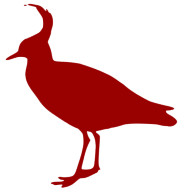
Another Year with CRIU

News from the Developers

Adrian Reber <areber@redhat.com>
Andrei Vagin <avagin@gmail.com>

Linux Plumbers Conference 2018
November 13, Vancouver

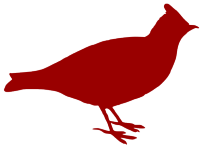
Checkpoint Restore In Userspace

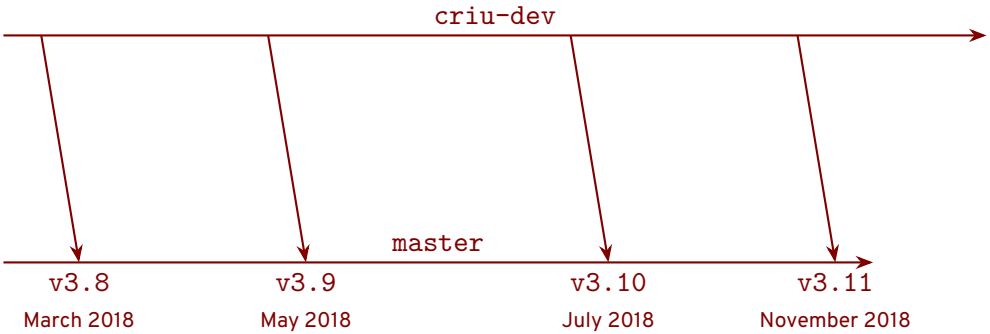


2011: Initial CRIU RFC

2012: crtools v0.1 release

2013: CRIU 1.0 release

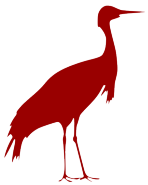



criu-dev
master
v3.8
March 2018
v3.9
May 2018
v3.10
July 2018
v3.11
November 2018

7 CRIU releases since LPC 2017

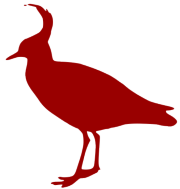
```
$ git diff v3.5..origin/master --stat | tail -n 1

516 files changed, 21907 insertions(+), 5443 deletions(-)
```



```
$ git log v3.5..origin/master| grep Author| sort| uniq| wc -l

37
```



Virtuozzo, IBM, Red Hat, Google,
Arista, INESC-ID, HP, etc

CRIU 3.5 - Clay Jay

- September 2017
- Lazy Migration - Userfaultfd

CRIU 3.6 - Alabaster Finch

- October 2017
- Added support to checkpoint and restore
  - Files sent over unix sockets
  - Threads with different credentials
  - IPv6 over IPv4 tunnel (SIT devices)

CRIU 3.7 - Vinyl Magpie

- December 2017
- Added support to checkpoint and restore
  - SO_REUSEPORT option
  - IPv4 mapped inet sockets
  - net_prio cgroups
  - Overmounted shared mountpoints

CRIU 3.8 - Snow Bunting

- March 2018
- Added support to checkpoint and restore
  - Multiple network namespaces
  - Overmounted tmpfs mounts
  - Unix and epoll descriptors in SCM messages

CRIU 3.9 - Sand Martin

- May 2018
- Added support to checkpoint and restore
  - TUN/TAP devices in sub network namespaces
  - File descriptors opened with O_TMPFILE

CRIU 3.10 - Granite Eagle

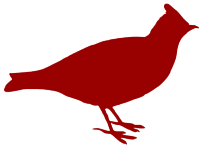
- July 2018
- Added Python 3 support
- Large pages support for aarch64/ppc64le
- Added support to checkpoint and restore
  - Per thread seccomp chains

CRIU 3.11 - Glass Flamingo

- November 2018
- Added support for configuration files
- Added support for external network namespaces
- cpuinfo: detect compact frames and handle noxsaves
  - epoll: add support for duped targets
  - tun: add support for multiple network namespaces
  - x86: support extendable fpu frames

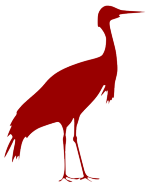
External Network Namespaces

Container Runtime creates Network Namespace

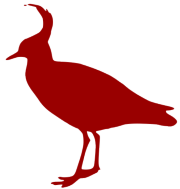


CRIU dumps container with Network Namespace

On restore CRIU *creates* a new Network Namespace



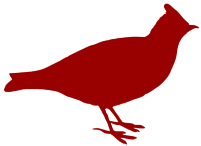
Podman does it differently



Podman uses CNI to create a Network Namespace

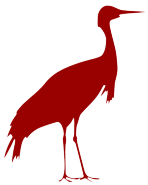
Podman tells runc to use that Network Namespace

CRIU dumps container with Network Namespace

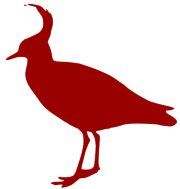


On restore CRIU *creates* a new Network Namespace

This, however, is a different Network Namespace



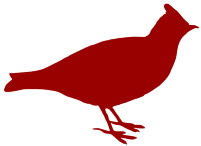
`criu restore` into an existing Network Namespace



```
criu dump --external net[<inode>]:netns-name -t <PID>
    criu restore --inherit-fd fd[<FD>]:netns-name
```

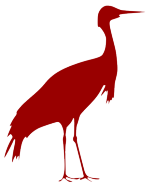
Configuration Files

Containers Runtimes are layered

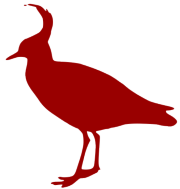


New CRIU features requires changes on all layers

Influence CRIU's behaviour via configuration files



Configuration Files Example



```
$ criu dump -t `pgrep -f 'tcp-howto 127.0.0.1 10000'`
Error (criu/sk-inet.c:188): inet: Connected TCP socket,
    consider using --tcp-established option.
```

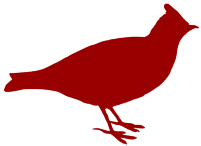
```
$ echo tcp-established > /etc/criu/default.conf
$ criu dump -t `pgrep -f 'tcp-howto 127.0.0.1 10000'`
Error (criu/tty.c:1861): tty: Found dangling tty with
    sid 16693 pgid 16711 (pts) on peer fd 0.
Task attached to shell terminal. Consider using --shell
    -job option. More details on http://criu.org/
    Simple_loop
```

```
$ echo shell-job >> /etc/criu/default.conf
$ criu dump -t `pgrep -f 'tcp-howto 127.0.0.1 10000'`
    && echo OK
OK
```

https://lisas.de/~adrian/posts/
2018-Nov-08-criu-configuration-files.html

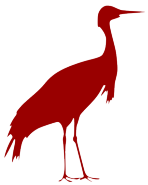
Container Runtimes

runc: pre-copy, post-copy



lxc/lxd: pre-copy

docker/podman: no optimization



GO Bindings

https://github.com/checkpoint-restore/go-criu

CRIU Hackathon 2018

- This Friday, 2018-11-16
- `http://ateliervancouver.com/`
- 319 W Hastings St #400, Vancouver, BC

Wishes?

Questions?

The end.

Thanks for listening.