Subject: Re: [Patch 1/3] Miscellaneous container fixes
Posted by Paul Menage on Tue, 05 Dec 2006 12:04:56 GMT
View Forum Message <> Reply to Message

On 12/1/06, Paul Jackson <pj@sgi.com> wrote:
> Read the comment in kernel/cpuset.c for the routine cpuset_destroy().
> It explains that update_flag() is called where it is (turning off
> the cpu_exclusive flag, if it was set), to avoid the calling sequence:
>
>   cpuset_destroy->update_flag->update_cpu_domains->lock_cpu_hotplug
>
> while holding the callback_mutex, as that could ABBA deadlock with the
> CPU hotplug code.

This particular race is gone in the -mm2 kernel since cpus_exclusive
no longer drives sched_domains - can we assume that this will be
reaching mainline some time soon?

>
> But with this container based rewrite of cpusets, it now seems that
> cpuset_destroy -is- called holding the callback_mutex (though I don't
> see any mention of that in the cpuset_destroy comment ;), so it would

And in fact I explicitly documented it as only holding manage_mutex,
not callback_mutex in Documentation/containers.txt. I think maybe this
slipped in during the multi-hierarchy rewrite. :-(

Looking at the various *_destroy() functions in the container
subsystems in my patch set, I think that it should be OK to call the
destructors prior to taking callback_mutex for the unlinking of the
container from its parents.

>
> I also notice that the comment for container_lock() in the file
> kernel/container.c only mentions its use in the oom code. That is
> no longer the only, or even primary, user of this lock routine.
> The kernel/cpuset.c code uses it frequently (without comment ;),
> and I wouldn't be surprised to see other future controllers calling
> container_lock() as well.

As was pointed out by Chandra Seetharaman, it would be nice if we
could avoid having all the container subsystems relying on
callback_mutex for their locking needs - particularly since that's
likely to be acquired at performance-sensitive times.

The cpu_acct and beancounters subsystems that I included in my patch
set both use their own per-container locks for synchronization, so
it's not completely necessary to use the central locks. There's

probably a happy medium between "one big lock" and "way too many small locks".

Paul