**Question 1 (20 points):** The access time to the main memory that serves data and instructions to a processor is 70 ns. For a given program, 20% of the instructions are loads or stores. The design team is considering two versions for the processor:

- **Design A**, with an 8 KB unified L1 cache that has a miss rate of 4.0%.
- **Design B**, with a 16 KB L1 unified cache that has a hit time of 1.07 ns.

In both designs, the hit time of the L1 cache determines the clock cycle of the processor. Both designs are equiped with a 4 MB L2 cache that has a miss rate of 75% and a hit time of 12 ns.

1. (**7 points**) The Average Memory Acccess Time (AMAT) for design **A** was measured to be 3.5 $ns$. What is the clock cycle for this design?

   The correct definition of AMAT (according to the textbook) is the average memory access time for each instruction executed in the program. When there is a hit both for the data and the instruction, both accesses happen in parallel and thus the only time spent is the hit time of L1. But the additional miss penalty for the 20% of the instructions that also access data needs to be taken into consideration. We will use T to represent time and R to represent rate.

$$\begin{aligned}
\text{AMAT} &= L1_{\text{hitT}} + (1.0 + \text{DataR}) \times [L1_{\text{missR}} \times (L2_{\text{hitT}} + L2_{\text{missR}} \times \text{Memory}_{\text{accessT}})] \\
L1_{\text{hitT}} &= \text{AMAT} - (1.0 + \text{DataR}) \times [L1_{\text{missR}} \times (L2_{\text{hitT}} + L2_{\text{missR}} \times \text{Memory}_{\text{accessT}})] \\
\text{ClockCycle}(\mathbf{A}) &= L1_{\text{hitT}}(\mathbf{A}) = 3.5\ ns - 1.2 \times [0.04 * (12\ ns + 0.75 * 70\ ns)] \\
&= 3.5\ ns - 3.096\ ns = 0.404\ ns
\end{aligned}$$

   An alternative definition for AMAT (which is not used in the textbook) is the average time required for each memory access. In this case the data access for a load or store instruction is considered a separate access and in the case of the unified cache, we do not need to take the data rate into consideration. With this definition the solution is:

$$\begin{aligned}
\text{AMAT} &= L1_{\text{hitT}} + L1_{\text{missR}} \times (L2_{\text{hitT}} + L2_{\text{missR}} \times \text{Memory}_{\text{accessT}}) \\
L1_{\text{hitT}} &= \text{AMAT} - L1_{\text{missR}} \times (L2_{\text{hitT}} + L2_{\text{missR}} \times \text{Memory}_{\text{accessT}}) \\
\text{ClockCycle}(\mathbf{A}) &= L1_{\text{hitT}}(\mathbf{A}) = 3.5\ ns - 0.04 * (12\ ns + 0.75 * 70\ ns) \\
&= 3.5\ ns - 2.58\ ns = 0.92\ ns
\end{aligned}$$

2. (**7 points**) The AMAT for design **B** was measured to be 3.0 $ns$. What is the miss rate, expressed as a percentage, for the L1 cache in this design?

   Using the correct definition for AMAT:

$$\begin{aligned}
\frac{\text{AMAT} - L1_{\text{hitT}}}{1.0 + \text{DataR}} &= L1_{\text{missR}} \times (L2_{\text{hitT}} + L2_{\text{missR}} \times \text{Memory}_{\text{accessT}}) \\
L1_{\text{missR}} &= \frac{\text{AMAT} - L1_{\text{hitT}}}{(1.0 + \text{DataR}) \times [L2_{\text{hitT}} + L2_{\text{missR}} \times \text{Memory}_{\text{accessT}}]} \\
L1_{\text{missR}}(\mathbf{B}) &= \frac{3.0\ ns - 1.07\ ns}{1.2 \times [12\ ns + 0.75 * 70\ ns]} = \frac{1.93}{77.4} = 0.025 = 2.5\%
\end{aligned}$$

Using the alternative definition for AMAT:

$$\begin{aligned} L1_{\text{missR}} &= \frac{\text{AMAT} - L1_{\text{hitT}}}{L2_{\text{hitT}} + L2_{\text{missR}} \times \text{Memory}_{\text{accessT}}} \\ L1_{\text{missR}}(\mathbf{B}) &= \frac{3.0\ ns - 1.07\ ns}{12\ ns + 0.75 * 70\ ns} = \frac{1.93}{64.5} = 0.03 = 3\% \end{aligned}$$

3. (**6 points**) A new member of the design team says that she can change the L1 of design **A** to reduce the clock cycle without affecting miss rate. As a manager, you are skeptical about her claim because you calculated how much faster the clock cycle would have to be to match design **B**. What should be the percent reduction in the hit time of L1 of design **A** for that design to have the same AMAT as design **B** for this program?

   Using the correct definition for AMAT:

$$\begin{aligned} L1_{\text{hitT}} &= \text{AMAT} - (1.0 + \text{DataR}) \times [L1_{\text{missR}} \times (L2_{\text{hitT}} + L2_{\text{missR}} \times \text{Memory}_{\text{accessT}})] \\ \text{ClockCycle}(\mathbf{newA}) &= L1_{\text{hitT}}(\mathbf{newA}) = 3.0\ ns - 1.2 \times [0.04 * (12\ ns + 0.75 * 70\ ns)] \\ &= 3.0\ ns - 3.096\ ns = -0.096\ ns \end{aligned} \tag{1}$$

   Therefore it is impossible to reduce hit time of design A, without affecting the miss rate, to match the AMAT of design B.

   Using the alternative definition for AMAT:

$$\begin{aligned} \text{ClockCycle}(\mathbf{newA}) &= L1_{\text{hitT}}(\mathbf{newA}) = 3.0\ ns - 0.04 * (12\ ns + 0.75 * 70\ ns) \\ &= 3.0\ ns - 2.58\ ns = 0.42\ ns \\ \text{Pencent reduction} &= \frac{0.92 - 0.42}{0.92} \times 100 = 54\% \end{aligned}$$