

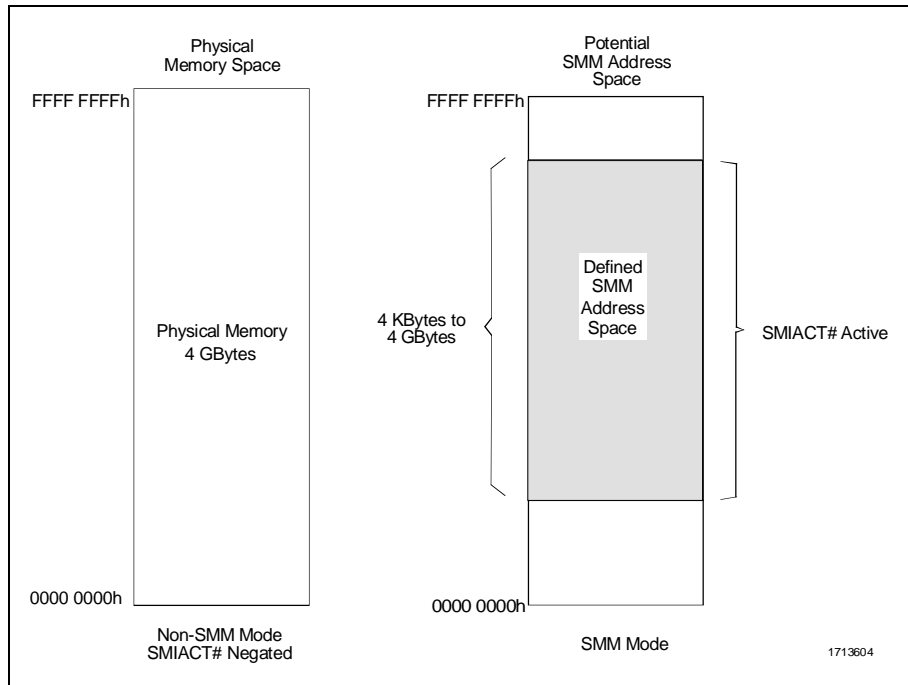
2.9 System Management Mode

System Management Mode (SMM) provides an additional interrupt which can be used for system power management or software transparent emulation of I/O peripherals. SMM is entered using the System Management Interrupt (SMI#) that has a higher priority than any other interrupt, including NMI. An SMI interrupt can also be triggered via software using an SMINT instruction. After an SMI interrupt, portions of the CPU state are automatically

saved, SMM is entered, and program execution begins at the base of SMM address space (Figure 2-37). Running in SMM address space, the interrupt routine does not interfere with the operating system or any application program.

Eight SMM instructions have been added to the x86 instruction set that permit software initiated SMM, and saving and restoring of the total CPU state when in SMM mode. Two SMM pins, SMI# and SMIACT#, support SMM functions.

Figure 2-37. System Management Memory Address Space



2.9.1 SMM Operation



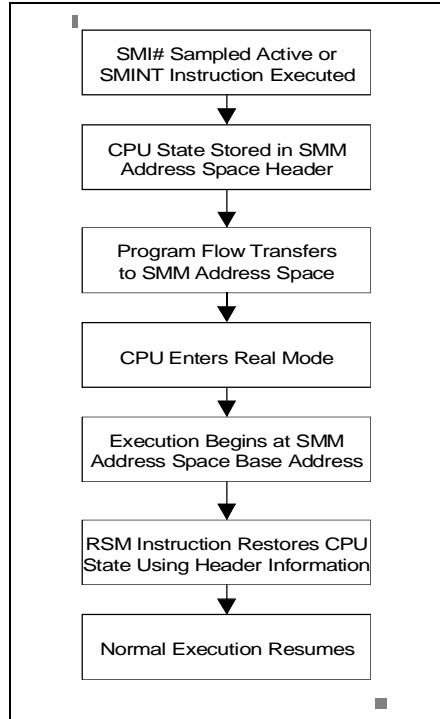
SMM operation is summarized in Figure 2-38. Entering SMM requires the assertion of the SMI# pin for at least two CLK periods or execution of the SMINT instruction. For the SMI# or SMINT instruction to be recognized, the following configuration register bits must be set as shown in Table 2-33. The configuration registers are discussed in detail earlier in this chapter.

Table 2-33. Requirements for Recognizing SMI# and SMINT

REGISTER (Bit)	SMI#	SMINT
SMI CCR1 (1)	1	1
SMAC CCR1 (2)	0	1
ARR3 SIZE (3-0)	> 0	> 0
SM3 CCR1 (7)	1	1

After recognizing SMI# or SMINT and prior to executing the SMI service routine, some of the CPU state information is changed. Prior to modification, this information is automatically saved in the SMM memory space header located at the top of SMM memory space. After the header is saved, the CPU enters real mode and begins executing the SMI service routine starting at the SMM memory base address.

The SMI service routine is user definable and may contain system or power management software. If the power management software forces the CPU to power down, or the SMI service routine modifies more than what is automatically saved, the complete CPU state information can be saved.



2.9.2 SMM Memory Space

Figure 2-38. SMI Execution Flow Diagram

Header

With every SMI interrupt or SMINT instruction, certain CPU state information is automatically saved in the SMM memory space header located at the top of SMM address space as shown Figure 2-39 and Table 2-34 (Page 2-66).

The header contains CPU state information that is modified when servicing an SMI interrupt. Included in this information are two pointers. The Current IP points to the instruction that was executing when the SMI was detected.

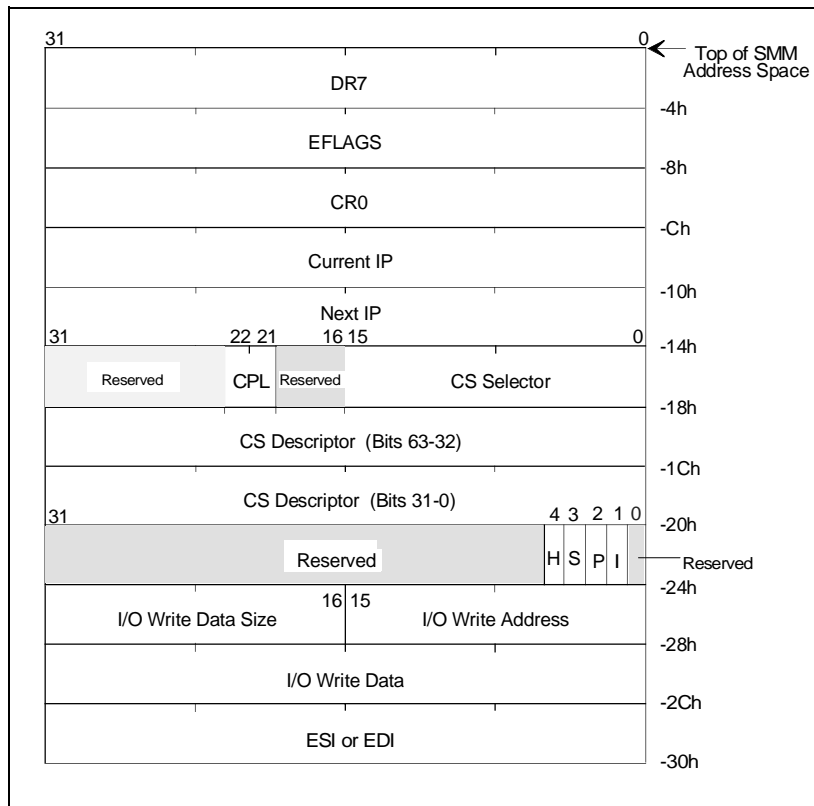


Figure 2-39. SMM Memory Space Header



The Next IP points to the instruction that will be executed after exiting SMM. Also saved are the contents of debug register 7 (DR7), the extended flags register (EFLAGS), and control register 0 (CR0). If SMM has been entered due to an I/O trap for a REP INSx or REP OUTSx instruction, the Current IP and Next IP fields contain the same addresses and the I and P field contain valid information.

If entry into SMM was caused by an I/O trap it is useful for the programmer to know the port address, data size and data value associated with that I/O operation. This information is also saved in the header and is only valid for an I/O write operation. The I/O write information is not restored within the CPU when executing a RSM instruction.

Table 2-34. SMM Memory Space Header

NAME	DESCRIPTION	SIZE
DR7	The contents of Debug Register 7.	4 Bytes
EFLAGS	The contents of Extended Flags Register.	4 Bytes
CR0	The contents of Control Register 0.	4 Bytes
Current IP	The address of the instruction executed prior to servicing SMI interrupt.	4 Bytes
Next IP	The address of the next instruction that will be executed after exiting SMM mode.	4 Bytes
CS Selector	Code segment register selector for the current code segment.	2 Bytes
CPL	Current privilege level for current code segment.	2 Bits
CS Descriptor	Code segment register descriptor for the current code segment.	8 Bytes
H	If set indicates the processor was in a halt or shutdown prior to servicing the SMM interrupt.	1 Bit
S	Software SMM Entry Indicator. S = 1, if current SMM is the result of an SMINT instruction. S = 0, if current SMM is not the result of an SMINT instruction.	1 Bit
P	REP INSx/OUTSx Indicator. P = 1 if current instruction has a REP prefix. P = 0 if current instruction does not have a REP prefix.	1 Bit
I	IN, INSx, OUT, or OUTSx Indicator. I = 1 if current instruction performed is an I/O WRITE. I = 0 if current instruction performed is an I/O READ.	1 Bit
I/O Write Data Size	Indicates size of data for the trapped I/O write. 01h = byte 03h = word 0Fh = dword	2 Bytes
I/O Write Address	Processor port used for the trapped I/O write.	2 Bytes
I/O Write Data	Data associated with the trapped I/O write.	4 Bytes
ESI or EDI	Restored ESI or EDI value. Used when it is necessary to repeat a REP OUTSx or REP INSx instruction when one of the I/O cycles caused an SMI# trap.	4 Bytes

Note: INSx = INS, INSB, INSW or INSD instruction.
Note: OUTSx = OUTS, OUTSB, OUTSW and OUTSD instruction.

2.9.3 SMM Instructions

The IBM 6x86 CPU automatically saves the minimal amount of CPU state information when entering SMM which allows fast SMI service routine entry and exit. After entering the SMI service routine, the MOV, SVDC, SVLDT and SVTS instructions can be used to save the complete CPU state information. If the SMI service routine modifies more than what is automatically saved or forces the CPU to power down, the complete CPU state information must be saved. Since the CPU is a static device, its internal state is retained when the input clock is stopped. Therefore, an entire CPU state save is not necessary prior to stopping the input clock.

The new SMM instructions, listed in Table 2-35, can only be executed if:

- 1) SMI# = 0
- 2) SM3 = 1
- 3) ARR3 SIZE > 0
- 4) Current Privilege Level = 0
- 5) SMAC bit is set or the CPU is in an SMI service routine.

If the above conditions are not met and an attempt is made to execute an SVDC, RSDC, SVLDT, RSLDT, SVTS, RSTS, SMINT or RSM instruction, an invalid opcode exception is generated. These instructions can be executed outside of defined SMM space provided the above conditions are met.

The SMINT instruction may be used as a software controlled mechanism to enter SMM.

Table 2-35. SMM Instruction Set

INSTRUCTION	OPCODE	FORMAT	DESCRIPTION
SVDC	0F 78 [mod sreg3 r/m]	SVDC mem80, sreg3	<i>Save Segment Register and Descriptor</i> Saves reg (DS, ES, FS, GS, or SS) to mem80.
RSDC	0F 79 [mod sreg3 r/m]	RSDC sreg3, mem80	<i>Restore Segment Register and Descriptor</i> Restores reg (DS, ES, FS, GS, or SS) from mem80. Use RSM to restore CS. Note: Processing "RSDC CS, Mem80" will produce an exception.
SVLDT	0F 7A [mod 000 r/m]	SVLDT mem80	<i>Save LDTR and Descriptor</i> Saves Local Descriptor Table (LDTR) to mem80.
RSLDT	0F 7B [mod 000 r/m]	RSLDT mem80	<i>Restore LDTR and Descriptor</i> Restores Local Descriptor Table (LDTR) from mem80.
SVTS	0F 7C [mod 000 r/m]	SVTS mem80	<i>Save TSR and Descriptor</i> Saves Task State Register (TSR) to mem80.
RSTS	0F 7D [mod 000 r/m]	RSTS mem80	<i>Restore TSR and Descriptor</i> Restores Task State Register (TSR) from mem80.
SMINT	0F 7E	SMINT	<i>Software SMM Entry</i> CPU enters SMM mode. CPU state information is saved in SMM memory space header and execution begins at SMM base address.
RSM	0F AA	RSM	<i>Resume Normal Mode</i> Exits SMM mode. The CPU state is restored using the SMM memory space header and execution resumes at interrupted point.

Note: mem80 = 80-bit memory location



All of the SMM instructions (except RSM and SMINT) save or restore 80 bits of data, allowing the saved values to include the hidden portion of the register contents.

2.9.4 SMM Memory Space

SMM memory space is defined by setting the SM3 bit and specifying the base address and size of the SMM memory space in the ARR3 register. The base address must be a multiple of the SMM memory space size. For example, a 32 KByte SMM memory space must be located at a 32 KByte address boundary. The memory space size can range from 4 KBytes to 4 GBytes.

SMM memory space accesses are always non-cacheable. SMM accesses ignore the state of the A20M# input pin and drive the A20 address bit to the unmasked value.

SMM memory space can be accessed while in normal mode by setting the SMAC bit in the CCR1 register. This feature may be used to initialize the SMM memory space.

2.9.5 SMI Service Routine Execution

Upon entry into SMM, after the SMM header has been saved, the CR0, EFLAGS, and DR7 registers are set to their reset values. The Code Segment (CS) register is loaded with the base, as defined by the ARR3 register, and a limit of 4 GBytes. The SMI service routine then begins execution at the SMM base address in real mode.

The programmer must save the value of any registers that may be changed by the SMI service routine. For data accesses immediately after entering the SMI service routine, the programmer must use CS as a segment override. I/O port access is possible during the routine but care must be taken to save registers modified by the I/O instructions. Before using a segment register, the register and the register's descriptor cache contents should be saved using the SVDC instruction. While executing in the SMM space, execution flow can transfer to normal memory locations.

Hardware interrupts, (INTRs and NMIs), may be serviced during a SMI service routine. If interrupts are to be serviced while executing in the SMM memory space, the SMM memory space must be within the 0 to 1 MByte address range to guarantee proper return to the SMI service routine after handling the interrupt.

INTRs are automatically disabled when entering SMM since the IF flag is set to its reset value. Once in SMM, the INTR can be enabled by setting the IF flag. NMI is also automatically disabled when entering SMM. Once in SMM, NMI can be enabled by setting NMI_EN in CCR3. If NMI is not enabled, the CPU latches one NMI event and services the interrupt after NMI has been enabled or after exiting SMM through the RSM instruction.

Within the SMI service routine, protected mode may be entered and exited as required, and real or protected mode device drivers may be called.

To exit the SMI service routine, a Resume (RSM) instruction, rather than an IRET, is executed. The RSM instruction causes the IBM 6x86 processor to restore the CPU state using the SMM header information and resume execution at the interrupted point. If the full CPU state was saved by the programmer, the stored values should be reloaded prior to executing the RSM instruction using the MOV, RSDC, RSLDT and RSTS instructions.

CPU States Related to SMM and Suspend Mode

The state diagram shown in Figure 2-40 (Page 2-70) illustrates the various CPU states associated with SMM and suspend mode. While in the SMI service routine, the IBM 6x86 CPU can enter suspend mode either by (1) executing a halt (HLT) instruction or (2) by asserting the SUSP# input.

During SMM operations and while in SUSP# initiated suspend mode, an occurrence of SMI#, NMI, or INTR is latched. (In order for INTR to be latched, the IF flag must be set.) The INTR or NMI is serviced after exiting suspend mode.

If suspend mode is entered via a HLT instruction from the operating system or application software, the reception of an SMI# interrupt causes the CPU to exit suspend mode and enter SMM.

2.10 Shutdown and Halt

The Halt Instruction (HLT) stops program execution and prevents the processor from using the local bus until restarted. The IBM 6x86 CPU then issues a special Stop Grant bus cycle and enters a low-power suspend mode if the SUSP_HLT bit in CCR2 is set. SMI, NMI, INTR with interrupts enabled (IF bit in EFLAGS=1), WM_RST or RESET forces the CPU out of the halt state. If interrupted, the saved code segment and instruction pointer specify the instruction following the HLT.

Shutdown occurs when a severe error is detected that prevents further processing. An NMI input can bring the processor out of shutdown if the IDT limit is large enough to contain the NMI interrupt vector and the stack has enough room to contain the vector and flag information. Otherwise, shutdown can only be exited by a processor reset.

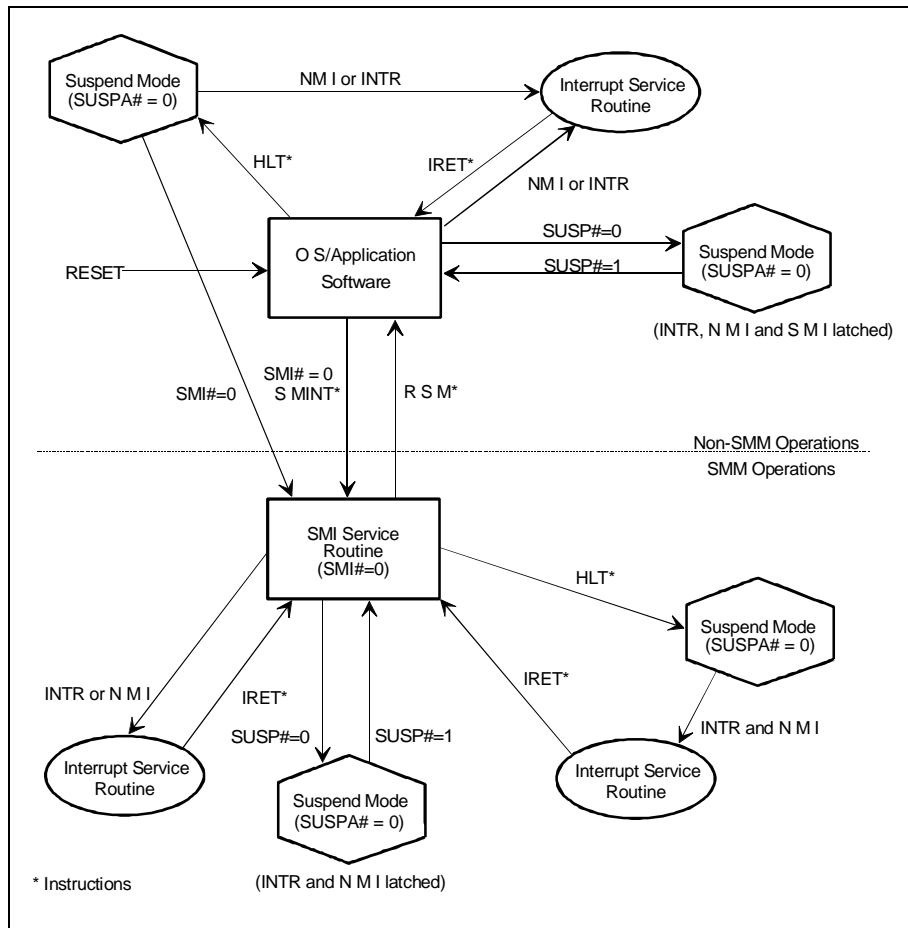


Figure 2-40. SMM and Suspend Mode State Diagram

2.11 Protection

Segment protection and page protection are safeguards built into the IBM 6x86 CPU protected mode architecture which deny unauthorized or incorrect access to selected memory addresses. These safeguards allow multitasking programs to be isolated from each other and from the operating system. Page protection is discussed earlier in this chapter. This section concentrates on segment protection.

Selectors and descriptors are the key elements in the segment protection mechanism. The segment base address, size, and privilege level are established by a segment descriptor. Privilege levels control the use of privileged instructions, I/O instructions and access to segments and segment descriptors. Selectors are used to locate segment descriptors.

Segment accesses are divided into two basic types, those involving code segments (e.g., control transfers) and those involving data accesses. The ability of a task to access a segment depends on the:

- segment type
- instruction requesting access
- type of descriptor used to define the segment
- associated privilege levels (described below).

Data stored in a segment can be accessed only by code executing at the same or a more privileged level. A code segment or procedure can only be called by a task executing at the same or a less privileged level.

2.11.1 Privilege Levels

The values for privilege levels range between 0 and 3. Level 0 is the highest privilege level (most privileged), and level 3 is the lowest privilege level (least privileged). The privilege level in real mode is effectively 0.

The **Descriptor Privilege Level (DPL)** is the privilege level defined for a segment in the segment descriptor. The DPL field specifies the minimum privilege level needed to access the memory segment pointed to by the descriptor.

The **Current Privilege Level (CPL)** is defined as the current task's privilege level. The CPL of an executing task is stored in the hidden portion of the code segment register and essentially is the DPL for the current code segment.

The **Requested Privilege Level (RPL)** specifies a selector's privilege level and is used to distinguish between the privilege level of a routine actually accessing memory (the CPL), and the privilege level of the original requestor (the RPL) of the memory access. The lesser of the RPL and CPL is called the effective privilege level (EPL). Therefore, if RPL = 0 in a segment selector, the effective privilege level is always determined by the CPL. If RPL = 3, the effective privilege level is always 3 regardless of the CPL.

For a memory access to succeed, the effective privilege level (EPL) must be at least as privileged as the descriptor privilege level ($EPL \leq DPL$). If the EPL is less privileged than the DPL ($EPL > DPL$), a general protection fault is generated. For example, if a segment has a DPL = 2, an instruction accessing the segment only succeeds if executed with an $EPL \leq 2$.



2.11.2 I/O Privilege Levels

The I/O Privilege Level (IOPL) allows the operating system executing at CPL=0 to define the least privileged level at which IOPL-sensitive instructions can unconditionally be used. The IOPL-sensitive instructions include CLI, IN, OUT, INS, OUTS, REP INS, REP OUTS, and STI. Modification of the IF bit in the EFLAGS register is also sensitive to the I/O privilege level. The IOPL is stored in the EFLAGS register.

An I/O permission bit map is available as defined by the 32-bit Task State Segment (TSS). Since each task can have its own TSS, access to individual processor I/O ports can be granted through separate I/O permission bit maps.

If $CPL \leq IOPL$, IOPL-sensitive operations can be performed. If $CPL > IOPL$, a general protection fault is generated if the current task is associated with a 16-bit TSS. If the current task is associated with a 32-bit TSS and $CPL > IOPL$, the CPU consults the I/O permission bitmap in the TSS to determine on a port-by-port basis whether or not I/O instructions (IN, OUT, INS, OUTS, REP INS, REP OUTS) are permitted, and the remaining IOPL-sensitive operations generate a general protection fault.

2.11.3 Privilege Level Transfers

A task's CPL can be changed only through intersegment control transfers using gates or task switches to a code segment with a different privilege level. Control transfers result from exception and interrupt servicing and from execution of the CALL, JMP, INT, IRET and RET instructions.

There are five types of control transfers that are summarized in Table 2-36 (Page 2-73). Control transfers can be made only when the operation causing the control transfer references the correct descriptor type. Any violation of these descriptor usage rules causes a general protection fault.

Any control transfer that changes the CPL within a task results in a change of stack. The initial values for the stack segment (SS) and stack pointer (ESP) for privilege levels 0, 1, and 2 are stored in the TSS. During a CALL control transfer, the SS and ESP are loaded with the new stack pointer and the previous stack pointer is saved on the new stack. When returning to the original privilege level, the RET or IRET instruction restores the less-privileged stack.

Table 2-36. Descriptor Types Used for Control Transfer

TYPE OF CONTROL TRANSFER	OPERATION TYPES	DESCRIPTOR REFERENCED	DESCRIPTOR TABLE
Intersegment within the same privilege level.	JMP, CALL, RET, IRET*	Code Segment	GDT or LDT
Intersegment to the same or a more privileged level.	CALL	Gate Call	GDT or LDT
Interrupt within task (could change CPL level).	Interrupt Instruction, Exception, External Interrupt	Trap or Interrupt Gate	IDT
Intersegment to a less privileged level (changes task CPL).	RET, IRET*	Code Segment	GDT or LDT
Task Switch via TSS	CALL, JMP	Task State Segment	GDT
Task Switch via Task Gate	CALL, JMP	Task Gate	GDT or LDT
	IRET**, Interrupt Instruction, Exception, External Interrupt	Task Gate	IDT

* NT (Nested Task bit in EFLAGS) = 0

** NT (Nested Task bit in EFLAGS) = 1

Gates

Gate descriptors provide protection for privilege transfers among executable segments.

Gates are used to transition to routines of the same or a more privileged level. Call gates, interrupt gates and trap gates are used for privilege transfers within a task. Task gates are used to transfer between tasks.

Gates conform to the standard rules of privilege. In other words, gates can be accessed by a task if the effective privilege level (EPL) is the same or more privileged than the gate descriptor's privilege level (DPL).

2.11.4 Initialization and Transition to Protected Mode

The IBM 6x86 microprocessor switches to real mode immediately after RESET. While operating in real mode, the system tables and registers should be initialized. The GDTR and IDTR must point to a valid GDT and IDT, respectively. The GDT must contain descriptors which describe the initial code and data segments.

The processor can be placed in protected mode by setting the PE bit in the CR0 register. After enabling protected mode, the CS register should be loaded and the instruction decode queue should be flushed by executing an intersegment JMP. Finally, all data segment registers should be initialized with appropriate selector values.



2.12 Virtual 8086 Mode

Both real mode and virtual 8086 (V86) mode are supported by the IBM 6x86 CPU allowing execution of 8086 application programs and 8086 operating systems. V86 mode allows the execution of 8086-type applications, yet still permits use of the IBM 6x86 CPU paging mechanism. V86 tasks run at privilege level 3. When loaded, all segment limits are set to FFFFh (64K) as in real mode.

2.12.1 V86 Memory Addressing

While in V86 mode, segment registers are used in an identical fashion to real mode. The contents of the segment register are multiplied by 16 and added to the offset to form the segment base linear address. The IBM 6x86 CPU permits the operating system to select which programs use the V86 address mechanism and which programs use protected mode addressing for each task.

The IBM 6x86 CPU also permits the use of paging when operating in V86 mode. Using paging, the 1-MByte address space of the V86 task can be mapped to anywhere in the 4-GByte linear address space of the IBM 6x86 CPU.

The paging hardware allows multiple V86 tasks to run concurrently, and provides protection and operating system isolation. The paging hardware must be enabled to run multiple V86 tasks or to relocate the address space of a V86 task to physical address space greater than 1 MByte.

2.12.2 V86 Protection

All V86 tasks operate with the least amount of privilege (level 3) and are subject to all of the IBM 6x86 CPU protected mode protection checks. As a result, any attempt to execute a privileged instruction within a V86 task results in a general protection fault.

In V86 mode, a slightly different set of instructions are sensitive to the I/O privilege level (IOPL) than in protected mode. These instructions are: CLI, INT n, IRET, POPF, PUSHF, and STI. The INT3, INTO and BOUND variations of the INT instruction are not IOPL sensitive.

2.12.3 V86 Interrupt Handling

To fully support the emulation of an 8086-type machine, interrupts in V86 mode are handled as follows. When an interrupt or exception is serviced in V86 mode, program execution transfers to the interrupt service routine at privilege level 0 (i.e., transition from V86 to protected mode occurs) and the VM bit in the EFLAGS register is cleared. The protected mode interrupt service routine then determines if the interrupt came from a protected mode or V86 application by examining the VM bit in the EFLAGS image stored on the stack. The interrupt service routine may then choose to allow the 8086 operating system to handle the interrupt or may emulate the function of the interrupt handler. Following completion of the interrupt service routine, an IRET instruction restores the EFLAGS register (restores VM=1) and segment selectors and control returns to the interrupted V86 task.

2.12.4 Entering and Leaving V86 Mode

V86 mode is entered from protected mode by either executing an IRET instruction at CPL = 0 or by task switching. If an IRET is used, the stack must contain an EFLAGS image with VM = 1. If a task switch is used, the TSS must contain an EFLAGS image containing a 1 in the VM bit position. The POPF instruction cannot be used to enter V86 mode since the state of the VM bit is not affected. V86 mode can only be exited as the result of an interrupt or exception. The transition out must use a 32-bit trap or interrupt gate which must point to a non-conforming privilege level 0 segment (DPL = 0), or a 32-bit TSS. These restrictions are required to permit the trap handler to IRET back to the V86 program.

2.13 Floating Point Unit Operations

The IBM 6x86 CPU includes an on-chip FPU that provides the user access to a complete set of floating point instructions (see Chapter 6). Information is passed to and from the FPU using eight data registers accessed in a stack-like manner, a control register, and a status register. The IBM 6x86 CPU also provides a data register tag word which improves context switching and performance by maintaining empty/non-empty status for each of the eight data registers. In addition, registers in the CPU contain pointers to (a) the memory location containing the current instruction word and (b) the memory location containing the operand associated with the current instruction word (if any).

FPU Tag Word Register. The IBM 6x86 CPU maintains a tag word register (Figure 2-41 (Page 2-76)) comprised of two bits for each physical data register. Tag Word fields assume one of four values depending on the contents of their associated data registers, Valid (00), Zero (01), Special (10), and Empty (11). Note: Denormal, Infinity, QNaN, SNaN and unsupported formats are tagged as “Special”. Tag values are maintained transparently by the IBM 6x86™ CPU and are only available to the programmer indirectly through the FSTENV and FSAVE instructions.

FPU Control and Status Registers. The FPU circuitry communicates information about its status and the results of operations to the programmer via the status register. The FPU status register is comprised of bit fields that reflect exception status, operation execution status, register status, operand class, and comparison results. The FPU status register bit definitions are shown in Figure 2-42 (Page 2-76) and Table 2-37 (Page 2-76).

The FPU Mode Control Register (MCR) is used by the CPU to specify the operating mode of the FPU. The MCR contains bit fields which specify the rounding mode to be used, the precision by which to calculate results, and the exception conditions which should be reported to the CPU via traps. The user controls precision, rounding, and exception reporting by setting or clearing appropriate bits in the MCR. The FPU mode control register bit definitions are shown in Figure 2-43 (Page 2-77) and Table 2-38 (Page 2-77).

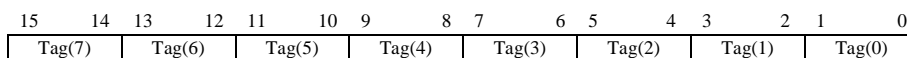


Figure 2-41. FPU Tag Word Register

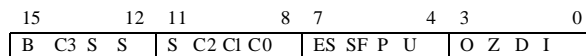


Figure 2-42. FPU Status Register

Table 2-37. FPU Status Register Bit Definitions

BIT POSITION	NAME	DESCRIPTION
15	B	Copy of the ES bit. (ES is bit 7 in this table.)
14, 10 - 8	C3 - C0	Condition code bits.
13 - 11	SSS	Top of stack register number which points to the current TOS.
7	ES	Error indicator. Set to 1 if an unmasked exception is detected.
6	SF	Stack Fault or invalid register operation bit.
5	P	Precision error exception bit.
4	U	Underflow error exception bit.
3	O	Overflow error exception bit.
2	Z	Divide by zero exception bit.
1	D	Denormalized operand error exception bit.
0	I	Invalid operation exception bit.

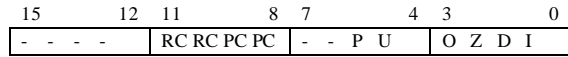


Figure 2-43. FPU Mode Control Register

Table 2-38. FPU Mode Control Register Bit Definitions

BIT POSITION	NAME	DESCRIPTION
11 - 10	RC	Rounding Control bits: 00 Round to nearest or even 01 Round towards minus infinity 10 Round towards plus infinity 11 Truncate
9 - 8	PC	Precision Control bits: 00 24-bit mantissa 01 Reserved 10 53-bit mantissa 11 64-bit mantissa
5	P	Precision error exception bit mask.
4	U	Underflow error exception bit mask.
3	O	Overflow error exception bit mask.
2	Z	Divide by zero exception bit mask.
1	D	Denormalized operand error exception bit mask.
0	I	Invalid operation exception bit mask.