Subject: [RFC PATCH v2] posix timers: allocate timer id per task
Posted by Stanislav Kinsbursky on Wed, 17 Oct 2012 13:17:41 GMT
View Forum Message <> Reply to Message

v2:
1) Hash table become RCU-friendly. Hash table search now done under RCU lock protection.
I've tested scalability on KVM with 4 CPU. The testing environment was build of 10 processes, each had 512 posix timers running (SIGSEV_NONE) and was calling timer_gettime() in loop. With all this stuff being running, I was measuring time of calling of syscall timer_gettime() 10000 times.

Without this patch: ~7ms
With this patch   : ~7ms

This patch is required CRIU project (www.criu.org).
To migrate processes with posix timers we have to make sure, that we can restore posix timer with proper id.
Currently, this is not true, because timer ids are allocated globally.
So, this is precursor patch and it's purpose is make posix timer id to be allocated per task.

Patch replaces global idr with global hash table for posix timers and makes timer ids unique not globally, but per task. Next free timer id is type of integer and stored on signal struct (posix_timer_id). If free timer id reaches negative value on timer creation, it will be dropped to zero and -EAGAIN will be returned to user.
Hash table is size of page (4KB).
Key is constructed as follows:
key = hash_ptr(current->signal) ^ hash_32(posix_timer_id);

Signed-off-by: Stanislav Kinsbursky <skinsbursky@parallels.com>

```
---
 include/linux/posix-timers.h |    1 
 include/linux/sched.h        |    4 +-
 kernel/posix-timers.c        |  112 +++++++++++++++++++++++++++--------------
 3 files changed, 78 insertions(+), 39 deletions(-)

diff --git a/include/linux/posix-timers.h b/include/linux/posix-timers.h
index 042058f..60bac69 100644
--- a/include/linux/posix-timers.h
+++ b/include/linux/posix-timers.h
@@ -55,6 +55,7 @@ struct cpu_timer_list {
 /* POSIX.1b interval timer structure. */
 struct k_itimer {
  struct list_head list;  /* free/ allocate list */
+ struct hlist_node t_hash;
  spinlock_t it_lock;
```

```
  clockid_t it_clock;  /* which timer type */
  timer_t it_id;   /* timer id */
diff --git a/include/linux/sched.h b/include/linux/sched.h
index 0dd42a0..dce1651 100644
--- a/include/linux/sched.h
+++ b/include/linux/sched.h
@@ -51,6 +51,7 @@ struct sched_param {
 #include <linux/cred.h>
 #include <linux/llist.h>
 #include <linux/uidgid.h>
+#include <linux/idr.h>

 #include <asm/processor.h>

@@ -536,7 +537,8 @@ struct signal_struct {
  unsigned int  has_child_subreaper:1;

  /* POSIX.1b Interval Timers */
- struct list_head posix_timers;
+ int   posix_timer_id;
+ struct list_head posix_timers;

  /* ITIMER_REAL timer for the process */
  struct hrtimer real_timer;
diff --git a/kernel/posix-timers.c b/kernel/posix-timers.c
index 69185ae..f43e2e6 100644
--- a/kernel/posix-timers.c
+++ b/kernel/posix-timers.c
@@ -47,31 +47,28 @@
 #include <linux/wait.h>
 #include <linux/workqueue.h>
 #include <linux/export.h>
+#include <linux/hash.h>

 /*
- * Management arrays for POSIX timers.  Timers are kept in slab memory
- * Timer ids are allocated by an external routine that keeps track of the
- * id and the timer.  The external interface is:
- *
- * void *idr_find(struct idr *idp, int id);           to find timer_id <id>
- * int idr_get_new(struct idr *idp, void *ptr);       to get a new id and
- *                                    related it to <ptr>
- * void idr_remove(struct idr *idp, int id);          to release <id>
- * void idr_init(struct idr *idp);                    to initialize <idp>
- *                                    which we supply.
- * The idr_get_new *may* call slab for more memory so it must not be
- * called under a spin lock.  Likewise idr_remore may release memory
- * (but it may be ok to do this under a lock...).
```

```
- * idr_find is just a memory look up and is quite fast.  A -1 return
- * indicates that the requested id does not exist.
+ * Management arrays for POSIX timers. Timers are now kept in static PAGE-size
+ * hash table.
+ * Timer ids are allocated by local routine, which selects proper hash head by
+ * key, constructed from current->signal address and per signal struct counter.
+ * This keeps timer ids unique per task, but now they can intersect between
+ * tasks.
  */

 /*
  * Lets keep our timers in a slab cache :-)
  */
 static struct kmem_cache *posix_timers_cache;
-static struct idr posix_timers_id;
-static DEFINE_SPINLOCK(idr_lock);
+
+#define POSIX_TIMERS_HASH_BITS  9
+#define POSIX_TIMERS_HASH_SIZE  (1 << POSIX_TIMERS_HASH_BITS)
+
+/* Hash table is size of PAGE currently */
+static struct hlist_head posix_timers_hashtable[POSIX_TIMERS_HASH_SIZE];
+static DEFINE_SPINLOCK(hash_lock);

 /*
  * we assume that the new SIGEV_THREAD_ID shares no bits with the other
@@ -152,6 +149,56 @@ static struct k_itimer *__lock_timer(timer_t timer_id, unsigned long
*flags);
  __timr;           \
 })

+static int hash(struct signal_struct *sig, unsigned int nr)
+{
+ int hash = hash_ptr(sig, POSIX_TIMERS_HASH_BITS);
+ return hash ^ hash_32(nr, POSIX_TIMERS_HASH_BITS);
+}
+
+static struct k_itimer *__posix_timers_find(struct hlist_head *head, struct signal_struct *sig,
timer_t id)
+{
+ struct hlist_node *node;
+ struct k_itimer *timer;
+
+ hlist_for_each_entry_rcu(timer, node, head, t_hash) {
+  if ((timer->it_signal == sig) && (timer->it_id == id))
+   return timer;
+ }
+ return NULL;
```

```
+}
+
+static struct k_itimer *posix_timer_by_id(timer_t id)
+{
+ struct signal_struct *sig = current->signal;
+ struct hlist_head *head = &posix_timers_hashtable[hash(sig, id)];
+
+ return __posix_timers_find(head, sig, id);
+}
+
+static int posix_timer_add(struct k_itimer *timer)
+{
+ struct signal_struct *sig = current->signal;
+ int next_free_id = sig->posix_timer_id;
+ struct hlist_head *head;
+ int ret = -ENOENT;
+
+ do {
+  spin_lock(&hash_lock);
+  head = &posix_timers_hashtable[hash(sig, sig->posix_timer_id)];
+  if (__posix_timers_find(head, sig, sig->posix_timer_id) == NULL) {
+   hlist_add_head_rcu(&timer->t_hash, head);
+   ret = sig->posix_timer_id++;
+  } else {
+   if (++sig->posix_timer_id < 0)
+    sig->posix_timer_id = 0;
+   if (sig->posix_timer_id == next_free_id)
+    ret = -EAGAIN;
+  }
+  spin_unlock(&hash_lock);
+ } while (ret == -ENOENT);
+ return ret;
+}
+
 static inline void unlock_timer(struct k_itimer *timr, unsigned long flags)
 {
  spin_unlock_irqrestore(&timr->it_lock, flags);
@@ -271,6 +318,7 @@ static __init int init_posix_timers(void)
   .timer_get = common_timer_get,
   .timer_del = common_timer_del,
  };
+ int i;

  posix_timers_register_clock(CLOCK_REALTIME, &clock_realtime);
  posix_timers_register_clock(CLOCK_MONOTONIC, &clock_monotonic);
@@ -282,7 +330,8 @@ static __init int init_posix_timers(void)
  posix_timers_cache = kmem_cache_create("posix_timers_cache",
      sizeof (struct k_itimer), 0, SLAB_PANIC,
```

```
      NULL);
- idr_init(&posix_timers_id);
+ for (i = 0; i < POSIX_TIMERS_HASH_SIZE; i++)
+  INIT_HLIST_HEAD(&posix_timers_hashtable[i]);
  return 0;
 }

@@ -504,9 +553,9 @@ static void release_posix_timer(struct k_itimer *tmr, int it_id_set)
 {
  if (it_id_set) {
   unsigned long flags;
-  spin_lock_irqsave(&idr_lock, flags);
-  idr_remove(&posix_timers_id, tmr->it_id);
-  spin_unlock_irqrestore(&idr_lock, flags);
+  spin_lock_irqsave(&hash_lock, flags);
+  hlist_del_rcu(&tmr->t_hash);
+  spin_unlock_irqrestore(&hash_lock, flags);
  }
  put_pid(tmr->it_pid);
  sigqueue_free(tmr->sigq);
@@ -552,22 +601,9 @@ SYSCALL_DEFINE3(timer_create, const clockid_t, which_clock,
   return -EAGAIN;

  spin_lock_init(&new_timer->it_lock);
- retry:
- if (unlikely(!idr_pre_get(&posix_timers_id, GFP_KERNEL))) {
-  error = -EAGAIN;
-  goto out;
- }
- spin_lock_irq(&idr_lock);
- error = idr_get_new(&posix_timers_id, new_timer, &new_timer_id);
- spin_unlock_irq(&idr_lock);
- if (error) {
-  if (error == -EAGAIN)
-   goto retry;
-  /*
-   * Weird looking, but we return EAGAIN if the IDR is
-   * full (proper POSIX return value for this)
-   */
-  error = -EAGAIN;
+ new_timer_id = posix_timer_add(new_timer);
+ if (new_timer_id < 0) {
+  error = new_timer_id;
   goto out;
  }

@@ -640,7 +676,7 @@ static struct k_itimer *__lock_timer(timer_t timer_id, unsigned long *flags)
  struct k_itimer *timr;
```

```
  rcu_read_lock();
- timr = idr_find(&posix_timers_id, (int)timer_id);
+ timr = posix_timer_by_id(timer_id);
  if (timr) {
   spin_lock_irqsave(&timr->it_lock, *flags);
   if (timr->it_signal == current->signal) {
```

Subject: Re: [RFC PATCH v2] posix timers: allocate timer id per task
Posted by Eric Dumazet on Wed, 17 Oct 2012 13:44:26 GMT
View Forum Message <> Reply to Message

On Wed, 2012-10-17 at 17:18 +0400, Stanislav Kinsbursky wrote:

> +static int hash(struct signal_struct *sig, unsigned int nr)
> +{
> + int hash = hash_ptr(sig, POSIX_TIMERS_HASH_BITS);
> + return hash ^ hash_32(nr, POSIX_TIMERS_HASH_BITS);
> +}
> +

This is quite expensive on 64 bit arches. I suggest instead :

return hash_32(hash32_ptr(sig) ^ nr, POSIX_TIMERS_HASH_BITS);

Subject: Re: [RFC PATCH v2] posix timers: allocate timer id per task
Posted by Eric Dumazet on Wed, 17 Oct 2012 13:57:26 GMT
View Forum Message <> Reply to Message

On Wed, 2012-10-17 at 17:18 +0400, Stanislav Kinsbursky wrote:

> +static int posix_timer_add(struct k_itimer *timer)
> +{
> + struct signal_struct *sig = current->signal;
> + int next_free_id = sig->posix_timer_id;
> + struct hlist_head *head;
> + int ret = -ENOENT;
> +
> + do {
> + spin_lock(&hash_lock);
> + head = &posix_timers_hashtable[hash(sig, sig->posix_timer_id)];
> + if (__posix_timers_find(head, sig, sig->posix_timer_id) == NULL) {
> +   hlist_add_head_rcu(&timer->t_hash, head);

Hmm...

```
> +   ret = sig->posix_timer_id++;
> +  } else {
> +   if (++sig->posix_timer_id < 0)
> +    sig->posix_timer_id = 0;
> +   if (sig->posix_timer_id == next_free_id)
> +    ret = -EAGAIN;
> +  }
> +  spin_unlock(&hash_lock);
> + } while (ret == -ENOENT);
> + return ret;
> +}
> +
```

You probably need to add a rcu_assign_pointer() or smp_wmb() before the :

new_timer->it_signal = current->signal;

in the following block :

```
        spin_lock_irq(&current->sighand->siglock);
        new_timer->it_signal = current->signal;
        list_add(&new_timer->list, &current->signal->posix_timers);
        spin_unlock_irq(&current->sighand->siglock);
```

Or else another thread can read outdated informations...

```
        spin_lock_irq(&current->sighand->siglock);
        list_add(&new_timer->list, &current->signal->posix_timers);
        spin_unlock_irq(&current->sighand->siglock);
 smp_wmb();
        new_timer->it_signal = current->signal;
```

Subject: Re: [RFC PATCH v2] posix timers: allocate timer id per task
Posted by Stanislav Kinsbursky on Wed, 17 Oct 2012 14:19:56 GMT
View Forum Message <> Reply to Message

> On Wed, 2012-10-17 at 17:18 +0400, Stanislav Kinsbursky wrote:
>
>> +static int hash(struct signal_struct *sig, unsigned int nr)
>> +{
>> + int hash = hash_ptr(sig, POSIX_TIMERS_HASH_BITS);
>> + return hash ^ hash_32(nr, POSIX_TIMERS_HASH_BITS);
>> +}

>> +
>
> This is quite expensive on 64 bit arches. I suggest instead :
>
> return hash_32(hash32_ptr(sig) ^ nr, POSIX_TIMERS_HASH_BITS);
>
>

I'm not familiar with hashing costs, so sure, I'll update.

--
Best regards,
Stanislav Kinsbursky

Subject: Re: [RFC PATCH v2] posix timers: allocate timer id per task
Posted by Stanislav Kinsbursky on Wed, 17 Oct 2012 14:26:13 GMT
View Forum Message <> Reply to Message

> On Wed, 2012-10-17 at 17:18 +0400, Stanislav Kinsbursky wrote:
>
>> +static int posix_timer_add(struct k_itimer *timer)
>> +{
>> + struct signal_struct *sig = current->signal;
>> + int next_free_id = sig->posix_timer_id;
>> + struct hlist_head *head;
>> + int ret = -ENOENT;
>> +
>> + do {
>> + spin_lock(&hash_lock);
>> + head = &posix_timers_hashtable[hash(sig, sig->posix_timer_id)];
>> + if (__posix_timers_find(head, sig, sig->posix_timer_id) == NULL) {
>> + hlist_add_head_rcu(&timer->t_hash, head);
>
> Hmm...
>
>> + ret = sig->posix_timer_id++;
>> + } else {
>> + if (++sig->posix_timer_id < 0)
>> + sig->posix_timer_id = 0;
>> + if (sig->posix_timer_id == next_free_id)
>> + ret = -EAGAIN;
>> + }
>> + spin_unlock(&hash_lock);
>> + } while (ret == -ENOENT);
>> + return ret;
>> +}

>> +
>
> You probably need to add a rcu_assign_pointer() or smp_wmb() before
> the :
>
> new_timer->it_signal = current->signal;
>
> in the following block :
>
> spin_lock_irq(&current->sighand->siglock);
> new_timer->it_signal = current->signal;
> list_add(&new_timer->list, &current->signal->posix_timers);
> spin_unlock_irq(&current->sighand->siglock);
>
> Or else another thread can read outdated informations...
>
>
> spin_lock_irq(&current->sighand->siglock);
> list_add(&new_timer->list, &current->signal->posix_timers);
> spin_unlock_irq(&current->sighand->siglock);
> smp_wmb();
> new_timer->it_signal = current->signal;
>
>

Sorry, but I don't understand what races you are talking about.
Could you specify, please?

--
Best regards,
Stanislav Kinsbursky