LINUX内核开发与调试

--内核转储

张银奎

2020/2

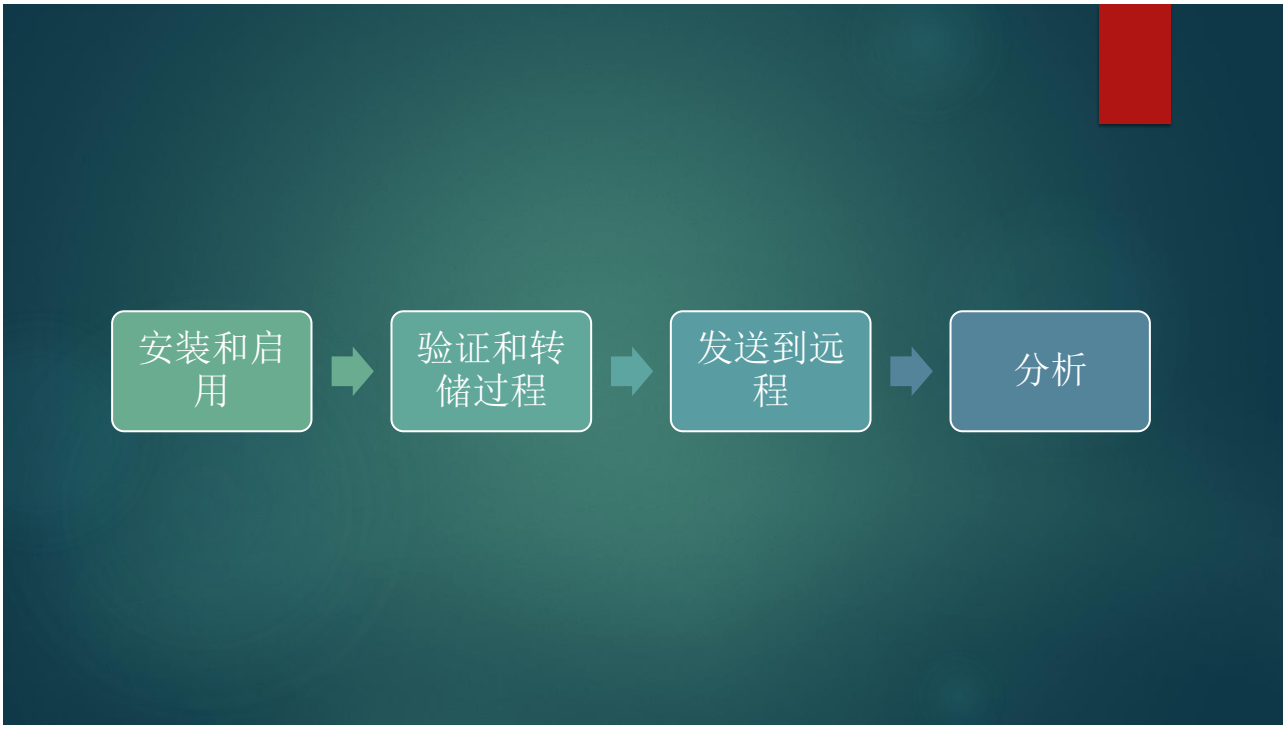
1

简历

- the LKCD (Linux Kernel Crash Dump) kernel patch from SGI in 1999
- Red Hat Netdump facility in 2002
- Red hat Diskdump facility in 2004
- Kdump crash dump facility in 2006

```
Copyright (C) 2002-2017  Red Hat, Inc.
Copyright (C) 2004, 2005, 2006, 2010  IBM Corporation
Copyright (C) 1999-2006  Hewlett-Packard Co
Copyright (C) 2005, 2006, 2011, 2012  Fujitsu Limited
Copyright (C) 2006, 2007  VA Linux Systems Japan K.K.
Copyright (C) 2005, 2011  NEC Corporation
Copyright (C) 1999, 2002, 2007  Silicon Graphics, Inc.
Copyright (C) 1999, 2000, 2001, 2002  Mission Critical Linux, Inc.
```

2

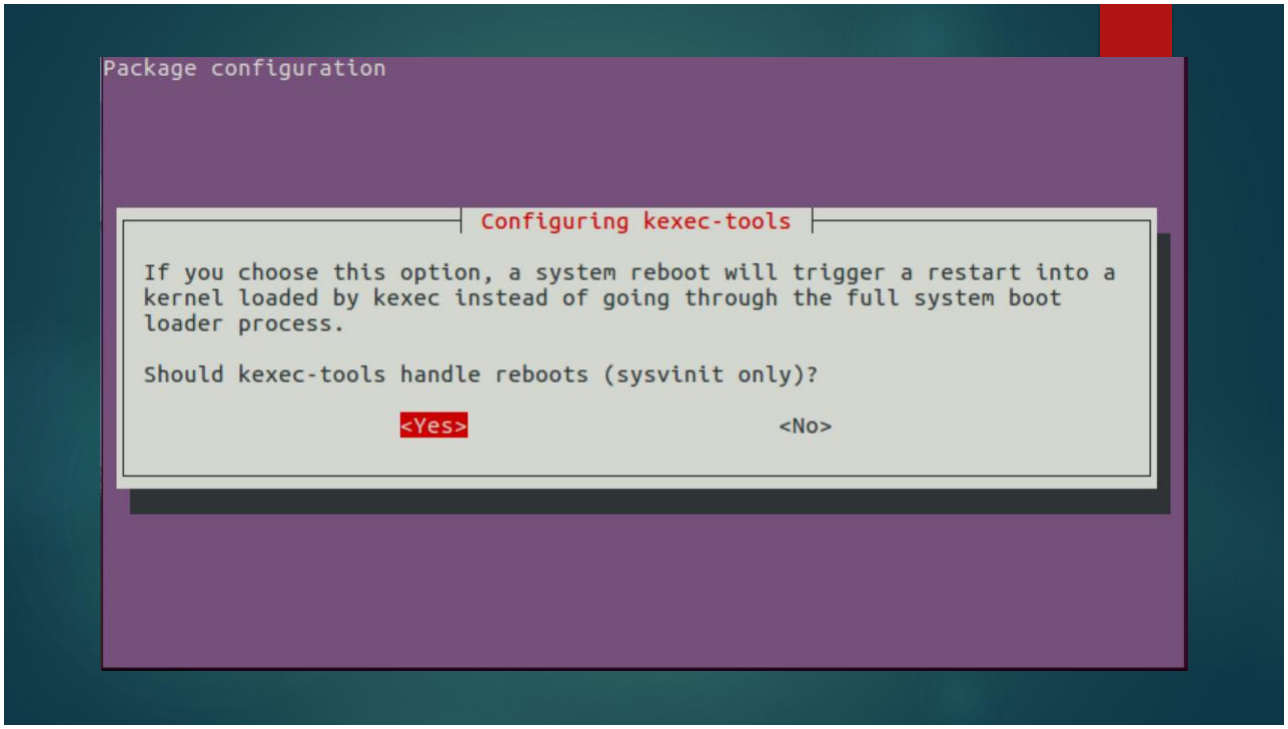


3

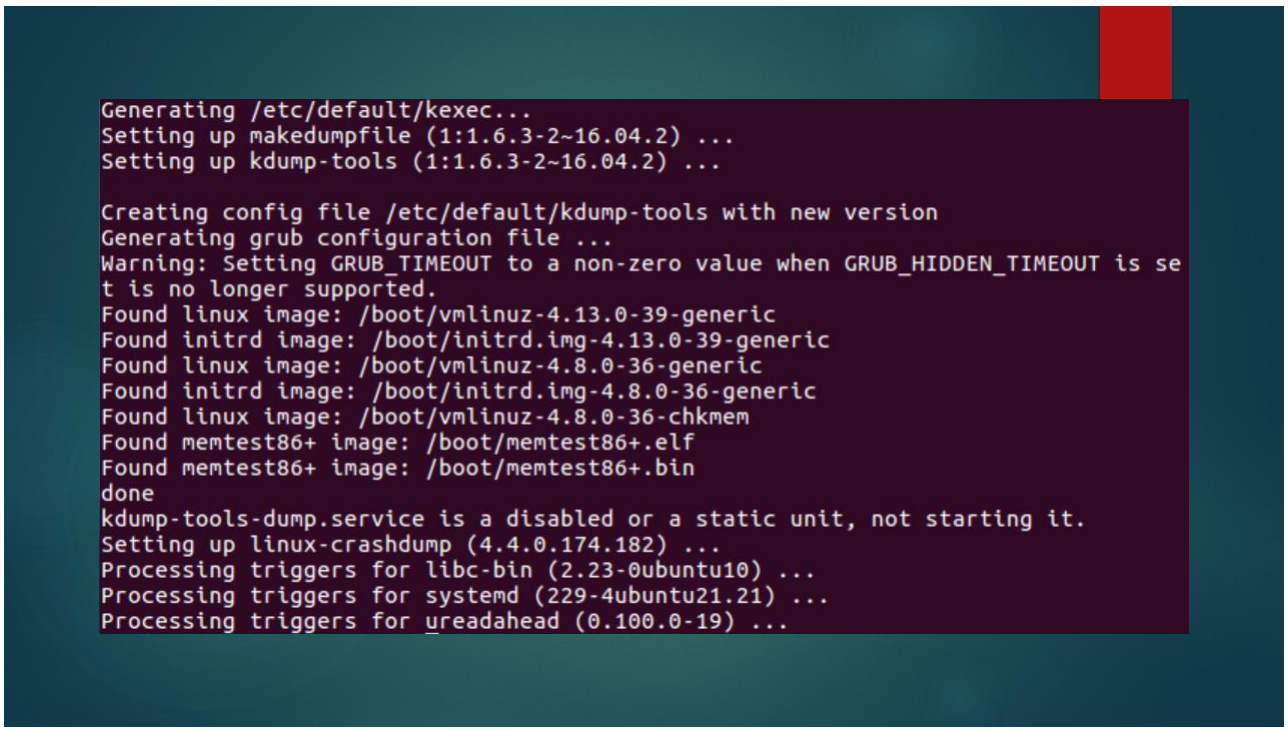
用途

- 严重的内核空间故障
  - 驱动程序
  - 内核本身
- KGDB/KDB不可用
- 希望永久保存事故现场

4


Package configuration
Configuring kexec-tools
If you choose this option, a system reboot will trigger a restart into a kernel loaded by kexec instead of going through the full system boot loader process.
Should kexec-tools handle reboots (sysvinit only)?
<Yes>
<No>

5


Generating /etc/default/kexec...
Setting up makedumpfile (1:1.6.3-2~16.04.2) ...
Setting up kdump-tools (1:1.6.3-2~16.04.2) ...
Creating config file /etc/default/kdump-tools with new version
Generating grub configuration file ...
Warning: Setting GRUB_TIMEOUT to a non-zero value when GRUB_HIDDEN_TIMEOUT is set is no longer supported.
Found linux image: /boot/vmlinuz-4.13.0-39-generic
Found initrd image: /boot/initrd.img-4.13.0-39-generic
Found linux image: /boot/vmlinuz-4.8.0-36-generic
Found initrd image: /boot/initrd.img-4.8.0-36-generic
Found linux image: /boot/vmlinuz-4.8.0-36-chkmem
Found memtest86+ image: /boot/memtest86+.elf
Found memtest86+ image: /boot/memtest86+.bin
done
kdump-tools-dump.service is a disabled or a static unit, not starting it.
Setting up linux-crashdump (4.4.0.174.182) ...
Processing triggers for libc-bin (2.23-0ubuntu10) ...
Processing triggers for systemd (229-4ubuntu21.21) ...
Processing triggers for ureadahead (0.100.0-19) ...

6

```
gedu@gedu-VirtualBox:~$ sudo apt install linux-crashdump
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  crash kdump-tools kexec-tools libsnappy1v5 makedumpfile
The following NEW packages will be installed:
  crash kdump-tools kexec-tools libsnappy1v5 linux-crashdump makedumpfile
0 upgraded, 6 newly installed, 0 to remove and 533 not upgraded.
Need to get 2,956 kB of archives.
After this operation, 9,059 kB of additional disk space will be used.
Do you want to continue? [Y/n]
Get:1 http://cn.archive.ubuntu.com/ubuntu xenial/main amd64 libsnappy1v5 amd64 1
.1.3-2 [16.0 kB]
Get:1 http://cn.archive.ubuntu.com/ubuntu xenial/main amd64 libsnappy1v5 amd64 1
.1.3-2 [16.0 kB]
Get:2 http://cn.archive.ubuntu.com/ubuntu xenial-updates/main amd64 crash amd64
7.2.3+real-1~16.04.1 [2,680 kB]
5% [2 crash 47.9 kB/2,680 kB 2%]
```

7

安装系统时安装

```
|------------------------| Configuring kdump-tools |------------------------|
|                                                                            |
|                                                                            |
| If you choose this option, the kdump-tools mechanism will be enabled. A    |
| reboot is still required in order to enable the crashkernel kernel         |
| parameter.                                                                 |
|                                                                            |
| Should kdump-tools be enabled by default?                                  |
|                                                                            |
|                     <Yes>                          <No>                    |
|                                                                            |
|----------------------------------------------------------------------------|
```

- Ubuntu 16.04开始安装程序可以自动安装kdump
- 手工启用：dpkg-reconfigure kdump-tools

8

命令行参数

- gedu@gedu-VirtualBox:~$ cat /proc/cmdline
- BOOT_IMAGE=/boot/vmlinuz-4.13.0-39-generic root=UUID=16b3f987-37b9-4b8f-b049-be9663cab9bc ro text quiet splash pti=off crashkernel=384M-:128M
- crashkernel=<range1>:<size1>[,<range2>:<size2>,...][@offset]
- range=start-[end] 'start' is inclusive and 'end' is exclusive.

- if the RAM is smaller than 384M, then don't reserve anything (this is the "rescue" case)
- if the RAM size is between 386M and 2G (exclusive), then reserve 64M
- if the RAM size is larger than 2G, then reserve 128M

9

从内核消息看实际情况

[0.000000] Reserving 128MB of memory at 528MB for crashkernel (System RAM: 1023MB)

- dmesg | grep crash

10

kdump-config show

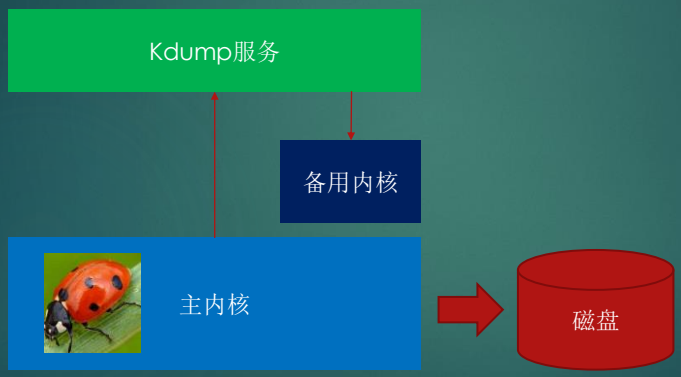
```
DUMP_MODE:        kdump
USE_KDUMP:        1
KDUMP_SYSCTL:     kernel.panic_on_oops=1
KDUMP_COREDIR:    /var/crash
crashkernel addr: 0x
  /var/lib/kdump/vmlinuz: symbolic link to /boot/vmlinuz-4.13.0-39-generic
kdump initrd:
  /var/lib/kdump/initrd.img: symbolic link to /var/lib/kdump/initrd.img-4.13.0-39-generic
current state:    ready to kdump

kexec command:
  /sbin/kexec -p --command-line="BOOT_IMAGE=/boot/vmlinuz-4.13.0-39-generic
root=UUID=16b3f987-37b9-4b8f-b049-be9663cab9bc ro text quiet splash pti=off reset_devices
nr_cpus=1 systemd.unit=kdump-tools-dump.service irqpoll nousb ata_piix.prefer_ms_hyperv=0" --
initrd=/var/lib/kdump/initrd.img /var/lib/kdump/vmlinuz
```

11

12

工作过程



1. 主内核Panic
2. 内核中的Kdump支持，通知Kdump服务
3. Kdump服务通过Kexec启动备用内核
4. 备用内核将主内核的内存数据转储到磁盘

13

Kexec

- 在当前内核环境下启动另一个内核
- 相当于特殊的boot loader
- 跳过BIOS过程，有人用作快速重启

加载：
kexec -l /boot/vmlinuz-2.6.18-194.11.4.el5
--initrd=/boot/initrd-2.6.18-194.11.4.el5.img
--command-line="$(cat /proc/cmdline)"

执行：
Kexec -e

14

备胎

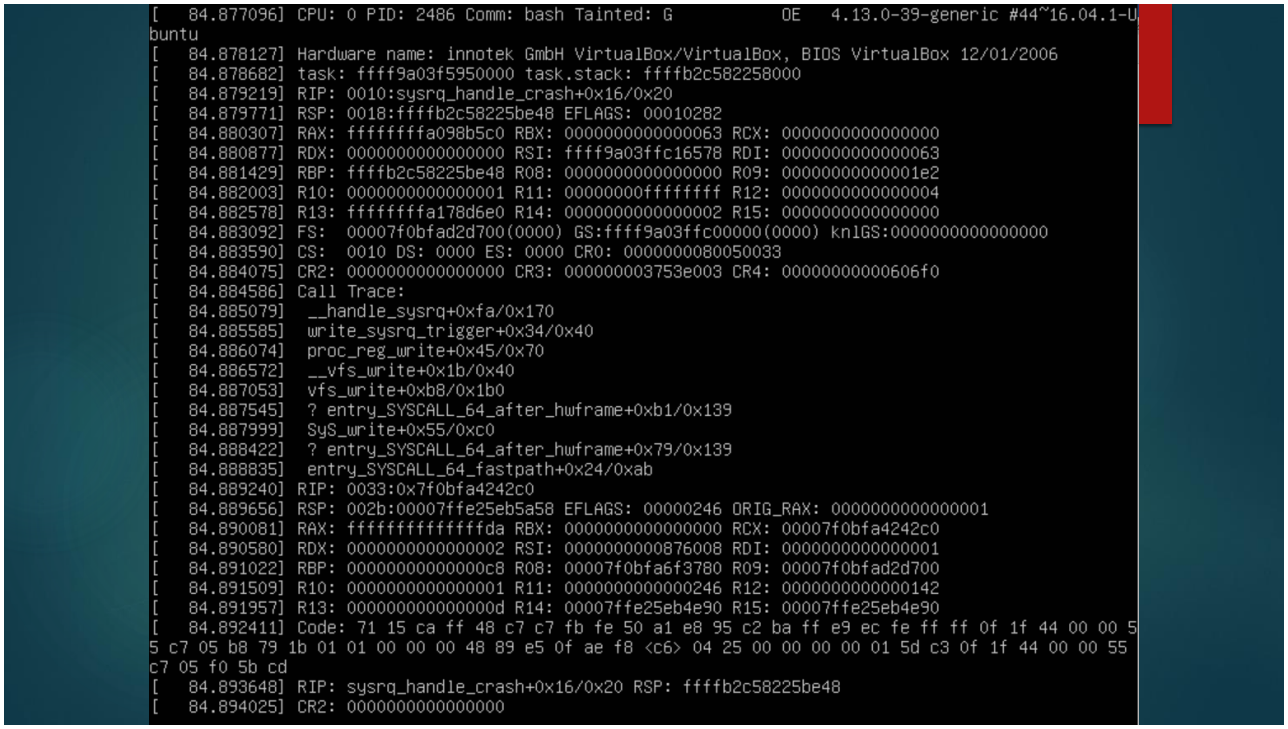
```
gedu@gedu-VirtualBox:/var/lib/kdump$ ll
total 21532
drwxr-xr-x  3 root root     4096 3月   5 21:01 ./
drwxr-xr-x 69 root root     4096 3月   4 23:15 ../
drwxr-xr-x  5 root root     4096 4月  22  2019 initramfs-tools/
lrwxrwxrwx  1 root root       43 3月   5 21:01 initrd.img -> /var/lib/kdump/initrd.img-4.13.0-
39-generic
-rw-r--r--  1 root root 22035913 3月   4 23:15 initrd.img-4.13.0-39-generic
lrwxrwxrwx  1 root root       31 3月   5 21:01 vmlinuz -> /boot/vmlinuz-4.13.0-39-generic
```

15

Please fasten your belt !

- sudo sysctl -w kernel.sysrq=1
- sudo –s
- # echo c > /proc/sysrq-trigger

16



17

- GRUB_CMDLINE_LINUX_DEFAULT="$GRUB_CMDLINE_LINUX_DEFAULT crashkernel=384M-:512M“

```
gedu@gedu-VirtualBox:/etc/default/grub.d$ cat kdump-tools.cfg
GRUB_CMDLINE_LINUX_DEFAULT="$GRUB_CMDLINE_LINUX_DEFAULT crashkernel=384M-:512M"
```

18

19

远程

```
# ----------------------------------------------------------------------
# Remote dump facilities:
# SSH - username and hostname of the remote server that will receive the dump
#       and dmesg files.
# SSH_KEY - Full path of the ssh private key to be used to login to the remote
#           server. use kdump-config propagate to send the public key to the
#           remote server
# HOSTTAG - Select if hostname of IP address will be used as a prefix to the
#           timestamped directory when sending files to the remote server.
#           'ip' is the default.
SSH="ubuntu@kdump-netcrash"
```

20

21

https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Deployment_Guide/s1-kdump-crash.html

32.3. ANALYZING THE CORE DUMP

To determine the cause of the system crash, you can use the **crash** utility, which provides an interactive prompt very similar to the GNU Debugger (GDB). This utility allows you to interactively analyze a running Linux system as well as a core dump created by `netdump`, `diskdump`, `xendump`, or `kdump`.

Important

To analyze the `vmcore` dump file, you must have the crash and kernel-debuginfo packages installed. To install the crash package in your system, type the following at a shell prompt as `root`:

```
yum install crash
```

To install the kernel-debuginfo package, make sure that you have the yum-utils package installed and run the following command as `root`:

```
debuginfo-install kernel
```

22

两种模式


```
Usage:

  crash [OPTION]... NAMELIST MEMORY-IMAGE[@ADDRESS]     (dumpfile form)
  crash [OPTION]... [NAMELIST]                          (live system form)
```

23

都需要符号

- 称为NAMELIST
- -g编译内核

```
Usage:

  crash [OPTION]... NAMELIST MEMORY-IMAGE[@ADDRESS]     (dumpfile form)
  crash [OPTION]... [NAMELIST]                          (live system form)
```

24

Red Hat的符号支持

In RHEL3 installations, the vmlinux file associated with the running kernel is split into two files, a stripped version found in the /boot directory; which has have the operating system release string appended to it, for example, vmlinux-2.4.21-4.ELsmp. The stripped file in /boot contains a link to its associated debuginfo file, which is located in the /usr/lib/debug/boot directory.

In RHEL4, RHEL5 and RHEL6 installations, the vmlinux file is part of the kernel debuginfo package, and is found in the relevant /usr/lib/debug/lib/modules/<release> dir

25

```
GNU gdb (GDB) 7.6
Copyright (C) 2013 Free Software Foundation, Inc.
License GPLv3+: GNU GPL version 3 or later <http://gnu.org/licenses/gpl.html>
This is free software: you are free to change and redistribute it.
There is NO WARRANTY, to the extent permitted by law.  Type "show copying"
and "show warranty" for details.
This GDB was configured as "x86_64-unknown-linux-gnu"...

crash: read error: kernel virtual address: ffffffff81e3b380  type: "page_offset_base"
crash: this kernel may be configured with CONFIG_STRICT_DEVMEM, which
       renders /dev/mem unusable as a live memory source.
crash: trying /proc/kcore as an alternative to /dev/mem

      KERNEL: vmlinux_gedu
    DUMPFILE: /proc/kcore
        CPUS: 1
        DATE: Sat Apr  4 22:18:32 2020
      UPTIME: 00:05:34
LOAD AVERAGE: 1.23, 0.52, 0.19
       TASKS: 379
    NODENAME: gedu-VirtualBox
     RELEASE: 4.8.0-36-generic
     VERSION: #36~16.04.1-Ubuntu SMP Sun Mar 3 16:40:19 CST 2019
     MACHINE: x86_64  (1992 Mhz)
      MEMORY: 3 GB
         PID: 2483
     COMMAND: "crash"
        TASK: ffff8800b81ac880  [THREAD_INFO: ffff8800b6bc4000]
         CPU: 0
       STATE: TASK_RUNNING (ACTIVE)

crash>
```

26

命令一览

```
crash> help

*               files           mod             runq            union
alias           foreach         mount           search          vm
ascii           fuser           net             set             vtop
bt              gdb             p               sig             waitq
btop            help            ps              struct          whatis
dev             irq             pte             swap            wr
dis             kmem            ptob            sym             q
eval            list            ptov            sys
exit            log             rd              task
extend          mach            repeat          timer

crash version: 4.0-8.11    gdb version: 6.1
For help on any command above, enter "help <command>".
For help on input options, enter "help input".
For help on output options, enter "help output".

crash>
```

27

基本用法

- crash <内核文件> <转储文件>
- crash vmlinux-2.4.20-2.1.15.entsmp vmcore

28

常用命令

- > h 命令执行历史
- > ps 显示进程
- > rd 观察内存
- > p 打印变量， pd十进制，px 十六进制
- > files 显示打开文件
- > bt 栈回溯

29

打开文件

```
crash> files
PID: 2483   TASK: ffff8800b81ac880  CPU: 0   COMMAND: "crash"
ROOT: /    CWD: /home/gedu/labs/linux-source-4.8.0
 FD       FILE            DENTRY           INODE       TYPE PATH
  0 ffff8800b944a500 ffff8800ba664000 ffff8800b9b85030 CHR  /dev/pts/18
  1 ffff8800b944a500 ffff8800ba664000 ffff8800b9b85030 CHR  /dev/pts/18
  2 ffff8800b944a500 ffff8800ba664000 ffff8800b9b85030 CHR  /dev/pts/18
  3 ffff8800b804cd00 ffff8800baeb5b40 ffff8800bb0123c0 CHR  /dev/null
  4 ffff8800b4d7df00 ffff8800b9a07c00 ffff8800ba68bc10 REG  /proc/kcore
  5 ffff8800b4d7d000 ffff8800b40bdc00 ffff8800b40d8958 REG  /home/gedu/labs/linux-source-4.8.0/vmlinux_gedu
  6 ffff8800b8e1d900 ffff8800b7e44b40 ffff8800ba6304a0 FIFO
  7 ffff8800b8e1de00 ffff8800b7e44b40 ffff8800ba6304a0 FIFO
  8 ffff8800b8680700 ffff8800b9a073c0 ffff8800b99e37c0 REG  /tmp/tmpfiBlboH
 10 ffff8800b8cefd00 ffff8800b7947cc0 ffff8800ba20b280 FIFO
```

30

```
crash> dev
CHRDEV    NAME                CDEV            OPERATIONS
   1      mem              ffff8800bb12f980  memory_fops
   4      /dev/vc/0        ffffffff8221bec0  console_fops
   4      tty              ffff8800bb12f700  tty_fops
   4      ttyS             ffff8800bb1c2000  tty_fops
   5      /dev/tty         ffffffff8221ad80  tty_fops
   5      /dev/console     ffffffff8221ad00  console_fops
   5      /dev/ptmx        ffffffff8221b000  ptmx_fops
   5      ttyprintk        ffff8800b3ebda00  tty_fops
   6      lp               ffff8800357c1800  lp_fops
   7      vcs              ffff8800bb12f400  vcs_fops
  10      misc             ffff8800bb05c480  misc_fops
  13      input            ffff8800b3d54378  joydev_fops
  21      sg               ffff88003549b480  sg_fops
  29      fb               ffff8800bb05c880  fb_fops
  89      i2c              ffff8800b8507b60  i2cdev_fops
  99      ppdev            ffff8800356b4d80  pp_fops
 108      ppp              ffff8800b3f87c80  ppp_device_fops
 116      alsa             ffff8800bb1bc900  snd_fops
 128      ptm              ffff8800bb1c2f80  tty_fops
 136      pts              ffff8800bb1c2180  tty_fops
 180      usb              ffff8800bb05cb80  usb_fops
 189      usb_device       ffffffff82228620  usbdev_file_operations
 204      ttyMAX                (none)
 226      drm              ffff8800b8510e80  drm_stub_fops
 246      aux              ffff8800b8510880  auxdev_fops
 247      hidraw           ffffffffc00ce3c0  hidraw_ops
 248      bsg              ffffffff82214ae0  bsg_fops
 249      watchdog              (none)
 250      rtc              ffff8800354b5320  rtc_dev_fops
 251      dimmctl          ffff8800bb05c180  nvdimm_fops
 252      ndctl            ffff8800bb05cf80  nvdimm_bus_fops
 253      tpm                   (none)
```

31

进程列表

```
crash> ps
   PID    PPID  CPU   TASK    ST  %MEM   VSZ   RSS  COMM
     0      0   0  c030a000  RU   0.0     0     0  [swapper]
     1      0   0  cff98000  IN   0.2  1412   468  init
     2      1   0  c1446000  IN   0.0     0     0  [keventd]
     3      1   0  cfffa000  IN   0.0     0     0  [kapm-idled]
     4      0   0  cfff8000  IN   0.0     0     0  [ksoftirqd_CPU0]
     5      0   0  cffee000  IN   0.0     0     0  [kswapd]
     6      0   0  cffec000  IN   0.0     0     0  [kreclaimd]
     7      0   0  c1826000  IN   0.0     0     0  [bdflush]
     8      0   0  c1824000  IN   0.0     0     0  [kupdated]
     9      1   0  cff90000  IN   0.0     0     0  [mdrecoveryd]
    13      1   0  cf07a000  IN   0.0     0     0  [kjournald]
    89      1   0  ce804000  IN   0.0     0     0  [khubd]
   184      1   0  ce4d4000  IN   0.0     0     0  [kjournald]
   572      1   0  cd938000  IN   0.0   440    48  dhcpcd
   637      1   0  ce4a4000  IN   0.2  1476   612  syslogd
   642      1   0  cd92c000  IN   0.2  2092   432  klogd
   663      1   0  ce2bc000  IN   0.2  1564   612  portmap
   691      1   0  cd84a000  IN   0.3  1652   668  rpc.statd
   803      1   0  cd756000  IN   0.2  1400   452  apmd
   828      1   0  cd6c2000  IN   0.3 18024   684  ypbind
   830    828   0  cd76e000  IN   0.3 18024   684  ypbind
   831    830   0  cd71c000  IN   0.3 18024   684  ypbind
-- MORE --  forward: <SPACE>, <ENTER> or j  backward: b or k  quit: q
```

32

White Paper: Red Hat Crash Utility

by David Anderson
<anderson@redhat.com>

Content

https://people.redhat.com/anderson/crash_whitepaper/

33

redhat

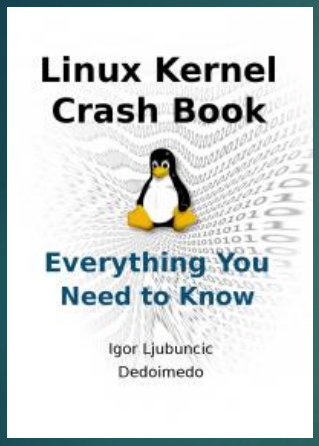
Basic Kernel Crash Dump Analysis

Hardik Vyas
Technical Support Engineer (Kernel)
19/04/2016

https://people.redhat.com/hvyas/Basics_of_Kernel_Panic_Hang_and_%20Kdump.pdf

34

Igor Ljubuncic aka Dedoimedo



- https://www.dedoimedo.com/computers/crash-analyze.html
- four technical works, eight novels, and five anthologies

35

Lister - [c:\bench\linux-4.4.14\Documentation\kdump\gdbmacros.txt]

File Edit Options Help

```
        printf "Trapno %ld, cr2 0x%lx, error_code %ld\n", $pid_task.thread.trap_no, \
                                $pid_task.thread.cr2, $pid_task.thread.error_code

end
document trapinfo
        Run info threads and lookup pid of thread #1
        'trapinfo <pid>' will tell you by which trap & possibly
        address the kernel panicked.
end


define dmesg
        set $i = 0
        set $end_idx = (log_end - 1) & (log_buf_len - 1)

        while ($i < logged_chars)
                set $idx = (log_end - 1 - logged_chars + $i) & (log_buf_len - 1)

                if ($idx + 100 <= $end_idx) || \
                   ($end_idx <= $idx && $idx + 100 < log_buf_len)
                        printf "%.100s", &log_buf[$idx]
                        set $i = $i + 100
                else
                        printf "%c", log_buf[$idx]
                        set $i = $i + 1
                end
        end
end
document dmesg
        print the kernel ring buffer
end
```

GDB macros

36

Can't use kdump or kexec for Linux virtual machines on Hyper-V

- This issue occurs because Hyper-V cannot host two simultaneous connections from the same synthetic driver that's running inside a virtual machine.
- When kdump is configured on a Linux virtual machine that's using the Linux Integration Services synthetic storage driver (also known as storvsc), the kexec kernel is configured to use the same driver. If the Linux virtual machine crashes, the synthetic storage driver that's hosted in the kexec kernel tries to open a connection to the Hyper-V storage provider. However, Hyper-V fails to establish the new connection because of the pre-existing connection to the same storage driver on the crashed Linux virtual machine. Therefore, the kexec kernel cannot dump the core for the crashed Linux virtual machine.

https://support.microsoft.com/en-us/help/2858695/can-t-use-kdump-or-kexec-for-linux-virtual-machines-on-hyper-v

37

切问而近思

欢迎关注格友公众号

38