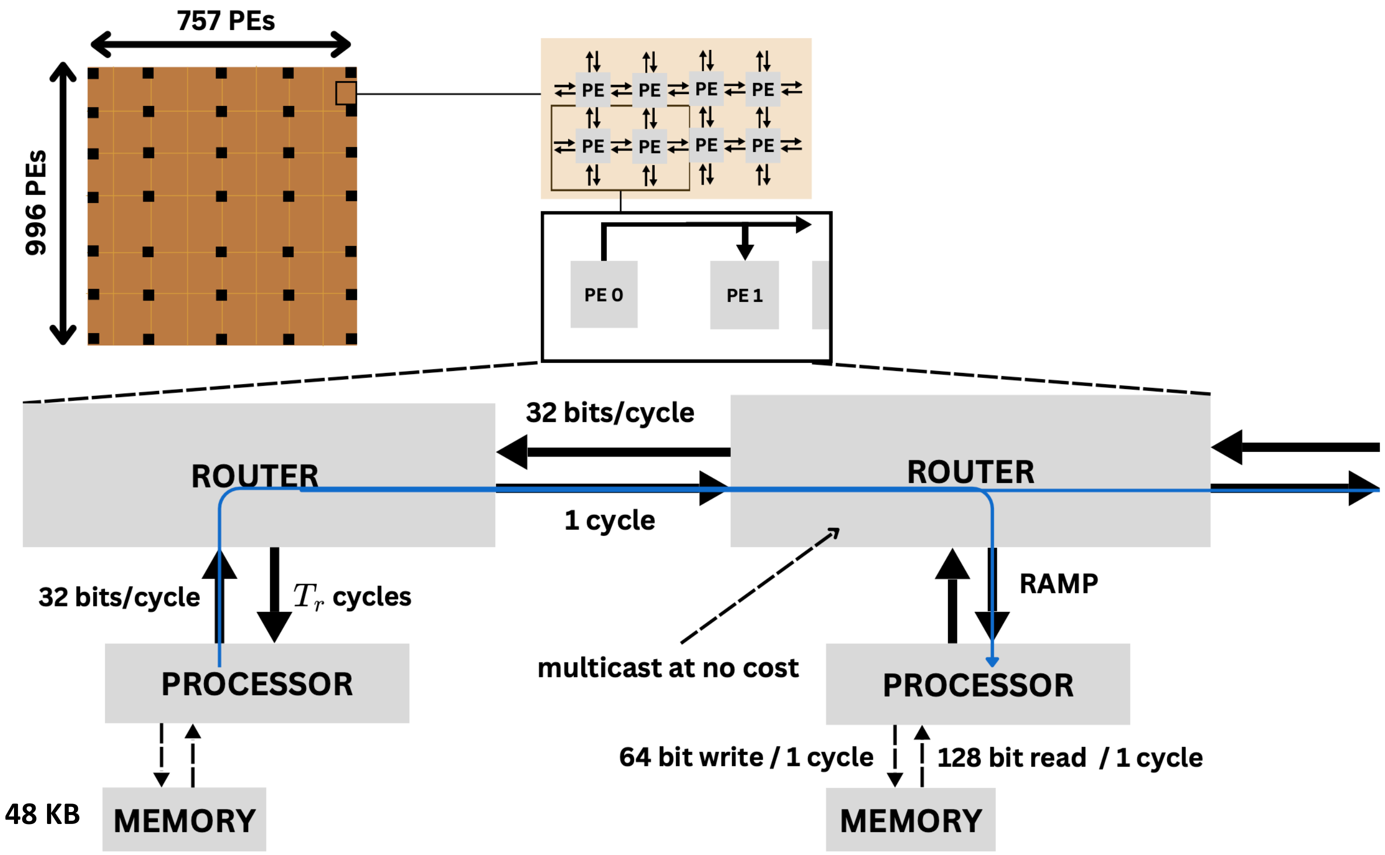
LUKAS GIANINAZZI, PIOTR LUCZYNSKI, LEIGHTON WILSON, P. IFF, D. DE SENSI, M. BESTA, S. ASHKBOOS, Y. BAUMANN, T. BEN-NUN, T. HOEFLER

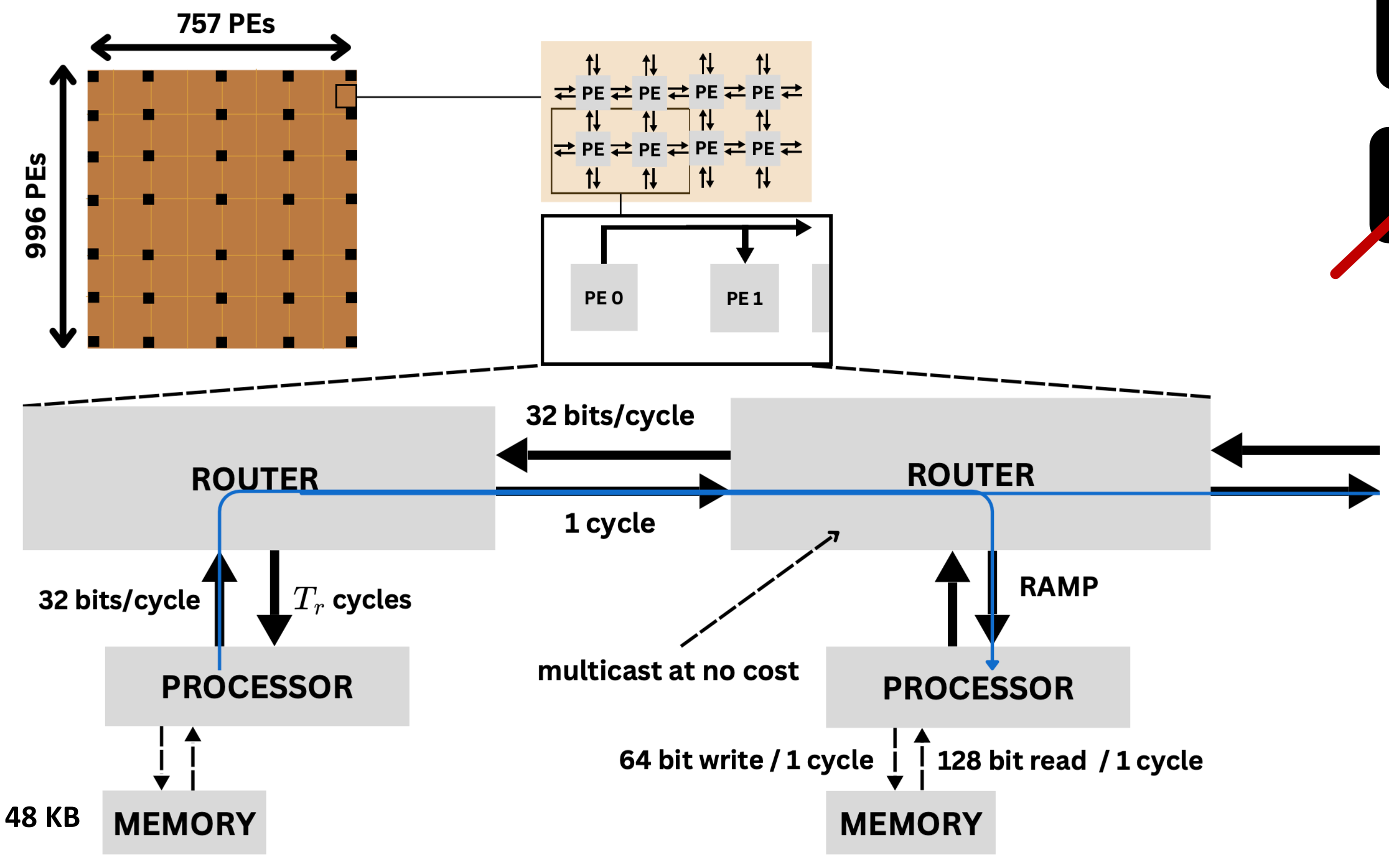
Performance Models Enable HPC on AI Accelerators

Modeling AI Accelerators



Cerebras CS-2 Wafer-Scale Engine

Modeling AI Accelerators

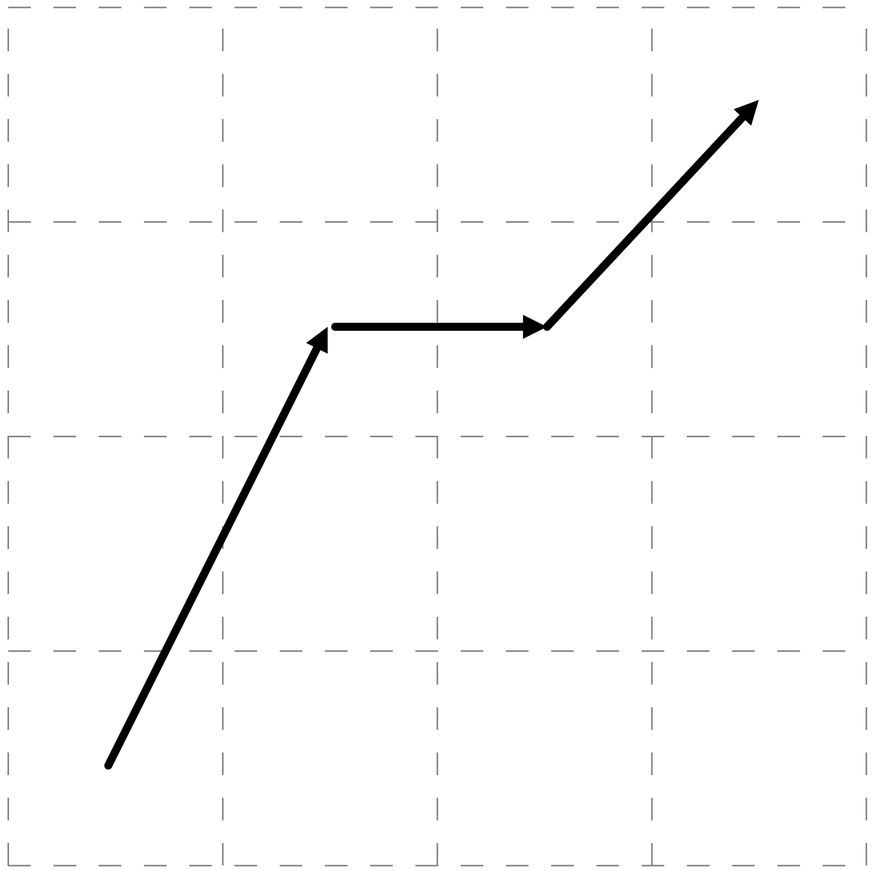


~~PRAM~~

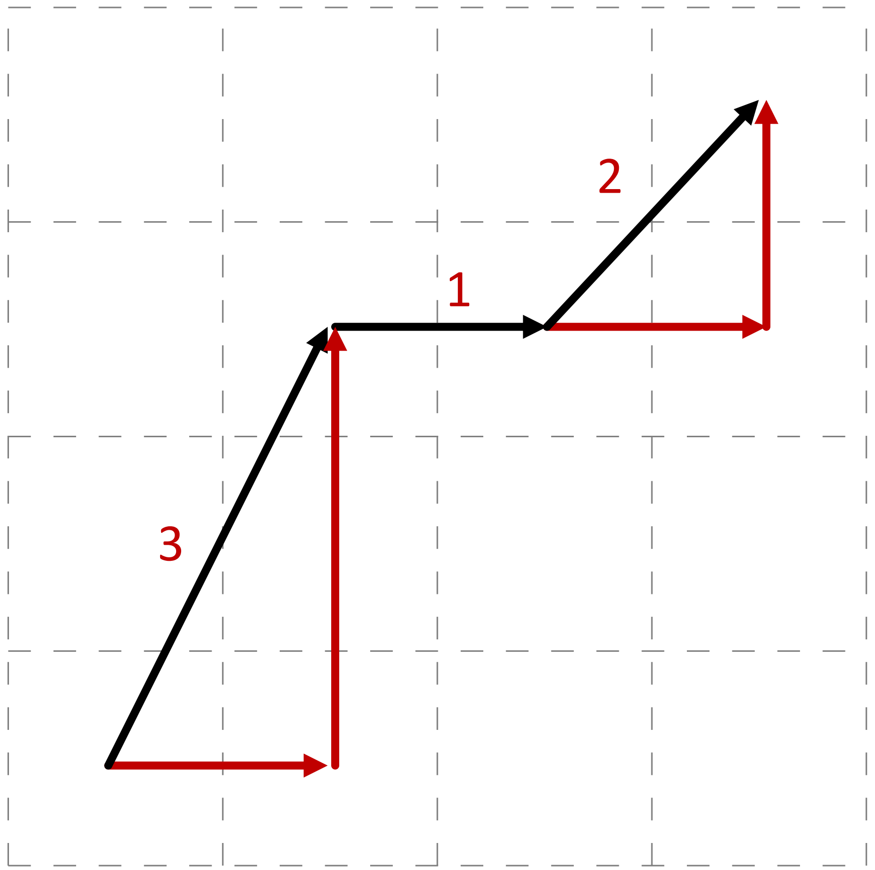
~~BSP~~

Cerebras CS-2 Wafer-Scale Engine

Modeling AI Accelerators – Spatial Model

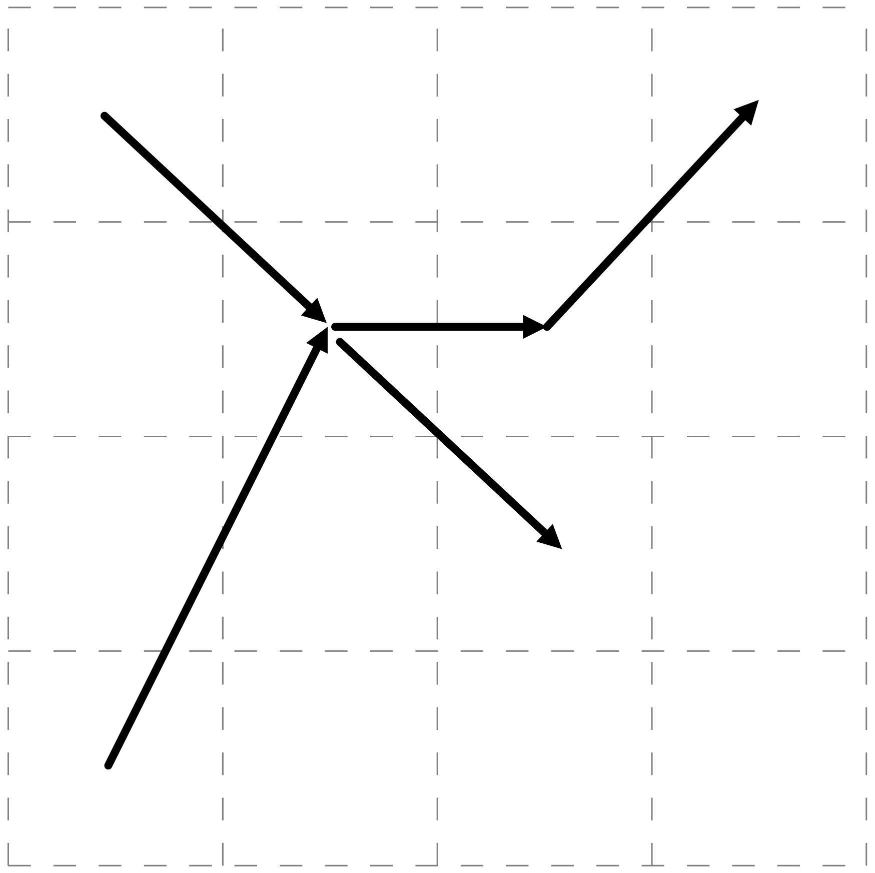


Modeling AI Accelerators – Spatial Model



Distance 6

Modeling AI Accelerators – Spatial Model

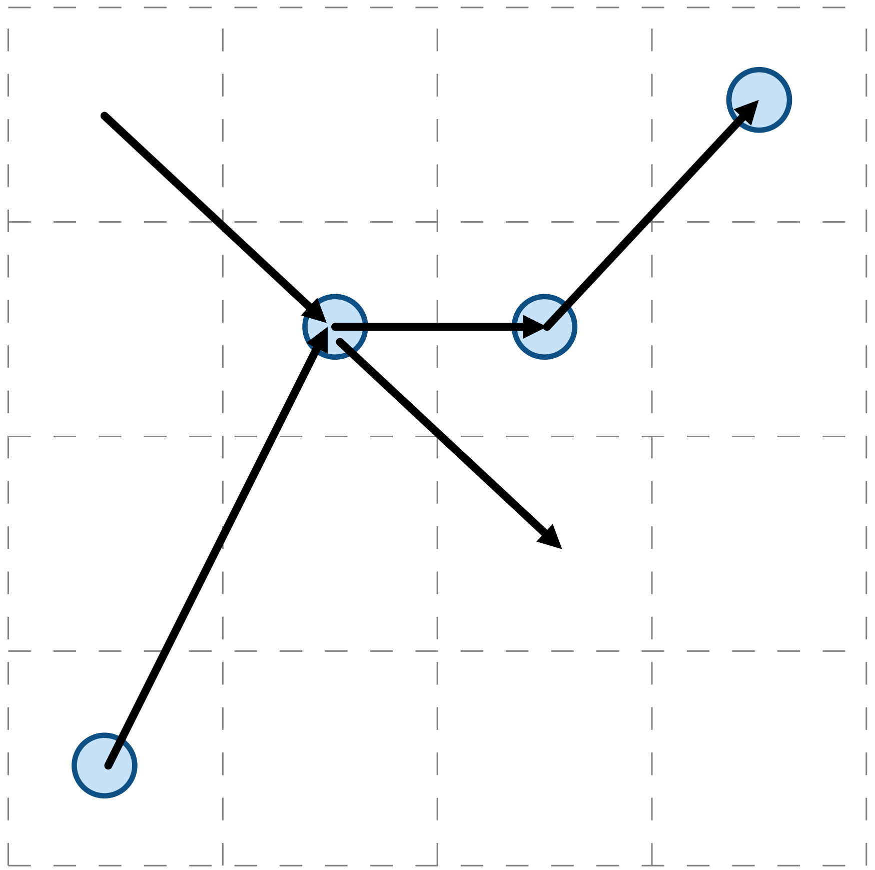


Distance 6

Maximum 6

Total 10

Modeling AI Accelerators – Spatial Model



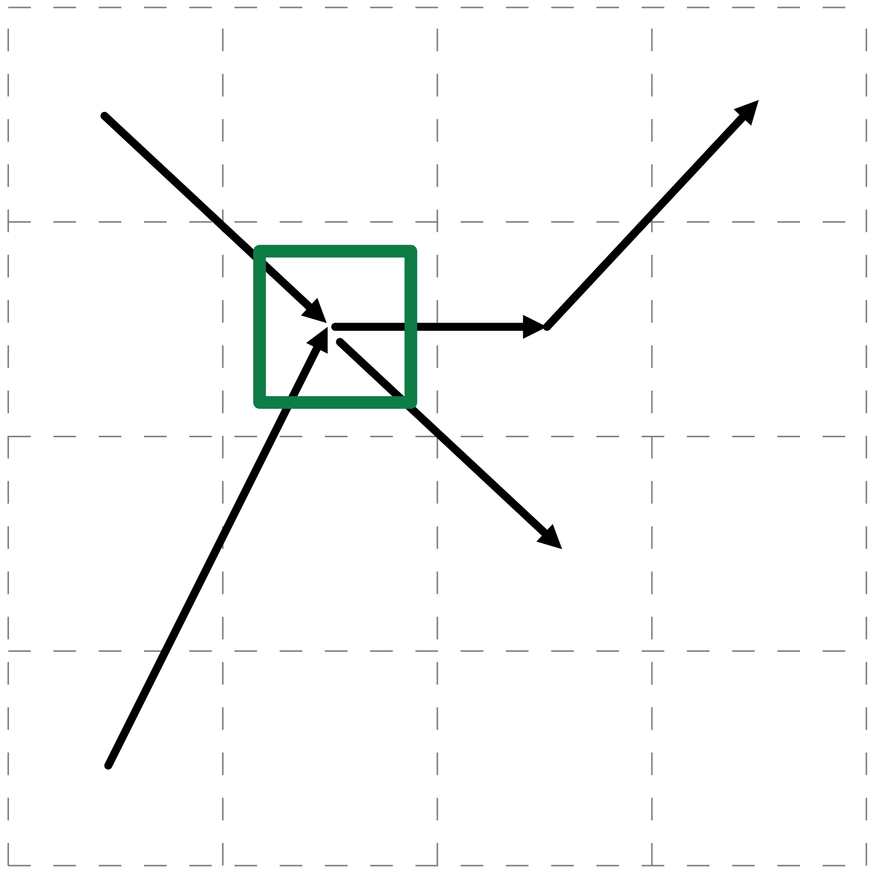
Distance 6

Maximum 6

Total 10

Depth 4

Modeling AI Accelerators – Spatial Model



Distance 6

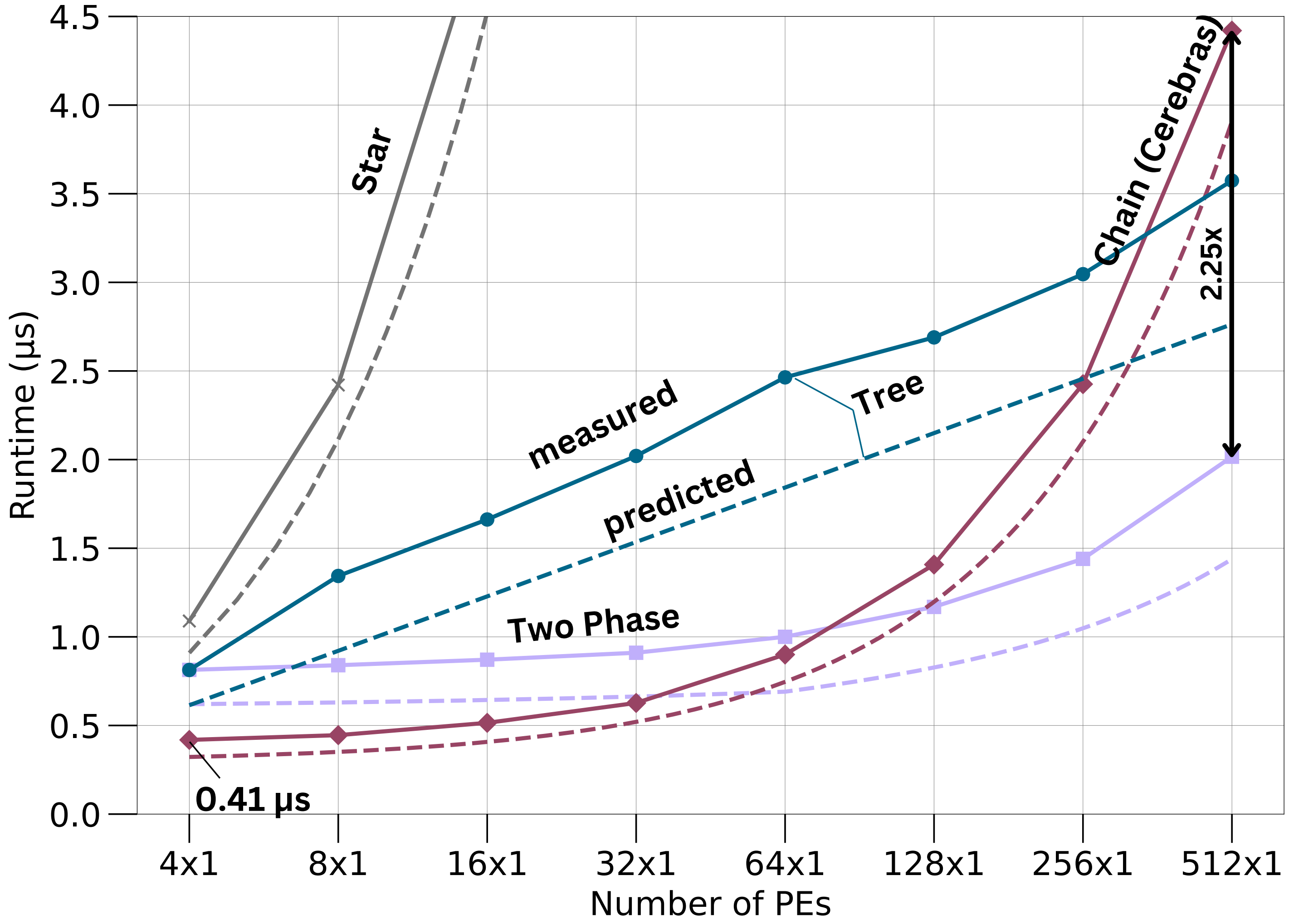
Maximum 6

Total 10

Depth 4

Contention 2

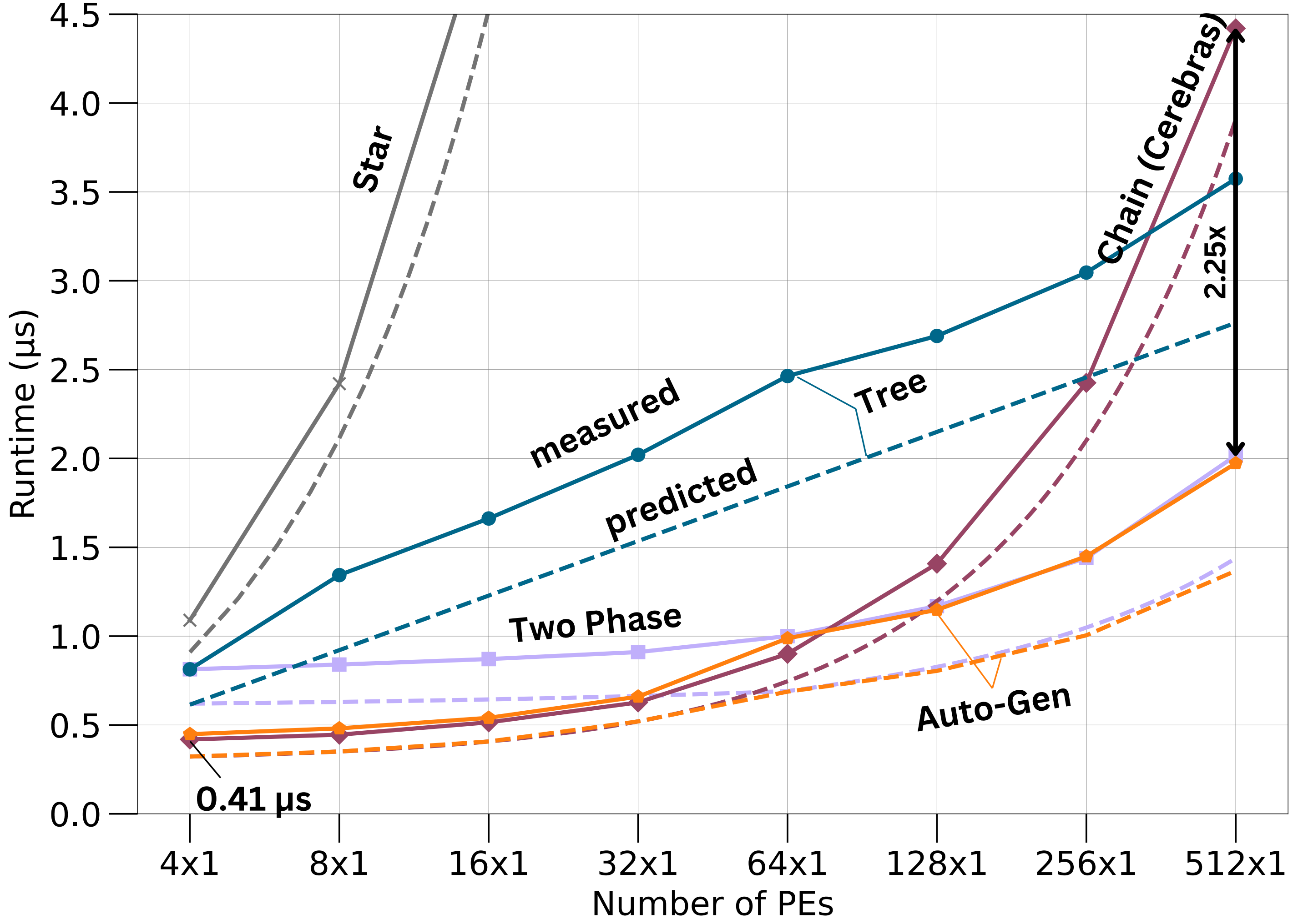
Communication Collectives on the CS-2



Reduce 1 KB per PE

Near-Optimal Wafer-Scale Reduce
https://arxiv.org/abs/2404.15888
to appear at HPDC 2024

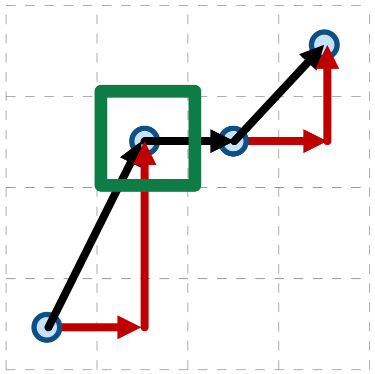
Communication Collectives on the CS-2



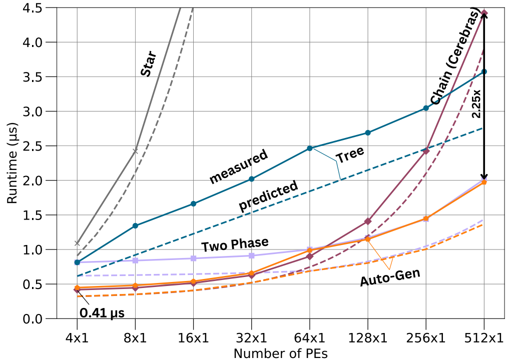
Reduce 1 KB per PE

Near-Optimal Wafer-Scale Reduce
https://arxiv.org/abs/2404.15888
to appear at HPDC 2024

Conclusions



Spatial Performance Model



Design Efficient Algorithms

Generate Efficient Kernels

More of SPCL's research:

youtube.com/@spcl 180+ Talks

twitter.com/spcl_eth 1.4K+ Followers

github.com/spcl 3.8K+ Stars

… or spcl.ethz.ch

Near-Optimal Wafer-Scale Reduce
Lucyznski & Gianinazzi et al.
https://arxiv.org/abs/2404.15888
to appear at HPDC 2024