Peering observations 2007 vs. 2016

Greg Hankins, AS ~~18508~~ 38016
~~<ghankins@force10networks.com>~~ <greg.hankins@nokia.com>
~~European~~ Global Peering Forum ~~2, Barcelona~~ 11, Hollywood, CA

Image source: http://as2914.net/

  Public

NOKIA

GPF 11 2016/04/13

Agenda

1. Introduction
2. Research questions
3. BGP configuration results
4. BGP route results
5. Route server results
6. ARP/ND scaling

Supporters

- Contributors: DE-CIX R&D
  - Thomas King <thomas.king@de-cix.net>
  - Christoph Dietzel <christoph.dietzel@de-cix.net>
- Thanks: everyone who peers with AS 38016!

My perspective on peering

Managed BGP routing and peering at AS 4355

Peering project with AS 18508 at PAIX

Peering project with AS 38016 at DE-CIX

1998 1999 … … … 2007 … … … 2016

Started working for router vendors

PAIX Palo Alto:
118 members

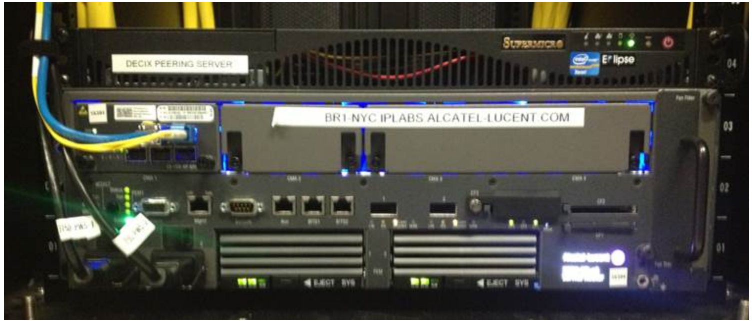
DE-CIX Frankfurt:
677 members
DE-CIX New York:
113 members

Changes in the Industry

Goal of this presentation: look at how peering is deployed from an operational perspective

Peering project with AS 38016 at DE-CIX

- DE-CIX offers space for an R&D project at DE-CIX New York
- IX Reach offers VLANs to DE-CIX Frankfurt and for transit
- Install a router and a server at 325 Hudson
- Update PeeringDB
- Start peering at DE-CIX Frankfurt and DE-CIX New York using R&D AS 38016 (Nokia IP Labs)
- Observe passive measurements
  - Being a passive observer doesn't influence behavior
  - Look for interesting things
  - Compare to what has changed since 2007



  Public

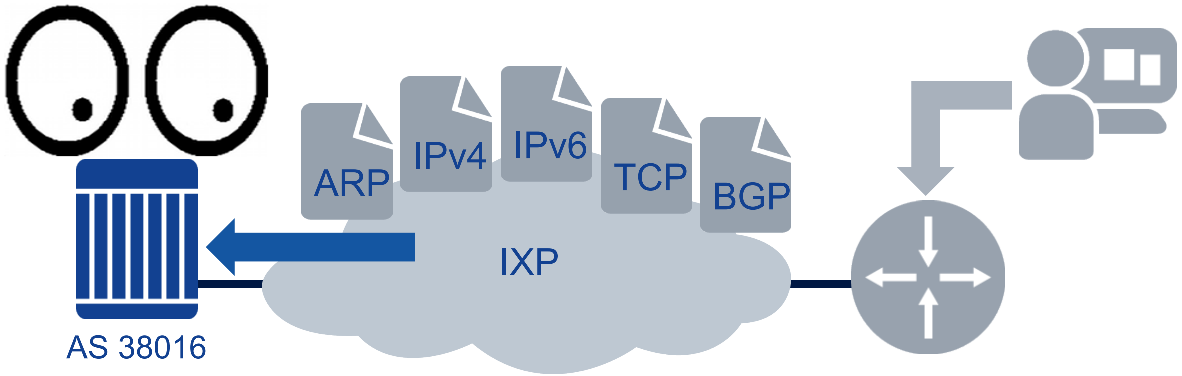
Research questions

- Very interested in how peering is deployed from a operational configuration and technical perspective
- What features are people using and not using at an IXP?
  - And why?
  - What are the implications?
  - Especially interesting to me as a router vendor because of feature impact
- Not really interested in prefixes or path attributes
  - Other excellent research projects* are looking at these things

```
A:br1-nyc.iplabs# configure
A:br1-nyc.iplabs>config# router
A:br1-nyc.iplabs>config>router# bgp
A:br1-nyc.iplabs>config>router>bgp# info
-------------------------------------------------
            enable-peer-tracking
            add-paths
                ipv4 send 1 receive
                ipv6 send 1 receive
            exit
            backup-path ipv4 ipv6
            best-path-selection
                always-compare-med strict-as zero
                deterministic-med
            exit
A:br1-nyc.iplabs>config>router>bgp#
```

*See references slide

What can you observe passively from BGP peering?



Can observe peering behavior	Cannot observe configuration
• Control plane packets	• Router configuration
• Protocols (IPv4/IPv6)	• Routing policies
• BGP capabilities	• Peering policies
• BGP routes and prefixes	• Network architecture
• BGP path attributes	• Not interested in these things

BGP configuration results: peering configuration

Configuration	2007 PAIX Palo Alto	2016 DE-CIX Frankfurt & New York	Observations
Unique BGP peers[1]	21	25	
Total BGP peers[2]	30	76	IPv6, multiple connections, route servers
IPv4 unicast BGP peers	21	39	
IPv6 unicast BGP peers	6	37	Almost 100% IPv6 peering!
IPv4 multicast BGP peers	3	0	No interdomain multicast
IPv6 multicast BGP peers	0	0	No interdomain multicast
MSDP peers	1	0	No interdomain multicast
BFD peers	0	0	Interesting!
Route server peers	0	4	Not available in 2007 at PAIX

[1]Unique BGP peers: unique peers based on peer ASN
[2]Total BGP peers: total operational BGP peers

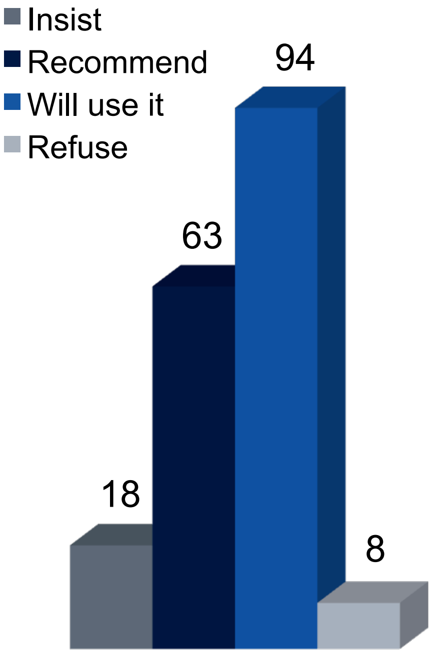
BGP security features in use

Feature	2007	2016	Observations
Unique BGP peers	21	25	
Protection of BGP Sessions via the TCP MD5 Signature Option (RFC 2385)	16	4	Rarely used or required now, but why?
The Generalized TTL Security Mechanism (GTSM) for BGP (RFC 5082)	Didn’t count	0	23 with TTL = 1 2 with TTL = 64
The TCP Authentication Option (RFC 5925)	Not available	0	No interest?

  Public

Protection of BGP Sessions via the TCP MD5 Signature Option

- RFC 2385 basics
  - Defines a TCP option for carrying an MD5 digest in a TCP segment
  - The digest is a signature for that segment
  - Protects BGP against spoofed TCP segments (RST)
- People generally dislike MD5, because it causes more operational problems than it solves
  - Password negotiation
  - Password management, storage and lost passwords
- Notable decrease in use, and rarely required now
  - 2007: 16 of 21 unique peers
  - 2016: 4 of 25 unique peers



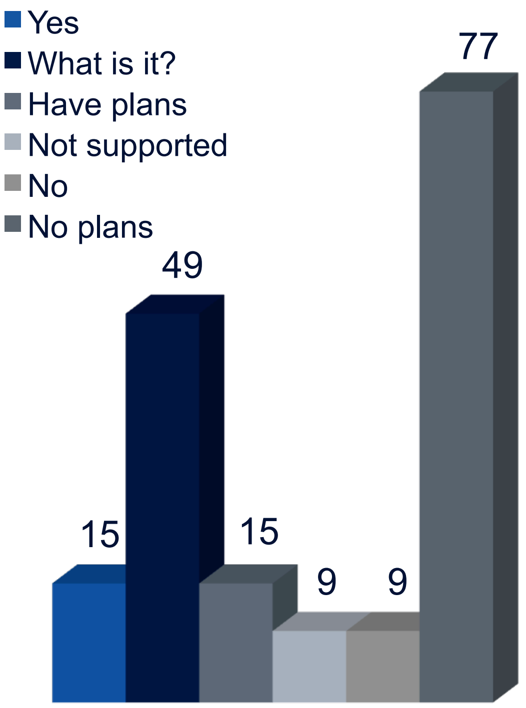
Peering Survey 2008 – MD5
169 respondents

Source: http://www.twoguys.org/~gregh/peering-survey-2008/peering-survey-2008-results.pdf

  Public

The Generalized TTL Security Mechanism (GTSM) for BGP

- RFC 5082 basics for BGP
  - EBGP peering is usually between routers that are on the same IP network, especially at an IXP
  - Use Time to Live (TTL) (IPv4) or Hop Limit (IPv6) to see if a packet came from a different IP network
  - Send TTL = 255, expect TTL = 255
  - TTL < 255 means the packet was spoofed, because routers always decrement the counter
- Requires manual negotiation and configuration
  - Commands and logic varies between BGP implementations
  - Must configure matching send and receive behavior on each router
- Not requested or in use by any peers
  - 23 of 25 unique peers sending TTL = 1
  - 2 of 25 unique peers sending TTL = 64



Peering Survey 2008 – GTSM
169 respondents

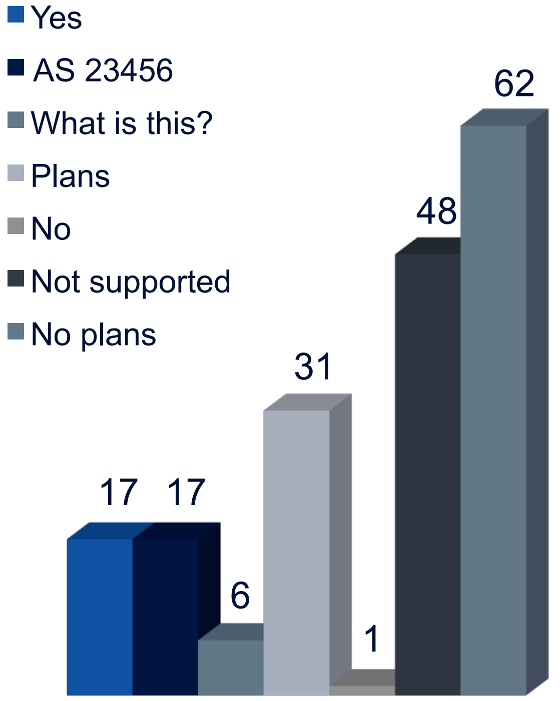
Source: http://www.twoguys.org/~gregh/peering-survey-2008/peering-survey-2008-results.pdf

BGP capabilities in use

Feature	2007	2016	Observations
Unique BGP peers	21	25	
Multiprotocol Extensions for BGP-4 (RFC 4760, Capability Code 1)	21	25	Router is capable of MP, probably IPv6
Multiple address families advertized	Didn't count	2	1 x "IPv4 IPv6", 1 x "IPv4 MCAST-IPv4"
Route Refresh Capability for BGP-4 (RFC 2918, Capability Code 2)	21	25	Enabled by default on most routers, but?
Old Cisco Route Refresh Capability (Capability Code 128)	21	0	Historical, not advertised anymore
Enhanced Route Refresh Capability for BGP-4 (RFC 7313, Capability Code 70)	Not available	0	Too early?
Outbound Route Filtering Capability for BGP-4 (RFC 5291, Capability Code 3)	0	0	Does it make sense at an IXP?
Graceful Restart Mechanism for BGP (RFC 4724, Capability Code 64)	2	0	Interesting!
BGP Support for Four-Octet Autonomous System (AS) Number Space (RFC 6793, Capability Code 65)	0	24	Broad support now
Advertisement of Multiple Paths in BGP (draft-ietf-idr-add-paths, Capability Code 69)	Not available	0	Too early?

BGP Support for Four-Octet Autonomous System (AS) Number Space

- RIRs issue lots of these ASNs now
- Supported by all current BGP implementations
- Required for peering now and broadly deployed
- Seen somewhere in the AS path for
  - 9.29% of IPv4 routes (152,432 of 1,639,291 routes)
  - 11.69% of IPv6 routes (18,154 of 201,884 routes)
  - Edge and transit ASNs
- Quite a change from previous results
  - 2007: 0 of 21 unique peers
  - 2015: 24 of 25 unique peers



Peering Survey 2008 – Four-Octet ASN
169 respondents

Source: http://www.twoguys.org/~gregh/peering-survey-2008/peering-survey-2008-results.pdf

Interesting observations based on protocol configurations

Category	Observations	Conclusions?
Security	• Given that the IXP network is often leaked into global routing table • MD5 is rarely used • No TCP-AO • No GTSM	• Don't care? • Too complicated? • Trust IXP and peers? • Is TCP/BGP packet security (still) an issue? • Use policy filtering and/or external databases instead?
Resiliency	• No BFD, peer aliveness detection via slow BGP mechanism • No BGP graceful restart • No ADD-PATH • BGP route refresh is enabled by default, is it used?	• Don't care? • Too complicated? • Rely on stability of IXPs or peering infrastructure? • Design around IXP and peer failures?

Little interest in member protocol security and resiliency at IXPs… but why?

BGP route results: BGP MULTI_EXIT_DISC (MED)

- MEDs are a metric used to influence traffic coming into your AS
- A large variety of MEDs are received on routes
- Generally considered a bad idea to compare MEDs between ASs, because they mean different things to different people
  - [RFC 4451 BGP MULTI_EXIT_DISC (MED) Considerations](), section 4.1
- Implementations vary in how MEDs are compared, and what values a missing and default MED means
  - Check your router configuration carefully

	IPv4 routes	IPv6 routes
Total RIB routes	1,639,291	133,852
No MED	220,507 (13.45%)	35,797 (26.74%)
MED = 0	138,050 (8.42%)	26,839 (20.05%)
MED > 0	1,280,734 (78.13%) Values 1 – 4,294,967,295	71,216 (53.21%) Values 1 – 48,512

Why is there such a big difference in IPv4 and IPv6 MED values?

IPv4 prefix distributions

- Received IPv4 prefix distributions correlate well with statistics from AS 131072 (APNIC R&D) and AS 6447 (Route Views) projects
- These projects have a lot more detailed and historical data, and this is not the focus of these observations
- However, it's interesting to note that a small percentage of prefixes with masks /25 - /32 are received at an IXP
  - Including many with blackhole next hop, mostly /32s
- This is not a big deal today, but something worth keeping an eye on as IPv4 address exhaustion continues

Prefix length	AS 38016 %	AS 131072 %	AS 6447 %
/22	11.22	11.02	11.38
/23	9.83	9.66	9.55
/24	55.93	54.98	54.01
/25	**0.02**	**0.10**	**0.20**
/26	**0.02**	**0.10**	**0.21**
/27	**0.00**	**0.08**	**0.15**
/28	**0.00**	**0.01**	**0.12**
/29	**0.01**	**0.01**	**0.17**
/30	**0.01**	**0.01**	**0.12**
/31	**0.00**	**0.00**	**0.01**
/32	**0.17**	**0.01**	**0.49**

/25 - /32: 0.23% or
3,900 of 1,639,291 total IPv4 RIB routes

1,906 routes have blackhole next hop

Source: http://bgp.potaroo.net/index-bgp.html, Tue Apr 5 18:10:17 2016 (UTC+1000)

IPv6 prefix distributions

Prefix length	AS 38016 %	AS 131072 %	AS 6447 %
/32	38.10	26.92	24.98
/48	46.78	46.12	43.51
/64	0.02	0.00	3.34
/125	**0.01**	**N/A**	**0.07**
/126	**0.24**	**N/A**	**1.57**
/127	**0.01**	**N/A**	**0.09**
/128	**0.01**	**N/A**	**0.22**

/125 - /128: 0.27% or 394 of 133,852 total IPv6 RIB routes

2 routes have blackhole next hop

Source: http://bgp.potaroo.net/index-bgp.html, Tue Apr 5 18:45:03 2016 (UTC+1000)

Route server results: route server peering

- Route server availability
  - Only at a few IXPs in the “early” days
  - A lot of development work and deployment in the mid/late-2000s
  - Now considered part of standard IXP infrastructure
- Can get large percentages* of the IPv4 and IPv6 Internet routing tables just by peering with route servers
- Some networks do not peer with route servers for their own reasons
- However, that’s a lot of reachability for very little effort

Route server	Member IPv4 peering (%)	IPv4 routes (% of full table)	Member IPv6 peering (%)	IPv6 routes (% of full table)
rs1 FRA	82.62	130,620 (21.81)	64.09	12,910 (45.70)
rs2 FRA	79.79	129,998 (21.71)	62.80	12,920 (45.73)
rs1 NYC	72.41	52,696 (8.80)	58.62	11,540 (40.85)
rs2 NYC	70.68	52,696 (8.80)	58.62	11,540 (40.85)

Why is there such a big difference in IPv4 and IPv6 peering and route percentages?

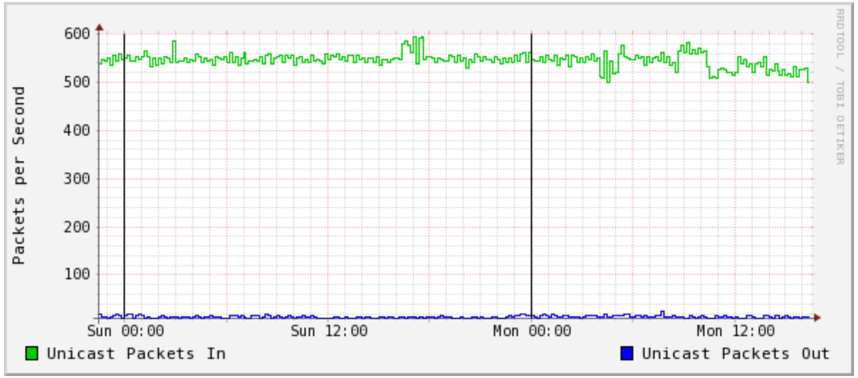
*Source: simple napkin math based on received routes and FIB size of 598,815 IPv4 routes and 28,248 IPv6 routes

ARP/ND scaling: problem statement

- Peering fabric networks are getting bigger as IXPs connect more members
  - DE-CIX Frankfurt, for example has grown from their original /24 ⇒ /23 in 2001 ⇒ /22 in 2008 ⇒ /20 in the future
  - Quadratic rate of increase of ARP/ND traffic volume measured*
- This starts to present a problem with ARP/ ND control plane packets that adds up
  - Packet/s: per member and total network
  - Kb/s: per member and total network
  - Router CPU: small % utilization per member
  - ARP and neighbor table entries: per member

Mask	Hosts
/24	254
/23	510
/22	1022
/21	2046
/20	4094

DE-CIX Frankfurt:
677 members



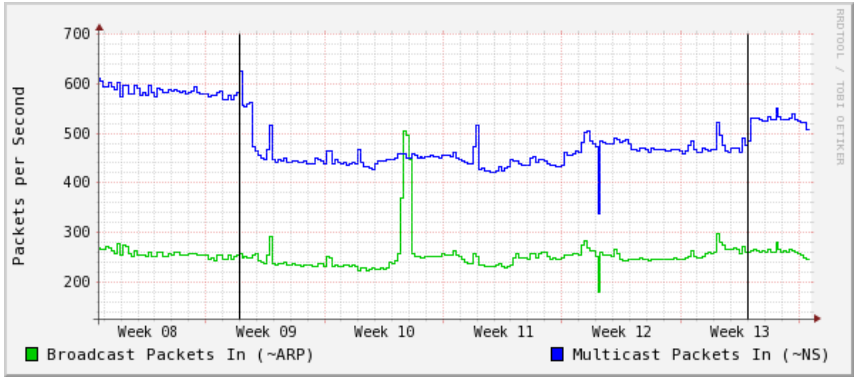
DE-CIX Frankfurt control plane traffic

*Source: https://www.euro-ix.net/s/legacy/documents/1236/original/ARP%20and%20ND%20Handling%20with%20VPLS.pdf

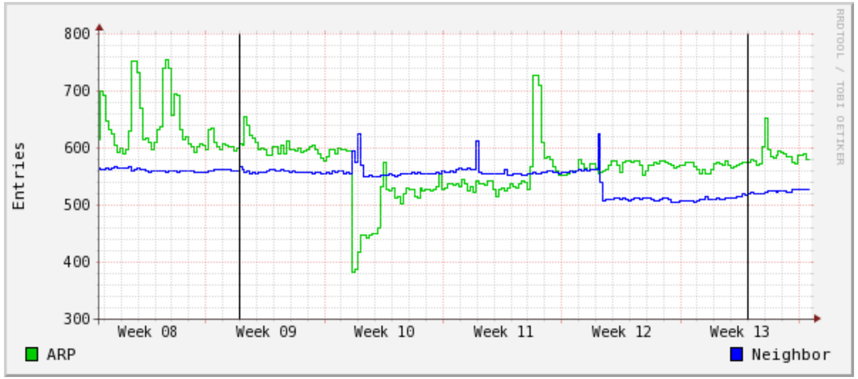
  Public

ARP/ND scaling

- Average rates of ~255 ARP/s and ~524 NS/s at DE-CIX Frankfurt with 677 members
  - Broadcast traffic is entirely ARP
  - Multicast traffic is mostly NS with some MLD
  - NS for global and link-local addresses is > ~2 x ARP traffic
- ~582 ARP and ~527 neighbor entries on my router, but I only have 76 total peers
- Mitigation solutions exist to “sponge” invalid requests, but not to reduce or eliminate valid requests
- EVPN* is a possible solution as described in [draft-ietf-bess-evpn-proxy-arp-nd]
  - Static IP/MAC configuration
  - Local proxy-ARP/ND from each IXP edge device
  - Eliminates ARP/ND from IXP core
  - Reduces ARP/ND to each member to only what is needed



DE-CIX Frankfurt ARP and NS packets/second



Total ARP and neighbor entries

*Ethernet VPN (EVPN) - Overlay Networks for Ethernet Services: https://ripe68.ripe.net/presentations/170-ripe-68-evpn.pdf

Conclusions on peering in 2007 vs. 2016

- Fundamentally, not much has changed
  - Almost 100% IPv6 peering
  - More prefixes and some longer masks
- Several members now have multiple connections each with multiple links
- Route servers now provide a simple way to immediately benefit from peering at an IXP
- Few new protocol features are used, mainly four-octet ASNs
- Low requirements for security and resiliency features
- ARP/ND scaling could become a problem at large IXPs

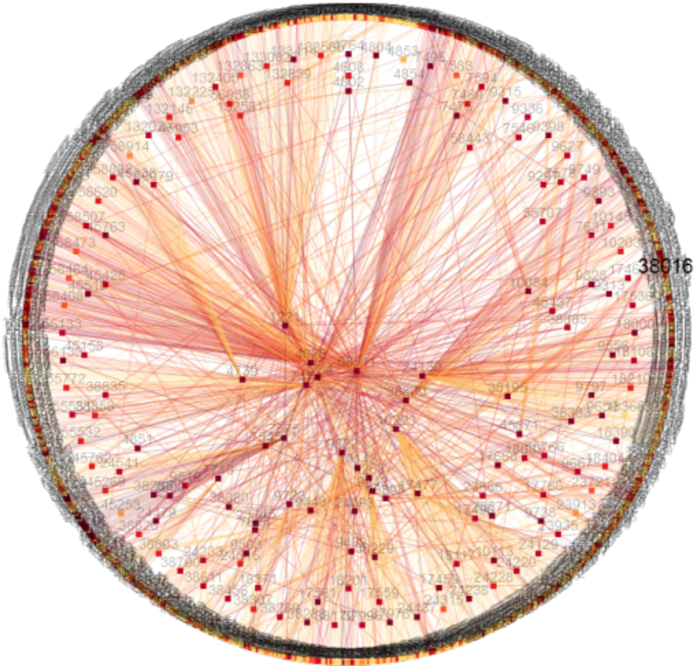


Image source: http://labs.apnic.net/vizas/

  Public

What's next?

- AS 38016 has a different vantage point at an IXP than AS 131072 (APNIC R&D) and AS 6447 (Route Views)
- Would like to focus on prefix and path analysis specifically as they related to bilateral and route server (multilateral) peering at IXPs
- We'll see what interesting results we can find…

References

- Peering Observations
  - Presented at EPF 2, Barcelona
  - Not available online anymore, ask me and I'll send it to you
- Peering Survey 2008 Announcement
  - https://www.nanog.org/meetings/nanog42/presentations/peering-bof-XVII.pdf
- Peering Survey 2008 Results
  - http://www.twoguys.org/~gregh/peering-survey-2008/peering-survey-2008-results.pdf
  - http://meetings.ripe.net/ripe-56/presentations/eix/Hankins-Peering_Survey_2008_Results.pdf
  - http://www.uknof.org.uk/uknof10/Davidson-Peering-survey.pdf
- All data as seen on Mon Apr 5 04:00:00 EDT 2016 by AS 38016
- Some excellent research projects and tools
  - APNIC R&D: http://bgp.potaroo.net/
  - Route Views: http://www.routeviews.org/
  - RIPEstat: https://stat.ripe.net/
  - BGPmon: http://www.bgpmon.net/
  - NLNOG RING: https://ring.nlnog.net/

Questions?
Answers?
Comments?
Observations?

  Public

NOKIA