NERO:
A Near High-Bandwidth Memory Stencil Accelerator for Weather Prediction Modeling

Gagandeep Singh, Dionysios Diamantopoulos, Christoph Hagleitner,
Juan Gómez-Luna, Sander Stuijk, Onur Mutlu, and Henk Corporaal

30th FPL, Sweden
31th August 2020

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY
ETH Zürich
SAFARI
IBM Research | Zurich

Executive Summary

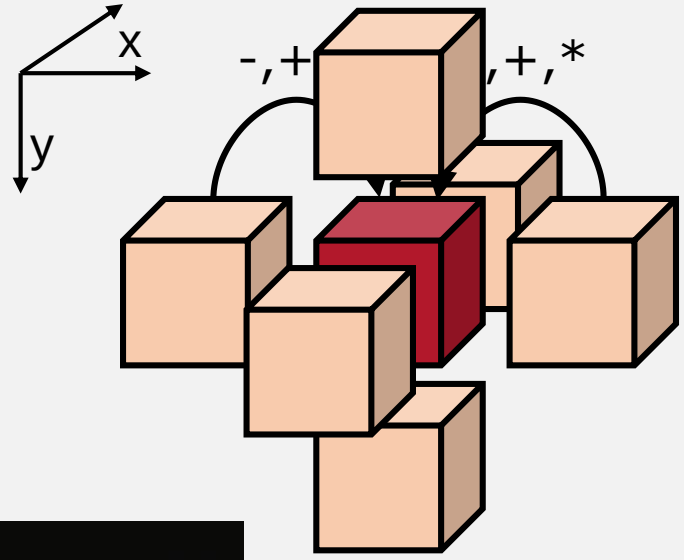
- **Motivation:** Stencil computation is an essential part of weather prediction applications
- **Problem:** Memory bound with limited performance and high energy consumption on multi-core architectures
- **Goal:** Mitigate the performance bottleneck of compound weather prediction kernels in an energy-efficient way
- **Our contribution: NERO**
  - First near High-Bandwidth Memory (HBM) FPGA-based accelerator for representative kernels from a real-world weather prediction application
  - Detailed roofline analysis to show weather prediction kernels are constrained by DRAM bandwidth on a state-of-the-art CPU system
  - Data-centric caching with precision-optimized tiling for a heterogeneous memory hierarchy
  - Scalability analysis for both DDR4 and HBM-based FPGA boards
- **Evaluation**
  - NERO outperforms a 16-core IBM POWER9 system by 4.2x and 8.3x when running two compound stencil kernels
  - NERO reduces energy consumption by 22x and 29x with an energy efficiency of 1.5 GFLOPS/Watt and 17.3 GFLOPS/Watt

Outline

- Background
- CPU Roofline Analysis
- FPGA-based Platform
- NERO: Near-HBM Accelerator for Weather Prediction Modeling
- Precision-optimized Tiling
- Evaluation
  - Performance Analysis
  - Energy Efficiency Analysis
- Summary

Stencil Computations and Applications

Stencil computations update values in a grid using a **fixed pattern** of grid points



e.g., 7-point Jacobi in 3D plane

Stencils are used in **~30% of high-performance computing applications**

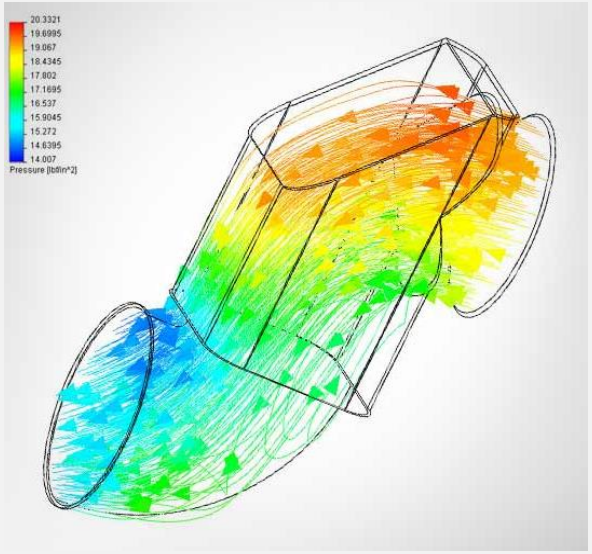


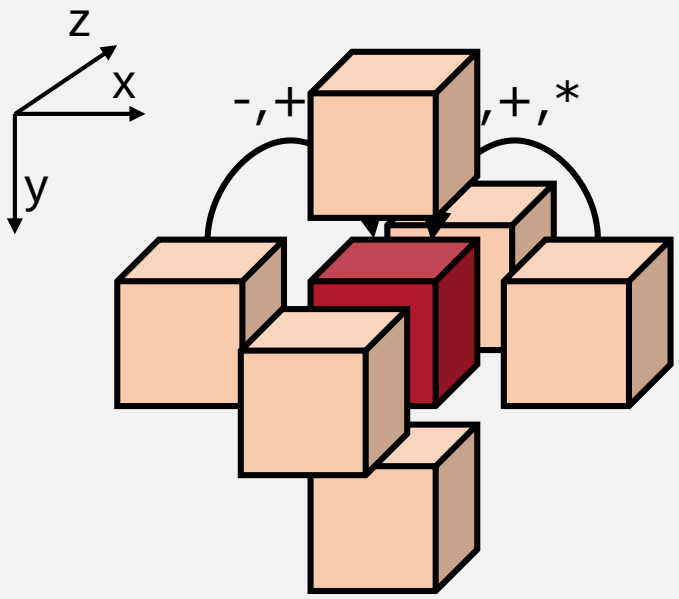
Image sources: http://www.flometrics.com/fluid-dynamics/computational-fluid-dynamics

Naoe, Kensuke et al. "Secure Key Generation for Static Visual Watermarking by Machine Learning in Intelligent Systems and Services" IJSSOE, 2010

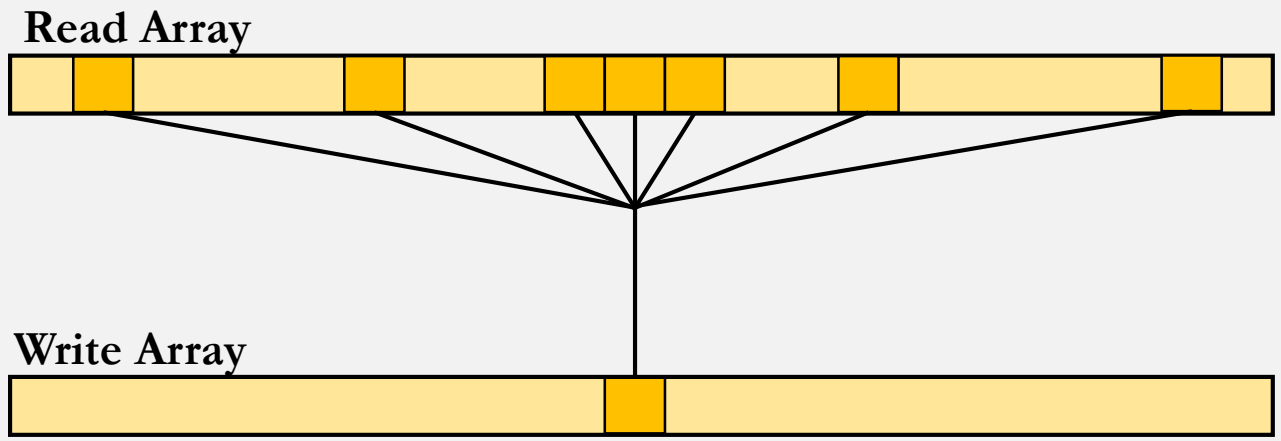
Stencil Characteristics

High-order stencil computations are cache unfriendly

- Limited arithmetic intensity
- Sparse and complex access pattern



e.g., 7-point Jacobi in 3D plane



Mapping of 7-point Jacobi from 3D plane onto 1D plane

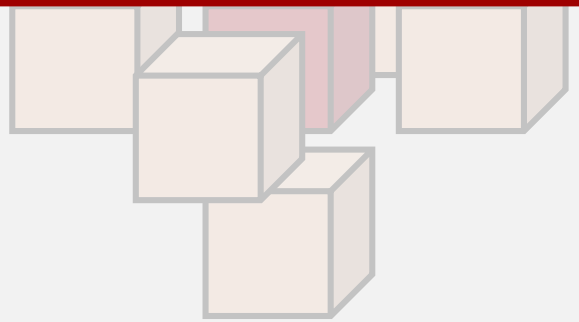
Image source: Xu, Jingheng et al. "Performance Tuning and Analysis for Stencil-Based Applications on POWER8 Processor" ACM TACO, 2018

Stencil Characteristics

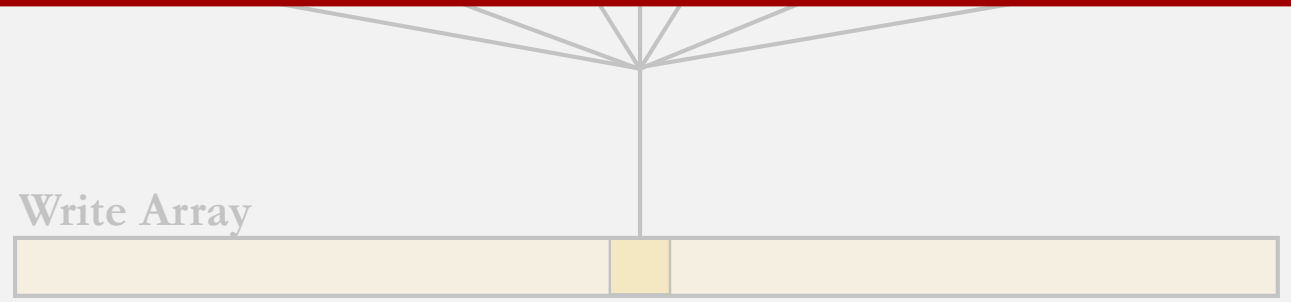
High-order stencil computations are cache unfriendly

- Limited arithmetic intensity
- Sparse and complex access pattern

Performance bottleneck



e.g., 7-point Jacobi in 3D plane



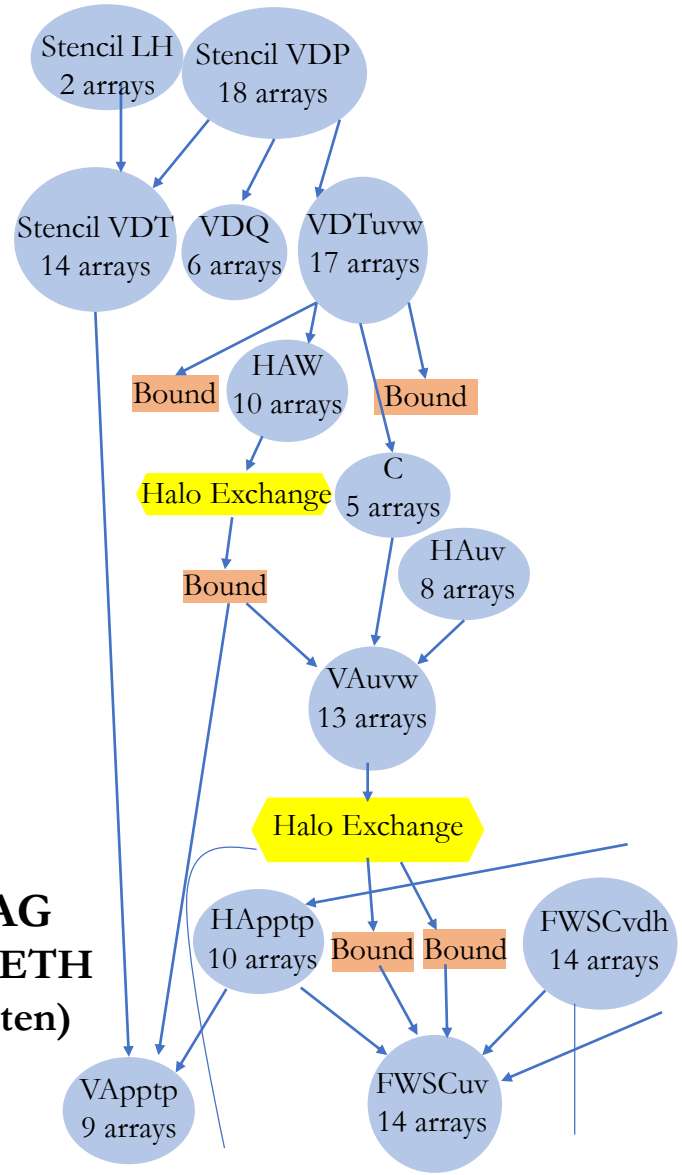
Mapping of 7-point Jacobi from 3D plane onto 1D plane

Image source: Xu, Jingheng et al. "Performance Tuning and Analysis for Stencil-Based Applications on POWER8 Processor" ACM TACO, 2018

Stencil Computations in Weather Applications

COSMO (Consortium for Small-Scale Modeling) weather prediction application

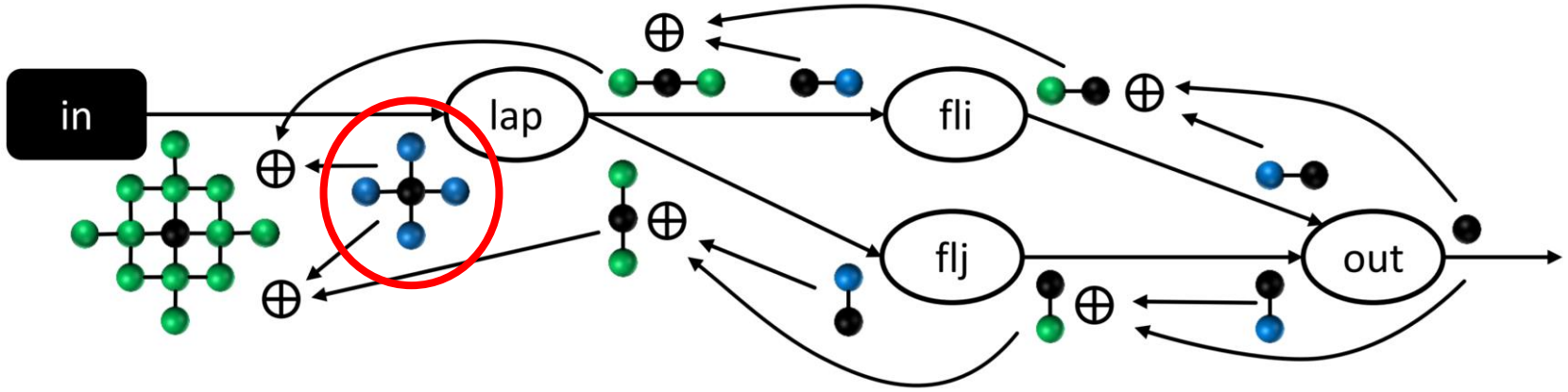
- The essential part of the weather prediction models is called **dynamical core**
- Around **80 different** stencil compute motifs
- ~30 variables and ~70 temporary arrays (3D grids)
- Horizontal diffusion and vertical advection
- **Complex stencil programs**

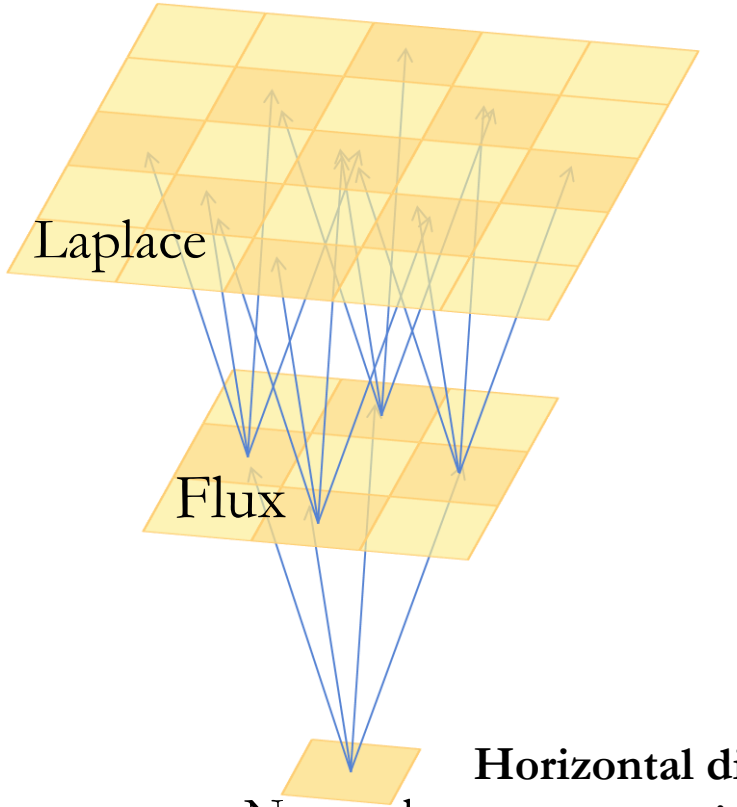


Section of COSMO CDAG (Courtesy CSCS/ETH and Ronald Luijten)

Example Complex Stencil: Horizontal Diffusion

- Compound stencil kernel consists of a **collection** of elementary stencil kernels
- Iterates over a 3D grid performing **Laplacian** and **flux** operations
- **Complex** memory access behavior and **low** arithmetic intensity





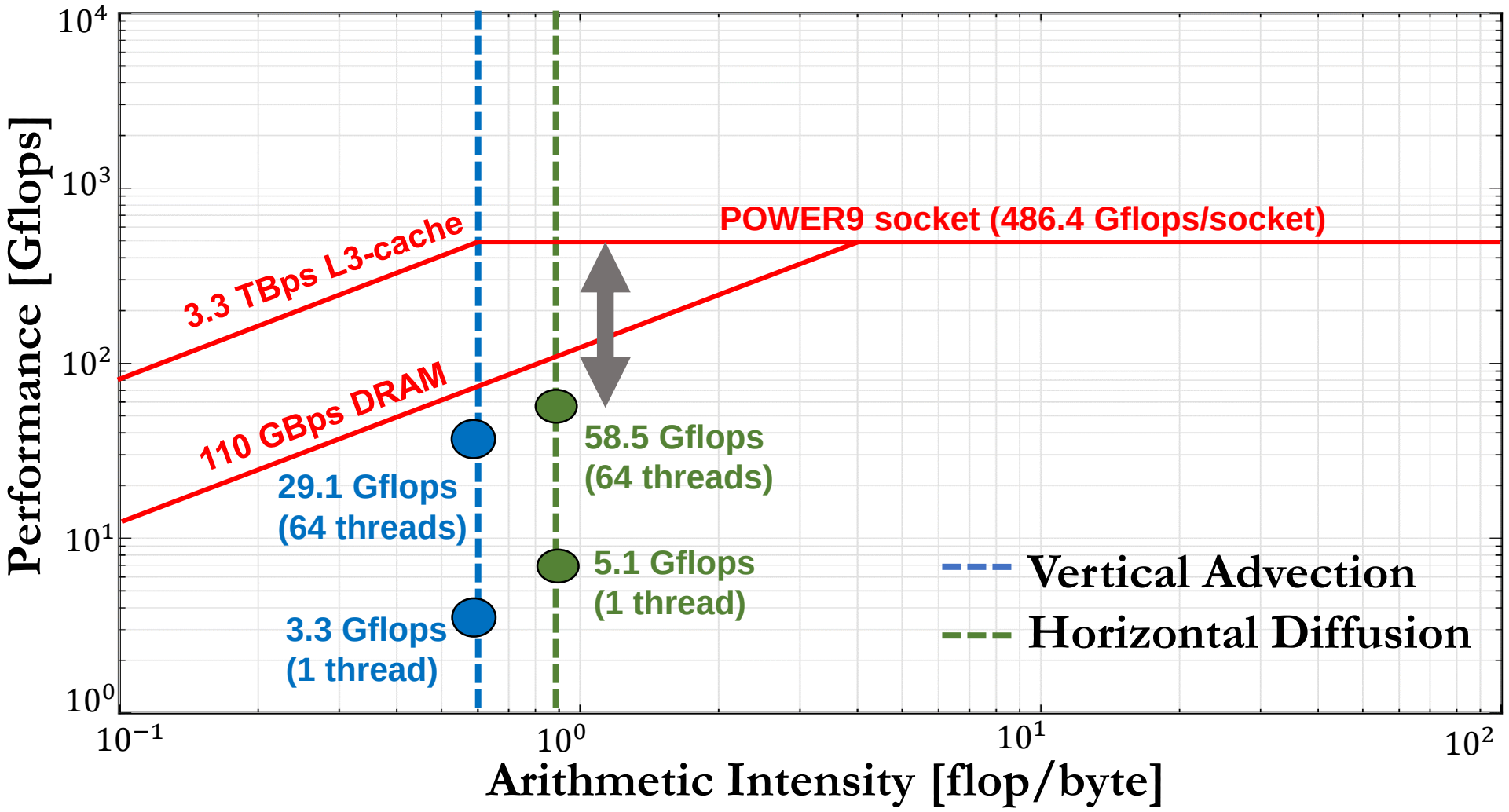
Horizontal diffusion composition (Courtesy CSCS/ETH and Ronald Luijten)

DFG source: Gysi, Tobias, et al. "MODESTO: Data-centric Analytic Optimization of Complex Stencil Programs on Heterogeneous Architectures" ICS, 2015

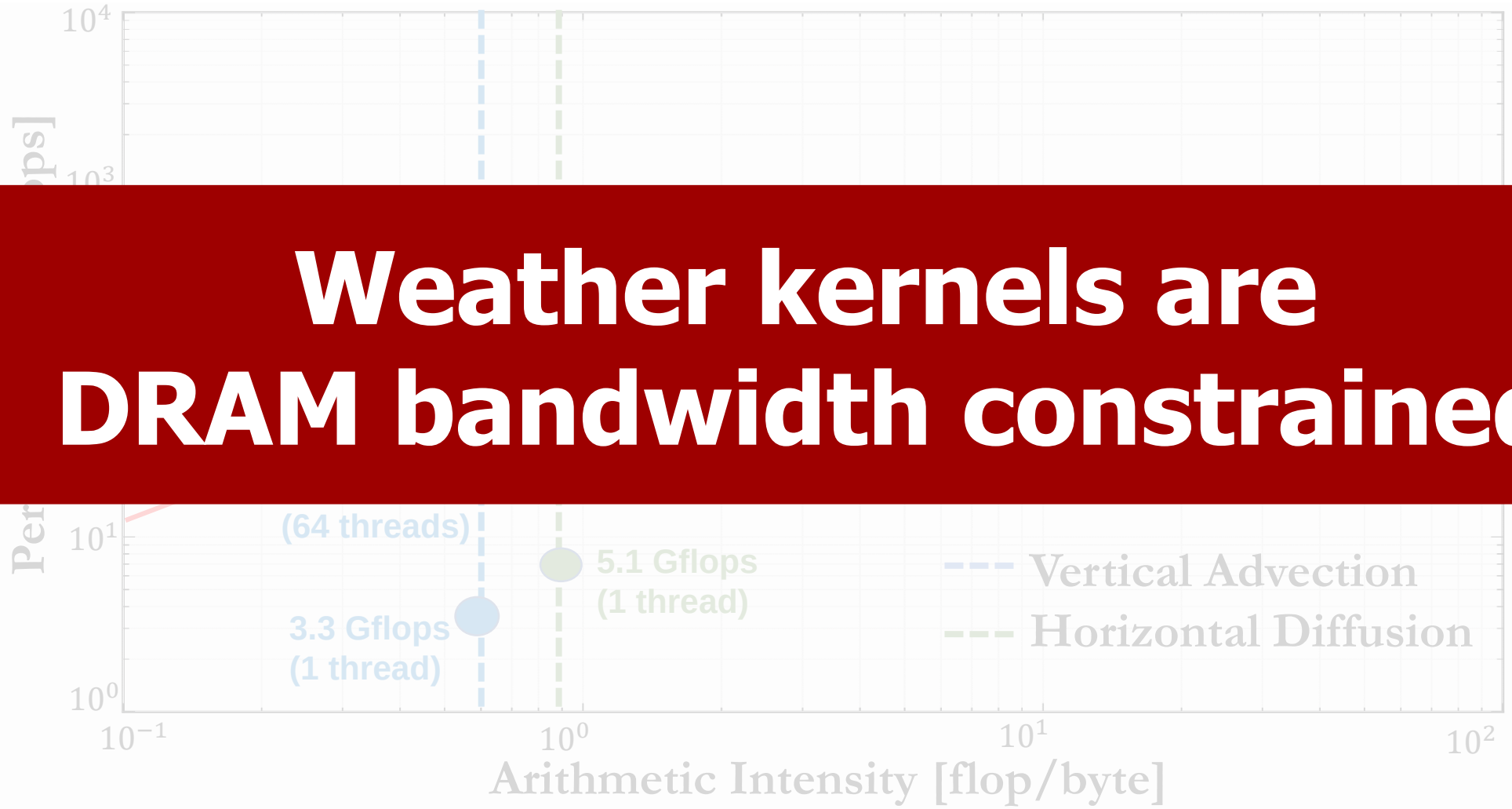
Outline

- Background
- CPU Roofline Analysis
- FPGA-based Platform
- NERO: Near-HBM Accelerator for Weather Prediction Modeling
- Precision-optimized Tiling
- Evaluation
  - Performance Analysis
  - Energy Efficiency
- Summary

IBM POWER9 Roofline Analysis



IBM POWER9 Roofline Analysis



Weather kernels are DRAM bandwidth constrained

Outline

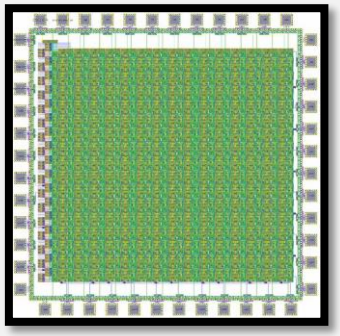
- Background
- CPU Roofline Analysis
- FPGA-based Platform
- NERO: Near-HBM Accelerator for Weather Prediction Modeling
- Precision-optimized Tiling
- Evaluation
  - Performance Analysis
  - Energy Efficiency Analysis
- Summary

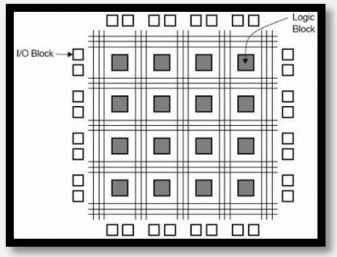
Silicon Alternatives

CPUs

GPUs

FPGAs





ASICs

FLEXIBILITY

EFFICIENCY

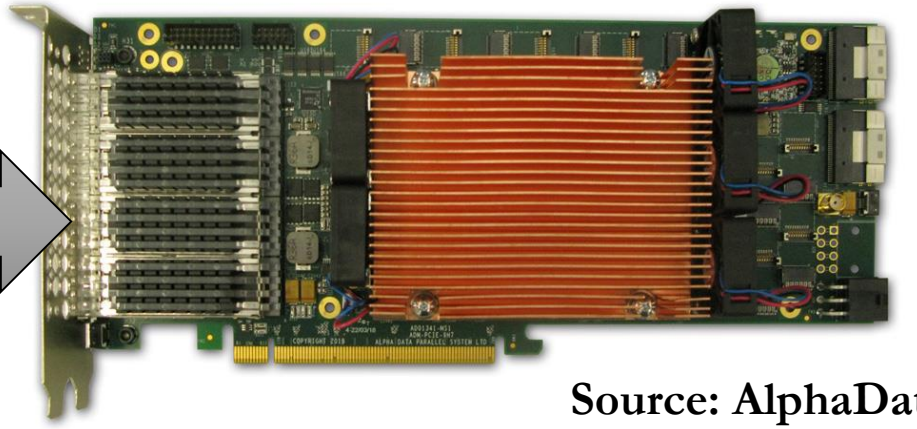
FPGAs are highly configurable!

Heterogeneous System: CPU+FPGA

Source: IBM

POWER9 AC922

CAPI2



Source: AlphaData

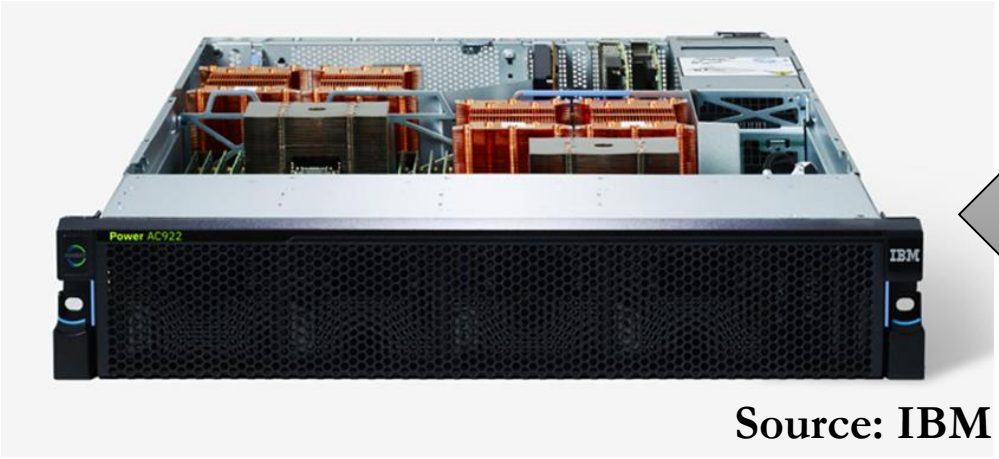
HBM-based AD9H7 board

We evaluate two POWER9+FPGA systems:

1. HBM-based board AD9H7

Xilinx Virtex Ultrascale+™ XCVU37P-2

Heterogeneous System: CPU+FPGA



Source: IBM

CAPI2

Source: AlphaData

POWER9 AC922

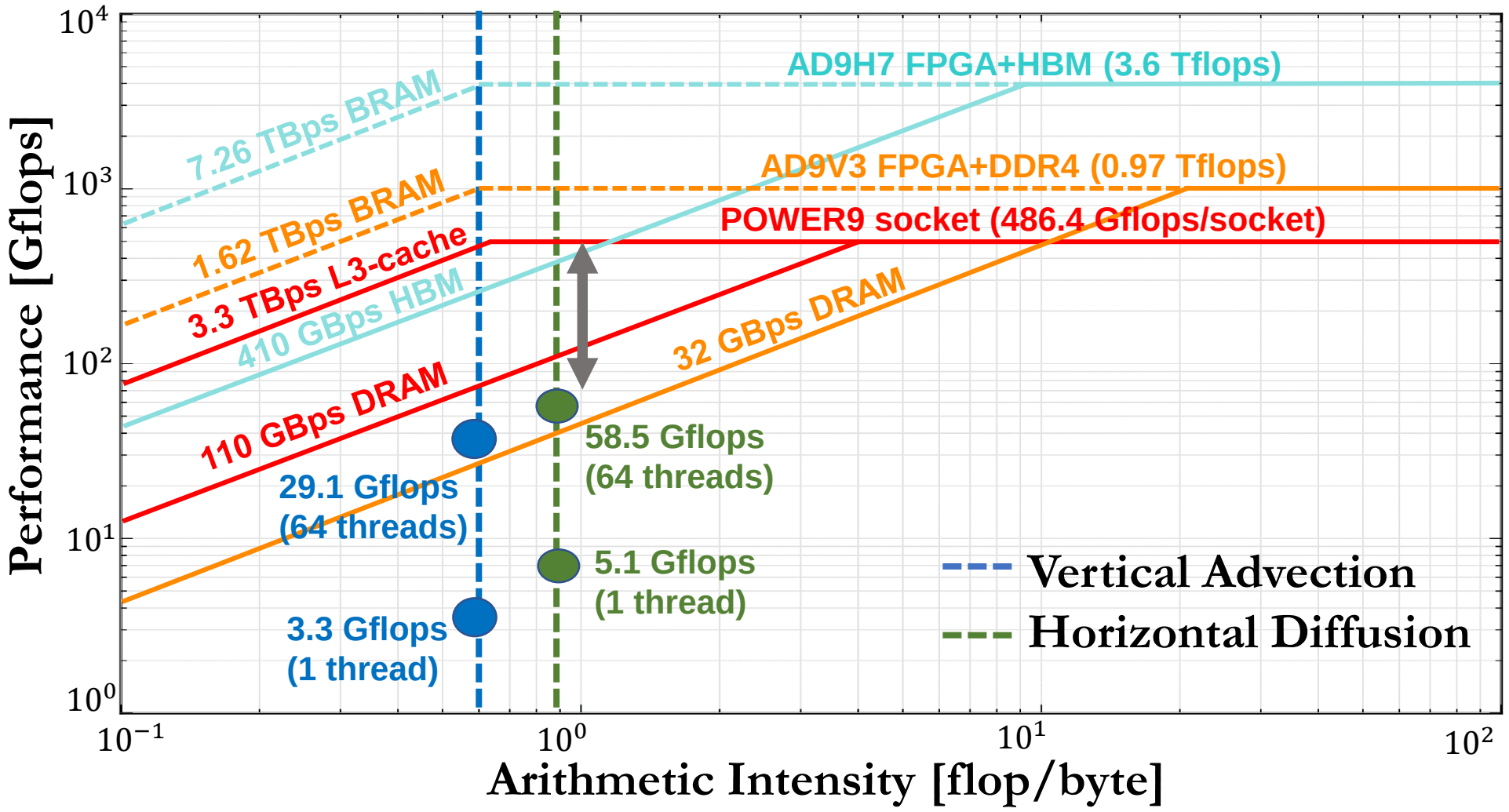
DDR4-based AD9V3 board

We evaluate two POWER9+FPGA systems:

1. HBM-based board AD9H7
Xilinx Virtex Ultrascale+™ XCVU37P-2

2. DDR4-based board AD9V3
Xilinx Virtex Ultrascale+™ XCVU3P-2

FPGAs Have Tremendous Potential



Outline

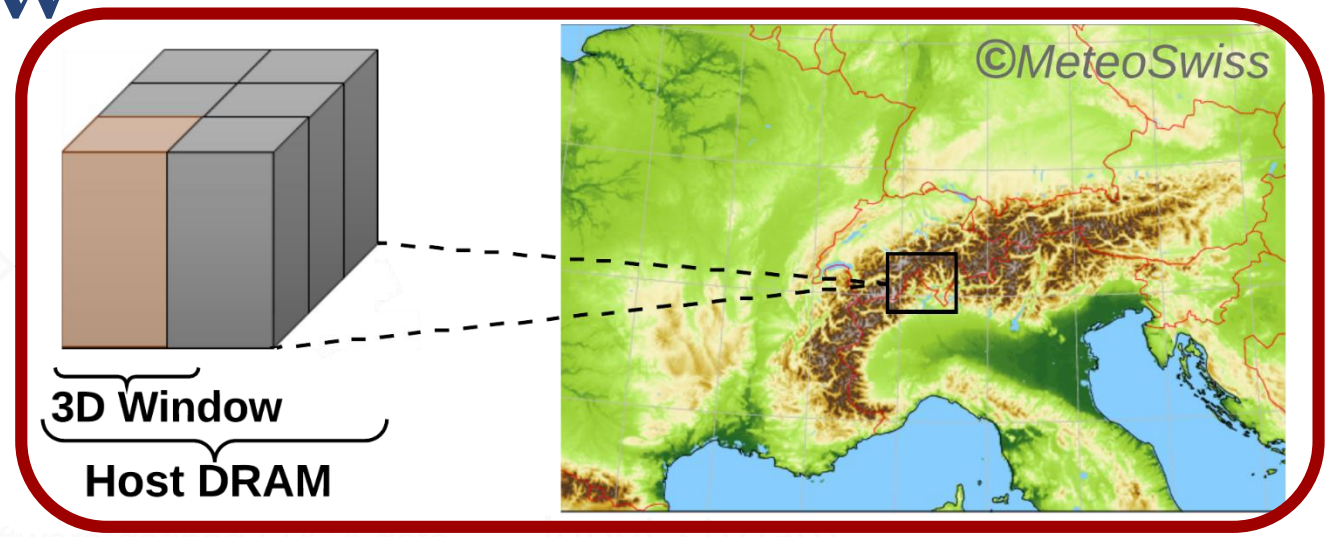
- Background
- CPU Roofline Analysis
- FPGA-based Platform
- **NERO: Near-HBM Accelerator for Weather Prediction Modeling**
- Precision-optimized Tiling
- Evaluation
  - Performance Analysis
  - Energy Efficiency Analysis
- Summary

NERO: A Near High-Bandwidth Memory Stencil Accelerator for Weather Prediction Modeling

- First **near-HBM FPGA-based** accelerator for representative kernels from a **real-world weather prediction application**
- Data-centric caching with **precision-optimized tiling** for a heterogeneous memory hierarchy
- In-depth **scalability analysis** for both DDR4 and HBM-based FPGA boards

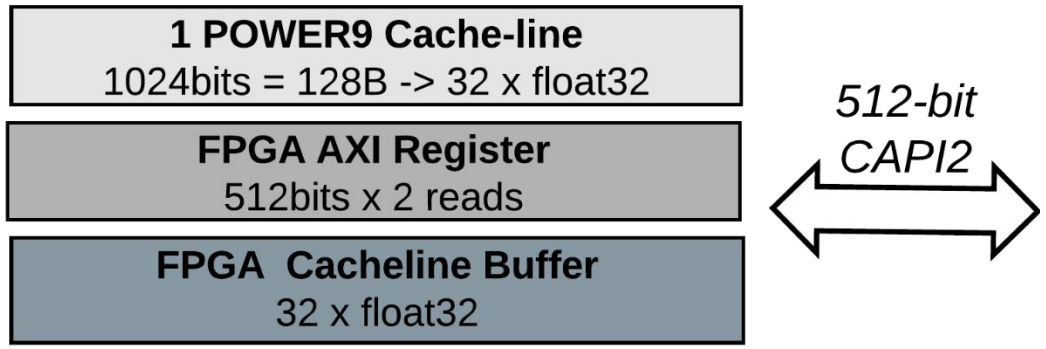
NERO Design Flow

NERO Design Flow



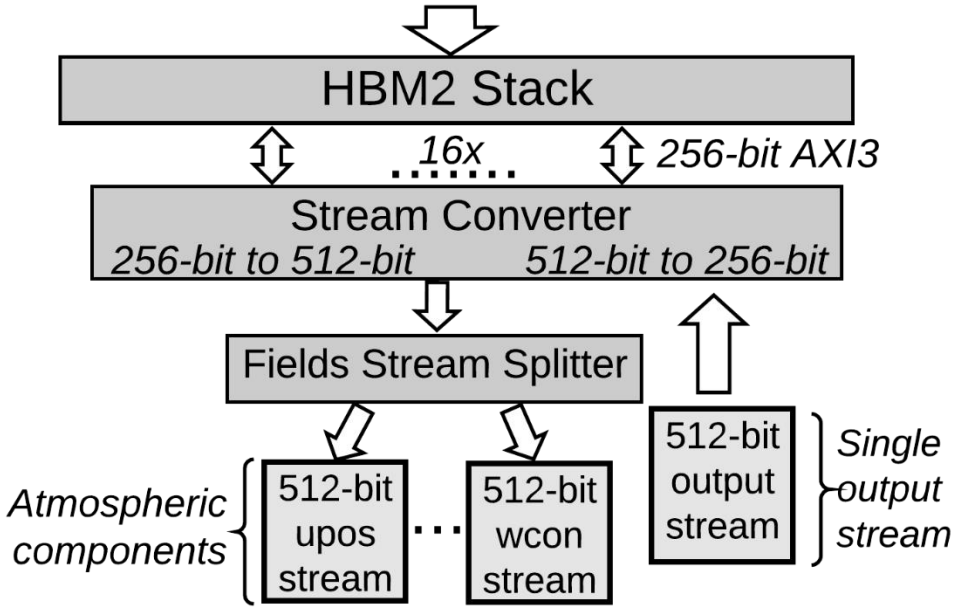
Weather data in the host DRAM

NERO Design Flow



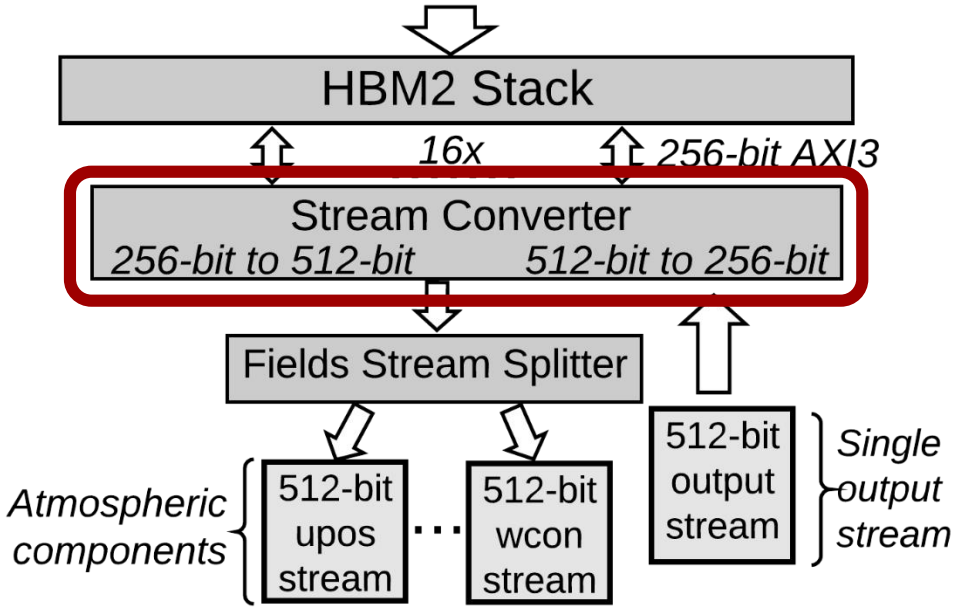
Cache-line transfer over CAPI2

NERO Design Flow



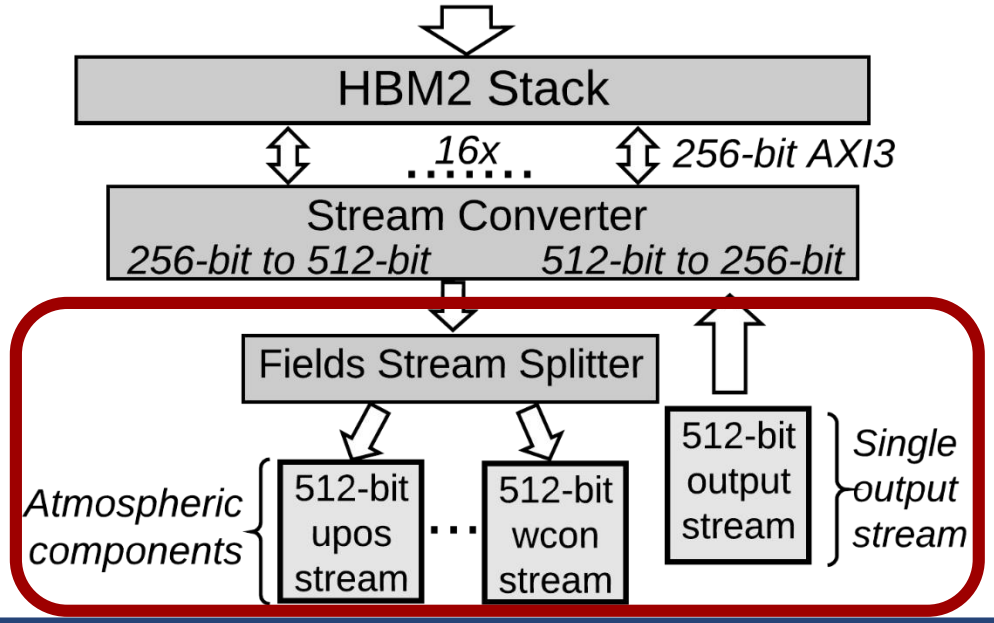
Data mapping onto HBM

NERO Design Flow



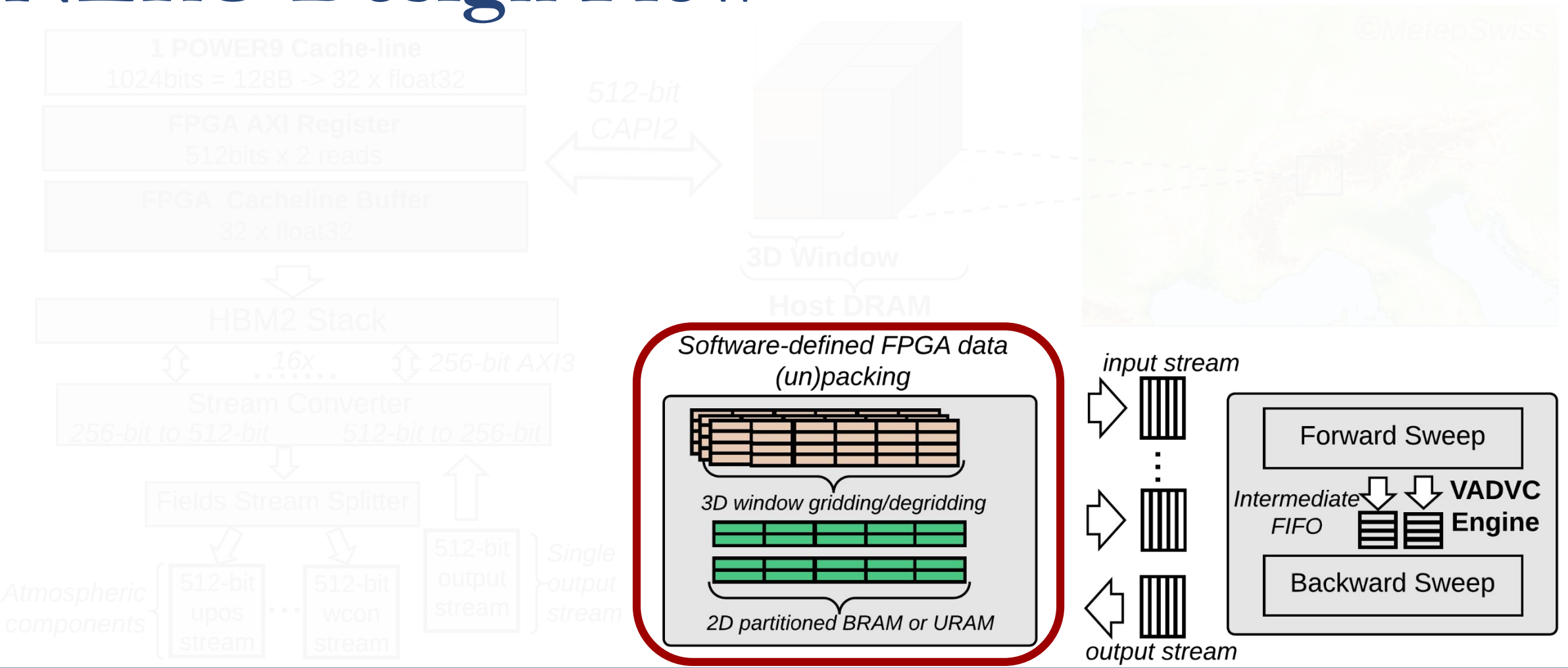
Data mapping onto HBM

NERO Design Flow



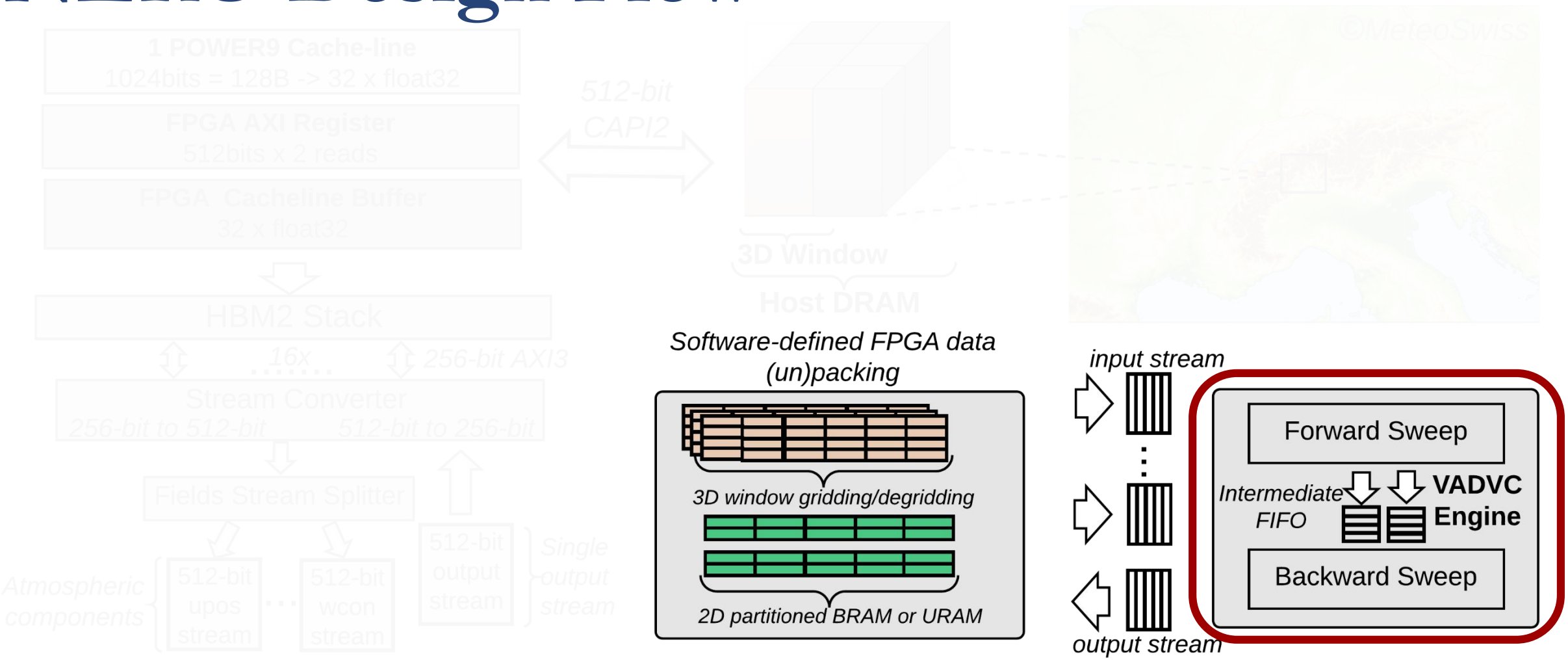
Data mapping onto HBM

NERO Design Flow



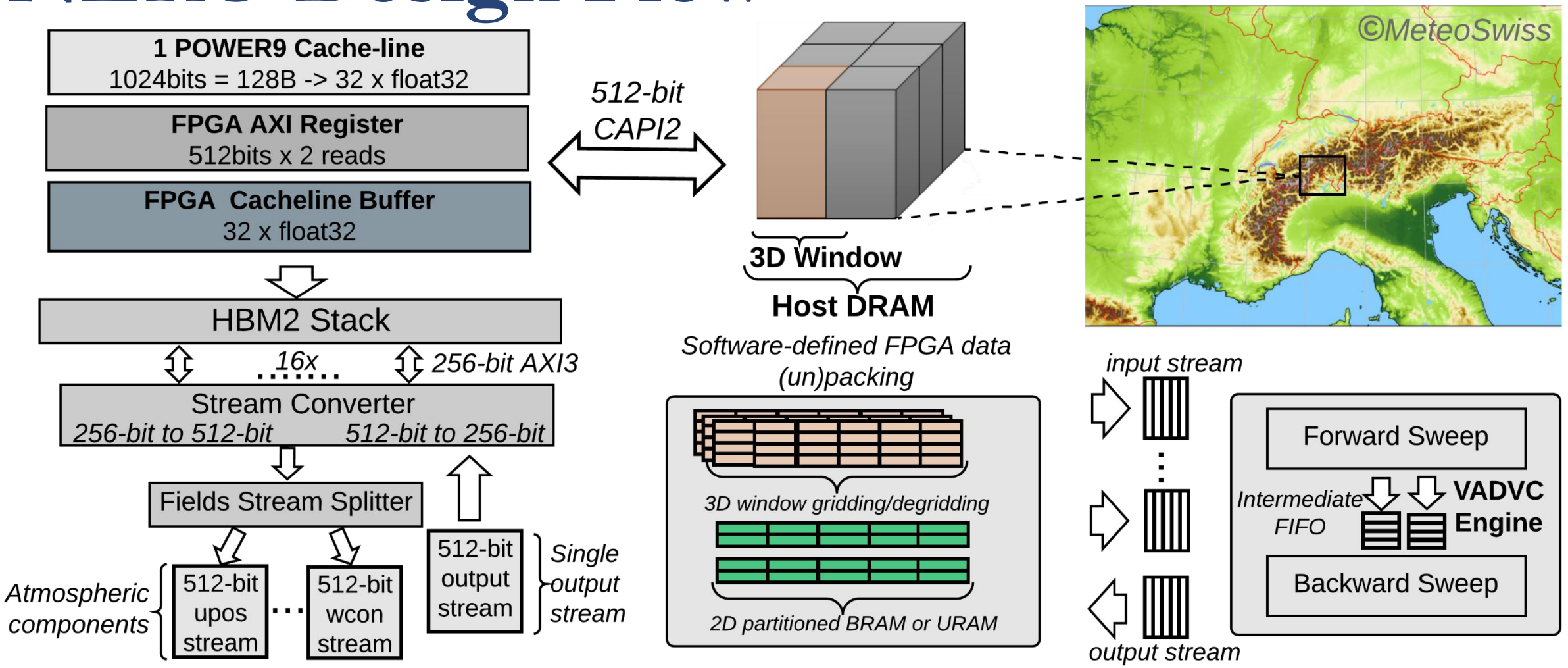
Main execution pipeline

NERO Design Flow



Main execution pipeline

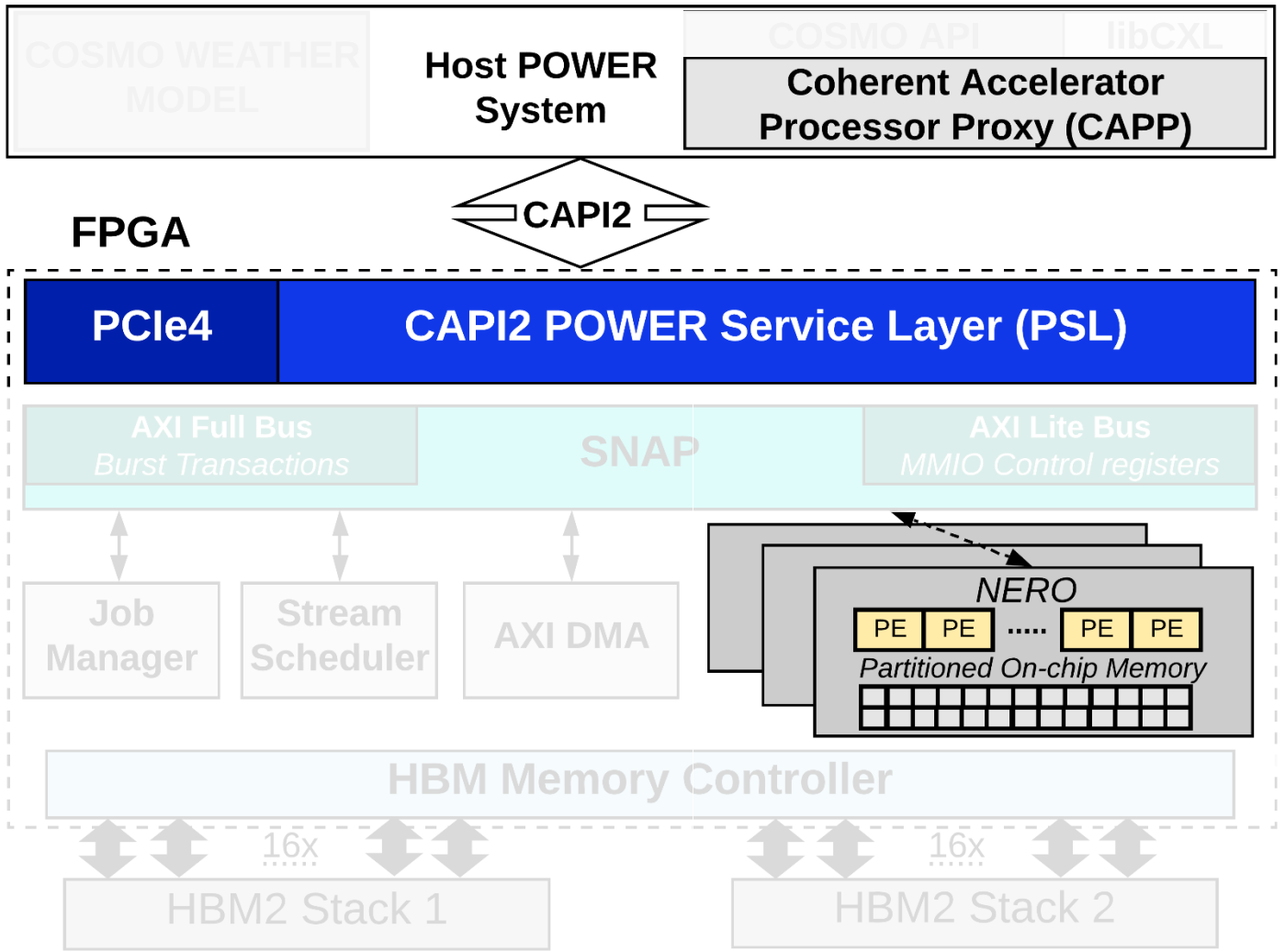
NERO Design Flow



Complete design flow

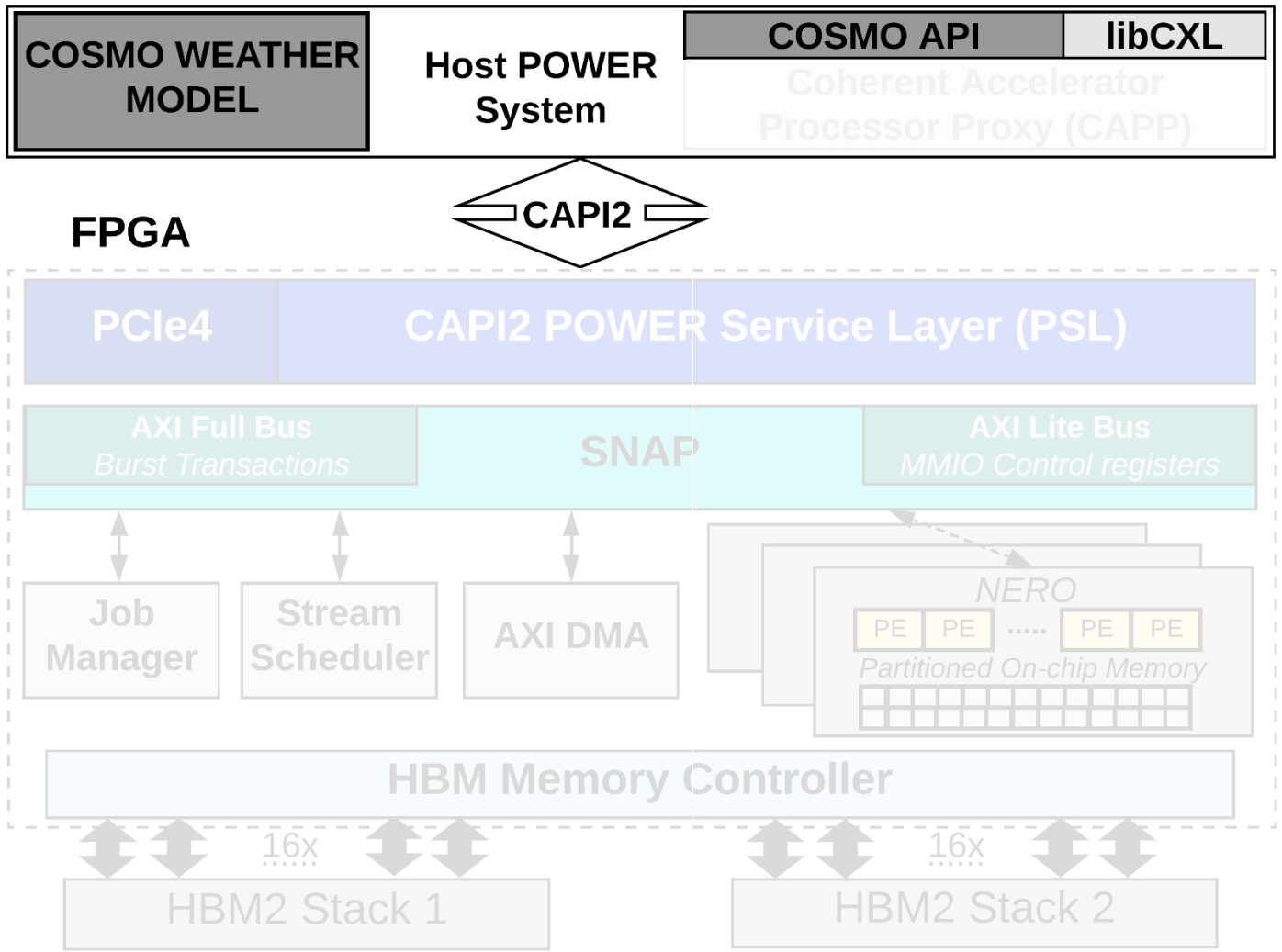
NERO Application Framework

- NERO communicates to Host over **CAPI2** (Coherent Accelerator Processor Interface)



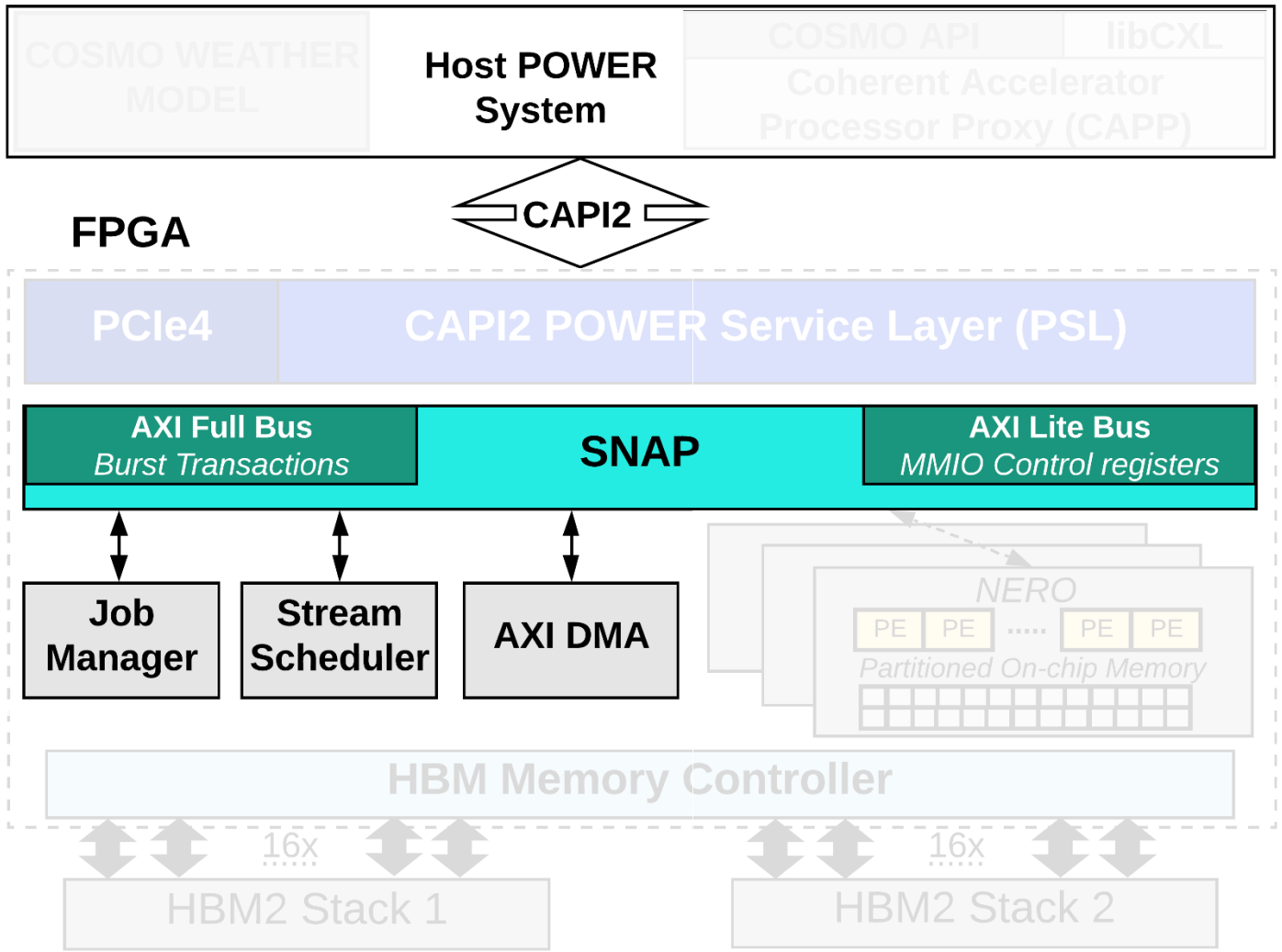
NERO Application Framework

- NERO communicates to Host over **CAPI2** (Coherent Accelerator Processor Interface)
- **COSMO API** handles offloading jobs to NERO



NERO Application Framework

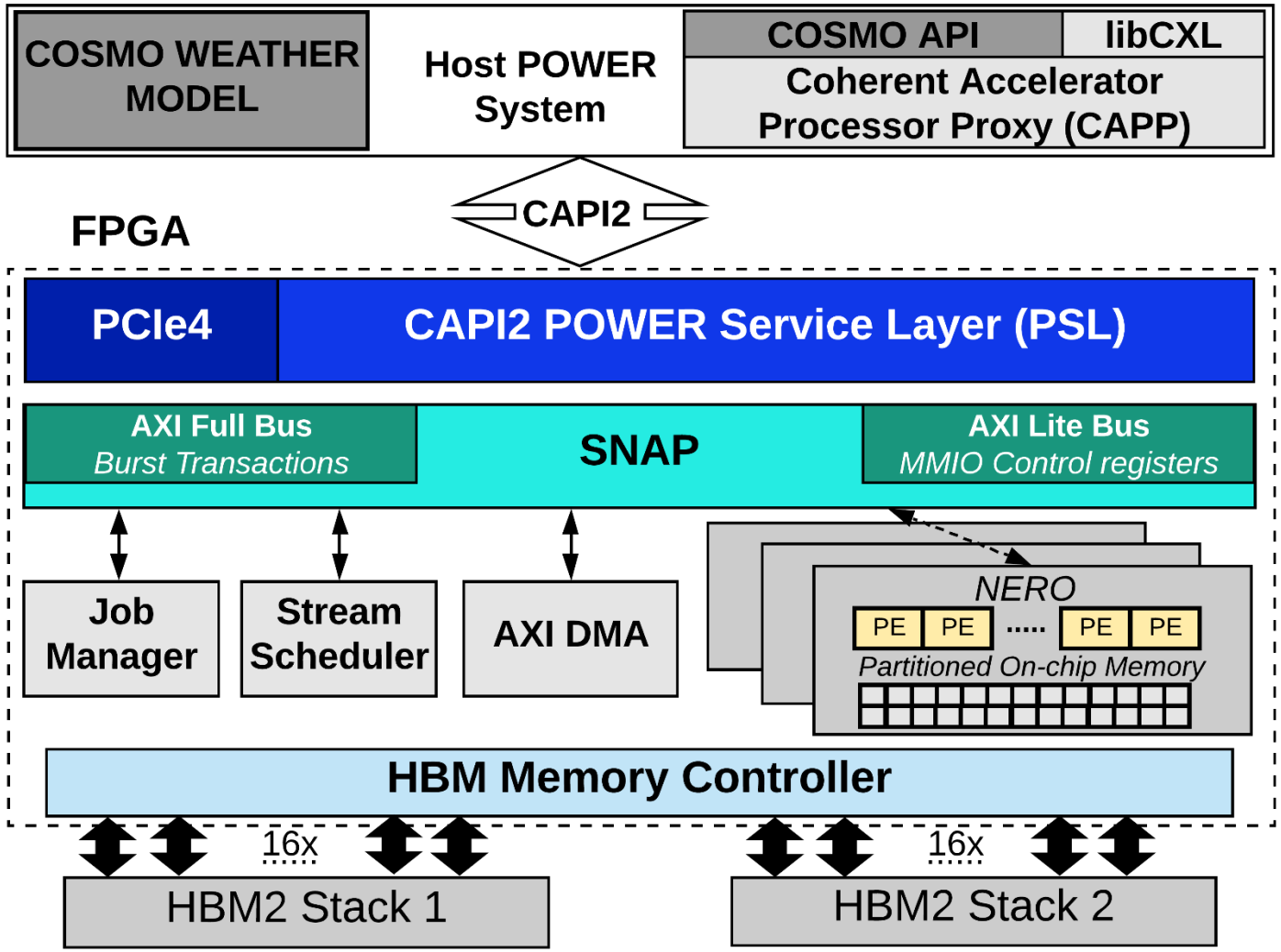
- NERO communicates to Host over **CAPI2** (Coherent Accelerator Processor Interface)
- **COSMO API** handles offloading jobs to NERO
- **SNAP** (Storage, Network, and Analytics Programming) allows for seamless integration of the COSMO API



https://github.com/open-power/snap

NERO Application Framework

- NERO communicates to Host over **CAPI2** (Coherent Accelerator Processor Interface)
- **COSMO API** handles offloading jobs to NERO
- **SNAP** (Storage, Network, and Analytics Programming) allows for seamless integration of the COSMO API



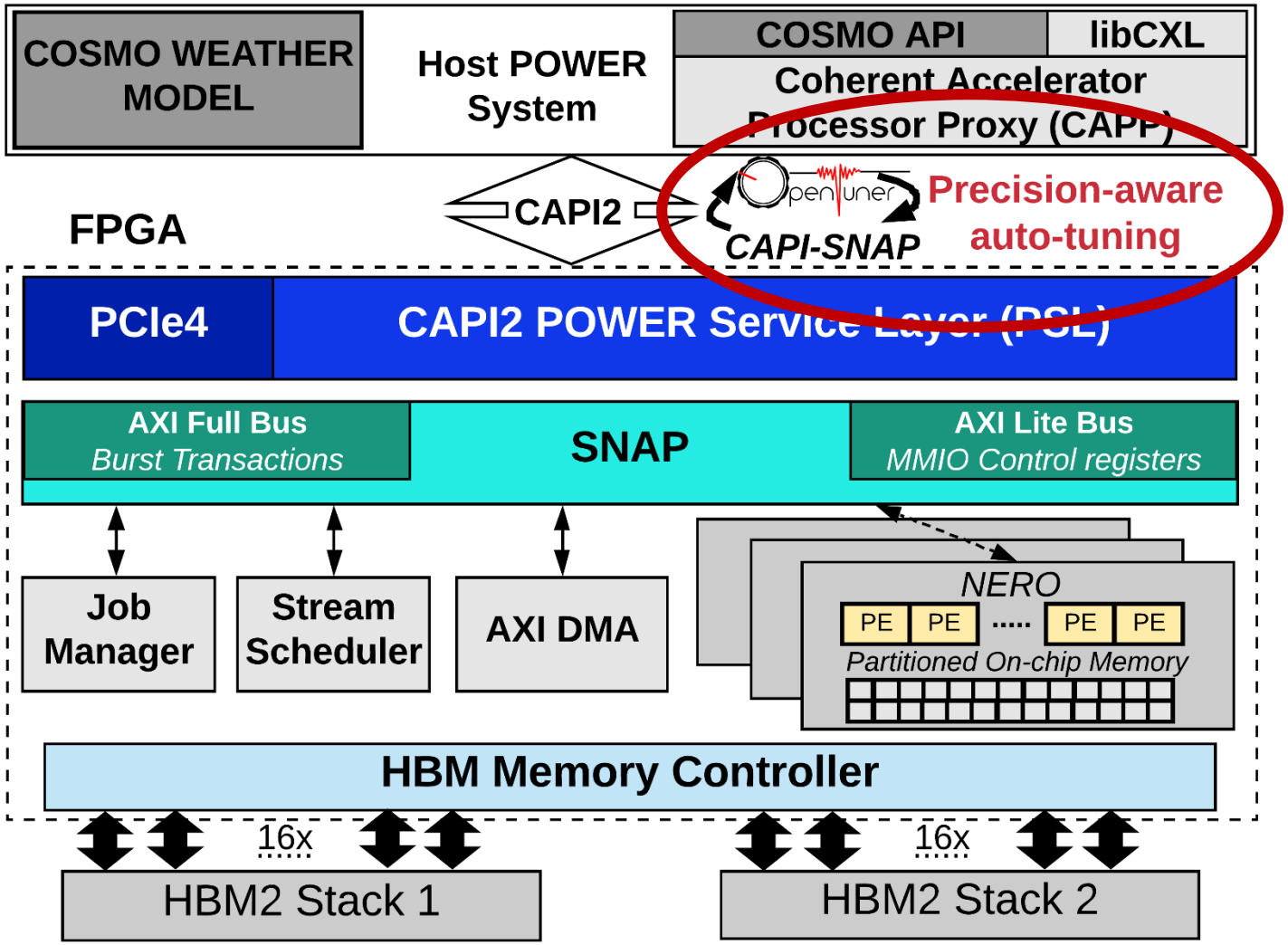
https://github.com/open-power/snap

Outline

- Background
- CPU Roofline Analysis
- FPGA-based Platform
- NERO: Near-HBM Accelerator for Weather Prediction Modeling
- **Precision-optimized Tiling**
- Evaluation
  - Performance Analysis
  - Energy Efficiency Analysis
- Summary

Precision-optimized Tiling

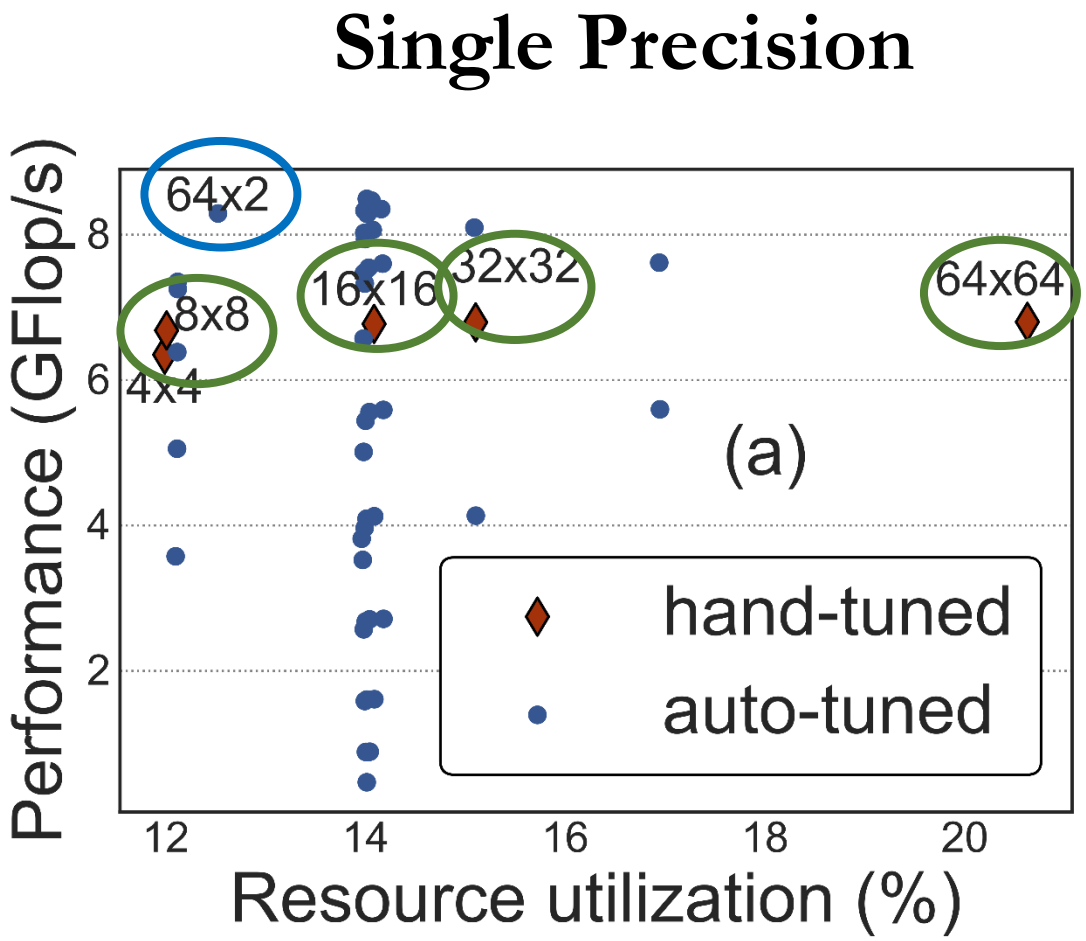
- The **best window size** is **critical**
- Formulate the search for the best window size as a multi-objective **auto-tuning** problem
- Taking into account the **datatype precision**
- We make use of **OpenTuner**



https://github.com/jansel/opentuner

Precision-optimized Tiling

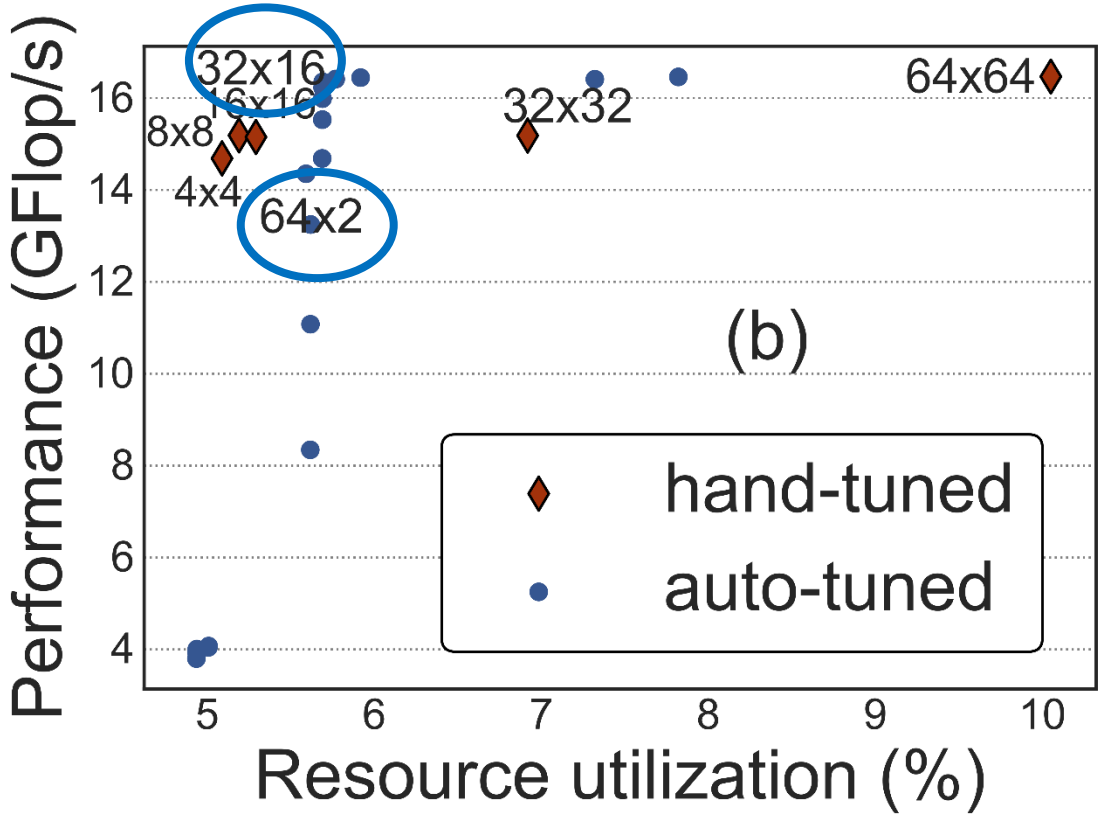
Precision-optimized Tiling



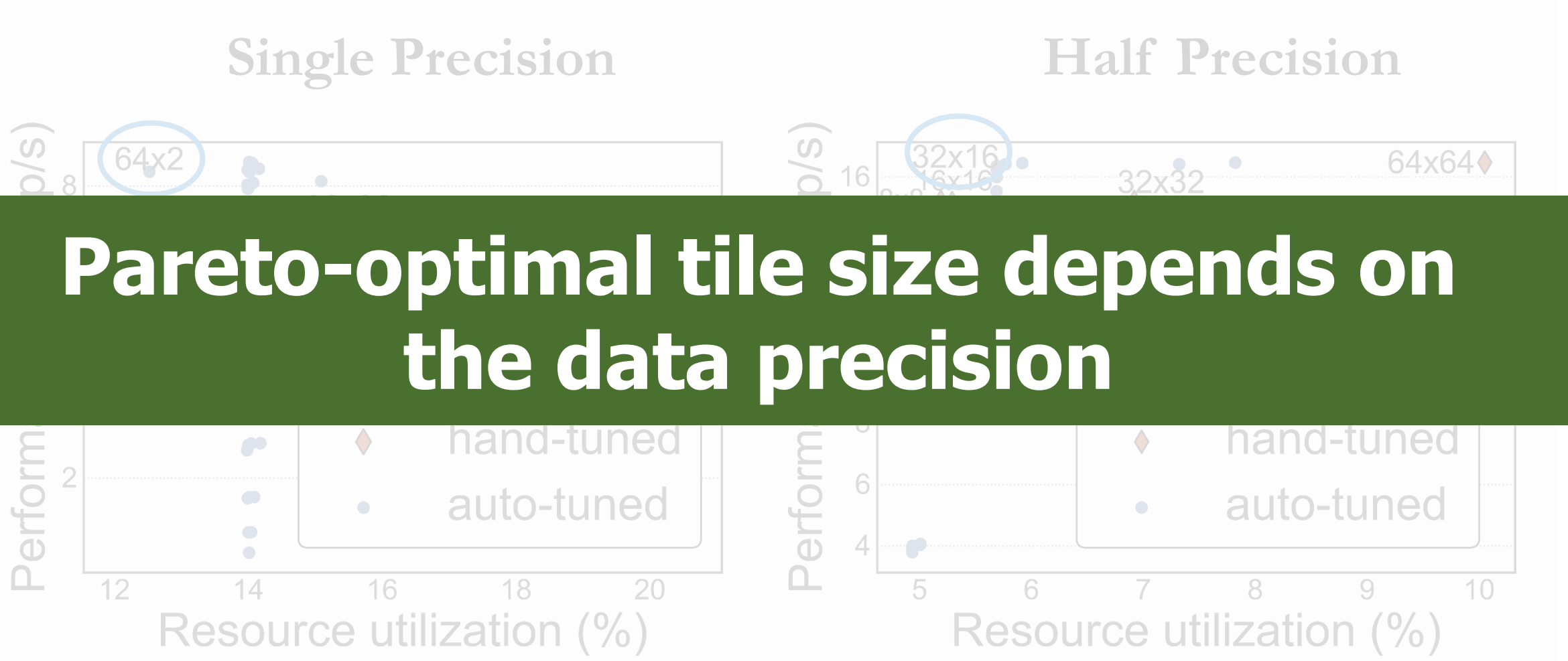
Precision-optimized Tiling

Single Precision

Half Precision



Precision-optimized Tiling

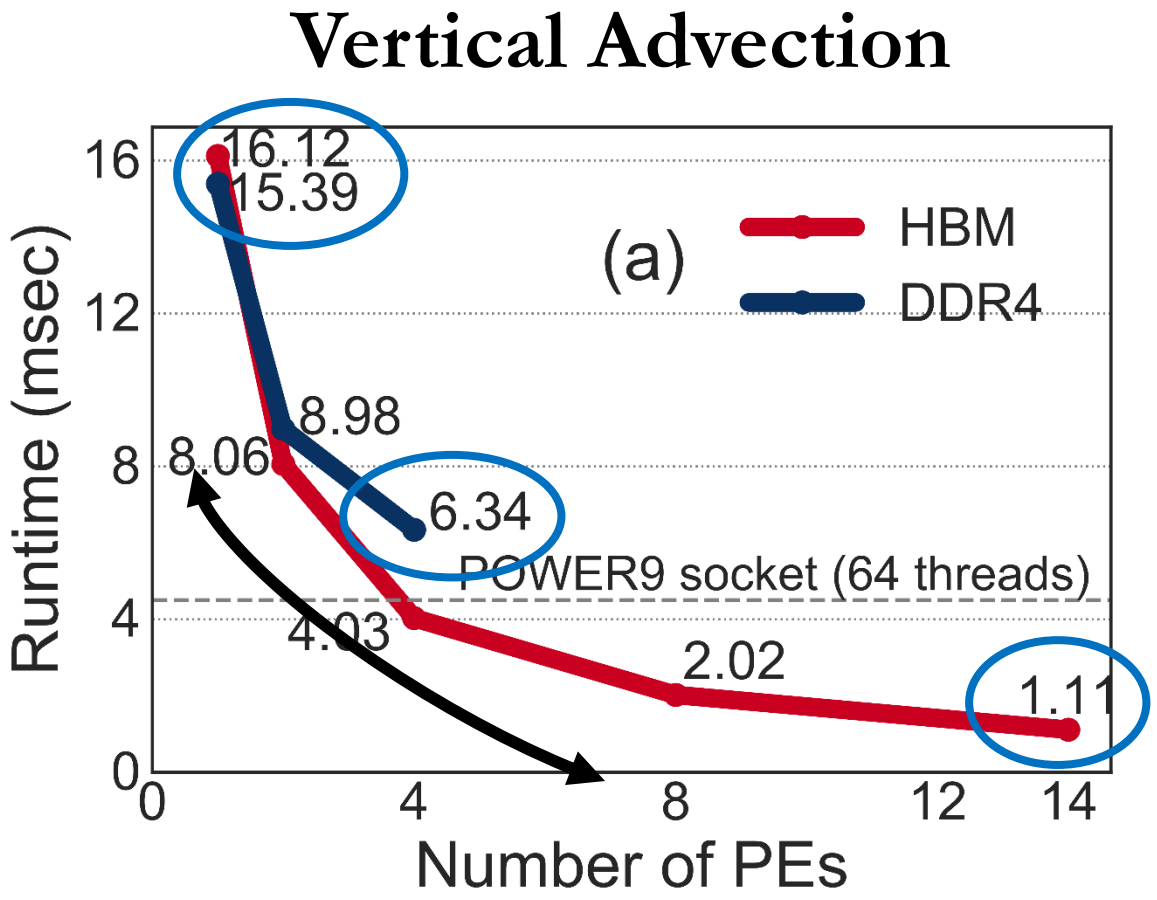


Pareto-optimal tile size depends on the data precision

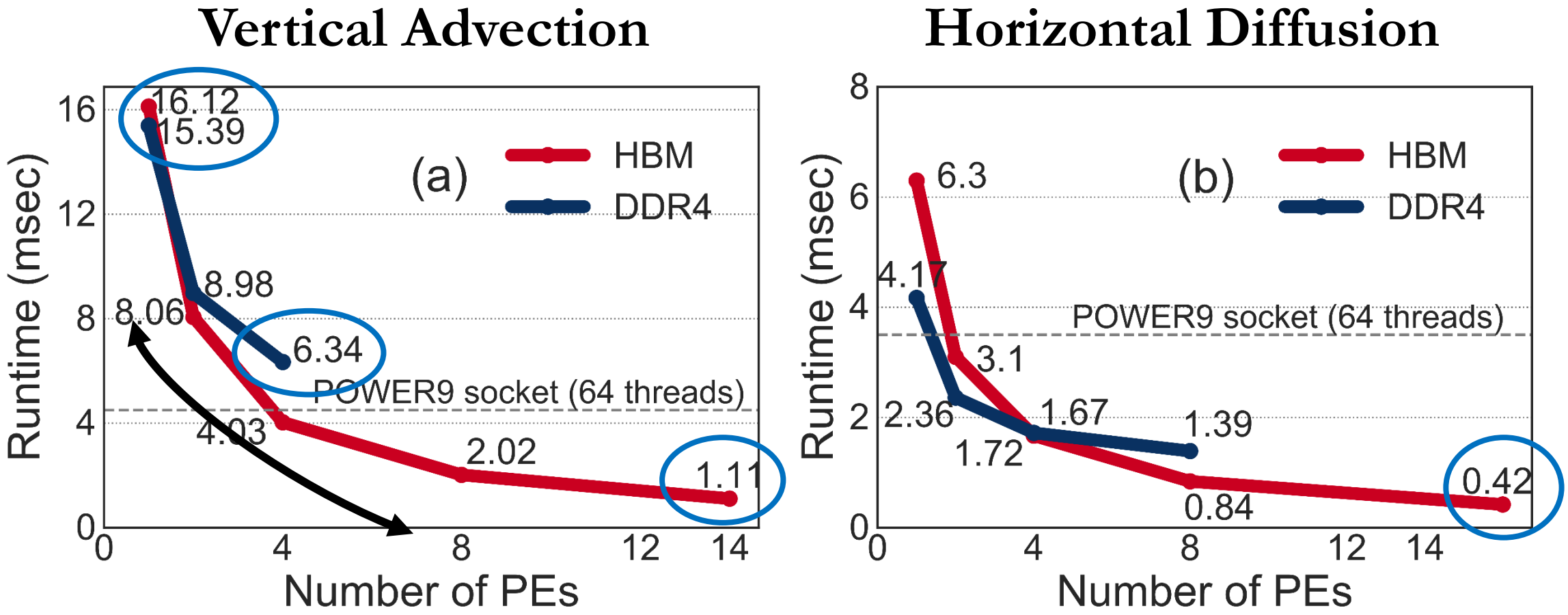
Outline

- Background
- CPU Roofline Analysis
- FPGA-based Platform
- NERO: Near-HBM Accelerator for Weather Prediction Modeling
- Precision-optimized Tiling
- Evaluation
  - Performance Analysis
  - Energy Efficiency Analysis
- Summary

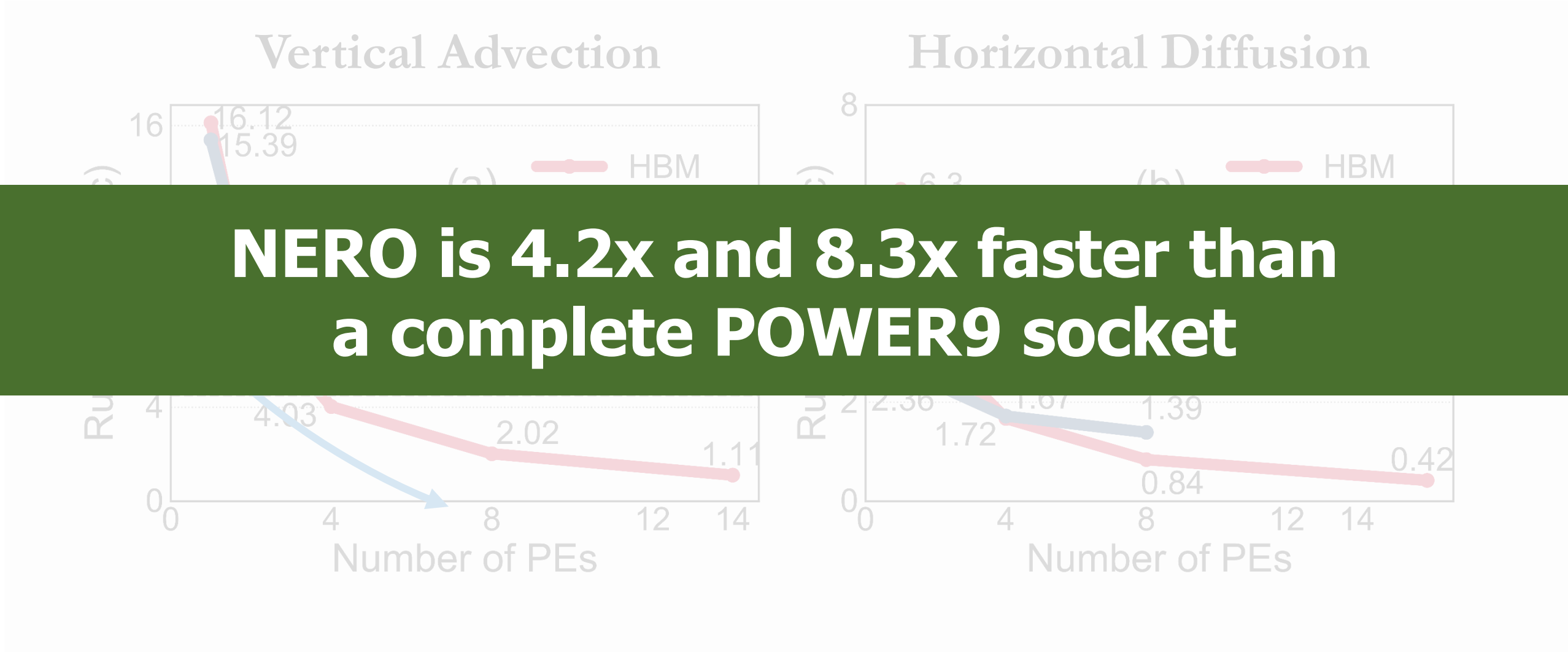
NERO Performance Analysis



NERO Performance Analysis



NERO Performance Analysis

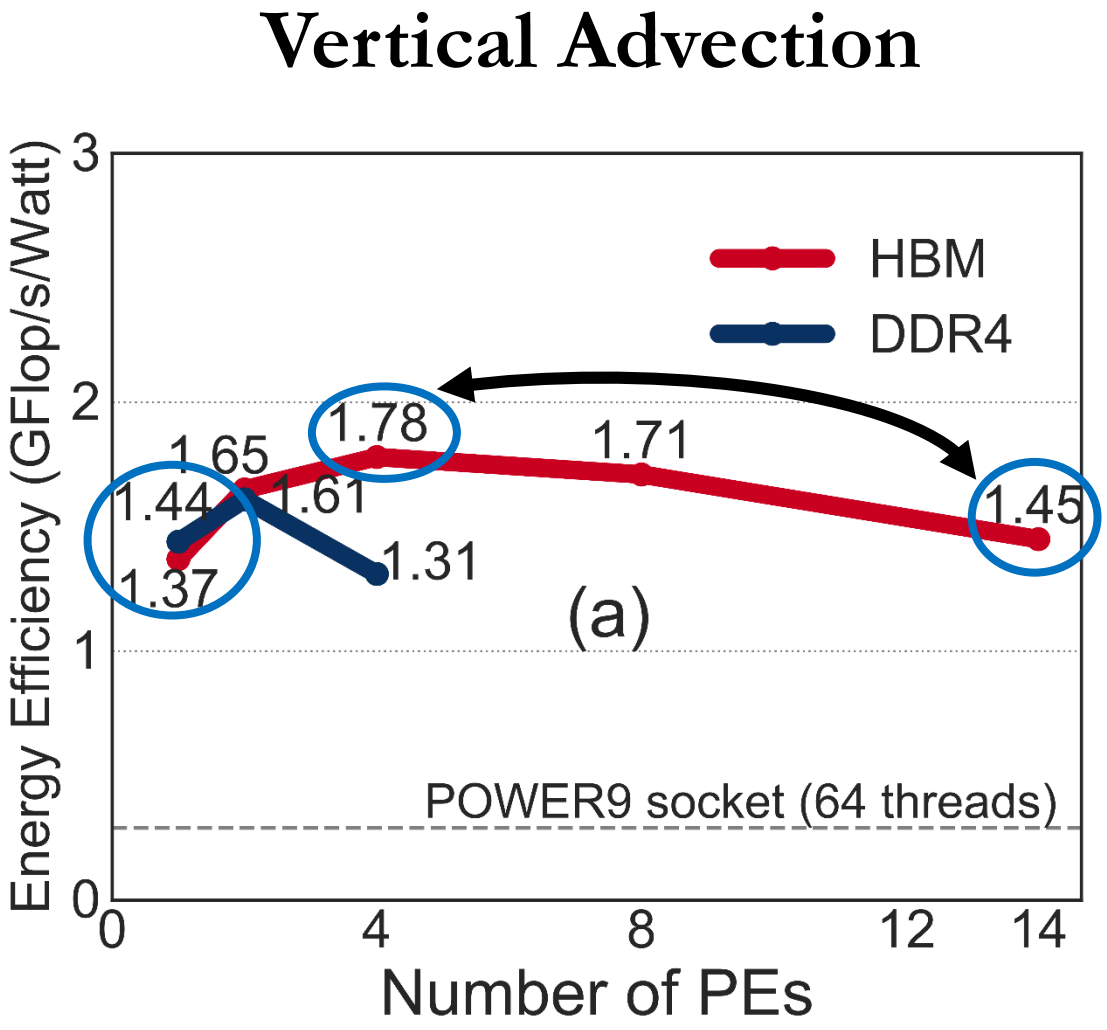


NERO is 4.2x and 8.3x faster than a complete POWER9 socket

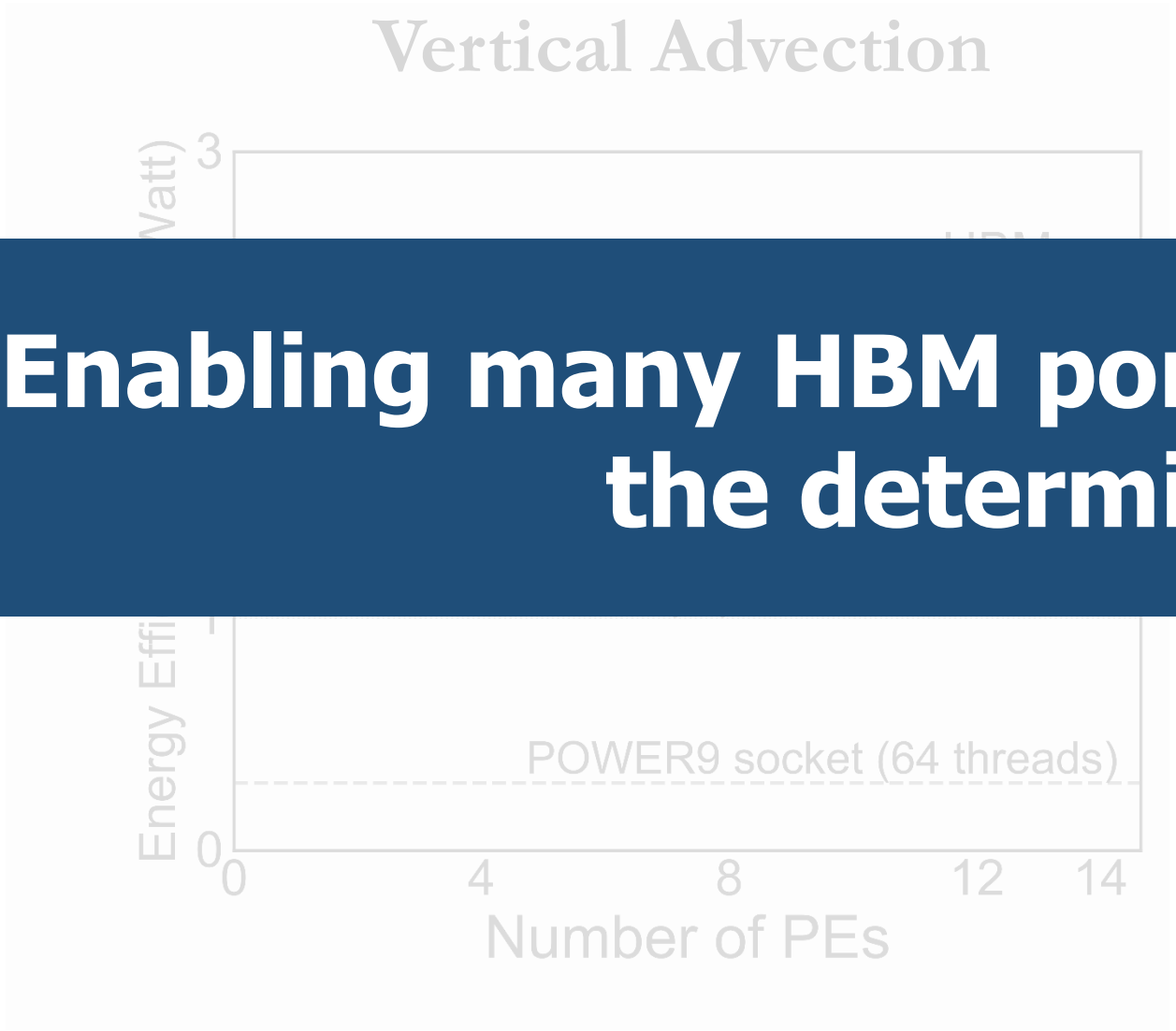
Outline

- Background
- CPU Roofline Analysis
- FPGA-based Platform
- NERO: Near-HBM Accelerator for Weather Prediction Modeling
- Precision-optimized Tiling
- Evaluation
  - Performance Analysis
  - Energy Efficiency Analysis
- Summary

How Energy Efficient is NERO?

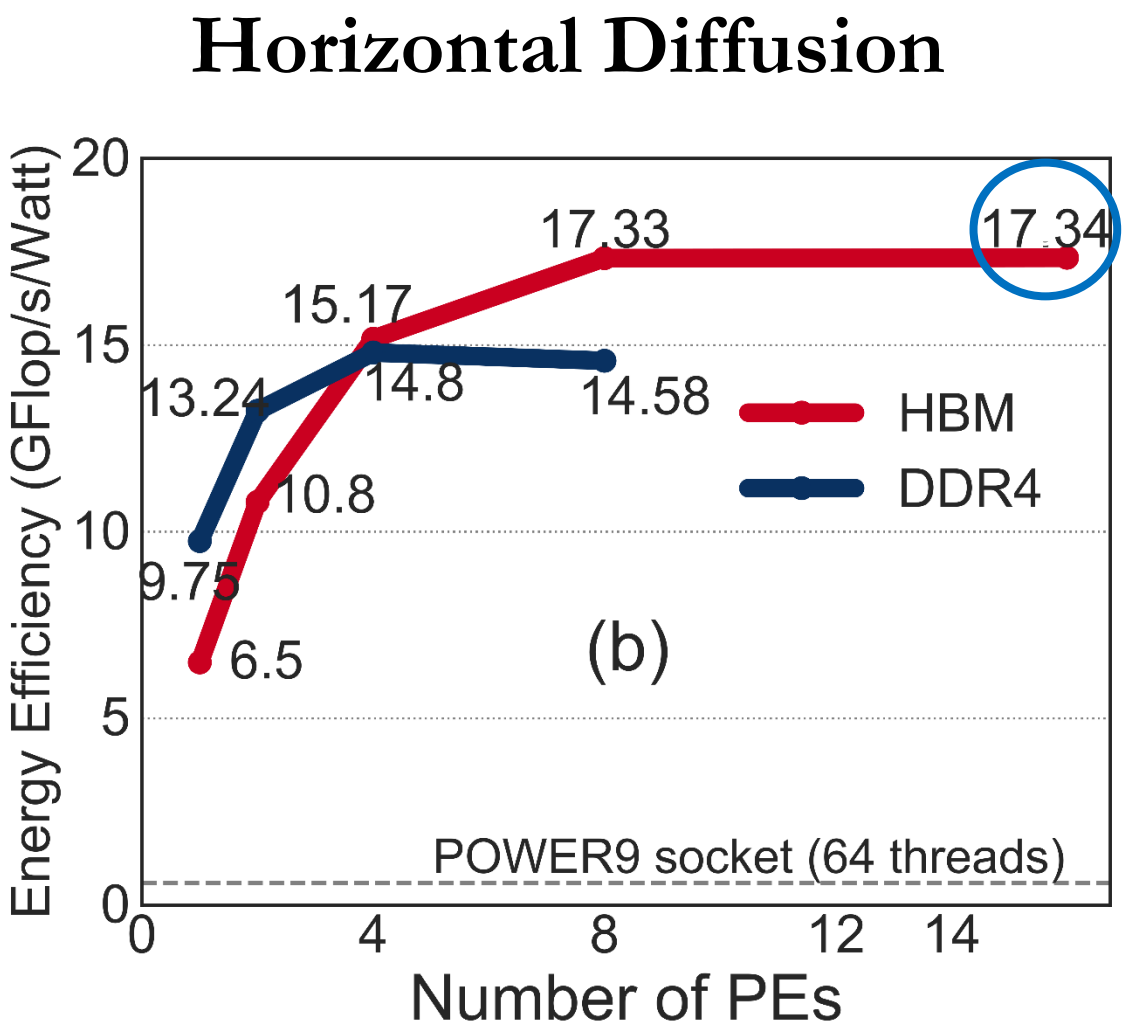


How Energy Efficient is NERO?



Enabling many HBM ports might not always be the determining factor

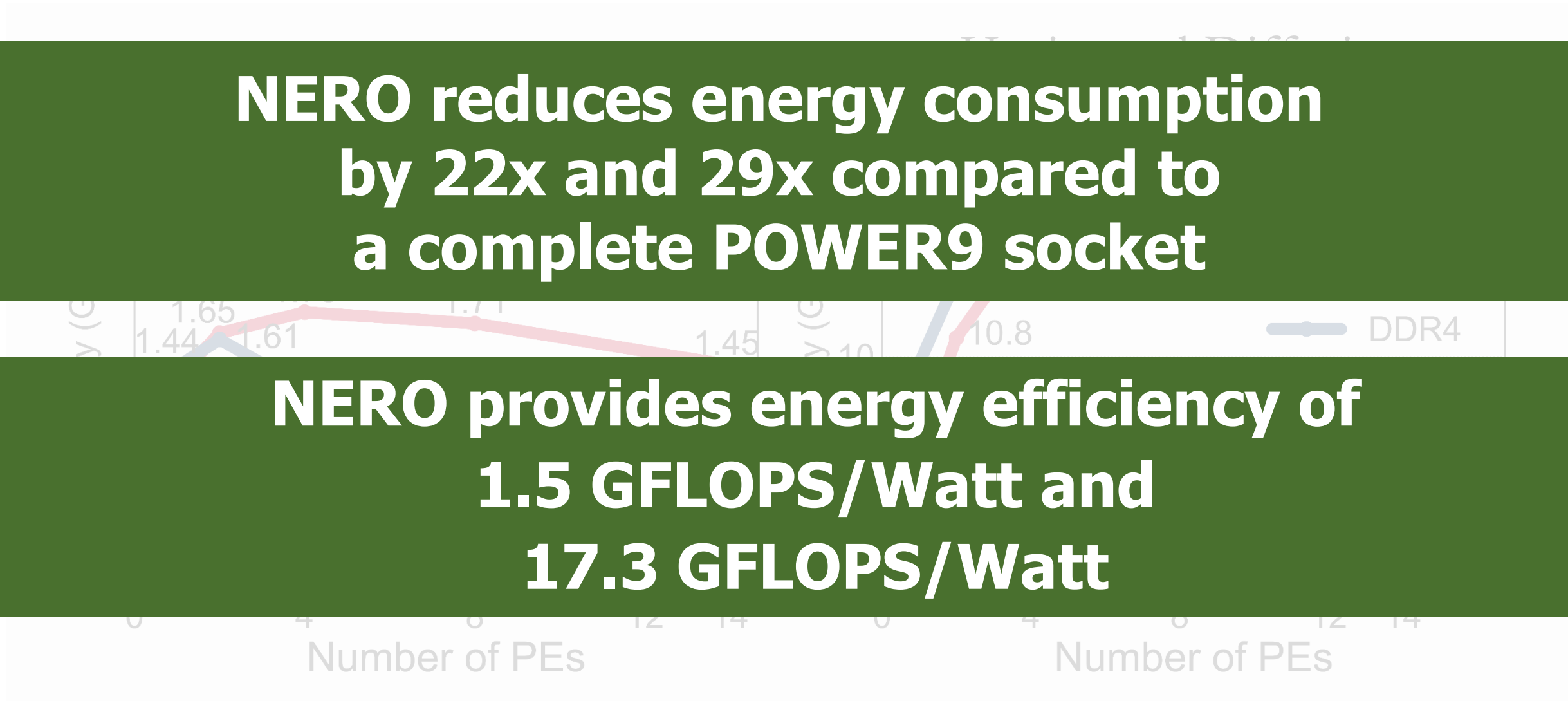
How Energy Efficient is NERO?



How Energy Efficient is NERO?

NERO reduces energy consumption by 22x and 29x compared to a complete POWER9 socket

NERO provides energy efficiency of 1.5 GFLOPS/Watt and 17.3 GFLOPS/Watt



Outline

- Background
- CPU Roofline Analysis
- FPGA-based Platform
- NERO: Near-HBM Accelerator for Weather Prediction Modeling
- Precision-optimized Tiling
- Evaluation
  - Performance Analysis
  - Energy Efficiency Analysis
- Summary

Summary

- **Motivation:** Stencil computation is an essential part of weather prediction applications
- **Problem:** Memory bound with limited performance and high energy consumption on multi-core architectures
- **Goal:** Mitigate the performance bottleneck of compound weather prediction kernels in an energy-efficient way
- **Our contribution: NERO**
  - First near High-Bandwidth Memory (HBM) FPGA-based accelerator for representative kernels from a real-world weather prediction application
  - Detailed roofline analysis to show weather prediction kernels are constrained by DRAM bandwidth on a state-of-the-art CPU system
  - Data-centric caching with precision-optimized tiling for a heterogeneous memory hierarchy
  - Scalability analysis for both DDR4 and HBM-based FPGA boards
- **Evaluation**
  - NERO outperforms a 16-core IBM POWER9 system by 4.2x and 8.3x when running two compound stencil kernels
  - NERO reduces energy consumption by 22x and 29x with an energy efficiency of 1.5 GFLOPS/Watt and 17.3 GFLOPS/Watt

NERO:
A Near High-Bandwidth Memory Stencil Accelerator for Weather Prediction Modeling

Gagandeep Singh, Dionysios Diamantopoulos, Christoph Hagleitner,
Juan Gómez-Luna, Sander Stuijk, Onur Mutlu, and Henk Corporaal

30th FPL, Sweden
31th August 2020

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY
ETH Zürich
SAFARI
IBM Research | Zurich