NYU Tandon School of Engineering

Fall 2022, ECE 6913

Homework Assignment 2

Instructor: Azeez Bhavnagarwala, email: ajb20@nyu.edu

Course Assistants

Varadraj Kakodkar (vns2008), Kartikay Kaushik (kk4332), Siddhanth Iyer (si2152), Swarnashri Chandrashekar (sc8781), Karan Sheth (kk4332), Haotian Zheng (hz2687), Haoren Zhang (kk4332), Varun Kumar (vs2411)

Homework Assignment 2 [released Tuesday September 20th 2022] [due Friday September 30th by 11:59PM]

You *are allowed* to discuss HW assignments with anyone. You are *not allowed* to share your solutions with other colleagues in the class. Please feel free to reach out to the Course Assistants or the Instructor during office hours or by appointment if you need any help with the HW.
Please enter your responses in this Word document after you download it from NYU Classes. *Please use the Brightspace portal to upload your completed HW.*

---

1. After graduating, you are asked to become the lead computer designer at Hyper Computers, Inc. Your study of usage of high-level language constructs suggests that procedure calls are one of the most expensive operations. You have invented a scheme that reduces the loads and stores normally associated with procedure calls and returns. The first thing you do is run some experiments with and without this optimization. Your experiments use the same state-of-the-art optimizing compiler that will be used with either version of the computer. These experiments reveal the following information:

- The clock rate of the unoptimized version is 5% higher.
- 30% of the instructions in the unoptimized version are loads or stores.
- The optimized version executes 2/3 as many loads and stores as the unoptimized version. For all other instructions the dynamic counts are unchanged.
- All instructions (including load and store) take one clock cycle.

Which is faster? Justify your decision quantitatively.

Ans]

CPU performance equation:

Execution Time = IC * CPI * Tcycle

The unoptimized version is 5% faster in clock cycle time

Tcycle_unop = 0.95 * Tcycle_op

Instruction Count (IC) of load/store instructions are 30% of total IC in the unoptimized version

IC_ld/st_unop = 0.3 * IC_unop

IC of load store instructions in optimized version is 0.67 of IC of load/store instructions in unoptimized version

$$IC\_ld/st\_op = 0.67 * IC\_ld/st\_unop$$

IC of all other (non load/store) instructions in optimized version = IC of all other (non load/store) instructions in unoptimized version

$$IC\_others\_unop = IC\_others\_op$$

$$CPI = 1$$

$$Execution\ Time\_unop = IC\_unop * Tcycle\_unop$$

$$= 0.95 * IC\_unop * Tcycle\_op$$

$$Execution\ Time\_op = IC\_op * Tcycle\_op$$

$$IC\_op = IC\_others\_op + IC\_ld/st\_op$$

$$IC\_op = IC\_others\_unop + IC\_ld/st\_op$$

$$IC\_op = 0.7 * IC\_unop + 0.67 * IC\_ld/st\_unop$$

$$IC\_op = 0.7 * IC\_unop + 0.67 * (0.3 * IC\_unop) = 0.7 * IC\_unop + 0.2*IC\_unop$$

$$= 0.9 * IC\_unop$$

$$Execution\ Time\_op = IC\_op * Tcycle\_op$$

$$= 0.9 * IC\_unop * 1.05 * Tcycle\_unop$$

$$= 0.945 * IC\_unop * Tcycle\_unop$$

Improvement of 5.5%

2. General-purpose processes are optimized for general-purpose computing. That is, they are optimized for behavior that is generally found across a large number of applications. However, once the domain is restricted somewhat, the behavior that is found across a large number of the target applications may be different from general-purpose applications. One such application is deep learning or neural networks. Deep learning can be applied to many different applications, but the fundamental building block of inference—using the learned information to make decisions—is the same across them all. Inference operations are largely parallel, so they are currently performed on graphics processing units, which are specialized more toward this type of computation, and not to inference in particular. In a quest for more performance per watt, Google has created a custom chip using tensor processing units to accelerate inference operations in deep learning.1 This approach can be used for speech recognition and image recognition, for example. This problem explores the trade-offs between this process, a general-purpose processor (Haswell E5-2699 v3) and a GPU (NVIDIA K80), in terms of performance and cooling. If heat is not removed from the computer efficiently, the fans will blow hot air back onto the computer, not cold air. Note: The differences are more than processor—on-chip memory and DRAM also come into play. Therefore statistics are at a system level, not a chip level.

a. If Google's data center spends 70% of its time on workload A and 30% of its time on workload B when running GPUs, what is the speedup of the TPU system over the GPU system?

b. Google's data center spends 70% of its time on workload A and 30% of its time on workload B when running GPUs, what percentage of Max IPS does it achieve for each of the three systems?

c. Building on (b), assuming that the power scales linearly from idle to busy power as IPS grows from 0% to 100%, what is the performance per watt of the TPU system over the GPU system?

d. If another data center spends 40% of its time on workload A, 10% of its time on workload B, and 50% of its time on workload C, what are the speedups of the GPU and TPU systems over the general-purpose system?

e. A cooling door for a rack cost $4000 and dissipates 14 kW (into the room; additional cost is required to get it out of the room). How many Haswell-, NVIDIA-, or Tensor-based servers can you cool with one cooling door, assuming TDP in Figures 1.27 and 1.28?

f. Typical server farms can dissipate a maximum of 200 W per square foot. Given that a server rack requires 11 square feet (including front and back clearance), how many servers from part (e) can be placed on a single rack, and how many cooling doors are required?

System	Chip	TDP	Idle power	Busy power
General-purpose	Haswell E5-2699 v3	504 W	159 W	455 W
Graphics processor	NVIDIA K80	1838 W	357 W	991 W
Custom ASIC	TPU	861 W	290 W	384 W

Figure 1.27 Hardware characteristics for general-purpose processor, graphical processing unit-based or custom ASIC-based system, including measured power

System	Chip	Throughput			% Max IPS		
		A	B	C	A	B	C
General-purpose	Haswell E5-2699 v3	5482	13,194	12,000	42%	100%	90%
Graphics processor	NVIDIA K80	13,461	36,465	15,000	37%	100%	40%
Custom ASIC	TPU	225,000	280,000	2000	80%	100%	1%

Figure 1.28 Performance characteristics for general-purpose processor, graphical processing unit-based or custom ASIC-based system on two neural-net workloads

2 a:

Performance gain obtained by improving some portion of the computer speed

= (Execution Time for entire task without using the enhancement) / (Execution Time for entire task using the enhancement)

Speedup of computer A over B = Execution Time of B / Execution Time of A – (1)

From Table 2:

Speedup of TPU over GPU for workload A = 225000/13461 = 16.7 – (2)

Speedup of TPU over GPU for workload B = 280000/36465 = 7.7 – (3)

From (1 –3), considering 70% of a GPUs Execution Time is for workload A, 30% for B

0.7 x Ex Time (A+B) of GPU = Ex Time (A) of TPU x 16.7 (for workload A)

0.3 x Ex Time (A+B) of GPU = Ex Time (B) of TPU x 7.7 (for workload B)

So, Ex Time (A+B) of TPU

= [0.7 x Ex Time (A+B) of GPU /16.7 + 0.3 x Ex Time (A+B) of GPU/7.7]

Ex Time (A+B) TPU = [0.7/16.7 + 0.3/7.7] Ex Time (A+B) GPU

So, [Ex Time (A+B) GPU / Ex Time (A+B) TPU] = speedup of TPU over GPU = [from (1) above]

= 1/ [0.7/16.7 + 0.3/7.7] = 1/[0.0419 + .03896] = 12.37

2 b: **IPS [Instructions/sec] = Clock rate [cycles/sec] / CPI [cycles/Instruction]**

Datacenter spends 70% of its time in workload A & 30% of its time in workload B

The general purpose CPU can accomplish only 42% of its maximum IPS in workload A and 100% of its maximum IPS in workload B

so, maximum IPS it can possibly achieve = 42% x 70% + 100% x 30% = 0.294 + 0.3 = 0.594

In GPU maximum IPS in workload A, B = 37%, 100%

so, 37% x 70% + 100% x 30% = 0.37 * 0.7 + 1 * 0.3 = 0.559

In Custom ASIC (TPU), maximum IPS in workload A, B = 80%, 100% so, 0.8 x 0.7 + 1 x 0.3 = 0.86

2 c: Linear dependence of power on IPS

at max IPS, processor is consuming 'busy' power

at 0 IPS, processor is consuming idle power

Idle power + [max power –idle power] x [% of max IPS] = Power

CPU: 159 W + (455 W-159 W) x0.594= 335 W

GPU: 357 W + (991 W-357 W) x0.559= 711 W

TPU: 290 W + (384 W-290 W) x0.86= 371 W

Ratio of performance per watt (PPW) of system1 to system 2

= [% max IPS/Power]1 / [% of max IPS/Power]2

for TPU:GPU comparison

PPW Ratio of TPU to GPU = 0.86 x 711 / 0.559 x 371 = 2.95

The TPU delivers almost 3X higher performance per watt for this workload

2 d: Another data center spends 40% of its time on workload A, 10% of its time on workload B, and

50% of its time on workload C

Speedup of GPU over CPU for workload A = 13461 / 5482

Speedup of GPU over CPU for workload B = 36465 / 13194

Speedup of GPU over CPU for workload C = 15000 / 12000

So, Speedup of GPU over general-purpose CPU is

Speedup of TPU over CPU for workload A = 225000 / 5482

Speedup of TPU over CPU for workload B = 280000 / 13194

Speedup of TPU over CPU for workload C = 2000 / 12000

So, Speedup of TPU over general-purpose CPU is:

2 e. TDP or Thermal Design Power is higher than the maximum power dissipation from a processor and thus equal to the minimum rate of heat removal that must be exceeded by cooling so that junction temperatures in the chip do not exceed their max spec. T

A 'cooling door ' removes 14kW of power from a rack holding multiple server boards and are expensive ($4K)

With a cooling door, 14kW/504W = # of CPUs that can be cooled = 27

With a cooling door, 14kW/1838W = # of GPUs that can be cooled = 7

With a cooling door, 14kW/861W = # of TPUs that can be cooled = 16

2 f: Maximum power per rack in a server farm: 200 W/ft2 x 11 ft2= 2200 W

Maximum number of servers per rack & number of cooling doors for this maximum:

CPU: 2200W / 504W = 4.37; 4 servers ;

GPU: 2200W / 1838W = 1.19; 1 server ;

TPU: 2200W / 861W = 2.55; 2 servers ;

One cooling door is enough to dissipate 2200W.

3. In this exercise, assume that we are considering enhancing a quad-core machine by adding encryption hardware to it. When computing encryption operations, it is 20 times faster than the normal mode of execution. We will define percentage of encryption as the percentage of time in the original execution that is spent performing encryption operations. The specialized hardware increases power consumption by 2%.

a. Draw a graph that plots the speedup as a percentage of the computation spent performing encryption. Label the y-axis "Net speedup" and label the x-axis "Percent encryption."

b. With what percentage of encryption will adding encryption hardware result in a speedup of 2?

c. What percentage of time in the new execution will be spent on encryption operations if a speedup of 2 is achieved?

Ans]

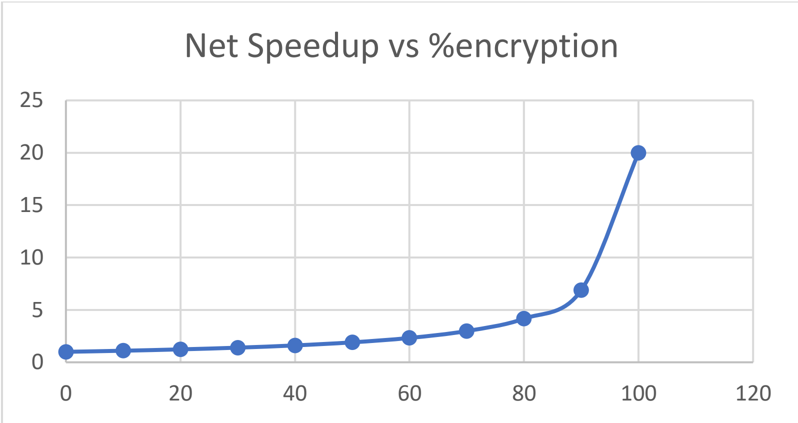
a. **Net Speedup** $= \boldsymbol{TE_{old}} / \boldsymbol{TE_{new}}$

$$= \frac{\text{TEold}}{[\%\text{encryption}]\ \textbf{TE}\text{old}/20 + [100\%-\%\text{encryption}]\ \textbf{TE}\text{old}/1}$$

$$= \frac{1}{0.05[\%\text{encryption}] + [1-\%\text{encryption}]}$$

$$= \frac{1}{1-0.95[\%\text{encryption}]}$$

%encryption	Net Speedup
0	1
10	1.104972376
20	1.234567901
30	1.398601399
40	1.612903226
50	1.904761905
60	2.325581395
70	2.985074627
80	4.166666667
90	6.896551724
100	20



b. Net Speedup = $\frac{1}{1-0.95[\%encryption]}$

1-0.95[%encryption] = 1/2

[%encryption] = **52.63%**

c. **% Time spent =** $\frac{x/20}{(\frac{x}{20})+(100\%-x)}$

Here, x=0.5263

$$= \frac{0.5263/20}{(\frac{0.5263}{20})+(1-0.5263)}$$

$$= \frac{0.0263}{0.0263+0.4737}$$

= **5.26%**

4. Assume that we make an enhancement to a computer that improves some mode of execution by a factor of 10. Enhanced mode is used 50% of the time, measured as a percentage of the execution time when the enhanced mode is in use. Recall that Amdahl's Law depends on the fraction of the original, unenhanced execution time that could make use of enhanced mode. Thus, we cannot directly use this 50% measurement to compute speedup with Amdahl's Law.

a. What is the speedup we have obtained from fast mode?

b. What percentage of the original execution time has been converted to fast mode?

Ans]

a. Let x be %enhancement

$$\textbf{\% Time spent} = \frac{\frac{x}{10}}{\left(\frac{x}{10}\right)+(100\%-x)}$$

$$\mathbf{0.5} = \frac{x/10}{\left(\frac{x}{10}\right)+(1-x)}$$

0.1x + 1 - x = 0.2x

x = 0.909

Net Speedup $= \boldsymbol{TE_{old}} / \boldsymbol{TE_{new}}$

$$= \frac{\text{TEold}}{[\%\text{enhancement}]\ \mathbf{TE}\text{old}/10 + [100\%-\%\text{enhancement}]\ \mathbf{TE}\text{old}/1}$$

$$= \frac{1}{0.1[0.909] + [1-0.909]}$$

$$= \frac{1}{0.1[0.901] + [0.099]}$$

$$= \frac{1}{0.1819}$$

= 5.497

b. (already calculated above)
To find the percentage of the original execution time which was accelerated, we plug these figures into Amdahl's Law again:

$$\text{Fraction vectorised} = \frac{\text{speeduprm overall X speedup accelerated} - \text{speedup accelerated}}{\text{speedup overll X speedup accelerated} - \text{speedup overall}}$$

= 5.5 X 10 -10 / 5.5 X 10 - 5.5
= 45 / 49.5
= 90.90%

5. When parallelizing an application, the ideal speedup is speeding up by the number of processors. This is limited by two things: percentage of the application that can be parallelized and the cost of communication. Amdahl's Law takes into account the former but not the latter.

a. What is the speedup with N processors if 80% of the application is parallelizable, ignoring the cost of communication?

b. What is the speedup with eight processors if, for every processor added, the communication overhead is 0.5% of the original execution time.

c. What is the speedup with eight processors if, for every time the number of processors is doubled, the communication overhead is increased by 0.5% of the original execution time?

d. What is the speedup with N processors if, for every time the number of processors is doubled, the communication overhead is increased by 0.5% of the original execution time?

e. Write the general equation that solves this question: What is the number of processors with the highest speedup in an application in which P% of the original execution time is parallelizable, and, for every time the number of processors is doubled, the communication is increased by 0.5% of the original execution time?

Ans]

a.

$$\text{Net Speedup} = \frac{1}{(1-0.8)+0.8/N} = \frac{1}{0.2+0.8/N}$$

b.

$$\text{Net Speedup} = \frac{1}{0.2+(8 \times \mathbf{0.005})+0.8/8} = \frac{1}{0.2+0.04+0.1} = \frac{1}{0.34} = 2.94$$

c.

we know that Log 8 = 3

$$\text{Net Speedup} = \frac{1}{0.2+(3 \times \mathbf{0.005})+0.8/8} = \frac{1}{0.2+0.015+0.1} = \frac{1}{0.315} = 3.17$$

d.

$$\text{Net Speedup} = \frac{1}{0.2+(\mathbf{log}\,\boldsymbol{N} \times \mathbf{0.005})+0.8/N}$$

e. Since the number of processors = N, Percentage of parallelization = P

Net Speedup = $\frac{1}{(1-\mathrm{P})+ (\mathbf{log}\,\boldsymbol{N} \text{ x } \mathbf{0.005}) + \mathrm{P/N}}$

For max speedup,

$$\frac{d}{dN}\text{ Net Speedup} = 0$$

$$\left(\frac{1}{(1-\mathrm{P})+ (\mathbf{log}\,\boldsymbol{N} \text{ x } \mathbf{0.005}) + \mathrm{P/N}}\right)^{-2} * \left((1-\mathrm{P}) + \left(\frac{0.005}{\mathrm{N} * (\ln 2)}\right) - \frac{\mathrm{P}}{\mathrm{N^{\wedge}2}}\right) = 0$$

$\mathrm{P/N} = 0.005/\ln 2$

Hence $\mathbf{N = 138.62*P}$

Where P is the % parallelization fraction value

6. Your company has just bought a new 22-core processor, and you have been tasked with optimizing your software for this processor. You will run four applications on this system, but the resource requirements are not equal. Assume the system and application characteristics listed in Table 1.1 below (from textbook)

Table 1.1 Four applications

Application	A	B	C	D
% resources needed	41	27	18	14
% parallelizable	50	80	60	90

The percentage of resources of assuming they are all run in serial. Assume that when you parallelize a portion of the program by X, the speedup for that portion is X.

a. How much speedup would result from running application A on the entire 22-core processor, as compared to running it serially?

b. How much speedup would result from running application D on the entire 22-core processor, as compared to running it serially?

c. Given that application A requires 41% of the resources, if we statically assign it 41% of the cores, what is the overall speedup if A is run parallelized but everything else is run serially?

d. What is the overall speedup if all four applications are statically assigned some of the cores, relative to their percentage of resource needs, and all run parallelized?

e. Given acceleration through parallelization, what new percentage of the resources are the applications receiving, considering only active time on their statically-assigned cores?

Ans]

a. Net Speedup = $\frac{1}{(1-0.5)+0.5/22}$

= 1.91

b. Net Speedup = $\frac{1}{(1-0.8)+0.8/22}$

= 7.096

c.

Number of cores required for application A = 0.41 * 22 = 9.02

Net Speedup for A = $\frac{1}{(1-0.5)+0.5/9}$

= 1.81

Overall Speedup = $\frac{1}{(1-0.41)+0.41/1.81}$

= 1.22

d.

Number of cores for A = 0.41 * 22 = 9.02

Number of cores for B = 0.27 * 22 = 5.94

Number of cores for C = 0.18 * 22 = 3.96

Number of cores for D = 0.14 * 22 = 3.08

Net Speedup for A = $\frac{1}{(1-0.5)+0.5/9}$ = 1.8

Net Speedup for B = $\frac{1}{(1-0.8)+0.8/5}$ = 3

Net Speedup for C = $\frac{1}{(1-0.6)+0.6/4}$ = 1.82

Net Speedup for D = $\frac{1}{(1-0.9)+0.9/3}$ = 2.5

Overall Speedup = $\frac{1}{\frac{0.41}{1.8}+\frac{0.27}{3}+\frac{0.18}{1.82}+\frac{0.9}{2.5}}$ = **2.12**

e.

With acceleration through parallelization,

% of resources received by app A = 0.41/speedup(A) = 0.41/1.8 = 0.2277 = **22.77%**

% of resources received by app B = 0.27/speedup(B) = 0.27/3 = 0.09 = **9%**

% of resources received by app C = 0.18/speedup(C) = 0.18/1.82 = 0.0989 = **9.89%**

% of resources received by app D = 0.14/speedup(D) = 0.14/2.5 = 0.056 = **5.6%**

7. When making changes to optimize part of a processor, it is often the case that speeding up one type of instruction comes at the cost of slowing down something else. For example, if we put in a complicated fast floating-point unit, that takes space, and something might have to be moved farther away from the middle to accommodate it, adding an extra cycle in delay to reach that unit. The basic Amdahl's Law equation does not take into account this trade-off.

a. If the new fast floating-point unit speeds up floating-point operations by, on average, 2x, and floating-point operations take 20% of the original program's execution time, what is the overall speedup (ignoring the penalty to any other instructions)?

b. Now assume that speeding up the floating-point unit slowed down data cache accesses, resulting in a 1.5x slowdown (or 2/3 speedup). Data cache accesses consume 10% of the execution time. What is the overall speedup now?

c. After implementing the new floating-point operations, what percentage of execution time is spent on floating-point operations? What percentage is spent on data cache accesses?

Ans]

a. Net Speedup = $\frac{1}{(1-0.2)+0.2/2}$

= 1.11

b. Data cache accesses = 10%
Floating point operations = 20%

Data cache access speedup = 2/3
Floating point operations speedup = 2

Net Speedup = $\frac{1}{(1-0.1-0.2)+0.1\left(\frac{3}{2}\right)+\frac{0.2}{2}}$

$= \frac{1}{(0.7)+(0.15)+(0.1)}$

= 1.05

c.

The fractional data cache access is 0.1*3/2=0.15
fractional floating point is is 0.2/2

% Time spent(FP) = $\frac{\frac{0.2}{2}}{\left(\frac{0.2}{2}\right)+\left(\frac{0.1}{\frac{2}{3}}\right)+(0.7)} = \mathbf{10.5\%}$

$$\textbf{\% Time spent(Data Cache)} = \frac{\frac{0.1}{\frac{2}{3}}}{\left(\frac{0.2}{2}\right)+\left(\frac{0.1}{\frac{2}{3}}\right)+(0.7)} = \mathbf{15.78\%}$$