ECE 3030: Physical Foundations of Computer Engineering

Spring 2024

Final Exam

May 1, 2024

Time: 2 hour 50 min

Instructor: Asif Khan

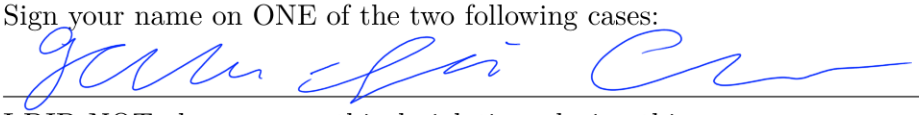
Instructions:

1. There are 15 pages in this test. Count the number of pages and notify the proctor if you are missing a page.
2. Read all the problems carefully and thoroughly before you begin working.
3. A list of constants and equations is provided on pages 14, 15.
4. You are required to answer all 5 questions. There are 100 total points. Observe the point value of each problem and allocate your time accordingly.

Q1	10 pts
Q2	25 pts
Q3	25 pts
Q4	20 pts
Q5	20 pts
Q6 (HW11)	25 pts
Q7 (HW12)	25 pts
Total	150 pts

5. Show all your work and circle/underline your final answer. For numerical answers, write the units. Write legibly. If I cannot read it, it will be considered a wrong answer. Do all work on the space provided; use scratch paper when necessary. Turn in all scratch paper, even if it did not lead to an answer.
6. Report any and all ethics violations to the instructor/proctor.

Sign your name on ONE of the two following cases:



I DID NOT observe any ethical violations during this exam:

I observed an ethical violation during this exam:

[Q1.1] **Physics of resistors:** The following table lists different properties of different interconnect metals. Rank them in terms of your preference for their use of as interconnect metals in a chip. Provide a brief explanation of your answer. [10 pts]

Material	Electron density (m^{-3})	Electron mobility (m^2/Vs)
Al	1.98×10^{29}	1.2×10^{-3}
Cu	8.5×10^{28}	4.32×10^{-3}
W	6×10^{28}	1.8×10^{-3}
AB	5×10^{28}	3×10^{-3}

Best interconnect : highest conductivity

Worst interconnect : lowest conductivity

$\sigma = nq\mu$, q is constant (charge of e^-)

$\sigma_{Al} = 1.98\times10^{29} \times 1.2\times10^{-3} \cdot q = 2.376\times10^{26}\, q$

$\sigma_{Cu} = 8.5\times10^{28} \times 4.32\times10^{-2} \cdot q = 3.672\times10^{27}\, q$

$\sigma_W = 6\times10^{28} \times 1.8\times10^{-3} \cdot q = 1.08\times10^{26}\, q$

$\sigma_{AB} = 5\times10^{28} \times 3\times10^{-3} \cdot q = 1.5\times10^{26}\, q$

$\sigma_{Cu} > \sigma_{Al} > \sigma_{AB} > \sigma_W$

Best interconnect

Cu , Al , AB , W

Worst interconnect

Q2 **MOSFETs and Delay and Power in Inverter:** If decrease the doping density N_A in a MOSFET with all the parameters unchanged, how will the following quantities change? [Total 25 pts]

[Q2.1] The MOSFET threshold voltage, V_t. [5 pts]

$$V_t = \frac{q N_A W}{C_{ox}} + 2q\psi_B$$

If N_A decrease, V_t also decrease

[Q2.2] The on-state current, I_{ON} of the MOSFET. [5 pts]

Since V_t decreases, the $I_D - V$ curve shifts to the left, which causes I_{ON} to increase.

[Q2.3] The off-state leakage current, I_{OFF} of the MOSFET. [5 pts]

Since V_t decreases, the $I_D - V$ curve shifts to the left, which causes I_{OFF} to increase.

[Q2.4] The corresponding inverter delay. [5 pts]

Since V_t decreases, effective resistance decreases, which means delay would also decrease.

[Q2.5] The active power in the corresponding inverter. The clock frequency did not change. [5 pts]

$P_{active} = C_L V_{DD}^2 f$

Since f stays the same, P_{active} stays the same.

Q3 **Scaling:** Consider that all three physical dimensions of MOSFETs (W, L, t_{ox}) are downscaled by factor of x and the power supply voltage, V_{DD} and the threshold voltage, V_t are decreased by a factor of y in every subsequent generation. In addition, the total area of the chip, A_c, also increases by 5% in every subsequent generation. [Total 25 pts]

[Q3.1] If, in every subsequent generation, the total number of transistors doubles, what is the nominal value of x? [5 pts]

$$N_n = \frac{A_c}{W_n \cdot L_n} \qquad \frac{1.05 A_c}{\frac{L_n}{x}\frac{W_n}{x}} = 1.05 x^2 \frac{A_c}{W_n \cdot L_n} = 2 \frac{A_c}{W_n \cdot L_n}$$

$$1.05 x^2 = 2 \rightarrow x = 1.38$$

[Q3.2] Consider the clock frequency and the total active/dynamic chip power of the n-th generation are f_n and $P_{chip,n}$. Find an expression of the ratio of total chip powers of two subsequent generation, $P_{chip,n+1}/P_{chip,n}$ in terms of x, y, f_n and f_{n+1}. [5 pts]

(Assume # of transistor does NOT double) $P_n = C_{L,n} V_{DD,n}^2 \cdot f_{,n} \rightarrow P_{chip,n} = C_{L,n} \cdot V_{DD,n}^2 \cdot f_n \cdot N_n$

$$N_{n+1} = \frac{A_{chip} \cdot 1.05}{\frac{1}{x^2} W_n L_n} = 1.05 x^2 N_n$$

$$P_{chip,n+1} = \frac{1}{xy^2} C_{L,n} \cdot V_{DD,n}^2 \cdot f_{n+1} \cdot 1.05 x^2 N_n$$

$$= \frac{1.05x}{y^2} f_{n+1} \cdot C_{L,n} V_{DD,n}^2 \cdot N_n$$

$$\frac{P_{chip,n+1}}{P_{chip,n}} = \boxed{\frac{1.05x}{y^2} \quad \frac{f_{n+1}}{f_n}}$$

[Space for Q3.2]

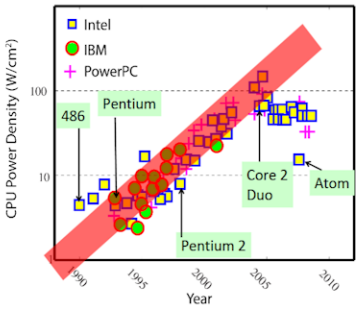
[Q3.3] What is the relation between f_{n+1} and f_n if you wanted the total active/dynamic chip power to remain the same across generations. [5 pts]

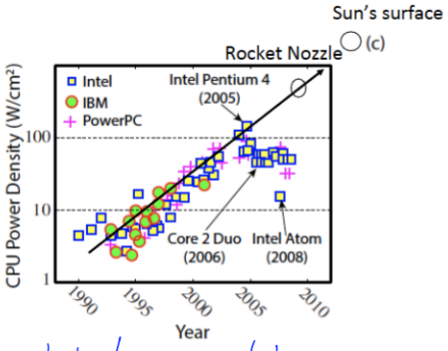
$$\frac{P_{chip,n+1}}{P_{chip,n}} = \frac{1.05X}{Y^2} \cdot \frac{f_{n+1}}{f_n} = 1$$

$$f_{n+1} = \frac{Y^2}{1.05X} f_n$$

f_n would increase by a factor of $\frac{Y^2}{1.05X}$ every subsequent generation

[Q3.4] In recent years, why is it not possible to reduce the threshold voltage of the MOSFETs significantly in subsequent generations? Use necessary figures. [Total 5 pts]





If Vt decreases, Ion increases, which would increase the power draw and power density of the chip. There is no way of effectively cool the chip with such high power density.

[Q3.5] In the face of the inability to downscale the threshold voltage significantly in successive generation, what should be the value of threshold voltage scaling factor y if you wanted to physical scaling to continue at the same rate x as you calculated in Q3.1 while keeping the total active/dynamic chip power and the clock frequency constant in each subsequent generation. [5 pts]

$$\frac{y^2}{1.05x} = 1 \quad , \quad x = 1.38$$

$$\frac{y^2}{1.05(1.38)} = 1 \quad , \quad y^2 = 1.05(1.38)$$

$$y = \sqrt{1.05(1.38)} = \boxed{1.204}$$

Q4 **Memory technologies.** [Total 20 pts]

[Q4.1] **SRAM Array:** Consider the SRAM array shown in figure 1. You want to read all the cells in row 2. What is the sequence of operation you will need to perform? Make sure that, after your prescribed operations, you keep the data in the cells you read intact. [Total 10 pts]

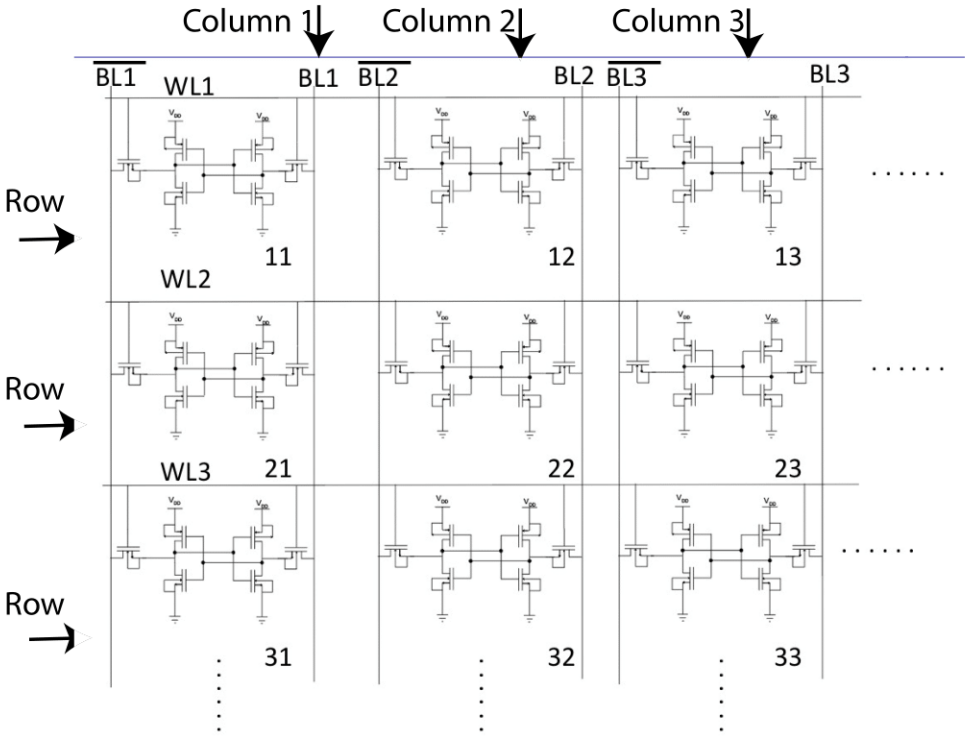


Figure 1: An SRAM array.

1) Charge BL1, $\overline{BL1}$, BL2, $\overline{BL2}$, BL3, $\overline{BL3}$ to V_{DD}

2) Set WL2 = V_{DD} and keep WL1 and WL3 at 0

3) Compare V_{BL1} and $V_{\overline{BL1}}$. If $V_{BL1} > V_{\overline{BL1}}$, read = 1, otherwise read = 0

4) Compare V_{BL2} to $V_{\overline{BL2}}$ and V_{BL3} to $V_{\overline{BL3}}$ to determine corresponding read values.

[Q4.2] **Magnetic Hard Drives:** Briefly explain why the magnetic hard drive technology is not classified as a random access memory technology by clearly stating what random memory access means. [5 pts]

Random access memory means you can access any location instantaneously, while sequential memory is not able to access data instantaneously, but instead sequentially. Since magnetic drives are classified as sequential memory and data on them cannot be accessed randomly and instantaneously, magnetic drives are not classified as RAM technology.

[Q4.3] Say you have deployed a bunch of IoT sensors to measure the distribution of temperature at different location in a forest which is then sent to a central server through internet telemetry. These sensors do not have any reliable source of power–*i.e.*, they do not have batteries; they harvest power from vibrations caused by the wind breeze, and the sensors may loose power at any moment. To store the temperature data, what kind of memory will you use: SRAM, DRAM, FLASH or magnetic hard drive? Explain your answer. [Total 5 pts]

FLASH

Want to use non-volatile (since no constant power source)
options FLASH / Magnetic

Since there is not much available power, and since FLASH uses less power than magnetic due to lack of moving parts, FLASH would be preferred over magnetic.

Q5 **Floating-gate transistor for storage:** Consider the following floating-gate transistor.

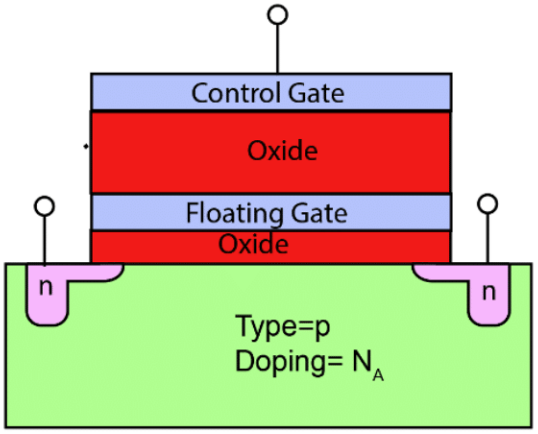


Figure 2: A floating gate transistor.

[Q5.1] How is data stored in this device. [Total 5 pts]

If there is a charge in the floating gate, the data stored is 0.

If there is no charge in the floating gate, the data stored is 1.

[Q5.2] How would you read the stored data in the device by reading the drain current? [Total 5 pts]

Read stored data by applying median voltage. If the transistor turns on, the data is 0; if transistor remains off, the data is 1.

[Q5.3] Say you want to design the device for two different application: (1) Portable device, for example, a smart phone which a consumer will use only for an average of 3 years, (2) a long term storage, for example, an external hard-drive, which a consumer will use only for an average of ~~3~~ longer years. What will you change in the device and what will be the trade-off between data retention and voltage required for writing data into the device. [Total 10 pts]

The oxide thickness would have to change.

For the portable device with shorter life span, thinner oxide layer would be preferred since data retention is shorter and requires less voltage.

For the external hard drive with longer life span, thicker oxide layer would be preferred since data retention is longer and requires more voltage.

Trade off:

Want longer data retention → thicker oxide layer → more voltage/power

Want lower voltage/power → thinner oxide layer → shorter data retention

Q6 **DRAM Array:** Consider the DRAM array shown in figure 3.

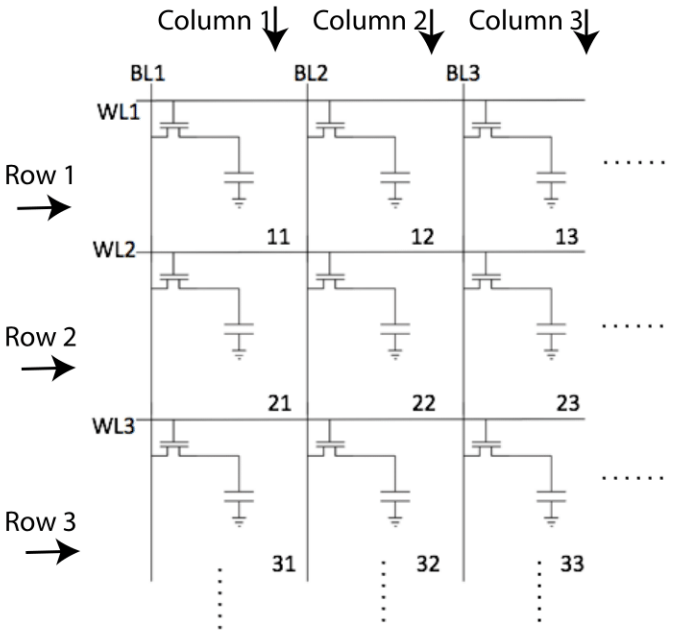


Figure 3: A DRAM array.

[6.1] Say you want to read all the cells in row 2. What is the sequence of operation you will need to perform? Make sure that, after your prescribed operations, you keep the data in the cells you read intact. [Total 5 pts]

1) Set BL1, BL2, and BL3 to $\frac{V_{DD}}{2}$

2) Set WL2 to V_{DD} while keeping WL1 and WL3 to 0

3) Read V_{BL} to determine read value

4) Apply read value back to the BL to keep data intact.

[6.2] Explain in short why it necessary to have cache memories in today's microprocessor technology. [Total 10 pts]

Cache memory is necessary for modern-day microprocessors because of the relatively slow speed of other memory types, such as DRAM, FLASH, and magnetic drives. If lower-speed memory were used, the time waiting to fetch memory would result in wasted clock cycles. Cache memory is faster and has quicker fetch times compared to other memory types, which means more clock cycles can be used for processing data.

[6.3] Compare and contrast the different memory technologies (SRAM, DRAM, Flash and magnetic disk) used in a memory hierarchy with respect to density, speed, and volatility [Total 10 pts]

Density:

Highest → Lowest

Magnetic disk → FLASH → DRAM → SRAM

Speed

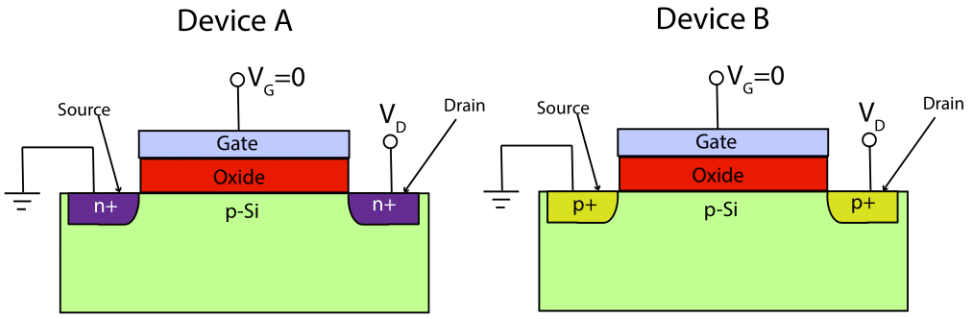
Highest → Lowest

SRAM → DRAM → FLASH → Magnetic disk

Volatility:

Volatile	Non-volatile
SRAM	FLASH
DRAM	Magnetic disk

Q7 **MOSFETS:** Consider the two devices shown in the following figure.



[7.1] Device A is the MOSFET structure that we discussed in the class and in which the source and drain are n^+-type (heavily doped n-type). On the other hand, in device B source and drain are p^+-type (heavily doped p-type). Based on what we discussed in class, do you think device B will behave like a switch–*i.e.* will device B be able to block current from flowing between the drain and the source terminal when the gate voltage V_G is zero (or less than the threshold voltage)? Provide justification for your answer. [Total 10 pts]

No, Device B will not act as a switch.

Since the semiconductor at the source/drain is the same type as the body of the transistor, current will always flow through the transistor.

[7.2] **Dynamic Voltage and Frequency Scaling (DVFS):** Consider the logic blocks shown in figure 4. Logic block 2 and 3 receives input from logic block 1 at the same time. Logic block 5 needs inputs from blocks 3 and 4 for generate the final output. Logic block 3 receives input from logic block 2. Logic block 2, 3 and 4 requires 50, 10, and 30 cycles, respectively, to generate the respective output.

Based on what you have learned in class, how will you apply DVFS in these system? For which logic block(s), will you increase or decrease the power supply voltage and by how much? How much energy will you save in your prescribed process? Explain your answer. Make necessary and simplest possible reasonable assumptions. [Total 10 pts]

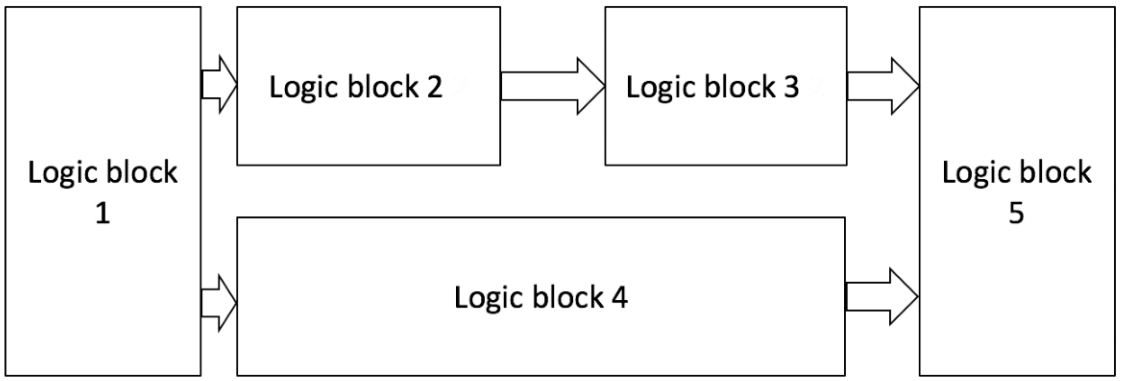


Figure 4: Dynamic Voltage and Frequency Scaling (DVFS).

Total cycles from block 2 & 3 : 50 + 10 = 60

Logic Block 4 should also match same cycle

Block 4 requires 30 cycles, so decrease frequency

Half frequency of Block 4 → also decrease V_{DD} by half

$P_{DVFS} = C_L(\frac{1}{2}V_{DD})^2 \frac{1}{2}f = \frac{1}{8}C_L V_{DD}^2 f$

$E_{DVFS} = \frac{P}{f}$ $E_{before} = C_L V_{DD}^2 f \cdot \frac{30}{f}$ $E_{after} = C_L V_{DD}^2 f \cdot \frac{1}{8} \cdot \frac{60}{f}$

$E_{saved} = 30 C_L V_{DD}^2 - \frac{60}{8} C_L V_{DD}^2 = \boxed{22.5 C_L V_{DD}^2}$

[7.3] **Energy, Power, and Heat** As we know in Moore's law, the transistor density of devices increases at an exponential rate. However, the power supply voltage, V_{DD}, did not decrease until after 1999. Provide a few reasons why modern semiconductors face these limitations. [Total 5 pts]

Leakage current: leakage current increases when transistors size decreases because the oxide thickness decreses. The higher power supply voltage will increase the leakage current, wasting energy and produce more heat.

Heat: increasing the power supply voltage will increase the energy consumption, which would increase power consumption and increase power density. The chip will not run efficiently, which defeats the purpose of making chips more efficient.