ECSE 425 Lecture 2: Trends in Computer Architecture

H&P, Chapter 1

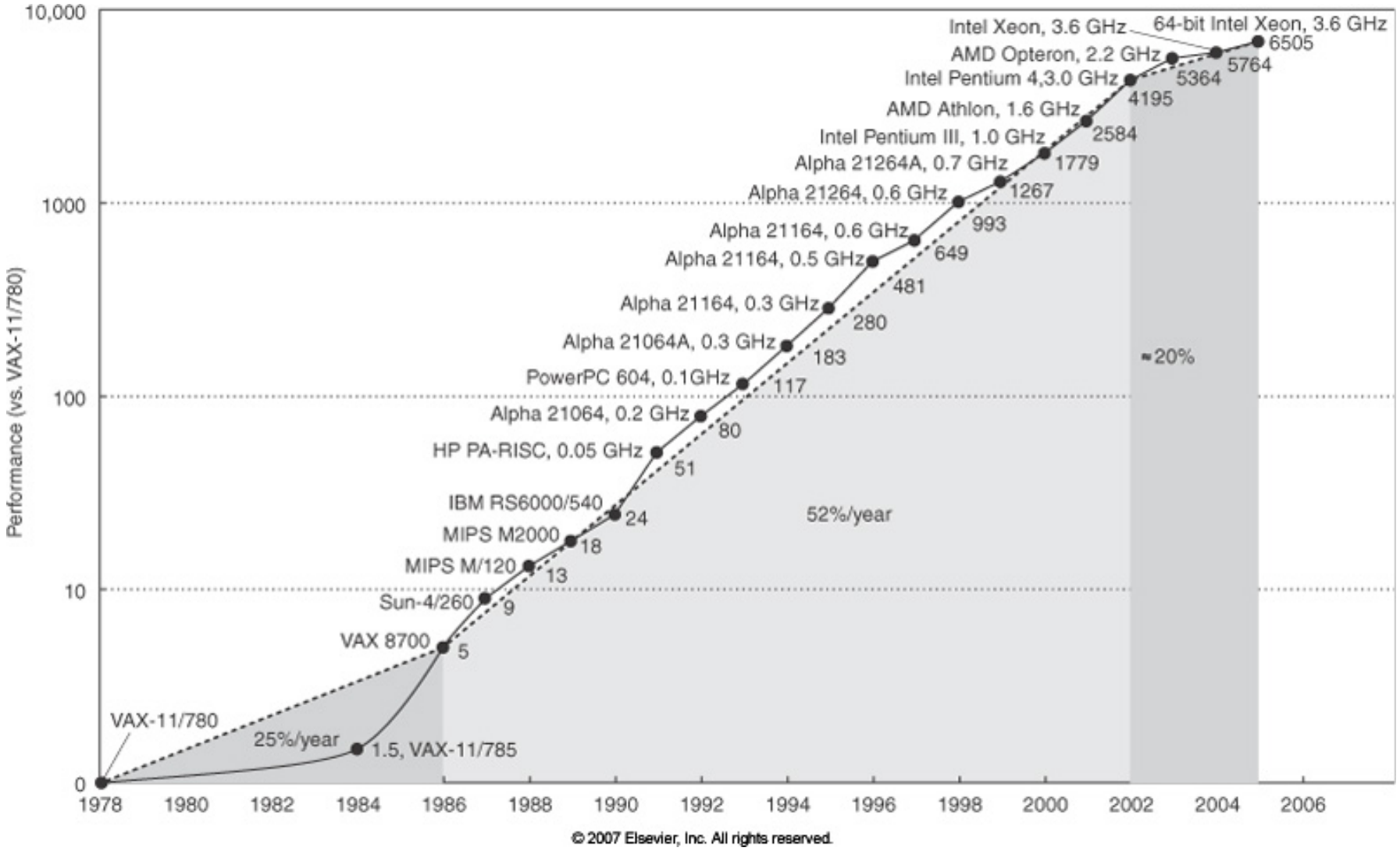
Administrative Matters

- Tutorial
  - Fridays, 3:35-4:25 PM, TR 0060
- Homework 1
  - Out today
    - WebCT
    - http://www.info425.ece.mcgill.ca
  - Due in class 9/19

Today

- Trends in Processor Performance
- Defining computer architecture
- Technology Trends
  - Bandwidth vs. Latency
  - Transistor and wire scaling
  - Power

Trends in Processor Performance



What Made This Improvement Possible?

- From 1986 to 2002, improvements in
  - Process technology (how computers are fabricated)
  - *Computer architecture* (how computers are designed)
  - Compilers (how software is prepared for computers)
- Architecture in this era: RISC
  - A few, simple instructions
  - Exploit instruction-level Parallelism (ILP)
  - Hide memory latency with multi-level cache hierarchy

What Happened in 2002?

- Since 2002, 20% / year; new challenges include
  - Power ⇒ temperature ⇒ reliability
  - Related limits in the ILP that can be exposed
  - Memory latency and bandwidth “wall”
- Architecture now: multi-processors
  - Integrate many, less complex processors on a chip
  - Exploit thread-level parallelism (TLP)
  - Exploit data-level parallelism (DLP)

What is Computer Architecture?

- Computer architecture is the art and science of
  - **selecting** *hardware components*, and
  - **interconnecting** *hardware components*, in order to
  - **satisfy** *application requirements*.
- Three aspects of computer architecture
  - Instruction set architecture (ISA) } Specification
  - Computer organization } Implementation
  - Computer hardware } Implementation

Instruction Set Architecture (ISA)

- An ISA is a contract between
  - Software developers, and
  - Hardware designers
- The ISA defines …
  - Functional behavior of a processor
  - HW interface exposed to software
- The ISA typically includes …
  - Assembly language definition
  - Programming model
- Examples of ISAs
  - 80x86, ARM, MIPS64, PowerPC, SPARC

Implementing an ISA

- Now, implementation >> ISA design
  - Computer organization
  - Computer hardware
- Organization (a.k.a., micro-architecture)
  - Pipelining, functional unit mix, memory hierarchy, branch prediction, etc.
- *Example*: AMD Opteron 64 and Intel Pentium 4
  - Same ISA, different organizations
  - Opteron: 12 stage int pipe, 1 MB L2
  - P4: 20 stage int pipe, *Hyper-Threading™*, 512 KB L2

Computer Hardware

- Logic design and packaging
  - Manufacturing technology
  - Circuit design strategy
  - Memory interface
- Has implications for
  - Clock rate
  - Power dissipation
  - Die area
  - Cooling requirements
- *Example*: Pentium 4, Mobile Pentium 4
  - Same organization, with different target applications

Architecture in Context

- Recall that computer architecture is
  - Instruction set architecture
  - Computer organization
  - Computer hardware
- An ISA may be used for decades (x86)
  - Future-proof your architecture!
- Architects must therefore be aware of trends in
  - Use: what do users want
  - Technology: what can hardware do

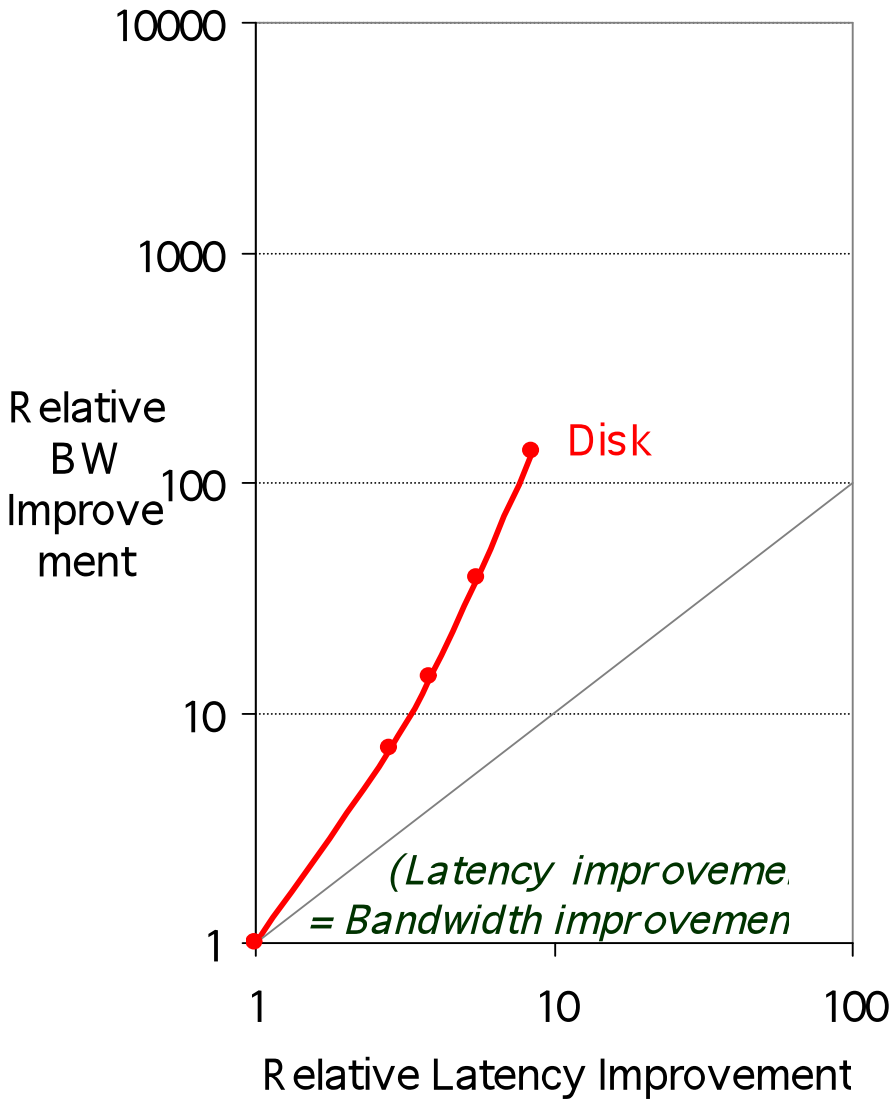
Trends in Computer Architecture

- Trends in Technology
  - Performance: bandwidth vs. latency
  - Transistor and wire scaling
- Trends in Power
- Trends in Cost
- Trends in Reliability
  - This is my area of research

Trends in Technology

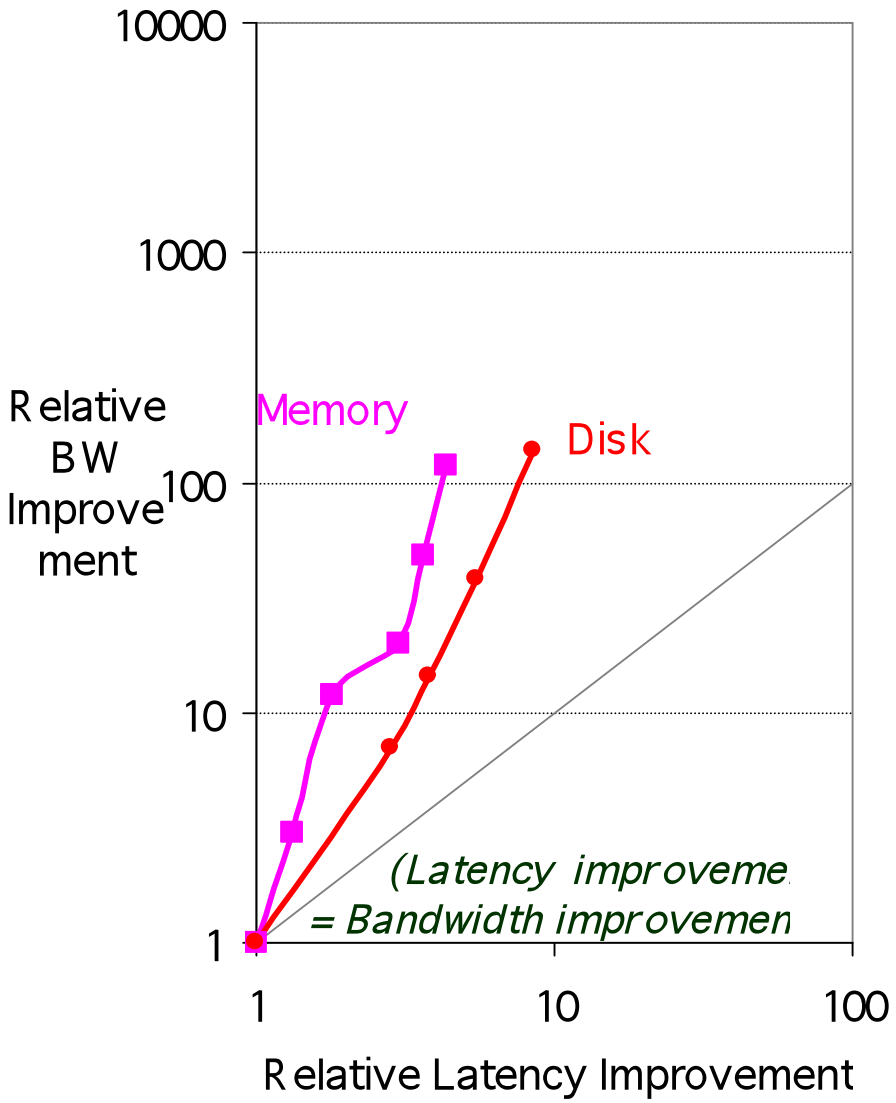
- Technologies that impact architecture:
  - **Logic** manufacturing technology
    - Transistor Density: ~ 35% / yr
    - Die size: ~ 10-20%
    - # transistors per chip: ~ 40-55%
  - **Memory** (DRAM) manufacturing technology
    - Capacity ~ 40%
  - **Storage** technology
    - Magnetic disk density ~ 30% (since 2004)
  - **Network** technology

Bandwidth vs. Latency



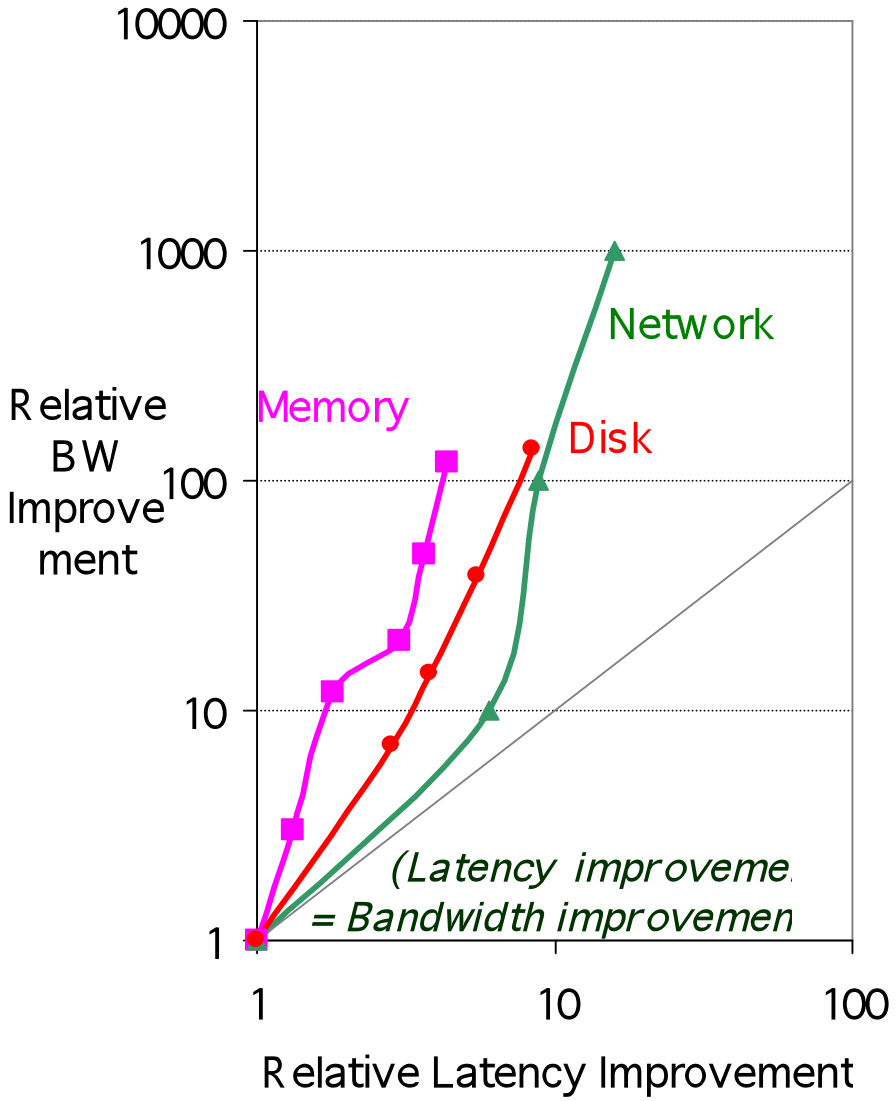
- Disk : 3600, 5400, 7200, 10000, 15000 RPM
  - (8x, 143x)

Bandwidth vs. Latency



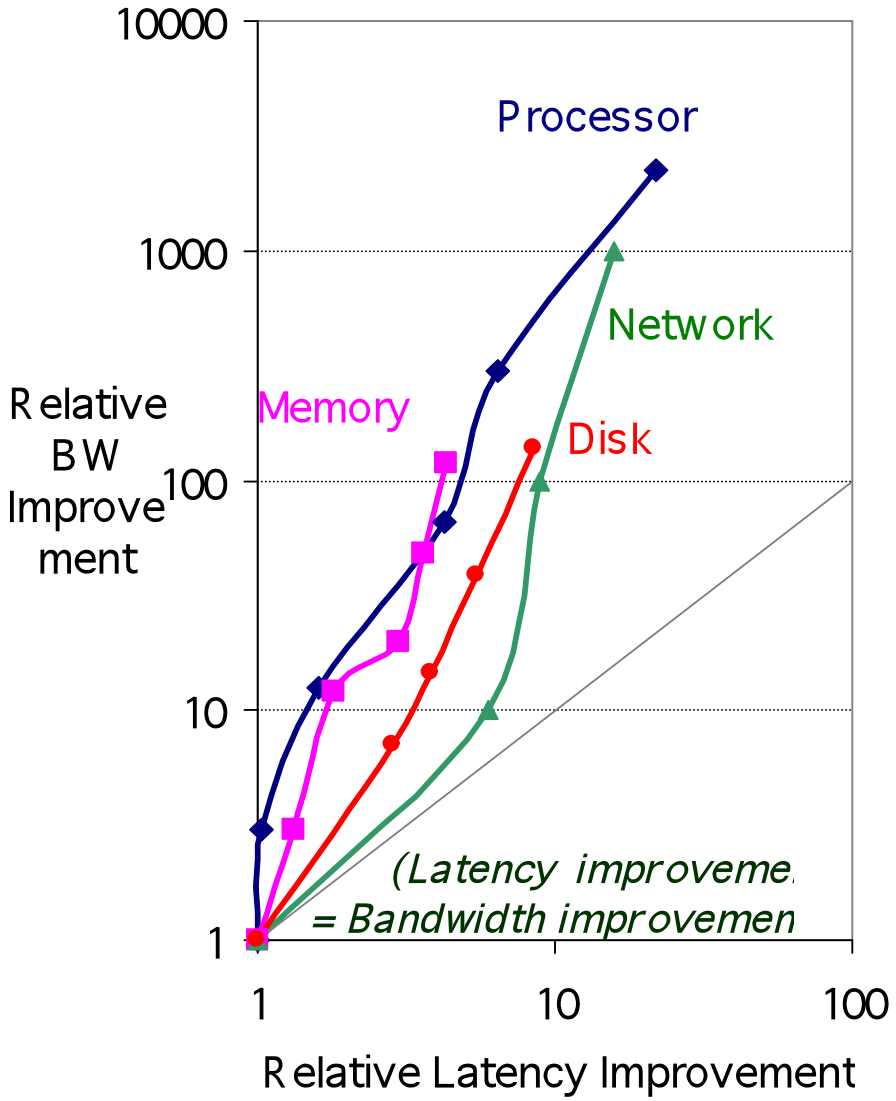
- Memory: 16bit plain DRAM, Page Mode DRAM, 32b, 64b, SDRAM, DDR
  - (4x,120x)
- Disk : 3600, 5400, 7200, 10000, 15000 RPM
  - (8x, 143x)

Bandwidth vs. Latency



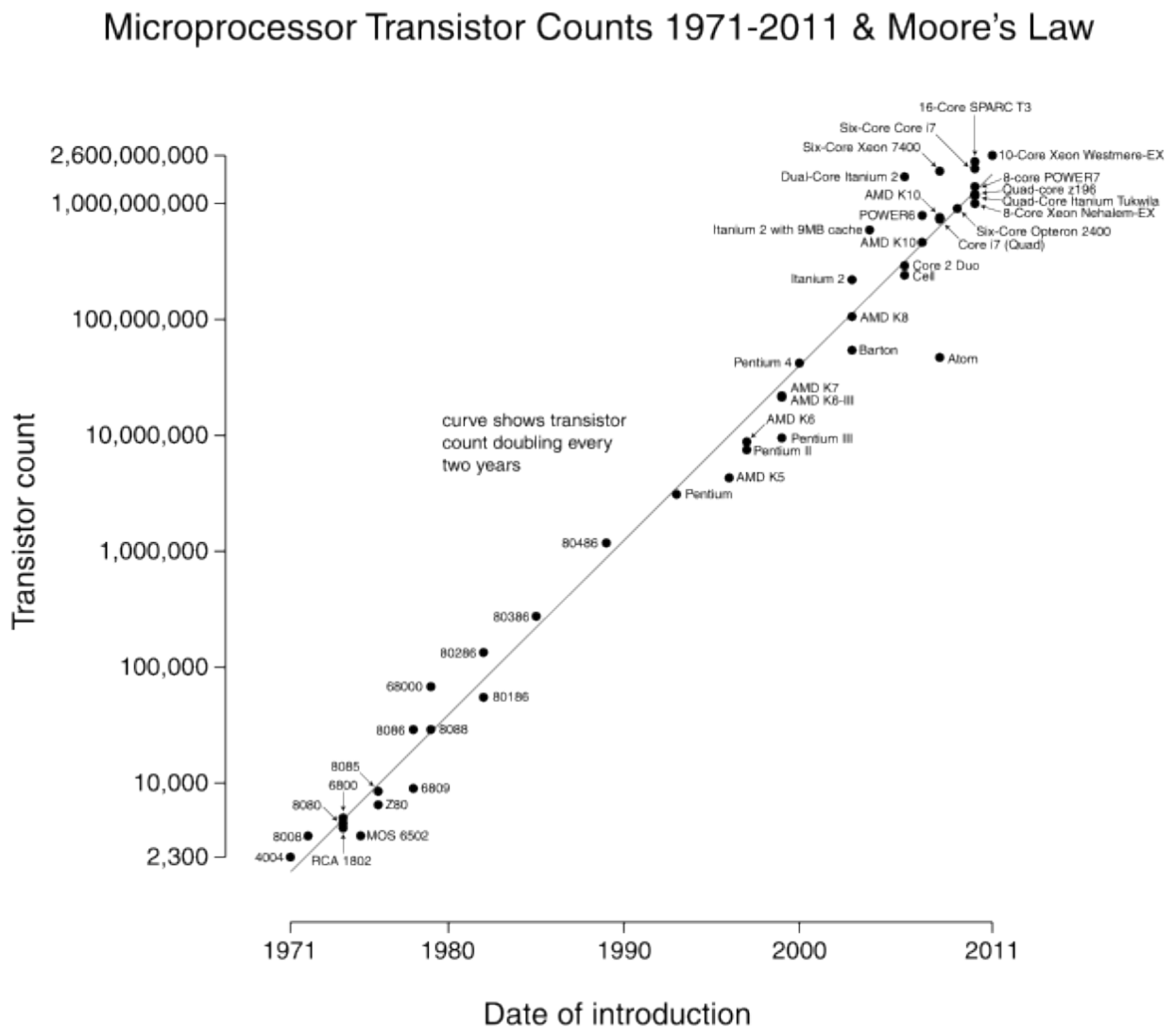
- Ethernet: 10Mb, 100Mb, 1000Mb, 10000 Mb/s
  - (16x,1000x)
- Memory: 16bit plain DRAM, Page Mode DRAM, 32b, 64b, SDRAM, DDR
  - (4x,120x)
- Disk : 3600, 5400, 7200, 10000, 15000 RPM
  - (8x, 143x)

Bandwidth vs. Latency



- Processor: ‘286, ‘386, ‘486, Pentium, Pentium Pro, Pentium 4
  - (21x,2250x)
- Ethernet: 10Mb, 100Mb, 1000Mb, 10000 Mb/s
  - (16x,1000x)
- Memory: 16bit plain DRAM, Page Mode DRAM, 32b, 64b, SDRAM, DDR
  - (4x,120x)
- Disk : 3600, 5400, 7200, 10000, 15000 RPM
  - (8x, 143x)

Scaling: Moore's Law



[Credit: Wgsimon, wikepedia.com]

Transistor and Wires Scaling

- Process technology nodes are defined by their *feature size*
  - Minimum width of a transistor or wire
- As feature size decreases linearly …
  - Quadratic increase in transistor density
  - Linear increase in transistor performance

Scaling and Architecture

- Datapath width
  - 4, 8, 16, 32, and 64 bit architectures (buses, ALU's)
- Datapath organization
  - Longer pipelines
  - More complex branch prediction
- On-chip memory
  - L1, L2 and L3 caches
- Propagation delay is a major problem
  - P4 dedicates two pipeline stages to propagation delay

Trends in IC Power

- Total Power = Dynamic Power + Static Power
- Dynamic (switching) power
  - Power consumed charging and discharging capacitances in integrated circuits (ICs)
- Static power
  - Power dissipated when circuits aren't switching
  - Also called "leakage" power

Dynamic Power

- $P_{dynamic} = \alpha f C V^2$
  - *alpha* = switching factor
  - *f* = clock frequency
  - *C* = load capacitance
  - *V* = supply voltage
- $E_{dynamic} = CV^2$
  - Half dissipated during charging
  - Half stored in the capacitor

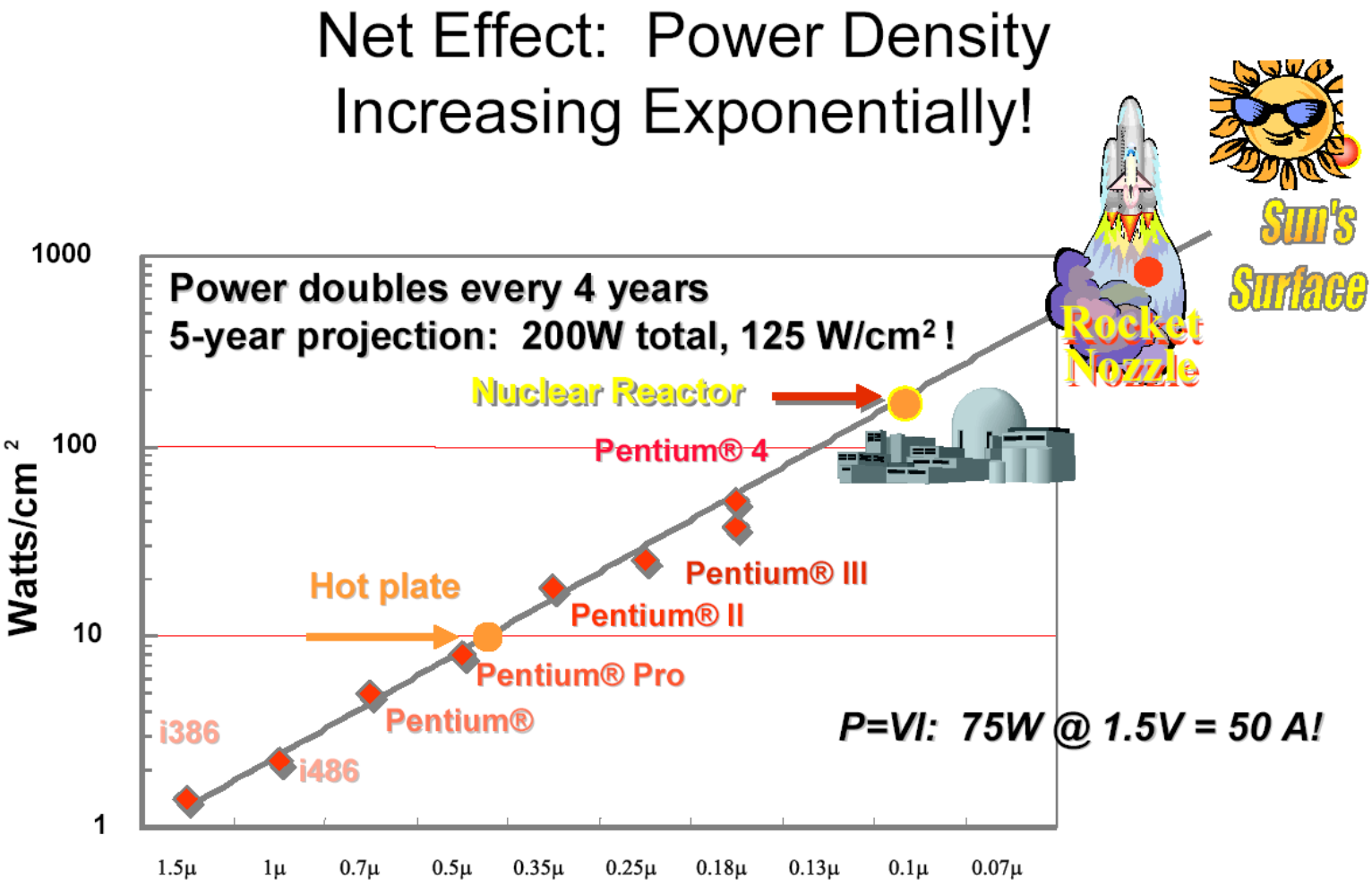
Static Power

- $P_{static} = I_{static} \cdot V$
- Contribution to *P* grows with each generation
- Exponentially dependent on
  - Junction temperature
  - V_{DD}, V_{Th}

Power Example

- Many microprocessors have adjustable supply voltage. Suppose a 15% reduction in voltage results in a 15% reduction in frequency.
- What is the effect on dynamic power?

Power and Architecture



* "New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies" – Fred Pollack, Intel Corp. Micro32 conference key note - 1999. Courtesy Avi Mendelson, Intel.

Summary

- Computer architecture
  - Instruction set architecture
  - Computer organization
  - Computer hardware
- Trends
  - Performance: increasing!
    - Processor technology, comp. architecture, compilers
  - Bandwidth vs. latency: easier to improve bandwidth
  - Power: increasing with scaling, performance
    - An important limit

Next Time

- More Trends
  - Cost
  - Dependability

Backup Slides

Briefly: Implications of ILP ⇒ TLP

- What are the implications of moving to exploit TLP rather than ILP?
  - Software
  - Hardware