Issues in developing the new generation high performance computers

Depei Qian
Beihang University
ICPP 2015，Beijing
Sep. 4, 2015

Outline

- A brief review
- Issues in developing next generation supercomputers
- Prospects

A brief review

The 863 program

- The most important high-tech R&D program of China since 1987
- Proposed by 4 senior scientists and approved by former leader Deng Xiaoping in March 1986
- A regular R&D program, named after the 5-year plan, current the 12th 5-year 863 program
- 8 areas, Information Technology is one of them
- Strategic, looking-ahead, frontier research on major technologies supporting China's development
- Emphasize technology transfer and adoption of research outcomes by industry
- Encourage enterprise participation

Evolution of 863's emphasis

- 1987: Intelligent computers
  - Influenced by the 5th generation computer program in Japan
- 1990: from intelligent computer to high performance computers
  - Emphasize practical HPC capability for research and industry
- 1998: from high performance computer system to HPC environment
  - Emphasize resource sharing and ease of access
  - Broaden usage of the HPC systems

Evolution of 863's emphasis

- 2006: from high performance to high productivity
  - Emphasize other metrics such as programmability, program portability, and reliability besides peak performance
- Current: from HPC environment to HPC application service environment
  - Emphasize integrated efforts on HPC systems, HPC environment, and HPC applications
  - Explore new mechanisms and business models for HPC services
  - Promote the emerging computing service industry

Three key projects on HPC

- 2002-2005：High Performance Computer and Core Software (863 key project)
  - Efforts on resource sharing and collaborative work
  - Developing grid-enabled applications in multiple areas
  - Successfully developed TFlops computers and China National Grid (CNGrid) testbed
- 2006-2010：High Productivity Computer and Service Environment (863 key project)
  - Emphasizing other system features besides the peak performance
    - Efficiency in program development
    - Portability of programs
    - Robust of the system
  - Addressing the service features of the HPC environment
  - Successfully developed Peta-scale computers, upgraded CNGrid into the national HPC service environment

The three key projects on HPC (cont'd)

- 2010-2016：High Productivity Computer and Application Service Environment（863 key project）
  - Exploring new operation models and mechanisms of CNGrid
  - Developing cloud-like application villages over CNGrid to promote applications
  - Developing world-class computer systems
    - Tianhe-2
    - Sunway-NG

- Emphasizing balanced development of HPC systems, HPC environment, and HPC applications

HPC systems developed in the past 20 years

- 1993：Dawning-I，shared memory SMP, 640 MIPS peak
  - Dawning 1000: MPP, 2.5GFlops（1995）
  - Dawning 1000A：cluster（1996）
  - Dawning 2000：111GFlops（1999）
  - Dawning 3000：400GFlops（2000）
- 2003：Lenovo DeepComp 6800，5.32TFlops peak, cluster
- 1993-2003: performance increased 8000+ times

Dawning I

Dawning 1000

Dawning 2000

Dawning 3000

DeepComp6800

HPC systems developed in the past 20 years (cont'd)

- 2004：Dawning 4000A，Peak performance 11.2TFlops, cluster
  - Lenovo 7000,150TFlops peak, Hybrid cluster and Dawning 5000A, 230TFlops, cluster (2008)
  - TH-1A, 4.7PFlops peak, 2.56PFlops LinPack, CPU+GPU (2010)
  - Dawning 6000, 3Pflops peak, 1.27 PFlops LinPack, CPU+GPU (2010）
  - Sunway-Bluelight，1.07PFlops peak，796TF LinPack, Homogeneous, implemented with China's multicore processors (2011）
- 2013: Tianhe-2，54PFlops peak and 33.9PFlops LinPack, CPU+MIC accelerated architecture
- 2003-2013: performance increased 10000 times
- 84,000,000 times in 20 years (1,000,000 times in TOP500)

Dawning 4000

DeepComp 7000

Dawning 5000

Dawning 6000

TH-1A

Sunway-Bluelight

First phase of TH-2

- Delivered in May 2013
- Hybrid system
  - 32000 Xeon, 48000Xeon Phis, 4096 FT CPUs
- 54.9PF peak, 33.86PF Linpack
- Interconnect
  - proprietary TH Express-2
- 1.4PB memory, 12PB disk
- Power: 17.8MW
- Installed at the National Supercomputing Center in Guangzhou

Second phase of TH-2

- The implementation scheme of the second phase of TH-2 was evaluated and approved in July of 2014
  - Upgrading interconnect (completed)
  - Increasing No. of computing nodes (completed)
  - Upgrading computing nodes
    - Upgrade the accelerator from Knight Conner to Knight Landing
    - Change the ratio of CPU to MIC from 2:3 to 2:2

Second phase of TH-2 (cont'd)

- The scheme has to be changed because of the new embargo regulation of the US government
- Completion of the second phase will be delayed
- The final TH-2 has to rely on indigenous FT processors, a stimulation to the R&D on kernel technologies in China
- The development of the new FT processors is on going

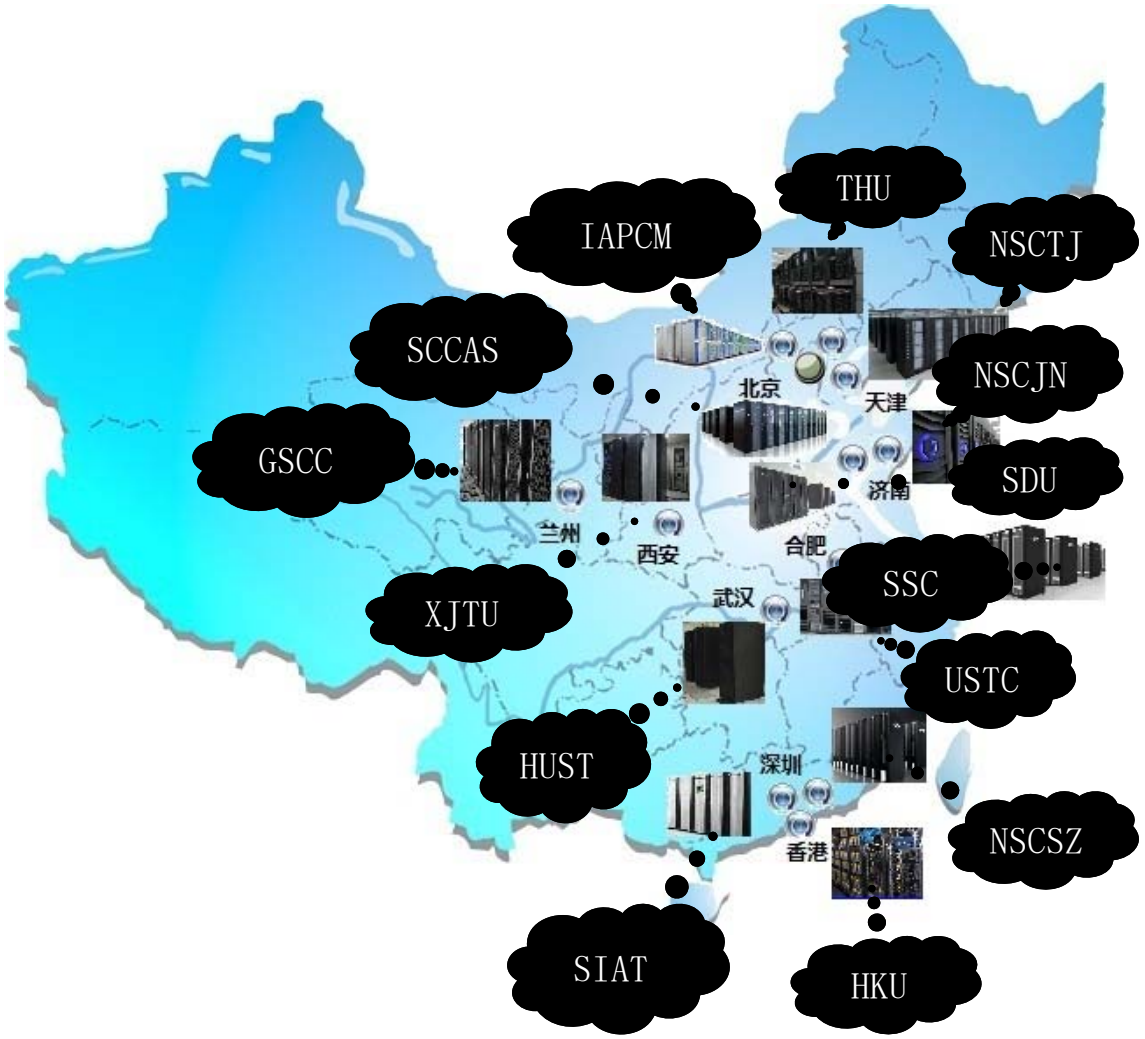
The second 100PF system

- The second 100PF system (Sunway-NG？) will be developed by the end of 2016
- A large system implemented with indigenous SW many-core processors in together with a smaller multicore system (1PF) implemented with commercial processors to meet the requirement of different applications
- The SW processor is under development

HPC environment development in the past 15+ years

- 1999-2000: National HPC Environment
  - 5 nodes
  - Equipped with Dawning computers
- 2002-2005: China National Grid (CNGrid), a testbed of new infrastructure
  - enabled by CNGrid GOS
  - 8 nodes
  - 18TF computing power
  - 200TB storage
- 2006-2010: CNGrid service environment, emphasizing service features
  - enabled by CNGrid Suite
  - 14 nodes
  - 8PF aggregated computing power
  - >15PB storage
  - >400 software and tools as services
  - supporting >1000 projects

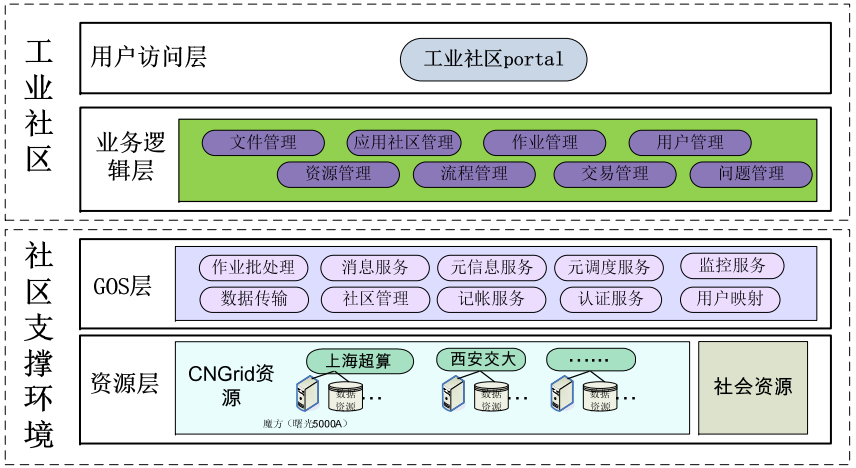
CNGrid sites

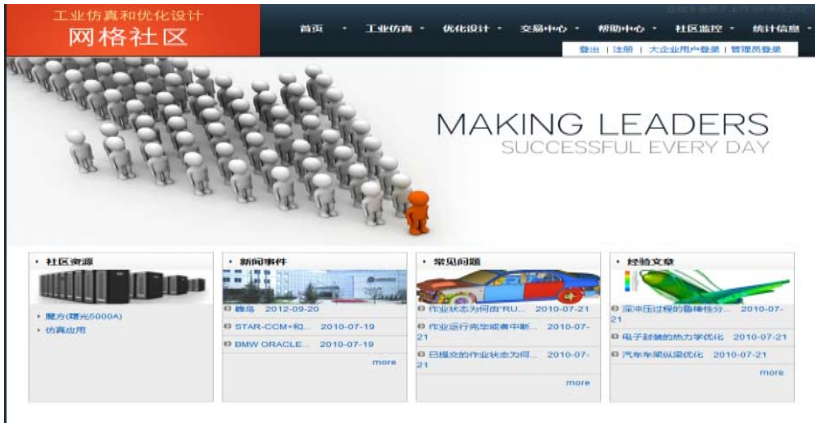


	CPU/GPU	Storage
SCCAS	157TF/300TF	1.4PB
SSC	200TF	600TB
NSC-TJ	1PF/3.7PF	2PB
NSC-SZ	716TF/1.3PF	9.2PB
NSC-JN	1.1PF	2PB
THU	104TF/64TF	1PB
IAPCM	40TF	80TB
USTC	10TF	50TB
XJTU	5TF	50TB
SIAT	30TF/200TF	1PB
HKU	23TF/7.7TF	130TB
SDU	10TF	50TB
HUST	3TF	22TB
GPCC	13TF/28TF	40TB

Current development of CNGrid (2011-2015)

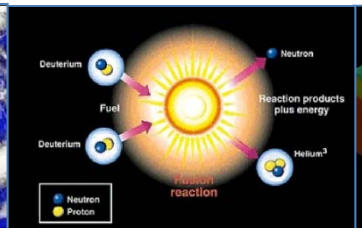
- CNGrid is treated as a layer of virtual resources available to the up-layers
- Establishing domain-oriented application villages (communities) on top of CNGrid, which provide services to the end users
- Developing business models and operation mechanisms between CNGrid and application villages
- Developing enabling technologies and platform supporting CNGrid transformation
- Application villages currently being developed
  - Industrial product design optimization
  - New drug discovery
  - Digital media

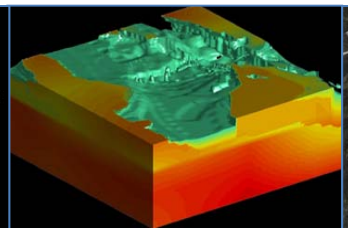
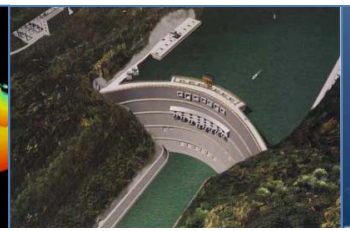


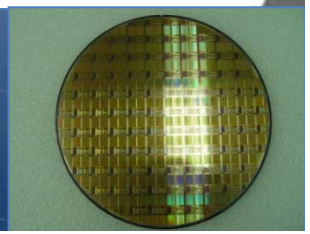


HPC applications in the past 15+ years

- 1999-2000: a set of demo-applications developed over the National HPC environment
- 2002-2005: productive applications
  - 11 grid-enabled applications in 4 selected areas
- 2006-2010: productive applications and large scale parallel software
  - 8 grid-enabled applications
  - 8 HPC-enabled applications
  - A parallel library
  - Up to 10,000-core level of parallelism achieved



Current development in parallel software (2011-2016)

- Application software development supported
  - Fusion simulation
  - CFD for aircraft design
  - Drug discovery
  - Rendering for Digital media
  - Structural mechanics for large machinery
  - Simulation of electro-magnetic environment
- Level of Parallelism required
  - Effective use of more 300,000 cores with >30% efficiency
- Must be productive systems in the real applications

Experiences

- Coordination between the national research programs and the development plans of the local government
  - Matching money for developing the computers
  - Joint effort in establishing national supercomputing centers
- Collaboration between industry, universities, research institutes, and application organizations
  - HPC centers played an important role in the development of high performance computers
  - Industry participated in system development
    - Inspur, Sugon, Lenovo actively participated in the development of PF- and 50PF-scale high performance computer systems
  - Application organizations led the development of application software

Weakness identified

- Lack of some kernel technologies, a lesson learned from the recent embargo
  - more R&D required on kernel technologies
  - stronger emphasis on self-controllable technologies
  - ecosystem for technology development is crucial
- HPC application is weak
  - rely on imported commercial software, also affected by embargo
  - needs for developing software in key application areas
  - open-source software will become more important
- HPC environment development is not sustainable
  - lack of long-term funding
  - need new models and mechanisms, unique condition in China
- Shortage of talents
  - need to develop new HPC-related curriculum in universities
  - continuous training during R&D

Issues in developing next generation supercomputers

Major challenges towards Exa-scale systems

- Major challenges
  - performance obtained by applications
  - power consumption limit
  - programmability
  - Resilience
- Addressing challenges by
  - Architectural support
  - Technical innovations
  - Hardware/software coordination

Constrained design principle

- Set up constraints on
  - Scale of the system
    - number of the nodes: 30,000-100,000
  - Footprint of the system
    - number of the cabinets < 300
  - Power consumption
    - energy efficiency of 30-50GF/W
  - Cost
    - relatively constant for every generation of supercomputers

Co-design based on application features

- Understand the workload characteristics of typical exa-scale applications
  - earth system modeling, fusion simulation, turbulence simulation, materials simulation, bio-informatics data analysis,
- Co-design based on application characteristics
  - propose architecture appropriate for major applications
  - Look for architectural support to major algorithms
- Develop metrics and benchmarks to understand how well the architecture adapts to the applications

1. Architecture

- Classifying architectures using “homogeneity/heterogeneity” and “CPU only/CPU+Accelerator”
- Homo-/Hetero refers to the “ISA”

	CPU only	CPU+Acc
Homogeneous	Sequoia K-computer Sunway/BL	Stampede TH-2
Heterogeneous	Dawning 6000/HPP (AMD+Loonson)	TH-1A TITAN Dawning 6000/Nebulae, Tsubame 2.5

1. Architecture consideration

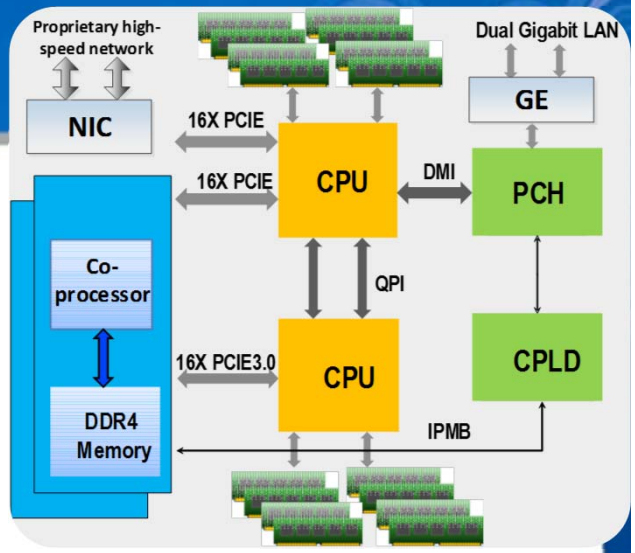
- Accelerated architecture vs homogeneous many-core architecture
  - To meet the requirements of a wide range of applications
  - Pairwise CPU/accelerator or stand-alone bulk accelerators?
    - CPU/GPU coordinated performance is only slightly higher than using GPU only in some applications
    - utilization of accelerator resources is sometimes low
- Reconfigurable architecture
  - Take the advantages of both special purpose or general purpose
  - Static or dynamic reconfigurable
  - Languages and tools to support reconfiguration are crucial

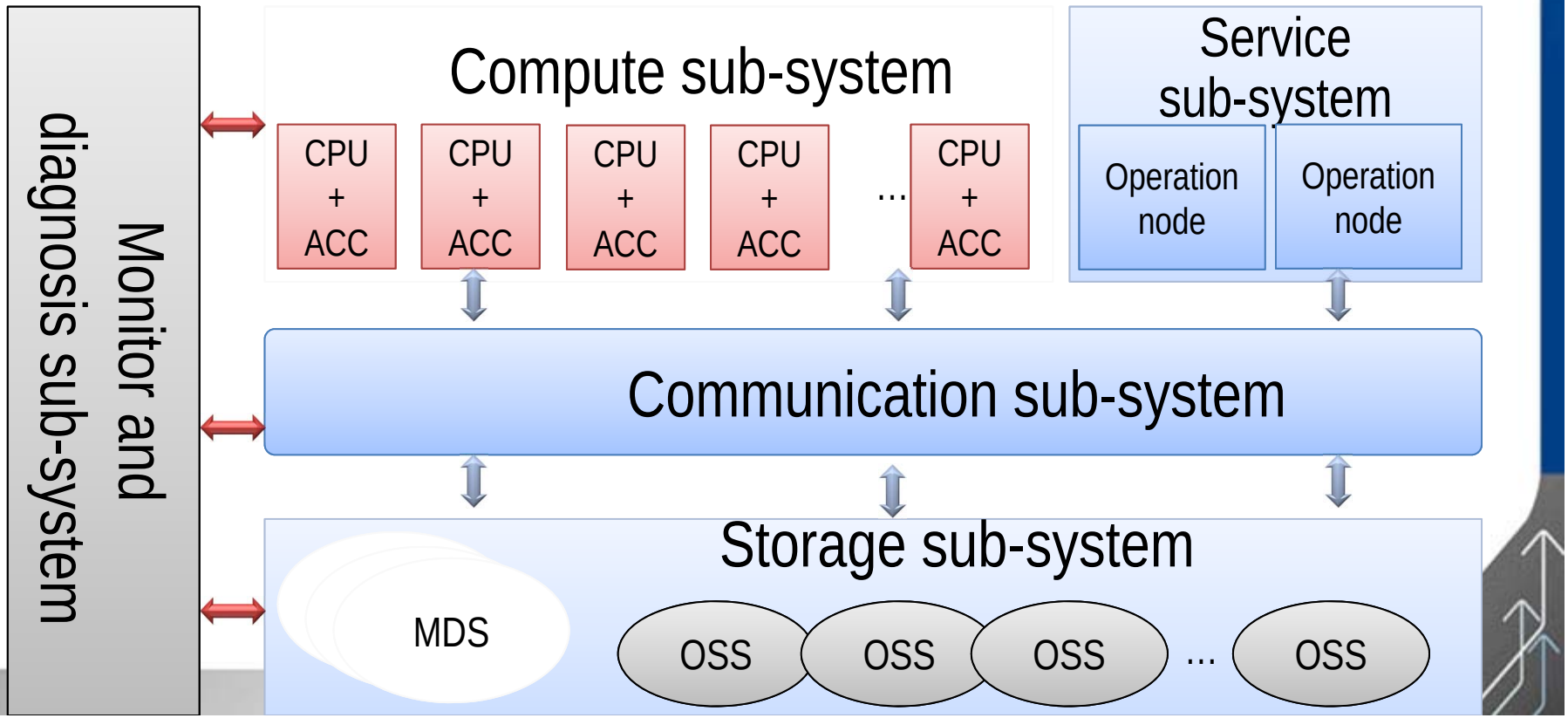
Architecture Designs

- Making trade-offs between performance, power consumption, programmability, resilience, and cost
  - Hybrid architecture (TH-1A & TH-2)
    - General purpose + high density computing (GPU or MIC)
  - HPP architecture (Dawning 6000/Loonson)
    - Enable different processors to co-exist
    - Support global address space
    - Multi-level of parallelism
  - Multi-conformation and Multi-scale adaptive architecture (SW/BL)
    - Cluster implemented with Intel processor for supporting commercial software
    - Homogeneous system implemented with domestic multicore processors for computing-intensive applications
    - Support parallelism at different levels

TH-1A/TH-2 architecture

- Hybrid system architecture
  - Computing sub-system (CPU+GPU/MIC)
  - Service sub-system
  - Communication networks
  - Storage sub-system
  - Monitoring and diagnosis sub-system

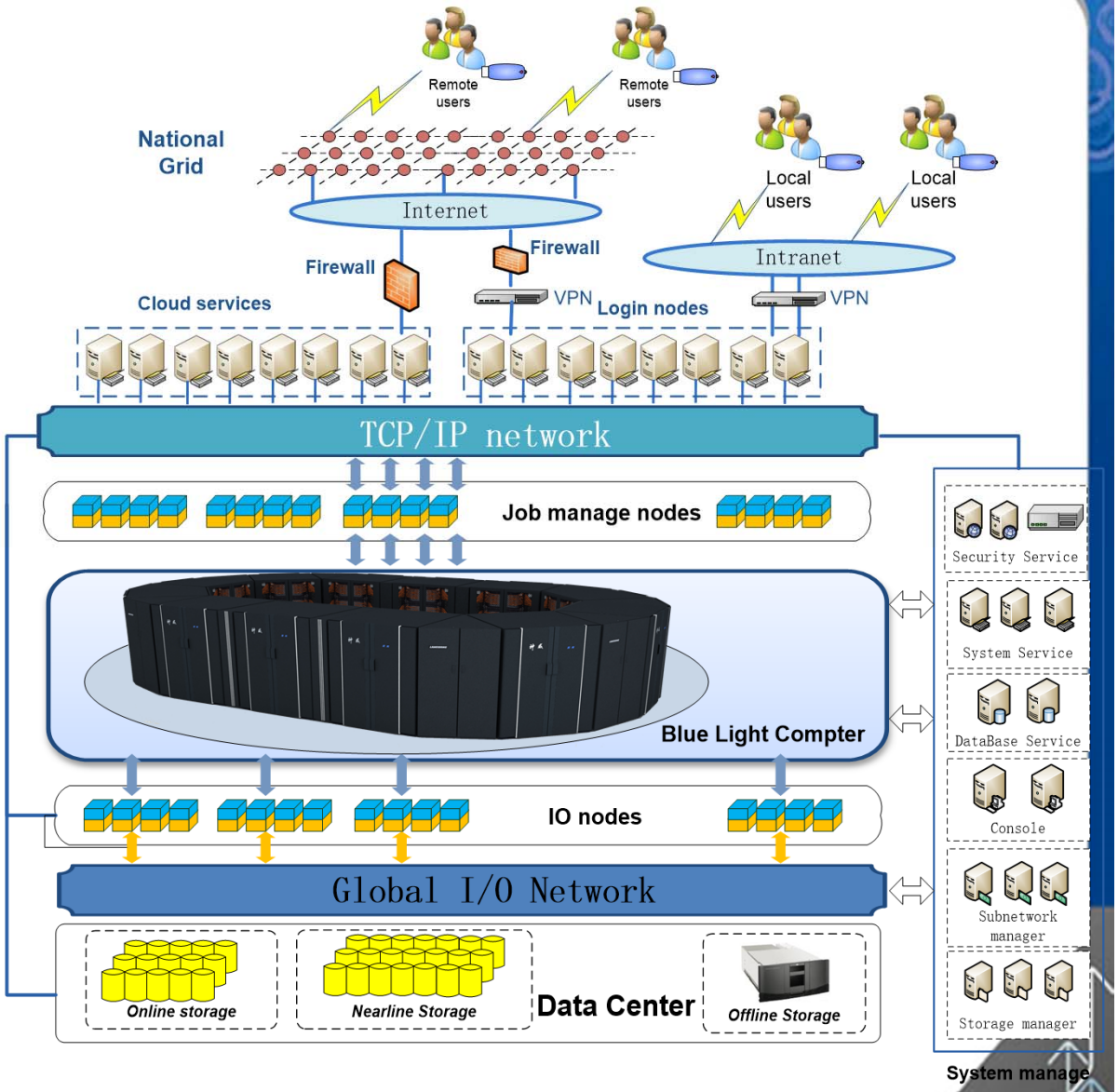




Sunway BlueLight Architecture

Main features

- SW1600 CPU: 16 cores/975~1100MHz/124.8~ 140.8Gflops;
- Fat-tree based interconnection, QDR 4×10Gbps high speed serial transmission between nodes, MPI message latency of 2μs;
- SWCC/C++/Fortran/ UPC/MPICC/Scientific library;
- Storage: 2PB, theoretical I/O bandwidth: 200GB/s, IOR(~60GB/s);



2. Processor

- Processor is the key element to achieve performance and energy efficiency
  - 20MW system power consumption requires processor with 100GF/W energy efficiency
  - very difficult to achieve
- Processor micro-architecture
  - heterogeneous many-core
  - trade-off between die area and speed
    - on-chip specialized processing units used when needed
- High memory access bandwidth
- On-chip memory and networks

3. Memory

- Byte/Flops ratio becomes very low
  - <3% for 100PF systems, even lowers for exaflops systems
  - Require fundamental change in algorithm
- Speed gap between the computing cores and the memory becomes larger
- Require both high bandwidth and low latency
  - Bandwidth
    - Wide memory data path and high cache hit rate
  - Latency
    - Improved by high cache hit + prefetching
    - Prefetch
      - when? where? and what size?
      - adaptive prefetching according to the nature of the programs
- Introducing new memory devices into memory hierarchy
  - increases memory space while reducing the power consumption
  - NVM as buffer: disk-NVM-DRAM-cache
  - NVM as part of main memory: hybrid memory
- 3D packaging is a solution to improve memory performance
  - introducing new problems in memory and cache organization
    - appropriate lay-out to shorten the wires

4. Interconnect

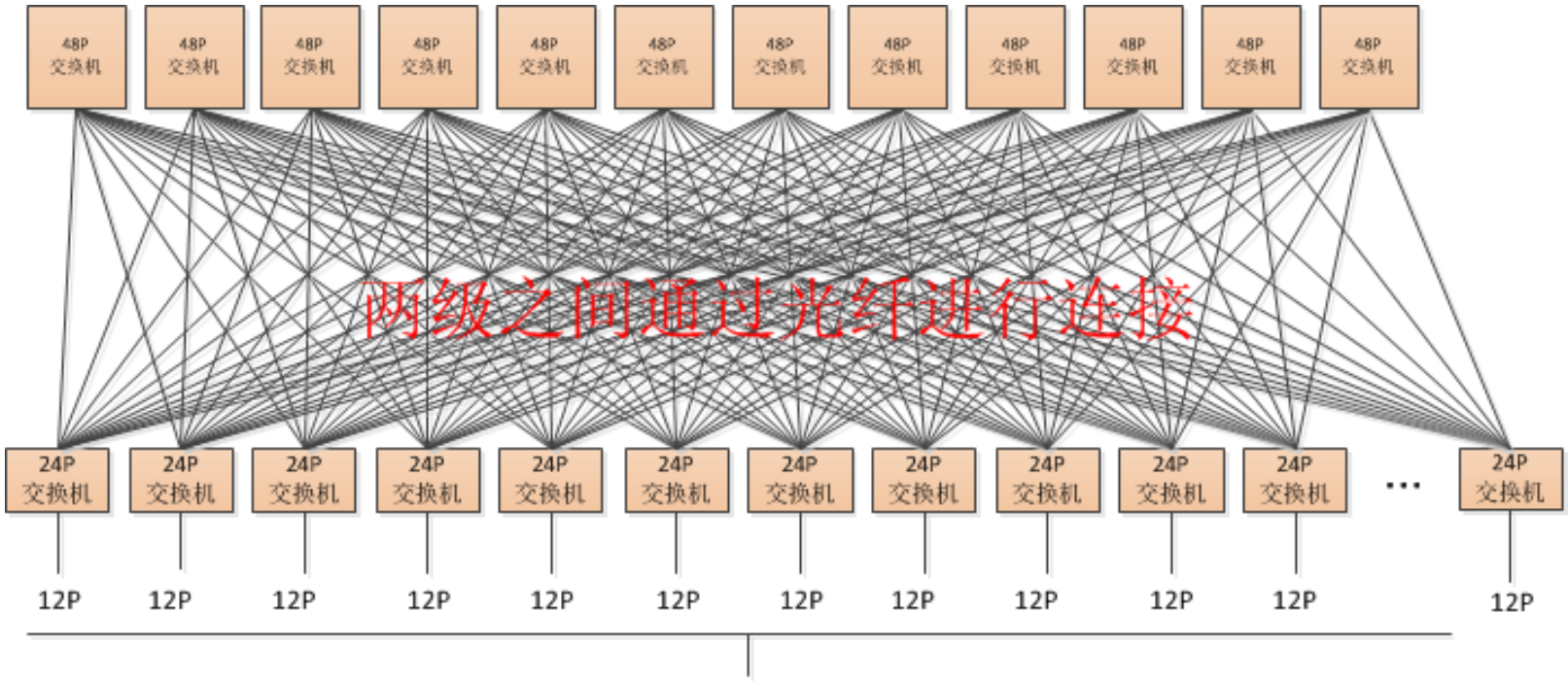
- Three major requirements
  - Scalability
    - scalable to support interconnect of large number of nodes
  - Bandwidth
    - high bandwidth for performance
  - Latency
    - Low latency is critical for synchronization and short message passing
- Affected by network topology, link technology, and the communication protocol
- Both system level interconnect and on-chip interconnect should be studied
- Energy consumption will limit the increase of link data rate
- New technology demanded
  - silicon photonic communication
  - new topology
  - new light weight protocol

TH-2 Interconnect

- High speed interconnect chipset
  - interconnect interface chip N10
    - PCIe 3.0 x16 host interface
    - Link rate 14Gbps
    - Network connection bandwidth 12GB/s
    - Low latency message passing and high bandwidth RDMA
  - high radix router chip HNR
    - 24 port，network port bandwidth 12GB/s
    - 376Tbps Single chip throughput
    - Support distributed multi-path adaptive routing

TH-2 Interconnect (cont'd)

- Fat-tree topology
- Maximum 18432 nodesal
- Optical-electric hybrid transport technology
- Proprietary network protocol

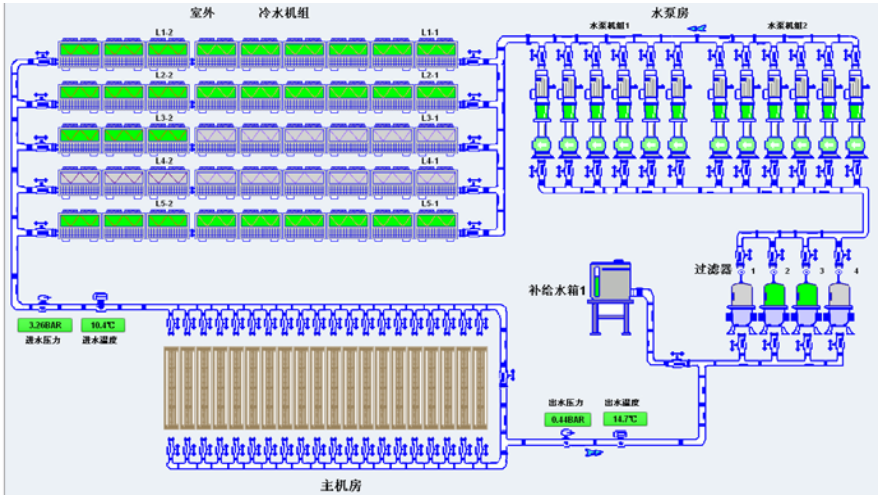


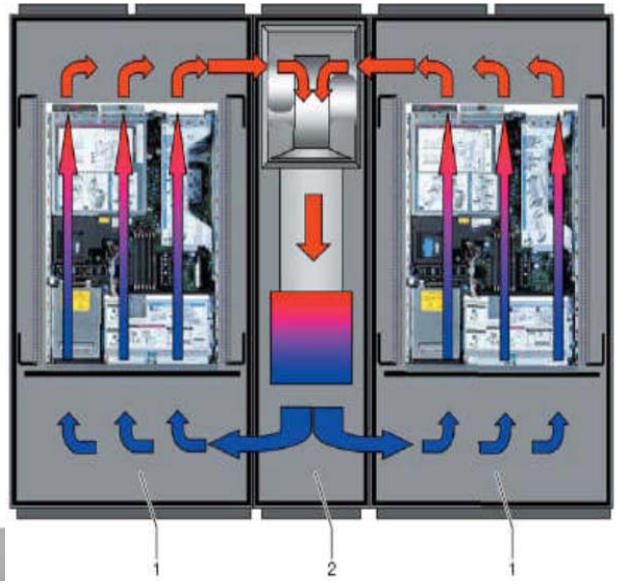
5. Energy efficiency

- Energy saving measures
  - At different level
    - devices/node/system
  - With different approaches
    - hardware/software/coordinated
  - In different aspects
    - computing system/cooling/power supply/computing room
  - At different time
    - static: energy-aware programming and tuning
      - energy-tuner might become a standard tool as the debugger and performance tool, they altogether address energy efficiency, correctness, performance of the program, respectively
    - dynamic: runtime scheduling and DVFS control

SW/BL cooling system

- **Efficient cooling**
  - **Water cooling to the node board**
  - **Energy-saving**
  - **Environment-friendly**
    - **High room temperature**
    - **Low noise**

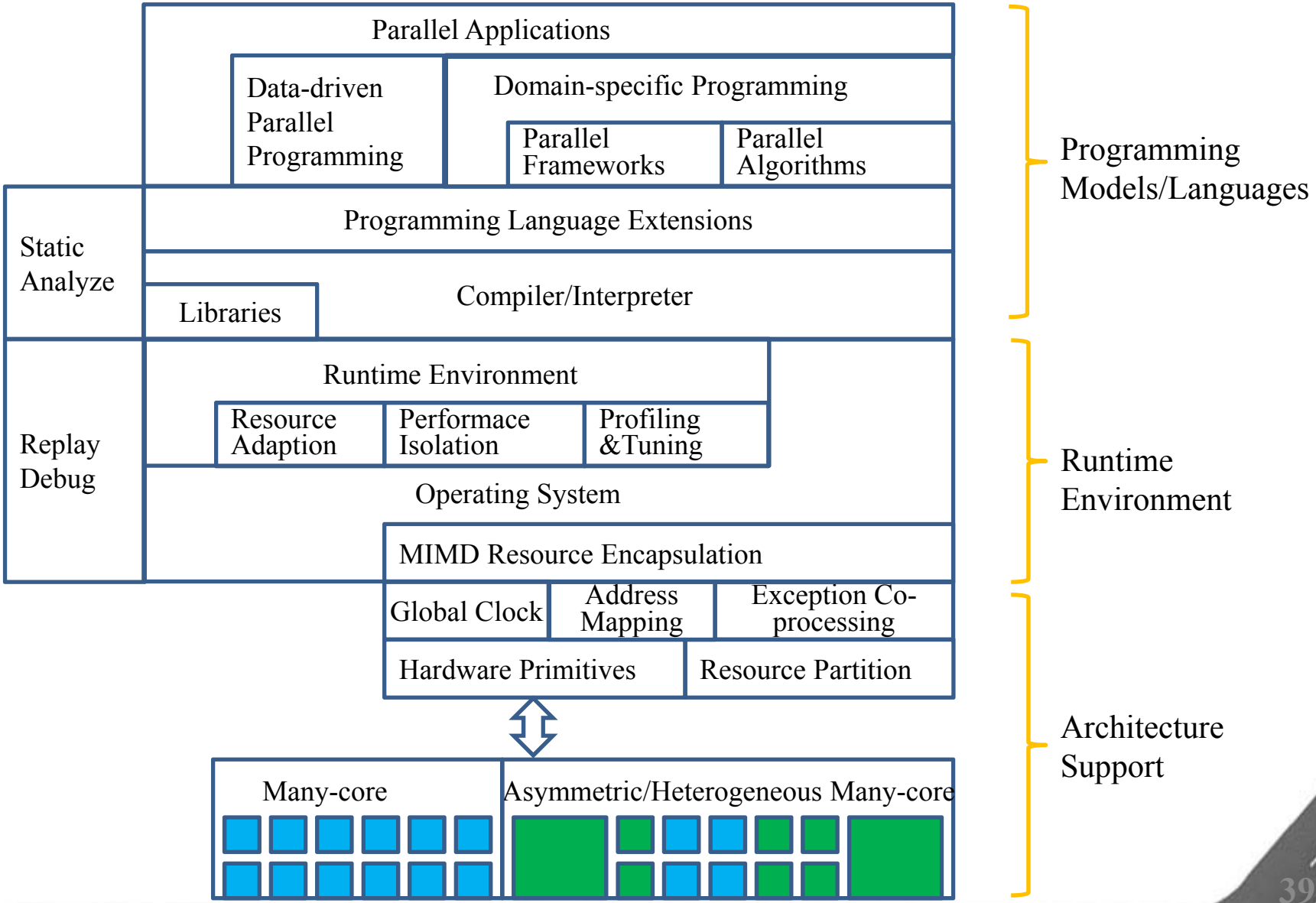




6. Programmability

- Parallel programming becomes a common practice for all application developers, not only HPC
- Difficulty introduced by heterogeneity, requires new language/compiler support, new performance tools
- A holistic approach in supporting many-core parallel programming needed

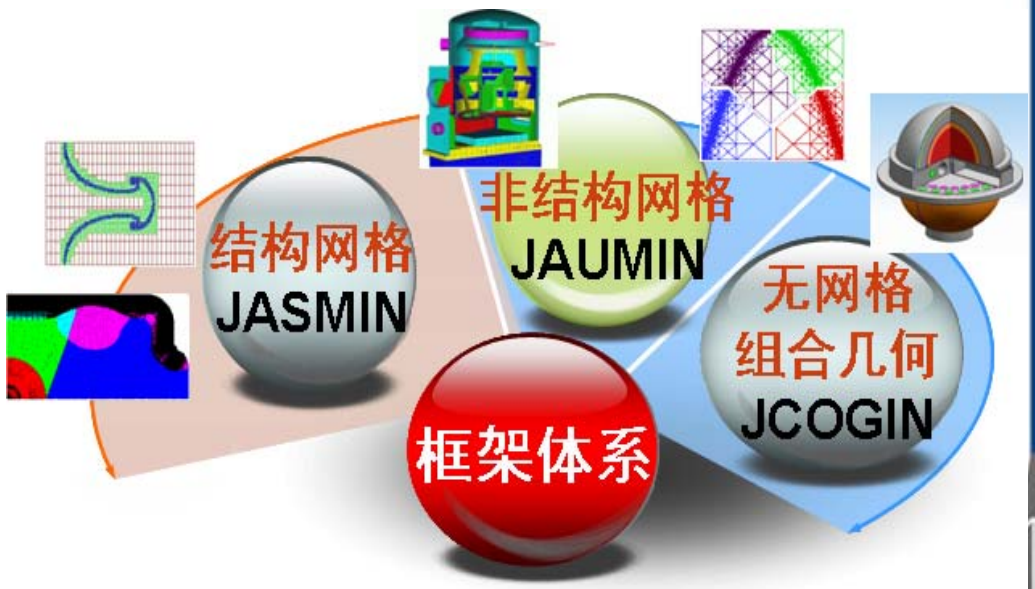
A holistic approach supporting manycore programming



Parallel programming framework

- Hiding the complexity of parallel programming
  - Integrating efficient implementation of fast parallel algorithms
  - Providing efficient data structures and solver libraries
  - Supporting software engineering practice in large scale parallel software development

- Four components
  - –JASMIN：adaptive structured mesh
  - –PHG： parallel hierarchical grid
  - –JAUMIN: adaptive unstructured mesh
  - –JCOGIN： mesh-free combinatory geometry



7. Parallel algorithms

- Algorithms must adapt to low ratio of memory capacity and computing capability
- Re-design algorithms to match the heterogeneous architecture
- Architecture-aware algorithms vs architecture-support to algorithm implementation
  - requires efforts from both sides: the programmers and architects
- Performance and power consumption are sensitive to data move, must reduce data move in algorithm implementation

8. Resilience

- Resillience measures
  - At different levels
    - device/component/system
  - In different aspects
  - hardware/software/coordinated
  - With different approaches
    - redundancy to hide failure
    - checkpointing to recover execution
    - new checkpointing mechanism to deal with very short MTBF
      - reduce the context
      - use new media to store
- Accurate monitoring
  - more embedded sensors to capture failures
  - efficient data collection for large-scale system monitoring
  - accurate data analysis to identify failures
- Fast recovery from the failure
  - accurate locating and isolation of faults
  - reconfiguration of the system

9. OS and runtime

- Effective management and efficient use of large scale heterogeneous resources to achieve more predictable performance
- Dynamic matching demand and resource by runtime
- Reduce interference among co-run programs in competing shared resources
  - Interference-aware scheduling and dynamic migration
- Resource virtualization to enable effective resource sharing

Prospects

The structure of the new national R&D Programs

- Five tracks in the new national research systems
  - Basic research program
  - Mega-research program
  - Key research and development program
  - Enterprises-oriented research program
  - Research centers and talents program

Key R&D program

- The track 3 is of the biggest change
  - including previous 863, 973, and ebabling programs, and other ministerial level R&D programs

- A transit period of 2015-2016
  - No new 863, 973 projects will be launched in 2015 and 2016

A new proposal on HPC

- Strategic studies have been organized jointly by the 863 key project and the Supercomputing Innovation Alliance under the guidance of the MoST
- A proposal for the key project on HPC in the 13th five-year plan has been submitted
- The decision has not been made

Motivation for next generation supercomputer

- The key value of developing Exa-scale computers identified
  - Addressing the grand challenge problems
    - energy, environment, climate change…
  - Enabling industry transformation
    - simulation and optimization for high speed train design, aircraft design, automobile design,…
    - support SME
  - Social development and people's benefit
    - drug discovery, weather forecast, precision medicine, digital media…
  - Scientific discovery
    - high energy physics, computational chemistry, new material, cosmology…
- Boost computer industry by technology transfer
- Self-controllable HPC technologies
  - A lesson learnt from the recent embargo regulation

Goal and major tasks

- Goal
  - Pursuing the leading position in HPC system development
  - Strengthen development of kernel technologies
  - Promote HPC applications
  - Build up HPC infrastructure with service features and explore the path to the HPC service industry
- Major tasks
  - Next generation supercomputer development
  - CNGrid upgrading and transformation
  - Domain HPC applications development

Risk and difficulties

- Un-secured matching funding
  - the funding from national R&D program is not enough
  - getting matching funding from the local government and user institutes is difficult to continue
- Energy efficiency metrics is difficult to achieve
  - The biggest obstacle to exe-scale computers is the power budget, difficult to achieve 50GF/W
- Ecosystem for indigenous processors has not been established
  - The system software and application software for indigenous processors is not enough
  - Establishing an eco-system for indigenous processors is crucial
- Collaborative research is crucial to success
  - Too much emphasis on competition, less collaboration because of the current evaluation system
  - Must find the way to collaboratively conduct R&D on the next generation supercomputer

Thank you!