# HPCA - recap and moving forward

Soumyajit Dey, Assistant Professor,
CSE, IIT Kharagpur

January 7, 2021

# What is Computer Architecture? [2]

- Several years ago, the term computer architecture often referred only to instruction set design. Implementation is considered separately. Leaves out much critical details
- What people advocate : think of the instruction set architecture (ISA) as boundary of HW and SW
- ISA : the programmer/compiler- visible instruction set
- Kind of ISA ; are instructions complex/simple in function, are they fast/slow to execute ?
- Kind of ISA : are instructions able to access memory directly ? Is access *aligned* [1] ??

[1] shall study later
[2] Ref: Hen Pat books

# Addressing modes

- Register – get data in register and register (e.g. \$r1) encoding is part of instruction. (direct addressing). Ex: add \$t0 \$t1 \$t2
- Displacement – indirect addressing. Ex: lw rd, i(rb),
- Immediate – Ex: addi \$t1 \$t1 1. Operand (limited by bit width) is a constant within the instruction itself.
- PC relative – address is the sum of the program counter and a constant in the instruction.

# Technology Trends

- Transistor density increases by about 35% per year. Increases in die size : 10% to 20% per year. The combined effect is a growth rate in transistor count on a chip of about 40% to 55% per year, or doubling every 18 to 24 months. This trend is popularly known as Moore's law.
- DRAM - the rate of improvement has continued to slow

| DRAM growth rate | Characterization of impact on DRAM capacity |
|---|---|
| 60%/year | Quadrupling every 3 years |
| 60%/year | Quadrupling every 3 years |
| 40%–60%/year | Quadrupling every 3 to 4 years |
| 40%/year | Doubling every 2 years |
| 25%–40%/year | Doubling every 2 to 3 years |

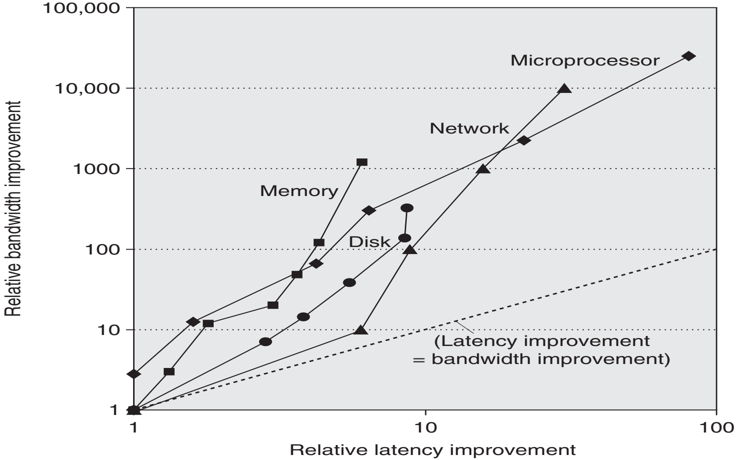
# Technology Trends

- Semiconductor Flash - Nonvolatile. Flash memory is 15 to 20 times cheaper per bit than DRAM. Capacity double roughly twice per year
- Magnetic disk - 15 to 25 times cheaper per bit than Flash. Density doubles every three years. Central to server and warehouse scale storage.

# Performance Trends

- Bandwidth or throughput is the total amount of work done in a given time, such as megabytes per second for a disk transfer.
- Latency or response time is the time between the start and the completion of an event, such as milliseconds for a disk access.
- Simple rule of thumb is that bandwidth grows by at least the square of the improvement in latency.

# Performance Trends



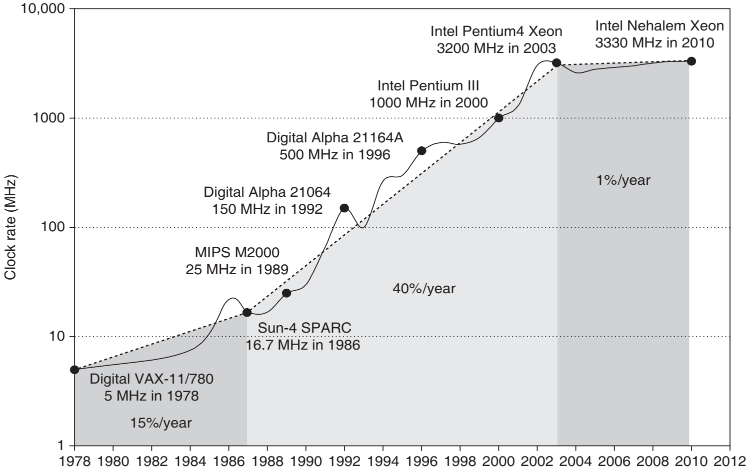
# Energy and Power consumption in Microprocessor

Transistor dynamic power consumption

- $Power_{dyn} = C \times V_{DD}^2 \times f$
- $f$ : switching frequency
- $C$ : capacitive load driven by the transistor
- $V_{DD}$ : operating voltage

The first microprocessors consumed less than a watt and the first 32-bit microprocessors used about 2 watts, while a 3.3 GHz Intel Core i7 consumes 130 watts. Given that this heat must be dissipated from a chip that is about 1.5 cm on a side, we have reached the limit of what can be cooled by air.

# Clock rate flattening out



# Optimizations

Modern microprocessors try to improve energy efficiency despite flat clock rates and constant supply voltages

- Clock gating: if no floating-point instructions are executing, the clock of the floating-point unit is disabled. If some cores are idle, their clocks are stopped.
- Dynamic Voltage-Frequency Scaling (DVFS)
- Mode based optimizations : DRAMs have a series of increasingly lower power modes to extend battery life. Disk can spin slowly when idle.

# Dependability

- $MTTF = \frac{1}{failure\ rate(\lambda)}$
- $Availability = \frac{MTTF}{MTTF+MTTR}$
- Let us check some examples from book