# PERFORMANCE, POWER, ENERGY

Mahdi Nazm Bojnordi

Assistant Professor

School of Computing

University of Utah

# Recall: Processor Performance

- Clock cycle time (CT = 1/clock frequency)
  - Influenced by technology and pipeline
- Cycles per instruction (CPI)
  - Influenced by architecture
  - IPC may be used instead (IPC = 1/CPI)
- Instruction count (IC)
  - Influenced by ISA and compiler
- CPU time = IC x CPI x CT

# Example: Clock Cycle Time

- I execute a scientific program with 1B instructions on my laptop. I observe an average cycle per instruction (CPI) of 4.5 for each run. Compute the CPU time if the clock frequency is 2GHz.

# Example: Clock Cycle Time

- I execute a scientific program with 1B instructions on my laptop. I observe an average cycle per instruction (CPI) of 4.5 for each run. Compute the CPU time if the clock frequency is 2GHz.

- CPU time = IC x CPI x CT
- $= 1 \times 10^9 \times 4.5 \times 0.5 \times 10^{-9}$
- = 2.25 seconds

# Example: Clock Cycle Time

- I execute a scientific program with 1B instructions on my laptop. I observe an average cycle per instruction (CPI) of 4.5 for each run. Compute the CPU time after overclocking to 3.2GHz.

# Example: Clock Cycle Time

- I execute a scientific program with 1B instructions on my laptop. I observe an average cycle per instruction (CPI) of 4.5 for each run. Compute the CPU time after overclocking to 3.2GHz.

- CPU time = IC x CPI x CT
- $=1 \times 10^9 \times 4.5 \times 0.3125 \times 10^{-9}$
- = 1.40625 seconds

# Example: Cycles Per Instruction

- Computer A: Cycle Time = 250ps, CPI = 2.0
- Computer B: Cycle Time = 500ps, CPI = 1.2
- Same ISA and same program
- Which is faster, and by how much?

# Example: Cycles Per Instruction

- Computer A: Cycle Time = 250ps, CPI = 2.0
- Computer B: Cycle Time = 500ps, CPI = 1.2
- Same ISA and same program
- Which is faster, and by how much?

$$\text{CPU Time}_A = \text{Instruction Count} \times \text{CPI}_A \times \text{Cycle Time}_A$$

$$= I \times 2.0 \times 250\text{ps} = I \times 500\text{ps}$$ ← A is faster…

$$\text{CPU Time}_B = \text{Instruction Count} \times \text{CPI}_B \times \text{Cycle Time}_B$$

$$= I \times 1.2 \times 500\text{ps} = I \times 600\text{ps}$$

$$\frac{\text{CPU Time}_B}{\text{CPU Time}_A} = \frac{I \times 600\text{ps}}{I \times 500\text{ps}} = 1.2$$ ← …by this much

# Example: Instruction Count

- There exist two algorithms for a scientific problem. Program A implements Algorithm A using 10B instructions. But, Program B needs only 2B instructions for Algorithm B. Compute the CPU times for an average IPC of 0.25 on a 4GHz processor.

# Example: Instruction Count

- There exist two algorithms for a scientific problem. Program A implements Algorithm A using 10B instructions. But, Program B needs only 2B instructions for Algorithm B. Compute the CPU times for an average IPC of 0.25 on a 4GHz processor.
- Program A: CPU time = $10 \times 10^9 \times 4 \times 0.25 \times 10^{-9}$
- = 10 seconds
- Program B: CPU time = $2 \times 10^9 \times 4 \times 0.25 \times 10^{-9}$
- = 2 seconds

# Measuring Performance

- What program to use for measuring performance?
- Benchmarks Suites
  - A set of representative programs that are likely relevant to the user
  - Examples:
    - SPEC CPU 2006: CPU-oriented programs (for desktops)
    - SPECweb: throughput-oriented (for servers)
    - EEMBC: embedded processors/workloads

# SPEC CPU Benchmark

- Programs used to measure performance
  - Supposedly typical of actual workload
- Standard Performance Evaluation Corp (SPEC)
  - Develops benchmarks for CPU, I/O, Web, …
- SPEC CPU2006
  - Elapsed time to execute a selection of programs
    - Negligible I/O, so focuses on CPU performance
  - Normalize relative to reference machine
  - Summarize as geometric mean of performance ratios
    - CINT2006 (integer) and CFP2006 (floating-point)

$$\sqrt[n]{\prod_{i=1}^{n} Execution\ time\ ratio_i}$$

# Improving Performance

- Consider an employee who is given a fix budget of $500 to enhance the performance their laptop. There exist two options for system upgrade: (a) make CPU 2x faster and (b) make memory 1.5x faster. Which one is upgrade option is better?

# Amdahl's Law

- The law of diminishing returns

$$\text{Execution time after improvement} = \frac{\text{Execution time affected by improvement}}{\text{Amount of improvement}} + \text{Execution time unaffected}$$

# Amdahl's Law

- The law of diminishing returns

$$\text{Execution time after improvement} = \frac{\text{Execution time affected by improvement}}{\text{Amount of improvement}} + \text{Execution time unaffected}$$

$$\text{Execution time}_{\text{new}} = \text{Execution time}_{\text{old}} \times \left( (1 - \text{Fraction}_{\text{enhanced}}) + \frac{\text{Fraction}_{\text{enhanced}}}{\text{Speedup}_{\text{enhanced}}} \right)$$

# Amdahl's Law

- The law of diminishing returns

$$\text{Execution time after improvement} = \frac{\text{Execution time affected by improvement}}{\text{Amount of improvement}} + \text{Execution time unaffected}$$

$$\text{Execution time}_{\text{new}} = \text{Execution time}_{\text{old}} \times \left( (1 - \text{Fraction}_{\text{enhanced}}) + \frac{\text{Fraction}_{\text{enhanced}}}{\text{Speedup}_{\text{enhanced}}} \right)$$

$$\text{Speedup}_{\text{overall}} = \frac{\text{Execution time}_{\text{old}}}{\text{Execution time}_{\text{new}}} = \frac{1}{(1 - \text{Fraction}_{\text{enhanced}}) + \frac{\text{Fraction}_{\text{enhanced}}}{\text{Speedup}_{\text{enhanced}}}}$$

# Improving Performance

- Consider an employee who is given a fix budget of $500 to enhance the performance their laptop. There exist two options for system upgrade: (a) make CPU 2x faster and (b) make memory 1.5x faster. Which one is upgrade option is better?
- Scenario 1: 20% CPU and 80% Memory
  - (a): speedup=1.11x (b): speedup=1.36x

# Improving Performance

- Consider an employee who is given a fix budget of $500 to enhance the performance their laptop. There exist two options for system upgrade: (a) make CPU 2x faster and (b) make memory 1.5x faster. Which one is upgrade option is better?
- Scenario 1: 20% CPU and 80% Memory
  - (a): speedup=1.11x (b): speedup=1.36x
- Scenario 2: 70% CPU and 30% Memory
  - (a): speedup=1.53x (b): speedup=1.11x

# Example Problem

- Our new processor is 10x faster on computation than the original processor. Assuming that the original processor is busy with computation 40% of the time and is waiting for IO 60% of the time, what is the overall speedup?

# Example Problem

- Our new processor is 10x faster on computation than the original processor. Assuming that the original processor is busy with computation 40% of the time and is waiting for IO 60% of the time, what is the overall speedup?

$f=0.4 \quad s=10$

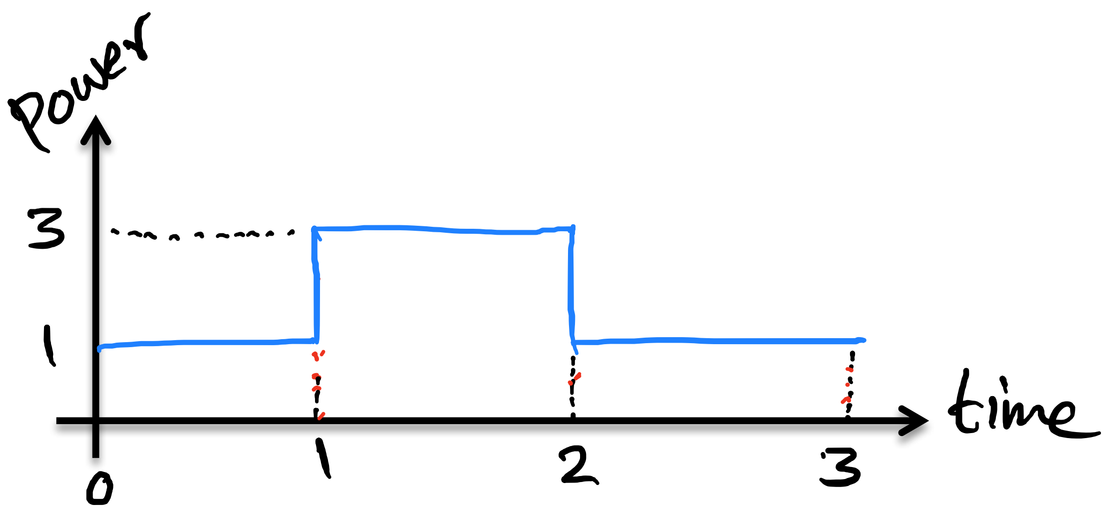
Speedup = 1 / (0.6 + 0.4/10) = 1/0.64 = 1.5625

# Power and Energy

- Power = Voltage x Current (P = VI)
  - Instantaneous rate of energy transfer (Watt)
- Energy = Power x Time (E = PT)
  - The cost of performing a task (Joule)

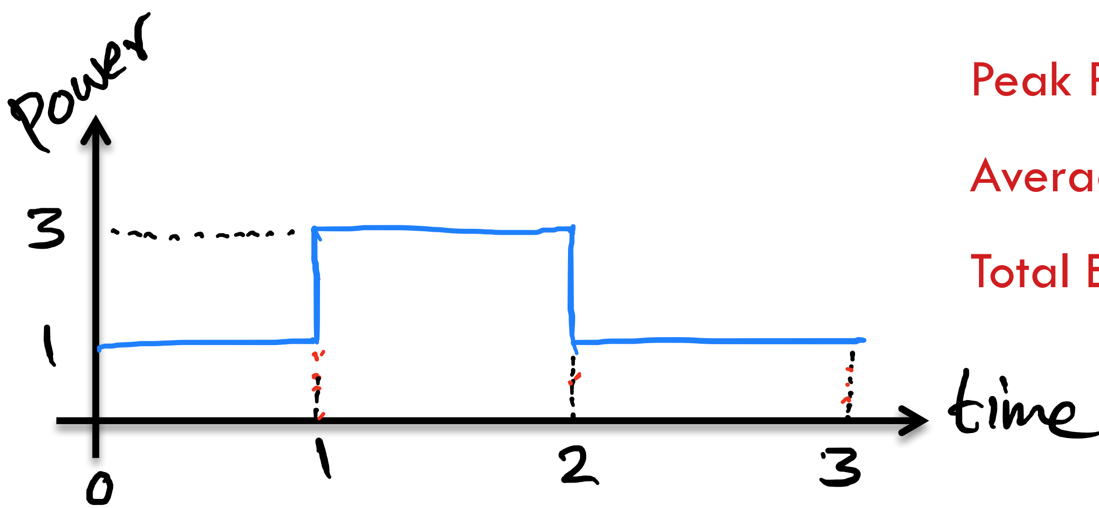
# Power and Energy

- Power = Voltage x Current (P = VI)
  - Instantaneous rate of energy transfer (Watt)
- Energy = Power x Time (E = PT)
  - The cost of performing a task (Joule)



# Power and Energy

- Power = Voltage x Current (P = VI)
  - Instantaneous rate of energy transfer (Watt)
- Energy = Power x Time (E = PT)
  - The cost of performing a task (Joule)




Peak Power = 3W

Average Power = 1.66W

Total Energy = 5J

# CPU Power and Energy

- All consumed energy is converted to heat
  - CPU power is the rate of heat generation
  - Excessive peak power may result in burning the chip
- Static and dynamic energy components
  - Energy = ($Power_{Static}$ + $Power_{Dynamic}$) x Time

# Example: Power and Energy

- Consider using Zoom for a 50-minute IVC meeting on your laptop that dissipates 75W dynamic power. Assume that your laptop dissipates 15W static power. Compute the total energy consumed for the meeting?

# Example: Power and Energy

- Consider using Zoom for a 50-minute IVC meeting on your laptop that dissipates 75W dynamic power. Assume that your laptop dissipates 15W static power. Compute the total energy consumed for the meeting?

- Energy = $(Power_{Static} + Power_{Dynamic}) \times Time$
- $= (15 + 75) \times 3000 = 270kJ$

# Example: Power and Energy

- Consider using Zoom for a 50-minute IVC meeting on your laptop that dissipates 75W dynamic power. Assume that your laptop dissipates 15W static power. Compute the total energy consumed for the meeting? Assuming an energy rate of 20 ¢/kWh, what's the cost of each meeting?
- Energy = $(Power_{Static} + Power_{Dynamic})$ x Time
- = (15 + 75) x 3000 = 270kJ

# Example: Power and Energy

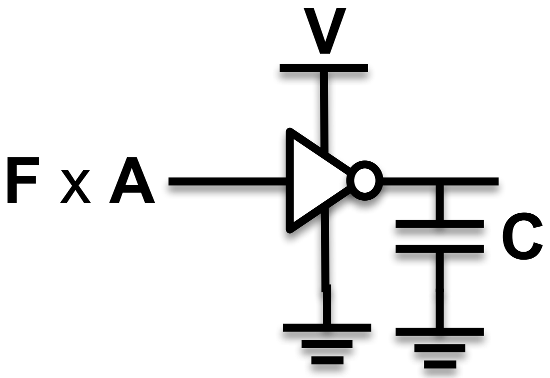
- Consider using Zoom for a 50-minute IVC meeting on your laptop that dissipates 75W dynamic power. Assume that your laptop dissipates 15W static power. Compute the total energy consumed for the meeting? Assuming an energy rate of 20 ¢/kWh, what's the cost of each meeting?
- Energy = $(Power_{Static} + Power_{Dynamic})$ x Time
- = (15 + 75) x 3000 = 270kJ
- 1kWh = 3,600kJ → Cost = 1.5¢

# CPU Power and Energy

- All consumed energy is converted to heat
  - CPU power is the rate of heat generation
  - Excessive peak power may result in burning the chip
- Static and dynamic energy components
    - $\text{Energy} = (\text{Power}_{\text{Static}} + \text{Power}_{\text{Dynamic}}) \times \text{Time}$
  - How to compute for CPU?
    - $\text{Power}_{\text{Static}} = \text{Voltage} \times \text{Current}_{\text{Static}}$
    - $\text{Power}_{\text{Dynamic}} = \text{Capacitance} \times \text{Voltage}^2 \times (\text{Activity} \times \text{Frequency})$

# Power Reduction Techniques

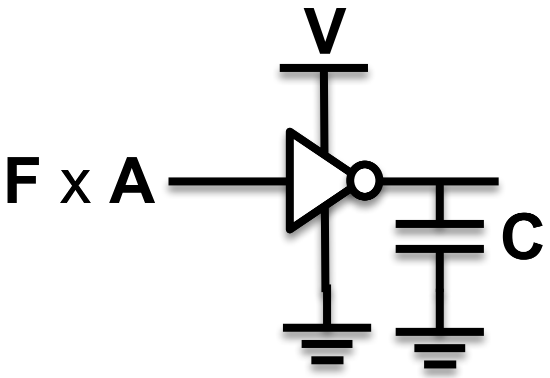
- Reducing capacitance (C)
- Reducing voltage (V)
- Reducing frequency (F)




- .

# Power Reduction Techniques

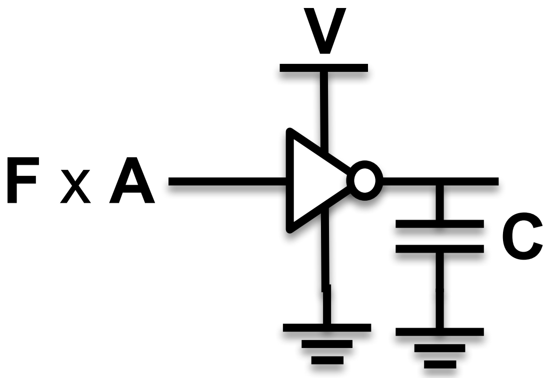
- Reducing capacitance (C)
  - Requires changes to physical layout and technology
- Reducing voltage (V)
- Reducing frequency (F)




  - .

# Power Reduction Techniques

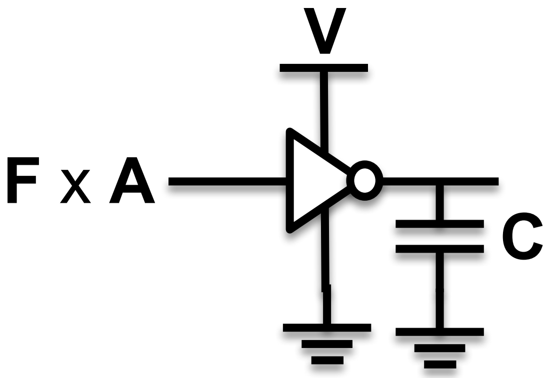
- Reducing capacitance (C)
  - Requires changes to physical layout and technology
- Reducing voltage (V)
  - Negative effect on frequency
  - Opportunistically power gating (wakeup time)
  - Dynamic voltage and frequency scaling
- Reducing frequency (F)




  - .

# Power Reduction Techniques

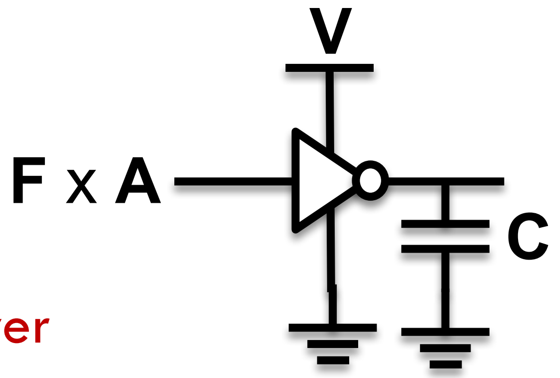
- Reducing capacitance (C)
  - Requires changes to physical layout and technology
- Reducing voltage (V)
  - Negative effect on frequency
  - Opportunistically power gating (wakeup time)
  - Dynamic voltage and frequency scaling
- Reducing frequency (F)
  - Negative effect on CPU time
  - Clock gating in unused resources




  - .

# Power Reduction Techniques

- Reducing capacitance (C)
  - Requires changes to physical layout and technology
- Reducing voltage (V)
  - Negative effect on frequency
  - Opportunistically power gating (wakeup time)
  - Dynamic voltage and frequency scaling
- Reducing frequency (F)
  - Negative effect on CPU time
  - Clock gating in unused resources
- Points to note
  - Utilization directly effects dynamic power
  - Lowering power does NOT mean lowering energy



# Example: Frequency Scaling

- Consider a processor consuming 80W dynamic power. By only reducing the frequency from 4GHz to 2GHz, what will be the new dynamic power?

# Example: Frequency Scaling

- Consider a processor consuming 80W dynamic power. By only reducing the frequency from 4GHz to 2GHz, what will be the new dynamic power?

- $Power_{Dynamic}$ = Capacitance x $Voltage^2$ x (Activity x Frequency)
- @4GHz $Power_{Dynamic}$ = 80W
- @2GHz $Power_{Dynamic}$ = 40W