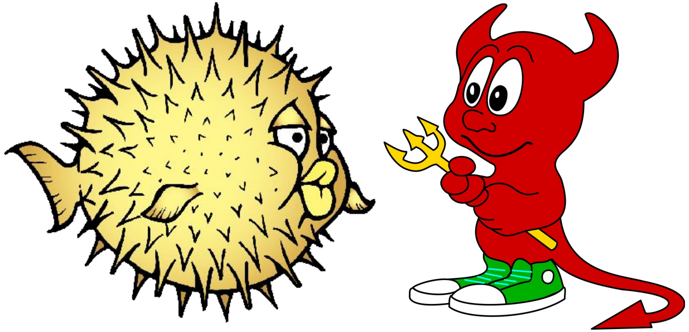
Introduction to global anycast using OpenBSD (on a budget)

TheLongCon 2024

The next ~20 minutes

- What is anycast?
- Why does it exist?
- Who does it?
- How expensive is it?
- How can I do it?
- Demo

If you take one thing away from this talk…



Puffy the OpenBSD mascot with Beastie, the BSD Daemon (© Kirk McKusick)

There are many other operating systems besides Linux

Anycast primer

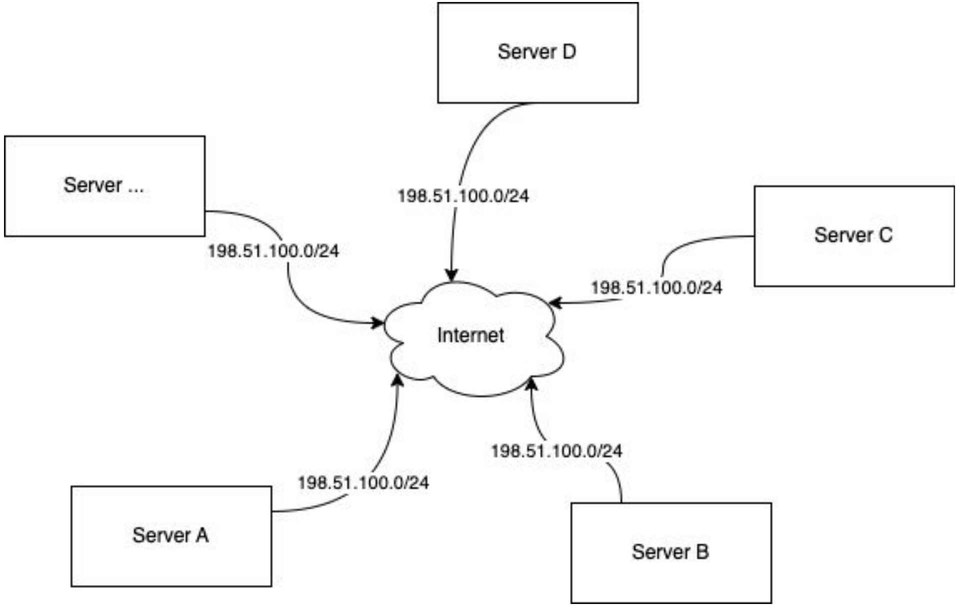
- Multiple servers
- Multiple locations
- Multiple ISPs
- The **same** IP addresses

Why?

- Lowering latency
- Load distribution
- DoS mitigation (single D)

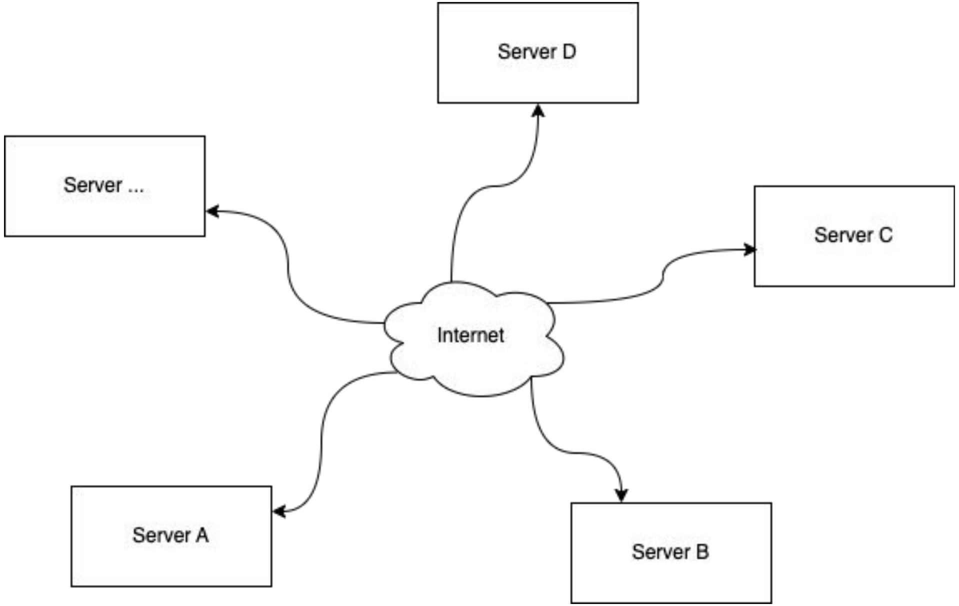
Note

- No control over inbound traffic



Why I cared…

- Single point of failure is administrative in nature.
- Volumetric denial of service attacks should be (relatively) geographically bounded.
- DDoS mitigations require scaling up the resources, but don’t require massive architecture changes.



Why you might care…

- Some things are outside your control
- Single points of administrative failure
- Resilience requires full stack, for some definition of full stack

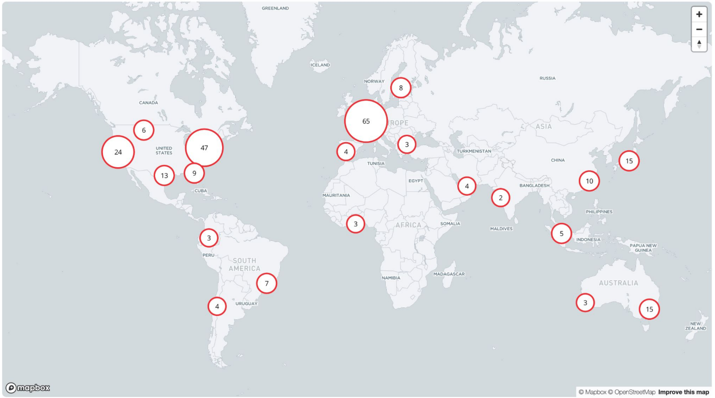
Any individual PoP will likely go down

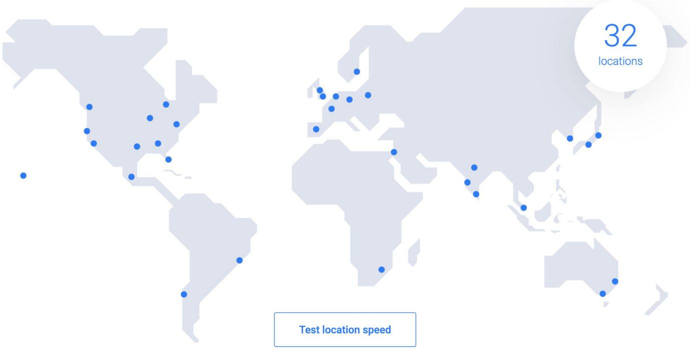
- Definitions for pedants
  - Available: “able to be used or obtained…”
  - Resilience: “provide and maintain an acceptable level of service in the face of…”
  - Reliable: “ability … to function under stated conditions for a specified period of time.”
- How many 9s?
  - Tier 2: 99.74% (~22 hours/year)
  - Tier 3: 99.98% (~1.6 hours/year)
  - Tier 4: 99.99% (~26 minutes/year)
- At the end of it all
  - “The website didn’t load”
  - “My email didn’t send”
  - “I couldn’t do X”

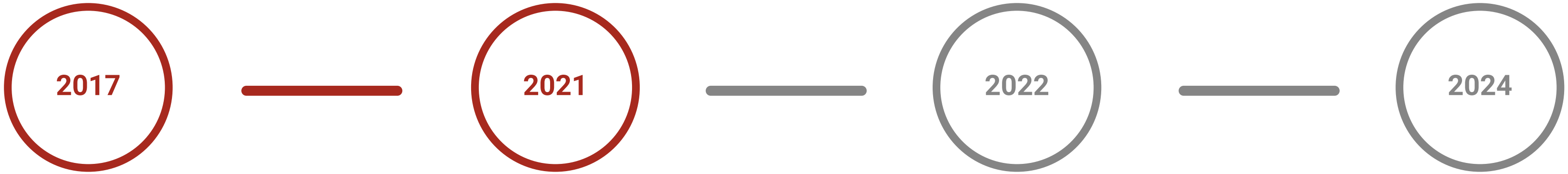
So have multiple PoPs

- You own the hardware
- Full control
- Generally fixed/known costs
  - Orders of magnitude savings at any serious scale

- You rent the resources
- Small/no setup time
- Usage based billing
- Easy to get started







2017

AWS

Roughly 4 hour global outage due to S3 configuration change and cascading failures

2021

Facebook/Meta

6-7 hour global outage due to BGP configuration change gone wrong

2022

Rogers

Roughly 15 hour outage due to BGP configuration change gone wrong

2024

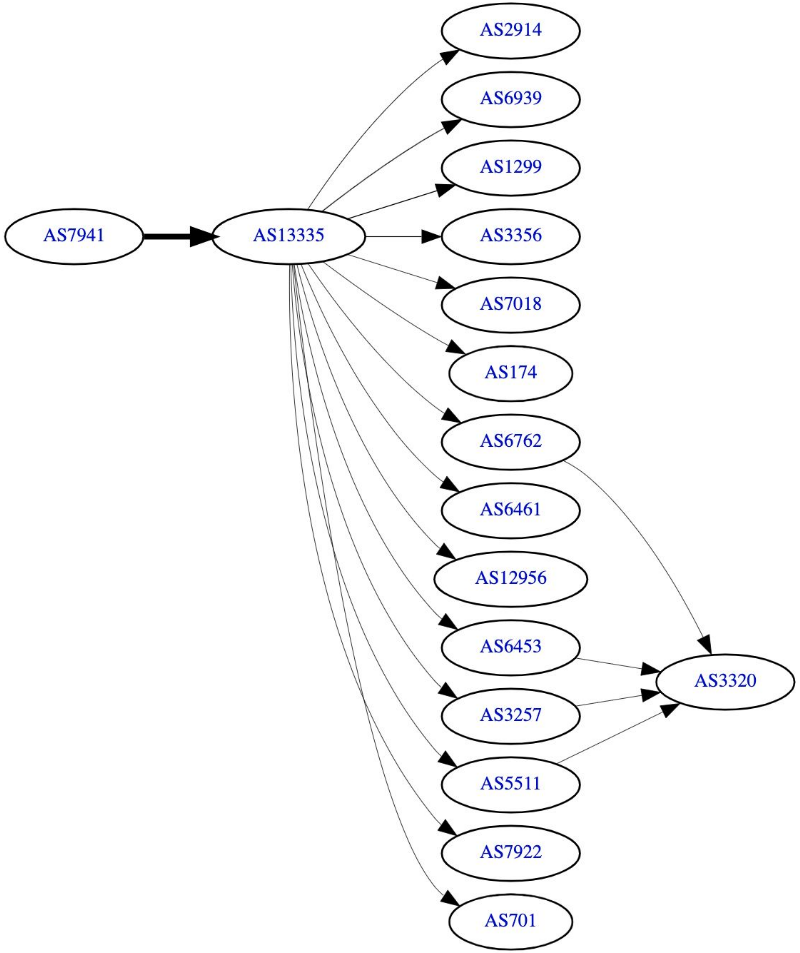
Cloudflare

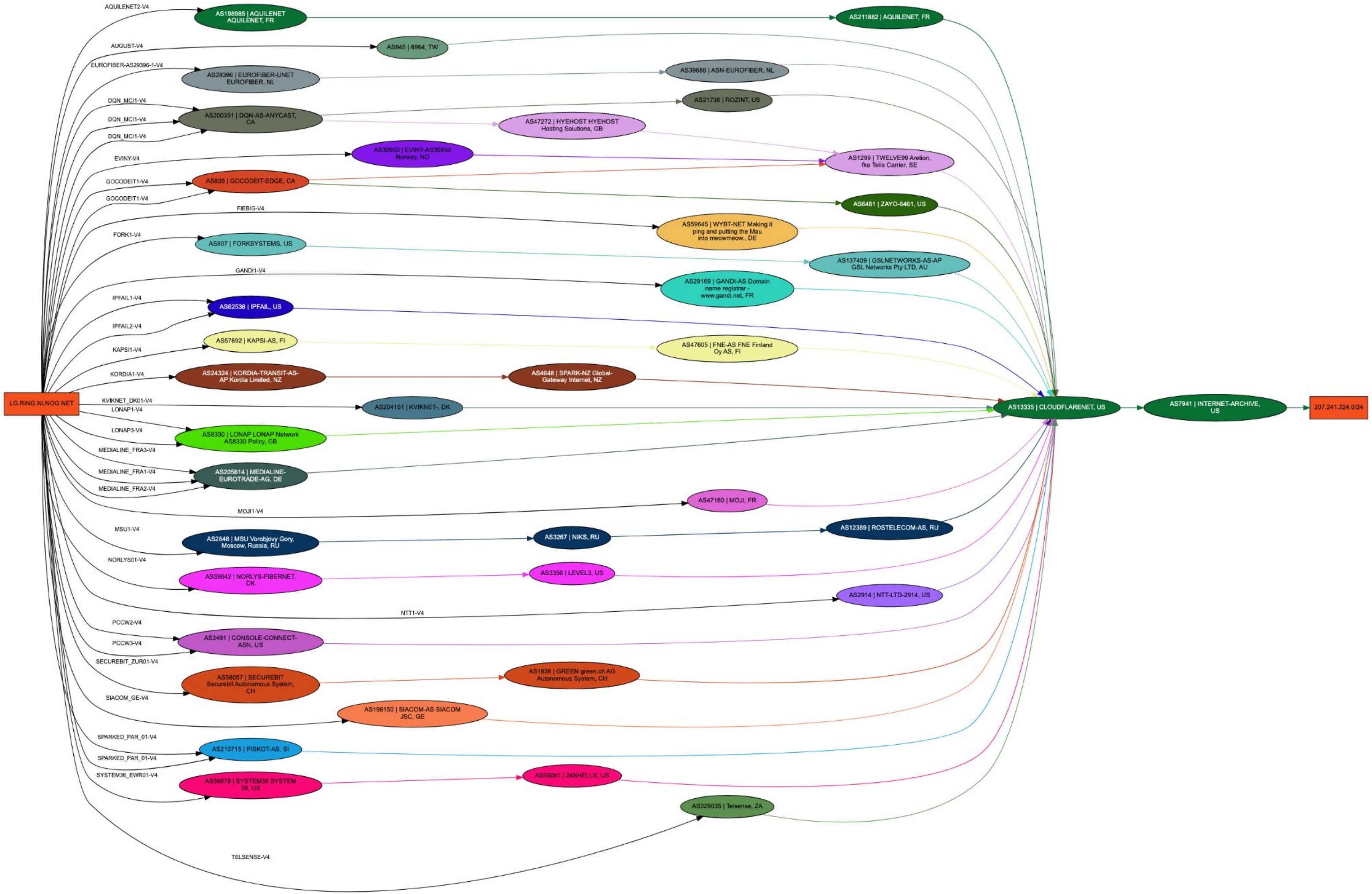
Roughly 4 hour disruption and degradation due to some ISP in Brazil pushing a BGP configuration change

Who does this (anycast)?

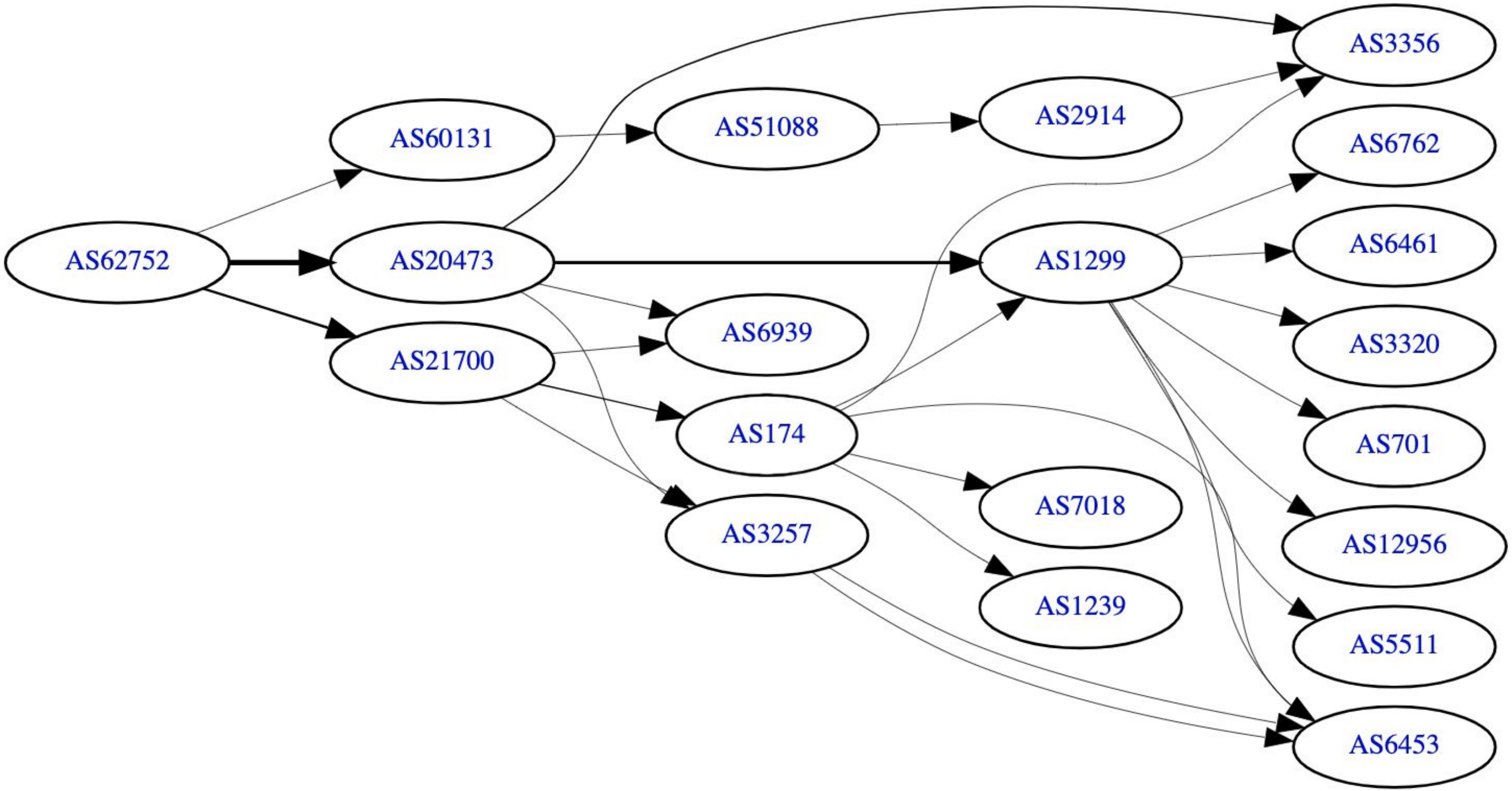
- Content delivery networks
- DDoS mitigation companies
- Critical infrastructure
  - Mainly DNS related
- “At Scale” infrastructure
  - Global scale
  - >100gbits
  - Usual suspects

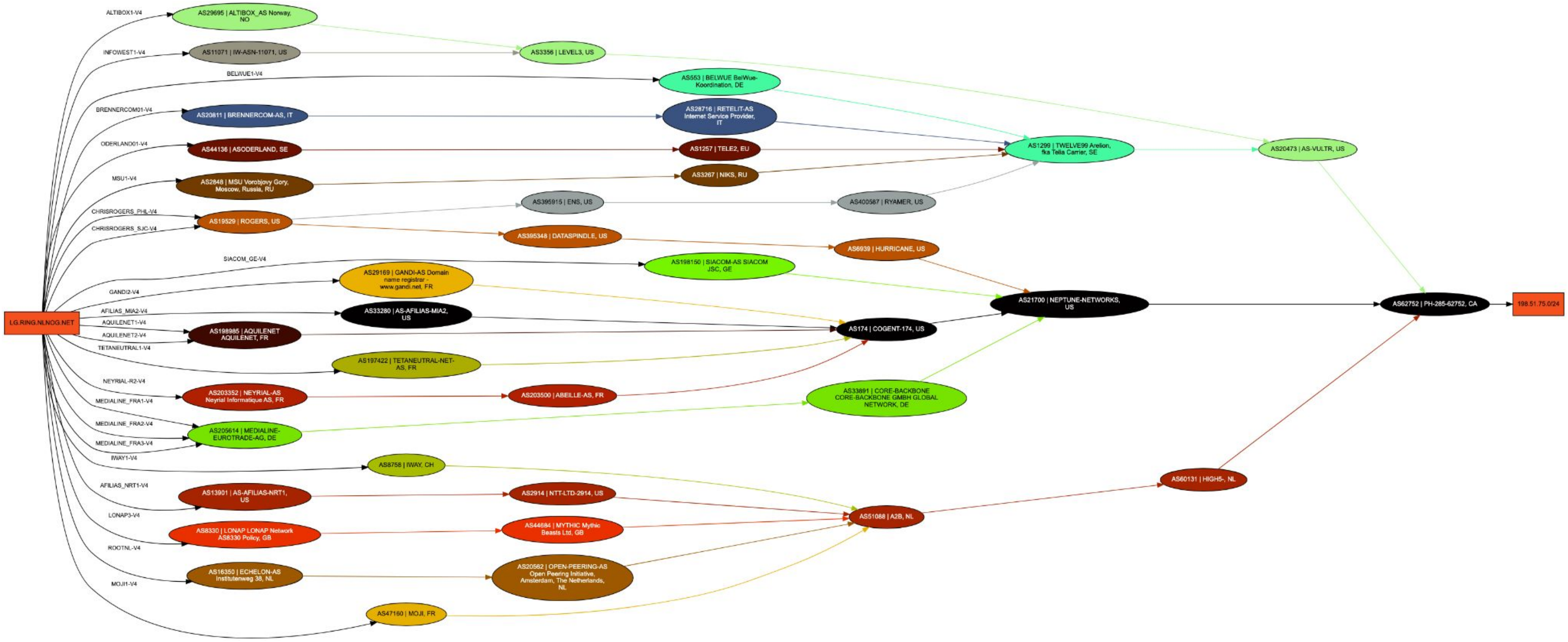
Origin AS7941




LG.RING.NLNOG.NET
AQUILENET2-V4
AS198985 | AQUILENET AQUILENET, FR
AS211882 | AQUILENET, FR
AUGUST-V4
AS945 | 8964, TW
EUROFIBER-AS29396-1-V4
AS29396 | EUROFIBER-UNET EUROFIBER, NL
AS39686 | ASN-EUROFIBER, NL
DQN_MCI1-V4
DQN_MCI1-V4
DQN_MCI1-V4
AS200351 | DQN-AS-ANYCAST, CA
AS21738 | ROZINT, US
AS47272 | HYEHOST HYEHOST Hosting Solutions, GB
EVINY-V4
AS30950 | EVINY-AS30950 Norway, NO
AS1299 | TWELVE99 Arelion, fka Telia Carrier, SE
GOCODEIT1-V4
GOCODEIT1-V4
AS835 | GOCODEIT-EDGE, CA
AS6461 | ZAYO-6461, US
FIEBIG-V4
AS59645 | WYBT-NET Making it ping and putting the Mau into meowmeow., DE
FORK1-V4
AS807 | FORKSYSTEMS, US
AS137409 | GSLNETWORKS-AS-AP GSL Networks Pty LTD, AU
GANDI1-V4
AS29169 | GANDI-AS Domain name registrar - www.gandi.net, FR
IPFAIL1-V4
IPFAIL2-V4
AS62538 | IPFAIL, US
KAPSI1-V4
AS57692 | KAPSI-AS, FI
AS47605 | FNE-AS FNE Finland Oy AS, FI
KORDIA1-V4
AS24324 | KORDIA-TRANSIT-AS-AP Kordia Limited, NZ
AS4648 | SPARK-NZ Global-Gateway Internet, NZ
KVIKNET_DK01-V4
AS204151 | KVIKNET-, DK
AS13335 | CLOUDFLARENET, US
AS7941 | INTERNET-ARCHIVE, US
207.241.224.0/24
LONAP1-V4
LONAP3-V4
AS8330 | LONAP LONAP Network AS8330 Policy, GB
MEDIALINE_FRA3-V4
MEDIALINE_FRA1-V4
MEDIALINE_FRA2-V4
AS205614 | MEDIALINE-EUROTRADE-AG, DE
MOJI1-V4
AS47160 | MOJI, FR
MSU1-V4
AS2848 | MSU Vorobjovy Gory, Moscow, Russia, RU
AS3267 | NIKS, RU
AS12389 | ROSTELECOM-AS, RU
NORLYS01-V4
AS39642 | NORLYS-FIBERNET, DK
AS3356 | LEVEL3, US
NTT1-V4
AS2914 | NTT-LTD-2914, US
PCCW2-V4
PCCW3-V4
AS3491 | CONSOLE-CONNECT-ASN, US
SECUREBIT_ZUR01-V4
AS58057 | SECUREBIT Securebit Autonomous System, CH
AS1836 | GREEN green.ch AG Autonomous System, CH
SIACOM_GE-V4
AS198150 | SIACOM-AS SIACOM JSC, GE
SPARKED_PAR_01-V4
SPARKED_PAR_01-V4
AS210715 | PISKOT-AS, SI
SYSTEM36_EWR01-V4
AS59678 | SYSTEM36 SYSTEM 36, US
AS55081 | 24SHELLS, US
AS329035 | Telsense, ZA
TELSENSE-V4

Origin AS62752





LG.RING.NLNOG.NET
ALTIBOX1-V4
INFOWEST1-V4
BELWUE1-V4
BRENNERCOM01-V4
ODERLAND01-V4
MSU1-V4
CHRISROGERS_PHL-V4
CHRISROGERS_SJC-V4
SIACOM_GE-V4
GANDI2-V4
AFILIAS_MIA2-V4
AQUILENET1-V4
AQUILENET2-V4
TETANEUTRAL1-V4
NEYRIAL-R2-V4
MEDIALINE_FRA1-V4
MEDIALINE_FRA2-V4
MEDIALINE_FRA3-V4
IWAY1-V4
AFILIAS_NRT1-V4
LONAP3-V4
ROOTNL-V4
MOJI1-V4
AS29695 | ALTIBOX_AS Norway, NO
AS11071 | IW-ASN-11071, US
AS3356 | LEVEL3, US
AS553 | BELWUE BelWue-Koordination, DE
AS20811 | BRENNERCOM-AS, IT
AS28716 | RETELIT-AS Internet Service Provider, IT
AS44136 | ASODERLAND, SE
AS1257 | TELE2, EU
AS2848 | MSU Vorobjovy Gory, Moscow, Russia, RU
AS3267 | NIKS, RU
AS395915 | ENS, US
AS400587 | RYAMER, US
AS19529 | ROGERS, US
AS395348 | DATASPINDLE, US
AS6939 | HURRICANE, US
AS1299 | TWELVE99 Arelion, fka Telia Carrier, SE
AS20473 | AS-VULTR, US
AS198150 | SIACOM-AS SIACOM JSC, GE
AS29169 | GANDI-AS Domain name registrar - www.gandi.net, FR
AS33280 | AS-AFILIAS-MIA2, US
AS198985 | AQUILENET AQUILENET, FR
AS174 | COGENT-174, US
AS21700 | NEPTUNE-NETWORKS, US
AS62752 | PH-285-62752, CA
198.51.75.0/24
AS197422 | TETANEUTRAL-NET-AS, FR
AS203352 | NEYRIAL-AS Neyrial Informatique AS, FR
AS203500 | ABEILLE-AS, FR
AS33891 | CORE-BACKBONE CORE-BACKBONE GMBH GLOBAL NETWORK, DE
AS205614 | MEDIALINE-EUROTRADE-AG, DE
AS8758 | IWAY, CH
AS13901 | AS-AFILIAS-NRT1, US
AS2914 | NTT-LTD-2914, US
AS51088 | A2B, NL
AS60131 | HIGH5-, NL
AS8330 | LONAP LONAP Network AS8330 Policy, GB
AS44684 | MYTHIC Mythic Beasts Ltd, GB
AS16350 | ECHELON-AS Institutenweg 38, NL
AS20562 | OPEN-PEERING-AS Open Peering Initiative, Amsterdam, The Netherlands, NL
AS47160 | MOJI, FR

~$10k

/24 of IPv4 space

Wait for years for the ability to pay, or go to the private market.

- Use IPv6.
- Option to lease at ~$100-200/month
- Run your services from Africa

$300

Annual fees

Paid to ARIN, the regional registrar.

- ~€1,800 RIPE region. Be thankful.

~$100

Monthly

For some amount of compute and network on other peoples machines.

- “Lower” costs by running your own hardware.

Pricing and costs - Location location location

- Name resources
  - ASN / IP space
- VMs
  - Compute resources
  - Outbound bandwidth
- Physical
  - Hardware costs
  - Colocation
  - Transit
  - IX

$10-25

Usually measured in bytes per month
Only egress

$100

$75

$200 - 1gbit

$100

Who do I talk to?

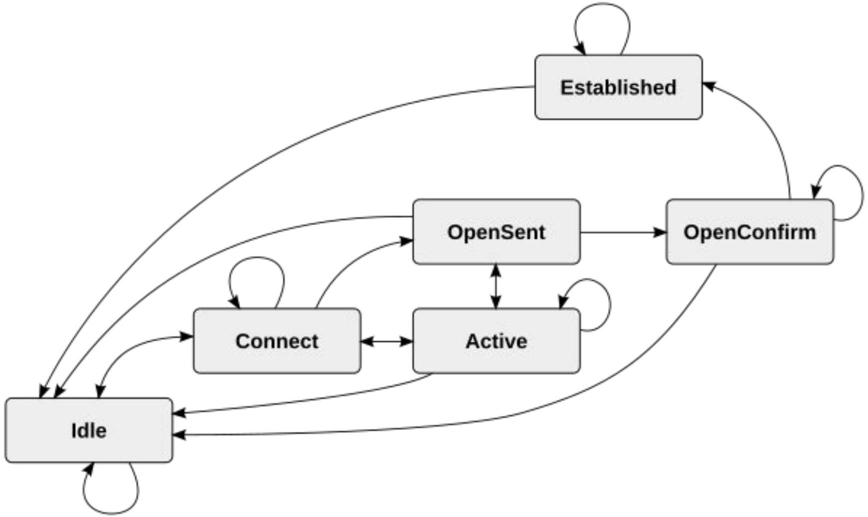
- **bgp.services (211 providers listed..)**
- datapacket.com (CDN77)
- vultr.com
- openbsd.amsterdam
- anexia.com
- neptunenetworks.org
- coloclue.net
- equinix.com
- he.net
- … take your pick

How can I do it?

- Advertise BGP
  - Border Gate Protocol
- Do that from multiple places

Some details not covered in this talk:

- RPKI and IRR
- Bogons
- Communities
- Prefix limits



External Networking - pf.conf(5)

```
set skip on lo
block in all
pass out keep state

…

ext_if = "vio0"
upstream_peers = “169.254.169.254 198.51.100.1”
pass in on $ext_if proto tcp from any to $ext_if:0 port 22 keep state (no-sync)
pass in on $exit_if proto tcp from $upstream_peers to $ext_if:0 port 179 keep state (no-sync)
```

rdomain and egress filtering

Lots of details omitted

pfsync

- BGP on port 179
- Remote access (ssh/ipsec/wg)

Internal Networking - VPN/Routing

- Do you need it?
  - Comes down to state
- tailscale / headscale / wg
  - Tailscale proper
  - Headscale (single “tailnet”)
  - Distribute the public keys and IPs yourself
- tinc
  - Distribute public keys and IPs
- ipsec
  - iked
  - isakmpd

Certificates!

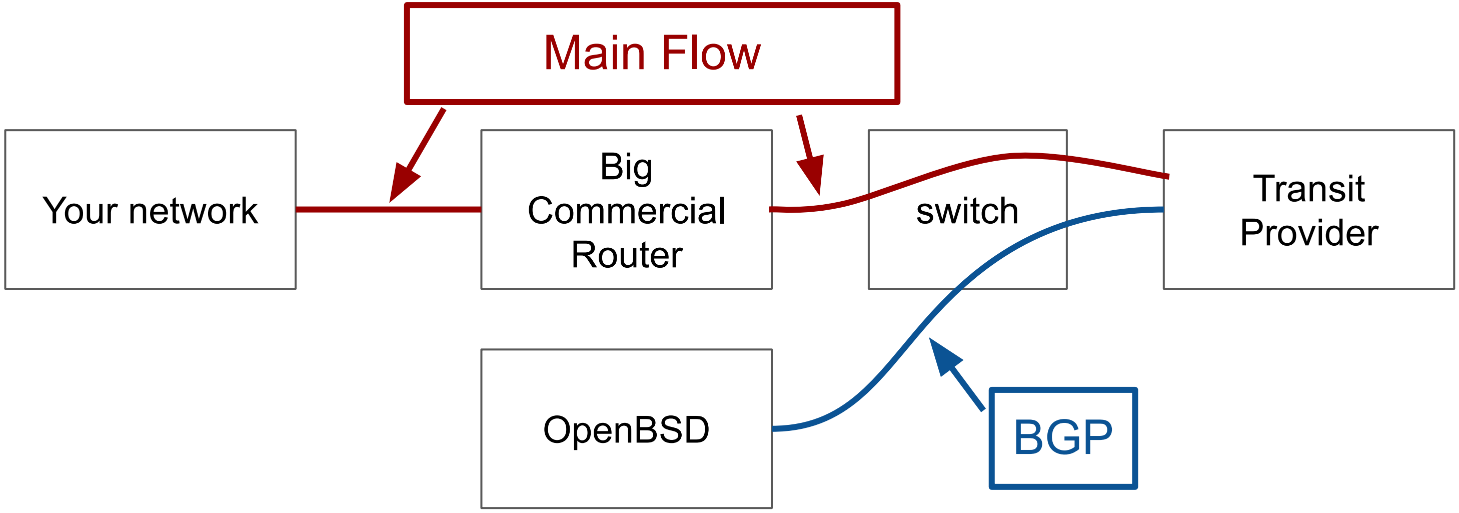
- etherip
  - Bridge all the things! (cringe, don’t do this)
- vxlan
  - learning vs static
- GRE

Internal Networking - pfsync(4)

- Protocol and daemons to sync firewall state
- Do you need it?
- Option to multicast to other nodes
- Use the *defer* option to attempt to ensure sync
  - Large number of nodes makes this improbable
  - Race condition at scale

External Networking - note about BGP speakers

- Transit connections will likely hand you a /30 or /29 of ipv4 space
- Can can point traffic at another host on the same network



elixir

- In packages (pkg_add elixir)
- functional, concurrent, high-level general-purpose
- Runs on the BEAM
  - You may have heard of erlang
- Processes and message passing
- Known for Phoenix
  - MVC framework supporting http and websocket
- Generally used for (soft) real-time systems

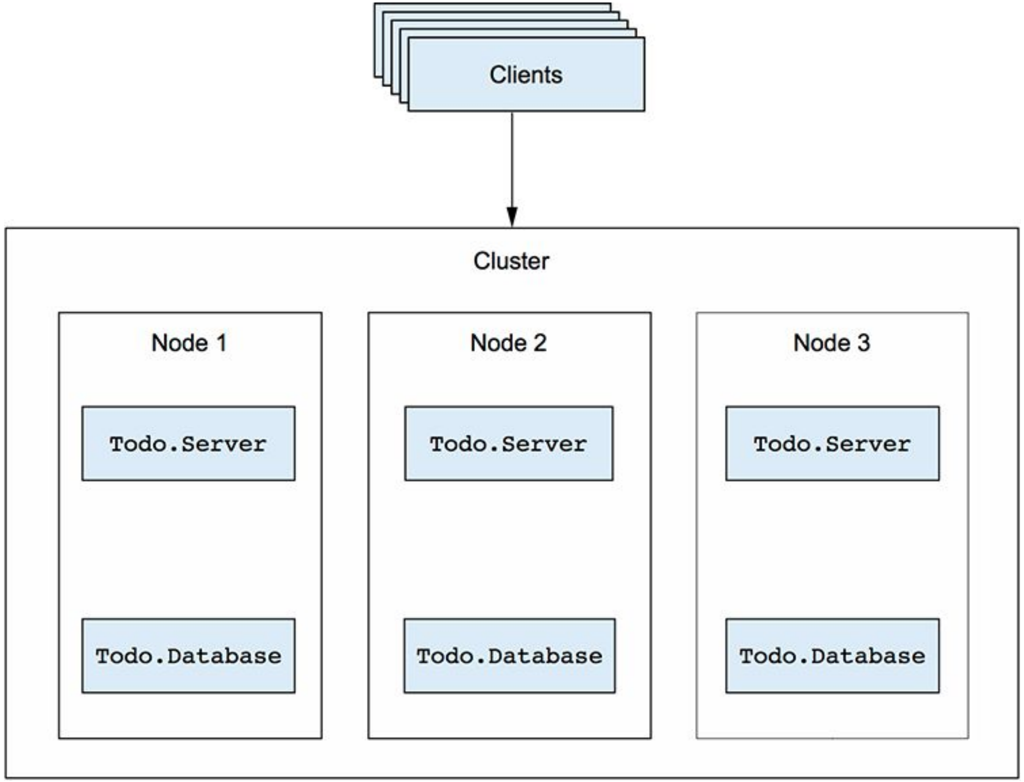


Image from Elixir In Action - Manning

Simple httpd(8) example

- Remember that binding to * is not always the answer
    - hint: rdomains exist

Preemptive questions

- Doesn’t anycast not work for TCP
  - It works fine, within certain bounds.
- What happens to a TCP session when routes change?
  - It resets.
- Why bother to do this?
  - I’m sorry I failed; See previous slide around latency and scale.
- How does this relate to security?
  - See denial of service notes in previous slide.

That’s it; Thanks.

Questions / Comments / Concerns?

Networking - rdomain(4)

- One routing domain per interface
  - Used to determine inbound rdomain
- One or more routing tables per routing domain
  - Forward path from machine

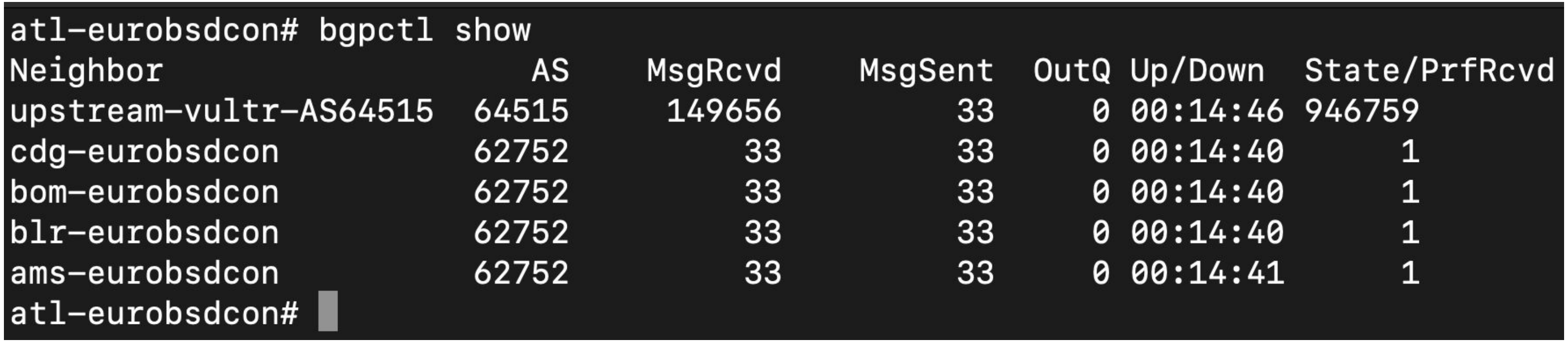
Use case in a small anycast environment:

- One rdomain for public external advertisement
  - relayd or similar services binding to some addresses
- One rdomain for VPN endpoints
  - Allows the same public v4 address to be used with pf.conf pass lines

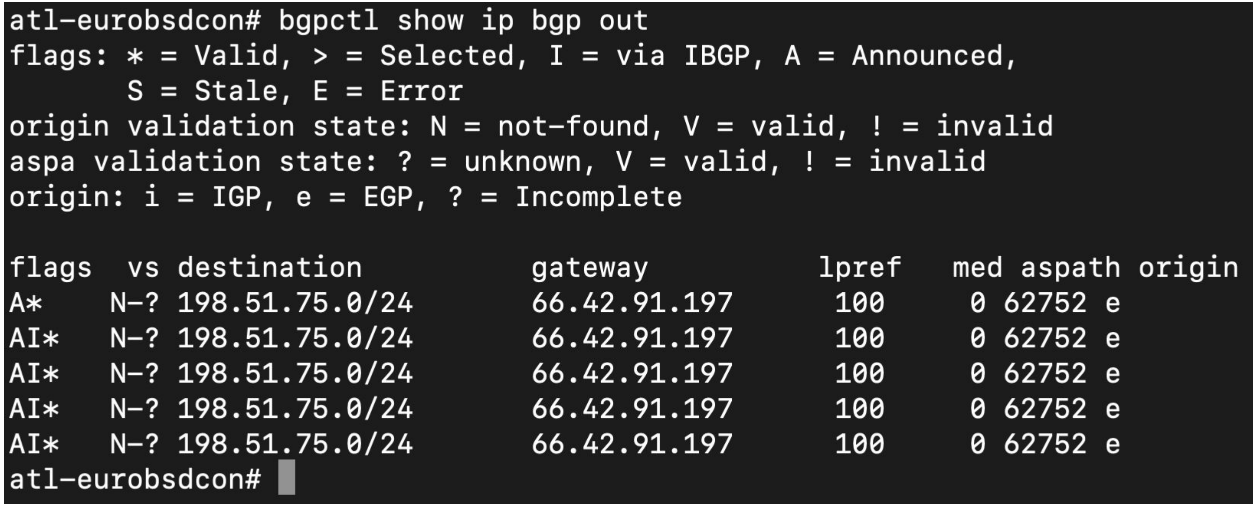
Networking - carp(4)

- Multiple hosts with a single IP
- Requires shared layer2 domain
- One inbound host is live at one time
- Can be setup with STONITH

Internal Networking - Full Mesh can get out of hand



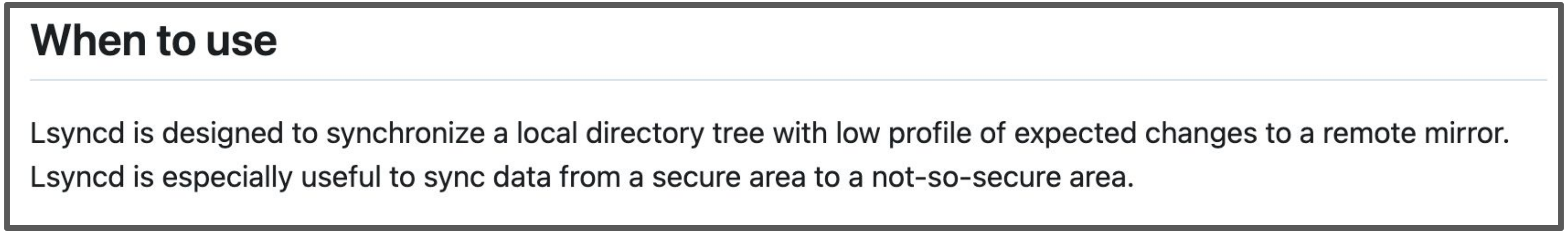
```
atl-eurobsdcon# bgpctl show
Neighbor                   AS    MsgRcvd    MsgSent  OutQ Up/Down  State/PrfRcvd
upstream-vultr-AS64515  64515     149656         33     0 00:14:46 946759
cdg-eurobsdcon          62752         33         33     0 00:14:40      1
bom-eurobsdcon          62752         33         33     0 00:14:40      1
blr-eurobsdcon          62752         33         33     0 00:14:40      1
ams-eurobsdcon          62752         33         33     0 00:14:41      1
atl-eurobsdcon# 
```



```
atl-eurobsdcon# bgpctl show ip bgp out
flags: * = Valid, > = Selected, I = via IBGP, A = Announced,
       S = Stale, E = Error
origin validation state: N = not-found, V = valid, ! = invalid
aspa validation state: ? = unknown, V = valid, ! = invalid
origin: i = IGP, e = EGP, ? = Incomplete

flags  vs destination        gateway          lpref   med aspath origin
A*    N-? 198.51.75.0/24      66.42.91.197       100     0 62752 e
AI*   N-? 198.51.75.0/24      66.42.91.197       100     0 62752 e
AI*   N-? 198.51.75.0/24      66.42.91.197       100     0 62752 e
AI*   N-? 198.51.75.0/24      66.42.91.197       100     0 62752 e
AI*   N-? 198.51.75.0/24      66.42.91.197       100     0 62752 e
atl-eurobsdcon# 
```

Storage - lsyncd

- Uses inotify under the hood
- Batches events on the filesystem up
- Executes one or more scripts that do things
  - Default is rsync
  - Ability to extend/modify fairly easily
- One direction - not bidirectional!



When to use

Lsyncd is designed to synchronize a local directory tree with low profile of expected changes to a remote mirror. Lsyncd is especially useful to sync data from a secure area to a not-so-secure area.

Storage - FUSE

- OpenBSD supports version 2.6
  - sshfs is supported
  - s3fs (and most fuse systems) aren’t supported
- Custom filesystem are doable
- cgo/rust/… bindings are available and function

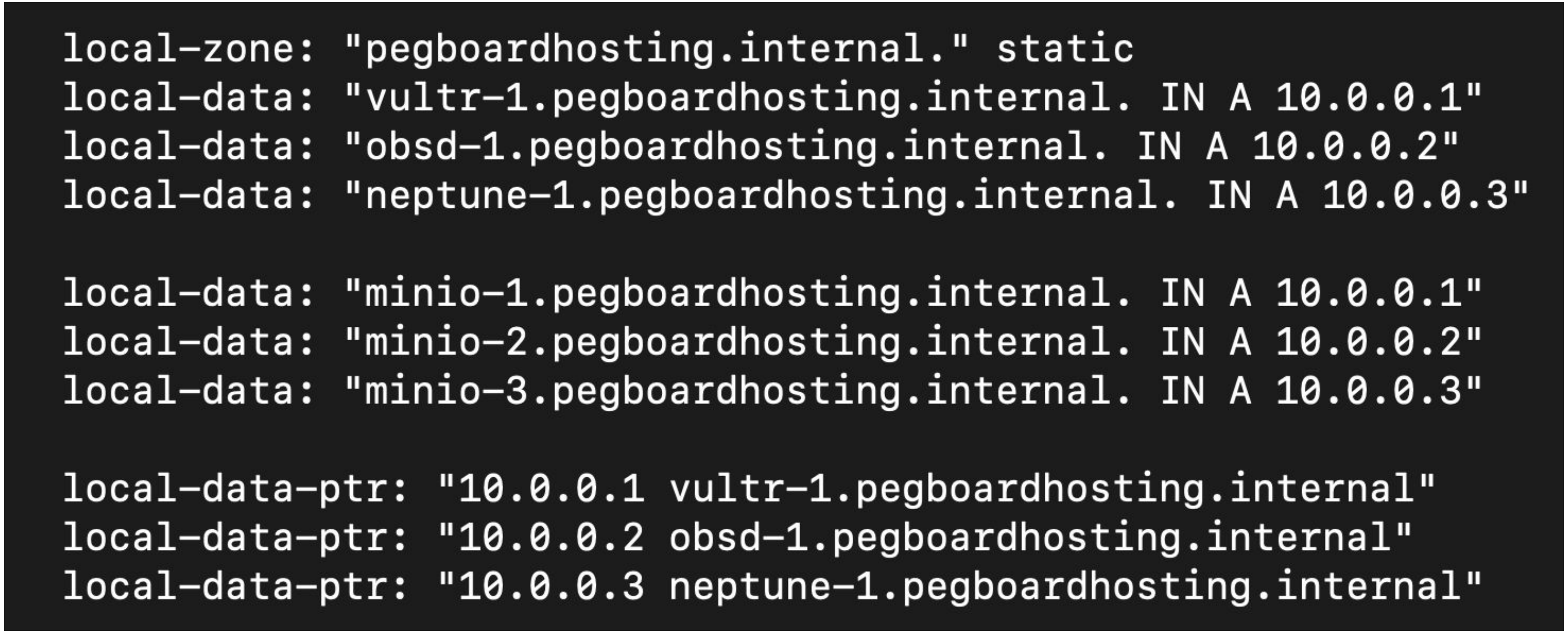
Opportunity for raft + FUSE passthrough

- Would be fairly FS / OS agnostic
- riverfs (>10 years abandoned)

Storage - minio

- S3 compatible API
- Written in Go
- Dual licensed (AGPL,Commercial)
- Has some not so nice requirements
    - Forces dns resolution in a particular scheme
    - Doesn’t handle adding/removing nodes/rebalancing
- Handles erasure coding, auto healing, hooks

Storage - minio



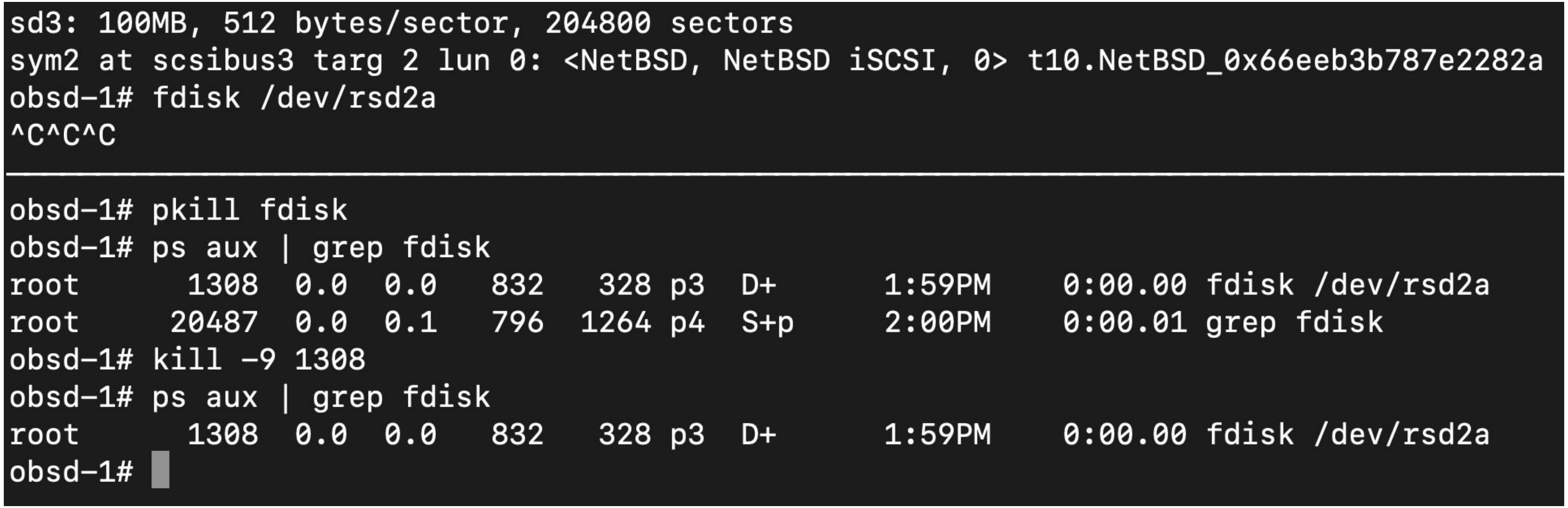
```
local-zone: "pegboardhosting.internal." static
local-data: "vultr-1.pegboardhosting.internal. IN A 10.0.0.1"
local-data: "obsd-1.pegboardhosting.internal. IN A 10.0.0.2"
local-data: "neptune-1.pegboardhosting.internal. IN A 10.0.0.3"

local-data: "minio-1.pegboardhosting.internal. IN A 10.0.0.1"
local-data: "minio-2.pegboardhosting.internal. IN A 10.0.0.2"
local-data: "minio-3.pegboardhosting.internal. IN A 10.0.0.3"

local-data-ptr: "10.0.0.1 vultr-1.pegboardhosting.internal"
local-data-ptr: "10.0.0.2 obsd-1.pegboardhosting.internal"
local-data-ptr: "10.0.0.3 neptune-1.pegboardhosting.internal"
```

snippet of unbound setup for minio

Storage - iscsi

- Block level remote storage
- Target (server) is available in packages/ports
    - pkg_add netbsd-iscsi-target
    - effectively scsi, over tcp, over ipsec
    - https://dataswamp.org/~solene/2019-02-21-iscsi-server.html
- Daemon (client) is built in
    - iscsi.conf(5) used to configure iscsid(8)
    - Results in /dev/vscsiX

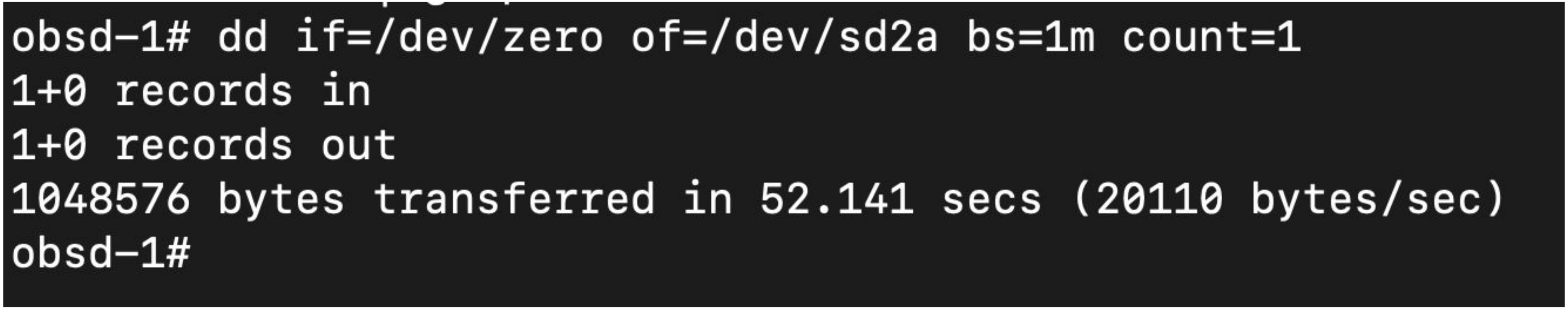
Storage - iscsi - here there be dragons



```
sd3: 100MB, 512 bytes/sector, 204800 sectors
sym2 at scsibus3 targ 2 lun 0: <NetBSD, NetBSD iSCSI, 0> t10.NetBSD_0x66eeb3b787e2282a
obsd-1# fdisk /dev/rsd2a
^C^C^C
```

```
obsd-1# pkill fdisk
obsd-1# ps aux | grep fdisk
root      1308  0.0  0.0   832   328 p3  D+      1:59PM    0:00.00 fdisk /dev/rsd2a
root     20487  0.0  0.1   796  1264 p4  S+p     2:00PM    0:00.01 grep fdisk
obsd-1# kill -9 1308
obsd-1# ps aux | grep fdisk
root      1308  0.0  0.0   832   328 p3  D+      1:59PM    0:00.00 fdisk /dev/rsd2a
obsd-1# 
```

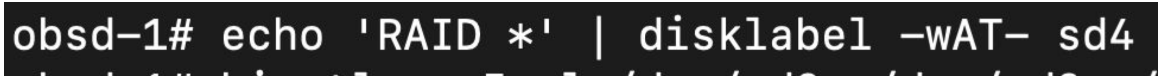
Storage - iscsi - performance



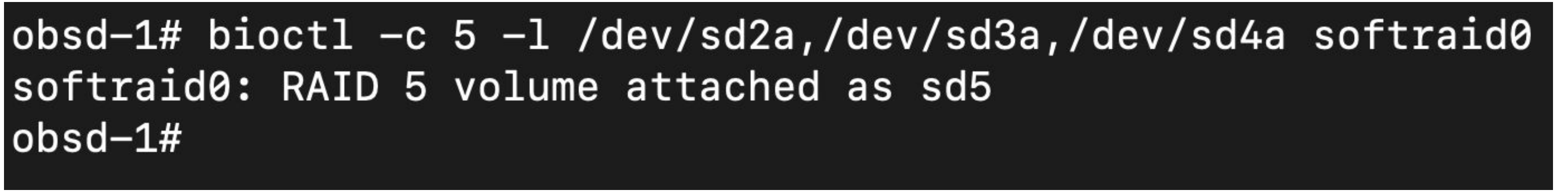
```
obsd-1# dd if=/dev/zero of=/dev/sd2a bs=1m count=1
1+0 records in
1+0 records out
1048576 bytes transferred in 52.141 secs (20110 bytes/sec)
obsd-1#
```

Storage - bioctl(8)

- software raid
- stripe / mirror / concat/ raid5



```
obsd-1# echo 'RAID *' | disklabel -wAT- sd4
```



```
obsd-1# bioctl -c 5 -l /dev/sd2a,/dev/sd3a,/dev/sd4a softraid0
softraid0: RAID 5 volume attached as sd5
obsd-1#
```

Storage - NFS

- Built in with nfsd(8)
- Provides single write path for filesystem
- Sane and off the shelf
- Works well with carp

External Networking - bgpd.conf(5)

- Advertise upstream
- Runs in providers IP space

```
peer1_as = “64512”
peer1_them = “203.0.113.1”
peer1_me = “203.0.113.2”
peer1_tcpmd5 = “foo”

prefix-set my_advertised {
        198.51.100.0/24
}

network prefix-set my_advertised set {
        nexthop $peer1_me, origin egp
}
```

```
group "upstreams" {
        neighbor $peer1_them {
                remote-as $peer1_as
                local-address $peer1_me
                descr "upstream-$peer1_name"
                tcp md5sig password $peer1_tcpmd5

                multihop 2
                announce IPv4 unicast
        }
}



allow to group upstreams prefix-set mypublic
```

External Networking - bgpipe

- BGP man-in-the-middle proxy that dumps all conversation
- bidirectional BGP to JSON bridge to a background process (filter or mirror mode)
- websocket + TLS transport of BGP sessions over the public Internet
- BGP listener on one side, connecting with a TCP-MD5 password on the other side
- BGP speaker that streams an MRT file after the session is established
- fast MRT to JSON converter (and back)
- IP prefix limits enforcer
- router control plane firewall (drop, modify, and synthesize BGP messages)
- https://github.com/bgpfix/bgpipe