Subject: Re: [ckrm-tech] [patch00/05]: Containers(V2)- Introduction
Posted by Chandra Seetharaman on Thu, 21 Sep 2006 01:45:20 GMT
View Forum Message <> Reply to Message

On Wed, 2006-09-20 at 17:42 -0700, Paul Menage wrote:
> On 9/20/06, Paul Jackson <pj@sgi.com> wrote:
> > Chandra wrote:
> > > AFAICS, That doesn't help me in over committing resources.
> >
> > I agree - I don't think cpusets plus fake numa ... handles over commit.
> > You might could hack up a cheap substitute, but it wouldn't do the job.
>
> I have some patches locally that basically let you give out a small
> set of nodes initially to a cpuset, and if memory pressure in
> try_to_free_pages() passes a specified threshold, automatically
> allocate one of the parent cpuset's unused memory nodes to the child
> cpuset, up to specified limit. It's a bit ugly, but lets you trade of
> performance vs memory footprint on a per-job basis (when combined with
> fake numa to give lots of small nodes).

Interesting. So you could set up the fake node with "guarantee" and let
it grow till "limit" ?

BTW, can you do these with fake nodes:
 - dynamic creation
 - dynamic removal
 - dynamic change of size

Also, How could we account when a process moves from one node to
another ?

>
> 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 01:52:30 GMT
View Forum Message <> Reply to Message

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.

>
> 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.

>
> 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.

Paul
