Funded by the Horizon 2020 Framework Programme of the European Union
MSCA-ITN-EID

NARMADA: Near-memory horizontal diffusion accelerator for scalable stencil computation

Gagandeep Singh, Dionysios Diamantopoulos, Sander Stuijk, Christoph Hagleitner, and Henk Corporaal
sin@zurich.ibm.com

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

IBM Research | Zurich

FPL, Barcelona
11th-September-2019

Stencil Computations and Applications

Stencils are used in ~30% of HPC applications:

- Fluid dynamics, image processing, atmospheric modelling

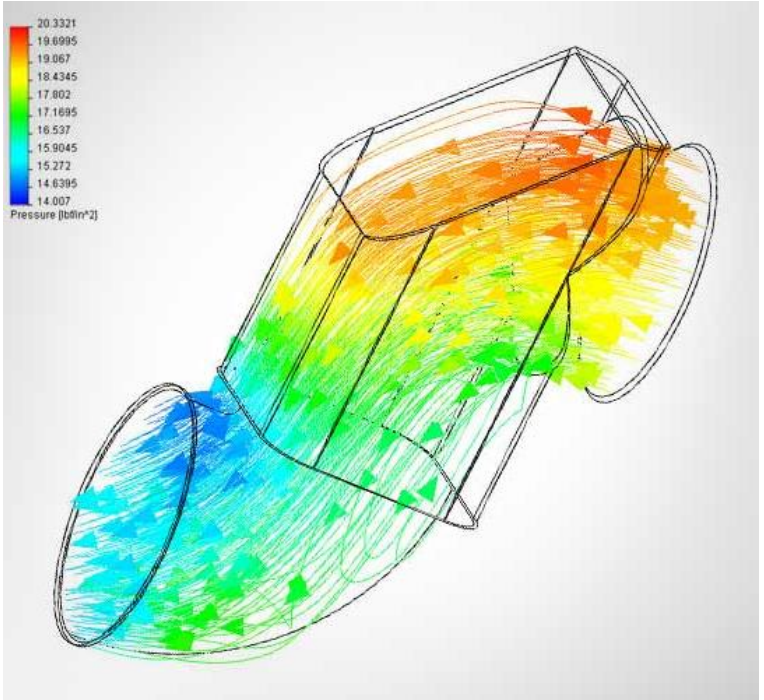


Image sources: http://www.flometrics.com/fluid-dynamics/computational-fluid-dynamics/
Naoe, Kensuke et al. "Secure Key Generation for Static Visual Watermarking by Machine Learning in Intelligent Systems and Services." *IJSSOE,* 2010

Workload Characteristics

High-order stencil computations are cache unfriendly

Workload Characteristics

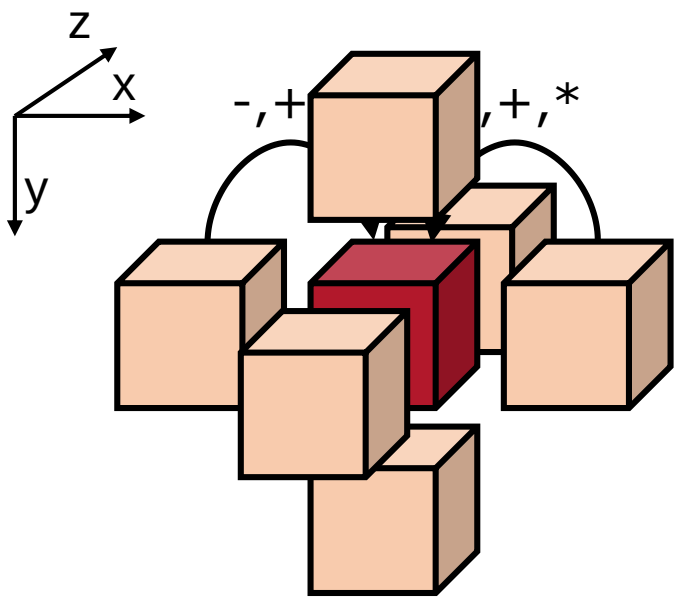
High-order stencil computations are cache unfriendly

- Limited arithmetic intensity: only reuse potential in neighboring pixels
- Sparse and complex access pattern

Workload Characteristics

High-order stencil computations are cache unfriendly

- Limited arithmetic intensity: only reuse potential in neighboring pixels
- Sparse and complex access pattern



7-point Jacobi in 3D plane

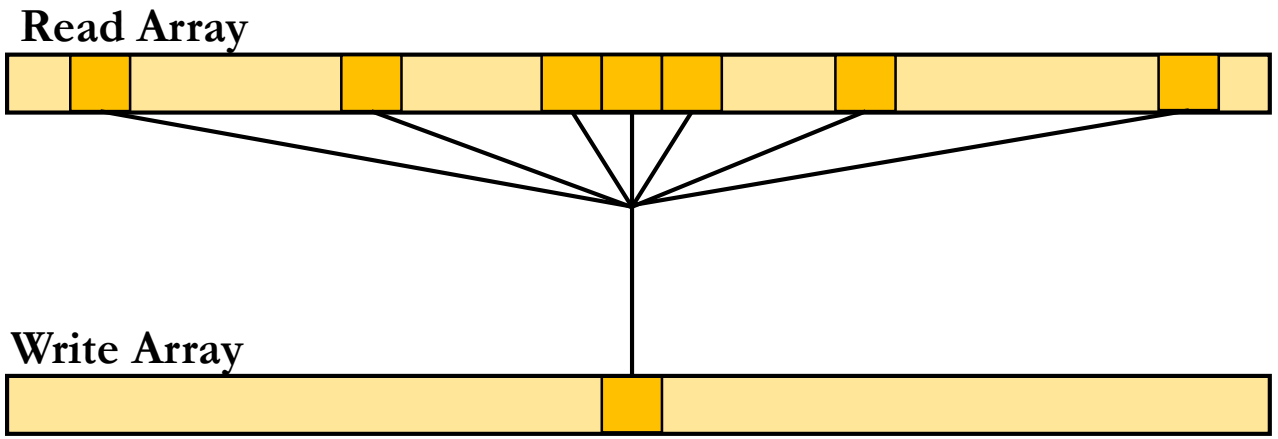


Image source: Performance Tuning and Analysis for Stencil-Based Applications on POWER8 Processor – Xu et al. (2018)

Workload Characteristics

High-order stencil computations are cache unfriendly

- Limited arithmetic intensity: only reuse potential in neighboring pixels

Data movement bottleneck

7-point Jacobi in 3D plane

Image source: Performance Tuning and Analysis for Stencil-Based Applications on POWER8 Processor – Xu et al. (2018)

Stencil use in COSMO

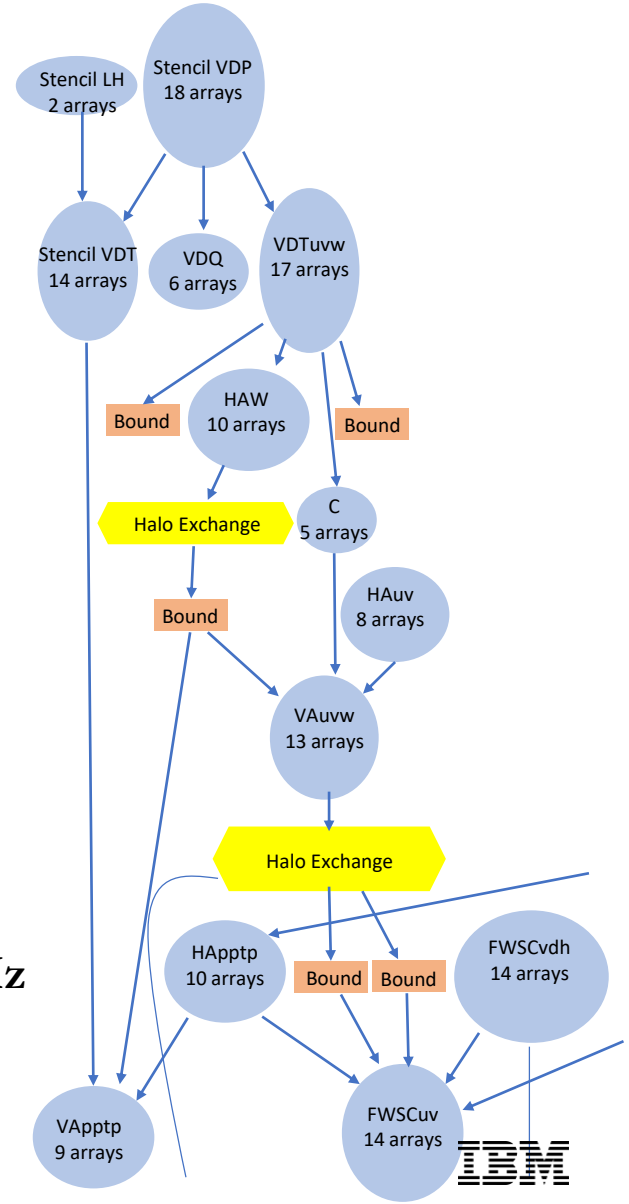
Stencil computing in weather/climate applications

- Dynamical core is the most essential part of the weather models

Stencil use in COSMO

Stencil computing in weather/climate applications

- Dynamical core is the most essential part of the weather models
- *O* (100) different stencil compute motifs
- ~30 variable- and ~70 temporary arrays (3D grids)

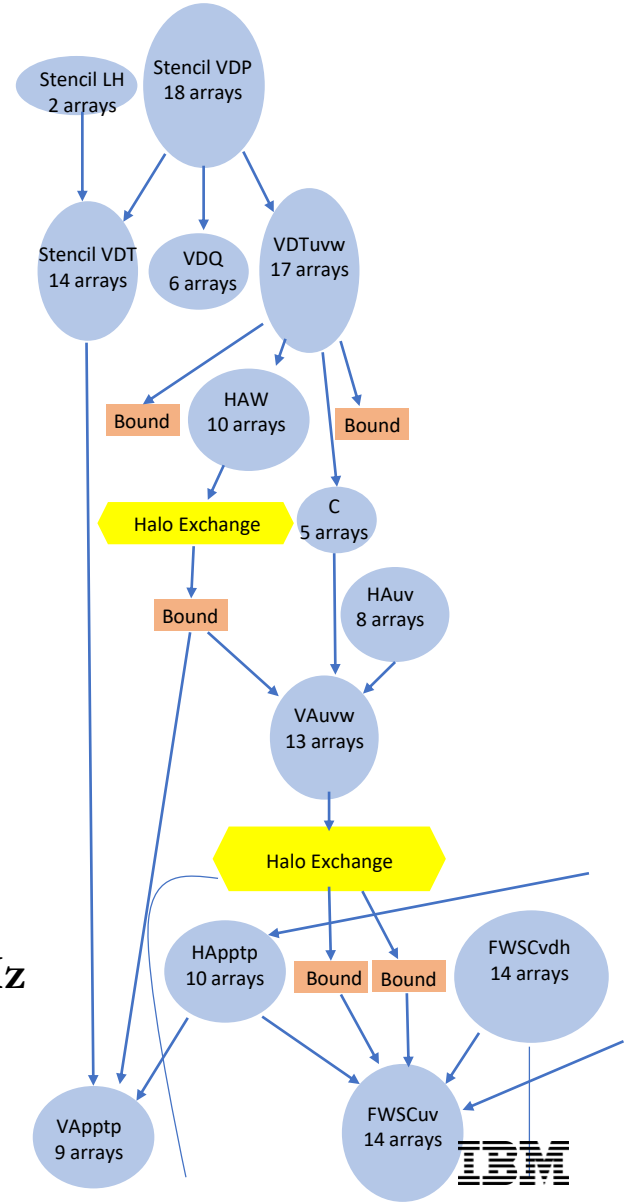


Section of
COSMO CDAG
(Courtesy CSCS/ETHz
and Ronald Luijten)

Stencil use in COSMO

Stencil computing in weather/climate applications

- Dynamical core is the most essential part of the weather models
- *O* (100) different stencil compute motifs
- ~30 variable- and ~70 temporary arrays (3D grids)
- **Complex stencil programs**



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

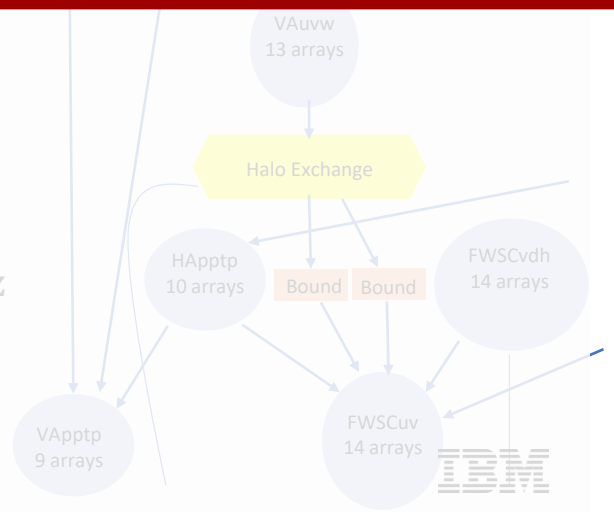
Stencil use in COSMO

Stencil computing in weather/climate applications

- Dynamical core is the most essential part of the weather models

Not another "elementary" stencil talk!

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

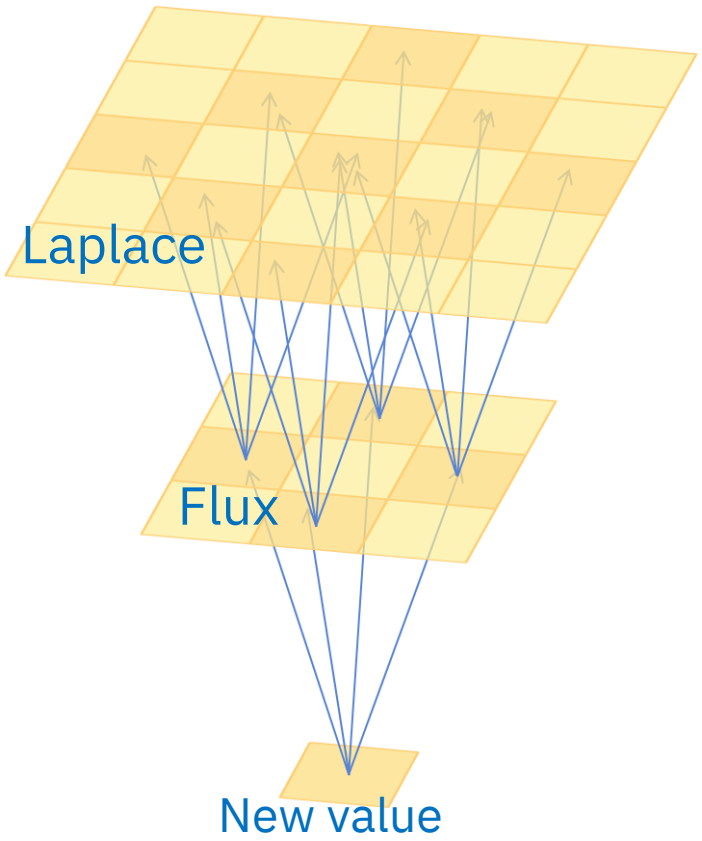


Horizontal Diffusion (“complex” stencil)

- Compound stencil kernel, consists of a collection of elementary stencil kernels

Horizontal Diffusion ("complex" stencil)

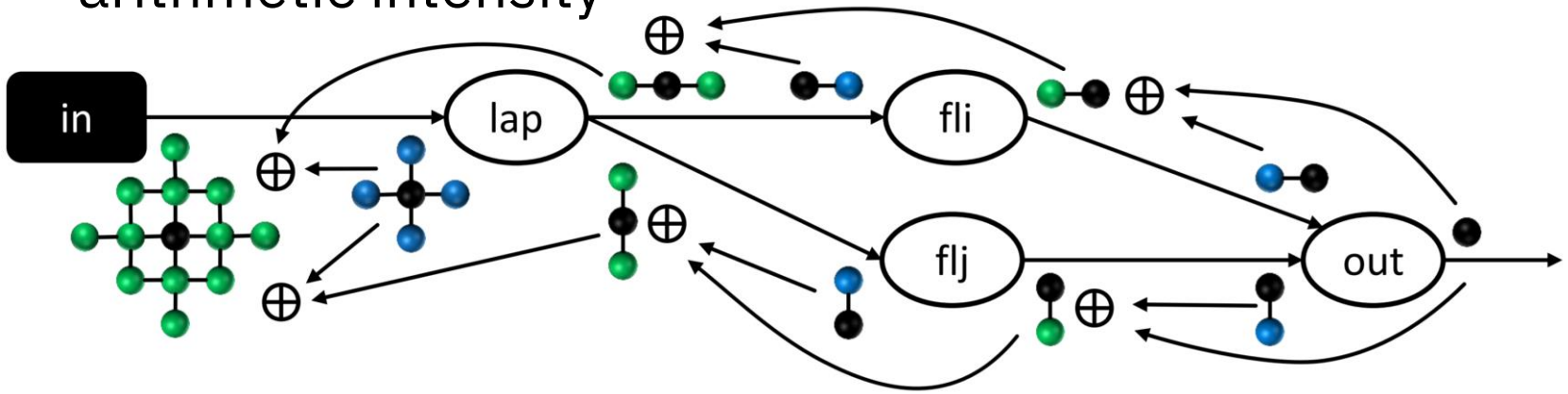
- Compound stencil kernel, consists of a collection of elementary stencil kernels
- Iterates over 3D grid that perform laplacian and flux operations
- Complex memory access behavior and low arithmetic intensity

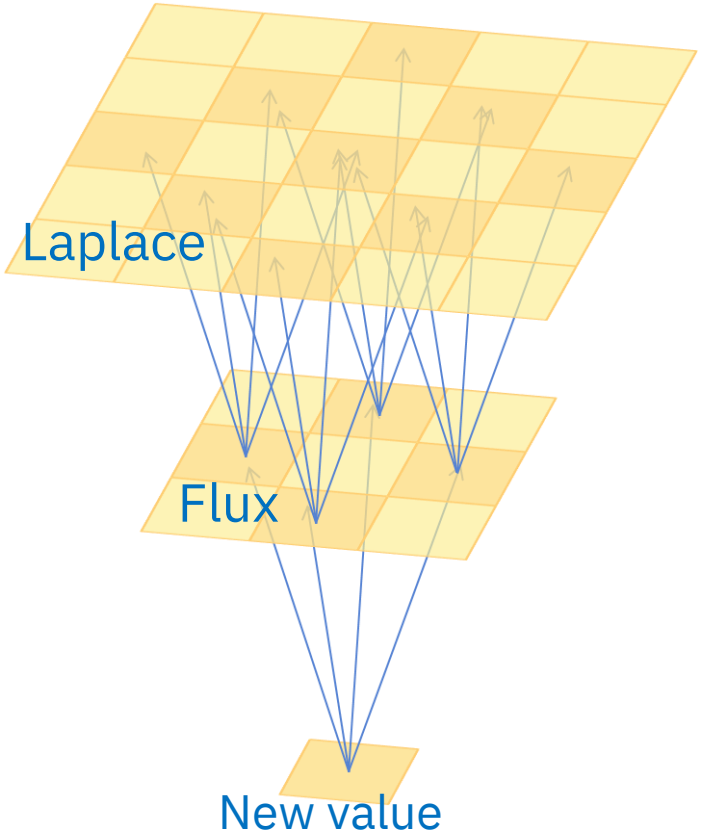


Stencil composition
(Courtesy CSCS/ETHz and Ronald Luijten)

Horizontal Diffusion ("complex" stencil)

- Compound stencil kernel, consists of a collection of elementary stencil kernels
- Iterates over 3D grid that perform laplacian and flux operations
- Complex memory access behavior and low arithmetic intensity

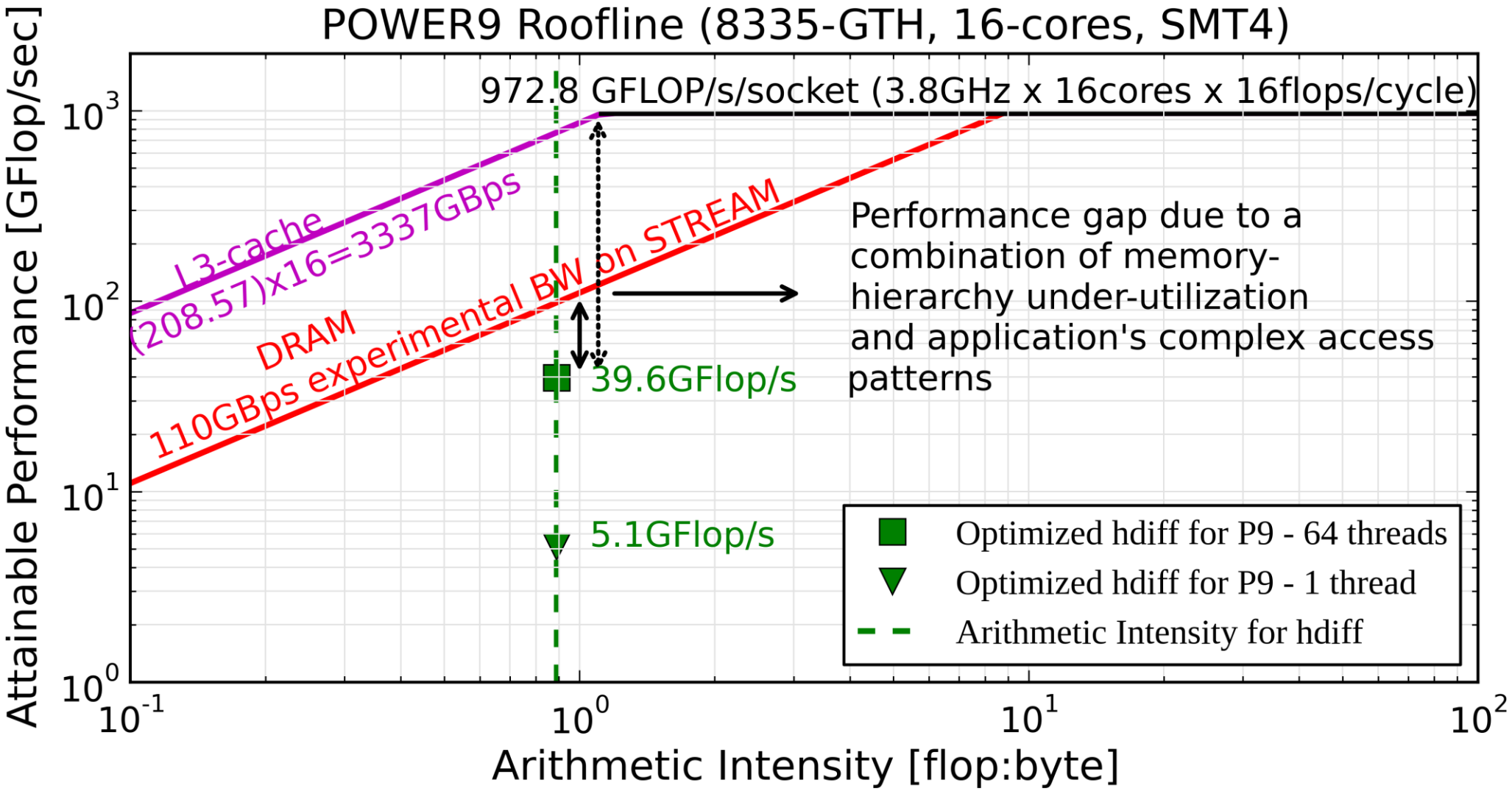




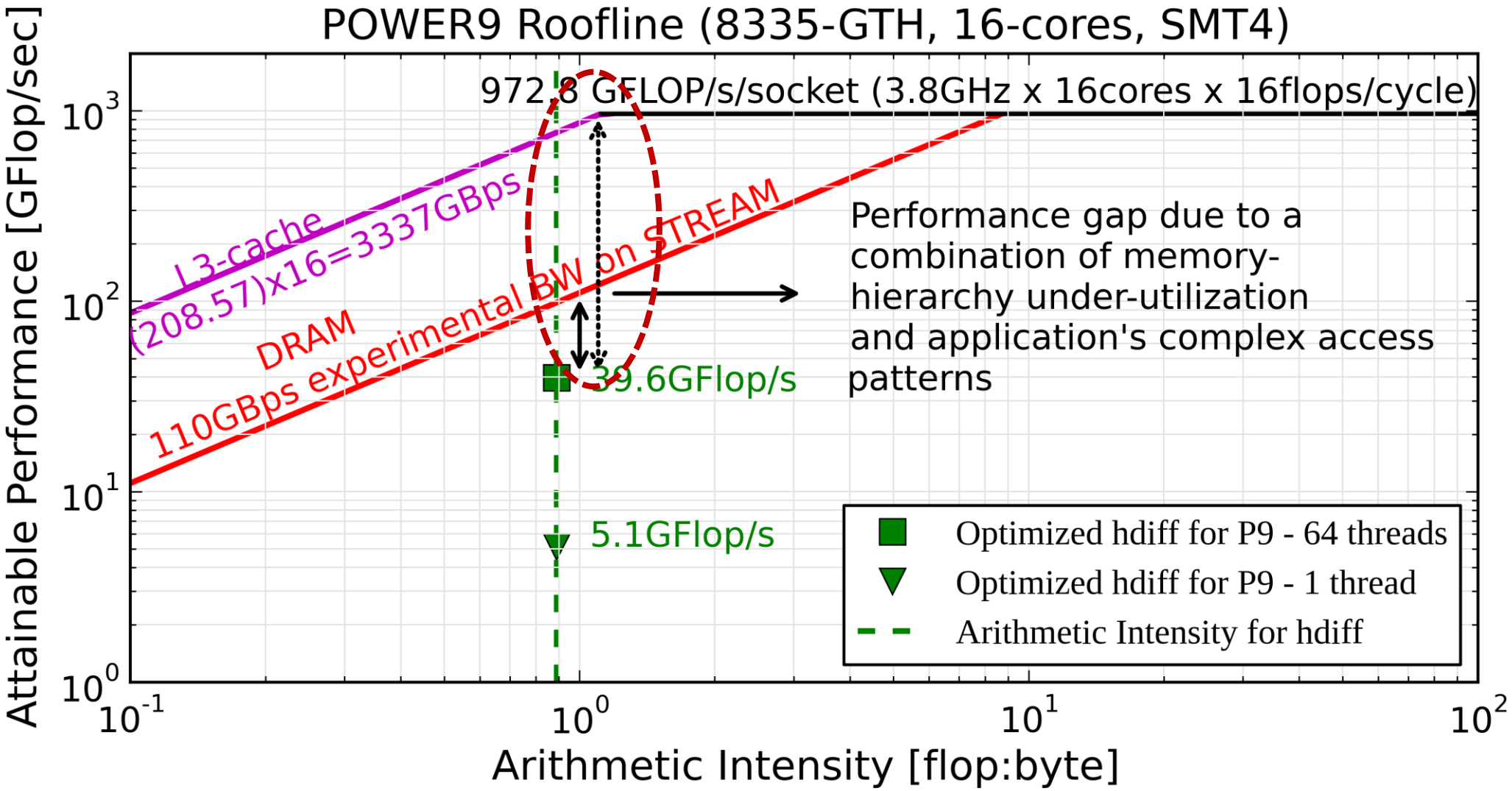
Stencil composition
(Courtesy CSCS/ETHz and Ronald Luijten)

T. Gysi, T. Grosser, T. Hoefler: MODESTO: Data-centric Analytic Optimization of Complex Stencil Programs on Heterogeneous Architectures, ICS'15
DFG from T. Hoefler

IBM POWER9 CPU Roofline

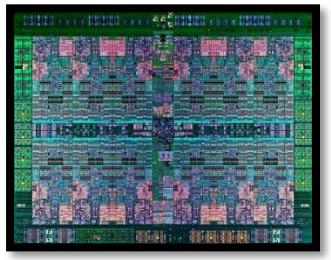


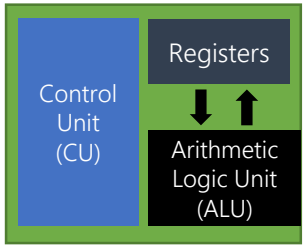
IBM POWER9 CPU Roofline



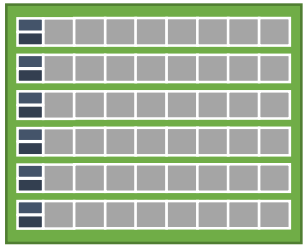
Alternative Platforms

CPUs

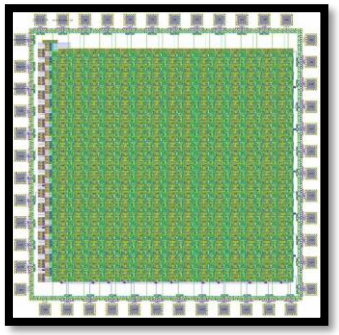


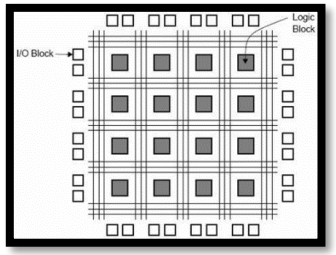


GPUs

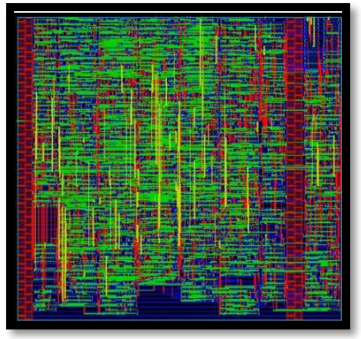


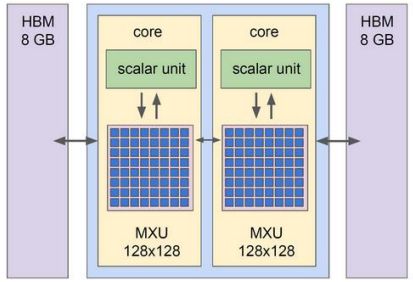
FPGAs





ASICs





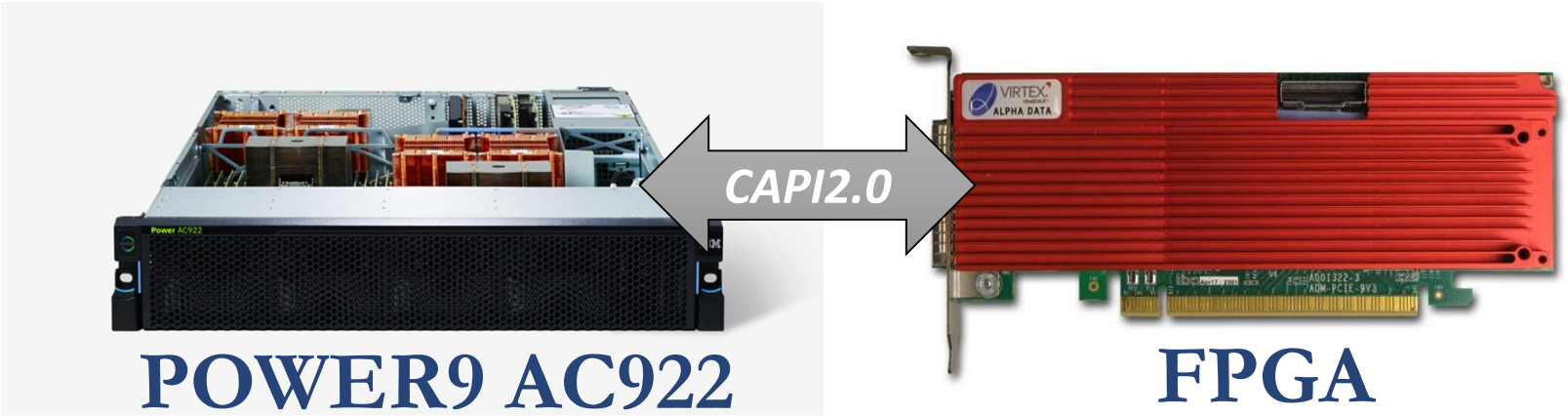
Processing Units ASICs for emerging workloads, e.g. Google TPU

FLEXIBILITY ← → **EFFICIENCY**

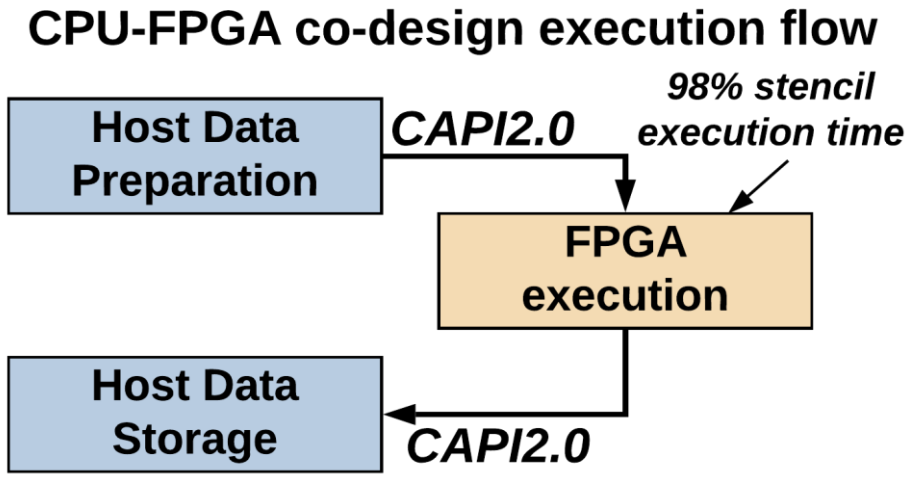
FPGAs ideal for adapting to rapidly evolving workloads!

Heterogeneous Architecture: CPU+FPGA

- Host System
  - IBM POWER9-16 core (64-threads)
  - Power: IBM AMESTER[1]

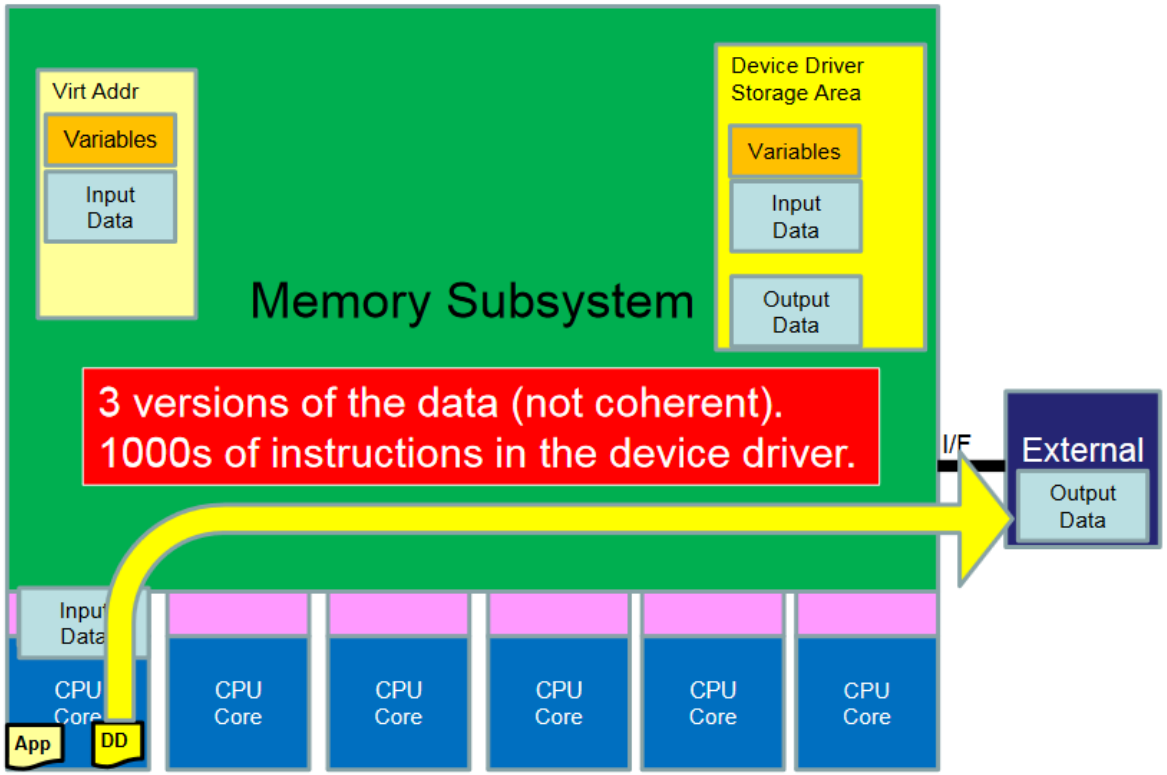


- FPGA board
  - Xilinx Virtex® Ultrascale+™ XCVU3P-2

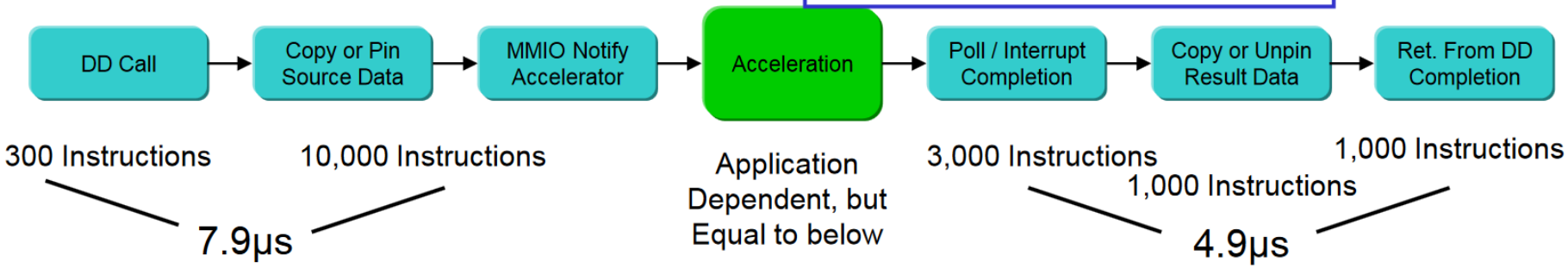


[1] https://github.com/open-power/amester

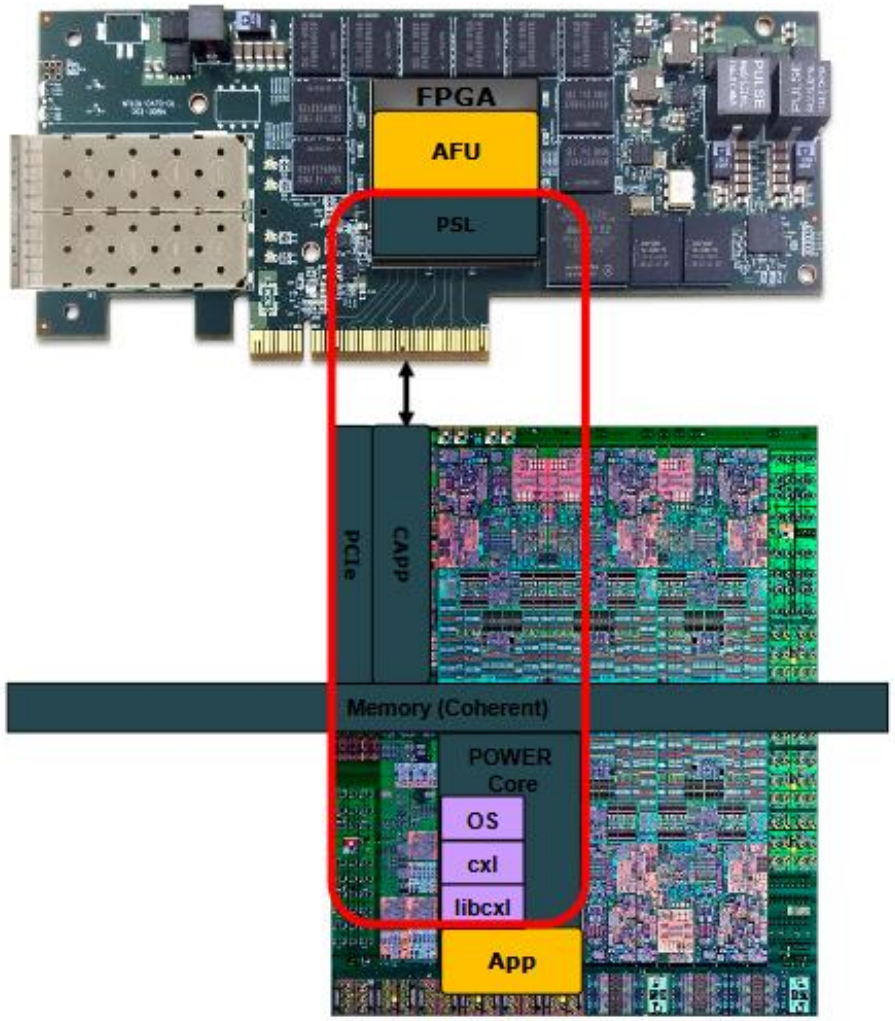
Traditional I/O Technology



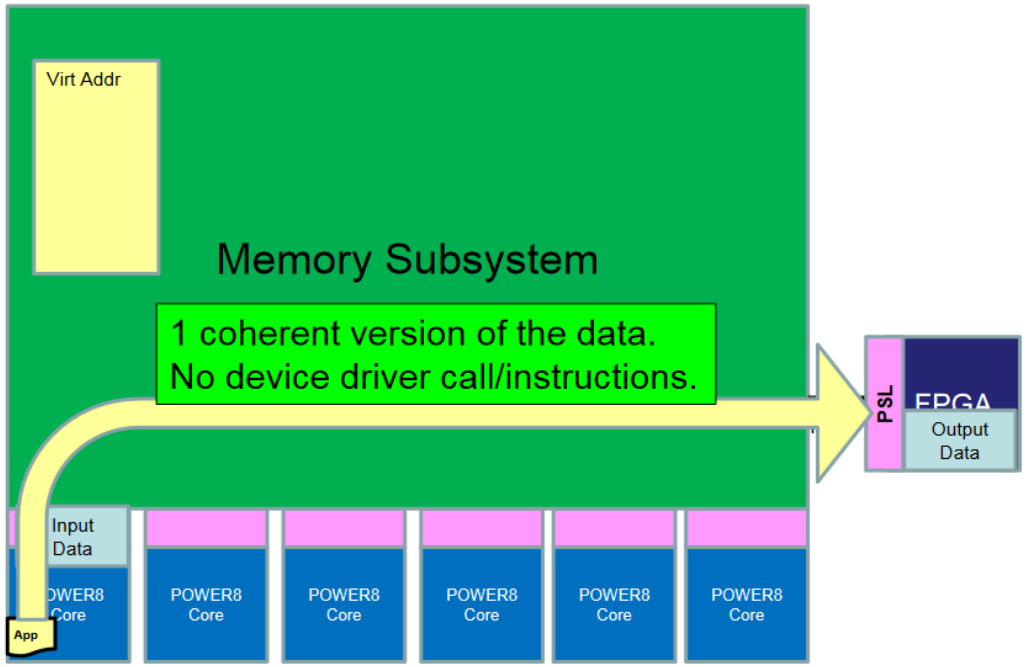
Typical I/O Model Flow: Total ~13µs for data prep



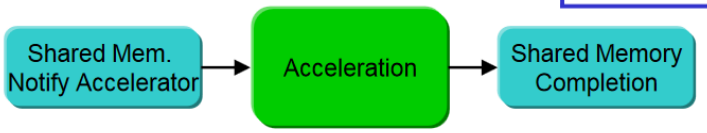
CAPI Technology Overview



POWER8 – POWER9 Processor



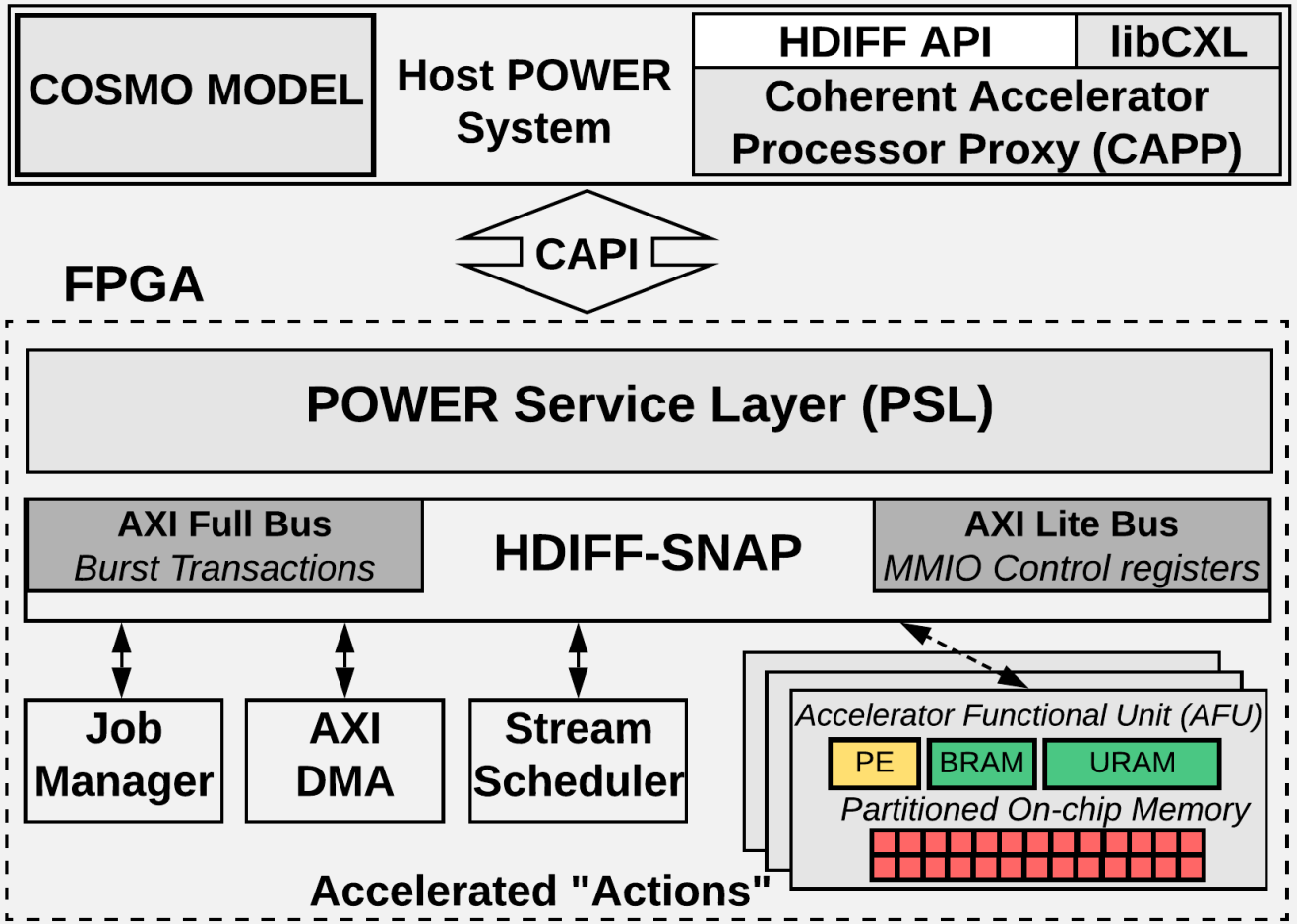
Flow with a CAPI Model: Total 0.36µs



400 Instructions	Application Dependent, but Equal to above	100 Instructions
0.3µs		0.06µs

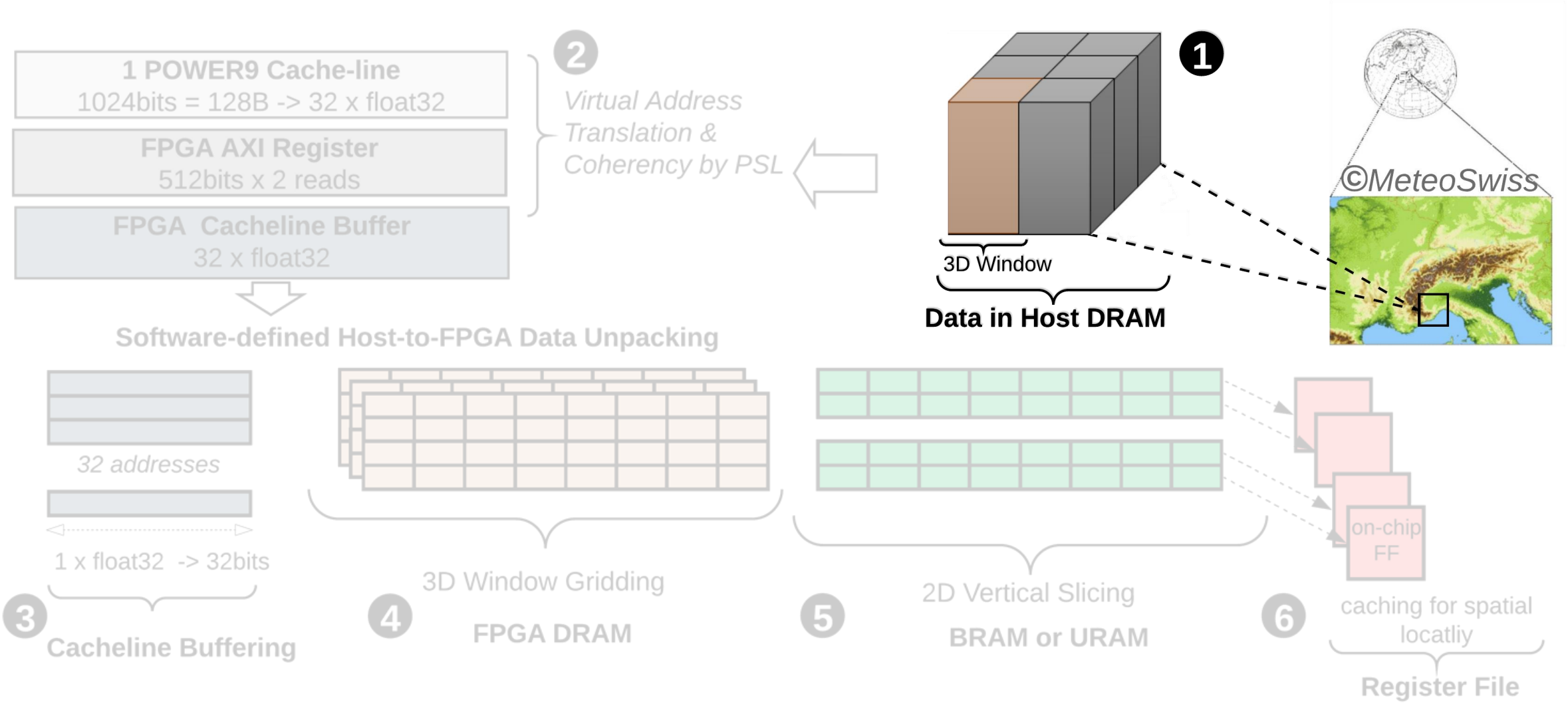
Accelerator Framework

- Accelerators are acting as peers to CPU
- High-performance cache-coherent link
- An interrupt-based queuing mechanism
- Minimal CPU usage (thus power) during FPGA use

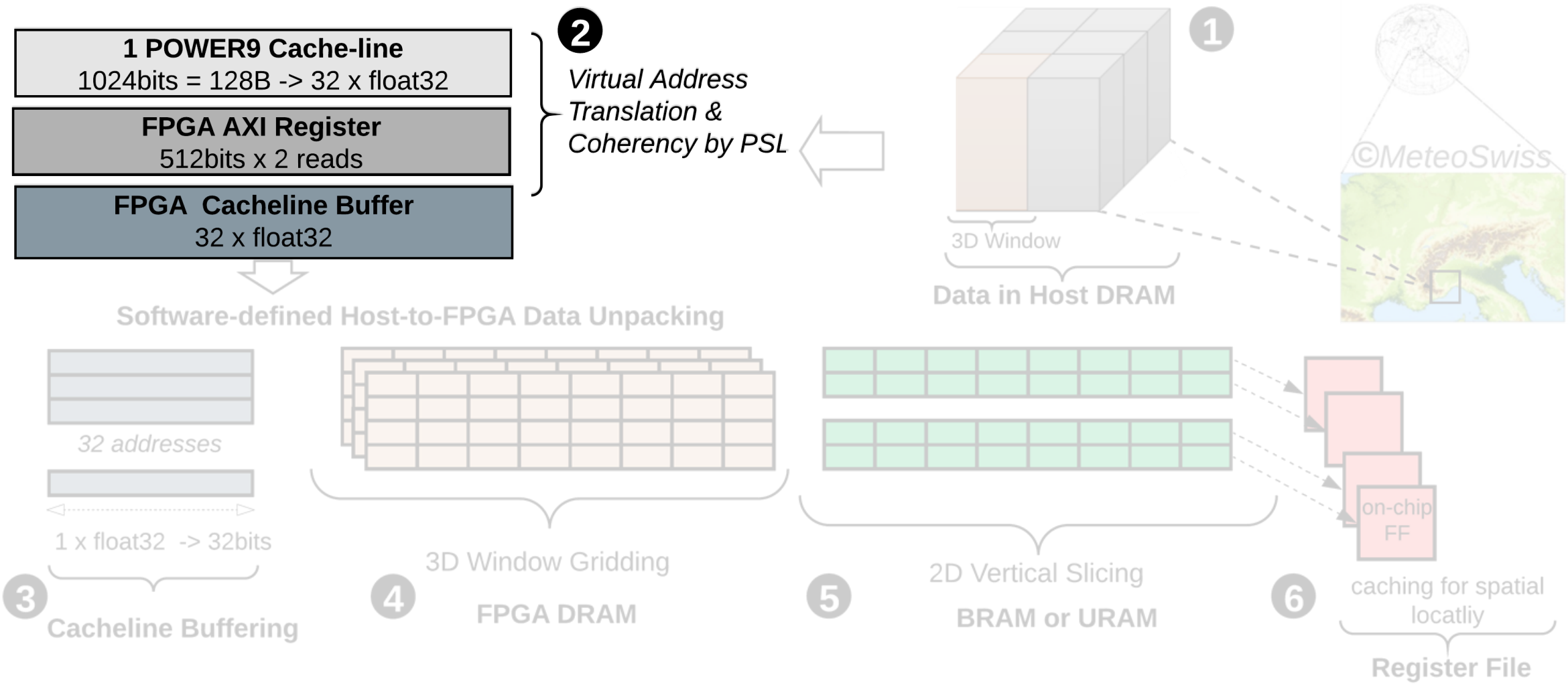


https://github.com/open-power/snap

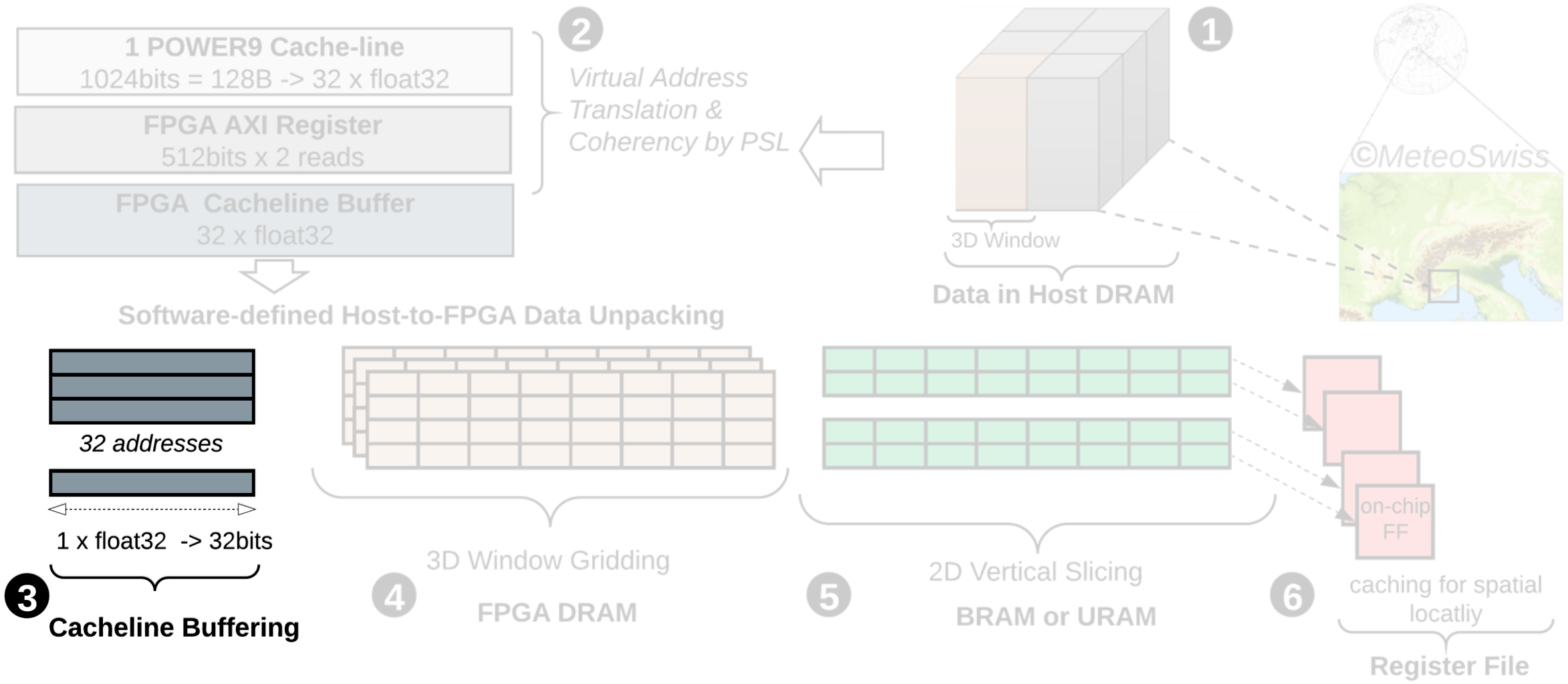
Dataflow Sequence



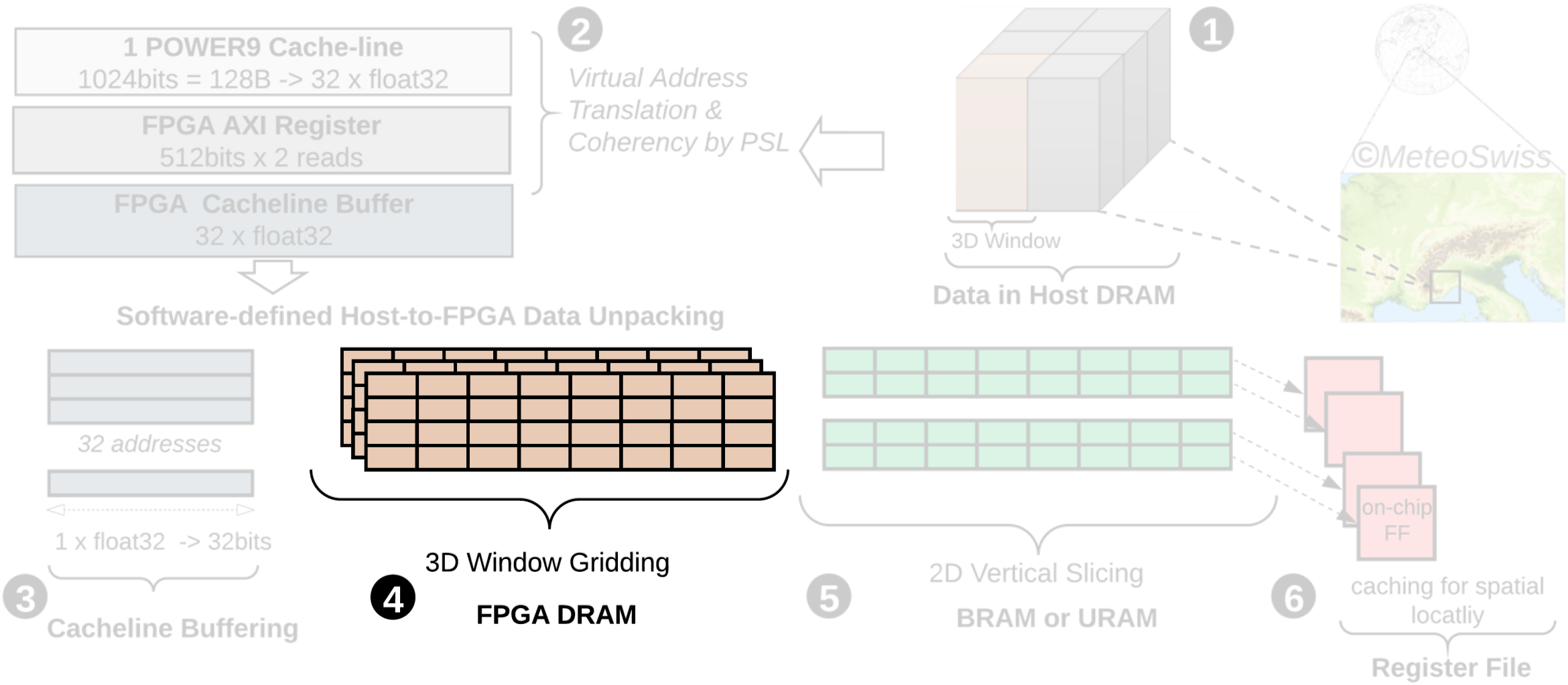
Dataflow Sequence



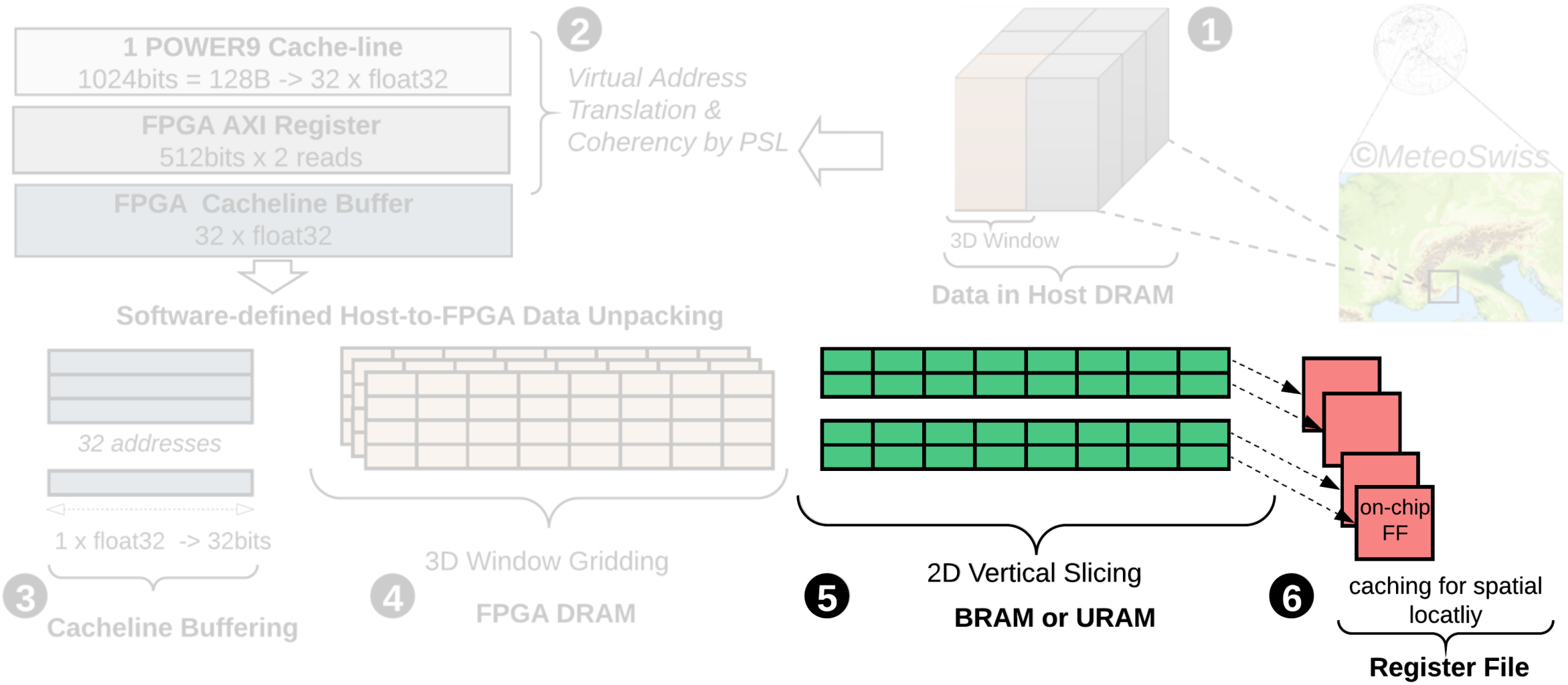
Dataflow Sequence



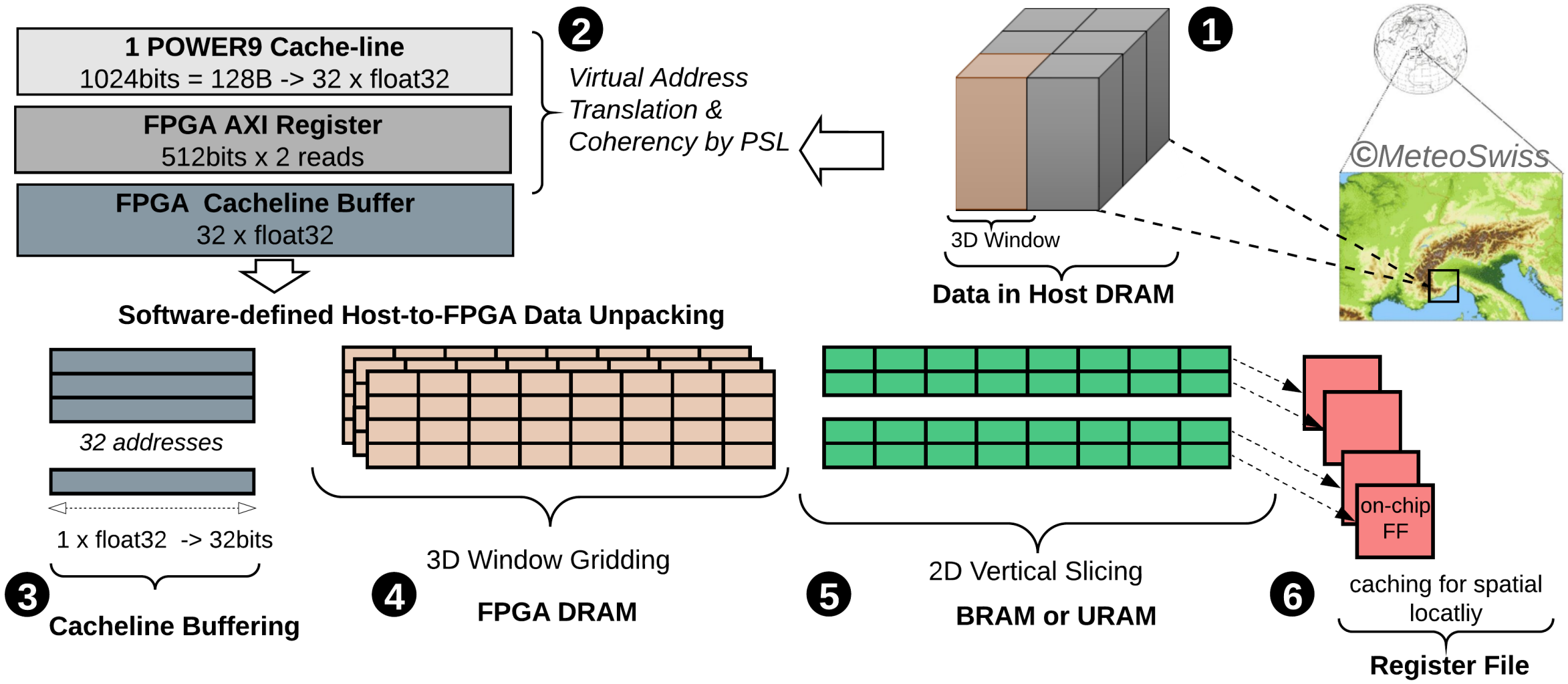
Dataflow Sequence



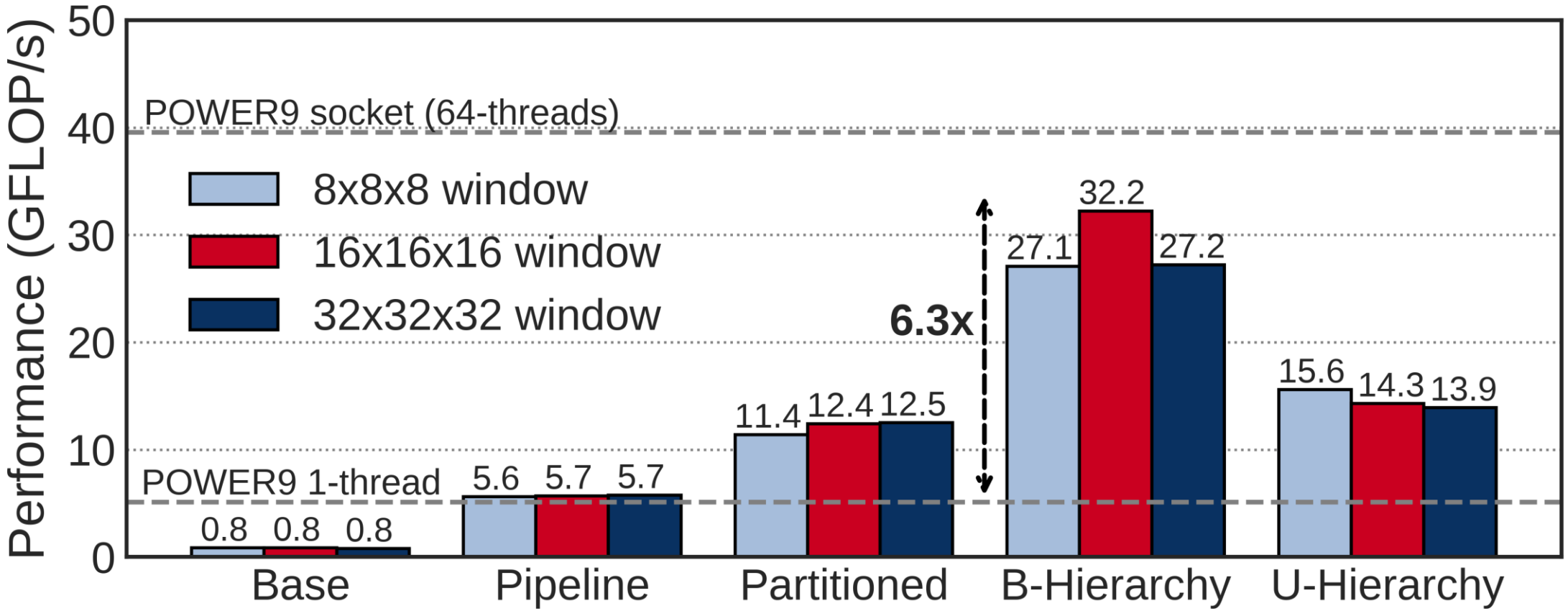
Dataflow Sequence



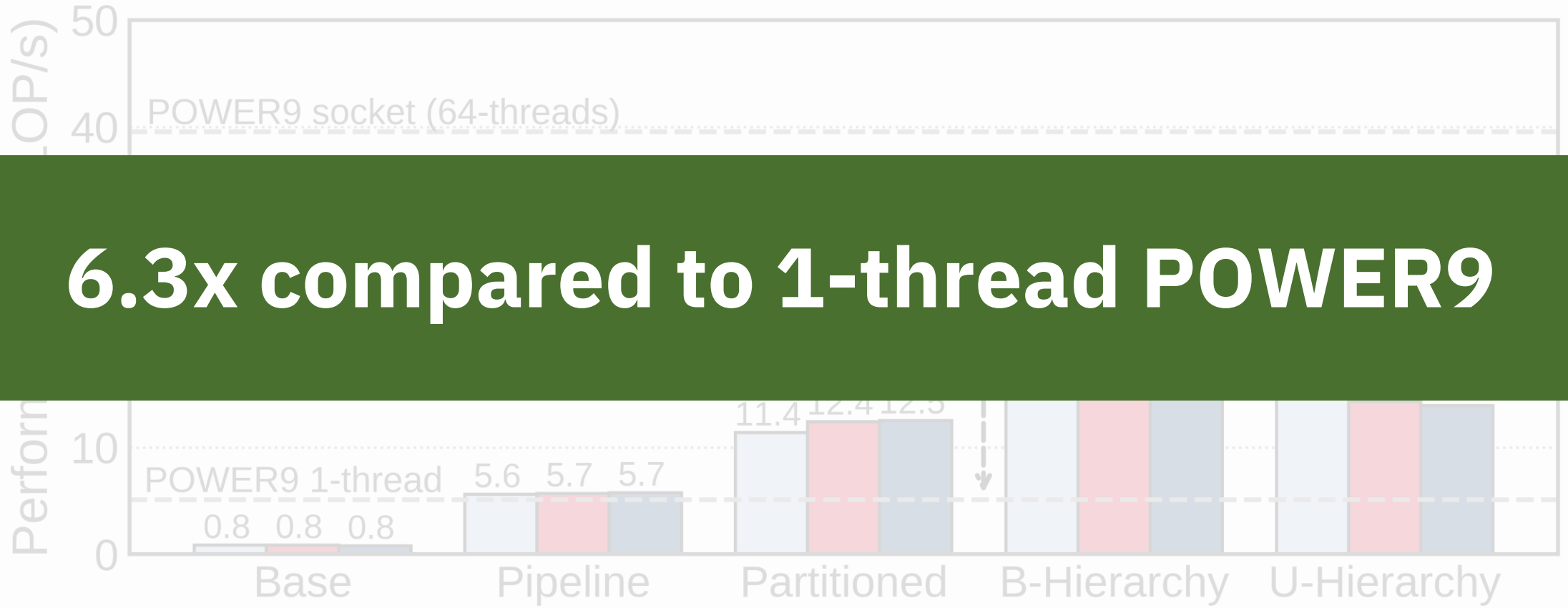
Dataflow Sequence



AFU Performance

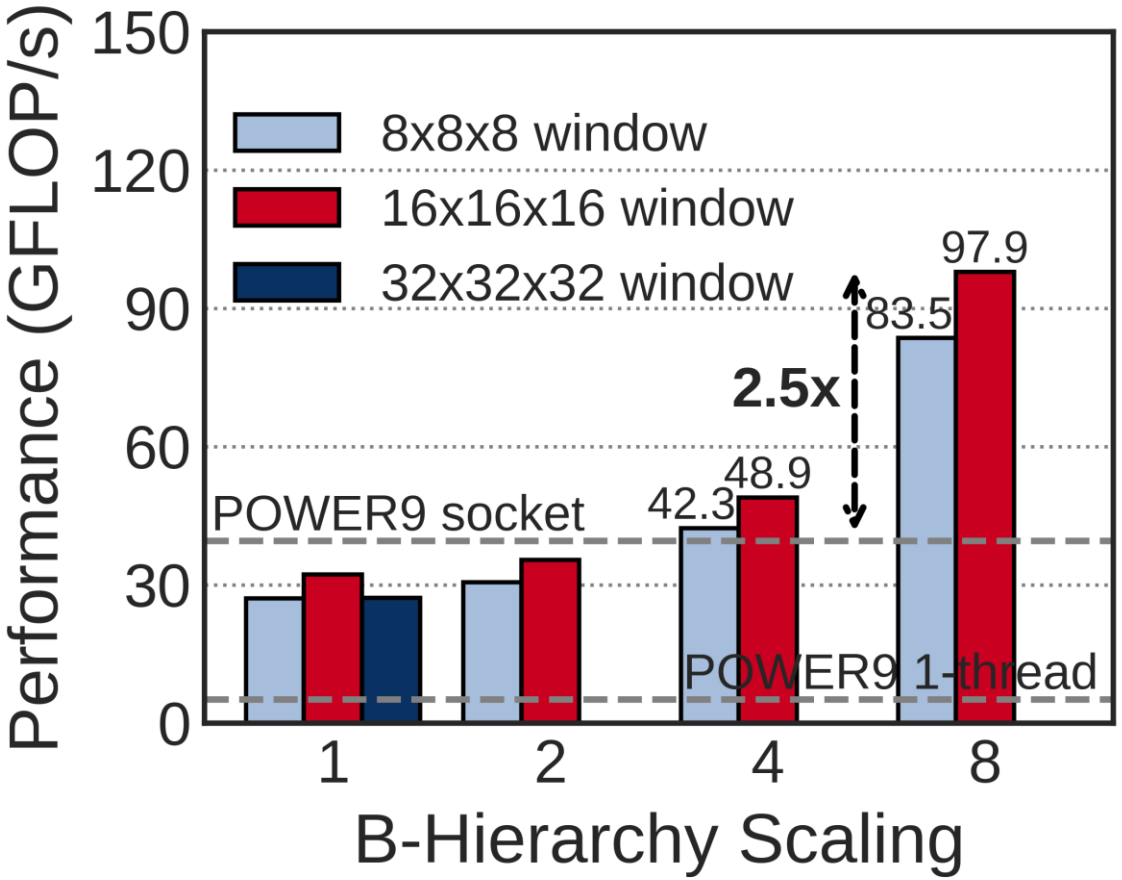


AFU Performance

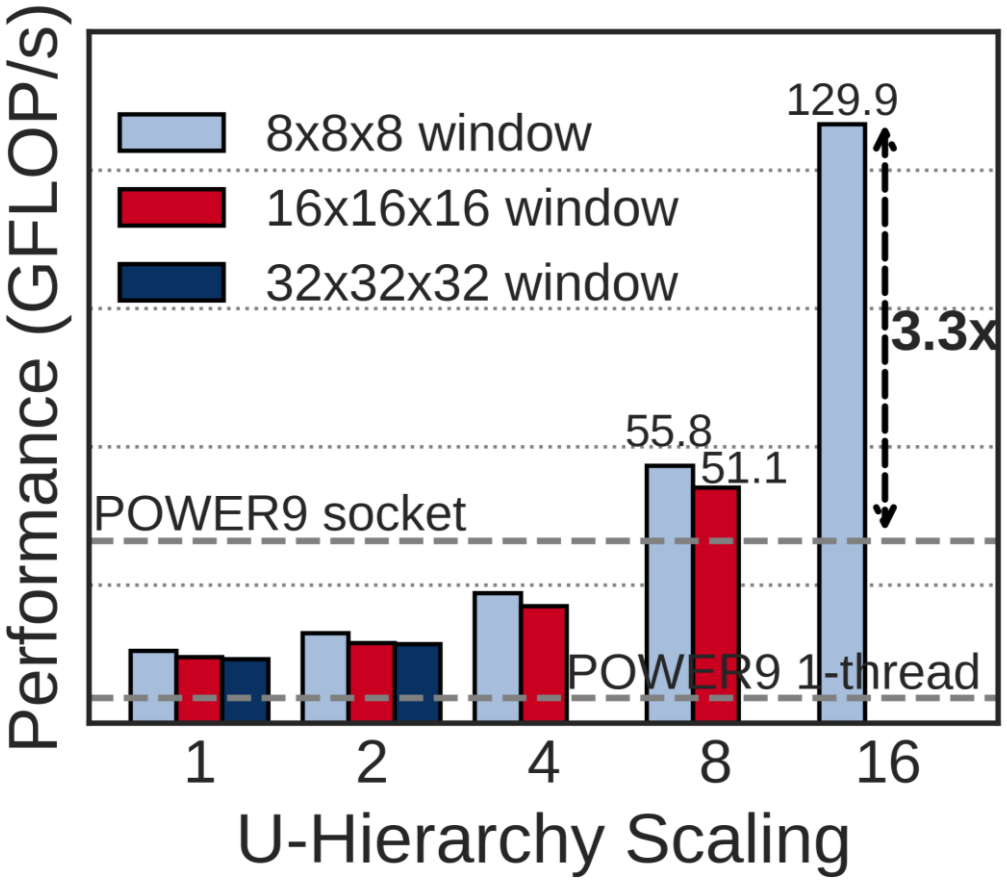


6.3x compared to 1-thread POWER9

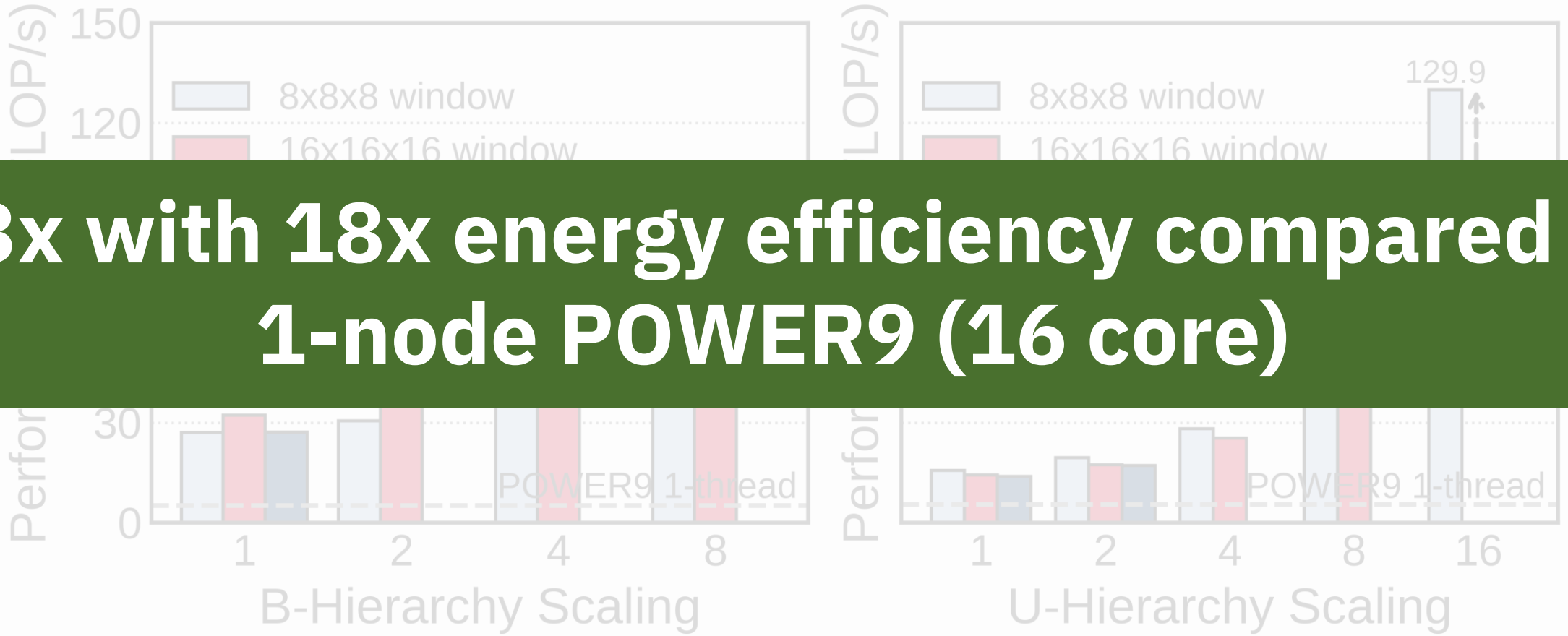
AFU Scaling Analysis



AFU Scaling Analysis



AFU Scaling Analysis



3.3x with 18x energy efficiency compared to 1-node POWER9 (16 core)

Accelerator Scaling Prediction

- **Goal:** Quick prediction for scaling AFUs on different FPGA boards
- Long run-time
- Back-of-the-envelope calculations cannot accurately predict complex design behavior
- Heuristics
- **Approach**: Empirical best-fit model using data collected from one device and predict for all the devices in an FPGA family

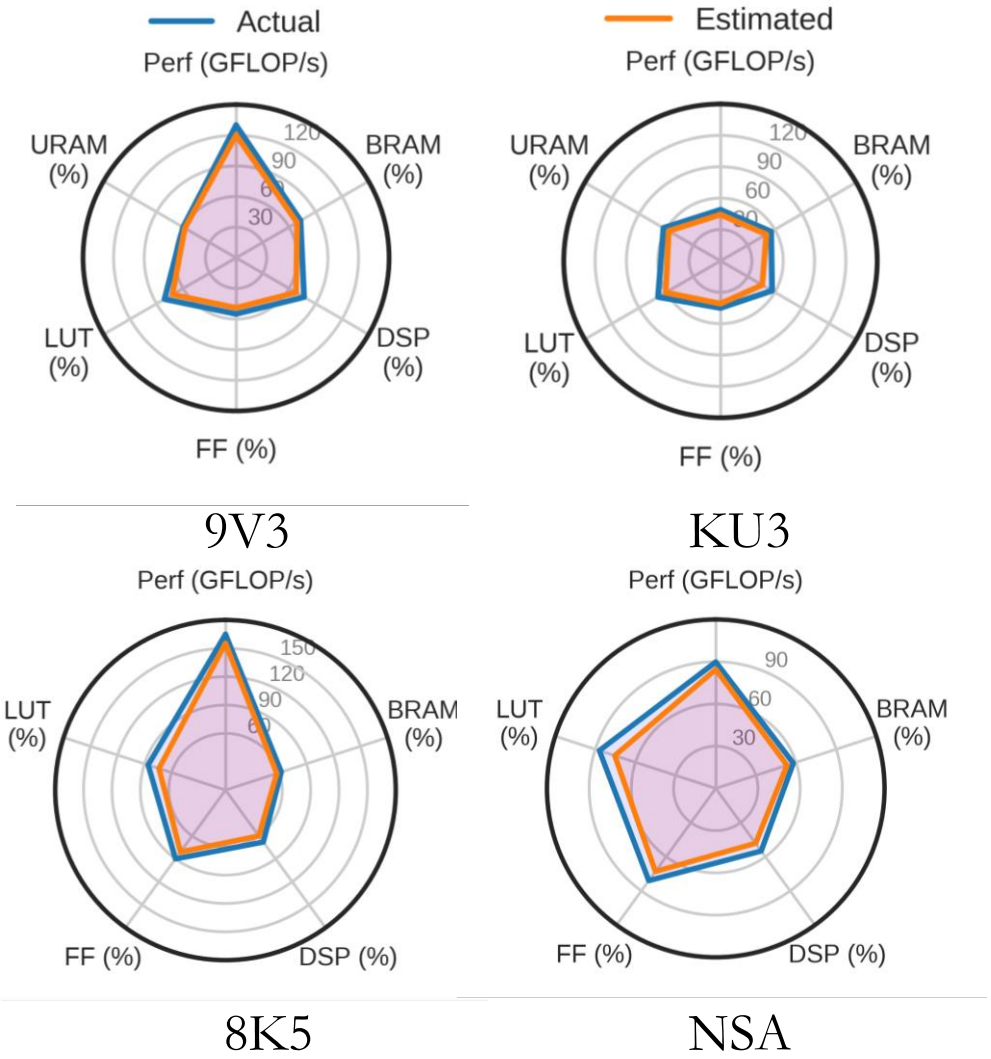
Accelerator Scaling Prediction

- Xilinx Ultrascale and Ultrascale+ families
  - CAPI enabled
  - With and without URAM

Accelerator Scaling Prediction

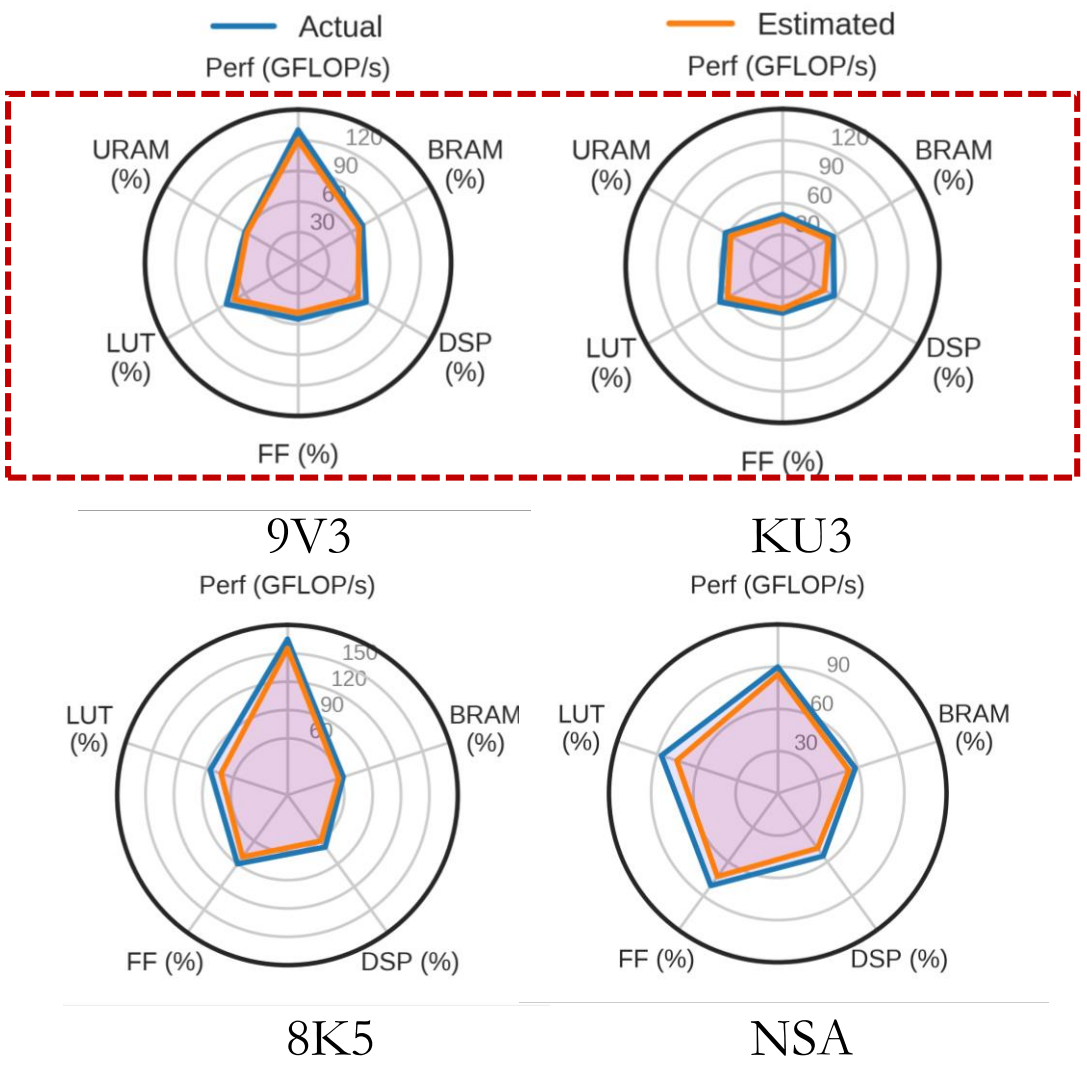
- Xilinx Ultrascale and Ultrascale+ families
  - CAPI enabled
  - With and without URAM

- MRE under 15.2%



Accelerator Scaling Prediction

- Xilinx Ultrascale and Ultrascale+ families
  - CAPI enabled
  - With and without URAM
- MRE under 15.2%
- URAM offers more heterogeneity and energy-efficiency with scaling



Future work

- Intra-node multi-FPGA scaling
- Implementation with OpenCAPI and HBM
- General purpose COSMO accelerator
- Trans-precision analysis
- Run-time adaptability for other stencils

Executive Summary

- **Motivation:** Stencil computation is essential part of HPC applications
- **Problem:** Limited performance on conventional architectures
- **Goal:** Study the applicability of compound stencils from real-world weather prediction application on reconfigurable architectures
- **Our contribution: NARMADA**
  - First implementation and optimization of horizontal diffusion kernel from COSMO application on modern heterogeneous system
  - A data-centric heterogeneous memory hierarchy caching scheme with scalability analysis
- **Results**
  - NARMADA has 6.3x performance compared to a 1-thread performance of the state-of-the-art IBM POWER9 CPU
  - 3.3x performance with 18x energy-efficiency compared to a complete IBM POWER9 node

IBM

Funded by the Horizon 2020 Framework Programme of the European Union
MSCA-ITN-EID

NARMADA: Near-memory horizontal diffusion accelerator for scalable stencil computation

Gagandeep Singh, Dionysios Diamantopoulos, Sander Stuijk, Christoph Hagleitner, and Henk Corporaal
sin@zurich.ibm.com

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

IBM Research | Zurich

FPL, Barcelona
11th-September-2019