CS/EE 6810: Computer Architecture

- Class format:
  - Most lectures on YouTube ***BEFORE*** class
  - Use class time for discussions, clarifications, problem-solving, assignments

Introduction

- Background: CS 3810 or equivalent, based on Hennessy and Patterson's Computer Organization and Design

- Text for CS/EE 6810: Hennessy and Patterson's Computer Architecture, A Quantitative Approach, 5th Edition

- Topics
  - Measuring performance/cost/power
  - Instruction level parallelism, dynamic and static
  - Memory hierarchy
  - Multiprocessors
  - Storage systems and networks

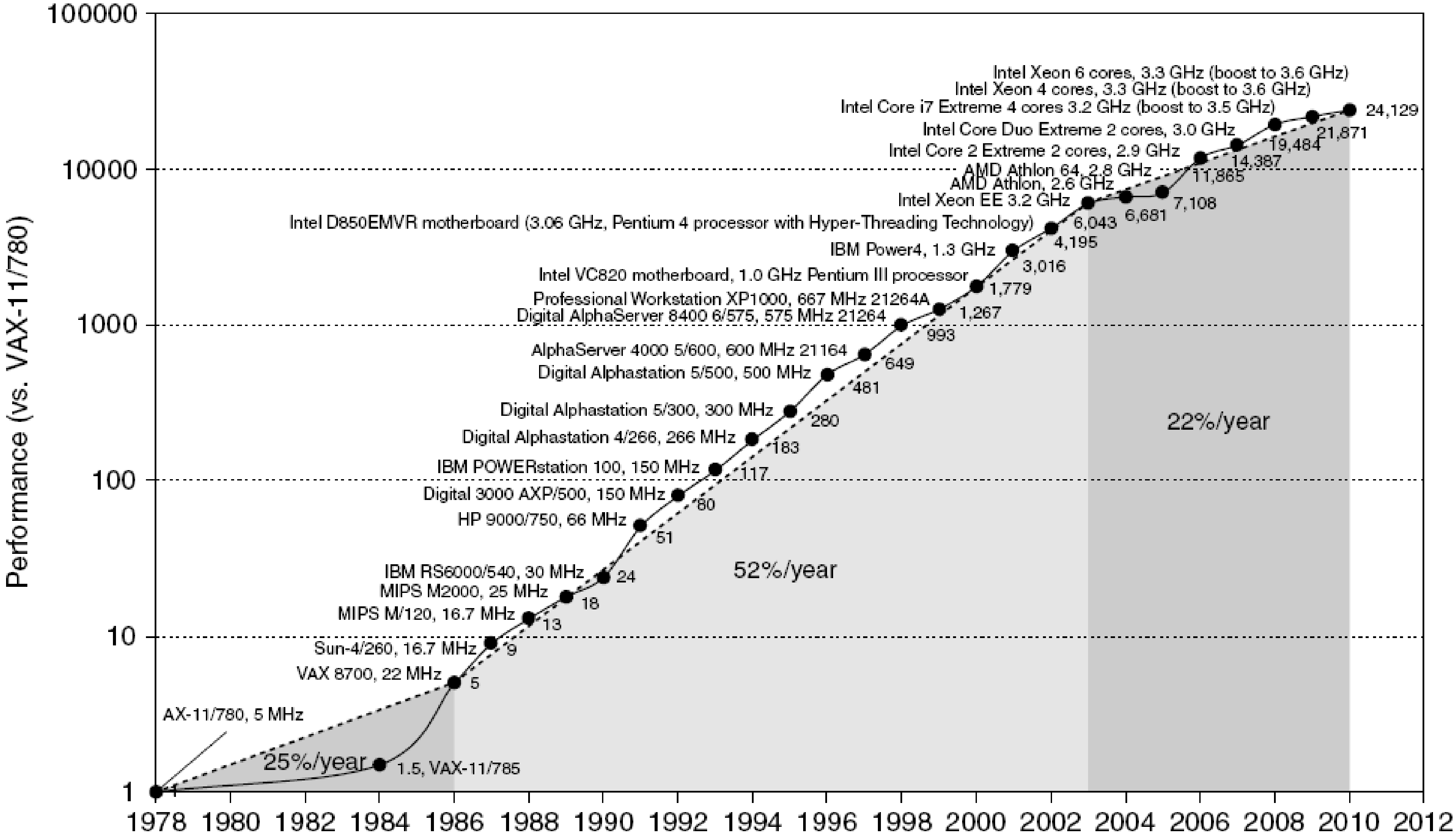
Organizational Issues

- Office hours, MEB 3414, by appointment
- TAs: Akhila Gundu, Sahil Koladiya, Shravanthi Manohar, see class webpage for office hrs
- Special accommodations, add/drop policies (see class webpage)
- Class web-page, slides, notes, and class mailing list at http://www.eng.utah.edu/~cs6810
  - Two exams, 25% each
  - Homework assignments, 50%, you may skip one
  - No tolerance for cheating

Lecture 1: Computing Trends, Metrics

- Topics: (Sections 1.1 - 1.5, 1.8 - 1.10)
  - Technology trends
  - Metrics (performance, energy, reliability)

Historical Microprocessor Performance

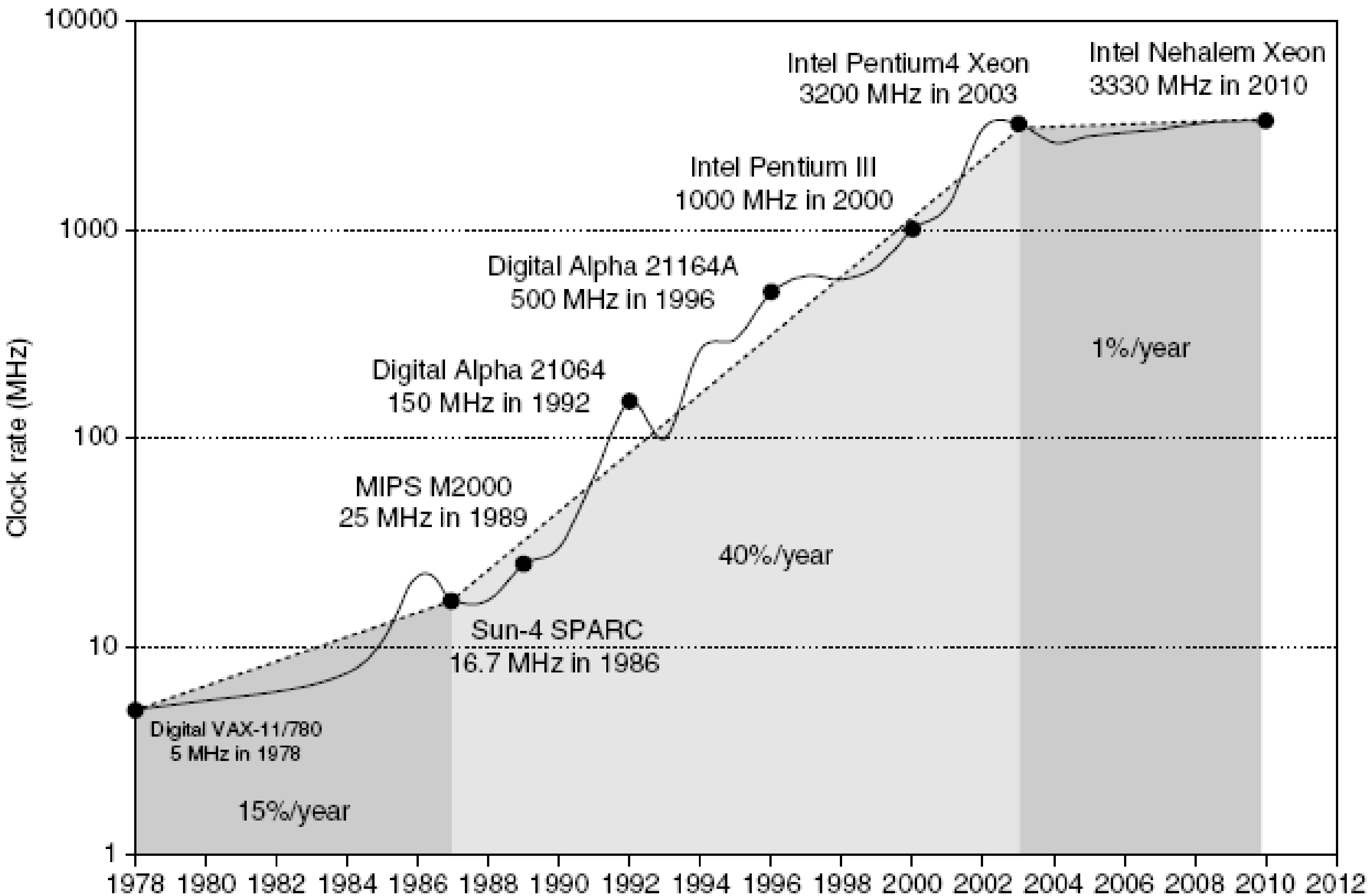


Source: H&P textbook

Points to Note

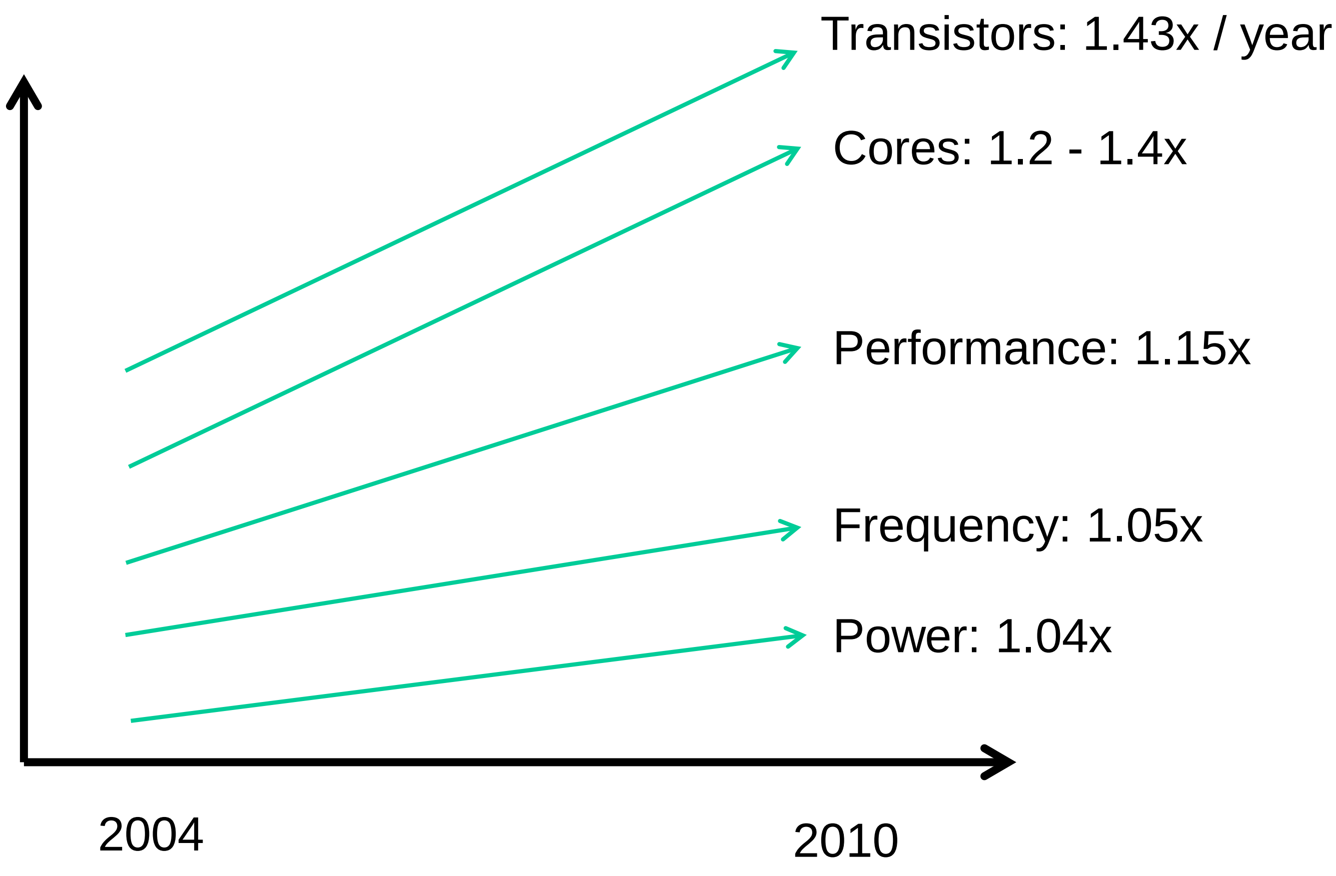
- The 52% growth per year is because of faster clock speeds and architectural innovations (led to 25x higher speed)
- Clock speed increases have dropped to 1% per year in recent years
- The 22% growth includes the parallelization from multiple cores
- Moore's Law: transistors on a chip double every 18-24 months

Clock Speed Increases



Source: H&P textbook

Recent Microprocessor Trends



Source: Micron University Symp.

Processor Technology Trends

- Transistor density increases by 35% per year and die size increases by 10-20% per year… more functionality
- Transistor speed improves linearly with size (complex equation involving voltages, resistances, capacitances)… can lead to clock speed improvements!
- Wire delays do not scale down at the same rate as logic delays
- The power wall: it is not possible to consistently run at higher frequencies without hitting power/thermal limits (Turbo Mode can cause occasional frequency boosts)

What Helps Performance?

- Note: no increase in clock speed

- In a clock cycle, can do more work -- since transistors are faster, transistors are more energy-efficient, and there's more of them

- Better architectures: finding more parallelism in one thread, better branch prediction, better cache policies, better memory organizations, more thread-level parallelism, etc.

Where Are We Headed?

- Modern trends:
  - Clock speed improvements are slowing
    - power constraints
  - Difficult to further optimize a single core for performance
  - Multi-cores: each new processor generation will accommodate more cores
  - Need better programming models and efficient execution for multi-threaded applications
  - Need better memory hierarchies
  - Need greater energy efficiency
  - In some domains, wimpy cores are attractive
  - Dark silicon, accelerators
  - Reduced data movement

Power Consumption Trends

- Dyn power α activity x capacitance x voltage^2 x frequency
- Capacitance per transistor and voltage are decreasing, but number of transistors is increasing at a faster rate; hence clock frequency must be kept steady
- Leakage power is also rising; is a function of transistor count, leakage current, and supply voltage
- Power consumption is already between 100-150W in high-performance processors today
- Energy = power x time = (dynpower + lkgpower) x time

Power Vs. Energy

- Energy is the ultimate metric: it tells us the true "cost" of performing a fixed task
- Power (energy/time) poses constraints; can only work fast enough to max out the power delivery or cooling solution
- If processor A consumes 1.2x the power of processor B, but finishes the task in 30% less time, its relative energy is 1.2 X 0.7 = 0.84; Proc-A is better, assuming that 1.2x power can be supported by the system

Reducing Power and Energy

- Can gate off transistors that are inactive (reduces leakage)
- Design for typical case and throttle down when activity exceeds a threshold
- DFS: Dynamic frequency scaling -- only reduces frequency and dynamic power, but hurts energy
- DVFS: Dynamic voltage and frequency scaling – can reduce voltage and frequency by (say) 10%; can slow a program by (say) 8%, but reduce dynamic power by 27%, reduce total power by (say) 23%, reduce total energy by 17%
  (Note: voltage drop → slow transistor → freq drop)

Other Technology Trends

- DRAM density increases by 40-60% per year, latency has reduced by 33% in 10 years (the memory wall!), bandwidth improves twice as fast as latency decreases

- Disk density improves by 100% every year, latency improvement similar to DRAM

- Emergence of NVRAM technologies that can provide a bridge between DRAM and hard disk drives

- Also, growing concerns over reliability (since transistors are smaller, operating at low voltages, and there are so many of them)

Defining Reliability and Availability

- A system toggles between
  - Service accomplishment: service matches specifications
  - Service interruption: services deviates from specs
- The toggle is caused by *failures* and *restorations*
- Reliability measures continuous service accomplishment and is usually expressed as mean time to failure (MTTF)
- Availability measures fraction of time that service matches specifications, expressed as MTTF / (MTTF + MTTR)

Cost

- Cost is determined by many factors: volume, yield, manufacturing maturity, processing steps, etc.
- One important determinant: area of the chip
- Small area ➔ more chips per wafer
- Small area ➔ one defect leads us to discard a small-area chip, i.e., yield goes up
- Roughly speaking, half the area ➔ one-third the cost

Title

- Bullet