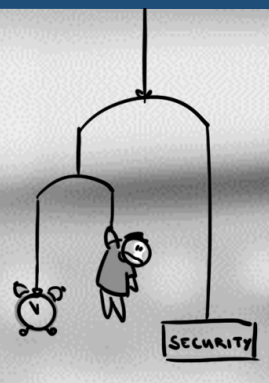


IPv4-with-IPv6 Next-Hop

Tobias Fiebig
Max-Planck Institut für Informatik

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

Stop Doing IPv4 Driven Addressing Plans

- IPv4 around for +40 years and still no solution for making clean addressing plans!
- Why is there CIDR if it makes IP?
- Want to work on Layer 3? Find the Layer 2 address first!
- "Yes, please give me 50 different prefixes in a SINGLE metro!"

- There is a solution: Embrace our lord an savior RFC8950!

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

supported by: RIPE NCC Community Projects Fund

ARP & NDP

ARP

- Figure out to which MAC address to send packets for an IPv4 address
- Construct ethernet frame with fitting destination MAC

NDP

- Figure out to which MAC address to send packets for an IPv6 address
- Construct ethernet frame with fitting destination MAC

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

supported by:
RIPE NCC Community Projects Fund
LWLCOM connecting you.
OPENFACTORY
Virtuacloud
DE-CIX
WIIT THE PREMIUM CLOUD

RFC5549, RFC8950, and a draft

RFC5549 & RFC8950

- RFC5549 in 2009: What if we just put an IPv6 address into the nexthop field of an IPv4 prefix in BGP?

draft-chroboczek-intarea-v4-via-v6-00

- Still active: How to actually handle an IPv4 route with an IPv6 nexthop; Essentially:
  - Ask for MAC you'd have to send packets for the IPv6 nexthop to
  - Send the IPv4 packet there

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

supported by: RIPE NCC Community Projects Fund · LWLCOM connecting you. · OPENFACTORY · Virtuacloud · DE-CIX · WIIT THE PREMIUM CLOUD

RFC8950 & Draft Vendor Support

BGP AFI v4 with v6 Nexthop

- JunOS
- Arista
- Cisco
- ExaBGP (no FIB)
- FRR
- Bird

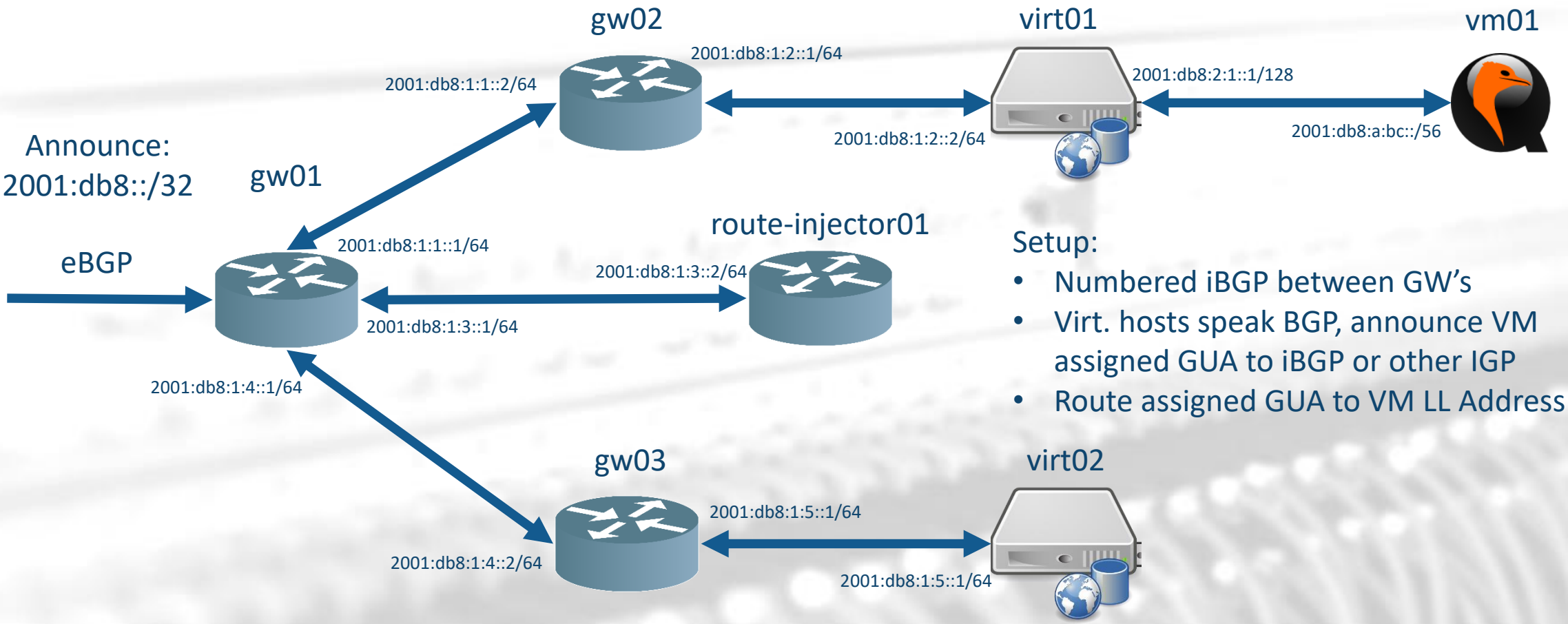
Actually sending packets

- JunOS
- Arista
- Cisco
- Linux (netlink)
- FreeBSD

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

supported by:
RIPE NCC Community Projects Fund
LWLCOM connecting you.
OPENFACTORY
Virtuacloud
DE-CIX
WIIT THE PREMIUM CLOUD

Addressing With Port... er v4-w-v6-nh

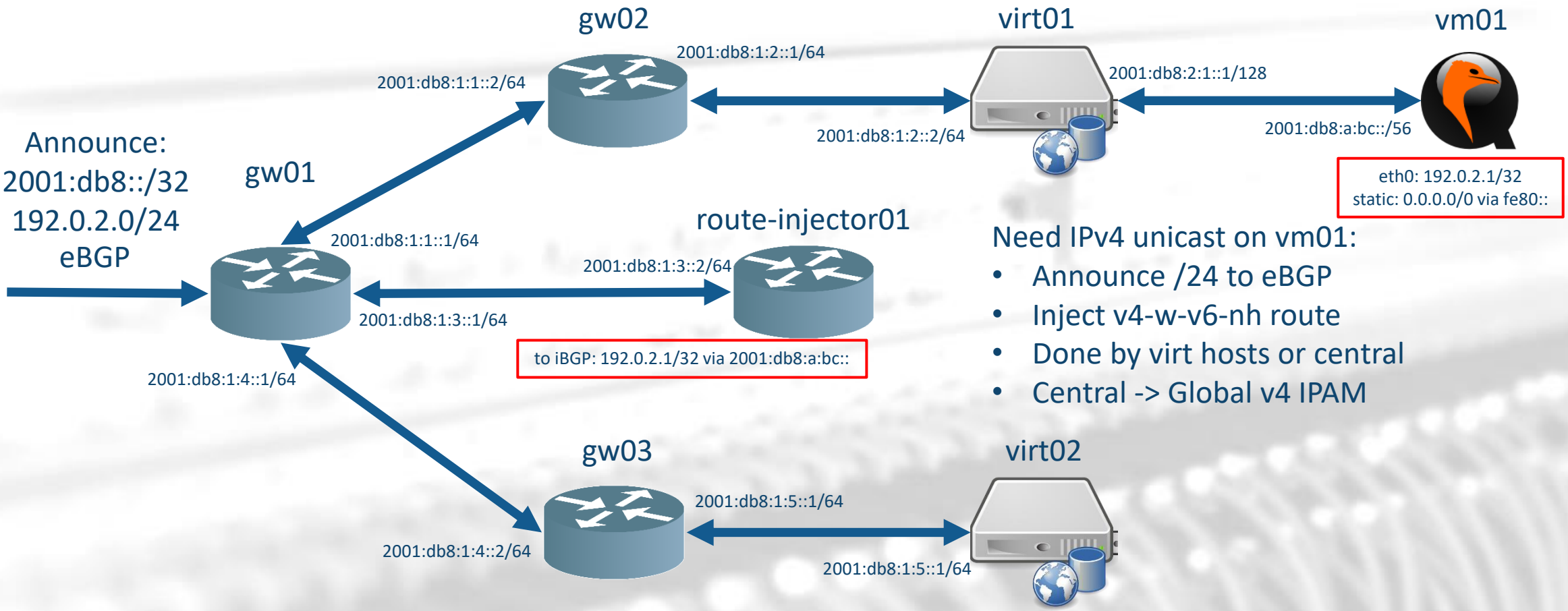


Setup:

- Numbered iBGP between GW's
- Virt. hosts speak BGP, announce VM assigned GUA to iBGP or other IGP
- Route assigned GUA to VM LL Address

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

Addressing With Port... er v4-w-v6-nh

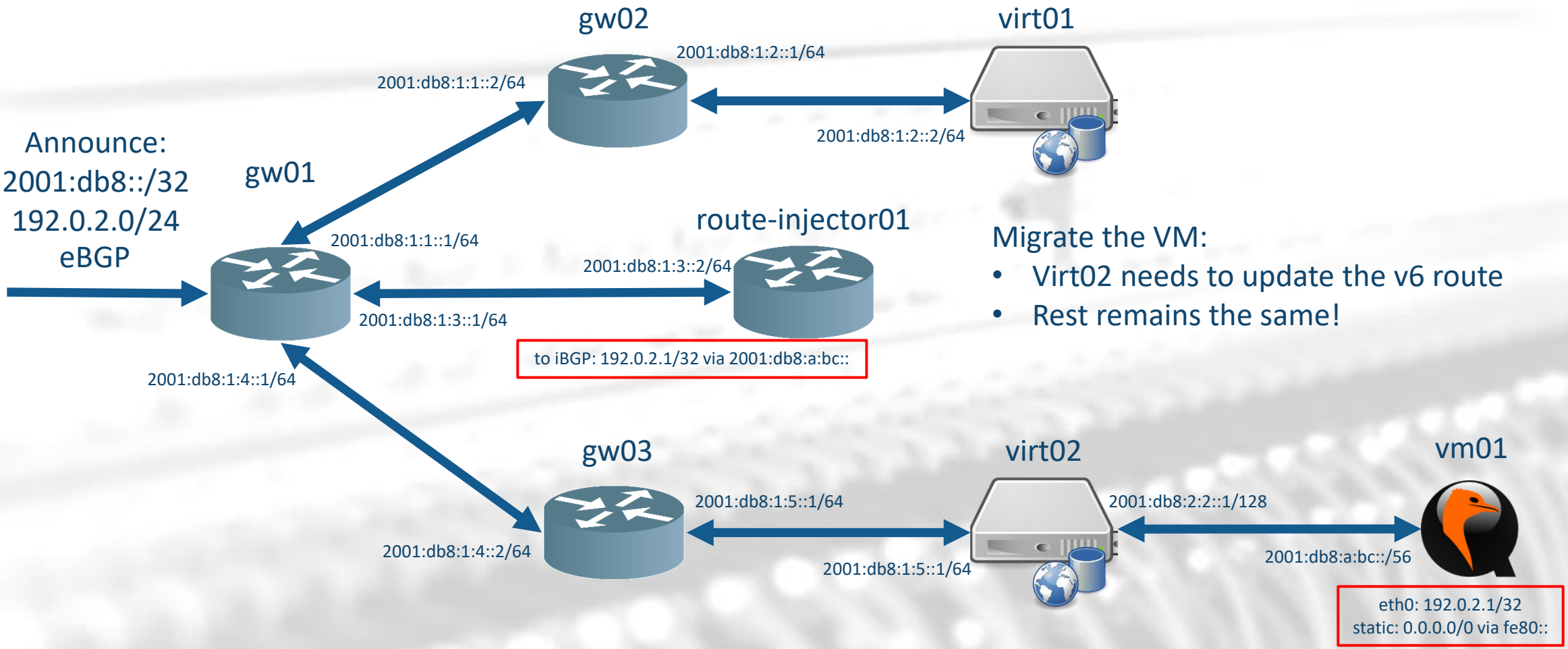


Need IPv4 unicast on vm01:

- Announce /24 to eBGP
- Inject v4-w-v6-nh route
- Done by virt hosts or central
- Central -> Global v4 IPAM

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

Addressing With Port... er v4-w-v6-nh



Migrate the VM:

- Virt02 needs to update the v6 route
- Rest remains the same!

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

Advantages of v4-w-v6-nh

- Fine grained (/32) routing of IPv4
- IPv4 as a complete add-on
- Clean IPv6 centric addressing plan
- No need for IPv4 transfer/router/network/broadcast addresses
  - This includes eBGP!
  - Really use *all* you IPv4 addresses
- IPv4 'follows' an IPv6 address (prefix) around
- *Technically* there does not even have to be a loopback IPv4 address on routers; Nice for traceroute/PMTUD though
- Legacy services can be connected behind v4-w-v6-nh transport
- This can be done partially (only for end-hosts, only for transport…)

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

supported by: RIPE NCC Community Projects Fund

RFC8950 on the IX

- Currently a Euro-IX WG looks into RFC8950 at the IX: https://github.com/euro-ix/rfc8950-ixp
- IXP prefixes should usually not be globally reachable anyway
- One *should* originate ICMP for IX interfaces from lo anyway
- There are no concerns about too many members anymore

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

supported by:
RIPE NCC Community Projects Fund
LWLCOM connecting you.
OPENFACTORY
Virtuacloud
DE-CIX
WIIT THE PREMIUM CLOUD

Caveats of v4-w-v6-nh

- Traceroutes are less (only router loopback) or completely useless (not even IPv4 loopback)
- Does not work for some client OSes yet
- Needs vendor support on routing infra
- Works best with a clean IPv6 addressing policy
- You still need a working IPv6 IPAM

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

Roadmap: What is needed

- Better vendor support (currently only Bird supports injection)
- Draft for DHCP4 via IPv6 GUA; Get v4 from a central location only if (when) needed
- Draft to let clients know about the local DHCP4 server via RA
- Make interface identification better
  - RFC4950 for v4-w-v6-nh
  - RFC5837(++)
- *SOME* way to actually fix PMTUD
- Finally an IPv4 free Internet core

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

supported by:
RIPE NCC Community Projects Fund
LWLCOM connecting you.
OPENFACTORY
Virtuacloud
DE-CIX
WIIT THE PREMIUM CLOUD

RFC8950 on Transport & eBGP for me

- Removed all transfer IPv4 in AS59645 (except for that one OpeBSD router); Works since over a year on JunOS / VyOS
- Using private ASN eBGP underlay for LL/Loopback distribution in AS59645
- Setup a dedicated test setup which uses the no-IPv4-except-on-leaf approach using FRR (edge/dist) and bird (injection)
- Currently five eBGP sessions without IPv4:
  - 2x Upstream from AS59645 to AS215250 (default route)
  - 1x Peering AS59645 to AS215250 (harvesting higher LPREF)
  - 1x Peering AS211286 to AS215250 (harvesting higher LPREF)
  - 1x BGP.tools route collector export (worked OOTB ;-))

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

V4LESS-AS: Testing RFC8950 in Practice

- The dedicated test setup:
  - AS215250
    45.91.12.0/24
    2a06:d1c3::/32
- Running different test scenarios:
  - lo with/without IPv4
  - On-path MTU break
  - eBGP
  - RPKI/IRR invalid IPv4 prefixes (TBD, to allow filter testing)
- RIPE Atlas & NLNOG RING nodes for introspection
- See: https://measurement.network/services/v4less-as/

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

supported by: RIPE NCC Community Projects Fund

V4LESS-AS: How it works

- Available at IXPs with pre-configured passive higher LPREF sessions for all members
- Fragmenting backhaul tunnels for clean 1500 MTU from the border
- You can establish an RFC8950 session and use the RIPE Atlas/NLNOG Ring nodes for testing

- Currently active IXPs
  - BCIX
  - FogIXP
  - FNC-IX
  - France-IX
    - Lille, Paris
  - DE-CIX
    - Dusseldorf, Frankfurt, Munich, Hamburg, New York, Istanbul, Madrid, Marseille
  - And soon an IX near you… ?

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

supported by:
RIPE NCC Community Projects Fund
LWLCOM connecting you.
OPENFACTORY
Virtuacloud
DE-CIX
WIIT THE PREMIUM CLOUD

Key Take-Aways

- RFC8950 / v4-w-v6-nh is the future
  - Allows you to fully leverage all IPv4 you have
  - Build a clean IPv6 centric addressing scheme/architecture
  - Have IPv4 as a flexible add-on with central IPv4 IPAM
- Setup a session to AS215250 at a common IX and test it out
- Reach out to contact@measurement.network to sponsor a presence at an IX near you!
  - Needs: IX port, VM with 4-8GB memory, 2 cores, additional IPv6 only interface with static routing for mgmt / backhaul

https://measurement.network/services/v4less-as/

V4LESS-AS URL

Tobias Fiebig – tfiebig@mpi-inf.mpg.de

supported by: RIPE NCC Community Projects Fund