

Caches

Computer Architecture: Design and Simulation
CMU 15-346, Spring 2022

Improving Cache Performance

- **Improving Access Time**
- **Improving Hit Rate**

Access Time

- **Accessing memory**
 - **Average access time**
 - **Hit time**
 - **Miss penalty**
 - **Miss rate**
- **$AAT = HT + MP * MR$**
 - **MP – can be recursive**
 - **$AAT = HT + (HT + MP_2 * MR_2) * MR_1$**

Hit Time

- **Small / simple cache**
- **VIPT**
 - **Access the TLB in parallel**

Miss rate Time

- **Basic miss rate techniques**
 - **Block size increase (to a point)**
 - **Better capture locality, but pollution and transfer time**
 - **Cache size increase (to a point)**
 - **Hold working set, but increase access time**
 - **Associativity increase (to a point)**
 - **8-way is close to full, but increase access time**
 - **Victim cache**
 - **Prefetch**
 - **Replacement**

Victim Cache

- **A little bit of extra associativity will go a long way**
- **Add 4-way cache with one set**
 - **Evicted lines are inserted into this cache**
 - **Check in parallel with L1**

Prefetch

- **Only way to avoid cold misses**
- **Predict which memory accesses will come next**
 - **Speculatively fetch those lines early**
- **Issues:**
 - **Using cache space**
 - **Using memory bandwidth**
 - **Timeliness of prefetch**

Compiler

- **Compilers can insert prefetches**
- **Compilers also control memory layout**
 - **Struct field order**
 - **Struct alignment**
 - **Instruction layout**

LRU, NRU, RRIP

▪ LRU

- Order the lines in a set
- On access, if hit, line is first, increase counter for other lines
- On access, if miss, evict the oldest line

▪ NRU

- 1-bit per block
- On hit, block's bit is zero
- On miss, evict a block with one, if none, reset bits to one and select

▪ RRIP

- K-bits per block
- On hit, zero block's bits
- On miss, evict a block with maximum value, else increment all blocks by one
- On insert, block gets a non-zero value (often $2^{(k)-2}$)

Miss penalty

- **Early restart / critical word first**
- **Sub-blocking**

Sub-blocking

- **Split a block into multiple sub-blocks**
 - **Track whether each sub-block is valid**
 - **Fetch the sub-block that is missed**
- **Advantages**
 - **Reduces tag overhead**
 - **Supports early restart / critical word first**