FreeBSD Networking in Virtualised Hosting

Patrick M. Hausen
EuroBSDCon 2021, not Vienna
Agenda

- Introduction
- Our Current Architecture
- Going IPv6 Only
- Layer 2 Challenges
- What's Wrong with Ethernet?
About Me

- Working in IT since 1986
- Minix 1.1 since 1989
- FreeBSD since 1993
- In charge of network and data centre operations at punkt.de
About Our Team

• mOps – the Magnificent Operators
• 3 (originally) operators
• 1 (originally) developer
About punkt.de

- Founded in 1996
- Started as an ISP
- Today: Hosting and development of web applications
- Roughly 100 Servers
- RIPE Member
- DENIC Member
- 4 development, 1 operations team
VIMAGE/VNET

- Introduces the epair(4) virtual interface
- Essentially a virtual patch cable
- One end inside the jail, other end on the host system
- Bridge, route, NAT to your heart’s content
System Architecture

Jail
vnet0

Jail
vnet0

Jail
vnet0

Jail
vnet0

Bridge Interface

igb0, ix0, …

IPv6 and IPv4

LAN
Inside View (Dual Stack)

epair0b: flags=8843 [...]  
options=8<VLAN_MTU>  
ether 12:15:08:76:d1:6c  
hwaddr 02:9c:87:4a:c2:0b  
inet6 fe80::1015:8ff:fe76:d16c%epair0b prefixlen 64 [...]  
inet6 2a00:b580:8000:11:d852:25bf:5d4d:c275 prefixlen 64  
inet 217.29.41.210 netmask 0xffffffff00 [...]  
groups: epair  
media: Ethernet 10Gbase-T (10Gbase-T <full-duplex>)  
status: active  
dx6 options=21<PERFORMNUD,AUTO_LINKLOCAL>
Going IPv6 Only

- **Gate64**
  - vnet0
  - IPv6 and IPv4

- **Jail**
  - vnet0
  - IPv6

- **Jail**
  - vnet0
  - IPv6

- **Jail**
  - vnet0
  - IPv6

**Bridge Interface**

- ipv0, ix0, ...

**IPv6 and IPv4**

**LAN**
Going IPv6 Only - Egress

- NAT64
- RFC 6052, 6146
- Uses the 64:ff9b::/96 address range
- Route that range through Gate64
- IPFW does the NAT
- Resolver needs to "lie" about AAAA records

Short demo
Going IPv6 Only - Ingress

- SNI proxy
- Supports HTTP and HTTPS
- HTTP is important for Letsencrypt
- AAAA Record points to jail proper
- A record points to gate64 jail
Going IPv6 Only - Ingress

- Connect to SNI proxy via IPv4
- Request with SNI (hostname)
- Proxy looks up IPv6 in DNS
- Checks if permitted address range
- If permitted, connects via TCP (not HTTP!)

Most common problem for customers: forget to set AAAA record
Going IPv6 Only - Ingress

- Native IPv6 or jumphost for SSH
- SSH tunnels for everything else (sorry!)

- What about QUIC?
Layer 2 Challenges

https://bugs.freebsd.org/bugzilla/show_bug.cgi?id=227100

• Epair bug – only happened in production
• Near impossible to reproduce
• Happened more frequently as the DC grew
• Interface stopped forwarding packets when "hardware" queue filled up
• So what's different in production?
Broadcasts!

40 percent of all packets sent or received to/from a single hosting server are broadcast/multicast!
Solutions

- Increase queue length
- Fix bug ;-)  
- Still not optimal  
- How can we get rid of the broadcasts?
Move the Bridge off the Wire

**Gate64**
- vnet0 IPv6 and IPv4

**Jail**
- vnet0 IPv6

**Jail**
- vnet0 IPv6

**Jail**
- vnet0 IPv6

**Bridge Interface** – acts as default gateway

Host Interface: igb0, ix0, ...

IPv6 and IPv4

LAN

Short demo
Routing might be a good idea!

- Jail/VM mobility is a problem with Layer 2
- All the big guys do it
- Start jail, announce routes via BGP
- Downside: needs dynamic routing protocol on each host
What's wrong with Ethernet?

It's this ...
What's wrong with Ethernet?

But we pretend it's this ...
What's wrong with Ethernet?

- It's all point to point links
- Full duplex
- With flow control
- Switches actively work to forward broadcasts
- MAC addresses are a relic of the past
So where to go?

- Can we have a VNET point to point IF? Please?

- No IP addresses on IF, just routes
- No leaking RFC 1918 transfer networks
- No arp/ndp cache depletion "attacks"
In the Meantime

- /32 and /128 epair
- IP addresses can be re-used with /32 ("unnumbered")
- Interface routes do the rest
- Redistribute with OpenBGPd, FRR, ...
- Mobility problem solved
vnet interface - no bridge!

Short demo
Open Issues

• Solve remaining routing problem
• Integrate with iocage and/or Bastille
• Discuss possibility of a true point to point interface
Questions/Discussion?
Thanks!