Real-time Data Processing for CMS Level-1 Trigger using **CXL Memory Lake Architecture**

Giovanna Lazzari Miotto, Thomas Owen James, Emilio Meschi

CERN-EP CMD - CMS Data Acquisition & Trigger

2024 CERN openlab Technical Workshop (indico)

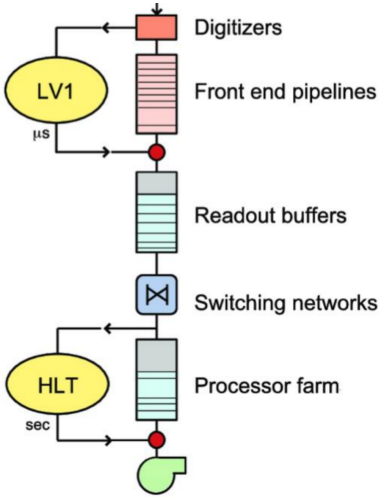
26 March 2024

openlab partner

Outline

- CMS Trigger System
- L1 Data Scouting
  - L1DS Demonstrator
  - L1DS Online Processing
  - Proposal for a Memory-Lake Architecture
- Compute Express Link
- Prototype
  - Validation
  - Benchmarks
- Conclusion

CMS Trigger System

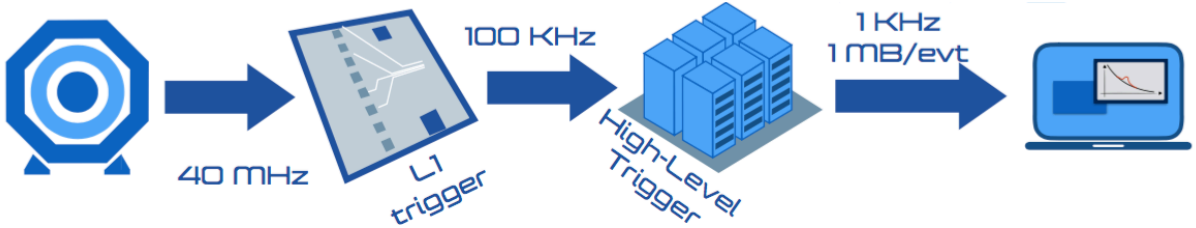


~2.4 billion collisions per second at the LHC full bunch-crossing rate (40 MHz):

- ~1.5 MB events → ~480 Tbps transfer rate
- Full readout is **not technically feasible**
- Offline storage and CPU budget limitations

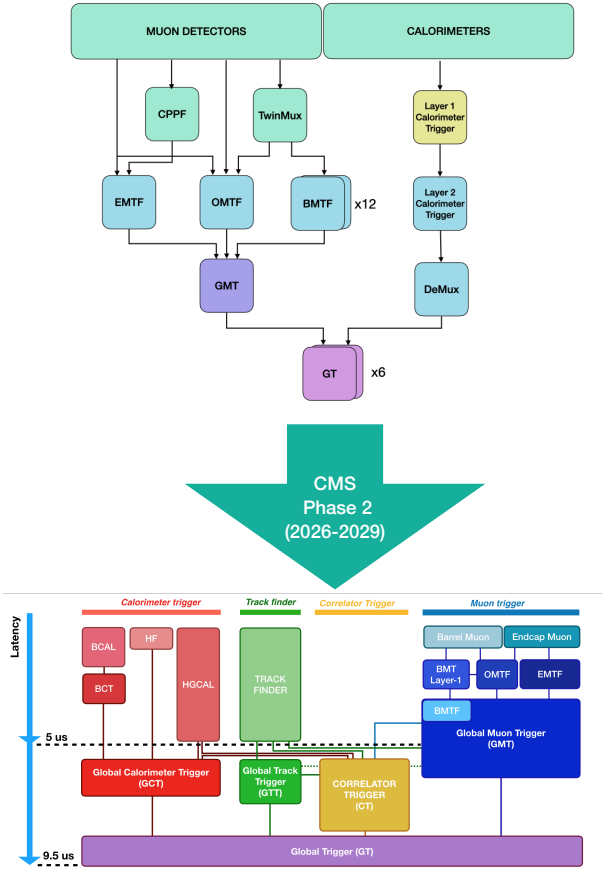
Two-tier mechanism with a fast trigger to select interesting events

- **Level-1 (L1) Trigger:** *fixed* latency of **3.2** microseconds → *FPGA*
- **High-Level Trigger (HLT):** flexible latency (~500ms) per event → CPU / GPU



L1 Data Scouting: Analysis at 40 MHz

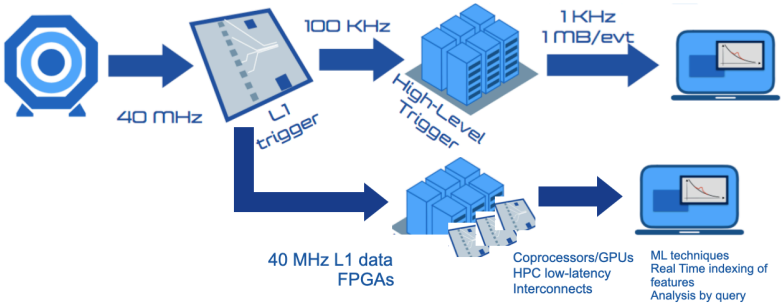
- We can still work with all bunch-crossings with **L1 trigger primitives**
- **These primitives will have improved resolution as of LHC's Run 4**



L1 Trigger, current (top) and post-Phase 2 (bottom)

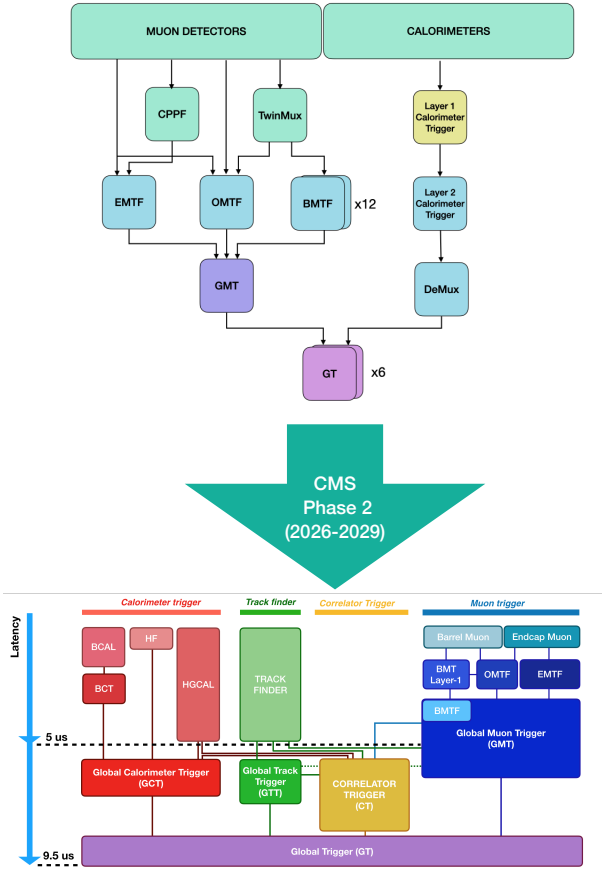
L1 Data Scouting: Analysis at 40 MHz

- We can still work with all bunch-crossings with **L1 trigger primitives**
- **These primitives will have improved resolution as of LHC's Run 4**



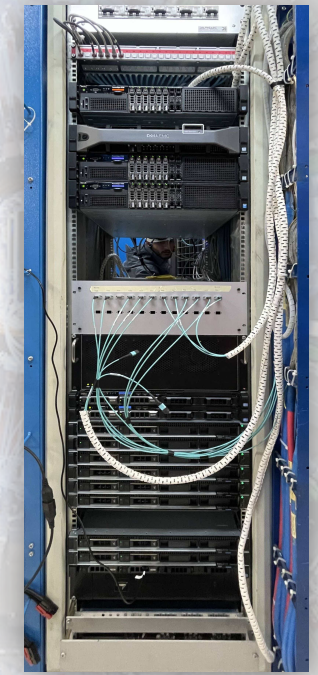
The Level-1 Data Scouting (L1DS) project leverages this reduced data:

- Full rate analysis of certain topologies
- Diagnostics and monitoring: bunch-crossing correlations, independent lumisection measurements at the bunch-crossing level
- Exploration of **new physics** not aligned to standard physics triggers
    - **Reduce bias** of physics research!



L1 Trigger, current (top) and post-Phase 2 (bottom)

L1DS Demonstrator

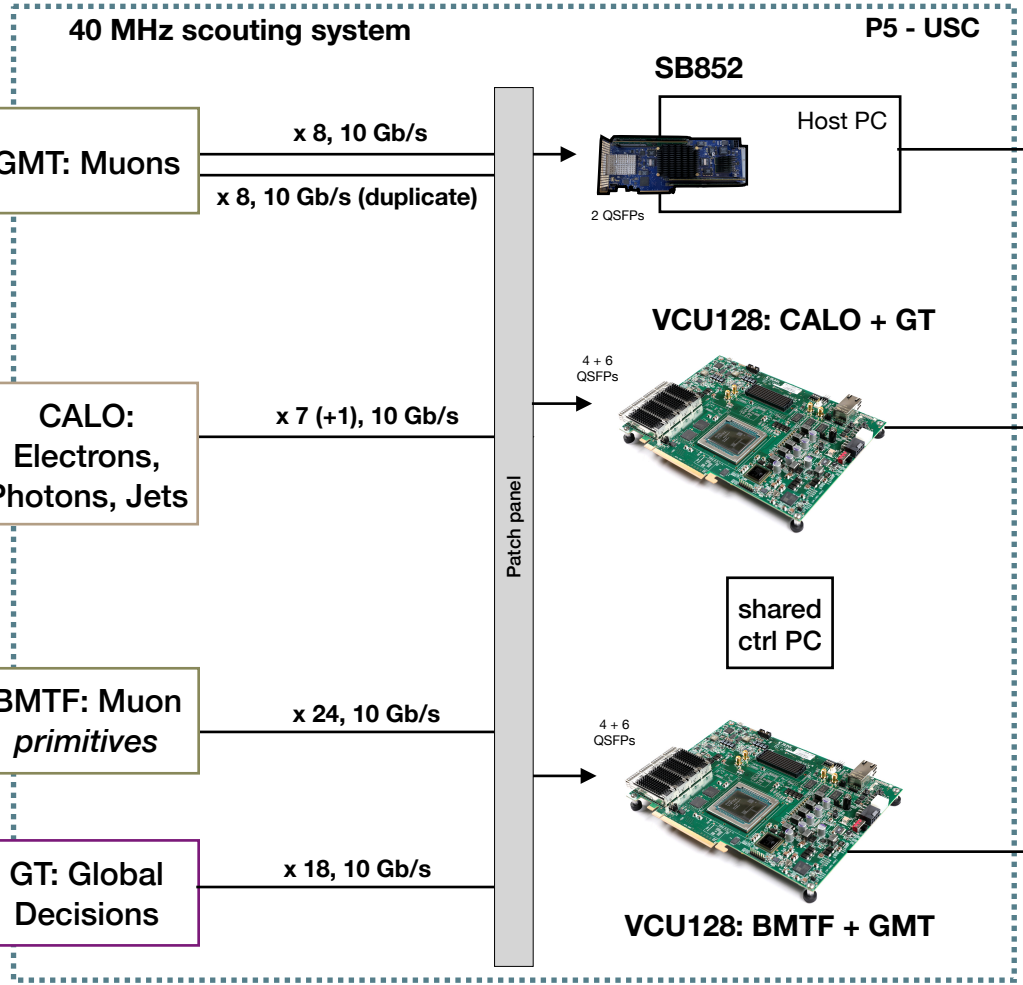


L1DS at CMS USC

GMT

CALO

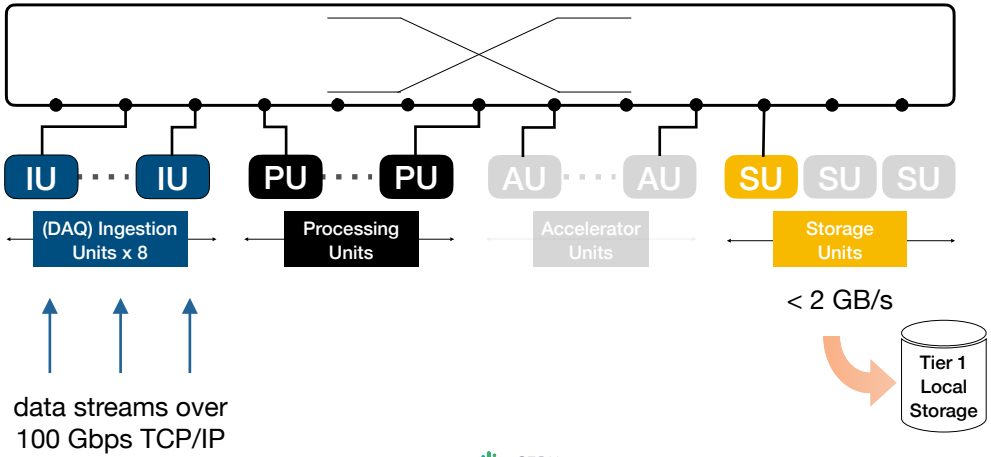
BMTF



L1DS Online Processing: Run 3

Experimental setup (in operation throughout LHC Run 3)

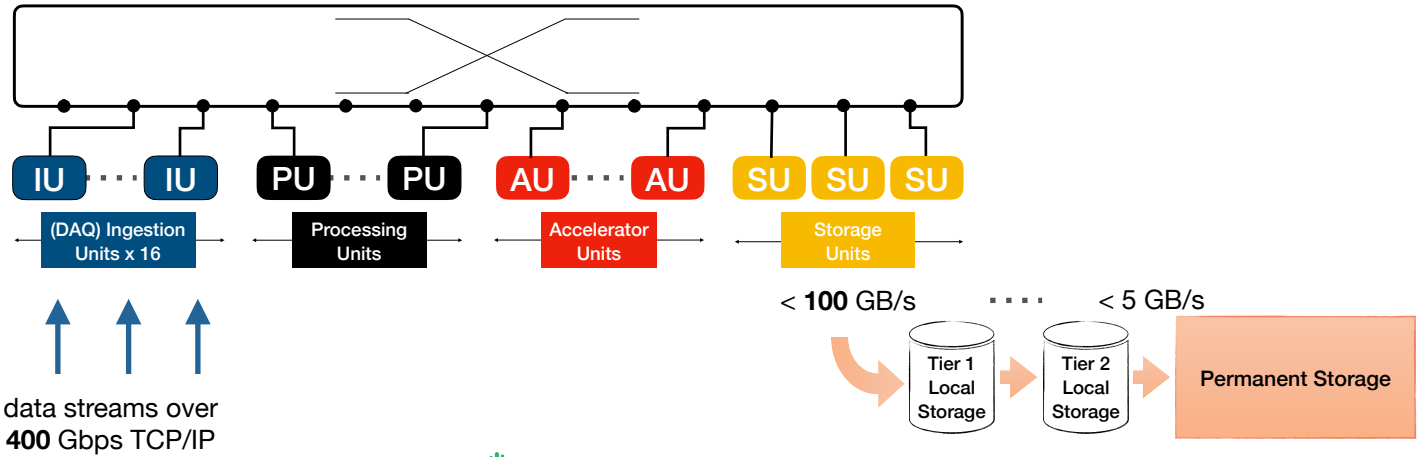
1. A few **ingestion servers** receive data from the L1DS readout boards over 100 Gb/s connections
2. The incoming data is stored to a **local ramdisk buffer**
3. Immediately available for access to **processing farm** (ramdisk mounted **over NFS**) ~2min latency window



L1DS Online Processing: Phase 2

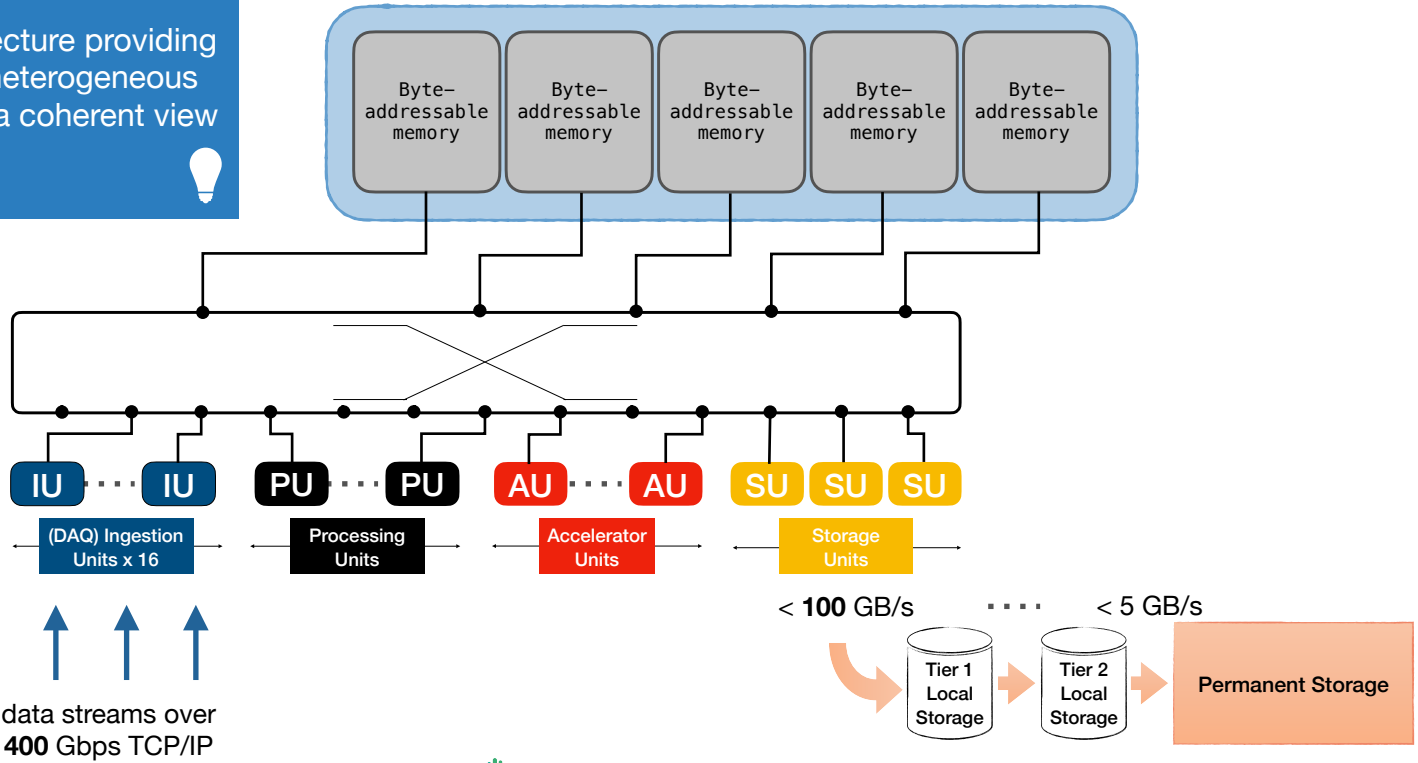
For CMS Phase 2 and LHC Run 4, upgrades will enable **more complex, real-time data analysis** on inbound data streams

- Including dedicated accelerator nodes, e.g., GPUs, FPGAs
- Significant increase in data volume, to be mitigated by reduction strategies
  - extract topological information, fake / zero rejection, invariant mass, histogramming
  - **store only analysis products**
- Data lake for memory expansion? --> **inefficient data copy over NFS**



L1DS Online Processing: Phase 2 **+ Memory Lake**

A *memory lake* architecture providing shared memory for heterogeneous computing units with a coherent view

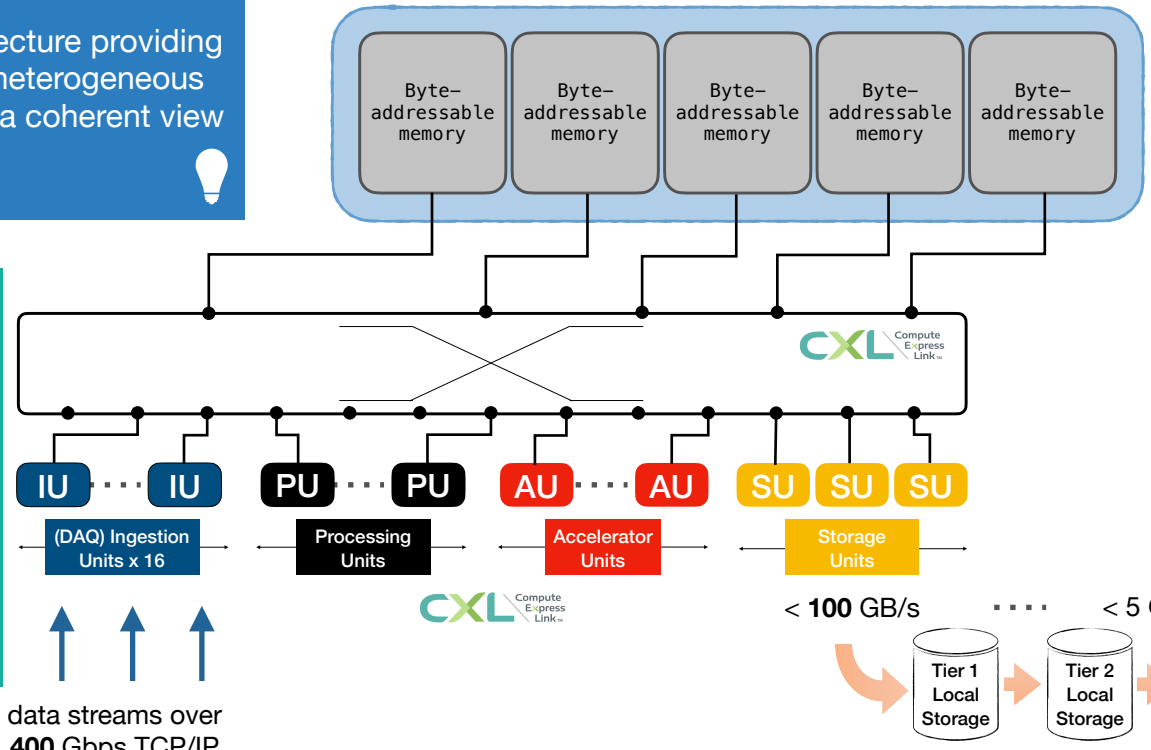


L1DS Online Processing: Phase 2 **+ Memory Lake**

A *memory lake* architecture providing shared memory for heterogeneous computing units with a coherent view

The CXL standard promises...

- Near-memory compute (NMC)
- High-bandwidth
- Low-latency
- Cache coherence



Memory lake *O*(100 TB)

CXL switch complex *O*(128 ports)

CXL-enabled devices

CERN openlab

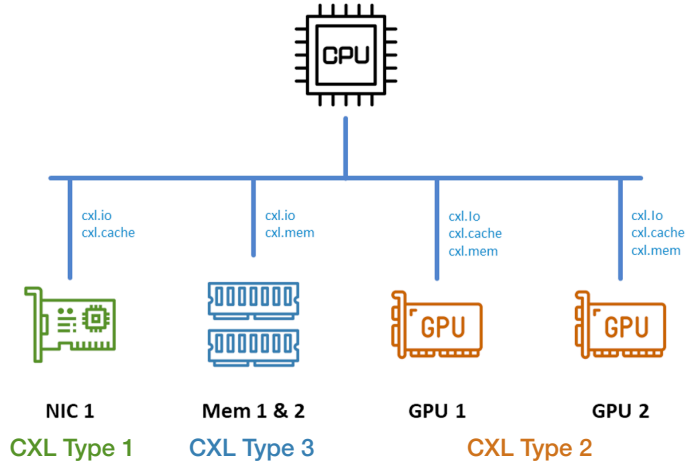
Compute Express Link

Emerging open standard for **high-bandwidth heterogeneous, disaggregated computing**

- **Unified, coherent** memory space across CPUs & devices
- Resource sharing. Shared & fabric-attached **memory pools**
- PCIe Gen 5 physical layer
- Improved data & operand movement between hosts, accelerators

Dynamic multiplexing of 3 protocols:

`CXL.io`: traditional PCIe block I/O
`CXL.mem`: device memory
`CXL.cache`: system memory



The three CXL device types and their protocols.
Adapted from H3 (source)

First CXL 2.0-compliant memory modules (CMMs) are already in the horizon, e.g., Micron CZ120

Don't miss the next talk:
"Solving HPC challenges with Micron CXL attached memory products"

Prototype at CMS

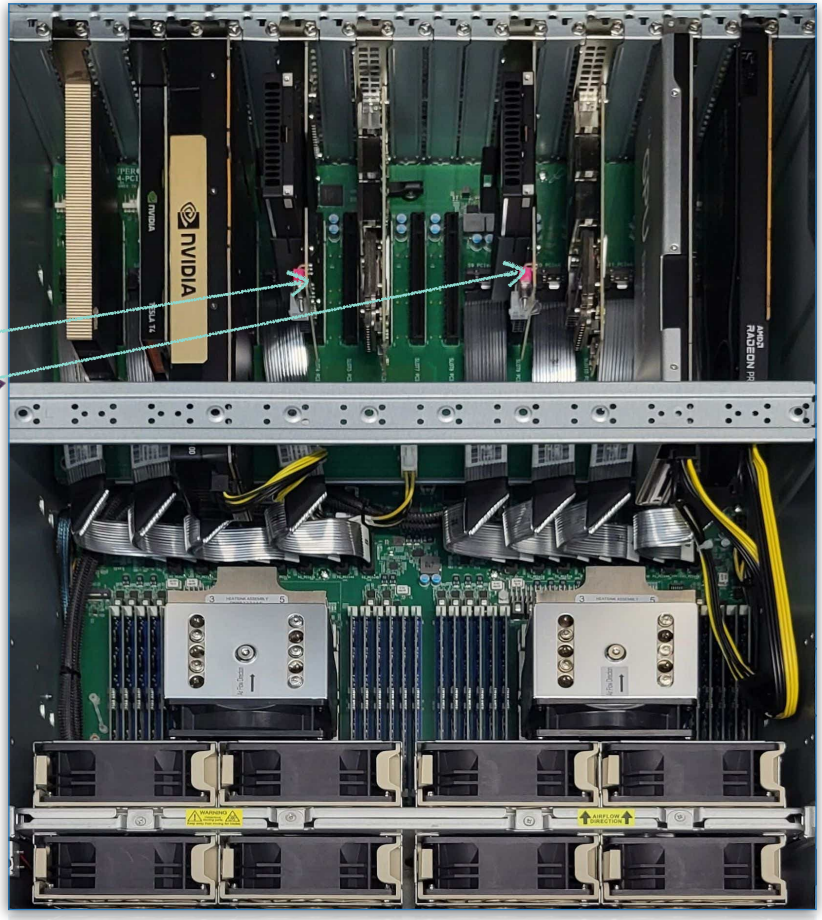
Supermicro server

2x **AMD EPYC 9454** 'Genoa' 48-Core @ 2.75 GHz, SMT on
2x **96 CPU**, 460 GB/s peak bandwidth per socket
2x **256 GB L3** shared cache
24x **16 GB DDR5**-4800 RDIMM

2x **Micron CZ120** 128 GB
'Type 3' CXL memory expander
PCIe Gen5, x8 data lanes
36 GB/s peak memory R/W bandwidth

Software:
Stock **RHEL 9.3** (kernel 5.14.0-362.8)
AutoNUMA on
`libcxl` v78

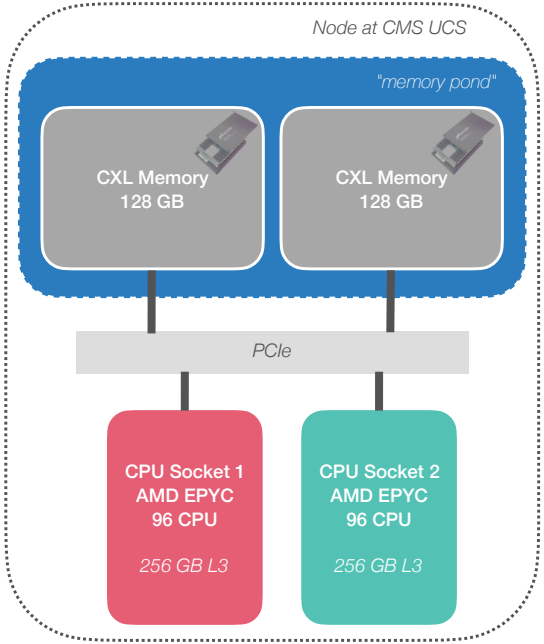
Note: WS1 'pre-engineering' samples. Units from an updated version are arriving soon



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

Prototype: validation and integration

"Memory pond"



Validated and benchmarked (*see next slide*) both as NUMA and DAX device

- As headless NUMA nodes (one per CMM):
  - Expanding memory capacity of each socket
  - Measured bandwidth and latency with MLC (Memory Latency Checker)
  - Memory manipulation with `numactl`
- As PCIe shared DAX devices:
  - Expanding shared NVM
  - Measured sustained bandwidth in processing workloads with STREAM
  - Validation with concurrent `daxio` operations

IU Ingestion Unit

`scdaq` experimental ingestion & buffering unit [Gitlab]

- Part of the scouting demonstrator's software stack
- Near 100 Gbps board-to-host link saturation with multiple streams
- Ongoing refactors, including the output sink: **shift to direct access + `mmap`**

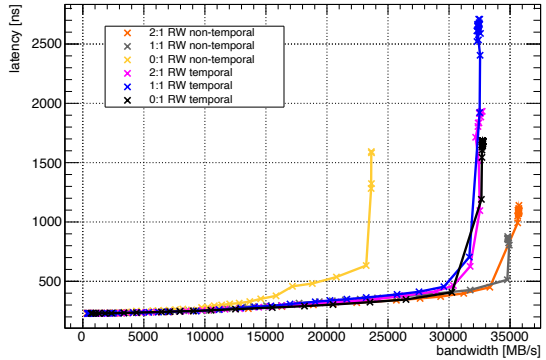
Prototype: benchmarks

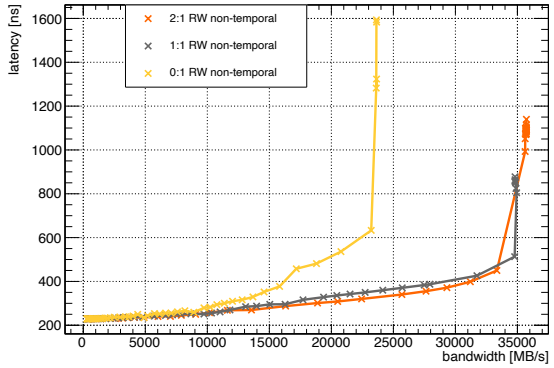
MLC Benchmark [1,3]

Intel Memory Latency Checker v3.11

Evaluate memory expansion as headless NUMA domain

Latency vs bandwidth

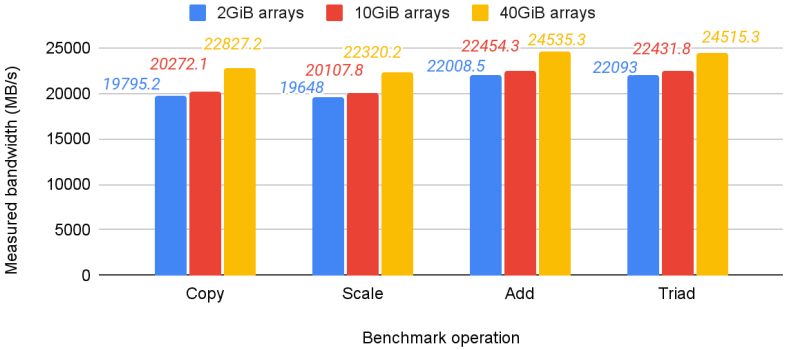




STREAM Benchmark [2,3]

Sustained main-memory bandwidth + computations:
copy (transfer), scale (+ arithmetic),
sum (+ load/store), triad (chained MUL+ADD)

Single CMM (DAX), 2 GiB arrays:



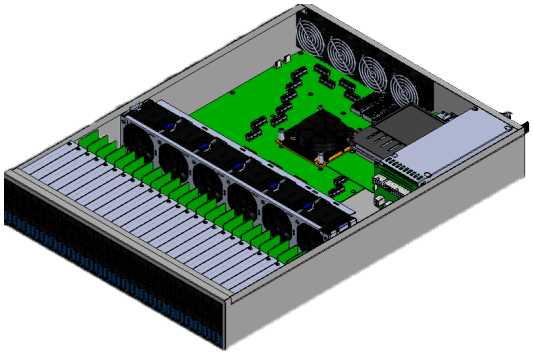
1 MLC
2 STREAM
3 CXL ResKit

CERN openlab

Moving forward

CXL Memory Pooling

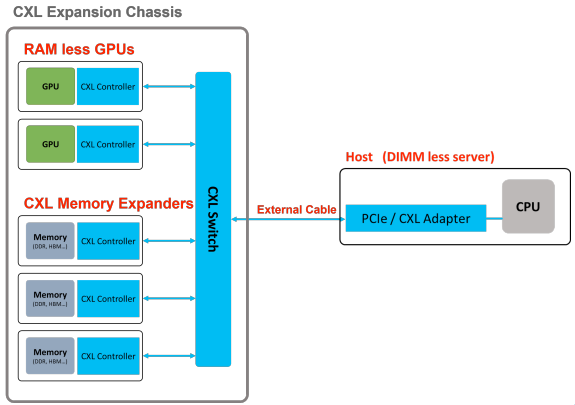
- Topology enabling **multi-host memory expansion chassis**
- **CXL switch, few host connections, few dozen CXL memory modules**
- E.g., Falcon C5024 (H3)



Falcon C5024 memory chassis, H3 Platform

Open problem: **data management in a multi-host system**

- As of CXL 3.0: **UUID-tagged memory allocation**
- Investigating **FAMFS ("Fabric-Attached Memory File System")** [RFC]
  - Scale-out shared-memory FS residing in DAX
  - Handles metadata, space allocation in a sharable way
  - Support for POSIX RW
  - Apt for serial-sharing and ephemeral storage



Summary

- At CMS, Level-1 Data Scouting (L1DS) seeks to enable **quasi-online, heterogeneous processing of L1 trigger primitives at the full 40 MHz rate**, exposing unbiased collision data that is otherwise lost in the selection process.
- For transiently storing this data and intermediate processing products, **CXL memory modules (CMMs) are a remarkable candidate to power a memory lake due to CXL's coherent, shared-memory protocols**.
- A **prototype has been set up at the CMS service cavern**, benchmarked and validated, with CMM integration into existing L1DS tooling well under way.

Next steps:

1. Study feasibility of replacing ramdisk with globally-accessible CXL-based memory (up to 300 TB), including existing data management solutions, e.g., FAMFS
2. Introduce CXL 'Type 2' accelerators & validate memory coherence
3. Validate CXL switches' support for 400 Gbps sustained bandwidth per port

CERN openlab

Micron™ openlab partner

CERN CMS Compact Muon Solenoid

Thank you

Acknowledgments:
Thanks to the **CERN openlab-Micron** collaboration for funding this ongoing project, and to our collaborators from Micron: **Jason Adlard, Tony Brewer, Glen Edwards, John Groves, Andrey Kudryavtsev.**

Any questions?

Giovanna Lazzari Miotto
CMD L1 Data Scouting
glazzari@cern.ch

To learn more about L1 Data Scouting:

Check out our TWiki and publications.

Ardino et al. "*A 40 MHz Level-1 Trigger Scouting System for the CMS Phase 2 Upgrade*". 2023. [doi]

The next **L1 Data Scouting Workshop** is taking place ~October 2024, near Geneva.

CERN openlab

Micron openlab partner

CERN CMS Compact Muon Solenoid


CERN
openlab