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

神威 SUNWAY

Enabling Nonlinear Earthquake Simulation for 18-Hz and 8-Meter Scenarios

Haohuan Fu[1,2], Conghui He[1,2], Bingwei Chen[1,2], Zekun Yin[3], Zhenguo Zhang[4], Wenqiang Zhang[5], Tingjian Zhang[3], Wei Xue[1,2], Wanwang Yin[6], Guangwen Yang[1,2], Xiaofei Chen[4]

1. Tsinghua University 2. NSCC-Wuxi 3. Shandong University
4. SUSTC 5. USTC 6. NRCPC

March 23rd, 2018 @ Taipei

Acknowledgement

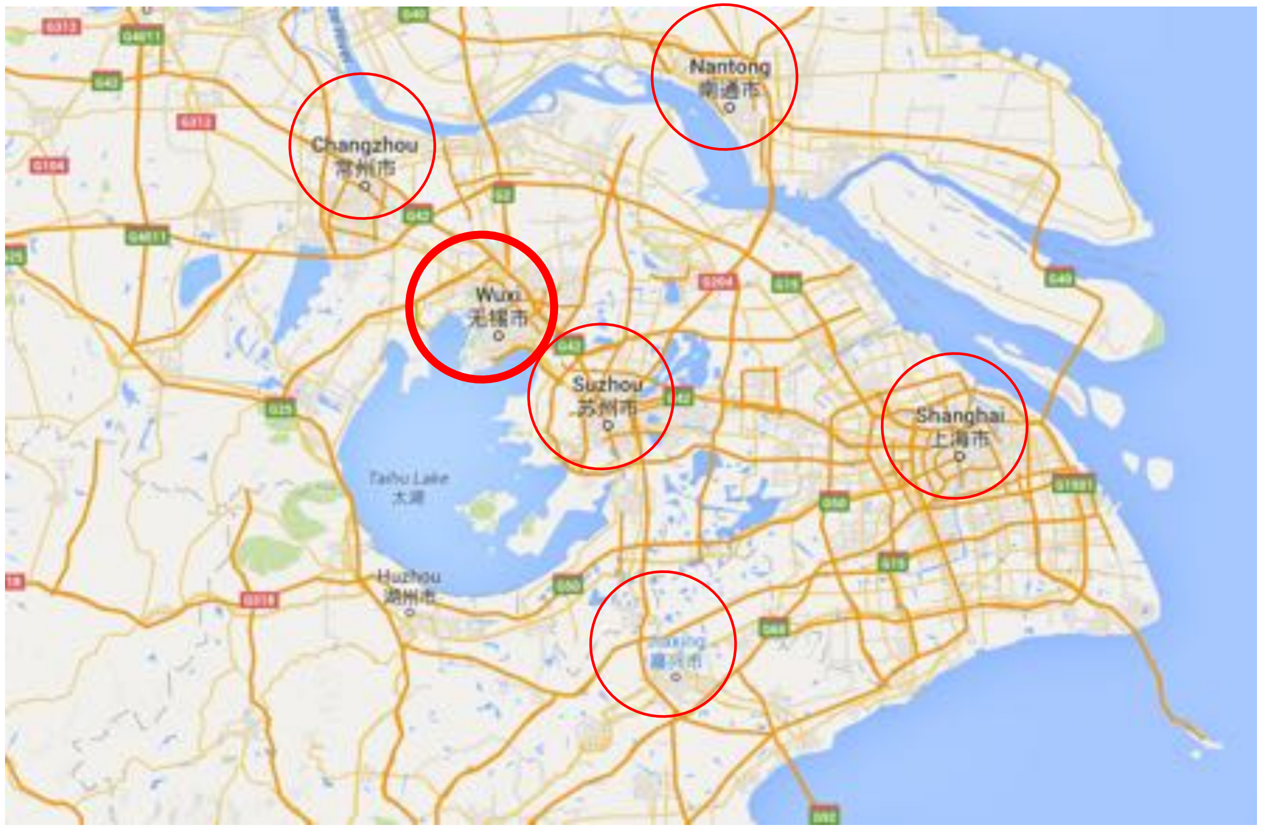
- Sponsor: MOST, Jiangsu Province, City of Wuxi
- Vendor: NRCPC (thousands of engineers working on the hardware, software, and integration of the system)
- Application and Library Developers

Sunway TaihuLight

神威

太湖之光

Sunway TaihuLight



City	Rank in Top100
Shanghai	1
Suzhou	7
Wuxi	14
Nantong	24
Changzhou	34
Jiaxing	50

The Design Process

Sunway-I:

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

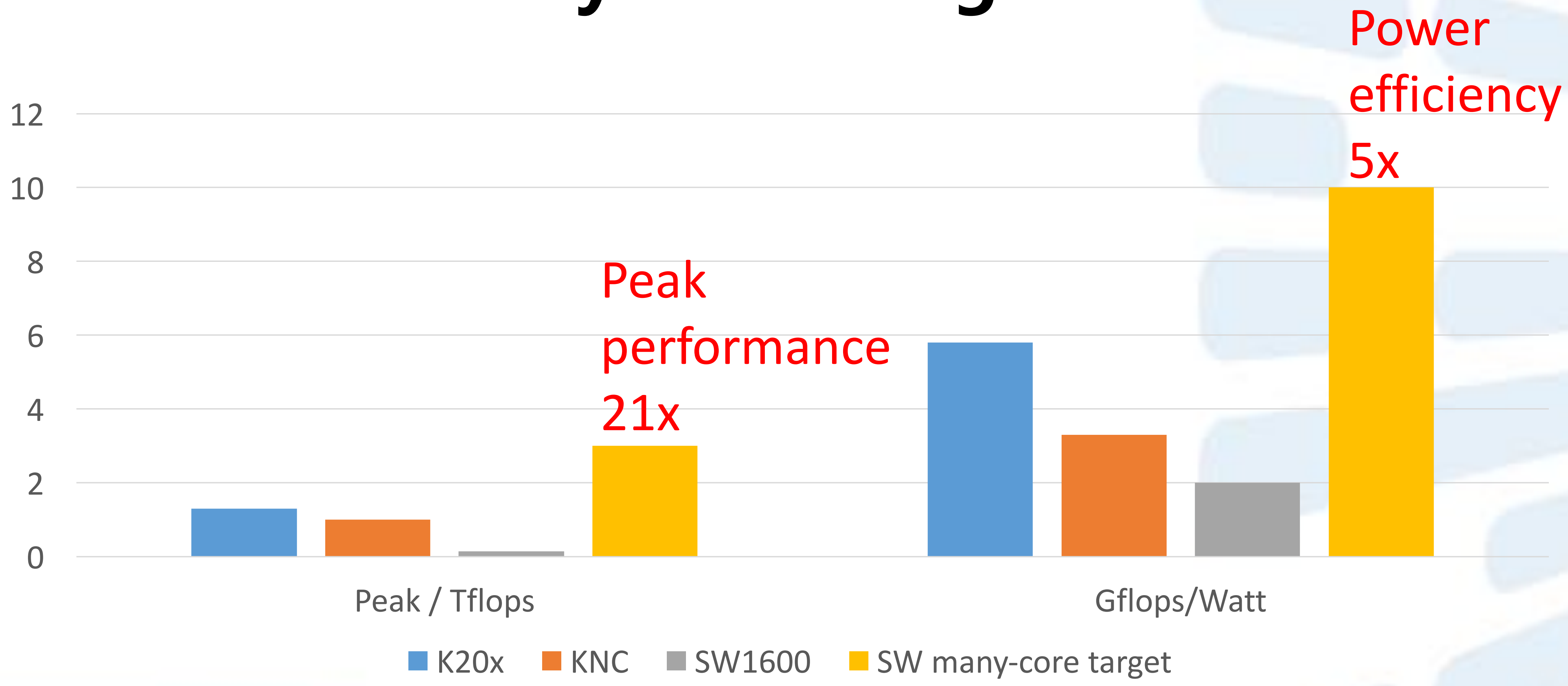
Sunway BlueLight:

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

Sunway TaihuLight:

- Peak > 100 Pflops
- homemade CPU

Sunway CPU Design Goal



CPU Design Strategy

Simple for more

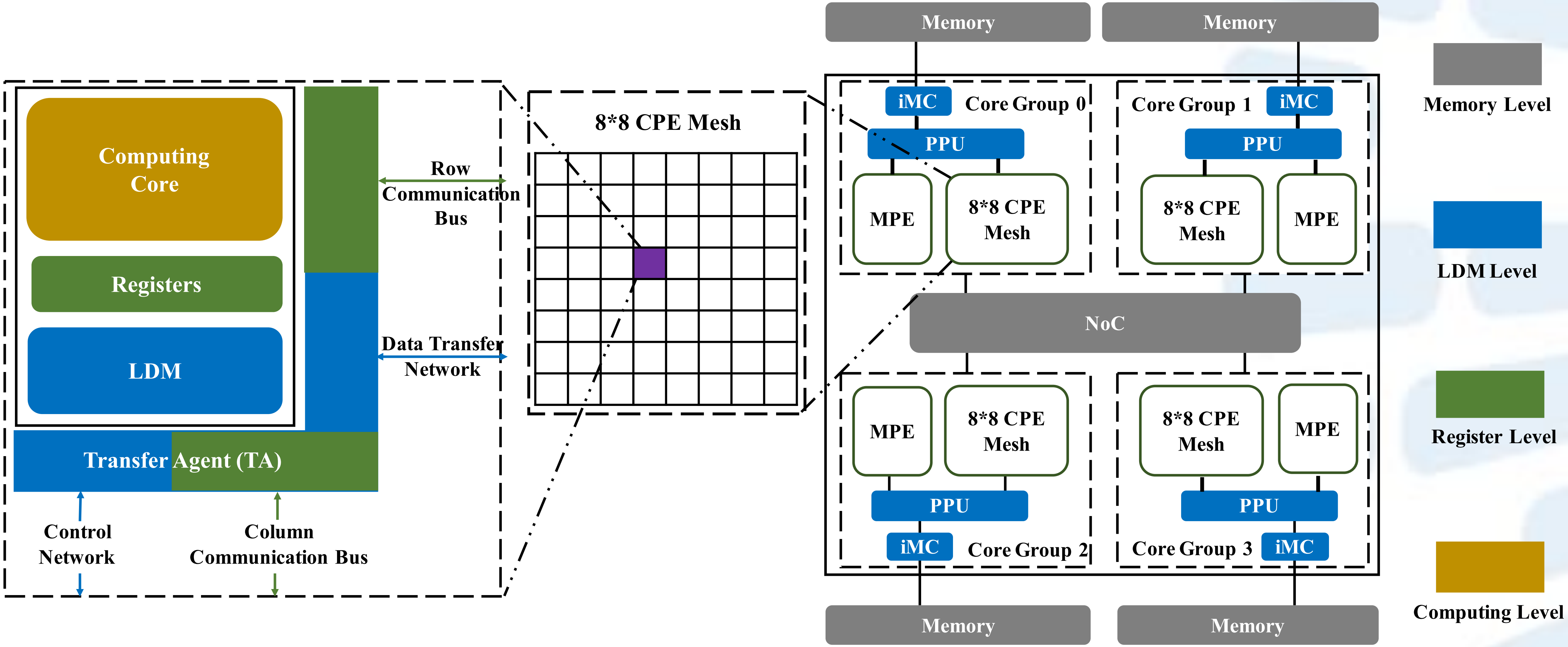
Wide vector

Scratchpad buffer instead of cache

Inter-core communication and synchronization support

Inherited core group structure (divide and conquer)

SW26010: Sunway 260-Core Processor



SW26010: Overview

Peak Performance	**3.06 TFlops**
Memory	32 GB
Memory Bandwidth	136.5 GB/s
# core group	4
# cores	260

High-Density Integration of the Computing System

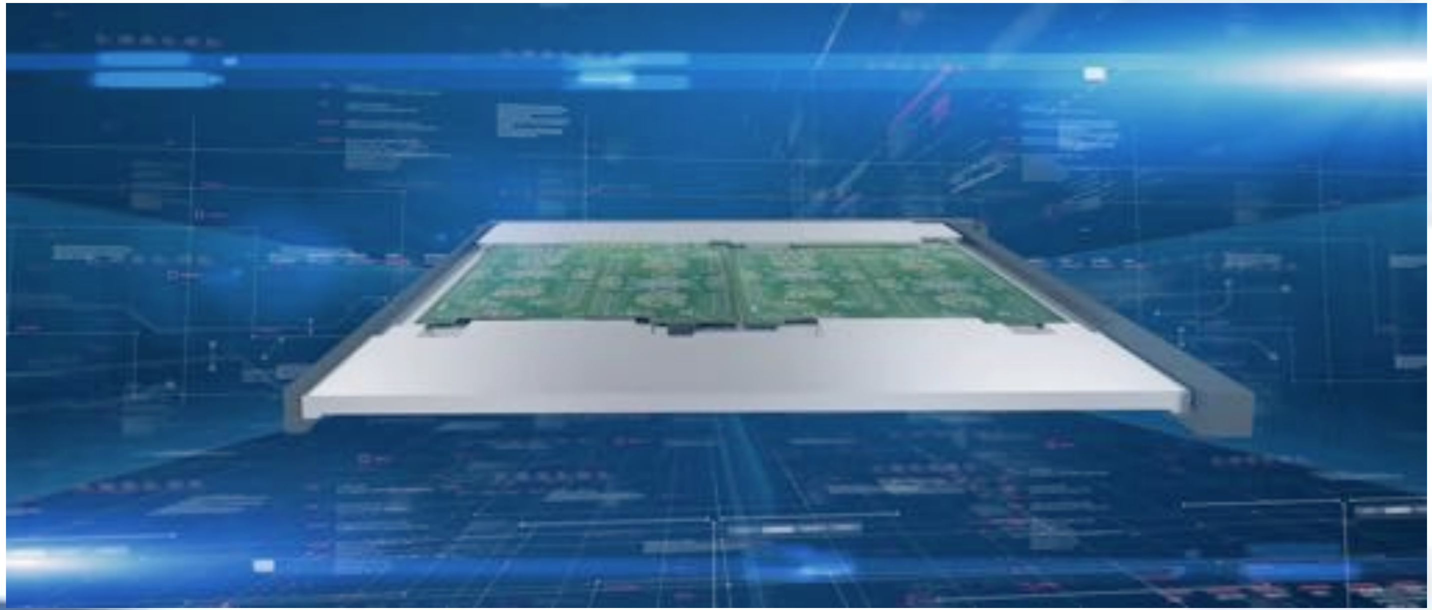
- A Five-Level Integration Hierarchy
  - computing node
  - computing board
  - super node
  - cabinet
  - entire computing system

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 Computing System

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

A System with Over 10 Million Cores

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

Sunway TaihuLight V.S. Other Systems

System	TaihuLight	Tianhe-2	Piz Daint	Titan	Sequoia	K
Peak Performance (PFlops)	**125.4**	54.9	36.2	27.1	20.1	11.3
Total Memory (TB)	**1310**	1024	340	710	1572	1410
Linpack Performance (PFlops)	**93.0(74%)**	33.9(62%)	19.6(54.1%)	17.6(65%)	17.2(85.3%)	10.5(93.2%)
Rank of Top500	**1**	2	3	4	5	8
Performance/Power (Mflops/W)	**6051.3**	1901.5	10398	2142.8	2176.6	1060
Rank of Green500	**17**	118	6	109	100	277
GTEPS	**23755.7**	2061.48	###	###	23751	38621
Rank of Graph500	**2**	8	###	###	3	1
HPCG (Pflops)	**0.48**	0.5801	0.48	0.3223	0.3304	0.6027
Rank of HPCG	**3**	2	3	8	7	1

The Sunway Machine Family

Sunway TaihuLight:
- NSCC-Wuxi, 2016
- 260-core processor
- 125 Pflops
- 1st of TOP500

Sunway Exa-Pilot System:
- 2018
- 5 ~ 10 Tflops per node
- 10 ~ 20 Gflops/W

Sunway Exa-Scale System
- 2021?
- 1000 Pflops
- 30 Gflops/W

Outline

- Motivation and State of the Art
- Major Challenges
- Our Contributions
- Performance and Simulation Results
- Summary and Outlook

The Earthquake Hazard

5.12 Wenchuan Earthquake:

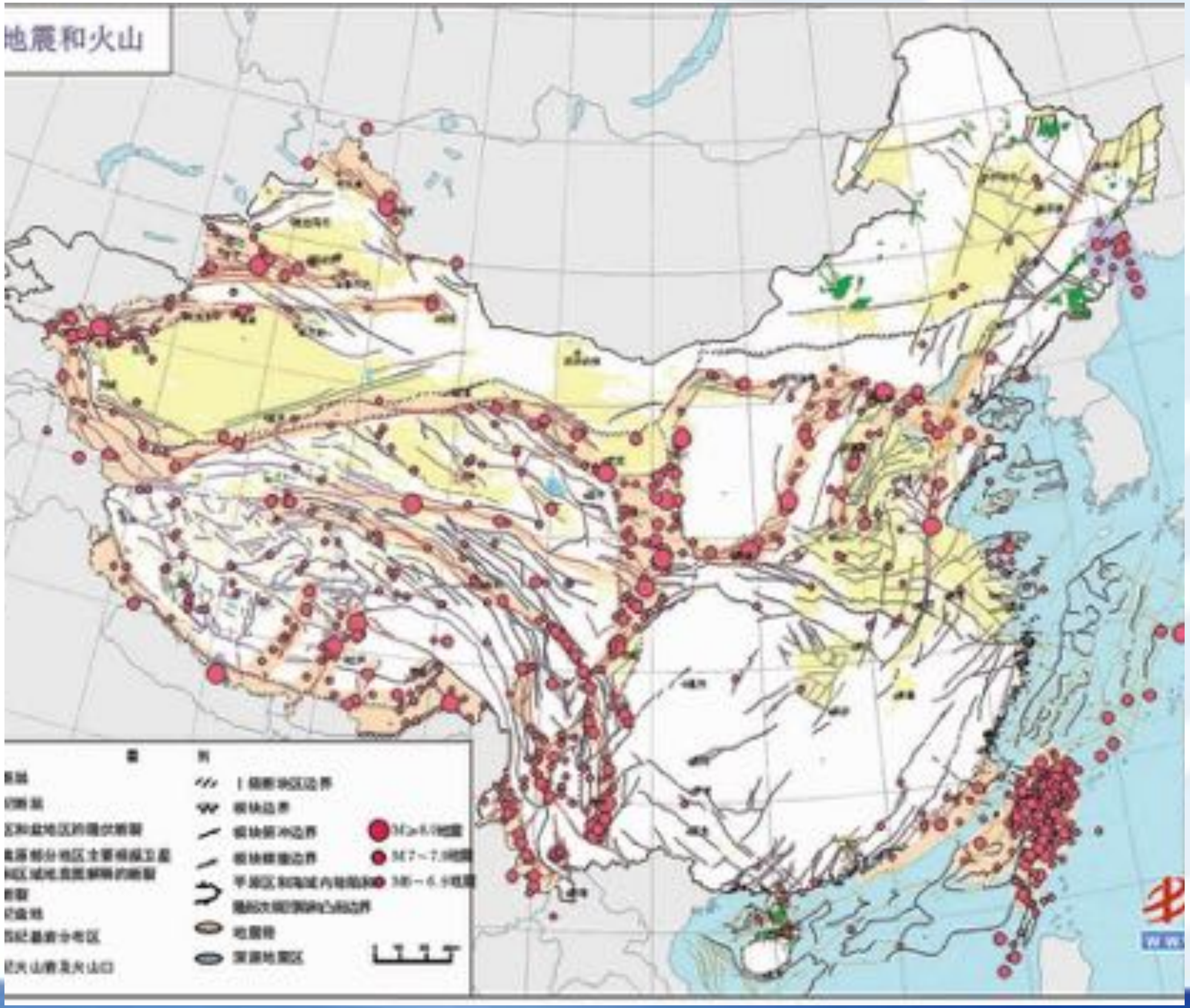
- ~90,000 dead or missing
- ~400,000 injured
- 100 billion CNY loss

The Earthquake Hazard

Earthquake Hazard in China

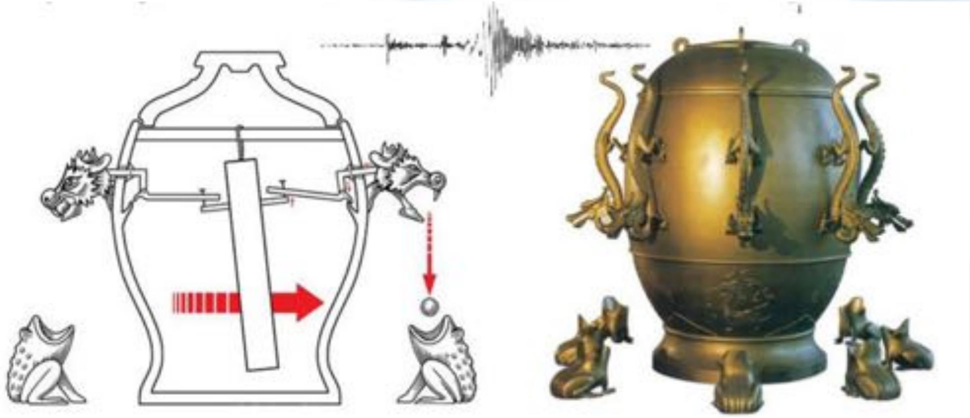
- 23 earthquake zones

- High intensity earthquake zones (M7 above) cover over 50% of the land
  - 20% major transportation lines
  - 21% population
  - 25% hydropower projects
  - 30% large mines



Numerical Earthquake Prediction: the Ultimate Dream

- Numerical Earthquake Prediction
  - extremely difficult if we target all three key elements (time, location, and magnitude) concurrently
- Sub-problems are feasible and still meaningful
  - target two of the three elements
  - reduce the hazard and risk



Zhang, Heng
AD 78-139

Examples of Meaningful Sub-Problems

- Aftershock prediction
  - known location, predict time and magnitude
  - much easier than earthquake prediction, but still unresolved
- Categorization of regional earthquake risks
  - no limit on time, focused on location and magnitude
  - long-term evaluation of risks
- Earthquake risk prediction (for heavily populated and important infrastructures) based on scenario simulations
  - scenario-oriented (location specified, and time independent)
  - accurate prediction of both the magnitude and the hazard distribution

Examples of Meaningful Sub-Problems

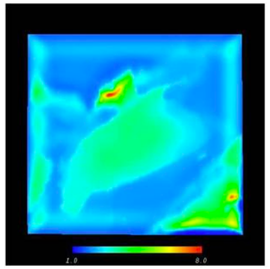
Much has been learned from this and other virtual earthquakes about how to reduce risk and improve resilience

- **Beats waiting to learn tragically from the real thing!**

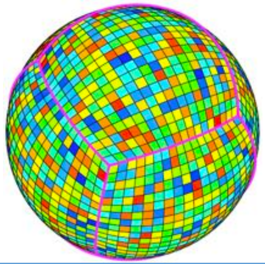
T. H. Jordan

  - no limit on time, focused on location and magnitude
  - long-term evaluation of risks

- Earthquake risk prediction (for heavily populated and important infrastructures) based on scenario simulations
  - scenario-oriented (location specified, and time independent)
  - accurate prediction of both the magnitude and the hazard distribution

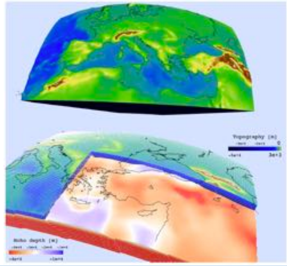


Cray T3D, 1996
- 256 CPUs
- 8 Gflops



SEM

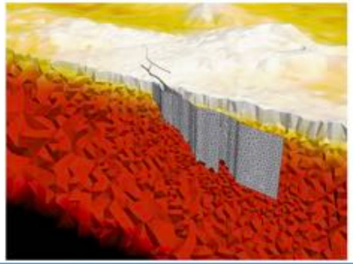
SPECFEM3D



Earth Simulator, 2003
- 1,944 CPUs
- 5 Tflops

Jaguar, 2008
- 29,000 CPUs
- 35.7 Gflops

Cray XK6, 2012
- 896 GPUs
- 135 Tflops



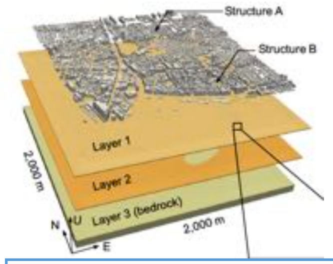
SeisSol

EDGE

Tianhe-2, 2014
- 1.5 million cores (KNC)
- 8.6 Pflops

Cori, 2017
- 612,000 cores (KNL)
- 10.4 Pflops

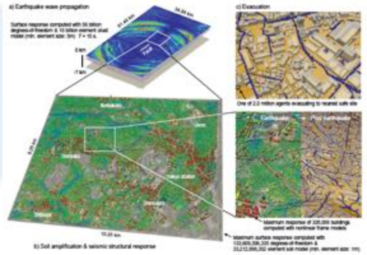
- SPECFEM3D
  - spectral element
- SeisSol to EDGE
  - discontinuous Galerkin finite element
- GAMERA to GOJIRA
  - implicit finite element
- AWP-ODC
  - finite difference
  - plasticity supported



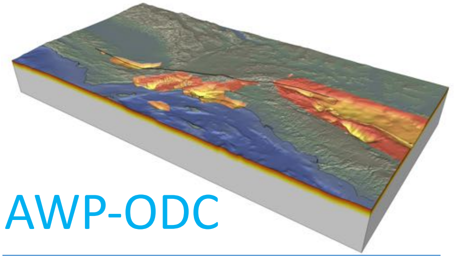
K Computer, 2014
- 663,552 cores
- 0.804 Pflops

GAMERA

GOJIRA

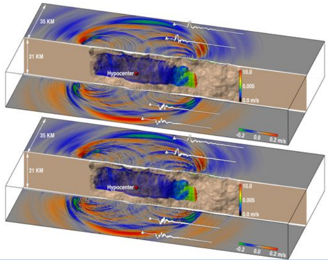


K Computer, 2015
- 663,552 cores
- 1.97 Pflops

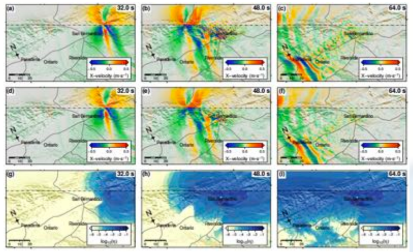


AWP-ODC

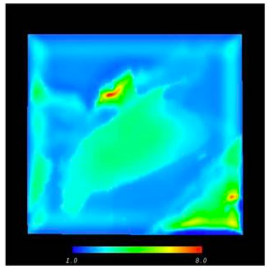
Jaguar, 2010
- 223,074 cores
- 220 Tflops



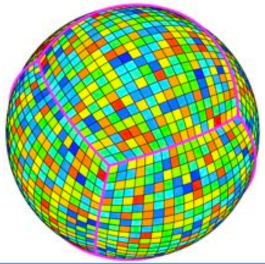
Titan, 2013
- 16,384 GPUs
- 2.33 Pflops GPU



Titan, 2016
- 8,192 GPUs
- 1.6 Pflops non-linear



Cray T3D, 1996
- 256 CPUs
- 8 Gflops

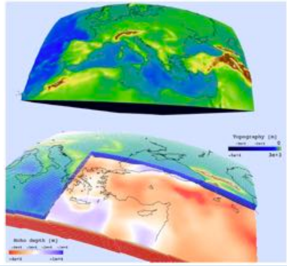


Earth Simulator, 2003
- 1,944 CPUs
- 5 Tflops

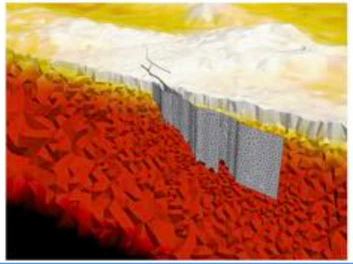
SEM

SPECFEM3D

Jaguar, 2008
- 29,000 CPUs
- 35.7 Gflops



Cray XK6, 2012
- 896 GPUs
- 135 Tflops

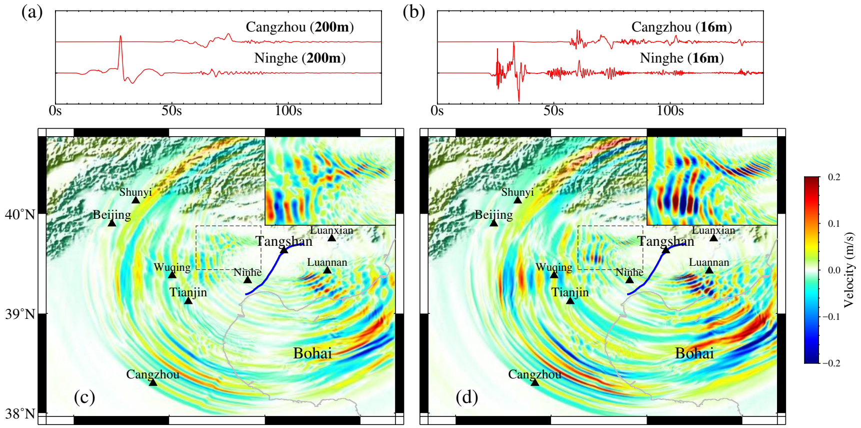


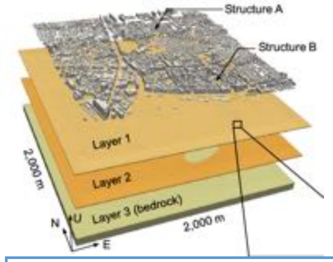
SeisSol

EDGE

Tianhe-2, 2014
- 1.5 million cores (KNC)
- 8.6 Pflops

Cori, 2017
- 612,000 cores (KNL)
- 10.4 Pflops

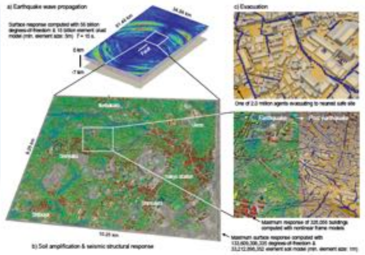




GAMERA

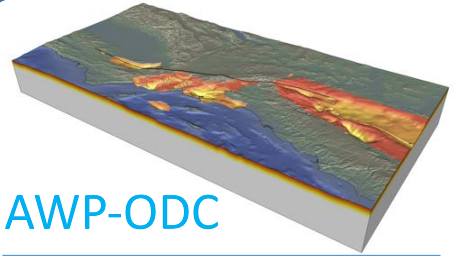
GOJIRA

K Computer, 2014
- 663,552 cores
- 0.804 Pflops

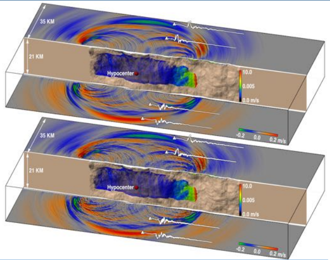


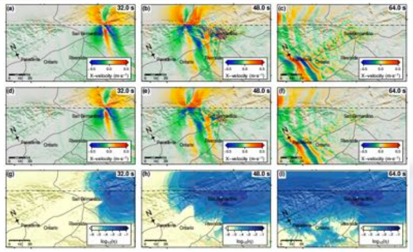
K Computer, 2015
- 663,552 cores
- 1.97 Pflops

Sunway TaihuLight, 2017
- 10,140,000 cores
- 15.2 Pflops without compression
- 18.9 Pflops with compression



AWP-ODC





Jaguar, 2010
- 223,074 cores
- 220 Tflops

Titan, 2013
- 16,384 GPUs
- 2.33 Pflops GPU

Titan, 2016
- 8,192 GPUs
- 1.6 Pflops non-linear

Outline

- Motivation and State of the Art
- Major Challenges
- Our Contributions
- Performance and Simulation Results
- Summary and Outlook

A Typical Earthquake Simulation Setup

300 km x 300 km x 50 km

15,000 x 15,000 x 2,500 (562.5 billion grids)

30~40 variables per grid

20 meter, 10 Hz

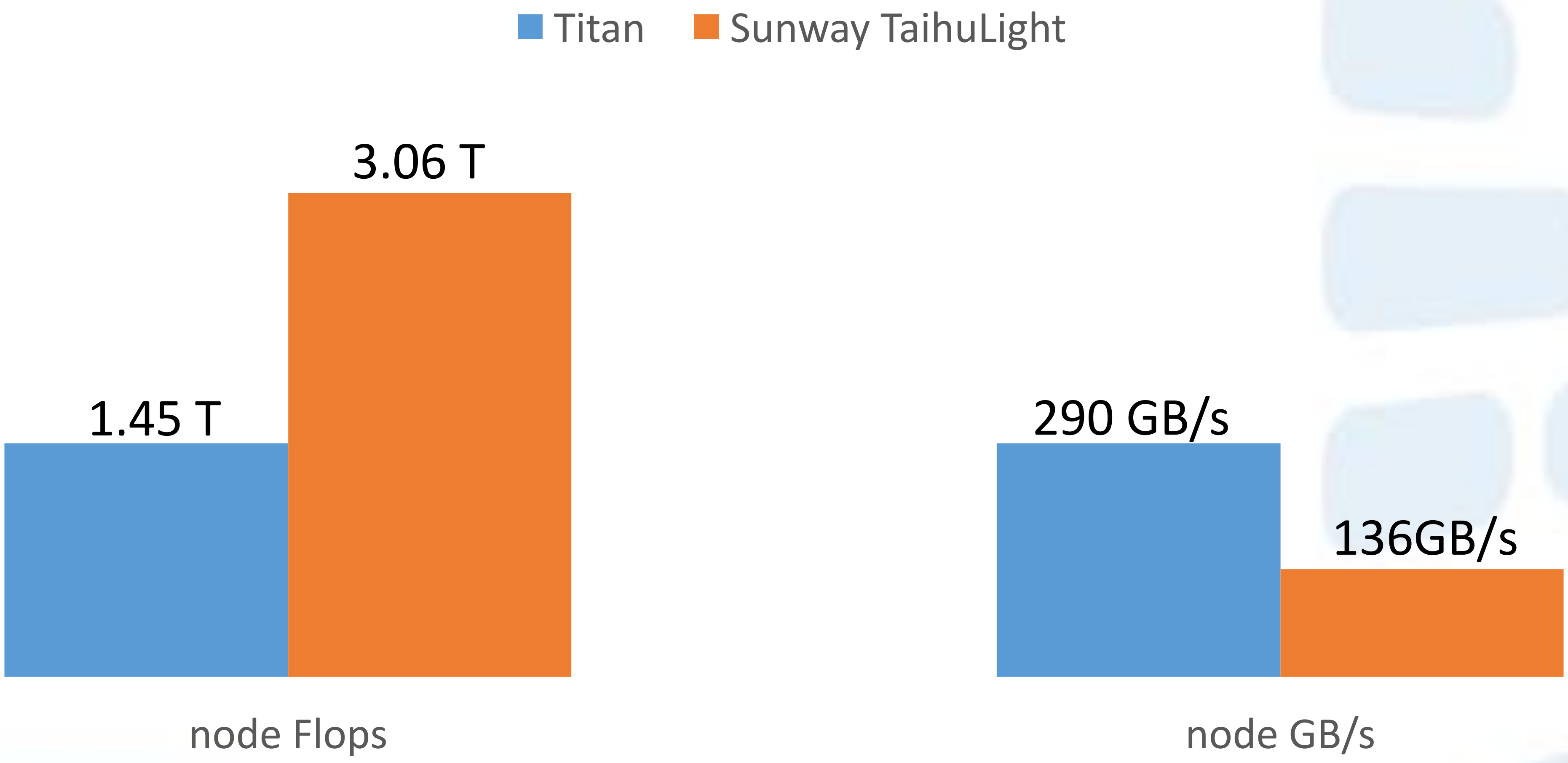
100 s / 0.001 s = 10^5 (time steps)

~500 FLOPs per grid

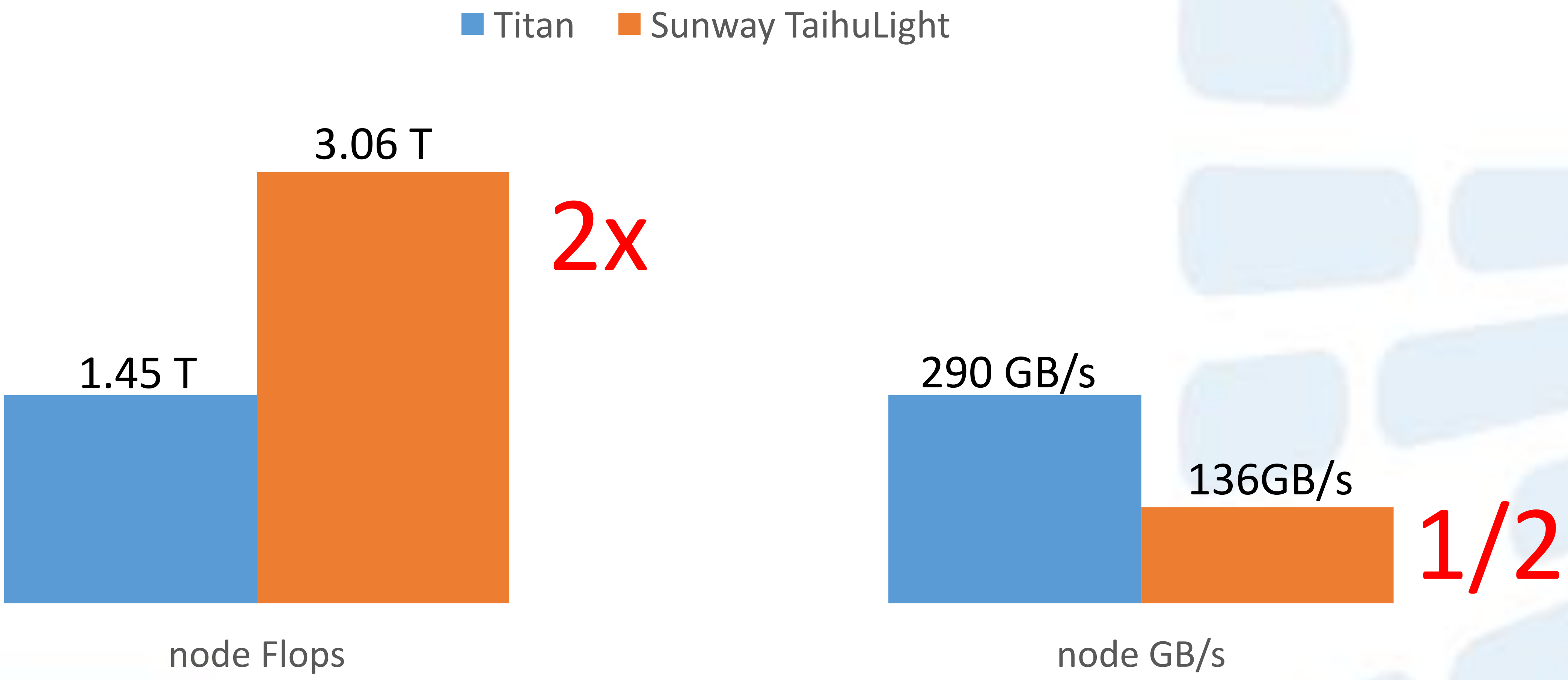
memory size: ~150 TB

total flop: 100 Eflop

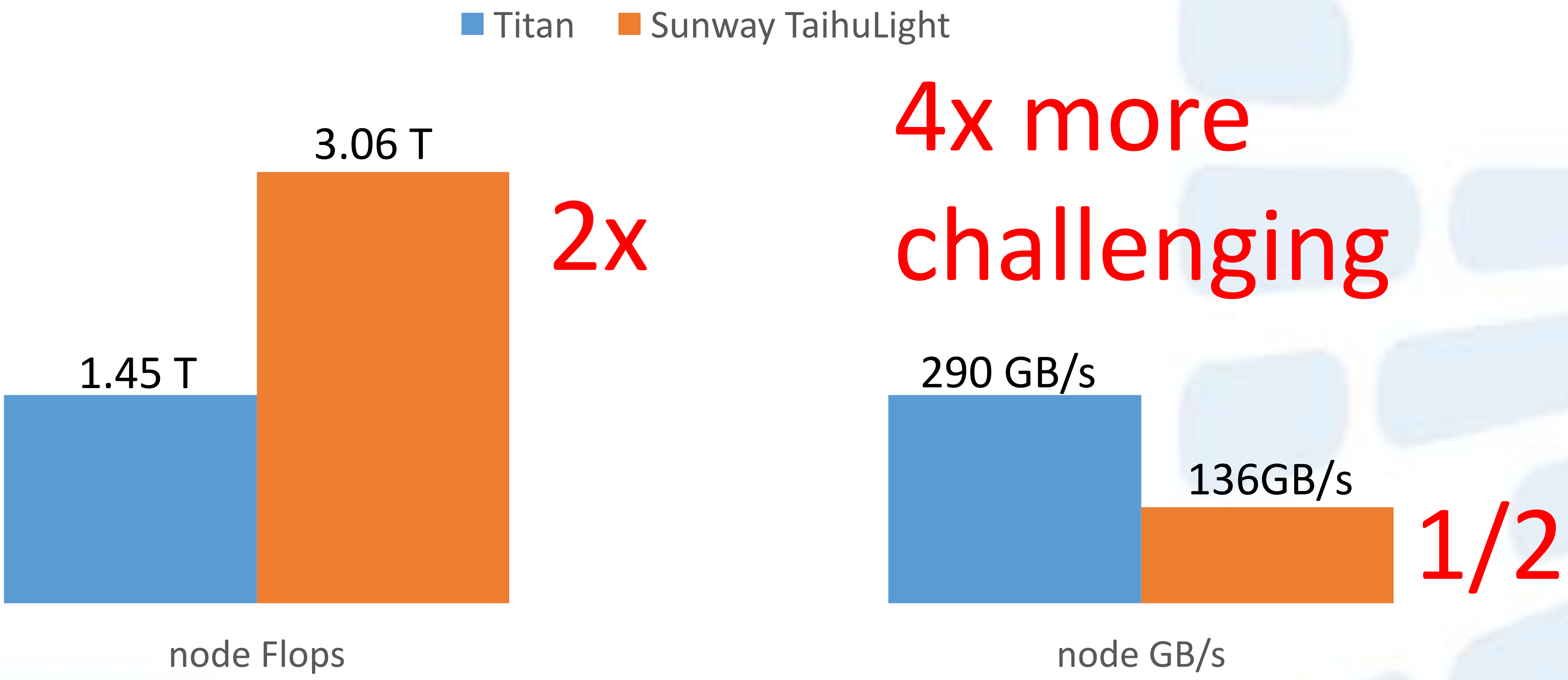
The Memory Barrier



The Memory Barrier



The Memory Barrier

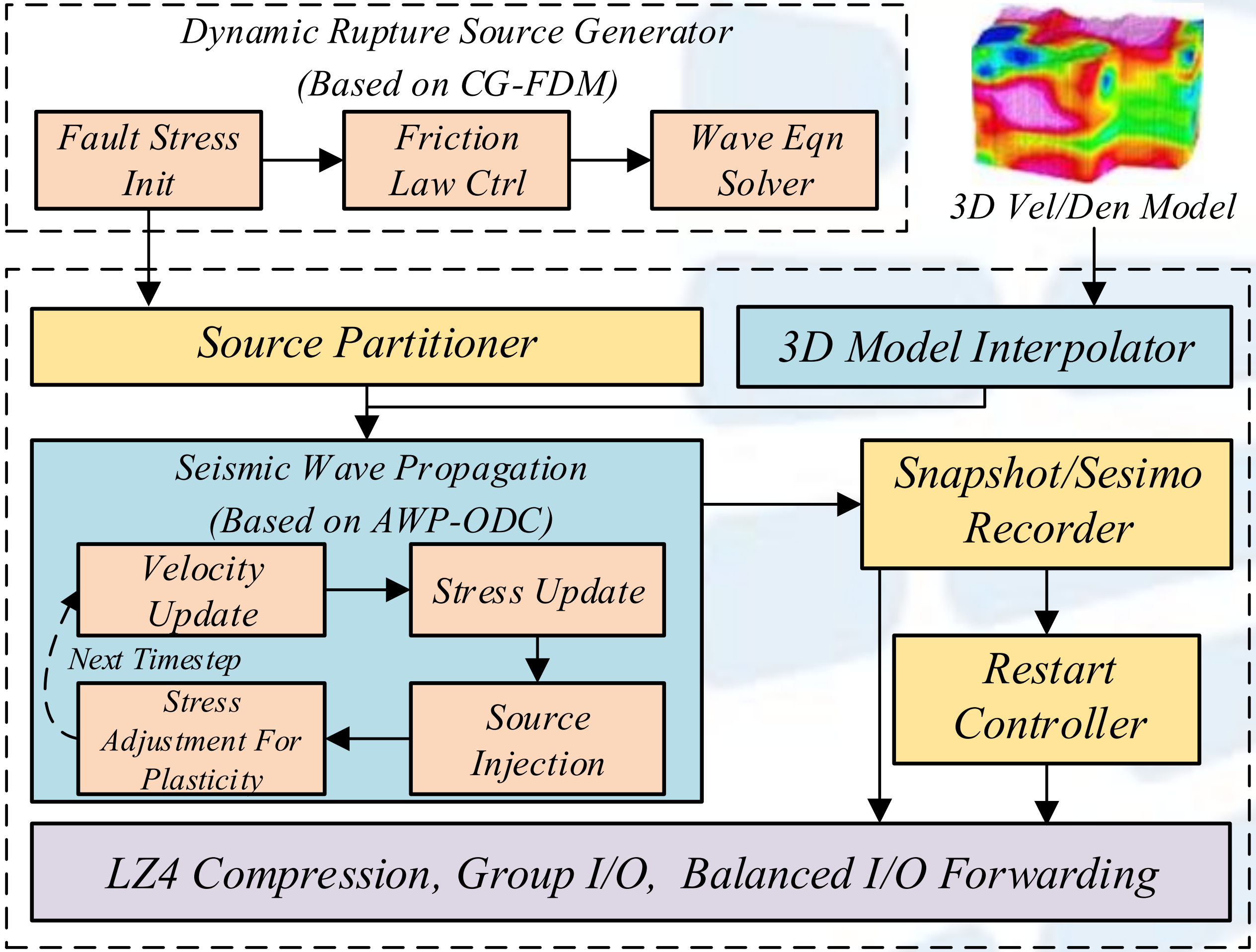


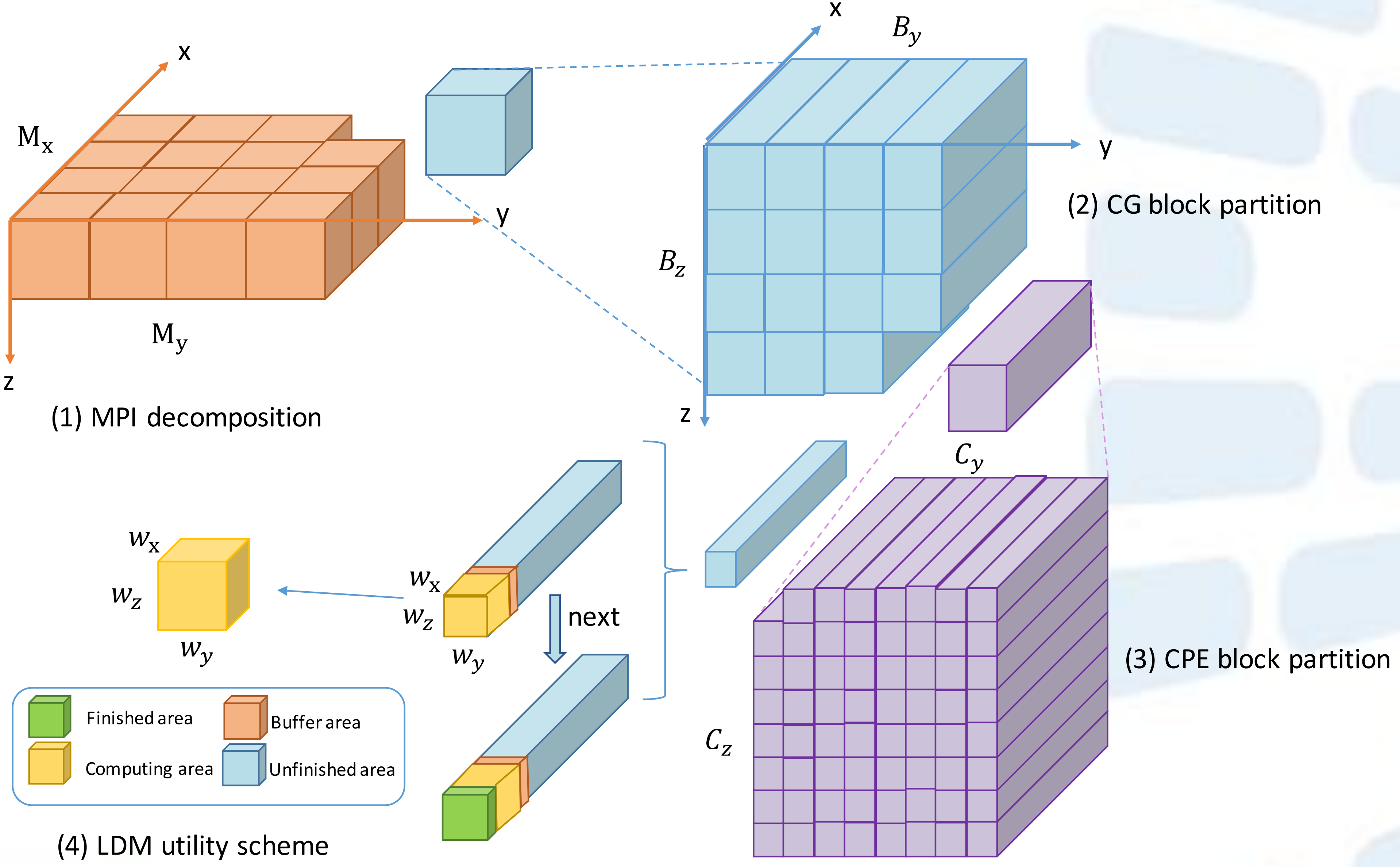
Outline

- Motivation and State of the Art
- Major Challenges
- Our Contributions
- Performance and Simulation Results
- Summary and Outlook

Our Earthquake Simulation Framework

- Dynamic rupture source generator (originated from CG-FDM)
- Seismic wave propagation (originated from AWP-ODC)
- Other utilities:
  - source partitioner (~70 TB input)
  - 3D Model Interpolator
  - Restart controller (~100 TB snapshot)

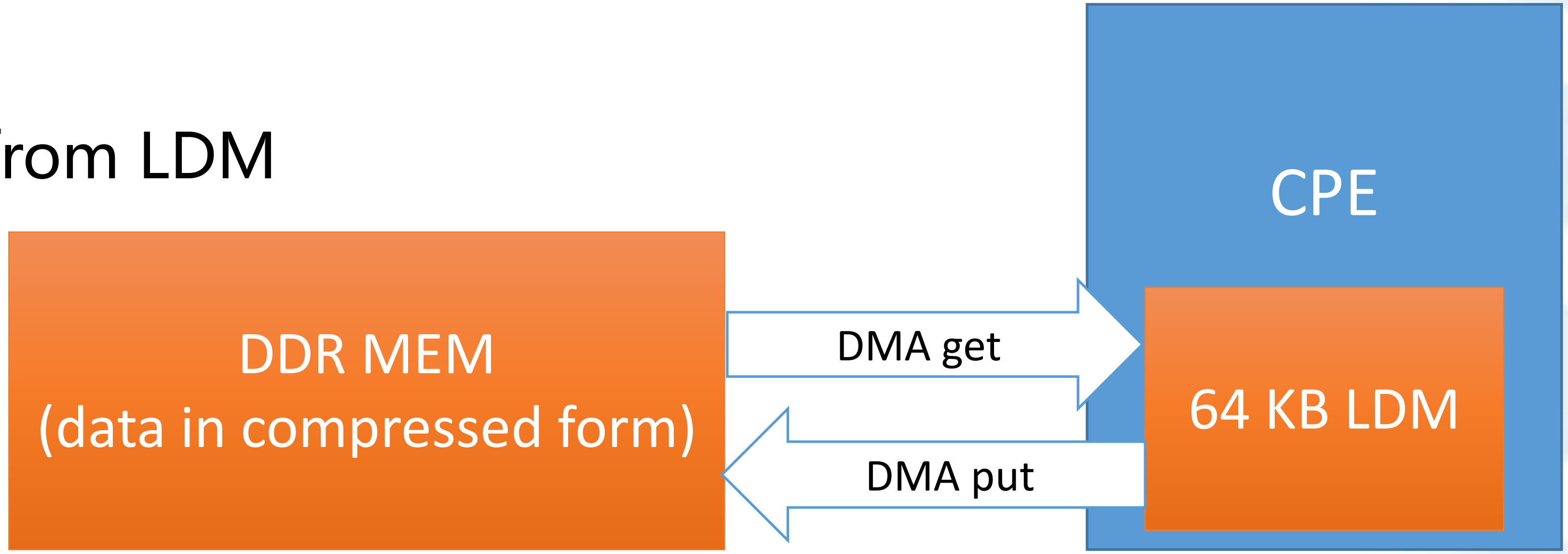




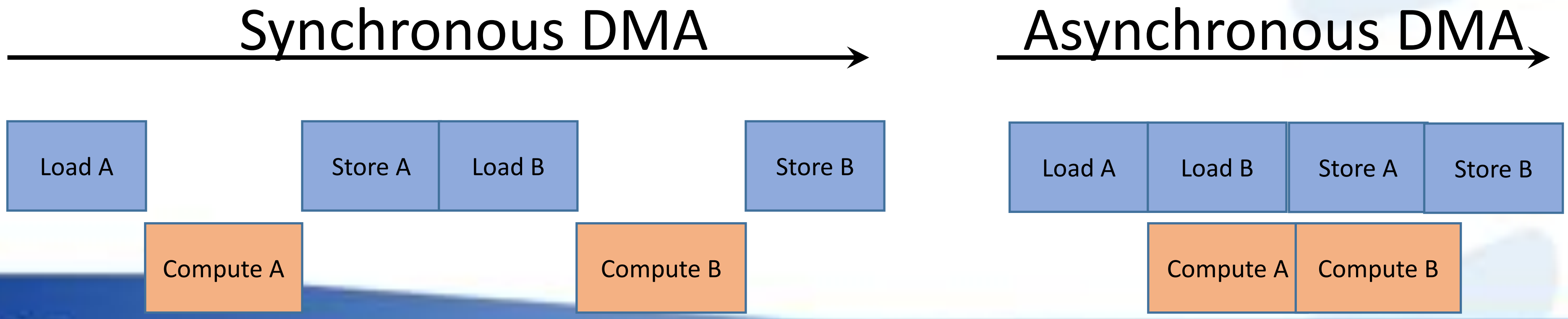
Three-Level Domain Decomposition

The DMA-based Memory Model

- Compute from LDM

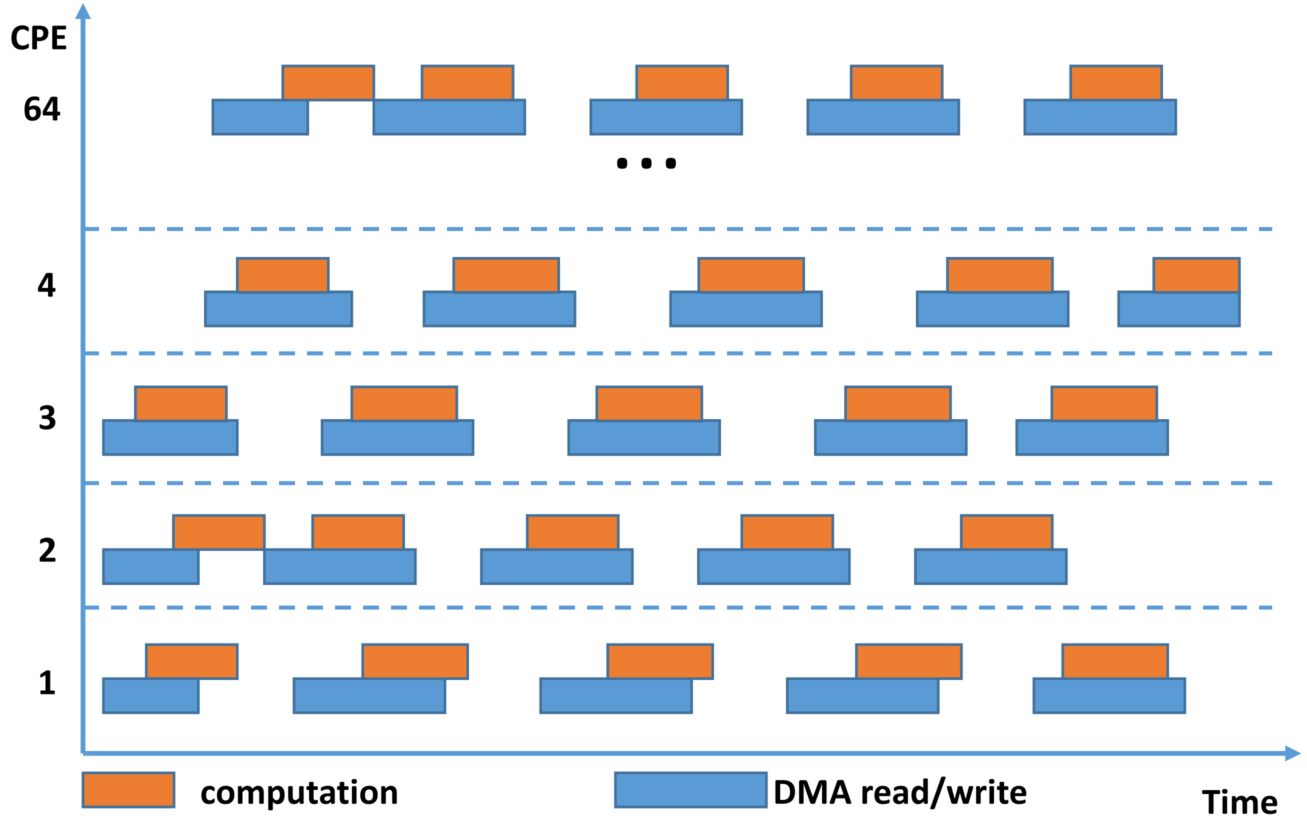


- A carefully designed DMA scheme to overlap DMA and compute

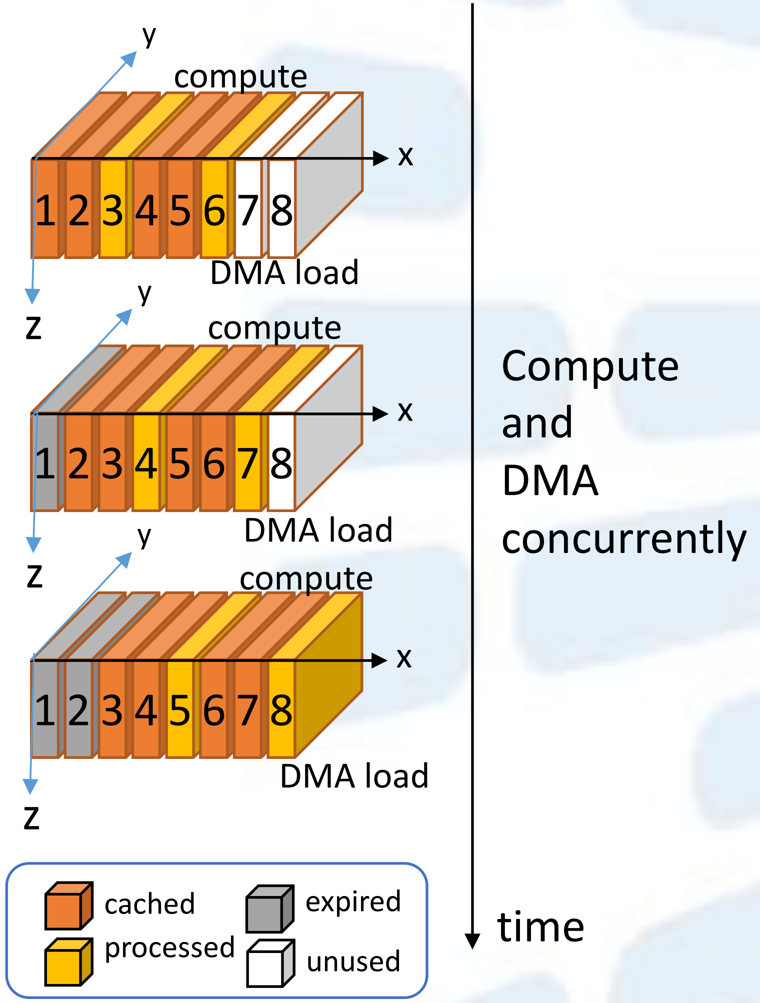


The DMA and Buffering Scheme

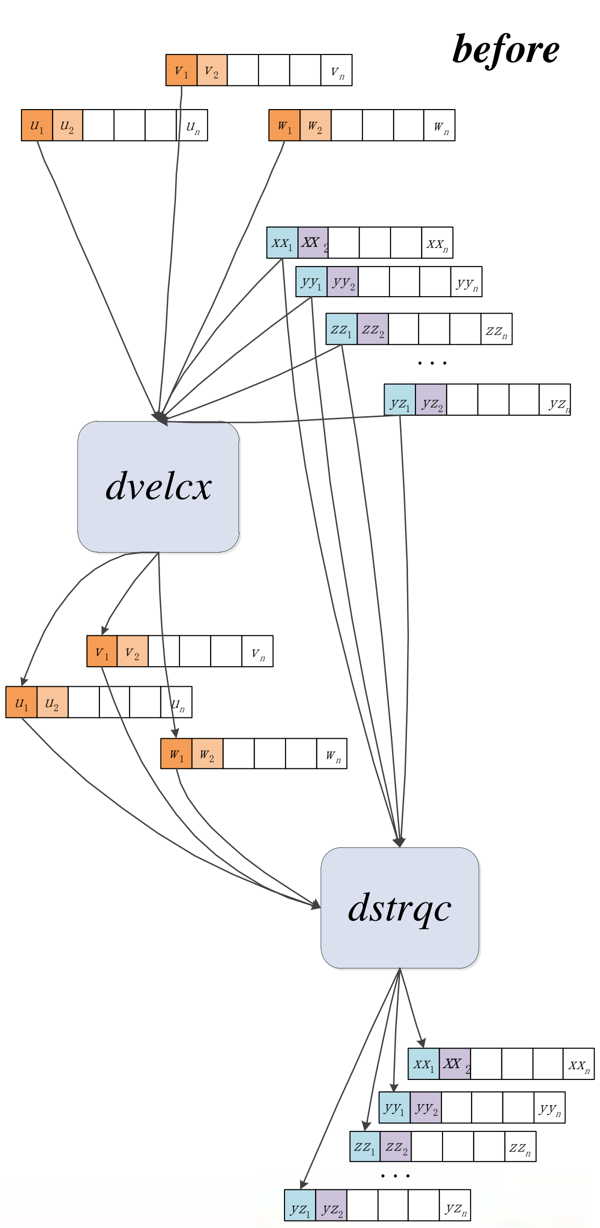
(1) Overlap in different CPEs



64 CPE share the DMA bandwidth, with multithreading to hide the latency.



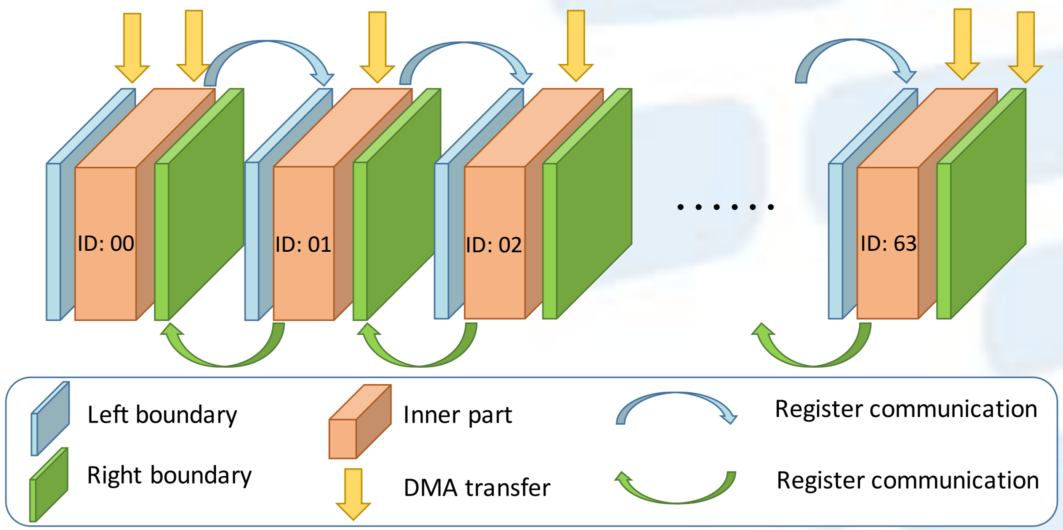
(2) Overlap inside CPE



(1) array fusion,

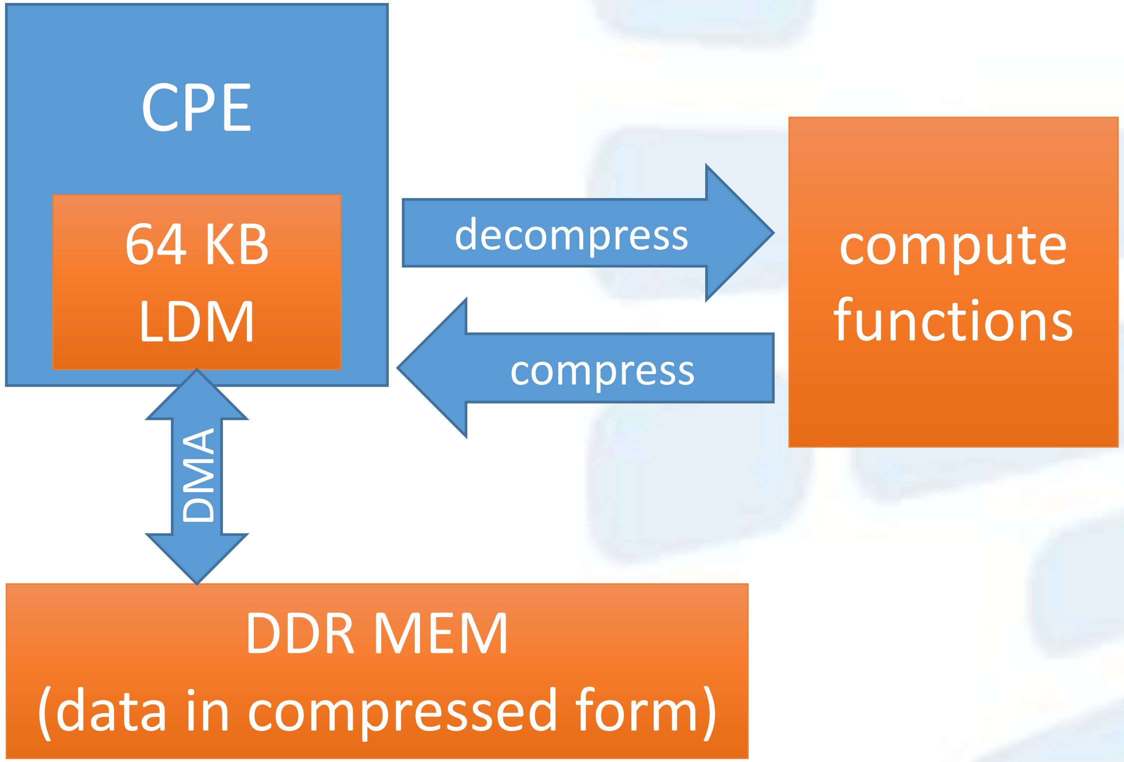
(2) halo exchange through register communication,

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



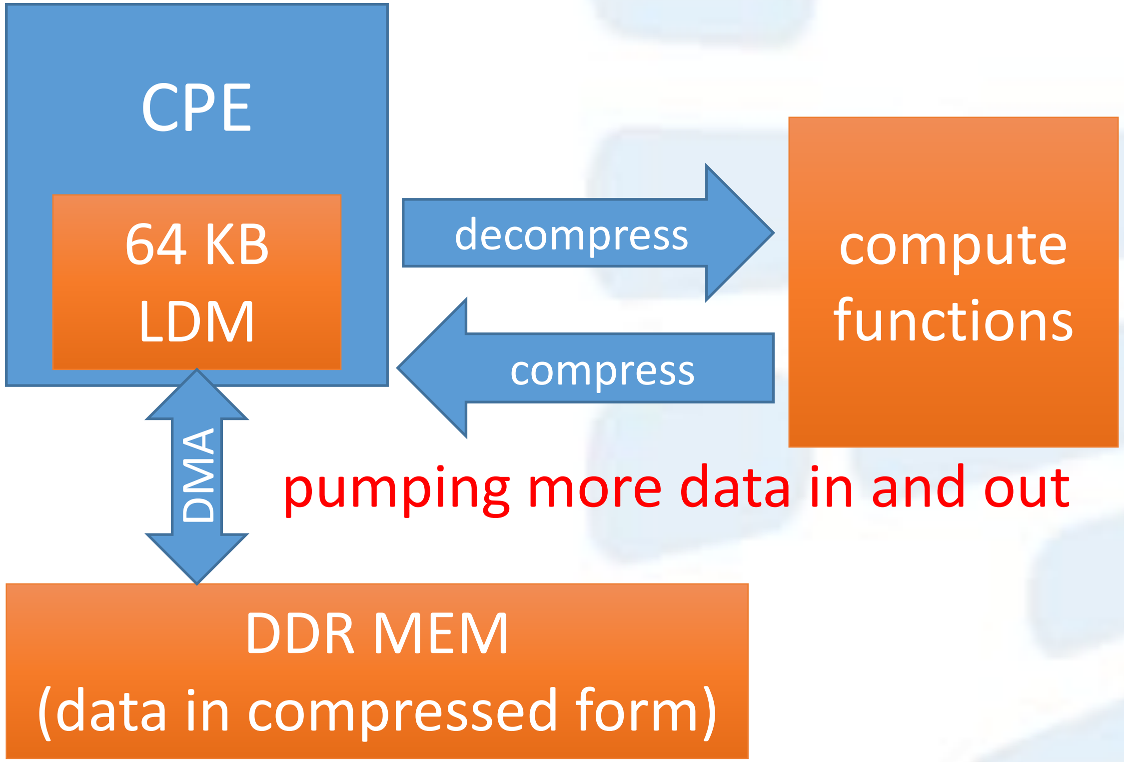
Compression: Squeezing Extra Performance

	Peak	**Utilized**	**%**
Flops	765 G	94.7 G	12.2%
Memory size	5 G	4.6 G	92%
Memory BW	34 GB/s	25 GB/s	73.5%
LDM size	64 KB	60 KB	93.8%



Compression: Squeezing Extra Performance

	Peak	Utilized	%
Flops	765 G	94.7 G	12.2%
Memory size	5 G	4.6 G	92%
Memory BW	34 GB/s	25 GB/s	73.5%
LDM size	64 KB	60 KB	93.8%



enable even larger problems

Compression: Not an Easy Task

Additional complexity and cost

Extra LDM read/write due to compression/decompression operations

Broken floating-point instruction pipeline

Compression: Further Optimization

Additional complexity and cost

Extra LDM read/write due to compression/decompression operations

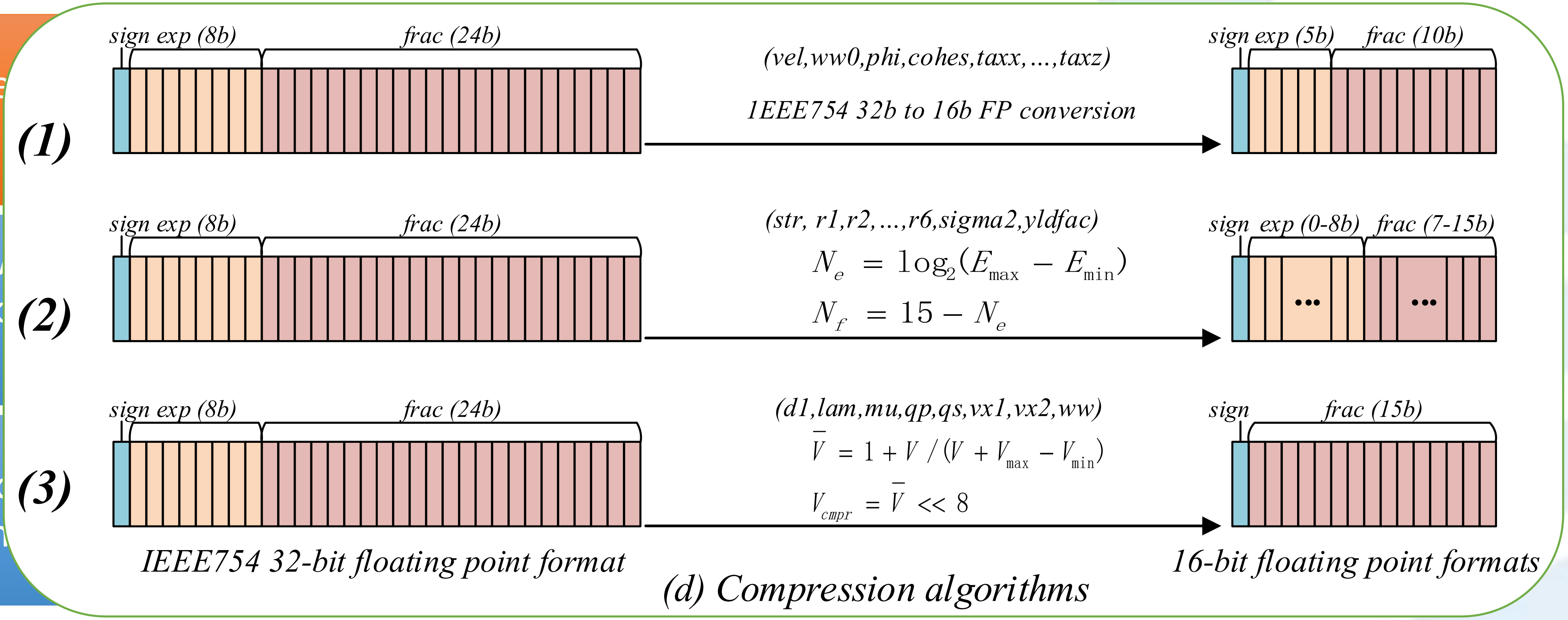
Broken floating-point instruction pipeline

Compression: Further Optimization

Additional comple
and cost

Extra LDM read/w
compression/deco
operations

Broken floating-po
instruction pipelin



(d) Compression algorithms

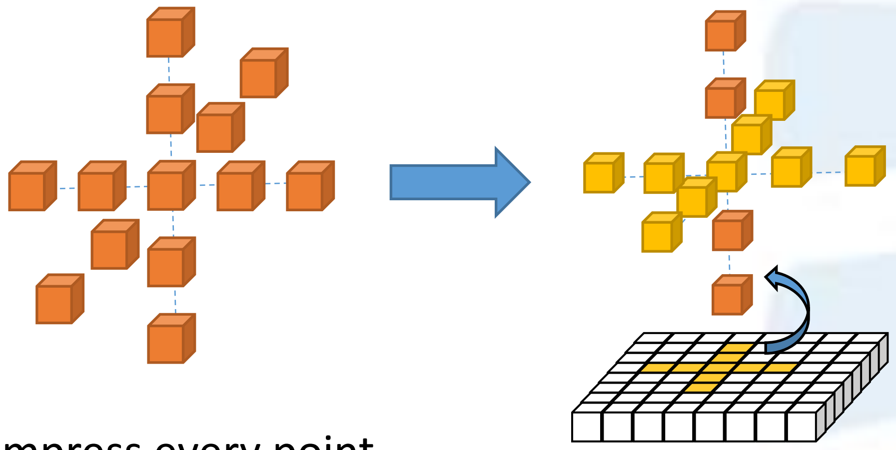
1/3 of original performance

Compression: Further Optimization

Additional complexity and cost

Extra LDM read/write due to compression/decompression operations

Broken floating-point instruction pipeline



compress every point on the fly

buffering a plane

1/3 to 90% of original performance

Compression: Further Optimization

Additional complexity and cost

Extra LDM read/write due to compression/decompression operations

Broken floating-point instruction pipeline

```
LOAD LDM1,$ra
SSL $ra, $ra
STORE $a, LDM1
LOAD LDM2,$rb
SSL $rb, $rb
STORE $rb, LDM2
LOAD LDM3, $rc
SSL $rc $rc
STORE $rc, LDM3
LOAD LDM1,$ra
LOAD LDM2,$rb
ADD $ra, $rb, $ra
LOAD LDM3, $rc
MUL $ra, $rc, $ra
STORE $ra, LDM2
```

→

```
LOAD LDM1,$ra
SSL $ra, $ra
LOAD LDM2,$rb
SSL $rb, $rb
LOAD LDM3, $rc
 $rc $rc
ADD $a, $b, $a
MUL $a, $c, $a
STORE $a, LDM2
```

switch the buffering of temporary variables from LDM to registers by using intrinsic assembly instructions, especially for function calls

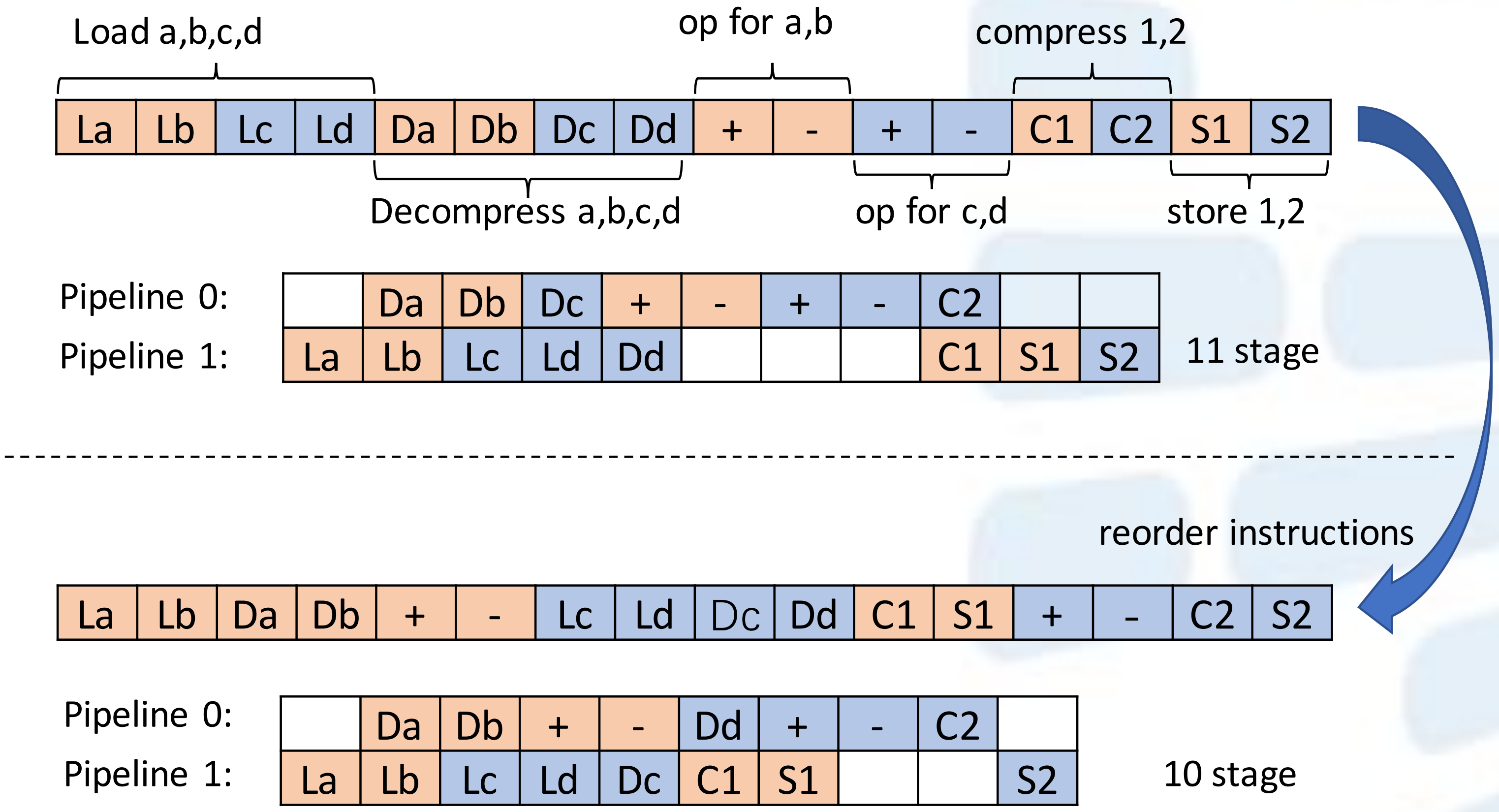
90% to 120% of original performance

Compression: Further Optimization

Additional complexity and cost

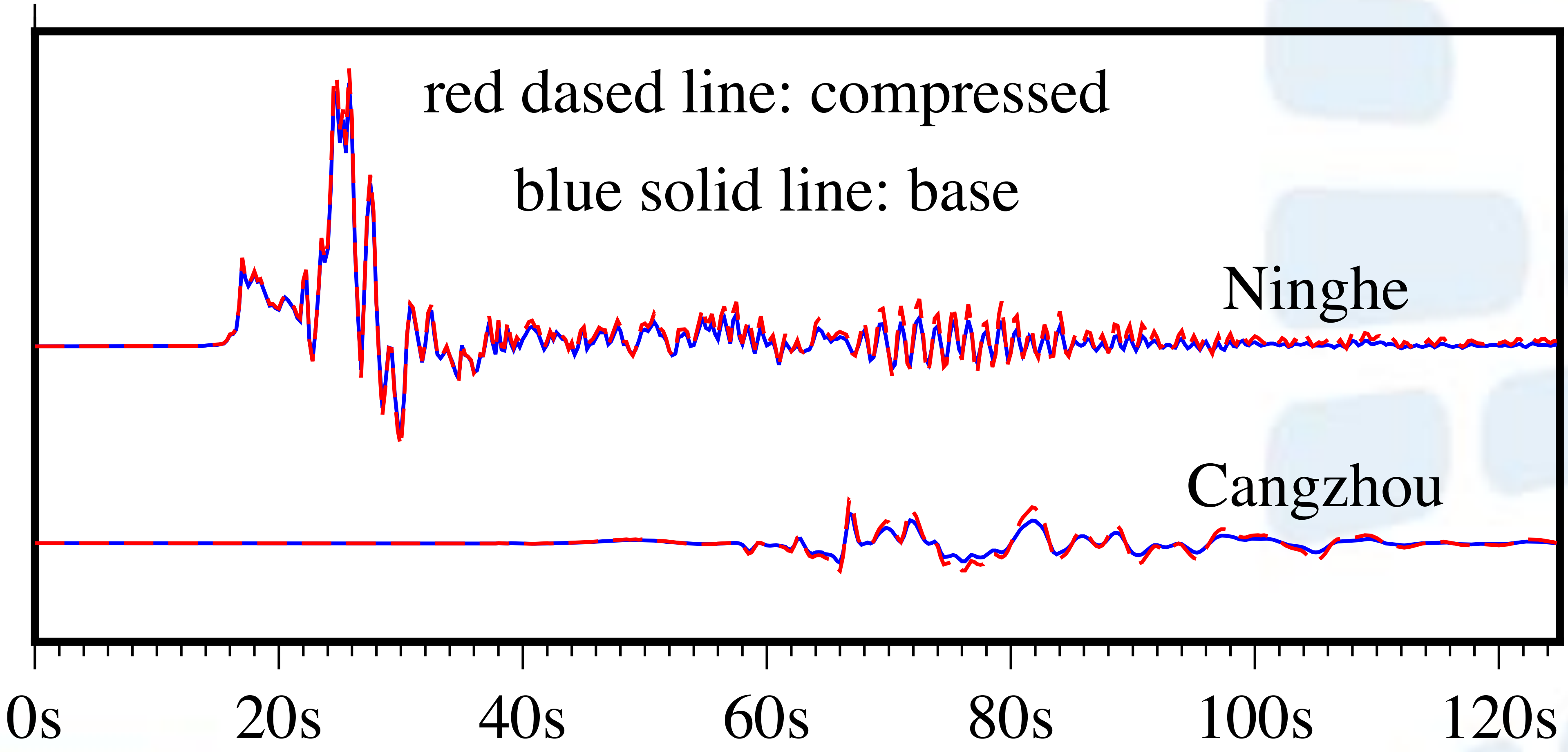
Extra LDM read/write due to compression/decompression operations

Broken floating-point instruction pipeline



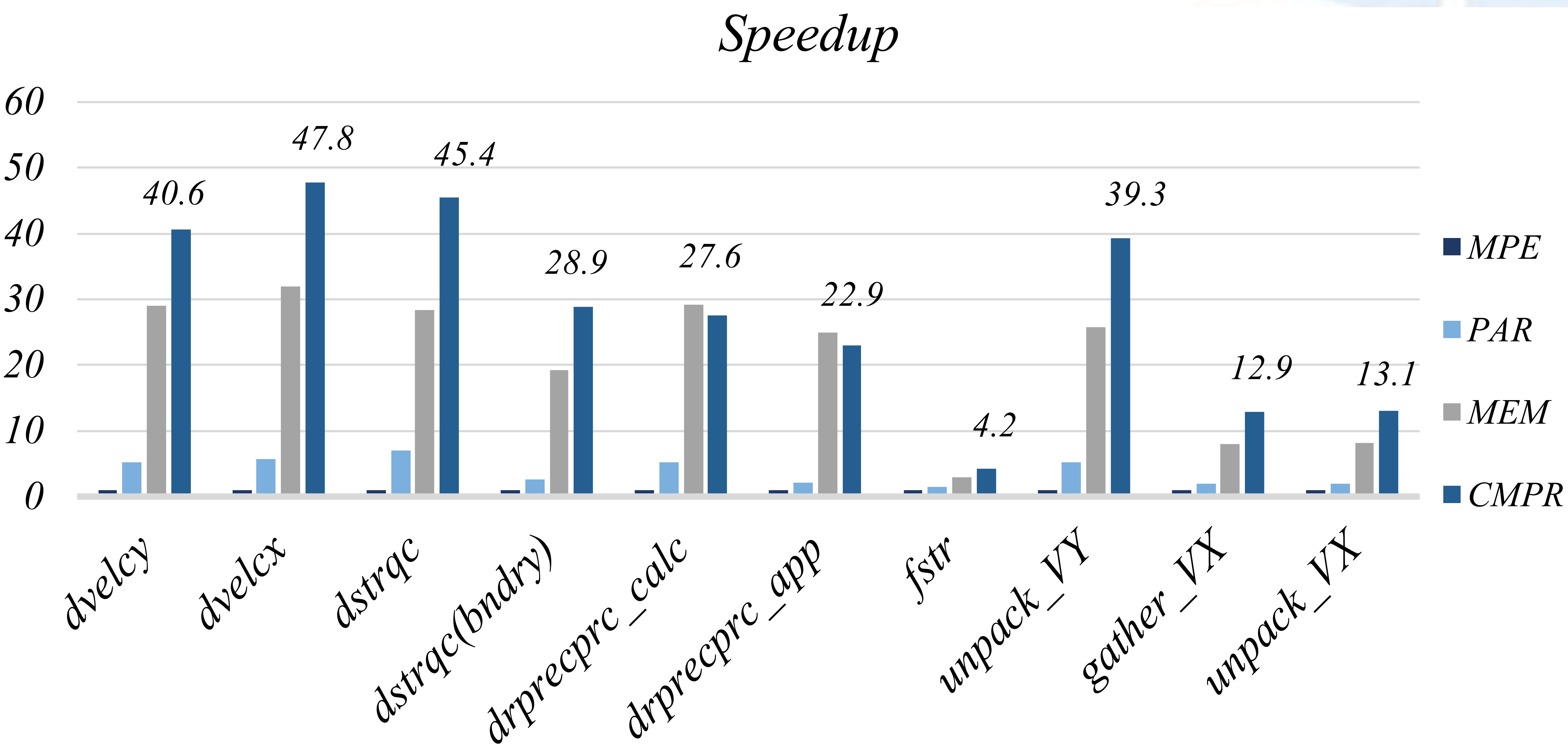
120% to 130% of original performance

On-the-fly Compression

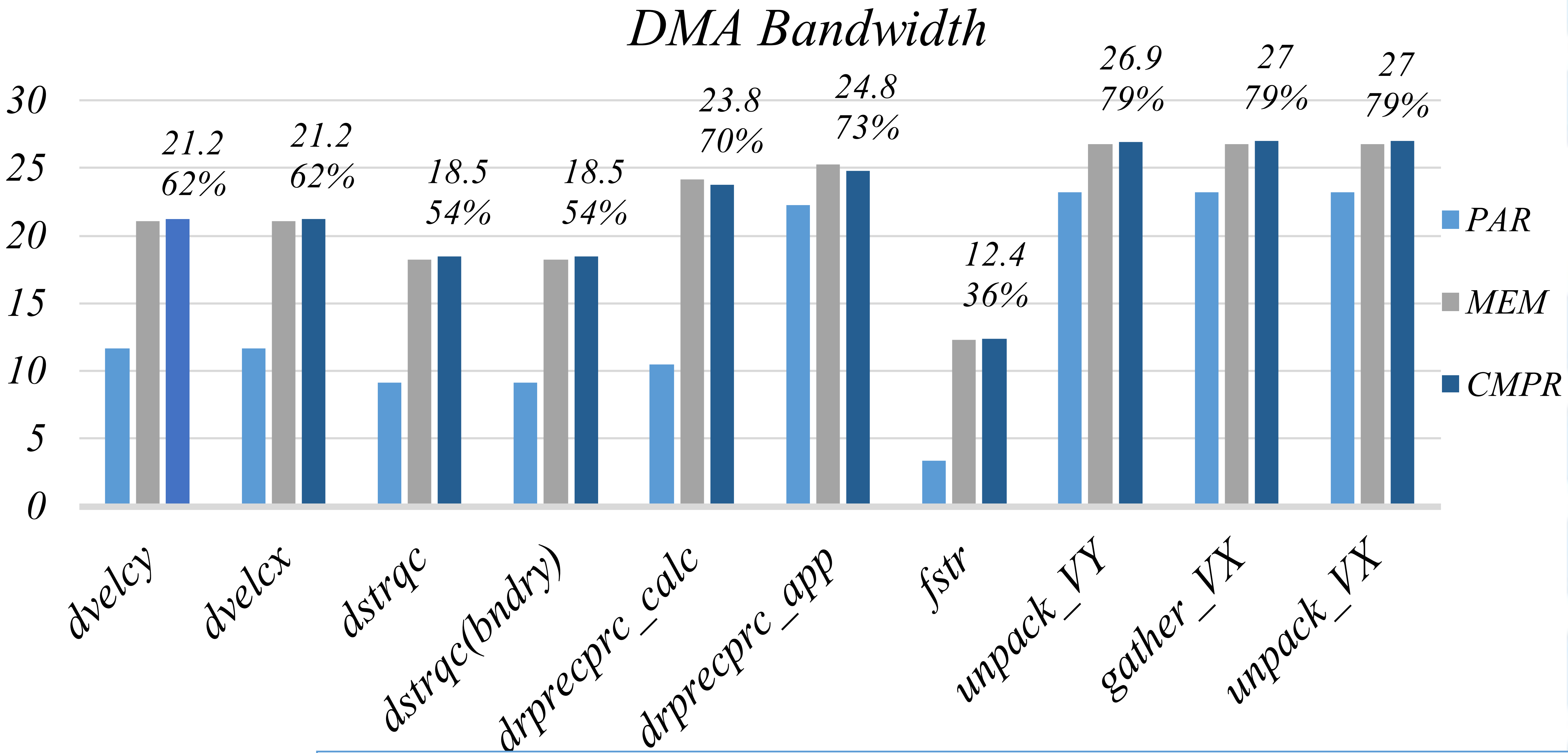


Outline

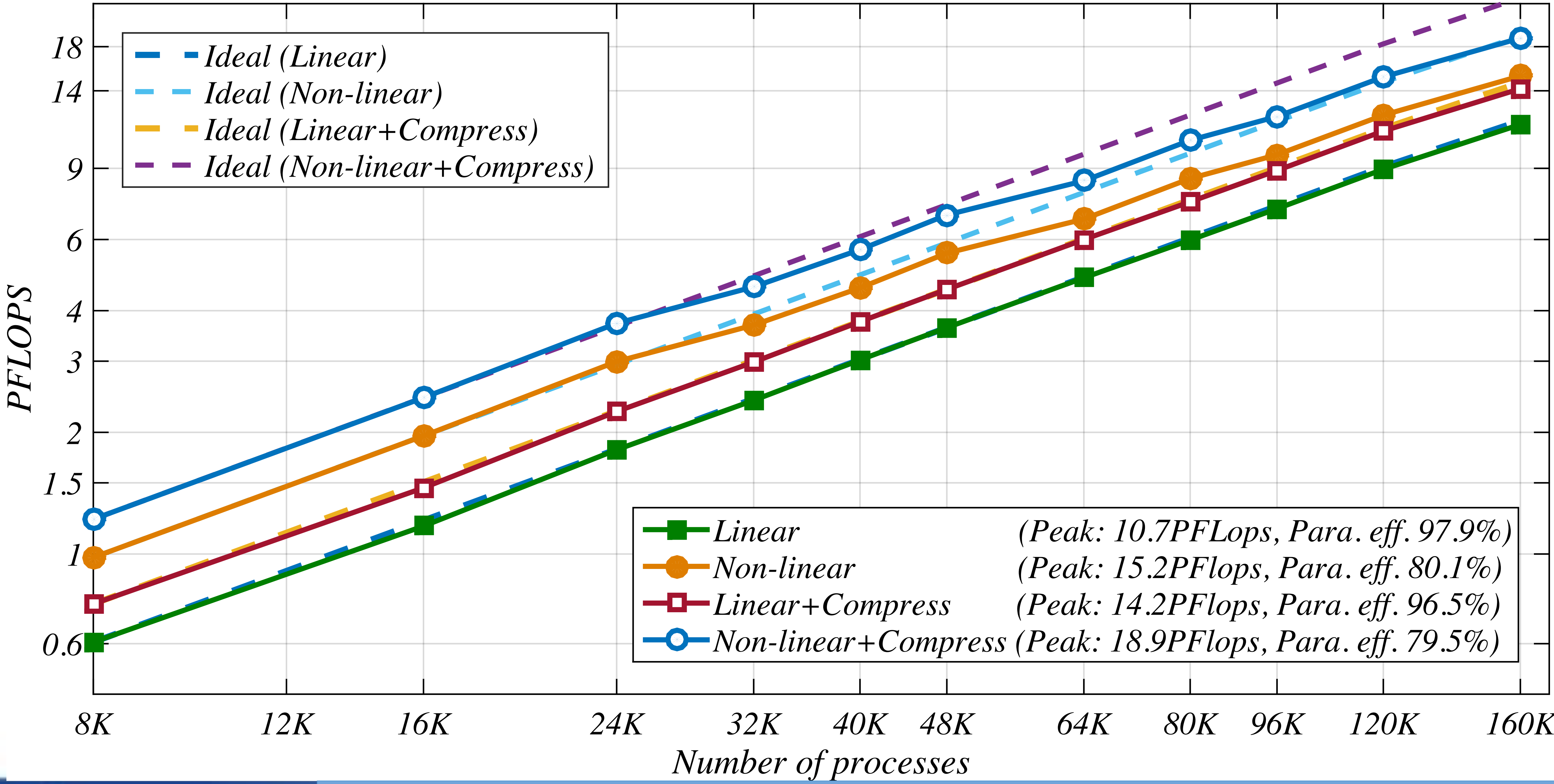
- Motivation and State of the Art
- Major Challenges
- Our Contributions
- Performance and Simulation Results
- Summary and Outlook



20~40x speedup when switching from 1 MPE to 64 CPE

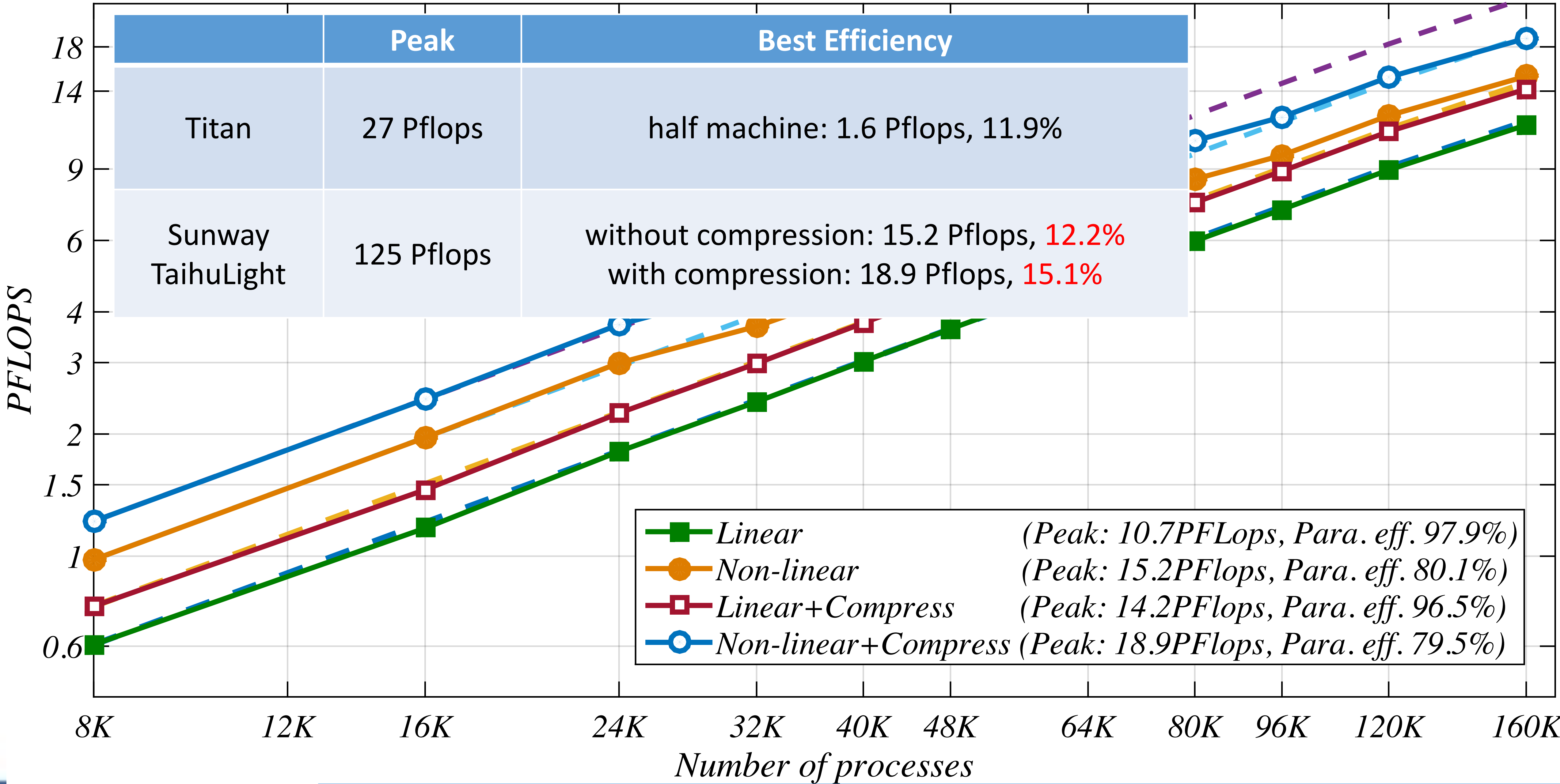


60%~79% memory bandwidth utilization

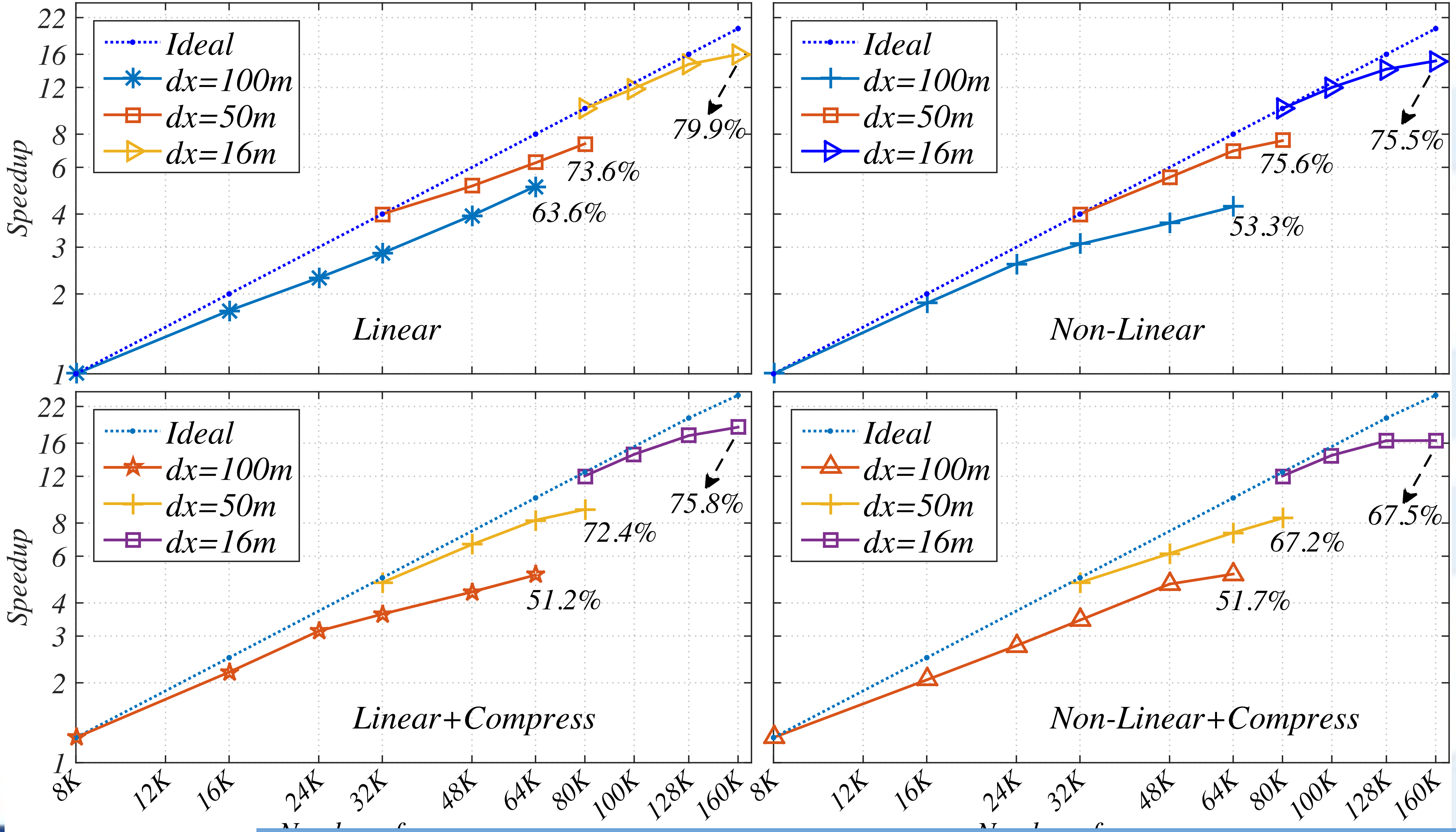


Weak Scaling

	Peak	Best Efficiency
Titan	27 Pflops	half machine: 1.6 Pflops, 11.9%
Sunway TaihuLight	125 Pflops	without compression: 15.2 Pflops, 12.2% with compression: 18.9 Pflops, 15.1%



Weak Scaling

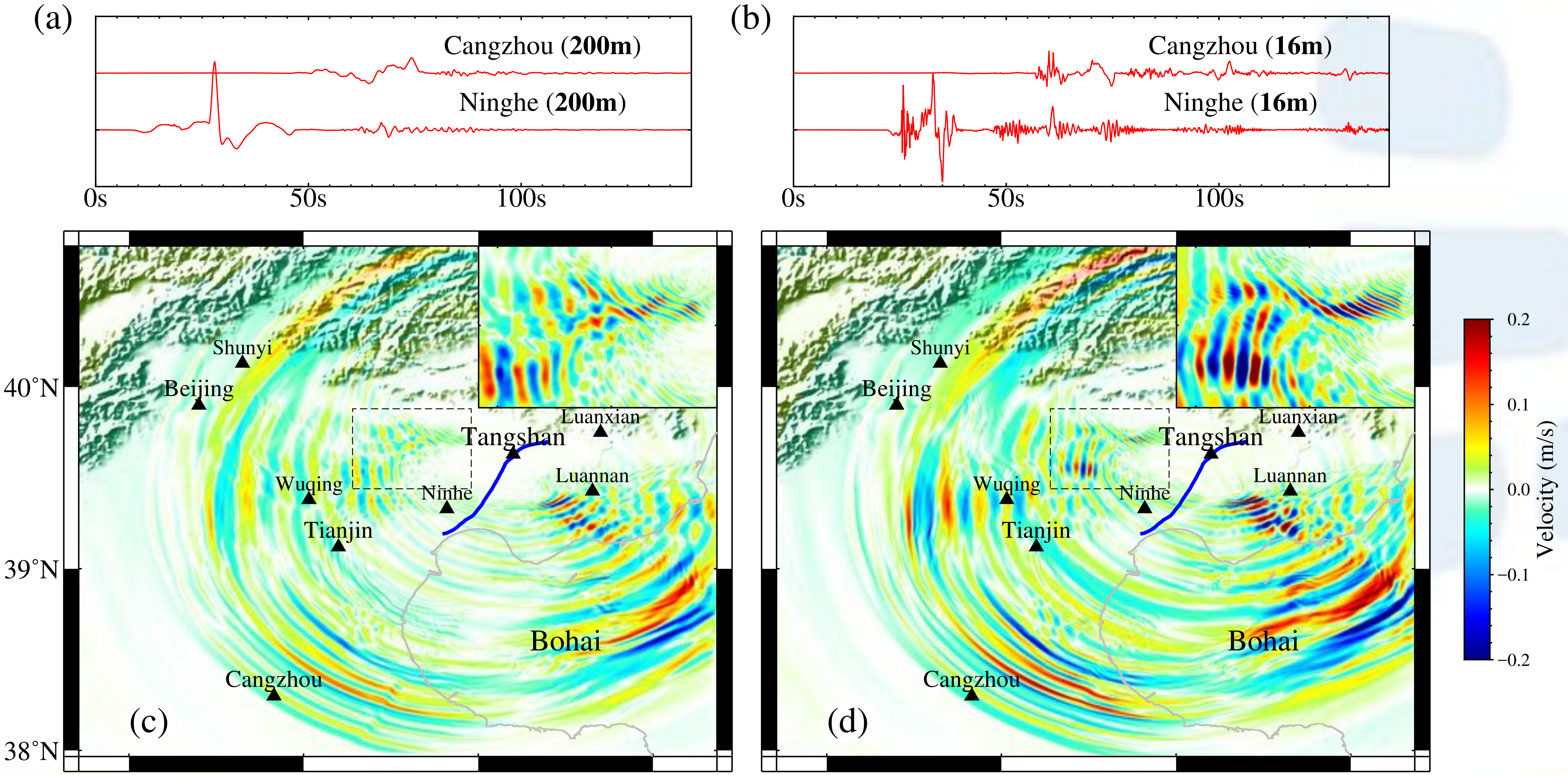


Strong Scaling

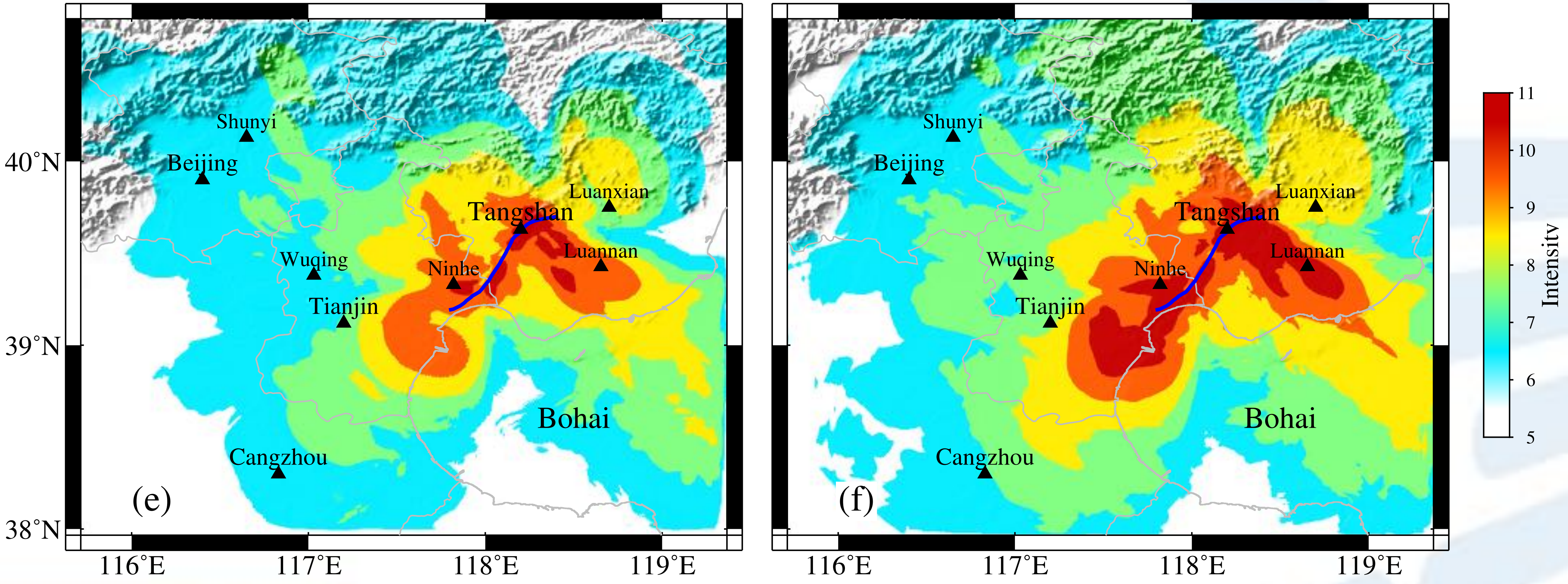
Simulation of the Tangshan Earthquake

- Tangshan earthquake, M7.2, 1976
- Simulation domain: 320 km x 312 km x 40 km

Simulation Results: 200m vs 16m



Simulation Results: 200m vs 16m



Outline

- Motivation and State of the Art
- Major Challenges
- Our Contributions
- Performance and Simulation Results
- Summary and Outlook

Nonlinear Earthquake Simulation on Sunway TaihuLight

- A complete framework with both dynamic rupture and seismic wave propagation modules
- An elaborate memory scheme that solves the memory constraint, and achieves a performance of 15.2 Pflops
  - a carefully designed DMA scheme with array fusion to coalesce the DMA operations
  - optimized blocking configuration guided by an analytic model
  - halo exchange through register communication
- On-the-fly compression to further improve the performance to 18.9 Pflops
- Future work
  - coupled simulation with mechanic model of buildings
  - generalization of the compression scheme for other scientific computing applications

Special Acknowledgements

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

- **Haohuan Fu**, Junfeng Liao, Jinzhe Yang, and et al., "The Sunway TaihuLight Supercomputer: system and applications", SCIENCE CHINA Information Sciences, 59.7 (2016): 072001.
- **Haohuan Fu**, Conghui He, Bingwei Chen, et al., "18.9-Pflops Nonlinear Earthquake Simulation on Sunway TaihuLight: Enabling Depiction of 18-Hz and 8-Meter Scenarios", in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC17), 12 pages, **ACM Gordon Bell Prize**, 2017.
- **Haohuan Fu**, Junfeng Liao, Nan Ding, et al., "Redesigning CAM-SE for Peta-Scale Climate Modeling Performance and Ultra-High Resolution on Sunway TaihuLight", in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC17), 12 pages, one out of the 3 **Gordon Bell finalists**, 2017.
- Chao Yang*, Wei Xue*, **Haohuan Fu***, et al., "10M-Core scalable fully-implicit solver for nonhydrostatic atmospheric dynamics", in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC16), **Gordon Bell Prize**, pp. 57-68, Salt Lake City, Utah, US, 2016.
- **Haohuan Fu**, Junfeng Liao, Wei Xue, Lanning Wang and et al., "Refactoring and Optimizing the Community Atmosphere Model (CAM) on the Sunway TaihuLight Supercomputer", in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC16), pp. 969-980, Salt Lake City, Utah, US, 2016
- Jiarui Fang, **Haohuan Fu**, Wenlai Zhao, Bingwei Chen, Weijie Zheng, and Guangwen Yang, "swDNN: A Library for Accelerating Deep Learning Applications on Sunway TaihuLight", in Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 615-624, May, 2017.

For more details, please refer to the above papers or contact haohuan@tsinghua.eud.cn.

THANK YOU