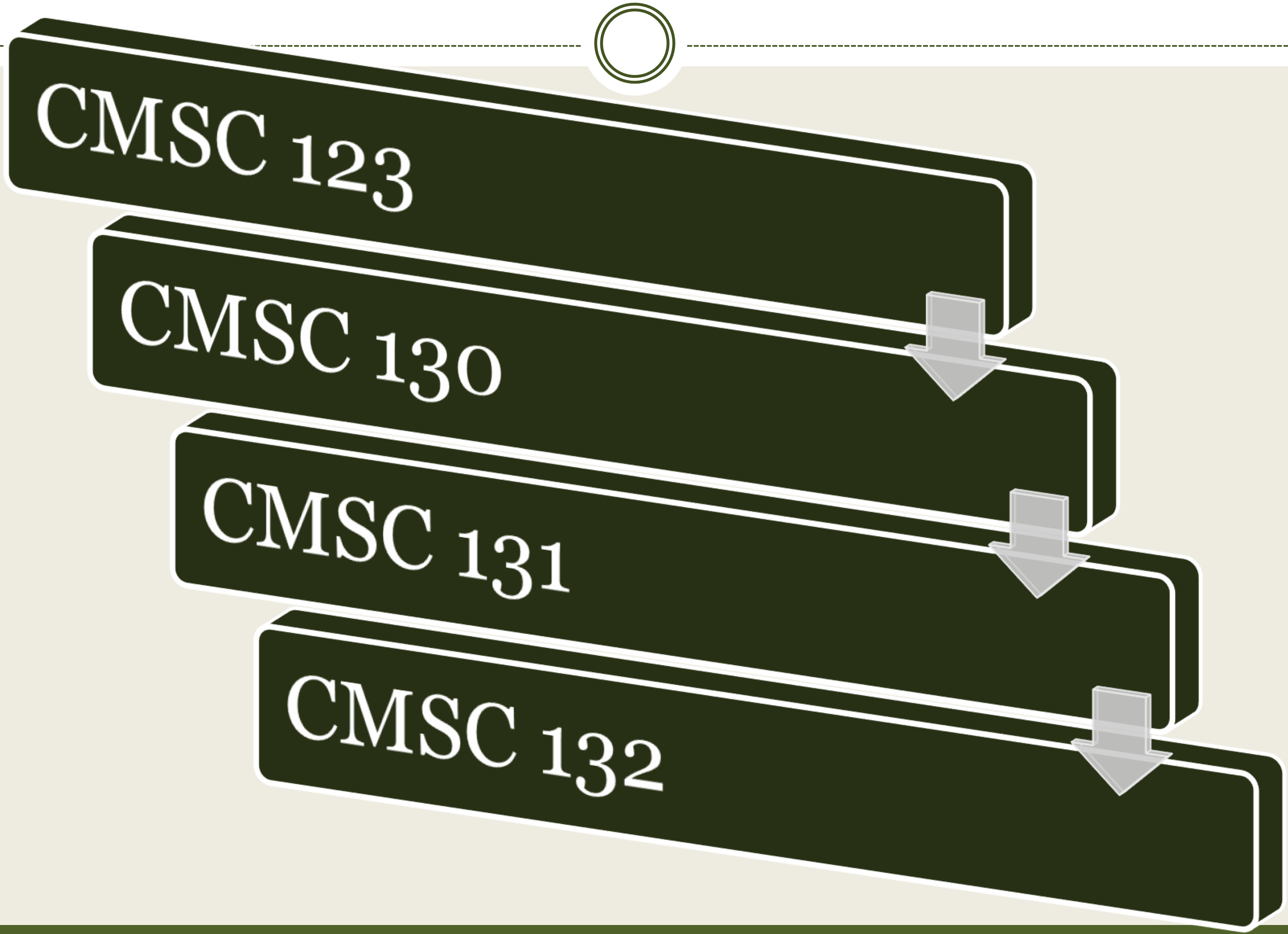
CMSC 132: Computer Architecture

Asst. Prof. Reginald Neil C. Recario
rncrecario@gmail.com
Institute of Computer Science
University of the Philippines Los Baños

Why study CMSC 132?

CMSC 132 Ladder*



CMSC 123

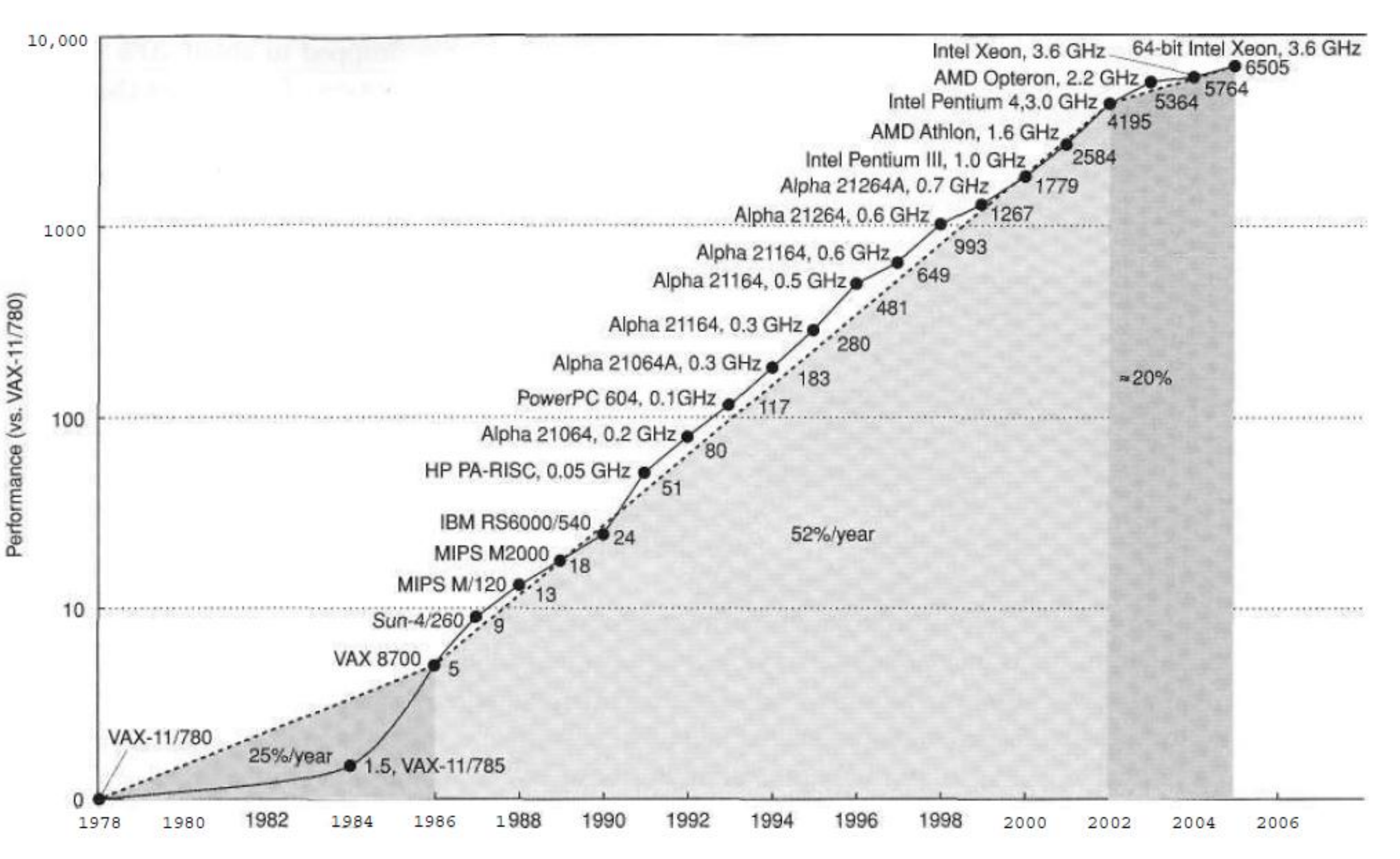
CMSC 130

CMSC 131

CMSC 132

**The arrows do not necessarily depict dependence being a prerequisite of a course*

Growth of processor performance



Reference: Computer Architecture: A Quantitative Approach. p3

RISC

- Reduced Instruction Set Computer
- RISC-based machines focused the attention of designers on the two critical performance techniques:
    - Instruction level parallelism
    - Use of caches
- Forced other architectures to keep up or disappear.

RISC

- Processor performance dropping in 2002 due to
  - Maximum power dissipation of air-cooled chips
  - Little instruction-level parallelism left to exploit efficiently
  - Almost unchanged memory latency
- Lead to high performance through multiple processors rather than uniprocessors.

History of Computers…

- 1960s – mainframes
- 1970s – minicomputers
- 1980s – desktop computers
- 1990s – emergence of Internet and WWW, first successful handheld digital computing devices and high performance digital consumer electronics

Desktop Computer

- Has the largest market
- Optimize *price-performance*
  - Compute performance and graphics performance
- Well-characterized in terms of applications and benchmarking

Servers

- Role of servers grew to provide larger-scale and more reliable file and computing services
- Important characteristics:
  - Dependability
  - Scalability
  - Efficient Throughput

Embedded Computers

- Fastest growing portion of the computer market
- Range from everyday machines
- Have the widest range or processing power and cost
  - Performance requirement is real-time execution
- A real-time performance is when a segment of the application has an absolute maximum execution time.

Embedded Computers

- Soft real-time
  - arise when it is possible to occasionally miss the time constraint on an event, as long as not too many are missed
- Need of embedded computers
  - Minimize memory
  - Minimize power

Defining Computer Architecture

- When defining a computer architecture, a computer designer is tasked to determine what attributes are important for a new computer
- Must design a computer to maximize performance while staying within cost, power and availability constraints

Computer Architecture

- In the past, computer architecture often referred only to instruction set design
- According to Wikipedia (2013), computer architecture is a set of disciplines that describes a computer system by specifying its parts and their relations.

Instruction Set Architecture (ISA)

- ISA refer to the actual programmer visible instruction set
  - Serves as the boundary between the software and hardware

Instruction Set Architecture (ISA)

- Seven dimensions of ISA
  - Class of ISA
  - Memory addressing
  - Addressing modes
  - Types and sizes of operands
  - Operations
  - Control flow instruction
  - Encoding an ISA

Class of ISA

- Nearly all ISAs are classified as general-purpose register architectures

Memory addressing

- Virtually all desktop and server computers, including the 80x86 and MIPS, use byte addressing to access memory operands.

Addressing modes

- Specify the address of a memory object.

Types and sizes of operands

- Support for operand sizes of 8-bit (ASCII character), 16-bit (Unicode character or half word), 32-bit (integer or word), 64-bit (double word or long integer), and IEEE 754 floating point in 32-bit (single precision) and 64-bit (double precision).

Operations

- The general categories of operations are data transfer, arithmetic logical, control (discussed next), and floating point.

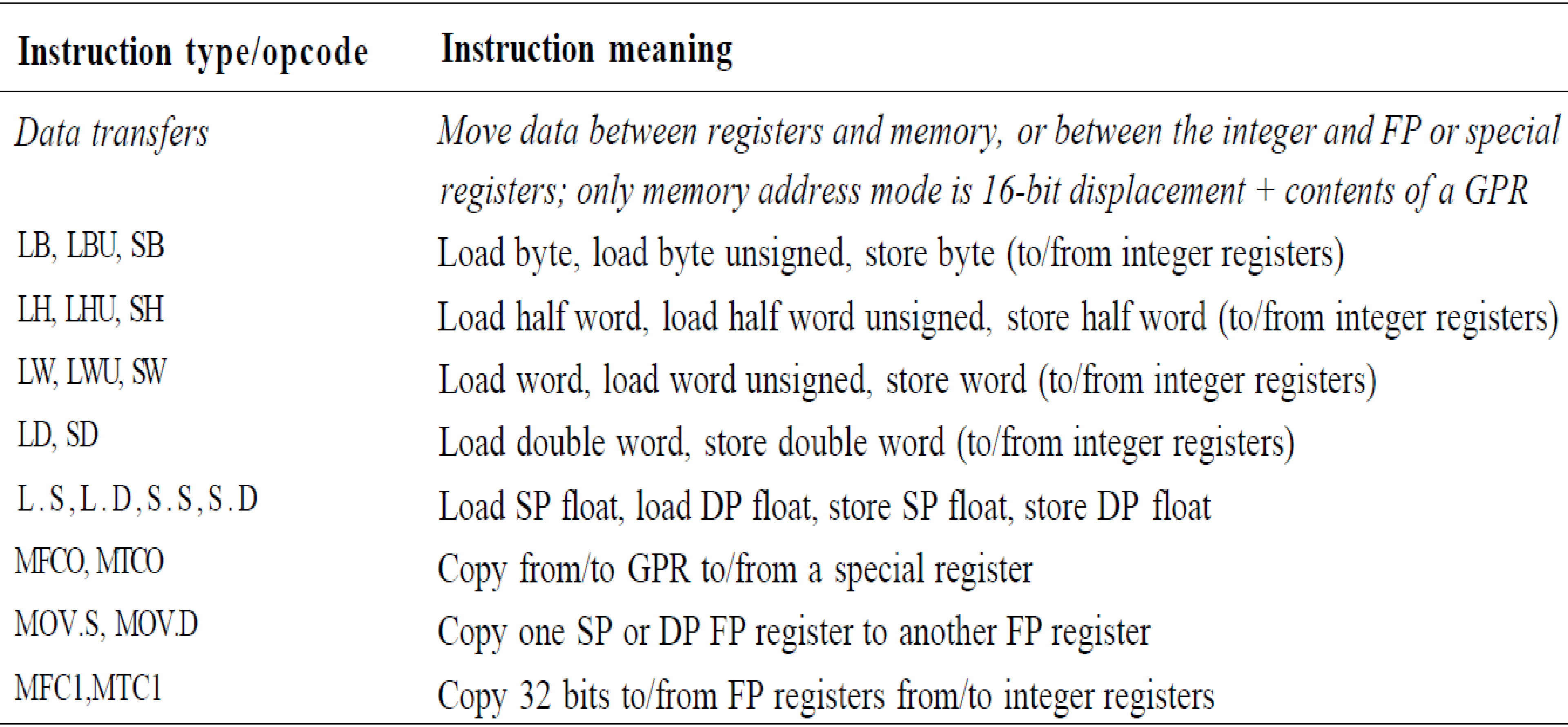
Control flow instruction

- Support for conditional branches, unconditional jumps, procedure calls, and returns.

Encoding an ISA

- Two basic choices on encoding: fixed length and variable length

Sample Instruction Set



Instruction type/opcode	Instruction meaning
Data transfers	*Move data between registers and memory, or between the integer and FP or special registers; only memory address mode is 16-bit displacement + contents of a GPR*
LB, LBU, SB	Load byte, load byte unsigned, store byte (to/from integer registers)
LH, LHU, SH	Load half word, load half word unsigned, store half word (to/from integer registers)
LW, LWU, SW	Load word, load word unsigned, store word (to/from integer registers)
LD, SD	Load double word, store double word (to/from integer registers)
L.S,L.D,S.S,S.D	Load SP float, load DP float, store SP float, store DP float
MFCO, MTCO	Copy from/to GPR to/from a special register
MOV.S, MOV.D	Copy one SP or DP FP register to another FP register
MFC1,MTC1	Copy 32 bits to/from FP registers from/to integer registers

Reference: Computer Architecture: A Quantitative Approach. p11

Trends in Technology

- Four technologies that are critical in implementation
  - Integrated circuit logic technology
  - Semiconductor DRAM
  - Magnetic disk technology
  - Network technology

Integrated Circuit Logic Technology

- The combined effect is a growth rate in transistor count on a chip of about 40% to 55% per year.

Semiconductor DRAM

- Capacity increases by about 40% per year

Magnetic Disk Technology

- Disks are still 50-100 times cheaper per bit than DRAM

Network Technology

- Network performance depends both on the performance of switches and on the performance of the transmission system

Moore's Law

- States that the number of transistors or ICs double in approximately two years
- Intel co-founder Gordon E. Moore

Bandwidth vs Latency

- Bandwidth or throughput is the total amount of work done in a given time, such as megabytes per second for a disk transfer

Bandwidth vs Latency

- Latency or response time is the time between the start and the completion of an event, such as milliseconds for a disk access

Bandwidth vs Latency

- Performance is the primary differentiator for microprocessors and networks
- Capacity is generally more important than performance for memory and disks
  - Greater bandwidth than latency

Bandwidth vs Latency

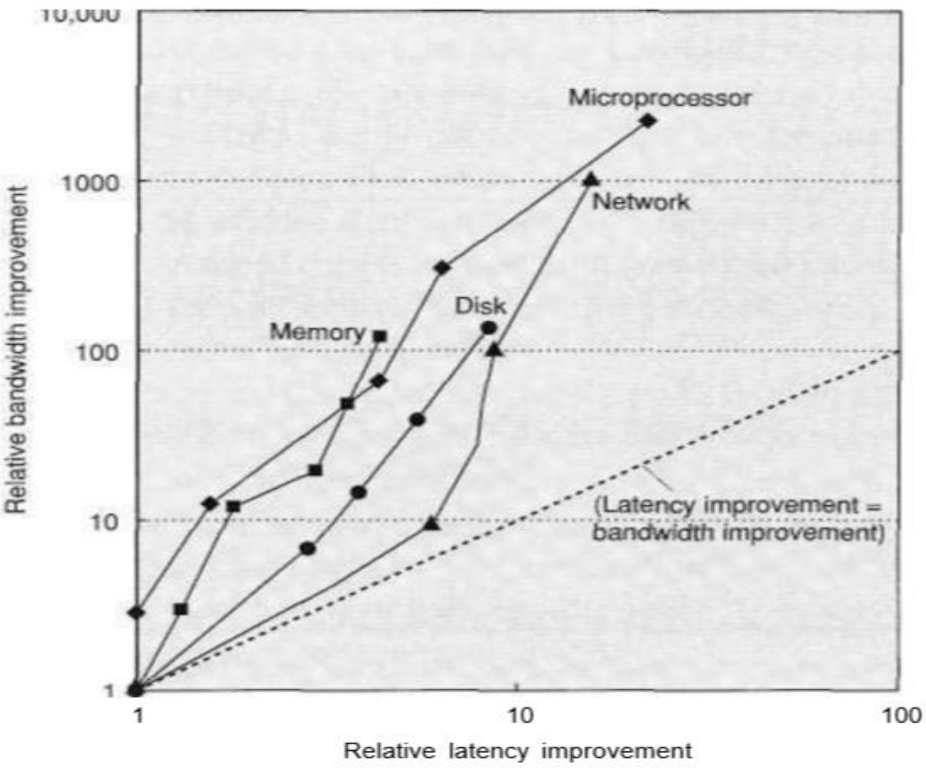


Figure 1.8 Log-log plot of bandwidth and latency milestones from Figure 1.9 relative to the first milestone. Note that latency improved about 10X while bandwidth improved about 100Xto 1000X.From Patterson [2004].

Reference: Computer Architecture: A Quantitative Approach. p15

Bandwidth vs Latency

- A simple rule of thumb is that bandwidth grows by at least the square of the improvement in latency.

Scaling of Transistor Performance and Wires

- Integrated circuit processes are characterized by the feature size, which is the minimum size of a transistor or a wire in either the x or y dimension.

Scaling of Transistor Performance and Wires

- 10 microns in 1971
- 0.09 microns (90 nanometers) in 2006
- And still getting smaller!

Scaling of Transistor Performance and Wires

- Transistor count per square millimeter of silicon is determined by the surface area of a transistor.
- The density of transistors increases quadratically with a linear decrease in feature size.

Scaling of Transistor Performance and Wires

- Increase in transistor performance is complex though.
  - Reduction in x or y dimension requires correction of operating voltage to maintain correct operation and reliability of the transistors
  - Complex relationship between size and performance

Scaling of Transistor Performance and Wires

- Although transistors generally improve in performance with decreased feature size, wires in an integrated circuit do not.

Scaling of Transistor Performance and Wires

- The signal delay for a wire increases in proportion to the product of its resistance and capacitance.
  - Some improvements like introduction of copper for wire delay

Trends in Power in Integrated Circuits

- Power also provides challenges as devices are scaled
  - Must be distributed within the chip
  - Power is dissipated as heat and must be removed

Trends in Power in Integrated Circuits

- For CMOS chips, the traditional dominant energy consumption has been in switching transistors, also called dynamic power.
- The power required per transistor is proportional to the product of the load capacitance of the transistor, the square of the voltage, and the frequency of switching, with watts being the unit

Trends in Power in Integrated Circuits

- Formula:

Power$_{\text{dynamic}}$ = ½ * Capacitive load * Voltage^2 * Frequency switched

Trends in Power in Integrated Circuits

- Mobile devices care about battery life more than power, so energy is the proper metric, measured in joules

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

Trends in Power in Integrated Circuits

- Power and energy can be reduced by reducing voltage.
- The capacitive load is a function of the number of transistors connected to an output and the technology, which determines the capacitance of the wires and the transistors.
- For a fixed task, slowing clock rate reduces power, but not energy.

Example:

- Some microprocessors today are designed to have adjustable voltage, so that a 15% reduction in voltage may result in a 15% reduction in frequency. What would be the impact on dynamic power?

Example:

- Answer :
- Since the capacitance is unchanged, the answer is the ratios of the voltages and frequencies:

Example:

$$\frac{\text{Power}_{\text{new}}}{\text{Power}_{\text{old}}} = \frac{(\text{Voltage} \times 0.85)^2 \times (\text{Frequency switched} \times 0.85)}{\text{Voltage}^2 \times \text{Frequency switched}} = 0.85^3 = 0.61$$

There by reducing power to about 60% of the original.

Quiz

- Assuming that the capacitance and frequency are kept constant, will doubling the amount of voltage increase or decrease the power of a newly developed microprocessor?
- Write **W** if it will increase, else write **O** and include your solution.

First Long Examination

- 1st LE on Mar 5 (Thu) class hours.
- Please bring calculators. Calculators are not required though.
- Phones and other gadgets will not be allowed as a substitute for calculators.
- You are not allowed to store formulas on your calculators.
- Type of exam: DNM (does not matter)

Trends in Power in Integrated Circuits

- The increase in the number of transistors switching, and the frequency with which they switch, dominates the decrease in load capacitance and voltage.
- This still leads to an overall growth in power consumption and energy.

Trends in Power in Integrated Circuits

- First microprocessors consumed 10th of a watt.
- 3.2 GHz Pentium 4 Extreme Edition consumes 135 watts.

Trends in Power in Integrated Circuits

- Problem? Heat dissipation
- We reach the limits of what can be cooled by air.
- Possible improvements:
  - Several Intel microprocessors have temperature diodes to reduce activity automatically if the chip gets too hot.

Trends in Power in Integrated Circuits

- Possible improvements:
  - Reduction in voltage and clock frequency or instruction issue rate
- Challenges
  - Distributing the power
  - Removing the heat
  - Preventing hot spots

Trends in Power in Integrated Circuits

- Power has become a limitation to transistors.
- Most microprocessors today turn off the clock of inactive modules to save energy and dynamic power.
    - Example, turn off FP unit if no FP instructions are being executed.

Trends in Power in Integrated Circuits

- Although dynamic power is the primary source of power dissipation in CMOS, static power is becoming an important issue because leakage current flows even when a transistor is off:

$$\mathbf{Power_{static} = Current_{static} * Voltage}$$

Trends in Power in Integrated Circuits

- Increase in transistor leads to increase in power and leakage current.
- As a result, very low power systems are even gating the voltage to inactive modules to control loss due to leakage.
- Goal for leakage is 25% in 2006.

Trends in Cost

- Supercomputers—cost-sensitive designs are of growing significance.
- The use of technology improvements to lower cost, as well as increase performance, has been a major theme in the computer industry.

Trends in Cost

- References (books) often ignore cost in cost-performance.
  - Cost is not the same across the industry segments.
- Costs and its factors help make in intelligent decisions for designers.
  - Should a ‘new’ feature be included?

The Impact of Time, Volume, and Commodification

- The cost of a manufactured computer component decreases over time even without major improvements in the basic implementation technology.
- Why?
  - Because of learning curve!

The Impact of Time, Volume, and Commodification

- The learning curve itself is best measured by change in yield—the percentage of manufactured devices that survives the testing procedure.

The Impact of Time, Volume, and Commodification

- Example:
  - Price per megabyte of DRAM
  - Price dropped in a long term by about 40% per year
  - Microprocessors price also dropped
    - However, it is less standardized compared to DRAM.

The Impact of Time, Volume, and Commodification

- Volume is also a factor.
- Increasing volumes affect cost in several ways.
  - First, they decrease the time needed to get down the learning curve.
  - Second, volume decreases cost, since it increases purchasing and manufacturing efficiency.

The Impact of Time, Volume, and Commodification

- Volume is also a factor.
- Increasing volumes affect cost in several ways.
  - First, they decrease the time needed to get down the learning curve.
  - Second, volume decreases cost, since it increases purchasing and manufacturing efficiency.

The Impact of Time, Volume, and Commodification

- As a rule of thumb, some designers have estimated that cost decreases about 10% for each doubling of volume.
- Moreover, volume decreases the amount of development cost that must be amortized by each computer, thus allowing cost and selling price to be closer.

The Impact of Time, Volume, and Commodification

- Commodities are products that are sold by multiple vendors in large volumes and are essentially identical.
- Competition among vendors affect prices.
  - Decreases the gap between cost and selling price.
  - Decreases cost.

The Impact of Time, Volume, and Commodification

- Why price reduction happens?
  - Reductions occur because a commodity market has both volume and a clear product definition

The Impact of Time, Volume, and Commodification

- Led to the low end of the computer business being able to achieve better price- performance than other sectors and yielded greater growth at the low end, although with very limited profits (as is typical in any commodity business).

Cost of an Integrated Circuit

- Integrated circuit costs are becoming a greater portion of the cost that varies between computers, especially in the high-volume, cost-sensitive portion of the market.

Cost of an Integrated Circuit

- Computer designers must understand the costs of chips to understand the costs of current computers.
- Basic process of silicon manufacture is unchanged: A wafer is still tested and chopped into dies .

Cost of an Integrated Circuit

- Thus the cost of a packaged integrated circuit is

$$\mathbf{IC_{cost} = \frac{CD + CTD + CPFT}{FTY}}$$

Cost of IC = Cost of die + Cost of Testing die +
Cost Packaging and final test

Cost of an Integrated Circuit

- Where
  - IC_{cost} is the cost of the IC
  - CD is the cost of die
  - CTD is the cost of testing die
  - CPFT is the cost of packaging and final testing
  - FTY is the final test yield

Cost of an Integrated Circuit

- Learning how to predict the number of good chips per wafer requires first learning how many dies fit on a wafer and then learning how to predict the percentage of those that will work.

Cost of an Integrated Circuit

- From there it is simple to predict cost:

$$\mathbf{CD} = \frac{\mathbf{CW}}{\mathbf{DW \times DY}}$$

Where

CD = cost of die

CW = cost of wafer

DW = Dies per wafer

DY = Die yield

Cost of an Integrated Circuit

- The number of dies per wafer is the approximately the area of the wafer divided by the area of the die.

Cost of an Integrated Circuit

- It can be accurately estimated by

$$\mathbf{DW} = \frac{\pi * (WD/2)^2}{DA} - \frac{\pi * WD}{\sqrt{(2*DA)}}$$

where

DW = dies per wafer

WD = wafer diameter

DA = Die area

Cost of an Integrated Circuit

- First ratio refers to the wafer to die area.
- Second ratio to compensate for the “square peg in a hole” problem

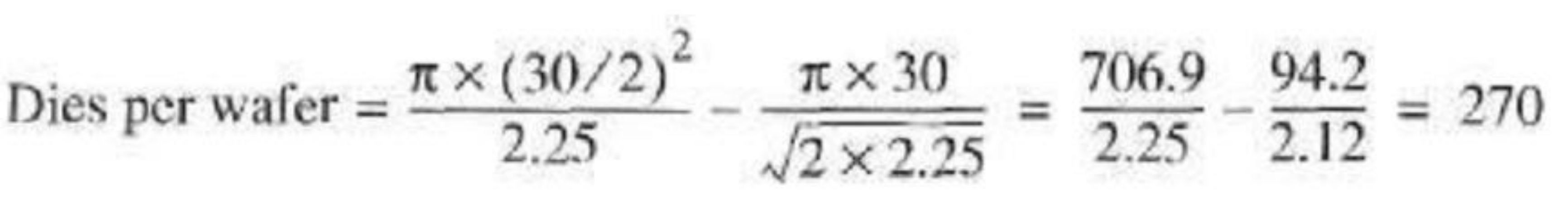
Cost of an Integrated Circuit

- Example:

Find the number of dies per 300 mm (30 cm) wafer for a die that is 1.5 cm on a side.

Cost of an Integrated Circuit

- Compute first for the die area which is 2.25 cm².
- Dies per wafer is



$$\text{Dies per wafer} = \frac{\pi \times (30/2)^2}{2.25} - \frac{\pi \times 30}{\sqrt{2 \times 2.25}} = \frac{706.9}{2.25} - \frac{94.2}{2.12} = 270$$

Cost of an Integrated Circuit

- Take note however that the value is just the number of wafers. The maximum number of wafers!
- The critical question is: What is the fraction of good dies on a wafer number, or the die yield!

Cost of an Integrated Circuit

- A simple model of integrated circuit yield, which assumes that defects are randomly distributed over the wafer and that yield is inversely proportional to the complexity of the fabrication process:

$$\text{Die yield} = \text{Wafer yield} \times \left(1 + \frac{\text{Defects per unit area} \times \text{Die area}}{\alpha}\right)^{-\alpha}$$

Cost of an Integrated Circuit

- For the sake of simplicity, we will assume that wafer yield is 100% unless otherwise stated.

Cost of an Integrated Circuit

- In 2006, this value is typically 0.4 defects per square centimeter for 90 nm, as it depends on the maturity of the process (recall the learning curve, mentioned earlier.

Cost of an Integrated Circuit

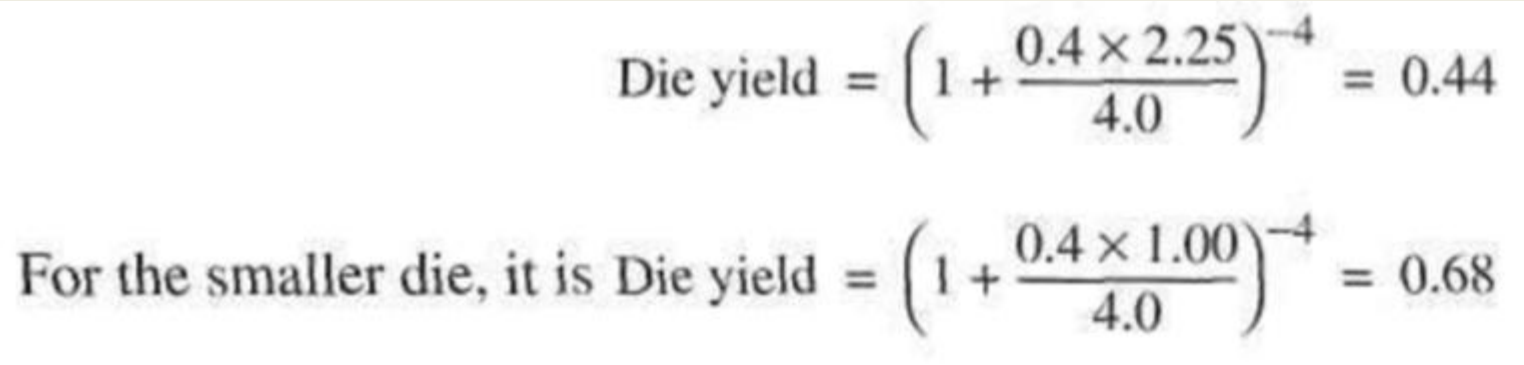
- Lastly, α is a parameter that corresponds roughly to the number of critical masking levels, a measure of manufacturing complexity.
- For multilevel metal CMOS processes in 2006, a good estimate is a = 4.0.
- For discussion and class purposes, we will use α=4.0 unless otherwise stated.

Cost of an Integrated Circuit

- Example:
- Find the die yield for dies that are 1.5 cm on a side and 1.0 cm on a side, assuming a defect density of 0.4 per cm2 and α is 4.

Cost of an Integrated Circuit

- The die has an area of 2.25 cm² and the second one has an area of 1.00 cm².



$$\text{Die yield} = \left(1 + \frac{0.4 \times 2.25}{4.0}\right)^{-4} = 0.44$$

For the smaller die, it is $\text{Die yield} = \left(1 + \frac{0.4 \times 1.00}{4.0}\right)^{-4} = 0.68$

- That is, less than half of all the large die are good but more than two-thirds of the small die are good.

Cost of an Integrated Circuit

- There are approximately 120 good 2.25 cm^2 dies from the 300 mm wafer.

Cost of an Integrated Circuit

- Example:
- Find the number of dies per 300 mm (30 cm) wafer for a die that is 1.0 cm on a side.
- Find the number of die yield.

Reference(s):

- **Hennessy, J.L., Patterson, D.A**. Computer Architecture: A Quantitative Approach (4th Ed)