Subject: Making a VE use a particular cpu
Posted by arghbis on Mon, 22 Dec 2008 09:38:45 GMT
View Forum Message <> Reply to Message

Hello everybody,

is it possible to limit a VE to a particular cpu or set of cpus?

I've a server with 2 Intel I7 quad-core cpus and i'd like to split the cores among the different VEs.

thanks for your help

Subject: Re: Making a VE use a particular cpu
Posted by maratrus on Mon, 22 Dec 2008 13:37:52 GMT
View Forum Message <> Reply to Message

Hello,

you cannot bind your VE to physical CPU but you can limit the number of CPUS inside VE and cpu time.
(man vzctl (--cpus parameters and --cpulimit))

Subject: Re: Making a VE use a particular cpu
Posted by arghbis on Mon, 22 Dec 2008 13:43:12 GMT
View Forum Message <> Reply to Message

thanks for the answer.

I already know the option --cpus, but it does not enforce the use of "separated" cpus. For instance, i have 2 VEs that run transcoding programs, each using 4 threads. To optimize the performances, i want these two VE to run on different cpus.

Maybe the kernel already handle this or maybe my idea is stupid?

Subject: Re: Making a VE use a particular cpu
Posted by piavlo on Mon, 22 Dec 2008 23:12:05 GMT
View Forum Message <> Reply to Message

You can do that using cpusets from HN (incase cpusets actually work in openvz kernel) you can put a specific VE init process in it's own cpuset.

You can use cpuset tool for easier administration of cpusets - http://developer.novell.com/wiki/index.php/Cpuset

Subject: Re: Making a VE use a particular cpu
Posted by piavlo on Tue, 23 Dec 2008 00:13:05 GMT
View Forum Message <> Reply to Message

I've just tried it and it works as should.

This is on 8 cpus system:

```
fire-ovz1 ~ # cset set --cpu=6 --set=blah
cset: --> created cpuset "blah"
fire-ovz1 ~ # cset proc --list blah
cset: "blah" cpuset of:         6 cpu, with:     0 tasks running
fire-ovz1 ~ #  cset proc --set=blah --exec vzctl start 103
cset: 1 tasks match criteria
cset: --> last message, executed args into cpuset "/blah", new pid is: 5842
Starting VE ...
VE is mounted
Setting CPU units: 100000
Configure meminfo: 2460180
Configure veth devices: veth103.0
VE start in progress...
fire-ovz1 ~ # cset proc --list blah
cset: "blah" cpuset of:         6 cpu, with:     8 tasks running
  USER       PID  PPID S TASK NAME
  -------- ----- ----- - ---------
  root      5855     1 S init [3]
  root      5936  5855 S /sbin/udevd --daemon
  root      6180  5855 S /sbin/rc default
  root      6190  5855 S /usr/sbin/syslog-ng
  root      6201  5855 S /usr/sbin/atd
  root      6204  6180 S /sbin/runscript /etc/init.d/net.eth0 start
  root      6205  6204 S /bin/sh /lib64/rc/sh/runscript.sh /etc/init.d/net.eth
  root      6264  6205 S dhcpcd -H -C resolv.conf -C ntp.conf -C yp.conf -G -h
fire-ovz1 ~ #
```

Subject: Re: Making a VE use a particular cpu
Posted by arghbis on Tue, 23 Dec 2008 09:29:32 GMT
View Forum Message <> Reply to Message

Thanks a lot for the tips!

i'll try it soon

Subject: Re: Making a VE use a particular cpu

Posted by maratrus on Tue, 23 Dec 2008 14:48:07 GMT
View Forum Message <> Reply to Message

Hello,

as far as I understand (and please correct me if I'm wrong) cpuset changes task_struct->cpu_allowed field according to its rules.
That's mean that we bind the task not to physical but virtual cpus (it is OpenVZ specific feature).
Virtual CPUs are bound to physical CPUs but this connection is not permanent. So, your task could be bound to virtual CPU (vcpu) but actually run on different physical CPUs (pcpu).

We can conduct the following experiment (on the test node!)
Set appropriate sets with cset command as described in previous post.
Run our VE and issue inside that VE the command e.g.
cat /dev/zero > /dev/null &

then on the HN trigger several times:

echo "p" > /proc/sysrq-trigger

and look in dmesg. There should be appear information concerning with the task running on that moment (physical and virtual CPUs numbers also should be shown). Our goal is catch "cat" process inside VE running on different physical CPUs.

Subject: Re: Making a VE use a particular cpu
Posted by piavlo on Tue, 23 Dec 2008 16:19:22 GMT
View Forum Message <> Reply to Message

maratrus wrote on Tue, 23 December 2008 16:48Hello,

as far as I understand (and please correct me if I'm wrong) cpuset changes task_struct->cpu_allowed field according to its rules.
That's mean that we bind the task not to physical but virtual cpus (it is OpenVZ specific feature).

Even if I do it in HN = VE0?

If i have 8 pcpus does it means that there are also exactly 8 vcpus?

Can you elaborate (or point me to documentstion) on why
this is needed in openvz kernel. Since using cpu affinity in openvz kernels is also usefull feature, and I don't understand why this vcpu & pcpu thing needed in VE0.

Quote:
Virtual CPUs are bound to physical CPUs but this connection is not permanent. So, your task could be bound to virtual CPU (vcpu) but actually run on different physical CPUs (pcpu).

You mean that the cpus i see inside HN , for example using htop
are also virtual, so if i monitnor a specific process bound to specific cpu using htop, I won't notice if it migrates to other pcpu(unless if it migrates to other vcpu)? While on vanilla kernel i can see if the process migrates to other cpu.

What about mpstat command, do i see vcpus with it or pcpus?
Is there a cpu load monitoring command what will show me pcpus?

From what it looks the sys entries at
/sys/devices/system/cpu/cpu?/ show me pcpus
since putting offline cpu with:
echo 0 > /sys/devices/system/cpu/cpu3/online

Removes cpu 3 from /proc/cpuinfo , which I hope shows me pcpus!

Quote:
We can conduct the following experiment (on the test node!)
Set appropriate sets with cset command as described in previous post.
Run our VE and issue inside that VE the command e.g.
cat /dev/zero > /dev/null &

then on the HN trigger several times:

echo "p" > /proc/sysrq-trigger

and look in dmesg. There should be appear information concerning with the task running on that moment (physical and virtual CPUs numbers also should be shown). Our goal is catch "cat" process inside VE running on different physical CPUs.

Is cpu 6 below a vcpu or pcpu?

---------- IPI show regs ----------
CPU 6:
Modules linked in: simfs vznetdev vzethdev vzrst vzcpt tun vzdquota vzmon vzdev i2c_i801 i2c_core button
Pid: 32363, comm: cat Not tainted 2.6.24-openvz-006-r5 #1 ovz006
RIP: 0010:[<ffffffff80381a06>] [<ffffffff80381a06>] __clear_user+0x16/0x40
RSP: 0018:ffff81023df8bee0 EFLAGS: 00000212
RAX: 0000000000000008 RBX: 0000000000001000 RCX: 00000000000000ef
RDX: 0000000000000000 RSI: 0000000000000000 RDI: 000000000060b888
RBP: 000000000060b000 R08: 00002acf45daf070 R09: ffff8101a6c73400
R10: 0000000000000000 R11: 0000000000000246 R12: 0000000000001000
R13: 0000000000000000 R14: 0000000000001000 R15: 0000000000000000
FS: 00002acf45db36f0(0000) GS:ffff81043fc09c00(0000) knlGS:0000000000000000
CS: 0010 DS: 0000 ES: 0000 CR0: 000000008005003b

CR2: 0000000000690000 CR3: 0000000435a2d000 CR4: 00000000000006e0
DR0: 0000000000000000 DR1: 0000000000000000 DR2: 0000000000000000
DR3: 0000000000000000 DR6: 00000000ffff0ff0 DR7: 0000000000000400

I've made about 20 tries with the sysrq thing and the 32326 task was always running on cpu 6 according to dmesg.

So to force migration i put cpu 6 offline
echo 0 > /sys/devices/system/cpu/cpu6/online

In htop i saw that the process migrated to cpu0 (or vcpu0 more correctly).

The next time i issued sysrq the machine froze Probably sysrq is a bad thing to do with offline cpus.

AFAIU from what I've seen by default vcpu0 is bound to pcpu0
vcpu1 to pcpu1 and etc...
And I saw that vcpus default mapping was NOT altered , during my sysrq test.(can you suggest on which loads this mapping will be altered?) So cpu affinity might probably work as should in 99%
of the time in openvz

Subject: Re: Making a VE use a particular cpu
Posted by maratrus on Wed, 24 Dec 2008 13:17:59 GMT
View Forum Message <> Reply to Message

Hello,

as far as I see your node is running 2.6.24 kernel.
Sorry for misleading but my previous message concerns with 2.6.18 kernels.

2.6.24 kernel doesn't contain all this stuff.

Quote:
Even if I do it in HN = VE0?

If i have 8 pcpus does it means that there are also exactly 8 vcpus?

Can you elaborate (or point me to documentstion) on why
this is needed in openvz kernel. Since using cpu affinity in openvz kernels is also usefull feature, and I don't understand why this vcpu & pcpu thing needed in VE0.

The main idea of all this stuff is providing fair scheduler.
If we are going to run a process of particular VE we cannot guarantee that there are processes of

that VE in a given CPU runqueue.

Subject: Re: Making a VE use a particular cpu
Posted by piavlo on Wed, 24 Dec 2008 14:29:08 GMT
View Forum Message <> Reply to Message

maratrus wrote on Wed, 24 December 2008 15:17Hello,

as far as I see your node is running 2.6.24 kernel.
Sorry for misleading but my previous message concerns with 2.6.18 kernels.

2.6.24 kernel doesn't contain all this stuff.

So in 2.6.24 cpusets work as in vanilla kernel?

Quote:
The main idea of all this stuff is providing fair scheduler.
If we are going to run a process of particular VE we cannot guarantee that there are processes of that VE in a given CPU runqueue.

Frankly, I don't quite undertand:
Is this Virtual CPU runqueue or Physical CPU runqueue?
Scheduler can move proccess from one runqueue to another anyway, so how is it differenet in fairscheduling.

And why this does not happen in 2.6.24 ovz?
Is it since openvz fairscheduler based on new default linux CFS scheduler? If so then why this does not happen in CFS?

And if i'm on 2.6.18 ovz and use mpstat to monitor cpu load
which cpus do i see vcpus or pcpus?

Thanks

Subject: Re: Making a VE use a particular cpu
Posted by maratrus on Wed, 24 Dec 2008 17:36:56 GMT
View Forum Message <> Reply to Message

Hello,

we have to provide an ability to divide processes into several groups (e.g. if VE1 has only one process and VE2 has hundred of them the last hundred processes in VE2 should get the same CPU time (in sum) as a single one in VE1).

O(1) scheduler make OpenVZ developers to create an additional layer to provide such functionality.

Using CFS scheduler allowed to get rid of some things because it was possible to use hierarchy without building a massive additional layer.

Quote:
And if i'm on 2.6.18 ovz and use mpstat to monitor cpu load
which cpus do i see vcpus or pcpus?

Information from proc concerning cpus should be represent as vcpu.

Subject: Re: Making a VE use a particular cpu
Posted by piavlo on Wed, 24 Dec 2008 18:31:06 GMT
View Forum Message <> Reply to Message

maratrus wrote on Wed, 24 December 2008 19:36
Information from proc concerning cpus should be represent as vcpu.

And information under /sys/devices/system/cpu/ is pcpu?

Thanks for the explanations

Subject: Re: Making a VE use a particular cpu
Posted by maratrus on Thu, 25 Dec 2008 07:28:44 GMT
View Forum Message <> Reply to Message

Quote:
And information under /sys/devices/system/cpu/ is pcpu?

Information under that directory is about physical CPUs.
BTW, that information is not available from inside the VE.

VCPU layer is not CPU emulation layer - it is scheduler layer.
Roughly speaking it is CPU mask virtualization.

Subject: Re: Making a VE use a particular cpu
Posted by piavlo on Thu, 25 Dec 2008 13:39:49 GMT
View Forum Message <> Reply to Message

maratrus wrote on Thu, 25 December 2008 09:28

Information under that directory is about physical CPUs.
BTW, that information is not available from inside the VE.

VCPU layer is not CPU emulation layer - it is scheduler layer.
Roughly speaking it is CPU mask virtualization.

Ok Thanks

Subject: Re: Making a VE use a particular cpu
Posted by maratrus on Fri, 26 Dec 2008 07:43:46 GMT
View Forum Message <> Reply to Message

Quote:

Russian language!
