Subject: Re: [ckrm-tech] [PATCH 1/9] Containers (V9): Basic container framework
Posted by Paul Jackson on Tue, 01 May 2007 18:40:30 GMT
View Forum Message <> Reply to Message

[[I have bcc'd one or more batch scheduler experts on this post.
They will know who they are, and should be aware that they are
not listed in the public cc list of this message. - pj]]

Balbir Singh, responding to Paul Menage's Container patch set on lkml, wrote:
>
> > +*** notify_on_release is disabled in the current patch set. It may be
> > +*** reactivated in a future patch in a less-intrusive manner
> > +
>
> Won't this break user space tools for cpusets?

Yes - disabling notify_on_release would definitely break some important uses of cpusets. This feature must be reactivated somehow before I'll sign up for putting this patch set in the main line.

Actually, after I posted a few days ago in another lkml post:
http://lkml.org/lkml/2007/4/29/66

that just the simplest cpuset command:
mount -t cpuset cpuset /dev/cpuset
mkdir /dev/cpuset/foo
echo 0 > /dev/cpuset/foo/mems

caused an immediate kernel deadlock (Srivatsa has proposed a fix), it is pretty clear that this container patch set is not getting the cpuset testing it will need for acceptance. That's partly my fault.

The batch scheduler folks, such as the variants of PBS, LSF and SGE are major user of cpusets on NUMA hardware.

This container based replacement for cpusets isn't ready for the main line until at least one of those schedulers can run through one of their test suites. I hesitate to even acknowledge this, as I might be the only person in a position to make this happen, and my time available to contribute to this patch set has been less than I would like.

But if it looks like we have all the pieces in place to base cpusets on containers, with no known regressions in cpuset capability, then we must find a way to ensure that one of these batch schedulers, using cpusets on a NUMA box, still works.

--

I won't rest till it's the best ...
Programmer, Linux Scalability
Paul Jackson <pj@sgi.com> 1.925.600.0401