

The Intel® processor roadmap for industry-standard servers

technology brief, 6th Edition



Abstract.....	2
Introduction.....	2
Xeon x86 processors on 180nm and 130nm technology.....	2
NetBurst architecture.....	3
Pentium 4 processors.....	4
Xeon processors.....	4
Hyper-Thread Technology.....	4
Xeon x86 processors on 90nm technology.....	6
Architecture.....	6
Prescott new instructions.....	7
64-bit extensions - EM64T.....	7
Dual-core technology.....	8
Conclusion.....	9
Call to action.....	10

Abstract

Intel is introducing significant changes to its product roadmap in the near future. Processors that include Intel's Hyper-Thread Technology have dramatically increased 32-bit processor performance, and Intel currently supports 64-bit extensions to its 32-bit processor family. This paper provides a roadmap for these processors and details some of the more important upcoming changes as they affect industry-standard enterprise servers.

Introduction

As standards-based computing has pushed into the enterprise server market, the demand for increased performance and greater variety in processor solutions has grown with it. To meet this demand, Intel continues to introduce processor innovations and new speeds at a pace that sometimes makes it difficult to keep track of all the developments. This paper attempts to clarify the direction and progress of Intel's processor development by summarizing the recent history and near-term plans for Intel processors as they relate to the industry-standard enterprise server market.

Xeon x86 processors on 180nm and 130nm technology

In the second half of 2000, Intel launched the NetBurst architecture, its seventh-generation 32-bit architecture. The single-processor version of this architecture is the Pentium 4. Versions intended for multi-processor environments are referred to as Xeon® architecture (for dual-processor systems) and Xeon MP architecture (for systems using more than two processors). These processors were manufactured on Intel's 180nm and 130nm silicon CMOS process technology. Table 1 includes the release dates and features of previously released Intel x86 processors as well as processors projected to be available through 2006.

Table 1. Intel x86 processors

Code Name	Market name	Line width (nm)	Description	Date available/ Projected availability	Cache	Bus speed ¹ (MT/s)
Northwood	Pentium 4	130	Die shrink of Pentium 4, targeting single-processor value systems.	2H2001	512-KB L2	800
Prestonia	Xeon or Xeon with 512-KB L2	130	Version of Northwood with Hyper-Thread. Designed for dual-processor applications.	1Q2002	512-KB L2	533
Gallatin	Xeon MP	130	Die shrink of Foster-MP with Hyper-Thread Technology. Designed for multi-processor applications; targeting performance systems.	2H2002 – 2MB 1Q2004 – 4MB	512-KB L2 1MB – 4MB L3	400
Prescott	Pentium 4	90	Die shrink of Pentium 4, with 64-bit extensions, 31-stage pipeline, HT, and Prescott new instructions (SSE3).	1H2004	1MB L2	800
Smithfield	Pentium D	90	Dual-core uni-processor	2H2005	1MB L2/Core	800
Presler	TBD	65	Dual-core uni-processor	1H2006	TBD	>800

Code Name	Market name	Line width (nm)	Description	Date available/ Projected availability	Cache	Bus speed ¹ (MT/s)
Nocona	Xeon	90	Dual-processing version of Prescott	2H2004	1MB L2	800
Irwindale	Xeon	90	2MB L2 version of Nocona	1Q2005	2MB L2	800
Cranford	Xeon MP	90	Xeon MP	1Q2005	1MB L2	667
Prescott 2M	Xeon	90	2MB L2 version of Prescott	1Q2005	2MB L2	800
Potomac	Xeon MP	90	Xeon MP	1Q2005	1MB L2, >8MB L3	667
Dempsey	Xeon	65	Dual-core Xeon	1H2006	TBD	800
Paxville	Xeon MP	90	Dual-core Xeon MP	1Q2006	TBD	667

¹ MT/s is an abbreviation for Mega-Transfers per second. A bus operating at 200 MHz and transferring four data packets on each clock (referred to as quad-pumped) would have 800 MT/s.

NetBurst architecture

One performance-enhancing feature of the NetBurst architecture is its *hyper-pipeline*, a 20-stage branch-prediction pipeline. Previous 32-bit processors had a 10-stage pipeline. The hyper-pipeline can contain more than 100 instructions at once and can handle up to 48 loads and stores concurrently. Specific to this NetBurst design is an improved branch-prediction algorithm aided by a large branch target array that stores branch predictions.

Other enhancements for NetBurst include:

- Higher bandwidth for instruction fetches
- 256-KB Level 2 (L2) cache with 64-byte cache lines
- NetBurst system bus: a 64-bit, 100-MHz bus capable of providing 3.2 GB/s of bandwidth by double pumping the address and quad pumping the data. The 100-MHz quad pumped data bus is also referred to as a 400-MHz data bus. To provide higher levels of performance, Intel added support for 533 MHz to the Pentium 4 (P4) and Xeon processors and later added support for 800 MHz to the P4.
- Integer arithmetic logic unit (ALU) running at twice the clock speed (double data rate)
- Modified floating point unit (FPU)
- Streaming SIMD extension 2 (SSE2): New instructions bring the total to 144 SIMD instructions to manage floating point, application, and multimedia performance.
- Advanced dynamic execution
- Deeper instruction window for out-of-order, speculative execution and improved branch prediction over the P6 dynamic execution core
- Execution trace cache (stores pre-decoded micro-operations)
- Enhanced floating point/multimedia engine
- Hyper-threading in Xeon processors and P4 processors

Pentium 4 processors

P4 processors are versions of the NetBurst architecture intended for low-cost, single-processor servers. The 180nm version of the P4 was known as Willamette, and the 130nm version was known as Northwood. HP uses the same P4 processor in commercial desktops, consumer desktops, and low-cost, single-processor servers. For use in the servers, however, the P4 processors are qualified with server chipsets under server workloads.

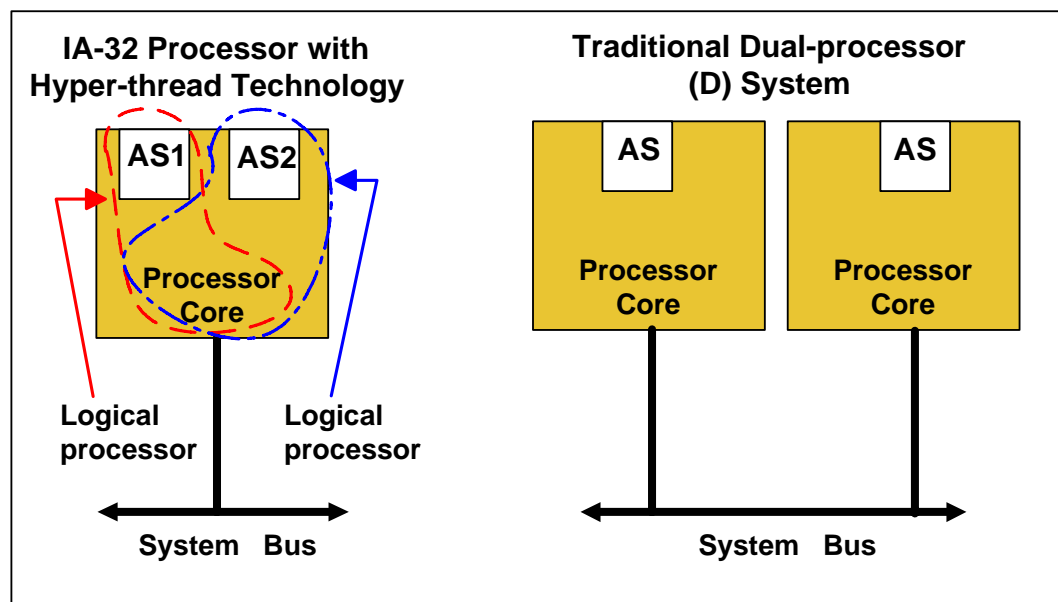
Xeon processors

The NetBurst processor version intended for use in dual-processor environments such as workstations and servers is called the Xeon processor. The version designed for multi-processor environments is called Xeon MP processor. Both the dual and multi-processor designs incorporate new features, such as Hyper-Thread Technology and larger caches, to facilitate multi-processor performance. The MP version also adds a third-level cache on the processor die to improve performance in multi-processing systems by reducing competition between processors for shared resources.

Hyper-Thread Technology

Intel Hyper-Thread Technology is a design enhancement for server environments. It takes advantage of the fact that, according to Intel estimates, the utilization rate for the execution units in a NetBurst processor is typically only about 35 percent. To improve the utilization rate, Hyper-Threading adds Multi-Thread-Level Parallelism (MTLP) to the design. In essence, MTLP means that the core receives two instruction streams from the operating system (OS) to take advantage of idle cycles on the execution units of the processor. For one physical processor to appear as two distinct processors to the OS, the new design replicates the pieces of the processor with which the OS interacts to create two logical processors in one package. These replicated components include the instruction pointer, the interrupt controller, and other general-purpose registers—all of which are collectively referred to as the Architectural State, or AS (see Figure 1).

Figure 1. Comparison of Hyper-Threading and traditional dual-processor



Since multi-processing operating systems such as Microsoft® Windows® and Linux® are designed to divide their workload into threads that can be independently scheduled, these OSes can send two distinct threads to work their way through execution in the same device. This provides the opportunity for a higher abstraction level of parallelism at the thread level rather than simply at the instruction level, as in the Pentium 4 design. To illustrate this concept, refer to Table 2, where it can be seen that instruction-level parallelism is able to take advantage of opportunities in the instruction stream to execute independent instructions at the same time. Thread-level parallelism, shown in Table 3, takes this a step further, since two independent instruction streams are available for simultaneous execution opportunities.

It should be noted that the performance gain from adding Hyper-Threading will not equal the expected gain from adding a second physical processor. The overhead to maintain the threads and the requirement to share processor resources will necessarily limit the Hyper-Thread performance. Nevertheless, Hyper-Threading is a valuable and cost-effective addition to the Pentium 4 design.

Table 2. Instruction-level parallelism

Instruction number	Instruction thread	Instruction-level parallelism
1	Read register A	Operations 1, 2, and 3 are independent and can execute simultaneously if resources permit.
2	Write register B	
3	Read register C	
4	Add A + B	This operation must wait for instructions 1 and 2 to complete, but it can execute in parallel with operation 3.
5	Inc A	This operation needs to wait for the completion of instruction 4 before executing.

Table 3. Thread-level parallelism

Instruction number	Instruction thread 1	Instruction number	Instruction thread 2	Thread-level parallelism
1a	Read A	1b	Add D + E	None of the instructions in Thread 2 depend on those in Thread 1 so, to the extent that execution units are available, any of them can execute in parallel with those in Thread 1.
2a	Read B	2b	Inc E	
3a	Read C	3b	Read F	As an example, instruction 2b must wait for instruction 1b, but does not need to wait for 1a. Similarly, if two arithmetic units are available, 4a and 4b can execute at the same time.
4a	Add A + B	4b	Add E+F	
5a	Inc A	5b	Write E	

According to Intel’s internal simulations, Hyper-Thread Technology achieves its objective of improving the microarchitecture utilization rate significantly. Improved performance is the real goal though, and Intel reports that the performance gain can be as high as 30 percent.

The performance gained by these design changes is limited by the fact that two threads now share and compete for processor resources, such as the execution pipeline and L1 and L2 caches. There is some risk that data that one thread needs can be replaced in a cache by data that the other is using, resulting in a higher turnover of cache data (referred to as thrashing) and a reduced hit rate. Hyper-Thread Technology also puts a heavier load on the OS to allocate threads and switch contexts on the device. The evaluation of the threads for parallelism and context switching are OS tasks and increase the operating overhead.

Currently, Hyper-Thread Technology presents little in the way of software licensing issues. Intel asserts that the Hyper-Thread design is still only a single-processor unit, so customers should not have to purchase two software licenses for each processor. This is true for Microsoft SQL Server 2000¹ and Windows Server 2003, which only require one license for each physical processor, regardless of how many logical processors it contains. However, Windows 2000 Server does not make this distinction between physical and logical processors and fills the licensing limit based on the number of processors the BIOS discovers at boot time.²

Xeon x86 processors on 90nm technology

In 2004, Intel introduced major changes to the P4 and Xeon processor lines. These processors, which are still marketed as P4 and Xeon, are manufactured on Intel's 90nm silicon CMOS process technology. Codenames for these processors include "Prescott," "Prescott 2M," "Smithfield," and "Presler" for the uni-processor P4; "Nocona," "Irwindale," and "Dempsey" for Xeon; and "Cranford," "Potomac," and "Paxville" for Xeon MP.

Enhancements for NetBurst on 90nm technology include:

- Die-shrink to 90nm (65 nm for Presler and Dempsey)
- Larger, more effective caches (L1 increased from 8KB, 4-way to 16KB, 8-way. L2 increased from 512KB to 1MB.)
- Faster processor bus: a 64-bit, 200-MHz bus capable of providing 6.4 GB/s of bandwidth by double pumping the address and quad pumping the data. The 200-MHz Quad-pumped data bus is also referred to as an 800-MHz data bus.
- Extended hyper-pipeline (31 stages versus 20 stages) to enable high CPU core frequencies
- Enhanced execution units including the additional of a dedicated integer multiplier, and support for shift and rotate instruction execution on a fast ALU
- Improved branch prediction to help compensate for longer pipeline
- Prescott New Instructions (described below)
- Larger execution schedulers and execution queues
- Improved hardware memory prefetcher
- Improved Hyper-Threading
- 64-bit extensions (described below)
- Dual-core (for Smithfield, Dempsey, and Paxville)

Architecture

In keeping with its history of regularly increasing processor frequencies, Intel has extended the hyper-pipeline queue from 20 (in the earlier Pentium 4 design) to 31 stages. The goal of the longer pipeline is to split the amount of work that must be done during a single clock period into smaller pieces so the work can be done faster and the device can be clocked at higher frequencies in the future. The biggest drawback to this approach is that, as the pipe gets longer, interruptions to the regular flow of instructions in the pipe become progressively more costly in terms of performance. For example, whenever the branch-prediction logic predicts the wrong instruction sequence, the entire contents of the pipeline must be flushed and the correct instruction sequence must be fetched in. This stalls the processor until the instructions propagate through the pipe. To mitigate such stalls, Intel improved the branch-prediction algorithm sufficiently to prevent this deeper pipeline from causing performance degradation.

¹ See paper on this topic at: www.microsoft.com/sql/howtobuy/SQLonHTT.doc

² For more information on Hyper-Threading technology, visit: www.microsoft.com/windows2000/docs/hyperthreading.doc

Prescott new instructions

The Prescott design adds 13 new instructions referred to as SSE3, which stands for Streaming Single-Instruction-Multiple-Data (SIMD) Extensions 3, or Prescott New Instructions. As they did in earlier processors, SIMD instructions provide the potential for improved performance because each instruction permits operation on multiple data items at the same time. For Prescott processors, there are new versions of arithmetic, graphics, and Hyper-Thread synchronization instructions.

The arithmetic group consists of one new instruction for converting x87 data into integer format, and five instructions that simplify the process of performing complex arithmetic. Complex numbers actually consist of two numbers, a real and an imaginary component. The new instructions facilitate complex operations because they are designed to operate on both parts of these complex pairs of numbers at the same time.³ Using these instructions also simplifies coding complex arithmetic operations because fewer instructions are needed to accomplish the goal.

The graphics group contains one instruction for video encoding and four that are specific to graphics operations. Finally, two instructions facilitate Hyper-Threaded operation, for example, by allowing one operational thread to be moved to a higher priority than another.

64-bit extensions - EM64T

In response to market demands, Intel has added 64-bit extensions to the x86 architecture that will be included in the next-generation Xeon, Xeon MP, and P4 processors. The key advantage of 64-bit processing is the ability for the system to address a much larger flat memory space (up to 16 Exabytes). Even though today's 32-bit architecture can actually access up to 64 GB of memory, access above the standard 4 GB limit must go through a slow and cumbersome windowing facility. Due to the complexities of this, most 32-bit applications have not made use of the higher address space. The ability to address memory above 4 GB might not seem like a high priority today, because few applications require more than one or two GB. However, as those familiar with the industry know, memory requirements have never stopped at the last artificial ceiling. Rather than forcing users to switch out both the hardware and software infrastructure would require, simply adding 64-bit extensions to the x86 processors provides the same addressing benefit at a much lower cost.

AMD was first to release 64-bit extensions—called AMD64—with its Opteron processor in early 2003. Within the year, Intel responded with its own plans to deliver a similar solution called Extended Memory 64 Technology, or EM64T, which is broadly compatible to AMD64. Both EM64T and AMD64 use the same register sets and definitions, and the 64-bit instructions are nearly identical. HP expects that any minor differences will be handled by the OS and compiler, so that the average application writer or customer should see no differences. New operating systems are required to make use of 64-bit extensions. Red Hat, SuSE, and Microsoft provide AMD64 support and EM64T support.

Even though the larger memory addressing capability is the primary advantage of 64-bit extensions, it is not the only one. 64-bit extensions also provide a larger register set with eight additional general-purpose registers (GPR) and 64-bit versions of the existing registers. With a total of 16 bGPRs, 64-bit extensions provide additional resources that compilers can use to increase performance. The 16-register limit was a tradeoff AMD chose as a good compromise between performance and cost.

³ See Intel's article on SSE3 and complex arithmetic at http://cache-www.intel.com/cd/00/00/06/67/66715_66715.pdf

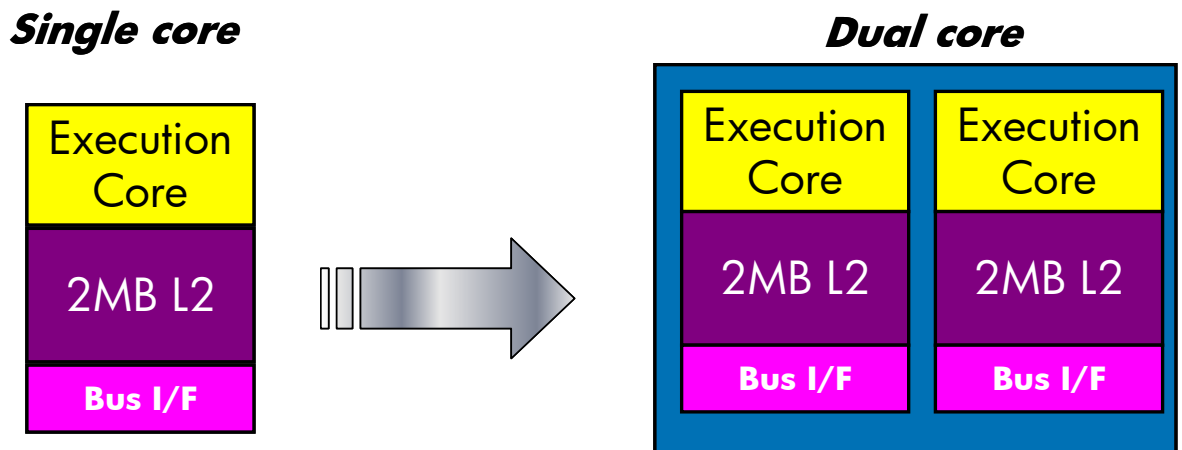
Dual-core technology

Current single-core architectures are being maximized and pushed to their limits in data centers. For instance, with each increase in frequency, single-core processors are becoming less cost effective. Increasing CPU core frequencies delivers lower incremental performance gains at the cost of increasing power requirements. Moreover, in multi-threading environments with only one processor, the multiple threads compete for available compute resources, limiting the increase in performance. These factors create significant barriers for single-core architectures to perform at an increasingly higher level and to keep pace with the growing needs of data centers.

In 2005, the latest enhancement to processor architecture is dual-core technology. This technology is designed to make server processors perform more efficiently in multi-threaded environments. Dual-core technology provides an additional degree of processor design flexibility that offers increased performance while addressing power and timing issues, without being cost prohibitive.

A dual-core processor is a single physical package that contains two, full processor cores per socket. The two cores share the same functional execution units and cache hierarchy; however, the OS recognizes each execution core as an independent processor. Figure 2 illustrates the execution cores of single-core and dual-core processors.

Figure 2. Execution cores of single core and dual-core processors



Dual-core processors with Hyper-Threading Technology have the ability to run two threads on each execution core, allowing these processors to run up to four threads simultaneously. The added capacity provided by the second execution core reduces competition for processor resources and allows for greater processor utilization. Thus, the performance improvement of a dual-core processor is in addition to the improvement that can be achieved with Hyper-Threading Technology.

Intel is currently developing dual-core processors for uni-processor, dual-processor, and multi-processor architecture. The dual-core dual-processor and dual-core multi-processor versions will allow for even greater utilization due to the increase in the number of execution cores. Table 4 includes the Intel dual-core processor roadmap for the near future.

Table 4. Dual-core roadmap

Code name	Processor architecture	Projected availability
Smithfield	Uni-processor	2H2005
Presler	Uni-processor	1H2006
Dempsey	Dual-processor	1H2006
Paxville	Multi-processor	1Q2006

Conclusion

The next few years will again mark dramatic increases in the processing capability of Intel processors. Newer and more powerful 32-bit processors with 64-bit extensions and dual-core technologies will provide additional horsepower to HP industry-standard server products and extend the life of x86-based systems.

Call to action

To help us better understand and meet your needs for ISS technology information, please send questions and comments about this paper to: TechCom@HP.com.

© 2002, 2005 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

Intel, Intel Xeon, Pentium and Itanium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries

Linux is a U.S. registered trademark of Linus Torvalds.

Microsoft and Windows are U.S. registered trademarks of Microsoft Corporation.

TC050402TB, 04/2005

