Dr. Haohuan Fu is an associate professor in the Ministry of Education Key Laboratory for Earth System Modeling, and Center of Earth System Science in Tsinghua University. He is also the associate director of the National Supercomputing Center in Wuxi. His research interests include design methodologies for highly efficient and highly scalable simulation applications that can take advantage of emerging multi-core, many-core, and reconfigurable architectures, and make full utilization of current Peta-Flops and future Exa-Flops supercomputers; and intelligent data management, analysis, and data mining platforms that combine the statistical methods and machine learning technologies.

Sunway TaihuLight: Extreme Computing and Big Data
Sunway TaihuLight: Extreme Computing and Big Data Analytics

Haohuan Fu
National Supercomputing Center in Wuxi
Department of Earth System Science, Tsinghua University

June 19th 2017 @ Frankfurt
Outline

Sunway Machine: the Challenges and Opportunities

Scientific Computing with 10 Million Cores

Machine Learning Platform: A First Step with swDNN
The Sunway Machine Family

Sunway-I:
- CMA service, 1998
- commercial chip
- 0.384 Tflops
- 48th of TOP500

Sunway BlueLight:
- NSCC-Jinan, 2011
- 16-core processor
- 1 Pflops
- 14th of TOP500

Sunway TaihuLight:
- NSCC-Wuxi, 2016
- 260-core processor
- 125 Pflops
- 1st of TOP500
SW26010: Sunway 260-Core Processor
10M-core system

\[ 40 \times 1,024 \times 4 \times 65 = 10,649,600 \]

- 163,840 processes
- 65 threads
- racks
- chips
- core-groups
- cores
- total number of cores

260-core Chip

Rack

System
Sunway TaihuLight: Challenges and Opportunities

Pros
- 100+ PF Computing Power
- 6+ Gflops/Watt
- New chip with new features

Cons
- 10 million heterogeneous cores
- the memory wall
- the porting challenge
Outline

1. Sunway Machine: the Architecture and the Programming Model
2. Scientific Computing with 10 Million Cores
An (Incomplete) List of Full-Scale Applications

**2016**
- Fully Implicit Solver Nonhydrostatic Atmospheric Dynamics
- Ultra-high Resolution Surface Wave Modeling
- Extreme-scale Phase Field Simulations of Coarsening Dynamics
- Peta-scale Atomistic Simulation of Silicon Nanowires
- Run-away Electron Trajectory Simulation of Tokamak Magnetic Confinement Nuclear Fusion
- Genome Functional Annotation and Homeotic Gene Building
- Spacecraft CFD Numerical Simulation

**2017**
- Extreme-scale Graph Processing Framework
- Extreme-scale Global Simulation of Planetary Rings
- Faithful Simulations of Highly Entangled Quantum Spin Liquid States via PEPS++
- Highly Scalable Molecular Dynamics Simulation of Condensed Covalent Materials
- Model Optimizer for cryo-EM Biological Macromolecule Structure Determination
- Redesigning CAM-SE for Peta-Scale Climate Modeling Performance on Sunway TaihuLight
- 15-Pflops Nonlinear Earthquake Simulation on Sunway TaihuLight: Enabling Depiction of Realistic 10 Hz Scenarios
An (Incomplete) List of Full-Scale Applications

2016
- Fully Implicit Solver Nonhydrostatic Atmospheric Dynamics
- Ultra-high Resolution Surface Wave Modeling
- Extreme-scale Phase Field Simulations of Coarsening Dynamics
- Peta-scale Atomistic Simulation of Silicon Nanowires
- Run-away Electron Trajectory Simulation of Tokamak Magnetic Confinement Nuclear Fusion
- Genome Functional Annotation and Homeotic Gene Building
- Spacecraft CFD Numerical Simulation

2017
- Extreme-scale Graph Processing Framework
- Extreme-scale Global Simulation of Planetary Rings
- Faithful Simulations of Highly Entangled Quantum Spin Liquid States via PEPS++
- Highly Scalable Molecular Dynamics Simulation of Condensed Covalent Materials
- Model Optimizer for cryo-EM Biological Macromolecule Structure Determination
- Redesigning CAM-SE for Peta-Scale Climate Modeling Performance on Sunway TaihuLight
- 15-Pflops Nonlinear Earthquake Simulation on Sunway TaihuLight: Enabling Depiction of Realistic 10 Hz Scenarios
## An (Incomplete) List of Full-Scale Applications

<table>
<thead>
<tr>
<th>Year</th>
<th>Applications</th>
</tr>
</thead>
</table>
| 2016 | - Fully Implicit Solver Nonhydrostatic Atmospheric Dynamics  
- Ultra-high Resolution Surface Wave Modeling  
- Extreme-scale Phase Field Simulations of Coarsening Dynamics  
- Peta-scale Atomistic Simulation of Silicon Nanowires  
- Run-away Electron Trajectory Simulation of Tokamak Magnetic Confinement Nuclear Fusion  
- Genome Functional Annotation and Homeotic Gene Building  
- Spacecraft CFD Numerical Simulation |
| 2017 | - Extreme-scale Graph Processing Framework  
- Extreme-scale Global Simulation of Planetary Rings  
- Faithful Simulations of Highly Entangled Quantum Spin Liquid States via PEPS++  
- Highly Scalable Molecular Dynamics Simulation of Condensed Covalent Materials  
- Model Optimizer for cryo-EM Biological Macromolecule Structure Determination  
- Redesigning CAM-SE for Peta-Scale Climate Modeling Performance on Sunway TaihuLight  
- 15-Pflops Nonlinear Earthquake Simulation on Sunway TaihuLight: Enabling Depiction of Realistic 10 Hz Scenarios |
An (Incomplete) List of Full-Scale Applications

2016
• Fully Implicit Nonhydrostatic Solver for Atmospheric Dynamics

2017
• Redesigning CAM-SE for Peta-Scale Climate Modeling Performance on Sunway TaihuLight
• 15-Pflops Nonlinear Earthquake Simulation on Sunway TaihuLight: Enabling Depiction of Realistic 10 Hz Scenarios
The Gap between Software and Hardware

China’s models
- pure CPU code
- scaling to hundreds or thousands of cores

China’s supercomputers
- heterogeneous systems with many-core chips
- millions of cores

- millions lines of legacy code
- poor scalability
- written for multi-core, rather than many-core
Our Research Goals

- highly scalable framework that can efficiently utilize many-core processors
- automated tools to deal with the legacy code

China’s models
- pure CPU code
- scaling to hundreds or thousands of cores

China’s supercomputers
- heterogeneous systems with many-core chips
- millions of cores

- millions lines of legacy code
- poor scalability
- written for multi-core, rather than many-core
Our Research Goals

- highly scalable framework that can efficiently utilize many-core processors
- automated tools to deal with the legacy code

China’s models
- pure CPU code
- scaling to hundreds or thousands of cores

China’s supercomputers
- heterogeneous systems with many-core chips
- millions of cores
163,840 processes

65 threads

racks | chips | core-groups | cores | total number of cores

\[ 40 \times 1,024 \times 4 \times 65 = 10,649,600 \]

Now let’s find a way to design a subdomain solver.
163,840 processes
65 threads

163,840 processes

racks | chips | core-groups | cores | total number of cores

40 \times 1,024 \times 4 \times 65 = 10,649,600

DD-MG $\kappa$-cycle

Two issues:
1. Data locality
2. Global synchronization

Parallel ILU with level-scheduling
40 × 1,024 × 4 × 65 = 10,649,600

Two issues:
1. Multiple sweeps needed
2. Global synchronization
163,840 processes

65 threads

**racks**  **chips**  **core-groups**  **cores**  **total number of cores**

$40 \times 1,024 \times 4 \times 65 = 10,649,600$

**DD-MG $\kappa$-cycle**

**Geometry-based pipelined ILU (GP-ILU)**

- Subdomain matrix of 1st-order with geometric index
- Our goal of design:
  1. Single sweep
  2. Synchronization-free
  3. Improved data-locality

$reg\_size = \frac{(num\_cores-1)+blk\_height}{cell\_size} \times dim_z$
The 3-km res run: 1.01 SYPD with 10.6M cores, dt=240s, I/O penalty <5%
Weak-scaling results

Resolution (km)

DOFs=772B

7.95 DP-PF

“Exa-scale” for exp

89.5X

23.66 DP-PF

SYPD

Total number of cores

The 488-m res run: 0.07 SYPD, 10.6M cores, dt=240s, 89.5X speedup over explicit
Our Research Goals

- highly scalable framework that can efficiently utilize many-core processors
- automated tools to deal with the legacy code

China’s models
- pure CPU code
- scaling to hundreds or thousands of cores

China’s supercomputers
- heterogeneous systems with many-core chips
- millions of cores

- millions lines of legacy code
- poor scalability
- written for multi-core, rather than many-core
The CESM Project on Sunway TaihuLight

- CAM5.0
- POP2.0
- CLM4.0
- CICE4.0
- CPL7
- CESM1.2.0

Tsinghua + BNU 30+ Professors and Students

- Four component models, millions lines of code
- Large-scale run on Sunway TaihuLight
  - 24,000 MPI processes
  - Over one million cores
- 10-20x speedup for kernels
- 2-3x speedup for the entire model

"Refactoring and Optimizing the Community Atmosphere Model (CAM) on the Sunway TaihuLight Supercomputer", in Proceedings of SC 2016.
Major Challenges

- A high complexity in application, and a heavy legacy in the code base (millions lines of code)
- An extremely complicated MPMD program with no hotspots (or hundreds of hotspots)
- Misfit between the in-place design philosophy and the new architecture
- Lack of people with interdisciplinary knowledge and experience
Workflow of CAM

After initialization, the physics and the dynamics are executed in turn during each simulation time-step.
Porting of CAM: General Idea

- Entire code base: 530,000 lines of code

- Components with regular code patterns
  - e.g. the CAM-SE dynamic core
  - manual OpenACC parallelization and optimization on code and data structures

- Components with irregular and complex code patterns
  - e.g. the CAM physics schemes
  - loop transformation tool to expose the right level of parallelism and code size
  - memory footprint analysis and reduction tool
Speedup of Major Kernels in CAM-SE

7x to 22x speedup for computing intensive kernels;
2x to 7x speedup for memory-bound kernels
The `microp_mg1_0` kernel demonstrates a significant speedup of 14x, as the intermediate variables and arrays provide a nice fit to the SPM of the CPE clusters after the automated optimizations.
CAM model: scalability and speedup

• million core scale, 2.81 SYPD
• many-core refactoring for the entire model
• competitive simulation speed to the same model on NCAR Yellowstone

![Simulation Speed Chart](chart.png)

Simulation Speed (Described in Model Year Per Day (MYPD))

Number of CGs (each CG includes 1 MPE and 64 CPEs)

- MPE only
- MPE+CPE for dynamic core
- MPE+CPE for both dynamic core and physics schemes

- 1024
- 2400
- 4096
- 5120
- 7350
- 9600
- 12000
- 24000

CAM model: scalability and speedup

• million core scale, 2.81 SYPD
• many-core refactoring for the entire model
• competitive simulation speed to the same model on NCAR Yellowstone

![Simulation Speed Chart](chart.png)

Simulation Speed (Described in Model Year Per Day (MYPD))

Number of CGs (each CG includes 1 MPE and 64 CPEs)

- MPE only
- MPE+CPE for dynamic core
- MPE+CPE for both dynamic core and physics schemes

- 1024
- 2400
- 4096
- 5120
- 7350
- 9600
- 12000
- 24000
A Complete Redesign of the Dynamic Core
15-Pflops Nonlinear Earthquake Simulation on Sunway TaihuLight: Enabling Depiction of Realistic 10 Hz Scenarios

- Fully optimized for Sunway TaihuLight
- Achieving the same efficiency as Titan, with only 1/3 bandwidth per node.
- High-fidelity simulation of the Tangshan earthquake using over 10 million cores  
  - 320 km x 320 km x 60 km  
  - 10Hz non-linear simulation
Outline

1. Sunway Machine: the Architecture and the Programming Model
2. Scientific Computing with 10 Million Cores
swCaffe: Distributed Caffe on Sunway TaihuLight Supercomputer

- Scale-up: swDNN library for high performance

- Scale-out:
  - Design MPI-based parameter server for distributed training
  - Design 4 CG parallel strategy on one worker based on shared memory
Scale-out

- Tensor Transformation with CPEs + swDNN Library

Caffe Layout: [B, C, H, W]

swDNN Layout: [B, C, H, W]

Training AlexNet with swCaffe

- Intel (24 cores)
  - total: 1.0x
  - convolution: 2.2x
  - fully connected: 3.5x

- swBLAS (1CG)
  - total: 1.0x
  - convolution: 7.1x
  - fully connected: 4.5x

- swDNN (1CG)
  - total: 1.0x
  - convolution: 8.5x
  - fully connected: 8.5x
MPI-based parameter server

- VGG16 with synchronous SGD
  - Weak Scalability: 22.4x speedup with 32 workers (Data Parallel)
  - Strong Scalability: 3x speedup with 4 workers, 4.7x with 16 workers due to small batch
4 CGs parallel design

- Each worker uses 4 CGs
- Shared memory allows data parallel without communication
- 4x speedup over 1CG worker
Long Term Plan

- **Traditional HPC Applications** *(Science -> Service)*
  - weather / climate service
  - seismic data processing service
  - CFD simulation framework for Advanced Manufacturing

- **Deep Learning Related Applications**
  - the swDNN framework
  - collaborating with face++ for face recognition applications
  - collaborating with Sogou for voice recognition and translation
  - strategic collaboration with Baidu on both software and hardware (Sunway learning chip?)

- **Sunway Micro**
  - commercial, standardized 1U/2U, customizable (test systems available in August, available from market in November)