RESEARCH AND READING ASSIGNMENT

1) [12] [Max 1 page] Read and briefly summarize the paper titled “A New Golden Age for Computer Architecture” and answer the following questions:

a) What are the reasons of switching from CISC to RISC?

b) In which situation VLIW will fail and in which will still be available?

c) What are the main current challenges for processor architecture?

d) List the approaches improving program performance in hardware technology.

Solution:

a) First, the RISC instructions were simplified so there was no need for a microcoded interpreter.
Second, the fast memory, formerly used for the microcode interpreter of a CISC ISA, was repurposed to be a cache of RISC instructions.
Third, register allocators based on Gregory Chaitin's graph-coloring scheme made it much easier for compilers to efficiently use registers, which benefited these register-register ISAs.
Finally, Moore's Law meant there were enough transistors in the 1980s to include a full 32-bit datapath, along with instruction and data caches, in a single chip.

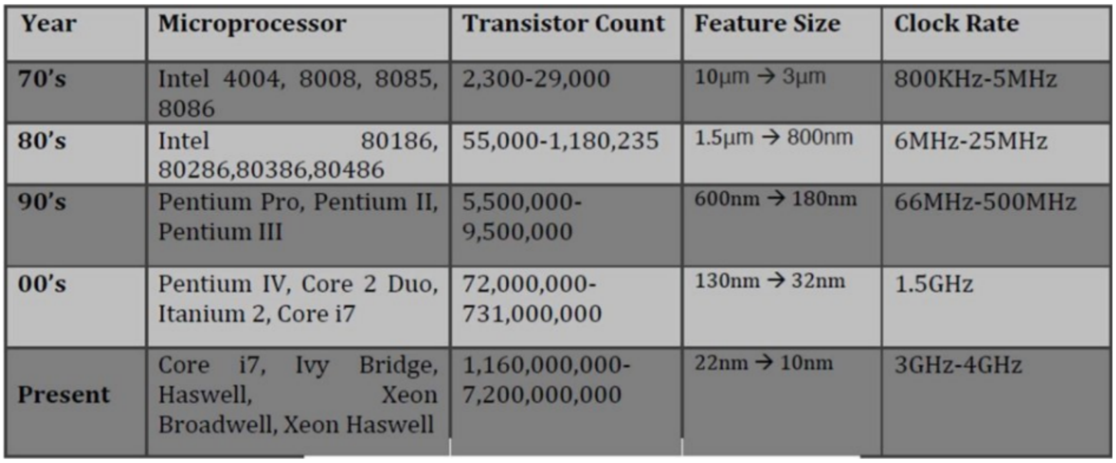
b) Although the EPIC approach worked well for highly structured floating-point programs, it struggled to achieve high performance for integer programs that had less predictable cache misses or less-predictable branches.
The good news is VLIW still matches narrower applications with small programs and simpler branches and omit caches, including digital-signal processing.

c) An era without Dennard scaling, along with reduced Moore's Law and Amdahl's Law in full effect means inefficiency limits improvement in performance to only a few percent per year and requires new architectural approaches that use the integrated-circuit capability much more efficiently.
Unfortunately, speculation introduced an unknown but significant security flaw into many processors. In particular, the Meltdown and Spectre security flaws led to new vulnerabilities that exploit vulnerabilities in the microarchitecture, allowing leakage of protected information at a high rate.

d) First, by improving the performance of modern high-level languages that are typically interpreted; and second, by building domain-specific architectures that greatly improve performance and efficiency compared to general-purpose CPUs. DSLs are another example of how to improve the hardware/software interface that enables architecture innovations like DSAs.

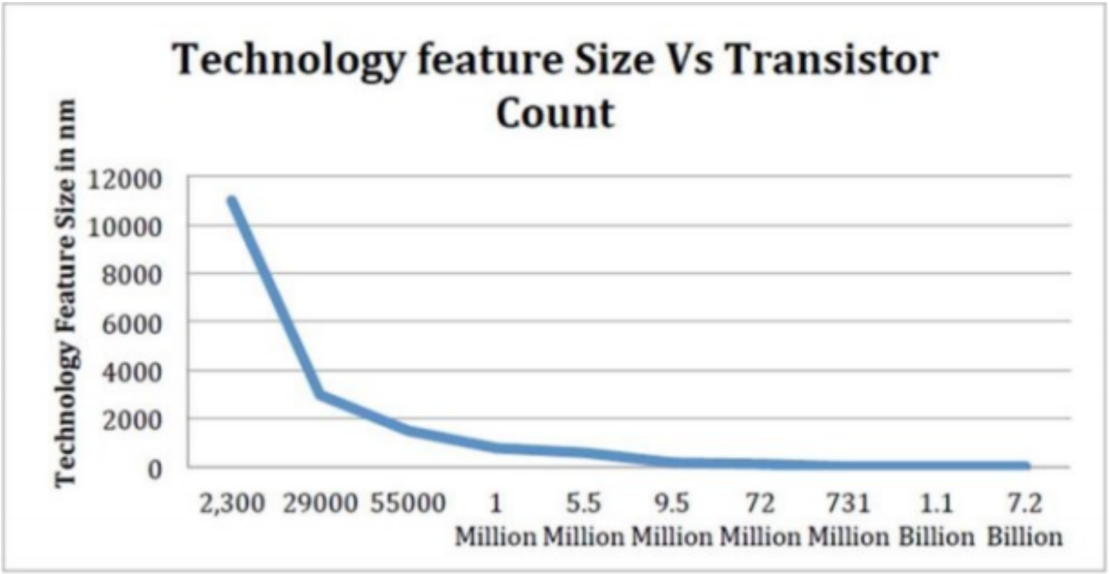
2) [12] [Max 1 page] Determine the rate of increase in transistor counts and clock frequencies in the 70's, 80's, 90's, 00's, and this decade. Also, create a plot of the number of transistors versus technology feature size using an MS Excel spreadsheet. (Hint: You can visit the Intel on-line microprocessor museum.)

Solution:



Year	Microprocessor	Transistor Count	Feature Size	Clock Rate
70's	Intel 4004, 8008, 8085, 8086	2,300-29,000	10μm → 3μm	800KHz-5MHz
80's	Intel 80186, 80286,80386,80486	55,000-1,180,235	1.5μm → 800nm	6MHz-25MHz
90's	Pentium Pro, Pentium II, Pentium III	5,500,000-9,500,000	600nm → 180nm	66MHz-500MHz
00's	Pentium IV, Core 2 Duo, Itanium 2, Core i7	72,000,000-731,000,000	130nm → 32nm	1.5GHz
Present	Core i7, Ivy Bridge, Haswell, Xeon Broadwell, Xeon Haswell	1,160,000,000-7,200,000,000	22nm → 10nm	3GHz-4GHz

The transistor count has increased from the 1970's to the present, at an average and approximate rate of 35% - 45% every year, which is in line with Moore's law.



Today's transistor count = 15,300,000,000 and clock frequency = 3.8 GHz

EXERCISES

1) [12] Table below shows relevant chip statistics that influence the cost of several processors. Explore the effect of different possible design decisions for the Processor A and answer the below questions.

Chip	Die size (mm^2)	Estimated defect rate (per cm^2)	N	Manufacturing size (nm)	Transistors (billions)
Processor A	180	0.03	12	10	7.5
Processor B	120	0.04	14	7	7.5
Processor C	200	0.04	14	7	12

a) What is the yield for Processor A?

Solution: Based on the Bose-Einstein model of yield,
$Die\ Yield = Wafer\ Yeild \times 1 / (1 + Defects\ per\ unit\ area \times Die\ area)$
We assumed that the wafer yield is 100%, no wafers are bad, i.e., Wafer yield = 1

N = 12
Estimated defect rate$_A$ = 0.03cm2
Die size = Die area = 180 mm2 = 1.8 cm2
Die Yield = 1 / (1+0.03*1.8) 12 = 0.532

Thus, Die Yield for processor A = 0.532

b) What might be the reasons that Processor A has a lower defect rate than the others?
Solution: It is fabricated in a larger technology, which is an older plant. As plants age, their process gets tuned, and the defect rate decreases.

2) [12] 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. 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?
Solution:
In 1990, processor performance of VAX-11/780 was 24MHZ, and performances increased by 52% per year. Therefore in 35 year it should be 1.52^35 times of 1990.

3) [10] A new 22-core processor runs four applications on this system, but the resource requirements are not equal. Assume the system and application characteristics as listed. 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.

Application	A	B	C	D
% Parallelizable	50	80	60	90

Solution:

$$\text{Speedup}_{\text{overall}} = \frac{1}{(1 - \text{Fraction}_{\text{enchanced}}) + \frac{\text{Fraction}_{\text{enhanced}}}{\text{Speedup}_{\text{enhanced}}}}$$

For 50% of A, which is parallelizable, the application can run on 41% of the total 22 cores rather than running on 1 core serially. Then, $\text{Speedup}_{\text{enhanced}}$ is 22*(.41) and $\text{Fraction}_{\text{enhanced}}$ is 0.5

$$\text{Speedup}_{\text{A}} = \frac{1}{0.5 + \frac{0.5}{22*(.41)}} = 1.8$$

Let us assume all the application run at same speed x before parallelization. (Other assumptions are also accepted with explanation)

$$\text{Speedup}_{\text{overall}} = \frac{4\text{x}}{3\text{x} + \frac{\text{x}}{1.8}} = 1.125$$

4) [12]One critical factor in powering a server farm is cooling. If heat is not removed from the computer efficiently, the fans will blow hot air back onto the computer, not cold air. Observe the effect of different design decisions on the necessary cooling, and thus the price, of a system.

A cooling door for a rack costs $4,000 and dissipates 14 KW (into the room; additional cost is required to get it out of the room). How many servers with a Processor P2, 1 GB 240-pin DRAM, and a single 7,200 rpm hard drive can you cool with one cooling door? Use the table below for your power calculations.

Component Type	Product	Performance	Power
Processor	P1	1.2 GHz	72-79 W peak
	P2	2 GHz	48.9 – 66 W
DRAM	MEM1	184-pin	3.7 W
	MEM2	240-pin	2.3 W
Hard disk drive	HDD1	5400 rpm	7.9 W read/seek, 2.9 W idle
	HDD2	7200 rpm	7.9 W read/seek, 4.0 W idle

Solution:
Total power dissipated by cooling rack = 14 KW
Power consumed by P2 processor = 60 W
Power consumed by 1 GB 240-pin DRAM = 2.3 W
Power consumed by a single 7200 rpm hard drive = 7.9 W
So, numbers of servers that can be cooled are,
Total power dissipated / Total power consumed = 14 KW / (60W + 2.3W + 7.9W) = 199 servers.
(183 for 66W P2, 236 for 48.9W P2)

CASE STUDIES

A) You have the following characteristics, as shown in the table below, on your company's processor for a certain benchmark, which runs at 600 MHz:

Instruction Type	Frequency (%)	Cycles
Arithmetic and logical	30	3
Load and Store	25	2
Branches	40	3
Floating Point (FP)	5	8

You are asked to consider a cheaper, lower-performance version of this processor, by removing some of the FP hardware to reduce the die size. The wafer has a diameter of 10 cm, costs $3,000, and has a defect rate of 2/(cm^2). This wafer has a 80% yield. The current chip has a die size of 12 mm^2. The new chip becomes 10 mm^2, and FP instructions will now take 12 cycles to execute.

a) [10] What are the old and new CPI (Cycles Per Instructions) and MIPS (Million Instructions Per Second) ratings running this benchmark?

b) [10] What would be the theoretical limit of the best possible overall speedup that we could ever get by only improving the FP unit, and what would be the CPI and MIPS ratings of this new processor?

c) [Bonus, 10] What are the old and new die yields? What are the old and new costs per (working) processor? Please comment on the overall effect of the proposed hardware change on the cost and the performance of the processor. (N = 4)

Solution:

a)

Old CPI = (3 * 0.30) + (2 * 0.25) + (3 * 0.40) + (8 * 0.05) = 3.0

New CPI = (3 * 0.30) + (2 * 0.25) + (3 * 0.40) + (12 * 0.05) = 3.2

MIPS = Clock rate / (CPI * 10^6)

Old MIPS = 600 MHz / (3.0 * 10^6) = 600 * 10^6 / (3.0 * 10^6) = 200

New MIPS = 600 MHz / (3.2 * 10^6) = 600 / 3.2 = 187.5

b)

Best possible way of Improving FP Unit is to make the FP Unit execution time 1 clock cycle compared to 8 clock cycles.

Old CPI = (3 * 0.30) + (2 * 0.25) + (3 * 0.40) + (8 * 0.05) = 3.0

New CPI = (3 * 0.30) + (2 * 0.25) + (3 * 0.40) + (1 * 0.05) = 2.65

Overall speedup = old CPI/new CPI = 3.0 / 2.65 = 1.13

New MIPS = 600 MHz / (2.65 * 10^6) = 600 / 2.65 = 226.42

c) We assumed that N = 4,

$$\text{Cost of Die} = \frac{Cost\ of\ wafer}{Dies\ per\ wafer\ \times Die\ yield}$$

$$\text{Dies per wafer} = \frac{\pi \times (Wafer\ diameter/2)^2}{Die\ area} - \frac{\pi \times Wafer\ diameter}{\sqrt{2 \times Die\ area}}$$

$$\text{Die per wafer}_{old} = \frac{\pi \times (10/2)^2}{0.12} - \frac{\pi \times 10}{\sqrt{2 \times 0.12}} = 590$$

$$\text{Die per wafer}_{new} = \frac{\pi \times (10/2)^2}{0.10} - \frac{\pi \times 10}{\sqrt{2 \times 0.10}} = 715$$

$$\text{Die Yield} = Wafer\ Yield \times 1/(1 + defects\ per\ unit\ area \times Die\ area)^N$$

$$\text{Die Yield}_{old} = 0.80 \times \frac{1}{(1 + 2 \times 0.12)^4} = 0.338$$

$$\text{Die Yield}_{new} = 0.80 \times \frac{1}{(1 + 2 \times 0.10)^4} = 0.386$$

$$Cost_{old} = \frac{3000}{590 \times 0.338} = \$15.04$$

$$Cost_{new} = \frac{3000}{715 \times 0.386} = \$10.87$$

B) [10] Your company produces a mobile device. To extend the battery life in the newer version of the device, you are asked to elaborate on the idea to simply reduce the processor clock speed by 25% and make no other changes. Stating your assumptions, describe whether this is a good idea or a bad idea, and why? Make sure to address both power and energy.

Solution:

$$Energy_{dynamic} = Capacitive\ load\ \times Voltage^2$$

$$Power_{dynamic} = \frac{1}{2} \times Capacitive\ load\ \times\ Voltage^2\ \times Frequency\ switched$$

Based on the above equations, forcing low CPU clock frequency reduces power consumption of your laptop, but as it makes programs to run longer time to complete the same amount of work, it can be **less energy efficient**. In other words, you may be able to accomplish less work with the same battery charge. It is not recommended to limit CPU clock frequency to the lowest for the sake of battery life.

There are some exceptions in which limiting CPU clock frequency can be beneficial:

- When you want to reduce the heat and fan noise from your laptop
- When you run a real-time high CPU utilization programs (e.g. games) and you can tolerate the quality degradation