ELE 475 PS#4 Solution

Problem # 1:

a)

Direct mapped: 0.86ns.
2-Way: 1.12ns.
4-Way: 1.37ns.

This makes the 2-way set associative cache 30% slower than the direct mapped cache, and the 4-way cache ~59% slower.

b)

16K: 1.27ns.
32K: 1.35ns.
64K: 1.37ns.

This makes the 32K cache 6% slower than the 16 cache, and the 64K cache ~8% slower.

c)

Average Access Time = hit% * hit_time + miss% * miss_time

Miss% = misses per instruction / reference per instruction

DM = 2.2%
2-way = 1.2%
4-way = 0.33%
8-way = 0.09%

Cache access time;

DM = 0.86ns @ 0.5ns cycle time = 2 cycles
2-way = 1.12ns @ 0.5ns cycle time = 3 cycles
4-way = 1.37 @ 0.83ns cycle time = 2 cycles
8-way = 0.09% @ 0.79 cycles = 3 cycles

Miss penalty:

DM = 20 cycles
2-way = 20 cycles
4-way = 13 cycles
8-way = 13 cycles

Average Access Time:

DM: ((1 – 0.022) * 2) + (0.022 * 20) = 2.396 cycles @ 0.5ns/cycle = 1.2ns.
2-way: ((1 – 0.012) * 3) + (0.012 * 20) = 3.2 cycles @ 0.5ns/cycle = 1.6ns.
4-way: ((1 – 0.0033) * 2) + (0.0033 * 13) = 2.036 cycles @ 0.83ns/cycle = 1.69ns.
8-way: ((1 – 0.0009) * 2) + (0.0009 * 13) = 3 cycles @ 0.79ns/cycle = 2.37ns.

Direct mapped is the best cache.

ELE 475 PS#4 Solution
Problem # 2:

a)
With critical word first, the miss time would be 120 cycles. With early restart, depending on where the needed word is in the block, it would take 120 cycles if the needed word is in the first block, 120 + (16 * 1) if the needed data is in the second block, 120 + (16 * 2) if the needed data is in the third block, and 120 + (16 * 3) if the needed data is in the fourth block. Without critical word first/early restart, the cache processor needs to wait for all of the data to return, or 168 cycles. We still assume that the data can bypass waiting to be written to the cache as is stated in the problem.

b)
It is likely that the L1 cache contributes to the average memory access time more, therefore critical word first and early restart is more likely to matter for L1 caches.

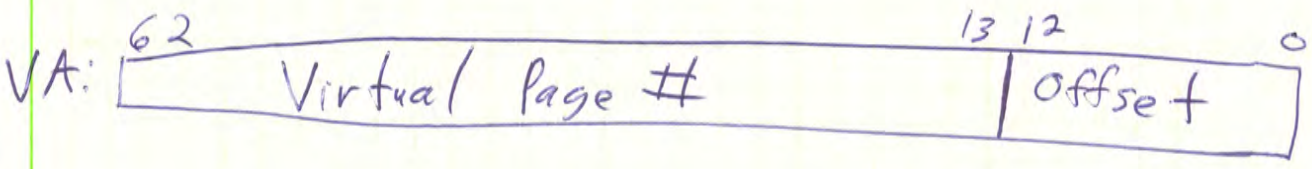
ELE 475 PS#4 Solution

Problem # 3:

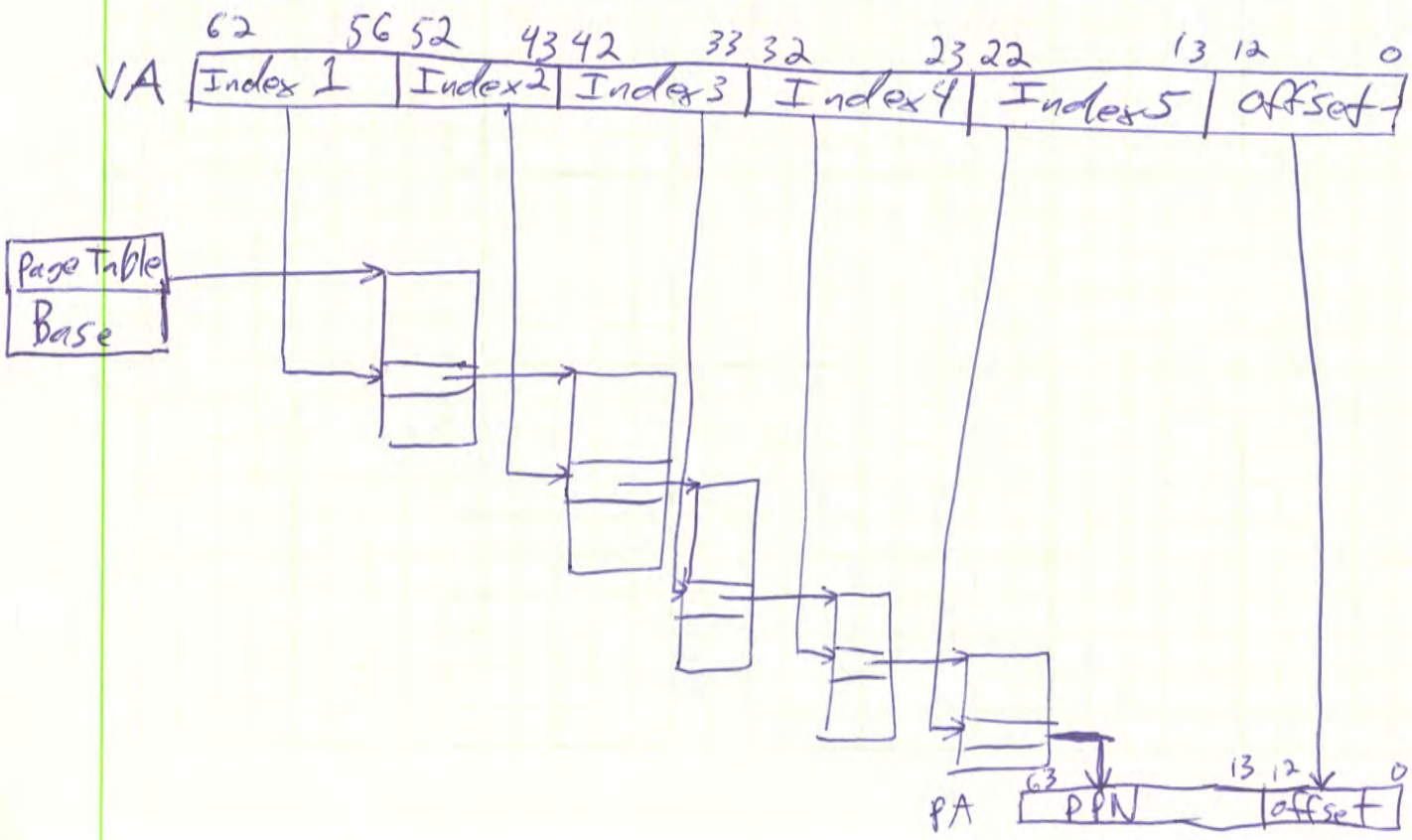
16 entries * 1MB/entry = 16MB reach maximum.

ECE 475 PS#4 David

Problem 4)



Each Page Table Entry (PTE) needs 51 bits of physical address and a valid bit. the lowest level entry (leaf) needs a dirty bit, kernel bit, and two reserved bits also. All of these fit nicely in 64 bits. 1024 PTE's fit within an 8KB page. Therefore, we look at the virtual address to see how many levels are needed. this needs 5-levels of page tables



Problem 10 (cont)

Need 5 pages just to build 5 levels of page tables. To increase reach, we have different pages pointed to by leaf page table which is 5 pages.
5 pages × 8 KB = 40 KB.

The minimum would be to use 0 or 1 page depending how literal the problem is taken. All 10 pages can be used in the different levels to point to one 8 KB page or no pages if all leafs are marked invalid.

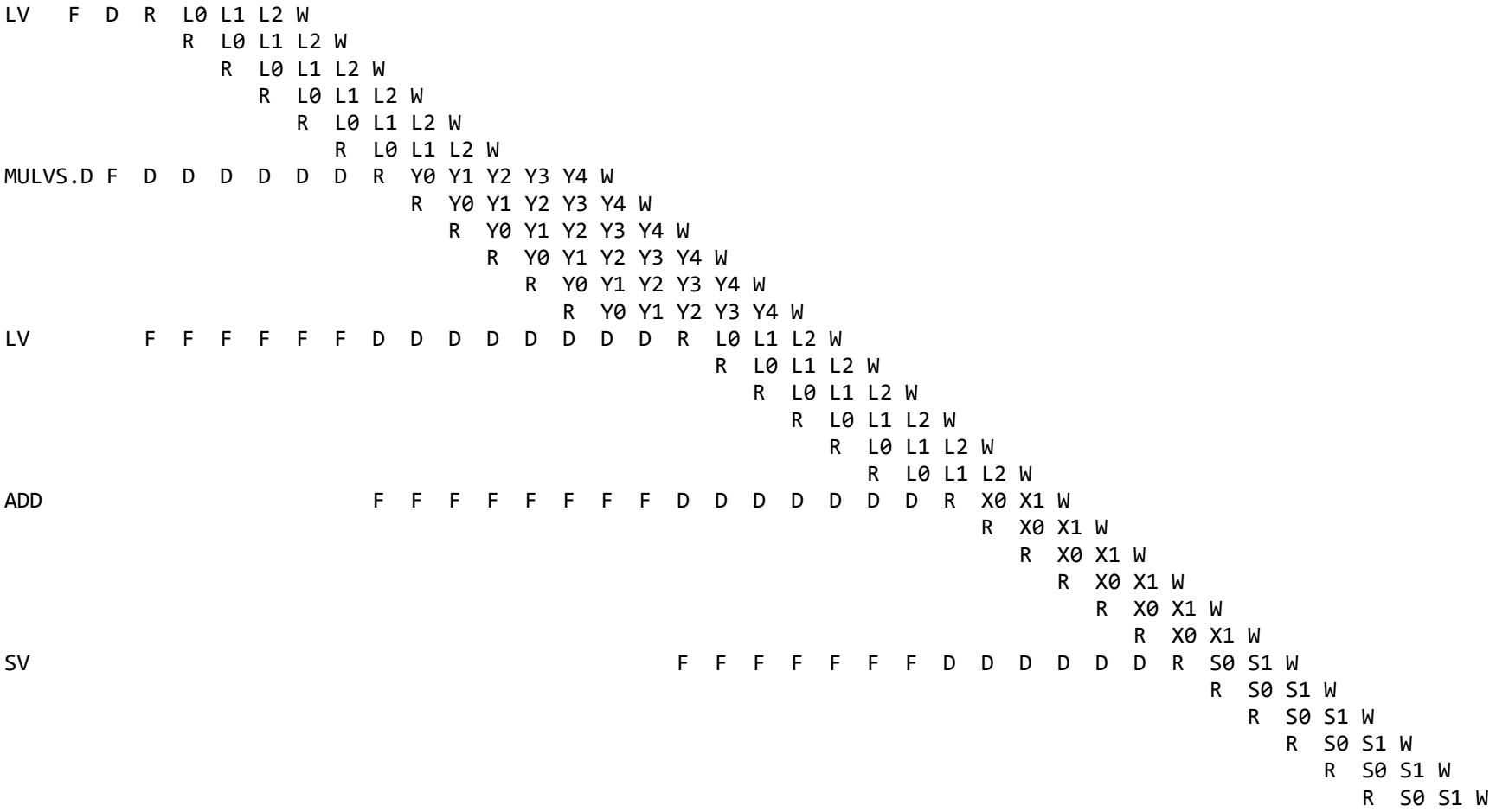
ELE 475 PS#4 Solution
Problem # 5:

On a machine with a software managed TLB, a TLB miss can occur because the either the page is not mapped into the page table or the page is simply not loaded into the TLB but is in the page table. A TLB miss does not always occur in a bus-error/segmentation fault with a software managed TLB because the page may simply need to be copied from the page table to the TLB.

On a machine with a hardware page-table walker, a miss in the page table may or may not cause a bus-error/segmentation fault. The OS definitely gets involved, but many times, the OS simply has not mapped in a page yet, as is the case for the heap and stack, or the page may be on disk. If none of these cases are true, then the process that causes the page fault is likely to receive a bus-error/segmentation fault.

ELE 475 PS#4 Solution
Problem # 6:

We only draw the vector code pipeline because it is common for the control processor to be decoupled in a different pipeline in a real vector machine.



ELE 475 PS#4 Solution
Problem # 7:

We only draw the vector code pipeline because it is common for the control processor to be decoupled in a different pipeline in a real vector machine.

```
LV    F  D  R  L0 L1 L2 W
               R  L0 L1 L2 W
                  R  L0 L1 L2 W
            R  L0 L1 L2 W
               R  L0 L1 L2 W
                  R  L0 L1 L2 W
MULVS.D F  D  D  D  D  D  R  Y0 Y1 Y2 Y3 Y4 W
                           R  Y0 Y1 Y2 Y3 Y4 W
                              R  Y0 Y1 Y2 Y3 Y4 W
                        R  Y0 Y1 Y2 Y3 Y4 W
                           R  Y0 Y1 Y2 Y3 Y4 W
                              R  Y0 Y1 Y2 Y3 Y4 W
LV          F  F  F  F  F  D  R  L0 L1 L2 W
                              R  L0 L1 L2 W
                                 R  L0 L1 L2 W
                              R  L0 L1 L2 W
                                 R  L0 L1 L2 W
                                    R  L0 L1 L2 W
ADD                        F  D  D  D  D  D  R  X0 X1 W
                                             R  X0 X1 W
                                                R  X0 X1 W
                                          R  X0 X1 W
                                             R  X0 X1 W
                                                R  X0 X1 W
SV                            F  F  F  F  F  D  D  D  D  R  S0 S1 W
                                                         R  S0 S1 W
                                                            R  S0 S1 W
                                                      R  S0 S1 W
                                                         R  S0 S1 W
                                                            R  S0 S1 W
```

ELE 475 PS#4 Solution
Problem # 8:

GPUs do not have vector length registers. Instead, they pack a vector into SIMD style operations. In order to handle the case where the vector length is not an integral size of the SIMD width, they use predication to disable operations. This is in effect how GPUs strip-mine a loop.

On a GPU, when two elements of data in a vector need different processing, operations are masked using a vector mask. This is a form of predication. A masked operation utilizes resources independent on whether the operation is enabled or disabled.