

---

# Introduction to Computer Architecture

## CSCE 4213

David Andrews

Rm 527 JBHT

[dandrews@uark.edu](mailto:dandrews@uark.edu)

CSCE University of Arkansas



# Agenda

---

- Technology Trends
- Types of Parallelism
- Measuring Performance



# Trends in Technology

---

- DRAM capacity: 25-40%/year (slowing)
  - 8 Gb (2014), 16 Gb (2019), possibly no 32 Gb
- Flash capacity: 50-60%/year
  - 8-10X cheaper/bit than DRAM
- Magnetic disk capacity: recently slowed to 5%/year
  - Density increases may no longer be possible, maybe increase from 7 to 9 platters
  - 8-10X cheaper/bit than Flash
  - 200-300X cheaper/bit than DRAM
- Emerging NVRAM technologies:
  - PCM, STTRAM, Memristors



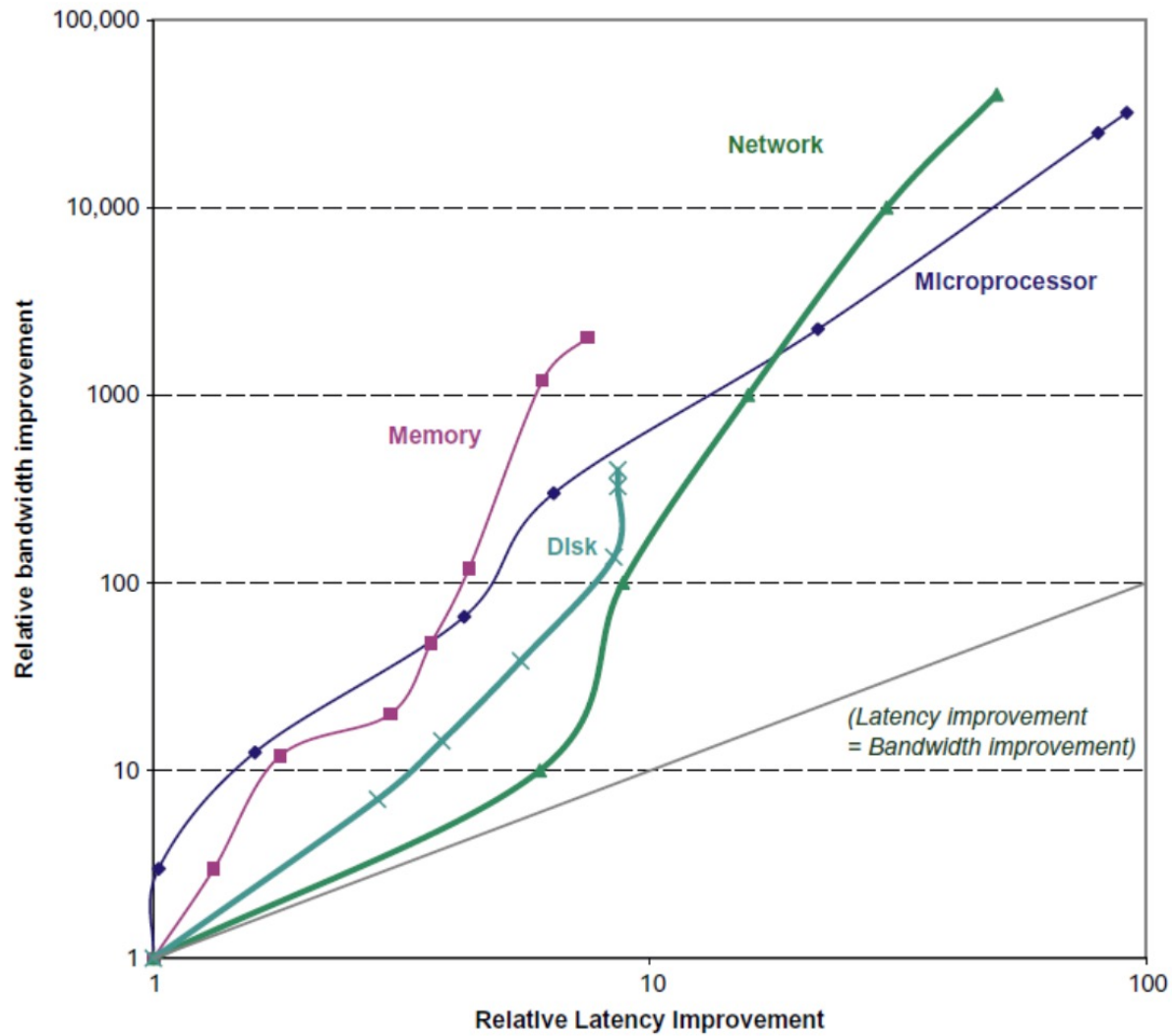
# What are we Optimizing ?

---

- Prior Design Goal: Keep processor(s) busy:
  - Focus: Bandwidth or throughput
    - Total work done in a given time
    - 32,000-40,000X improvement for processors
    - 300-1200X improvement for memory & disks
- New Design Goal: Machine Learning Latency
  - Latency or response time
  - Time between start and completion of an event
  - 50-90X improvement for processors
  - 6-8X improvement for memory and disks



# Bandwidth and Latency



Log-log plot of bandwidth and latency milestones



# How Do We Compare Performance?

---

- (Amdahl's Law) Compute Speedup ( $S_p$ ) of X relative to Y  
Execution time<sub>y</sub> / Execution time<sub>x</sub>
  - Unitless Ratio
- What is Execution Time ?
  - Depends on What You Need to Evaluate.
  - Choose to Remove Unknowns-focus on Only 1 Variable
  - Examples:
    - Wall clock time: includes all system overheads
    - CPU time: only computation time
- Use Meaningful Benchmarks
  - Kernels (e.g. matrix multiply)
  - Toy programs (e.g. sorting)
  - Synthetic benchmarks (e.g. Dhrystone)
  - Benchmark suites (e.g. SPEC06fp, TPC-C)



# Amdahl's Law

---

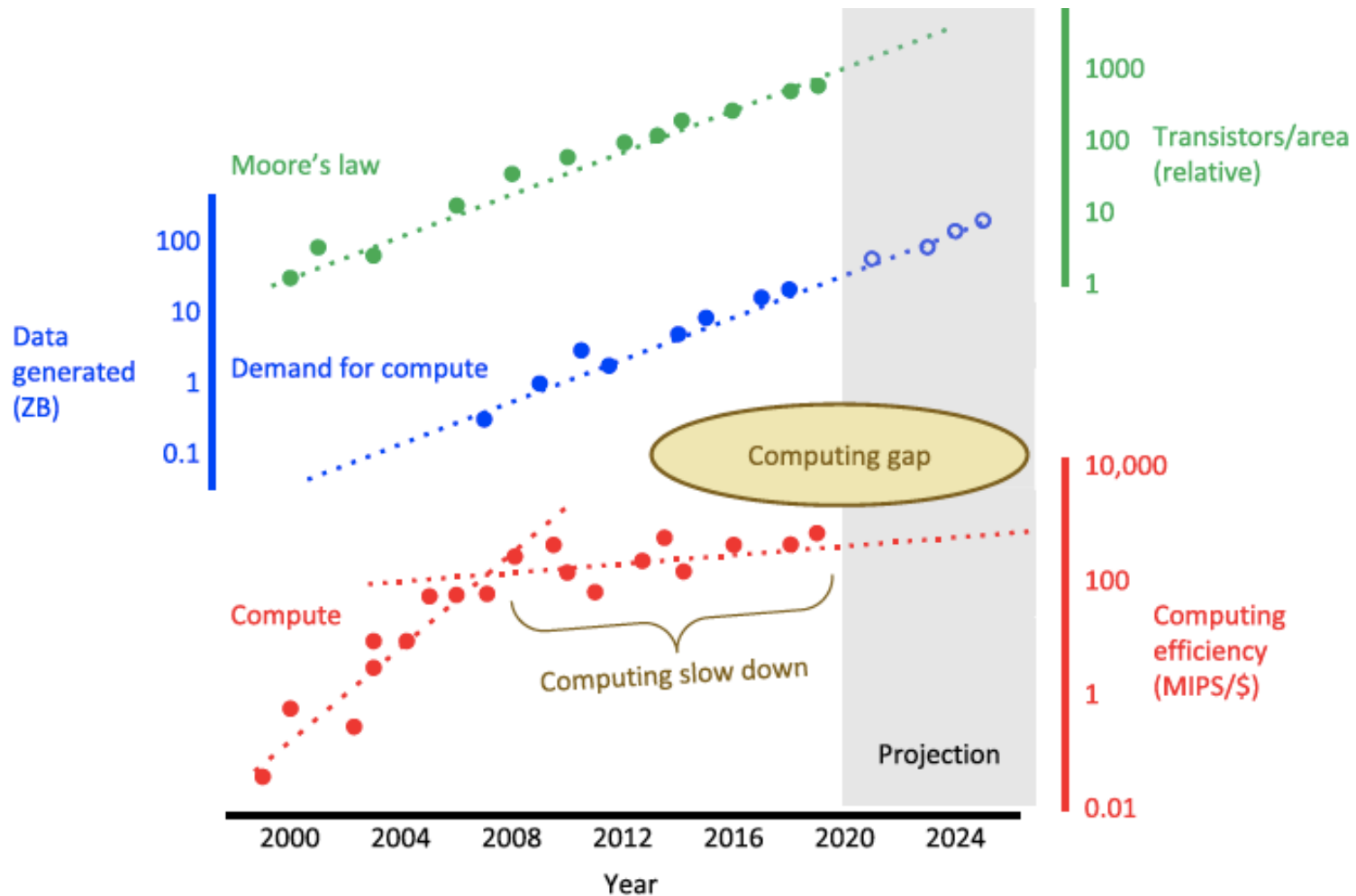
$$\text{Speedup } S_p = \frac{T_{old}}{T_{new}} = \frac{\text{Execution Time}_{old}}{\text{Execution Time}_{new}}$$

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

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



# We Created a Monster That Needs Continual Feeding !!!





# Performance Is Thru Parallelism !

---

- **Cannot Clock Faster so Do More In Parallel**
  - Apply Transistors to Exploit Parallelism
  - Parallelism Exists at Different "Granularities"
  - Circuit, Data, Instruction, Procedural, Program....
- **(Ch 3) Implicit Parallelism within a Single Processor**
  - Out of Order Instruction-Level parallelism (ILP)
  - Speculation
- **From the Application Program on Multiple Processors**
  - Data-level parallelism (DLP) (Ch 4)
  - Thread-level parallelism (TLP) (Ch 5)
  - Domain Specific Acceleration (DSA) (Ch 7)



# Flynn's Taxonomy

---

- Single instruction stream, single data stream (SISD)
- Single instruction stream, multiple data streams (SIMD)
  - Vector architectures
  - Multimedia extensions
  - Graphics processor units
- Multiple instruction streams, single data stream (MISD)
  - No commercial implementation
- Multiple instruction streams, multiple data streams (MIMD)
  - Tightly-coupled MIMD
  - Loosely-coupled MIMD



# Review CSCE 2214: Processor Performance Equation

*CPU time = CPU clock cycles for a program × Clock cycle time*

$$\text{CPU time} = \frac{\text{CPU clock cycles for a program}}{\text{Clock rate}}$$

$$\text{CPI} = \frac{\text{CPU clock cycles for a program}}{\text{Instruction count}}$$

*CPU time = Instruction count × Cycles per instruction × Clock cycle time*

$$\frac{\text{Instructions}}{\text{Program}} \times \frac{\text{Clock cycles}}{\text{Instruction}} \times \frac{\text{Seconds}}{\text{Clock cycle}} = \frac{\text{Seconds}}{\text{Program}} = \text{CPU time}$$



# Review CSCE 2214:

## Processor Performance Equation

---

- Different instruction types having different CPIs

$$\text{CPU clock cycles} = \sum_{i=1}^n IC_i \times CPI_i$$

$$\text{CPU time} = \left( \sum_{i=1}^n IC_i \times CPI_i \right) \times \text{Clock cycle time}$$



# Review CSCE 2214:

## Processor Performance Equation

- Different instruction types having different CPIs

$$\text{CPU clock cycles} = \sum_{i=1}^n IC_i \times CPI_i$$

$$\text{CPU time} = \left( \sum_{i=1}^n IC_i \times CPI_i \right) \times \text{Clock cycle time}$$

