White Paper

Inside Intel® Core™ Microarchitecture

Setting New Standards for Energy-Efficient Performance

Ofri Wechsler
Intel Fellow, Mobility Group Director,
Mobility Microprocessor Architecture
Intel Corporation

Introduction

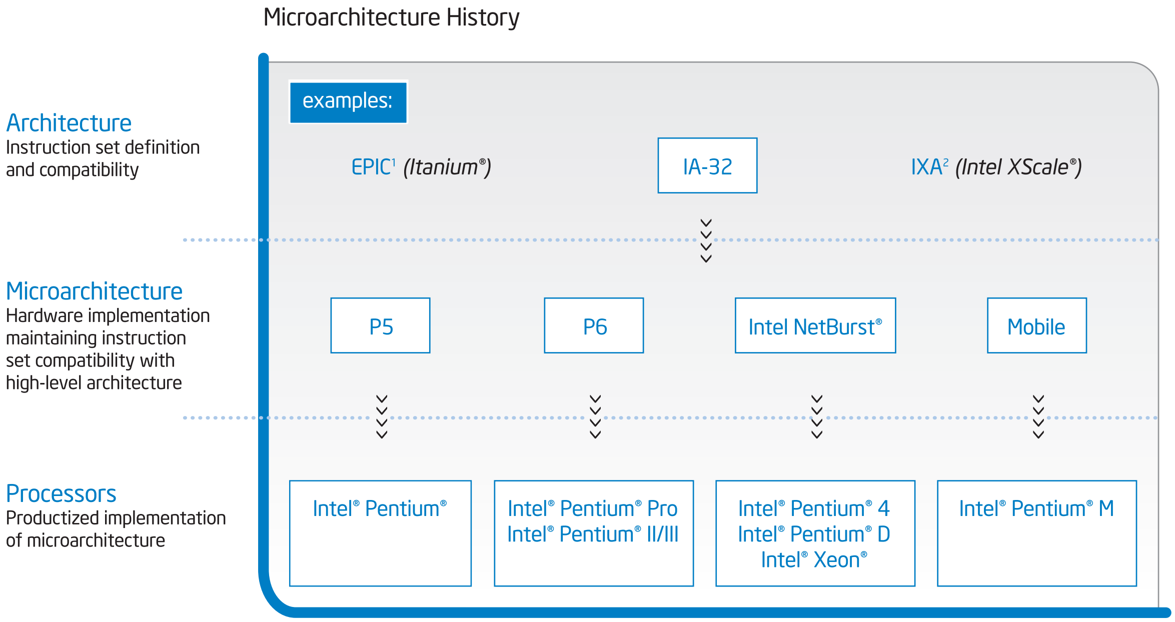
The Intel® Core™ microarchitecture is a new foundation for Intel® architecture-based desktop, mobile, and mainstream server multi-core processors. This state-of-the-art multi-core optimized and power-efficient microarchitecture is designed to deliver increased performance and performance-per-watt—thus increasing overall energy efficiency. This new microarchitecture extends the energy efficient philosophy first delivered in Intel's mobile microarchitecture found in the Intel® Pentium® M processor, and greatly enhances it with many new and leading edge microarchitectural innovations as well as existing Intel NetBurst® microarchitecture features. What's more, it incorporates many new and significant innovations designed to optimize the power, performance, and scalability of multi-core processors.

The Intel Core microarchitecture shows Intel's continued innovation by delivering both greater energy efficiency and compute capability required for the new workloads and usage models now making their way across computing.

With its higher performance and low power, the new Intel Core microarchitecture will be the basis for many new solutions and form factors. In the home, these include higher performing, ultra-quiet, sleek and low-power computer designs, and new advances in more sophisticated, user-friendly entertainment systems. For IT, it will reduce space and electricity burdens in server data centers, as well as increase responsiveness, productivity and energy efficiency across client and server platforms. For mobile users, the Intel Core microarchitecture means greater computer performance combined with leading battery life to enable a variety of small form factors that enable world-class computing "on the go." Overall, its higher performance, greater energy efficiency, and more responsive multitasking will enhance user experiences in all environments—in homes, businesses, and on the go.

Intel® Core™ Microarchitecture Design Goals

Intel continues to drive platform enhancements that increase the overall user experience. Some of these enhancements include areas such as connectivity, manageability, security, and reliability, as well as compute capability. One of the means of significantly increasing compute capability is with Intel® multi-core processors delivering greater levels of performance and performance-per-watt capabilities. The move to multi-core processing has also opened the door to many other micro-architectural innovations to continue to even further improve performance. Intel Core microarchitecture is one such state-of-the-art microarchitectural update that was designed to deliver increased performance combined with superior power efficiency. As such, Intel Core microarchitecture is focused on enhancing existing and emerging application and usage models across each platform segment, including desktop, server, and mobile.



1. EPIC (Explicitly Parallel Instruction Computing) 2. IXA (Intel® Internet Exchange Architecture)

Figure 1. This diagram shows the difference between processor architecture and microarchitecture. Processor Architecture refers to the instruction set, registers, and memory data-resident data structures that are public to a programmer. Processor architecture maintains instruction set compatibility so processors will run code written for processor generations, past, present, and future. Microarchitecture refers to the implementation of processor architecture in silicon. Within a family of processors, the microarchitecture is often enhanced over time to deliver improvements in performance and capability, while maintaining compatibility to the architecture.

Delivering Energy-Efficient Performance

In the microprocessor world, performance usually refers to the amount of time it takes to execute a given application or task, or the ability to run multiple applications or tasks within a given period of time. Contrary to a popular misconception, it is not clock frequency (GHz) alone or the number of instructions executed per clock cycle (IPC) alone that equates to performance. True performance is a combination of both clock frequency (GHz) and IPC.[1] As such, performance can be computed as a product of frequency and instructions per clock cycle:

Performance = Frequency x Instructions per Clock Cycle

This shows that the performance can be improved by increasing frequency, IPC, or possibly both. It turns out that frequency is a function of both the manufacturing process and the microarchitecture. At a given clock frequency, the IPC is a function of processor microarchitecture and the specific application being executed. Although it is not always feasible to improve both the frequency and the IPC, increasing one and holding the other close to constant with the prior generation can still achieve a significantly higher level of performance.

In addition to the two methods of increasing performance described above, it is also possible to increase performance by reducing the number of instructions that it takes to execute the specific task being measured. Single Instruction Multiple Data (SIMD) is a technique used to accomplish this. Intel first implemented 64-bit integer SIMD instructions in 1996 on the Intel® Pentium® processor with MMX™ technology and subsequently introduced 128-bit SIMD single precision floating point, or Streaming SIMD Extensions (SSE), on the Pentium III processor and SSE2 and SSE3 extensions in subsequent generations. Another innovative technique that Intel introduced in its mobile microarchitecture is called microfusion. Intel's microfusion combines many common micro-operations or micro-ops (instructions internal to the processor) into a single micro-op, such that the total number of micro-ops that need to be executed for a given task is reduced.

As Intel has continued to focus on delivering capabilities that best meet customer needs, it has also become important to look at delivering optimal performance combined with energy efficiency—to take into account the amount of power the processor will consume to generate the performance needed for a specific task. Here power consumption is related to the dynamic

1. Performance also can be achieved through multiple cores, multiple threads, and using special purpose hardware. Those discussions are beyond the scope of this paper. Please refer to Intel's white paper: *Platform 2015: Intel® Processor and Platform Evolution for the Next Decade* for further details.

capacitance (the ratio of the electrostatic charge on a conductor to the potential difference between the conductors required to maintain that charge) required to maintain IPC efficiency times the square of the voltage that the transistors and I/O buffers are supplied with times the frequency that the transistors and signals are switching at. This can be expressed as:

Power = Dynamic Capacitance x Voltage x Voltage x Frequency

Taking into account this power equation along with the previous performance equation, designers can carefully balance IPC efficiency and dynamic capacitance with the required voltage and frequency to optimize for performance and power efficiency. The balance of this paper will explain how Intel's new microarchitecture delivers leadership performance and performance-per-watt using this foundation.

Intel® Core™ Microarchitecture Innovations

Intel has long been the leader in driving down power consumption in laptops. The mobile microarchitecture found in the Intel Pentium M processor and Intel® Centrino® mobile technology has consistently delivered an industry-leading combination of laptop performance, performance-per-watt, and battery life. Intel NetBurst microarchitecture has also delivered a number of innovations enabling great performance in the desktop and server segments.

Now, Intel's new microarchitecture will combine key industry-leading elements of each of these existing microarchitectures, along with a number of new and significant performance and power innovations designed to optimize the performance, energy efficiency, and scalability of multi-core processors.

The balance of this paper will discuss these key Intel Core microarchitecture innovations:

- Intel® Wide Dynamic Execution
- Intel® Intelligent Power Capability
- Intel® Advanced Smart Cache
- Intel® Smart Memory Access
- Intel® Advanced Digital Media Boost

Intel® Wide Dynamic Execution

Dynamic execution is a combination of techniques (data flow analysis, speculative execution, out of order execution, and super scalar) that Intel first implemented in the P6 microarchitecture used in the Pentium Pro processor, Pentium II processor, and Pentium III processors. For Intel NetBurst microarchitecture, Intel introduced its Advanced Dynamic Execution engine, a very deep, out-of-order speculative execution engine designed to keep the processor's execution units executing instructions. It also featured an enhanced branch-prediction algorithm to reduce the number of branch mispredictions.

Now with the Intel Core microarchitecture, Intel significantly enhances this capability with Intel Wide Dynamic Execution. It enables delivery of more instructions per clock cycle to improve execution time and energy efficiency. Every execution core is wider, allowing each core to fetch, dispatch, execute, and return up to four full instructions simultaneously. (Intel's Mobile and Intel NetBurst microarchitectures could handle three instructions at a time.) Further efficiencies include more accurate branch prediction, deeper instruction buffers for greater execution flexibility, and additional features to reduce execution time.

One such feature for reducing execution time is macrofusion. In previous generation processors, each incoming instruction was individually decoded and executed. Macrofusion enables common instruction pairs (such as a compare followed by a conditional jump) to be combined into a single internal instruction (micro-op) during decoding. Two program instructions can then be executed as one micro-op, reducing the overall amount of work the processor has to do. This increases the overall number of instructions that can be run within any given period of time or reduces the amount of time to run a set number of instructions. By doing more in less time, macrofusion improves overall performance and energy efficiency

.The Intel Core microarchitecture also includes an enhanced Arithmetic Logic Unit (ALU) to further facilitate macrofusion. Its single cycle execution of combined instruction pairs results in increased performance for less power.

The Intel Core microarchitecture also enhances micro-op fusion—an energy-saving technique Intel first used in the Pentium M processor. In modern mainstream processors, x86 program instructions (macro-ops) are broken down into small pieces, called micro-ops, before being sent down the processor pipeline to be processed. Micro-op fusion "fuses" micro-ops derived from the same macro-op to reduce the number of micro-ops that need to be executed. Reduction in the number of micro-ops results in more efficient scheduling and better performance at lower power. Studies have shown that micro-op fusion can reduce the number of micro-ops handled by the out-of-order logic by more than ten percent. With the Intel Core microarchitecture, the number of micro-ops that can be fused internally within the processor is extended.

Intel® Wide Dynamic Execution

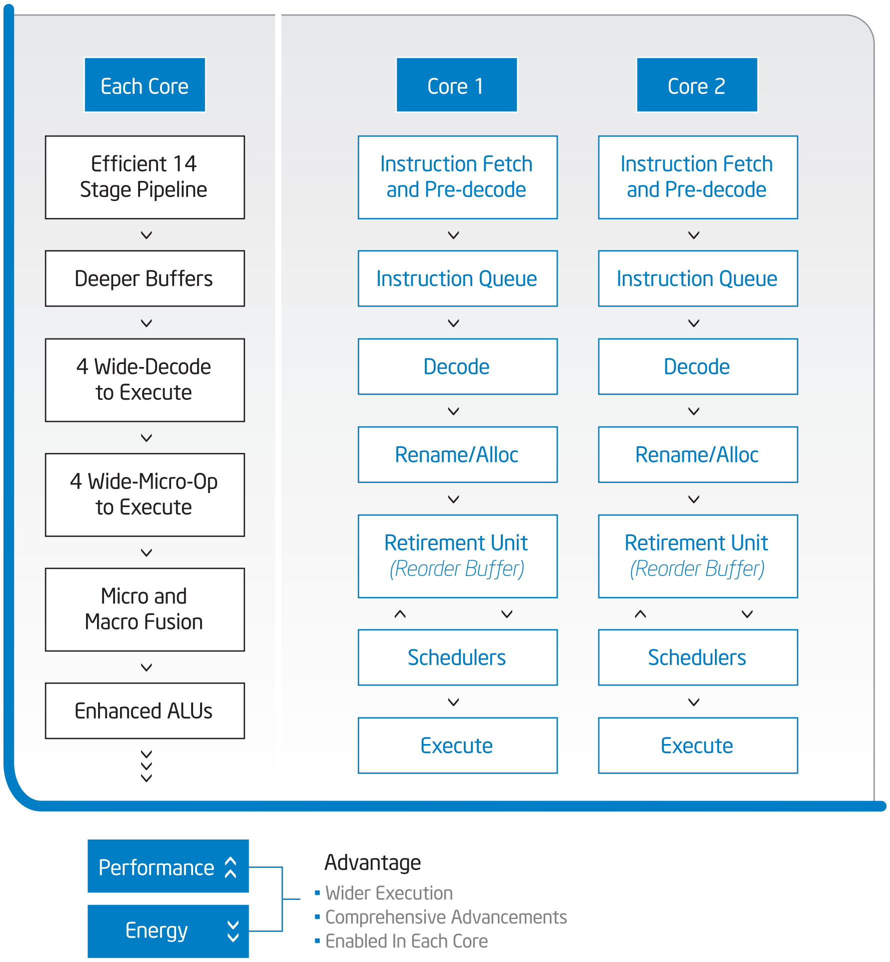


Figure 2. With the Intel Wide Dynamic Execution of the Intel Core microarchitecture, every execution core in a multi-core processor is wider. This allows each core to fetch, dispatch, execute, and return up to four full instructions simultaneously. A single multi-core processor with four cores could fetch, dispatch, execute, and return up to 16 instructions simultaneously.

Intel® Intelligent Power Capability

Intel Intelligent Power Capability is a set of capabilities designed to reduce power consumption and design requirements. This feature manages the runtime power consumption of all the processor's execution cores. It includes an advanced power gating capability that allows for an ultra fine-grained logic control that turns on individual processor logic subsystems only if and when they are needed. Additionally, many buses and arrays are split so that data required in some modes of operation can be put in a low power state when not needed.

In the past, implementing power gating has been challenging because of the power consumed in the powering down and ramping back up, as well as the need to maintain system responsiveness when returning to full power. Through Intel Intelligent Power Capability, we've been able to satisfy these concerns, ensuring both significant power savings without sacrificing responsiveness. The result is excellent energy optimization enabling the Intel Core microarchitecture to deliver more energy-efficient performance for desktop PCs, mobile PCs, and servers.

Intel® Advanced Smart Cache

The Intel Advanced Smart Cache is a multi-core optimized cache that improves performance and efficiency by increasing the probability that each execution core of a dual-core processor can access data from a higher-performance, more-efficient cache subsystem. To accomplish this, Intel shares L2 cache between cores.

To understand the advantage of this design, consider that most current multi-core implementations don't share L2 cache among execution cores. This means when two execution cores need the same data, they each have to store it in their own L2 cache. With Intel's shared L2 cache, the data only has to be stored in one place that each core can access. This better optimizes cache resources.

By sharing L2 caches among each core, the Intel Advanced Smart Cache also allows each core to dynamically utilize up to 100 percent of available L2 cache. When one core has minimal cache requirements, other cores can increase their percentage of L2 cache, reducing cache misses and increasing performance. Multi-Core Optimized Cache also enables obtaining data from cache at higher throughput rates.

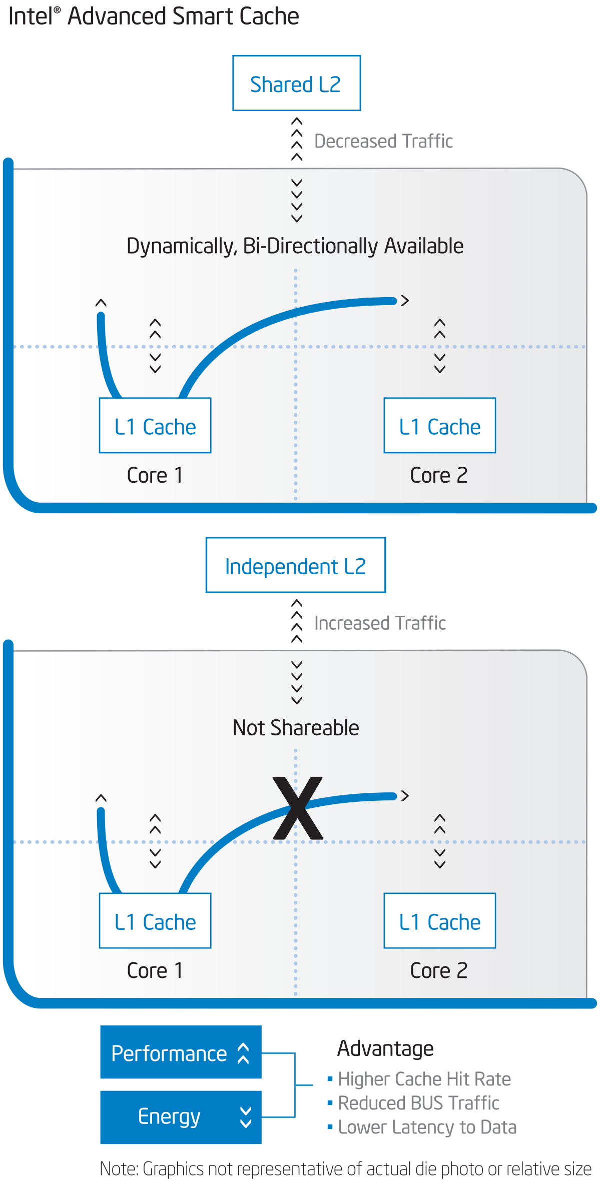


Figure 3. In a multi-core processor where two cores don't share L2 cache, an idle core also means idle L2 cache space. This is a critical waste of resources, especially when another core may be suffering a performance hit because its L2 cache is too full. Intel's shared L2 cache design enables the working core to dynamically take over the entire L2 cache and maximize performance.

Intel® Smart Memory Access

Intel Smart Memory Access improves system performance by optimizing the use of the available data bandwidth from the memory subsystem and hiding the latency of memory accesses. The goal is to ensure that data can be used as quickly as possible and that this data is located as close as possible to where it's needed to minimize latency and thus improve efficiency and speed.

Intel Smart Memory Access includes an important new capability called memory disambiguation, which increases the efficiency of out-of-order processing by providing the execution cores with the built-in intelligence to speculatively load data for instructions that are about to execute BEFORE all previous store instructions are executed. To understand how this works, we have to look at what happens in most out-of-order microprocessors.

Normally when an out-of-order microprocessor reorders instructions, it can't reschedule loads ahead of stores because it doesn't know if there are any data location dependencies it might be violating. Yet in many cases, loads don't depend on a previous store and really could be loaded before, thus improving efficiency. The problem is identifying which loads are okay to load and which aren't.

Intel's memory disambiguation uses special intelligent algorithms to evaluate whether or not a load can be executed ahead of a preceding store. If it intelligently speculates that it can, then the load instructions can be scheduled before the store instructions to enable the highest possible instruction-level parallelism. If the speculative load ends up being valid, the processor spends less time waiting and more time processing, resulting in faster execution and more efficient use of processor resources. In the rare event that the load is invalid, Intel's memory disambiguation has built-in intelligence to detect the conflict, reload the correct data and re-execute the instruction.

In addition to memory disambiguation, Intel Smart Memory Access includes advanced prefetchers. Prefetchers do just that—"prefetch" memory contents before they are requested so they can be placed in cache and then readily accessed when needed. Increasing the number of loads that occur from cache versus main memory reduces memory latency and improves performance.

To ensure data is where each execution core needs it, the Intel Core microarchitecture uses two prefetchers per L1 cache and two prefetchers per L2 cache. These prefetchers detect multiple streaming and strided access patterns simultaneously. This enables them to ready data in the L1 cache for "just-in-time" execution. The prefetchers for the L2 cache analyze accesses from cores to ensure that the L2 cache holds the data the cores may need in the future.

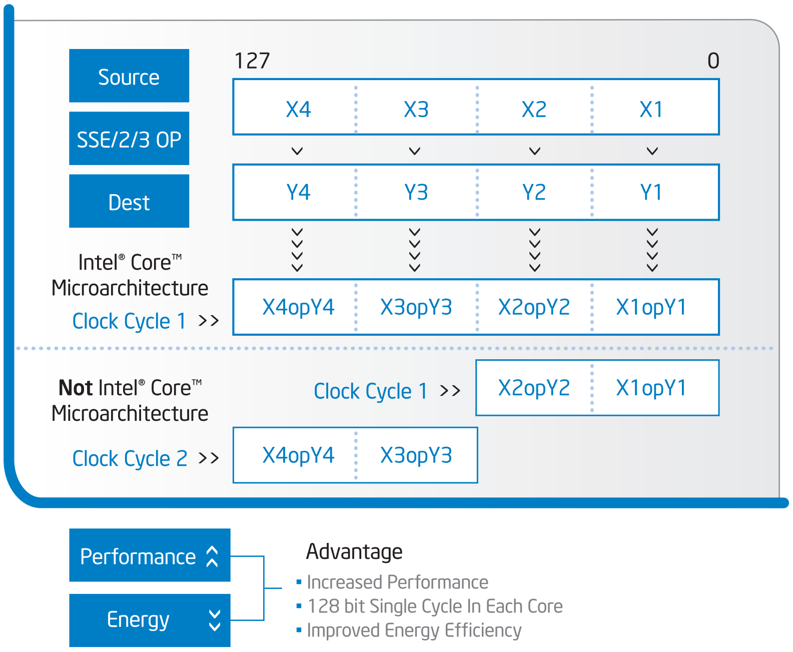
Combined, the advanced prefetchers and the memory disambiguation result in improved execution throughput by maximizing the available system-bus bandwidth and hiding latency to the memory subsystem.

Intel® Advanced Digital Media Boost

The Intel Advanced Digital Media Boost is a feature that significantly improves performance when executing Streaming SIMD Extension (SSE) instructions. 128-bit SIMD integer arithmetic and 128-bit SIMD double-precision floating-point operations reduce the overall number of instructions required to execute a particular program task, and as a result can contribute to an overall performance increase. They accelerate a broad range of applications, including video, speech and image, photo processing, encryption, financial, engineering, and scientific applications. SSE instructions enhance the Intel architecture by enabling programmers to develop algorithms that can mix packed, single-precision, floating-point, and integers, using both SSE and MMX instructions respectively.

On many previous generation processors, 128-bit SSE, SSE2 and SSE3 instructions were executed at a sustained rate of one complete instruction every two clock cycles—for example, the lower 64 bits in one cycle and the upper 64 bits in the next. The Intel Advanced Digital Media Boost feature enables these 128-bit instructions to be completely executed at a throughput rate of one per clock cycle, effectively doubling the speed of execution for these instructions. This further adds to the overall efficiency of Intel Core microarchitecture by increasing the number of instructions handled per cycle. Intel Advanced Digital Media Boost is particularly useful when running many important multimedia operations involving graphics, video and audio, and processing other rich data sets that use SSE, SSE2 and SSE3 instructions.

Intel® Advanced Digital Media Boost
SSE Operation (SSE/SSE2/SSE3)



Note: Graphics not representative of actual die photo or relative size

Figure 4. With Intel Single Cycle SSE, 128-bit instructions can be completely executed at a throughput rate of one per clock cycle, effectively doubling the speed of execution for these instructions.

Intel® Core™ Microarchitecture and Software

Intel expects that the majority of existing applications will see immediate benefits when running on processors that are based upon the Intel Core microarchitecture. For more information on software and the Intel Core microarchitecture, please visit the Intel® Software Network on the Intel Web site at **www.intel.com/software**.

www.intel.com

Summary

The Intel Core microarchitecture is a new, state-of-the-art, multi-core optimized microarchitecture that delivers a number of new and innovative features that will set new standards for energy-efficient performance. This energy-efficient, low power, high-performing, and scaleable blueprint will be the foundation for future Intel-based server, desktop, and mobile multi-core processors.

This new microarchitecture extends the energy efficient philosophy first delivered in Intel's mobile microarchitecture found in the Intel® Pentium® M processor, and greatly enhances it with many new and leading edge microarchitectural innovations as well as existing Intel NetBurst® microarchitecture features. Products based on Intel Core microarchitecture will enter the market in the second half of 2006, and will enable a wave of innovation across desktop, server, and mobile platforms. Desktops can deliver greater compute performance as well as ultra-quiet, sleek and low-power designs. Servers can deliver greater compute density, and laptops can take the increasing compute capability of multi-core to new mobile form factors.

Learn More

You can discover much more by visiting these Intel Web sites:

Intel® Core™ Duo processors
www.intel.com/products/processor/coreduo

Intel® Platforms
www.intel.com/platforms

Intel Multi-Core
www.intel.com/multi-core

Intel Architectural Innovation
www.intel.com/technology/architecture

Energy-Efficient Performance
www.intel.com/technology/eep

Author Biography

Ofri Wechsler is an Intel Fellow in the Mobility Group and director of Mobility Microprocessor Architecture at Intel Corporation. In this role, Wechsler is responsible for the architecture of the new Intel® Core™ Duo processor, the upcoming processor code named "Merom," and the architecture development of other next-generation CPUs. Previously, Wechsler served as manager for the IDC AV, responsible for the validation of the P55C. Wechsler joined Intel in 1989 as a design engineer for i860. He received his bachelor's degree in electrical engineering from Ben Gurion University, Beer Sheva, Israel, in 1998. He has four U.S. patents.

Printed in the United States. 0306/RMR/HBD/2K 311830-001US