2.9  a. The average memory access time of the current (4-way 64KB) cache is 1.69ns. 
64KB direct mapped cache access time = .86ns @ .5 ns cycle time = 2 cycles 
Way-predicted cache has cycle time and access time similar to direct mapped 
cache and miss rate similar to 4-way cache. 
The AMAT of the way-predicted cache has three components: miss, hit with 
way prediction correct, and hit with way prediction mispredict: 
\[0.0033 \times (20) + (0.80 \times 2 + (1 - 0.80) \times 3) \times (1 - 0.0033) = 2.26 \text{ cycles} = 1.13 \text{ ns}\]

b. The cycle time of the 64KB 4-way cache is 0.83ns, while the 64KB direct-
mapped cache can be accessed in 0.5ns. This provides 0.83/0.5 = 1.66 or 66% 
 faster cache access.

c. With 1 cycle way misprediction penalty, AMAT is 1.13ns (as per part a), but 
with a 15 cycle misprediction penalty, the AMAT becomes: 
\[0.0033 \times 20 + (0.80 \times 2 + (1 - 0.80) \times 15) \times (1 - 0.0033) = 4.65 \text{ cycles} or 2.3 \text{ ns}\]

d. The serial access is 2.4ns/1.59ns = 1.509 or 51% slower.

Problem 2.9 needs results from 2.8. Solutions for 2.8 is given below:

2.8  a. The access time of the direct-mapped cache is 0.86ns, while the 2-way and 
4-way are 1.12ns and 1.37ns respectively. This makes the relative access 
times 1.12/0.86 = 1.30 or 30% more for the 2-way and 1.37/0.86 = 1.59 or 
59% more for the 4-way.

b. The access time of the 16KB cache is 1.27ns, while the 32KB and 64KB are 
1.35ns and 1.37ns respectively. This makes the relative access times 1.35/ 
1.27 = 1.06 or 6% larger for the 32KB and 1.37/1.27 = 1.078 or 8% larger for 
the 64KB.

c. Avg. access time = hit% \times \text{hit time} + miss% \times \text{miss penalty}, miss% = misses 
per instruction/references per instruction = 2.2% (DM), 1.2% (2-way), 0.33% 
(4-way), .09% (8-way).

Direct mapped access time = .86ns @ .5ns cycle time = 2 cycles 
2-way set associative = 1.12ns @ .5ns cycle time = 3 cycles
4-way set associative = 1.37ns @ .83ns cycle time = 2 cycles
8-way set associative = 2.03ns @ .79ns cycle time = 3 cycles
Miss penalty = (10/.5) = 20 cycles for DM and 2-way; 10/.83 = 13 cycles for
4-way; 10/.79 = 13 cycles for 8-way.

Direct mapped – (1 – .022) × 2 + .022 × (20) = 2.396 cycles => 2.396 × .5 = 1.2ns
2-way – (1 – .012) × 3 + .012 × (20) = 3.2 cycles => 3.2 × .5 = 1.66ns
4-way – (1 – .0033) × 2 + .0033 × (13) = 2.036 cycles => 2.06 × .83 = 1.69ns
8-way – (1 – .0009) × 3 + .0009 × 13 = 3 cycles => 3 × .79 = 2.37ns

Direct mapped cache is the best.

2.11  a. With critical word first, the miss service would require 120 cycles. Without
critical word first, it would require 120 cycles for the first 16B and 16 cycles
for each of the next 3 16B blocks, or 120 + (3 × 16) = 168 cycles.

   b. It depends on the contribution to Average Memory Access Time (AMAT) of
      the level-1 and level-2 cache misses and the percent reduction in miss service
times provided by critical word first and early restart. If the percentage reduc-
tion in miss service time provided by critical word first and early restart is
roughly the same for both level-1 and level-2 miss service, then if level-1
misses contribute more to AMAT, critical word first would likely be more
important for level-1 misses.

2.12  a. 16B, to match the level 2 data cache write path.

   b. Assume merging write buffer entries are 16B wide. Since each store can
      write 8B, a merging write buffer entry would fill up in 2 cycles. The level-2
      cache will take 4 cycles to write each entry. A non-merging write buffer
      would take 4 cycles to write the 8B result of each store. This means the
      merging write buffer would be 2 times faster.

   c. With blocking caches, the presence of misses effectively freezes progress
      made by the machine, so whether there are misses or not doesn’t change the
      required number of write buffer entries. With non-blocking caches, writes can
      be processed from the write buffer during misses, which may mean fewer
      entries are needed.