faelix.link/vuknof3

SALT + NETBOX + VYOS = NETWORK AUTOMATION + ROUTING SECURITY

MAREK ISALSKI — FAELIX

faelix.link/vuknof3

VyOS SaltStack YAML Netbox BGP OSPF FRR RPKI IRR XDP bgpq3 UTRS RTBH NetFlow

MAREK ISALSKI — FAELIX

Who?

- Marek and Lou and Laura
- A bunch of @NetworkMoose
- Giants:
  - FRR, Routinator 3000, VyOS+contributors, Linux, netfilter/iptables, Salt+contributors, RIPE+RIRs, Team Cymru, DigitalOcean+Netbox-contributors...

The Story So Far...

- 2019-04: UKNOF43 MikroTik IPv6 routing vuln talk
  - Hinted about AS41495's future plans

https://faelix.link/uknof43

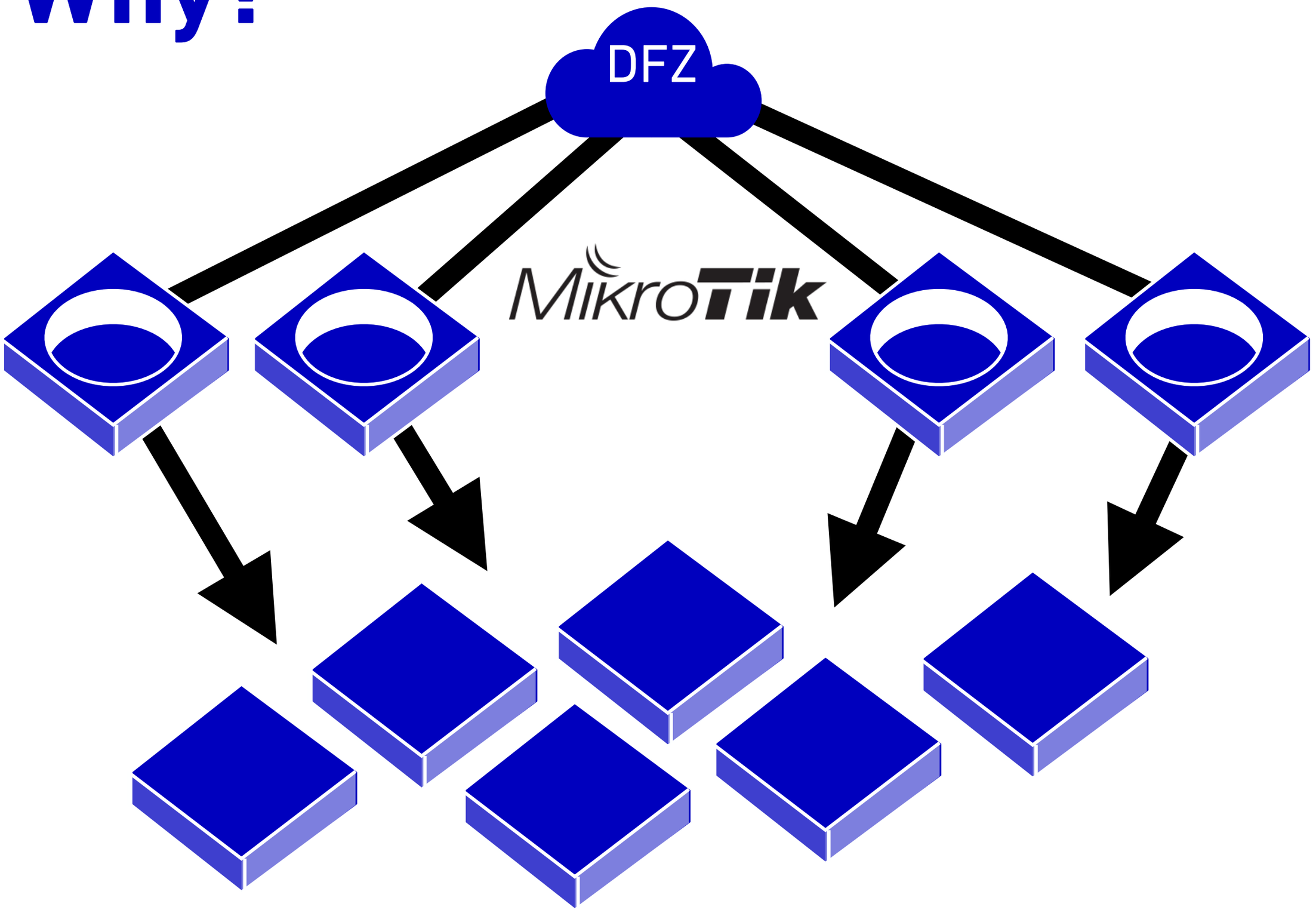
SCANNING IPv6 ADDRESS SPACE...

AND THE REMOTE VULNERABILITIES IT UNCOVERS

The Story So Far...

- 2019-04: UKNOF43 MikroTik IPv6 routing vuln talk
  - Hinted about AS41495's future plans
- 2019-10: NetMcr40 post-migration tech talk
  - Mentioned shortcomings and follow-on actions
- 2020-04: UKNOF46 talk on operational experience
  - ...postponed till next time?

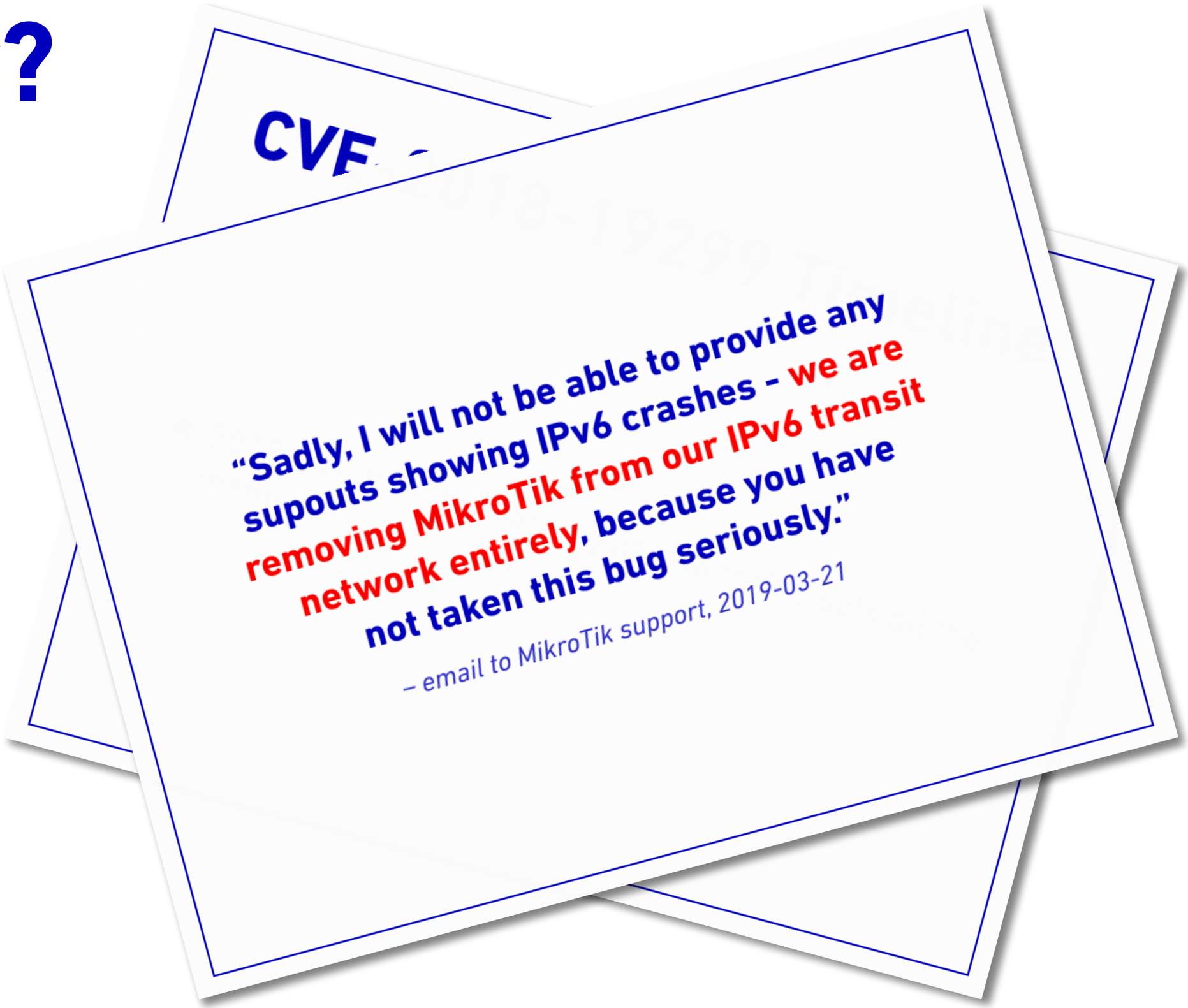
Why?



Why?


CVE-2018-19299 Timeline
2018-04-19 — "forwarding of ipv6 traffic eats all the memory"

Why?


CVE
"Sadly, I will not be able to provide any supouts showing IPv6 crashes - we are removing MikroTik from our IPv6 transit network entirely, because you have not taken this bug seriously."
– email to MikroTik support, 2019-03-21

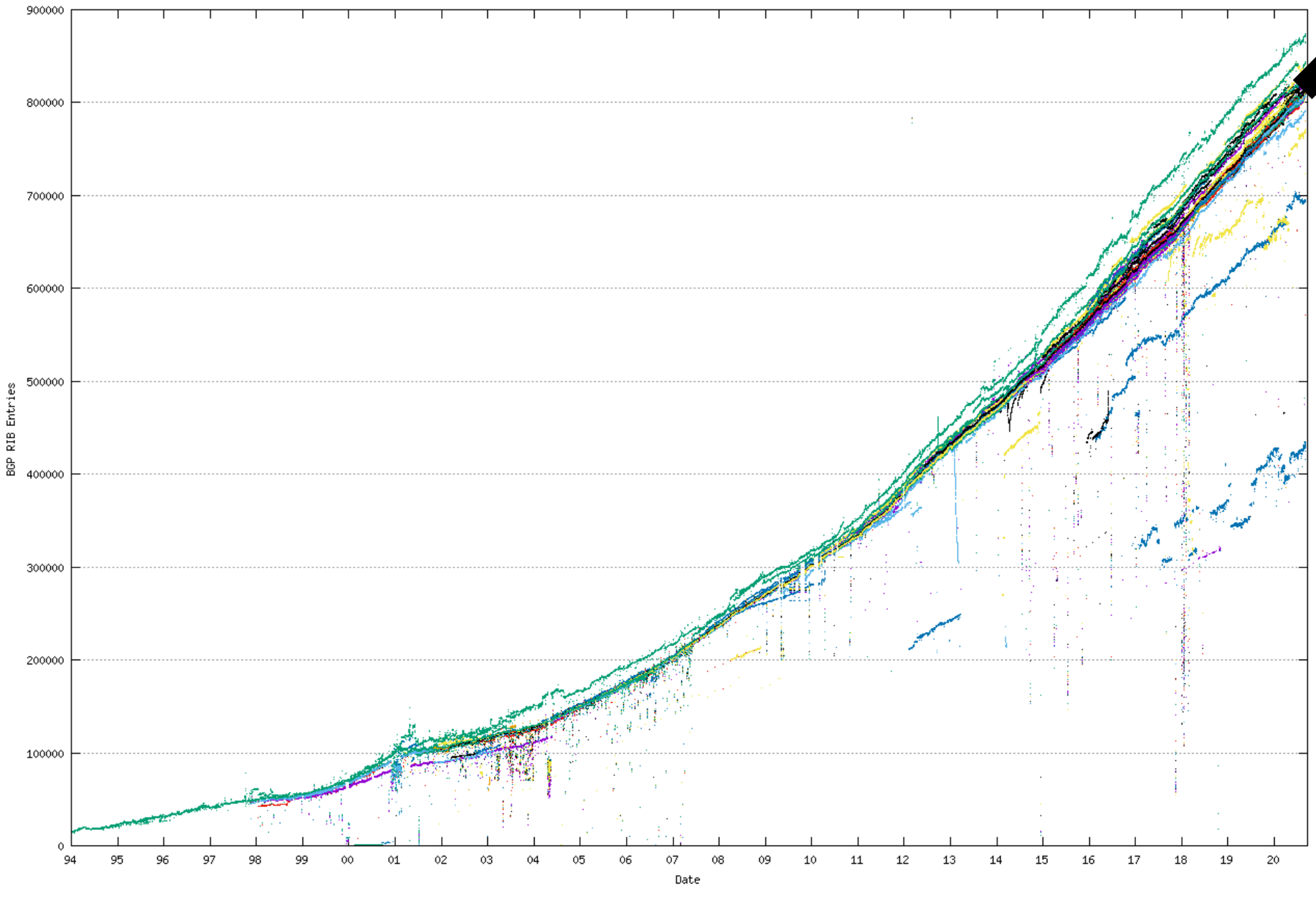
Why?

actually involved
MX routers, but cautionary tale
applies to Cisco 7600/6500 and
MikroTik CCR too!

Why?

- RIPE NCC Update 2019-10-02
- Today we allocated the last of our contiguous /22 IPv4 address blocks. We still have approximately **one million addresses available, in the form of /23s and /24s**, and we will continue making /22-equivalent allocations made up of these smaller blocks. Once we can no longer allocate the equivalent of a /22, we will announce that we have reached run-out. We expect this to occur in November 2019.

Why?

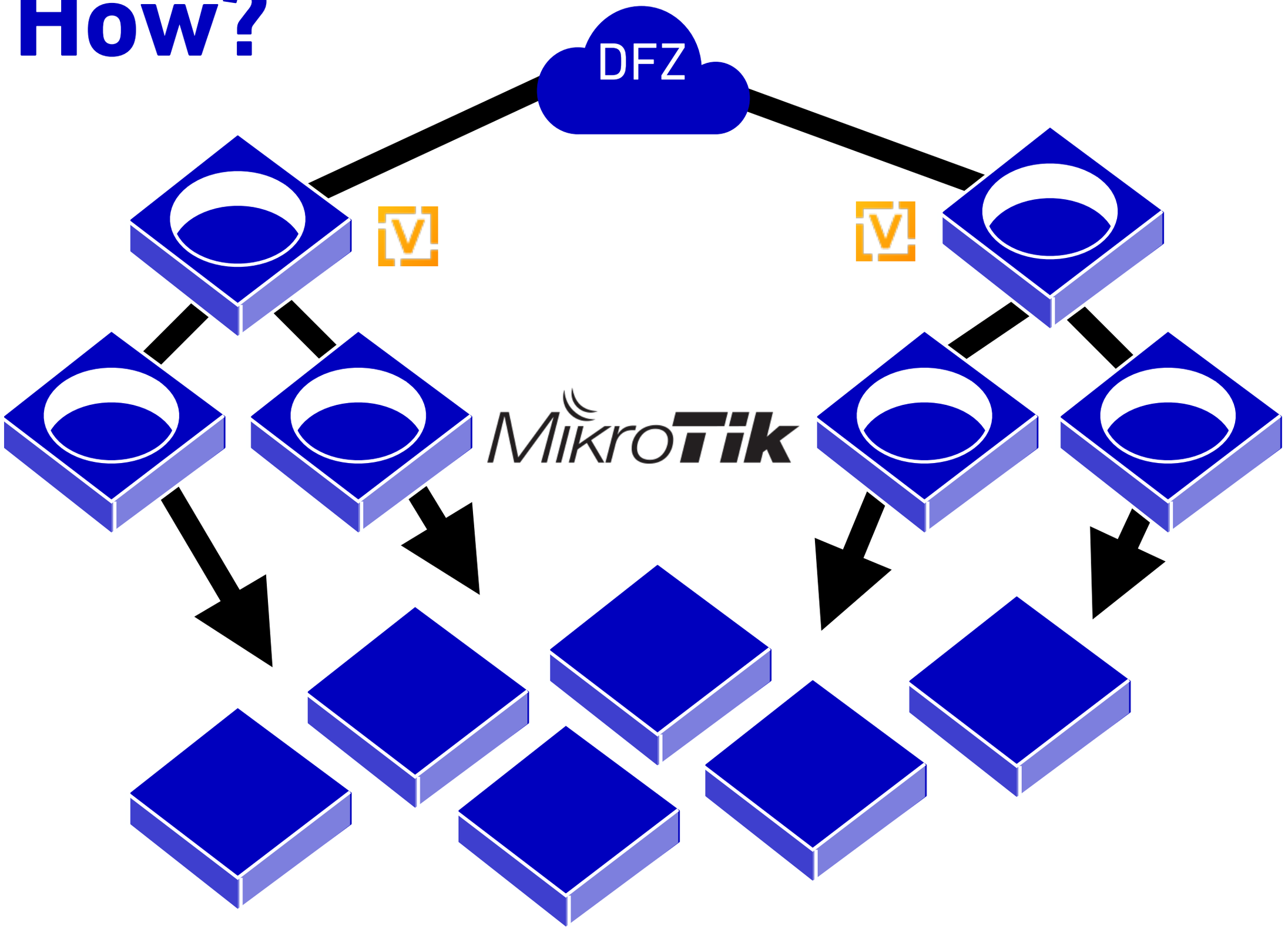


1M routes in 2021?

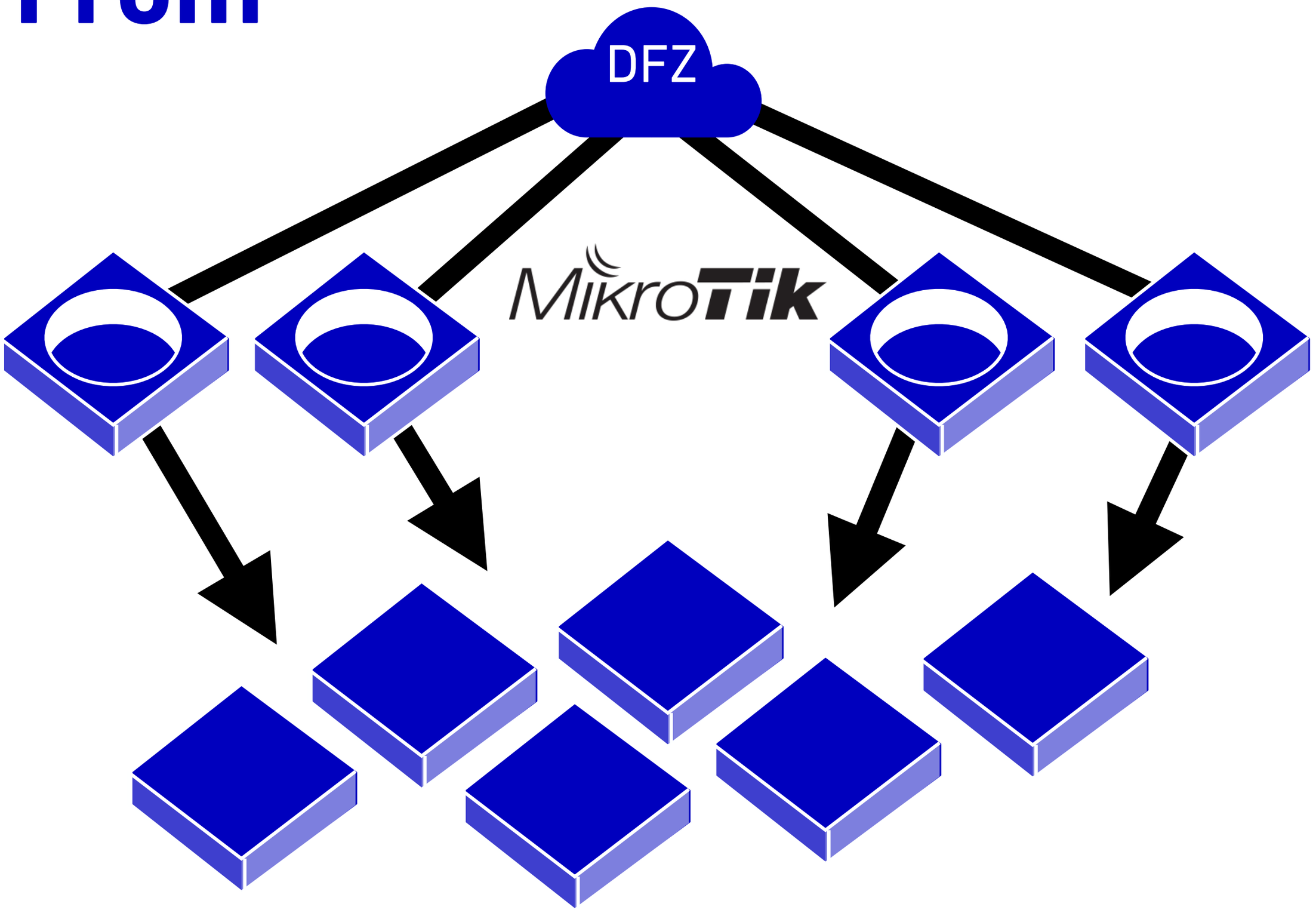
Where?

- This story all took place in:
  - Reynolds House (Equinix MA2)
  - Williams House (Equinix MA1)
- ...and then "laps of honour" in:
  - AQL DC2
  - Telehouse West and North
  - Interxion LON2

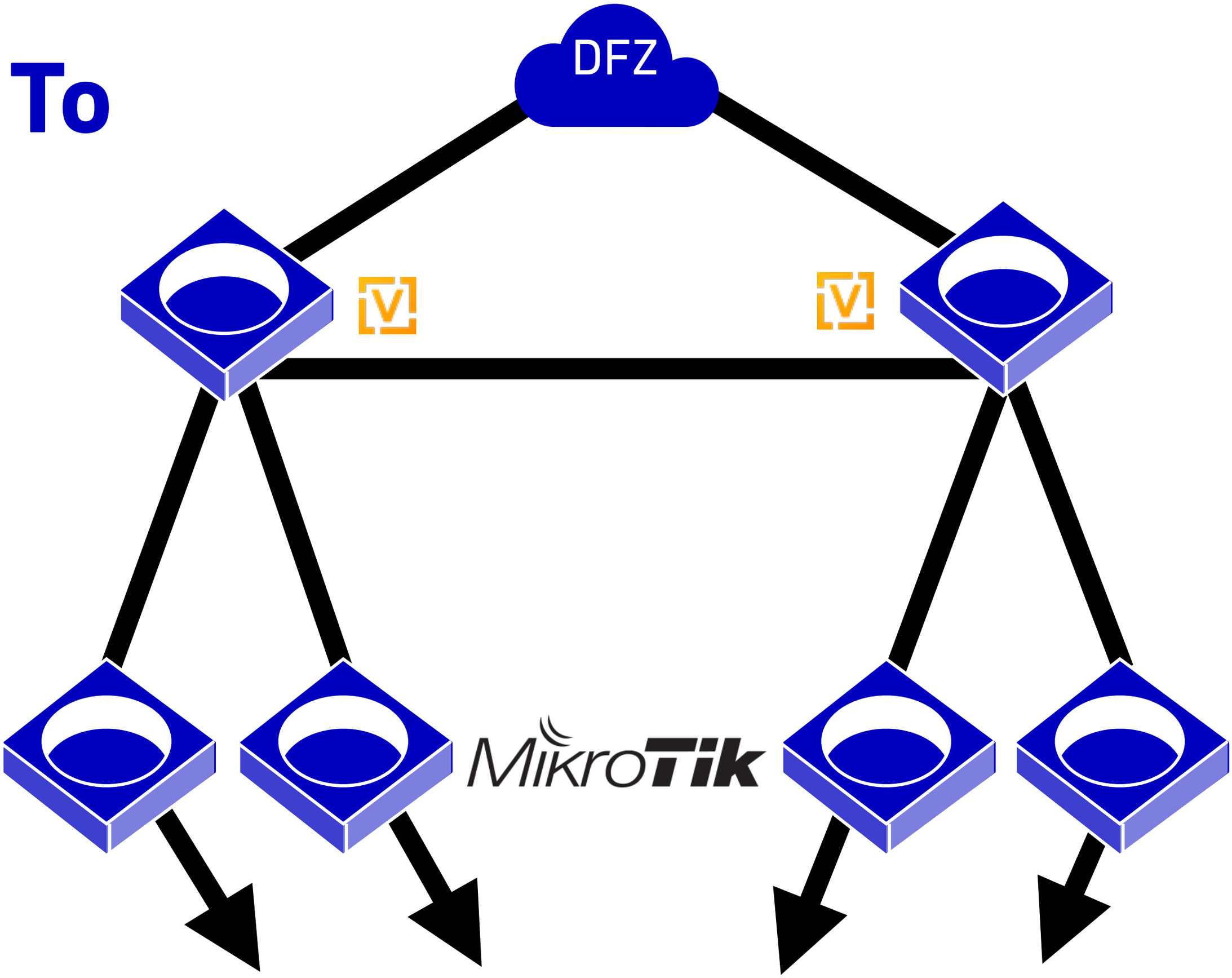
How?



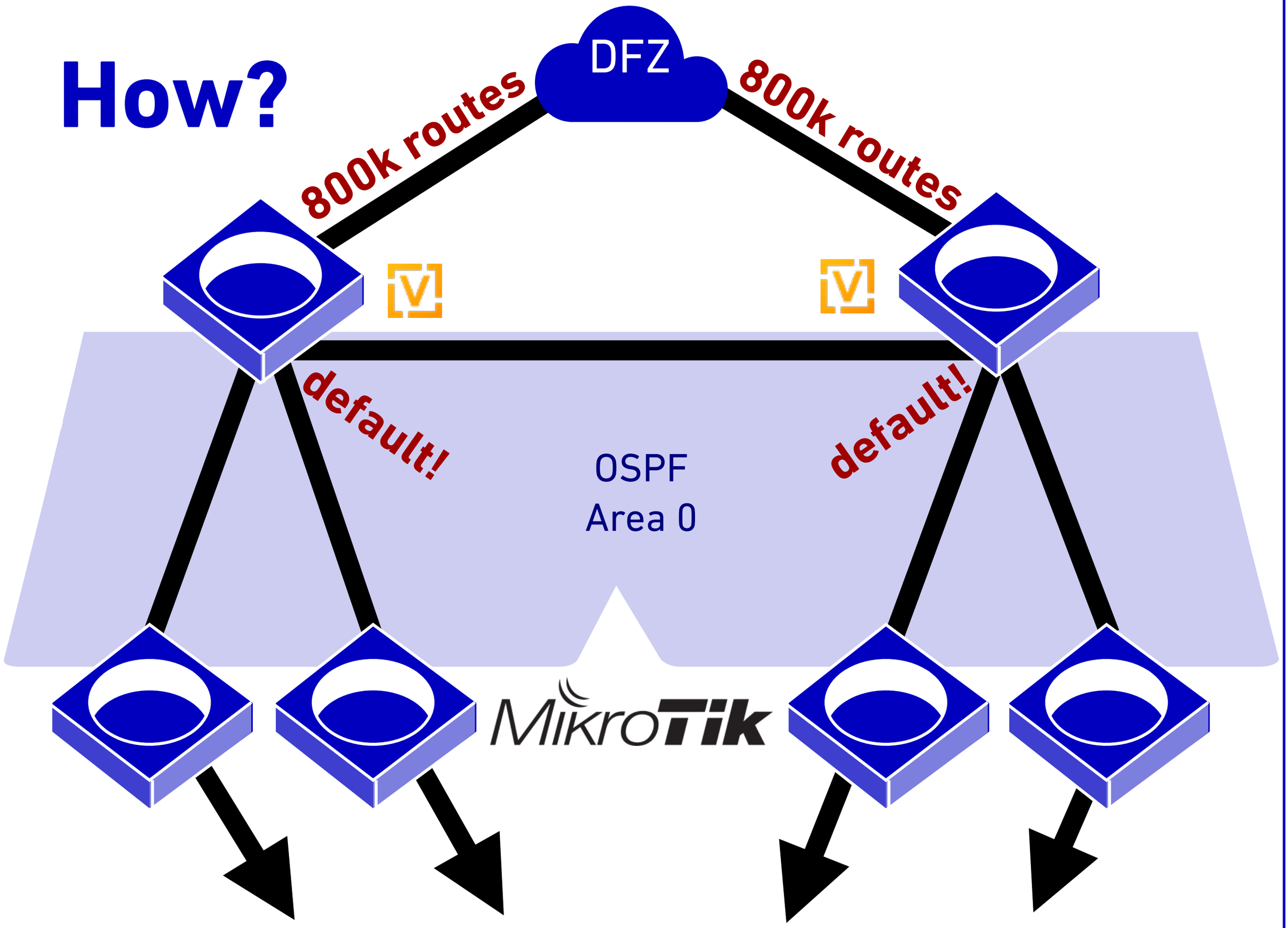
From



To



How?


DFZ
800k routes
800k routes
default!
default!
OSPF
Area 0
MikroTik

HARDWARE

Commodity

Network Peering Hardware

Brand	% of IXPs	# of Devices
Cisco Systems, Inc	21.7	3306
Juniper Networks	18.7	[illegible]
Routerboard.com	10.1	1541
unknown	8.6	1304
Super Micro Computer, Inc.	3.5	534
Intel Corporate	2.5	386
Arista Networks	2.5	386
VMware, Inc.	2.5	378
HUAWEI TECHNOLOGIES CO.,LTD	1.7	260
Brocade Communications Systems LLC	1.2	186

10.1%

Euro-IX IXPDB confirms it

How?

- 4x 10G SFP+ ports (must not be RJ45, cannot be QSFP+ with SFP+ breakout)
- At least 2x 1G RJ45 ports
- Low CPU power draw
- Prefer GHz over cores (single-thread performance)
- ~16 Gb RAM
- Ideally 1U
- Ideally dual PSU and IPMI

How?

- SM Chassis: Front I/O, Redundant 400W PSUs
- SM X11SCM-F (2x RJ45 Intel® I210-AT + OOB/IPMI)
- Xeon E3-1265Lv5 (4c/8t, 2.9GHz, 45W)
- 16GB DDR4 2666 ECC, 2x 240Gb SSDs
- Intel X710-DA4 (4x SFP+ NIC)

How?

Sentral

DESCRIPTION	QTY	RATE	AMOUNT
CSE-515-R407 SM Fixed Bay Chassis & Front I/O c/w Redundant 400w PSU's + Ra s (CSE-515-R407) Motherboard: SM X11SCM-F (onboard Dual LAN with Intel® Ethernet C ntroller I210-AT) CPU(s): Intel® Xeon E3-1265Lv5 (4/8*2.9Ghz/8MB/8GT's) - (Tray) RAM: 16GB DDR4 2666 ECC UDIMM (2*8GB) SSD: 2* 240GB Intel s4510 Expansion: RSC-RR1U-E16 Riser + Intel X710-DA4	2	1,480.00	2,960.00
Delivery Shipping & Handling	2	15.00	30.00

PAID

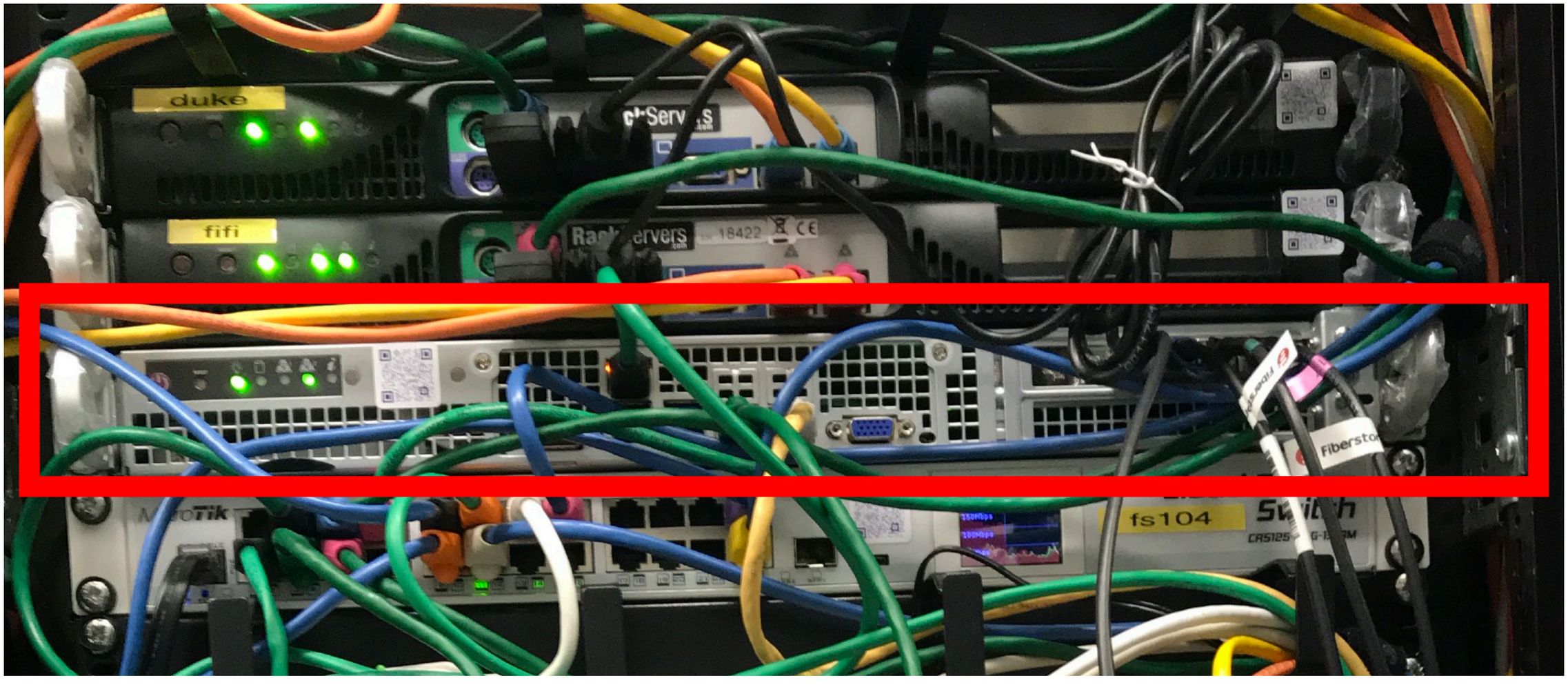
Bank Details:
IT Resolve Limited
Barclays Bank PLC
67a Above Bar, High Street
Southampton, SO14 3NZ
Account:
Sort Code:
IBAN:
SWIFTBIC:

SUBTOTAL	2,990.00
VAT TOTAL	598.00
TOTAL	3,588.00
PAYMENT	3,588.00
BALANCE DUE	**GBP 0.00**

VAT SUMMARY

RATE	VAT	NET
VAT @ 20%	598.00	2,990.00

How?



SOFTWARE

VyOS

- Linux
- FRR (Quagga fork) for routing protocols
- iproute2, iptables, ipset, etc

- VyOS 1.3: planned XDP (eXpress Data Path) support
  - Like DPDK / FDIO / VPP (but less tap/vif mess)
  - Up towards 20Mpps per CPU core...!

SaltStack

- Big investment in Salt @faelix
  - Use it for bare metal (virtualisation clusters)
  - Use it for our ISP services and NFVs
  - Use it for full-/part-managed customer VPSs
- ...why not use it for routers?
  - VyOS includes salt-minion (agent)

Introducing: hphr

- "Halophile Router" (salt-loving)
- Uses Netbox for single source-of-truth
- Uses SLS (YAML) for anything not possible in Netbox
- Generates a VyOS router configuration file
- Loads, compares, commits, saves to config.boot

- Released open-source: github.com/faelix/hphr

Addressing

Interfaces

	Name	LAG	Description
☐	⇄ eth0 AC:1F:6B:94:1F:12		—

IP Address	Status/Role
10.13.0.57/22	Active

	Name	LAG	Description
☐	⇄ eth1 AC:1F:6B:94:1F:13		LINX Manchester

IP Address	Status/Role
195.66.244.97/24	Active
2001:7f8:4:2::a217:1/64	Active

```
Last login: Mon Oct  7 15:11:31 2019 from 10.13.0.143
vyos@bly.w.faelix.net:~$ show interfaces
Codes: S - State, L - Link, u - Up, D - Down, A - Admin Down
Interface        IP Address                     S/L  Description
---------        ----------                     ---  -----------
eth0             10.13.0.57/22                  u/u  -
eth1             195.66.244.97/24               u/u  LINX Manchester
                 2001:7f8:4:2::a217:1/64
```

Connections

⇄ eth4 3C:FD:FE:D0:20:32	—	—	—	#625	cao.w.faelix.net	xe1-bly

IP Address	Status/Role	VRF	Descriptio
46.227.203.225/30	Active	Global	—
2a01:9e00:a217:2:2022:2031:1:1/126	Active	Global	—

⇄ eth5 3C:FD:FE:D0:20:33	—	—	—	#623	rae.w.faelix.net	xe0-bly

IP Address	Status/Role	VRF	Descriptio
46.227.203.229/30	Active	Global	—
2a01:9e00:a217:2:2022:2032:1:1/126	Active	Global	—

Cable

Type	Direct Attach Copper (Passive)
Status	Connected
Label	—
Color	
Length	2 Meters

Termination A

Device	bly.w.faelix.net
Type	Interface
Component	eth4

Termination B

Device	cao.w.faelix.net
Type	Interface
Component	xe1-bly

```
eth4            46.227.203.225/30                    u/u  - (xe1-bly @ cao.w.faelix.net)
                2a01:9e00:a217:2:2022:2031:1:1/126
eth5            46.227.203.229/30                    u/u  - (xe0-bly @ rae.w.faelix.net)
                2a01:9e00:a217:2:2022:2032:1:1/126
```

Sub-Interfaces

802.1Q Mode: Tagged

Access: One untagged VLAN
Tagged: One untagged VLAN and/or one or more tagged VLANs
Tagged All: Implies all VLANs are available (w/optional untagged VLAN)

Untagged vlan: 1

Tagged vlans: × 337 (AS3257 GTT Transit)

☐	⇄ eth3 3C:FD:FE:D0:20:31	—	—	Tagged
☐	● eth3.337	GTT AS3257 TRANSIT	—	—

IP Address	Status/Role	VRF
77.67.124.102/30	Active	Global
2001:668:0:3:ffff:1:0:4fe/126	Active	Global

```
eth3              -                                   u/u  - (xe12 @ fs116.w.faelix.net)
eth3.337          77.67.124.102/30                    u/u  AS3257 GTT Transit => GTT AS3257 TRANSIT
                  2001:668:0:3:ffff:1:0:4fe/126
```

OSPF, BGP, etc?

jeremystretch commented on 19 Apr 2017 Contributor

Marking this as a duplicate of #127. BGP communities are likely going to be out of scope for NetBox.

jeremystretch closed this on 19 Apr 2017

jeremystretch commented on 20 Jul 2016 Contributor

Circuits are intended only to represent physical connections. I definitely wouldn't try to use them for VPN tunnels.

A VPN tunnel is a pretty abstract concept, and probably delves too far into the realm of configuration management to be applicable to NetBox.

jeremystretch closed this on 20 Jul 2016

Netbox ≈ "physical infra"

- Using SaltStack's "pillar" in two ways:
  - Netbox for DCIM / IPAM
  - SLS (YAML-syntax) files for OSPF, BGP, etc

Peering/Transit Interface

```
interfaces:
    eth3.337:
        bcp38:
            source:
                bgpq3:
                    IPv4: AS-FAELIX
                    IPv6: AS-FAELIX
        ip:
            ospf:
                passive: True
        ipv6:
            ospfv3:
                passive: True
        netflow: True
```

P/T Interface: BCP38

```
*filter
:INPUT ACCEPT [23923954:3646493851]
:FORWARD ACCEPT [7683969829:4464797908411]
:OUTPUT ACCEPT [45590736:10129789709]
-A INPUT -p tcp -m tcp --dport 179 -m set --match-set control-plane-bgp-v4 src -j ACCEPT
-A INPUT -p tcp -m tcp --dport 179 -m tcp -j REJECT --reject-with tcp-reset
-A FORWARD -m set --match-set bcp38-cone-oface-v4 src,dst -j ACCEPT
-A FORWARD -m set --match-set bcp38-else-oface-v4 src,dst -j DROP
COMMIT
```

```
bgpq3:
    IPv4: AS-FAELIX
    IPv6: AS-FAELIX
```

bgpq3

```
as-set:         AS-FAELIX
descr:          Faelix Limited AS Set
members:        AS41495
members:        AS-EMIZON
members:        AS-AVR
members:        AS-IQIP
members:        AS-NAMESERVERPARTY
members:        AS-CSLDUAL
members:        AS-METROSECURITY
members:        AS6647
org:            ORG-FLX1-RIPE
```

```
46.227.200.0/21,eth3.337
46.227.200.0/21,eth2.391
46.227.200.0/21,eth2.1302
46.227.200.0/21,eth2.1303
46.227.200.0/21,eth1
139.28.100.0/22,eth3.337
139.28.100.0/22,eth2.391
139.28.100.0/22,eth2.1302
139.28.100.0/22,eth2.1303
139.28.100.0/22,eth1
```

BCP38 Implementation

- Throws away entire VyOS firewall stack
  - Native commands are off-limits ⇒ fragility
- Currently using iptables
  - Will transition to nft ⇒ it's the future, offload, etc
- Eventually move this into XDP
  - Along with routing decision ⇒ more PPS

P/T Interface: NetFlow

```
-A PREROUTING -j NOTRACK
-A PREROUTING -i eth3.337 -j NFLOG --nflog-group 2 --nflog-range 64 --nflog-threshold 10
```

```
netflow:
    sampling-rate: 16
    servers:
        - "10.13.1.246:2055"
```

BGP Peering

```
"195.66.244.22":
    remote-as: 786
    description: AS786 (Janet) @ LINX Manchester
    address-family:
        ipv4-unicast:
            maximum-prefix: 320
            prefix-list:
                export: auto-AS-FAELIX
                import: auto-AS-JANETPLUS
            route-map:
                import: LINXMAN-  -IPv4
                export: TRANSIT-  t-IPv4
            soft-reconfiguration
                - inbound
"2001:7f8:4:2::312:1":
    remote-as: 786
    description: AS786 (Janet) @  INX Manchester
    address-family:
```

bgpq3

BGP Peering

```
"195.66.244.22":
    remote-as: 786
    description: AS786 (Janet) @ LINX Manchester
    address-family:
        ipv4-unicast:
            maximum-prefix: 320
            prefix-list:
                export: auto-AS-EAFLIX
                import: auto-AS-JANETPLUS
            route-map:
                import: LINXMAN-in-IPv4
                export: TRANSIT-out-IPv4
            soft-reconfiguration:
                - inbound
"2001:7f8:4:2::312:1":
    remote-as: 786
    description: AS786 (Janet) @ LINX Manchester
    address-family:
```

"aggregate"

bgpq3 -A AS-LLNW | wc

18425 106080 792306

wc /config/config.boot

558568

Long Bootup Times!

BGP Peering

```
"195.66.244.22":
    remote-as: 786
    description: AS786 (Janet) @ LINX Manchester
    address-family:
        ipv4-unicast:
            maximum-prefix: 320
            prefix-list:
                export: auto-AS-FAELIX
                import: auto-AS-JANETPLUS
            route-map:
                import: LINXMAN-in-IPv4
                export: TRANSIT-out-IPv4
            soft-reconfiguration:
                - inbound
"2001:7f8:4:2::312:1":
    remote-as: 786
    description: AS786 (Janet) @ LINX Manchester
    address-family:
```

RPKI via RTRR

```
LINXMAN-in:
    IPv4:
        - match-rpki: invalid
          action: deny
        - match-prefix-list: nphr-DFZ-IPv4
          set-community: 41495:64701
          continue: next
        - match-prefix-list: auto-AS41495-CH
          set-local-preference: 700
          set-community: 41495:5459
        - match-prefix-list: nphr-DFZ-IPv4
          set-local-preference: 600
          set-community: 41495:5459
    IPv6:
        - match-rpki: invalid
          action: deny
```

```
TRANSIT-out:
    IPv4:
        - match-community: AS41495-upstream
          action: deny
        - match-community: AS41495-peer
          action: deny
        - match-community: AS41495-customer
        - match-community: AS41495-internal
        - match-prefix-list: auto-AS-FAELIX
    IPv6:
        - match-community: AS41495-upstream
```

BGP via PeeringDB (v2)

```
ix:
    IXLeeds:
        peeringdb_ixlan: 435

        address-family:
            ipv4-unicast:
                route-map:
                    import: IXLEEDS-in-IPv4
            ipv6-unicast:
                route-map:
                    import: IXLEEDS-in-IPv6
        asn:
            33920: {}
            20940: {}
            41230: {}
            57099: {}
            30740: {}
            56595: {}
            786: {}
            3856: {}
            42: {}
            25178: {}
            15692: {}
            26415: {}
            20738: {}
            43013: {}
            36040: {}
            6939:
                default:
                    address-family:
                        ipv4-unicast:
                            prefix-list:
                                import: hphr-DFZ-IPv4
                        ipv6-unicast:
                            prefix-list:
                                import: hphr-DFZ-IPv6
                            route-map:
                                import: AS6939-IXLEEDS-in-IPv6
```

peeringdb API via salt

defaults

peering by ASN

copy-paste is dead

overrides

BGP Upstream

```
"80.68.83.0":
    remote-as: 35425
    description: Bytemark AS35425 transit
    update-source: 80.68.83.1
    address-family:
        ipv4-unicast:
            allowas-in: 1
            prefix-list:
                export: auto-AS-FAELIX
                import: hphr-DFZ-IPv4
            route-map:
                import: AS35425-in-IPv4
                export: TRANSIT-out-IPv4
            soft-reconfiguration:
                - inbound
"2001:41c8:2000:3::1":
    remote-as: 35425
    description: Bytemark AS35425 transit
    update-source: 2001:668:0:3:ffff:1:0:4fe
    address-family:
        ipv6-unicast:
```

RPKI via RTRR

```
AS35425-in:
    IPv4:
        - match-rpki: invalid
          action: deny
        - match-prefix-list: hphr-DFZ-IPv4
          set-community: 41495:64700
          continue: next
        - match-prefix-list: auto-AS41495-CH
          set-local-preference: 200
          set-community: 41495:35425
        - match-prefix-list: hphr-DFZ-IPv4
          set-local-preference: 200
          set-community: 41495:35425
    IPv6:
        - match-rpki: invalid
          action: deny
        - match-prefix-list: hphr-DFZ-IPv6
```

RPKI

Routinator 3000

```
protocols:
    rpki:
        cache:
            rey_man_uk_lg_faelix_net:
                address: 46.227.200.12
                port: 3323
            wil_man_uk_lg_faelix_net:
                address: 46.227.203.12
                port: 3323
            sat_gen_ch_lg_faelix_net:
                address: 185.134.196.14
                port: 3323
```

GOTCHAS

BGP C/P Protection

```
Sep 29 02:48:03 bly bgpd[1234]: [EC 33554454] 195.66.244.231 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:50:03 bly bgpd[1234]: [EC 33554454] 2001:7f8:4:2::220a:2 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:50:03 bly bgpd[1234]: [EC 33554454] 195.66.244.231 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:52:03 bly bgpd[1234]: [EC 33554454] 2001:7f8:4:2::220a:2 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:52:03 bly bgpd[1234]: [EC 33554454] 195.66.244.231 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:53:00 bly bgpd[1234]: [EC 100663313] SLOW THREAD: task bgpd_sync_callback (7fdb4984ed60) ran for 50172ms (cpu time 48285ms)
Sep 29 02:53:42 bly watchfrr[1167]: [EC 268435457] bgpd state -> unresponsive : no response yet to ping sent 90 seconds ago
Sep 29 02:53:42 bly watchfrr[1167]: [EC 100663303] Forked background command [pid 1284]: /usr/lib/frr/watchfrr.sh restart bgpd
Sep 29 02:53:48 bly bgpd[1234]: [EC 100663313] SLOW THREAD: task bgpd_sync_callback (7fdb4984ed60) ran for 48218ms (cpu time 46469ms)
Sep 29 02:53:48 bly bgpd[1234]: Terminating on signal
Sep 29 02:54:02 bly zebra[1230]: [EC 4043309117] Client 'vnc' encountered an error and is shutting down.
Sep 29 02:54:02 bly zebra[1230]: [EC 4043309117] Client 'bgp' encountered an error and is shutting down.
Sep 29 02:54:02 bly watchfrr[1167]: [EC 268435457] bgpd state -> down : unexpected read error: Connection reset by peer
Sep 29 02:54:02 bly zebra[1230]: client 30 disconnected. 0 vnc routes removed from the rib
Sep 29 02:54:02 bly zebra[1230]: zebra/zebra_ptm.c:1345 failed to find process pid registration
Sep 29 02:54:02 bly watchfrr[1167]: bgpd state -> up : connect succeeded
Sep 29 02:54:02 bly zebra[1230]: client 20 disconnected. 847117 bgp routes removed from the rib
Sep 29 02:54:02 bly zebra[1230]: client 20 says hello and bids fair to announce only bgp routes vrf=0
Sep 29 02:54:02 bly zebra[1230]: client 32 says hello and bids fair to announce only vnc routes vrf=0
Sep 29 03:15:30 bly ospfd[1249]: ASBR[Status:0]: Update
Sep 29 03:20:37 bly ospfd[1249]: ASBR[Status:1]: Update
```

bgpd state -> unresponsive : no response yet to ping sent 90 seconds ago

BGP C/P Protection

bgpd restarting, we're gonna be ok?!

```
Sep 29 02:48:03 bly bgpd[1234]: [EC 33554454] 195.66.244.231 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:50:03 bly bgpd[1234]: [EC 33554454] 2001:7f8:4:2::220a:2 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:50:03 bly bgpd[1234]: [EC 33554454] 195.66.244.231 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:52:03 bly bgpd[1234]: [EC 33554454] 2001:7f8:4:2::220a:2 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:52:03 bly bgpd[1234]: [EC 33554454] 195.66.244.231 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:53:00 bly bgpd[1234]: [EC 100663313] SLOW THREAD: task bgpd_sync_callback (7fdb4984ed60) ran for 50172ms (cpu time 48285ms)
Sep 29 02:53:42 bly watchfrr[1167]: [EC 268435457] bgpd state -> unresponsive : no response yet to ping sent 90 seconds ago
Sep 29 02:53:42 bly watchfrr[1167]: [EC 100663303] Forked background command [pid 1284]: /usr/lib/frr/watchfrr.sh restart bgpd
Sep 29 02:53:48 bly bgpd[1234]: [EC 100663313] SLOW THREAD: task bgpd_sync_callback (7fdb4984ed60) ran for 48218ms (cpu time 46469ms)
Sep 29 02:53:48 bly bgpd[1234]: Terminating on signal
Sep 29 02:54:02 bly zebra[1230]: [EC 4043309117] Client 'vnc' encountered an error and is shutting down.
Sep 29 02:54:02 bly zebra[1230]: [EC 4043309117] Client 'bgp' encountered an error and is shutting down.
Sep 29 02:54:02 bly watchfrr[1167]: [EC 268435457] bgpd state -> down : unexpected read error: Connection reset by peer
Sep 29 02:54:02 bly zebra[1230]: client 30 disconnected. 0 vnc routes removed from the rib
Sep 29 02:54:02 bly zebra[1230]: zebra/zebra_ptm.c:1345 failed to find process pid registration
Sep 29 02:54:02 bly watchfrr[1167]: bgpd state -> up : connect succeeded
Sep 29 02:54:02 bly zebra[1230]: client 20 disconnected. 847117 bgp routes removed from the rib
Sep 29 02:54:02 bly zebra[1230]: client 20 says hello and bids fair to announce only bgp routes vrf=0
Sep 29 02:54:02 bly zebra[1230]: client 32 says hello and bids fair to announce only vnc routes vrf=0
Sep 29 03:15:30 bly ospfd[1249]: ASBR[Status:0]: Update
Sep 29 03:20:37 bly ospfd[1249]: ASBR[Status:1]: Update
```

847177 bgp routes removed from the rib

[NARRATOR VOICE]
THEY WERE NOT "OK"

```
bly$ show ip bgp summary
% BGP instance not found
bly$
```

Architecture of VyOS

/config/config.boot

So: Reboot the Router

- https://phabricator.vyos.net/T1514
  - FRR's running-config not "refreshable" if a daemon crashes, can only reboot router to make VyOS send new running-config
  - or: delete protocols bgp 41495, commit, load /config/config.boot, commit
- Rebooting a router is annoying, but we're in a maintenance window, the other router is ok, so we're gonna be ok.

[NARRATOR VOICE]
THEY WERE NOT "OK"

sorry, no slides online for this one!

THE WORST MAINTENANCE OF MY LIFE (SO FAR)

MAREK ISALSKI — FAELIX

sorry, no slides online for this one!

ANNOUNCING 0.0.0.0/0 TO LON1 (BY ACCIDENT)

MAREK ISALSKI — FAELIX

FRR Converges BGP Fast

- Router booted up
- Began configuring itself
  - Brought interfaces up
  - Brought BGP instance up
- Cogent transit session established
- Peering sessions on LINX established
- ...then VyOS applied prefix-lists and route-maps :-(

FRR Converges BGP Fast

- Rou
- Beg
- Peering
- ...then VyOS applied p aps :-(

ANNOUNCING 800K ROUTES TO LINX MAN (BY ACCIDENT)

BUGS!

ANNOUNCING 800K ROUTES

TO LINX MANUBY A

~~T148~~ + ~~T944~~

~~T1698~~ + ~~T1144~~

Achievement Unlocked :(

FRR Impedance Mismatch

- router bgp 41495
- neighbor 195.66.244.6 remote-as 6939
- neighbor 195.66.244.6 soft-reconfiguration inbound
- neighbor 195.66.244.6 maximum-prefix 200000
- neighbor 195.66.244.6 prefix-list hphr-DFZ-IPv4 in
- neighbor 195.66.244.6 prefix-list auto-AS-FAELIX out
- neighbor 195.66.244.6 route-map AS6939-LINX-in-IPv4 in
- neighbor 195.66.244.6 route-map TRANSIT-out-IPv4 out

FRR Impedance Mismatch

- router bgp 41495
  - neighbor 195.66.244.6 remote-as 6939
  - neighbor 195.66.244.6 soft-reconfiguration inbound
  - neighbor 195.66.244.6 maximum-prefix 200000
  - neighbor 195.66.244.6 prefix-list hphr-DFZ-IPv4 in
  - neighbor 195.66.244.6 prefix-list auto-AS-FAELIX out
  - neighbor 195.66.244.6 route-map AS6939-LINX-in-IPv4 in
  - neighbor 195.66.244.6 route-map TRANSIT-out-IPv4 out

neighbor shutdown

no shutdown

FRR Impedance Mismatch

- **"Since FRR lacks a command for creating peers in down state, no other workaround is possible."** Bug: T1148

- router bgp 41495
- neighbor 195.66.244.6 remote-as 6939
- neighbor 195.66.244.6 soft-reconfiguration inbound
- neighbor 195.66.244.6 maximum-prefix 200000
- neighbor 195.66.244.6 prefix-list hphr-DFZ-IPv4 in
- neighbor 195.66.244.6 prefix-list auto-AS-FAELIX out
- neighbor 195.66.244.6 route-map AS6939-LINX-in-IPv4 in
- neighbor 195.66.244.6 route-map TRANSIT-out-IPv4 out

FRR Impedance Mismatch

- **"Since FRR lacks a con… …g peers in down state, no other workaround is possible."** Bug: FR…
- router bgp 41495
  - no bgp default ipv4-unicast
  - neighbor 195.66.244.6 remote-as 6939
  - neighbor 195.66.244.6 soft-reconfiguration inbound
  - neighbor 195.66.244.6 maximum-prefix 200000
  - address-family ipv4
    - neighbor 195.66.244.6 prefix-list hphr-DFZ-IPv4 in
    - neighbor 195.66.244.6 prefix-list auto-AS-FAELIX out

Well, actually…

BGP C/P Protection

```
Sep 29 02:48:03 bly bgpd[1234]: [EC 33554454] 195.66.244.231 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:50:03 bly bgpd[1234]: [EC 33554454] 2001:7f8:4:2::220a:2 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:50:03 bly bgpd[1234]: [EC 33554454] 195.66.244.231 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:52:03 bly bgpd[1234]: [EC 33554454] 2001:7f8:4:2::220a:2 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:52:03 bly bgpd[1234]: [EC 33554454] 195.66.244.231 [Error] bgp_read_packet error: Connection reset by peer
Sep 29 02:53:00 bly bgpd[1234]: [EC 100663313] SLOW THREAD: task bgpd_sync_callback (7fdb4984ed60) ran for 50172ms (cpu time 48285ms)
Sep 29 02:53:42 bly watchfrr[1167]: [EC 268435457] bgpd state -> unresponsive : no response yet to ping sent 90 seconds ago
Sep 29 02:53:42 bly watchfrr[1167]: [EC 100663303] Forked background command [pid 1284]: /usr/lib/frr/watchfrr.sh restart bgpd
Sep 29 02:53:48 bly bgpd[1234]: [EC 100663313] SLOW THREAD: task bgpd_sync_callback (7fdb4984ed60) ran for 48218ms (cpu time 46469ms)
Sep 29 02:53:48 bly bgpd[1234]: Terminating on signal
Sep 29 02:54:02 bly zebra[1230]: [EC 4043309117] Client 'vnc' encountered an error and is shutting down.
Sep 29 02:54:02 bly zebra[1230]: [EC 4043309117] Client 'bgp' encountered an error and is shutting down.
Sep 29 02:54:02 bly watchfrr[1167]: [EC 268435457] bgpd state -> down : unexpected read error: Connection reset by peer
Sep 29 02:54:02 bly zebra[1230]: client 30 disconnected. 0 vnc routes removed from the rib
Sep 29 02:54:02 bly zebra[1230]: zebra/zebra_ptm.c:1345 failed to find process pid registration
Sep 29 02:54:02 bly watchfrr[1167]: bgpd state -> up : connect succeeded
Sep 29 02:54:02 bly zebra[1230]: client 20 disconnected. 847117 bgp routes removed from the rib
Sep 29 02:54:02 bly zebra[1230]: client 20 says hello and bids fair to announce only bgp routes vrf=0
Sep 29 02:54:02 bly zebra[1230]: client 32 says hello and bids fair to announce only vnc routes vrf=0
Sep 29 03:15:30 bly ospfd[1249]: ASBR[Status:0]: Update
Sep 29 03:20:37 bly ospfd[1249]: ASBR[Status:1]: Update
```

BGP C/P Protection

added to hphr

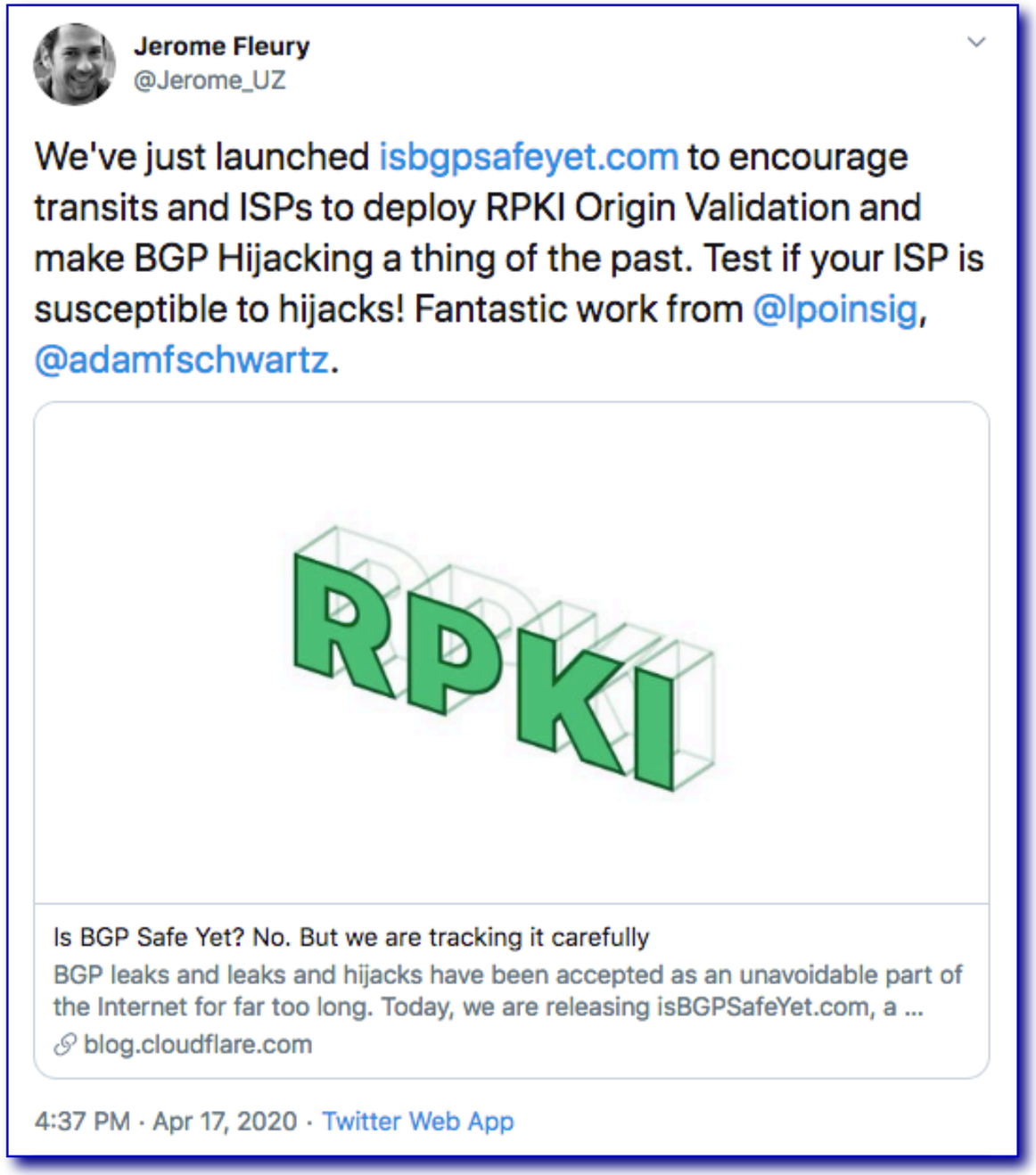
```
control-plane-protection:
    bgp:
        IPv4:
            "46.227.200.0/21": AS41495
            "185.134.196.0/22": AS41495
            "185.1.101.0/24": EquinixIX manchester
            "195.66.244.0/24": LINX manchester
            "31.217.131.64/29": AS43531
            "80.68.83.0/31": AS35425
            "149.6.10.128/29": AS174
            "77.67.124.100/30": AS3257
        IPv6:
            "2a01:9e00::/29": AS41495
            "2001:7f8:bc::/64": EquinixIX
            "2001:7f8:4:2::/64": LINX manchester
            "2a02:27f0:2:391::/64": AS43531
            "2001:668:0:3:ffff:1:0:4fc/126": AS35425
            "2001:978:2:24::/125": AS174
            "2001:41c8:2000:3::/126": AS3257
```

```
Sep 29 02:48:03 bly bg                peer
Sep 29 02:50:03 bly bg                set by peer
Sep 29 02:50:03 bly bg                peer
Sep 29 02:52:03 bly bg                set by peer
Sep 29 02:52:03 bly bg                peer
Sep 29 02:53:00 bly bg                2ms (cpu time 48285ms)
Sep 29 02:53:42 bly wa                seconds ago
Sep 29 02:53:42 bly wa                sh restart bgpd
Sep 29 02:53:48 bly bg                3ms (cpu time 46469ms)
Sep 29 02:53:48 bly bg
Sep 29 02:54:02 bly ze
Sep 29 02:54:02 bly ze
Sep 29 02:54:02 bly wa                by peer
Sep 29 02:54:02 bly ze
Sep 29 02:54:02 bly ze
Sep 29 02:54:02 bly wa
Sep 29 02:54:02 bly ze
Sep 29 02:54:02 bly ze
Sep 29 02:54:02 bly ze
Sep 29 03:15:30 bly os
Sep 29 03:20:37 bly os
```

INTERLUDE

“ IS BGP SAFE YET ”

isbgpsafeyet.com



Pride comes...


FAELIX
@faelix
Guess we were doing this before it was cool. #rpki #bgp
Jerome Fleury @Jerome_UZ · Apr 17
We've just launched isbgpsafeyet.com to encourage transits and ISPs to deploy RPKI Origin Validation and make BGP Hijacking a thing of the past. Test if your ISP is susceptible to hijacks! Fantastic work from @lpoinsig, @adamfschwartz. blog.cloudflare.com/is-bgp-safe-ye...
9:45 PM · Apr 17, 2020 · Twitter for iPhone

...before a segfault

```
Apr 17 08:22:42 aebi bgpd[1238]: /usr/lib/x86_64-linux-gnu/frr/libfrr.so.0(zlog_backtrace_sigsafe+0x67) [0x7f43747983d7]
Apr 17 08:22:42 aebi bgpd[1238]: /usr/lib/x86_64-linux-gnu/frr/libfrr.so.0(zlog_signal+0x113) [0x7f4374798833]3747983d7]
Apr 17 08:22:42 aebi bgpd[1238]: /usr/lib/x86_64-linux-gnu/frr/libfrr.so.0(+0x712e5) [0x7f43747b92e5]74798833]3747983d7]
Apr 17 08:22:42 aebi bgpd[1238]: /lib/x86_64-linux-gnu/libpthread.so.0(+0xf890) [0x7f43735c2890]92e5]74798833]3747983d7]
Apr 17 08:22:42 aebi bgpd[1238]: /usr/lib/frr/bgpd(+0x3cb6c) [0x5576bbdeeb6c]_main+0xf5) [0x5576bbe42415]92e5]74798833]3747983d7]
Apr 17 08:22:42 aebi bgpd[1238]: /usr/lib/x86_64-linux-gnu/frr/libfrr.so.0(+0x712e5) [0x7f43747b92e5]74798833]3747983d7]
Apr 17 08:22:42 aebi bgpd[1238]: /usr/lib/x86_64-linux-gnu/libpthread.so.0(+0xf890) [0x7f43735c2890]92e5]74798833]3747983d7]
Apr 17 08:22:42 aebi bgpd[1238]: /usr/lib/frr/bgpd(main+0x2ff) [0x5576bbdeeb6c]_main+0xf5) [0x5576bbe42415]92e5]74798833]3747983d7]
Apr 17 08:22:42 aebi bgpd[1238]: /lib/x86_64-linux-gnu/libc.so.6(__libc_start_main+0xf3) [0x7f43735c2b45]b00]]3747983d7]
Apr 17 08:22:42 aebi bgpd[1238]: /usr/lib/frr/bgpd(+0x3cb6c) [0x5576bbdeeb6c]_main+0xf5) [0x5576bbe42415]92e5]74798833]3747983d7]
Apr 17 08:22:42 aebi bgpd[1238]: in thread bgpd_sync_callback scheduled from bgpd/bgp_rpki.c:509#0 (0x5), aborting...
```

RPKI Kicked Our Butt!

... bgpd_sync_callback from bgpd/bgp_rpki.c

Shout out to @wolf480pl@mstdn.io for brainstorming this with me to be sure!

Fixed in FRRouting 7.2.1

BACK TO SCHEDULED CONTENT

...before an OOM

```
Sep  3 04:15:21 bly bgpd[1228]: [EC 100663313] SLOW THREAD: task bgpd_sync_callback (7fe7131dcd60) ran for 51696ms (cpu time 51688ms)
Sep  3 04:16:00 bly watchfrr[1167]: [EC 268435457] bgpd state -> unresponsive : no response yet to ping sent 90 seconds ago
Sep  3 04:16:00 bly watchfrr[1167]: [EC 100663303] Forked background command [pid 4836]: /usr/lib/frr/watchfrr.sh restart bgpd
Sep  3 04:16:12 bly bgpd[1228]: [EC 100663313] SLOW THREAD: task bgpd_sync_callback (7fe7131dcd60) ran for 50939ms (cpu time 50929ms)
Sep  3 04:16:12 bly bgpd[1228]: Terminating on signal
Sep  3 04:16:20 bly watchfrr[1167]: Warning: restart bgpd child process 4836 still running after 20 seconds, sending signal 15
Sep  3 04:16:20 bly watchfrr[1167]: restart bgpd process 4836 terminated due to signal 15
Sep  3 04:16:31 bly zebra[1224]: [EC 4043309117] Client 'vnc' encountered an error and is shutting down.
Sep  3 04:16:31 bly zebra[1224]: [EC 4043309117] Client 'bgp' encountered an error and is shutting down.
Sep  3 04:16:31 bly watchfrr[1167]: [EC 268435457] bgpd state -> down : unexpected read error: Connection reset by peer
Sep  3 04:16:31 bly zebra[1224]: client 30 disconnected. 0 vnc routes removed from the rib
Sep  3 04:16:31 bly zebra[1224]: zebra/zebra_ptm.c:1345 failed to find process pid registration
Sep  3 04:16:31 bly zebra[1224]: client 20 disconnected. 906341 bgp routes removed from the rib
Sep  3 04:17:21 bly watchfrr[1167]: [EC 100663303] Forked background command [pid 5576]: /usr/lib/frr/watchfrr.sh restart bgpd
Sep  3 04:17:21 bly zebra[1224]: client 20 says hello and bids fair to announce only bgp routes vrf=0
Sep  3 04:17:21 bly zebra[1224]: client 32 says hello and bids fair to announce only vnc routes vrf=0
Sep  3 04:17:21 bly watchfrr[1167]: bgpd state -> up : connect succeeded
```

restart bgpd process 4836 terminated due to signal 15
client 20 disconnected. 906341 bgp routes removed from the RIB

...BACK TO SCHEDULED CONTENT...?

...before a segfault

```
Sep  3 09:32:16 aebi watchfrr[1302]: [EC 268435457] ospf6d state -> down : read
returned EOF
Sep  3 09:32:16 aebi zebra[1332]: [EC 4043309121] Client 'ospf6' encountered an
error and is shutting down.
Sep  3 09:32:16 aebi zebra[1332]: client 52 disconnected. 47 ospf6 routes removed
from the rib
Sep  3 09:32:21 aebi watchfrr[1302]: [EC 100663303] Forked background command [pid
11593]: /usr/lib/frr/watchfrr.sh restart ospf6d
```

Client 'ospf6' encountered an error and is shutting down.
client 52 disconnected. 47 ospf6 routes removed from the rib

...before a segfault

```
Sep  3 09:32:16 aebi watchfrr[1302]: [EC 268435457] ospf6 state -> down : read
returned EOF
Sep  3 09:32:16 aebi zebra[1332]: [EC 4043309121] Client 'ospf6' encountered an
error and is shutting down.
Sep  3 09:32:16 aebi zebra[1332]: client 52 disconnected. 47 ospf6 routes removed
from the rib
Sep  3 09:32:21 aebi watchfrr[1321]: [EC 100663303] Forked background command [pid
11593]: /usr/lib/frr/watchfrr.sh restart ospf6d
```

FRR #6735 + #6086

Client 'ospf6' encountered an error and is
client 52 disconnected. 47 ospf6 routes removed from the rib

Fixed in FRRouting 7.5

THIS STORY ISN'T OVER...

XXXTODO

- Add features to FRR, provide VyOS workarounds, or make Salt states for router reboot maintenance
- Speed-up VyOS' prefix-list insertion into FRR ⇒ T2425
- VyOS to refresh config into FRR's daemons ⇒ T2175
- Fragility of VyOS firewalling when replaced with barebones iptables from hphr
- NIC parameter and bare metal router sysctl tuning
- XDP (hopefully we get time to follow on from @atoonk)

What We Achieved

- Two very long nights of maintenance (plus London)
- Full table BGP on VyOS converge time in seconds
- Routing on MikroTiks converges near-instantly
- BCP38 (customers cannot spoof source address)
- IRR filtering* (only accept where route/route6 object)
- RPKI (will not accept invalid routes from P/T)
- Templated configuration (repeatable, automated)
- Single source of truth (the docs become the config)

VYOS & RPKI AT THE BGP AS EDGE

E: marek @ faelix . net
T: @maznu

E: lou @ faelix . net
T: @gremlinlou