History of Processor Performance

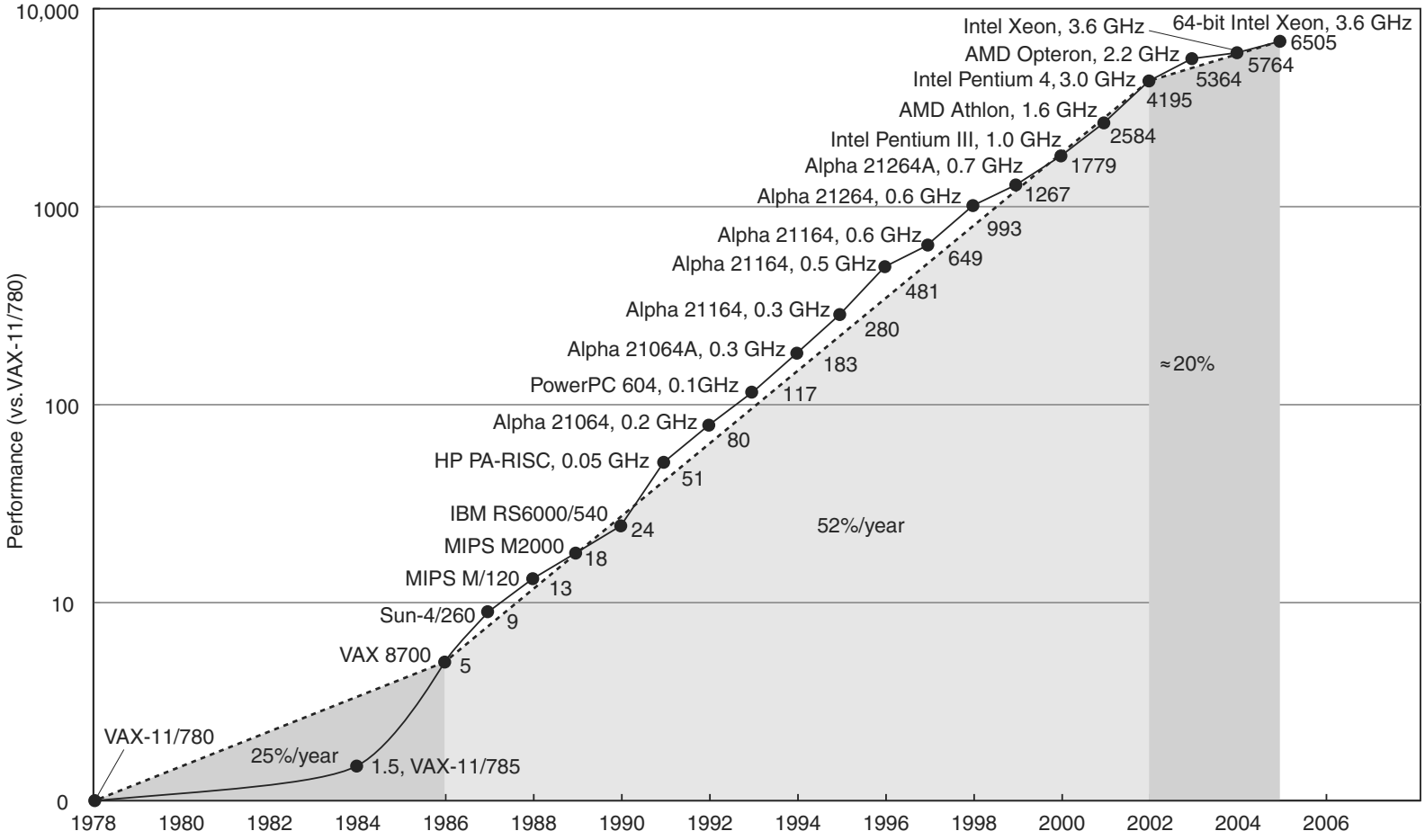


FIGURE 1.16 Growth in processor performance since the mid-1980s. This chart plots performance relative to the VAX 11/780 as measured by the SPECint benchmarks (see Section 1.8). Prior to the mid-1980s, processor performance growth was largely technology-driven and averaged about 25% per year. The increase in growth to about 52% since then is attributable to more advanced architectural and organizational ideas. By 2002, this growth led to a difference in performance of about a factor of seven. Performance for floating-point-oriented calculations has increased even faster. Since 2002, the limits of power, available instruction-level parallelism, and long memory latency have slowed uniprocessor performance recently, to about 20% per year. Copyright © 2009 Elsevier, Inc. All rights reserved.

History of Processor Performance

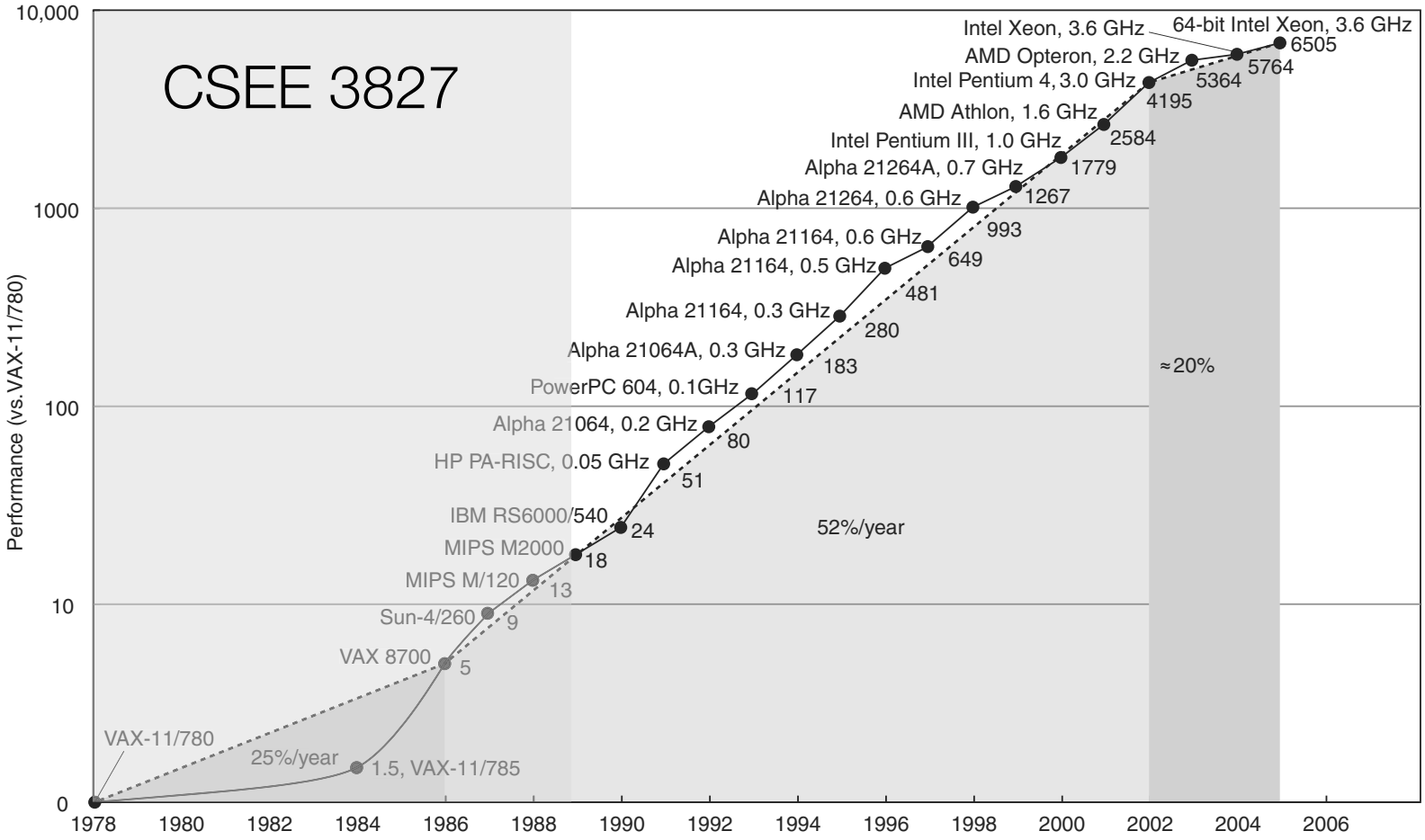


FIGURE 1.16 Growth in processor performance since the mid-1980s. This chart plots performance relative to the VAX 11/780 as measured by the SPECint benchmarks (see Section 1.8). Prior to the mid-1980s, processor performance growth was largely technology-driven and averaged about 25% per year. The increase in growth to about 52% since then is attributable to more advanced architectural and organizational ideas. By 2002, this growth led to a difference in performance of about a factor of seven. Performance for floating-point-oriented calculations has increased even faster. Since 2002, the limits of power, available instruction-level parallelism, and long memory latency have slowed uniprocessor performance recently, to about 20% per year. Copyright © 2009 Elsevier, Inc. All rights reserved.

History of Processor Performance

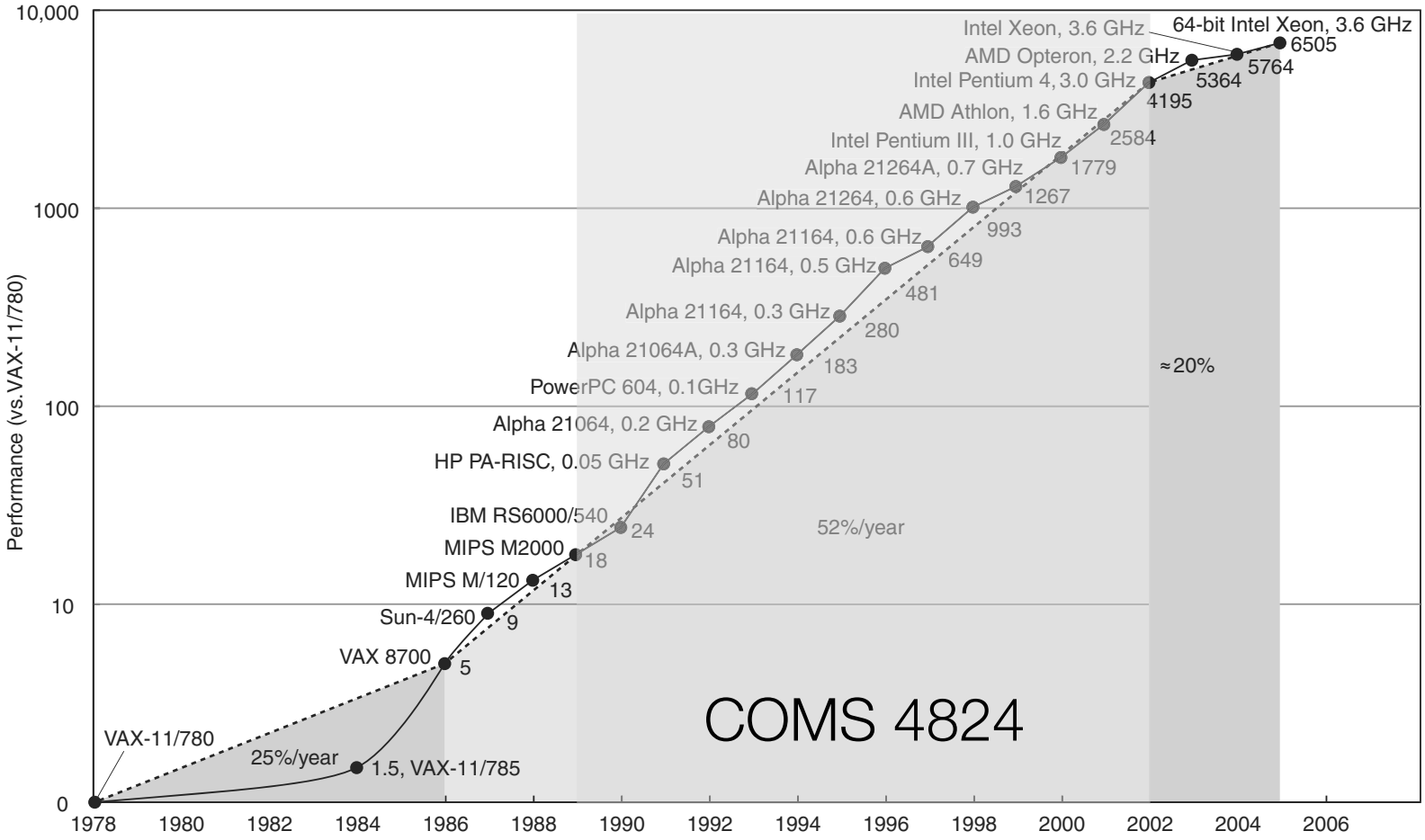
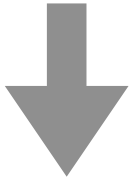


FIGURE 1.16 Growth in processor performance since the mid-1980s. This chart plots performance relative to the VAX 11/780 as measured by the SPECint benchmarks (see Section 1.8). Prior to the mid-1980s, processor performance growth was largely technology-driven and averaged about 25% per year. The increase in growth to about 52% since then is attributable to more advanced architectural and organizational ideas. By 2002, this growth led to a difference in performance of about a factor of seven. Performance for floating-point-oriented calculations has increased even faster. Since 2002, the limits of power, available instruction-level parallelism, and long memory latency have slowed uniprocessor performance recently, to about 20% per year. Copyright © 2009 Elsevier, Inc. All rights reserved.

Abstract Stages of Execution

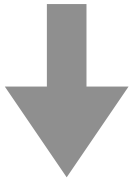
Instruction Fetch

(Instructions fetched from memory into CPU)



Instruction Issue / Execution

(Instructions executed on ALU or other functional unit)



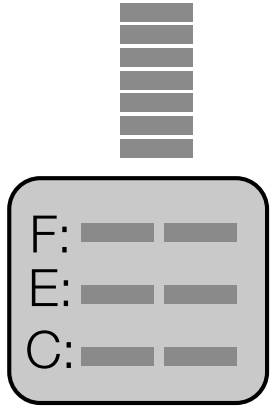
Instruction Completion / Commit

(Architectural state updated, i.e., regfile or memory)

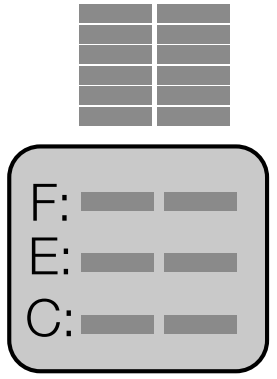
Multiple Instruction Issue Processors

Multiple instructions fetched, executed, and committed in each cycle

In superscalar processors instructions are scheduled by the HW



In VLIW processors instructions are scheduled by the SW



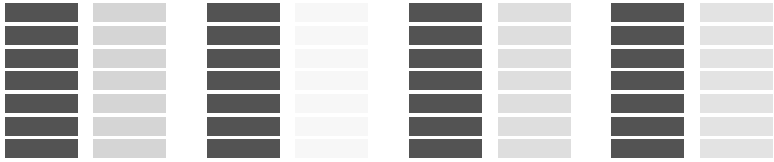
In all cases, the goal is to exploit instruction-level parallelism (ILP)

Flynn's Taxonomy

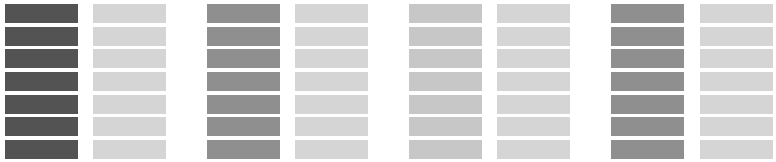
Single Instruction, Single Data (SISD)



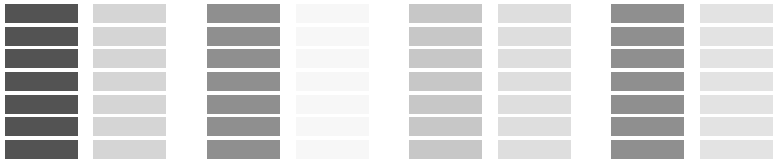
Single Instruction, Multiple Data (SIMD)



Multiple Instruction, Single Data (MISD)



Multiple Instruction, Multiple Data (MIMD)



Exploits instr-level parallelism (ILP)

Exploits data-level parallelism (DLP)

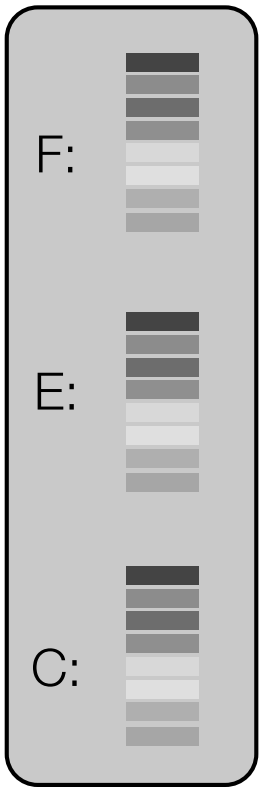
Out-of-order execution

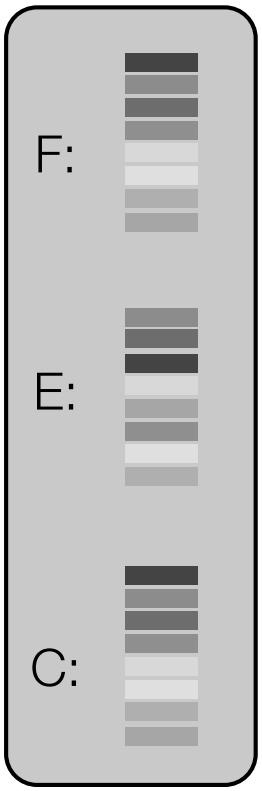
In in-order execution, instructions are fetched, executed, and committed in compiler order

In out-of-order execution (OOO), instructions are fetched, and committed in compiler, order; may be executed in some other order

One stalls, they all stall

Relatively simple HW





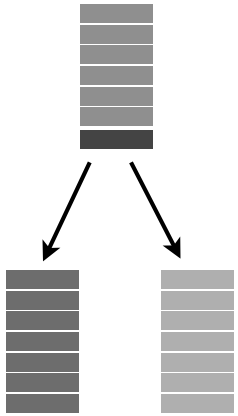
One stalls, independent instrs may proceed

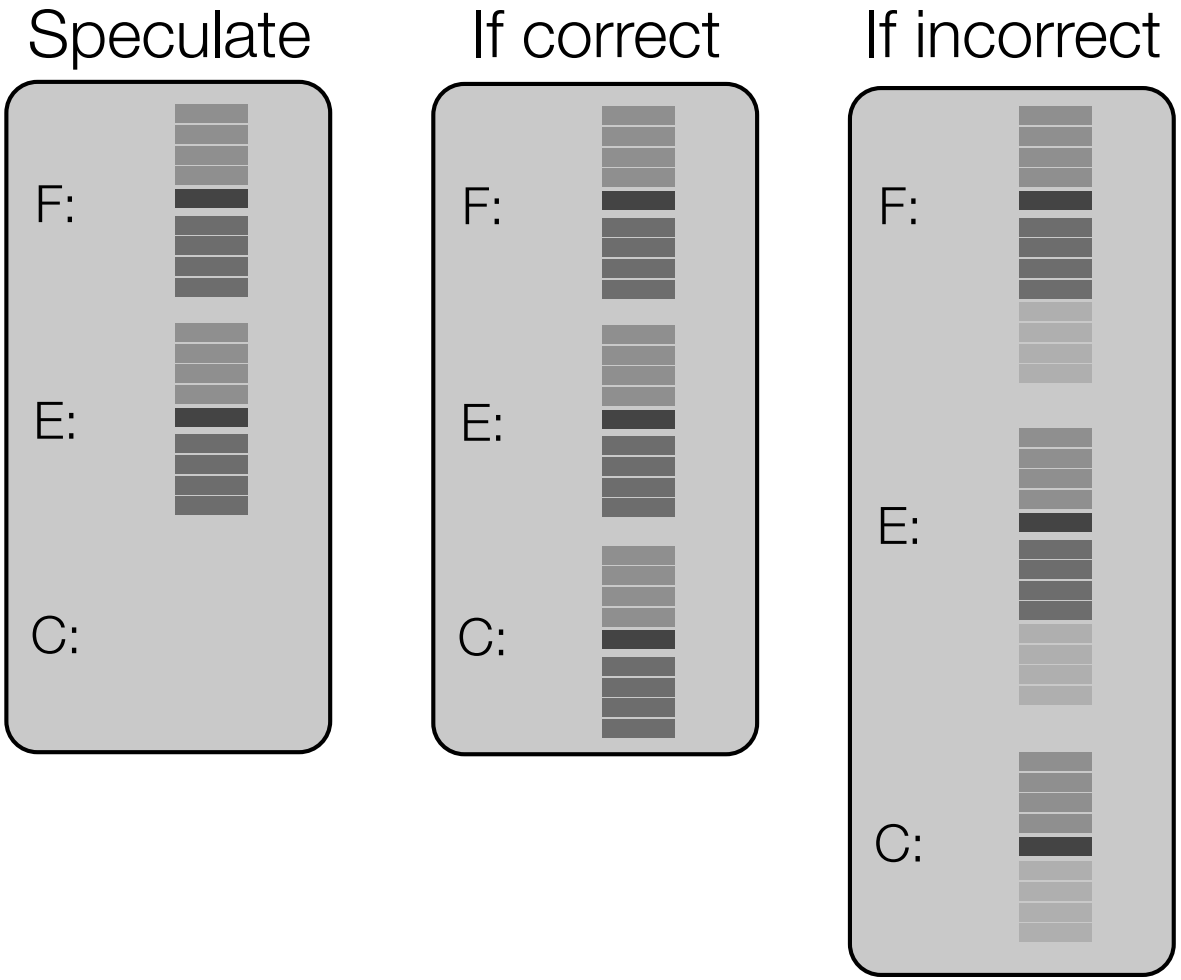
Additional hardware required for reordering

Another way to exploit instruction-level parallelism (ILP)

Speculation

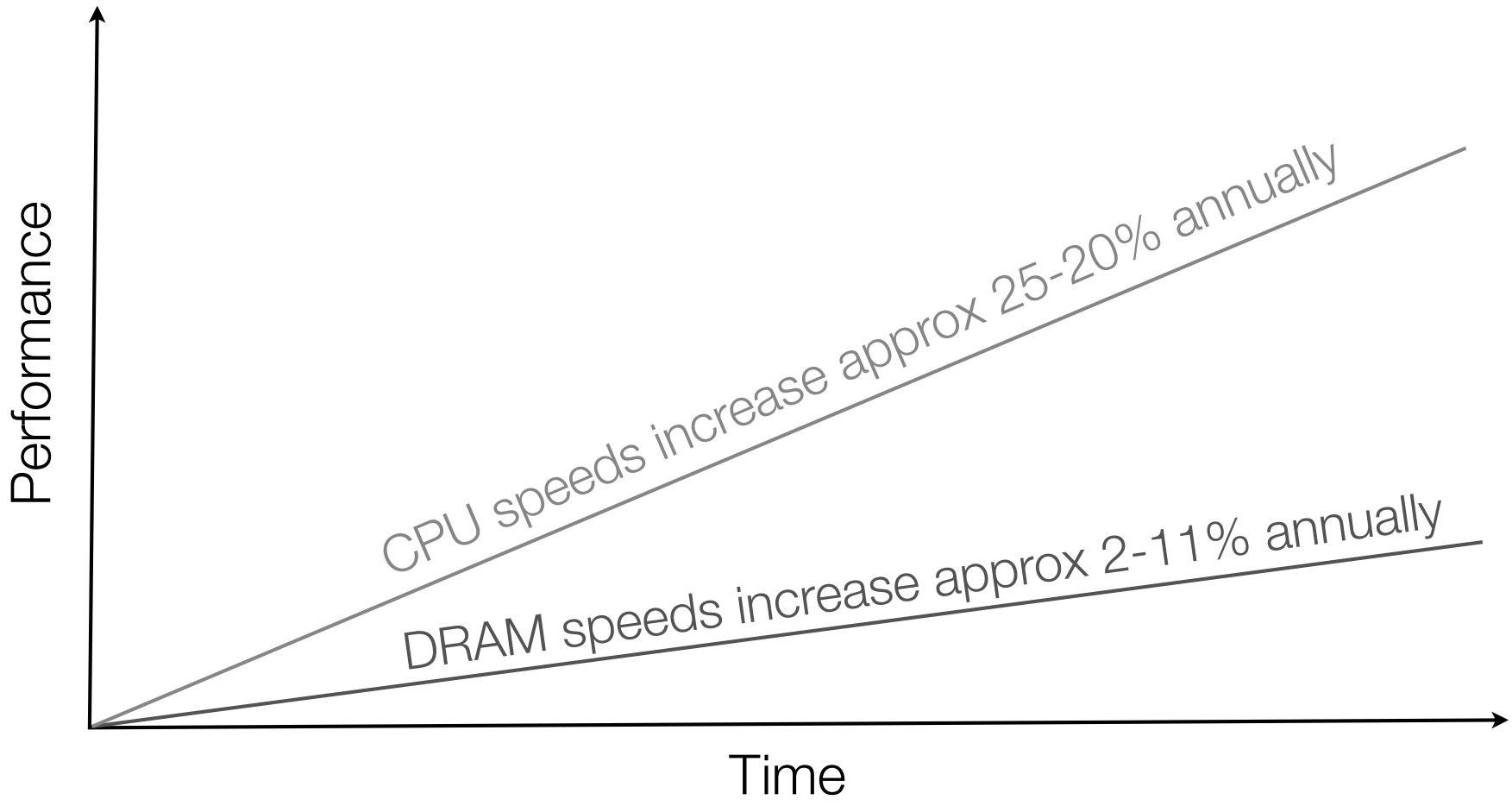
Speculation is executing an instruction before it is known that it should be executed





Mis-speculation executes excess instructions, costing time and power

The Memory Wall



A result of this gap is that cache design has increased in importance over the years. This has resulted in innovations such as victim caches and trace caches.

Modern Processor Performance

While single threaded performance has leveled, multithreaded performance potential scaling.

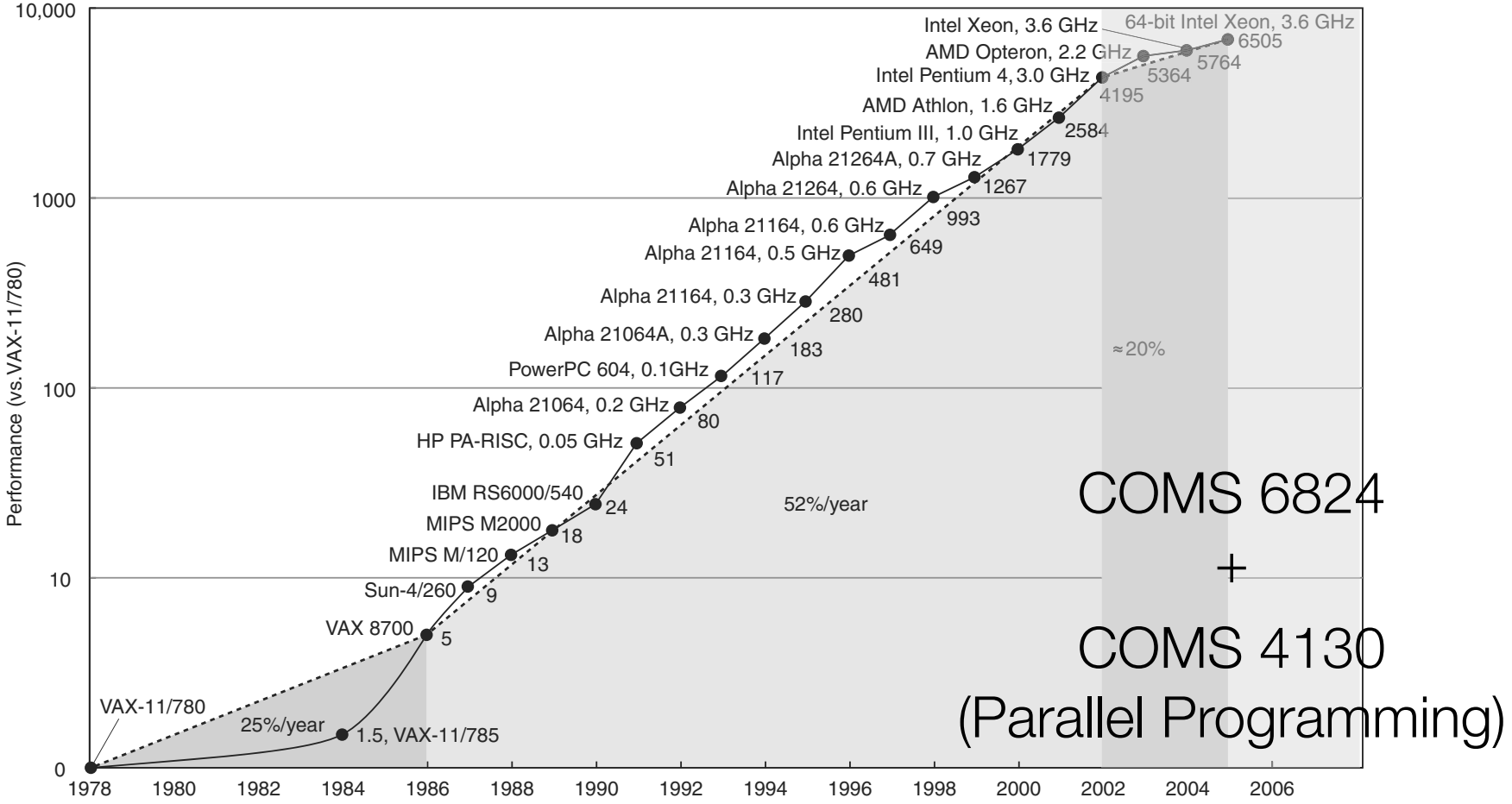
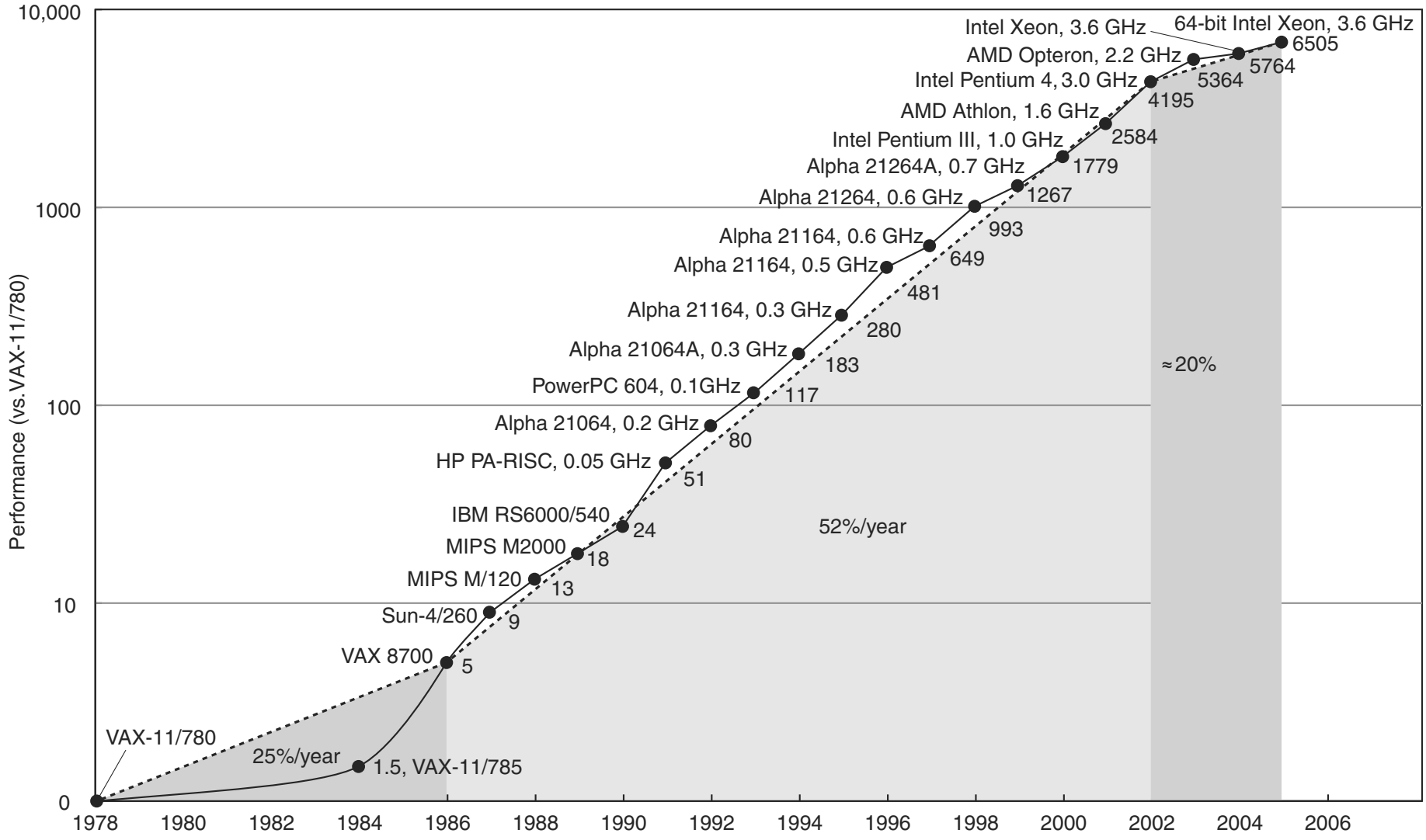
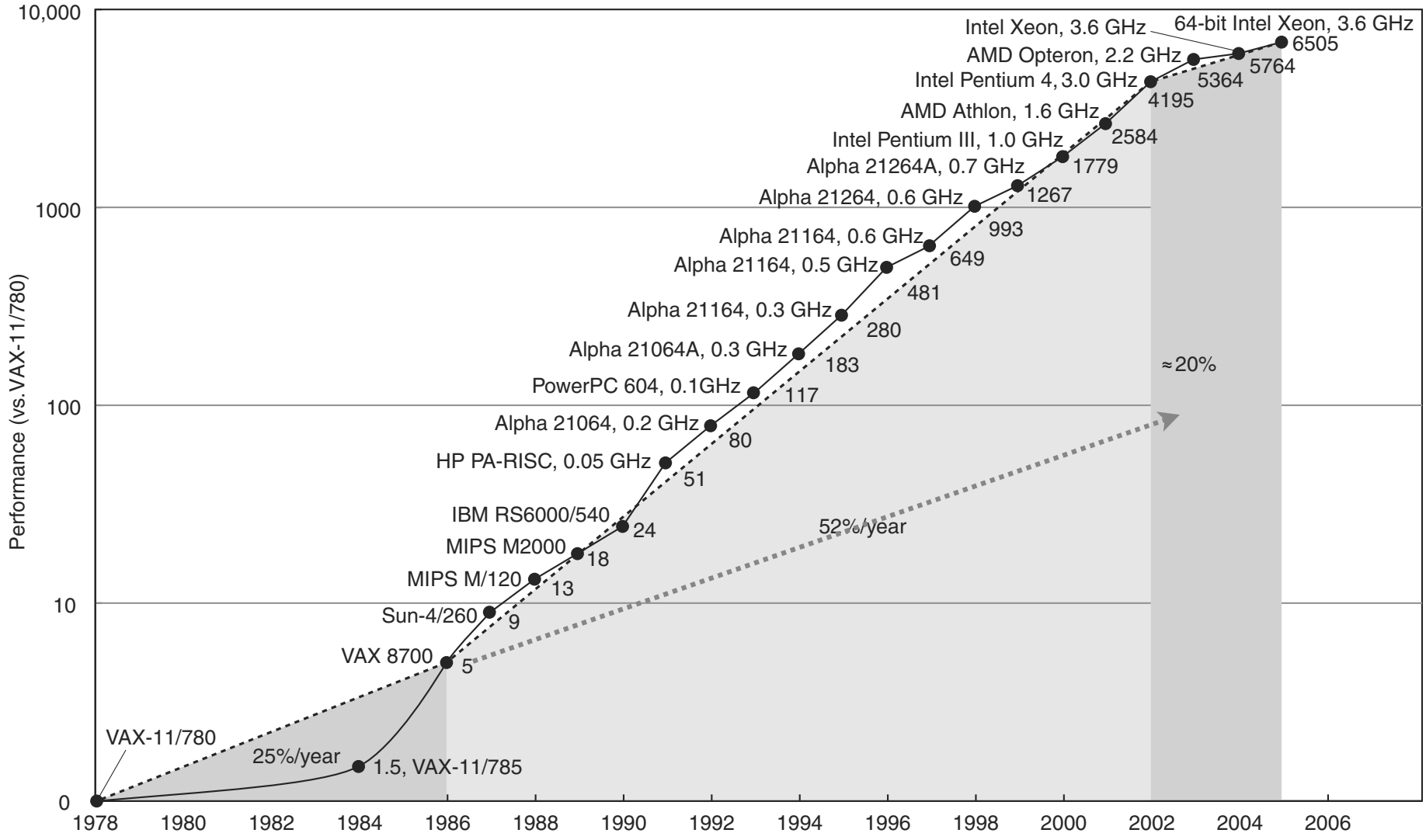


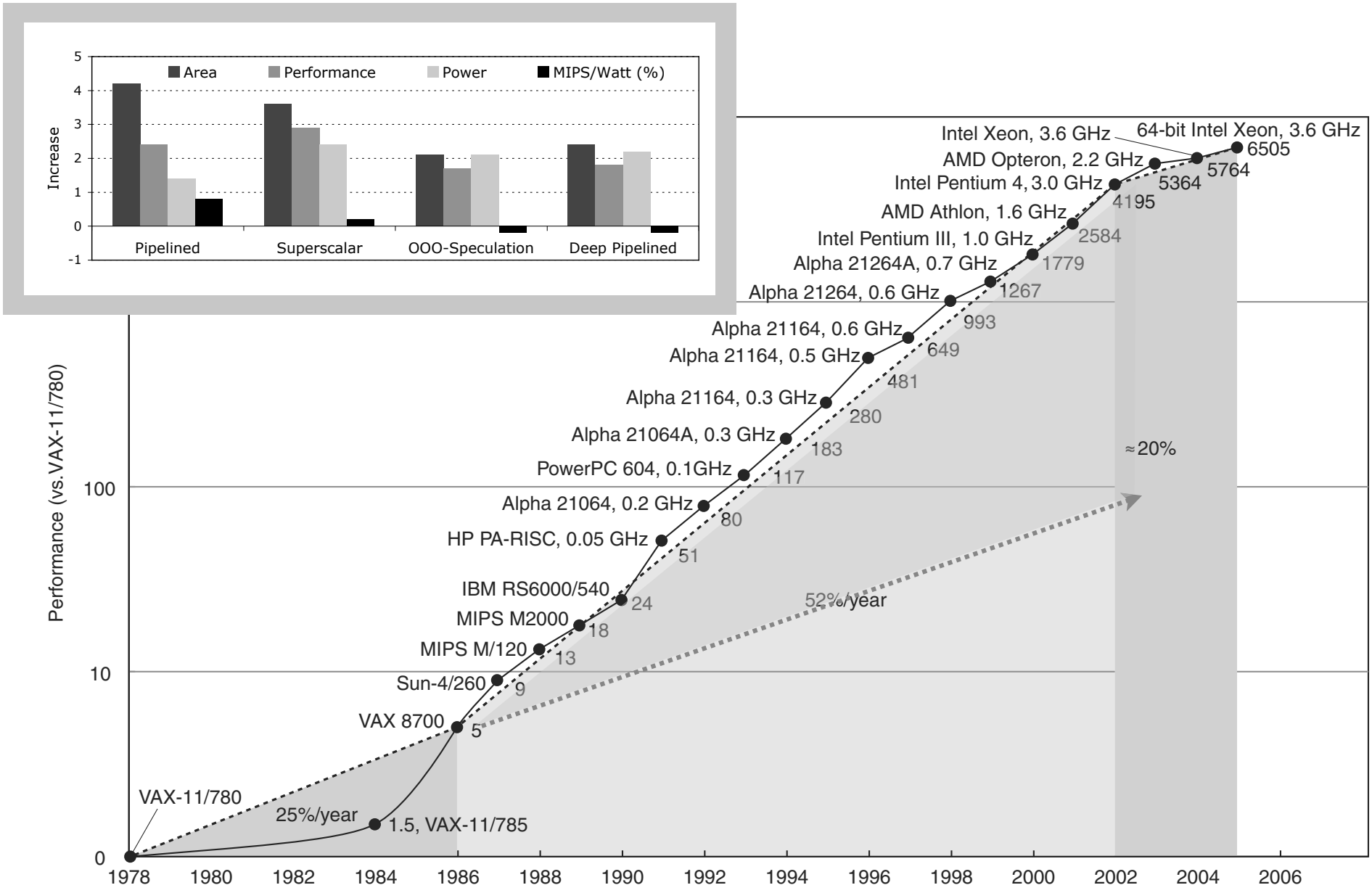
FIGURE 1.16 Growth in processor performance since the mid-1980s. This chart plots performance relative to the VAX 11/780 as measured by the SPECint benchmarks (see Section 1.8). Prior to the mid-1980s, processor performance growth was largely technology-driven and averaged about 25% per year. The increase in growth to about 52% since then is attributable to more advanced architectural and organizational ideas. By 2002, this growth led to a difference in performance of about a factor of seven. Performance for floating-point-oriented calculations has increased even faster. Since 2002, the limits of power, available instruction-level parallelism, and long memory latency have slowed uniprocessor performance recently, to about 20% per year. Copyright © 2009 Elsevier, Inc. All rights reserved.



Source: Hennessy and Patterson, "Computer Architecture: A Quantitative Approach"



Source: Hennessy and Patterson, "Computer Architecture: A Quantitative Approach"



Source: Hennessy and Patterson, "Computer Architecture: A Quantitative Approach"

The Power Wall

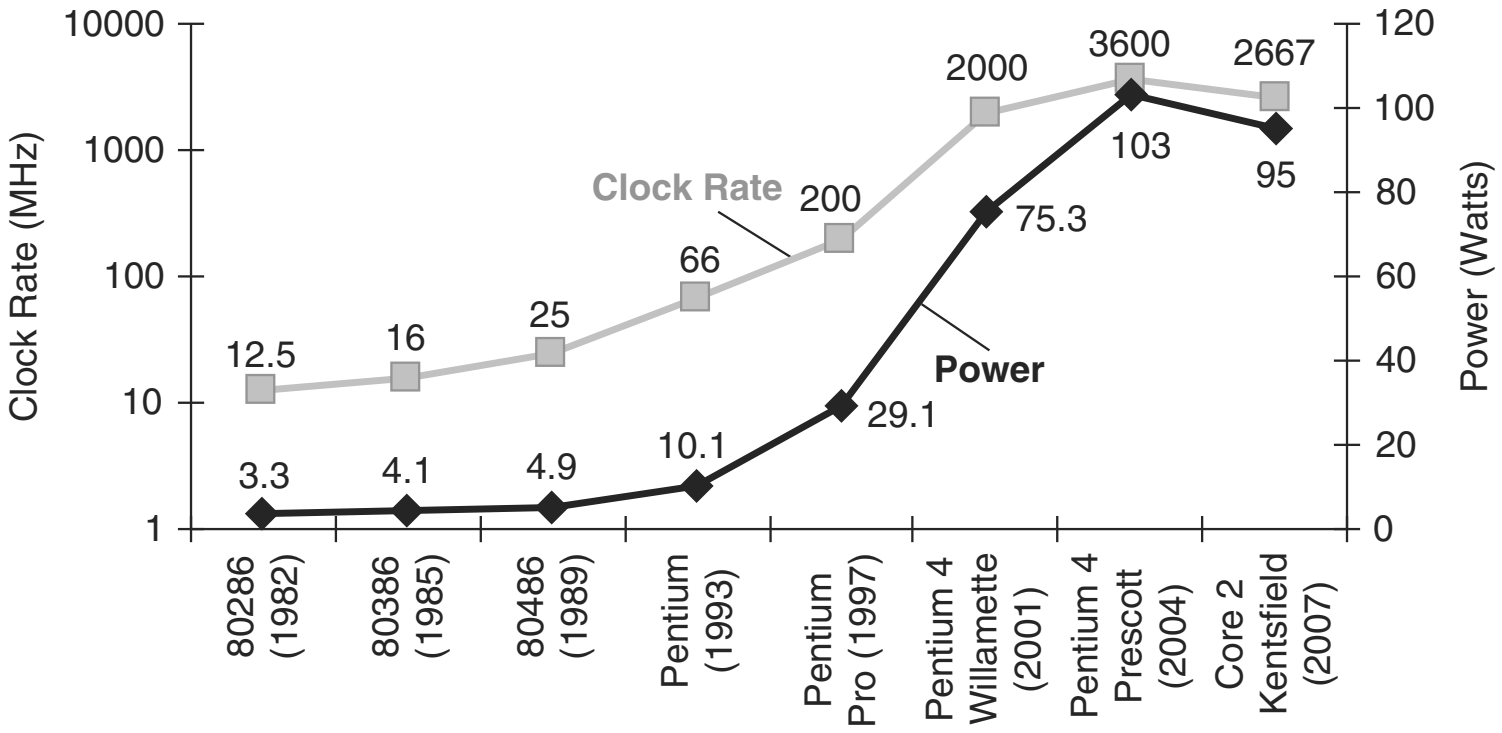
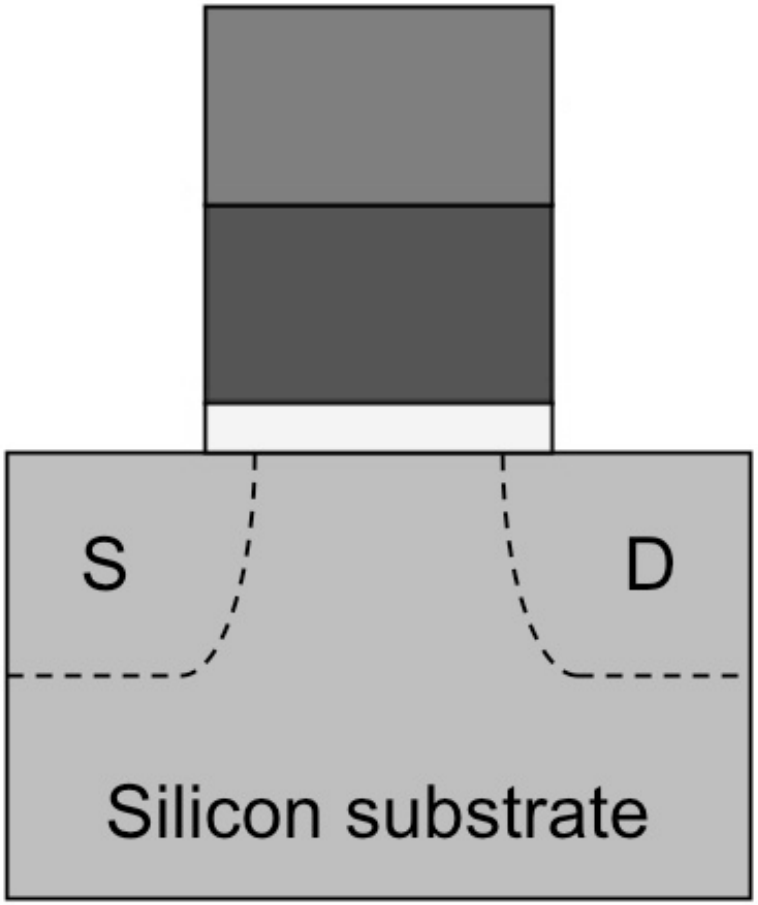


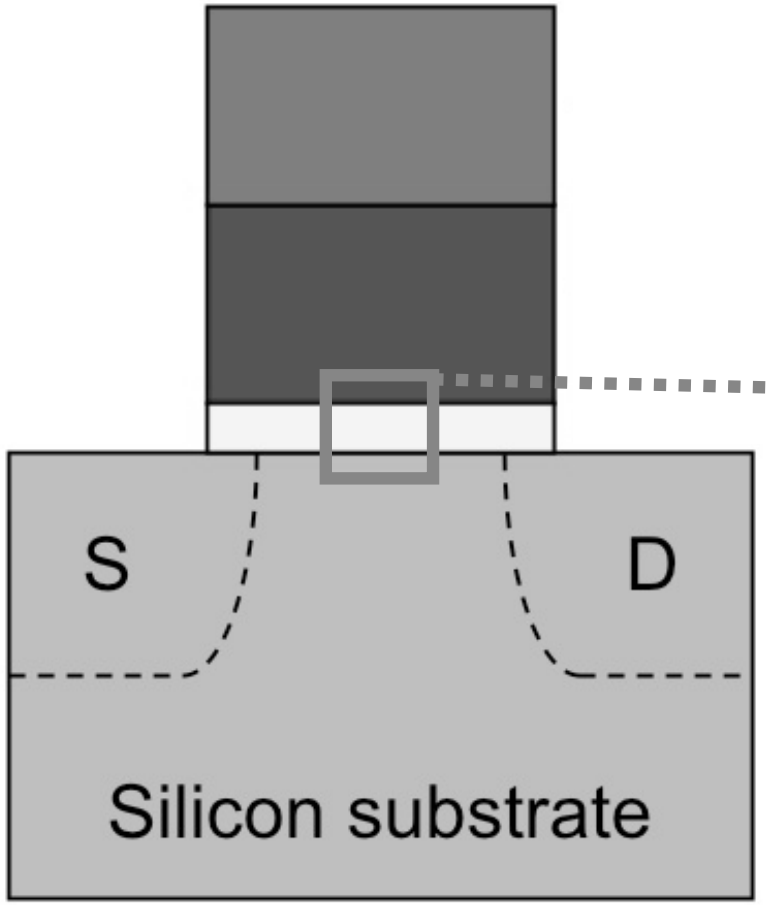
FIGURE 1.15 Clock rate and Power for Intel x86 microprocessors over eight generations and 25 years. The Pentium 4 made a dramatic jump in clock rate and power but less so in performance. The Prescott thermal problems led to the abandonment of the Pentium 4 line. The Core 2 line reverts to a simpler pipeline with lower clock rates and multiple processors per chip. Copyright © 2009 Elsevier, Inc. All rights reserved.

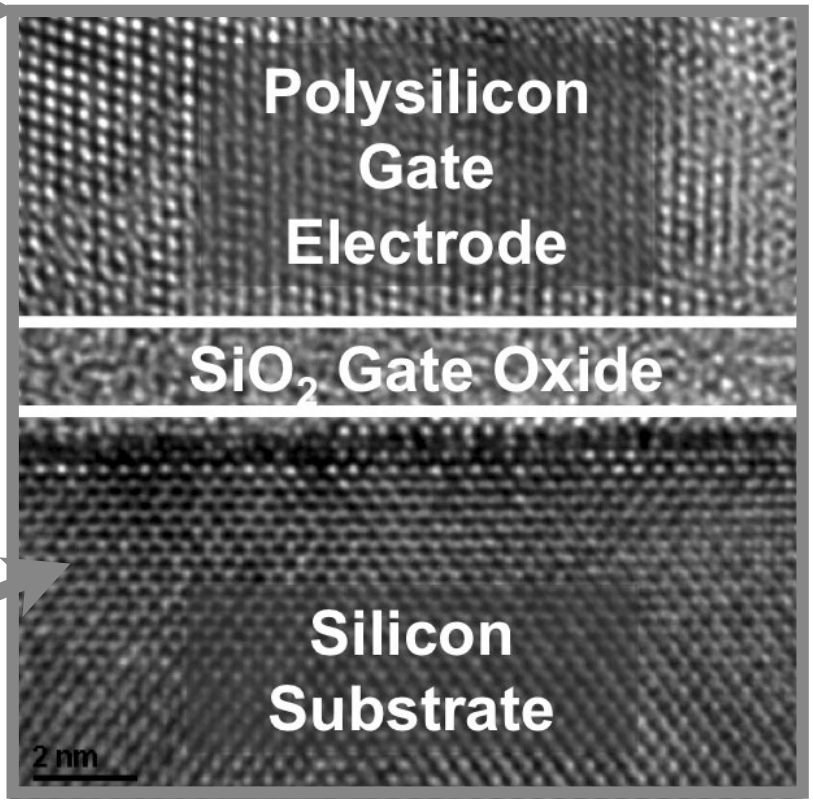
Much of it goes back to the transistor



Source: Intel press foils

Much of it goes back to the transistor

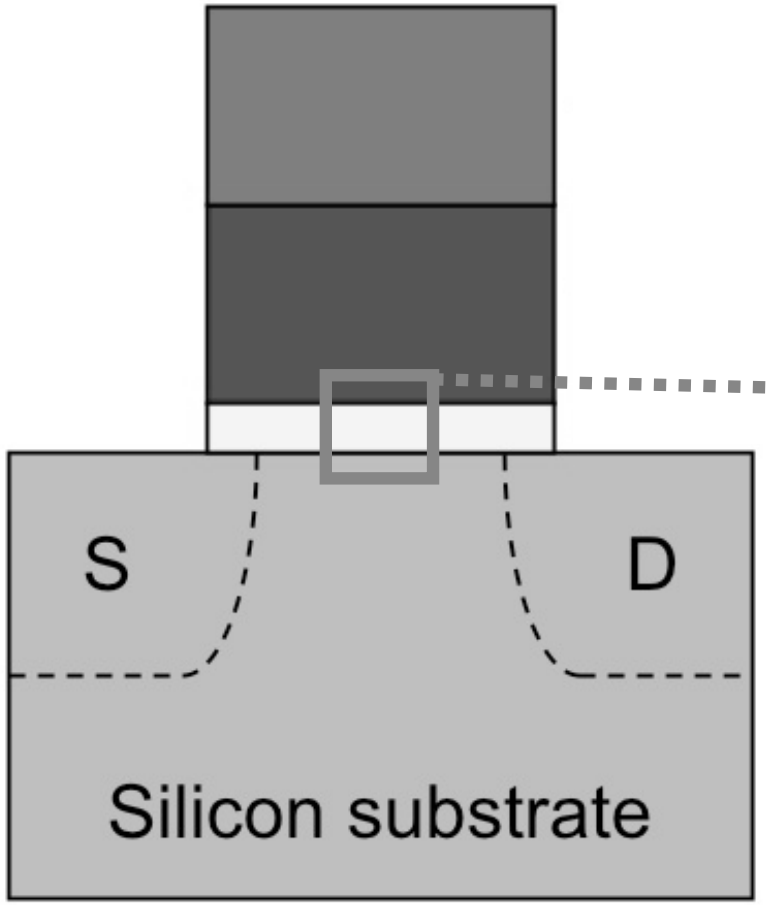


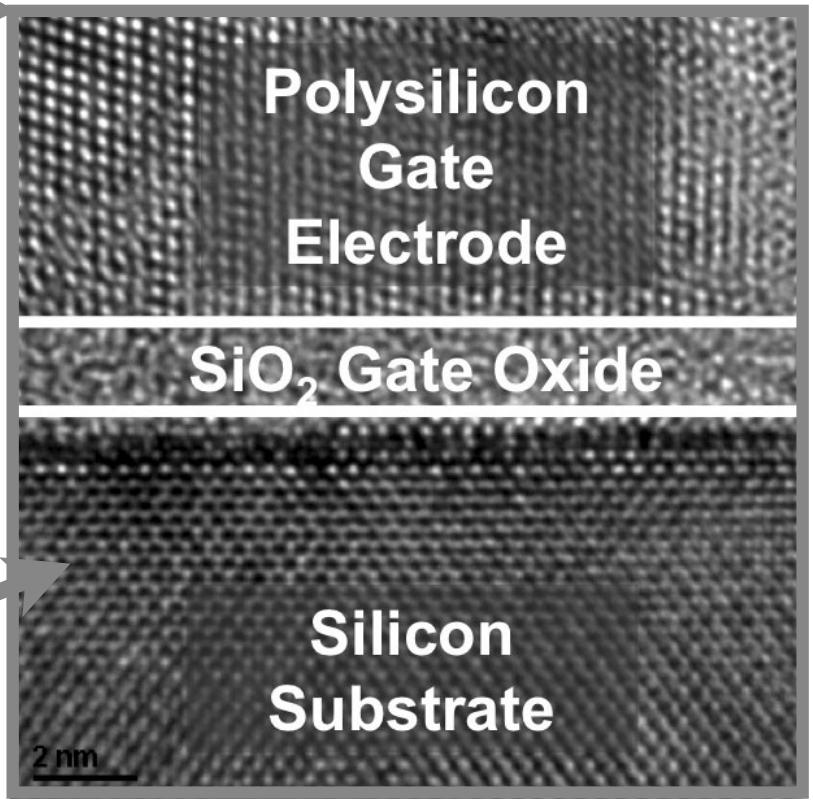


individual atoms!

Source: Intel press foils

Much of it goes back to the transistor



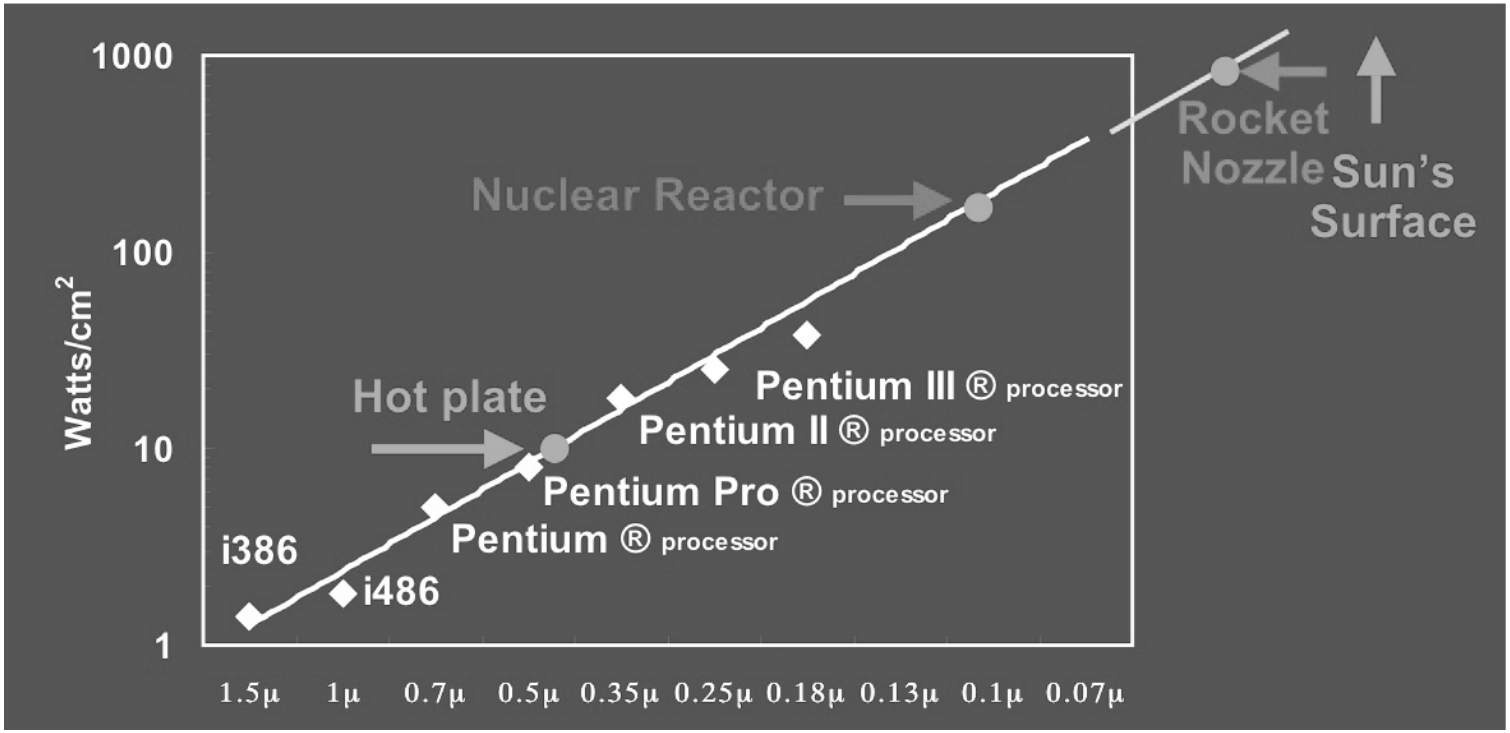


individual atoms!

= leakage current + defects

Source: Intel press foils

A model of power

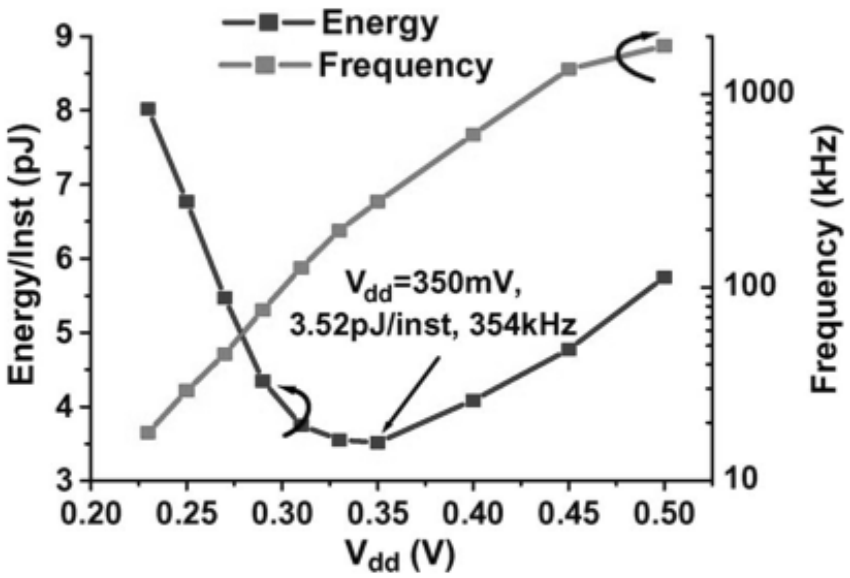


$$P = P_{switch} + P_{leakage}$$

$$P_{switch} = E_{switch} \times F = (C \times Vdd^2) \times F$$

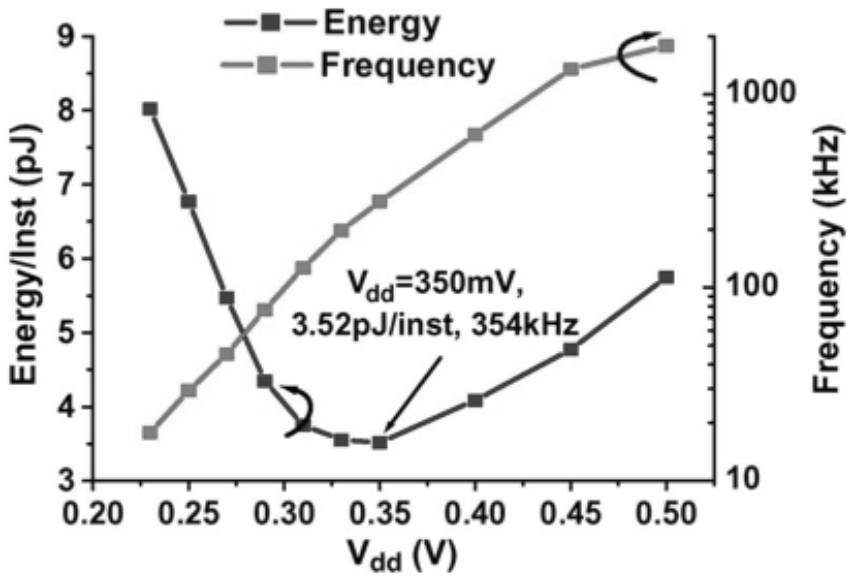
$$P_{leakage} = Vdd \times I$$

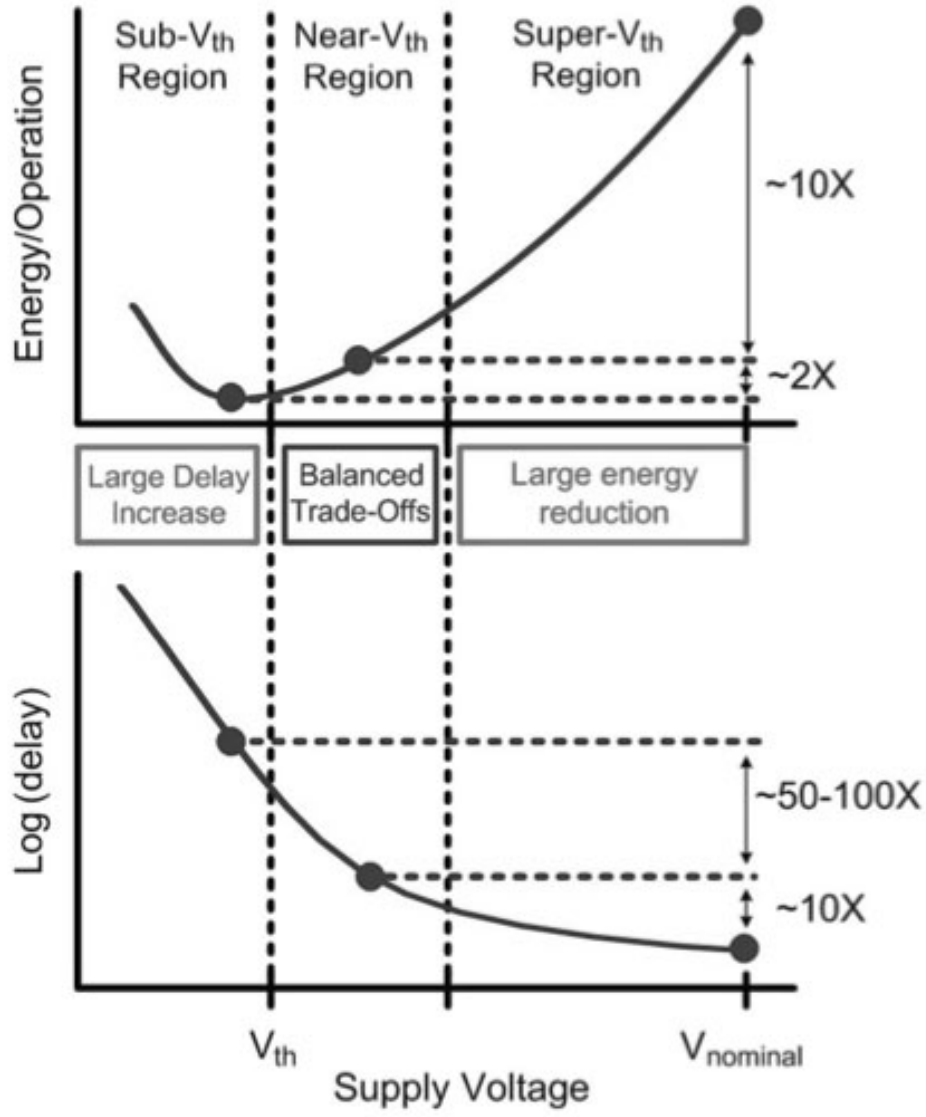
Voltage Scaling: DVFS + Near-Threshold Computing



[Source: Dreslinski et al.: Near-Threshold Computing: Recplaiming Moore's Law Through Energy Efficient Integrated Circuits]

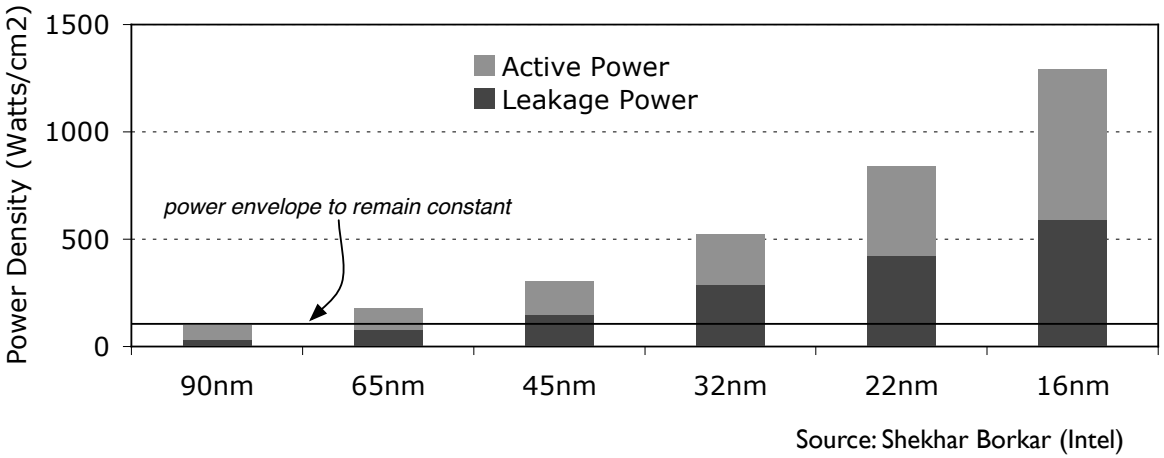
Voltage Scaling: DVFS + Near-Threshold Computing





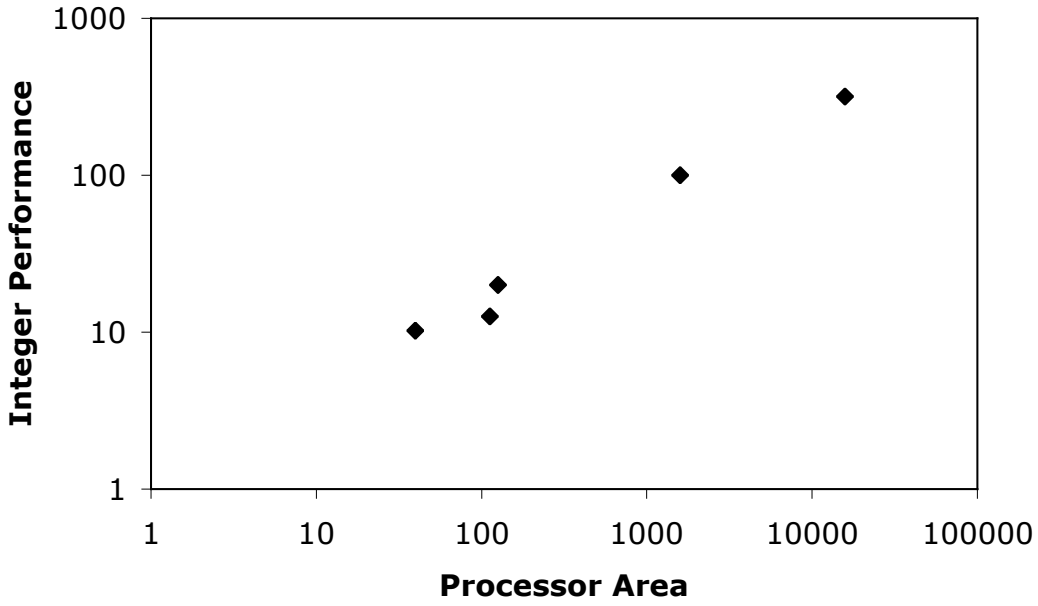
[Source: Dreslinski et al.: Near-Threshold Computing: Recplaiming Moore's Law Through Energy Efficient Integrated Circuits]

Chip Area and Power Consumption



Source: Shekhar Borkar (Intel)

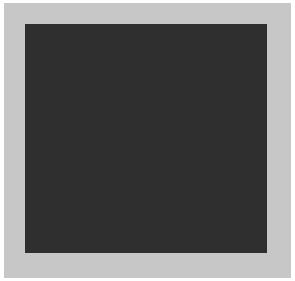
With leakage power dominating, power consumption roughly proportional to transistor count



Source: Shekhar Borkar (Intel)

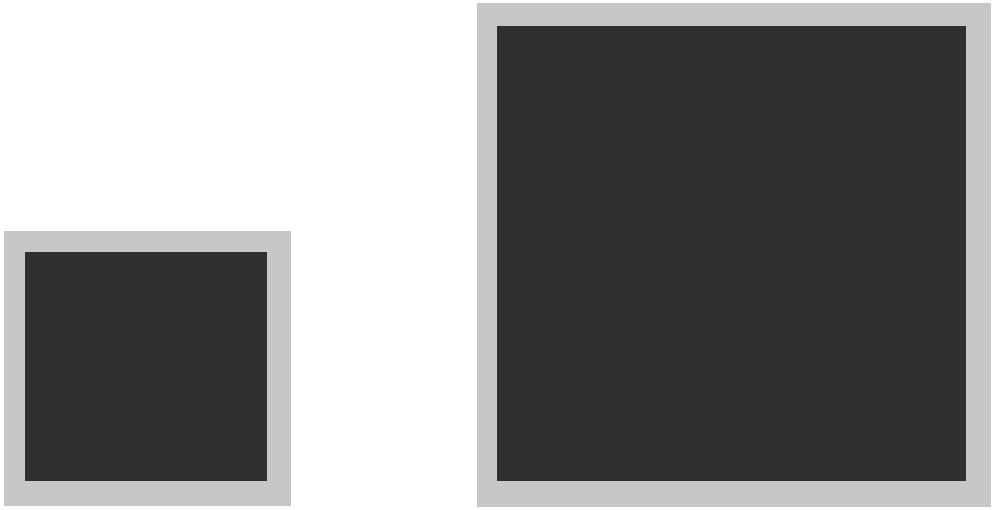
Pollack's Law: Processor performance grows with sqrt of area

The Resulting Shift to Multicore



Perf = 1
Power = 1

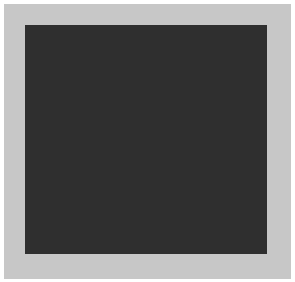
The Resulting Shift to Multicore



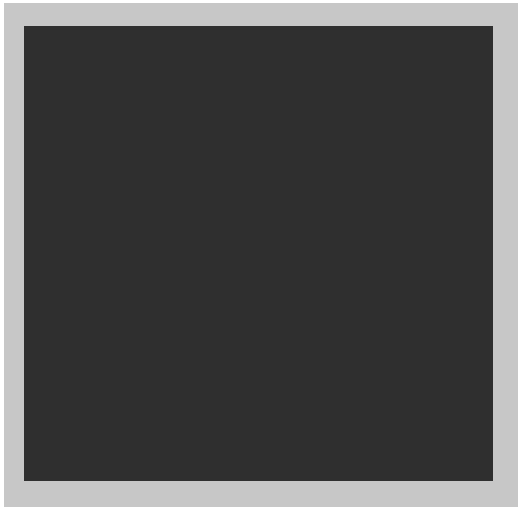
Perf = 1
Power = 1

Perf = 2
Power = 4

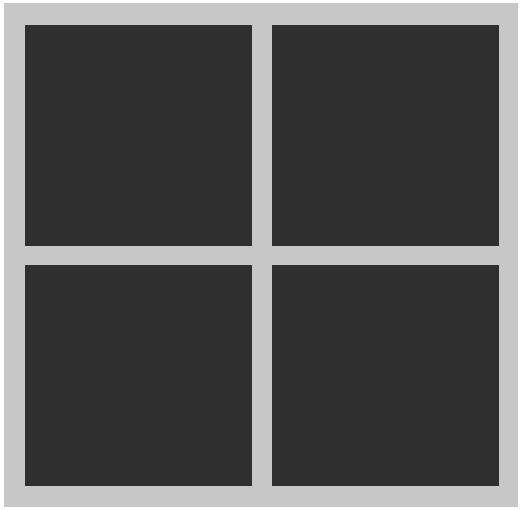
The Resulting Shift to Multicore



Perf = 1
Power = 1



Perf = 2
Power = 4



Perf = 4
Power = 4

Sea Change in Architecture: Multicore

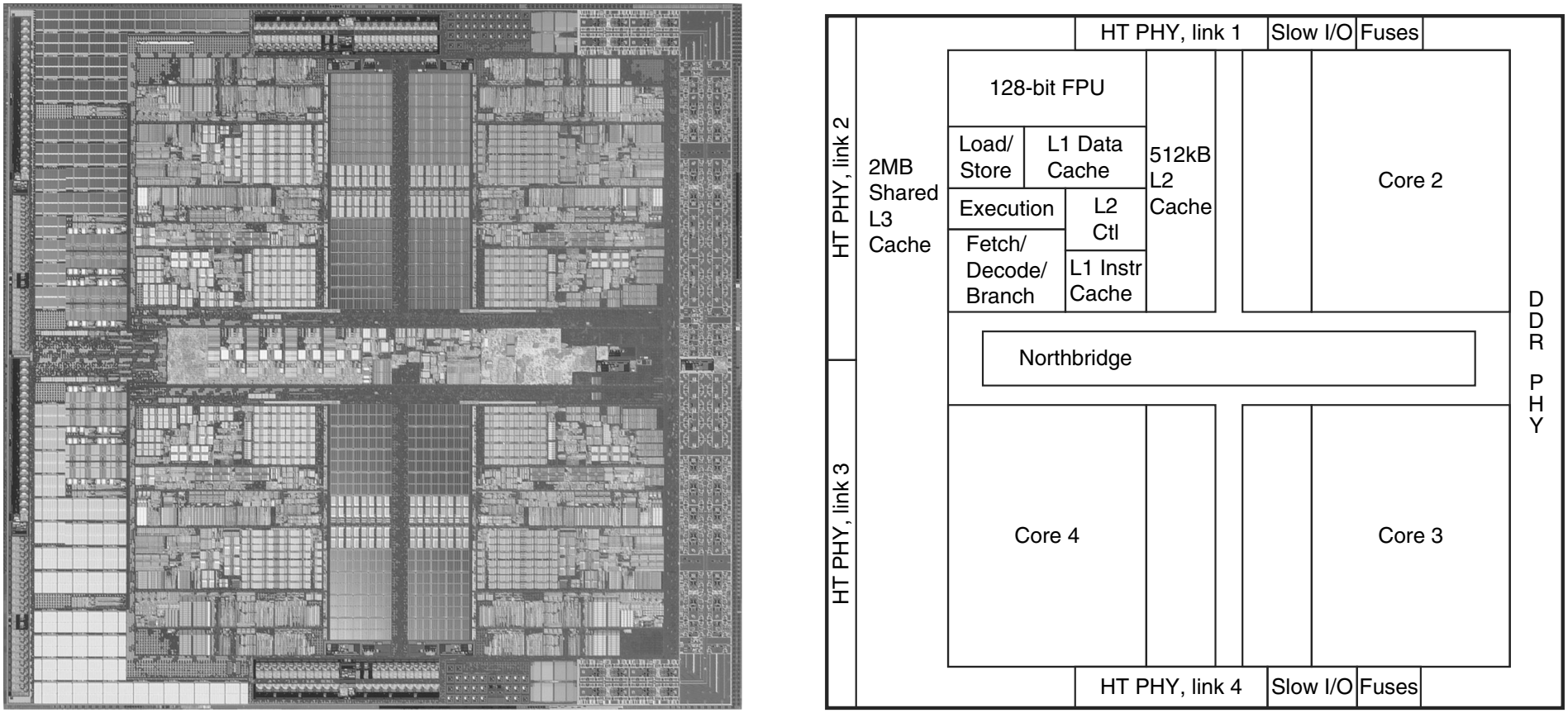


FIGURE 1.9 Inside the AMD Barcelona microprocessor. The left-hand side is a microphotograph of the AMD Barcelona processor chip, and the right-hand side shows the major blocks in the processor. This chip has four processors or "cores". The microprocessor in the laptop in Figure 1.7 has two cores per chip, called an Intel Core 2 Duo. Copyright © 2009 Elsevier, Inc. All rights reserved.

x86 64-bit Architecture Evolution

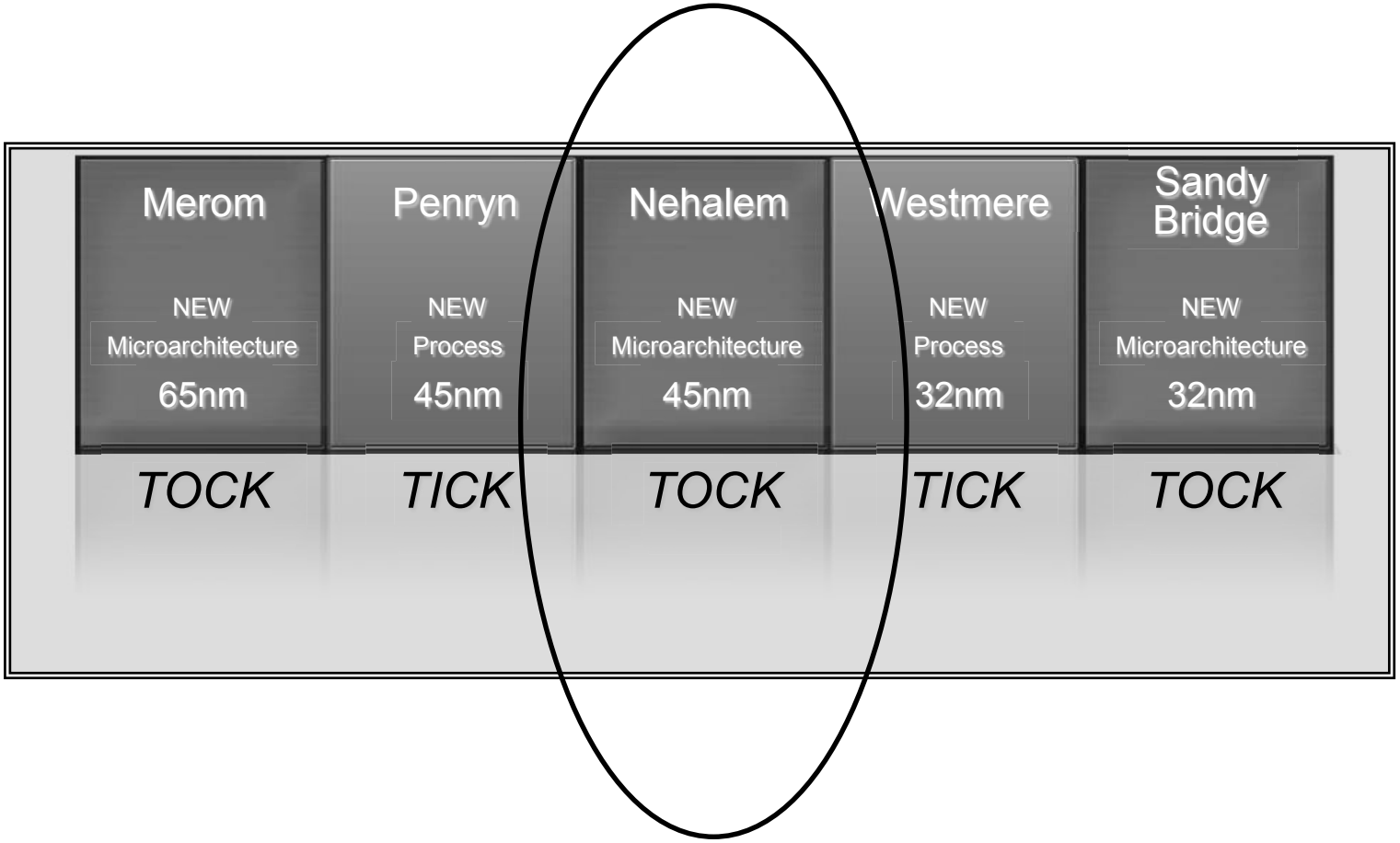
	2003	2005	2007	2008	2009	2010
	AMD Opteron™	**AMD Opteron™**	**"Barcelona"**	**"Shanghai"**	**"Istanbul"**	**"Magny-Cours"**
Mfg. Process	90nm SOI	90nm SOI	65nm SOI	45nm SOI	45nm SOI	45nm SOI
CPU Core	K8	K8	Greyhound	Greyhound+	Greyhound+	Greyhound+
L2/L3	1MB/0	1MB/0	512kB/2MB	512kB/6MB	512kB/6MB	512kB/12MB
Hyper Transport™ Technology	3x 1.6GT/.s	3x 1.6GT/.s	3x 2GT/s	3x 4.0GT/s	3x 4.8GT/s	4x 6.4GT/s
Memory	2x DDR1 300	2x DDR1 400	2x DDR2 667	2x DDR2 800	2x DDR2 1066	4x DDR3 1333

Max Power Budget Remains Consistent

AMD The future is fusion

[Source: HotChips '09]

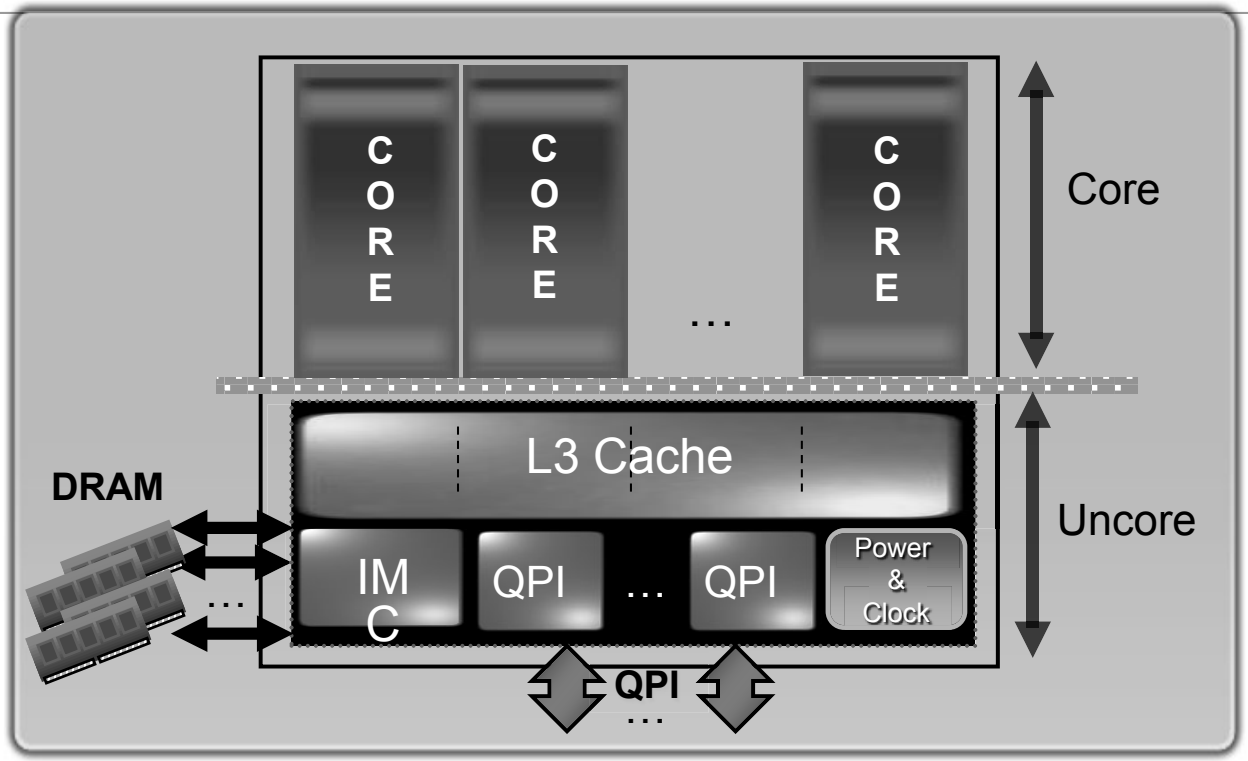
Tick-Tock Development Model



All products, dates, and figures are preliminary and are subject to change without notice.

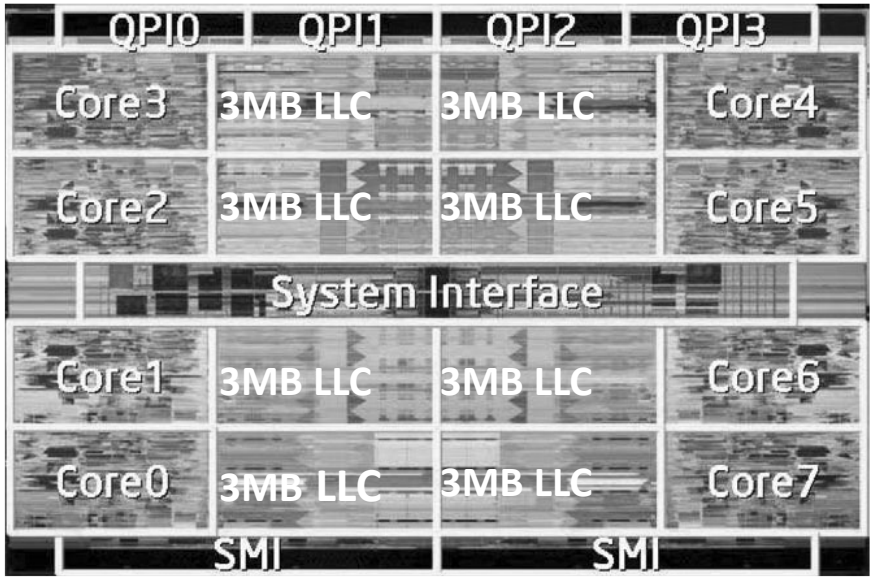
[Source: HotChips '09]

Nehalem Core/Uncore Modularity



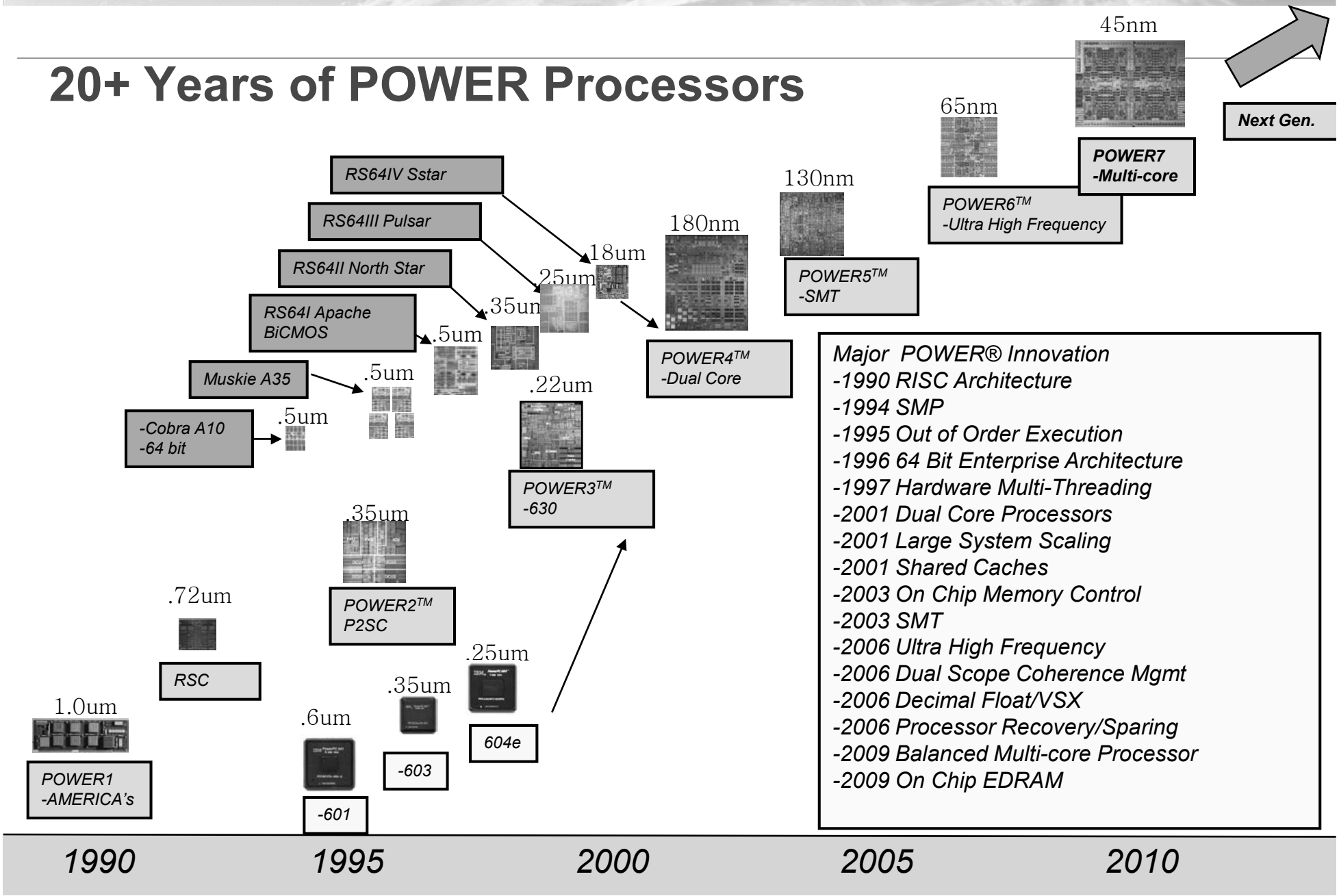
- Common core for client and server CPUs
  - http://www.intel.com/technology/architecture-silicon/next-gen/whitepaper.pdf
  - Some unique features only on NHM-EX
- Uncore differentiates different segment specific CPUs
  - Scalable Core/Uncore gasket interface
  - Decouples core and uncore operation

Nehalem-EX CPU



- Monolithic single die CPU
- 8 Nehalem cores, 16 threads
- 24MB shared L3 cache
- 2 integrated memory controllers
- Scalable Memory Interconnect (SMI) with support for up to 8 DDR channels
- 4 Quick Path Interconnect (QPI) links with up to 6.4GT/s
- Supports 2, 4 and 8 socket in glueless configs and larger systems using Node Controller (NC)
- Intel 45nm process technology
- 2.3 Billion transistors

20+ Years of POWER Processors

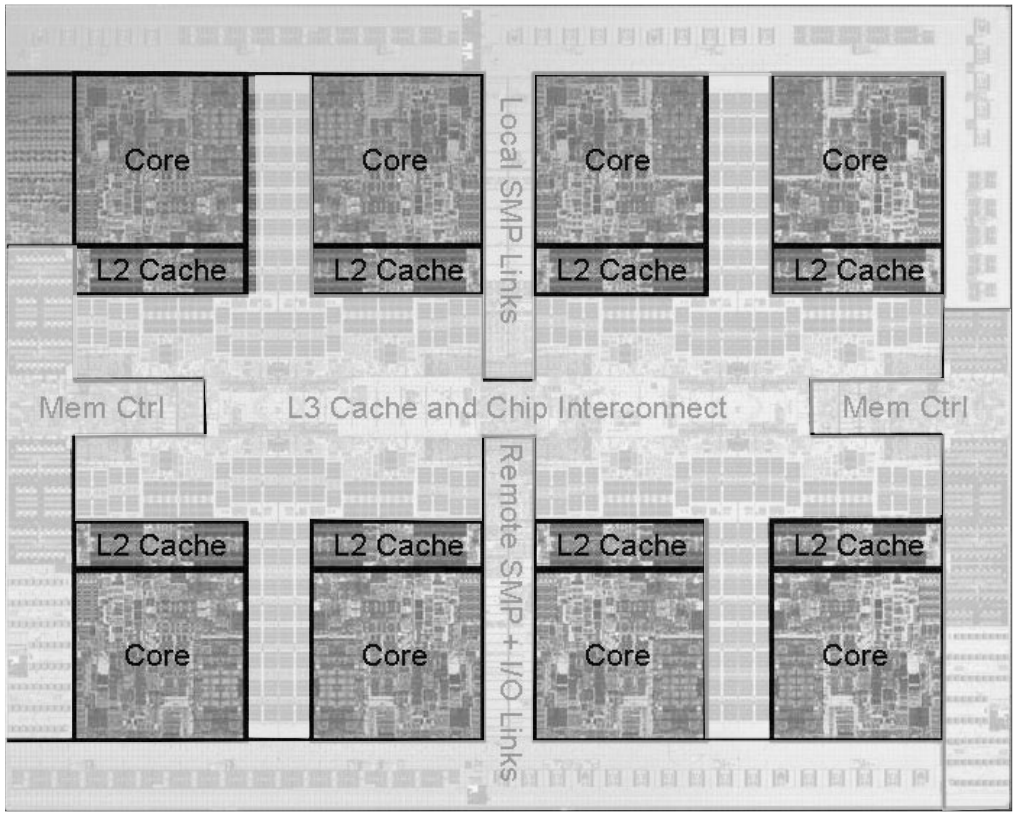


* Dates represent approximate processor power-on dates, not system availability

[Source: HotChips '09]

POWER7 Processor Chip

- 567mm^2 Technology: 45nm lithography, Cu, SOI, eDRAM
- 1.2B transistors
  - Equivalent function of 2.7B
  - eDRAM efficiency
- Eight processor cores
  - 12 execution units per core
  - 4 Way SMT per core
  - 32 Threads per chip
  - 256KB L2 per core
- 32MB on chip eDRAM shared L3
- Dual DDR3 Memory Controllers
  - 100GB/s Memory bandwidth per chip sustained
- Scalability up to 32 Sockets
  - 360GB/s SMP bandwidth/chip
  - 20,000 coherent operations in flight
- Advanced pre-fetching Data and Instruction
- Binary Compatibility with POWER6



* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.

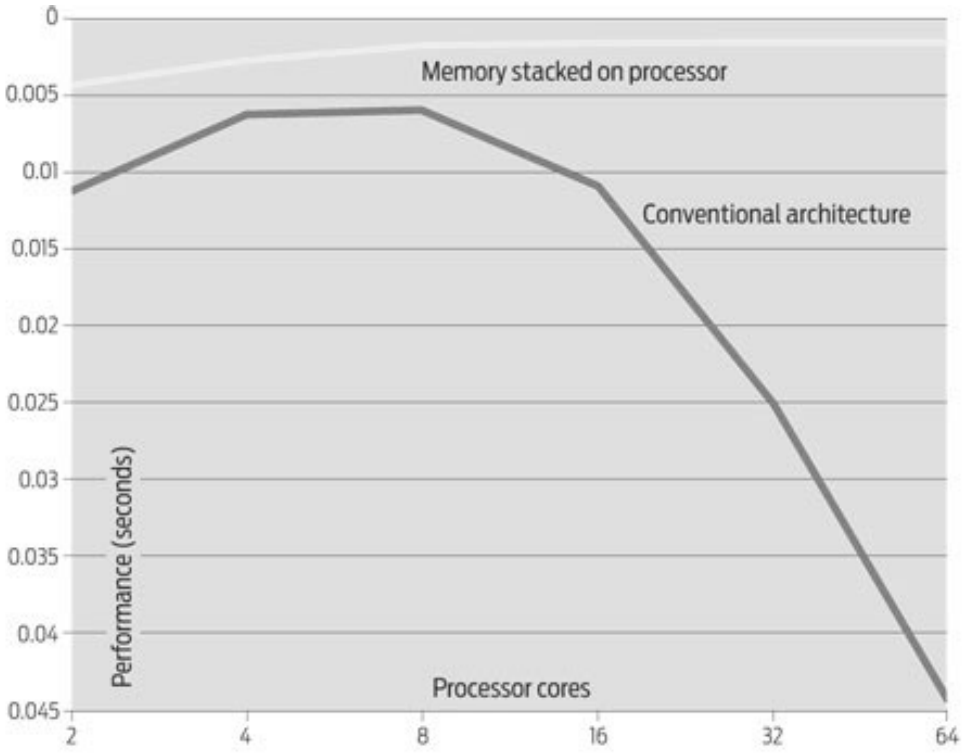
[Source: HotChips '09]

Challenges in Multicore

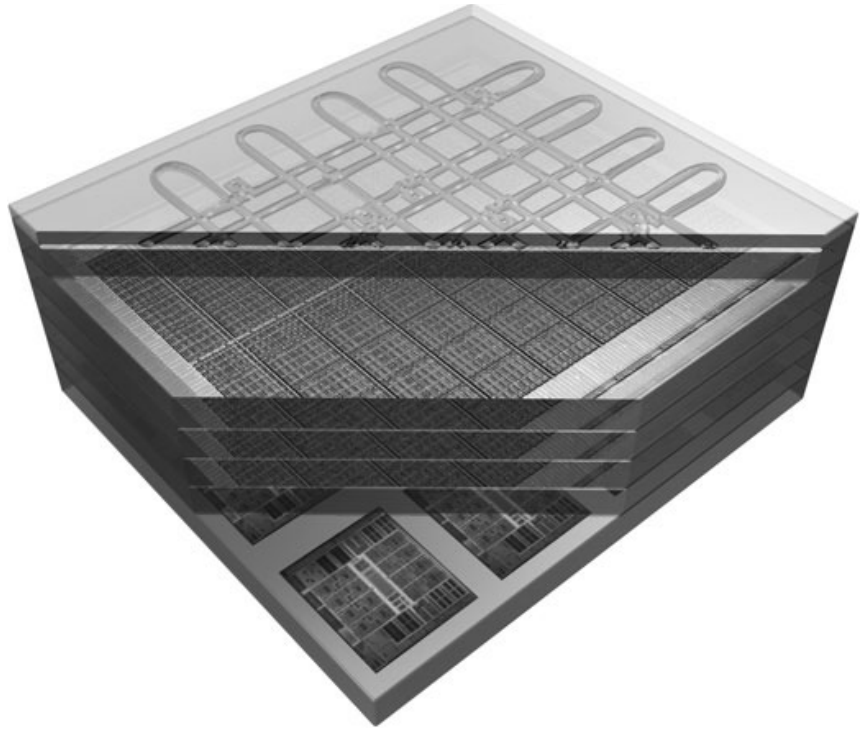
1. Performance dependent on parallel codes

2. Memory bandwidth ("feeding the beast")

3. Communication and coherence

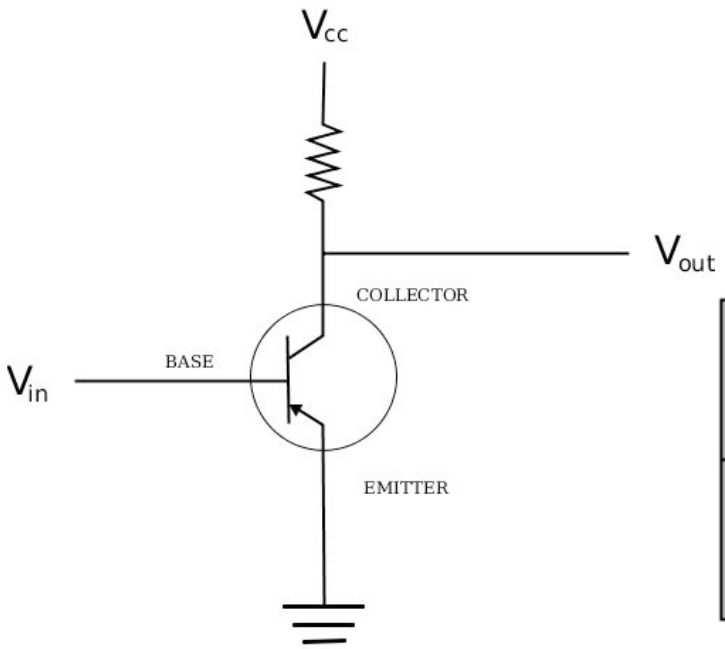


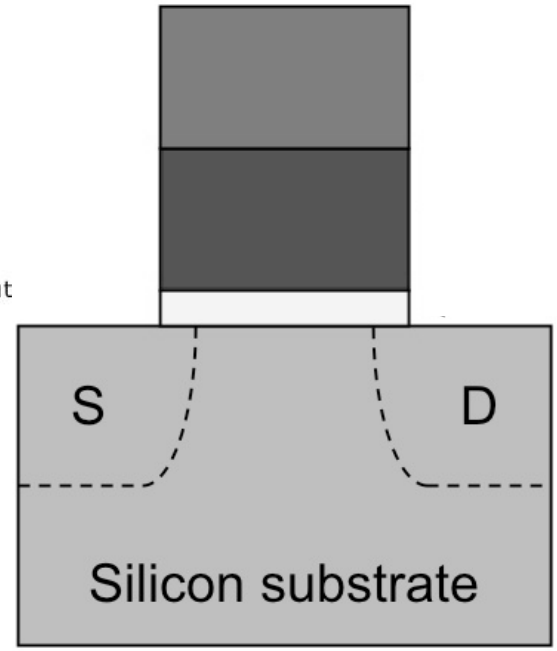
[Source: Sandia NL]

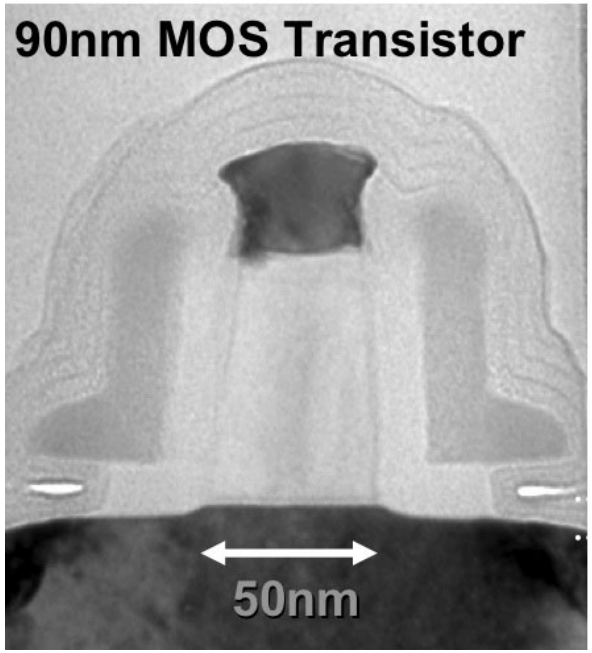


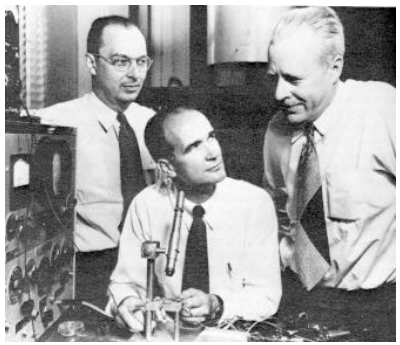
Computer architecture: the long view

The Transistor



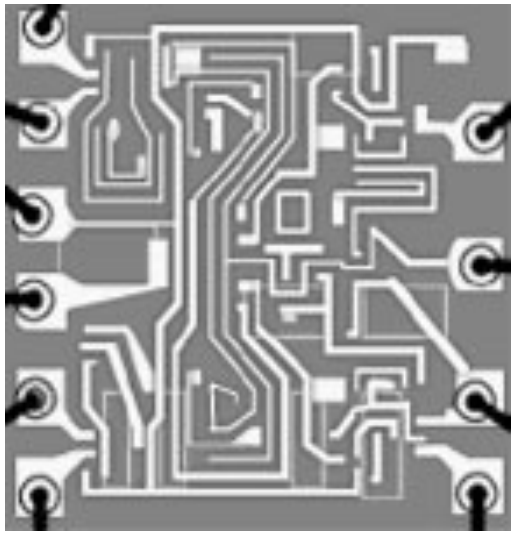


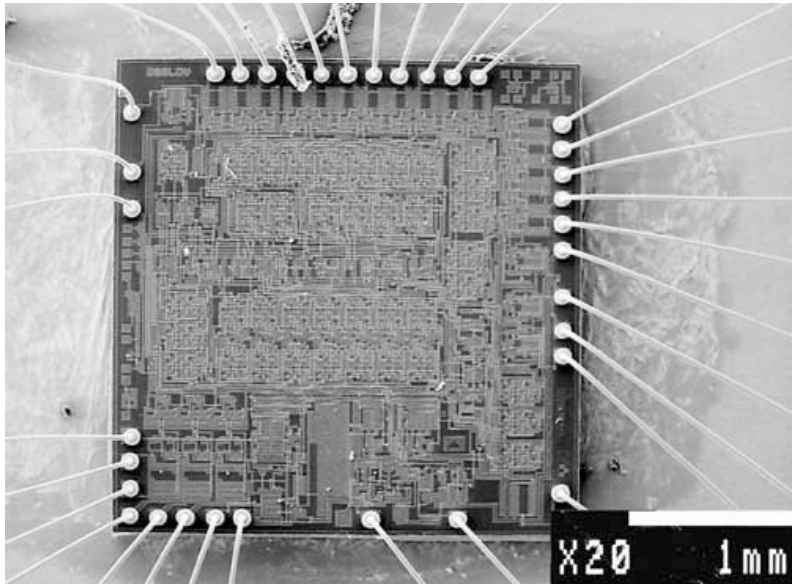




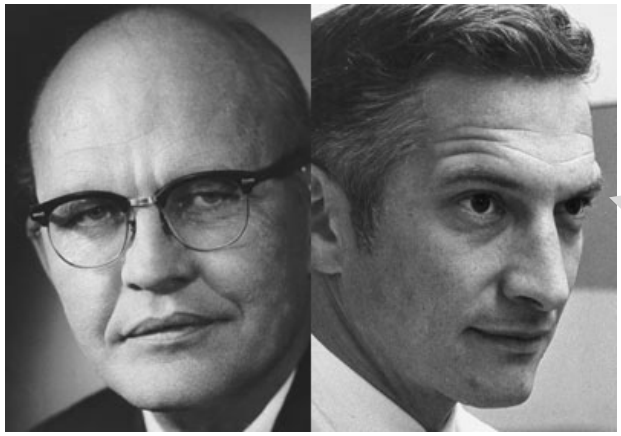
John Bardeen, Walter Brattain, William Shockley
Nobel Prize in Physics in 1965

The Integrated Circuit









Jack Kilby at Texas Instruments and Robert Noyce at Fairchild Camera

2000 Nobel Prize in Physics

Technology:

- Logic gates
- SRAM
- DRAM
- Circuit technologies
- Packaging
- Magnetic storage
- Flash memory
- Biochips
- 3D stacking

[Credit: Milo Martin, UPenn]

Computing Devices



















Technology:

Logic gates
SRAM
DRAM
Circuit technologies
Packaging
Magnetic storage
Flash memory
Biochips
3D stacking

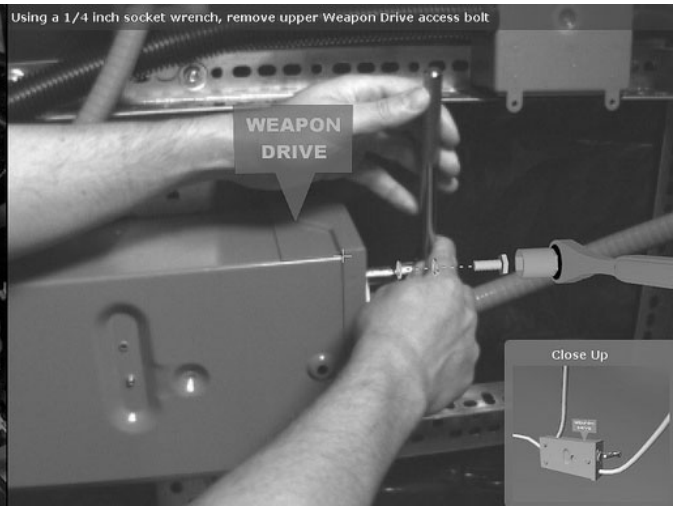
Application domains:

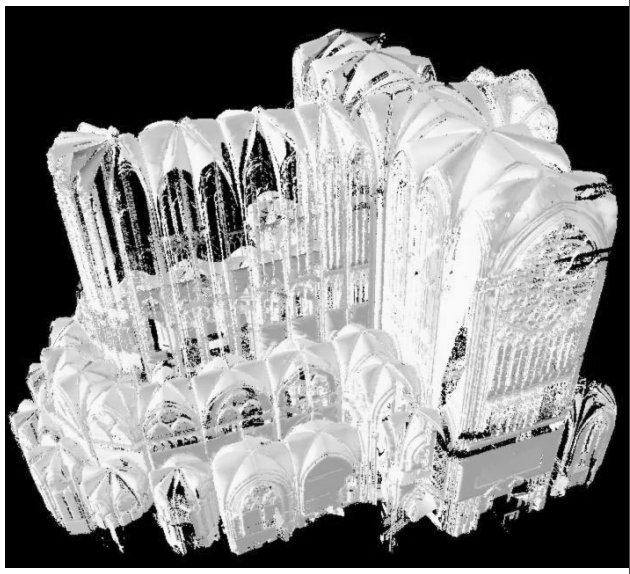
PCs
Servers
PDAs
Mobile phones
Supercomputers
Game consoles
Embedded

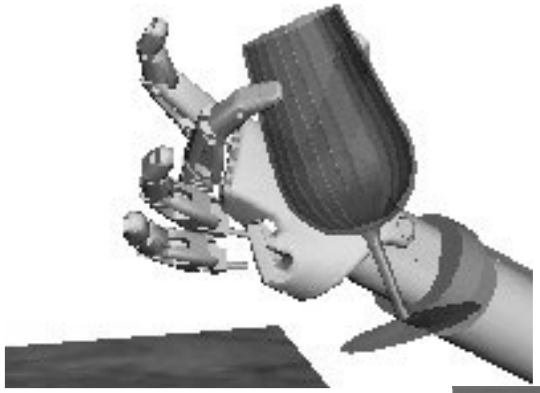
[Credit: Milo Martin, UPenn]

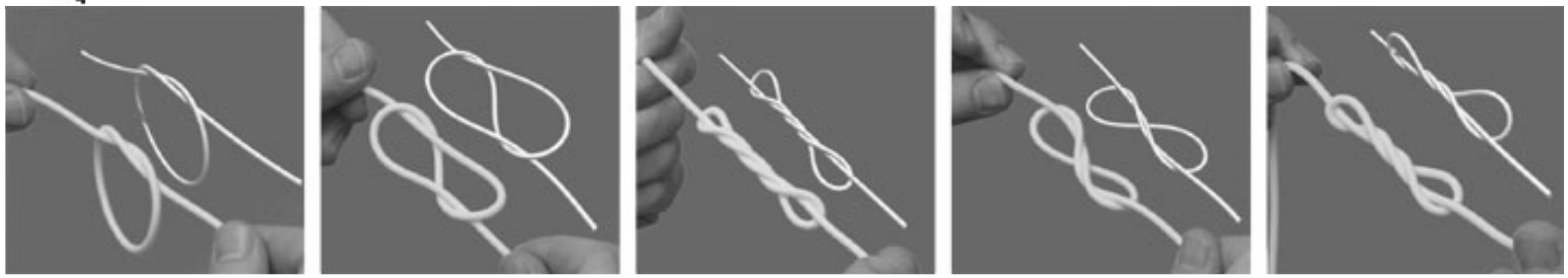
Computing Applications











Technology:

Logic gates
SRAM
DRAM
Circuit technologies
Packaging
Magnetic storage
Flash memory
Biochips
3D stacking

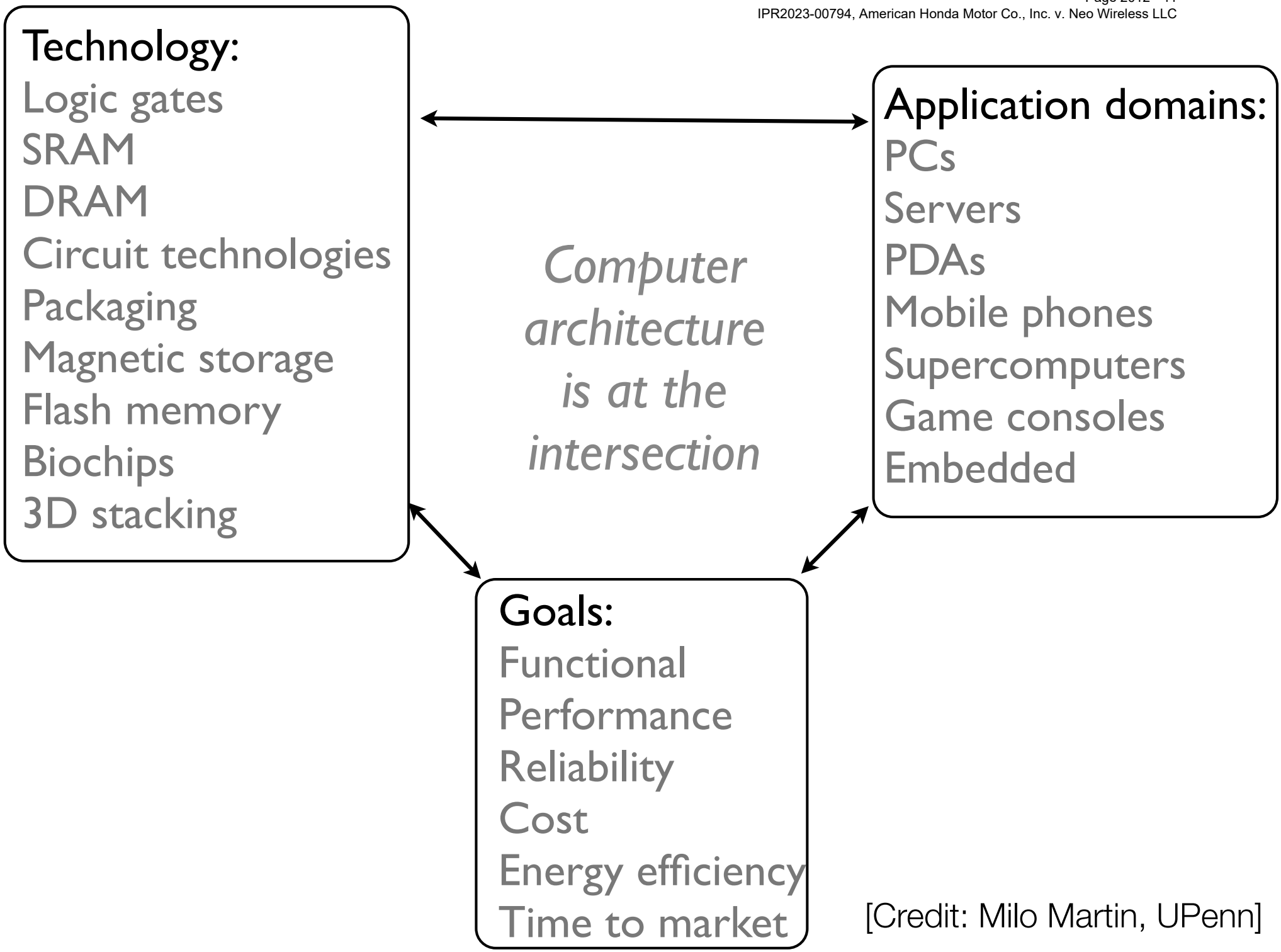
Application domains:

PCs
Servers
PDAs
Mobile phones
Supercomputers
Game consoles
Embedded

Goals:

Functional
Performance
Reliability
Cost
Energy efficiency
Time to market

[Credit: Milo Martin, UPenn]



[Credit: Milo Martin, UPenn]