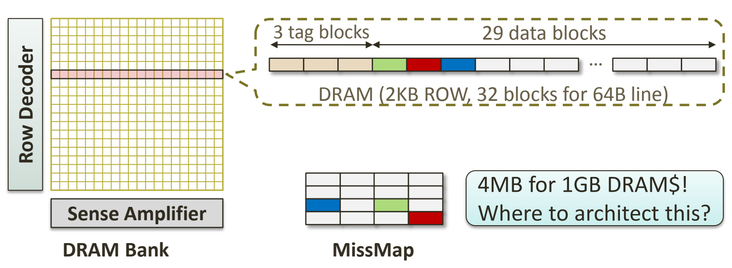
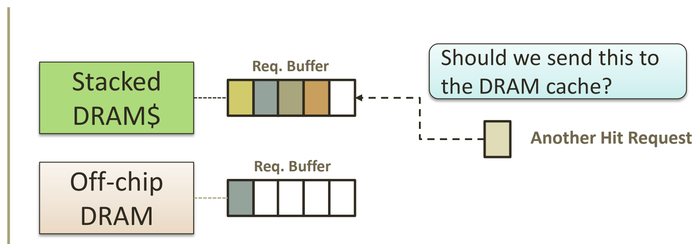
A Mostly-Clean DRAM Cache for Effective Hit Speculation and Self-Balancing Dispatch

comparch Jaewoong Sim Hyesoon Kim AMD Research Gabriel H. Loh Mike O'Connor
PURDUE UNIVERSITY Mithuna Thottethodi

Motivation:

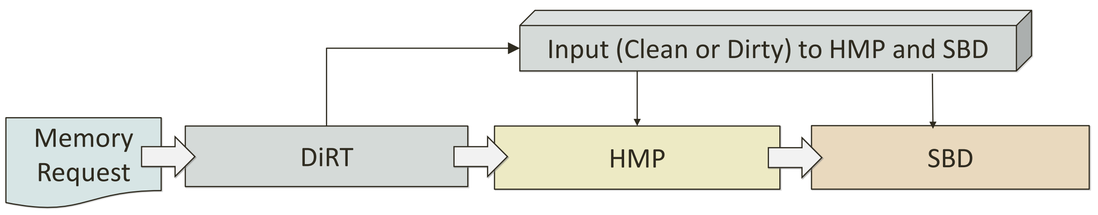
- The cache line tracking structure (MissMap) to avoid a DRAM cache access on a miss is too expensive and less practical
- Applying a conventional cache organization to DRAM caches makes the aggregate system bandwidth under-utilized
- Dirty data in DRAM caches severely restrict the effectiveness of speculative techniques





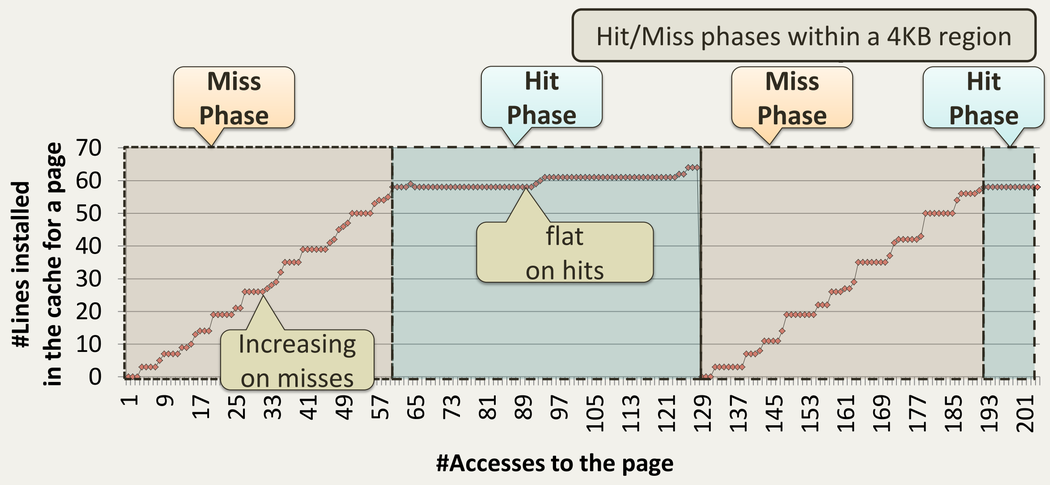
Our Solution: HMP + SBD + DiRT

- Use a low-cost hit-miss predictor to avoid a DRAM cache access on a miss (HMP)
- Steer hit requests to either a DRAM cache or off-chip memory based on the expected latency of both memory sources (SBD)
- Maintain a mostly-clean DRAM cache via region-based WT/WB to guarantee the cleanliness of a memory request (DiRT)



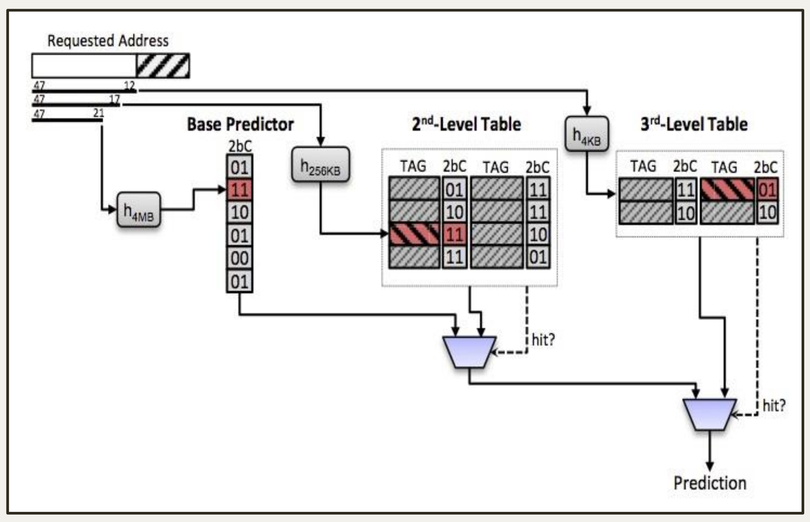
HMP$_{region}$: Region-Based Prediction

- Break up memory space into coarser-grained regions (e.g. 4KB)
- Index into the HMP$_{region}$ with a hash of the region's base address



A simple 2-bit counter will be effective for making hit-miss predictions for the region!

HMP$_{MG}$: Multi-Granular Hit-Miss Predictor



- Base Predictor
  - Prediction for 4MB regions
  - Default Prediction
- 2nd-Level Table
  - Prediction for 256KB regions
  - On Tag Matching
- 3rd-Level Table
  - Prediction for 4KB regions
  - On Tag Matching
- Overrides predictions from larger-granularity predictor tables

HMP$_{MG}$ provides 95%+ prediction accuracy at a less-than-1KB cost!

Self-Balancing Dispatch (SBD)

- Steer hit requests to a DRAM cache or off-chip memory based on the expected latency of both memory sources

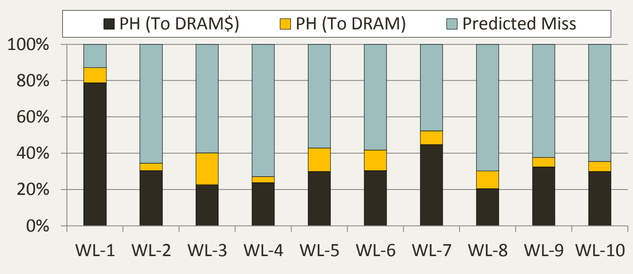
Algorithm: Self-Balancing Dispatch

$N_{Off-Chip}$: # of requests already waiting for the same bank in the off-chip memory
$L_{Off-Chip}$: Typical latency of one off-chip memory request, excluding queuing delays
$E_{Off-Chip}$: $N_{Off-Chip} \times L_{Off-Chip}$ (total expected queuing delay for off-chip)

N_{DRAM_Cache}: # of requests already waiting for the same bank in the DRAM cache
L_{DRAM_Cache}: Typical latency of one DRAM cache request, excluding queuing delays
E_{DRAM_Cache}: $N_{DRAM\_Cache} \times L_{DRAM\_Cache}$ (total expected queuing delay for DRAM cache)

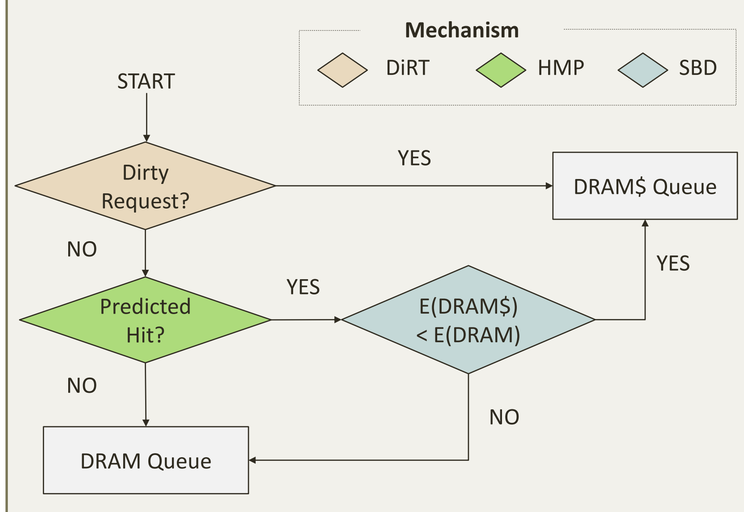
➔ $E_{Off-Chip} < E_{DRAM\_Cache}$: send request to off-chip

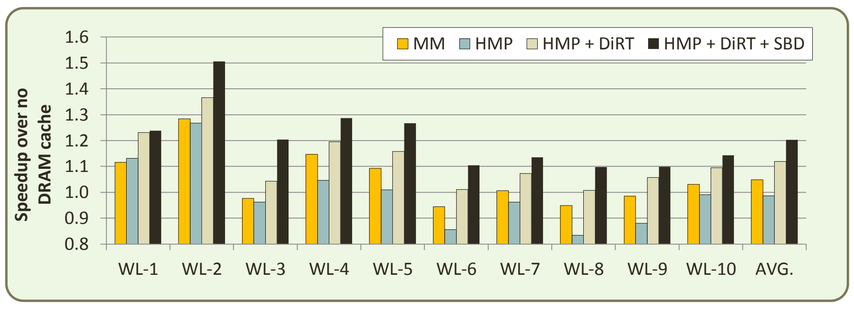
➔ $E_{Off-Chip} \geq E_{DRAM\_Cache}$: send request to DRAM$



Putting It All Together

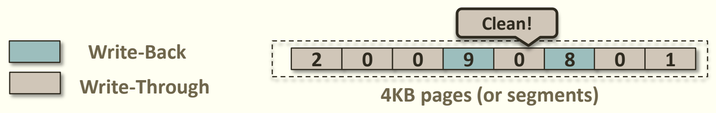
- HMP, SBD, and the DiRT can be accessed in parallel
- Based on the outcomes of the mechanisms, memory requests are sent to either DRAM$ or off-chip DRAM



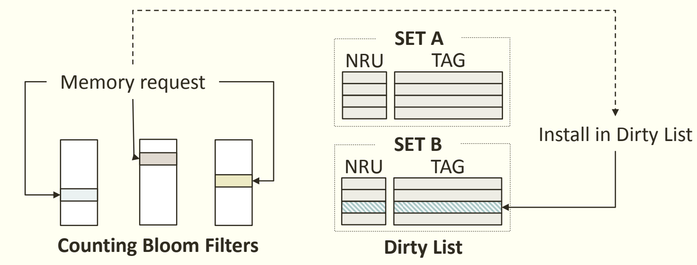


Dirty Region Tracker (DiRT)

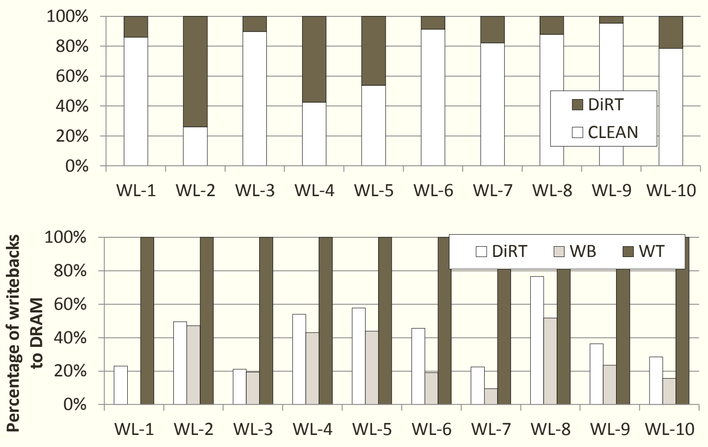
- Region-Based WT/WB policy



- DiRT = Counting Bloom Filters (CBFs) + Dirty List
  - CBFs: Track the number of writes to different pages
  - Dirty List: Record most write-intensive pages



Write-back policy is applied to the pages in the Dirty List



The 45th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-45)