Subject: Re: [PATCH v6 02/10] ipc: "use key as id" functionality for resource get system ca
Posted by ebiederm on Tue, 16 Oct 2012 09:03:29 GMT
View Forum Message <> Reply to Message

Stanislav Kinsbursky <skinsbursky@parallels.com> writes:

>> Hmm. Come to think of it I don't see why you need to set the id at all.
>> We are using an idr allocator which effectively offers the semantics
>> that the lowest available id will be allocated. The same semantics we
>> have for file descriptors.
>>
>> So it should be possible at least for the first pass at
>> checkpoint/restart to implement the restoration of sysv ipc without
>> IPC_PRESET at all.
>>
>> So IPC_PRESET should just be an optimization, not a necessary feature.
>>
>
> CRIU was designed to suspend/restore not only containers with it's own IPC
> namespace, but also for single process and process tree.
> So we have to restore IPC objects with proper id and key.

I was not suggesting restoring IPC objects without the proper id and
key.

I was pointing that since the algorithm for id assignment is known, that
by simply allocating objects with care you can guarantee that you
allocate objects with the proper id and key.

As for the case when you don't have your own ipc namespace, in general
that case will fail. Especially with the current id assignment
algorithm chances that your id is already taken are very high. So since
in general and most of the time it will fail to restore an ipc namespace
that is not shared outside of the container there needs to be a
compelling reason to support to even consider it.

I am not comfortable with the key<->id association changing on active
objects. That gives rise to new and scary races in userspace, and a
much more complex set of interactions to analyze to see if the both
userspace and the kernel code is correct.

I am not particularly comfortable adding a new allocation pathway for
sysv ipc objects when that new pathway appears completely unnecessary.

Stanislav please concentrate on something that works for the easy case

of an ipc namespace per container today. Once there is a working
checkpoint/restart and people can use it and see what the limitations
it becomes much easier to see if extra optimizations like your
IPC_PRESET code are worth it.

Eric