Poster Lightning Talks

Rob Aitken, Session Chair

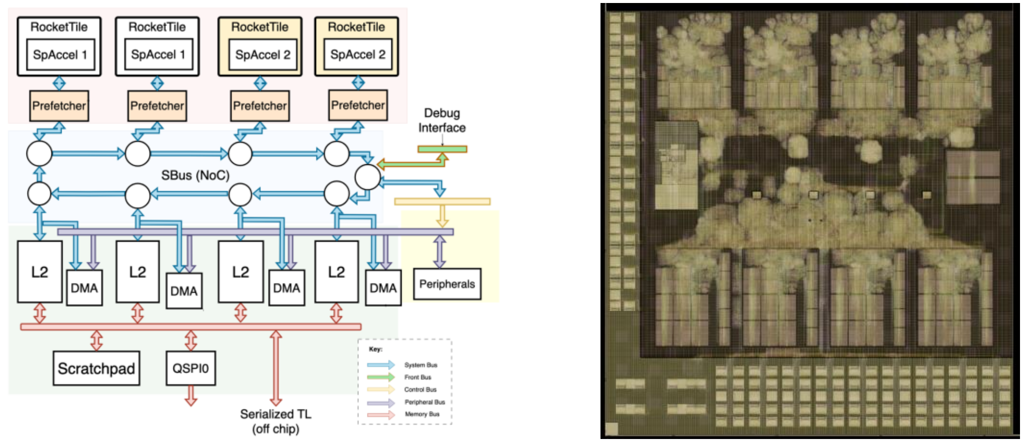
NeCTAr and RASoC: Intel 16 SoCs for Language Model Interference and Robotics

Viansa Schmulbach

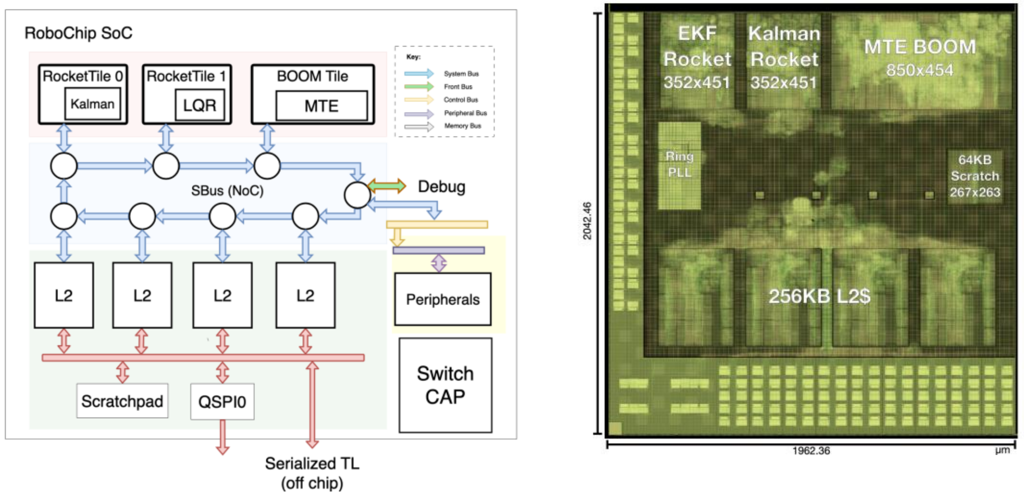
University of California, Berkeley

Chips At-A-Glance

- Chip design needs to be agile in response to rapidly evolving machine learning requirements
- Two 16nm heterogeneous system-on-chip designs developed from concept to chip tapeout in **<15 weeks** by a team of mostly undergraduate students with **no prior chip design experience**
- NeCTAR: **Near Cache Transformer Accelerator**
  - RISC-V SoC optimized for ML applications
  - Near-memory compute engines
  - CPU-coupled sparse matrix accelerators with L1 and L2 cache access
  - L2 data cache best-offset prefetchers
- RASoC: **Robotics Application System-on-Chip**
  - RISC-V SoC suitable for controlling robotics systems
  - Kalman filter and LQR accelerators
  - Improves software security by adding a new Memory Tagging Extension to a medium BOOM tile
  - Digitally controlled switched capacitor voltage regulator



NeCTAr block diagram & Die shot



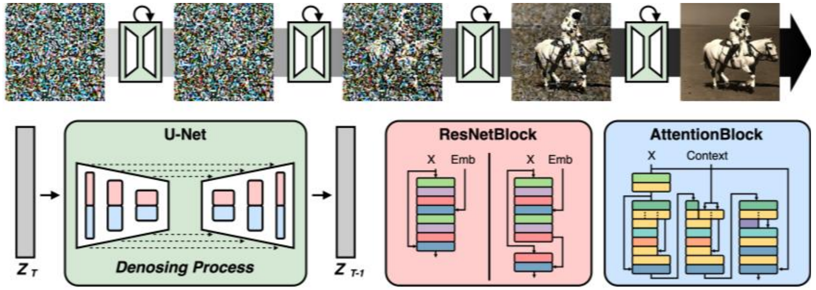
RASoC block diagram & Die shot

Picasso: An Area/Energy-Efficient End-to-End Diffusion Accelerator with Hyper-Precision Data Type

Sungyeob Yoo
KAIST

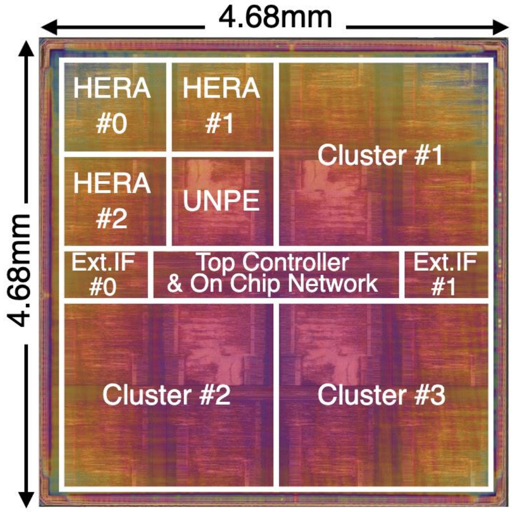
Picasso: Diffusion Accelerator

- Challenge
  - Accelerating **diffusion models** requires carefully balancing accuracy and bit precision in **low-bit quantization**, and effectively addressing latency in **non-matrix operations**.



Diffusion Model Overview

- Picasso
  - A 28nm end-to-end diffusion accelerator designed to enhance hardware efficiency without sacrificing accuracy.
  - Features include:
    - Hyper-Precision Data Type (**HYP8**)
    - Hyper-Efficient Reconfigurable Arrays (**HERA**)
    - Unified Non-Matrix Processing Engine (**UNPE**)



Chip Photograph

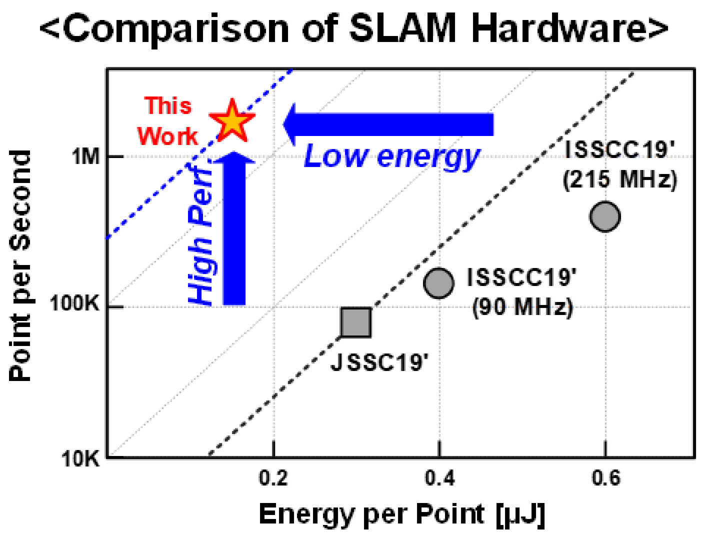
- Learn More
  - Join us at the poster session to explore how **Picasso achieves up to 26.8x better performance and 30.5x better area efficiency** over prior accelerators.

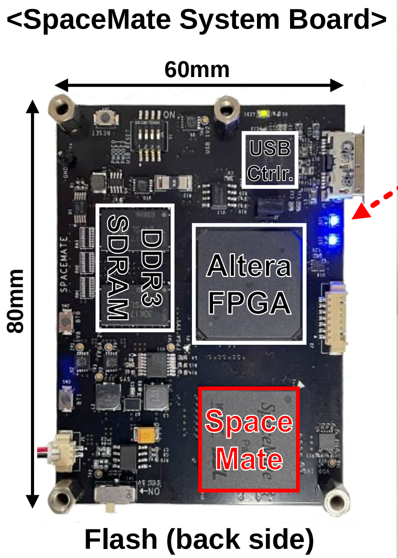
Space-Mate: A 303.5mW Real-Time NeRF SLAM Processor with Sparse-Mixture-of-Experts-based Acceleration

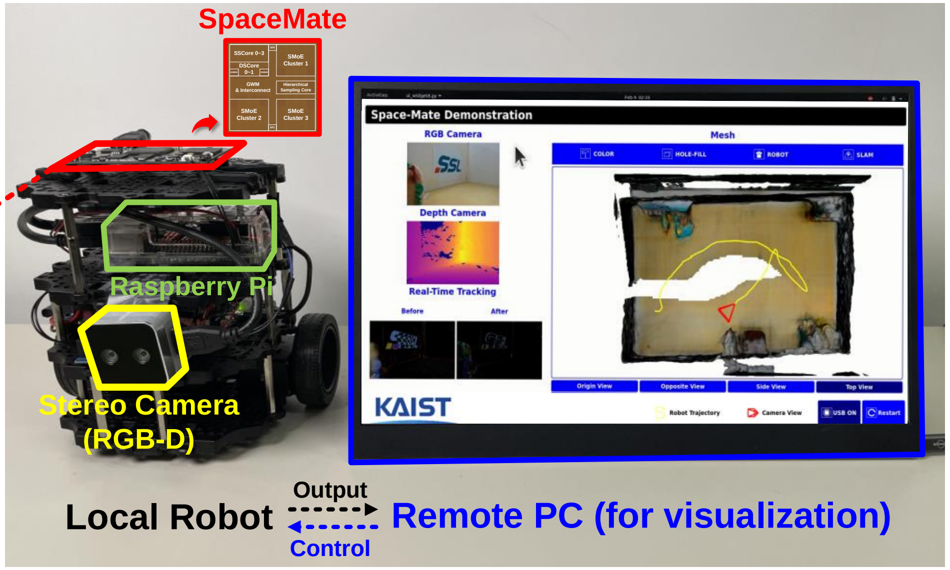
Gwangtae Park
School of EE, KAIST

Summary of Space-Mate

- HW-SW co-optimization to resolve NeRF-SLAM's **computation bottleneck**
- SW sol. : batch-level **conditional computing** w/ Sparse Mixture-of-Experts
- HW sol.: out-of-order dataflow to **handle irregular batch access**
- SOTA performance & demonstrated w/ open-source robot platform







RISC-V-based System-on-Chips for IoT Applications

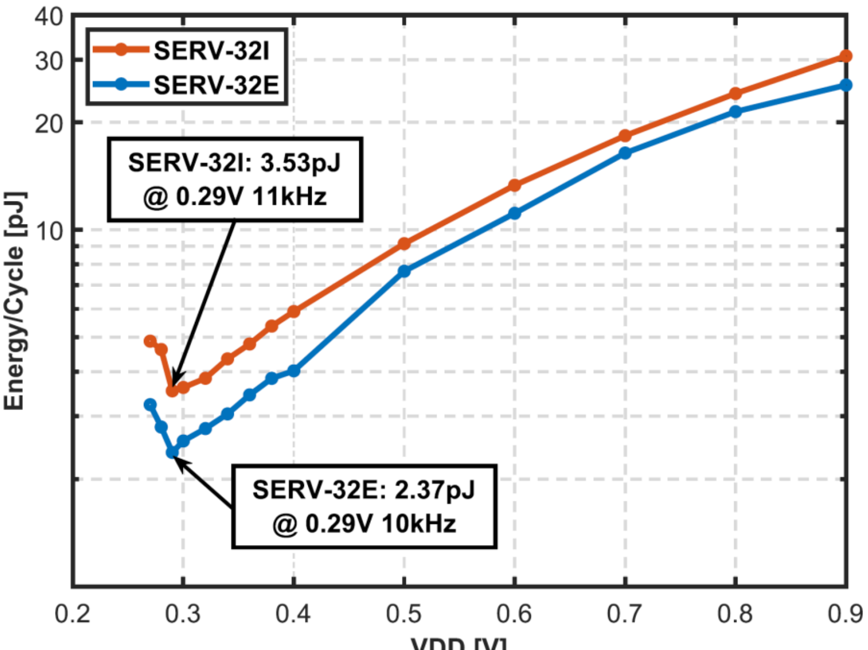
Khai-Duy Nguyen, Tuan-Kiet Dang, Binh Kieu-Do-Nguyen, Cong-Kha Pham, and Trong-Thuc Hoang

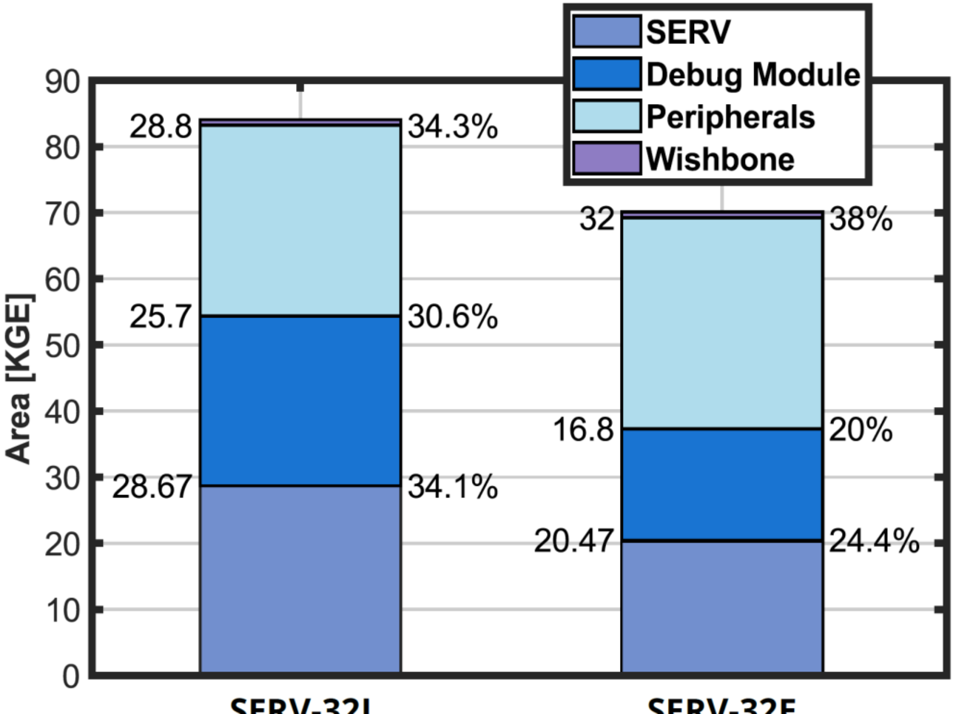
The University of Electro-Communications (UEC), Tokyo, Japan

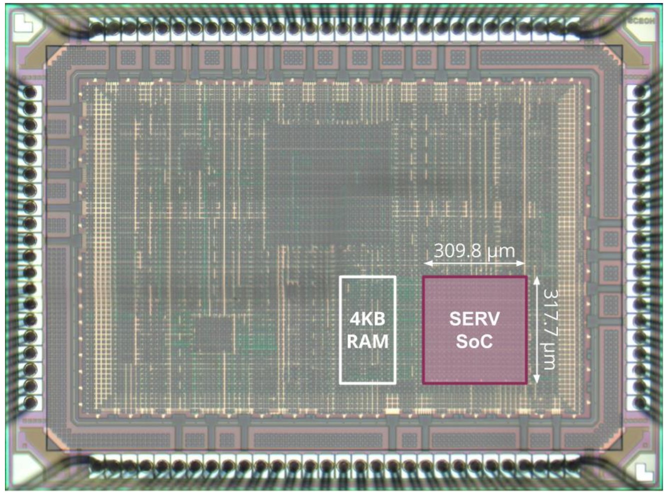
Problem: System-on-Chip for battery-less applications

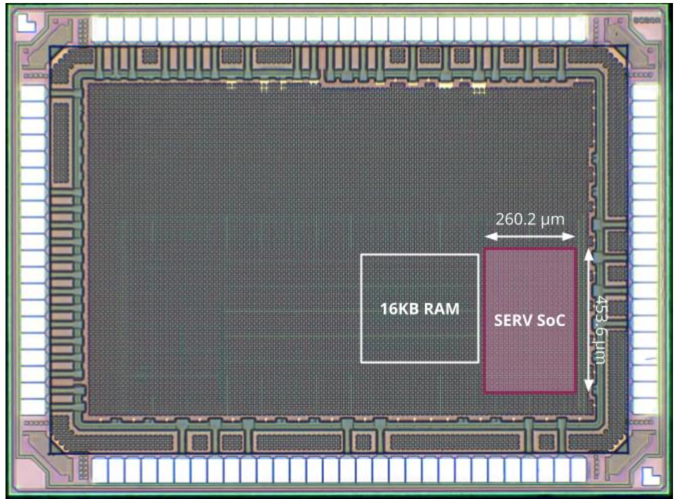
Solutions:

- SoC with bit-serial microprocessor
- Fabricated in low-power SOTB 65nm process









A Trusted Execution Environment RISC-V System on Chip

Binh Kieu-Do-Nguyen

University of Electro-Communications (UEC), Tokyo, Japan

Pham Laboratory
Integrated circuit design laboratory

A Trusted Execution Environment RISC-V System on Chip

Problems: RISC-V puts more challenges on security issues.

- Popularity → need to protect sensitive data.
- Openess → face to diverse threats from third-party IP.

Solution: Isolate sensitive data from rich environment.

- Isolated sub-system performs boot process.
- Secured key management scheme for boot procedure.
- Crypto-accelerators for security operations.

Join us for a stimulating discussion and a chance to delve deeper into the implications of this research.

CogniVision: A mW Power envelope SoC for Always-on Smart Vision in 40nm

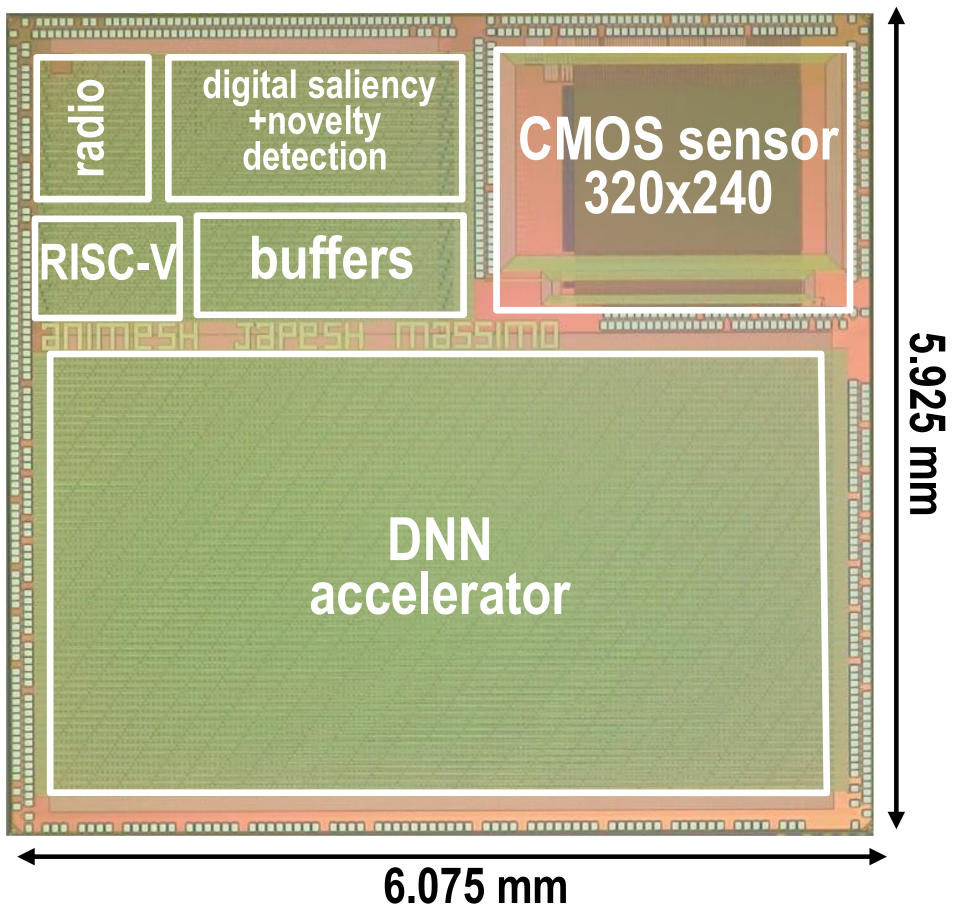
Animesh Gupta*, Japesh Vohra* and Massimo Alioto,
National University of Singapore (* Equal Contributing Authors)

CogniVision

End-to-end hierarchical execution from imager to comms
➔ early gating of readout/compute for true system power reduction

Cognitive system (AI) at low activity
➔ traditionally power-hungry DNN pushed to minor power contribution

Attentive system: software-programmable + wake-up receiver
➔ system can be updated over-the-air for many-camera coordination at scale
(vision settings, DNN)



Visit our poster session for more details!

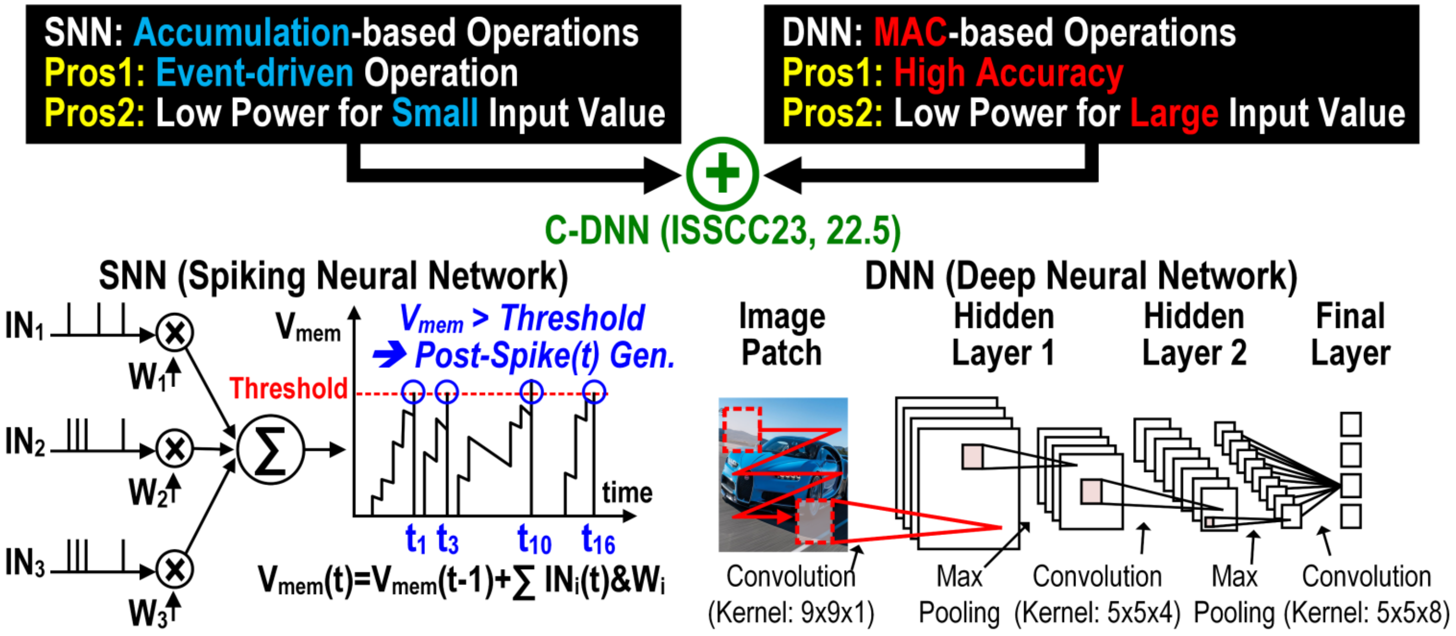
A Low-power Large-Language-Model Processor with Big-Little Network and Implicit-Weight-Generation for On-device AI

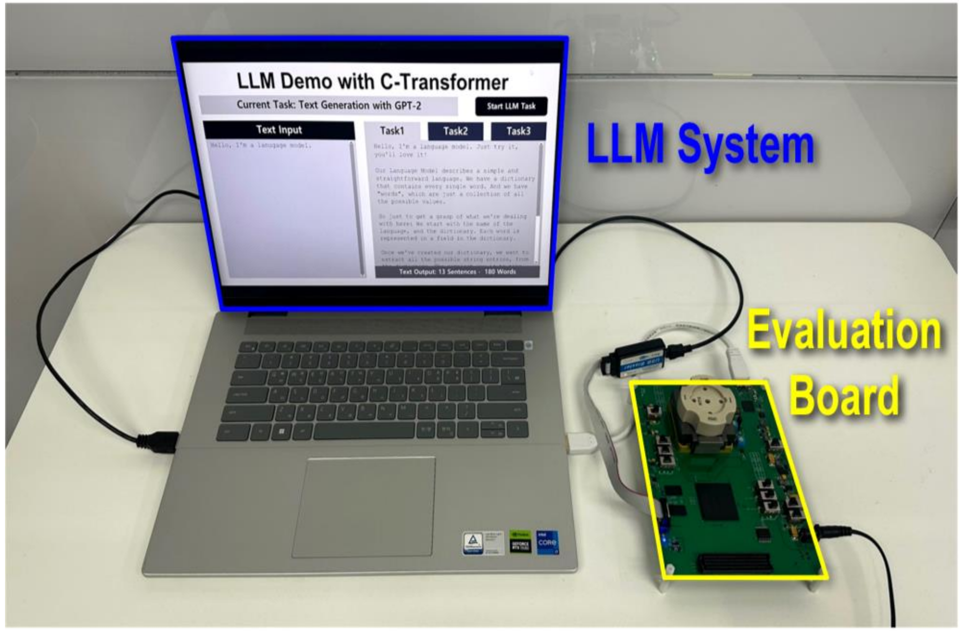
Sangyeob Kim, Sangjin Kim, Wooyoung Jo, Soyeon Kim, Seongyon Hong, Nayeong Lee, and Hoi-Jun Yoo

KAIST, Daejeon, Republic of Korea

Energy-efficient LLM System

- Reducing computations and parameters of large language model (LLM)
- Solution1: Adopting neuromorphic computing for LLM
- Solution2: Using Big-little Model and Implicit-Weight Generation





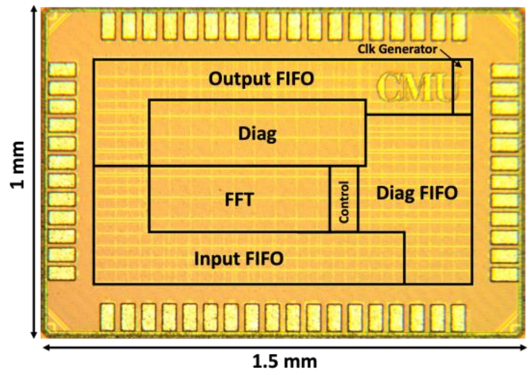
A 1.19GHz 9.52Gsamples/sec Radix-8 FFT Hardware Accelerator in 28nm

Larry Tang

Carnegie Mellon University

Towards an Automated FFT Accelerator Design Methodology

- **Problem**
  - FFT hardware accelerators are typically designed as standalone designs
  - Relatively fixed functionality, challenging to integrate into systems
- **Approach**
  - Hardware accelerate only a software FFTW ‘codelet’ to enable flexibility
  - Software stack integration and hardware generation using SPIRAL
  - Test chip: A Radix-8 Twiddle Codelet Accelerator in a TSMC 28nm process
- Feel free to stop by the poster for any discussion or Q&A



PACE: A Scalable and Energy Efficient CGRA in a RISC-V SoC for Edge Computing Applications

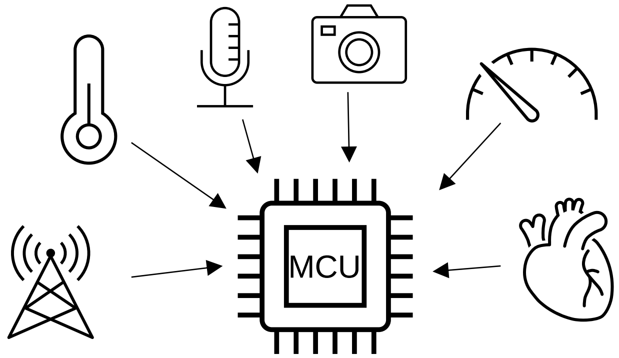
Vishnu P. Nambiar[1], Yi Sheng Chong[1], Thilini Kaushalya Bandara[2], Dhananjaya Wijerathne[2], Zhaoying Li[2], Rohan Juneja[2], Li-Shiuan Peh[2], Tulika Mitra[2], Anh Tuan Do[1]

[1]Institute of Microelectronics, Agency for Science, Technology and Research (A*STAR), Singapore
[2]School of Computing, National University of Singapore, Singapore

[Poster] *PACE: A Scalable and Energy Efficient CGRA in a RISC-V SoC for Edge Computing Applications*

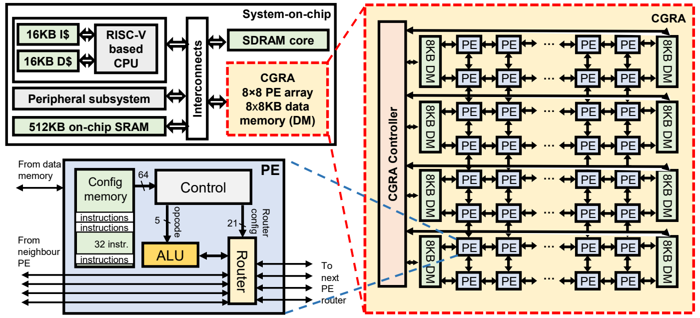
Problems

- Edge devices have lots of **sensors**
- A **CPU** is versatile yet **low efficiency**



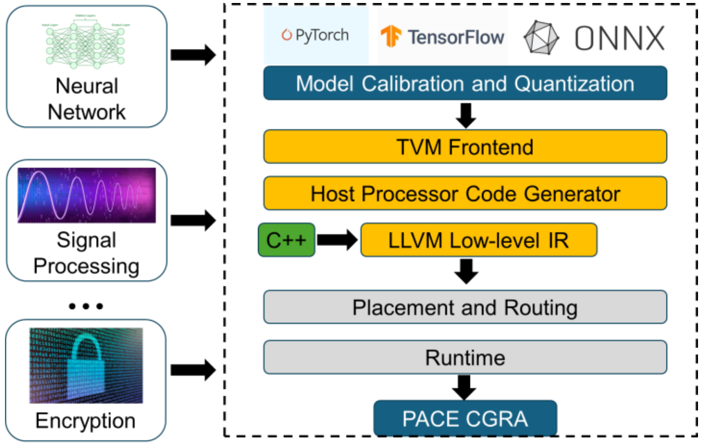
Solutions

- **CGRAs act as co-processors** to offload tasks and increase efficiency

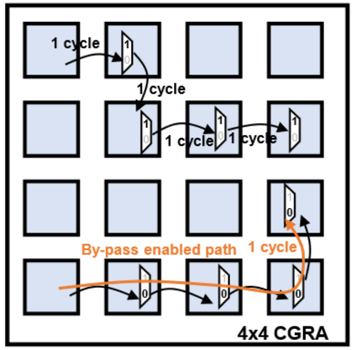


Major contributions

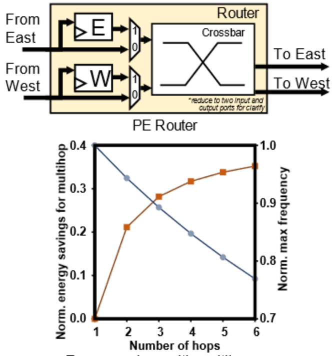
1. End-to-end **software toolchain** to various workloads



2. **Data multi-hop** to improve efficiency in hardware and mapping



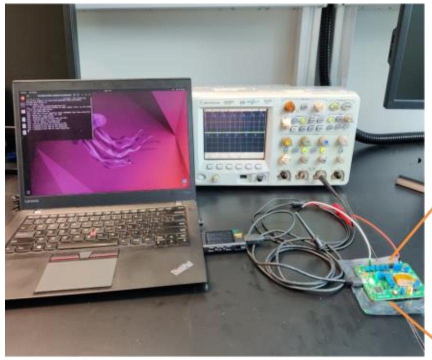
Flit data transfer in one cycle using bypath pass



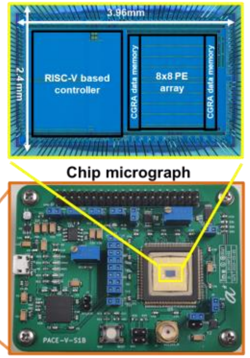
Energy savings with multihop

Results

- The proposed CGRA attains peak measured **efficiency of 360GOPS/W**, which is 1.2 to 4.6 times higher

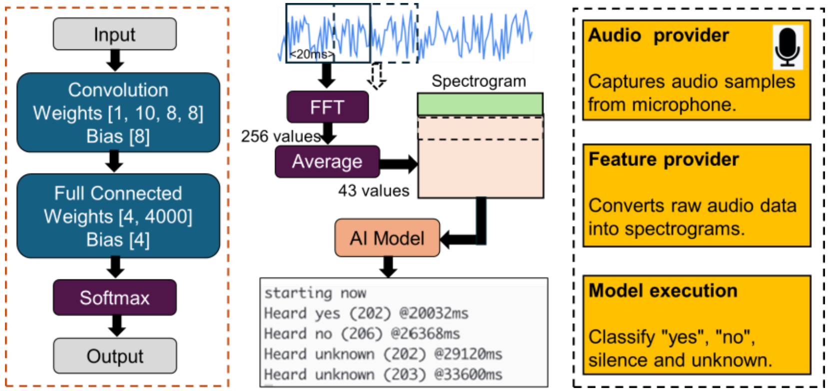


Test laptop, oscilloscope and the PCB board



PACE assembled on PCB board

- **Demonstration**: A keyword spotting application is run using the proposed CGRA



Visit our poster and demonstration to learn more! Thank you.

LSPU: A 20.7ms Low-latency Point Neural Network-based 3D Perception and Semantic LiDAR SLAM System-on-Chip for Autonomous Driving System

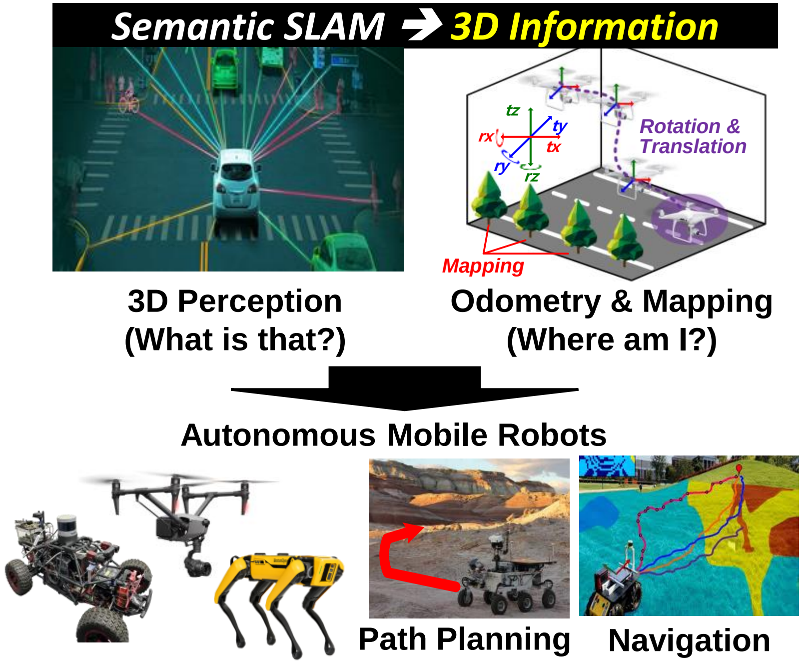
Jueun Jung[1], Seungbin Kim[1], Bokyoung Seo[1], Wuyoung Jang[1], Sangho Lee[1], Jeongmin Shin[1], Donghyeon Han[2], and Kyuho Jason Lee[1]

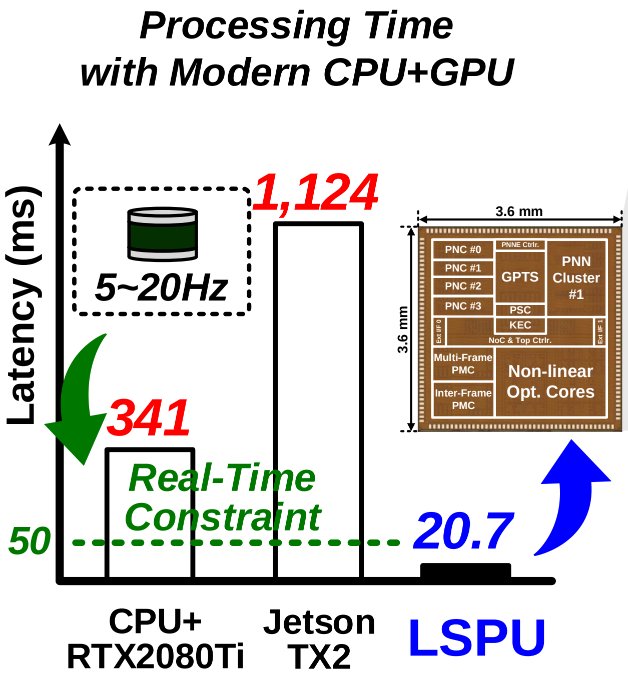
[1]Intelligent Systems Lab., Department of EE, UNIST, Korea

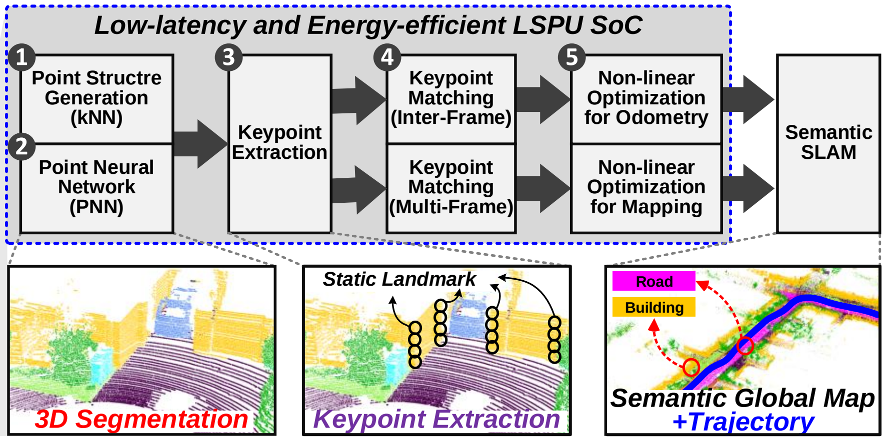
[2]Department of EECS, MIT, USA

LSPU: End-to-end Semantic SLAM SoC

- Semantic LiDAR SLAM is essential component for advanced autonomous driving, but CPU+GPU **fails real-time processing with LiDAR (<50ms).**
- Our Chip is the first semantic LiDAR SLAM processor (LSPU) **integrating point neural network (PNN)-based 3D perception into SLAM pipeline** with **5 heterogeneous accelerators** to fully support each algorithm.







➔ ***A Mobile Semantic SLAM Processor with 349.6mW and 20.7ms Low-latency***

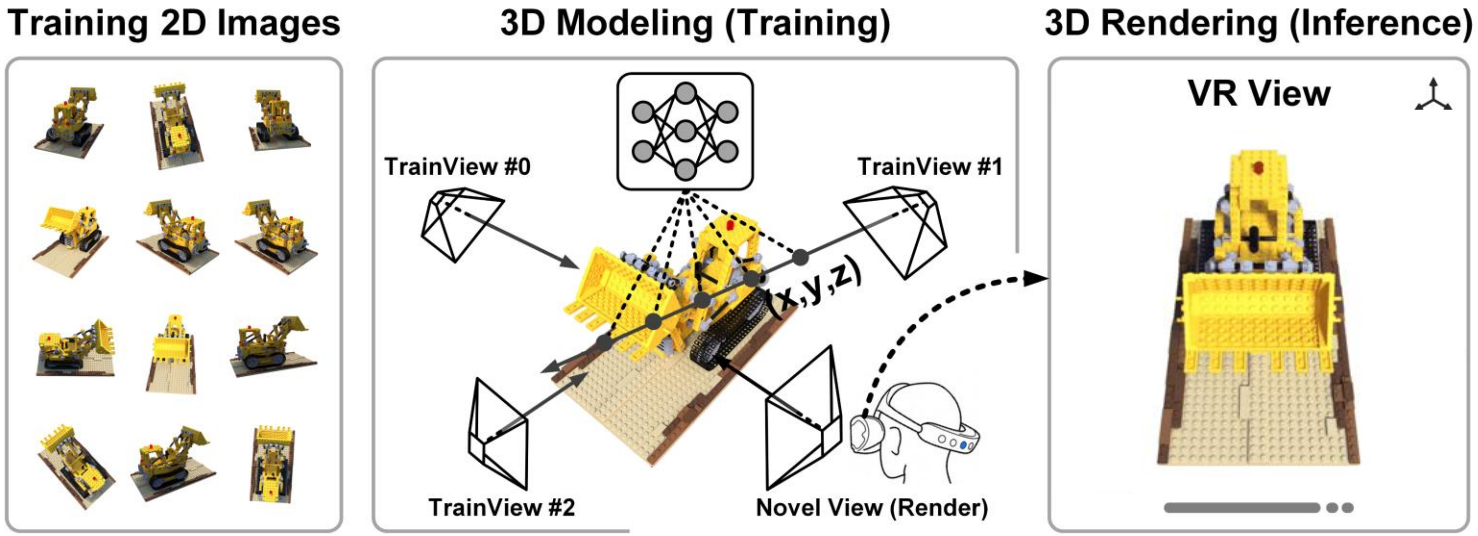
NeuGPU: A Neural Graphics Processing Unit for Instant Modeling and Real-Time Rendering on Mobile AR/VR Devices

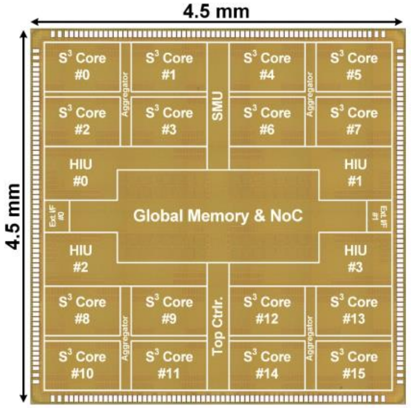
Junha Ryu[1], Hankyul Kwon[1], Wonhoon Park[1], Zhiyong Li[1], Beomseok Kwon[1], Donghyeon Han[2], Dongseok Im[1], Sangyeob Kim[1], Hyungnam Joo[1], Minsung Kim[1], and Hoi-Jun Yoo[1]

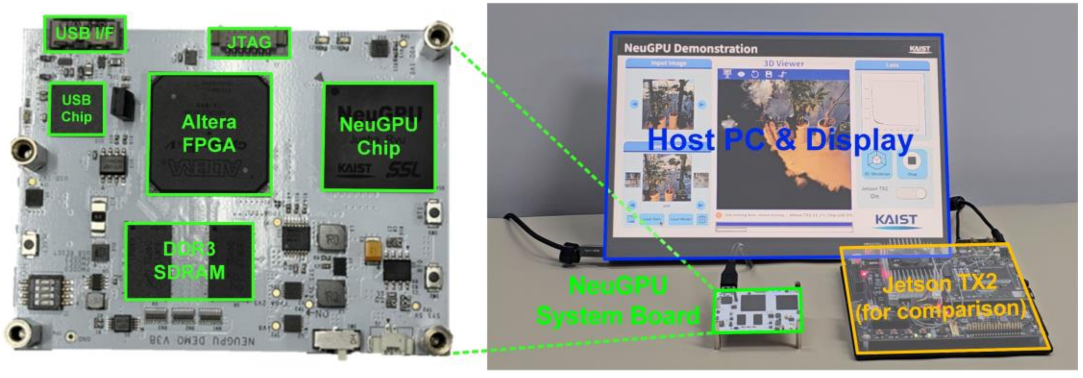
[1]School of EE, Korea Advanced Institute of Science and Technology
[2]Dept. of EECS, Massachusetts Institute of Technology

NeuGPU: Neural Rendering Accelerator

Overview of Proposed System

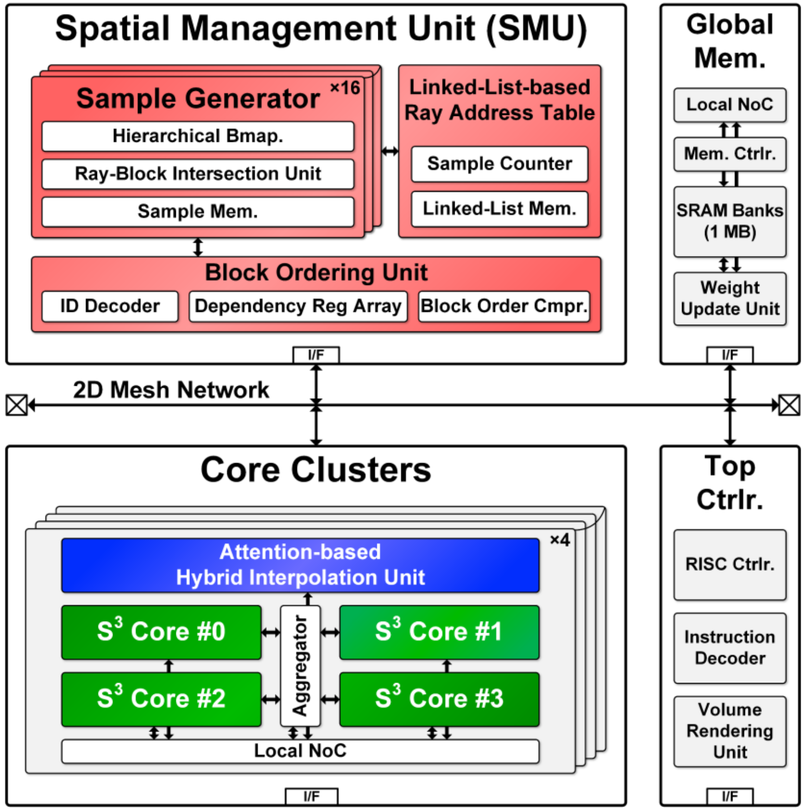






Overall Architecture

1. Taming Irregular Off-chip Mem. Access
2. Low Power Interpolation Unit
3. MLP Core w/ Coarse-Grained Skipping



A 40-nm 13.88-TOPS/W FC-DNN Engine for 16-bit Intelligent Audio Processing Featuring Weight-Sharing and Approximate Computing

Tay-Jyi Lin, Ze Li, Yun-Cheng Chen, **Chien-Tung Liu**, Tien-Fu Chen*, and Jinn-Shyan Wang

National Chung Cheng University and *National Yang Ming Chiao Tung University, Taiwan

Supported by National Science Council, Taiwan (MOST 111-2221-E-194-047-MY3 & NSTC 113-2640-E-194-001)

國立中正大學

National Chung Cheng University

WS-based FC-DNN Engine for *16b* Intelligent Audio Processing using Digital-IMC (DIMC)

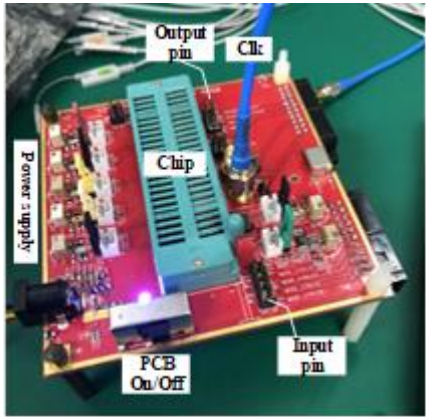
Application example of 16-bit intelligent audio processing:

Real-time Dysarthric Voice Conversion (DVC)

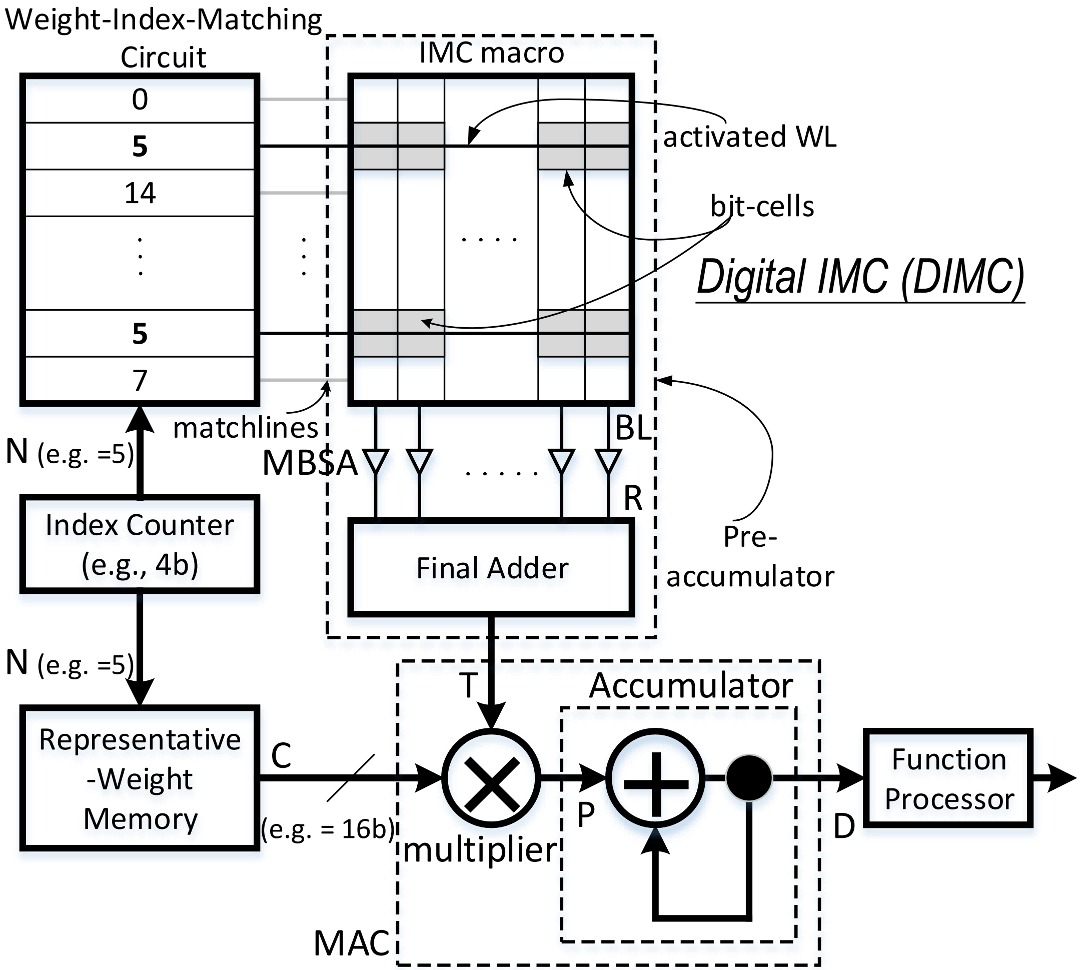
We welcome your questions and insights at our poster session!

Before conversion

After conversion



New DIMC-based FC-DNN engine for 17X improvement of energy efficiency compared to the work in [Ref]



Ref: T.-J. Lin, C.-Z. Liao, Y.-J. Hu, W.-C. Hsu, Z.-X. Wu, S.-Y. Wang, C.-M. Huang, Y.-H. Lai, C. Yeh, and J.-S. Wang, "A 40nm CMOS SoC for real-time dysarthric voice conversion of stroke patients," in Proc. IEEE Asia and South Pacific Design Automation Conference (ASP-DAC), Jan. 17-20, 2022.

A Smart Cache for a SmartNIC!

Scaling End-Host Networking to 400Gbps and Beyond

Annus Zulfiqar, Ali Imran, Venkat Kunaparaju, Ben Pfaff[1], Gianni Antichi[2], Muhammad Shahbaz

Purdue University, [1]Feldera, [2]Politecnico di Milano

Gigaflow – A Line-Rate, Pipeline-Aware Cache

Problem:

- SmartNICs offload only a subset of the SDN *traversal* cache into their limited HW resources
- Increasing link rates and diverse workloads strain this cache and misses incur high end-to-end traffic latency

Insight:

- SDN packet-processing pipelines and SmartNIC hardware are *both programmable*
- This enables us to build a *sub-traversal* cache, called **Gigaflow**, allowing sub-traversal sharing among flows

Evaluation:

- Gigaflow improves the SmartNIC **cache hit rate by up to 51%** and **misses by up to 90%**
- It captures **1000x more flow space** while using **18% lesser number of cache entries!**

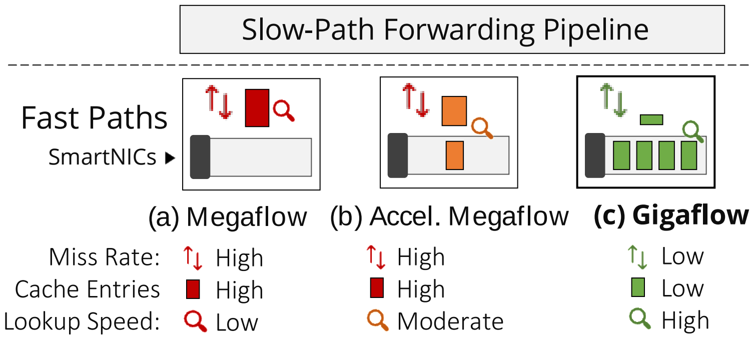


Figure 1: Comparison of OVS cache miss rate, entries and lookup speed with cache evolution

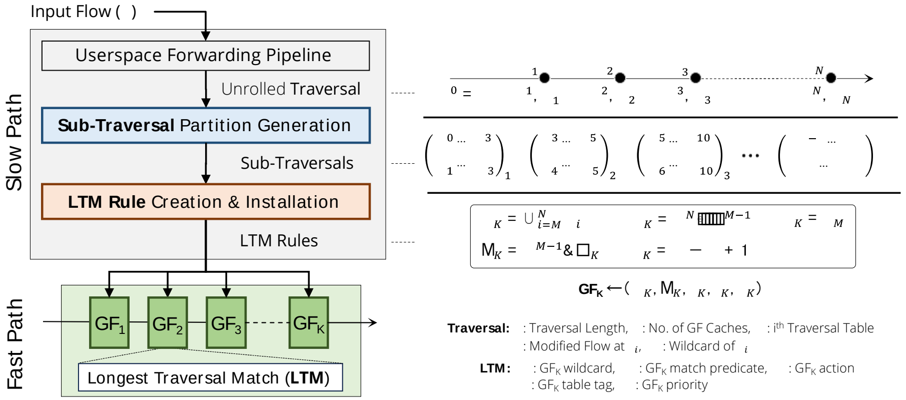
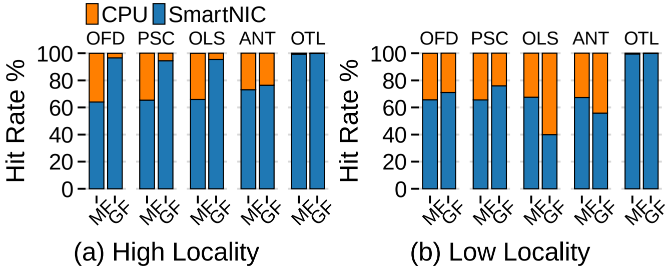


Figure 2: A high-level view of slow-path processing for cache misses with **Gigaflow**



	OFD	PSC	OLS	ANT	OTL
Megaflow	32K	32K	32K	32K	32K
Gigaflow	14.7M	4.9M	10.8M	1.3M	48K

Figure 3: **Gigaflow** performance compared to traditional Megaflow cache in SmartNICs

Questions?
Go see the posters and ask, or use Slack