Subject: OpenVZ IO Isolation ?
Posted by andre on Thu, 28 Apr 2011 14:22:12 GMT
View Forum Message <> Reply to Message

Hello,

We have been working with OpenVZ for some years on some servers and have found a problematic behavior with one of our customers.

There are just 2 VE's on this hardware (let's call them VE1 and VE2). Each one with 25% of the resources (to give room for more 2 VEs that would be activated later). It is a Core2Quad with 8GB RAM and RAID1 SATA.

The problem:

VE1 is able to slow down the whole server, making IOWAIT for all VEs really high. To make sure it was not hardware related, we moved the VEs to another hardware (with latest stable kernel: 2.6.18-238.5.1.el5.028stab085.2) and got the same behavior.

Each VE has CPULIMIT=90% and CPUS=1

Reading "IOACCT", read/write/nice/etc information seems pretty "normal" but DIRTY BYTES is high.. So I thought: Could OpenVZ IO scheduler not be counting this ?

I am not a kernel expert (nor have basic kernel coding knowledge), just trying to understand what could cause this and if there are any solutions. At this moment we are keeping VE1 alone at a hardware so it won't impact other customers.

IOWAIT shows VE1's mysqld as the responsible. One VE's process being able to make the whole node near 100% IOWAIT is worrying

Thanks.

Subject: Re: OpenVZ IO Isolation ?
Posted by seanfulton on Sun, 01 May 2011 18:10:44 GMT
View Forum Message <> Reply to Message

We have experienced many I/O problems over the years from kernel to kernel. It seems like they get fixed in one kernel, then a couple of months later, things show up broken again in another kernel.

I had problems with 85.2 and 85.1, but have been running 88.1 (testing) for about a week with a noticable improvement. I booted up the new stable, 89.1 yesterday on a machine that had been running 85.2 and noticed a much speedier system almost immediately.

I would try a new kernel and report your results. 88.1 and 89.1 work for me!

sean

Subject: Re: OpenVZ IO Isolation ?
Posted by seanfulton on Tue, 03 May 2011 20:40:48 GMT
View Forum Message <> Reply to Message

Check that: 89.1 is slooooooooow. I upgraded a machine over the weekend from 85.2 to the new 89.1 stable and a particular process we run on that machine was a 2900 per hour. The client complained that it had been at about 5,000 per hour. I down-graded them to 88.1--with no other changes, and they are up around 5,400 per hour.

So 89.1 seems to have a problem.

sean

Subject: Re: OpenVZ IO Isolation ?
Posted by swindmill on Thu, 05 May 2011 17:07:18 GMT
View Forum Message <> Reply to Message

Are you using ioprio for your containers?

Also, have you considered ionice'ing the mysql process that's causing your grief? This really isn't different from mysql on a standard linux host causing disk I/O contention and killing performance for other processes.

Subject: Re: OpenVZ IO Isolation ?
Posted by seanfulton on Thu, 05 May 2011 18:28:04 GMT
View Forum Message <> Reply to Message

I'm not using iopro or ionice. And I'm not even talking about nuanced performance improvements.

The I/O problems I see manifest themselves pretty ham-fistedly.

It's pretty simple to reproduce. Just take any big mysqldump file that is compressed, and try to restore it (talking 5G or more):

zcat file.sql.gz | mysql database

in VE1 can cause *all* disk activity in the other VE's to halt. System load peaks as the processes lock in a wait state, and things get really ugly.

Reboot in a different kernel, no other changes, and everything is fine.

What I don't understand is why I see this in one kernel, upgrade and the problem goes away. Then a few months (and kernel upgrades) later after I've forgotten about it, the problem re-appears.

sean

Subject: Re: OpenVZ IO Isolation ?
Posted by swindmill on Thu, 05 May 2011 18:37:09 GMT
View Forum Message <> Reply to Message

And this is significantly different than running a similar process on similar hardware without OpenVZ being involved?

Subject: Re: OpenVZ IO Isolation ?
Posted by seanfulton on Thu, 05 May 2011 18:46:40 GMT
View Forum Message <> Reply to Message

Absolutely! I know of no Linux system that will block *all* disk I/O from all other processes until one process finishes it's task. A DOS system, maybe. But Linux has a scheduler that hops back and forth between processes to give everybody a little slice of heaven from time to time. That's where you would use ionice to determine how big a slice each player gets.

No, this is really odd because it just shuts down everything until the MySQL restore is either killed or finished.

First post I had on this was a bug report and I thought it was related to an NFS file transfer, because it cropped up on the HN while doing a backup onto an NFS volume. In the year's since, I've seen it reappear over and over again, but since most of the "heavy" stuff is being done within the VE's, I haven't noticed whether it happens on the HN itself.

sean

Subject: Re: OpenVZ IO Isolation ?
Posted by andre on Tue, 30 Aug 2011 00:58:26 GMT
View Forum Message <> Reply to Message

Hello,

Do you know if there was any improvement on the OVZ RHEL6 kernel when compared to the

OVZ RHEL5 kernel regarding IO isolation?

Sometimes it just looks like as if there is no IO guarantee.

Thanks!

Subject: Re: OpenVZ IO Isolation ?
Posted by seanfulton on Tue, 30 Aug 2011 15:26:43 GMT
View Forum Message <> Reply to Message

There is a big improvement in RHEL6. I was able to do a test where I had a large database that was being built in one VE by doing:

zcat database.sql | mysql database

I ran it on an Intel i7 machine with 24G of RAM, RHEL5, and it would lock up IO in the other VEs (they would just wait). I then moved the VE to a machine with identical hardware running RHEL6 (scientific Linux), did a repeat, and there was no impact on the other VEs. You couldn't even tell it was running.

I have other issues with RHEL 6 right now, specifically a strange routing issue, but in this respect, it is a big improvement.

sean

Subject: Re: OpenVZ IO Isolation ?
Posted by andre on Wed, 31 Aug 2011 20:27:29 GMT
View Forum Message <> Reply to Message

Hello Sean,

Gonna give it a try.

Thanks for your feedback.

Subject: Re: OpenVZ IO Isolation ?
Posted by mustardman on Fri, 02 Sep 2011 17:04:03 GMT
View Forum Message <> Reply to Message

Why are you not considering trying ioprio as others have suggested? As far as I can determine,

no ioprio means all containers get the same by default and any container can hog all I/O if it wanted to.
http://wiki.openvz.org/I/O_priorities_for_containers

Subject: Re: OpenVZ IO Isolation ?
Posted by seanfulton on Fri, 02 Sep 2011 18:06:41 GMT
View Forum Message <> Reply to Message

I think the VZ containers should prevent activity inside one container from stopping IO on all other containers. Otherwise, what is the point of containers? For that matter, if the Linux OS can't prevent one task from robbing IO from all other tasks on the system, what is the point of a multi-tasking, multi-user OS?

I think this is a defect, and am not sure why people are so quick to say it is normal Linux or OpenVZ behavior.

Again, the purpose of a multitasking OS to *share* resources. The purpose of OpenVZ is to *share* a server. If a normal, non-privileged user inside one container can launch a non-privileged task that blocks all IO from other containers, including the HN, I think that is a defect.

I have done a bit of research on this and it appears to be a bug in the 2.6.18 kernel source, likely in the x86_64 branch. I've seen many reports from non-OpenVZ users who have problems like this but they seem to resolve themselves through driver updates, firmware updates to the RAID subsystem they are using, etc. Users who are not using a RAID subsystem seem to continue to have the problem. But all of them start with the 2.6.18 kernel, x86_64 version.

Also, this is not a problem in the RHEL6 releases. I have tested by moving a VE from a RHEL5 machine where the problem is demonstrated to a RHEL6 machine and the problem goes away.

sean

Subject: Re: OpenVZ IO Isolation ?
Posted by seanfulton on Fri, 02 Sep 2011 18:11:07 GMT
View Forum Message <> Reply to Message

Also, the link you point to talks about setting the IO priority for a container. All of our containers are at the default (4), meaning they all have EQUAL IO priority, and so IO should be equally shared by all of the containers. But it is not. It gets hogged by a single container, which has equal IO priority as all of the other containers.

Hence, I think it is a defect. The IO Priority setting is not being honored.

sean

Subject: Re: OpenVZ IO Isolation ?
Posted by mustardman on Fri, 02 Sep 2011 20:37:17 GMT
View Forum Message <> Reply to Message

Defect or not, why not try set the offending container to ioprio 3 and/or the other containers to ioprio 5 and see what happens.

I'm kind of surprised you have not tried this yet. I'm not saying it will solve your problem. I'm just saying that it wouldn't hurt to try even if you don't think you should have to do that under normal circumstances.

I have had some I/O problems in the past when certain containers are doing certain things and ioprio has helped me. Saying you don't have that problem under RHEL6 is not conclusive imho. There could be other reasons for that.

Subject: Re: OpenVZ IO Isolation ?
Posted by andre on Tue, 06 Sep 2011 02:32:25 GMT
View Forum Message <> Reply to Message

Whenever I experienced this issue, I played with IOPRIO (giving less priority to the io leecher and more priority to the rest) without success. Even if it worked, I do agree with seanfulton, "isolation" means that one container should not be able to affect the guaranteed resources of the other container:

Equal share of IO on a server with 2 containers should mean that each of them have 50% io guarantee and if one of them is able to make the other one unusable (100% iowait) there is no io isolation at all.

I still have not had the opportunity to confirm if RHEL6 kernel fixed this but based on sean report I believe it is ok !

Subject: Re: OpenVZ IO Isolation ?
Posted by raver119 on Tue, 13 Sep 2011 00:10:35 GMT
View Forum Message <> Reply to Message

I had the same issues on debian squeeze with apt-packaged openvz kernal.

1 any process, on VE0 or any VEx could lock all IO on quite powerful server.

After upgrading to RHEL6 2.6.32-042stab035.1 problem is gone.