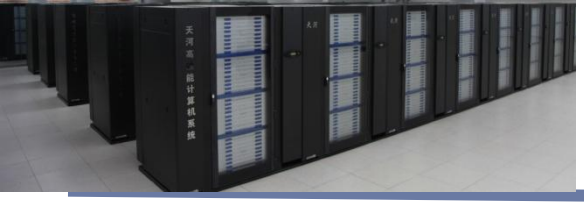


Tianhe
天河

Hybrid System & Application

Yutong Lu
School of Computer Science
NUDT China
ytlu@nudt.edu.cn

Outline

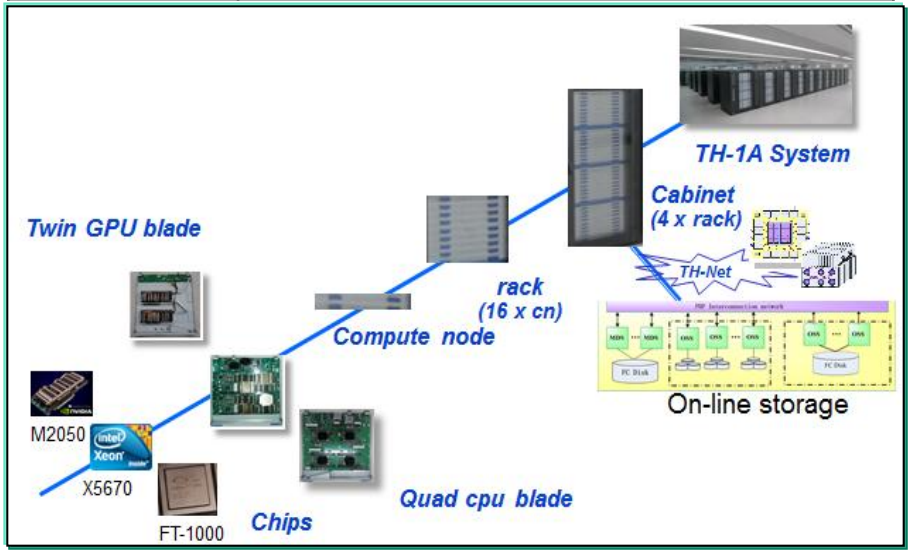
- Overview of Tianhe-1A
- Status of Application
- Prospect of next generation of Tianhe system

Overview of TianHe-1A

Motivation

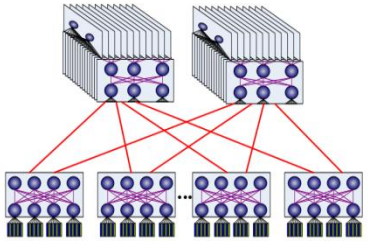
- High productivity petaflops system
  - High performance
  - High bandwidth and low latency communication
  - High throughput and large capacity I/O
  - Low power consumption
  - Maintenance and Usability
- NSCC-TJ & Others
  - Open platform for research and education
  - Public information infrastructure

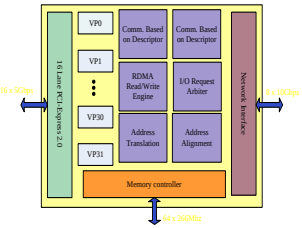
	Specification (2010.11)
Processors	7168nodes,14336 Intel CPUs + 7168 nVIDIA GPUs +2048FT CPUS, Peak 4.7PF, Linpack 2.57PF
Interconnect	Proprietary high-speed interconnection network TH-net,10GB/s per direction
Memory	262TB in total
Storage	Global shared parallel storage system, 2PB
Cabinets	140 compute / communication/storage Cabinets
Power	4.04MW (635.15MF/W)
Cooling	Water cooling system

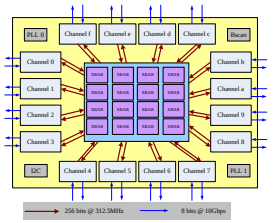


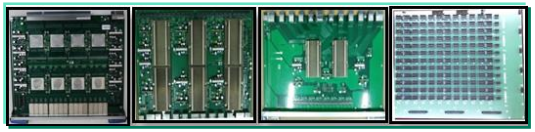
Overview of TianHe-1A

- TH-Net
  - High frequency 10Gbps signal transmission
  - Optimized Channel bonding (8 Lane x 10Gbps)
  - High radix route
  - Topology: Hierarchy fat-tree structure
- Network interface ASIC：NIC
  - Implement bulk data transfer such as RDMA and MP
- High radix router ASIC：NRC
  - 16ports, Throughput of single NRC: 2.56Tbps
  - Communication protocol
- Performance
  - P2P
    - bi-BW: 20GB/s
    - latency: 22ns
  - System
    - Aggregate BW: 1228.8Tbps
    - bi-section BW: 307.2Tbps
  - Switch
    - Ports: 384
    - Throughput: 61.44Tbps





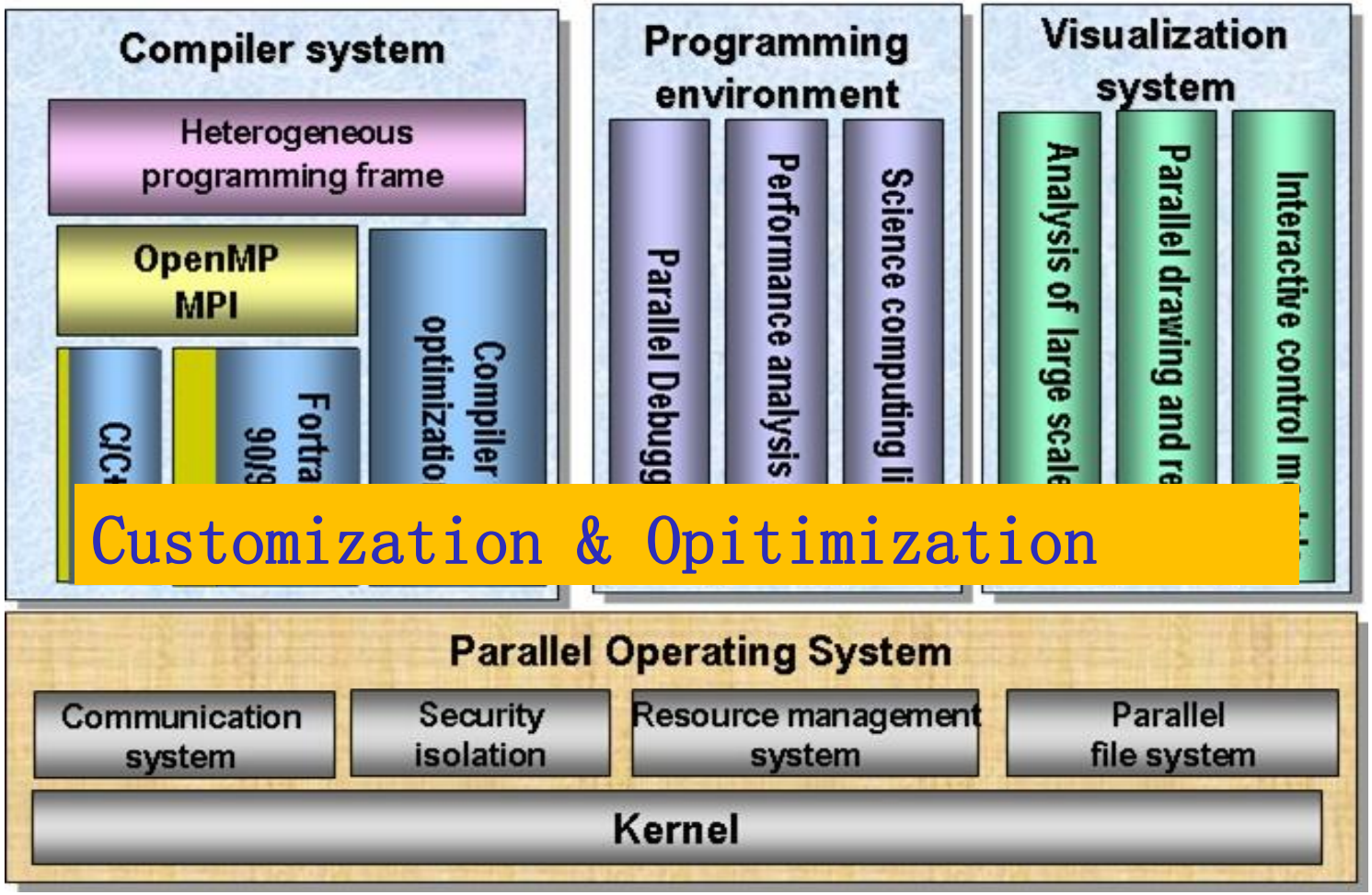




Overview of TianHe-1A

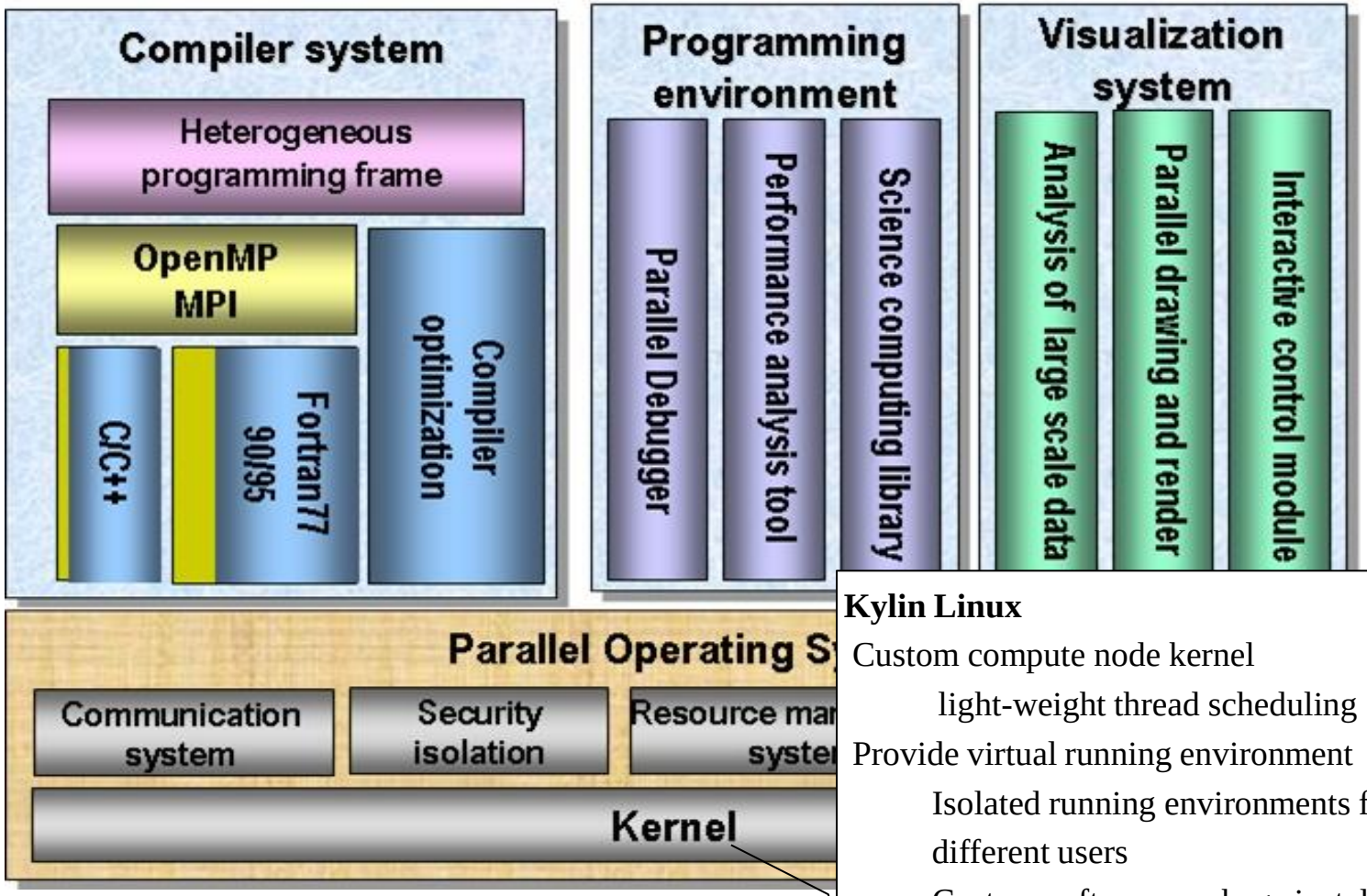
Tianhe 天河

TH-1A software stack



Overview of TianHe-1A

TH-1A software stack



Kylin Linux

Custom compute node kernel

light-weight thread scheduling

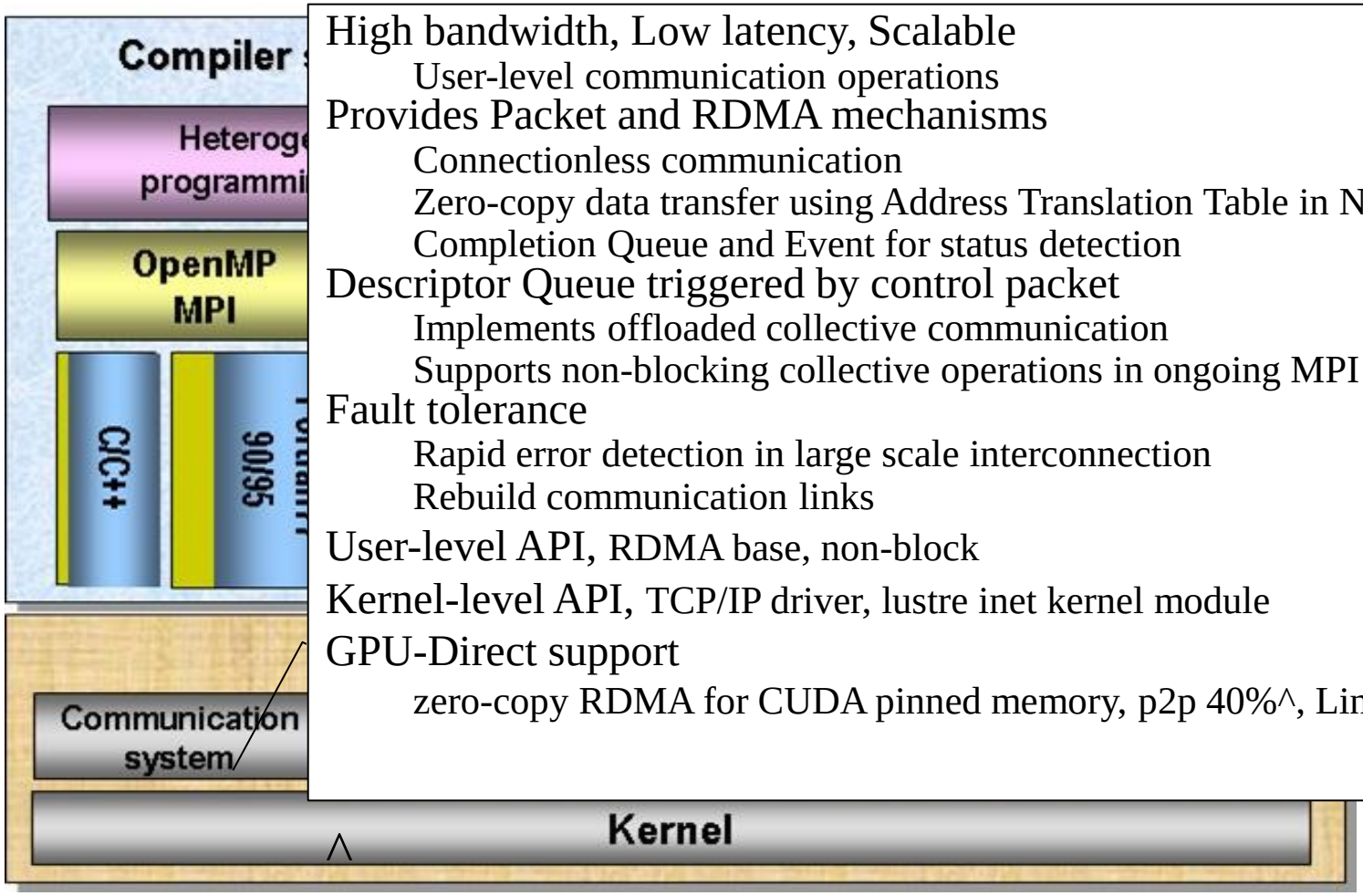
Provide virtual running environment

Isolated running environments for different users

Custom software package installation

Overview of TianHe-1A

TH-1A software stack



High bandwidth, Low latency, Scalable
- User-level communication operations

Provides Packet and RDMA mechanisms
- Connectionless communication
- Zero-copy data transfer using Address Translation Table in NIC
- Completion Queue and Event for status detection

Descriptor Queue triggered by control packet
- Implements offloaded collective communication
- Supports non-blocking collective operations in ongoing MPI standard

Fault tolerance
- Rapid error detection in large scale interconnection
- Rebuild communication links

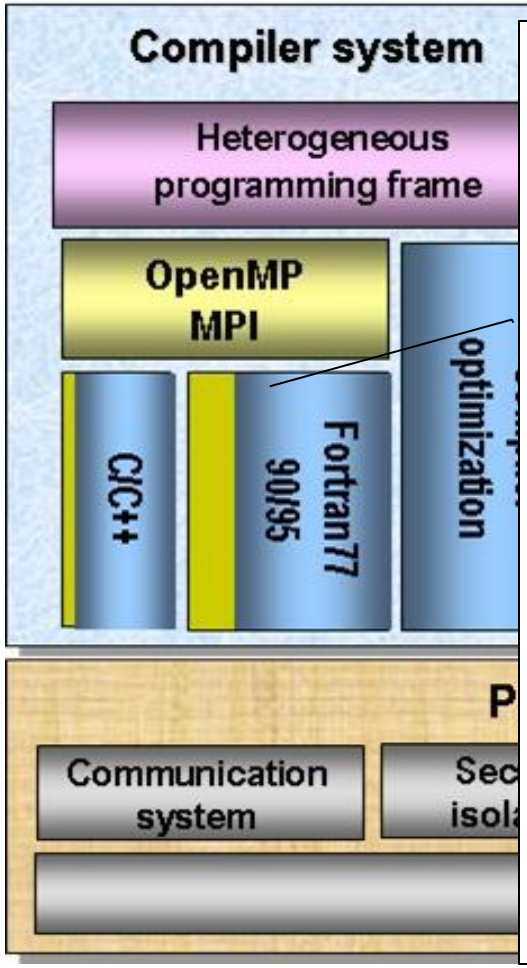
User-level API, RDMA base, non-block

Kernel-level API, TCP/IP driver, lustre inet kernel module

GPU-Direct support
- zero-copy RDMA for CUDA pinned memory, p2p 40%^, Linpack 5%^

Overview of TianHe-1A

TH-1A software stack



MPI-2.2 standard compliance

P2P: Latency: 1.57 us; Bandwidth: 6340 MB/s

Multiple channel approach

- Shared memory channel
- Exclusive RDMA channel
- Shared RDMA channel

Hybrid channels

- Utilizing the characteristic of nearest-neighbor communication in many applications
- Improve performance with limited costs of hardware resources

Zero-copy rendezvous protocol

- Pin-down Cache to reduce the overhead of memory registration

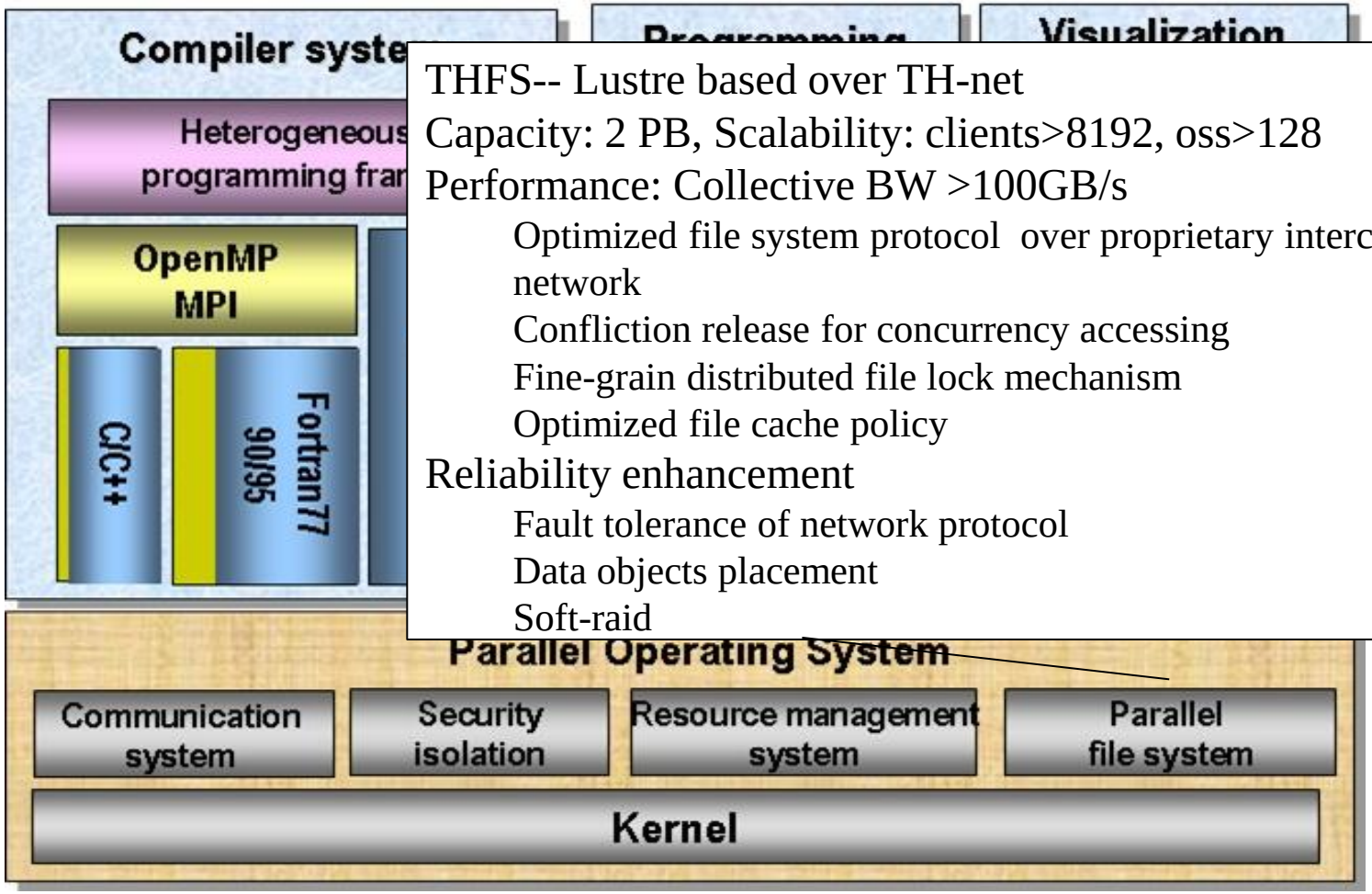
Efficient collective communication based on hardware supporting

MPI-IO

- ADIO driver for FS

Overview of TianHe-1A

TH-1A software stack



THFS-- Lustre based over TH-net

Capacity: 2 PB, Scalability: clients>8192, oss>128

Performance: Collective BW >100GB/s

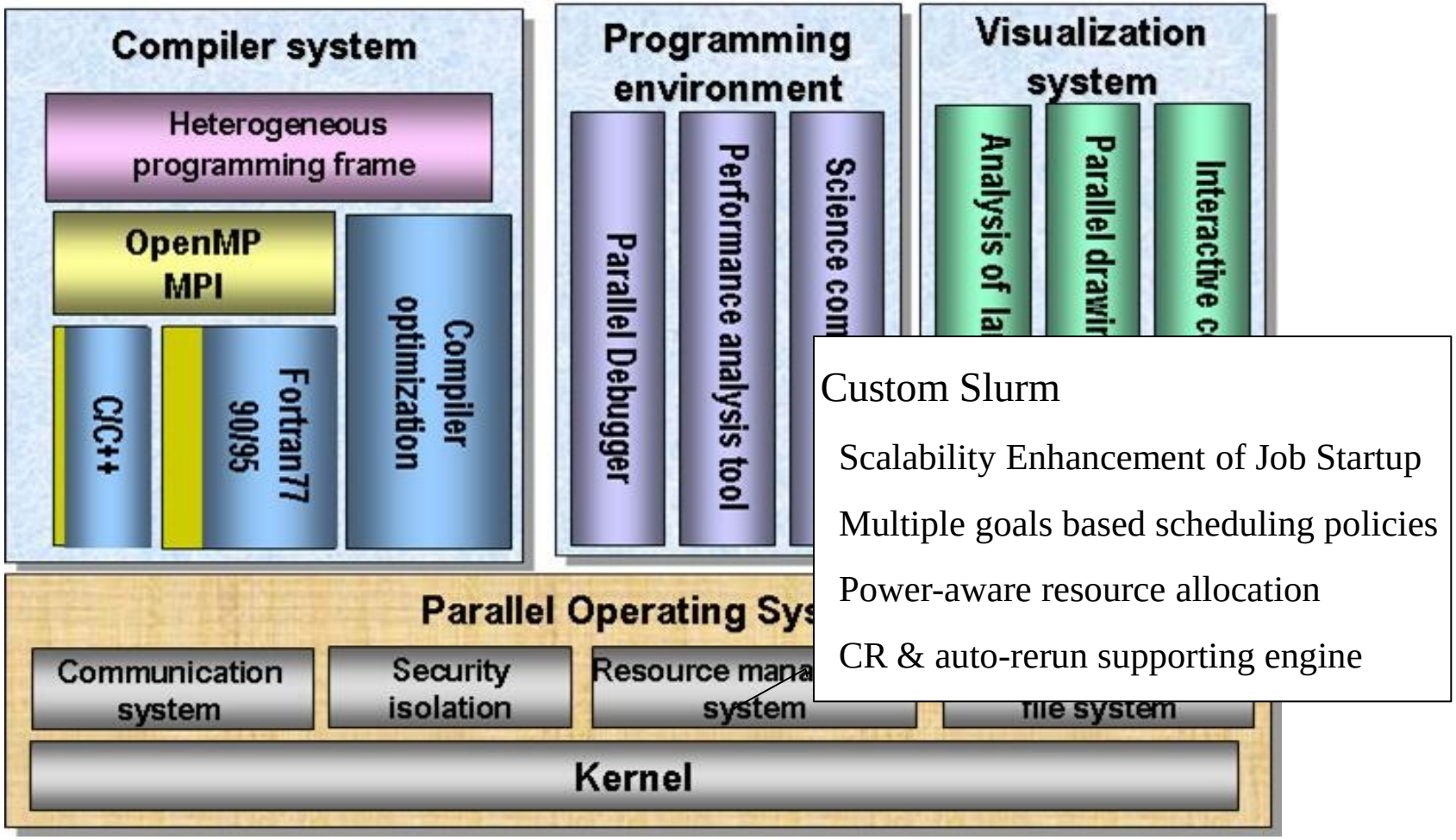
- Optimized file system protocol over proprietary interconnection network
- Confliction release for concurrency accessing
- Fine-grain distributed file lock mechanism
- Optimized file cache policy

Reliability enhancement

- Fault tolerance of network protocol
- Data objects placement
- Soft-raid

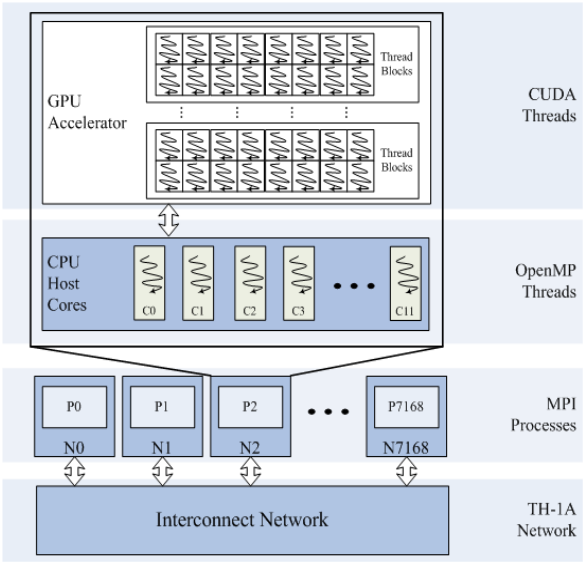
Overview of TianHe-1A

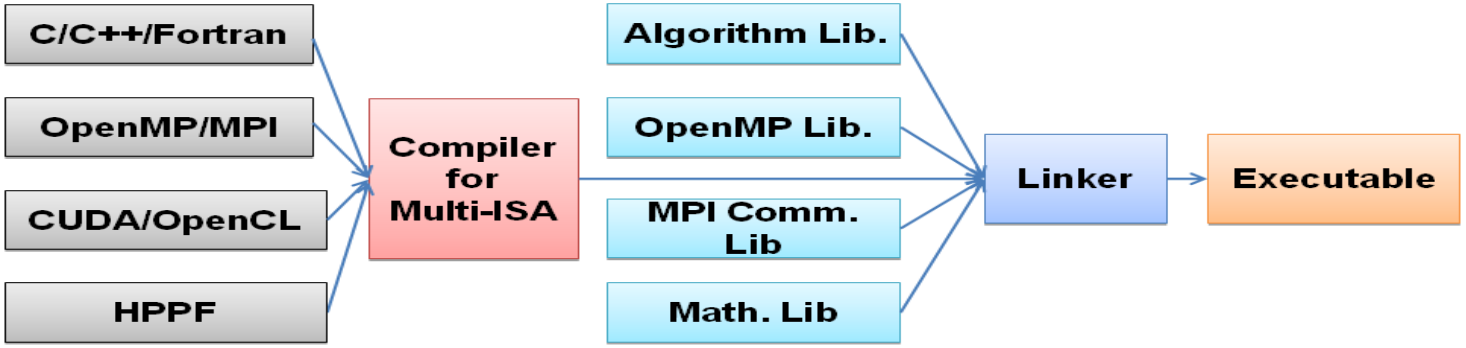
TH-1A software stack



Overview of TianHe-1A

- Traditional Parallel programming on TH-1A
  - C, Fortran, C++, Java
  - CUDA/OpenCL + OpenMP + MPI
  - CUDA/OpenCL + PThread + MPI
- Heterogeneous programming framework
  - Accelerate the large scale, complex applications
  - Develop applications efficiently
  - Including:
    - Inter-node homogeneous parallel programming (users)
    - Intra-node heterogeneous parallel computing (experts)





Overview of TianHe-1A

- Inter-node: Homogeneous framework (JASMIN)
  - Hiding parallel programming details for large scale applications
  - Patch-based objects data structures
  - MPI communication, dynamic load balancing
  - Zero-copy optimization in communication library

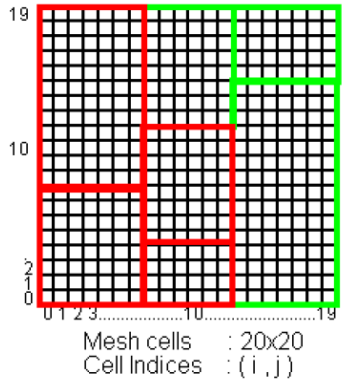
A parallel software infrastructure

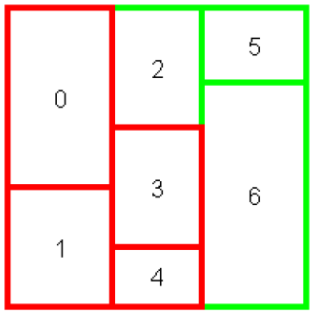
JASMIN

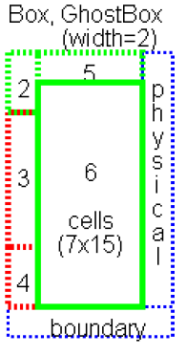
(J Adaptive Structured Mesh applications INfrastructure)

Supports the large scale parallel simulations on adaptive structured mesh (SAMR) using massively parallel processing machines (MPP).

http:://www.iapcm.ac.cn/jasmin



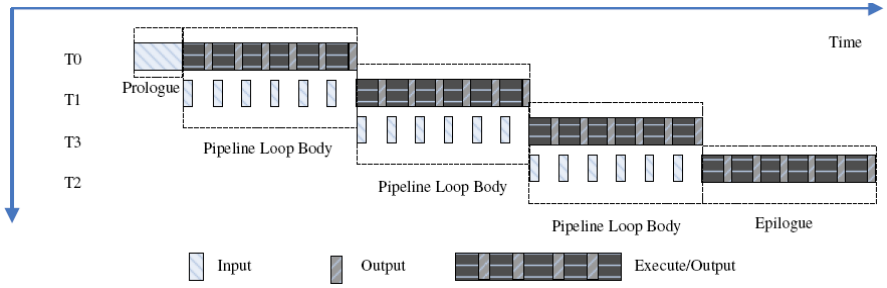




Overview of TianHe-1A

- Intra-node: Heterogeneous framework
  - Hiding GPU programming
  - Optimizations including
    - Adaptive partitioning, balance the workloads between CPUs and GPU
    - Asynchronous data transfer / computing, overlap CPU operations with GPU operations
    - Software pipelining, overlap GPU computing with data transfer between host and GPU device memory

$$Gsplit = \frac{P_{GPU}}{P_{GPU} + P_{CPU}}$$

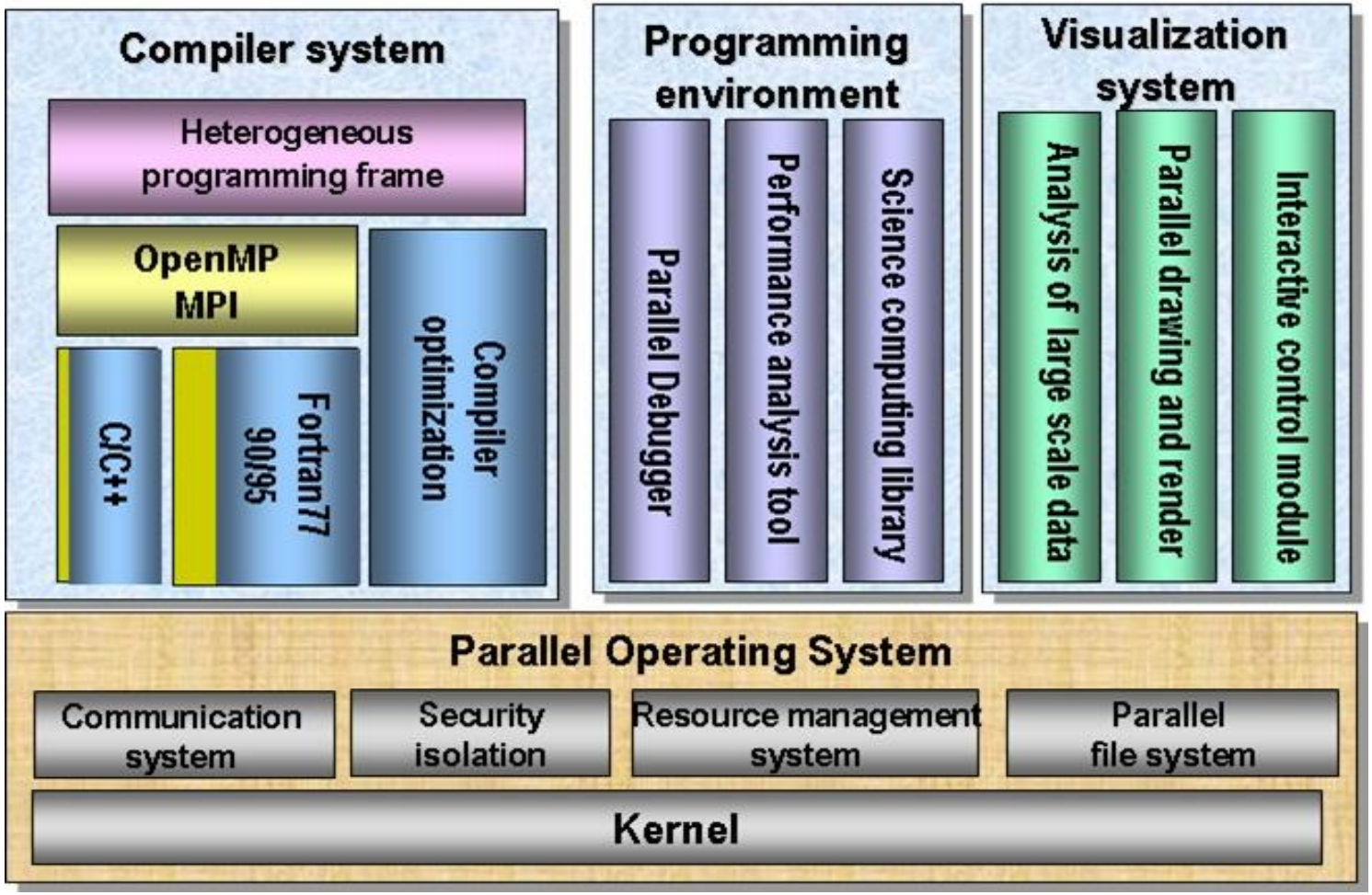


Improve the performance of various applications 10~22%

Overview of TianHe-1A

Tianhe 天河

TH-1A software stack



Status of Application

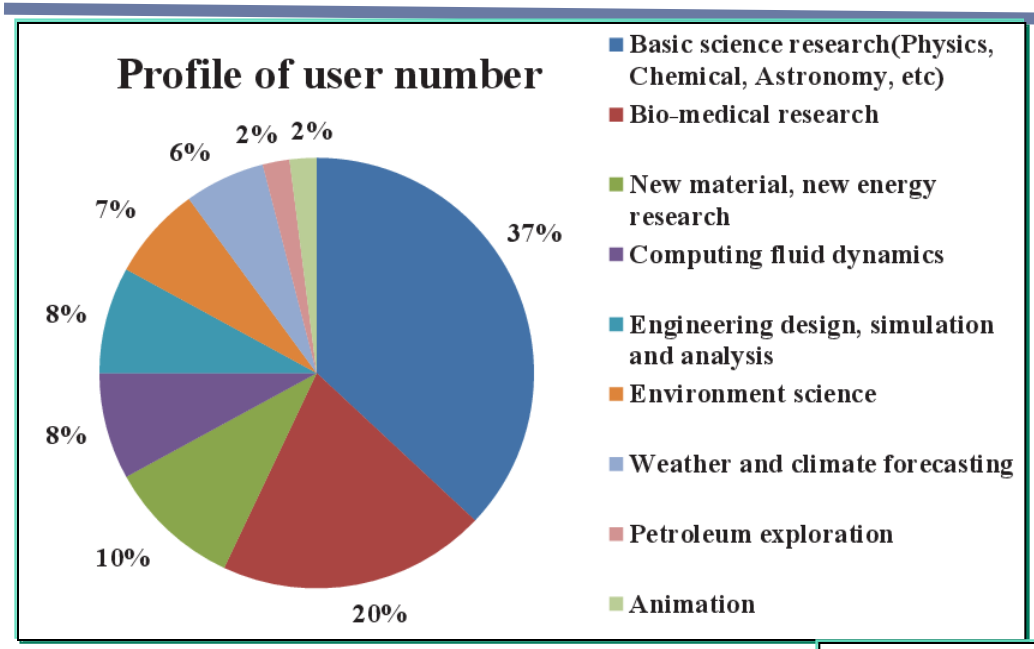
Tianhe 天河

- NSCC-Tianjin (Rank 5)
  - Rpeak 4700TFlops， Rmax2566TFlops
    - 7168 nodes, 14336CPU+7168GPU, 4MW
  - 2010.11

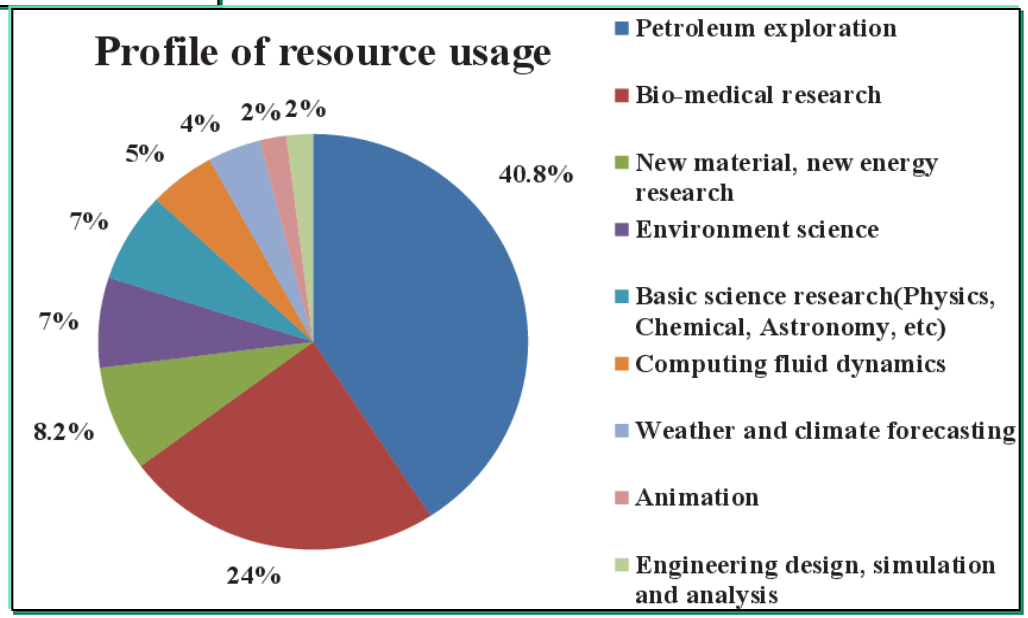
- NSCC-Changsha (Rank 28)
  - Rpeak 1343TFlops， Rmax 771.7TFlops
    - 2048 nodes, 4096CPU+2048GPU, 1.1MW
  - 2011.7

- NSCC-Guangzhou (Rank 80)
  - Rpeak335.8TFlops Rmax211.7TFlops
    - 512 nodes, 1024CPU+512GPU, 289KW
  - 2012.5

Status of Application



- Statistic in NSCC-TJ
  - >130,000 Jobs
  - ~76% Utilization



Status of Application

5 + X Scheme

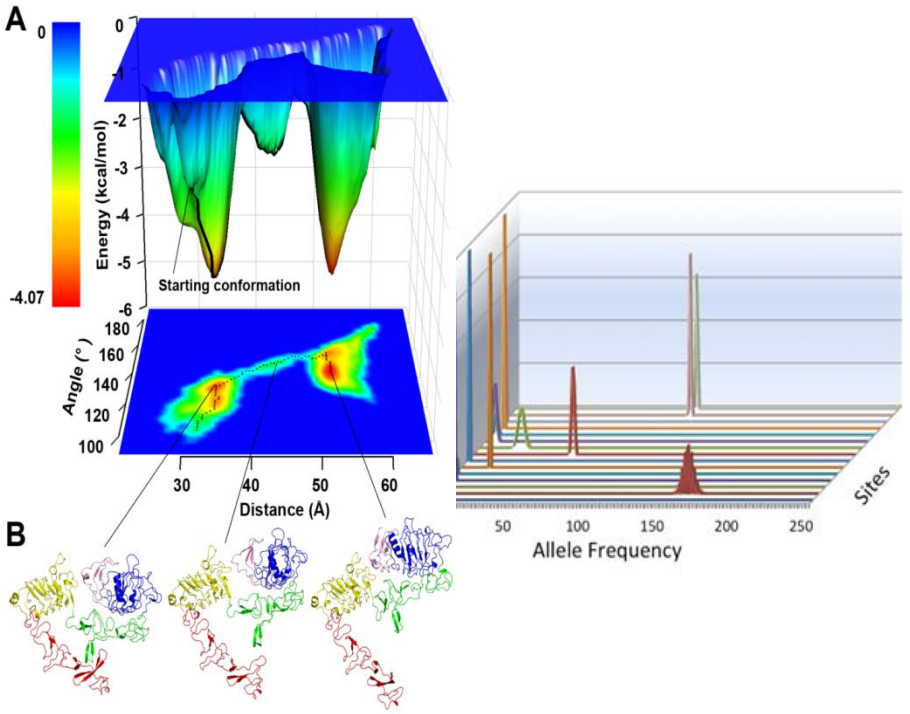
- Five industry platforms
  - Petroleum seismic processing
  - Biologic medicine computing
  - High-end equipment manufacture
  - Animation and image
  - Geographic information processing
- X: initiative of Sci & Tech
  - Climate
  - Energy
  - Dynamic ……

Case study

- Petroleum seismic processing
  - Seismic Imaging, Reverse Time Migration(RTM)
  - Maximize the production of discovered reservoirs
  - Explore new ones in complex domains

Survey detail	Running detail
1050 sq.km 700GB seismic data	70000 shots, 7100 nodes 16 hours
680 sq.km 1.4TB seismic data	80000 shots, 7000 nodes 40 hours
2600 sq.km 2.2TB seismic data	217900 shots，2000 nodes 65 hours

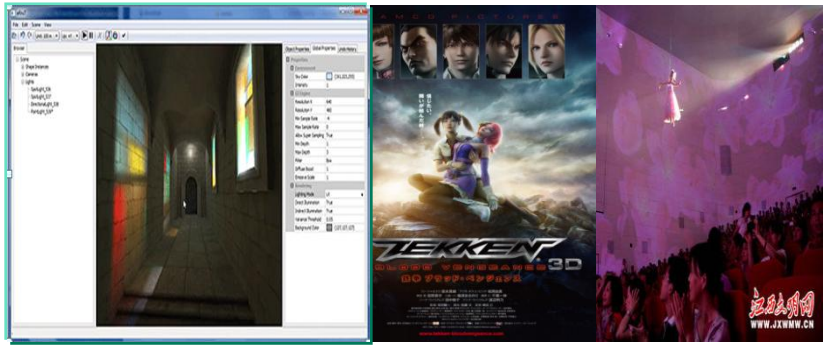
- Biology and life science
  - Structure-Function Relationship of Drug Targets
  - Gene analysis and sorting
    - Human gene HR-analysis (500samples, 5 hours)
    - Rice gene sorting (2000 samples, 3days)



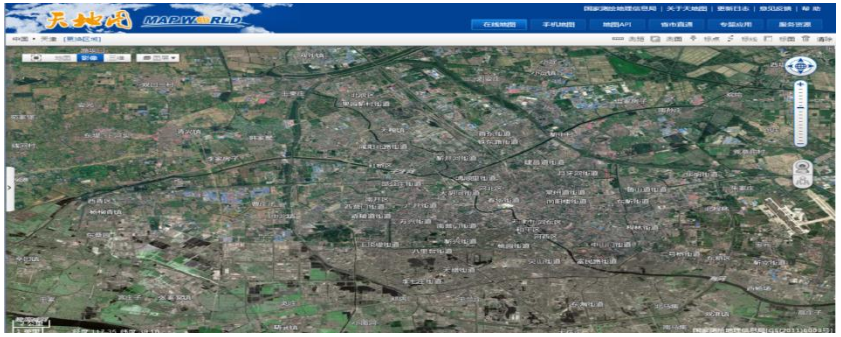
Case study

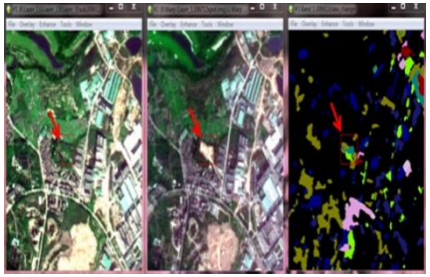
Tianhe 天河

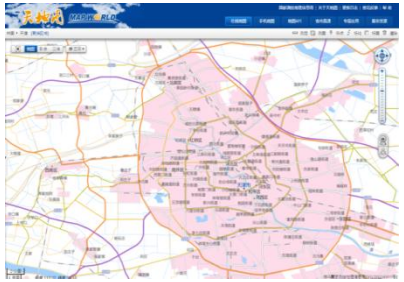
- Animation and image
  - Super rendering capability, 12,000 cores (expect up to 80,000 cores)
  - To build the cloud digital stunt syntheses island and the game middleware of distributed computing system



- Geographic information system
  - Massive geographic information data storage and processing
  - National basic geographic information center
  - Geographic monitor and investigation
  - Contingency responding support system
  - Sky-ground map services
  - 3D city services

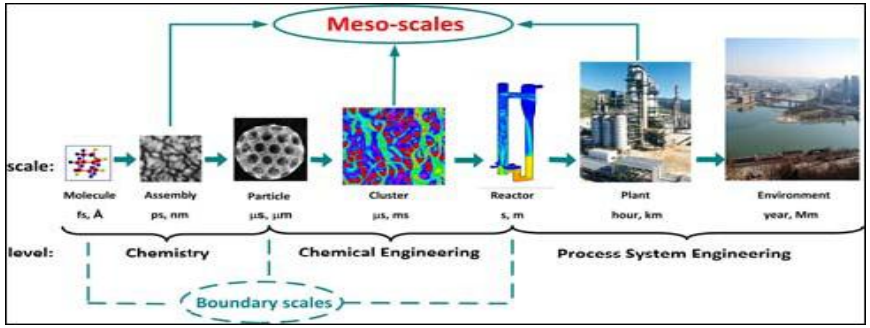




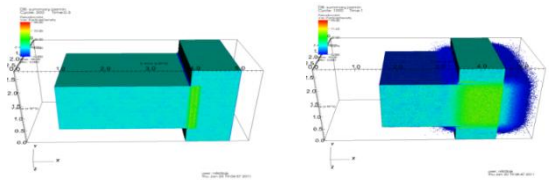


Case study

- **Trans-scale Simulation of Silicon Deposition Process**
  - scalable bond-order potential (BOP) for the molecular dynamics simulation of crystalline silicon
  - 1.17Pflops in SP plus 92.1Tflops in DP on 7168 GPUs and 86016 CPUs
  - 1.87Pflops in single precision (SP) on 7168 GPUs
    - 25.3% of its peak performance
    - 80% of its instruction throughput

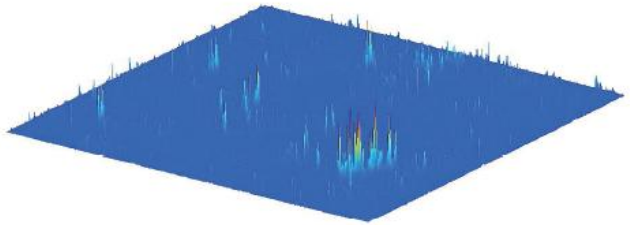


- **High speed collision system**
  - Force calculation is accelerated by GPU
  - 21.9x speedup on a single GPU compared to a single CPU core
  - Excellent weak and strong scalability with up to 4096 nodes (106,496 cpu/gpu cores) for problems with up to 11.16 billion atoms on Tianhe-1A
  - Embedded Atom Method potential. scale to the whole system is expected

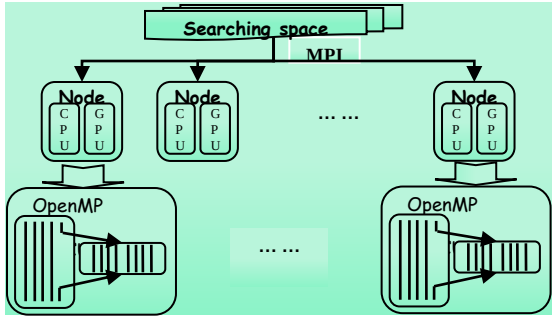


Case study

- **Direct Numerical Simulation of Turbulent Flow**
  - GPU-accelerated FFT solver (PKUFFT)
  - Taylor micro-scale Reynolds number up to 1164
  - with the grid resolution up to 8192^3
  - 7168nodes, >3.2million cuda cores(>100,000 gpu cores)
  - 30TFlops FFT sustained performance（SP）



- **Crypt cracker, Brute-force attack**
  - Number of passwords checked on single node
    - Without GPU, 50Kilo/s, With GPU, 250Kilo/s
  - Whole system（186368 cores）lineal scalable
  - Number of passwords checked on Tianhe-1A
    - 1.8 Billion per second

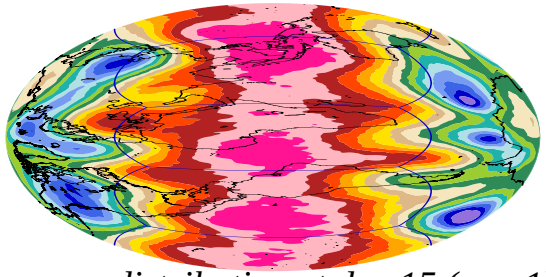


Key length	6	7	8	9
Alpha & Digits	16s	16.7m	17.3h	44.5d

The expected time for a successful attack on Tianhe-1A

Case study

- Climate
  - Evolution of global climate, human surviving environment
  - Marine environment, economical value
  - Global shallow water model
    - On up to 3,750 nodes (45,000 CPU cores + 52,500 GPU cores)
    - Aggregate performance 809 Tflops (32% of peak) in double-precision

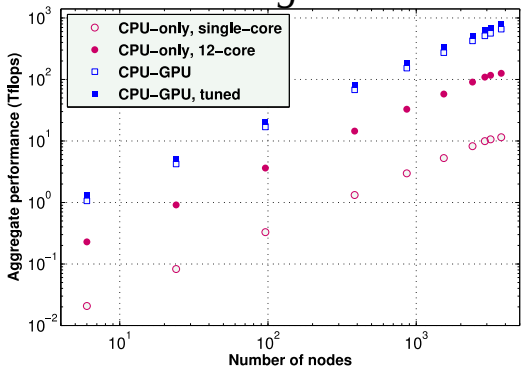


Pressure distribution at day 15 (res: 1km) using real topography data of the Earth

Strong scaling results

#nodes	384	1536	2400	3750
Time (s)	56.9	14.4	9.4	5.9
Efficiency	1.00	0.99	0.97	0.98
Agg. Tflops	84.5	335.1	513.4	809.6

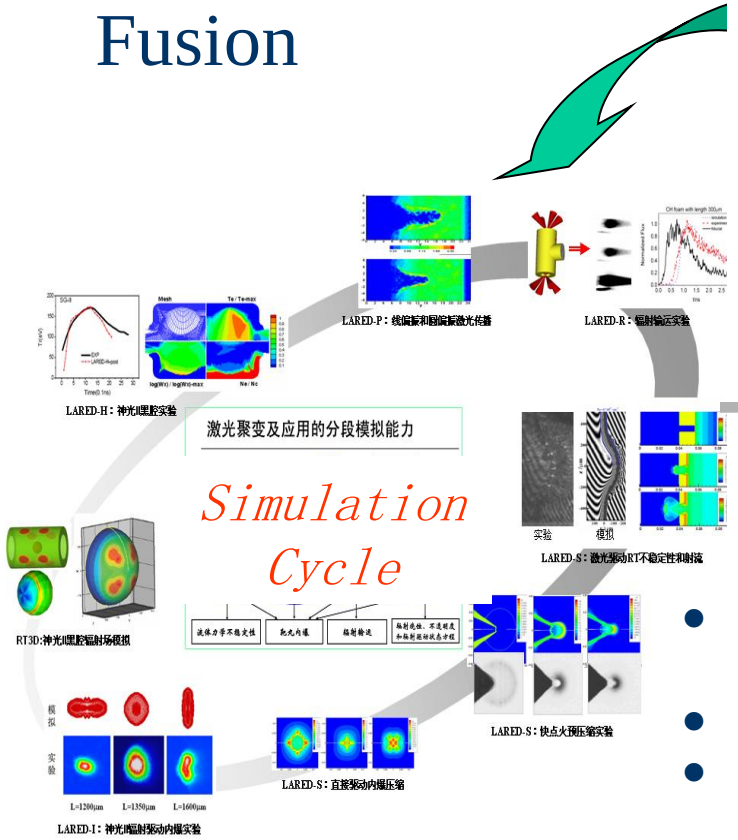
Weak scaling results

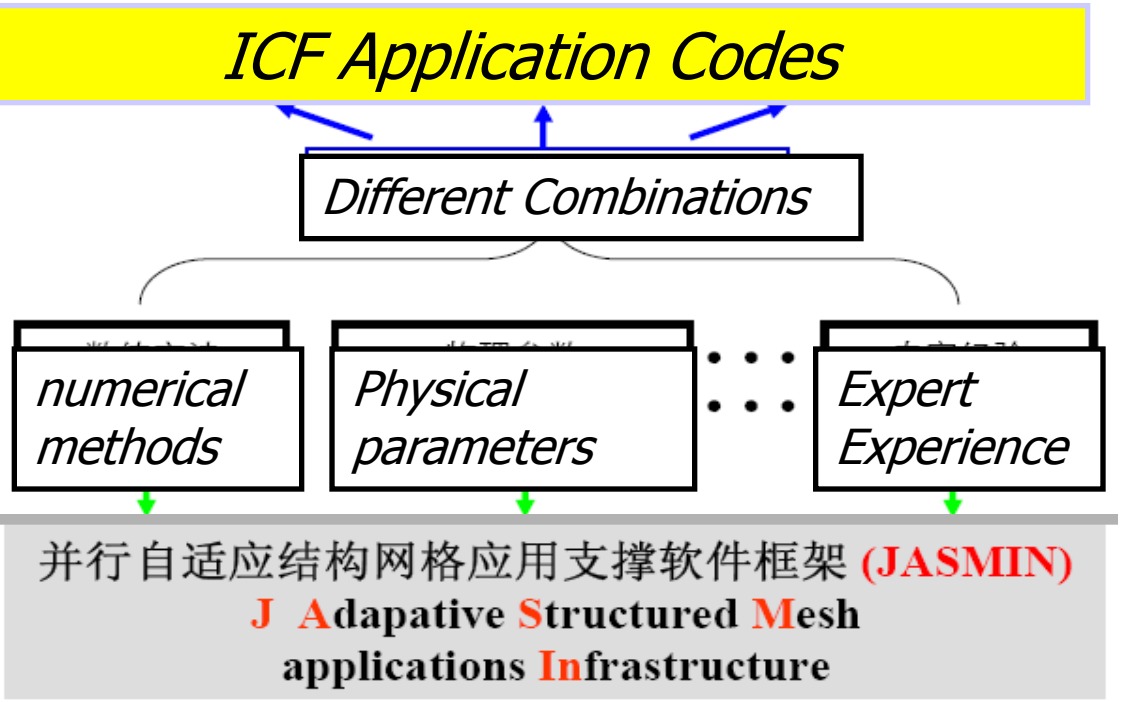


--from Dr. Chao Yang, Chinese Academy of Sciences

Case study

- Inertial Confinement Fusion





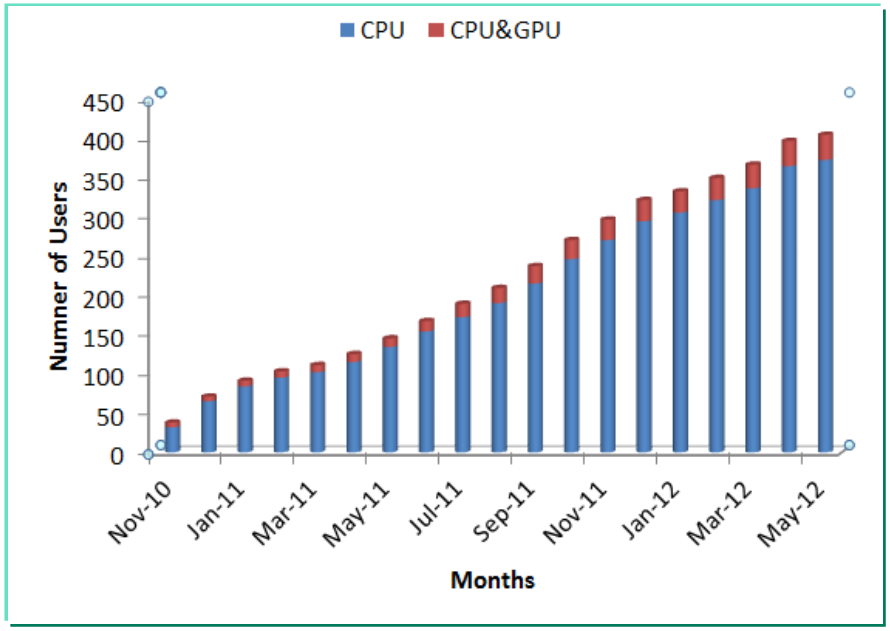
- Hides parallel computing and adaptive implementations using tens of thousands of CPU cores;
- Provides efficient data structures, algorithms and solvers;
- Support software engineering for code extensibility.

Program	Application domain	#Cores	Parallel efficiency
LARED-P	Inertial confinement fusion	84,000	73%
LAP3D	Inertial confinement fusion	42,000	63%
LARED-S	Inertial confinement fusion	84,000	47%
MD3D	Material science	84,000	47%
FDTD3D	High-power microwave	42,000	21%

--from Prof. Zeyao Muo, IAPCM, China

Status of Application

- Changes of Parallel application Scale
  - CPU + GPU
    - >30 efficient apps now
    - Maximum >80,000 cpu core+ 100,000 gpu core)
    - 10s apps on the way
  - Climate
  - Energy
  - Material
  - Biology and life science
  - Seismic data processing
  - CFD
  - Animation

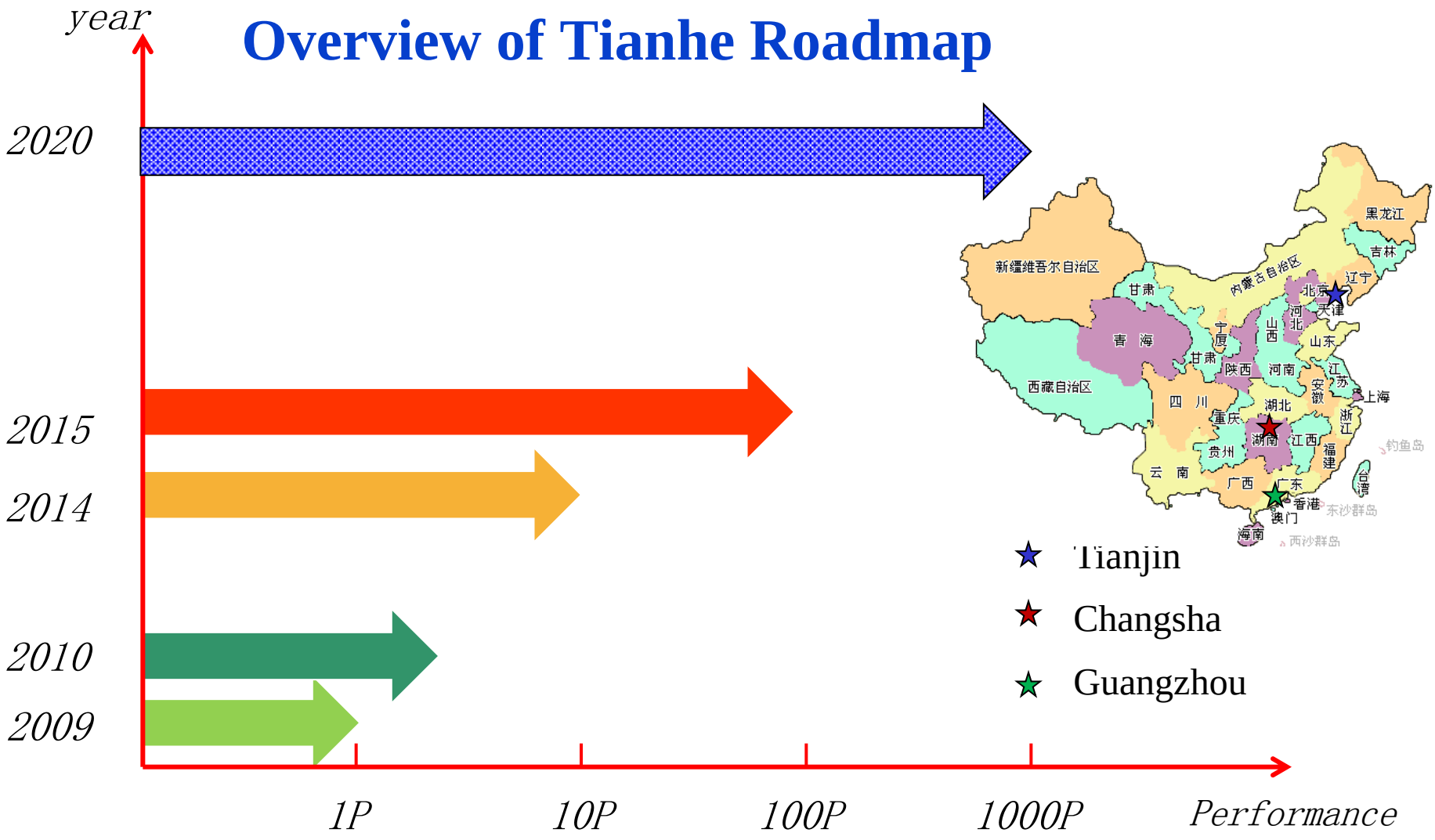


Status of Application

- Our principle for HPC
  - Practicality and Usability
  - Mature technology，correctness and functionality
  - Optimization technique，improve performance , scalability and reliability
- Need to improve
  - New GPU-like architecture
    - Data moving
    - Memory/cache architecture
    - Flexible execution mode
  - New programming model and tools
  - Change minds
    - Rethink the basic application algorithm
    - Redesign the parallel programs

Prospect of Next-Gen Tianhe

Overview of Tianhe Roadmap

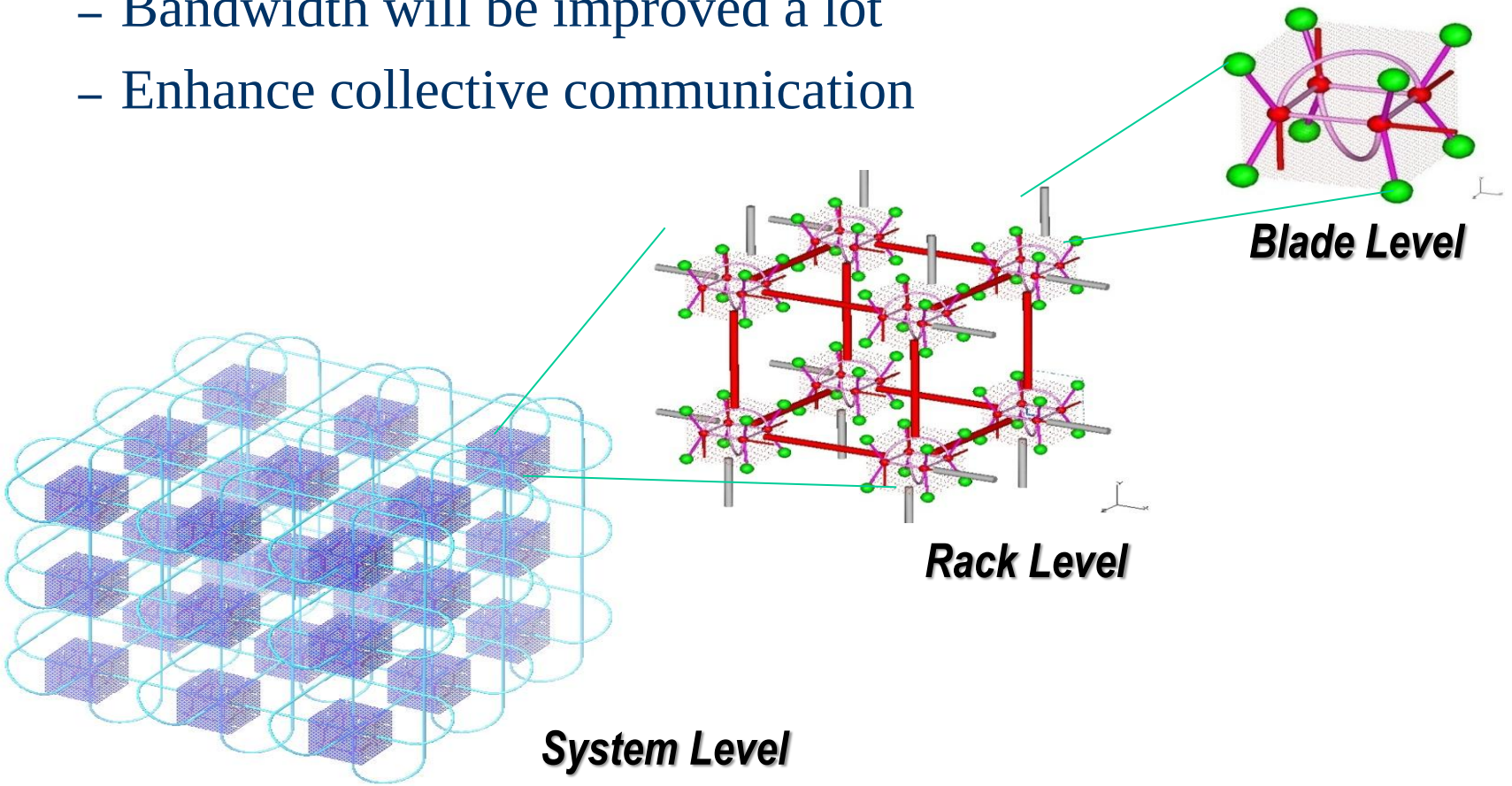


Prospect of Next-Gen Tianhe

- Highlights of next generation of Tianhe system
  - Heterogeneous parallel architecture
  - Multiple-dimension interconnection network
  - Hierarchy I/O storage system
  - Autonomic fault tolerant management
  - Adaptive power aware computing
  - Domain specific programming framework

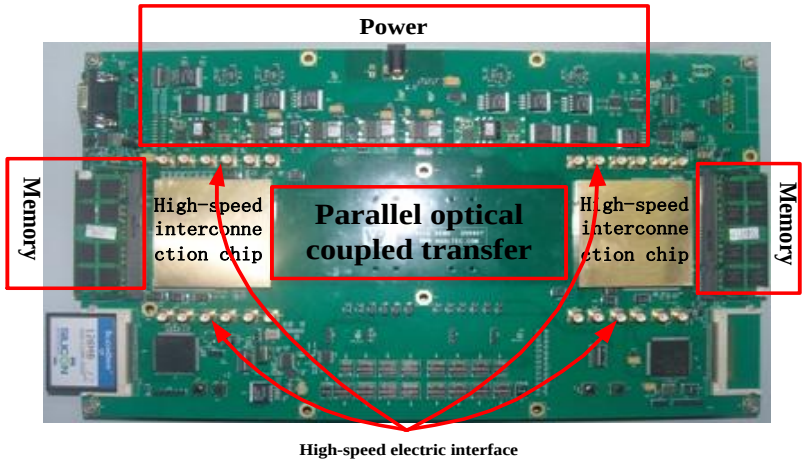
Prospect of Next-Gen Tianhe

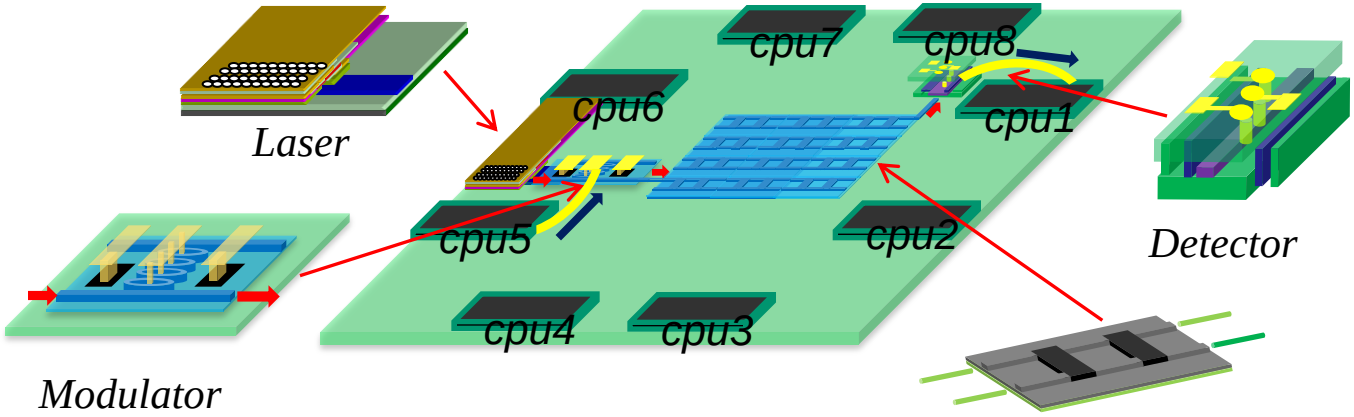
- Multiple-dimension interconnection network
  - Support more than 100,000 nodes
  - Bandwidth will be improved a lot
  - Enhance collective communication



Prospect of Next-Gen Tianhe

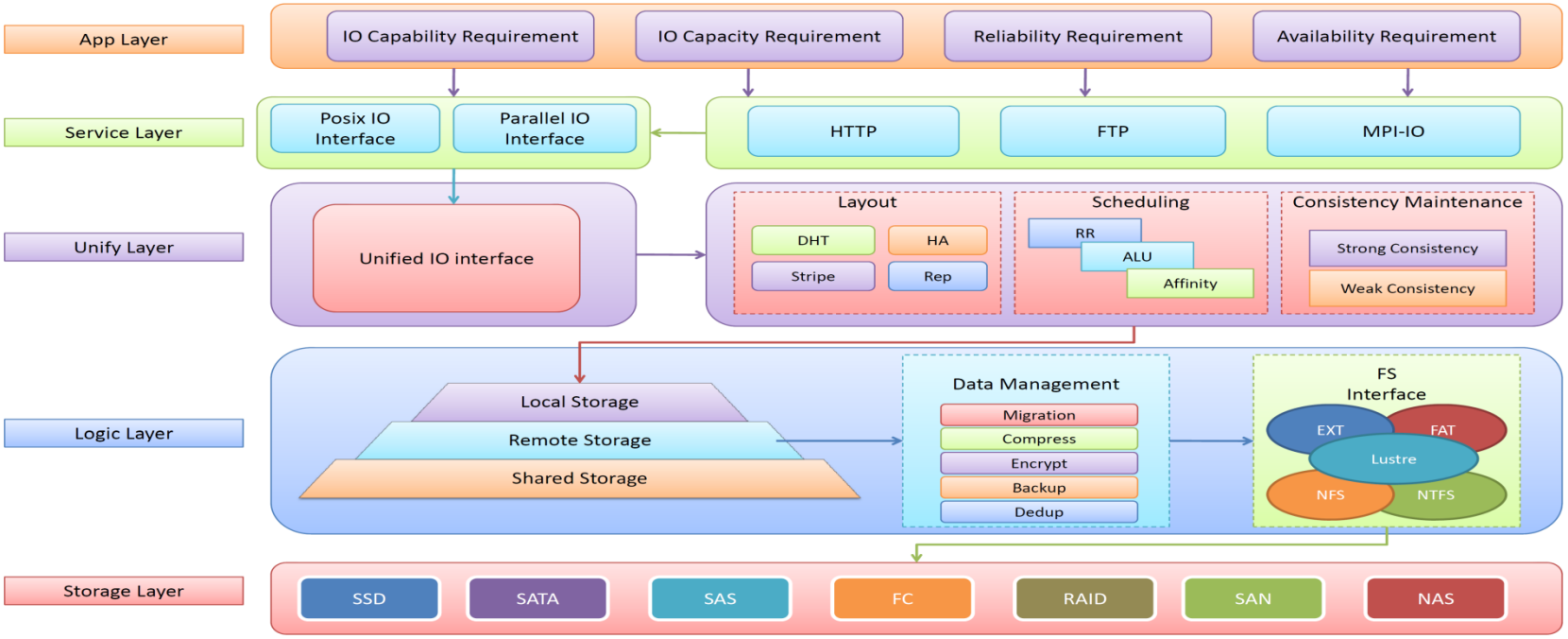
- Inter-chip optical connection
  - Optical interface between processors
  - Optical switch between CPUs is under research





Prospect of Next-Gen Tianhe

Large-scale Hybrid Tiered File System Architecture

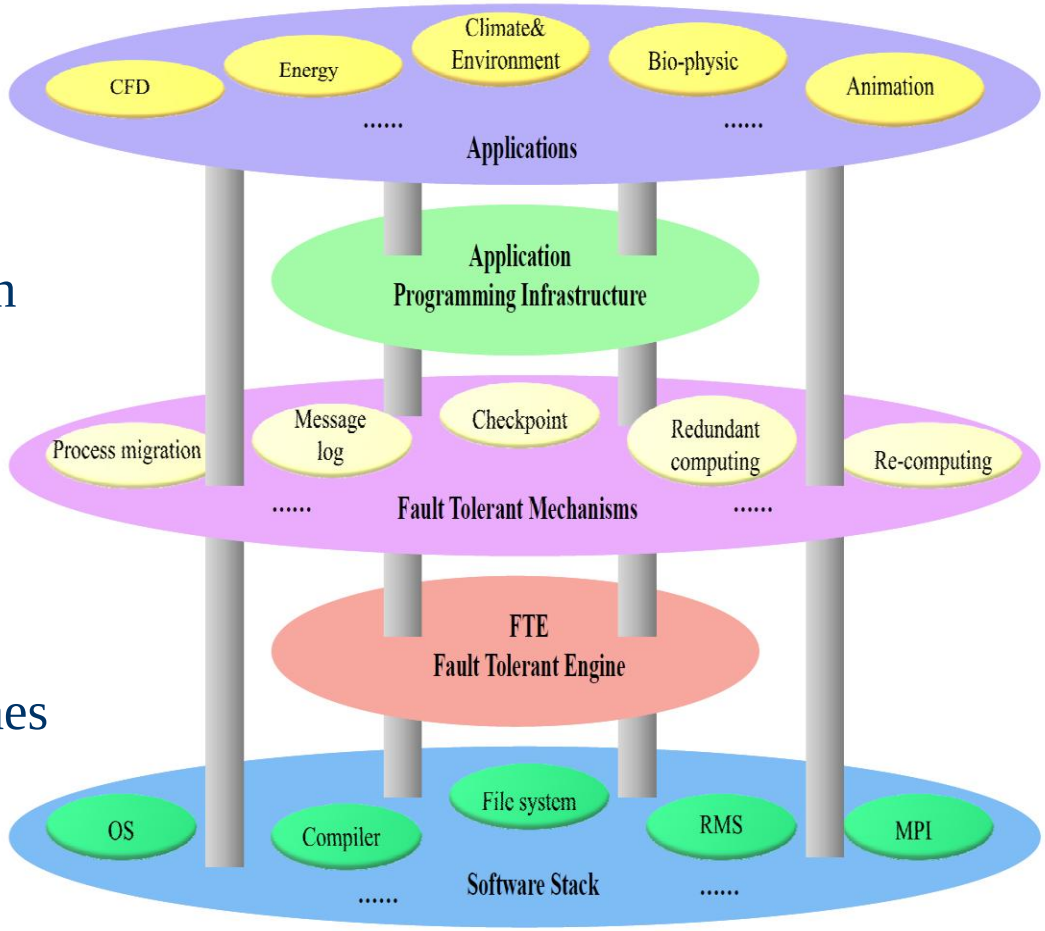


- Scalability to achieve >1TB/s I/O bandwidth by leveraging spatial locality
- Usability by federating multi-level storage into unified name space
- Flexibility by key components re-configuration for application optimization
- Applicability for supercomputers and clusters with hybrid infrastructure

Prospect of Next-Gen Tianhe

Resilience computing Framework

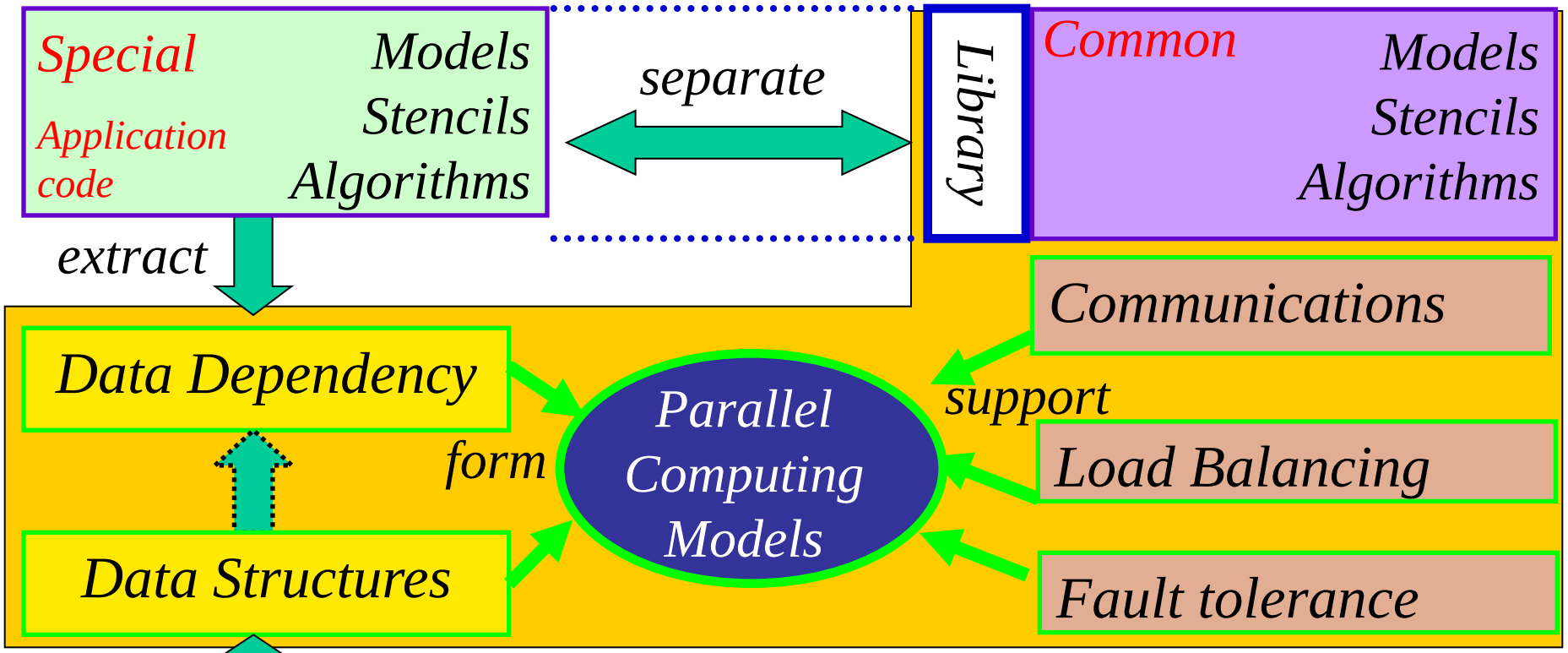
- Capability to support extreme large scale computing
- Collaboration with whole system software stack
- Coherent fault detection
- Coordinate fault tolerant decision
- Cooperation of multiple fault recovery mechanics
- Combination of proactive and reactive strategies
- Customizable fault detection, prediction and recovery approaches
- Support various parallel models



Prospect of Next-Gen Tianhe

Tianhe 天河

Parallel programming Framework

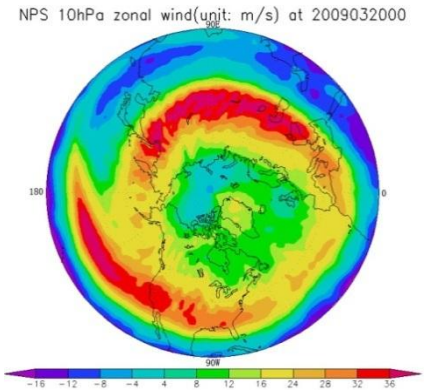


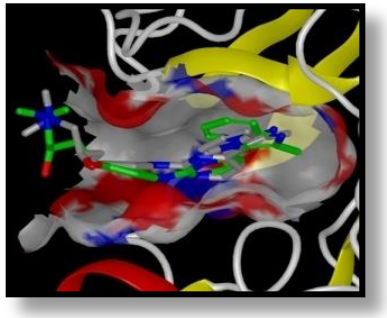
Hide the complexity of programming(mil cores/hybrid)
Integrate the efficient parallel fast numerical algorithms
Provide efficient data structures and solver libraries
Support software engineering for code reusing *USERS*

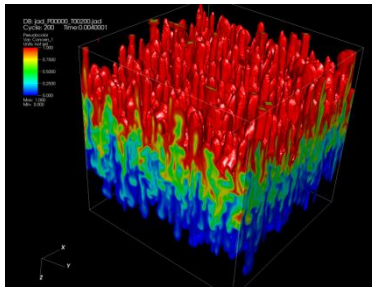
Optimization interface
Heterogeneous issues
Communication organization
Load balance
Fault tolerance *System software developers*

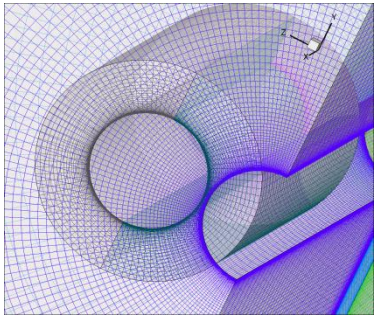
Prospect of Next-Gen Tianhe

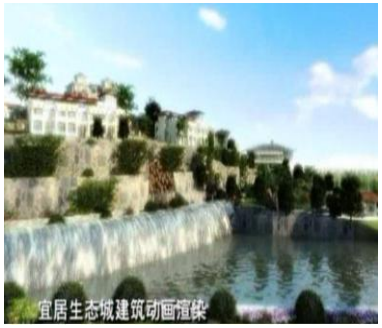
- Five priority areas
  - Climate & Environment
  - Bio-medicine
  - New energy
  - Civil engineering
  - Animation











Summary

Tianhe 天河

- Towards next generation of Tianhe system
  - Heterogeneous architecture
  - New enabling technology
  - High performance scalable interconnection
  - Balance the computation and data access
  - Feasible fault tolerant and power management mechanics
  - Usable domain-specific programming framework
  - Selected priority application areas