

关于云计算部分产品网卡RDMA功能触发Linux内核BUG导致服务器重启问题的技术公告

CVM CVK H3Cloud CloudOS H3Cloud CAS OSD 孙亚华 2020-12-10 发表

问题描述

【产品型号】

H3C CAS
H3C ONESstor
H3C CloudOS
H3C CMP

【涉及版本】

CAS E0703、E0705版本 (CentOS) ;
ONESstor 3.0 R2130P01 (不含) 之前所有版本, 及E3113、E3115、E3115P01、E3116版本;
CloudOS 3.0 E3109 (不含) 之前所有版本, 5.0 E5102H01 (不含) 之前所有版本, 及由以上版本升级上来的版本;
CMP E1158 (不含) 之前所有版本;
UIS产品未安装RDMA软件包, 不会加载i40iw驱动, 不涉及该问题;
VDI系列产品:
| Workspace和Learning底层是UIS或E0706及以上版本的CAS, 不涉及该问题;
| Desktop、云学院和云学堂底层是ubuntu系统的CAS, 不涉及该问题;
大数据系列产品未安装RDMA软件包, 不会加载i40iw驱动, 不涉及该问题。

【问题描述】

相关产品部署在带有RDMA功能网卡的服务器上, 系统会自动加载该功能对应的i40iw驱动, 该驱动在运行过程中会概率性触发服务器重启。

原因分析

服务器重启后在系统的/var/crash/路径下会生成vmcore文件, 对该文件解析后会显示如下标红错误: "BUG: unable to handle paging request at ffffa97f8e9a7868"。

```
KERNEL: /home/g21593/kernel-build-4.14/RPMS/x86_64/usr/lib/debug/lib/modules/4.14.0-49.x86_64/vmlinux
DUMPFILE: vmcore [PARTIAL DUMP]
CPUS: 40
DATE: Wed Jan 15 13:50:06 2020
UPTIME: 124 days, 10:20:26
LOAD AVERAGE: 13.91, 12.78, 11.36
TASKS: 8165
NODENAME: R1A1-MGST1-3065
RELEASE: 4.14.0-49.el7.centos.x86_64 内核版本
VERSION: #1 SMP Wed Jul 11 07:57:04 UTC 2018
MACHINE: x86_64 (2200 Mhz)
MEMORY: 255.7 GB
PANIC: BUG: unable to handle paging request at ffffa97f8e9a7868 -bug类型
PID: 1604
COMMAND: "irqbalance" 出错的程序
TASK: ffff9bfd7c4bbc80 [THREAD_INFO: ffff9bfd7c4bbc80]
CPU: 19
STATE: TASK_RUNNING (PANIC)
```

系统加载i40iw驱动并在注册中断地址值时, 出现地址值错误。在相关产品的软件运行过程中, 如果有其他应用程序 (如irqbalance) 在读取该异常地址值时会触发内核bug, 导致内核崩溃、服务器重启。

规避措施/解决方案

规避方案:

若涉及的云计算产品未使用RDMA功能, 可手动卸载i40iw驱动规避该问题。具体操作步骤如下:

1、登录相关产品所有节点后台使用lsmod | grep i40iw命令确认是否安装i40iw驱动。若未安装该驱动, 不用做任何操作; 若有安装该驱动, 则执行第2-4步进行卸载。

```
[root@localhost ~]# lsmod | grep i40iw    ###Step1: 该命令确认是否有i40iw驱动
i40iw                184320  0          ###有该返回信息, 则表明有安装该驱动, 需要执行Step2操作
ib_core              229376  1 i40iw
i40e                 311296  1 i40iw
```

2、卸载i40iw驱动并检查是否卸载成功, 若lsmod | grep i40iw 命令无回显说明卸载成功。

```
[root@localhost ~]# rmmod i40iw          ###Step2: 卸载该驱动
[root@localhost ~]# lsmod | grep i40iw   ###检查是否卸载成功
```

3、使用modinfo i40iw命令获取该驱动文件存储路径信息, 下图中的下划线标注部分的信息即为该存储的绝对路径信息。

```
[root@ localhost ~]# modinfo i40iw ###Step3: 获取驱动绝对路径信息↵
filename:      /lib/modules/4.14.0-49.9.x86_64/kernel/drivers/infiniband/hw/i40iw/i40iw.ko↵
license:       Dual BSD/GPL↵
description:   Intel(R) Ethernet Connection X722 iWARP RDMA Driver↵
author:        Intel Corporation, <e1000-rdma@lists.sourceforge.net>↵
rhelversion:   7.5↵
srcversion:    BA81619670A93D0BEA4907C↵
depends:        ib_core,i40e↵
intree:        Y↵
name:          i40iw↵
vermagic:      4.14.0-49.9.x86_64 SMP mod_unload modversions ↵
```

4、修改rc.local文件，确保服务器重启后不会再次加载该驱动。操作命令如下：

chmod +x /etc/rc.d/rc.local

vi /etc/rc.d/rc.local

在rc.local文件内容的最后一行增加：rmmod <第3步获取到的路径信息>，如下截图：

```
[root@ localhost ~]# chmod +x /etc/rc.d/rc.local ###Step4: 修改文件权限↵
[root@ localhost ~]# vi /etc/rc.d/rc.local ###Step4: 修改文件，并保存（最后一行增加rmmod + step3获取到的路径信息）↵
[root@ localhost ~]# cat /etc/rc.d/rc.local ###修改后的文件见如下标红处↵
#!/bin/bash↵
# THIS FILE IS ADDED FOR COMPATIBILITY PURPOSES↵
#↵
# It is highly advisable to create own systemd services or udev rules↵
# to run scripts during boot instead of using this file.↵
#↵
# In contrast to previous versions due to parallel execution during boot↵
# this script will NOT be run after all other services.↵
#↵
# Please note that you must run 'chmod +x /etc/rc.d/rc.local' to ensure↵
# that this script will be executed during boot.↵
touch /var/lock/subsys/local↵
if [ ! -f /opt/tcpserver.log ]; then↵
    su root -c "sh /opt/tcpserver/install.sh" >>/opt/tcpserver.log 2>&1↵
fi↵
service firewalld stop↵
setMinionId > /dev/null &↵
tcpServer > /dev/null &↵
if [ ! -d "/var/log/matrix" ]; then↵
    mkdir /var/log/matrix↵
fi↵
if [ ! -f /opt/bin/link_templet.sh ]; then↵
    /bin/bash /opt/bin/common/link_templet.sh >>/var/log/matrix/set_link_templet.log 2>&1↵
fi↵
/opt/bin/common/iSCSILoginAtStart.sh &↵
/opt/bin/common/ServicesCheckAtStart.sh &↵
/opt/bin/common/ScheduleAtStart.sh &↵
rmmod /lib/modules/4.14.0-49.9.x86_64/kernel/drivers/infiniband/hw/i40iw/i40iw.ko↵
```

解决方案：

CAS 升级至E0706及以后版本；

ONESstor 升级至R2130P01或E3322（待发布）版本；

CloudOS 3.0的E3109（不含）之前版本及由之前版本升级上来的版本，或5.0的E5102H01（不含）之前版本及由之前版本升级上来的版本请按照上述规避方案处理。新安装的E3109及之后版本或E5102H01及之后版本不涉及该问题。