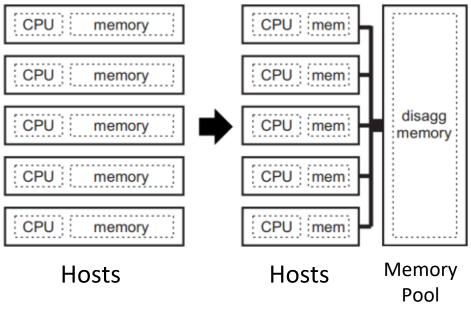
Memory Page Prefetching for Disaggregated Memory Systems

Kobayash and Sato Lab, C2IM1503, Zecheng Li

Background & Challenges

Memory Disaggregation



- Memory are **tightly coupled within each server** in traditional datacenters, causeing wide **resource underutilization**.
- Memory disaggregation is to **leverage available memory outside server boundaries like the memory from other server nodes**.
- The memory **latency increases** when the memory system is placed **far** from the hosts, leading to **performance degradation**.

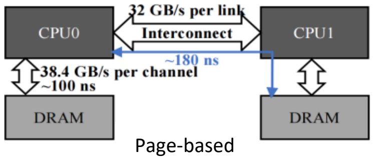
Current Approaches and Challenges

Kernel-based systems

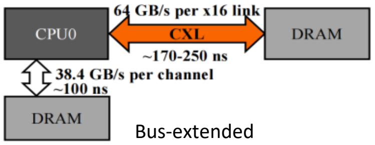
- Relies on the **virtual memory** systems, and do not need extra hardware in current remote accessing systems. Remote memory pages are accessed by triggering **page faults** through the OS.

Bus-extended systems

- Relies on the extra **coherent interconnects hardware** for remote accessing (e.g., **CXL-based** system). Results in more efficient **cache-line** access within a disaggregated memory systems.



Page-based



- ◆ ***Kernel-based*** systems causes **overhead** in **dirty data tracking** and **eviction** at page level, causes a **significant waste.**
- ◆ ***CXL-based*** systems overly focuses on **cache-line accesses** that it overlooks the overall **hit rate of memory pages**.

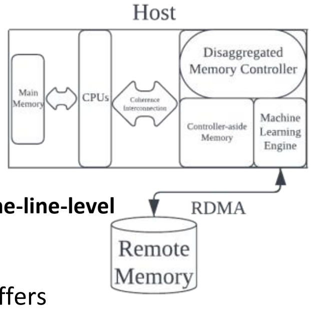
◆ To sum up, excessively favoring either **cache-lines-level** or **page-level** accesses is **not advisable**, a **balance** should be struck between them.

Approach

Multi-granularity ML-based page prefetching for CXL-based memory disaggregation

Prefetcher Architecture

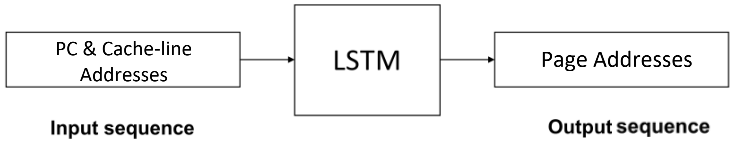
- ● Propose a cache-line based page prefetcher based on CXL-based system.



-Memory access analysis: **cache-line-level**
-Prefetching: **page-level**

- **CXL cache coherence** offers dataflow to train ML engine.
- Utilize controller-aside & host memory as an **additional cache** for remote pages.
- Page prefetching is initiated once **a certain quantity of cache lines** is observed by the prediction engine.

ML-based Prediction Model



- The **LSTM** model has proven effective in **cache behavior prediction** using the sequences of the addresses.[Zhan Shi et al., MICRO 52].
- Analyzing the **attention layer** reveals that caching decisions are primarily **influenced** by a **few key addresses** within the long history. [Zhan Shi et al., MICRO 52].
- This information allows us to create a **smaller model** to handle individual input **values** with **significant influence** with even **better precision** like **SVM**.

Conclusions

- This work aims to improve the **page prefetching** of **CXL-based disaggregated memory** systems using **cache-line** information, through **machine learning** , enhancing the **hit rate** of remote prefetching pages.
- This work will conducting **benchmark** tests like SPEC2017 on **CXL-based disaggregated** system using simulator, analyzing cache **access patterns** and page prefetching relationships, exploring **machine learning models**, with the ultimate goal of **deploying the method** on a CXL-capable simulator to validate the findings.