How far a naive FreeBSD container implementation can go
Menu of today

- What is a container
- What is pot and what we did so far
- Thoughts about containers on FreeBSD
whoami

Luca Pizzamiglio

pizzamig@FreeBSD.org

Port committer since 2017
What is a container?

A way to distribute application

- The image
- The runtime

Developer focused at first
pot

pot is a jail framework to provide containers to FreeBSD

Initial goal: to prove that FreeBSD has all the technologies needed by a container-alike environment

This is why pot is written in `/bin/sh`

Educational goal: to learn how containers work

Project started in November 2017, presented at FOSDEM in 2018 and 2020
What is pot?

One ZFS dataset with everything you need

FreeBSD base, packages and your application

flavour: scripts to imitate Dockerfile

pot image: `zfs snapshot && zfs send | xz`

focus on

  non persistent jails

no `sh /etc/rc` but only one process per jail
What is pot - a bit of runtime

Once the image is available, it runs via jail

jail provides a clean and isolated runtime environment for the container

rctl is used to provide resource limits

VNET is used to provide a new network stack, if wanted

bridge can be used to provide network abstraction

pf is used to provide NAT and redirection if needed
pot network

pot supports different network setups

- inherit: inherit the stack of the host
- alias: different IP on the network card
- public-bridge: use a bridge, shared between jails, to attach VNET based pot
- private-bridge: use a bridge, to attach VNET based pot, dedicated to few jails

IPv4 address allocation for bridges requires potnet, a third party application

pot can support different IP stacks

- IPv4 only
- IPv6 only
- Dual stack
pot network - bridge and stack

One bridge per stack

Bridge and IPv4

  The bridge and all jails lives in a detached internal network

  pf provides connectivity via nat (outbound) and redirect (inbound)

Bridge and IPv6

  Nat and redirect are so IPv4, put the network interface on the bridge

  limitation: support for promiscuous mode (no wlan)
pot and nomad

nomad is an open source container orchestrator developed by HashiCorp

A nomad driver for pot has been implemented to provide jails orchestration in a “cloud native” way

   Original implementation written by Esteban Barrios

Enables a kubernetes like experience
pot and nomad

Feel free to play with it with sysutils/minipot

It uses traefik as ingress
potluck

potluck is the image registry for pot

Originally implemented by Stephan Lichtenauer

From a collection of flavors, it generates binary images

Flavors repository: https://github.com/bsdpot/potluck

Registry URL: https://potluck.honeyguide.net/
pot and nomad - how far?
pot and nomad - this far!

Select a task to start your session.

Multiple instances of this task are running. The allocation below was selected by random draw.

Customize your command, then hit ‘return’ to run.

$ nomad alloc exec -i -t -task server-B 47425999 sh
# ls -al
total 36
  drwxr-x--- 2 root wheel 7 Apr 9 2021.
  drwx------- 18 root wheel 23 Sep 6 12:50 ..
  -rw-r--r-- 2 root wheel 1023 Apr 9 2021 .cshrc
  -rw-r--r-- 1 root wheel 80 Apr 9 2021 .k5login
  -rw-r--r-- 1 root wheel 328 Apr 9 2021 .login
  -rw-r--r-- 2 root wheel 507 Apr 9 2021 .profile
  -rw-r--r-- 1 root wheel 865 Apr 9 2021 .shrc
# uname -a
FreeBSD server-B_e72e3b80_47425999-65f5-5de4-4df7-2b0d3d89f2a3 13.0-RELEASE-p5
FreeBSD 13.1-STABLE #43 stable/13-n252233-108b6eb670ec: Fri Sep 2 08:32:00 UTC 2022 root@jailhost-b:/usr/obj/usr/src/amd64.amd64/sys/GENERIC amd64
#
pot and nomad: a community effort

Esteban wrote the driver

grembo@ runs a cluster in a professional environment

  Many corner cases with issues that have been addressed

  Additional use cases, like batch jobs

Stephan maintains a registry with pre-built images, available for everyone

  More flavours can be added!

(almost) regular updates in the Quarterly Status Report
Latest features

pot (from COVID until today) [0.15.2]

- Layered images
- Dns configuration when cloning
- Custom directory with flavors
- Garbage collect POSIX shared memory (fixed in CURRENT, tho)
- Fix concurrency for start/stop race conditions
- Support to encrypted ZFS dataset (thanks to ZFS support)

nomad pot driver

- Support for batch jobs and periodic batch jobs
- Support for signals and exec
Current FreeBSD issues/differences

pf redirect from the same host not working as expected

  It solvable using a reflect jail, but it’s still above my comprehension

ability to nullfs-mount a file

vnet/epair destroy / jail stop race conditions

  Many has been solved, but we still have a sleep because sometimes it still happens

rctl won’t kill the jail (OOM) in case of higher memory consumption
What next?

Initial assumptions are a constant source of pain

Every new big feature needs a lot of work

- No initial design for images and no OCI support
- No good pot lifecycle support
  - Clean up after non persistent jails exit
  - Nomad-pot-driver is currently taking care of it
- Needs to manage jails through a supervisor
  - Ability to starts containers as user, without sudo
- sh is a lot of fun, until it’s not
- Log needs proper support
  - As stdout not as syslogd
Personal thoughts

To evolve, pot needs a profound redesign, some reimplementation and dropping some features.

The FreeBSD community seems to ask for container support.

Only a community can implement and support it.

- Many subsystems involved (ZFS, jails, network)
- Many different ways to use (stress) containers

Use a programming language with a rich set of useful libraries.

- GO seems the natural choice for containers, but …

Need of emulation on other OS for local development.
Thanks!

Thanks to everyone contributed, every PR makes a difference!

Thanks to you for listening!

Any question?

Reach out for any additional question to pizzamig@FreeBSD.org