Subject: Re: [RFC] Virtualization steps
Posted by ebiederm on Tue, 28 Mar 2006 16:15:56 GMT
View Forum Message <> Reply to Message

Nick Piggin <nickpiggin@yahoo.com.au> writes:

> Kirill Korotaev wrote:
>> Nick, will be glad to shed some light on it.
>>
>
> Thanks very much Kirill.
>
> I don't think I'm qualified to make any decisions about this,
> so I don't want to detract from the real discussions, but I
> just had a couple more questions:
>
>> First of all, what it does which low level virtualization can't:
>> - it allows to run 100 containers on 1GB RAM
>>   (it is called containers, VE - Virtual Environments,
>>   VPS - Virtual Private Servers).
>> - it has no much overhead (<1-2%), which is unavoidable with hardware
>>   virtualization. For example, Xen has >20% overhead on disk I/O.
>
> Are any future hardware solutions likely to improve these problems?

This isn't a direct competition, both solutions coincide nicely.

The major efficiency differences are fundamental to the approaches and
can only be solved in software and not hardware. The fundamental efficiency
limits of low level virtualization are not sharing resources between
instances well (think how hard memory hotplug is to solve), the fact
that running a kernel takes at least 1MB for just the kernel, the
fact that no matter how good your hypervisor is there will be some
hardware interface it doesn't virtualize.

Whereas what we are aiming at are just enough modifications to the kernel
to allow multiple instances of user space. We aren't virtualizing anything
that isn't already virtualized in the kernel.

>> OS kernel virtualization
>> ~~~~~~~~~~~~~~~~~~~~~~~~
>
> Is this considered secure enough that multiple untrusted VEs are run
> on production systems?

Kirill or Herbert can give a better answer but that is of the major
points of BSD Jails and their kin is it not?

> What kind of users want this, who can't use alternatives like real
> VMs?

Well that question assumes a lot. The answer that assumes a lot in the other direction is that adding an additional unnecessary layers just complicates the problem and slows things down for no reason while making it so you can't assume the solution is always present. In addition to doing it in a non-portable way so it is only available on a few platforms.

I can't even think of a straight answer to the users question.

My users are in the high performance computing realm, and for that subset it is easy. Xen and it's kin don't virtualize the high bandwidth low latency communication hardware that is used, and that may not even be possible. Using a hypervisor in a situation like that certainly isn't general or easily maintainable. (Think about what a challenge it has been to get usable infiniband drivers merged).

>> Summary of previous discussions on LKML
>> ~~~~~~
>
> Have their been any discussions between the groups pushing this
> virtualization, and important kernel developers who are not part of
> a virtualization effort? Ie. is there any consensus about the
> future of these patches?

Yes, but just enough to give us hope :)

Unless you count the mount namespace as part of this in which case pieces are already merged.

The challenging is that writing kernel code that does this is easy. Writing kernel code that is mergeable and that the different groups all agree meets their requirements is much harder. It has taken us until now to have a basic approach that we all agree on. Now we get to beat each other up over the technical details :)

Eric