Input output

Polling vs interrupt, Interrupt controllers, interrupt descriptor table, interrupt handlers, Direct memory access, hard disks

Abhilash Jindal

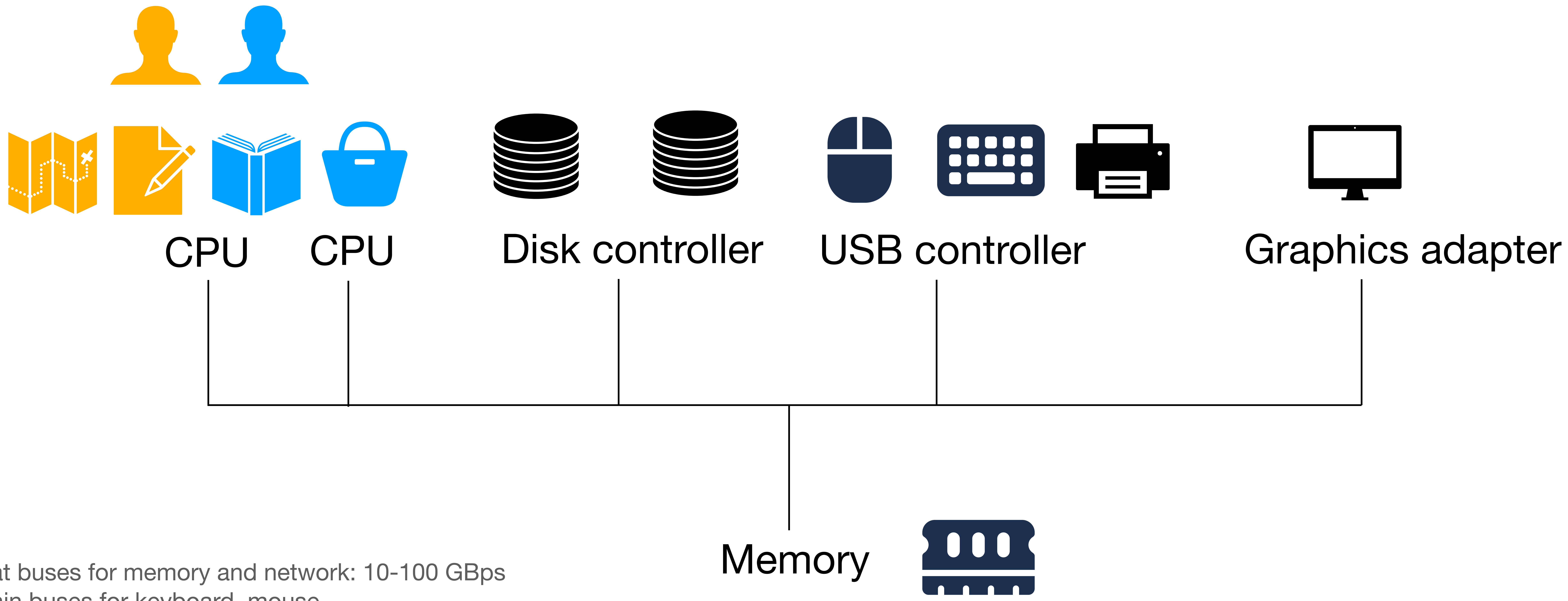
Agenda

- Overview of IO devices (OSTEP Ch. 36): Polling, Interrupts, Direct memory access
- Interrupt handling (xv6 Ch. 3): interrupt controllers, interrupt descriptor table
- Hard disk drives (OSTEP Ch. 37): disk geometry, disk scheduling
- Redundant Array of Inexpensive Disks (OSTEP Ch. 38): improve capacity, throughput, fault tolerance

Overview of IO devices

OSTEP Ch. 36

Computer organization

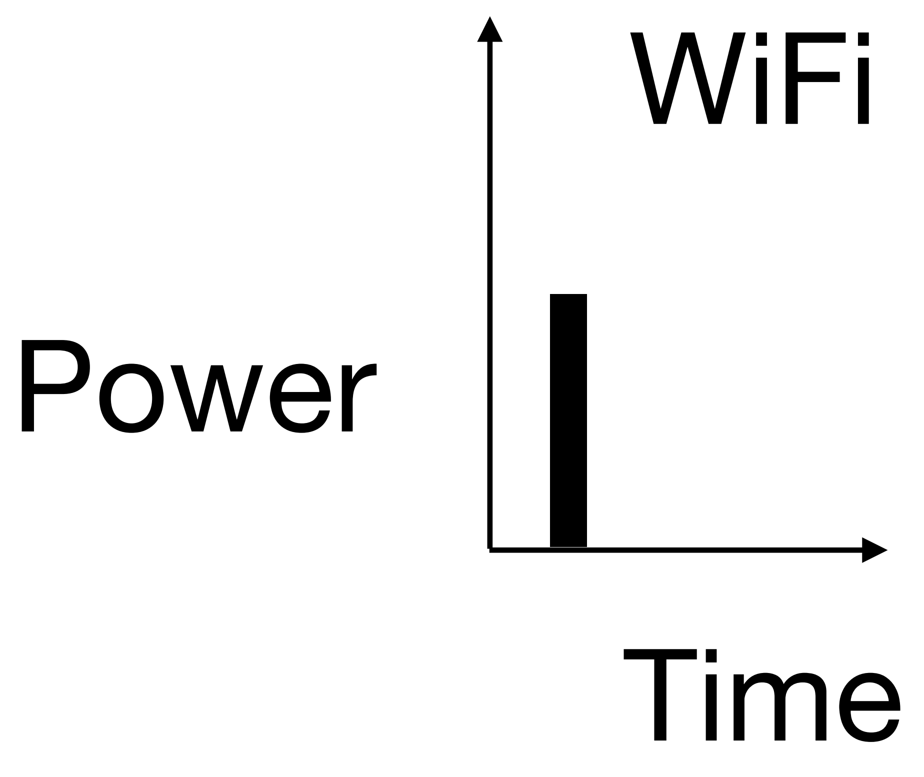


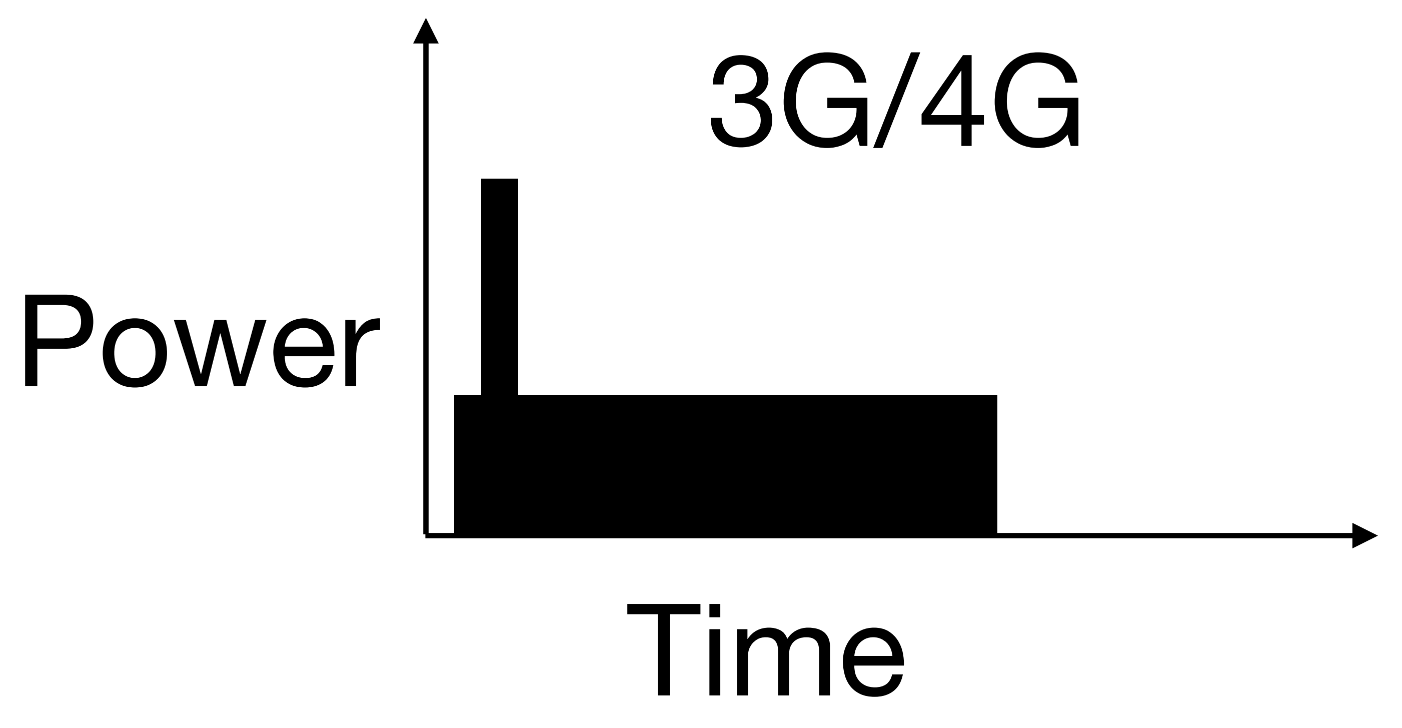
Fat buses for memory and network: 10-100 GBps
Thin buses for keyboard, mouse

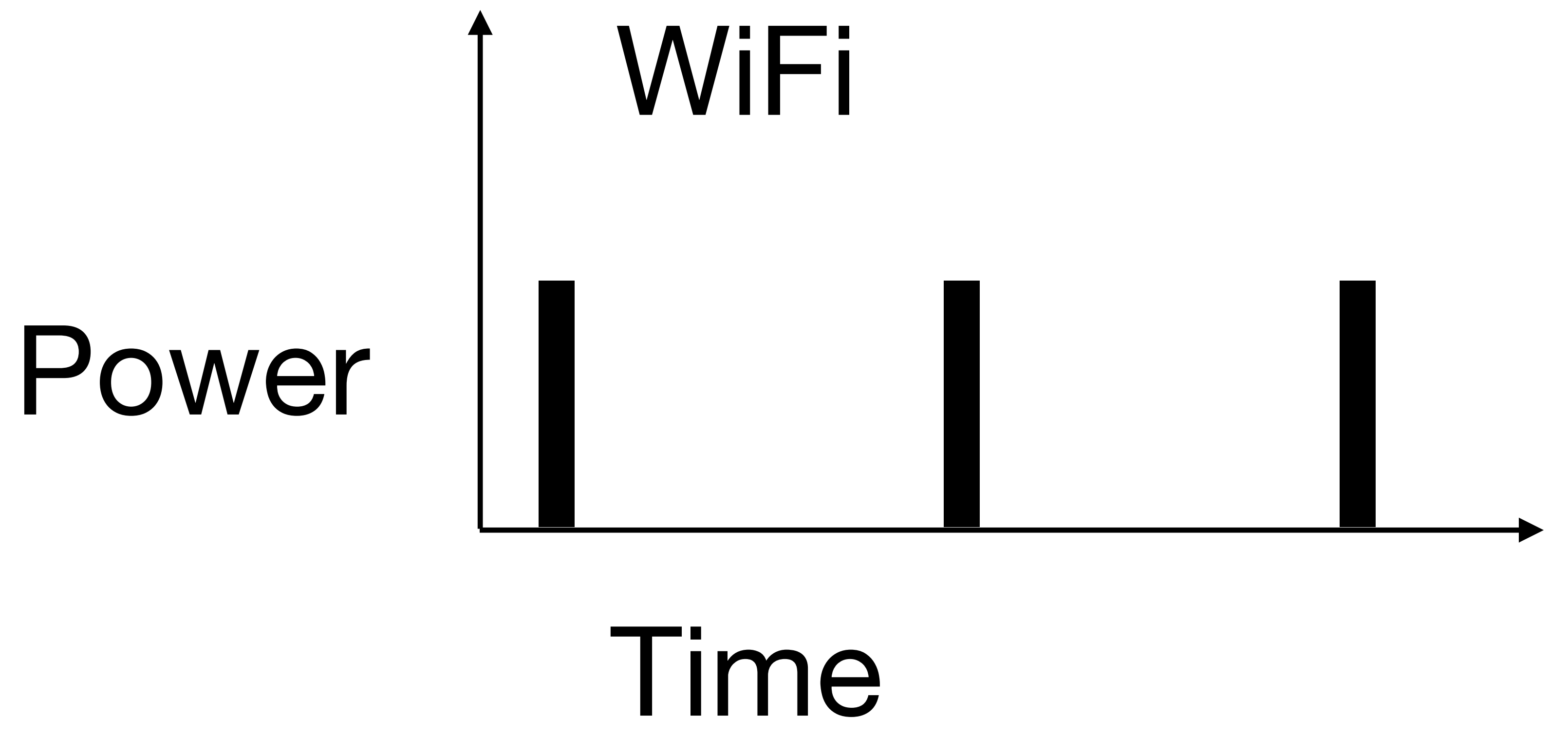
Fitting into the OS

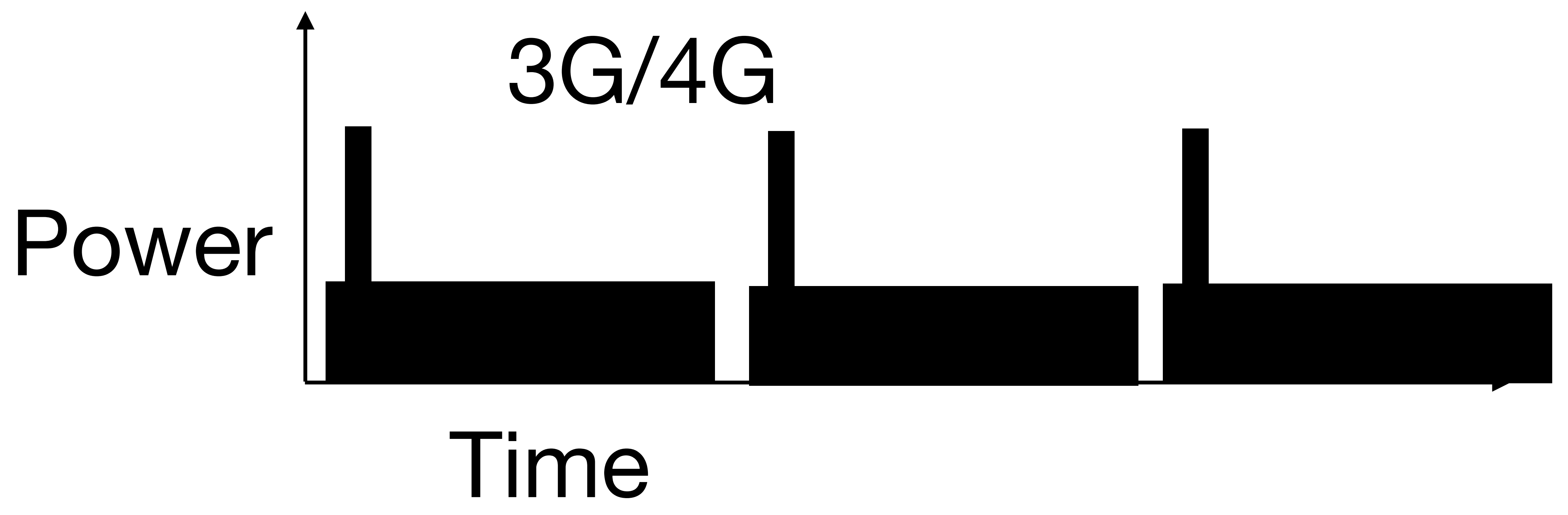
Hide device specific details in device driver

- Abstraction allows OS and applications to stay device-neutral
- Abstraction can hurt.
    - Example: 3G/4G are more efficient for bursty traffic
- > 70% of OS code is device drivers. Tend to have most number of bugs









Memory-mapped IO and Port-mapped IO

- Memory mapped:
    - Regular memory access instructions
    - Reads and writes are routed to appropriate device
    - Does not behave like memory! Reading same location twice can change due to external events
- Port mapped:
    - Special IN and OUT instructions

```
+------------------+  <- 0xFFFFFFFF (4GB)
|      32-bit      |
|  memory mapped   |
|     devices      |
|                  |
/\/\/\/\/\/\/\/\/\/\

/\/\/\/\/\/\/\/\/\/\
|                  |
|      Unused      |
|                  |
+------------------+  <- depends on amount of RAM
|                  |
|                  |
| Extended Memory  |
|                  |
|                  |
+------------------+  <- 0x00100000 (1MB)
|     BIOS ROM     |
+------------------+  <- 0x000F0000 (960KB)
|  16-bit devices, |
|  expansion ROMs  |
+------------------+  <- 0x000C0000 (768KB)
|   VGA Display    |
+------------------+  <- 0x000A0000 (640KB)
|                  |
|    Low Memory    |
|                  |
+------------------+  <- 0x00000000
```

Canonical protocol

- Poll device until it is ready
- CPU cannot do anything else.
- Example: CPU needs to spend ~1 million instructions waiting for disk
- Ok for bootloader. It does not have anything else to do.
- Not ok for OS. It can run other processes.

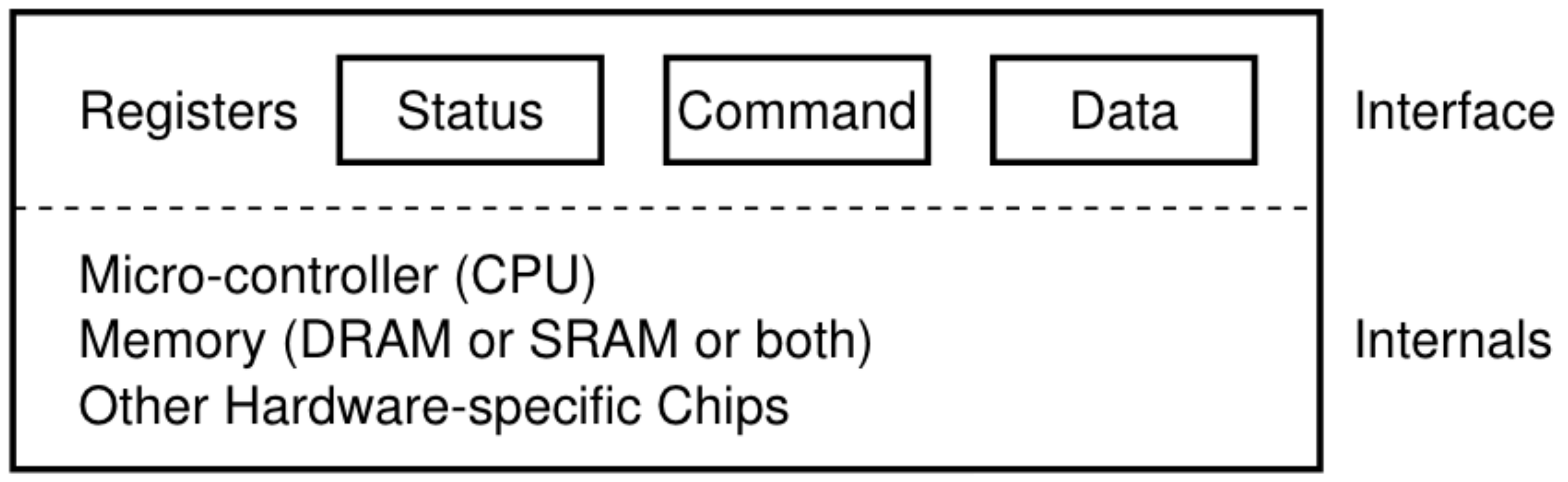


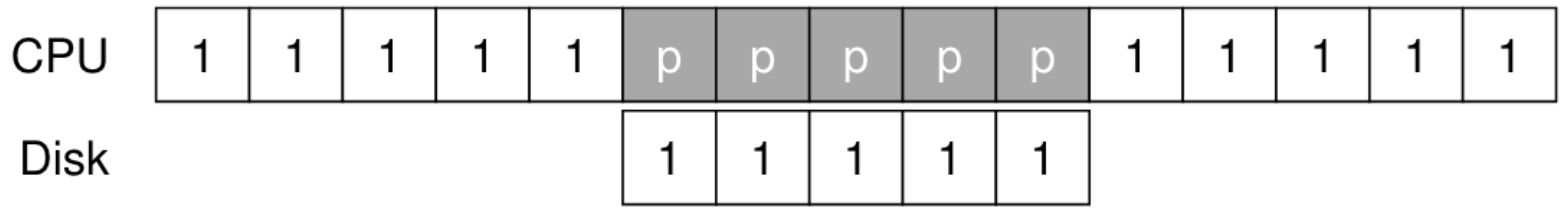
Figure 36.3: **A Canonical Device**

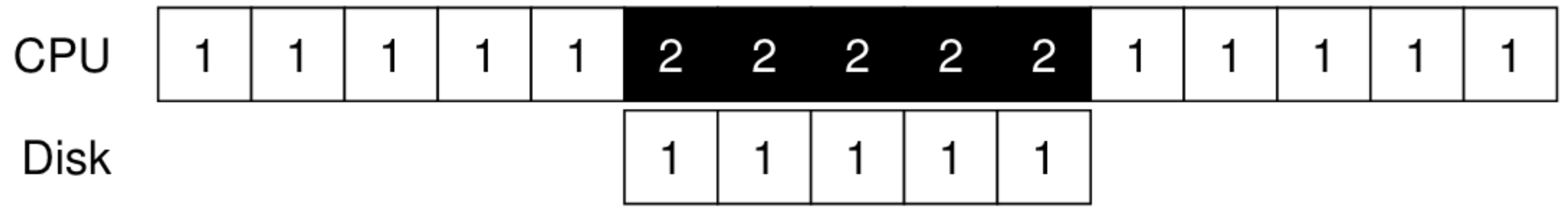
bootmain.c

```
void waitdisk(void){
  // Wait for disk ready.
  while((inb(0x1F7) & 0xC0) != 0x40);
}

// Read a single sector at offset into dst.
void readsect(void *dst, uint offset) {
  // Issue command.
  waitdisk();
  outb(0x1F2, 1);   // count = 1
  ..
  outb(0x1F7, 0x20);  // cmd 0x20 - read sectors

  // Read data.
  waitdisk();
  insl(0x1F0, dst, SECTSIZE/4);
}
```

Lowering CPU overheads with interrupts

- Device sends an interrupt that it is ready
- CPU runs another process in the meantime
- Better CPU utilisation
- Not a good idea if device is fast.
    - If first poll finds that the device is ready, unnecessary overhead of switching processes





More efficient data movement

Direct Memory Access (DMA)

CPU	1	1	1	2	2	c	c	c	1
Disk				1	1				

CPU	1	1	1	2	2	2	2	2	1
DMA						c	c	c	
Disk				1	1				

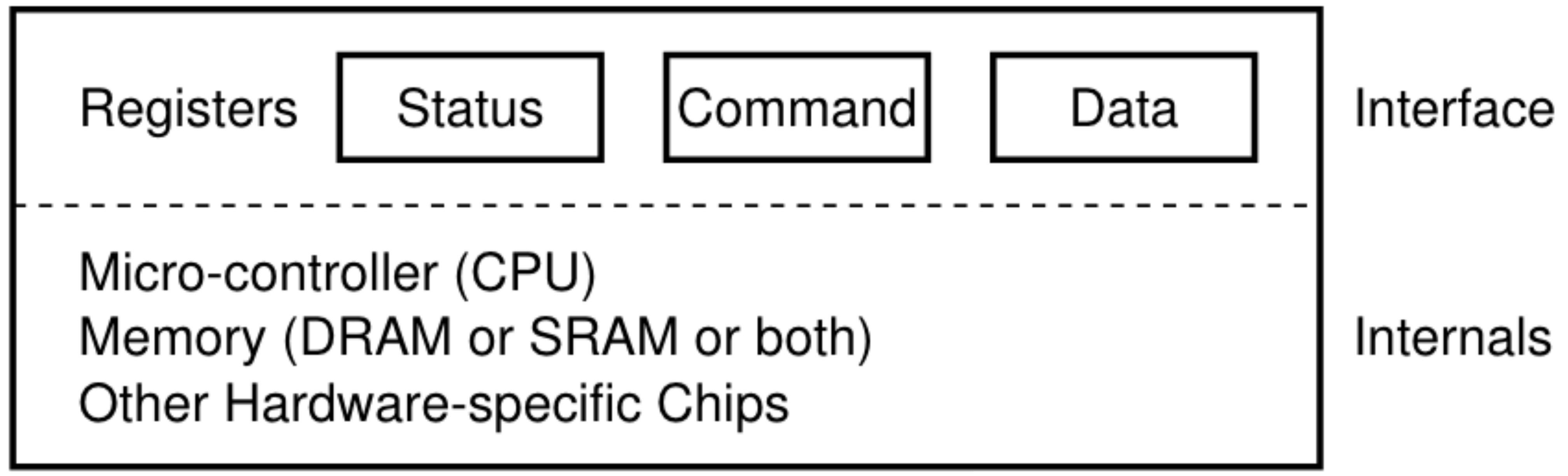


Figure 36.3: **A Canonical Device**

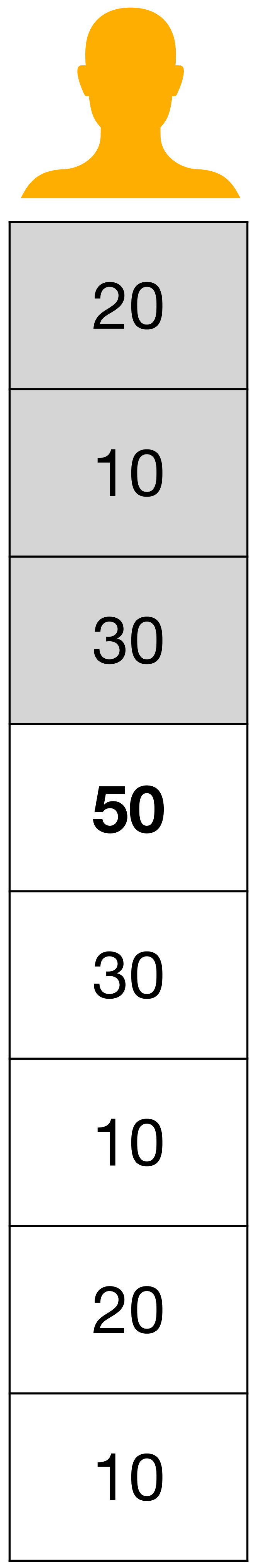
bootmain.c

```
void waitdisk(void){
  // Wait for disk ready.
  while((inb(0x1F7) & 0xC0) != 0x40);
}

// Read a single sector at offset into dst.
void readsect(void *dst, uint offset) {
  // Issue command.
  waitdisk();
  outb(0x1F2, 1);   // count = 1
   ..
  outb(0x1F7, 0x20);  // cmd 0x20 - read sectors

  // Read data.
  waitdisk();
  insl(0x1F0, dst, SECTSIZE/4);
}
```

Interrupt controllers, interrupt handling

xv6 Ch. 3 “Code: interrupts”

Calculator analogy



- 2 0 = (move pointer to 10)
- + 1 0 = (move pointer to 30)
- + 3 0 = (move pointer to 50)

Interrupt

Give me the calculator!

- 3*2 = 6

End of Interrupt

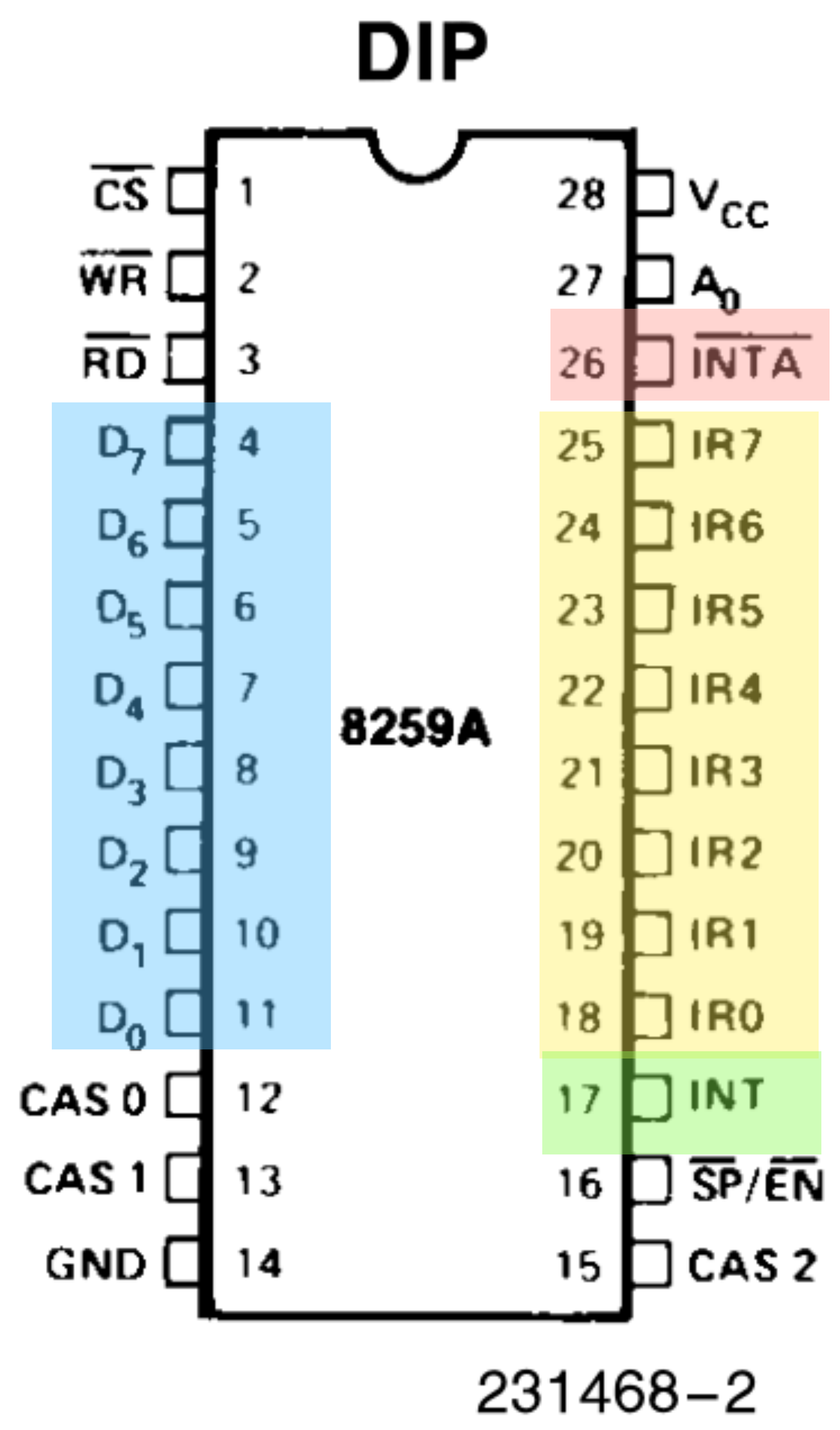
Ok, you can have it back

- + 5 0 = (move pointer to 30)
- + 3 0 = (move pointer to 10)
- + 1 0 = (move pointer to 20)
- + 2 0 = (move pointer to 10)

Programmable interrupt controllers (PIC)

Example: Intel 8259A

- Devices connect to IR0-IR7 pins.
  Device enables its pin to raise interrupt
- INT pin connects to CPU.
- PIC sends an 8-bit "interrupt vector" to CPU via D0-D7 pins
- CPU acknowledges that it is now working on interrupt on INTA pin
- CPU acknowledges "end-of-interrupt" on INTA pin



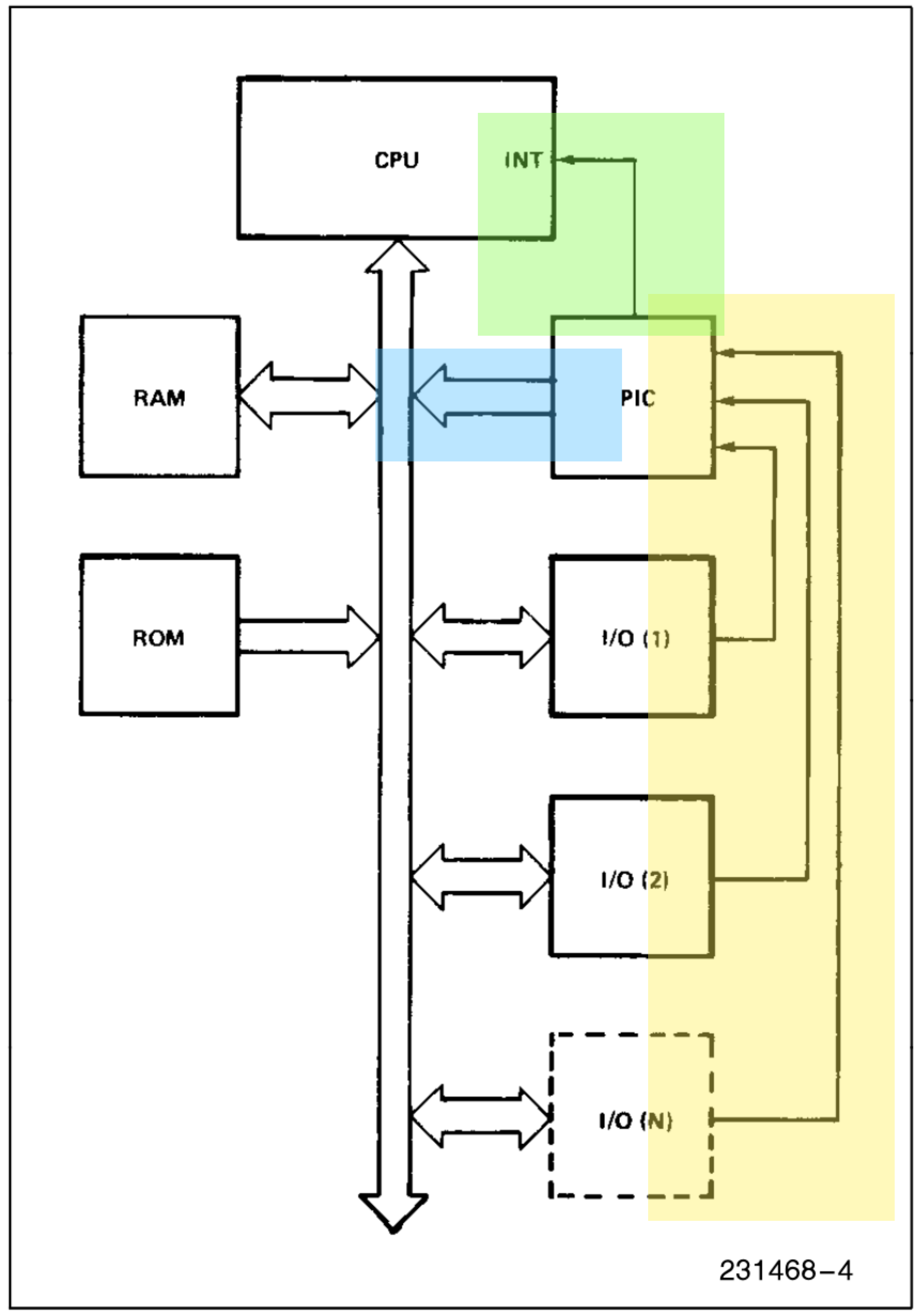
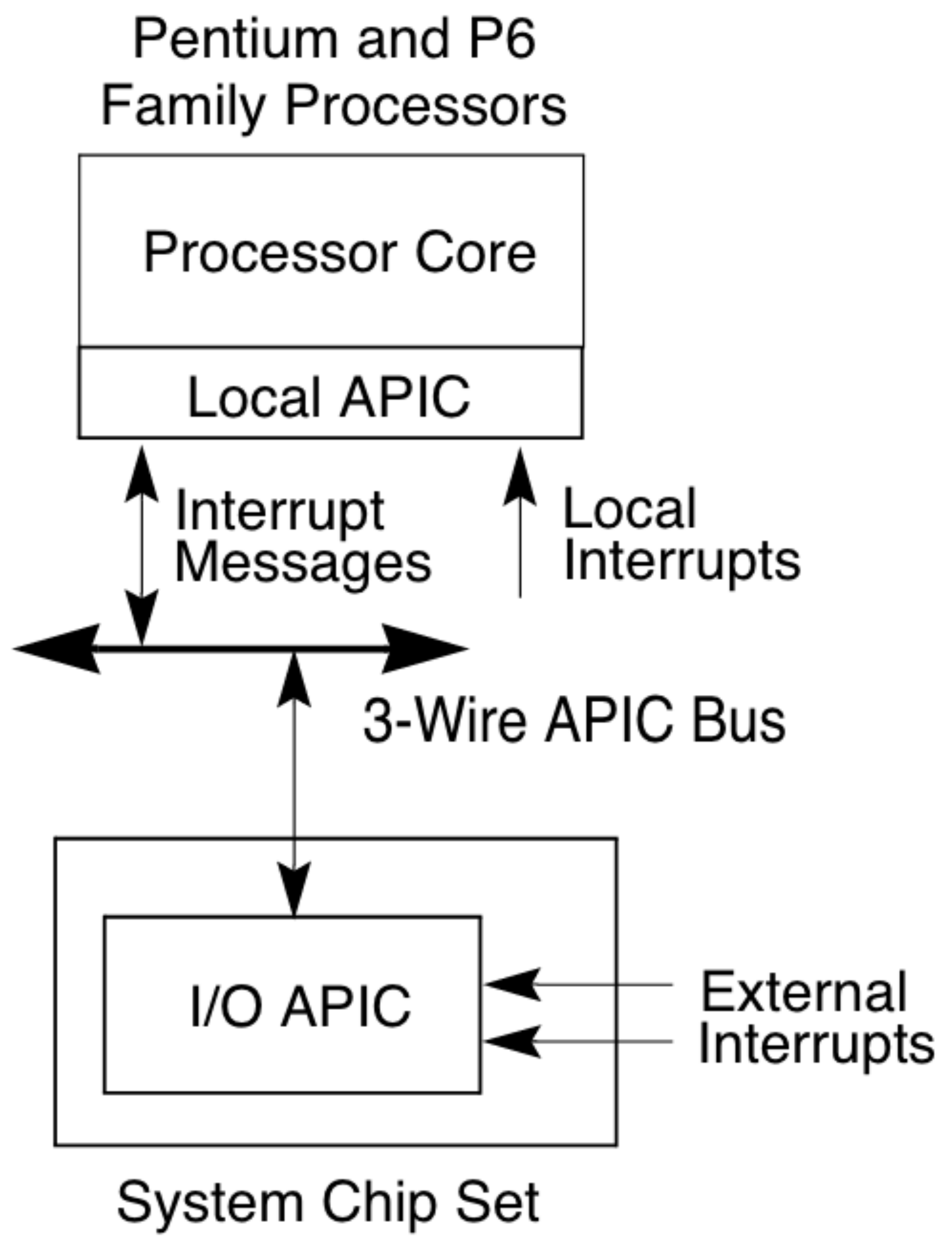


Figure 3b. Interrupt Method

Advanced programmable interrupt controllers (APIC)

- Each CPU can have local APICs for handling *local interrupts* like timer, thermal sensor, etc.
- A separate IO APIC receives external interrupts like keyboard, mouse, disk, etc and forwards it to a particular CPU
  - Example: Route keyboard interrupts to CPU-0, disk interrupts to CPU-1



p3-pic

Code walkthrough

- main.c calls lapicinit, picinit, ioapicinit
- lapicinit enables timer interrupt at every 10ms. lapicw is just writing to memory location (MMIO)
- picinit just disables PIC using outb instructions (PMIO)
- ioapicinit initialises IO APIC with MMIO
- Bootloader had disabled interrupt with cli. We will not receive interrupts yet.

Interrupt enable flag

- cli: Clear interrupt flag
  - PIC is not allowed to interrupt
- sti: Set interrupt flag

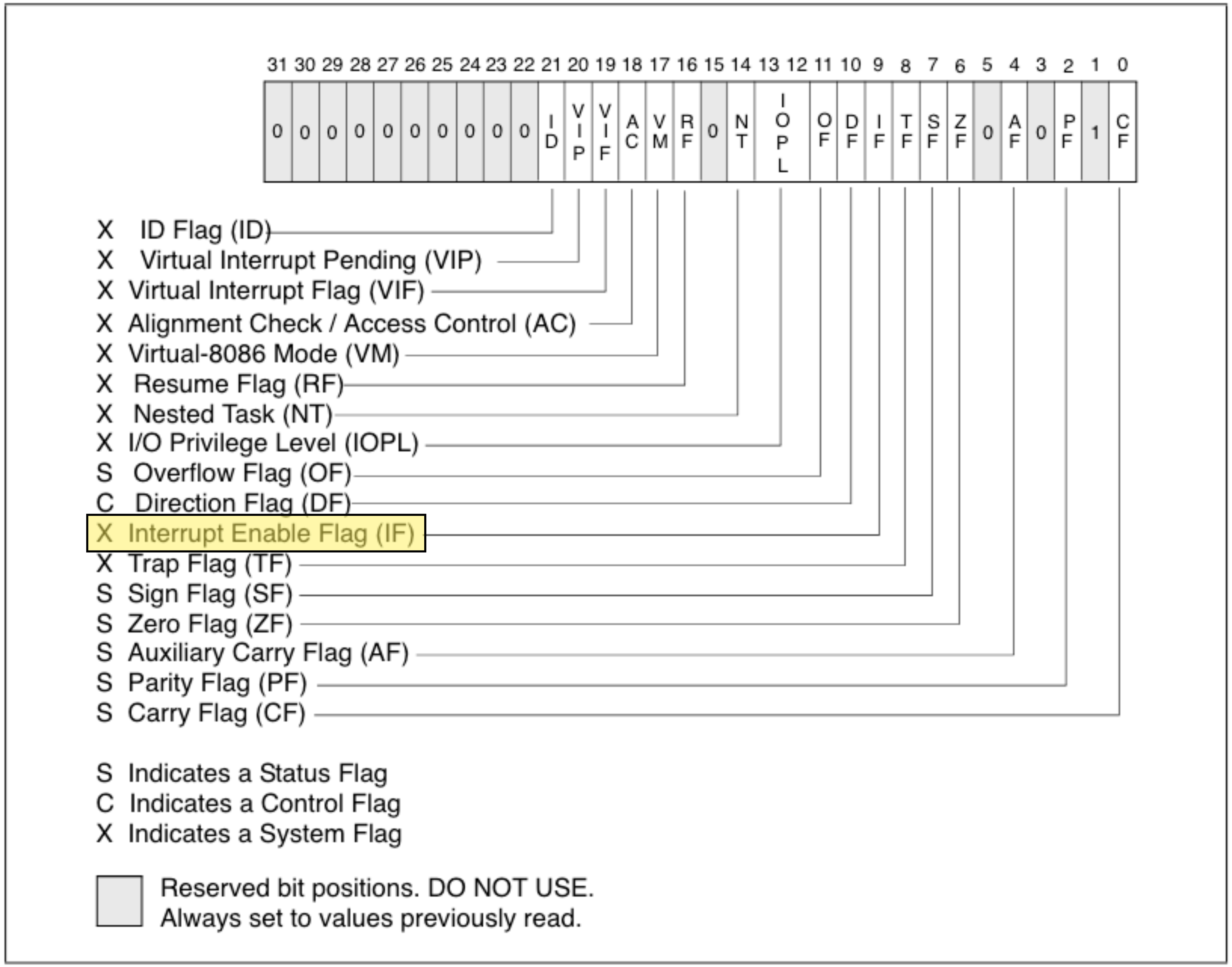
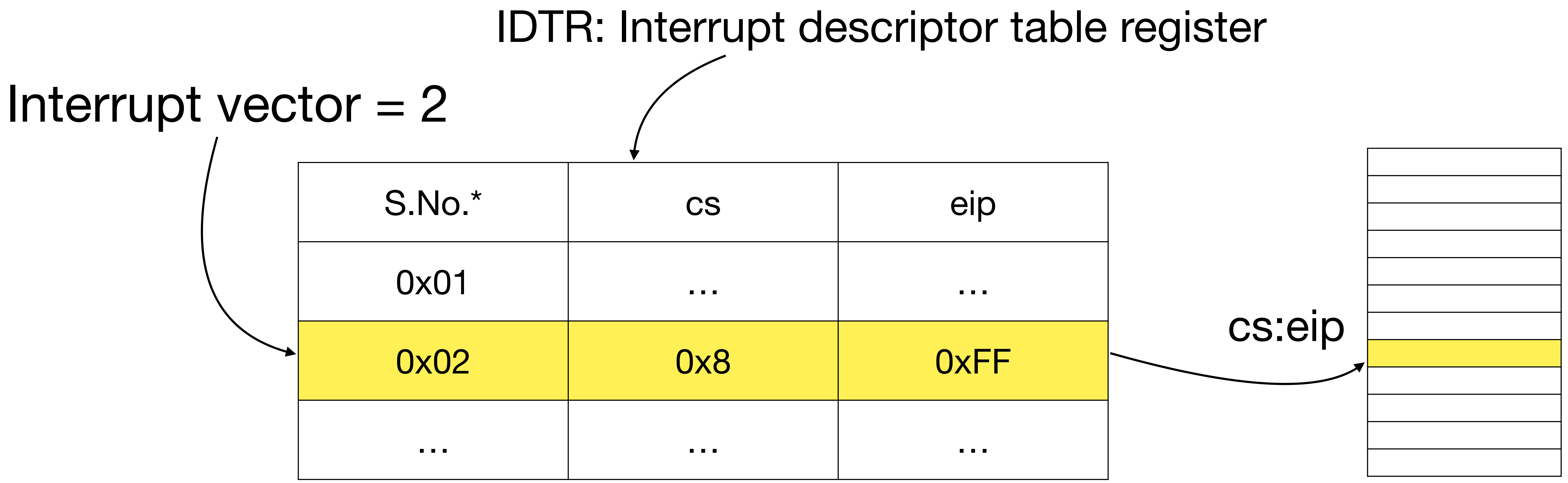

31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0
0 0 0 0 0 0 0 0 0 0 ID VIP VIF AC VM RF 0 NT IOPL OF DF IF TF SF ZF 0 AF 0 PF 1 CF
X ID Flag (ID)
X Virtual Interrupt Pending (VIP)
X Virtual Interrupt Flag (VIF)
X Alignment Check / Access Control (AC)
X Virtual-8086 Mode (VM)
X Resume Flag (RF)
X Nested Task (NT)
X I/O Privilege Level (IOPL)
S Overflow Flag (OF)
C Direction Flag (DF)
X Interrupt Enable Flag (IF)
X Trap Flag (TF)
S Sign Flag (SF)
S Zero Flag (ZF)
S Auxiliary Carry Flag (AF)
S Parity Flag (PF)
S Carry Flag (CF)
S Indicates a Status Flag
C Indicates a Control Flag
X Indicates a System Flag
Reserved bit positions. DO NOT USE.
Always set to values previously read.

Figure 3-8. EFLAGS Register

Interrupt handling in a nutshell

- OS sets up “interrupt descriptor table” (IDT)
- Points IDTR to IDT using LIDT instruction
- When interrupt occurs, jump %eip to interrupt handler, handle interrupt, tells LAPIC about end of interrupt, resume what we were doing



S.No.*	cs	eip
0x01	…	…
0x02	0x8	0xFF
…	…	…

Interrupt descriptor table

- Interrupt descriptor table register (IDTR) points to interrupt descriptor table in memory
- OS sets up IDT and initialises IDTR using LIDT instruction
- Interrupt descriptor table has one entry for each interrupt vector (upto 2^8=256)

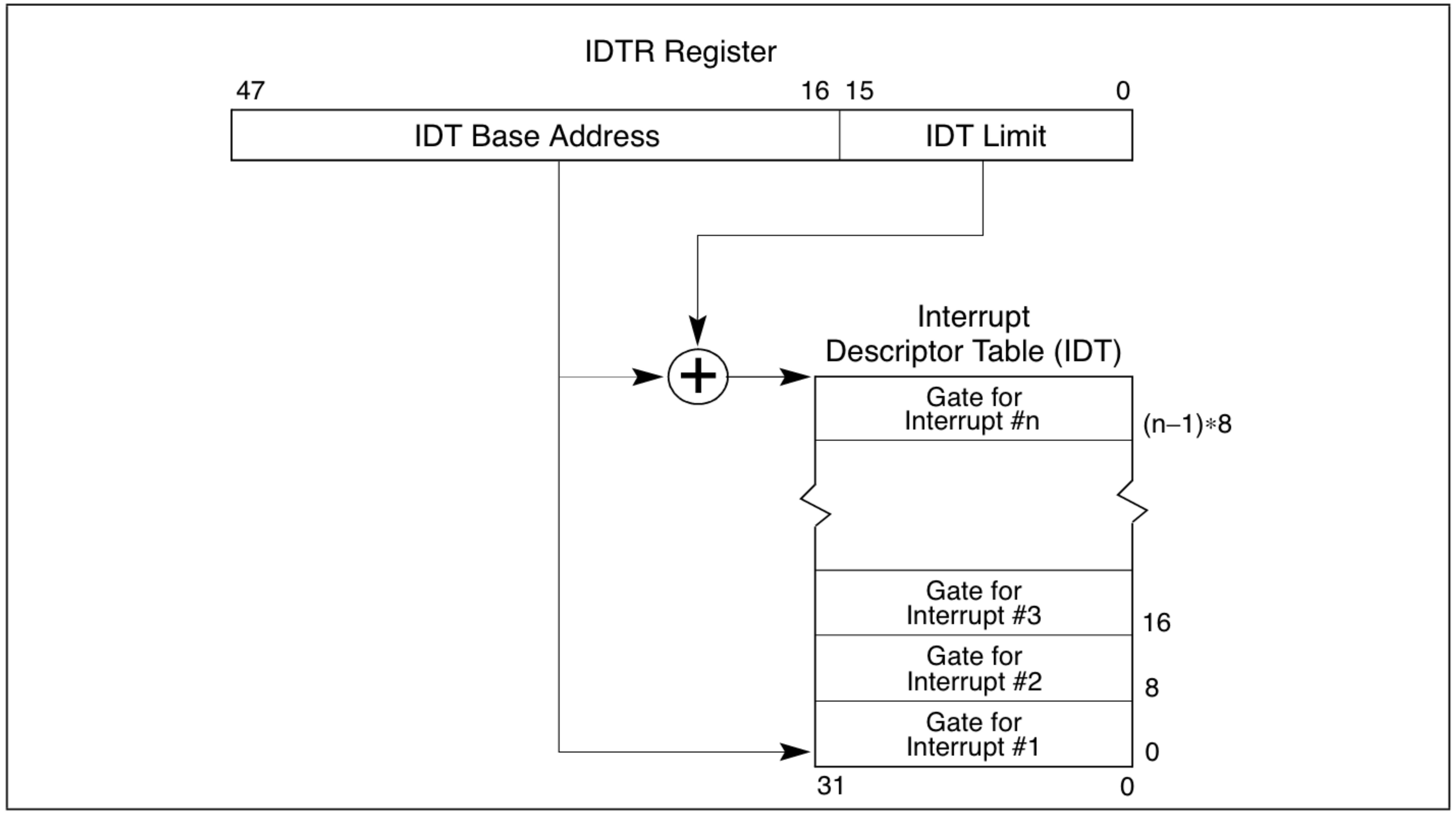
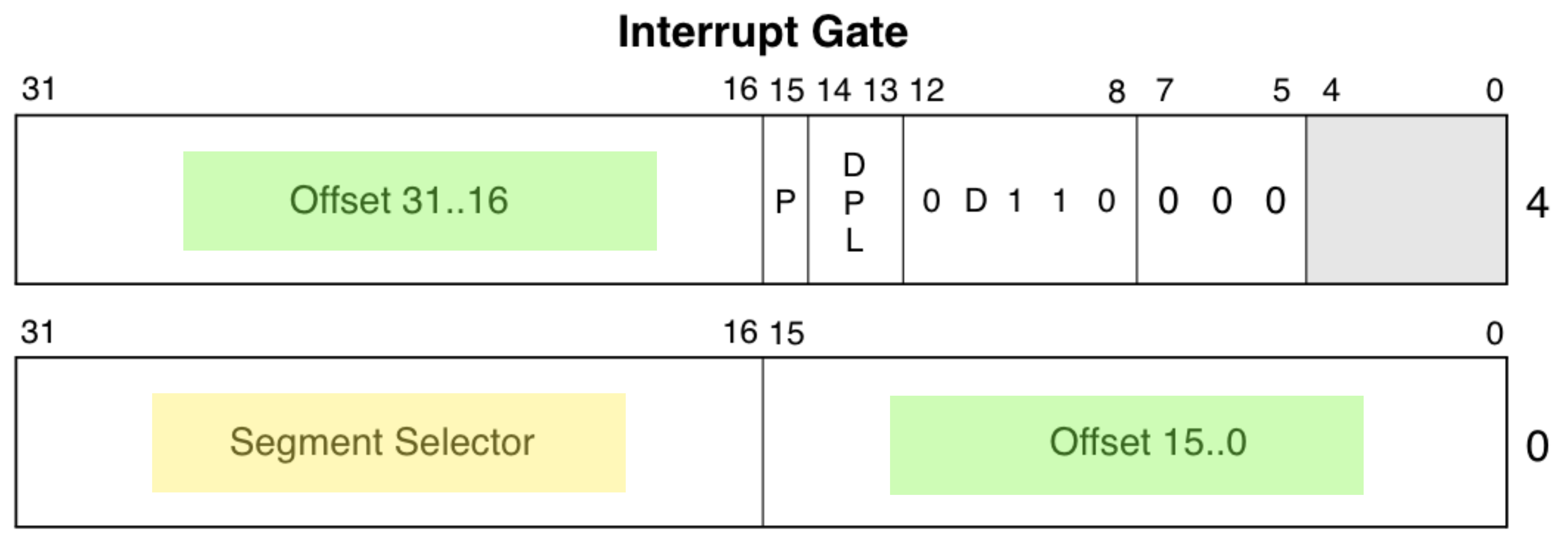


Figure 6-1. Relationship of the IDTR and IDT

Interrupt descriptor table (2)

- Each IDT entry is 64-bits. Contains code segment and eip
- When interrupt appears, hardware changes CS and EIP to the one pointed by IDT entry



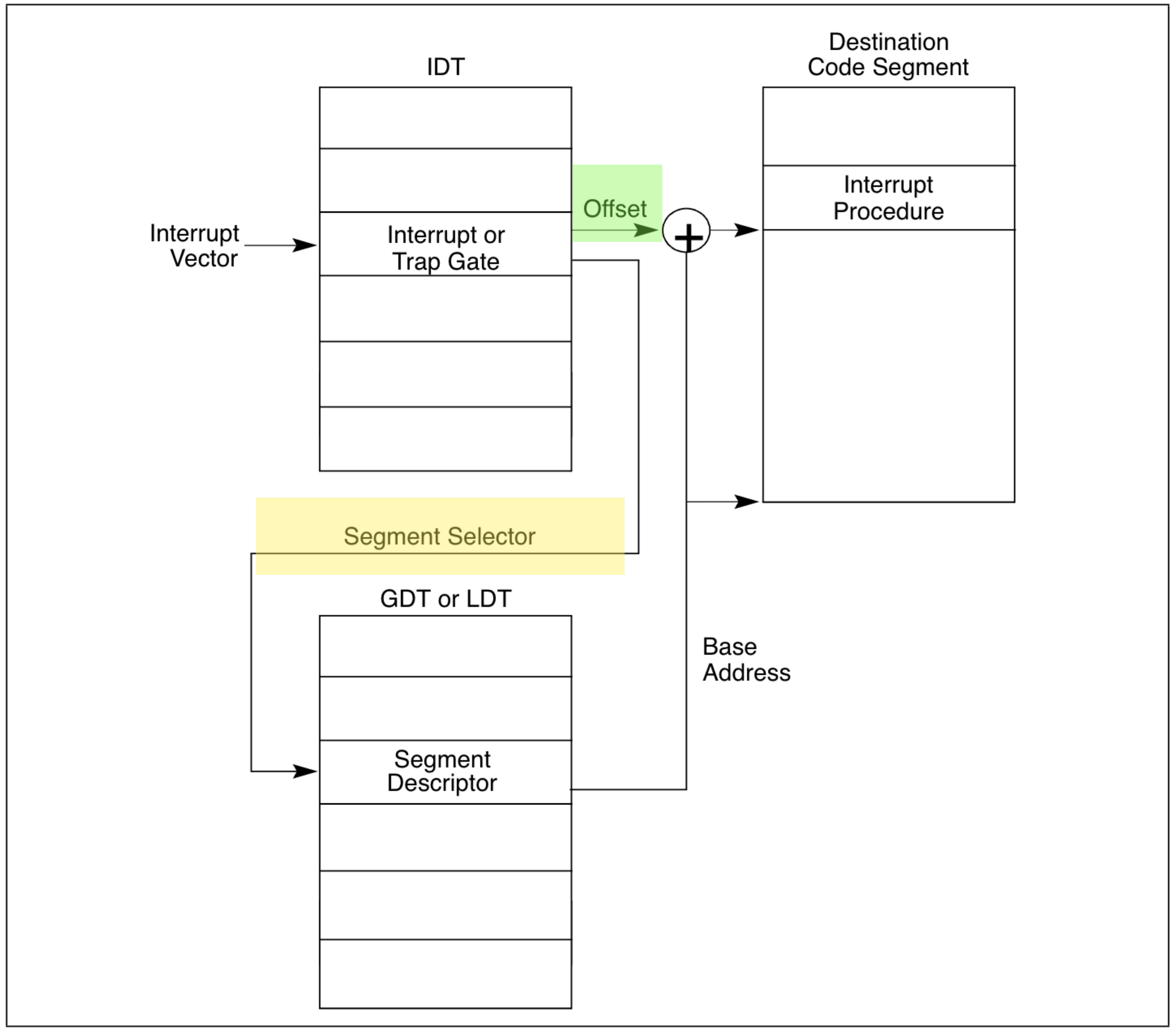
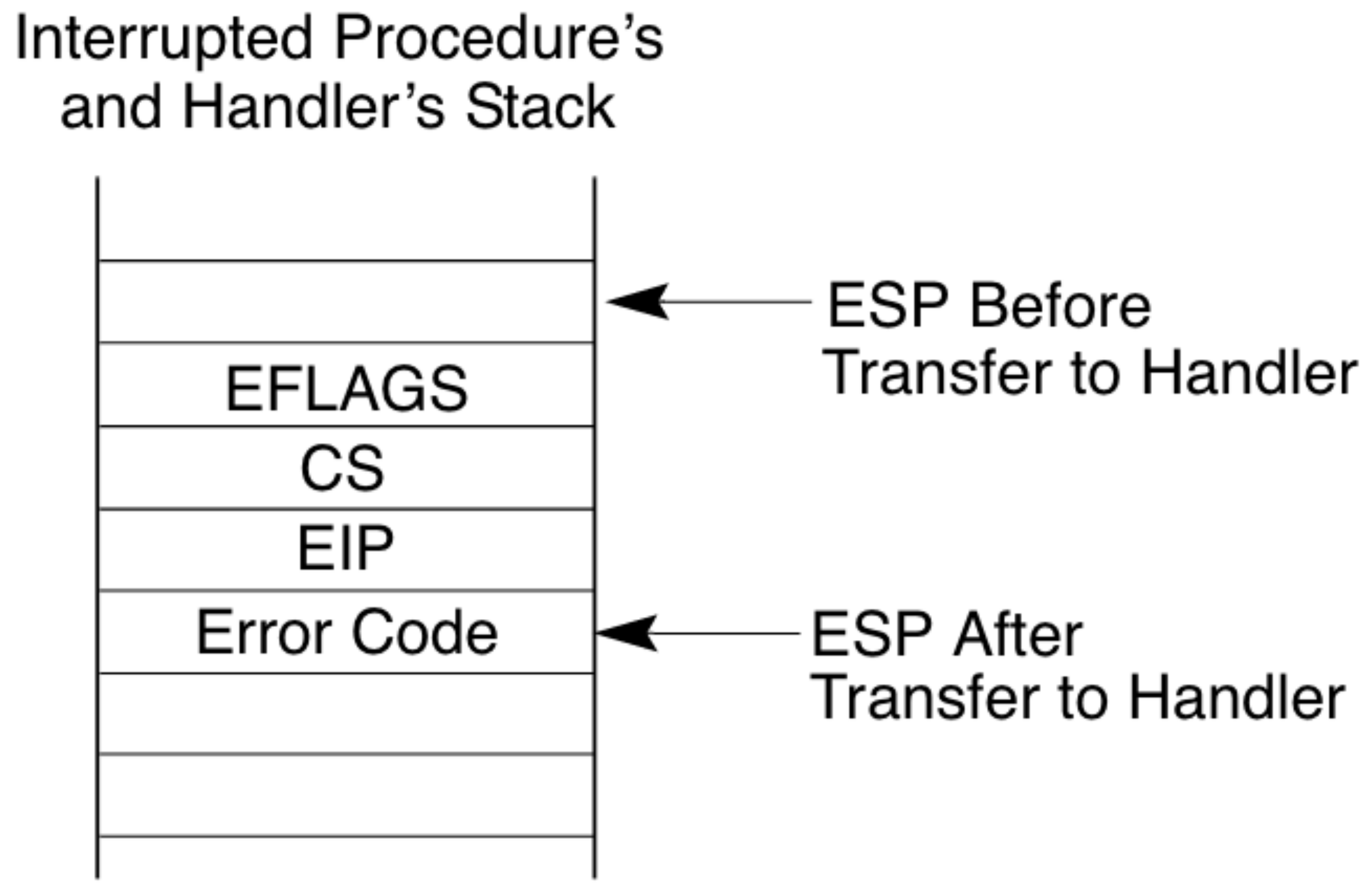


Figure 6-3. Interrupt Procedure Call

Interrupt handling

- On an interrupt, hardware pushes old EFLAGS, CS and EIP on the stack
- Jumps CS and EIP according to IDT
- IRET instruction (similar to RET instruction) restores CS, EIP, EFLAGS, ESP
- Interrupt handler may push more registers, like eax etc. on the stack.


Interrupted Procedure's and Handler's Stack
EFLAGS
CS
EIP
Error Code
ESP Before Transfer to Handler
ESP After Transfer to Handler

Code walkthrough

p4-traps

- vectors.pl creates 256 IDT entries. ‘i’th entry write ‘i’ on top of the stack and jumps to ‘alltraps’
- main.c calls tvinit and idtinit to setup interrupt descriptor table to populate the 256 entries and point IDTR to IDT
- ‘alltraps’ in trapasm.S runs ‘pushal’ to save general purpose registers. Then it calls ‘trap’ with the trapframe.
- ‘trap’ in ‘trapasm.S’ reads trapno saved by vectors.S to find out which interrupt occurred. It handles timer and spurious interrupts. It signals EOI to LAPIC when it is done with interrupt.
- trapasm recovers registers with popal, backs up esp above err code and trap number, executes iret to jump back to whatever OS was doing earlier

Visualizing interrupt handling

eip

```
for(;;)
;
```

vectors.S

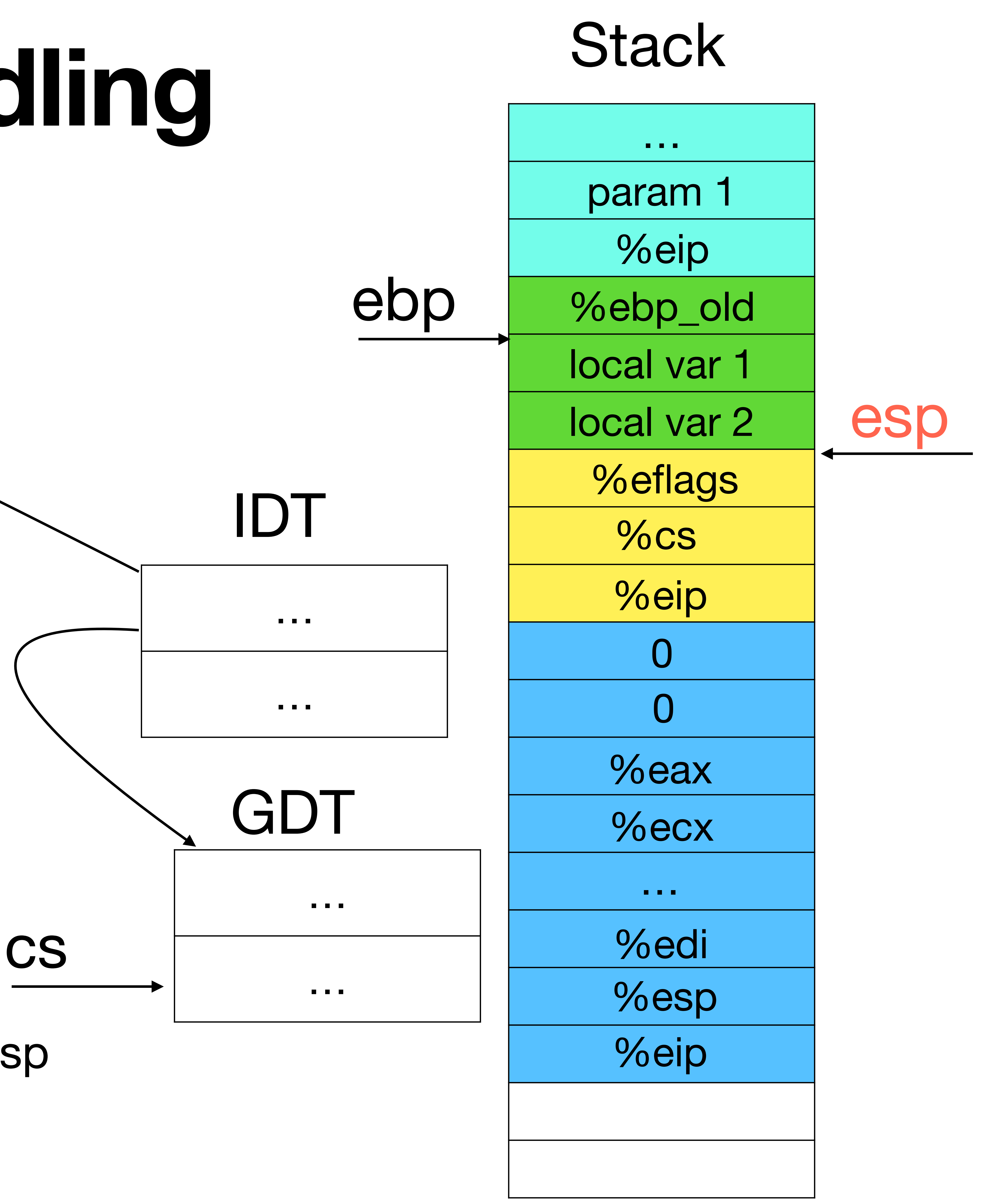
```
.globl vector0
vector0:
  pushl $0
  pushl $0
  jmp alltraps
```

trap.c

```
void
trap(struct trapframe *tf)
{
  switch(tf->trapno){
  case T_IRQ0 + IRQ_TIMER:
    ticks++;
    cprintf("Tick! %d\n\0", ticks);
    lapiceoi();
..
return
```

trapasm.S

```
alltraps:
  pushal
  pushl %esp
  call trap
  addl $4, %esp
  popal
  addl $0x8, %esp
  iret
```



Stack
...
param 1
%eip
ebp
%ebp_old
local var 1
local var 2
esp
%eflags
%cs
%eip
0
0
%eax
%ecx
...
%edi
%esp
%eip
IDT
...
...
GDT
...
...
CS