Subject: Re: [Fwd: DELIVERY FAILURE: 554 Service unavailable; Client host [32.97.110.153] blocked using black

Posted by Herbert Poetzl on Tue, 13 Mar 2007 14:24:08 GMT

View Forum Message <> Reply to Message

> From: Hansen donotmail <hansendc@us.ibm.com>
> To: Herbert Poetzl <herbert@13thfloor.at>
> Cc: Andrew Morton <akpm@linux-foundation.org>,
> containers@lists.osdl.org, linux-kernel@vger.kernel.org,
> menage@google.com, xemul@sw.ru
> Subject: Re: [RFC][PATCH 2/7] RSS controller core
> Date: Mon, 12 Mar 2007 16:02:08 -0700
>
> On Mon, 2007-03-12 at 23:41 +0100, Herbert Poetzl wrote:
> > On Mon, Mar 12, 2007 at 11:42:59AM -0700, Dave Hansen wrote:
> > > How about we drill down on these a bit more.
> > >
> > > On Mon, 2007-03-12 at 02:00 +0100, Herbert Poetzl wrote:
> > > > - shared mappings of 'shared' files (binaries
> > > >   and libraries) to allow for reduced memory
> > > >   footprint when N identical guests are running
> > >
> > > So, it sounds like this can be phrased as a requirement like:
> > >
> > > "Guests must be able to share pages."
> > >
> > > Can you give us an idea why this is so?
> >
> > sure, one reason for this is that guests tend to
> > be similar (or almost identical) which results
> > in quite a lot of 'shared' libraries and executables
> > which would otherwise get cached for each guest and
> > would also be mapped for each guest separately
> >
> > > On a typical vserver system,
> >
> > there is nothing like a typical Linux-VServer system :)
> >
> > > how much memory would be lost if guests were not permitted
> > > to share pages like this?
> >
> > let me give a real world example here:
> >
> > - typical guest with 600MB disk space
> > - about 100MB guest specific data (not shared)
> > - assumed that 80% of the libs/tools are used
>
> I get the general idea here, but I just don't think those numbers are

> very accurate. My laptop has a bunch of gunk open (xterm, evolution,
> firefox, xchat, etc...). I ran this command:
>
> lsof | egrep '/(usr/|lib.*\.so)' | awk '{print $9}' | sort \
> | uniq | xargs du -Dcs
>
> and got:
>
> 113840 total
>
> On a web/database server that I have (ps aux | wc -l == 128), I just ran
> the same:
>
> 39168   total

> That's assuming that all of the libraries are fully read in and
> populated, just by their on-disk sizes.

> Is that not a reasonable measure of the kinds of things that we
> can expect to be shared in a vserver?

no, you have to add the binaries too, as they are mapped
read only/executeable and shared between guests .. same
goes for unmodified data files and the inode cache ...

> If so, it's a long way from 400MB.

when I find the time, I can actually set up 100 guests
with and without sharing and check the difference on
an otherwise unloaded system ...

> Could you try a similar measurement on some of your machines? Perhaps
> mine are just weird.

mail server here: binaries and libraries, no shared
files (as they are harder to figure) sum up to

  28904

while the total RSS used inside the guest is at

  102256

so we have roughly 1/3rd shared here between guests
(assumed they are sharing the data)

> > > > - 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)
> > >
> > > Is this a really hard requirement?
> >
> > no, not hard, but a reasonable optimization ...
> >
> > let me note once again, that for full isolation
> > you better go with Xen or some other Hypervisor
> > because if you make it work like Xen, it will
> > become as slow and resource hungry as any other
> > paravirtualization solution ...
>
> Believe me, _I_ don't want Xen. :)
>
> > > It seems a bit fluffy to me.
> >
> > most optimizations might look strange at first
> > glance, but when you check what the limitting
> > factors for OS-Level virtualizations are, you
> > will find that it looks like this:
> >
> > (in order of decreasing relevance)
> >
> > - I/O subsystem
> > - available memory
> > - network performance
> > - CPU performance
> >
> > note: this is for 'typical' guests, not for
> > number crunching or special database, or pure
> > network bound applications/guests ...
>
> I don't doubt this, but doing this two-level page-out thing
> for containers/vservers over their limits is surely something
> that we should consider farther down the road, right?

we should consider it now, and implement it lateron :)

but in general, I'm against adding code to mainline which
makes the system crawl and leave the fixup for later ...
we have too many cases in 2.6 which are the result of
such doing, and nobody knows if they will get fixed up
ever, because nobody can identify the introduced overhead

> It's important to you, but you're obviously not doing any of
> the mainline coding, right?

I have no problem working on mainline code, after all, the
Linux-VServer patches _are_ against mainline, it just takes
a little more time for me to code something up, and I spend
more time thinking about it in the first place ...

don't forget, I'm doing that in my spare time and I do a
lot of testing of the mainline changes in Linux-VServer,
which always follows mainline development quite closely

> > > What are the consequences if this isn't done? Doesn't
> > > a loaded system eventually have all of its pages used
> > > anyway, so won't this always be a temporary situation?
> >
> > let's consider a quite limited guest (or several
> > of them) which have a 'RAM' limit of 64MB and
> > additional 64MB of 'virtual swap' assigned ...
> >
> > if they use roughly 96MB (memory footprint) then
> > having this 'fluffy' optimization will keep them
> > running without any effect on the host side, but
> > without, they will continously swap in and out
> > which will affect not only the host, but also the
> > other guests ...
>
> All workloads that use $limit+1 pages of memory will always
> pay the price, right? :)

when they need a different page, yes, when they stick to
$limit pages for some time, no. quite similar to a real
system being slightly over RAM :)

best,
Herbert

> -- Dave

____
Containers mailing list
Containers@lists.osdl.org
https://lists.osdl.org/mailman/listinfo/containers