# 【现网问题】ceph15.2.17版本出现osd.99\_markmedown问题

## 1.问题现象



是云扬直接群里发了截图，提示有这个告警，这个告警我没有收到，不清楚是通过什么途径发送的，后期加了邮件，可以关注起来，经云扬反馈，这个现象在现网出现的次数不多。

## **2.问题背景**

集群信息：出现问题OSD.99所在节点是172.29.224.90

## **3.问题分析**

1.查看问题节点osd.99出现问题时间点的日志如下:

\*\* File Read Latency Histogram By Level [default] \*\*

2025-04-02T15:33:39.611+0800 7f5dea664700 -1 osd.99 226153 heartbeat\_check: no reply from 172.29.224.88:6814 osd.75 since back 2025-04-02T15:33:16.351935+0800 front 2025-04-02T15:33:16.351900+0800 (oldest deadline 2025-04-02T15:33:39.251595+0800)

OSD.99 未收到OSD.75的消息，经黎宇反馈，OSD.75出现nvme掉盘现象，已提单处理

2.从message日志中可以看到当时出现了OSD.99 网络线程msgr-worker-1 17340在CPU 37上长时间处于内核态，无法正常切回用户态，此时CPU处于软死锁，所以网络线程无法正常响应业务消息，从而被mon标down

日志如下：

Apr 2 15:49:01 rg52-blk009.ceph.csyc4.iflytek.net systemd: Starting Session 875151 of user root.

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: NMI watchdog: BUG: soft lockup - CPU#37 stuck for 23s! [msgr-worker-1:17340]

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: Modules linked in: xt\_nat veth udp\_diag xt\_conntrack ipt\_MASQUERADE nf\_nat\_masquerade\_ipv4 nf\_conntrack\_netlink nfnetlink xt\_addrtype iptable\_filter iptable\_nat nf\_conntrack\_ipv4 nf\_defrag\_ipv4 nf\_nat\_ipv4 nf\_nat nf\_conntrack br\_netfilter bridge overlay(T) ipmi\_poweroff ipmi\_watchdog tcp\_diag inet\_diag 8021q garp mrp stp llc bonding dm\_mirror dm\_region\_hash dm\_log dm\_mod vfat fat coretemp kvm\_intel kvm iTCO\_wdt iTCO\_vendor\_support irqbypass crc32\_pclmul ghash\_clmulni\_intel aesni\_intel lrw gf128mul glue\_helper ablk\_helper cryptd ipmi\_ssif pcspkr ses enclosure scsi\_transport\_sas joydev sg mei\_me mei i2c\_i801 shpchp wmi ipmi\_si ipmi\_devintf ipmi\_msghandler acpi\_power\_meter acpi\_cpufreq ip\_tables xfs libcrc32c sd\_mod crc\_t10dif crct10dif\_generic ast drm\_kms\_helper syscopyarea sysfillrect

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: mlx5\_core sysimgblt fb\_sys\_fops ttm igb drm i40e crct10dif\_pclmul ahci crct10dif\_common crc32c\_intel libahci libata devlink dca nvme i2c\_algo\_bit nvme\_core ptp i2c\_core pps\_core megaraid\_sas(OE)

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: CPU: 37 PID: 17340 Comm: msgr-worker-1 Tainted: G W OEL ------------ T 3.10.0-693.el7.x86\_64 #1

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: Hardware name: Nettrix R620 G40/60WA32, BIOS 0PYH001052 04/11/2023

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: task: ffff881f4b709fa0 ti: ffff881f76a90000 task.ti: ffff881f76a90000

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: RIP: 0010:[<ffffffff81589443>] [<ffffffff81589443>] \_\_dev\_queue\_xmit+0x103/0x550

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: RSP: 0018:ffff88203f3439f8 EFLAGS: 00000286

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: RAX: 0000000000002000 RBX: ffff881027a40000 RCX: 00000000000005ea

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: RDX: ffff8800109f26c0 RSI: ffff8800109f2000 RDI: ffff881027a40000

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: RBP: ffff88203f343a40 R08: 00000000000005c8 R09: 0000000000000000

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: R10: ffffffff815ae260 R11: ffff881145274698 R12: ffff88203f343968

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: R13: ffffffff816b5c1d R14: ffff88203f343a40 R15: ffff88203a68fa00

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: FS: 00007f5deeeeb700(0000) GS:ffff88203f340000(0000) knlGS:0000000000000000

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: CS: 0010 DS: 0000 ES: 0000 CR0: 0000000080050033

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: CR2: 000055f481c808c0 CR3: 0000001f8d6a8000 CR4: 00000000003407e0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: DR0: 0000000000000000 DR1: 0000000000000000 DR2: 0000000000000000

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: DR3: 0000000000000000 DR6: 00000000fffe0ff0 DR7: 0000000000000400

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: Stack:

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: ffff881025a00780 0000000300000000 fffffff4ffcdda74 00000000bc318364

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: ffff8800090740f8 0000000000000002 ffff8800090740f8 00000000000005ea

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: ffff88203d4e58c0 ffff88203f343a50 ffffffff815898a0 ffff88203f343a70

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: Call Trace:

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: <IRQ>

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel:

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815898a0>] dev\_queue\_xmit+0x10/0x20

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffffc03bfbd2>] bond\_dev\_queue\_xmit+0x32/0x80 [bonding]

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffffc03c17a6>] bond\_start\_xmit+0x1b6/0x410 [bonding]

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff81586771>] dev\_hard\_start\_xmit+0x171/0x3b0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff81589788>] \_\_dev\_queue\_xmit+0x448/0x550

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815898a0>] dev\_queue\_xmit+0x10/0x20

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffffc0398de5>] vlan\_dev\_hard\_start\_xmit+0x95/0x120 [8021q]

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff81586771>] dev\_hard\_start\_xmit+0x171/0x3b0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff81589788>] \_\_dev\_queue\_xmit+0x448/0x550

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815898a0>] dev\_queue\_xmit+0x10/0x20

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff81595647>] neigh\_connected\_output+0xb7/0x100

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815cfaa7>] ip\_finish\_output+0x297/0x780

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815d0293>] ip\_output+0x73/0xe0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815cf810>] ? \_\_ip\_append\_data.isra.48+0xa00/0xa00

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815cdf31>] ip\_local\_out\_sk+0x31/0x40

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815ce2a3>] ip\_queue\_xmit+0x143/0x3a0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815e7c0f>] tcp\_transmit\_skb+0x4cf/0x9b0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815e824a>] tcp\_write\_xmit+0x15a/0xce0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815e8e05>] tcp\_tsq\_handler.part.31+0x35/0x40

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815e8fe8>] tcp\_tasklet\_func+0x108/0x140

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff8109084d>] tasklet\_action+0x7d/0x140

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff81090b3f>] \_\_do\_softirq+0xef/0x280

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff816b6a5c>] call\_softirq+0x1c/0x30

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: <EOI>

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel:

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff8102d3c5>] do\_softirq+0x65/0xa0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff8108ffdb>] \_\_local\_bh\_enable\_ip+0x9b/0xb0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff81090007>] local\_bh\_enable+0x17/0x20

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815cfaaf>] ip\_finish\_output+0x29f/0x780

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815d0293>] ip\_output+0x73/0xe0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815cf810>] ? \_\_ip\_append\_data.isra.48+0xa00/0xa00

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815cdf31>] ip\_local\_out\_sk+0x31/0x40

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815ce2a3>] ip\_queue\_xmit+0x143/0x3a0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815e7c0f>] tcp\_transmit\_skb+0x4cf/0x9b0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815e824a>] tcp\_write\_xmit+0x15a/0xce0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815e904e>] \_\_tcp\_push\_pending\_frames+0x2e/0xc0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815d77dc>] tcp\_push+0xec/0x120

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815db0f8>] tcp\_sendmsg+0xc8/0xc30

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815d0293>] ? ip\_output+0x73/0xe0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815cf810>] ? \_\_ip\_append\_data.isra.48+0xa00/0xa00

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff81606893>] inet\_sendmsg+0x63/0xb0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff815e7c0f>] ? tcp\_transmit\_skb+0x4cf/0x9b0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff8156a580>] sock\_sendmsg+0xb0/0xf0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff810cdc7c>] ? dequeue\_entity+0x11c/0x5d0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff8156ae29>] \_\_\_sys\_sendmsg+0x3a9/0x3c0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff8124adf1>] ? ep\_send\_events\_proc+0xb1/0x1d0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff816a866e>] ? schedule\_hrtimeout\_range\_clock+0xbe/0x150

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff8124b6d9>] ? ep\_scan\_ready\_list.isra.7+0x1b9/0x1f0

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff8124b853>] ? ep\_poll+0x123/0x360

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff8156b5f1>] \_\_sys\_sendmsg+0x51/0x90

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff8156b642>] SyS\_sendmsg+0x12/0x20

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: [<ffffffff816b4fc9>] system\_call\_fastpath+0x16/0x1b

Apr 2 15:49:18 rg52-blk009.ceph.csyc4.iflytek.net kernel: Code: 0f 1f 44 00 00 48 83 3b 00 0f 84 19 02 00 00 41 8b 96 dc 00 00 00 49 8b b6 e0 00 00 00 4c 8d ab 9c 00 00 00 41 8b 4e 68 48 01 f2 <41> 89 4e 28 0f b7 7a 02 66 85 ff 74 58 45 0f b7 86 c0 00 00 00

Apr 2 15:49:19 rg52-blk009.ceph.csyc4.iflytek.net sh: abrt-dump-oops: Found oopses: 1

Apr 2 15:49:19 rg52-blk009.ceph.csyc4.iflytek.net sh: abrt-dump-oops: Creating problem directories

Apr 2 15:49:19 rg52-blk009.ceph.csyc4.iflytek.net sh: abrt-dump-oops: Not going to make dump directories world readable because PrivateReports is on

Apr 2 15:49:20 rg52-blk009.ceph.csyc4.iflytek.net abrt-server: Looking for kernel package

Apr 2 15:49:20 rg52-blk009.ceph.csyc4.iflytek.net abrt-server: Kernel package kernel-3.10.0-693.el7.x86\_64 found

Apr 2 15:49:20 rg52-blk009.ceph.csyc4.iflytek.net abrt-dump-oops: Reported 1 kernel oopses to Abrt

Apr 2 15:49:21 rg52-blk009.ceph.csyc4.iflytek.net dbus[1138]: [system] Activating service name='org.freedesktop.problems' (using servicehelper)

Apr 2 15:49:21 rg52-blk009.ceph.csyc4.iflytek.net dbus-daemon: dbus[1138]: [system] Activating service name='org.freedesktop.problems' (using servicehelper)

Apr 2 15:49:21 rg52-blk009.ceph.csyc4.iflytek.net dbus[1138]: [system] Successfully activated service 'org.freedesktop.problems'

Apr 2 15:49:21 rg52-blk009.ceph.csyc4.iflytek.net dbus-daemon: dbus[1138]: [system] Successfully activated service 'org.freedesktop.problems'

Apr 2 15:49:21 rg52-blk009.ceph.csyc4.iflytek.net abrt-server: Email address of sender was not specified. Would you like to do so now? If not, 'user@localhost' is to be used [y/N]

Apr 2 15:49:21 rg52-blk009.ceph.csyc4.iflytek.net abrt-server: Email address of receiver was not specified. Would you like to do so now? If not, 'root@localhost' is to be used [y/N]

Apr 2 15:49:21 rg52-blk009.ceph.csyc4.iflytek.net abrt-server: Sending an email...

Apr 2 15:49:21 rg52-blk009.ceph.csyc4.iflytek.net abrt-server: Sending a notification email to: root@localhost

Apr 2 15:49:21 rg52-blk009.ceph.csyc4.iflytek.net abrt-server: Email was sent to: root@localhost

Apr 2 15:49:22 rg52-blk009.ceph.csyc4.iflytek.net ceph-osd: 2025-04-02T15:49:22.872+0800 7f5de8e61700 -1 osd.99 226178 set\_numa\_affinity unable to identify public interface '' numa node: (2) No such file or directory

Apr 2 15:49:25 rg52-blk009.ceph.csyc4.iflytek.net ceph-osd: 2025-04-02T15:49:25.953+0800 7fd6caf6d700 -1 osd.100 226181 get\_health\_metrics reporting 1 slow ops, oldest is osd\_op(client.170322608.0:16947452 5.8e384e56 5:6a721c71:::rbd\_data.14ca0a61d3d2c0.000000000014c4e7:head [write 0~2531328 in=2531328b] snapc 0=[] ack+ondisk+write+known\_if\_redirected e226175)

Apr 2 15:49:26 rg52-blk009.ceph.csyc4.iflytek.net ceph-osd: 2025-04-02T15:49:26.908+0800 7fd6caf6d700 -1 osd.100 226181 get\_health\_metrics reporting 1 slow ops, oldest is osd\_op(client.170322608.0:16947452 5.8e384e56 5:6a721c71:::rbd\_data.14ca0a61d3d2c0.000000000014c4e7:head [write 0~2531328 in=2531328b] snapc 0=[] ack+ondisk+write+known\_if\_redirected e226175)

Apr 2 15:49:27 rg52-blk009.ceph.csyc4.iflytek.net ceph-osd: 2025-04-02T15:49:27.931+0800 7fd6caf6d700 -1 osd.100 226181 get\_health\_metrics reporting 1 slow ops, oldest is osd\_op(client.170322608.0:16947452 5.8e384e56 5:6a721c71:::rbd\_data.14ca0a61d3d2c0.000000000014c4e7:head [write 0~2531328 in=2531328b] snapc 0=[] ack+ondisk+write+known\_if\_redirected e226175)

Apr 2 15:49:28 rg52-blk009.ceph.csyc4.iflytek.net ceph-osd: 2025-04-02T15:49:28.927+0800 7fd6caf6d700 -1 osd.100 226181 get\_health\_metrics reporting 1 slow ops, oldest is osd\_op(client.170322608.0:16947452 5.8e384e56 5:6a721c71:::rbd\_data.14ca0a61d3d2c0.000000000014c4e7:head [write 0~2531328 in=2531328b] snapc 0=[] ack+ondisk+write+known\_if\_redirected e226175)

这是这一次出现问题时的内核堆栈

3.但是从dmesg日志看，这台节点依次出现过类似的问题，如下所所示：

2024年06月21号 当时是CPU0 PID 1进程出现有问题

[Fri Jun 21 14:49:26 2024] Switched APIC routing to cluster x2apic.

[Fri Jun 21 14:49:26 2024] ..TIMER: vector=0x30 apic1=0 pin1=2 apic2=-1 pin2=-1

[Fri Jun 21 14:49:26 2024] random: fast init done

[Fri Jun 21 14:49:26 2024] smpboot: CPU0: Intel(R) Xeon(R) Silver 4310 CPU @ 2.10GHz (fam: 06, model: 6a, stepping: 06)

[Fri Jun 21 14:49:26 2024] TSC deadline timer enabled

[Fri Jun 21 14:49:26 2024] Performance Events: no PEBS fmt4+, generic architected perfmon,

[Fri Jun 21 14:49:26 2024] ------------[ cut here ]------------

[Fri Jun 21 14:49:26 2024] WARNING: CPU: 0 PID: 1 at arch/x86/events/intel/core.c:3905 intel\_pmu\_init+0xd1d/0xed9

[Fri Jun 21 14:49:26 2024] **hw perf events fixed 4 > max(3), clipping!**

[Fri Jun 21 14:49:26 2024] Modules linked in:

[Fri Jun 21 14:49:26 2024] CPU: 0 PID: 1 Comm: swapper/0 Not tainted 3.10.0-693.el7.x86\_64 #1

[Fri Jun 21 14:49:26 2024] Hardware name: Nettrix R620 G40/60WA32, BIOS 0PYH001052 04/11/2023

[Fri Jun 21 14:49:26 2024] ffff880169ab7d48 00000000e9cbffc5 ffff880169ab7cf8 ffffffff816a3d91

[Fri Jun 21 14:49:26 2024] ffff880169ab7d38 ffffffff810879c8 00000f41819f9020 ffffffff819f9020

[Fri Jun 21 14:49:26 2024] ffff88017fc04660 ffffffff81b46f4e 0000000000000000 0000000000000000

[Fri Jun 21 14:49:26 2024] Call Trace:

[Fri Jun 21 14:49:26 2024] [<ffffffff816a3d91>] dump\_stack+0x19/0x1b

[Fri Jun 21 14:49:26 2024] [<ffffffff810879c8>] \_\_warn+0xd8/0x100

[Fri Jun 21 14:49:26 2024] [<ffffffff81b46f4e>] ? merge\_attr+0x8d/0x8d

[Fri Jun 21 14:49:26 2024] [<ffffffff81087a4f>] warn\_slowpath\_fmt+0x5f/0x80

[Fri Jun 21 14:49:26 2024] [<ffffffff81b48c41>] intel\_pmu\_init+0xd1d/0xed9

[Fri Jun 21 14:49:26 2024] [<ffffffff81b46f4e>] ? merge\_attr+0x8d/0x8d

[Fri Jun 21 14:49:26 2024] [<ffffffff81b46f98>] init\_hw\_perf\_events+0x4a/0x54c

[Fri Jun 21 14:49:26 2024] [<ffffffff81b46f4e>] ? merge\_attr+0x8d/0x8d

[Fri Jun 21 14:49:26 2024] [<ffffffff810020e8>] do\_one\_initcall+0xb8/0x230

[Fri Jun 21 14:49:26 2024] [<ffffffff81b45153>] kernel\_init\_freeable+0xd2/0x219

[Fri Jun 21 14:49:26 2024] [<ffffffff81692c60>] ? rest\_init+0x80/0x80

[Fri Jun 21 14:49:26 2024] [<ffffffff81692c6e>] kernel\_init+0xe/0xf0

[Fri Jun 21 14:49:26 2024] [<ffffffff816b4f18>] ret\_from\_fork+0x58/0x90

[Fri Jun 21 14:49:26 2024] [<ffffffff81692c60>] ? rest\_init+0x80/0x80

[Fri Jun 21 14:49:26 2024] ---[ end trace ca7586ad0ac05b5d ]---

[Fri Jun 21 14:49:26 2024] full-width counters, Intel PMU driver.

[Fri Jun 21 14:49:26 2024] ... version: 5

[Fri Jun 21 14:49:26 2024] ... bit width: 48

[Fri Jun 21 14:49:26 2024] ... generic registers: 8

[Fri Jun 21 14:49:26 2024] ... value mask: 0000ffffffffffff

[Fri Jun 21 14:49:26 2024] ... max period: 00007fffffffffff

[Fri Jun 21 14:49:26 2024] ... fixed-purpose events: 3

[Fri Jun 21 14:49:26 2024] ... event mask: 00000007000000ff

[Fri Jun 21 14:49:26 2024] NMI watchdog: enabled on all CPUs, permanently consumes one hw-PMU counter.

[Fri Jun 21 14:49:26 2024] smpboot: Booting Node 0, Processors #1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 OK

[Fri Jun 21 14:49:26 2024] smpboot: Booting Node 1, Processors #12 #13 #14 #15 #16 #17 #18 #19 #20 #21 #22 #23 OK

[Fri Jun 21 14:49:26 2024] smpboot: Booting Node 0, Processors #24 #25 #26 #27 #28 #29 #30 #31 #32 #33 #34 #35 OK

[Fri Jun 21 14:49:26 2024] smpboot: Booting Node 1, Processors #36 #37 #38 #39 #40 #41 #42 #43 #44 #45 #46 #47 OK

[Fri Jun 21 14:49:26 2024] Brought up 48 CPUs

从报错信息来看，ds给出的结论如下:





这里有个疑问，为什么我们要使用内核Linux rbd\_perf\_24 3.10.0-693.el7.x86\_64 #1 SMP Tue Aug 22 21:09:27 UTC 2017 x86\_64 x86\_64 x86\_64 GNU/Linux 版本，这个是谁规定的，目前来看，内核版本和CPU型号的确是存在一些兼容性问题，只是说这个问题暂时不影响内核的启动、以及存储服务业务的使用

[Sun Jun 23 23:30:08 2024] ctnetlink v0.93: registering with nfnetlink.

[Sun Jun 23 23:30:08 2024] IPv6: ADDRCONF(NETDEV\_UP): docker0: link is not ready

[Mon Jun 24 19:33:58 2024] perf: interrupt took too long (2507 > 2500), lowering kernel.perf\_event\_max\_sample\_rate to 79000

[Tue Jun 25 01:54:18 2024] perf: interrupt took too long (3135 > 3133), lowering kernel.perf\_event\_max\_sample\_rate to 63000

[Tue Jun 25 15:01:47 2024] perf: interrupt took too long (3919 > 3918), lowering kernel.perf\_event\_max\_sample\_rate to 51000

[Wed Jun 26 17:52:16 2024] perf: interrupt took too long (4914 > 4898), lowering kernel.perf\_event\_max\_sample\_rate to 40000

[Fri Jun 28 14:01:31 2024] perf: interrupt took too long (6148 > 6142), lowering kernel.perf\_event\_max\_sample\_rate to 32000

[Sat Jun 29 18:49:46 2024] perf: interrupt took too long (7799 > 7685), lowering kernel.perf\_event\_max\_sample\_rate to 25000

[Mon Jul 1 10:01:49 2024] perf: interrupt took too long (10206 > 9748), lowering kernel.perf\_event\_max\_sample\_rate to 19000

[Sun Jul 7 17:36:35 2024] perf: interrupt took too long (12872 > 12757), lowering kernel.perf\_event\_max\_sample\_rate to 15000

[Tue Jul 16 18:08:13 2024] perf: interrupt took too long (16684 > 16090), lowering kernel.perf\_event\_max\_sample\_rate to 11000

[Wed Jul 17 18:44:36 2024] perf: interrupt took too long (21573 > 20855), lowering kernel.perf\_event\_max\_sample\_rate to 9000

[Thu Jul 18 18:17:29 2024] hrtimer: interrupt took 406484 ns

[Fri Jul 19 19:10:22 2024] perf: interrupt took too long (27804 > 26966), lowering kernel.perf\_event\_max\_sample\_rate to 7000

[Tue Jul 23 01:45:35 2024] NMI watchdog: BUG: soft lockup - CPU#11 stuck for 22s! [migration/11:63]

[Tue Jul 23 01:45:35 2024] Modules linked in: udp\_diag xt\_conntrack ipt\_MASQUERADE nf\_nat\_masquerade\_ipv4 nf\_conntrack\_netlink nfnetlink xt\_addrtype iptable\_filter iptable\_nat nf\_conntrack\_ipv4 nf\_defrag\_ipv4 nf\_nat\_ipv4 nf\_nat nf\_conntrack br\_netfilter bridge overlay(T) ipmi\_poweroff ipmi\_watchdog tcp\_diag inet\_diag 8021q garp mrp stp llc bonding dm\_mirror dm\_region\_hash dm\_log dm\_mod vfat fat coretemp kvm\_intel kvm iTCO\_wdt iTCO\_vendor\_support irqbypass crc32\_pclmul ghash\_clmulni\_intel aesni\_intel lrw gf128mul glue\_helper ablk\_helper cryptd ipmi\_ssif pcspkr ses enclosure scsi\_transport\_sas joydev sg mei\_me mei i2c\_i801 shpchp wmi ipmi\_si ipmi\_devintf ipmi\_msghandler acpi\_power\_meter acpi\_cpufreq ip\_tables xfs libcrc32c sd\_mod crc\_t10dif crct10dif\_generic ast drm\_kms\_helper syscopyarea sysfillrect mlx5\_core sysimgblt



这个要么是内核版本和cpu硬件的不匹配问题，要么就是当时的系统资源紧张，但是这个环境没有atop，无法继续查看当时的系统资源，现在无法继续明确。

[Thu Jul 18 18:17:29 2024] hrtimer: interrupt took 406484 ns

[Fri Jul 19 19:10:22 2024] perf: interrupt took too long (27804 > 26966), lowering kernel.perf\_event\_max\_sample\_rate to 7000

[Tue Jul 23 01:45:35 2024] NMI watchdog: BUG: soft lockup - CPU#11 stuck for 22s! [migration/11:63]

[Tue Jul 23 01:45:35 2024] Modules linked in: udp\_diag xt\_conntrack ipt\_MASQUERADE nf\_nat\_masquerade\_ipv4 nf\_conntrack\_netlink nfnetlink xt\_addrtype iptable\_filter iptable\_nat nf\_conntrack\_ipv4 nf\_defrag\_ipv4 nf\_nat\_ipv4 nf\_nat nf\_conntrack br\_netfilter bridge overlay(T) ipmi\_poweroff ipmi\_watchdog tcp\_diag inet\_diag 8021q garp mrp stp llc bonding dm\_mirror dm\_region\_hash dm\_log dm\_mod vfat fat coretemp kvm\_intel kvm iTCO\_wdt iTCO\_vendor\_support irqbypass crc32\_pclmul ghash\_clmulni\_intel aesni\_intel lrw gf128mul glue\_helper ablk\_helper cryptd ipmi\_ssif pcspkr ses enclosure scsi\_transport\_sas joydev sg mei\_me mei i2c\_i801 shpchp wmi ipmi\_si ipmi\_devintf ipmi\_msghandler acpi\_power\_meter acpi\_cpufreq ip\_tables xfs libcrc32c sd\_mod crc\_t10dif crct10dif\_generic ast drm\_kms\_helper syscopyarea sysfillrect mlx5\_core sysimgblt

[Tue Jul 23 01:45:35 2024] fb\_sys\_fops ttm igb drm i40e crct10dif\_pclmul ahci crct10dif\_common crc32c\_intel libahci libata devlink dca nvme i2c\_algo\_bit nvme\_core ptp i2c\_core pps\_core megaraid\_sas(OE)

[Tue Jul 23 01:45:35 2024] CPU: 11 PID: 63 Comm: migration/11 Tainted: G W OE ------------ T 3.10.0-693.el7.x86\_64 #1

[Tue Jul 23 01:45:35 2024] Hardware name: Nettrix R620 G40/60WA32, BIOS 0PYH001052 04/11/2023

[Tue Jul 23 01:45:35 2024] task: ffff880168f1cf10 ti: ffff880168f58000 task.ti: ffff880168f58000

[Tue Jul 23 01:45:35 2024] RIP: 0010:[<ffffffff8111680a>] [<ffffffff8111680a>] multi\_cpu\_stop+0x4a/0x100

[Tue Jul 23 01:45:35 2024] RSP: 0000:ffff880168f5bd98 EFLAGS: 00000246

[Tue Jul 23 01:45:35 2024] RAX: 0000000000000001 RBX: ffff881f4b661fa0 RCX: dead000000000200

[Tue Jul 23 01:45:35 2024] RDX: ffff88103fecffb0 RSI: 0000000000000286 RDI: ffff881f5d37bb70

[Tue Jul 23 01:45:35 2024] RBP: ffff880168f5bdc0 R08: ffff881f5d37bb10 R09: 0000000000000001

[Tue Jul 23 01:45:35 2024] R10: 000000000000b826 R11: 0000000000000001 R12: ffffffff810be571

[Tue Jul 23 01:45:35 2024] R13: ffff880168f5bd20 R14: 0000000000000000 R15: ffff88203cb2f6b4

[Tue Jul 23 01:45:35 2024] FS: 0000000000000000(0000) GS:ffff88103fec0000(0000) knlGS:0000000000000000

[Tue Jul 23 01:45:35 2024] CS: 0010 DS: 0000 ES: 0000 CR0: 0000000080050033

[Tue Jul 23 01:45:35 2024] CR2: 00005579de9a2ab4 CR3: 00000000019f2000 CR4: 00000000003407e0

[Tue Jul 23 01:45:35 2024] DR0: 0000000000000000 DR1: 0000000000000000 DR2: 0000000000000000

[Tue Jul 23 01:45:35 2024] DR3: 0000000000000000 DR6: 00000000fffe0ff0 DR7: 0000000000000400

[Tue Jul 23 01:45:35 2024] Stack:

[Tue Jul 23 01:45:35 2024] ffff88103fecffa8 ffff881f5d37bb70 ffff88103fecffa0 ffff881f5d37bb98

[Tue Jul 23 01:45:35 2024] ffffffff811167c0 ffff880168f5be90 ffffffff81116ab7 ffff880168f5bfd8

[Tue Jul 23 01:45:35 2024] ffff88103fecffb0 0000000000000000 0000000000000000 ffff88103fed6cc0

[Tue Jul 23 01:45:35 2024] Call Trace:

[Tue Jul 23 01:45:35 2024] [<ffffffff811167c0>] ? cpu\_stop\_should\_run+0x50/0x50

[Tue Jul 23 01:45:35 2024] [<ffffffff81116ab7>] cpu\_stopper\_thread+0x97/0x150

[Tue Jul 23 01:45:35 2024] [<ffffffff816a8f8d>] ? \_\_schedule+0x39d/0x8b0

[Tue Jul 23 01:45:35 2024] [<ffffffff810b909f>] smpboot\_thread\_fn+0x12f/0x180

[Tue Jul 23 01:45:35 2024] [<ffffffff810b8f70>] ? lg\_double\_unlock+0x40/0x40

[Tue Jul 23 01:45:35 2024] [<ffffffff810b098f>] kthread+0xcf/0xe0

[Tue Jul 23 01:45:35 2024] [<ffffffff810b08c0>] ? insert\_kthread\_work+0x40/0x40

[Tue Jul 23 01:45:35 2024] [<ffffffff816b4f18>] ret\_from\_fork+0x58/0x90

[Tue Jul 23 01:45:35 2024] [<ffffffff810b08c0>] ? insert\_kthread\_work+0x40/0x40

[Tue Jul 23 01:45:35 2024] Code: 1f 44 00 00 49 89 c5 48 8b 47 18 48 85 c0 0f 84 ab 00 00 00 0f a3 18 19 db 85 db 41 0f 95 c6 45 31 ff 31 c0 0f 1f 44 00 00 f3 90 <41> 8b 5c 24 20 39 c3 74 5d 83 fb 02 74 68 83 fb 03 75 05 45 84

[Tue Jul 23 01:45:35 2024] NMI watchdog: BUG: soft lockup - CPU#38 stuck for 23s! [msgr-worker-1:20489]

[Tue Jul 23 01:45:35 2024] Modules linked in: udp\_diag xt\_conntrack ipt\_MASQUERADE nf\_nat\_masquerade\_ipv4 nf\_conntrack\_netlink nfnetlink xt\_addrtype iptable\_filter iptable\_nat nf\_conntrack\_ipv4 nf\_defrag\_ipv4 nf\_nat\_ipv4 nf\_nat nf\_conntrack br\_netfilter bridge overlay(T) ipmi\_poweroff ipmi\_watchdog tcp\_diag inet\_diag 8021q garp mrp stp llc bonding dm\_mirror dm\_region\_hash dm\_log dm\_mod vfat fat coretemp kvm\_intel kvm iTCO\_wdt iTCO\_vendor\_support irqbypass crc32\_pclmul ghash\_clmulni\_intel aesni\_intel lrw gf128mul glue\_helper ablk\_helper cryptd ipmi\_ssif pcspkr ses enclosure scsi\_transport\_sas joydev sg mei\_me mei i2c\_i801 shpchp wmi ipmi\_si ipmi\_devintf ipmi\_msghandler acpi\_power\_meter acpi\_cpufreq ip\_tables xfs libcrc32c sd\_mod crc\_t10dif crct10dif\_generic ast drm\_kms\_helper syscopyarea sysfillrect mlx5\_core sysimgblt

[Tue Jul 23 01:45:35 2024] fb\_sys\_fops ttm igb drm i40e crct10dif\_pclmul ahci crct10dif\_common crc32c\_intel libahci libata devlink dca nvme i2c\_algo\_bit nvme\_core ptp i2c\_core pps\_core megaraid\_sas(OE)

该问题ds给出的解释如下：



从出现问题的进程PID: 63 Comm: migration上看，似乎和业务进程无关，这里暂时排除是业务自身的问题。

[Tue Jul 23 01:45:35 2024] fb\_sys\_fops ttm igb drm i40e crct10dif\_pclmul ahci crct10dif\_common crc32c\_intel libahci libata devlink dca nvme i2c\_algo\_bit nvme\_core ptp i2c\_core pps\_core megaraid\_sas(OE)

[Tue Jul 23 01:45:35 2024] CPU: 38 PID: 20489 Comm: msgr-worker-1 Tainted: G W OEL ------------ T 3.10.0-693.el7.x86\_64 #1

[Tue Jul 23 01:45:35 2024] Hardware name: Nettrix R620 G40/60WA32, BIOS 0PYH001052 04/11/2023

[Tue Jul 23 01:45:35 2024] task: ffff88201173eeb0 ti: ffff880f72ae4000 task.ti: ffff880f72ae4000

[Tue Jul 23 01:45:35 2024] RIP: 0010:[<ffffffff815895c2>] [<ffffffff815895c2>] \_\_dev\_queue\_xmit+0x282/0x550

[Tue Jul 23 01:45:35 2024] RSP: 0018:ffff88203f383a28 EFLAGS: 00000286

[Tue Jul 23 01:45:35 2024] RAX: 0000000000000000 RBX: ffffffff815af476 RCX: 0000000000000000

[Tue Jul 23 01:45:35 2024] RDX: ffff880f72ae7fd8 RSI: 0000000000000200 RDI: ffffffff815895a5

[Tue Jul 23 01:45:35 2024] RBP: ffff88203f383a40 R08: 0000000000000001 R09: 0000000000000000

[Tue Jul 23 01:45:35 2024] R10: 0000000000000040 R11: ffffffff813498a0 R12: ffff88203f383998

[Tue Jul 23 01:45:35 2024] R13: ffffffff816b5c1d R14: ffff88203f383a40 R15: ffff8803f28206f8

[Tue Jul 23 01:45:35 2024] FS: 00007fab5b84d700(0000) GS:ffff88203f380000(0000) knlGS:0000000000000000

[Tue Jul 23 01:45:35 2024] CS: 0010 DS: 0000 ES: 0000 CR0: 0000000080050033

[Tue Jul 23 01:45:35 2024] CR2: 000055c2a0efa480 CR3: 00000010247b6000 CR4: 00000000003407e0

[Tue Jul 23 01:45:35 2024] DR0: 0000000000000000 DR1: 0000000000000000 DR2: 0000000000000000

[Tue Jul 23 01:45:35 2024] DR3: 0000000000000000 DR6: 00000000fffe0ff0 DR7: 0000000000000400

[Tue Jul 23 01:45:35 2024] Stack:

[Tue Jul 23 01:45:35 2024] ffff8803f28206f8 00000000000005ea ffff88203d4e58c0 ffff88203f383a50

[Tue Jul 23 01:45:35 2024] ffffffff815898a0 ffff88203f383a70 ffffffffc03bfbd2 ffff882025ba7bc0

[Tue Jul 23 01:45:35 2024] 0000000000000002 ffff88203f383ab8 ffffffffc03c17a6 ffff8809202ea4d0

[Tue Jul 23 01:45:35 2024] Call Trace:

[Tue Jul 23 01:45:35 2024] <IRQ>

[Tue Jul 23 01:45:35 2024] [<ffffffff815898a0>] dev\_queue\_xmit+0x10/0x20

[Tue Jul 23 01:45:35 2024] [<ffffffffc03bfbd2>] bond\_dev\_queue\_xmit+0x32/0x80 [bonding]

[Tue Jul 23 01:45:35 2024] [<ffffffffc03c17a6>] bond\_start\_xmit+0x1b6/0x410 [bonding]

[Tue Jul 23 01:45:35 2024] [<ffffffff81586771>] dev\_hard\_start\_xmit+0x171/0x3b0

[Tue Jul 23 01:45:35 2024] [<ffffffff81589788>] \_\_dev\_queue\_xmit+0x448/0x550

[Tue Jul 23 01:45:35 2024] [<ffffffff815898a0>] dev\_queue\_xmit+0x10/0x20

[Tue Jul 23 01:45:35 2024] [<ffffffffc0398de5>] vlan\_dev\_hard\_start\_xmit+0x95/0x120 [8021q]

[Tue Jul 23 01:45:35 2024] [<ffffffff81586771>] dev\_hard\_start\_xmit+0x171/0x3b0

[Tue Jul 23 01:45:35 2024] [<ffffffff81589788>] \_\_dev\_queue\_xmit+0x448/0x550

[Tue Jul 23 01:45:35 2024] [<ffffffff815898a0>] dev\_queue\_xmit+0x10/0x20

[Tue Jul 23 01:45:35 2024] [<ffffffff81595647>] neigh\_connected\_output+0xb7/0x100

[Tue Jul 23 01:45:35 2024] [<ffffffff815cfaa7>] ip\_finish\_output+0x297/0x780

[Tue Jul 23 01:45:35 2024] [<ffffffff815d0293>] ip\_output+0x73/0xe0

[Tue Jul 23 01:45:35 2024] [<ffffffff815cf810>] ? \_\_ip\_append\_data.isra.48+0xa00/0xa00

[Tue Jul 23 01:45:35 2024] [<ffffffff815cdf31>] ip\_local\_out\_sk+0x31/0x40

[Tue Jul 23 01:45:35 2024] [<ffffffff815ce2a3>] ip\_queue\_xmit+0x143/0x3a0

[Tue Jul 23 01:45:35 2024] [<ffffffff815e7c0f>] tcp\_transmit\_skb+0x4cf/0x9b0

[Tue Jul 23 01:45:35 2024] [<ffffffff815e824a>] tcp\_write\_xmit+0x15a/0xce0

[Tue Jul 23 01:45:35 2024] [<ffffffff815e8e05>] tcp\_tsq\_handler.part.31+0x35/0x40

[Tue Jul 23 01:45:35 2024] [<ffffffff815e8fe8>] tcp\_tasklet\_func+0x108/0x140

[Tue Jul 23 01:45:35 2024] [<ffffffff8109084d>] tasklet\_action+0x7d/0x140

[Tue Jul 23 01:45:35 2024] [<ffffffff81090b3f>] \_\_do\_softirq+0xef/0x280

[Tue Jul 23 01:45:35 2024] [<ffffffff816b6a5c>] call\_softirq+0x1c/0x30

[Tue Jul 23 01:45:35 2024] <EOI>

[Tue Jul 23 01:45:35 2024] [<ffffffff8102d3c5>] do\_softirq+0x65/0xa0

[Tue Jul 23 01:45:35 2024] [<ffffffff8108ffdb>] \_\_local\_bh\_enable\_ip+0x9b/0xb0

[Tue Jul 23 01:45:35 2024] [<ffffffff81090007>] local\_bh\_enable+0x17/0x20

[Tue Jul 23 01:45:35 2024] [<ffffffffc0085d06>] ipt\_do\_table+0x336/0x710 [ip\_tables]

[Tue Jul 23 01:45:35 2024] [<ffffffff81090007>] ? local\_bh\_enable+0x17/0x20

[Tue Jul 23 01:45:35 2024] [<ffffffffc057d515>] ? \_\_nf\_conntrack\_find\_get+0x185/0x360 [nf\_conntrack]

[Tue Jul 23 01:45:35 2024] [<ffffffffc0540020>] ? iptable\_nat\_ipv4\_fn+0x20/0x20 [iptable\_nat]

[Tue Jul 23 01:45:35 2024] [<ffffffffc059e036>] iptable\_filter\_hook+0x36/0x80 [iptable\_filter]

[Tue Jul 23 01:45:35 2024] [<ffffffff815bfe60>] nf\_iterate+0x70/0xb0

[Tue Jul 23 01:45:35 2024] [<ffffffff815bff48>] nf\_hook\_slow+0xa8/0x110

[Tue Jul 23 01:45:35 2024] [<ffffffff815cdef6>] \_\_ip\_local\_out\_sk+0xf6/0x100

[Tue Jul 23 01:45:35 2024] [<ffffffff815cd5c0>] ? ip\_forward\_options+0x1c0/0x1c0

[Tue Jul 23 01:45:35 2024] [<ffffffff815cdf1b>] ip\_local\_out\_sk+0x1b/0x40

[Tue Jul 23 01:45:35 2024] [<ffffffff815ce2a3>] ip\_queue\_xmit+0x143/0x3a0

[Tue Jul 23 01:45:35 2024] [<ffffffff815e7c0f>] tcp\_transmit\_skb+0x4cf/0x9b0

[Tue Jul 23 01:45:35 2024] [<ffffffff815e824a>] tcp\_write\_xmit+0x15a/0xce0

[Tue Jul 23 01:45:35 2024] [<ffffffff815e904e>] \_\_tcp\_push\_pending\_frames+0x2e/0xc0

[Tue Jul 23 01:45:35 2024] [<ffffffff815d77dc>] tcp\_push+0xec/0x120

[Tue Jul 23 01:45:35 2024] [<ffffffff815db0f8>] tcp\_sendmsg+0xc8/0xc30

[Tue Jul 23 01:45:35 2024] [<ffffffff815d0293>] ? ip\_output+0x73/0xe0

[Tue Jul 23 01:45:35 2024] [<ffffffff815cf810>] ? \_\_ip\_append\_data.isra.48+0xa00/0xa00

[Tue Jul 23 01:45:35 2024] [<ffffffff81606893>] inet\_sendmsg+0x63/0xb0

[Tue Jul 23 01:45:35 2024] [<ffffffff815e7c0f>] ? tcp\_transmit\_skb+0x4cf/0x9b0

[Tue Jul 23 01:45:35 2024] [<ffffffff8156a580>] sock\_sendmsg+0xb0/0xf0

[Tue Jul 23 01:45:35 2024] [<ffffffff811de581>] ? \_\_slab\_free+0x81/0x2f0

[Tue Jul 23 01:45:35 2024] [<ffffffff8156ae29>] \_\_\_sys\_sendmsg+0x3a9/0x3c0

[Tue Jul 23 01:45:35 2024] [<ffffffff816a7722>] ? mutex\_lock+0x12/0x2f

[Tue Jul 23 01:45:35 2024] [<ffffffff8120019d>] ? do\_sync\_read+0x8d/0xd0

[Tue Jul 23 01:45:35 2024] [<ffffffff8156b5f1>] \_\_sys\_sendmsg+0x51/0x90

[Tue Jul 23 01:45:35 2024] [<ffffffff8156b642>] SyS\_sendmsg+0x12/0x20

[Tue Jul 23 01:45:35 2024] [<ffffffff816b4fc9>] system\_call\_fastpath+0x16/0x1b

[Tue Jul 23 01:45:35 2024] Code: 85 78 02 00 00 44 89 65 cc e8 4b 6a b0 ff 8b 45 cc 48 8b 4d d0 65 48 33 0c 25 28 00 00 00 0f 85 3d 02 00 00 48 83 c4 20 5b 41 5c <41> 5d 41 5e 41 5f 5d c3 66 0f 1f 44 00 00 48 89 de 4c 89 f7 ff



[Fri Jul 26 05:48:08 2024] fb\_sys\_fops ttm igb drm i40e crct10dif\_pclmul ahci crct10dif\_common crc32c\_intel libahci libata devlink dca nvme i2c\_algo\_bit nvme\_core ptp i2c\_core pps\_core megaraid\_sas(OE)

[Fri Jul 26 05:48:08 2024] CPU: 36 PID: 20067 Comm: msgr-worker-2 Tainted: G W OEL ------------ T 3.10.0-693.el7.x86\_64 #1

[Fri Jul 26 05:48:08 2024] Hardware name: Nettrix R620 G40/60WA32, BIOS 0PYH001052 04/11/2023

[Fri Jul 26 05:48:08 2024] task: ffff881eb9bb2f70 ti: ffff880f766a8000 task.ti: ffff880f766a8000

[Fri Jul 26 05:48:08 2024] RIP: 0010:[<ffffffff815e7488>] [<ffffffff815e7488>] tcp\_schedule\_loss\_probe+0xa8/0x1e0

[Fri Jul 26 05:48:08 2024] RSP: 0018:ffff88203f303e28 EFLAGS: 00000216

[Fri Jul 26 05:48:08 2024] RAX: 0000000000000007 RBX: ffff881e517d1f00 RCX: 0000000000000007

[Fri Jul 26 05:48:08 2024] RDX: 0000000000000007 RSI: 0000000000000004 RDI: 0000000000000f1f

[Fri Jul 26 05:48:08 2024] RBP: ffff88203f303e38 R08: 0000000000000001 R09: 00000000000005a8

[Fri Jul 26 05:48:08 2024] R10: 0000000021ead097 R11: 0000000000000000 R12: ffff88203f303d98

[Fri Jul 26 05:48:08 2024] R13: ffffffff816b5c1d R14: ffff88203f303e38 R15: ffff881e517d1f00

[Fri Jul 26 05:48:08 2024] FS: 00007f66e6b9e700(0000) GS:ffff88203f300000(0000) knlGS:0000000000000000

[Fri Jul 26 05:48:08 2024] CS: 0010 DS: 0000 ES: 0000 CR0: 0000000080050033

[Fri Jul 26 05:48:08 2024] CR2: 000055b6bf45fe00 CR3: 0000000faab93000 CR4: 00000000003407e0

[Fri Jul 26 05:48:08 2024] DR0: 0000000000000000 DR1: 0000000000000000 DR2: 0000000000000000

[Fri Jul 26 05:48:08 2024] DR3: 0000000000000000 DR6: 00000000fffe0ff0 DR7: 0000000000000400

[Fri Jul 26 05:48:08 2024] Stack:

[Fri Jul 26 05:48:08 2024] ffff881e517d1f00 ffff880003f69200 ffff88203f303ea0 ffffffff815e83a0

[Fri Jul 26 05:48:08 2024] ffff88203f303ed0 0000000100000000 00000000815e7243 ffff881e517d2038

[Fri Jul 26 05:48:08 2024] ffff880100003ec8 000000206da56100 ffff881e517d1f00 0000000000000001

[Fri Jul 26 05:48:08 2024] Call Trace:

[Fri Jul 26 05:48:08 2024] <IRQ>

[Fri Jul 26 05:48:08 2024] [<ffffffff815e83a0>] tcp\_write\_xmit+0x2b0/0xce0

[Fri Jul 26 05:48:08 2024] [<ffffffff815e8e05>] tcp\_tsq\_handler.part.31+0x35/0x40

[Fri Jul 26 05:48:08 2024] [<ffffffff815e8fe8>] tcp\_tasklet\_func+0x108/0x140

[Fri Jul 26 05:48:08 2024] [<ffffffff8109084d>] tasklet\_action+0x7d/0x140

[Fri Jul 26 05:48:08 2024] [<ffffffff81090b3f>] \_\_do\_softirq+0xef/0x280

[Fri Jul 26 05:48:08 2024] [<ffffffff816b6a5c>] call\_softirq+0x1c/0x30

[Fri Jul 26 05:48:08 2024] <EOI>

[Fri Jul 26 05:48:08 2024] [<ffffffff8102d3c5>] do\_softirq+0x65/0xa0

[Fri Jul 26 05:48:08 2024] [<ffffffff8108ffdb>] \_\_local\_bh\_enable\_ip+0x9b/0xb0

[Fri Jul 26 05:48:08 2024] [<ffffffff81090007>] local\_bh\_enable+0x17/0x20

[Fri Jul 26 05:48:08 2024] [<ffffffffc057d515>] \_\_nf\_conntrack\_find\_get+0x185/0x360 [nf\_conntrack]

[Fri Jul 26 05:48:08 2024] [<ffffffffc057efed>] nf\_conntrack\_in+0x27d/0x5f0 [nf\_conntrack]

[Fri Jul 26 05:48:08 2024] [<ffffffffc053366f>] ipv4\_conntrack\_local+0x4f/0x60 [nf\_conntrack\_ipv4]

[Fri Jul 26 05:48:08 2024] [<ffffffff815bfe60>] nf\_iterate+0x70/0xb0

[Fri Jul 26 05:48:08 2024] [<ffffffff815bff48>] nf\_hook\_slow+0xa8/0x110

[Fri Jul 26 05:48:08 2024] [<ffffffff815cdef6>] \_\_ip\_local\_out\_sk+0xf6/0x100

[Fri Jul 26 05:48:08 2024] [<ffffffff815cd5c0>] ? ip\_forward\_options+0x1c0/0x1c0

[Fri Jul 26 05:48:08 2024] [<ffffffff815cdf1b>] ip\_local\_out\_sk+0x1b/0x40

[Fri Jul 26 05:48:08 2024] [<ffffffff815ce2a3>] ip\_queue\_xmit+0x143/0x3a0

[Fri Jul 26 05:48:08 2024] [<ffffffff815e7c0f>] tcp\_transmit\_skb+0x4cf/0x9b0

[Fri Jul 26 05:48:08 2024] [<ffffffff815e824a>] tcp\_write\_xmit+0x15a/0xce0

[Fri Jul 26 05:48:08 2024] [<ffffffff815e904e>] \_\_tcp\_push\_pending\_frames+0x2e/0xc0

[Fri Jul 26 05:48:08 2024] [<ffffffff815d77dc>] tcp\_push+0xec/0x120

[Fri Jul 26 05:48:08 2024] [<ffffffff815db0f8>] tcp\_sendmsg+0xc8/0xc30

[Fri Jul 26 05:48:08 2024] [<ffffffff81606893>] inet\_sendmsg+0x63/0xb0

[Fri Jul 26 05:48:08 2024] [<ffffffff8156a580>] sock\_sendmsg+0xb0/0xf0

[Fri Jul 26 05:48:08 2024] [<ffffffff810cdc7c>] ? dequeue\_entity+0x11c/0x5d0

[Fri Jul 26 05:48:08 2024] [<ffffffff8156ae29>] \_\_\_sys\_sendmsg+0x3a9/0x3c0

[Fri Jul 26 05:48:08 2024] [<ffffffff816a7722>] ? mutex\_lock+0x12/0x2f

[Fri Jul 26 05:48:08 2024] [<ffffffff8120019d>] ? do\_sync\_read+0x8d/0xd0

[Fri Jul 26 05:48:08 2024] [<ffffffff8156b5f1>] \_\_sys\_sendmsg+0x51/0x90

[Fri Jul 26 05:48:08 2024] [<ffffffff8156b642>] SyS\_sendmsg+0x12/0x20

[Fri Jul 26 05:48:08 2024] [<ffffffff816b4fc9>] system\_call\_fastpath+0x16/0x1b

[Fri Jul 26 05:48:08 2024] Code: ea 04 84 d2 74 ca f6 83 d0 04 00 00 3f 75 c1 8b 83 34 06 00 00 01 c8 2b 83 a4 06 00 00 2b 83 a8 06 00 00 39 83 74 06 00 00 76 0c <31> c0 48 83 bb 68 02 00 00 00 75 99 83 f9 01 44 8d 24 36 0f 84

[Fri Jul 26 09:53:12 2024] NMI watchdog: BUG: soft lockup - CPU#47 stuck for 22s! [msgr-worker-0:19482]

[Fri Jul 26 09:53:12 2024] Modules linked in: udp\_diag xt\_conntrack ipt\_MASQUERADE nf\_nat\_masquerade\_ipv4 nf\_conntrack\_netlink nfnetlink xt\_addrtype iptable\_filter iptable\_nat nf\_conntrack\_ipv4 nf\_defrag\_ipv4 nf\_nat\_ipv4 nf\_nat nf\_conntrack br\_netfilter bridge overlay(T) ipmi\_poweroff ipmi\_watchdog tcp\_diag inet\_diag 8021q garp mrp stp llc bonding dm\_mirror dm\_region\_hash dm\_log dm\_mod vfat fat coretemp kvm\_intel kvm iTCO\_wdt iTCO\_vendor\_support irqbypass crc32\_pclmul ghash\_clmulni\_intel aesni\_intel lrw gf128mul glue\_helper ablk\_helper cryptd ipmi\_ssif pcspkr ses enclosure scsi\_transport\_sas joydev sg mei\_me mei i2c\_i801 shpchp wmi ipmi\_si ipmi\_devintf ipmi\_msghandler acpi\_power\_meter acpi\_cpufreq ip\_tables xfs libcrc32c sd\_mod crc\_t10dif crct10dif\_generic ast drm\_kms\_helper syscopyarea sysfillrect mlx5\_core sysimgblt

[Fri Jul 26 09:53:12 2024] fb\_sys\_fops ttm igb drm i40e crct10dif\_pclmul ahci crct10dif\_common crc32c\_intel libahci libata devlink dca nvme i2c\_algo\_bit nvme\_core ptp i2c\_core pps\_core megaraid\_sas(OE)

以上几个都是内核线程，从这个内核软死锁报错开始就是ceph业务线程卡住

[Fri Jul 26 09:53:12 2024] CPU: 47 PID: 19482 Comm: msgr-worker-0 Tainted: G W OEL ------------ T 3.10.0-693.el7.x86\_64 #1

[Fri Jul 26 09:53:12 2024] Hardware name: Nettrix R620 G40/60WA32, BIOS 0PYH001052 04/11/2023

[Fri Jul 26 09:53:12 2024] task: ffff880fe3e52f70 ti: ffff881eb9bb8000 task.ti: ffff881eb9bb8000

[Fri Jul 26 09:53:12 2024] RIP: 0010:[<ffffffff81586644>] [<ffffffff81586644>] dev\_hard\_start\_xmit+0x44/0x3b0

[Fri Jul 26 09:53:12 2024] RSP: 0018:ffff88203f5c3778 EFLAGS: 00000282

[Fri Jul 26 09:53:12 2024] RAX: 0000000000000000 RBX: ffff88203d72a5d0 RCX: ffff88203f5c37fc

[Fri Jul 26 09:53:12 2024] RDX: ffff881023245000 RSI: ffff88203d4e2000 RDI: ffff881338c1fcf8

[Fri Jul 26 09:53:12 2024] RBP: ffff88203f5c37d8 R08: 0000000000000000 R09: 00000000000005dc

[Fri Jul 26 09:53:12 2024] R10: ffffffff815ae260 R11: ffff88170a235f58 R12: ffff88203f5c36e8

[Fri Jul 26 09:53:12 2024] R13: ffffffff816b5c1d R14: ffff88203f5c37d8 R15: ffff88203d4e2000

[Fri Jul 26 09:53:12 2024] FS: 00007fee28a46700(0000) GS:ffff88203f5c0000(0000) knlGS:0000000000000000

[Fri Jul 26 09:53:12 2024] CS: 0010 DS: 0000 ES: 0000 CR0: 0000000080050033

[Fri Jul 26 09:53:12 2024] CR2: 000055ad96a09000 CR3: 0000001eb8f9f000 CR4: 00000000003407e0

[Fri Jul 26 09:53:12 2024] DR0: 0000000000000000 DR1: 0000000000000000 DR2: 0000000000000000

[Fri Jul 26 09:53:12 2024] DR3: 0000000000000000 DR6: 00000000fffe0ff0 DR7: 0000000000000400

[Fri Jul 26 09:53:12 2024] Stack:

[Fri Jul 26 09:53:12 2024] 0000000067d972ab ffff88203f5c37fc ffff881023245000 ffff88203d4e2000

[Fri Jul 26 09:53:12 2024] 0000000000000002 ffff88203f5c37d8 0000000067d972ab 000000000000002f

[Fri Jul 26 09:53:12 2024] ffff88203d4e2000 0000000000000002 ffff881338c1fcf8 ffff881fae70b4c0

[Fri Jul 26 09:53:12 2024] Call Trace:

[Fri Jul 26 09:53:12 2024] <IRQ>

[Fri Jul 26 09:53:12 2024] [<ffffffff81589788>] \_\_dev\_queue\_xmit+0x448/0x550

[Fri Jul 26 09:53:12 2024] [<ffffffff815898a0>] dev\_queue\_xmit+0x10/0x20

[Fri Jul 26 09:53:12 2024] [<ffffffff81595647>] neigh\_connected\_output+0xb7/0x100

[Fri Jul 26 09:53:12 2024] [<ffffffff815cfaa7>] ip\_finish\_output+0x297/0x780

[Fri Jul 26 09:53:12 2024] [<ffffffff815d0293>] ip\_output+0x73/0xe0

[Fri Jul 26 09:53:12 2024] [<ffffffff815cf810>] ? \_\_ip\_append\_data.isra.48+0xa00/0xa00

[Fri Jul 26 09:53:12 2024] [<ffffffff815cdf31>] ip\_local\_out\_sk+0x31/0x40

[Fri Jul 26 09:53:12 2024] [<ffffffff815ce2a3>] ip\_queue\_xmit+0x143/0x3a0

[Fri Jul 26 09:53:12 2024] [<ffffffff815e7c0f>] tcp\_transmit\_skb+0x4cf/0x9b0

[Fri Jul 26 09:53:12 2024] [<ffffffff815e824a>] tcp\_write\_xmit+0x15a/0xce0

[Fri Jul 26 09:53:12 2024] [<ffffffff815e904e>] \_\_tcp\_push\_pending\_frames+0x2e/0xc0

[Fri Jul 26 09:53:12 2024] [<ffffffff815e457b>] tcp\_rcv\_established+0x14b/0x760

[Fri Jul 26 09:53:12 2024] [<ffffffff815ef15a>] tcp\_v4\_do\_rcv+0x10a/0x340

[Fri Jul 26 09:53:12 2024] [<ffffffff815f08a8>] tcp\_v4\_rcv+0x798/0x9a0

[Fri Jul 26 09:53:12 2024] [<ffffffffc0533306>] ? ipv4\_confirm+0x86/0x100 [nf\_conntrack\_ipv4]

[Fri Jul 26 09:53:12 2024] [<ffffffff815ca054>] ip\_local\_deliver\_finish+0xb4/0x1f0

[Fri Jul 26 09:53:12 2024] [<ffffffff815ca339>] ip\_local\_deliver+0x59/0xd0

[Fri Jul 26 09:53:12 2024] [<ffffffff815c9fa0>] ? ip\_rcv\_finish+0x350/0x350

[Fri Jul 26 09:53:12 2024] [<ffffffff815c9cda>] ip\_rcv\_finish+0x8a/0x350

[Fri Jul 26 09:53:12 2024] [<ffffffff815ca666>] ip\_rcv+0x2b6/0x410

[Fri Jul 26 09:53:12 2024] [<ffffffff815c9c50>] ? inet\_del\_offload+0x40/0x40

[Fri Jul 26 09:53:12 2024] [<ffffffff81586f22>] \_\_netif\_receive\_skb\_core+0x572/0x7c0

[Fri Jul 26 09:53:12 2024] [<ffffffff815f5944>] ? tcp4\_gro\_receive+0x134/0x1b0

[Fri Jul 26 09:53:12 2024] [<ffffffff81607087>] ? inet\_gro\_receive+0x247/0x290

[Fri Jul 26 09:53:12 2024] [<ffffffff81587188>] \_\_netif\_receive\_skb+0x18/0x60

[Fri Jul 26 09:53:12 2024] [<ffffffff81587210>] netif\_receive\_skb\_internal+0x40/0xc0

[Fri Jul 26 09:53:12 2024] [<ffffffff81588318>] napi\_gro\_receive+0xd8/0x130

[Fri Jul 26 09:53:12 2024] [<ffffffffc0366e5e>] mlx5e\_handle\_rx\_cqe\_mpwrq+0x53e/0x870 [mlx5\_core]

[Fri Jul 26 09:53:12 2024] [<ffffffffc036720b>] mlx5e\_poll\_rx\_cq+0x7b/0x810 [mlx5\_core]

[Fri Jul 26 09:53:12 2024] [<ffffffffc0367f29>] mlx5e\_napi\_poll+0x89/0x370 [mlx5\_core]

[Fri Jul 26 09:53:12 2024] [<ffffffff8158799d>] net\_rx\_action+0x16d/0x380

[Fri Jul 26 09:53:12 2024] [<ffffffff81090b3f>] \_\_do\_softirq+0xef/0x280

[Fri Jul 26 09:53:12 2024] [<ffffffff816b6a5c>] call\_softirq+0x1c/0x30

[Fri Jul 26 09:53:12 2024] <EOI>

[Fri Jul 26 09:53:12 2024] [<ffffffff8102d3c5>] do\_softirq+0x65/0xa0

[Fri Jul 26 09:53:12 2024] [<ffffffff8108ffdb>] \_\_local\_bh\_enable\_ip+0x9b/0xb0

[Fri Jul 26 09:53:12 2024] [<ffffffff81090007>] local\_bh\_enable+0x17/0x20

[Fri Jul 26 09:53:12 2024] [<ffffffffc057d515>] \_\_nf\_conntrack\_find\_get+0x185/0x360 [nf\_conntrack]

[Fri Jul 26 09:53:12 2024] [<ffffffffc057efed>] nf\_conntrack\_in+0x27d/0x5f0 [nf\_conntrack]

[Fri Jul 26 09:53:12 2024] [<ffffffffc053366f>] ipv4\_conntrack\_local+0x4f/0x60 [nf\_conntrack\_ipv4]

[Fri Jul 26 09:53:12 2024] [<ffffffff815bfe60>] nf\_iterate+0x70/0xb0

[Fri Jul 26 09:53:12 2024] [<ffffffff815bff48>] nf\_hook\_slow+0xa8/0x110

[Fri Jul 26 09:53:12 2024] [<ffffffff815cdef6>] \_\_ip\_local\_out\_sk+0xf6/0x100

[Fri Jul 26 09:53:12 2024] [<ffffffff815cd5c0>] ? ip\_forward\_options+0x1c0/0x1c0

[Fri Jul 26 09:53:12 2024] [<ffffffff815cdf1b>] ip\_local\_out\_sk+0x1b/0x40

[Fri Jul 26 09:53:12 2024] [<ffffffff815ce2a3>] ip\_queue\_xmit+0x143/0x3a0

[Fri Jul 26 09:53:12 2024] [<ffffffff815e7c0f>] tcp\_transmit\_skb+0x4cf/0x9b0

[Fri Jul 26 09:53:12 2024] [<ffffffff815e824a>] tcp\_write\_xmit+0x15a/0xce0

[Fri Jul 26 09:53:12 2024] [<ffffffff815e904e>] \_\_tcp\_push\_pending\_frames+0x2e/0xc0

[Fri Jul 26 09:53:12 2024] [<ffffffff815d77dc>] tcp\_push+0xec/0x120

[Fri Jul 26 09:53:12 2024] [<ffffffff815db65d>] tcp\_sendmsg+0x62d/0xc30

[Fri Jul 26 09:53:12 2024] [<ffffffff815d0293>] ? ip\_output+0x73/0xe0

[Fri Jul 26 09:53:12 2024] [<ffffffff815cf810>] ? \_\_ip\_append\_data.isra.48+0xa00/0xa00

[Fri Jul 26 09:53:12 2024] [<ffffffff81606893>] inet\_sendmsg+0x63/0xb0

[Fri Jul 26 09:53:12 2024] [<ffffffff815e7c0f>] ? tcp\_transmit\_skb+0x4cf/0x9b0

[Fri Jul 26 09:53:12 2024] [<ffffffff8156a580>] sock\_sendmsg+0xb0/0xf0

[Fri Jul 26 09:53:12 2024] [<ffffffff810cdc7c>] ? dequeue\_entity+0x11c/0x5d0

[Fri Jul 26 09:53:12 2024] [<ffffffff8156ae29>] \_\_\_sys\_sendmsg+0x3a9/0x3c0

[Fri Jul 26 09:53:12 2024] [<ffffffff811b0b49>] ? do\_numa\_page+0x159/0x1e0

[Fri Jul 26 09:53:12 2024] [<ffffffff811b1f37>] ? handle\_mm\_fault+0x6b7/0xfa0

[Fri Jul 26 09:53:12 2024] [<ffffffff8120019d>] ? do\_sync\_read+0x8d/0xd0

[Fri Jul 26 09:53:12 2024] [<ffffffff816affd1>] ? \_\_do\_page\_fault+0x171/0x450

[Fri Jul 26 09:53:12 2024] [<ffffffff8156b5f1>] \_\_sys\_sendmsg+0x51/0x90

[Fri Jul 26 09:53:12 2024] [<ffffffff8156b642>] SyS\_sendmsg+0x12/0x20

[Fri Jul 26 09:53:12 2024] [<ffffffff816b4fc9>] system\_call\_fastpath+0x16/0x1b

[Fri Jul 26 09:53:12 2024] Code: 83 ec 38 48 89 55 b0 48 89 4d a8 65 48 8b 04 25 28 00 00 00 48 89 45 d0 31 c0 48 85 ff 0f 84 e9 02 00 00 0f 1f 40 00 49 8b 04 24 <49> c7 04 24 00 00 00 00 48 85 c0 48 89 45 b8 0f 95 45 c7 48 81

[Fri Jul 26 10:25:04 2024] NMI watchdog: BUG: soft lockup - CPU#21 stuck for 22s! [msgr-worker-2:16053]

[Fri Jul 26 10:25:04 2024] Modules linked in: udp\_diag xt\_conntrack ipt\_MASQUERADE nf\_nat\_masquerade\_ipv4 nf\_conntrack\_netlink nfnetlink xt\_addrtype iptable\_filter iptable\_nat nf\_conntrack\_ipv4 nf\_defrag\_ipv4 nf\_nat\_ipv4 nf\_nat nf\_conntrack br\_netfilter bridge overlay(T) ipmi\_poweroff ipmi\_watchdog tcp\_diag inet\_diag 8021q garp mrp stp llc bonding dm\_mirror dm\_region\_hash dm\_log dm\_mod vfat fat coretemp kvm\_intel kvm iTCO\_wdt iTCO\_vendor\_support irqbypass crc32\_pclmul ghash\_clmulni\_intel aesni\_intel lrw gf128mul glue\_helper ablk\_helper cryptd ipmi\_ssif pcspkr ses enclosure scsi\_transport\_sas joydev sg mei\_me mei i2c\_i801 shpchp wmi ipmi\_si ipmi\_devintf ipmi\_msghandler acpi\_power\_meter acpi\_cpufreq ip\_tables xfs libcrc32c sd\_mod crc\_t10dif crct10dif\_generic ast drm\_kms\_helper syscopyarea sysfillrect mlx5\_core sysimgblt

[Fri Jul 26 10:25:04 2024] fb\_sys\_fops ttm igb drm i40e crct10dif\_pclmul ahci crct10dif\_common crc32c\_intel libahci libata devlink dca nvme i2c\_algo\_bit nvme\_core ptp i2c\_core pps\_core megaraid\_sas(OE)

[Fri Jul 26 10:25:04 2024] CPU: 21 PID: 16053 Comm: msgr-worker-2 Tainted: G W OEL ------------ T 3.10.0-693.el7.x86\_64 #1

[Fri Jul 26 10:25:04 2024] Hardware name: Nettrix R620 G40/60WA32, BIOS 0PYH001052 04/11/2023

[Fri Jul 26 10:25:04 2024] task: ffff881f967d4f10 ti: ffff881f8c7d4000 task.ti: ffff881f8c7d4000

[Fri Jul 26 10:25:04 2024] RIP: 0010:[<ffffffff815c42f5>] [<ffffffff815c42f5>] ipv4\_dst\_check+0x5/0x30

[Fri Jul 26 10:25:04 2024] RSP: 0018:ffff88203f243d70 EFLAGS: 00000286

[Fri Jul 26 10:25:04 2024] RAX: ffff881f78246900 RBX: ffff880006415540 RCX: ffff880019815000

[Fri Jul 26 10:25:04 2024] RDX: ffffffff81adee40 RSI: 0000000000000000 RDI: ffff881f78246900

[Fri Jul 26 10:25:04 2024] RBP: ffff88203f243d90 R08: 0000000000000000 R09: 0000000000000000

[Fri Jul 26 10:25:04 2024] R10: 0000000000000000 R11: 0000000000000004 R12: ffff88203f243ce8

[Fri Jul 26 10:25:04 2024] R13: ffffffff816b5c1d R14: ffff88203f243d90 R15: ffff881f78246900

[Fri Jul 26 10:25:04 2024] FS: 00007f1cfbc99700(0000) GS:ffff88203f240000(0000) knlGS:0000000000000000

[Fri Jul 26 10:25:04 2024] CS: 0010 DS: 0000 ES: 0000 CR0: 0000000080050033

[Fri Jul 26 10:25:04 2024] CR2: 0000556d0e510000 CR3: 000000203e47c000 CR4: 00000000003407e0

[Fri Jul 26 10:25:04 2024] DR0: 0000000000000000 DR1: 0000000000000000 DR2: 0000000000000000

[Fri Jul 26 10:25:04 2024] DR3: 0000000000000000 DR6: 00000000fffe0ff0 DR7: 0000000000000400

[Fri Jul 26 10:25:04 2024] Stack:

[Fri Jul 26 10:25:04 2024] ffffffff8156c302 ffffffff813305f1 ffff8800362982f8 ffff880006415540

[Fri Jul 26 10:25:04 2024] ffff88203f243dc8 ffffffff815ce3ca ffff880006415540 ffff8800362982f8

[Fri Jul 26 10:25:04 2024] 0000000000000044 0000000000000020 0000000000300000 ffff88203f243e38

[Fri Jul 26 10:25:04 2024] Call Trace:

[Fri Jul 26 10:25:04 2024] <IRQ>

[Fri Jul 26 10:25:04 2024] [<ffffffff8156c302>] ? \_\_sk\_dst\_check+0x42/0xa0

[Fri Jul 26 10:25:04 2024] [<ffffffff813305f1>] ? csum\_partial+0x11/0x20

[Fri Jul 26 10:25:04 2024] [<ffffffff815ce3ca>] ip\_queue\_xmit+0x26a/0x3a0

[Fri Jul 26 10:25:04 2024] [<ffffffff815e7c0f>] tcp\_transmit\_skb+0x4cf/0x9b0

[Fri Jul 26 10:25:04 2024] [<ffffffff815e824a>] tcp\_write\_xmit+0x15a/0xce0

[Fri Jul 26 10:25:04 2024] [<ffffffff815e8e05>] tcp\_tsq\_handler.part.31+0x35/0x40

[Fri Jul 26 10:25:04 2024] [<ffffffff815e8fe8>] tcp\_tasklet\_func+0x108/0x140

[Fri Jul 26 10:25:04 2024] [<ffffffff8109084d>] tasklet\_action+0x7d/0x140

[Fri Jul 26 10:25:04 2024] [<ffffffff81090b3f>] \_\_do\_softirq+0xef/0x280

[Fri Jul 26 10:25:04 2024] [<ffffffff816b6a5c>] call\_softirq+0x1c/0x30

[Fri Jul 26 10:25:04 2024] <EOI>

[Fri Jul 26 10:25:04 2024] [<ffffffff8102d3c5>] do\_softirq+0x65/0xa0

[Fri Jul 26 10:25:04 2024] [<ffffffff8108ffdb>] \_\_local\_bh\_enable\_ip+0x9b/0xb0

[Fri Jul 26 10:25:04 2024] [<ffffffff81090007>] local\_bh\_enable+0x17/0x20

[Fri Jul 26 10:25:04 2024] [<ffffffff815cfaaf>] ip\_finish\_output+0x29f/0x780

[Fri Jul 26 10:25:04 2024] [<ffffffff815d0293>] ip\_output+0x73/0xe0

[Fri Jul 26 10:25:04 2024] [<ffffffff815cf810>] ? \_\_ip\_append\_data.isra.48+0xa00/0xa00

[Fri Jul 26 10:25:04 2024] [<ffffffff815cdf31>] ip\_local\_out\_sk+0x31/0x40

[Fri Jul 26 10:25:04 2024] [<ffffffff815ce2a3>] ip\_queue\_xmit+0x143/0x3a0

[Fri Jul 26 10:25:04 2024] [<ffffffff815e7c0f>] tcp\_transmit\_skb+0x4cf/0x9b0

[Fri Jul 26 10:25:04 2024] [<ffffffff815e824a>] tcp\_write\_xmit+0x15a/0xce0

[Fri Jul 26 10:25:04 2024] [<ffffffff815e904e>] \_\_tcp\_push\_pending\_frames+0x2e/0xc0

[Fri Jul 26 10:25:04 2024] [<ffffffff815d77dc>] tcp\_push+0xec/0x120

[Fri Jul 26 10:25:04 2024] [<ffffffff815db0f8>] tcp\_sendmsg+0xc8/0xc30

[Fri Jul 26 10:25:04 2024] [<ffffffff815d0293>] ? ip\_output+0x73/0xe0

[Fri Jul 26 10:25:04 2024] [<ffffffff81606893>] inet\_sendmsg+0x63/0xb0

[Fri Jul 26 10:25:04 2024] [<ffffffff8156a580>] sock\_sendmsg+0xb0/0xf0

[Fri Jul 26 10:25:04 2024] [<ffffffff810cdc7c>] ? dequeue\_entity+0x11c/0x5d0

[Fri Jul 26 10:25:04 2024] [<ffffffff8156ae29>] \_\_\_sys\_sendmsg+0x3a9/0x3c0

[Fri Jul 26 10:25:04 2024] [<ffffffff816a7722>] ? mutex\_lock+0x12/0x2f

[Fri Jul 26 10:25:04 2024] [<ffffffff8120019d>] ? do\_sync\_read+0x8d/0xd0

[Fri Jul 26 10:25:04 2024] [<ffffffff8156b5f1>] \_\_sys\_sendmsg+0x51/0x90

[Fri Jul 26 10:25:04 2024] [<ffffffff8156b642>] SyS\_sendmsg+0x12/0x20

[Fri Jul 26 10:25:04 2024] [<ffffffff816b4fc9>] system\_call\_fastpath+0x16/0x1b

[Fri Jul 26 10:25:04 2024] Code: 00 48 03 04 cd e0 fd b0 81 c3 0f 1f 40 00 31 c0 5d c3 0f 1f 40 00 0f 1f 44 00 00 55 48 89 e5 5d c3 0f 1f 44 00 00 0f 1f 44 00 00 <55> 31 c0 66 83 7f 66 ff 48 89 e5 75 1b 48 8b 57 18 48 8b 92 e8

[Fri Jul 26 10:59:21 2024] perf: interrupt took too long (34824 > 34755), lowering kernel.perf\_event\_max\_sample\_rate to 5000

[Fri Jul 26 16:21:08 2024] NMI watchdog: BUG: soft lockup - CPU#18 stuck for 22s! [msgr-worker-0:16051]

[Fri Jul 26 16:21:08 2024] Modules linked in: udp\_diag xt\_conntrack ipt\_MASQUERADE nf\_nat\_masquerade\_ipv4 nf\_conntrack\_netlink nfnetlink xt\_addrtype iptable\_filter iptable\_nat nf\_conntrack\_ipv4 nf\_defrag\_ipv4 nf\_nat\_ipv4 nf\_nat nf\_conntrack br\_netfilter bridge overlay(T) ipmi\_poweroff ipmi\_watchdog tcp\_diag inet\_diag 8021q garp mrp stp llc bonding dm\_mirror dm\_region\_hash dm\_log dm\_mod vfat fat coretemp kvm\_intel kvm iTCO\_wdt iTCO\_vendor\_support irqbypass crc32\_pclmul ghash\_clmulni\_intel aesni\_intel lrw gf128mul glue\_helper ablk\_helper cryptd ipmi\_ssif pcspkr ses enclosure scsi\_transport\_sas joydev sg mei\_me mei i2c\_i801 shpchp wmi ipmi\_si ipmi\_devintf ipmi\_msghandler acpi\_power\_meter acpi\_cpufreq ip\_tables xfs libcrc32c sd\_mod crc\_t10dif crct10dif\_generic ast drm\_kms\_helper syscopyarea sysfillrect mlx5\_core sysimgblt

[Fri Jul 26 16:21:08 2024] fb\_sys\_fops ttm igb drm i40e crct10dif\_pclmul ahci crct10dif\_common crc32c\_intel libahci libata devlink dca nvme i2c\_algo\_bit nvme\_core ptp i2c\_core pps\_core megaraid\_sas(OE)

再回到本次出现问题的堆栈信息，只要是业务线程挂住，一定会影响业务，但是走到这里，似乎没有什么思路了，一般来说，内核不会有什么太大的问题，大概率是用户态线程导致的一直常占着该CPU，但是现在开源ceph直接使用，也不太可能是软件bug，这个ceph版本是15.2.17，如果有bug，那所有的集群可能都会有，所以这个可能性也排除了，除非是系统资源的确吃紧；

继续查看监控数据 ， url如下：

[node - Monitor - Dashboards - Grafana](http://10.103.251.42:3000/d/_XG4IueGk/node?orgId=1&refresh=30s&var-prom=rg15_hfb3&var-host=172.29.224.90&from=now-7d&to=now)

翻看对应时间点2025/04/02 15:4

基本上没什么大的差异

继续分析

sar -f sa02 -b 输出如下： //磁盘io情况

Linux 3.10.0-693.el7.x86\_64 (rg52-blk009.ceph.csyc4.iflytek.net) 04/02/2025 \_x86\_64\_ (48 CPU)

12:00:01 AM tps rtps wtps bread/s bwrtn/s

02:40:01 PM 11364.37 6601.49 4762.88 476106.77 303694.29

02:50:01 PM 7102.06 2302.02 4800.04 129336.23 286431.24

03:00:01 PM 7962.71 3136.14 4826.58 216402.08 278229.06

03:10:01 PM 7028.90 2169.13 4859.77 132695.56 294837.45

03:20:01 PM 7271.27 2545.52 4725.75 176554.11 277082.07

03:30:01 PM 7912.54 2888.37 5024.17 187318.70 318583.96

03:40:01 PM 7178.80 2283.81 4894.99 143235.80 301669.24

03:50:01 PM 7605.33 2676.61 4928.73 179465.68 321998.47

04:00:01 PM 8053.85 3103.27 4950.58 204672.77 375928.91

04:10:01 PM 9756.20 5078.78 4677.42 327307.99 345773.23

04:20:01 PM 11888.91 7174.05 4714.86 475340.13 380406.65

04:30:01 PM 10616.96 6172.52 4444.45 388852.31 309944.47

04:40:01 PM 10338.32 5799.35 4538.97 367889.64 303770.97

04:50:01 PM 10211.19 5458.73 4752.45 362016.09 342044.27

05:00:01 PM 8360.23 3729.11 4631.12 250137.33 338823.32

05:10:01 PM 10563.93 5963.47 4600.46 396868.57 345212.53

从03:50:01 PM 来看TPS(每秒事务数) 也一般，不是很大，前面02:40:01 PM 都达到了11K，也没有出现集群异常的情况，bwrtn/s（写入吞吐量）约 321MB/s，bread/s (读吞吐量)约为179MB/s, 写和读和02:40:01 PM对比相对是较小的，说明此时io负载不是很高，io压力适中，未达瓶颈，这一点也和监控数据相一致。

sar -f sa02 -r //内存使用情况

Linux 3.10.0-693.el7.x86\_64 (rg52-blk009.ceph.csyc4.iflytek.net) 04/02/2025 \_x86\_64\_ (48 CPU)

12:00:01 AM kbmemfree kbmemused %memused kbbuffers kbcached kbcommit %commit kbactive kbinact kbdirty

03:00:01 PM 1613212 129867660 98.77 89269164 5515840 91492496 69.59 98120428 27703732 44

03:10:01 PM 1295480 130185392 99.01 89065392 5460040 91437224 69.54 99296052 26849988 812

03:20:01 PM 1375592 130105280 98.95 89164296 5461892 91433580 69.54 99008512 27066768 152

03:30:01 PM 1490944 129989928 98.87 89083084 5457772 91434108 69.54 98633476 27314552 48

03:40:01 PM 858260 130622612 99.35 89189460 5456932 91434156 69.54 99314592 27259984 140

03:50:01 PM 997800 130483072 99.24 89470228 5466832 91459604 69.56 99309560 27121140 160

04:00:01 PM 1217888 130262984 99.07 89301716 5463564 91433108 69.54 98866556 27355312 100

04:10:01 PM 681836 130799036 99.48 89263160 5490156 91434364 69.54 99326828 27412320 96

04:20:01 PM 1079016 130401856 99.18 89330248 5493404 91436320 69.54 98915668 27435880 112

看对应时间点03:50:01 PM kbcached（缓存）占用约 5.4GB，kbbuffers（缓冲区）高达 89GB，实际可用内存（kbmemfree + kbcached）约 6.4GB， 且 %commit（内存提交率）仅为 69.54%，无 OOM 风险。此时内存还是充足的，这一点和普罗米修斯监控数据是一致的。

但是从内核系统日志报错来看，有一个麦洛斯网卡的报错

[Wed Apr 2 15:45:06 2025] mlx5\_core sysimgblt fb\_sys\_fops ttm igb drm i40e crct10dif\_pclmul ahci crct10dif\_common crc32c\_intel libahci libata devlink dca nvme i2c\_algo\_bit nvme\_core ptp i2c\_core pps\_core megaraid\_sas(OE)

通过如下命令输出

[root@rg52-blk009 sa]# lspci | grep -i net

32:00.0 Ethernet controller: Intel Corporation Ethernet Controller X710 for 10GbE SFP+ (rev 02)

32:00.1 Ethernet controller: Intel Corporation Ethernet Controller X710 for 10GbE SFP+ (rev 02)

98:00.0 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)

98:00.1 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)

b1:00.0 Ethernet controller: Mellanox Technologies MT27710 Family [ConnectX-4 Lx] -----------------------> 和混闪研发测试环境172.30.240.24 相比，多了这个

b1:00.1 Ethernet controller: Mellanox Technologies MT27710 Family [ConnectX-4 Lx]

可以确认服务器上安装了 三种不同类型的网卡，具体型号和规格如下：

来自于ds

1. Intel X710 万兆网卡（10GbE SFP+）

PCIe 设备 ID：32:00.0 和 32:00.1

型号：Ethernet Controller X710 for 10GbE SFP+

规格：

速率：10 Gbps（万兆）

接口类型：SFP+（光纤模块或DAC线缆）

用途：高性能网络场景（如数据中心、存储网络）。

驱动模块：i40e（Intel 官方驱动）

2. Intel I350 千兆网卡（1GbE）

PCIe 设备 ID：98:00.0 和 98:00.1

型号：I350 Gigabit Network Connection

规格：

速率：1 Gbps（千兆）

接口类型：RJ45（电口）

用途：常规网络连接或管理口（如BMC/iDRAC）。

驱动模块：igb（Intel 官方驱动）

3. Mellanox ConnectX-4 Lx 高速网卡（25/40/50/100GbE）

PCIe 设备 ID：b1:00.0 和 b1:00.1

型号：MT27710 Family [ConnectX-4 Lx]

规格：

速率：支持 25/40/50/100 Gbps（具体速率取决于固件和模块）

接口类型：QSFP28（高速光纤）

用途：超低延迟、高吞吐场景（如HPC、AI集群、RDMA）。

驱动模块：mlx5\_core（NVIDIA/Mellanox 官方驱动

这里就很奇怪了，讯飞云存储单服务器都是单张网卡，为什么这里出现了三张网卡？ 再对比一下混闪研发测试环境172.30.240.24 如下所示：

[root@rbd\_perf\_24 ~]# lspci | grep -i net

4b:00.0 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)

4b:00.1 Ethernet controller: Intel Corporation I350 Gigabit Network Connection (rev 01)

4c:00.0 Ethernet controller: Intel Corporation Ethernet Controller X710 for 10GbE SFP+ (rev 02)

4c:00.1 Ethernet controller: Intel Corporation Ethernet Controller X710 for 10GbE SFP+ (rev 02)

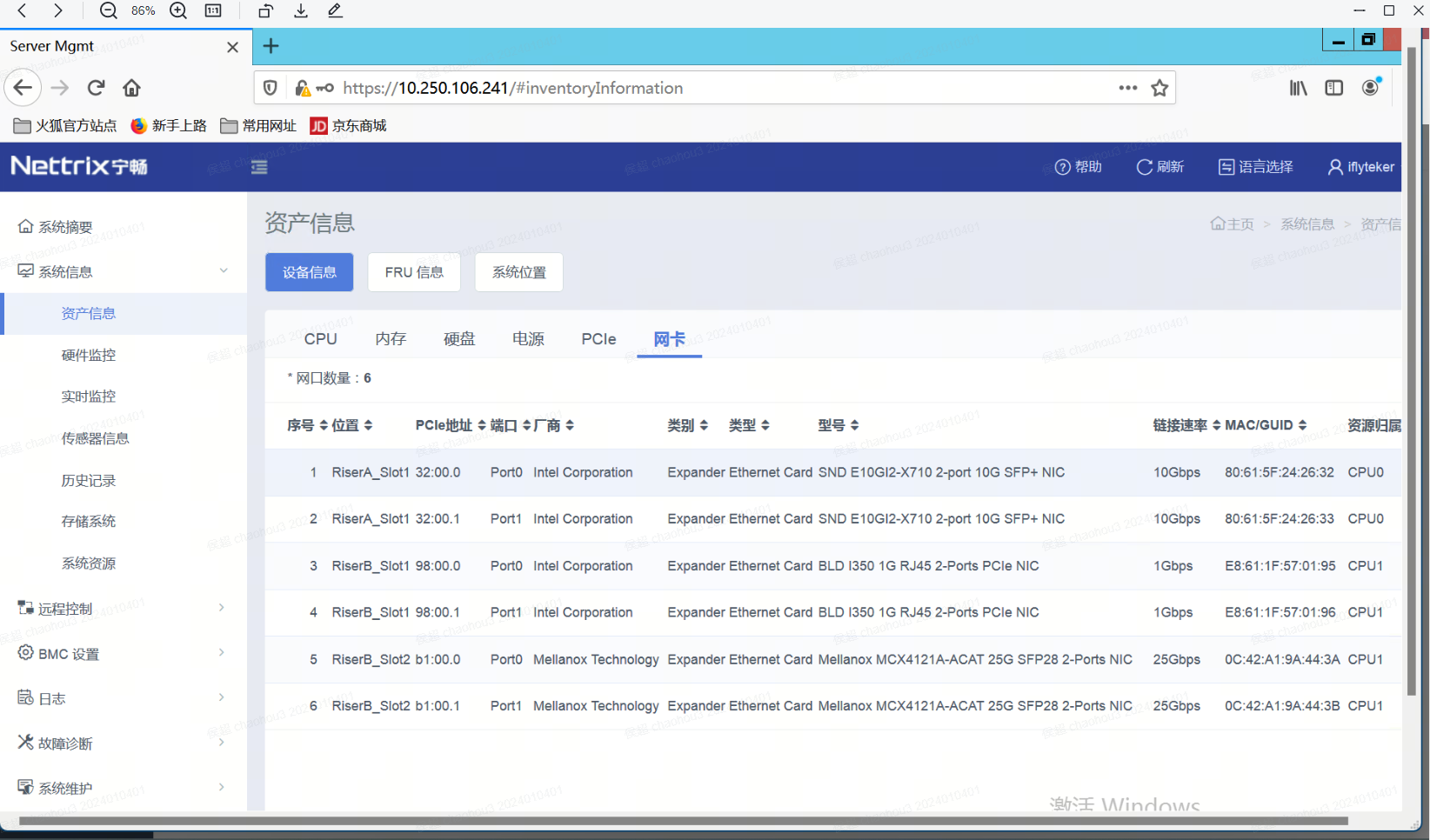
所以这里就有两个疑问点：

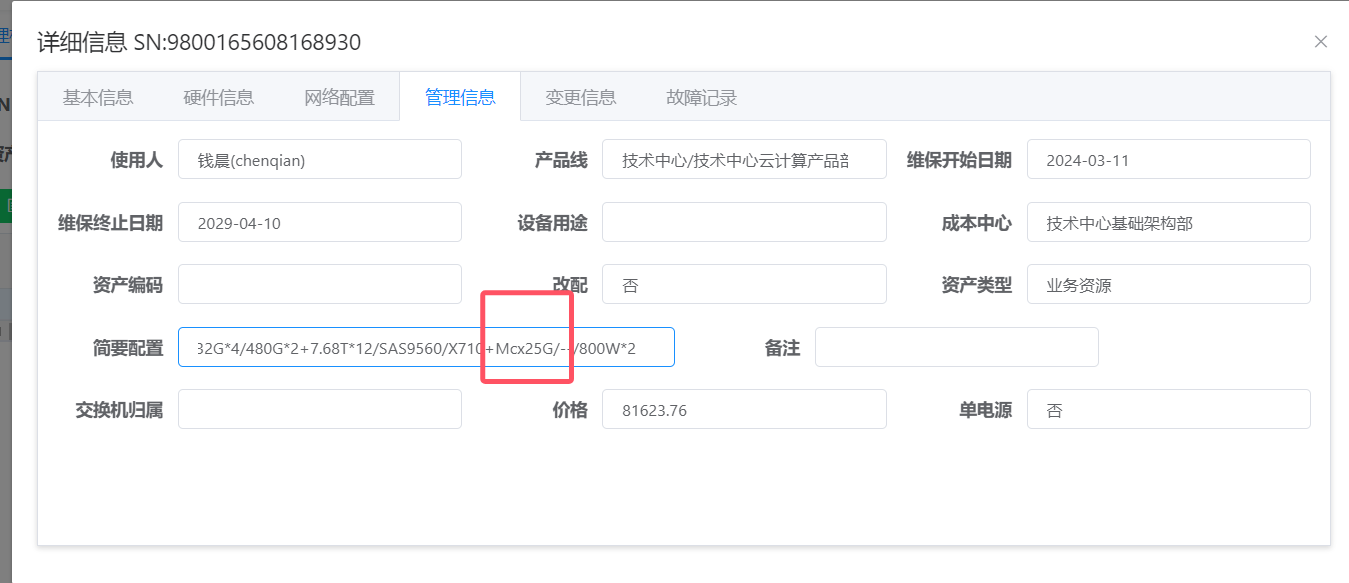
1. 讯飞云存储明明单服务器是单张网卡，为什么执行lspci却可以输出两张或三张网卡？
2. 出问题环境还多了一个高速网卡设备型号，而且内核打堆栈的也是这个 mlx5\_core， 只能怀疑这个环境不是很干净，曾经搞过别的东西，相关的驱动没有卸载干净就拿来做存储节点了

出问题环境SN号是9800165608168930

只要搞清楚这两个问题基本就清楚当前讯飞云存储网卡配置的问题了。

确认该节点有两张网卡，千兆网卡和万兆网卡





已联系曙光驻场 yaoyuankang，现在要确认如下两个问题：

1. 服务器出场是几张网卡，据我所知，咱们现在讯飞云存储都是单张网卡，如果是这样， 那就和照片不一致，多出的网卡是自己插上去的

服务器驻场工程师结论是这台服务器出厂配置是两张网卡，I350 千兆网卡和X710万兆网卡，只是我们自己为成本最小化，只用了万兆网卡，将两个万兆网口组bond0，连到一个交换机使用

1. 当前服务器节点是否有插麦洛斯网卡，需要现场确认

从CMDB里也确认这张mcx网卡是后面扩容的，属于出厂后，改配了机器，继续咨询外协值班同事(电话:18130066190) 当时提工单如下：





这个工单里没有明确指出mcx网卡驱动版本号，外协同事反馈可能是私下沟通确认，但是通过如下命令，可以确认mcx网卡驱动版本如下：

[root@rg52-blk009 osd.99\_bug]# cat /proc/net/bonding/bond0 | grep Sl --------> 该命令确认bond0口是25Gbmcx网卡的两个网口

Slave Interface: ens18f0

Slave queue ID: 0

Slave Interface: ens18f1

Slave queue ID: 0

[root@rg52-blk009 osd.99\_bug]# ethtool -i ens18f0 --------------------------> 查看网卡驱动版本信息

driver: mlx5\_core

version: 3.0-1 (January 2015)

firmware-version: 14.27.1016

expansion-rom-version:

bus-info: 0000:b1:00.0

supports-statistics: yes

supports-test: yes

supports-eeprom-access: no

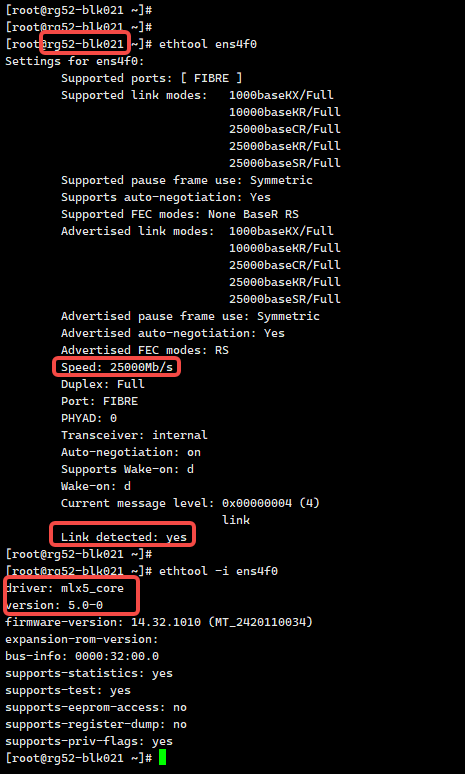
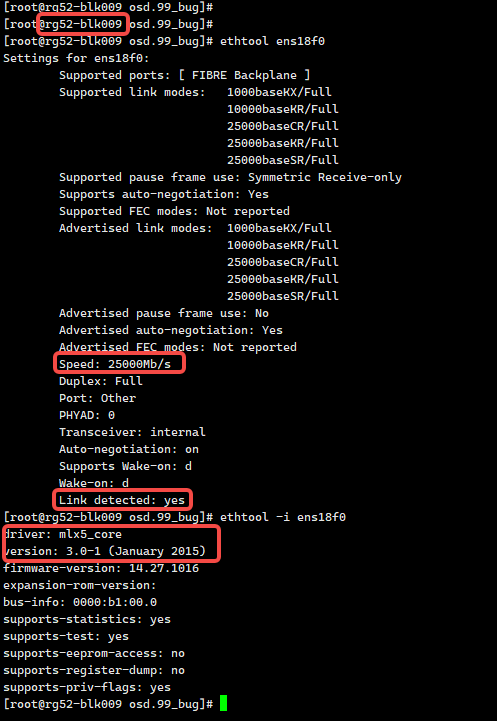
supports-register-dump: no

supports-priv-flags: yes

[root@rg52-blk009 osd.99\_bug]#

至此，可以确认当前节点使用的是25000Mb/s 麦洛斯网卡，并没有使用10GB X710网卡和千兆网卡。

所以问题大概是在麦洛斯网卡或网卡驱动这个方向上，继续查看一下问题节点172.29.224.90节点和172.29.224.139节点的25Gb/s麦洛斯网卡驱动版本，如下所示：



发现麦洛斯网卡驱动版本不一致，问题节点的驱动版本是3.0-1， 正常节点是5.0-0， 问题大概是网卡驱动版本导致的不兼容或者BUG。

## **4.问题结论**

1.该节点2025.04.02 15:45:06 时间点出现osd.99 marked me down，从内核日志报错信息来看，该osd进程网络线程msgr-worker-1一直卡在CPU37核上，此时CPU37核无法正常调度，导致网络线程一直卡着，无法退出，最后watchdog线程喂狗失败，打印soft lockup，网络线程一直卡住，该osd无法响应其他peer osd和mon节点，最终被mon标down

2.从sar命令分析OSD.99故障时磁盘io压力并不大，同时内存资源也较充足，未出现内存紧张情况，所以和当时该节点io压力和该节点的内存情况无关；

3.继续排查网络这条线索，协调外协和服务器驻场一起分析，确认当时节点172.29.224.90 一共是三张网卡，服务器厂商出厂提供的标配是两张网卡(千兆网卡和万兆网卡)，根据提单记录信息，后面业务自己改配加装了麦洛斯25Gb/s网卡，而我们现在组bond0的就是麦洛斯网卡；

4.通过对比分析172.29.224.90节点和172.29.224.139节点的25Gb/s麦洛斯网卡驱动版本，问题节点网卡驱动版本是version: 3.0-1 (January 2015)，正常节点是version: 5.0-0， 即问题节点网卡驱动版本太低。

## **5.问题方案**

建议升级一下节点172.29.224.90的麦洛斯网卡驱动版本，再观测问题现象是否存在。