China's HPC development in the next 5 years

Depei Qian
Beihang University/Sun Yat-sen University
HiPC 2016, Hyderabad, India
Dec 21, 2016

Outline

- A Brief review
- Current status
- Challenges and considerations
- Efforts in the 13th 5-year plan

A Brief review

The high-tech program (863 program)

- The most important high-tech R&D program of China since 1986
- Proposed by 4 senior scientists and approved by former leader Deng Xiaoping in March 1986
- Covers 8 areas, Information Technology is one of them
- Emphasis on strategic and frontier research on major technologies supporting country's development
- Also emphasize technology transfer and promotion to industry

Major changes of 863's research emphasis in computing

- 1987: Intelligent computers
  - Influenced by the 5th generation computer program in Japan
- 1990: from intelligent computer to parallel computers
  - Emphasizing practical HPC capability for research and industry
  - Developing SMP & MPP
- 1998: from single HPC system to HPC environment
  - Emphasizing resource sharing and ease of access
  - Promoting the usage of the HPC systems
- 2006: from high performance to high productivity
  - Emphasizing other metrics such as programmability, program portability, and reliability besides peak performance
- Current:
  - Emphasizing integrated and balanced development of systems, environment, and applications
  - Exploring new mechanisms and business models for HPC services
  - Establishing eco-system for HPC applications

Three key projects on HPC

- 2002-2005：High Performance Computer and Core Software
  - Resource sharing and collaborative work
  - Grid-enabled applications in multiple areas
  - TFlops computers and China National Grid (CNGrid) testbed
- 2006-2010：High Productivity Computer and Grid Service Environment
  - High productivity
    - Application performance
    - Efficiency in program development
    - Portability of programs
    - Robust of the system
  - Service features of the HPC environment
  - Peta-scale computers
- 2010-2016：High Productivity Computer and Application Service Environment
  - 100PF computers
  - Large scale HPC applications
  - Upgrading CNGird

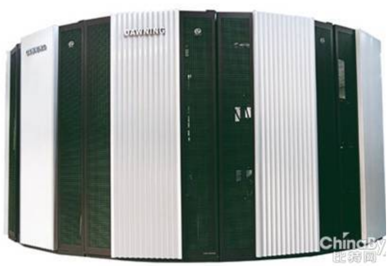
HPC systems developed (1993-2011)

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

Dawning I

Dawning 1000

Dawning 2000



Dawning 3000

DeepComp6800

HPC systems developed (1993-2011)

- 2004：Dawning 4000A, Peak performance 11.2TFlops, cluster (No 10 in TOP500)
- 2008：Lenovo DeepComp 7000,150TFlops peak, Hybrid cluster
- 2008：Dawning 5000A, 230TFlops, cluster (2008)
- 2010：TH-1A, 4.7PFlops peak, 2.56PFlops LinPack, CPU+GPU (No 1 in TOP500)
- 2010：Dawning 6000, 3Pflops peak, 1.27 PFlops LinPack, CPU+GPU
- 2011：Sunway Bluelight, 1.07PFlops peak, 796TF LinPack, implemented with China's multicore processors

Dawning 4000

DeepComp 7000

曙光5000A总体效果图

Dawning 5000

Dawning 6000

TH-1A

Sunway-Bluelight

Current status

Current key project on HPC

- High Productivity Computer and Application Service Environment (2011-2016)
  – Developing world-class computer systems
    - Tianhe-2
    - Sunway TaihuLight
  – Upgrading CNGrid and exploring new operation model and mechanism
  – Developing large scale parallel application software

First phase of TH-2

- Delivered in May 2013
- Hybrid system
  - 32000 Xeon, 48000 Xeon Phi, 4096 FT CPU
- 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 evaluated and approved in July of 2014
  - Upgrading interconnect (completed)
  - Increasing the number of computing nodes (completed)
  - Upgrading computing nodes (delayed)
    - Upgrade the accelerator, replacing Knight Conner by Knight Landing

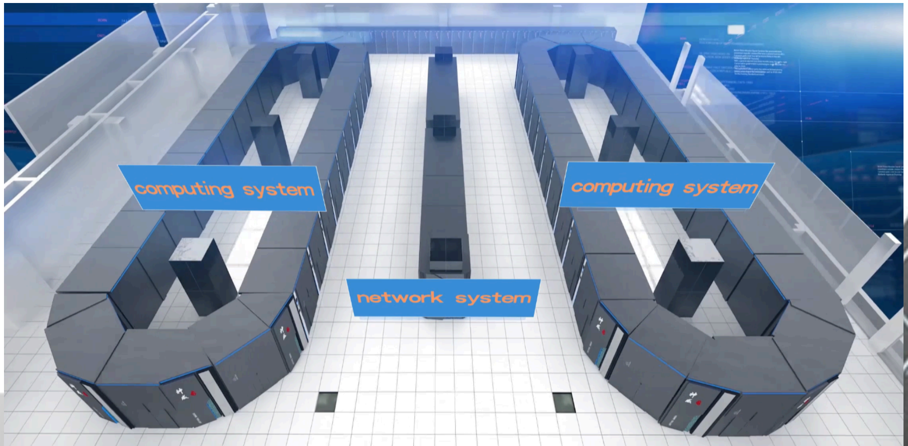
Second phase of TH-2 (cont'd)

- The scheme has to be changed because of the embargo regulation of the US government
- The upgrading of TH-2 has to rely on indigenous FT processors/accelerators
- Completion of the second phase of TH-2 is delayed until the next year
- The development of the new FT processors/accelerators is still an on-going effort

Sunway Taihulight

- The second 100PF system, Sunway TaihuLight, was delivered in April 2016 and installed at the National Supercomputing Center in Wuxi.
- Implemented with indigenous SW 26010 260-core processors

Entire System	
Peak Performance	125 PFlops
Linpack Performance	93 PFlops
Total Memory	1310.72 TB
Total Memory Bandwidth	5591.45 TB/s
# nodes	40,960
# cores	10,649,600

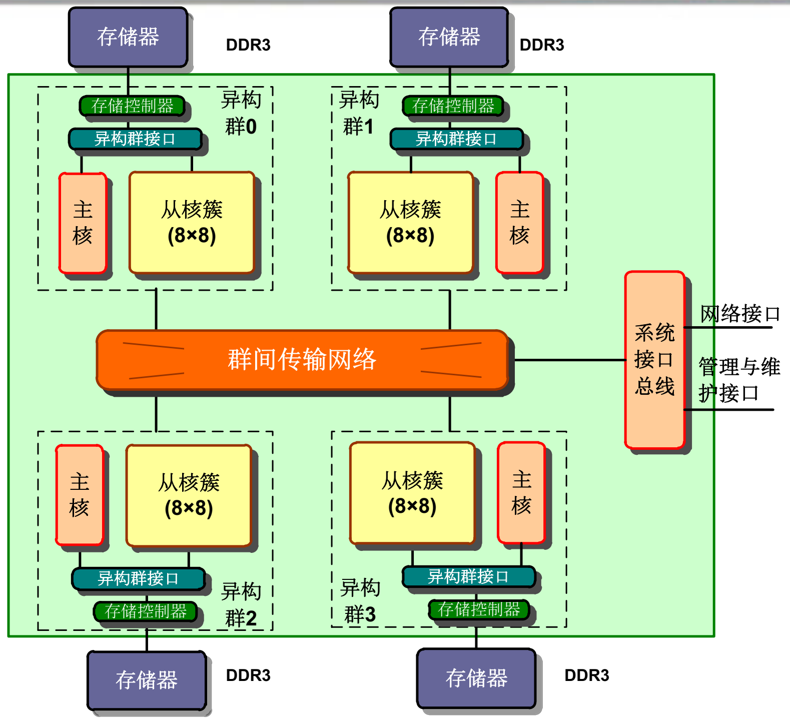


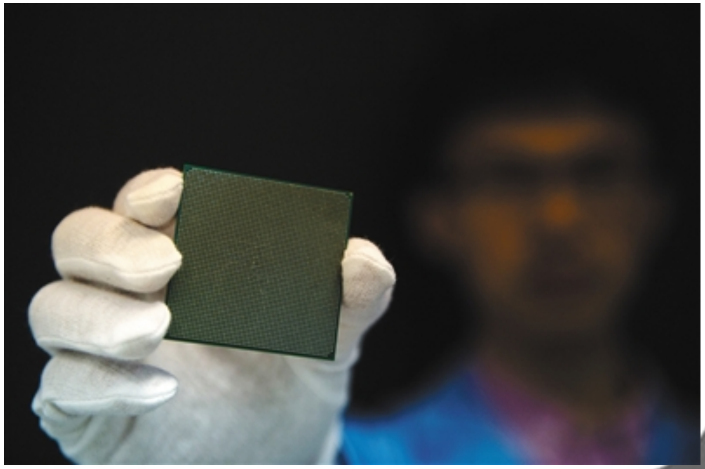
Sunway Taihulight (cont'd)

- Technology innovation achieved
  - High performance many-core processor
  - Low-power design
  - Very compact system
    - 40+ cabinets for 125PF
    - 1024 processors/cabinet
  - Efficient cooling: indirect water cooling to the chips
  - Efficient power supplies
  - Fault tolerant mechanism
    - detection and automatic replacement of the failed nodes
  - Many-core compiler support

SW 26010 Processor

- Core frequency ≈1.5 GHz
- DP Float peak performance ≈ 3.0 TFlops
- Energy efficiency ≈ 10 GFlops/w
- Heterogeneous many-core architecture
  - 260 cores (4 main cores and 256 computing cores with local memory)
  - on-chip integrated memory controllers and network interface





HPC environment development

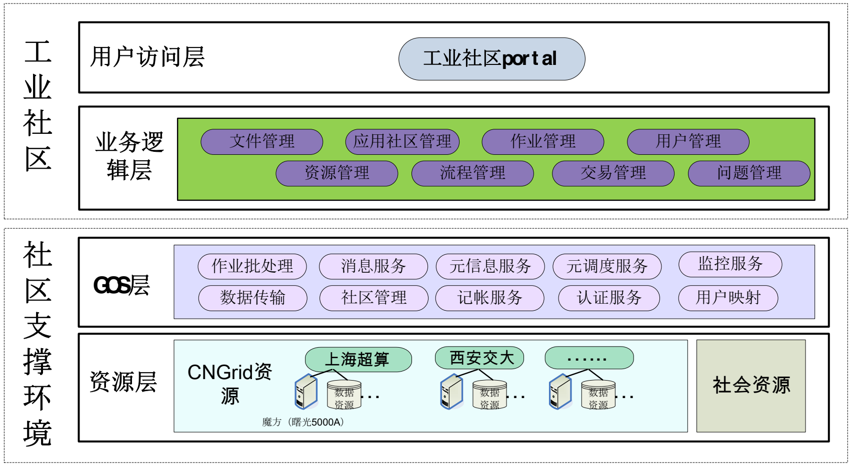
- CNGrid service environment established with service features
  - enabled by software CNGrid Suite
  - 14 nodes currently, will be 16 next year
  - 8PF aggregated computing power, will be upgraded by integration of two 100PF systems
  - >15PB storage
  - >400 software and tools as services
  - supported >3000 projects

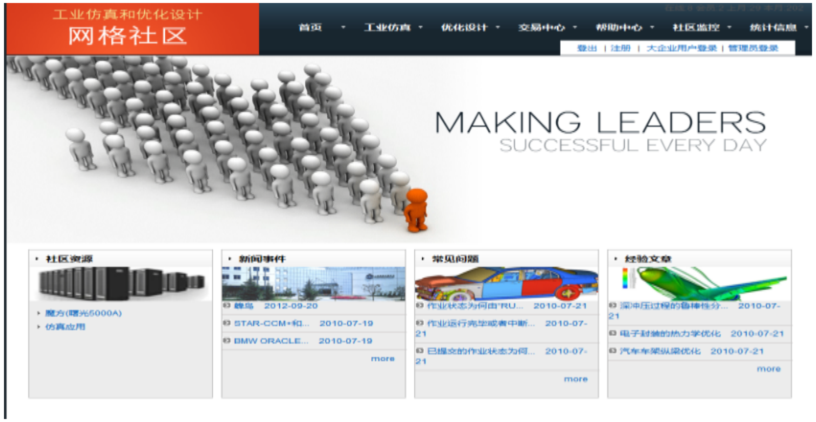
CNGrid sites

	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

Application villages over CNGrid

- Establishing domain-oriented application villages on top of CNGrid, providing services to the end users
- Set up business models and mechanisms between CNGrid and app villages
- Developing enabling technologies and platform supporting CNGrid transformation
- App villages currently being developed
  - Industrial product design optimization
  - New drug discovery
  - Digital media





Parallel software development

- 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
  - Effective use of more 300,000 cores with >30% efficiency required
  - Several reach more than million-core parallelism
- Must be used in the productive system for domain applications

Problems identified

- China is still weak in kernel HPC technologies
  - processor/accelerator
  - novel devices (new memory, storage, and network)
  - large scale parallel algorithms and programs implementation
- Weak in application software
  - Applications rely on imported commercial software
    - expensive
    - small scale parallelism
    - limited by export regulation
- Shortage in cross-disciplinary talents
  - No enough talents with both domain and IT knowledge
- Lack of multi-disciplinary collaboration

Challenges and considerations

Major Challenges to exa-scale systems

- Power consumption
  - Biggest obstacle
- Performance obtained by applications
- Programmability
  - Dealing with massive parallelism and Heterogeneity
- Resilience

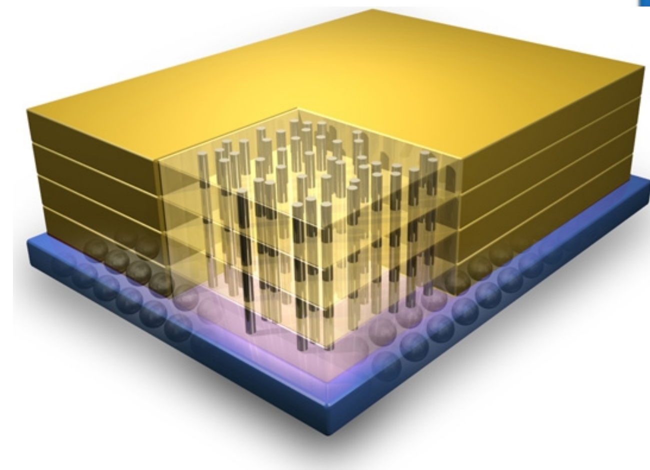
- How to make tradeoffs between performance, power consumption, and programmability?
  - Could we sacrifice programmability for higher energy efficiency?
- How to achieve continuous no-stop operation?
- How to adapt to a wide range of applications with reasonable efficiency?

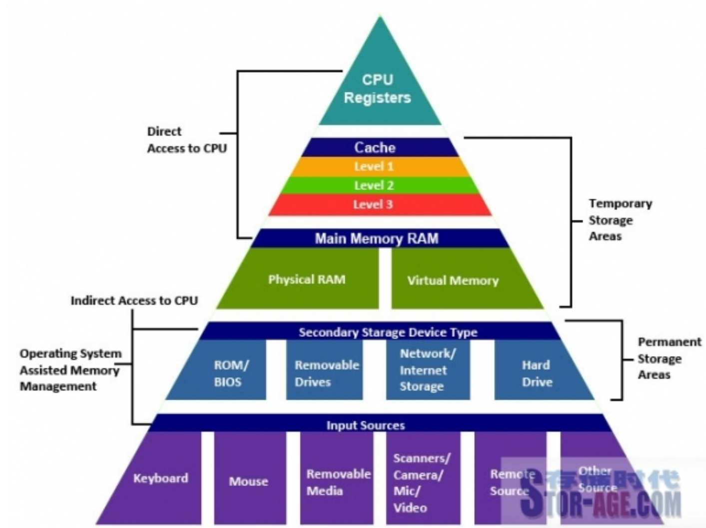
Architecture

- Novel architecture beyond the current heterogeneous accelerated/manycore-based expected
- Co-processor or partitioned heterogeneous architecture?
  - Low utilization of the co-processor in some applications, using CPU only
  - Bottleneck in moving data between CPU and co-processor
- Application-aware architecture
  - On-chip integration of special and general purpose units (idea from Prof. Andrew Chien), using the most efficient specific units when needed
  - Dynamic reconfigurable, how to program?
- Reducing data access and movement
  - Algorithm redesign
  - Energy-aware programming

Memory system

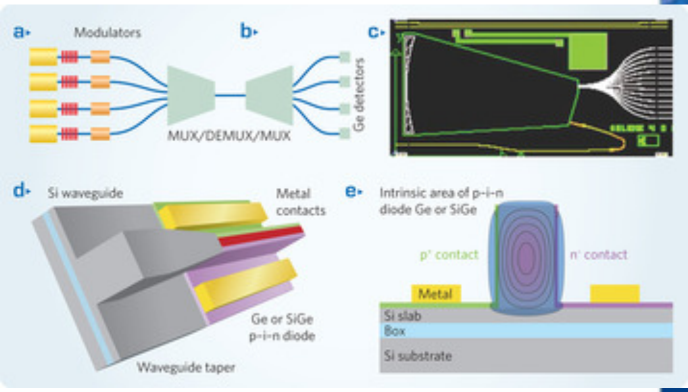
- Achieving large capacity, low latency, high bandwidth
- Increase capacity and lower power consumption by using DRAM/NVM together
  - Data placement issue
  - Handle the high write cost and limited lifetime of NVM due to write
- Bring the data closer to the processing
  - HBM near processor
  - On-chip DRAM
  - Simple functions in memory
- Using 3D stack technology
  - improving bandwidth and latency
  - match the physical and logical layout and reduce the distance of data moving
- Unified memory space in heterogeneous architecture

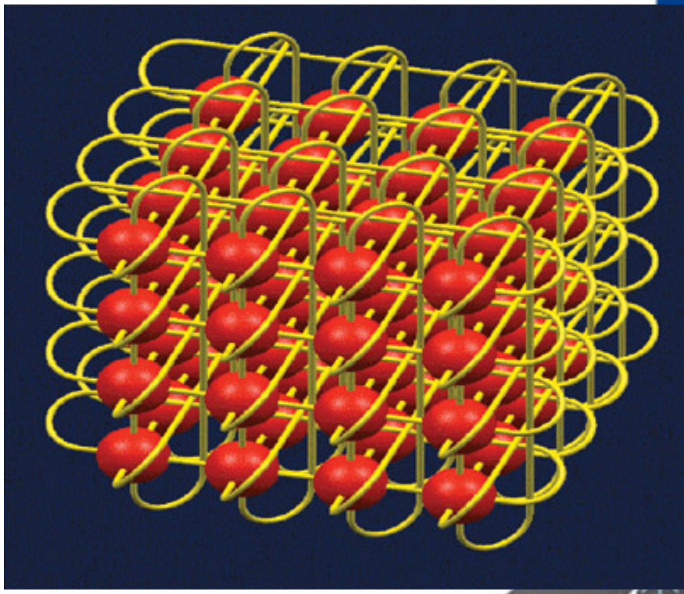




Interconnect

- Pursuing low latency, high bandwidth and low energy consumption
- New technologies
  - Silicon photonics communication
  - Optical interconnect/communication
  - 3D packaging
  - Miniature optical devices
- High scalability adapt to exa-scale
  - Interconnect for 10,000+ nodes
  - Low hop, low latency topology
  - Reliable and intelligent routing



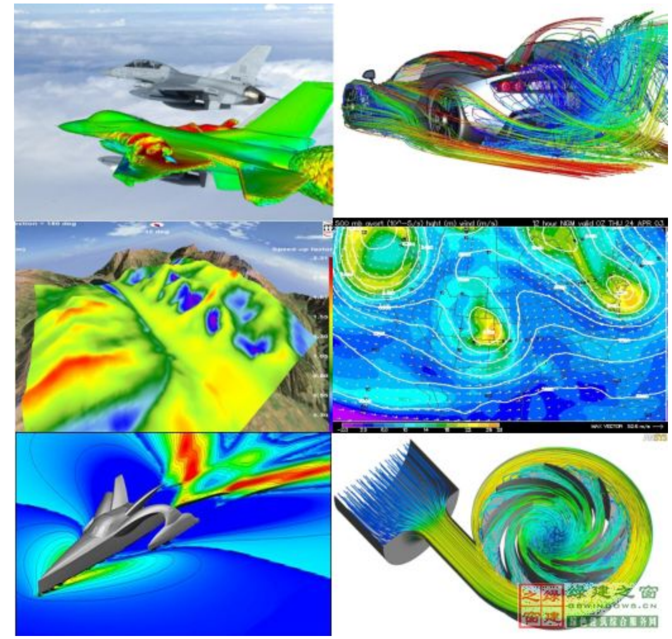


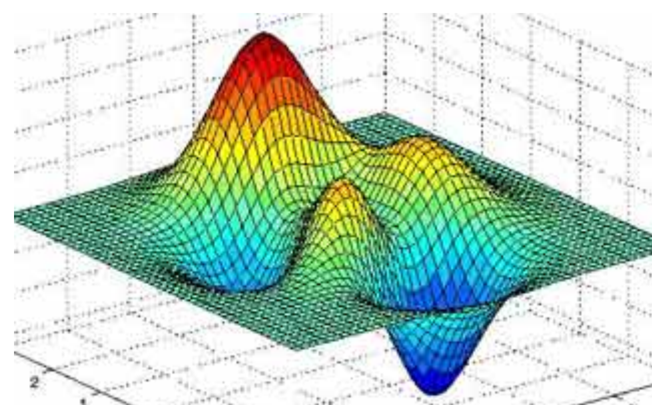
Programming the heterogeneous systems

- Parallel programming for heterogeneous systems
  - for efficient expression of parallelism, dependence, data sharing, execution semantics
  - facilitating problem decomposition on heterogeneous systems
- A holistic approach proposed to deal with the difficulties in programming and uncertainty in performance
  - Programming models
  - Programming languages and compiler
  - debugging
  - Runtime optimization
  - Architectural support

Modeling methods and algorithms

- Full chain innovation
  - Mathematical methods
  - Algorithms
  - Algorithm implementation and optimization
- Good mathematical method is often more fundamental than hardware improvement and algorithm optimization
- Architecture-aware algorithm implementation and optimization is necessary
- Domain-specific libraries for improving software performance, productivity, and reducing the programming barrier





Resilience

- Resilience is one of the key issues of the exa-scale system
  - Large scale of system
    - 50K to 100K nodes
    - Huge amount of components
  - Very short MTBF
  - Long time non-stop operation required for solving large scale problems
- Reliability measures at different levels, including device, node, and system level
- Software/hardware coordination
  - Checkpointing requires fast context saving and recovery to avoid domino roll-back
  - Fault-tolerance at the algorithm and application software level

Importance of the tools

- Development and optimization of large scale parallel software require support of tools
- Particularly important for systems implemented with self-developed processors
- Three default tools required
  - Parallel debugger for correctness
  - Performance tuner for performance
  - Energy optimizer for energy efficiency

Need for an eco-system

- Eco-system for exa-scale system based on indigenous processors is in a urgent need
  - System software
  - Tool software
  - Application development support
  - Application software
- How to attract the third party software developers
  - Need product lines instead of a single machine
- Collaboration between academia and industry required
- Multi-disciplinary collaboration required

Efforts in the 13th 5-year plan

Reform of research system in China

- The national research and development system is being reformed
  - Merge 100+ different national R&D programs/initiatives into 5 tracks of national programs
    - Basic research program (NSFC)
    - Mega-science and technology programs
    - Key R&D program (former 863, 973, enabling programs)
    - Enterprise innovation program
    - Facility/talent program

A New key project on HPC

- High performance computing has been identified as a priority subject under the key R&D program
- Strategic studies and planning have been conducted since 2014
- A proposal on HPC in the 13^{th} five-year plan was submitted in early 2015 and approved by the end of 2015 by a multi-government agent committee lead by the MOST
- The key project on high performance computing was launched in Feb. of 2016

Motivations for the new key project

- The key value of exa-scale computers identified
  - Addressing the grand challenge problems
    - Energy shortage, pollution, climate change…
  - Enabling industry transformation
    - Using simulation and optimization to support important systems and products
      - high speed train, commercial aircraft, automobile design…
    - support economy transformation
  - For social development and people's benefit
    - new drug discovery, precision medicine, digital media…
  - Enabling scientific discovery
    - high energy physics, computational chemistry, new material, astrophysics…
- Promote computer industry by technology transfer
- Developing HPC systems by self-controllable technologies
  - a lesson learnt from the recent embargo regulation

Goals and major tasks

- Goals
  - Strengthening R&D of kernel technologies and pursuing the leading position in high performance computer development
  - Promoting HPC applications
  - Building up an HPC infrastructure with service features and exploring the path to the HPC service industry
- Major tasks
  - Next generation supercomputer development
  - HPC applications development
  - CNGrid upgrading and transformation
- Each task will cover basic research, key technology development, and application demonstration

Task 1: Exa-scale computer development

- **Activities**
  - R&D on novel architectures and key technologies of the next generation supercomputers
  - Development of an exa-scale computer based on domestic processors
  - Technology transfer to promote development of high-end servers

Task 1: Exa-scale computer development

- **Basic research**
  - Novel high performance interconnect
    - Research on theories and implementation technologies of the novel interconnect
      - based on the enabling technologies of 3D chips, silicon photonics and on-chip networks
  - Programming&execution models for exa-scale systems
    - developing new programming models for heterogeneous systems
    - enhancing efficiency in programming
    - exploitation of advantages of the heterogeneous architectures

Task 1: Exa-scale computer development

- **Key technology**
  - Prototype systems for verification of the exa-scale technologies
    - candidate architectures for exa-scale computer
    - major implementation technologies
    - technologies for improving energy efficiency
    - prototype system
      - 512 nodes
      - 5-10TFlops/node
      - 10-20Gflops/W
      - point to point bandwidth>200Gbps
      - MPI latency<1.5us
      - Emphasis on self-controllable technologies
    - system software for prototypes
    - 3 typical applications to verify the design

Task 1: Exa-scale computer development

- **Key technology**
  - Architecture optimized for multi-objectives
    - exa-scale architecture under the constraints of performance, energy consumption, programmability, reliability, and cost
  - energy efficient computing node
    - 50-100TFlops/node
    - 30^{+}GFlops/w
  - high performance processor/accelerator design
    - 20TFlops/chip
    - 40^{+}GFlops/W
    - Support multiple programming models

Task 1: Exa-scale computer development

- **Key technology**
  - exa-scale system software
    - node OS
    - runtime
    - program development environment
    - system management
    - parallel debugger and performance analysis tool
  - highly scalable interconnect
    - high bandwidth, low latency
    - support interconnection of tens-of-million cores
  - scalable parallel I/O
    - multi-layer storage architecture
    - fault-tolerant techniques

Task 1: Exa-scale computer development

- **Key technology**
  - exa-scale infrastructure
    - high density assembling
    - high efficient power supply
    - high efficient cooling
  - energy efficiency
    - cross-layer strategy
    - hardware and software coordination
  - exa-scale system reliability

Task 1: Exa-scale computer development

- **Exa-scale computer system development**
  - exaflops in peak
  - Linpack efficiency >60%
  - 10PB memory
  - EB storage
  - 30GF/w energy efficiency
  - interconnect >500Gbps
  - large scale system management and resource scheduling
  - easy-to-use parallel programming environment
  - system monitoring and fault tolerance
  - support large scale applications

Task 1: Exa-scale computer development

- **Technology transfer**
  - High-end domain-oriented servers based on exa-scale system technologies
    - high performance computing node
    - high speed interconnect
    - scalable I/O
    - energy efficient
    - high reliability
    - application software

Task 2: HPC application development

- **Activities**
  – Basic research on exa-scale modeling and parallel algorithms
  – Developing high performance application software
  – Establishing the HPC application eco-system

Task 2: HPC application development

- **Basic research**
  - computable modeling and novel computational methods for exa-scale systems
  - scalable high-efficient parallel algorithms and parallel libraries for exa-scale systems

Task 2: HPC application development

- **Key technology**
  - programming framework for exa-scale software development, including framework for
    - structured mesh
    - unstructured mesh
    - mesh-free combinatory geometry
    - finite element
    - graph computing

    - supporting development of at least 40 million-core software

Task 2: HPC application development

- **Key technology and demo applications**
  - Numerical devices and their applications
    - numerical nuclear reactor
      - four components: Including reactor core particle transport, thermal hydraulics, structural mechanics and material optimization,
      - non-linear coupling of multi-physics processes
    - numerical aircraft
      - multi-disciplinary optimization covering aerodynamics, structural strength and fluid solid interaction
    - numerical earth
      - earth system modeling for studying climate change
      - non-linear coupling of multi-physical and chemical processes covering atmosphere, ocean, land, and sea ice
    - numerical engine
      - high fidelity simulation system for numerical prototyping of commercial aircraft engine
      - enabling fast and accurate virtual airworthiness experiments

Task 2: HPC application development

- **Key technology and demo applications**
  - high performance application software for domain applications
    - electromagnetic environment simulation
    - energy-efficient design of large fluid machinery
    - drug discovery
    - ship design
    - complex engineering project and critical equipment
    - energy exploration
    - numerical simulation of ocean
    - digital media rendering
    - large scale hydrological simulation
  - high performance application software for scientific research
    - material science
    - high energy physics
    - astrophysics
    - life science

Task 2: HPC application development

- **Eco-system for HPC application software development**
  - establishing a national-level R&D center for HPC application software
  - build up of a platform for HPC software development and optimization
  - tools for performance/energy efficiency and pre-/post-processing
  - build up software resource repository
  - developing typical domain application software

  - a joint effort involving national supercomputing centers, universities, and institutes

Task 3: HPC environment development

- **Activities**
  - Developing aystem-level software and operation platform for the national high performance computing environment
  - Upgrading CNGrid with leading computing resources and service capability
  - Developing service systems based on the national HPC environment

Task 3: HPC environment development

- **Key technology**
  - service mechanism and technical platform for the national HPC environment
    - new mechanisms and enabling technologies required by service–mode operation
    - upgrading the national HPC environment (CNGrid)
      - >500PF computing resources
      - >500PB storage
      - >500 application software and tools
      - >5000 users (team users)

Task 3: HPC environment development

- **Demo applications**
  - service systems based on the national HPC environment
    - integrated business platform, e.g.
      - complex product design
      - HPC-enabled EDA platform
    - application villages
      - innovation and optimization of industrial products
      - drug discovery
      - SME computing and simulation platform
    - platform for HPC education
      - provide computing resources and services to undergraduate and graduate students

Calls for proposals

- The first call for proposal was issued in Feb. , 2016. 19 projects have passed the evaluation and been launched
- The second call (for 2017) was issued in Oct., 2016, the pre-proposal submission ended last month and the evaluation process is on going
- These two rounds of call cover most of the subjects of the key project except the exa-scale system development.
- The exa-scale system development will be started after completion of the three prototypes

Thank you!