To Kill or to Checkpoint - That is the Question

Mike Rapoport
<rppt@linux.vnet.ibm.com>

Adrian Reber
<areber@redhat.com>

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 688386
Cloud native era

- Applications are
  - Adaptable
  - Resilient
  - Stateless

- No need for migration, just kill here and start there
CRIU today

- Supports several architectures
- Integrated with major container engines
  - Docker
  - LXC/D
  - OpenVZ
  - podman
  - runc
- Used in production
  - IBM Spectrum LSF
  - Google
  - Virtuozzo
Simple example

- Shell script
  - print the date each second

```bash
#!/bin/sh

while :: do
    sleep 1
    date
    done
```

- [https://criu.org/Simple_loop](https://criu.org/Simple_loop)
- [https://asciinema.org/a/c5j7RTcYkcQqFGpsqiuIkoaj](https://asciinema.org/a/c5j7RTcYkcQqFGpsqiuIkoaj)
Simple example

```bash
criu@criu-dev:/examples/simple_loop
$ ls
Loop.sh
criu@criu-dev:/examples/simple_loop
$ cat ./loop.sh
#!/bin/sh

while :: do
    sleep 1
    date
done

criu@criu-dev:/examples/simple_loop
$ ./loop.sh
Thu Oct 18 08:33:49 UTC 2018
Thu Oct 18 08:33:50 UTC 2018
Thu Oct 18 08:33:51 UTC 2018
Thu Oct 18 08:33:52 UTC 2018
Thu Oct 18 08:33:53 UTC 2018
Thu Oct 18 08:33:54 UTC 2018
Thu Oct 18 08:33:55 UTC 2018
Thu Oct 18 08:33:56 UTC 2018
Thu Oct 18 08:33:57 UTC 2018
Thu Oct 18 08:33:58 UTC 2018
Thu Oct 18 08:33:59 UTC 2018
Thu Oct 18 08:36:00 UTC 2018
Thu Oct 18 08:36:01 UTC 2018
Thu Oct 18 08:36:02 UTC 2018
Thu Oct 18 08:36:03 UTC 2018
Thu Oct 18 08:36:04 UTC 2018
Thu Oct 18 08:36:05 UTC 2018
Thu Oct 18 08:36:06 UTC 2018
Thu Oct 18 08:36:07 UTC 2018
Thu Oct 18 08:36:08 UTC 2018
Thu Oct 18 08:36:09 UTC 2018
Thu Oct 18 08:36:10 UTC 2018
Killed
criu@criu-dev:/examples/simple_loop
$
```
TCP echo server and client

```c
/* client */

int val = 1;

while (1) {
    write(sk, &val, sizeof(val));
    read(sk, &rval, sizeof(rval));
    printf("PP %d -> %d\n", val, rval);
    sleep(2);
    val++;
}

/* server */

int val;

while (1) {
    read(sk, &val, sizeof(val));
    write(sk, &val, sizeof(val));
}
```

- [https://criu.org/Simple_TCP_pair](https://criu.org/Simple_TCP_pair)
- [https://asciinema.org/a/yjC1IK4S0ZG8Dz1UFI3CK0Vnc](https://asciinema.org/a/yjC1IK4S0ZG8Dz1UFI3CK0Vnc)
Live TCP

$ sudo crio dump -v0 --shell-job --tcp-established -t $(pidof client) 66 echo OK
OK
$ sudo crio restore -v0 --shell-job --tcp-established
PP 22 -> 22
PP 23 -> 23

Killed

$ crio @cni-dev:/examples/tcp_pair

09:17:24.164352 IP 127.0.0.1.9876 > 127.0.0.1.53252: tcp 4
09:17:24.164364 IP 127.0.0.1.53252 > 127.0.0.1.9876: tcp 8
09:17:26.164651 IP 127.0.0.1.53252 > 127.0.0.1.9876: tcp 4
09:17:26.164731 IP 127.0.0.1.9876 > 127.0.0.1.53252: tcp 4
09:17:26.164742 IP 127.0.0.1.53252 > 127.0.0.1.9876: tcp 8

09:17:42.220738 IP 127.0.0.1.53252 > 127.0.0.1.9876: tcp 8
09:17:42.220749 IP 127.0.0.1.9876 > 127.0.0.1.53252: tcp 8
09:17:42.221147 IP 127.0.0.1.53252 > 127.0.0.1.9876: tcp 4
09:17:42.221214 IP 127.0.0.1.9876 > 127.0.0.1.53252: tcp 4
09:17:42.221240 IP 127.0.0.1.53252 > 127.0.0.1.9876: tcp 8
09:17:44.221490 IP 127.0.0.1.53252 > 127.0.0.1.9876: tcp 4
09:17:44.221562 IP 127.0.0.1.9876 > 127.0.0.1.53252: tcp 4
09:17:44.221584 IP 127.0.0.1.53252 > 127.0.0.1.9876: tcp 8
Iterative dump

- Checkpoint - restore of memcached server
  - Several pre-dump iterations
  - Final dump and server freeze
  - Lazy restore

- https://asciinema.org/a/4olkCRLaMGa4GhxTEvbzyTtv4
Iterative dump

GET: 17: 100 values
GET: 18: 100 values
GET: 19: 100 values
GET: 20: 100 values
GET: 21: 100 values
GET: 22: 100 values
GET: 23: 100 values
GET: 24: 100 values
GET: 25: 100 values
GET: 26: 100 values
GET: 27: 100 values
GET: 28: 100 values
GET: 29: 100 values
GET: 30: 100 values
GET: 31: 100 values

Putting 100 more values

GET: 0: 200 values
GET: 1: 200 values
GET: 2: 200 values
GET: 3: 200 values
GET: 4: 200 values
GET: 5: 200 values
GET: 6: 200 values
GET: 7: 200 values
GET: 8: 200 values
GET: 9: 200 values
GET: 10: 200 values
GET: 11: 200 values
GET: 12: 200 values
GET: 13: 200 values
GET: 14: 200 values
GET: 15: 200 values
GET: 16: 200 values
GET: 17: 200 values
GET: 18: 200 values
GET: 19: 200 values
GET: 20: 200 values
GET: 21: 200 values
GET: 22: 200 values
GET: 23: 200 values
GET: 24: 200 values

[0]: bash*
Iterative dump with LXD

- Iterative container migration using LXD
  - First migration threshold: 90%
  - Second migration threshold: 40%
  - Different number of pre-copy runs depending on these thresholds
- [https://asciinema.org/a/mUxKQHnPbRLdNMNukE2WuJku7](https://asciinema.org/a/mUxKQHnPbRLdNMNukE2WuJku7)
  - Hosts: RHEL7
  - Container: Alpine 3.7
Iterative dump with LXD

```
| NAME | STATE | IPV4 | IPV6 | TYPE | SNAPSHOTS |
| +-----+-------+-------+-------+------+------------|
| root@rhel01 -# lxc move alpine rhel02:alpine
| root@rhel01 -# lxc list
| +-----+-------+-------+-------+------+------------|
| NAME | STATE | IPV4 | IPV6 | TYPE | SNAPSHOTS |
| +-----+-------+-------+-------+------+------------|
| root@rhel01 -# lxc list rhel02:
| +-----+-------+-------+-------+------+------------|
| NAME | STATE | IPV4 | IPV6 | TYPE | SNAPSHOTS |
| +-----+-------+-------+-------+------+------------|
| alpine | RUNNING | | | PERSISTENT | |
| +-----+-------+-------+-------+------+------------|
| root@rhel01 -# lxc mv rhel02:alpine alpine
| root@rhel01 -# lxc list
| +-----+-------+-------+-------+------+------------|
| NAME | STATE | IPV4 | IPV6 | TYPE | SNAPSHOTS |
| +-----+-------+-------+-------+------+------------|
| alpine | RUNNING | | | PERSISTENT | |
| +-----+-------+-------+-------+------+------------|
| root@rhel01 -# lxc list
| 0 adrian@dcbz:~
| CRUI pages skipped 0
| CRUI pages skipped percentage 0%
| Doing another pre-dump in 001
| CRUI pages written 42849
| CRUI pages skipped 00400
| CRUI pages skipped percentage 59%
| Doing another pre-dump in 003
| CRUI pages written 10750
| CRUI pages skipped 03707
| CRUI pages skipped percentage 82%
| Doing another pre-dump in 005
| CRUI pages written 10880
| CRUI pages skipped 01577
| CRUI pages skipped percentage 90%
| Doing another pre-dump in 007
| CRUI pages written 0843
| CRUI pages skipped 04414
| CRUI pages skipped percentage 93%
| Memory pages skipped (93%) due to pre-copy is larger than threshold (90%)
| This was the last pre-dump; next dump is the final dump
| Using maximal 10 iterations for pre-dumping
| The other side does support pre-copy
| Doing another pre-dump in
| CRUI pages written 102457
| CRUI pages skipped 0
| CRUI pages skipped percentage 0%
| Doing another pre-dump in 001
| CRUI pages written 55897
| CRUI pages skipped 46560
| CRUI pages skipped percentage 46%
| Memory pages skipped (46%) due to pre-copy is larger than threshold (40%)
| This was the last pre-dump; next dump is the final dump
```
Application snapshot

- Checkpoint/restore using Podman
  - Apache Tomcat running in a container
  - Container is checkpointed
  - System rebooted
  - Container is restored

- https://asciinema.org/a/FsTbx9mZkzeuhCM2pFOr1tujiM
Connection to rhel01 closed.
[root@rhel02 ~]# ssh rhel01

root@rhel01:~# podman container restore --keep podman-criu-test 9f4b568a05e98297763ebc6bcbcebedede351ab781dd55f3a7a792e4253b95e01a80

root@rhel01:~# reboot
PolicyKit daemon disconnected from the bus.
We are no longer a registered authentication agent.
Connection to rhel01 closed by remote host.
Connection to rhel01 closed.
[root@rhel02 ~]# ssh rhel01

root@rhel01:~# podman container restore --keep podman-criu-test 9f4b568a05e98297763ebc6bcbcebedede351ab781dd55f3a7a792e4253b95e01a80

root@rhel01:~#

[root@rhel02 ~]# curl 10.22.0.53:8080/examples/servlets/servlet/HelloWorldExample
1
[root@rhel02 ~]# curl 10.22.0.53:8080/examples/servlets/servlet/HelloWorldExample
2
[root@rhel02 ~]# curl 10.22.0.53:8080/examples/servlets/servlet/HelloWorldExample
3
[root@rhel02 ~]# curl 10.22.0.53:8080/examples/servlets/servlet/HelloWorldExample
4
[root@rhel02 ~]# curl 10.22.0.53:8080/examples/servlets/servlet/HelloWorldExample
5
[root@rhel02 ~]# curl 10.22.0.53:8080/examples/servlets/servlet/HelloWorldExample
3
[root@rhel02 ~]# curl 10.22.0.53:8080/examples/servlets/servlet/HelloWorldExample
4
[root@rhel02 ~]#
CRIU and kexec

- Checkpoint OpenVPN server
  - While client connected via TCP
- kexec into different kernel
- Restore OpenVPN server
  - Client still connected
- https://asciinema.org/a/i6PF2bJo79QP6pnhqJwX4Wfs2
CRIU and kexec

available by default. Update your scripts to load br_netfilter if you need this.

* 6.709406 nf_conntrack version 0.5.8 (65536 buckets, 2024 max)
* 6.709767 ip6_tables: (C) 2000-2008 Netfilter Core Team
* 6.781887 Bits 55-68 of /proc/PID/pagemap entries are about to stop being page-size
* shifted some time soon. See the linux/documentation/vm/pagemap.txt for details.
* 6.784485 The pagemap bits 55-68 have changed their meaning! See the linux/documentation/vm/pagemap.txt for details.
* 6.835753 tun: Universal TUN/TAP device driver, 1.6
* 6.837770 tun: (C) 1999-2004 Max Kransynsky <maxx@qualcomm.com>
* 6.881899 cgroup: new mount options do not match the existing superblock, will be ignored
* 979365 IPv6: ADDRCONF(NETDEV_UP): veth1: link is not ready
* 7.111814 Netfilter messages via NETLINK v0.30.
* 7.126345 ctetlink v0.93: registering with nfnetlink.
* 7.145896 br0: port 1(veth0) entered blocking state
* 7.145896 br0: port 1(veth0) entered disabled state
* 7.147290 device veth0 entered promiscuous mode
* 7.149511 br0: port 1(veth0) entered blocking state
* 7.149597 br0: port 1(veth0) entered forwarding state
* 7.150855 IPv6: ADDRCONF(NETDEV_CHANGE): veth1: link becomes ready

0 Adrian/Adrians:~
Wed Oct 18 08:44:16 2018 OPTIONS IMPORT: route options modified
Wed Oct 18 08:44:16 2018 OPTIONS IMPORT: peer-id set
Wed Oct 18 08:44:16 2018 OPTIONS IMPORT: adjusting link mtu to 1024
Wed Oct 18 08:44:16 2018 OPTIONS IMPORT: data channel crypto options modified
Wed Oct 18 08:44:16 2018 Data Channel: using negotiated cipher 'AES-256-GCM'
Wed Oct 18 08:44:16 2018 Outgoing Data Channel: Cipher 'AES-256-GCM' initialized with 256 bit key
Wed Oct 18 08:44:16 2018 Incoming Data Channel: Cipher 'AES-256-GCM' initialized with 256 bit key
Wed Oct 18 08:44:16 2018 ROUTE_GATEWAY 192.168.122.1/255.255.255.0 IFACE=eht0 HWADDR: 52:54:00:35:0b:01
Wed Oct 18 08:44:16 2018 TUN/TAP device tun0 opened
Wed Oct 18 08:44:16 2018 /sbin/ifup link dev tun0 up mtu 1500
Wed Oct 18 08:44:16 2018 do_ifconfig, tt->old_ifconfig ipv6 setup=0
Wed Oct 18 08:44:16 2018 /sbin/ifup link dev tun0 up mtu 1500
Wed Oct 18 08:44:16 2018 /sbin/ifup addr add dev tun0 local 172.31.0.6 peer 172.31.0.5
Wed Oct 18 08:44:16 2018 /sbin/ifup route add dev 172.31.0.6/24 via 172.31.0.5
Wed Oct 18 08:44:16 2018 WARNING: this configuration may cache passwords in memory...
    use the auth-no-cache option to prevent this
Wed Oct 18 08:44:16 2018 Initialization Sequence Completed

64 bytes from 172.31.0.1: icmp_seq=2 ttl=64 time=15.8 ms
64 bytes from 172.31.0.1: icmp_seq=3 ttl=64 time=16.0 ms
64 bytes from 172.31.0.1: icmp_seq=4 ttl=64 time=16.4 ms
64 bytes from 172.31.0.1: icmp_seq=5 ttl=64 time=15.6 ms
64 bytes from 172.31.0.1: icmp_seq=6 ttl=64 time=15.8 ms
64 bytes from 172.31.0.1: icmp_seq=7 ttl=64 time=15.9 ms
64 bytes from 172.31.0.1: icmp_seq=8 ttl=64 time=16.0 ms
64 bytes from 172.31.0.1: icmp_seq=9 ttl=64 time=15.3 ms
64 bytes from 172.31.0.1: icmp_seq=10 ttl=64 time=15.9 ms
64 bytes from 172.31.0.1: icmp_seq=11 ttl=64 time=15.7 ms
64 bytes from 172.31.0.1: icmp_seq=12 ttl=64 time=15.5 ms
64 bytes from 172.31.0.1: icmp_seq=13 ttl=64 time=15.9 ms
64 bytes from 172.31.0.1: icmp_seq=14 ttl=64 time=16.2 ms
64 bytes from 172.31.0.1: icmp_seq=15 ttl=64 time=16.0 ms
64 bytes from 172.31.0.1: icmp_seq=16 ttl=64 time=15.0 ms
64 bytes from 172.31.0.1: icmp_seq=17 ttl=64 time=17.5 ms
64 bytes from 172.31.0.1: icmp_seq=18 ttl=64 time=15.7 ms
64 bytes from 172.31.0.1: icmp_seq=19 ttl=64 time=16.0 ms
64 bytes from 172.31.0.1: icmp_seq=20 ttl=64 time=16.7 ms
64 bytes from 172.31.0.1: icmp_seq=21 ttl=64 time=15.4 ms

2 Adrian/Adrians:~
Serverless

- Openwhisk action runtimes
  - Python
  - Java
  - Node.js

```
docker run vs docker start --checkpoint
```

<table>
<thead>
<tr>
<th></th>
<th>python3action</th>
<th>nodejs6action</th>
<th>java8action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start time (sec)</td>
<td>0.919969</td>
<td>0.743099</td>
<td>0.566213</td>
</tr>
<tr>
<td>Restore time (sec)</td>
<td>0.078663</td>
<td>0.091419</td>
<td>0.089104</td>
</tr>
</tbody>
</table>
Data store optimizations

- **Populate vs migrate**
  - Store with 100000 short entries

<table>
<thead>
<tr>
<th></th>
<th>memcached</th>
<th>redis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Populate time (sec)</td>
<td>1.238</td>
<td>6.254</td>
</tr>
<tr>
<td>Migrate time (sec)</td>
<td>0.806</td>
<td>1.671</td>
</tr>
</tbody>
</table>

- **C/R and migration help:**
  - Keep the caches
  - Reduce footprint for master-slave configuration
  - Run redis with transparent huge pages ;-)
More use cases

- **Application forensics**
  - Get application state from production
  - Investigate misbehaviour offline

- **Fault tolerant and highly available systems**
  - Use CRIU as state replication engine

- **Load balancing**

- **Move applications into** *screen*/*tmux*
Future: CRIU and kubernetes

- kubectl drain --migrate
- k8s replica sets with restore:
  ```yaml
  apiVersion: apps/v1
  kind: ReplicaSet
  metadata:
    name: datastore
  spec:
    replicas: 5
    startOptions:
      - checkpointRestore
  ```
Thank you!