Lecture 2

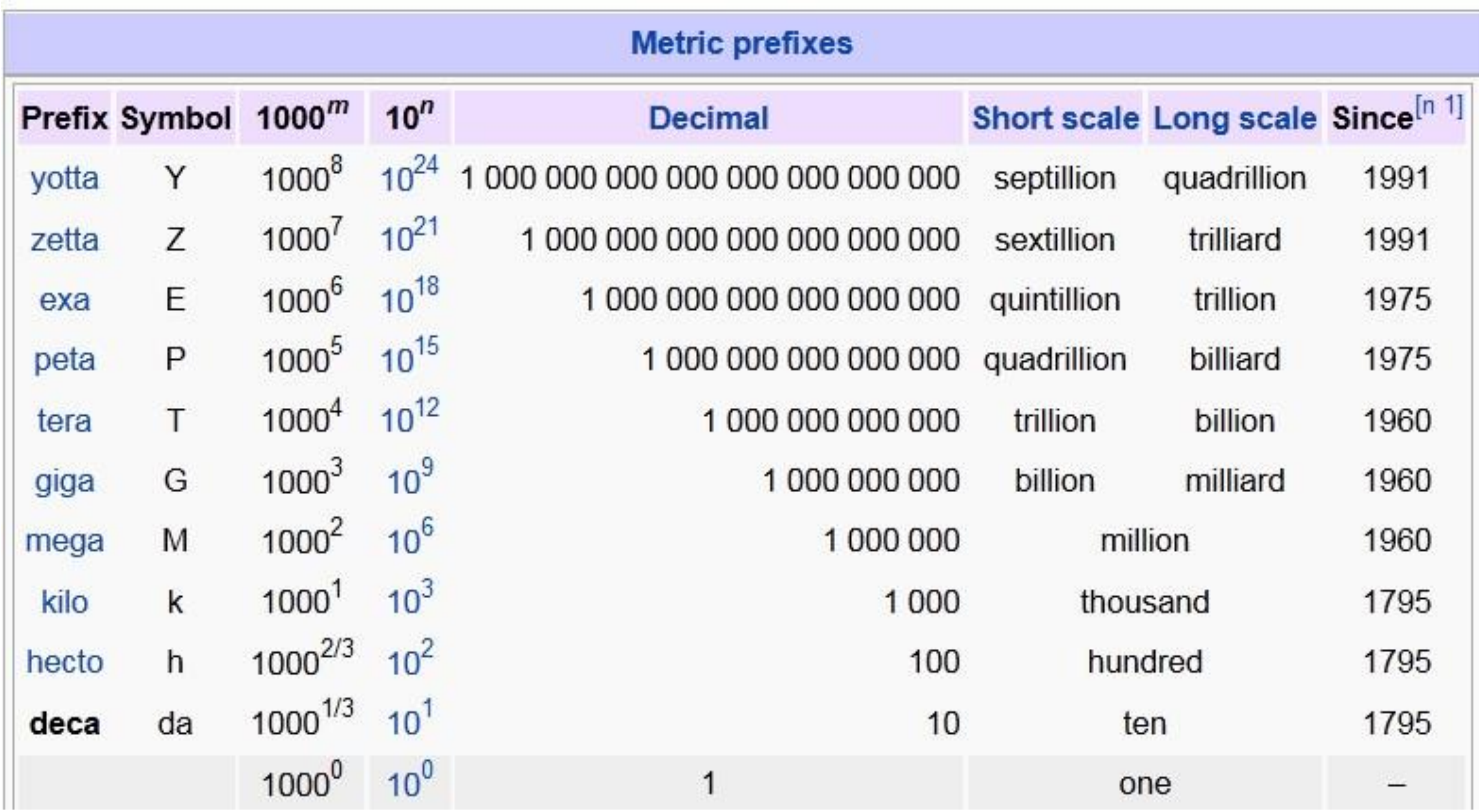
Computer Architecture

(Spring 2020)

Introduction

Dr. Duo Liu (刘铎)
Office: Main Building 0626
Email: liuduo@cqu.edu.cn

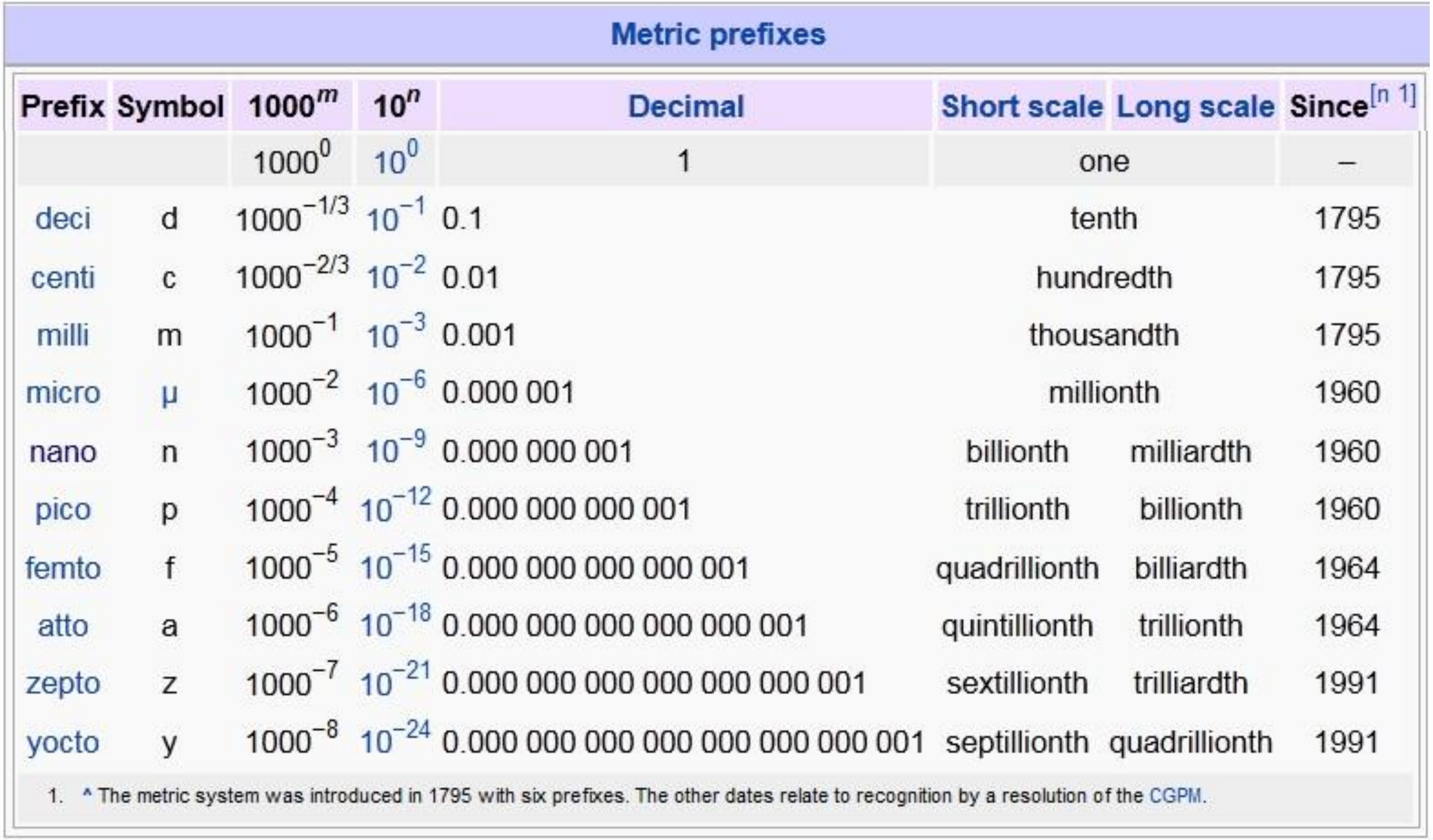
Review: Some Basic Definitions



Metric prefixes

Prefix	Symbol	1000^m	10^n	Decimal	Short scale	Long scale	Since[n 1]
yotta	Y	1000^8	10^{24}	1 000 000 000 000 000 000 000 000	septillion	quadrillion	1991
zetta	Z	1000^7	10^{21}	1 000 000 000 000 000 000 000	sextillion	trilliard	1991
exa	E	1000^6	10^{18}	1 000 000 000 000 000 000	quintillion	trillion	1975
peta	P	1000^5	10^{15}	1 000 000 000 000 000	quadrillion	billiard	1975
tera	T	1000^4	10^{12}	1 000 000 000 000	trillion	billion	1960
giga	G	1000^3	10^9	1 000 000 000	billion	milliard	1960
mega	M	1000^2	10^6	1 000 000	million		1960
kilo	k	1000^1	10^3	1 000	thousand		1795
hecto	h	$1000^{2/3}$	10^2	100	hundred		1795
deca	da	$1000^{1/3}$	10^1	10	ten		1795
		1000^0	10^0	1	one		–

Review: Some Basic Definitions



Metric prefixes

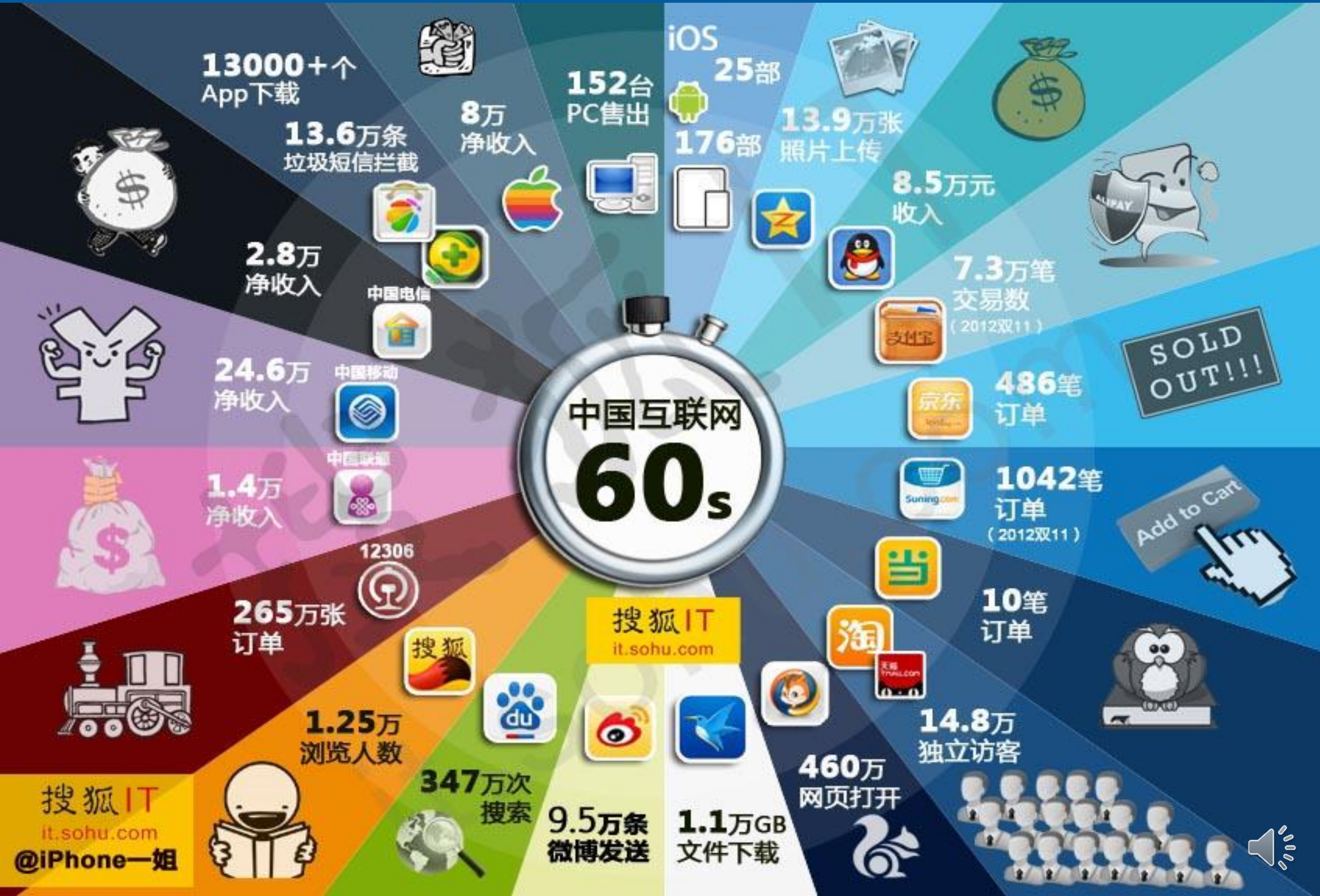
Prefix	Symbol	1000^m	10^n	Decimal	Short scale	Long scale	Since[n 1]
		1000^0	10^0	1	one		–
deci	d	$1000^{-1/3}$	10^{-1}	0.1	tenth		1795
centi	c	$1000^{-2/3}$	10^{-2}	0.01	hundredth		1795
milli	m	1000^{-1}	10^{-3}	0.001	thousandth		1795
micro	µ	1000^{-2}	10^{-6}	0.000 001	millionth		1960
nano	n	1000^{-3}	10^{-9}	0.000 000 001	billionth	milliardth	1960
pico	p	1000^{-4}	10^{-12}	0.000 000 000 001	trillionth	billionth	1960
femto	f	1000^{-5}	10^{-15}	0.000 000 000 000 001	quadrillionth	billiardth	1964
atto	a	1000^{-6}	10^{-18}	0.000 000 000 000 000 001	quintillionth	trillionth	1964
zepto	z	1000^{-7}	10^{-21}	0.000 000 000 000 000 000 001	sextillionth	trilliardth	1991
yocto	y	1000^{-8}	10^{-24}	0.000 000 000 000 000 000 000 001	septillionth	quadrillionth	1991

1. ^ The metric system was introduced in 1795 with six prefixes. The other dates relate to recognition by a resolution of the CGPM.

Review: Binary Prefix (1998)

- Kilobyte – 2^{10} or 1,024 bytes
- Megabyte– 2^{20} or 1, 024 Kilobytes
  - sometimes “rounded” to 10^{6} or 1,000,000 bytes
- Gigabyte – 2^{30} or 1, 024 Megabytes
  - sometimes rounded to 10^{9} or 1,000,000,000 bytes
- Terabyte – 2^{40} or 1, 024 Gigabytes
  - sometimes rounded to 10^{12} or 1,000,000,000,000 bytes
- Petabyte – 2^{50} or 1024 Terabytes
  - sometimes rounded to 10^{15} or 1,000,000,000,000,000 bytes
- Exabyte – 2^{60} or 1024 Petabytes
  - Sometimes rounded to 10^{18} or 1,000,000,000,000,000,000 bytes
- Zettabyte – 2^{70} or 1024 Exabytes
  - Sometimes rounded to 10^{21} or 1,000,000,000,000,000,000, 000 bytes

What Happens in 60s?



Data Explosion

1 ZB = 1,024 Exabyte (EB)
1 EB = 1,024 Petabyte (PB)
1 PB = 1,024 Terabyte (TB)
1 TB = 1,024 Gigabyte (GB)

847ZB/Year

218ZB/Year

Data Created
2016

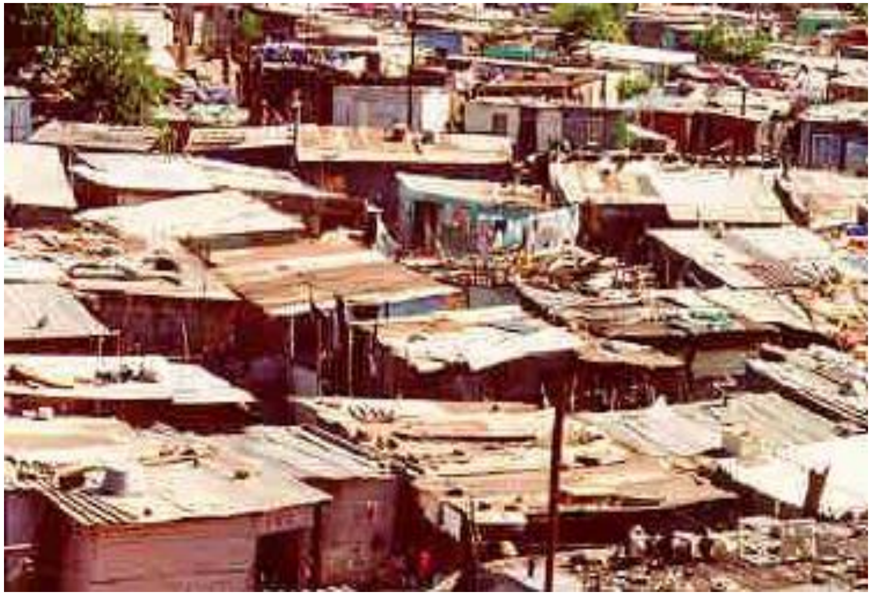
Data Created
2021

[Source: Cisco Global Cloud Index, 2016-2021]

Computer Architectures

- Computer architecture – The conceptual design and fundamental operational structure of a computer system.
  - CPU and Instruction Set
  - Access mode to memory
  - Components and their interconnection
  - …

Habitat '67, by Moshe Safdie, at Montreal, Canada, 1967 © Artifice, Inc

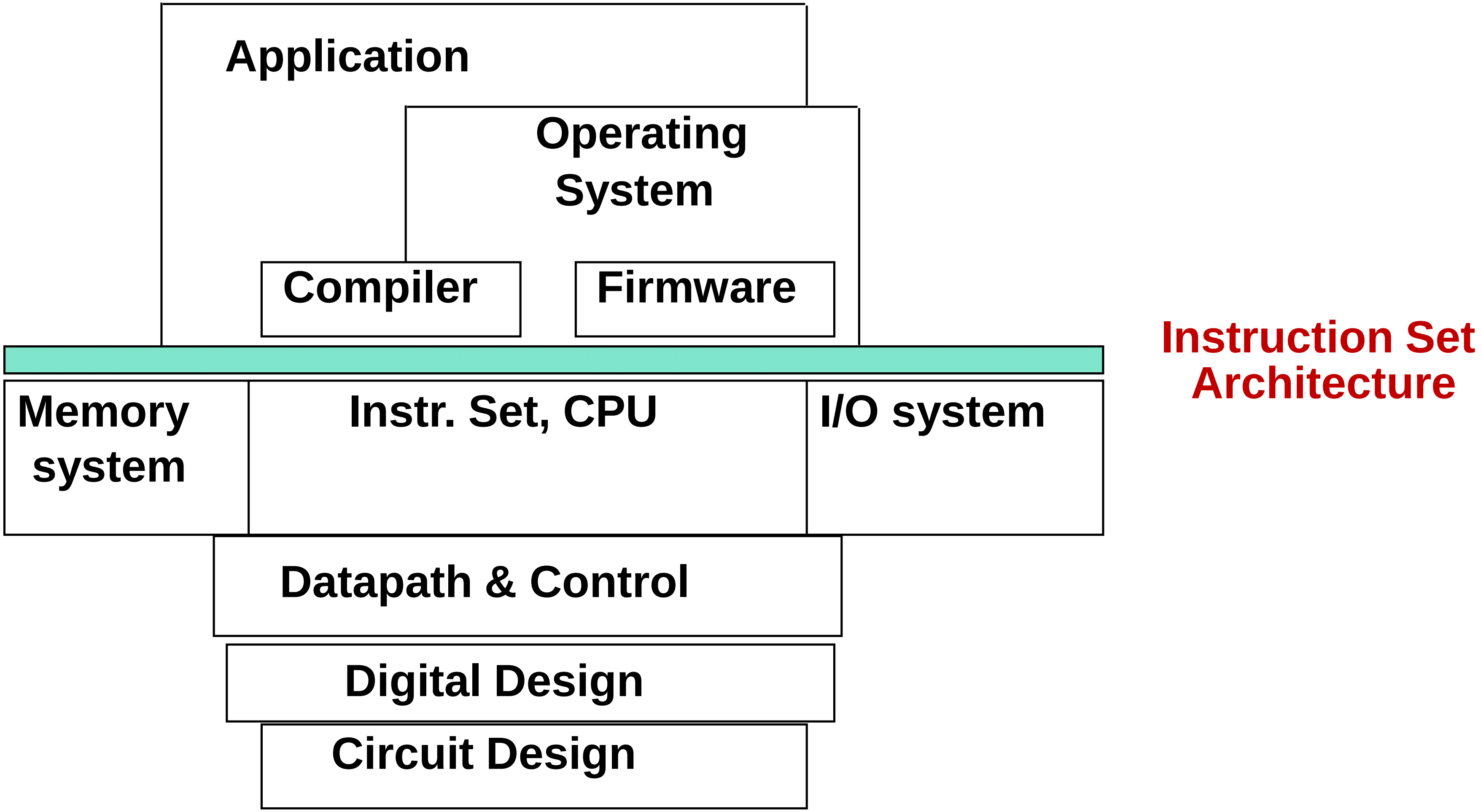


[Sangiovanni-Vincentelli Vincentelli 04]

What is Computer Architecture?

- **Computer Architecture is those aspects of the instruction set available to programmers, independent of the hardware on which the instruction set was implemented.**
- **The term computer architecture was first used in 1964 by Gene Amdahl, G. Anne Blaauw, and Frederick Brooks, Jr., the designers of the IBM System/360.**
- **The IBM/360 was a family of computers all with the same architecture, but with a variety of organizations(implementations).**

Where “Computer Architectures” Are?



Defining Computer Architecture

- **“Old” view of computer architecture:**
  - Instruction Set Architecture (ISA) design
  - i.e. decisions regarding:
    - registers, memory addressing, addressing modes, instruction operands, available operations, control flow instructions, instruction encoding

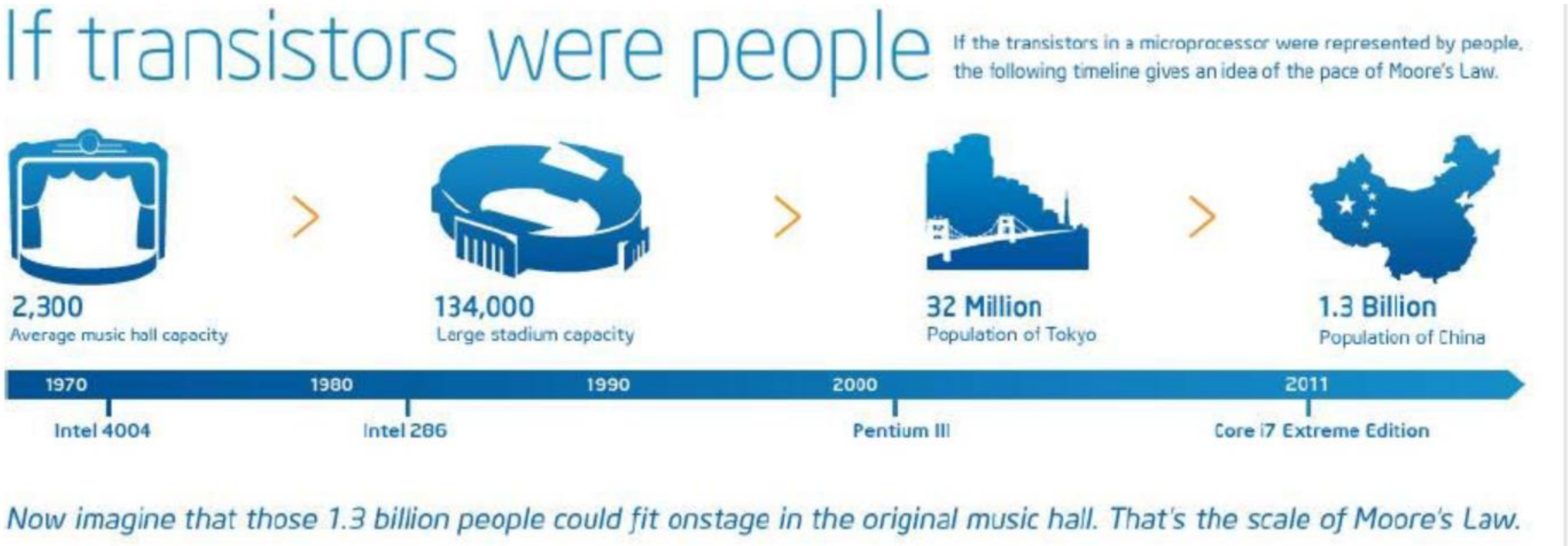
- **“Real” computer architecture:**
  - Specific requirements of the target machine
  - Design to maximize performance within constraints: cost, power, and availability
  - Includes ISA, microarchitecture, hardware

Computer Technology

- **Performance improvements:**
  - Improvements in semiconductor technology
    - Integrated circuit logic, DRAM, Flash, Disk
    - Scaling of transistor: Feature size, clock speed
  - Improvements in computer architectures
    - Enabled by HLL compilers, UNIX
    - Lead to RISC architectures
  - Together have enabled:
    - Lightweight computers
    - Productivity-based managed/interpreted programming languages

Moore's Law

- **Gordon Moore, one of the founders of Intel**
  - In 1965 he predicted the doubling of the number of transistors per chip every couple of years for the next ten years
  - http://www.intel.com/research/silicon/mooreslaw.htm

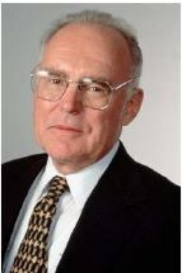


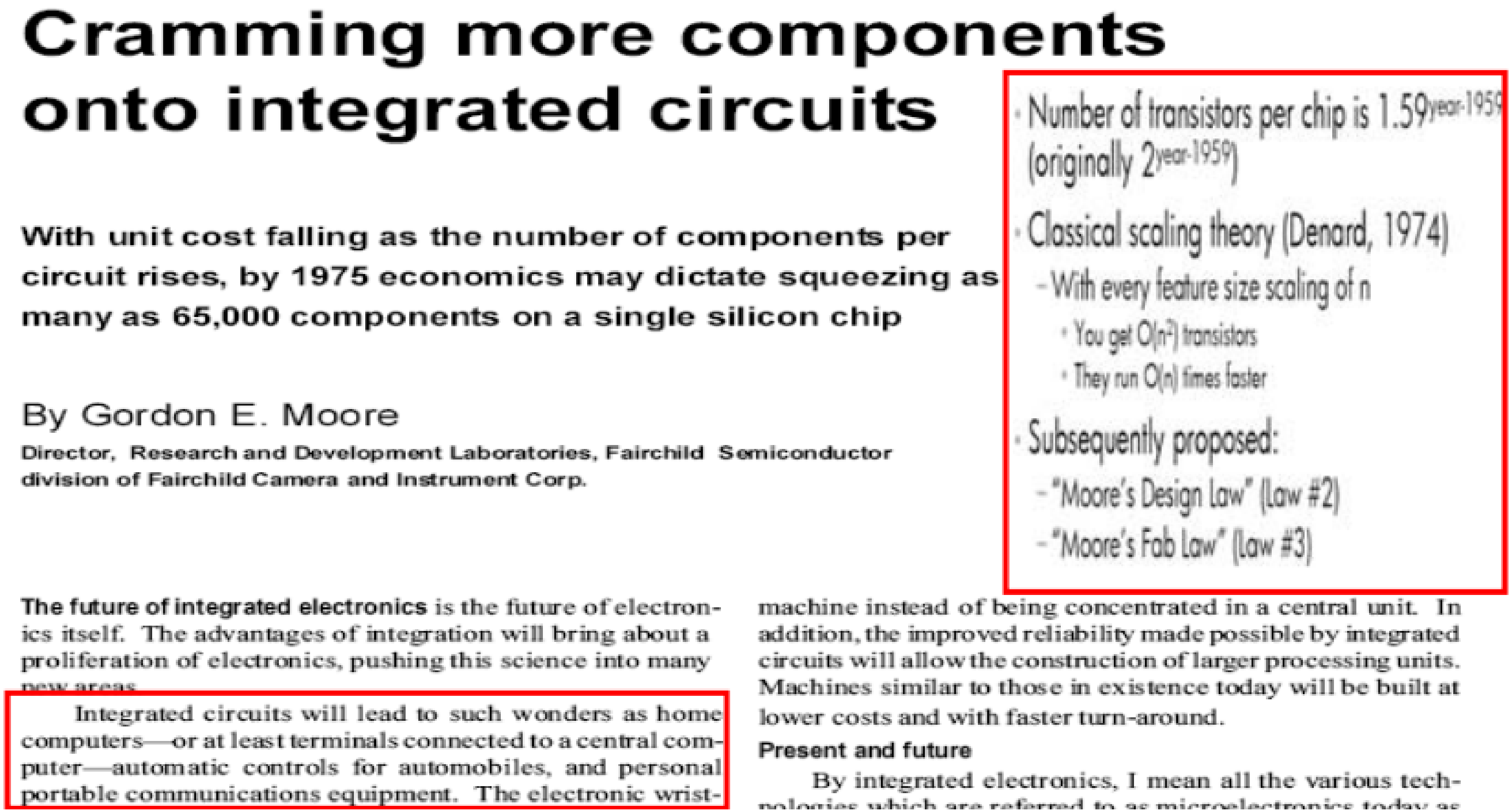
If transistors were people

If the transistors in a microprocessor were represented by people, the following timeline gives an idea of the pace of Moore's Law.

Now imagine that those 1.3 billion people could fit onstage in the original music hall. That's the scale of Moore's Law.

Moore's Law [Electronics, April 19, 1965]





Cramming more components onto integrated circuits

With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip

By Gordon E. Moore

Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.

- Number of transistors per chip is $1.59^{year-1959}$ (originally $2^{year-1959}$)
- Classical scaling theory (Denard, 1974)
  - With every feature size scaling of n
    - You get $O(n^2)$ transistors
    - They run $O(n)$ times faster
- Subsequently proposed:
  - "Moore's Design Law" (Law #2)
  - "Moore's Fab Law" (Law #3)

The future of integrated electronics is the future of electronics itself. The advantages of integration will bring about a proliferation of electronics, pushing this science into many new areas.

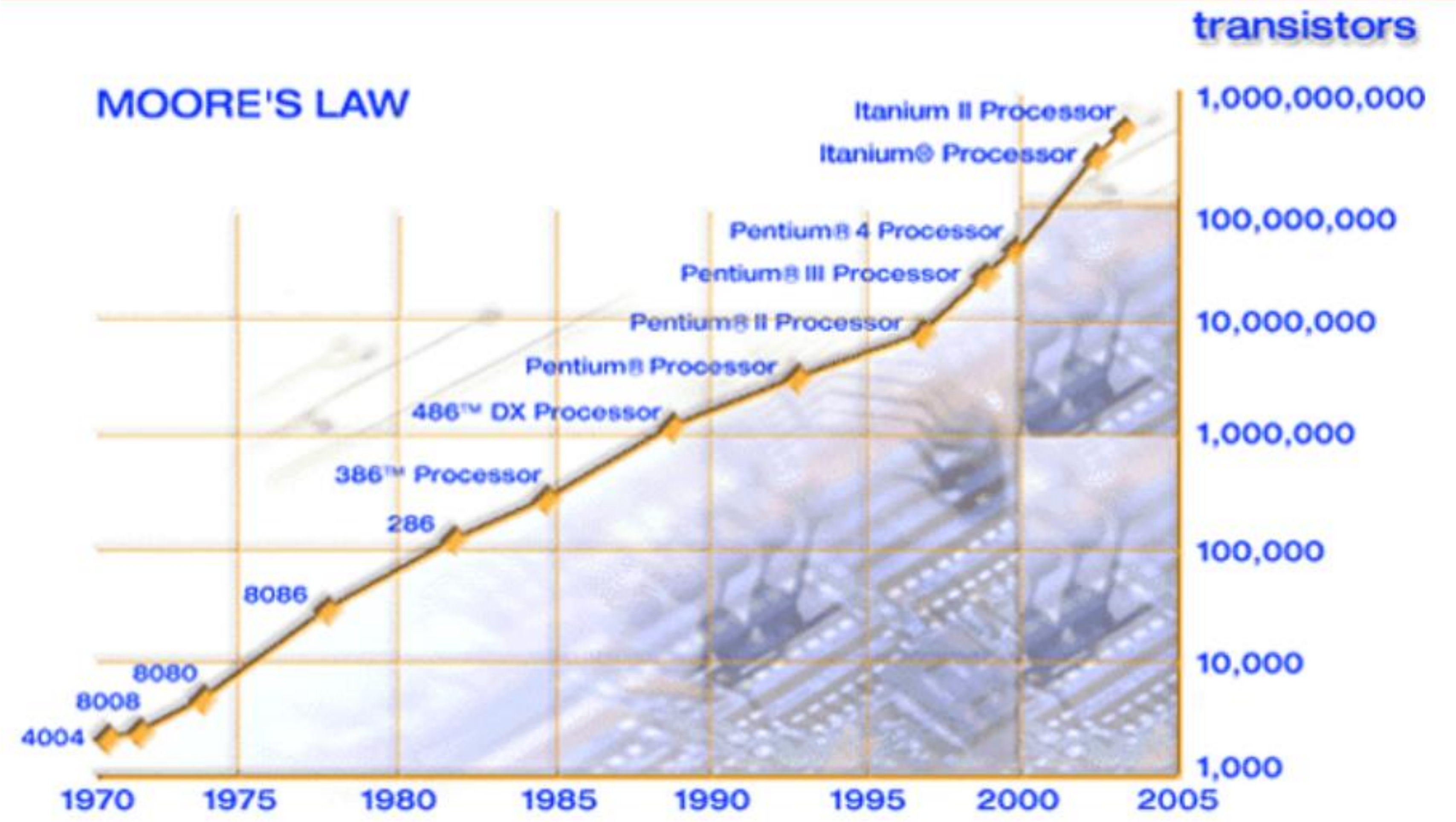
Integrated circuits will lead to such wonders as home computers—or at least terminals connected to a central computer—automatic controls for automobiles, and personal portable communications equipment. The electronic wristwatch needs only a display to be feasible today. machine instead of being concentrated in a central unit. In addition, the improved reliability made possible by integrated circuits will allow the construction of larger processing units. Machines similar to those in existence today will be built at lower costs and with faster turn-around.

Present and future

By integrated electronics, I mean all the various technologies which are referred to as microelectronics today as

http://www.intel.com/research/silicon/mooreslaw.htm

Moore’s Law [Intel Microprocessors]



http://www.intel.com/research/silicon/mooreslaw.htm

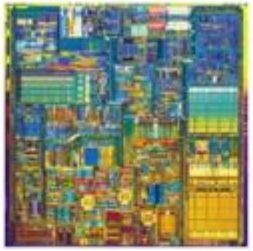
Technology Scaling and IT Industry Progress

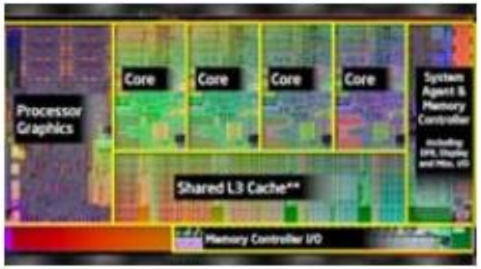
- 1971 (10μ)
- 4004 μP
- 5k trans.
- 4mm^2 die
- 108Khz
- 0.2W
- Busicom calculator

- 1982 (1.5μ)
- 286 μP
- 134k trans.
- 47mm^2 die
- 8Mhz
- 3W
- 15M PCs sold in 6yr

- 1994 (0.6μ)
- Pentium®
- 3.2M trans.
- 147mm^2 die
- 100Mhz
- 10W
- sound, images

- 2000 (0.18μ)
- Pentium® 4
- 42M trans.
- 217mm^2 die
- 1.5Ghz
- 58W
- internet



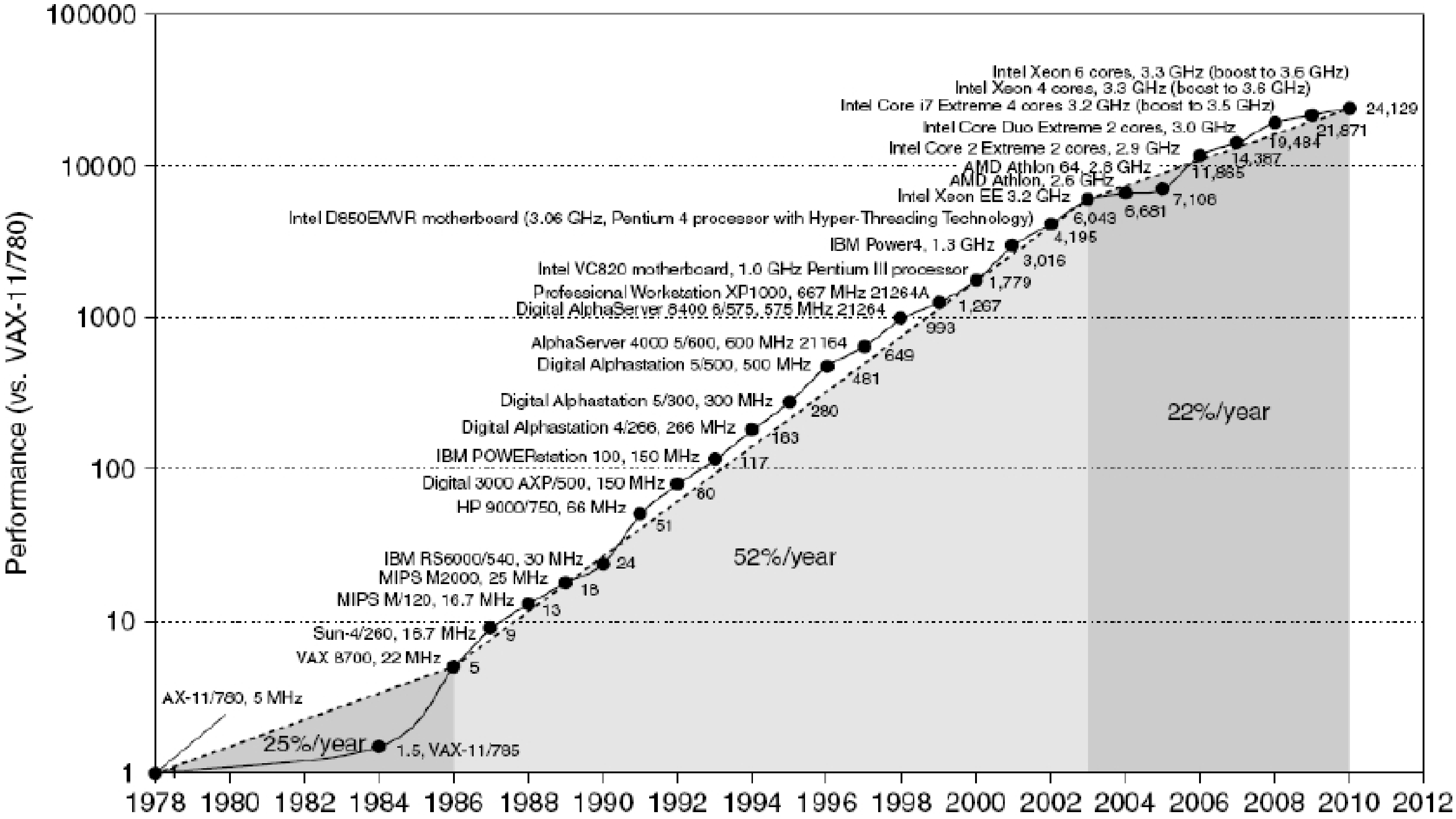


- 2011 (0.035μ)
- Core® (Sandy Bridge)
- 995M trans.
- 216mm^2 die
- 3.6Ghz
- 95W
- content creation, immersive gaming, pervasive computing

Transistors and Wires

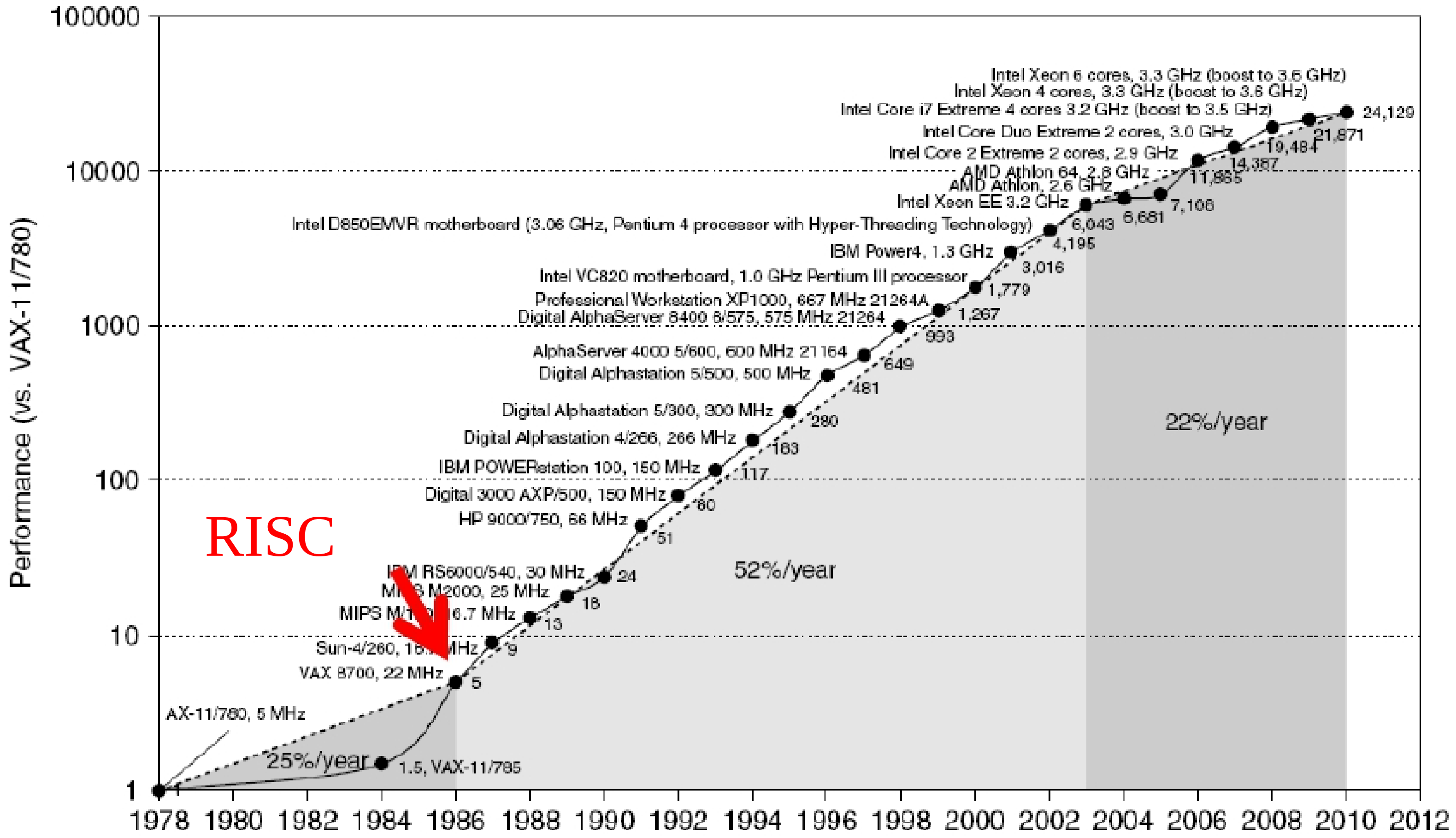
- **Feature size**
  - Minimum size of transistor or wire in x or y dimension
  - 10 microns in 1971 to .032 microns in 2011
  - Transistor performance scales linearly
    - Wire delay does not improve with feature size!
  - Integration density scales quadratically

Sequential Processor Performance



From Hennessy and Patterson Ed. 5 Image Copyright © 2011, Elsevier Inc. All rights Reserved.

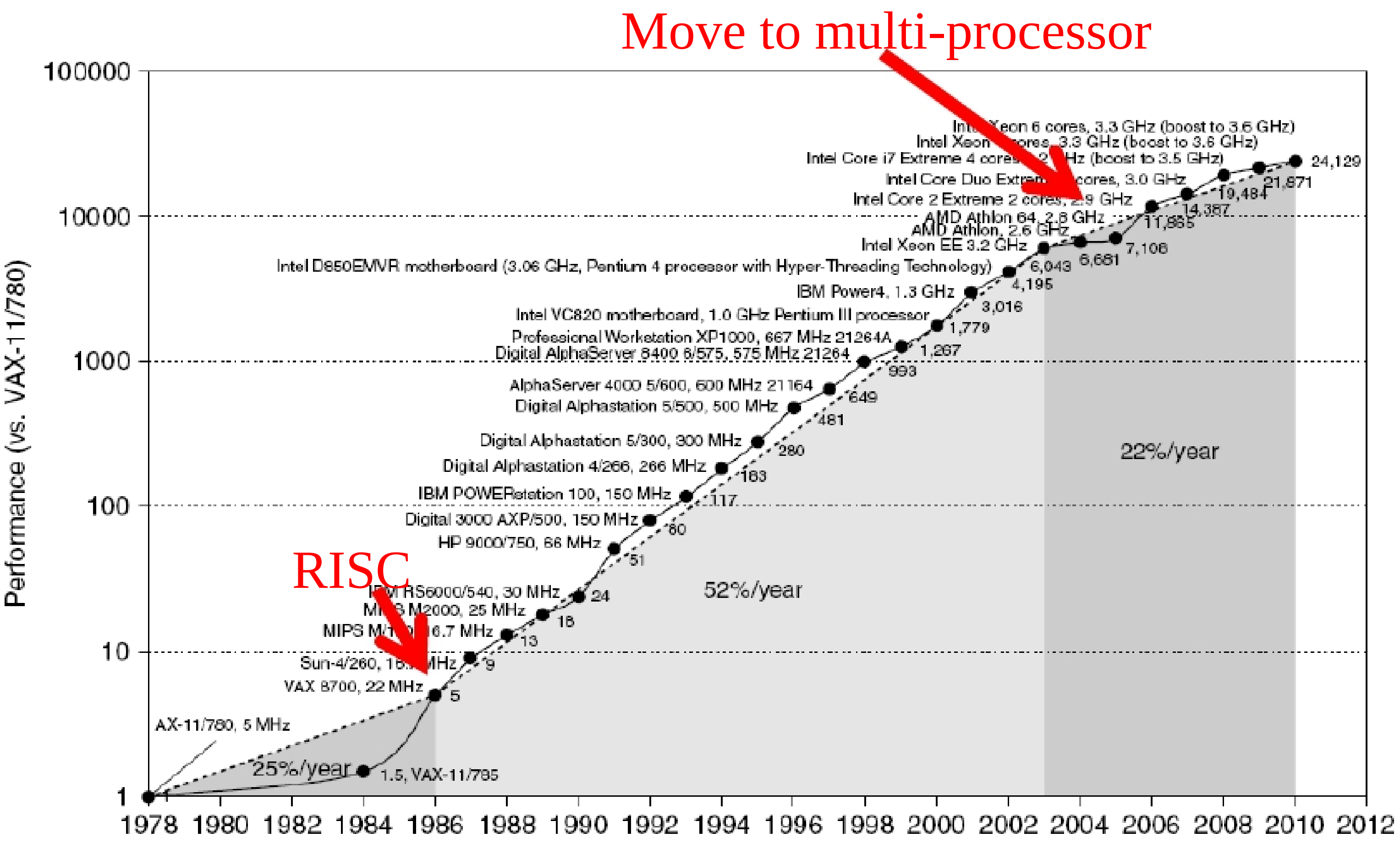
Sequential Processor Performance



RISC

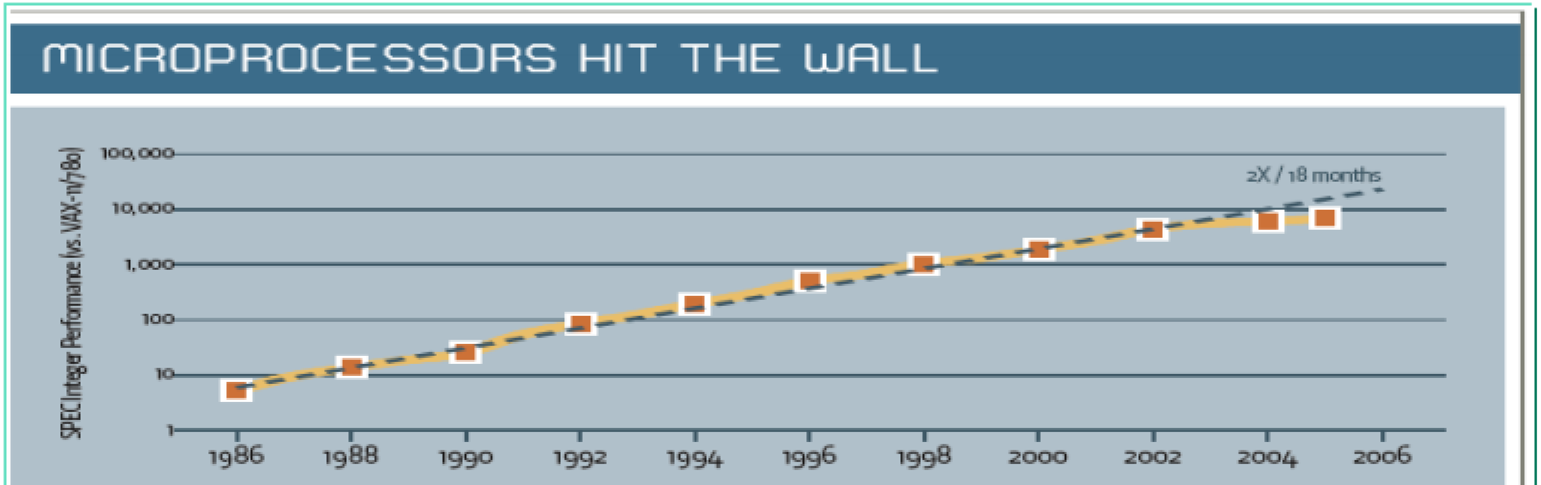
From Hennessy and Patterson Ed. 5 Image Copyright © 2011, Elsevier Inc. All rights Reserved.

Sequential Processor Performance



From Hennessy and Patterson Ed. 5 Image Copyright © 2011, Elsevier Inc. All rights Reserved.

Microprocessor Hit the Power Wall

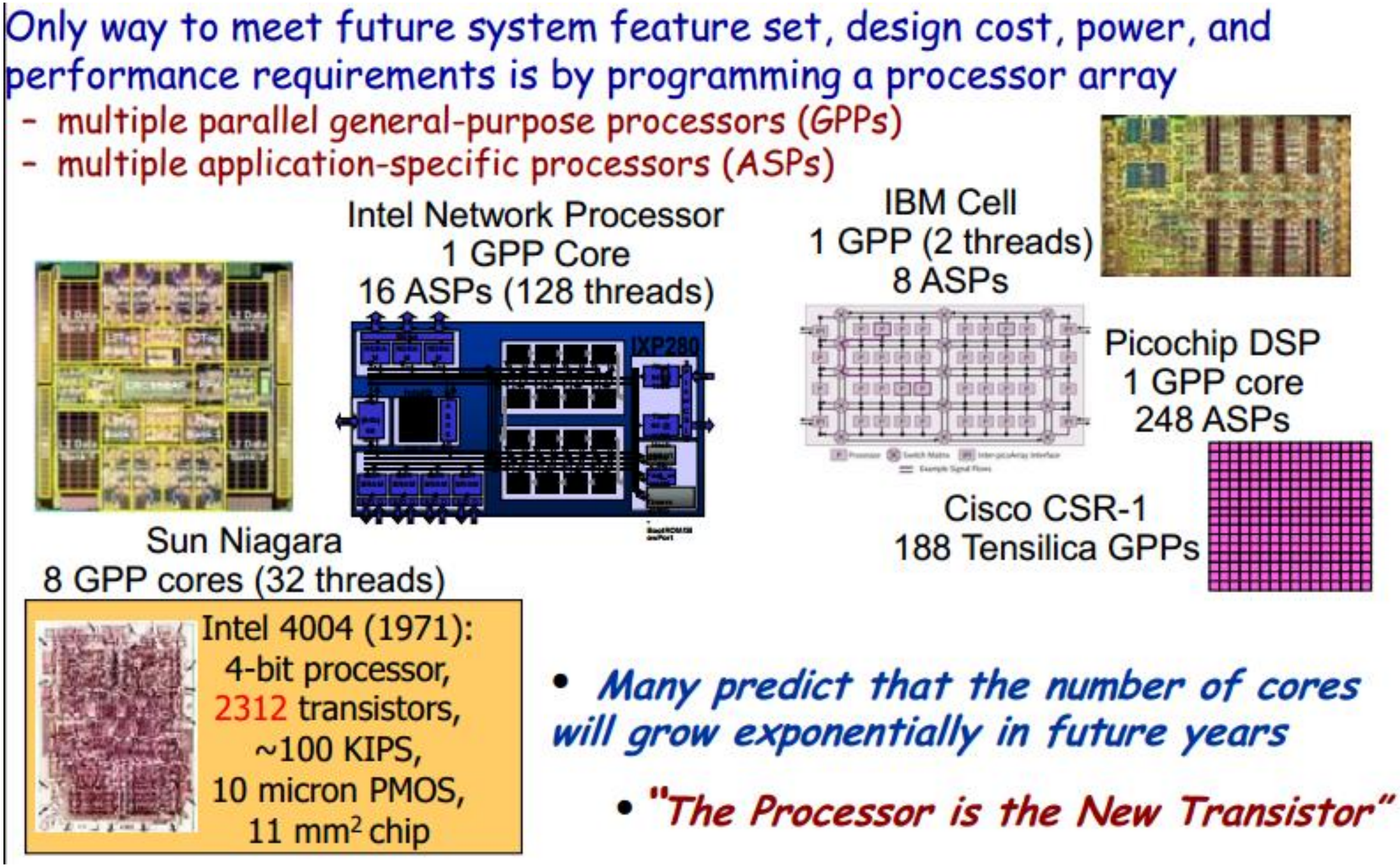


MICROPROCESSORS HIT THE WALL

Due to power dissipation, microprocessors can't sustain the pace set by Moore's Law: that is, doubling performance every 18 months. Parallelism promises greater efficiency—as long as programmers know how to exploit it.

– “The media's mischaracterization of Moore's Law is now evident. Gordon Moore predicted the regular doubling of the number of transistors on a chip. The job of computer architects was to turn twice as many transistors into twice as much performance . Between 1986 and 2002 architects succeeded, and we saw the greatest sustained increase in performance in computing history. The problem was that they kept increasing the power dissipated per chip, and in 2004 it was obvious that the industry had hit a power wall. Today, microprocessors are about a factor of three slower than if we could keep increasing power and doubling performance every 18 months Thus, while Moore's Law continues, power dissipation hit the wall” [D. A. Patterson 2007]

The Processor is the New Transistor



Only way to meet future system feature set, design cost, power, and performance requirements is by programming a processor array

- multiple parallel general-purpose processors (GPPs)
- multiple application-specific processors (ASPs)

Intel Network Processor
1 GPP Core
16 ASPs (128 threads)

IBM Cell
1 GPP (2 threads)
8 ASPs

Picochip DSP
1 GPP core
248 ASPs

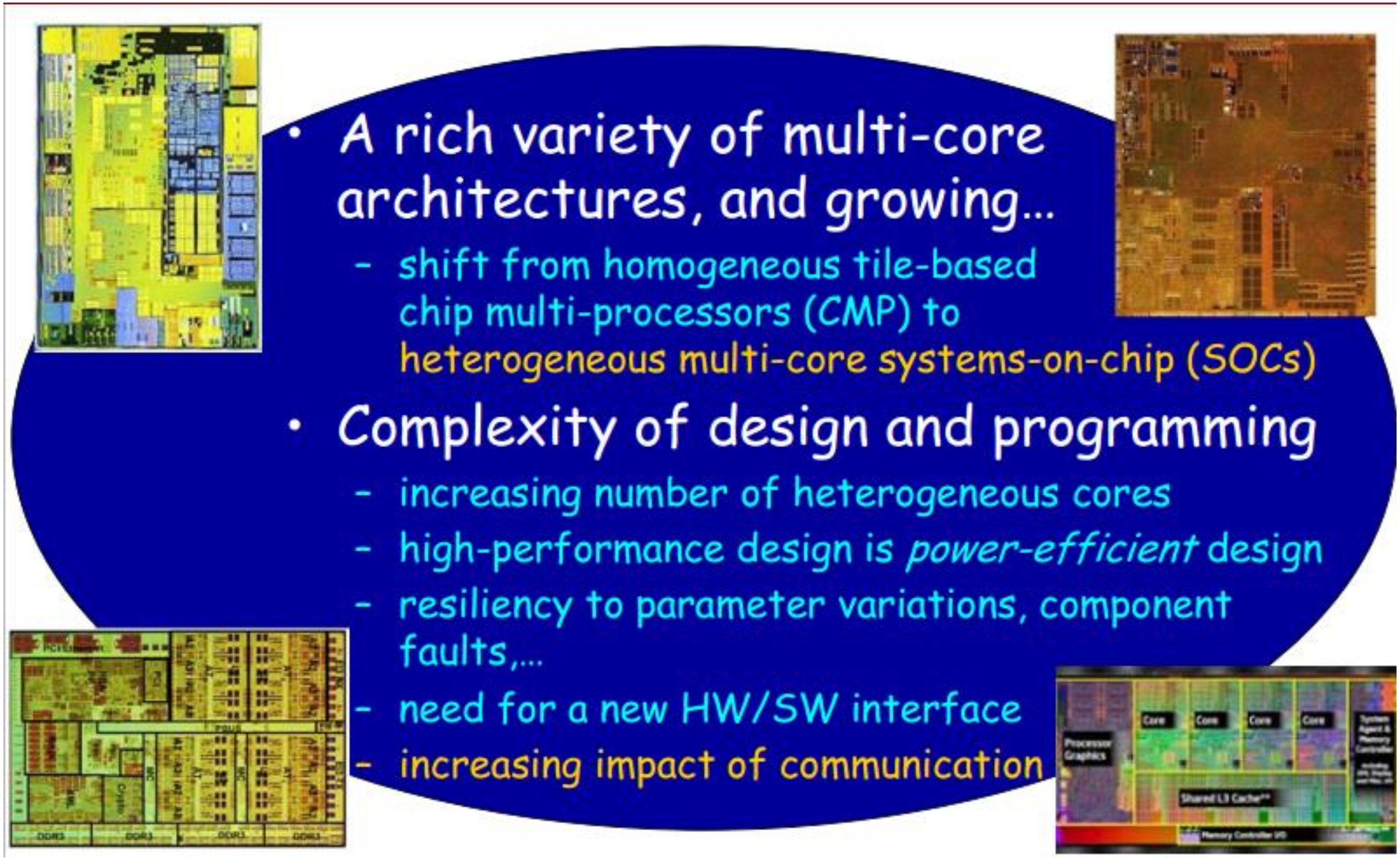
Cisco CSR-1
188 Tensilica GPPs

Sun Niagara
8 GPP cores (32 threads)

Intel 4004 (1971):
4-bit processor,
2312 transistors,
~100 KIPS,
10 micron PMOS,
11 mm^2 chip

- *Many predict that the number of cores will grow exponentially in future years*
    - *"The Processor is the New Transistor"*

The Era of Heterogeneous Multi-Core Systems-on-Chip (SoC)



- A rich variety of multi-core architectures, and growing…
  - shift from homogeneous tile-based chip multi-processors (CMP) to heterogeneous multi-core systems-on-chip (SOCs)
- Complexity of design and programming
  - increasing number of heterogeneous cores
  - high-performance design is *power-efficient* design
  - resiliency to parameter variations, component faults,…
  - need for a new HW/SW interface
  - increasing impact of communication

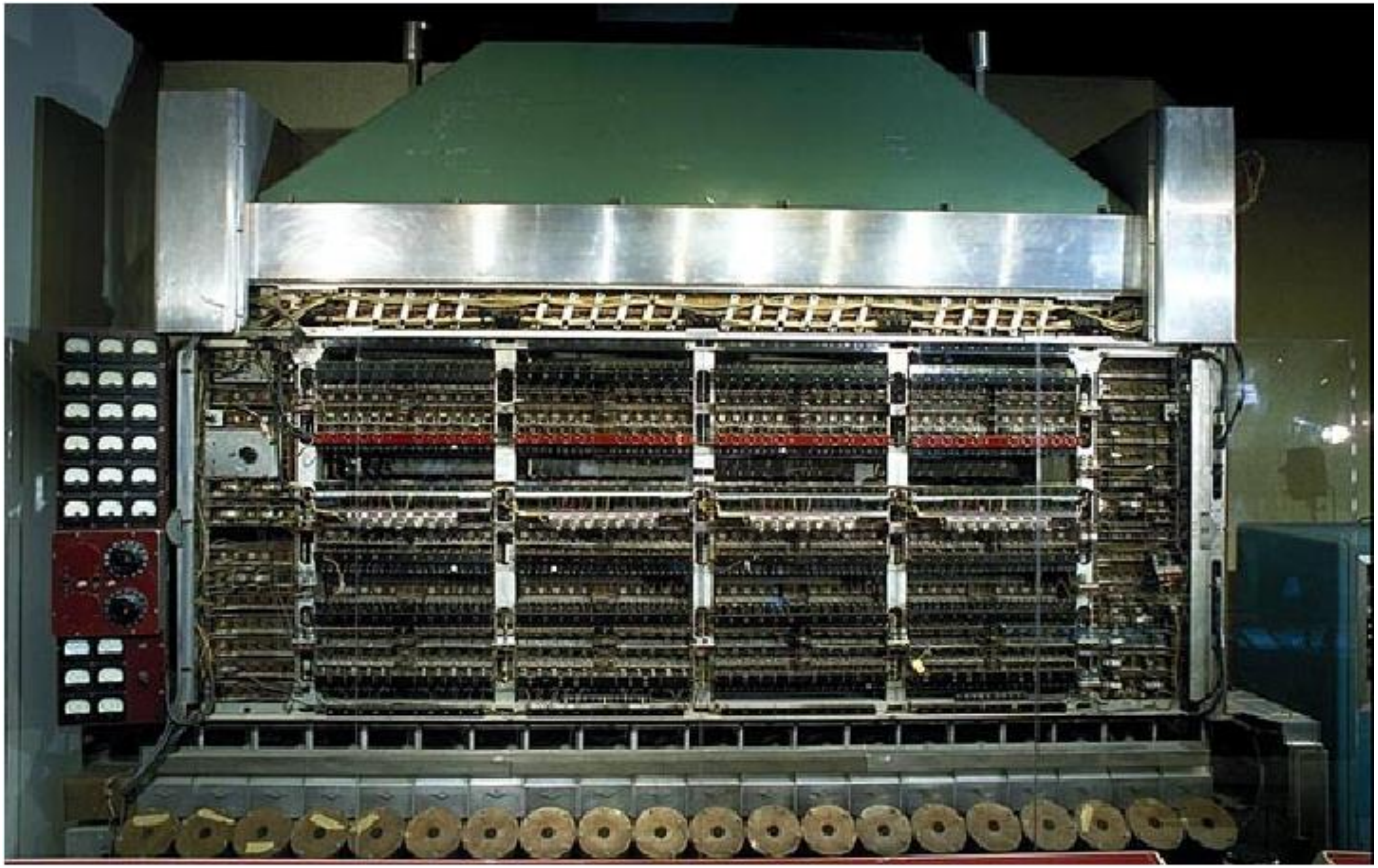
Current Trends in Architecture

- **Cannot continue to leverage Instruction-Level parallelism (ILP)**
  - Single processor performance improvement ended in 2003

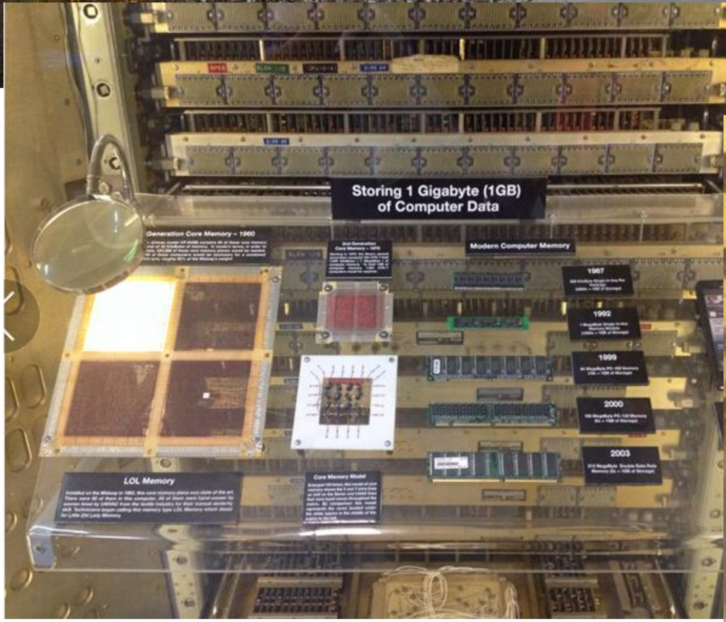
- **New models for performance:**
  - Data-level parallelism (DLP)
  - Thread-level parallelism (TLP)
  - Request-level parallelism (RLP)

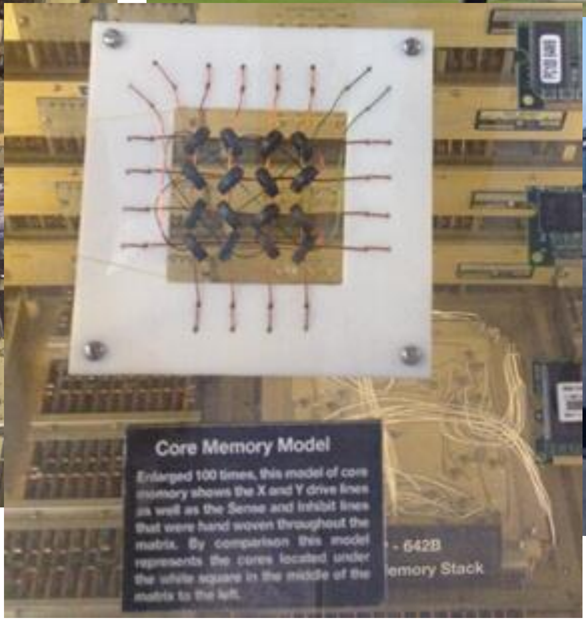
- **These require explicit restructuring of the application**

Computers Then…



IAS Machine. Design directed by John Von Nuemann.
First booted in Princeton NJ in 1952
Smithsonian Institution Archives (Smithsonian Image 95-06151)

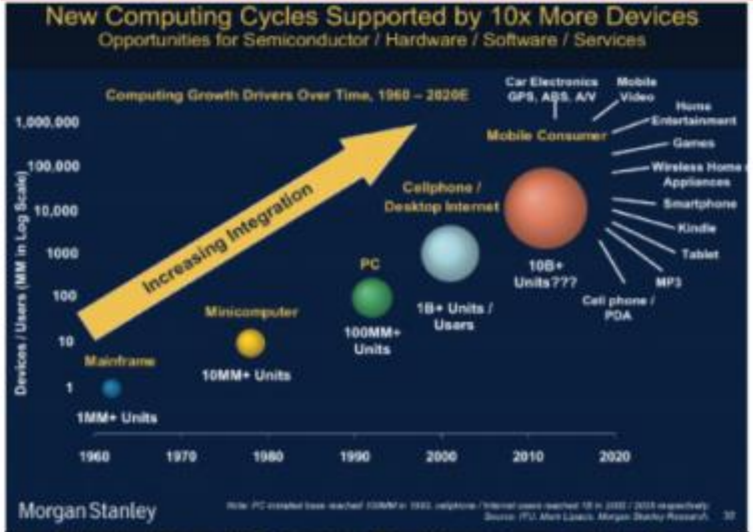

Storing 1 Gigabyte (1GB) of Computer Data
Modern Computer Memory
1987
1992
1999
2000
2003
LOL Memory
Core Memory Model

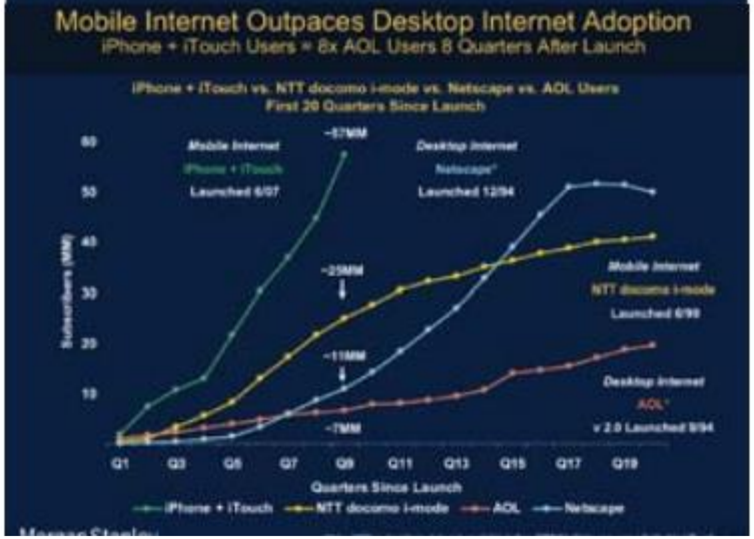

Core Memory Model
Enlarged 100 times, this model of core memory shows the X and Y drive lines as well as the Sense and Inhibit lines that were hand woven throughout the matrix. By comparison this model represents the cores located under the white square in the middle of the matrix to the left.
- 642B
Memory Stack

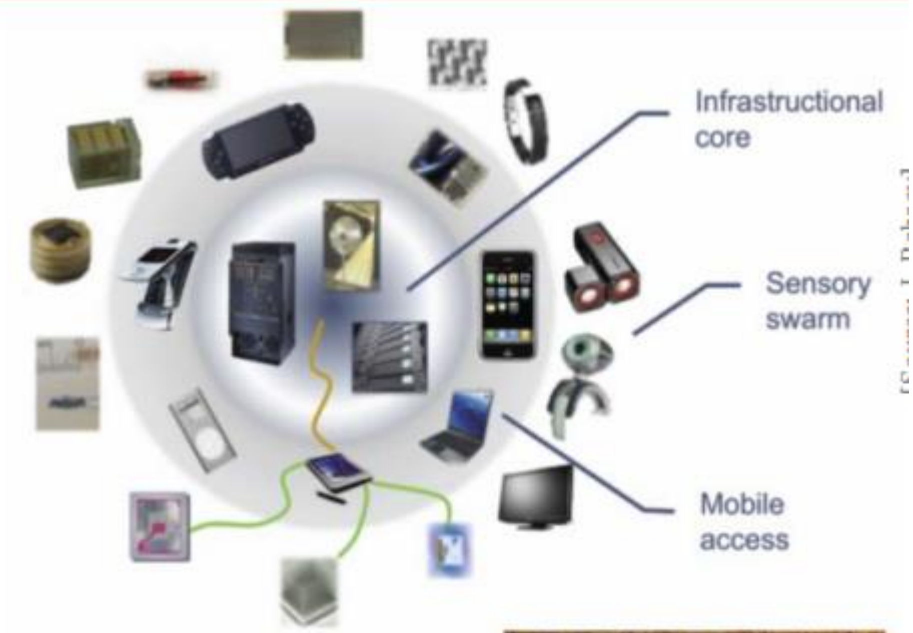
Computers Now

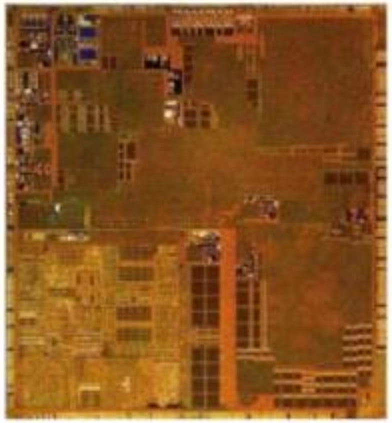
- Sensor Networks
- Cameras
- Smartphones
- Mobile Audio Players
- Laptops
- Autonomous Cars
- Servers
- Game Players
- Routers
- Flying UAVs
- GPS
- eBooks
- Tablets
- Set-top Boxes

The Emerging IT Scene and The Emerging Computing Platform

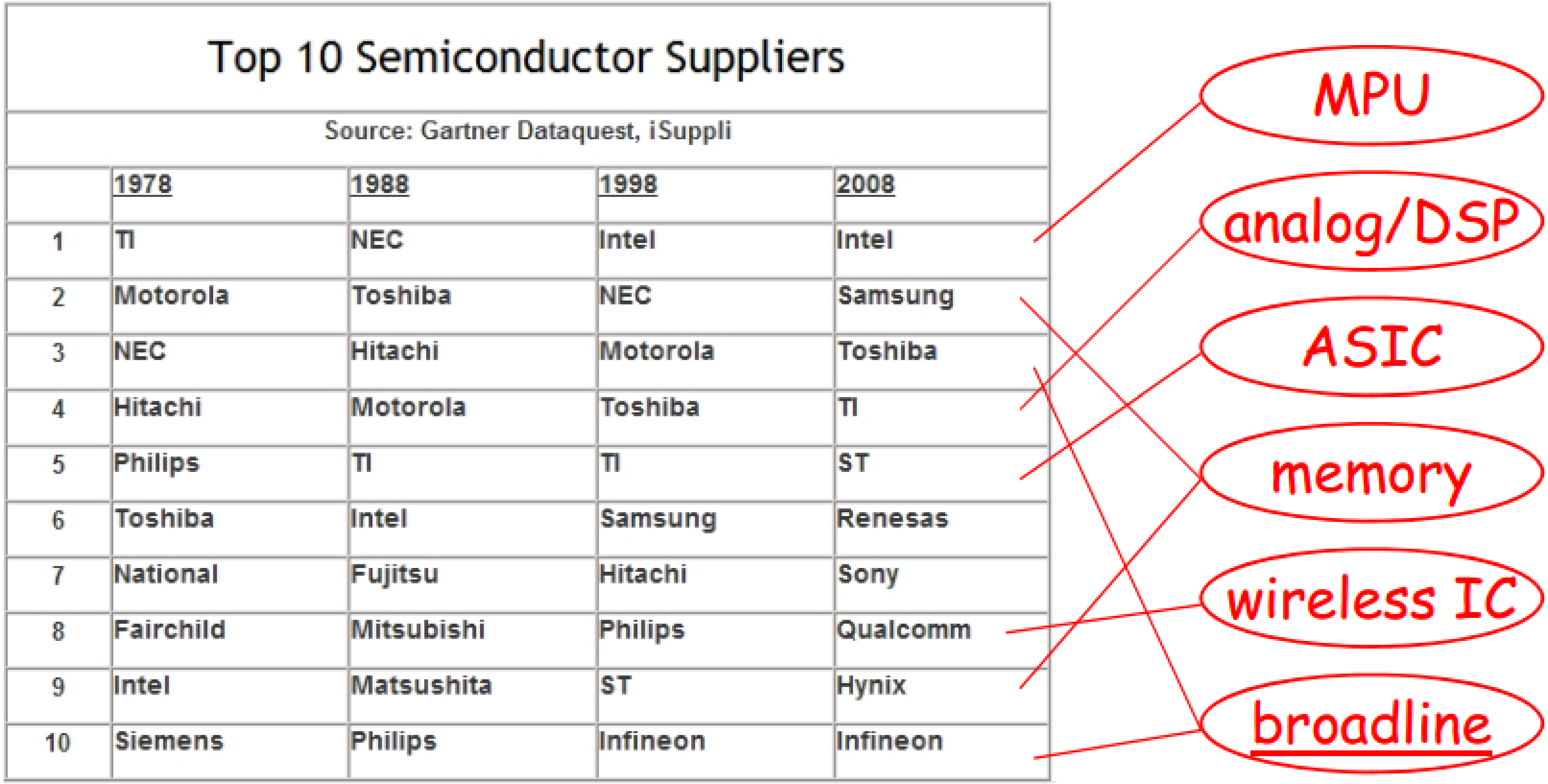








Top 10 Semiconductor Suppliers over the Last Forty Years



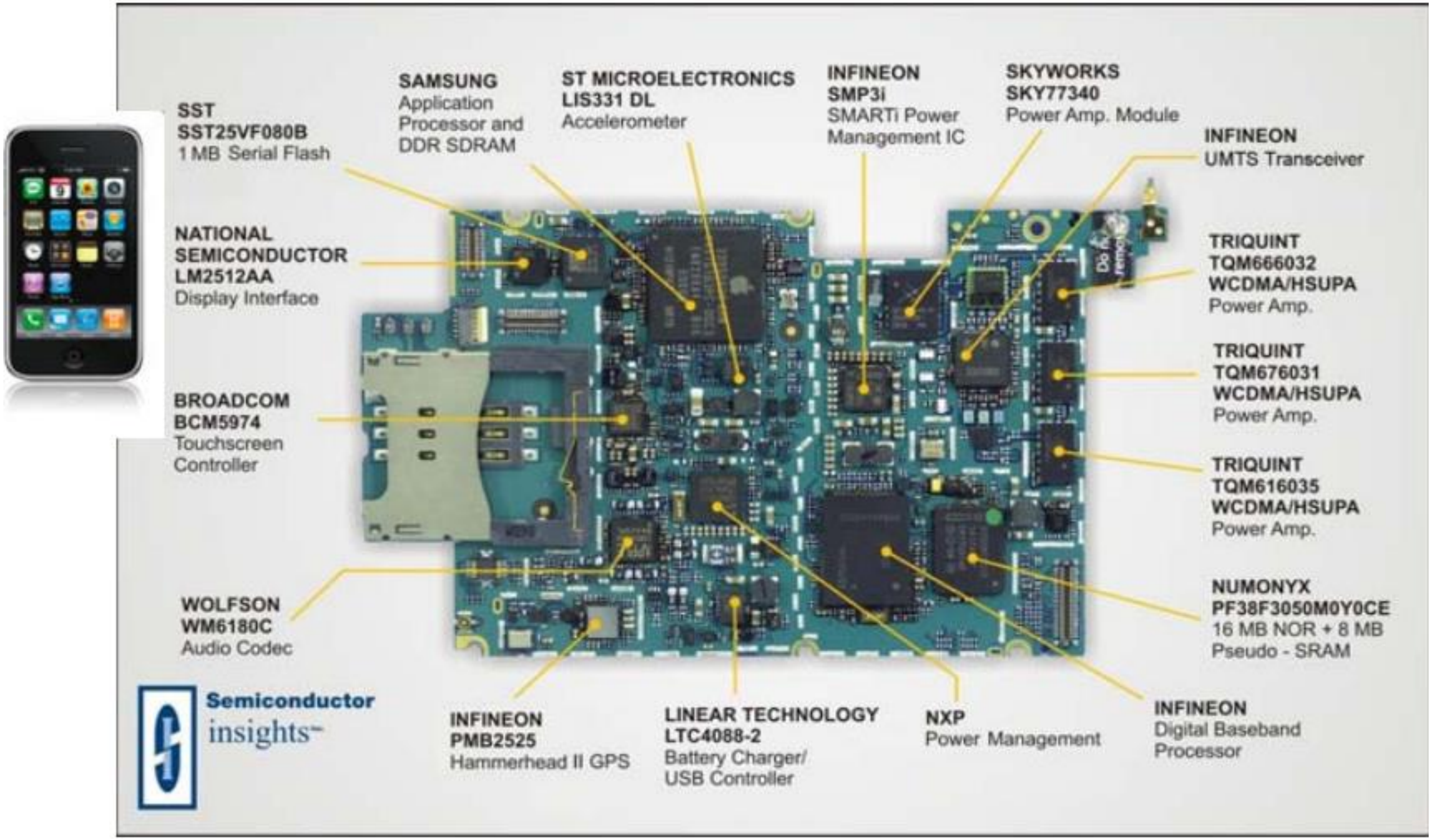
Top 10 Semiconductor Suppliers

Source: Gartner Dataquest, iSuppli

	1978	1988	1998	2008
1	TI	NEC	Intel	Intel
2	Motorola	Toshiba	NEC	Samsung
3	NEC	Hitachi	Motorola	Toshiba
4	Hitachi	Motorola	Toshiba	TI
5	Philips	TI	TI	ST
6	Toshiba	Intel	Samsung	Renesas
7	National	Fujitsu	Hitachi	Sony
8	Fairchild	Mitsubishi	Philips	Qualcomm
9	Intel	Matsushita	ST	Hynix
10	Siemens	Philips	Infineon	Infineon

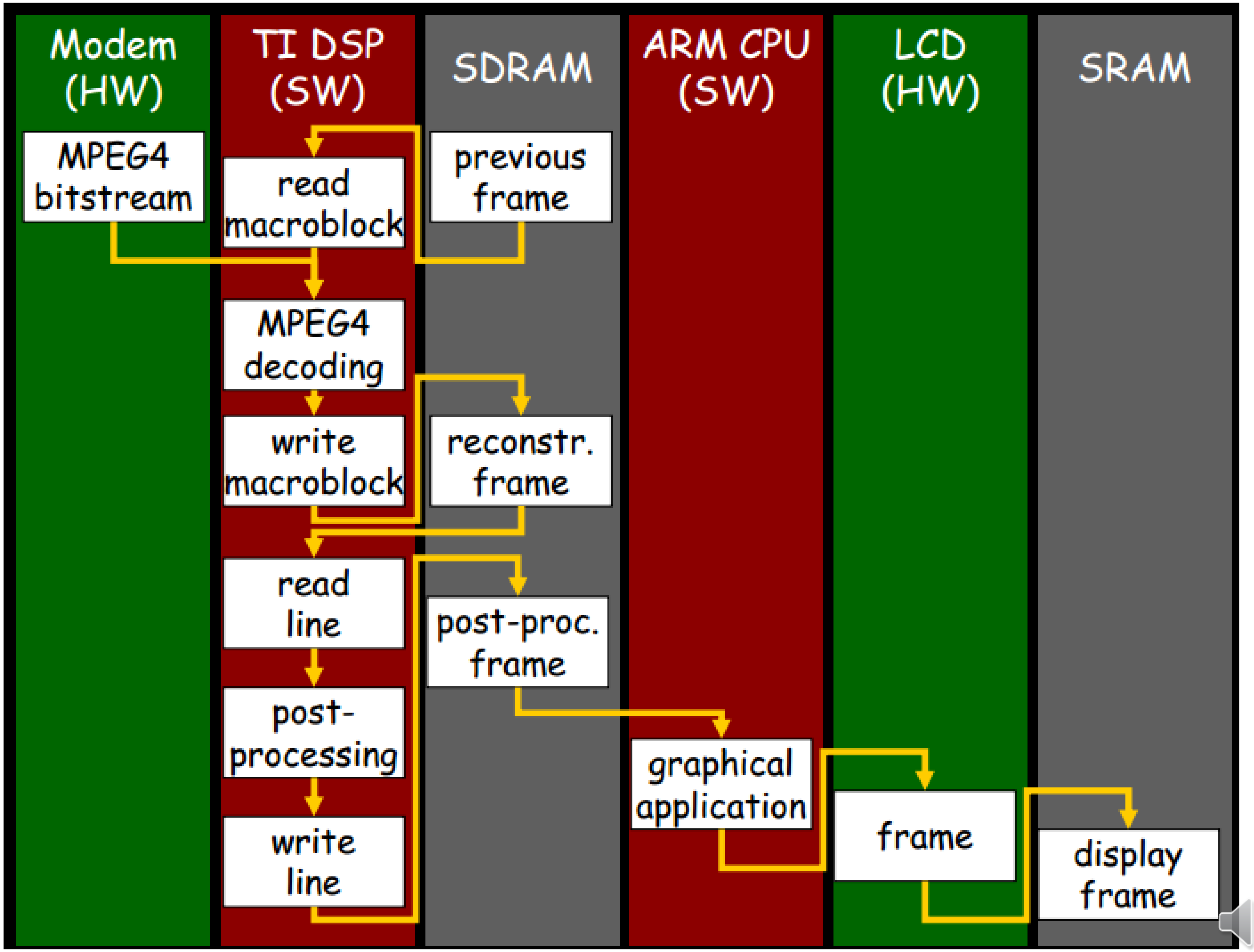
Trends towards focused companies (as opposed to having a broad product line) due to high cost of fabs, R&D, and IC design

Inside a Mobile Phone: the Apple I-Phone 3G

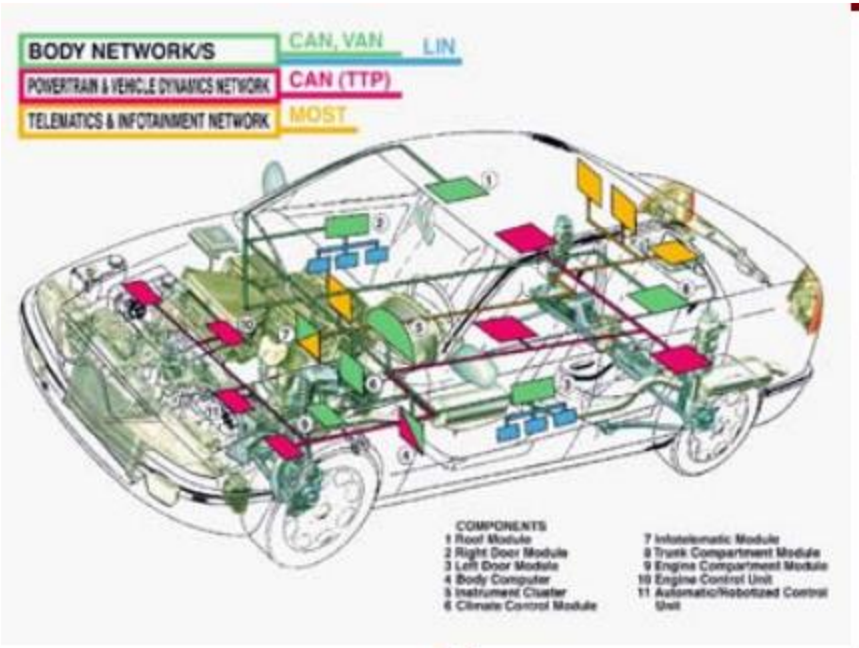


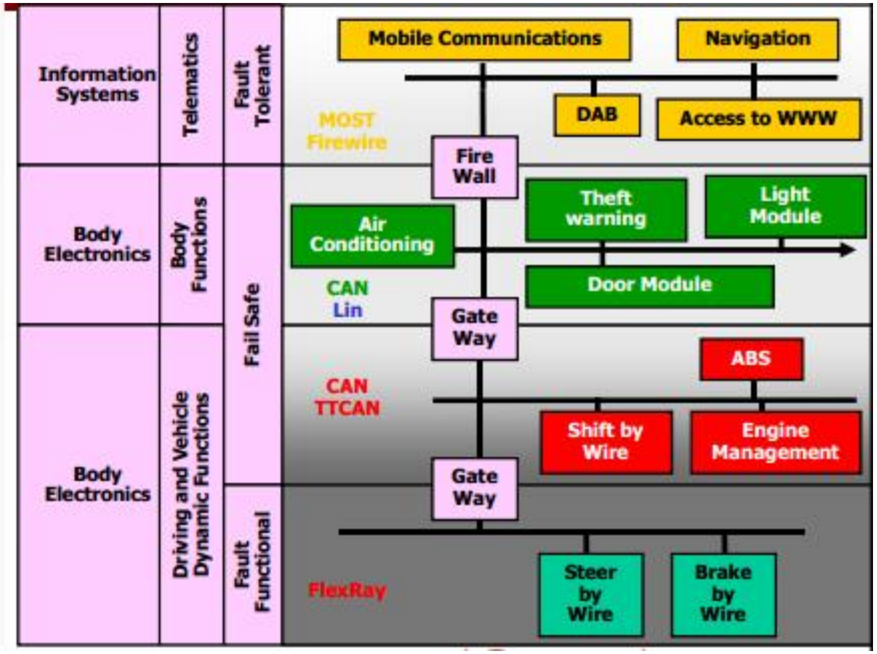
Heterogeneous Systems-on-Chip: SOCs for High-End Wireless Phones Market

- Dual-Core
  - ARM CPU
    - general purpose tasks
    - OS/UI processing
  - TI DSP
    - real-time signal processing tasks
- Example
  - basic dataflow for video decoding



Heterogeneous Embedded Systems: Electronics for the Car





- Up to 70 Electronic Computing Units (ECUs) in a modern car like a BMW Series 7
  - Heterogeneous communication networks
  - DSC (dynamic stability control) contains ABS as one of 15 sub-functionalities

Three Main Computing Classes (Year 2000)

Feature	Desktop	Server	Embedded
Price of system	\$500- \$10,000	\$10,000 - \$10,000,000	\$10-\$100,000 (including routers)
Price of μP module	\$100-\$1000	\$200-\$2000 (per μP)	\$0.2-\$200 (per μP)
μP sales per year (2000)	150M	4M	300M (32/64-bit only)
Critical Design issues	price/perf. graphics performance	throughput, availability, scalability	price, power-consumption, “performance”

Classes of Computers (by 2010)

- **Personal Mobile Device (PMD)**
  - e.g. smart 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**
  - Used for “Software as a Service (SaaS)”
  - Emphasis on availability and price-performance
  - Sub-class: Supercomputers, emphasis: floating-point performance and fast internal networks
- **Embedded Computers**
  - Emphasis: price, power

Parallelism

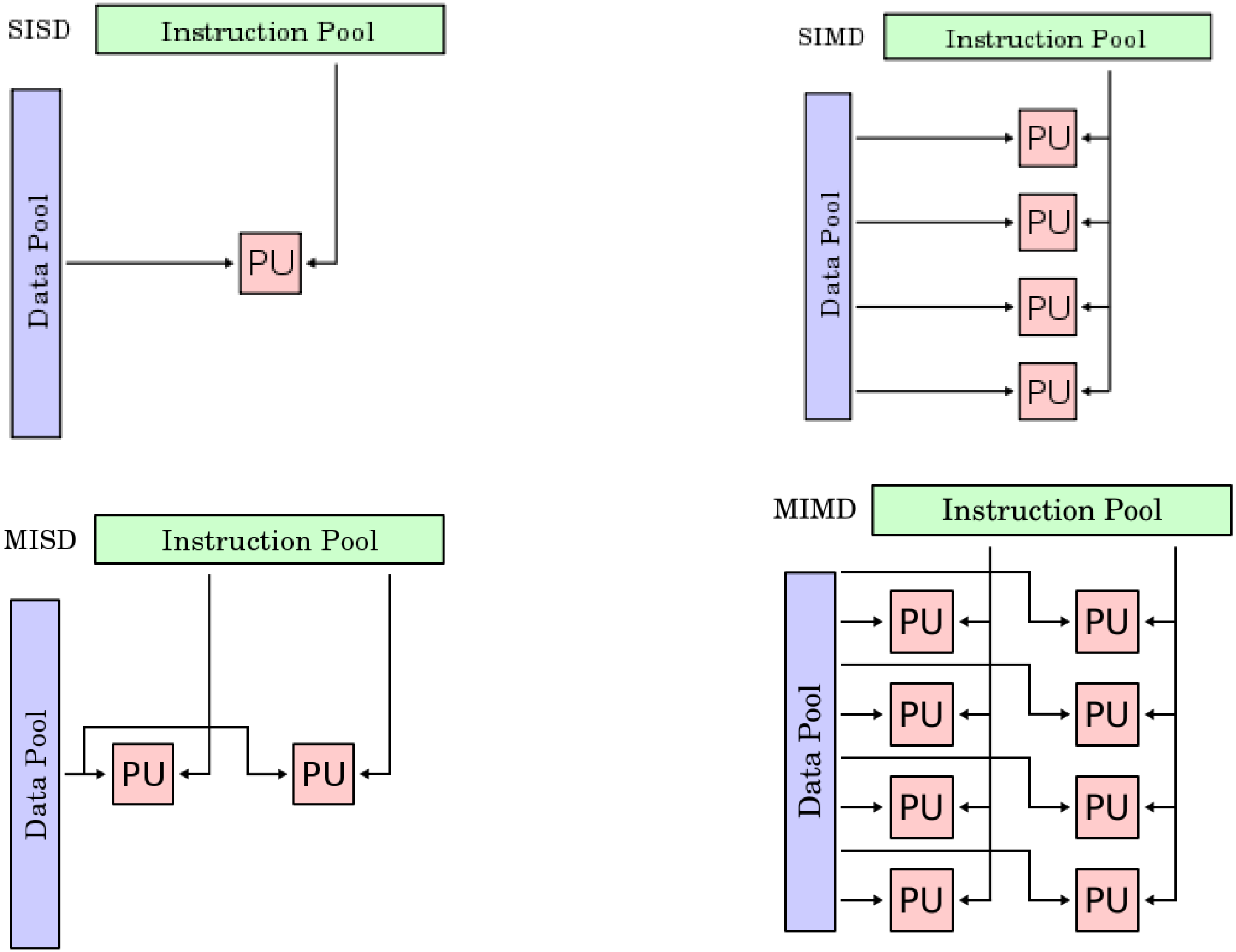
- **Classes of parallelism in applications:**
  - Data-Level Parallelism (DLP)
  - Task-Level Parallelism (TLP)

- **Classes of architectural parallelism:**
  - Instruction-Level Parallelism (ILP)
  - Vector architectures/Graphic Processor Units (GPUs)
  - Thread-Level Parallelism
  - Request-Level Parallelism

Flynn’s Taxonomy

- **Single instruction stream, single data stream (SISD)**
- **Single instruction stream, multiple data streams (SIMD)**
  - Vector architectures
  - Multimedia extensions
  - Graphics processor units
- **Multiple instruction streams, single data stream (MISD)**
  - No commercial implementation
- **Multiple instruction streams, multiple data streams (MIMD)**
  - Tightly-coupled MIMD
  - Loosely-coupled MIMD

Flynn’s Taxonomy



Trends in Technology

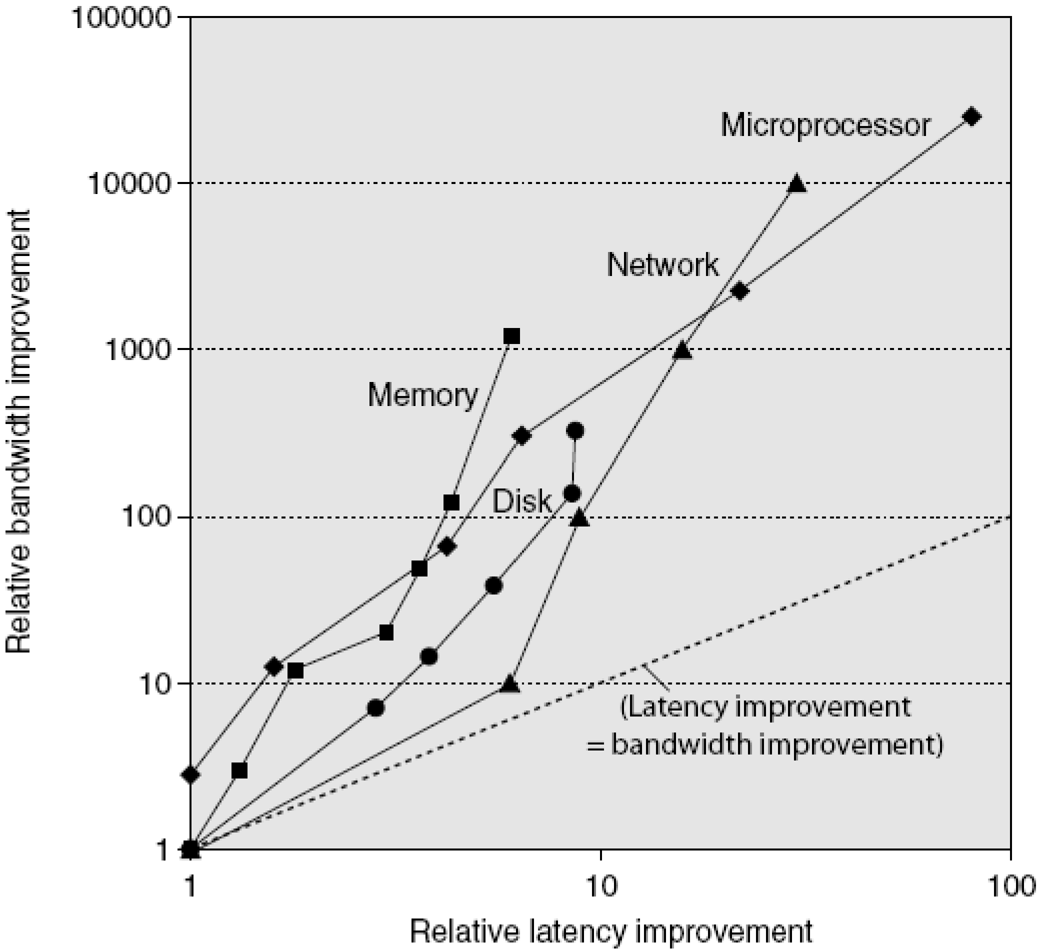
- **Integrated circuit technology**
  - Transistor density: 35%/year
  - Die size: 10-20%/year
  - Integration overall: 40-55%/year
- **DRAM capacity: 25-40%/year (slowing)**
- **Flash capacity: 50-60%/year**
  - 15-20X cheaper/bit than DRAM
- **Magnetic disk technology: 40%/year**
  - 15-25X cheaper/bit than Flash
  - 300-500X cheaper/bit than DRAM

Bandwidth 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

Bandwidth and Latency



Log-log plot of bandwidth and latency milestones