



AI4HW and HW4AI: The PLENA Ecosystem Journey

HAORAN WU¹, CAN XIAO², JIAYI NIE¹, XUAN GUO², BINGLEI LOU², JEFFREY T.H. WONG²,
ZHIWEN MO², CHENG ZHANG², PRZEMYSŁAW FORYS², CHENGYANG AI³, TIMI ADENIRAN¹, WAYNE LUK²,
HONGXIANG FAN², JIANYI CHENG³, TIMOTHY M. JONES¹, RIKA ANTONOVA¹, ROBERT MULLINS¹, AARON ZHAO²

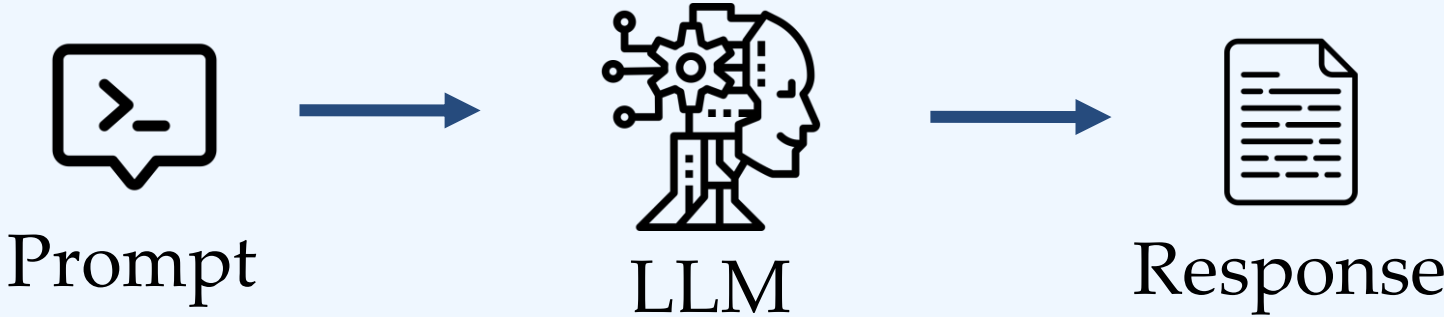


SEP 2026

WILLIAM GATES BUILDING
15 JJ THOMSON AVE
CAMBRIDGE, UK

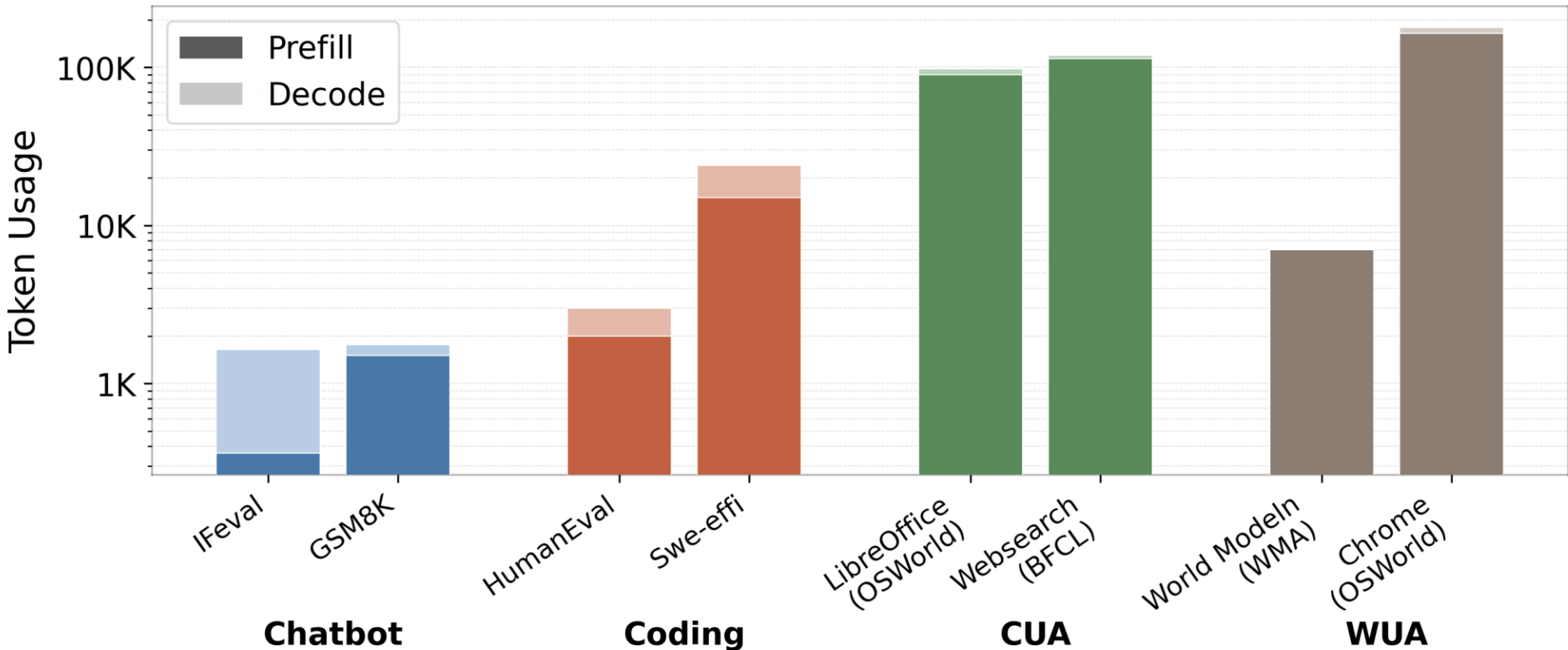
Towards the Agentic Era

Traditional Workloads

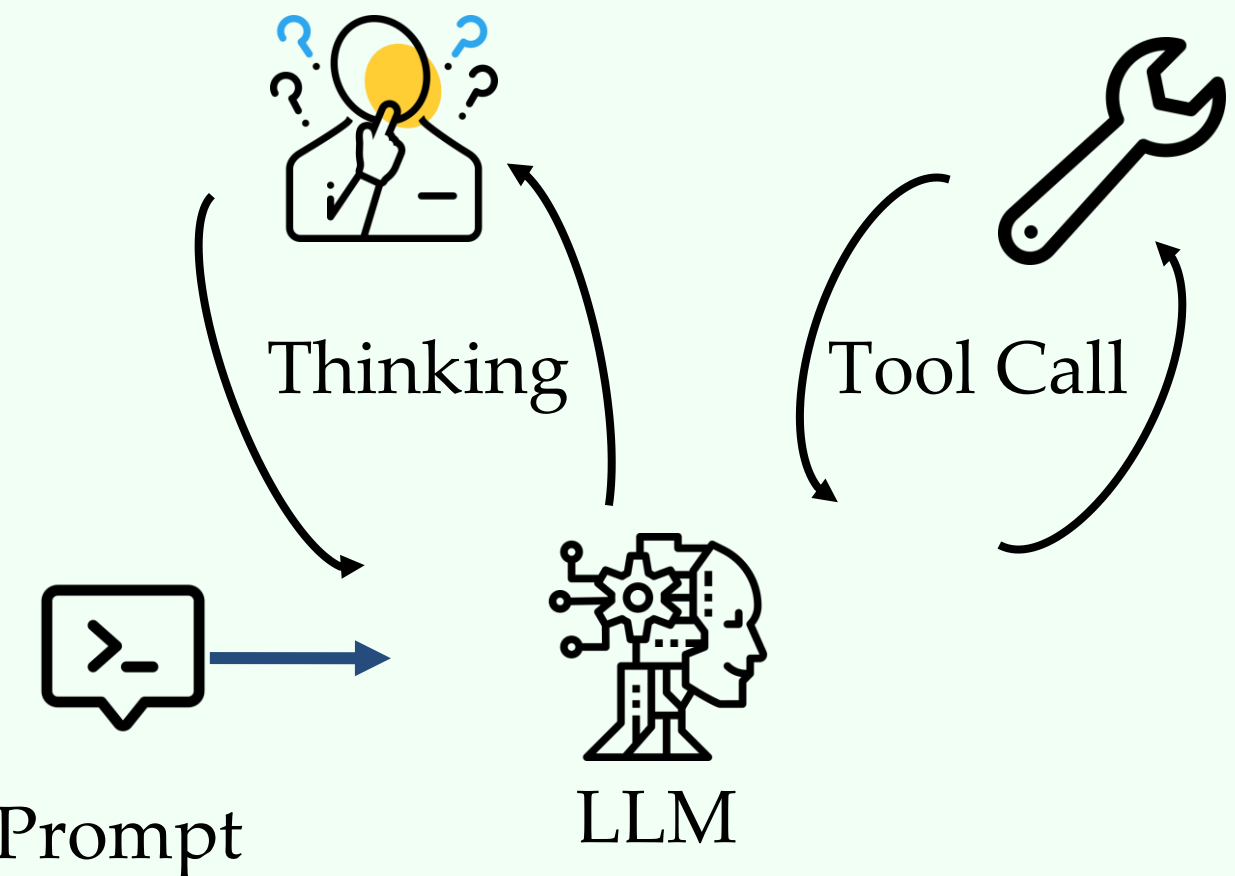


General ChatBot QA, Translation etc.

- Agentic workloads have significantly higher token usage

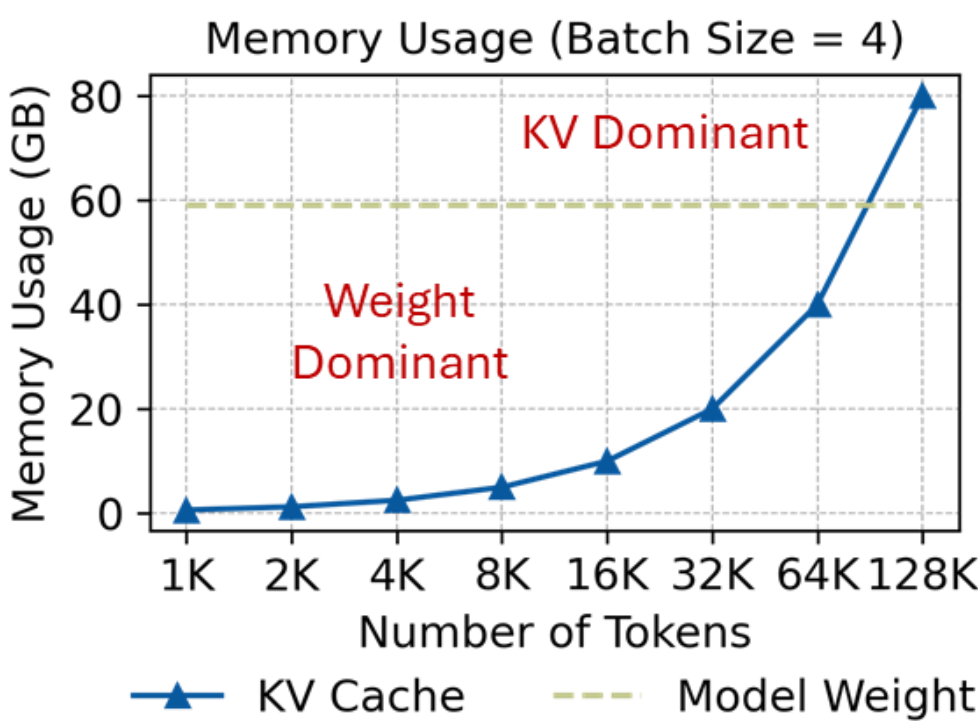
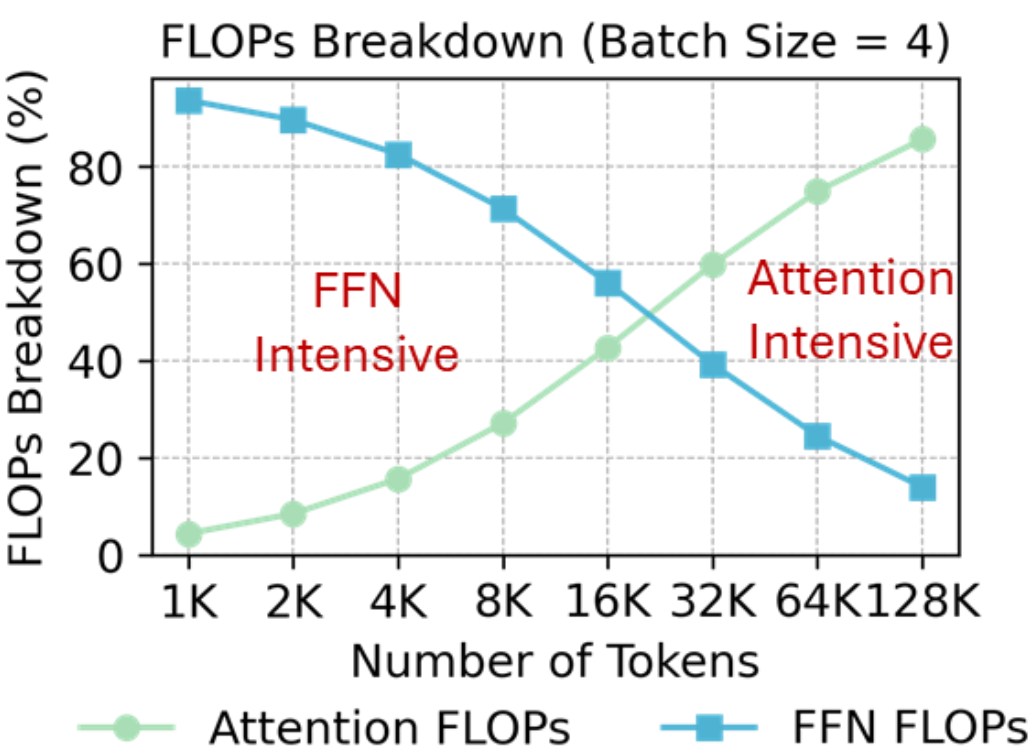


Agentic Workloads



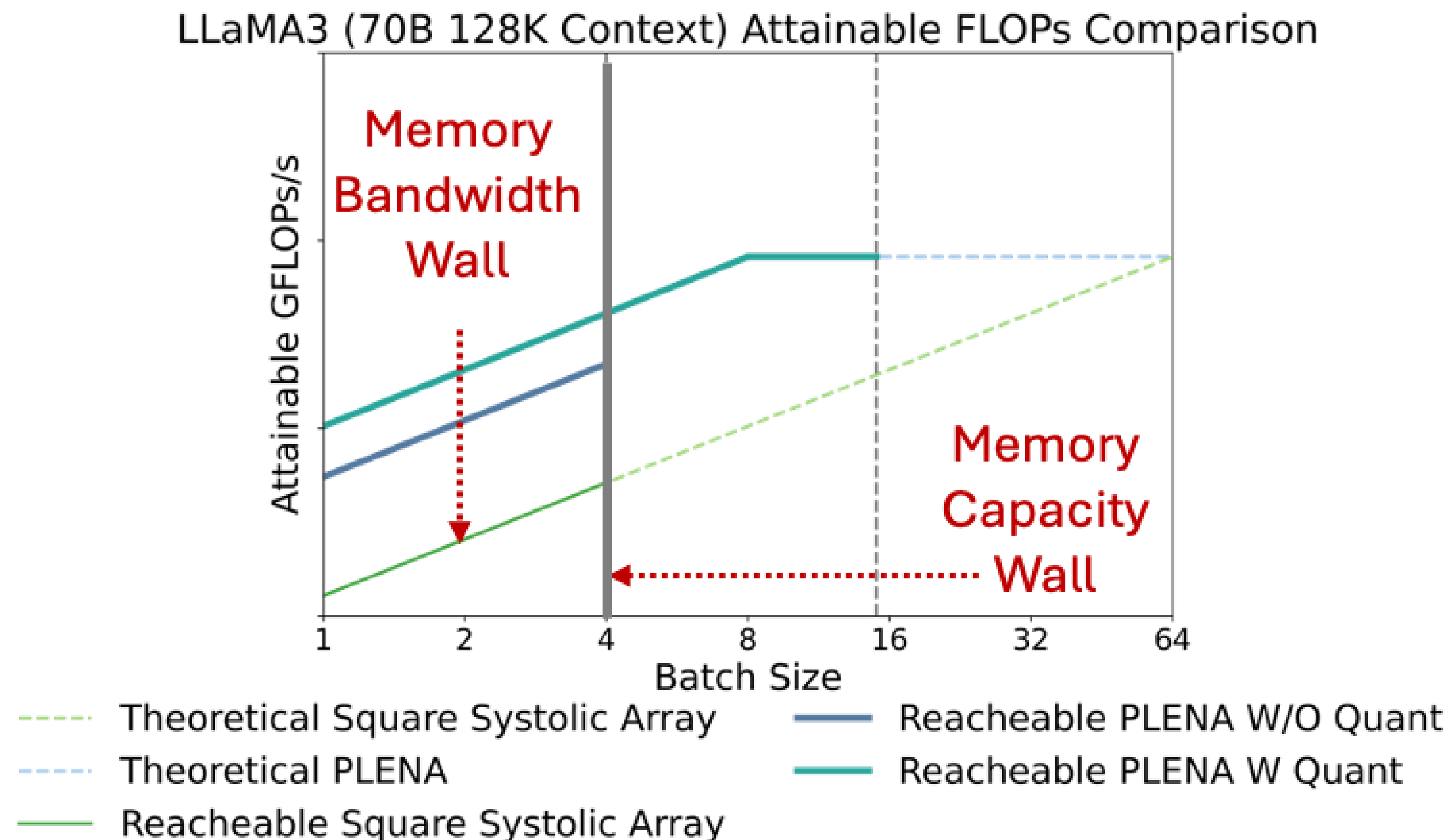
Coding Agent, Computer Use Agent etc.

- Agentic workloads fundamentally reshape the **compute** and **memory** characteristics of LLM inference



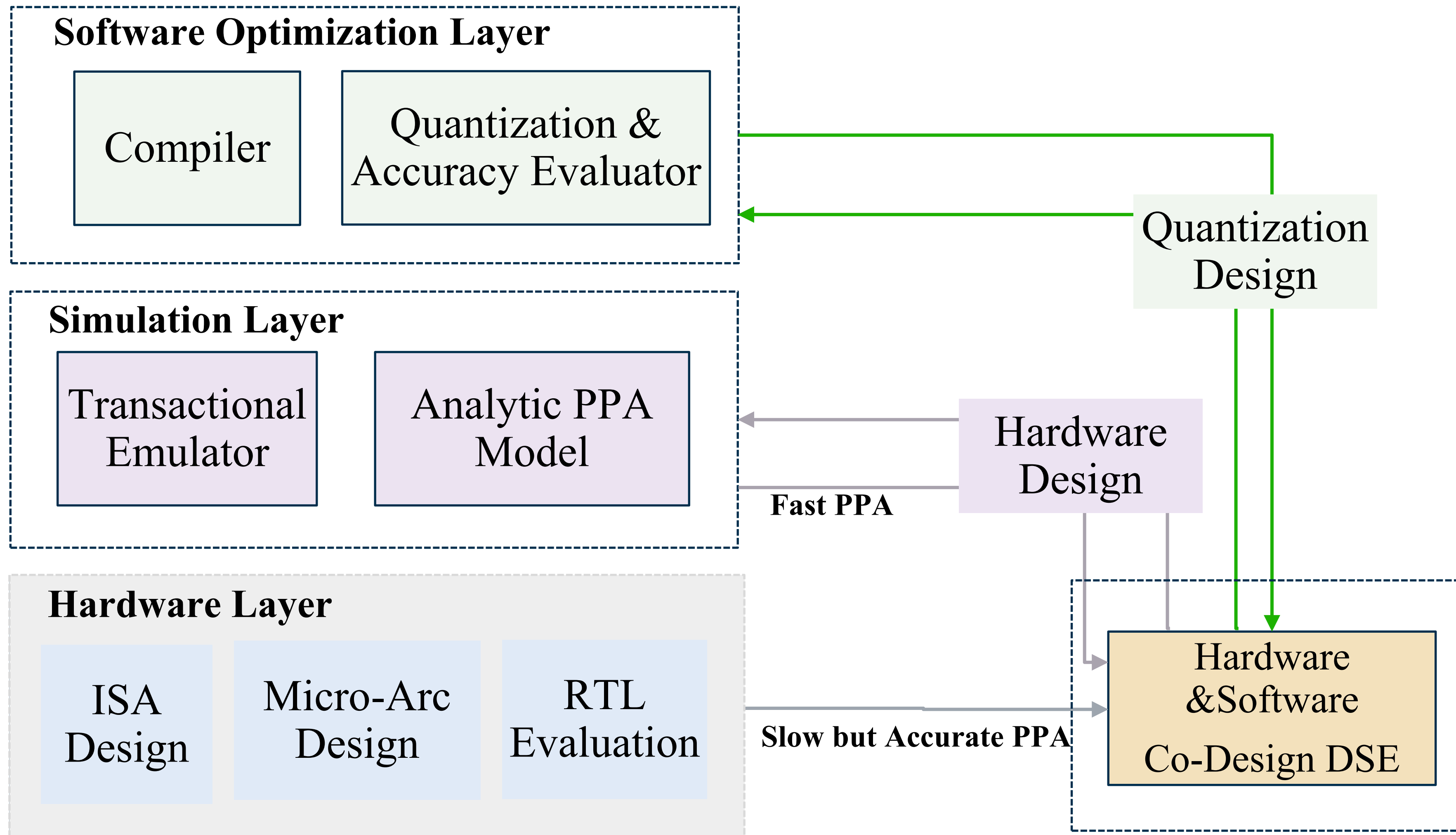
Memory Walls in Agentic Workloads

- **Bandwidth Wall:** Off-chip bandwidth limits on-chip compute.
 - **Capacity Wall:** Off-chip capacity limits the KV-cache batch size.
- Limits** compute utilization in traditional square-shaped systolic arrays.



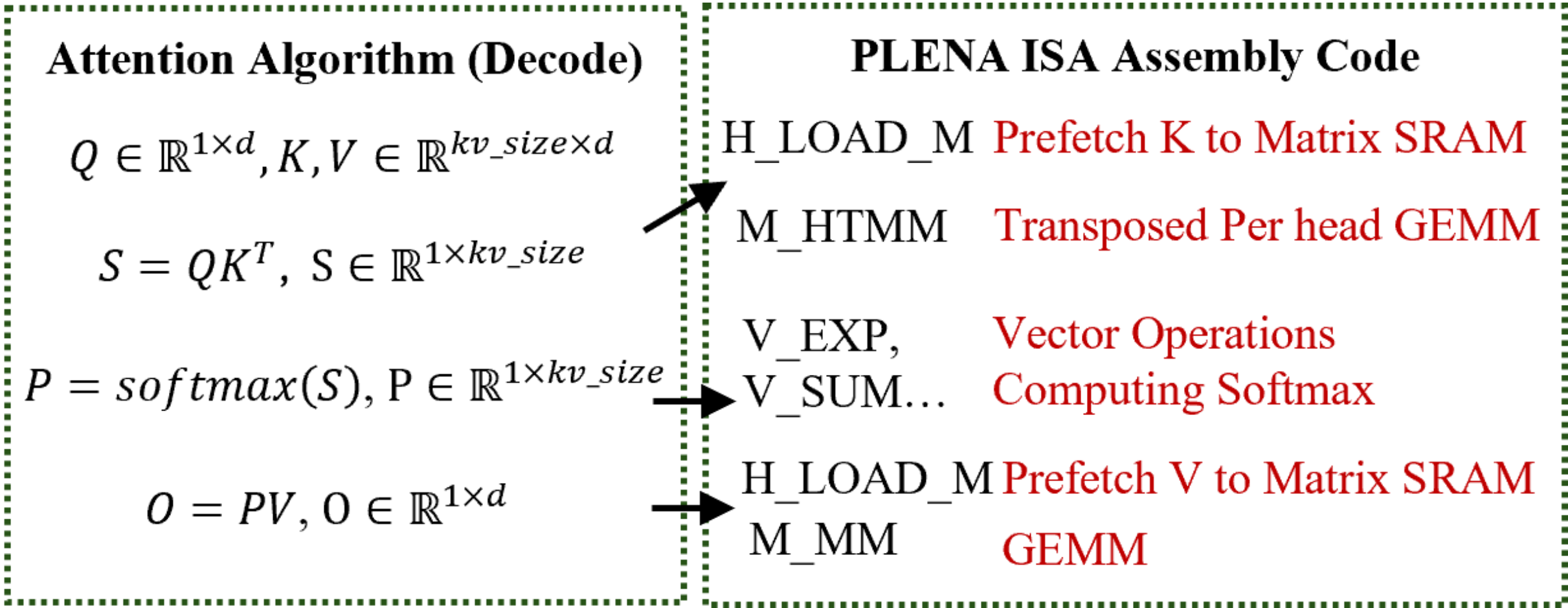
To solve these problems, we propose **PLENA** — an *open-source, hardware–software co-designed* NPU for agentic inference, spanning the *full stack* from model to hardware.

We open-source the whole stack – Model to Hardware

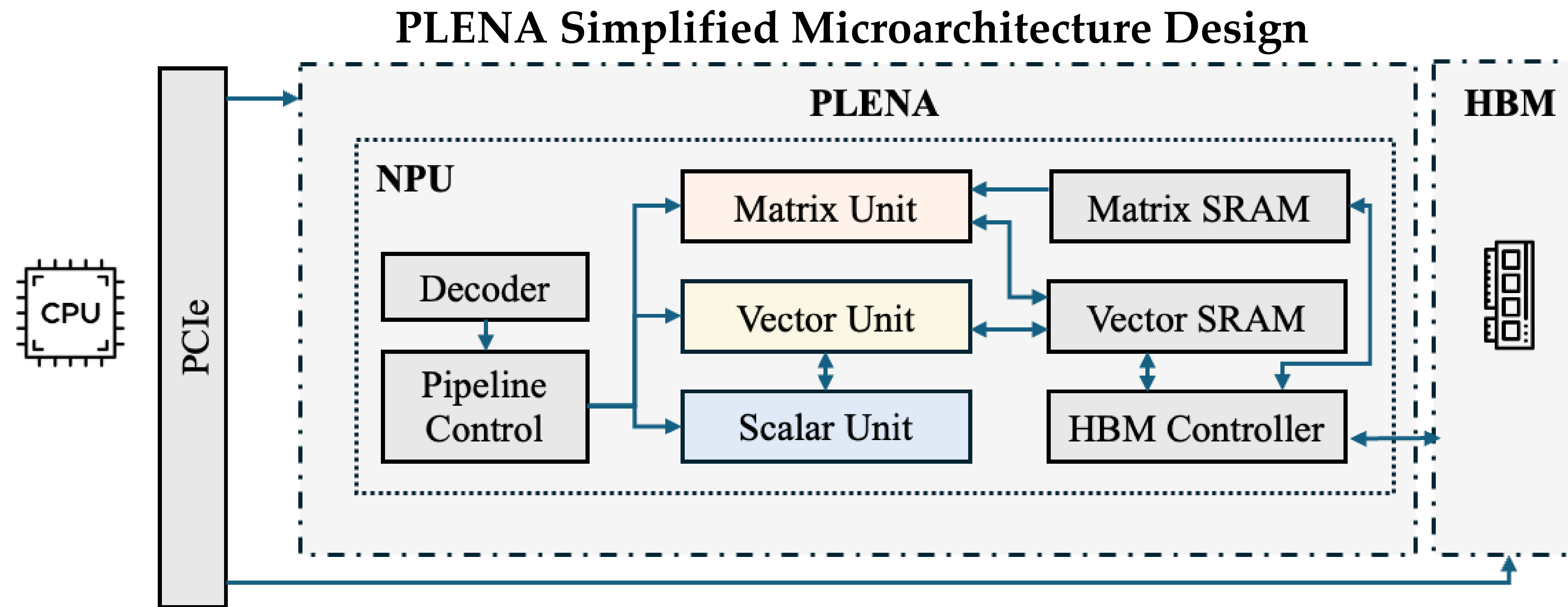


PLENA ISA Design

Types	Descriptions	# Instr.
Matrix(M)	Controls GEMM and GEMV operations, with or without matrix transposition	12
Vector(V)	Performs elementwise and reduction operations, and rotation for quantization	12
Scalar(S)	Performs scalar INT and FP arithmetic	17
HBM(H)	Handles data transfers between HBM and the Matrix/Vector SRAMs	3
Control(C)	Defines operation settings such as the HBM address, nested-loop configuration, and other execution parameters	8



The Optimization Pathways



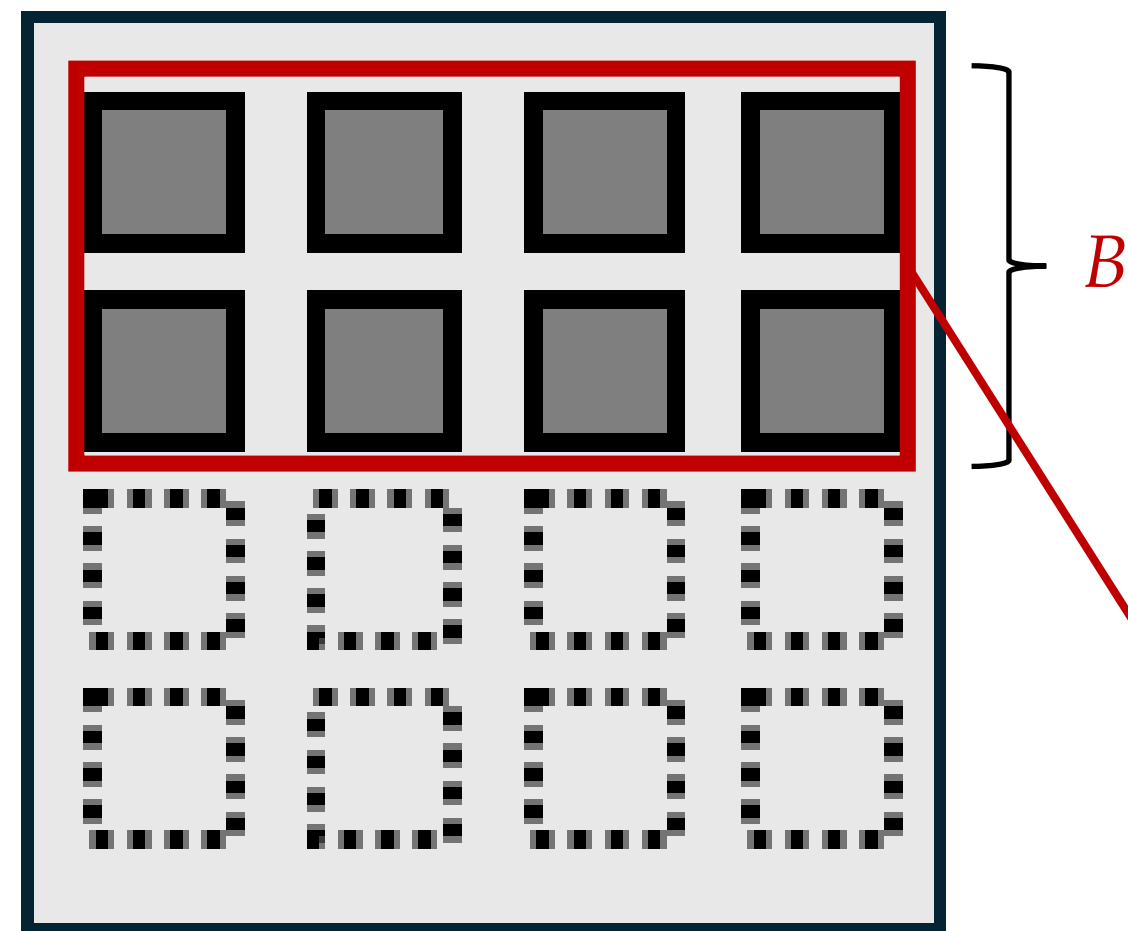
Optimization Pathways:

1. Flattened Systolic Array
2. Mixed Precision Supported
3. Supporting FlashAttention

Pathway 1: Flattened Systolic Array

Traditional Systolic Array No Longer Suitable

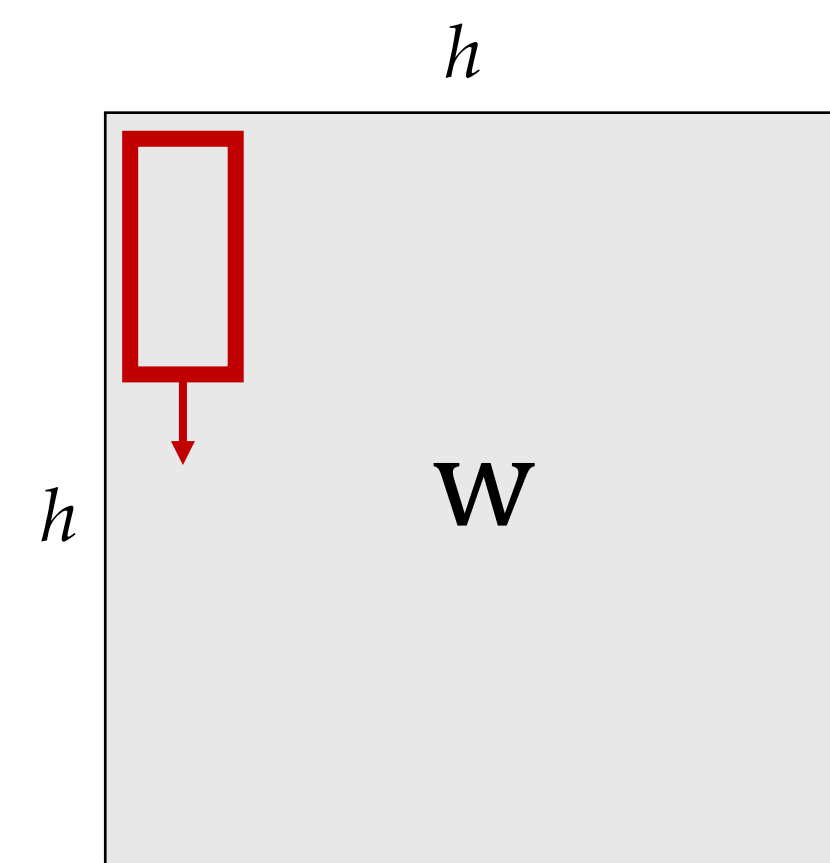
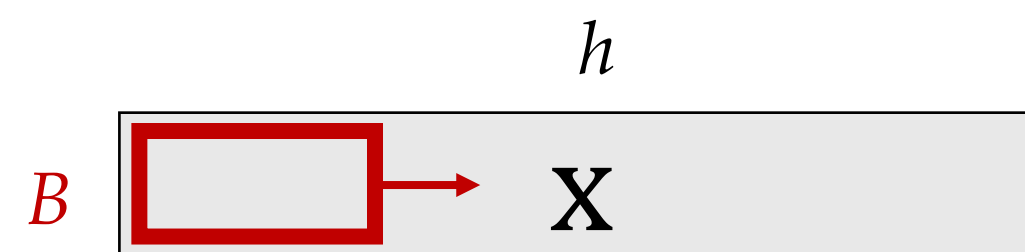
Square Shaped Systolic Array



Limited Batch Size B due to capacity wall

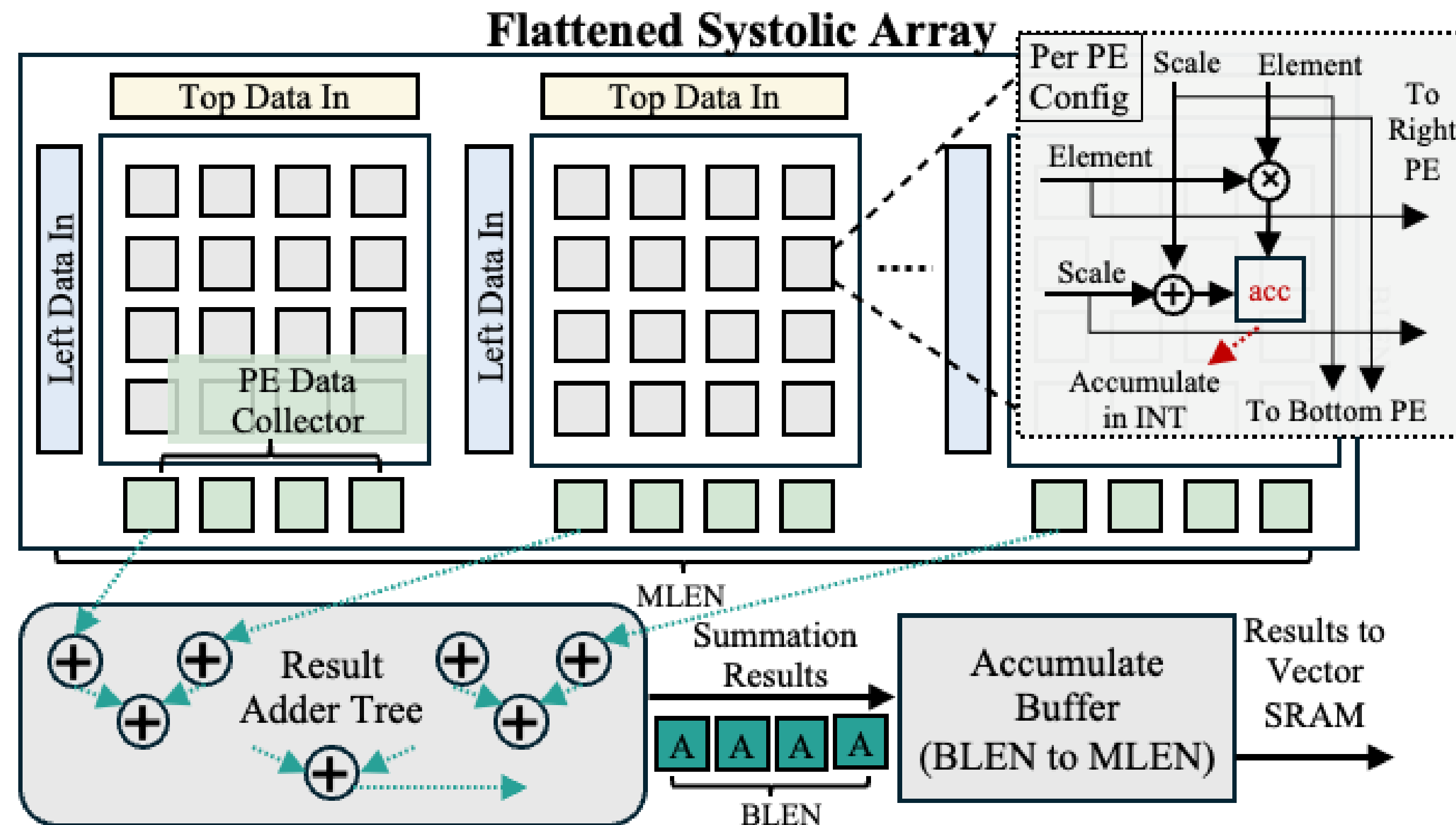
In Projection:

$$y = xw$$

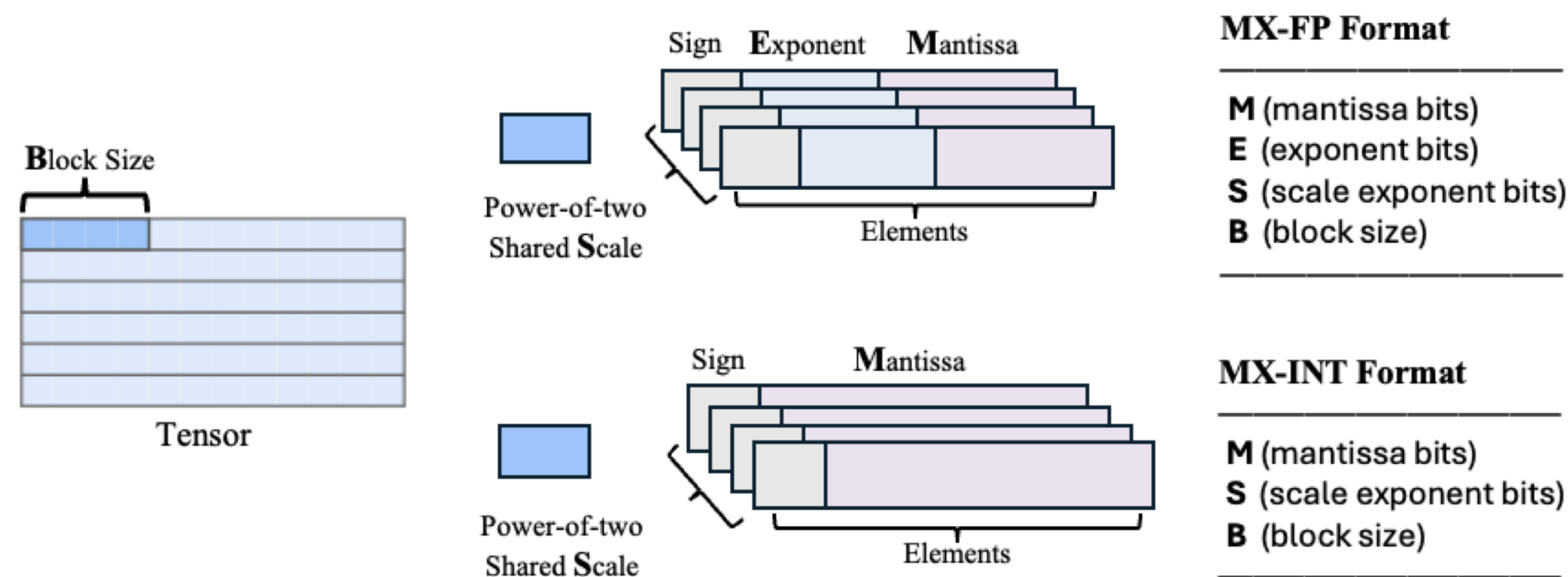
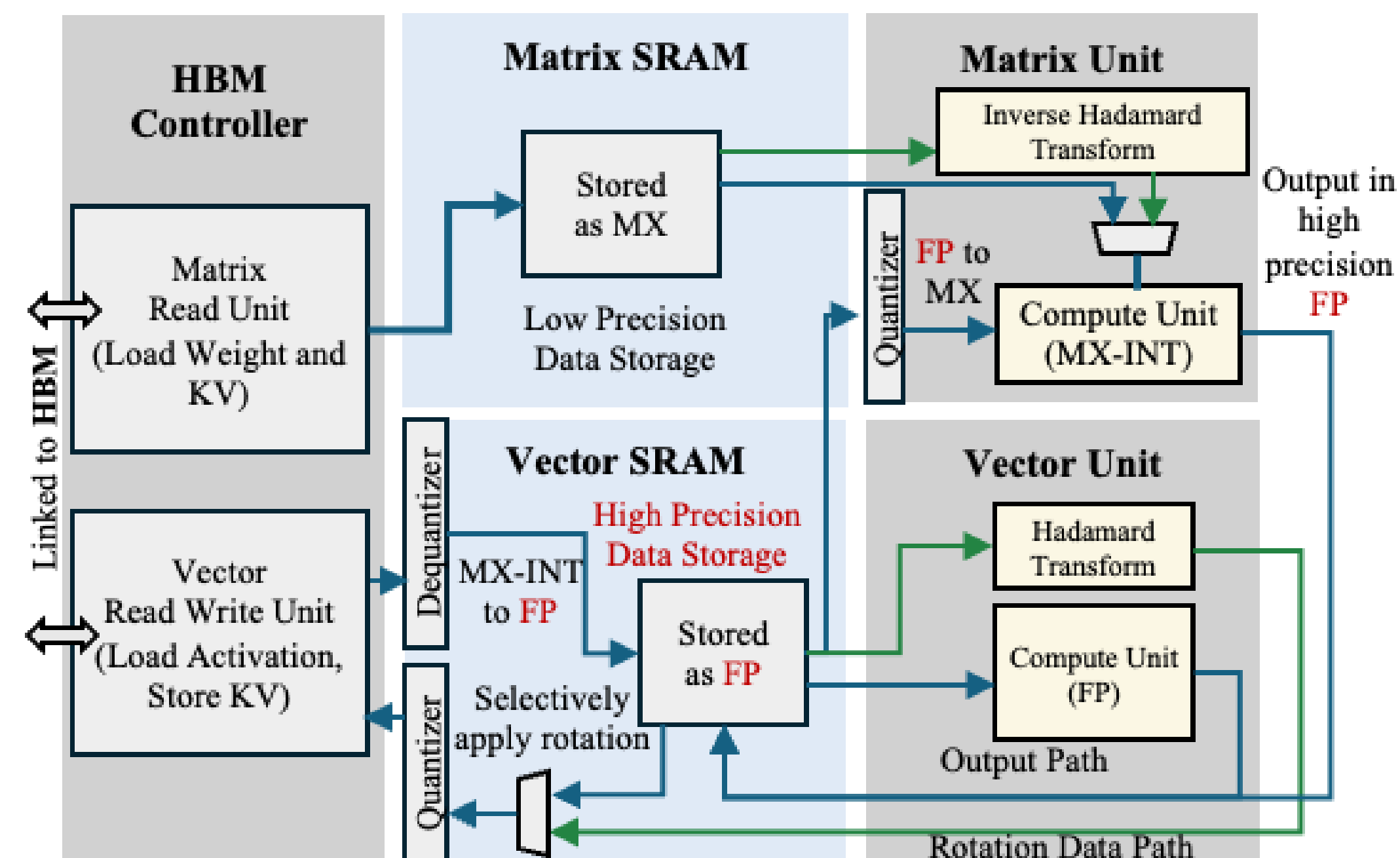


Low Systolic Array Utilization

Pathway 1: Flattened Systolic Array



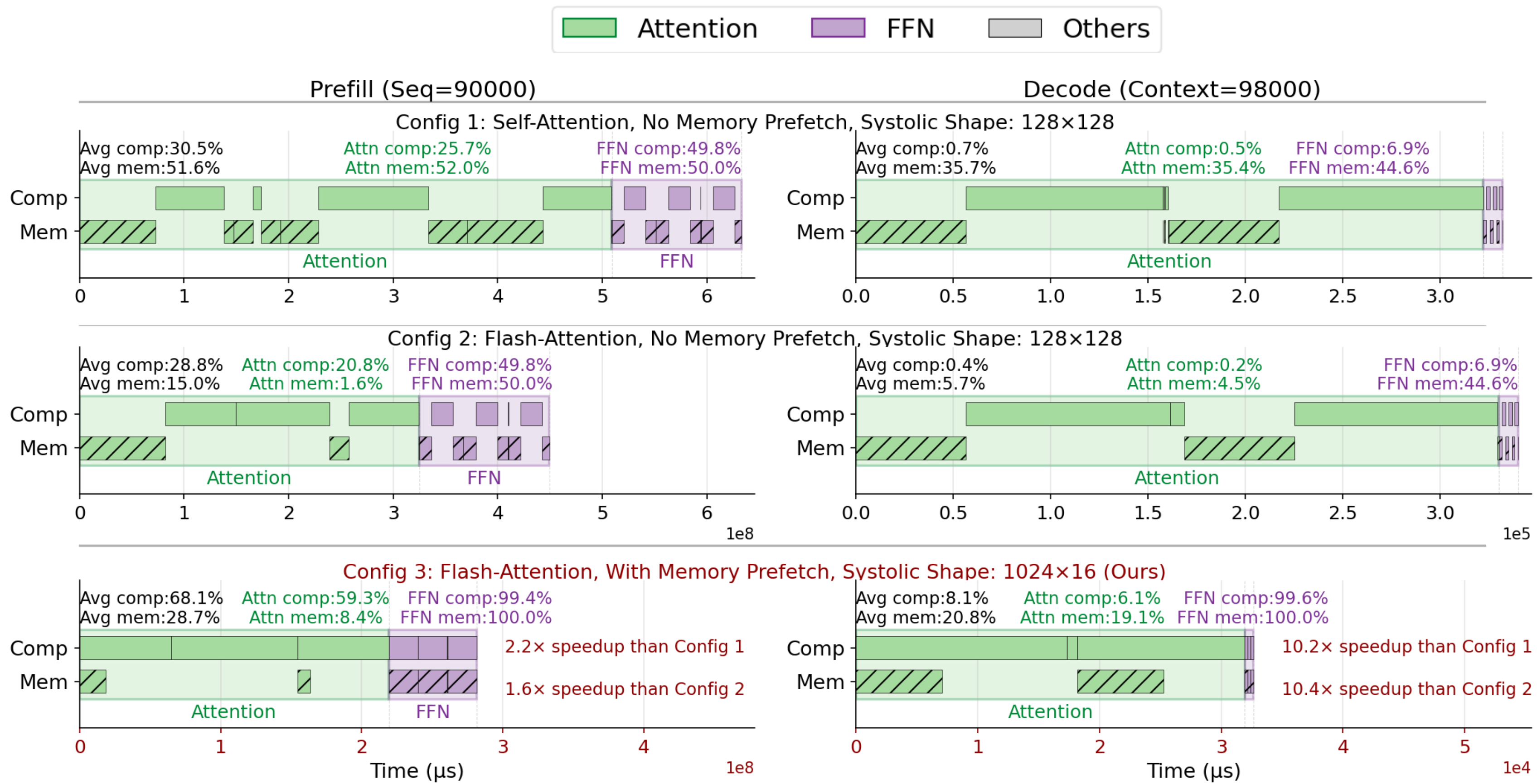
Pathway 2: Mixed Precision Supported



Quantization Configurations: definition of micro-scaling arithmetic data format

Pathway 3: Supporting FlashAttention

- Reduce the memory traffic
- Overlapping the KV cache read / write process to the HBM
- Matrix SRAM that could support low-cost transposed matrix read to support QK^T



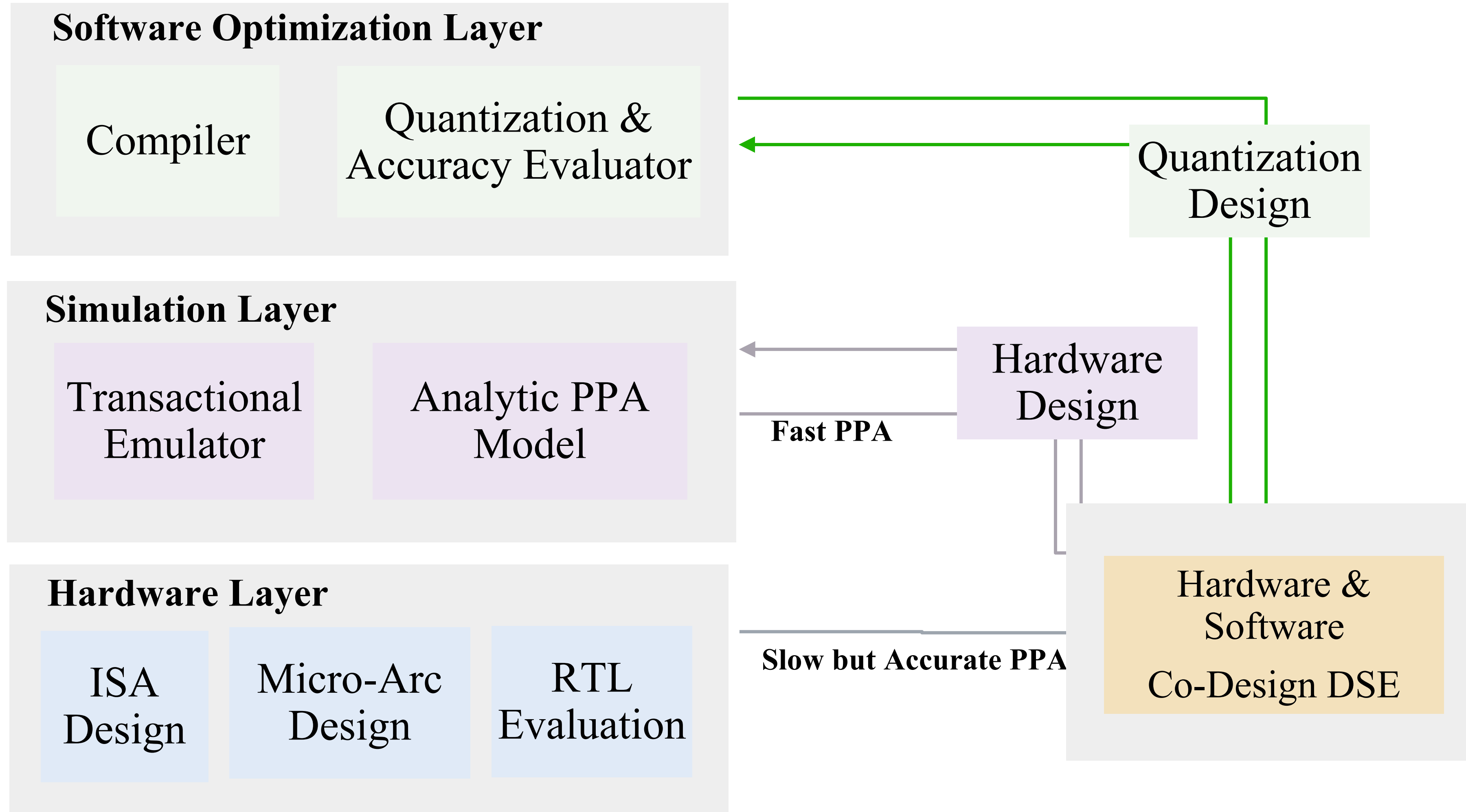
Self-Attention,
No Memory Prefetch

Flash-Attention,
No Memory Prefetch

Flash-Attention,
With Memory Prefetch

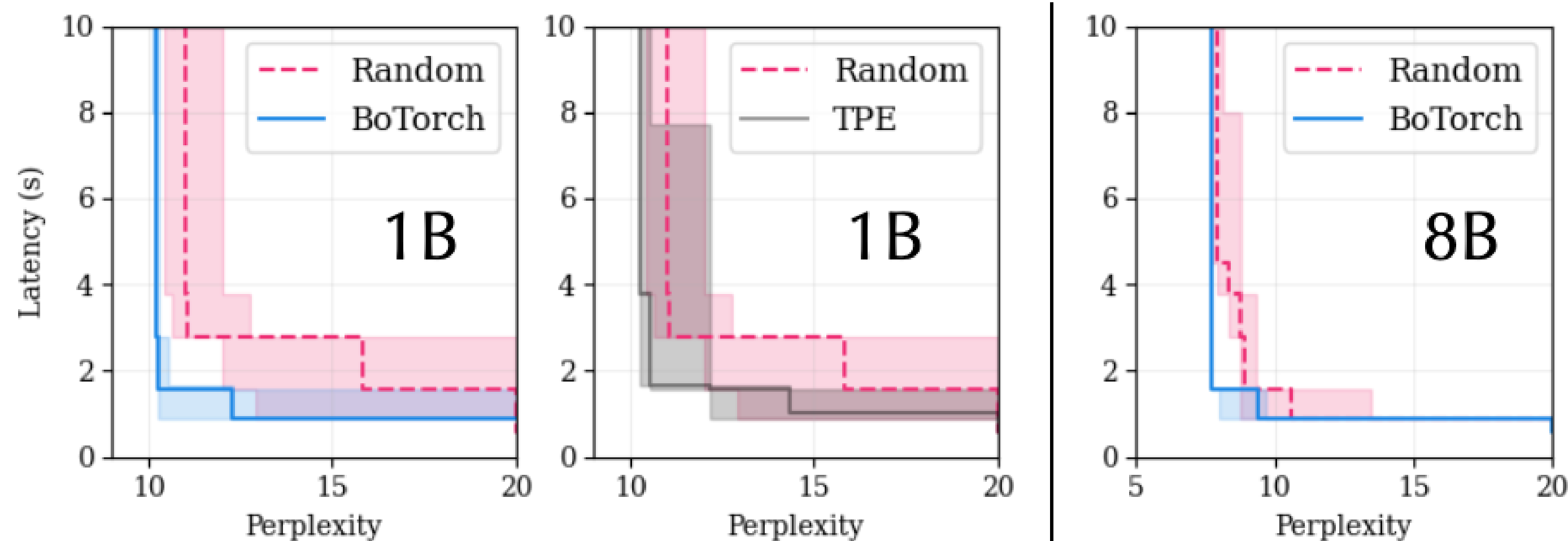


We open-source the whole stack – Model to Hardware



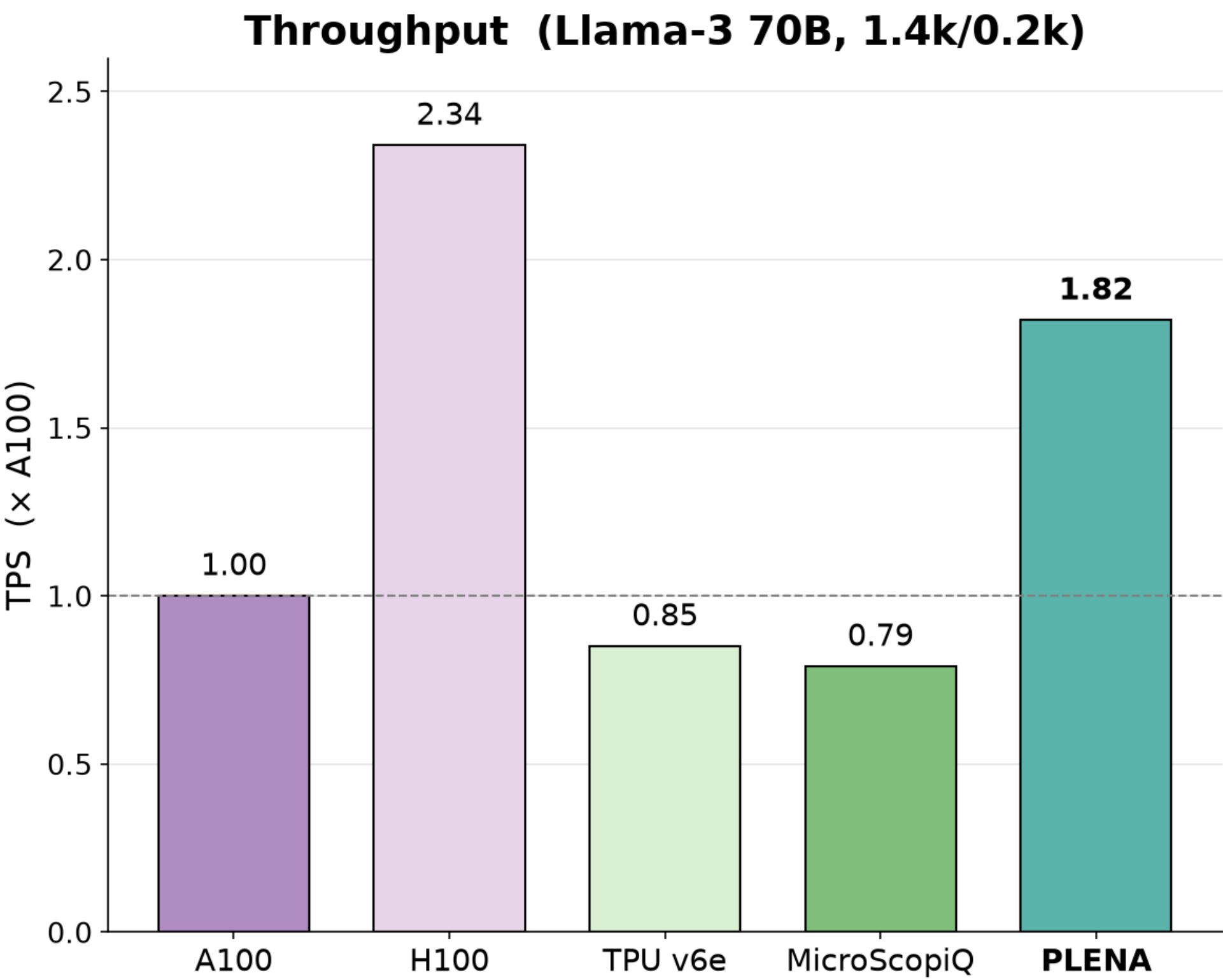
Balancing Area, Latency and Perplexity via Co-design

Multi-objective optimization to jointly optimize workload accuracy (perplexity↓), latency(↓) and area(↓).

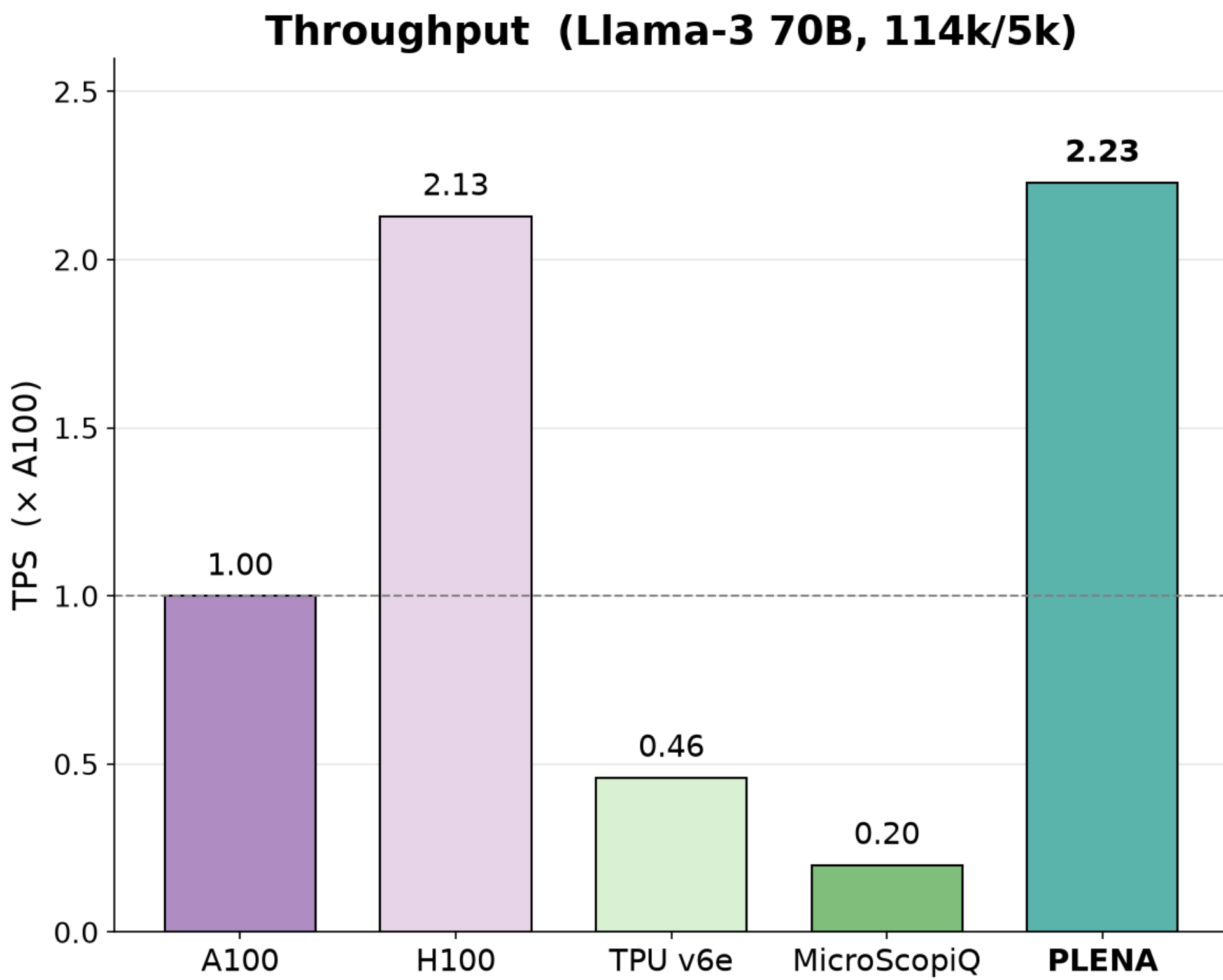


PLENA delivers **higher throughput** than leading accelerators / GPU devices with the same amount of multipliers.

Chatbot Workloads



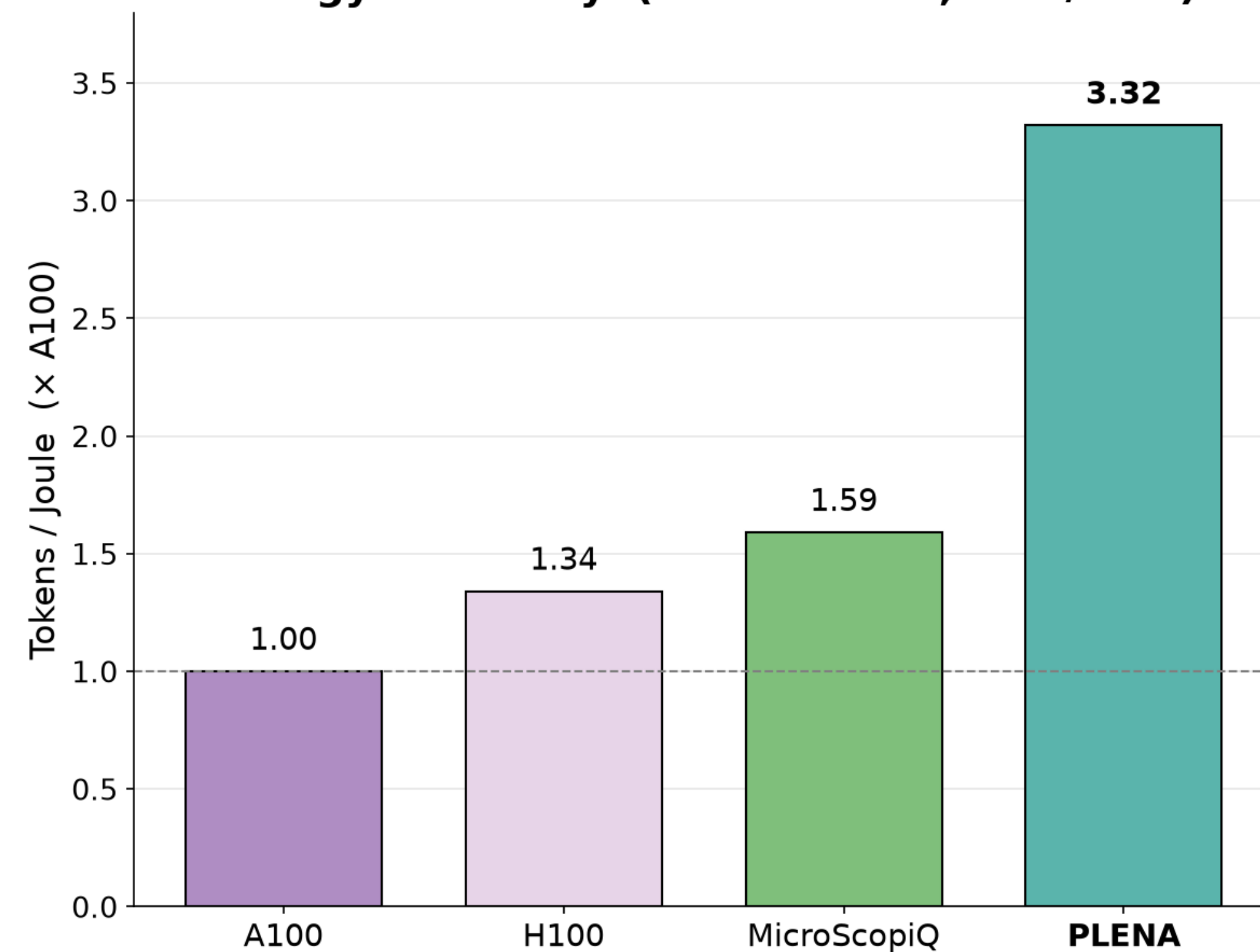
Agentic Workloads



PLENA delivers the **highest energy efficiency** than leading accelerators / GPU devices

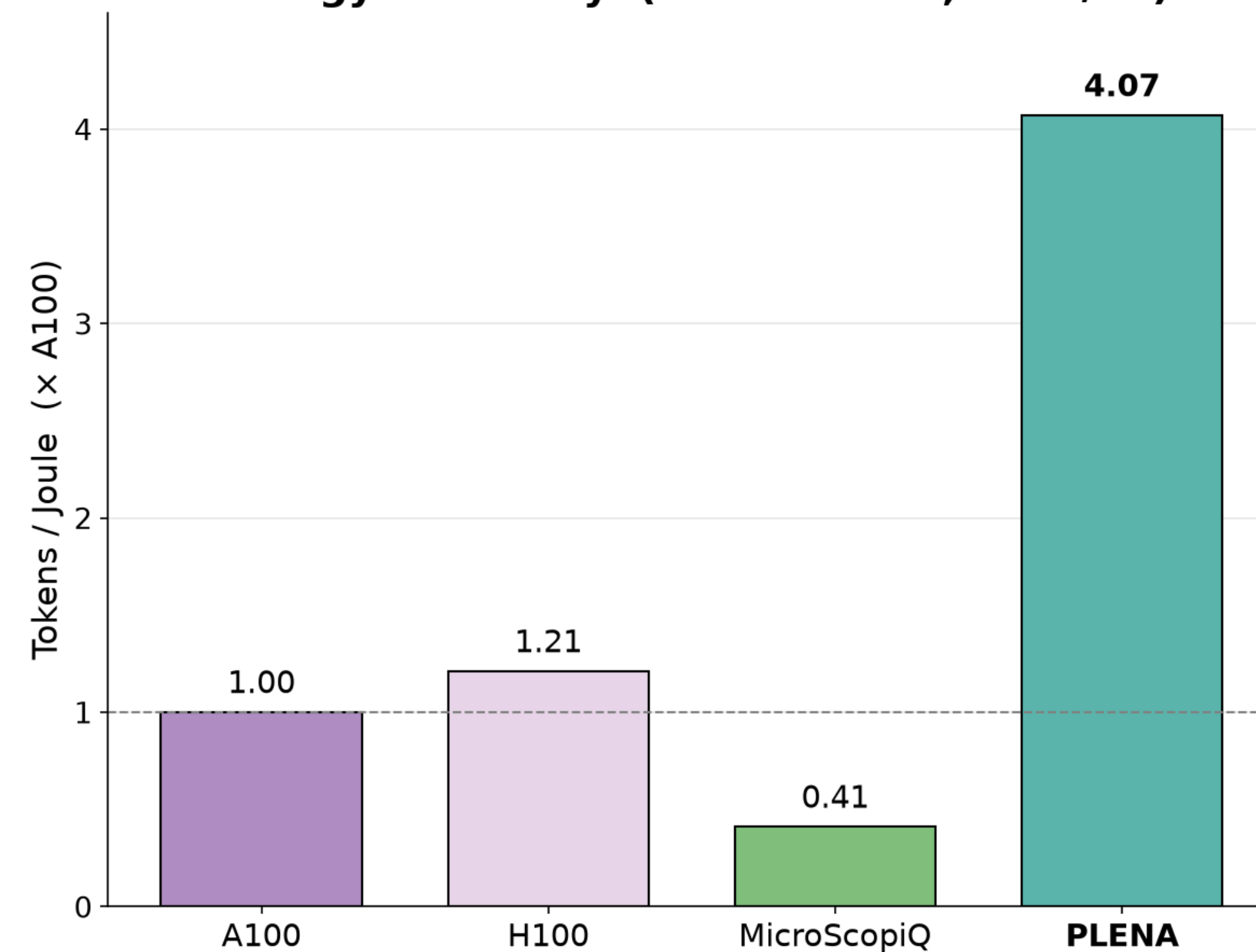
Chatbot Workloads

Energy Efficiency (Llama-3 70B, 1.4k/0.2k)

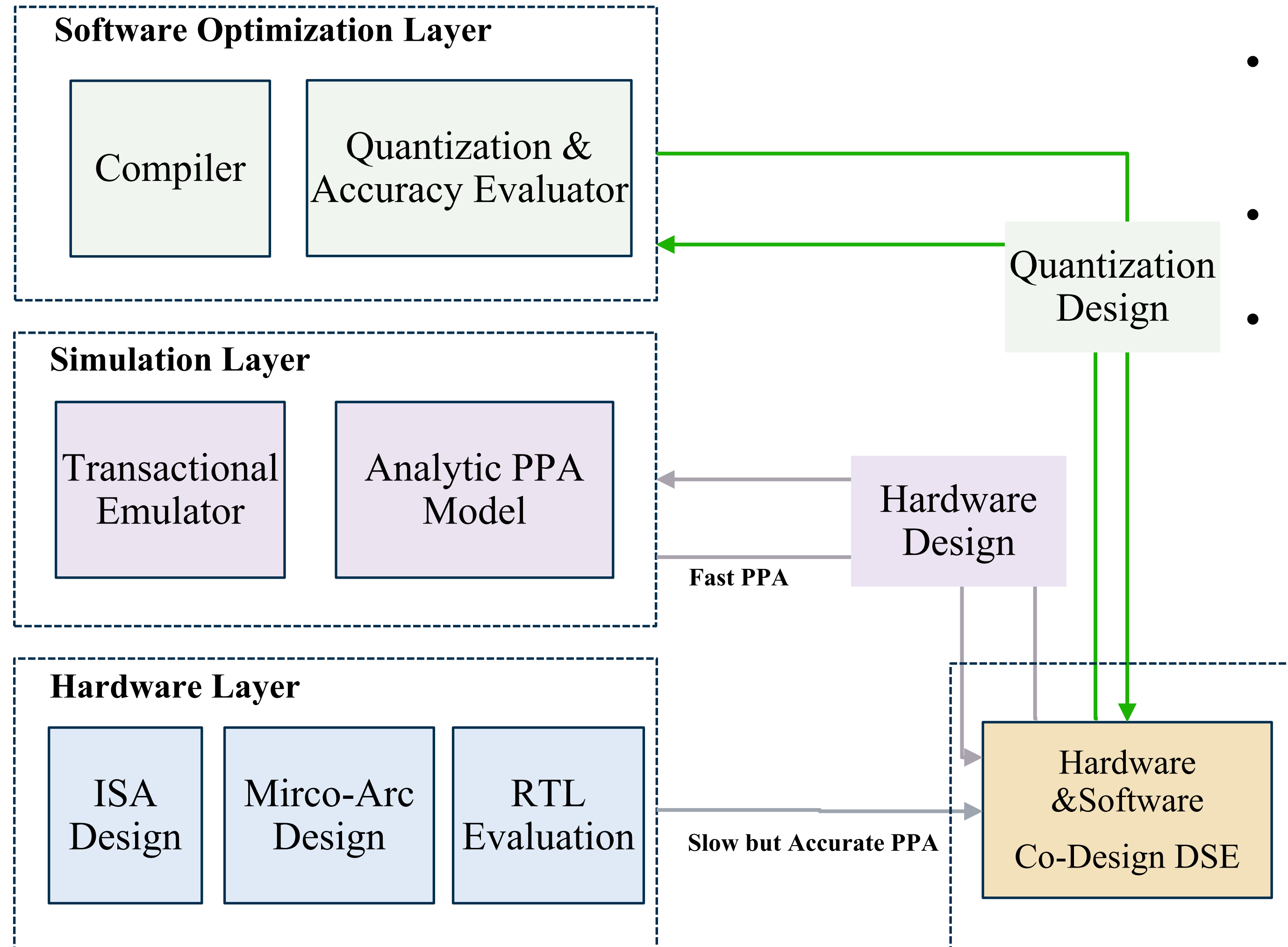


Agentic Workloads

Energy Efficiency (Llama-3 70B, 114k/5k)



PLENA Overview



- **Compiler:** Custom ISA and Aten based PyTorch-to-PLENA Compiler
- **Simulators:** Multi-Fidelity Simulations
- **RTL:** Ready for ASIC EDA tool flow



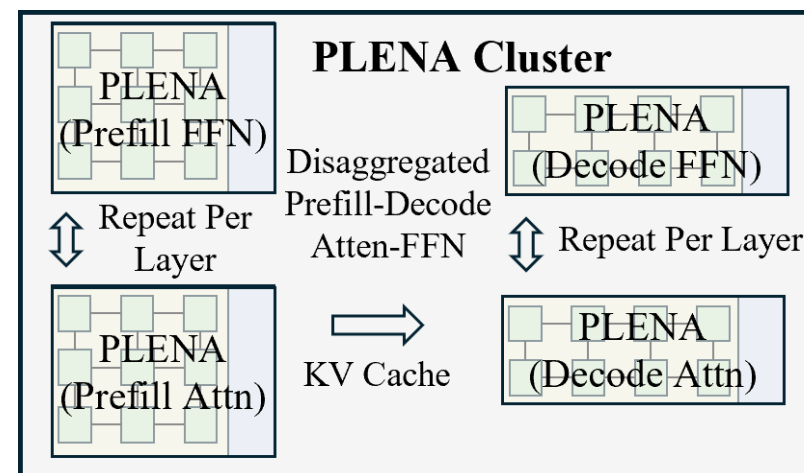
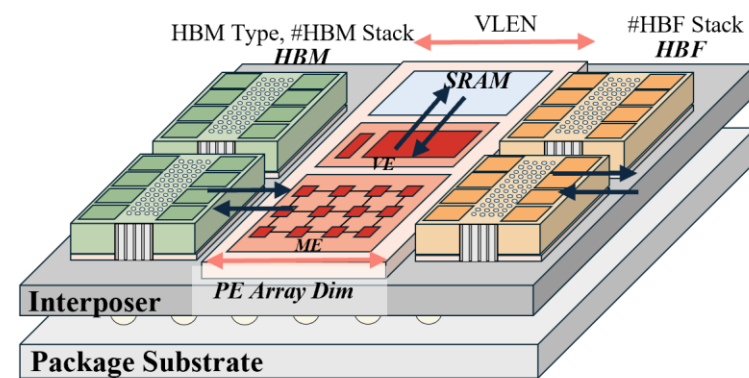
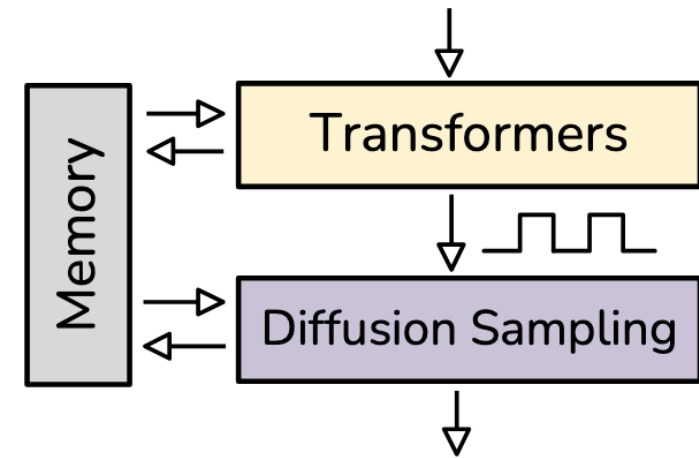
PLENA artifacts can be found at:

https://aicrosssim.github.io/PLENA_Doc/

PLENA Published at **ISCA 2026**



Beyond PLENA, What We've Built Next



KernelCraft: First benchmark for agentic LLM generation of close-to-metal assembly kernels on PLENA and other emerging accelerators. (presenting today)

DART: Diffusion NPU Accelerator for Refinement-based Text Generation

MemExplorer: A new memory system synthesizer for heterogeneous NPU systems

HeteroPanacea: Simulating future Heterogeneous AI systems (eg. GPU+NPU); with optimization like prefill-decode and attention-FFN disaggregations.

