P2-24: Intelligent Systems

Mission-Critical Computing
NSF CENTER FOR SPACE, HIGH-PERFORMANCE, AND RESILIENT COMPUTING (SHREC)

SHREC Annual Workshop (SAW23-24)

University of Pittsburgh · BYU Brigham Young University · Virginia Tech · UF University of Florida

January 17-18, 2024

Dr. Alan George
Mickle Chair Professor of ECE
University of Pittsburgh

Dr. David Langerman
Researcher
University of Pittsburgh

Marika Schubert
Evan Gretok
Stephen Palli
Josh Poravanthattil
Eileen Wang
Diego Wildenstein
Graduate Students
University of Pittsburgh

Number of requested memberships ≥ 6

Overview

Goal: Investigate **emerging machine learning** paradigms and devices for space and other embedded applications

Motivation: AI promises to expand capabilities for edge-system sensing and processing without compromising performance

Challenges: Space apps are subject to **SWaP-C** and **reliability constraints,** which pose novel complexity for emerging systems

Tasks for 2024

T1

Few-Shot Learning for Space

- Assess performance of few-shot learning onboard space-grade devices
- Enable more accurate classification of unknown classes without retraining

T2

Neuromorphic Vision and Computing

- Characterize resiliency of event-driven SNNs
- Explore tradeoff between biological plausibility and computational efficiency

T3

Novel Processor Architectures

- Characterize performance and reliability of Gemini APU in-memory processors
- Evaluate and optimize DL models for PIM architectures

T4

Deep-Learning Kernel Benchmarks

- Explore implementation statistics of DL kernels in vision models
- Create new DL benchmarks to better reflect expected performance

APU: Associative Processing Unit
DL: Deep Learning
PIM: Processing-in-Memory
SNN: Spiking Neural Network

T1: Few-Shot Learning ...in Space!!

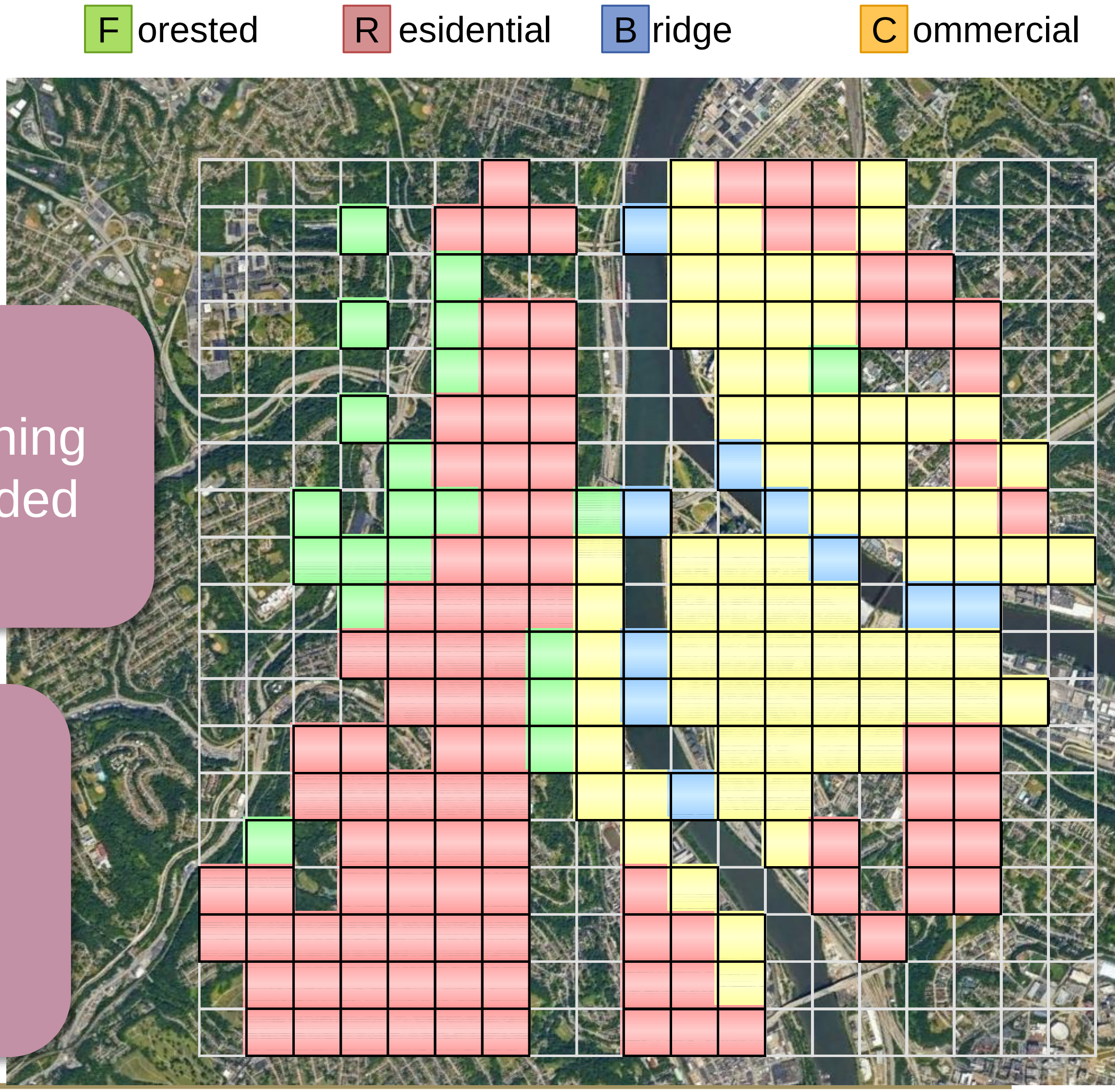
Evan Gretok, Eileen Wang
ewg13@pitt.edu elw96@pitt.edu

T1: Few-Shot Learning

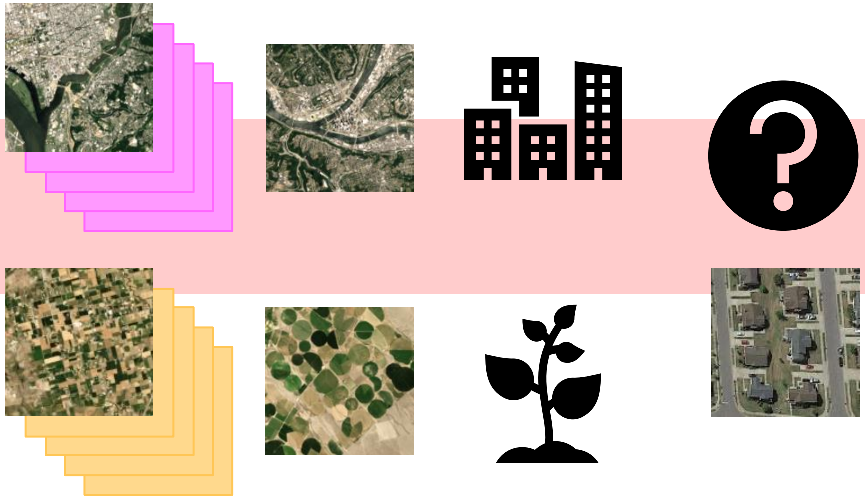
What is Few-Shot Learning?
- Training with a **small number** of samples per class

How is Few-Shot Learning Different?
- **No large dataset** to train as with supervised learning
- X-way, Y-shot for X classes and **Y samples** provided
- Small **query set** of images used for testing

What Can Few-Shot Learning Enable in Space?
- Can **reduce labelling** need, especially useful as vast majority of Earth-observation data is unlabeled
- Enable best guess of never-before-seen image classes on orbit **without retraining**



T1: Few-Shot Learning

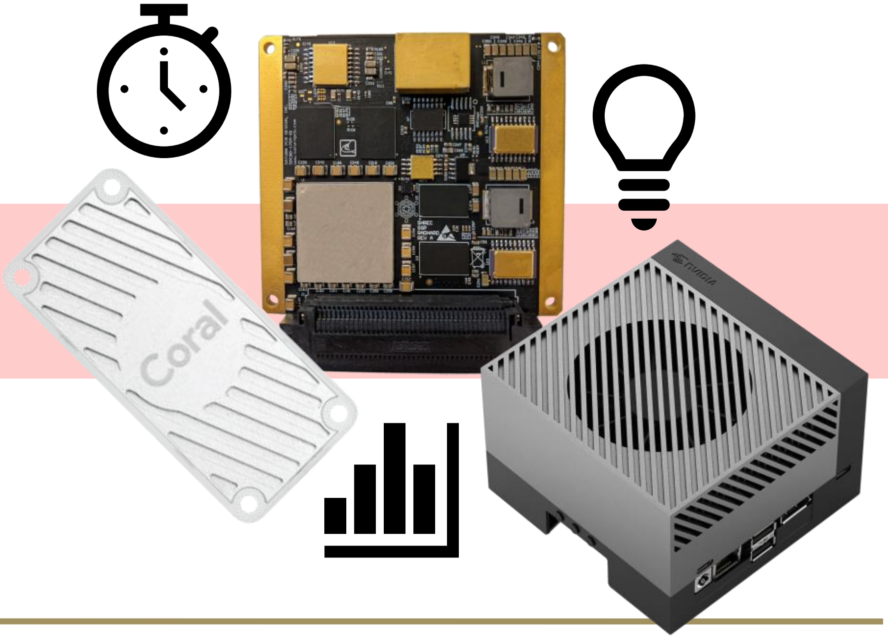


With What Will We Experiment?

- Leveraging existing **Earth-observation** datasets
- Evaluating different few-shot learning **algorithms**
- Varying number of **classes and samples** provided
- Exploring responses to **never-before-seen** classes

What Will We Measure?

- **Accuracy** of few-shot learning approach taken
- Runtime, memory use, and energy consumption of few-shot **inference** onboard space-grade hardware
- Algorithm-specific traits, such as **inter-class distances** for prototypical networks



T2: Neuromorphic Vision and Computing

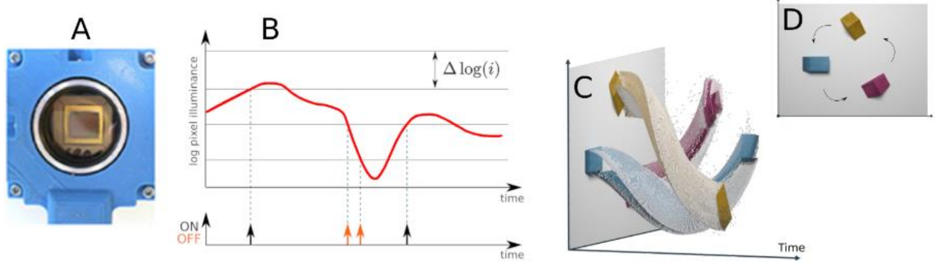
Joshua Poravanthattil
jbp51@pitt.edu

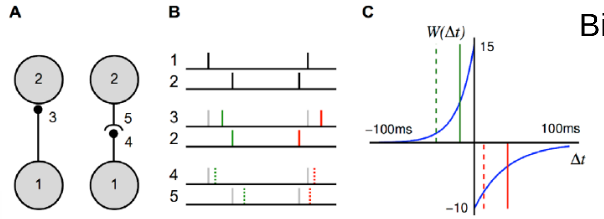
P2

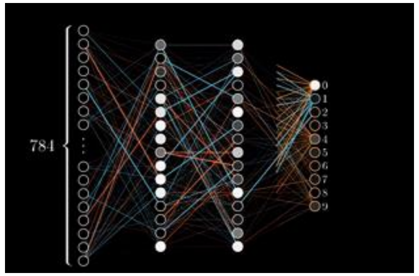
T2: Neuromorphic Vision and Computing

Why Event-Based Sensors and Algorithms?

- SNNs are **powerful** and **efficient**, especially when paired with **event-based** sensor data
- Prior simulation suggests that backprop SNNs exhibit **intrinsic reliability** to radiation-induced noise
- Many **learning methods** and **neuron models** to explore!







Biologically Plausible

BindsNET

snnTorch

Computationally Efficient

Resiliency Exploration

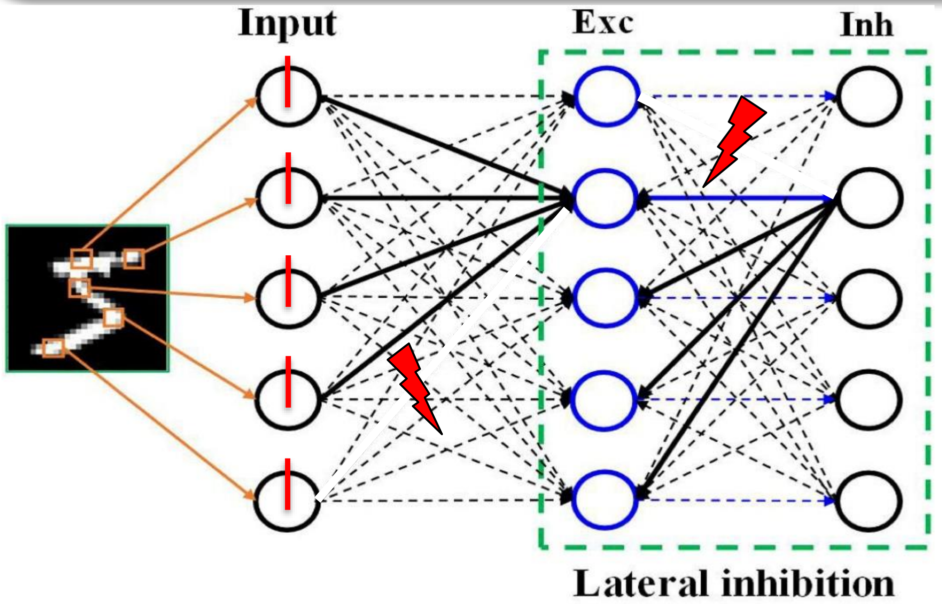
- SNN architectures can vary from **biologically plausible** to **computationally efficient**
- How can this **tradeoff** be exploited to make the most **resilient networks**?

T2: Neuromorphic Vision and Computing

How Will Radiation Tolerance Be Assessed?

- Vary the **neuron model** and **learning method** from biologically plausible to computationally efficient
- Inject data and processor faults on pretrained networks and analyze **performance hits**
- Investigate state-of-the-art **filtering** methodology

Los Alamos NATIONAL LABORATORY



What Will We Measure?

- **Accuracy** and **loss** metrics across neuron models and learning methods
- **Runtime** and **power utilization** statistics for hardware implementations

<u>**Exc**</u>: Excitatory Neurons
<u>**Inh:**</u> Inhibitory Neurons

T3: Novel Processing Architectures

Diego Wildenstein, Stephen Palli

Diego.Wildenstein@pitt.edu

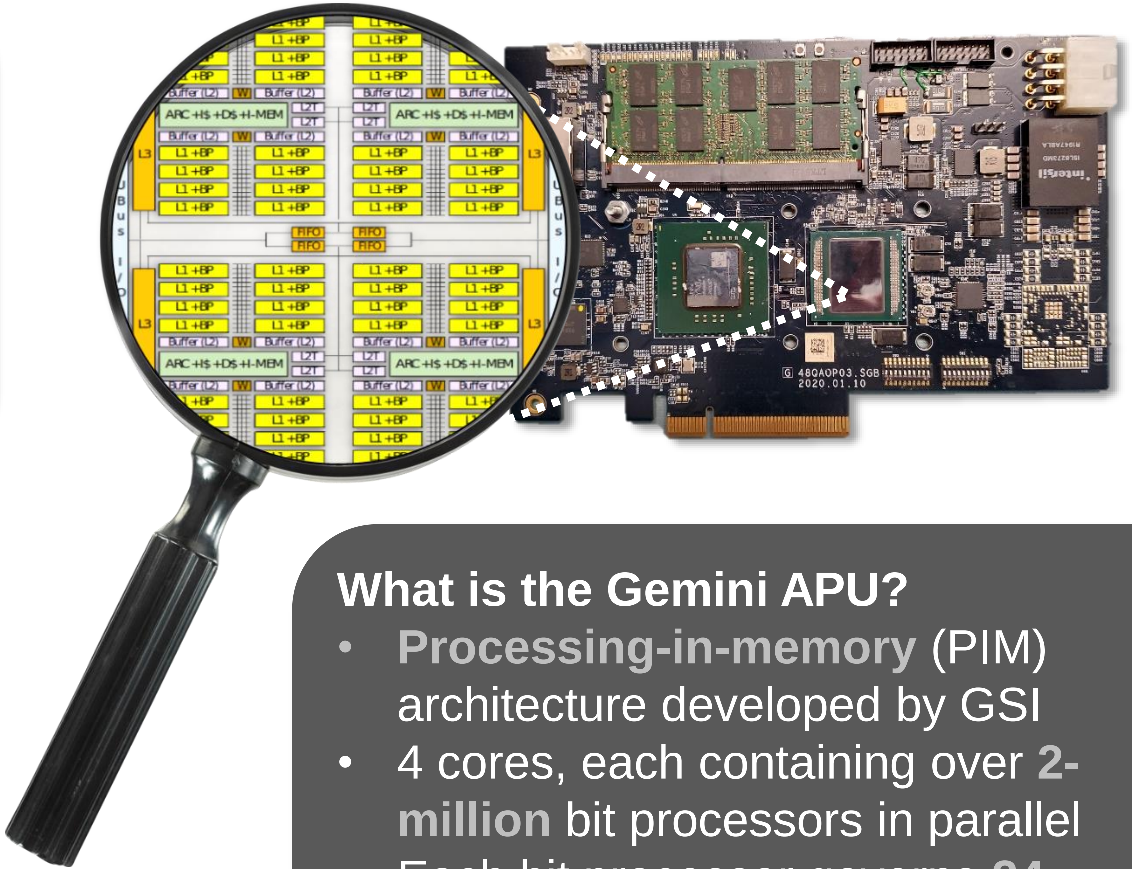
Stephen.Palli@pitt.edu

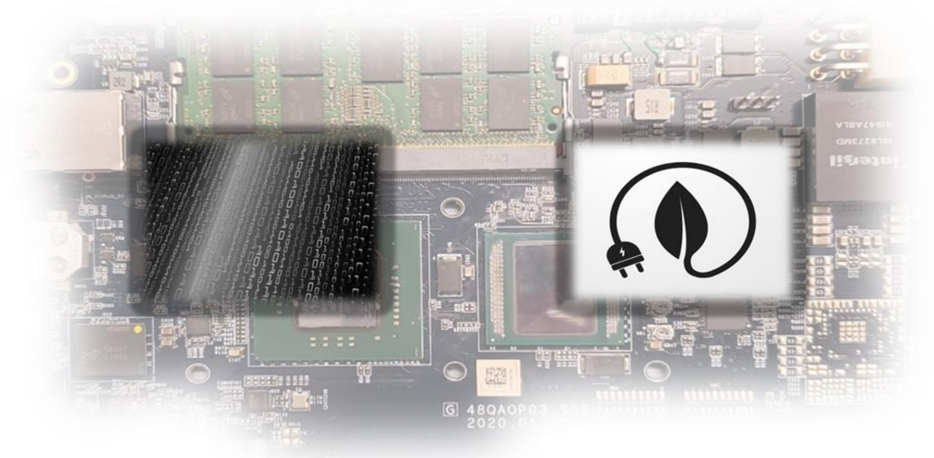
P2

T3: Novel Processing Architectures

What is PIM?

- Shared memory between device and CPU **reduces data transfers**
- High **data throughput** performance on **large scale** problems
- Energy consumption is **fractionally less** than modern CPUs and GPUs



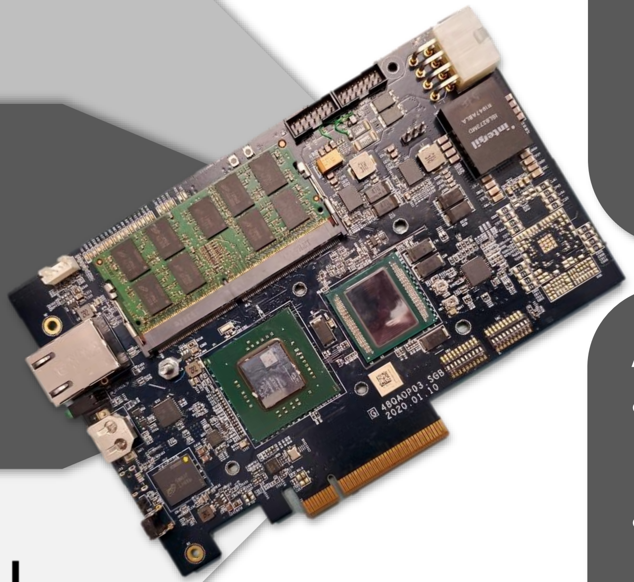


What is the Gemini APU?

- **Processing-in-memory** (PIM) architecture developed by GSI
- 4 cores, each containing over **2-million** bit processors in parallel
- Each bit processor governs **24 individual memory cells**

PIM: Processing-in-Memory
APU: Associative Processing Unit

T3: Novel Processing Architectures



APU Assessment for Future Missions

- Benchmark low-level linear algebra **compute kernels** at **various precision** data types
- Compare APU **performance** and **power efficiency** with modern and upcoming **space flight hardware**
- Assess **radiation tolerance** and **susceptibility** to single event effects on APU devices

APU Accelerators for Deep Learning

- Determine what deep-learning **models** and **apps** are best suited for **APU architecture**
- Leverage PIM architecture for **optimization** of common **deep-learning operations**
- Compare performance of **deep-learning apps** on APU with CPU and GPU implementations

<u>PIM</u>: Processing-in-Memory
<u>APU</u>: Associative Processing Unit

T4: Deep-Learning Kernel Benchmarks

Marika Schubert
marika.schubert@pitt.edu

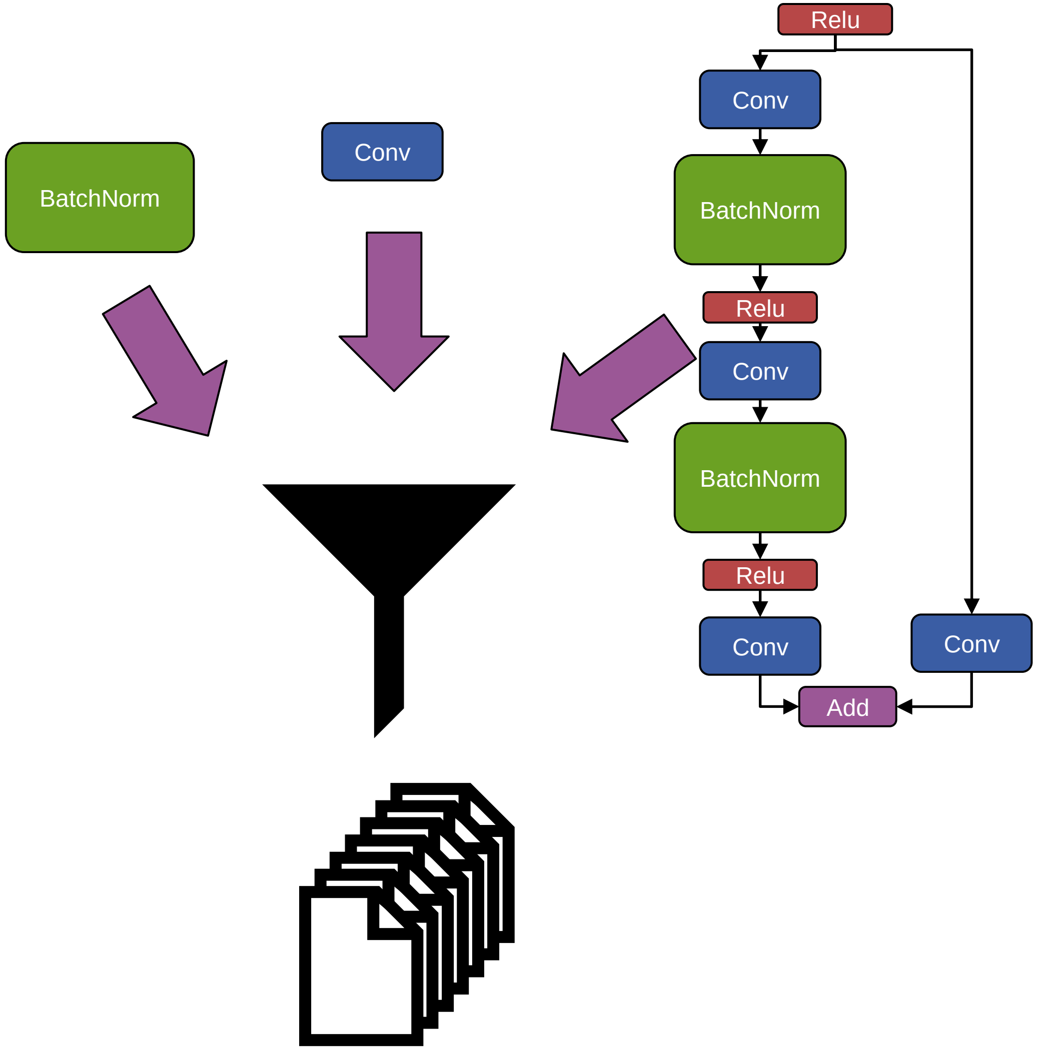
T4: Deep-Learning Kernel Benchmarks

Why Do We Benchmark DL Inference?

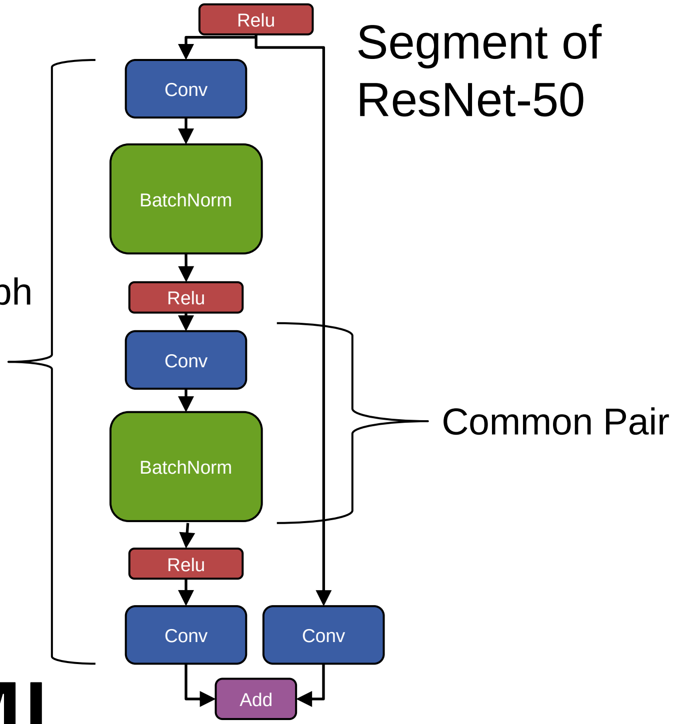
- Test system **latency, memory use, software compatibility**
- Determine how to **improve models** for given hardware platform

Why do We Need a Granular Benchmark?

- Most DL benchmarks (MLCommons, MLMark) test **full model** (great if you care about Resnet-50)
- Baidu's DeepBench tests kernels, but **from optimized libraries**
- Need **kernel benchmark for PyTorch performance**



T4: Deep-Learning Kernel Benchmarks



MLCommons

Baidu Research

Identification of Kernel Sequences and Subgraphs

- Identify **common kernel pairs and subgraphs** to describe vision models from a more granular perspective

Development and Comparison of Benchmark

- Use **summary statistics** to create new kernel-based/sub-graph-based benchmark
- Compare kernel coverage of benchmark to similar **DeepBench** kernel-based benchmark, coverage of **MLCommons**

Evaluation of Benchmark on Devices

- Run benchmark on **CPU/GPU devices**
- Run benchmark on **novel accelerators** that support PyTorch/ONNX (SambaNova, Tenstorrent, etc)

Milestones, Deliverables, Budget

Milestones

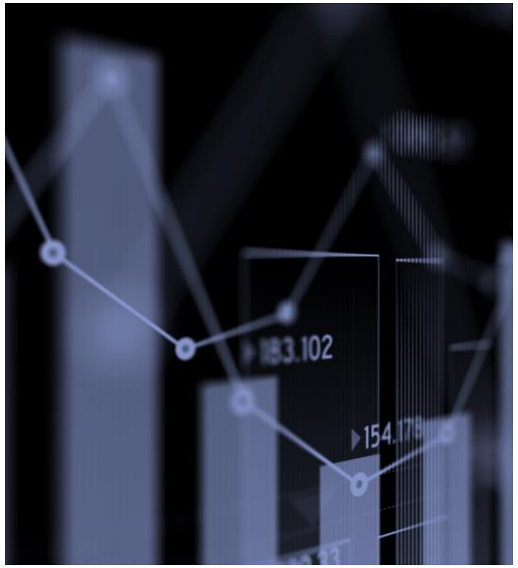
- SMW24 (06/24 or 07/24): Showcase **preliminary results** on all project tasks
- SAW24-25 (01/25): **Completion** of all project tasks

Deliverables

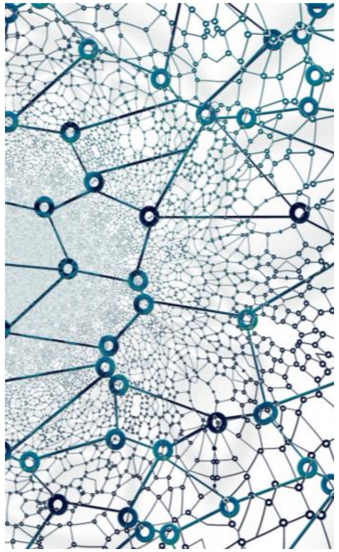
- Monthly **progress reports** from all projects
- Midyear and end-of-year **full reports** from all projects
- 4-5 conference or journal **publications**

Budget

- Minimum recommended: **Six (6) memberships** (300 Votes)



Conclusions and Member Benefits



Conclusions

- **Few-shot learning** enables accurate onboard classification with less labelled data and can even generalize its training to never-before-seen samples
- **Neuromorphic architectures** can be designed to provide resilient and efficient inferencing capabilities
- **In-memory processing architectures** can increase performance of deep-learning apps and expand onboard computation capability
- Deep-learning kernel benchmarks can be used to **explore optimizations** and improve **model selection** for embedded devices

Member Benefits

- Direct influence over **processors and frameworks** studied
- Direct influence over **apps and datasets** studied
- Direct benefit from new **methods, data, code, models, and insights from metrics, benchmarks, and emulations**