Subject: Re: swsusp done by migration (was Re: [RFC][PATCH 1/5] Virtualization/containers: startup)
Posted by Kyle Moffett on Fri, 10 Feb 2006 08:29:57 GMT
View Forum Message <> Reply to Message

On Feb 09, 2006, at 23:31, Sam Vilain wrote:
> Kyle Moffett wrote:
>> <wishful thinking>
>> I can see another extension to this functionality. With
>> appropriate changes it might also be possible to have a container
>> exist across multiple computers using some cluster code for
>> synchronization and fencing. The outermost container would be
>> the system boot container, and multiple inner containers would
>> use some sort of network- container-aware cluster filesystem to
>> spread multiple vservers across multiple servers, distributing
>> CPU and network load appropriately.
>> </wishful thinking>
>
> Yeah. If you fudged/virtualised /dev/random, the system clock, etc
> you could even have Tandem-style transparent High Availability.
> </more wishful thinking>
>
> Actually there is relatively little difference between a NUMA
> system and a cluster...

Yeah, a cluster is just a multi-tiered multi-address-space RNUMA
(*Really* Non-Uniform Memory Architecture) :-D. With some kind of
RDMA infiniband card and the right kernel and userspace tools, that
kind of cluster could be practical.

I _suspect_ (never really considered the issue before) that a
properly virtualized container could even achieve extremely high
fault tolerance by allowing systems to "vote" on correct output. If
you synchronize /dev/random and network IO across the system
correctly such that each instance of each userspace process on each
system sees _exactly_ the same virtual inputs and virtual clock in
the exact same order, then you could binary-compare the output of 3
different servers. If one didn't agree, it could be discarded and
marked as failing.

Cheers,
Kyle Moffett

--
There are two ways of constructing a software design. One way is to
make it so simple that there are obviously no deficiencies. And the
other way is to make it so complicated that there are no obvious
deficiencies. The first method is far more difficult.

-- C.A.R. Hoare