Fundamentals of Computer Design

(Computer Architecture: Chapter 1)

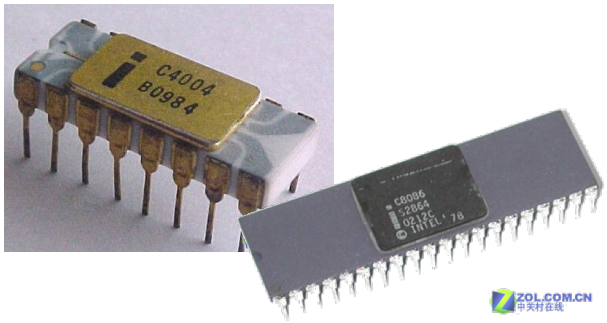
Yanyan Shen

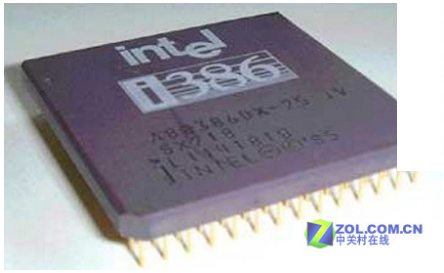
Department of Computer Science and Engineering

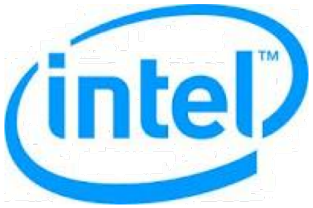
Agenda

- **1.1 Introduction**
- Classes of Computers
- Defining Computer Architecture
- Trends in Technology
- Trends in Power and Energy in ICs
- Trends in Cost
- Dependability
- Measuring Performance
- Quantitative Principles

Evolvement of Processors







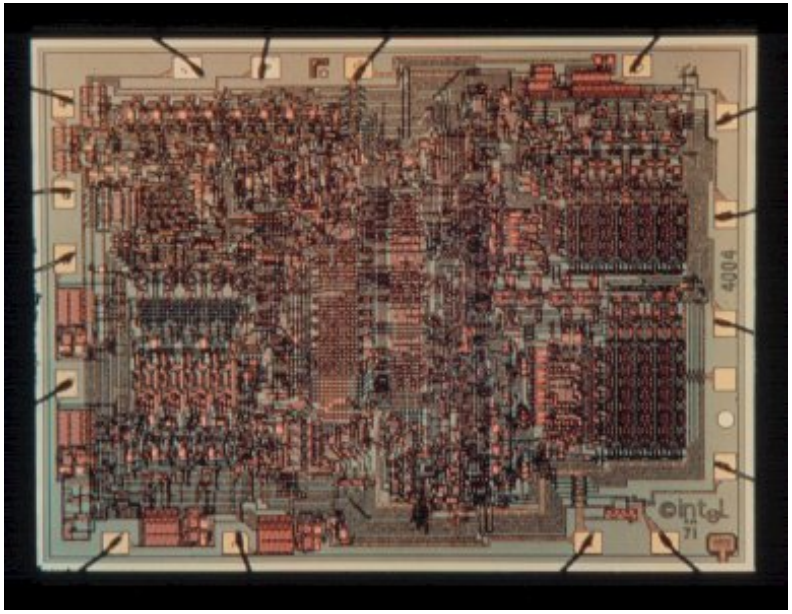
x86 Manufacturers

- Intel
- AMD
- VIA

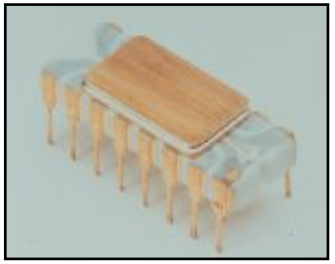
(In the past)

- Transmeta (discontinued its x86 line)
- Rise Technology (acquired by SiS)
- IDT (Centaur Technology x86 division acquired by VIA)
- National Semiconductor (sold the x86 PC designs to VIA and later the x86 embedded designs to AMD)
- Cyrix (acquired by National Semiconductor)
- NexGen (acquired by AMD)
- Chips and Technologies (acquired by Intel)
- IBM (discontinued its own x86 line)
- UMC (discontinued its x86 line)
- NEC (discontinued its x86 line)

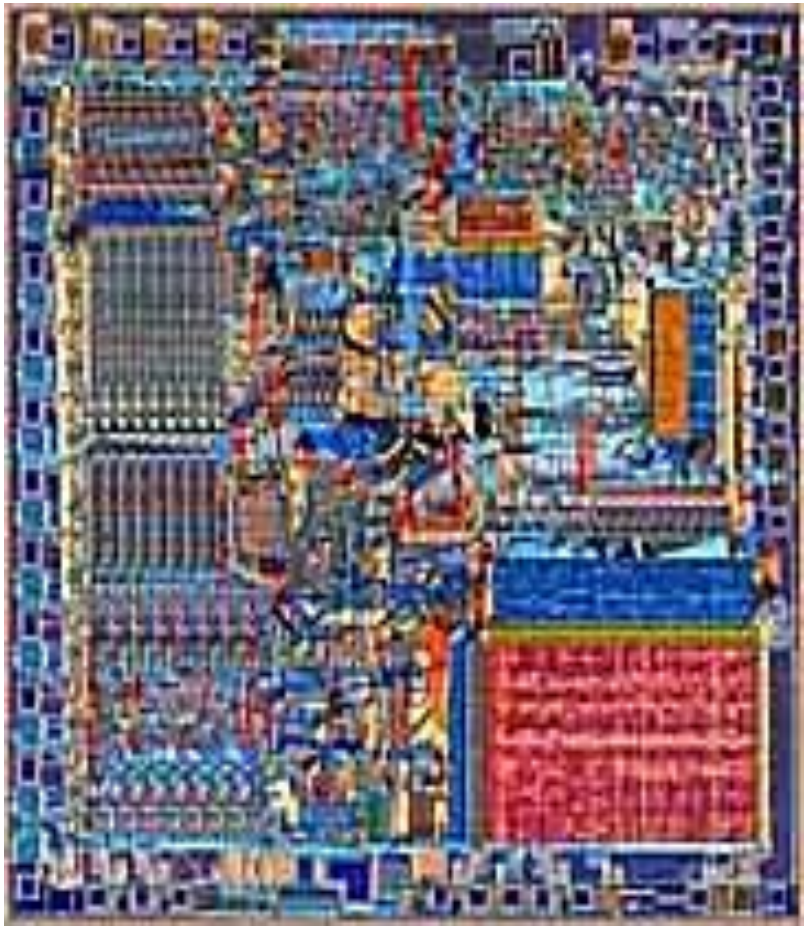
Intel 4004 Die Photo



- Introduced in 1970
  - First microprocessor
- 2,250 transistors
- 12 mm² (die size)
- 108 KHz

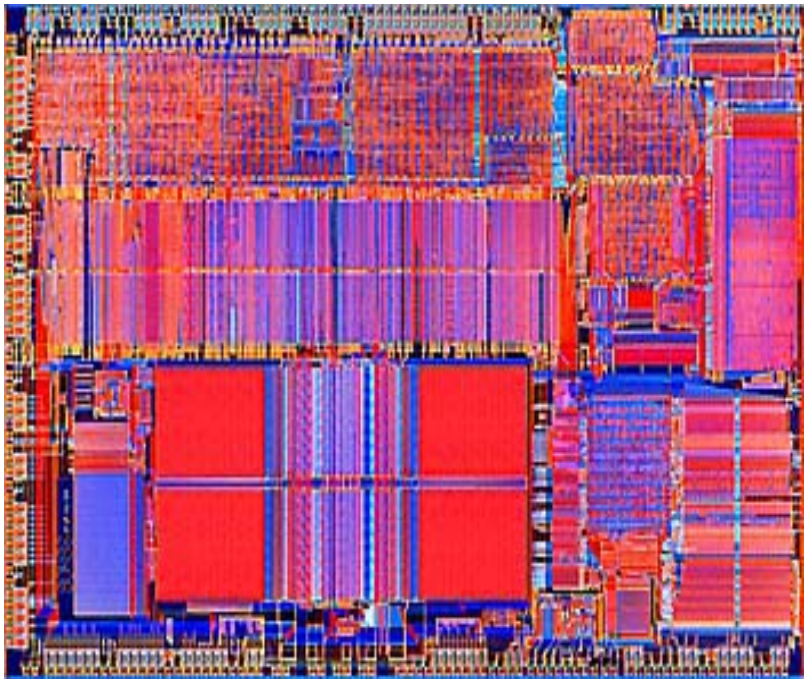


Intel 8086 Die Scan



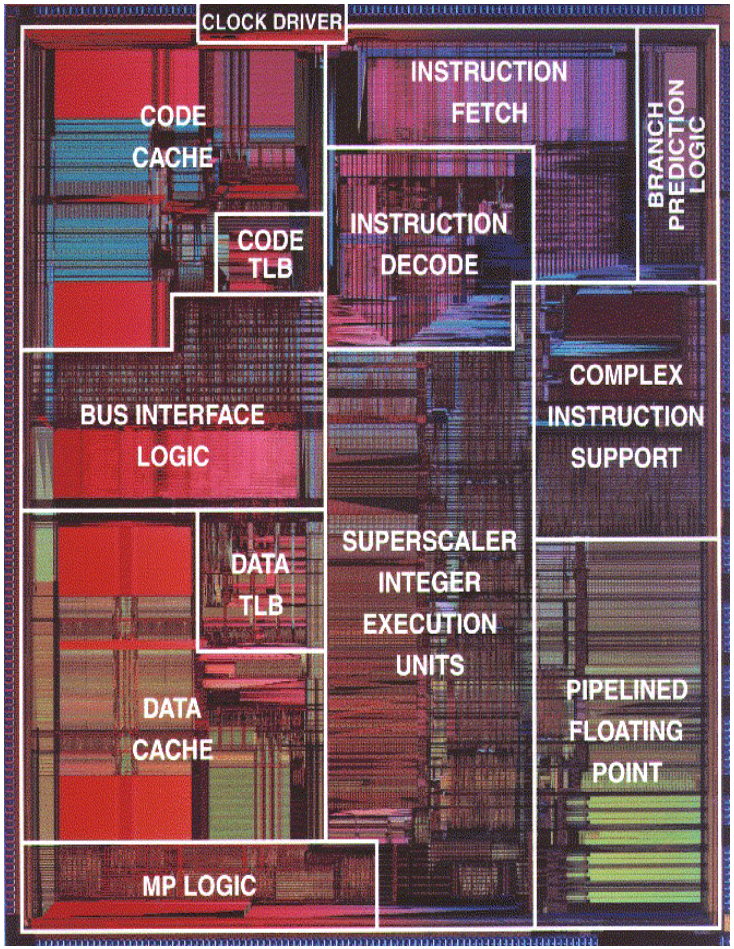
- Introduced in 1979
  - Basic architecture of the IA32 PC
- 29,000 transistors
- 33 mm^2
- 5 MHz

Intel 80486 Die Scan



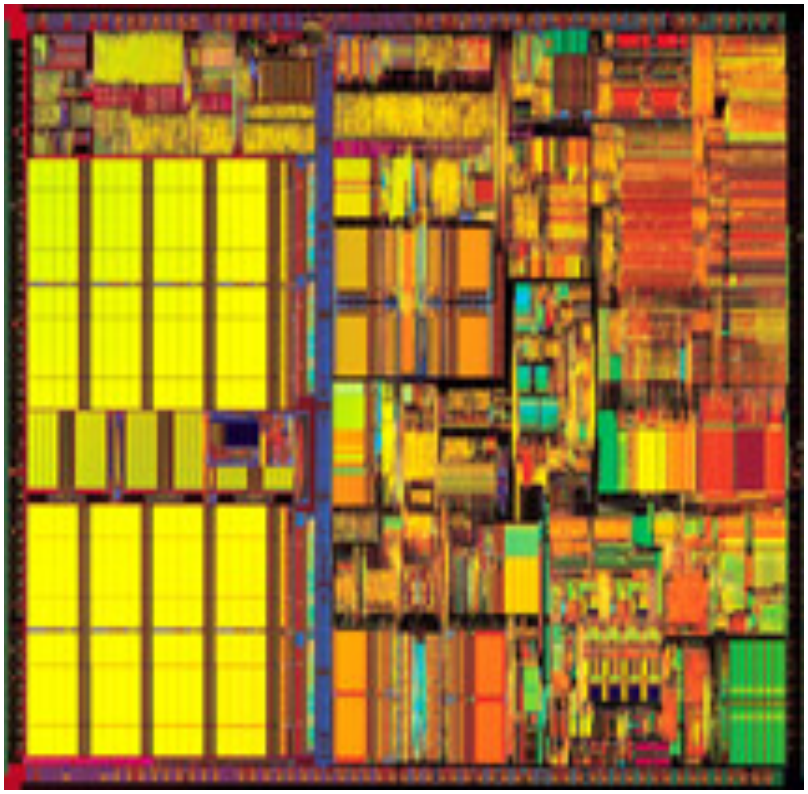
- Introduced in 1989
  - 1st pipelined implementation of IA32
- 1,200,000 transistors
- 81 mm²
- 25 MHz

Pentium Die Photo



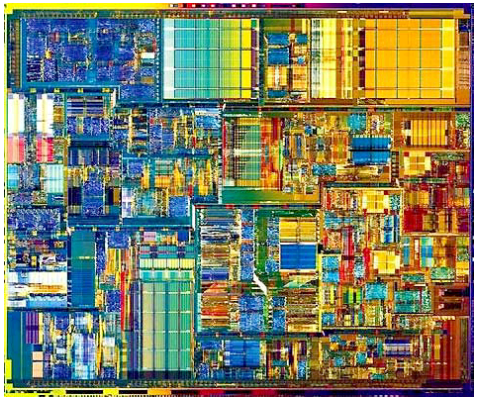
- Introduced in 1993
  - 1st superscalar implementation of IA32
- 3,100,000 transistors
- 296 mm²
- 60 MHz

Pentium III

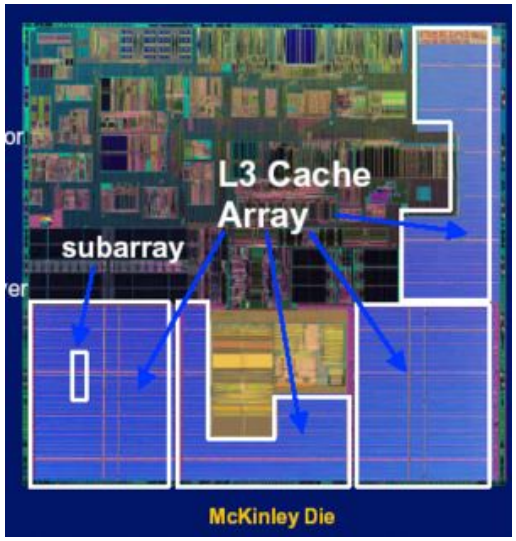


- Introduced in 1999
- 9,500,000 transistors
- 125 mm²
- 450 MHz

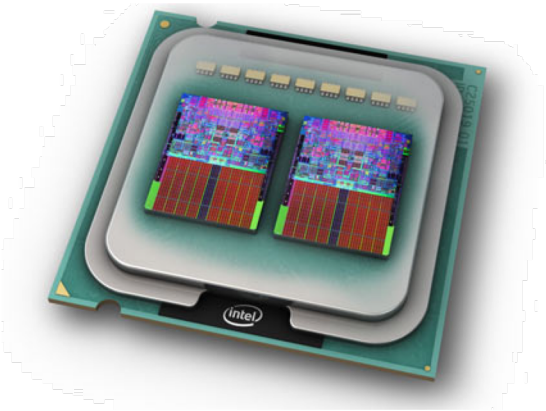
Pentium IV and Duo



Intel P4 – 55M tr
(2001)



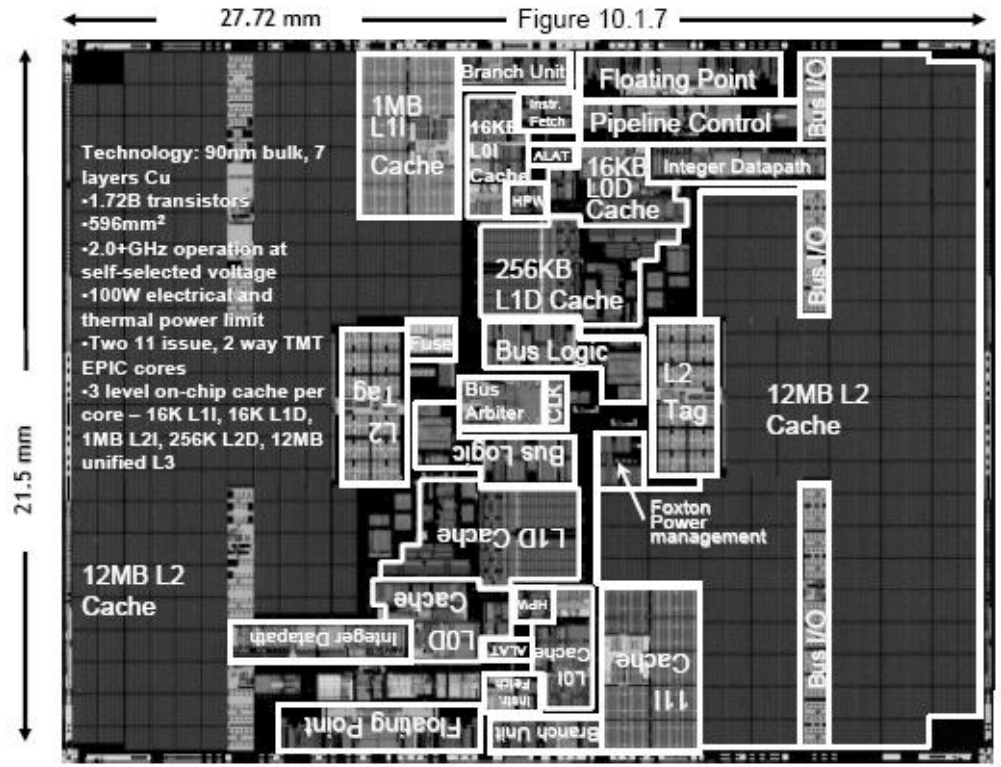
Intel Itanium
– 221M tr.
(2001)



Intel Core 2 Extreme
Quad-core 2x291M tr.
(2006)

Dual-Core Itanium 2 (Montecito)

- 1.72 B Transistors
- 2 GHz frequency



List of Intel Microprocessors

Comparison of Intel Processors

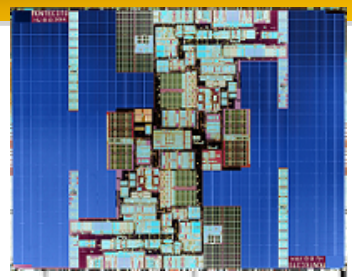
Processor	Series Nomenclature	Code Name	Clock Rate	Socket	Fabrication	TDP	Number of Cores	Bus Speed	L2 Cache	L3 Cache
Intel Pentium	N/A	P5, P54C, P54CTB, P54CS	60 MHz - 200 MHz	Socket 2, Socket 3,Socket 4, Socket 5,Socket 7	800 nm - 350 nm	Unknown	Single	50 MHz - 66 MHz	N/A	N/A
Intel Pentium MMX	N/A	P55C, Tillamook	120 MHz - 300 MHz	Socket 7	350 nm - 250 nm	Unknown	Single	60 MHz - 66 MHz	N/A	N/A
Intel Atom	Z5xx, Z6xx, N2xx, 2xx, 3xx, N4xx, D4xx, D5xx, N5xx, D2xxx, N2xxx	Diamondville,Pineview, Silverthorne, Lincroft, Cedarview, Medfield, Clover Trail	800 MHz - 2.13 GHz	Socket PBGA437,Socket PBGA441, Socket micro-FCBGA8 559	32 nm, 45 nm	0.65 W - 13 W	Single,Double	400 MHz, 533 MHz, 667 MHz, 2.5 GT/s	512 KiB - 1 MiB	-
Intel Celeron	3xx, 4xx, 5xx	Banias,Cedar Mill, Conroe, Coppermine, Covington, Dothan, Mendocino,Northwood, Prescott, Tualatin, Willamette, Yonah	266 MHz - 3.6 GHz	Slot 1,Socket 370, Socket 478, Socket 479,Socket 495, LGA 775,Socket M, Socket T	45 nm, 65 nm, 90 nm, 130 nm, 180 nm, 250 nm	5.5 W - 86 W	Single,Double	66 MHz, 100 MHz, 133 MHz, 400 MHz, 533 MHz, 800 MHz	0 KiB - 1 MiB	-
Intel Pentium Pro	52x	P6	150 MHz - 200 MHz	Socket 8	350 nm, 500 nm	29.2 W - 47 W	Single	60 MHz, 66 MHz	256 KiB, 512 KiB, 1024 KiB	-
Intel Pentium II	52x	Klamath,Deschutes, Tonga, Dixon	233 MHz - 450 MHz	Slot 1, MMC-1, MMC-2,Mini-Cartridge	250 nm, 350 nm	16.8 W - 38.2 W	Single	66 MHz, 100 MHz	256 KiB - 512 KiB	-
Intel Pentium III	52x, 53x	Katmai,Coppermine, Tualatin	450 MHz - 1.4 GHz	Slot 1, Socket 370	130 nm, 180 nm, 250 nm	17 W - 34.5 W	Single	100 MHz, 133 MHz	256 KiB - 512 KiB	-
Intel Xeon	n3xxx, n5xxx, n7xxx	Allendale,Cascades, Clovertown, Conroe, Cranford, Dempsey, Drake, Dunnington,Foster, Gainestown, Gallatin, Harpertown, Irwindale, Kentsfield,Nocona, Paxville, Potomac, Prestonia, Sossaman, Tanner, Tigerton, Tulsa, Wolfdale, Woodcrest	400 MHz - 4.4 GHz	Slot 2,Socket 603, Socket 604, Socket J, Socket T, Socket B LGA 1156,LGA 1366	45 nm, 65 nm, 90 nm, 130 nm, 180 nm, 250 nm	16 W - 165 W	Single,Double, Quad, Hexa, Octa	100 MHz, 133 MHz, 400 MHz, 533 MHz, 667 MHz, 800 MHz, 1066 MHz, 1333 MHz, 1600 MHz, 4.8 GT/s, 5.86 GT/s, 6.4 GT/s	256 KiB - 12 MiB	4 MiB - 16 MiB
Pentium 4	5xx, 6xx	Cedar Mill,Northwood, Prescott, Willamette	1.3 GHz - 3.8 GHz	Socket 423,Socket 478, LGA 775,Socket T	65 nm, 90 nm, 130 nm, 180 nm	21 W - 115 W	Single	400 MHz, 533 MHz, 800 MHz, 1066 MHz	256 KiB - 2 MiB	-
Pentium 4 Extreme Edition	5xx, 6xx	Gallatin,Prescott 2M	3.2 GHz - 3.73 GHz	Socket 478,Socket T	90 nm, 130 nm	92 W - 115 W	Single	800 MHz, 1066 MHz	512 KiB - 1 MiB	0 KiB - 2 MiB
Pentium M	7xx	Banias,Dothan	800 MHz - 2.266 GHz	Socket 479	90 nm, 130 nm	5.5 W - 27 W	Single	400 MHz, 533 MHz	1 MiB - 2 MiB	-
Pentium D/EE	8xx, 9xx	Smithfield,Presler	2.66 GHz - 3.73 GHz	Socket T	65 nm, 90 nm	95 W - 130 W	Double	533 MHz, 800 MHz, 1066 MHz	2×1 MiB - 2×2 MiB	-
Intel Pentium Dual-Core	E2xxx, E3xxx, E5xxx, T2xxx, T3xxx	Allendale,Penryn, Wolfdale, Yonah	1.6 GHz - 2.93 GHz	Socket 775,Socket M, Socket P,Socket T		45 nm, 65 nm	10 W - 65 W	Double	533 MHz, 667 MHz, 800 MHz, 1066 MHz	1 MiB - 2 MiB
Intel Pentium New	E5xxx, E6xxx, T4xxx, SU2xxx, SU4xxx, G69xx, P6xxx, U5xxx, G6xx, G8xx, B9xx	Penryn,Wolfdale, Clarkdale, , Sandy Bridge,	1.2 GHz - 3.33 GHz	Socket 775,Socket P, Socket T, LGA 1156, LGA 1155,	32 nm, 45 nm, 65 nm	5.5 W - 73 W	Single,Double	800 MHz, 1066 MHz, 2.5GT/s, 5 GT/s	2x256 KiB - 2 MiB	0 KiB - 3 MiB
Intel Core	Txxxx, Lxxxx, Uxxxx	Yonah	1.06 GHz - 2.33 GHz	Socket M	65 nm	5.5 W - 49 W	Single,Double	533 MHz, 667 MHz	2 MiB	-
Intel Core 2	Uxxxx, Lxxxx, Exxxx, Txxxx, P7xxx, Xxxxx, Qxxxx, QXxxxx	Allendale,Conroe, Merom, Penryn, Kentsfield, Wolfdale, Yorkfield	1.06 GHz - 3.33 GHz	Socket 775,Socket M, Socket P,Socket T, Socket T	45 nm, 65 nm	5.5 W - 150 W	Single,Double, Quad	533 MHz, 667 MHz, 800 MHz, 1066 MHz, 1333 MHz, 1600 MHz	1 MiB - 12 MiB	-
Intel Core i3	i3-xxx, i3-2xxx, i3-3xxx	Arrandale,Clarkdale, Sandy Bridge, Ivy Bridge	2.4 GHz - 3.4 GHz	LGA 1156, LGA 1155	22 nm, 32 nm	35 W - 73 W	Double	1066 MHz, 1600 MHz, 2.5 - 5 GT/s	256 KiB	3 MiB - 4 MiB
Intel Core i5	i5-7xx, i5-6xx, i5-2xxx, i5-3xxx	Arrandale,Clarkdale, Clarksfield, Lynnfield, Sandy Bridge, Ivy Bridge	1.06 GHz - 3.46 GHz	LGA 1156, LGA 1155	22 nm, 32 nm, 45 nm	17 W - 95 W	Double,Quad	2.5 - 5 GT/s	256 KiB	4 MiB - 8 MiB
Intel Core i7	i7-6xx, i7-7xx, i7-8xx, i7-9xx, i7-2xxx, i7-37xx, i7-38xx, i7-47xx	Bloomfield,Nehalem, Clarksfield, Clarksfield XM, Lynnfield, Sandy Bridge, Sandy Bridge-E, Ivy Bridge, Haswell	1.6 GHz - 3.6 GHz	LGA 1156,LGA 1155, LGA 1366,LGA 2011	22 nm, 32 nm, 45 nm	45 W - 130 W	Quad	4.8 GT/s, 6.4 GT/s	4×256 KiB	6 MiB - 10 MiB
Intel Core i7	i7-970, i7-980, i7-980x, i7-990x, i7-39xx, i7-38xx	Gulftown,Sandy Bridge-E	3.2 GHz - 3.46 GHz	LGA 1366,LGA 2011	32 nm	130 W	Hexa	6.4 GT/s	6x256 KiB	12 MiB - 15 MiB
Processor	Series Nomenclature	Code Name	Clock Rate	Socket	Fabrication	TDP	Number of Cores	Bus Speed	L2 Cache	L3 Cache

Processor Transistor Count (from http://en.wikipedia.org/wiki/Transistor_count)

Processor	Transistor count	Date of intro-duction	Manufactu-rer
Intel 4004	2300	1971	Intel
Intel 8008	2500	1972	Intel
Intel 8080	4500	1974	Intel
Intel 8088	29 000	1978	Intel
Intel 80286	134 000	1982	Intel
Intel 80386	275 000	1985	Intel
Intel 80486	1 200 000	1989	Intel
Pentium	3 100 000	1993	Intel
AMD K5	4 300 000	1996	AMD
Pentium II	7 500 000	1997	Intel
AMD K6	8 800 000	1997	AMD
Pentium III	9 500 000	1999	Intel
AMD K6-III	21 300 000	1999	AMD
AMD K7	22 000 000	1999	AMD
Pentium 4	42 000 000	2000	Intel

Processor	Transistor count	Date of introdu-ction	Manufacturer
Itanium	25 000 000	2001	Intel
Barton	54 300 000	2003	AMD
AMD K8	105 900 000	2003	AMD
Itanium 2	220 000 000	2003	Intel
Itanium 2 with 9MB cache	592 000 000	2004	Intel
Cell	241 000 000	2006	Sony/IBM/ Toshiba
Core 2 Duo	291 000 000	2006	Intel
Core 2 Quadro	582 000 000	2006	Intel
Dual-Core Itanium 2	1 700 000 000	2006	Intel

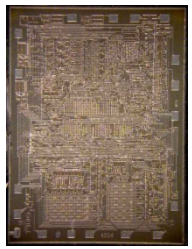
Moore's Law



0.09 µm
596 mm²

1.7 billions
Montecito

transistors



10 µm
13.5mm²

2,250

MOORE'S LAW

Dual-Core Intel® Itanium® 2 Processor
Intel® Itanium® 2 Processor
Intel® Itanium® Processor
Intel® Pentium® 4 Processor
Intel® Pentium® III Processor
Intel® Pentium® II Processor
Intel® Pentium® Processor
Intel486™ Processor
Intel386™ Processor
286
8086
8080
8008
4004

42millions

Exponential growth

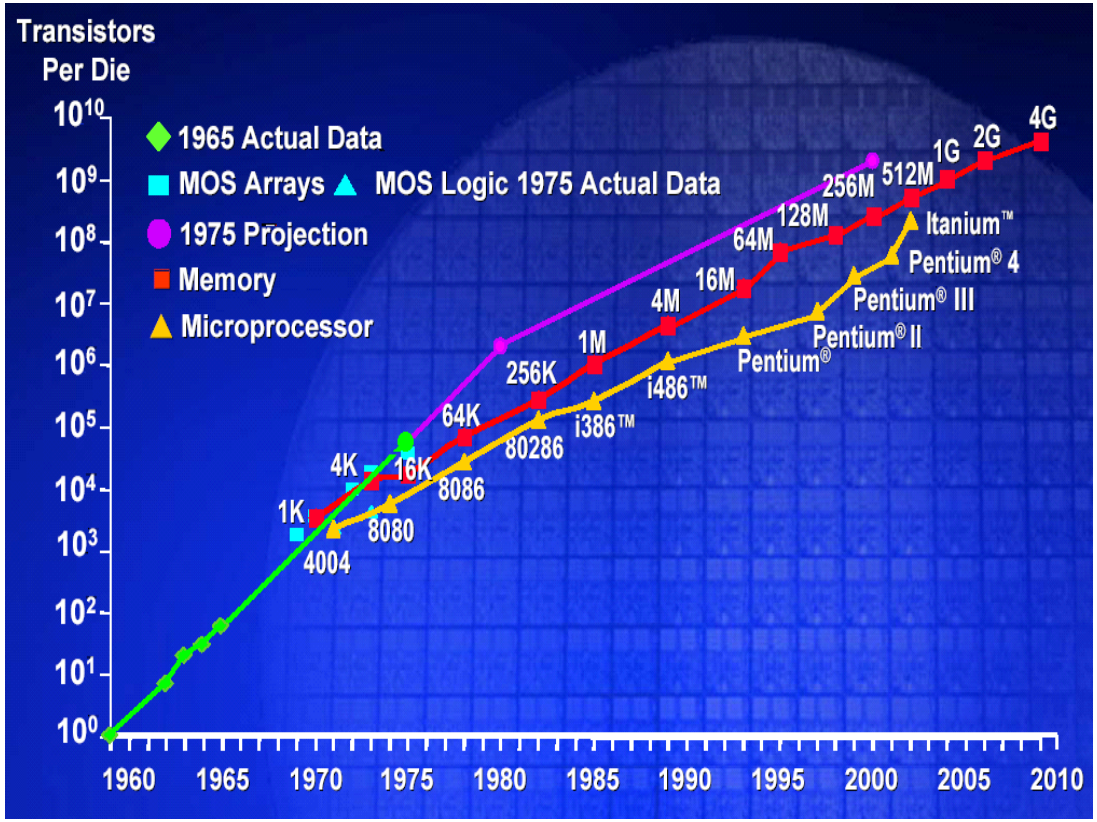
10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000

1970 1975 1980 1985 1990 1995 2000 2005 2010

Transistor count will be doubled every 18 months

— Gordon Moore, Intel co-founder

Memory Capacity (Single Chip DRAM)



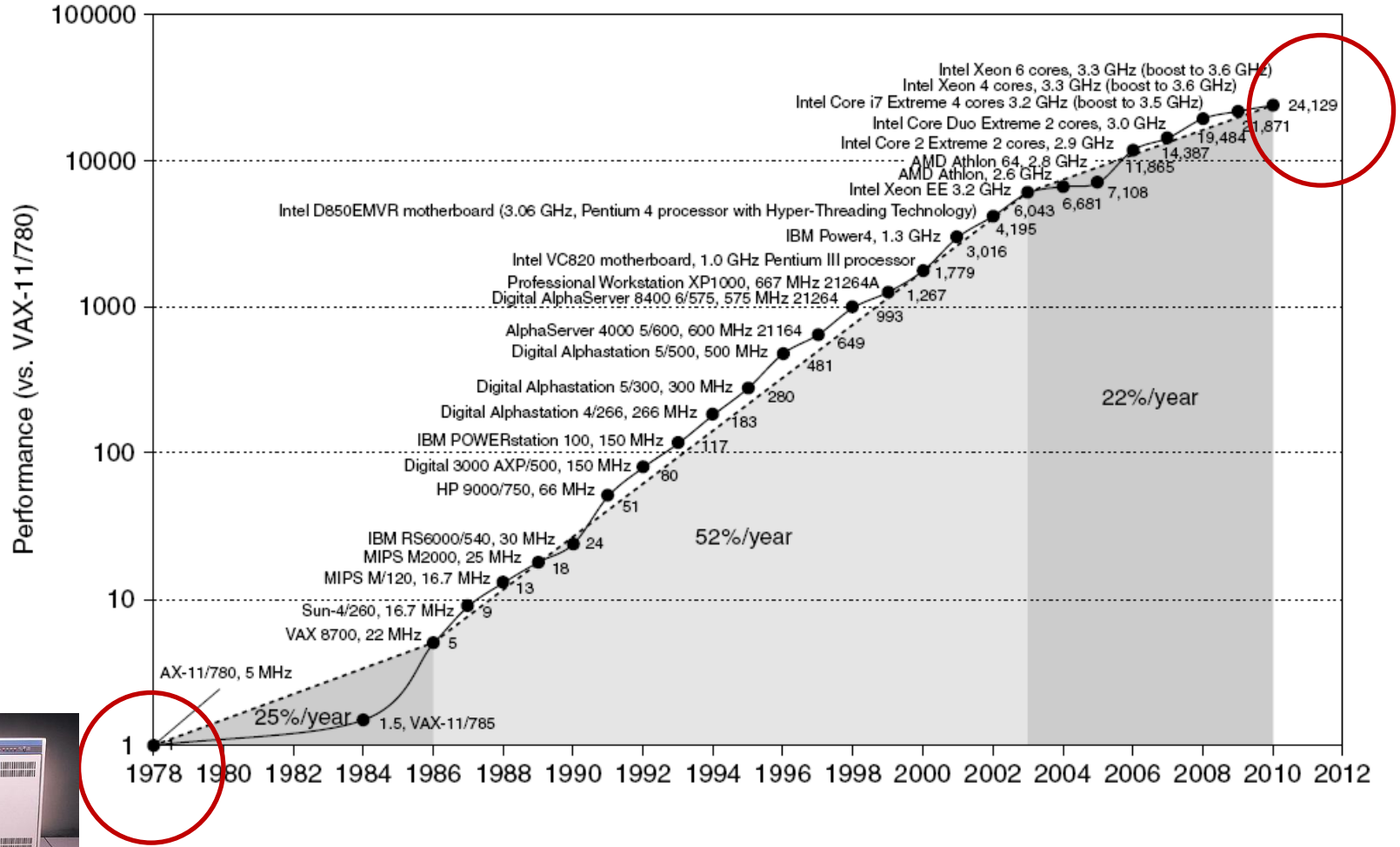
year	size(Mb)	cyc time
1980	0.0625	250 ns
1983	0.25	220 ns
1986	1	190 ns
1989	4	165 ns
1992	16	145 ns
1996	64	120 ns
2000	256	100 ns
2007	2G	52 ns

Moore's Law for Memory: capacity increases by 4x every 3 years

Trends in Technology

- Trends in Technology followed closely Moore's Law "*Transistor density of chips doubles every 1.5-2.0 years*"
- As a consequence of Moore's Law:
  - Processor speed doubles every 1.5-2.0 years
  - DRAM size doubles every 1.5-2.0 years
- These constitute a target that the computer industry aim for.

Rapid Improvement



Rapid Improvements

Understand the unprecedented innovations in computers!

What have been making computers faster and cheaper? And how to continue the innovations?

- **50 years of non-stop innovation – a technological miracle!**

	Capacity (memory)	Speed (CPU)	Price
IBM7030 (Stretch)1961 	128KB	1.2 MIPS	US$13,500,000
Pentium 4 Desktop 2000 	1G (typical)	1000 MIPS	US$800

Automobile and Computer

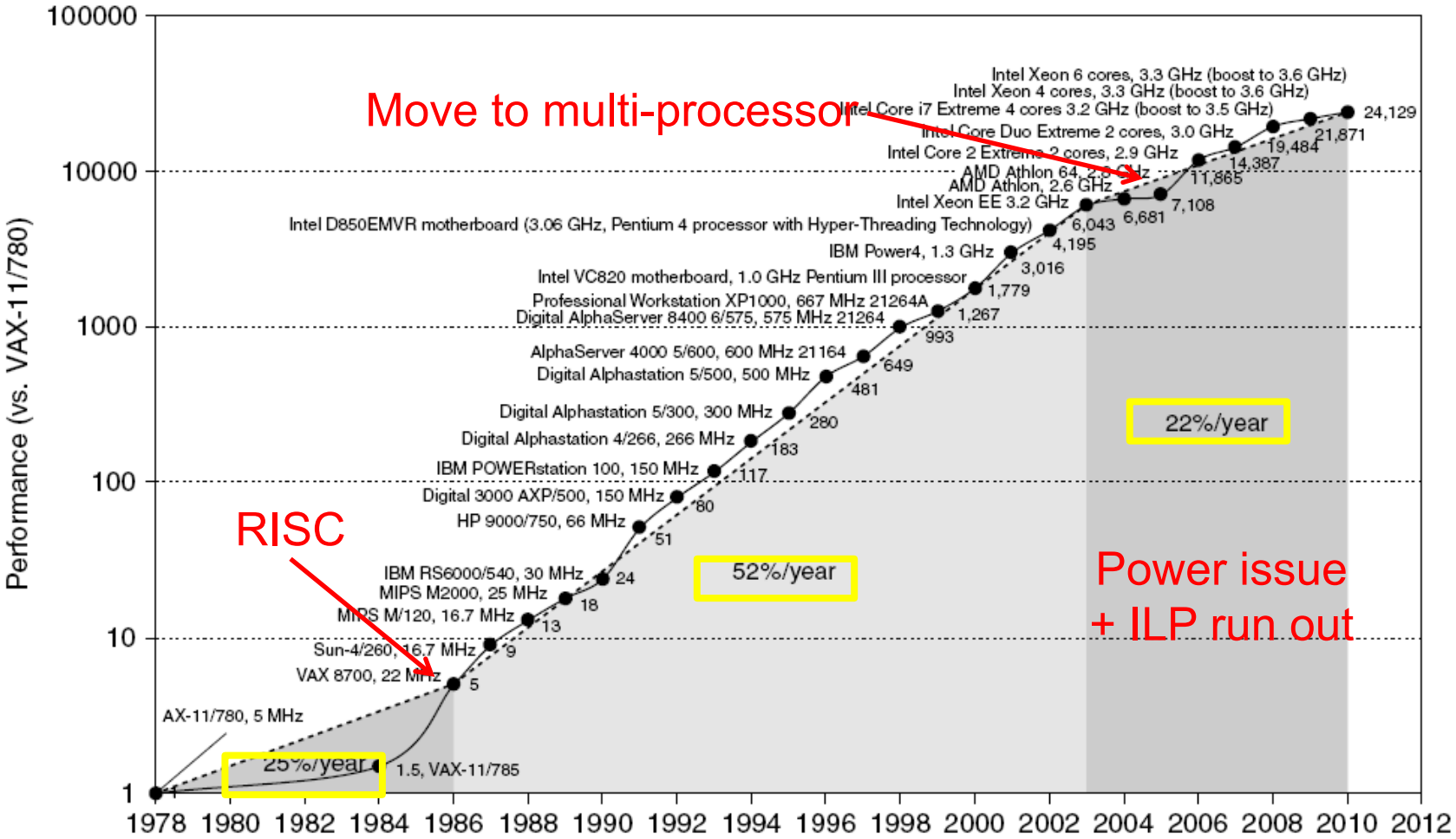
- ***What would cars be like if they had innovated like computer?***

	Capacity (passengers)	Speed (in KM/hour)	Price
A good car in the 1960's	5	100	US$5,000
A car in the 2000's had it developed like computers	39062	83333	US$0.29

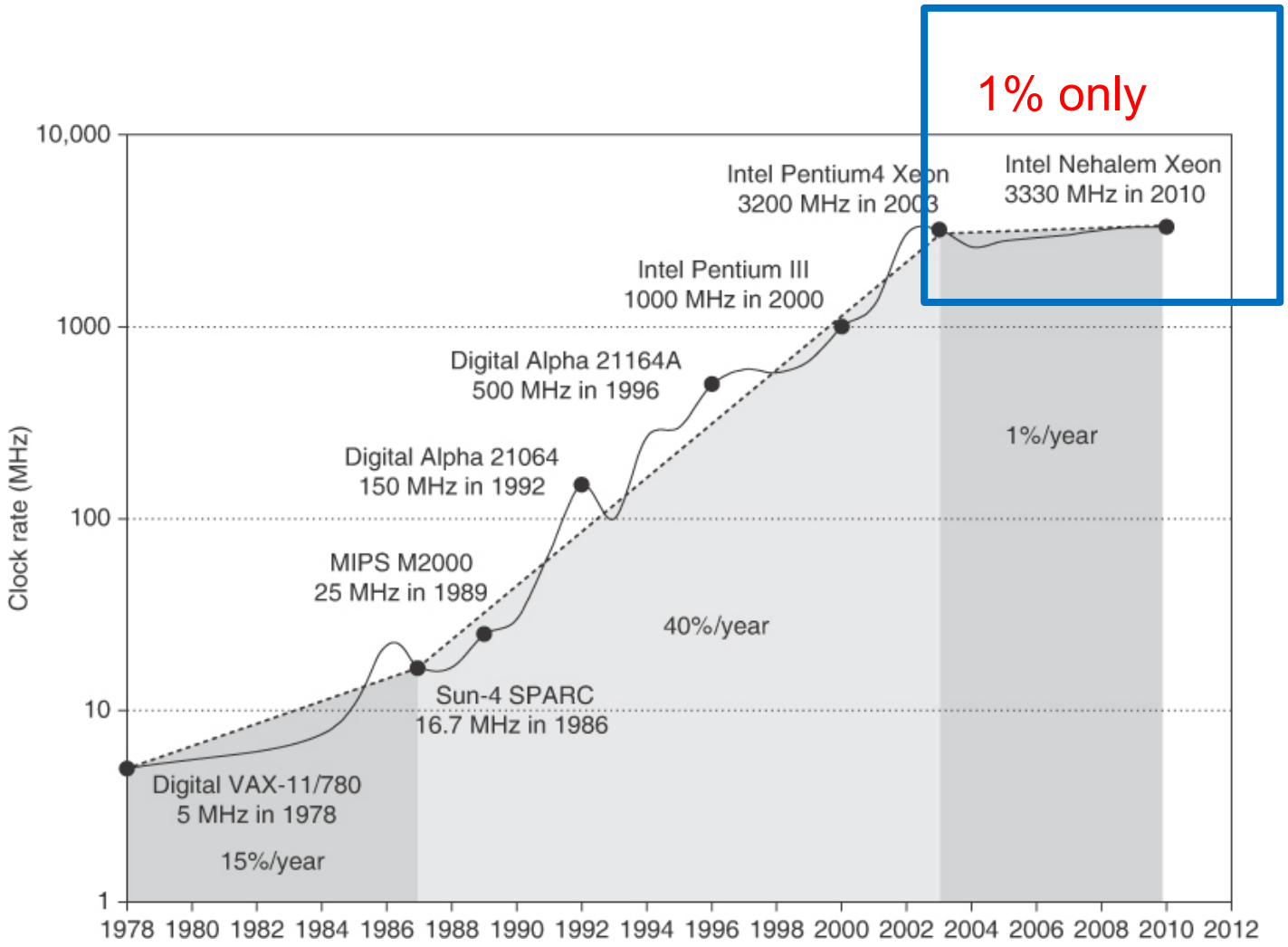
How to Make Computers Faster?

- **Option 1**: To increase the clock rate or main frequency
  - Today's mainstream: 3 GHz
- **Option 2**: To increase the logic density (number of gates in a chip)
  - Today's mainstream: 32 nm (2012)
  - Intel shipped 14nm for Core M family (2015)
  - In comparison, 10 microns (1971)
- **Always useful?**

Single Processor Performance



Growth in Clock Rate



Effects of Dramatic Growth

- 1) It has significantly enhanced the capability of a computer to users
- 2) The dramatic improvement in cost-performance leads to new classes of computers
  - Personal computers/ workstations emerged in 1980
  - Smart cell phones and tablets
  - Warehouse-scale computers
- 3) We see the dominance of *microprocessor-based* computers across the entire range of computer design
  - Minicomputers, made from off-the-self logic or gate arrays, disappear
  - Even mainframes and supercomputers made from microprocessors
- 4) Impact of software development.
  - Trade performance (C/C++) for productivity (Java, C#)

Current Trends in Architecture

- Historic switch in 2003: from uniprocessor to multiprocessor per chip. Single processor performance improvement ended in 2003
- Walls encountered
  - Power issue
  - No more Instruction-Level parallelism (ILP)

- This signals the change from *solely relying on instruction-level parallelism* to *exploiting more coarse-grained parallelism*
  - Data-level parallelism (DLP)
  - Thread-level parallelism (TLP)
  - Request-level parallelism (RLP)
- **These require explicit restructuring of the applications**

Agenda

- Introduction
- 1.2 Classes of Computers
- Defining Computer Architecture
- Trends in Technology
- Trends in Power and Energy in ICs
- Trends in Cost
- Dependability
- Measuring Performance
- Quantitative Principles

Five Classes of Computers

- **1) Personal Mobile Device (PMD)**
  - e.g. smart phones, tablet computers
  - Emphasis on energy efficiency and real-time
- **2) Desktop Computing**
  - Emphasis on price-performance
- **3) Servers**
  - Emphasis on availability, scalability, throughput
- **4) Clusters / Warehouse Scale Computers**
  - Used for "Software as a Service (SaaS)"
  - Emphasis on availability and price-performance
- **5) Embedded Computers**
  - Emphasis: price

System Characteristics

Feature	Personal mobile device (PMD)	Desktop	Server	Clusters/warehouse-scale computer	Embedded
Price of system	$100–$1000	$300–$2500	$5000–$10,000,000	$100,000–$200,000,000	$10–$100,000
Price of micro-processor	$10–$100	$50–$500	$200–$2000	$50–$250	$0.01–$100
Critical system design issues	Cost, energy, media performance, responsiveness	Price-performance, energy, graphics performance	Throughput, availability, scalability, energy	Price-performance, throughput, energy proportionality	Price, energy, application-specific performance

Personal Mobile Device (PMD)

- PMD: Collection of wireless devices with multimedia user interfaces
- Energy efficiency is critical
  - Most of the devices are driven by batteries
  - There is no fan for cooling the processor
- Flash memory instead of **disks** is used
  - Energy and size requirements
- Responsiveness or real-time performance
  - Hard real time
  - Soft real time

Desktop Computing

- The **largest** market in dollar terms
  - How about now?

- The metric of price-performance: combination of both price and performance
  - Compute performance
  - Graphics performance

Servers

- Servers are backbone of large-scale enterprise computing
- **Availability:** the percentage of time that a server is operational
- **Scalability:** scale up to increasing demand of services
- **Energy**

Cost of Downtime

Application	Cost of downtime per hour	Annual losses with downtime of		
		1% (87.6 hrs/yr)	0.5% (43.8 hrs/yr)	0.1% (8.8 hrs/yr)
Brokerage operations	$6,450,000	$565,000,000	$283,000,000	$56,500,000
Credit card authorization	$2,600,000	$228,000,000	$114,000,000	$22,800,000
Package shipping services	$150,000	$13,000,000	$6,600,000	$1,300,000
Home shopping channel	$113,000	$9,900,000	$4,900,000	$1,000,000
Catalog sales center	$90,000	$7,900,000	$3,900,000	$800,000
Airline reservation center	$89,000	$7,900,000	$3,900,000	$800,000
Cellular service activation	$41,000	$3,600,000	$1,800,000	$400,000
Online network fees	$25,000	$2,200,000	$1,100,000	$200,000
ATM service fees	$14,000	$1,200,000	$600,000	$100,000

Clusters/Warehouse-Scale Computers (WSC)

- **Software as a Service (SaaS):** search, social networking, video sharing
- **Cluster:** Collection of desktop computers or servers by local area networks, each of which runs its own OS
- **Energy:** 80% of a warehouse is associated with power and cooling of computers
- **Availability**
- **Throughput:** the amount of work that is done in a unit time

Supercomputers vs. WSC

- Supercomputer: a large pool of processors
- Both supercomputers and WSCs are expensive
- **Differences** between supercomputers and WSCs
  - Supercomputers run large, communication-intensive applications, and thus faster internal networks required
  - Supercomputers emphasize floating-point performance
  - WSCs emphasize interactive applications, large-scale storage, dependability and high bandwidth

Embedded Computers

- They can be found almost in every machine: washing machine, microwave, and printers
- Can be 8-, 16-, and 32-bit

- An embedded computer is usually application-specific
  - It does not run various, changing applications

Flynn's Taxonomy

Single instruction stream, single data stream (SISD)

Single instruction stream, multiple data streams (SIMD)

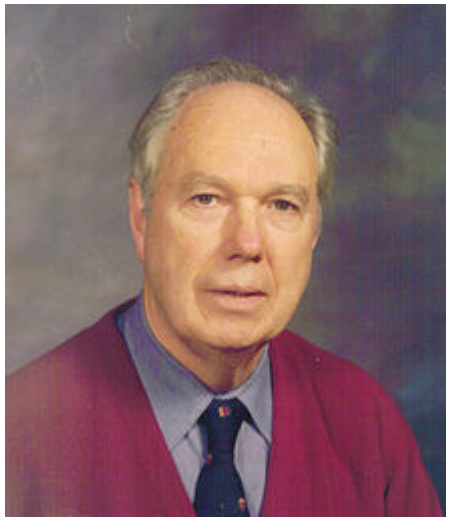
- Vector architectures
- Multimedia extensions
- Graphics processor units (GPU)

Multiple instruction streams, single data stream (MISD)

- No commercial implementation

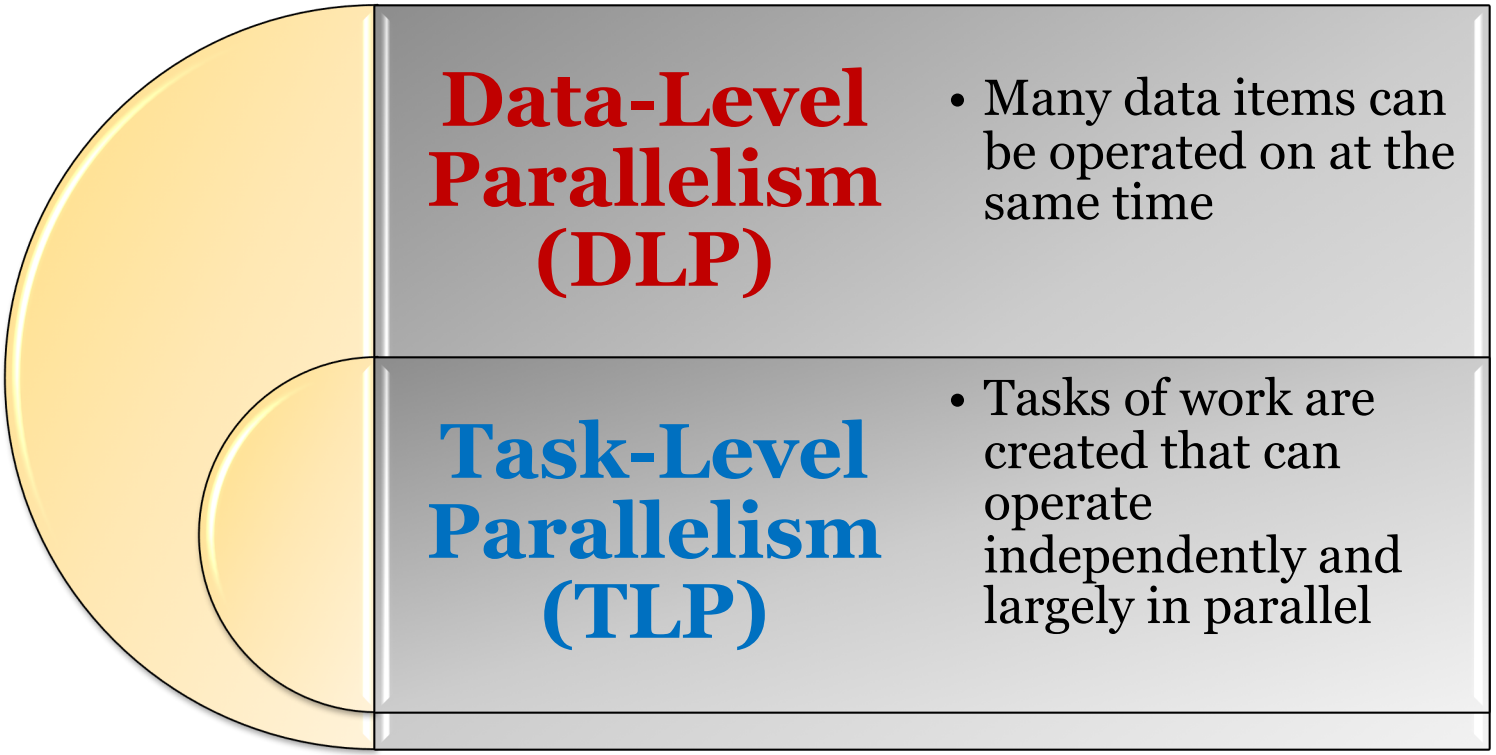
Multiple instruction streams, multiple data streams (MIMD)

- Tightly-coupled MIMD
- Loosely-coupled MIMD



Michael J. Flynn

Two Kinds of Parallelism in Applications



Four Major Ways for Exploiting Parallelism

Instruction-Level Parallelism (ILP)- e.g., Pipelining

- Data-level parallelism

Vector architectures/Graphic Processor Units (GPUs)

- Data-level parallelism

Thread-Level Parallelism - e.g., Multicore

- Data-level parallelism or task-level parallelism

Request-Level Parallelism – e.g., Clusters

- Data-level parallelism or task-level parallelism

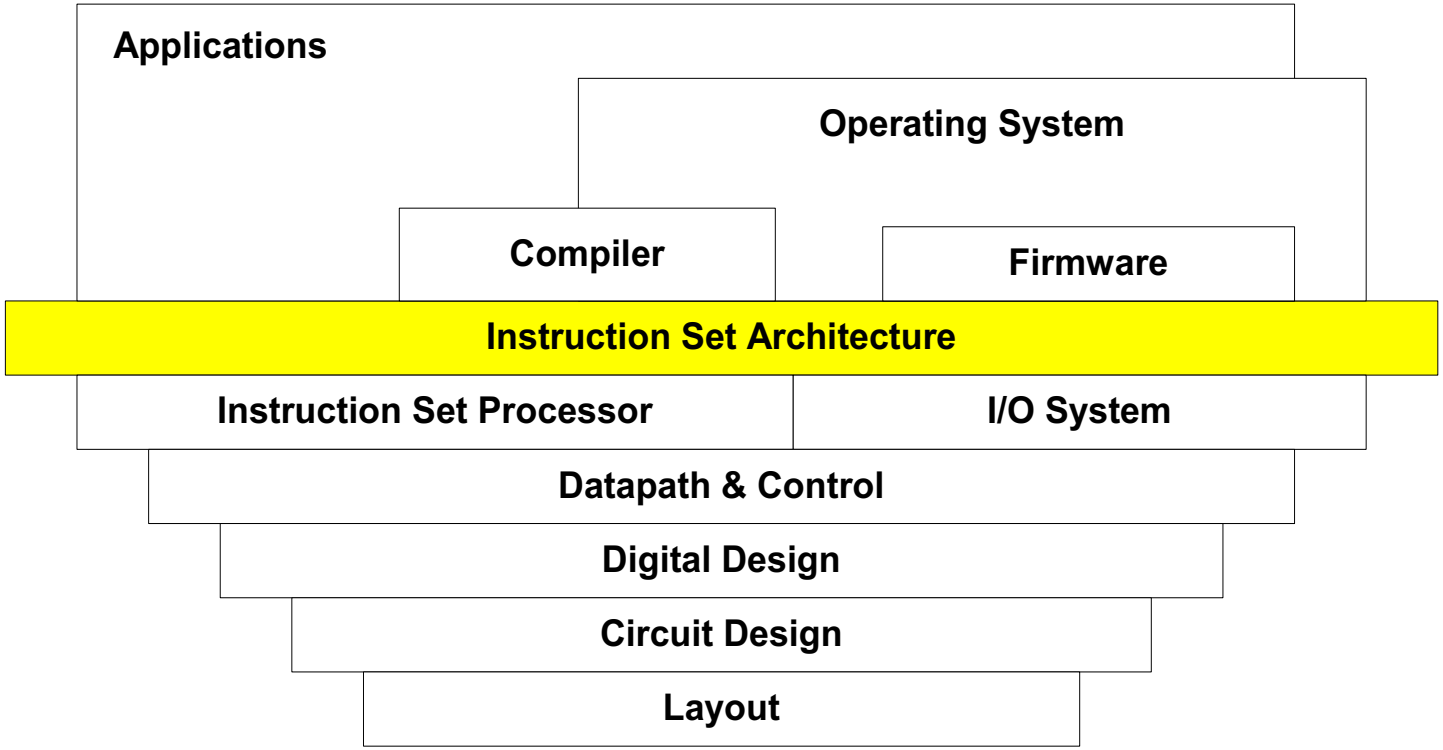
Agenda

- Introduction
- Classes of Computers
- 1.3 Defining Computer Architecture
- Trends in Technology
- Trends in Power and Energy in ICs
- Trends in Cost
- Dependability
- Measuring Performance
- Quantitative Principles

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
- The ISA is roughly the same as the programming model of a processor as seen by an assembly language programmer or compiler writer.

Instruction Set Architecture



S/W and H/W consists of hierarchical layers of abstraction, each hides details of lower layers from the above layer

The instruction set arch. abstracts the H/W and S/W interface and allows many implementation of varying cost and performance to run the same S/W

Contents of Instruction Set Architecture

- Registers
- Class of ISA
- Memory addressing
  - Byte addressing
- Addressing modes
- Types and sizes of operands
- Operations
- Control flow instructions
- Encoding an ISA
  - Fixed length and variable length

Addressing Modes

Example Instruction	Meaning	When used
Register	Add R4,R3	R4 <- R4 + R3
Immediate	Add R4, #3	R4 <- R4 + 3
Displacement	Add R4, 100(R1)	R4 <- R4 + M[100+R1]
Register deferred	Add R4,(R1)	R4 <- R4 + M[R1]
Indexed	Add R3, (R1 + R2)	R3 <- R3 + M[R1+R2]
Direct	Add R1, (1001)	R1 <- R1 + M[1001]
Memory deferred	Add R1, @(R3)	R1 <- R1 + M[M[R3]]
Auto-increment	Add R1, (R2)+	R1 <- R1 +M[R2] R2 <- R2 + d
Auto-decrement	Add R1,-(R2)	R2 <-R2-d R1 <- R1 + M[R2]
Scaled	Add R1, 100(R2)[R3]	R1<-R1+M[100+R2+R3*d]

Example ISAs

- ARM
  - ARMv1-v7
  - ARMv7 extensions
    - Thumb-2
    - NEON – media acceleration technology[8]
    - VFP v3
- HP
  - HP 2100
  - PA-RISC
- IBM
  - IBM 8100
  - IBM Series
  - Power Architecture
    - POWER
    - PowerPC
    - PowerPC AS
- MIPS

- Intel
  - IA-64
  - X86
  - x86 extensions
    - FPU (x87) – Floating-point-unit (FPU) instructions
    - MMX – MMX SIMD instructions
    - MMX Extended – extended MMX SIMD instructions
    - SSE – streaming SIMD extensions (SSE) instructions (70 instructions)
    - SSE2 – streaming SIMD extensions 2 instructions (144 new instructions)
    - SSE3 – streaming SIMD extensions 3 instructions (13 new instructions)
    - SSSE3 – supplemental streaming SIMD extensions (16 instructions)
    - SSE4.1 – streaming SIMD extensions 4, Penryn subset (47 instructions)
    - SSE4.2 – streaming SIMD extensions 4, Nehalem subset (7 instructions)
    - SSE4 – All streaming SIMD extensions 4 instructions (both SSE4.1 and SSE4.2)
    - SSE4a – streaming SIMD extensions 4a (AMD)
    - SSE5 – streaming SIMD extensions 5 (170 instructions)
    - XSAVE – XSAVE instructions
    - AVX – advanced vector extensions instructions
    - FMA – fused multiply-add instructions
    - AES – Advanced Encryption Standard instructions
    - CLMUL – Carry-less mtiply (PCLMULQDQ) instruction
    - 3DNow![citation needed] – 3DNow! instructions (21 instructions)
    - 3DNow! Extended – extended 3DNow! instructions (5 instructions)
    - Cyrix – Cyrix-specific instructions
    - AMD – AMD-specific instructions (older than K6)
    - SMM – System management mode instructions
    - SVM – Secure virtual machine instructions
    - PadLock – VIA PadLock instructions

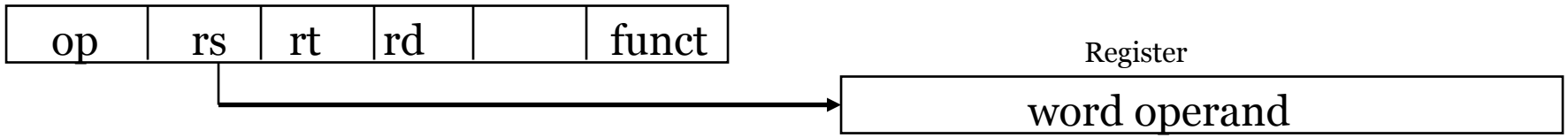
MIPS Instruction Fields

- MIPS fields are given names to make them easier to refer to

op	6-bits	opcode that specifies the operation
rs	5-bits	register file address of the first source operand
rt	5-bits	register file address of the second source operand
rd	5-bits	register file address of the result's destination
shamt	5-bits	shift amount (for shift instructions)
funct	6-bits	function code augmenting the opcode

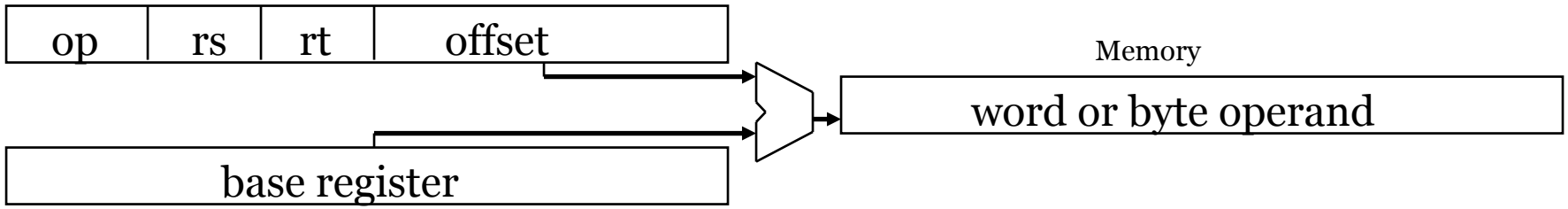
MIPS Addressing Modes Examples

1. Register addressing

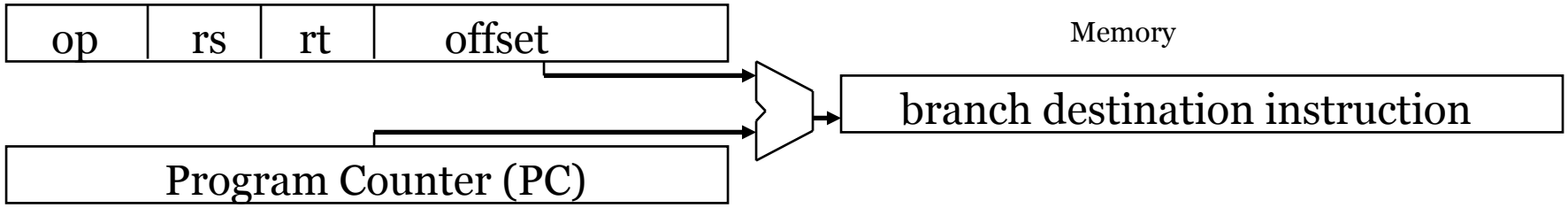


2. Immediate addressing

3. Base (displacement) addressing

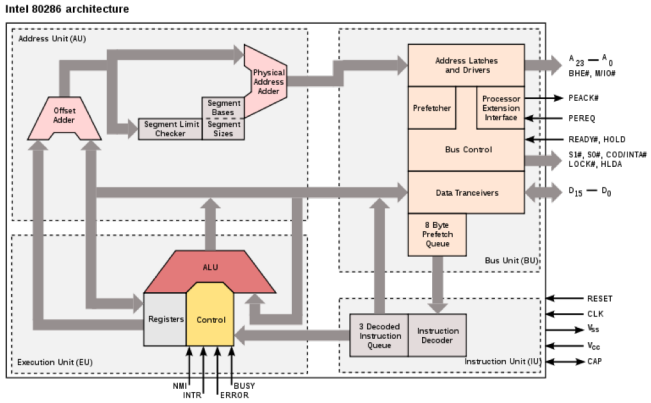


4. PC-relative addressing



Microarchitecture

- **microarchitecture**, also called **computer organization**, is the way a given ISA is implemented on a processor.
  - A given ISA may be implemented with different microarchitectures.
- The microarchitecture includes the constituent parts of the processor and how these interconnect and interoperate to implement the ISA.



Intel i80286 microarchitecture

Microarchitectural Concepts

- Instruction pipelining
- Hierarchical memory organization
- Cache
- Cache coherence
- Branch prediction
- Superscalar
- Out-of-order execution
- Register renaming
- Multiprocessing and multithreading

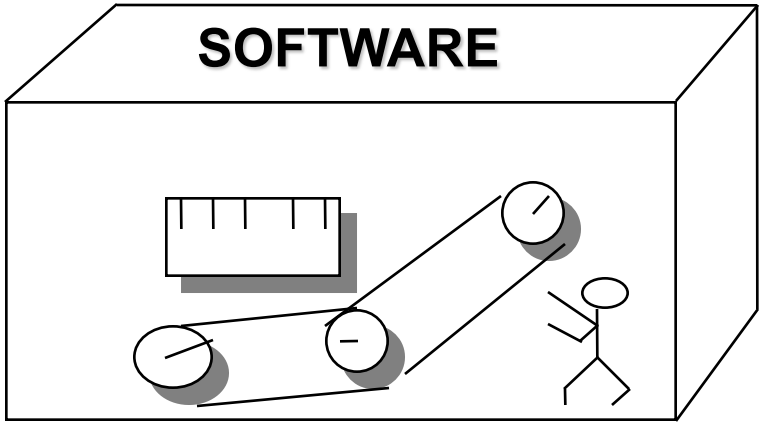
Computer Architecture in General

When we construct a building numerous practical considerations need to be taken into account:

- Available materials
- Worker skills
- Budget
- Space

Similarly, Computer Architecture is about working within constraints:

- What will the market buy?
- Cost/Performance
- Tradeoffs in materials and processes



Genuine Computer Architecture

- **Genuine definition**: *Designing the organization and hardware to meet goals and functional requirements*
- The implementation of a computer involves two components
  - 1) **organization** (or microarchitecture)
  - 2) **hardware**: detailed logic design and packaging
- In this course, **computer architecture** covers all the three:
  - (1) ***ISA***,
  - (2) ***microarchitecture*** *or* ***organization***, and
  - (3) ***hardware***

Agenda

- Introduction
- Classes of Computers
- Defining Computer Architecture
- 1.4 Trends in Technology
- Trends in Power and Energy in ICs
- Trends in Cost
- Dependability
- Measuring Performance
- Quantitative Principles

Rapid Changes in Technology

- If an ISA is successful, it should survive rapid changes in *implementation technology*
- The designer of a computer should be aware of such rapid changes

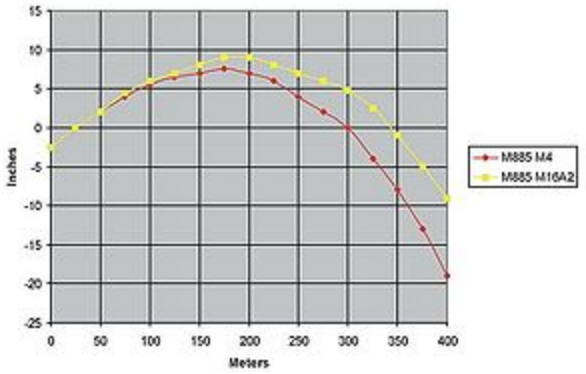
- **Five implementation technologies** are critical to model implementations of computer design

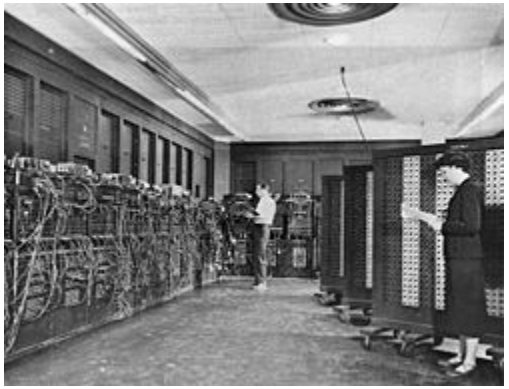
Five Critical Implementation Techs

Integrated circuit logic technology	• As Moore's law, the transistors count doubles every 18-24 months
Semiconductor DRAM	• The capacity per DRAM chip doubles every two to three years
Semiconductor Flash	• Capacity doubles roughly every two years
Magnetic disk technology	• Capacity doubles every three years
Network technology	• Bandwidth increases quickly

ENIAC - background

- Electronic Numerical Computer
- Eckert and Mauchly, University of Pennsylvania
- Purpose: *Trajectory tables for weapons*
- Started 1943, Finished 1946
  - Too late for war effort
  - Used until 1955





John Mauchly (1907-1980) &
J. Presper Eckert (1919-1995)

ENIAC - details

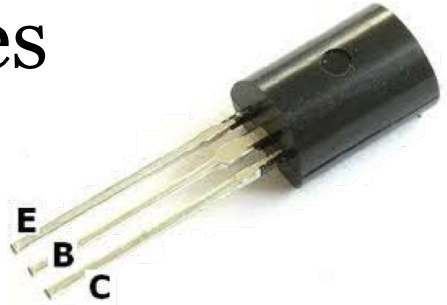
- Decimal (not binary)
- 20 accumulators of 10 digits
- Programmed manually by switches
- 18,000 **vacuum tubes**
- 30 tons
- 15,000 square feet
- 140 kW power consumption
- 5,000 additions per second

Straight ring/Overbeck counter

State	Q0	Q1	Q2	Q3
0	1	0	0	0
1	0	1	0	0
2	0	0	1	0
3	0	0	0	1
0	1	0	0	0
1	0	1	0	0
2	0	0	1	0
3	0	0	0	1
0	1	0	0	0

Transistors – the 2nd generation

- Replaced vacuum tubes
- Advantages
  - Smaller
  - Cheaper
  - Less heat dissipation
- Solid state device
- Made from Silicon (Sand)
- Invented 1947 at **Bell Labs**
- William Shockley et al.



First working transisto

John Bardeen, William Shockley and Walter Brattain at Bell Labs, 1948.

Transistor Based Computers

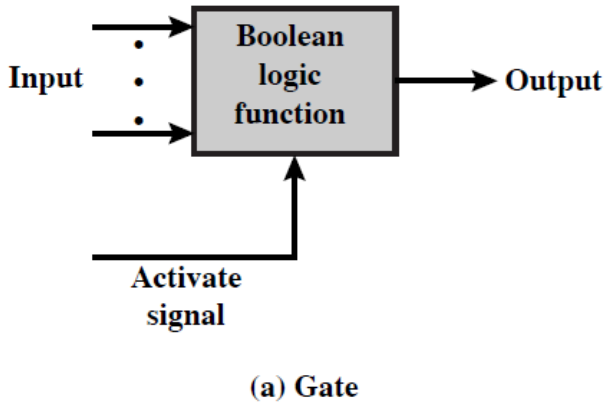
- Second generation machines
- NCR & RCA produced small transistor machines
- IBM
  - Produced IBM 7000
- DEC - 1957
  - Produced PDP-1

IBM 7000

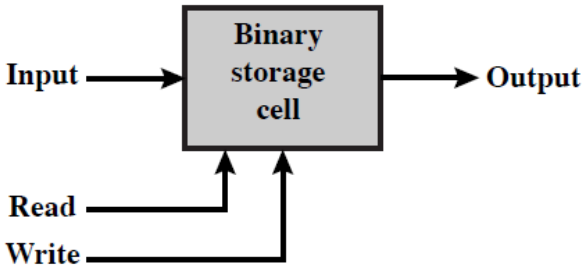
PDP-1

Microelectronics – 3rd Generation

- Literally - "small electronics"
- A computer is made up of gates, memory cells and interconnections
- These can be manufactured on a **semiconductor**
  - e.g. silicon wafer



(a) Gate



(b) Memory cell

Two Metrics: 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

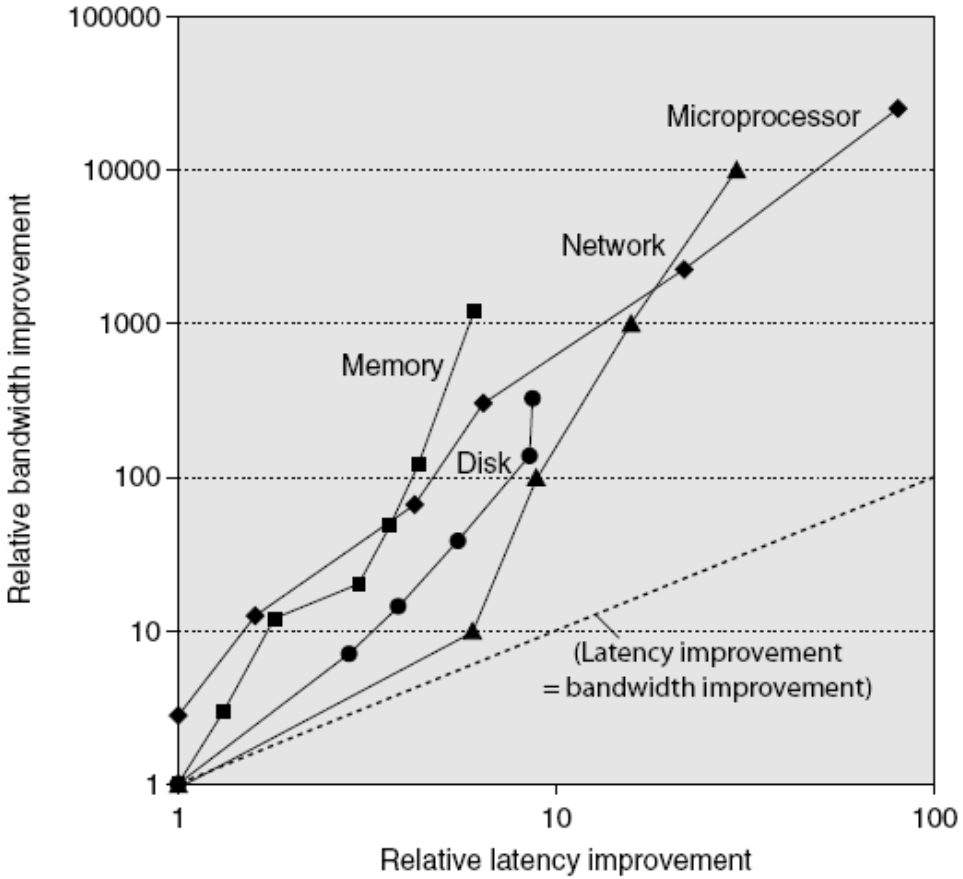
Microprocessor	16-bit address/ bus, microcoded	32-bit address/ bus, microcoded	5-stage pipeline, on-chip I & D caches, FPU	2-way superscalar, 64-bit bus	Out-of-order 3-way superscalar	Out-of-order superpipelined, on-chip L2 cache	Multicore OOO 4-way on chip L3 cache, Turbo
Product	Intel 80286	Intel 80386	Intel 80486	Intel Pentium	Intel Pentium Pro	Intel Pentium 4	Intel Core i7
Year	1982	1985	1989	1993	1997	2001	2010
Die size (mm^2)	47	43	81	90	308	217	240
Transistors	134,000	275,000	1,200,000	3,100,000	5,500,000	42,000,000	1,170,000,000
Processors/chip	1	1	1	1	1	1	4
Pins	68	132	168	273	387	423	1366
Latency (clocks)	6	5	5	5	10	22	14
Bus width (bits)	16	32	32	64	64	64	196
Clock rate (MHz)	12.5	16	25	66	200	1500	3333
Bandwidth (MIPS)	2	6	25	132	600	4500	50,000
Latency (ns)	320	313	200	76	50	15	4
Memory module	DRAM	Page mode DRAM	Fast page mode DRAM	Fast page mode DRAM	Synchronous DRAM	Double data rate SDRAM	DDR3 SDRAM
Module width (bits)	16	16	32	64	64	64	64
Year	1980	1983	1986	1993	1997	2000	2010
Mbits/DRAM chip	0.06	0.25	1	16	64	256	2048
Die size (mm^2)	35	45	70	130	170	204	50
Pins/DRAM chip	16	16	18	20	54	66	134
Bandwidth (MBytes/s)	13	40	160	267	640	1600	16,000
Latency (ns)	225	170	125	75	62	52	37
Local area network	Ethernet	Fast Ethernet	Gigabit Ethernet	10 Gigabit Ethernet	100 Gigabit Ethernet		
IEEE standard	802.3	803.3u	802.3ab	802.3ac	802.3ba		
Year	1978	1995	1999	2003	2010		
Bandwidth (Mbits/sec)	10	100	1000	10,000	100,000		
Latency (µsec)	3000	500	340	190	100		
Hard disk	3600 RPM	5400 RPM	7200 RPM	10,000 RPM	15,000 RPM	15,000 RPM	
Product	CDC WrenI 94145-36	Seagate ST41600	Seagate ST15150	Seagate ST39102	Seagate ST373453	Seagate ST3600057	
Year	1983	1990	1994	1998	2003	2010	
Capacity (GB)	0.03	1.4	4.3	9.1	73.4	600	
Disk form factor	5.25 inch	5.25 inch	3.5 inch	3.5 inch	3.5 inch	3.5 inch	
Media diameter	5.25 inch	5.25 inch	3.5 inch	3.0 inch	2.5 inch	2.5 inch	
Interface	ST-412	SCSI	SCSI	SCSI	SCSI	SAS	
Bandwidth (MBytes/s)	0.6	4	9	24	86	204	
Latency (ms)	48.3	17.1	12.7	8.8	5.7	3.6	

Milestones for

1) microprocessors,
2) memory,
3) networks, and
4) disks

Bandwidth and Latency

- Latency improved 6x to 80x
- Throughput improved 300x to 25,000x
- **Bandwidth has outpaced latency!**



Log-log plot of bandwidth and latency milestones

Feature Size and Its Impacts

Feature size is the minimum size of transistor or wire in x or y dimension

- **Integration density** scales quadratically with feature size
  - 10 microns in 1971 to .032 microns in 2011
- **Wire delay** does not scale well with feature size!
  - Wire delay is proportional to the product of resistance and capacitance

Agenda

- Introduction
- Classes of Computers
- Defining Computer Architecture
- Trends in Technology
- 1.5 Trends in Power and Energy in ICs
- Trends in Cost
- Dependability
- Measuring Performance
- Quantitative Principles

Power and Energy

Problem: Get power in, get power or heat out

- ***Thermal Design Power (TDP)***
  - Characterizes sustained power consumption
  - Used as target for power supply and cooling system
  - Lower than peak power, higher than average power consumption

- **Clock rate** can be reduced dynamically to limit power consumption

Dynamic Energy and Power

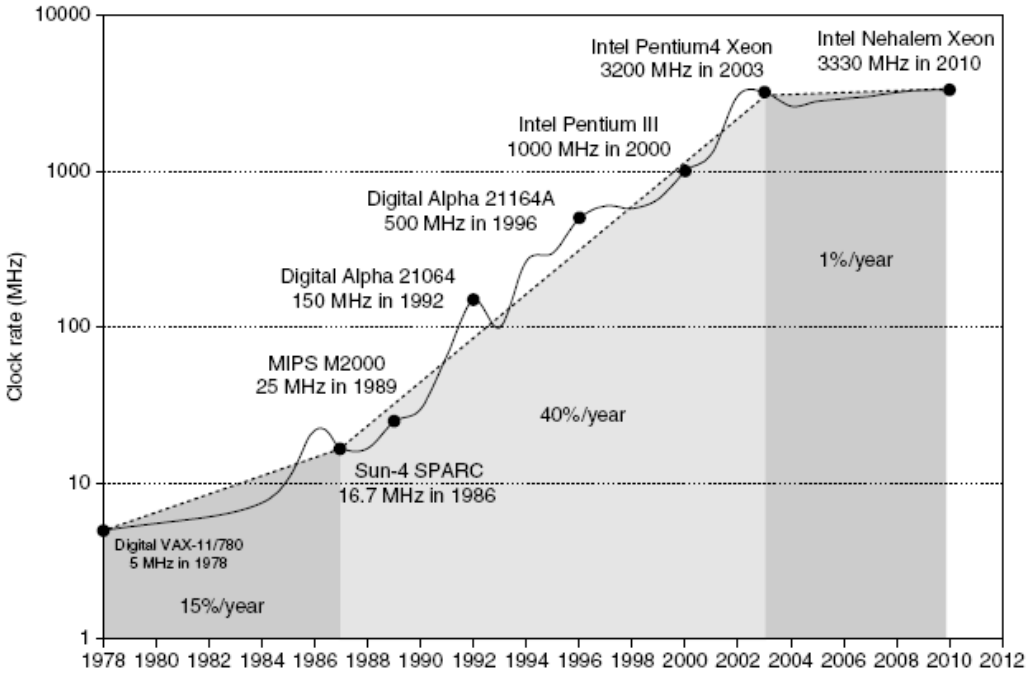
- ***Dynamic energy per transistor***
  - Used for a transistor switching from 0 -> 1 or 1 -> 0
  - ½ x Capacitive load x Voltage^2
  - The capacitive load is a function of *number of transistors* connected to output and *the technology* which determines the capacitance of wires and transistors
- ***Dynamic power*** *per transistor*
  - ½ x Capacitive load x Voltage^2 x Frequency switched

- **Voltage is the key**
  - The voltage of processors has become lower and lower

Power

- **Intel 80386** consumed ~ 2 W
- 3.3 GHz **Intel Core i7** consumes **130 W**
- Heat must be dissipated from 1.5 x 1.5 cm chip

- **This is the limit of what can be cooled by air**



Power vs. Energy

- **Power** is defined as the energy consumed in a unit time
- Question: Which metric should we use to compare processors: power or energy?

- It is better to use energy used for a specific task
  - Energy is tied to the a specific task and the time required for that task

Techniques for Reducing Power

Turn off the clock	• Turn off the clock of inactive modules
Dynamic Voltage-Frequency Scaling	• In periods of low activity, operate at lower frequency
Low power state for DRAM, disks	• In low power states, accesses are not allowed
Overclocking, turning off cores	• To turn off other cores and just run on a few cores

Static Power

- **Cause: The leakage current flows even when a transistor is off**

- $Power_{static} = Current_{static} \times Voltage$
  - Scales with number of transistors

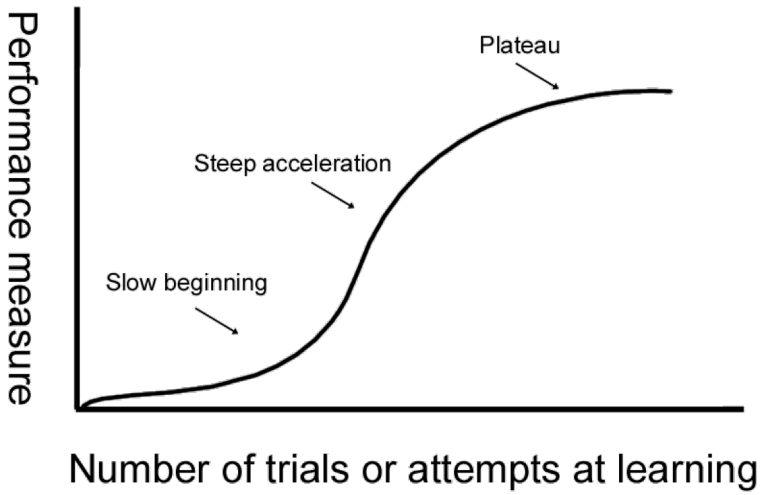
- To reduce static power: ***power gating***, <u>turning off the power supply</u>

Agenda

- Introduction
- Classes of Computers
- Defining Computer Architecture
- Trends in Technology
- Trends in Power and Energy in ICs
- 1.6 Trends in Cost
- Dependability
- Measuring Performance
- Quantitative Principles

Impact of Time, Volume, Commoditization

- ***Time***: The cost of a computer component decreases over time, why?
  - The learning curve!
  - ***Yield***: the percentage of manufactured devices that survives the testing procedure
- ***Volume*** is another factor in determining cost, why?
  - Amortized cost per computer
  - Microprocessors: price depends on volume, 10% less for each doubling of volume



learning curve

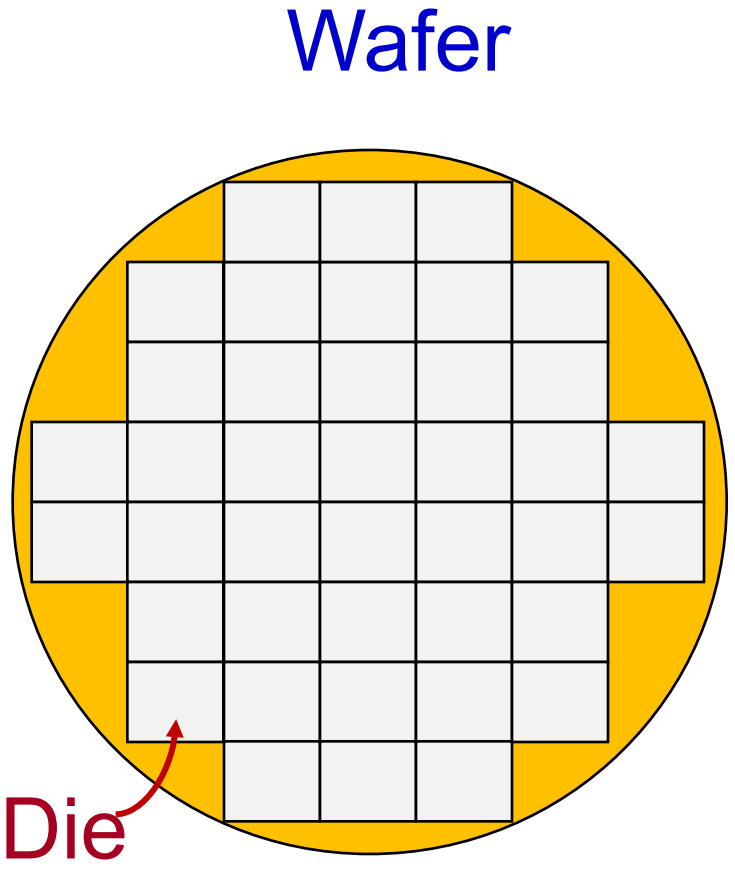
Impact of Time, Volume, Commoditization

Commodities are products that are sold by multiple vendors in large volumes and are identical

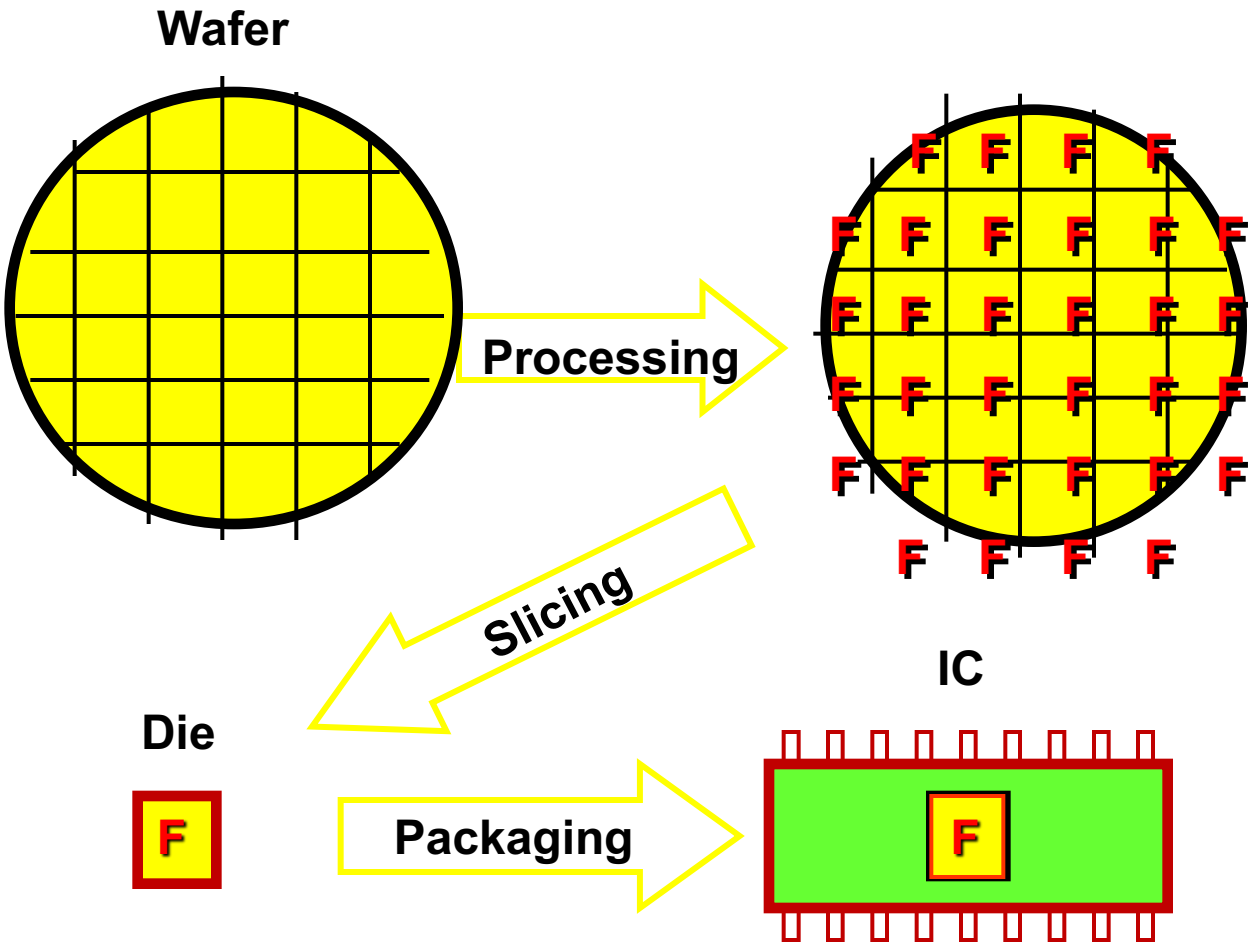
- Commoditization means that the market is highly competitive
  - 1) The gap between cost and price is **narrowed**
  - 2) It helps clearly define a product, and it increases the **competition** among the suppliers

Costs of Integrated Circuit (ICs)

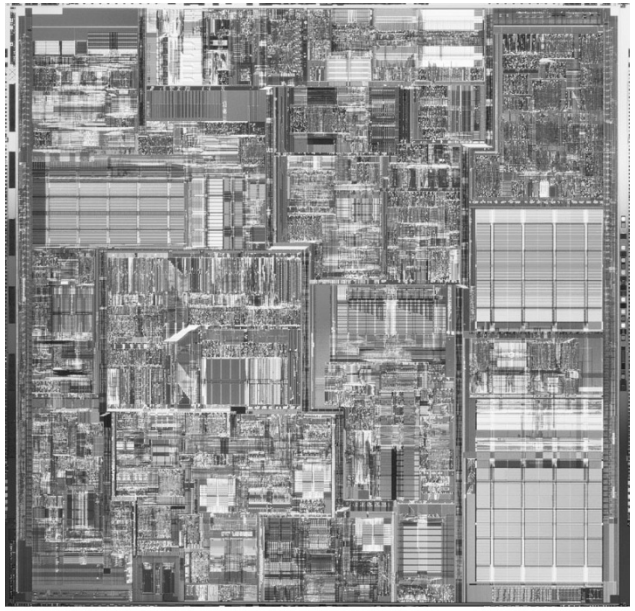
- Each copy of the integrated circuit appears in a ***die***
- Multiple dies are placed on each ***wafer***
- After fabrication, the individual dies are *separated, tested, and packaged*



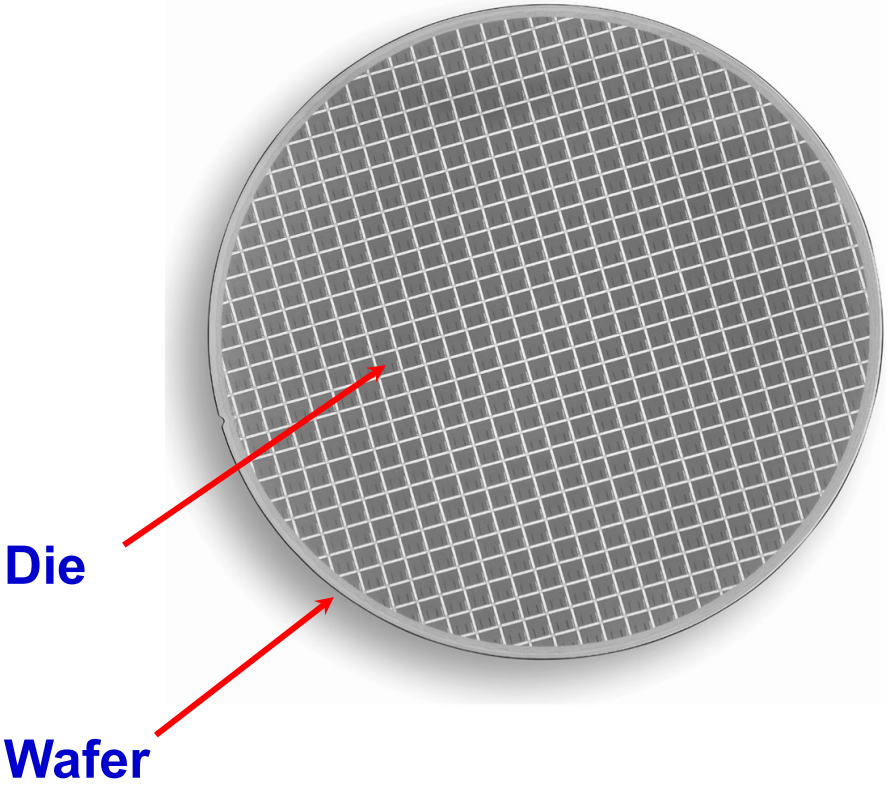
Wafer, Die, IC



Example of Pentium 4



Pentium 4 Processor



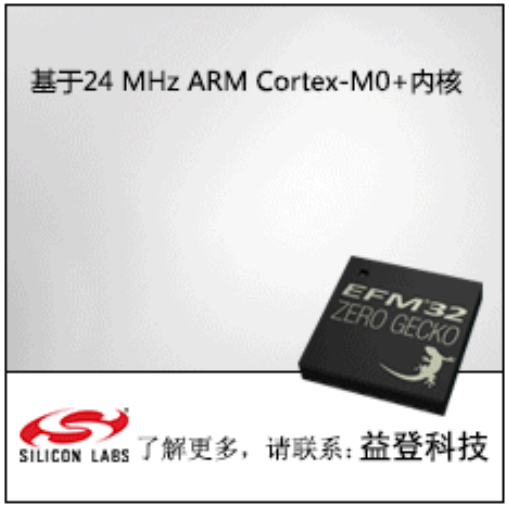
Typical Size of Industrial Wafers

首页 > 新闻 > 正文

全球首座18寸晶圆厂12月就绪

关键词： 18寸晶圆厂，半导体，台积电，晶圆

时间：2012-09-11 07:03:45　来源：苹果日报

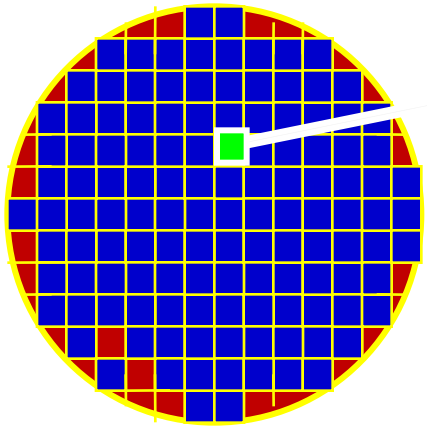


2012年国际半导体展闭幕，450mm（18寸）供应链论坛邀请到台积电、全球450mm联盟、应用材料、KLA-Tencor、LamResearch等深度探讨450mm未来发展蓝图，并率先预告世界第1座450mm晶圆厂将于今年12月准备就绪。

台积电派往全球450mm联盟的技术中心处长林进祥特别回台参加这次的论坛，他指出，世界第1座450mm晶圆厂将于今年12月准备就绪，而全球450mm联盟希望在2015年~2016年间建立18寸晶圆的测试生产线，可望陆续开始生产品质较好的生产晶圆。

Integrated Circuit Costs

$$\text{IC Cost} = \frac{\text{Die Cost} + \text{Testing cost} + \text{Packaging Cost}}{\text{Final Test Yield}}$$



Die

Processing

Testing

Packaging

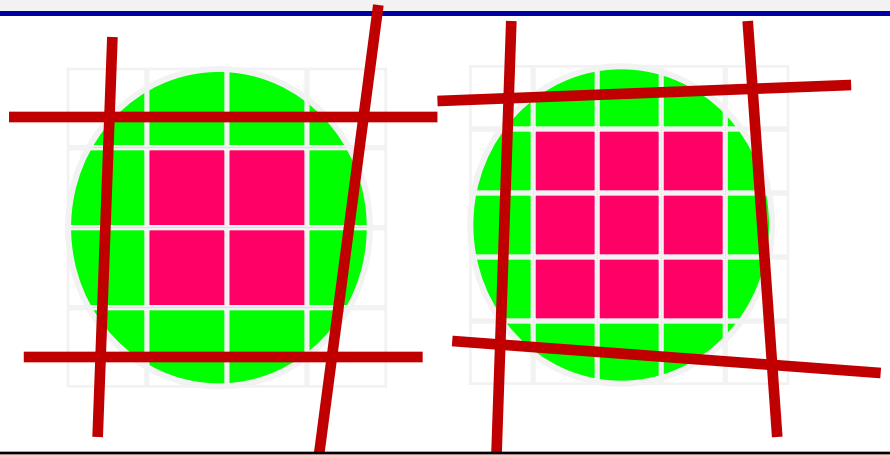
Testing

$$\text{Yields} = \frac{\text{good dies}}{\text{processed dies}}$$

Integrated Circuits Costs

$$\text{IC cost} = \frac{\text{Die cost} + \text{Testing cost} + \text{Packaging cost}}{\text{Final test yield}}$$

$$\text{Die cost} = \frac{\text{Wafer cost}}{\text{Dies per Wafer} \times \text{Die yield}}$$



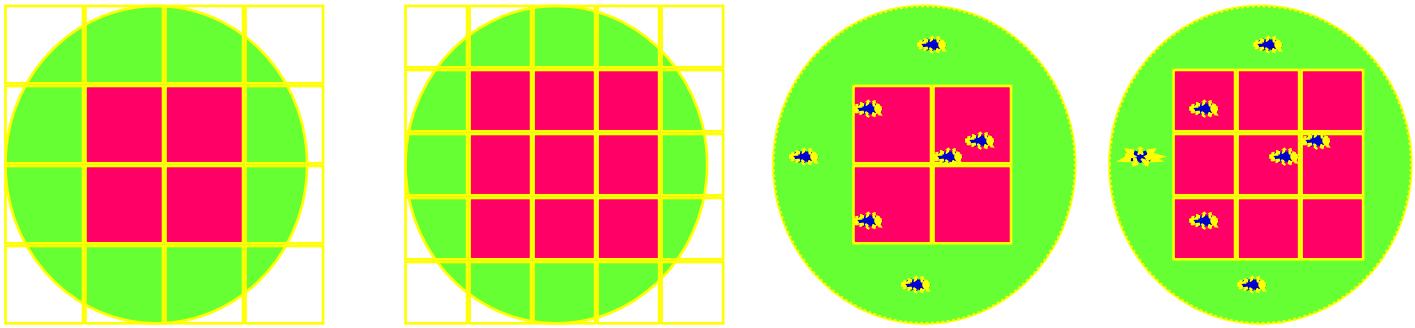
$$\text{Dies per Wafer} = \frac{\pi \left(\text{Wafer diameter}/2\right)^2}{\text{Die Area}} - \frac{\pi \left(\text{Wafer diameter}\right)}{\left(2 * \text{Die Area}\right)^{\frac{1}{2}}}$$

Example

- Find the number of dies per 20-cm wafer for a die that is 1.0 cm on a side and a die that is 1.5 cm on a side
- Answer

$$\text{Dies per Wafer} = \frac{\pi \left(\text{Wafer diameter}/2\right)^2}{\text{Die Area}} - \frac{\pi \left(\text{Wafer diameter}\right)}{\left(2 * \text{Die Area}\right)^{\frac{1}{2}}}$$

Integrated Circuit Cost



$$\text{Die yield} = \text{Wafer yield} \times 1/(1 + \text{Defects per unit area} \times \text{Die area})^{N}$$

Where N is a parameter inversely proportional to the number of mask Levels, which is a measure of the manufacturing complexity.

For today's CMOS process, good estimate is N = 11.5-15.5

Other Costs

$$\text{Die Test Cost} = \frac{\text{Test equipment Cost} * \text{Ave. Test Time}}{\text{Die Yield}}$$

Packaging Cost: depends on pins, heat dissipation, beauty, ...

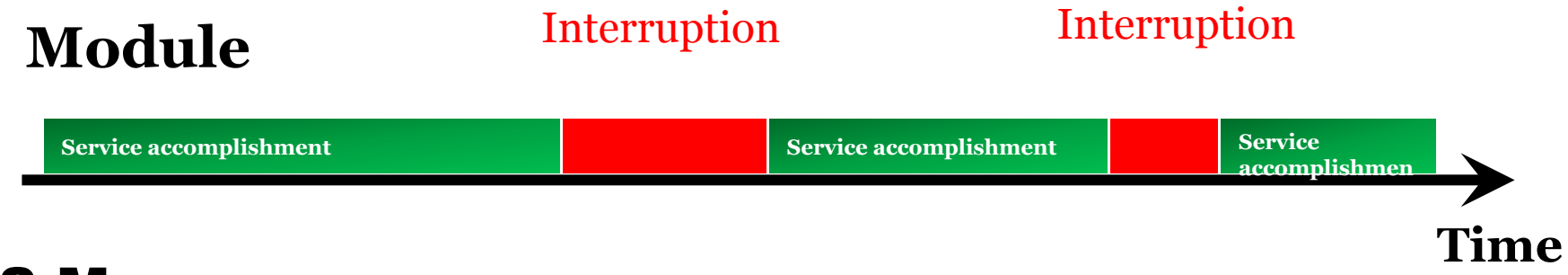
Chip	Die cost	Package			Test & assembly cost	Total
		pins	type	cost		
486DX2	$12	168	PGA	$11	$12	$35
Power PC 601	$53	304	QFP	$3	$21	$77
HP PA 7100	$73	504	PGA	$35	$16	$124
DEC Alpha	$149	431	PGA	$30	$23	$202
Super SPARC	$272	293	PGA	$20	$34	$326
Pentium	$417	273	PGA	$19	$37	$473

QFP: Quad Flat Package
PGA: Pin Grid Array
BGA: Ball Grid Array

Agenda

- Introduction
- Classes of Computers
- Defining Computer Architecture
- Trends in Technology
- Trends in Power and Energy in ICs
- Trends in Cost
- 1.7 Dependability
- Measuring Performance
- Quantitative Principles

Dependability



2 Measures:

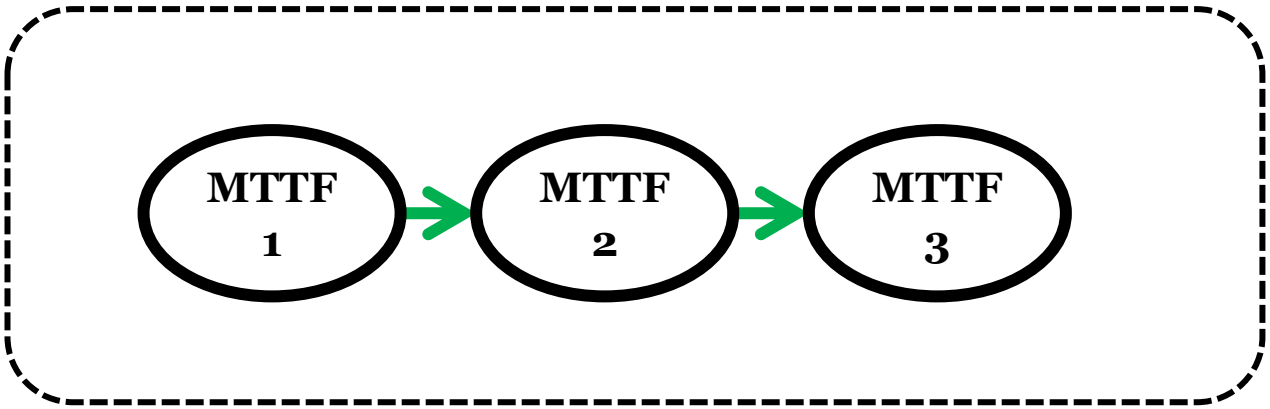
Module reliability

- Mean time to failure (MTTF)
- Mean time to repair (MTTR)
- Mean time between failures (MTBF) = MTTF + MTTR

Module availability

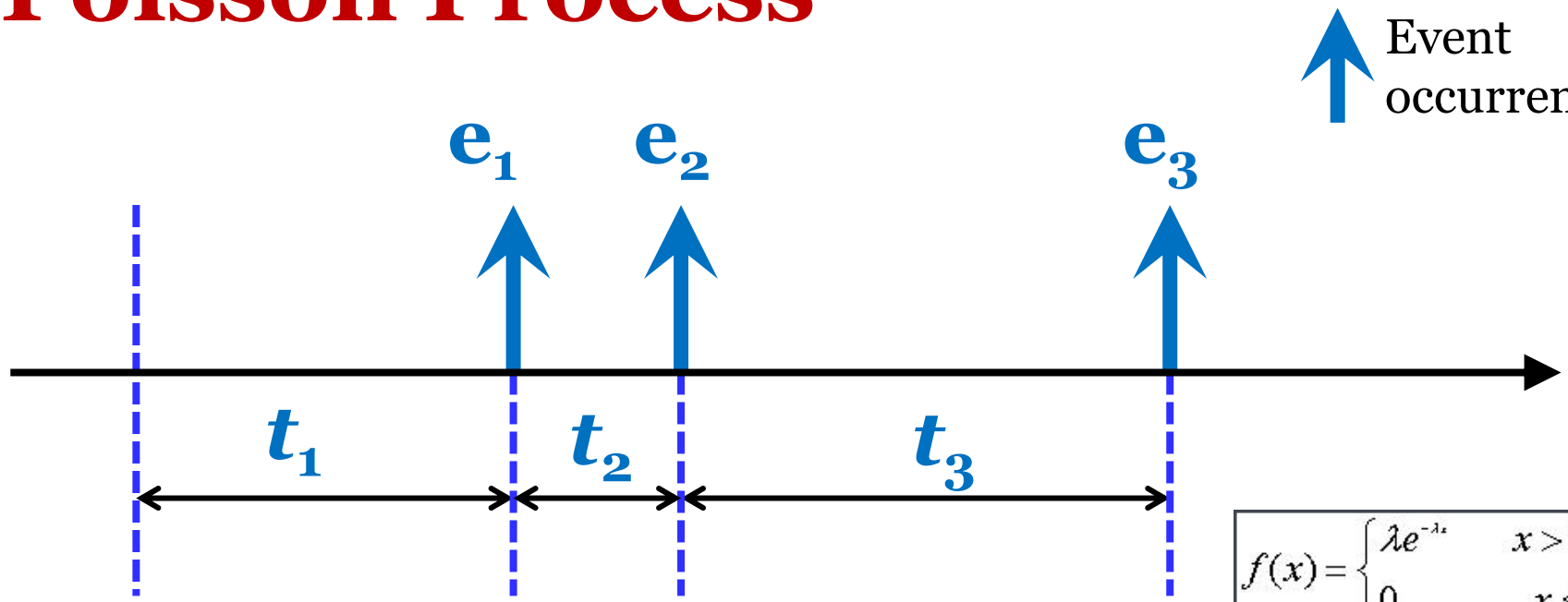
- Availability = MTTF / MTBF

MTTF of a System



- The MTTF of each module is *exponentially distributed* (independent of age)
- Modules are independent of each other
- What is the MTTF of the system?

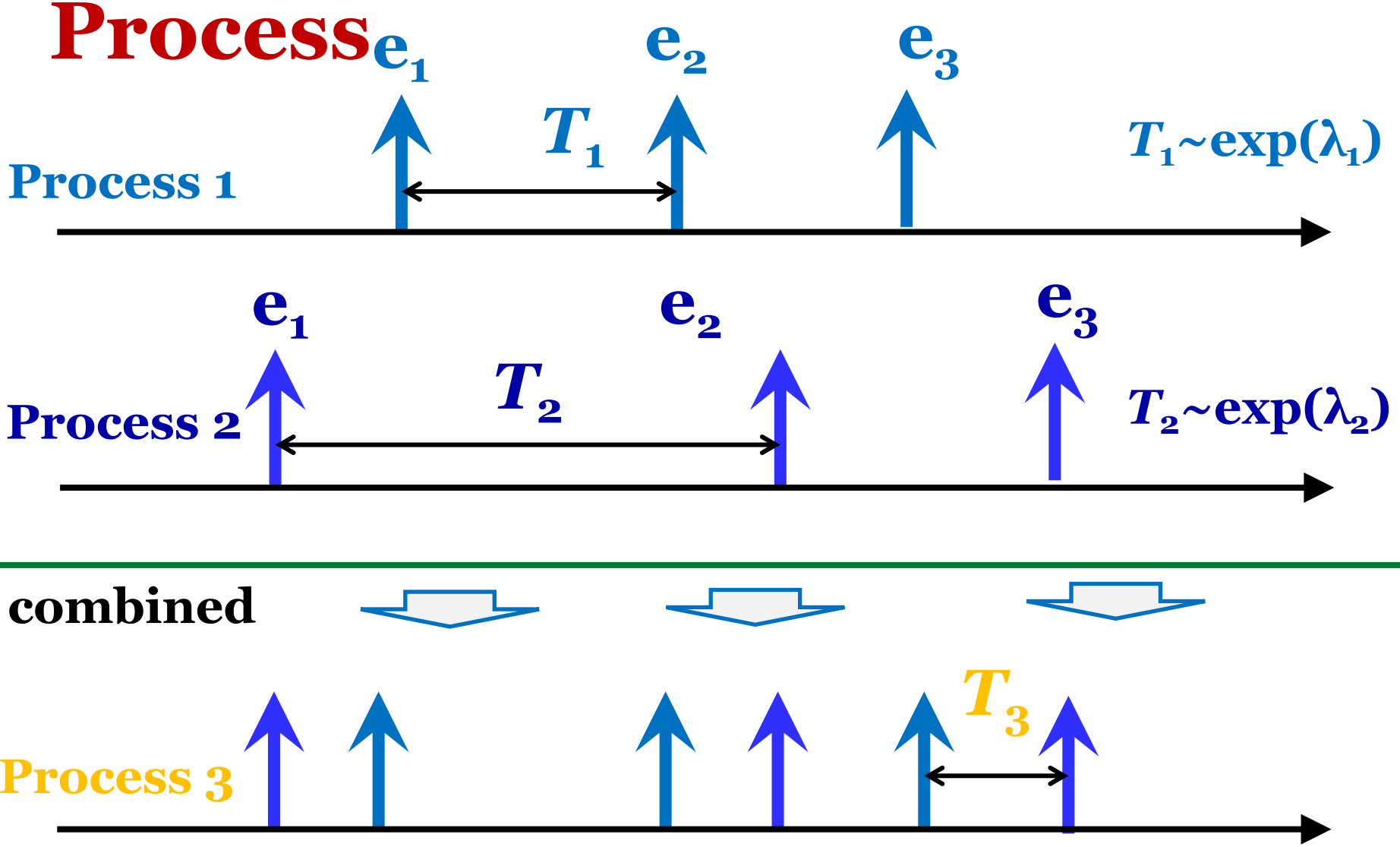
Exponential Distribution and Poisson Process



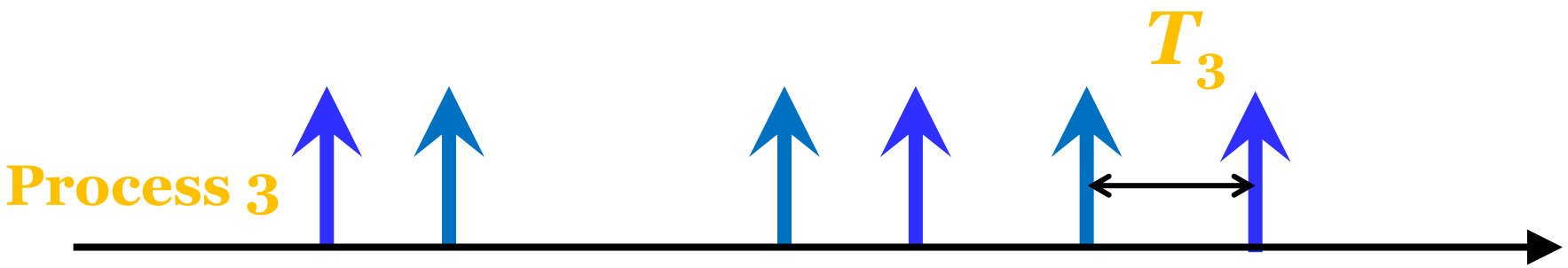
$$f(x)=\begin{cases}\lambda e^{-\lambda x} & x>0\\ 0 & x\le 0\end{cases}$$

- The time, *T*, between two events is a R.V. following the *exponential distribution*
- The arrivals of events follow the *Poisson Process*. Example, incoming calls of a hotline service

Characteristics of Poisson Process



Characteristics of Poisson Process

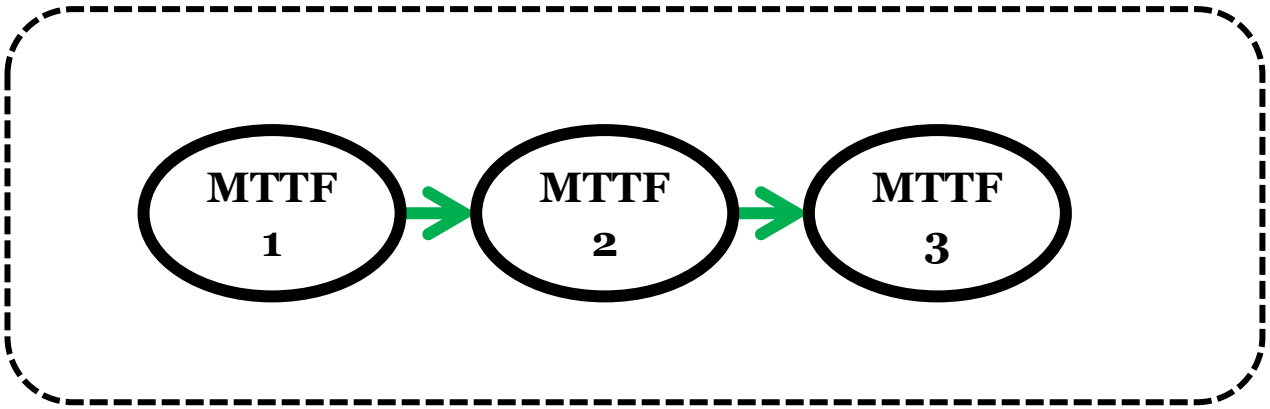


What process? T_3 follows what distribution?

$$T_3 \sim \exp(\lambda_3)$$

$$\lambda 3 = \lambda 1 + \lambda 2$$

MTTF of a System



System

What is the MTTF of the system?

$$MTTF_{system} = \frac{1}{\frac{1}{MTTF_1} + \frac{1}{MTTF_2} + \frac{1}{MTTF_3}}$$

Exercise

Assume a disk subsystem with the following components and MTTF:

- 10 disks, each rated at 1,000,000-hour MTTF
- 1 ATA controller, 500,000-hour MTTF
- 1 power supply, 200,000-hour MTTF
- 1 fan, 200,000-hour MTTF
- 1 ATA cable, 1,000,000-hour MTTF

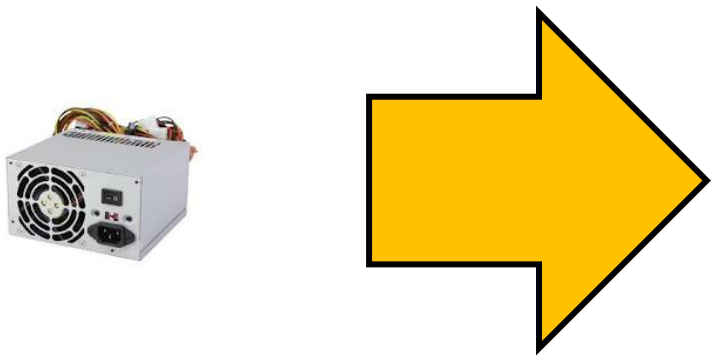
Assumptions: MTTF of the whole System?

- Exponential distributed
- Failures are independent

How to Increase Dependability?

Redundancy!

- In time (repeat the operation to see if it is still erroneous)
- In resources (duplicate resources)

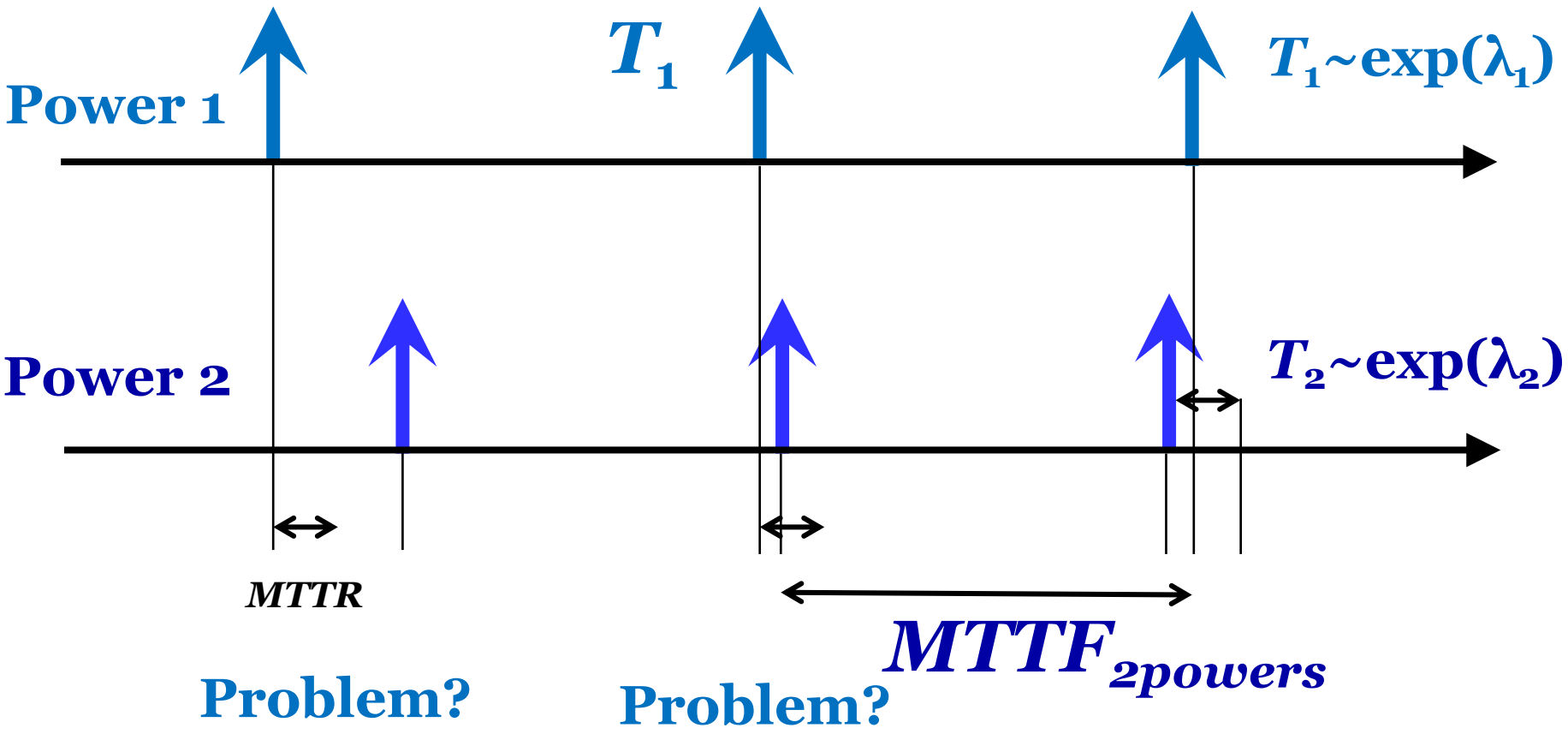


200,000-hour MTTF

10-hour MTTR

MTTF of two power supplies?

Computing the MTTF of Two Suppliers



$$MTTF_{2powers} = \frac{\frac{MTTF}{2}}{\frac{MTTR}{MTTF}}$$

Agenda

- Introduction
- Classes of Computers
- Defining Computer Architecture
- Trends in Technology
- Trends in Power and Energy in ICs
- Trends in Cost
- Dependability
- 1.8 Measuring Performance
- Quantitative Principles

Typical Performance Metrics

Response time (or execution time)

- The time between the start and the completion of an event

Throughput

- The total amount of work done in a unit time

Benchmarks

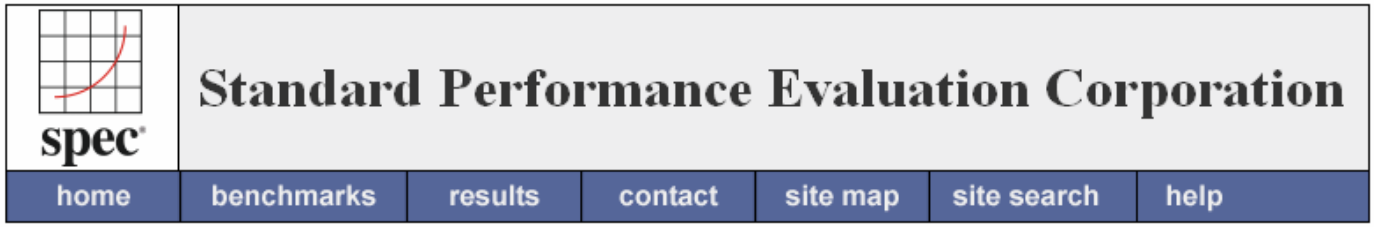
Benchmark: a common program for testing the execution times of computers

Programs lead to poor performance indication

- Kernels (e.g. matrix multiply)
- Toy programs (e.g. sorting)
- Synthetic benchmarks (e.g. Dhrystone)

Benchmark suit: collection of benchmark programs

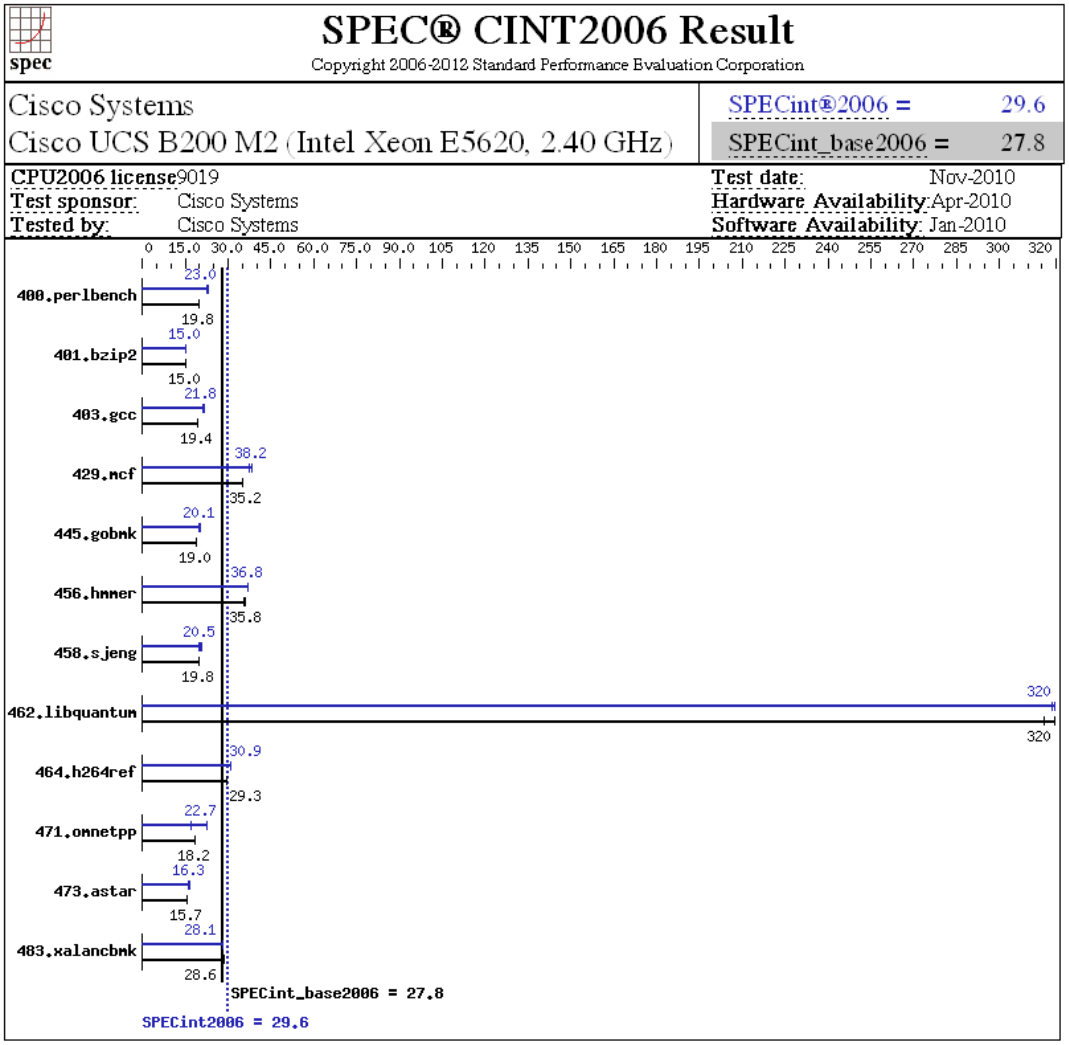
SPEC



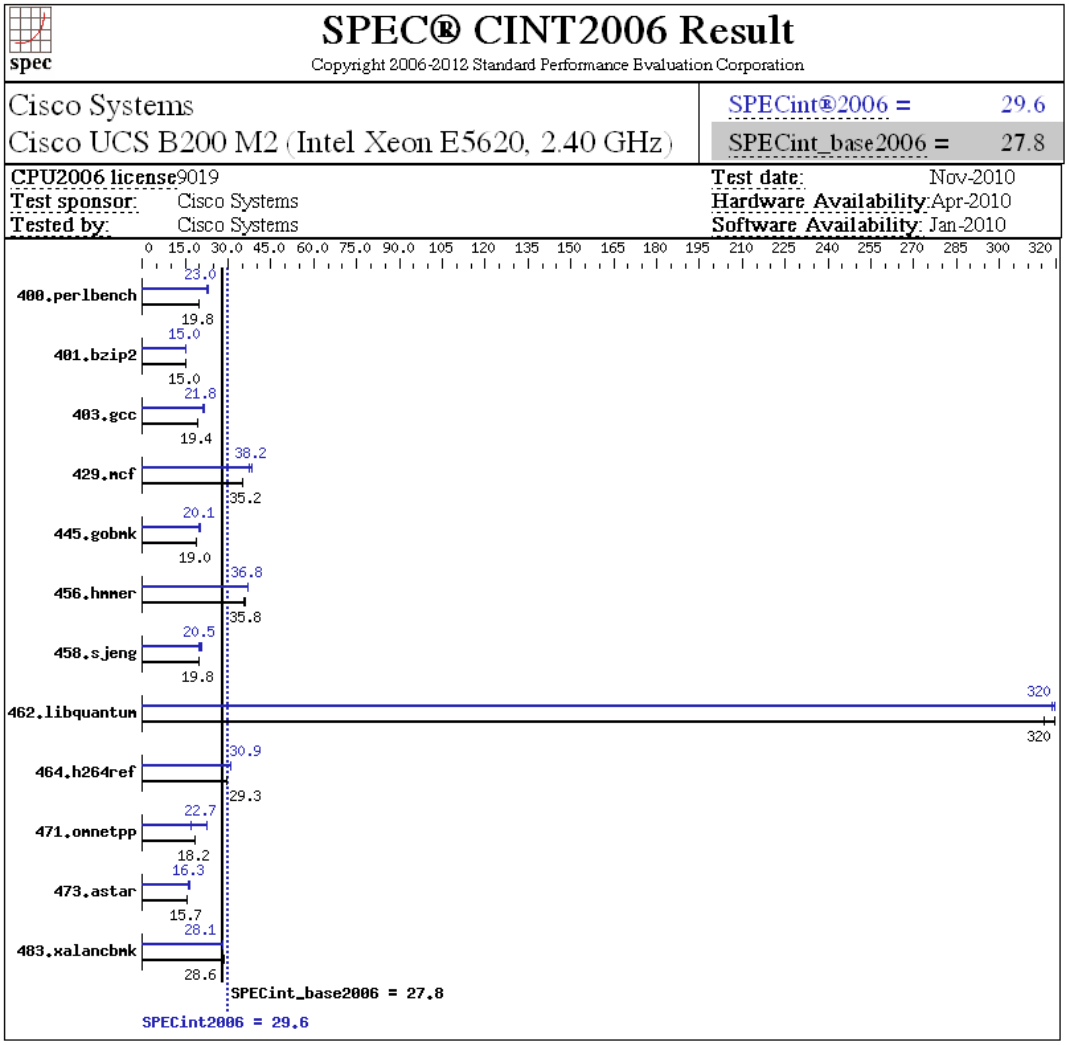
SPEC: Standard Performance Evaluation Corporation

- *A non-profit corporation* formed to establish, maintain and endorse a standardized set of relevant benchmarks that can be applied to the *newest generation of high-performance computers*.
- SPEC develops *benchmark suites* and also reviews and publishes *submitted results* from member organizations and other benchmark licensees.

Example SPEC Result Report



Example SPEC Result Report



400.perlbench
SPEC CPU2006 Benchmark Description

Benchmark Name

400.perlbench

Benchmark Authors

Larry Wall, et. al.

Benchmark Program General Category

Programming language

Benchmark Description

400.perlbench is a cut-down version of Perl v5.8.7, the popular scripting language. SPEC's version of Perl has had most of OS-specific features removed. In addition to the core Perl interpreter, several third-party modules are used:

- SpamAssassin v2.61
- Digest-MD5 v2.33
- HTML-Parser v3.35
- MHonArc v2.6.8
- IO-stringy v1.205
- MailTools v1.60
- TimeDate v1.16

Sources for all of the freely-available components used in 400.perlbench can be found in $SPEC/redistributable_sources/original/400.perlbench/ on your SPEC CPU2006 DVD.

Reporting Performance Results

How do we summarize performance, given the execution times of a set of benchmarks?

- Example as shown on the right.

Benchmarks	Opteron time (sec)	Itanium 2 time (sec)
wupwise	51.5	56.1
swim	125.0	70.7
mgrid	98.0	65.8
applu	94.0	50.9
mesa	64.6	108.0
galgel	86.4	40.0
art	92.4	21.0
equake	72.6	36.3
facerec	73.6	86.9
ammp	136.0	132.0
lucas	88.8	107.0
fma3d	120.0	131.0
sixtrack	123.0	68.8
apsi	150.0	231.0

Option 1: Arithmetic Mean

- The arithmetic mean of $x_1, x_2, \ldots, x_n$ is

$$Mean_{arith} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The ***problem*** of using arithmetic mean?
  - Benchmark programs with longer execution times would become more important
  - Example: 4 numbers **(5, 6, 5, 7, 100)**

Option 2: Weighted Arithmetic Mean

- To add a ***weighting factor*** to each benchmark program
- The *question*: how do we set weight factors?

- ***Possible solution***: use weights to make programs execute an equal time on a reference computer.
- ***Problem***: the reference computer would become crucial

SPEC Approach: Performance Ratio

- Instead of using absolute execution times, use ***the ratio*** of performance to *a reference computer*
- **A good property of using performance ratio:** the selection of reference computer is **irrelevant**

$$1.25 = \frac{\text{SPECRatio}_A}{\text{SPECRatio}_B} = \frac{\dfrac{\text{Execution time}_{\text{reference}}}{\text{Execution time}_A}}{\dfrac{\text{Execution time}_{\text{reference}}}{\text{Execution time}_B}} = \frac{\text{Execution time}_B}{\text{Execution time}_A} = \frac{\text{Performance}_A}{\text{Performance}_B}$$

SPEC Approach: Geometric Mean (Cont)

- The ***geometric mean*** of $x_1, x_2, \ldots, x_n$ is

$$Mean_{Geometric} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- Two **good properties** of geometric mean
  - 1) The geometric mean of the ratios = the ratio of geometric means
  - 2) The choice of the reference computer is irrelevant

Agenda

- Introduction
- Classes of Computers
- Defining Computer Architecture
- Trends in Technology
- Trends in Power and Energy in ICs
- Trends in Cost
- Dependability
- Measuring Performance
- **1.9 Quantitative Principles**

Principles for Computer Design

Take advantage of parallelism

- Data level parallelism and task level parallelism
- Piplelining, set-associative caches
- Multicore, multiprocessor, vector

Principle of locality

- Program intends to reuse data and instructions they have used recently
- Temporal locality and spatial locality

Focus on common case

- Favor the frequent case over the infrequent case

Amdahl's Law

The law defines the speedup by using a particular feature

$$\text{Execution time}_{new} = \text{Execution time}_{old} \times \left((1 - \text{Fraction}_{enhanced}) + \frac{\text{Fraction}_{enhanced}}{\text{Speedup}_{enhanced}} \right)$$

$$\text{Speedup}_{overall} = \frac{\text{Execution time}_{old}}{\text{Execution time}_{new}} = \frac{1}{(1 - \text{Fraction}_{enhanced}) + \frac{\text{Fraction}_{enhanced}}{\text{Speedup}_{enhanced}}}$$

Amdahl's law states that the performance improvement of using a new feature is limited by the fraction of the time the new feature can be used.

Processor Performance

$$\text{CPU time} = \text{CPU clock cycles for a program} \times \text{Clock cycle time}$$

$$\text{CPU time} = \frac{\text{CPU clock cycles for a program}}{\text{Clock rate}}$$

DEFINITION:

$$\text{CPI} = \frac{\text{CPU clock cycles for a program}}{\text{Instruction count}}$$

$$\text{CPU time} = \text{Instruction count} \times \text{Cycles per instruction} \times \text{Clock cycle time}$$

$$\frac{\text{Instructions}}{\text{Program}} \times \frac{\text{Clock cycles}}{\text{Instruction}} \times \frac{\text{Seconds}}{\text{Clock cycle}} = \frac{\text{Seconds}}{\text{Program}} = \text{CPU time}$$

Three Factors for Processor Improvement

Clock cycle time	Hardware technology and organization
CPI	Organization and instruction set architecture
Instruction count	Instruction set architecture and compiler technology

Different instruction types having different CPIs

$$\text{CPU clock cycles} = \sum_{i=1}^{n} \text{IC}_i \times \text{CPI}_i$$

$$\text{CPU time} = \left(\sum_{i=1}^{n} \text{IC}_i \times \text{CPI}_i \right) \times \text{Clock cycle time}$$

$$\text{CPI} = \frac{\sum_{i=1}^{n} \text{IC}_i \times \text{CPI}_i}{\text{Instruction count}} = \sum_{i=1}^{n} \frac{\text{IC}_i}{\text{Instruction count}} \times \text{CPI}_i$$