

Zero Volume 0 - The Kernel
Design and Implementation
DRAFT 1, Revision 11

Tuomo Petteri Venäläinen

August 24, 2013

Contents

I	Overview	5
1	Preface	7
1.1	Acknowledgements	7
1.2	Background	8
2	System Concepts	9
2.1	General Terminology	9
2.2	X86 Terminology	10
3	System Features	13
3.1	UNIX Features	13
3.2	POSIX Features	13
3.3	Zero Features	13
3.4	Compile-Time Configuration	14
II	Basic Kernel	15
4	Kernel Layout	17
5	Kernel Environment	19
5.1	Processor Support	19
5.1.1	Thread Scheduler	19
5.1.1.1	Thread Classes	19
5.1.1.2	Priorities	19
5.1.1.3	Data Structures	20
5.1.2	Interrupt Vector	22
5.1.2.1	Interrupt Descriptors	23
5.2	Memory	24
5.2.1	Overview	24
5.2.2	Segment Descriptor Tables	24
5.2.2.1	Segment Descriptors	24
5.2.3	Paging Data Structures	24
5.2.3.1	Page Directory Entry	24
5.2.3.2	Page Table Entry	24
5.2.3.3	Page Directory	24
5.2.3.4	Page Tables	24
5.2.4	Page Daemon	24

5.2.5	Zone Allocator	24
5.2.5.1	Page Replacement Algorithm	25
6	System Call Interface	27
6.1	Conventions	27
6.1.1	System Call Mechanism	27
6.2	Process Control	27
6.2.1	sysctl	27
6.2.2	exit	28
6.2.3	abort	28
6.2.4	fork	28
6.2.5	exec	28
6.2.6	thorp	29
6.2.7	pctl	29
6.2.8	sigop	30
6.3	Memory Interface	30
6.3.1	brk	30
6.3.2	map	30
6.3.3	umap	31
6.3.4	mhint	31
6.4	Shared Memory	31
6.4.1	shmget	32
6.4.2	shmat	32
6.4.3	shmdt	32
6.4.4	shmctl	32
6.5	Semaphores	33
6.6	Message Queues	33
6.7	Events	33
6.8	I/O	33
III	User Environment	35
6.9	Process Environment	37
6.10	Memory Map	37

Part I

Overview

Chapter 1

Preface

Goal

The goal of the Zero project is a new, portable, high performance, UNIX-inspired operating system. Such systems typically consist of a [relatively] small kernel and supporting user software such as editors, compilers, linkers and loaders, and other software development tools.

Rationale

Whereas different UNIX-like operating systems are doing strong for many users, the world is a different place from when UNIX and some of the related operating systems had their initial designs laid out. We have outstanding graphics (and physics) processors, high quality audio interfaces, lots of memory, plenty of disk space, and so forth. Also, the trend is leaning towards multiprocessor systems with new requirements and possibilities. I think and feel it's worth designing a new operating system for modern computers.

Design

Zero is a multithreaded multiprocessor-enabled kernel. Design goals include fast response to user actions as well as high multimedia performance.

1.1 Acknowledgements

Contributors

At the moment, I have kept Zero a one-man project on purpose. I feel it's too much of a moving target to spend other people's time working on things that may change any moment. I will do my best to bring Zero ready for others to work on - I have been offered help, and I want to thank you guys (who know yourself) here. :)

Open Source Community

First and foremost, I want to thank **the developers** of open and free software for their work and courage to release their work for others to use and modify. Keep the spirit strong!

TODO: thanks and greetings etc.

1.2 Background

Zero has its roots in old, still simple and elegant versions of the UNIX operating system. I still see many good, timeless things about the UNIX design worth reusing in a new operating system. One thing of particular attraction is the "everything is a file" philosophy; I plan to use and possibly extend that idea in Zero.

Chapter 2

System Concepts

This chapter is a glossary of some terminology used throughout the book.

2.1 General Terminology

Buffer

Buffers are used to avoid excess I/O system calls when reading and writing to and from I/O devices. This memory is allocated from a separate buffer cache, which can be either static or dynamic size. Buffer pages are 'wired' to memory; they will never get paged out to disk or other external devices, but instead buffer eviction leads to writing the buffer to a location on some device; typically a disk.

Event

Events are a means for the kernel and user processes to communicate with each other. As an example, user keyboard input needs to be encoded to a well known format (Unicode or in the case of a terminal, ISO 8859-1 or UTF-8 characters) and then dispatched to the event queues of the listening processes. Other system events include creation and destroyal of files and directories.

Zero schedules certain events on timers separate from the timer interrupt used for scheduling threads. At the time of such an event, it is dispatched to the event queues of [registered] listening processes; if an event handler thread has been set, it will be given a short time slice of the event timer. Event time slices should be shorter than scheduler time slices to not interfere with the rest of the system too much.

Interval Task

Short-lived, frequent tasks such as audio and video buffer synchronisation are scheduled with priority higher than normal tasks. It is likely such tasks should be given a slice of time shorter than other threads.

Page

A page is the base unit of memory management. Hardware protection works on per-page basis. Page attributes include read-, write-, and execute-permissions. For typical programs, the text (code) pages would have read- and execute-permissions, whereas the program [initialised] data pages as well as [uninitialised] BSS-segment pages would have read- and write- but not execute-permissions. This approach facilitates protection against overwriting code and against trying to execute code from the data and BSS-segments.

Process

A process is a running instance of a program. A process may consist of several threads. Threads of a process share the same address space, but have individual execution stacks.

Segment

Processes typically consist of several segments. These include a text segment for [read-only] code, a data segment for initialised global data, a BSS-segment for uninitialised global data, and different debugging-related segments. Note that the BSS-segment is runtime- allocated, whereas the data segment is read from the binary image. Also Note that in a different context, segments are used to refer to hardware memory management.

Task

In the context of Zero, the term task is synonymous the term process.

Thread

Threads are the basic execution unit of programs. To utilise multiprocessor-parallelism, a program may consist of several threads of execution, effectively letting it do several computation and I/O operations at the same time.

Trap

A trap is a hardware- or software generated event. Other names for traps include **interrupts**, **exceptions**, **faults**, and **aborts**. As an example, keyboard and mouse input may generate interrupts to be handled by interrupt service routines (**ISRs**).

Virtual Memory

Virtual memory is a technique to map physical memory to per-process address space. The processes see their address spaces as if they were in total control of the machine. These per-process address spaces are protected from being tampered by other processes. Address translations are hardware-level (the kernel mimics them, though), so virtual memory is transparent to application, and most of the time, even kernel programmers.

2.2 X86 Terminology

APIC

APIC stands for [CPU-local] 'advanced programmable interrupt controller.'

GDT

A GDT is 'global descriptor table', a set of memory segments with different permissions.

HPET

HPET is short for high precision event timer (aka multimedia timer).

IDT

IDT means interrupt descriptor table (aka interrupt vector).

ISR

ISR stands for interrupt service routine, i.e. interrupt handler.

LDT

An LDT is local descriptor table', a set of memory segments with different permissions.

PIT

PIT stands for programmable interrupt timer. This timer may be used to schedule threads on uniprocessor systems; for modern PCs, look into **APIC**.

Chapter 3

System Features

3.1 UNIX Features

Concepts

Zero is influenced and inspired by AT&T and BSD UNIX systems. As I think many of the ideas in these operating systems have stood the test of time nicely, it feels natural to base Zero on some well-known and thoroughly-tested concepts.

Nodes

Nodes are similar with UNIX 'file' descriptors. All I/O objects, lock structures needed for IPC and multithreading, as well as other types of data structures are called nodes, collectively. Their [64-bit] IDs are typically per-host kernel **memory addresses** (pointer values) for kernel **descriptor data structures**.

3.2 POSIX Features

Threads

Perhaps the most notable POSIX-influenced feature in Zero kernel is threads. POSIX- and C11-threads can be thought of as light-weight 'processes' sharing the parent process address space but having unique execution stacks. Threads facilitate doing several operations at the same time, which makes into better utilisation of today's multicore- and multiprocessor-systems.

3.3 Zero Features

Events

Possibly the most notable extension to traditional UNIX-like designs in Zero is the event interface. Events are interprocess communication messages between kernel and user processes. Events are used to notify of user device (keyboard,

mouse/pointer) input, filesystem actions such as removal and destroyal of files or directories, as well as to communicate remote procedure calls and data between two processes (possibly on different hosts).

Events are communicated using message passing; the fastest transport available is chosen to deliver messages from a process to another one; in a scenario like a local display connection, messages can be passed by using a shared memory segment mapped to the address spaces of both the kernel and the desktop server.

3.4 Compile-Time Configuration

The table below lists some features of the Zero kernel that you can configure at compile-time. The list is not complete; for more settings, consult `<kern/-conf.h>`.

Macro	Brief	Notes
SMP	multiprocessor support	Not functional yet
HZ	timer frequency	default value is 250
ZEROSCHED	default scheduler	do not change yet
NPROC	maximum simultaneous processes	default is 256
NTHR	maximum simultaneous threads	default is 4096
NCPU	number of CPU units supported	default is 8 if SMP is non-zero

Part II

Basic Kernel

Chapter 4

Kernel Layout

Monolithic Kernel

Zero is a traditional, monolithic kernel. It consists of several code modules, some notable ones of which are listed below.

Module	Description
tmr	hardware timer interface
proc	process interface
thr	thread scheduler
vm	virtual memory manager
page	page daemon
mem	kernel memory allocator
io	I/O primitives
buf	block/buffer cache management

The code modules above will be discussed in-depth in the later parts of this book.

Chapter 5

Kernel Environment

This chapter describes the kernel-mode execution environment. Hardware-specific things are described for the IA-32 and X86-64 architectures.

5.1 Processor Support

Initially, Zero shall support 32-bit and 64-bit X86 architectures. The plan is to port Zero at least to ARM in addition.

5.1.1 Thread Scheduler

5.1.1.1 Thread Classes

The following code snippet lists thread classes in decreasing order of priority.

```
#define THRRT      0 // real-time thread
#define THRUSER    1 // user thread
#define THRBATCH   2 // batch thread
#define THRIDLE    3 // idle thread
#define THRNCLASS  4
#define THRNPRIO   64 // # of priorities per class
```

5.1.1.2 Priorities

Priority Queue

Zero scheduler uses a queue with $THRNCLASS * THRNPRIO$ priorities.

Priority Adjustment

Every time a thread stops running, its priority is adjusted with the function below.

```

/* adjust thread priority */
static __inline__ long
thradjprio(struct thr *thr)
{
    long class = thr->class;
    long prio = thr->prio;

    if (class != THRRRT) {
        /* wrap around back to 0 at THRNPRIO / 2 */
        prio++;
        prio &= (THRNPRIO >> 1) - 1;
        prio = (THRNPRIO * class) + (THRNPRIO >> 1) + prio + thr->nice;
        prio = max(THRNPRIO * THRNCLASS - 1, prio);
        thr->prio = prio;
    }

    return prio;
}

```

5.1.1.3 Data Structures

The structures **struct proc** and **struct thr** are declared in `<kern/obj.h>`. The example implementations below are for the **IA-32** architecture.

struct proc

The process data structure, of the type **struct proc** consists of per-process book-keeping data shared between all threads of the process.

IA-32 Implementation

```

/* process */
struct proc {
    struct thr      *thr;           // current running thread
    long            nthr;           // # of threads
    /* round-robin queue */
    struct thrq     thrq;           // queue of ready threads
    /* page directory */
    pde_t           *pdir;          // page directory address
    /* stacks */
    uint8_t         *ustk;          // user-mode stack
    uint8_t         *kstk;          // kernel-mode stack
    long            class;
    /* memory attributes */
    uint8_t         *brk;
    /* process credentials */
    pid_t           pid;            // process ID
    pid_t           parent;         // parent process
    uid_t           ruid;           // real user ID
    gid_t           rgid;           // real group ID
    uid_t           euid;           // effective user ID
    gid_t           egid;           // effective group ID
    /* descriptor tables */
    desc_t          *dtab;
    /* signal state */
    sigset_t        sigmask;        // signal mask
    sigset_t        sigpend;        // pending signals
    signalhandler_t *sigvec[NSIG];
    /* runtime arguments */
    long            argc;           // argument count
    char            **argv;         // argument vector
    char            **envp;         // environment strings
    /* memory management */
    struct slabhdr   *vmtab[PTRBITS];
    /* event queue */
    struct ev        *evq;
} PACK() ALIGNED(PAGESIZE);

```

struct thr

The thread data structure, of the type **struct thr**, consists of machine-specific and portable members.

IA-32 Implementation

```

struct thr {
    /* thread control block */
    struct m_tcb  m_tcb;           // context
    /* thread stacks */
    uintptr_t     ustk;           // user-mode stack
    uintptr_t     kstk;           // kernel-mode stack
    /* state */
    long          state;          // thread state
    /* wait channel */
    uintptr_t     wchan;          // wait channel
    /* linkage */
    struct proc   *proc;          // owner process
    /* queue linkage */
    struct thr    *prev;          // previous in queue
    struct thr    *next;          // next in queue
    long          id;
    /* scheduler parameters */
    long          nice;           // priority adjustment
    long          class;          // thread class
    long          prio;           // priority
    long          runtime;        // run time
    /* system call context */
    struct syscall syscall;       // current system call
} PACK();

```

5.1.2 Interrupt Vector

The interrupt vector is an array of interrupt descriptors. The descriptors contain interrupt service routine base address for the function to be called to handle the interrupt.

The symbolic [macro] names for interrupts and IRQs (interrupt requests) are declared in `<kern/unix/x86/trap.h>`.

Interrupts

The following table lists X86 interrupts along with their associated signals.

Macro	Number	Description	Signal
TRAPDE	0x00	divide error; fault	SIGFPE
TRAPDB	0x01	reserved; fault/trap	N/A
TRAPNMI	0x02	non-maskable interrupt; interrupt	N/A
TRAPBP	0x03	breakpoint; trap	SIGTRAP
TRAPOF	0x04	overflow; trap	N/A
TRAPBR	0x05	bound range exceeded; fault	SIGBUS
TRAPUD	0x06	invalid opcode; fault	SIGILL
TRAPNM	0x07	no FPU; fault	SIGILL
TRAPDF	0x08	double-fault; fault, error == 0	N/A
TRAPRES1	0x09	reserved	N/A
TRAPTS	0x0a	invalid tss; fault + error	N/A
TRAPNP	0x0b	segment not present; fault + error	SIGSEGV
TRAPSS	0x0c	stack-segment fault; fault + error	SIGSTKFLT
TRAPGP	0x0d	general protection; fault + error	SIGSEGV
TRAPPF	0x0e	page-fault; fault + error	N/A
TRAPRES2	0x0f	reserved	N/A
TRAPMF	0x10	FPU error; fault + number	SIGFPE
TRAPAC	0x11	alignment check; fault, error == 0	SIGBUS
TRAPMC	0x12	machine check; abort	SIGABRT
TRAPXF	0x13	SIMD exception; fault	SIGFPE

IRQs

The following table lists default use of IRQs 0 through 0x0f (15). Zero redirects these to interrupts 0x20 through 0x2f (32 to 47).

Macro	Number	Description
IRQTIMER	0	timer interrupt
IRQKBD	1	keyboard interrupt
IRQCOM2AND4	3	serial ports 2 and 4
IRQCOM1AND3	4	serial ports 1 and 3
IRQLPT	5	parallel port
IRQFD	6	floppy disk drive
IRQRTC	8	real-time clock
IRQMOUSE	12	mouse
IRQFPU	13	floating point unit
IRQIDE0	14	IDE controller 1
IRQIDE1	15	IDE controller 2

5.1.2.1 Interrupt Descriptors

Entries in the interrupt vector, i.e. interrupt descriptor table (**IDT**), are called interrupt descriptors. These descriptors, whereas a bit hairy format-wise, consist

of interrupt handler address, protection ring (system or user), and certain other attribute flags.

5.2 Memory

5.2.1 Overview

5.2.2 Segment Descriptor Tables

A process may use either a global descriptor table with its physical address in the GDT-register or a local one with the address in the LDT-register. Processes are required to declare a handful of entries into their descriptor tables.

5.2.2.1 Segment Descriptors

A segment descriptor is a CPU-specific data structure. On **IA-32** and **X86-64** architectures, the descriptors have base address and limit fields, permission bits, and other such values.

The following code snippet illustrates important values in segment descriptors on **IA-32**.

```
#define SEGDEFBITS (SEG32BIT | SEG4KGRAN | SEGPRES)
#define SEGTTSS   (SEGAVAILTSS | SEGUSER | SEGPRES)
```

5.2.3 Paging Data Structures

Protection and other control of memory appears on per-page basis.

5.2.3.1 Page Directory Entry

5.2.3.2 Page Table Entry

5.2.3.3 Page Directory

5.2.3.4 Page Tables

5.2.4 Page Daemon

5.2.5 Zone Allocator

The Zero memory manager was crafted to use aggressive buffering of allocations. The higher buffer level is a **Bonwick**-style **magazine** layer consisting of allocation stacks for sub-slab blocks. The lower level is a typical **slab allocator**.

5.2.5.1 Page Replacement Algorithm

Chapter 6

System Call Interface

TODO

Keep in mind, that the interface described here is currently **incomplete**; therefore, please consult the final interface later.

The most notable missing things at the moment are support for sockets as well as semaphores.

6.1 Conventions

6.1.1 System Call Mechanism

Interrupt 0x80

On **IA-32** architectures, all of up to 3 system call parameters are passed in registers; the system call number is in **EAX**. On **X86-64** all system call arguments are passed in registers. On both architectures, system calls return a value for **errno**; the value is returned in **EAX** on IA-32, and in **RAX** on X86-64. **Failures** are indicated by setting the **CF-bit** (carry) in the **EFLAGS**-register. System calls are, in the first implementation, triggered by **interrupt 0x80**.

TODO: `sysenter` + `sysexit`

6.2 Process Control

6.2.1 `sysctl`

```
long sys__sysctl(long cmd, long parm, void *arg);
```

The `sysctl` system call is used to trigger certain system control operations. The commands, their arguments, and the return values are listed in the next table.

cmd	parm	arg	Returns
SYS_INIT	runlevel	N/A	new runlevel

TODO: **SYS_INIT** (**SYS_HALT**, **SYS_REBOOT**), ...

6.2.2 exit

```
long sys_exit(long val, long flg);
```

The exit system call terminates the calling process. The process returns **val** as its exit status. If **flg** has the **EXIT_DUMPACCT** bit set, process accounting information is dumped into the **/var/log/acct.log** system log file.

6.2.3 abort

```
void sys_abort(void);
```

The abort system call terminates the calling process in an abnormal way. If the limit for core dump size is set to be big enough, a memory image of the process is dumped into a **core** file. The location of this file may be either the local directory or one configured in **/etc/proc/core.cfg**.

6.2.4 fork

```
long sys_fork(long flg);
```

The fork system call creates a new child process. The return value is the process ID of the new child process in the parent, and 0 (zero) in the child process. If **flg** has the **FORK_VFORK** bit set, the new process shall share the parent's address space; otherwise, the child's address space will be a clone of the parent's address space. If **flg** has the **FORK_COW** (copy on write) bit set, the new process will only clone pages as they are written on.

6.2.5 exec

```
long sys_exec(char *path, char *argv[], ...);
```

The exec system call replaces the calling process by an instance of the program **path**. The argument vector **argv** holds argument strings for the program to be executed; the table must be terminated by a final **NULL** pointer.

If a third argument is given, it shall be **char **** used as **environment** strings for the program; the table must be terminated by a final **NULL** pointer.

6.2.6 throp

long sys_throp(long cmd, long parm, void *arg);

The throp system call provides thread control. The **cmd** argument is one of the values in the following table.

cmd	parm	arg	Notes
THR_NEW	class	struct thrarg *	pthread_create()
THR_JOIN	thrid	struct thrjoin *	pthread_join()
THR_DETACH	N/A	N/A	pthread_detach()
THR_EXIT	N/A	N/A	pthread_exit()
THR_CLEANUP	N/A	N/A	cleanup; pop and execute handlers etc.
THR_KEYOP	cmd	struct thrkeyop *	create, delete
THR_SYSTOP	cmd	struct thrsys *	atfork, sigmask, sched, scope
THR_STKOP	thrid	struct thrstk *	stack; addr, size, guardsize
THR_RTOP	thrid	struct thrrtop *	realtime thread settings
THR_SETATR	thrid	struct thratr *	set other attributes

6.2.7 pctl

long sys_pctl(long cmd, long parm, void *arg);

TODO: pid as parm, special values for the choices

The pctl system call provides process operations. The following tables list provided functionality.

cmd	parm	Notes
PROC_GETPID	N/A	getpid()
PROC_GETPGRP	N/A	getpggrp()
PROC_WAIT	-1	wait for any child
	pid > 0	wait for pid
	pid == 0	wait for children in the group of caller
	pid < -1	wait for children in the group abs(pid)
PROC_USLEEP	milliseconds	usleep()
PROC_NANOSLEEP	nanoseconds	nanosleep()

Notes

For **PROC_WAIT**, if **arg** is NULL or **struct rusage *** for collecting resource usage statistics.

cmd	parm	arg	Notes
PROC_SETSCHED	attribute bits	struct syssched *	set scheduling parameters
PROC_STAT	pid or RUSAGE_SELF	struct rusage *	get resource usage statistics
PROC_GETLIM	limit bits	struct rlimit *	read process limits
PROC_SETLIM	limit bits	struct rlimit *	set process limits

6.2.8 sigop

```
long sys_sigop(long cmd, long parm, void *arg);
```

The sigop system call provides control over signals and related program behavior. The different values for **cmd** as well as related values for **parm** are shown in the following table.

cmd	parm	arg	Notes
SIG_WAIT	N/A	N/A	pause()
SIG_SETFUNC	sig	struct sigarg *	signal()/sigaction()
SIG_SETMASK	N/A	sigset_t *	sigsetmask()
SIG_SEND	N/A	sigset_t *	raise() etc.
SIG_SETSTK	N/A	struct sigstk *	sigaltstack()
SIG_SUSPEND	N/A	sigset_t *	sigsuspend(), sigpause()

Structure Declarations

```
/* flg bits */
#define SIG_NOCLDSTOP 0x01 // no SIGCHLD on stop or cont
#define SIG_ONSTACK 0x02 // use sigaltstk() stack
#define SIG_RESETHAND 0x04 // reset handler to SIG_DFL
#define SIG_RESTART 0x08 // no EINTR behavior
#define SIG_SIGINFO 0x10 // func(int, siginfo_t, void *)
struct sigarg {
    long sig; // signal ID
    long flg; // see SIG_-macros above
    void *func; // signal disposition
};
```

6.3 Memory Interface

6.3.1 brk

```
long sys_brk(void *adr);
```

The brk system call sets the current break, i.e. top of heap, of the calling process to **adr**. The return value is 0 on success, -1 on failure.

6.3.2 map

```
void *sys_map(long desc, long flg, struct sysmem *arg);
```

The map system call is used to map [zeroed] anonymous memory or files to the calling process's virtual address space. For compatibility with existing systems, mapping the device special file **/dev/zero** is similar to using the **flg** value of **MAP_ANON**.

flag	Notes
MAP_FILE	object is a file (may be /dev/zero)
MAP_ANON	map anonymous memory set to zero
MAP_SHARED	changes are shared
MAP_PRIVATE	changes are private
MAP_FIXED	map to provided address
MAP_SINGLE	map buffer to single user process and kernel
MEM_NORMAL	normal behavior
MEM_SEQUENTIAL	sequential I/O buffer
MEM_RANDOM	random-access buffer
MEM_WILLNEED	don't unmap after use; keep in buffer cache
MEM_DONTNEED	unmap after use
MEM_DONTFORK	do not share with child processes

Structure Declarations

```
struct sysmem {
    void *base; // base address
    long ofs; // offset in bytes
    long len; // length in bytes
    long perm; // permission bits
};
```

6.3.3 umap

```
long sys_umap(void *adr, size_t size);
```

The umap system call unmaps memory regions mapped with sys_map().

6.3.4 mhint

```
long sys_mhint(void *adr, long flag, struct sysmem *arg);
```

The mhint system call is used to hint the kernel of a memory region use patterns. The possible bits for **flag** are shown in the table below; for **struct sysmem** declaration, see **map**.

MEM_NORMAL	default behavior
MEM_SEQUENTIAL	sequential I/O buffer
MEM_RANDOM	random-access buffer
MEM_WILLNEED	don't unmap after use; keep in buffer cache
MEM_DONTNEED	unmap after use
MEM_DONTFORK	do not share with forked child processes

6.4 Shared Memory

The shared memory interface of Zero is modeled after the POSIX interface.

6.4.1 shmget

```
long sys_shmget(long key, size_t size, long flag);
```

The `shmget` system call maps a shared memory segment; it returns a shared memory identifier (usually a long-cast of a kernel virtual memory address). If `key` is `IPC_PRIVATE`, a new segment and its associated book-keeping data are created. If `flag` has the `IPC_CREAT` bit set and there's no segment associated with `key`, a new segment is created in concert with the relevant data.

6.4.2 shmat

```
void *sys_shmat(long id, void *adr, long flag);
```

The `shmat` system call attaches the shared memory segment identified by `id` to the address space of the calling process. If `adr` is `NULL`, the segment is attached to the first address selected by the system. If `adr` is not `NULL` and `flag` has the `SHM_RND` bit set, the segment is mapped to `adr` rounded down to the previous multiple of `SHMLBA`. If `adr` is not `NULL` and `flag` does **not** have the `SHM_RND` bit set, the segment is mapped to `adr`. If `flag` has the `SHM_RDONLY` bit set, the segment is attached **read-only**; otherwise, provided the user process has read and write permissions, the segment is attached **read-write**.

6.4.3 shmdt

```
long sys_shmdt(void *adr);
```

The `shmdt` system call detaches the shared memory segment at `adr` from the address space of the calling process.

6.4.4 shmctl

```
sys_shmctl(long id, long cmd, void *arg);
```

The `shmctl` system call provides control operations for shared memory segments. The possible values for `cmd` are listed in the following table.

cmd	arg	brief
<code>IPC_STAT</code>	struct <code>shmid_ds *</code>	read segment attributes
<code>IPC_SET</code>	struct <code>shmid_ds *</code>	set segment permissions (uid, gid, mode)
<code>IPC_RMID</code>	N/A	destroy shared memory segment

TODO: shared memory, message queues, semaphores, events

6.5 Semaphores

6.6 Message Queues

6.7 Events

6.8 I/O

Part III

User Environment

6.9 Process Environment

6.10 Memory Map

Segment	Brief	Parameters
stack	process stack	read, write, grow-down
map	memory-mapped regions	read, write
heap	process heap (sbrk())	read, write
bss	uninitialised data	read, write, allocate
data	initialised data	read, write
text	process code	read, execute

Notes

- memory regions are shown from highest to lowest address, i.e. the addresses grow upwards
- the stack segment grows downwards in memory
- the BSS segment is allocated at run-time
- segments are shown in descending address order