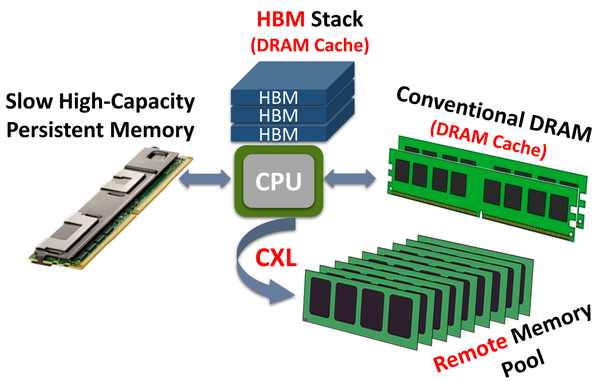
Toward Rethinking the Management Techniques in Emerging Memory Systems

Maryam Babaie, Ayaz Akram, Jason Lowe-Power
DArchR Research Group, Computer Science Department, University of California Davis

UC DAVIS
UNIVERSITY OF CALIFORNIA

Motivation

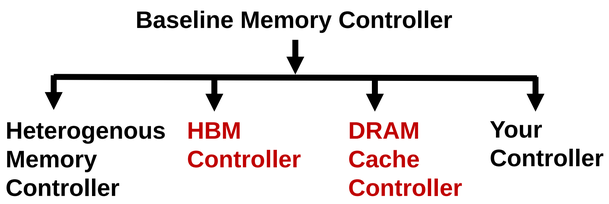
- ❑ HPC systems rely on **heterogeneous** memories.
  - *Intel*'s *Knights Landing, Cascade Lake, Sapphire Rapids*
- ❑ In these systems, fast memories can be used as **DRAM cache** to slow memories.
- ❑ **Disaggregated** memory resources also will use local DRAM as a cache to a remote memory.
- ❑ We need to rethink the memory managements.
- ❑ However, there is not an accurate model in the research community.



We extend gem5 to enable design space exploration of future heterogeneous/disaggregated memory systems. We add support for:

1. DRAM Cache
2. HBM2 interface and controller
3. Modular memory controller design

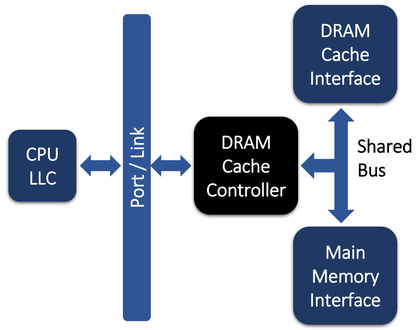
We refactored gem5's memory controller to extend modularity, as follows:



gem5 DRAM Cache Support

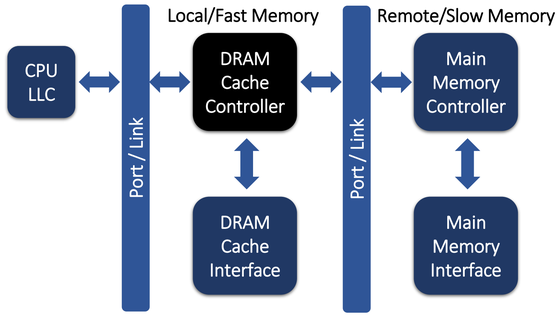
A. Dedicated DRAM Cache Controller

1. Unified Cache/Memory Controller (UDCC)
- Tightly-coupled DRAM cache and main memory
- Connection: shared bus
- Models Intel's Cascade Lake



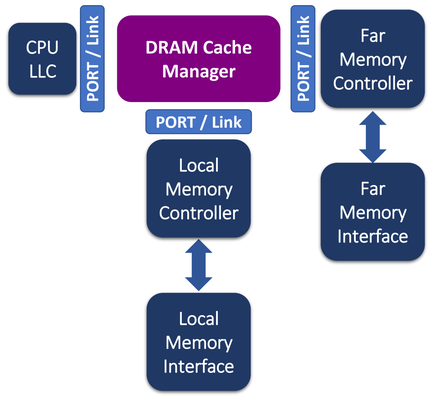
Goal: Enabling cycle-level analysis of heterogeneous and disaggregated memory systems to develop *new data management* schemes.

2. Disaggregated DRAM Cache Controller
- Flexible combination of interfaces
- Connection: configurable link



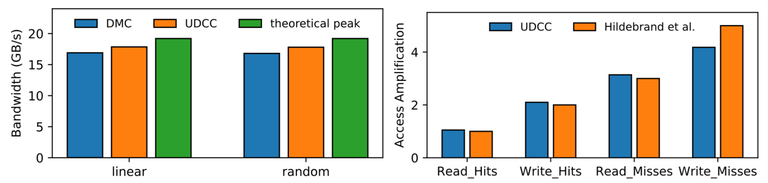
B. Decoupled DRAM Cache Policy Manager

- Supports multiple cache architectures in parallel



Verification

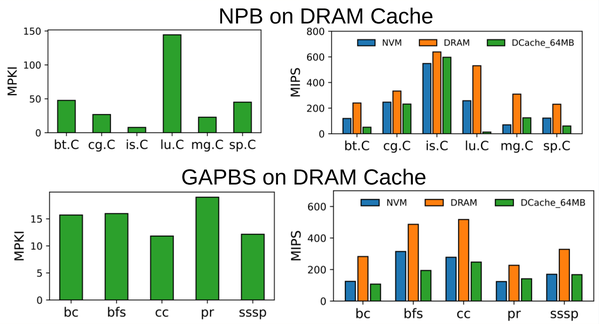
- Left: Peak BW of the DRAM cache VS gem5's default memory controller (DMC).
- Right: Access amplification of our model VS the real hardware [1].
  - ➢ Write_Misses Case: in real hardware write-fills and data-writes are not merged, where in our model they are.



Performance of HPC Applications

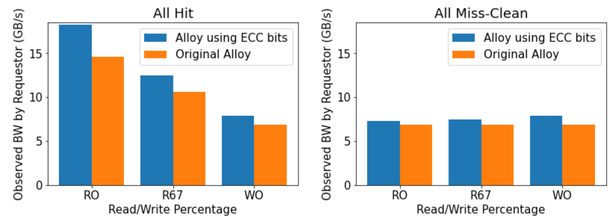
- **NPB** and **GAPBS**
- Million instructions per second (MIPS) and the DRAM cache misses per thousand instructions (MPKI) values are reported.
- UDCC was configured as Cascade Lake.

On DRAM cache, most workloads performed worse than DRAM and NVM main memory, due to high miss rate caused by the rigid cache architecture [1].



Performance of Alloy DRAM Cache

- Decoupled DRAM Cache Policy Manager configured as the architecture of Intel's Cascade Lake, without partial-writes.
- The DRAM cache interfaces: Original Alloy vs Alloy using ECC bits
- Main memory interface: DDR4
- The results show up-to 20% BW degradation for Original Alloy compared to Alloy using ECC bits.



Link Latency Case Study

- Using an HBM cache, backed by a (i) DDR4 and (ii) NVM main memory through a link, for a read-only miss-clean traffic.

On lower link latency, far NVM performs better than far DDR4. For higher link latency, NVM performs closely to the DDR4.

Far Mem.	Link Latency (µs)	Tot. BW (GB/s)	Avg. Resp. Time (µs)
DDR4	**No Link**	**6.31**	**1.484**
	0.2	5.86	1.599
	1	5.61	1.833
	2.5	5.20	2.985
	5	3.04	5.346
NVM	**No Link**	**6.03**	**2.489**
	0.2	6.03	2.487
	1	6.03	2.491
	2.5	4.86	3.269
	5	2.99	5.460

Conclusion

- In this work, we introduced heterogenous memory modeling support in gem5.
- The models we described in this work enable research opportunities for next generation of heterogenous and disaggregated HPC systems.

Reference

[1] M. Hildebrand, J. T. Angeles, J. Lowe-Power, and V. Akella, "A Case Against Hardware Managed DRAM Caches for NVRAM Based Systems," in 2021 ISPASS.