National Science Foundation's Industry/University Cooperative Research (I/UCRC) Program

F1-25: Device & Architecture Studies for Compute Cache Systems

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

SHREC Annual Workshop (SAW24-25)

University of Pittsburgh
BYU BRIGHAM YOUNG UNIVERSITY
VIRGINIA TECH.
UF UNIVERSITY of FLORIDA

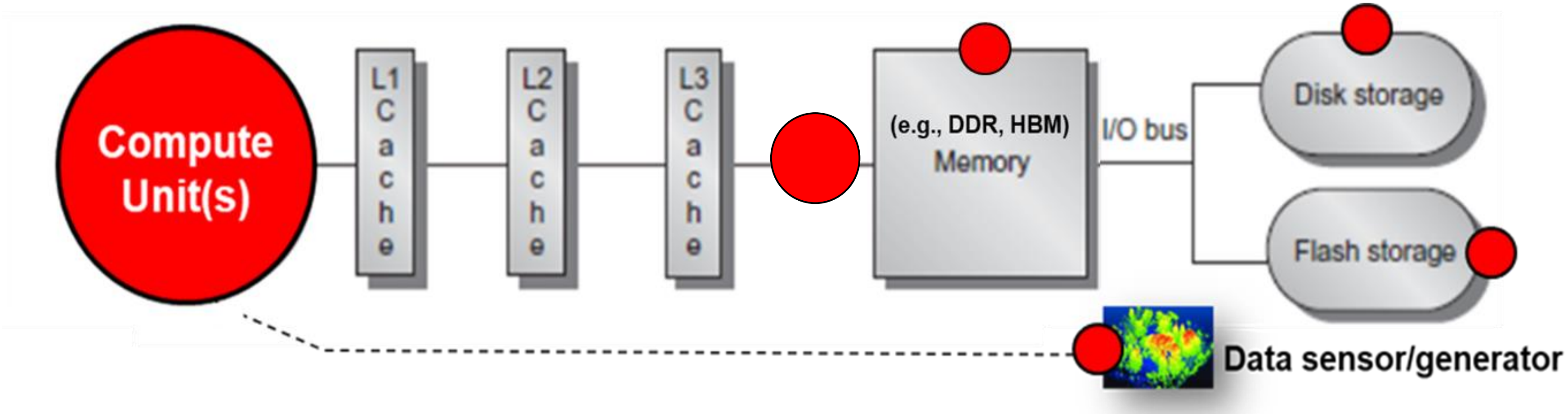
January 14-15, 2025

Dr. Herman Lam
Assoc. Professor of ECE

Y. Gao, P. Gupta
D. Klein, J. Madden
Research Students
University of Florida

Number of requested memberships 3 to 4

Device & Architecture Studies for Compute Cache Systems



Motivation ***Data bottleneck**:* Bring *compute close to data* for *data-intensive,* data-analytics applications

Goal Perform ***acceleration*** and ***scaling*** studies on devices, applications, and platforms for compute cache systems

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

T2: Profiling, Verification, and Rapid Prototyping Toolchain for MemCp Studies

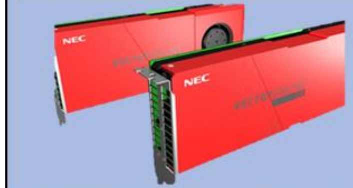
T3: Heterogeneous Compute Cache Architecture & Systems

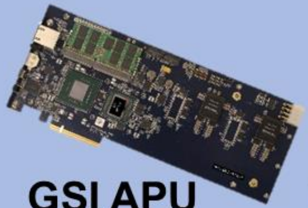
F1

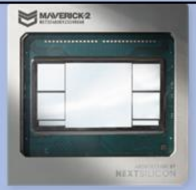
T1: Acceleration & Scaling Studies for MemCP* Devices & Accelerators

Expanding Device Support

Complete adaptation of FireHose and Circus Tent for other devices to understand performance on different workloads.


NEC Vector Engine

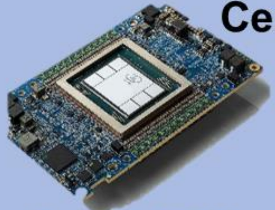

GSI APU


NextSilicon Maverick

SK Hynix AiMx

Cerebras WSE-3

UPMEM DPU

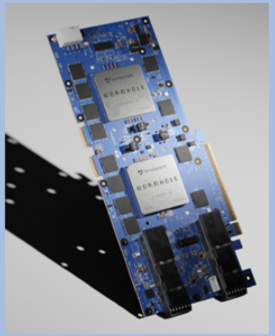

Gaudi TPU

SambaNova RDU

Upcoming Collaborations

Begin collaborations with other interested vendors and labs to investigate promising memory compute devices.

Tenstorrent
RISC-V-based accelerators with a fabric of compute tiles. Open-source friendly with large developer community.



Argonne Labs
National lab with ongoing projects regarding accelerating HPC applications on memory compute hardware.

Argonne NATIONAL LABORATORY

* MemCp: Memory Compute Devices

F1

T1: Acceleration & Scaling Studies for MemCP* Devices & Accelerators

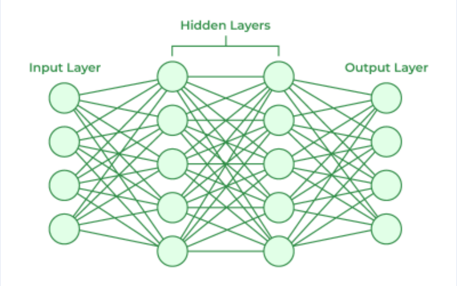
ML Benchmarking

Introduce ML Benchmarking tools to determine the upstream performance benefits of optimizing HPC applications

If we build a better matrix multiply…

$$\begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ a_7 & a_8 & a_9 \end{bmatrix} \begin{bmatrix} b_1 & b_2 & b_3 \\ b_4 & b_5 & b_6 \\ b_7 & b_8 & b_9 \end{bmatrix} = \begin{bmatrix} c_1 & c_2 & c_3 \\ c_4 & c_5 & c_6 \\ c_7 & c_8 & c_9 \end{bmatrix}$$

…can we get more efficient neural networks?



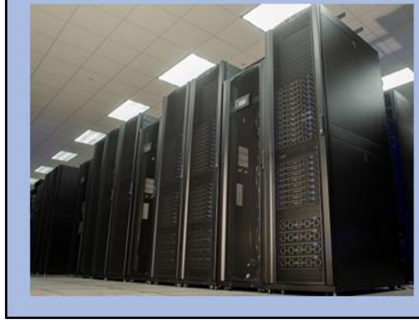
Expanded Scaling Studies

Utilize access to current research clusters to expand current scaling studies. Research obstructions in scaling due to topology.

ALCF AI Testbed

Research cluster with the latest from NextSilicon, Cerebras, and others.

TAMU ACES



NSF ACCESS cluster with several devices from Intel and Graphcore

* MemCp: Memory Compute Devices

T2: Profiling, Verification, and Rapid Prototyping Toolchain for MemCp* Studies

Rapid Prototyping Toolchain

Continued development of our rapid prototyping tool. Necessary to profile, verify, and implement workloads in a timely manner on heterogeneous clusters.

Design / Conversion

JSON File Parameters | Universal Developer API

Compilation / Implementation

Link to Device Libraries | Application Optimization

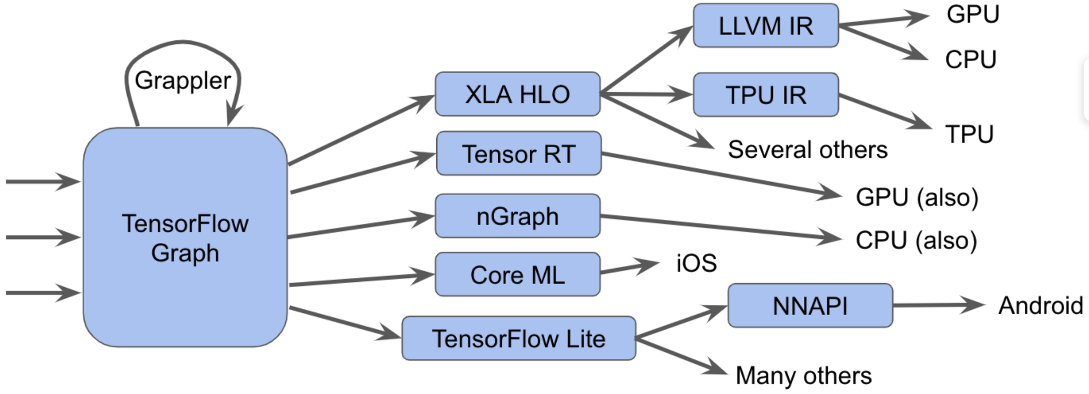
Verification / Experimental Results

Correctness in Output | Memory Usage

Routing / Communication | Timing Requirements

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.

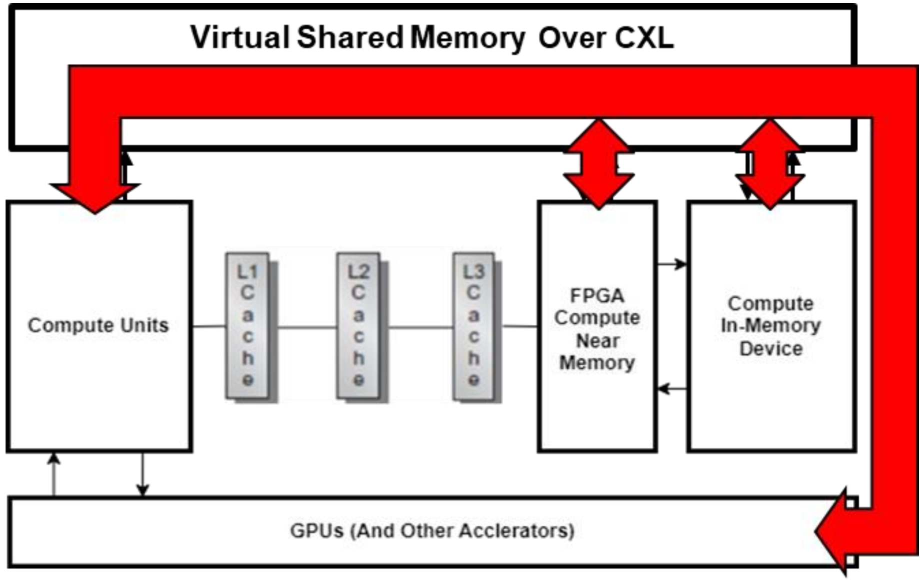


* MemCp: Memory Compute Devices

T3: Heterogenous Compute Cache (HCC) Development

T3a: Architecture Modelling Studies

Continued verification of the heterogenous compute cache architecture using various performance modelling and emulation methods.



Modelling & Verification Stack

QEMU
Qemu Emulator

gem5
Gem5 Simulator

T3b: Expand Interconnect Studies

Expand interconnect support for emerging technologies like UCIe for chiplet-based designs and UALink and UEC for scaling-up and scale-out the architecture.

CXL Compute Express Link™
UCIe Universal Chiplet Interconnect Express

Intra-node Inter-device Communication Protocols

Ultra Ethernet Consortium
ULTRA ACCELERATOR LINK™

Inter-node Fabric Scale-up Protocols

T3: Heterogenous Compute Cache (HCC) Development

T3c: Emulating HCC Architecture on QEMU

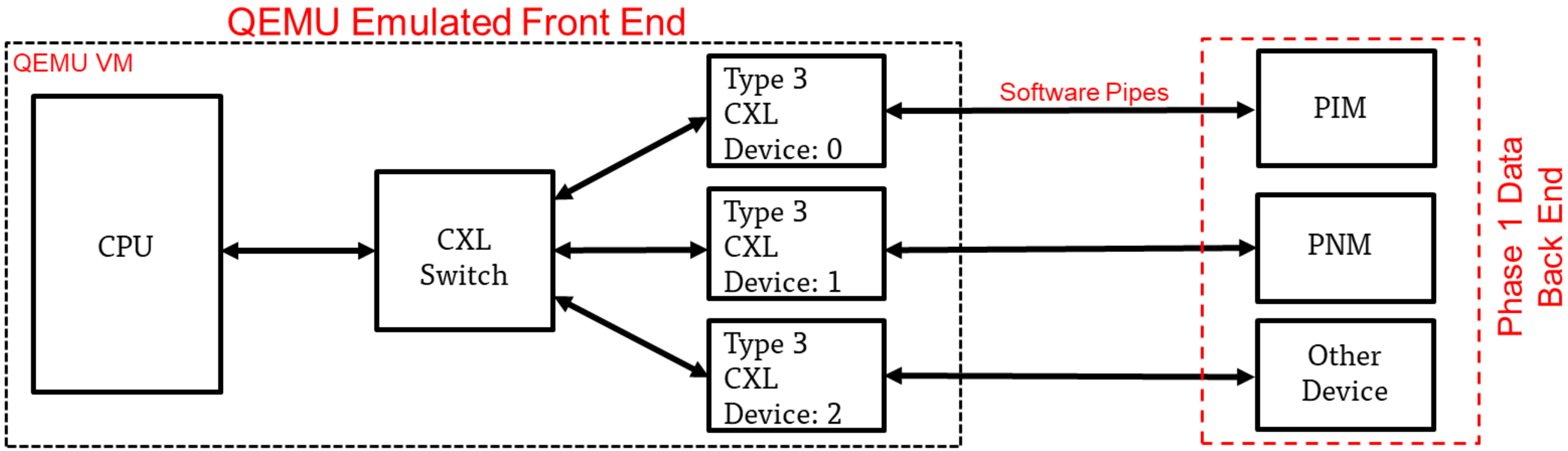


Fig: Proposed System using CXL over QEMU

QEMU provides robust support for CXL* device emulation and allows custom backends to execute sub-routines.

*(Type 1 and Type 3)

HCC Architecture Emulator Studies

1. Validate inter-device data sharing model using emulated system
2. Implement large workloads on the emulated system to evaluate their performance
3. Leverage inter-device data transfers to pipeline data across each device as shown on the right

Time

Data

	T0	T1	T2	T3	T4	T5
Chunk 1	PIM	PNM	Accel	CPU		
Chunk 2		PIM	PNM	Accel	CPU	
Chunk 3			PIM	PNM	Accel	CPU

Milestones, Deliverables & Budget

Milestones

- SMW25: Showcase midway progress on framework, platform, and interconnect exploration
- SAW25-26: 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 perform *acceleration* and *scaling* studies on devices, applications, and platforms for compute cache systems
  - Perform acceleration & scaling studies for MemCp devices & accelerators
  - Develop profiling, verification, & rapid prototyping toolchain for MemCp studies
  - Develop heterogeneous compute cache architecture & systems

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