High-throughput simulations of two-phase flows on Blue Gene/Q

Panos Hadjidoukas

Chair for Computational Science
ETH Zurich

Symposium on Multi-system Application Extreme-scaling Imperative
Edinburgh, 3-4 September 2015

with: Diego Rossinelli, Fabian Wermelinger, Jonas Sukys, Ursula Rasthofer,
Christian Conti, Babak Hejazialhosseini, Petros Koumoutsakos

CSE**lab**
Computational Science & Engineering Laboratory
http://www.cse-lab.ethz.ch

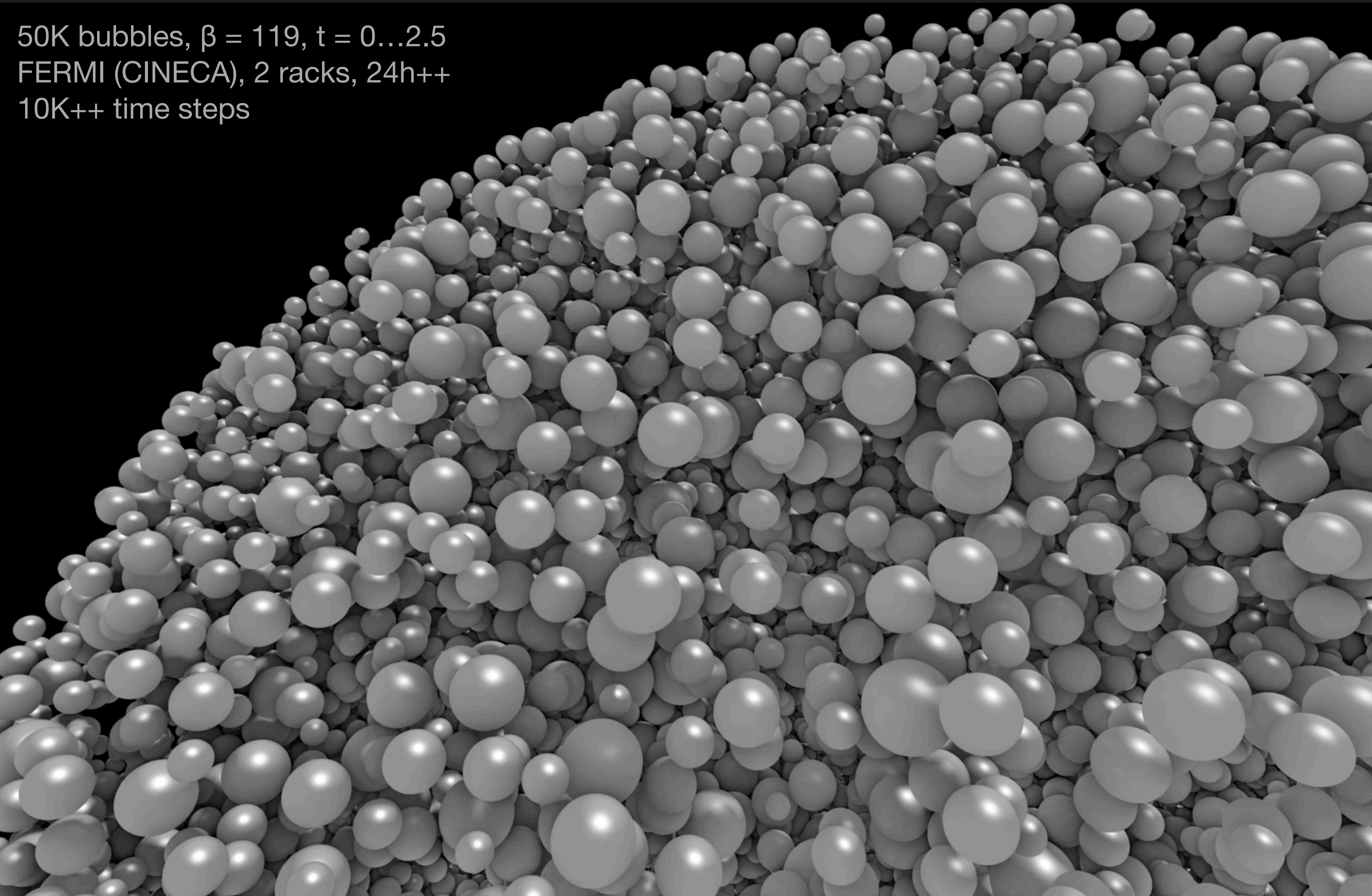
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Background

- CUBISM-MPCF: a finite volume, two phase flow solver at 14.4 PFLOP/s
    - 72% of the peak performance - 13 Trillion cells, Unprecedented time to solution
    - Bubble collapse simulations on 1.6M cores of Sequoia IBM BG/Q supercomputer
    - ACM Gordon Bell Prize 2013 (for peak performance)
- We will see how we managed to reach this performance

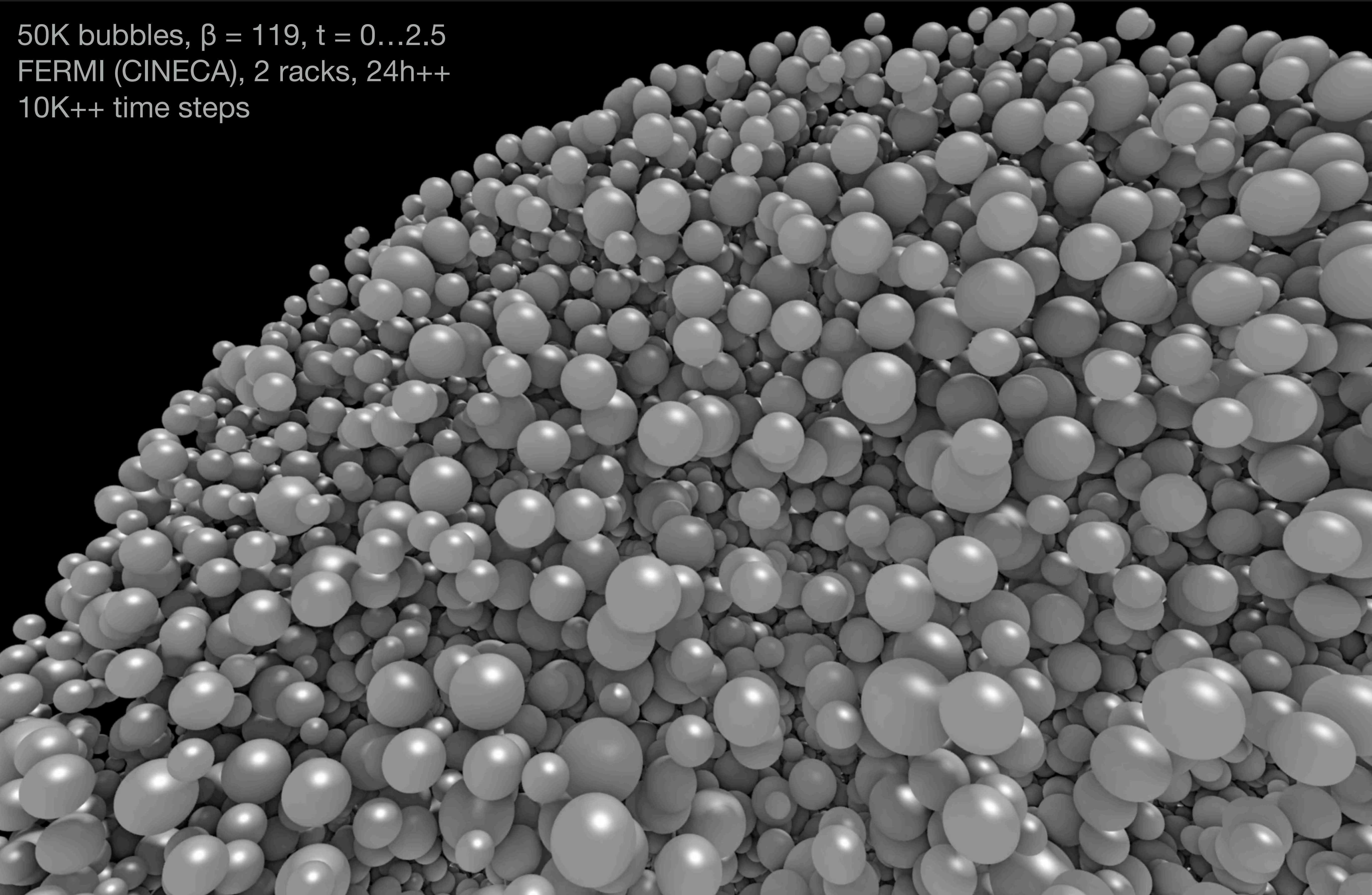
Cloud cavitation collapse

50K bubbles, β = 119, t = 0…2.5
FERMI (CINECA), 2 racks, 24h++
10K++ time steps



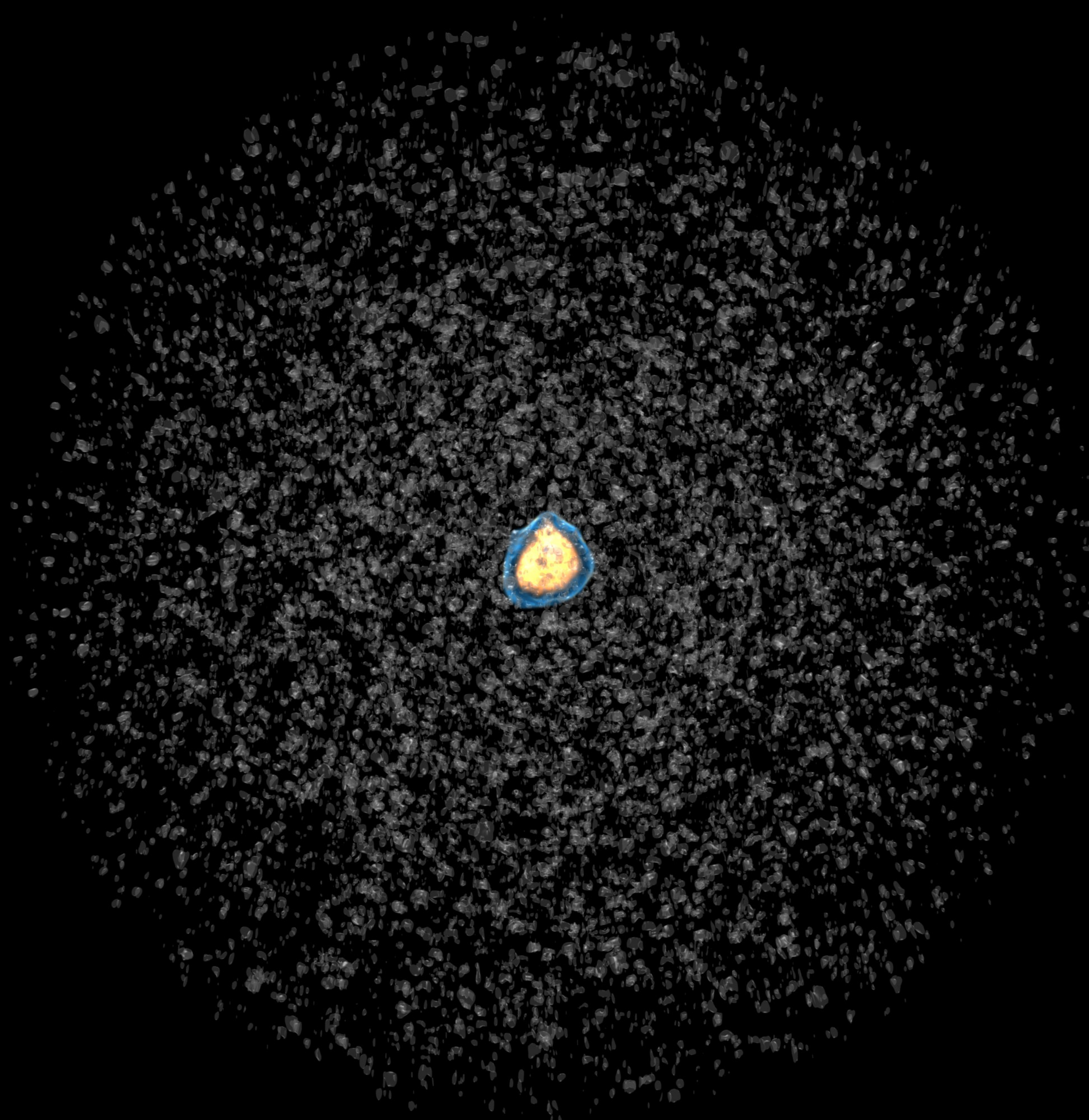
Cloud cavitation collapse

50K bubbles, β = 119, t = 0…2.5
FERMI (CINECA), 2 racks, 24h++
10K++ time steps



Initial cloud

Final stage of the collapse



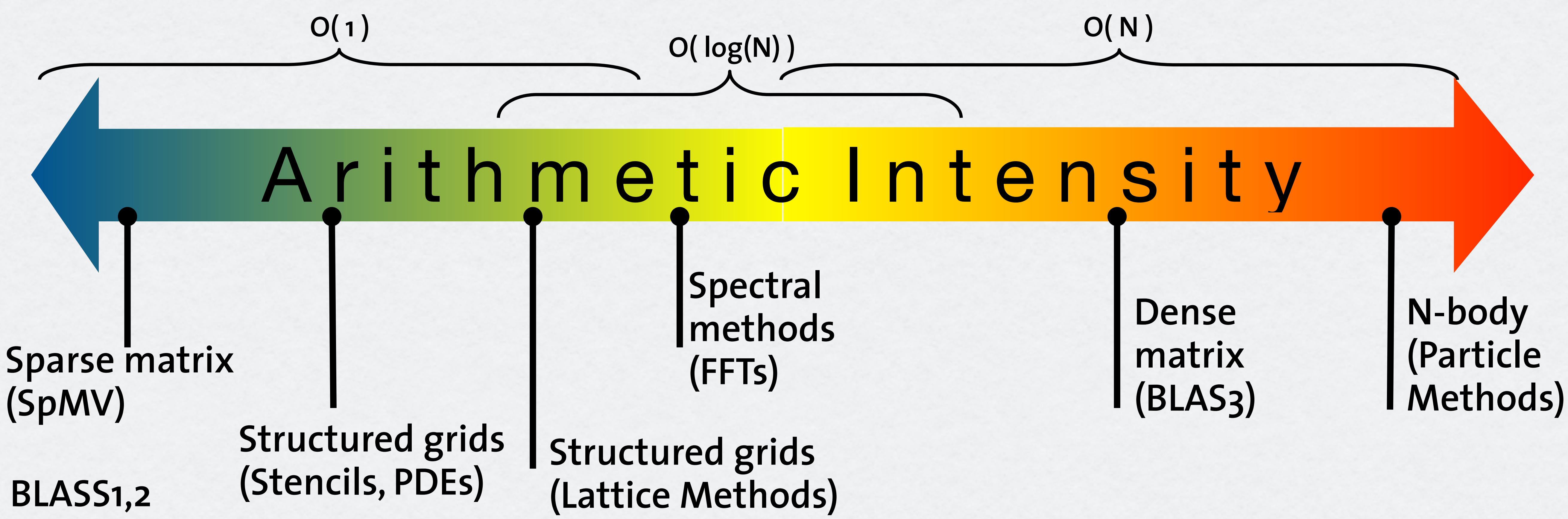
Setting the state of the art

- Peak Performance
  - 14.4 PFLOPS, 72% of peak
  - 1.3%-6.4% (Stanford)
- Time to solution
  - Tw = 1.8
  - Tw = 16.3 - 39.0 (Stanford)
- Computational Elements
  - 13.2 Trillion - 15k Bubbles
  - 0.4 Trillion - Turbulence (Stanford)

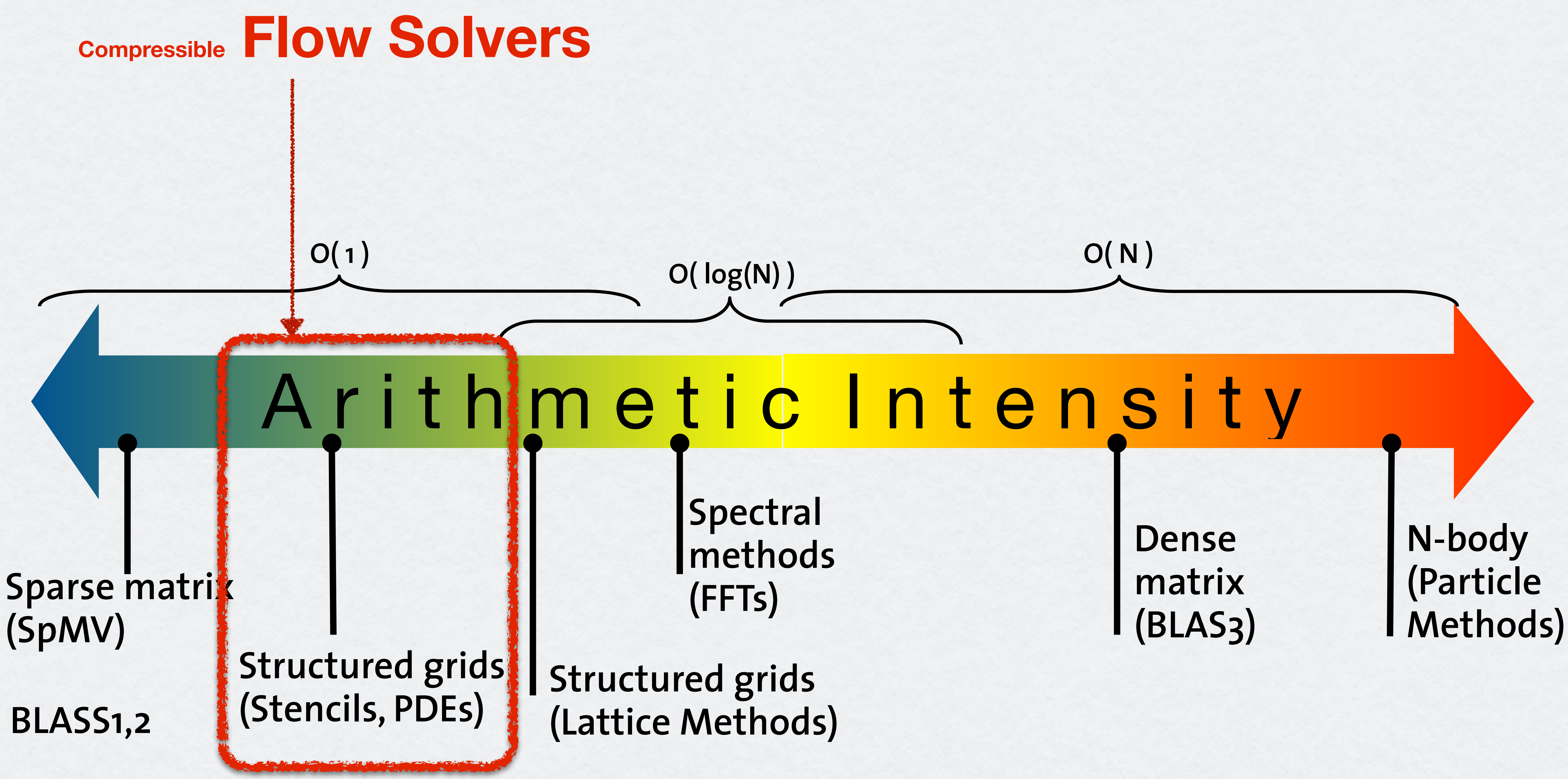
$$T_w = \Delta^{wt} * \frac{N_c}{N_p}$$ (Stanford paper)

(*) Stanford: SC'13, Bermejo-Moreno et al, Compressible Homogeneous Turbulence

Roofline and the 7 dwarfs

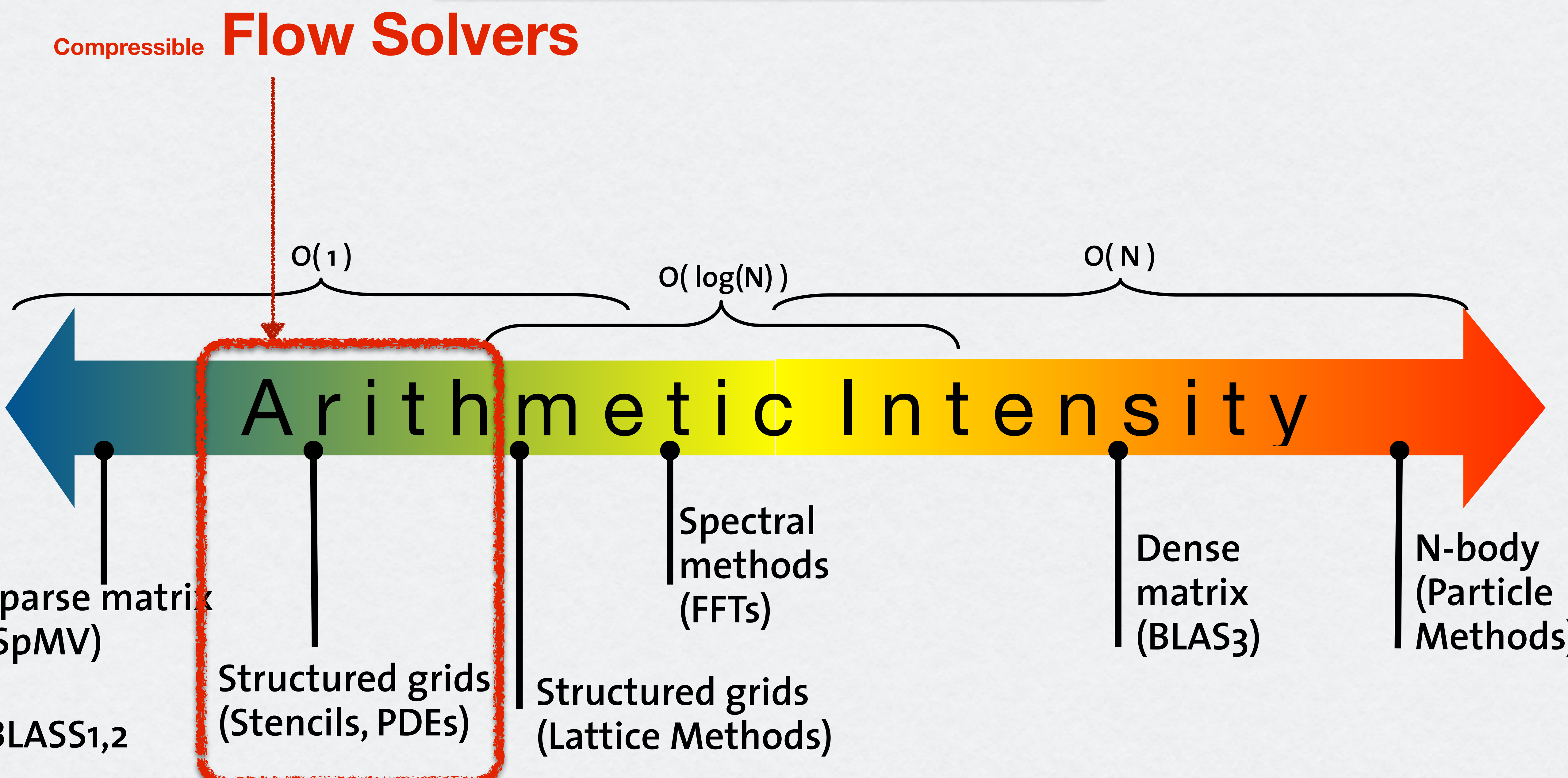


Roofline and the 7 dwarfs



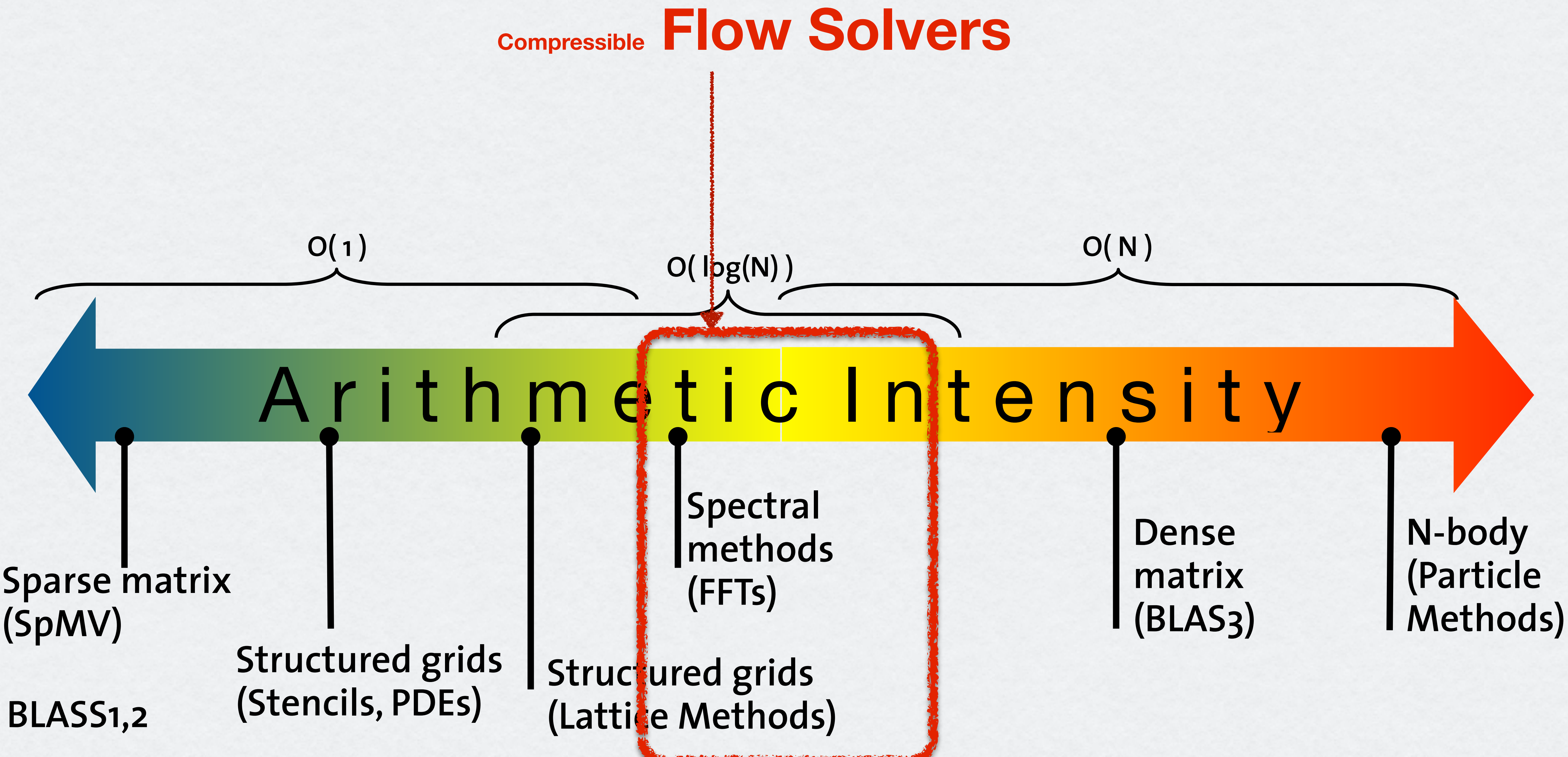
Roofline and the 7 dwarfs

- **ALGORITHMS & DATA STRUCTURES**
  - Operational Intensity (FLOP/Byte ratio)
  - FLOP/Instruction density



Roofline and the 7 dwarfs

- **ALGORITHMS & DATA STRUCTURES**
  - Operational Intensity (FLOP/Byte ratio)
  - FLOP/Instruction density



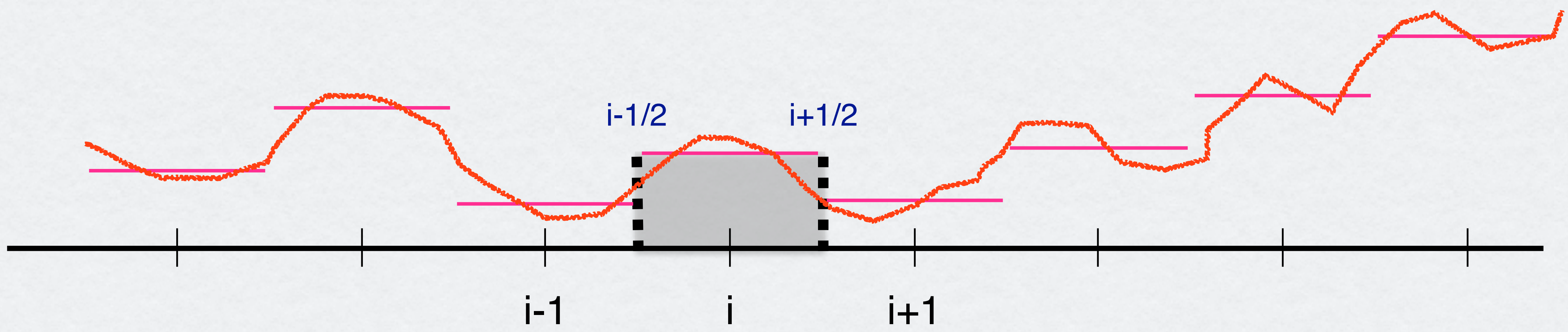
Numerics: Finite Volume Method

Reconstructions and Fluxes

Finite Volume and Conservation Laws

$$\int_V \frac{\partial \rho}{\partial t} dV = -(\rho_2 u_2 A_2 - \rho_1 u_1 A_1)$$

$$\tilde{U}_i^{n+1} = \tilde{U}_i^n + \frac{\delta t}{\delta x}\left[F(U^n_{i-\frac{1}{2}}) - F(U^n_{i+\frac{1}{2}})\right]$$



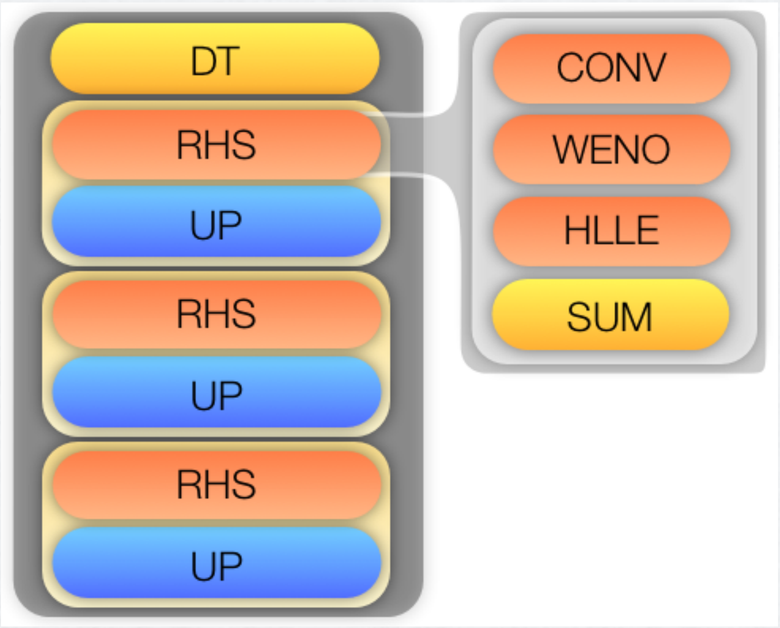
STEP 1 : Reconstruct $\mathrm{U}^n_{i-1/2}$, $\mathrm{U}^n_{i+1/2}$ from cell averages $\tilde{U}_i^n$ → **WENO5**

STEP 2: Solve the Riemann problem at cell interfaces $i \pm 1/2$. Get $F^n_{i\pm1/2}$ → **HLLE**

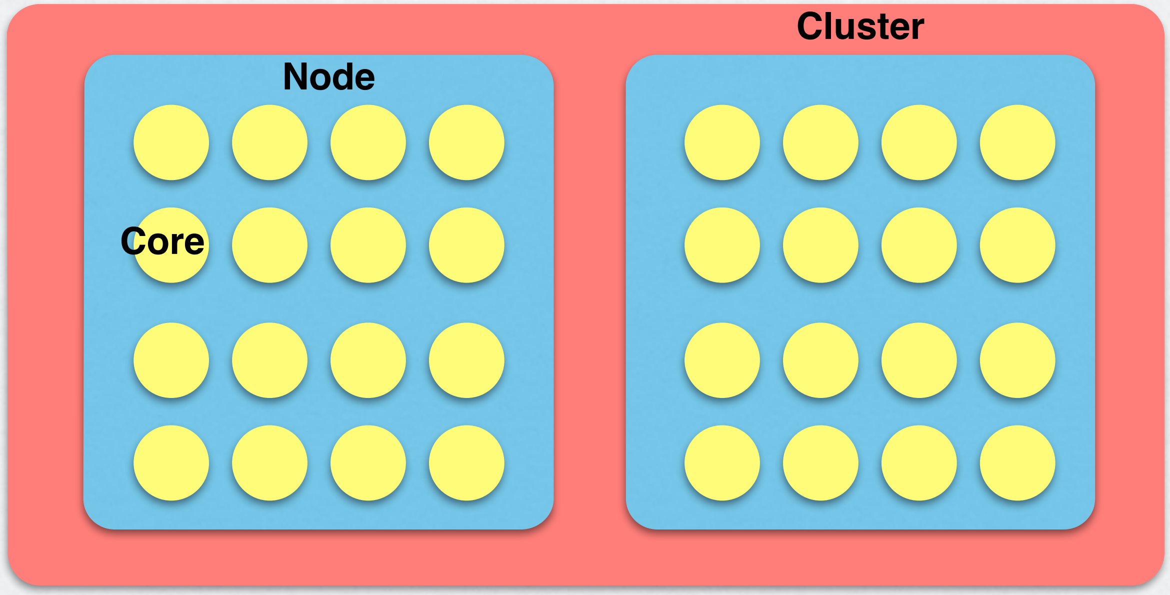
STEP 3: Update Cell Averages : $\tilde{U_i}^{n+1}$ → **LS-RK3**

Software: CUBISM-MPCF

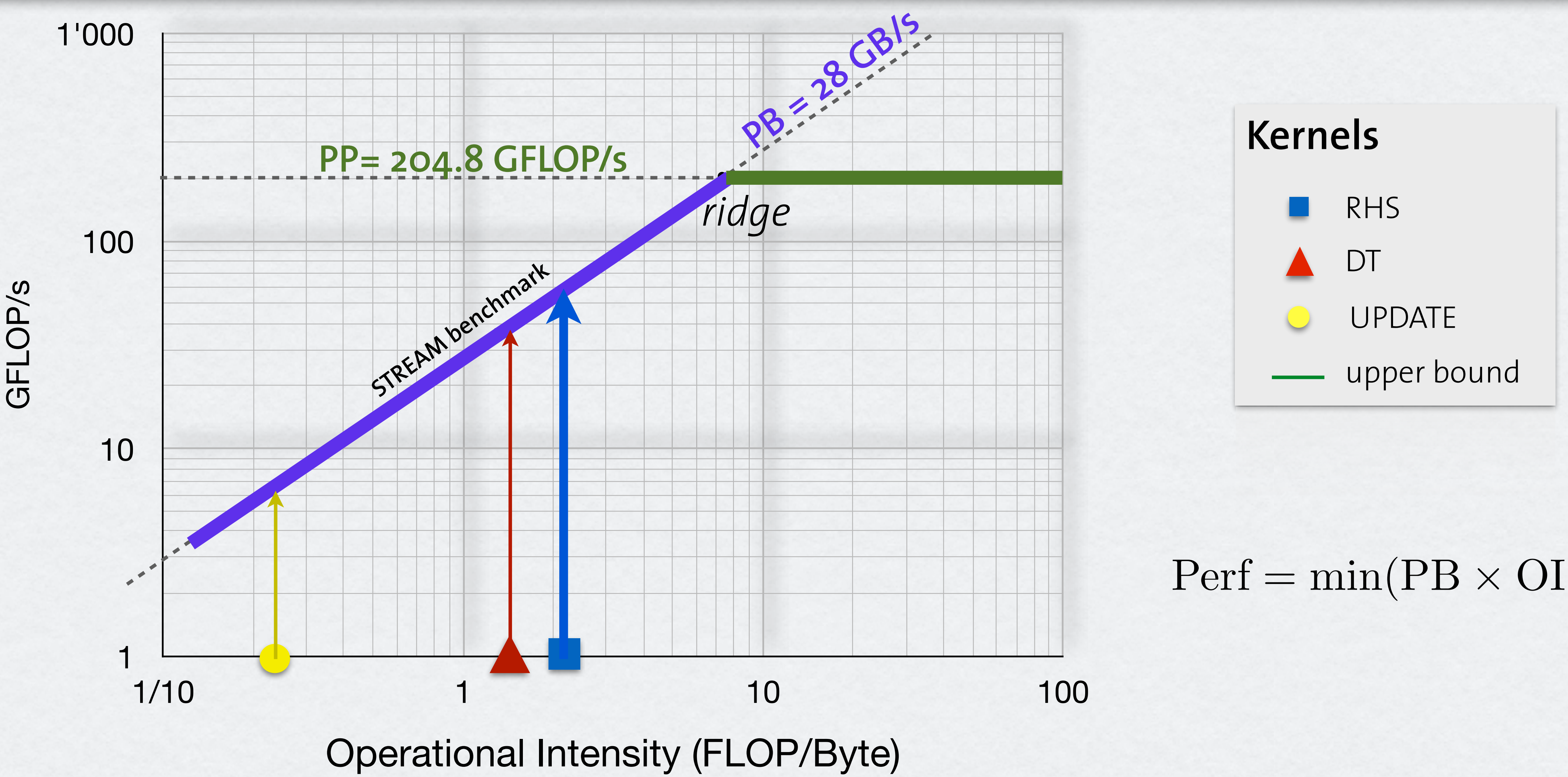
- Compute kernels:
  - DT
  - RHS (90%)
    - WENO (83%)
    - HLLE (13%)
  - UPDATE



- Software Layout
  - Core (SIMD)
  - Node (OpenMP)
  - Cluster (MPI)



The Roofline of BG/Q

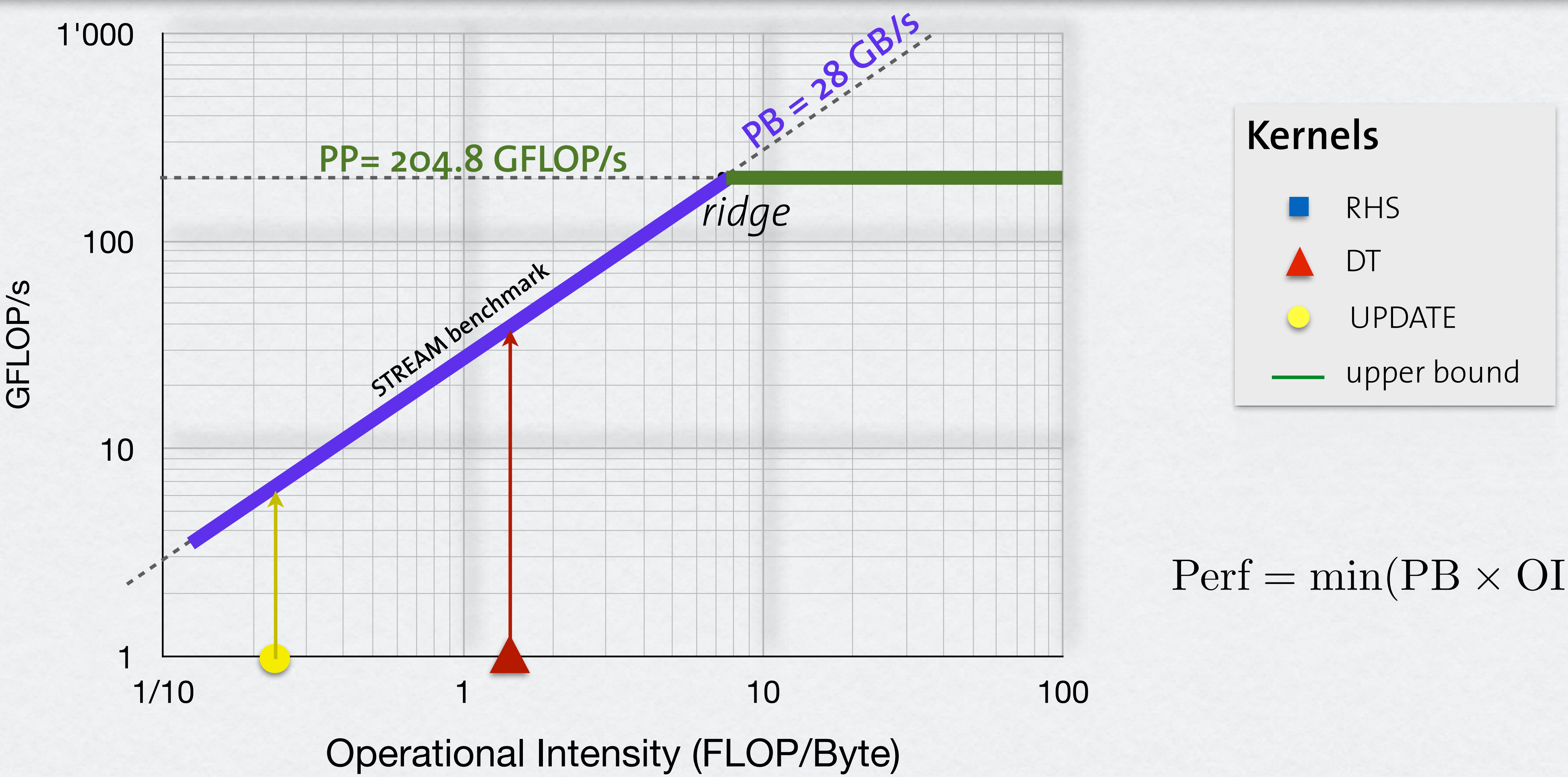


$$\text{Perf} = \min(\text{PB} \times \text{OI},\ \text{PP})$$

Operational Intensity: FLOP count over off-chip memory transfer

BG/Q **node** ridge point: (7.3 FLOP/Byte, 204.8 GFLOP/s)

The Roofline of BG/Q

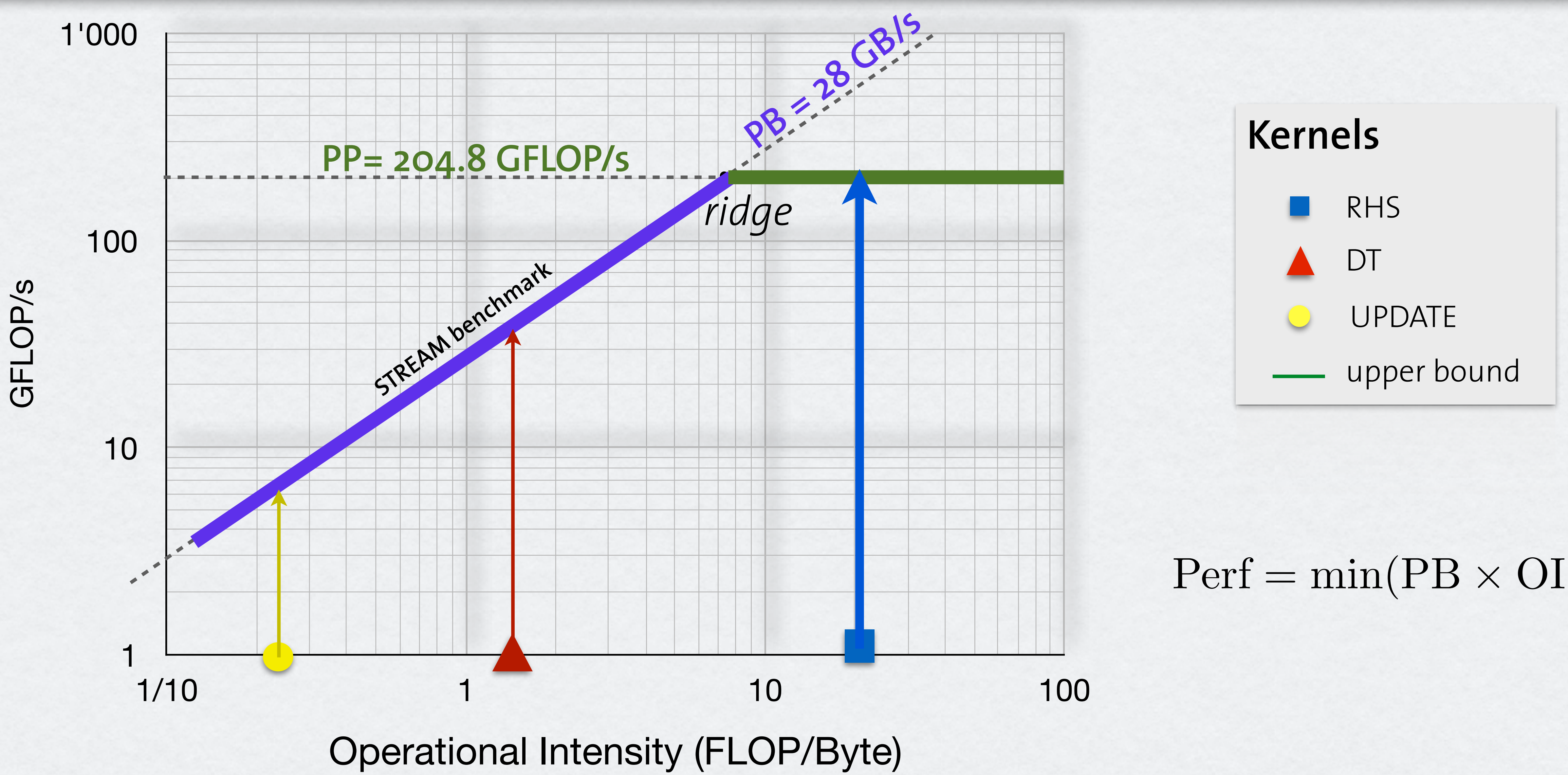


$$\mathrm{Perf} = \min(\mathrm{PB} \times \mathrm{OI},\ \mathrm{PP})$$

Operational Intensity: FLOP count over off-chip memory transfer

BG/Q **node** ridge point: (7.3 FLOP/Byte, 204.8 GFLOP/s)

The Roofline of BG/Q



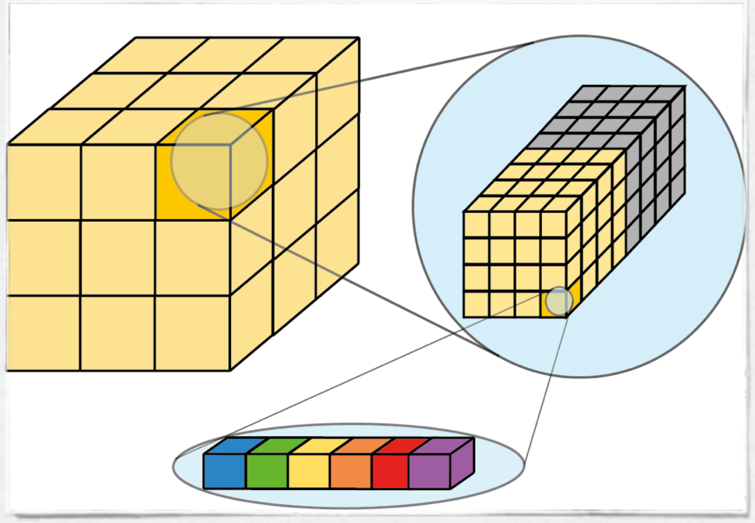
$$\mathrm{Perf} = \min(\mathrm{PB} \times \mathrm{OI},\ \mathrm{PP})$$

Operational Intensity: FLOP count over off-chip memory transfer

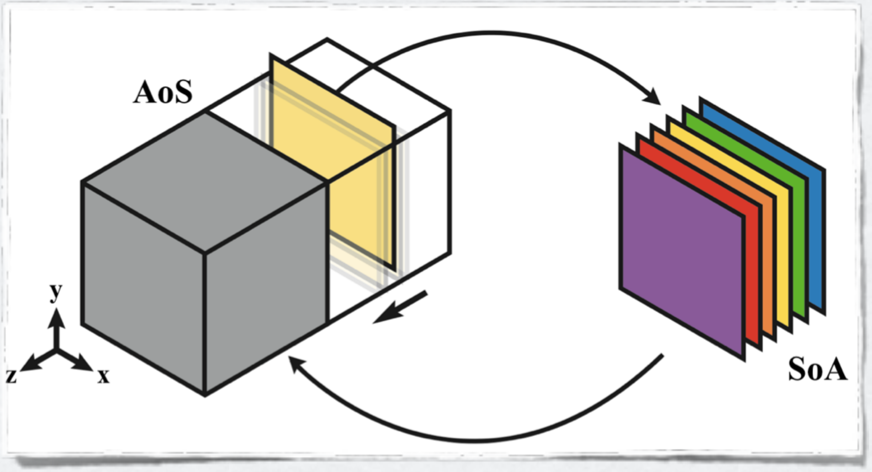
BG/Q **node** ridge point: (7.3 FLOP/Byte, 204.8 GFLOP/s)

Core layer

- Block-based memory layout
  - Increases spatial locality

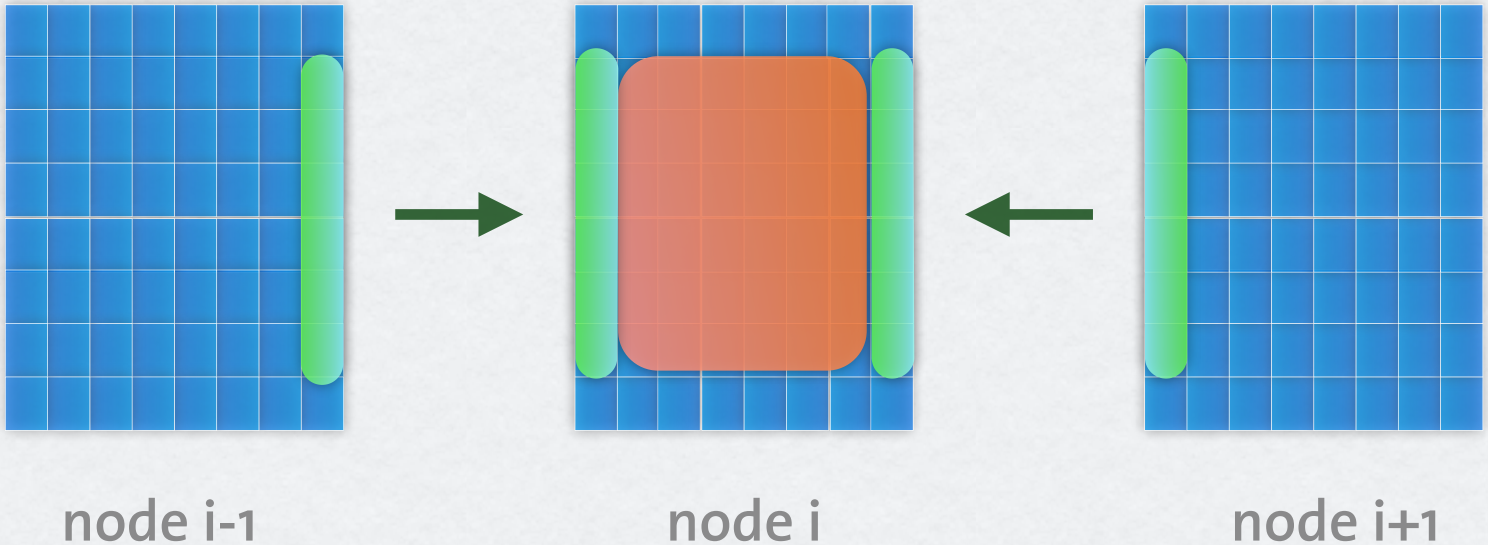


- IL and DL parallelism
  - 1 thread exclusively process 1 block
  - SoA -> explicit vectorization of all kernels
  - Identify common subexpressions in RHS
- Temporal locality
  - Buffers for active data-slices
    - e.g. in WENO, HLLE
  - Fusion of the RHS substages



Node and Cluster layers

- OpenMP parallelization - 64 threads
  - Depth-first thread placement
  - Dynamic loop scheduling
  - Direct memory reuse (no allocations)



- MPI parallelization:
  - Non-blocking P2P communication for halo blocks
  - 6 messages to neighbor ranks, size: 3-30MB
  - Communication Time ~ O(Time for processing 1-2 blocks)

BG/Q features

- Asynchronous progress communication
  - efficient C/T overlap
- QPX instructions
  - expose as many FMAs as possible (through code fusion)
  - vec_madd(a, b, c) = a*b+c
- Data (L1P) prefetching
  - confirmed policy with depth 1 (or 2) but not the default (3)
  - better cache utilization
- Block memory copy (builtin __bcopy)
  - faster ghost reconstruction, data packing and boundary conditions

Accuracy and more

- Numerical accuracy of reciprocal and divisions
  - Newton-Raphson scheme with one or two iterations
  - vec_swdiv(): uses two iterations
  - Square roots: custom implementation or vec_swsqrt()
- Other options
  - Loop unrolling
    - register spilling and unnecessary load/store instructions
    - #pragma unroll (1) around WENO
  - Careful selection of compilation flags
  - Stack size of OpenMP threads

Performance

- **RHS**: **14.4 PFLOP/s**, **72%** of peak
- **OVERALL**: **12.1 PFLOP/s, 65%** of peak

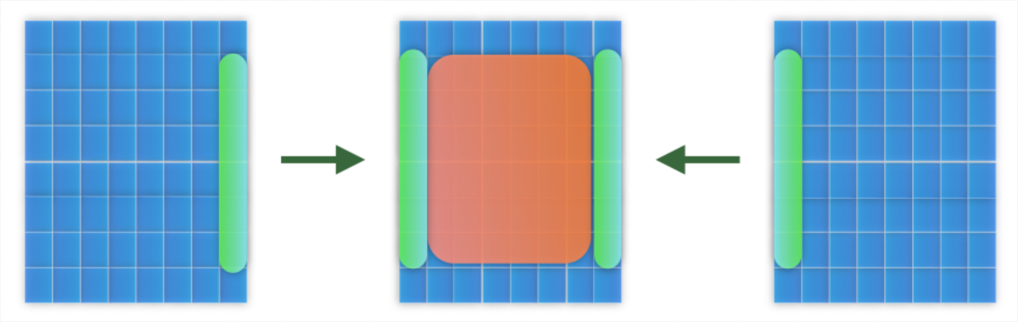
KERNEL	Node	Sequoia	Reason
RHS	**72.3%**	**71.8%**	**efficient C/T overlap**
DT	**19.9%**	**13.2%**	**global reduction (MPI_Allreduce)**
UPDATE	**2.3%**	**2.3%**	**local operations**
ALL	**65.5%**	**64.8%**	

- 1 rank / node, 64 OpenMP threads
- 16^3=4096 blocks, 32^3 grid points per block
- Accuracy with 2 Newton-Raphson iterations (NR=2)

Profiling: HPM library by Dr. Bob Walkup (IBM)

C/T overlap

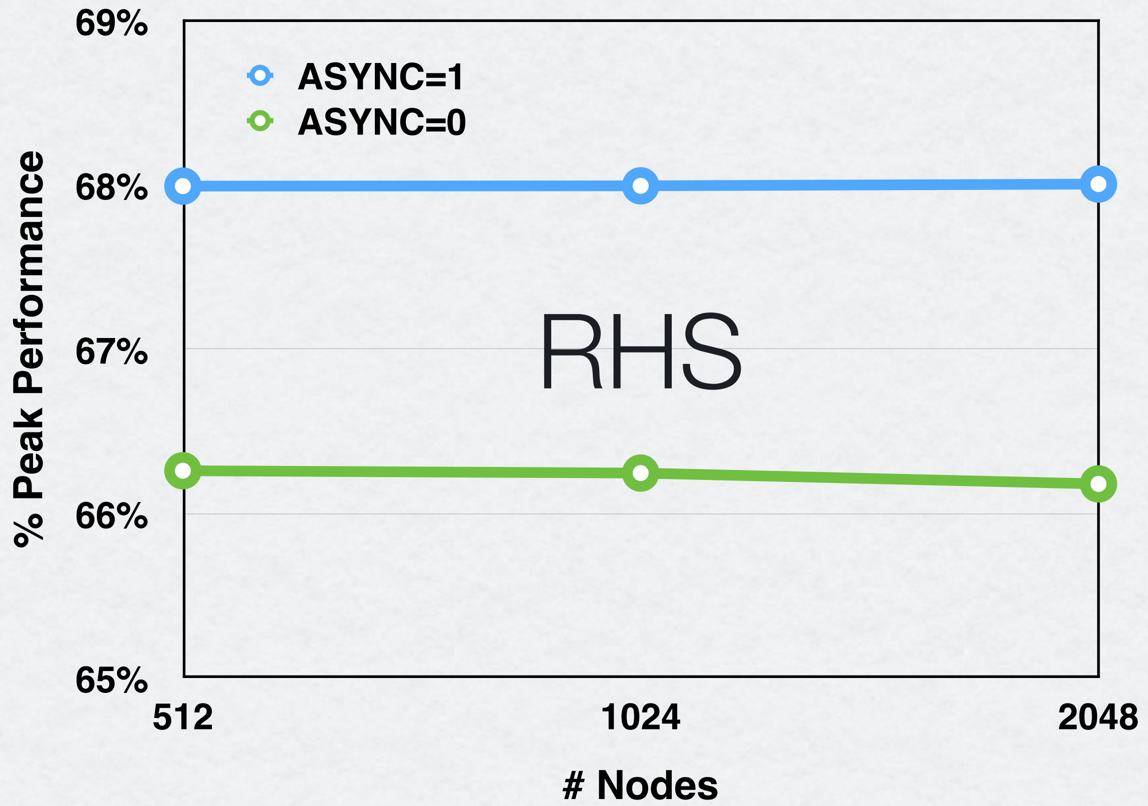
- Issue: time of **MPI_Waitall()** was not negligible!
  - Observed when the number of compute nodes increases
- Solution: asynchronous progress communication at the PAMID layer of the BGQ MPI implementation
  - One communication thread used for MPI asynchronous progress

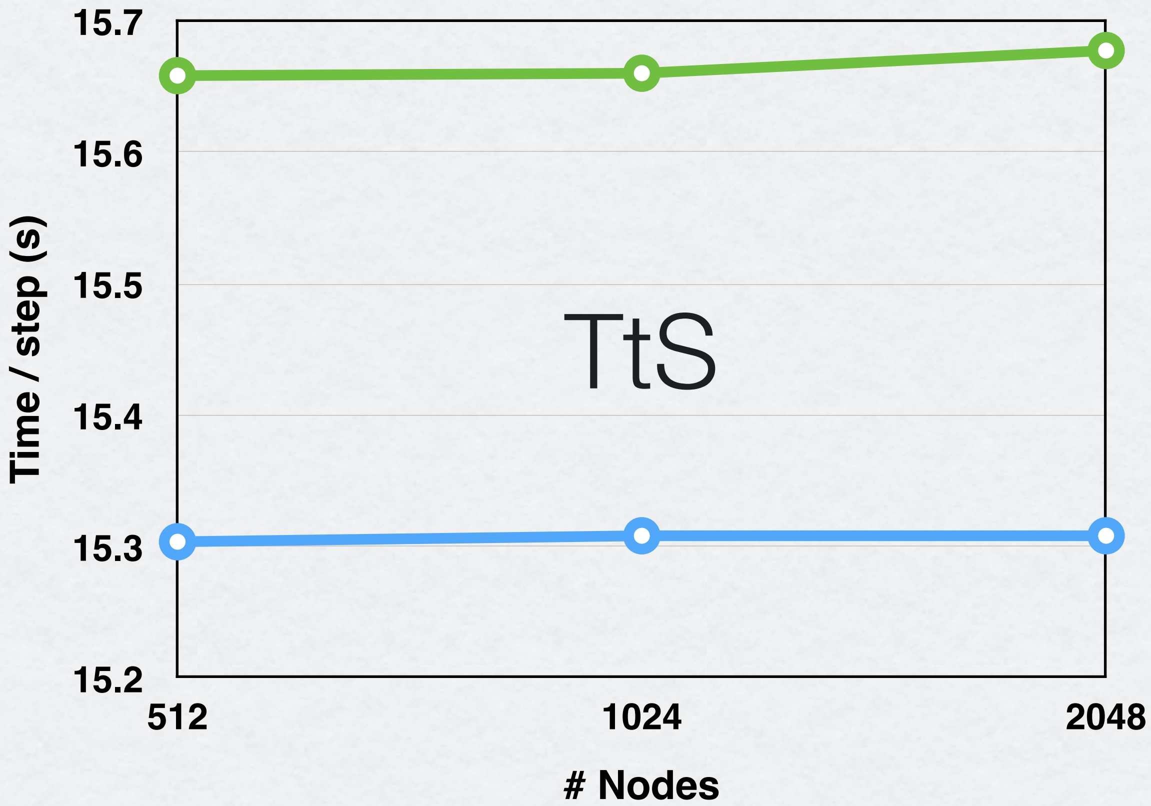


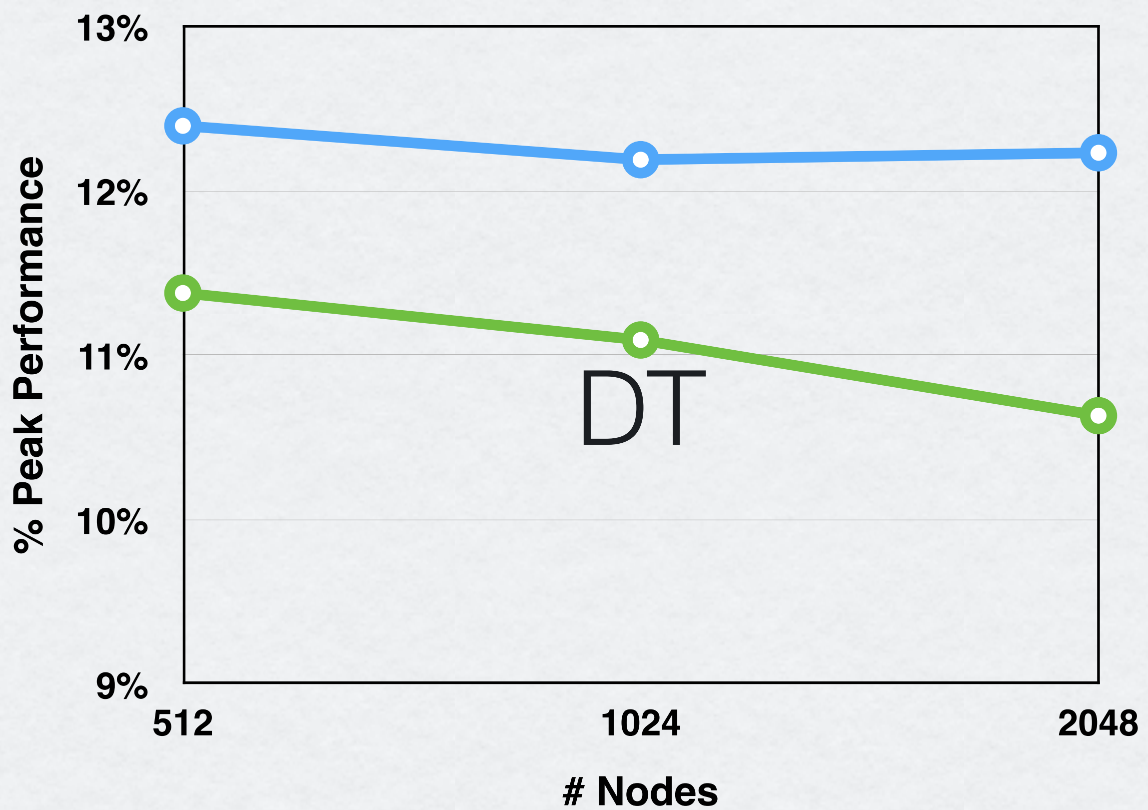
- How it works:

1. The main thread issues the necessary **MPI_Irecv/Isend** calls
2. OpenMP parallel region with 63 threads: 1 thread - 1 inner block (or more)
3. After this parallel region, the main thread calls **MPI_Waitall()**.
4. OpenMP parallel region with 64 threads: rest of the inner blocks + halo blocks processed using an OpenMP for loop.

C/T overlap (NR=1)





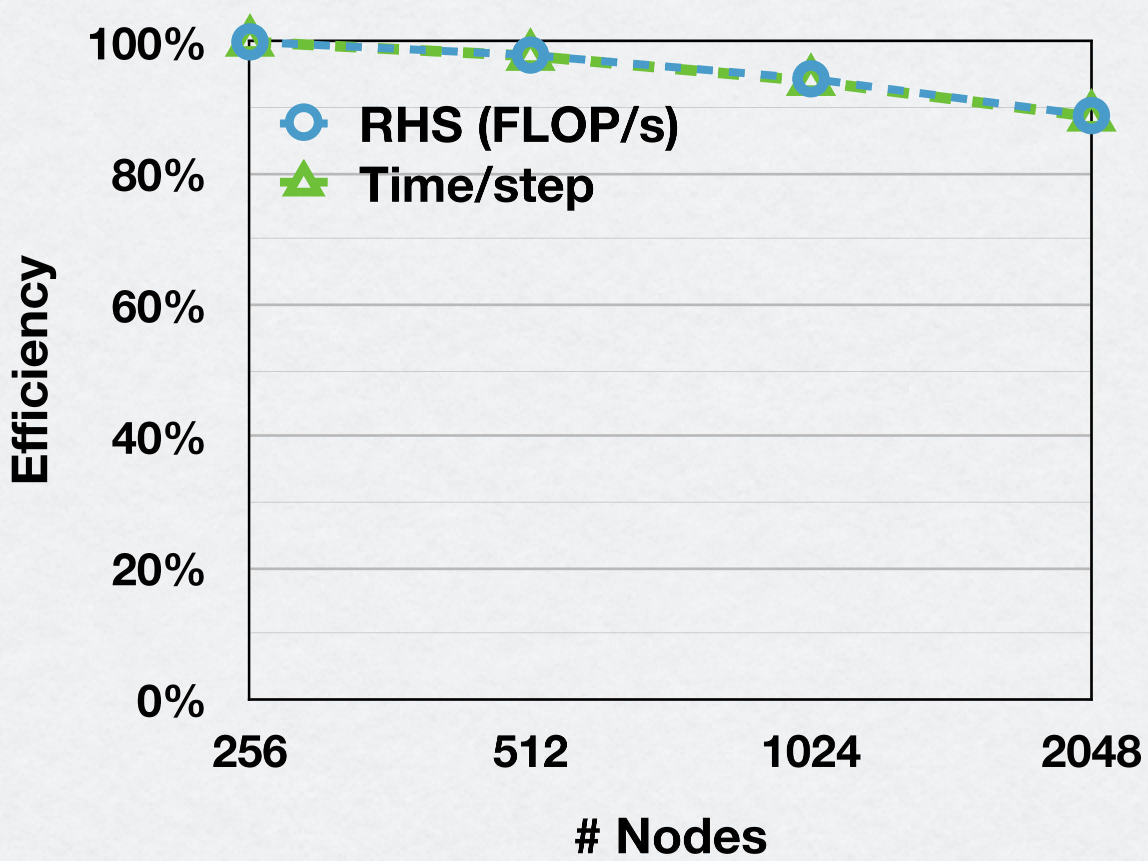


Performance degradation in very initial version (submission) without asynchronous comm. progress

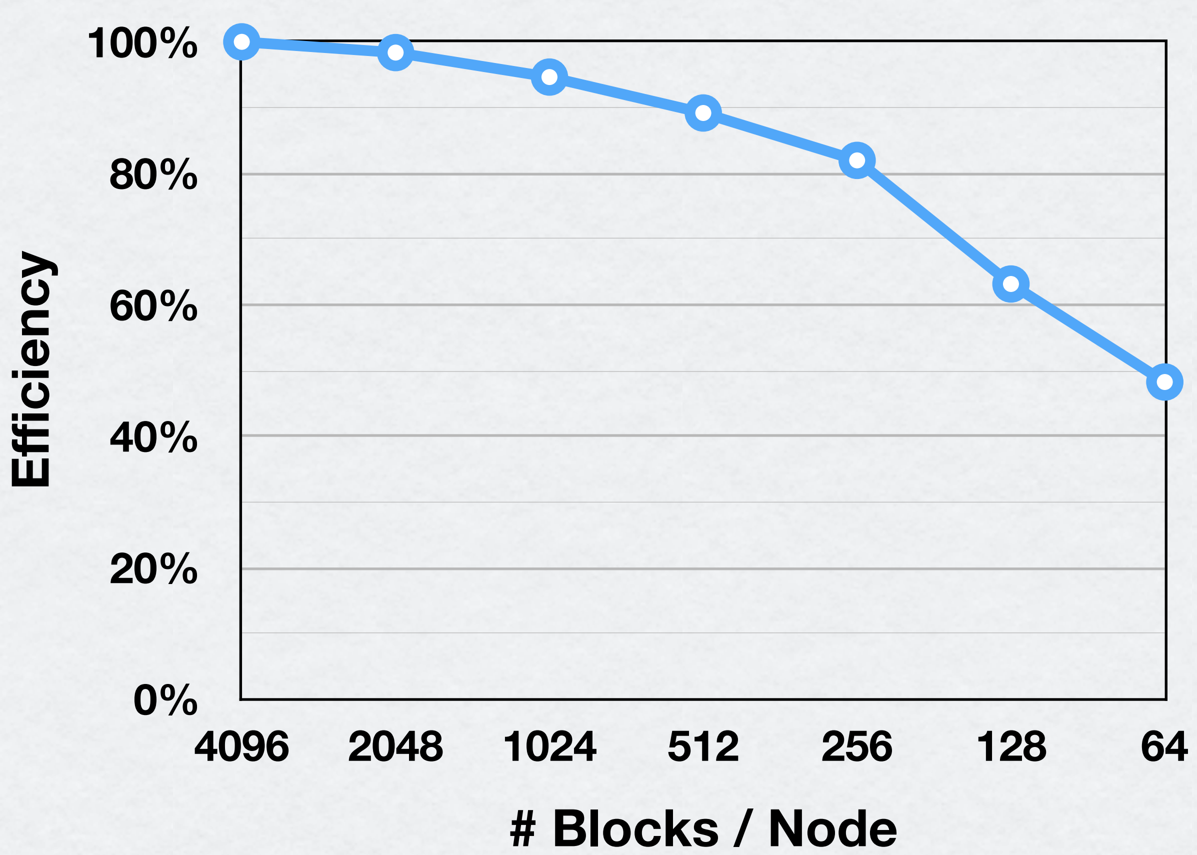
- 1 rack: 60%
- 24 racks: 58%
- 96 racks: 55%

Strong scaling (NR=1)

4096 blocks

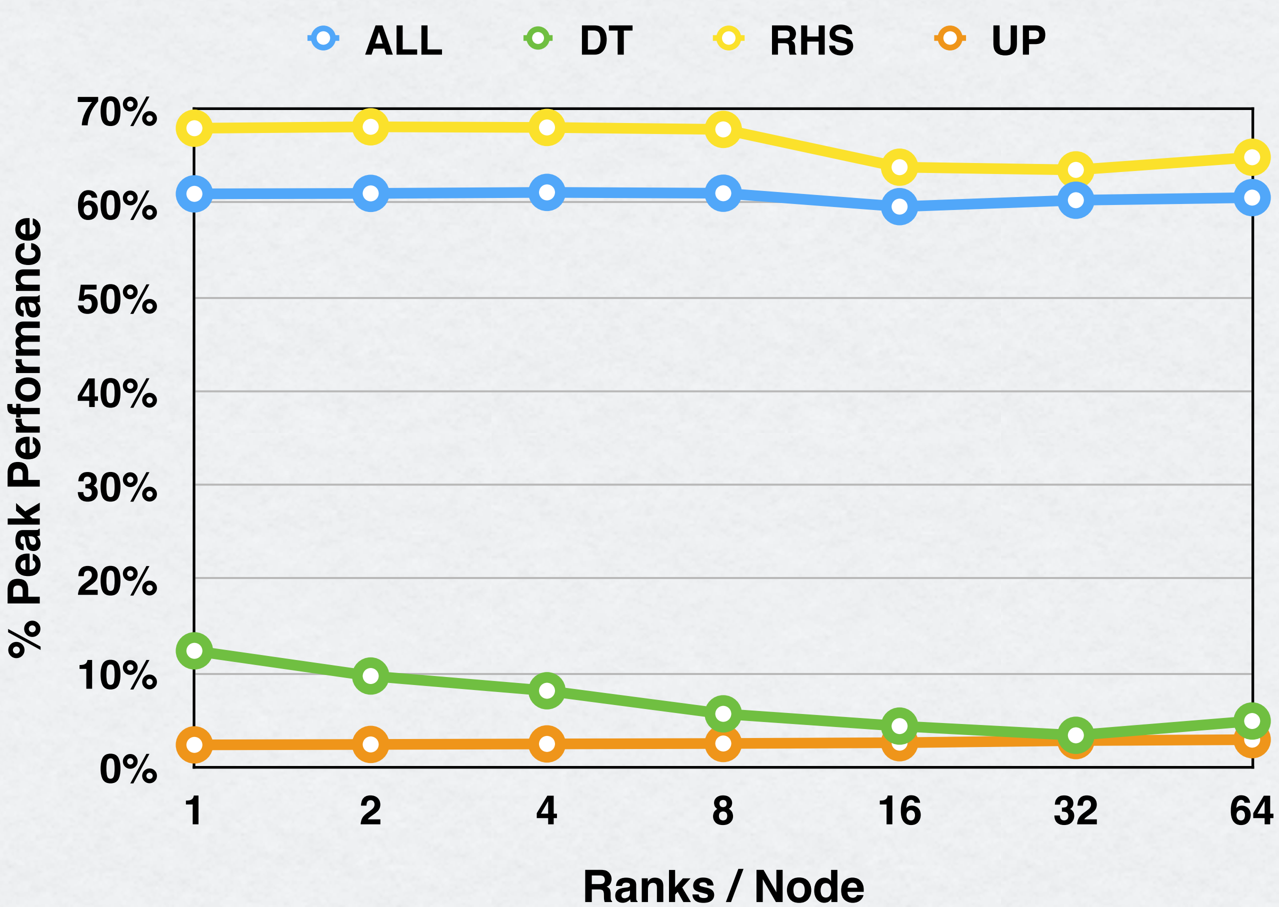


512 nodes

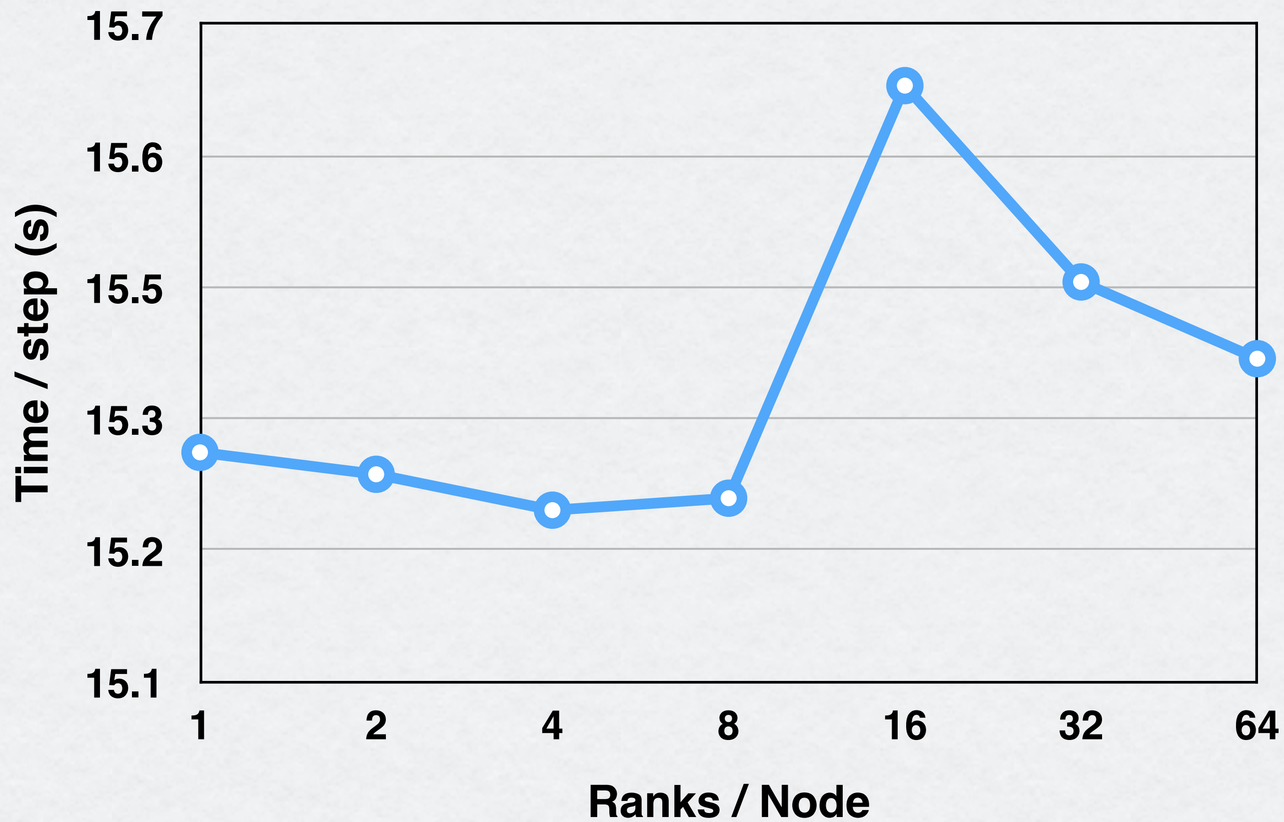


Multiple ranks per node (NR=1)

1024 nodes



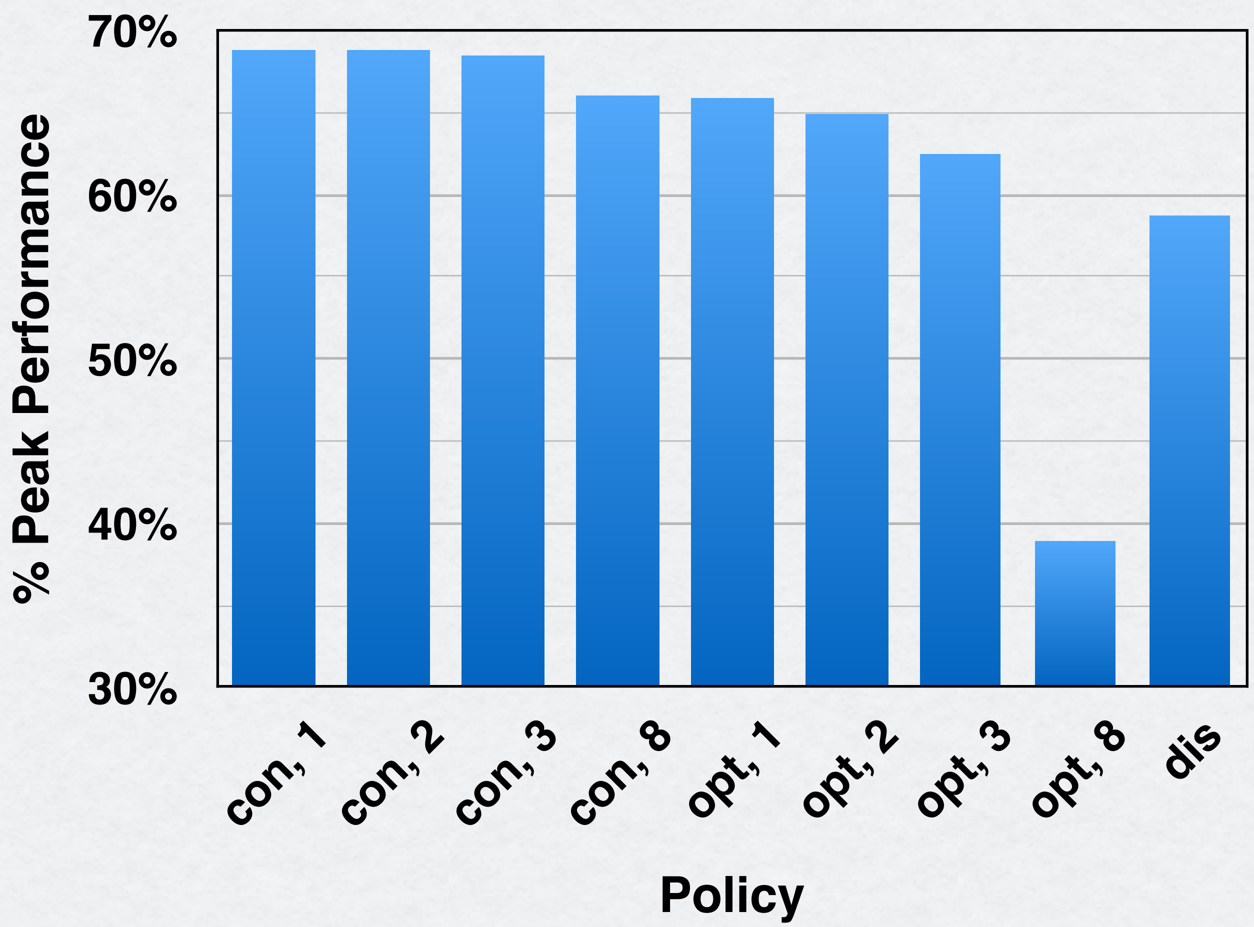
Peak Performance (RHS)



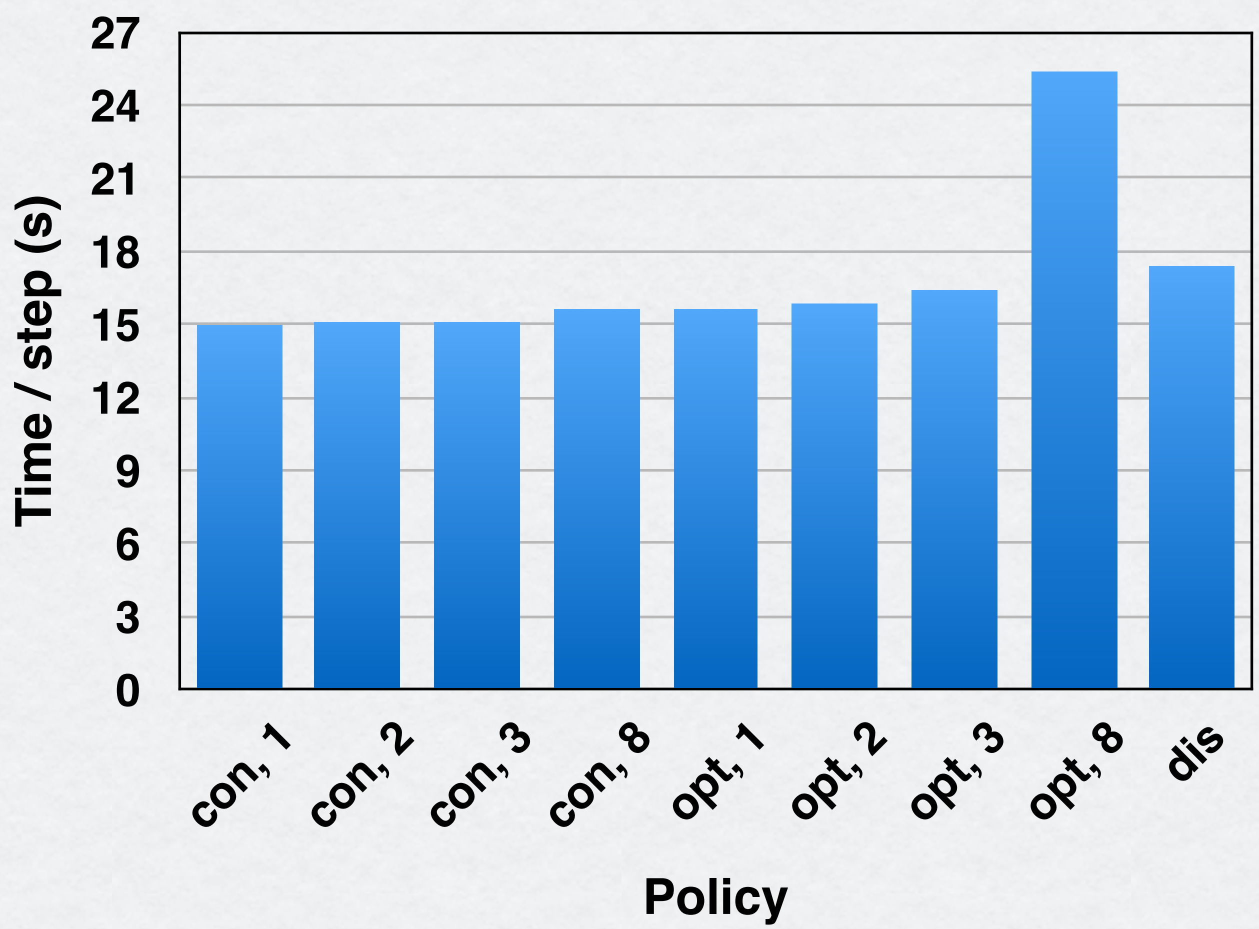
Time-to-Solution

Data prefetching

1 node, NR=1



Peak Performance (RHS)

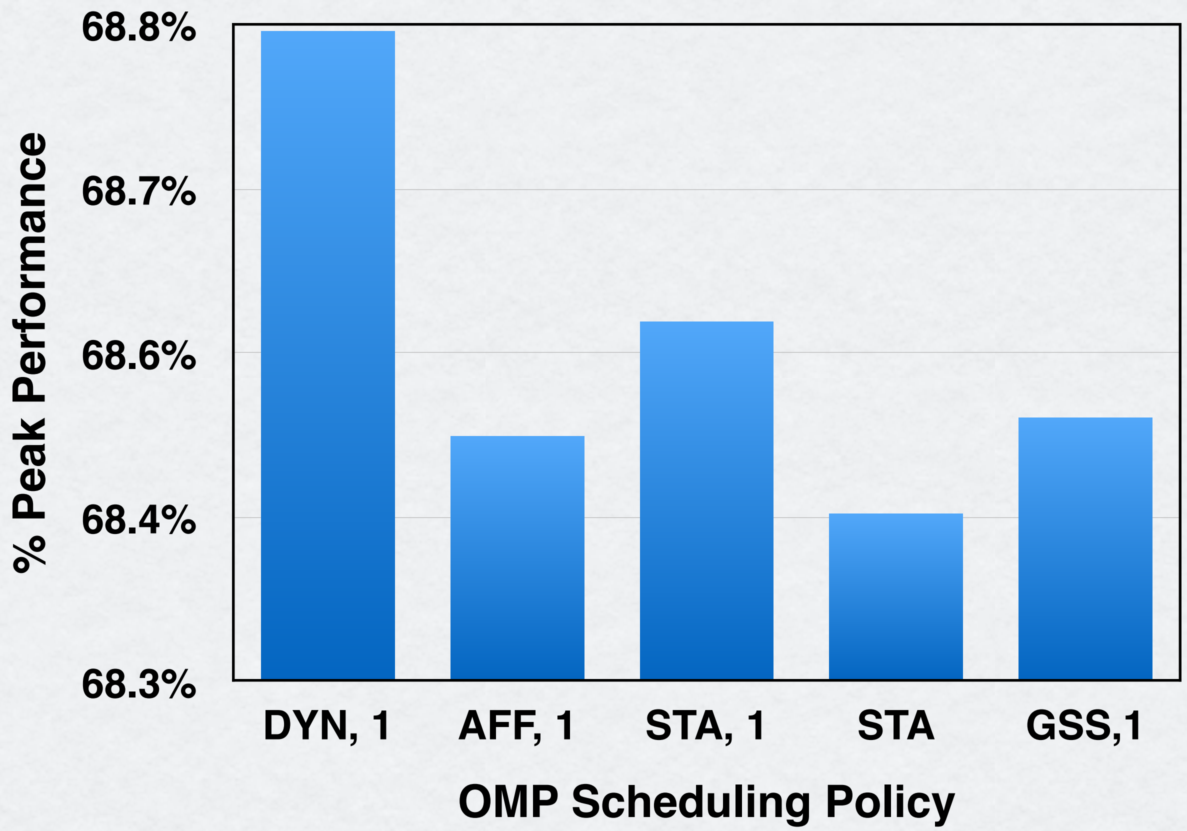


Time-to-Solution

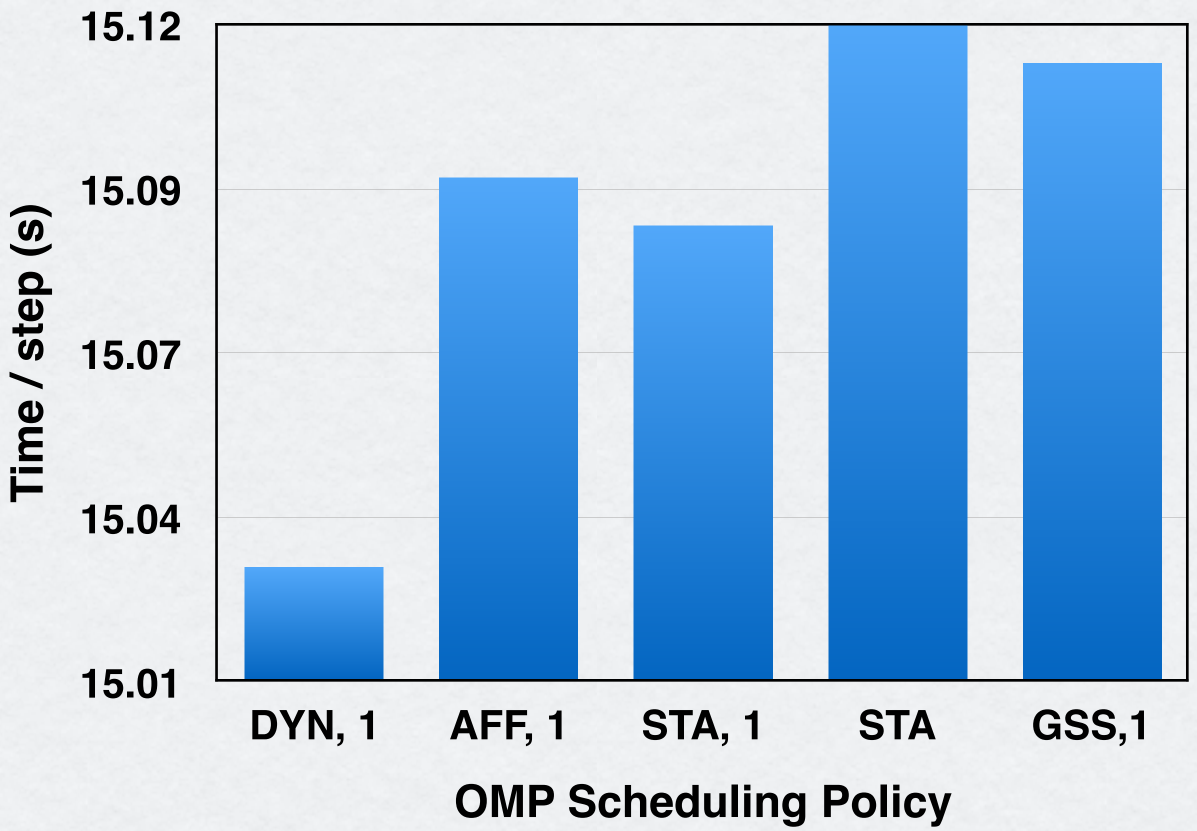
con: L1P_stream_confirmed
opt: L1P_stream_optimistic
dis: L1P_stream_disable

Loop scheduling

1 node, NR=1

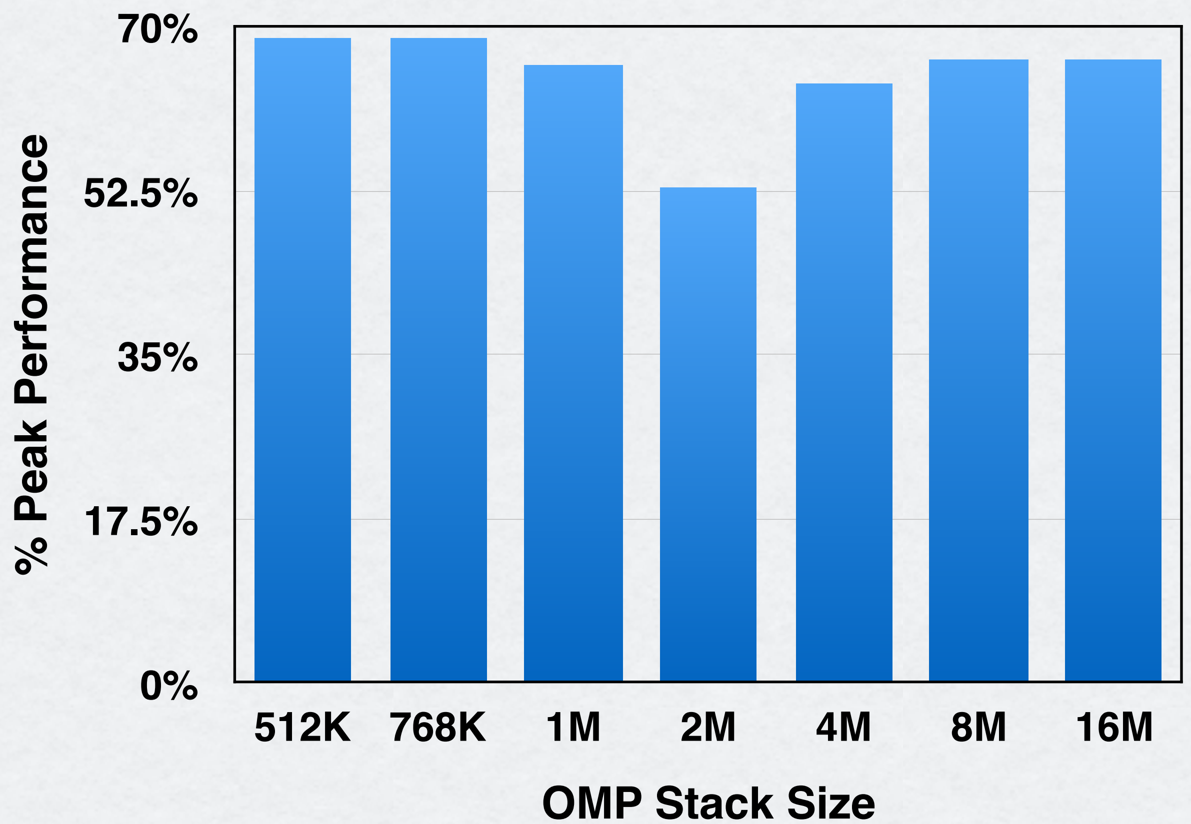


Peak Performance (RHS)

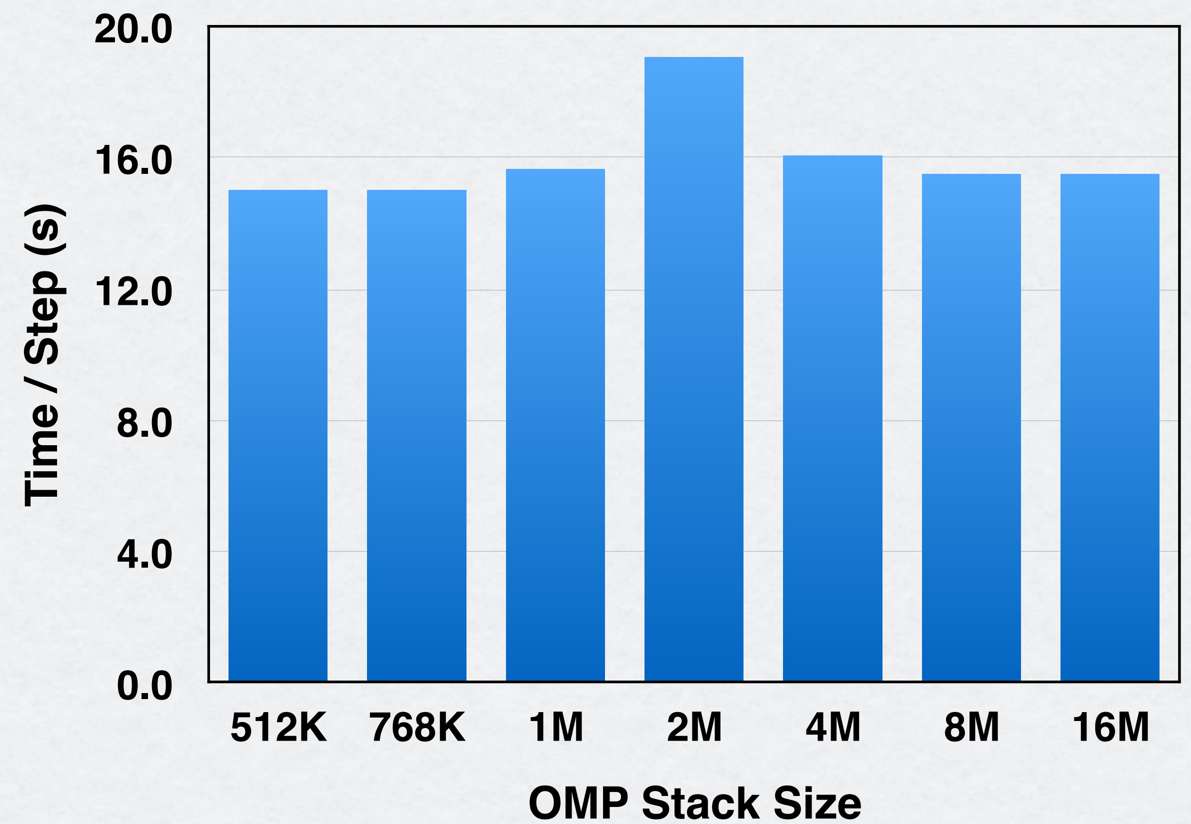


Time-to-Solution

Stack size / data alignment (NR=1)



Peak Performance (RHS)



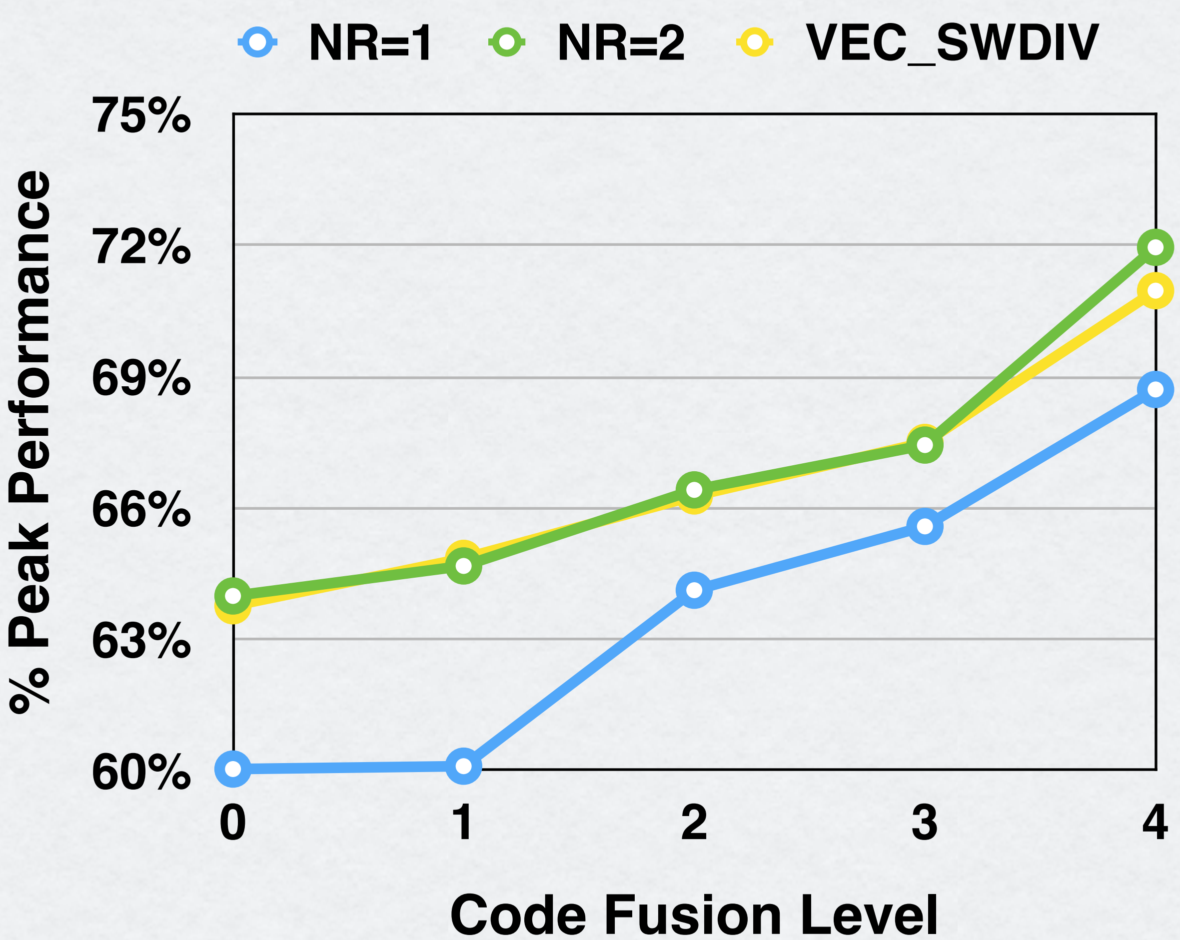
Time-to-Solution

Data alignment

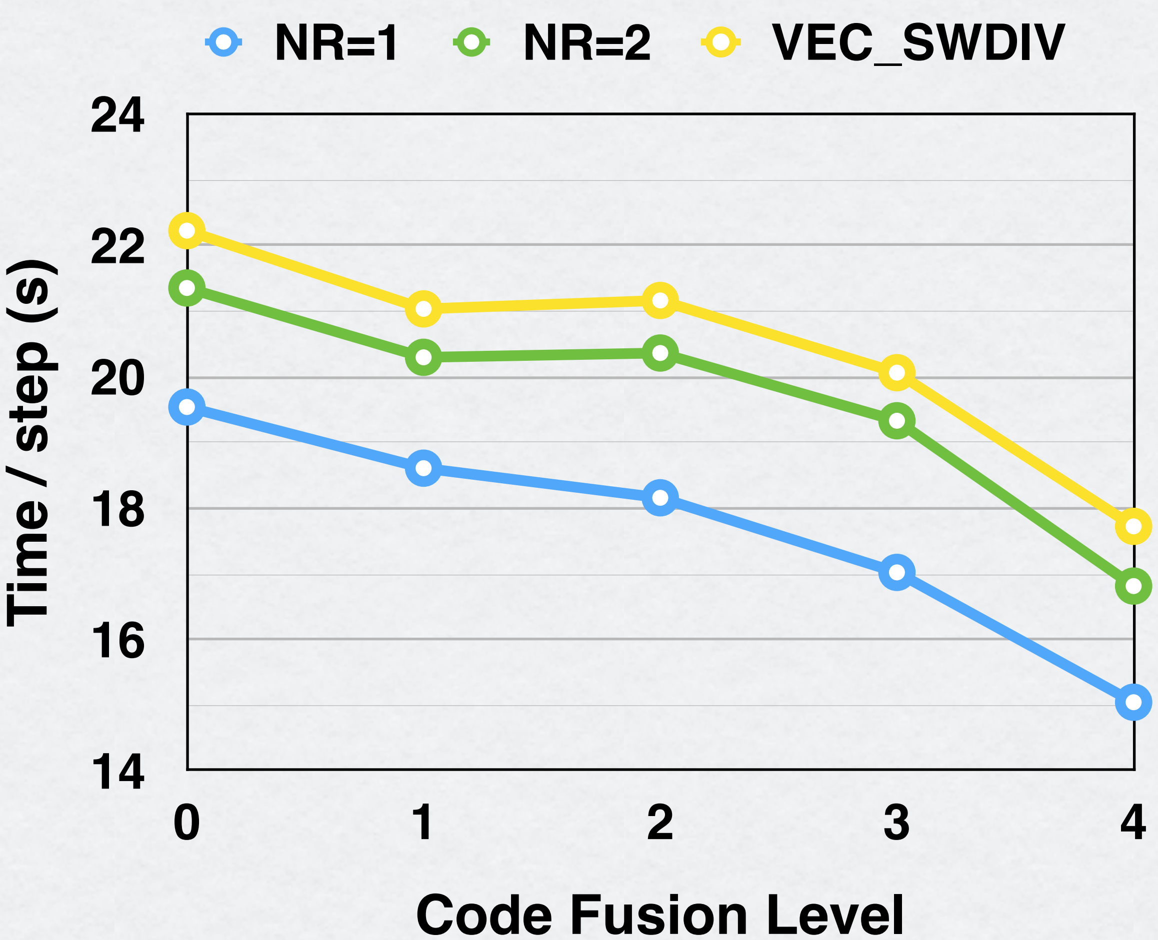
Alignment	Time/Step (s)	RHS (% peak)
16	15.03	68.79%
32	17.07	60.56%

Code Fusion (WENO, HLLE)

1 node



Peak Performance (RHS)



Time-to-Solution

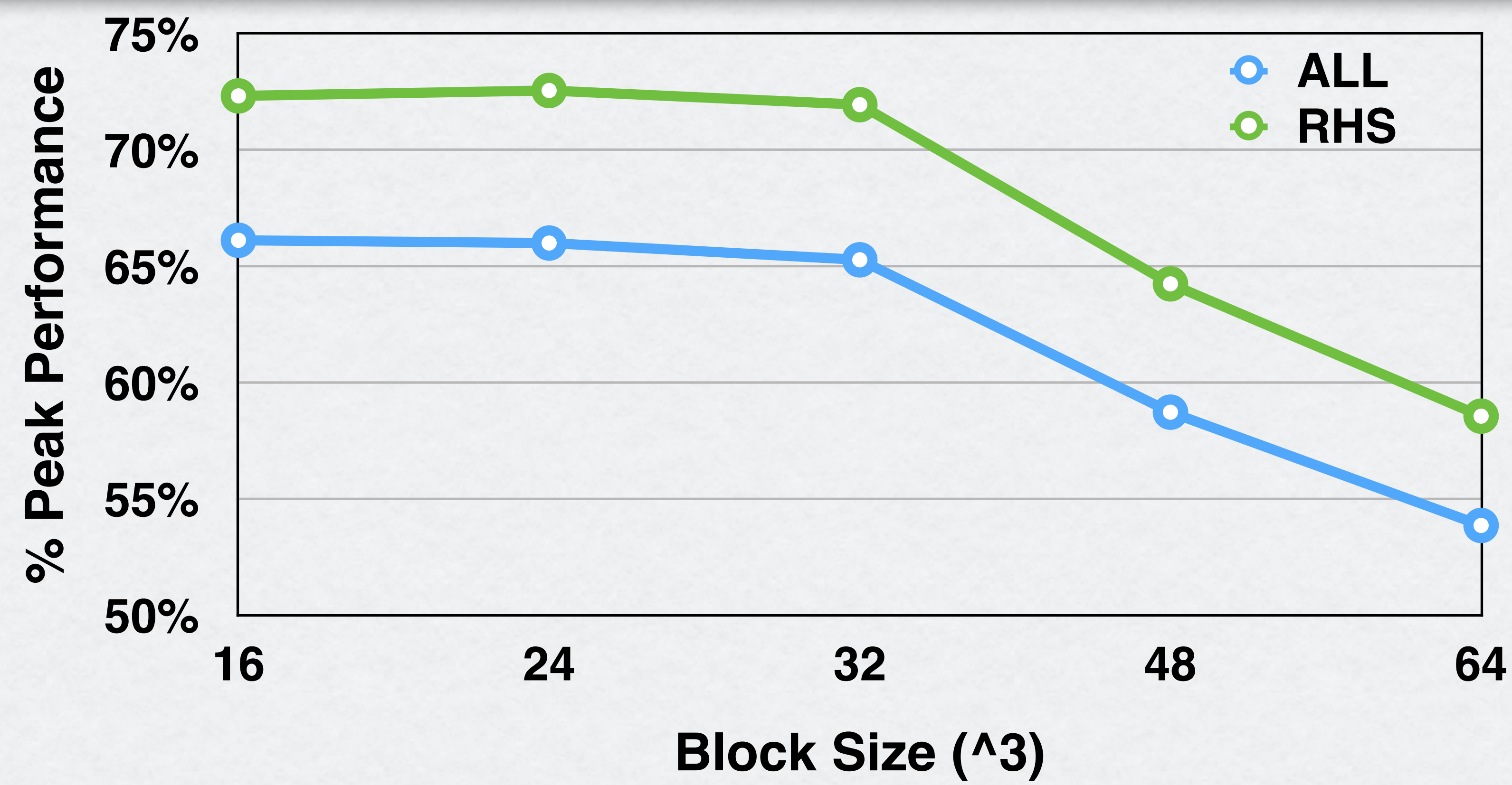
QPX vs CPP

1 node, NR=2

Full fusion — No code fusion

	QPX	QPX	CPP
RHS	**72.0%**	**64.0%**	**17.5%**
DT	19.7%	19.8%	12.2%
UP	2.3%	2.3%	2.2%
ALL	65.3%	59.3%	17.2%
Time/Step (s)	**16.8**	**21.4**	**84.1**

Block Size (NR=2)



BS^3	**16**	**24**	**32**	**48**	**64**
# Blocks	32768	9702	4096	1216	512
Time/Step (s)	18.3	17.2	16.8	18.1	19.4
GP/s (x10^6)	**7.33**	**7.81**	**7.98**	**7.41**	**6.93**

BS^3	**24**	**32**
#Blocks	15625	6859
RHS	**72.7%**	**72.4%**
GP/s (x10^6)	**7.83**	**8.03**

Going further

WENO5 vs WENO3

	NR=1		NR=2	
	WENO5	WENO3	WENO5	WENO3
Time/step (s)	15.0	10.7	16.8	11.6
RHS (%peak)	68.8%	45.1%	72.0%	52.7%

HLLE vs HLLC

	NR=1		NR=2	
	HLLE	HLLC	HLLE	HLLC
Time/step (s)	15.0	16.4	16.8	18.4
RHS (%peak)	68.8%	69.5%	72.0%	72.8%

Avoiding divisions in WENO5

	NR=1		NR=2	
	WENO5	NEW	WENO5	NEW
Time/step (s)	15.0	14.6	16.8	14.8
RHS (%peak)	68.8%	64.3%	72.0%	64.7%

Ongoing work

- Simulations of turbulent fields
- Lossy and lossless compression of 3D simulation data
- CUBISM-MPCF on heterogeneous platforms

- Web resources
    - www.cse-lab.ethz.ch
    - github.com/cselab/CUBISM-MPCF

Thank you for your attention!

JÜLICH FORSCHUNGSZENTRUM

CINECA

PRACE

Argonne NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY INCITE LEADERSHIP COMPUTING

CSCS Centro Svizzero di Calcolo Scientifico Swiss National Supercomputing Centre