Subject: Re: [PATCH 0/5] per-cpu/cpuacct cgroup scheduler statistics
Posted by Glauber Costa on Thu, 16 Feb 2012 10:06:11 GMT
View Forum Message <> Reply to Message

On 02/15/2012 02:31 AM, Serge Hallyn wrote:
> Quoting Glauber Costa (glommer@parallels.com):
>> On 02/02/2012 06:19 PM, Glauber Costa wrote:
>>> Hi,
>>>
>>> Here is my new attempt to get a per-container version of some
>>> /proc data such as /proc/stat and /proc/uptime.
>>>
>>> In this series I solved the visibility problem, which is,
>>> the problem of how and when to show /proc/stat data per-cgroup,
>>> by declaring it not a problem.
>>>
>>> This can probably be done in userspace with other aids, like mounting
>>> a fuse overlay that simulates /proc from outside a container, to a
>>> container location.
>>>
>>> Here, we should have most of the data needed to do that. They are drawn
>> >from both the cpu cgroup, and cpuacct. Each cgroup exports the data it
>>> knows better, and I am not really worried here about bindings between them.
>>>
>>> In this first version, I am using clock_t units, being quite proc-centric.
>>> It made my testing easier, but I am happy to show any units you guys would
>>> prefer.
>>>
>>> Besides that, it still has some other minor issues to be sorted out.
>>> But I verified the general direction to be working, and would like to know
>>> what you think.
>>>
>>
>> Hi,
>>
>> Did someone had any chance to take a look at this already?
>>
>> Thanks
>
> Hi,
>
> By declaring proc visibility not a problem and sticking to io stats,
> you sort of left me where I don't know what I'm talking about :)

Heh. Do you at least agree with the approach of just dumping the
information in cgroup files, so we can join them later? (be it via
userspace or in a follow up kernel patch if the need really arises from
real workloads?)

Do you have any comments on any preferred format?

> So
> let me just say, on patch 2, "store number of iowait events in a task_group",
> my initial reaction is "boy that's a lot more work. What is the performance
> impact?"
Yeah, The first thing I need to do if I'm carrying this forward is to
measure that.

>
> It'd be possible to move the extra processing out of the hot-path by
> only changing the # for the deepest cgroup, and pulling it into
> ancestor cgroups only when someone is viewing the stats or the child
> cgroup goes away.

In principle, should be doable. We discussed this briefly (me and Peter)
once, and the problem is that it of course imposes a hit on the readers.
If you're reading often enough (can be the case for things polling
/proc/stat), this can be a problem.

But if we are really doing this, we can very well do it for all stats,
not only iowait...

> But if you have #s showing statistically negligable
> performance impact anyway then that wouldn't be worth it.
>
Need to work on that.