CSCE 692 – Chapter 1 Homework Problems

2nd Lt David Crow – 6 January 2019

(100 Points)

Due: NLT 1100 Tuesday, 15 January 2019

**Problems: 1.1, 1.2, 1.4, 1.7, 1.8, 1.16 (5 pts / subproblem)**

**Instructions:**

* Print your name on each page
* Clearly indicate your answer
* Clearly show your work such that I can understand your thinking
* Explain any assumptions and provide references (sources) for any additional information you had to research in order to complete any problem.

Recommendations (i.e., how to score max points)

* Read through the problem statement in this handout, which includes important clarifications from the text.
* Think through and work out the problems on scratch paper first, then copy your answers (electronically, or by hand) into this handout.

**Case Study 1: Chip Fabrication Cost** 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Chip** | **Die Size (mm2)** | **Estimated defect rate**  **(per cm2)** | **N** | **Manufacturing size (nm)** | **Transistors (billion)** | **Cores** |
| BlueDragon | 180 | 0.03 | 12 | 10 | 7.5 | 4 |
| RedDragon | 120 | 0.04 | 14 | 7 | 7.5 | 4 |
| Phoenix | 200 | 0.04 | 14 | 7 | 12 | 8 |

**Figure 1.26 ­Manufacturing cost factors for several hypothetical current and future processors**

1. [5/5] <1.6> Figure 1.26 gives the hypothetical relevant chip statistics that influence the cost of several current chips. In the next few exercises, you will be exploring the effect of different possible design decisions for the Intel chips.
   1. [5] <1.6> What is the yield, or percentage of good dies, for the Phoenix chip?

According to the book, . If we let , then will represent the percentage of good dies. In using our numbers, we see that , which is equal to . In other words, of dies for the Phoenix chip are good.

* 1. [5] <1.6> Why does the Phoenix chip have a higher defect rate than BlueDragon?

Because the BlueDragon chip has a larger manufacturing size than that of the Phoenix, we know that the BlueDragon chip has been around for a longer period of time. For this reason, we know that those responsible for producing BlueDragon have had more time to refine the manufacturing process and thus lower the defect rate.

1. [5/5] <1.6> They will sell a range of chips from that factory, and they need to decide how much capacity to dedicate to each chip. Imagine that they will sell two chips. Phoenix is a completely new architecture designed with 7 nm technology in mind, whereas RedDragon is the same architecture as their 10 nm BlueDragon. Imagine that RedDragon will make a profit of $15 per defect-free chip. Phoenix will make a profit of $30 per defect-free chip. Each wafer has a 450 mm (45 cm) diameter.
   1. [5] <1.6> How much profit do they make on each wafer of Phoenix chips?

The book tells us that . We can use our numbers to show that . Because of these dies are defect-free, we have a total profit of .

* 1. [5] <1.6> How much profit do they make on each wafer of RedDragon chips?

We can calculate the yield like we did in 1.1a. Here, the yield for the RedDragon chips is .

Similarly, each wafer produces . Thus, the total profit is .

**Case Study 2: Power Consumption in Computer Systems**

1.4 [5/5/5/5] <1.5> A cell phone performs very different tasks, including streaming music, streaming video, and reading email. These tasks perform very different computing tasks. Battery life and overheating are two common problems for cell phones, so reducing power and energy consumption are critical. In this problem, we consider what to do when the user is not using the phone to its full computing capacity. For these problems, we will evaluate an unrealistic scenario in which the cell phone has no specialized processing units. Instead, it has a quad-core, general-purpose processing unit. Each core uses 0.5 W at full use. For email-related tasks, the quad-core is 8x as fast as necessary.

1. [5] <1.5> How much dynamic energy and power are required compared to running at full power? First, suppose that the quad-core operates for 1/8 of the time and is idle for the rest of the time. That is, the clock is disabled for 7/8 of the time, with no leakage occurring during that time. Compare total dynamic energy as well as dynamic power while the core is running.

If the quad-core runs for of the time (as compared to running the whole time), then it will use of the power.

Similarly, we know that , so using the same amount of power for of the time means that this task will use of the energy (as compared to running the whole time).

1. [5] <1.5> How much dynamic energy and power are required using frequency and voltage scaling? Assume frequency and voltage are both reduced to 1/8 the entire time.

As the books says, , so reducing the voltage to gives . This corresponds with using of the energy.

The book also tells us that is equal to . If we reduce both frequency and voltage to , we see that , which is the same as . This corresponds with using of the power.

1. [5] <1.6, 1.9> Now assume the voltage may not decrease below 50% of the original voltage. This voltage is referred to as the voltage floor, and any voltage lower than that will lose the state. Therefore, while the frequency can keep decreasing, the voltage cannot. What are the dynamic energy and power savings in this case?

If we reduce frequency to as before and now reduce voltage to , we see that gives a power usage of . Furthermore, we see that gives an energy usage of .

1. [5] <1.5> How much energy is used with a dark silicon approach? This involves creating specialized ASIC hardware for each major task and power gating those elements when not in use. Only one general-purpose core would be provided, and the rest of the chip would be filled with specialized units. For email, the one core would operate for 25% the time and be turned completely off with power gating for the other 75% of the time. During the other 75% of the time, a specialized ASIC unit that requires 20% of the energy of a core would be running.

Because , we can say that , where is the required amount of time to complete a given task. Furthermore, we can say that .

We know that the general-purpose core uses for of the time, so we say that . Because , we can rewrite this as .

Furthermore, we know that the ASIC uses of the energy as before, so we say .

Thus, . In other words, the new configuration only uses of the energy that the old configuration requires.

**Exercises**

1.7 [5/5/5/5/5] <1.4, 1.5> One challenge for architects is that the design created today will require several years of implementation, verification, and testing before appearing on the market. This means that the architect must project what the technology will be like several years in advance. Sometimes, this is difficult to do.

1. [5] <1.4> According to the trend in device scaling historically observed by Moore’s Law, the number of transistors on a chip in 2025 should be how many times the number in 2015?

If we use Gordon Moore’s amended law (that is, his 1975 claim that the number of transistors on a chip would double every two years), we should expect the number of transistors on a chip to double in 2017, 2019, 2021, 2023, and 2025. We should thus expect a chip in 2025 to have times as many transistors as a chip in 2015.

1. [5] <1.5> The increase in performance once mirrored this trend. Had performance continued to climb at the same rate as in the 1990s, approximately what performance would chips have over the VAX-11/780 in 2025?

According to figure 1.1 in the textbook, performance increased at a rate of every year during the 1990s. The -era, as I’ll call it, ended in 2003 with an Intel Xeon chip, which performed times better than the VAX-11/780. We can extrapolate to the year 2025 to predict a performance of times the VAX-11/780.

1. [5] <1.5> At the ~~current~~ rate of increase of the mid-2000s, what is a more updated projection of performance in 2025?

If we instead consider the fact that a 2011 chip outperformed the VAX-11/780 by times at the end of the -era, we should expect a 2025 chip to outperform the same VAX by times.

1. [5] <1.4> What has limited the rate of growth of the clock rate, and what are architects doing with the extra transistors now to increase performance?

The book says that, since 2004 or so, “current and voltage couldn’t keep dropping and still maintain the dependability of integrated circuits” (5). Put another way, we can’t further reduce the voltage over a chip because it needs to overcome the laws of physics.

Furthermore, we can’t really increase the power to a chip. At its current levels, “we are near the limit of what can be cooled by air” (26). We can no longer increase the power to a chip and retain the ability to effectively dissipate the added heat.

We can’t decrease the voltage, and we can’t increase the power, so we can’t really increase the clock rate. Computer architects now look to improve energy efficiency by a) dynamically turning off the clock, b) employing DVFS, c) designing for typical use-cases, and d) allowing for overclocking.

Alternatively, computer architects frequently design special-purpose processors that can perform certain tasks much faster and in a much more efficient way than a general-purpose processor.

Finally, computer architects now place multiple cores on each chip. In doing so, they can utilize the extra transistors and continue to improve performance.

1. [5] <1.4> The rate of growth for DRAM capacity has also slowed down. For 20 years, DRAM capacity improved by 60% each year. If 8 Gbit DRAM was first available in 2015, and 16 Gbit is not available until 2019, what is the current DRAM growth rate?

Clearly, we see that DRAM capacity doubles over the four years between 2015 and 2019. Mathematically, we can write that , and thus corresponds to a growth rate of . In other words, we can expect DRAM to increase by each year. Of course, we’ve seen that increases in other areas can suddenly drop, so it’s certainly possible that an increase each year is not feasible.

1.8 [5/5] <1.5> You are designing a system for a real-time application in which specific deadlines must be met. Finishing the computation faster gains nothing. You find that your system can execute the necessary code, in the worst case, twice as fast as necessary.

1. [5] <1.5> How much energy do you save if you execute at the current speed and turn off the system when the computation is complete?

If we execute at the current speed, and if we shut off the system twice as fast as necessary, we can halve the required energy usage. In other words, we can save of the expected energy costs.

1. [5] <1.5> How much energy do you save if you set the voltage and frequency to be half as much?

The book says that . It’s clear from this equation that frequency does not affect the energy usage. If we halve the voltage, though, we can expect , which is equal to . In other words, by halving the voltage, we reduce our energy usage to one-fourth the original usage.

1.16 [5/5/5/5/5] <1.10> 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.

1. [5] <1.10> What is the speedup with N processors if 80% of the application is parallelizable, ignoring the cost of communication?

The book says . Thus, .

1. [5] <1.10> What is the speedup with eight processors if, for every processor added, the communication overhead is 0.5% of the original execution time?

Our answer here is similar. However, we must add (one overhead of for each new processor) to our new execution time. Thus, our new .

1. [5] <1.10> 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?

Our answer here is nearly identical. However, instead of adding for each of the eight processors, we’ll add it for each time we double the number of processors (i.e. ). This tells us that .

1. [5] <1.10> 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?

We know that is equal to the number of times we can double before hitting . We thus need to add for each of our doublings. Thus, we can say that .

1. [5] <1.10> 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?

First, let’s generalize the above equation for processors and a process that is parallelizable. That general equation is . Because we can find maximum values by setting the derivative equal to , we can find the ideal number of processors using the equation .