Did you get your floppy yet?

Modirum - EuroBSDCon 2023
This site is hosted from the retro PC at our sponsor table. Visit us to receive your very own bootable 90mm (3.5") floppy disk with a self-hosted web server, an IRC client, and various network tools. See you!

(Made with TheDraw for DOS and converted to HTML using ansifilter)
Hi !
WHO ARE YOU AGAIN?

- We authenticate on-line payments
  - using FreeBSD and open source software
  - on our own infrastructure, also built with open source
- Facing (un)usual challenges
  - Security requirements up the wazoo, DDoS attacks, etc.
  - "The Blame Game" means "make sure you can blame someone else" (..that'd be us..)
WHY AM I HERE?

- Used FreeBSD since ~2000
- Love open source
- Been working in the payment industry since 2003
- Have been radicalised by the Internet
  - I *really* love open source
  - ..and floppies.
PREVIOUSLY ON THIS SHOW:

- The Blame Game: FreeBSD and the absurdities of security compliance
- The Blame Game continues: Getting up from under the bus
- Today: Using BSD to do "real work" - CorporateBSD
CorporateBSD

NOT A HOW-TO!
Guy from Netflix: "I think you're wrong"

(Ill get back to this)
WHAT I WANT TO COVER

- How we use FreeBSD and OSS
- What makes the community so good for us
- We want to contribute more - how?
- War stories (because it never ends)
THE PROBLEMS CHALLENGES

- Lots of traffic:
  - 100s of payment transactions/sec
  - 5-15K write operations/sec to DB
  - >80K TLS handshakes/sec in software (DDoS)
- Lots of data: ~40TBx5 nodes actively-queried SQL
- Security, redundancy, compliance, having fun
THROWING HARDWARE AT THE PROBLEM

- Routers and FWs: 28-core AMD (overkill!)
- Application servers: 2x 16-core AMD, enough RAM
- DB servers: 2x 12-core Intel, 256GB RAM
  - 2x 8TB SATA SSD (MySQL logs)
  - 14x 4TB SATA SSD (Data pool)
  - 2x 200GB NVMe SSD (ZIL for both of the above)
DECENT HARDWARE, BUT..

None of it is exactly high-end.

- BSDRP make the routers tick; OPNsense ditto on firewalls
- Applications are nginx+Tomcat
  - Not super-efficient, but well-understood
- MySQL should not run on ZFS, we're told
  - BS, we call

In short: The software is the star of the show!
Wrap me if you can:

JAILS
DOING WEIRD SH*T IN JAILS

We want VNET, but classic jails have some security benefits:

- Cannot modify its own network stack
- No raw packets, no packet capturing, limited `/dev`

Solution: Wrap a classic jail in a VNET jail

- We get pf, `lo0`, and can do mounts, monitoring, etc
- Parent jail filesystems unmounted after starting child
- Has proven flexible and stable
ORCHESTRATING YOUR PETS / HERDING CATS

- Using Puppet on hosts and in jails
- Automatic ZFS snapshots during Puppet runs (nrpe)
- Live-configure and re-configure VNET pf with Puppet
- Auto-create pf rules based on in-jail nginx config
  - Using Puppet 'exported resources'
- Kernel audit logs and FIM on hosts
  - Stealthy observation, very very sneaky
Snap me baby one more time:

MySQL+ZFS
ABOUT CLUSTERS

- Galera is a cluster engine used in MySQL, MariaDB, Percona
- Synchronous: All nodes "certify" each transaction: Check locks, constraints, etc.
- High QPS requires very low latency on and between nodes
  - Multi-site clusters effectively impossible (we tried)
- A long-running ALTER locks the entire cluster (don't ask)
  - Percona-toolkit to the rescue!
PERFORMANCE SHOULD SUCK, RIGHT?

- ZFS isn't so bad after all:
  - Compression (3-4x in our case) makes good use of CPU, and ZFS encryption is awesome
  - ZIL makes NAND wear largely a non-issue, especially with large record sizes
  - Snapshots + jails enables recovery from stupid in minutes
- Bottleneck is when ZFS flushes to disk
  - Increased interval (10s) helped a lot
  - SATA has unpredictable latency; moving to NVMe
Where all the garbage lands
MORE NGINX CONFIG HACKS

- nginx can take a beating - but also mess with attackers
- More protection for your upstream application:
  - Brutally drop bad requests (444); this is no time for protocol
  - Integrate with application to rate-limit based on session tokens or similar
- LUA plugin is insanely powerful
- Hard limit: TLS handshakes per second
  - Cannot "outsource" to cloud for various reasons
- Session token in hostname?
- Abusing server name indicator (SNI)
- Rate-limit in TLS handshake: `ssl_client_hello_by_lua_*`
- DNS server load might be an issue..
- Work in progress, brainstorming with smart(er) people
... mumble ... reuseport ... don't do that ... <ALARM BELL> ... 

... oh thank the Great Maker, it was just a dream ...
I think you're wrong
(Or something to that effect)

Netflix person to slightly star-struck impostor
Awesome debugging session around our table in Vienna

Turns out jails were skipped over for SO_REUSEPORT_LB

One-liner change to fix for VNET

Bigger change for all jails in CURRENT (D37029)

Lots of smart people from lots of places involved. Thank you!
A STORY IN SEVERAL FRAGMENTS:

WHO'S EATING OUR PACKETS?!?
RANDOM PACKET DELAY IS RANDOM

- Constant trickle of retransmits for no good reason
- In-house troubleshooting got us ~nowhere
- HW supplier provided ~free test rig
- Klara folks reproduced and proposed fix
- Other people chimed in with more improvements
- reviews.freebsd.org-conversation was inspiring (D38843)
  - ...and fixed this particular problem
BUT WAIT, THERE ARE MORE MISSING PACKETS!

- Turns out this wasn't the only bug
- `relayd` used for L3 load balancing (NAT)
- Reloading config kills (some) traffic
- Klara helped update `relayd` port to match upstream
- Turns out backend hosts are marked down until health check has run
  - Any OpenBSD people here who can help?
COMPLAINTS AND WHINES

- Not much has changed:
  - Document `sysctl` and `pf` defaults (include rationales and implications of changing them!)
  - PF syncookies should mirror kernel syncookies
  - Jail/container management
- Console logs (oom kills, network/pf errors, whatever) should state jail ID!
- Contributing ports is still too frustrating
THANK YOU ALL!

- Contributors of all kinds
- Organisers of this event
- Everyone working to make the community tick
- My esteemed colleagues
- All the unsung OSS heroes (there's an XKCD for that)
  - Now if we could get that sponsor bidding war going..

See you next year!