CS203 Winter ’16 - Project 3

Due: Thursday, February 25 midnight.
All projects will be turned in to iLearn.

1. Objective

The goal of project 3 is to explore memory hierarchy design choices, with a focus on L1 cache design. Try to keep and organize your result files, as you may need to reuse results for different questions. This will help avoid unnecessary simulation runs.

2. Exploring Cache Design

The configuration for L1 data cache can be found in `simu.conf.apache` as shown below. Configurations of interest that are covered in class are highlighted in blue.

```
[DL1_core]
deviceType        = 'cache'
coreCoupledFreq   = true
inclusive         = true
directory         = false
blockName         = "dcache"
numBanks          = 1                       # Number of banks
sendFillPortOccp  = 1
sendFillNumPorts  = 1
maxRequests       = 8                       # Max number of pending requests
size              = 64*1024                 # Cache size (bytes)
assoc             = 4                       # Associativity (n-Ways)
skew              = false
bsize             = 64                      # Cache block size (bytes)
replPolicy        = 'LRU'                   # Replacement policy
bkNumPorts        = 1
bkPortOccp        = 0
hitDelay          = 4                       # Hit delay
missDelay         = 4                       # Miss delay
lowerLevel        = "PrivL2 L2 sharedby 2"
bankShift         = 4
fillBuffSize      = 4
pfetchBuffSize    = 16
wbBuffSize        = 16
```

Currently, the caches are fairly large for the given workloads. In order to observe interesting cache behaviors, let’s first decrease the cache size to 8KB.

```
size              = 8*1024                  # Cache size (bytes)
```

a) For this cache configuration, how many blocks and sets does the cache contain?

b) Run this configuration for our four workloads. What is the L1 average memory access time and cache miss rate for each workload?

The L1 data cache average memory access time and miss rate is presented in the result report and is given by the following highlighted in blue:

```
*****************************************************************************************************
# File  : esesc_noname.M4fWnD  : Mon Jan 11 02:19:05 2016
*****************************************************************************************************
Sampler 0 (Procs  0 1)
         Rabbit     Warmup    Detail    Timing    Total  KIPS
  KIPS   44359       1760       757       765      3794
  Time     5.0%      83.3%      1.6%     10.1%             : Sim Time (s) 131.789 Exe   4.739 ms Sim (1700MHz)
  Inst    59.0%      38.7%      0.3%      2.0%            : Approx Total Time 288.533 ms Sim (1700MHz)
*****************************************************************************************************
Proc : Avg.Time :  BPType           :  Total  :          RAS        :  BPred  :         BTB         :  BTAC   : WasteRatio
   0 :   24.993 :  2bit             :  61.11% : (100.00% of 9.58%) :  68.11% : ( 54.37% of 29.30%) :  33.74% :  38.41%
-----------------------------------------------------------------------------------------------------------------------------
Proc :  nCommit   :   nInst   :  AALU   :  BALU   :  CALU   :  LALU   :  SALU   :  LD Fwd :    Replay    : Worst Unit  (clk)
   0 :  10139922  :  10139922 :  48.38% :  18.22% :   0.05% :  25.76% :   7.60% :  0.00%  :      N/A     : MUNIT_AALU  0.22
-----------------------------------------------------------------------------------------------------------------------------
--------
Proc  IPC     uIPC    Active        Cycles    Busy   LDQ   STQ  IWin   ROB  Regs    IO  maxBr  MisBr Br4Clk brDelay
   0   1.26   1.26   1.00         8056756     62.9   0.2   1.8   1.8   0.1   0.7   0.0   0.0   0.0    0.0   13.1
*****************************************************************************************************
Cache           Occ  AvgMemLat MemAccesses  MissRate  (  RD ,    WR,    BUS)
IL1(0)          0.0  2.2          3608033       0.18%  ( 99.8%,  0.0%,  0.0%)
-----------------------------------------------------------------------------------------------------
DL1(0)          0.0  4.9          3347212       0.23%    ( 99.8%, 97.2%,  0.0%)
-----------------------------------------------------------------------------------------------------
ITLB(0)         0.0  2.2       3607979       0.00%    (100.0%,  0.0%,  0.0%)
L2(0)           0.0  43.7      14186        24.61%    ( 75.4%,  0.0%,  0.0%)
Memory(0)       0.0  0.0       0             0.00%    (100.0%,  0.0%,  0.0%)
PTLB(0)         0.0  4.9       3339493       0.16%    ( 99.8%,  0.0%,  0.0%)
*****************************************************************************************************
```

c) In simple 5-stage in-order pipelines, the miss penalty is typically equivalent to the miss latency, due to the pipeline stalling. In more complex architectures, such as out-of-order, speculative and multithreaded processors, the miss penalty is difficult to calculate since instructions can still execute during a miss. For the current cache configuration (8KB, 4-way), what is the miss penalty for each of the four workloads? Do different workloads exhibit different miss penalties? Why or why not?

d) In the course textbook, there are two general rules of thumb regarding cache design. The first is that an 8-way set associative cache for all practical purposes is as effective in reducing misses as a fully associative cache. Do you observe if that is the case? Support your answer with experimental results.

e) The second rule of thumb is called the 2:1 cache rule of thumb. This rule of thumb states that a direct-mapped cache of size N has about the same miss rate as a 2-way set associative cache of size N/2. Assuming N = 8KB, does this rule of thumb hold in your experiments? Again, support your answer with experimental results.

f) Measuring cold, capacity, and conflict misses. (See “Measuring/Classifying Misses” in slides). In this problem, we will identify the types of misses for an 8KB, 4-way set associative cache (our initial cache configuration). For each benchmark, provide a breakdown of the type of cache misses. You can provide the breakdown in terms of the miss rate. For example, if the 8KB, 4-way cache has a 20% miss rate, an infinite size cache have a 1% miss rate, and a fully associative cache have a 10% miss rate, then 1% is due to cold misses, 9% is due to capacity misses, and 10% is due to conflict misses.