

Performance Analysis of GEMM Workloads on the AMD Versal Platform

ISFPGA
2025

Kaustubh Mhatre, Prashant Mulleti, Curt Bansil, Endri Taka, Aman Arora

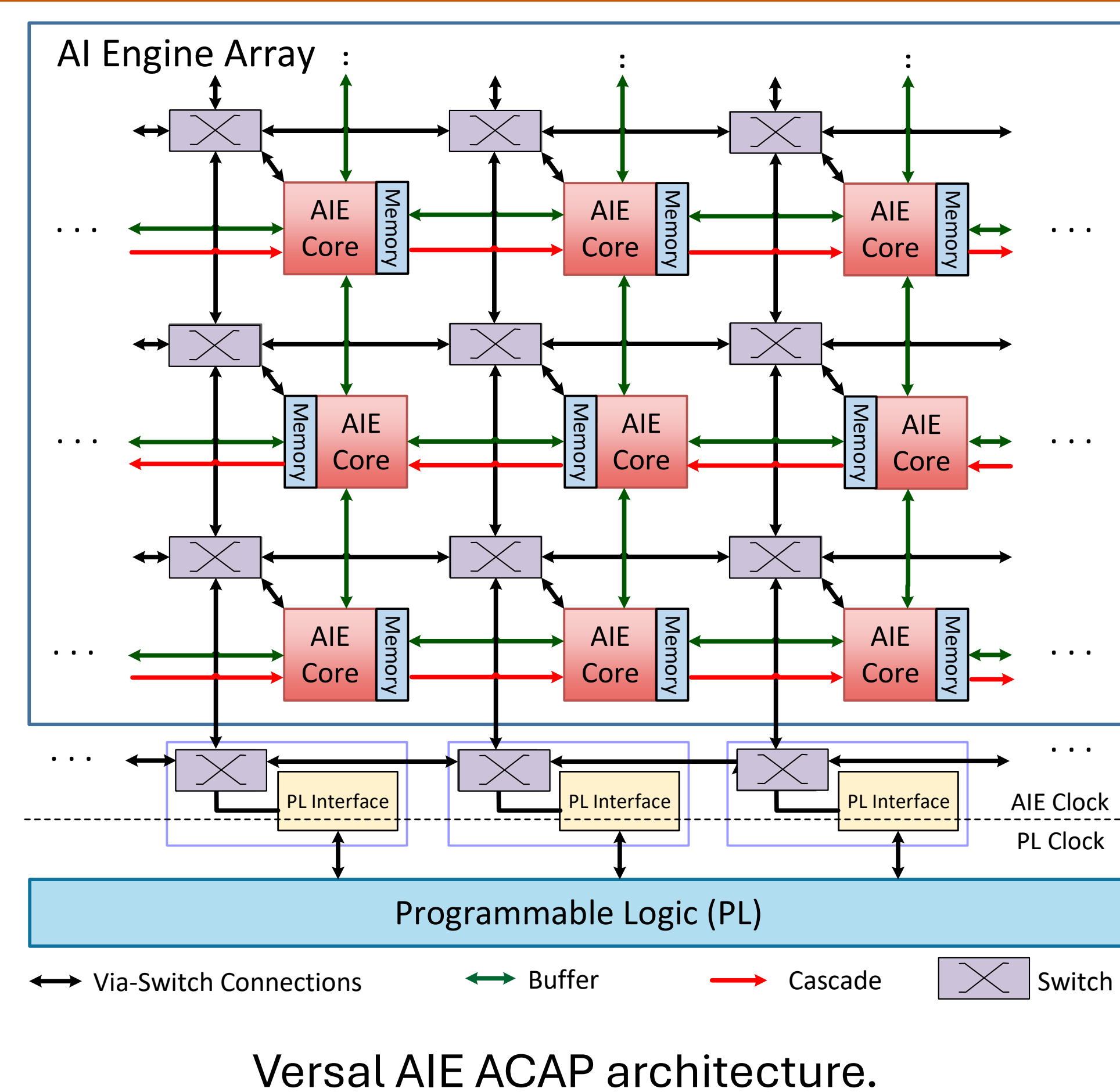
Advent Lab
Arizona State University



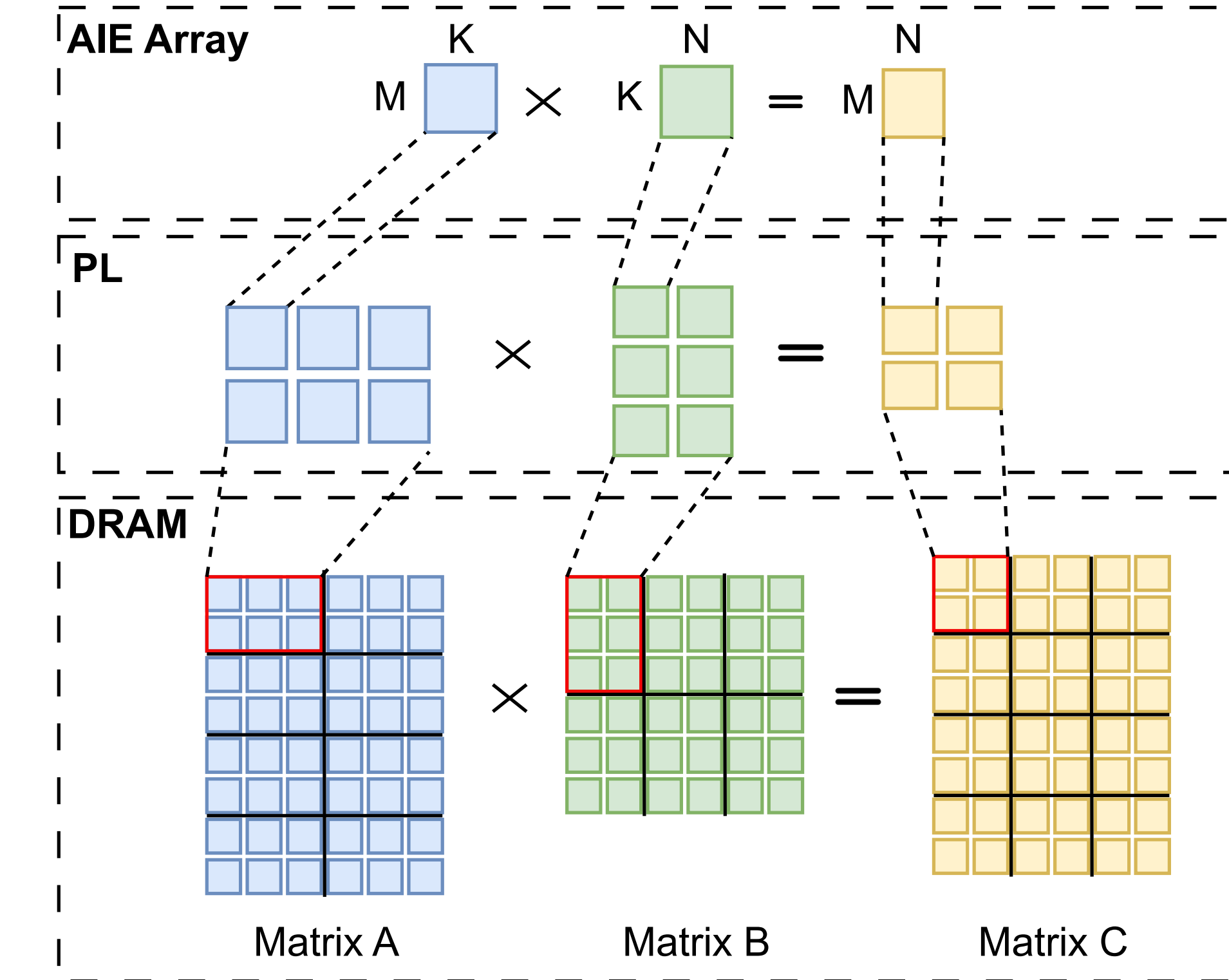
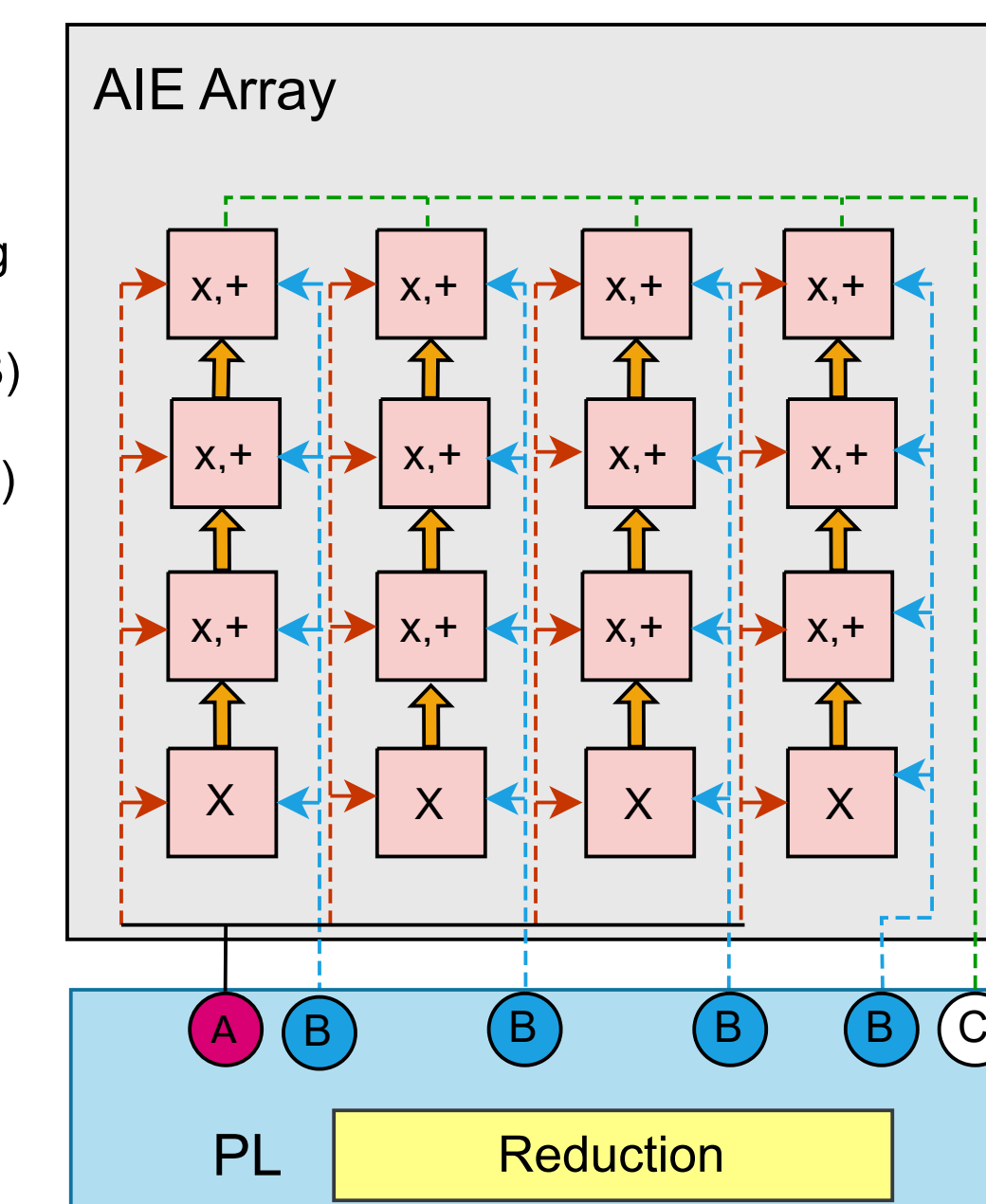
Introduction

- AMD Versal is a new heterogeneous computing hardware architecture comprised of adaptive intelligence (AI) engines, programmable logic, and a processing system.
- General Matrix Multiplication (GEMM) is the fundamental building block of modern deep learning (DL) applications such as Chat-GPT, and GEMM workloads can be mapped onto Versal in different ways, each with distinct trade-offs.
- We present a thorough analysis of GEMM workloads of different shapes and sizes, showcasing performance artifacts associated with the AMD Versal architecture.
- Focusing on the unique aspects of the Versal architecture, We analyzed workload scaling, sensitivity to the hardware architecture parameters and, system efficiency.

Versal Architecture and GEMM Mapping

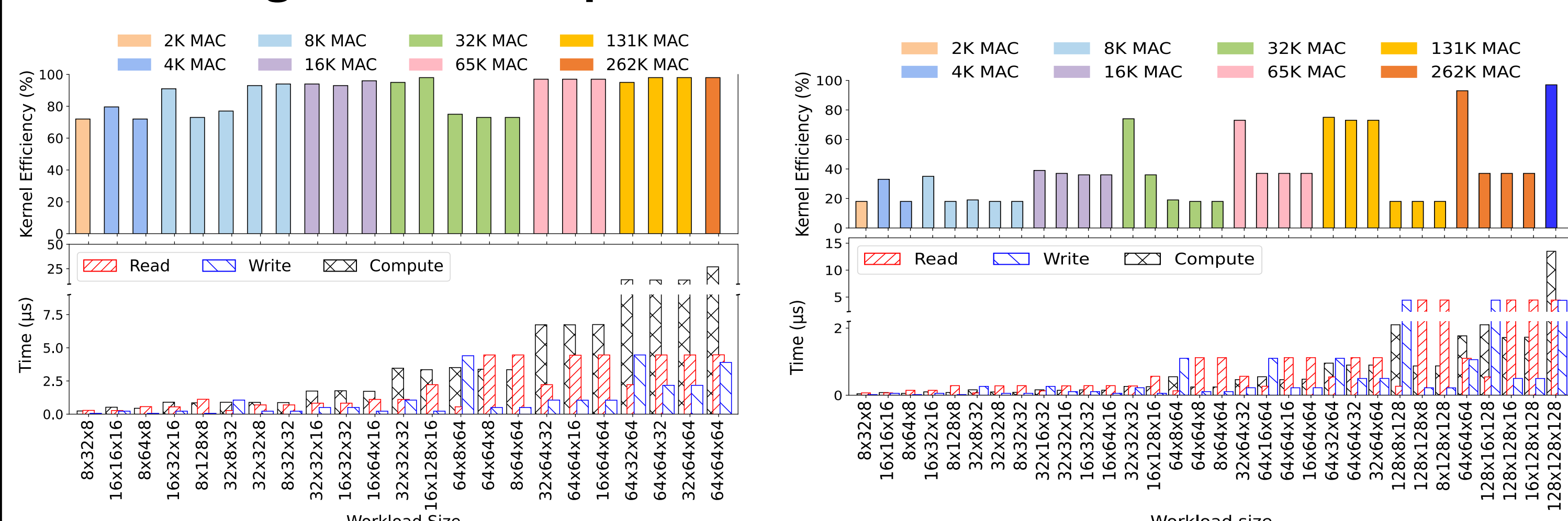


- Packet switching (out)
- Broadcast/Circuit switching
- Packet Switching (Matrix B)
- Packet Switching (Matrix A)
- Cascaded connection

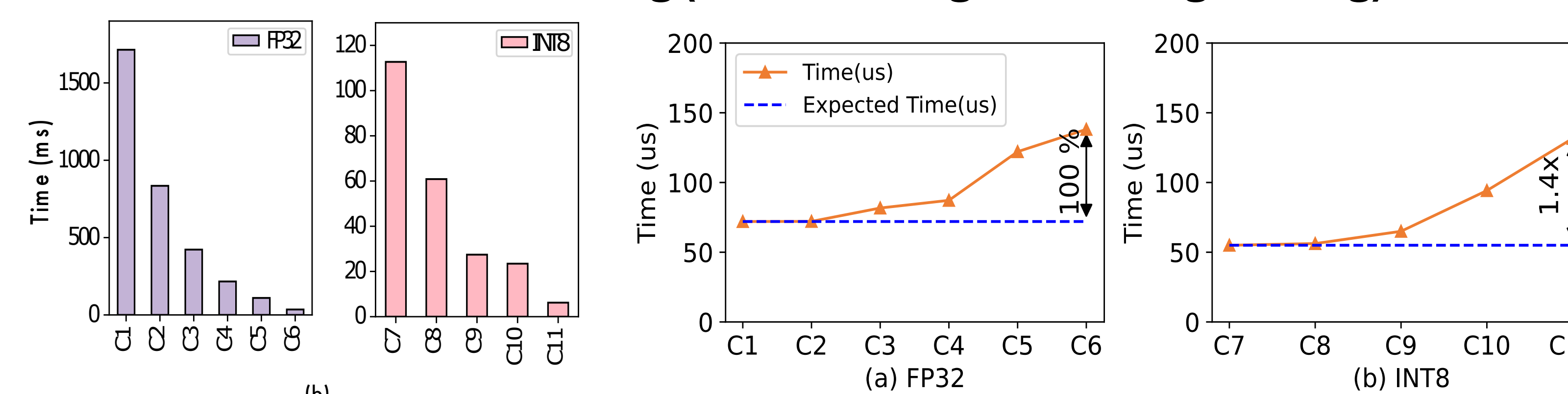


Analysis and Results

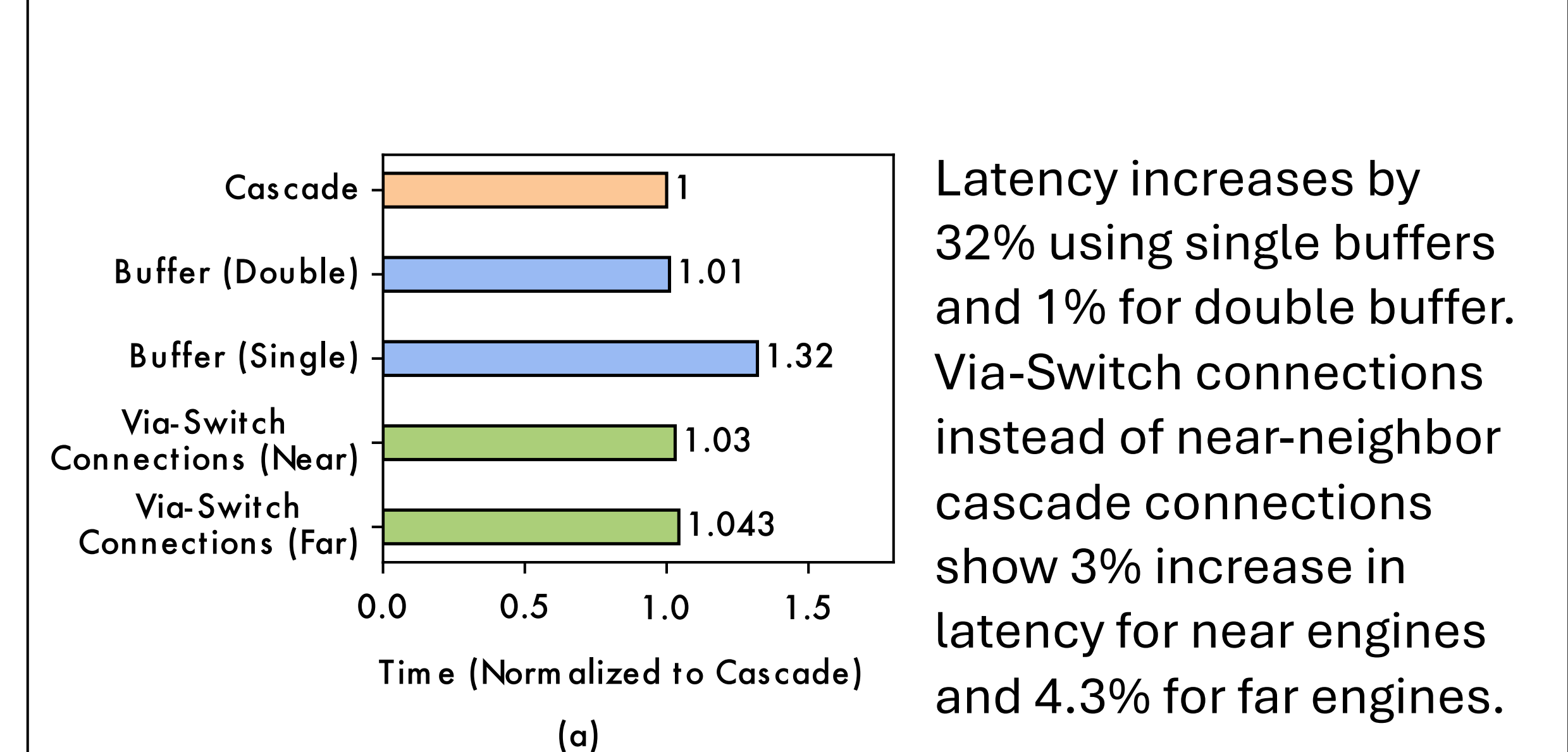
Single AIE kernel performance for different matrix size



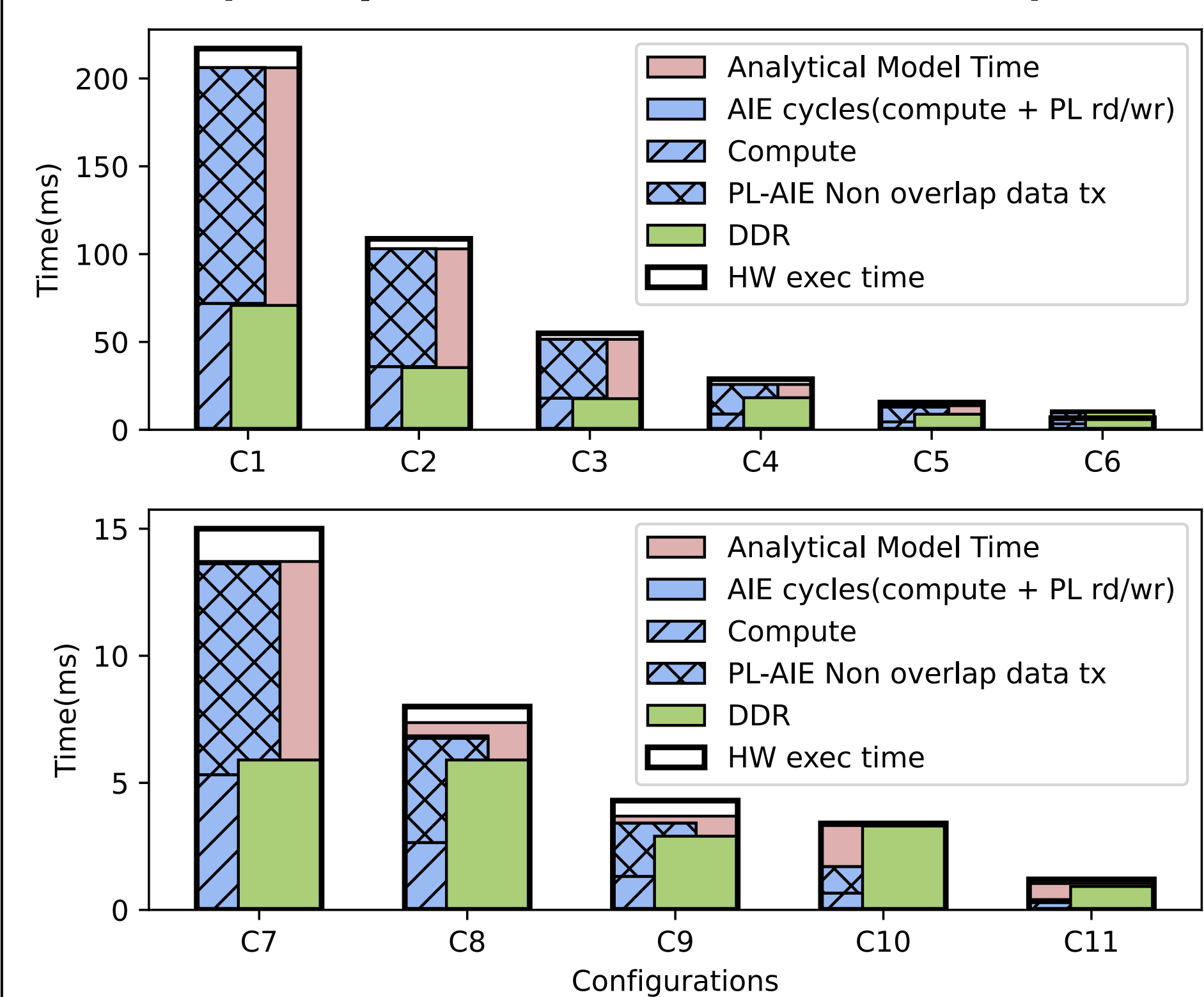
Performance scaling (weak scaling and strong scaling)



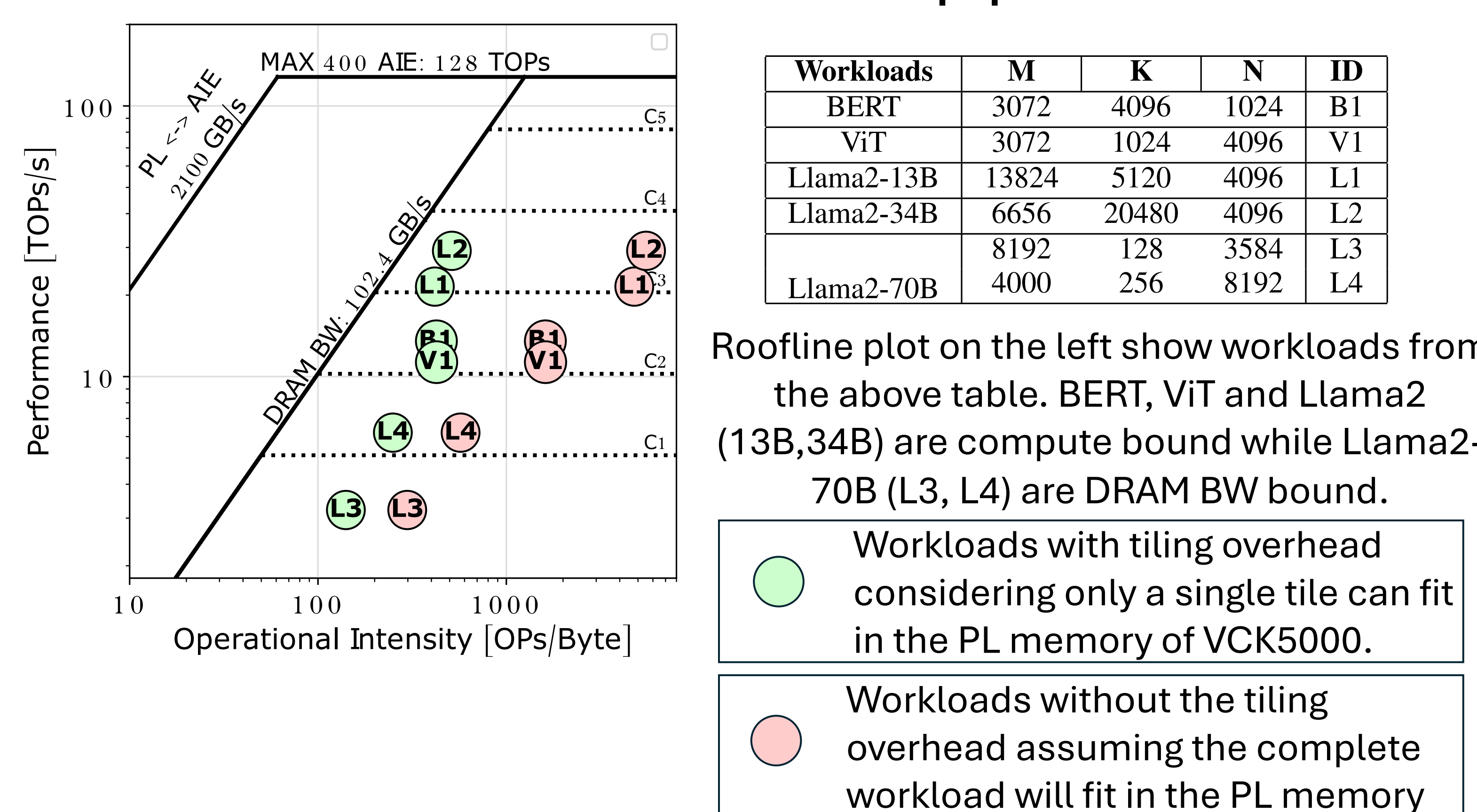
Impact of different communication schemes between AIEs on GEMM



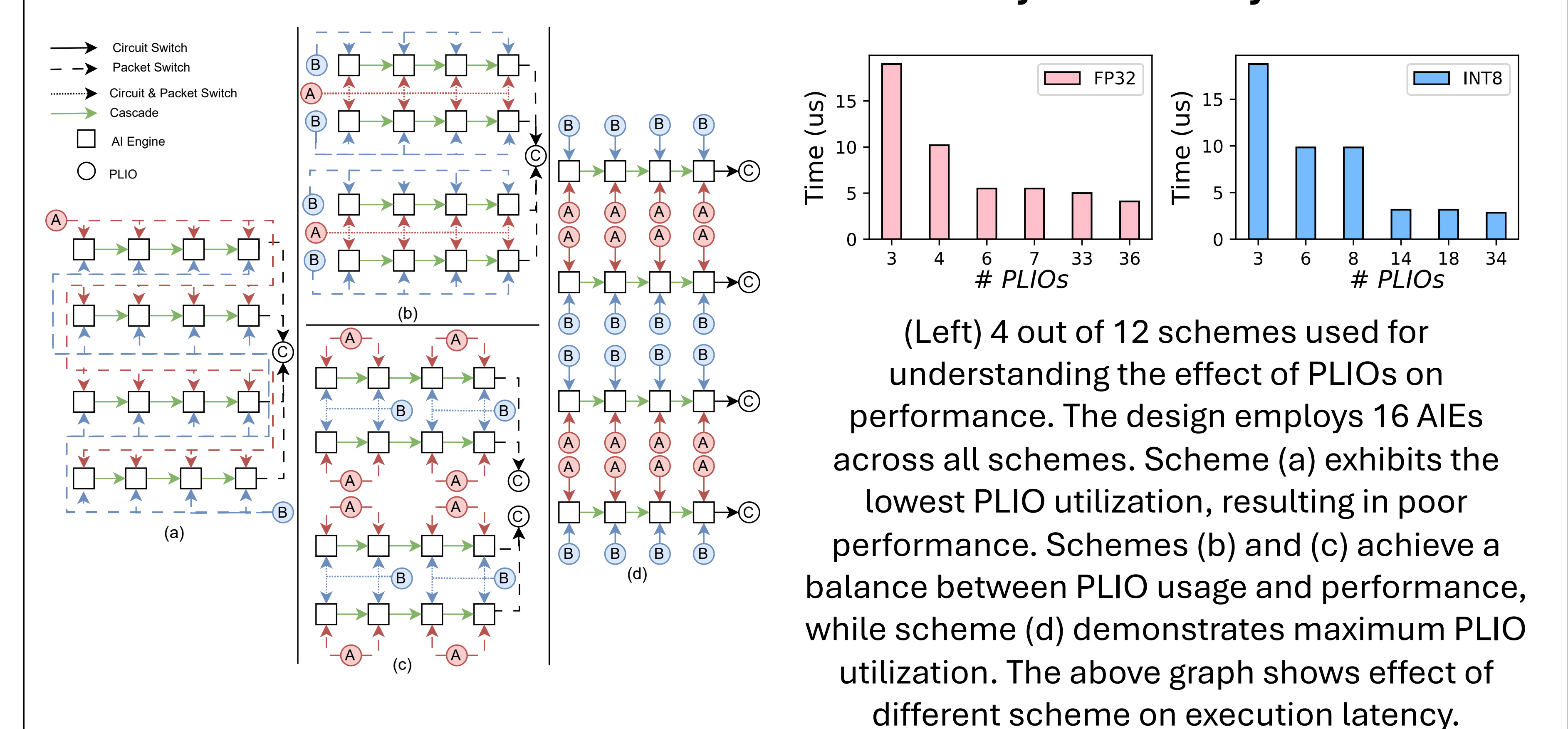
Execution breakdown on VCK5000 (compute vs. communication)



Performance for workloads from popular DNNs



Performance effect of PLIO on AIE array connectivity



Configuration	Precision	# AIEs	Native Size	# PLIOs
C1	FP32	16	32x128x128	7
C2	FP32	32	64x128x128	10
C3	FP32	64	128x128x128	20
C4	FP32	128	128x256x128	36
C5	FP32	256	256x128x256	64
C6	FP32	384	384x128x256	96
C7	INT8	16	128x256x128	14
C8	INT8	32	128x256x256	20
C9	INT8	64	256x256x256	40
C10	INT8	128	256x512x256	72
C11	INT8	256	256x512x512	112

Conclusion

- This study provides a thorough examination of AMD Versal's performance for GEMM workloads.
- The findings demonstrate the effect of architectural parameters on the performance informed by carefully designed experiments.
- We used SOTA implementations and their variations to perform this analysis.
- The analysis yields comprehensive set of guidelines to assist developers in designing more efficient designs.

Design Guidelines

- **AIE kernel size:** Choose kernels with highest efficiency (Good communication to compute balance)
- **AIE - AIE communication interface:** Cascade for low-latency streaming needs and buffer for non-streaming needs.
- **# of AIEs for a problem:** Always consider DRAM bandwidth and AIE to PL interface bandwidth when designing.
- **PLIO usage:** Sharing PLIO using packet switching and circuit switching can save PLIO resources.

References

- [1] J. Zhuang et. al. "CHARM: Composing Heterogeneous Accelerators for Matrix Multiply on Versal ACAP Architecture,"
- [2] AMD/Xilinx, "Versal ACAP AI Engine Architecture Manual (AM009)," 2021.



Advent Lab,
ASU



Kaustubh Mhatre,
kmhatre@asu.edu