CRIU Introduction

Chao Ye
cye@redhat.com

Outline

- CRIU
  - What’s CRIU
  - Under the hood
  - CRIU cmdline usage
  - CRIU usage scenarios
  - What cannot be checkpointed
  - CRIU support
- CRIU x Container
  - CRIU Demo
  - Containerized App support status

Introduce CRIU

What’s CRIU

Checkpoint/Restore In Userspace, or CRIU (pronounced kree-oo, IPA: /krɪʊ/, Russian: криу), is a software tool for Linux operating system. Using this tool, you can freeze a running application (or part of it) and checkpoint it to a hard drive as a collection of files. You can then use the files to restore and run the application from the point it was frozen at. The distinctive feature of the CRIU project is that it is mainly implemented in user space.

https://criu.org/Main_Page

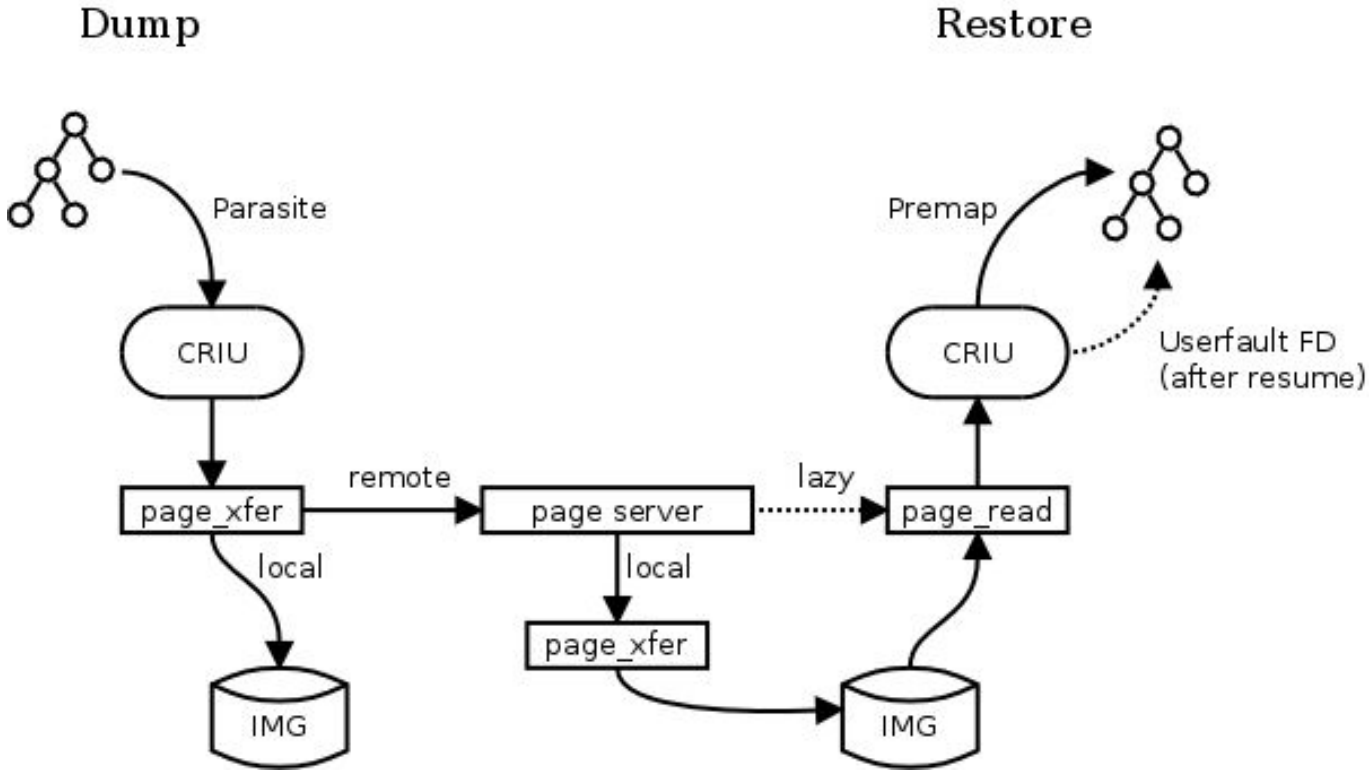
Introduce CRIU

Under the hood

- Kernel
  - ptrace
  - mmap
  - parasite code
- Userspace
  - criu

https://criu.org/Category:Under_the_hood

Introduce CRIU



Introduce CRIU

CRIU cmdline Usage:

criu dump|pre-dump -t PID [<options>]

criu restore [<options>]

criu check [--feature FEAT]

criu exec -p PID <syscall-string>

criu page-server

criu service [<options>]

criu dedup

Introduce CRIU

CRIU usage scenarios

- Live migration
- Load balance
- High availability
- ...

Introduce CRIU

What cannot be checkpointed

- Dumped with special option
  - External resources
  - File locks
  - Invisible files
- Cannot be dumped (yet)
  - Devices
  - Tasks with debugger attached
  - Task from a different user (for non-root)
  - Task running in compat mode (x86-64)
  - Sockets other than TCP, UDP, UNIX, packet and netlink
  - Packetized pipes
  - Cork-ed UDP sockets
  - Files sent over unix sockets
  - Half-opened UNIX connections
  - SysVIPC memory segment w/o IPC namespace

Introduce CRIU

CRIU Architecture support

- x86_64, Intel/AMD
- PPC64LE, IBM Power 8
- ARM, AARCH64

Enterprise Linux Support

- RHEL: TechPreview Since RHEL 7.2
- SLE: Kernel support enabled on SLE12-SP2
- Ubuntu: Live Migration in LXD

CRIU x Container

- Application use various network, filesystem, cgroups, etc.
- Baremetal OS env is too complicated
- Containerized application are isolated with the rest of OS/Applications
- Contianerized applications need similar KVM live migration technolgy

CRIU Demo

- Checkpoint/Restore application on Baremetal OS
- Checkpoint/Restore application with TCP connection
- Checkpoint/Restore VNC Server
- Checkpoint/Restore runc/Docker container

CRIU Demo - eatmem

[root@Fedora24-Server eatmem]# sh demo-eatmem.sh checkpoint
1
2
3
4
5
6 < ===== Checkpointed

[root@Fedora24-Client eatmem]# sh demo-eatmem.sh restore
7 < ===== Restored from image
8
9
10

CRIU Demo - eatmem

[root@Fedora24-Server eatmem]# sh demo-eatmem.sh checkpoint
1
2
3
4
5
6 < ===== Checkpointed

[root@Fedora24-Client eatmem]# sh demo-eatmem.sh restore
7 < ===== Restored from image
8
9
10

CRIU Demo - eatmem

```
[root@Fedora24-Server eatmem]# crit show pstree.img
{
    "magic": "PSTREE",
    "entries": [
        {
            "pid": 1183,
            "ppid": 0,
            "pgid": 1177,
            "sid": 1108,
            "threads": [
                1183
            ]
        }
    ]
}
```

CRIU Demo - vsftpd

```
[root@Fedora24-Server vsftpd]# sh demo-vsftpd.sh checkpoint
criu dump --images-dir /demo/images/vsftpd --shell-job --ext-unix-sk
--tcp-established --file-locks --tree $(ps -A | grep vsftpd | awk '{print $1}')
0aa71e0d2c3a1daff204d8c51ad98b4b
/var/ftp/pub/Fedora-Workstation-netinst-x86_64-24-1.2.iso
< ===== Checkpointed

[root@Fedora24-Client vsftpd]# sh demo-vsftpd.sh download
Fedora-Workstation-netinst-x86_64-24-1.2.iso                    21%[=====>
]  94.94M  --.-KB/s    eta 62s
```

CRIU Demo - vsftpd

[root@Fedora24-Server vsftpd]# sh demo-vsftpd.sh restore
  1276: Window parameters are not restored
  1278: Window parameters are not restored
< ===== TCP reestablished

[root@Fedora24-Client vsftpd]# sh demo-vsftpd.sh download
0aa71e0d2c3a1daff204d8c51ad98b4b
/root/Fedora-Workstation-netinst-x86_64-24-1.2.iso

CRIU Demo - vncserver

1) Start VNC Server
2) Start VNC Client

```
TigerVNC Viewer 64-bit v1.7.0
Built on: 2016-09-12 08:27
Copyright (C) 1999-2016 TigerVNC Team and many others (see README.txt)
See http://www.tigervnc.org for information on TigerVNC.

Sat Oct 22 14:39:59 2016
 DecodeManager: Detected 1 CPU core(s)
 DecodeManager: Decoding data on main thread
 CConn:       connected to host 127.0.0.1 port 5925
 CConnection: Server supports RFB protocol version 3.8
 CConnection: Using RFB protocol version 3.8
 CConnection: Choosing security type None(1)
 X11PixelBuffer: Using default colormap and visual, TrueColor, depth 24.
 CConn:       Using pixel format depth 24 (32bpp) little-endian rgb888
 CConn:       Using Tight encoding
 CConn:       Enabling continuous updates
```

CRIU Demo - vncserver

3) Check ps tree

```
systemd,1 --switched-root --system --deserialize 23
  └─vnc_server.sh,1266,ipc,pid ./vnc_server.sh icewm
      ├─Xvnc,1267 :25 -v -geometry 800x600 -SecurityTypes none
      └─icewm,1270
          └─xterm,1274
              └─bash,1276
                  └─sh,1296 loop.sh
                      └─sleep,1302 1
```

< ===== pid, ipc namespace used here

4) Dump VNC Server

5) Check ps tree again

```
root      1260  0.0  0.2 117664  3040 pts/0    S+   14:39   0:00 sh demo-vnc.sh
root      1271  0.3  1.2 280656 12248 pts/0    Sl+  14:39   0:00 vncviewer 127.0.0.1:25
root      1334  0.0  0.0 117144   932 pts/0    S+   14:40   0:00 grep vnc
```

CRIU Demo - vncserver

6) Restore VNC Server
7) Check ps tree

```
systemd,1 --switched-root --system --deserialize 23
  └─vnc_server.sh,1339,ipc,pid ./vnc_server.sh icewm
      ├─Xvnc,1342 :25 -v -geometry 800x600 -SecurityTypes none
      └─icewm,1341
          └─xterm,1343
              └─bash,1344
                  └─sh,1345 loop.sh
                      └─sleep,1346 1
```

Sat Oct 22 14:40:41 2016
CConn: End of stream

CRIU Demo - runc

[root@Fedora24-Server ~]# create_runc_bundle
docker export 0371a53370063a71012892fa81a9e85e48ee12383df855084da3009e26387e66 | tar -C /root/runc/httpd/rootfs -xf -
[root@Fedora24-Server ~]# start_runc_container httpd
runc start -b /root/runc/httpd httpd
AH00557: httpd: apr_sockaddr_info_get() failed for runc
AH00558: httpd: Could not reliably determine the server's fully qualified domain name, using 127.0.0.1. Set the 'ServerName' directive globally to suppress this message
AH00557: httpd: apr_sockaddr_info_get() failed for runc
AH00558: httpd: Could not reliably determine the server's fully qualified domain name, using 127.0.0.1. Set the 'ServerName' directive globally to suppress this message
[Sat Oct 22 06:37:23.231146 2016] [mpm_event:notice] [pid 1:tid 140271485646720] AH00489: Apache/2.4.23 (Unix) configured -- resuming normal operations
[Sat Oct 22 06:37:23.232110 2016] [core:notice] [pid 1:tid 140271485646720] AH00094: Command line: '/usr/local/apache2/bin/httpd -D FOREGROUND'

CRIU Demo - runc

```
[root@Fedora24-Server ~]# checkpoint_runc_container
runc list
ID          PID         STATUS      BUNDLE             CREATED
httpd       1592        running     /root/runc/httpd   2016-10-22T06:37:23.198579161Z
runc checkpoint --image-path /root/runc/httpd/image --work-path /root/runc/httpd/log --shell-job --ext-unix-sk --file-locks
--tcp-established httpd
runc list
ID          PID         STATUS      BUNDLE      CREATED

[root@Fedora24-Server ~]# restore_runc_container
runc list
ID          PID         STATUS      BUNDLE      CREATED
runc restore -b /root/runc/httpd --image-path /root/runc/httpd/image --work-path /root/runc/httpd/log --tcp-established
--file-locks --ext-unix-sk --detach httpd
runc list
ID          PID         STATUS      BUNDLE             CREATED
httpd       1811        running     /root/runc/httpd   0001-01-01T00:00:00Z
```

CRIU Demo - Live Migration

[root@Fedora24-Client ~]# ./p.haul-wrap service
Waiting for connection...

[root@Fedora24-Server ~]# ./p.haul-wrap client 192.168.122.12 pid 1937
Establish connection...
Exec p.haul: ./p.haul pid 1937 --to 192.168.122.12 --fdrpc 3 --fdmem 4
……
14:47:01.694: 2020: Migration succeeded
14:47:01.695: 2020:   total time is ~1.47 sec
14:47:01.696: 2020:   frozen time is ~0.61 sec (['0.01', '0.00', '0.60'])
14:47:01.697: 2020:   restore time is ~0.08 sec
14:47:01.697: 2020:   img sync time is ~0.21 sec
14:47:01.698: 2020: Removing images

Containerized App support status

- Top 10 Applications
  - dnsmasq
  - httpd
  - vsftpd
  - sendmail
  - tomcat
  - mongo
  - mysql
  - mariadb
  - postgres
  - oracle

Container support status

- LXC
  - Build-in support in latest lxc
- Docker
  - Experienmental branch
- runc
  - Build-in support in latest runc

Credits

- The Linux kernel code (https://www.kernel.org/)
- Tejun Heo's ptrace-parasite (https://code.google.com/p/ptrace-parasite/)
- CRIU (https://criu.org/Main_Page)
- P.Haul (https://criu.org/P.Haul)
- runC (https://runc.io/)
- Docker (https://www.docker.com/)
- LXC (https://linuxcontainers.org/)

Q & A