

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

INTRODUCTION

When the conversation of CPU speed or performance comes up, the clock speed of the microprocessor usually dominates the discussion. After all, that is what determines how quickly your CPU performs, right? The faster the clock speed, the faster and more powerful the CPU. Actually, this is a very common misconception. True, clock speed is a significant factor, but certainly not the determining factor. However, it can be easy to confuse the increase in MHZ directly to the increase in performance. One typically sees the advancement in microprocessor technology strictly in these terms.

This white paper will look at all the factors that determine the performance and speed of a CPU. As you will see, there are ten in all that I have identified. Certainly, there are others such as the mainboard chipset. But, we must keep this reasonable. However first, it will review the various components of the microprocessor that make up its operations. Really, there is more to them than meets the eye. The paper will then conclude with a historical review of the development of the microprocessor. This will serve to apply and help further understand just how the performance factors affect the performance of a CPU. To really understand how the microprocessor performs today, this perspective is necessary.

The specifications for most CPU's can be seen in a nutshell in the corresponding whitepaper, ***Microprocessor/CPU Specifications***. In order to achieve the proper comparative perspective and detail, this paper has been done in the form of a Microsoft Excel 7.0 for 95 spreadsheet.

Lets begin our venture with a brief look at the basic parts of the CPU.

BASIC PARTS OF THE CPU (3)

- Input/Output Unit
- Control Unit
- Arithmetic/Logic Unit

All three parts of the microprocessor interact together.. The ***I/O Unit*** is under the control of the ***Control Unit***, and the operation of the ***Control Unit*** may be determined by the ***Arithmetic/Logic Unit***.

- The ***I/O Unit*** determines the bus width of the microprocessor, which influences how quickly data and instructions can be moved in and out of the microprocessor.
- The ***Control Unit*** operates the microprocessor's internal clock, which determines the rate at which the chip operates.
- The ***Arithmetic/Logic Control Unit*** and its registers within determine how much data the microprocessor can operate on at one time.

The Input/Output Unit

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

- Links the microprocessor to the rest of the circuitry on the computer, passing along program instructions and data to the registers of the control unit and the arithmetic/logic unit.
- Each signal leaving the microprocessor goes through a signal buffer in the I/O unit that boosts its current capacity.
- There are two kinds of external connections to the Input/Output unit. These are called the **data bus** and the **address bus**.
- The **data bus** conveys data and instructions. The number of bits in the data bus directly influences how quickly it can move information. Data bus widths range from 8 bit, to 64 bit in the latest Pentium processors.
- The **address bus** sends memory locations of data or instructions to or from the microprocessor. The number of bits available on the address bus influences how much memory the microprocessor can address. A micro processor with 16 address lines can directly work with 2¹⁶ addresses, or 65,536 (64K) different memory locations (1x2¹⁶ binary). The different microprocessors used in various PC's span a range of address bus widths from 20 to 32 bits.

The Control Unit

- Consists of the Data Cache, the Code Cache, the Branch Predictor Unit, and the Instruction Prefetch Buffer and Decode Unit.
- The **Control Unit** is a clocked logic circuit that controls the operation of the entire chip.
- Follows the instructions contained in an external program and tells the **Arithmetic/Logic Unit** what to do.
- Receives instructions from the **I/O Unit**, translates them into a form that can be understood by the **Arithmetic/Logic Unit**, and keeps track of which step of the program is being executed.

The Arithmetic/Logic Unit

- The **Arithmetic/Logic Unit** handles all the decision making (the mathematical computations and logic functions) that are performed by the microprocessor. The unit takes the instructions decoded by the control unit and either carries them out directly or executes the appropriate microcode to modify the data contained in its registers. The results are passed back out of the microprocessor through the **I/O Unit**.
- There are two ways to speed up the execution of instructions (other than internal clock rate increase). These both involve reducing the number of internal steps required for execution. This can be done by making the microprocessor more complex so that steps

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

can be combined (CISC) or by making the instructions simpler so that fewer steps are required (RISC).

- Other ways of trimming the cycles required by programs is to operate on more than one instruction simultaneously. Three approaches to processing more instructions at once are **pipelining**, **Superscalar**, and **dynamic execution** architectures.

CPU SPEED AND PERFORMANCE FACTORS

There are eleven main factors that determine the speed and performance of a CPU. They are as follows:

1. **Internal Clock Speed** - Measured in MHZ (millions of cycles per second), this is sometimes referred to as the internal bus speed. The first IBM computer, the 8088 PC, had an internal clock speed of 4.77 MHZ. To prevent the microprocessor from reacting to invalid electrical signals, the chip waits for an indication that it has a valid command to carry out. This indication is provided by the system clock. The microprocessor checks the instructions given to it each time it receives a clock pulse. Except for the latest Superscalar, pipelined microprocessors, most chips carry out one instruction every clock cycle. Despite different frequencies inside (internal clock) and outside (external, or system bus clock), the system clock is used to synchronize logic operations. The only time that clock speed gives a reliable indication of relative performance is when you compare two identical chip designs that operate at different frequencies, i.e. 133 and 166 MHZ.
2. **Microcode Efficiency** - The number of steps required to multiply two numbers. Microcode contain the instructions that tell the processor what to do, and in what order. The unit of measure is in **clocks**. Table 1 below shows the variations in microcode efficiency between some of the INTEL X86 CPU's.

TABLE 1 MICROCODE EFFICIENCY		
CPU	Parameter	Clocks
8088	Add Two Numbers	12
80386	Add Two Numbers	6
80486	Add Two Numbers	2
Pentium	Add Two Numbers	1
Pentium Pro	Add Three Numbers	1

Microcode has made machines more complex, but it has also made them more flexible by allowing backward compatibility. This allows the data processing circuitry of the chip to be designed independently of the instructions it must carry out. In effect, the

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

microcode in a microprocessor is a secondary set of instructions that run invisibly inside the chip on a nanoprocessor (essentially a microprocessor within a microprocessor). However, microcodes makes computers and microprocessors more complicated. **RISC (Reduced Instruction Set Computers)** design computers have a reduced need for microcode and have reduced instructions sets. **CISC (Complex Instruction Set Computers)** computer require more complex instructions sets and more microcode.

3. **Word Size** - The largest number (bits) that can be worked on at one time by the CPU. Also commonly referred to as the **register size** or **register width**. To put this in perspective, it is necessary to fully understand the terms bit, nibble, byte, word, and double word. (Please refer to these definitions in **Glossary A** at the end of this paper). A register size can range from 16 bits to 32 bits. With the forthcoming Inter Merced CPU currently under design, 64 bit registers are right around the corner.

The register is a part of the CPU that acts as both memory and a workbench. It holds bit patterns until they can be worked on or output. In modern microprocessors, all registers are nearly equal in composition. Most modern RISC computers have 32 registers. The width of the registers has a substantial impact on the performance of the CPU. The more bits assigned to each register, the more information that can be processed in every microprocessor operation. The performance advantage of using wider registers depends on the software being run. For example, if a computer program tells the microprocessor to work on 16 bits at a time, the full power of the 32 bit registers will not be used.

4. **Floating Point Capability** - The Floating Point Unit (FPU), also referred to as math coprocessor, off loads complex math calculations and operations, freeing up the CPU for other tasks. The concern her is, does the CPU have a math coprocessor, and if so, what are its capabilities.
5. **Cache RAM** - Does the CPU have Level 1 (on board) or Level 2 (system board) cache RAM? If so, what is the size and the quality? Cache Ram enables the CPU to have both data and instructions readily available, so it can avoid going to the slower, and more distant, main memory. RAM cache sizes may range from 8K Megabytes, to 32K Megabytes found in the AMD K6 and Pentium II processors. Most cache RAM is of the static (SRAM) design, but it can be either asynchronous or synchronous. Synchronous SRAM is normally quick enough in nanoseconds (Ns) to synchronize operation with the CPU.
6. **Data Bus** - This should say data bus width, or width of the data bus. The number of data lines reflects how much information can be carried to the CPU at one time, or the largest number of bits that can enter the CPU at one time. Data bus widths range from 8bit, to 64 bit found in the Pentium and Pentium Pro processors.

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

7. **Address Bus** - The size or width of the address bus or address path. This determines how much memory the CPU can address or use. The larger the address, the more memory the CPU can address. Address bus widths range from 20 to 32 bit in the most modern microprocessors. Table 2 shows how the address bus width affects CPU memory addressing capabilities. Widths may range from 20 bit to 32 bit. In the not to future, 64 bit address bus's will be a reality.

8. **Pipeline, Superscalar, and Dynamic Execution Processes**

Pipelining enables a microprocessor to read an instruction, start to process it, and then, before finishing with the first instruction, read another instruction. Because every instruction requires several steps each in a different part of the chip, several instructions can be worked on at once, and passed along through the chip like a bucket brigade. Intel's Pentium has four levels of pipelining, so up to four different instructions may be undergoing different phases of execution at one time inside the chip.

In general, the more stages of pipelining, the greater the acceleration. But, the bigger the pipelining, the more time wasted. The waste resulting from branching begins to outweigh the benefits of bigger pipelines in the vicinity of five stages. The most powerful microprocessors use a technology called branch prediction logic, where the microprocessor makes its best guess at which branch a program will take as it is filling up the pipeline.

Superscalar systems provide two or more execution paths for programs, and thus, can process two or more program parts simultaneously. The Pentium has two parallel pipelined execution paths.

Dynamic Execution involves optimally adjusting instruction execution by predicting program flow, analyzing the program's dataflow graph to choose the best order to execute the instructions, then speculatively executing instructions in the preferred order.

TABLE 2 CPU ADDRESSABLE MEMORY	
Address Bus Width (Bits)	Addressable Memory (MB)
20	1
24	16
32	4 GB
64	1.84467E+19

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

9. *Main (Core) Memory Type & Timing*

When the CPU is unable to find the instructions or data it wants in the RAM cache, it then must go to the main memory to find what it needs to proceed onward. The speed at which the CPU can access this information has a large affect on its performance. Hence, the type of memory will determine the memory's ability to send data and instructions to the CPU expeditiously. Important factors here are the memory timing sequences, and the speed of the memory in nanoseconds. Table 3 summarizes the various types of DRAM and their performance factors.

**TABLE 3
MEMORY TIMING & SPEED**

RAM Type	Common Name	Timing	Speed	Bandwidth	NOTES
DRAM	Dynamic RAM				One transistor & one capacitor per memory location
FPM	Fast Page Mode DRAM	5-3-3-3	70-60	32 & 64 bit	60 Ns must be used with 66 MHz system bus
EDO	Extended Data Out DRAM	5-2-2-2	70-60-50	32 & 64 bit	50 Ns can be used with Triton HX or VX chipsets
BEDO	Burst Extended Data Out	5-1-1-1		32 & 64 bit	Internal address counter; limited to 66 MHz bus
SDRAM	Synchronous DRAM	5-1-1-1	10	32 & 64 bit	Supports a 100 MHz system bus speed; VX chipset
RDRAM	Rambus DRAM	5-1-1-1	3.75	8 bit (16)	Capable of 533 MHz clock. Two channel configuration
SRAM	Static RAM		12-8.5		Two transistors per memory location (off or on)
ASRAM	Asynchronous SRAM		20-15-12		Not fast enough for synchronous CPU access
SSRAM	Synchronous Burst SRAM	2-1-1-1			Performance declines with < 66 MHz system bus
PB SRAM	Pipelined Burst SRAM	3-1-1-1	8-4.5		Good for system bus speeds of 75 to 133 MHz
Clock (MHz)	Cycle (NS)				
25	40				
33	30				
50	20				
66	15				
100	10				
200	5				

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

11. *Density*

Density refers to the transistor technology involved in the design and manufacture of the microprocessor. Considering that CPU's consist of millions of tiny transistors, their size (in microns) and spacing (also in microns) has much to do with the performance of the device. Smaller transistors can be more tightly packed. Consequently, they can be spaced closer together. The closer the spacing, the less distance electrical signals have to travel. Thus, the relative speed of the CPU increases. Table 4 shows the relative change in transistor technology over the years. The current standard is .35 micron. However, .28 micron, and even .25 micron chips are currently under design and being tested.

CPU	Technology (microns)
8008	10
8080	6
8088	3
80286	1.5
80386	1
80486	.8
Pentium	.35
Pentium Pro	.35
Pentium II	.28

CPU EVOLUTION

To better understand how these ten performance factors apply in the real world, we will take a short look at how the microprocessor has developed historically. I will attempt to keep this in very basic terms. However, when the Pentium, Pentium Pro, and AMD K6 processors are reviewed, the discussion may become a little more technical. With these chips, the culmination and integration of performance enhancing technology will be seen and realized. It must be kept in mind that these chips are of an advanced and complex architecture. It would be impossible to describe how they work and how they perform in simple language. For this reason, a ***Glossary of Microprocessor Terminology*** (Appendix A) has been included at the end of this paper.

Intel 4004/8080 Family

- Introduced in 1971
- 4 bit bus and registers
- Designed to run at 108 KHz
- 8080 had the register and data bus width doubled to 8 bits

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

Intel 8086 Family

- Introduced in 1978
- Doubled the register size to 16 bits
- Doubled the data bus to 16 bits
- 20 bit address bus
- Designed to run at 4.77 MHz

INTEL 286 Family

- Introduced in 1984
- Full 16 bit data bus and registers
- Designed to run from 6 to 10 MHz
- Address lines increased to 24
- Protected mode memory available

INTEL 386 Family

- Introduced in 1985
- Full 32 bit processor (address bus, data bus, and registers)
- Protected and virtual memory with enhanced MMU
- Designed to run at 16 to 33 MHz
- Floating Point Unit optional
- Added 16 byte L1 cache

INTEL 486 Family

- Introduced in 1989
- Full 32 bit
- Improved microcode efficiency with reduced instruction set execution time
- Piplined internal structure
- Designed to run at 33 MHz
- Utilized clock doubling and clock tripling
- 8K L1 cache
- Built in synchronous math coprocessor
- VL bus introduced

Intel Pentium

- Introduced in 1993
- Doubled data bus to 64 bits
- Address bus increased to 36 bits
- 32 bit registers (word size processing)
- Utilizes pipelining, and Superscalar architecture. Consequently, the Pentium is essential two 486 CPU together, with only one half having a math coprocessor
- L2 RAM cache added

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

Here is how the Pentium works in the simplest terms:

The **bus interface unit** (BIU) sends and receives both data and coded instructions from the computer's random access memory (RAM), or L2 (Level 2) cache. The processor is connected to RAM by the PC's motherboard circuits, which are known as the bus. In this case, we are speaking of the system, or external bus. Data moves into the processor 64 bits at a time.

The bus interface unit sends and receives data and code along two separate paths that can each handle 64 bits at a time. One path leads to an 8K storage unit, or cache, used for data. The other path leads to an identical cache used only for code that tells the processor what to do with the data. The data and code stay in the caches until other parts of the microprocessor need them.

While the code is waiting in its cache, another part of the CPU called the **branch prediction unit** inspects the instructions to determine which of the two **arithmetic logic units (ALUs)** can handle them more efficiently. This inspection ensures that one of the ALUs isn't waiting while the other ALU finishes execution another instruction.

The **instruction prefetch buffer** retrieves the code tagged by the branch prediction unit and the decode unit translates the software code into the type of instructions that the ALUs can understand.

If any floating point numbers, or numbers with decimal fractions need processing, they are passed to the specialized processor called the **FPU**, or **Floating Point Unit**.

Within the execution unit, two arithmetic logic units process all data consisting only of integers. Each of the ALUs receives instructions up to 32 bits at a time from the instruction decode unit. Each ALU processes its own instructions simultaneously using data moved from the data cache to the electronic scratch pad known as a register.

The two arithmetic logic units and the floating point unit send the results of their processing to the data cache. The data cache sends the results to the bus interface unit, which, in turn, sends the results to RAM, or the L2 cache.

Intel Pentium Pro Processor

The new approach used by the Pentium Pro processor removes the constraint of linear instruction sequencing between the traditional **fetch** and **execute** phases, and opens up a wide instruction window using an instruction pool. This approach allows the **execute** phase of the Pentium Pro processor to have much more visibility into the program's instruction stream so that better scheduling may take place. It requires the instruction **fetch/decode** phase of the Pentium Pro processor to be much more intelligent in terms of predicting program flow. Optimized scheduling requires the fundamental **execute** phase to be replaced by decoupled

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

dispatch/execute and *retire* phases. This allows instructions to be started in any order but always be completed in the original program order.

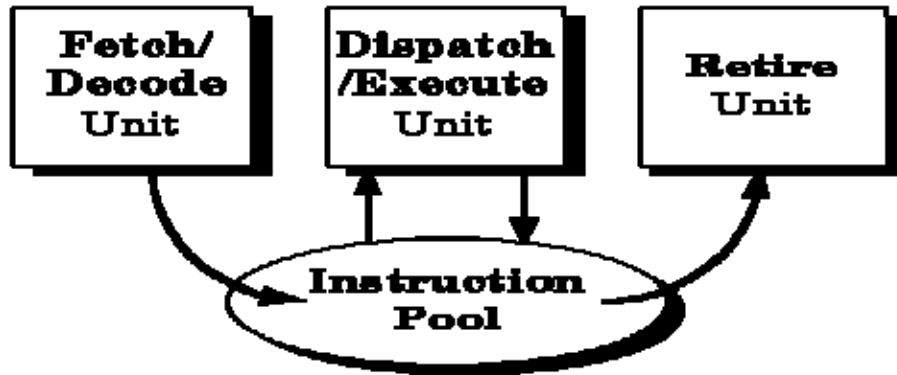


Figure 1. The P6 is implemented as three independent engines that communicate using an instruction pool.

The Pentium Pro processor is implemented as three independent engines coupled with an instruction pool as shown in Figure 1. It is important to note why this three independent-engine approach was taken. One of the fundamental facts of today's microprocessor implementations is that most CPU cores are not fully utilized. Consider the code fragment (courtesy of Intel) in Figure 2 below.

The first instruction in this example is a load of r1 that, at run time, causes a cache miss. A traditional CPU core must wait for its bus interface unit to read this data from main memory and return it before moving on to instruction 2. The CPU stalls while waiting for this data and is thus being under-utilized.

```
r1 <= mem [r0]           /* Instruction 1 */  
r2 <= r1 + r2           /* Instruction 2 */  
r5 <= r5 + 1           /* Instruction 3 */  
r6 <= r6 - r3           /* Instruction 4 */
```

Figure 2. A typical code fragment.

While CPU speeds have increased at least 10 times over the past 10 years, the speed of main memory devices has only increased by 60 percent. This increasing memory latency, relative to the CPU core speed, is a fundamental problem that the Pentium Pro processor solves. The Pentium Pro processor is designed from an overall system implementation perspective which will allow higher performance systems to be designed with cheaper memory subsystem designs.

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

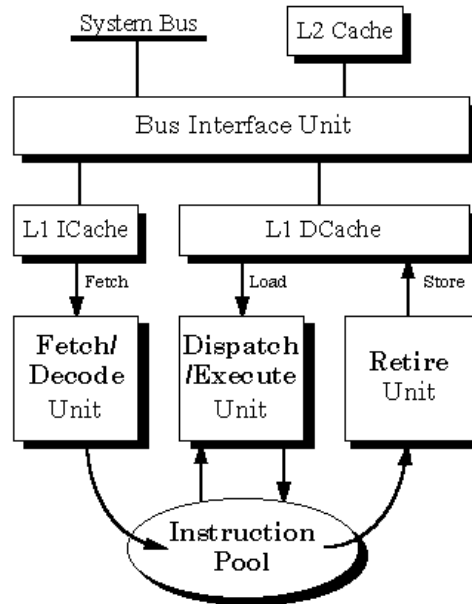


Figure 3. The three core engines interface with the memory subsystem using SK/SK unified caches.

To avoid this memory latency problem the Pentium Pro processor looks-ahead into its instruction pool at subsequent instructions and will do useful work rather than be stalled. In the example, instruction 2 is not executable since it depends upon the result of instruction 1. But, both instructions 3 and 4 are executable. The Pentium Pro processor speculatively executes instructions 3 and 4. Since it cannot commit the results of this execution, the results are instead stored back in the instruction pool awaiting in-order retirement. The core executes instructions depending upon their readiness to execute and not on their original program order. This approach has the side effect that instructions are typically executed out-of-order.

The cache miss on instruction 1 will take many internal clocks, so the Pentium Pro processor core continues to look ahead for other instructions that could be speculatively executed. The fetch/decode unit must correctly predict if the **dispatch/execute** unit is to do useful work. The register set and results are only committed to permanent machine state when it removes completed instructions from the pool in original program order.

Dynamic Execution Technology can be summarized as optimally adjusting instruction execution by predicting program flow, analyzing the program's dataflow graph to choose the best order to execute the instructions, and having the ability to speculatively execute instructions in the preferred order. The Pentium Pro processor dynamically adjusts its work, as defined by the incoming instruction stream to minimize overall execution time. The Pentium Pro processor has the unique combination of improved branched prediction (the offering of the core many instructions), data flow analysis (choosing the most efficient order), and speculative execution (executing instructions in the preferred order) that enables it to

LEGACY CPU PERFORMANCE FACTORS

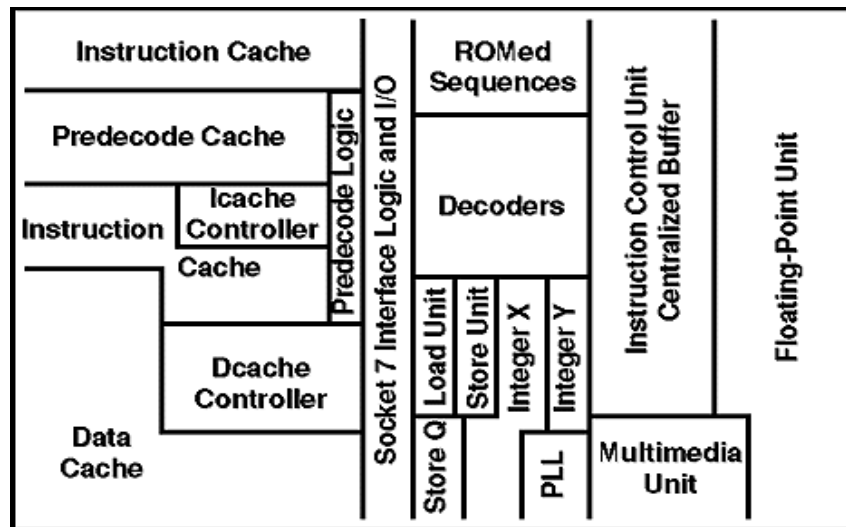
By Mark E. Donaldson

deliver a 33% performance boost over the Pentium processor. This unique combination is called Dynamic Execution, and similar in impact as Superscalar was to the Pentium) P5 generation of processors.

AMD K6

The K6 is in large part comparable to the Pentium Pro in terms of design elements, but it has trappings of the Pentium processor as well. Like the Pentium Pro, the K6 can perform speculative and out-of-order execution of instructions. The big difference between the two chips is in cache architecture. Each chip has an on-chip L-1 cache, but the K6's is 64 KB instead of the Pentium Pro's 16 KB. The Pentium Pro does have an advantage in that its L-2 cache is part of the processor package and runs at the chip's core frequency. A large L-1 cache yields a significant performance boost, as can be seen when comparing the performance of Pentium and MMX Pentium processors on applications that haven't been optimized for the MMX instruction set.

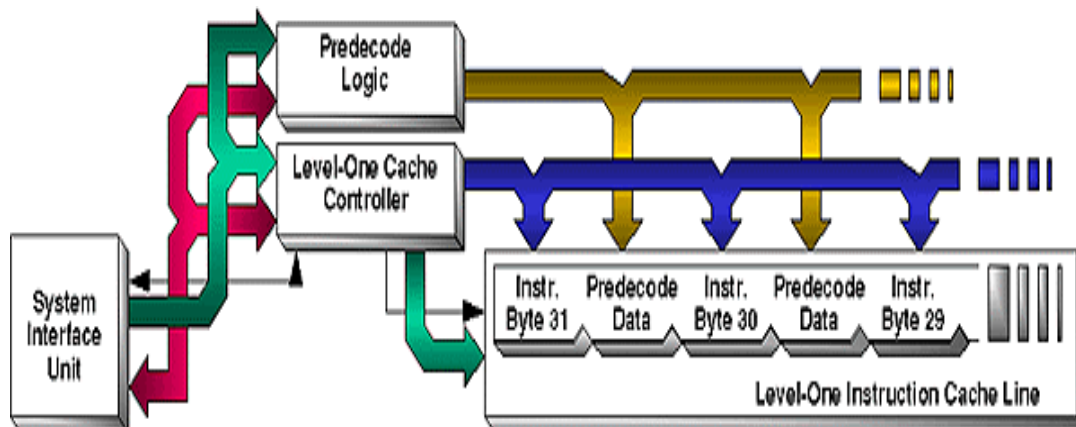
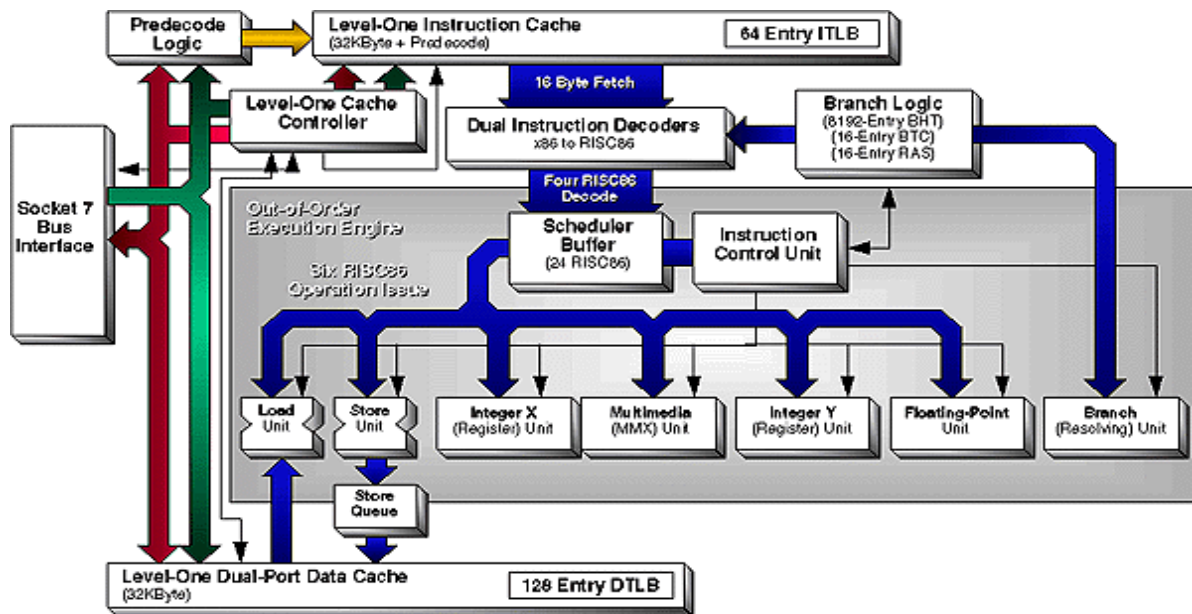
The Pentium-like element of the K6 processor is its Socket 7 pinout, which means the K6 can use the same processor socket as the Pentium processor. This means systems manufacturers will be able to easily integrate the chip without making major modifications to their motherboard design.



The design, architecture, and flow of this advanced chip is best explained by viewing the diagrams below.

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson



Intel Pentium II

Pentium II evolved from the existing Pentium Pro. The Pentium II CPU provides better 16-bit and 32-bit performance for Windows 95 than its predecessor, as well as MMX instructions and an internal cache that is doubled from 16K to 32K. Intel also placed the secondary cache inside the Pentium II's **Single Edge Contact (SEC)** cartridge to help take advantage of core CPU clock speeds of 233-MHz and higher. Like the K6, the Pentium II is based on a .35-micron manufacturing process but will be moved to a tighter, .25 micron process. The .25-micron CPU, with the code name **Deschutes**, is expected near the end of 1997.

The most important change the Pentium II will offer is a new connection to the motherboard. Since the days of the 386, there have always been specially designed CPU sockets on the motherboard where the microprocessor chips **pin grid array** (PGA) design used to fit right in. Ya know, it's that open standards stuff. A short time after the 486 was released, the

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

motherboards started being equipped with the so called **Zero Insertion Force**, or **ZIF** sockets, which made it quite simple to change the CPU. The Pentium CPU, as well as the Pentium Pro CPU, also plugged into these ZIF sockets, and hence there is no problem when changing from one CPU to another.

After the P54C was released, the Socket 5 was first used. Later, the Socket 7 specification was used for Pentium CPU's. This socket was not licensed by Intel. All AMD and Cyrix had to do was apply to the Socket 7 specification, and one could easily swap an Intel Pentium CPU for an AMD or Cyrix chip. The Socket 8 specification for the Pentium Pro should be the first one licensed by Intel. Competitors won't be able to build chips to fit in this socket. Something important to know is the upcoming Pentium II will **NOT** fit in a normal PGA CPU ZIF socket.

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

APPENDIX A

GLOSSARY OF MICROPROCESSOR TERMINOLOGY

Architectural State - The value of registers, flags and memory as viewed by the programmer.

Associative Memory - A table that is accessed not via an explicit index, but by the data it contains. If no entries of the associative memory match the input data, a "miss" signal is asserted. If any entries of the associative memory match the input data, the associative memory indicates the match, and produces any related data that was stored with that entry. This is often termed Content Addressable Memory.

Bit - A single binary digit, either a 0 or a 1.

Branch Prediction - Pipelined machines must fetch the next instruction before they have completely executed the previous instruction. If the previous instruction was a branch, then the next-instruction fetch could have been from the wrong place. Branch prediction is a technique that attempts to infer the proper next instruction address, knowing only the current one, typically using an associative memory called a BTB.

Branch Recovery - When a branch is mispredicted, the speculative state of the machine must be flushed and fetching restarted from the correct target address. We call this activity branch recovery.

Branch Target Buffer - A small (typically 128-512 entry) associative memory that watches the I-Cache index and tries to predict which I-Cache index should be accessed next, based on branch history. Optimizing the actual algorithm used in retaining the history of each entry is an area of ongoing research. Pentium and Pentium Pro processors uses a variant of Yeh's algorithm (IEEE Micro-24 conference proceedings-1991).

Bus Interface Unit - A partially ordered unit responsible for connecting the three internal units to the real world. The bus interface unit communicates directly with the L2 cache supporting up to four concurrent cache accesses. The bus interface unit also controls a transaction bus, with MESI snooping protocol, to system memory.

Content Addressable Memory - Synonym for Associative Memory. A table that is accessed not via an explicit index, but by the data it contains. If no entries of the associative memory match the input data, a "miss" signal is asserted. If any entries of the associative memory match the input data, the associative memory indicates the match, and produces any related data that was stored with that entry.

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

Dispatch/Execute Unit - An out-of-order unit that accepts the dataflow stream, schedules execution of the uops subject to data dependencies and resource availability, and temporarily stores the results of these speculative executions.

Double Word (dword) - Four bytes or 32 bits. It takes a full 32 bits to fill the registers of all modern CPU's such as the Pentium, Pentium Pro, the AMD K5 and K6, and the Cyrix 686 series.

Dynamic Execution - Dynamic Execution involves optimally adjusting instruction execution by predicting program flow, analyzing the program's dataflow graph to choose the best order to execute the instructions, then speculatively executing instructions in the preferred order.

Fetch/Decode Unit - An in-order unit that takes as input the user program instruction stream from the instruction cache, and decodes them into a series of micro-operations (**uops**) that represent the dataflow of that instruction stream. The program pre-fetch is itself speculative.

I-Cache (Instruction Cache) - A fast local memory that holds the instructions to be executed. When a program tries to access an instruction that is not yet (or no longer) in the cache, the CPU must wait until hardware fetches the desired instructions from another cache (or memory itself) downstream. These stalls in the fetch/decode unit of the Pentium Pro processor are typically overlapped by the other units that are processing independently.

Instruction Pool - A metaphor used to describe the mechanism used by three independent Pentium Pro processor units to communicate. The pool is implemented as a CAM called the ReOrder Buffer (ROB).

L2 Cache - Caches exist in a "memory hierarchy." There is a small but very fast L1 cache; if that misses, then the access is passed on to the bigger but slower L2 cache, and if that misses, the access goes to main memory (or L3 cache if the system has one).

Nibble - Four bits, or a half byte.

Pipelining - A microarchitecture design technique that divides the execution of an instruction into sequential steps, using different microarchitectural resources at each step. Pipelined machines have multiple IA instructions executing at the same time, but at different stages in the machine.

RAT (Register Alias Table) - Renames programmer visible register references to internal physical registers. This mapping is done at run time.

Reservation Station - A generalized mechanism where uops wait for their dependent component parts. Once a uop has all of its operands, the uop is dispatched to a resource unit for execution.

LEGACY CPU PERFORMANCE FACTORS

By Mark E. Donaldson

Retire Unit - An in-order unit that knows how and when to commit the temporary, speculative results to permanent architectural state.

Resource Unit - The Pentium Pro processor has multiple resources that are scheduled by the Reservation Station: they include 2 integer units, a full floating point arithmetic unit, a floating point multiplier, divider, and shifter, and two address generation units.

Retirement - A generalized mechanism that removes a completed uop from the ROB and commits its state to whatever permanent architectural state was designated by the Intel architecture instruction.

ROB (Re-Order Buffer) - The Pentium Pro processor functional unit where initial uops wait, speculative results are collected, and then are retired.

Speculative Execution - A generalized mechanism that permits instructions to be started "early," i.e., ahead of their normal execution sequence. Results of this speculation are stored temporarily (in the ROB) since they may be discarded due to a change in program flow.

Store Buffer - A queue that receives write requests from the CPU and sends them to the memory subsystem. This store buffer is snooped by pending loads.

Superscalar - The ability to process more than one instruction per clock. The Pentium processor has two execution pipes (U and V) so it is Superscalar level 2. The Pentium Pro processor can dispatch and retire 3 instructions per clock so it is Superscalar level 3.

UOP - A micro operation. The three decoders translate Intel architecture instructions into fixed length uops that are easier to schedule by the dispatch/execute unit. Most IA instructions translate to single uops, some need up to four, and the complex instructions (e.g., Enter, Leave) need microcode support.

Word - Two bytes, or 16 bits.