Designing an Efficient Multi-GPU Coastal Ocean Model

Bridging the gap between
global changes and local impacts

De Le Court *Miguel*

Lambrechts *Jonathan*

Legat *Vincent*

✉ miguel.delecourt@uclouvain.be

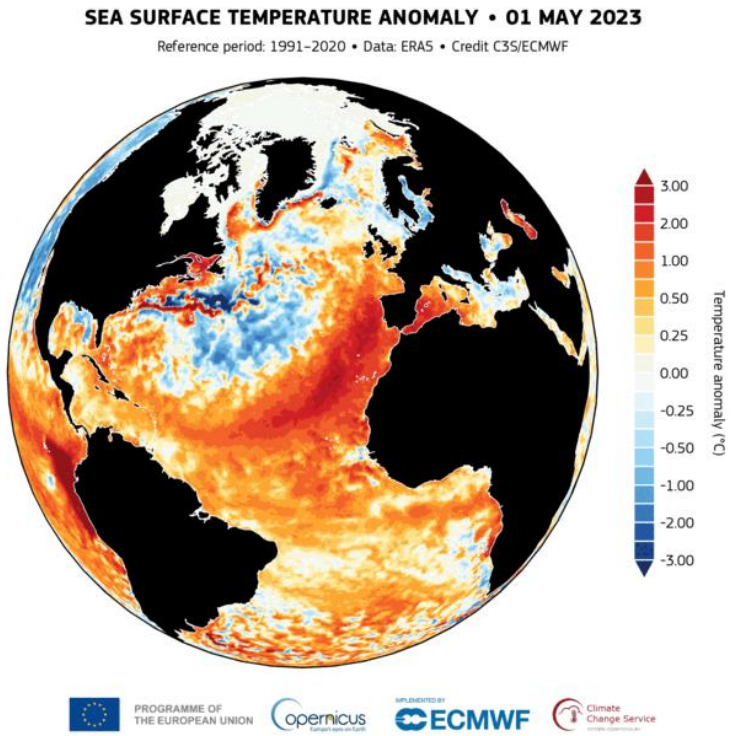


UCLouvain

AGU ADVANCING EARTH AND SPACE SCIENCE

Coastal ecosystems are under increasing pressure

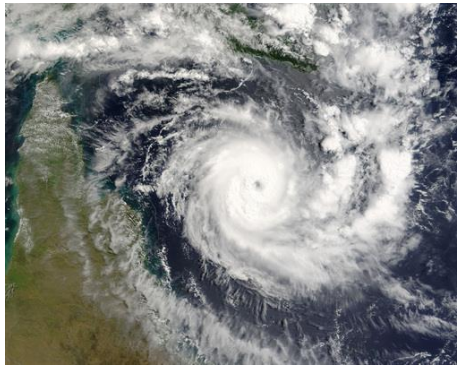




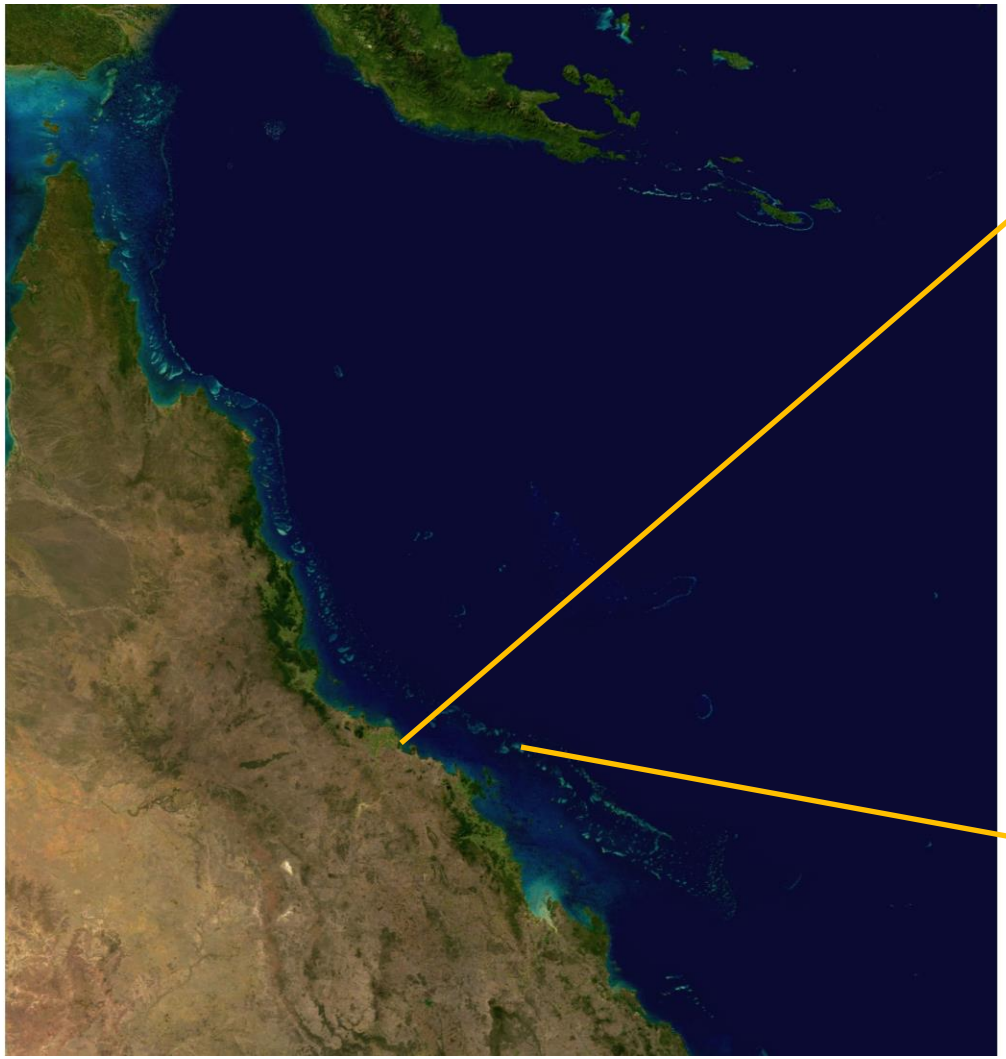








Process scales range across 4 orders of magnitude

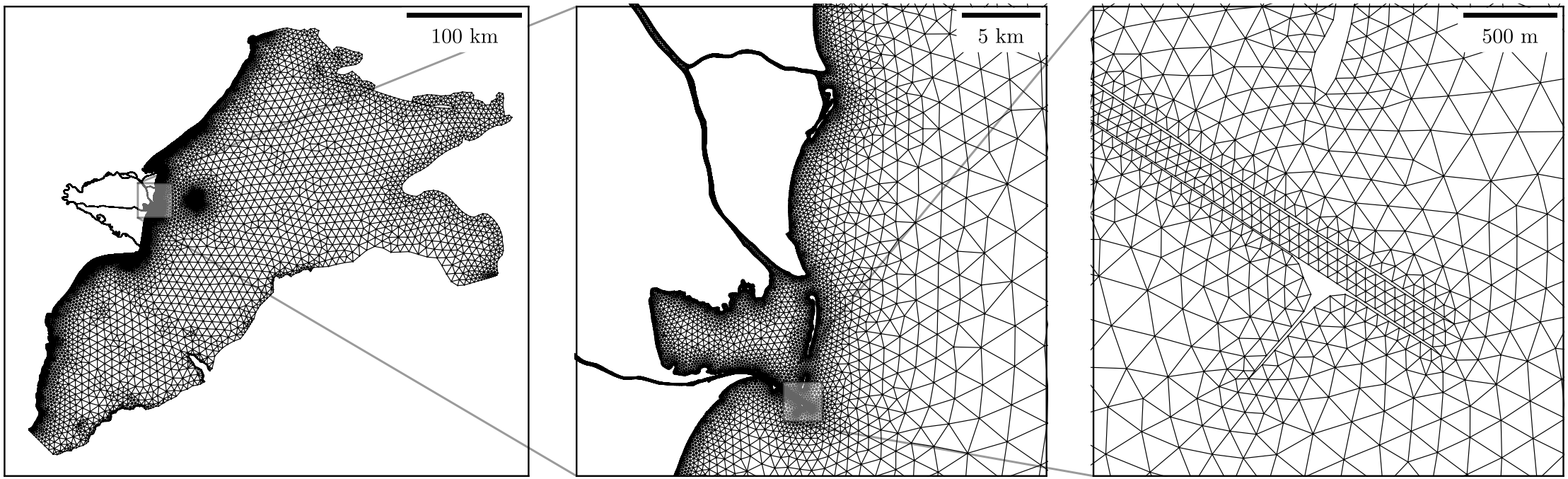






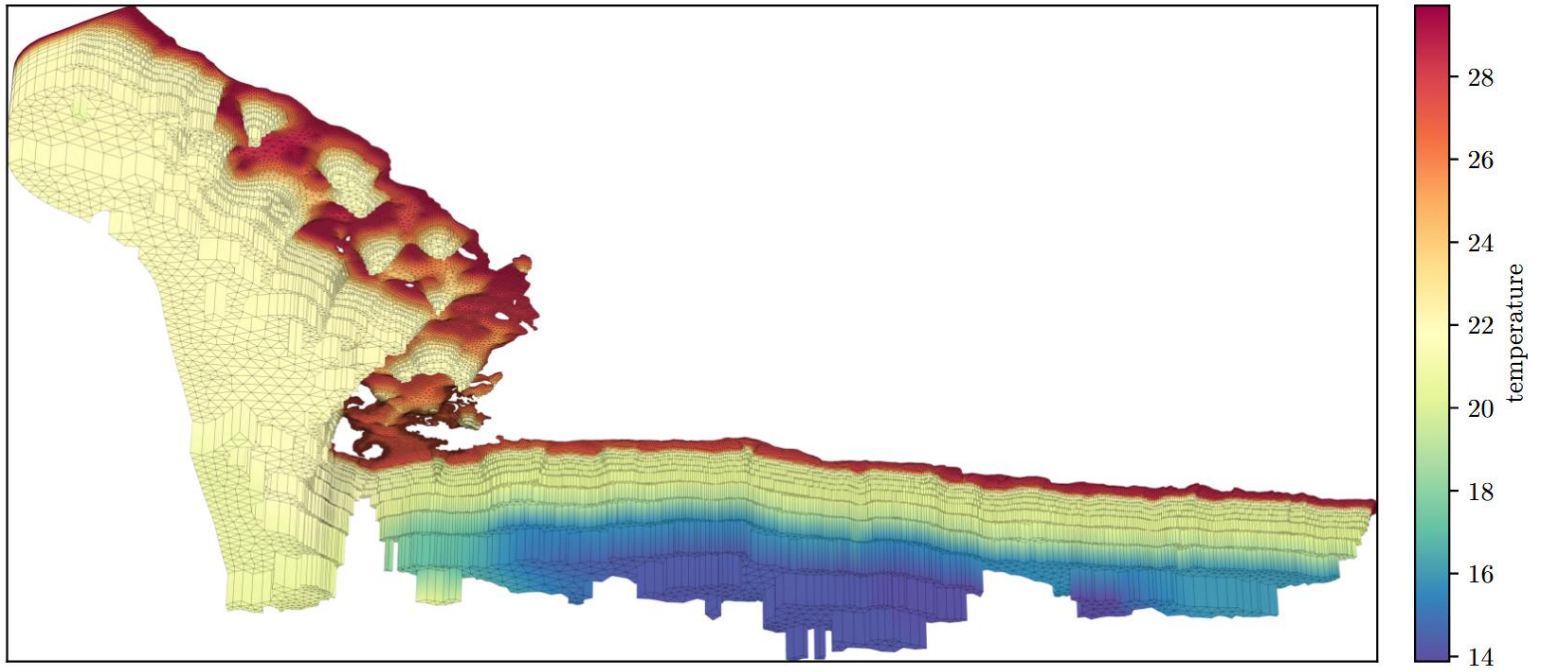
SLIM : Second-generation Louvain-la-Neuve Ice-ocean Model

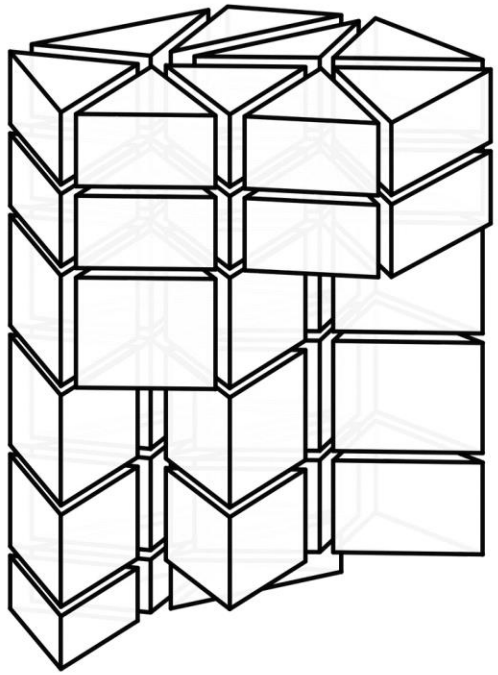
- Discontinuous Galerkin FEM
- Unstructured triangular mesh



SLIM : Second-generation Louvain-la-Neuve Ice-ocean Model

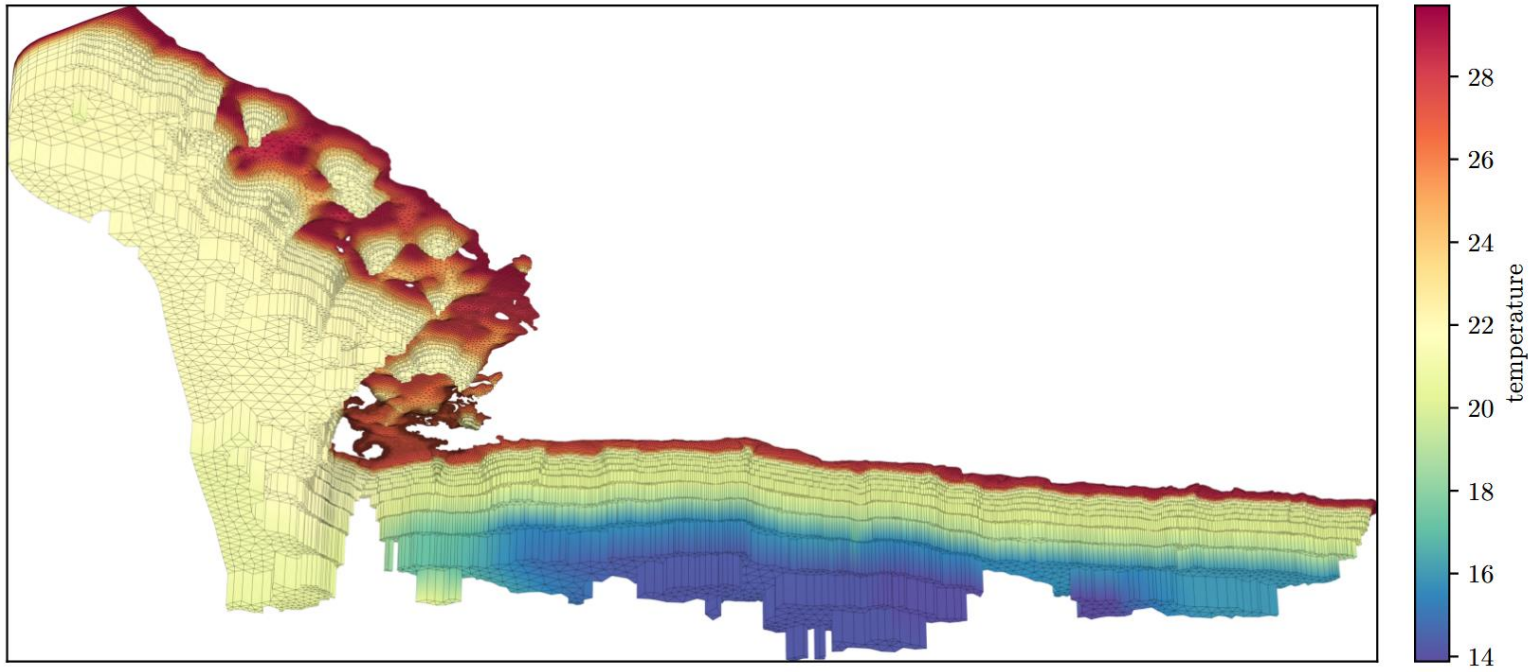
- Discontinuous Galerkin FEM
- Unstructured triangular mesh
- Vertically extruded prisms
- Split-imex temporal scheme

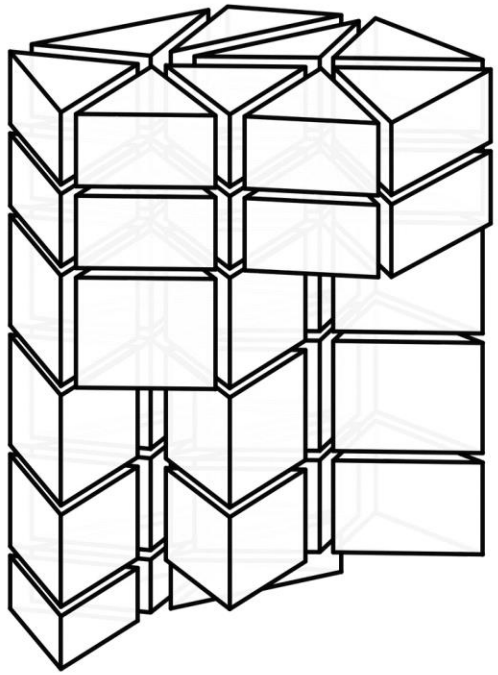




SLIM : A multi-GPU ocean model

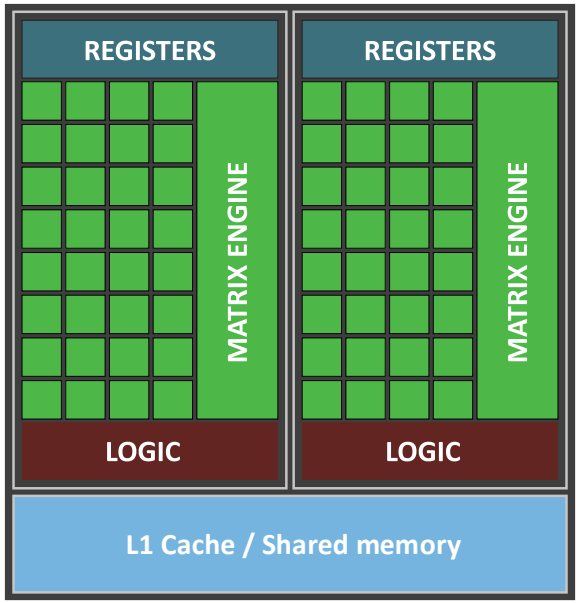
- Discontinuous Galerkin FEM : Elements can be processed in parallel ✓
- Unstructured triangular mesh : Indirection in lateral memory access ✗
- Vertically extruded prisms : Structured vertical direction ✓
- Split-imex temporal scheme : Linear system for each column ✗



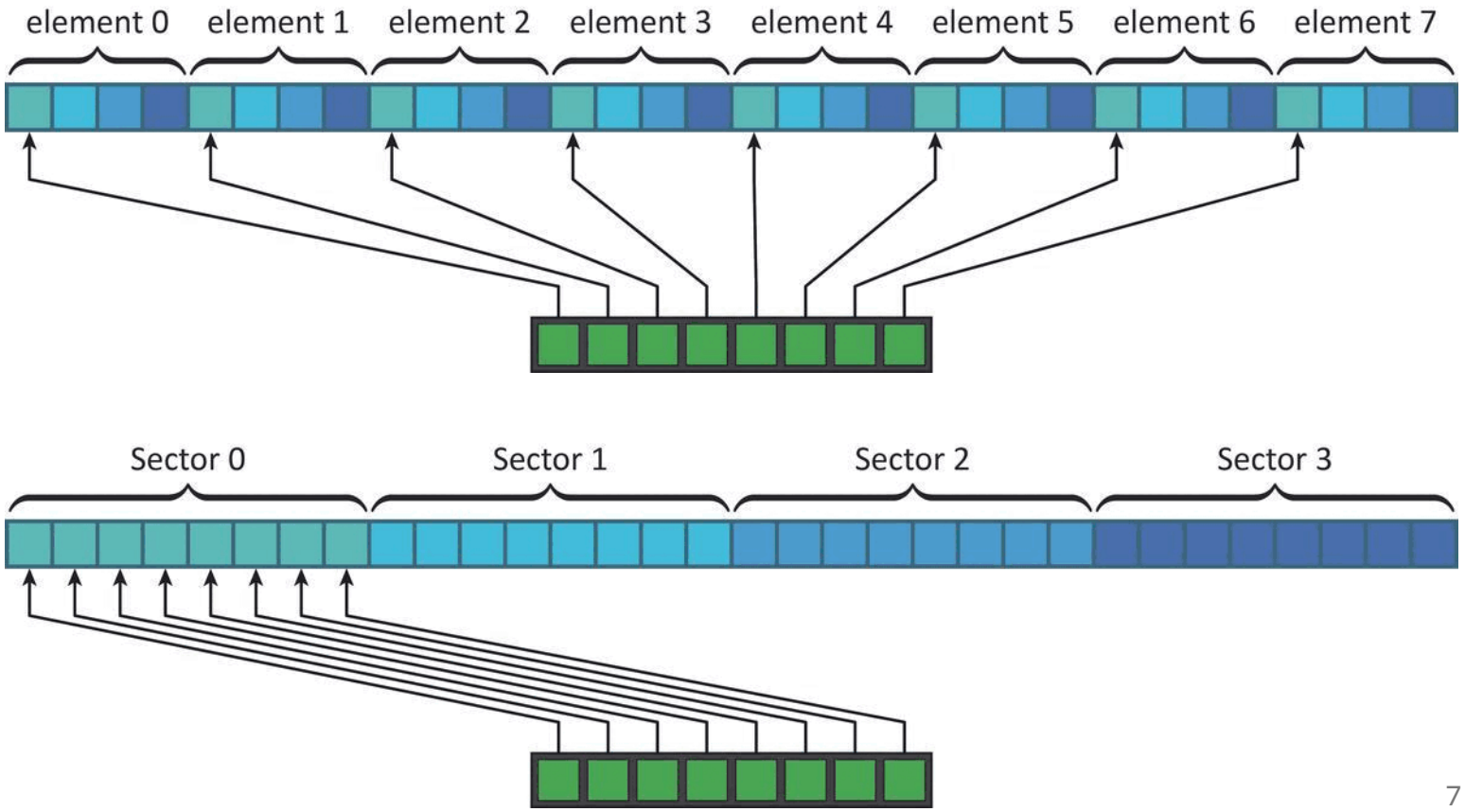


GPUs require coalescent memory access

- On the CPU : we want to maximize locality
- On the GPU : data is accessed simultaneously
- Much smaller cache per core : data will not fit! → Excessive loads
- Optimal pattern : all cores from a CU read same sector : **One sector read per access**



compute unit



Getting the most out of the GPU

Elements can be processed in parallel

- Assign 1 GPU thread per element*

Indirection in lateral memory access

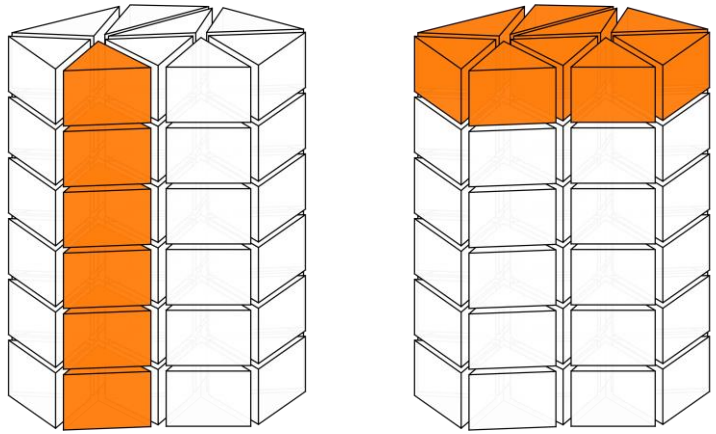
- 2D mesh is reordered following a Hilbert curve

Structured vertical direction

- Structured memory accesses

Linear system for each column

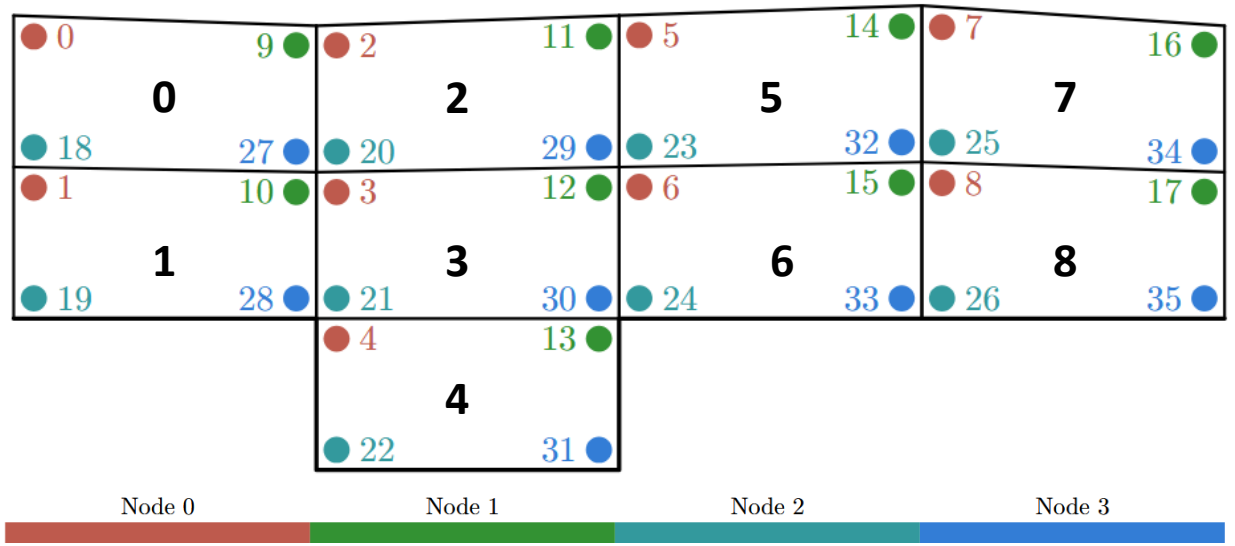
- Cell memory layout



The “cell” layout

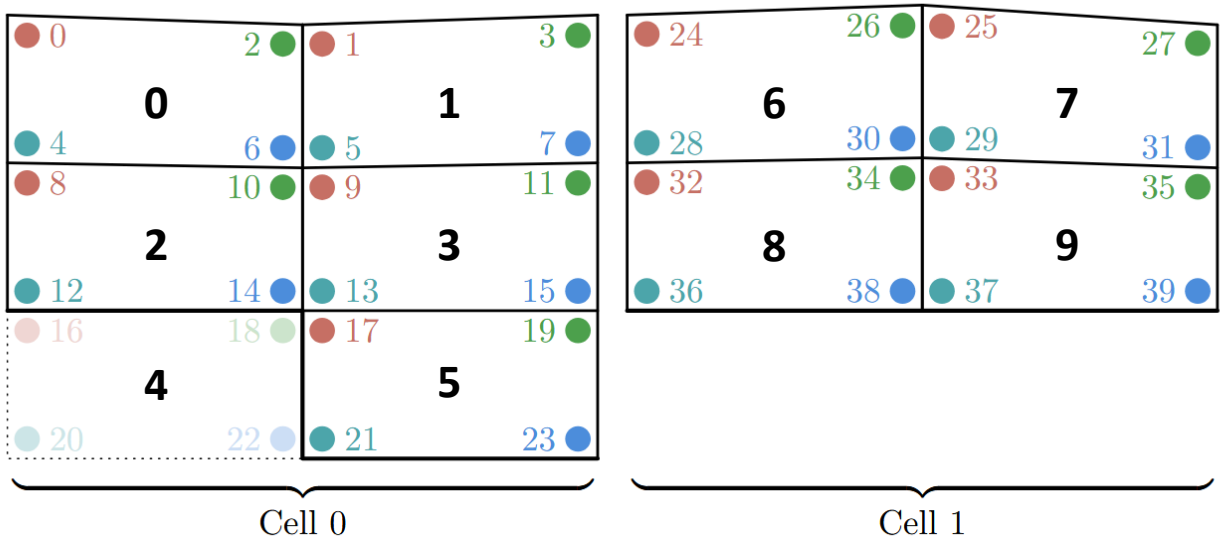
Default memory layout :

- Efficient accesses to neighboring values
- Best for assembly of matrices

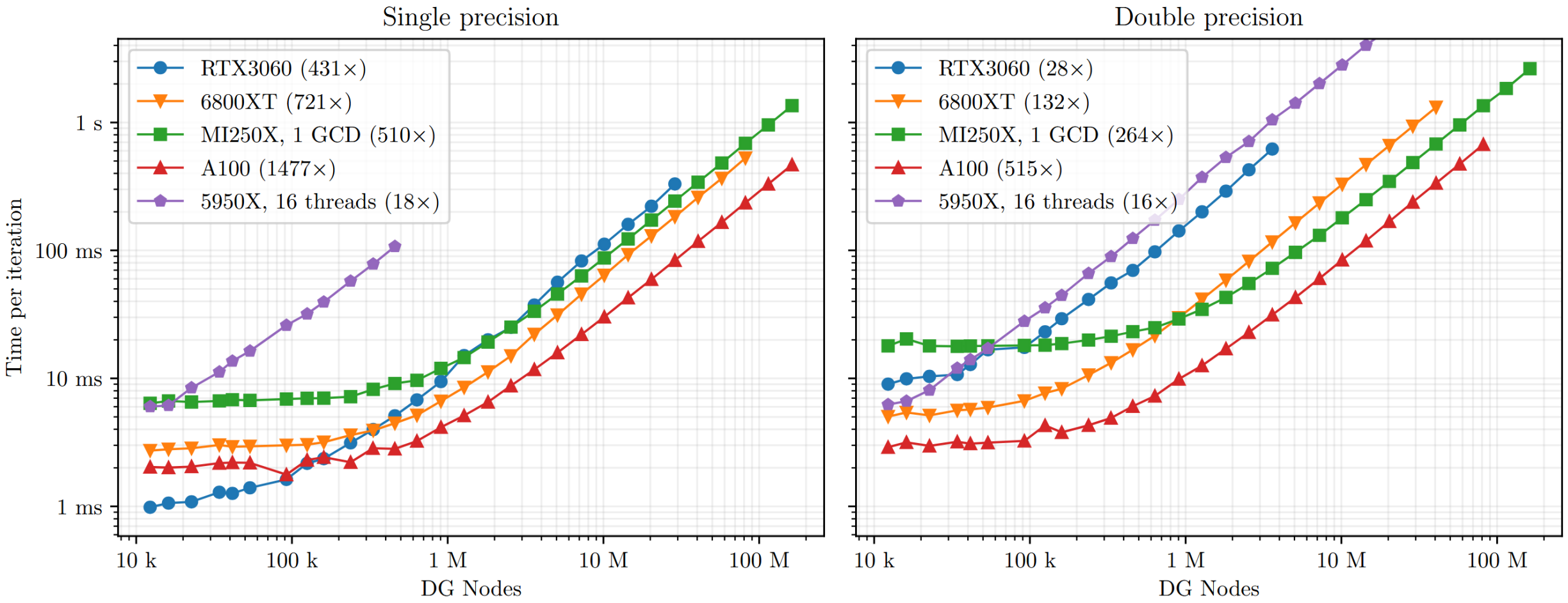


Cell memory layout :

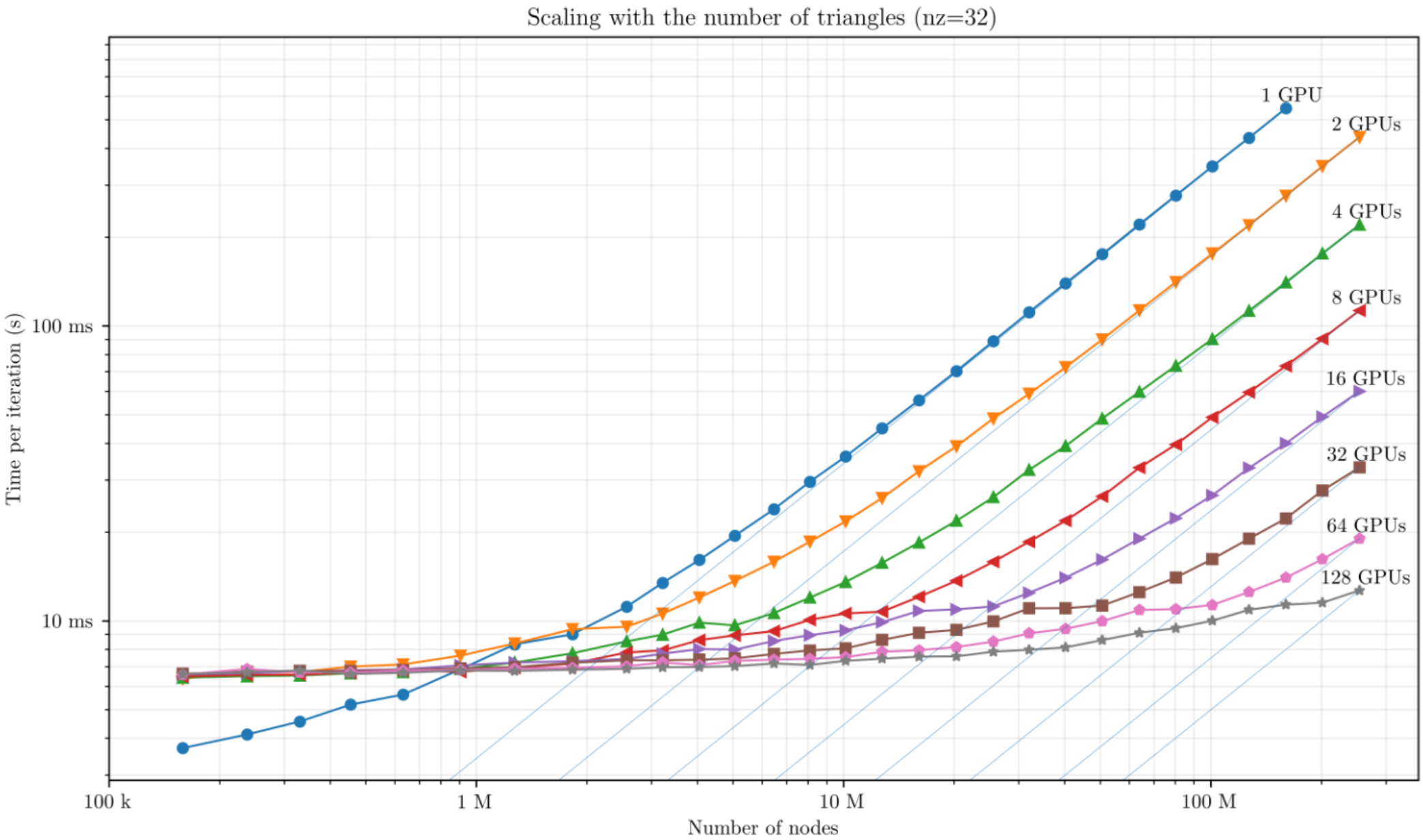
- Used to solve linear systems
- Optimal memory layout when solving
- Achieves peak bandwidth



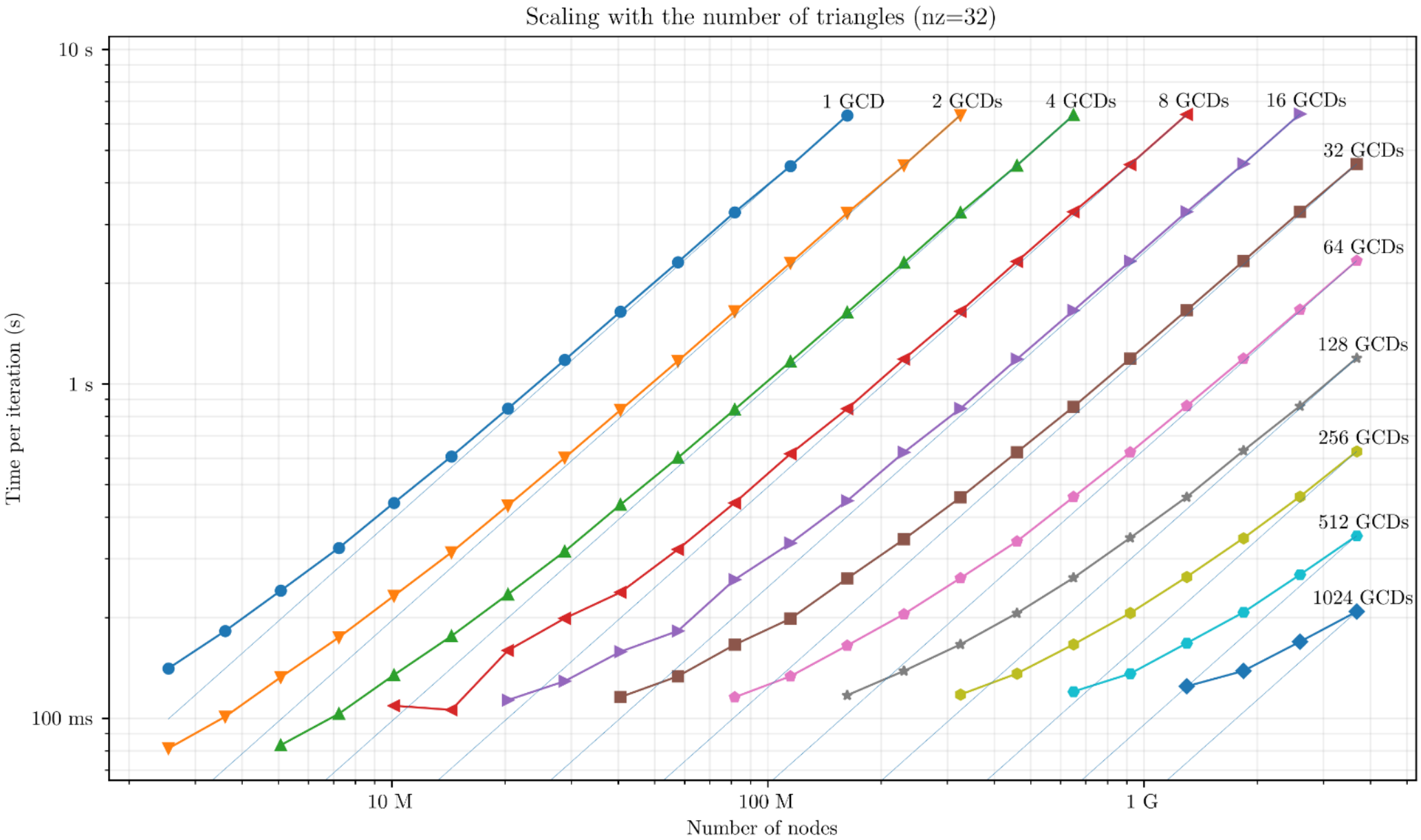
Performance of SLIM : various architectures



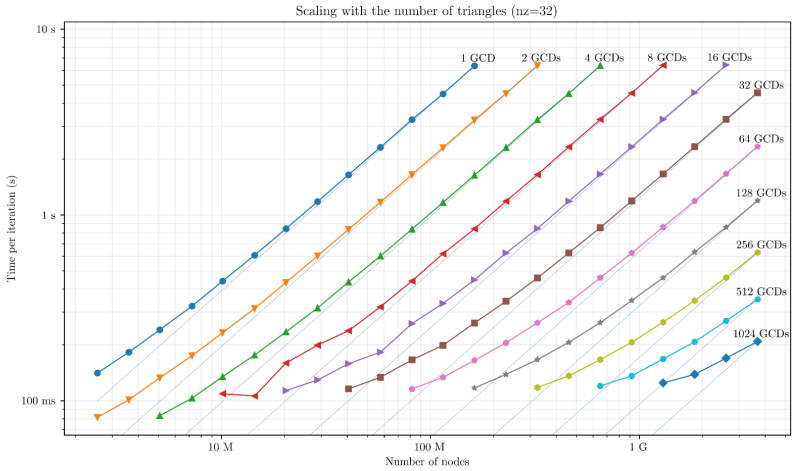
Performance of SLIM : Scaling on NVIDIA A100

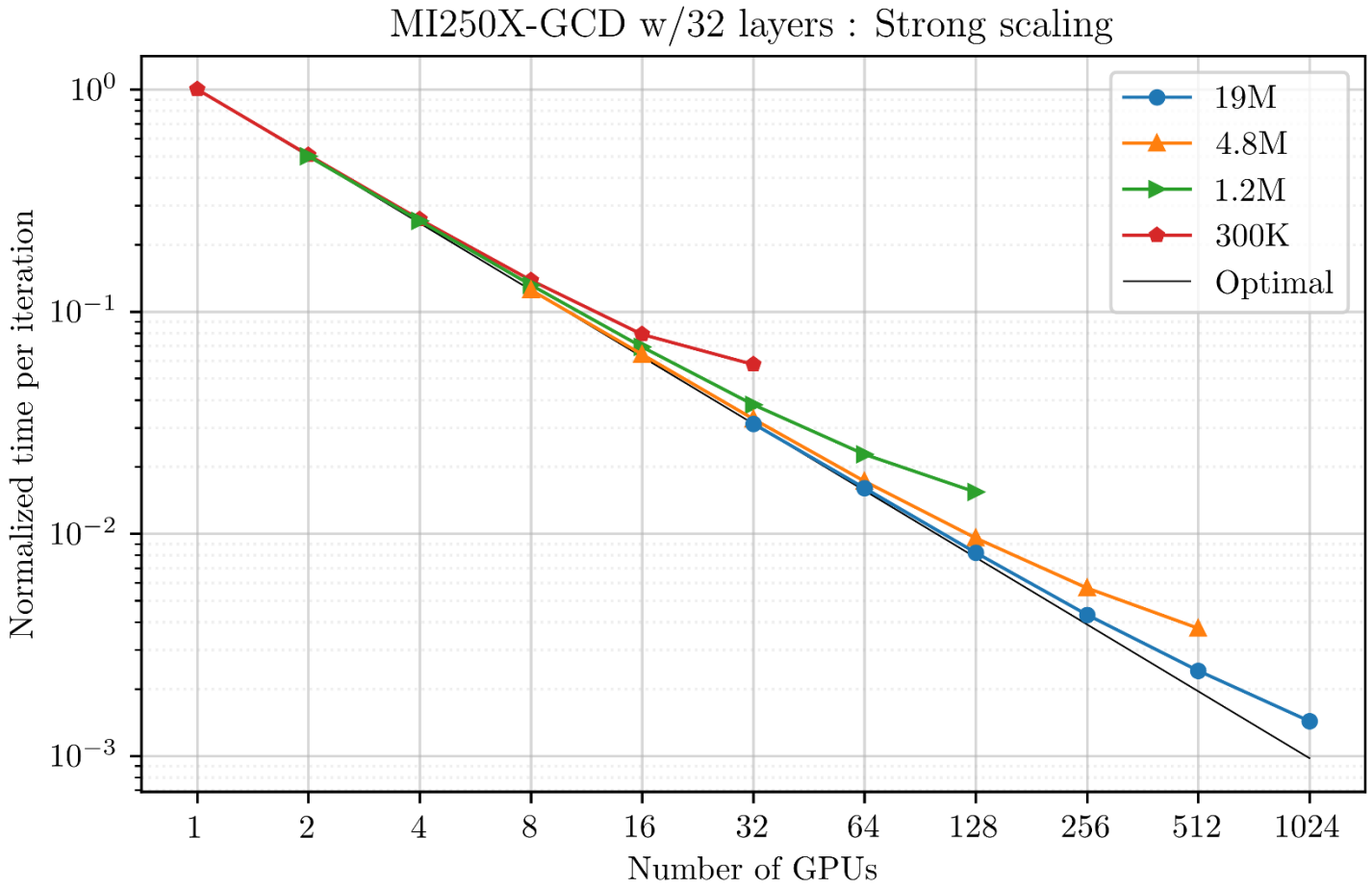


Performance of SLIM : Scaling on AMD MI250X



Scaling across multiple GPUs (AMD)

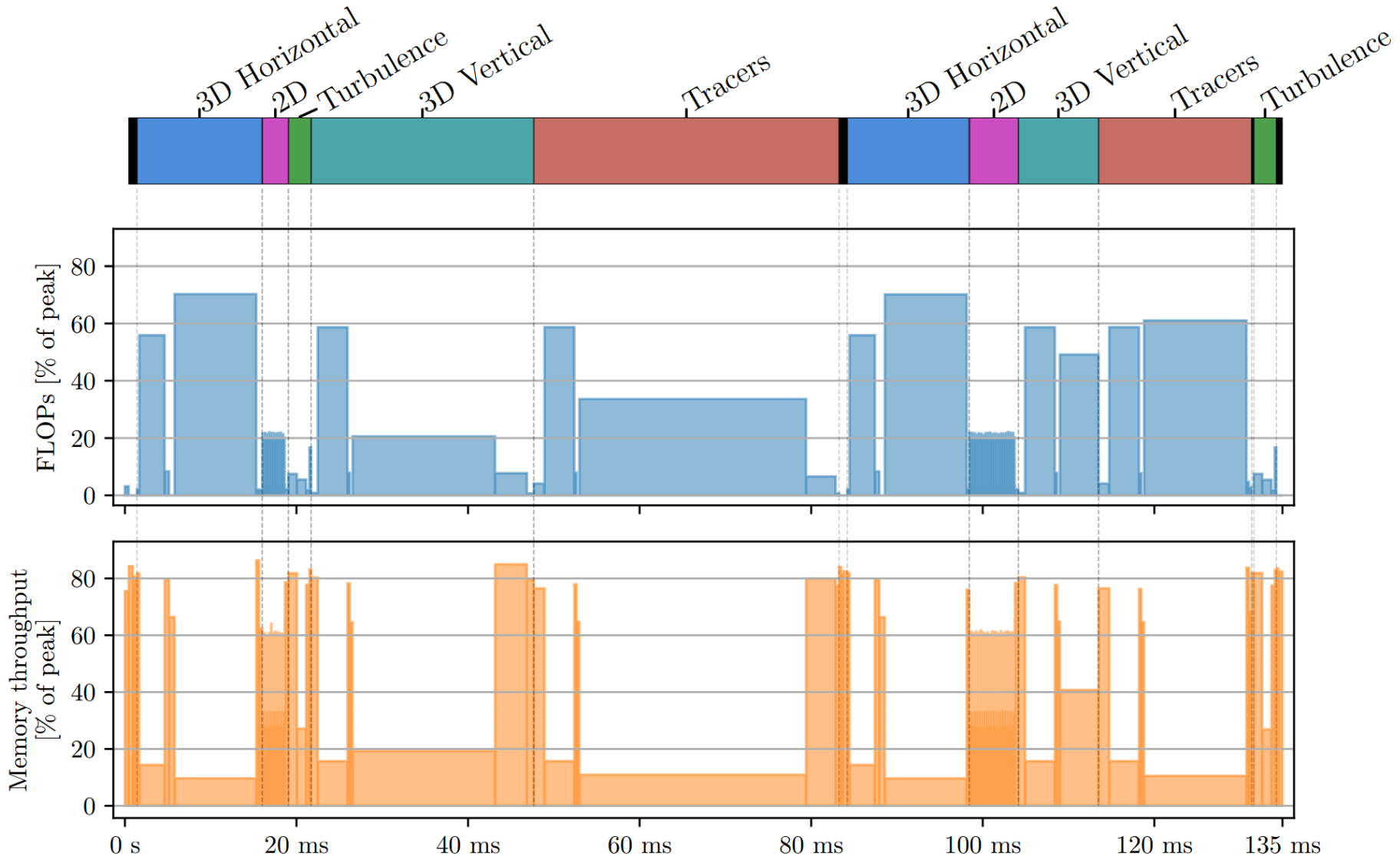




Performance of SLIM : Efficient use of resources

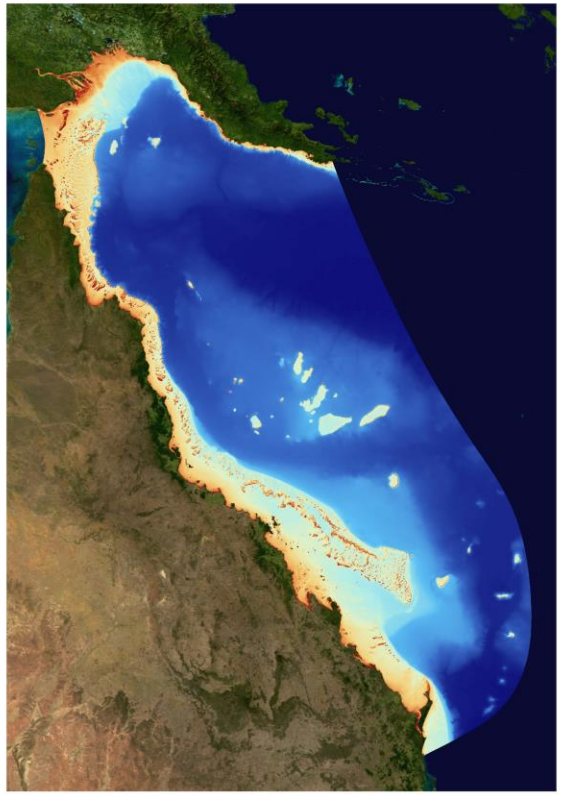
AVG : 30% of peak FP32

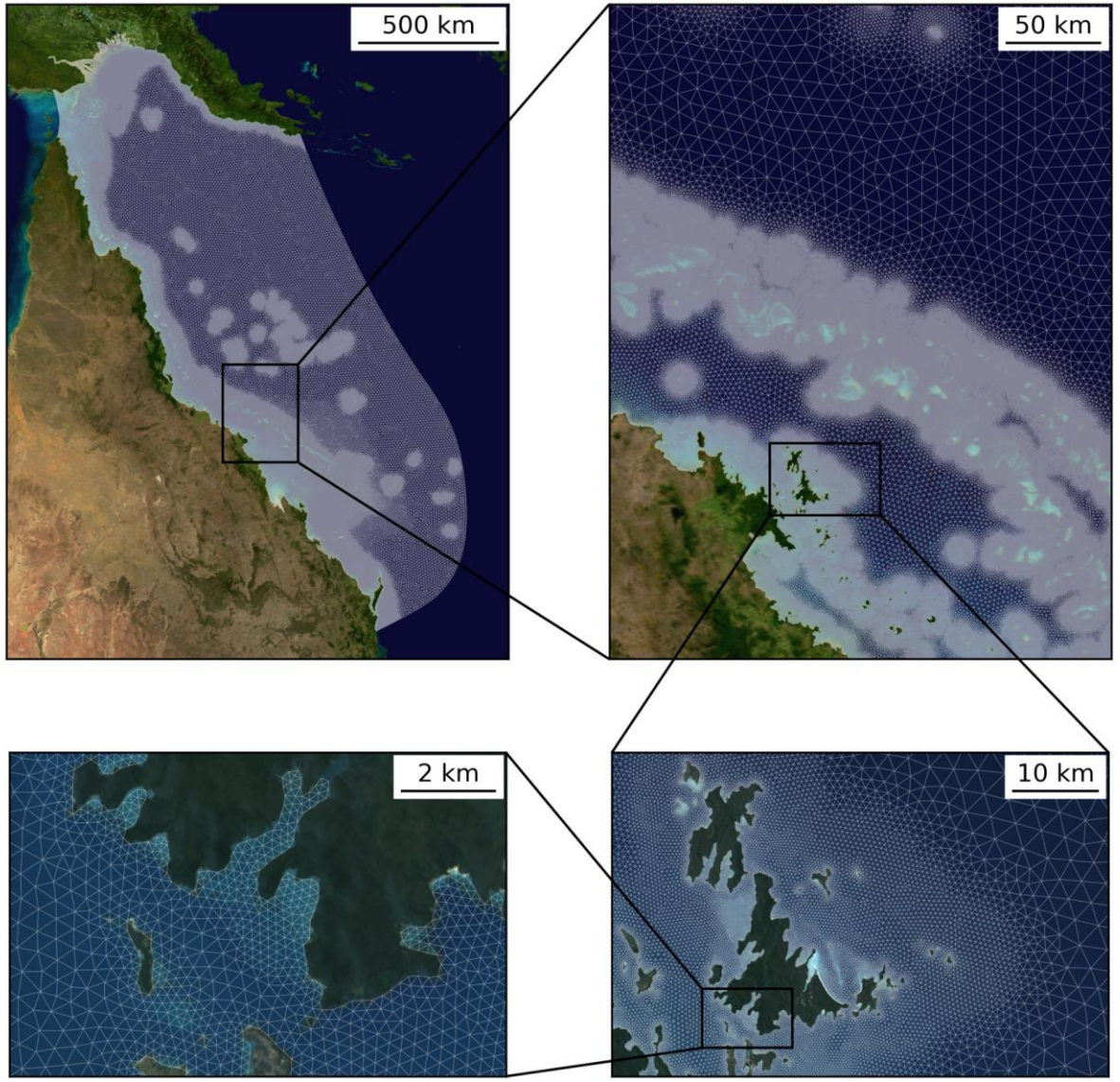
AVG: 30% of peak bandwidth



What the Multi-GPU SLIM Model Unlocks : Large-scale sims & more

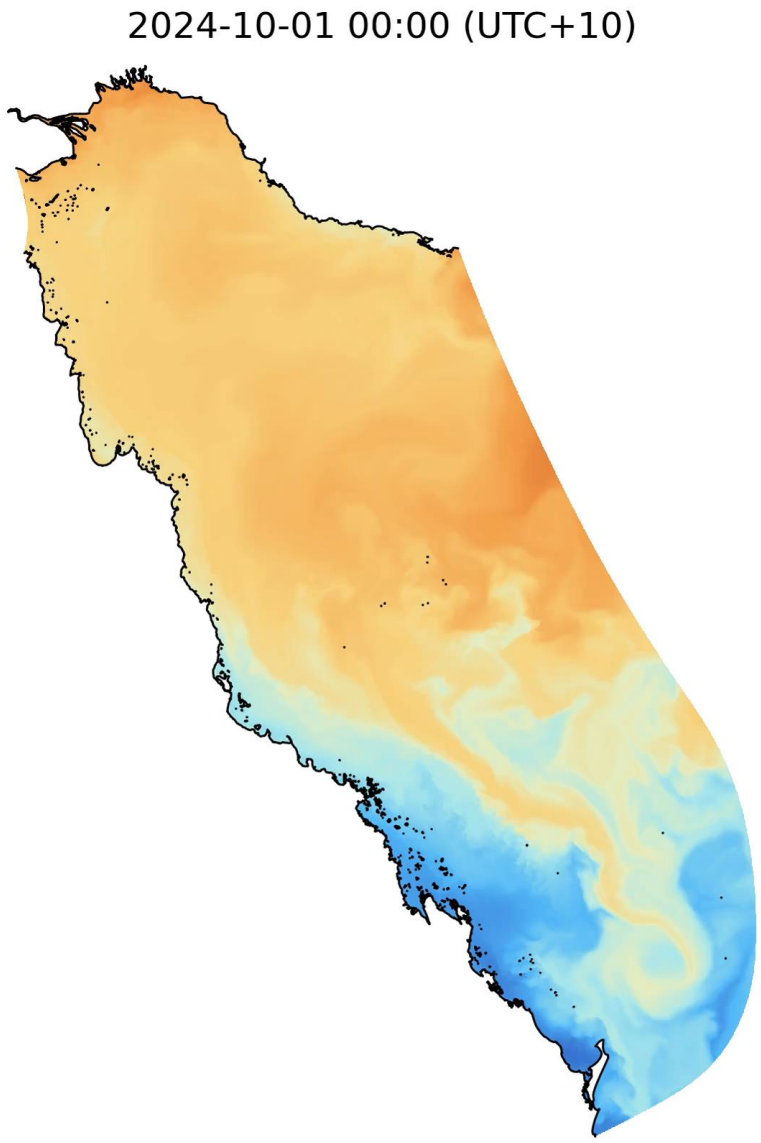
- 30M prisms, 200M nodes, 1B DoFs
- 2200km long, 700km wide
- 200m resolution on reefs
- Speed ratio of ~100 on 64 MI20X GCDs
- 400GB of outputs per simulated month





What the Multi-GPU SLIM Model Unlocks : Large-scale sims & more

- 30M prisms, 200M nodes, 1B DoFs
- 2200km long, 700km wide
- 200m resolution on reefs
- Speed ratio of ~100 on 64 MI20X GCDs
- 400GB of outputs per simulated month



Happy to chat about questions, ideas, solutions !

✉ miguel.delecourt@uclouvain.be