Subject: Re: [patch00/05]: Containers(V2)- Introduction
Posted by Paul Jackson on Wed, 20 Sep 2006 23:18:31 GMT
View Forum Message <> Reply to Message

Chistroph, responding to Alan:

> > I'm also not clear how you handle shared pages correctly under the fake
> > node system, can you perhaps explain that further how this works for say
> > a single apache/php/glibc shared page set across 5000 containers each a
> > web site.
>
> Cpusets can share nodes. I am not sure what the problem would be? Paul may
> be able to give you more details.

Cpusets share pre-assigned nodes, but not anonymous proportions of the total system memory.

So sharing an apache/php/glibc page set across 5000 containers using cpusets would be awkward. Unless I'm missing something, you'd have to prepage in that page set, from some task allowed that many pages in its own cpuset, then you'd run each of the 5000 web servers in smaller cpusets that allowed space for the remainder of whatever that web server was provisioned, not counting the shared pages. The shared pages wouldn't count, because cpusets doesn't ding you for using a page that is already in memory -- it just keeps you from allocating fresh pages on certain nodes. When it came time to do rolling upgrades to new versions of the software, and add a marketing driven list of 57 additional applications that the customers could use to build their website, this could become an official nightmare.

Overbooking (selling say 10 Mbs of memory for each server, even though there is less than 5000 * 10 Mb total RAM in the system) would also be awkward. One could simulate with overlapping sets of fake numa nodes, as I described in an earlier post today (the one that gave each task some four of the five 20 MB fake cpusets.) But there would still be false resource conflicts, and the (ab)use of the cpuset apparatus for this seems unintuitive, in my opinion.

I imagine that a web site supporting 5000 web servers would be very interested in overbooking working well. I'm sure the $7.99/month cheap as dirt virtual web servers of which I am a customer overbook.

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