Question 1 (30 points): You are working in the design team for a computer that has a memory hierarchy with the following characteristics: It has a base average number of clocks per instruction (CPI) of 1.0 (the base CPI is computed assuming that all the memory accesses hit in the first level of cache). This computer operates at the clock frequency of 4.0 GHz (1 GHz = 10^9 Hertz), and has an L1 instruction cache with an access time of 1 cycle and with a hit rate of 99%. It also has an L1 data cache with a hit rate of 92% and an access time of 1 cycle. The design team has determined that in the most important workload for this machine, 25% of the instructions executed are load or store. An access to the main memory requires 25 *nano*seconds (1 *ns* = 10^{-9} seconds).

a. (**10 points**) What is the effective CPI when the miss rate and the memory access penalty for L1 and D1 are taken into account?

$$
\begin{aligned}
\text{Clock cycle} &= \frac{1}{4 \times 10^9 Hz} = 0.25 \; ns \\
\text{Memory access} &= \frac{25 \; ns}{0.25 \; ns} = 100 \text{ clock cycles} \\
\text{CPI}_{\text{IL1andDL1}} &= \text{CPI}_{\text{base}} + \text{CPI}_{\text{IL1}} + \text{CPI}_{\text{DL1}} \\
\text{CPI}_{\text{IL1}} &= IL1_{\text{miss rate}} \times IL1_{\text{miss penalty}} = 0.01 \times 100 = 1.0 \; \frac{\text{clocks}}{\text{instruction}} \\
\text{CPI}_{\text{DL1}} &= \text{Data}_{\text{rate}} \times DL1_{\text{miss rate}} \times DL1_{\text{miss penalty}} = 0.25 \times 0.08 \times 100 = 2.0 \; \frac{\text{clocks}}{\text{instruction}} \\
\text{CPI}_{\text{IL1andDL1}} &= 1.0 + 1.0 + 2.0 = 4.0 \; \frac{\text{clocks}}{\text{instruction}}
\end{aligned}
$$

b. (**10 points**) How much faster the machine becomes when an unified L2 cache is added to the design? This L2 cache has a 75% local hit rate, and it has an access time of 5 cycles. The access to the L2 **does not** occur simultaneously to the access to the lower levels of the memory.

Both instructions and data will benefit from the addition of L2. Therefore the effective CPI with L2 can be computed as:

$$
\begin{aligned}
\text{CPI}_{\text{IL1,DL1,L2}} &= 1.0 + (0.01 + 0.25 \times 0.08) \times (5 + 0.25 \times 100) \\
&= 1.0 + 0.03 \times 30 = 1.9 \frac{\text{clocks}}{\text{instruction}}
\end{aligned}
$$

With L2 the machine will be $\frac{4.0}{1.9} = 2.1$ times faster.

1

c. (**10 points**) The head of the design team wants the CPI to be reduced to $1.5 \frac{\text{clocks}}{\text{instruction}}$ in order to beat a competing machine that is expected to hit the market at the same time as this design. The design team decides to attempt to achieve this goal by adding an L3 cache to the design — the L2 from part (b) remains in the hierarchy. A preliminary study indicates that the L3 cache will have an access time of 20 cycles and that the access to L3 access **cannot proceed** in parallel with the memory access. What is the minimum L3 hit rate that the team has to achieve in order to reach the CPI goal?

The new goal is:

$$\text{CPI}_{\text{IL1,DL1,L2,L3}} \leq 1.5 \frac{\text{clock}}{\text{instruction}}$$

$$(1)$$

Adding L3, the expression for CPI becomes:

$$
\begin{aligned}
\text{CPI}_{\text{IL1,DL1,L2,L3}} &= 1.0 + (0.01 + 0.25 \times 0.08) \times (5 + 0.25 \times (\text{L3}_{\text{access}} + L3_{\text{miss rate}} \times 100)) \\
\text{CPI}_{\text{IL1,DL1,L2,L3}} &= 1.0 + 0.03 \times (5 + 0.25 \times (\text{L3}_{\text{access}} + L3_{\text{miss rate}} \times 100)) \\
\text{CPI}_{\text{IL1,DL1,L2,L3}} &= 1.0 + 0.15 + 0.0075 \times (\text{L3}_{\text{access}} + L3_{\text{miss rate}} \times 100) \\
\text{CPI}_{\text{IL1,DL1,L2,L3}} &= 1.15 + 0.0075 \times \text{L3}_{\text{access}} + 0.75 \times L3_{\text{miss rate}} \\
\text{CPI}_{\text{IL1,DL1,L2,L3}} &= 1.15 + 0.0075 \times 20 + 0.75 \times L3_{\text{miss rate}} \\
\text{CPI}_{\text{IL1,DL1,L2,L3}} &= 1.15 + 0.15 + 0.75 \times L3_{\text{miss rate}}
\end{aligned}
$$

$$(2)$$

Therefore:

$$
\begin{aligned}
\text{CPI}_{\text{IL1,DL1,L2,L3}} &= 1.3 + 0.75 \times L3_{\text{miss rate}} \leq 1.5 \\
L3_{\text{miss rate}} &\leq \frac{1.5 - 1.3}{0.75} = \frac{0.2}{0.75} = 0.266 \Rightarrow L3_{\text{hit rate}} \geq 0.734
\end{aligned}
$$

Thus the hit rate of L3 must be at least 73.4%.