Computer Architecture

Lecture 24: Cutting-Edge Research in Computer Architecture III

Dr. Gagandeep Singh

Postdoctoral Researcher

December 23rd 2021

SAFARI

ETH Zürich

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

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

NERO: Weather Prediction Accelerator [FPL 2020]

- **Gagandeep Singh**, Dionysios Diamantopoulos, Christoph Hagleitner, Juan Gómez-Luna, Sander Stuijk, Onur Mutlu, and Henk Corporaal,
  "NERO: A Near High-Bandwidth Memory Stencil Accelerator for Weather Prediction Modeling"
  Proceedings of the 30th International Conference on Field-Programmable Logic and Applications (***FPL***),
  Gothenburg, Sweden, September 2020.
  [Slides (pptx) (pdf)]
  [Lightning Talk Slides (pptx) (pdf)]
  [Talk Video (23 minutes)]
  One of the four papers nominated for the Stamatis Vassiliadis Memorial Best Paper Award.

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

Gagandeep Singh[a,b,c] Dionysios Diamantopoulos[c] Christoph Hagleitner[c] Juan Gómez-Luna[b]
Sander Stuijk[a] Onur Mutlu[b] Henk Corporaal[a]
[a]Eindhoven University of Technology [b]ETH Zürich [c]IBM Research Europe, 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 upto 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

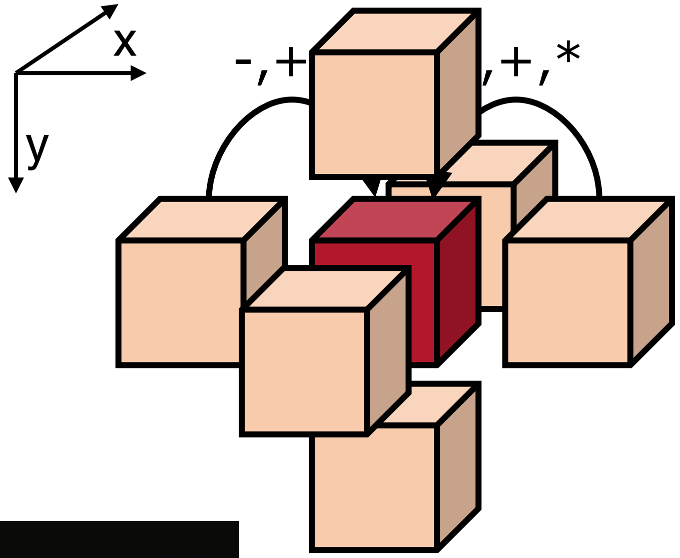
Energy Efficiency Analysis

Summary

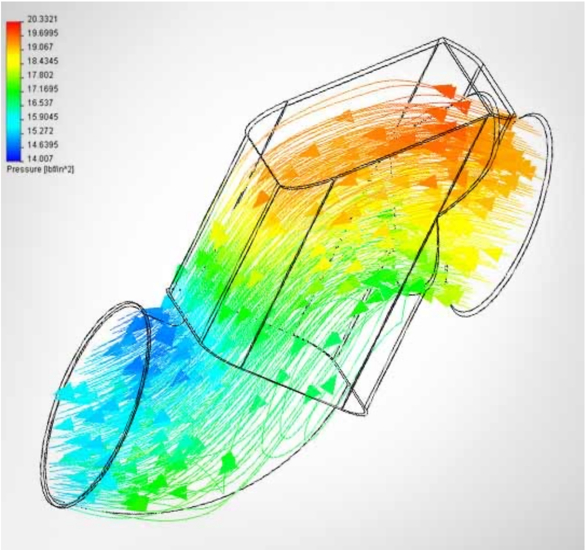
Stencil Computations and Applications

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

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



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



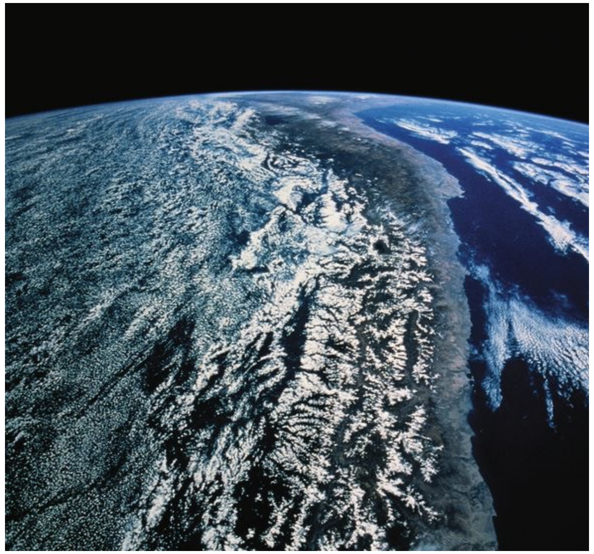
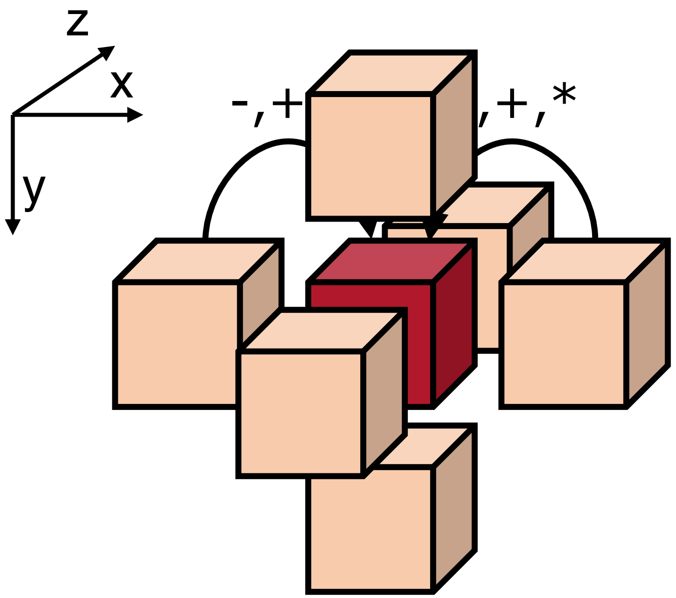


Image sources: http://www.flometrics.com/fluid-dynamics/computational-fluid-dynamics
Naoe, Kensuke et al. "Secure Key Generation for Static Visual Watermarking by Machine Learning in Intelligent Systems and Services" IJSSOE, 2010

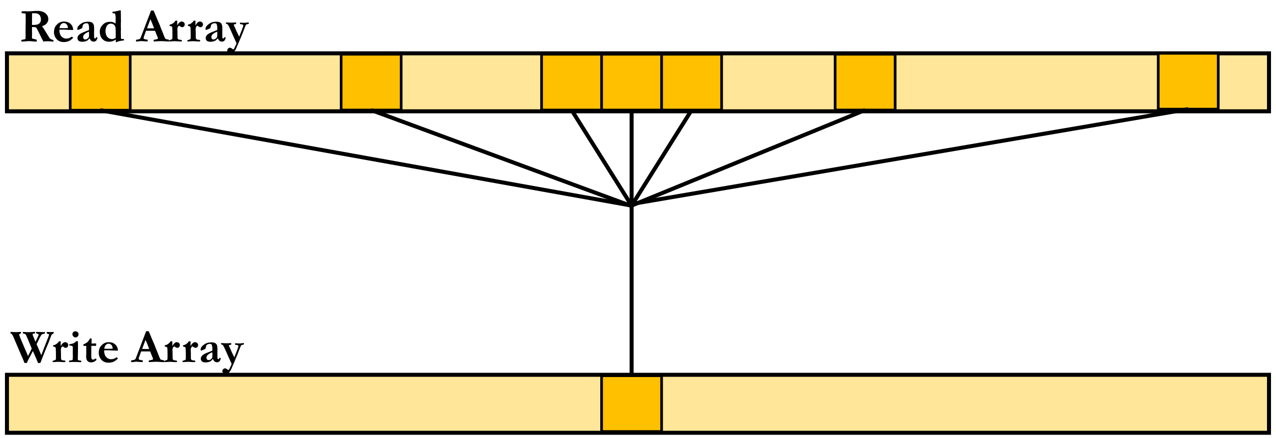
Stencil Characteristics

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

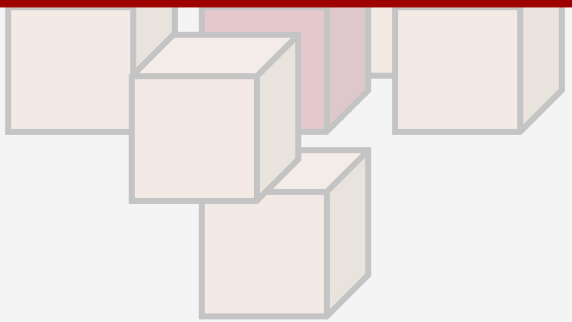
Image source: Xu, Jingheng et al. "Performance Tuning and Analysis for Stencil-Based Applications on POWER8 Processor" ACM TACO, 2018

Stencil Characteristics

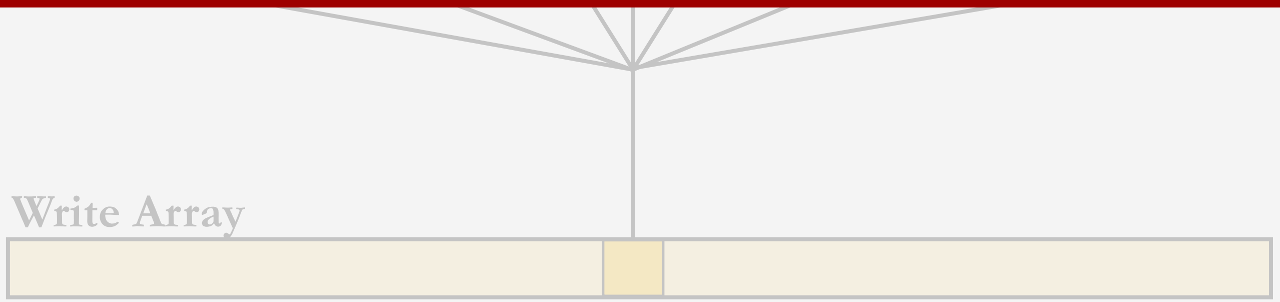
High-order stencil computations are cache unfriendly

- Limited arithmetic intensity
- Sparse and complex access pattern

Performance bottleneck



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



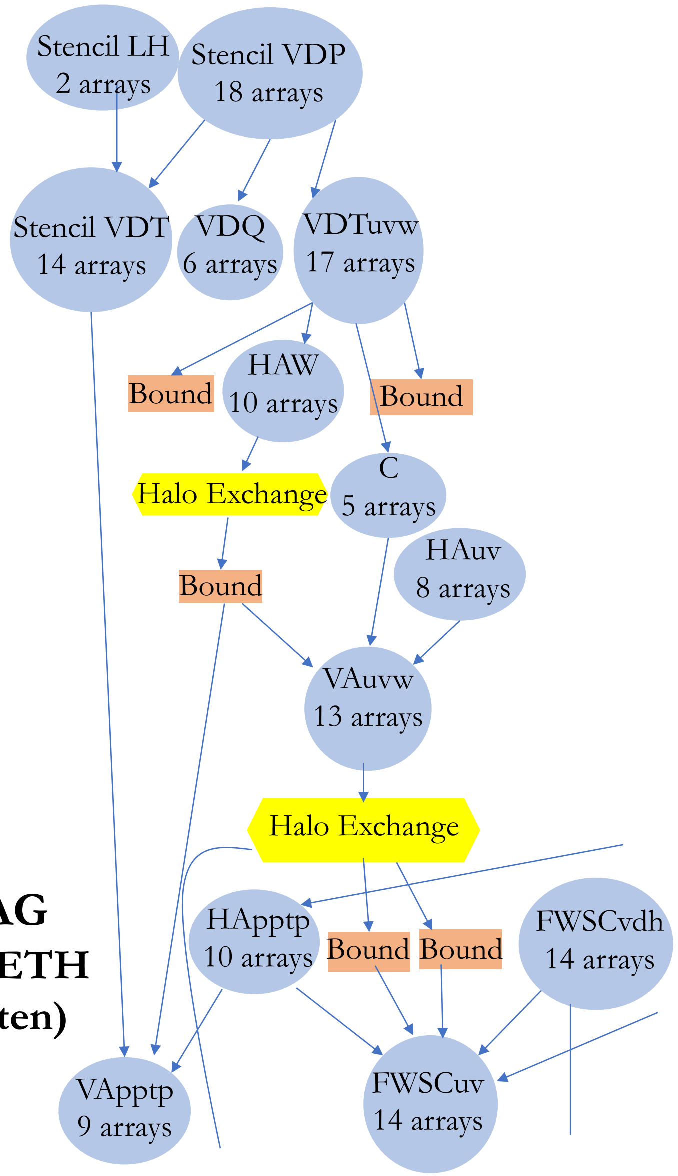
Mapping of 7-point Jacobi from 3D plane onto 1D plane

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**



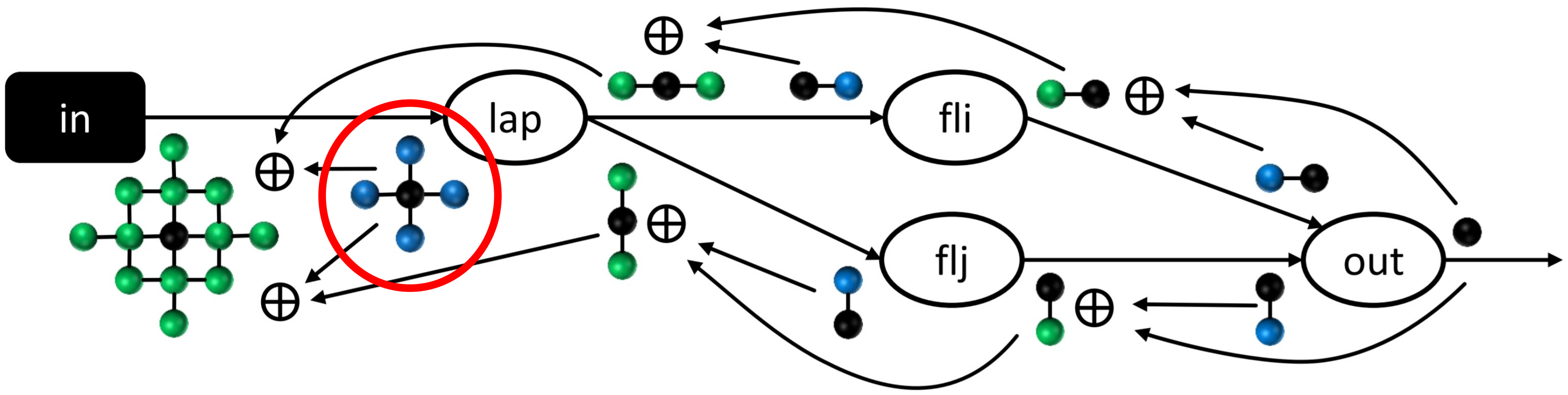
Section of COSMO CDAG (Courtesy CSCS/ETH and Ronald Luijten)

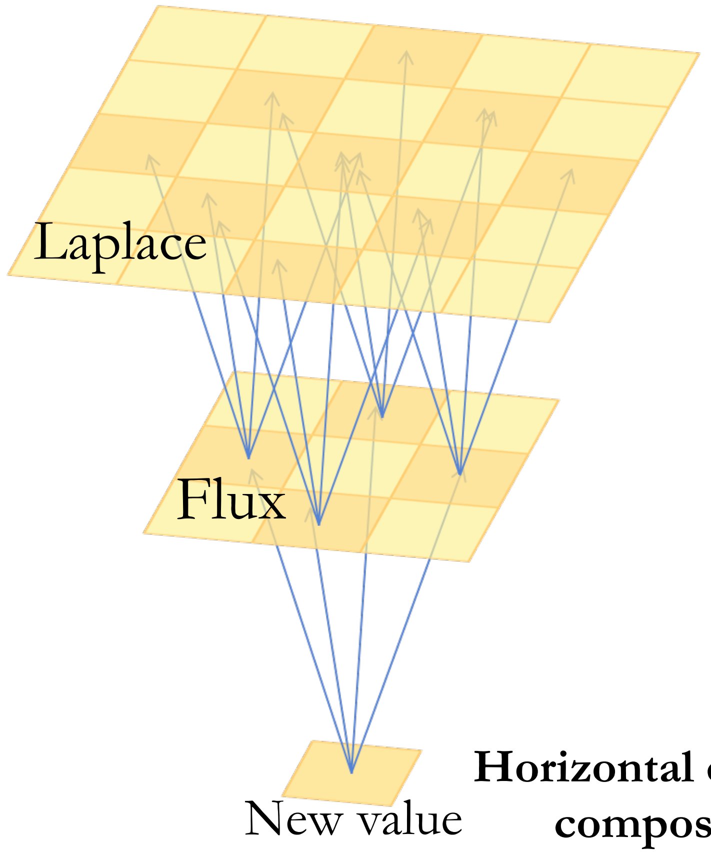
Example Complex Stencil: Horizontal Diffusion

- 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





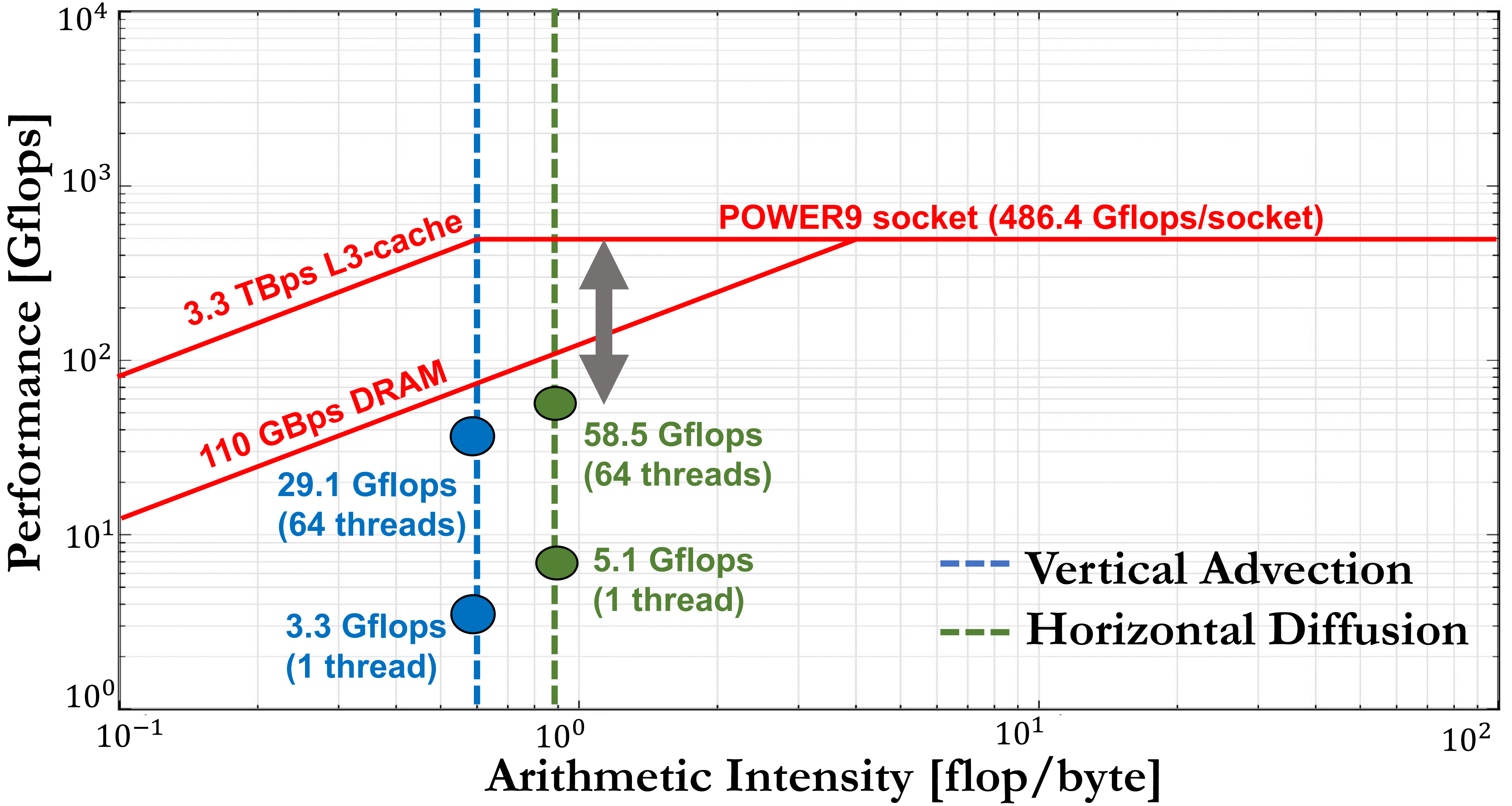
Horizontal diffusion composition (Courtesy CSCS/ETH and Ronald Luijten)

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

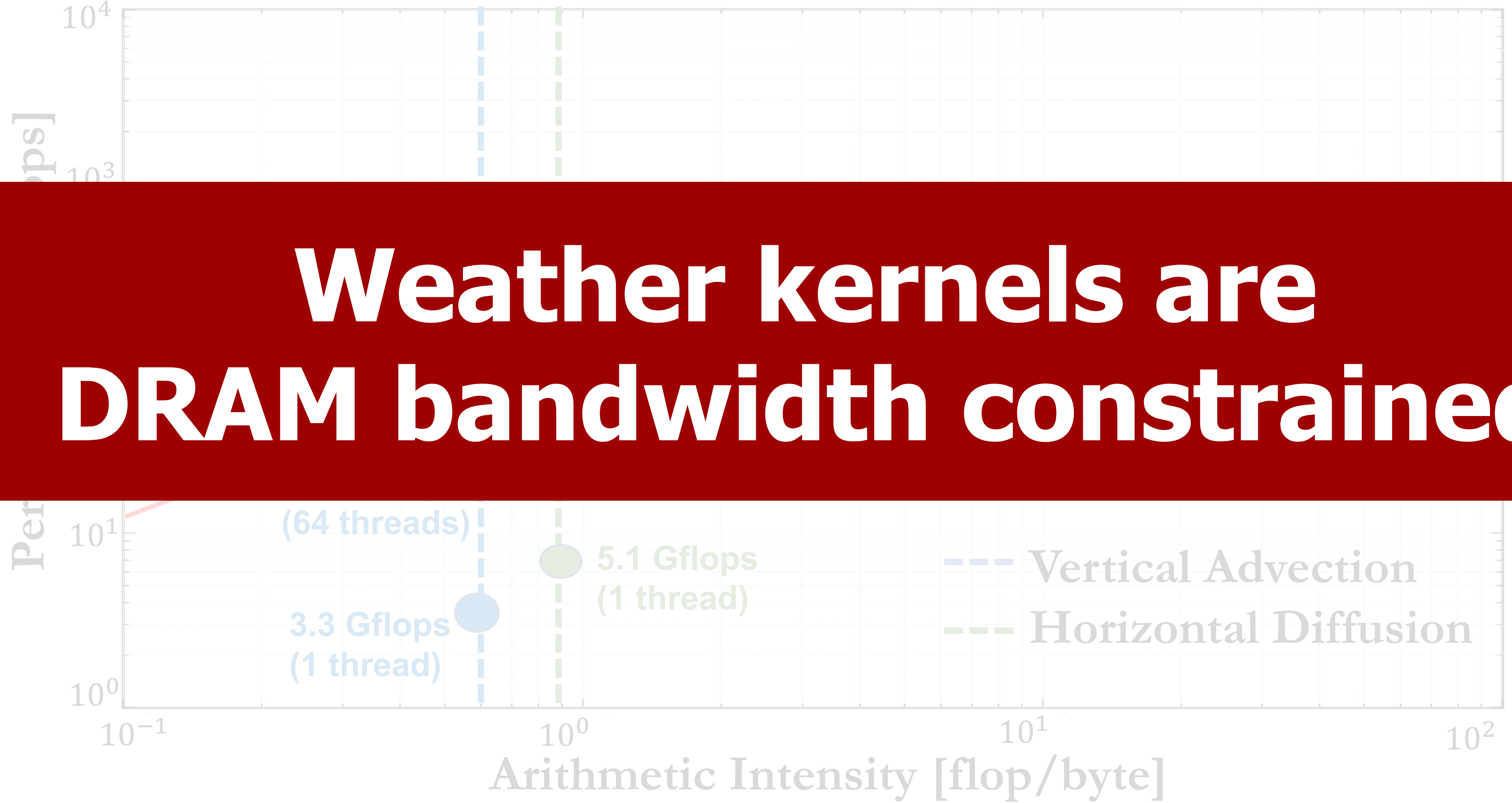
Outline

- 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

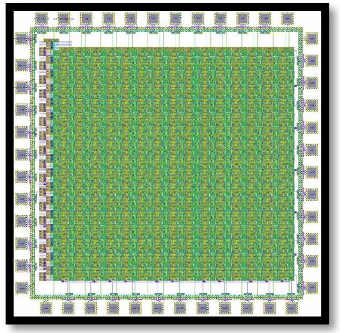
Outline

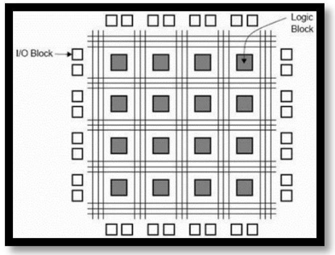
Silicon Alternatives

CPUs

GPUs

FPGAs





ASICs

FLEXIBILITY

EFFICIENCY

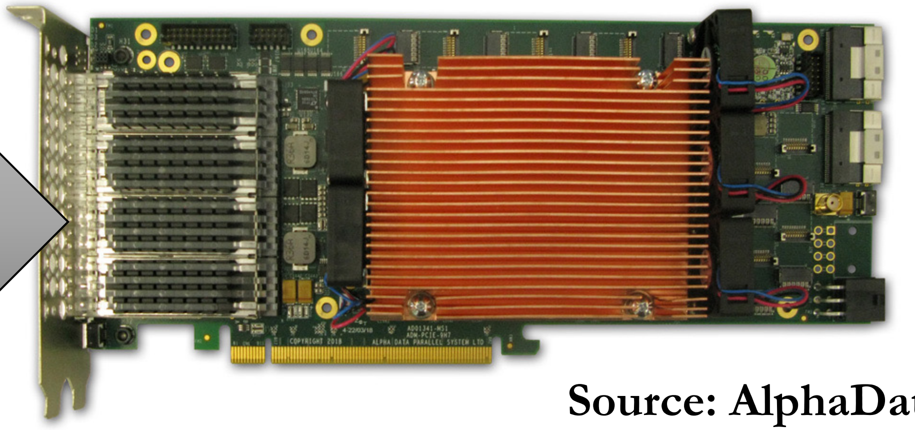
FPGAs are highly configurable!

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

POWER9 AC922

Source: AlphaData

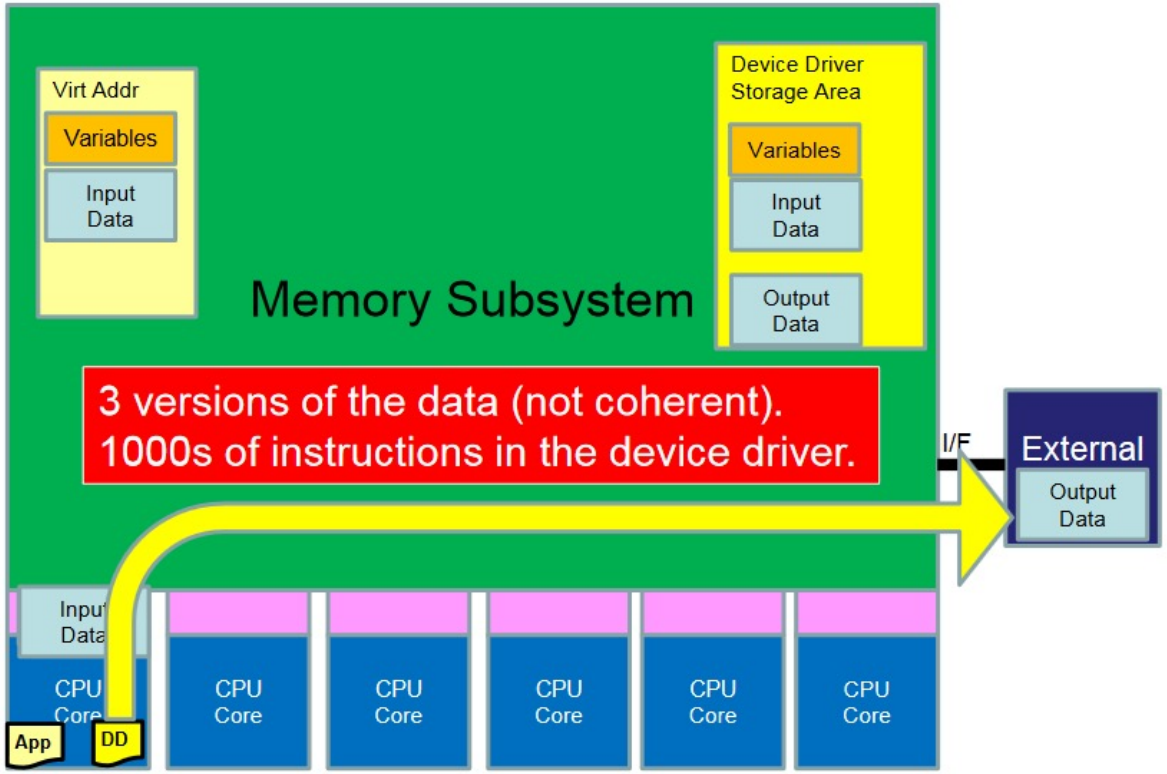
DDR4-based AD9V3 board

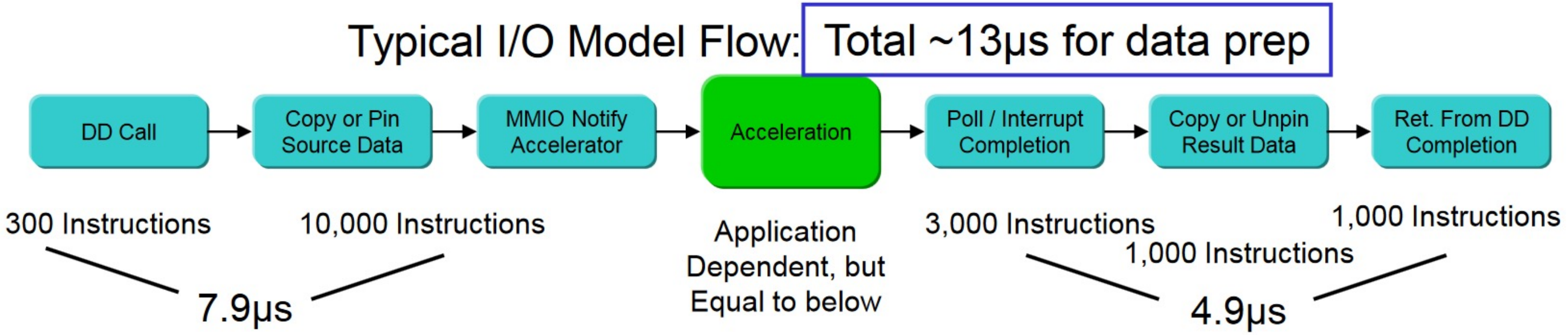
We evaluate two POWER9+FPGA systems:

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

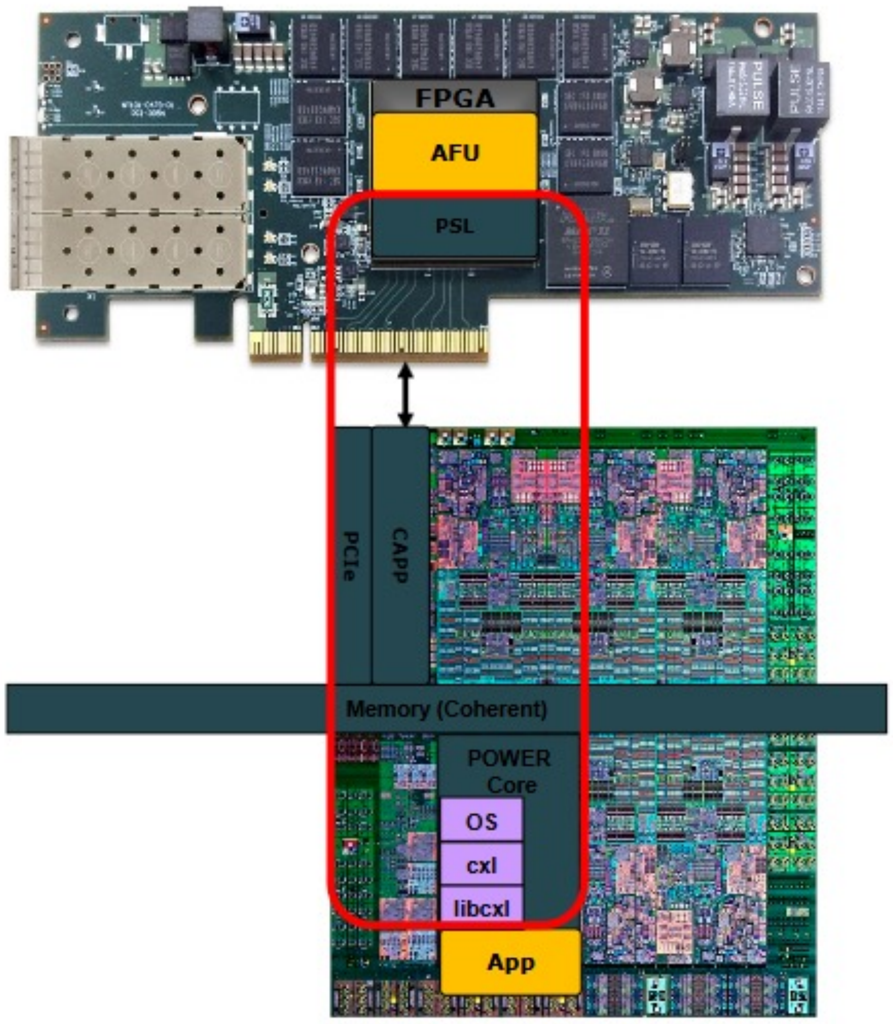
2. DDR4-based board AD9V3
Xilinx Virtex Ultrascale+™ XCVU3P-2

Background: Traditional I/O Technology

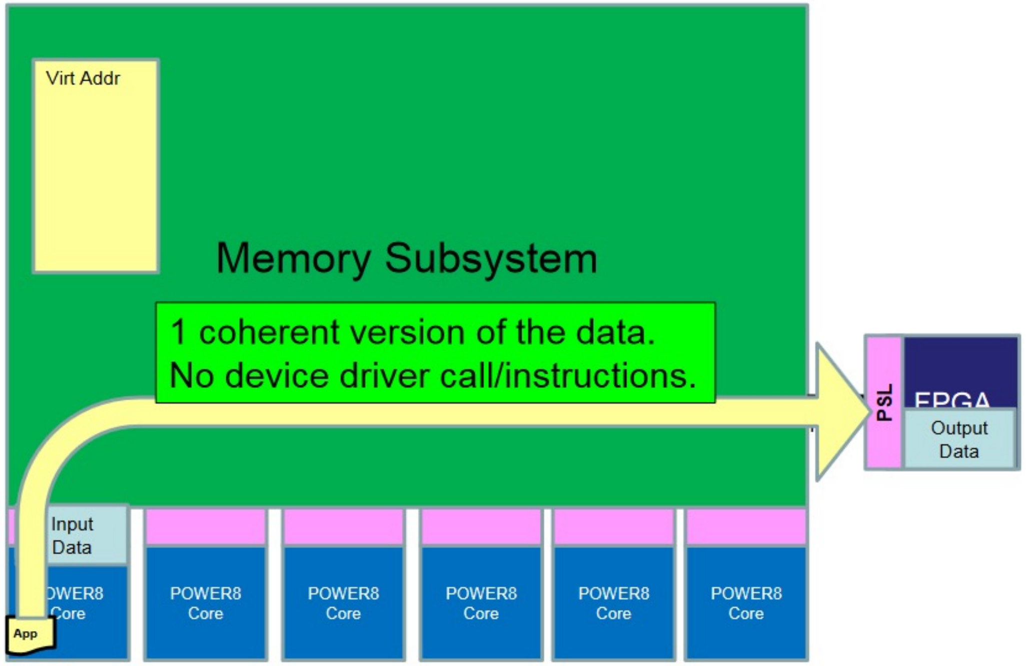




CAPI Overview



POWER8 – POWER9 Processor



Flow with a CAPI Model: **Total 0.36μs**

Shared Mem. Notify Accelerator →	Acceleration →	Shared Memory Completion
400 Instructions	Application Dependent, but Equal to above	100 Instructions
0.3μs		0.06μs

Outline

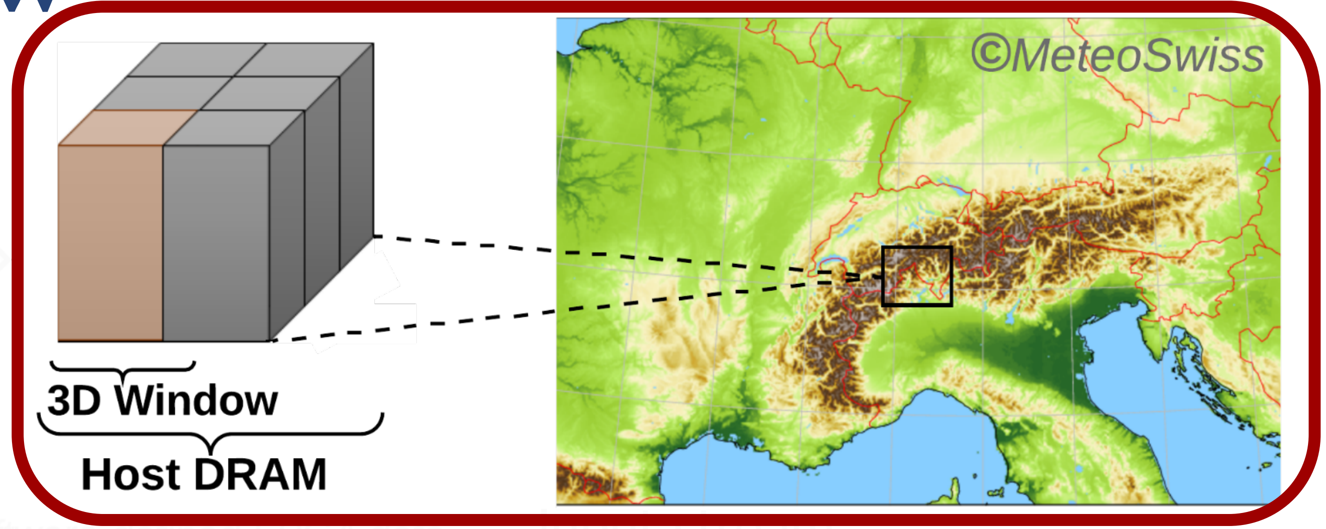
- Background
- CPU Roofline Analysis
- FPGA-based Platform
- **NERO: Near-HBM Accelerator for Weather Prediction Modeling**
- Precision-optimized Tiling
- Evaluation
  - Performance Analysis
  - Energy Efficiency Analysis
- Summary

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

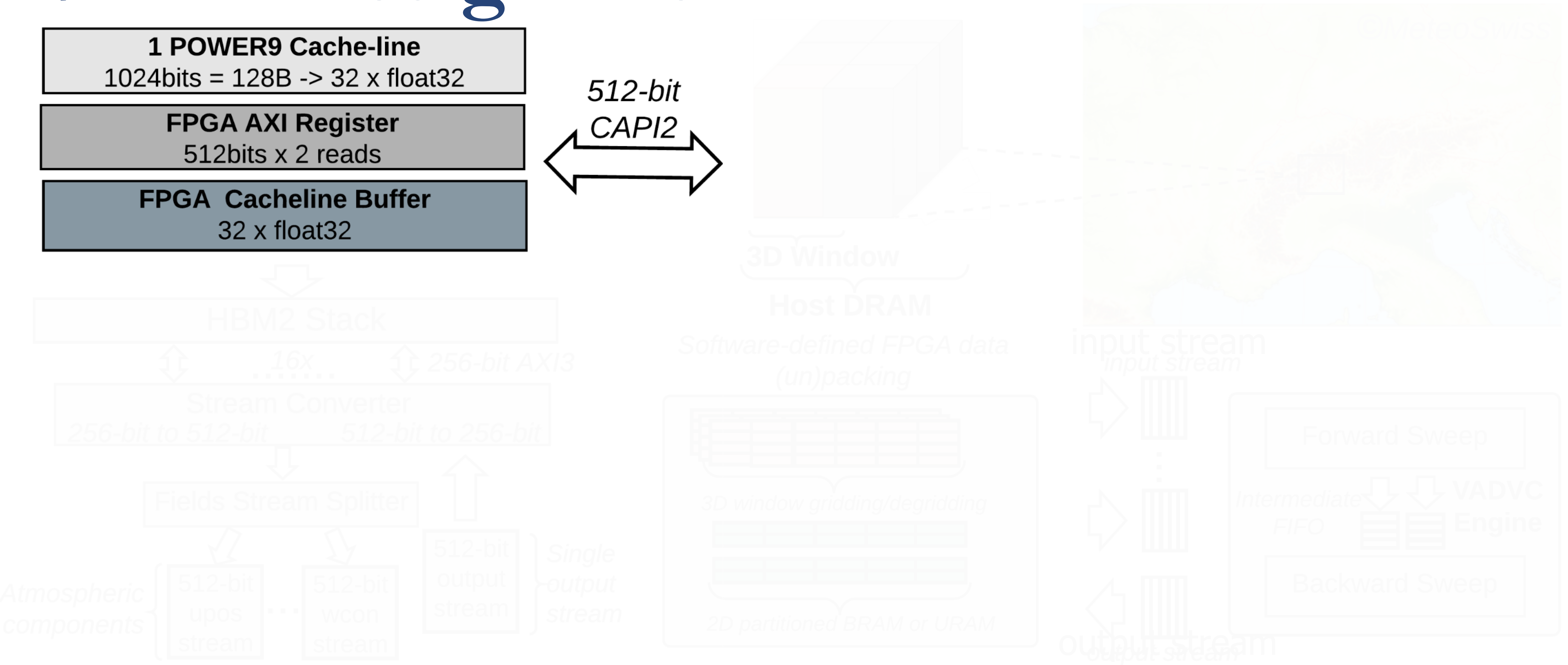
NERO Design Flow

NERO Design Flow



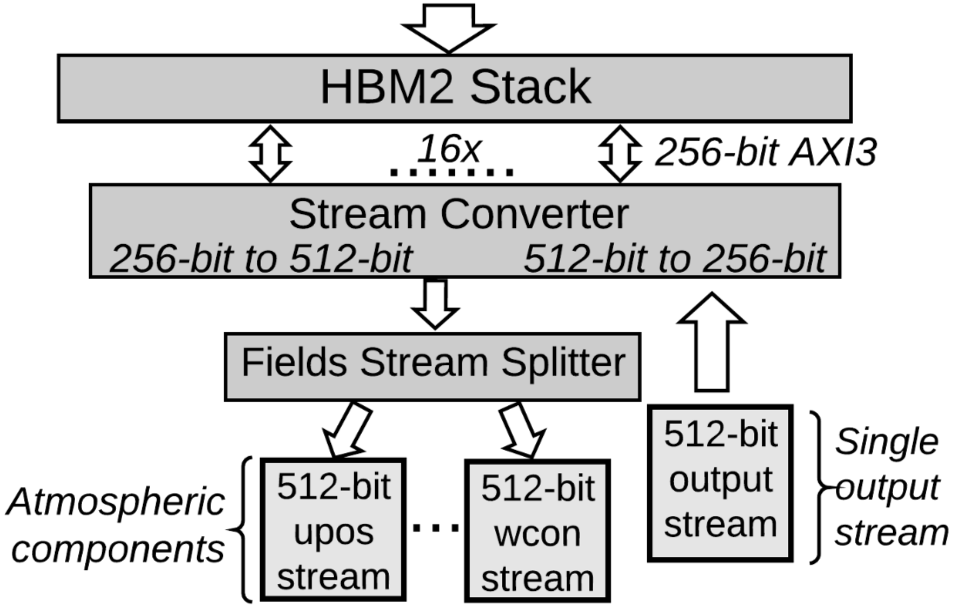
Weather data in the host DRAM

NERO Design Flow



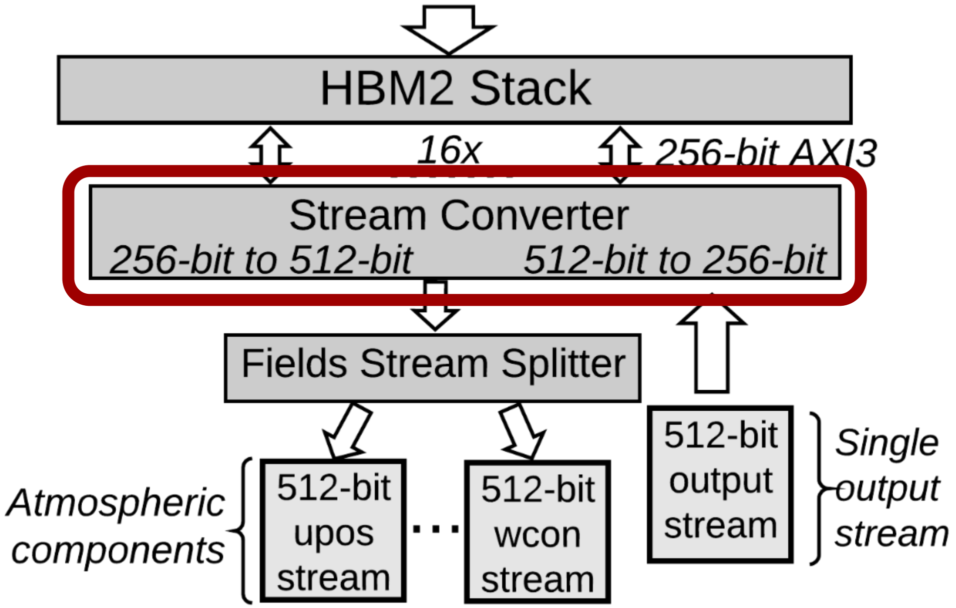
Cache-line transfer over CAPI2

NERO Design Flow



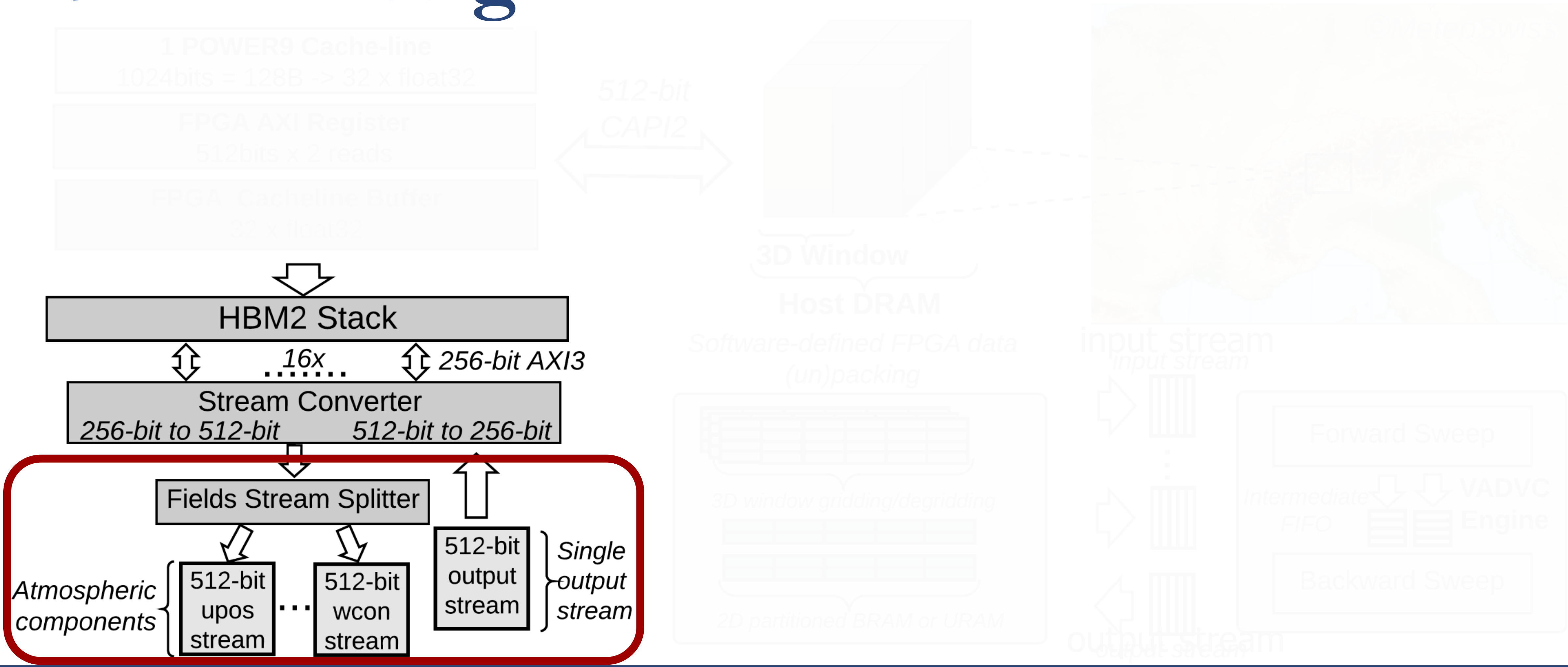
Data mapping onto HBM

NERO Design Flow



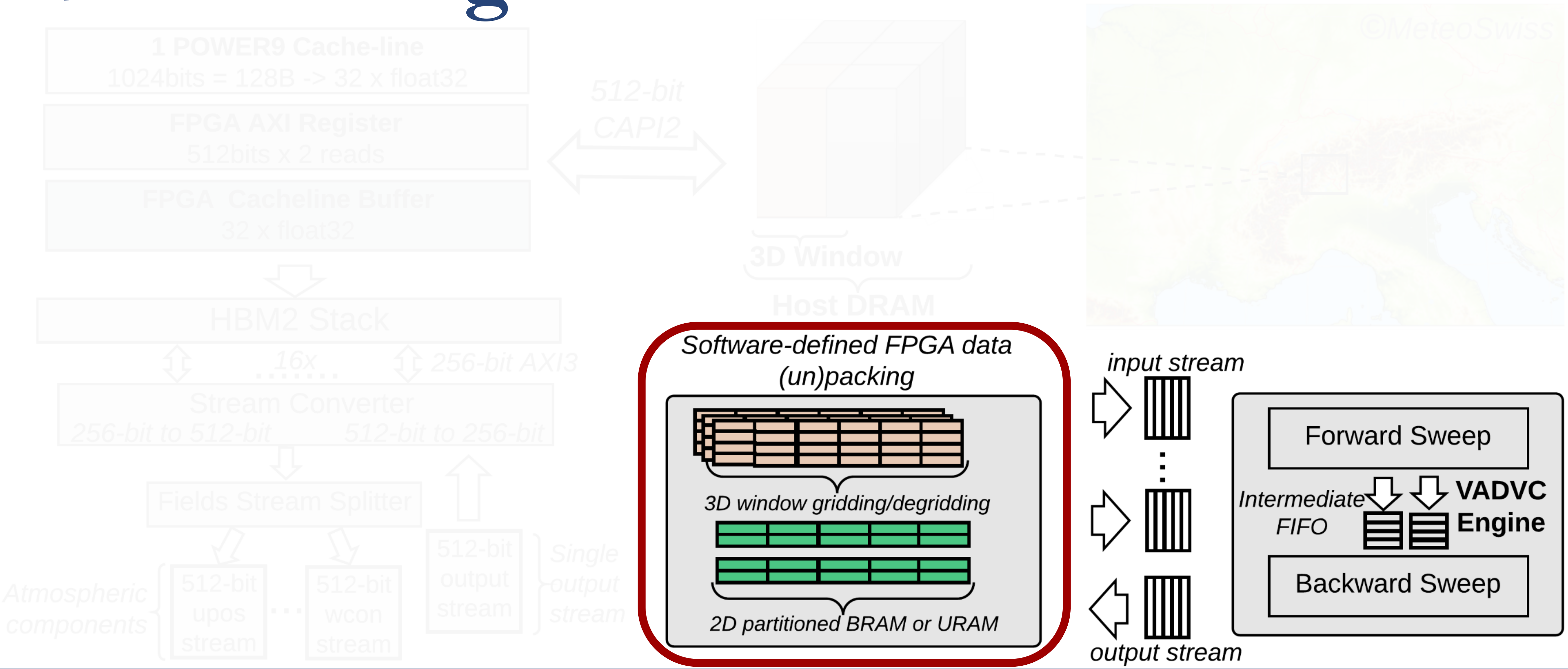
Data mapping onto HBM

NERO Design Flow



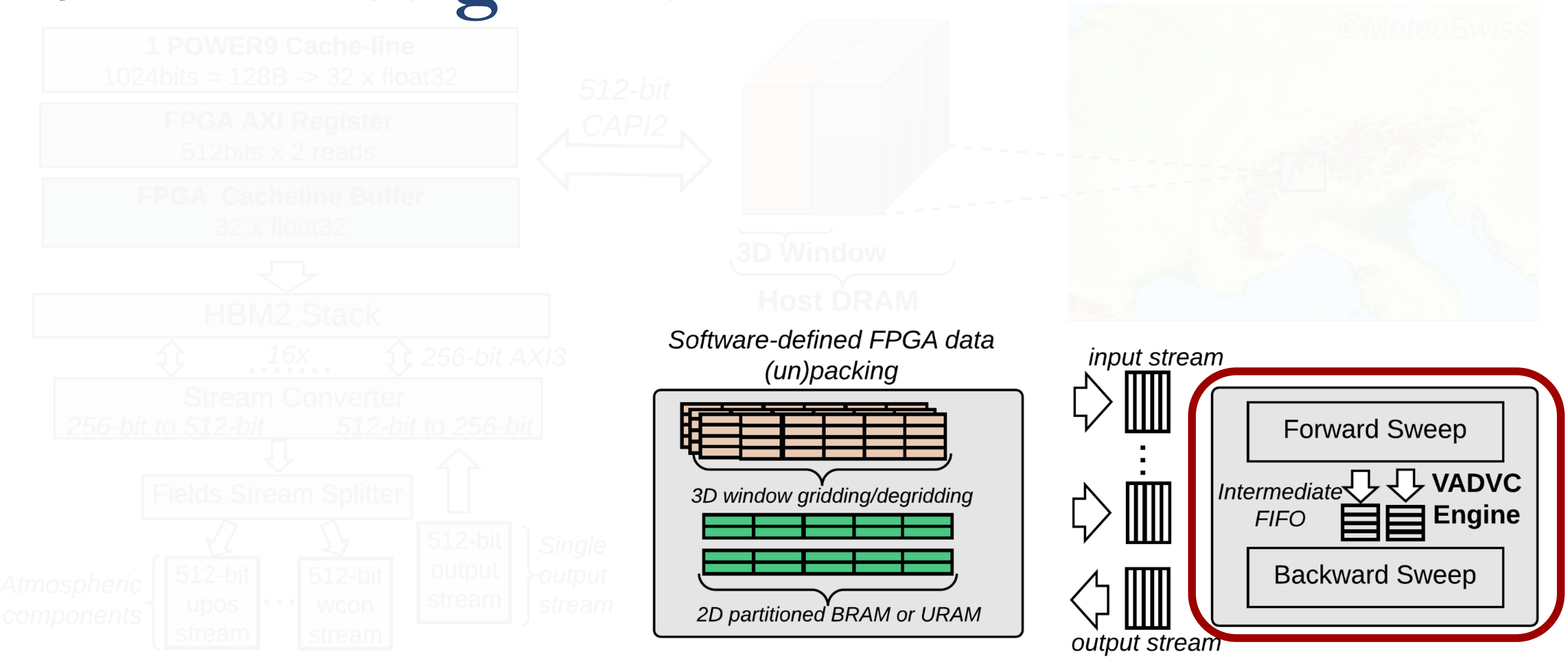
Data mapping onto HBM

NERO Design Flow



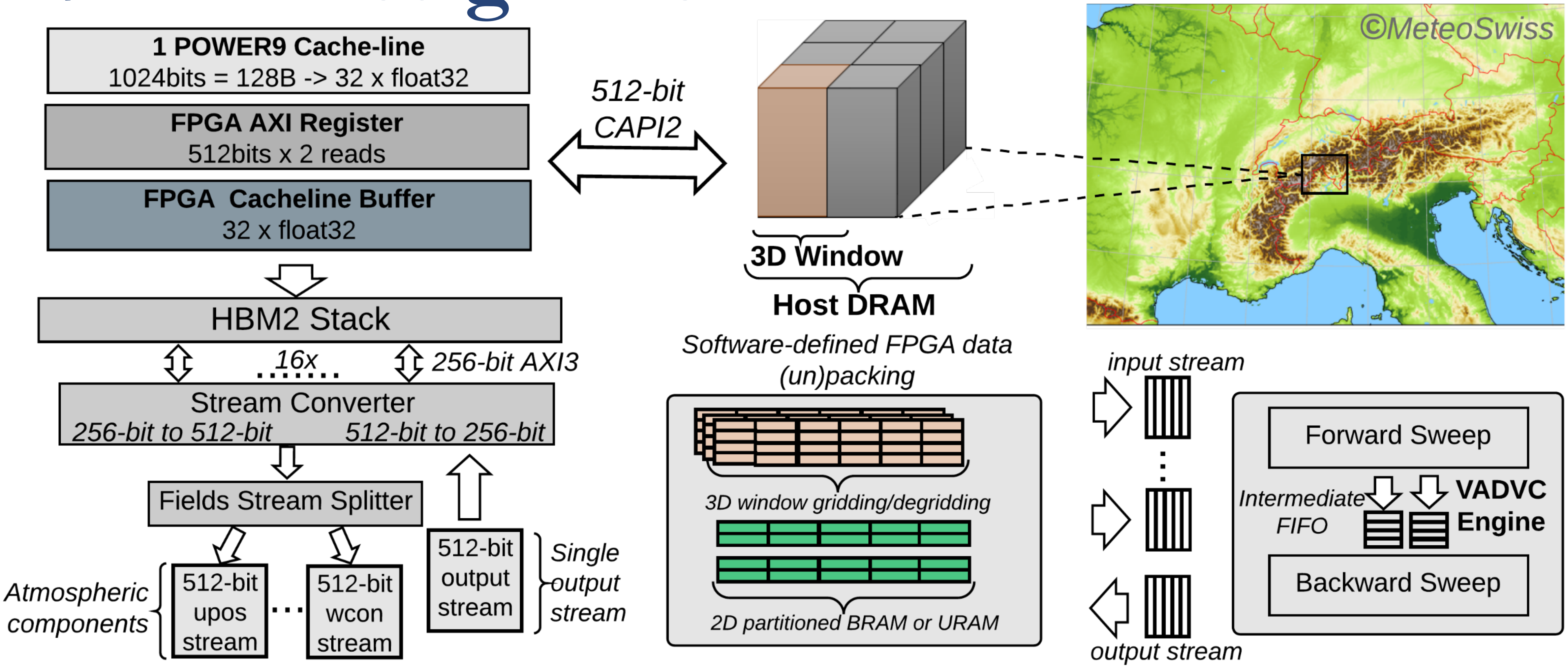
Main execution pipeline

NERO Design Flow



Main execution pipeline

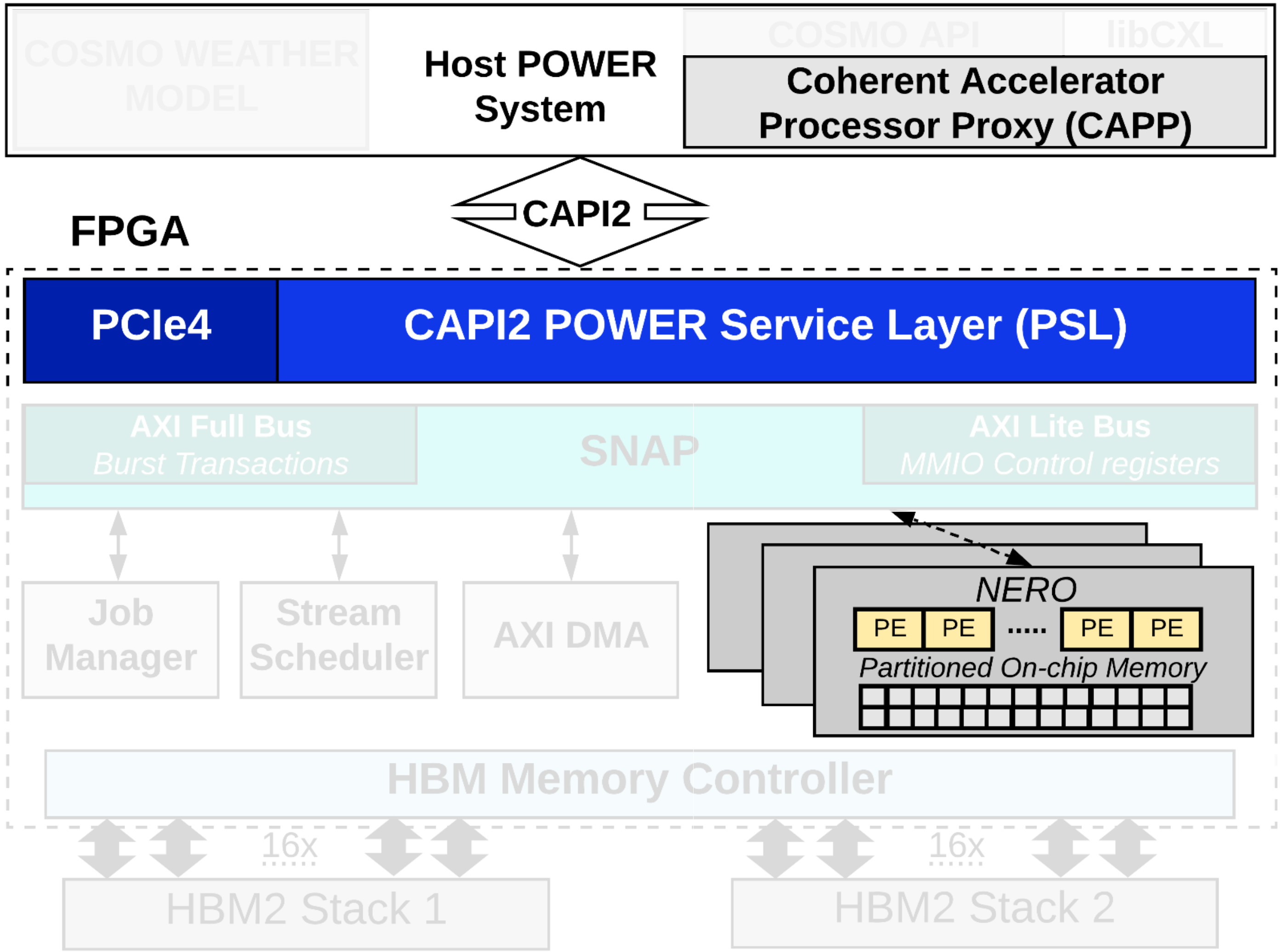
NERO Design Flow



Complete design flow

NERO Application Framework

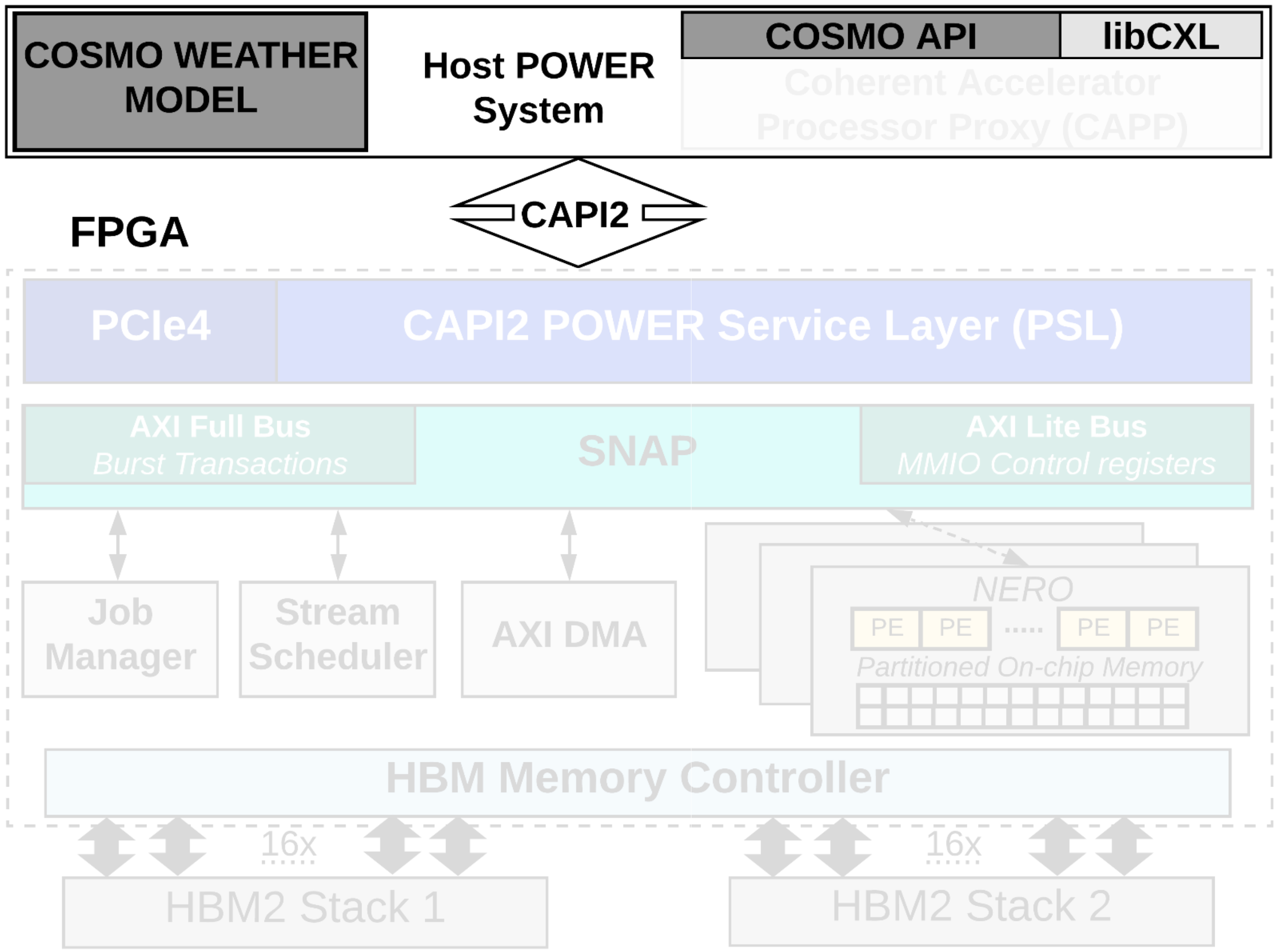
- NERO communicates to Host over **CAPI2** (Coherent Accelerator Processor Interface)



NERO Application Framework

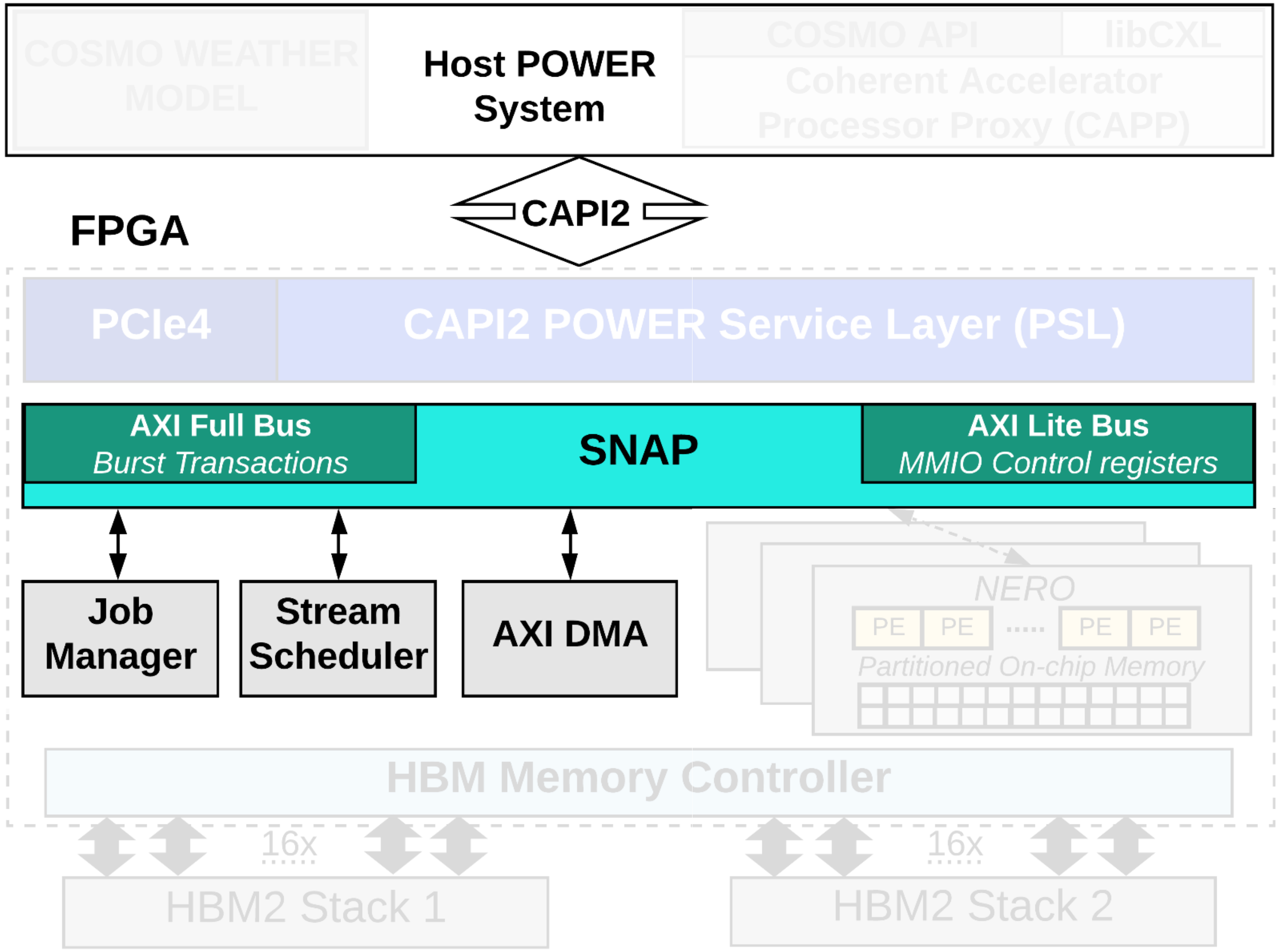
- NERO communicates to Host over **CAPI2** (Coherent Accelerator Processor Interface)

- **COSMO API** handles offloading jobs to NERO



NERO Application Framework

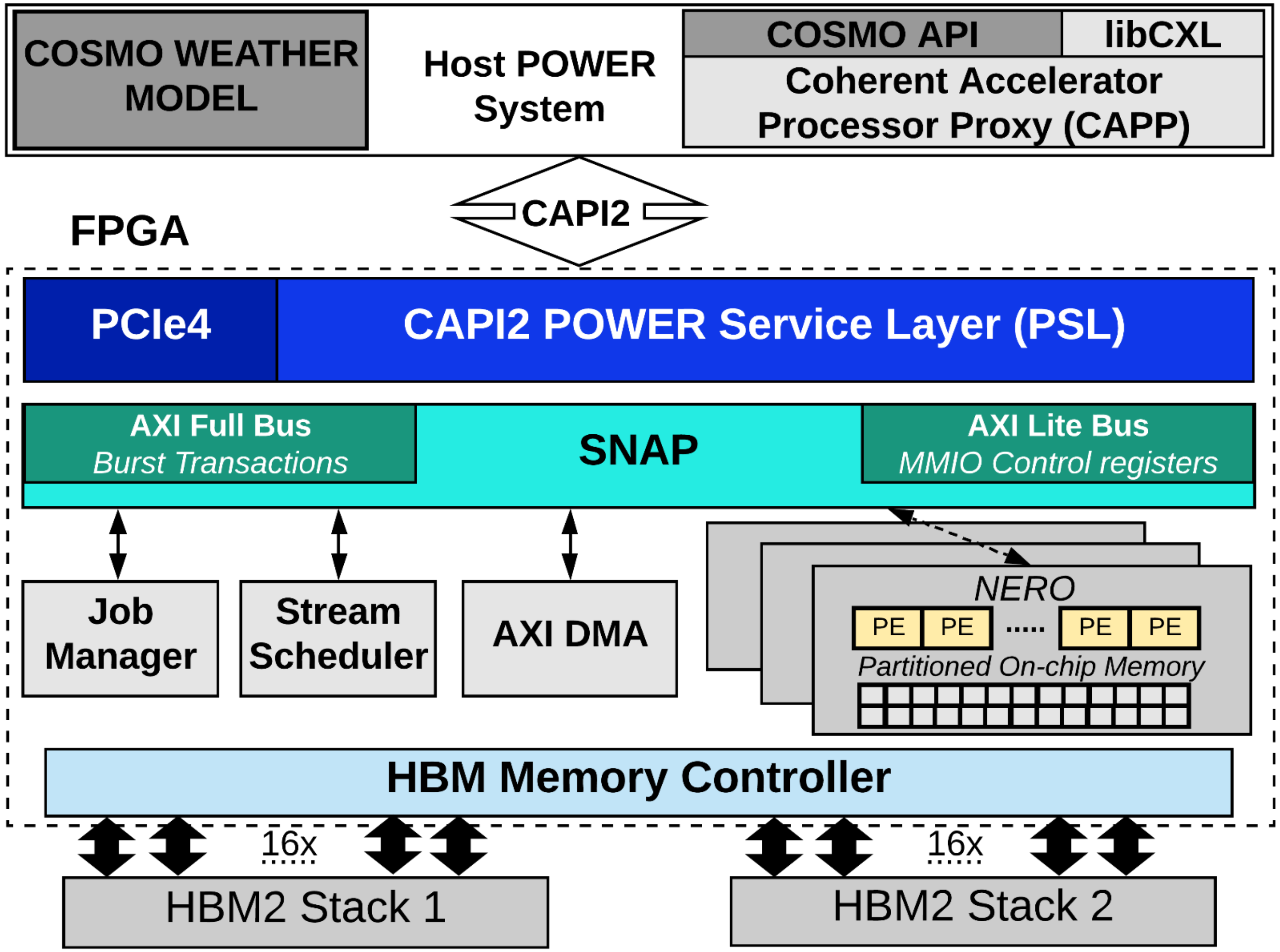
- NERO communicates to Host over **CAPI2** (Coherent Accelerator Processor Interface)
- **COSMO API** handles offloading jobs to NERO
- **SNAP** (Storage, Network, and Analytics Programming) allows for seamless integration of the COSMO API



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

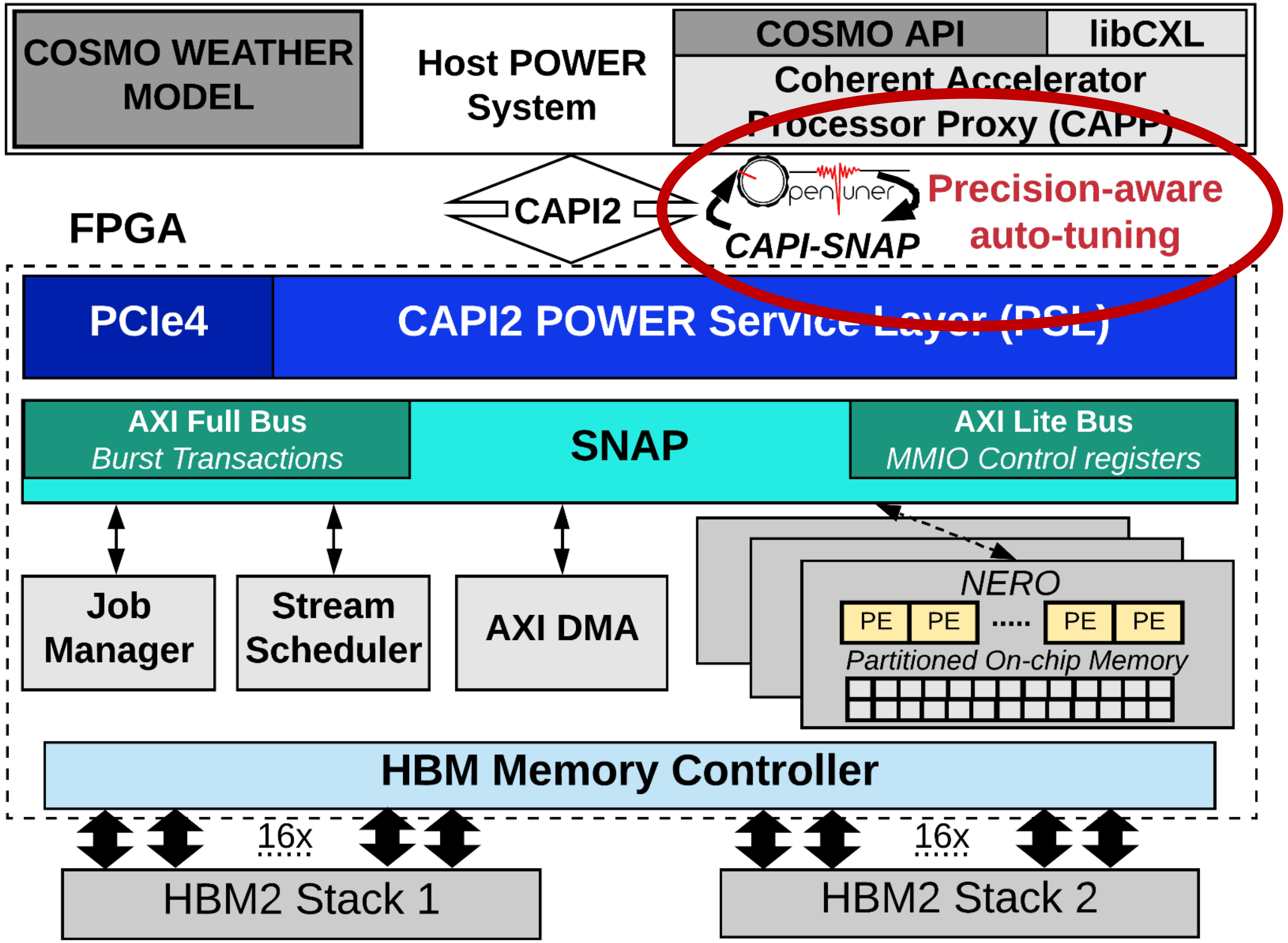


https://github.com/open-power/snap

Outline

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**

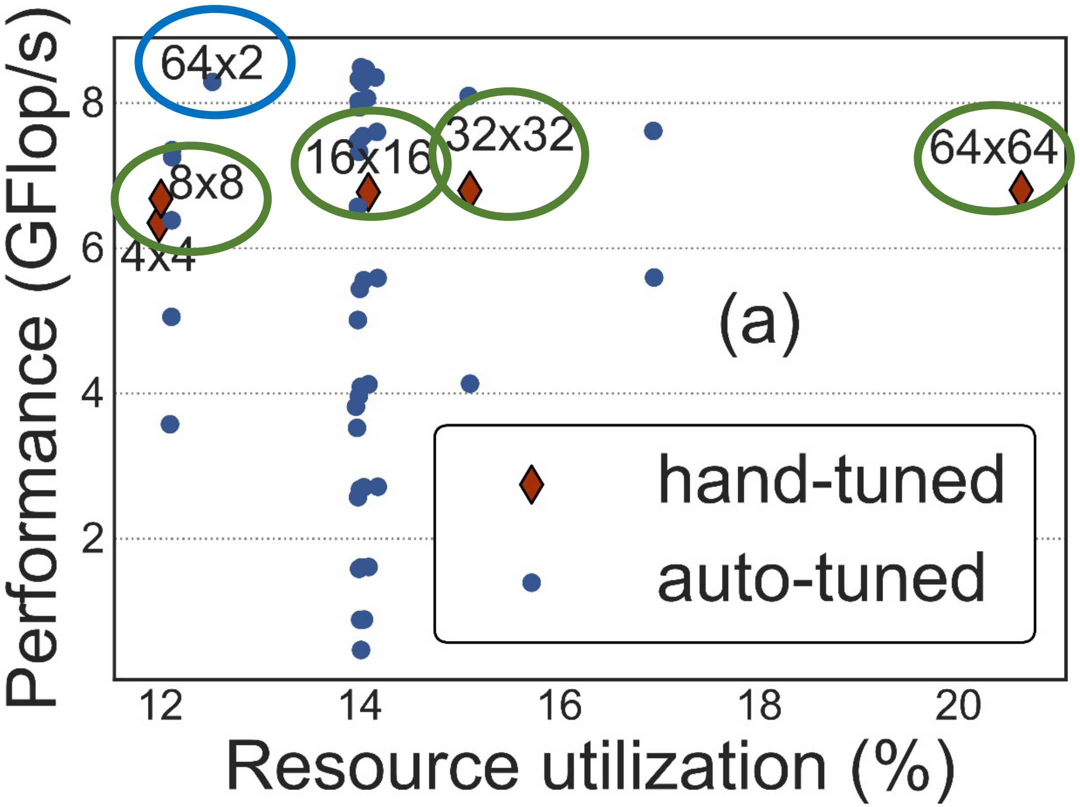


https://github.com/jansel/opentuner

Precision-optimized Tiling

Precision-optimized Tiling

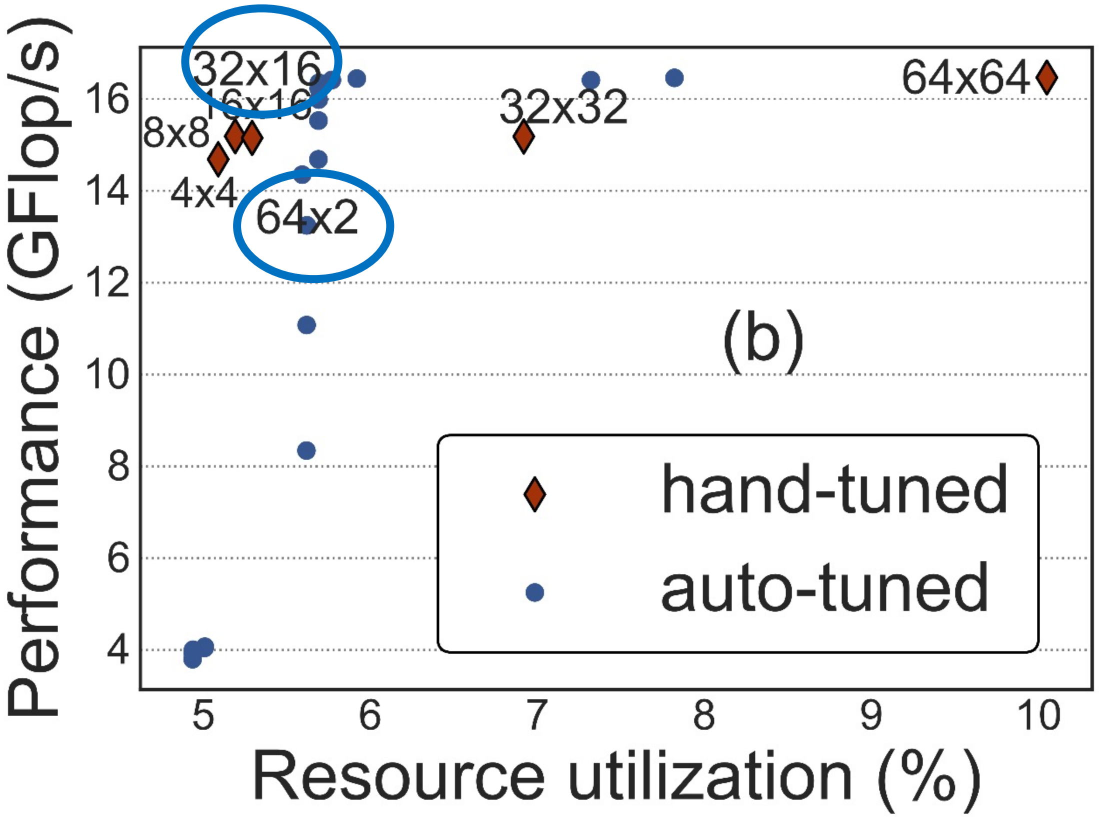
Single Precision



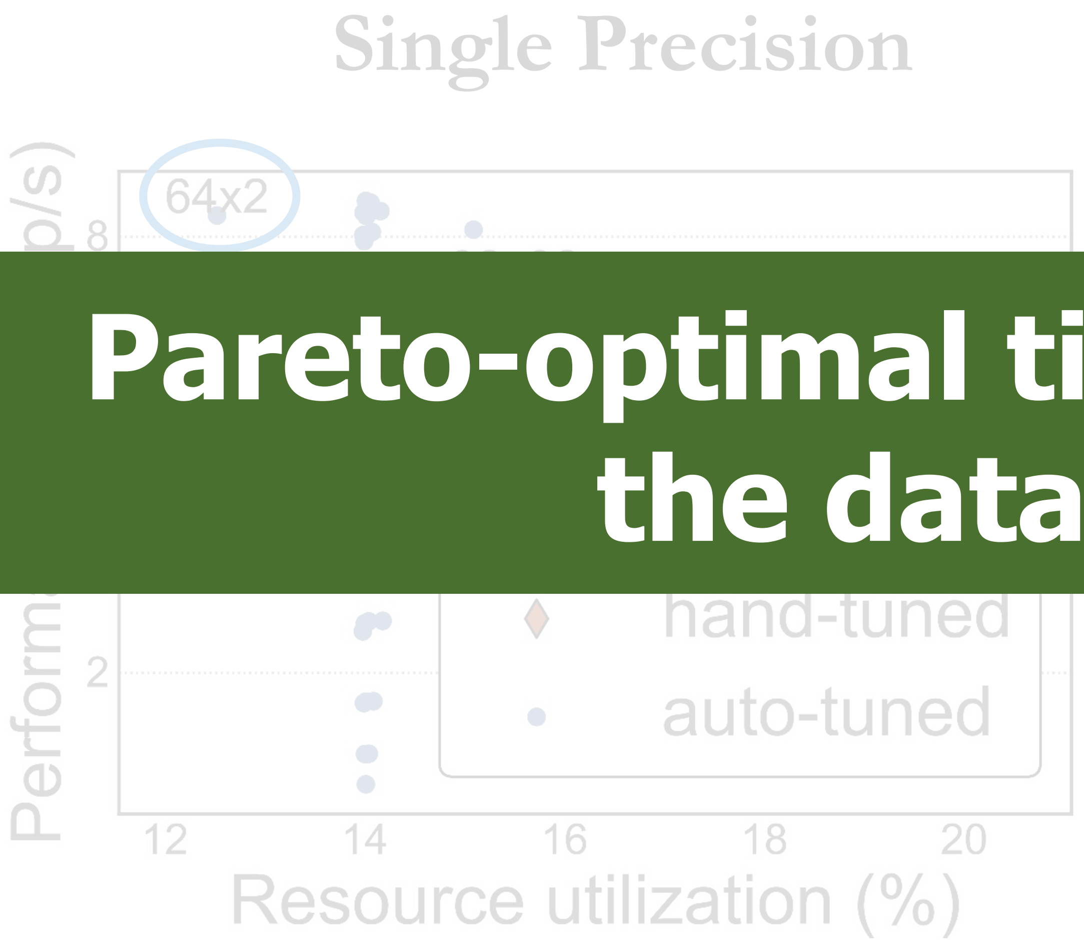
Precision-optimized Tiling

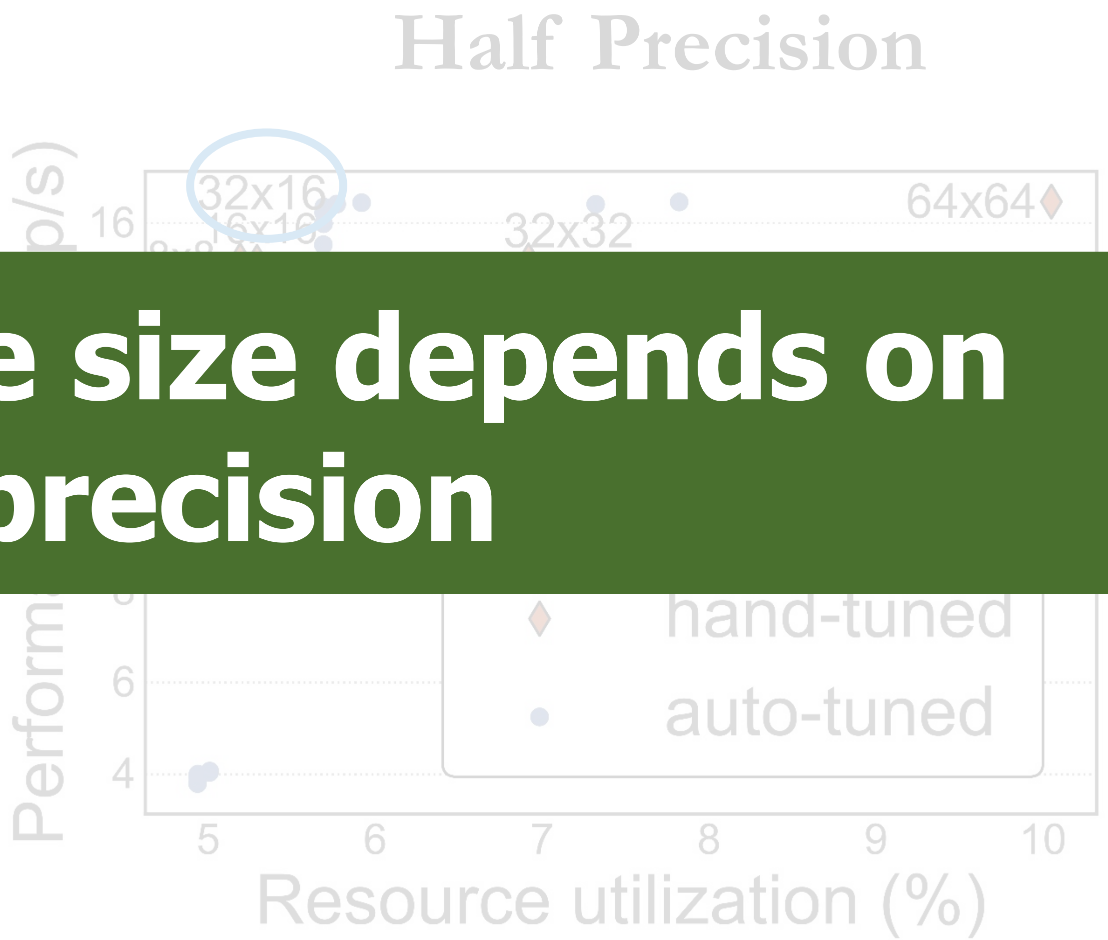
Single Precision

Half Precision



Precision-optimized Tiling

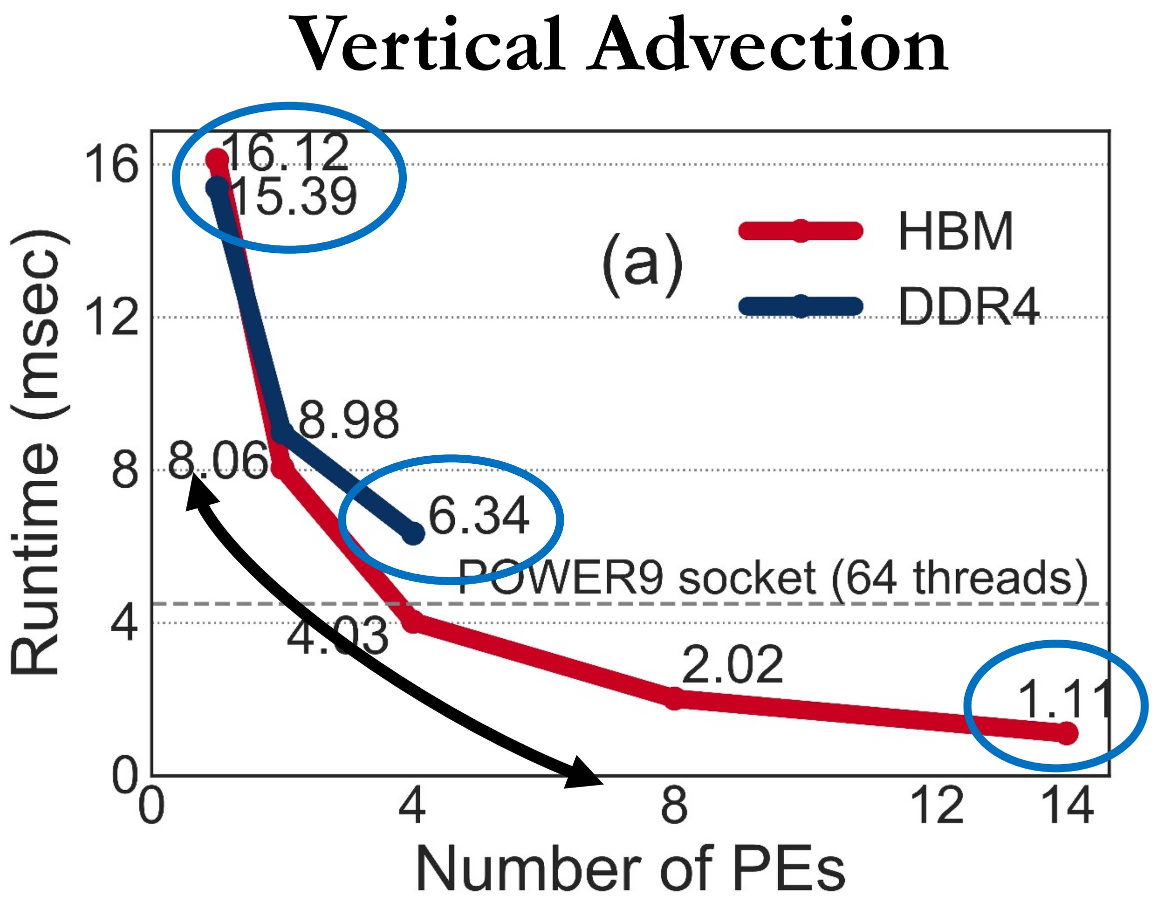




Pareto-optimal tile size depends on the data precision

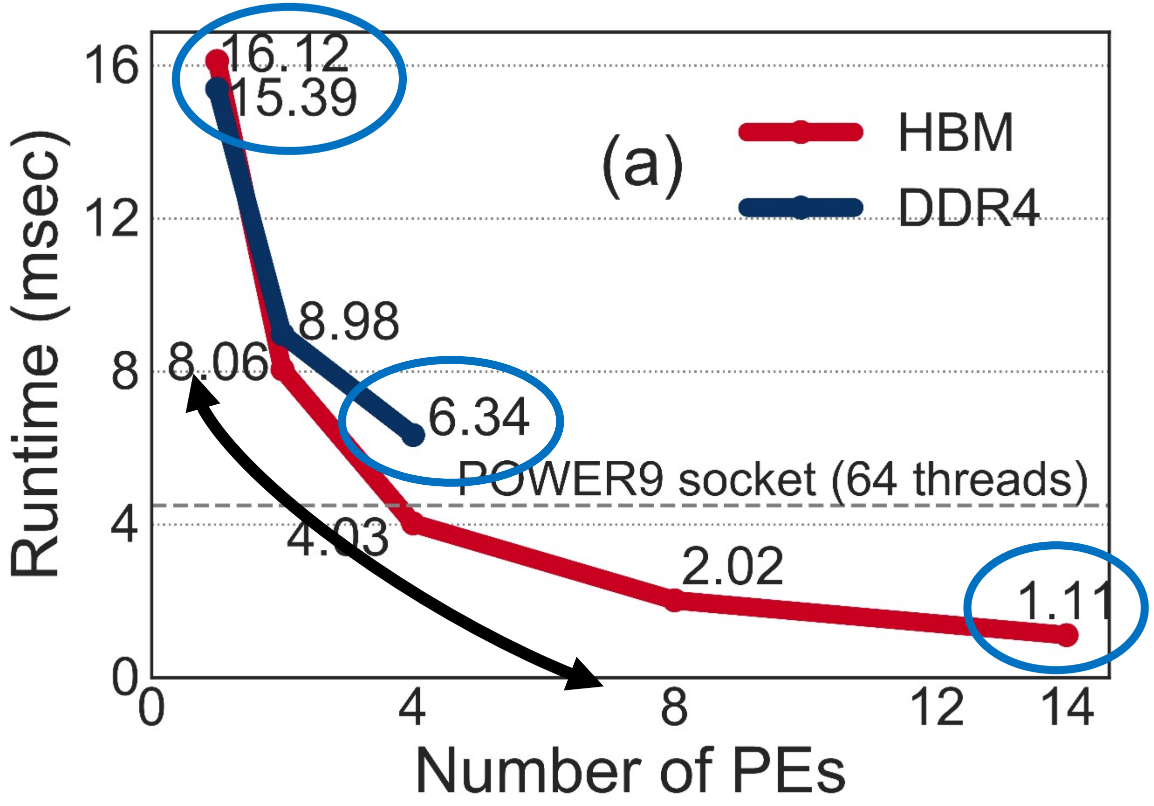
Outline

NERO Performance Analysis

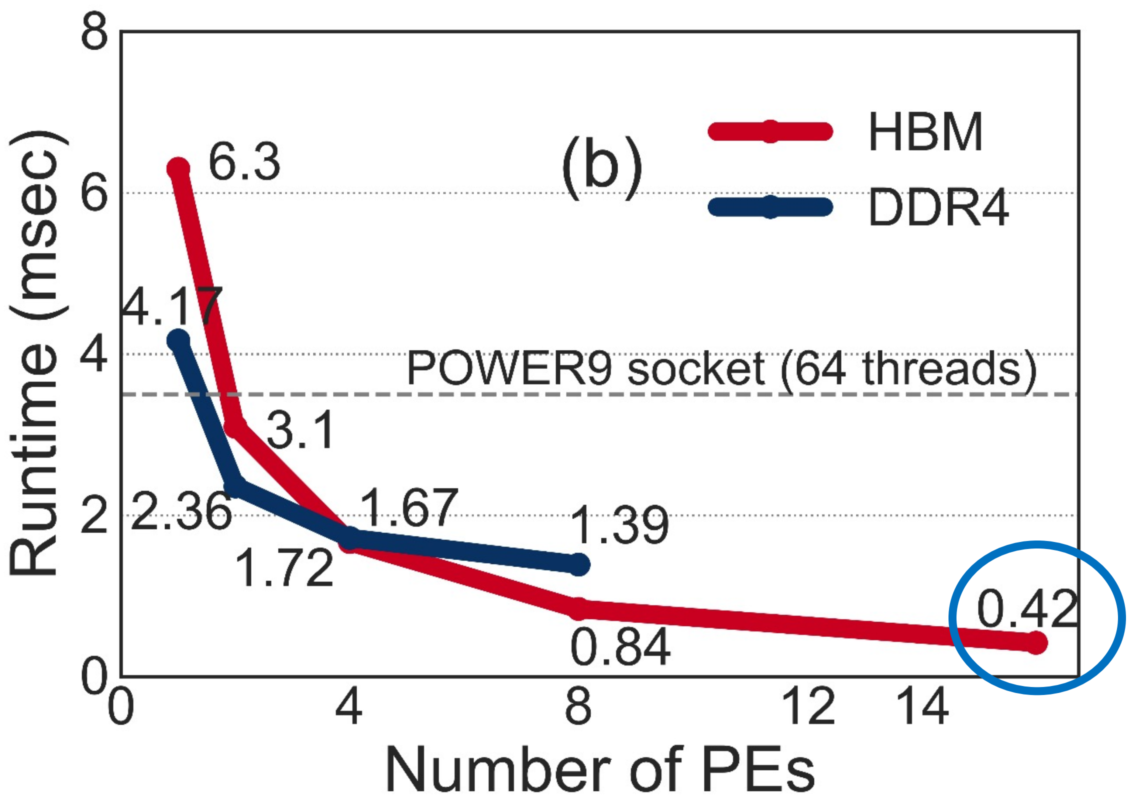


NERO Performance Analysis

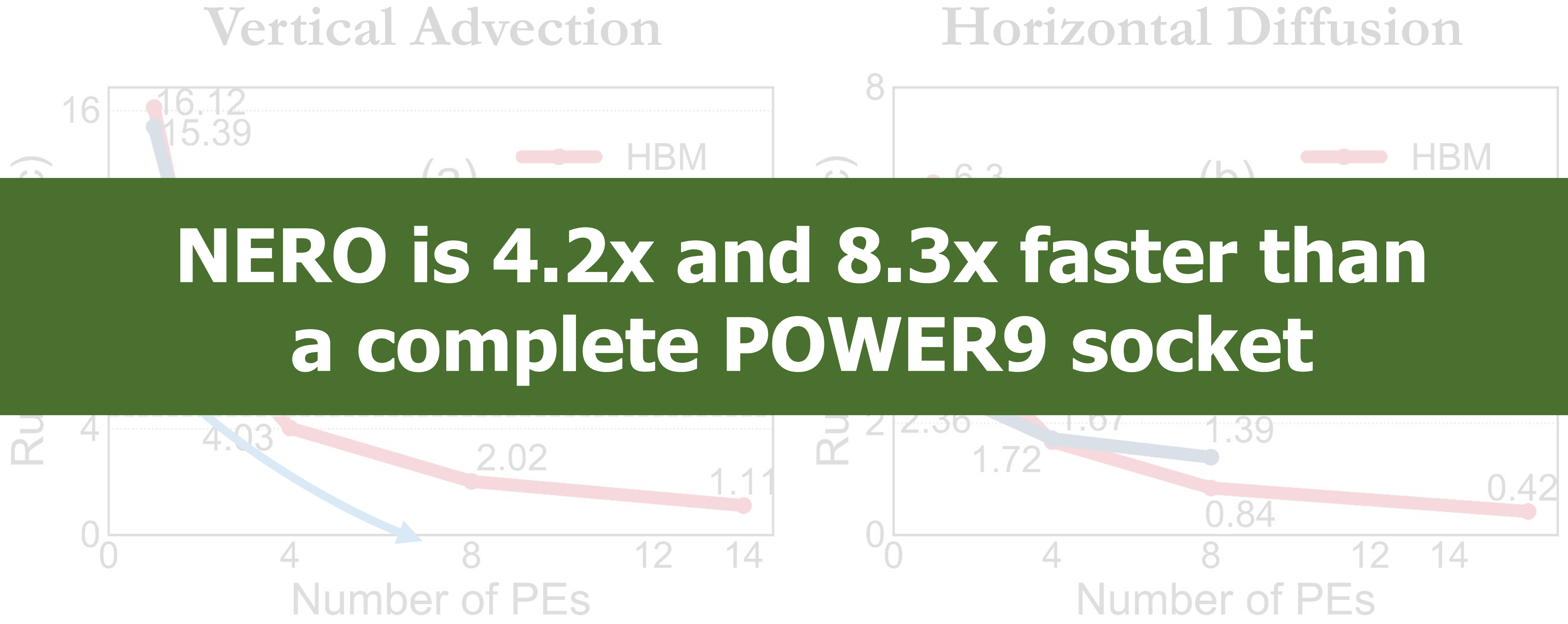
Vertical Advection



Horizontal Diffusion



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

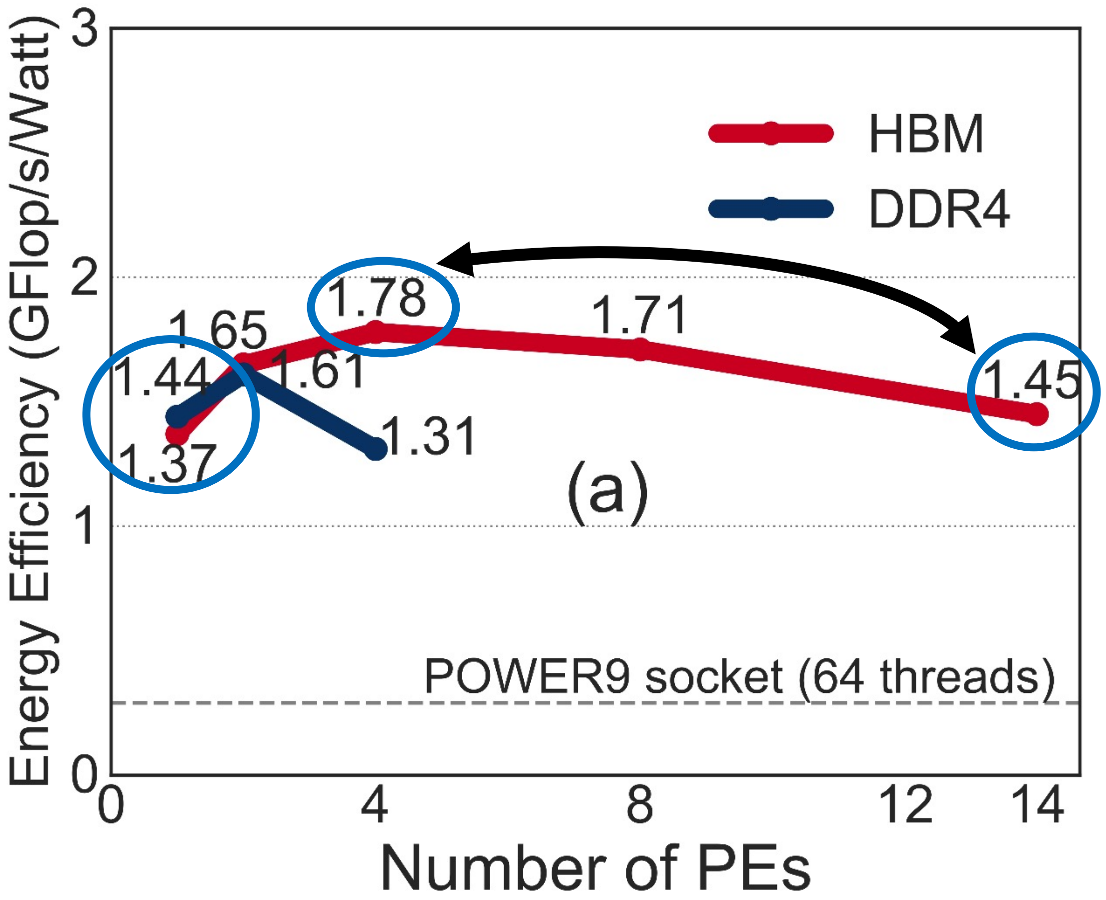
Evaluation

- Performance Analysis
- Energy Efficiency Analysis

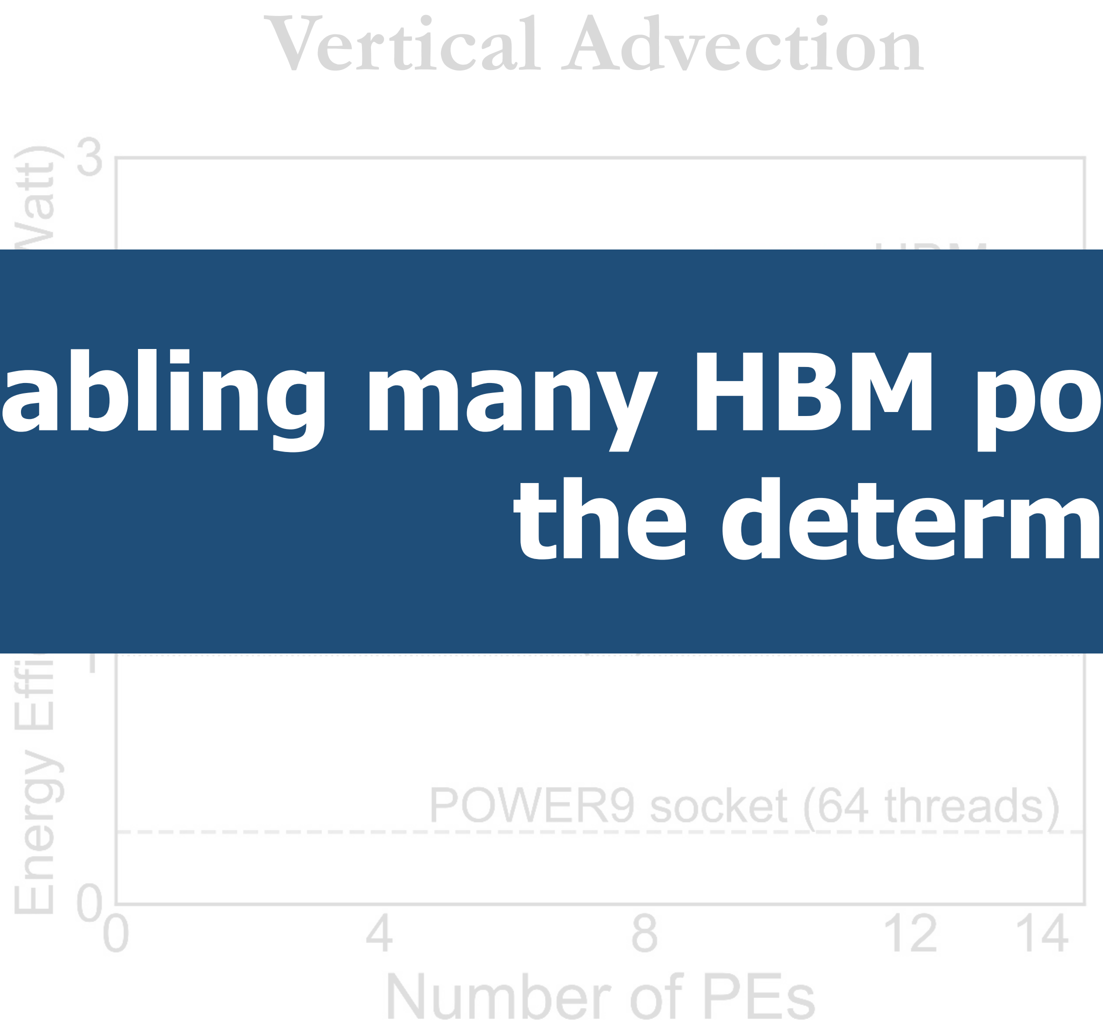
Summary

How Energy Efficient is NERO?

Vertical Advection



How Energy Efficient is NERO?

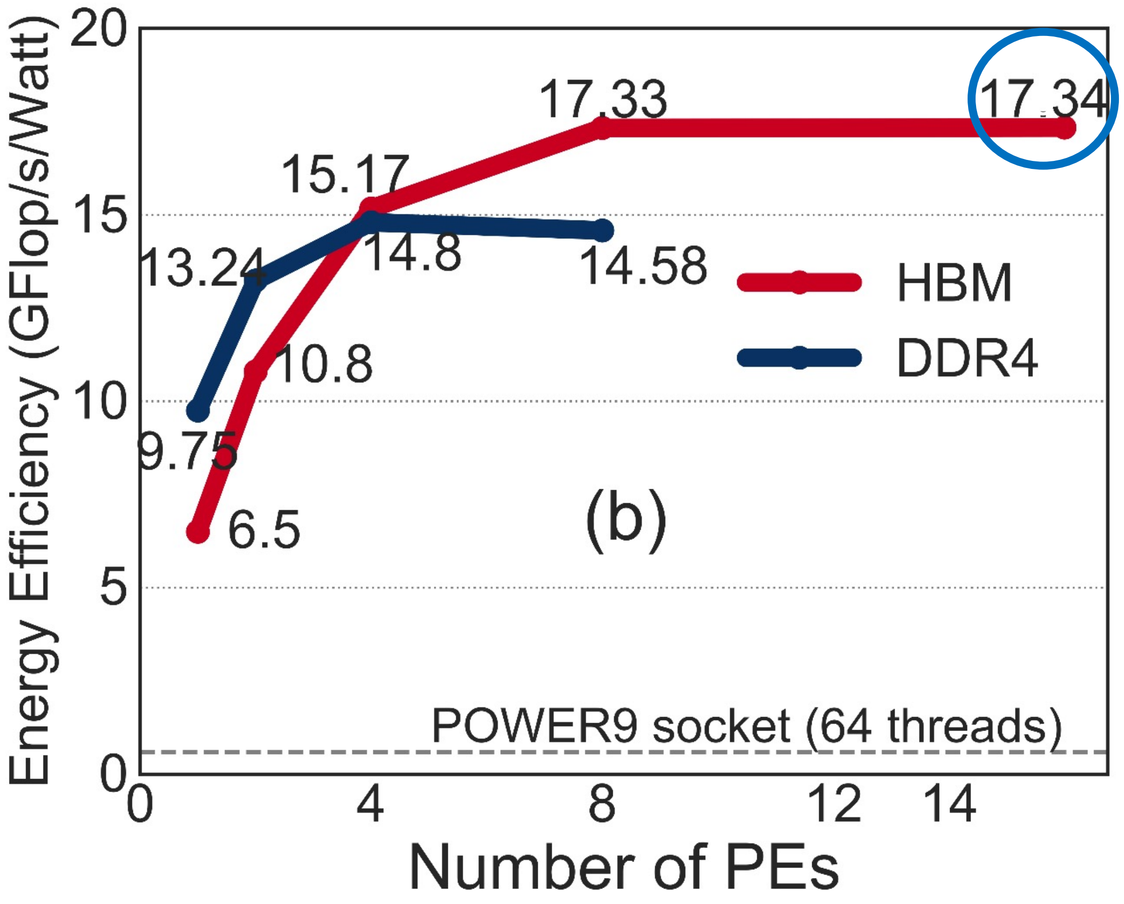


Enabling many HBM ports might not always be the determining factor

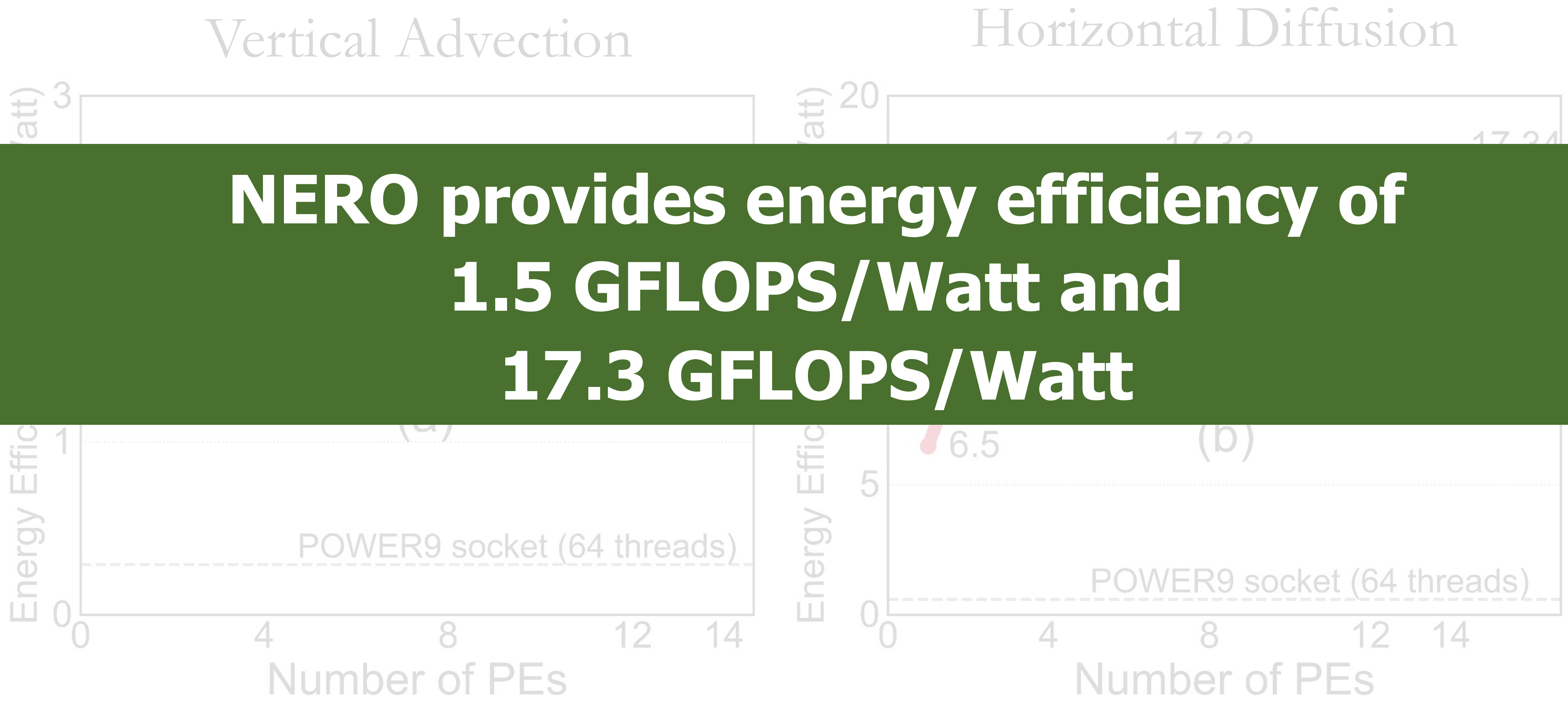
How Energy Efficient is NERO?

Vertical Advection

Horizontal Diffusion



How Energy Efficient is NERO?



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 upto 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

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

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications

Gagandeep Singh, Mohammed Alser, Damla Senol Cali,
Dionysios Diamantopoulos, Juan Gómez-Luna, Henk Corporaal, and
Onur Mutlu

Near-Memory Acceleration [IEEE Micro 2021]

- Gagandeep Singh, Mohammed Alser, Damla Senol Cali, Dionysios Diamantopoulos, Juan Gomez-Luna, Henk Corporaal, Onur Mutlu,
  "FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications"
  IEEE Micro, 2021.
  [Source Code]

Previous
Next
Table of Contents
Past Issues

Home / Magazines / IEEE Micro / 2021.04

IEEE Micro

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications

July-Aug. 2021, pp. 39-48, vol. 41
DOI Bookmark: 10.1109/MM.2021.3088396

Authors

Gagandeep Singh, ETH Zürich, Zürich, Switzerland
Mohammed Alser, ETH Zürich, Zürich, Switzerland
Damla Senol Cali, Carnegie Mellon University, Pittsburgh, PA, USA
Dionysios Diamantopoulos, Zürich Lab, IBM Research Europe, Rüschlikon, Switzerland
Juan Gomez-Luna, ETH Zürich, Zürich, Switzerland
Henk Corporaal, Eindhoven University of Technology, Eindhoven, The Netherlands
Onur Mutlu, ETH Zürich, Zürich, Switzerland

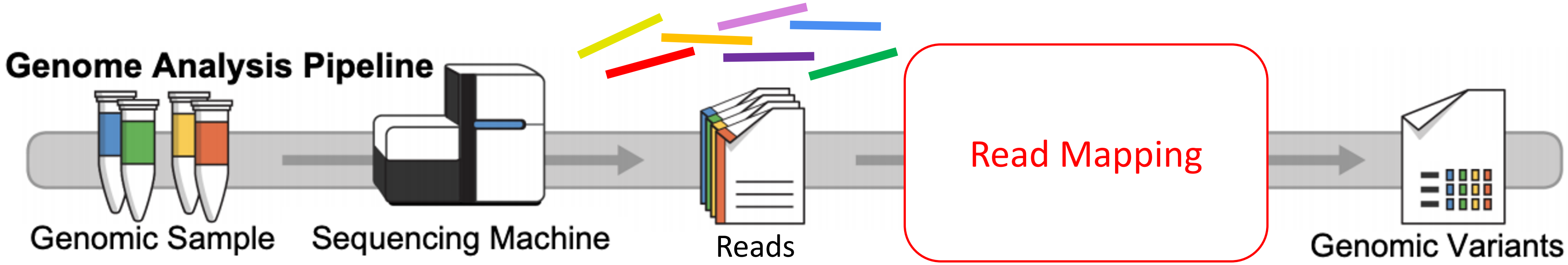
How to Analyze a Genome?

NO

machine gives the **complete sequence** of genome as output

>CCTCCTCAGTGCCACCCAGCCCACTGGCAGCTCCCAAACAGGCTCTTATTAAAACACCCTGTTCCCTGCCCCTTGGAGTGAGGTGTCAAG
GACCTAAACTAAAAAAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTCTT
CATGTCAAGGACCTAATGTGCTAAACAGCACTTTTTTGACCATTATTTTGGATCTGAAAGAAATCAAGAATAAATGAAGGACTTGATACATTG
GAAGAGGAGAGTCAAGGACCTACAGAAAAAAAAAAAAAAAGAAAAAGAAAAGAAAAAGA**A**TTTAAAATTTAAGTAATTCTTTGAAAAAA
ACTAATTTCTAAGCTTCTT**C**ATGTCAAGGACCTAATGTCTGTGTTGCAGGTCTTCTTGCATTTCCCTGTCAAAAGAAAAAGAATTTAAAATTT
AAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTCAGGCCAAGAGTTGCAAAAAAAAAAAAAAGAAAAA
GAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTAGCCAGAATGG
TTGTGGGATGGGAGCCTCTGTGGACCGACCAGGTAGCTCTCTTTTCCACACTGTAGTCTCAAAGCTTCTTCATGTGGTTTCTCTGAGTGAAA
AAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTT**T**TTCATGTCAAGGACC
TAATGTAGCTATACTGAACGTTATCTAGGGGAAAGATTGAAGGGGAGCTCTAAGGTCAACACACCACCACTTCCCAGAAAGCTTCTTCA……

Genome Analysis in Real Life



Current sequencing machine provides
small randomized fragments
of the original DNA sequence

Alser+, "Technology dictates algorithms: Recent developments in read alignment", Genome Biology, 2021

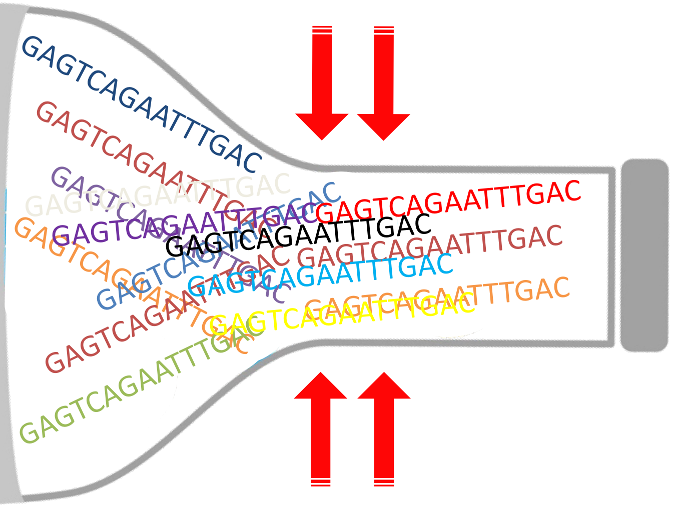
Bottlenecked in Read Mapping!!

48 Human whole genomes

at 30× coverage

in about 2 days

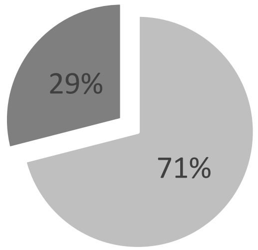
Illumina NovaSeq 6000



1 Human genome

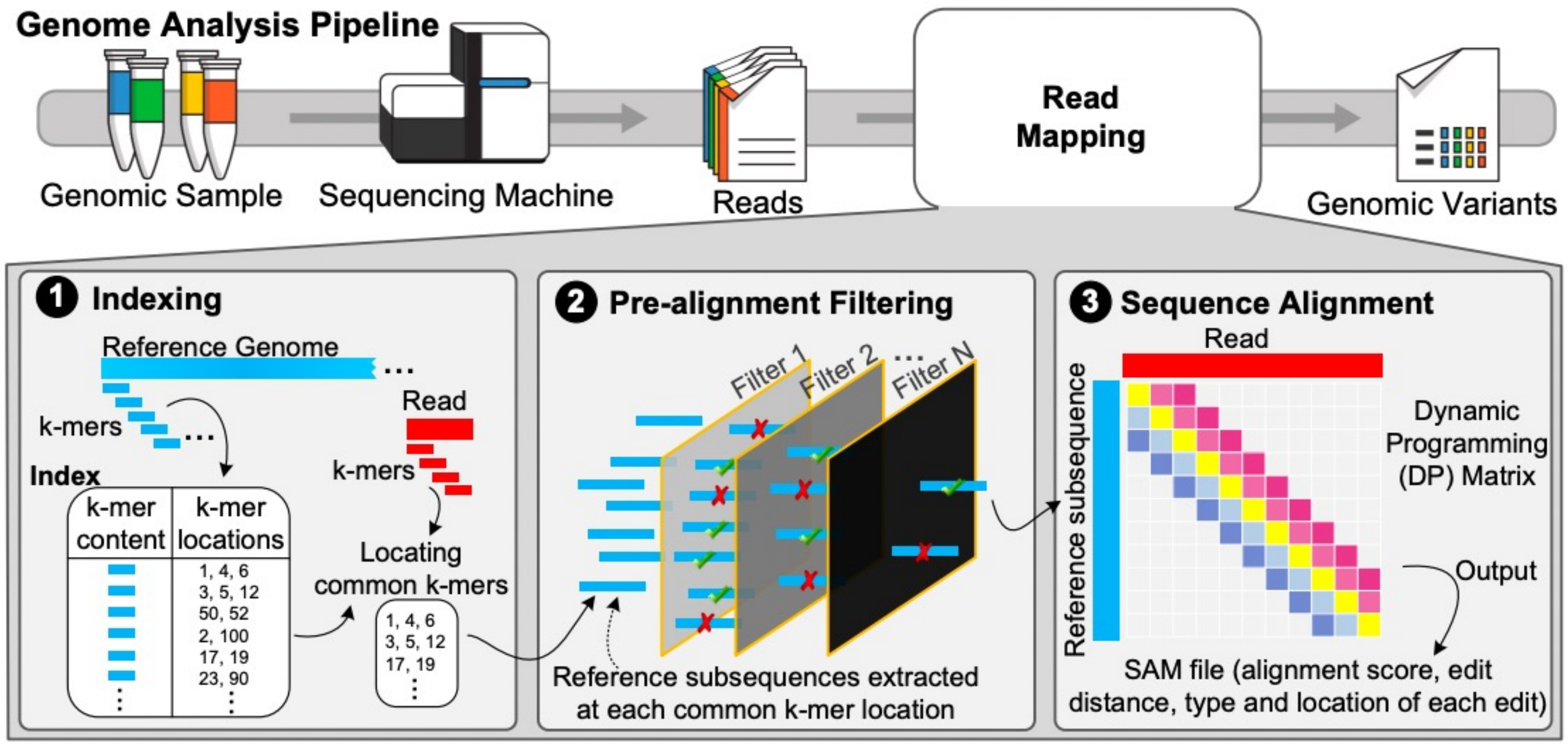
32 CPU hours

on a 48-core processor



Goyal+, "[Ultra-fast next generation human genome sequencing data processing using DRAGENTM bio-IT processor for precision medicine](#)", *Open Journal of Genetics*, 2017.

Accelerating Read Mapping

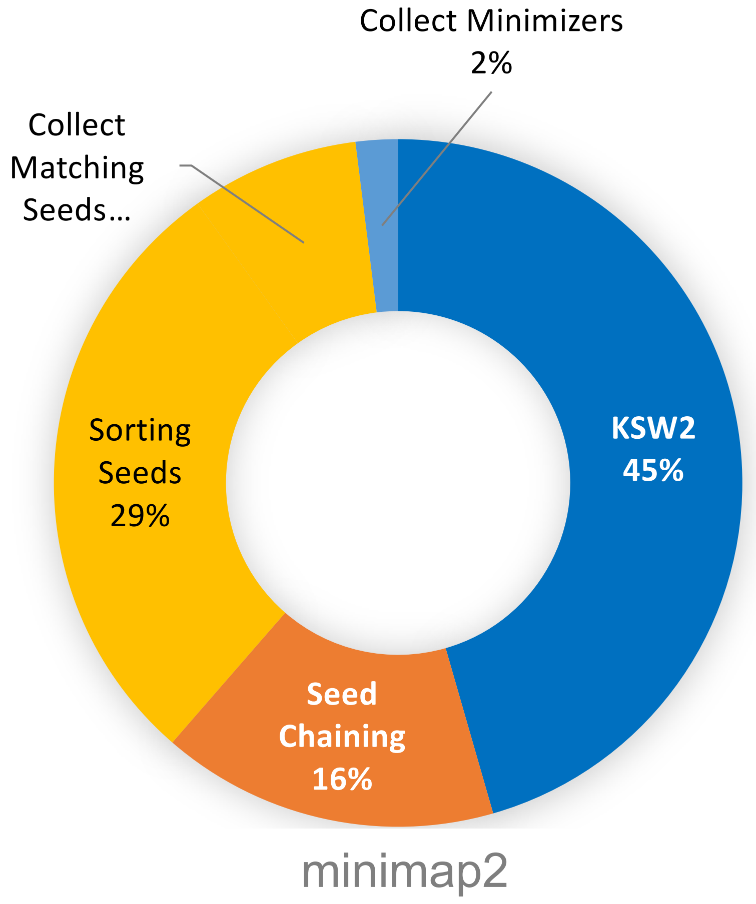


Alser+, "Accelerating Genome Analysis: A Primer on an Ongoing Journey", IEEE Micro, 2020.

Read Mapping Execution Time

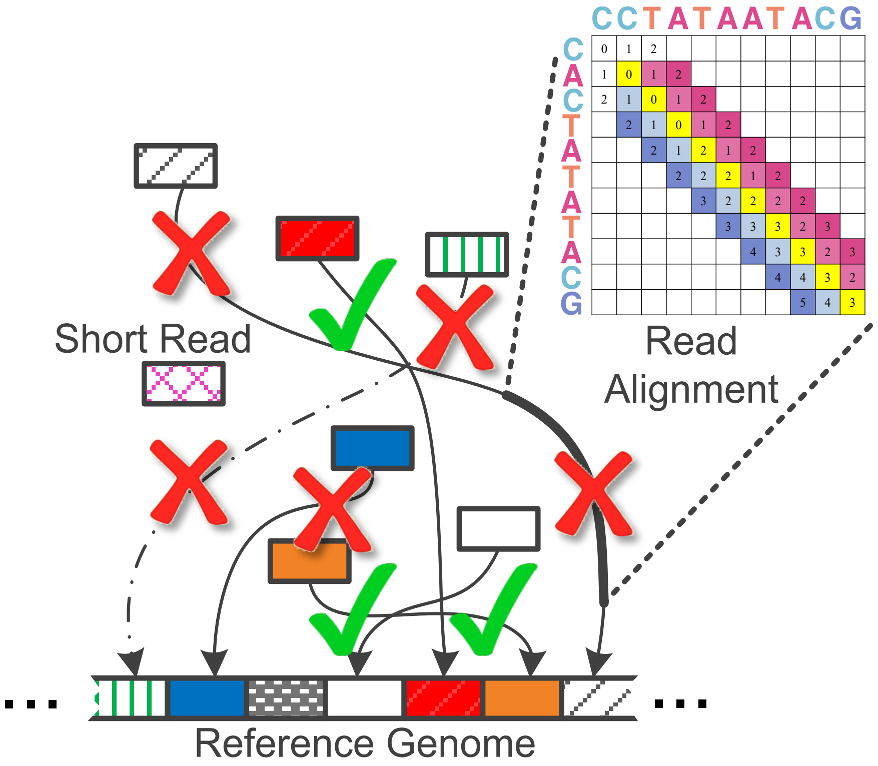
>60%

of the read mapper's execution time is spent in sequence alignment



ONT FASTQ size: 103MB (151 reads), Mean length: 356,403 bp, std: 173,168 bp, longest length: 817,917 bp

Large Search Space for Mapping Location

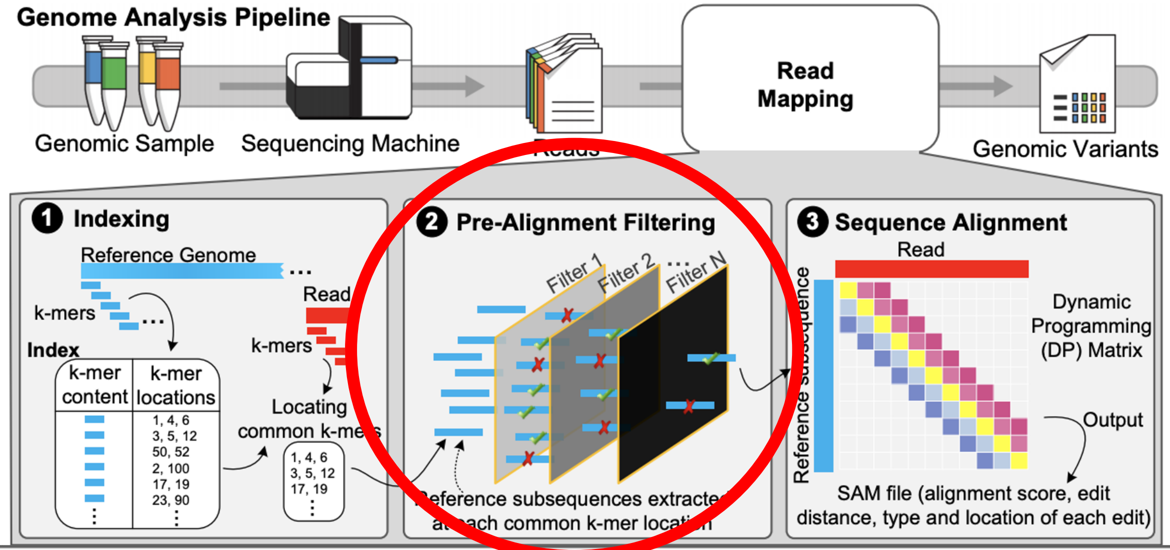


98%

of candidate locations have high dissimilarity with a given read

Cheng *et al*, *BMC bioinformatics* (2015)
Xin *et al, BMC genomics* (2013)

Accelerating Read Mapping



Accelerating Indexing

- Reducing the number of seeds
- Reducing data movement during indexing

Accelerating Pre-Alignment Filtering

- q-gram filtering
- Pigeonhole principle
- Base counting
- Sparse DP

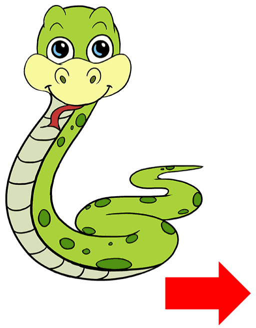
Accelerating Alignment

- Accurate alignment accelerators
- Heuristic-based alignment accelerators

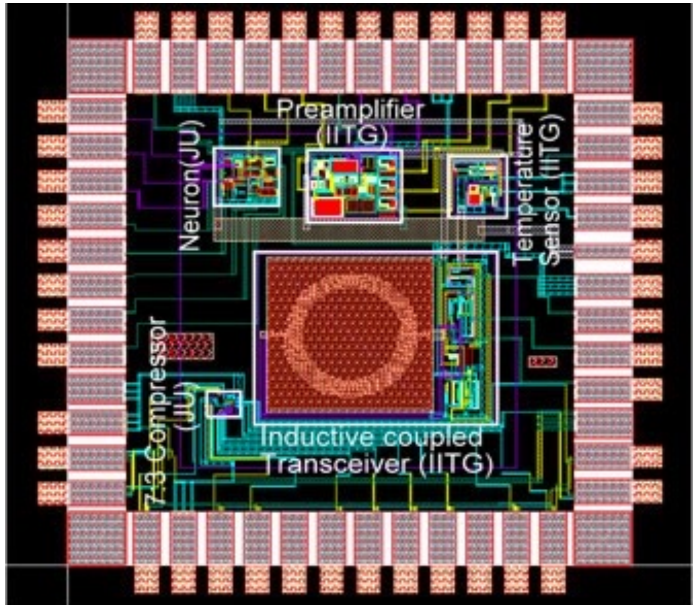
Alser+, "Accelerating Genome Analysis: A Primer on an Ongoing Journey", IEEE Micro, 2020.

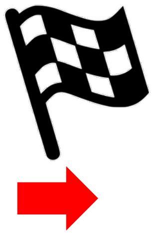
SneakySnake

- **Key observation:**
  - Correct alignment is a sequence of non-overlapping long matches
- **Key idea:**
  - Approximate edit distance calculation is similar to Single Net Routing problem in VLSI chip



VLSI chip layout





Alser+, "SneakySnake: a fast and accurate universal genome pre-alignment filter for CPUs, GPUs and FPGAs", Bioinformatics, 2020.

Stencil Computation in Weather Modeling

COSMO (Consortium for Small-Scale Modeling)

- Around **80 complex stencils**

- Horizontal diffusion
  Vertical advection

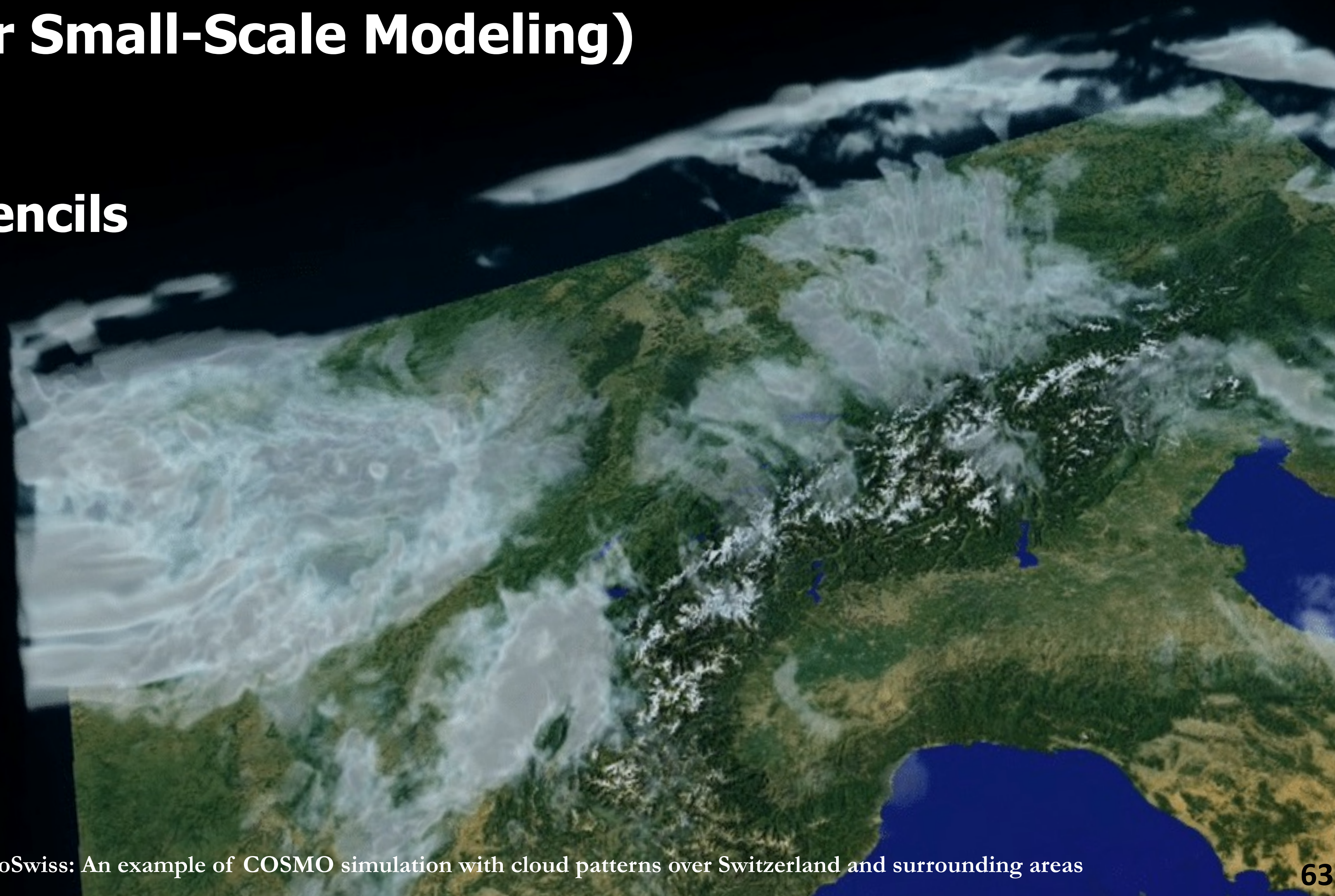
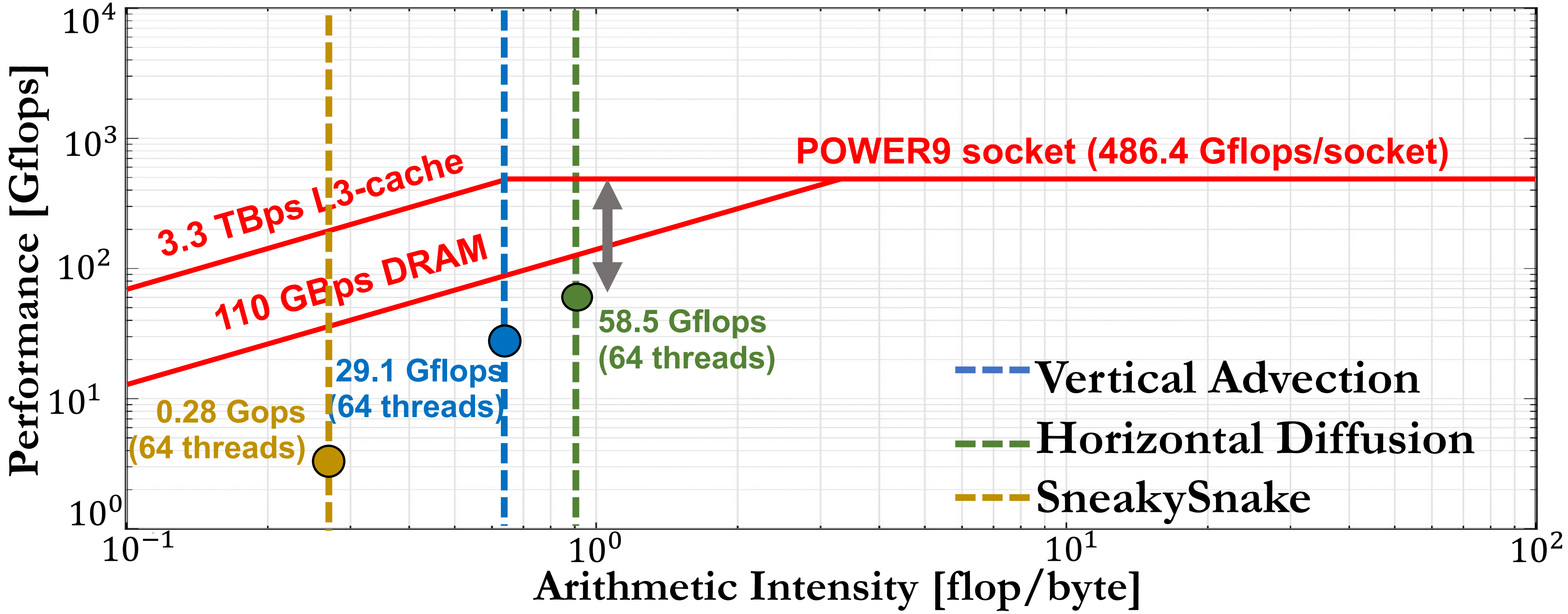


Image Source: NVIDIA/MeteoSwiss: An example of COSMO simulation with cloud patterns over Switzerland and surrounding areas

Motivation and Goal

Memory bound with **limited performance** and **high energy consumption** on **IBM POWER9 CPU**



Goal:

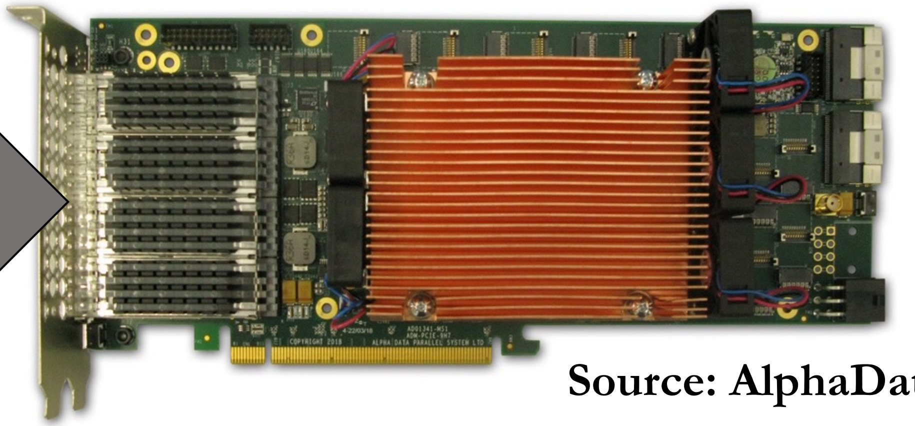
- **Mitigate** the **performance bottleneck** of modern data-intensive applications in an **energy-efficient way**
- Evaluate the use of **near-memory acceleration** using a **FPGA+HBM** connected through **IBM CAPI2** (Coherent Accelerator Processor Interface)/**OCAPI** (OpenCAPI)

Heterogeneous System: CPU+FPGA

Source: IBM

POWER9 AC922

OCAPI



Source: AlphaData

HBM-based AD9H7 board

We evaluate:

I. Two POWER9+FPGA systems:

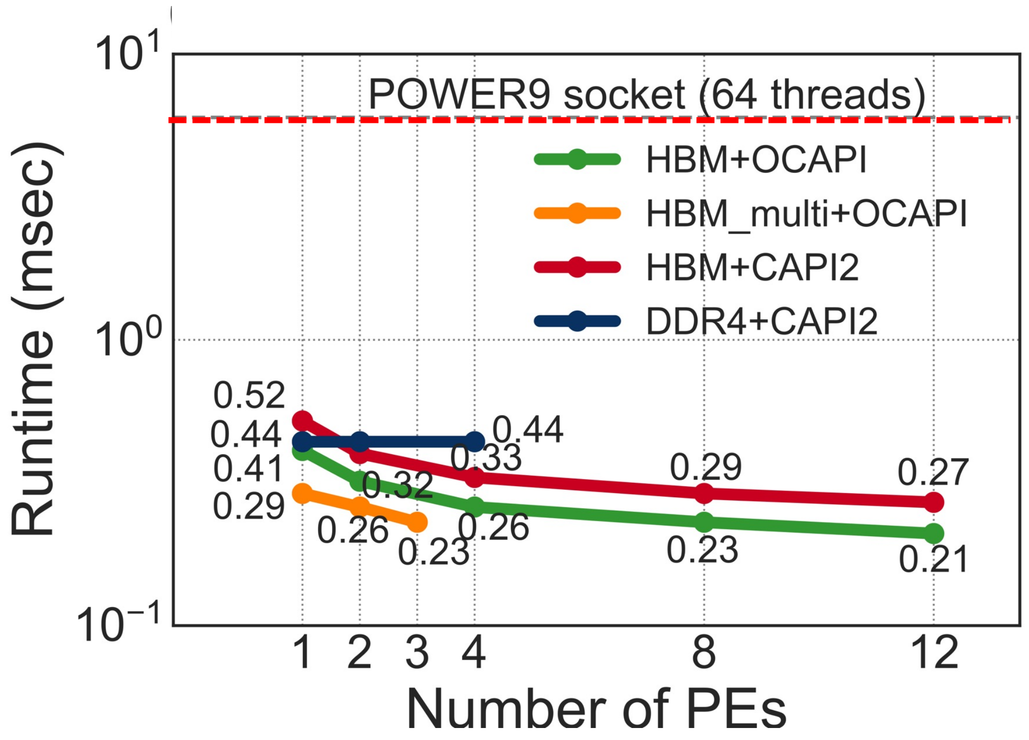
1. HBM-based AD9H7 board
Xilinx Virtex Ultrascale+™ XCVU37P-2

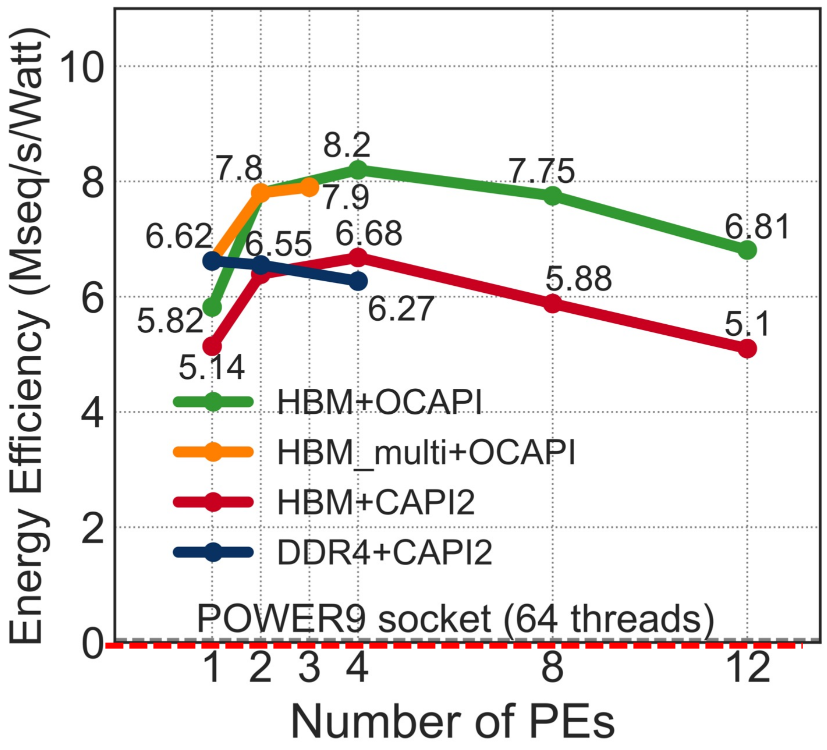
2. DDR4-based AD9V3 board
Xilinx Virtex Ultrascale+™ XCVU3P-2

II. Two interconnect technologies: **CAPI2** and **OCAPI**

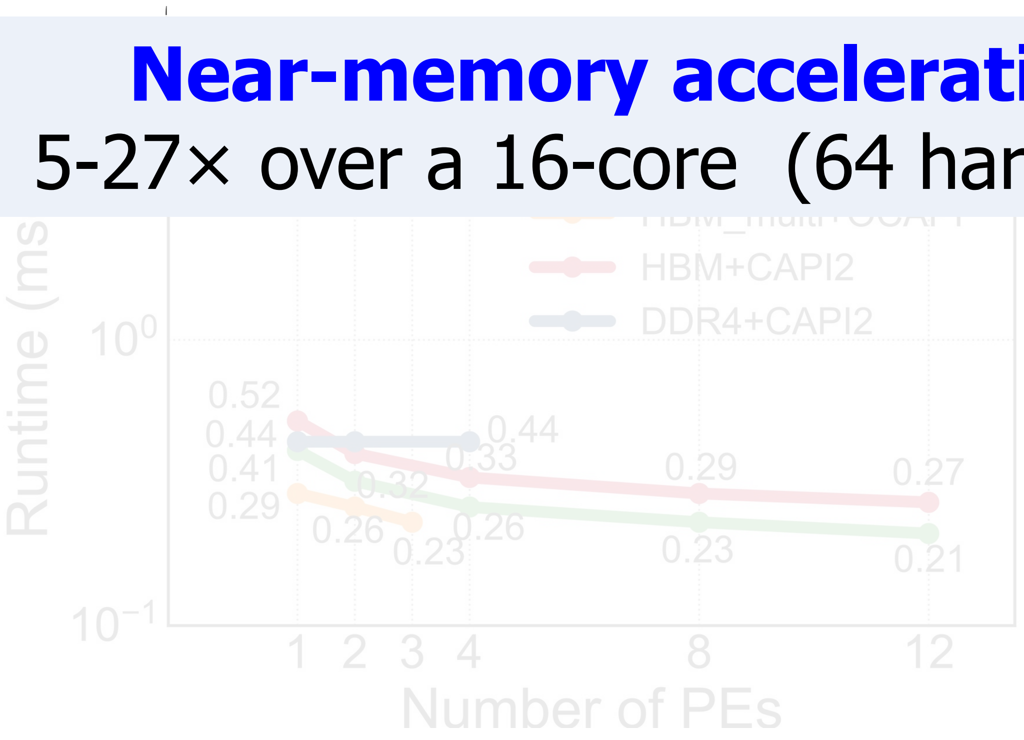
III. Two processing element (PE) designs: **single channel** and **multiple channel**

Results: Performance Comparison





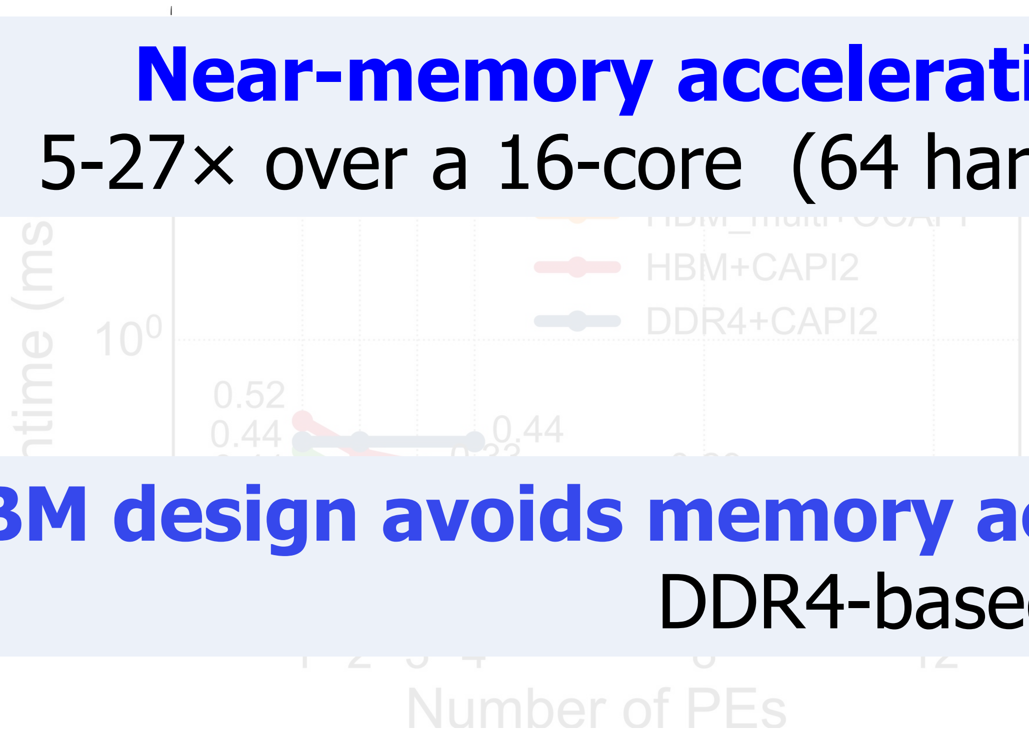
Results: Performance Comparison

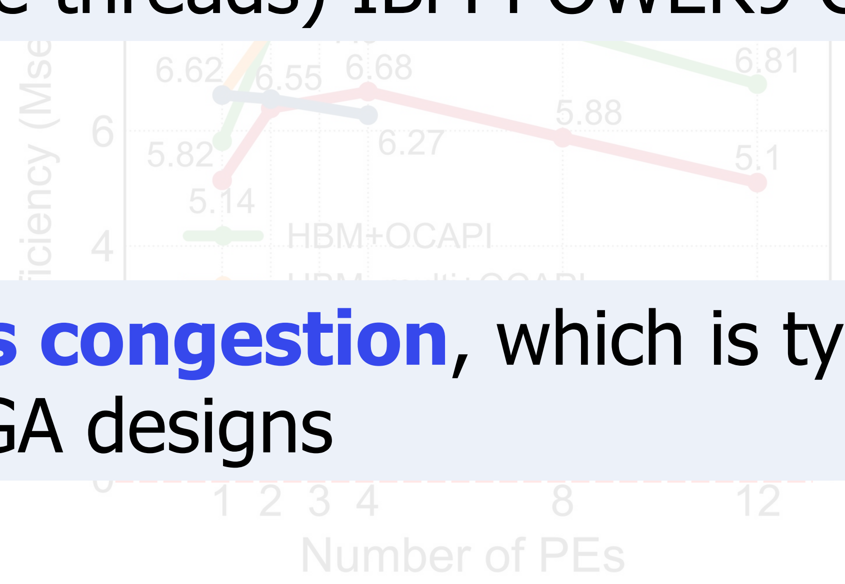


Near-memory acceleration improves **performance** by
5-27× over a 16-core (64 hardware threads) IBM POWER9 CPU

Results: Performance Comparison

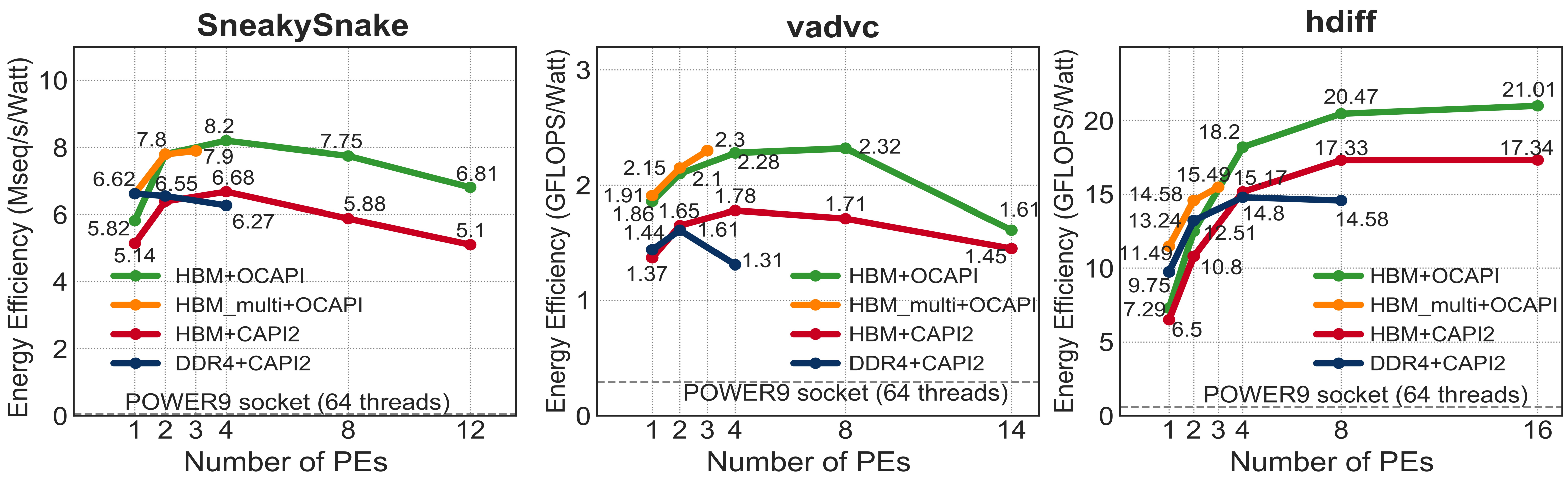
Near-memory acceleration improves **performance** by
5-27× over a 16-core (64 hardware threads) IBM POWER9 CPU





HBM design avoids memory access congestion, which is typical in
DDR4-based FPGA designs

Results: Energy Efficiency Comparison



Results: Energy Efficiency Comparison

Near-memory acceleration improves **energy efficiency** by 12-133×, respectively, over a 16-core (64 hardware threads) IBM POWER9 CPU

Results: Energy Efficiency Comparison

Near-memory acceleration improves **energy efficiency** by 12-133×, respectively, over a 16-core (64 hardware threads) IBM POWER9 CPU

Single channel & multiple channel HBM designs
Open-source: https://github.com/CMU-SAFARI

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications

Gagandeep Singh, Mohammed Alser, Damla Senol Cali,
Dionysios Diamantopoulos, Juan Gómez-Luna, Henk Corporaal, and Onur Mutlu

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

Computer Architecture
Lecture 24: Cutting-Edge Research in Computer Architecture III

Dr. Gagandeep Singh
Postdoctoral Researcher

December 23rd 2021

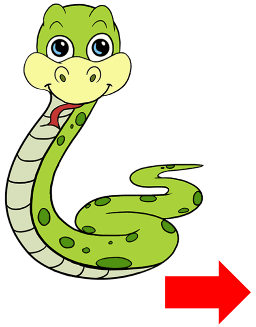
SAFARI

ETH Zürich

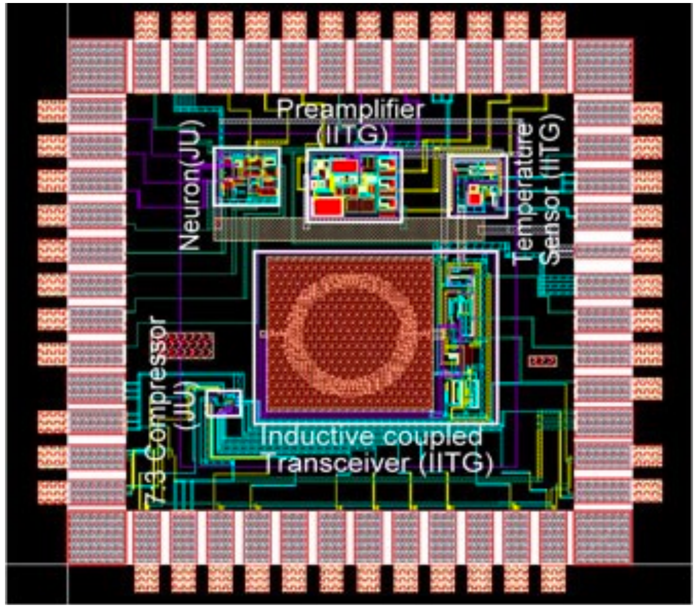
Backup

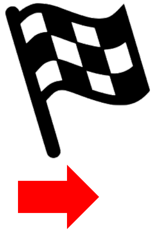
SneakySnake

- **Key observation:**
  - Correct alignment is a sequence of non-overlapping long matches
- **Key idea:**
  - Approximate edit distance calculation is similar to Single Net Routing problem in VLSI chip



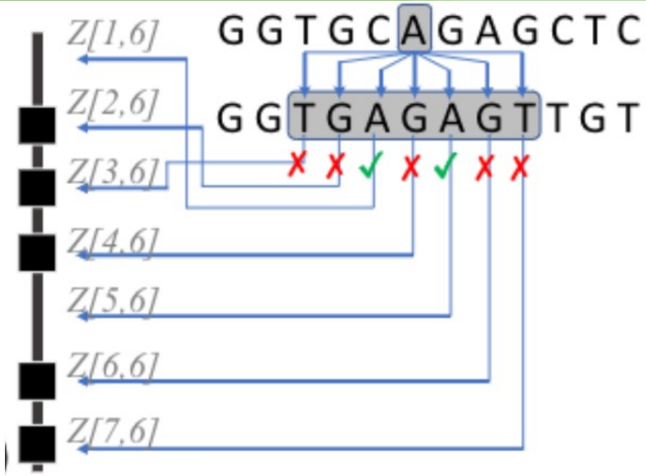
VLSI chip layout





SneakySnake Walkthrough

Building Neighborhood Map → Finding the Routing Travel Path → Examining the Snake Survival



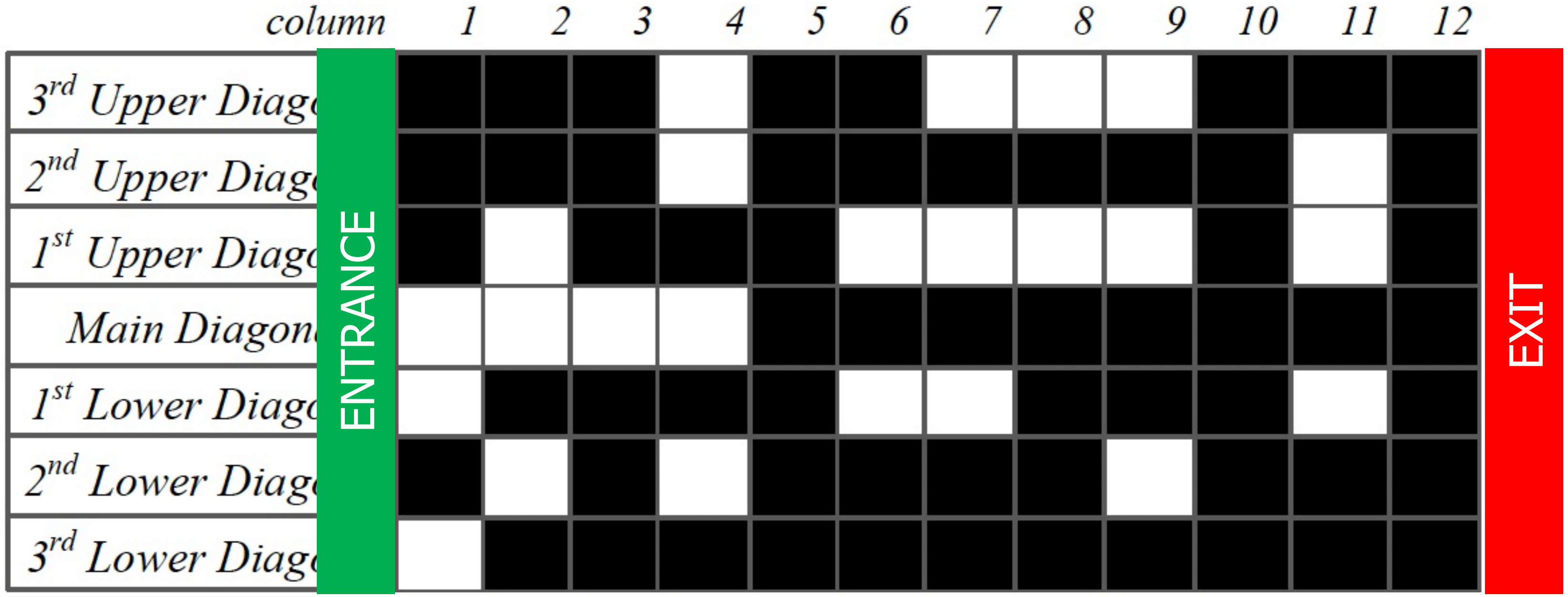
$E = 3$

column	1	2	3	4	5	6	7	8	9	10	11	12
3rd Upper Diagonal	1	1	1	0	1	1	0	0	0	1	1	1
2nd Upper Diagonal	1	1	1	0	1	1	1	1	1	1	0	1
1st Upper Diagonal	1	0	1	1	1	0	0	0	0	1	0	1
Main Diagonal	0	0	0	0	1	1	1	1	1	1	1	1
1st Lower Diagonal	0	1	1	1	1	0	0	1	1	1	0	1
2nd Lower Diagonal	1	0	1	0	1	1	1	1	0	1	1	1
3rd Lower Diagonal	0	1	1	1	1	1	1	1	1	1	1	1

SneakySnake Walkthrough

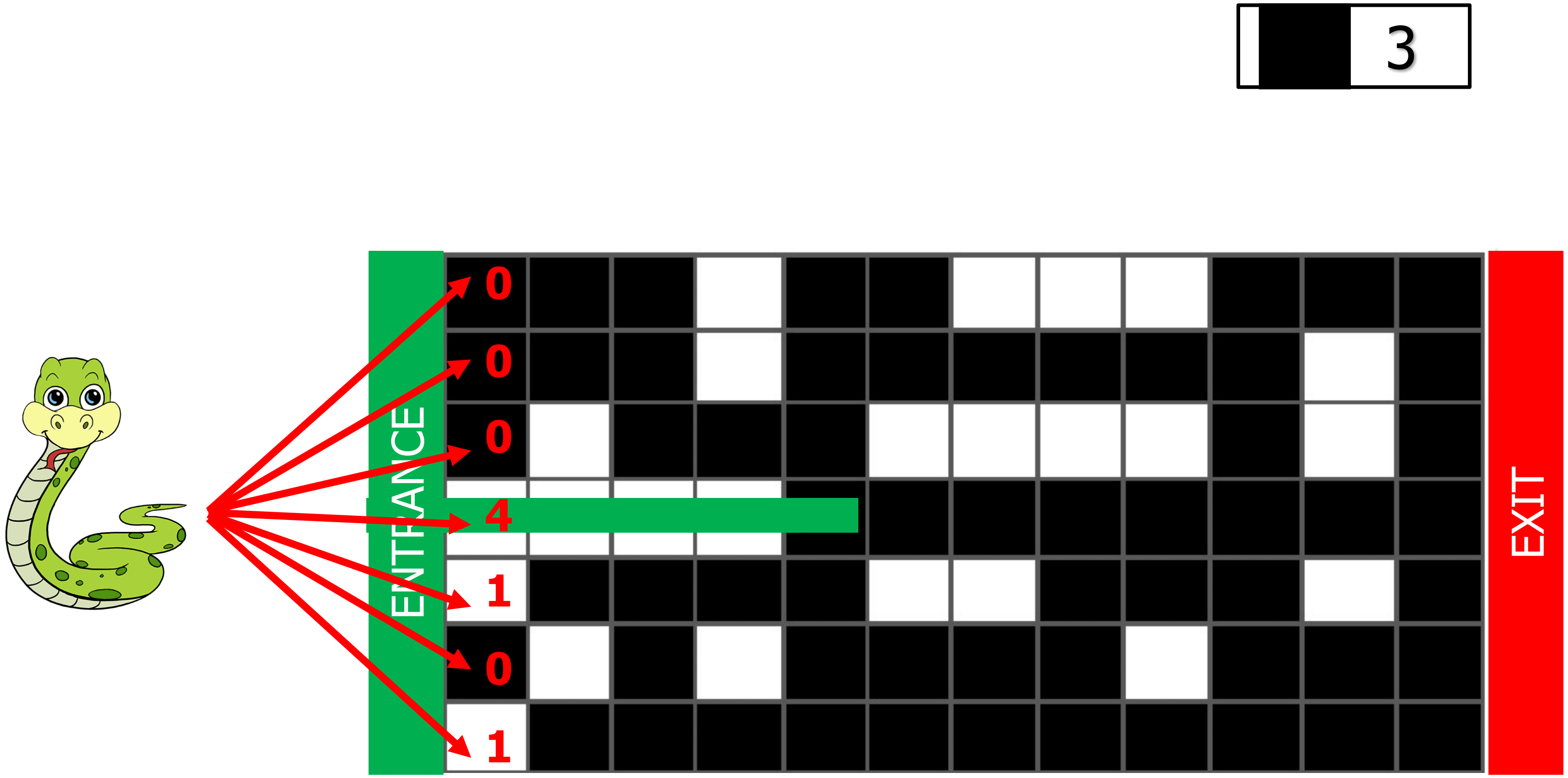
Building Neighborhood Map | Finding the Routing Travel Path | Examining the Snake Survival

$E = 3$



SneakySnake Walkthrough

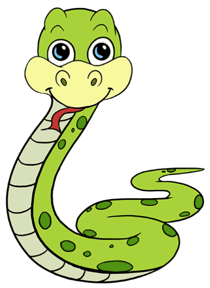
Building Neighborhood Map → Finding the Routing Travel Path → Examining the Snake Survival

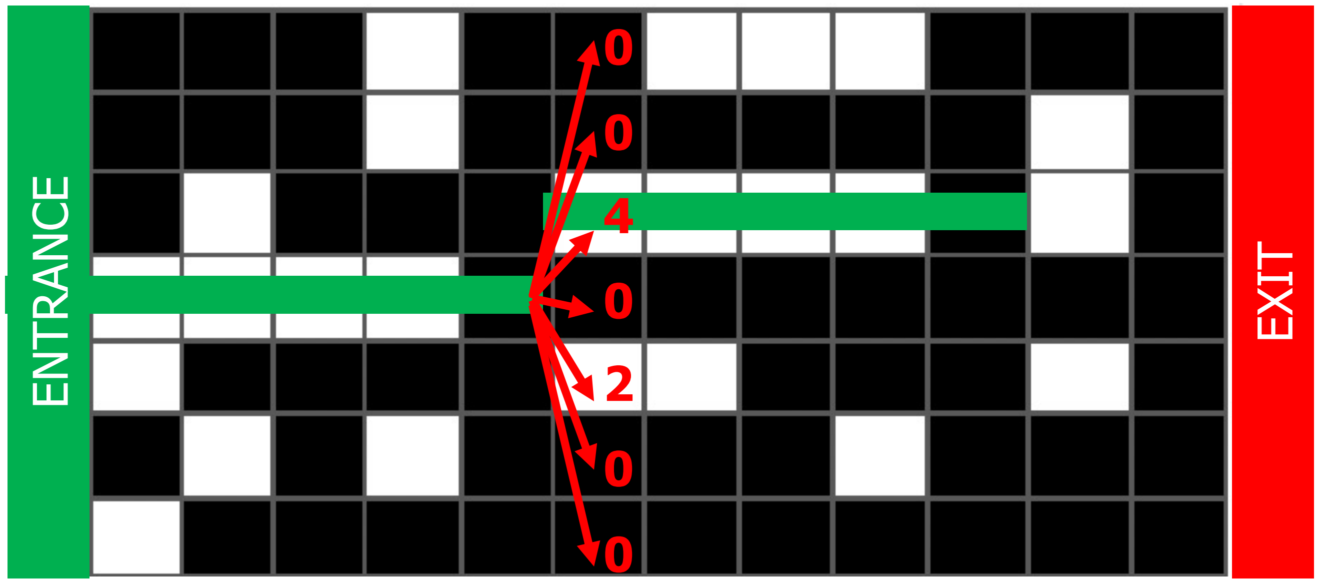


SneakySnake Walkthrough

Building Neighborhood Map → Finding the Routing Travel Path → Examining the Snake Survival

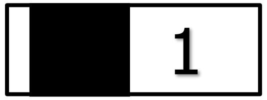
2

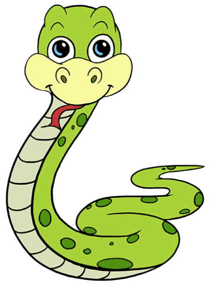


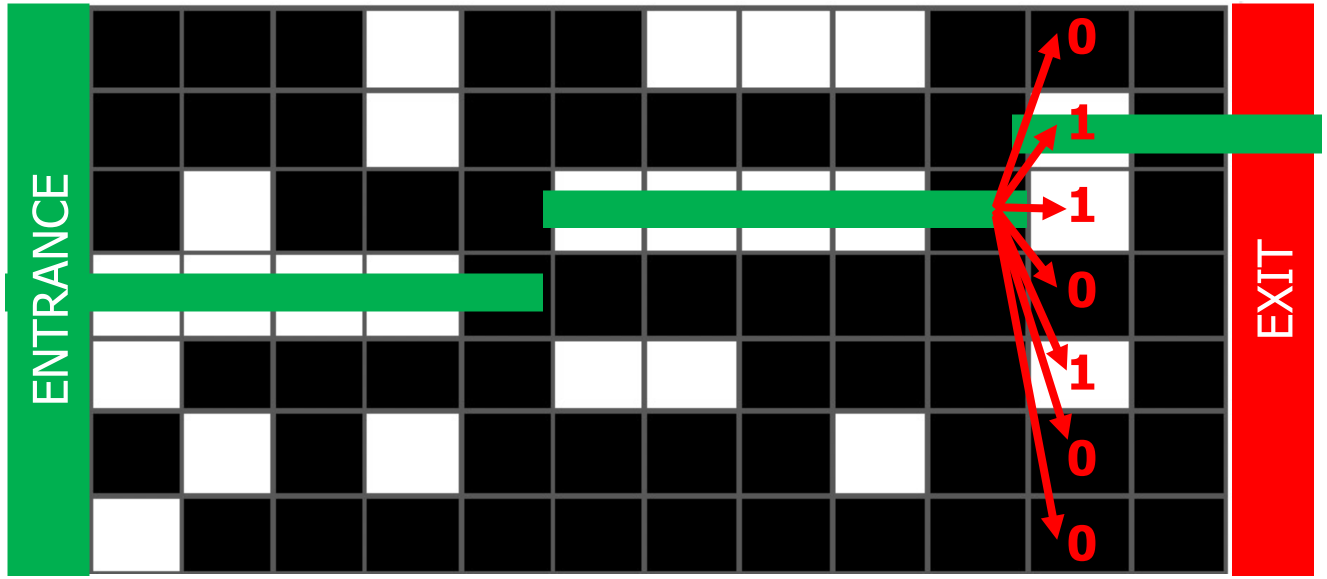


SneakySnake Walkthrough

Building Neighborhood Map → Finding the Routing Travel Path → Examining the Snake Survival

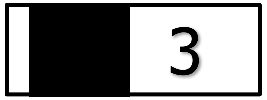




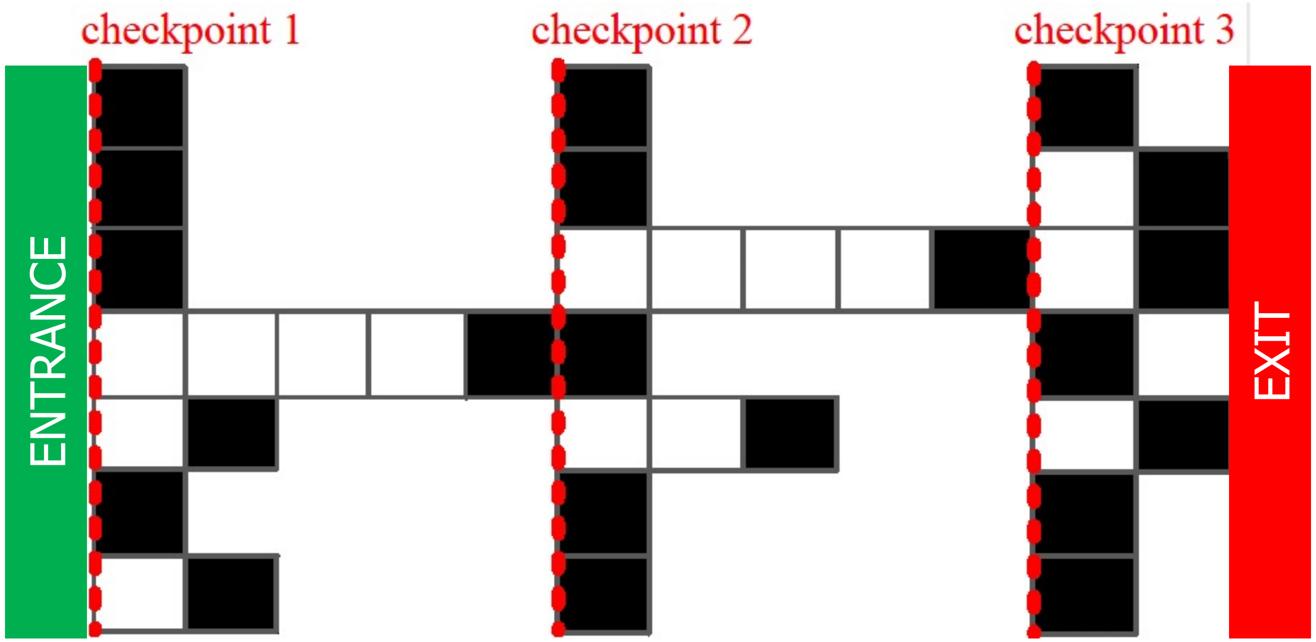


SneakySnake Walkthrough

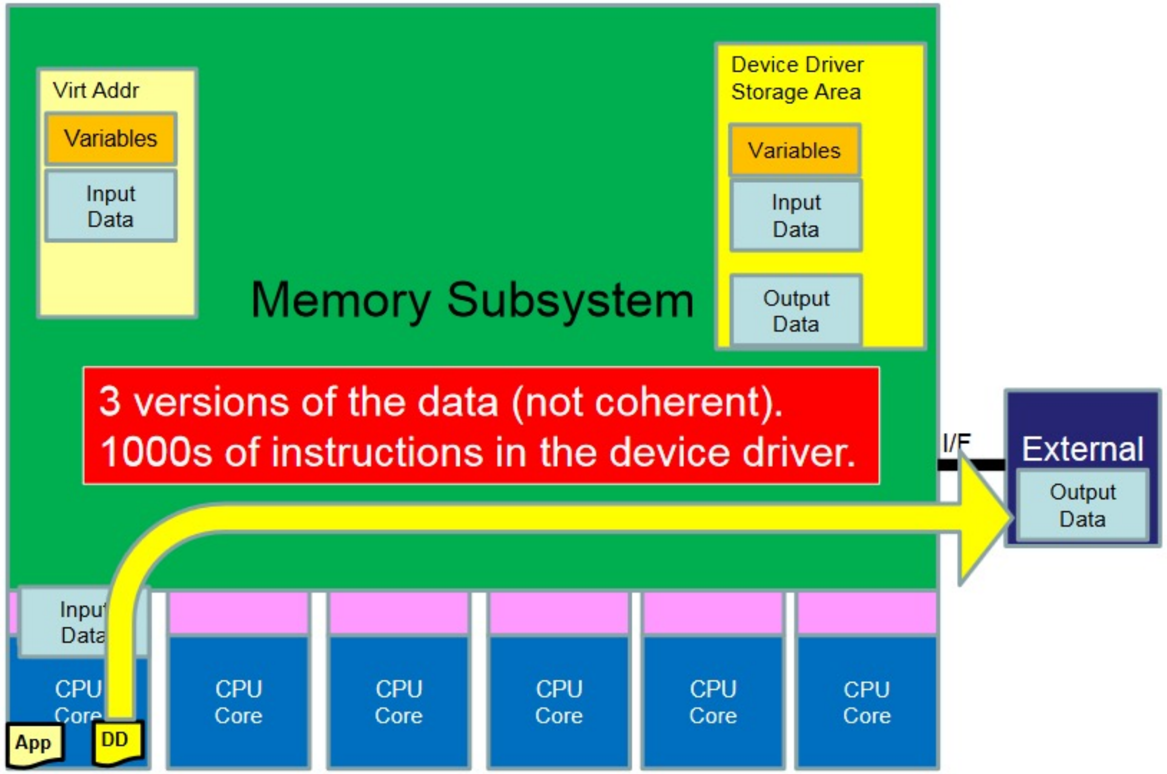
Building Neighborhood Map → Finding the Routing Travel Path → Examining the Snake Survival

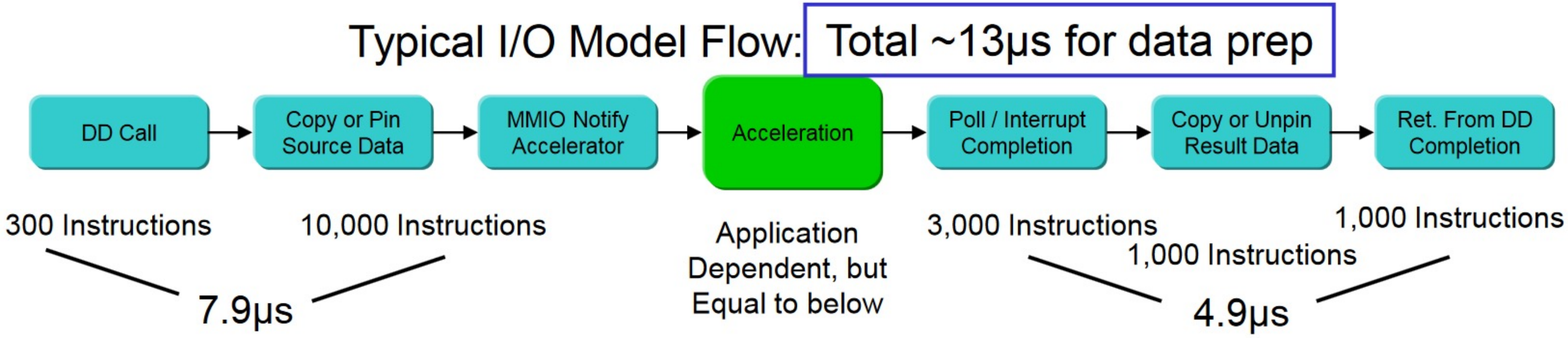


This is what you actually need to build and it can be done on-the-fly!

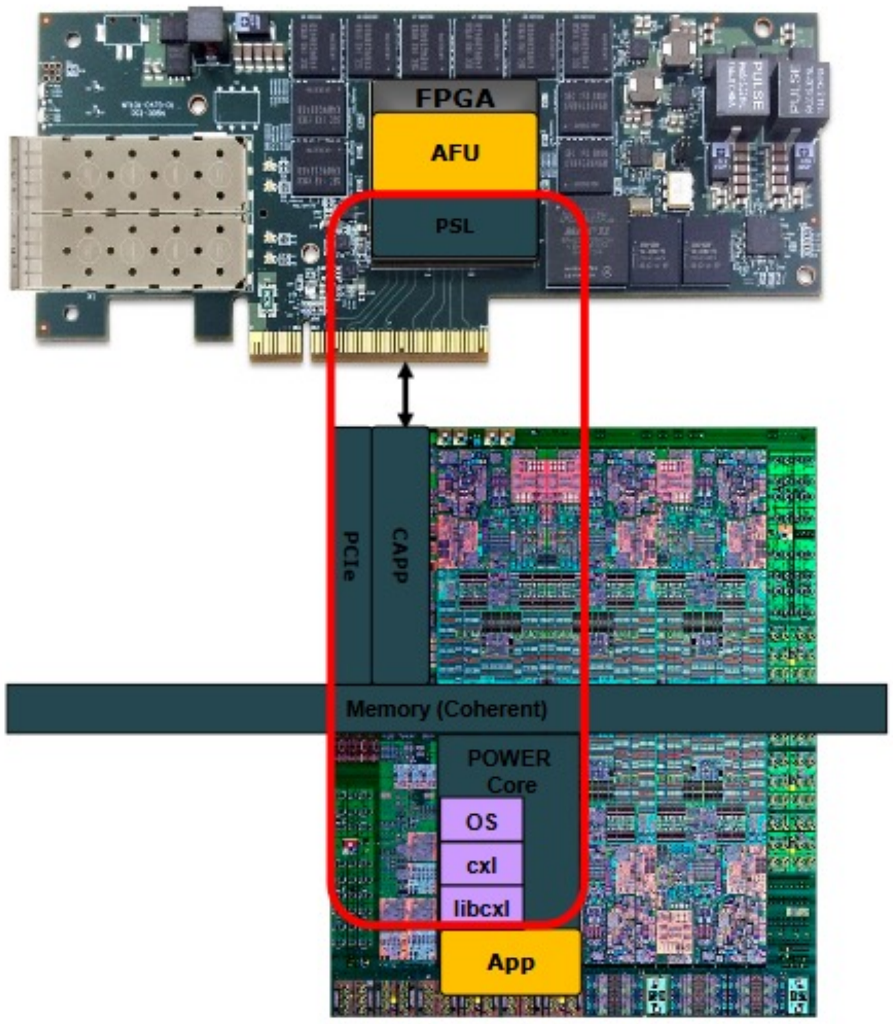


Background: Traditional I/O Technology

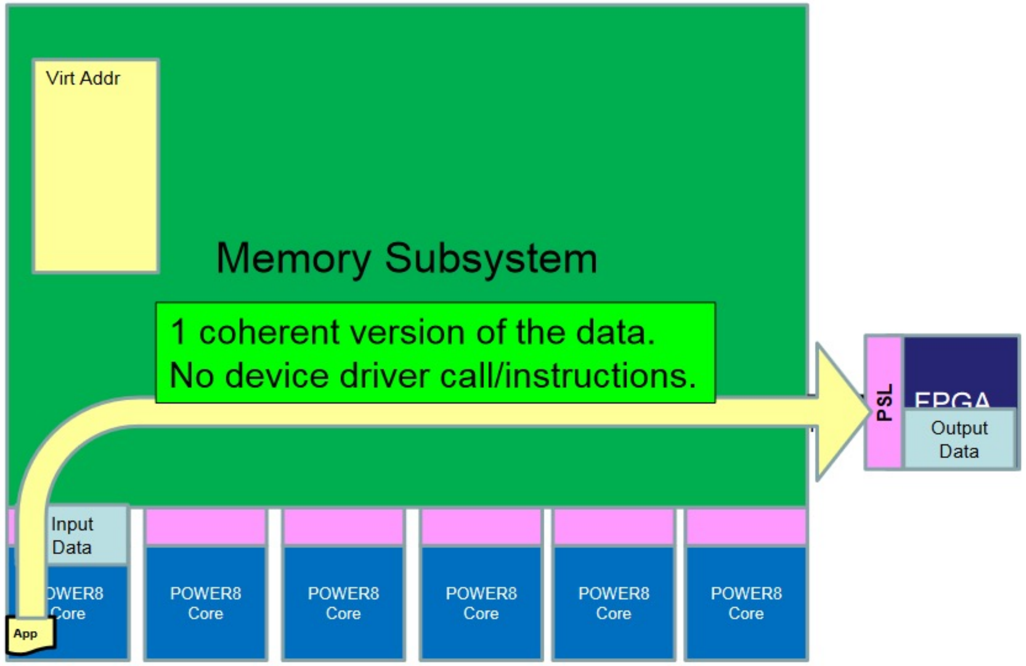


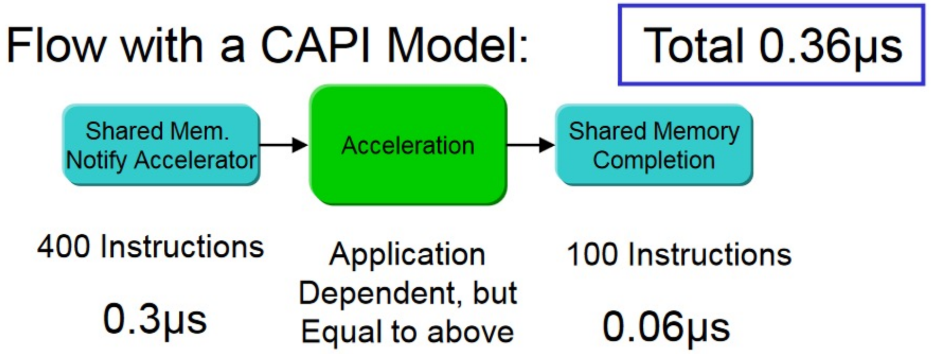


CAPI Overview



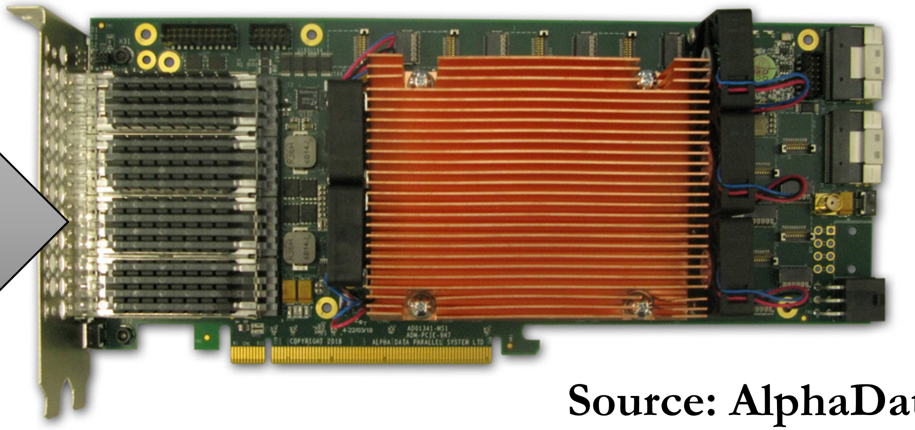
POWER8 – POWER9 Processor





C1 Mode for Weather Acceleration

Source: IBM



Source: AlphaData

POWER9 AC922

HBM-based AD9H7 board

- **Host System**
  IBM POWER9-16 core (64-threads)

- **FPGA board**
  Xilinx Virtex® Ultrascale+™ XCVU37P-2

Gagandeep Singh et al. "NERO: A Near High-Bandwidth Memory Stencil Accelerator for Weather Prediction Modeling." FPL, 2020