Row Buffer Locality-Aware Hybrid Memory Caching Policies

HanBin Yoon, Justin Meza, Rachata Ausavarungnirun, Rachael Harding, Onur Mutlu (Carnegie Mellon University)

1. Motivation / Background

- DRAM scaling is becoming difficult
- Memories like Phase Change Memory (PCM) offer scalability, but have drawbacks

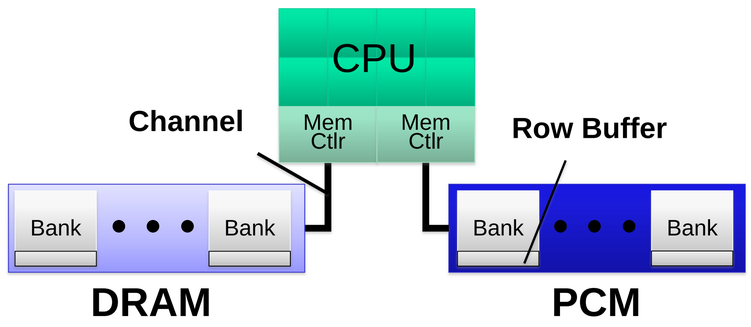
- Use DRAM as a cache to PCM

	PCM	DRAM
Data storage	Resistance	Charge
Scalability	High	Low
Latency (R/W)	~4x/~12x	1x
Energy (R/W)	~2x/~40x	1x
Endurance	10^8 writes	N/A

2. Key Insight

- DRAM and PCM both employ row buffers
- Similar row hit latency, different row miss latencies

- Store data which miss in the row buffer and are reused frequently in DRAM



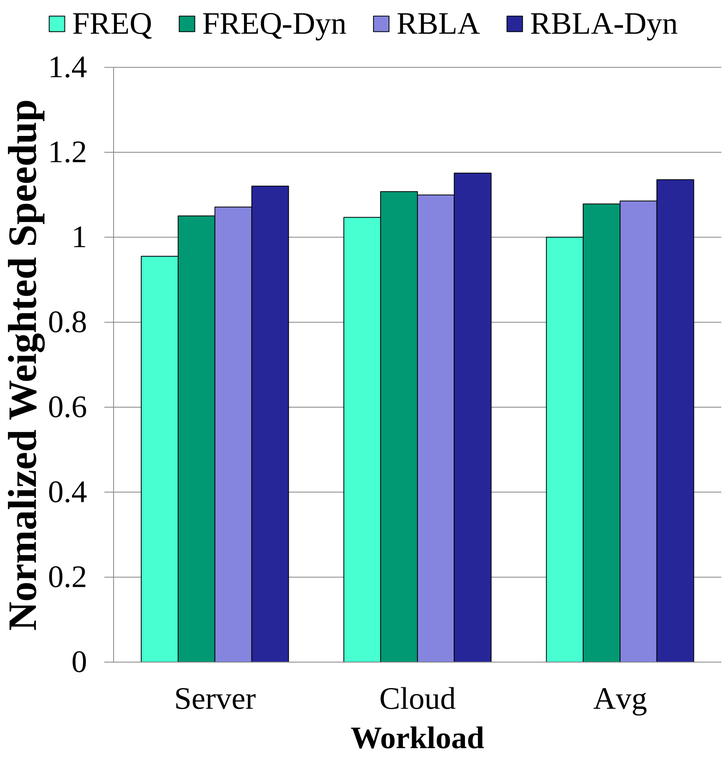
	PCM	DRAM
Row buffer hit	40 ns	40 ns
Row buffer miss	128−368 ns	80 ns

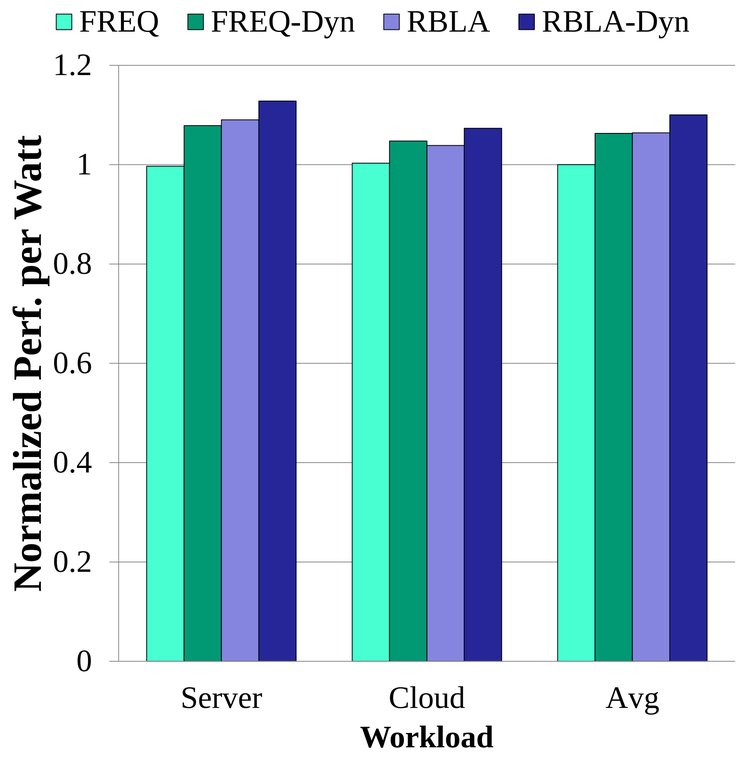
3. Mechanism

- For recently accessed rows in PCM,
  - Track misses to predict future locality
  - Track accesses to predict future reuse
  - Cache data after a threshold number of misses and accesses in an interval
  - Dynamically adjust threshold to adapt to runtime characteristics

4. Evaluation

- 16-core system, 32/512 KB L1/L2 per core
- Separate DRAM and PCM controllers
- 1 GB DRAM, 16 GB PCM (both 8 banks)





Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University

Georgia Tech

intel

PRINCETON UNIVERSITY

UC Berkeley