Subject: Problem checkpointing OpenVZ VMs under Xen
Posted by giles on Thu, 07 Jul 2011 18:23:15 GMT
View Forum Message <> Reply to Message

Hi there,

First of all, thanks to OpenVZ's creators for producing such an useful-looking tool!

We're building an application that needs to run under Xen (for development) and on Amazon EC2 (live). Our users need to run multiple processes each, and be isolated from each other. So far we've been using chroot to enforce filesystem isolation, but we need to improve that so that they can't see each others' process, and we'd like to be able to checkpoint/restore their stuff so that we can migrate everything from machine to machine. OpenVZ looks like the perfect solution.

I've been experimenting with running the default Debian Squeeze amd64 OpenVZ package, linux-image-openvz-amd64, which supplies an OpenVZ-enabled kernel version 2.6.32-5. The package maintainer says (correctly) that it works as a Xen domU kernel. vzctl reports that it is version 3.0.24. I was able to boot a Xen image with this kernel running, and create and start an OpenVZ VM within it. So, for those who are keeping track, I had computer running Xen that had a domU with the OpenVZ kernel, which was itself hosting an OpenVZ virtual machine. Kind of a virtualization Inception. I'm very impressed at how easily it all worked.

Unfortunately checkpointing doesn't seem to work so well. Here's the output I got:

root@localhost:/var/lib/vz/template/cache# vzctl chkpnt 101
Setting up checkpoint...
suspend...
dump...
Can not dump container: Invalid argument
Message from syslogd@localhost at Jul 7 17:55:31 ...
kernel:[582.902928] general protection fault: 0000 [#1] SMP
Message from syslogd@localhost at Jul 7 17:55:31 ...
kernel:[582.902946] last sysfs file: /sys/devices/virtual/net/lo/operstate
Message from syslogd@localhost at Jul 7 17:55:31 ...
kernel:[582.903178] Stack:
Message from syslogd@localhost at Jul 7 17:55:31 ...
kernel:[582.903233] Call Trace:
Message from syslogd@localhost at Jul 7 17:55:31 ...
kernel:[582.903304] Code: 8b 4c 24 10 4c 8b 44 24 18 48 8b 44 24 20 48 8b 4c 24 28 48 8b 54 24 30 48 8b
74 24 38 48 8b 7c 24 40 48 83 c4 48 48 83 c4 30 c3 <6a> 00 48 89 f8 48 89 f7 ff d0 48 89 c7 e8 ec c2 f4 e0
00 00 48
Error: iptables-save terminated
Checkpointing failed
root@localhost:/var/lib/vz/template/cache#

/var/log/syslog has more details:

Jul 7 17:55:31 localhost kernel: [582.902928] general protection fault: 0000 [#1] SMP
Jul 7 17:55:31 localhost kernel: [582.902946] last sysfs file: /sys/devices/virtual/net/lo/operstate
Jul 7 17:55:31 localhost kernel: [582.902954] CPU 0
Jul 7 17:55:31 localhost kernel: [582.902961] Modules linked in: vzethdev vznetdev simfs vzrst vzcpt vzd
quota vzmon vzdev xt_tcpudp xt_length xt_hl xt_tcpmss xt_TCPMSS iptable_mangle iptable_filter xt_multiport
xt_limit xt_dscp ipt_REJECT ip_tables x_tables vzevent xfs exportfs evdev ext3 jbd mbcache xen_netfront xen
_blkfront
Jul 7 17:55:31 localhost kernel: [582.903037] Pid: 1496, comm: vzctl Not tainted 2.6.32-5-openvz-amd64 #
1 feoktistov
Jul 7 17:55:31 localhost kernel: [582.903046] RIP: 0010:[<ffffffffa0105744>] [<ffffffffa0105744>] child
_rip+0x0/0x14 [vzcpt]
Jul 7 17:55:31 localhost kernel: [582.903063] RSP: 0003:ffff88000249bf58 EFLAGS: 00000200
Jul 7 17:55:31 localhost kernel: [582.903071] RAX: 0000000000000000 RBX: 00000000ffffffff0 RCX: 000000000
0000000
Jul 7 17:55:31 localhost kernel: [582.903081] RDX: 0000000000004011 RSI: ffff880009407ce8 RDI: ffffffffa
01035ec
Jul 7 17:55:31 localhost kernel: [582.903090] RBP: 0000000000000002 R08: 0000000000000040 R09: 000080d00
9407fd8
Jul 7 17:55:31 localhost kernel: [582.903099] R10: 0000000000000000 R11: ffffffff8122c942 R12: 000000000
0000000
Jul 7 17:55:31 localhost kernel: [582.903108] R13: 0000000000004011 R14: ffff880009407ce8 R15: ffffffffa
01035ec
Jul 7 17:55:31 localhost kernel: [582.903120] FS: 00007feadc849700(0000) GS:ffff88000311b000(0000) knlG
S:0000000000000000
Jul 7 17:55:31 localhost kernel: [582.903131] CS: e033 DS: 0000 ES: 0000 CR0: 000000008005003b
Jul 7 17:55:31 localhost kernel: [582.903139] CR2: 00007feadbd43760 CR3: 0000000009ba6000 CR4: 000000000
0002660
Jul 7 17:55:31 localhost kernel: [582.903149] DR0: 0000000000000000 DR1: 0000000000000000 DR2: 000000000
0000000
Jul 7 17:55:31 localhost kernel: [582.903158] DR3: 0000000000000000 DR6: 00000000ffff0ff0

DR7: 000000000
0000400
Jul 7 17:55:31 localhost kernel: [582.903168] Process vzctl (pid: 1496, veid=0, threadinfo ffff88000249a
000, task ffff88000fe2d000)
Jul 7 17:55:31 localhost kernel: [582.903178] Stack:
Jul 7 17:55:31 localhost kernel: [582.903183] ffffffffa01035ec ffff880009407ce8 0000000000004011 000000
0000000000
Jul 7 17:55:31 localhost kernel: [582.903197] <0> ffffffff81010e31 ffffffff810115dd ffffffff8122c942 000
0000000000000
Jul 7 17:55:31 localhost kernel: [582.903214] <0> 000080d009407fd8 0000000000000040 0000000000000000 000
0000000000000
Jul 7 17:55:31 localhost kernel: [582.903233] Call Trace:
Jul 7 17:55:31 localhost kernel: [582.903242] [<ffffffffa01035ec>] ? dumpfn+0x0/0x106 [vzcpt]
Jul 7 17:55:31 localhost kernel: [582.903254] [<ffffffff81010e31>] ? int_ret_from_sys_call+0x7/0x1b
Jul 7 17:55:31 localhost kernel: [582.903264] [<ffffffff810115dd>] ? retint_restore_args+0x5/0x6
Jul 7 17:55:31 localhost kernel: [582.903276] [<ffffffff8122c942>] ? sock_destroy_inode+0x0/0x10
Jul 7 17:55:31 localhost kernel: [582.903285] [<ffffffff8122c942>] ? sock_destroy_inode+0x0/0x10
Jul 7 17:55:31 localhost kernel: [582.903296] [<ffffffffa0105744>] ? child_rip+0x0/0x14 [vzcpt]
Jul 7 17:55:31 localhost kernel: [582.903304] Code: 8b 4c 24 10 4c 8b 44 24 18 48 8b 44 24 20 48 8b 4c 2
4 28 48 8b 54 24 30 48 8b 74 24 38 48 8b 7c 24 40 48 83 c4 48 48 83 c4 30 c3 <6a> 00 48 89 f8 48 89 f7 ff d
0 48 89 c7 e8 ec c2 f4 e0 00 00 48
Jul 7 17:55:31 localhost kernel: [582.903418] RIP [<ffffffffa0105744>] child_rip+0x0/0x14 [vzcpt]
Jul 7 17:55:31 localhost kernel: [582.903429] RSP <ffff88000249bf58>
Jul 7 17:55:31 localhost kernel: [582.903437] ---[end trace ff752e26828c9f92]---
Jul 7 17:55:31 localhost kernel: [582.903674] CPT ERR: ffff880001953000,101 :iptables-save terminated

Is this an unavoidable problem when running under Xen? Or perhaps a bug? Or something that I can fix with appropriate configuration? Or am I just being stupid? I really hope it's one of the latter two :-)

Thanks in advance for any help.

Giles

Subject: Re: Problem checkpointing OpenVZ VMs under Xen
Posted by kir on Mon, 11 Jul 2011 09:01:43 GMT
View Forum Message <> Reply to Message

I suggesting retrying with our RHEL6-based kernel, which is way better than plain 2.6.32.

Subject: Re: Problem checkpointing OpenVZ VMs under Xen
Posted by giles on Tue, 12 Jul 2011 10:19:15 GMT
View Forum Message <> Reply to Message

Thanks, I'll give that a try. I'm using Debian, though -- is that likely to cause problems?

Subject: Re: Problem checkpointing OpenVZ VMs under Xen
Posted by giles on Tue, 12 Jul 2011 12:38:21 GMT
View Forum Message <> Reply to Message

I tried using the RHEL6 kernel. Installed the vzctl, vzquota and vzdump packages from the normal Debian packages, then downloaded the vzkernel-2.6.32-042stab020.1.x86_64.rpm, used alien to convert it to a .deb, installed it, then built an initrd.img using update-initramfs.

I was able to boot a Xen domU VM using the new kernel, and again creating and starting OpenVZ virtual machines within it worked fine. However, checkpointing worked even less well than it did before; it caused a kernel panic which made the Xen VM reboot.

The domU's console showed this:

[1069.573718] general protection fault: 0000 [#1] SMP
[1069.573737] last sysfs file: /sys/devices/virtual/net/lo/operstate
[1069.573746] CPU 0
[1069.573750] Modules linked in: vzethdev vznetdev simfs vzrst nf_nat nf_conntrack_ipv4 nf_conntrack nf_defrag_ipv4 vzcpt nfs lockd fscache nfs_acl auth_rpcgss sunrpc vzdquota vzmon vzdev xt_length xt_hl xt_tcpmss xt_TCPMSS iptable_mangle iptable_filter xt_multiport xt_limit xt_dscp ipt_REJECT ip_tables ipv6 vzevent xfs exportfs ext3 jbd mbcache xen_netfront xen_blkfront [last unloaded: scsi_wait_scan]
[1069.573845]
[1069.573850] Modules linked in: vzethdev vznetdev simfs vzrst nf_nat nf_conntrack_ipv4 nf_conntrack nf_defrag_ipv4 vzcpt nfs lockd fscache nfs_acl auth_rpcgss sunrpc vzdquota vzmon vzdev xt_length xt_hl xt_tcpmss xt_TCPMSS iptable_mangle iptable_filter xt_multiport xt_limit xt_dscp ipt_REJECT ip_tables ipv6 vzevent xfs exportfs ext3 jbd mbcache xen_netfront xen_blkfront [last unloaded: scsi_wait_scan]
[1069.573944] Pid: 1584, comm: vzctl Not tainted 2.6.32-042stab020.1 #1 042stab020
[1069.573954] RIP: 0010:[<ffffffffa0303924>] [<ffffffffa0303924>] child_rip+0x0/0x13a [vzcpt]
[1069.573972] RSP: 0003:ffff880003a81f58 EFLAGS: 00000200
[1069.573980] RAX: 0000000000000000 RBX: ffffffffa02f6b00 RCX: 0000000000000000
[1069.573990] RDX: 0000000000004011 RSI: ffff880003abfba8 RDI: ffffffffa02f6b00

[1069.574000] RBP: ffff880003abfb78 R08: ffff880003814210 R09: 0000000000000000
[1069.574008] R10: 0000000000000007 R11: 0000000000000000 R12: ffff880003abfba8
[1069.574008] R13: 0000000000004011 R14: 0000000000000002 R15: ffff880003abfd00
[1069.574008] FS: 00007f6319fc2700(0000) GS:ffff8800041a0000(0000) knlGS:0000000000000000
[1069.574008] CS: e033 DS: 0000 ES: 0000 CR0: 000000008005003b
[1069.574008] CR2: 00007f63194bc760 CR3: 00000000039cc000 CR4: 0000000000002660
[1069.574008] DR0: 0000000000000000 DR1: 0000000000000000 DR2: 0000000000000000
[1069.574008] DR3: 0000000000000000 DR6: 00000000ffff0ff0 DR7: 0000000000000400
[1069.574008] Process vzctl (pid: 1584, veid=0, threadinfo ffff880003a80000, task ffff880003f8d300)
[1069.574008] Stack:
[1069.574008] ffff880003abfd00 0000000000000002 0000000000004011 ffff880003abfba8
[1069.574008] <0> ffffffff8100b463 ffffffff8100bc1d 0000000000000000 0000000000000007
[1069.574008] <0> 0000000000000000 ffff880003814210 0000000000000000 0000000000000000
[1069.574008] Call Trace:
[1069.574008] [<ffffffff8100b463>] ? int_ret_from_sys_call+0x7/0x1b
[1069.574008] [<ffffffff8100bc1d>] ? retint_restore_args+0x5/0x6
[1069.574008] [<ffffffffa0303924>] ? child_rip+0x0/0x13a [vzcpt]
[1069.574008] Code: 8b 4c 24 10 4c 8b 44 24 18 48 8b 44 24 20 48 8b 4c 24 28 48 8b 54 24 30 48 8b 74 24 38 48 8b 7c 24 40 48 83 c4 48 48 83 c4 30 c3 <6a> 00 48 89 f8 48 89 f7 ff d0 48 89 c7 e8 fa 99 d6 e0 00 00 c7
[1069.574008] RIP [<ffffffffa0303924>] child_rip+0x0/0x13a [vzcpt]
[1069.574008] RSP <ffff880003a81f58>
[1069.574008] ---[end trace 5dcfcbade405cc62]---
[1069.574008] Kernel panic - not syncing: Fatal exception
[1069.574008] Pid: 1584, comm: vzctl Tainted: G   D   ----------------   2.6.32-042stab020.1 #1
[1069.574008] Call Trace:
[1069.574008] [<ffffffff814bd79b>] ? panic+0x78/0x143
[1069.574008] [<ffffffff814c091c>] ? _spin_unlock_irqrestore+0x1c/0x20
[1069.574008] [<ffffffff814c1a94>] ? oops_end+0xe4/0x100
[1069.574008] [<ffffffff8100f3eb>] ? die+0x5b/0x90
[1069.574008] [<ffffffff814c15d2>] ? do_general_protection+0x152/0x160
[1069.574008] [<ffffffff81007a7f>] ? xen_restore_fl_direct_end+0x0/0x1
[1069.574008] [<ffffffff814c0da5>] ? general_protection+0x25/0x30
[1069.574008] [<ffffffffa02f6b00>] ? dumptmpfs+0x0/0x1e0 [vzcpt]
[1069.574008] [<ffffffffa02f6b00>] ? dumptmpfs+0x0/0x1e0 [vzcpt]
[1069.574008] [<ffffffffa0303924>] ? child_rip+0x0/0x13a [vzcpt]
[1069.574008] [<ffffffff8100b0f3>] ? ret_from_fork_tail+0x0/0x17
[1069.574008] [<ffffffff8100b463>] ? int_ret_from_sys_call+0x7/0x1b
[1069.574008] [<ffffffff8100bc1d>] ? retint_restore_args+0x5/0x6
[1069.574008] [<ffffffffa0303924>] ? child_rip+0x0/0x13a [vzcpt]

Again, any help would be much appreciated!