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

F1-24: Device & Architecture Studies for Compute Cache 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. Herman Lam

Assoc. Professor of ECE

Y. Gao, P. Gupta

R. Hernandez-Lopez

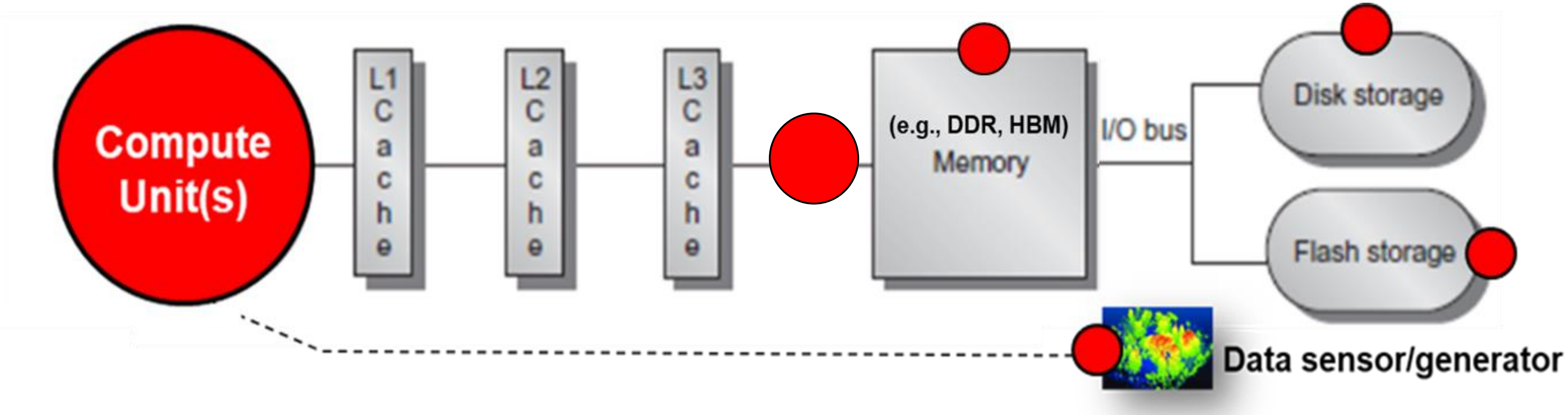
D. Klein, J. Madden

Research Students

University of Florida

Number of requested memberships ≥ 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: Device & Algorithm Studies for Compute Cache**:** Acceleration and scaling studies on devices, applications of interest.

T2: oneAPI Cross-Platform & ModSim Studies for Compute Cache

T3: Compute Cache for Edge Computing for FNN (Fire Neural Net)

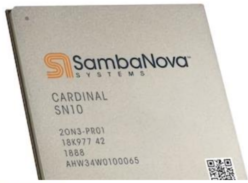
F1

Device & Algorithm Studies for Compute Cache

Acceleration Studies

- Continued development & evaluation of devices on FireHose* benchmark:
  - **FY2023 accelerators**: Graphcore IPU, UPMEM DPU
  - **New devices (FY2024)** Habana Gaudi TPU, SambaNova RDU
- **New benchmark:** Circus Tent**

Suggestions from SHREC Members?



Scaling Studies

- **Expand current benchmark development for scaling studies**
- Resources available via ALCF clusters at *Argonne National Lab*
  - Graphcore IPU-POD$_{64}$
  - SambaNova Datascale SN30
  - GroqRack

Graphcore IPU-POD$_{64}$

- 64 Gen 2 IPUs
- 22 PFLOPS @ FP16.16
- 2D Torus Configuration

Datascale SN30

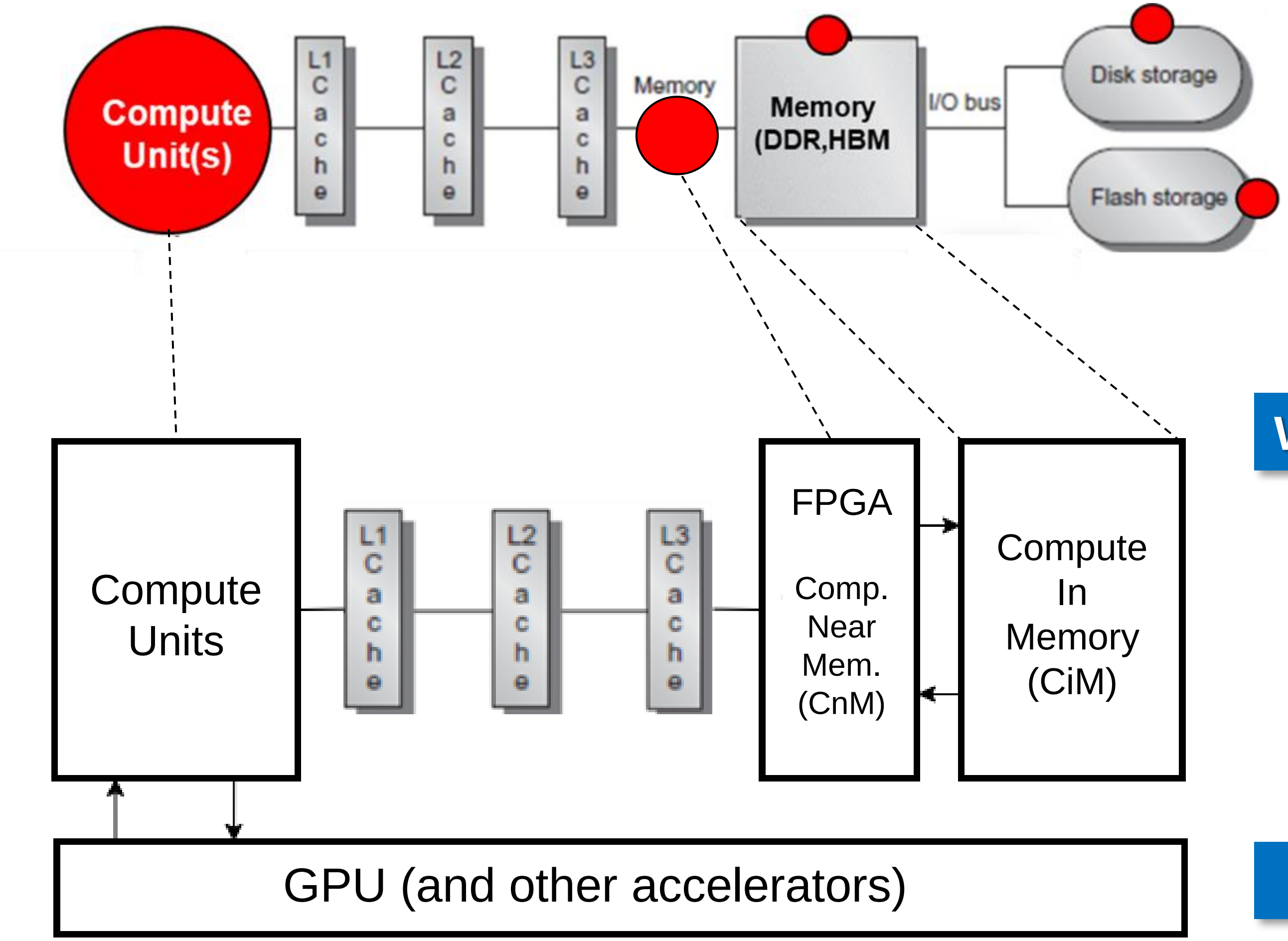
- 8 SambaNova RDUs
- 100s of TFLOPS of compute
- 8 TB total memory

*FireHose: https://stream-benchmarking.github.io/firehose/

** Circus Tent: https://github.com/tactcomplabs/circustent

T2: OneAPI Cross-Platform & ModSim Studies

OneAPI Cross Platform Studies



Goals

- Develop compute-cache architectures on heterogeneous systems using Intel OneAPI
- Perform ModSim studies for notional compute-cache systems
- Assess architecture performance across diverse workloads

Why OneAPI ?

- OneAPI offers multi-platform support for programming CPU, FPGAs, GPUs, etc. together.
- HLS programs can facilitate experimentation: easy kernel scheduling, workload division and memory management

Tools

oneAPI SYCL AdaptiveCpp

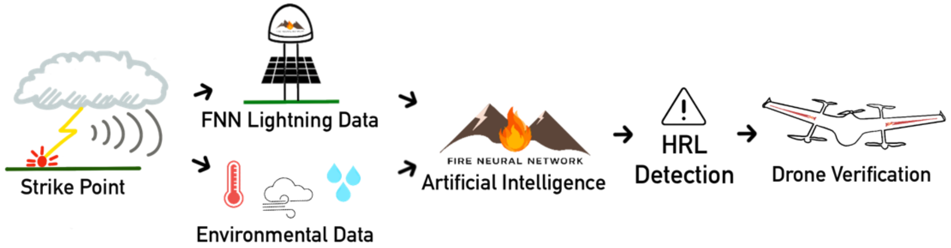
F1

T3: Compute Cache for Edge Computing: FNN (Fire Neural Net)

Overview of FNN

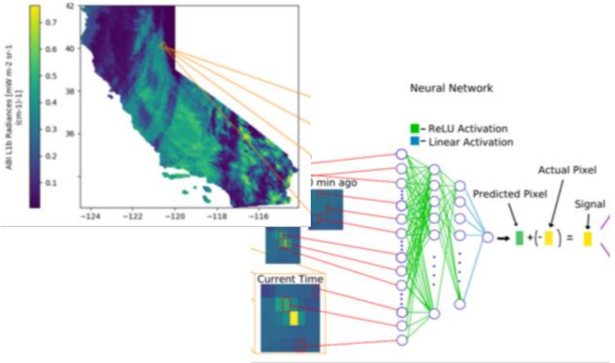
Detection

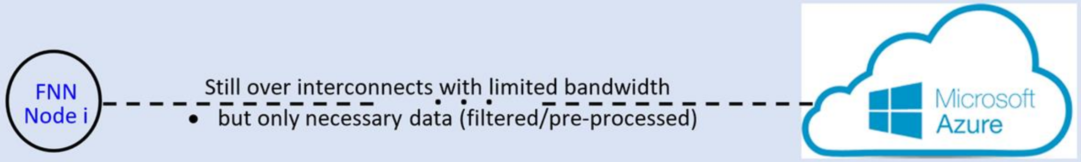
- FNN™ uses its **proprietary detector** technology along with environmental data from satellites and ground stations to recognize ***High-Risk-Lighting™ (HRL™)*** events and issues relevant ignition alerts.



Verification and Mapping

- FNN™ works with its partners to verify and monitor wildfire ignition using ***satellite feeds***, wildfire ***camera systems***, and ***dedicated UAVs***.




FNN Node i

Still over interconnects with limited bandwidth
- but only necessary data (filtered/pre-processed)

Microsoft Azure

Goal & Approach

- Study and develop *compute-cache architectures* toward satisfying the real-time FNN requirement

At the FNN nodes on the computing edge

- Bring compute close to data on the edge for FNN
  - *Heterogeneous computing* using available and emerging technologies for edge computing (FPGAs, CPUs, PIM*, ...)
  - *Minimize the transfer of unnecessary data to servers:* filter and pre-process raw data at FNN nodes
  - *Distribute (off-load) computation and intelligence* as appropriate to nodes

At the cloud servers

- Servers focus on processing (inference) using *already filtered and pre-processed data*
- Explore the use of other heterogeneous & emerging computing technologies and platforms *(e.g., GPU, TPU**)*

*PIM: Process-In-Memory

**TPU: Tensor Processing Unit

Milestones, Deliverables & Budget

Milestones

- SMW24: Showcase midway progress on framework, platform, and interconnect exploration
- SAW24-25: 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 ≥ 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 and scaling studies on *devices* and *applications* of interest
  - Perform *oneAPI* cross-platform & ModSim studies for compute cache
  - Develop a compute cache system for *edge computing* for FNN (*Fire Neural Net*)

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