Structural problems in the IPv6 routing

Bernhard Schmidt
Leibniz Supercomputing Centre, Munich, Germany

schmidt@lrz.de
berni@birkenwald.de

RIPE-56, Berlin, May 2008

User experience key to IPv6 success

making a connection (pseudocode)

```
addrlist = getaddrinfo(host, AF_UNSPEC)
for addr in addrlist:
     if (socket = connect(addr)):
          break

write(socket, ....)
```

- Ordering of results depends on local implementation, usually prefers IPv6 over IPv4
- Serialized connection attempts

Multiple possible error cases

- good – immediate next try
  - ICMPv6 Type 1 Code 0 – no route to destination
  - ICMPv6 Type 1 Code 3 – Address unreachable (ND timeout)
  - TCP RST - connection actively refused (service not listening on IPv6)
- bad
  - no reply at all – connection timeout (up to several minutes)
- worse – connection established but
  - extremely bad throughput or high latency
  - pMTU issues
  - TCP never recovers!

Webforums (eew!) are loaded with the advice to disable IPv6 first thing on connection problems.

Evolution of routing policy

- fulltable swaps on worldwide tunnels
- prefer shorter tunnels, keep traffic regional
- limit amount of routes being exchanged
- transit-peer-downstream relations (BGP policy)
- transit-peer-customer relations (economic)

but some networks are still operating like in 1996

- get all paths between two large ASNs (e.g. AS3257 and AS6939) from GRH
- if a direct *peering* exists, all customer prefixes behind AS6939 should be visible through the direct AS-path “3257 6939 <customer>”

```
grh.sixxs.net> sh bgp ipv6 regexp ^3257_6939_
2001:278::/32      3257 6939 4725 i
2001:288::/32      3257 6939 9264 1659 17717 i
2001:388:28::/48   3257 6939 7575 i
2001:388:1034::/48 3257 6939 7575 i
[...]
Total number of prefixes 97
```

- get all paths between two large ASNs (e.g. AS3257 and AS6939) from GRH
- if a direct *peering* exists, all customer prefixes behind AS6939 should be visible through the direct AS-path “3257 6939 <customer>”

```
grh.sixxs.net> sh bgp ipv6 regexp ^3257_6939_
2001:278::/32      3257 6939 4725 i
2001:288::/32      3257 6939 9264 1659 17717 i
2001:388:28::/48   3257 6939 7575 i
2001:388:1034::/48 3257 6939 7575 i
[...]
Total number of prefixes 97
```

- if you can still see indirect paths “3257 <transit>+ 6939” those prefixes are not treated as customer prefixes by AS6939 and exported as part of a fulltable to a downstream
- the downstream neighbor advertises it to other neighbors and the paths get eventually advertised to AS3257 – we have a leak

```
grh.sixxs.net> sh bgp ipv6 regexp ^3257_.+_6939_
2001:220:4000::/34  3257 6175 17715 6435 6939 2516 7660 24287 24490 9270 38128 i
2001:220:8000::/33  3257 6175 17715 6435 6939 2516 7660 24287 24490 9270 38128 i
[...]
Total number of prefixes 8
```

- 97 direct paths on “3257 6939”
  - HE.net customer prefixes advertised on their peering with Tiscali
- 8 indirect paths on “3257 .+ 6939”
  - HE.net peering/upstream prefixes advertised to a downstream client as part of a fulltable and then eventually leaked
  - 5 paths going through “3257 6175 17715 6435 6939 .+”
  - 3 paths going through “3257 2497 2500 4725 6939 .+”
- result for 3257 -> 6939: **97** / **8**
- do the same analysis for every combination, get a peering matrix
- try to do some educated guesses on the data

Peering matrix

Direct paths (“Source-AS Destination-AS”) / indirect paths (“Source-AS Transit-AS Destination-AS”)

Source	Destination									
	1273	2914	3257	3356	3549	6453	6939	12702 [1]	30071	Total
1273		104 / -	132 / 1	32 / -	71 / -	64 / -	95 / -	25 / -	- / 56	934
2914	20 / -		79 / 13	65 / -	41 / 3	47 / -	64 / 13	13 / -	72 / -	987
3257	20 / -	102 / -		66 / -	94 / -	45 / -	101 / 8	21 / -	86 / 10	1099
3356	24 / -	112 / -	431 / -		- / 48	72 / -	- / 97	14 / -	111 / -	1099
3549	106 / -	90 / -	147 / -	- / -		56 / -	145 / 5	8 / -	12 / 60	1034
6453	52 / -	143 / -	157 / -	68 / -	78 / -		- / 2	23 / -	- / 56	865
6939	65 / 1	84 / 14	62 / 33	- / 22	78 / 27	- / 23		20 / -	76 / 1	979
12702	35 / -	75 / -	69 / -	57 / -	36 / -	50 / -	5 / -		78 / -	n/a
30071	- / -	100 / -	106 / 23	78 / -	111 / 6	- / -	132 / 10	28 / -		1045

30071 getting upstream from 3257, but filtering routes to 1273/6453

[1] including 701

Peering matrix

Direct paths (“Source-AS Destination-AS”) / indirect paths (“Source-AS Transit-AS Destination-AS”)

Source	Destination									
	1273	2914	3257	3356	3549	6453	6939	12702 [1]	30071	Total
1273		104 / -	132 / 1	32 / -	71 / -	64 / -	95 / -	25 / -	- / 56	934
2914	20 / -		79 / 13	65 / -	41 / 3	47 / -	64 / 13	13 / -	72 / -	987
3257	20 / -	102 / -		66 / -	94 / -	45 / -	101 / 8	21 / -	86 / 10	1099
3356	24 / -	112 / -	431 / -		- / 48	72 / -	- / 97	14 / -	111 / -	1099
3549	106 / -	90 / -	147 / -	- / -		56 / -	145 / 5	8 / -	12 / 60	1034
6453	52 / -	143 / -	157 / -	68 / -	78 / -		- / 2	23 / -	- / 56	865
6939	65 / 1	84 / 14	62 / 33	- / 22	78 / 27	- / 23		20 / -	76 / 1	979
12702	35 / -	75 / -	69 / -	57 / -	36 / -	50 / -	5 / -		78 / -	n/a
30071	- / -	100 / -	106 / 23	78 / -	111 / 6	- / -	132 / 10	28 / -		1045

30071 getting upstream from 3257, but filtering routes to 1273/6453

3356 getting a fulltable from 3257, but no transit

[1] including 701

Peering matrix

Direct paths (“Source-AS Destination-AS”) / indirect paths (“Source-AS Transit-AS Destination-AS”)

Source	Destination									
	1273	2914	3257	3356	3549	6453	6939	12702 [1]	30071	Total
1273		104 / -	132 / 1	32 / -	71 / -	64 / -	95 / -	25 / -	- / 56	934
2914	20 / -		79 / 13	65 / -	41 / 3	47 / -	64 / 13	13 / -	72 / -	987
3257	20 / -	102 / -		66 / -	94 / -	45 / -	101 / 8	21 / -	86 / 10	1099
3356	24 / -	112 / -	431 / -		- / 48	72 / -	- / 97	14 / -	111 / -	1099
3549	106 / -	90 / -	147 / -	- / -		56 / -	145 / 5	8 / -	12 / 60	1034
6453	52 / -	143 / -	157 / -	68 / -	78 / -		- / 2	23 / -	- / 56	865
6939	65 / 1	84 / 14	62 / 33	- / 22	78 / 27	- / 23		20 / -	76 / 1	979
12702	35 / -	75 / -	69 / -	57 / -	36 / -	50 / -	5 / -		78 / -	n/a
30071	- / -	100 / -	106 / 23	78 / -	111 / 6	- / -	132 / 10	28 / -		1045

30071 getting upstream from 3257, but filtering routes to 1273/6453

3356 getting a fulltable from 3257, but no transit

peering issues at GBLX

[1] including 701

Peering matrix

Direct paths ("Source-AS Destination-AS") / indirect paths ("Source-AS Transit-AS Destination-AS")

Source	Destination									
	1273	2914	3257	3356	3549	6453	6939	12702 [1]	30071	Total
1273		104 / -	132 / 1	32 / -	71 / -	64 / -	95 / -	25 / -	- / 56	934
2914	20 / -		79 / 13	65 / -	41 / 3	47 / -	64 / 13	13 / -	72 / -	987
3257	20 / -	102 / -		66 / -	94 / -	45 / -	101 / 8	21 / -	86 / 10	1099
3356	24 / -	112 / -	431 / -		- / 48	72 / -	- / 97	14 / -	111 / -	1099
3549	106 / -	90 / -	147 / -	- / -		56 / -	145 / 5	8 / -	12 / 60	1034
6453	52 / -	143 / -	157 / -	68 / -	78 / -		- / 2	23 / -	- / 56	865
6939	65 / 1	84 / 14	62 / 33	- / 22	78 / 27	- / 23		20 / -	76 / 1	979
12702	35 / -	75 / -	69 / -	57 / -	36 / -	50 / -	5 / -		78 / -	n/a
30071	- / -	100 / -	106 / 23	78 / -	111 / 6	- / -	132 / 10	28 / -		1045

30071 getting upstream from 3257, but filtering routes to 1273/6453

3356 getting a fulltable from 3257, but no transit

peering issues at GBLX

no filler feeds

[1] including 701

- about 50 prefixes without any proper upstream exist (exact numbers are hard to tell)
- most common for NRENs in LACNIC and APNIC region and among old 6bone participants
- sometimes political reasons
- but mostly lazyness
- to have a more or less complete routingtable you need ...

Fillerfeeds

- some large ASNs (“Tier-1”) get an unfiltered fulltable from smaller networks to have a “transit-of-last-resort”
  - usually low BGP localpref, filters
  - no reasonable policy beyond this point
  - outbound announcements cannot be controlled
- some just don’t

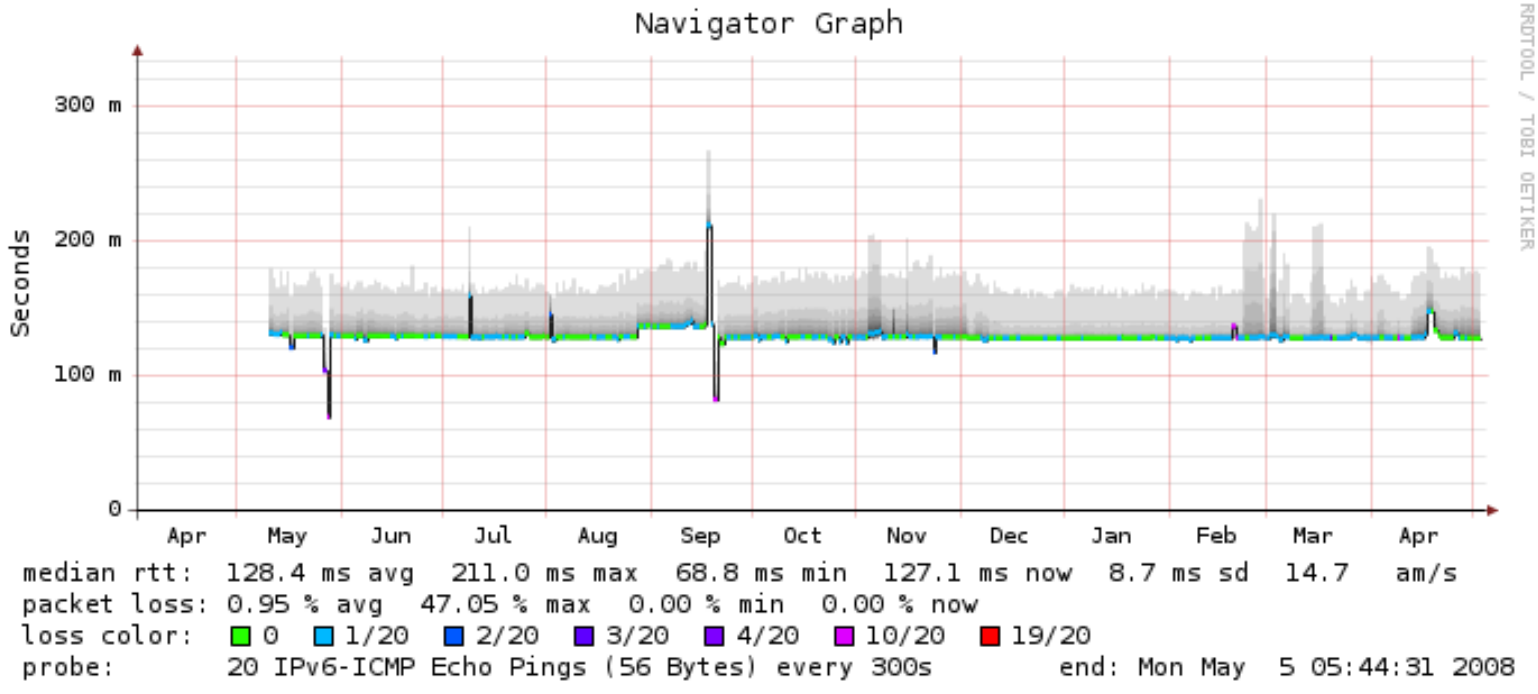
1273	none
2914	7660 (52 prefixes)
3257	2497 (82 prefixes, also peering) and 6175 (100 prefixes, also peering)
3356	3257 (one-way transit)
3549	18084 (94 prefixes)
6453	none
12702	unknown (possibly none)

Example: AS6509, 2001:410::/32, CA*net4

- CA*net4 (AS6509) peers with several research networks natively, fast and stable connectivity
- but, no commercial upstream for political reasons
- outbound connectivity generally through China

Peer	to 6509	6509 to peer
1273	no path	17579 23911 1273
2914	7660 22388 11537 6509	17579 23911 7660 2914
3257	2497 2500 7660 22388 11537	17579 23911 7660 2500 2497 3257
3356	3257 2497 2500 7660 22388 11537 6509	17579 23911 7660 2500 4725 3356
3549	18084 2500 7660 22388 11537 6509	17579 23911 7660 2500 4725 6939 3549
6175	17715 6435 278 18592 6509	17579 23911 7660 2500 2497 6175
6453	no path	17579 23911 7660 2500 4725 6175 6453
12702	no path	17579 23911 7660 2500 4725 701 12702
30071	6175 17715 6435 278 18592 6509	17579 23911 7660 7684 18084 3257 30071

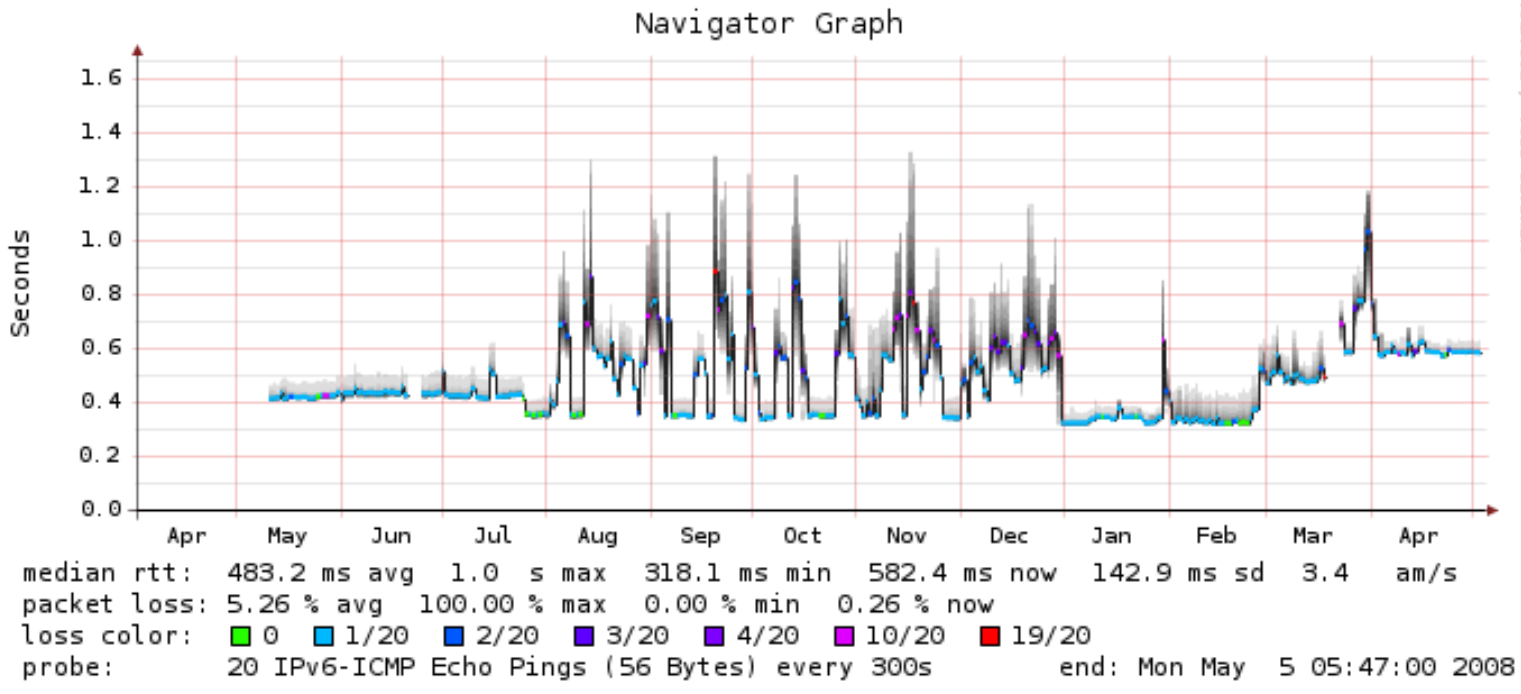
RTT measurement from AS12816 (NREN)



stable path

12816 680 20965 6509 – 6509 20965 680 12816

RTT measurement from AS29259 (commercial)



path mostly stable for the last couple of weeks

29259 286 6175 2497 2500 7660 22388 11537 6509 – 6509 17579 23911 1273 8767 29259

Throughput measurement

iperf to/from a static host inside CA*net4 (ryouko.imsb.nrc.ca) in single TCP stream mode and three parallel TCP stream mode (-P 3)

	send	receive	RTT
NREN	5.97 Mbit/s	61.6 Mbit/s	130ms
NREN (-P 3)	9.43 Mbit/s	44.5 Mbit/s	
commercial	3.16 Mbit/s	2.62 Mbit/s	330ms
commercial (-P 3)	4.75 Mbit/s	4.89 Mbit/s	
RIPE-56	56 kbit/s	unknown	800ms
RIPE-56 (-P 3)	6 kbit/s	unknown	10% PL

Paths taken during RIPE-56 monday plenary:
2121 286 6175 17715 6435 278 18592 6509 – 6509 17579 23911 1273 5539 2121

Something changed, latency and throughput from AS29259 through the roof starting Monday afternoon, forward path changed to the one used at RIPE-56 through Lavanet

Count-to-infinity

- BGP update propagation speed is limited, leads to paths getting longer on global BGP withdraw
- longer paths that are usually hidden by BGP best-path selection algorithm can be seen

start	680 8767 29259
0:00	680 3549 8767 29259
0:30	680 3549 1273 286 286 29259
1:00	WITHDRAWN

in current model every ASN in any visible path is transit for at least one of source/destination ASN (BGP policy, not economic relationship)

Count-to-infinity

- BGP update propagation speed is limited, leads to paths getting longer on global BGP withdraw
- longer paths that are usually hidden by BGP best-path selection algorithm can be seen

start	680 8767 29259
0:00	680 3549 8767 29259
0:30	680 3549 1273 286 286 29259
1:00	WITHDRAWN

in current model every ASN in any visible path is transit for at least one of source/destination ASN (BGP policy, not economic relationship)

1:30	680 20965 11537 22388 7660 2500 2497 3257 3549 8767 29259
2:00	680 20965 11537 17579 23911 7660 2500 2497 3257 3549 8767 29259
2:30	680 20965 11537 17579 23911 24489 24490 24490 9270 7660 2500 2497 3257 3549 8767 29259
3:00	680 20965 11537 17579 23911 24489 24490 24490 9270 7660 2500 4725 17715 6435 6939 2497 3257 3549 8767 29259
3:30	WITHDRAWN

bad/bogus paths can stay there for a long time due to *BGP ghosting*

small set of ASNs visible in most bad paths

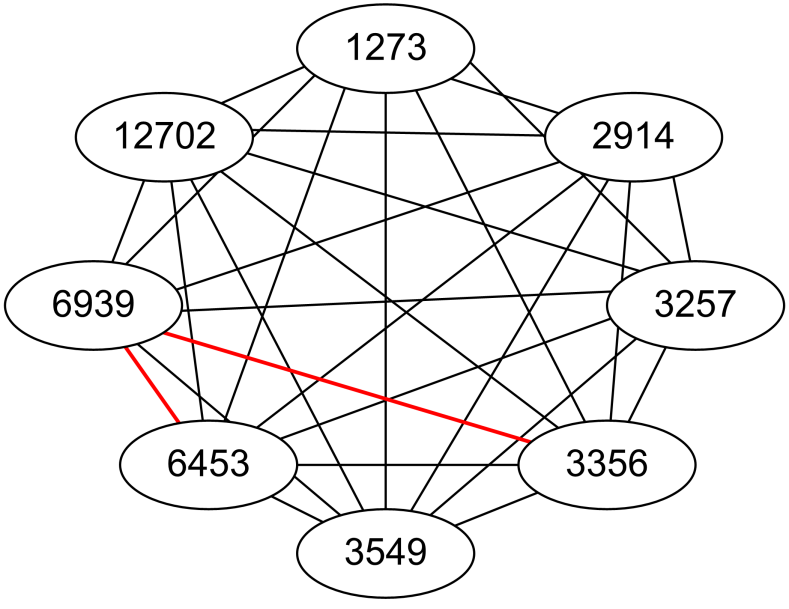
- 278 – UNAM-MX
- 2500 – WIDE
- 4725 – SOFTBANK
- 6435 – Lavanet
- 7660 – APAN-JP
- 17715 – ChungHwa Telecom

Does anyone have responsive contacts?

some progress ...

- 3549 has depreffed 18084 and applied filters to 6175 – thanks to Steve Powell
- 2914 has applied filters to 2500, looking at 7660 now – thanks to Brad Dreisbach
- 6939 has dropped fulltable exports to 6bone ASNs 4555 and 278 – thanks to Mike Leber

Proposal



- almost full mesh between major ASNs - only two missing peerings
- can be seen as “Tier-1” in the IPv6 world - drop all 6bone connections?
- very dangerous term, peering wars are sure to start
- interim idea: everyone filter their big peers from their filler feeds
- stops bad paths from being generated and eventually ghost on prefix withdrawal
- but might not put enough pressure on AS without transit

- ipv6-ops mailing list – `http://lists.cluenet.de/mailman/listinfo/ipv6-ops`
- SixXS GRH – `http://grh.sixxs.net`
- IPv6-enabled webproxy (Squid 3-HEAD) – proxytest.mucip.net:3128