Subject: Re: [ckrm-tech] [patch00/05]: Containers(V2)- Introduction
Posted by Chandra Seetharaman on Thu, 21 Sep 2006 20:06:26 GMT
View Forum Message <> Reply to Message

On Wed, 2006-09-20 at 18:52 -0700, Paul Menage wrote:
> On 9/20/06, Chandra Seetharaman <sekharan@us.ibm.com> wrote:
> >
> > Interesting. So you could set up the fake node with "guarantee" and let
> > it grow till "limit" ?
>
> Sure - that works great. (Theoretically you could do this all in
> userspace - start by assigning "guarantee" nodes to a
> container/cpuset and when it gets close to its memory limit assign
> more nodes to it. But in practice userspace can't keep up with rapid
> memory allocators.
>
I agree, especially when one of your main object is resource
utilization. Think about the magnitude of this when you have to deal
with 100s of containers.

> >
> > BTW, can you do these with fake nodes:
> > - dynamic creation
> > - dynamic removal
> > - dynamic change of size
>
> The current fake numa support requires you to choose your node layout
> at boot time - I've been working with 64 fake nodes of 128M each,
> which gives a reasonable granularity for dividing a machine between
> multiple different sized jobs.

It still will not satisfy what OpenVZ/Container folks are looking for:
100s of containers.

>
> >
> > Also, How could we account when a process moves from one node to
> > another ?
>
> If you want to do that (the systems I'm working on don't really) you
> could probably do it with the migrate_pages() syscall. It might not be
> that efficient though.

Totally agree, that will be very costly.
>
> Paul
--

Chandra Seetharaman | Be careful what you choose....
- sekharan@us.ibm.com | .....you may get it.

Subject: Re: [ckrm-tech] [patch00/05]: Containers(V2)- Introduction
Posted by Paul Menage on Thu, 21 Sep 2006 20:10:22 GMT
View Forum Message <> Reply to Message

On 9/21/06, Chandra Seetharaman <sekharan@us.ibm.com> wrote:
> > The current fake numa support requires you to choose your node layout
> > at boot time - I've been working with 64 fake nodes of 128M each,
> > which gives a reasonable granularity for dividing a machine between
> > multiple different sized jobs.
>
> It still will not satisfy what OpenVZ/Container folks are looking for:
> 100s of containers.

Right - so fake-numa is not the right solution for everyone, and I never suggested that it is. (Having said that, there are discussions underway to make the zone-based approach more practical - if you could have dynamically-resizable nodes, this would be more applicable to openvz).

But, there's no reason that the OpenVZ resource control mechanisms couldn't be hooked into a generic process container mechanism along with cpusets and RG.

Paul