

MIT6.828 Lab3: User Environments

Zhuofan Zhang

Jan 2022

Contents

1	User Environments and Exception Handling	1
1.1	Environment State	1
1.2	Allocating the Environments Array	2
1.3	Creating and Running Environments	3
1.4	Handling Interrupts and Exceptions	12
1.5	Basics of Protected Control Transfer	13
1.6	Types of Exceptions and Interrupts	14
1.7	Nested Exceptions and Interrupts	15
1.8	Setting Up the IDT	15
2	Page Faults, Breakpoints Exceptions, and System Calls	19
2.1	Handling Page Faults	19
2.2	The Breakpoint Exception	20
2.3	System calls	21
2.4	User-mode startup	24
2.5	Page faults and memory protection	25

Chapter 1

User Environments and Exception Handling

本次 Lab 的第一部分主要处理 JOS 对进程的抽象及异常处理两部分的内容。

1.1 Environment State

JOS 对进程（Process/Environment）的抽象位于 `inc/env.h` 及 `kern/env.c` 中，包括结构体 `struct Env` 及一系列接口。系统使用三个全局变量：`envs`, `curenv`, `env_free_list` 对所有用户进程及当前进程进行管理。

对于每一个用户进程，JOS 使用 `struct Env` 表示：

```
struct Env {
    struct Trapframe env_tf; // Saved registers
    struct Env *env_link;    // Next free Env
    envid_t env_id;          // Unique environment identifier
    envid_t env_parent_id;   // env_id of this env's parent
    enum EnvType env_type;   // Indicates special system environments
    unsigned env_status;     // Status of the environment
    uint32_t env_runs;       // Number of times environment has run

    // Address space
    pde_t *env_pgdir;        // Kernel virtual address of page dir
};
```

其中，当空闲 `env` 被放入 `env_free_list` 中时依靠 `env_link` 构建链表。

JOS 的 Environment 组成与 `*nix` 系统相似，由 **Thread** 和 **Address Space** 两部分概念组成：前者由 `env_tf` 中的寄存器描述（即进程切换时需保存的现场），后者由 `env_pgdir` 指向的页表目录描述。

注：*JOS* 与 *xv6* 设计存在差异。*JOS* 采用的是 *Single Kernel Stack* 的设计，一次只能有一个进程陷入内核；而 *xv6* 的每个进程都拥有独立的内核栈（*xv6* 的进程由 `struct proc` 描述）。

在上述 Env 结构体中，需要注意用于记录进程当前状态的 env_status 变量。在 JOS 的设计中进程总共有 5 种可能的状态：

Status	Annotation
ENV_FREE	表示进程尚未被分配，此时应处于 env_free_list 中
ENV_RUNNABLE	表示进程已被分配且可在下次进程切换时被调度
ENV_RUNNING	表示为当前执行中的进程
ENV_NOT_RUNNABLE	表示进程活跃但不可调度：可能在等待 IPC 等状态
ENV_DYING	Zombie 进程。详细内容将在 Lab4 展开

1.2 Allocating the Environments Array

Exercise 1

Modify mem_init() in kern/pmap.c to allocate and map the envs array. This array consists of exactly NENV instances of the Env structure allocated much like how you allocated the pages array. Also like the pages array, the memory backing envs should also be mapped user read-only at UENVS (defined in inc/memlayout.h) so user processes can read from this array.

You should run your code and make sure check_kern_pgdir() succeeds.

第一个练习要求我们为 envs 数组分配物理空间，并将其映射至内核地址空间。分配方法与 Lab2 中分配与页管理数组 pages 方式接近，我们根据 memlayout.h 文件的提示，在 mem_init() 中增加如下代码：

```
////////////////////////////////////
// Make 'envs' point to an array of size 'NENV' of 'struct Env'.
// LAB 3: Your code here.
envs = (struct Env *) boot_alloc(NENV * sizeof(struct Env));

...
////////////////////////////////////
// Map the 'envs' array read-only by the user at linear address UENVS
// (ie. perm = PTE_U | PTE_P).
// Permissions:
//   - the new image at UENVS -- kernel R, user R
//   - envs itself -- kernel RW, user NONE
// LAB 3: Your code here.
boot_map_region(kern_pgdir, UENVS, PTSIZE, PADDR(envs), PTE_U);
```

1.3 Creating and Running Environments

本节的内容是构建进程管理的 API，包括实现进程创建分配、加载可运行镜像的功能。

Exercise 2

In the file `env.c`, finish coding the following functions:

`env_init()`

`env_setup_vm()`

`region_alloc()`

`load_icode()`

`env_create()`

`env_run()`

As you write these functions, you might find the new `cprintf` verb `%e` useful – it prints a description corresponding to an error code. For example,

```
r = -E_NO_MEM;
```

```
panic("env_alloc: %e", r);
```

will panic with the message "env_alloc: out of memory".

`env_init`

`env_init` 函数是实现对已在 `mem_init` 中分配的 `envs` 数据进行初始化，并将未分配进程放入空闲列表中。与页管理的 `pages` 数组及其空闲列表的管理模式相同。需要注意的是，实验对进程放入空闲列表的顺序有要求，根据实验提示完成即可：

```
void
env_init(void)
{
    assert(envs != NULL); // Make sure the envs is allocated successfully
    assert(env_free_list == NULL);
    env_free_list = envs;
    envs[0].env_id = 0;
    envs[0].env_link = NULL; // Not necessary
    envs[0].env_status = ENV_FREE;
    for(int i = 1; i < NENV; ++i)
    {
        envs[i].env_id = 0;
        envs[i-1].env_link = &envs[i];
        envs[i-1].env_status = ENV_FREE;
    }
    envs[NENV-1].env_link = NULL;
    env_init_percpu();
}
```

env_setup_vm

`env_setup_vm` 函数是在分配新进程时，设置进程的内核部分地址空间的函数。所有进程高位处（即内核地址映射）均是相同的，根据提示，可以使用已经构建完整映射的内核页目录 `kern_pgdir` 作为模板构建（可直接使用 `memcpy`，注意地址转换）。此外，根据实验提示，需将本页目录自身映射到地址空间的 UVPT 处。实现如下：

```
static int
env_setup_vm(struct Env *e)
{
    struct PageInfo *p = NULL;

    // Allocate a page for the page directory
    if (!(p = page_alloc(ALLOC_ZERO)))
        return -E_NO_MEM;

    // Now, set e->env_pgdir and initialize the page directory.
    //
    // Hint:
    //   - The VA space of all envs is identical above UTOP
    //     (except at UVPT, which we've set below).
    //   See inc/memlayout.h for permissions and layout.
    //   Can you use kern_pgdir as a template? Hint: Yes.
    //   (Make sure you got the permissions right in Lab 2.)
    //   - The initial VA below UTOP is empty.
    //   - You do not need to make any more calls to page_alloc.
    //   - Note: In general, pp_ref is not maintained for
    //     physical pages mapped only above UTOP, but env_pgdir
    //     is an exception -- you need to increment env_pgdir's
    //     pp_ref for env_free to work correctly.
    //   - The functions in kern/pmap.h are handy.

    // LAB 3: Your code here.
    e->env_pgdir = page2kva(p);
    memcpy(e->env_pgdir, kern_pgdir, PGSIZE);

    // UVPT maps the env's own page table read-only.
    // Permissions: kernel R, user R
    e->env_pgdir[PDX(UVPT)] = PADDR(e->env_pgdir) | PTE_P | PTE_U;

    // increment the pp_ref
    p->pp_ref++;

    return 0;
}
```

region_alloc

当进程申请物理空间并请求将其映射到要求的虚拟地址时调用该函数。注意请求地址及请求内存长度可以是非页表大小对齐的，因此实际分配前需完成对齐。

注：根据提示，该函数仅为 *load_icode* 映射 *ELF* 内容至进程虚拟空间调用

```
static void
region_alloc(struct Env *e, void *va, size_t len)
{
    // LAB 3: Your code here.
    // (But only if you need it for load_icode.)
    //
    // Hint: It is easier to use region_alloc if the caller can pass
    //      'va' and 'len' values that are not page-aligned.
    //      You should round va down, and round (va + len) up.
    //      (Watch out for corner-cases!)
    uintptr_t start = ROUNDDOWN((uintptr_t)(va), PGSIZE);
    uintptr_t end = ROUNDUP((uintptr_t)(va) + len, PGSIZE);

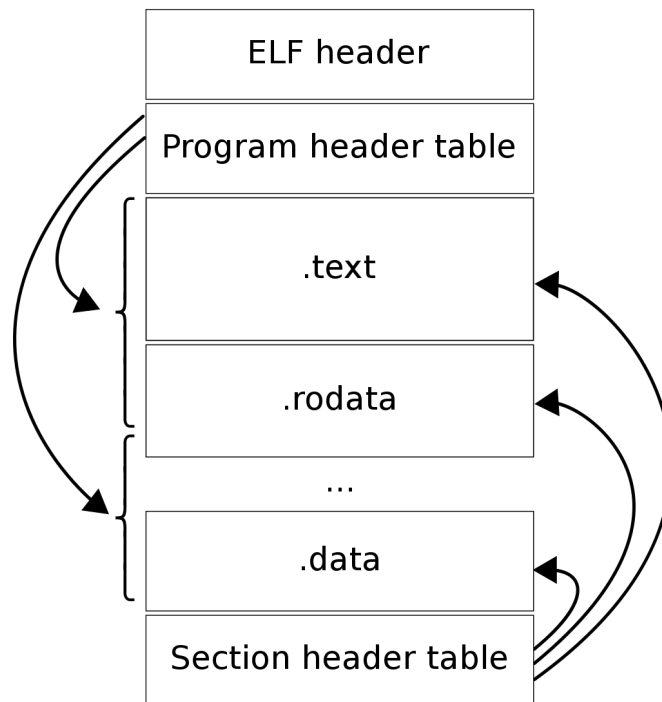
    if(!e)
        panic("region_alloc: env is NULL.\n");

    while(start < end)
    {
        struct PageInfo *p = page_alloc(0);
        if(!p)
            panic("region_alloc: page_alloc failed.\n");
        page_insert(e->env_pgdir, p, (void *)start, PTE_U | PTE_W);
        start += PGSIZE;
    }
}
```

load_icode

load_icode() 函数实现从 ELF 文件中加载信息，即实现 exec 的核心内容。

ELF 格式的可执行文件的文件结构如下图所示，其中与文件内容到运行时虚拟内存空间映射相关的是 Program header table 部分：它提供了每个部分到虚拟内存空间的映射关系。Program header table 在 ELF 文件中位置由 e_phoff 给出，条目数目由 e_phnum 给出。



代码中还有几点需要注意：(1) 由于映射时需使用 memcpy/memset 函数，两个函数在地址翻译时使用的是当前 CR3 寄存器所指定的页表，因此需临时切换至当前进程的页目录，完成映射后再切回内核页目录；(2) 注意到 ELF 镜像通过一个指针传递，我们可以查看 load_icode 的调用链，发现其由 env_create 负责调用，而 env_create 调用时使用 env.h 中定义的宏 ENV_PASTE3，将调用的位置设置为 _binary_obj_[NAME]_start，其中 NAME 即为二进制程序的名字。关于这一系列变量的解释在 Lab 中也有给出：这是一种将二进制可执行文件嵌入内核代码的一种方式。

具体地，在 kern/Makeflag 我们可以看到内核链接部分的 flags 中有 -b binary 选项，这个选项意味着在链接过程中将文件按 raw-format（理论上可以是任意二进制格式）链接而不按.o 形式解析，被原封不动嵌入可执行文件中。此外，链接器还会为这些嵌入的二进制文件引入符号注明其位置，即上文提到的 _binary_obj_[NAME]_start 的形式：JOS 使用 nm（打印二进制文件中的符号表）命令将其输出到了 kernel.sym 文件中以便使用者查阅，我们也可以看到这些符号。因此，当内核被加载后，这些被嵌入的 raw-ELF 也被加载到内存中，并且其位置由上述链接器提供的全局符号标志，故 load_icode（以及后文出现的 env_create()）可以直接使用这些全局符号找到这些嵌入的 ELF。


```

static void
load_icode(struct Env *e, uint8_t *binary)
{
    // LAB 3: Your code here.
    struct Elf *elf = (struct Elf *)binary;
    if(elf->e_magic != ELF_MAGIC)
        panic("load_icode: e_magic != ELF_MAGIC.\n");

    struct Proghdr *pht = (struct Proghdr *)(binary + elf->e_phoff);

    lcr3(PADDR(e->env_pgdir)); // for valid-use of memcpy/memset
    for(int i = 0; i < elf->e_phnum; ++i)
    {
        if(pht[i].p_type == ELF_PROG_LOAD)
        {
            // Allocate physical regions
            region_alloc(e, (void *)pht[i].p_va, pht[i].p_memsz);
            // Copy the contents from ELF
            memcpy(
                (void *)pht[i].p_va,
                (void *) (binary) + pht[i].p_offset,
                pht[i].p_filesz
            );
            // Set the rest to be zero
            if(pht[i].p_filesz < pht[i].p_memsz)
                memset(
                    (void *)pht[i].p_va + pht[i].p_filesz,
                    0,
                    pht[i].p_memsz - pht[i].p_filesz
                );
        }
    }
    lcr3(PADDR(kern_pgdir));

    // Setup the entry
    e->env_tf.tf_eip = elf->e_entry;

    // Now map one page for the program's initial stack
    // at virtual address USTACKTOP - PGSIZE.
    // LAB 3: Your code here.
    region_alloc(e, (void *) (USTACKTOP - PGSIZE), PGSIZE);
}

```

env_create & env_run

env_create 用于创建新进程并载入二进制程序，env_run 功能则为切换当前执行的进程。代码实现如下：

```
void
env_create(uint8_t *binary, enum EnvType type)
{
    // LAB 3: Your code here.
    struct Env *new_env = NULL;
    if(env_alloc(&new_env, 0) < 0)
        panic("env_create: env_alloc failed.\n");
    new_env->env_type = type;
    load_icode(new_env, binary);
}

void
env_run(struct Env *e)
{
    // Step 1: If this is a context switch (a new environment is running):
    //   1. Set the current environment (if any) back to
    //       ENV_RUNNABLE if it is ENV_RUNNING (think about
    //       what other states it can be in),
    //   2. Set 'curenv' to the new environment,
    //   3. Set its status to ENV_RUNNING,
    //   4. Update its 'env_runs' counter,
    //   5. Use lcr3() to switch to its address space.
    // Step 2: Use env_pop_tf() to restore the environment's
    //   registers and drop into user mode in the
    //   environment.
    // ...
    // LAB 3: Your code here.
    if(curenv)
    {
        if(curenv->env_status == ENV_RUNNING)
            curenv->env_status = ENV_RUNNABLE;
    }
    curenv = e;
    curenv->env_status = ENV_RUNNING;
    curenv->env_runs++;
    // !Use physic addr of pgdir
    lcr3(PADDR(curenv->env_pgdir));
    env_pop_tf(&(curenv->env_tf));
}
```

在上述部分代码完成后，Lab 让我们实现一个调试：首先在 `env_pop_tf` 处设置断点后执行，通过查阅 `kernel` 源码我们知道当前内核启动后会创建第一个用户进程 `hello` 并使用 `env_run` 运行它，`env_pop_tf` 就发生在这次调用中。当 `env_pop_tf` 完成后，执行流进入用户进程的运行状态。此时 Lab 要求我们在 `hello` 进程中的 `int $0x30` 处打断点并执行，再单步执行后可发现内核发生 `fault`。

```
[00000000] new env 00001000
EAX=00000000 EBX=00000000 ECX=0000000d EDX=eebfde78
ESI=00000000 EDI=00000000 EBP=eebfde20 ESP=eebfdde8
EIP=00800f06 EFL=00000016 [---AP-] CPL=3 II=0 A20=1 SMM=0 HLT=0
ES =0023 00000000 ffffffff 00cff300 DPL=3 DS [-WA]
CS =001b 00000000 ffffffff 00cffa00 DPL=3 CS32 [-R-]
SS =0023 00000000 ffffffff 00cff300 DPL=3 DS [-WA]
DS =0023 00000000 ffffffff 00cff300 DPL=3 DS [-WA]
FS =0023 00000000 ffffffff 00cff300 DPL=3 DS [-WA]
GS =0023 00000000 ffffffff 00cff300 DPL=3 DS [-WA]
LDT=0000 00000000 00000000 00008200 DPL=0 LDT
TR =0028 f0191b20 00000067 00408900 DPL=0 TSS32-avl
GDT= f011e500 0000002f
IDT= f0191320 000007ff
CR0=80050033 CR2=00000000 CR3=003bc000 CR4=00000000
DR0=00000000 DR1=00000000 DR2=00000000 DR3=00000000
DR6=ffff0fff DR7=00000400
EFER=0000000000000000
Triple fault. Halting for inspection via QEMU monitor.
```

查看 `hello.c` 与 `hello.asm` 并结合内核打印的信息我们可以得知该 `fault` 在打印“hello world!”语句时发生；检查调用链可知，`vcprintf` 在调用 `sys_cputs()` 时产生了一个 `syscall`，该 `syscall` 调用出现问题的 `int $0x30` 指令。由此我们基本可以确认，这个 `fault` 是由于缺少处理 `syscall` 的中断处理程序而产生的。

此处我们还需厘清一个问题是：JOS 是在哪里完成对用户态/内核态的区分的？也即，我们分配的进程在哪里被设置为用户态程序？

我们知道在 `x86` 中，一段内容处于 `Ring x` 记录在当前 `CS` 寄存器中的 `CPL` 位，这一位通常与该段内容所处段的段选择子 `DPL` 位相等，因此设置一段内容的 `Ring x` 权限，必然涉及 `GDT` 的设置。我们使用 IDE 的搜索功能查询 `GDT` 相关关键字，很快就能找到相关代码：`env_init_percpu()`：

```

// Load GDT and segment descriptors.
void
env_init_percpu(void)
{
    lgdt(&gdt_pd);
    // The kernel never uses GS or FS, so we leave those set to
    // the user data segment.
    asm volatile("movw %%ax,%%gs" : : "a" (GD_UD|3));
    asm volatile("movw %%ax,%%fs" : : "a" (GD_UD|3));
    // The kernel does use ES, DS, and SS. We'll change between
    // the kernel and user data segments as needed.
    asm volatile("movw %%ax,%%es" : : "a" (GD_KD));
    asm volatile("movw %%ax,%%ds" : : "a" (GD_KD));
    asm volatile("movw %%ax,%%ss" : : "a" (GD_KD));
    // Load the kernel text segment into CS.
    asm volatile("ljmp %0,$1f\n 1:\n" : : "i" (GD_KT));
    // For good measure, clear the local descriptor table (LDT),
    // since we don't use it.
    lldt(0);
}

```

这段代码使用 `gdt_pd` 变量的内容更新了当前的 GDT，同时设置了当前内核代码的 CS，ES，DS，SS 寄存器为 GD_KD 段选择子的位置。我们查看 `gdt_pd` 的内容：

```

struct Segdesc gdt[] =
{
    // 0x0 - unused (always faults -- for trapping NULL far pointers)
    SEG_NULL,
    // 0x8 - kernel code segment
    [GD_KT >> 3] = SEG(STA_X | STA_R, 0x0, 0xffffffff, 0),
    // 0x10 - kernel data segment
    [GD_KD >> 3] = SEG(STA_W, 0x0, 0xffffffff, 0),
    // 0x18 - user code segment
    [GD_UT >> 3] = SEG(STA_X | STA_R, 0x0, 0xffffffff, 3),
    // 0x20 - user data segment
    [GD_UD >> 3] = SEG(STA_W, 0x0, 0xffffffff, 3),
    // 0x28 - tss, initialized in trap_init_percpu()
    [GD_TSS0 >> 3] = SEG_NULL
};

struct Pseudodesc gdt_pd = {
    sizeof(gdt) - 1, (unsigned long) gdt
};

```

从上述代码我们就可以看到，新的 GDT 仍然采用 flat-pattern 的方式来忽略分段机制，但利用了段选择子的权限位设置实现内核态和用户态的权限区分（内核段为 Ring0，代码段为 Ring3）。所有用户态进程均有内核分配及初始化，因此内核的权限设置工作应当也由内核代码实现。同样进行搜索后发现，在分配进程的函数 `env_alloc()` 中就进行了相应设置，由此我们便可以理解操作系统如何提供运行在 Ring3 的用户态进程了：

```
int
env_alloc(struct Env **newenv_store, env_id_t parent_id)
{
    ...
    // Set up appropriate initial values for the segment registers.
    // GD_UD is the user data segment selector in the GDT, and
    // GD_UT is the user text segment selector (see inc/memlayout.h).
    // The low 2 bits of each segment register contains the
    // Requestor Privilege Level (RPL); 3 means user mode. When
    // we switch privilege levels, the hardware does various
    // checks involving the RPL and the Descriptor Privilege Level
    // (DPL) stored in the descriptors themselves.
    e->env_tf.tf_ds = GD_UD | 3;
    e->env_tf.tf_es = GD_UD | 3;
    e->env_tf.tf_ss = GD_UD | 3;
    e->env_tf.tf_esp = USTACKTOP;
    e->env_tf.tf_cs = GD_UT | 3;
    ...
}
```

1.4 Handling Interrupts and Exceptions

Exercise 3

Read Chapter 9, Exceptions and Interrupts in the 80386 Programmer's Manual (or Chapter 5 of the IA-32 Developer's Manual), if you haven't already.

Interrupts & Exceptions

通常情况下，外部中断（**External Interrupts**）可以被分为两类：中断（**Interrupts**）和异常（**Exceptions**）。其中两者可以被进一步细分。

- **中断（Interrupts）**：分为可屏蔽中断（**Maskable Interrupts**）与不可屏蔽中断（**Nonmaskable Interrupts**），前者由 CPU 的 INTR 引脚发出信号产生，后者由 NMI 引脚产生。一般来说不可屏蔽中断都由致命错误引发，通常的处理方式不是为其设置 Handler，而是直接发出错误警告终止程序。

一般认为中断信号是异步信号，由 CPU 外设备产生。

- **异常（Exceptions）**：异常通常分为 (1) 由 CPU 自行检测到的异常，会被进一步分类为 trap/fault/abort，例如 divide zero 异常；(2) 还有部分信号是可编程的，例如 INT 0x3 指令触发，也被称为软中断（**Software Interrupts**）。

与中断相对的，异常信号通常为同步信号，即由当前 CPU 运行中的代码产生的。

Enabling and Disabling Interrupts

中断与异常屏蔽方式有以下几种：

- **NMI Masks Further NMIs:**

当一个 NMI 处理程序进行中时，其他的 NMI 会被暂时屏蔽，直到处理程序执行结束。

- **IF Masks INTR:**

通过设置 IF(interrupt-enable flag) 可以开启/关闭可屏蔽中断。x86 提供 CLI/STI 指令设置该位（bootloader.S 中使用过该指令）。

- **RF Masks Debug Faults:**

设置 EFLAGS 中的 RF 位可以开启/关闭 debug fault。

- **MOV or POP to SS Masks Some Interrupts and Exceptions:**

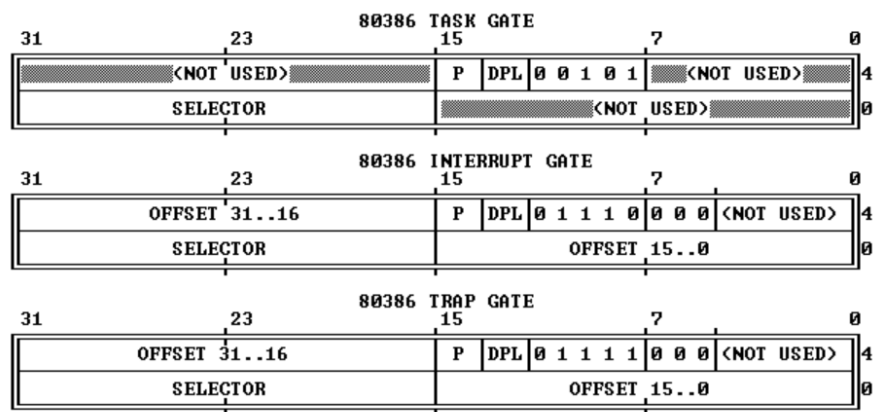
这是一类应对不正常行为的中断屏蔽。如果在修改 SS 寄存器与 ESP 寄存器之间产生了中断，会产生 SS:ESP 的不一致性，可能影响中断处理程序的行为。因此，在修改 SS 寄存器时（MOV/POP）80386 架构会禁止 NMI、INTR 及异常中断，仅有 page fault 与 general protection fault 会发生。改为使用 LSS 指令可以避开这个问题处理。

1.5 Basics of Protected Control Transfer

为实现从用户态到内核态切换的同时保证对用户态可执行内容的限制，x86 提供了两种机制：中断向量表（**Interrupt Descriptor Table, IDT**）及任务状态段（**Task State Segment, TSS**）。

Interrupt Description Table (IDT)

IDT 的形式与 GDT 类似，地址存储在专门的寄存器 IDTR 中。IDT 提供一个 256 entries 的表记录处理不同类型中断/异常的函数入口，当异常发生时调用对应的函数（加载 handler 的 EIP/CS）。IDT 中包含三种不同的表项，如下图所示：



Task State Segment (TSS)

当调用中断处理程序时，需要我们保存当前执行流的现场以便完成异常处理后恢复执行。而且应当注意的是，执行流现场必须保存到高特权级的位置，否则可能遭到其他用户态带 bug 或恶意的代码的影响。

当发生中断/异常且 handler 向更高特权级转换时，x86 会切换当前使用的栈到高特权级栈（内核栈）。同时，x86 定义了 TSS 作为保存现场的数据结构，发生异常时将记录执行流相关信息的 TSS 压入内核栈中。

TSS 结构保存着大量信息，JOS 仅使用 ESP0 及 SS0 两个域，用于指定切换的内核栈位置。TSS 的地址被保存在 TR 寄存器中，可以通过 LTR/STR 指令进行操作。JOS 在 trap_init_percpu() 中执行了 TSS 段的初始化加载。

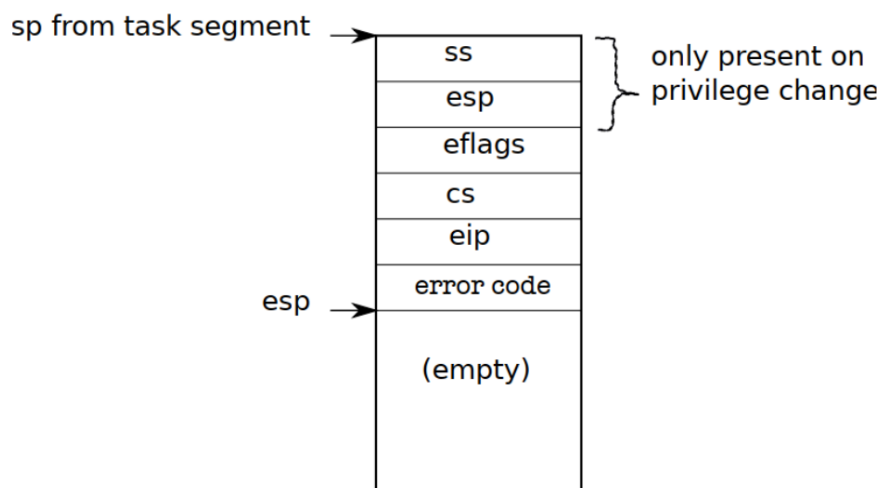
1.6 Types of Exceptions and Interrupts

在 x86 中，所有的（同步）异常被分配到 0-31 号中断向量上，即 IDT 的前 32 个 entries，例如 `page-fault` 就使用 14 号中断向量。

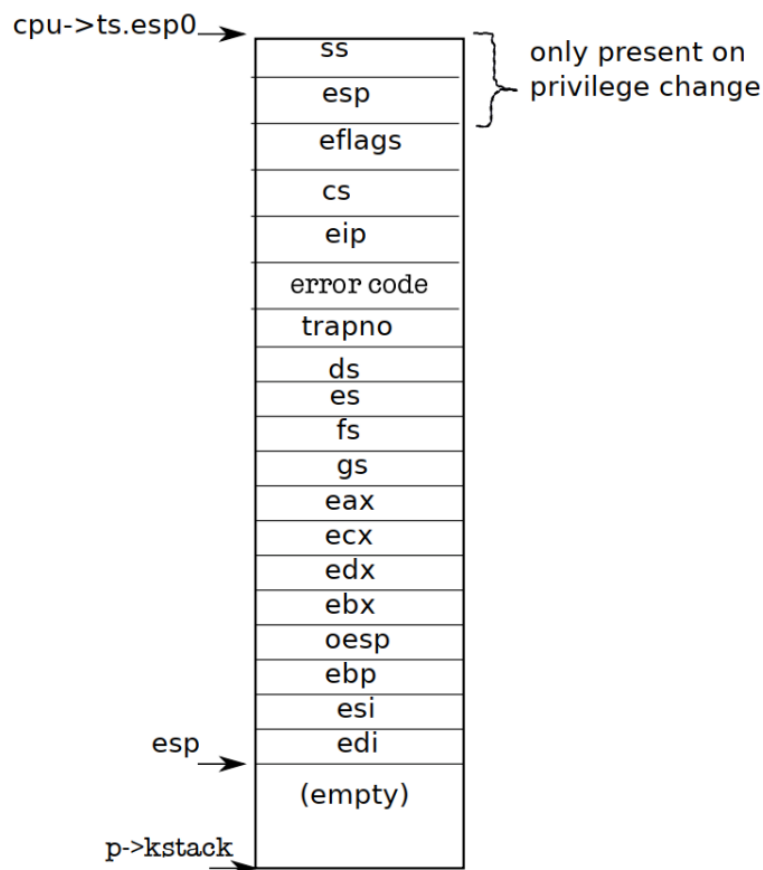
大于 31 号的中断向量被用于处理软件中断，可能的来源是 `int` 指令发送的中断及外部设备中断等异步中断。在 Lab3 中，我们会实现 JOS 对 0-31 号所有同步异常处理入口的设置以及系统调用中断的设置（JOS 使用了中断号 0x30，这个选择是可任意的）；在 Lab4 中会进一步拓展至可以处理时钟中断等外部中断。x86 中发射中断向量的指令是 `int n` 指令，它完成以下任务：

- 根据索引 `n` 在 IDT 中找到处理函数的入口；
- 检查 CS 寄存器的权限位（CPL）；
- 将当前 ESP 寄存器及 SS 寄存器存放入 CPU 内部空闲寄存器；
- 从 TSS 中加载内核栈的 ESP 寄存器及 SS 寄存器；
- 将原有的 ESP 寄存器、SS 寄存器（两者当前暂时存放于 CPU 其他寄存器中）等入内核栈；
- 将 EFLAGS 寄存器、当前进程的 CS 寄存器及 IP 寄存器入（内核栈）；
- 将 CPU 控制交给处理函数（处理函数负责调用 `iret` 命令从中断处理中返回）。

到这里，`int` 指令已经完成了填充内核栈信息的一部分工作，内核栈内容如下图所示：



为进一步填充完整信息（形成 `TrapFrame`），`trap` 入口函数提供进一步操作，如下图所示。JOS 的做法是在每个入口函数将 `trapno` 入栈后，统一进入 `_alltrap` 函数处理（Exercise 4 内容）。



1.7 Nested Exceptions and Interrupts

当已经处于内核态时，也可以继续嵌套处理中断和异常，此时会在内核栈中继续装填处理帧，唯一不同的是不会再将 `SS` 和 `ESP` 寄存器入栈。

1.8 Setting Up the IDT

Exercise 4

Edit `trapentry.S` and `trap.c` and implement the features described above. The macros `TRAPHANDLER` and `TRAPHANDLER_NOEC` in `trapentry.S` should help you, as well as the `T_` defines in `inc/trap.h`. You will need to add an entry point in `trapentry.S` (using those macros) for each trap defined in `inc/trap.h`, and you'll have to provide `_alltraps` which the `TRAPHANDLER` macros refer to. You will also need to modify `trap_init()` to initialize the `idt` to point to each of these entry points defined in `trapentry.S`; the `SETGATE` macro will be helpful here.

Test your trap handling code using some of the test programs in the user directory that cause exceptions before making any system calls, such as `user/divzero`. You should be able to get make grade to succeed on the `divzero`, `softint`, and `badsegment` tests at this point.

Part A 最后一部分要求修改 `trapentry.S` 及 `trap.c` 文件，对 IDT 进行初始化。`trapentry.S` 提供了两个宏：TRAPHANDLER 与 TRAPHANDLER_NOEC 用于生成处理函数入口（与 xv6 中的 `vectors.pl` 功能相近）。设置代码如下：

```
/*
 * Lab 3: Your code here for generating entry points for the different traps.
 */
TRAPHANDLER_NOEC(DIVIDE, T_DIVIDE)
TRAPHANDLER_NOEC(DEBUG, T_DEBUG)
TRAPHANDLER_NOEC(NMI, T_NMI)
TRAPHANDLER_NOEC(BRKPT, T_BRKPT)
TRAPHANDLER_NOEC(OFLOW, T_OFLOW)
TRAPHANDLER_NOEC(BOUND, T_BOUND)
TRAPHANDLER_NOEC(ILLOP, T_ILLOP)
TRAPHANDLER_NOEC(DEVICE, T_DEVICE)
TRAPHANDLER(DBLFLT, T_DBLFLT)
TRAPHANDLER(TSS, T_TSS)
TRAPHANDLER(SEGNP, T_SEGNP)
TRAPHANDLER(STACK, T_STACK)
TRAPHANDLER(GPFLT, T_GPFLT)
TRAPHANDLER(PGFLT, T_PGFLT)
TRAPHANDLER_NOEC(FPERR, T_FPERR)
TRAPHANDLER(ALIGN, T_ALIGN)
TRAPHANDLER_NOEC(MCHK, T_MCHK)
TRAPHANDLER_NOEC(SIMDERR, T_SIMDERR)
TRAPHANDLER_NOEC(SYSCALL, T_SYSCALL)
TRAPHANDLER_NOEC(DEFAULT, T_DEFAULT)
```

观察两个宏的定义，发现设置后的入口均会跳转至 `__alltraps` 入口进行通用处理。这是所有处理函数的通用入口，它填充内核栈使其具有完整的 `Trapframe` 结构，并根据 Lab 提示加载 DS、ES 寄存器，并将 ESP 寄存器入栈。参考 xv6 源码，对 `__alltraps` 补充如下。需要注意的是，JOS 与 x86 的 `tf` 结构略有不同，需进行一定修改：使用了 `pushal` 指令，即 `push all general registers`：

```
.global __alltraps
__alltraps:
    pushl %ds
    pushl %es
    pushal
    movw $GD_KD, %ax
    movw %ax, %ds
    movw %ax, %es
    pushl %esp
    call trap
```

观察可知，这段代码最终跳转至 `trap.c/trap()`，由该函数进行分发与中断处理。参考 `xv6`，我们需要正确设置 `trap` 的入口使其可以跳转至 `handler`，也即设置 IDT。这一步骤在 `trap.c/trap_init()` 函数中完成：

```
void
trap_init(void)
{
    extern struct Segdesc gdt[];

    // LAB 3: Your code here.
    extern void DIVIDE();
    extern void DEBUG();
    ...
    extern void SYSCALL();
    extern void DEFAULT();

    SETGATE(idt[T_DIVIDE], 0, GD_KT, DIVIDE, 0);
    SETGATE(idt[T_DEBUG], 0, GD_KT, DEBUG, 0);
    SETGATE(idt[T_NMI], 0, GD_KT, NMI, 0);
    SETGATE(idt[T_BRKPT], 0, GD_KT, BRKPT, 3); // !
    SETGATE(idt[T_OFLOW], 0, GD_KT, OFLOW, 0);
    SETGATE(idt[T_BOUND], 0, GD_KT, BOUND, 0);
    SETGATE(idt[T_ILLOP], 0, GD_KT, ILLOP, 0);
    SETGATE(idt[T_DEVICE], 0, GD_KT, DEVICE, 0);
    SETGATE(idt[T_DBLFLT], 0, GD_KT, DBLFLT, 0);
    SETGATE(idt[T_TSS], 0, GD_KT, TSS, 0);
    SETGATE(idt[T_SEGNP], 0, GD_KT, SEGNP, 0);
    SETGATE(idt[T_STACK], 0, GD_KT, STACK, 0);
    SETGATE(idt[T_GPFLT], 0, GD_KT, GPFLT, 0);
    SETGATE(idt[T_PGFLT], 0, GD_KT, PGFLT, 0);
    SETGATE(idt[T_FPERR], 0, GD_KT, FPERR, 0);
    SETGATE(idt[T_ALIGN], 0, GD_KT, ALIGN, 0);
    SETGATE(idt[T_MCHK], 0, GD_KT, MCHK, 0);
    SETGATE(idt[T_SIMDERR], 0, GD_KT, SIMDERR, 0);
    SETGATE(idt[T_SYSCALL], 0, GD_KT, SYSCALL, 3); // !
    SETGATE(idt[T_DEFAULT], 0, GD_KT, DEFAULT, 0);
    // Per-CPU setup
    trap_init_percpu();
}
```

Question

Answer the following questions:

1. What is the purpose of having an individual handler function for each exception/interrupt? (i.e., if all exceptions/interrupts were delivered to the same handler, what feature that exists in the current implementation could not be provided?)
2. Did you have to do anything to make the user/softint program behave correctly? The grade script expects it to produce a general protection fault (trap 13), but softint's code says `int $14`. Why should this produce interrupt vector 13? What happens if the kernel actually allows softint's `int $14` instruction to invoke the kernel's page fault handler (which is interrupt vector 14)?

尝试回答上述问题。

1. 不同的中断对应不同的问题，应当采取不同的处理方式，很难进行统一化处理。

2. `user/softint` 程序原本应该产生 13 号中断（尽管其 `int` 指令发射的是 14 号中断），因为该程序运行在 Ring 3 权限级，而根据 IDT 的定义我们知道 14 号中断需要 `DPL=0`，因此在用户态发射 14 号中断将引发越权异常，从而触发 13 号中断。如果内核允许用户态代码使用 14 号 `Page Fault` 中断，可能导致有 bug 的或者恶意程序耗竭系统的内存资源导致崩溃。

Chapter 2

Page Faults, Breakpoints Exceptions, and System Calls

2.1 Handling Page Faults

Exercise 5

Modify `trap_dispatch()` to dispatch page fault exceptions to `page_fault_handler()`. You should now be able to get make grade to succeed on the `faultread`, `faultreadkernel`, `faultwrite`, and `faultwritekernel` tests. If any of them don't work, figure out why and fix them. Remember that you can boot JOS into a particular user program using *make run-x* or *make run-x-nox*. For instance, *make run-hello-nox* runs the hello user program.

这一节练习的内容较为简单，要求我们将 PageFault 异常分配给它对应的处理函数。我们仅需修改 `trap_dispatch` 函数即可：

```
static void
trap_dispatch(struct Trapframe *tf)
{
    // Handle processor exceptions.
    // LAB 3: Your code here.
    switch(tf->tf_trapno)
    {
        case T_PGFLT:
            page_fault_handler(tf);
            break;
    }
    ...
}
```

2.2 The Breakpoint Exception

Exercise 6

Modify `trap_dispatch()` to make breakpoint exceptions invoke the kernel monitor. You should now be able to get make grade to succeed on the breakpoint test.

这一步实验也十分简单，我们在上一个 Exercise 的代码基础上追加即可：

```
static void
trap_dispatch(struct Trapframe *tf)
{
    // Handle processor exceptions.
    // LAB 3: Your code here.
    switch(tf->tf_trapno)
    {
        case T_PGFLT:
            page_fault_handler(tf);
            break;
        case T_BRKPT:
            monitor(tf);
            break;
    }
    ...
}
```

Challenge!

Modify the JOS kernel monitor so that you can 'continue' execution from the current location (e.g., after the `int3`, if the kernel monitor was invoked via the breakpoint exception), and so that you can single-step one instruction at a time. You will need to understand certain bits of the EFLAGS register in order to implement single-stepping.

【待完成】

Question

3. The break point test case will either generate a break point exception or a general protection fault depending on how you initialized the break point entry in the IDT (i.e., your call to SETGATE from trap_init). Why? How do you need to set it up in order to get the breakpoint exception to work as specified above and what incorrect setup would cause it to trigger a general protection fault?

4. What do you think is the point of these mechanisms, particularly in light of what the user/softint test program does?

尝试回答上述问题。

3. 在 SETGATE 阶段如果将 DPL-bit 设置为 3 则触发 breakpoint，否则（DPL=0）则触发 general protection fault。这是因为 breakpoint 用户程序处在 Ring 3 权限级，若设置 DPL!=3 将导致其无权限发送 int \$3 指令，从而触发地址保护异常。

4. 上述所有的机制都是为用户态代码提供“有限”访问权限的保障。

2.3 System calls

Exercise 7

Add a handler in the kernel for interrupt vector T_SYSCALL. You will have to edit kern/trapentry.S and kern/trap.c's trap_init(). You also need to change trap_dispatch() to handle the system call interrupt by calling syscall() (defined in kern/syscall.c) with the appropriate arguments, and then arranging for the return value to be passed back to the user process in %eax. Finally, you need to implement syscall() in kern/syscall.c. Make sure syscall() returns -E_INVALID if the system call number is invalid. You should read and understand lib/syscall.c (especially the inline assembly routine) in order to confirm your understanding of the system call interface. Handle all the system calls listed in inc/syscall.h by invoking the corresponding kernel function for each call.

Run the user/hello program under your kernel (make run-hello). It should print "hello, world" on the console and then cause a page fault in user mode. If this does not happen, it probably means your system call handler isn't quite right. You should also now be able to get make grade to succeed on the testbss test.

这一节实验要补全系统调用的完整框架，我们需要理清 JOS 系统调用的设计。

用户态系统调用的接口由 lib/syscall.c 的 syscall() 函数提供，它会使用 int 指令发射 30 号中断。我们注意 syscall() 中内联汇编的写法：

```

// lib/syscall.c
static inline int32_t
syscall(int num, int check, uint32_t a1, uint32_t a2, uint32_t a3, uint32_t a4, uint32_t
    a5)
{
    ...
    asm volatile("int %1\n"
        : "=a" (ret)
        : "i" (T_SYSCALL),
          "a" (num),      // eax
          "d" (a1),       // edx
          "c" (a2),       // ecx
          "b" (a3),       // ebx
          "D" (a4),       // edi
          "S" (a5)        // esi
        : "cc", "memory");
    ...
}

```

查看 C 语言内联汇编的语法可知，该 `int` 命令发送时将 `syscall-no` 放入 EAX 寄存器中，并在其他寄存器中存放参数。`int` 命令发生后，根据前面的内容可知执行流将交给 `kern/trap.c` 中的 `trap()` 函数，并进一步交给 `trap_dispatch` 分发任务。我们需要在此处补全 `dispatch` 部分，将系统调用交给 `kern/syscall.c` 的 `syscall()` 处理（注意此处 `switch` 分支直接 `return`，不再执行后续内容）：

```

static void
trap_dispatch(struct Trapframe *tf)
{
    switch(tf->tf_trapno)
    {
        ...
        case T_SYSCALL:
            tf->tf_regs.reg_eax = syscall(
                tf->tf_regs.reg_eax,    // syscall-no
                tf->tf_regs.reg_edx,
                tf->tf_regs.reg_ecx,
                tf->tf_regs.reg_ebx,
                tf->tf_regs.reg_edi,
                tf->tf_regs.reg_esi
            );
            return;    // !
    }
    ...
}

```


然后我们进一步完善内核的 `syscall()` 函数，这个函数的主要功能是根据 `syscall-no` 将任务分发给真正的处理函数，并将参数传递给它们：

```
int32_t
syscall(uint32_t syscallno, uint32_t a1, uint32_t a2, uint32_t a3, uint32_t a4, uint32_t
    a5)
{
    // Call the function corresponding to the 'syscallno' parameter.
    // Return any appropriate return value.
    // LAB 3: Your code here.

    // panic("syscall not implemented");

    switch (syscallno)
    {
        case SYS_cputs:
            sys_cputs((char *)a1, a2);
            return 0;
        case SYS_cgetc:
            return sys_cgetc();
        case SYS_getenvid:
            return sys_getenvid();
        case SYS_env_destroy:
            return sys_env_destroy(a1);
        default:
            return -E_INVALID;
    }
}
```

至此，我们可以总结出自用户态调用JOS提供系统调用函数的完整流程：

1. 调用 `lib` 中的 `syscall` 函数，该函数使用 `int` 指令发送中断；
2. 程序陷入 JOS 内核，根据 IDT 中记录的函数，执行流交给 `trap` 函数；
3. `trap` 函数调用 `trap_dispatch` 分发任务；
4. `trap_dispatch` 将任务及函数参数传递给 `kern` 中的 `syscall` 函数；
5. `syscall` 函数进一步调用真正执行任务的函数（如 `sys_cputs`）。

2.4 User-mode startup

完善基础设施后，我们可以尝试运行用户进程，并尝试使用 JOS 目前提供的系统调用了。

Exercise 8

Add the required code to the user library, then boot your kernel. You should see user/hello print "hello, world" and then print "i am environment 00001000". user/hello then attempts to "exit" by calling `sys_env_destroy()` (see `lib/libmain.c` and `lib/exit.c`). Since the kernel currently only supports one user environment, it should report that it has destroyed the only environment and then drop into the kernel monitor. You should be able to get **make grade** to succeed on the hello test.

根据实验的提示，我们需要补充 `user-lib` 的内容，而具体是 `libmain.c` 的内容。实验参考材料告诉我们，所有用户态程序以 `lib/entry.S` 处代码为入口并进入 `libmain.c` 的 `libmain()` 函数。该函数应当设置当前用户进程的目标，并将执行权交给真正的用户态函数 `umain`（在执行真正用户态程序前的各类准备处理与真实操作系统的 `libc` 有相似之处）。我们修改 `libmain` 函数：

```
void
libmain(int argc, char **argv)
{
    // set thisenv to point at our Env structure in envs[].
    // LAB 3: Your code h(ere).
    thisenv = envs + ENVX(sys_getenvid());

    // save the name of the program so that panic() can use it
    if (argc > 0)
        binaryname = argv[0];

    // call user main routine
    umain(argc, argv);

    // exit gracefully
    exit();
}
```

我们用 **make run-hello** 命令查看 `hello` 程序运行结果。可以看到，在新进程被分配给用户态后，分别产生了两次中断号为 48（0x30）的中断（即系统调用），根据代码我们也可以知道分别为对 `getenvid` 及 `cputs` 的调用：

```
[00000000] new env 00001000
Incoming TRAP frame at 0xeffffbcb and trap-no 48
Incoming TRAP frame at 0xeffffbcb and trap-no 48
hello, world
```

2.5 Page faults and memory protection

实验的最后一部分内容是处理 page fault 的异常。根据 Exercise 5 的内容，所有 page fault 都被交给 `page_fault_handler` 函数处理。实验要求我们区分处理内核引发/用户态引发的 page fault。

Exercise 9

Change `kern/trap.c` to panic if a page fault happens in kernel mode.

Hint: to determine whether a fault happened in user mode or in kernel mode, check the low bits of the `tf_cs`.

Read `user_mem_assert` in `kern/pmap.c` and implement `user_mem_check` in that same file. Change `kern/syscall.c` to sanity check arguments to system calls.

Boot your kernel, running `user/buggyhello`. The environment should be destroyed, and the kernel should not panic.

Finally, change `debuginfo_eip` in `kern/kdebug.c` to call `user_mem_check` on `usd`, `stabs`, and `stabstr`. If you now run `user/breakpoint`, you should be able to run backtrace from the kernel monitor and see the backtrace traverse into `lib/libmain.c` before the kernel panics with a page fault. What causes this page fault? You don't need to fix it, but you should understand why it happens.

我们首先修改 `page_fault_handler` 函数，使其将所有的内核页错误 panic 处理：

```
void
page_fault_handler(struct Trapframe *tf)
{
    ...
    // LAB 3: Your code here.
    if(tf->tf_cs == GD_KT)
        panic("Page fault from kernel.");
    ...
}
```

然后我们再根据实验要求，构建 `user_mem_check` 函数。该函数供 `user_mem_assert` 函数调用，用于检查用户态程序是否产生内存访问的问题。`user_mem_assert` 函数实现完成后供 `syscall` 的实际功能函数使用。根据实验提示，`user_mem_check` 函数实现如下：

```

int
user_mem_check(struct Env *env, const void *va, size_t len, int perm)
{
    // LAB 3: Your code here.
    uintptr_t va_start = (uintptr_t)ROUNDDOWN(va, PGSIZE);    // first-page
    uintptr_t va_end = (uintptr_t)ROUNDDOWN(va+len, PGSIZE);    // last-page
    for(uintptr_t pva = va_start; pva <= va_end; pva += PGSIZE)
    {
        pte_t *ppte;
        // struct PageInfo *ppage;
        // ppage = page_lookup(env->env_pgdir, (void *)pva, &ppte);
        ppte = pgdir_walk(env->env_pgdir, (void *)pva, 0);
        if(!ppte || pva > ULIM || ((*ppte & perm) != perm))
        {
            if(pva < (uintptr_t)va)
                user_mem_check_addr = (uintptr_t)va;
            else
                user_mem_check_addr = pva;
            return -E_FAULT;
        }
    }

    return 0;
}

```

再检查当前需要调用 `user_mem_assert` 的函数 `sys_cputs`:

```

static void
sys_cputs(const char *s, size_t len)
{
    // Check that the user has permission to read memory [s, s+len).
    // Destroy the environment if not.

    // LAB 3: Your code here.
    user_mem_assert(curenv, (void *)s, len, 0);

    // Print the string supplied by the user.
    cprintf("%.*s", len, s);
}

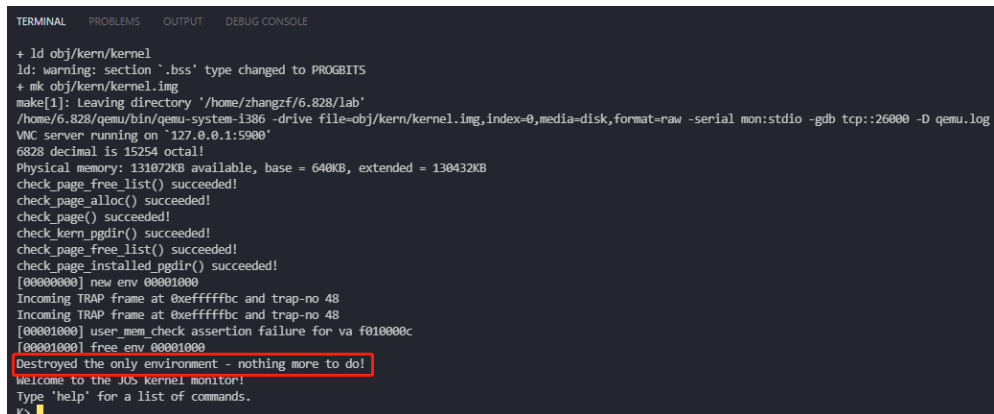
```

最后再根据要求修改 `debuginfo_eip` 函数即可：

```
...  
// Make sure this memory is valid.  
// Return -1 if it is not. Hint: Call user_mem_check.  
// LAB 3: Your code here.  
if(user_mem_check(curenv, usd, sizeof(struct UserStabData), PTE_U))  
    return -1;  
...
```

Exercise 10

Boot your kernel, running `user/evilhello`. The environment should be destroyed, and the kernel should not panic.



```
TERMINAL  PROBLEMS  OUTPUT  DEBUG CONSOLE  
+ ld obj/kern/kernel  
ld: warning: section `.bss' type changed to PROGBITS  
+ mk obj/kern/kernel.img  
make[1]: Leaving directory '/home/zhangzf/6.828/lab'  
/home/6.828/qemu/bin/qemu-system-i386 -drive file=obj/kern/kernel.img,index=0,media=disk,format=raw -serial mon:stdio -gdb tcp::26000 -D qemu.log  
VMC server running on '127.0.0.1:5900'  
6828 decimal is 15254 octal!  
Physical memory: 131072KB available, base = 640KB, extended = 130432KB  
check_page_free_list() succeeded!  
check_page_alloc() succeeded!  
check_page() succeeded!  
check_kern_pgdir() succeeded!  
check_page_free_list() succeeded!  
check_page_installed_pgdir() succeeded!  
[00000000] new env 00001000  
Incoming TRAP frame at 0xeffffbfc and trap-no 48  
Incoming TRAP frame at 0xeffffbfc and trap-no 48  
[00001000] user_mem_check assertion failure for va f010000c  
[00001000] free env 00001000  
Destroyed the only environment - nothing more to do!  
Welcome to the JOS kernel monitor!  
Type 'help' for a list of commands.  
K> |
```