Multi-Core Microprocessor Chips: Motivation & Challenges

Dileep Bhandarkar, Ph. D.
Architect at Large
Digital Enterprise Group
Intel Corporation

May 2006

Agenda

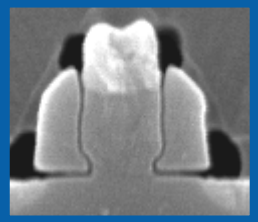
- Semiconductor Technology Evolution
- Design Challenges
- Why Multi-Core Processor Chips?
- Power/Performance Trade-Offs
- CMP Directions
- Beyond CMP
- Summary

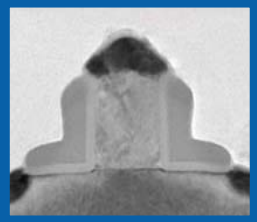
Intel only: On-time "2-year-cycle"

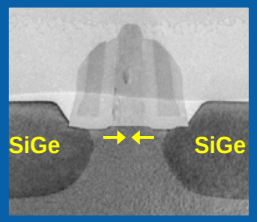
	180nm	130nm	90nm	65nm	45nm
Wafer Size (mm):	200	200/300	300	300	300
1st Production:	1999	2002	2003	2005	2007

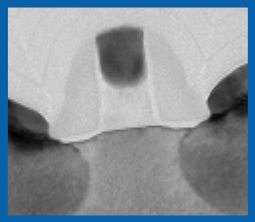
Transistors:

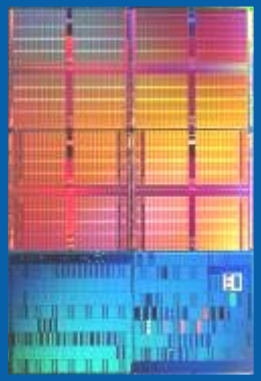
Interconnects:

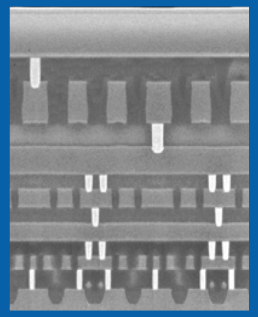


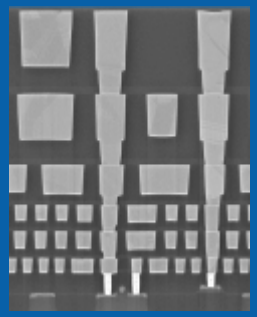


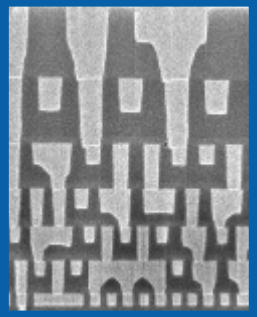


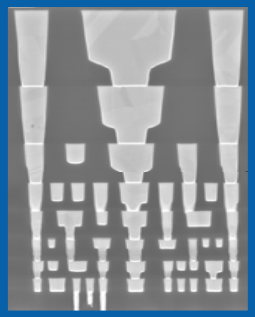












100nm L_G	70nm L_G	50nm L_G	35nm L_G	Details Coming!
$CoSi_2$	$CoSi_2$	NiSi	NiSi	
		Strain Si	Strain Si	
6 Al	6 Cu	7 Cu	8 Cu	
SiOF	SiOF	Low-k	Low-k	

45 nm Logic Process on Track for Delivery in 2007

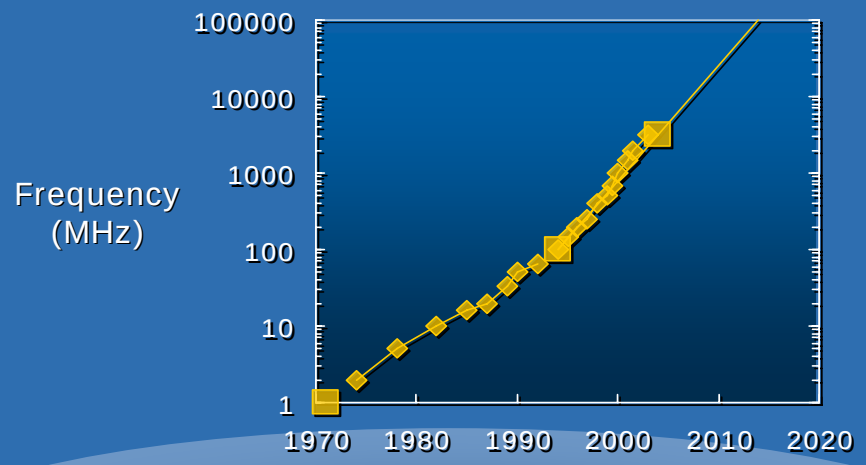
Process Name	P1262	P1264	P1266	P1268
Lithography	90 nm	65 nm	45 nm	32 nm
1st Production	2003	2005	2007	2009

Moore's Law continues!

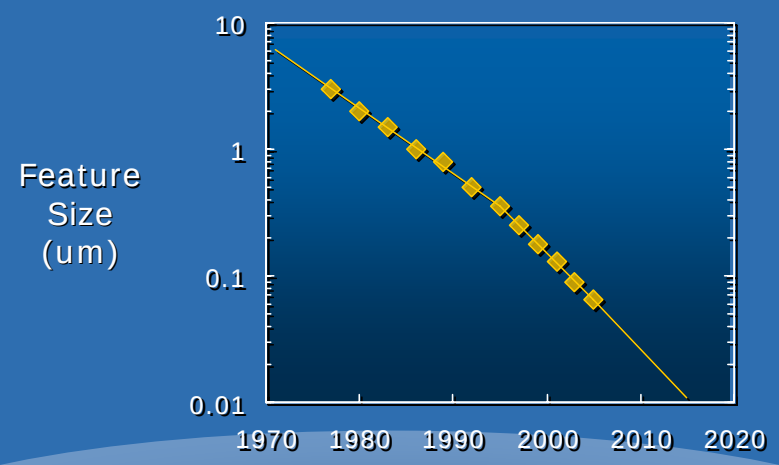
Intel continues to develop a new technology generation every 2 years

Historical Driving Forces

Increased Performance via Increased Frequency

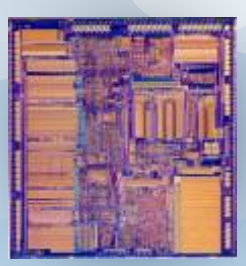


Shrinking Geometry



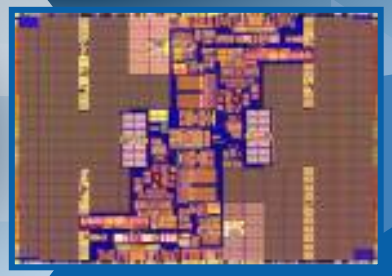
1971
4004 Processor
2300 Transistors

1978
8008 Processor
IBM PC



1985
i386 Processor
32-bit

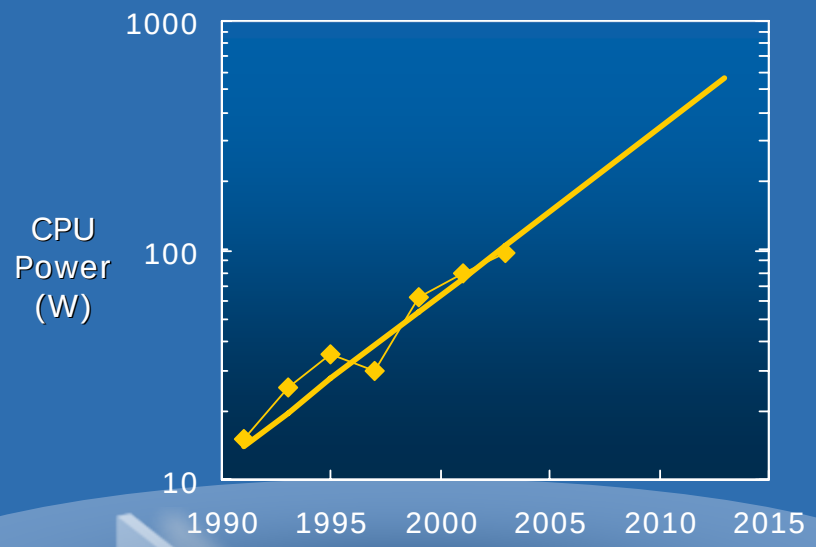
1993
Pentium Processor
3.1M transistors



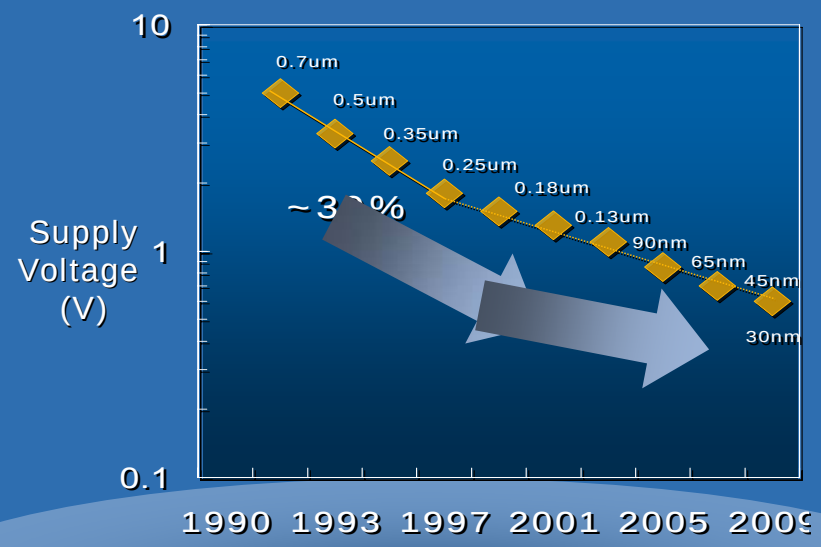
2005
Montecito
1.7B Transistors

The Challenges

Power Limitations



Diminishing Voltage Scaling



Power = Capacitance x Voltage^2 x Frequency
also
Power ~ Voltage^3

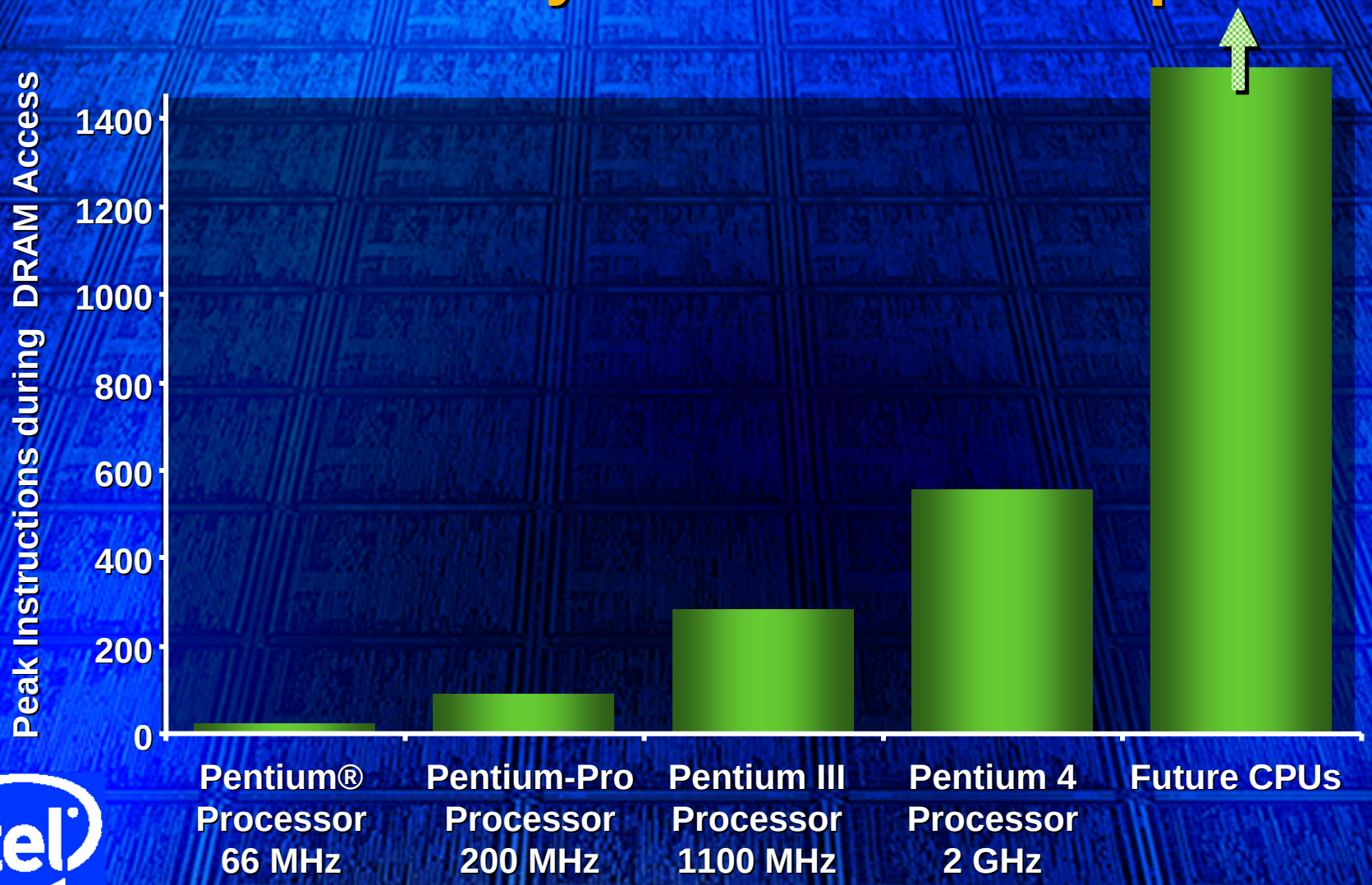
Agenda

- Semiconductor Technology Evolution
- Design Challenges
- Why Multi-Core Processor Chips?
- Power/Performance Trade-Offs
- CMP Directions
- Beyond CMP
- Summary

Design Challenges

- Memory latency not scaling as fast as processor speed
- Power growing non-linearly with single thread performance
- Designer productivity lagging design complexity
- Ability to validate and test complex design
- Keeping up with new process technology every two years

Long Latency DRAM Accesses: Needs Latency Tolerant Techniques



DRAM Latency Tolerance

- Continue building even larger caches
  - Every semiconductor process generation provides opportunity to double cache size
  - Cache becomes larger part of die
- Hide multiple threads of execution behind memory latency
- Intel implemented simultaneous multi-threading in 2000
- Implement multi-core products as Moore's Law allows

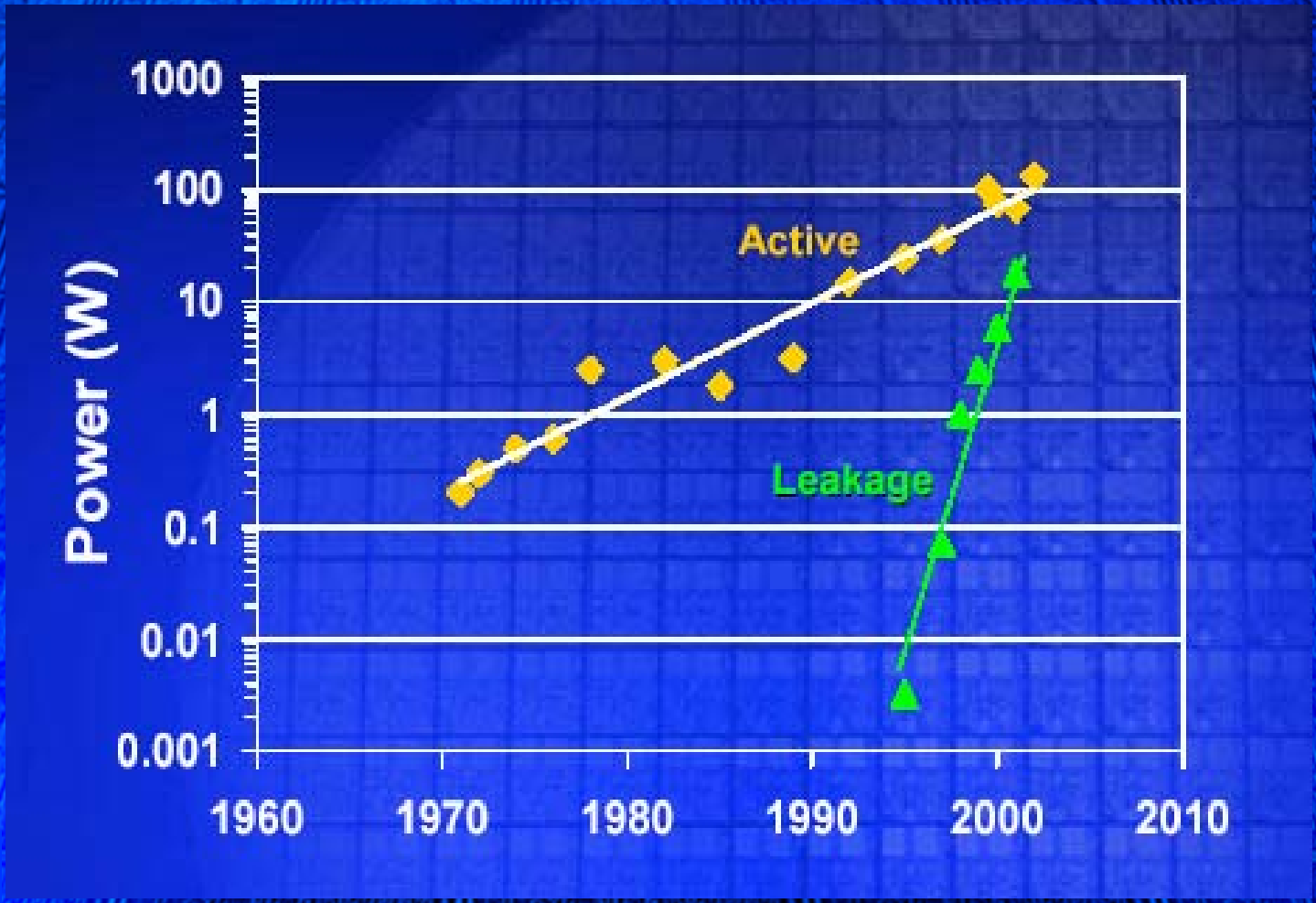
Agenda

- Semiconductor Technology Evolution
- Design Challenges
- Why Multi-Core Processor Chips?
- Power/Performance Trade-Offs
- CMP Directions
- Beyond CMP
- Summary

Situational Analysis

- With Each Process Generation transistor density doubles
  - Frequency has increased by ~1.5X; ~1.3x in future
  - Vcc has scaled by about ~0.8x; ~0.9x in future
  - Capacitance has scaled by 0.7x
  - Total power may not scale down due to increased leakage
- Instruction Level Parallelism harder to find
- Increasing single-stream performance often requires non-linear increase in design complexity
- Many server applications are inherently parallel
- Parallelism exists in multimedia applications
- Multi-tasking usage models becoming popular

Processor Power



Design Complexity and Productivity factors

- Huge transistor budgets stress ability to design and verify complex chips
- Multi-core fits well with increasing transistor budgets
- Multi-core design addresses density/designer gap

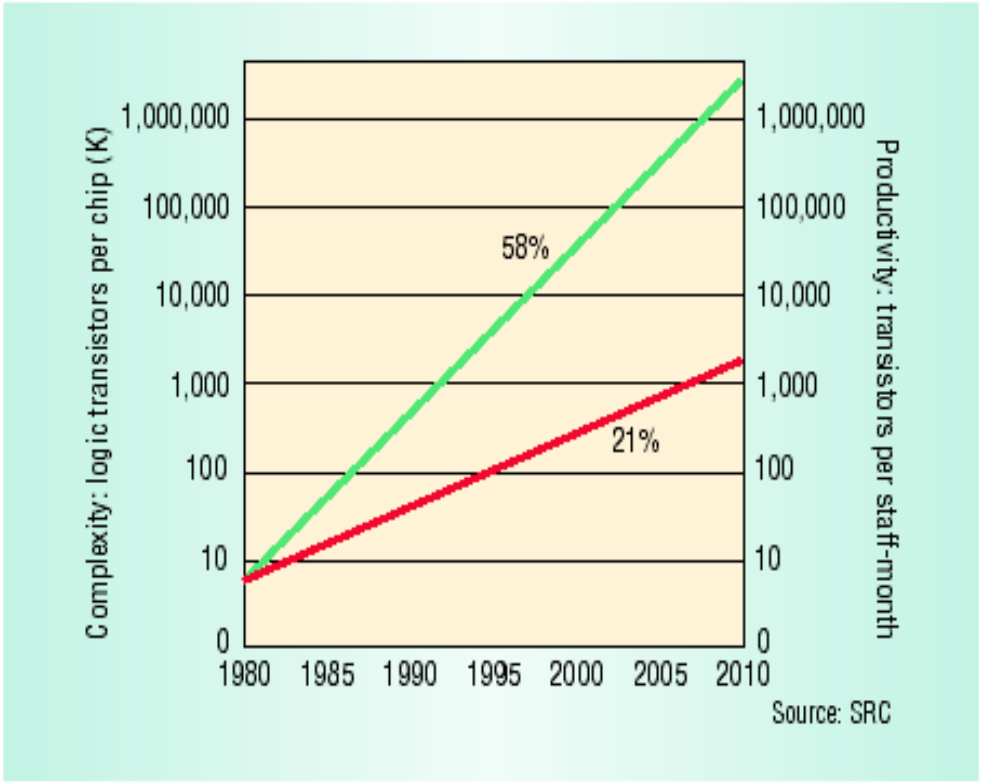


Figure 2. Design complexity and designer productivity. Since 1980, the design gap between growth in chip complexity and productivity growth in logic design tools has widened each year.

Agenda

- Semiconductor Technology Evolution
- Design Challenges
- Why Multi-Core Processor Chips?
- Power/Performance Trade-Offs
- CMP Directions
- Beyond CMP
- Summary

Iron Law of Performance

- Execution Time is the product of
  - Path Length
  - Cycles Per Instruction (CPI)
  - Cycle Time
- CPI is the sum of
  - infinite-cache core cpi
  - miss rate * effective memory latency
- Bad (good) news is that performance does not scale up (down) linearly with frequency

The Magic of Voltage Scaling

- Power = Capacitance * Voltage^2 * Frequency
- Frequency α Voltage in region of interest
- Power increases as the cube of Frequency
- Good news is that voltage scaling works
- 10% reduction in voltage yields
  - 10% reduction in frequency
  - 30% reduction in power
  - less than 10% reduction in performance

Simple Dual Core Example

- Assume Single Core processor at 100W
  - 80W for core, 20W for cache and I/O
  - 50% die are is core
- Dual core within same power envelop
  - 20W for I/O and cache
  - 40W per core
  - Die size increases by 50%
  - Reduce voltage by 21% to reduce core power to 40W
  - Frequency reduces by ~20%
  - Single thread perf reduces by ~15%
  - Throughput increases by 70-80%

Possible Improvements

- Develop new power efficient core
  - E.g. extensive clock gating
  - Big power savings with little or no performance loss
- Design a smaller core with lower performance
  - Area and power savings much greater than performance loss
  - Use larger number of cores
- Adjust frequency and power of each core with load factor
  - Inactive cores can be put in sleep mode
  - Maintain overall die power constant

A New Era...

THE OLD

Performance
Equals Frequency

Unconstrained Power

Voltage Scaling

THE NEW

Performance
Equals IPC

Multi-Core

Power Efficiency

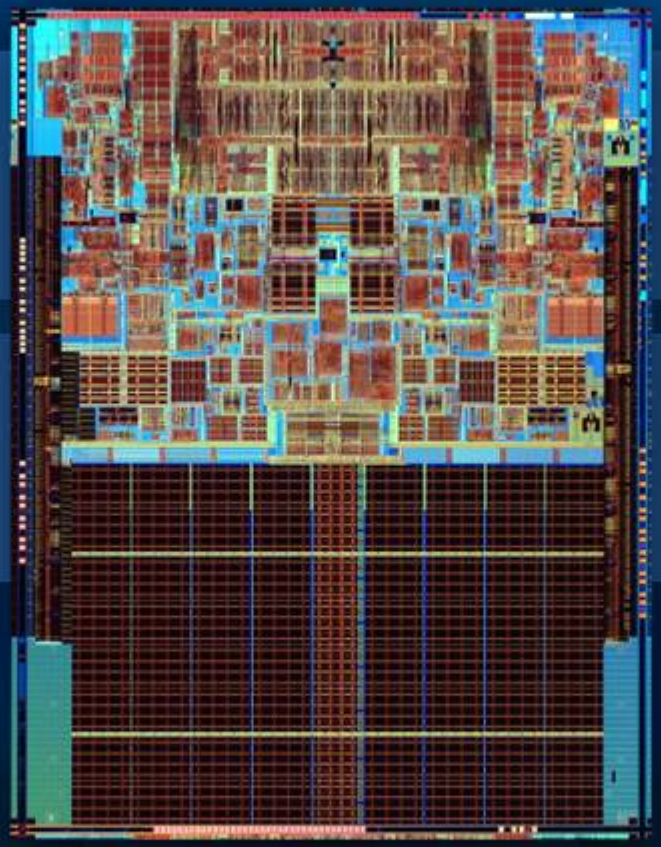
Microarchitecture
Advancements

intel

Intel Core Micro-architecture
Five Key Innovations

Intel® Wide Dynamic Execution

Intel® Intelligent Power Capability

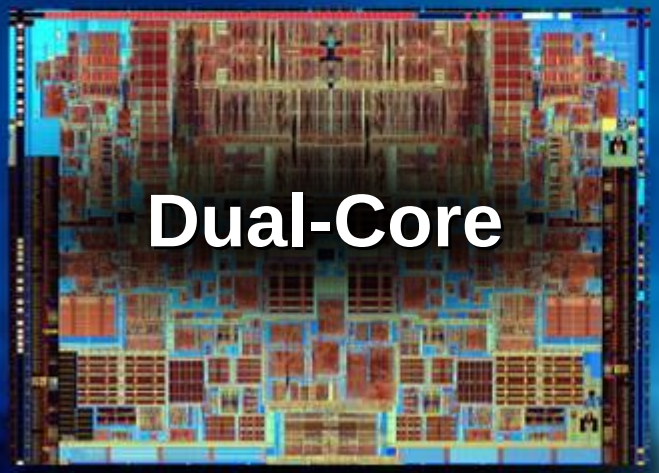


Intel® Advanced Digital Media Boost

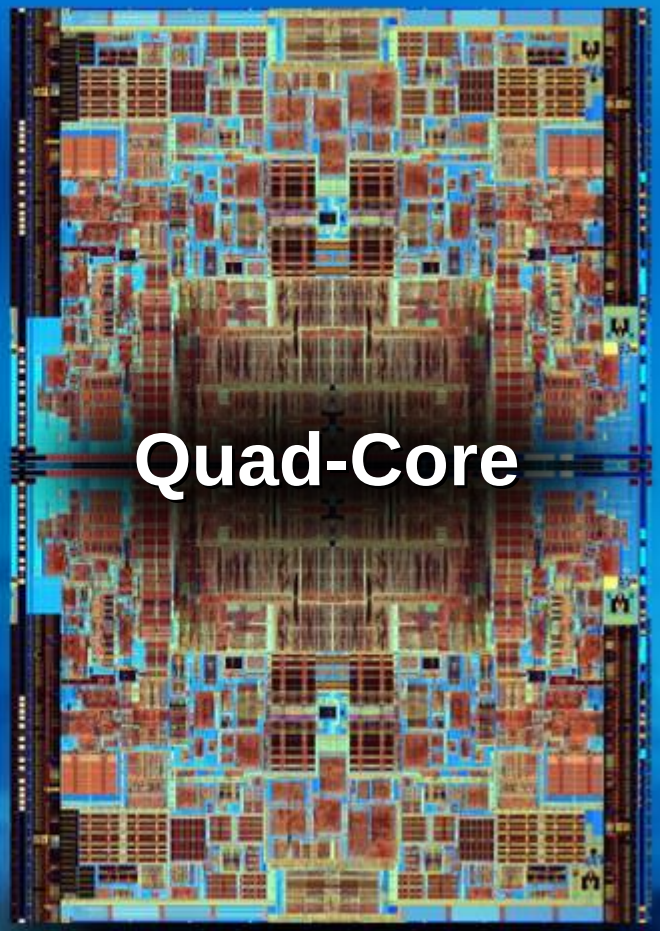
Intel® Smart Memory Access

Intel® Advanced Smart Cache

Multi-Core Trajectory

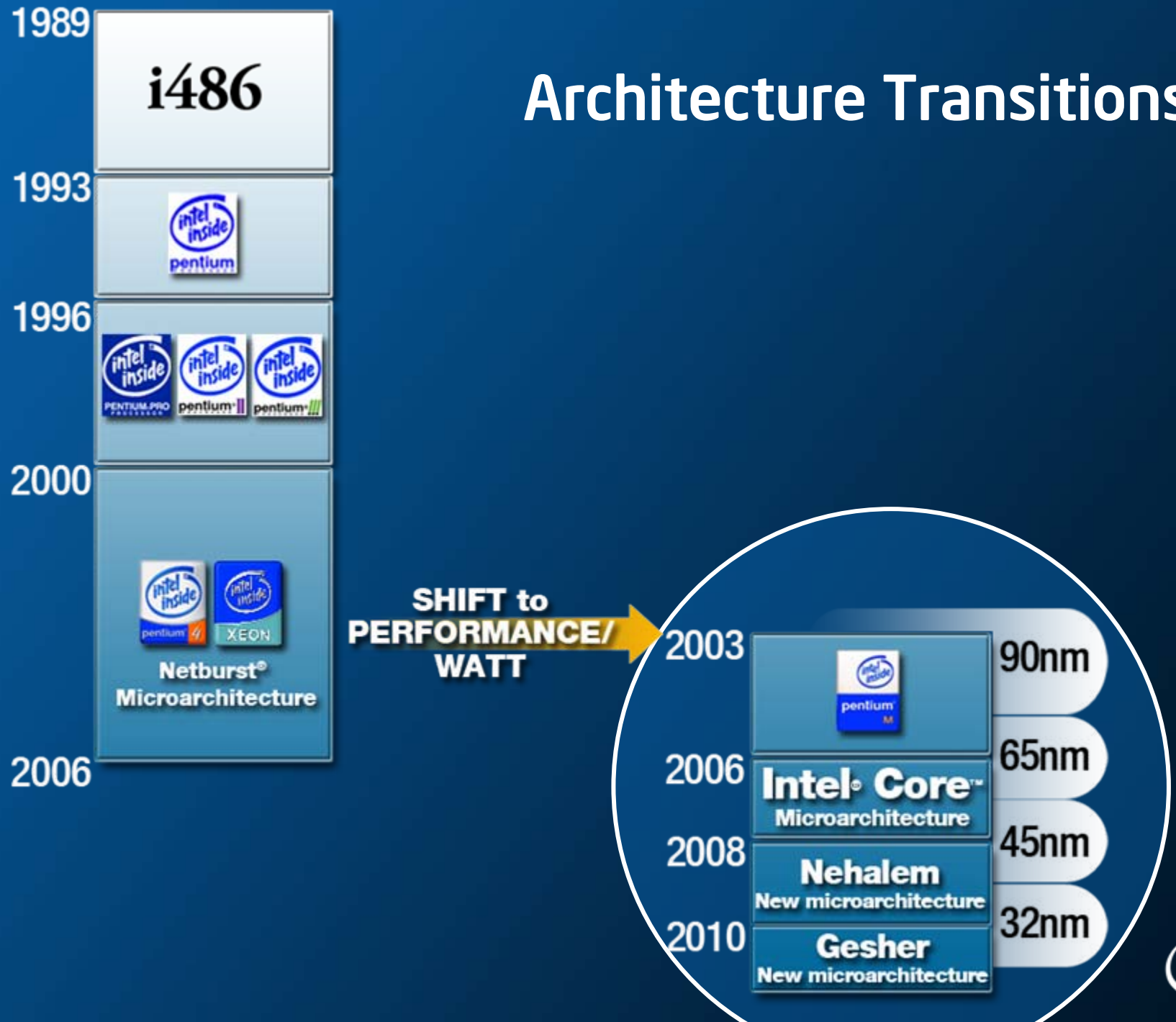


2H 2006

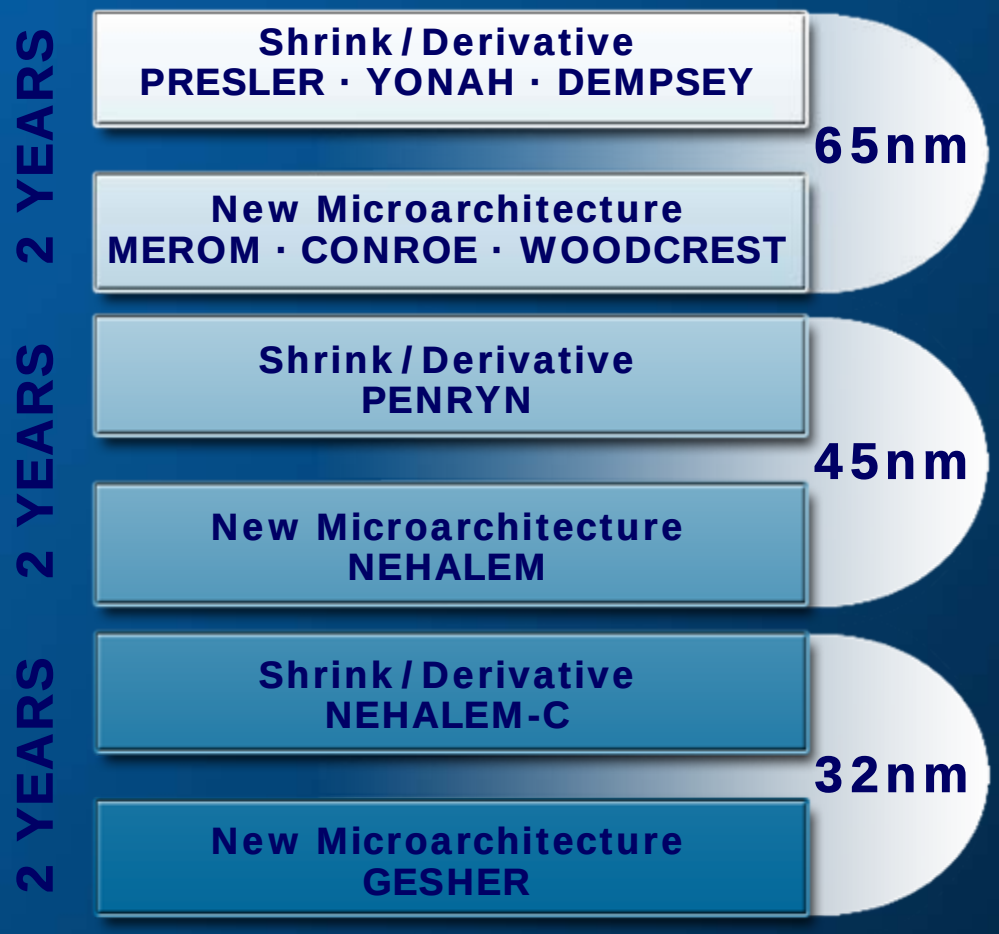


1H 2007

Architecture Transitions



Microprocessor Design Model



PRINCIPLES

1. One micro-architecture for all high volume market segments
2. Optimized for performance/watt
3. Parallel design teams
4. No waiting on new process technology
5. Chipset cadence offset for fast ramp

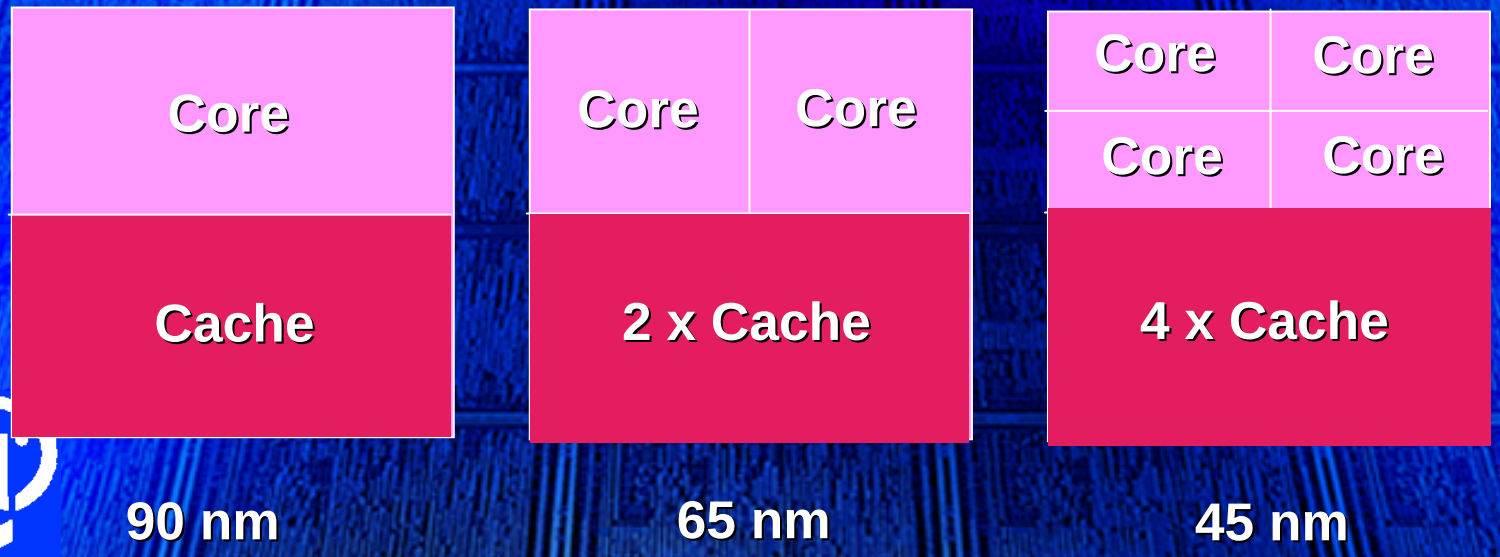
OBJECTIVE: Sustained Technology Leadership

Agenda

- Semiconductor Technology Evolution
- Design Challenges
- Why Multi-Core Processor Chips?
- Power/Performance Trade-Offs
- CMP Directions
- Beyond CMP
- Summary

Possible Evolution

- Transistor density doubles with each process generation
- New generation enables complex new core
- Possible alternative design point
  - Double the cache capacity in same area
  - Double the number of processor cores
  - Frequency improves with process technology



Ramping Multi-core Everywhere

	2005	2006*	2007*
Desktop Mainstream/Performance	Shipping	>70%	>90%
Mobile Mainstream/Performance	Shipping	>70%	>90%
Server	Shipping	>85%	~100%

Desktop Client

Mobile Client

Server & Workstation

* Data is projected run rate exiting the year. Source: Intel

Expect to ship >60 million multi-core processors by end of 2006

CMP Challenges

- How much Thread Level Parallelism is there in most workloads?
- Ability to generate code with lots of threads & performance scaling
- Thread synchronization
- Operating systems for parallel machines
- Single thread performance tradeoff
- Power limitations
- On-chip interconnect/cache infrastructure
- Memory and I/O bandwidth required

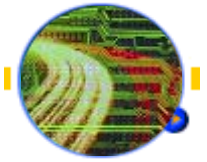
Intel's Software Tools and Support

Thread Checker
Thread Profiler

Solutions, Blueprints,
Sizing/Scaling Guides

Math Kernel Libraries
Performance Primitives

Driver Optimization
Labs



Compilers

Solution Services
Developer Services

VTune™ Analyzers

Software College
Early Access Programs

How Many Cores?

- Where does the doubling stop?
  - Driven by software issues
- Today Microsoft Windows supports only 64 threads!
- How many applications scale to 64 threads?
- How well does performance scale with thread count?

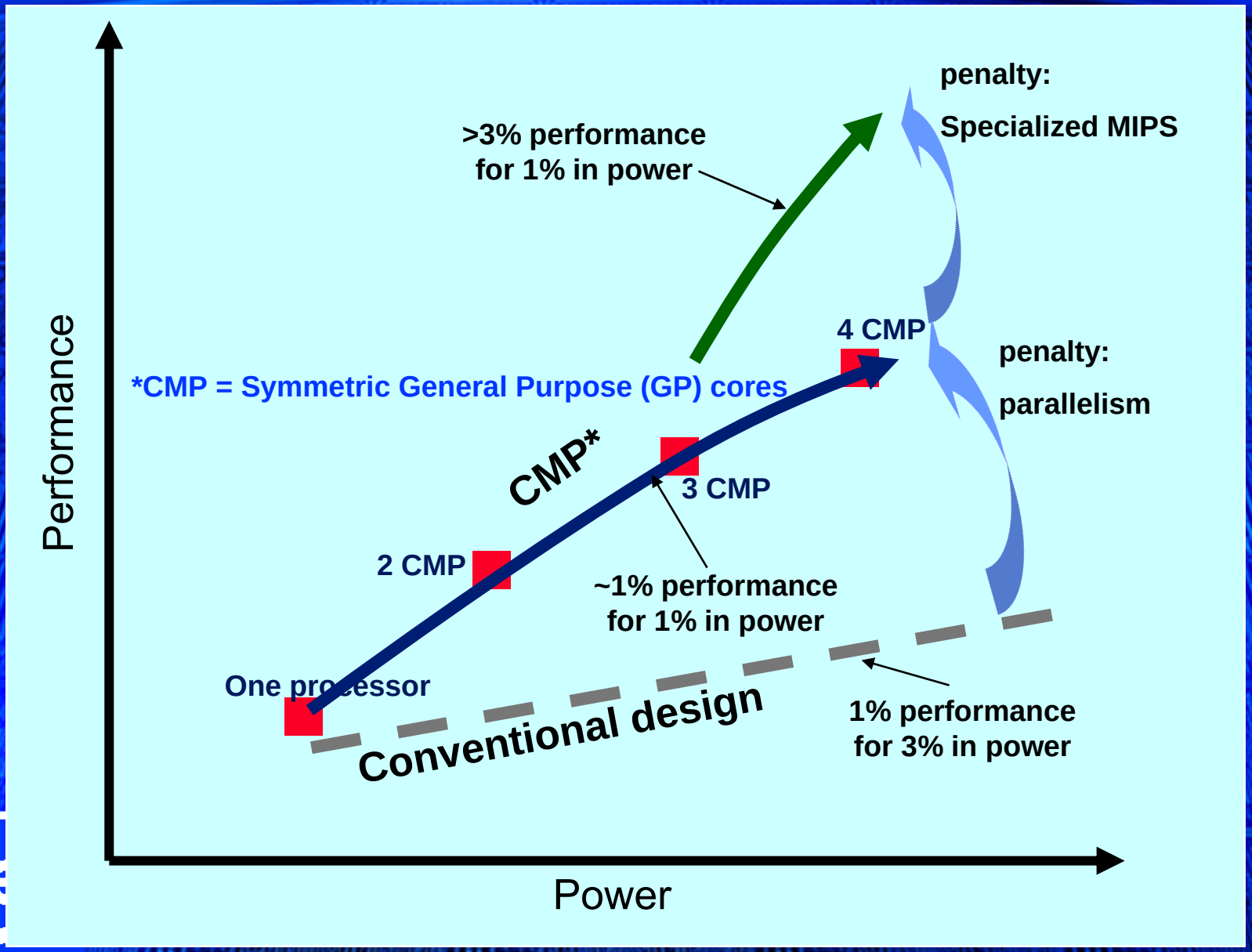
Agenda

- Semiconductor Technology Evolution
- Design Challenges
- Why Multi-Core Processor Chips?
- Power/Performance Trade-Offs
- CMP Directions
- Beyond CMP
- Summary

Looking Beyond CMP

- How far do we push the number of general purpose cores?
- Is there are role for application specific engines?
- Programming model for heterogeneous cores

Improving Power Efficiency



Application Specific Engines

- Can achieve better power efficiency than general purpose cores
- Simpler design due to targeted application and lack of support for full operating system
- Challenge
  - Needs to support high volume application
  - Reconfigurable?
- Graphics and Multimedia engines are good candidates

Agenda

- Semiconductor Technology Evolution
- Design Challenges
- Why Multi-Core Processor Chips?
- Power/Performance Trade-Offs
- CMP Directions
- Beyond CMP
- Summary

Summary

- One billion transistors are here already!
- Chip Level Multiprocessing and large caches can exploit Moore's Law
- Amount of parallelism in future microprocessor systems will increase
- Heterogeneous cores may emerge eventually
- Need applications and tools that can exploit parallelism
- Design challenges and software issues remain

Collaborate, Innovate, Lead!

Closing Thought

"Don't be encumbered by past history, go off and do something wonderful."

- Robert Noyce
Intel Co-founder