

# 6.828 Fall 2006 Lab 3: User Environments

Handed out Wednesday, September 27, 2006

Part A due Thursday, October 5, 2006

Part B due Thursday, October 12, 2006

## Introduction

In this lab you will implement the basic kernel facilities required to get a protected user-mode environment (i.e., "process") running. You will enhance the JOS kernel to set up the data structures to keep track of user environments, create a single user environment, load a program image into it, and start it running. You will also make the JOS kernel capable of handling any system calls the user environment makes and handling any other exceptions it causes.

**Note:** In this lab, the terms *environment* and *process* are interchangeable - they have roughly the same meaning. We introduce the term "environment" instead of the traditional term "process" in order to stress the point that JOS environments do not provide the same semantics as UNIX processes, even though they are roughly comparable.

## Getting Started

Download our reference code for lab 3 (lab3-handout.gz) from the labs section of this course, untar it into your 6.828 directory, and merge the changes between lab2 and lab3 into your working kernel source tree as you did before for lab2. As we mentioned before, the `diff` and `patch` utilities can be very useful for this purpose, as well as their "big brother" `cvs`.

Lab 3 contains a number of new source files, which you should browse through as you merge them into your kernel:

inc/	env.h	Public definitions for user-mode environments
	trap.h	Public definitions for trap handling
	syscall.h	Public definitions for system calls from user environments to the kernel
	lib.h	Public definitions for the user-mode support library
kern/	env.h	Kernel-private definitions for user-mode environments
	env.c	Kernel code implementing user-mode environments
	trap.h	Kernel-private trap handling definitions
	trap.c	Trap handling code
	trapentry.S	Assembly-language trap handler entry-points

syscall.h	Kernel-private definitions for system call handling
syscall.c	System call implementation code
lib/ Makefrag	Makefile fragment to build user-mode library, obj/lib/libuser.a
entry.S	Assembly-language entry-point for user environments
libmain.c	User-mode library setup code called from entry.S
syscall.c	User-mode system call stub functions
console.c	User-mode implementations of <code>putchar</code> and <code>getchar</code> , providing console I/O
exit.c	User-mode implementation of <code>exit</code>
panic.c	User-mode implementation of <code>panic</code>
user/ *	Various test programs to check kernel lab 3 code

In addition, a number of the source files we handed out for lab2 are modified in lab3. To see the differences, you can type:

```
$ diff -ur lab2 lab3
```

## Lab Requirements

This lab is divided into two parts, A and B. Part A is due after the first week of the lab period; you should just `gmake handin` your lab before the Part A deadline, even though your code may not yet pass all of the grade script tests. (If it does, great!) You only need to have all the grade script tests passing by the Part B deadline at the end of the second week.

As in lab 2, you will need to do all of the regular exercises described in the lab and *at least one* challenge problem. Additionally, you will need to write up brief answers to the questions posed in the lab and a short (e.g., one or two paragraph) description of what you did to solve your chosen challenge problem. If you implement more than one challenge problem, you only need to describe one of them in the write-up, though of course you are welcome to do more. Place the write-up in a file called `answers.txt` (plain text) or `answers.html` (HTML format) in the top level of your `lab2` directory before handing in your work.

## Debugging tips

For all its faults, Bochs is still a much more hospitable debugging environment than a real processor. Put it to work for you!

- The command `info idt` will print the current interrupt descriptor table (IDT). This is useful for checking whether you set it up correctly.
- The `vb` command sets a breakpoint at a particular `CS:EIP` address. Since the kernel code segment selector is 8, `vb 8:0xf0101234` sets a breakpoint at the

given kernel address. Similarly, since the user segment selector is 0x1b, `vb 0x1b:0x80020` sets a breakpoint at the given user address.

Finally, note that passing all the `gmake grade` tests does not mean your code is perfect. It may have subtle bugs that will only be tickled by future labs. In a perfect world, `gmake grade` would find all your bugs, but no one builds operating systems in a perfect world anyway. Keep in mind that debugging an operating system is a very holistic task -- there are abstraction boundaries, but you can't necessarily place much trust in them since nothing is really enforcing them. If you get all sorts of weird crashes that don't seem to be explainable by a single bug in the layer you're working on, it's likely that they're explainable by a single bug in a different layer.

## Inline Assembly

In this section you may find GCC's inline assembly language feature useful, although it is also possible to complete the lab without using it. At the very least, you will need to be able to understand the fragments of inline assembly language ("asm" statements) that already exist in the source code we gave you. For the "definitive" reference to GCC inline assembly language, type `info gcc`, select the "C Extensions" chapter, and then the "Extended Asm" section. You can find several other sources of information on GCC inline assembly language on the class reference materials (syllabus section) page.

## Hand-In Procedure

As before, you can test your code against our test scripts by running `gmake grade`. When you are done, run `gmake handin` to tar up and hand in your source tree.

# Part A: User Environments and Exception Handling

The new include file `inc/env.h` contains basic definitions for user environments in JOS; you should familiarize yourself with it right away. The kernel uses the `Env` data structure to keep track of critical data pertaining to each user environment. In this lab you will initially create just one environment, but you will need to design the JOS kernel to support multiple environments; lab 4 will take advantage of this feature by allowing a user environment to `fork` other environments.

As you can see in `kern/env.c`, the kernel maintains three main global variables pertaining to environments:

```
struct Env *envs = NULL;           /* All environments */
struct Env *curenv = NULL;        /* the current env */
static struct Env_list env_free_list; /* Free list */
```

Once JOS gets up and running, the `envs` pointer points to an array of `Env` structures representing all the environments in the system. In our design, the JOS kernel will support a maximum of `NENV` simultaneously active environments, although there will

typically be far fewer running environments at any given time. (`NENV` is a constant `#define'd` in `inc/env.h`.) Once it is allocated, the `envs` array will contain a single instance of the `Env` data structure for each of the `NENV` possible environments.

The JOS kernel keeps all of the inactive `Env` structures on the `env_free_list`. This design allows extremely quick and efficient allocation and deallocation of environments, as they merely have to be added to or removed from the free list.

The kernel uses the `curenv` variable to keep track of the *currently executing* environment at any given time. During boot up, before the first environment is run, `curenv` is initially set to `NULL`.

## Environment State

The `Env` structure is defined in `inc/env.h` as follows (although more fields will be added in future labs):

```
struct Env {
    struct Trapframe env_tf;           // Saved registers
    LIST_ENTRY(Env) env_link;         // Free list link pointers
    u_int env_id;                     // Unique environment
    identifier
    u_int env_parent_id;              // env_id of this env's parent
    u_int env_status;                // Status of the environment

    // Address space
    Pde *env_pgdir;                  // Kernel virtual address of
page dir
    u_int env_cr3;                    // Physical address of page dir
};
```

We now briefly describe the state kept by the kernel for each user environment.

### **env\_tf:**

This structure, defined in `inc/trap.h`, holds the saved register values for the environment while that environment is *not* running: i.e., when the kernel or a different environment is running. The kernel saves these when switching from user to kernel mode for any reason, so that the environment can later be resumed where it left off.

### **env\_link:**

This is a pair of pointers allowing the `Env` to be placed on the `env_free_list`. See `inc/queue.h` for details.

### **env\_id:**

The kernel stores here a value that uniquely identifies the environment currently using this `Env` structure (i.e., using this particular slot in the `envs` array). After a user environment terminates, the kernel may subsequently re-allocate the same `Env` structure to a different environment - but in this case the new environment will still have a different `env_id` from the old one even though the new environment is re-using the same slot in the `envs` array.

**env\_parent\_id:**

The kernel stores here the `env_id` of the environment that created this environment. In this way the environments can form a "family tree," which will be useful for making security decisions about which environments are allowed to do what to whom.

**env\_status:**

This variable holds one of the following values:

`ENV_FREE:`

Indicates that the `Env` structure is inactive, and therefore on the `env_free_list`.

`ENV_RUNNABLE:`

Indicates that the `Env` structure represents a currently active environment, and the environment is waiting to run on the processor.

`ENV_NOT_RUNNABLE:`

Indicates that the `Env` structure represents a currently active environment, but it is not currently ready to run: for example, because it is waiting for an interprocess communication (IPC) from another environment.

**env\_pgdir:**

This variable holds a *virtual address* pointer to this environment's page directory.

**env\_cr3:**

This variable holds the corresponding *physical address* for this environment's page directory.

Like a Unix process, a JOS environment couples the concepts of "thread" and "address space". The thread is defined primarily by the saved registers (the `env_tf` field), and the address space is defined by the page directory and page tables pointed to by `env_pgdir` and `env_cr3`. To run an environment, the kernel must set up the CPU with *both* the saved registers and the appropriate address space.

Our `struct Env` is analogous to `struct proc` in xv6. Both structures hold the environment's (i.e., process's) user-mode register state directly within the `env_tf` substructure (dubbed simply `tf` in the case of xv6). In JOS, individual environments do not have their own kernel stacks as processes do in xv6. Instead, all JOS kernel code runs on a *single* kernel stack.

## Allocating the Environments Array

In lab 2, you allocated memory in `i386_vm_init()` for the `pages[]` array, which is a table the kernel uses to keep track of which pages are free and which are not. You will now need to modify `i386_vm_init()` further to allocate a similar array of `Env` structures, called `envs`.

**Exercise 1.** Modify `i386_vm_init()` in `kern/pmap.c` to allocate and map the `envs` array. This array consists of exactly `NENV` instances of the `Env` structure, laid out consecutively in the kernel's virtual address space starting at address `UENVS` (defined in `inc/memlayout.h`). The

physical pages that these virtual addresses map to do not have to be contiguous, since the kernel only ever uses virtual addresses to access the `envs` array. You should be able to allocate and map this array in exactly the same way as you did for the `pages` array.

## Creating and Running Environments

You will now write the code in `kern/env.c` necessary to run a user environment. Because we do not yet have a filesystem, we will set up the kernel to load a static binary image that is *embedded within the kernel itself*, much like xv6 does with `load_icode` on line 1356. As does xv6, JOS embeds these binaries in the kernel as *real* ELF executable images. By contrast, the Unix V6 OS studied in 6.828 in previous years used short hand-assembled code fragments.

The Lab 3 `GNUmakefile` generates a number of binary images in the `obj/user/` directory. If you look at `kern/Makefrag`, you will notice some magic that "links" these binaries directly into the kernel executable as if they were `.o` files. The `-b` binary option on the linker command line causes these files to be linked in as "raw" uninterpreted binary files rather than as regular `.o` files produced by the compiler. (As far as the linker is concerned, these files do not have to be ELF images at all - they could be anything, such as text files or pictures!) If you look at `obj/kern/kernel.sym` after building the kernel, you will notice that the linker has "magically" produced a number of funny symbols with obscure names like `_binary_obj_user_hello_start`, `_binary_obj_user_hello_end`, and `_binary_obj_user_hello_size`. The linker generates these symbol names by mangling the file names of the binary files; the symbols provide the regular kernel code with a way to reference the embedded binary files.

In `i386_init()` in `kern/init.c` you'll see code to run one of these binary images in an environment. However, the critical functions to set up user environments are not complete; you will need to fill them in.

**Exercise 2.** In the file `env.c`, finish coding the following functions:

```
env_init():
    initialize all of the Env structures in the envs array and add
    them to the env_free_list.
env_setup_vm():
    allocate a page directory for a new environment and initialize
    the kernel portion of the new environment's address space.
segment_alloc():
    allocates and maps physical memory for an environment
load_icode():
    you will need to parse an ELF binary image, much like the boot
    loader already does, and load its contents into the user address
    space of a new environment.
env_create():
```

allocate an environment with `env_alloc` and call `load_icode` to load an ELF binary into it.

`env_run()`:

start a given environment running in user mode.

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".

Once you are done you should compile your kernel and run it under Bochs. If all goes well, your system should crash in the user program while it is trying to make a system call, since you haven't implemented system calls yet; you should examine what happened as explained below. Below is a call graph of the code up to the point where the user code is invoked. Make sure you understand the purpose of each step.

- start (kern/entry.S)
- i386\_init
  - cons\_init
  - i386\_detect\_memory
  - i386\_vm\_init
  - page\_init
  - env\_init
  - idt\_init (still incomplete at this point)
  - env\_create
  - env\_run
    - env\_pop\_tf

Set a Bochs breakpoint at `env_pop_tf`, which should be the last function you hit before actually entering user mode. Step through this function; the processor should enter user mode after the `iret` instruction. You should then see the first instruction in the user environment's executable, which is the `cmp1` instruction at the label `start` in `lib/entry.S`. You should be able to single-step through this user mode environment code until you first hit an `int $0x30` instruction, which is the instruction that user-mode code executes to make a system call. Unless you've changed the code in `i386_init`, the first system call you'll hit is `sys_cputs`, which is called by `cprintf` to put characters to the console. (See `lib/syscall.c` for more details.) If you cannot get to this point, then something is wrong with your address space setup or program loading code; go back and fix it before continuing.

## Handling Interrupts and Exceptions

At this point, the first `int $0x30` system call instruction in user space is a dead end: once the processor gets into user mode, there is no way to get back out. You will now need to implement basic exception and system call handling, so that it is possible for the kernel to recover control of the processor from user-mode code. The first thing you should do is thoroughly familiarize yourself with the x86 interrupt and exception mechanism.

**Exercise 3.** Read Chapter 9, Exceptions and Interrupts in the 80386 Programmer's Manual (or Chapter 5 of the IA-32 Developer's Manual) (see readings), if you haven't already.

In this lab we generally follow Intel's terminology for interrupts, exceptions, and the like. However, be aware that terms such as exceptions, traps, interrupts, faults and aborts have no standardized meaning across architectures or operating systems, and often used rather loosely without close regard to the subtle distinctions between them on a particular architecture such as the x86. When you see these terms outside of this lab, the meanings might be slightly different.

## Basics of Protected Control Transfer

Exceptions and interrupts are both "protected control transfers," which cause the processor to switch from user to kernel mode without giving the user-mode code any opportunity to interfere with the functioning of the kernel or other environments. In Intel's terminology, an *interrupt* is a protected control transfer that is caused by an asynchronous event usually external to the processor, such as notification of external device I/O activity. An *exception*, in contrast, is a protected control transfer caused synchronously by the currently running code, for example due to a divide by zero or an invalid memory access.

In order to ensure that these protected control transfers are actually *protected*, the processor's interrupt/exception mechanism is designed so that the code currently running when the interrupt or exception occurs *does not get to choose arbitrarily where the kernel is entered or how*. Instead, the processor ensures that the kernel can be entered only under carefully controlled conditions. On the x86, this protection is provided on the basis of two particular mechanisms:

1. **The Interrupt Descriptor Table.** The processor ensures that interrupts and exceptions can only cause the kernel to be entered at a few specific, well-defined entry-points *determined by the kernel itself*, and not by the code currently running when the interrupt or exception is taken.

In particular, x86 interrupts and exceptions are differentiated into up to 256 possible "types", each associated with a particular *interrupt number* (often referred to synonymously as an *exception number* or *trap number*). Once the processor identifies a particular interrupt or exception to be taken, it uses the interrupt number as an index into the processor's *interrupt descriptor table* (IDT), which is a special table that the kernel sets up in kernel-private memory, much like the GDT. From the appropriate entry in this table the processor loads:



- the value to load into the instruction pointer (`EIP`) register, pointing to the kernel code designated to handle that type of exception.
  - the value to load into the code segment (`CS`) register, which includes in bits 0-1 the privilege level at which the exception handler is to run. (In JOS, all exceptions are handled in kernel mode, or privilege level 0.)
2. **The Task State Segment.** In addition to having a well-defined entry-point in the kernel for an interrupt or exception handler, the processor also needs a place to save the *old* processor state before the interrupt or exception occurred, such as the original values of `EIP` and `CS` before the processor invoked the exception handler, so that the exception handler can later restore that old state and resume the interrupted code from where it left off. But this save area for the old processor state must in turn be protected from unprivileged user-mode code; otherwise buggy or malicious user code could easily compromise the kernel.

For this reason, when an x86 processor takes an interrupt or trap that causes a privilege level change from user to kernel mode, it not only loads new values into `EIP` and `CS`, but also loads new values into the stack pointer (`ESP`) and stack segment (`SS`) registers, effectively switching to a new stack private to the kernel. The processor then pushes the original values of *all* of these registers, along with the contents of the `EFLAGS` register, onto this new kernel stack before starting to run the kernel's exception handler code. The new `ESP` and `SS` do *not* come from the IDT like the `EIP` and `CS` do, but instead from a separate structure called the *task state segment* (TSS).

Although the TSS is a somewhat large and complex data structure that can potentially serve a variety of purposes, in JOS it will *only* be used to define the kernel stack that the processor should switch to when it transfers from user to kernel mode. Since "kernel mode" in JOS is privilege level 0 on the x86, the processor uses the `ESP0` and `SS0` fields of the TSS to define the kernel stack when entering kernel mode; none of the other fields in the TSS will ever be used in JOS.

## Types of Exceptions and Interrupts

All of the synchronous exceptions that the x86 processor can generate internally use interrupt numbers between 0 and 31, and therefore map to IDT entries 0-31. For example, the page fault handler is "hard-wired" by Intel to interrupt number 14. Interrupt numbers greater than 31 are only used by *software interrupts*, which can be generated by the `INT` instruction, or asynchronous *hardware interrupts*, caused by external devices when they need attention.

In this section we will extend JOS to handle the internally generated x86 exceptions in the 0-31 that are currently defined by Intel. In addition, in the next section we will also make JOS handle software interrupt number 0x30, which JOS (fairly arbitrarily) uses as its system call interrupt number. In Lab 4 we will extend JOS to handle externally generated hardware interrupts such as the clock interrupt.

## An Example

Let's put these pieces together and trace through an example. Let's say the processor is executing code in a user environment and encounters a divide instruction that attempts to divide by zero.

1. The processor switches to the stack defined by the `SS0` and `ESP0` fields of the TSS, which in JOS will hold the values `GD_KD` and `KSTACKTOP`, respectively.
  2. The processor pushes the exception parameters on the kernel stack, starting at address `KSTACKTOP`:
- ```
3.          +-----+ KSTACKTOP
4.          | 0x00000  old SS  | " - 4
5.          |         old ESP | " - 8
6.          |         old EFLAGS | " - 12
7.          | 0x00000 | old CS  | " - 16
8.          |         old EIP  | " - 20 <----- ESP
9.          +-----+
```
10. Because we're handling a divide error, which is interrupt number 0 on the x86, the processor reads IDT entry 0 and sets `CS:EIP` to point to the handler function defined there.
  11. The handler function takes control and handles the exception, for example by terminating the user environment.

For certain types of x86 exceptions, in addition to the "standard" five words above, the processor pushes onto the stack another word containing an *error code*. The page fault exception, number 14, is an important example. See the 80386 manual to determine for which exception numbers the processor pushes an error code, and what the error code means in that case. When the processor pushes an error code, the stack would look as follows at the beginning of the exception handler when coming in from user mode:

```
+-----+ KSTACKTOP
| 0x00000  old SS  | " - 4
|         old ESP | " - 8
|         old EFLAGS | " - 12
| 0x00000 | old CS  | " - 16
|         old EIP  | " - 20
|         error code | " - 24 <----- ESP
+-----+
```

## Nested Exceptions and Interrupts

The processor can take exceptions and interrupts both from kernel and user mode. It is only when entering the kernel from user mode, however, that the x86 processor automatically switches stacks before pushing its old register state onto the stack and invoking the appropriate exception handler through the IDT. If the processor is *already* in kernel mode when the interrupt or exception occurs (the low 2 bits of the `CS` register are

already zero), then the kernel just pushes more values on the same kernel stack. In this way, the kernel can gracefully handle *nested exceptions* caused by code within the kernel itself. This capability is an important tool in implementing protection, as we will see later in the section on system calls.

If the processor is already in kernel mode and takes a nested exception, since it does not need to switch stacks, it does not save the old `SS` or `ESP` registers. For exception types that do not push an error code, the kernel stack therefore looks like the following on entry to the exception handler:

```

+-----+ <----- old ESP
|      old EFLAGS      | |      " - 4
| 0x000000 | old CS   | |      " - 8
|      old EIP      | |      " - 12
+-----+

```

For exception types that push an error code, the processor pushes the error code immediately after the old `EIP`, as before.

There is one important caveat to the processor's nested exception capability. If the processor takes an exception while already in kernel mode, and *cannot push its old state onto the kernel stack* for any reason such as lack of stack space, then there is nothing the processor can do to recover, so it simply resets itself. Needless to say, any decent kernel should be designed so that this will never happen unintentionally.

## Setting Up the IDT

You should now have the basic information you need in order to set up the IDT and handle exceptions in JOS. For now, you will set up the IDT to handle all the to handle interrupt numbers 0-31 (the processor exceptions) and interrupts 32-47 (the device IRQs). We may add additional interrupts later.

The header files `inc/trap.h` and `kern/trap.h` contain important definitions related to interrupts and exceptions that you will need to become familiar with. The file `kern/trap.h` contains trap-related definitions that will remain strictly private to the kernel, while the companion header file `inc/trap.h` contains general definitions that may also be useful to user-level programs and libraries in the system.

Note: Some of the exceptions in the range 0-31 are defined by Intel to be reserved. Since they will never be generated by the processor, it doesn't really matter how you handle them. Do whatever you think is cleanest.

The overall flow of control that you should achieve is depicted below:

```

IDT                trapentry.S                trap.c
+-----+

```

```

|   &handler1   |-----> handler1:      trap (struct Trapframe
*tf)            |                // do stuff      {
|                |                call trap        // handle the
exception/interrupt
|                |                // undo stuff   }
+-----+
|   &handler2   |-----> handler2:
|                |                // do stuff
|                |                call trap
|                |                // undo stuff
+-----+
.
.
.
+-----+
|   &handlerX   |-----> handlerX:
|                |                // do stuff
|                |                call trap
|                |                // undo stuff
+-----+

```

Each exception or interrupt should have its own handler in `trapentry.S` and `idt_init()` should initialize the IDT with the addresses of these handlers. Each of the handlers should build a `struct Trapframe` (see `inc/trap.h`) on the stack and call into `trap()` (in `trap.c`) with a pointer to the `Trapframe`.

After control is passed to `trap()`, that function handles the exception/interrupt or dispatches the exception/interrupt to a specific handler function. If and when the `trap()` function returns, the code in `trapentry.S` restores the old CPU state saved in the `Trapframe` and then uses the `iret` instruction to return from the exception.

**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`. You will also need to modify `idt_init()` to initialize the `idt` to point to each of these entry points defined in `trapentry.S`; the `SETGATE` macro will be helpful here.

Hint: your code should perform the following steps:

1. push values to make the stack look like a `struct Trapframe`
2. load `GD_KD` into `%ds` and `%es`
3. `pushl %esp` to pass a pointer to the `Trapframe` as an argument to `trap()`
4. `call trap`
5. `pop` the values pushed in steps 1-3

6. `iret`

Consider using the `pushal` and `popal` instructions; they fit nicely with the layout of the `struct Trapframe`.

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.

*Challenge!* You probably have a lot of very similar code right now, between the lists of `TRAPHANDLER` in `trapentry.S` and their installations in `trap.c`. Clean this up. Change the macros in `trapentry.S` to automatically generate a table for `trap.c` to use. Note that you can switch between laying down code and data in the assembler by using the directives `.text` and `.data`.

## Questions

Answer the following questions in your `answers.txt`:

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 (i.e., as the `grade` script expects)? *Why* is this the correct behavior? What happens if the kernel actually allows `softint`'s `int $14` instruction to invoke the kernel's page fault handler (which is interrupt number 14)?

## Part B: Page Faults, Breakpoints Exceptions, and System Calls

Now that your kernel has basic exception handling capabilities, you will refine it to provide important operating system primitives that depend on exception handling.

### Handling Page Faults

The page fault exception, interrupt number 14 (`T_PGFLT`), is a particularly important one that we will exercise heavily throughout this lab and the next. When the processor takes a page fault, it stores the linear address that caused the fault in a special processor control register, `CR2`. In `trap.c` we have provided the beginnings of a special function, `page_fault_handler()`, to handle page fault exceptions.

**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.

You will further refine the kernel's page fault handling below, as you implement system calls.

## The Breakpoint Exception

The breakpoint exception, interrupt number 3 (`T_BRKPT`), is normally used to allow debuggers to insert breakpoints in a program's code by temporarily replacing the relevant program instruction with the special 1-byte `int3` software interrupt instruction. In JOS we will abuse this exception slightly by turning it into a primitive pseudo-system call that any user environment can use to invoke the JOS kernel monitor. This usage is actually somewhat appropriate if we think of the JOS kernel monitor as a primitive debugger. The user-mode implementation of `panic()` in `lib/panic.c`, for example, performs an `int3` after displaying its panic message.

**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.

*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.

*Optional:* If you're feeling really adventurous, find some x86 disassembler source code - e.g., by ripping it out of Bochs, or out of GNU binutils, or just write it yourself - and extend the JOS kernel monitor to be able to disassemble and display instructions as you are stepping through them. Combined with the symbol table loading from lab 2, this is the stuff of which real kernel debuggers are made.

## Questions

Answer the following questions in your `answers.txt`:

1. The break point test case will either generate a break point exception of a general protect fault depending on how you initialized the break point entry in the IDT (i.e., your call to `SETGATE` from `idt_init`). Why? How did you need to set it in order to get the breakpoint exception to work as specified above?

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

## System calls

User processes ask the kernel to do things for them by invoking system calls. When the user process invokes a system call, the processor enters kernel mode, the processor and the kernel cooperate to save the user process's state, the kernel executes appropriate code in order to carry out the system call, and then resumes the user process. The exact details of how the user process gets the kernel's attention and how it specifies which call it wants to execute vary from system to system.

In the JOS kernel, we will use the `int` instruction, which causes a processor interrupt. In particular, we will use `int $0x30` as the system call interrupt. We have defined the constant `T_SYSCALL` to `0x30` for you. You will have to set up the interrupt descriptor to allow user processes to cause that interrupt. Note that interrupt `0x30` cannot be generated by hardware, so there is no ambiguity caused by allowing user code to generate it.

We will pass the system call number and the system call arguments in registers. This way, we don't need to grub around in the user environment's stack or instruction stream. The system call number will go in `%eax`, and the arguments (up to five of them) will go in `%edx`, `%ecx`, `%ebx`, `%edi`, and `%esi`, respectively. The kernel passes the return value back in `%eax`. The assembly code to invoke a system call has been written for you, in `syscall()` in `lib/syscall.c`. You should read through it and make sure you understand what is going on.

**Exercise 7.** Add a handler in the kernel for interrupt number `T_SYSCALL`. You will have to edit `kern/trapentry.S` and `kern/trap.c`'s `idt_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. You may also find it helpful to read `inc/syscall.h`.

Run the `user/hello` program under your kernel. 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.

*Challenge!* Implement system calls using the `sysenter` and `sysexit` instructions instead of using `int 0x30` and `iret`.

The `sysenter/sysexit` instructions were designed by Intel to be faster than `int/iret`. They do this by using registers instead of the stack and by making assumptions about how the segmentation registers are used. The exact details of these instructions can be found in Volume 2B of the Intel reference manuals.

The easiest way to add support for these instructions in JOS is to add a `sysenter_handler` in `kern/trapentry.S` that creates the same trap frame that is normally created by an `int 0x30` instruction (being sure to save the correct return address and stack pointer provided by the user environment). Then, instead of calling into `trap`, push the arguments to `syscall` and call `syscall` directly. Once `syscall` returns, set everything up for and execute the `sysexit` instruction.

You will also need to add code to `kern/init.c` to set up the necessary model specific registers (MSRs). Look at the `enable_sep_cpu` function in [this diff](#) for an example of this, and you can find an implementation of `wrmsr` to add to `/inc/x86.h` [here](#)). Finally, `lib/syscall.c` must be changed to support making a system call with `sysenter`. Here is a possible register layout for the `sysenter` instruction:

|                                 |                                       |
|---------------------------------|---------------------------------------|
| <code>eax</code>                | - <code>syscall</code> number         |
| <code>edx, ecx, ebx, edi</code> | - <code>arg1, arg2, arg3, arg4</code> |
| <code>esi</code>                | - return pc                           |
| <code>ebp</code>                | - return esp                          |
| <code>esp</code>                | - trashed by <code>sysenter</code>    |

GCC's inline assembler does not support directly loading values into `ebp`, so you will need to add code to save (push) and restore (pop) it yourself (and you may want to do the same thing for `esi` as well). The return address can be put into `esi` by using an instruction like `leal after_sysenter_label, %%esi`.

Note that this only supports 4 arguments, so you will need to leave the old method of doing system calls around if you want to support 5 argument system calls as well.

Finally, in order for Bochs to support these instructions, it must be compiled with the `--enable-sep` option, in addition to the other options listed on the tools page.

## User-mode startup

The user programs start running at the top of `lib/entry.S`. After some setup, this code calls `libmain()`, in `lib/libmain.c`. The `libmain()` function needs to initialize a global pointer `env` to point at this environment's `struct Env` in the `envs[]` array. (Note that `lib/entry.S` has already defined `envs` to point at the `UENVS` mapping you set up in lab 2.) Hint: look in `inc/env.h` and use `sys_getenvid`.



`libmain()` then calls `umain`, which, in the case of the hello program, is in `user/hello.c`. Note that after printing "hello, world", it tries to access `env->env_id`. This is why it faulted earlier. Now that you've initialized `env` properly, it should not fault. If it still faults, you probably haven't mapped the `UENVS` area user-readable (back in lab 1 in `pmap.c`; this is the first time we've actually used the `UENVS` area).

**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 00000800". `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.

## Page faults and memory protection

Memory protection is a crucial feature of an operating system. By using memory protection, the operating system can ensure that bugs in one program cannot corrupt other programs or corrupt the operating system itself.

Typically, operating systems rely on hardware support to implement memory protection. The OS keeps the hardware informed about which virtual addresses are valid and which are not. When a program tries to access an invalid address or one for which it has no permissions, the processor stops the program at the instruction causing the fault and then traps into the kernel with information about the attempted operation. If the fault is fixable, the kernel can fix it and let the program continue running. If the fault is not fixable, then the program cannot continue, since it will never get past the instruction causing the fault.

As an example of a fixable fault, consider an automatically extended stack. In many systems the kernel initially allocates a single stack page, and then if a program faults accessing pages further down the stack, the kernel will allocate those pages automatically and let the program continue. By doing this, the kernel only allocates as much stack memory as the program needs, but the program can work under the illusion that it has an arbitrarily large stack.

System calls present an interesting problem for memory protection. Most system call interfaces let user programs pass pointers to the kernel. These pointers point at user buffers to be read or written. The kernel then dereferences these pointers while carrying out the system call. There are two problems with this:

1. A page fault in the kernel is taken a lot more seriously than a page fault in a user program. If the kernel page faults, that's usually a kernel bug, and the fault handler will panic the kernel (and hence the whole system). In a system call, when the kernel is dereferencing pointers to the user's address space, we need a way to

remember that any page faults these dereferences cause are actually on behalf of the user program.

2. The kernel typically has more memory permissions than the user program. The user program might ask the kernel to read from or write to a location in kernel memory that the user program cannot access but that the kernel can. If the kernel is not careful, a buggy or malicious user program can trick the kernel into using its greater privilege in unintended ways, possibly so as to destroy the integrity of the kernel completely.

For both of these reasons the kernel must be extremely careful when handling pointers presented by user programs.

You will now implement solutions to these two problems with a single mechanism. The gist of the approach that this year's JOS kernel uses (which is different from last year's) is to scrutinize all pointers passed from userspace into the kernel, and to perform in kernel code what the processor does in hardware. That is, when the user passes the kernel a pointer, the kernel will check that it's a user-accessible address, and that the user page tables are correctly configured to allow the memory operation to go through.

Thus, the kernel ought never induce a page fault when handling a system call. If the kernel does page fault, it should panic and terminate. (Why wouldn't this behavior be acceptable for a "commercial" OS such as Linux or FreeBSD?)

**Exercise 9.** Change `kern/trap.c`'s 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.

Change `kern/init.c` to run `user/buggyhello` instead of `user/hello`. Compile your kernel and boot it. The environment should be destroyed, and the kernel should *not* panic. You should see:

```
[00001000] user_mem_check assertion failure for
va 00000001
[00001000] free env 00001000
Destroyed the only environment - nothing more to
do!
```

Note that the same mechanism you just implemented also works for malicious user applications (such as `user/evilhello`).

### Exercise 10.

Change `kern/init.c` to run `user/evilhello`. Compile your kernel and boot it. The environment should be destroyed, and the kernel should not panic. You should see:

```
      [00000000] new env 00001000
      [00001000] user_mem_check assertion failure for
va f0100020
      [00001000] free env 00001000
```

**This completes the lab.** Make sure you pass all the `gmake grade` tests, and hand in your work with `gmake handin`.

---