ECE 3140/CS 3420
Computer Organization
Spring 2009

Intel IA-32

Announcements

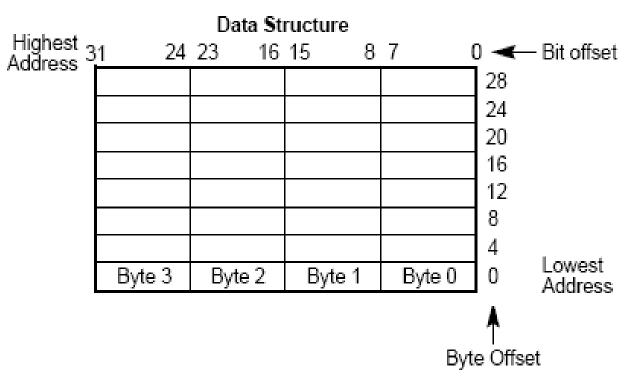
- Project 4b
  - Due Fri, April 23, 10:00am
- Project 4c
  - Due Sat, May 2, 10:00pm
- Guest Lecturer
  - Tue. Apr 28, in class
  - Dave Albonesi
  - MIPS R10K

The Intel IA-32 Legacy

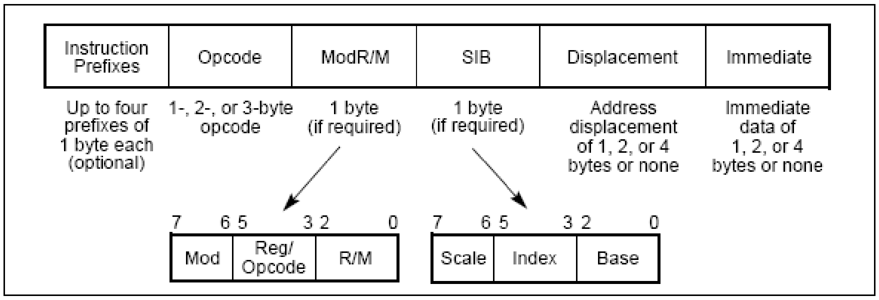
- "Object code created for processors released as early as 1978 still executes on the latest processors in the IA-32 architecture family."
  – ***IA-32 Intel® Architecture Software Developer's Manual, Volume 1: Basic Architecture***
- "[It's] checkered history has led to an architecture that is difficult to explain and impossible to love."
  – ***Patterson and Henessy***

In the beginning…

- IA-32 processors are "little endian" machines
- The bytes of a word are numbered starting from the least significant byte



IA-32 Instruction Format



- IA-32 Instructions can be from 1 to 17 bytes in length
- All IA-32 processors prefetch instructions
  - Different processor generations use different prefetching mechanisms

ECE3140/CS3420 Cornell University

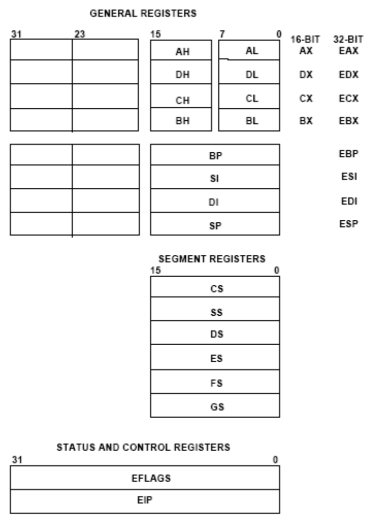
Instruction Prefixes

- LOCK prefix
  - Forces an operation that ensures exclusive use of shared memory in a multiprocessor environment
- Repeat prefixes
  - Cause an instruction to be repeated for each element of a string.
  - Use these prefixes only with string instructions
- Branch hint prefixes
  - Allow a program to give a hint to the processor about the most likely code path for a branch
  - Use these prefixes only with conditional branch instructions (Jcc)
- Operand-size override prefix
  - Allows a program to switch between 16- and 32-bit operand sizes
- Address-size override prefix
  - Allows programs to switch between 16- and 32-bit addressing.

ECE3140/CS3420 Cornell University

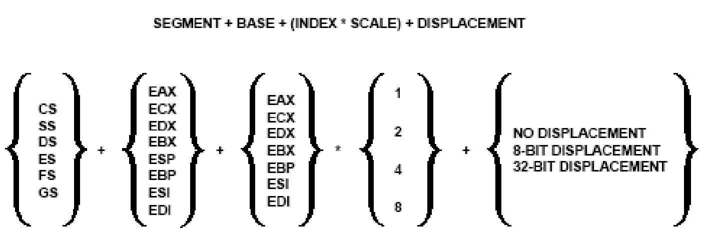
Application Register Set

- Eight 16-bit registers (AX, BX, CX, DX, BP, SP, SI, and DI)
  - Derived from 8086 names Low 16 bits of corresponding general registers
- Eight 8-bit registers (AH, BH, CH, and DH and AL, BL, CL, and DL)
  - High and Low bytes of corresponding 16-bit registers
- Eight 32-bit general registers (EAX, EBX, ECX, EDX, EBP, ESP, ESI, and EDI)

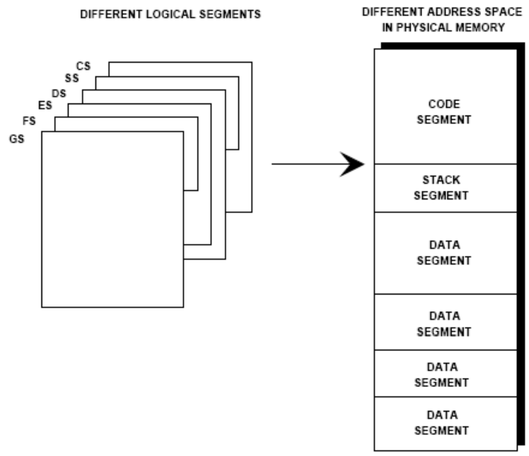


Effective Address Calculation

- The modR/M byte provides the most flexible (and common) form of addressing.
- The offset which results from adding some combination of these components is called an *effective address*.
- Each of these components can have either a positive or negative value, with the exception of the scaling factor.

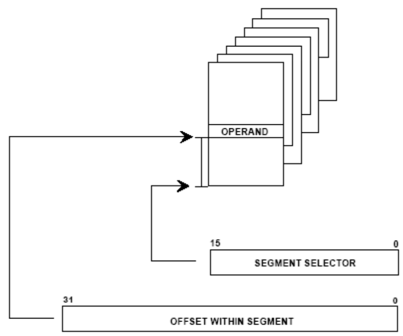


Segmented Memory Model

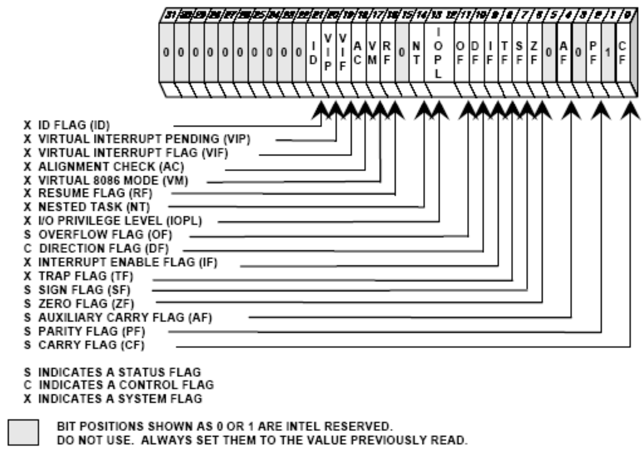


Segmented Addressing

- The logical address space
  - 16,383 segments of up to 4 gigabytes each
  - Up to 2^{46} bytes (64 terabytes).
  - Maps onto the physical address space
- A pointer into a segmented address space consists of two parts.
  - A 16-bit *segment selector* contains the *base address* of the segment.
  - A 32-bit *offset* within a segment



EFLAGS Register



16-bit Processors and Segmentation (1978)

- The IA-32 architecture family was preceded by 16-bit processors, the 8086 and 8088
  - The 8086 has 16-bit registers and a 16-bit external data bus, with 20-bit addressing giving a 1-MByte address space
  - The 8088 is similar to the 8086 except it has an 8-bit external data bus
- The 8086/8088 introduced segmentation to the IA-32 architecture
  - A 16-bit segment register contains a pointer to a memory segment of up to 64 KBytes
  - Using four segment registers at a time, 8086/8088 processors are able to address up to 256 KBytes without switching between segments
  - The 20-bit addresses that can be formed using a segment register and an additional 16-bit pointer provide a total address range of 1 MByte

The Intel 286 Processor (1982)

- The Intel 286 processor introduced protected mode operation into the IA-32 architecture.
  - Protected mode uses the segment register content as selectors or pointers into descriptor tables.
- Descriptors provide 24-bit base addresses with a physical memory size of up to 16 Mbytes
  - Support for virtual memory management on a segment swapping basis
- Protection mechanisms include:
  - Segment limit checking
  - Read-only and execute-only segment options
  - Four privilege levels.

ECE3140/CS3420 Cornell University

The Intel 386 Processor (1985)

- The Intel386 processor was the first 32-bit processor in the IA-32 architecture family
- Eight 32-bit registers
  - The lower half of each 32-bit Intel386 register retains the properties of the 16-bit registers of earlier generations, permitting backward compatibility.
- Virtual-8086 mode that allows for even greater efficiency when executing programs created for 8086/8088 processors.
- In addition, the Intel386 processor has support for:
  - A 32-bit address bus that supports up to 4-GBytes of physical memory
  - A segmented-memory model and a flat memory model
  - Paging, with a fixed 4-KByte page size providing a method for virtual memory management
  - Support for parallel stages

ECE3140/CS3420 Cornell University

The Intel 486 Processor (1989)

- The Intel486 processor expanded instruction decode and execution units into five pipelined stages.
  - Up to five instructions in different stages of execution.
- In addition, the processor added:
  - An 8-KByte on-chip first-level cache that increased the percent of instructions that could execute at the scalar rate of one per clock
  - An integrated x87 FPU
  - Power saving and system management capabilities

The Intel Pentium Processor (1993)

- Superscalar processor (two execution pipelines, known as u and v, together can execute two instructions per clock).
  - Branch prediction with an on-chip branch table
  - 8 KByte on-chip first-level code cache and 8 KByte on-chip first-level data cache
  - More efficient virtual-8086 mode; allow s 4-MByte as well as 4-KByte pages
  - Internal data paths of 128 and 256 bits; burstable external 64 bit data bus
  - Multiple processors support
- Pentium MMX technology
  - Single-instruction, multipledata execution model
  - Perform s parallel computations on packed integer datacontained in 64-bit registers

The P6 Family of Processors (1995-1999)

- The Intel Pentium Pro processor is three-way superscalar.
  - Decodes, dispatches, and completes execution of (retire) three instructions per clock cycle.
  - Dynamic execution (microdata flow analysis, out-of-order execution, superior branch prediction, and speculative execution)
  - 256 KByte Level 2 cache
- Pentium II processor added Intel MMX Technology
  - MMX SIMD execution model on packed integer data contained in 64-bit registers
  - Level l data and instruction caches 16 KBytes each,
  - Level 2 cache sizes of 256 KBytes, 512 KBytes, and 1 MByte
  - Multiple low-power states
- Pentium III processor added Streaming SIMD Extensions (SSE)
  - SSE extensions provide a new set of 128-bit registers and the ability to perform SIMD operations on packed single-precision floating-point values.

Pentium 4 (2000) and Pentium 4 HT (2004)

- The Intel Pentium 4 processor is based on Intel NetBurst® microarchitecture
- Introduced the following major feature sets:
  - Streaming SIMD Extensions 2 (SSE2)
  - Streaming SIMD Extensions 3 (SSE3)
- Hyper-Threading (HT) Technology enables a single physical processor to execute two or more separate code streams (threads) concurrently
  - Support for two or more logical processors, each of which has its own state.
  - Each logical processor consists of a full set of IA-32 data registers, segment registers, control registers, debug registers and most of the MSRs.

The Intel Xeon Processor (2001-2004)

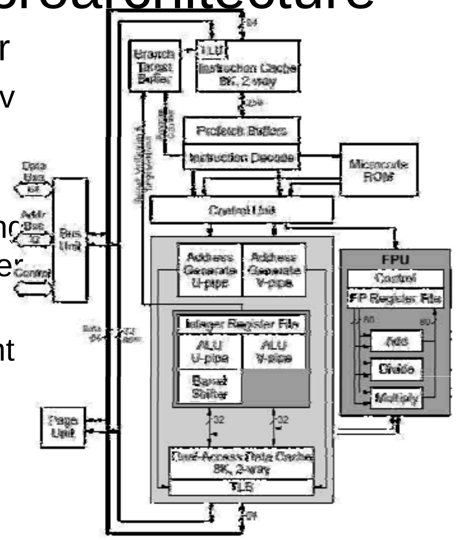
- The Intel Xeon processor is also based on the Intel NetBurst microarchitecture;
- Designed for use in multi-processor server systems and high-performance workstations.
- The Intel Xeon processor MP introduced support for Hyper-Threading Technology

The Intel Pentium M Processor (2003-2004)

- High performance, low power mobile processor family with micro-architectural enhancements
- This family is designed to
  - Extend battery life
  - Enable extended mobility, ultra thin form-factors, and integrated wireless networking.

Pentium Microarchitecture

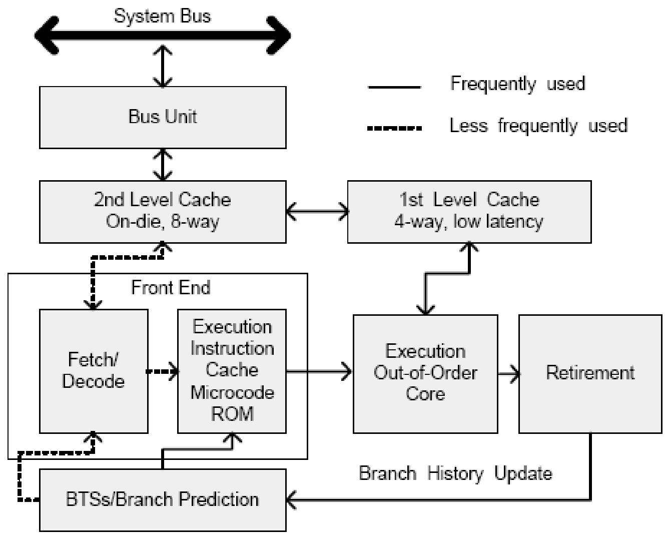
- Superscalar processor
  - Dual pipelines, u and v
  - Separate instruction and data caches
  - Branch prediction using the branch target buffer (BTB),
  - Pipelined floating-point unit
  - 64-bit external data bus.



The P6 Family Microarchitecture

- Three-way superscalar, pipelined microarchitecture
- The processor is able on average to decode, dispatch, and complete execution of (retire) three instructions per clock cycle.
- The P6 processor family uses a decoupled, 12-stage superpipeline that supports out-of-order instruction execution.

The P6 Family Microarchitecture



P6 Dynamic execution

- Deep branch prediction
  - Decodes instructions beyond branches to keep the instruction pipeline full
  - Highly optimized branch prediction algorithms
- Dynamic data flow analysis
  - Real-time analysis of the flow of data through the processor
  - Detect opportunities for out-of-order instruction execution
  - Executes instructions in the order that best optimizes the use of the processor's multiple execution unit
  - Maintains data integrity
- Speculative execution
  - Execution of instructions that lie beyond an unresolved conditional branch
  - Dispatch and execution of instructions is decoupled from the commitment of results.

P6 Dynamic execution (cont.)

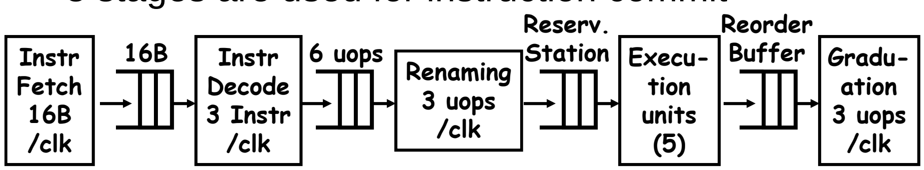
- The out-of-order execution core
  - Data-flow analysis to execute all available instructions in the instruction pool and store the results in temporary registers.
- The retirement unit searches the instruction pool for completed instructions
  - I.e., no data dependencies with other instructions or unresolved branch predictions
- The retirement unit commits the results of completed instructions
  - Stores results to memory and/or the IA-32 registers in the order they were originally issued
  - Retires the instructions from the instruction pool

Dynamic Scheduling in P6

- P6 doesn't pipeline 80x86 instructions
- P6 decode unit translates the Intel instructions into 72-bit micro-operations (~ MIPS)
- Sends micro-operations to reorder buffer & reservation stations
- Many instructions translate to 1 to 4 micro-operations
- Complex 80x86 instructions are executed by a conventional microprogram (8K x 72 bits) that issues long sequences of micro-operations
- 14 clocks in total pipeline (~ 3 state machines)

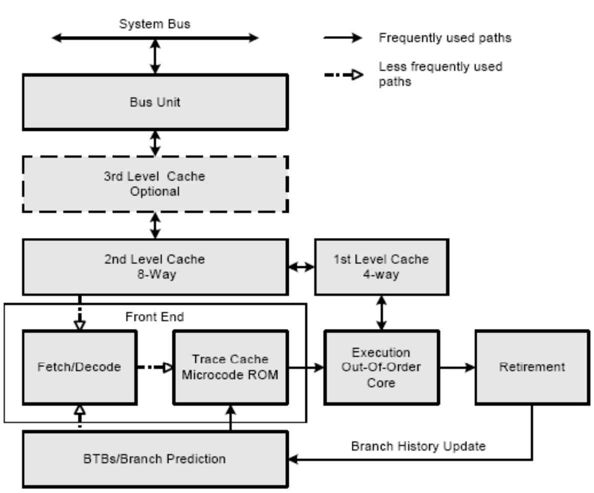
P6 Pipeline

- 8 stages are used for in-order instruction fetch, decode, and issue
  - Takes 1 clock cycle to determine length of 80x86 instructions + 2 more to create the uops
- 3 stages are used for out-of-order execution in one of 5 separate functional units
- 3 stages are used for instruction commit



Cornell University

The Intel NetBurst Microarchitecture



Cornell University

The Front End Pipeline

- Supplies instructions in program order to the out-of-order execution core
  - Prefetches IA-32 instructions that are likely to be executed
  - Fetches instructions that have not already been prefetched
  - Decodes IA-32 instructions into micro-operations
  - Generates microcode for complex instructions and special-purpose code
  - Delivers decoded instructions from the execution trace cache
  - Predicts branches using highly advanced algorithm

Out-Of-Order Execution Core

- The out-of-order execution core
  - Reorders instructions so that if one μop is delayed, other μops may proceed around it
  - Several buffers to smooth the flow of μops
- The core can dispatch up to six μops per cycle (this exceeds trace cache and retirement μop bandwidth)
  - Most pipelines can start executing a new μop every cycle, so several instructions can be in flight at a time for each pipeline.
- A number of arithmetic logical unit (ALU) instructions can start at two per cycle; many floating-point instructions can start once every two cycles

Retirement Unit

- The retirement unit
  - Receives the results of the executed μops from the out-of-order execution core and
  - Processes the results so that the architectural state updates according to the original program order
- When a μop completes and writes its result, it is retired
  - Up to three μops may be retired per cycle
  - The Reorder Buffer (ROB) buffers completed μops, updates the architectural state in order, and manages the ordering of exceptions.
  - The retirement section also keeps track of branches and sends updated branch target information to the BTB. The BTB then purges pre-fetched traces that are no longer needed

Pentium, Pentium Pro, Pentium 4 Pipeline

Prefetch	Decode	Decode	Execute	Write-back

P5 Microarchitecture

Fetch	Fetch	Decode	Decode	Decode	Rename	ROB Rd	Rdy/Sch	Dispatch	Execute

P6 Microarchitecture

TC Nxt IP	TC Fetch	Drive	Alloc	Rename	Queue	Schedule

Schedule	Schedule	Dispatch	Dispatch	Reg File	Reg File	Execute	Flags	Branch Ck	Drive

NetBurst Microarchitecture

- Pentium (P5) = 5 stages
- Pentium Pro, II, III (P6) = 10 stages
- Pentium 4 (NetBurst) = 20 stages