CS/EE 6810: Computer Architecture
Fall 2022

Rajeev Balasubramonian

Mon/Wed 1:25pm - 2:45pm

CS/EE 6810: Computer Architecture

- 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 or 6th Edition

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

Lectures and Office Hours

- Class format:
  - Most lectures pre-recorded and posted on YouTube
  - Regular lectures every Mon/Wed
  - Allocate time every week to do video review, perhaps as you’re working on that week’s assignment
  - Masks, vaccines strongly encouraged; inform me in case of a positive covid test; stay home if you’re unwell
  - Office hours mentioned on class webpage; also available for a few minutes right after every lecture; email me to set up any other meetings
  - TA office hours – TBA – Tues, Wed, Thurs

Organizational Issues

- Canvas for hw submissions, announcements, grades. Assignment almost every week (due Wed/Thu).
- Special accommodations, add/drop policies, preferred names/pronouns
- Class web-page, slides, notes, and videos at https://www.cs.utah.edu/~rajeev/cs6810

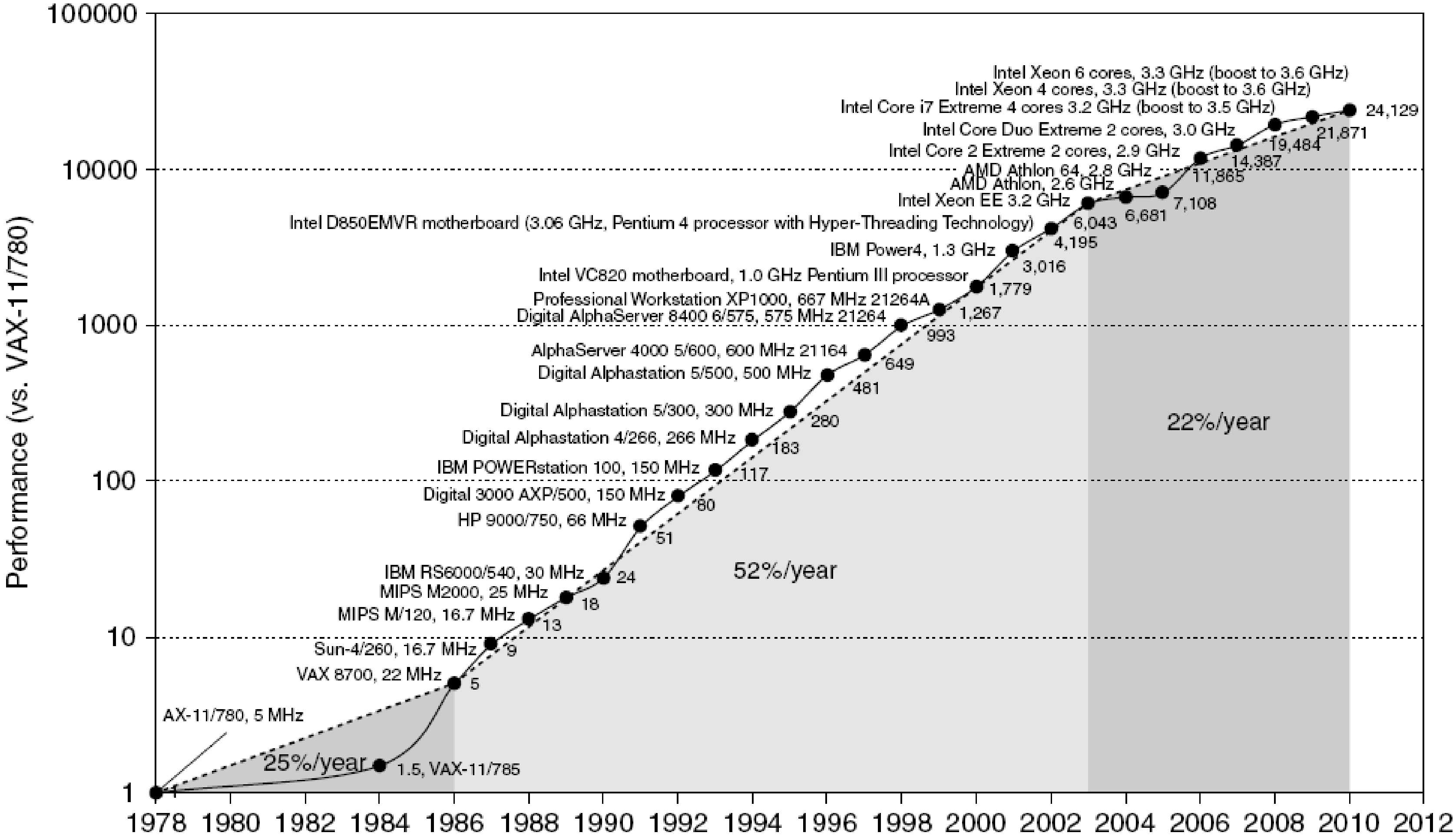
Grading

- Midterm (30%), Final exam (30%), Homeworks (40%)
- We will drop your two lowest homework scores
- No tolerance for cheating

Lecture 1: Computing Trends, Metrics

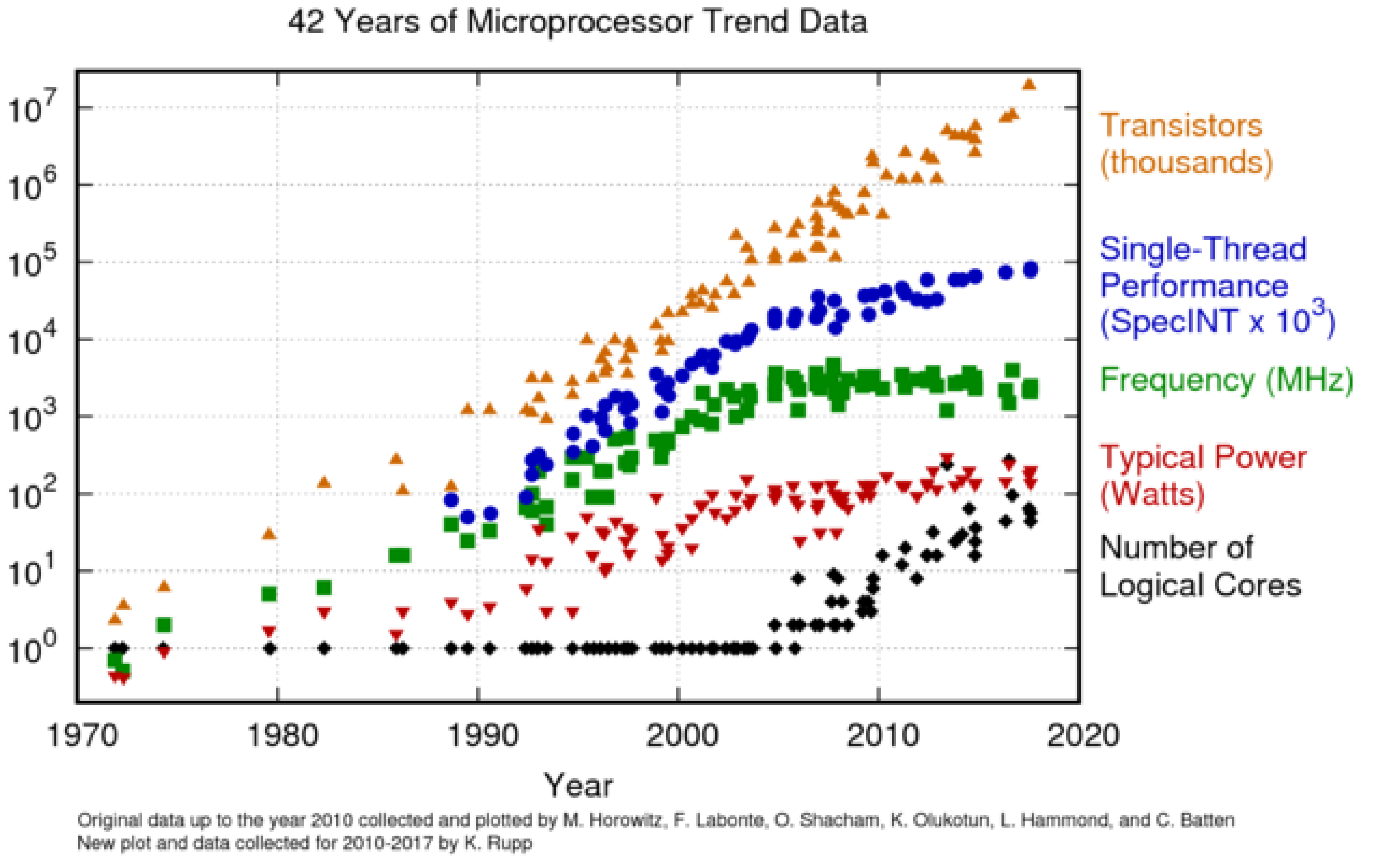
- Topics:
  - Technology trends
  - Metrics (performance, energy, reliability)

Historical Microprocessor Performance



Source: H&P textbook

Microprocessor Performance



Source: karlrupp.net

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)
- 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; fancy cooling required beyond ~150W

What Helps Performance?

- 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, moving computations to memory, accelerating some kernels, …

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

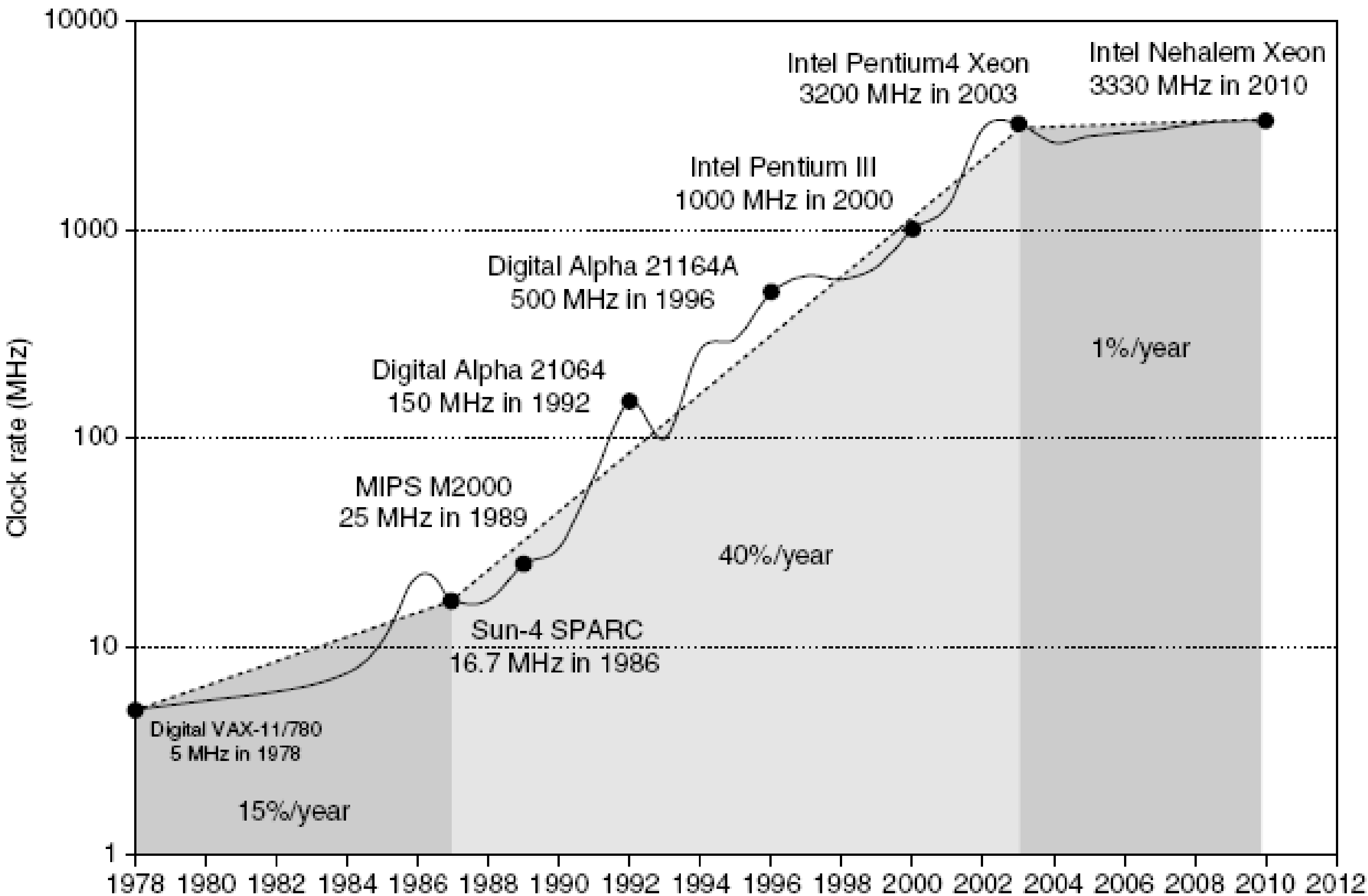
Power Impact

- Dynamic Power proportional to activity x C x V^2 x f
- Power wall: fancy cooling required beyond ~150W
- Increasing frequency led to power wall in early 2000s
- Frequency has stagnated since then
- End of voltage (Dennard) scaling in early 2010s
- Has led to dark silicon and dim silicon (occasional turbo)

Performance Stagnation

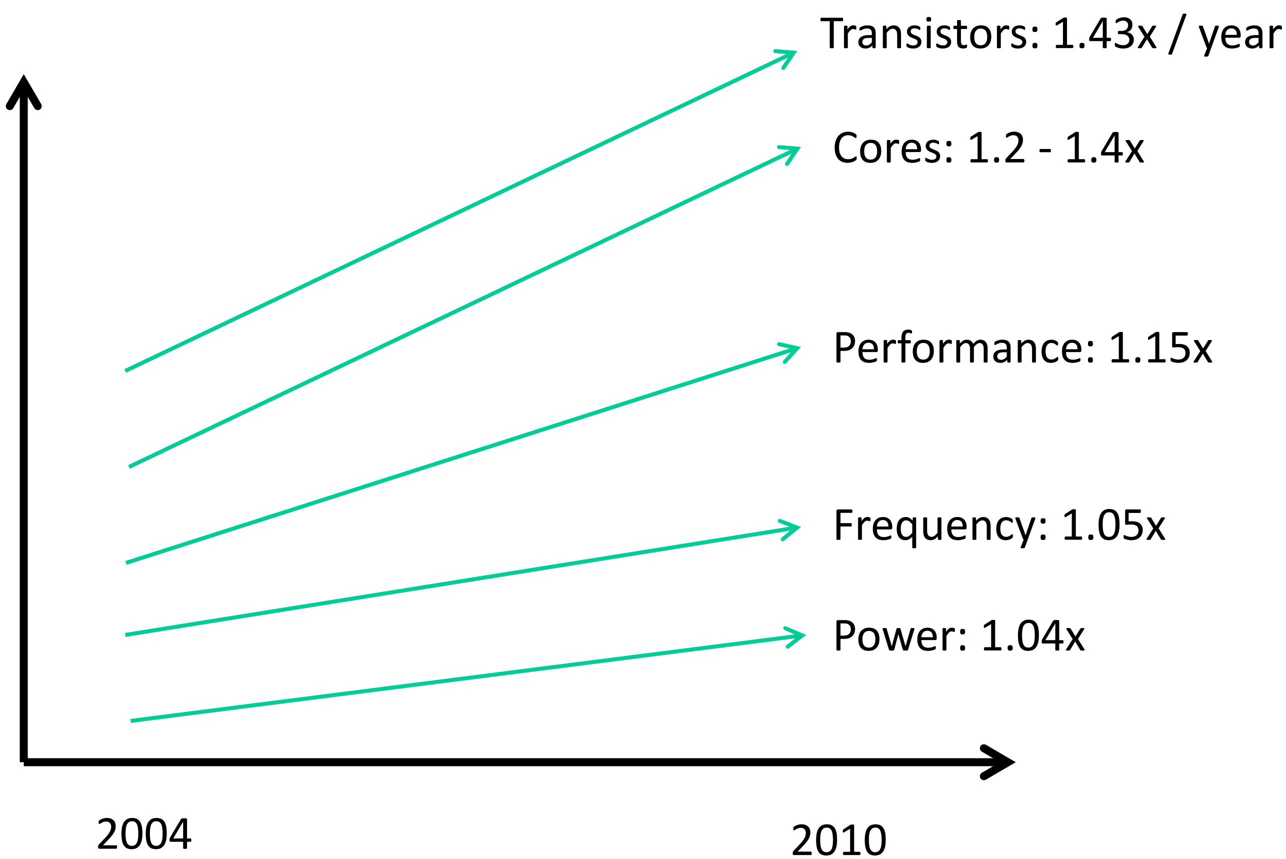
- Running out of ideas to improve single thread performance
- Power wall makes it harder to add complex features
- Power wall makes it harder to increase frequency
- Transistor count will stagnate shortly
- Additional performance provided by: more cores, occasional spikes in frequency, accelerators

Clock Speed Increases



Source: H&P textbook

Recent Microprocessor Trends



Source: Micron University Symp.

More Diverse Platforms

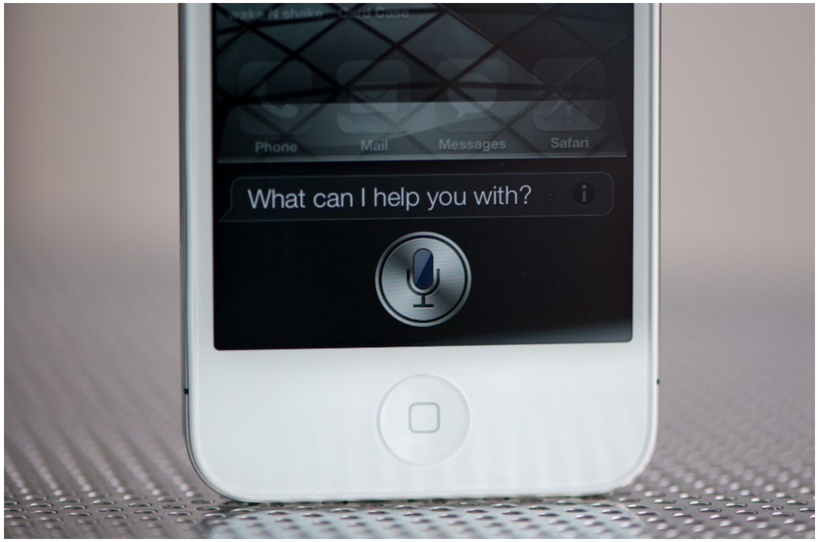


Image credits: uber, extremetech, anandtech

New Design Concerns

Image credits: gizmodo

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
- Dark silicon, accelerators
- Reduced data movement
- Emergence of new metrics: security, reliability
- Emergence of new workloads: ML, graphs, genomics