

Sincronización en el kernel Linux

Arquitectura Interna de Linux - 2016





Contenido



1 Origen de la concurrencia en Linux

2 Primitivas de sincronización del kernel Linux



Contenido



1 Origen de la concurrencia en Linux

2 Primitivas de sincronización del kernel Linux



Flujos de ejecución en el kernel



- Las funciones del kernel Linux se invocan en respuesta a *eventos*
 - El kernel se comporta como un "servidor" que responde peticiones
 - 1 Llamadas al sistema
 - 2 Excepciones de los programas de usuario
 - 3 Interrupciones de los dispositivos de E/S
- Flujo de ejecución del kernel: secuencia de instrucciones que se ejecutan en modo kernel
 - Más ligero que un proceso (menos contexto)
 - No siempre desempeña acciones en nombre de un proceso de usuario



Ejecución en una CPU



- En un instante *t* cada CPU del sistema puede estar ejecutando:
 - Código de un proceso en modo usuario (Contexto de Proceso en modo usuario)
 - Código del kernel en una llamada al sistema o manejador de excepción (Contexto de Proceso en modo kernel)
 - Código de un kernel thread (Contexto de Proceso en modo kernel)
 - Código que realiza procesamiento ligado a una interrupción (Contexto de Interrupción)

1	I	ı	ı	I	ı	ı	ı	I	I
user		user		user		l I	l I	user	User Mode
	kernel		i – – – I	i I	kernel		kernel		Kernel Mode
		i I	intr			intr		 	



Recuerda: Expropiación de usuario



- El SO puede expulsar a un proceso de la CPU mientras se ejecuta en modo usuario y poner a otro proceso en su lugar
 - Planificadores expropiativos: RR, Colas multinivel, CFS, ...
- La expropiación de usuario requiere soporte HW
 - El SO configura el temporizador del sistema para que genere interrupciones periódicas (tick)
 - Interrupción del timer → transición a modo kernel → activación del planificador (scheduler_tick())
 - Al retornar a modo usuario tras la interrupción se podría ejecutar otro proceso



Expropiación de kernel



- El kernel Linux es **expropiativo** (*preemptive*)
 - Cualquier proceso ejecutándose en modo kernel (contexto de proceso en modo kernel) podría ser reemplazado por otro proceso casi en cualquier instante
 - Por ejemplo, expropiación durante la ejecución de una llamada al sistema (handler function)
- Objetivo expropiación: reducir el tiempo de respuesta (latencia) de los procesos de usuario
 - Latencia: tiempo transcurrido desde que un proceso entra en el estado "listo para ejecutar" hasta que se pone a ejecutar en una CPU
- Durante la ejecución de una llamada al sistema se puede producir expropiación en el kernel siempre y cuando no se haya deshabilitado explícitamente la expropiación



preempt_disable(), preempt_enable()

Origen de la concurrencia en Linux



- Situaciones que provocan la ejecución de flujos concurrentes/entrelazados en el kernel:
 - 1 Interrupciones
 - Una interupción puede ocurrir de forma asíncrona en casi cualquier instante, interrumpiendo al código que está actualmente en ejecución
 - 2 Bottom-half processing
 - 3 Expropiación de kernel
 - Como el kernel es expropiativo, un flujo de ejecución del kernel puede expropiar a otro
 - 4 Bloqueos y sincronización con el modo usuario
 - Un proceso que se ejecuta en modo kernel puede bloquearse y por tanto invocar al planificador (schedule()) para permitir la ejecución de otro proceso
 - 5 Multiprocesamiento simétrico (SMP)
 - Dos flujos de ejecución del kernel podrían ejecutarse al mismo tiempo en distintas CPUs



Situaciones que requieren sincronización



- Una condición de carrera puede ocurrir cuando el resultado de un cómputo/procesamiento depende de cómo se entrelace la ejecución de instrucciones de dos o más flujos de ejecución del kernel
- Para evitar condiciones de carrera debe garantizarse exclusión mutua entre secciones críticas (SCs)
 - En sistemas monoprocesador, desactivar interrupcciones dentro de la SC garantiza exclusión mutua
 - En sistemas multiprocesador/multicore es más complicado garantizar exclusión mutua
 - Necesario usar primitivas de sincronización



Contenido



1 Origen de la concurrencia en Linux

2 Primitivas de sincronización del kernel Linux



Primitivas de sincronización en Linux

ArTeC.



Técnica	Descripción	Ámbito
Deshabilitar localmente las interrupciones	Ignorar temporalmente las interrupciones en una CPU	Local
Operaciones Atómicas	Instrucciones RMW (read-modify-write) atómicas	Global
Variables <i>Per-CPU</i>	Replicar una estructura de datos en distintas CPUs	Global
Barreras de memoria	Evitar el reordenamiento de instrucciones	Global
Spin locks	Cerrojos de espera activa	Global
Semáforos	Mecanismo de sincronización de espera bloqueante	Global
Read-copy-update (RCU)	Acceso sin cerrojos a datos compartidos vía punteros	Global
Seqlocks	Cerrojo basado en contador de acceso	Global
Variables completion	Colas de espera	Global

Deshabilitar interrupciones



Exclusión mutua en UP

```
local_irq_disable();
.. Sección crítica ..
local_irq_enable();
```

Exclusión mutua en UP (seguro)

```
unsigned long flags=0;
local_irq_save(flags);
.. Sección crítica ..
local_irq_restore(flags);
```

Operaciones <linux/irqflags.h>

- local_irq_disable(): Deshabilita las interrupciones en la CPU local
- local_irq_enable(): Habilita las interrupciones en la CPU local
- local_irq_save(flags): Almacena en flags el estado actual de las interrupciones y deshabilita las interrupciones en la CPU local
- local_irq_restore(flags): Restaura el antiguo estado de las interrupciones almacenado en flags



Operaciones atómicas sobre enteros



- Tipo atomic_t: contador atómico de 32 bits
 - Uso en fragmentos de código del kernel que acceden concurrentemente a una variable entera compartida
 - atomic_t soporta conjunto de operaciones atómicas
 - Declarado en <asm/atomic.h>
 - Implementación dependiente de arquitectura mediante instrucciones RMW (read-modify-write)
- Tipo atomic64_t: contador atómico de 64 bits
 - Uso análogo a atomic_t
 - Declarado en <asm/atomic64.h>



Ejemplo: Operaciones atómicas sobre enteros



int contador=0;

CPU 0 CPU 1

contador++; contador+=2;



Ejemplo: Operaciones atómicas sobre enteros



```
int contador=0;
```

CPU 0 CPU 1

contador++; contador+=2;

atomic_t contador=ATOMIC_INIT(0);

CPU 0 CPU 1

atomic_inc(&contador); atomic_add(2,&contador);



Operaciones atómicas: atomic_t



Función/Macro	Descripción
ATOMIC_INIT(i)	En declaración, inicializa el contador a i
$atomic_read(v)^1$	Devuelve el valor entero asociado
atomic_set(v,i)	Asigna i al contador
atomic_add(i,v)	Incrementa el contador en i unidades
atomic_sub(i,v)	Decrementa i unidades el contador
atomic_sub_and_test(i,v)	Decrementa i unidades el contador y devuelve 1 si el resultado es 0
atomic_inc(v)	Incrementa el contador
atomic_dec(v)	Decrementa el contador
atomic_dec_and_test(v)	Decrementa el contador y devuelve 1 si el resultado es 0
atomic_inc_and_test(v)	Incrementa el contador y devuelve 1 si el resultado es 0
atomic_add_negative(v)	Incrementa el contador en ${\tt i}$ unidades y devuelve ${\tt l}$ si el resultado es negativo

 $^{^{1}\}mathrm{v}$ ha de ser un puntero a atomic_t (valor_por referencia).

Operaciones atómicas: atomic64_t



	Función/Macro	Descripción
	ATOMIC64_INIT(i)	En declaración, inicializa el contador a i
	$atomic64_read(v)^2$	Devuelve el valor entero asociado
	atomic64_set(v,i)	Asigna i al contador
	atomic64_add(i,v)	Incrementa el contador en i unidades
	atomic64_sub(i,v)	Decrementa i unidades el contador
	atomic64_sub_and_test(i,v)	Decrementa i unidades el contador y devuelve 1 si el resultado es 0
	atomic64_inc(v)	Incrementa el contador
	atomic64_dec(v)	Decrementa el contador
	atomic64_dec_and_test(v)	Decrementa el contador y devuelve 1 si el resultado es 0
	atomic64_inc_and_test(v)	Incrementa el contador y devuelve ${\bf 1}$ si el resultado es ${\bf 0}$
ArTeC.	atomic64_add_negative(v)	Incrementa el contador en i unidades y devuelve 1 si el resultado es negativo
ArTeC.		0 Incrementa el contador en i unidades y devuelve 1 s

 $^{^2} v$ ha de ser un puntero a atomic64_t (valor por referencia).

Recuerda: Cerrojo o mutex



- Mecanismo ideal para resolver el problema de la sección crítica
- Podemos ver un cerrojo como un objeto con dos estados:
 - abierto ó cerrado
- Soporta dos operaciones atómicas
 - Adquirir el cerrojo (lock)

Liberar el cerrojo (unlock)

```
unlock() { estado=abierto; despertar flujo bloqueado; }
```

lo ejecuta el flujo de ejecución que lo cerró con lock() antes (propietario)



Spin locks (I)



- *Spin lock*: Cerrojo de espera activa
 - Sólo puede ser adquirido por un flujo de ejecución del kernel al mismo tiempo
 - Si un flujo de ejecución intenta adquirir un cerrojo que no está libre, esperará a que se libere usando un bucle de espera (busy wait/spin)
 - En Linux representado por el tipo de datos spinlock_t
 - Operaciones declaradas en linux/spinlock.h>
 - El tipo de cerrojo más utilizado en el kernel
 - Única alternativa de cerrojo en aquellos sitios donde no es posible realizar llamadas bloqueantes (p. ej., manejadores de interrupción)



Spin locks (II)



 Garantizan exclusión mutua en regiones de código que acceden a estructuras de datos compartidas entre flujos de ejecución

```
DEFINE_SPINLOCK(sp);
```

```
spin_lock(&sp);
.. Sección crítica 1 ..
spin_unlock(&sp);
```

.. Sección crítica 2 .. spin_unlock(&sp);

Fluio de eiecución 1

Flujo de ejecución 2

spin_lock(&sp);

2 consideraciones:

- spin_lock() inhibe la expropiación de kernel dentro de la sección crítica
- 2 No debemos ejecutar funciones bloqueantes (como vmalloc()) dentro de spin_lock(); ... spin_unlock();



Spin locks (III)



Función/Macro	Descripción
DEFINE_SPINLOCK(name)	Declara e inicializa un <i>spin lock</i> como variable (global) con nombre name
spin_lock_init(sp) ³	Inicializa un spin lock
spin_lock(sp)	Adquiere el spin lock
spin_unlock(sp)	Libera el <i>spin lock</i>
spin_trylock(sp)	Intenta adquirir el <i>spin lock</i> . Si no está libre actualmente, el <i>spin lock</i> no se adquiere y la función devolverá un valor distinto de cero
spin_is_locked(sp)	Devuelve cero si el <i>spin lock</i> está libre actualmente, y un valor distinto de cero en otro caso



 $^{^3 {\}rm sp}$ ha de ser un puntero a spinlock_t (valor por referencia).

Spin locks (IV)



- Si usamos un spin lock para proteger estructura de datos compartida entre manejador de interrupción y cualquier otro flujo de ejecución del kernel → desactivar localmente las interrupciones antes de intentar adquirir el lock
 - No hacer esto puede provocar interbloqueo (deadlock)

Funciones adicionales

	Función/Macro	Descripción
	spin_lock_irq(sp)	Desactiva las interrupciones en la CPU actual y adquiere el <i>spin lock</i>
	spin_unlock_irq(sp)	Reactiva las interrupciones en la CPU actual y libera el <i>spin lock</i>
	spin_lock_irqsave(sp,flags)	Guarda en flags (unsigned long) el estado actual de las interrupciones en la CPU actual, deshabilita localmente las interrupciones y adquiere el <i>spin lock</i>
r	spin_unlock_irqrestore(sp,flags)	Libera el <i>spin lock</i> y restaura el antiguo estado de las interrupciones locales (almacenado en flags)

Código no SMP-safe



```
typedef struct{
   int data:
   struct list head links;
}list item t;
LIST HEAD(shared list);
void list_insert_front(list_item_t* item)
      list_add(&item->links,&shared_list);
void list_print(char* buf)
      char* dst=buf;
      list_item_t* cur_entry=NULL;
       struct list head *cur node;
      list_for_each(cur_node,&shared_list)
          cur_entry = list_entry(cur_node, list_item_t, links);
          dst+=sprintf(dst,"%i\n",cur_entry->data);
```

Spin locks: ejemplo



```
LIST HEAD(shared list);
DEFINE_SPINLOCK(sp);
void list insert front(list item t* item)
      spin lock(&sp);
      list add(&item->links.&shared list):
      spin unlock(&sp):
void list print(char* buf)
      char* dst=buf:
      list item t* cur entry=NULL;
       struct list head *cur node;
      spin lock(&sp);
      list_for_each(cur_node,&shared_list)
          cur entry = list entry(cur node, list item t, links);
          dst+=sprintf(dst, "%i\n", cur_entry->data);
       spin unlock(&sp);
```

Reader-Writer Spin Locks



- Tipo especial de spin lock que permite incrementar el grado de concurrencia en una sección crítica
 - Paradigma lectores-escritores:
 - Permite lecturas simultáneas pero sólo una escritura
 - Exclusión mutua entre lectura-escritura y escritura-escritura
 - En Linux representado mediante el tipo de datos rwlock_t
 - Los lectores tienen prioridad → inanición de los escritores

Funciones/Macros (<linux/rwlock.h>)

- Inicialización: DEFINE_RWLOCK(), rwlock_init()
- Lector: read_lock(), read_unlock(), read_trylock(), read_lock_irq(), read_unlock_irq(), read_lock_irqsave(), read_unlock_irqrestore().
- Escritor: write_lock(), write_unlock(), write_trylock(),
 write_lock_irq(), write_unlock_irq(), write_lock_irqsave(),
 write_unlock_irqrestore().



Reader-Writer Spin Locks: ejemplo



```
LIST HEAD(shared list);
DEFINE RWLOCK(rwl):
void list insert front(list item t* item)
      write lock(&rwl);
      list add(&item->links.&shared list):
       write unlock(&rwl):
void list print(char* buf)
      char* dst=buf:
      list item t* cur entry=NULL;
       struct list head *cur node;
      read lock(&rwl);
      list_for_each(cur_node,&shared_list)
          cur entry = list entry(cur node, list item t, links);
          dst+=sprintf(dst, "%i\n", cur_entry->data);
      read unlock(&rwl);
```

Semáforos (I)



Dos tipos de semáforos en Linux:

- 1 Semáforos POSIX: Utilizados por procesos de usuario
- 2 Semáforos del kernel: Usados por flujos de ejecución del kernel

Semáforos del kernel

- Semáforos generales
 - Contador asociado
 - Dos operaciones básicas:
 - down(): operación wait() del semáforo
 - up(): operación signal() del semáforo
- Espera bloqueante
 - Si un flujo de ejecución ejecuta operación wait() cuando contador del semáforo ≤ 0, el flujo se bloquea (sleep)
 - Solo puede utilizarse en Contexto de Proceso en modo kernel



Semáforos (II)



```
wait(s){
 s = s - 1;
 if(s < 0){}
    <Bloquear al proceso>
signal(s){
 s = s + 1:
  if (hay_procesos_bloqueados){
    <Desbloquear a un proceso</pre>
         bloqueado por wait>
```

Significado de s

- $s \ge 0 \to s$ es el número de veces que se puede invocar a wait() sin que ningún proceso invocador se bloquee
- $s \le 0 \to |s|$ es el número de procesos bloqueados en el semáforo

Semáforos (III)



Semáforos del kernel (Cont.)

- En Linux representado mediante struct semaphore
 - Operaciones declaradas en linux/semaphore.h>
- Muy utilizados en implementación de llamadas al sistema (handler functions)
- Uso típico de los semáforos:
 - Garantizar exclusión mutua
 - Inicializar contador del semáforo a 1
 - Encerrar las secciones críticas entre down() y up()
 - 2 Implementar sincronización basada en condiciones sobre números enteros
 - 3 Cola de espera
 - Inicializar contador del semáforo a 0
 - Bloqueo incondicional de proceso con down()
 - Despertar a proceso bloqueado con up()



Semáforos (IV)



Función/Macro	Descripción
DEFINE_SEMAPHORE(name)	Declara e inicializa a 1 un semaforo (struct semaphore) como variable (global) con nombre name
sema_init(sem,n)4	Inicializa un semáforo y establece su contador a n
down(sem)	Operación wait() del semáforo. En el caso de que se produzca un bloqueo, el kernel pone el proceso en estado TASK_UNINTERRUPTIBLE (el proceso NO responde a señales). No es recomendable utilizar esta operación
up(sem)	Operación signal() del semáforo.
<pre>down_interruptible(sem)</pre>	Variante de la operación <i>wait()</i> del semáforo. En el caso de que se produzca un bloqueo, el kernel pone el proceso en estado TASK_INTERRUPTIBLE (el proceso SÍ responde a señales.). Si el proceso es despertado por una señal antes de decrementar el semáforo, la función devuelve un valor distinto de 0 y 0 en caso contrario.

⁴sem ha de ser un puntero a struct semaphore (valor por referencia).

Semáforos (V)



Función/Macro	Descripción			
down_killable(sem)	Variante de la operación wait() del semáforo. En el caso de que se produzca un bloqueo, el proceso se pondrá en estado especial TASK_KILLABLE (el proceso solo responde a señales que fuerzan su terminación). Sí el proceso es despertado por una señal antes de decrementar el semáforo, la función devuelve un valor distinto de 0 y 0 en caso contrario.			
down_trylock(sem)	Intenta decrementar atómicamente el valor del semáforo y devuelve 1 si lo consigue. Si no es posible (contador \leq 0) devuelve 0.			



Semáforos: Ejemplo 1



```
LIST_HEAD(shared_list);
DEFINE_SEMAPHORE(sem);
int list_insert_front(list_item_t* item)
{
    if (down_interruptible(&sem)){
        return -EINTR;
    }
    list_add(&item->links,&shared_list);
    up(&sem);
    return 0;
}
```



Semáforos: Ejemplo 1 (cont.)



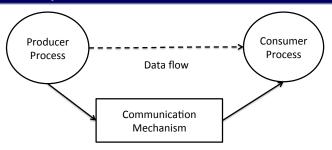
```
int list_print(char* buf)
      char* dst=buf:
      list_item_t* cur_entry=NULL;
      struct list head *cur node;
      if (down_interruptible(&sem)){
          return -EINTR;
      list_for_each(cur_node,&shared_list)
          cur_entry = list_entry(cur_node, list_item_t, links);
          dst+=sprintf(dst,"%i\n",cur_entry->data);
      up(&sem);
      return 0;
```



Semáforos: Ejemplo 2



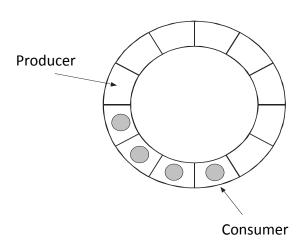
Problema del productor-consumidor



- El consumidor se bloquea hasta que haya datos que leer
- El productor se bloqueará al intentar enviar datos a través del mecanismo de sincronización cuando esté lleno

Productor-consumidor: buffer circular







Semáforos: Ejemplo 2



Productor/consumidor

- Implementado mediante un módulo del kernel que gestiona buffer circular de enteros
- Módulo exporta entrada /proc/prodcons
 - Productor: \$ echo 7 > /proc/prodcons
 - Inserta elemento al final del buffer
 - Se bloquea si no hay espacio en el buffer
 - Consumidor: \$ cat /proc/prodcons
 - Consume el primer elemento del buffer y lo devuelve al programa de usuario
 - Si no hay elementos que consumir se bloquea



Semáforos: Ejemplo 2 (Inic.)



```
#define MAX ITEMS CBUF 5
  static struct proc_dir_entry *proc_entry;
  static cbuffer t* cbuf;
  struct semaphore elementos.huecos:
   int init prodcons module( void )
    cbuf = create_cbuffer_t(MAX_ITEMS_CBUF);
    sema init(&elementos,0);
    sema init(&huecos.MAX ITEMS CBUF):
    if (!cbuf)
      return -ENOMEM:
    proc_entry = proc_create("prodcons",0666, NULL, &proc_entry_fops);
    if (proc_entry == NULL) {
        destroy_cbuffer_t(cbuf);
       return -ENOMEM;
    return 0;
Ar}
```

Semáforos: Ejemplo 2 (Productor)



```
ssize_t prodcons_write(struct file *filp, const char __user *buf,
                           size t len. loff t *off)
 char kbuf[MAX_CHARS_KBUF+1];
 int val=0:
 .. copy_from_user() + Convertir char* a entero y almacenarlo en val ..
 if (down interruptible(&huecos))
   return -EINTR:
 insert cbuffer t(cbuf.val):
 up(&elementos);
 return len;
```



Semáforos: Ejemplo 2 (Consumidor)



```
ssize_t prodcons_read(struct file *filp, char __user *buf, size_t len, loff_t *off)
 char kbuf[MAX CHARS KBUF+1];
 int nr_bytes=0;
 int val:
 if (offset>0)
   return 0;
 if (down interruptible(&elementos))
   return -EINTR;
 val=remove_cbuffer_t(cbuf);
 up(&huecos):
 nr_bytes=sprintf(kbuf, "%i\n", val);
 ... copy to user() ...
 return nr_bytes;
```

Semáforos: Ejemplo 2



- El código anterior no garantiza exclusión mutua en el acceso al buffer
 - Si la estructura de datos no soportara acceso concurrente → potencial condición de carrera
 - Solución: Añadir semáforo adicional para garantizar exclusión mutua

Productor

```
if (down_interruptible(&huecos))
    return -EINTR;

/* Entrar a la SC */
if (down_interruptible(&mtx)) {
        up(&huecos);
        return -EINTR;
}
    insert_cbuffer_t(cbuf,val);
/* Salir de la SC */
up(&mtx);
up(&elementos);
```

Consumidor

```
if (down_interruptible(&elementos))
    return -EINTR;

/* Entrar a la SC */
if (down_interruptible(&mtx)){
    up(&elementos);
    return -EINTR;
}

val=remove_cbuffer_t(cbuf);
/* Salir de la SC */
up(&mtx);
up(&huecos);
```

ArTe

Referencias



- Linux Kernel Development
 - Cap. 9 "An Introduction to Kernel Synchronization"
 - Cap. 10 "Kernel Synchronization Methods"
- Understanding the Linux Kernel
 - Cap. 11 "Kernel Synchronization"



Licencia



Arquitectura Interna de Linux - Sincronización en el kernel Linux Versión 0.3

©J.C. Sáez

This work is licensed under the Creative Commons Attribution-Share Alike 3.0 Spain License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/3.0/es/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105,USA.

Esta obra está bajo una licencia Reconocimiento-Compartir Bajo La Misma Licencia 3.0 España de Creative Commons. Para ver una copia de esta licencia, visite http://creativecommons.org/licenses/by-sa/3.0/es/ o envíe una carta a Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

Este documento (o uno muy similar) está disponible en https://cv4.ucm.es/moodle/course/view.php?id=70009



