

Project: Cache Simulator

Overview:

You will write a cache simulator to evaluate aspects of cache architecture and the effect different cache parameters have on performance. Modeling is often the first step in the design process. This project will give you the opportunity to model cache alternatives.

Your simulator will read address traces from different programs. You will use C as the programming language and start with the template and provided materials on the Github repository.

We will assume that this is for an L1 cache closely associated with the CPU core. We will assume that there is an L2 downstream which mitigates extreme penalties for main memory access. This assumption can be changed with the miss penalty multiplier (see below).

Additionally, you will investigate the implications of LRU design in hardware for set associative caches.

If working with a partner(s), all must complete the entire project together. Each person must fully understand the code and the results.

Grading:

Simulator (50 points)

Questions (40 points). This assumes you have data.

LRU policy (10 points)

The address traces:

The address traces are representative of load/store instruction address targets. The traces were generated by a simulator of a RISC processor from the SPEC benchmarks. You do not need to know the data associated with the instruction. You do need to know read/write. You will also be given a value indicating the number of instructions executed between previous memory access and the current (including the load or store instruction itself).

Because these are only load/store instructions, you are simulating a data cache; however, the techniques and your C code can be used, obviously for an instruction cache. For an instruction only cache, you would treat all addresses as reads, for example.

Format of the address traces:

```
# LS ADDRESS IC
```

Where LS is a 0 for a load (read) and 1 for a store (write), ADDRESS is an 8-character hexadecimal number, and IC is the number of instructions executed between the previous memory access and this one (including the load or store instruction itself). There is a single space between each field. The instruction count information will be used to calculate execution time (or at least cycle count). A sample address trace starts out like this:

```
# 0 7ffffed80 1      Note: Read with 0 instructions prior to this + one.  
# 0 10010000 10     Note: Read with 9 instructions prior to this + one.  
# 0 10010060 3  
# 1 10010030 4      Note: Write  
# 0 10010004 6  
# 0 10010064 3  
# 1 10010034 4
```

The prior trace shows a total of 31 instructions with 5 reads and 2 writes to data.

WARNING: Each trace contains memory accesses of just over 5 million instructions. So, probably can't edit them 😊

Simulator “Knobs”

Your simulator should support:

- 1) Block size: This is the row size in bytes for data stored in the cache. A 16-byte block size has a cache with 16 bytes of data per row in the cache.
 - a. Blocks are multiples of 2. Min 16 to max 128
- 2) Associativity: This is the number of “bedrooms in the house”
 - a. Sets are multiples of 2. Min = 1 (Direct mapped), Max = 8
 - b. You will use an LRU replacement policy for 2/4/8 way (See additional at end of assignment)
- 3) Allocate policy
 - a. We will assume write allocate and write back. Reads always allocate.
 - b. A write allocate cycle is a read miss penalty to fill the cache line (same as the read allocate).
- 4) Cache size for the data storage bytes
 - a. Multiple of 2: Min of 16KB, max of 128KB
 - b. This is total cache size; therefore, watch number of rows based on associativity
- 5) Miss penalty
 - a. This is the penalty in clock cycles for a cache miss causing a new cache line fetch to the cache. This will be an integer value equal to the clock rate in GHz x 15 for the base value and add to this the additional due to block size (see below). For

example, clock rate 2GHz, MP = $2*15=30$. Clock rate 4GHz, MP = $4*15 = 60$. This is low if access is to main memory, but we will assume that there is an L2 that mitigates this value.

- b. Dirty line eviction on write back for dirty line is 2 clock penalty (see below)

Architecture assumptions

Nothing comes for free. So, larger sizes take longer to respond.

Assume a base cycle time (clock rate) of 2Ghz. You can make this a parameter in your program. This will be used in calculating AMAT and elapsed time.

All instructions take one cycle. All hits take one cycle. Loads add miss penalties. The miss penalty for a hit is zero.

All misses result in a stall for miss-penalty cycles.

A writeback of a dirty line is mostly hidden by downstream buffering and/or caches. So use an extra 2-cycle delay to write a dirty line to a write buffer. A new load evicting a clean line takes the miss delay (30, for example). A load with a dirty line eviction takes the miss delay + 2 (32, for example).

Cache size implications. Cache sizes slow down the cycle time!

- 1) Starting with 16KB as base, add 5% cycle time for access for 32KB, 10% for 64KB, and 15% for 128KB.
 - a. Note, this does NOT affect number of cycles. It affects the total time!
- 2) Associativity also takes longer. Add 5% cycle time for 2 way, 7.5% for 4 way, 10% for 8 way.
 - a. Note, this does NOT affect number of cycles. It affects the total time!
- 3) Block size clock cycle adder to miss penalty. Larger blocks take longer to load.
 - a. 16 is base. 32 bytes is +2, 64 bytes is +6, 128 bytes is +12
 - b. For example the 30 cycle mp in the default case goes to 32 for 64 bytes and 36 for 128 bytes.
- 4) Add penalties. 32KB with 4-Way is 12.5% cycle time adder.

Defaults to assume:

- 1) No cache. Run the address trace assuming no cache. **All loads** take the miss penalty. All stores are 2 clocks (no miss penalty). We assume that there is still a downstream buffer. And, we have no cache to load lines after eviction
- 2) **Base cache (default):** 16 bytes block, 16KB size, direct, miss penalty 30.

Results

Upload your code and results. Keep your work effort in a github repository on the class organization.

Your pdf document should include:

Make sure code is commented.

Flow diagram of code

Your results (details, spreadsheets, graphs). Provide a header with the results output for each run.

Your answers to the questions (based on your data)

Results in a spreadsheet for each of the three traces (art, mcf, and swim) for each of:

- 1) No cache. There are no hits. All read accesses use the mp. Can still use write buffer delay for writes.
- 2) Default cache
- 3) Cache size: 16KB, 32KB, 64KB, 128KB (default with change in cache size)

Choose best two results for each trace based on cycle time and continue

- 4) Associativity: 1, 2, 4, 8 (Using best two from prior). You already ran direct map (1) prior. Run with 2, 4, and 8 using the chosen cache sizes.

Choose best the result for each trace based on cycle time and continue.

- 5) Block size 16, 32, 64, 128 bytes. You already ran block size 16.

You should now have 4 final results for each of the three traces to compare.

Keep track of:

All of the initial parameters described above, plus show results similar to the example results below:

Results Output

- 1) Execution time in cycles and in elapsed time.
- 2) Total instructions
- 3) Memory accesses
- 4) Overall miss rate
- 5) Read/load miss rate
- 6) CPI
- 7) Average memory access time in cycles
 - a. Cycles per memory access, 0 cycles for a cache hit, mp for cache miss, plus write buffer penalty for dirty line evicts).
 - b. Example for – a 4 run below: $((475672+20015)*30 + 60015*2)/1957764 = 7.66$ cycles
- 8) Dirty evictions
- 9) Load misses
- 10) Store misses
- 11) Load hits
- 12) Store Hits

Using your results:

For each trace address trace (art, mcf, and swim), Plot graphs of

- 1) Graph Miss rate vs block size vs associativity and vs cache size
 - a. Y axis is miss rate
 - b. X axis is cache size
 - c. Colors for associativity
 - d. Line types for Block size
 - e. Label clearly your graph
- 2) Graph execution time vs block size vs associativity vs cache size prior b-e and a as time.
- 3) Do your results for Total CPI match miss rate results? In other words do you see better overall Total CPI with better overall miss rate? Discuss.
- 4) Does lowest miss rate correlate with overall best execution time? Is it a good indicator of performance? Discuss
- 5) In what case(s) did the option with lowest miss rate not have the lowest execution time? Why?
- 6) Were results uniform across the three programs? In what cases did different programs give different conclusions? Speculate as to why this may be true
- 7) Choose what you believe to be the best case design for the cache assuming no limits. Justify choice. What is the difference compared to two default cases of no cache and

default cache (CPI, execution time, and AMAT). What is the final clock rate? Why did you choose this design?

- 8) Now choose best design for fastest overall clock rate. The marketing folks want to advertise the highest clock speed. Now, what is your choice? What is the overall CPI, AMAT, and execution time and final clock rate. Justify choice. Compare to #7. What is your fastest overall clock rate?

LRU Design (10 points)

It is fairly straightforward to implement an LRU policy in a high-level language like “C”. However, hardware implementation can be quite a challenge.

Describe how to implement the LRU replacement policy hardware for your cache. Design for a 2-, 4-, and 8-way set associative cache.

For each, describe your solution (in words) and provide Verilog for it as well (if you'd like). You should assume a static design (clocked). Since your design is for the main L1 cache, assume the LRU decision must be completed in 1 clock cycle.

Estimate the complexity in delay stages and 2-input equivalents as a way to compare 2 vs 4 vs 8-way LRU policy in hardware. Don't forget the data selection from the ways.

NOTE: The traces and aspects of the lab come from Oberlin.edu Cynthia Taylor site.

Think about how to intelligently debug and test your program. Running immediately on the entire input gives you little insight on whether it is working (unless it is way off). To do this create separate memory tests (you can see the text format above) to ensure cache size, cache associativity, blocksize, and miss penalty are functioning correctly. You do not need to turn them in, but they will help tremendously.

Speed matters. These simulations should take a couple minutes (actually, much less) on an unloaded machine. If it is taking much more than that, do yourself a favor and think about what you are doing inefficiently.

Simulations are not the same as hardware. If your tag only takes 16 bits, feel free to use an integer for that value. Other time-saving optimizations along these lines might be useful.

Quick check on code

Assuming a miss penalty of 30 cycles and base sizes, the first instructions from the trace above of 31 instructions will take 151 cycles with 4 cache misses and 3 cache hits for 5 loads and 2 stores with 30 cycle miss penalty.

Correctness check (I have not checked these, but I assume correct). Results for the sample trace above:

```
$ java Cache -s 32 -a 4 -b 32 -mp 30 traces/test.trace
Cache parameters:
  Cache Size (kB)          32
  Cache Associativity      4
  Cache Block Size (bytes) 32
  Miss penalty (cycles)    30

Simulation results:
  execution time            151 cycles
  instructions               31
  memory accesses            7
  overall miss rate          0.57
  load miss rate             0.60
  CPI                        4.87
  average memory access time 17.14 cycles
  dirty evictions            0
  load_misses                3
  store_misses                1
  load_hits                  2
  store_hits                  1
```

Additional results to check your algorithms.

```

gunzip -c art.trace.gz | java cache -a 1 -s 16 -l 16 -mp 30
    Cache parameters:
    Cache Size (KB)           16
    Cache Associativity       1
    Cache Block Size (bytes) 16
    Miss penalty (cyc)        30

    Simulation results:
    execution time 21857966 cycles
    instructions 5136716
    memory accesses 1957764
    overall miss rate 0.28
    read miss rate 0.30
    memory CPI 3.26
    total CPI 4.26
    average memory access time 8.54 cycles
    dirty evictions 60540
    load_misses 523277
    store_misses 30062
    load_hits 1208606
    store_hits 195819

```

```

$ java Cache -s 32 -a 4 -b 32 -mp 30 traces/art.trace.gz
Cache parameters:
    Cache Size (kB)           32
    Cache Associativity       4
    Cache Block Size (bytes) 32
    Miss penalty (cycles)    30

```

```

Simulation results:
    execution time              20127356 cycles
    instructions                 5136716
    memory accesses              1957764
    overall miss rate            0.25
    load miss rate               0.27
    CPI                          3.92
    average memory access time   7.66 cycles
    dirty evictions              60015
    load_misses                  475672
    store_misses                 20015
    load_hits                    1256211
    store_hits                   205866

```

```

> gunzip -c mcf.trace.gz | java cache -a 8 -s 64 -l 32 -mp 42
    Cache parameters:
    Cache Size (KB)           64
    Cache Associativity       8

```

Cache Block Size (bytes)	32
Miss penalty (cyc)	42

Simulation results:
execution time 143963250 cycles
instructions 19999998
memory accesses 6943857
overall miss rate 0.42
read miss rate 0.36
memory CPI 6.20
total CPI 7.20
average memory access time 17.85 cycles
dirty evictions 995694
load_misses 2036666
store_misses 867426
load_hits 3552806
store_hits 486959

LRU Design (10 points)

It is fairly straightforward to implement an LRU policy in a high-level language like “C”. However, hardware implementation can be quite a challenge.

Describe how to implement the LRU replacement policy hardware for your cache. Design for a 2-, 4-, and 8-way set associative cache.

For each, describe your solution (in words) and provide Verilog for it as well (if you'd like). You should assume a static design (clocked). Since your design is for the main L1 cache, assume the LRU decision must be completed in 1 clock cycle.

Estimate the complexity in delay stages and 2-input equivalents as a way to compare 2 vs 4 vs 8-way LRU policy in hardware. Don't forget the data selection from the ways.