
National Supercomputing Centers, Supercomputer Systems and Applications.

Exascale Supercomputer in China: From Prototype to E-System

Some Exploratory Work of Lattice QCD on E-prototype System
<table>
<thead>
<tr>
<th>#</th>
<th>Site</th>
<th>Manufacturer</th>
<th>Computer</th>
<th>Country</th>
<th>Cores</th>
<th>Rmax (TFlops)</th>
<th>Power (MW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Oak Ridge National Laboratory</td>
<td>IBM</td>
<td>Summit IBM Power System, P9 22C 3.07GHz, Mellanox EDR, NVIDIA GV100</td>
<td>USA</td>
<td>2,414,592</td>
<td>148.6</td>
<td>10.1</td>
</tr>
<tr>
<td>2</td>
<td>Lawrence Livermore National Laboratory</td>
<td>IBM</td>
<td>Sierra IBM Power System, P9 22C 3.1GHz, Mellanox EDR, NVIDIA GV100</td>
<td>USA</td>
<td>1,572,480</td>
<td>94.6</td>
<td>7.4</td>
</tr>
<tr>
<td>3</td>
<td>National Supercomputing Center in Wuxi</td>
<td>NRCPC</td>
<td>Sunway TaihuLight NRCP Sunway SW26010, 260C 1.45GHz</td>
<td>China</td>
<td>10,649,600</td>
<td>93.0</td>
<td>15.4</td>
</tr>
<tr>
<td>4</td>
<td>National University of Defense Technology</td>
<td>NUDT</td>
<td>Tianhe-2A ANUDT TH-IVB-FEP, Xeon 12C 2.2GHz, Matrix-2000</td>
<td>China</td>
<td>4,981,760</td>
<td>61.4</td>
<td>18.5</td>
</tr>
<tr>
<td>5</td>
<td>Texas Advanced Computing Center / Univ. of Texas</td>
<td>Dell</td>
<td>Frontera Dell C6420, Xeon Platinum 8289 26C 2.7GHz, Mellanox HDR</td>
<td>USA</td>
<td>448,448</td>
<td>23.5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Swiss National Supercomputing Centre (CSCS)</td>
<td>Cray</td>
<td>Piz Daint Cray XC50, Xeon E5 12C 2.6GHz, Aries, NVIDIA Tesla P100</td>
<td>Switzerland</td>
<td>387,872</td>
<td>21.2</td>
<td>2.38</td>
</tr>
<tr>
<td>7</td>
<td>Los Alamos NL / Sandia NL</td>
<td>Cray</td>
<td>Trinity Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries</td>
<td>USA</td>
<td>979,072</td>
<td>20.2</td>
<td>7.58</td>
</tr>
<tr>
<td>8</td>
<td>National Institute of Advanced Industrial Science and Technology</td>
<td>Fujitsu</td>
<td>Al Bridging Cloud Infrastructure (ABCi) PRIMEGY CX2550 M4, Xeon Gold 20C 2.4GHz, IB-EDR, NVIDIA V100</td>
<td>Japan</td>
<td>391,680</td>
<td>19.9</td>
<td>1.65</td>
</tr>
<tr>
<td>9</td>
<td>Leibniz Rechenzentrum</td>
<td>Lenovo</td>
<td>SuperMUC-NG ThinkSystem SD530, Xeon Platinum 8174 24C 3.1GHz, Intel Omni-Path</td>
<td>Germany</td>
<td>305,856</td>
<td>19.5</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Lawrence Livermore National Laboratory</td>
<td>IBM</td>
<td>Lassen IBM Power System, P9 22C 3.1GHz, Mellanox EDR, NVIDIA Tesla V100</td>
<td>USA</td>
<td>288,288</td>
<td>18.2</td>
<td></td>
</tr>
</tbody>
</table>
Global Progress in Supercomputer
Supercomputer System Share

**COUNTRIES / SYSTEM SHARE**

- United States, 23%
- China, 44%
- United Kingdom, 3%
- France, 4%
- Japan, 6%
- Netherlands, 2%
- Germany, 3%
- Ireland, 3%
- Others, 10%
全球超级计算机TOP500入围数量排行

1994年6月
Global Progress in Supercomputer

TOP500 List

COUNTRIES (TOP50) / SYSTEM SHARE

- United States, 38%
- Japan, 18%
- China, 6%
- Germany, 8%
- United Kingdom, 6%
- France, 10%
- Others, 10%
# Global Progress in Supercomputer

**IBM SUMMIT**

<table>
<thead>
<tr>
<th>Application Performance</th>
<th>122 PF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Nodes</td>
<td>4356</td>
</tr>
<tr>
<td>Node performance</td>
<td>42 TF</td>
</tr>
<tr>
<td>Memory per Node</td>
<td>512 GB DDR4 + 96 GB HBM2</td>
</tr>
<tr>
<td>NV memory per Node</td>
<td>1600 GB</td>
</tr>
<tr>
<td>Total System Memory</td>
<td>&gt;10 PB DDR4 + HBM2 + Non-volatile</td>
</tr>
<tr>
<td>Processors</td>
<td>2 IBM POWER9™, 9,216 CPUs, 6 NVIDIA Volta™, 27,648 GPUs</td>
</tr>
<tr>
<td>File System</td>
<td>250 PB, 2.5 TB/s, GPFS™</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>8.8</td>
</tr>
<tr>
<td>Interconnect</td>
<td>Mellanox EDR 100G InfiniBand</td>
</tr>
<tr>
<td>Operating System</td>
<td>Red Hat Enterprise Linux (RHEL) version 7.4</td>
</tr>
</tbody>
</table>
Global Progress in Supercomputer

IBM SUMMIT

<table>
<thead>
<tr>
<th>Tesla Product</th>
<th>Tesla K40</th>
<th>Tesla M40</th>
<th>Tesla P100</th>
<th>Tesla V100</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>GK10 (Kepler)</td>
<td>GM200 (Maxwell)</td>
<td>GP100 (Pascal)</td>
<td>GV100 (Volta)</td>
</tr>
<tr>
<td>Boost Clock</td>
<td>810/875 MHz</td>
<td>1114 MHz</td>
<td>1480 MHz</td>
<td>1455 MHz</td>
</tr>
<tr>
<td>Peak FP32 TFLOP/s</td>
<td>5.04</td>
<td>6.8</td>
<td>10.6</td>
<td>15</td>
</tr>
<tr>
<td>Peak FP64 TFLOP/s</td>
<td>1.68</td>
<td>2.1</td>
<td>5.3</td>
<td>7.5</td>
</tr>
<tr>
<td>Tensor Core TFLOP/s</td>
<td>NA</td>
<td>NA</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>Texture Units</td>
<td>240</td>
<td>192</td>
<td>224</td>
<td>320</td>
</tr>
<tr>
<td>Memory Interface</td>
<td>384-bit GDDR5</td>
<td>384-bit GDDR5</td>
<td>4096-bit HBM2</td>
<td>4096-bit HBM2</td>
</tr>
<tr>
<td>Memory Size</td>
<td>Up to 12 GB</td>
<td>Up to 24 GB</td>
<td>16 GB</td>
<td>16 GB</td>
</tr>
<tr>
<td>L2 Cache Size</td>
<td>1516 KB</td>
<td>3072 KB</td>
<td>4096 KB</td>
<td>6144 KB</td>
</tr>
<tr>
<td>Shared Memory Size / SM</td>
<td>16 KB/32 KB/48 KB</td>
<td>96 KB</td>
<td>64 KB</td>
<td>Configurable up to 96 KB</td>
</tr>
<tr>
<td>Register File Size / SM</td>
<td>256 KB</td>
<td>256 KB</td>
<td>256 KB</td>
<td>256 KB</td>
</tr>
<tr>
<td>Register File Size / GPU</td>
<td>3840 KB</td>
<td>6144 KB</td>
<td>14336 KB</td>
<td>20480 KB</td>
</tr>
<tr>
<td>TDP</td>
<td>235 Watts</td>
<td>250 Watts</td>
<td>300 Watts</td>
<td>300 Watts</td>
</tr>
<tr>
<td>Transistors</td>
<td>7.1 billion</td>
<td>8 billion</td>
<td>15.3 billion</td>
<td>21.1 billion</td>
</tr>
<tr>
<td>GPU Die Size</td>
<td>551 mm²</td>
<td>601 mm²</td>
<td>610 mm²</td>
<td>815 mm²</td>
</tr>
<tr>
<td>Manufacturing Process</td>
<td>28 nm</td>
<td>28 nm</td>
<td>12 nm</td>
<td>12 nm FFN</td>
</tr>
</tbody>
</table>
TH-1A Supercomputer
“TH-1A” was ranked the first on Top500 list released in Nov 2010.
三大技术创新

Three technical innovations

- CPU+GPU异构融合体系结构
  - CPU+GPU heterogeneous architecture
- 64位多核多线程自主飞腾1000 CPU
  - 64 bits Multi-core and Multi-thread CPU
- 自主高速互连通信技术
  - Self-developed High-speed interconnect communication technology

国际学术界的评价

Remark from international academia

中国的“天河一号”采取的CPU与GPU融合的结构，代表了未来超级计算机的发展趋势。随着计算机规模的不断拓展，这种结构虽然不是唯一的解决方法，但目前看来是最好的。
The architecture is not the only solution, but the best so far.

-- 美国田纳西大学教授杰克•唐加拉
Prof. Jack Dongarra, Tennessee University

中国的“天河一号”采取的CPU与GPU融合的结构是美国田纳西大学教授杰克•唐加拉所说，代表了未来超级计算机的发展趋势。随着计算机规模的不断拓展，这种结构虽然不是唯一的解决方法，但目前看来是最好的。

“天河一号”的运算速度比橡树岭国家实验室的要快大约40%，这是运算速率的极大提升。中国同时研制了一种互联技术，让这些处理器相互联系，这不是美国的技术，而是中国自己的技术。这是一个创举。
The technology belongs to China, not U.S., it’s a pioneering work.

-- 美国斯坦福大学计算机系主任比尔•戴利
Billy Daley, Computer Science College, Stanford University
Past of Supercomputer in China

Roadmap of TH-1A
Global Progress in Supercomputer

Development History

From 160MFlops to 200PFlops; Performance improve >1000000000

The first supercomputer

2018, summit, 187PFlops

TH-1A, 2010年, 4.7PFlops

Cray-YMP 1988年，2.3Gflops

Cray-1 1976 160MFlops

IBM红色选择 1999年，3万亿次

日本地球模拟器 2002年，40TFlops

走鹃, 2008年, 1千万亿次

From 160MFlops to 200PFlops; Performance improve >1000000000
1978年，由小平批准，国防科大开始研制“银河-1”
从此，开始了中国研制超级计算机的艰难和辉煌的历程

Approved by Xiao-ping DENG, the NUDT started to develop the “YH -1”.
From then on, began a difficult and glorious history of Chinese Supercomputer research.
Past of Supercomputer in China

100 Pflops system in China

From 2013 to 2017, keep ten times ranked No.1

Tianhe-2A

Sunway TaihuLight
Plan of Future of Supercomputer

MOST projects on HPC: 5 year Plan

- **Supercomputer R&D**
  - 建立协同研发的机制，集中优势力量，突破核心关键技术
  - Exascale Supercomputer in China (1000 Pflops)

- **HPC applications Support**
  - 建立适应不同行业的国家高性能计算应用软件中心
  - 部署行业能力型重大应用软件系统的研发，构建能力型行业重大应用数值模拟软件平台
  - 部署容量型普适推广应用课题，部署于国家超算中心等，培育、吸引和稳定一批自主应用软件系统的用户

- **HPC environment Construction**
  - 建立具有世界一流资源能力和服务水平的、支撑国家创新发展的国家高性能计算环境
International Collaboration

Supercomputing Development in China

Managed by UT-Battelle for the Department of Energy
Contents


3. Exascale Supercomputer in China: From Prototype to E-System

4. Some Exploratory Work of Lattice QCD on E-prototype System
National Supercomputing Centers
Supercomputer System Share

天津超算中心和“天河一号”主机
National Supercomputer Center in Tianjin and “Tianhe-1A”

深圳超算中心和“曙光星云”主机
National Supercomputing Center in Shenzhen and “Dawning Nebulae”

济南超算中心和“神威蓝光”主机
ShanDong Computer Science Center and “Sunway BlueLight MPP”
National Supercomputing Centers
Supercomputer System Share

长沙超算中心和“天河一号”主机
National Supercomputing Center in Changsha and “Tianhe-1”

广州超算中心和“天河二号”主机
National Supercomputer Center in Guangzhou and “Tianhe-2”

无锡超算中心和“神威·太湖之光”主机
National Supercomputing Center in Wuxi and “Sunway TaihuLight”
Application on China's Supercomputers

Applications

- Bio-informatics
- Oil Exploration
- Animation & Movie
- Engineering Design
- Remote Data
- Metrology & Climate
- Aero and Space Craft Design
- Environment Science
Application on China's Supercomputers
Large-Scale Computing
Application on China's Supercomputers

Engineering simulation and design
Application on China's Supercomputers
Large-Scale Data processing and analysis
Application on China's Supercomputers
Integrated Platform of Artificial Intelligence

Supporting Environment  Platform  Application

应用  平台  算力
智能医疗  智能制造  智慧城市  智能语音  智能机器人  自动驾驶
一体化创新平台  可视化训练平台  模块化集成平台  高可弹训练引擎
“天河一号”超级计算机  “天河三号”E级原型机
1 Development of High Performance Computing in China and Plan.

2 National Supercomputing Centers, Supercomputer Systems and Applications.

3 Exascale Supercomputer in China: From Prototype to E-System

4 Some Exploratory Work of Lattice QCD on E-prototype System
Exascale Supercomputer in China

Key-Technology Prototype → Exascale Supercomputer
Exascale Supercomputer in China
Prototype system
Exascale Supercomputer in China

Prototype system

**System Composition**

- 512 MT-2000+ computing nodes
- 6 Computing Cabinets
- Performance: 3.146 PFlops
- 128 FT-2000+ nodes
Exascale Supercomputer in China

ARM HPC
Exascale Supercomputer in China
System Architecture--Hybrid

Sunway chip level
Sugon board level
Tianhe system level
Exascale Supercomputer in China

Exascale computer: Flexible architecture design

- Master slave acceleration mode
- Multi-core computing mode
- Many-core computing mode: AI computing

High-speed interconnection network

CPU
Exascale Supercomputer in China
Diversity Future
Exascale Supercomputer in China

Software Definition System
- cloud
- Workflow
- Digital twins

USER

Supercomputer
Contents

3. Exascale Supercomputer in China: From Prototype to E-System
4. Some Exploratory Work of Lattice QCD on E-prototype System
Lattice QCD on E-prototype System

IHEP, CAS: Ming Gong, etc

ITP, CAS: Yibo Yang

Peking University: Xu Feng, etc
The one node performance is close to the KNL(TGCC Irene).
Current version:

- A D-slash function and a simple inverter for Wilson fermion.
- Two-level parallelization with MPI for MPEs and Athread for CPEs.
- Fixed-sized data partition: $16^4$ lattice block on each core group.
- Compression on transferring SU(3) matrix between MPEs and CPEs.
- Fully optimized with SIMD instructions.

Working in progress:

- Adopting our “C Virtual Machine” (CVM) to achieve more FLOPs.
  - CVM is a preliminary optimization platform for Sunway architecture.
  - Gene algorithm is adopted to seek potential accelerations.
  - It is designed for lattice QCD and can be used for other applications.
- Developing an interface to QOPQDP and Chroma.
Much more work on road!
Thank You!

mengxf@nscc-tj.cn
http://www.nscc-tj.cn