Ever-rising load on Debian jessie + DRBD8 + LXC host pairs

The Problem

Setup description

We are currently operating 5 pairs of pizza boxes with the following setup:

- Debian jessie as Host system
- DRBD to mirror partitions (LVM volumes), one partition for each virtual guest
- LXC (replacing OpenVZ in early 2016) for running kernel-virtualized Linux servers (mostly Debian squeeze - jessie, a few CentOS 6 - 7)

In addition we use VLANs to separate World LAN, 2 DMZs, Admin LAN. DRBD traffic uses a dedicated connection between the eth1's of either host of each pair (i.e. not connected to a switch).

Furthermore we use software bridging, using the hosts' `br0`/`br1`/... for `lxc.network.link`. Maybe we should use the hosts' `eth0`/`eth0.x`/... instead? Not sure.

Problem: Occasional LXC guest gets unreachable and load starts to rise forever

Several times a single LXC guest became unreachable and the host's "load" started to rise slowly but continously up to ~1000 (!) and more.

Fortunately in every single occation the host still allowed to login via `ssh`.

But, also in every single case, stopping the LXC guest affected always got the "lxc stop" stuck too. Sometimes we managed to stop other LXC guests, but usually those stop attempts got stuck too.

Unfortunately the only one way out of this mess it the Reset button (via BMCs' Web interface, IPMI, or physically on-site). Ouch. WTF.

Occurences

Case 1

Debian kernel 4.4 (jessie-backports), triggered by heavy I/O probably igniting the old disk write overcommitment bomb ("blocked for more than 120 seconds" is a classic)::

Kernel problem case 1

```
Jul 25 11:03:44 host3 kernel: [10002093.973317] drbd www-live-jessie: peer( Secondary -> Unknown ) conn(
Connected -> Disconnecting ) pdsk( UpToDate -> DUnknown )
Jul 25 11:03:44 host3 kernel: [10002093.973370] block drbd10: new current UUID D6ABC588F9BB7CD9:
FAA0156944BBECD1:EB0D0FA7736CECC4:0000000000000004
Jul 25 11:03:44 host3 kernel: [10002093.973436] drbd www-live-jessie: asender terminated
Jul 25 11:03:44 host3 kernel: [10002093.973440] drbd www-live-jessie: Terminating drbd_a_www-live
Jul 25 11:03:44 host3 kernel: [10002094.010958] drbd www-live-jessie: Connection closed
Jul 25 11:03:44 host3 kernel: [10002094.011142] drbd www-live-jessie: receiver terminated
Jul 25 11:04:14 host3 kernel: [10002124.533457] INFO: task jbd2/drbd5-8:28127 blocked for more than 120 seconds.
Jul 25 11:04:14 host3 kernel: [10002124.535804] "echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables
this message.
Jul 25 11:04:14 host3 kernel: [10002124.536912]  ffff880c093f8e40 ffff880c08f1e200 ffff880c079fc000
ffff880c079fbe50
Jul 25 11:04:14 host3 kernel: [10002124.536920]  ffffffff8158e971 ffff880c05abc8b8 ffffffffa014f272
0000000000015d80
Jul 25 11:04:14 host3 kernel: [10002124.536936]  [<ffffffff8158e971>] ? schedule+0x31/0x80
Jul 25 11:04:14 host3 kernel: [10002124.536959]  [<ffffffff810abed5>] ? put_prev_entity+0x35/0x710
Jul 25 11:04:14 host3 kernel: [10002124.536971]  [<ffffffff810dc2c9>] ? try_to_del_timer_sync+0x59/0x80
Jul 25 11:04:14 host3 kernel: [10002124.536983]  [<ffffffff810b7280>] ? wait_woken+0x90/0x90
Jul 25 11:04:14 host3 kernel: [10002124.536998]  [<ffffffff81095baf>] ? kthread+0xdf/0x100
Jul 25 11:04:14 host3 kernel: [10002124.537009]  [<ffffffff81592adf>] ? ret_from_fork+0x3f/0x70
Jul 25 11:04:14 host3 kernel: [10002124.537019] INFO: task rs:main Q:Reg:14895 blocked for more than 120
seconds.
Jul 25 11:04:14 host3 kernel: [10002124.540380]  [<ffffffffa014c31f>] ? add_transaction_credits+0x21f/0x2a0
[jbd2]
Jul 25 11:04:14 host3 kernel: [10002124.540392]  [<ffffffffa014c4ff>] ? start_this_handle+0x10f/0x420 [jbd2]
Jul 25 11:04:14 host3 kernel: [10002124.540403]  [<ffffffffa014cb79>] ? jbd2__journal_start+0xe9/0x1f0 [jbd2]
Jul 25 11:04:14 host3 kernel: [10002124.540439]  [<ffffffff8120608a>] ? __mark_inode_dirty+0x17a/0x370
Jul 25 11:04:14 host3 kernel: [10002124.540449]  [<ffffffff811f337d>] ? file_update_time+0xbd/0x110
Jul 25 11:04:14 host3 kernel: [10002124.540471]  [<ffffffffa027a07f>] ? ext4_file_write_iter+0x21f/0x450 [ext4]
Jul 25 11:04:14 host3 kernel: [10002124.540481]  [<ffffffff811e177e>] ? pipe_write+0x2fe/0x410
```

```
Jul 25 11:04:14 host3 kernel: [10002124.540492]  [<ffffffff811d8ec4>] ? new_sync_write+0xa4/0xf0
Jul 25 11:04:14 host3 kernel: [10002124.540499]  [<ffffffff811da2c2>] ? SyS_write+0x52/0xc0
Jul 25 11:04:14 host3 kernel: [10002124.540562] INFO: task mongod:34929 blocked for more than 120 seconds.
Jul 25 11:04:14 host3 kernel: [10002124.544201]  [<ffffffff81592736>] ? system_call_fast_compare_end+0xc/0x6b
Jul 25 11:04:14 host3 kernel: [10002124.545455]       Not tainted 4.4.0-0.bpo.1-amd64 #1
Jul 25 11:04:14 host3 kernel: [10002124.546557] "echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables
this message.
Jul 25 11:04:14 host3 kernel: [10002124.547684]  ffff881517fe2400 ffff8807e7fd8240 ffff880f4de9c000
00000000018c3b50
Jul 25 11:04:14 host3 kernel: [10002124.547691]  ffffffff8158e971 ffff880c05abc800 ffffffffa0153d3f
0000000000000000
Jul 25 11:04:14 host3 kernel: [10002124.547699]  [<ffffffff8158e971>] ? schedule+0x31/0x80
Jul 25 11:04:14 host3 kernel: [10002124.547711]  [<ffffffff810b7280>] ? wait_woken+0x90/0x90
Jul 25 11:04:14 host3 kernel: [10002124.547733]  [<ffffffff8120b3f8>] ? do_fsync+0x38/0x60
Jul 25 11:04:14 host3 kernel: [10002124.547740]  [<ffffffff81592736>] ? system_call_fast_compare_end+0xc/0x6b
Jul 25 11:04:14 host3 kernel: [10002124.548906]       Not tainted 4.4.0-0.bpo.1-amd64 #1
Jul 25 11:04:14 host3 kernel: [10002124.551308]  [<ffffffff81168ed6>] ? add_to_page_cache_lru+0x56/0x90
Jul 25 11:04:14 host3 kernel: [10002124.551340]  [<ffffffff811b68aa>] ? alloc_pages_current+0x8a/0x110
Jul 25 11:04:14 host3 kernel: [10002124.551350]  [<ffffffff8117615e>] ? ondemand_readahead+0xce/0x230
Jul 25 11:04:14 host3 kernel: [10002124.551356]  [<ffffffff81169da0>] ? generic_file_read_iter+0x490/0x600
Jul 25 11:04:14 host3 kernel: [10002124.551366]  [<ffffffff811d8dad>] ? new_sync_read+0x9d/0xe0
Jul 25 11:04:14 host3 kernel: [10002124.551372]  [<ffffffff811da3b6>] ? SyS_pread64+0x86/0xb0
Jul 25 11:04:14 host3 kernel: [10002124.551381] INFO: task mysqld:35224 blocked for more than 120 seconds.
Jul 25 11:04:14 host3 kernel: [10002124.553739] "echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables
this message.
Jul 25 11:04:14 host3 kernel: [10002124.555057]  [<ffffffff811da3b6>] ? SyS_pread64+0x86/0xb0
Jul 25 11:04:14 host3 kernel: [10002124.555080] INFO: task zdd:15555 blocked for more than 120 seconds.
Jul 25 11:04:14 host3 kernel: [10002124.557585] "echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables
this message.
Jul 25 11:04:14 host3 kernel: [10002124.558813] zdd             D ffff880c0fa15d80     0 15555  15553 0x00000100
Jul 25 11:04:14 host3 kernel: [10002124.558820]  ffff88144099c56c ffff8801fd7ae2c0 00000000ffffffff
ffff88144099c570
Jul 25 11:04:14 host3 kernel: [10002124.558826] Call Trace:
Jul 25 11:04:14 host3 kernel: [10002124.558834]  [<ffffffff8158ebfa>] ? schedule_preempt_disabled+0xa/0x10
Jul 25 11:04:14 host3 kernel: [10002124.558840]  [<ffffffff811e371c>] ? unlazy_walk+0xbc/0x150
Jul 25 11:04:14 host3 kernel: [10002124.558847]  [<ffffffff811e6ec7>] ? path_openat+0x517/0x1510
Jul 25 11:04:14 host3 kernel: [10002124.558853]  [<ffffffff811e9581>] ? do_filp_open+0x91/0x100
Jul 25 11:04:14 host3 kernel: [10002124.558863]  [<ffffffff811d860a>] ? do_sys_open+0x13a/0x230
Jul 25 11:04:14 host3 kernel: [10002124.558870] INFO: task zdd:15557 blocked for more than 120 seconds.
Jul 25 11:04:14 host3 kernel: [10002124.561326] "echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables
this message.
Jul 25 11:04:14 host3 kernel: [10002124.562591]  ffff880adfec8e40 ffff880c08f32f80 ffff88024d310000
ffff880c05abc870
Jul 25 11:04:14 host3 kernel: [10002124.562598]  ffffffff8158e971 ffff880c05abc800 ffffffffa014c096
0000000000000000
Jul 25 11:04:14 host3 kernel: [10002124.562605]  [<ffffffff8158e971>] ? schedule+0x31/0x80
Jul 25 11:04:14 host3 kernel: [10002124.562616]  [<ffffffff810b7280>] ? wait_woken+0x90/0x90
Jul 25 11:04:14 host3 kernel: [10002124.562646]  [<ffffffffa014c4ff>] ? start_this_handle+0x10f/0x420 [jbd2]
Jul 25 11:04:14 host3 kernel: [10002124.562681]  [<ffffffffa014cb79>] ? jbd2__journal_start+0xe9/0x1f0 [jbd2]
Jul 25 11:04:14 host3 kernel: [10002124.564016]       Not tainted 4.4.0-0.bpo.1-amd64 #1
Jul 25 11:04:14 host3 kernel: [10002124.565261] "echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables
this message.
Jul 25 11:04:14 host3 kernel: [10002124.566533] zdd             D ffff88180f355d80     0 15560  15553 0x00000100
Jul 25 11:04:14 host3 kernel: [10002124.566540]  ffff880c05abc888 ffff880c05abc824 ffff880cac743e78
ffff880c05abc8a0
Jul 25 11:04:14 host3 kernel: [10002124.566546] Call Trace:
Jul 25 11:04:14 host3 kernel: [10002124.566558]  [<ffffffffa0153d3f>] ? jbd2_log_wait_commit+0x9f/0x120 [jbd2]
Jul 25 11:04:14 host3 kernel: [10002124.566586]  [<ffffffffa0156765>] ? __jbd2_journal_force_commit+0x55/0x90
[jbd2]
Jul 25 11:04:14 host3 kernel: [10002124.567893]       Not tainted 4.4.0-0.bpo.1-amd64 #1
Jul 25 11:04:14 host3 kernel: [10002124.569148] "echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables
this message.
Jul 25 11:04:14 host3 kernel: [10002124.570480] zsd             D ffff88180f375d80     0 29409  15679 0x00000100
Jul 25 11:04:14 host3 kernel: [10002124.570484]  ffff8818052eb0c0 ffff880c08f8a440 ffff880be4314000
ffff880c05abc870
Jul 25 11:04:14 host3 kernel: [10002124.570490]  ffffffff8158e971 ffff880c05abc800 ffffffffa014c096
0000000000000000
Jul 25 11:04:14 host3 kernel: [10002124.570496]  [<ffffffff8158e971>] ? schedule+0x31/0x80
Jul 25 11:04:14 host3 kernel: [10002124.570513]  [<ffffffff810b7280>] ? wait_woken+0x90/0x90
Jul 25 11:04:14 host3 kernel: [10002124.570530]  [<ffffffffa014c31f>] ? add_transaction_credits+0x21f/0x2a0
[jbd2]
```

```
Jul 25 11:04:14 host3 kernel: [10002124.570540]  [<ffffffff8101fa25>] ? sched_clock+0x5/0x10
Jul 25 11:04:14 host3 kernel: [10002124.570548]  [<ffffffffa014c4ff>] ? start_this_handle+0x10f/0x420 [jbd2]
Jul 25 11:04:14 host3 kernel: [10002124.570560]  [<ffffffff811f2c21>] ? inode_init_always+0x101/0x1b0
```

Case 2

Debian kernel 4.6, RIP in or around DRBD module:

Kernel problem case 2

```
Aug  6 05:00:14 host18 kernel: [730487.320583] RIP: 0010:[<ffffffff813201e6>]  [<ffffffff813201e6>]
memcpy_erms+0x6/0x10
Aug  6 05:00:14 host18 kernel: [730487.339579] FS:  0000000000000000(0000) GS:ffff88103fb00000(0000) knlGS:
0000000000000000
Aug  6 05:00:14 host18 kernel: [730487.353038]  000000000000faf0 ffff88203650fc48 0000000000000000
ffffffff81517a12
Aug  6 05:00:14 host18 kernel: [730487.369167]  [<ffffffffc063e119>] ? drbd_send+0xc9/0x1e0 [drbd]
Aug  6 05:00:14 host18 kernel: [730487.387607]  [<ffffffffc063c2f0>] ? drbd_destroy_connection+0xf0/0xf0 [drbd]
Aug  6 05:00:14 host18 kernel: [730487.403314]  RSP <ffff88203650fb40>
```

Case 3

Debian kernel 4.6, RSP in or around DRBD module:

Kernel problem case 3

```
Aug 27 05:25:43 host19 kernel: [2547757.533648] RSP: 0018:ffff8801c93b3b40  EFLAGS: 00010292
Aug 27 05:25:43 host19 kernel: [2547757.579013]  00004000000005b4 00000000000005b4 00000000000008a0
0000000000000800
Aug 27 05:25:43 host19 kernel: [2547757.623329]  [<ffffffffc060a2f0>] ? drbd_destroy_connection+0xf0/0xf0 [drbd]
```

Case 4

Debian kernel 4.6, RSP in or around DRBD module:

Kernel problem case 4

```
Aug 27 04:26:29 host14 kernel: [478357.442244] Modules linked in: pci_stub(E) vboxpci(OE) vboxnetadp(OE)
vboxnetflt(OE) vboxdrv(OE) nfsv3(E) rpcsec_gss_krb5(E) nfsv4(E) dns_resolver(E) tcp_diag(E) inet_diag(E)
ipt_REJECT(E) nf_reject_ipv4(E) nf_log_ipv6(E) ip6t_rt(E) veth(E) drbd(E) ipmi_devintf(E) xt_multiport(E)
nf_log_ipv4(E) nf_log_common(E) xt_LOG(E) xt_limit(E) xt_tcpudp(E) nf_conntrack_ipv4(E) nf_defrag_ipv4(E)
ip6table_filter(E) xt_conntrack(E) xt_state(E) iptable_filter(E) ip_tables(E) nf_conntrack_ipv6(E)
nf_defrag_ipv6(E) nf_conntrack(E) ip6_tables(E) x_tables(E) nfsd(E) auth_rpcgss(E) nfs_acl(E) nfs(E) lockd(E)
grace(E) fscache(E) sunrpc(E) 8021q(E) garp(E) mrp(E) bridge(E) stp(E) llc(E) lru_cache(E) libcrc32c(E)
crc32c_generic(E) x86_pkg_temp_thermal(E) intel_powerclamp(E) coretemp(E) kvm_intel(E) kvm(E) irqbypass(E)<4>
[478357.455870] Hardware name: Thomas-Krenn.AG X9DR3-F/X9DR3-F, BIOS 3.0a 07/31/2013
Aug 27 04:26:29 host14 kernel: [478357.460567] RSP: 0018:ffff8808534f7b40  EFLAGS: 00010292
Aug 27 04:26:29 host14 kernel: [478357.465530] RBP: ffff8808534f7c58 R08: ffff880858d28af0 R09: 0000000000000000
Aug 27 04:26:29 host14 kernel: [478357.470727] FS:  0000000000000000(0000) GS:ffff88085fa00000(0000) knlGS:
0000000000000000
Aug 27 04:26:29 host14 kernel: [478357.476167] Stack:
Aug 27 04:26:29 host14 kernel: [478357.483832] Call Trace:
Aug 27 04:26:29 host14 kernel: [478357.489826]  [<ffffffff814acf80>] ? sock_sendmsg+0x30/0x40
Aug 27 04:26:29 host14 kernel: [478357.498117]  [<ffffffffc07b9efd>] ? w_send_dblock+0x9d/0x1c0 [drbd]
Aug 27 04:26:29 host14 kernel: [478357.506710]  [<ffffffffc07d02f0>] ? drbd_destroy_connection+0xf0/0xf0 [drbd]
Aug 27 04:26:29 host14 kernel: [478357.515569] Code: 90 90 90 90 90 eb 1e 0f 1f 00 48 89 f8 48 89 d1 48 c1 e9
03 83 e2 07 f3 48 a5 89 d1 f3 a4 c3 66 0f 1f 44 00 00 48 89 f8 48 89 d1 <f3> a4 c3 0f 1f 80 00 00 00 00 48 89
f8 48 83 fa 20 72 7e 40 38
Aug 27 04:26:29 host14 kernel: [478357.538514] ---[ end trace 0d23089f3d6f0d23 ]---
```

Case 5

Debian kernel 4.6 with DRBD module 4.8.4-1. DRBD again or RAM ("unable to handle kernel paging request")?

Kernel problem case 5

```
Sep  9 04:26:34 host14 kernel: [1056355.854359] BUG: unable to handle kernel paging request at 0000000000001000
Sep  9 04:26:34 host14 kernel: [1056355.854480] PGD 0
Sep  9 04:26:34 host14 kernel: [1056355.854536] Modules linked in: drbd(OE) ipt_REJECT(E) nf_reject_ipv4(E)
tcp_diag(E) inet_diag(E) nfsv3(E) rpcsec_gss_krb5(E) nfsv4(E) dns_resolver(E) ip6t_rt(E) veth(E) ipmi_devintf
(E) xt_multiport(E) pci_stub(E) vboxpci(OE) vboxnetadp(OE) vboxnetflt(OE) nf_log_ipv4(E) nf_log_common(E) xt_LOG
(E) xt_limit(E) vboxdrv(OE) xt_tcpudp(E) nf_conntrack_ipv4(E) nf_defrag_ipv4(E) xt_conntrack(E) xt_state(E)
ip6table_filter(E) nf_conntrack_ipv6(E) nf_defrag_ipv6(E) nf_conntrack(E) ip6_tables(E) iptable_filter(E)
ip_tables(E) x_tables(E) nfsd(E) auth_rpcgss(E) nfs_acl(E) nfs(E) lockd(E) grace(E) fscache(E) sunrpc(E) 8021q
(E) garp(E) mrp(E) bridge(E) stp(E) llc(E) lru_cache(E) libcrc32c(E) crc32c_generic(E) x86_pkg_temp_thermal(E)
intel_powerclamp(E) coretemp(E) kvm_intel(E) kvm(E) irqbypass(E) crct10dif_pclmul(E)<4>[1056355.855837] CPU: 0
PID: 21195 Comm: drbd_w_bs Tainted: G           OE   4.6.0-0.bpo.1-amd64 #1 Debian 4.6.4-1~bpo8+1
Sep  9 04:26:34 host14 kernel: [1056355.864470] RIP: 0010:[<ffffffff81320246>]  [<ffffffff81320246>]
memcpy_erms+0x6/0x10
Sep  9 04:26:34 host14 kernel: [1056355.873051] RDX: 00000000000004d8 RSI: 0000000000001000 RDI:
ffff88084f40b9e8
Sep  9 04:26:34 host14 kernel: [1056355.881789] R13: 00000000000004d8 R14: ffff88084f40bec0 R15:
ffff880859cdfc60
Sep  9 04:26:34 host14 kernel: [1056355.890683] CR2: 0000000000001000 CR3: 0000000001a06000 CR4:
00000000001426f0
Sep  9 04:26:34 host14 kernel: [1056355.899824]  0000000000000faf0 ffff880859cdfc40 0000000000000000
ffffffff81517ae2
Sep  9 04:26:34 host14 kernel: [1056355.909042]  [<ffffffff81324dcf>] ? copy_from_iter+0x22f/0x250
Sep  9 04:26:34 host14 kernel: [1056355.918209]  [<ffffffffc0aa3e49>] ? drbd_send+0xc9/0x1e0 [drbd]
Sep  9 04:26:34 host14 kernel: [1056355.927262]  [<ffffffffc0aa4027>] ? __send_command.isra.42+0xc7/0x1b0 [drbd]
Sep  9 04:26:34 host14 kernel: [1056355.936213]  [<ffffffffc0aa1f50>] ? drbd_destroy_connection+0xf0/0xf0 [drbd]
Sep  9 04:26:34 host14 kernel: [1056355.944993]  [<ffffffff81099ecf>] ? kthread+0xdf/0x100
Sep  9 04:26:34 host14 kernel: [1056355.953598] Code: 90 90 90 90 90 eb 1e 0f 1f 00 48 89 f8 48 89 d1 48 c1 e9
03 83 e2 07 f3 48 a5 89 d1 f3 a4 c3 66 0f 1f 44 00 00 48 89 f8 48 89 d1 <f3> a4 c3 0f 1f 80 00 00 00 00 48 89
f8 48 83 fa 20 72 7e 40 38
Sep  9 04:26:34 host14 kernel: [1056355.965431] CR2: 0000000000001000
[...]
Sep  9 04:27:52 host14 kernel: [1056433.917513] block drbd60: Remote failed to finish a request within 84108ms
> ko-count (7) * timeout (60 * 0.1s)
```

Case 6

Debian kernel 4.6 with DRBD module 4.8.4-1. DRBD again:

```
Sep 28 04:27:25 host14 kernel: [1628982.714636] task: ffff881058992fc0 ti: ffff881057e6c000 task.ti:
ffff881057e6c000
Sep 28 04:27:25 host14 kernel: [1628982.721567] RAX: ffff880667953600 RBX: 0000000000000590 RCX:
00000000000000c0
Sep 28 04:27:25 host14 kernel: [1628982.730672] R13: 00000000000000c0 R14: ffff8806679536c0 R15:
ffff881057e6fc60
Sep 28 04:27:25 host14 kernel: [1628982.742133]  ffffffff81324dcf ffff88042083a080 ffff880775c7bc00
0000000000000590
Sep 28 04:27:25 host14 kernel: [1628982.753478]  [<ffffffff81517ae2>] ? tcp_sendmsg+0x5f2/0xb00
Sep 28 04:27:25 host14 kernel: [1628982.764619]  [<ffffffffc05b5027>] ? __send_command.isra.42+0xc7/0x1b0 [drbd]
Sep 28 04:27:25 host14 kernel: [1628982.775427]  [<ffffffffc05b2f50>] ? drbd_destroy_connection+0xf0/0xf0 [drbd]
Sep 28 04:27:25 host14 kernel: [1628982.788215] RIP  [<ffffffff81320246>] memcpy_erms+0x6/0x10
```

This goes on DRBD and LXC mailing lists when I'm awake again.

Case 7

Debian kernel 4.6 with DRBD module 4.8.4-1. DRBD again:

```
Sep 30 14:12:45 host14 kernel: [203230.540687] Oops: 0000 [#1] SMP
Sep 30 14:12:45 host14 kernel: [203230.541997] CPU: 0 PID: 4211 Comm: drbd_w_bs Tainted: G           OE   4.6.0-
0.bpo.1-amd64 #1 Debian 4.6.4-1~bpo8+1
Sep 30 14:12:45 host14 kernel: [203230.542186] RIP: 0010:[<ffffffff81320246>]  [<ffffffff81320246>]
memcpy_erms+0x6/0x10
Sep 30 14:12:45 host14 kernel: [203230.542344] RDX: 00000000000003b0 RSI: 0000000000000003 RDI: ffff88080a616040
Sep 30 14:12:45 host14 kernel: [203230.542619] CS:  0010 DS: 0000 ES: 0000 CR0: 0000000080050033
Sep 30 14:12:45 host14 kernel: [203230.542863]  00004000000005b4 00000000000005b4 0000000000000a70
0000000000000a00
Sep 30 14:12:45 host14 kernel: [203230.543108]  [<ffffffffc04fde49>] ? drbd_send+0xc9/0x1e0 [drbd]
Sep 30 14:12:45 host14 kernel: [203230.554230]  [<ffffffffc04fbf50>] ? drbd_destroy_connection+0xf0/0xf0 [drbd]
Sep 30 14:12:45 host14 kernel: [203230.564960]  [<ffffffff81099df0>] ? kthread_park+0x50/0x50
Sep 30 14:12:45 host14 kernel: [203230.584805] ---[ end trace 2335d6e97c28a203 ]---
```

Cases 8-11

Same same.

Dismissed solution ideas (after case 4): DRBD9? Commercial support?

Our first idea was to try DRBD9, maybe with commercial support.

So we filled out Linbit's contact form, and got called back quickly.

Conclusion 1: Don't migrate from DRBD8 to DRBD9 unless you need >2 nodes

DRBD9 is for multi-node operation.

For 2-node operation DRBD8 is fine and recommended, and there will be getting support for at least years (or so they said 2016-08-29).

Conclusion 2: Commercial support prices not for us

We mirror within several pairs "pizza boxes" from the shelf, not between 2 top-of-the-line rack-high elephants.
This doesn't fit with their pricing that's based on per-node (and per-year).

So long, staying with pure Open Source approach then.

Trying out DRBD9 for curiousity's sake

One of our server pairs is due to be decommisioned soon and hosts nothing of relevance, under Debian jessie.

So we used Linbit's semi-official Ubuntu PPA Linbit's semi-official Ubuntu PPA to upgrade to DRBD9 the Open Source way.

We managed to do it, somehow. But I don't recommend it for production systems. There are too many hickups and not too much to gain unless you want to migrate to multi-node setups in which case I strongly recommend using new nodes anyway.

Solution attempt (after case 4): Up-to-date upstream DRBD Kernel Module, 8.4.8-1

Conclusion chain

Without checking mailing list archives it seems unlikely that an eventual severe problem with DRBD 8.4.6 (almost 16 months old now) still is not fixed in the current 8.x module eversion.

DRBD's contact area with the rest of the kernel or userspace (libc6) always has been quite thin (Unix rules, Linux without SystemDisabler still is a unix).

Building the module via DKMS (Thanks Dell!?) is no rocket science and widely documented, i.e. in Proxmox PVE's Wiki page on how to Build DRBD kernel module (Proxmox VE is a nice LXC+KVM+HA virtualization distro, using Debian OS, Ubuntu LTS kernels, and ther own unified adminstration tools & UI).

And, after all, DRBD sources are still Open Source.

So, let's build ;-)

Up-to-date DRBD 8 packages in Clazzes.org' Debian repository

See Debian jessie builds of DKMSed upstream DRBD8 Kernel Module.

Failure ;-(

Unfortunately the whole problem of an ever-rising load occured with DRBD module 4.8.4-1 too, as shown in case 5 above.

Refused Idea: Debian jessie non-backport Kernel 3.16 with upstream DRBD module 8.4.8-1

Unfortunately kernel 3.16 fell victim to someone trying to "fix" the VLAN encapsulation. In fact that fix made the kernel drop packets occationally enough to render this kernel unusable.

Probable Solution

Fixing drbd_main.c rg. Kernels 4.0+

We finally entrusted a Kernel specialist, Richard Weinberger from Sigma-Star.at.

We believe that his 0001-drbd-Fix-kernel_sendmsg-usage.patch solves the problem for Kernels 4.0 to 4.9 and have included this in our drbd8-dkms package (See Debian jessie builds of DKMSed upstream DRBD8 Kernel Module and the debian repository's pool directory http://deb.clazzes.org/debian/pool/jessie-drbdpkg-8/).

Kernel 4.10 will get a rewrite of that code and should solve the problem once and for all for everybody.

Update: The patch has also been sent to the drbd-dev list and should be picked up unless they go right to the rewrite for 4.10.

Conclusion

DRBD is faulty with Kernels 4.0-4.9.

Linbit didn't believe it and didn't care.

We had a professional kernel developer fix it.