



SPARTA

Spatial Acceleration for Efficient and Scalable Horizontal Diffusion Weather Stencil Computation

Gagandeep Singh, Alireza Khodamoradi, Kristof Denolf, Jack Lo, Juan Gómez-Luna, Joseph Melber, Andra Bisca, Henk Corporaal, and Onur Mutlu

37th International Conference on Supercomputing (ICS)
Orlando, Florida

SAFARI
SAFARI Research Group
safari.ethz.ch

ETH zürich

TU/e

EINDHOVEN
UNIVERSITY OF
TECHNOLOGY

AMD 
together we advance_

Talk Outline

Background and Motivation

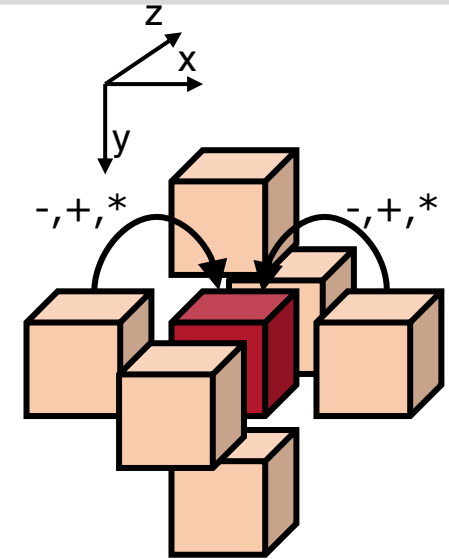
SPARTA: Design and Implementation

Evaluation of SPARTA and Key Results

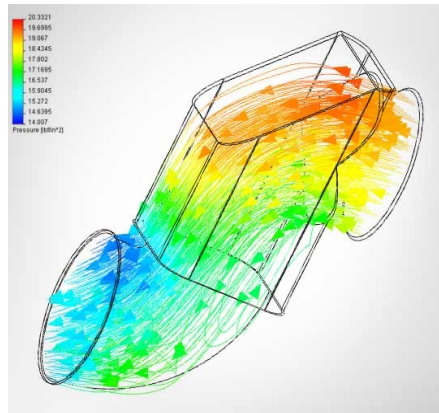
Summary

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



Fluid Dynamics



Image Processing

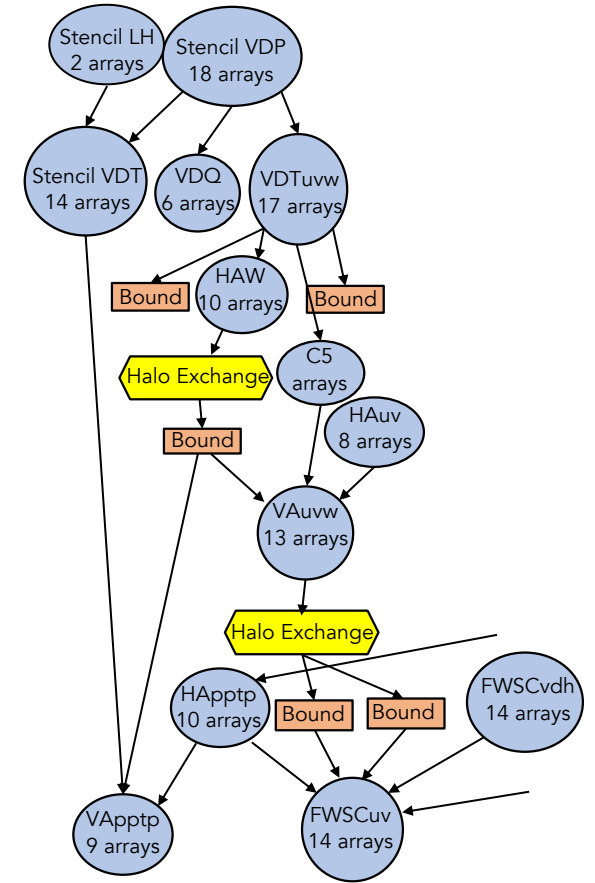


Climate/Weather Simulations

Stencil Computations in Weather Applications

COSMO (Consortium for Small-Scale Modeling) weather prediction application [Thaler+, PASC'19]

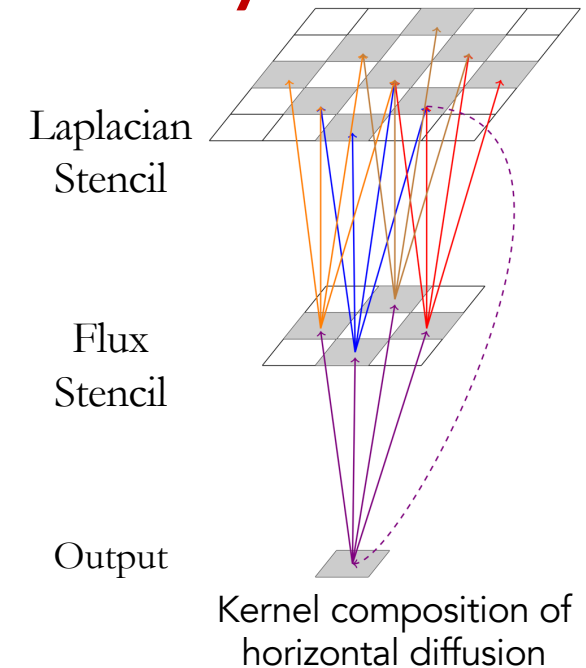
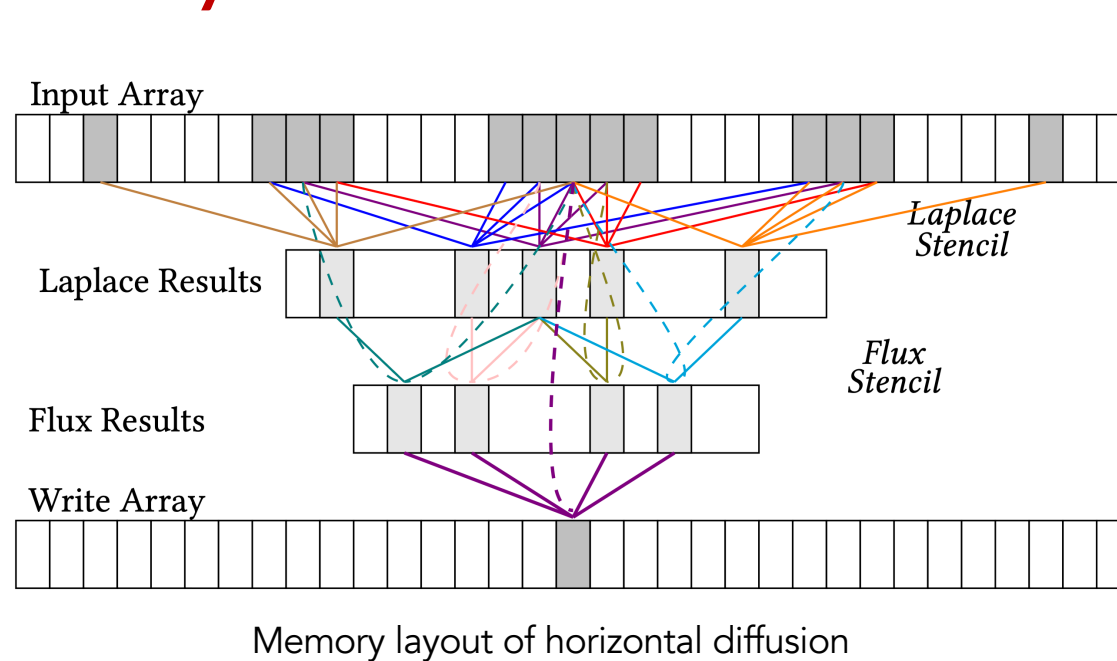
- The essential part of the weather prediction models is called the **dynamical core (dycore)**
- Around **80 different** stencil compute motifs
- ~30 variables and ~70 temporary arrays
- **Complex stencil computation**



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

Fundamental Complex Stencil: Horizontal Diffusion

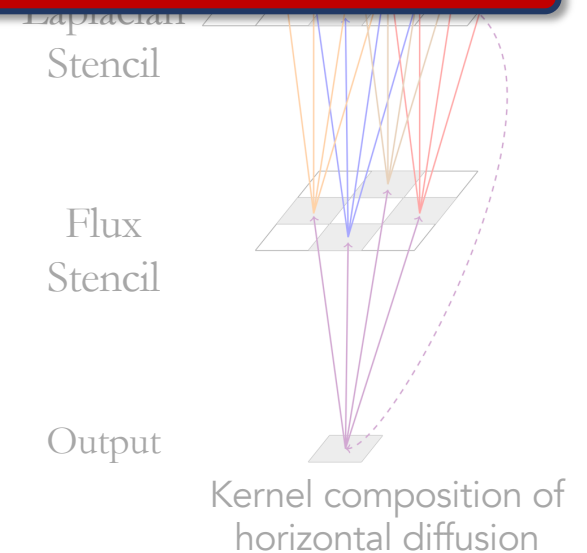
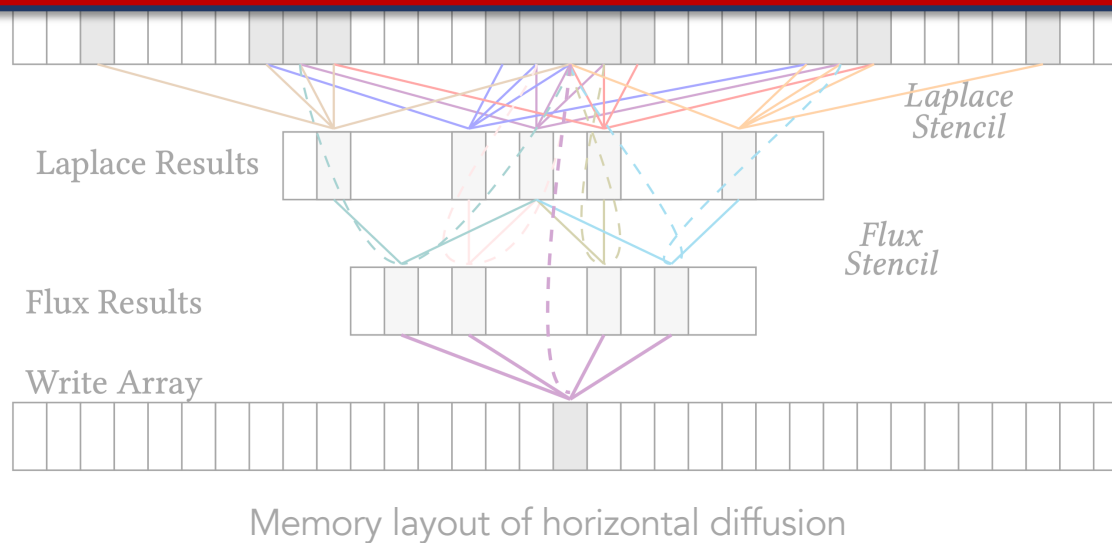
- Compound stencil kernel consists of a **collection** of elementary stencil kernels
- Iterates over a 3D grid performing **Laplacian** and **flux** operations
- **Complex memory access behavior** with **low arithmetic intensity**



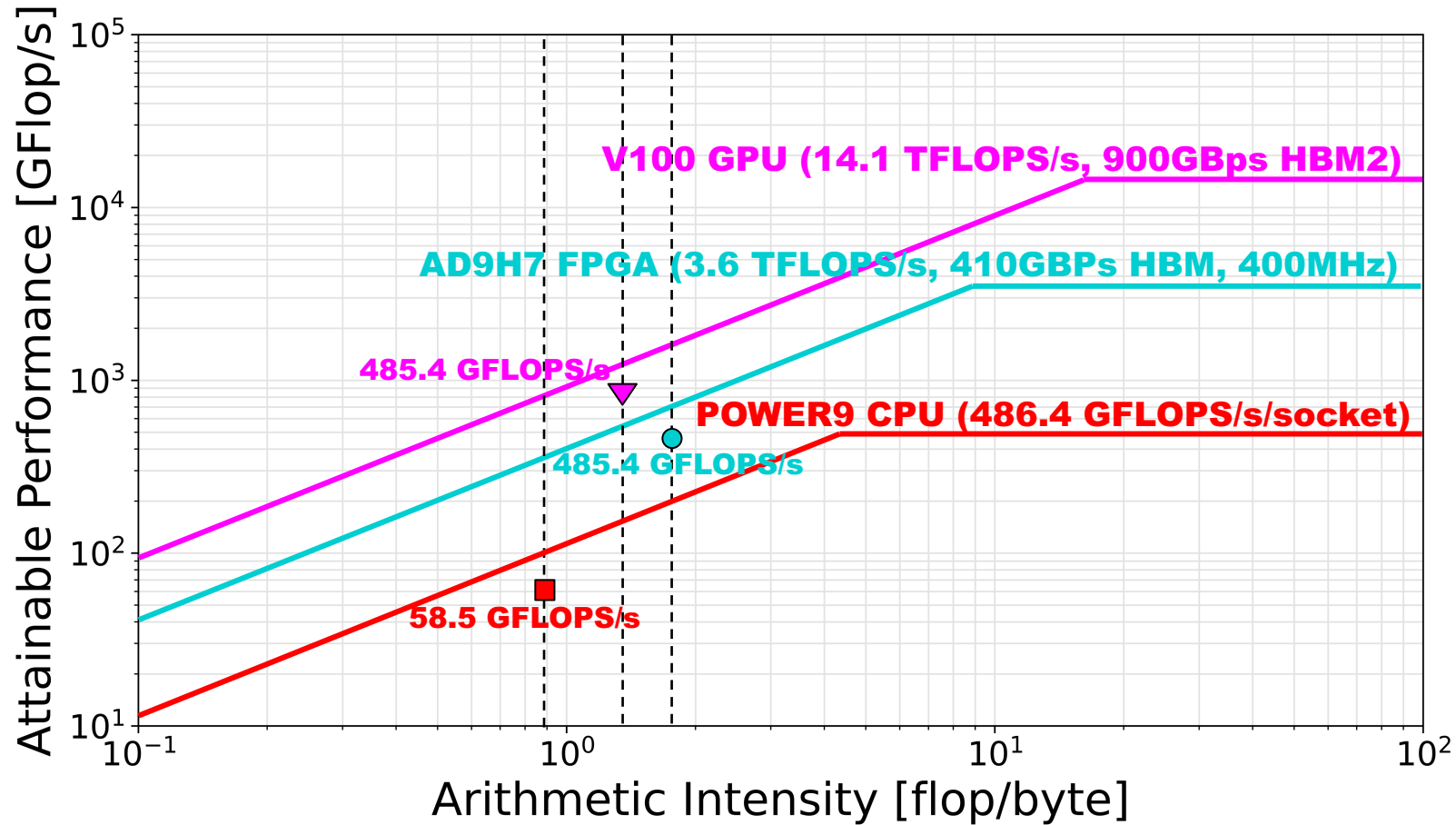
Fundamental Complex Stencil: Horizontal Diffusion

- Compound stencil kernel consists of a **collection** of elementary stencil kernels
- Iterates over a 3D grid performing **Laplacian** and **flux** operations

Performance bottleneck



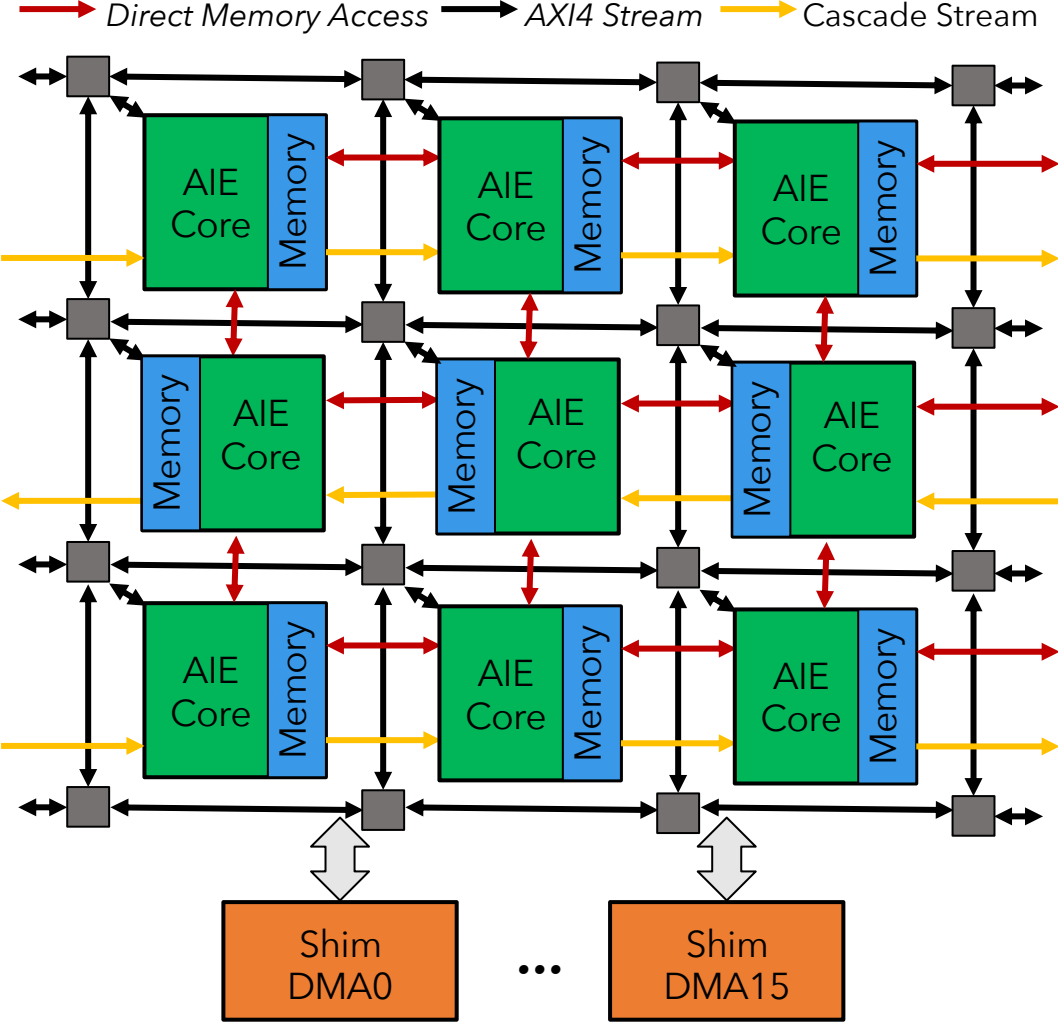
Roofline Analysis



<13.5% peak floating-point performance
on current computing systems

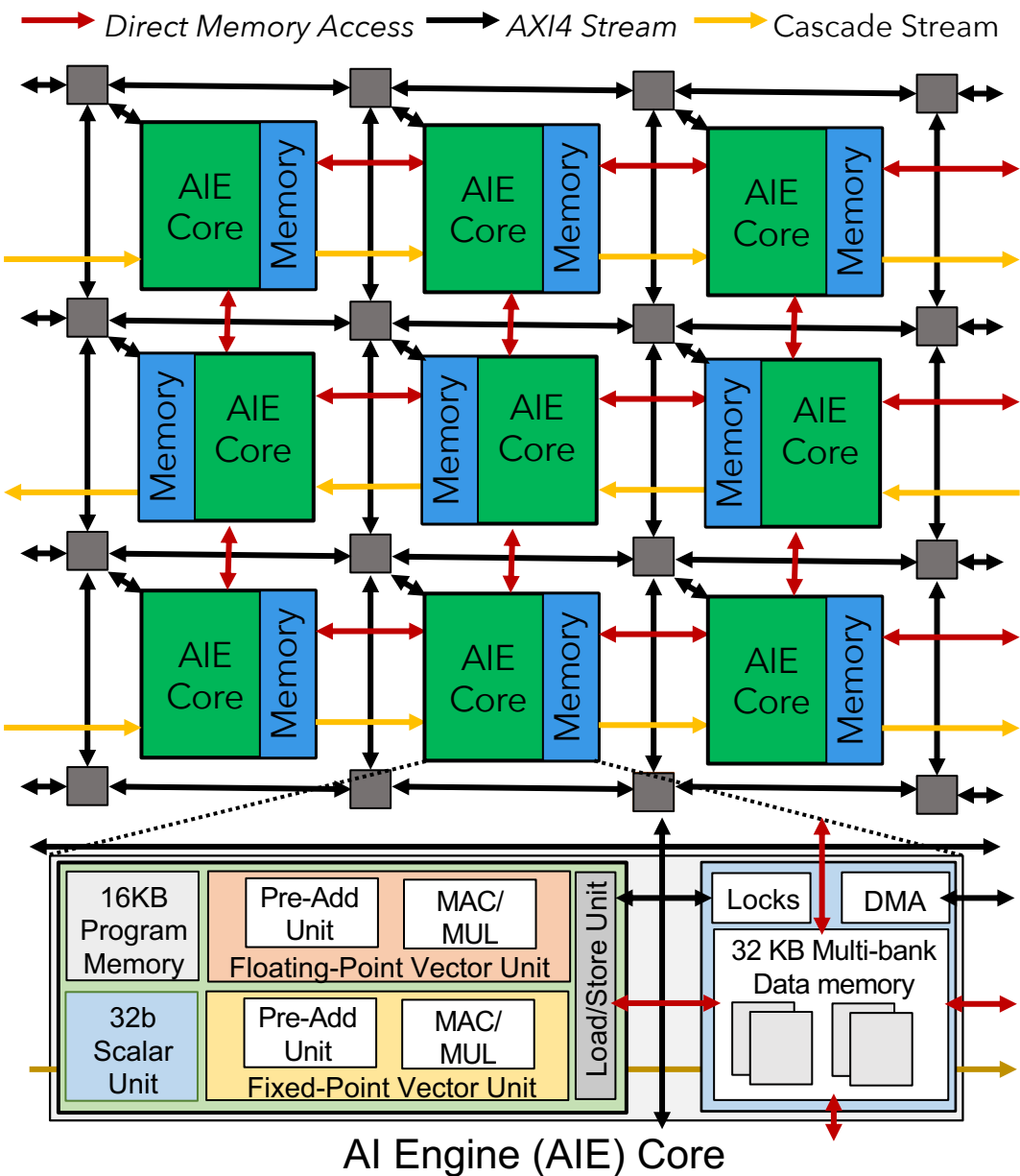
Spatial Architecture: AI Engine

- High compute density
- Allows for **tailoring of the dataflow** to optimize data movement



Spatial Architecture: AI Engine

- High compute density
- Allows for **tailoring of the dataflow** to optimize data movement
- 2D layout of spatial architectures **maps well to processing multi-dimensional stencil grids**



Our Goal

Mitigate the performance bottleneck of compound weather prediction kernels by taking advantage of the characteristics of **spatial computing systems**

Our Proposal



SPARTA

Novel spatial accelerator for
efficient and scalable
horizontal diffusion
weather stencil computation

Talk Outline

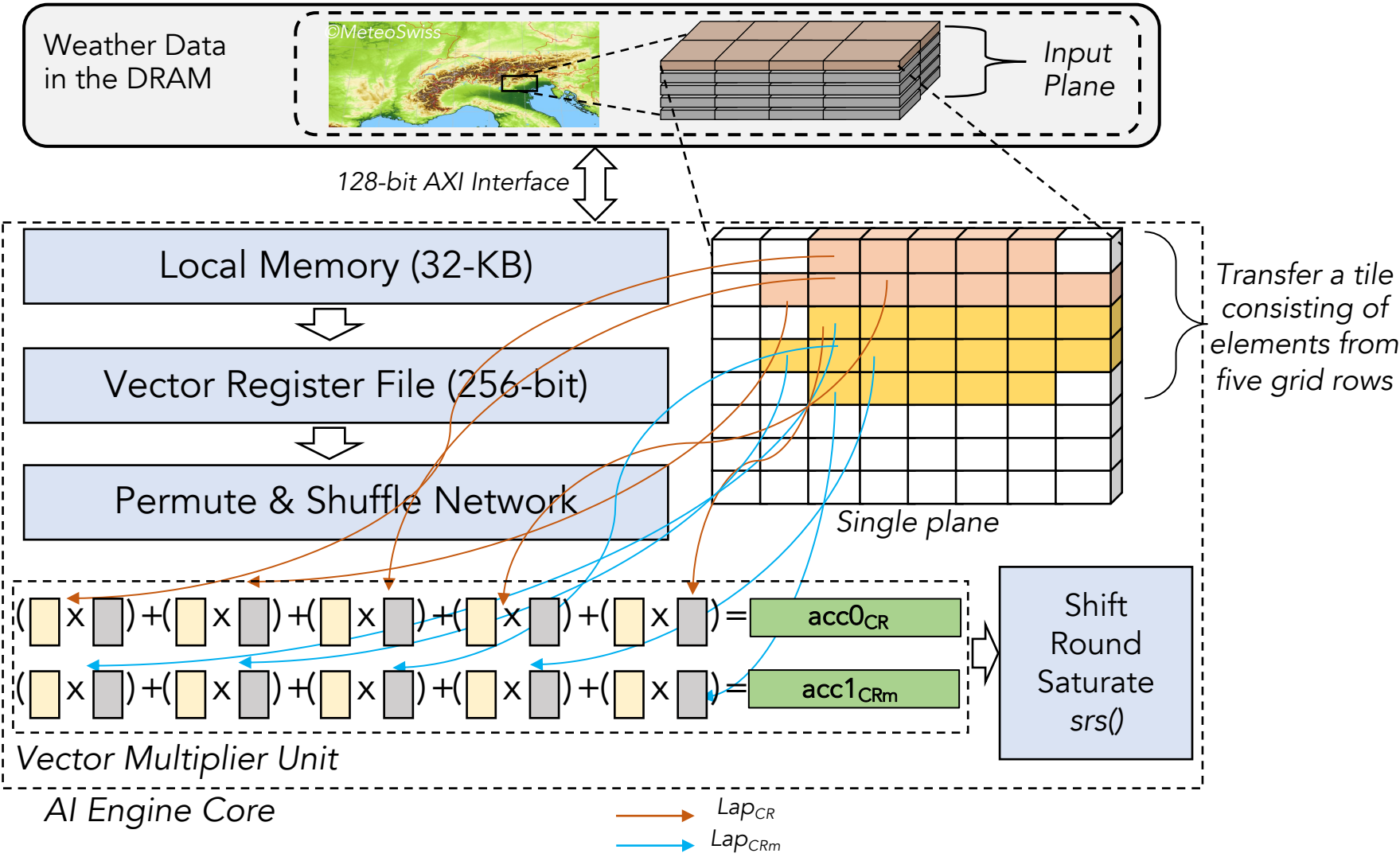
Background and Motivation

SPARTA: Design and Implementation

Evaluation of SPARTA and Key Results

Summary

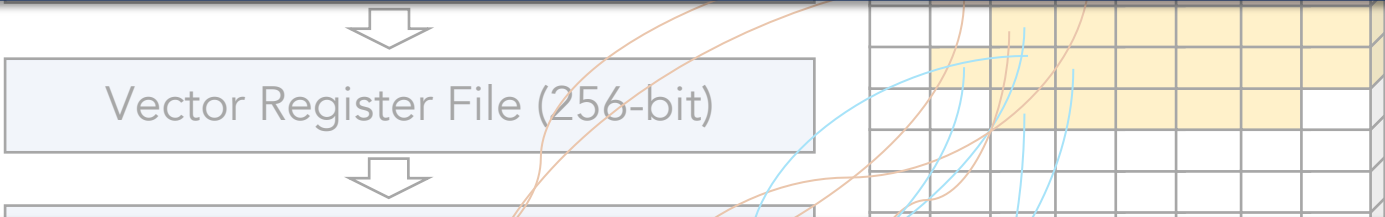
SPARTA Design: Single AIE Core Mapping



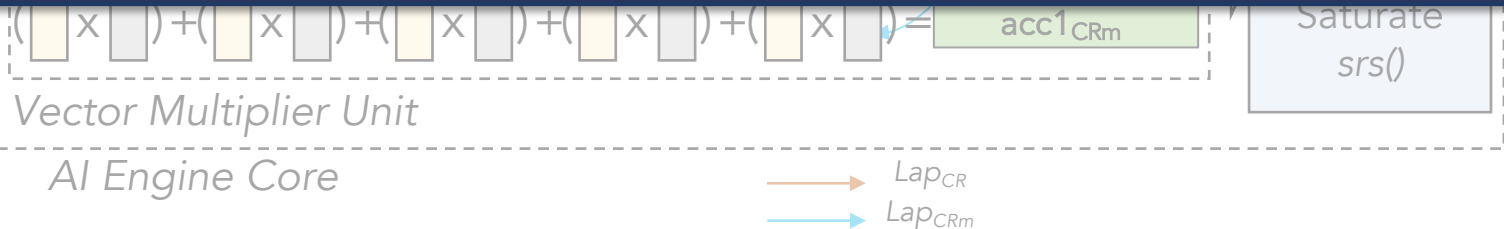
SPARTA Design: Single AIE Core Mapping



Imbalanced computation and memory demands

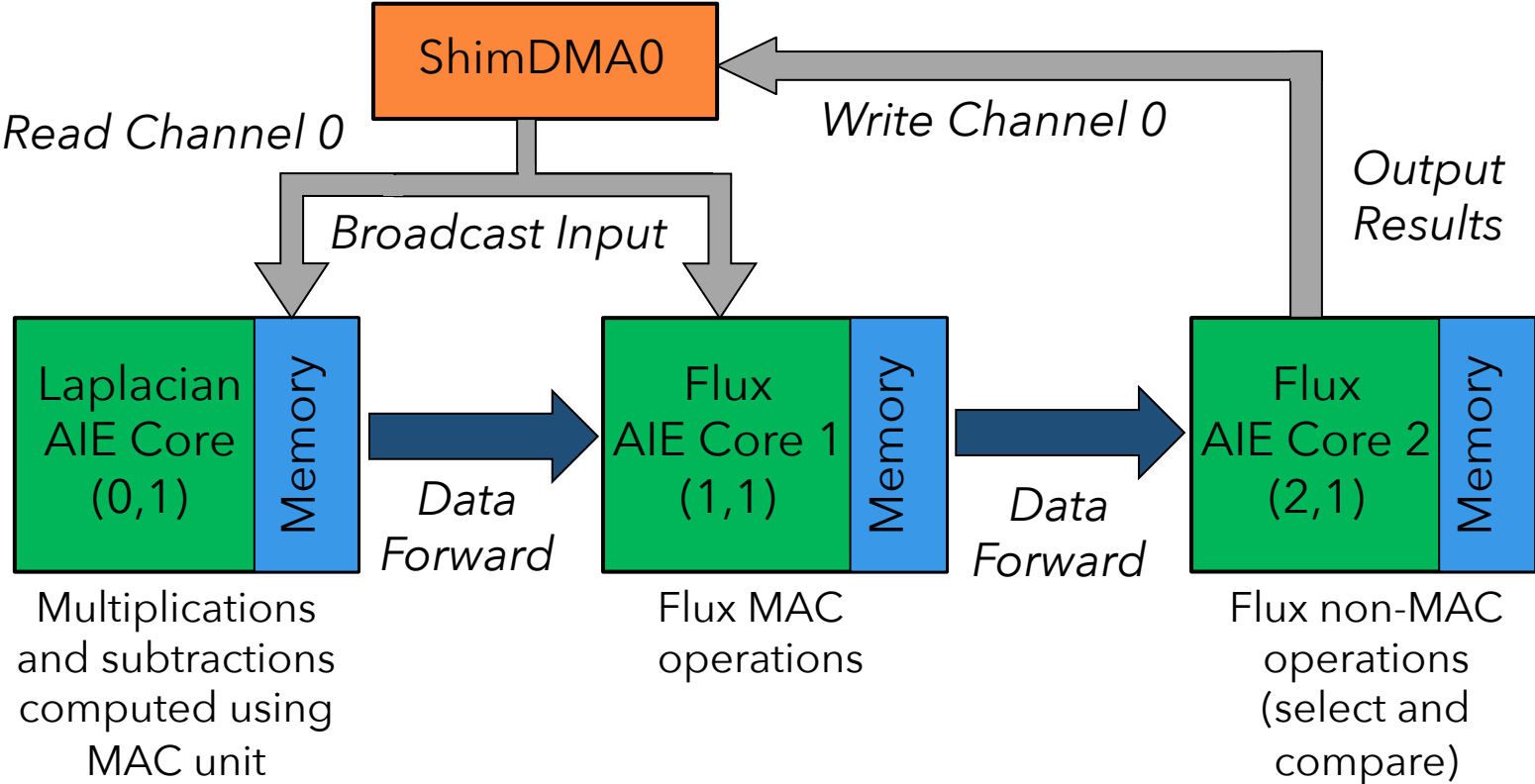


Flux stencils have lower compute-to-memory ratio than Laplacian stencils



SPARTA Design: Multi-AIE Core Mapping

Divide computation over **multiple AIE Cores**: dual-AIE and tri-AIE design



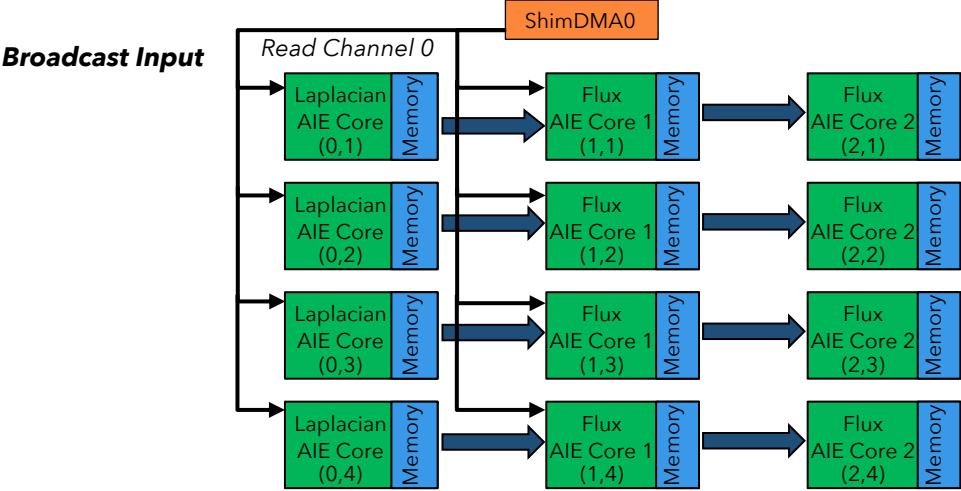
✓ **Compute-bound distributed among multiple cores**

✓ **Parallel execution of multiple AIE cores per shimDMA**

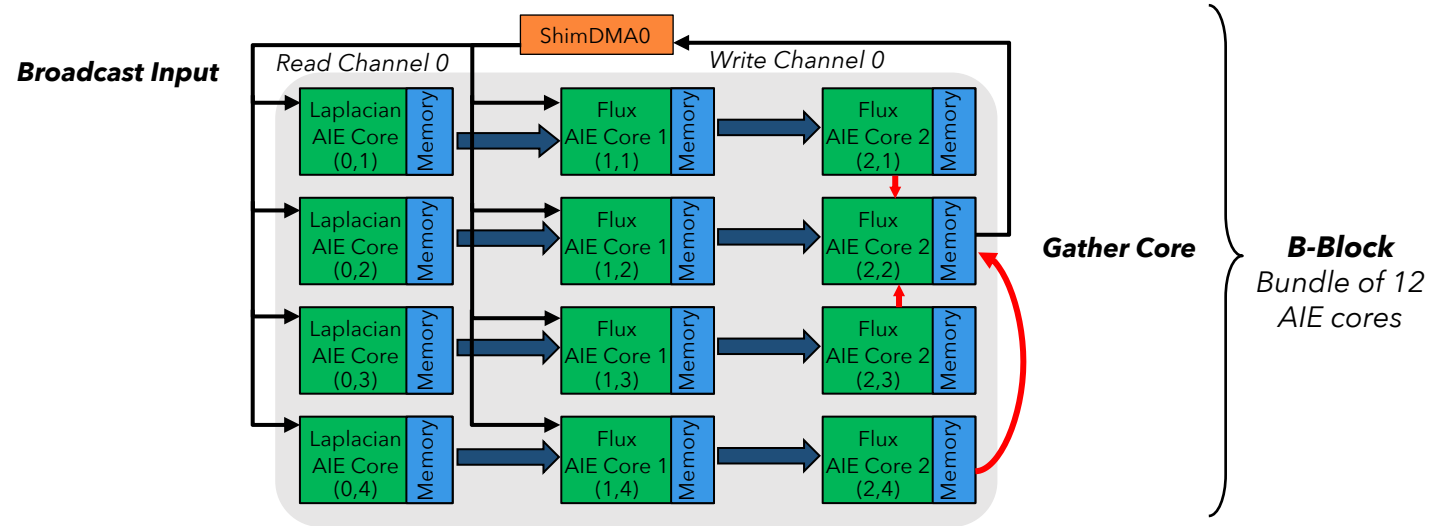
SPARTA Design: Scaling Challenges

1. **Balancing computation and memory resources**
2. **Limited external memory channels**
3. **Gathering and ordering of calculated results** before sending them back to the external memory
4. **Placing input and output cores** close to the external memory interface to optimize data transfer

SPARTA Design: Scaling Accelerator Mapping

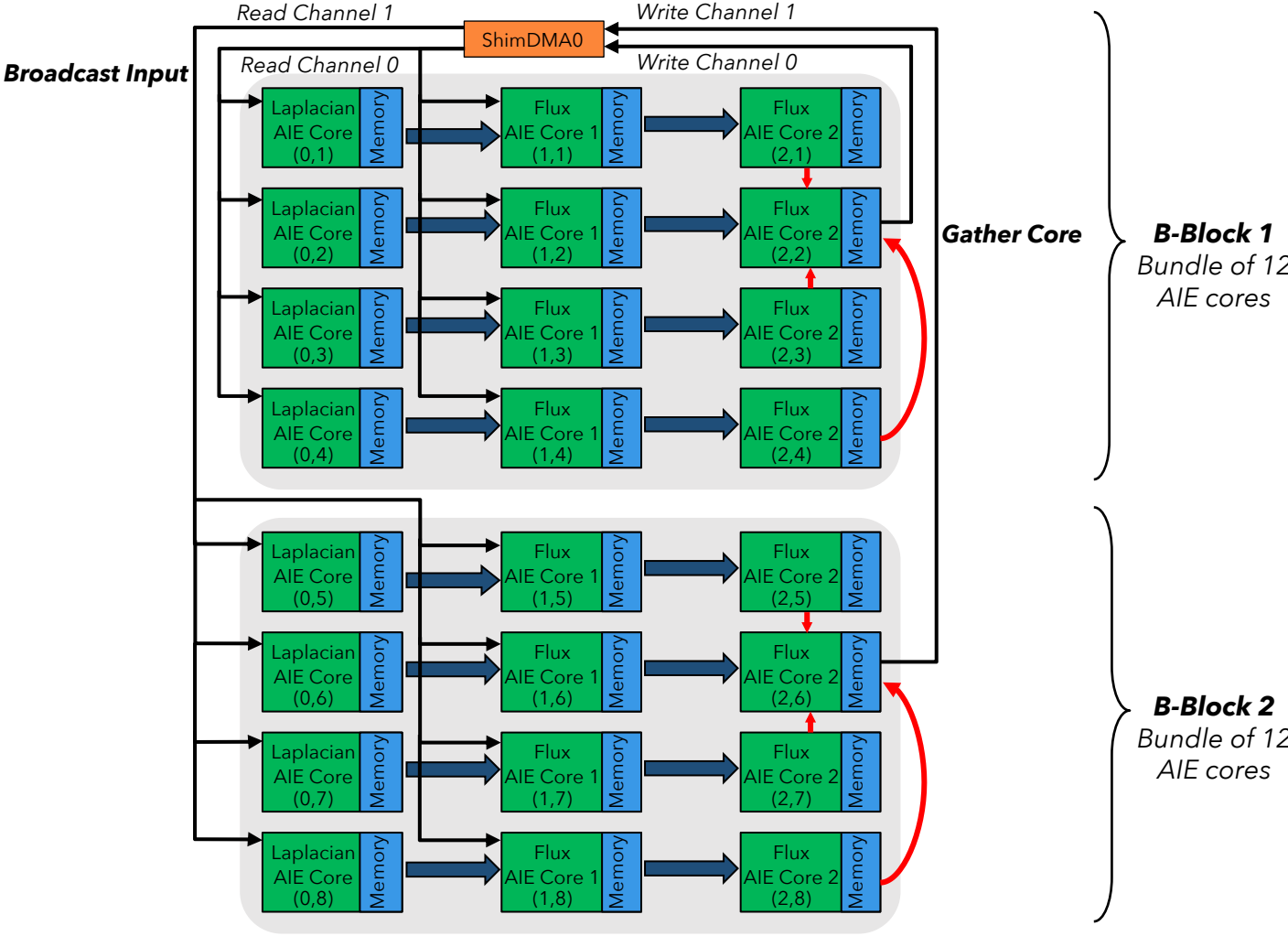


SPARTA Design: Scaling Accelerator Mapping

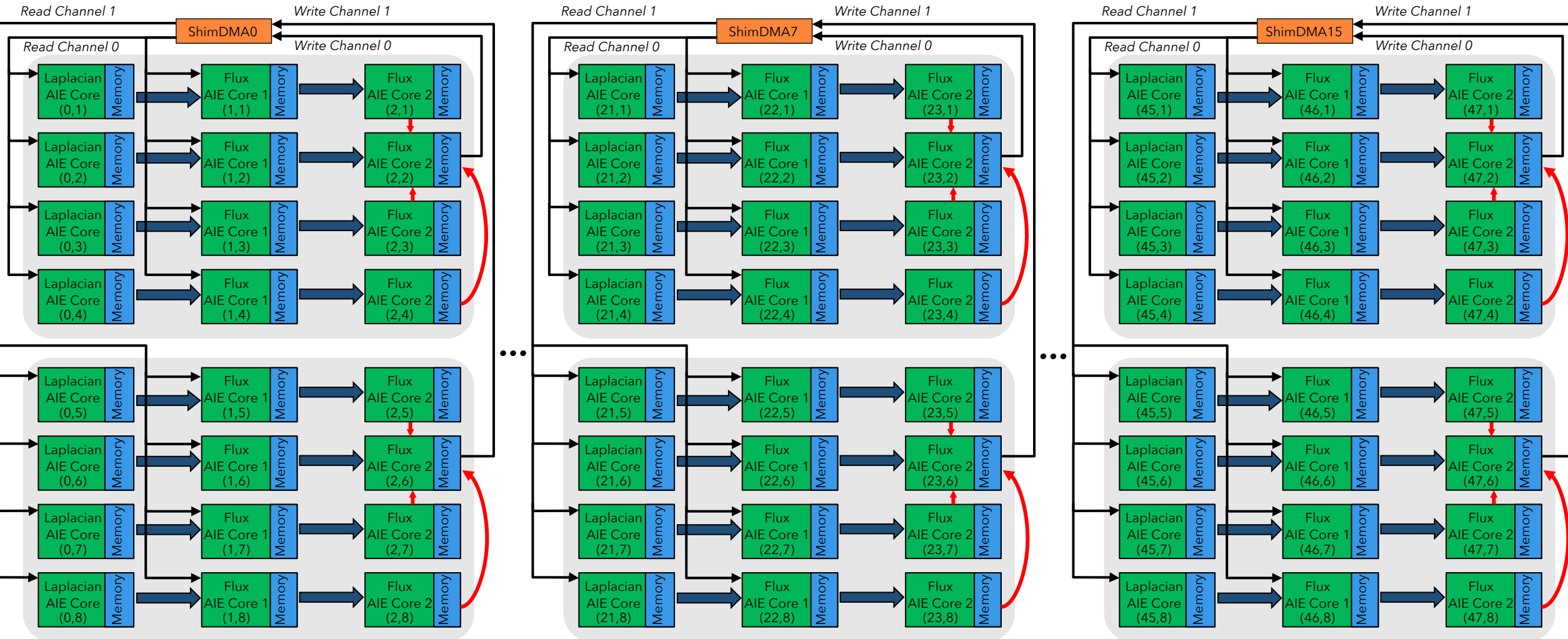


- ✓ Balance compute and memory resources
- ✓ Maximize the usage of shimDMA channels
- ✓ Efficient gathering and reordering of output
- ✓ Exploit data-reuse by broadcasting input data to multiple cores

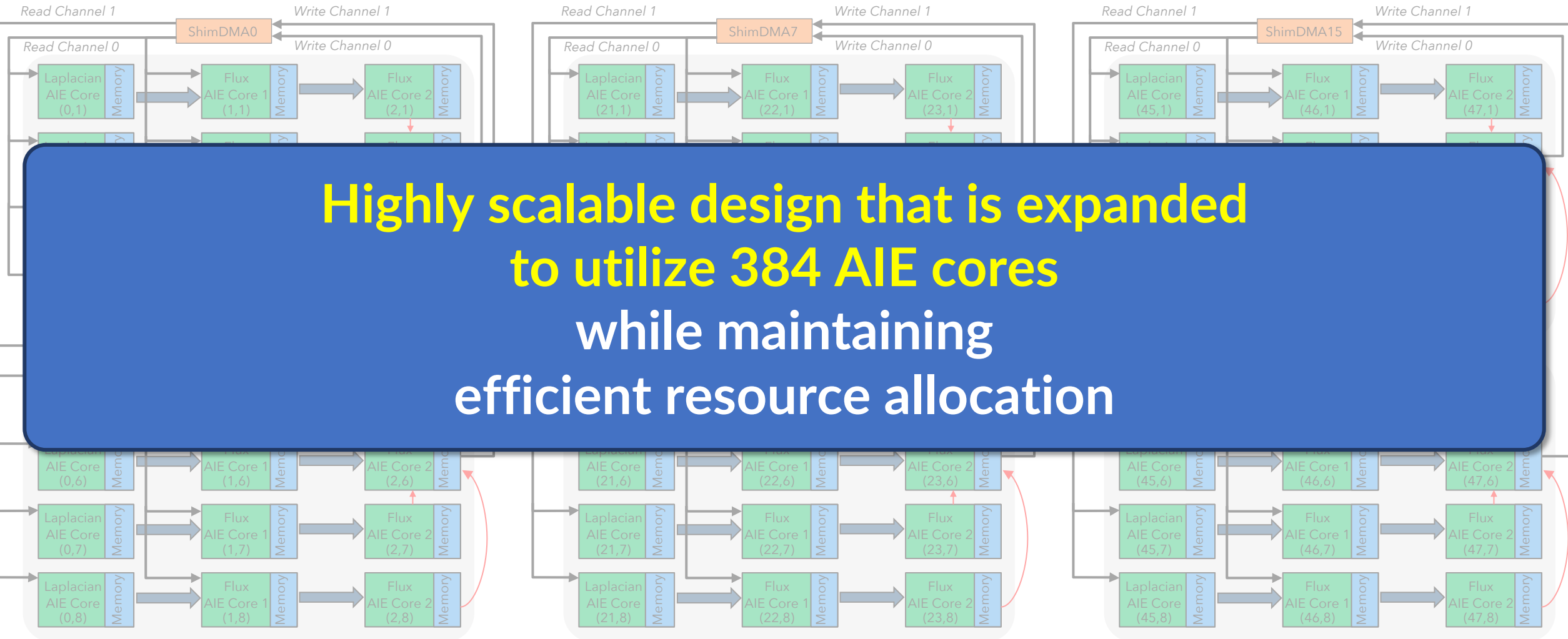
SPARTA Design: Scaling Accelerator Mapping



SPARTA Design: Scaling Accelerator Mapping

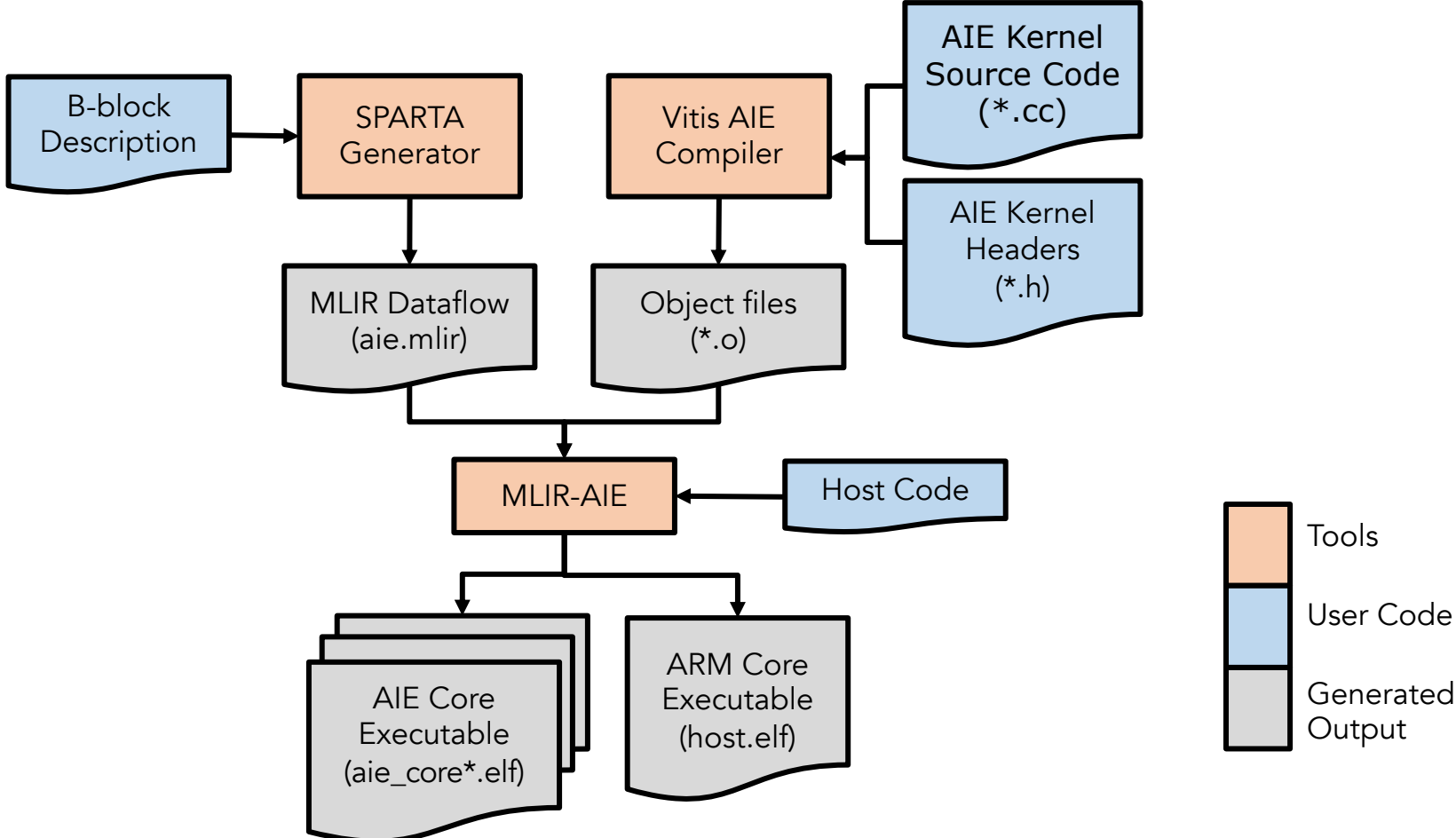


SPARTA Design: Scaling Accelerator Mapping



SPARTA Application Toolflow

MLIR (Multi-Level Intermediate Representation) to separate the AIE core computation optimization and the effective dataflow management



<https://github.com/Xilinx/mlir-aie>

Talk Outline

Background and Motivation

SPARTA: Design and Implementation

Evaluation of SPARTA and Key Results

Summary

Evaluation Methodology (1/2)

Real system evaluation

Versal AIE Configuration

- Frequency: 1 GHz
- Cores: 400

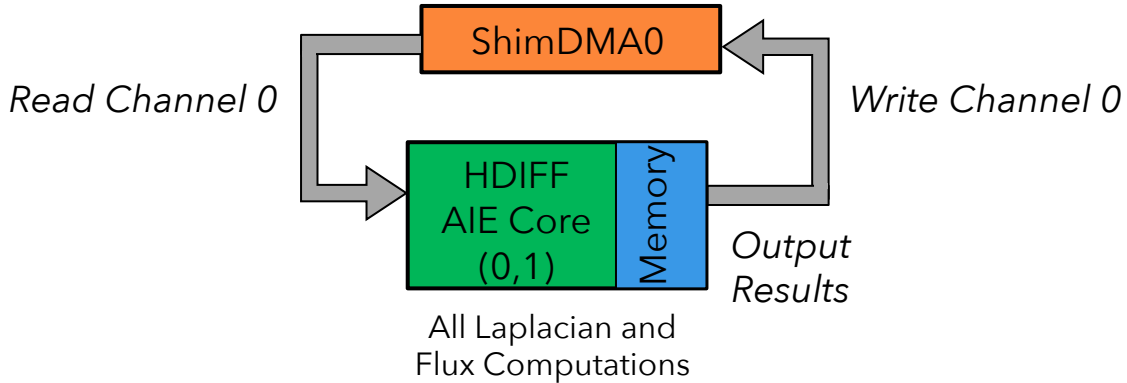


Evaluation Methodology (2/2)

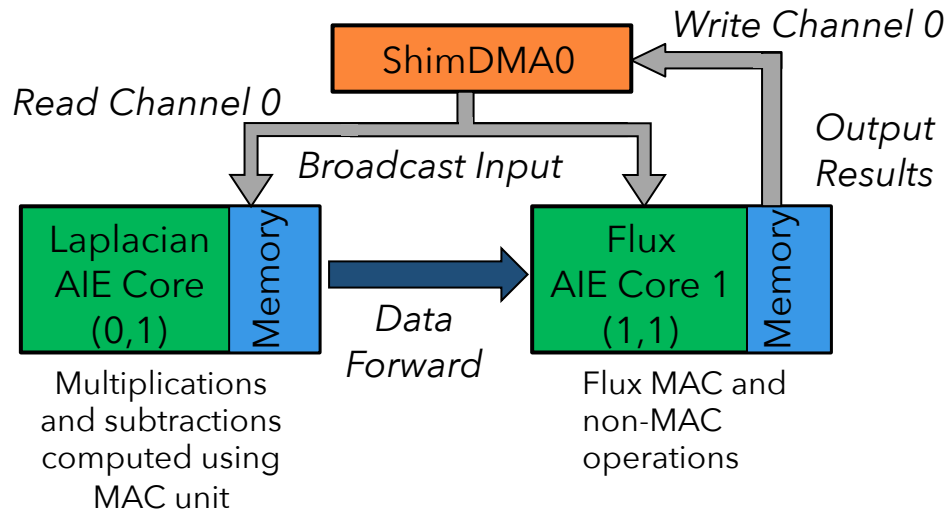
- **State-of-the-art baselines** on all major computing platforms
 - CPU [Singh+, FPL'21]
 - GPU [Licht+, CGO'21]
 - FPGA [Singh+, FPL'21]
- **Programming tools**
 - MLIR-AIE
 - Vitis Chess Compiler v2022.2
- **Elementary stencil benchmarks**
 - jacobi-1d
 - jacobi-2d-3pt
 - Laplacian
 - jacobi-2d-9pt
 - seidel-2d

SPARTA Performance: Single and Multi-AIE Design (1/2)

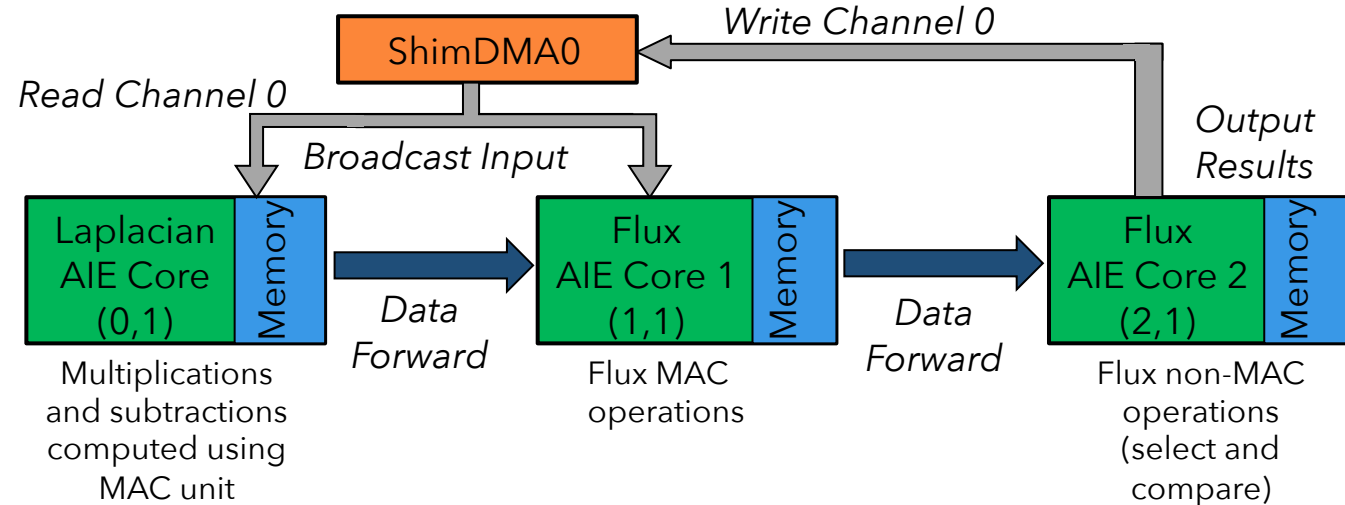
Single AIE



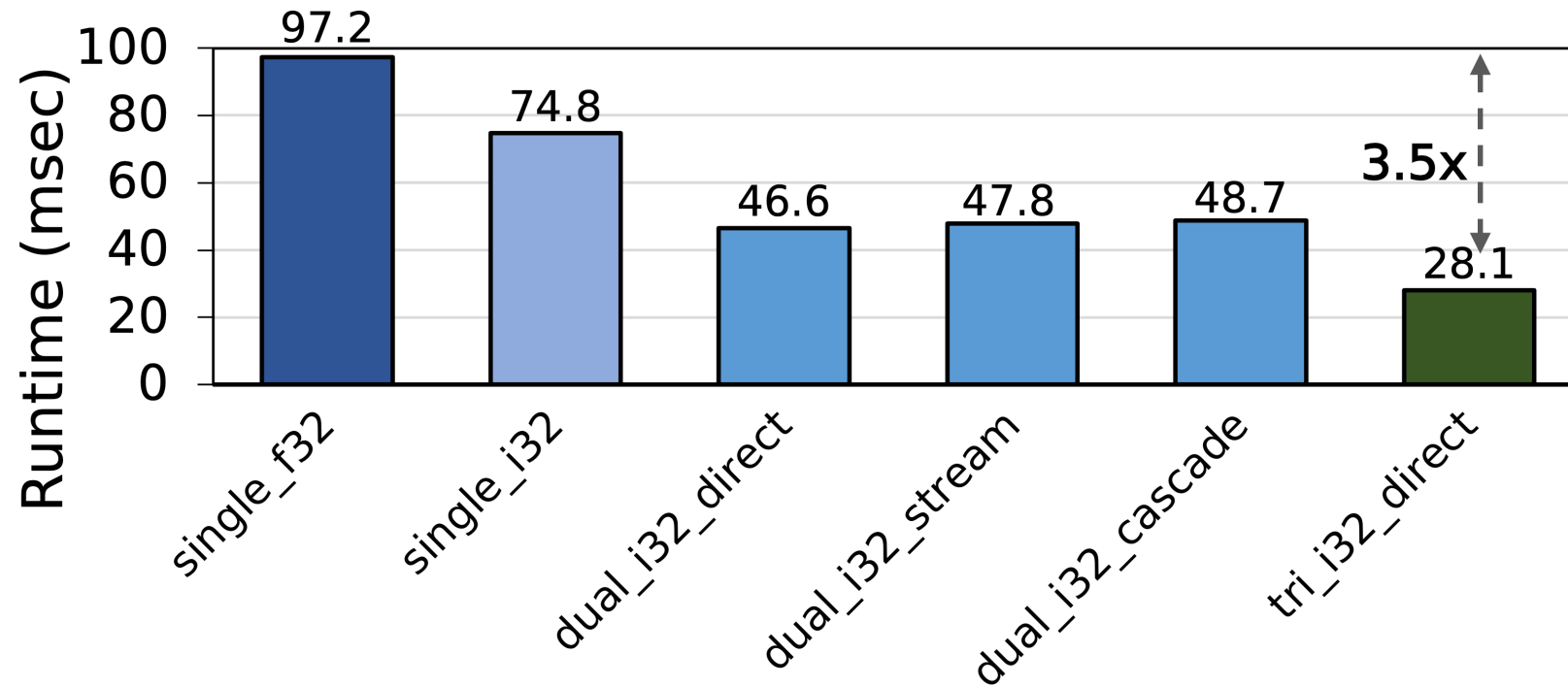
Dual-AIE



Tri-AIE

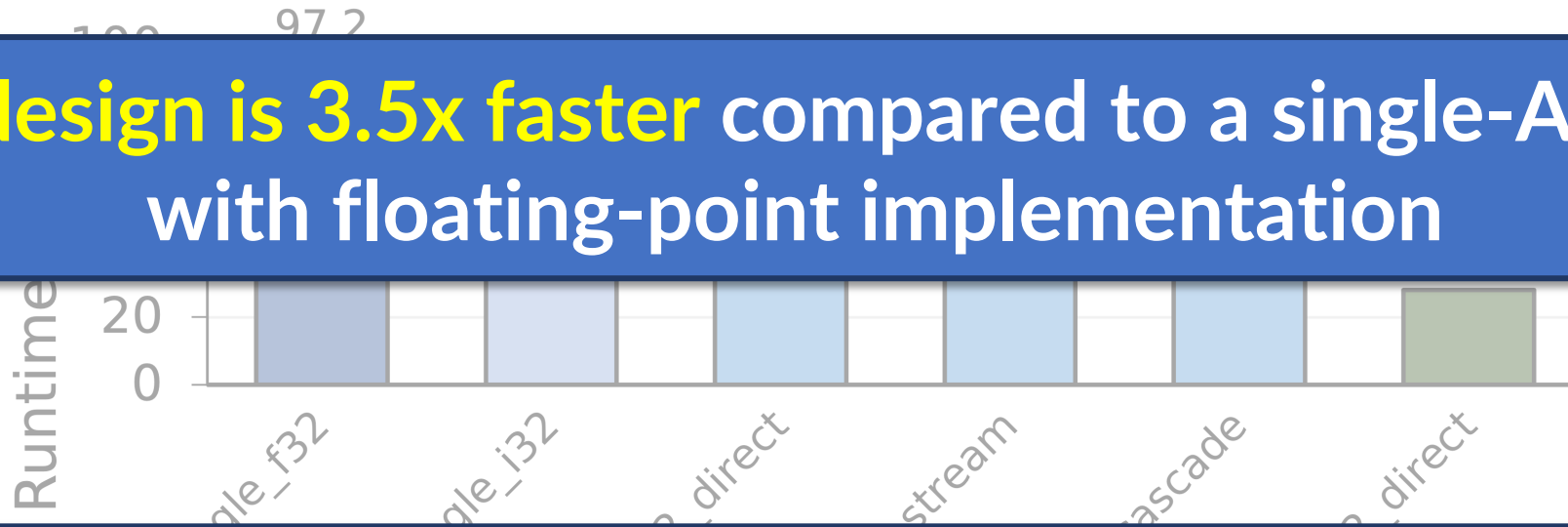


SPARTA Performance: Single and Multi-AIE Design (1/2)



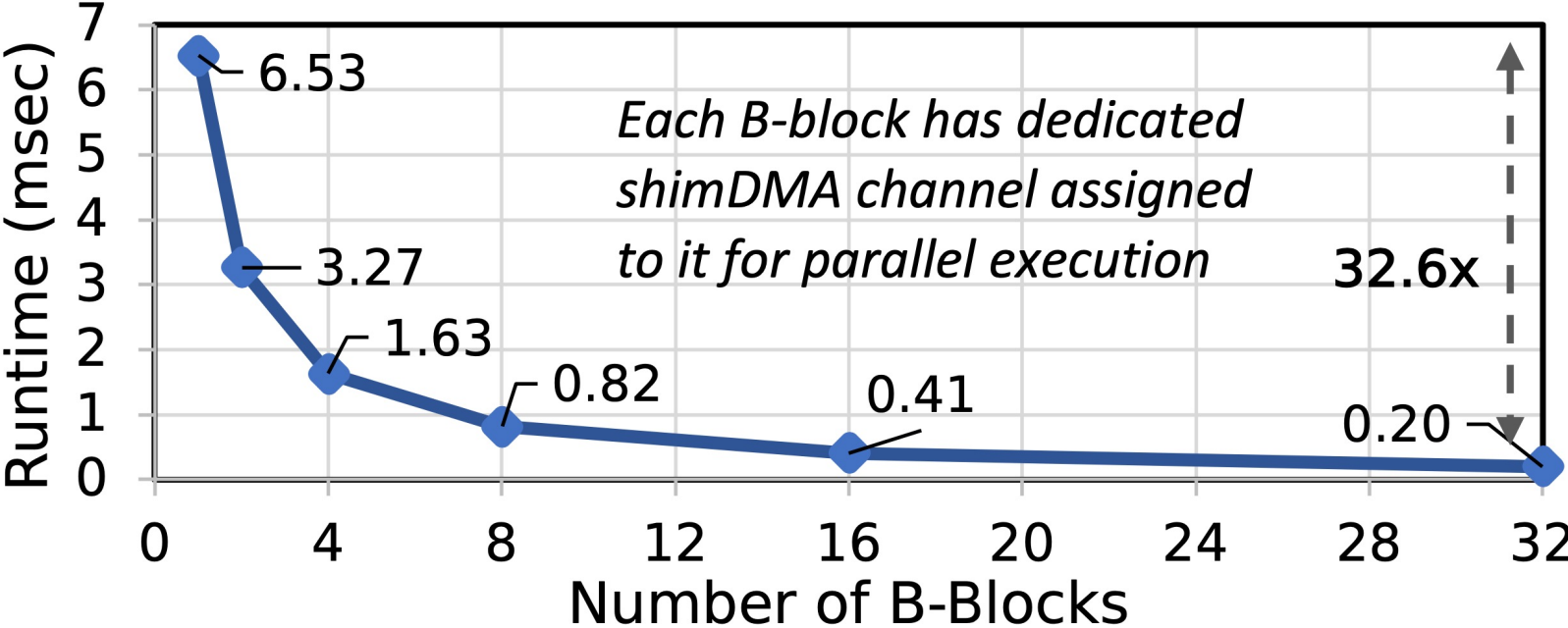
SPARTA Performance: Single and Multi-AIE Design (1/2)

Tri-AIE design is 3.5x faster compared to a single-AIE design with floating-point implementation



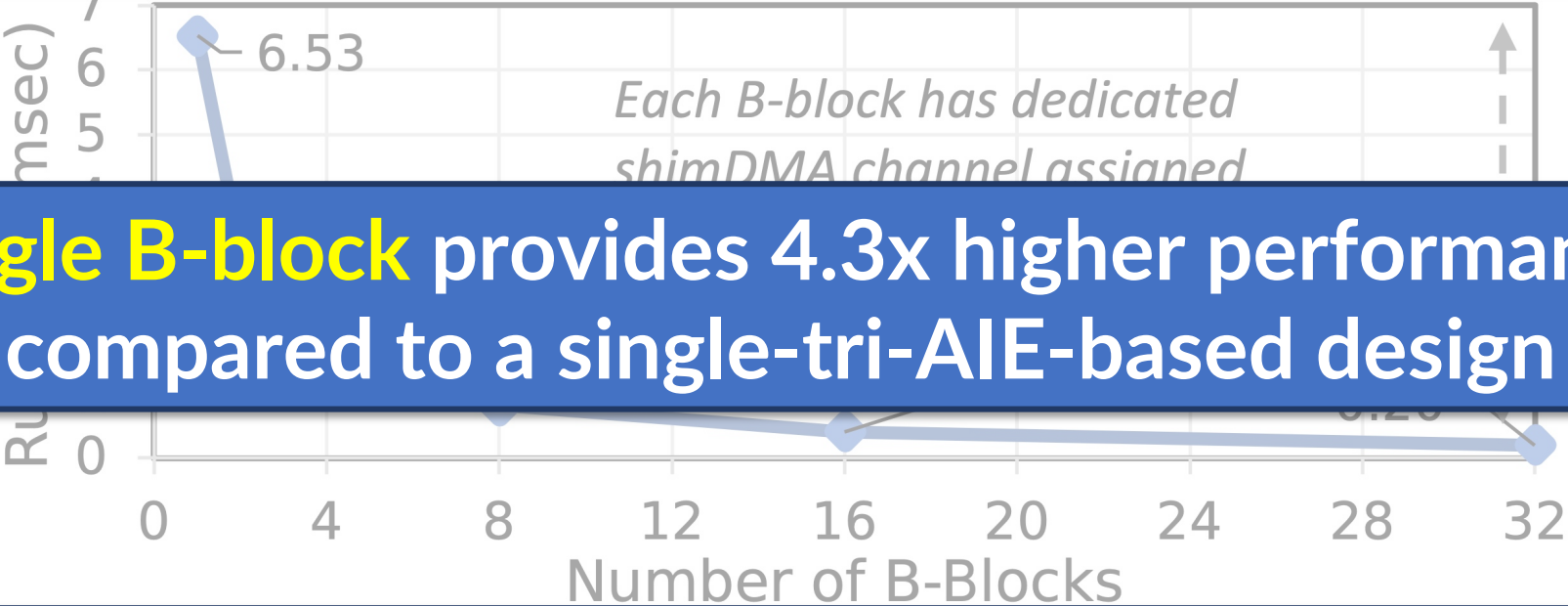
Dual-AIE designs provide 1.9x-2.1x performance than a single-AIE design depending upon the data forwarding interface

SPARTA Performance: Scaling Accelerator Design



SPARTA Performance: Scaling Accelerator Design

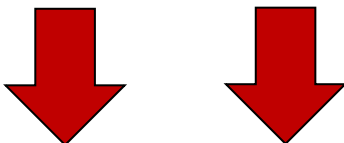
SPARTA is 32.6x faster than single-B-block-based design



Single B-block provides 4.3x higher performance compared to a single-tri-AIE-based design

Performance scales linearly with the number of B-blocks

SPARTA Performance: Comparison to SOTA



Stencil	Work	Year	Platform	Device	Mem. Tech.	Peak Perf. (TFLOPS)	Peak B/W (GB/s)	Perf. (GOp/s)	Arch. Roof. (%)
hdiff	[23]	2019	FPGA	XCVU3P [97]	DDR4	0.97	25.6	129.9	13.4%
hdiff	[16]	2021	CPU	Xeon E5-2690V3 [98]	DDR4	0.24	68.0	32.0	13.0%
hdiff	[24]	2021	CPU	POWER9 [31]	DDR4	0.49	110.0	58.5	11.8%
hdiff	[16]	2021	GPU	V100 [32]	HBM2	14.1	900.0	849.0	6.1%
hdiff	[16]	2021	FPGA	Stratix 10 [99]	DDR4	9.2	76.8	145.0	1.6%
hdiff	[24]	2021	FPGA	XCVU37P [97]	HBM	3.6	410.0	485.4	13.5%
hdiff	SPARTA	2023	AIE	XCVC1902 [83]	DDR4	3.1	25.6	995.7	32.2%

SPARTA outperforms state-of-the-art CPU, GPU, and FPGA implementations by 17.1x, 1.2x, and 2.1x, respectively

SPARTA Performance: Comparison to SOTA

Stencil	Work	Year	Platform	Device	Mem. Tech.	Peak Perf. (TFLOPS)	Peak B/W (GB/s)	Perf. (GOp/s)	Arch. Roof. (%)
hdiff	[23]	2019	FPGA	XCVU3P [97]	DDR4	0.97	25.6	129.9	13.4%
hdiff	[16]	2021	CPU	Xeon E5-2690V3 [98]	DDR4	0.24	68.0	32.0	13.0%
hdiff	[24]	2021	CPU	POWER9 [31]	DDR4	0.49	110.0	58.5	11.8%
hdiff	[16]	2021	GPU	V100 [32]	HBM2	14.1	900.0	849.0	6.1%
hdiff	[16]	2021	FPGA	Stratix 10 [99]	DDR4	9.2	76.8	145.0	1.6%
hdiff	[24]	2021	FPGA	XCVU37P [97]	HBM	3.6	410.0	485.4	13.5%
hdiff	SPARTA	2023	AIE	XCVC1902 [83]	DDR4	3.1	25.6	995.7	32.2%

State-of-the-art implementations achieve only 1.6%-13.5% of the peak theoretical performance of a platform

SPARTA achieves the highest peak roofline performance of 32.2%

SPARTA Performance: Comparison to SOTA

Stencil	Work	Year	Platform	Device	Mem. Tech.	Peak Perf. (TFLOPS)	Peak B/W (GB/s)	Perf. (GOp/s)	Arch. Roof. (%)
hdiff	[23]	2019	FPGA	XCVU3P [97]	DDR4	0.97	25.6	129.9	13.4%
hdiff	[16]	2021	CPU	Xeon E5-2690V3 [98]	DDR4	0.24	68.0	32.0	13.0%
hdiff	[24]	2021	CPU	POWER9 [31]	DDR4	0.49	110.0	58.5	11.8%
hdiff	[16]	2021	GPU	V100 [32]	HBM2	14.1	900.0	849.0	6.1%
hdiff	[16]	2021	FPGA	Stratix 10 [99]	DDR4	9.2	76.8	145.0	1.6%
hdiff	[24]	2021	FPGA	XCVU37P [97]	HBM	3.6	410.0	485.4	13.5%
hdiff	SPARTA	2023	AIE	XCVC1902 [83]	DDR4	3.1	25.6	995.7	32.2%

SPARTA is 2.4x more energy-efficient with 42.2 GOps/Watt than the state-of-the-art FPGA design

More in the Paper

- Results for **elementary stencil benchmarks**
- **Analytical modeling** for computation and memory requirements
- **Implementation details** for single and multi-AIE core mapping
- **Managing data transfer using MLIR**
- **Discussion and key takeaways**

More in the Paper

- Results for elementary stencil benchmarks

SPARTA: Spatial Acceleration for Efficient and Scalable Horizontal Diffusion Weather Stencil Computation

Gagandeep Singh^{a,b} Alireza Khodamoradi^a Kristof Denolf^a Jack Lo^a
Juan Gómez-Luna^b Joseph Melber^a Andra Bisca^a
Henk Corporaal^c Onur Mutlu^b
^aAMD Research ^bETH Zürich ^cEindhoven University of Technology



Full Paper

<https://arxiv.org/pdf/2303.03509.pdf>

- Discussion and key takeaways

SPARTA is Open Sourced

CMU-SAFARI / SPARTA Public

Notifications Fork 2 Star 8

Code Issues Pull requests Actions Projects Security Insights

main 1 branch 0 tags Go to file Code

singagan Update README.md e4a76ee on May 9 6 commits

HDIFF_dual_AIE_obj...	hdiff code	4 months ago
HDIFF_single_AIE_o...	hdiff code	4 months ago
HDIFF_single_AIE_o...	hdiff code	4 months ago
HDIFF_single_AIE_o...	hdiff code	4 months ago
HDIFF_tri_AIE_objec...	hdiff code	4 months ago
HDIFF_tri_AIE_objec...	hdiff code	4 months ago
img	hdiff code	4 months ago
LICENSE	hdiff code	4 months ago
README.md	Update README.md	last month

About

A novel spatial accelerator for horizontal diffusion weather stencil computation, as described in ICS 2023 paper by Singh et al. (<https://arxiv.org/pdf/2303.03509.pdf>)

- Readme
- MIT license
- 8 stars
- 5 watching
- 2 forks

Report repository

Releases

No releases published



SPARTA Code

<https://github.com/CMU-SAFARI/SPARTA>

<https://github.com/Xilinx/mlir-ai>

Talk Outline

Background and Motivation

SPARTA: Design and Implementation

Evaluation of SPARTA and Key Results

Summary

Summary

Mitigate the performance bottleneck of compound weather prediction kernels by taking advantage of the characteristics of spatial computing systems

SPARTA is a **novel spatial accelerator** for efficient and scalable **horizontal diffusion weather stencil computation**

SPARTA outperforms state-of-the-art CPU, GPU, and FPGA horizontal diffusion implementations by 17.1x, 1.2x, and 2.1x, respectively



SPARTA

Spatial Acceleration for Efficient and Scalable Horizontal Diffusion Weather Stencil Computation

Gagandeep Singh, Alireza Khodamoradi, Kristof Denolf, Jack Lo, Juan Gómez-Luna, Joseph Melber, Andra Bisca, Henk Corporaal, and Onur Mutlu

37th International Conference on Supercomputing (ICS)
Orlando, Florida

SAFARI
SAFARI Research Group
safari.ethz.ch

ETH zürich

TU/e

EINDHOVEN
UNIVERSITY OF
TECHNOLOGY

AMD 
together we advance_