Subject: Re: [patch00/05]: Containers(V2)- Introduction
Posted by Rohit Seth on Wed, 20 Sep 2006 17:26:47 GMT
View Forum Message <> Reply to Message

On Wed, 2006-09-20 at 09:25 -0700, Christoph Lameter wrote:
> On Tue, 19 Sep 2006, Rohit Seth wrote:
>
> > For example, a user can run a batch job like backup inside containers.
> > This job if run unconstrained could step over most of the memory present
> > in system thus impacting other workloads running on the system at that
> > time. But when the same job is run inside containers then the backup
> > job is run within container limits.
>
> I just saw this for the first time since linux-mm was not cced. We have
> discussed a similar mechanism on linux-mm.
>
> We already have such a functionality in the kernel its called a cpuset. A
> container could be created simply by creating a fake node that then
> allows constraining applications to this node. We already track the
> types of pages per node. The statistics you want are already existing.
> See /proc/zoneinfo and /sys/devices/system/node/node*/*.
>
> > We use the term container to indicate a structure against which we track
> > and charge utilization of system resources like memory, tasks etc for a
> > workload. Containers will allow system admins to customize the
> > underlying platform for different applications based on their
> > performance and HW resource utilization needs. Containers contain
> > enough infrastructure to allow optimal resource utilization without
> > bogging down rest of the kernel. A system admin should be able to
> > create, manage and free containers easily.
>
> Right thats what cpusets do and it has been working fine for years. Maybe
> Paul can help you if you find anything missing in the existing means to
> control resources.

cpusets provides cpu and memory NODES binding to tasks. And I think it
works great for NUMA machines where you have different nodes with its
own set of CPUs and memory. The number of those nodes on a commodity HW
is still 1. And they can have 8-16 CPUs and in access of 100G of
memory. You may start using fake nodes (untested territory) to
translate a single node machine into N different nodes. But am not sure
if this number of nodes can change dynamically on the running machine or
a reboot is required to change the number of nodes.

Though when you want to have in access of 100 containers then the cpuset
function starts popping up on the oprofile chart very aggressively. And
this is the cost that shouldn't have to be paid (particularly) for a
single node machine.

And what happens when you want to have cpuset with memory that needs to be even further fine grained than each node.

Containers also provide a mechanism to move files to containers. Any further references to this file come from the same container rather than the container which is bringing in a new page.

In future there will be more handlers like CPU and disk that can be easily embeded into this container infrastructure.

-rohit
