41 | IPC (中): 不同项目组之间抢资源, 如何协调?

2019-07-01 刘紹

趣谈Linux操作系统

讲入课程 >



讲述: 刘超

时长 17:32 大小 16.06M



了解了如何使用共享内存和信号量集合之后, 今天我们来解 析一下, 内核里面都做了什么。

不知道你有没有注意到,咱们讲消息队列、共享内存、信号量的机制的时候,我们其实能够从中看到一些统一的规律:它们在使用之前都要生成 key,然后通过 key 得到唯一的id,并且都是通过 xxxget 函数。

在内核里面,这三种进程间通信机制是使用统一的机制管理起来的,都叫 ipcxxx。

为了维护这三种进程间通信进制,在内核里面,我们声明了一个有三项的数组。

我们通过这段代码,来具体看一看。

■ 复制代码

←

根据代码中的定义,第 0 项用于信号量,第 1 项用于消息队列,第 2 项用于共享内存,分别可以通过 sem_ids、msg_ids、shm_ids 来访问。

这段代码里面有 ns,全称叫 namespace。可能不容易理解,你现在可以将它认为是将一台 Linux 服务器逻辑的隔离为多台 Linux 服务器的机制,它背后的原理是一个相当大的话题,我们需要在容器那一章详细讲述。现在,你就可以简单的认为没有 namespace,整个 Linux 在一个namespace下面,那这些 ids 也是整个 Linux 只有一份。

接下来,我们再来看 struct ipc ids 里面保存了什么。

首先, in_use 表示当前有多少个 ipc; 其次, seq 和 next_id 用于一起生成 ipc 唯一的 id, 因为信号量, 共享内存, 消息队列, 它们三个的 id 也不能重复; ipcs_idr 是一棵基数树, 我们又碰到它了, 一旦涉及从一个整数查找一个对象, 它都是最好的选择。

```
1 struct ipc_ids {
2     int in_use;
3     unsigned short seq;
4     struct rw semaphore rwsem;
```

```
5     struct idr ipcs_idr;
6     int next_id;
7 };
8
9 struct idr {
10     struct radix_tree_root idr_rt;
11     unsigned int idr_next;
12 };
```

也就是说,对于 sem_ids、msg_ids、shm_ids 各有一棵基数树。那这棵树里面究竟存放了什么,能够统一管理这三类ipc 对象呢?

通过下面这个函数 ipc_obtain_object_idr, 我们可以看出端倪。这个函数根据 id, 在基数树里面找出来的是 struct kern_ipc_perm。

■ 复制代码

•

如果我们看用于表示信号量、消息队列、共享内存的结构,就会发现,这三个结构的第一项都是 struct kern ipc perm。

```
1 struct sem array {
           struct kern ipc perm
2
                                  sem perm;
                                                  /* perm
3
          time t
                                   sem ctime;
                                                /* crea
           struct list head
                                   pending alter; /* pend
5
                                   pending_const; /* pend
           struct list head
7
                                                   /* that
          struct list head
                                   list id;
                                                  /* undc
8
           int
                                   sem nsems;
                                                  /* no.
                                   complex count; /* penc
10
          int
                                   use global lock;/* >0:
11
          unsigned int
12
13
           struct sem
                                   sems[];
14 } randomize layout;
15
16 struct msg queue {
           struct kern_ipc_perm q_perm;
17
           time_t q_stime;
18
                                           /* last msgsnd
          time t q rtime;
                                           /* last msgrcv
19
          time t q ctime;
                                          /* last change
20
          unsigned long q cbytes;
                                          /* current numb
21
          unsigned long q qnum;
                                          /* number of me
22
                                           /* max number c
23
          unsigned long q qbytes;
                                           /* pid of last
24
          pid t q lspid;
           pid t q lrpid;
                                           /* last receive
25
27
           struct list head q messages;
           struct list head q receivers;
28
```

```
29
           struct list head q senders;
30 } randomize layout;
31
  struct shmid kernel /* private to the kernel */
33 {
34
           struct kern ipc perm
                                    shm perm;
                                    *shm file;
           struct file
                                    shm nattch:
           unsigned long
37
           unsigned long
                                    shm segsz;
38
           time t
                                    shm atim;
39
           time t
                                    shm dtim;
40
           time t
                                    shm ctim:
41
           pid t
                                    shm cprid;
           pid t
                                    shm lprid;
42
43
           struct user struct
                                    *mlock user;
45
           /* The task created the shm object. NULL if th
           struct task struct
                                   *shm creator;
46
47
           struct list head
                                   shm clist;
                                                /* list
48 } randomize layout;
```

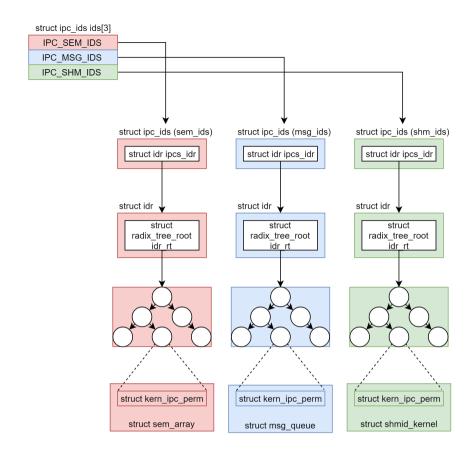
也就是说,我们完全可以通过 struct kern_ipc_perm 的指针,通过进行强制类型转换后,得到整个结构。做这件事情的函数如下:

```
static inline struct sem_array *sem_obtain_object(struc

{
struct kern_ipc_perm *ipcp = ipc_obtain_object_
return container_of(ipcp, struct sem_array, sem_array, sem_array
```

```
5 }
7 static inline struct msg queue *msq obtain object(struc
8 {
           struct kern ipc perm *ipcp = ipc obtain object
9
           return container of(ipcp, struct msg queue, q p
10
11 }
12
13 static inline struct shmid kernel *shm obtain object(st
14 {
15
           struct kern ipc perm *ipcp = ipc obtain object
16
           return container of(ipcp, struct shmid kernel,
17 }
```

通过这种机制,我们就可以将信号量、消息队列、共享内存抽象为 ipc 类型进行统一处理。你有没有觉得,这有点儿面向对象编程中抽象类和实现类的意思?没错,如果你试图去了解 C++ 中类的实现机制,其实也是这么干的。



有了抽象类,接下来我们来看共享内存和信号量的具体实现。

如何创建共享内存?

首先,我们来看创建共享内存的的系统调用。

```
2 {
           struct ipc namespace *ns;
           static const struct ipc ops shm ops = {
 5
                    .getnew = newseg,
                    .associate = shm security,
                    .more checks = shm more checks,
8
           };
9
           struct ipc params shm params;
           ns = current->nsproxy->ipc ns;
10
11
           shm params.key = key;
12
           shm params.flg = shmflg;
13
           shm params.u.size = size;
14
           return ipcget(ns, &shm ids(ns), &shm ops, &shm
15 }
```

这里面调用了抽象的 ipcget、参数分别为共享内存对应的 shm_ids、对应的操作 shm_ops 以及对应的参数 shm_params。

如果 key 设置为 IPC_PRIVATE 则永远创建新的,如果不是的话,就会调用 ipcget_public。ipcget 的具体代码如下:

```
7
                    return ipcget public(ns, ids, ops, para
8 }
9
  static int ipcget public(struct ipc namespace *ns, stru
11 {
12
           struct kern ipc perm *ipcp;
13
           int flg = params->flg;
14
           int err:
           ipcp = ipc_findkey(ids, params->key);
15
           if (ipcp == NULL) {
16
                    if (!(flg & IPC_CREAT))
17
18
                            err = -ENOENT:
                    else
                            err = ops->getnew(ns, params);
21
           } else {
                    if (flg & IPC_CREAT && flg & IPC EXCL)
22
23
                            err = -EEXIST;
24
                    else {
25
                            err = 0;
                            if (ops->more checks)
27
                                     err = ops->more checks(
28 .....
                    }
30
           return err;
31
32 }
```

在 ipcget_public 中,我们会按照 key,去查找 struct kern_ipc_perm。如果没有找到,那就看是否设置了 IPC_CREAT; 如果设置了,就创建一个新的。如果找到了,就将对应的 id 返回。

我们这里重点看,如何按照参数 shm_ops,创建新的共享内存,会调用 newseg。

```
1 static int newseg(struct ipc namespace *ns, struct ipc
2 {
 3
           key t key = params->key;
           int shmflg = params->flg;
4
           size t size = params->u.size;
 5
           int error:
 7
           struct shmid kernel *shp;
           size t numpages = (size + PAGE SIZE - 1) >> PAGE
8
           struct file *file;
9
           char name[13];
10
           vm flags t acctflag = 0;
11
12 .....
           shp = kvmalloc(sizeof(*shp), GFP KERNEL);
13
14 .....
          shp->shm perm.key = key;
15
          shp->shm perm.mode = (shmflg & S IRWXUGO);
           shp->mlock user = NULL;
17
18
           shp->shm perm.security = NULL;
19
20 .....
           file = shmem kernel file setup(name, size, acct
21
22 .....
23
           shp->shm cprid = task tgid vnr(current);
           shp->shm lprid = 0;
24
           shp->shm atim = shp->shm dtim = 0;
25
           shp->shm ctim = get seconds();
26
27
           shp->shm segsz = size;
           shp->shm nattch = 0;
28
           shp->shm file = file;
29
```

```
30
           shp->shm creator = current;
31
           error = ipc addid(&shm ids(ns), &shp->shm perm,
32
33 .....
34
           list add(&shp->shm clist, &current->sysvshm.shm
35 .....
           file inode(file)->i ino = shp->shm perm.id;
37
38
          ns->shm tot += numpages;
          error = shp->shm perm.id;
39
40 .....
41
          return error:
42 }
```

newseg 函数的第一步,通过 kvmalloc 在直接映射区分配一个 struct shmid_kernel 结构。这个结构就是用来描述共享内存的。这个结构最开始就是上面说的 struct kern_ipc_perm 结构。接下来就是填充这个 struct shmid_kernel 结构,例如 key、权限等。

newseg 函数的第二步,共享内存需要和文件进行关联。** 为什么要做这个呢?我们在讲内存映射的时候讲过,虚拟地址空间可以和物理内存关联,但是物理内存是某个进程独享的。虚拟地址空间也可以映射到一个文件,文件是可以跨进程共享的。 咱们这里的共享内存需要跨进程共享,也应该借鉴文件映射的思路。只不过不应该映射一个硬盘上的文件,而是映射到一个内存文件系统上的文件。mm/shmem.c 里面就定义了这样一个基于内存的文件系统。这里你一定要注意区分shmem 和 shm 的区别,前者是一个文件系统,后者是进程通信机制。

在系统初始化的时候, shmem_init 注册了 shmem 文件系统 shmem_fs_type, 并且挂在到了 shm_mnt 下面。

```
1 int init shmem init(void)
2 {
           int error;
3
           error = shmem init inodecache();
4
           error = register filesystem(&shmem fs type);
           shm mnt = kern mount(&shmem fs type);
8
          return 0;
9 }
10
11 static struct file system type shmem fs type = {
                           = THIS_MODULE,
12
           .owner
13
                          = "tmpfs",
           .name
                          = shmem mount,
14
           .mount
           .kill sb
                          = kill litter super,
15
           .fs flags
                          = FS USERNS MOUNT,
17 };
```

接下来, newseg 函数会调用 shmem_kernel_file_setup, 其实就是在 shmem 文件系统里面创建一个文件。

```
1 /**
* shmem kernel file setup - get an unlinked file livir
   * @name: name for dentry (to be seen in /proc/<pid>/ma
* @size: size to be set for the file
* @flags: VM NORESERVE suppresses pre-accounting of th
6 struct file *shmem kernel file setup(const char *name,
           return shmem file setup(name, size, flags, S
8
9 }
10
11 static struct file * shmem file setup(const char *name
                                         unsigned long fl
12
13 {
14
          struct file *res;
         struct inode *inode;
15
       struct path path;
          struct super block *sb;
17
18
          struct qstr this;
19 .....
          this.name = name;
20
          this.len = strlen(name);
21
          this.hash = 0; /* will go */
22
23
          sb = shm mnt->mnt sb;
          path.mnt = mntget(shm mnt);
24
          path.dentry = d alloc pseudo(sb, &this);
25
          d set d op(path.dentry, &anon_ops);
26
27 .....
           inode = shmem get inode(sb, NULL, S IFREG | S I
28
           inode->i flags |= i flags;
29
```

__shmem_file_setup 会创建新的 shmem 文件对应的 dentry 和 inode,并将它们两个关联起来,然后分配一个 struct file 结构,来表示新的 shmem 文件,并且指向独特的 shmem_file_operations。

```
static const struct file operations shmem file operatic
2
           .mmap
                           = shmem mmap,
           .get unmapped area = shmem get unmapped area,
4 #ifdef CONFIG TMPFS
           .llseek
                           = shmem file llseek,
5
                           = shmem file read iter,
           .read iter
                           = generic file write iter,
           .write iter
           .fsvnc
                           = noop fsync,
8
           .splice read = generic file splice read,
9
           .splice write = iter file splice write,
10
                          = shmem fallocate,
           .fallocate
11
12 #endif
13 };
```

newseg 函数的第三步,通过 ipc_addid 将新创建的 struct shmid_kernel 结构挂到 shm_ids 里面的基数树上,并返回相应的 id,并且将 struct shmid_kernel 挂到 当前进程的 sysvshm 队列中。

至此, 共享内存的创建就完成了。

如何将共享内存映射到虚拟地址空间?

从上面的代码解析中,我们知道,共享内存的数据结构 struct shmid_kernel,是通过它的成员 struct file *shm_file,来管理内存文件系统 shmem 上的内存文件 的。无论这个共享内存是否被映射, shm file 都是存在的。

接下来,我们要将共享内存映射到虚拟地址空间中。调用的是 shmat,对应的系统调用如下:

```
SYSCALL_DEFINE3(shmat, int, shmid, char __user *, shmad

unsigned long ret;

long err;

err = do_shmat(shmid, shmaddr, shmflg, &ret, SHMLBA force_successful_syscall_return();

return (long)ret;

}
```

```
10 long do shmat(int shmid, char user *shmaddr, int shmf
                 ulong *raddr, unsigned long shmlba)
11
12 {
13
           struct shmid kernel *shp;
14
           unsigned long addr = (unsigned long)shmaddr;
           unsigned long size;
           struct file *file;
16
17
           int err:
18
           unsigned long flags = MAP SHARED;
           unsigned long prot;
19
           int acc mode;
20
21
           struct ipc namespace *ns;
22
           struct shm file data *sfd;
23
           struct path path;
24
           fmode t f mode;
           unsigned long populate = 0;
25
26 .....
           prot = PROT READ | PROT WRITE;
27
28
          acc mode = S IRUGO | S IWUGO;
29
           f mode = FMODE READ | FMODE WRITE;
30 .....
           ns = current->nsproxy->ipc ns;
31
           shp = shm obtain object check(ns, shmid);
33 .....
           path = shp->shm file->f path;
34
           path get(&path);
           shp->shm nattch++;
37
           size = i size read(d inode(path.dentry));
38 .....
           sfd = kzalloc(sizeof(*sfd), GFP KERNEL);
39
40 .....
           file = alloc file(&path, f mode,
41
                              is file hugepages(shp->shm fi
42
                                    &shm file operations hu
43
                                    &shm file operations);
44
```

```
45 .....
           file->private data = sfd;
46
           file->f mapping = shp->shm file->f mapping;
47
           sfd->id = shp->shm perm.id;
48
49
           sfd->ns = get ipc ns(ns);
           sfd->file = shp->shm file;
50
           sfd->vm ops = NULL;
51
52 .....
           addr = do mmap pgoff(file, addr, size, prot, fl
53
           *raddr = addr;
54
          err = 0;
56 .....
57
           return err;
58 }
```

在这个函数里面,shm_obtain_object_check 会通过共享内存的 id,在基数树中找到对应的 struct shmid_kernel 结构,通过它找到 shmem 上的内存文件。

接下来,我们要分配一个 struct shm_file_data,来表示这个内存文件。将 shmem 中指向内存文件的 shm_file 赋值给 struct shm_file_data 中的 file 成员。

然后,我们创建了一个 struct file,指向的也是 shmem 中的内存文件。

为什么要再创建一个呢?这两个的功能不同,shmem中 shm_file 用于管理内存文件,是一个中立的,独立于任何一个进程的角色。而新创建的 struct file 是专门用于做内存映射的,就像咱们在讲内存映射那一节讲过的,一个硬盘上的文件要映射到虚拟地址空间中的时候,需要在 vm_area_struct 里面有一个 struct file *vm_file 指向硬盘上的文件,现在变成内存文件了,但是这个结构还是不能少。

新创建的 struct file 的 private_data, 指向 struct shm_file_data, 这样内存映射那部分的数据结构, 就能够通过它来访问内存文件了。

新创建的 struct file 的 file_operations 也发生了变化,变成了 shm file operations。

接下来,do_mmap_pgoff 函数我们遇到过,原来映射硬盘上的文件的时候,也是调用它。这里我们不再详细解析了。它会分配一个 vm_area_struct 指向虚拟地址空间中没有分配的区域,它的 vm_file 指向这个内存文件,然后它会调用 shm_file_operations 的 mmap 函数,也即 shm_mmap 进行映射。

```
■ 复制代码
```

shm_mmap 中调用了 shm_file_data 中的 file 的 mmap 函数,这次调用的是 shmem_file_operations 的 mmap,也即 shmem_mmap。

```
1 static int shmem_mmap(struct file *file, struct vm_area
2 {
```

```
file_accessed(file);

vma->vm_ops = &shmem_vm_ops;

return 0;

}
```

这里面, vm_area_struct 的 vm_ops 指向 shmem_vm_ops。等从 call_mmap 中返回之后, shm_file_data 的 vm_ops 指向了 shmem_vm_ops,而 vm_area_struct 的 vm_ops 改为指向 shm_vm_ops。

我们来看一下, shm_vm_ops 和 shmem_vm_ops 的定义。

```
1 static const struct vm_operations_struct shm_vm_ops = {
2
          .open = shm open,
                                  /* callback for a new v
           .close = shm close,
                                  /* callback for when th
3
           .fault = shm fault,
5 };
7 static const struct vm operations struct shmem vm ops =
           .fault
                         = shmem fault,
8
           .map pages
                         = filemap map pages,
10 };
```

它们里面最关键的就是 fault 函数,也即访问虚拟内存的时候,访问不到应该怎么办。

当访问不到的时候,先调用 vm_area_struct 的 vm_ops,也即 shm_vm_ops 的 fault 函数 shm_fault。然后它会转而调用 shm_file_data 的 vm_ops,也即 shmem_vm_ops的 fault 函数 shmem_fault。

```
■ 复制代码
```

虽然基于内存的文件系统,已经为这个内存文件分配了 inode,但是内存也却是一点儿都没分配,只有在发生缺页 异常的时候才进行分配。

```
6 .....
    error = shmem getpage gfp(inode, vmf->pgoff, &v
8
                                     gfp, vma, vmf, &ret):
9 .....
10 }
11
12 /*
* shmem getpage gfp - find page in cache, or get from
14
   * If we allocate a new one we do not mark it dirty. Th
15
   * vm. If we swap it in we mark it dirty since we also
16
17
   * entry since a page cannot live in both the swap and
18
   * fault mm and fault type are only supplied by shmem f
19
20
   * otherwise they are NULL.
21
    */
22 static int shmem getpage gfp(struct inode *inode, pgoff
23
           struct page **pagep, enum sgp type sgp, gfp t g
24
           struct vm area struct *vma, struct vm fault *vm
25 {
26 .....
   page = shmem alloc and acct page(gfp, info, sbinfo,
27
                                           index, false);
28
29 .....
30 }
                                                        •
```

shmem_fault 会调用 shmem_getpage_gfp 在 page cache 和 swap 中找一个空闲页,如果找不到就通过 shmem_alloc_and_acct_page 分配一个新的页,他最终会

调用内存管理系统的 alloc_page_vma 在物理内存中分配一个页。

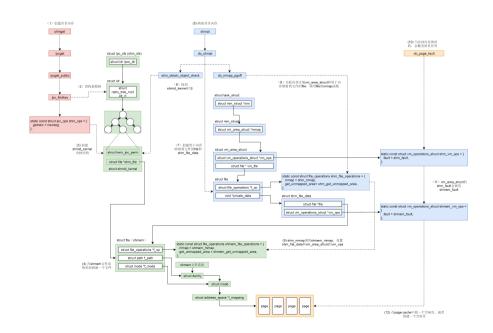
至此, 共享内存才真的映射到了虚拟地址空间中, 进程可以像访问本地内存一样访问共享内存。

总结时刻

我们来总结一下共享内存的创建和映射过程。

- 1. 调用 shmget 创建共享内存。
- 2. 先通过 ipc_findkey 在基数树中查找 key 对应的共享内存对象 shmid_kernel 是否已经被创建过,如果已经被创建,就会被查询出来,例如 producer 创建过,在consumer 中就会查询出来。
- 3. 如果共享内存没有被创建过,则调用 shm_ops 的 newseg 方法,创建一个共享内存对象 shmid_kernel。 例如,在 producer 中就会新建。
- 4. 在 shmem 文件系统里面创建一个文件, 共享内存对象 shmid_kernel 指向这个文件, 这个文件用 struct file 表示, 我们姑且称它为 file1。
- 5. 调用 shmat,将共享内存映射到虚拟地址空间。
- 6. shm_obtain_object_check 先从基数树里面找到 shmid kernel 对象。

- 7. 创建用于内存映射到文件的 file 和 shm_file_data,这里的 struct file 我们姑且称为 file2。
- 8. 关联内存区域 vm_area_struct 和用于内存映射到文件的 file, 也即 file2, 调用 file2的 mmap 函数。
- 9. file2 的 mmap 函数 shm_mmap, 会调用 file1 的 mmap 函数 shmem_mmap, 设置 shm_file_data 和 vm_area_struct 的 vm_ops。
- 10. 内存映射完毕之后,其实并没有真的分配物理内存,当访问内存的时候,会触发缺页异常 do_page_fault。
- 11. vm_area_struct 的 vm_ops 的 shm_fault 会调用 shm_file_data 的 vm_ops 的 shmem_fault。
- 12. 在 page cache 中找一个空闲页,或者创建一个空闲页。



课堂练习

在这里,我们只分析了 shm_ids 的结构,消息队列的程序 我们写过了,但是 msg_ids 的结构没有解析,你可以试着 解析一下。

欢迎留言和我分享你的疑惑和见解, 也欢迎可以收藏本节内容, 反复研读。你也可以把今天的内容分享给你的朋友, 和他一起学习和进步。



© 版权归极客邦科技所有,未经许可不得传播售卖。 页面已增加防盗追踪,如有侵权极客邦将依法追究其法律责任。

上一篇 40 | IPC (上): 不同项目组之间抢资源, 如何协调?

下一篇 42 | IPC (下): 不同项目组之间抢资源, 如何协调?

精选留言(2)





Amark

2019-07-02

老师有没有什么通俗易懂的资料,您将的太专业了 _{展开}~







不一样的烟火

2019-07-01

听完了 快点更新 ₩

展开~

作者回复: 牛



