国家超级计算无锡中心
National Supercomputing Center in Wuxi

Sunway TaihuLight: Designing and Tuning Scientific Applications at the Scale of 10 Million Cores

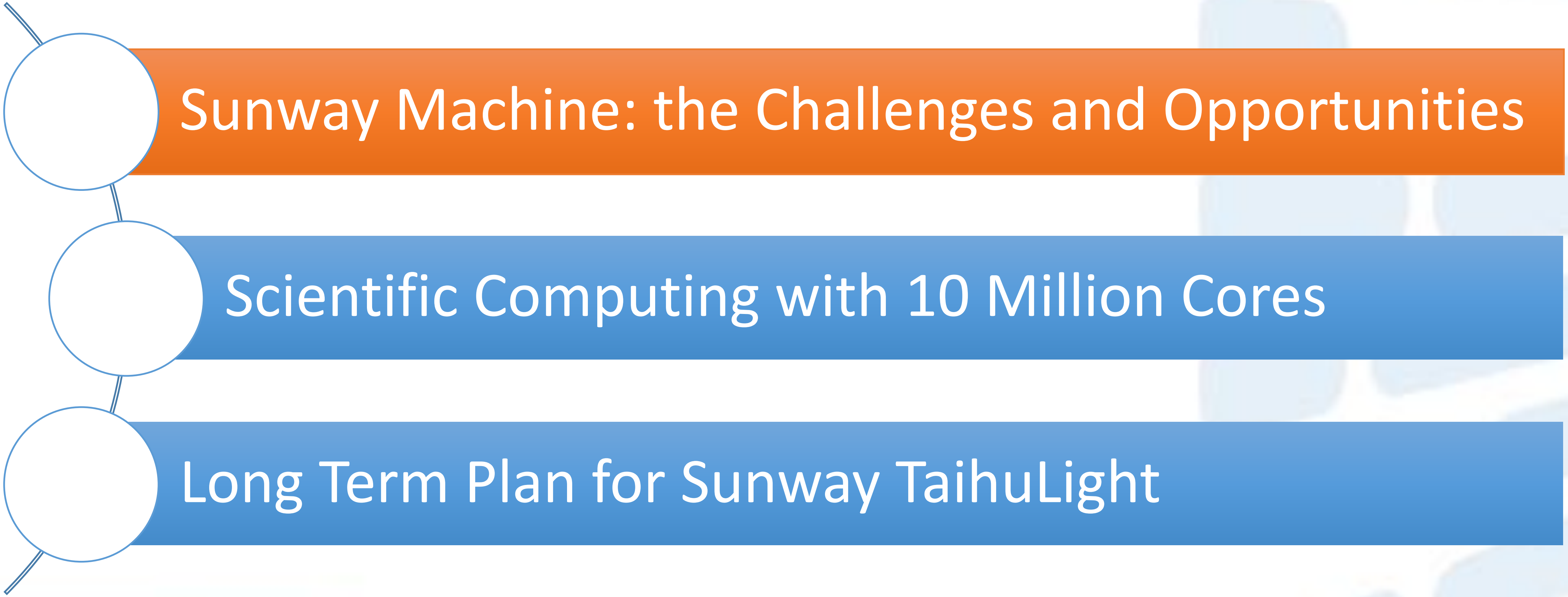
Haohuan Fu

National Supercomputing Center in Wuxi

Department of Earth System Science, Tsinghua University

September 13th 2017 @ ICAS

Outline



- Sunway Machine: the Challenges and Opportunities
- Scientific Computing with 10 Million Cores
- Long Term Plan for Sunway TaihuLight

The Sunway Machine Family

Sunway-I:

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

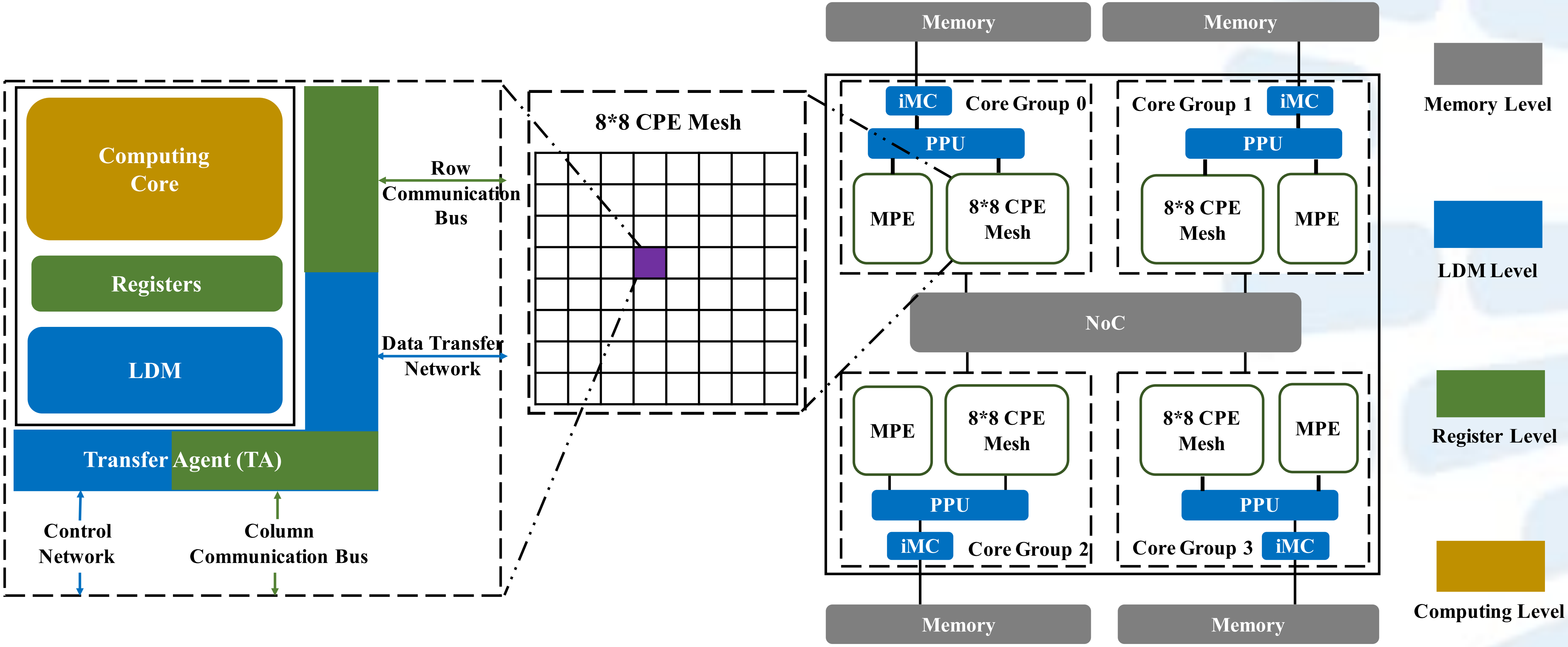
Sunway BlueLight:

- 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



High-Density Integration of the Computing System

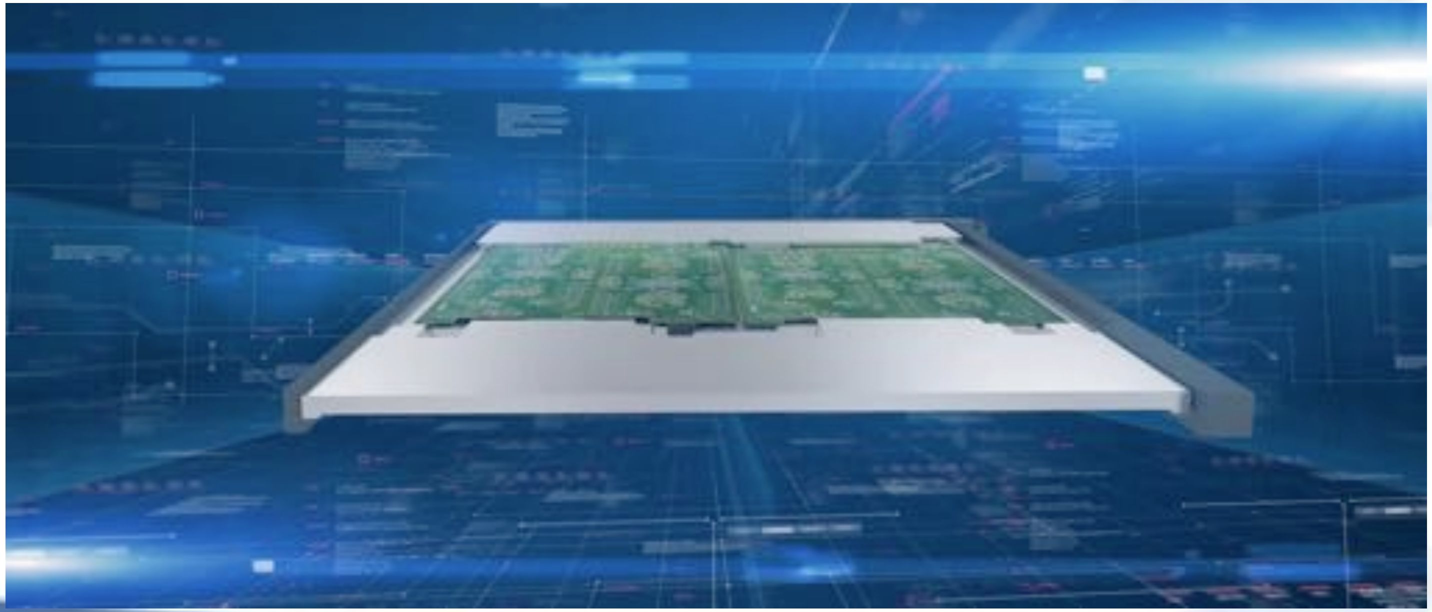
- A Five-Level Integration Hierarchy
  - computing node
  - computing board
  - super node
  - cabinet
  - entire computing system

High-Density Integration of the Computing System

- A Five-Level Integration Hierarchy
  - computing node
  - computing board
  - super node
  - cabinet
  - entire computing system

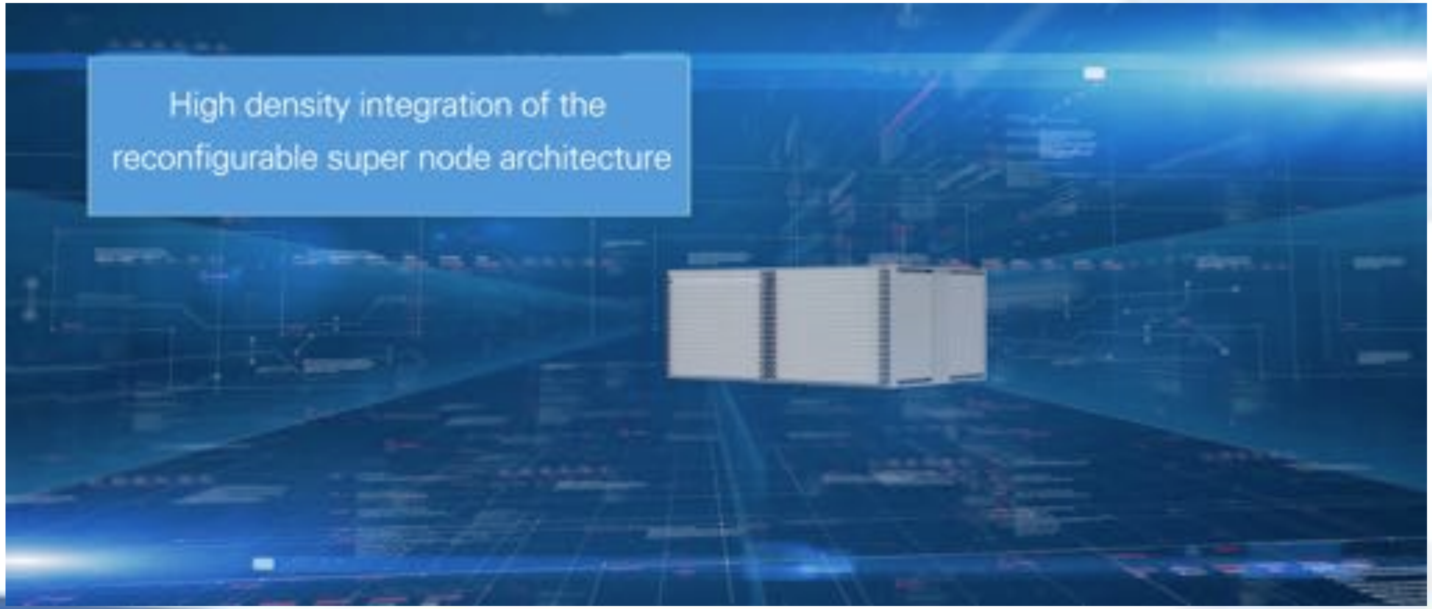
High-Density Integration of the Computing System

- A Five-Level Integration Hierarchy
  - computing node
  - computing board
  - super node
  - cabinet
  - entire computing system



High-Density Integration of the Computing System

- A Five-Level Integration Hierarchy
  - computing node
  - computing board
  - super node
  - cabinet
  - entire computing system


High density integration of the reconfigurable super node architecture

High-Density Integration of the Computing System

- A Five-Level Integration Hierarchy
  - computing node
  - computing board
  - super node
  - cabinet
  - entire computing system

How to Connect the 10 Million Cores?

$$40 \times 4 \times 256 \times 4 \times (1+8 \times 8) = 10{,}649{,}600$$

How to Connect the 10 Million Cores?

$$40\times4\times256\times4\times(1+\underbrace{8\times8}_{\substack{\text{2D core array}\\ \text{with row and}\\ \text{column buses}}}) = 10,649,600$$

How to Connect the 10 Million Cores?

$$40\times4\times256\times\underbrace{4\times(1+\underbrace{8\times8}_{\text{2D core array with row and column buses}})}_{\text{Network on Chip}} = 10,649,600$$

How to Connect the 10 Million Cores?

$$40\times4\times256\times4\times(1+\underbrace{8\times8}) = 10{,}649{,}600$$

2D core array with row and column buses

Network on Chip

Customized Network Board to Fully Connect 256 Nodes

How to Connect the 10 Million Cores?

$$40 \times 4 \times 256 \times 4 \times (1+8 \times 8) = 10{,}649{,}600$$

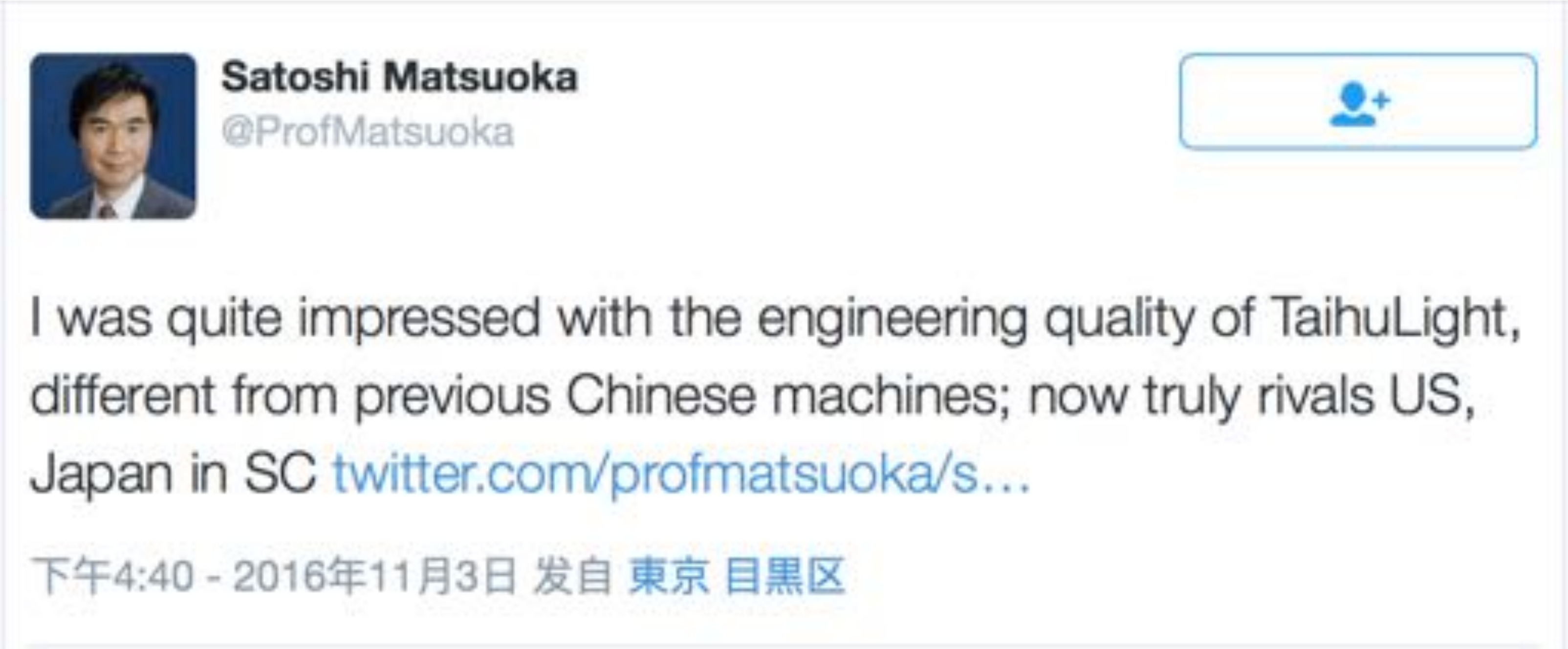
2D core array with row and column buses

Network on Chip

Customized Network Board to Fully Connect 256 Nodes

Sunway Net

Tweet Comments from Prof. Satoshi Matsuoka

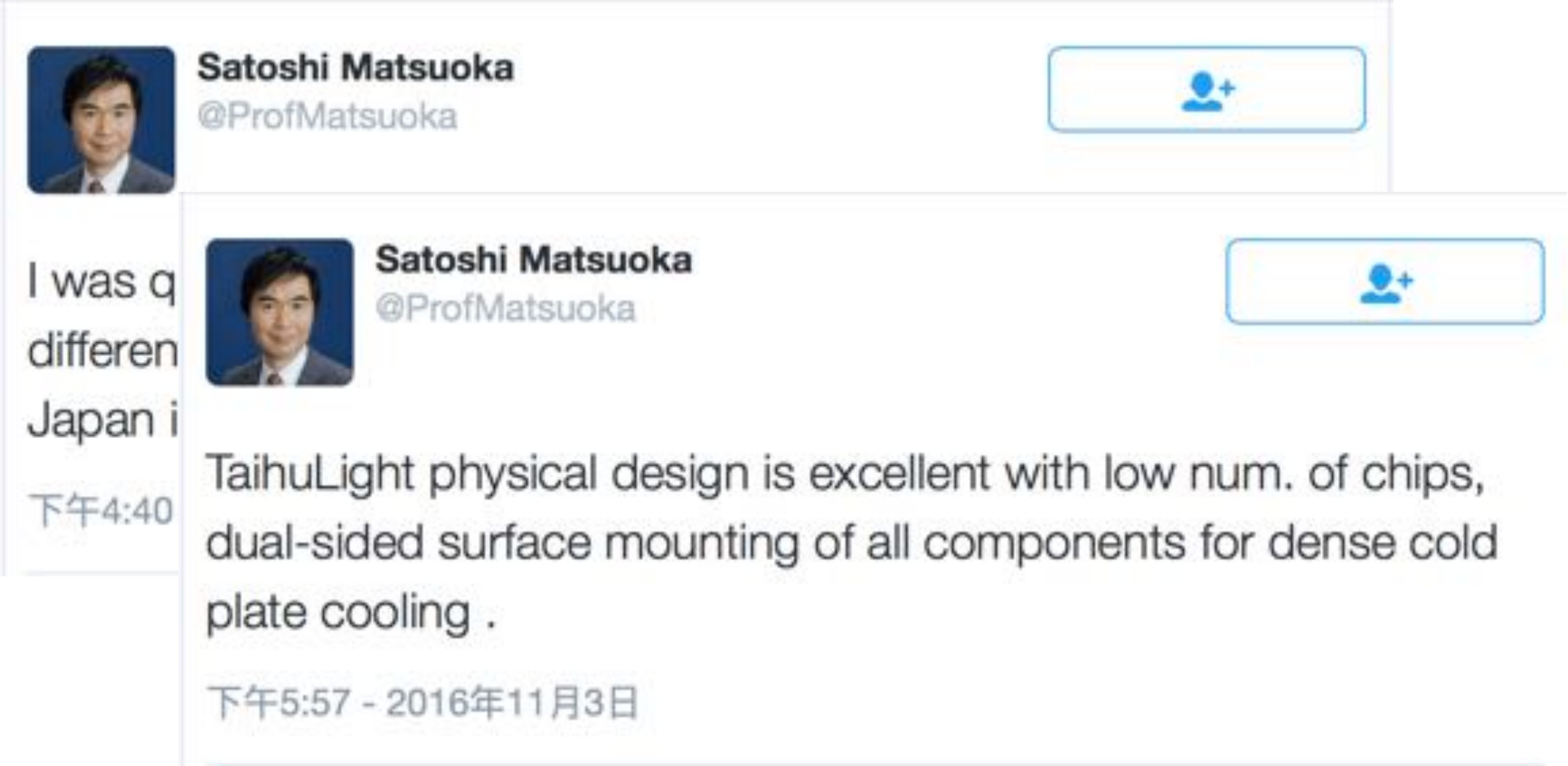


Satoshi Matsuoka
@ProfMatsuoka

I was quite impressed with the engineering quality of TaihuLight, different from previous Chinese machines; now truly rivals US, Japan in SC twitter.com/profmatsuoka/s…

下午4:40 - 2016年11月3日 发自 東京 目黒区

Tweet Comments from Prof. Satoshi Matsuoka



Satoshi Matsuoka
@ProfMatsuoka

I was q
differen
Japan i

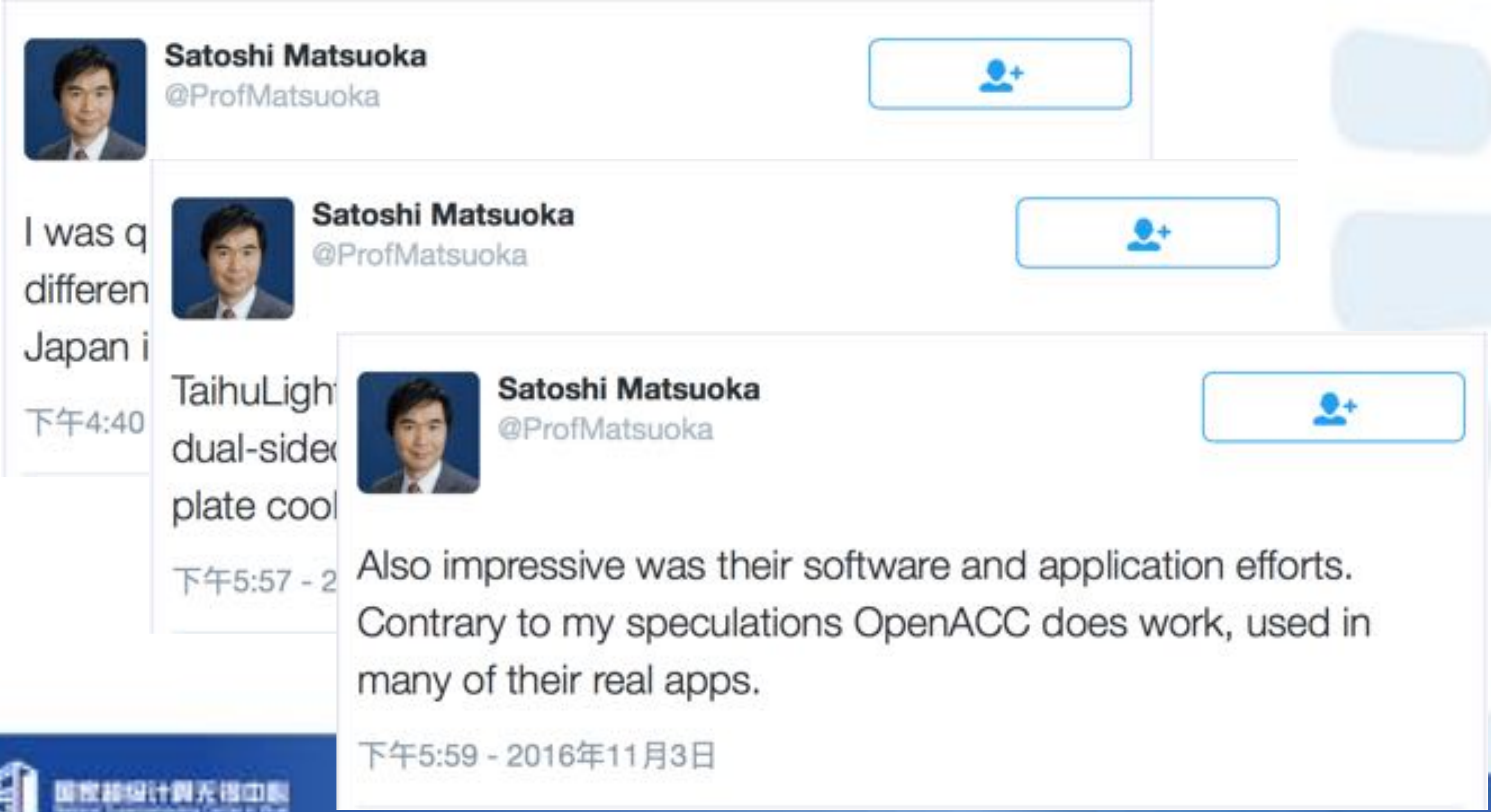
下午4:40

Satoshi Matsuoka
@ProfMatsuoka

TaihuLight physical design is excellent with low num. of chips, dual-sided surface mounting of all components for dense cold plate cooling .

下午5:57 - 2016年11月3日

Tweet Comments from Prof. Satoshi Matsuoka



Satoshi Matsuoka
@ProfMatsuoka

I was q
differen
Japan i

下午4:40

Satoshi Matsuoka
@ProfMatsuoka

TaihuLigh
dual-sided
plate cool

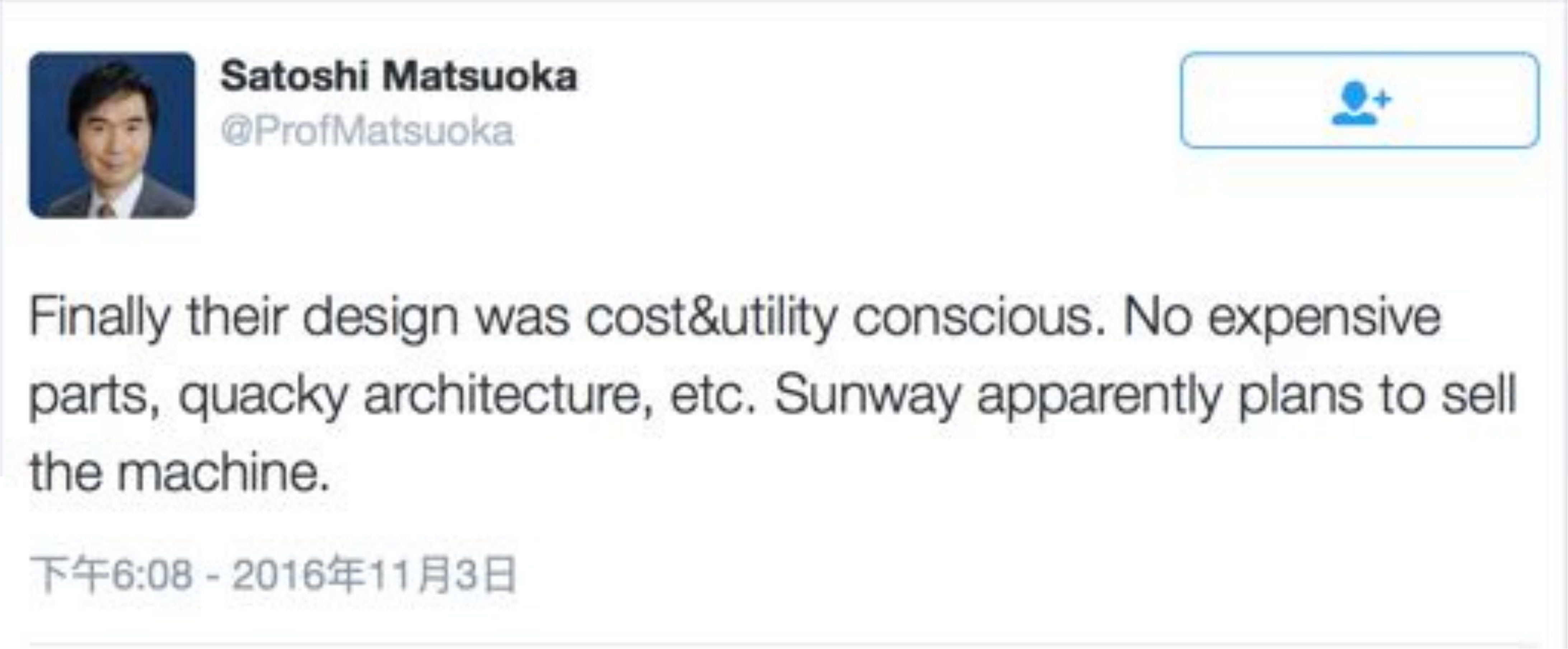
下午5:57 - 2

Satoshi Matsuoka
@ProfMatsuoka

Also impressive was their software and application efforts. Contrary to my speculations OpenACC does work, used in many of their real apps.

下午5:59 - 2016年11月3日

Tweet Comments from Prof. Satoshi Matsuoka



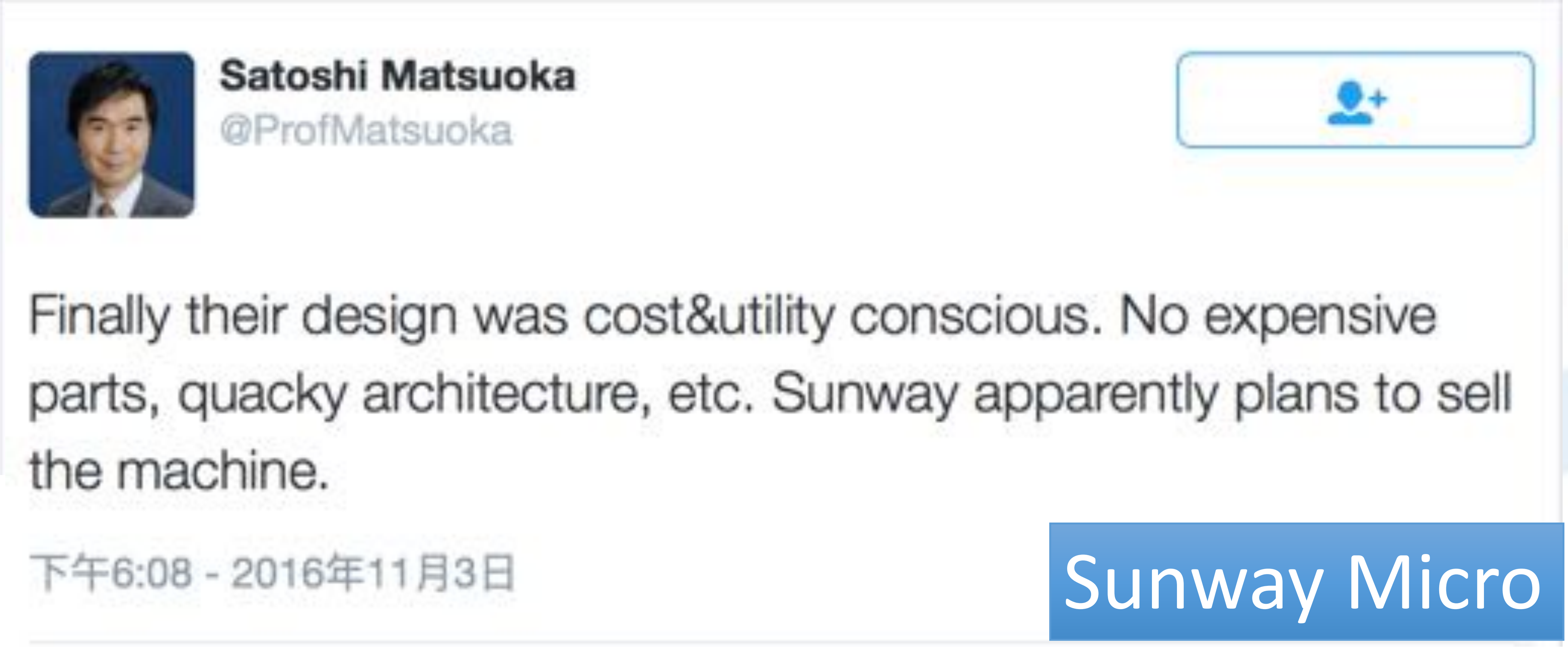
Satoshi Matsuoka

@ProfMatsuoka

Finally their design was cost&utility conscious. No expensive parts, quacky architecture, etc. Sunway apparently plans to sell the machine.

下午6:08 - 2016年11月3日

Tweet Comments from Prof. Satoshi Matsuoka



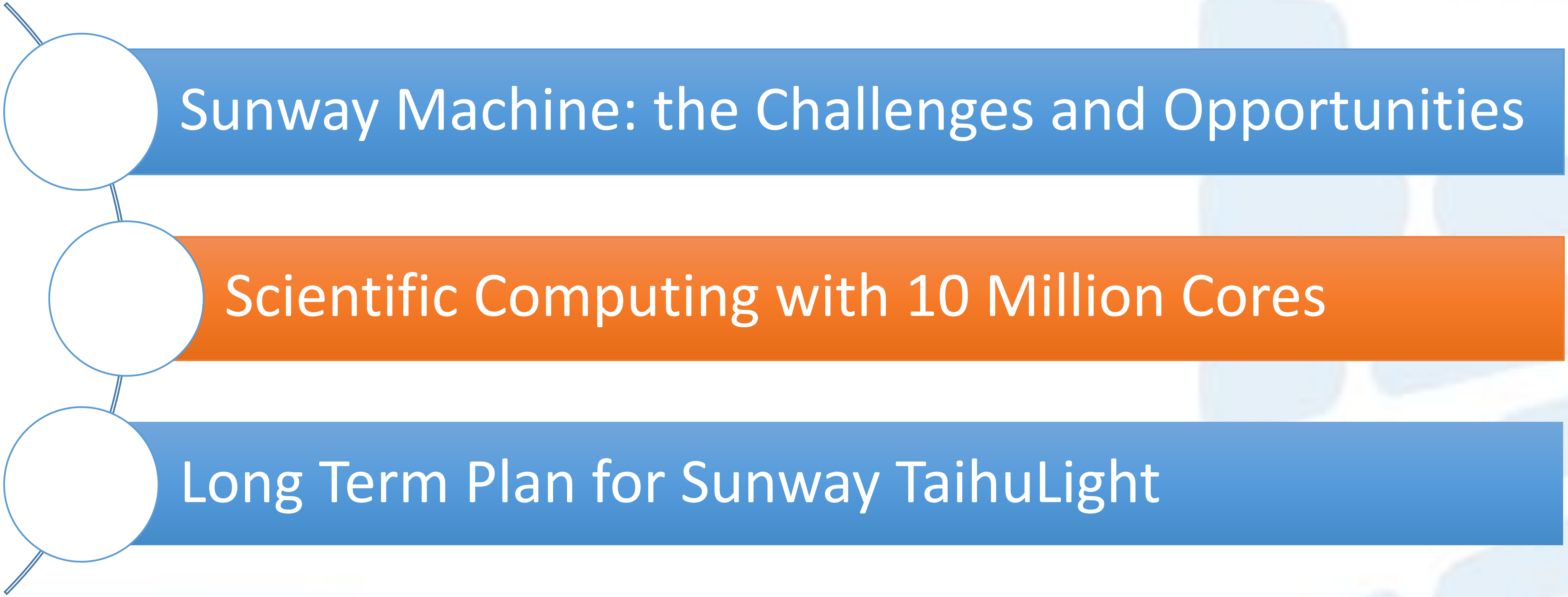
Satoshi Matsuoka
@ProfMatsuoka

Finally their design was cost&utility conscious. No expensive parts, quacky architecture, etc. Sunway apparently plans to sell the machine.

下午6:08 - 2016年11月3日

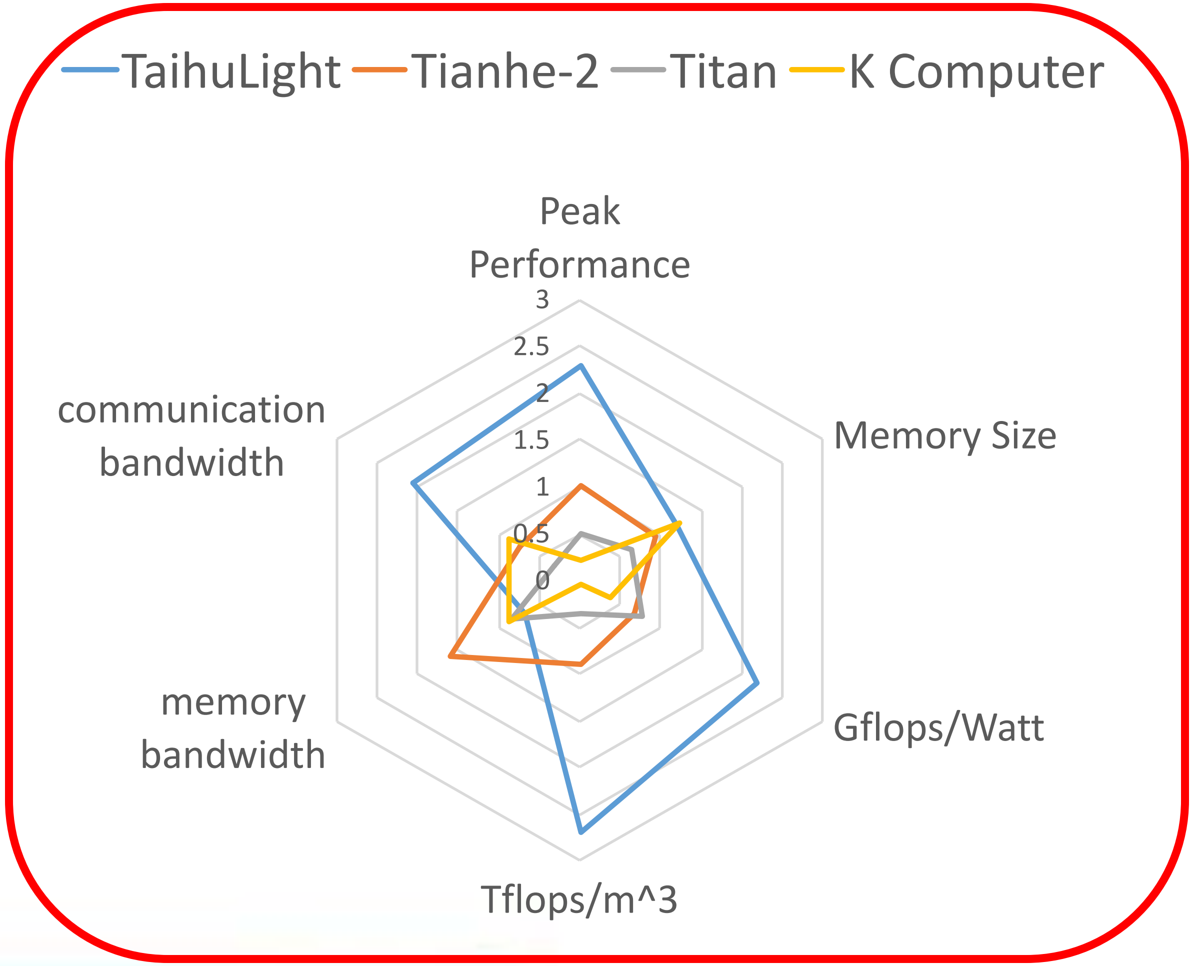
Sunway Micro

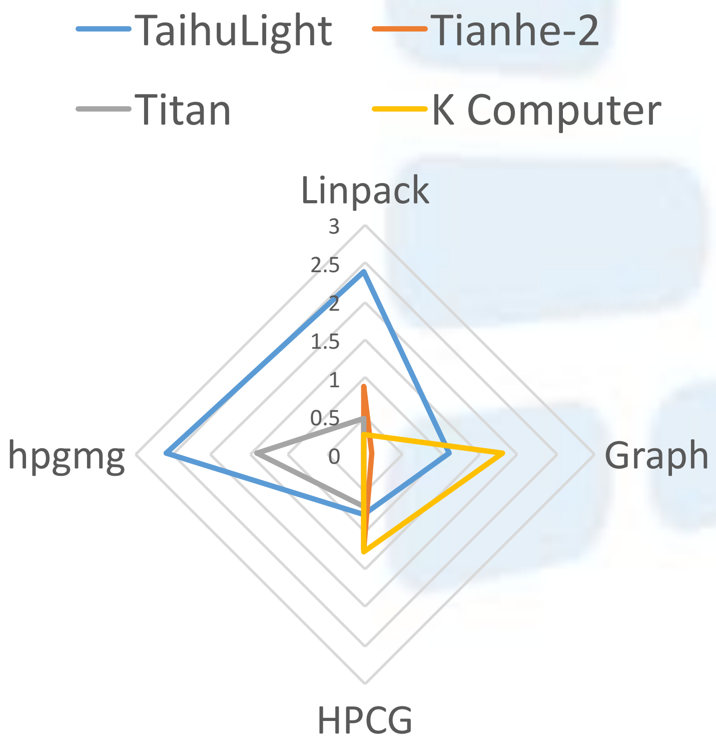
Outline



- Sunway Machine: the Challenges and Opportunities
- Scientific Computing with 10 Million Cores
- Long Term Plan for Sunway TaihuLight

Machine Capability Comparison





Major Features to Consider

Sunway TaihuLight

125 Pflops

10 million cores

user-controlled 64 KB LDM

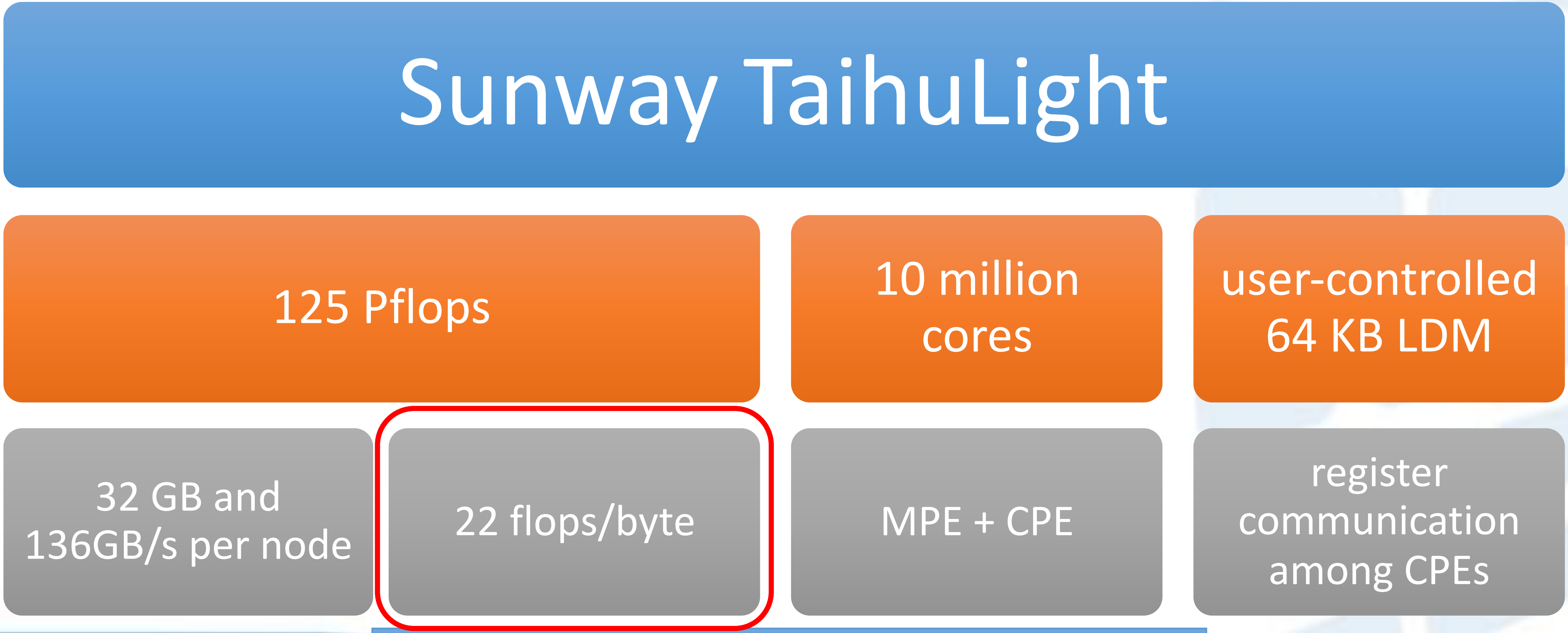
32 GB and 136GB/s per node

22 flops/byte

MPE + CPE

register communication among CPEs

Major Features to Consider



Sunway TaihuLight

125 Pflops

10 million cores

user-controlled 64 KB LDM

32 GB and 136GB/s per node

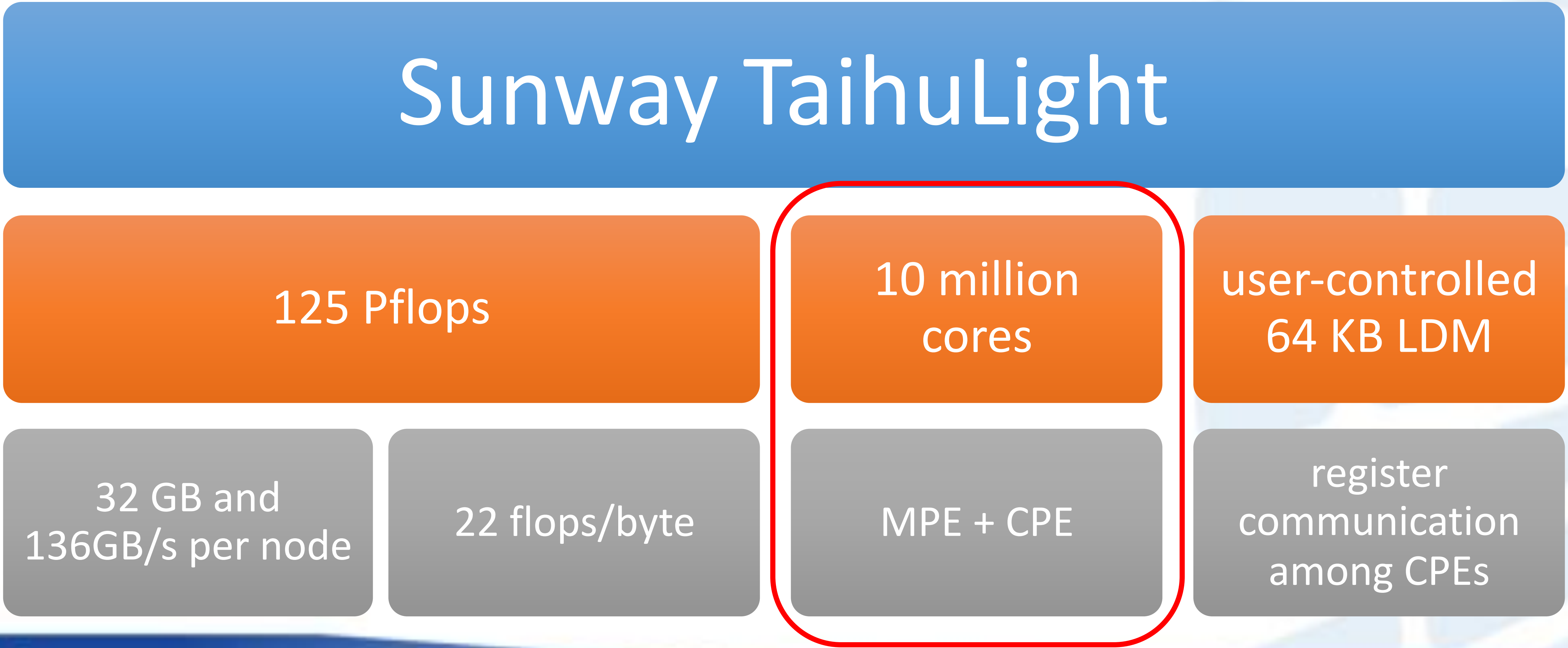
22 flops/byte

MPE + CPE

register communication among CPEs

Intel KNL 7250 of Cori:	6.5 flops/byte
NVIDIA P100 of Piz Daint:	7.2 flops/byte

Major Challenge #1: Scaling



Sunway TaihuLight

125 Pflops

10 million cores

user-controlled 64 KB LDM

32 GB and 136GB/s per node

22 flops/byte

MPE + CPE

register communication among CPEs

Major Challenge #2: Memory Wall

Sunway TaihuLight

125 Pflops

10 million cores

user-controlled 64 KB LDM

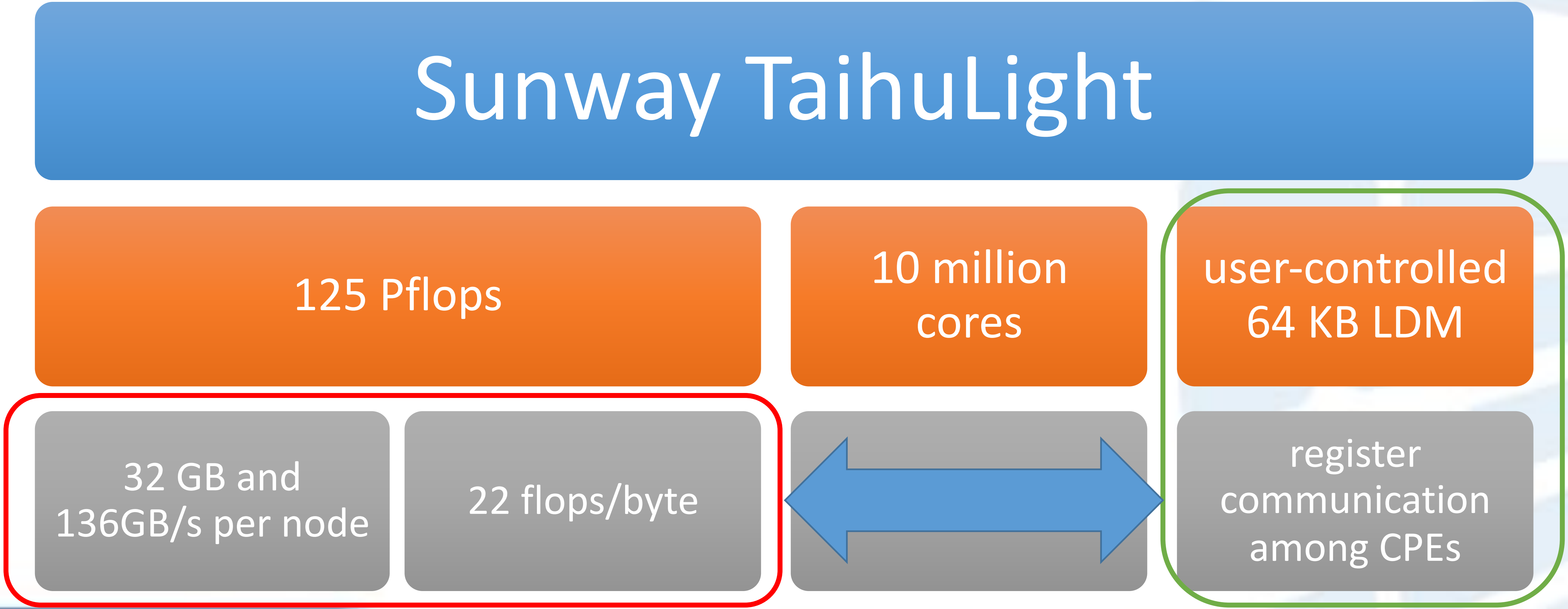
32 GB and 136GB/s per node

22 flops/byte

MPE + CPE

register communication among CPEs

Major Challenge #2: Memory Wall



Sunway TaihuLight

125 Pflops

10 million cores

user-controlled 64 KB LDM

32 GB and 136GB/s per node

22 flops/byte

register communication among CPEs

Refactoring and Redesigning

An (Incomplete) List of Full-Scale Applications

2016	2017
Fully Implicit Solver for Atmospheric Dynamics	Extreme-scale Graph Processing Framework
Surface Wave Modeling	Simulation of Planetary Rings
Phase Field Simulations of Coarsening Dynamics	Simulations of Quantum Spin Liquid States via PEPS++
Atomistic Simulation of Silicon Nanowires	Molecular Dynamics Simulation of Condensed Covalent Materials
Run-away Electron Trajectory Simulation	cryo-EM Macromolecule Structure Determination
Genome Functional Annotation and Homeotic Gene Building	Redesigning CAM-SE
Spacecraft CFD Numerical Simulation	Nonlinear Earthquake Simulation

An (Incomplete) List of Full-Scale Applications

2016 Gordon Bell Finalists

- Fully Implicit Solver for Atmospheric Dynamics
- Surface Wave Modeling
- Phase Field Simulations of Coarsening Dynamics
- Atomistic Simulation of Silicon Nanowires
- Run-away Electron Trajectory Simulation
- Genome Functional Annotation and Homeotic Gene Building
- Spacecraft CFD Numerical Simulation

2017

- Extreme-scale Graph Processing Framework
- Simulation of Planetary Rings
- Simulations of Quantum Spin Liquid States via PEPS++
- Molecular Dynamics Simulation of Condensed Covalent Materials
- cryo-EM Macromolecule Structure Determination
- Redesigning CAM-SE
- Nonlinear Earthquake Simulation

An (Incomplete) List of Full-Scale Applications

2016 Gordon Bell Prize	2017
Fully Implicit Solver for Atmospheric Dynamics	Extreme-scale Graph Processing Framework
Surface Wave Modeling	Simulation of Planetary Rings
Phase Field Simulations of Coarsening Dynamics	Simulations of Quantum Spin Liquid States via PEPS++
Atomistic Simulation of Silicon Nanowires	Molecular Dynamics Simulation of Condensed Covalent Materials
Run-away Electron Trajectory Simulation	cryo-EM Macromolecule Structure Determination
Genome Functional Annotation and Homeotic Gene Building	Redesigning CAM-SE
Spacecraft CFD Numerical Simulation	Nonlinear Earthquake Simulation

An (Incomplete) List of Full-Scale Applications

2016 Gordon Bell Prize

- Fully Implicit Solver for Atmospheric Dynamics
- Surface Wave Modeling
- Phase Field Simulations of Coarsening Dynamics
- Atomistic Simulation of Silicon Nanowires
- Run-away Electron Trajectory Simulation
- Genome Functional Annotation and Homeotic Gene Building
- Spacecraft CFD Numerical Simulation

2017 Gordon Bell Finalists

- Extreme-scale Graph Processing Framework
- Simulation of Planetary Rings
- Simulations of Quantum Spin Liquid States via PEPS++
- Molecular Dynamics Simulation of Condensed Covalent Materials
- cryo-EM Macromolecule Structure Determination
- Redesigning CAM-SE
- Nonlinear Earthquake Simulation

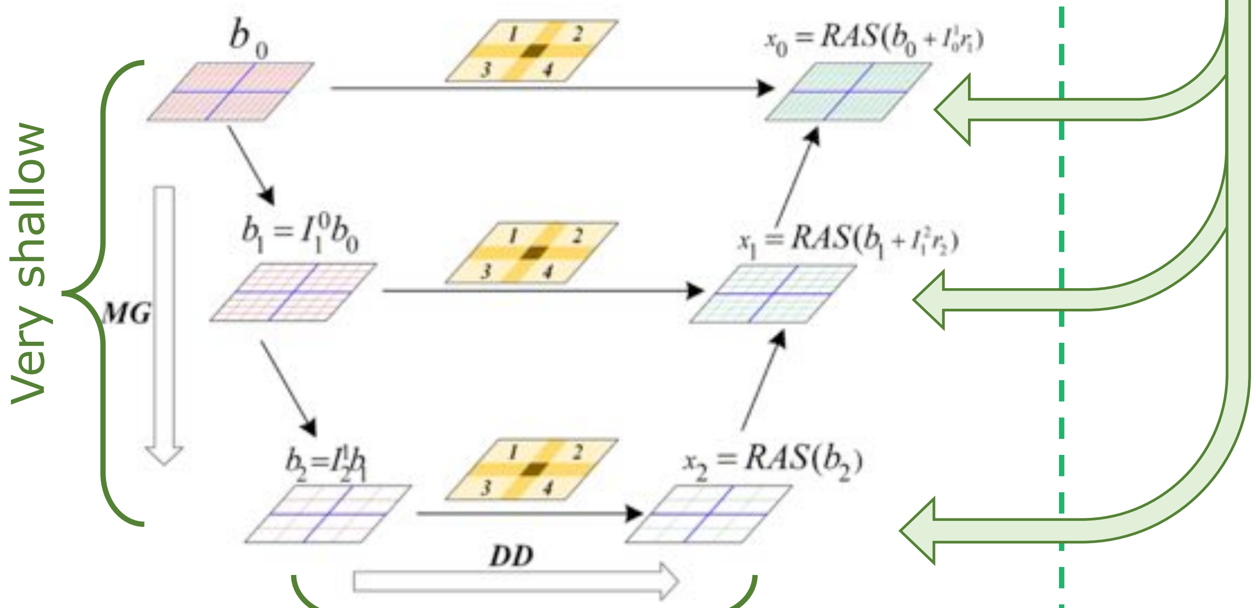
163,840 processes

65 threads

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

$$\underbrace{40 \times 1{,}024 \times 4}_{\text{163,840 processes}} \times \underbrace{65}_{\text{65 threads}} = 10{,}649{,}600$$

DD-MG $\mathcal{K}$-cycle



Very shallow

Uniform DD

Plug & Play

Now let's find a way to design a subdomain solver.

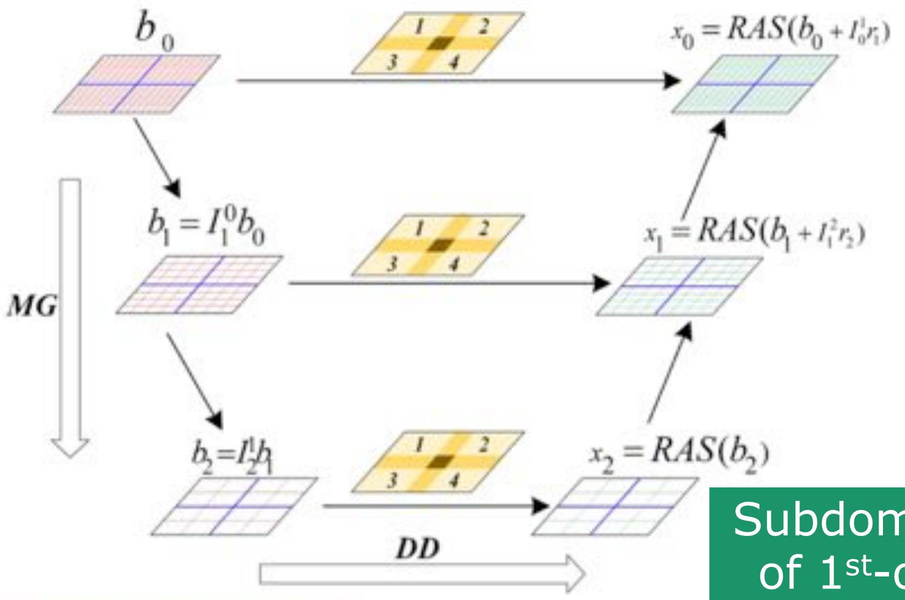
Application (I): Implicit Solver for Atmospheric Dynamics

163,840 processes | 65 threads

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

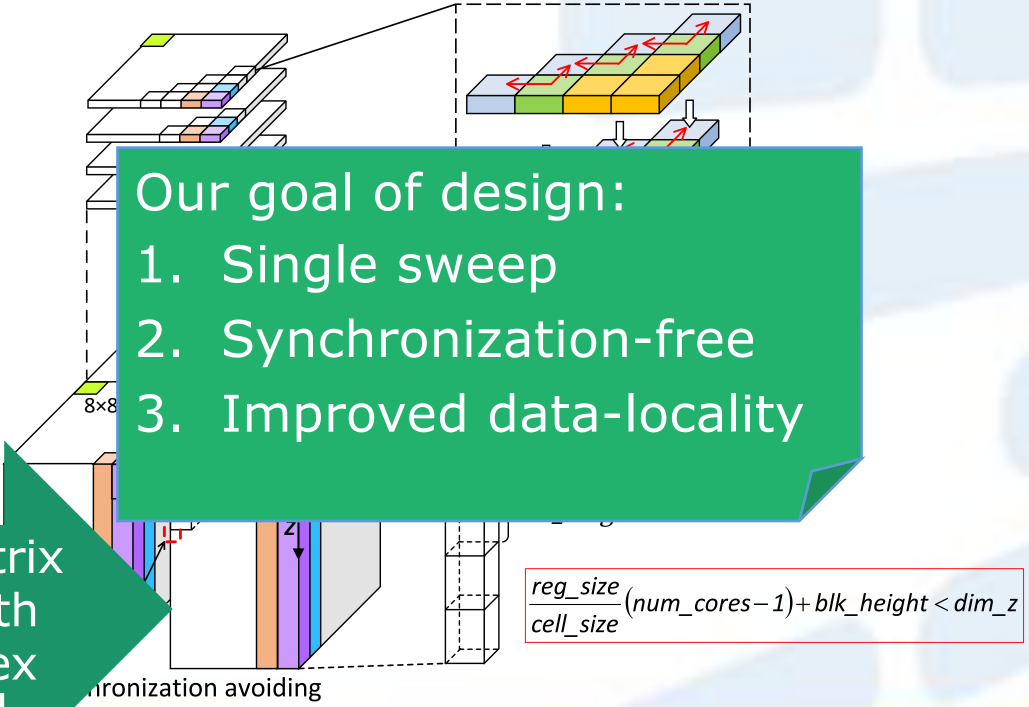
40 × 1,024 × 4 × 65 = 10,649,600

DD-MG $\mathcal{K}$-cycle



Subdomain matrix of 1st-order with geometric index

Geometry-based pipelined ILU (GP-ILU)

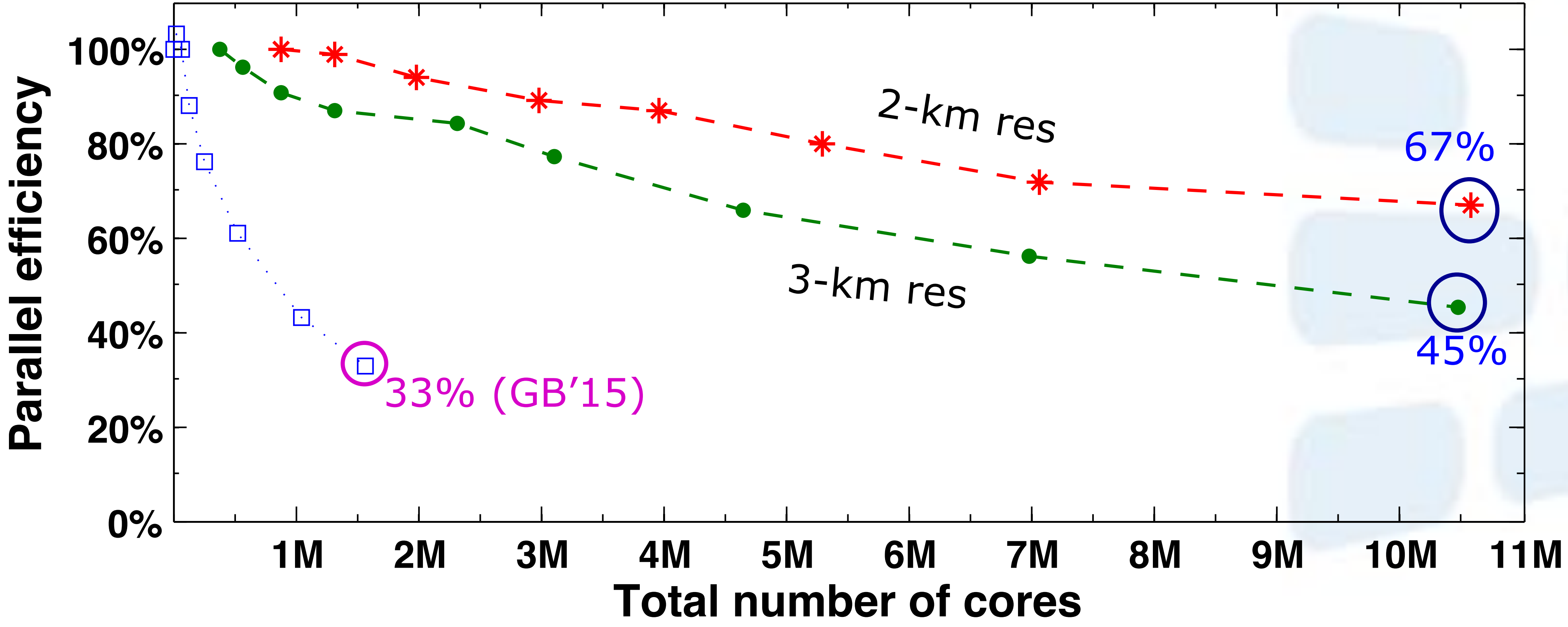


8×8
z
ronization avoiding
$\frac{reg\_size}{cell\_size}(num\_cores-1)+blk\_height<dim\_z$

Our goal of design:

1. Single sweep
2. Synchronization-free
3. Improved data-locality

Strong-scaling results



The 3-km res run: 1.01 SYPD with 10.6M cores, dt=240s, I/O penalty <5%

Weak-scaling results

DOFs=772B

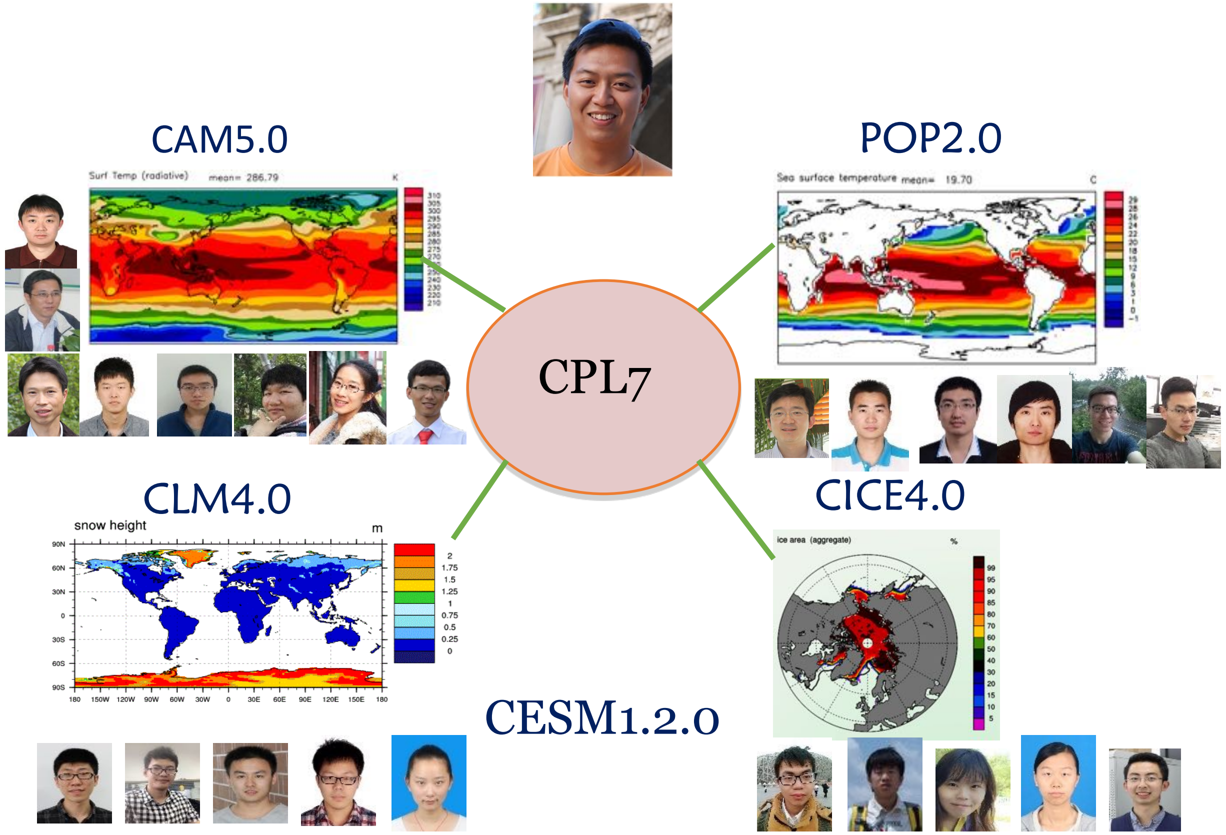
7.95 DP-PF

“Exa-scale” for exp

23.66 DP-PF

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

Application (II): Porting CESM and Redesigning CAM-SE for Sunway TaihuLight


CAM5.0

POP2.0

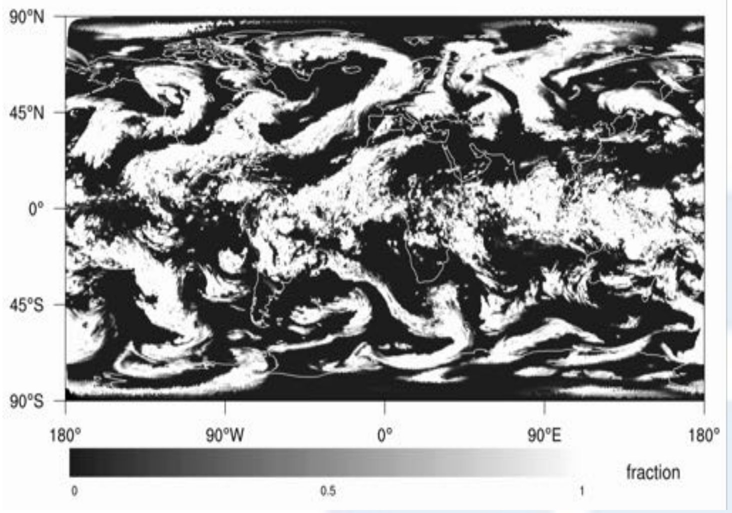
CPL7

CLM4.0

CICE4.0

CESM1.2.0
Surf Temp (radiative) mean= 286.79 K
Sea surface temperature mean= 19.70 C
snow height m
ice area (aggregate) %

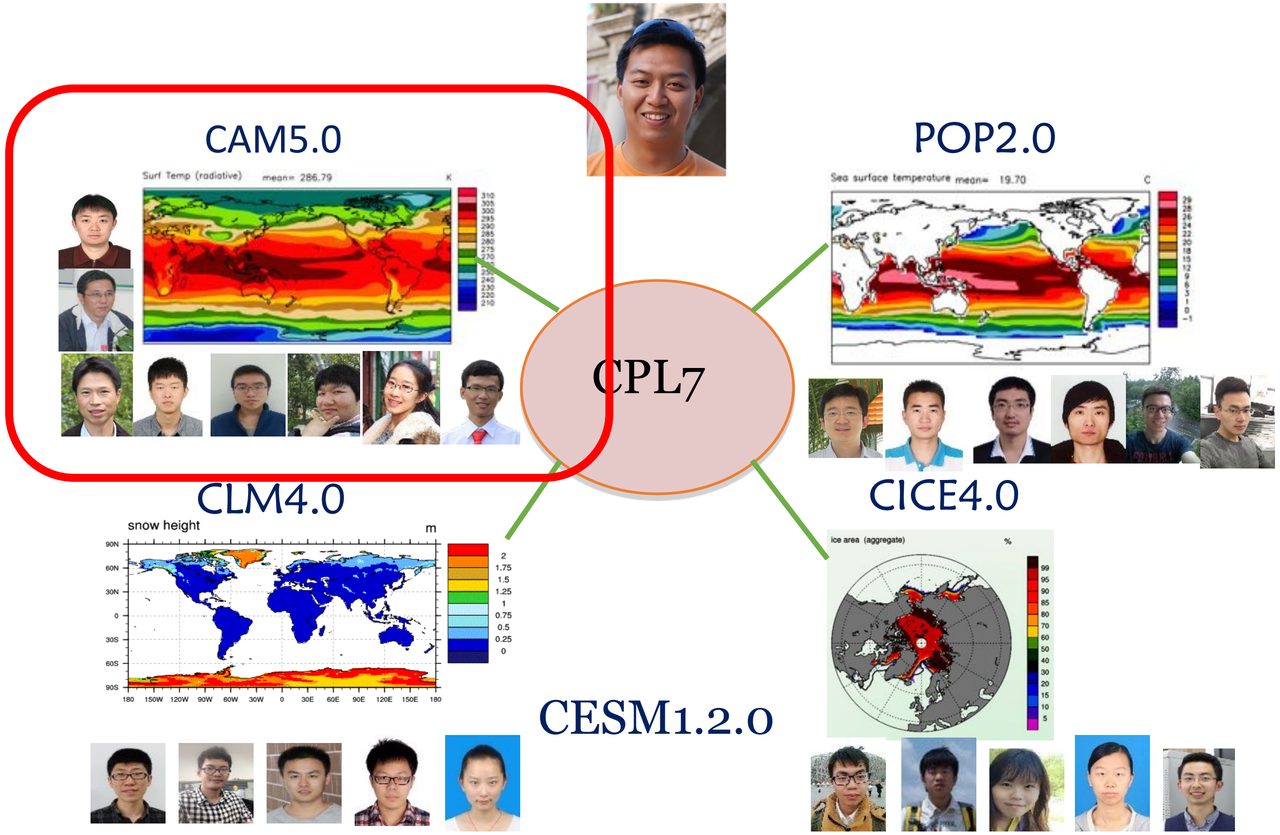
Tsinghua + BNU 30+ Professors and Students

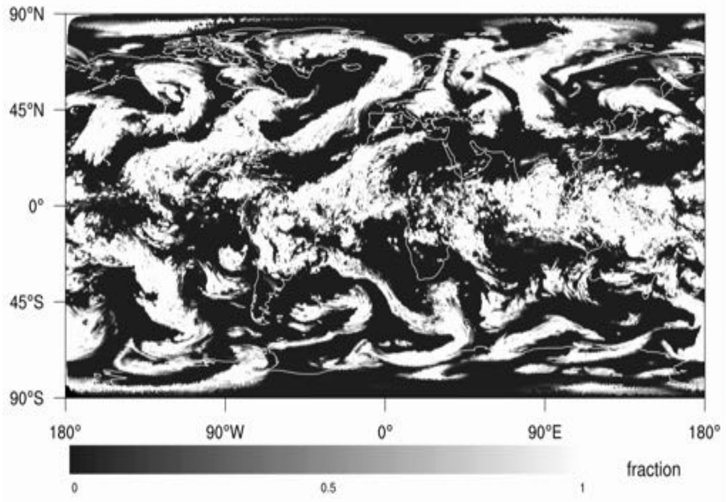


- Four component models, millions lines of code
- Large-scale run on Sunway TaihuLight
  - 24,000 MPI processes
  - Over one million cores
- 10-20x speedup for kernels
- 2-3x speedup for the entire model

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

Application (II): Porting CESM and Redesigning CAM-SE for Sunway TaihuLight



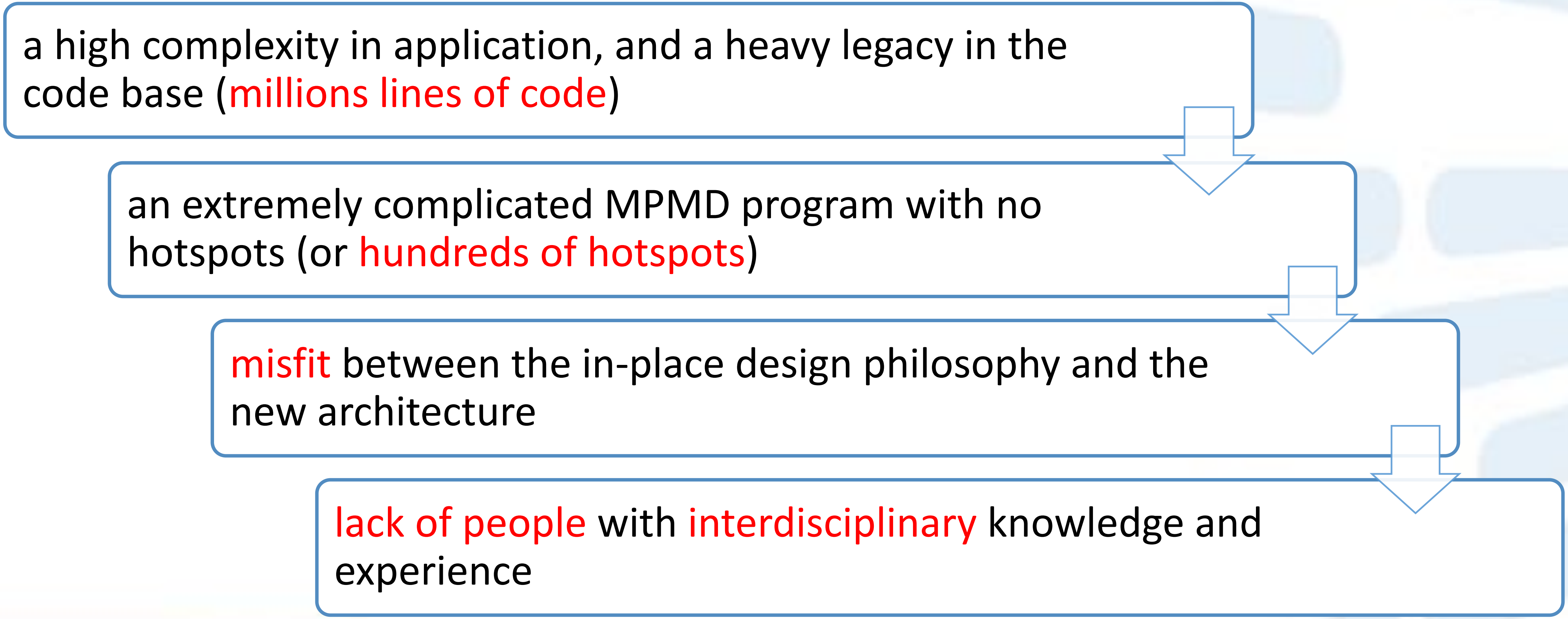


Tsinghua + BNU 30+ Professors and Students

- 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



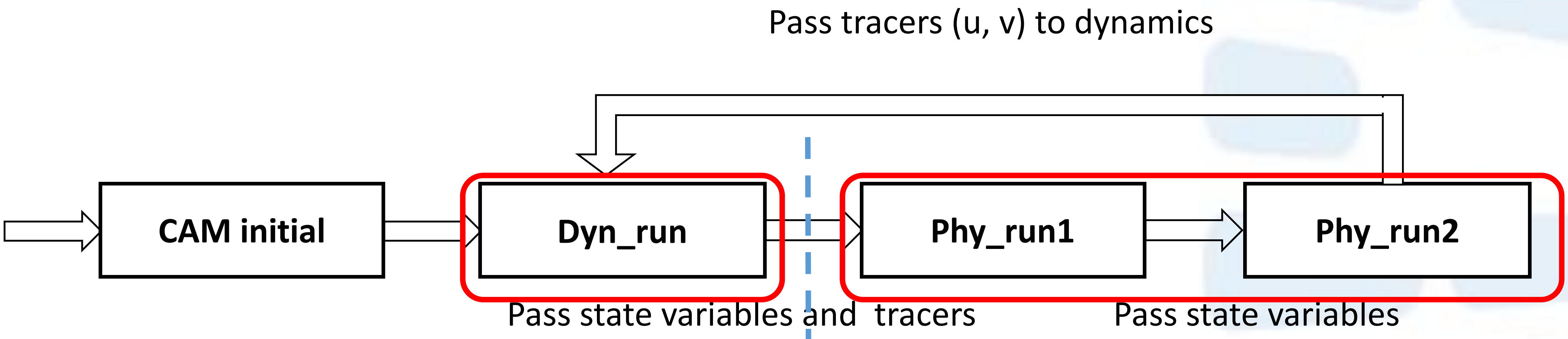
a high complexity in application, and a heavy legacy in the code base (millions lines of code)

an extremely complicated MPMD program with no hotspots (or hundreds of hotspots)

misfit between the in-place design philosophy and the new architecture

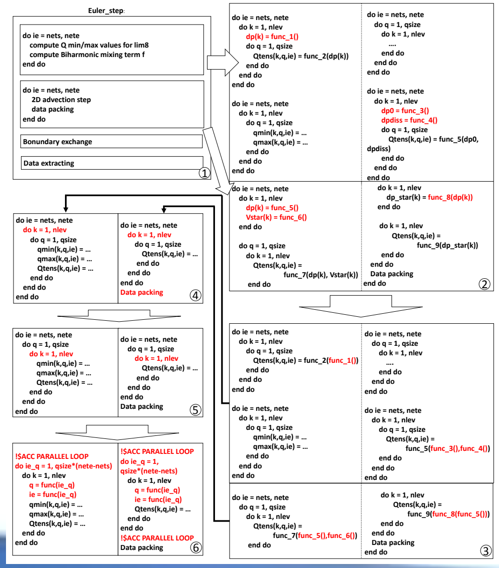
lack of people with interdisciplinary knowledge and experience

OpenACC-based Refactoring of CAM



Pass tracers (u, v) to dynamics
CAM initial
Dyn_run
Phy_run1
Phy_run2
Pass state variables and tracers
Pass state variables

- manual transformation of loops
- manual OpenACC parallelization and optimization on code and data structures



Euler_step:

```
do ie = nets, nete
  compute Q min/max values for lim8
  compute Biharmonic mixing term f
end do

do ie = nets, nete
   2D advection step
   data packing
end do

Bonundary exchange

Data extracting
```

①

```
do ie = nets, nete
  do k = 1, nlev
    dp(k) = func_1()
    do q = 1, qsize
      Qtens(k,q,ie) = func_2(dp(k))
    end do
  end do
end do

do ie = nets, nete
  do k = 1, nlev
    do q = 1, qsize
      qmin(k,q,ie) = …
      qmax(k,q,ie) = …
    end do
  end do
end do

do ie = nets, nete
  do k = 1, nlev
    dp(k) = func_5()
    Vstar(k) = func_6()
  end do

  do q = 1, qsize
    do k = 1, nlev
      Qtens(k,q,ie) =
                func_7(dp(k), Vstar(k))
     end do
```

```
do ie = nets, nete
  do q = 1, qsize
    do k = 1, nlev
      ....
    end do
  end do
end do

do ie = nets, nete
  do k = 1, nlev
    dp0 = func_3()
    dpdiss = func_4()
    do q = 1, qsize
      Qtens(k,q,ie) = func_5(dp0,
dpdiss)
    end do
  end do
end do
   do k = 1, nlev
     dp_star(k) = func_8(dp(k))
   end do

   do k = 1, nlev
       Qtens(k,q,ie) =
                 func_9(dp_star(k))
     end do
  end do
  Data packing
end do
```

②

```
do ie = nets, nete
  do k = 1, nlev
    do q = 1, qsize
      Qtens(k,q,ie) = func_2(func_1())
    end do
  end do
end do

do ie = nets, nete
  do k = 1, nlev
    do q = 1, qsize
      qmin(k,q,ie) = …
      qmax(k,q,ie) = …
    end do
  end do
end do

do ie = nets, nete
  do q = 1, qsize
    do k = 1, nlev
      Qtens(k,q,ie) =
              func_7(func_5(),func_6())
     end do
```

```
do ie = nets, nete
  do q = 1, qsize
    do k = 1, nlev
      ....
    end do
  end do
end do

do ie = nets, nete
  do k = 1, nlev
    do q = 1, qsize
      Qtens(k,q,ie) =
              func_5(func_3(),func_4())
    end do
  end do
end do
    do k = 1, nlev
       Qtens(k,q,ie) =
            func_9(func_8(func_5()))
    end do
  end do
  Data packing
end do
```

③

```
do ie = nets, nete
  do k = 1, nlev
    do q = 1, qsize
      qmin(k,q,ie) = …
      qmax(k,q,ie) = …
      Qtens(k,q,ie) = …
    end do
  end do
end do
```

```
do ie = nets, nete
  do k = 1, nlev
    do q = 1, qsize
      Qtens(k,q,ie) = …
    end do
  end do
end do
Data packing
```

④

```
do ie = nets, nete
  do q = 1, qsize
    do k = 1, nlev
      qmin(k,q,ie) = …
      qmax(k,q,ie) = …
      Qtens(k,q,ie) = …
    end do
  end do
end do
```

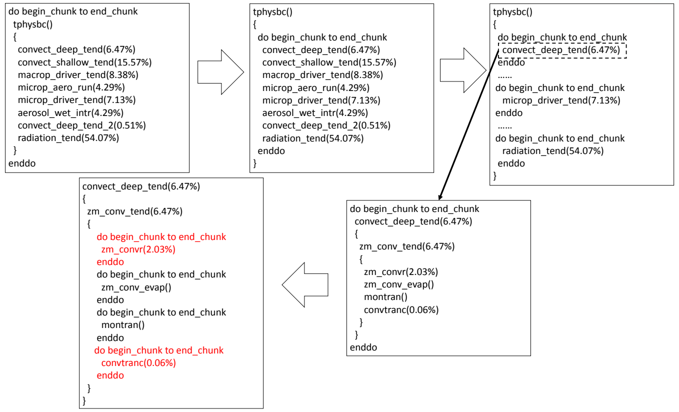
```
do ie = nets, nete
  do q = 1, qsize
    do k = 1, nlev
      Qtens(k,q,ie) = …
    end do
  end do
end do
Data packing
```

⑤

```
!$ACC PARALLEL LOOP
do ie_q = 1, qsize*(nete-nets)
  do k = 1, nlev
    q = func(ie_q)
    ie = func(ie_q)
    qmin(k,q,ie) = …
    qmax(k,q,ie) = …
    Qtens(k,q,ie) = …
  end do
end do
```

```
!$ACC PARALLEL LOOP
do ie_q = 1,
qsize*(nete-nets)
  do k = 1, nlev
    q = func(ie_q)
    ie = func(ie_q)
    Qtens(k,q,ie) = …
  end do
end do
!$ACC PARALLEL LOOP
Data packing
```

⑥



```
do begin_chunk to end_chunk
 tphysbc()
 {
  convect_deep_tend(6.47%)
  convect_shallow_tend(15.57%)
  macrop_driver_tend(8.38%)
  microp_aero_run(4.29%)
  microp_driver_tend(7.13%)
  aerosol_wet_intr(4.29%)
  convect_deep_tend_2(0.51%)
  radiation_tend(54.07%)
 }
enddo
```

```
tphysbc()
{
 do begin_chunk to end_chunk
  convect_deep_tend(6.47%)
  convect_shallow_tend(15.57%)
  macrop_driver_tend(8.38%)
  microp_aero_run(4.29%)
  microp_driver_tend(7.13%)
  aerosol_wet_intr(4.29%)
  convect_deep_tend_2(0.51%)
  radiation_tend(54.07%)
 enddo
}
```

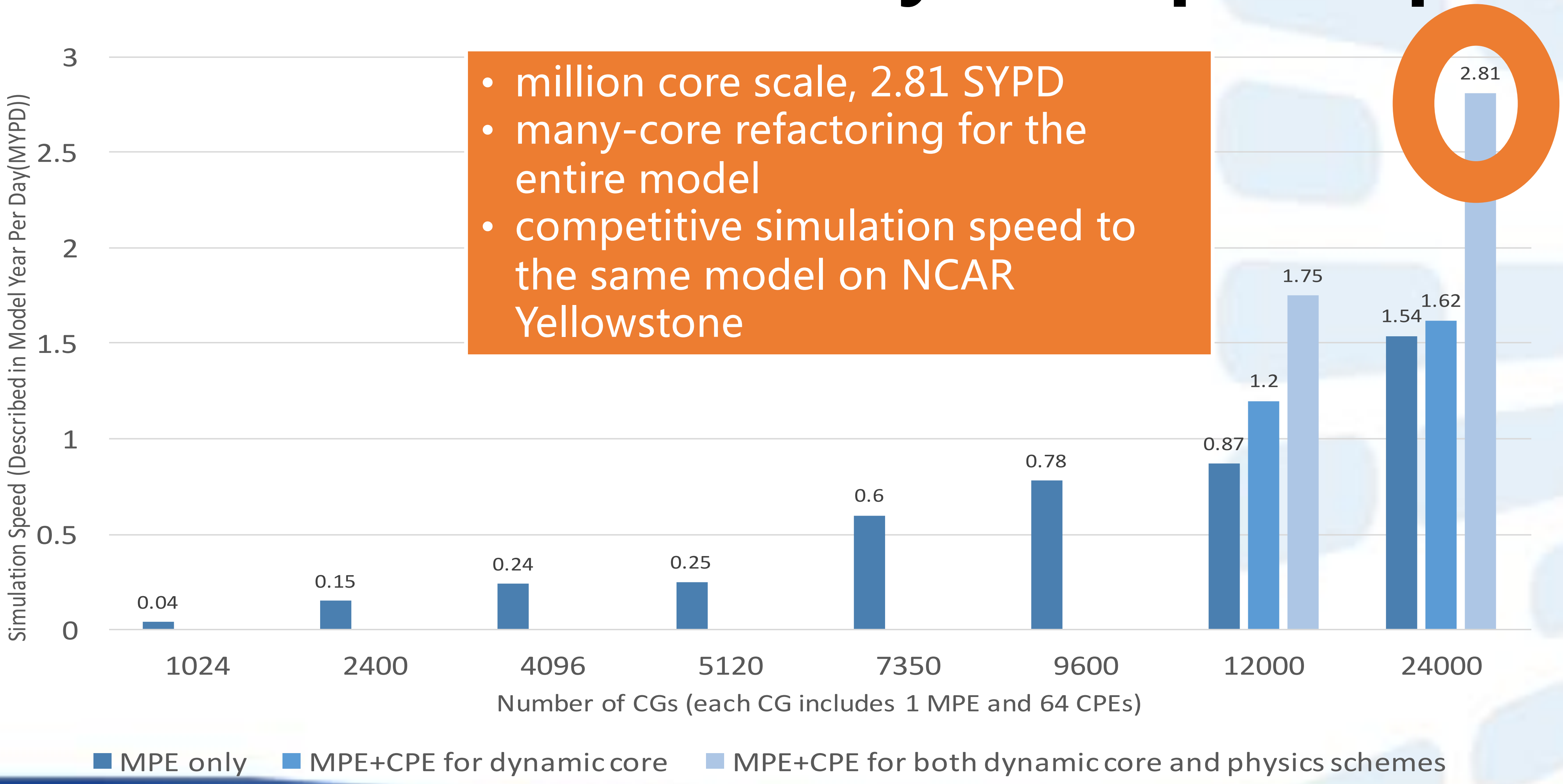
```
tphysbc()
{
 do begin_chunk to end_chunk
  convect_deep_tend(6.47%)
 enddo
 ……
do begin_chunk to end_chunk
  microp_driver_tend(7.13%)
enddo
 ……
do begin_chunk to end_chunk
  radiation_tend(54.07%)
 enddo
}
```

```
do begin_chunk to end_chunk
 convect_deep_tend(6.47%)
 {
  zm_conv_tend(6.47%)
  {
   zm_convr(2.03%)
   zm_conv_evap()
   montran()
   convtranc(0.06%)
  }
 }
enddo
```

```
convect_deep_tend(6.47%)
{
 zm_conv_tend(6.47%)
 {
   do begin_chunk to end_chunk
    zm_convr(2.03%)
   enddo
   do begin_chunk to end_chunk
    zm_conv_evap()
   enddo
   do begin_chunk to end_chunk
    montran()
   enddo
  do begin_chunk to end_chunk
    convtranc(0.06%)
   enddo
 }
}
```

- tool based transformation of loops

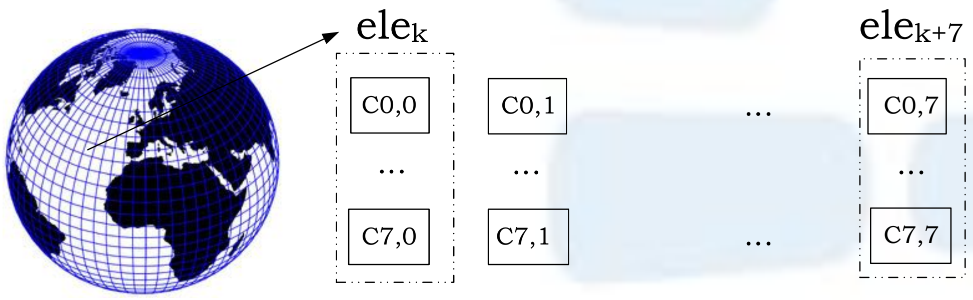
CAM model: scalability and speedup



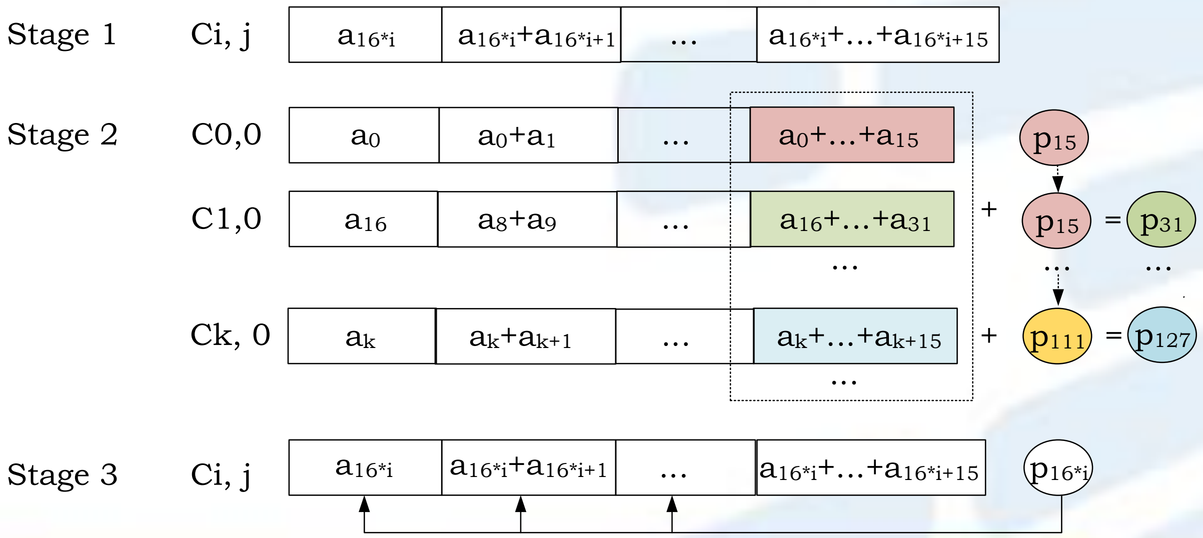
- million core scale, 2.81 SYPD
- many-core refactoring for the entire model
- competitive simulation speed to the same model on NCAR Yellowstone

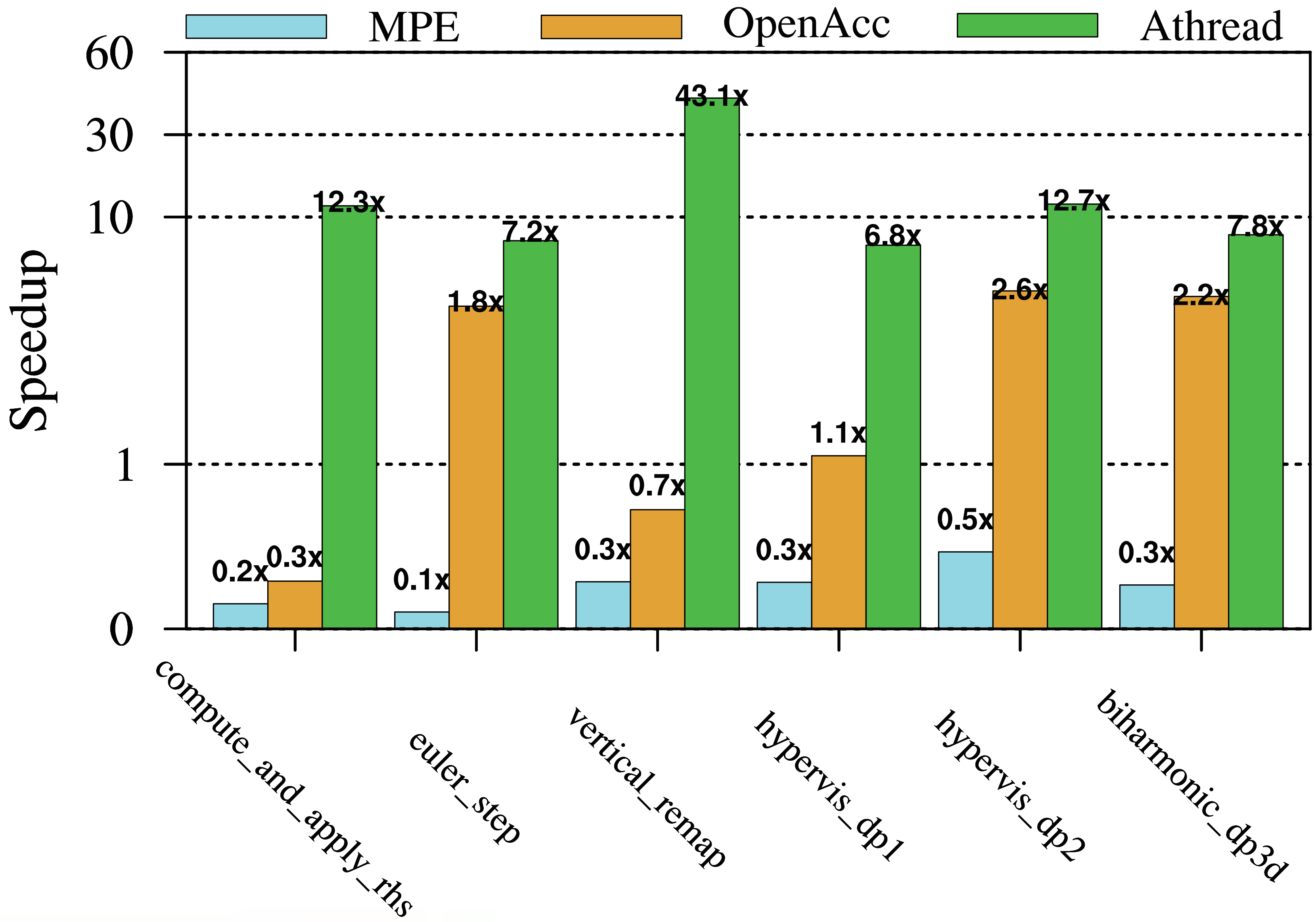
Athread-based Fine-grained Redesign

- Step 1: rewrite of Fortran OpenACC code to Athread C code
  - finer memory control through a specific DMA scheme
  - more efficient vectorization



- Step 2: register-communication based redesign
  - remove data dependency
  - expose more parallelism





1 Sunway CG (64 CPEs) could be equivalent to

0.1x Intel Core

or

1.8x Intel Core

or

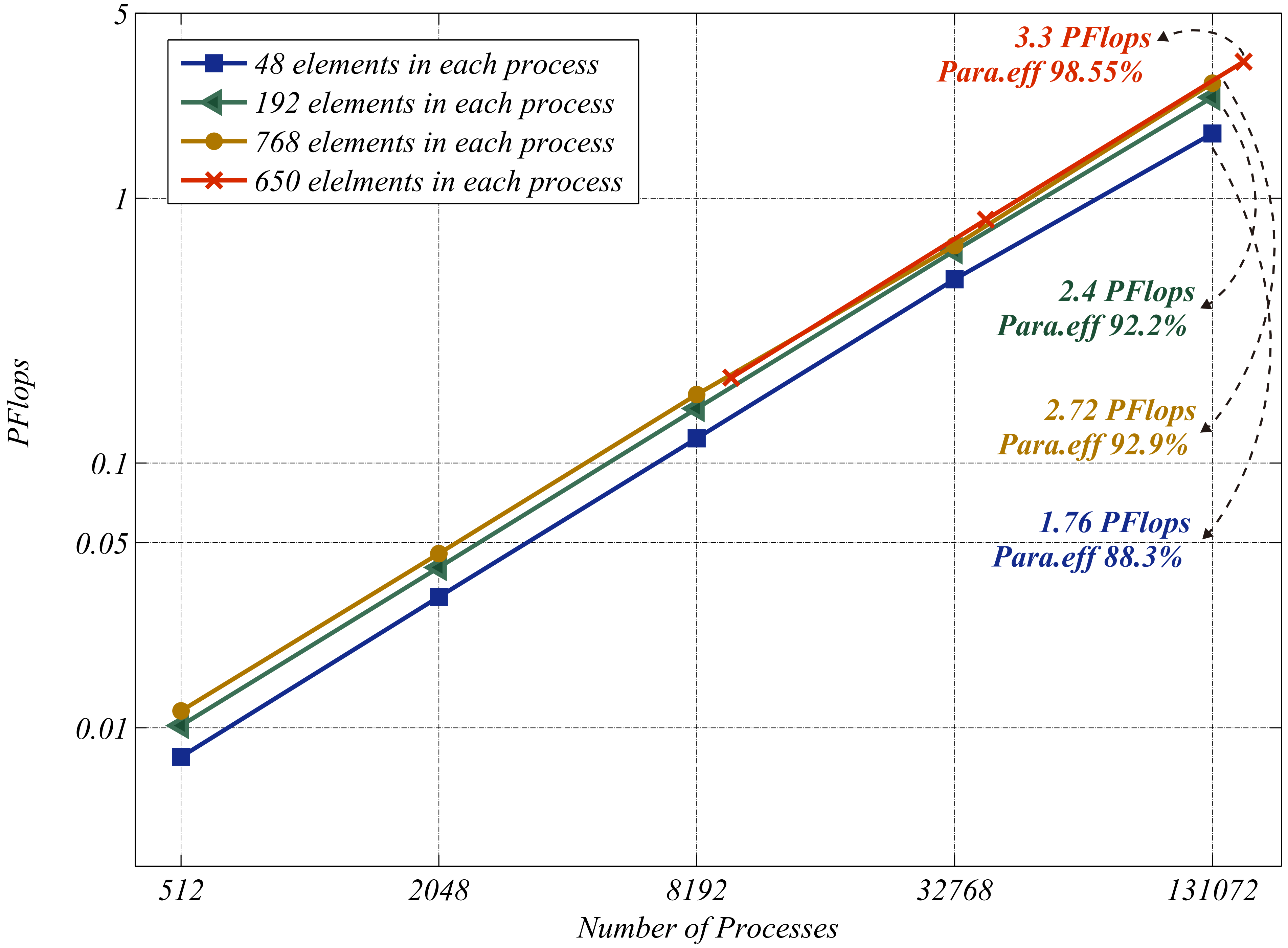
7.2x Intel Core

or in certain cases

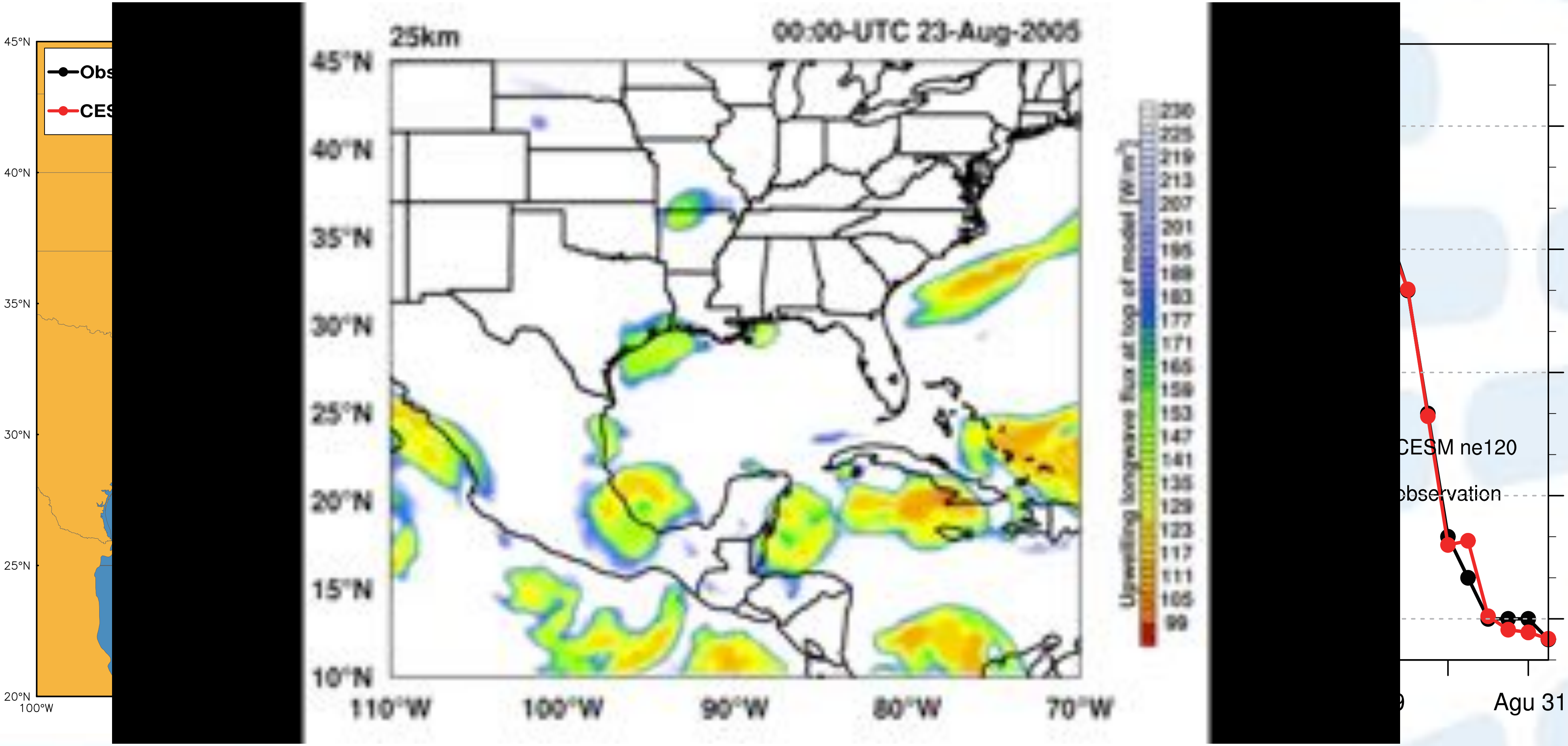
43.1x Intel Core

Performance Improvement through Redesign

Scaling the Dynamic Core to Millions of Cores

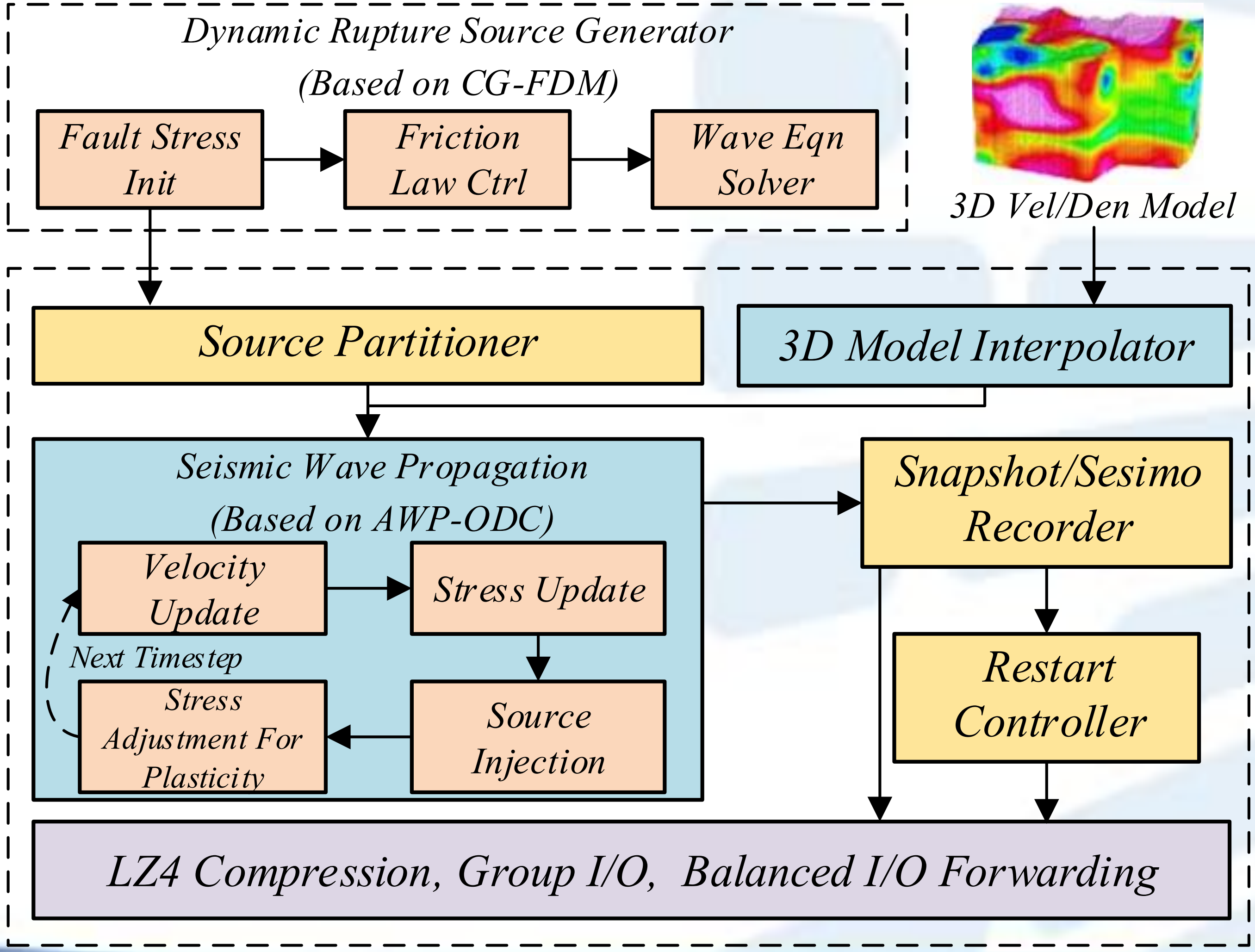


Simulation of Hurricane Katrina



Application (III): Nonlinear Earthquake Simulation on Sunway TaihuLight

- Dynamic rupture source generator (originated from CG-FDM)
- Seismic wave propagation (originated from AWP-ODC)
- Other utilities:
  - source partitioner
  - 3D Model Interpolator
  - Restart controller



Dynamic Rupture Source Generator (Based on CG-FDM)

Fault Stress Init

Friction Law Ctrl

Wave Eqn Solver

3D Vel/Den Model

Source Partitioner

3D Model Interpolator

Seismic Wave Propagation (Based on AWP-ODC)

Velocity Update

Stress Update

Next Timestep

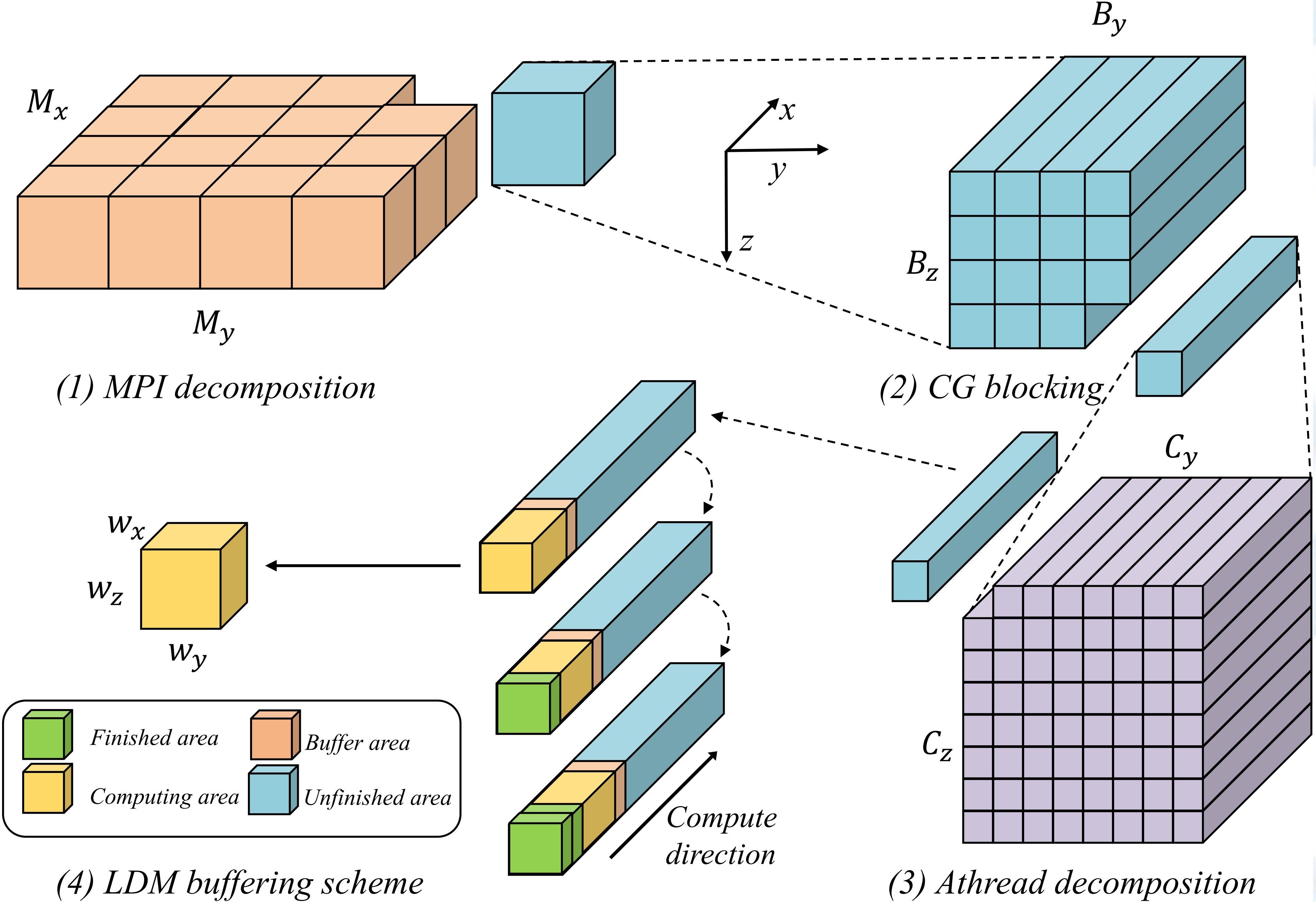
Stress Adjustment For Plasticity

Source Injection

Snapshot/Sesimo Recorder

Restart Controller

LZ4 Compression, Group I/O, Balanced I/O Forwarding



(1) MPI decomposition

(2) CG blocking

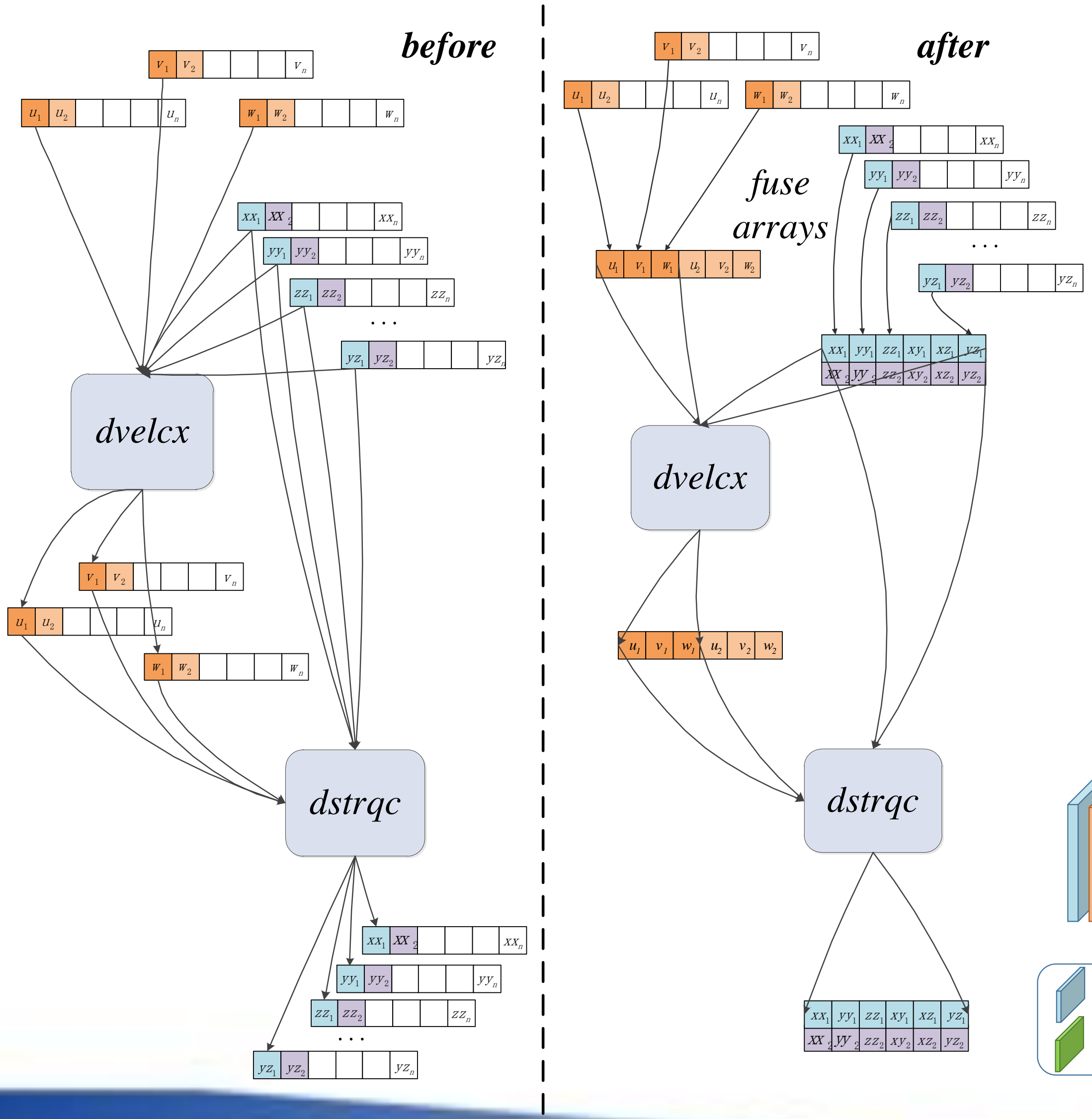
(4) LDM buffering scheme

(3) Athread decomposition

国家超级计算无锡中心

Multi-Level Domain Decomposition

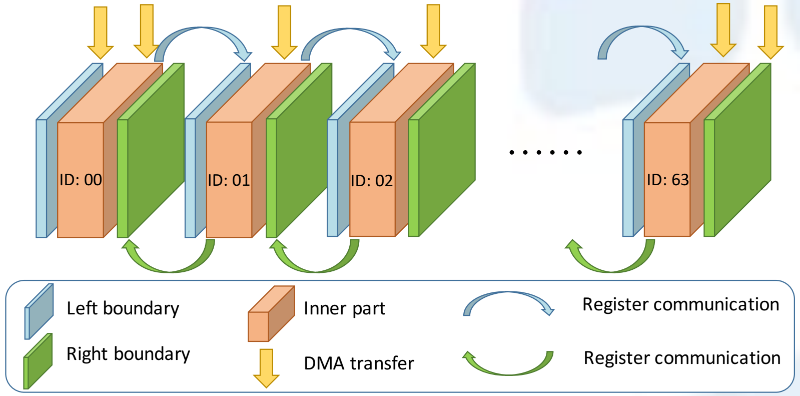
A Balanced Memory Scheme



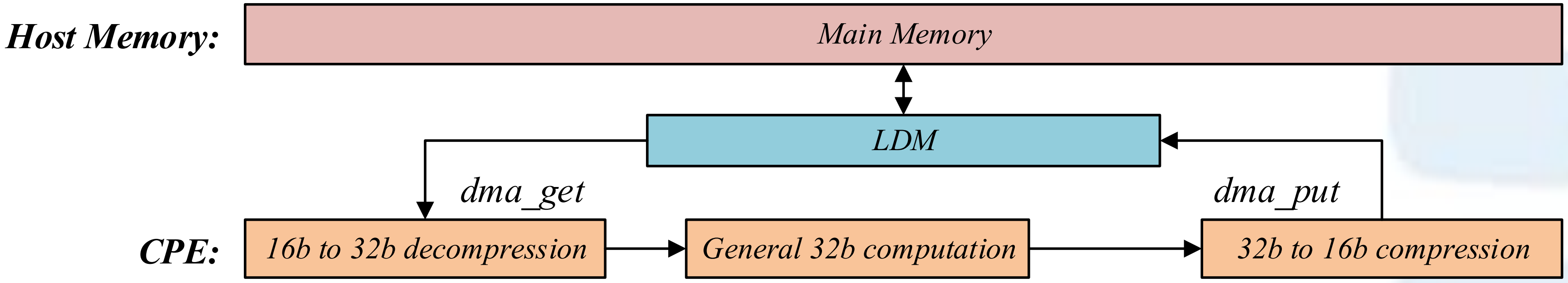
(1) array fusion,

(2) halo exchange through register communication,

(3) and optimized blocking configuration guided by an analytical model



On-the-fly Compression



Host Memory:

Main Memory

LDM

dma_get

dma_put

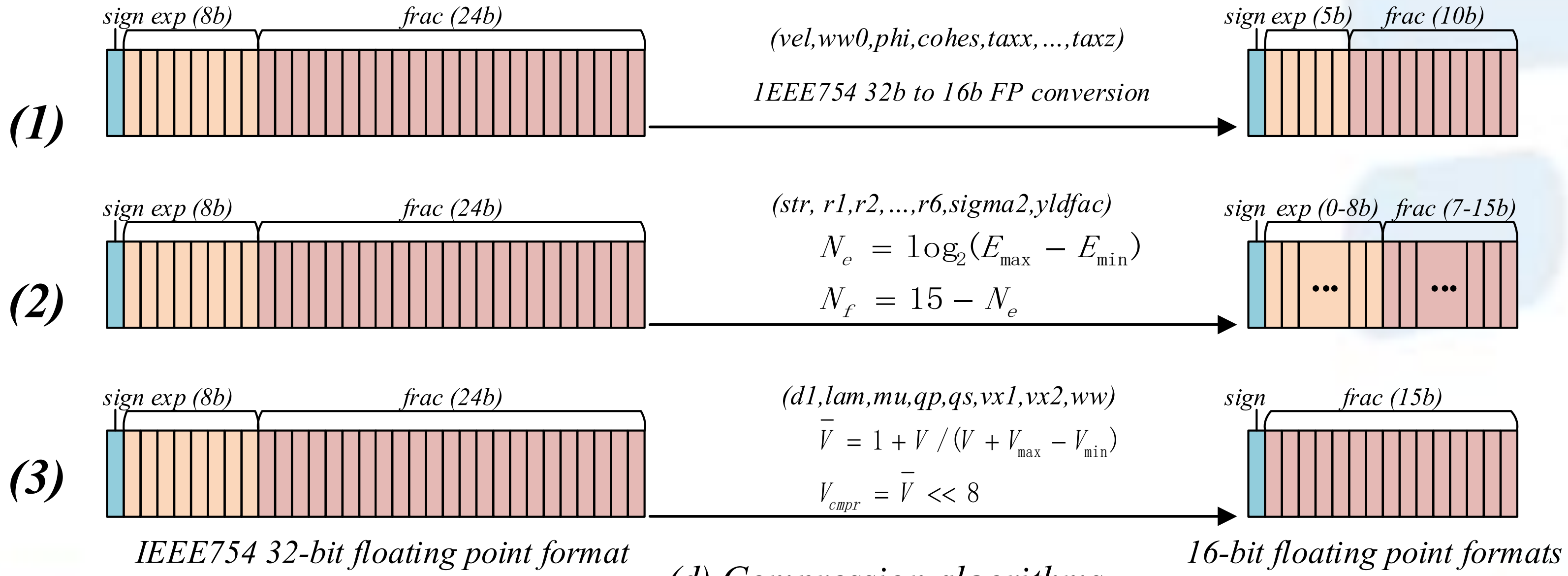
CPE:

16b to 32b decompression

General 32b computation

32b to 16b compression

(b) Computation workflow



(1)

sign exp (8b) frac (24b)

(vel,ww0,phi,cohes,taxx,…,taxz)

IEEE754 32b to 16b FP conversion

sign exp (5b) frac (10b)

(2)

sign exp (8b) frac (24b)

(str, r1,r2,…,r6,sigma2,yldfac)

$$N_e = \log_2(E_{\max} - E_{\min})$$

$$N_f = 15 - N_e$$

sign exp (0-8b) frac (7-15b)

(3)

sign exp (8b) frac (24b)

(d1,lam,mu,qp,qs,vx1,vx2,ww)

$$\overline{V} = 1 + V / (V + V_{\max} - V_{\min})$$

$$V_{cmpr} = \overline{V} << 8$$

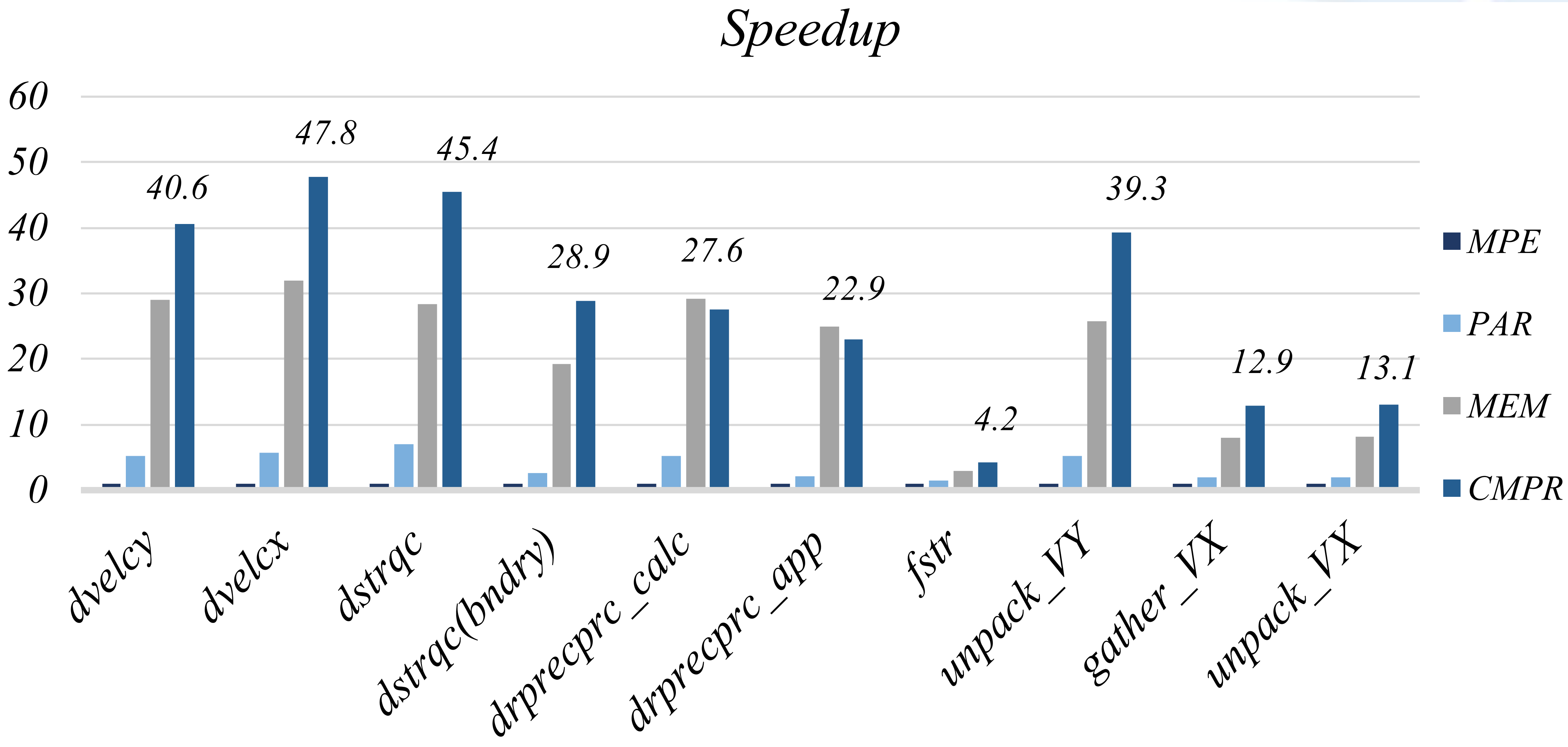
sign frac (15b)

IEEE754 32-bit floating point format

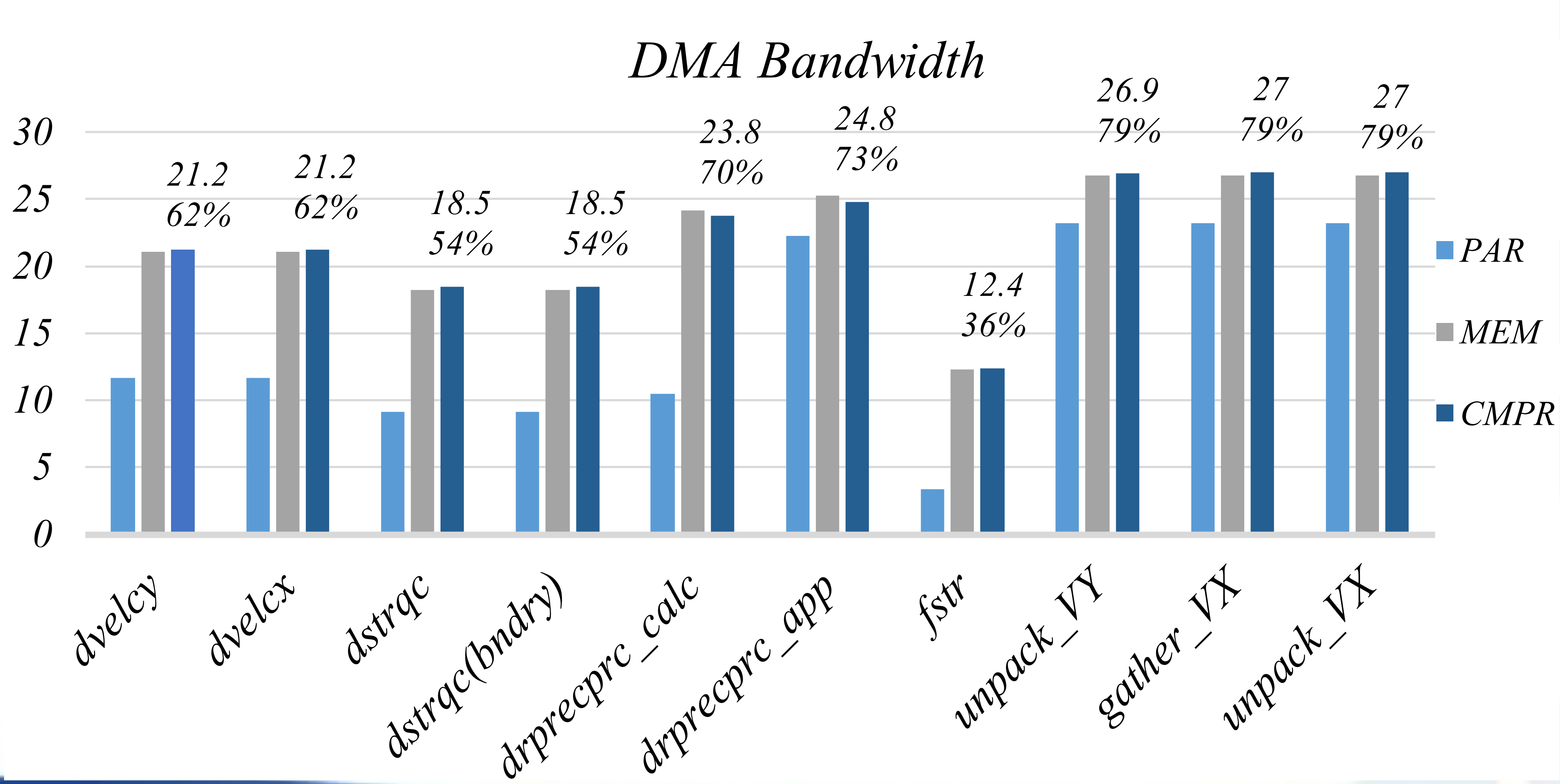
16-bit floating point formats

(d) Compression algorithms

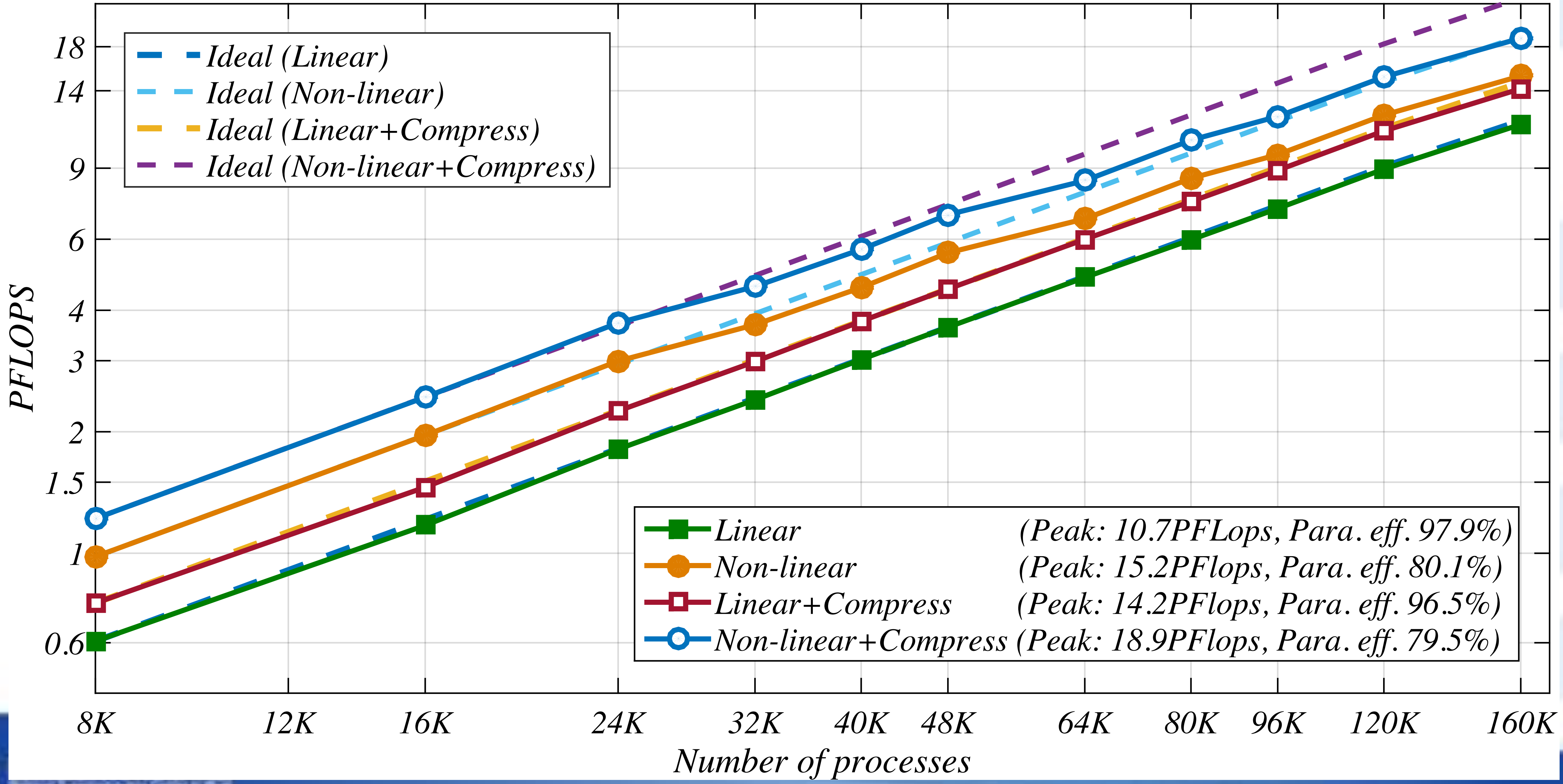
Speedup: 64 CPE vs 1 MPE



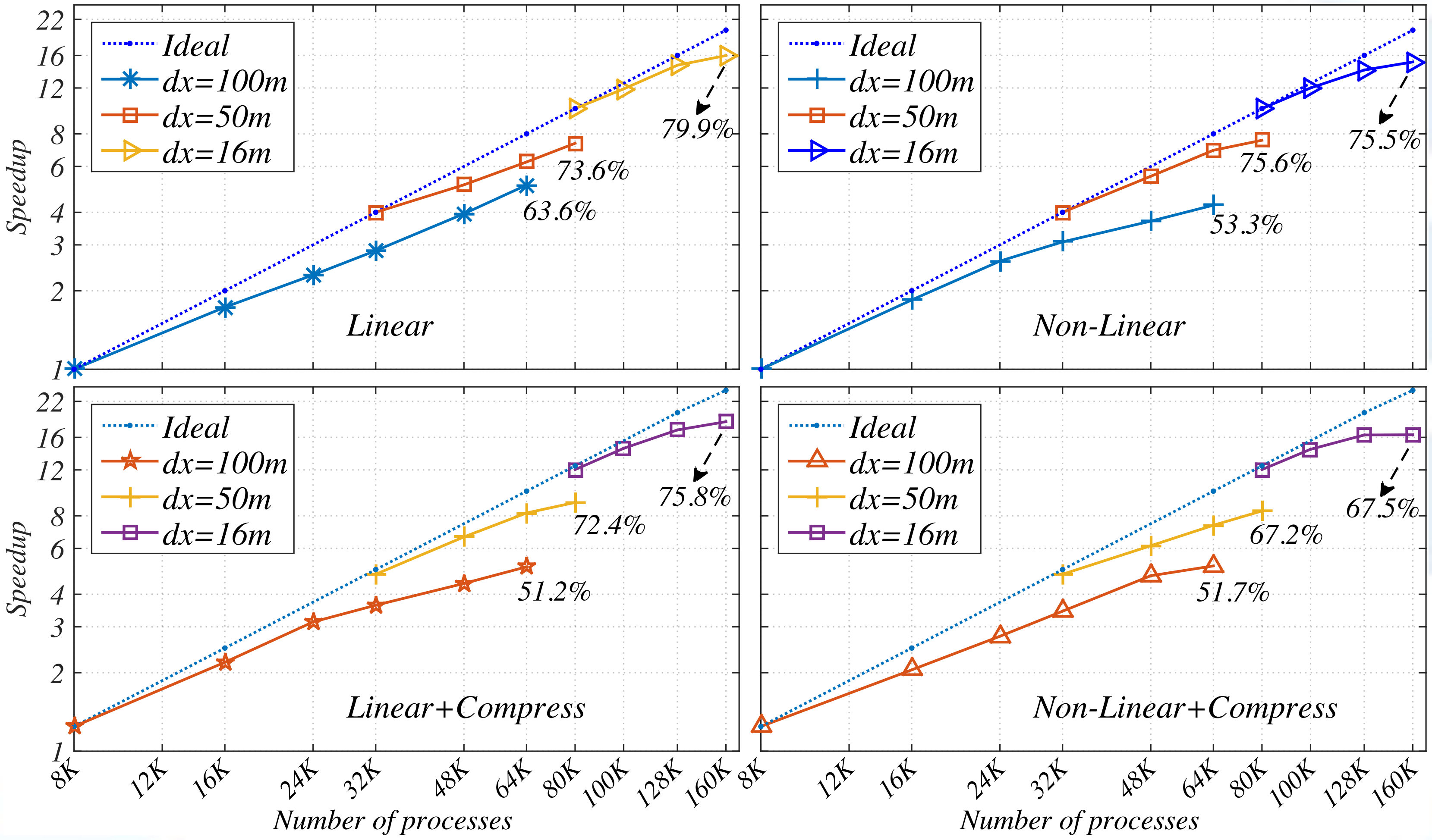
Memory Bandwidth Utilization



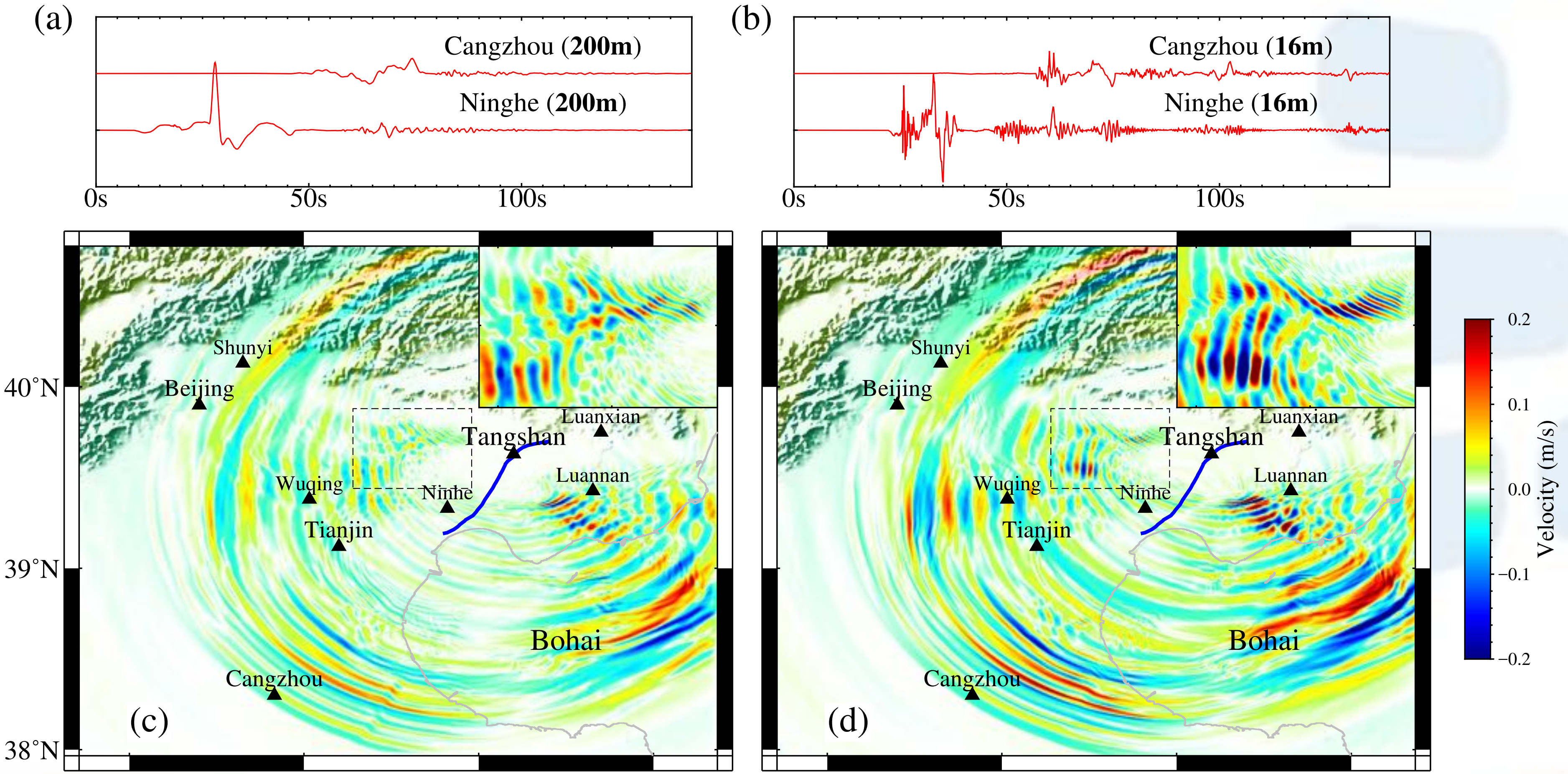
Weak Scaling



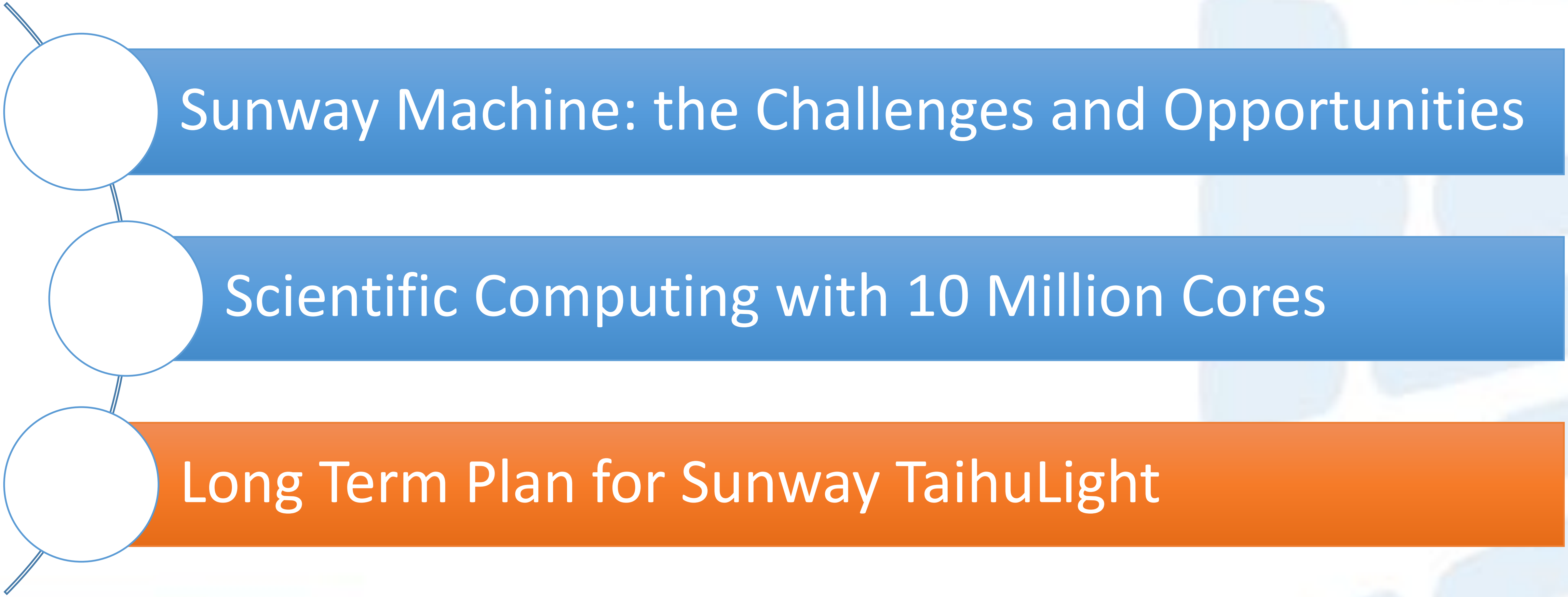
Strong Scaling



Simulation Results: 200m vs 16m

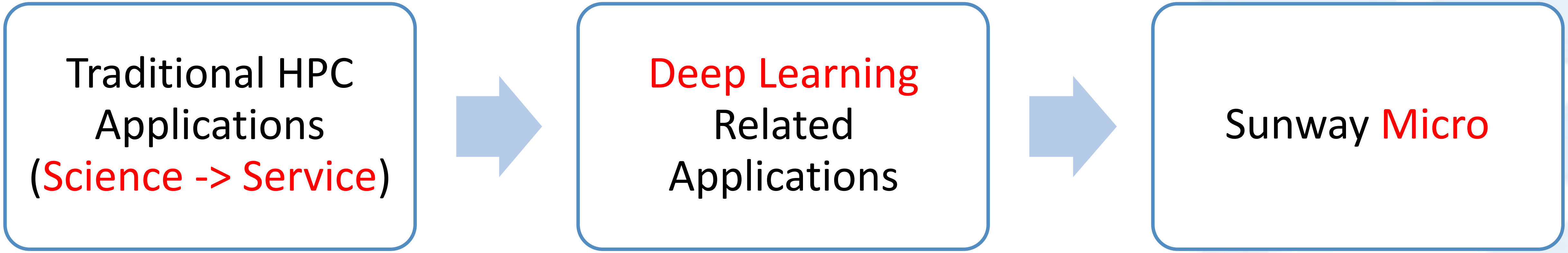


Outline



- Sunway Machine: the Challenges and Opportunities
- Scientific Computing with 10 Million Cores
- Long Term Plan for Sunway TaihuLight

Long Term Plan



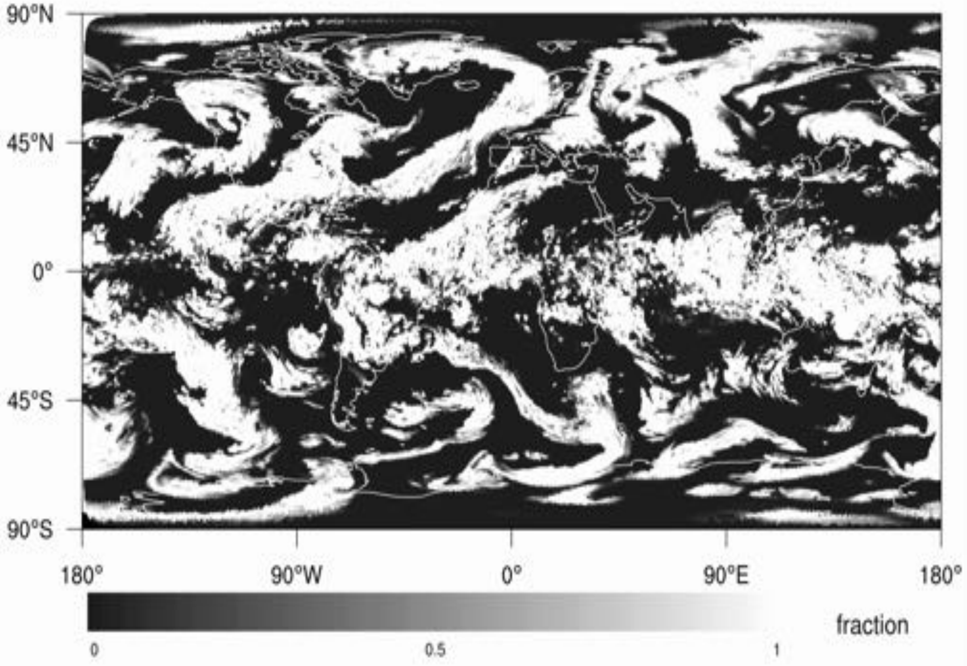
Traditional HPC Applications (Science -> Service)

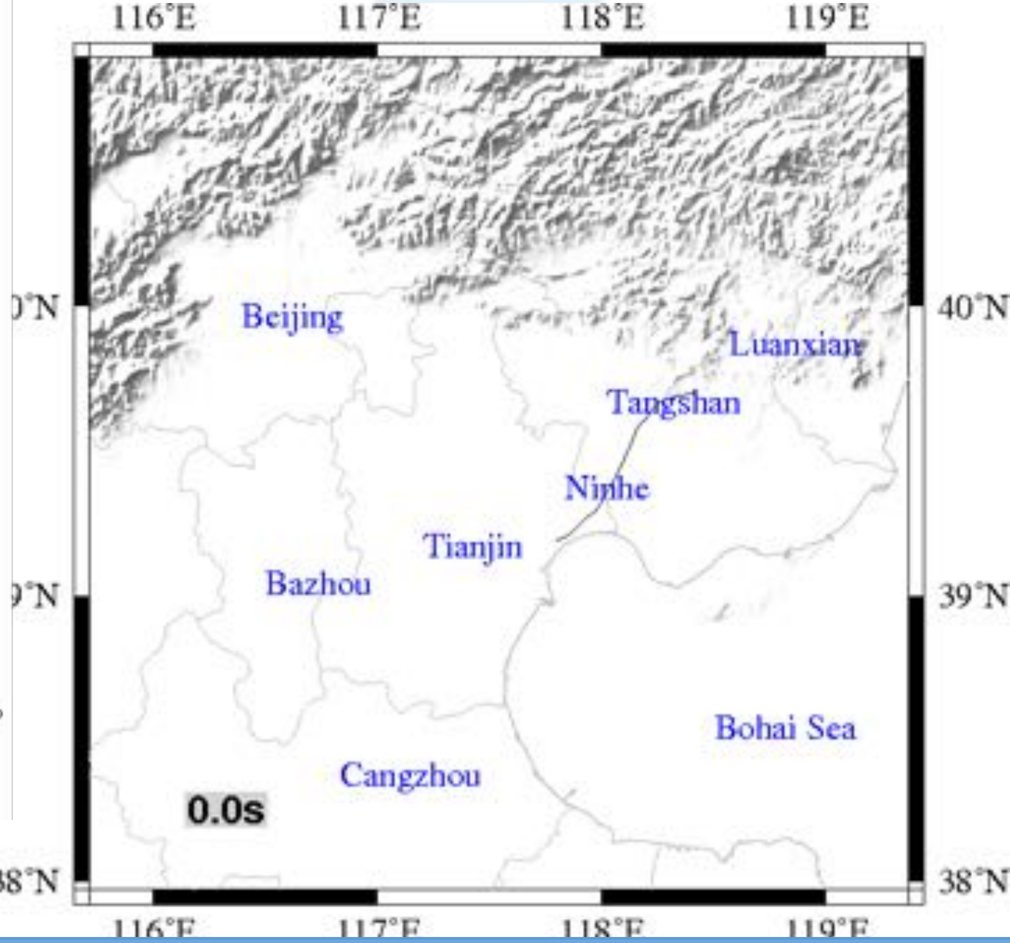
Deep Learning Related Applications

Sunway Micro

Long Term Plan

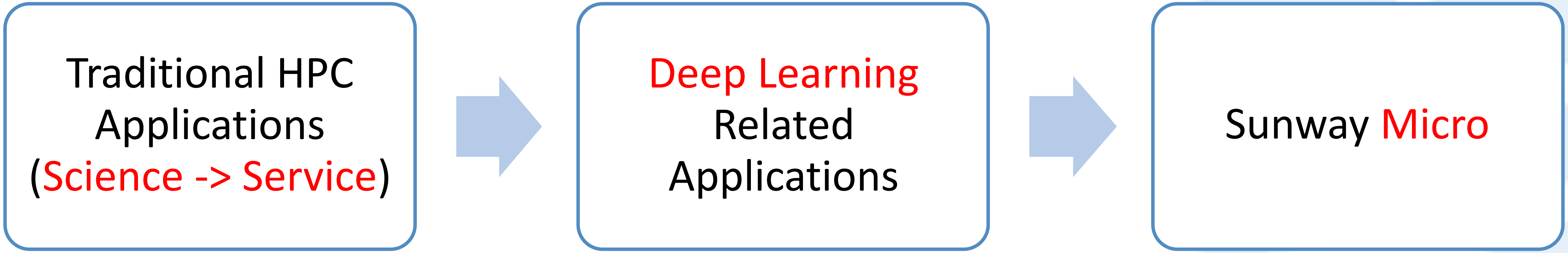
Traditional HPC Applications (Science -> Service)





"15-Pflops Nonlinear Earthquake Simulation on Sunway TaihuLight: Enabling Depiction of Realistic 10 Hz Scenarios", Gordon Bell Prize Finalist, SC 2017.

Long Term Plan



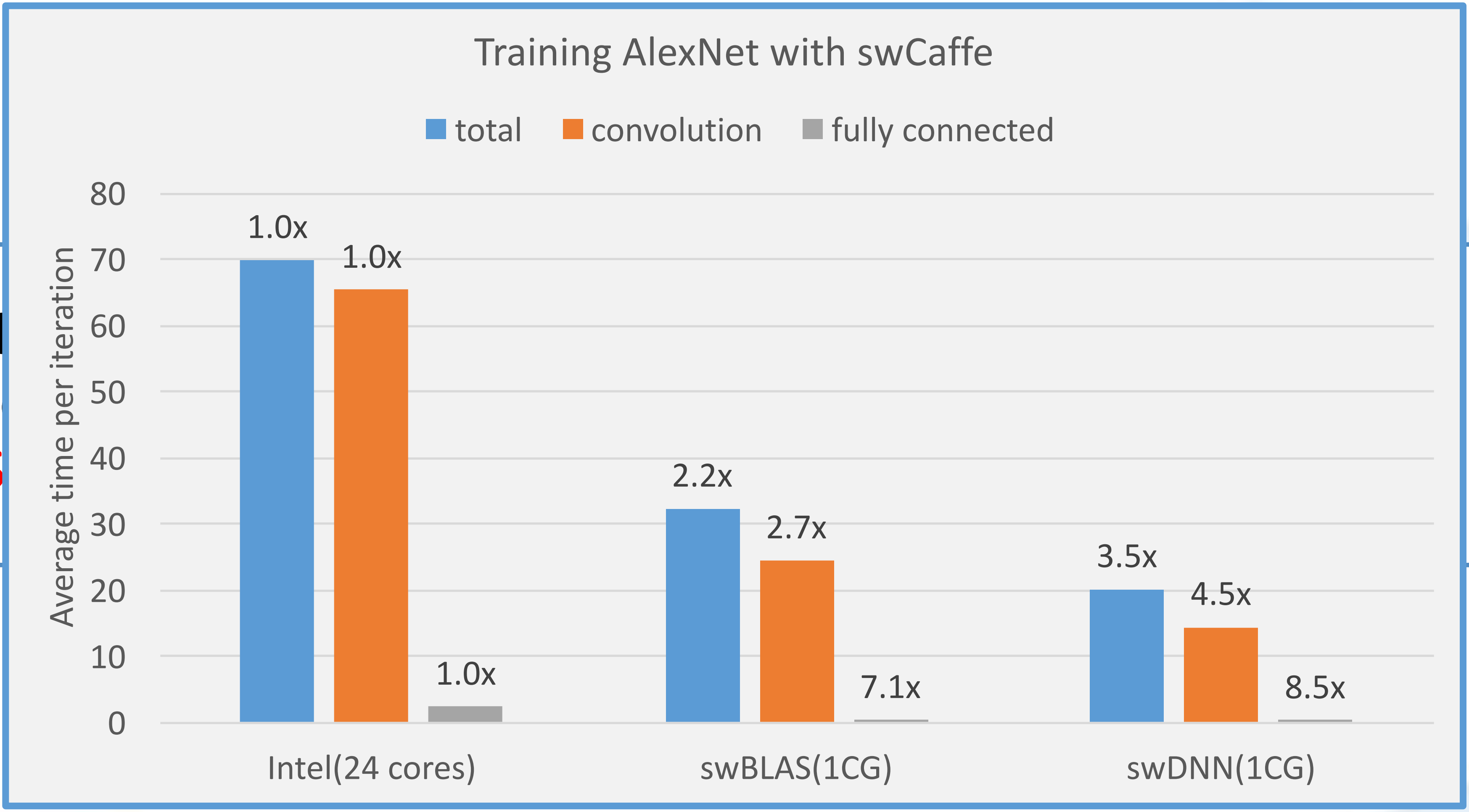
Traditional HPC Applications (Science -> Service)

Deep Learning Related Applications

Sunway Micro

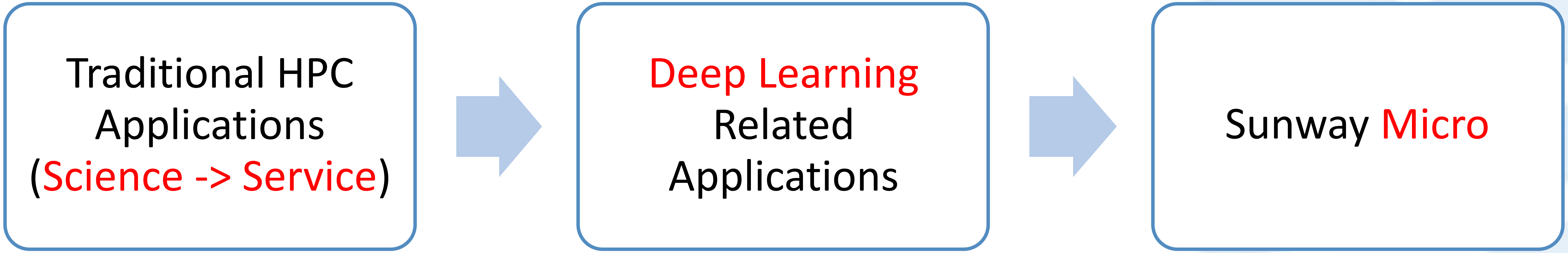
Long Term Plan

Traditional
Applicati
(Science -> S



Micro

Long Term Plan



Traditional HPC Applications (Science -> Service)

Deep Learning Related Applications

Sunway Micro

Acknowledgements

- MOST, China: major sponsors of the HPC hardware and software development

- NRCPC: vendor of the machine

- NCAR: Rich Loft, John Dennis, Allison Baker, Haiying Xu (support and advice on the CAM-SE work)

- SCEC: Yifeng Cui, Steve Day, Daniel Roten, Kim Olsen, Josh Tobin, Alex Breuer, and Dawei Mu (discussion and advice on the earthquake simulation work)

THANK YOU