Inside the NetBurst™ Micro-Architecture of the Intel® Pentium® 4 Processor

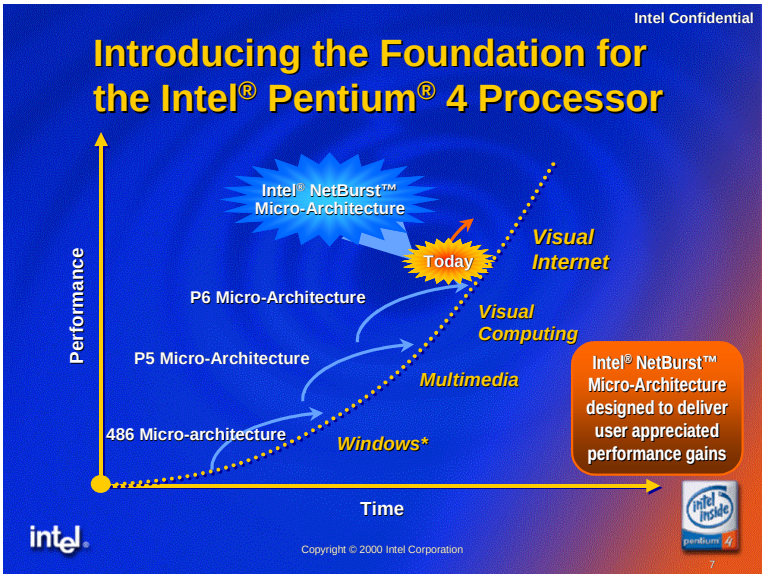
Revision 1.0

Introduction

The Intel® NetBurst™ micro-architecture is the foundation for the Intel Pentium® 4 processor. It includes several important new features and innovations that will allow the Intel Pentium 4 processor and future IA-32 processors to deliver industry leading performance for the next several years. This paper describes the most important features and innovations included in the Intel NetBurst micro-architecture.

> ***Processor architecture versus micro-architecture***
> The architecture of a processor refers to the instruction set, registers, and memory-resident data structures that are public to a programmer and are maintained and enhanced from one generation of architecture to the next. The micro-architecture of a processor refers to implementation of a processor architecture in silicon. Within a family of processors, like the Intel IA-32 processors, the micro-architecture typically changes from one processor generation to the next, while implementing the same public processor architecture. Intel's IA-32 architecture is based on the x86 instruction set and registers. It has been enhanced and extended through generations of IA-32 processors, while maintaining backwards compatibility for code written to run on the earliest IA-32 processors.

As shown in the following figure, new micro-architectures have historically been required to drive increases in processor performance for a particular processor architecture. The early life cycle of each micro-architecture generation delivers a large performance gain over time. However, as the micro-architectural design matures, the performance delivered starts to diminish, requiring new micro-architectural advances in order to maintain the performance trajectory expected by the marketplace. The Intel NetBurst micro-architecture is the latest, true micro-architectural generation from Intel that implements the IA-32 architecture. This micro-architecture, along with several extensions to the IA-32 architecture, have been designed not only to increase the raw instruction processing speed of IA-32 processors, but also to unleash the richness of the visual internet. The Intel NetBurst micro-architecture allows the Pentium 4 processor to deliver this next-generation performance so it can be fully experienced and appreciated by the user, rather than focusing on simply speeding up applications such as word and spreadsheet processing. These type of applications need only to keep pace with a human level of response time, unlike multimedia applications which have an almost unbounded need for performance



Historical view of Intel® micro-architectures and the overall performance trajectory.

What Determines True Processor Performance?

The only measure of performance that really matters is the amount of time it takes to execute a given application. Contrary to a popular misconception, it is not clock frequency (MHz) alone or the number of instructions executed per clock (IPC) alone that equates to performance. True performance is a combination of both clock frequency (MHz) and IPC:

Performance = MHz x IPC

This shows that the performance can be improved by increasing frequency, IPC or optimally both. It turns out that frequency is a function of both the manufacturing process and the micro-architecture. At a given clock frequency, the IPC is a function of processor micro-architecture 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 Pentium® processor with MMX™ technology and subsequently introduced 128bit SIMD single precision floating point (SSE) on the Pentium III processor

Applications can be broadly divided into two categories: integer/basic office productivity applications, and floating point/multimedia applications. The IPC achievable by these different application categories varies greatly, and this variance is strongly affected by the number of branches that the application code typically takes and the predictability of these branches. The more branches taken that are difficult to predict, the higher the possibility of mis-predicting and performing nonproductive work.

Integer and basic office productivity applications, such as word and spreadsheet processing, tend to have many branches in the code that are difficult to predict, thus reducing overall IPC potential. As a result, performance increases on these applications are more resistant to improvements in micro-architectural means, such as deeper pipelines. Also, significantly raising the performance level on these types of applications does not necessarily increase the user's experience, as these types of applications only need to keep pace with the human level of read and write response time and today's higher end Pentium III processors satisfy this requirement.

Floating point and multimedia applications tend to have branches that are very predictable, and thus naturally have a higher average IPC potential. As a result, these types of applications generally scale very well with frequency and are inclined to benefit greatly from deeper pipelines. In addition, the processing power required by these applications tends to be unbounded: the more performance that is available, the better the user's experience.

The NetBurst Micro-Architecture of the Intel Pentium 4 Processor

The Pentium 4 processor, utilizing the NetBurst micro-architecture, is a complete processor re-design that delivers new technologies and capabilities while advancing many of the innovative features, such as "out-of-order speculative execution" and "super-scalar execution", introduced on prior Intel micro-architectural generations. Many of these new innovations and advances were made possible with the improvements in processor technology, process technology and circuit design and could not previously be implemented in high-volume, manufacturable solutions. The features and resulting benefits of the new micro-architecture are defined in the following sections.

Designed for Performance

A focused architectural definition effort was used to study the benefits of many advanced processor technologies and determine the best approach to improve the overall performance of the processor for many years to come. The result of this definition effort was a micro-architecture that significantly increased frequency capabilities to well above 40% higher than that of the P6 micro-architecture (on the same manufacturing process) while maintaining an average IPC that was within approximately 10% to 20% of the P6 micro-architecture. In this design, although the IPC is lower, the increase in frequency capability more than makes up *(Performance = frequency x IPC)* and delivers overall higher performance capability to the end user. This was done in the NetBurst micro-architecture by implementing a **hyper-pipelined technology** where the depth of the pipeline was doubled from that of the P6 micro-architectural generation. Although this deeper pipeline delivers significantly higher levels of frequency, the potential performance impacts associated with the longer pipeline were comprehended and overcome in the design. The design effort focused on the following:

- ***Minimizing the Penalty Associated with Branch Mis-predicts***

  Explanation of Branch Mis-predict Penalty: As with the P6 generation, the NetBurst micro-architecture takes advantage of out-of-order, speculative execution. This is where the processor routinely uses an internal branch prediction algorithm to predict the result of branches in the program code and then speculatively executes instructions down the predicted code branch. Although branch prediction algorithms are highly accurate, they are not 100% accurate. If the processor mis-predicts a branch, all the speculatively executed instructions must be flushed from the processor pipeline in order to restart the instruction execution down the correct program branch. On more deeply pipelined designs, more instructions must be flushed from the pipeline, resulting in a longer recovery time from a branch mis-predict. The net result is that applications that have many, difficult to predict, branches will tend to have a lower average IPC.

  Minimization of mis-predict penalty: To minimize the branch mis-prediction penalty and maximize the average IPC, the deeply pipelined NetBurst micro-architecture greatly reduces the number of branch mis-predicts and provides a quick method of recovering from any branches that have been mis-predicted. To minimize this penalty, the NetBurst micro-architecture has implemented an **Advanced Dynamic Execution** engine and an **Execution Trace Cache.** These features are both described later in this paper.

- ***Keeping the High-Frequency Execution Units Busy (vs. Sitting Idle)***

  Although a processor may have a high frequency capability, it must provide a means to ensure that the execution units (integer and floating point) are continually being supplied with instructions for execution. This ensures that these high-frequency units are executing instructions (not sitting idle). With the high frequency of these execution units in the NetBurst micro-architecture and the implementation of the **Rapid Execution Engine**, where the Arithmetic Logic Units are running at two times the core frequency, Intel has implemented a number of features that ensure that these execution units have a continuous stream of instructions to execute. Intel has implemented a **400-MHz system bus**, an **Advanced Transfer Cache**, an **Execution Trace Cache**, an **Advanced Dynamic Execution** engine and a low-latency **Level 1 Data Cache.** These features work together to quickly provide instructions and data to the processor's high-performance execution units, thus keeping them executing code instead of just idling at high frequency.

- ***Reducing the Number of Instructions Needed to Complete a Task or Program***

  Many applications often perform repetitive operations on large sets of data. Further, the data sets involved in these operations tend to be small values that can be represented with a small number of bits. These two observations can be combined to improve application performance by both compactly representing data sets and by implementing instructions that can operate in these compact data sets. This type of operation is called Single Instruction Multiple Data (SIMD) and can reduce the overall number of instructions that a program is required to execute. The NetBurst micro-architecture implements 144 new SIMD instructions, called **Streaming SIMD Extensions 2 (SSE2)**. The SSE2 instruction set enhances the SIMD instructions previously delivered with MMX technology and SSE technology. These new instructions support 128-bit SIMD integer operations and 128-bit SIMD double-precision floating-point operations. By doubling the amount of data on which a given instruction can operate, only half the number of instructions in a code loop need to be executed.

Intel NetBurst Micro-architecture Feature Details

Hyper-Pipelined Technology: The hyper-pipelined technology of the NetBurst micro-architecture doubles the pipeline depth, compared to the P6 micro-architecture. One of the key pipelines, the branch prediction/recovery pipeline, is implemented with a 20 stage pipeline in the NetBurst micro-architecture, compared to the equivalent pipeline in the P6 micro-architecture, which was implemented with a 10 stage pipeline. This technology significantly increases processor performance and frequency scalability of the base micro-architecture.

Execution Trace Cache: The Execution Trace Cache is an innovative way to implement a Level 1 instruction cache. It caches decoded x86 instructions (micro-ops), thus removing the latency associated with the instruction decoder from the main execution loops. In addition, the Execution Trace Cache stores these micro-ops in the path of program execution flow, where the results of branches in the code are integrated into the same cache line. This increases the instruction flow from the cache and makes better use of the overall cache storage space (12K micro-ops) since the cache no longer stores instructions that are branched over and never executed. The result is a means to deliver a high volume of instructions to the processor's execution units and a reduction in the overall time required to recover from branches that have been mis-predicted.

Rapid Execution Engine: Through a combination of architectural, physical and circuit designs, the simple Arithmetic Logic Units (ALUs) within the processor run at two times the frequency of the processor core. This allows the ALUs to execute certain instructions with a latency that is ½ the duration of the core clock and results in higher execution throughput as well as reduced latency of execution.

***400-MHz System Bus*:** Through a physical signaling scheme of quad pumping the data transfers over a 100-MHz clocked system bus and a buffering scheme allowing for sustained 400-MHz data transfers, the Pentium 4 processor supports Intel's highest performance desktop system bus delivering 3.2GB of data per second in and out of the processor. This compares to 1.06GB/s delivered on the Pentium III processor's 133-MHz system bus.

***Advanced Dynamic Execution**:* The Advanced Dynamic Execution engine is a very deep, out-of-order speculative execution engine that keeps the execution units executing instructions. It does so by providing a very large window of instructions from which the execution units can choose. The large out-of-order instruction window allows the processor to avoid stalls that can occur while instructions are waiting for dependencies to resolve. One of the more common forms of stalls is waiting for data to be loaded from memory on a cache miss. This aspect is very important in high frequency designs, as the latency to main memory increases relative to the core frequency. The NetBurst micro-architecture can have up to 126 instructions in this window (in flight) vs. the P6 micro-architecture's much smaller window of 42 instructions.

The Advanced Dynamic Execution engine also delivers an enhanced branch prediction capability that allows the Pentium 4 processor to be more accurate in predicting program branches. This has the net effect of reducing the number of branch mis-predictions by about 33% over the P6 generation processor's branch prediction capability. It does this by implementing a 4KB branch target buffer that stores more detail on the history of past branches, as well as by implementing a more advanced branch prediction algorithm. This enhanced branch prediction capability is one of the key design elements that reduce the overall sensitivity of the NetBurst micro-architecture to the branch mis-prediction penalty.

Advanced Transfer Cache: The Level 2 Advanced Transfer Cache is 256KB in size and delivers a much higher data throughput channel between the Level 2 cache and the processor core. The Advanced Transfer Cache consists of a 256-bit (32-byte) interface that transfers data on each core clock. As a result, a 1.4-GHz Pentium 4 processor can deliver a data transfer rate of 44.8GB/s (32 bytes x 1 (data transfer per clock) x 1.4 GHz = 44.8GB/s). This compares to a transfer rate of 16GB/s on the Pentium III processor at 1 GHz and contributes to the Pentium 4 processor's ability to keep the high-frequency execution units executing instructions vs. sitting idle.

Streaming SIMD Extensions 2 (SSE2): With the introduction of SSE2, the NetBurst micro-architecture now extends the SIMD capabilities that MMX technology and SSE technology delivered by adding 144 new instructions that deliver 128-bit SIMD integer arithmetic operation and 128-bit SIMD Double-Precision Floating Point. These new instructions deliver the capability to 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.

Resultant Performance Expectations

The Pentium 4 processor shows immediate performance improvements across most existing software applications available today, with performance levels varying depending on the application category type and the application's tendency to execute instructions and instruction sequences that are optimally executed on the new micro-architecture.

Over time, as more applications are optimized, either specifically for the micro-architecture via assembler-level optimizations, or are revised using the latest NetBurst micro-architecture optimized compilers and libraries, we will continue to see even greater levels of performance scaling when the software runs on the Pentium 4 processor.

In summary, the Pentium 4 processor, based upon the NetBurst micro-architecture, delivers an acceleration of performance across the applications and usages where users will truly be able to experience and appreciate it. These usages include: 3D visualization, gaming, video, speech, and image photo processing, encryption, financial, engineering and Scientific Applications.