FreeBSD Networking in Virtualised Hosting

Patrick M. Hausen

EuroBSDCon 2021, not Vienna

Agenda

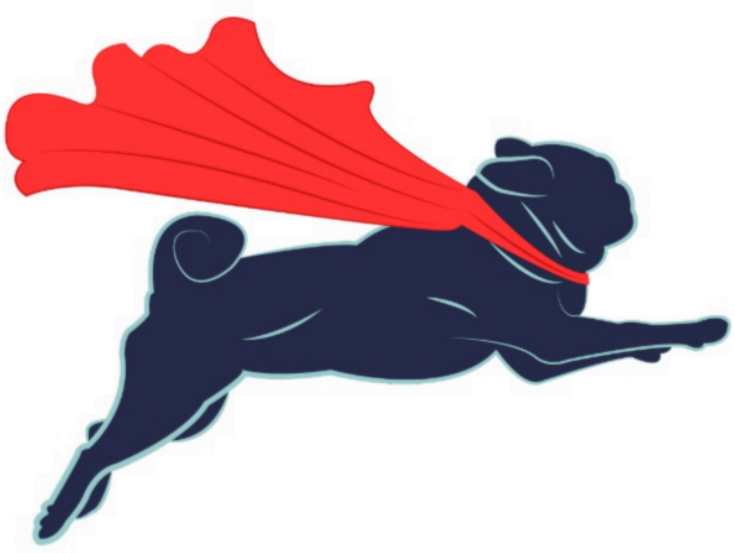
- Introduction
- Our Current Architecture
- Going IPv6 Only
- Layer 2 Challenges
- What's Wrong with Ethernet?

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



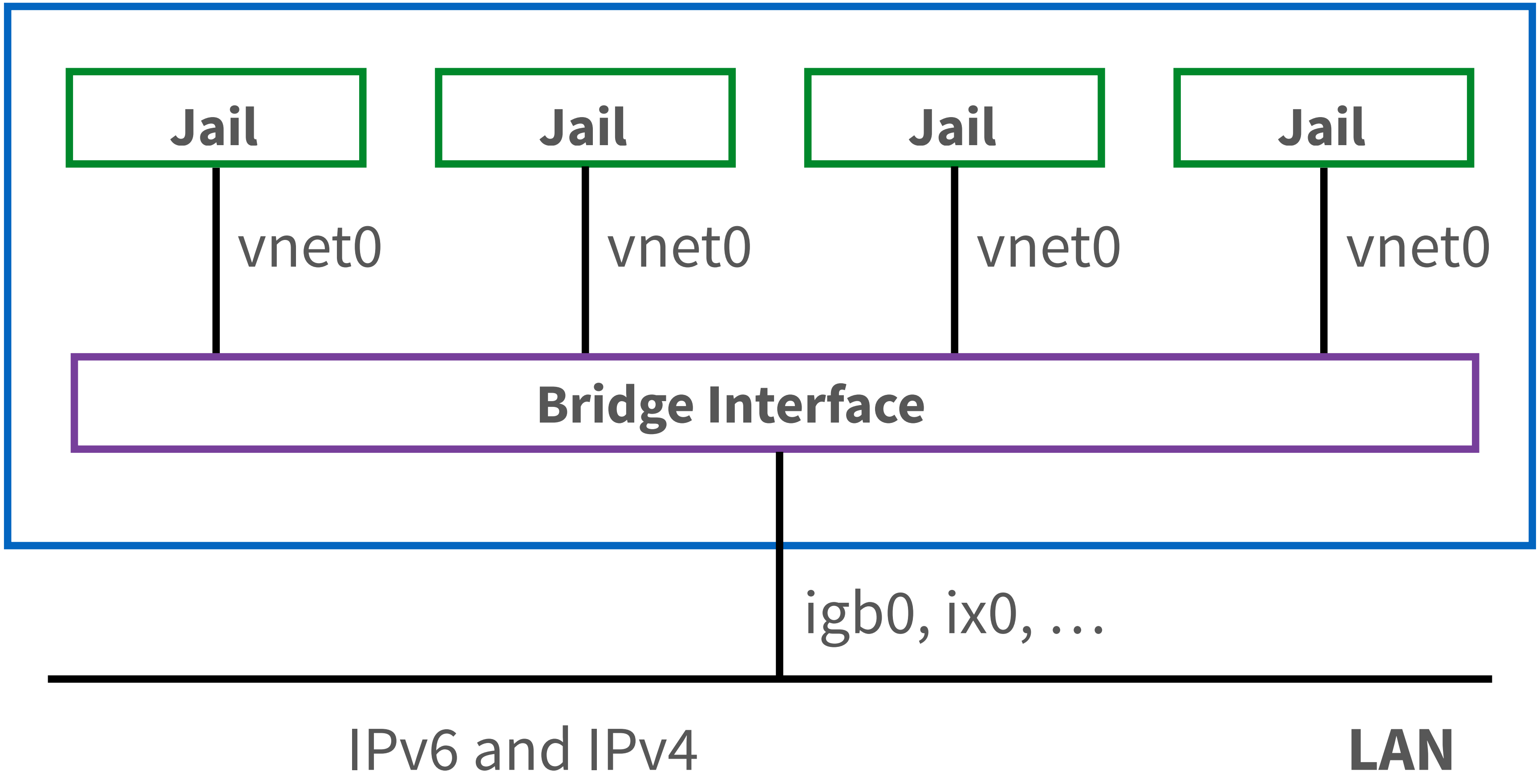
About punkt.de

- 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

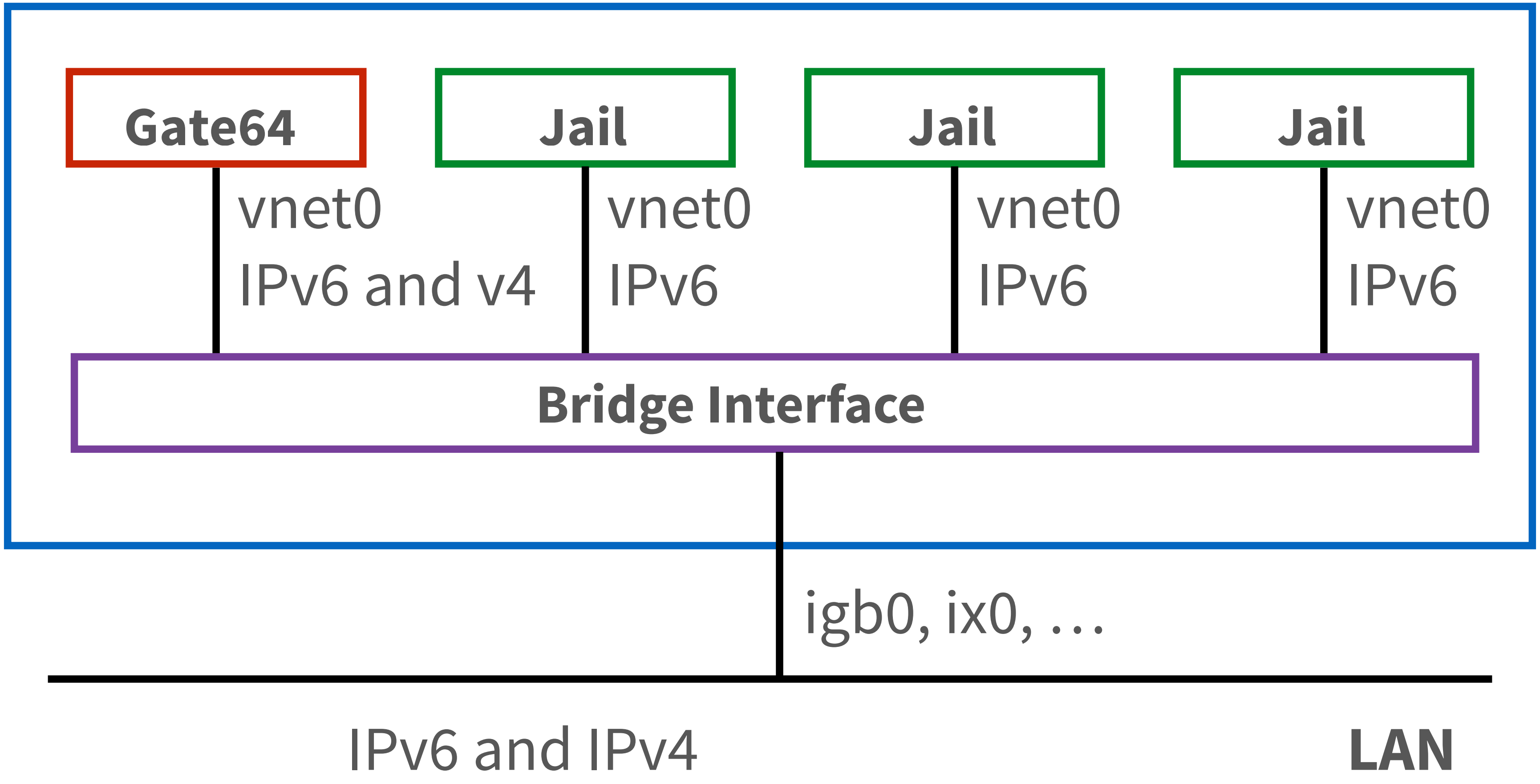
System Architecture



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 0xffffff00 [...]
  groups: epair
  media: Ethernet 10Gbase-T (10Gbase-T <full-duplex>)
  status: active
  nd6 options=21<PERFORMNUD,AUTO_LINKLOCAL>
```

Going IPv6 Only



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!

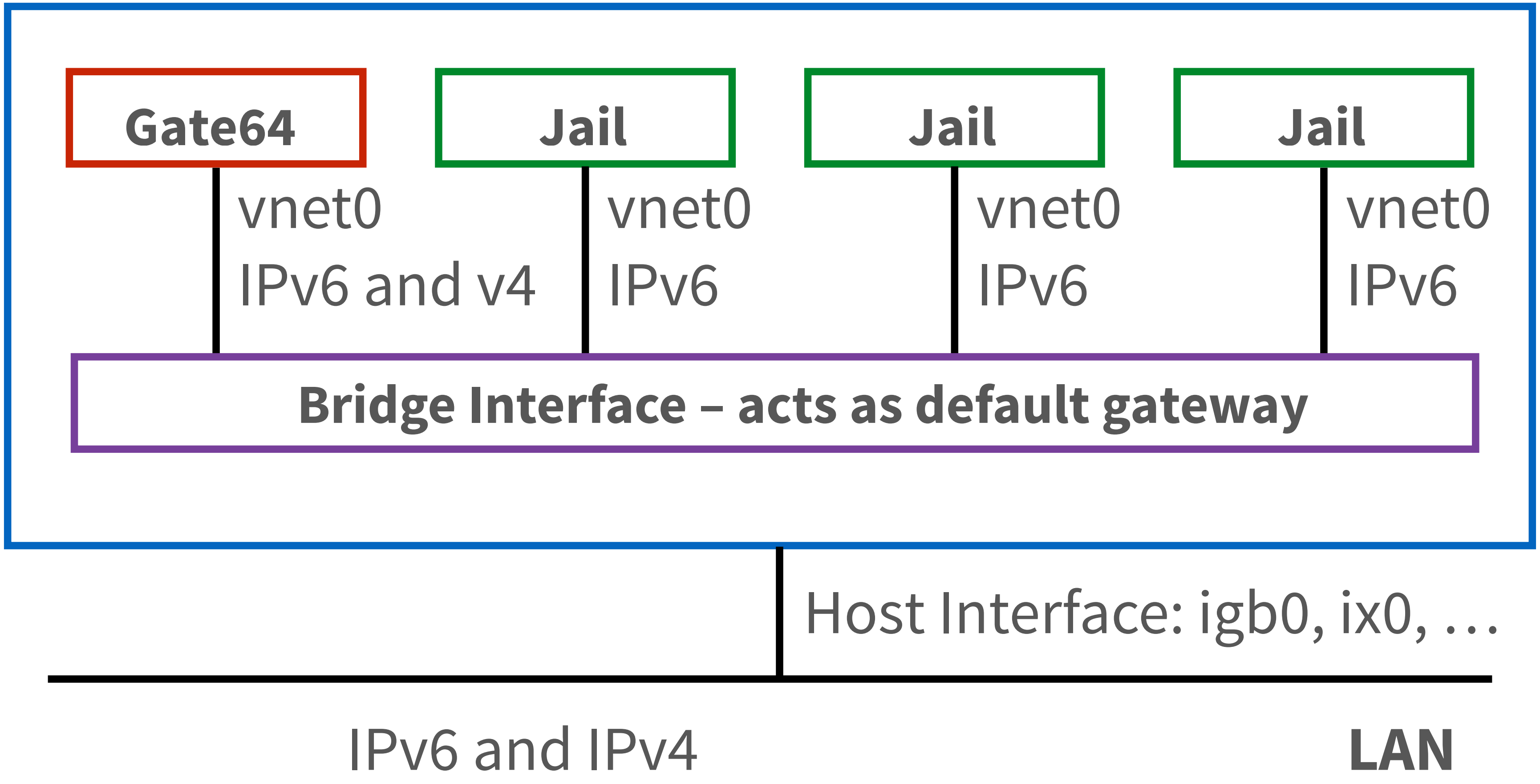
40 percent of all packets sent or received to/from a single hosting server are broadcast/multicast!

Solutions

- Increase queue length
- Fix bug ;-)

- Still not optimal
- How can we get rid of the broadcasts?

Move the Bridge off the Wire

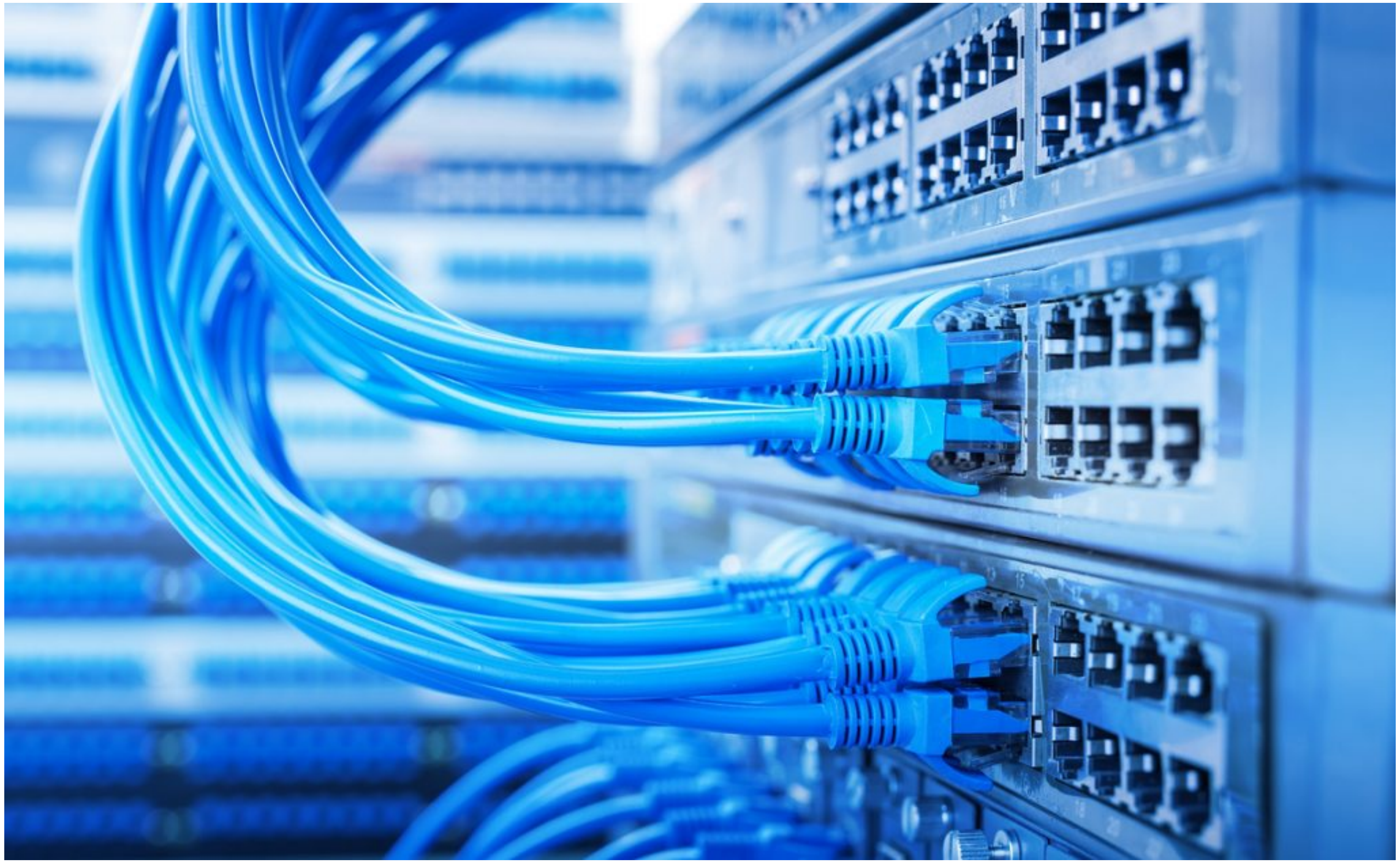


Short demo

Routing might be a good idea!

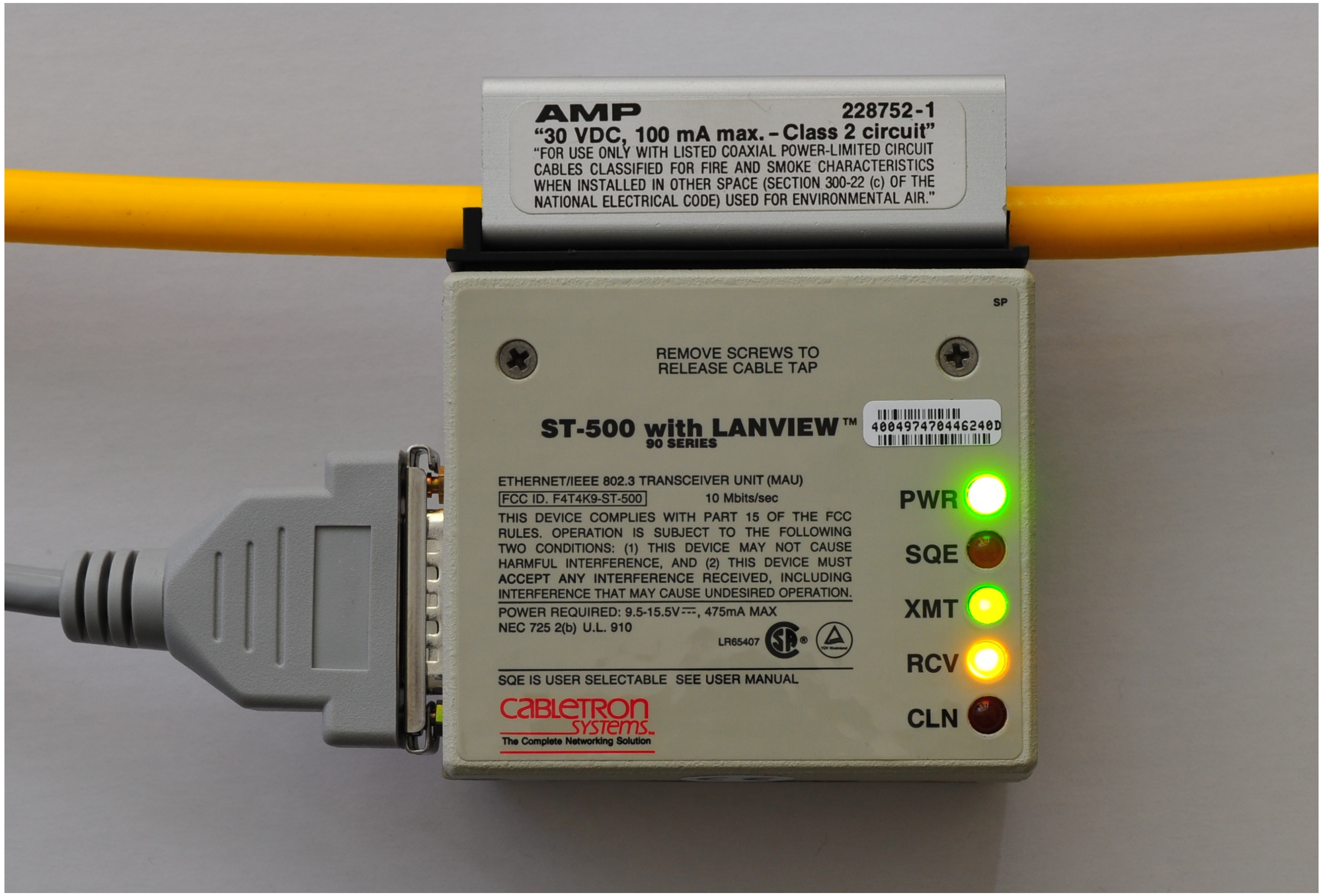
- Jail/VM mobility is a problem with Layer 2
- All the big guys do it
- Start jail, announce routes via BGP
- Downside: needs dynamic routing protocol on each host

What's wrong with Ethernet?



It's this ...

What's wrong with Ethernet?


AMP 228752-1
"30 VDC, 100 mA max. - Class 2 circuit"
"FOR USE ONLY WITH LISTED COAXIAL POWER-LIMITED CIRCUIT CABLES CLASSIFIED FOR FIRE AND SMOKE CHARACTERISTICS WHEN INSTALLED IN OTHER SPACE (SECTION 300-22 (c) OF THE NATIONAL ELECTRICAL CODE) USED FOR ENVIRONMENTAL AIR."
SP
REMOVE SCREWS TO RELEASE CABLE TAP
ST-500 with LANVIEW™
90 SERIES
400497470446240D
ETHERNET/IEEE 802.3 TRANSCEIVER UNIT (MAU)
FCC ID. F4T4K9-ST-500
10 Mbits/sec
THIS DEVICE COMPLIES WITH PART 15 OF THE FCC RULES. OPERATION IS SUBJECT TO THE FOLLOWING TWO CONDITIONS: (1) THIS DEVICE MAY NOT CAUSE HARMFUL INTERFERENCE, AND (2) THIS DEVICE MUST ACCEPT ANY INTERFERENCE RECEIVED, INCLUDING INTERFERENCE THAT MAY CAUSE UNDESIRED OPERATION.
POWER REQUIRED: 9.5-15.5V⎓, 475mA MAX
NEC 725 2(b) U.L. 910
LR65407
SQE IS USER SELECTABLE SEE USER MANUAL
CABLETRON SYSTEMS
The Complete Networking Solution
PWR
SQE
XMT
RCV
CLN

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?

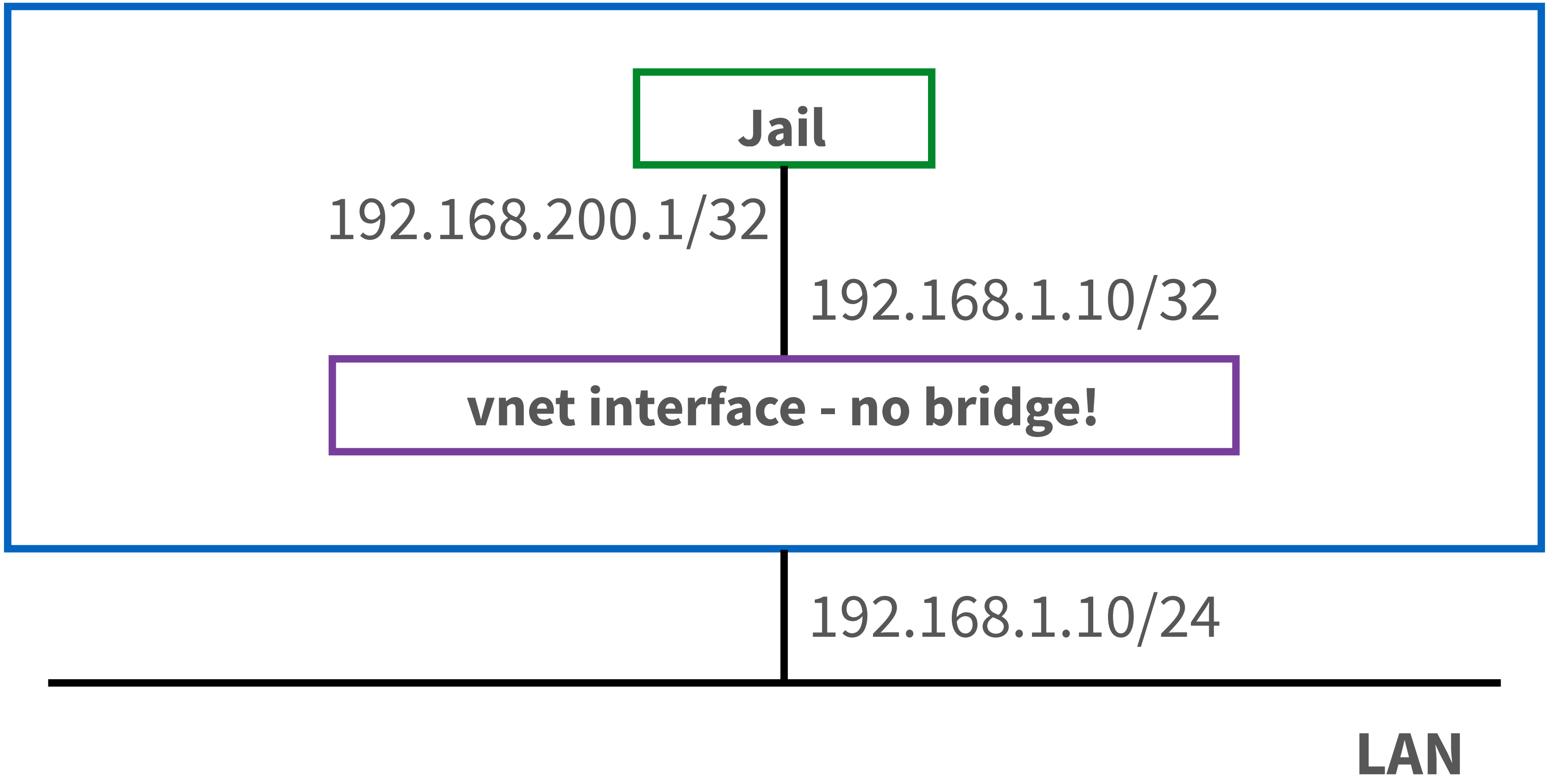
- Can we have a VNET point to point IF?
- Please?

- 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

/32 epair



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!