商城 游戏



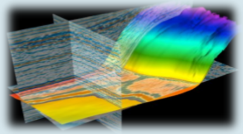
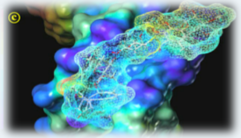
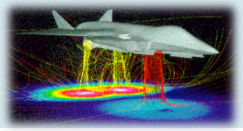
国防科学技术大学

National University of Defense Technology

HPCL

Application scale in next 5 years

天河

Applications	Current Scale in China	Scale in next 5 years
 **Seismic Exploration**	2600km^2 , 5km depth 217900 shots 2.2TB data	Millions of shots
Genomics Research	2PB bioinformatics data	100PB bio data
New Energy (Magnetic Confinement Fusion)	2 billion ions 0.83 billion electrons	100 billion electrons
 **Drug Design**	200-300ns Molecular Dynamics simulations	10 Million molecular 1000ns/day
 **CFD** (Aircraft Design)	3.5 billion mesh points	100 billion mesh points
Universal Evolution (neutrinos)	110 billion particles	Trillion particles
Smart City (Urban Electromagnetic Spectrum Monitoring System)	Area (Guangzhou city): 200km^2 Grid size:1.0km*1.0km	Grid Size: 100m*100m

Status of Tianhe System

天河

System	Tianhe-1A	Tianhe-2	Tianhe-2A
System Peak(PF)	4.7	54.9	~100
Peak Power(MW)	4.04	17.8	~18
Total System Memory	262 TB	1.4 PB	~3PB
Node Performance(TF)	0.655	3.431	~6
Node processors	Xeon X5670 Nvidia M2050	Xeon E5 2692 Xeon Phi	Xeon or China CPU China Accelerator
System size(nodes)	7,168 nodes	16,000 nodes	~18,000
System Interconnect	TH Express-1	TH Express-2	TH Express-2+
File System	2 PB Lustre	12.4PB H^2FS +Lustre	~30PB H^2FS+TDM

China Accelerator

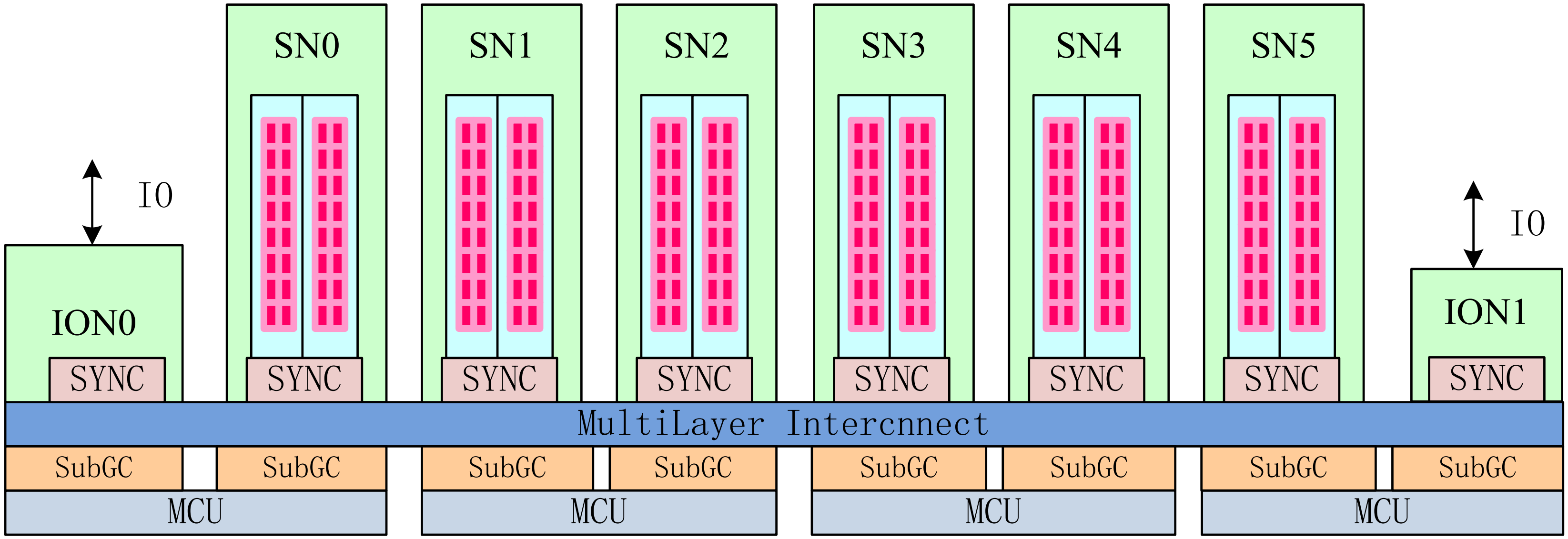
Matrix2000 GPDSP

❒ High Performance

- 64bit Supported
- ~2.4/4.8TFlops(DP/SP)
- 1GHz, ~200W

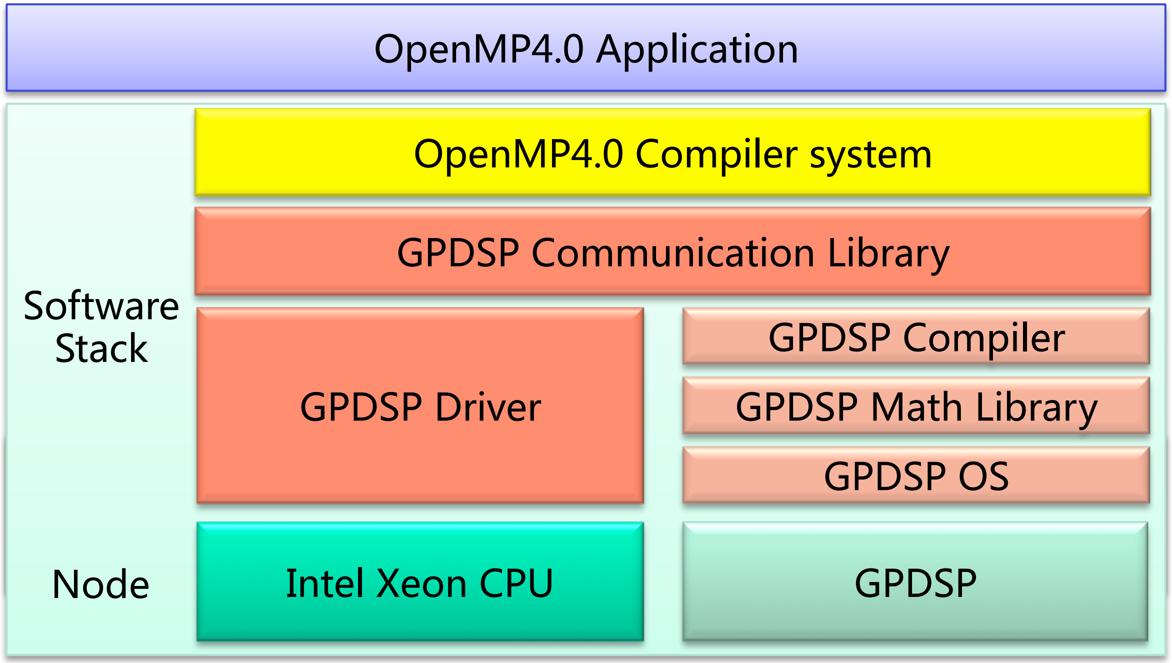
❒ High Throughput

- High-bandwidth Memory
- 32~64GB
- PCIE 3.0, 16x



China Accelerator

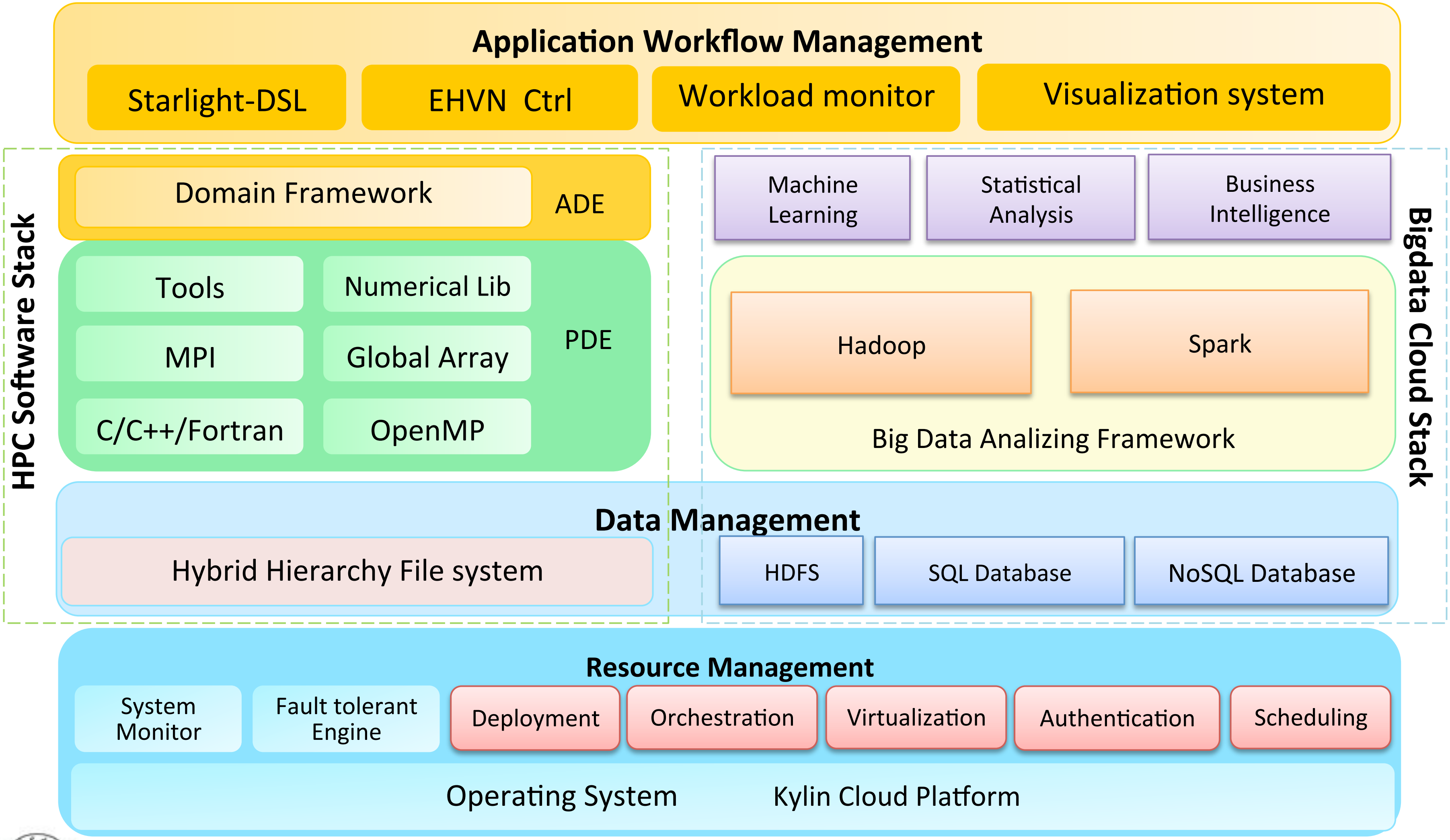
- ❐ Software stack
  - ➢ OpenMP4.0
  - ➢ OS, Compiler and Math Library on GPDSP
  - ➢ GPDSP Driver, Communication Library



HPC & Bigdata convergence Stack

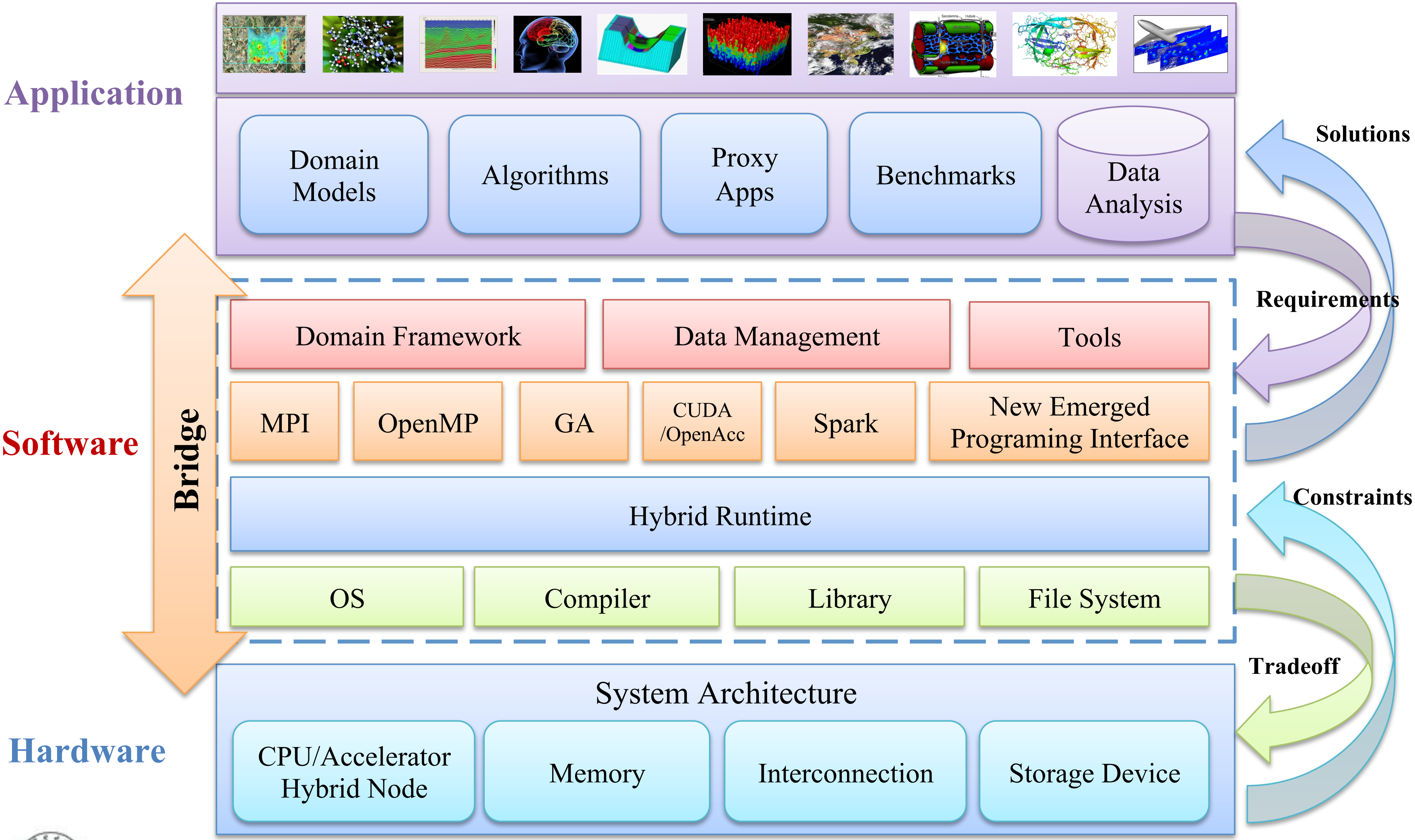
天河

❐ Starlight Software Stack



Co-design Eco-system

天河



Application

Domain Models

Algorithms

Proxy Apps

Benchmarks

Data Analysis

Software

Bridge

Domain Framework

Data Management

Tools

MPI

OpenMP

GA

CUDA /OpenAcc

Spark

New Emerged Programing Interface

Hybrid Runtime

OS

Compiler

Library

File System

Hardware

System Architecture

CPU/Accelerator Hybrid Node

Memory

Interconnection

Storage Device

Solutions

Requirements

Constraints

Tradeoff

Summary

□ Heterogeneous Architecture

□ Adaptive system and software design

□ HPC & Bigdata Convergence

□ Supercomputing Eco-system

Thanks