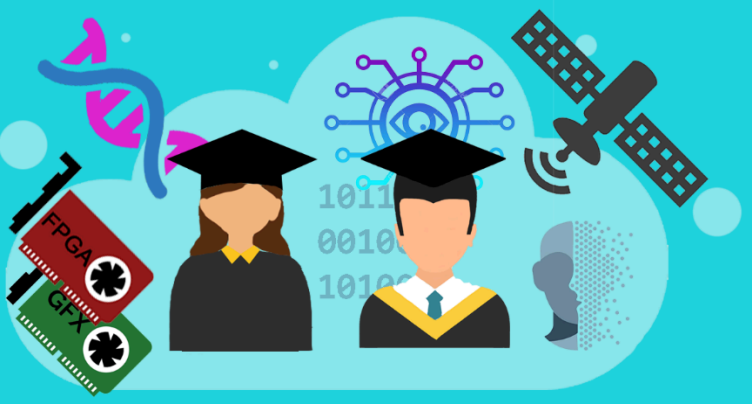
Diploma Thesis

Microproccessors and Digital Systems Laboratory



CXL Tier Profiler: In-Band Performance Telemetry for Contention-Aware Page Placement in Multi-Host CXL Memory Systems

The rapid growth of memory-intensive workloads such as AI inference, in-memory databases and graph analytics has made memory capacity and bandwidth a first-order constraint in modern datacenters. Provisioning sufficient local DRAM for every server is increasingly expensive, and provisioned memory is often poorly utilized: large portions of DRAM remain idle on some servers while others exhaust their capacity. This mismatch has motivated a shift from server-local memory toward disaggregated and pooled memory architectures.

Compute Express Link (CXL) has emerged as the main interconnect enabling this shift. Built on top of PCIe, CXL allows a host processor to access external memory devices coherently, through ordinary load and store instructions, exposing them to the operating system as additional NUMA nodes. With CXL switches and multi-headed memory devices, a single memory expander can be shared by several hosts simultaneously, forming a memory pool that is dynamically partitioned across servers. Pooling improves utilization and reduces cost. However, it also means that each host now accesses a memory tier whose performance depends not only on its own accesses, but also on the traffic generated by every other host attached to the same device.

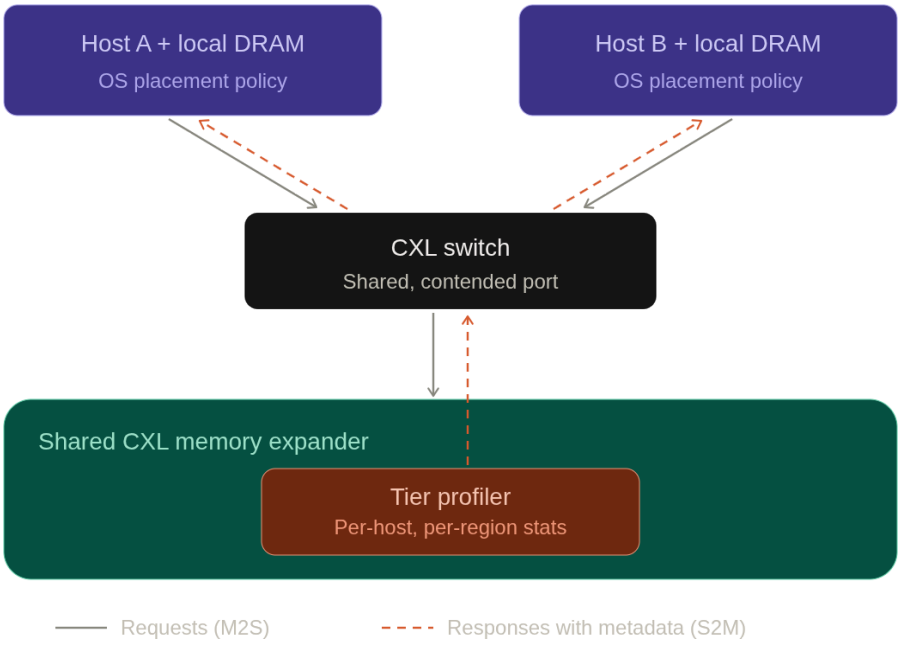


Fig. 1: *Multi-host CXL system with a device-side tier profiler. Performance metadata is carried in-band within CXL.mem device-to-host (S2M) responses, over the same links used for memory traffic.*

As a result, the effective latency and bandwidth of a shared CXL memory tier are not static properties. When multiple hosts contend on the same CXL memory port, queueing within the device controller and the switch increases access latency well beyond its unloaded value and reduces the bandwidth available to each host. Existing tiered-memory policies, such as Linux NUMA balancing in tiering mode and TPP, make placement decisions using only information observed by the local host, namely page access recency sampled through page faults or hardware counters, combined with a fixed view of each tier's latency. A host therefore cannot distinguish whether degraded CXL performance originates from its own access pattern or from contention caused by other hosts. Consequently, it may continue promoting and demoting pages based on hotness alone, while the actual bottleneck lies in a shared device queue that it cannot observe. Under contention, such placement decisions can negate the benefits of pooling: latency-critical pages remain in a congested tier, while local DRAM capacity is occupied by pages that are not performance-critical.

Addressing this problem involves three main challenges:

1. **Contention observability.** Contention on a shared CXL memory port is only visible at the point where traffic from multiple hosts converges. The memory device is therefore the only component that can accurately capture aggregate behavior, including per-host request rates, queue occupancy and per-region access frequency. This thesis will identify which device-side metrics most reliably reflect contention and its impact on host-perceived performance.
2. **In-band metadata transport.** The collected information must be delivered to each host with minimal overhead and without introducing a separate management channel. The CXL protocol already provides a basis for this, through the QoS telemetry load indication carried in device-to-host responses and the metadata fields defined in CXL.mem messages. Recent specification revisions have also introduced device-side performance and hotness monitoring units. This thesis will investigate how richer performance metadata can be encoded within CXL.mem messages, and evaluate the trade-off between metadata granularity and interconnect overhead.
3. **Placement policy design.** The host operating system must use this metadata to decide which pages reside in local DRAM and which in the shared CXL tier. It must also determine when pages should be migrated, and how migration activity should be restricted when the device is already congested. The policy must remain effective under varying contention levels while keeping its own runtime and migration overhead low.

Since access to multi-host CXL hardware with configurable devices remains expensive and limited, the gem5 simulator will be used to model the system. gem5 is a modular, open-source full-system simulator widely used in computer architecture research, and prior work has already extended it with CXL memory models. Its modularity allows the modeling of multiple hosts sharing a CXL memory device with arbitrary latency, bandwidth and queueing parameters.

This thesis addresses the problem of contention-aware page placement in multi-host CXL memory systems. The goal is to design a CXL tier profiler, a device-side monitoring unit that tracks per-host and per-region performance behavior on the shared CXL memory port and conveys this information to each host in-band, through CXL protocol messages. A host-side OS or driver policy will use the received metadata to guide page placement between local DRAM and the shared CXL tier, issuing `mbind` and `move_pages` calls to apply its decisions. The approach will be evaluated on a simulated multi-host CXL topology in gem5 and compared against host-only baselines, including first-touch, interleaving and TPP.

The scope of this thesis includes:

- A characterization of multi-host contention on a shared CXL memory port in gem5, quantifying how latency, bandwidth and placement quality degrade as the number of contending hosts

and their traffic intensity increase, and identifying the device-side metrics that best capture contention.
- The design and implementation of a CXL tier profiler within the simulated memory device, which maintains per-host and per-region statistics, such as request rate, queue occupancy and access hotness, and encodes a compact summary of them as metadata within CXL.mem messages.
- An OS/driver placement policy that combines host-local access information with the device-provided metadata to drive page promotion, demotion and migration throttling, preventing latency-critical pages from residing in a congested shared tier.
- An empirical evaluation on multi-tenant workloads, such as databases, key-value stores and graph analytics, comparing the proposed approach against host-only placement policies in terms of application performance, fairness across hosts, migration overhead and the interconnect overhead of the metadata itself.

PREREQUISITES:

Familiarity with:

- Computer Architecture
- Operating Systems
- Coding Languages (C++, Python)

RELATED MATERIAL:

- TPP: Transparent Page Placement for CXL-Enabled Tiered Memory (ASPLOS '23) https://dl.acm.org/doi/10.1145/3582016.3582063
- Pond: CXL-Based Memory Pooling Systems for Cloud Platforms (ASPLOS '23) https://dl.acm.org/doi/10.1145/3575693.3578835
- Demystifying CXL Memory with Genuine CXL-Ready Systems and Devices (MICRO '23) https://dl.acm.org/doi/10.1145/3613424.3614256

CONTACT INFORMATION:

Spyros Chalkias: schalkias@microlab.ntua.gr

Manolis Xenos: mxenos@microlab.ntua.gr

Dr. Manolis Katsaragakis: mkatsaragakis@microlab.ntua.gr

Prof. Dimitrios Soudris: dsoudris@microlab.ntua.gr

Asst. Prof. Sotirios Xydis: sxydis@microlab.ntua.gr