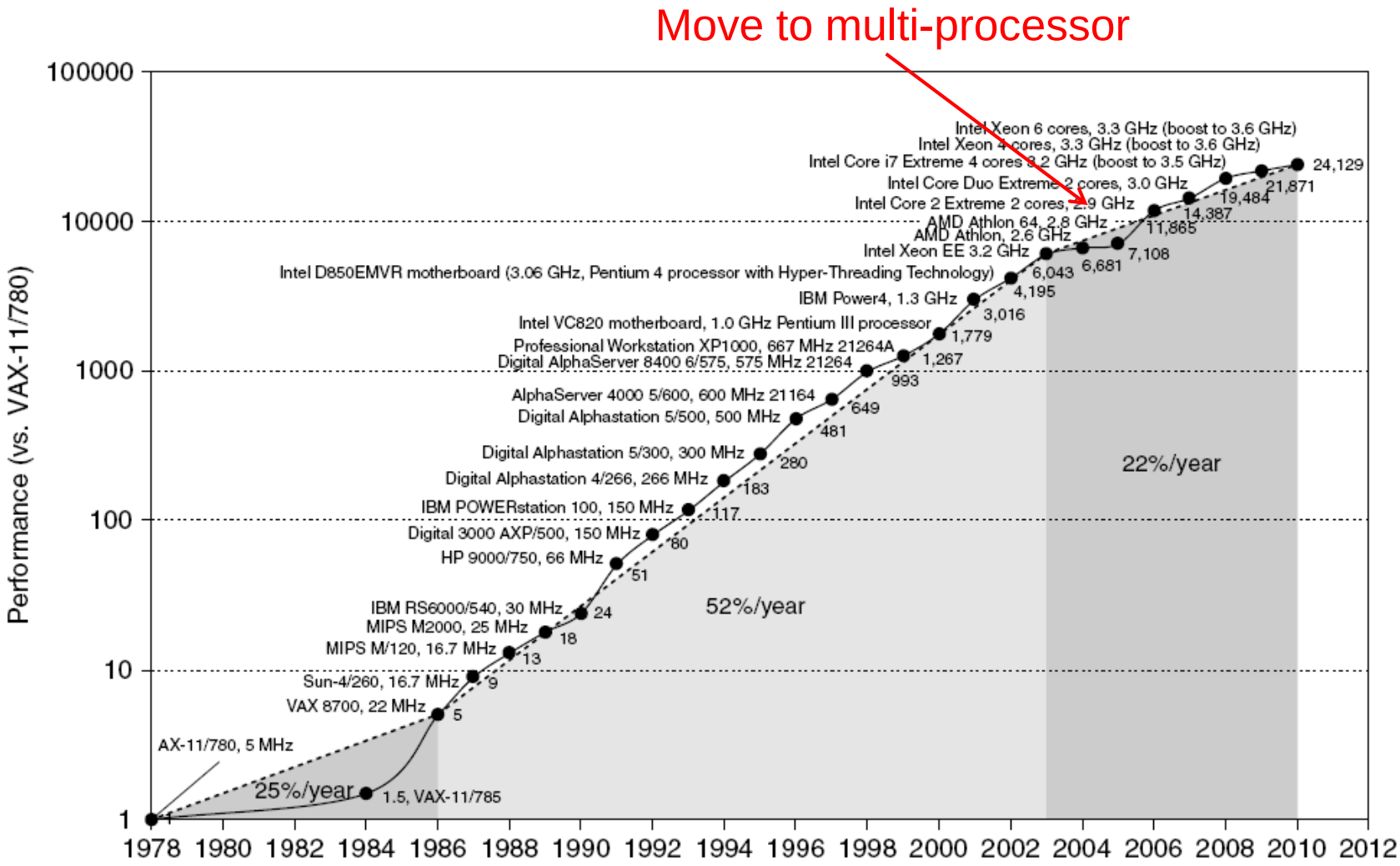
FUNDAMENTALS OF QUANTITATIVE DESIGN

Contents and objective

- Getting familiar with terms & concepts
- Overall trends
- Figure of merit (metrics)
- Computer design principles

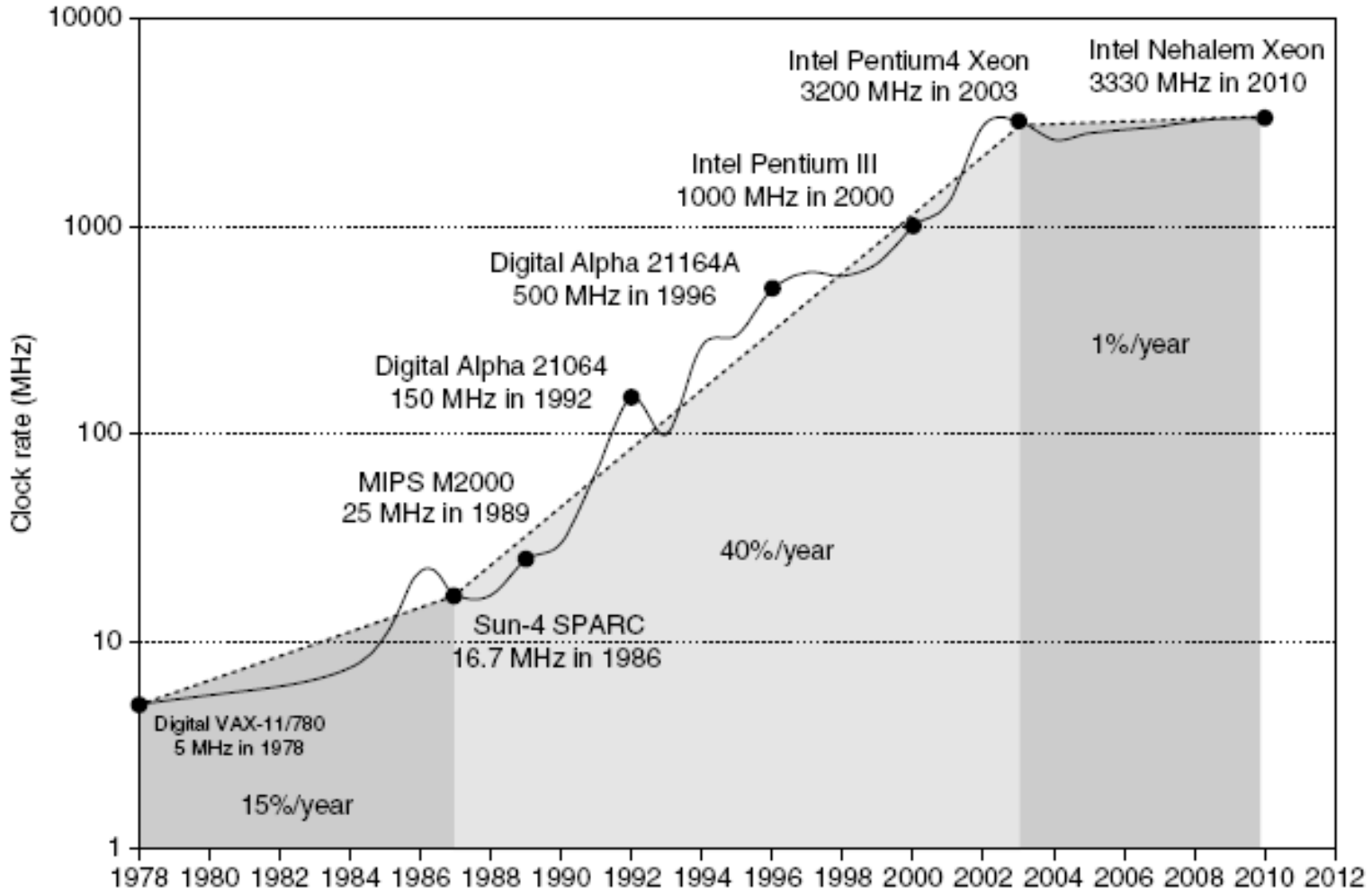
MicroProcessor Performance



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

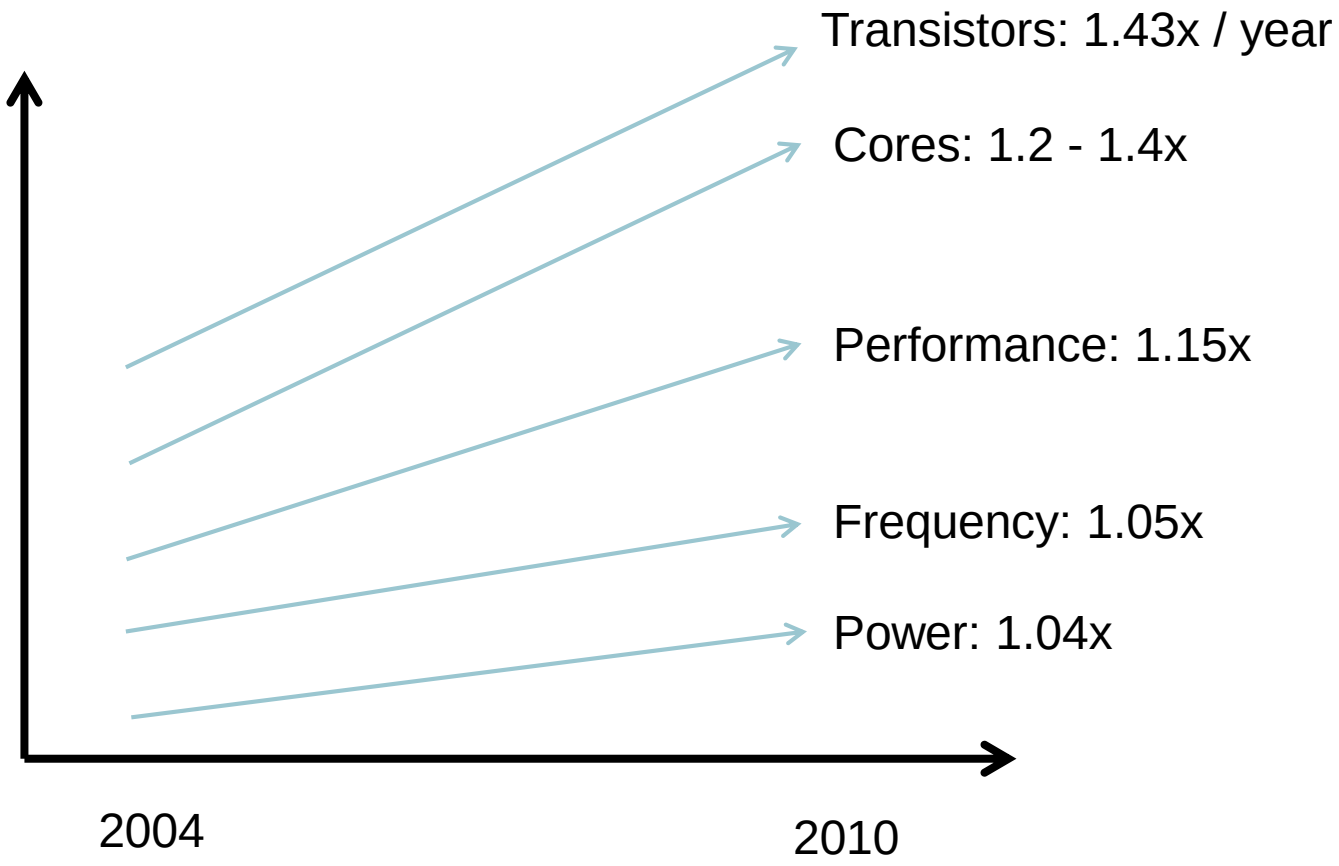
Clock Speed Increases



Source: H&P textbook

5

Recent Microprocessor Trends



Source: Micron University Symp.

6

Classes of Computers

- Personal Mobile Device (PMD)
  - e.g. start phones, tablet computers
  - Emphasis on energy efficiency and real-time
- Desktop Computing
  - Emphasis on price-performance
- Servers
  - Emphasis on availability, scalability, throughput
- Clusters / Warehouse Scale Computers
  - Emphasis on availability and price-performance
  - Sub-class: Supercomputers, emphasis: floating-point performance and fast internal networks
- Embedded Computers
  - Emphasis: price

Choosing Programs to Evaluate Perf.

- Toy benchmarks
  - e.g., quicksort, puzzle
  - No one really runs. Scary fact: used to prove the value of RISC in early 80's
- Synthetic benchmarks
  - Attempt to match average frequencies of operations and operands in real workloads.
  - e.g., Whetstone, Dhrystone
  - Often slightly more complex than kernels; But do not represent real programs
- Kernels
  - Most frequently executed pieces of real programs
  - Good for focusing on individual features not big picture
  - Tend to over-emphasize target feature
- Real programs
  - e.g., gcc, spice, SPEC2006 (standard performance evaluation corporation), TPCC, TPCD, PARSEC, SPLASH

METRICS

Transistor dimension (Area)

- Feature size
  - Also called geometry, process node
  - Minimum size of transistor or wire in x or y dimension
  - 10 microns in 1971 to .032 microns in 2011, .016 microns in 2016
  - Leads to chip-miniaturization
  - Allows fitting more transistors on a chip

Throughput and Latency

- Bandwidth or throughput
  - Total work done in a given time
  - 10,000-25,000X improvement for processors
  - 300-1200X improvement for memory and disks
- Latency or response time
  - Time between start and completion of an event
  - 30-80X improvement for processors
  - 6-8X improvement for memory and disks

Measuring Performance

- Typical performance metrics:
  - Response time
  - Throughput
- Speedup of X relative to Y
  - Execution time$_Y$ / Execution time$_X$

- Execution time
  - Wall clock time: includes all system overheads
  - CPU time: only computation time