ROW BUFFER LOCALITY-AWARE CACHING POLICY FOR HYBRID MEMORIES

HanBin Yoon, Justin Meza, Rachata Ausavarungnirun, Rachael Harding, Onur Mutlu (CMU)

MOTIVATION / BACKGROUND

- **DRAM scalability is reaching its limits**
- **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

KEY INSIGHT

- **DRAM and PCM both employ *row buffers***
- ***Same* hit latency, *different* row miss latencies**
- **Store data which *miss in the row buffer* and are *reused frequently* in DRAM**

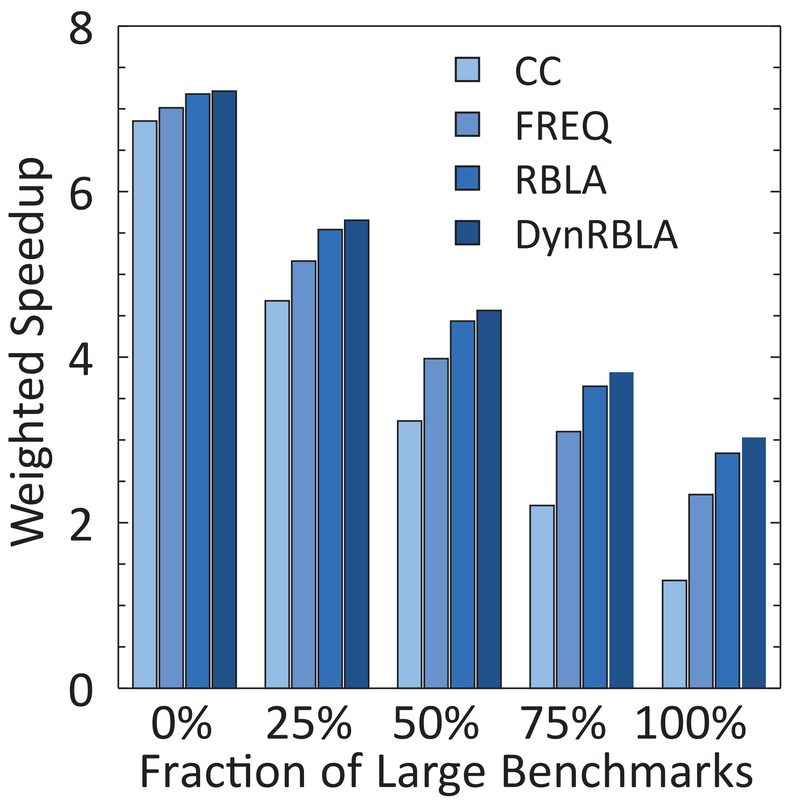
	PCM	DRAM
Row buffer hit	40 ns	40 ns
Row buffer miss	128−368 ns	80 ns

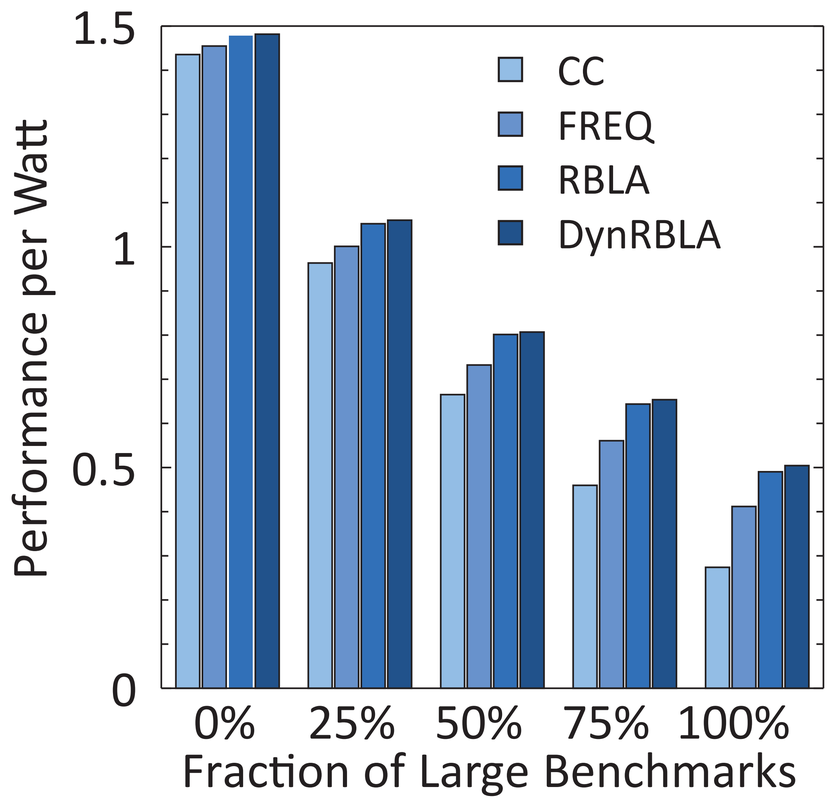
MECHANISM

- **For a *subset* of 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**

EVALUATION

- **16-core system, 32/512 KB L1/L2 per core**
- **16 MB DRAM / 512 MB PCM per core**
- **DDR3-1066, 8/16 banks DRAM/PCM**





Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University · Georgia Tech · intel · Princeton University · UC Berkeley