# 添加系统调用

# 实验目的

学习重建Linux内核。

学习Linux内核的系统调用，理解、掌握Linux系统调用的实现框架、用户界面、参数传递、进入/返回过程。阅读Linux内核源代码，通过添加一个简单的系统调用实验，进一步理解Linux操作系统处理系统调用的统一流程。了解Linux操作系统缺页处理，进一步掌握task\_struct结构的作用。

# 实验内容

在现有的系统中添加一个不用传递参数的系统调用。这个系统调用的功能是实现统计操作系统缺页总次数，当前进程的缺页次数，以及每个进程的“脏”页面数。严格来说这里讲的“缺页次数”实际上是页错误次数，即调用do\_page\_fault函数的次数。实验主要内容：

* 在Linux操作系统环境下重建内核
* 添加系统调用的名字
* 利用标准C库进行包装
* 添加系统调用号
* 在系统调用表中添加相应表项
* 修改统计缺页次数、“脏”页相关的内核结构和函数
* sys\_mysyscall的实现
* 编写用户态测试程序

# 实验指导

下面的指导是以ubuntu 16.04和kernel 4.6.0为例，不同的Linux发行版本和内核版本，实验方法可能会有所不同。添加新的系统调用的步骤：

1. 下载一份内核源代码

Linux受GNU通用公共许可证（GPL）保护，其内核源代码是完全开放的。现在很多Linux的网站都提供内核代码的下载。推荐你使用Linux的官方网站：http://www.kernel.org。在这里你可以找到所有的内核版本。

1. 部署内核源代码

此过程比较机械、枯燥，因而容易出错。请严格按下述步骤来操作。

首先，把linux-4.6.tar.xz包放在主目录下，解开linux-4.6.tar.xz包：

tar –xvf linux-4.6.tar.xz

解压出来的内核代码存放在/usr/src/linux-4.6目录下。为了方便操作及一致性，可以通过路径/usr/src/linux去访问它，这只要建一个符号链接：

ln -s /usr/src/linux-4.6/ /usr/src/linux

1. 配置内核

**第1次编译内核的准备：**

在ubuntu环境下，用命令make menuconfig对内核进行配置时，需要用终端模式下的字符菜单支持软件包libncurses5-dev，因此你是第一次重建内核，需要下载并安装该软件包，下载并安装命令如下：

sudo apt-get install libncurses5-dev

若上面这一步提示错误信息，则输入下面的命令sudo apt-get –f install ，建立库依赖关系。

**查看README文件：**

在你进行这项工作之前，不妨先看一看/usr/src/linux目录下内核源代码自带的README文件。在这份文件中，对怎样进行内核的解压，配置，安装都进行了详细的讲解。

**配置内核：**

在编译内核前，一般来说都需要对内核进行相应的配置。配置是精确控制新内核功能的机会。配置过程也控制哪些需编译到内核的二进制映像中(在启动时被载入)，哪些是需要时才装入的内核模块（module）。

cd /usr/src/linux

第一次编译的话，有必要将内核源代码树置于一种完整和一致的状态。因此，我们推荐执行命令make mrproper。它将清除目录下所有配置文件和先前生成核心时产生的.o文件：

make mrproper

为了与正在运行的操作系统内核的运行环境匹配，可以先把当前已配置好的文件复制到当前目录下，新的文件名为.config文件：

cp /boot/config-`uname -r` .config

这里，命令`uname –r`得到当前内核版本号。然后：

make menuconfig

进行配置时，大部分选项可以使用其缺省值，只有小部分需要根据用户不同的需要选择。例如，如果硬盘分区采用ext2文件系统（或ext3文件系统），则配置项应支持ext2文件系统（ext3文件系统）。又例如，系统如果配有SCSI总线及设备，需要在配置中选择SCSI卡的支持。

对每一个配置选项，用户有三种选择，它们分别代表的含义如下：

* “<\*>”或“[\*]” － 将该功能编译进内核
* “[ ]” － 不将该功能编译进内核
* “[M]” － 将该功能编译成可以在需要时动态插入到内核中的模块

将与核心其它部分关系较远且不经常使用的部分功能代码编译成为可加载模块，有利于减小内核的长度，减小内核消耗的内存，简化该功能相应的环境改变时对内核的影响。许多功能都可以这样处理，例如像上面提到的对SCSI卡的支持，等等。

1. 添加系统调用号

系统调用号在文件unistd.h里面定义。这个文件可能在你的Linux系统上会有两个版本：一个是C库文件版本，出现的地方是在 ubuntu 16.04 只带的/usr/include/asm-generic/unistd.h；另外还有一个版本是内核自己的unistd.h，出现的地方是在你解压出来的内核代码的对应位置（比如include/uapi/asm-generic/unistd.h）。当然，也有可能这个C库文件只是一个对应到内核文件的连接。现在，你要做的就是在文件unistd.h中添加我们的系统调用号：\_\_NR\_mysyscall，x86体系架构的系统调用号223没有使用，我们新的系统调用号定义为223号，如下所示：

ubuntu 16.04为：/usr/include/asm-generic/unistd.h

kernel 4.6为：include/uapi/asm-generic/unistd.h

在/usr/include/asm-generic/unistd.h文件中的查找定义223号的行，作如下修改：

-- #define \_\_NR3264\_fadvise64 223

--\_\_SC\_COMP(\_\_NR3264\_fadvise64, sys\_fadvise64\_64, compat\_sys\_fadvise64\_64)

++ #define \_\_NR\_mysyscall 223

++ \_\_SYSCALL(\_\_NR\_mysyscall, sys\_mysyscall)

在文件include/uapi/asm-generic/unistd.h中做同样的修改

注意：不同版本的内核，需要修改路径（子目录名和文件名）可能不一样，根据实际版本号查找这些.h文件。系统调用号也可能不一样，可以根据内核版本不同对系统调用号进行修改。

添加系统调用号之后，系统才能根据这个号，作为索引，去找syscall­\_table中的相应表项。所以说，我们接下来的一步就是：

1. 在系统调用表中添加或修改相应表项

我们前面讲过，系统调用处理程序（system\_call）会根据eax中的索引到系统调用表（sys\_call\_table）中去寻找相应的表项。所以，我们必须在那里添加我们自己的一个值。

arch/x86/entry/syscalls/syscall\_32.tbl

**# 222 is unused**

**223 i386 mysyscall sys\_mysyscall**

**224 i386 gettid sys\_gettid**

**225 i386 readahead sys\_readahead**

到现在为止，系统已经能够正确地找到并且调用sys\_mysyscall。剩下的就只有一件事情，那就是sys\_mysyscall的实现。

1. 修改统计系统缺页次数和进程缺页次数的内核代码

由于每发生一次缺页都要进入缺页中断服务函数do\_page\_fault一次，所以可以认为执行该函数的次数就是系统发生缺页的次数。可以定义一个全局变量pfcount作为计数变量，在执行do\_page\_fault时，该变量值加1。在当前进程控制块中定义一个变量pf记录当前进程缺页次数，在执行do\_page\_fault时，这个变量值加1。

先在include/linux/mm.h文件中声明变量pfcount：

++ extern unsigned long pfcount;

要记录进程产生的缺页次数，首先在进程task\_struct中增加成员pf，在include/linux/sched.h文件中（第1394行jjj）的task\_struct结构中添加pf字段：

++ unsigned long pf;

统计当前进程缺页次数需要在创建进程是需要将进程控制块中的pf设置为0，在进程创建过程中，子进程会把父进程的进程控制块复制一份，实现该复制过程的函数是kernel/fork.c文件中的dup\_task\_struct()函数，修改该函数将子进程的pf设置成0：

static struct task\_struct \*dup\_task\_struct(struct task\_struct \*orig)

{

…..

tsk = alloc\_task\_struct\_node(node);

if (!tsk)

return NULL;

……

++ tsk->pf=0;

……

}

在arch/x86/mm/fault.c文件中定义变量pfcount；并修改arch/x86/mm/fault.c中do\_page\_fault()函数。每次产生缺页中断，do\_page\_fault()函数会被调用，pfcount变量值递增1,记录系统产生缺页次数，current->pf值递增1，记录当前进程产生缺页次数：

…

++ unsigned long pfcount;

\_\_do\_page\_fault(struct pt\_regs \*regs, unsigned long error\_code)

{

…

++ pfcount++;

++ current->pf++;

…

}

1. sys\_mysyscall的实现

我们把这一小段程序添加在kernel/sys.c里面。在这里，我们没有在kernel目录下另外添加自己的一个文件，这样做的目的是为了简单，而且不用修改Makefile，省去不必要的麻烦。

mysyscall系统调用实现输出系统缺页次数、当前进程缺页次数，及每个进程的“脏”页面数。

asmlinkage int sys\_mysyscall(void)

{

……

//printk(“当前进程缺页次数：%lu,current->pf”)

//每个进程的“脏”页面数

return 0;

}

1. 编译内核和重启内核

用make工具编译内核：

sudo make bzImage –j2

编译内核需要较长的时间，具体与机器的硬件条件及内核的配置等因素有关。完成后产生的内核文件bzImage的位置在/usr/src/linux/arch/i386/boot目录下，当然这里假设用户的CPU是Intel x86型的，并且你将内核源代码放在/usr/src/linux目录下。

如果编译过程中产生错误，你需要检查第4、5、6、7步修改的代码是否正确，修改后要再次使用make命令编译，直至编译成功。

如果选择了可加载模块，编译完内核后，要对选择的模块进行编译，然后安装。用下面的命令编译模块并安装到标准的模块目录中：

sudo make modules –j2

sudo make modules\_install

通常，Linux在系统引导后从/boot目录下读取内核映像到内存中。因此我们如果想要使用自己编译的内核，就必须先将启动文件安装到/boot目录下。安装内核命令：

sudo make install

grub是管理[ubuntu](http://www.linuxidc.com/topicnews.aspx?tid=2" \o "Ubuntu" \t "_blank)系统启动的一个程序。想运行刚刚编译好的内核，就要修改对应的grub。

sudo mkinitramfs 4.6.0 -o /boot/initrd.img-4.6.0

sudo update-grub2

我们已经编译了内核bzImage，放到了指定位置/boot。现在，请你重启主机系统，期待编译过的Linux操作系统内核正常运行！

sudo reboot

1. 编写用户态程序

要测试新添加的系统调用，需要编写一个用户态测试程序（test.c）调用mysyscall系统调用。mysyscall系统调用中printk函数输出的信息在/var/log/messages文件中(ubuntu 为/var/log/kern.log文件)。/var/log/messages (ubuntu 为/var/log/kern.log文件)文件中的内容也可以在shell下用dmesg命令查看到。

用户态程序

#include <linux/unistd.h>

# include <sys/syscall.h>

#define \_\_NR\_ mysyscall 223

int main()

{

syscall(\_\_NR\_mysyscall);

//……

}

* 用gcc编译源程序

gcc –o test test.c

* 运行程序

./test

# 完成实验后回答问题：

1. 多次运行test程序，每次运行test后记录下系统缺页次数和当前进程缺页次数，给出这些数据。test程序打印的缺页次数是否就是操作系统原理上的缺页次数？有什么区别？
2. 除了通过修改内核来添加一个系统调用外，还有其他的添加或修改一个系统调用的方法吗？如果有，请论述。
3. 对于一个操作系统而言，你认为修改系统调用的方法安全吗？请发表你的观点。

# 参考文献

[1] ubuntu内核，在<http://kernel.ubuntu.com/~kernel-ppa/mainline/>

[2] Linux官方内核，在<https://www.kernel.org/pub/linux/kernel/>

[3] Linux公社<http://www.linuxidc.com/Linux/2016-06/132707.htm>

[4] ubuntu上安装 Linux Kernel 4.4，http://www.linuxidc.com/Linux/2016-01/127383.htm

# 撰写实验报告的要求

1. 按照实验报告模板格式撰写。
2. 整个实验过程的截图。
3. 源程序的修改部分，运行结果的截图。
4. 必须撰写实验讨论（即心得体会），内容为实验过程中遇到的问题及解决方法等。否则扣除本实验20%分数。
5. 实验报告中需要附上mysyscall()和test.c的源代码，完成后把实验报告上传到“作业系统”中。