CXL Memory Management for the CMS L1 Scouting System and Beyond

Guilherme Paulino

CERN Openlab Summer Student

July 8 – September 6, 2024

Supervisors:

Giovanna Lazzari Miotto

Thomas Owen James

CERN openlab

Micron openlab partner

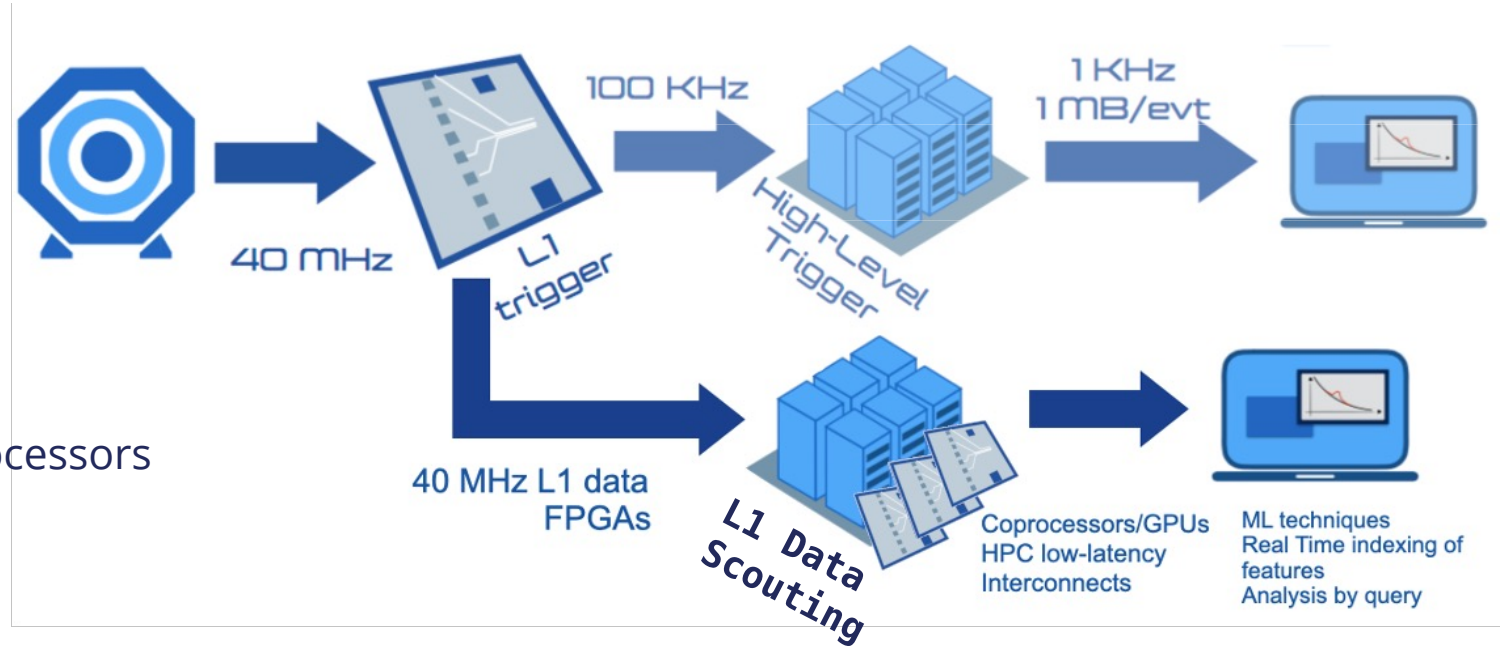
L1 Data Scouting

CMS

- 40MHz bunch crossing rate
- Not feasible for full readout
- L1 Trigger selects by signatures

L1 Data Scouting

- Capture L1 primitives from trigger processors
  - For every bunch crossing
- **Full access to physics**
  - Unconstrained by the L1 latency and accept rate



https://indico.cern.ch/event/1356148/contributions/5818261/

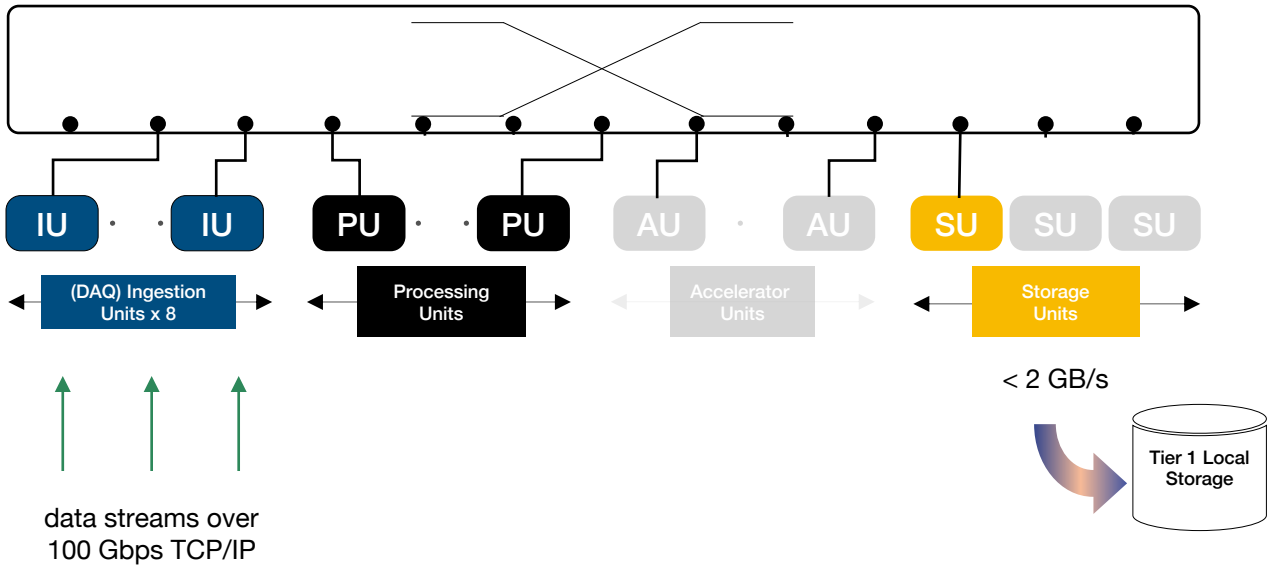
L1DS Online Processing

- Upcoming Hi-Lumi LHC upgrade will substantially increase data volumes, requiring advanced data processing techniques.

L1DS: Run 3

1. A few **ingestion servers** receive data from the L1DS over 100G
2. Incoming data stored to a local **ramdisk buffer**
3. Immediately available to processing farm (mounted **over NFS**) ~2min latency

Experimental setup (in operation)

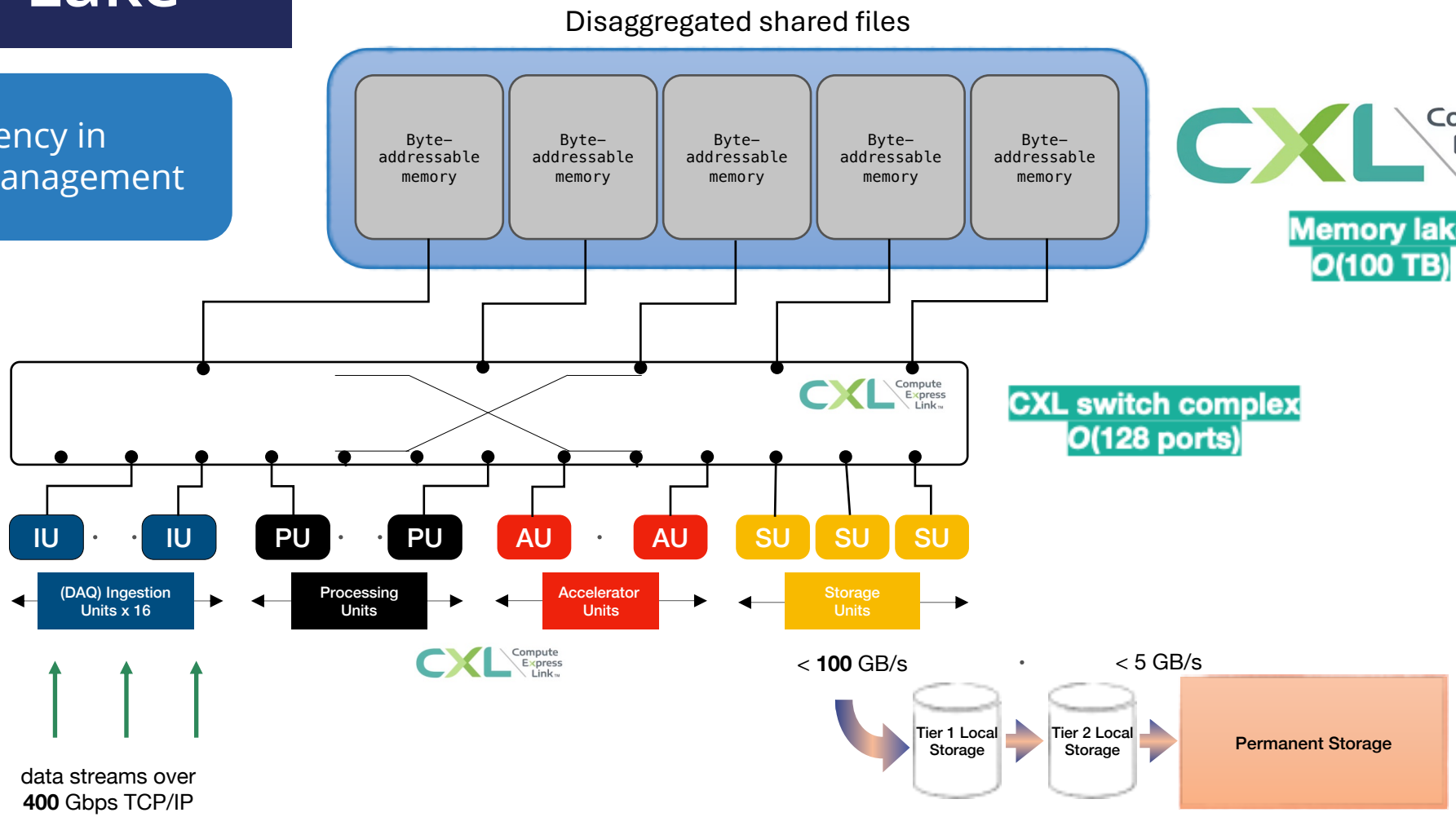


https://indico.cern.ch/event/1356148/contributions/5818261/

Memory Lake

Cache coherency in
Multi host data management

High Bandwidth
Low latency



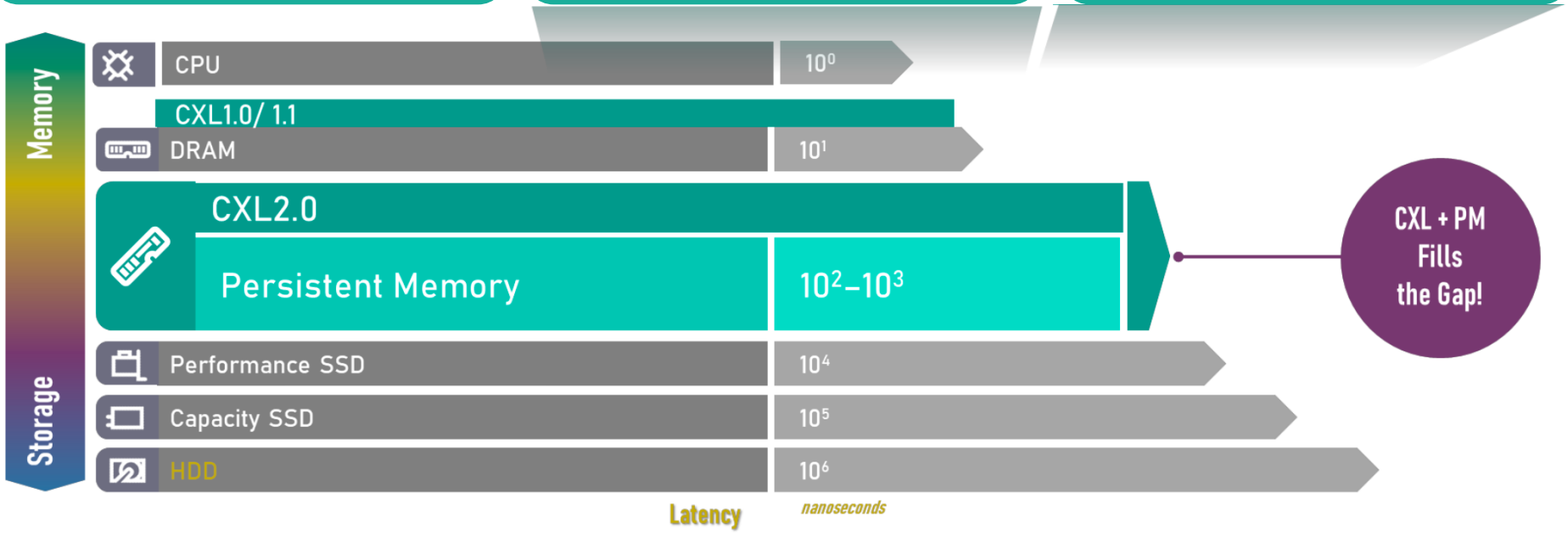
https://indico.cern.ch/event/1356148/contributions/5818261/

Memory Tiering

Standardized Memory Management and Interfaces

Pmem capability to function as both memory and Storage

DAX direct access eliminates OS-level file caching



Adapted from CXL 2.0 White paper
https://computeexpresslink.org/resource-library/

CMS Prototype

Supermicro server

2x AMD EPYC 9454 Genoa

- 96 CPU cores, 460 GB/s peak bandwidth per socket
- 24x 16 GB DDR5-4800 RDIMM
- PCIe gen 5.0 x128

2x Micron CZ120 256 GB

- CXL 2.0 compliant Type 3 memory expansion
- Support to CXL.mem transactions to attached DDR4 media and CXL.io for inband device management
- PCIe Gen5 x8 data lanes
- 36 GB/s peak memory R/W bandwidth

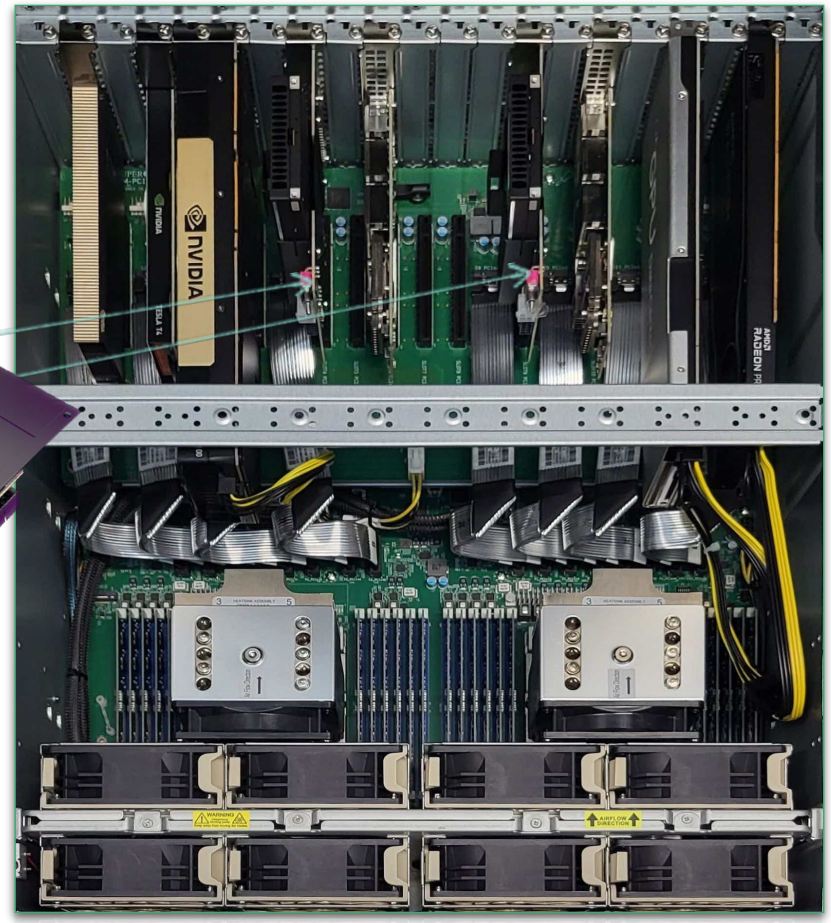
Software:

Linux RHEL 9.3 (kernel 6.8-rc4 **Famfs**-enabled)

https://github.com/cxl-micron-reskit/famfs

CZ120

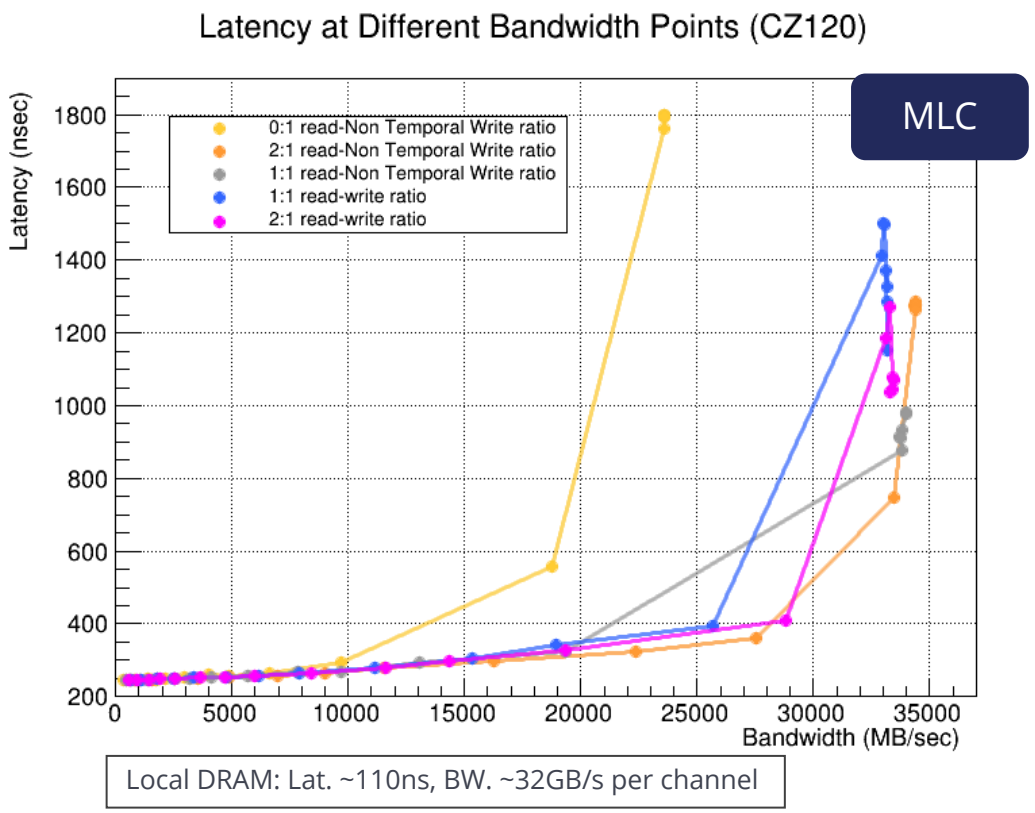
Provided by Micron openlab partner



CXL host node at CMS USC.
*Server supplied by **E4 Computer Engineering***

CZ120 Benchmark Results

- **NUMA mode (System-RAM)**
  - Intel Memory Latency Checker (MLC)
- **Modified to support DAX attached memory**
  - STREAM
  - Multichase
  - StressAppTest



Local DRAM: Lat. ~110ns, BW. ~32GB/s per channel

STREAM in DAX mode using **Famfs files**

Function	Rate MB/s	Avg time	Min time	Max time
Copy:	20060.3	0.807057	0.797596	0.816493
Scale:	19964.4	0.809129	0.801426	0.816096
Add:	22306.9	1.082312	1.075900	1.101913
Triad:	22317.9	1.081636	1.075371	1.090843

Array size = 1G

Thank you!

Guilherme Paulino
gpaulino@cern.ch
ra117119@students.ic.unicamp.br

CERN openlab

Micron openlab partner

Backup Slides

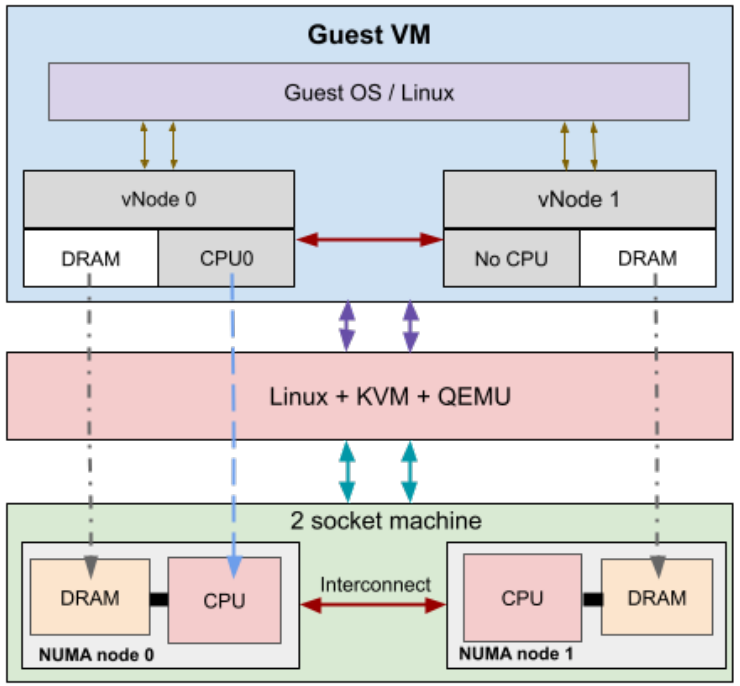
CERN openlab

Micron openlab partner

What's next?

Next steps:

- Compare performance using shared files
  - ramdisk + NFS
  - **CXL + Famfs**
- Define benchmark test cases
  - Workload type
  - File sizes, access patterns, concurrency levels
- Emulation (QEMU + KVM)
  - Multi-host environment
- Integrate SCDAQ software for demonstration
- System design
  - FPGA-based AI Accelerators



emucxl Framework
https://arxiv.org/html/2404.08311v1