Jail Developers/Production Users Call Minutes

These calls take place each Wednesday at 10AM Pacific barring conflicts with community events and holidays. The focus alternates between Developers and Production Users each week The Zoom meeting link is available and you are welcome to contact Antranig V. (antranigv@freebsd dot am) and Michael Dexter (editor@callfortesting dot org) with questions, concerns, and to be added or removed from the invitation mailing list. This regular call began on 2023-03-08.

2023-03-08 Production Users
Attendees: Phillip V, Michael D, Alan S, Antranig V, Dave C, Igor A, Phillip V, Goran M, Dan L, Dominik J

Welcome!

Introductions

Alan Somers: Developer, been using iocage and ezjail before that. Disappointed that iocage is under-maintained. Concerned that Bastille and pot are making: Written in sh, making it hard to debug, resulting in zero test suite.

Phill V: Pharmacist using Jails for internal development. Following Lucas' book tools but always had short-life Jails and the tools were not too important. Now needs a long term strategy.

Igor Antonov: FreeBSD enthusiasts, Linux $dayjob for ten plus years. Knows Linux containers. Interest is not seeing Jails repeat the mistakes that Linux containers have made.

Dave C: Usually "**DCH**". Started with FreeBSD in 2010. Saw the cracks in OpenSSL resulting in key leakage. Company wanted Kubernetes and he moved the solution to FreeBSD. Watched the journey of **`iocage`** from sh to py to go... Agreed that we need a community solution. Wrote his own tool... Wrote a tutorial "DIY jails" that shows what the tools are doing under the hood. Interested in seeing the Handbook updated. Has a wish list of items to put into Jails:

1. Top complaint: Lack of a state machine to clearly follow transitions, to help build add-ons and extensions.
2. Seeks to add “live” Jail metadata in the kernel, as opposed to static jail.conf.

DIY jails - slides https://freeside.skunkwerks.at/~dch/FreeBSD/diy-jails/tutorial.html sometimes off overnight. Src is here https://git.sr.ht/~dch/diy-jails-tutorial/ contributions and corrections welcomed!

Antranig: “You can call me Frank” :) Ran his country’s CERT, using FreeBSD. Deployed honeypots around the country. Created his own honeypot solution based on FreeBSD. Produced Jailer, YAJMT (yet another Jail Management Tool). Desire to identify the common steps.

Dan Langille: Runs BSDCan, PGCon, BSD Diary, and Fresh Ports. Been using jails for 10+ years. Writing a talk about how those more jails you use, the more jails you need. Also helped with mkjail from Felder. Note that it does not destroy jails. “rmjail”? Note that it is ZFS aware. “Most Jail managers do too much.”

Goran M: Co-founded a hacker space “Tilde Center”. Mostly working on FreeBSD Audio and has a home audio/photography studio. Experimenting with a Jail utility. He’s certain we can do better. Working with Alan. Learning DTrace.

Dominik J: Switching from Linux to FreeBSD.

Michael Dexter: “RPM Hell” drove me to Jail and the j-tools in FreeBSD 5.1 sent me on a 20 year multiplicity/virtualization journey that culminated this week with a short and sweet four page paper for my upcoming talks outlining the huge progress that has been made.

Minutes

Question: Should libxo be added? Note the preexec command: Could be a shell script. Need environment variables passed in there for global variables. jail_attach is the only kernel resource related to this. How to one JID to a VNET instance?

Note kern_jail in the source tree. This is under-documented. Note the 130 GOTOs.

```
$ grep goto kern_jail.c | wc -l
    137
```

Note that a jail can have a '.' in the name, which indicates that it is nested which supports dependencies. One line of documentation describes this.

Ihor: Is name.name the only way to show a child/parent relationship between jails?
Answer: This is how the configuration file determines the relationship. A nested jail not worry about its networking as it is handled by the parent.

Freshports.org uses this for clean build environments. "jls" can be ambiguous.

Question: Is anyone using Jails in multiple datasets? Most have one zroot according to an IRC poll.

Note the -e flag in addition to -f. Will print the configuration file(s). Document!

"We're all doing the same work in different repositories - so much more could be in base and we augment it with ports and packages." Consider upstreaming: Tools to assign epairs to Jails. `"vnet=epair"` and it is auto-created. Note the `/etc/start_if.<device name>` i.e. `/etc/start_if.epair4b` for an epair with a MAC address. Could start a database. Useful for WiFi drivers. Also available on OpenBSD. Should also support /etc/stop_if.<device>

If the /etc/start_if.⟨interface⟩ file is present, it is read and executed by the sh(1) interpreter before configuring the interface as specified in the ifconfig_⟨interface⟩ and ifconfig_⟨interface⟩_alias⟨n⟩ variables.

Dave: Two broad directions/initiatives: Build a list of desired in-kernel Jail resources/functionality. Seek development and funding. Both need consensus. Specifically, the state machine mentioned above for cleanly correcting interrupted states. When ZFS support arrived in Jails, there was metadata for dataset etc. The kernel interface supports additional metadata. The kernel could issue a UUID that will not be reused and cannot conflict. Currently facing the risk of race conditions. Would like userspace Jails, created by unprivileged users. “Rootless Jails” Linux does this and it has various security implications. Igor has notes and thoughts on this. Some amount of restriction would be required. Could have certain sysctls disabled, restrict hard links... Could use the MAC framework.

Antranig: Re: Funding: His company could assist but is an Elixir, rather than C shop. Rootless jails: We have rootless chroots, no? Yes. This may be a starting place. He’s using a variable UUID in the Jail conf to handle this. Note how Zone restart behavior differs.

How much promise does runj have ?

dfr's implementation is much further ahead (podman/buildah etc) than runj.

Note the Open Container Initiative and OCI for FreeBSD

Notes on dfr's podman port

Goran’s current patch: https://github.com/mekanix/freebsd-src/tree/feature/jail
Patches the jail utility which is aware of all applicable configuration paths.
Note “wildcard” jails - global config and other jails can override one or more above it.
“/etc/jail.conf” could be at the end... Start with jail.conf.d? Thoughts on the order of the parser? Dan: Sounds like /etc/defaults jail.conf is really jail.conf.defaults that you manually override. “This is a setting you cannot override” might be desired.
Note: * `{ exec.start = “”; exec.stop = “”; }`

Note: `/etc/jail.<name>.conf`
Note: `jail -r -f /path/to/jail.conf ‘*’`

Dave: The kernel, not configuration files should manage the true state of Jails. Need a description syntax that is handed to the kernel. Note UCL/libucl. Not sure if libucl supports variables but perhaps it can be added. Is there any UCL support in Jail yet?

Note that jls supports libxo JSON etc.

If you have an integer for a Jail name, it will use that as the jail ID. Feature or bug? What happens when they collide? Dave: Every jail is based on a commit hash… check the hash.

Note the possibility of integer mix-ups here. This will give Jail ID “something” “3”:

```
* { $id = 1; }
something { $id = 2; }
* { $id = 3; }
```

Anything global needs to be last, not first. Feature? Bug?

Note `/usr/sbin/ctld` use a local configuration file for a UCL example.

Is there an in-base UCL parser equivalent to jq(1)? libucl is there.

Is it helpful to inject runtime options that are not persistent (in a configuration file)?

You *can* obtain the runtime configuration from a running jail. Not obvious in the manual page: Try this command: `jls -n -j JID`

`jls -vh -j app0` will give values, -h for headers.

The order of flags matters! -vh and -hv do different things!

Question: Is the -h header output consistent?

`jls -d` will show deceased jails and will give a history. Jails never truly die! Is this a ZFS issue where resources are not released? There is a GOTO loop that monitors the Jail and ZFS subsystems. This might be only a 14-CURRENT issue. Issue went away on UFS. Saw the same issue on VNET. Needed -vnet flag. This was found with OccamBSD.

Dan's patch (link?): Prefer not specify running Jail names and prefer have global variables in jail.conf, rather than repeating information. Note this Ansible playbook for some of this.

Antranig: Note the sysctl.d review.

Note cron.d that is often used by packages in `/usr/local/etc` ... Note syslog.d linked from syslog.conf, and the "new syslog" syntax. The syntax changed in 13!

Question: How to stagger Jail operations in cron? Dan: Use jitter?

Poll: Who is using DHCP with Jails? Dan is, Goran runs it in Jails with IPv4/IPv6. Run rtdvd/rtsol on the router... (advertisements/solicitation) Expirations in 10 minutes. Can result in two default routes. The rtsol route does not disappear.

Question: Should FreeBSD OCI-compliant and follow their standard? Too many implementations become standards. To what degree is runj following the standard? Note Bryan Cantrill's talk on debugging Docker in Production. Consider a Linux-compatibility shim? The Go-based control plane is quite decoupled. Nomad scales much better - produced a 2000 node cluster. Mesos scales much better than Kubernetes or Nomad Link Link

Note: Brendan Gregg's comments on tracing Zones and Jails compared to Linux containers.

Quote: PHK: "I used a script to create lots of jails and got bored at 64k"

Question: Is Netgraph the only system that virtualizes SCTP? Does Netgraph need better coupling with PF?

Note https://github.com/freedbsd/meetings
The Desktop group is using hack.md - Desktop Group Notes on HackMD

Question: What are the top use cases for Jail? Note the app jail forum post. Appliances? What do we want Jails to do, a high level? Note the TrueNAS Plugins example. How best set up Nextcloud? See: YUNOHOST. Note the need for SoHo/Business appliances. Web interface?

Note: hello.systems AppImage Linux app images in jails. Note Snapcraft. Igor has much to share about this. Note NixOS.

Distinguish:

- desktop and server workloads,
- Ephemeral(data goes away when container dies) vs stateful (i.e. light VMs)

Priorities and Takeaways

- Unprivileged Jail launch
- Document `jail -e`
- Document `/etc/start|stop_if.<device name>`
- Document the commonalities of all Jail management tools
    - Stop reinventing the wheel and start upstreaming
- Review order of configuration file overrides and immutability
- Jail UCL support: Note bhyve_config/nvlist support
- Proper state machine for Jail start up and tear down state
- Identify common kernel resources to implement such as a true UUID, exposed metadata, and much more
- Jail audit! What does the official Jail test suite do? Note the Kyua files.
- https://github.com/allanjude/uclcmd - “jq” for UCL. Figure out way to move it into base

Michael! Add UTC times to the announcements!

Adjourn: 12:56 PM Pacific 20:56 UTC
Next Meeting: 2023-03-15 Developers