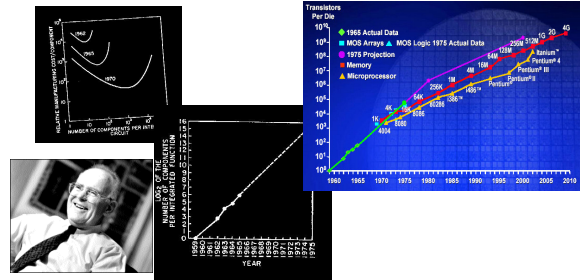


CMSC 411
Computer Systems Architecture
Lecture 2
Trends in Technology

Moore's Law: 2X transistors / "year"



- "Cramming More Components onto Integrated Circuits"
 - Gordon Moore, Electronics, 1965
- # on transistors / cost-effective integrated circuit double every N months ($12 \leq N \leq 24$)

CMSC 411 - 3 (from Patterson)

2

Tracking Technology Performance Trends

- Drill down into 4 technologies:
 - Disks,
 - Memory,
 - Network,
 - Processors
- Compare ~1980 Archaic (Nostalgic) vs. ~2000 Modern (Newfangled)
 - Performance Milestones in each technology
- Compare for Bandwidth vs. Latency improvements in performance over time
- Bandwidth: number of events per unit time
 - E.g., Mbits / second over network, Mbytes / second from disk
- Latency: elapsed time for a single event
 - E.g., one-way network delay in microseconds, average disk access time in milliseconds

CMSC 411 - 3 (from Patterson)

3

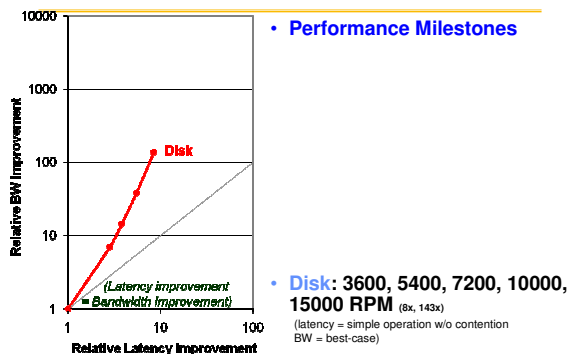
Disks: Archaic(Nostalgic) v. Modern(Newfangled)

- | | |
|-----------------------------|--|
| • CDC Wren I, 1983 | • Seagate 373453, 2003 |
| • 3600 RPM | • 15000 RPM (4X) |
| • 0.03 GBytes capacity | • 73.4 GBytes (2500X) |
| • Tracks/Inch: 800 | • Tracks/Inch: 64000 (80X) |
| • Bits/Inch: 9550 | • Bits/Inch: 533,000 (60X) |
| • Three 5.25" platters | • Four 2.5" platters (in 3.5" form factor) |
| • Bandwidth: 0.6 MBytes/sec | • Bandwidth: 86 MBytes/sec (140X) |
| • Latency: 48.3 ms | • Latency: 5.7 ms (8X) |
| • Cache: none | • Cache: 8 MBytes |

CMSC 411 - 3 (from Patterson)

4

Latency Lags Bandwidth (for last ~20 years)



CMSC 411 - 3 (from Patterson)

5

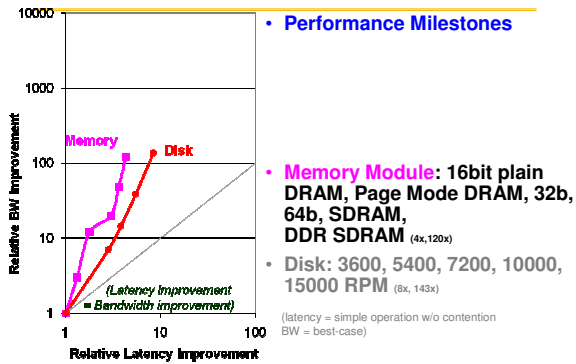
Memory: Archaic (Nostalgic) v. Modern (Newfangled)

- | | |
|--|--|
| • 1980 DRAM (asynchronous) | • 2000 Double Data Rate Synchr. (clocked) DRAM |
| • 0.06 Mbits/chip | • 256.00 Mbits/chip (4000X) |
| • 64,000 xtors, 35 mm ² | • 256,000,000 xtors, 204 mm ² |
| • 16-bit data bus per module, 16 pins/chip | • 64-bit data bus per DIMM, 66 pins/chip (4X) |
| • 13 Mbytes/sec | • 1600 Mbytes/sec (120X) |
| • Latency: 225 ns | • Latency: 52 ns (4X) |
| • (no block transfer) | • Block transfers (page mode) |

CMSC 411 - 3 (from Patterson)

6

Latency Lags Bandwidth (last ~20 years)

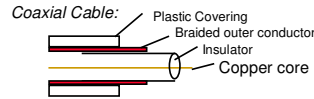


CMSC 411 - 3 (from Patterson)

7

LANs: Archaic (Nostalgic)v. Modern (Newfangled)

- | | |
|---|---|
| <ul style="list-style-type: none"> • Ethernet 802.3 • Year of Standard: 1978 • 10 Mbits/s link speed • Latency: 3000 μsec • Shared media • Coaxial cable | <ul style="list-style-type: none"> • Ethernet 802.3ae • Year of Standard: 2003 • 10,000 Mbits/s (1000X) link speed • Latency: 190 μsec (15X) • Switched media • Category 5 copper wire |
|---|---|



"Cat 5" is 4 twisted pairs in bundle
Twisted Pair:

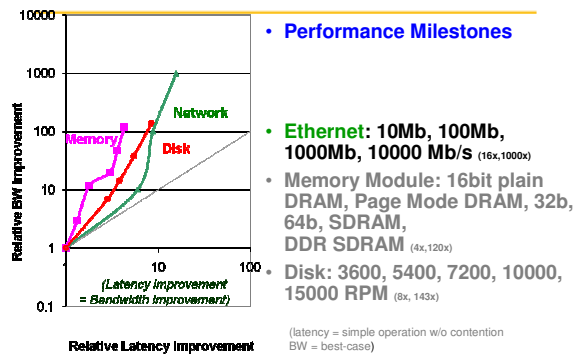


Copper, 1mm thick, twisted to avoid antenna effect

CMSC 411 - 3 (from Patterson)

8

Latency Lags Bandwidth (last ~20 years)



CMSC 411 - 3 (from Patterson)

9

CPUs: Archaic (Nostalgic) v. Modern (Newfangled)

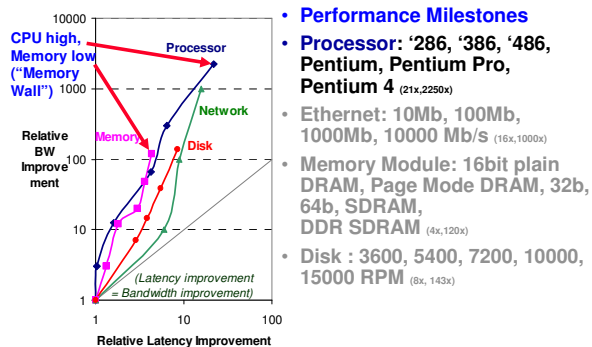
- | | |
|--|---|
| <ul style="list-style-type: none"> • 1982 Intel 80286 • 12.5 MHz • 2 MIPS (peak) • Latency 320 ns • 134,000 xtors, 47 mm² • 16-bit data bus, 68 pins • Microcode interpreter, separate FPU chip • (no caches) | <ul style="list-style-type: none"> • 2001 Intel Pentium 4 • 1500 MHz(120X) • 4500 MIPS (peak) (2250X) • Latency 15 ns (20X) • 42,000,000 xtors, 217 mm² • 64-bit data bus, 423 pins • 3-way superscalar, Dynamic translate to RISC, Superpipelined (22 stage), Out-of-Order execution • On-chip 8KB Data caches, 96KB Instr. Trace cache, 256KB L2 cache |
|--|---|



CMSC 411 - 3 (from Patterson)

10

Latency Lags Bandwidth (last ~20 years)



CMSC 411 - 3 (from Patterson)

11

Rule of Thumb for Latency Lagging BW

- In the time that bandwidth doubles, latency improves by no more than a factor of 1.2 to 1.4 (and capacity improves faster than bandwidth)
- Stated alternatively:
Bandwidth improves by more than the square of the improvement in Latency

CMSC 411 - 3 (from Patterson)

12

6 Reasons Latency Lags Bandwidth

1. Moore's Law helps BW more than latency

- **Faster transistors, more transistors, more pins help Bandwidth**
 - » MPU Transistors: 0.130 vs. 42 M xtors (300X)
 - » DRAM Transistors: 0.064 vs. 256 M xtors (4000X)
 - » MPU Pins: 68 vs. 423 pins (6X)
 - » DRAM Pins: 16 vs. 66 pins (4X)
- **Smaller, faster transistors but communicate over (relatively) longer wires: limits latency**
 - » Feature size: 1.5 to 3 vs. 0.18 micron (8X,17X)
 - » MPU Die Size: 35 vs. 204 mm² (ratio sqrt \Rightarrow 2X)
 - » DRAM Die Size: 47 vs. 217 mm² (ratio sqrt \Rightarrow 2X)

CMSC 411 - 3 (from Patterson)

13

6 Reasons Latency Lags Bandwidth (cont'd)

2. Distance limits latency

- **Size of DRAM block \Rightarrow long bit and word lines \Rightarrow most of DRAM access time**
- **Speed of light and computers on network**

3. Bandwidth easier to sell ("bigger=better")

- **E.g., 10 Gbits/s Ethernet ("10 Gig") vs. 10 μ sec latency Ethernet**
- **4400 MB/s DIMM ("PC4400") vs. 50 ns latency**
- **Even if just marketing, customers now trained**
- **Since bandwidth sells, more resources thrown at bandwidth, which further tips the balance**

CMSC 411 - 3 (from Patterson)

14

6 Reasons Latency Lags Bandwidth (cont'd)

4. Latency helps BW, but not vice versa

- **Spinning disk faster improves both bandwidth and rotational latency**
 - » 3600 RPM \Rightarrow 15000 RPM = 4.2X
 - » Average rotational latency: 8.3 ms \Rightarrow 2.0 ms
 - » Things being equal, also helps BW by 4.2X
- **Lower DRAM latency \Rightarrow More access/second (higher bandwidth)**
- **Higher linear density helps disk BW (and capacity), but not disk Latency**
 - » 9,550 BPI \Rightarrow 533,000 BPI \Rightarrow 60X in BW

CMSC 411 - 3 (from Patterson)

15

6 Reasons Latency Lags Bandwidth (cont'd)

5. Bandwidth hurts latency

- **Queues help Bandwidth, hurt Latency (Queuing Theory)**
- **Adding chips to widen a memory module increases Bandwidth but higher fan-out on address lines may increase Latency**

6. Operating System overhead hurts Latency more than Bandwidth

- **Long messages amortize overhead; overhead bigger part of short messages**

CMSC 411 - 3 (from Patterson)

16

Summary of Technology Trends

- **For disk, LAN, memory, and microprocessor, bandwidth improves by square of latency improvement**
 - In the time that bandwidth doubles, latency improves by no more than 1.2X to 1.4X
- **Lag probably even larger in real systems, as bandwidth gains multiplied by replicated components**
 - Multiple processors in a cluster or even in a chip
 - Multiple disks in a disk array
 - Multiple memory modules in a large memory
 - Simultaneous communication in switched LAN
- **HW and SW developers should innovate assuming Latency Lags Bandwidth**
 - If everything improves at the same rate, then nothing really changes
 - When rates vary, require real innovation

CMSC 411 - 3 (from Patterson)

17

TRENDS IN SILICON COSTS

CMSC 411 - 1

18

Costs

- From Figure 1.9 in H&P 3/e
- The cost of components in a \$1000 PC in 2001 are:
 - CPU – 22%
 - Monitor – 19%
 - Hard drive – only 9%
 - DRAM – only 5% (for 128MB)
 - Software – 20% (OS & basic office suite)

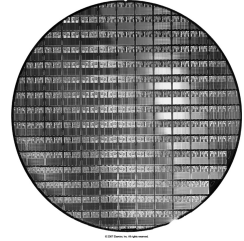
CMSC 411 - 2

19

Manufacture of DRAM and other chips

- Chips are manufactured on wafers - circular disks containing many dies (chips).
- The wafer is tested and chopped into dies.

Fig. 1.12 in H&P
117 AMD Opterons



CMSC 411 - 2

20

Wafers and dies

- To find the cost of a die:
 - Number of dies per wafer is *at most* the area of the wafer divided by the area of the die.
 - The cost of the wafer divided by the number of working dies per wafer is the cost of each die.
- The fraction of working dies is called the *die yield*, which decreases as the area of the die increases.
- Rule of thumb (p. 20): Cost of die is proportional to the square of the die area

CMSC 411 - 2

21