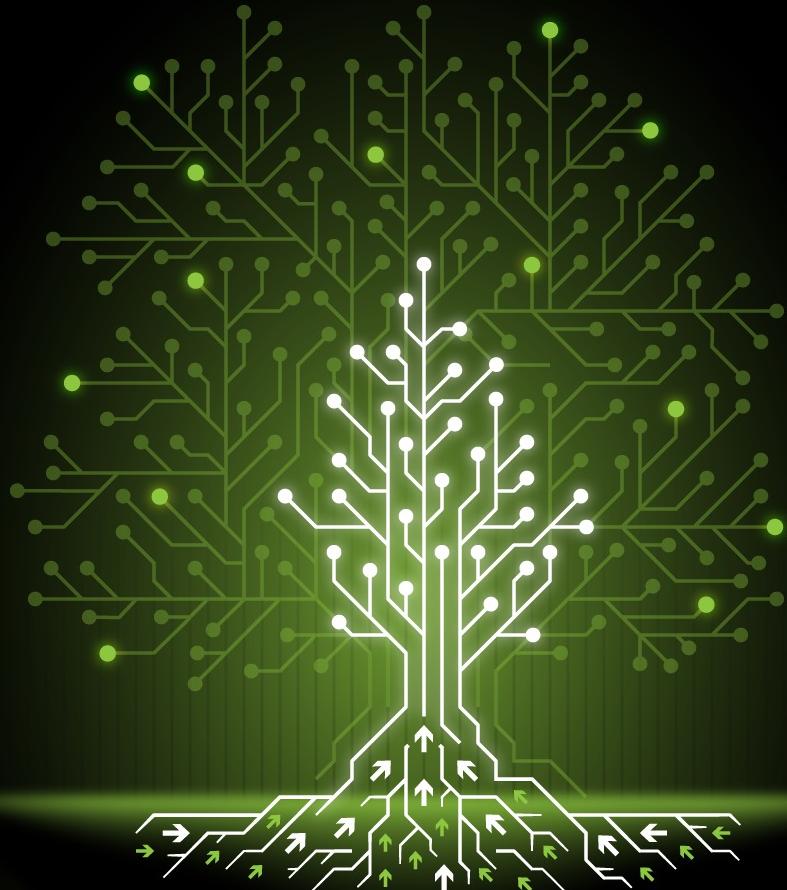


OCP

FUTURE TECHNOLOGIES SYMPOSIUM

OCP Global Summit
October 18, 2023 | San Jose, CA



Architectural Challenges and Innovation for Compute Infrastructure Co-Design

Peipei Zhou

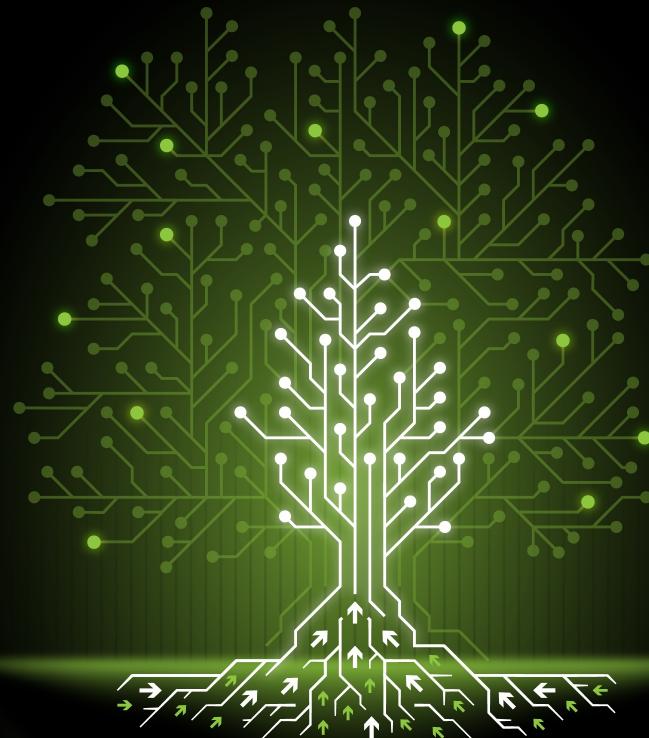
Assistant Professor, University of Pittsburgh



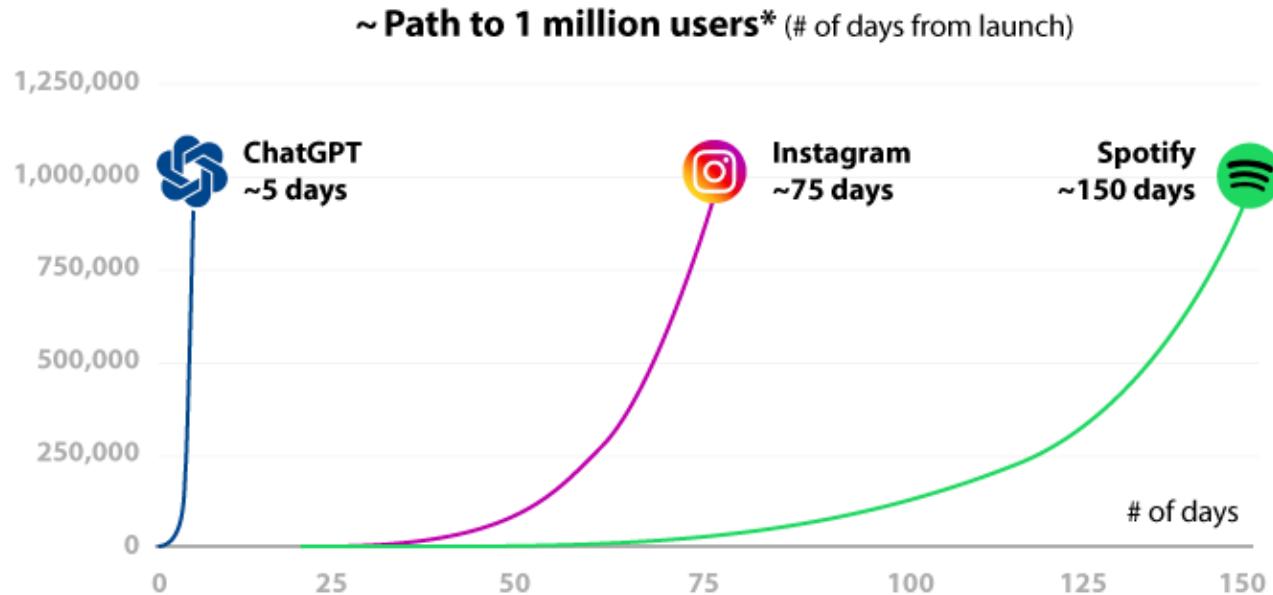
OCP
GLOBAL
SUMMIT

OCTOBER 17-19, 2023
SAN JOSE, CA

Scaling Innovation Through Collaboration



Generative AI Models: ChatGPT

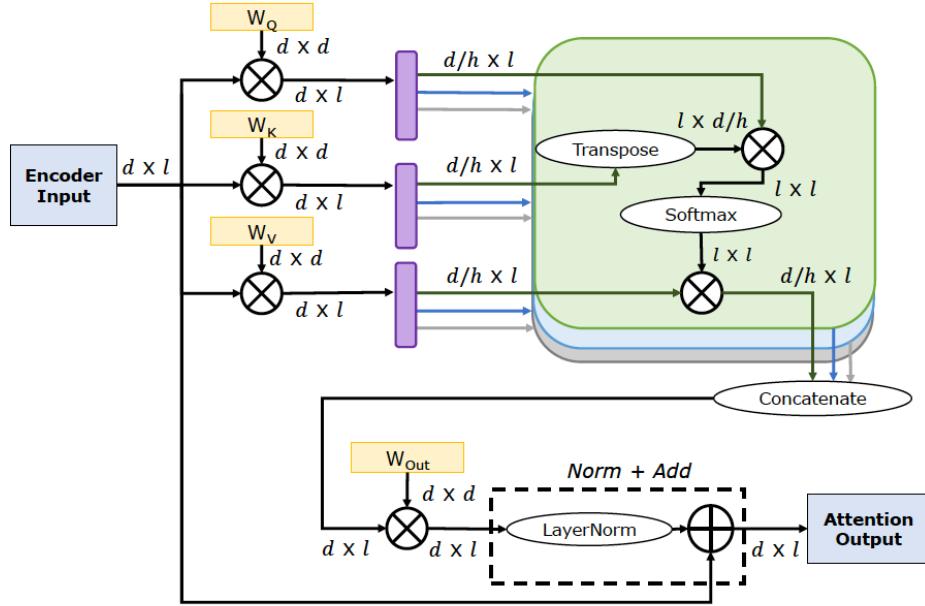


Sources: Google, Subredditstats, Media Reports

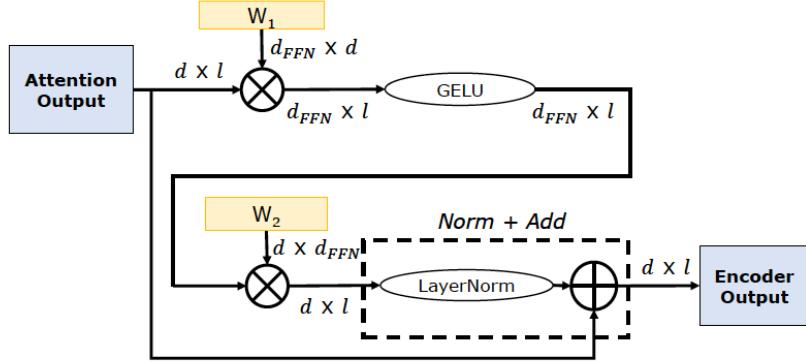
Generative AI Models: Stable Diffusion, Dall-E



Transformer Models



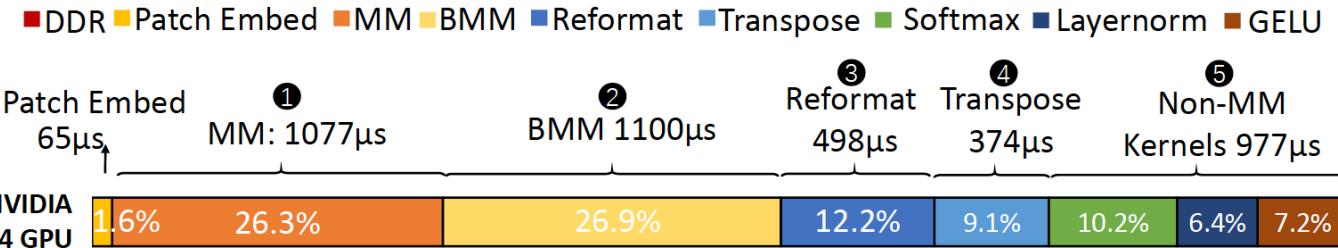
Muti-Head Attention (MHA) Module



Feed-Forward Network (FFN) Module

Kernel Breakdown

- Profiling Transformer based model, DeiT-T, on Nvidia GPU T4 (TSMC 12 nm)

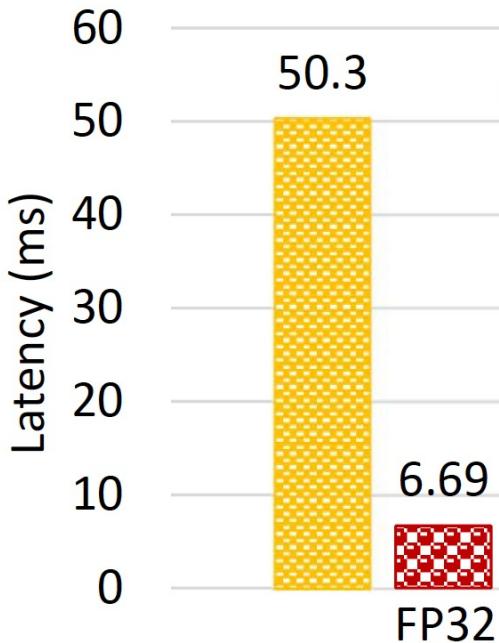


- ① Low Tensor Cores utilization for INT8 MM kernels.
- ② TensorRT adopts an implicit quantization policy, which leads to BMM computing in FP32, which could originally be in INT8.
- ③ The quan/dequan between FP32 and INT8 consumes non-negligible GPU cycles
- ④ The data layout change also consumes nonnegligible GPU cycles
- ⑤ The nonlinear kernels, e.g., Softmax, GeLU, Layernorm, take significant GPU cycles

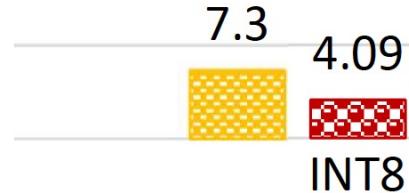
FPGA vs. GPU?

▣ FPGA U250, HeatViT

☒ GPU T4, TensorRT



Hardware Specification	FP32	INT8	Off-chip BW
AMD FPGA U250 [35]	1.2 T	6.95 T	77 GB/s
Nvidia GPU T4 [36]	8.1 T	130 T	320 GB/s

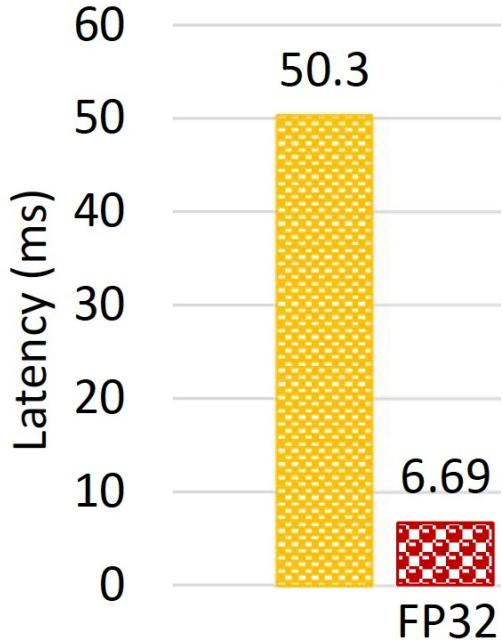


GPU+FPGA?

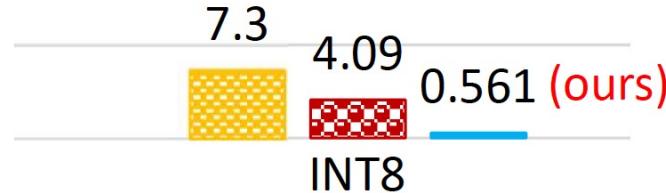
FPGA U250, HeatViT

GPU T4, TensorRT

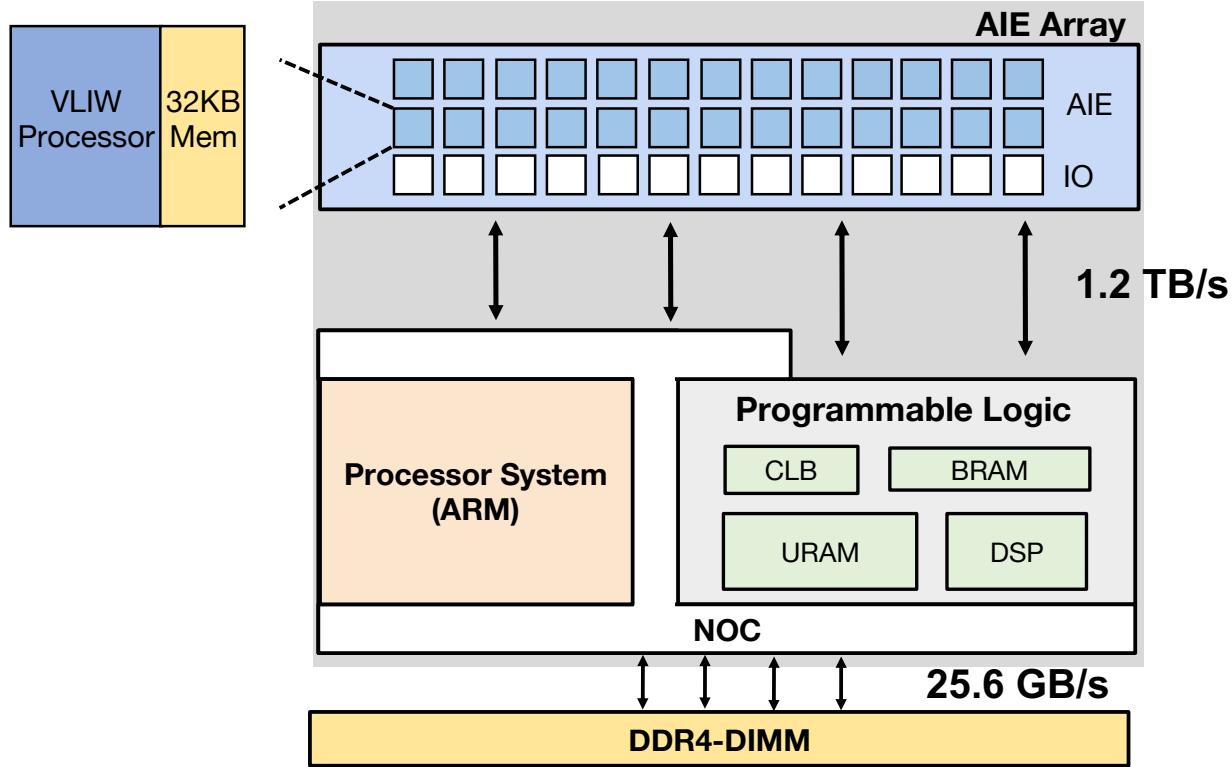
ACAP VCK190, EQ-ViT (ours)



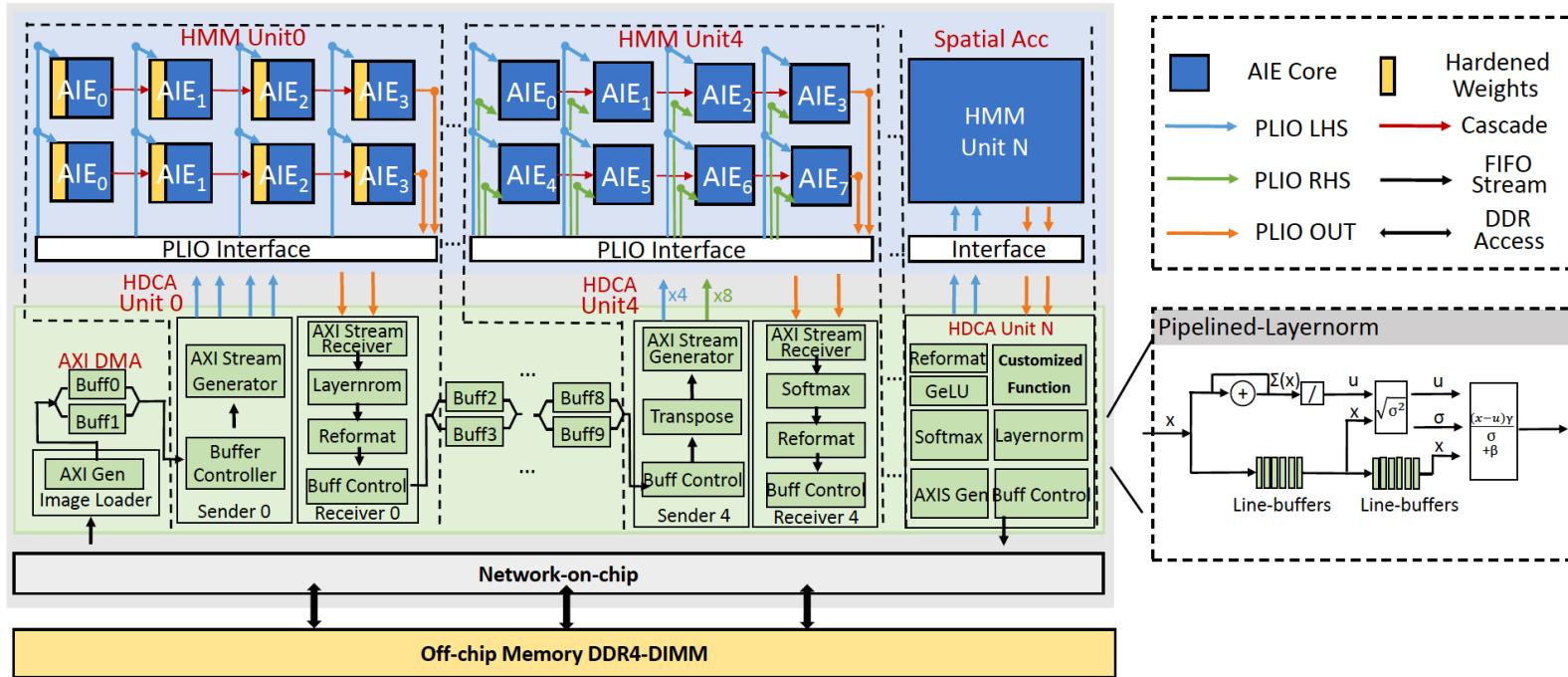
Hardware Specification	FP32	INT8	Off-chip BW
AMD FPGA U250 [35]	1.2 T	6.95 T	77 GB/s
Nvidia GPU T4 [36]	8.1 T	130 T	320 GB/s
AMD ACAP VCK190 [37]	6.4 T	102.4 T	25.6 GB/s



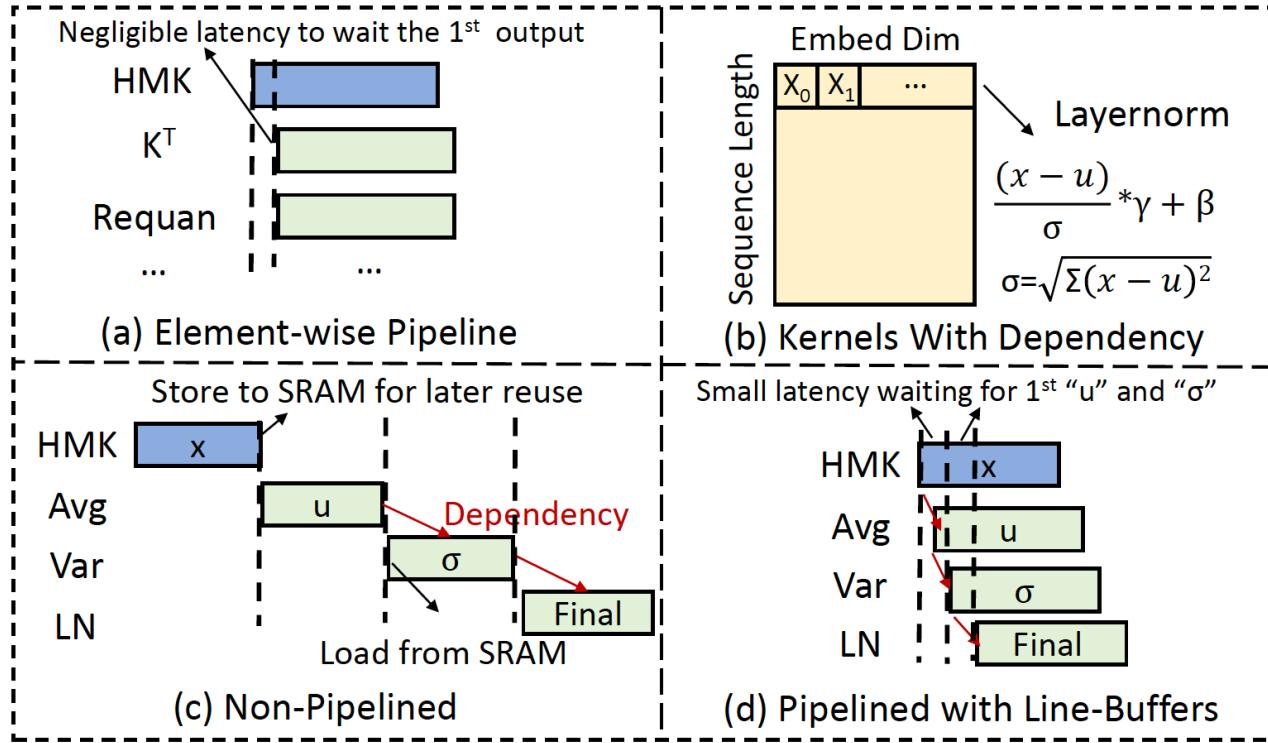
Versal ACAP Architecture



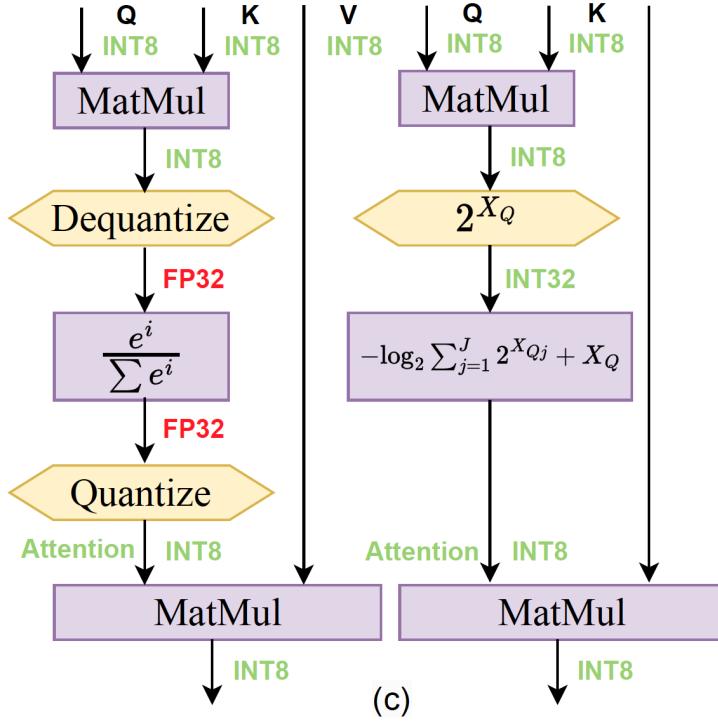
Heterogeneous Accelerator Architecture



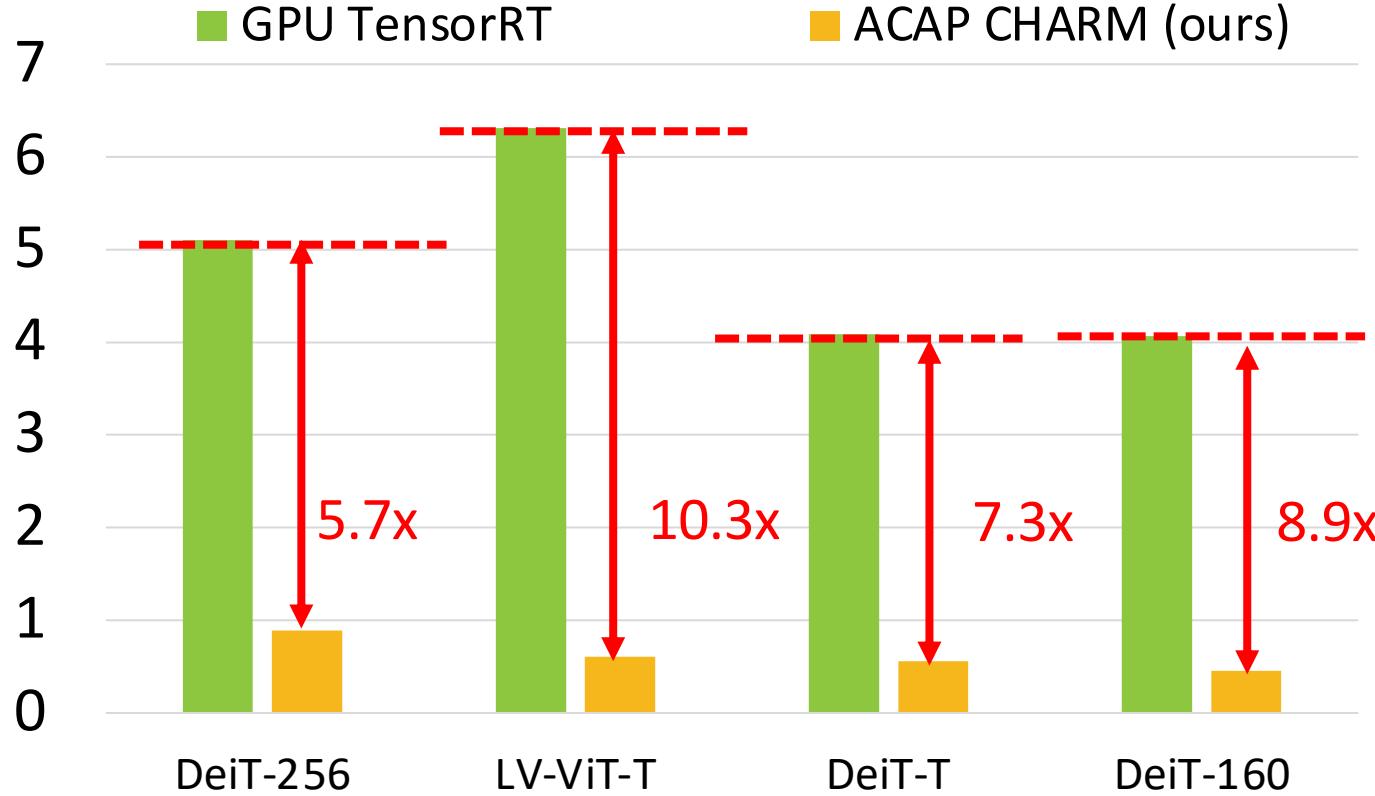
Fine-Grained Pipeline



INT Non-linear Functions (Softmax, GELU)

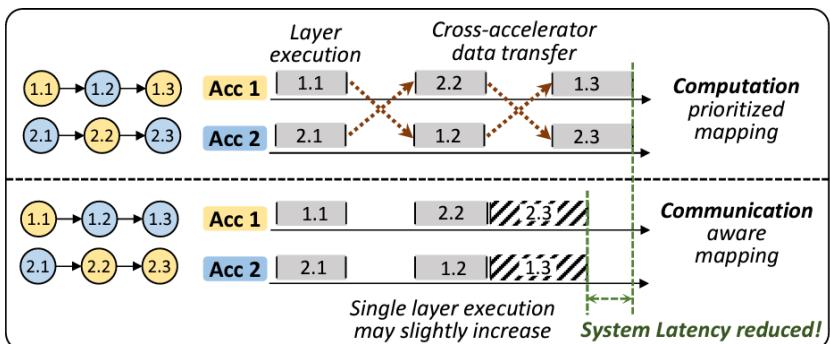


Reduces Latency by 10x over Nvidia GPU T4

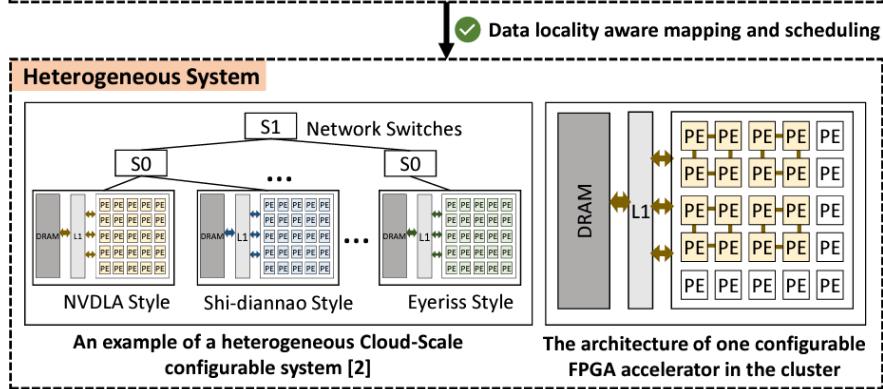
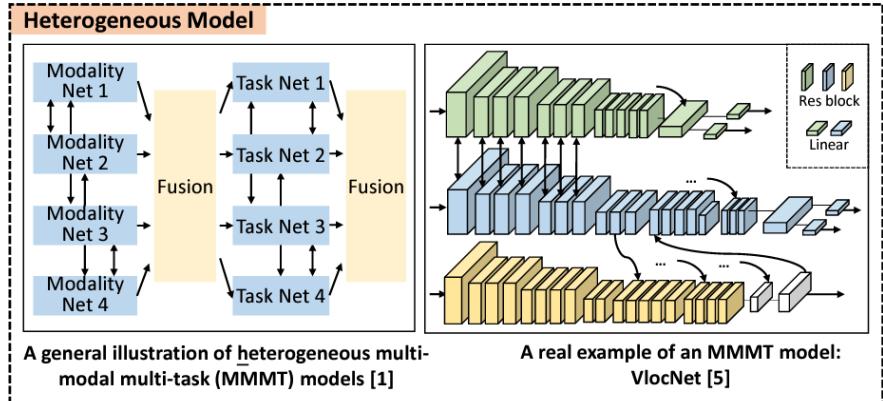


Scale-Out?

- From Heterogeneous Models to Heterogeneous System
- Computation-Communication Aware



H2H: heterogeneous model to heterogeneous system mapping with computation and communication awareness, DAC 2022



Lower Latency, Lower Energy

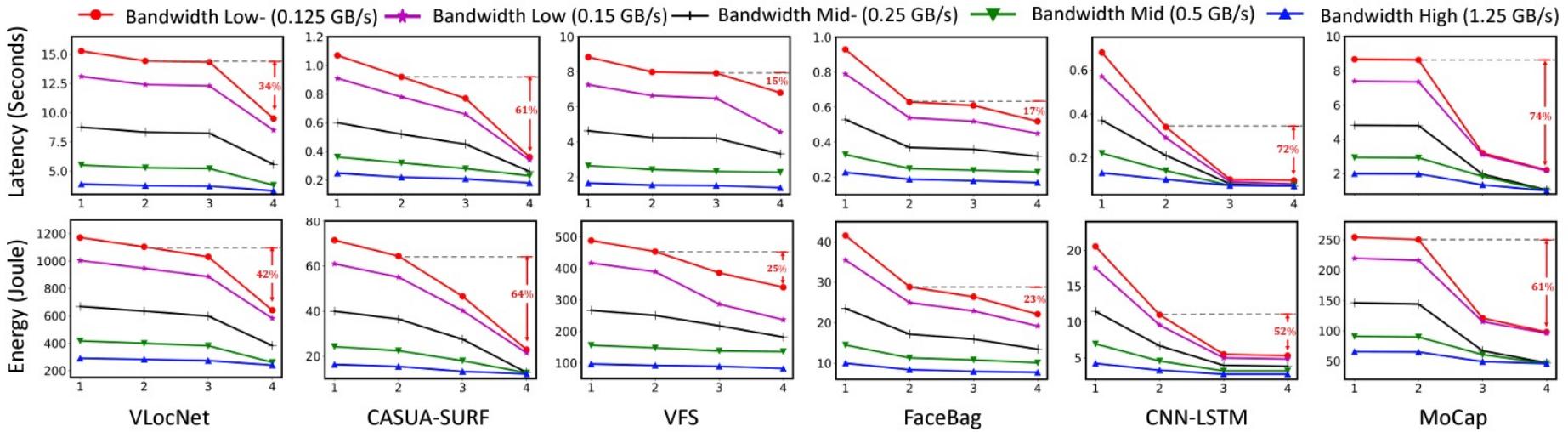


Figure 4: The latency and energy performance comparison.

H2H: heterogeneous model to heterogeneous system mapping with computation and communication awareness, DAC 2022



Open Source?

- <https://github.com/arc-research-lab/CHARM>
- <https://dl.acm.org/doi/10.1145/3543622.3573210>

RESEARCH ARTICLE

OPEN ACCESS •



CHARM: Composing Heterogeneous Accelerators for Matrix Multiply on Versal ACAP Architecture

Authors: Jinming Zhuang, Jason Lau, Hanchen Ye, Zhuoping Yang, Yubo Du, Jack Lo, Kristof Denolf, Stephen Neuendorffer, Alex Jones, Jingtong Hu, Deming Chen, Jason Cong, Peipei Zhou [Authors Info & Claims](#)

FPGA '23: Proceedings of the 2023 ACM/SIGDA International Symposium on Field Programmable Gate Arrays • February 2023 •
Pages 153–164 • <https://doi.org/10.1145/3543622.3573210>

Published: 12 February 2023 [Publication History](#)



0 1,450



About

CHARM: Composing Heterogeneous Accelerators for Matrix Multiply on Versal ACAP Architecture (Full Paper accepted to FPGA2023!)

fpga deeplearning

design-space-exploration versal

high-level-synthesis

electronic-design-automation

heterogeneous-computing acap

domain-specific-architecture versalacap

Readme

MIT license

Activity

85 stars

5 watching

11 forks

[Report repository](#)

Chiplet?

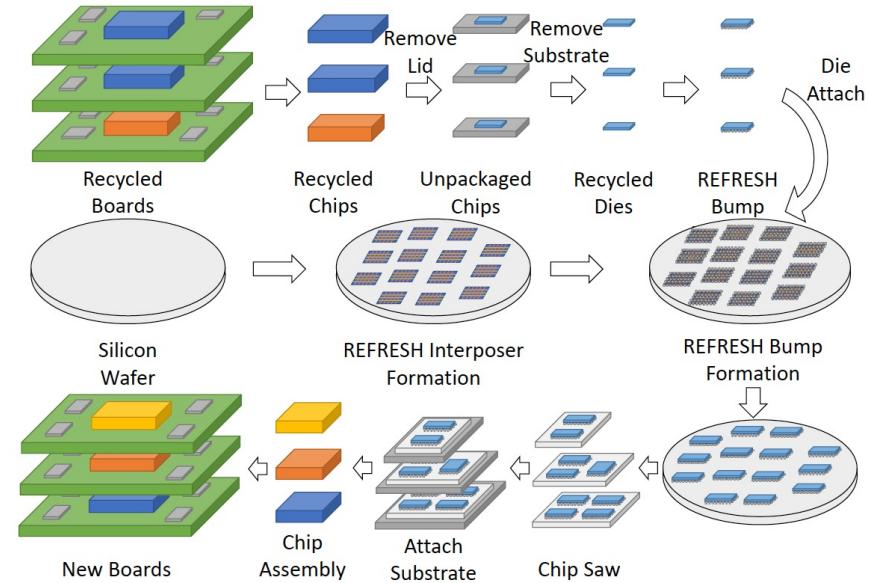
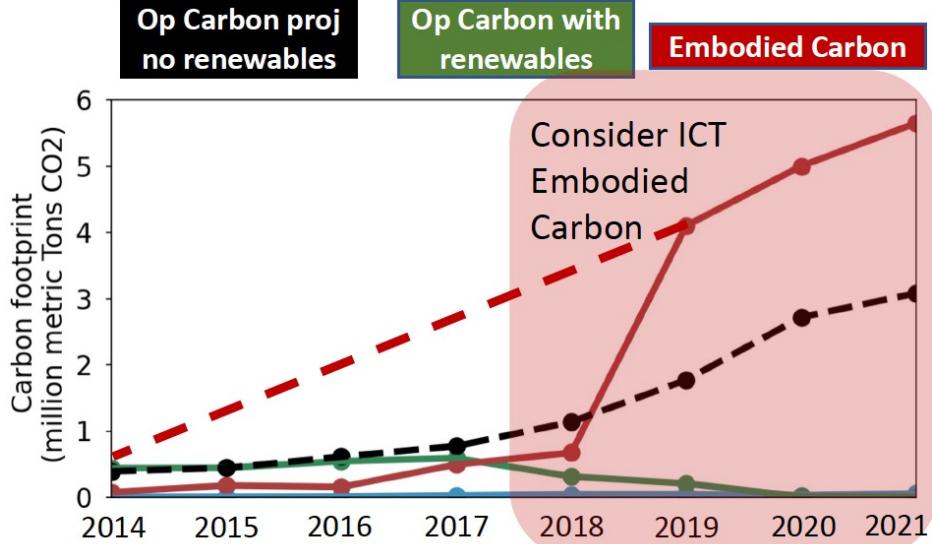
- H2H-> H2H2H
- Heterogeneous Models to Heterogeneous Chiplet Systems with Heterogeneous Components
- Computation & Communication Aware
- Hierarchical Scheduling & Mapping
- Latency vs Throughput



OCTOBER 18, 2023
SAN JOSE, CA

Scaling Innovation Through Collaboration

Sustainability?



Rereading Chiplets

Source of CO₂e from Meta Datacenters

NSF CCF#2324864: Collaborative Research: DESC: Type II: REFRESH: Revisiting Expanding FPGA Real-estate for Environmentally Sustainability Heterogeneous-Systems

Sustainability?

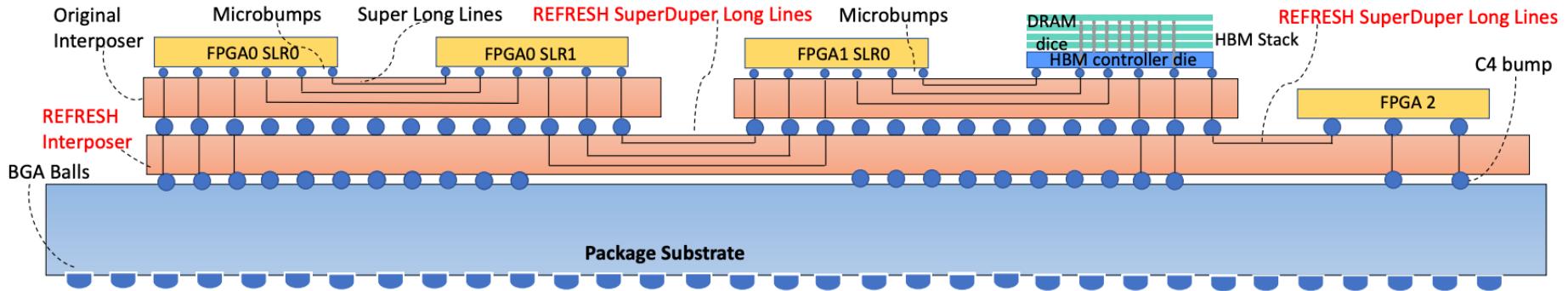


Fig. 2: REFRESH interposer for integration of homogeneous and heterogeneous monolithic and/or chiplet-based FPGAs.

NSF CCF#2324864: Collaborative Research: DESC: Type II: REFRESH: Revisiting
Expanding FPGA Real-estate for Environmentally Sustainability Heterogeneous-Systems

Peipei Zhou is an assistant professor of the Electrical Computer Engineering department at the University of Pittsburgh. Her research interests include design automation, hardware/software co-design, AI chip design, etc. She has participated in >\$11M Federal Funds (>\$2M as Lead PI).

Her work in FPGA acceleration for deep learning won the 2019 Donald O. Pederson Best Paper Award from the IEEE Council for Design Automation (CEDA). Her works have also won 2018 ISPASS Best Paper Nominee and 2018 ICCAD Best Paper Nominee.

<https://peipeizhou-eecs.github.io/>
peipei.zhou@pitt.edu





OCP

FUTURE TECHNOLOGIES SYMPOSIUM

OCP Global Summit | October 18, 2023 | San Jose, CA