Subject: Re: [RFC][PATCH 2/7] RSS controller core
Posted by Herbert Poetzl on Mon, 12 Mar 2007 21:11:11 GMT
View Forum Message <> Reply to Message

On Mon, Mar 12, 2007 at 12:02:01PM +0300, Pavel Emelianov wrote:
> >>> Maybe you have some ideas how we can decide on this?
> >> We need to work out what the requirements are before we can
> >> settle on an implementation.
> >
> > Linux-VServer (and probably OpenVZ):
> >
> > - shared mappings of 'shared' files (binaries
> >   and libraries) to allow for reduced memory
> >   footprint when N identical guests are running
>
> This is done in current patches.

nice, but the question was about _requirements_
(so your requirements are?)

> > - virtual 'physical' limit should not cause
> >   swap out when there are still pages left on
> >   the host system (but pages of over limit guests
> >   can be preferred for swapping)
>
> So what to do when virtual physical limit is hit?
> OOM-kill current task?

when the RSS limit is hit, but there _are_ enough
pages left on the physical system, there is no
good reason to swap out the page at all

 - there is no benefit in doing so (performance
   wise, that is)

 - it actually hurts performance, and could
   become a separate source for DoS

what should happen instead (in an ideal world :)
is that the page is considered swapped out for
the guest (add guest penality for swapout), and
when the page would be swapped in again, the guest
takes a penalty (for the 'virtual' page in) and
the page is returned to the guest, possibly kicking
out (again virtually) a different page

> > - accounting and limits have to be consistent
> >   and should roughly represent the actual used

> > memory/swap (modulo optimizations, I can go
> > into detail here, if necessary)
>
> This is true for current implementation for
> booth - this patchset ang OpenVZ beancounters.
>
> If you sum up the physpages values for all containers
> you'll get the exact number of RAM pages used.

hmm, including or excluding the host pages?

> > - OOM handling on a per guest basis, i.e. some
> > out of memory condition in guest A must not
> > affect guest B
>
> This is done in current patches.

> Herbert, did you look at the patches before
> sending this mail or do you just want to
> 'take part' in conversation w/o understanding
> of hat is going on?

again, the question was about requirements, not
your patches, and yes, I had a look at them _and_
the OpenVZ implementations ...

best,
Herbert

PS: hat is going on? :)

> > HTC,
> > Herbert
> >
> >> Sigh. Who is running this show? Anyone?
> >>
> >> You can actually do a form of overcommittment by allowing multiple
> >> containers to share one or more of the zones. Whether that is
> >> sufficient or suitable I don't know. That depends on the requirements,
> >> and we haven't even discussed those, let alone agreed to them.
> >>
> >> _____
> >> Containers mailing list
> >> Containers@lists.osdl.org
> >> https://lists.osdl.org/mailman/listinfo/containers
> >

_____
Containers mailing list