Subject: A question about Node RAM
Posted by max0181 on Fri, 06 Jan 2012 16:59:05 GMT
View Forum Message <> Reply to Message

Hello,

I've a question for this mailing-list ^^

My enterprise is going to order a 128Gb of RAM server.
I saw that the OpenVZ Kernel can only support 64Gb.

That's because the wiki isn't up to date ?
What's about that ?
How to bypass this limit ? Can we ?
Recompiling the kernel.. ?

It's important for us =)

Thanks !

Subject: Re: A question about Node RAM
Posted by Kirill Korotaev on Fri, 06 Jan 2012 17:39:39 GMT
View Forum Message <> Reply to Message

Sure, it's old information and likely it was about 32bit kernels which are limited to 64GB just because CPUs are... :)
64bit kernels are not limited anyhow and OpenVZ is not different in this regard from standard Linux.

fixed a couple of places I found with 64GB mentioning:
http://wiki.openvz.org/Different_kernel_flavors_(UP,_SMP,_ENTERPRISE,_ENTNOSPLIT)
http://wiki.openvz.org/FAQ#How_scalable_is_OpenVZ.3F

Thanks,
Kirill

On Jan 6, 2012, at 20:59 , Quentin MACHU wrote:

> Hello,
>
> I've a question for this mailing-list ^^
>
> My enterprise is going to order a 128Gb of RAM server.
> I saw that the OpenVZ Kernel can only support 64Gb.
>
> That's because the wiki isn't up to date ?
> What's about that ?

> How to bypass this limit ? Can we ?
> Recompiling the kernel.. ?
>
> It's important for us =)
>
> Thanks !

Subject: Re: A question about Node RAM
Posted by max0181 on Fri, 06 Jan 2012 18:08:13 GMT
View Forum Message <> Reply to Message

Hello,

Thanks for this answer.
So, we can use 128Gb/256Gb server ? =]

Actually, we're working on Debian 6.
Do you have any tips on Distro / Kernel ?

Debian 6 + Kernel from Debian repos is really stable ? Debian 5 more maybe ?

We'll have 6*3To HardDrive SAS in RAID 10 to improve I/O
And Two *Opteron 6128 8 cores* Magny-Cours 8x 2Ghz.

Do you think it's ok for something like 126 VM with 1Gb of RAM ? =)

Thanks for all :)

2012/1/6 Kirill Korotaev <dev@parallels.com>

> Sure, it's old information and likely it was about 32bit kernels which are
> limited to 64GB just because CPUs are... :)
> 64bit kernels are not limited anyhow and OpenVZ is not different in this
> regard from standard Linux.
>
> fixed a couple of places I found with 64GB mentioning:
>
> http://wiki.openvz.org/Different_kernel_flavors_(UP,_SMP,_ENTERPRISE,_ENTNOSPLIT)
> http://wiki.openvz.org/FAQ#How_scalable_is_OpenVZ.3F
>
> Thanks,
> Kirill
>
> On Jan 6, 2012, at 20:59 , Quentin MACHU wrote:
>
> > Hello,

> >
> > I've a question for this mailing-list ^^
> >
> > My enterprise is going to order a 128Gb of RAM server.
> > I saw that the OpenVZ Kernel can only support 64Gb.
> >
> > That's because the wiki isn't up to date ?
> > What's about that ?
> > How to bypass this limit ? Can we ?
> > Recompiling the kernel.. ?
> >
> > It's important for us =)
> >
> > Thanks !
--
Cordialement,
MACHU Quentin

Subject: Re: A question about Node RAM
Posted by jjs - mainphrame on Fri, 06 Jan 2012 18:18:23 GMT
View Forum Message <> Reply to Message

I'm running openvz on Debian 6 and recently switched to the rhel6-based kernel which provides the vswap configuration option. That was a big improvement, and the rhel kernel rpms were very easy to convert to debs which worked like a charm.

Joe

On Fri, Jan 6, 2012 at 10:08 AM, Quentin MACHU <quentin.machu@gmail.com>wrote:

> Hello,
>
> Thanks for this answer.
> So, we can use 128Gb/256Gb server ? =]
>
> Actually, we're working on Debian 6.
> Do you have any tips on Distro / Kernel ?
>
> Debian 6 + Kernel from Debian repos is really stable ? Debian 5 more maybe
> ?
>
> We'll have 6*3To HardDrive SAS in RAID 10 to improve I/O
> And Two *Opteron 6128 8 cores* Magny-Cours 8x 2Ghz.
>
> Do you think it's ok for something like 126 VM with 1Gb of RAM ? =)
>

> Thanks for all :)
>
>
>
> 2012/1/6 Kirill Korotaev <dev@parallels.com>
>
>> Sure, it's old information and likely it was about 32bit kernels which
>> are limited to 64GB just because CPUs are... :)
>> 64bit kernels are not limited anyhow and OpenVZ is not different in this
>> regard from standard Linux.
>>
>> fixed a couple of places I found with 64GB mentioning:
>>
>> http://wiki.openvz.org/Different_kernel_flavors_(UP,_SMP,_ENTERPRISE,_ENTNOSPLIT)
>> http://wiki.openvz.org/FAQ#How_scalable_is_OpenVZ.3F
>>
>> Thanks,
>> Kirill
>>
>> On Jan 6, 2012, at 20:59 , Quentin MACHU wrote:
>>
>> > Hello,
>> >
>> > I've a question for this mailing-list ^^
>> >
>> > My enterprise is going to order a 128Gb of RAM server.
>> > I saw that the OpenVZ Kernel can only support 64Gb.
>> >
>> > That's because the wiki isn't up to date ?
>> > What's about that ?
>> > How to bypass this limit ? Can we ?
>> > Recompiling the kernel.. ?
>> >
>> > It's important for us =)
>> >
>> > Thanks !
> --
> Cordialement,
> MACHU Quentin
>
>

Subject: Re: A question about Node RAM
Posted by Martin Dobrev on Fri, 06 Jan 2012 18:23:23 GMT
View Forum Message <> Reply to Message

Sounds really massive. I'm not really sure if the I/O will fit, even with RAID 10, but of course it only

depends on how I/O intensive your VEs will be. On the other hand, as Kiril already mentioned, OpenVZ kernel is not so different from mainstream kernels and as such there should be no limit in the number of the running VEs.

Martin Dobrev

Sent from iPhonespam SPAMSPAM 4

On 06.01.2012, at 20:08, Quentin MACHU <quentin.machu@gmail.com> wrote:

> Hello,
>
> Thanks for this answer.
> So, we can use 128Gb/256Gb server ? =]
>
> Actually, we're working on Debian 6.
> Do you have any tips on Distro / Kernel ?
>
> Debian 6 + Kernel from Debian repos is really stable ? Debian 5 more maybe ?
>
> We'll have 6*3To HardDrive SAS in RAID 10 to improve I/O
> And Two Opteron 6128 8 cores Magny-Cours 8x 2Ghz.
>
> Do you think it's ok for something like 126 VM with 1Gb of RAM ? =)
>
> Thanks for all :)
>
>
> 2012/1/6 Kirill Korotaev <dev@parallels.com>
> Sure, it's old information and likely it was about 32bit kernels which are limited to 64GB just because CPUs are... :)
> 64bit kernels are not limited anyhow and OpenVZ is not different in this regard from standard Linux.
>
> fixed a couple of places I found with 64GB mentioning:
> http://wiki.openvz.org/Different_kernel_flavors_(UP,_SMP,_ENTERPRISE,_ENTNOSPLIT)
> http://wiki.openvz.org/FAQ#How_scalable_is_OpenVZ.3F
>
> Thanks,
> Kirill
>
> On Jan 6, 2012, at 20:59 , Quentin MACHU wrote:
>
> > Hello,
> >
> > I've a question for this mailing-list ^^
> >
> > My enterprise is going to order a 128Gb of RAM server.

> > I saw that the OpenVZ Kernel can only support 64Gb.
> >
> > That's because the wiki isn't up to date ?
> > What's about that ?
> > How to bypass this limit ? Can we ?
> > Recompiling the kernel.. ?
> >
> > It's important for us =)
> >
> > Thanks !
> --
> Cordialement,
> MACHU Quentin
>

Subject: Re: A question about Node RAM
Posted by Kirill Korotaev on Fri, 06 Jan 2012 19:13:52 GMT
View Forum Message <> Reply to Message

>From RAM/CPU perspective this configuration is fine.
But if you plan to run I/O intensive apps you may want to have more HDD drives (maybe with less capacity each) to make your raid capable to handle more IOPS.

Kirill

On Jan 6, 2012, at 22:08 , Quentin MACHU wrote:

> Hello,
>
> Thanks for this answer.
> So, we can use 128Gb/256Gb server ? =]
>
> Actually, we're working on Debian 6.
> Do you have any tips on Distro / Kernel ?
>
> Debian 6 + Kernel from Debian repos is really stable ? Debian 5 more maybe ?
>
> We'll have 6*3To HardDrive SAS in RAID 10 to improve I/O
> And Two Opteron 6128 8 cores Magny-Cours 8x 2Ghz.
>
> Do you think it's ok for something like 126 VM with 1Gb of RAM ? =)
>
> Thanks for all :)
>
>
> 2012/1/6 Kirill Korotaev <dev@parallels.com>
> Sure, it's old information and likely it was about 32bit kernels which are limited to 64GB just

because CPUs are... :)
> 64bit kernels are not limited anyhow and OpenVZ is not different in this regard from standard Linux.
>
> fixed a couple of places I found with 64GB mentioning:
> http://wiki.openvz.org/Different_kernel_flavors_(UP,_SMP,_ENTERPRISE,_ENTNOSPLIT)
> http://wiki.openvz.org/FAQ#How_scalable_is_OpenVZ.3F
>
> Thanks,
> Kirill
>
> On Jan 6, 2012, at 20:59 , Quentin MACHU wrote:
>
> > Hello,
> >
> > I've a question for this mailing-list ^^
> >
> > My enterprise is going to order a 128Gb of RAM server.
> > I saw that the OpenVZ Kernel can only support 64Gb.
> >
> > That's because the wiki isn't up to date ?
> > What's about that ?
> > How to bypass this limit ? Can we ?
> > Recompiling the kernel.. ?
> >
> > It's important for us =)
> >
> > Thanks !
> --
> Cordialement,
> MACHU Quentin
>
> <ATT00001.c>

Subject: Re: A question about Node RAM
Posted by max0181 on Fri, 06 Jan 2012 19:35:33 GMT
View Forum Message <> Reply to Message

Hello,

Thanks again!

You mean that we should use for exemple this stable kernel :
 http://download.openvz.org/kernel/branches/rhel6-2.6.32/042s
tab044.11/vzkernel-2.6.32-042stab044.11.i686.rpmto
get a lot of stability ? By following this little guide :
http://wiki.openvz.org/Install_kernel_from_rpm_on_debian.

The apps won't be so disk IO-vore. Tons of VM are for... LAMP / VocalServer / Minecraft & other game servers...

Another tips to have an OpenVZ Stable Node ?

I think we should use vzsplit -n 128 to get a UBC-configuration.
I don't know anything else.

We've 12 dedicted servers on Debian 6. Some of them aren't so stable, using the kernel from repos. We sometimes need to make an hard-reboot.
We'll migrate these 12 servers to one.

Thanks !

PS: Sorry If I post wrong, first time use of mailing-lists.

2012/1/6 Kirill Korotaev <dev@parallels.com>

> >From RAM/CPU perspective this configuration is fine.
> But if you plan to run I/O intensive apps you may want to have more HDD
> drives (maybe with less capacity each) to make your raid capable to handle
> more IOPS.
>
> Kirill
>
> On Jan 6, 2012, at 22:08 , Quentin MACHU wrote:
>
> > Hello,
> >
> > Thanks for this answer.
> > So, we can use 128Gb/256Gb server ? =]
> >
> > Actually, we're working on Debian 6.
> > Do you have any tips on Distro / Kernel ?
> >
> > Debian 6 + Kernel from Debian repos is really stable ? Debian 5 more
> maybe ?
> >
> > We'll have 6*3To HardDrive SAS in RAID 10 to improve I/O
> > And Two Opteron 6128 8 cores Magny-Cours 8x 2Ghz.
> >
> > Do you think it's ok for something like 126 VM with 1Gb of RAM ? =)
> >
> > Thanks for all :)
> >
> >
> > 2012/1/6 Kirill Korotaev <dev@parallels.com>

> > Sure, it's old information and likely it was about 32bit kernels which
> are limited to 64GB just because CPUs are... :)
> > 64bit kernels are not limited anyhow and OpenVZ is not different in this
> regard from standard Linux.
> >
> > fixed a couple of places I found with 64GB mentioning:
> >
> http://wiki.openvz.org/Different_kernel_flavors_(UP,_SMP,_ENTERPRISE,_ENTNOSPLIT)
> > http://wiki.openvz.org/FAQ#How_scalable_is_OpenVZ.3F
> >
> > Thanks,
> > Kirill
> >
> > On Jan 6, 2012, at 20:59 , Quentin MACHU wrote:
> >
> > > Hello,
> > >
> > > I've a question for this mailing-list ^^
> > >
> > > My enterprise is going to order a 128Gb of RAM server.
> > > I saw that the OpenVZ Kernel can only support 64Gb.
> > >
> > > That's because the wiki isn't up to date ?
> > > What's about that ?
> > > How to bypass this limit ? Can we ?
> > > Recompiling the kernel.. ?
> > >
> > > It's important for us =)
> > >
> > > Thanks !
> > --
> > Cordialement,
> > MACHU Quentin
> >
> > <ATT00001.c>
>
>
--
Cordialement,
MACHU Quentin

Subject: Re: A question about Node RAM
Posted by dowdle on Fri, 06 Jan 2012 19:55:41 GMT
View Forum Message <> Reply to Message

Greetings,

----- Original Message -----
> I'm running openvz on Debian 6 and recently switched to the
> rhel6-based kernel which provides the vswap configuration option.
> That was a big improvement, and the rhel kernel rpms were very easy
> to convert to debs which worked like a charm.

Just a few comments. If you want to run the RHEL6 kernel, and that's what I'm running, why not run it on RHEL6 or a RHEL6 clone?

Debian isn't supported very long... about 3 years (from initial release). RHEL6 will be supported for some time.

On an OpenVZ host node installing services and adding users isn't recommended and you want to keep your host node pretty minimal. I know one of the advantages of Debian is that it is a fantastic server OS with a large library of software... which you don't care about for an OpenVZ host node.

I'm just saying. :)

TYL,
--
Scott Dowdle
704 Church Street
Belgrade, MT 59714
(406)388-0827 [home]
(406)994-3931 [work]

Subject: Re: A question about Node RAM
Posted by jjs - mainphrame on Fri, 06 Jan 2012 20:08:51 GMT
View Forum Message <> Reply to Message

I started out using debian, but I'm building new openvz servers on centos.
There are some who for ideological reasons prefer to stick with debian, and
I completely understand that. The rhel-based kernel provides the best
results for existing debian openvz servers. Personally I'm pragmatic, and
prefer what works best, but changing distros can take awhile when
production servers are involved.

Joe

On Fri, Jan 6, 2012 at 11:55 AM, Scott Dowdle <dowdle@montanalinux.org>wrote:

> Greetings,
>
> ----- Original Message -----
> > I'm running openvz on Debian 6 and recently switched to the
> > rhel6-based kernel which provides the vswap configuration option.

> > That was a big improvement, and the rhel kernel rpms were very easy
> > to convert to debs which worked like a charm.
>
> Just a few comments. If you want to run the RHEL6 kernel, and that's what
> I'm running, why not run it on RHEL6 or a RHEL6 clone?
>
> Debian isn't supported very long... about 3 years (from initial release).
> RHEL6 will be supported for some time.
>
> On an OpenVZ host node installing services and adding users isn't
> recommended and you want to keep your host node pretty minimal. I know one
> of the advantages of Debian is that it is a fantastic server OS with a
> large library of software... which you don't care about for an OpenVZ host
> node.
>
> I'm just saying. :)
>
> TYL,
> --
> Scott Dowdle
> 704 Church Street
> Belgrade, MT 59714
> (406)388-0827 [home]
> (406)994-3931 [work]

Subject: Re: A question about Node RAM
Posted by Tim Small on Fri, 06 Jan 2012 20:19:46 GMT
View Forum Message <> Reply to Message

On 06/01/12 19:35, Quentin MACHU wrote:
> Hello,
>
> Thanks again!
>
> You mean that we should use for exemple this stable kernel :
> http://download.openvz.org/kernel/branches/rhel6-2.6.32/042s
tab044.11/vzkernel-2.6.32-042stab044.11.i686.rpm
> to get a lot of stability ? By following this little guide :
> http://wiki.openvz.org/Install_kernel_from_rpm_on_debian.
>
> The apps won't be so disk IO-vore. Tons of VM are for... LAMP /
> VocalServer / Minecraft & other game servers...

Isn't that "putting all your eggs in one basket"? What happens if that
machine has a hardware fault? Personally, I'd perhaps favour going for
e.g. 4 or 5 Sandy Bridge based machines, each being quad core, and with
32G RAM (maybe something like a Dell R210 II), and use some sort of

clustering system (maybe pacemaker with drbd, or glusterfs, or sheepdog) to distribute the storage between the nodes, and allow moving VMs between nodes.

May well be cheaper too, but almost certainly more reliable... Larger numbers of simpler cheaper machines is how Google, Amazon etc. do it - big fat machines like the one you've described are usually trouble in my experience...

Tim.

--
South East Open Source Solutions Limited
Registered in England and Wales with company number 06134732.
Registered Office: 2 Powell Gardens, Redhill, Surrey, RH1 1TQ
VAT number: 900 6633 53 http://seoss.co.uk/ +44-(0)1273-808309

Subject: Re: A question about Node RAM
Posted by Tim Small on Fri, 06 Jan 2012 20:34:35 GMT
View Forum Message <> Reply to Message

On 06/01/12 19:55, Scott Dowdle wrote:

Fair points, but FWIW...

We run OpenVZ hardware nodes on Debian, because:

> Debian isn't supported very long... about 3 years (from initial release). RHEL6 will be supported for some time.
>

Debian lets you easily and reliably upgrade in-place to the next release.

> I know one of the advantages of Debian is that it is a fantastic server OS with a large library of software... which you don't care about for an OpenVZ host node.
>

I'm not sure where EPEL is these days, but we run the following packages on our hardware nodes, which aren't packaged with EL5 (not so sure about RHEL6 - maybe a couple of those are in there now):

smartd
logcheck
munin
ipmitool
pacemaker+heartbeat
drbd

puppet
arno-iptables-firewall

... and probably a few others which I've forgotten about. I know that several of those are 3rd-party packaged for RHEL5/6, but then you're losing a lot of the "it's-guaranteed-supported-for-ages" benefit anyway, and you've got to fart about researching and tracking down software, and you're arguably worsening your security as a result too? No?

Cheers,

Tim.

--
South East Open Source Solutions Limited
Registered in England and Wales with company number 06134732.
Registered Office: 2 Powell Gardens, Redhill, Surrey, RH1 1TQ
VAT number: 900 6633 53 http://seoss.co.uk/ +44-(0)1273-808309

Subject: Re: A question about Node RAM
Posted by Sharp on Fri, 06 Jan 2012 20:54:17 GMT
View Forum Message <> Reply to Message

On Fri, Jan 06, 2012 at 08:34:35PM +0000, Tim Small wrote:
> I'm not sure where EPEL is these days, but we run the following packages
> on our hardware nodes, which aren't packaged with EL5 (not so sure about
> RHEL6 - maybe a couple of those are in there now):

Just looking at el6.

> smartd
smartmontools is in the RHEL itself.

> logcheck
> munin
EPEL has those.

> ipmitool
RHEL has that.

> pacemaker+heartbeat
pacemaker is in RHEL and hearbeat is in EPEL.

> drbd
Can find only this:
drbdlinks.noarch : A program for managing links into a DRBD shared
partition

> puppet
EPEL surely has it.

> arno-iptables-firewall
It's absent. But I did a google search about that and I can understand
why there isn't a package such as that.

>
> ... and probably a few others which I've forgotten about. I know that
> several of those are 3rd-party packaged for RHEL5/6, but then you're
> losing a lot of the "it's-guaranteed-supported-for-ages" benefit anyway,
> and you've got to fart about researching and tracking down software, and
> you're arguably worsening your security as a result too? No?
Usually you have all you need inside stock RHEL or with the addition of
EPEL. If there is something you need and it is absent from EPEL -- you
can always became a package maintainer for it. Fedora community is a
great gang and they will always help you.

--
SY, Ilya A. Otyutskiy aka Sharp

Subject: Re: A question about Node RAM
Posted by jjs - mainphrame on Fri, 06 Jan 2012 22:59:16 GMT
View Forum Message <> Reply to Message

On Fri, Jan 6, 2012 at 12:34 PM, Tim Small <tim@seoss.co.uk> wrote:

>
> pacemaker+heartbeat

Interesting idea, I wonder about the tradeoffs. I tend to keep the host
node pretty lean and run heartbeat/corosync/pacemaker in the CTs, if
anywhere.

Joe

Subject: Re: A question about Node RAM
Posted by Kirill Korotaev on Sat, 07 Jan 2012 08:33:25 GMT
View Forum Message <> Reply to Message

On Jan 7, 2012, at 00:19 , Tim Small wrote:

> On 06/01/12 19:35, Quentin MACHU wrote:

>> Hello,
>>
>> Thanks again!
>>
>> You mean that we should use for exemple this stable kernel : http://download.openvz.org/kernel/branches/rhel6-2.6.32/042stab044.11/vzkernel-2.6.32-042stab044.11.i686.rpm to get a lot of stability ? By following this little guide : http://wiki.openvz.org/Install_kernel_from_rpm_on_debian.
>>
>> The apps won't be so disk IO-vore. Tons of VM are for... LAMP / VocalServer / Minecraft & other game servers...
>
> Isn't that "putting all your eggs in one basket"? What happens if that machine has a hardware fault? Personally, I'd perhaps favour going for e.g. 4 or 5 Sandy Bridge based machines, each being quad core, and with 32G RAM (maybe something like a Dell R210 II), and use some sort of clustering system (maybe pacemaker with drbd, or glusterfs, or sheepdog) to distribute the storage between the nodes, and allow moving VMs between nodes.

do not recomment gluster or sheepdog - they are nowhere near production quality. So speaking about reliability - a described HW with SAS drives RAID is by far more reliable.

> May well be cheaper too, but almost certainly more reliable... Larger numbers of simpler cheaper machines is how Google, Amazon etc. do it - big fat machines like the one you've described are usually trouble in my experience...
>
> Tim.
> --
> South East Open Source Solutions Limited
> Registered in England and Wales with company number 06134732.
> Registered Office: 2 Powell Gardens, Redhill, Surrey, RH1 1TQ
> VAT number: 900 6633 53
> http://seoss.co.uk/ +44-(0)1273-808309
> <ATT00001.c>

Subject: Re: A question about Node RAM
Posted by Tim Small on Sat, 07 Jan 2012 17:02:00 GMT
View Forum Message <> Reply to Message

On 06/01/12 22:59, jjs - mainphrame wrote:
> On Fri, Jan 6, 2012 at 12:34 PM, Tim Small <tim@seoss.co.uk
> <mailto:tim@seoss.co.uk>> wrote:
>
>
>
> pacemaker+heartbeat
>
>
> Interesting idea, I wonder about the tradeoffs. I tend to keep the

> host node pretty lean and run heartbeat/corosync/pacemaker in the CTs,
> if anywhere.
>

We have a few machines where we put the OpenVZ container backing stores on drbd and use heartbeat+pacemaker (we had some issues with corosync during testing when we initially set things up a few years ago, but it's probably fine now) to manage the OpenVZ containers as cluster resources.

Disk writes are relatively expensive so it's not perfect for all workloads, but it works well overall, and has survived real hardware failures (e.g. motherboard failure) with minimal downtime.

It also allows you to move nodes around easily and should allow you to carry out things like host node kernel updates without bringing down containers (using live migration to other HNs) - although we've not gotten around to testing this.

Our machines are in pairs, but really it'd be better to have them in something like groups of four, so that when a HN fails, the remaining 3 HNs each end up running a third of the evicted containers... This would require corosync instead of heartbeat of course (heartbeat supports 2 nodes only).

Tim.

--
South East Open Source Solutions Limited
Registered in England and Wales with company number 06134732.
Registered Office: 2 Powell Gardens, Redhill, Surrey, RH1 1TQ
VAT number: 900 6633 53 http://seoss.co.uk/ +44-(0)1273-808309

Subject: Re: A question about Node RAM
Posted by Tim Small on Sat, 07 Jan 2012 17:11:12 GMT
View Forum Message <> Reply to Message

On 06/01/12 20:54, Ilya A. Otyutskiy wrote:
> On Fri, Jan 06, 2012 at 08:34:35PM +0000, Tim Small wrote:
>
>> I'm not sure where EPEL is these days, but we run the following packages
>> on our hardware nodes, which aren't packaged with EL5 (not so sure about
>> RHEL6 - maybe a couple of those are in there now):
>>
> Just looking at el6....
>

Thanks for the research on that - it's handy to know and overall that

picture is certainly better than the last time I tried setting a node up with a similar set of software under EL5 (although I wonder how good logcheck ends up being when it's not a core part of the distro - for us it's a really key piece of software) - I should say tho' that I'd expect to end up doing more backporting and overall fiddling about with EL6 than I would with Debian.

That having been said, I'd expect to get a slightly more stable kernel out of Redhat, as it's hard to better their engineering team, but then again I've not really seen any more problems on the Debian nodes which I've managed than I have on the Redhat ones...

Cheers,

Tim.

--
South East Open Source Solutions Limited
Registered in England and Wales with company number 06134732.
Registered Office: 2 Powell Gardens, Redhill, Surrey, RH1 1TQ
VAT number: 900 6633 53 http://seoss.co.uk/ +44-(0)1273-808309

Subject: Re: A question about Node RAM
Posted by Tim Small on Sat, 07 Jan 2012 17:21:10 GMT
View Forum Message <> Reply to Message

On 07/01/12 08:33, Kirill Korotaev wrote:
> > (maybe pacemaker with drbd, or glusterfs, or sheepdog) to distribute the storage between the nodes, and allow moving VMs between nodes.
>
> do not recomment gluster or sheepdog - they are nowhere near production quality. So speaking about reliability - a described HW with SAS drives RAID is by far more reliable.
>

I've not done much with gluster, but I know people who have it in production and are happy with it. Sheepdog looks very promising - we've had a bit of a play with it and plan to do more investigation in the future... We have drbd in production and are happy with it.

Over the years, I've had so much trouble with hardware RAID, that I now avoid it if at all possible.

My experience with real top-end hardware has been that you get to find lots of interesting new bugs (both software and hardware) because you're using relatively unusual hardware - you end up with a machine which is like 0.01% of the global machines running Linux instead of 5% or

whatever. When you do hit such bugs, often the developers can't
reproduce the issue because they don't have access to the same hardware...

RAISe - Rudundant Array of Inexpensive Servers! :-)

Tim.

--
South East Open Source Solutions Limited
Registered in England and Wales with company number 06134732.
Registered Office: 2 Powell Gardens, Redhill, Surrey, RH1 1TQ
VAT number: 900 6633 53 http://seoss.co.uk/ +44-(0)1273-808309

Subject: Re: A question about Node RAM
Posted by Kirill Korotaev on Sat, 07 Jan 2012 17:29:21 GMT
View Forum Message <> Reply to Message

On Jan 7, 2012, at 21:21 , Tim Small wrote:

> On 07/01/12 08:33, Kirill Korotaev wrote:
>>> (maybe pacemaker with drbd, or glusterfs, or sheepdog) to distribute the storage between the nodes, and allow moving VMs between nodes.
>>
>> do not recomment gluster or sheepdog - they are nowhere near production quality. So speaking about reliability - a described HW with SAS drives RAID is by far more reliable.
>>
>
> I've not done much with gluster, but I know people who have it in
> production and are happy with it. Sheepdog looks very promising - we've
> had a bit of a play with it and plan to do more investigation in the
> future... We have drbd in production and are happy with it.

How big total storage these people are running with gluster?

> Over the years, I've had so much trouble with hardware RAID, that I now
> avoid it if at all possible.
>
> My experience with real top-end hardware has been that you get to find
> lots of interesting new bugs (both software and hardware) because you're
> using relatively unusual hardware - you end up with a machine which is
> like 0.01% of the global machines running Linux instead of 5% or
> whatever. When you do hit such bugs, often the developers can't
> reproduce the issue because they don't have access to the same hardware...
>
> RAISe - Rudundant Array of Inexpensive Servers! :-)
>
> Tim.

>
> --
> South East Open Source Solutions Limited
> Registered in England and Wales with company number 06134732.
> Registered Office: 2 Powell Gardens, Redhill, Surrey, RH1 1TQ
> VAT number: 900 6633 53 http://seoss.co.uk/ +44-(0)1273-808309
>

Subject: Re: A question about Node RAM (gluster)
Posted by dowdle on Sat, 07 Jan 2012 19:35:52 GMT
View Forum Message <> Reply to Message

Greetings,

----- Original Message -----
> > I've not done much with gluster, but I know people who have it in
> > production and are happy with it. Sheepdog looks very promising -
> > we've had a bit of a play with it and plan to do more investigation in
> > the future... We have drbd in production and are happy with it.
>
> How big total storage these people are running with gluster?

I too have heard good things about Gluster from people (Research Computing Group at Montana State Bozeman) using it in production. I don't know exactly how much storage they have but I believe it is in the triple-digit TBs. They definitely plan to grow it as their needs increase.

TYL,
--
Scott Dowdle
704 Church Street
Belgrade, MT 59714
(406)388-0827 [home]
(406)994-3931 [work]

Subject: Re: A question about Node RAM
Posted by Aleksandar Ivanisevic on Mon, 06 Feb 2012 15:43:22 GMT
View Forum Message <> Reply to Message

Tim Small <tim@seoss.co.uk> writes:

> It also allows you to move nodes around easily and should allow you to
> carry out things like host node kernel updates without bringing down
> containers (using live migration to other HNs) - although we've not
> gotten around to testing this.

I've tested this and its terrible ;) Migration across two drbd volumes syncing at the same time -- disaster in terms of latency and I/O speed for the remaining node(s) in the cluster.

> Our machines are in pairs, but really it'd be better to have them in
> something like groups of four, so that when a HN fails, the remaining 3
> HNs each end up running a third of the evicted containers... This would
> require corosync instead of heartbeat of course (heartbeat supports 2
> nodes only).

Groups of four might work ok provided that the drbd devices are on separate disks and you are careful always to migrate to an unrelated machine that doesn't have the standby volume from the source.
