TrustForge: A Cryptographically Secure Enclave

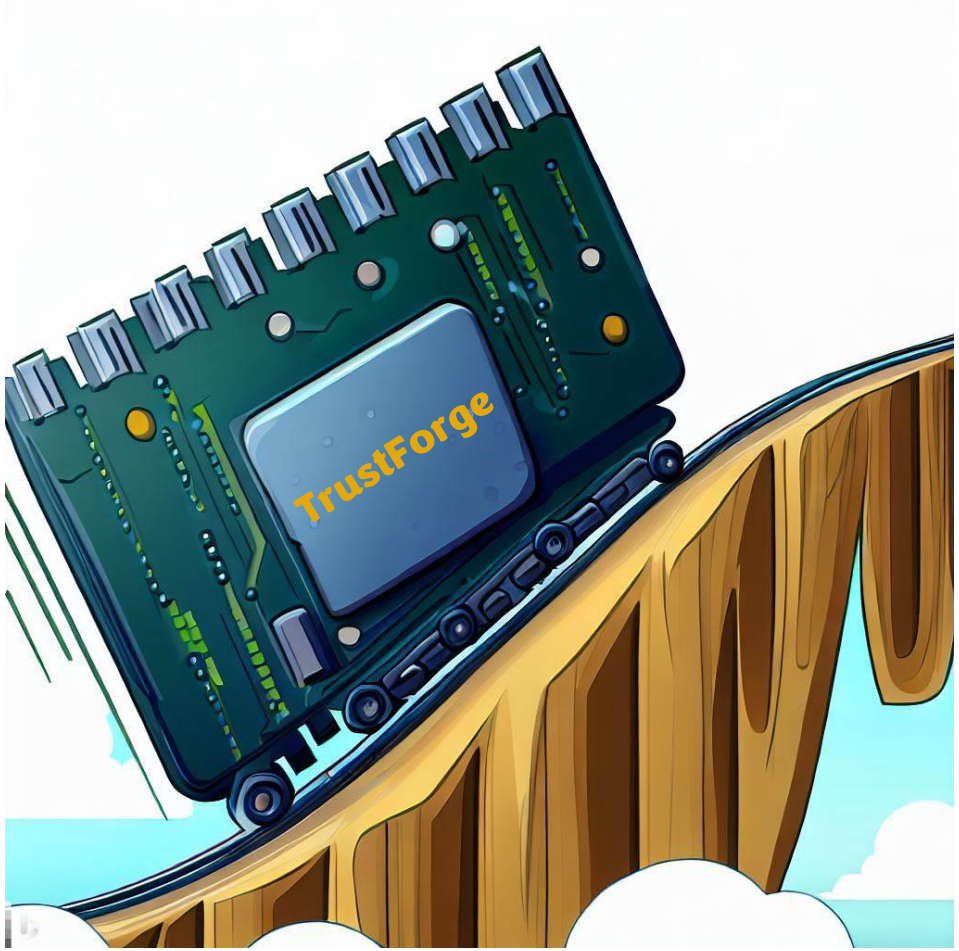
Todd Austin, CEO
Valeria Bertacco, Chief Scientist
Alex Kisil, Director of Engineering

austin@agitalabs.com

AGITA LABS

The Little (TrustForge) Enclave that Could…

- Execute RISC-like instructions directly on encrypted data
- Transform a hackable CPU into next-gen privacy tech that surpasses FHE and ZKP
- Nullify software and hardware hacking
- Pass commercial red teaming and full formal verification with zero vulnerabilities
- Stop data breaches once and for all!
- Advance data privacy by denying software access to sensitive data it is processing
- Do this with only a 190k-gate functional unit!



Drop by my poster to learn more!

An Open-Source 130-nm Fusion-Enabled Deconvolution Kernel Generator IC For Real-Time mmWave Radar Platform Motion Compensation

Nikhil Poole, Priyanka Raina, and Amin Arbabian

August 29, 2023

Problem/Motivation

Autonomous Vehicles

High-Performance GPUs/TPUs

Smart Infrastructures

Modern sensor networks are powerful, consuming 100s of W of power to achieve high resolution/performance.

But how can we process the same quantity of data on resource-limited edge devices with minimal latency?

Micro-Drones

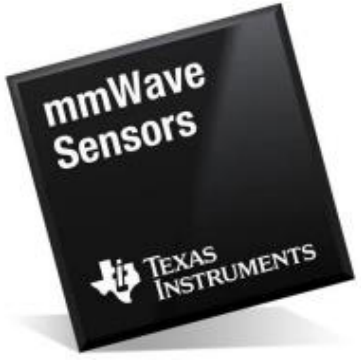
AR Glasses

Wearable/Portable Aids

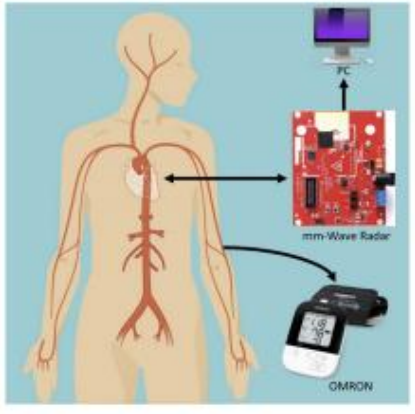
Goal:
Real-time edge sensing providing high-resolution with minimal power consumption.

mmWave Radar: An Optimal Sensing Modality

- cm-Scale Range Resolution ✓
- High Velocity Resolution ✓
- Mm-Scale Accuracy/Sensitivity ✓
- High Maximum Range + Velocity ✓
- Small, Easily Integrable Form Factor ✓
- Environmentally Robust ✓



Widely Applicable ✓



Primary Challenge With High-Resolution Edge Sensing

Drone Micromotion

mmWave Sensor Platform Motion

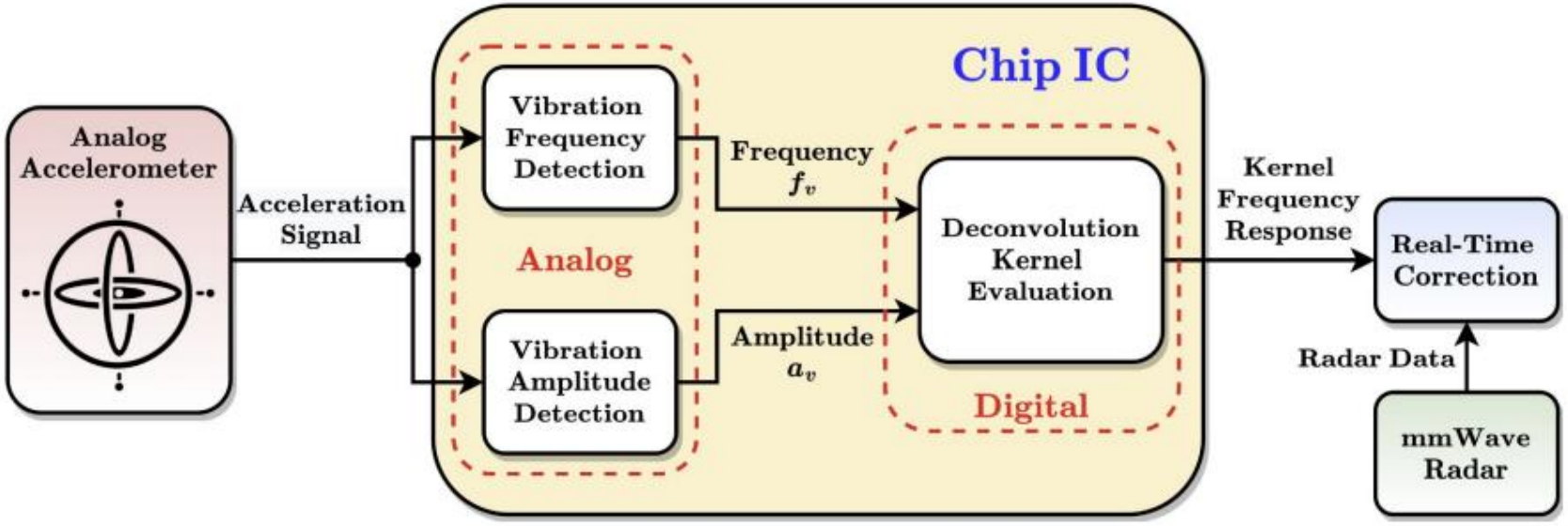
Engine Vibration

Wearable Devices

Portable Devices

Solution

First custom IC enabling <u>real-time</u>, <u>resource-efficient</u> FMCW mmWave radar platform vibratory motion compensation, via early sensor fusion.



26-35 Vibration Suppression

FULLY OPEN-SOURCE

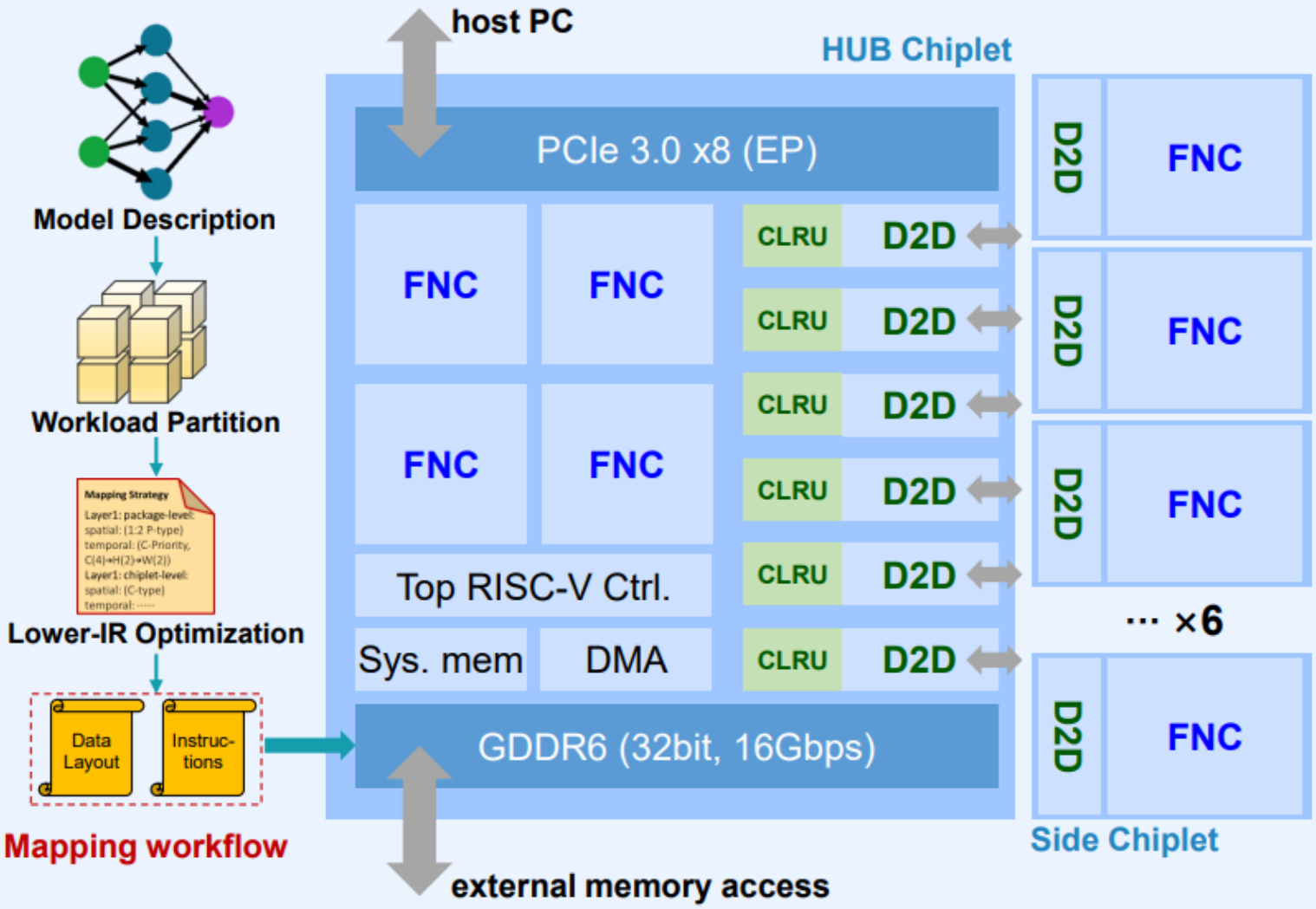
< 95-ms Latency

A Scalable Multi-Chiplet Deep Learning Accelerator with Hub-Side 2.5D Heterogeneous Integration

Zhanhong Tan[1], Yifu Wu[2], Yannian Zhang[2], Haobing Shi[2], Wuke Zhang[2], Kaisheng Ma[1]

[1]Tsinghua University,
[2]Polar Bear Tech

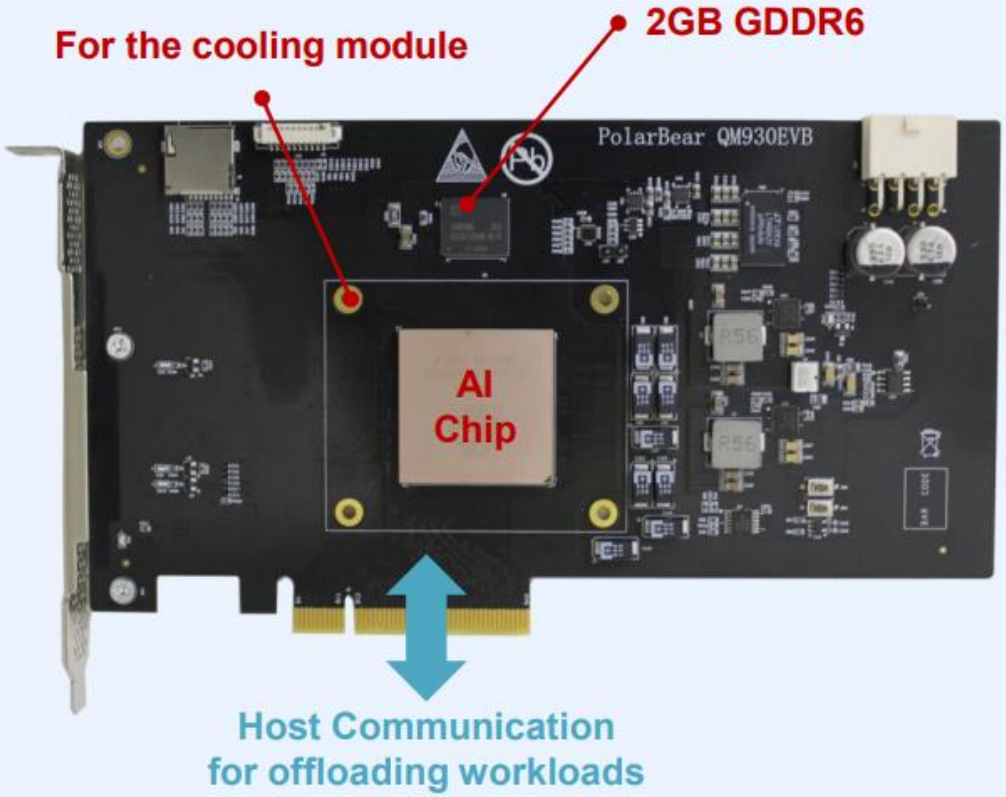
Overall Architecture



- **Flexible Neural Core (FNC)**
  - Reconfigurable architecture for the shape diversity
- **Mapping dataflow**
  - Die-to-Die communication-aware workload generator
- **Interconnection**
  - High-bandwidth Die-to-Die based on 2.5D package
  - Efficient chiplet routing unit (CLRU)

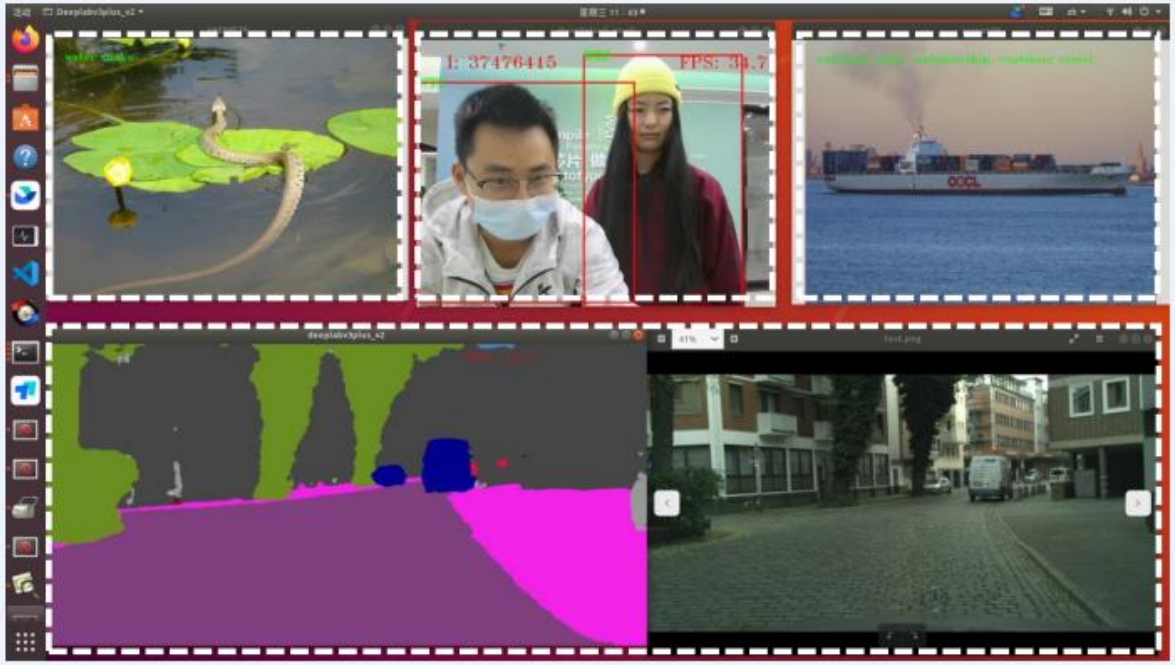
System Board and Demo

■ System PCIe-based Board



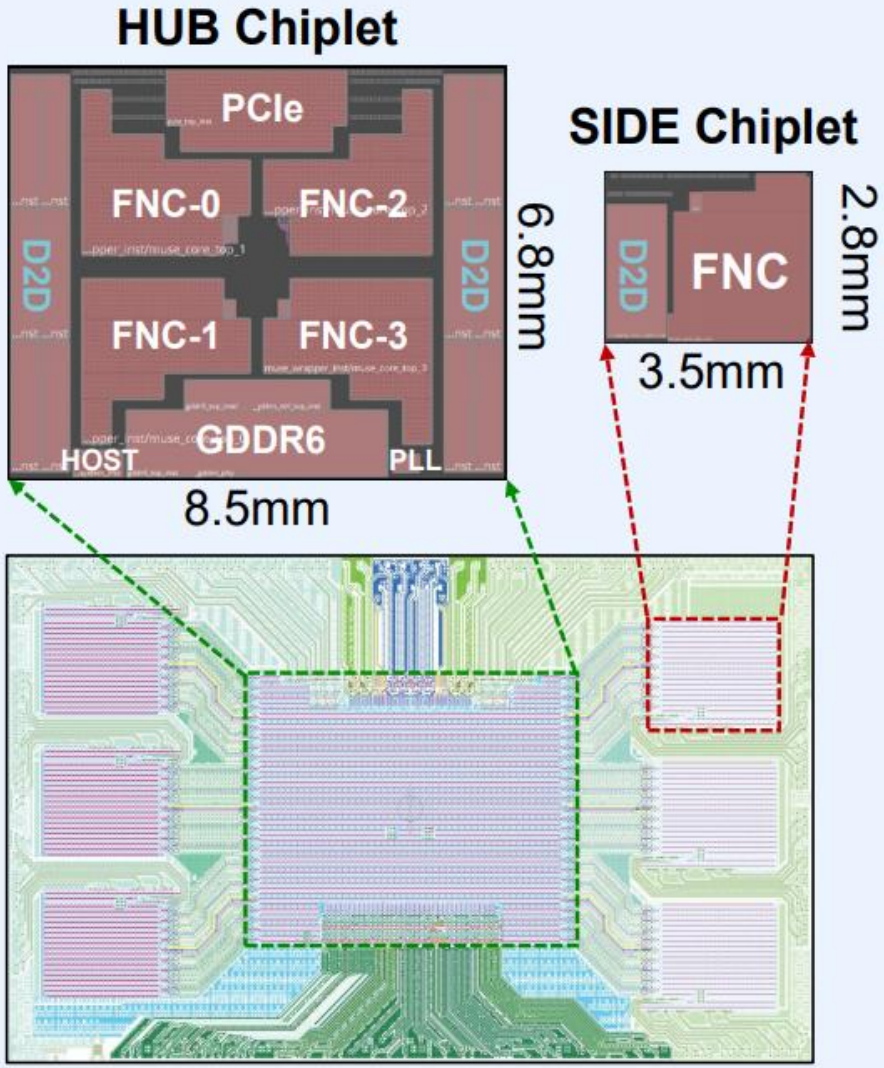
■ Demo for running concurrent 4 models

Model 1: VGG-16 Model 2: YOLO v5s Model 3: ResNet50



Model 4: DeepLab v3+

Chip Summary



RDL Layer for 2.5D Package

Items		Specifications
Technology		CMOS 12nm
Die Area	HUB Chiplet	8.5mm × 6.8mm = 57.8mm²
	SIDE Chiplet	3.5mm × 2.8mm = 9.8mm²
Supply Voltage		0.8V ~ 1.2V
Frequency		100MHz – 1GHz
Peak Performance	INT4	40TOPS (8b-A × 4b-W)
	INT8	20TOPS (8b-A × 8b-W)
	INT16	10TOPS (16b-A × 8b-W)
NPU Core Efficiency		2.02TOPS/W
Power		4.5W ~ 12W
D2D Bandwidth		6×24GB/s for TX/RX
External Memory Bandwidth		64GB/s (GDDR6)
Bump Pitch for 2.5D Pkg		55μm

PHEP: Paillier Homomorphic Encryption Processors for Privacy-Preserving Applications in Cloud Computing

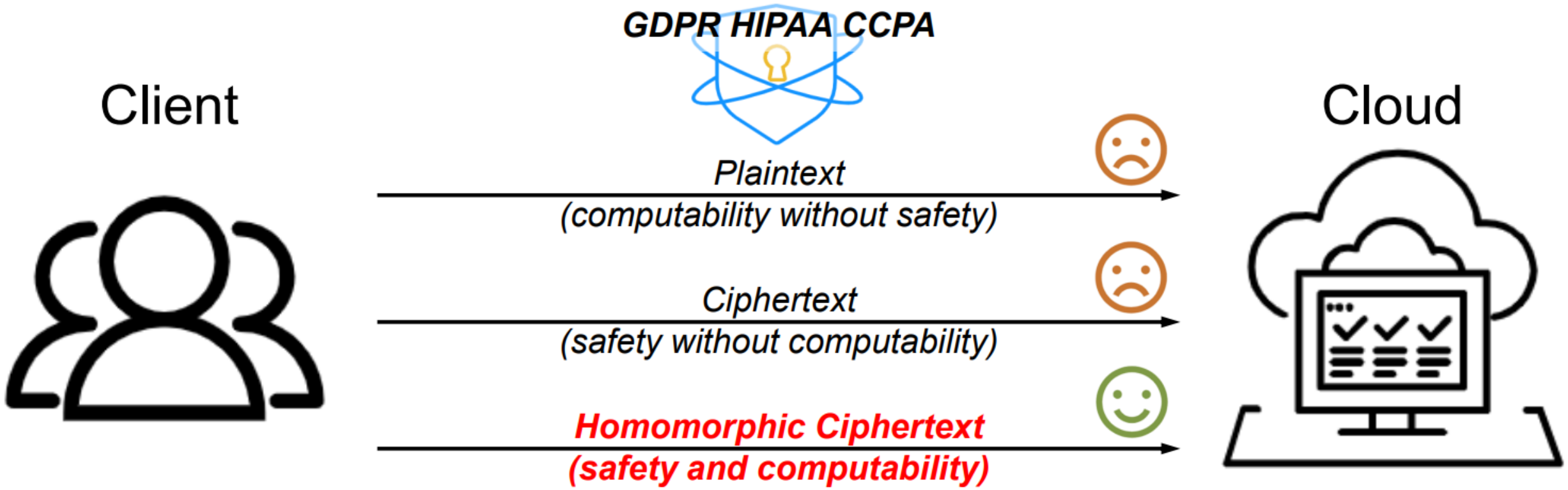
Guiming Shi[1], Yi Li[2], Xueqiang Wang[2], Zhanhong Tan[1], Dapeng Cao[3], Jingwei Cai[1], Yuchen Wei[1], Zehua Li[3], Wuke Zhang[4], Yifu Wu[4], Wei Xu[1*], and Kaisheng Ma[1*]

[1]Tsinghua University
[2]HuaKong TsingJiao
[3]Xi'an JiaoTong University
[4]Polar Bear Tech

*Corresponding Authors

Homomorphic Encryption in Cloud Computing

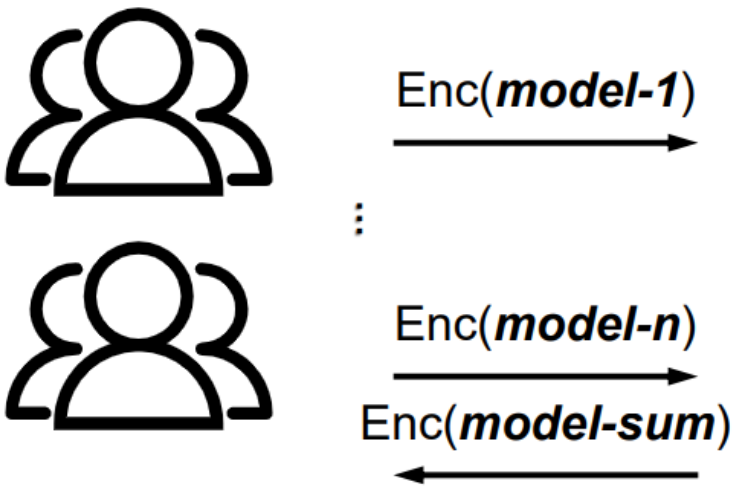
- Data privacy is a critical problem in Cloud Computing.
- Paillier Homomorphic Encryption can protect the privacy of the data and enable computing on the ciphertext without decryption first.

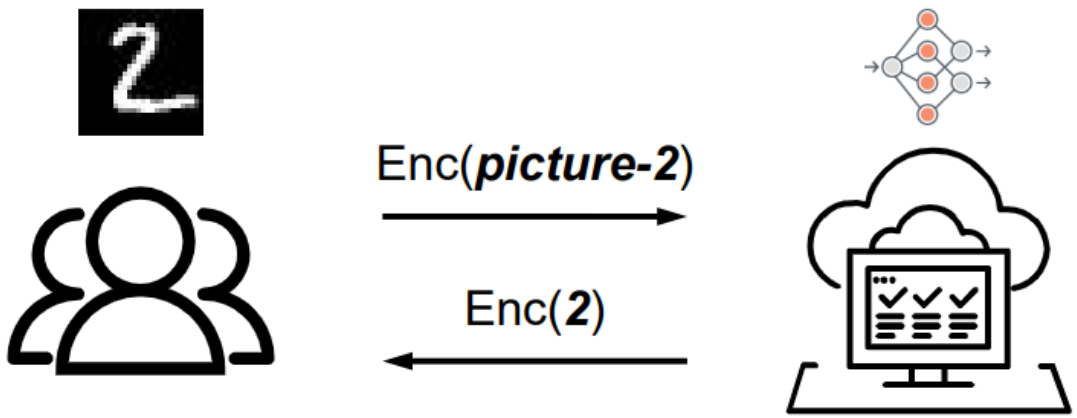


Bottleneck in Training and Inference Applications are Different

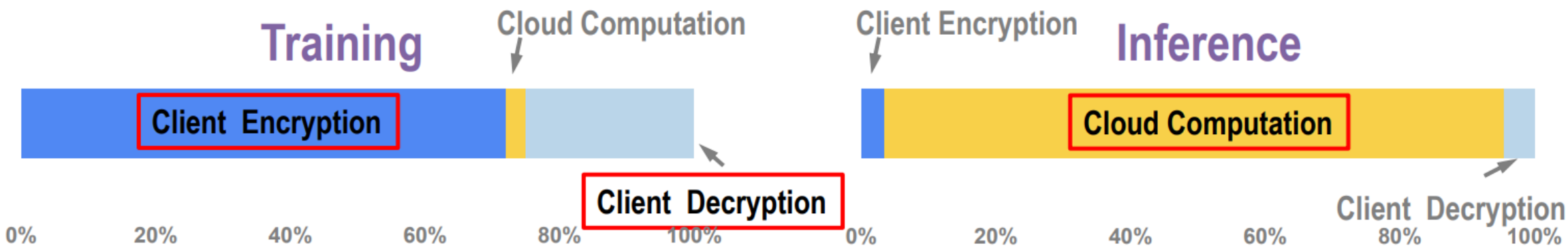
Application-Training, Bottleneck: Client Encryption and Decryption

Application-Inference, Bottleneck: Cloud Computation





Latency Breakdown On CPU



Comparison of the Two Engines: Specification

Fabricated on the Same Wafer: Significantly Reduces NRE of the Engine Chips

Same Area：43mm^2 @ UMC 28nm HPC+

- **Optimized for Parallelism**

Items/PE	Montgomery Unit
Algorithm	Montgomery Multiplication
Arithmetic	3*128 Bit Multiplier

- **Optimized for Performance**

Items/PE	Montgomery Unit	Stein Unit
Algorithm	Montgomery Multiplication	Stein Modular Inversion
Arithmetic	3*256 Bit Multiplier	3*4102 Bit Adder

High Performance Paillier Homomorphic Encryption Processors

PHEP Engine-1

- ***480 TOPS (INT8)***
- ***Client Encryption: 84KOPs***
- ***Cloud Computation: 402KOPs***
- ***Client Decryption: 106KOPs***

PHEP Engine-2

- ***192 TOPS (INT8)***
- ***Client Encryption: 52KOPs***
- ***Cloud Computation: 47MOPs***
- ***Client Decryption: 48KOPs***

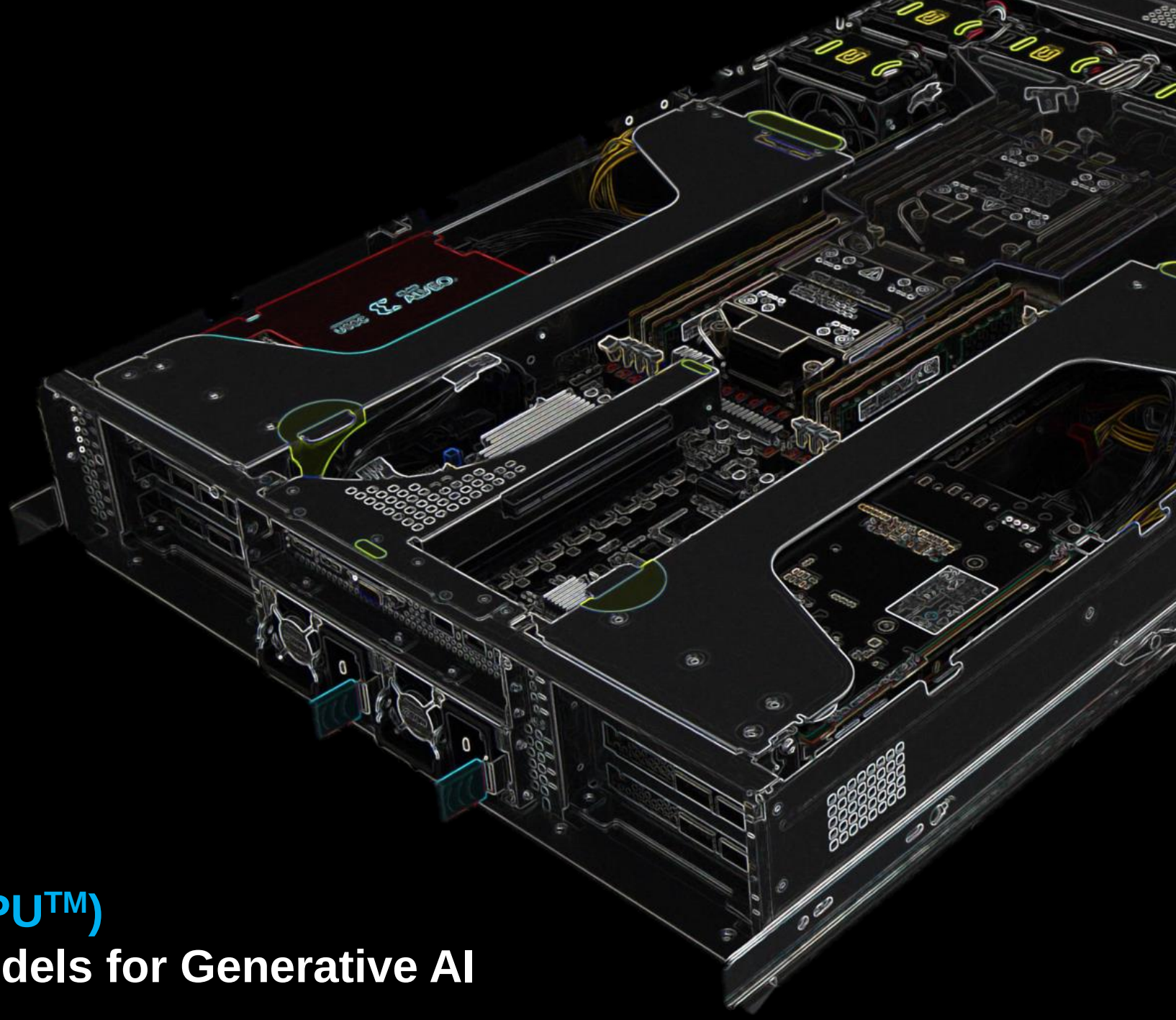
$Bit\ width\ of\ ciphertext = 4096, Bit\ width\ of\ plaintext = 64, Bit\ width\ of\ weight\ in\ Conv = 8.$
$Maximum\ Performance\ in\ Optimized\ Applications.$

Thank You

shigm21@mails.tsinghua.edu.cn

HOT CHIPS

HYPER ACCEL



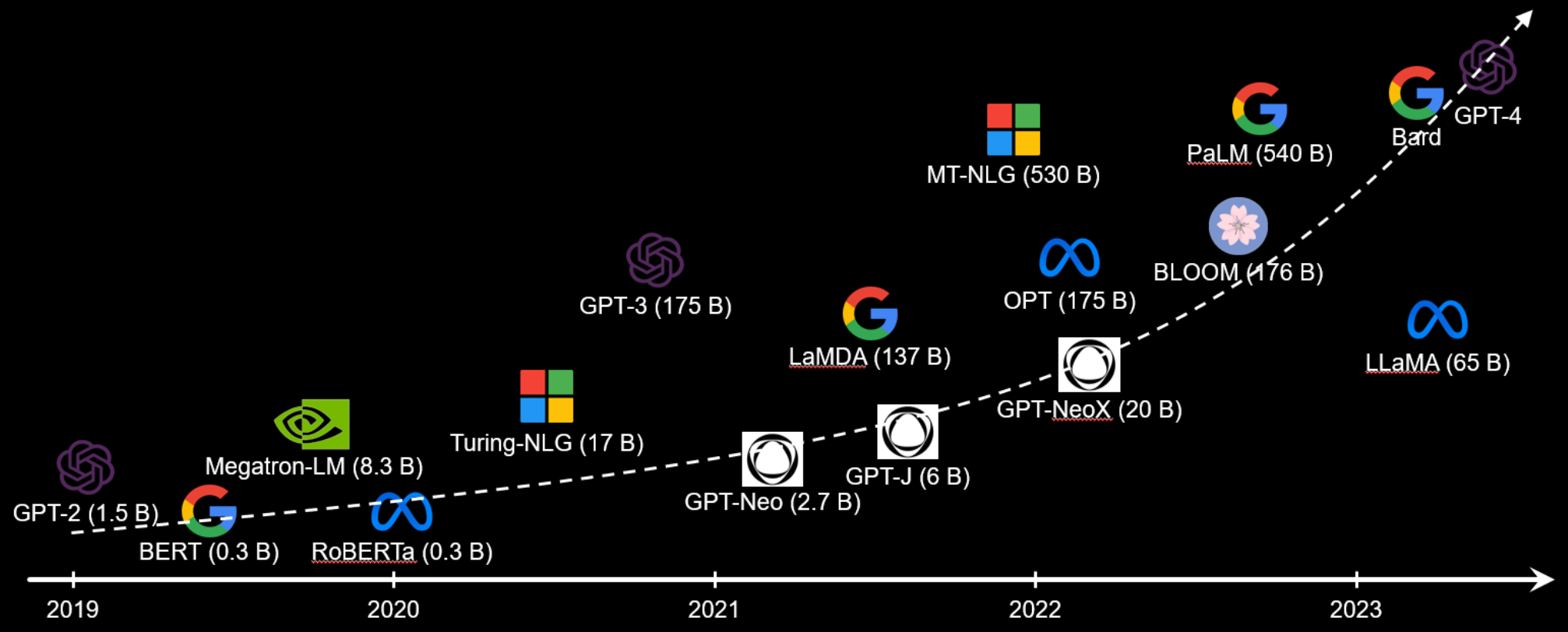
5

Seungjae Moon
Co-Founder, Research Scientist

Latency Processing Unit (LPU™)
Accelerating Hyperscale Models for Generative AI

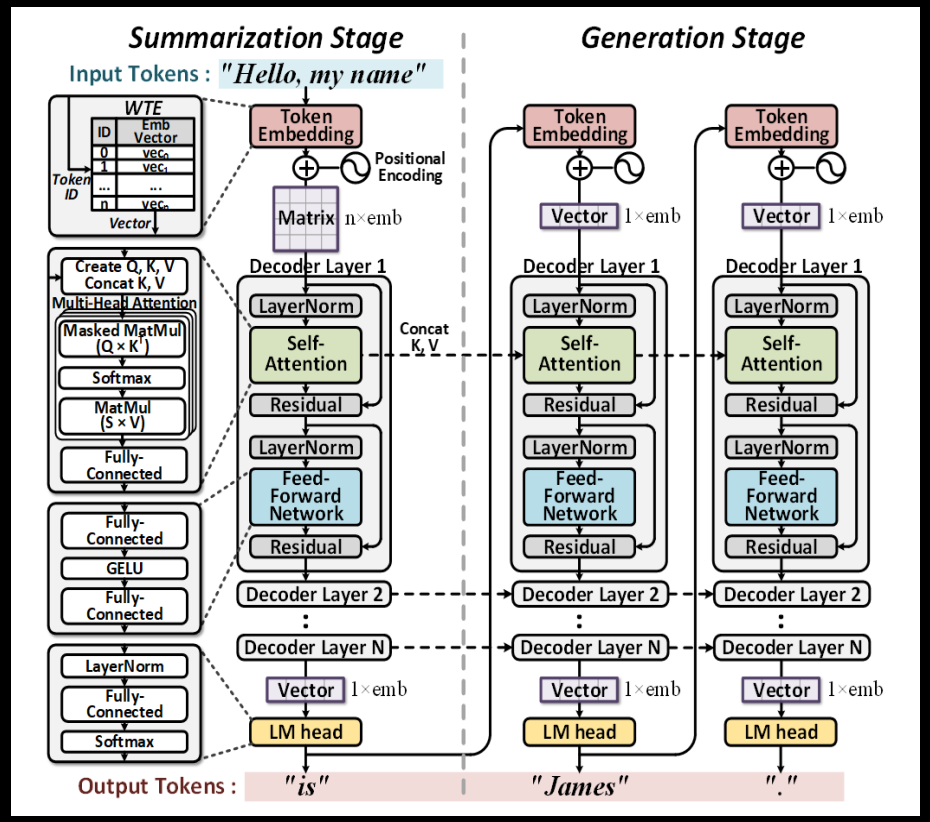
Large Language Model (LLM)

- Demand for highly scalable model inference hardware to support growing LLM



HYPER ACCEL

LLM Training & Inference Characteristics



Training

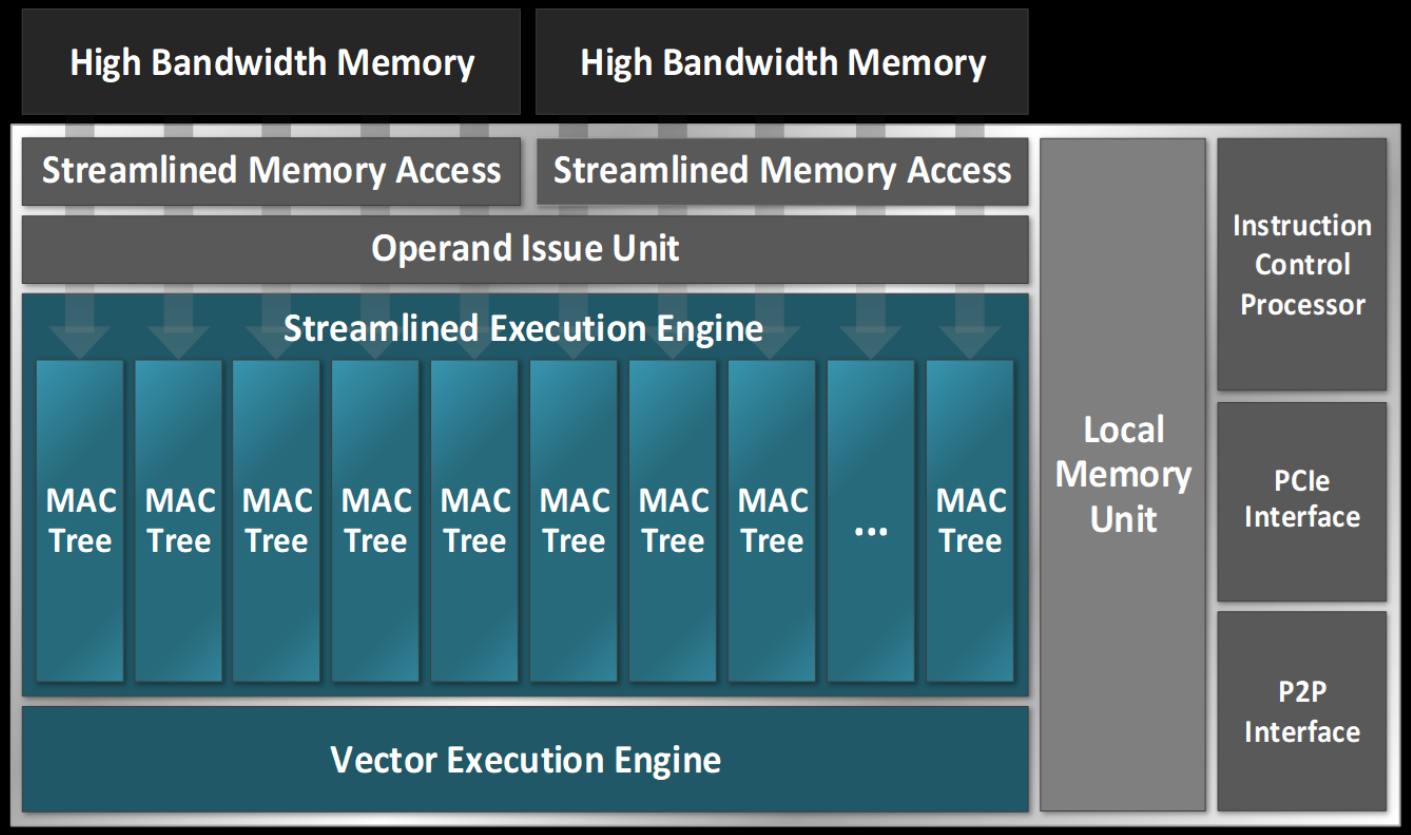
- Large batch
- Compute-intensive
- Throughput-oriented hardware

Inference

- Small batch
- Memory-intensive
- Latency-oriented hardware

HYPER ACCEL

Latency Processing Unit (LPU)



WORLD-FIRST HARDWARE DEDICATED FOR END-TO-END INFERENCE OF LLM

HYPER ACCEL

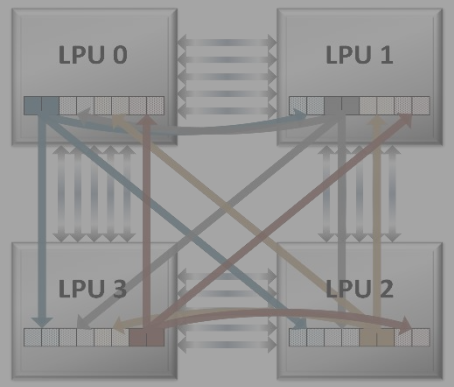
LPU Key Features

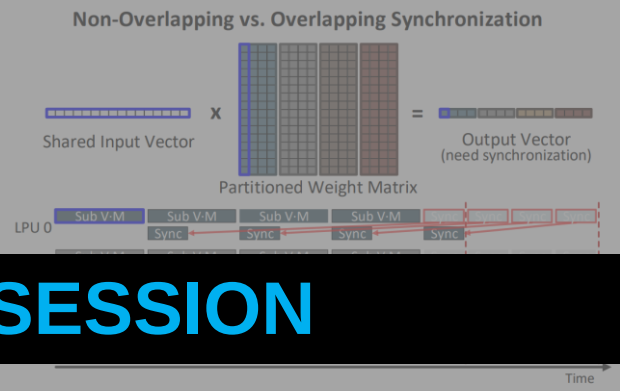
Streamlined Execution Engine

- Perfectly balances memory bandwidth and compute logic to maintain bandwidth usage of ~90%

Expandable Synchronization Link

- Lightweight full-duplex peer-to-peer communication technology with custom protocol
- Performs data synchronization with low latency and latency hiding to achieve near-perfect scalability

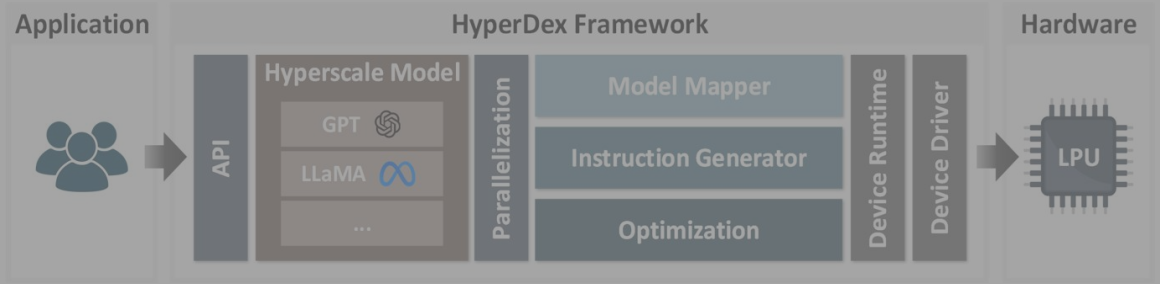




DETAILS AT OUR POSTER SESSION

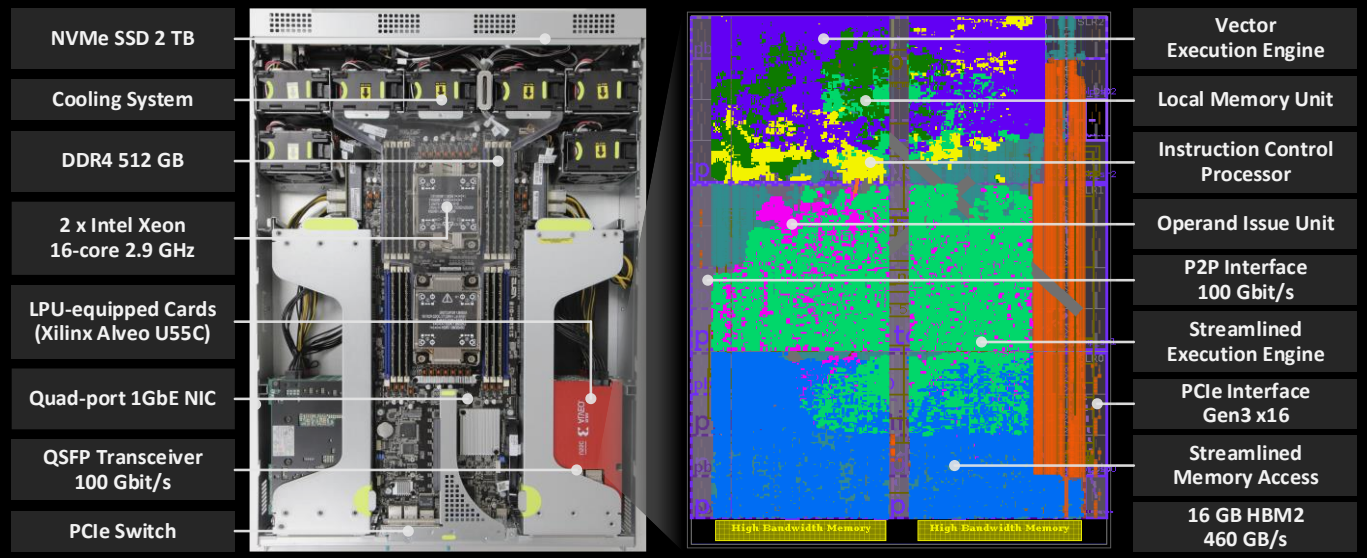
HyperDex Software Stack

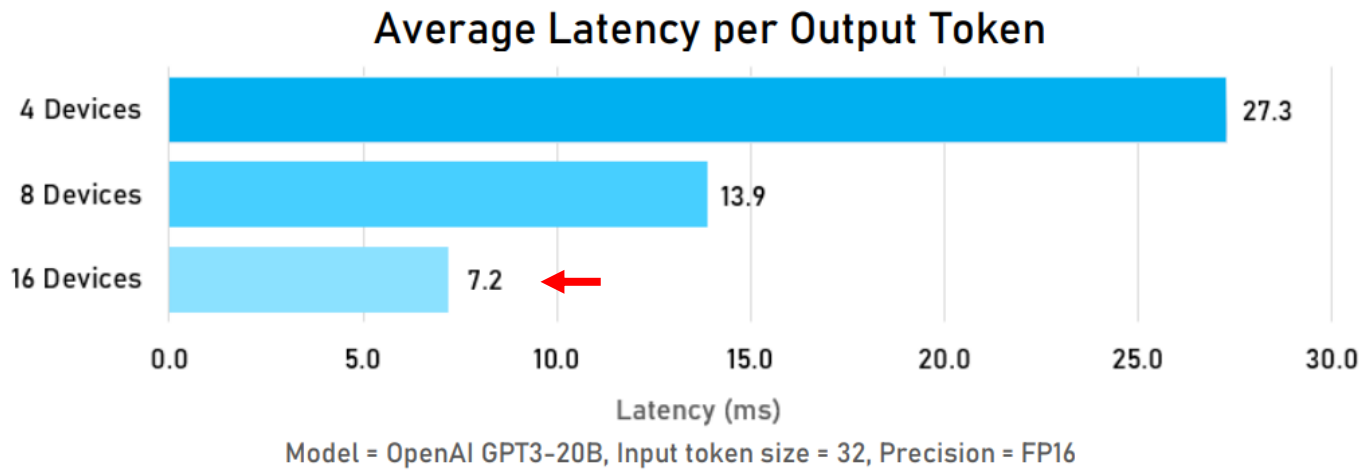
- Software framework that supports various language model structures and user requests
- Runs automated tools to create prerequisite data (e.g., parallelization, mapping, and instructions) to run the LPU at the application level



HYPER ACCEL

Performance





140 TOKENS PER SECOND FOR TEXT GENERATION WORKLOAD

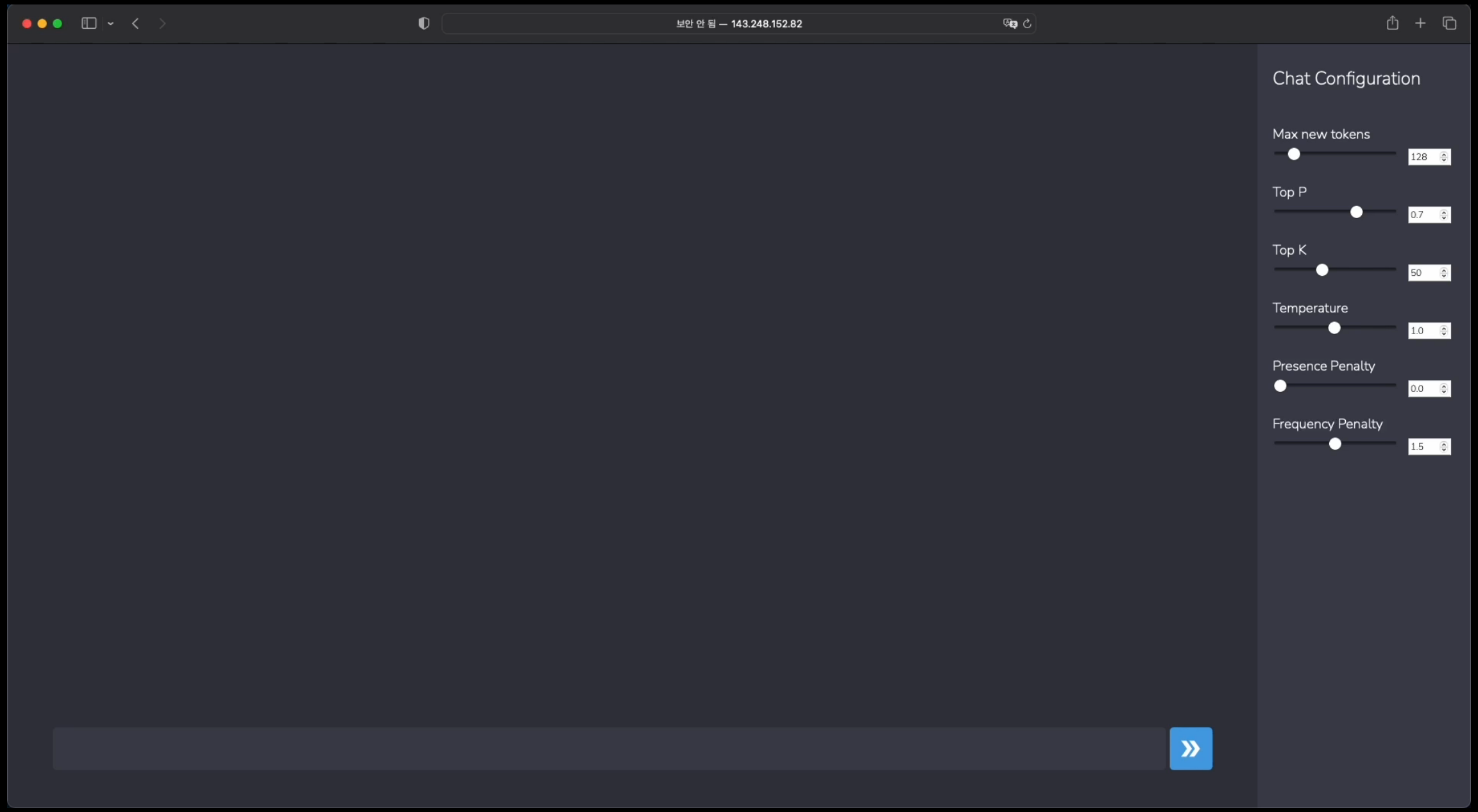
HYPER ACCEL

1 LPU vs. 8 LPUs on Meta AI LLaMA2 7B Model

1 LPU (49.6sec)

8 LPUs (8.7sec)

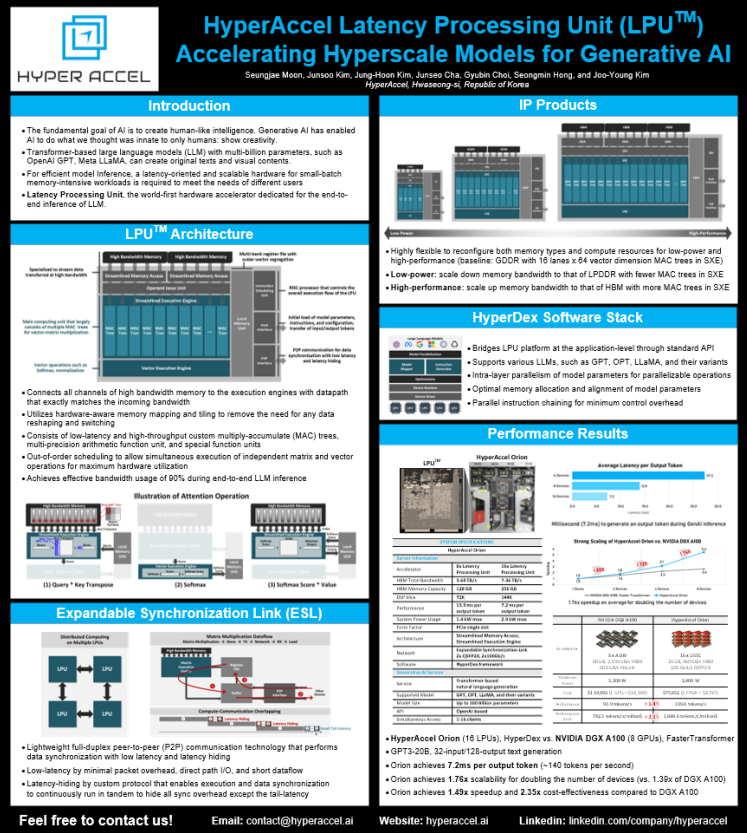
Chatbot UI


보안 안 됨 — 143.248.152.82
Chat Configuration
Max new tokens
128
Top P
0.7
Top K
50
Temperature
1.0
Presence Penalty
0.0
Frequency Penalty
1.5

HYPER ACCEL

- HyperAccel, a startup that provides hyper-accelerated silicon IP and solutions for emerging applications

- Currently supported models (up to ~100B parameters)
  - OpenAI GPT 1, 2, 3
  - Meta AI LLaMA 1, 2
  - Meta AI OPT
  - Stanford Alpaca
  - EleutherAI Polyglot
  - EleutherAI GPT Neo-X



For more information, please visit our poster session!

Website: www.hyperaccel.ai
Email : sj.moon@hyperaccel.ai | contact@hyperaccel.ai

MLSoC™ - An overview

Hot Chips 35, August 28-29, 2023

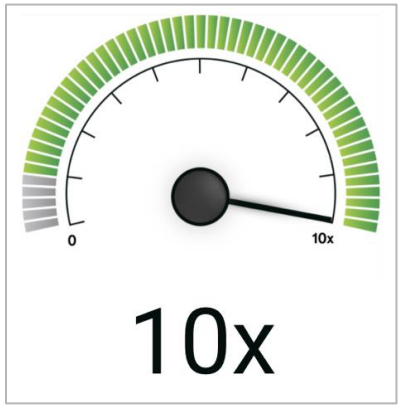
Srivi Dhruvanarayan, Victor Bittorf

Our Vision: Effortless machine learning for the embedded edge

Run any computer vision application, any network, any model, any framework, any sensor, any resolution.

Wide variety of end2end CV applications; Accuracy and common operator support

Any



Application

MLSoC: Run entire CV application efficiently

Efficient ML handling: Tile architecture.

Scheduling: Operations efficiently pipelined, scheduled

Performance

Data handling: On chip dec/enc, compliant with AVC, HEVC

Pre/Post processing: Dedicated CVU

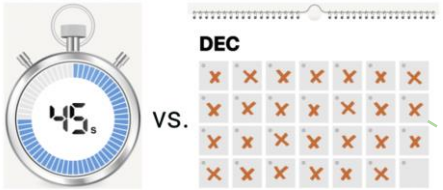
Efficient static scheduling: Maximize compute while minimizing data movement;

Low power

Quantization scheme: Patented accurate low-power scheme

Fully INT8 inference: 100% of total compute dedicated

Patented cache usage: Manage memory hierarchy



Pushbutton

Low code productization; Ready to use models, production ready platforms and dev kits

SiMa.ai key innovations

ANY and 10x

Highly flexible ML accelerator

Fully programmable MLA

Supports full complement of CV applications at lowest power

SW controlled data movement, scheduling & synchronization

SW control of cache/memory hierarchy, data movement: minimal data movement, small cache, high compute efficiency and lowest power

Secure, self-contained SoC

Seamless heterogeneous compute

Enables ML capabilities for legacy apps; future proofs applications

Optimized end-to-end SoC pipeline

Highly optimized building blocks enable best in class end-to-end performance

Pushbutton

Simple to develop & deploy

Effortless customer integration

Low code ML
Customers can develop & deploy applications without having to understand details of HW

Low code customer evaluation

Vision Development Platform (VDP)

Allows customers to build apps without having to write code

Purpose built for ML edge at embedded edge

Acquire ANY Data

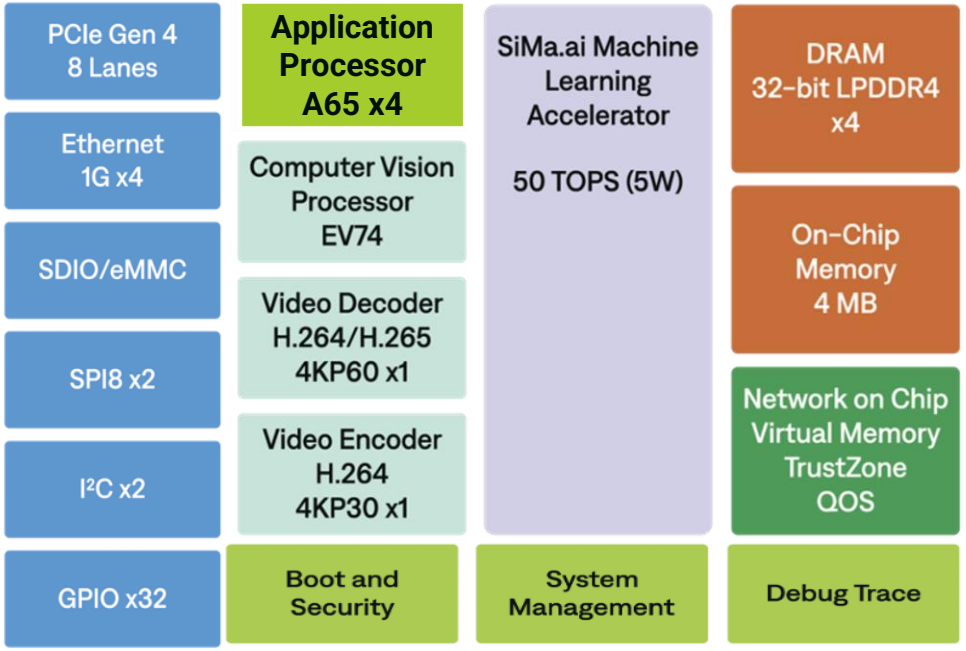
PCIe Gen4, 1GB Ethernet
Dedicate: I2C, SPI, GPIO, SDIO

10X ML Processing

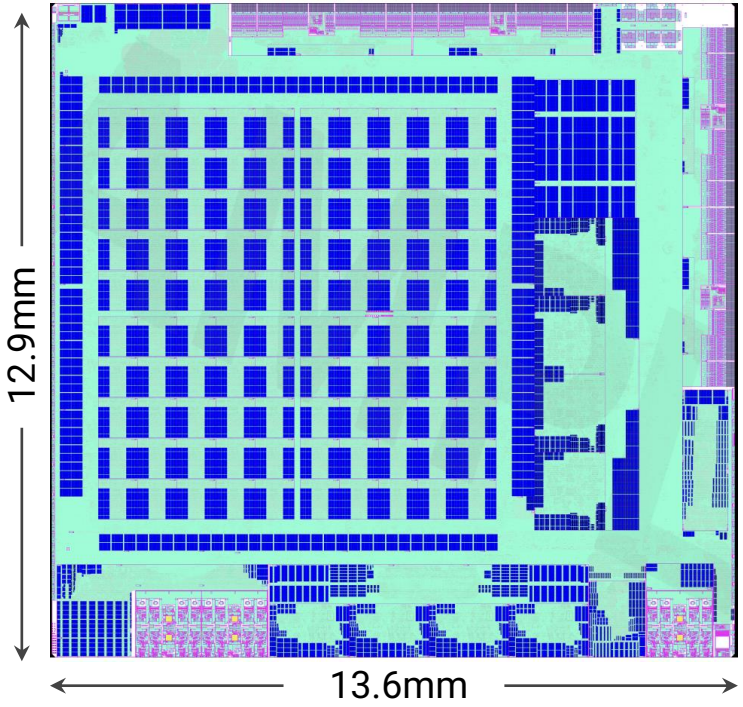
Video, CV & ML Processors
ML Accelerator - 50 TOPS INT8
Quad Vector DSP - 600 GOPS
HW video encode/decode
16GB LPDDR4

Decide, Control & Update

Quad A65E ARM8.3
Dedicated secure boot processor



TSMC - 16nm
TDP - 15-20W
Typical ML Workloads, CV Pipelines - 8-10W

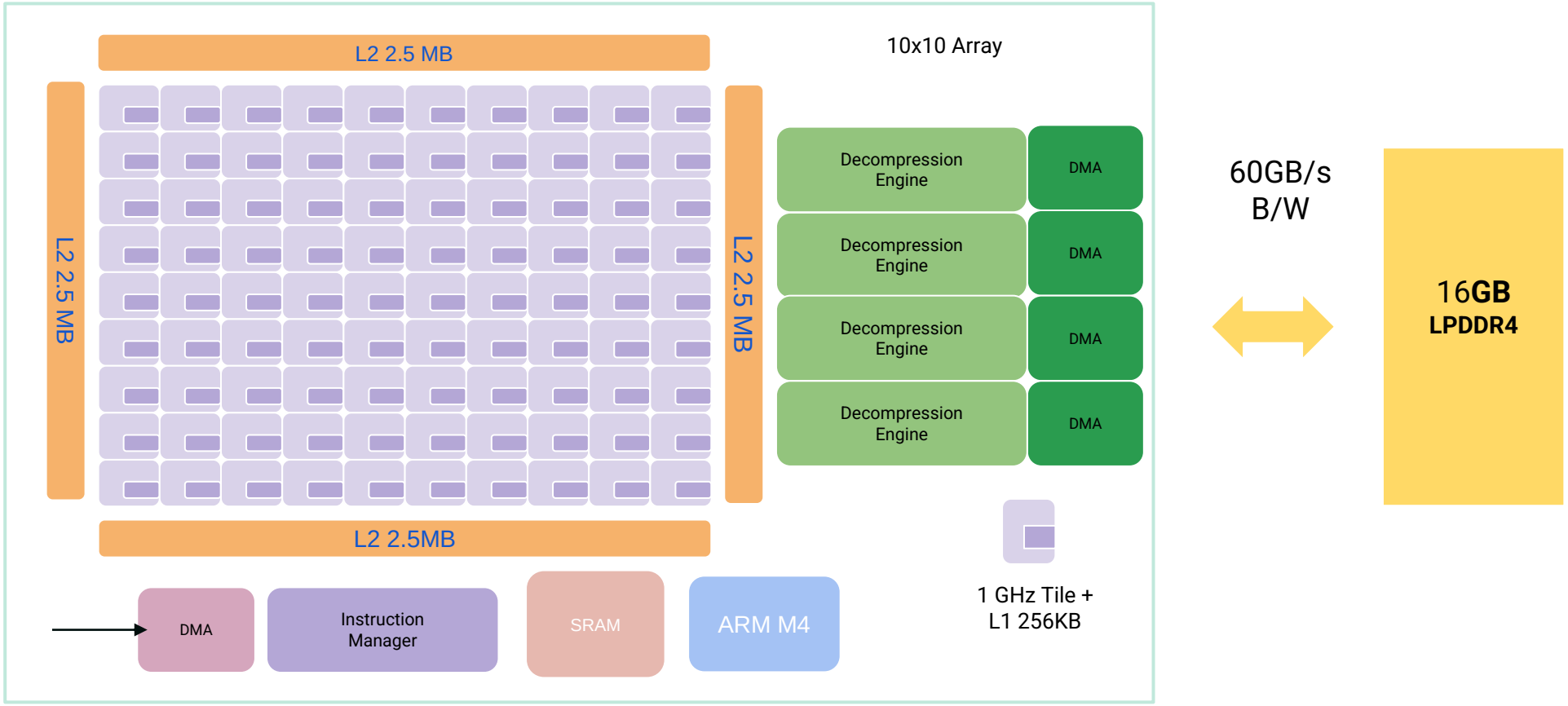


MLSoC™ Machine Learning System-on-Chip

Silicon Overview - 10x Performance for ML Processing

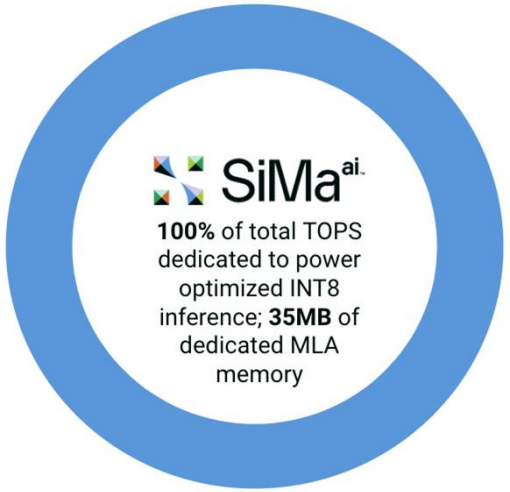
Machine Learning Accelerator

50 INT8 TOPS

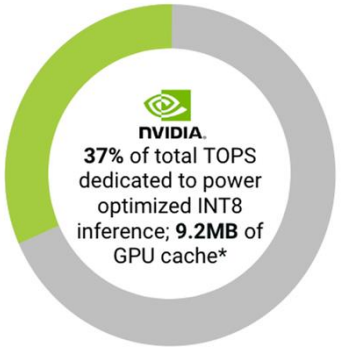


MLPerf: SiMa.ai delivers advantage over NVIDIA

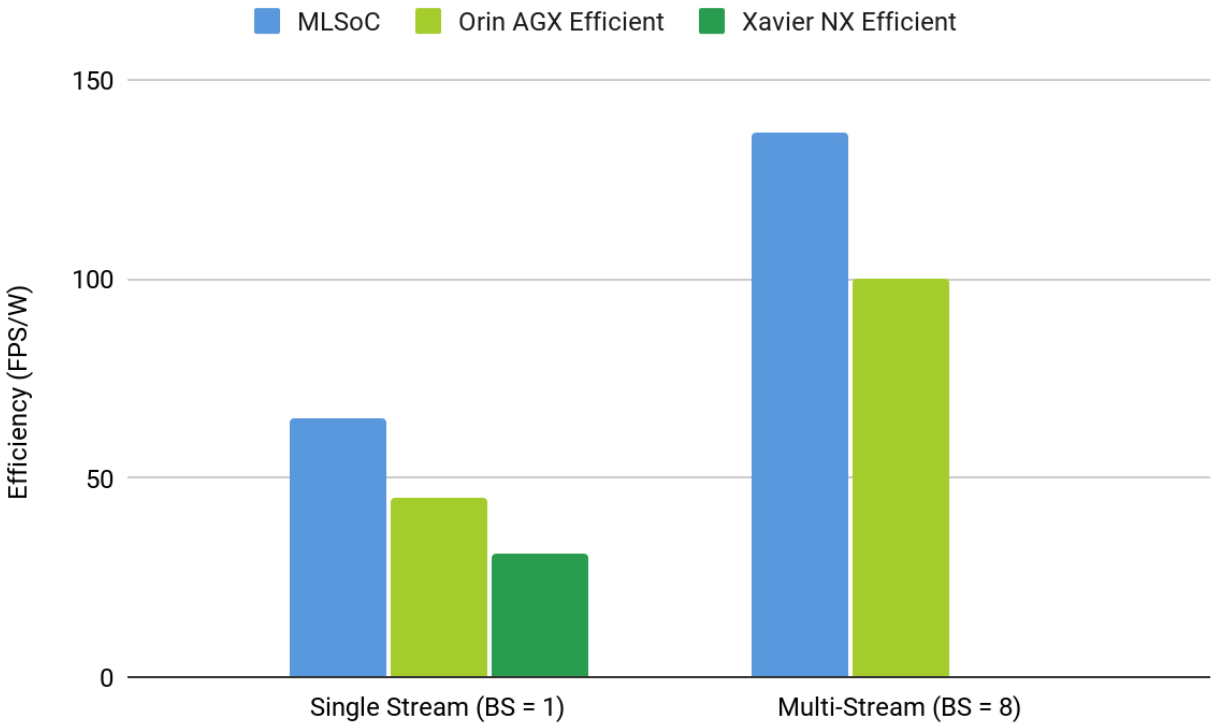
SiMa.ai MLSoC (N16) compiled results unseats Orin (8nm) ***on both performance and power***



Purpose built
Low power + small area



General purpose (GPU)
High power + large area

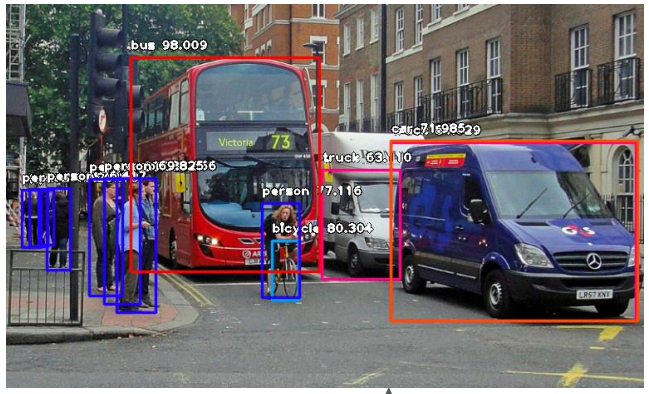


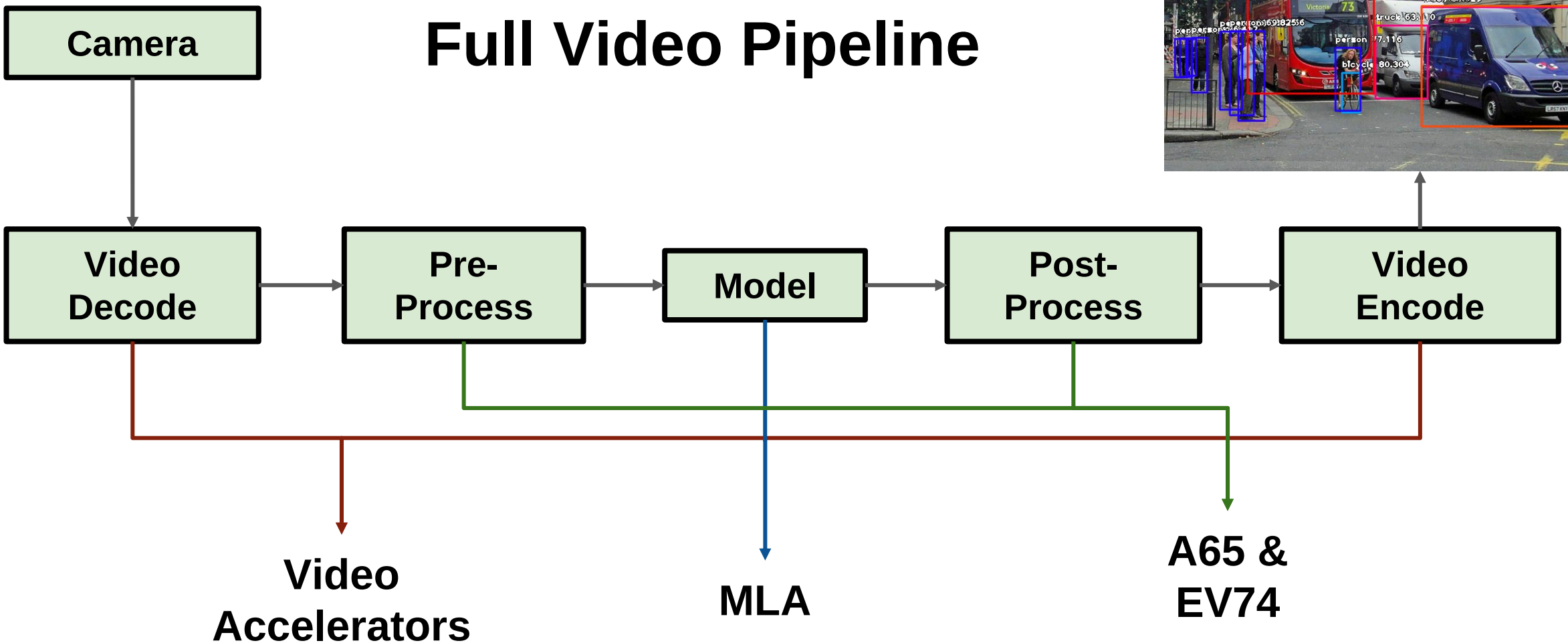
1 Camera
1.4x Orin
2.1x Xavier

8 Cameras
1.37x Orin
Xavier data not published

Link to MLPerf results

Full Video Pipeline





Video Accelerators

Industry Standard Video codecs

MLA

Custom Designed AI Accelerator

A65 & EV74

DSP & General Purpose Execution

Shaheen: An Open, Secure, and Scalable RV64 SoC for Autonomous Nano-UAVs

University of Bologna

L. Valente, A. Veeran, M. Sinigaglia, Y. Tortorella, A. Nadalini, N. Wistoff, B. Sá, A. Garofalo, R. Psiakis, M. Tolba, A. Kulmala,N. Limaye, O. Sinanoglu, S. Pinto, D. Palossi, L. Benini, B. Mohammad, D. Rossi

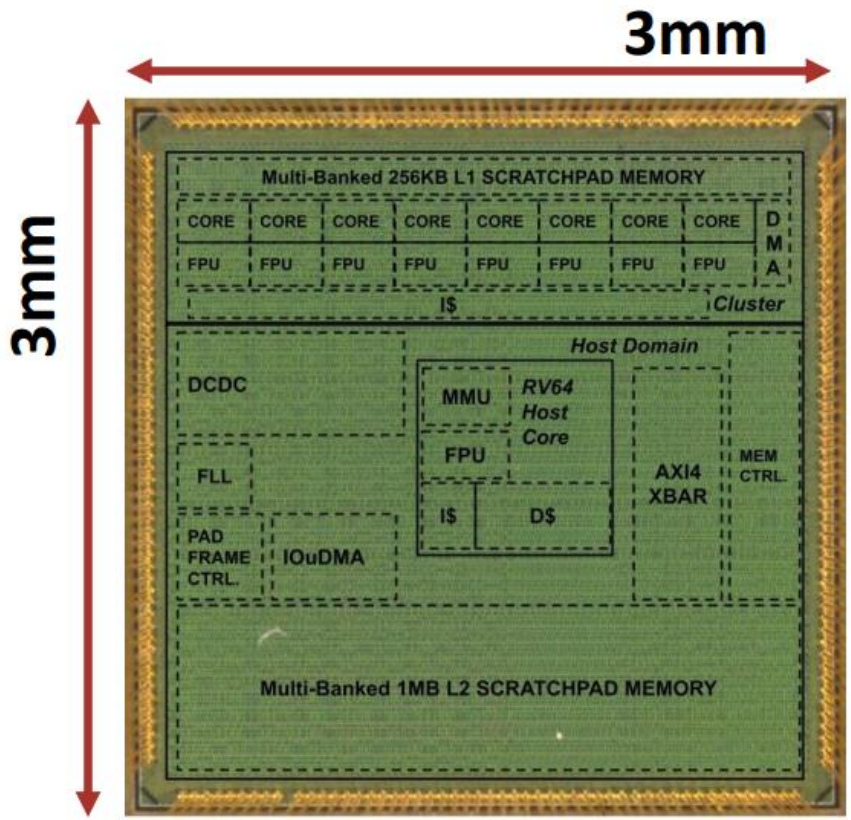
luca.valente@unibo.it

Autonomous Nano-UAVs

- **Versatility, safety, and cost-effet:**
  - small and agile
  - ideal for accessing hard-to-reach areas or tight spaces (inspection/maintenance)
  - relatively inexpensive to produce and operate
- **Requirements for future generation of nano-UAVs:**
  - Run increasingly complex multi-tasking workloads with large memory footprint
  - Within a few hundred mW power budget
  - Support for virtualization and secure operations in uncontrolled/hostile scenarios

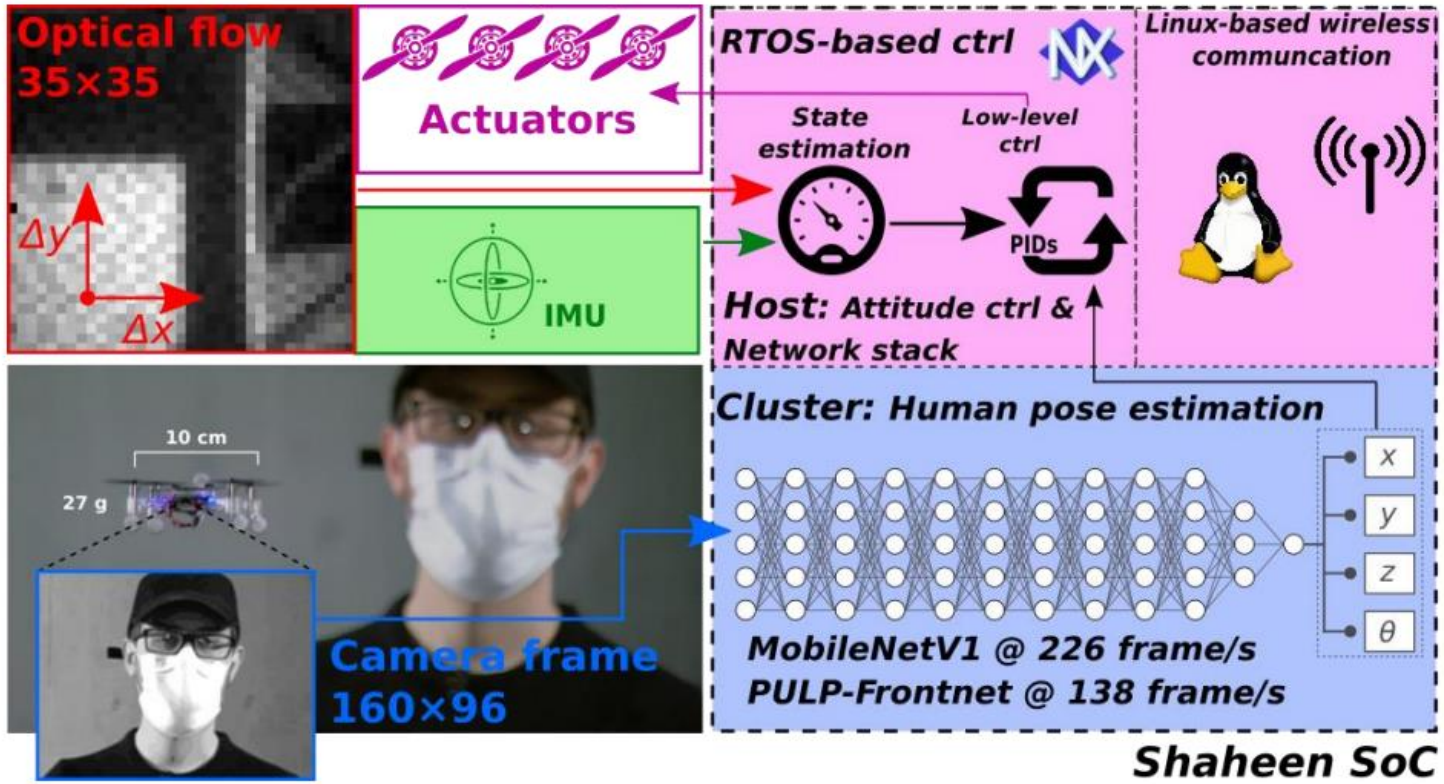
Shaheen: an Open, Secure, and Scalable RV64 SoC for Autonomous Nano-UAVs

- **9mm² SoC in 22nm FDSOI technology with:**
  - A **RV64** Linux-capable CPU enhanced with
    - **Hypervisor** support
    - **Timing-channels mitigation**
  - An **energy efficient** programmable multi-core accelerator (PMCA) based on **8 RV32 cores with ML and DSP extensions**
  - Up to **512MB** of low-power off-chip **main memory**
  - **Logic locking** on key IPs within the architecture
  - **200mW power envelope**



RV64 and custom RV32: the best of both worlds

- **Host:**
  - On top of the **hypervisor**:
    - Attitude control (**RTOS-based**)
    - **Linux-based** legacy software such as wireless network stack.
- **PMCA:**
  - The PMCA runs the **CNN-based** pose estimation task fed by a low-resolution front-looking camera.



Luca Valente luca.valente@unibo.it

Institut für Integrierte Systeme – ETH Zürich
Gloriastrasse 35
Zürich, Switzerland

DEI – Universitá di Bologna
Viale del Risorgimento 2
Bologna, Italy

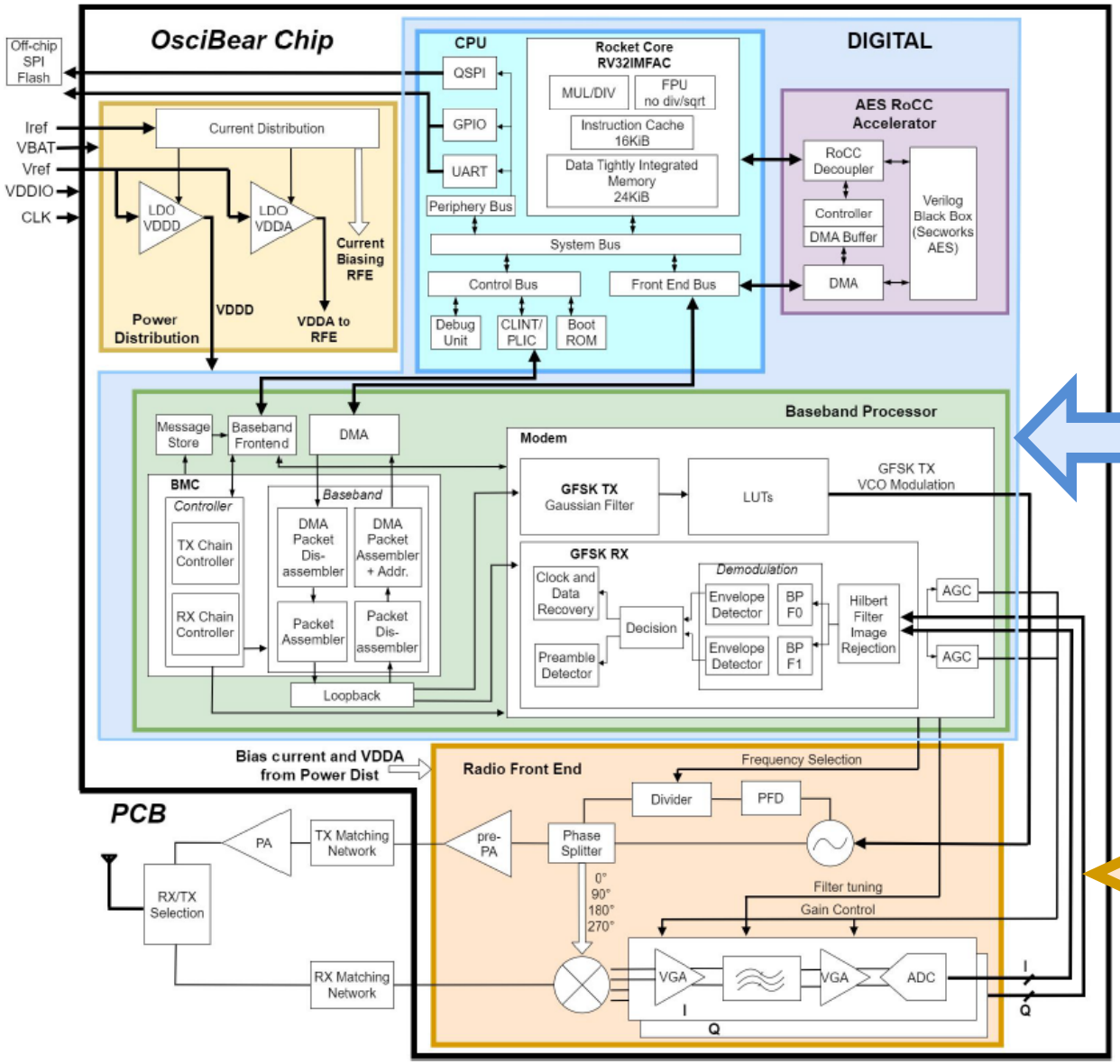
Berkeley
Wireless Research Center

A Heterogeneous SoC for Bluetooth LE in 28nm

Felicia Guo, Nayiri Krzysztofowicz, Alex Moreno, Jeffrey Ni, Daniel Lovell, Yufeng Chi, Kareem Ahmad, Sherwin Afshar, Josh Alexander, Dylan Brater, Cheng Cao, Daniel Fan, Ryan Lund, Jackson Paddock, Griffin Prechter, Troy Sheldon, Shreesha Sreedhara, Anson Tsai, Eric Wu, Kerry Yu, Daniel Fritchman, Aviral Pandey, Ali Niknejad, Kristofer Pister, and Borivoje Nikolic

University of California, Berkeley

Chip in a Semester



- 19 students in 14 weeks
- 1 mm^2 chip in TSMC 28nm
- Berkeley developed tools

CHIPYARD

HAMMER

CHISEL

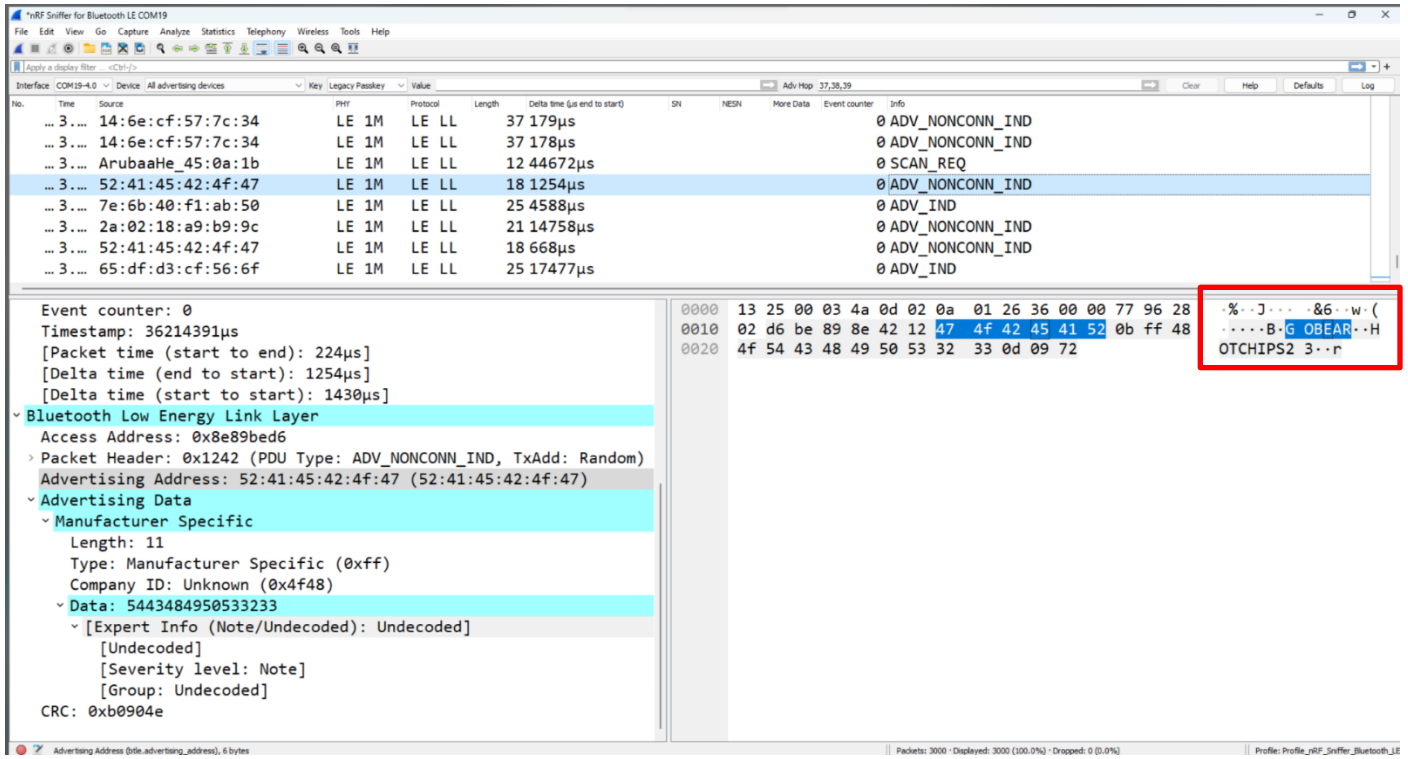
BAG

Berkeley Analog Generator

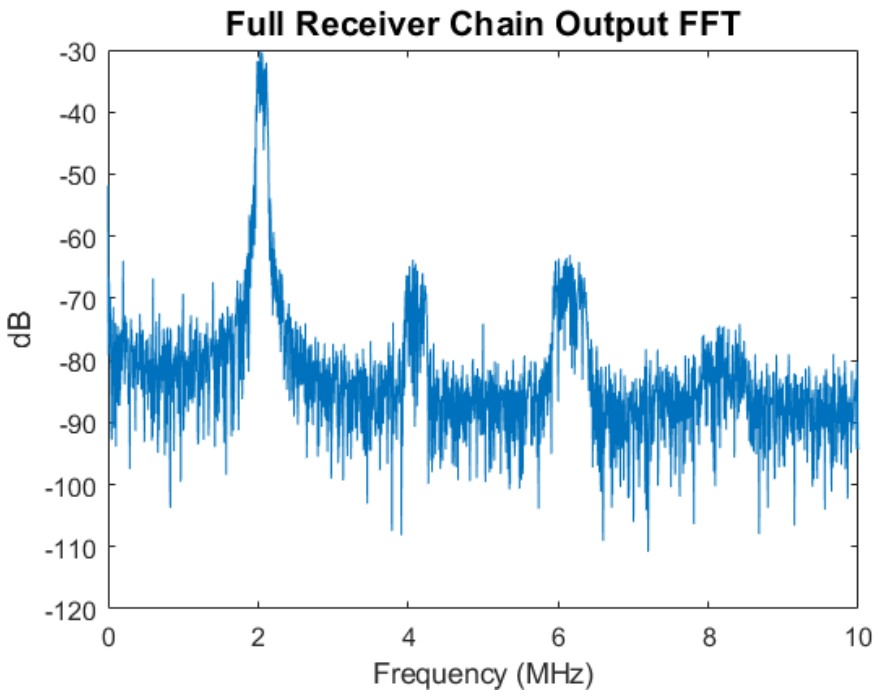
(ADC Only)

BLE Measurement Results

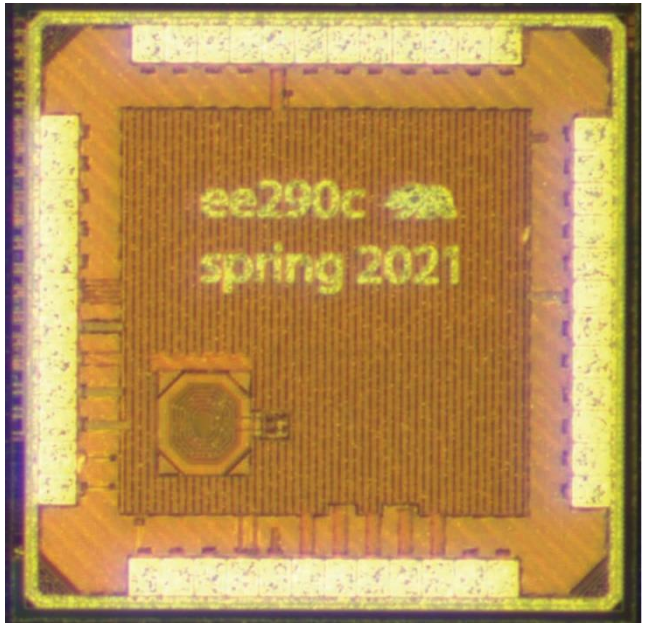
- Transmit packet decoding done two ways:
  1. Software Testing
  2. Commercial (Nordic nRF52840 DK) receiver
- Receiver tested from PCB antenna port to differential output of final VGA (pre-ADC)

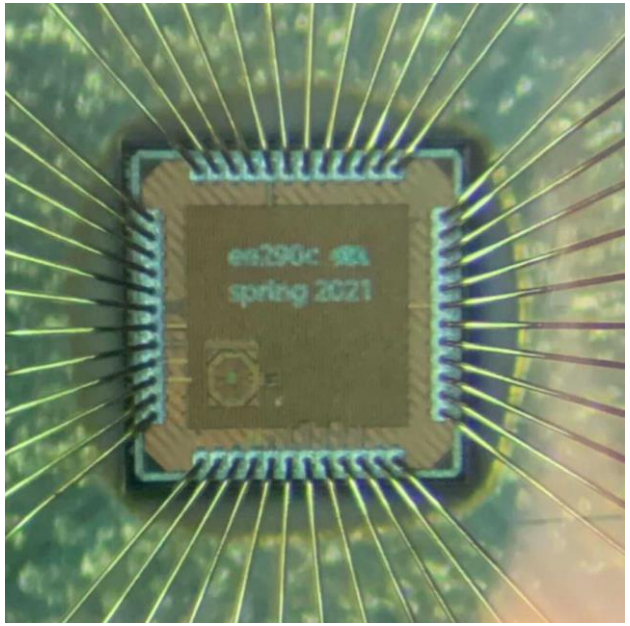


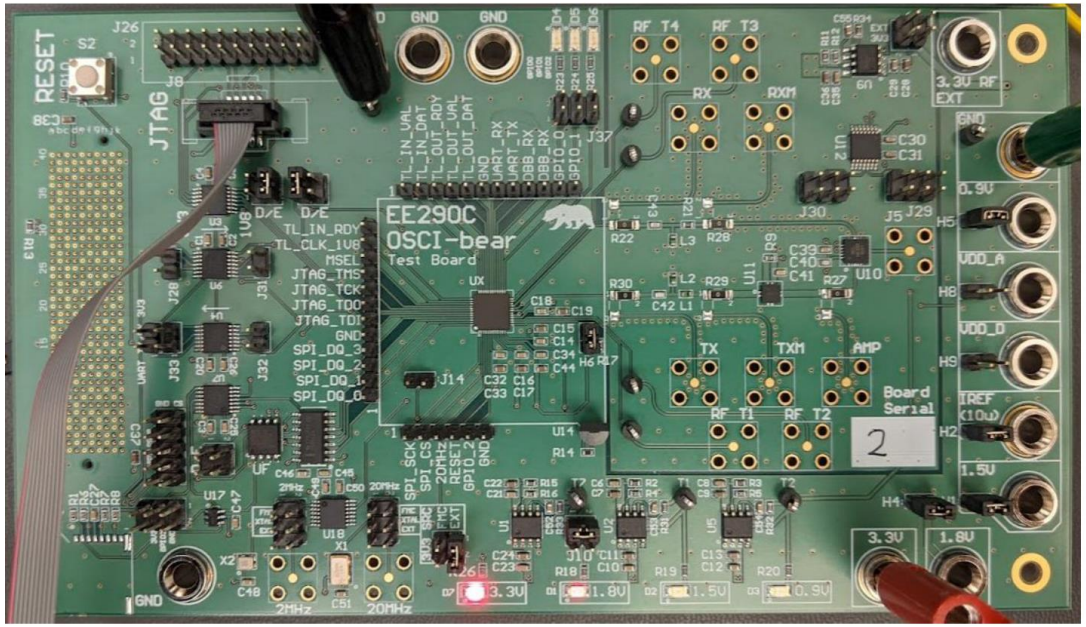
Commercial receiver decoding correctly formed packet sent with "GOBEAR HOTCHIPS23" as payload



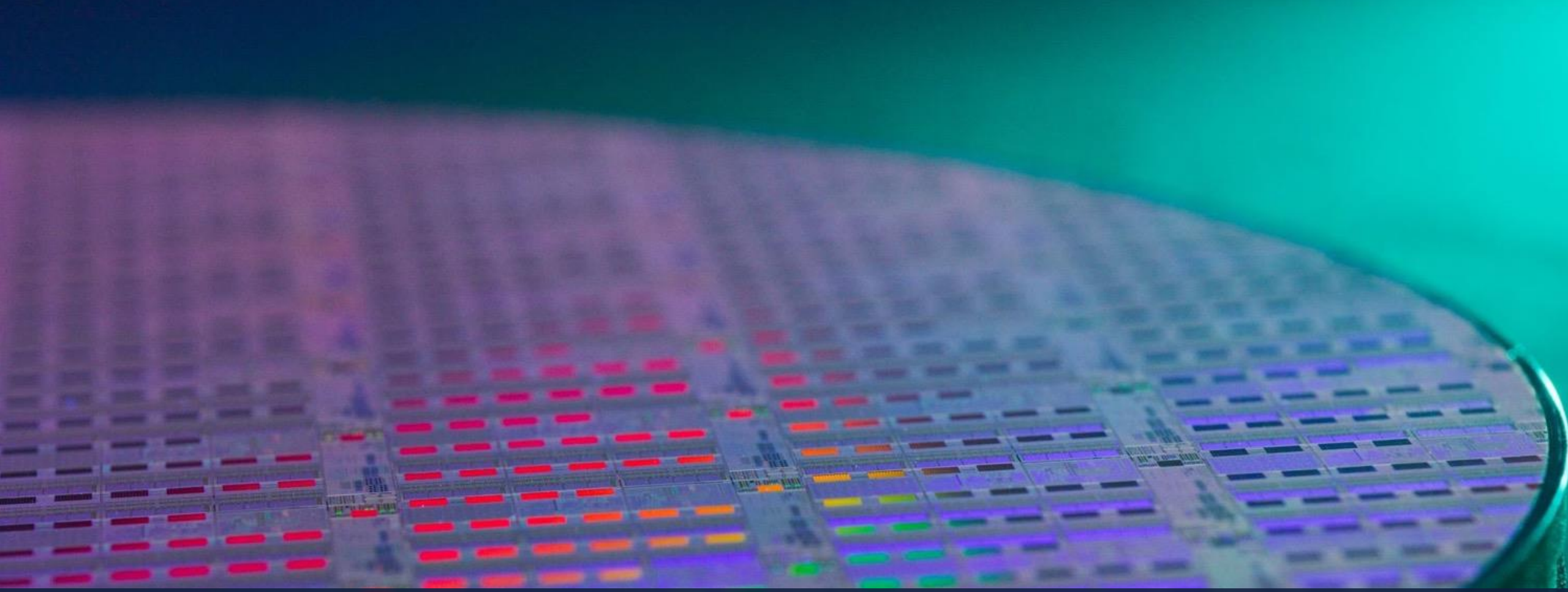
Thank You!







We would like to acknowledge and thank Apple for supporting integrated circuit engineering classes at UC Berkeley EECS department through Apple's New Silicon Initiative program. This work is also funded by NSF CCRI ENS Chipyard Award #2016662.



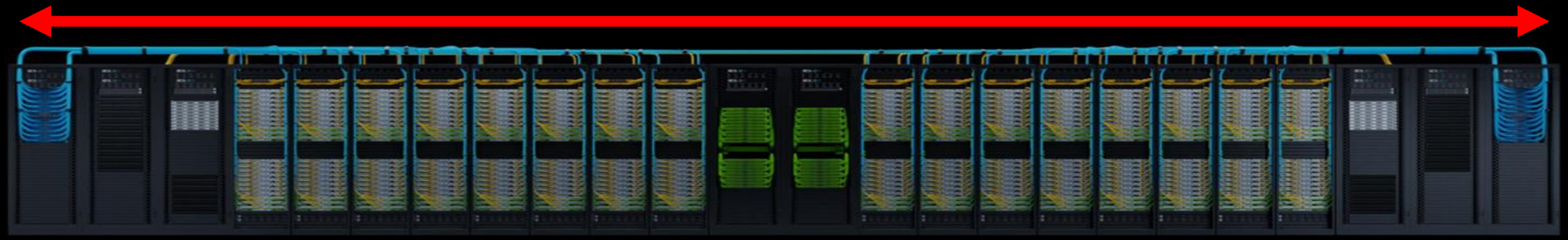
Driving Compute Scale-out Performance with Optical I/O Chiplets in Advanced System-in-Package Platforms

Mark Wade, PhD | President, CTO, Co-Founder | August 28, 2023

Hundreds/Thousands of Sockets Computing as One

Model Size >> Node Size → Distribute across many nodes → Interconnect fabric becomes major bottleneck → Optical I/O Needed (reach, power, latency)

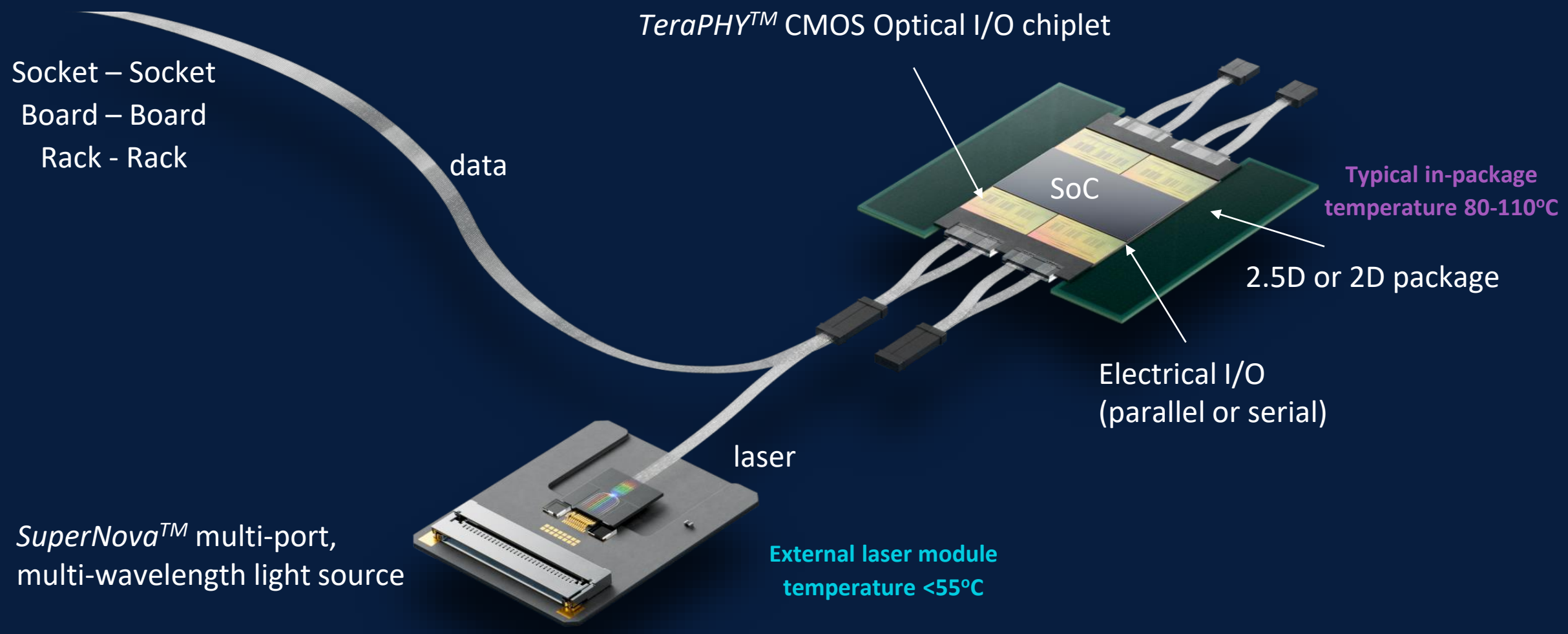
15 meters (50 feet)



DGX GH200 SuperPOD (Source: Nvidia)

NVidia DGX GH200: 256 GPUs acting as 1 “Mega GPU”

Ayar Labs Optical I/O (OIO) Chiplet

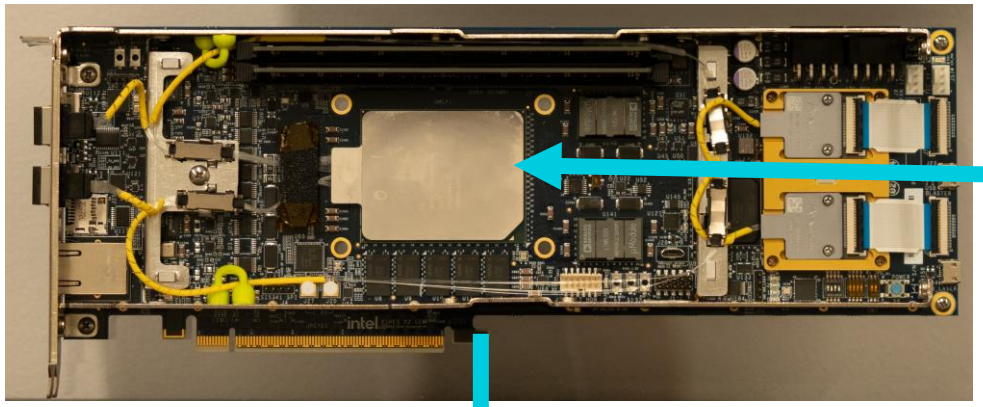


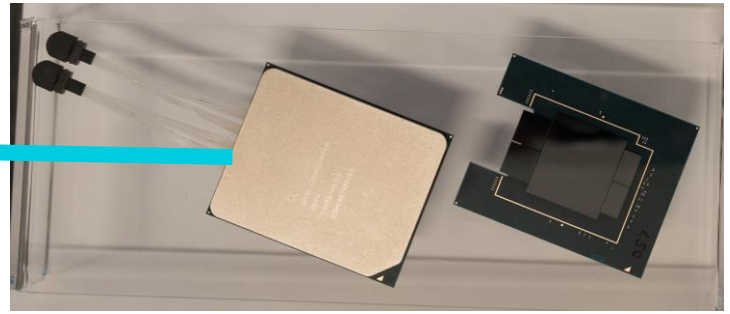
Optical FPGA PCIe Card

2 x 4Tbps Optical I/O Chiplets

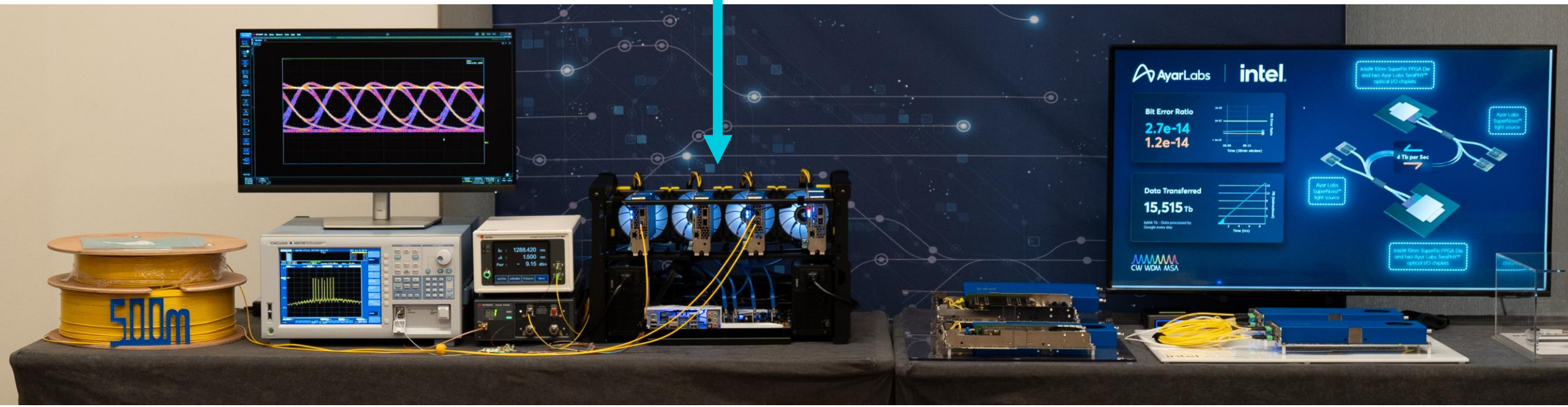
<1e-12 BER pre-FEC

10ns + TOF latency





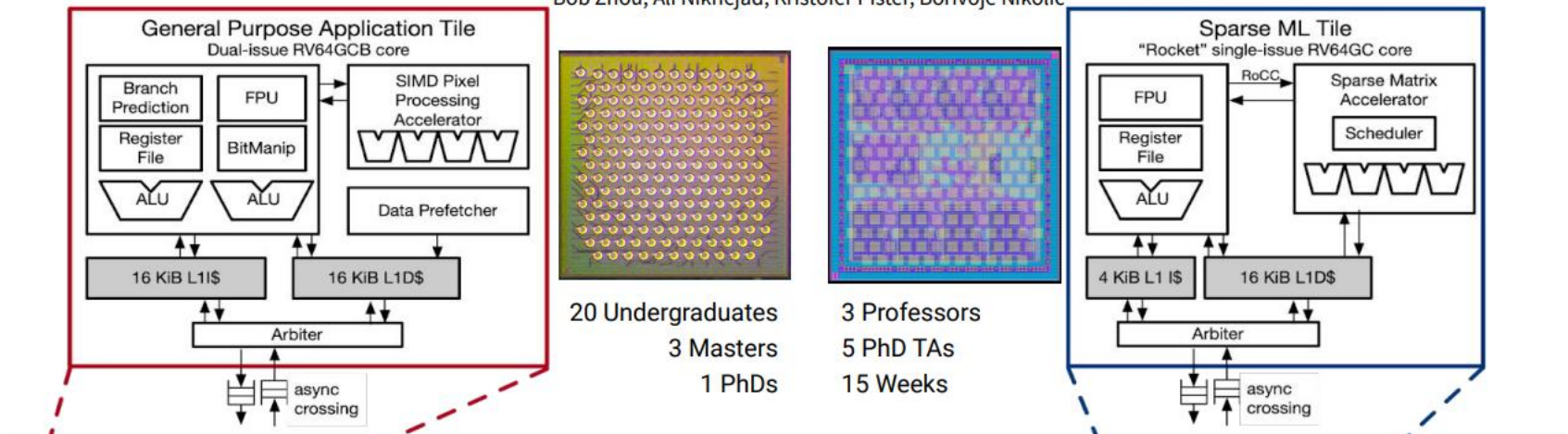
Running Live Outside



Optical I/O Enables Fully Disaggregated Distributed Compute

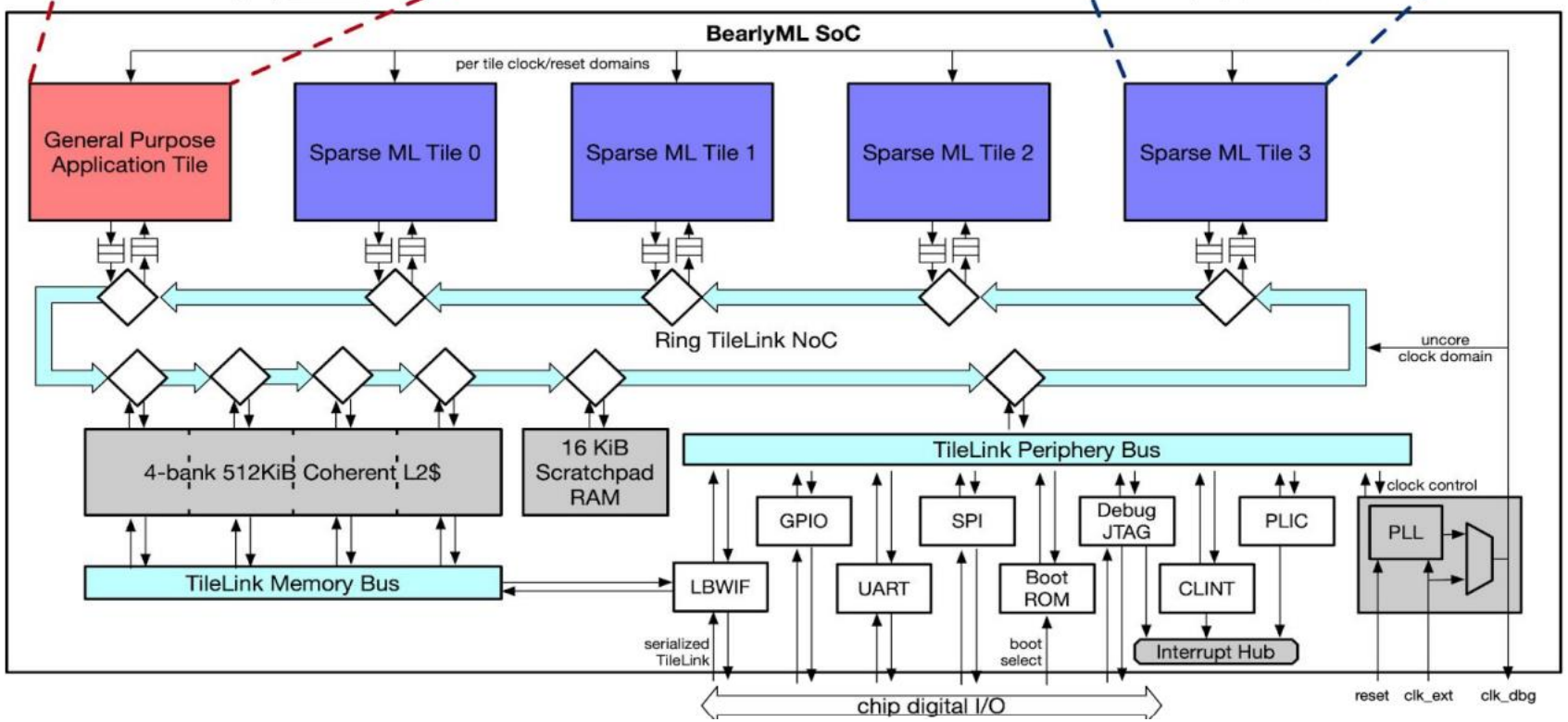
A Heterogeneous RISC-V SoC for ML Applications in Intel 16 Technology

Yufeng Chi, Franklin Huang, Raghav Gupta, Ella Schwarz, Jennifer Zhou, Reza Sajadiany, Animesh Agrawal, Max Banister, Michelle Boulos, Jason Chandran, Jessica Dowdall, Leena Elzeiny, Claire Gantan, Anthony Han, Roger Hsiao, Chadwick Leung, Edwin Lim, Jose Rodriguez, Tushar Sondhi, Mitchell Twu, Rongyi Wang, Mike Xiao, Ruohan Yan, Paul Kwon, Zhaokai Liu, Jerry Zhao, Bob Zhou, Ali Niknejad, Kristofer Pister, Borivoje Nikolić

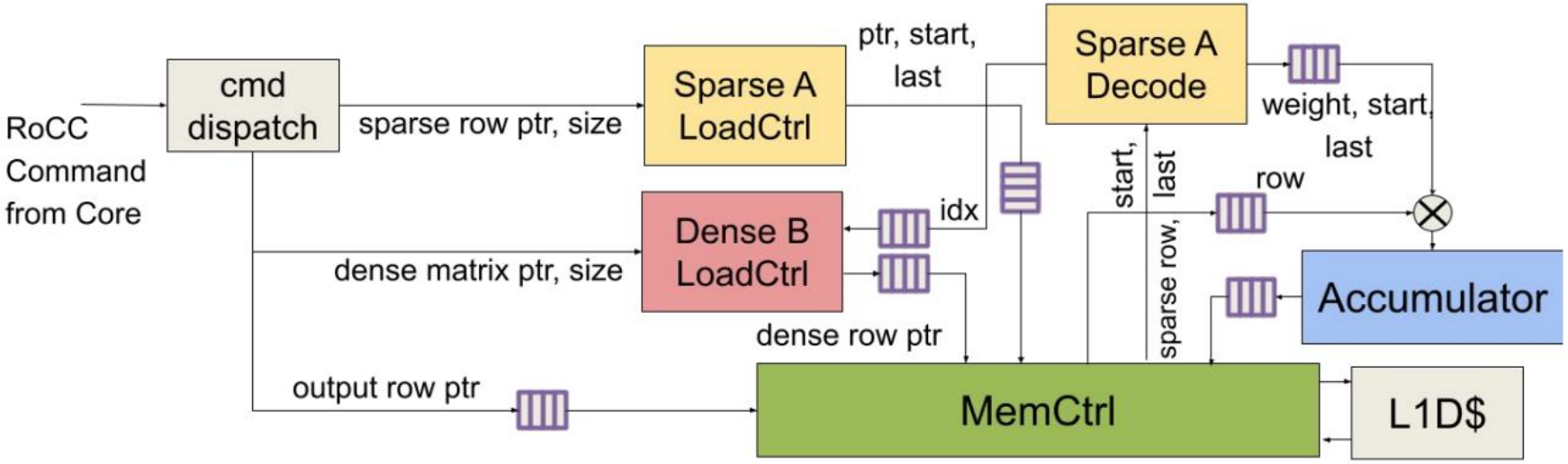


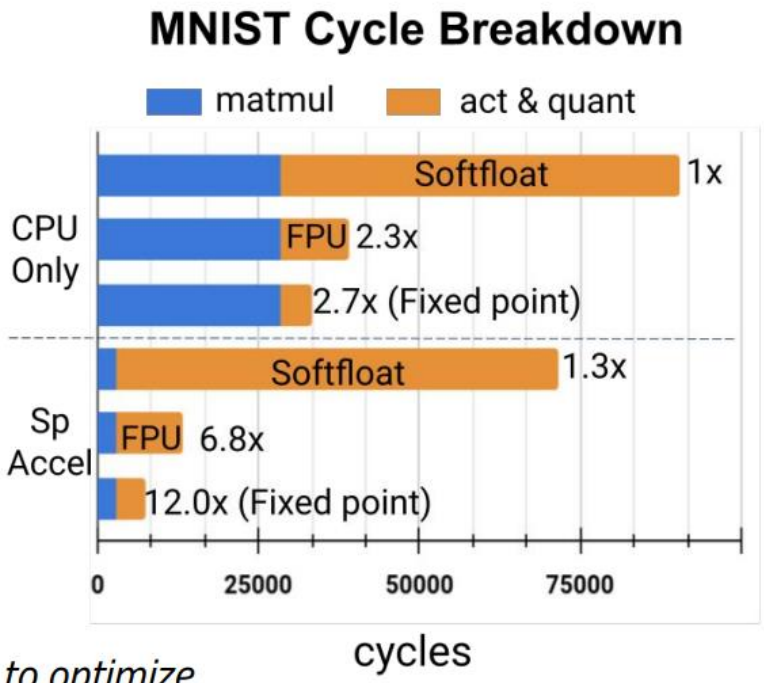
20 Undergraduates
3 Masters
1 PhDs

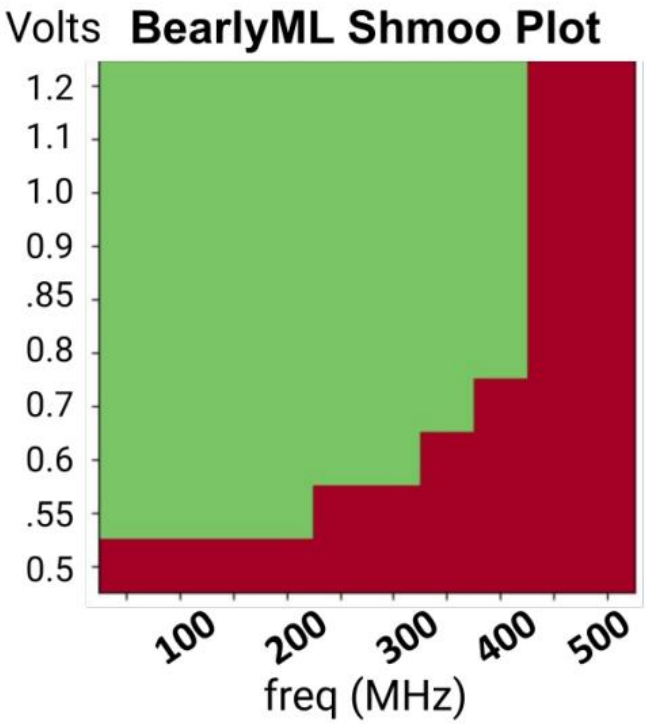
3 Professors
5 PhD TAs
15 Weeks



A Heterogeneous RISC-V SoC for ML Applications in Intel 16 Technology







don't forget to optimize non-matmul operators!

End of Poster Lightning Talks