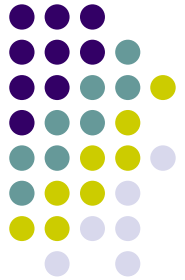
HPC System in NUDT

Prof. Lu Yutong

National university of Defense Technology, China

email: ytlu@nudt.edu.cn

Outline



- Background
- Tianhe-1A brief
- Current system software stack
- Exascale computing

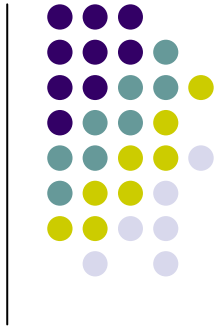
National University of Defense Technology

8 Colleges/Schools

Located in Changsha, Hunan Province

- Aerospace and Material Engineering
- Science
- Mechanics Engineering and Automation
- Electronic Science and Engineering
- Information System and Management
- Computer Science
- Photo-Electronic Science and Engineering
- Social Sciences

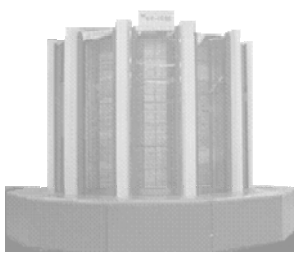
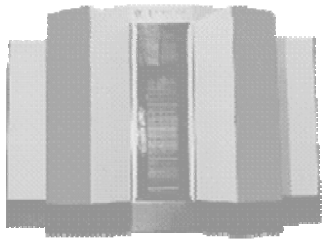
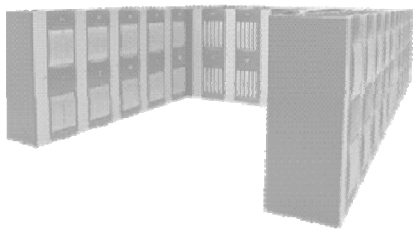
School of Computer Science



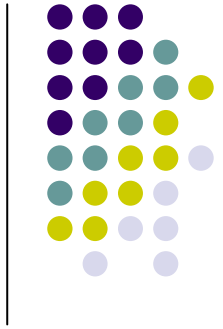
- Scientific Research Directions
  - High Performance Computing
    - Tianhe-1/Tianhe-1A super computer
    - Galaxy high performance computers
  - Mirco-processors
    - CPUs, and DSPs
  - System Software
    - Operating system, compiler, middleware
  - Network and communications
    - Galaxy high-speed interconnect network
    - Galaxy routers

School of Computer Science

- Achievements on High Performance Computing
  - 1983, Galaxy, 100Mflops, the First Supercomputer in China
  - 1992, Galaxy, 1Gflops,the First Gflops supercomputer in China
  - 2000, Galaxy, 1Tflops, the First TFlops supercomputer in China
  - 2009, Tianhe-1, 1.2Pflops, Top5
  - 2010, Tianhe-1A, 4.7Pflops, Top1


TH-1A system



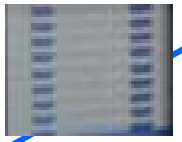
- Hybrid MPP architecture: CPU & GPU
- Custom software stack
- Peak performance 4.7PF, Linpack 2.57PF, No.1 Top500
- Power consumption 4.04MW(635.15MF/W), No.12 Green500

Items	Configuration
Processors	14336 Intel CPUs + 7168 nVIDIA GPUs + 2048FT CPUs
Memory	262TB in total
Interconnect	Proprietary high-speed interconnecting network
Storage	Global shared parallel storage system, 2PB
Cabinets	120 Compute / service Cabinets
	14 Storage Cabinets
	6 Communication Cabinets

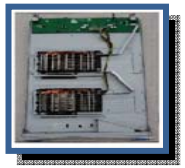
From chips to Entire system

TH-1A System

Cabinet
(4 x rack)



rack
(16 x cn)



Compute node

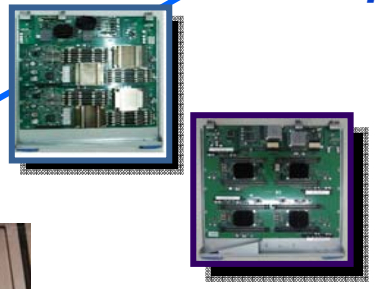
Twin GPU blade

M2050

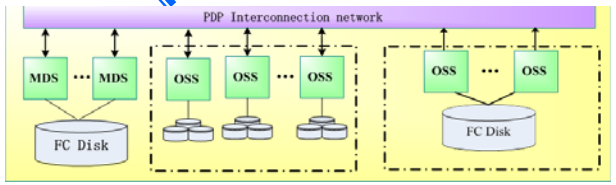
X5670

FT-1000
8c*8t,1GHz

Chips



Quad cpu blade



On-line storage

Hardware

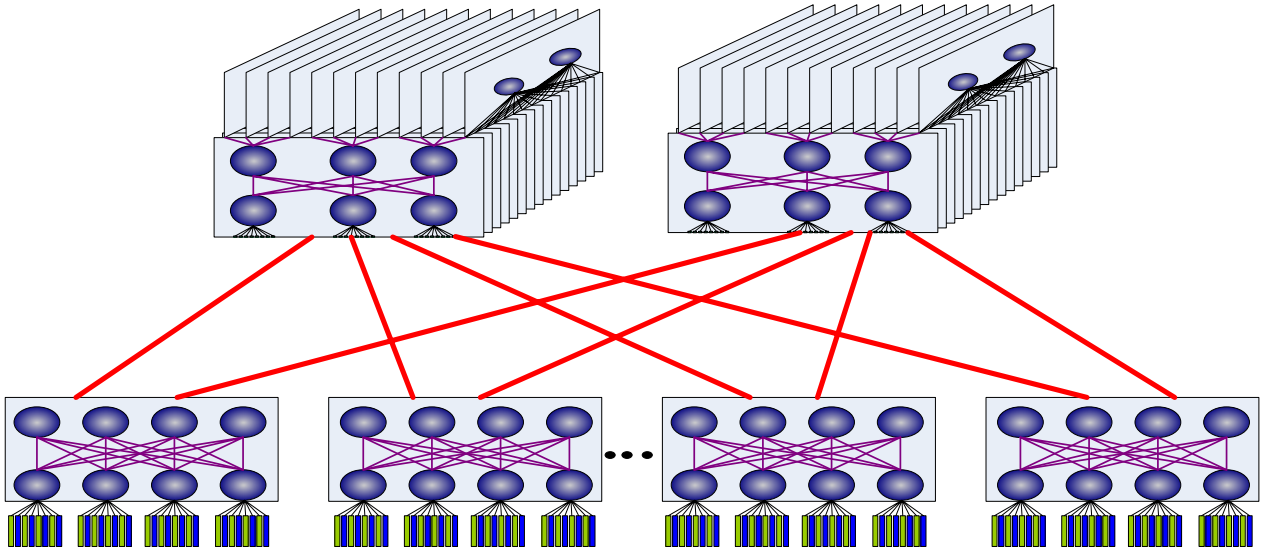
- 3 kinds of LSI chips
  - CPU: FT-1000(PSoC)
  - High radix router ASIC：NRC
  - Network interface ASIC：NIC

- 15 kinds of PCB boards

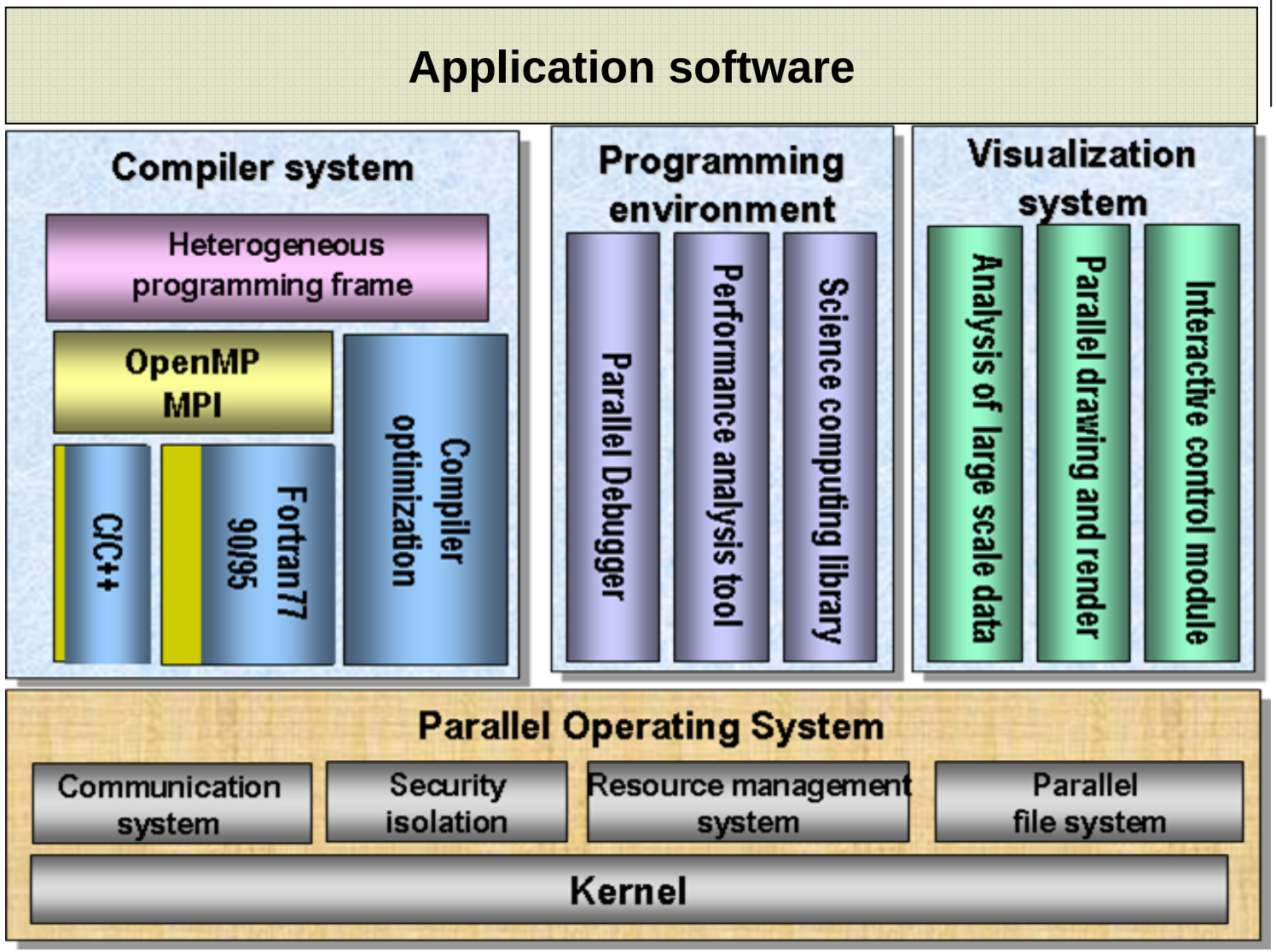
- 4 kinds of nodes, 2 sets of networks
- Custom cabinet
- water cooling system

Interconnection network

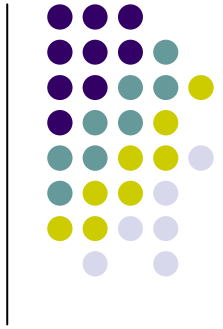
- Optimized Channel bonding (8 Lane x 10Gbps)
- bi-BW– 160Gbps (2xIB QDR)
- Topology: Hierarchy fat-tree structure
  - First stage: 16 nodes connected by 16-port switching board
  - Second stage: all parts connected to eleven 384-port switches



TH-1A software stack

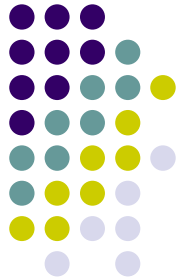


What’s the point



- Customization
- Optimization

Operating system

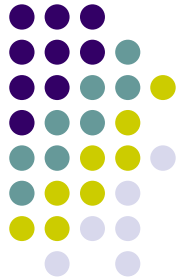


- Kylin Linux
- compute node kernel
- Provide virtual running environment
  - Isolated running environments for different users
  - Custom software package installation
- QoS support
- Power aware computing

Glex communicating system

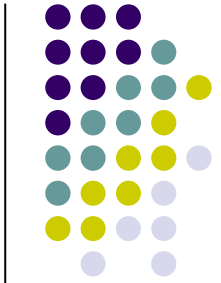
- Proprietary Interconnection based on high radix router
- High bandwidth packet and RDMA communication
  - Zero copy user space RDMA
  - MPI base on GLEX: Bandwidth 6.3GB/s
  - Accelerate collective operation with hardware support in communication interface
- Fault tolerance
  - Rapid error detection in large scale interconnection
  - Rebuild communication links

Resource management system



- Resource management, job scheduler (slurm based)
  - Heterogenous resources management and topology-aware scheduling
- Large scale parallel job launcher (custom)
  - Improve system structure, optimized protocol, network performance, file system performance
  - **logN, klogN**（**whole system HPL launching time less than 2 mins)**
- System power management
- Automatic CR supporting
- Accounting Enhance

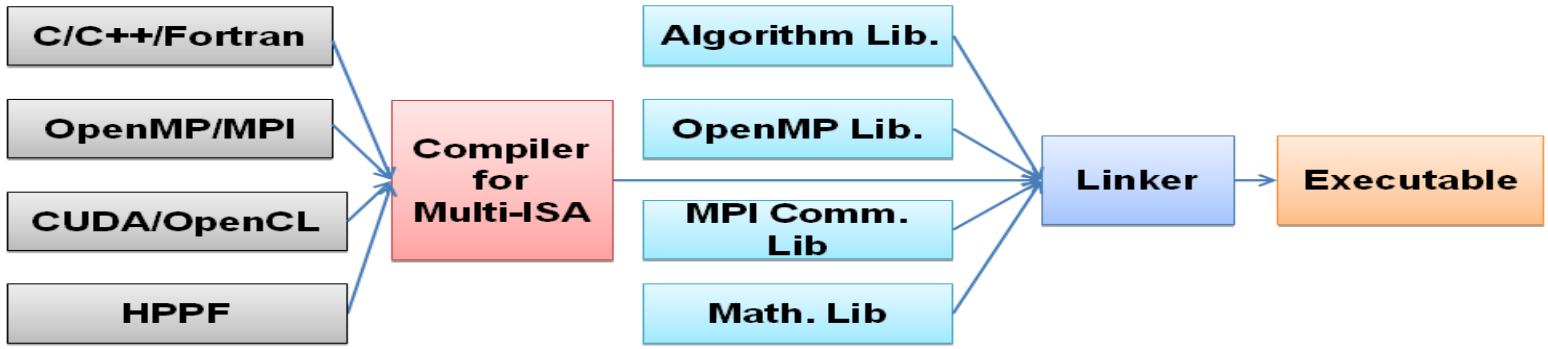
Global parallel file system



- Object storage architecture (Lustre based)
- Capacity: 2 PB, Scalability: clients>8192, oss>128
- Performance: Collective BW (IOR): >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

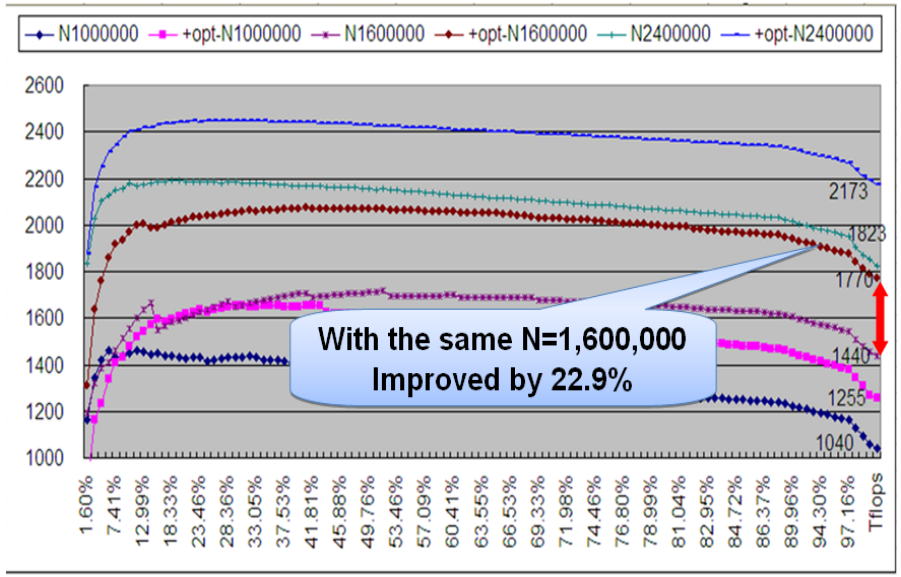
Compiler system

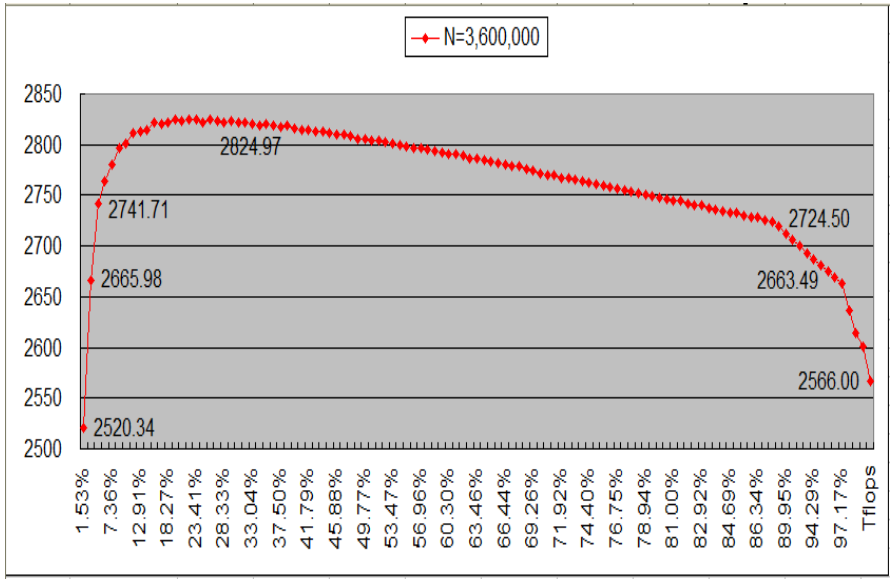
- C, C++, Fortran, Java
- OpenMP, MPI, OpenMP/MPI
- CUDA, OpenCL
- Heterogeneous programming framework
  - Accelerate the large scale, complex applications,
  - Use the computing power of CPUs and GPUs, hide the GPU programming to users
  - Inter-node homogeneous parallel programming (users)
  - Intra-node heterogeneous parallel computing (experts)



Compiler Optimized

- Accelerating HPL (MPI(custom)+OpenMP+Cuda)
  - Adaptive partition
  - Asynchronous data transfer
  - Software pipeline
  - Affinity scheduling
  - Zero-copy





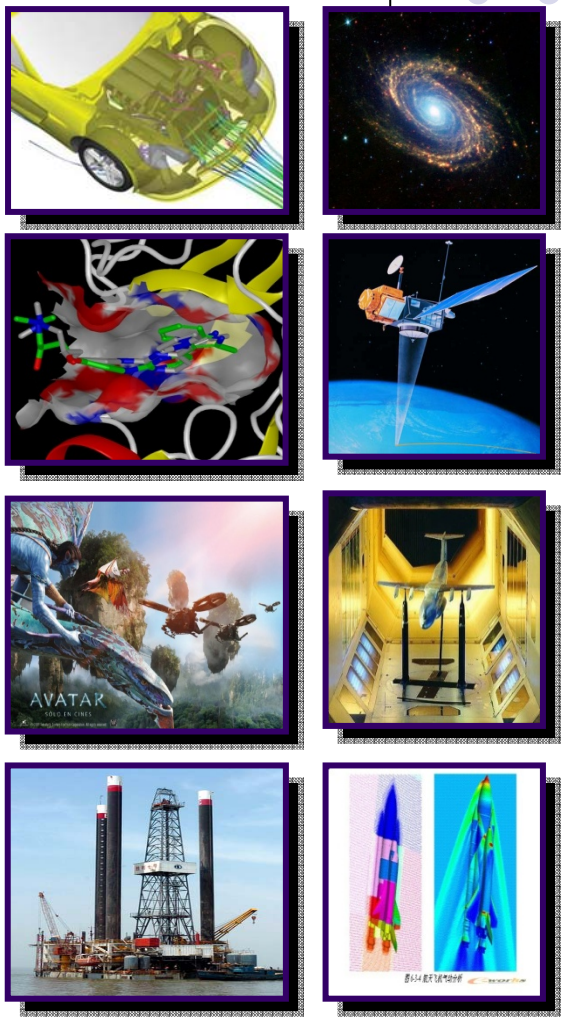
Ratio: 54.6%

Programming environment

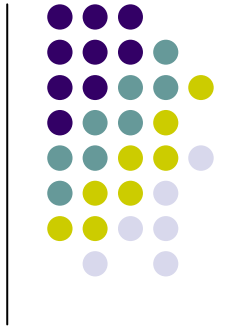
- Virtual running environments
  - Provide services on demand
- Parallel toolkits
  - Based on Eclipse
  - To integrate all kinds of tools
  - Editor, debugger, profiler
- Work flow support
  - Support QoS negotiate

Applications

- Weather and climate forecasting
- Oil exploration
- Bio-medical research
- High-end equipment development
- New energy research
- Animation design
- New material research
- Engineering design, simulation and analysis
- Remote sensing data processing
- Financial risk analysis
- ……

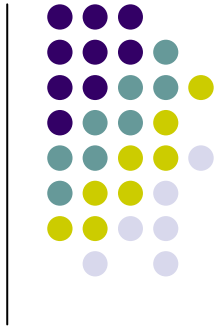


Current

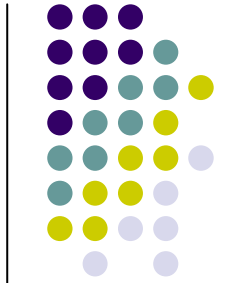


- Our principle
  - Practicality and Usability
- HPC system in NUDT
  - Mature technology: correctness and functionality
  - Optimization technique: improve performance , scalability and reliability
  - Long-term accumulation
    - Various architectures
    - Various programming models and frameworks
    - Various applications supporting

Exascale computing



- Key issues of System software
  - Fault tolerance
  - Scalability
  - Power consumption
- Way…
  - Whole stack solution and optimization
    - Theory
    - Technique
- Goal…
  - Implementation of platform independent software system



Thanks