F1-26: Device and Architecture Studies for Large AI Systems

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

SHREC Annual Workshop (SAW25-26)

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

January 13-14, 2026

Dr. Herman Lam
Assoc. Professor of ECE

Dr. Janise McNair
Professor of ECE

P. Gupta, J. Lewis

J. Madden, A. Rice-Bladykas

Research Students
University of Florida

Number of requested memberships 3 to 4

Project Goal & Approach

Goal

Optimize and advance key technologies that will accelerate performance of *large AI* and *mission-critical* systems

- Perform *acceleration and scaling studies* on devices, applications, and benchmarks for development of *heterogeneous compute cache architectures/systems*
- *Machine learning-based optimizations for* multi-radio access technology 5G/6G mission-critical deployments

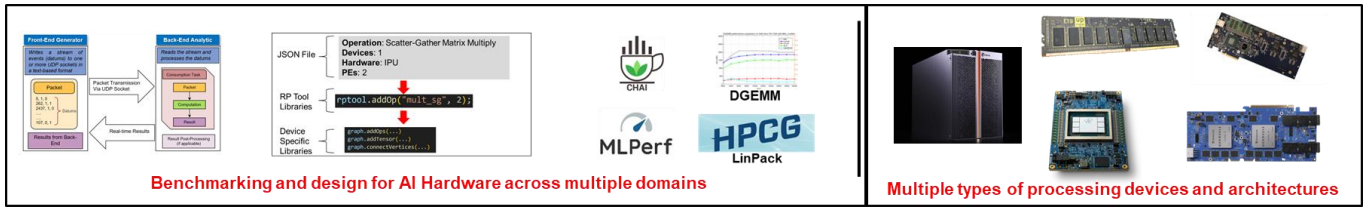
R&D Approach and F2 Tasks

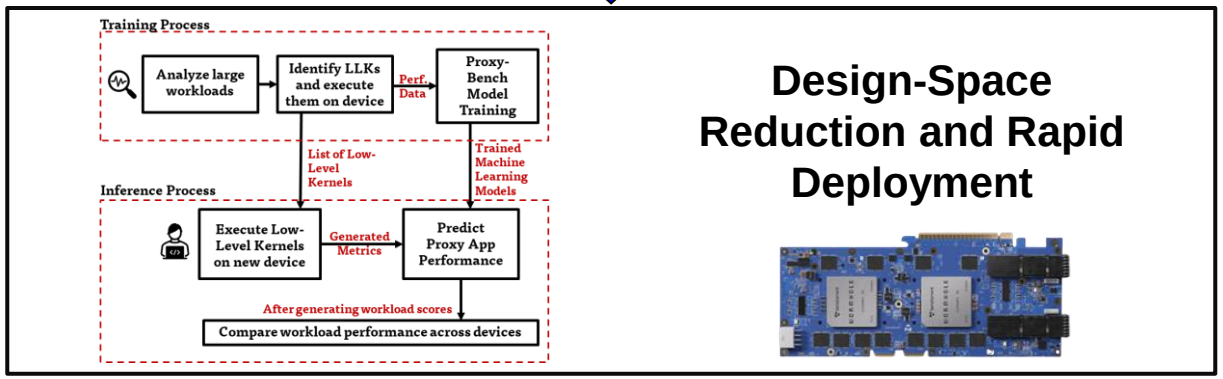
- Development of heterogeneous compute cache (HCC) architectures and systems
  - Acceleration & Scaling Studies for Memory Compute (MemCp) Devices & Accelerators
  - Methods and tools for Design-Space Reduction and Rapid Deployment
  - Prototype and Deployment of HCC architectures and systems
- Develop adaptive and responsive SDN[1]-managed 5G interoperable networks.

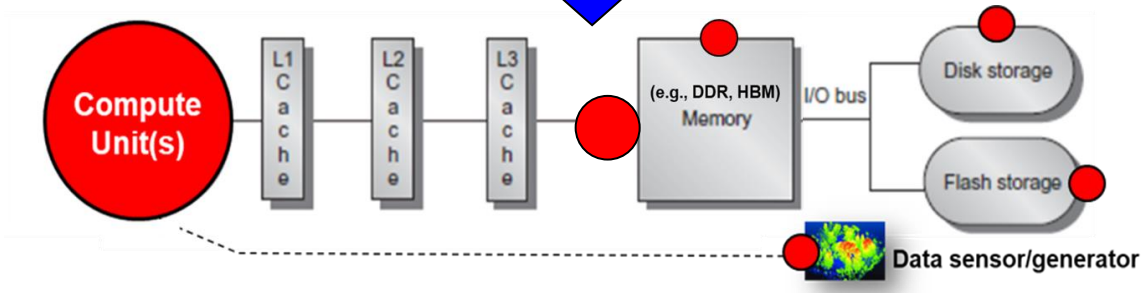
[1] SDN: Software-defined networks

[2] OnDemand (https://openondemand.org/)
[3] Prometheus Monitoring System (https://prometheus.io/)Grafana
[4] Data Visualization System (https://grafana.com/)

Task Overview for Development of HCC Systems







T1: Acceleration & Scaling Studies for Memory Compute (MemCp) Devices & Accelerators

T2(a): Machine Learning-Aided Rapid Deployment

T2(b): Performance Modelling & Prediction for Design-Space Reduction

T3: Development and Deployment of HCC Architectures & Systems

T1: Memory Compute Devices & Accelerators

- **Tenstorrent Wormhole**
  - Investigate additional HPC/ML operations and formats
  - Evaluate Wormhole's potential as a control plane device
  - Scale large applications across multiple Tenstorrent devices

- **GSI Associative Processing Unit**
  - Explore use of GSI APUs to accelerate RAG (Retrieval Augment Generation) and other ML ops of interests.
  - Explore STTR* and other programs for collaboration

s

- **Cerebras WS3**
  - Complete/evaluate transpilation studies on Cerebras device
  - Publish paper regarding transpiled application performance

- **Other Emerging Accelerators and Devices**
  - Benchmark the performance of emerging devices and accelerate of interest various kernels and workloads
  - Understand scaling strategies and apply them on workloads

Approach

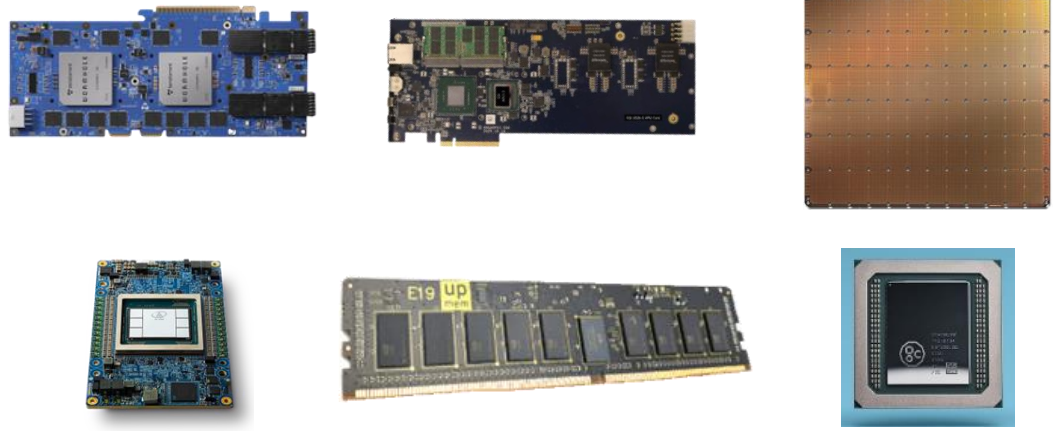
Acceleration Studies

- Identify and Study Kernels of interest
- Determine productivity issues in the development of benchmark workloads

Scaling Studies

- Re-factor algorithms to scale efficiently in a multi-core environment
- Identify architectural barriers for scaling

Accelerators under research

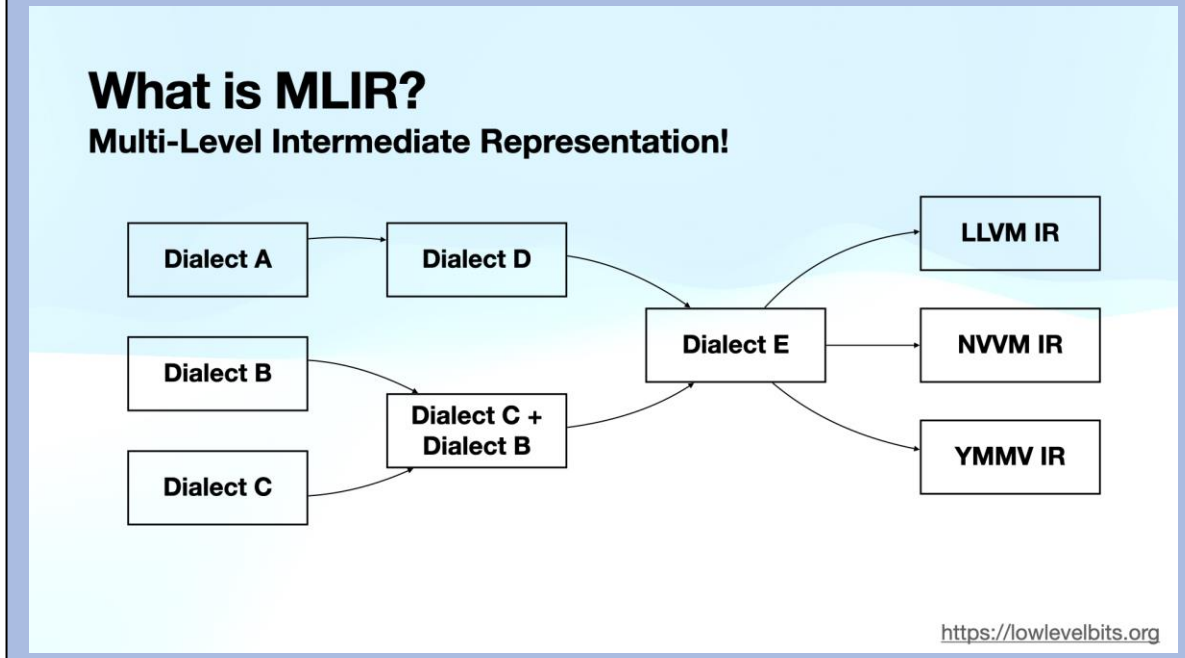


* STTR: Small Business Technology Transfer

T2(a) Machine Learning-Aided Rapid Deployment

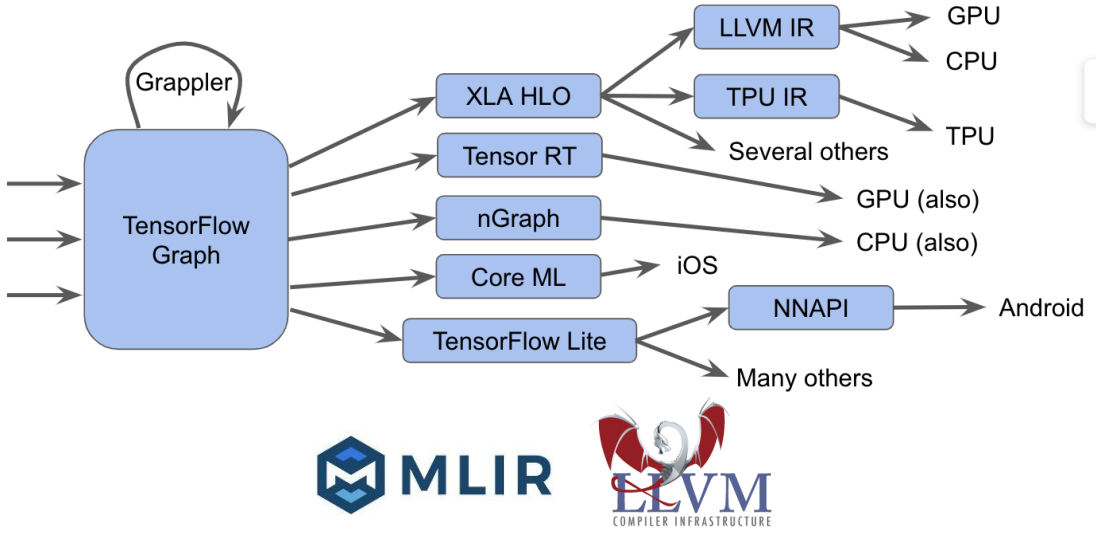
Generalized MLIR Framework

Development of an MLIR Framework that with primitives better suited for dataflow accelerators. Pre-existing dialects will be used as a base.



Write Once, Deploy Multiple

Re-design benchmarks using high-level MLIR dialects which can define programs in a hardware agnostic way. Different backend targets can then generate device-specific code.

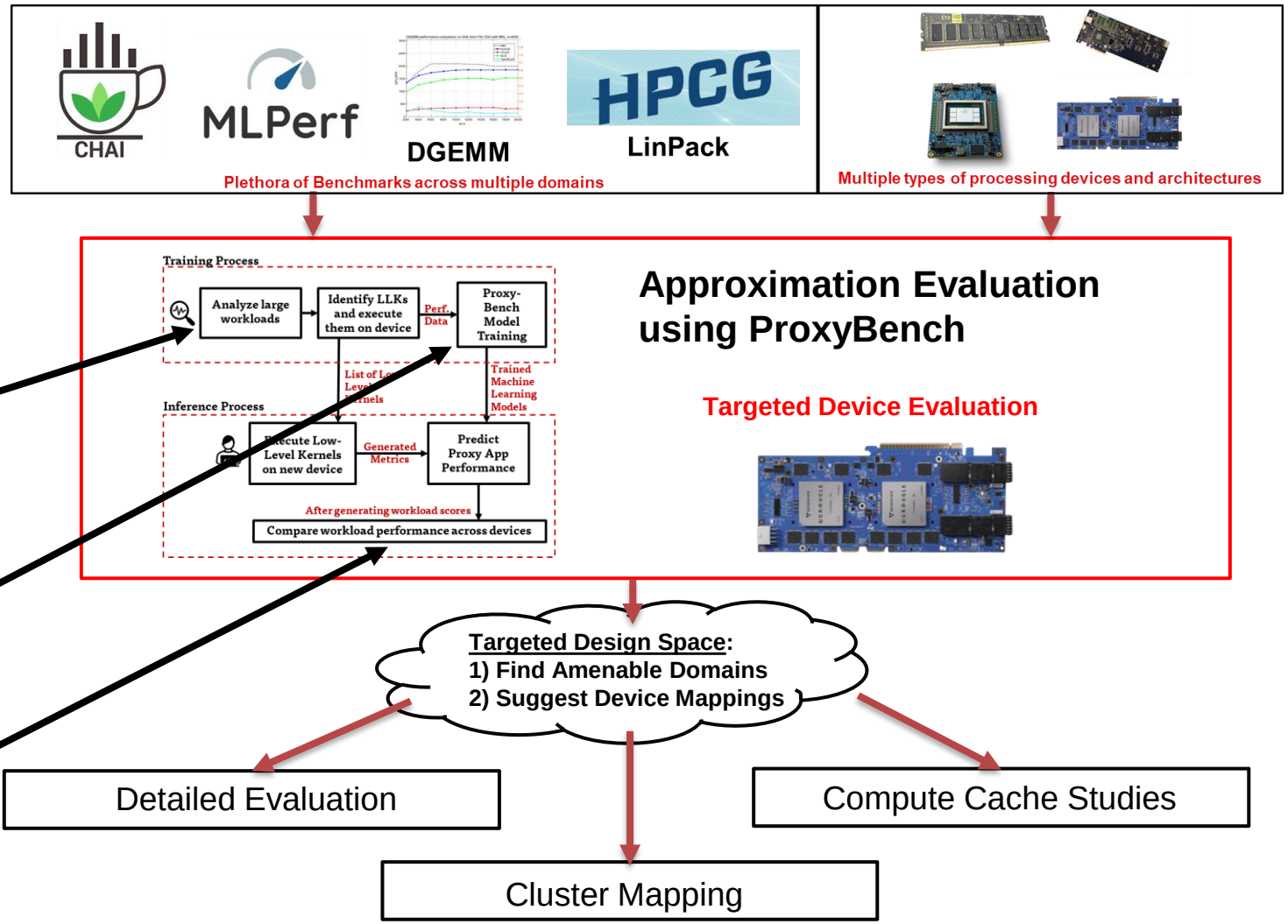


T2(b) ProxyBench: Reduction of Design Space

Proxybench reduces the large design space introduced by multiple benchmarks across many domains

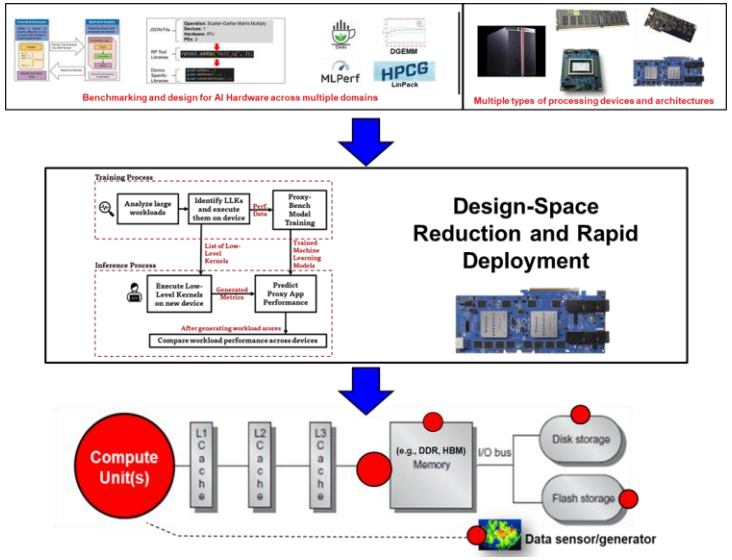
Research Thrusts

- **Automated Analysis of Large Workloads**
  Automatic DAG based analysis of large applications and workloads for low-level kernel identification and extraction
- **Training Proxybench Models**
  Investigate better models, update model parameters and focus on model portability for similar devices
- **Model Interpretability and Inference**
  Understand the decision-making mechanism of the models and apply results towards building workload management and scheduling tools



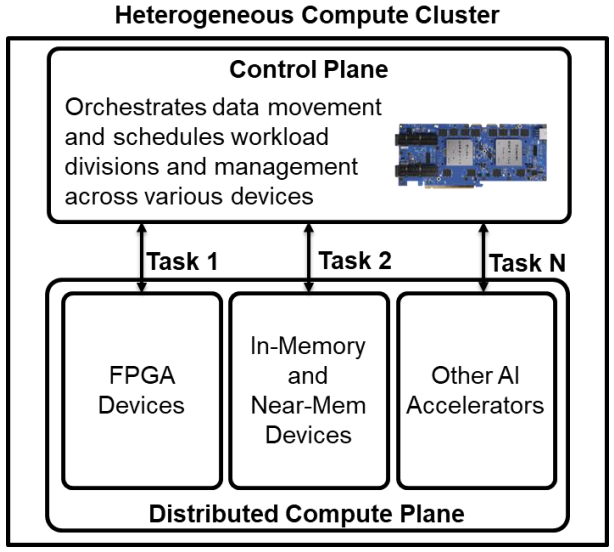
T3: Heterogeneous Compute Cache Architecture

Heterogeneous Compute Cache (HCC) Architecture investigates the design and implementation of a custom accelerator-centric heterogeneous cluster



Research Thrusts

- **Prototype & Deployment of Apps on Accelerators of Interest**
  Integrate transpilation, benchmarking, and MLIR frameworks for a comprehensive evaluation of application migrations from HPC hardware to dataflow architectures



- **HCC Architecture Studies Integrating Implementation w/ Emulators & Simulators**
  Integrate implementation, emulation, and simulation frameworks for an **integrated study** of existing and emerging accelerators to generate a hybrid verification stack for a complete HCC architecture

- **Investigate Novel Control Planes**
  Investigate non-traditional control planes which reduce CPU dependency for data management and orchestration for large workloads in a cluster

- **Emerging Interconnect Studies**
  Evaluate novel interconnect technologies such as UALink and CXL which can enable faster and more efficient data transfers for large data workloads

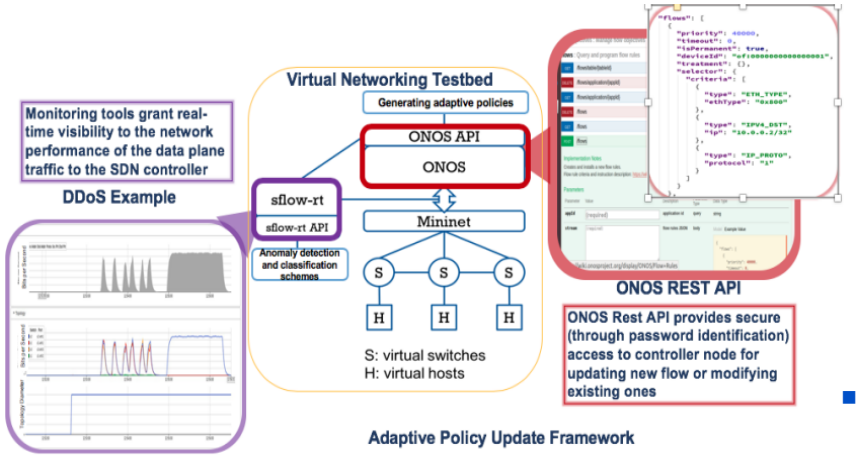
T4: *Reinforcement learning* techniques for *next generation 5G/6G networks.*

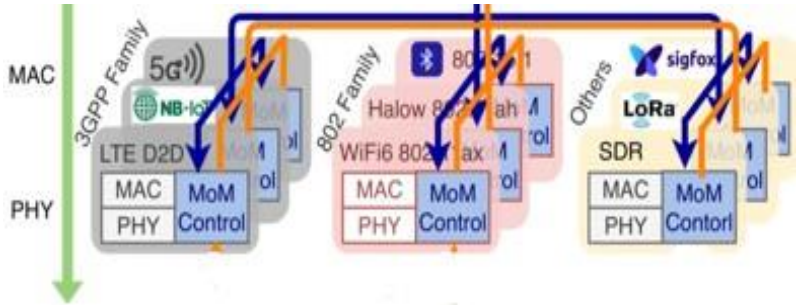
Research Thrust 1

Machine-Learning Approaches

Explores the application of a machine-learning based algorithms in 5G/6G multi-radio access technology mobility.

- **Shortest Distance Algorithm:** Address the performance disparity according to the size of the satellite constellations.
- **Training**: Train machine learning model on failure conditions, including device, link and signal failures.
- **Analysis:** Investigate using reinforcement learning or some other machine learning approach for mobility management for various multi-radio access technology environments.





Research Thrust 2

Next Generation Network Performance Analysis

Examine new tools for more accurate performance evaluation

- **SDN-based Approach:** Using SDN controllers to manage multi-radio access technology mobility.
- **Tools:** Onboard simulation tools for multi-radio access performance, such as Omnet++, ns-3, and Mininet.
- **Metrics**: Collect connection times, duration, delay, transition time, from surrounding multi-tier towers and access points.

Milestones, Deliverables & Budget

Milestones

- SMW26: Showcase midway progress on framework, platform, and interconnect exploration
- SAW26-27: Present completed project results

Deliverables

- Application source code and technology-transfer support
- Progress reports documenting research methods, progress, results, and analysis
- Several conference and/or journal publications

Membership Budget

- Requesting 3 to 4 memberships

Conclusions & Member Benefits

Conclusions

- The goal is to optimize and advance key technologies that will accelerate performance of large AI and mission-critical systems
  - Perform acceleration & scaling studies for MemCp devices & accelerators
  - Develop profiling, verification, & rapid prototyping toolchain for MemCp studies
  - Develop heterogeneous compute cache architecture & systems
  - Creating network architectures for applications, *such as security and quality of service,* that can generate and leverage *real-time situational awareness* through new network profile *data sets*, network *models*, and *machine learning and AI-based* protocols.

Member Benefits

- Direct influence over selected architecture, app, and inter-connect studies
- Technology transfer of accelerated archs/apps/techniques of interest to members
- Key insights and lessons learned from design space explorations & tradeoff analyses