Homework 1

1. Hennessy and Patterson page 60-61 problem 1.2

   The time spent by the unenhanced computer on work that could be enhanced is 10 times 50\% of the overall enhanced time, since the enhanced mode is 10 times faster and 50\% of the overall enhanced time is spent in enhanced mode. The time spent by the unenhanced computer on the rest of its work is the other 50\% of the overall enhanced time. Therefore,

   \[
   \text{Overall Unenhanced Time} = ((10)(50\%) + (50\%)) \text{Overall Enhanced Time}
   \]

   from which it follows that the answer to part (a) is

   \[
   \text{Speedup} = \frac{\text{Overall Unenhanced Time}}{\text{Overall Enhanced Time}} = (10)(50\%) + (50\%) = 5.0 + 0.5 = 5.5
   \]

   To solve part (b), observe that in

   \[
   \text{Overall Unenhanced Time} = (10)(50\%) \text{Overall Enhanced Time} + (50\%) \text{Overall Enhanced Time}
   \]

   the first term represents the portion of Unenhanced Time that could be enhanced and the second represents the portion of Unenhanced Time that could not. Therefore, the ratio \( f \) of original time for work that could be enhanced to the total original time, which is the fraction of enhancement \( f \) in the traditional Amdahl’s law formula, is

   \[
   f = \frac{(10)(50\%)}{(10)(50\%) + (50\%)} = \frac{5.0}{5.5} \approx 91\%.
   \]

   The above solution method required mental insight into how to formulate the problem: That is hard to think of but easy to understand once it is done. If all else fails, an alternative approach is to write the given facts in mathematical form and try to solve the equations for quantities you can use to calculate the answer to the question.

   **Equations:**

   \[
   T_{\text{Orig}} = T_{\text{Orig,Other}} + T_{\text{Orig,Enhance-able}}
   \]

   \[
   T_{\text{New}} = T_{\text{Orig,Other}} + T_{\text{Orig,Enhance-able}}/10
   \]

   \[
   T_{\text{Orig,Enhance-able}}/10 = (50\%)T_{\text{New}}
   \]

   Notice that there are 3 homogeneous equations in 4 unknowns.

1.2a **Question:** Speedup = \( T_{\text{Orig}}/T_{\text{New}} \)

   Let’s first eliminate \( T_{\text{Orig,Other}} \):

   \[
   T_{\text{Orig}} - T_{\text{New}} = (1 - \frac{1}{10})T_{\text{Orig,Enhance-able}}
   \]

   \[
   T_{\text{Orig}} - T_{\text{New}} = (0.90)T_{\text{Orig,Enhance-able}}
   \]
Next, eliminate $T_{\text{Orig,Enhance-able}}$:

$$T_{\text{Orig,Enhance-able}} = 10 \cdot 50\% T_{\text{New}} = (5)T_{\text{New}}$$
$$T_{\text{Orig}} - T_{\text{New}} = (0.90)(5)T_{\text{New}} = (4.5)T_{\text{New}}$$

Finally, express the answer:

$$T_{\text{Orig}} = (4.5)T_{\text{New}} + T_{\text{New}} = (5.5)T_{\text{New}}$$
$$\frac{T_{\text{Orig}}}{T_{\text{New}}} = 5.5$$

1.2b Question: “percentage of original execution time converted to fast mode”

is $T_{\text{Orig,Enhance-able}}/T_{\text{Orig}} \times 100\%$

$$T_{\text{Orig,Enhance-able}}/T_{\text{Orig}} = \frac{(5)T_{\text{New}}}{(5.5)T_{\text{New}}}$$
$$T_{\text{Orig,Enhance-able}}/T_{\text{Orig}} = \frac{5}{5.5} = 0.909 \approx 91\%$$

2. Problem 1.3

- Page 31 says FPSQR is responsible for 20% of the original execution time. Page 33 says the frequency of FPSQR instructions is 2% and their average CPI is 20 CPI; hence the contribution to the average CPI due to FPSQR is $(2\%)(20) = 0.4$ CPI. Page 34 shows how to calculate that the average overall CPI is 2.0 CPI. Hence the fraction of cycles used, on the average, for FPSQR compared to total cycles used per instruction, is $(0.4)/(2.0)$ which is 20%. Since the clock cycle time is fixed, this 20% figure is the fraction of original execution time spent for FPSQR, consistent with page 31.

- Page 31 says proposal one is to speed up the FPSQR operation by a factor of 10. Page 33 says the CPI for FPSQR is 20; proposal one of page 34 says to reduce it to 2. Therefore, the speedup is $20/2 = 10$.

- Page 31 says FP operations are responsible for 50% of the total execution time. Page 33 says the frequency of FP operations is 25% and their average CPI is 4.0; hence the contribution to the average CPI due to all FP operations is $(25\%)(4.0) = 1.0$ CPI. Compared to the overall CPI of 2.0 calculated on page 34, this is 50%, consistent with page 31.

- Page 31 says the design team’s second alternative is to speed up overall FP operation by a factor of 2. Page 33 says the average CPI of FP operations is 4.0 CPI, the second alternative on page 34 is to reduce this to 2.0 CPI, again a factor of 2 speedup.

- We saw that every assumption given in the page 31 problem holds true for the page 33-34 problem. Hence, any question that can be answered from the page 31 assumptions will have the same answer for the page 33-34 situation. It is in that sense they are “the same.”

3. Hennessy and Patterson page 61 problem (1.4). We put all the information from Figure 1.17 (page 45) and the problem statement on a chart:
<table>
<thead>
<tr>
<th>Original</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inst. Type</td>
<td>Freq.</td>
</tr>
<tr>
<td>modifiable ALU ops</td>
<td>43%*25%</td>
</tr>
<tr>
<td>other ALU ops</td>
<td>43%*75%</td>
</tr>
<tr>
<td>Loads</td>
<td>21%</td>
</tr>
<tr>
<td>Stores</td>
<td>12%</td>
</tr>
<tr>
<td>Branches</td>
<td>24%</td>
</tr>
</tbody>
</table>

*Beware: this Freq. is relative to the Instruction Count of the Original system. Since some Load instructions will be removed when the modifiable ALU ops are replaced by new ALU-mem ops, the total of this column is less than 100%.

The “modifiable ALU ops” are the 25% of the (dynamic) instructions that are ALU ops that directly use a loaded operand that is not used again. The New computer will have ALU instructions with a memory addressing mode for one source operand; these instructions are not present in the Old computer. We assume that the compiler for the New computer will use one new ALU-mem instruction in place of each (Old compiler’s) pair of instructions composed of a Load followed by an ALU op that uses the loaded operand that is not used again.

Let IC denote the (dynamic) Instruction Count for some program on the Old system. The following spreadsheet shows how the total Cycle Counts can be calculated for the Old and New systems.

<table>
<thead>
<tr>
<th>Original</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>(.43)(.25)IC</td>
<td>1</td>
</tr>
<tr>
<td>(.43)(.75)IC</td>
<td>1</td>
</tr>
<tr>
<td>(.21)IC</td>
<td>2</td>
</tr>
<tr>
<td>(.12)IC</td>
<td>2</td>
</tr>
<tr>
<td>(.24)IC</td>
<td>2</td>
</tr>
<tr>
<td>IC</td>
<td>(1.57)IC</td>
</tr>
</tbody>
</table>

We conclude that even though the New system as a smaller Instruction Count, the Old system still has a smaller Cycle Count. Since the Clock Cycle Time is assumed to be unchanged, the Old system is faster than the New system by a factor of (1.7025/1.57) ≈ 1.08; its performance is better by about 8%. Hence adding memory ALU op instructions to a load store architecture is, under the given assumptions, a “bad idea.”

**Supplemental Questions and Answers**

1. a. Suppose you could run at the rate of 10 km/hr and drive at the rate of 127 km/hr. Ignore the fact that maximum average speed varies with the overall time or distance, especially for running.

0. Illustrate with a qualitative graph and explain in English how and why average running speed depends on the overall distance; include very short distances as well as very long distances. Go to the circle path or the track to do experiments if necessary.
Average speed depends on overall distance because of distance is required to accelerate and warm up to reach maximum speed for short distances, and limits to anaerobic muscle energy, exhaustion, etc., reduce overall speed for longer distances. The distance is limited due to limits on the maximum time a human can keep running. It does not approach 0 because we assume that the overall measurement stops when the runner halts.

The qualitative features shown on the graph are borne out by citations of world class athletic records cited by a couple of students.

Students who wrote formulas, vector diagrams, numbers (other than reports of actual athletic performance to back up their report) or plotted distance against time missed the point of this problem.

aa. Suppose you run \( n_1 \) kilometers and then drive another \( n_2 \) kilometers. Q1: Derive a formula for your average speed for the whole trip in terms of \( n_1 \) and \( n_2 \). Q2: Does your formula give the weighted arithmetic mean (i.e., average) or the weighted harmonic mean of the rates 10 km/hr and 127 km/hr? Q3: What are the weights?

Q1:

\[
\text{Average speed} = \frac{\text{Total distance}}{\text{Total time}}
\]

\[
= \frac{(n_1 + n_2) \text{ km}}{(\frac{n_1}{10} + \frac{n_2}{127}) \text{ hr}}
\]

\[
= \frac{1}{\frac{n_1}{(n_1 + n_2)} \frac{1}{10} \text{ km/hr} + \frac{n_2}{(n_1 + n_2)} \frac{1}{127} \text{ km/hr}}
\]

Q2: This is the weighted harmonic mean of 10 km/hr and 127 km/hr with Q3: weights respectively \( n_1/(n_1 + n_2) \) with weight \( n_2/(n_1 + n_2) \).

ab. Suppose you run for \( t_1 \) hours and then drive another \( t_2 \) hours. Answer the three questions from part (aa.)

Q1:

\[
\text{Average speed} = \frac{(10t_1 + 127t_2) \text{ km}}{(t_1 + t_2) \text{ hr}}
\]

\[
= \left( \frac{t_1}{t_1 + t_2} \right) (11 \text{ km/hr}) + \left( \frac{t_2}{t_1 + t_2} \right) (120 \text{ km/hr})
\]

Q2: This is the weighted arithmetic mean of 10 km/hr and 127 km/hr with Q3: weights respectively \( t_1/(t_1 + t_2) \) and \( t_2/(t_1 + t_2) \).

b. Suppose (somewhat different) running and driving rates were specified by times: 328 seconds to run 1 kilometer and 29 seconds to drive 1 kilometer.
ba. Answer the questions of part (aa.) in terms of the unit times 328 sec/km and 29 sec/km.

Q1:

Average speed = \frac{(n_1 + n_2) \text{ km}}{(n_1 \text{ km})(328 \text{ sec/km}) + (n_2 \text{ km})(29 \text{ sec/km})}

2,3: This is *neither* the weighted arithmetic mean nor the weighted arithmetic mean of the unit time quantities 328 sec/km and 29 sec/km. Beware of poorly formulated problems wherever you encounter them! You can convert the per unit time quantities to speeds and just use the answer to part (aa). If we wish to combine unit time quantities, we should report

Average time per kilometer = (Average speed)^{-1}

= \frac{(n_1 \times 328 + n_2 \times 29) \text{ sec}}{(n_1 + n_2) \text{ km}}

= \left( \frac{n_1}{n_1 + n_2} \right)(328 \text{ sec/km}) + \left( \frac{n_2}{n_1 + n_2} \right)(29 \text{ sec/km})

This is the weighted arithmetic mean of 328 sec/km with weight \( n_1/(n_1 + n_2) \) and 29 sec/km with weight \( n_2/(n_1 + n_2) \).

bb. Answer the questions of part (ab.) in terms of the unit times 328 sec/km and 29 sec/km.

Q1:

Average speed = \frac{(t_1/328 + t_2/29)\text{km}}{(t_1 + t_2)\text{sec}}

Q2,3: Again, this is neither the weighted arithmetic mean nor the weighted harmonic mean of the quantities 328 sec/km and 29 sec/km! One could, as before, convert to speeds and use the answer to part (ab). However,

Average time per kilometer = (Average speed)^{-1}

= \frac{(t_1 + t_2) \text{ sec}}{\left( \frac{t_1}{328 \text{ sec/km}} + \frac{t_2}{29 \text{ sec/km}} \right) \text{ km}}

= \frac{1 \text{ sec/km}}{\left( \frac{t_1}{t_1 + t_2} \right)(330 \text{ sec/km}) + \left( \frac{t_2}{t_1 + t_2} \right)(30 \text{ sec/km})}

is the weighted harmonic mean of 328 sec/km with weight \( t_1/(t_1 + t_2) \) and 29 sec/km with weight \( t_2/(t_1 + t_2) \) sec/km.

Please observe: Weights are non-dimensional fractions; no units are given for them.

**Mathematical Summary**

There are \( m \) categories of operations, numbered by \( j = 1, \ldots, m \). For each operation \( j \) there is quantity of “results” \( N_j \) and a quantity of “effort” or \( D_j \). The total amount of results is

\[ N = N_1 + N_2 + \cdots + N_m, \]
and “effort”

\[ D = D_1 + D_2 + \cdots + D_m. \]

For each category \( j \), the “rate" \( r_j \) is

\[ r_j = \frac{N_j}{D_j}. \]

The overall or average “rate” \( r \) is

\[ r = \frac{N}{D} = \frac{N_1 + N_2 + \cdots + N_m}{D_1 + D_2 + \cdots + D_m}. \]

In performance reporting or analysis, the individual operation “rates” \( r_j \) are usually determined or estimated by analysis of the operation’s architecture. Frequencies are measured from benchmarks. Sometimes frequencies are given in terms of “efforts"

\[ w_j = \frac{D_j}{D_1 + D_2 + \cdots + D_m} = \frac{D_j}{D} \]

and sometimes in terms of “results”

\[ f_j = \frac{N_j}{N_1 + N_2 + \cdots + N_m} = \frac{N_j}{N}. \]

These definitions of \( w_j \) and \( f_j \) imply that they are frequency distributions: \( \sum w_j = 1 \) and \( 0 \leq w_j \leq 1 \); similarly for \( f_j \).

When an overall “rate” is a total of “results” divided by a total of “efforts,” optimizing the rate means maximizing it. (“Get the biggest Bang for the Buck, or greatest performance for a given price.”)

Conclusions:

- The overall “rate" \( R \) is the weighted arithmetic mean of the \( r_j \) when the frequencies \( w_j \) are given in terms of “efforts”.
- The overall “rate" \( R \) is the weighted harmonic mean of the \( r_j \) when the frequencies \( f_j \) are given in terms of “results”.

Proof for “effort frequencies”:

\[ R = \frac{N_1}{D} + \frac{N_2}{D} + \cdots + \frac{N_m}{D} \]

\[ R = \frac{N_1 D_1}{D_1 D} + \frac{N_2 D_2}{D_2 D} + \cdots + \frac{N_m D_m}{D_m D} \]

\[ R = r_1 w_1 + r_2 w_2 + \cdots + r_m w_m \]

Proof for “result frequencies”:

\[ \frac{1}{R} = \frac{D}{N} = \frac{D_1}{N} + \frac{D_2}{N} + \cdots + \frac{D_m}{N} \]

\[ \frac{1}{R} = \frac{D_1 N_1}{N} + \frac{D_2 N_2}{N} + \cdots + \frac{D_m N_m}{N} \]

\[ \frac{1}{R} = \frac{1}{r_1} f_1 + \frac{1}{r_2} f_2 + \cdots + \frac{1}{r_m} f_m \]

\[ R = \frac{1}{f_1} + \frac{1}{f_2} + \cdots + \frac{1}{f_m} \]

6
Application: For travelling, \( N_t \) signifies the distance and \( D_t \) signifies time. For computing performance, \( N_t \) signifies jobs and \( D_t \) signifies time. For CPI of computers, \( N_t \) signifies number of cycles consumed and \( D_t \) signifies the number of instructions that consumed them. When \( r \) reports CPI, smaller numbers are better.

2. (a) If you invest \$1000. in a savings account that has an interest rate of 3.5% per year compounded annually, how much money will be in the account after 10 years, assuming none is withdrawn during that time?

\[
\$1000 \cdot (1.035)^{10} = \$1000 \cdot (1.4105988) = \$1410.60
\]

(b) If you wish to double your money in 5 years, what annual interest rate must you get, assuming annual compounding and no withdrawals?

Find \( x \) for which \((1 + x)^5 = 2.\) That is the \( x \) for which \(1 + x = 2^{1/5} \). \(2^{1/5} = 1.148698,\) so an annual interest rate of approximately 0.1487 or 14.87% is required.

The next two parts refer to figure 1.1 on page 2. of Hennessy and Patterson.
(c) Explain why the grey curve marked “1.35x per year” is not a straight line?

Both the “x” and “y” axis scales are linear, and the grey curve represents an exponential, not a linear function. The derivative of an exponential function is proportional to the function value, so the slope is not constant.

(d) What is the average rate of growth per year of performance between the years 1987 and 1990? Between the years 1990 and 1994? Give the best estimate you can based on the data you can read from the graph.

Read roughly from the graph, the SPECint ratings are 12 for 1987 (SUN4 which is the first SPARC), 24 for 1990 (IBM Power1) and 125 for 1994 (DEC Alpha). Growth during the 4 years between 1990 and 1994 was by overall factor of 2, so the average growth factor per year was \(2^{(1/4)}\) or 1.19, that is 19% per year.

Between 1990 and 1994, overall growth factor 125/24 \(\approx 5\) averages to \(5(1/4) \approx 1.495\) times per year, that is, around 50% per year.

Pitfall: The overall performance improvement between 1987 and 1990 was 100% (that’s true). However, it is false to report the average growth as \(100%/4 = 25%\) per year.

Indeed our example calculates average yearly growth based on the assumption that growth is compounded annually. Under the model of continuous compounding of growth, \(dP/dt = gP\) where \(P\) is performance, \(t\) is time measured in years, and \(g\) is (continuously compounded) growth per year. The solution to this differential equation is \(P = C \exp(gt)\), so overall growth by a factor of 2 in 4 years corresponds to the \(g\) value for which \(\exp(g \cdot 4) = 2\). Such \(g = \ln(2)/4 \approx 17%\). Growth by a factor of 5 in 4 years corresponds to a rate of \(\ln(5)/4 \approx 40.2\%\), compounded continuously.