

**Case Study: IPv6 Packet Loss to CH ccTLD Name Server**

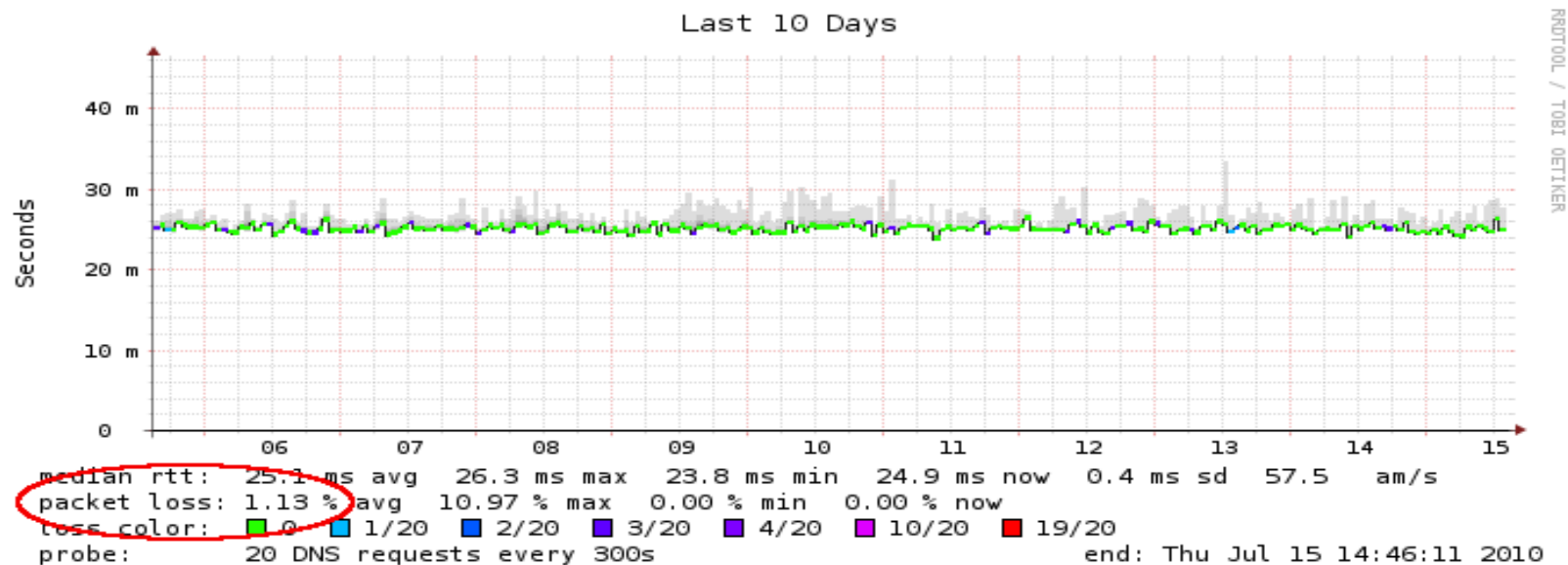
SWITCH has outsourced one of their secondary name servers for the CH ccTLD to Netnod, the Swedish company that also runs i.root-servers.net. The zone is served from 2001:67c:1010::/48 using anycast.

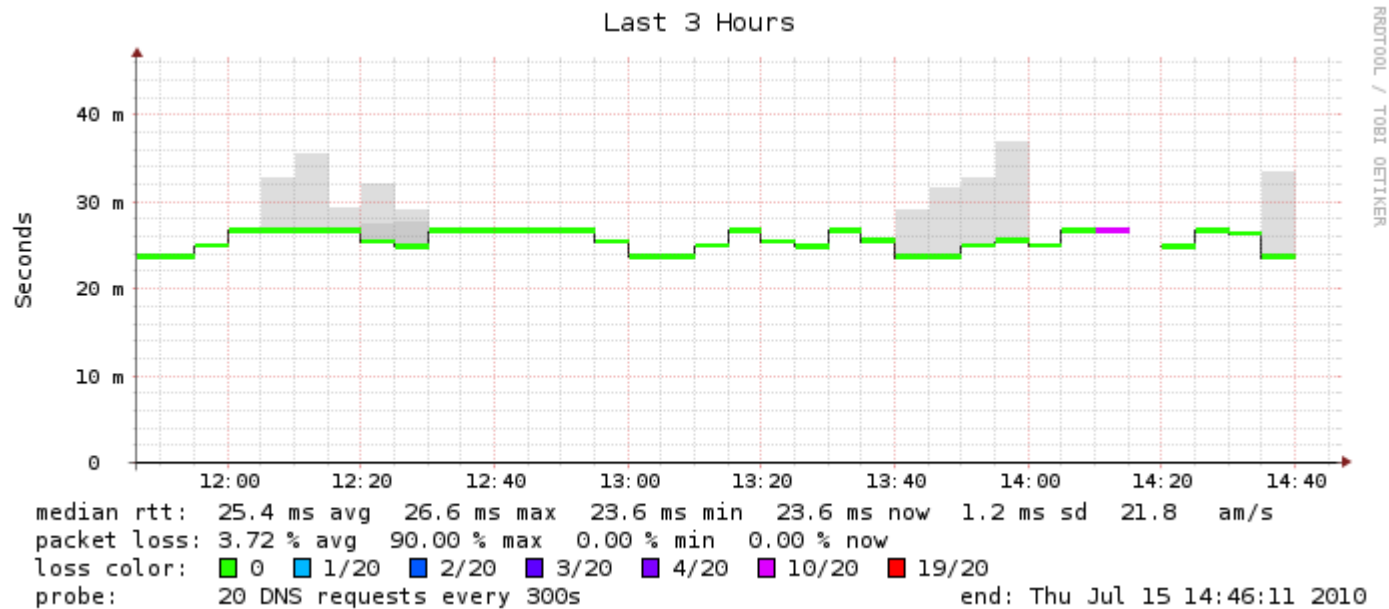
In July 2010, only the sites in Stockholm and London (LINX) were IPv6-enabled, but Stockholm announced a shorter prefix → all traffic was attracted by the LINX site.

# Symptom: Packet Loss



SmokePing probe using DNS queries:





- Tri-modal RTT
  - Either no loss or heavy/total loss
- What could a reasonable next step be?

# First Step: Check with ICMP



Test with `ping` (ICMP) shows neither loss nor staggered RTT.

**Hypothesis:** problem at or near the DNS server, only affecting DNS traffic (and IPv4 only).

Opened a ticket with Netnod NOC (July 9). They reported

- No load-balancer or other devices in front of server
- IPv4/IPv6 served on dual-stack host
- Performed local test and did not observe loss

# Broaden the Scope of Investigation



Report from Netnod contradicts hypothesis, we need to consider other possibilities after all.

Look at network path from SWITCH to the server

traceroute to ch1.dnsnode.net (2001:67c:1010:2::53), 30 hops max, 40 byte packets

- 1 swiCE3-V300.switch.ch (2001:620:0:113::1) 0.392 ms
- 2 swiZH2-10GE-1-1.switch.ch (2001:620:0:c027::2) 4.106 ms
- 3 swiLX1-10GE-1-3.switch.ch (2001:620:0:c015::1) 4.193 ms
- 4 swiLX2-10GE-4-4.switch.ch (2001:620:0:c008::2) 4.157 ms
- 5 swissix-glb.init7.net (2001:7f8:24::7) 14.682 ms
- 6 r1fra1.core.init7.net (2001:1620:2::6) 10.825 ms
- 7 ge-4-0-4-295.fra21.ip6.tinet.net (2001:668:0:3::2000:111) 11.424 ms
- 8 xe-2-0-0.lon21.ip6.tinet.net (2001:668:0:2::131) 21.792 ms
- 9 xe-11-0-0.lon20.ip6.tinet.net (2001:668:0:2::1:1b02) 21.787 ms
- 10 xe-11-0-0.lon10.ip6.tinet.net (2001:668:0:2::1:1051) 21.779 ms
- 11 g0-0-123.tr2.tfm7.thn.linx.net (2001:668:0:3::4000:82) 26.275 ms
- 12 2a01:40:1003:2::3 (2a01:40:1003:2::3) 26.232 ms
- 13 ch1.dnsnode.net (2001:67c:1010:2::53) 23.315 ms

Apart from SWITCH and Netnod, packets transit three autonomous systems

- Init7 (AS13030): Swiss ISP (Zurich → Frankfurt)
- Tinet (AS3257): Global carrier, formerly Tiscali, now owned by Neutral Tandem (Frankfurt → London)
- LINX (AS5359): London Internet Exchange, provides IPv6 transit to Netnod

While doing traceroutes, I noticed that the path within AS3257 changed with each trace.

## Linux traceroute has some cool features

```
: gall@atitlan[gall]; traceroute -V
```

Modern traceroute for Linux, version 2.0.11, Jun 20 2008

Copyright (c) 2006 Dmitry Butskoy, License: GPL v2 or any later

```
: gall@atitlan[gall]; traceroute6 -s 2001:620:0:113:21b:78ff:fe30:297e --sport=1234 -U -q1 -f 8 -m 9 -N1 ch1.dnsnode.net
```

traceroute to ch1.dnsnode.net (2001:67c:1010:2::53), 10 hops max, 40 byte packets

```
8 xe-7-3-0.lon20.ip6.tinet.net (2001:668:0:2::1:1242) 24.759 ms
```

```
9 xe-0-3-0.lon21.ip6.tinet.net (2001:668:0:2::1:1ae2) 24.784 ms
```

**-U use UDP**

**--sport set source UDP port**

**-q1 only send out one query per hop**

**-f 8 initial TTL**

**-m 9 maximum TTL**

**-N 1 don't send probes in parallel**

## Hops #8 and #9 vary with the source port

```
: gall@atitlan[gall]; traceroute6 -s 2001:620:0:113:21b:78ff:fe30:297e --sport=1234 -U -q1 -f 8 -m 9 -N1 ch1.dnsnode.net
traceroute to ch1.dnsnode.net (2001:67c:1010:2::53), 10 hops max, 40 byte packets
 8 xe-7-3-0.lon20.ip6.tinet.net (2001:668:0:2::1:1242) 24.759 ms
 9 xe-0-3-0.lon21.ip6.tinet.net (2001:668:0:2::1:1ae2) 24.784 ms
: gall@atitlan[gall]; traceroute6 -s 2001:620:0:113:21b:78ff:fe30:297e --sport=1235 -U -q1 -f 8 -m 9 -N1 ch1.dnsnode.net
traceroute to ch1.dnsnode.net (2001:67c:1010:2::53), 10 hops max, 40 byte packets
 8 xe-9-1-0.lon20.ip6.tinet.net (2001:668:0:2::1:1671) 21.842 ms
 9 lon20.ip6.tinet.net (2001:668:0:2::1:1b12) 21.787 ms
: gall@atitlan[gall]; traceroute6 -s 2001:620:0:113:21b:78ff:fe30:297e --sport=1236 -U -q1 -f 8 -m 9 -N1 ch1.dnsnode.net
traceroute to ch1.dnsnode.net (2001:67c:1010:2::53), 10 hops max, 40 byte packets
 8 xe-9-1-0.lon20.ip6.tinet.net (2001:668:0:2::1:1671) 21.799 ms
 9 lon20.ip6.tinet.net (2001:668:0:2::1:1b12) 21.790 ms
: gall@atitlan[gall]; traceroute6 -s 2001:620:0:113:21b:78ff:fe30:297e --sport=1237 -U -q1 -f 8 -m 9 -N1 ch1.dnsnode.net
traceroute to ch1.dnsnode.net (2001:67c:1010:2::53), 10 hops max, 40 byte packets
 8 xe-2-0-0.lon21.ip6.tinet.net (2001:668:0:2::131) 21.759 ms
 9 xe-11-0-0.lon20.ip6.tinet.net (2001:668:0:2::1:1b02) 21.741 ms
```

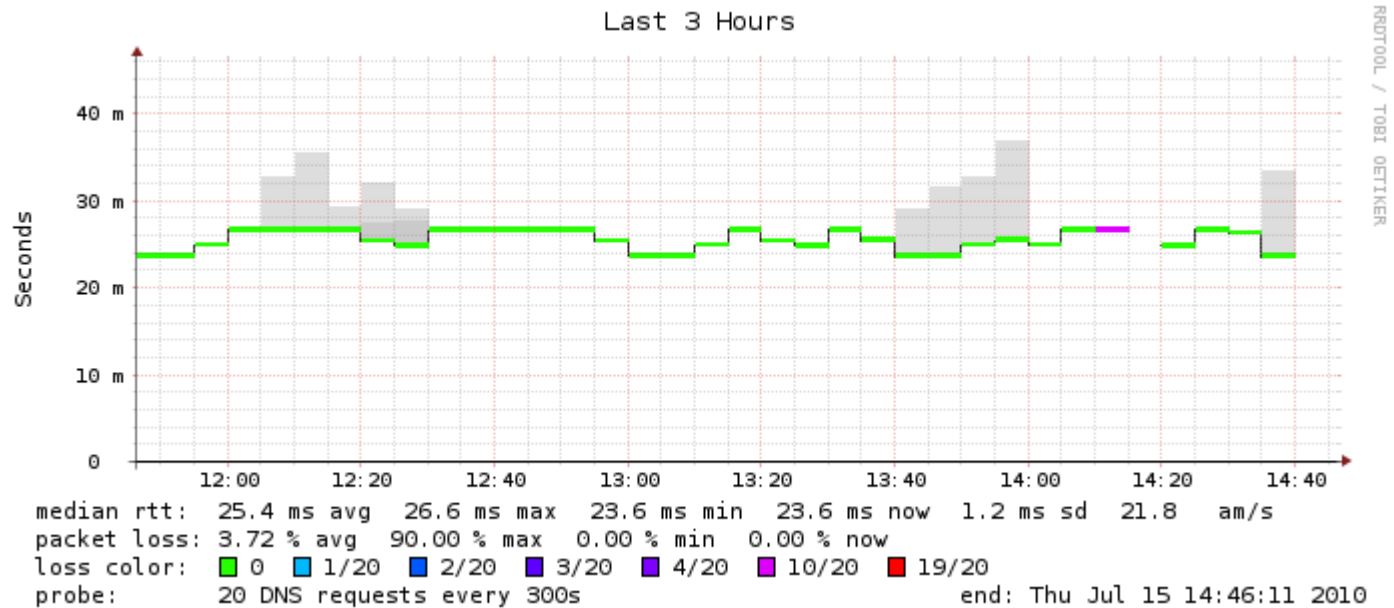
What could cause this?

More than one path with same “distance” (metric) to a given destination exist in the routing table. Without multipathing, some tie-breaker is used to select a single best path.

With multipathing, all paths are used. How are the packets spread across the paths?

- Per-packet: use next path for each packet
- Per flow: map flows to paths (hash), keep packets of one flow on same path. “Flow” depends on implementation/configuration
  - Source/Destination IP only
  - 5-tuple Source IP/Destination IP/Protocol/Source Port/Destination Port
  - ...

## Could multipathing explain one of the anomalies?



## Variance of RTT

**Hypothesis:** there are 3 or 4 equal-cost paths within AS3257 of substantially different length (3ms ~ 600km optical fiber). SmokePing probes select one of them with random source port. ICMP is not affected, because all packets are mapped to the same path.

High jitter of DNS traffic is a little annoying but harmless.

**But what about the loss?**

# Path Investigation (6): BINGO!



**Hypothesis:** one of the alternate paths is lossy, so loss depends on the (random) DNS query port.

Can be easily checked with dig by fixing the source port

```
:gall@atitlan[gall]; dig -b2001:620:0:114:21b:78ff:fe30:2974#1234
```

This hypothesis turned out to be wrong. BUT: when going through all source ports from 1025 to 65535, I discovered that *all* queries with source port in the range 35072 – 35327 were dropped! Also: same thing for TCP.

How can we find out where the drops are happening?

- Forward or return path?
- In which AS?

# Narrowing Down the Scope (1)



Distinguishing between forward and return path is difficult

- Looking glass in remote AS
- Access to remote site
- Source-routing (sadly mostly unusable due to security)

We asked Netnod to check on their server: packets in question do not arrive there → **Loss occurs on forward path.**

Can traceroute help us find out more?

# Narrowing Down the Scope (2)



```
: gall@atitlan[gall]; traceroute6 -s 2001:620:0:113:21b:78ff:fe30:297e --sport=35072 -q1 -N1 -w1 ch1.dnsnode.net
```

```
traceroute to ch1.dnsnode.net (2001:67c:1010:2::53), 30 hops max, 40 byte packets
```

```
 1 swiCE3-V300.switch.ch (2001:620:0:113::1) 0.301 ms
 2 swiZH2-10GE-1-1.switch.ch (2001:620:0:c027::2) 4.144 ms
 3 swiIX1-10GE-1-3.switch.ch (2001:620:0:c015::1) 4.312 ms
 4 swiIX2-10GE-4-4.switch.ch (2001:620:0:c008::2) 4.178 ms
 5 swissix-glb.init7.net (2001:7f8:24::7) 4.151 ms
 6 r1fra1.core.init7.net (2001:1620:2::6) 18.958 ms
 7 *
 8 *
 9 *^C
```

OK, time to open a case with Tinet (July 12).

- July 13: “You should see an improvement”, but nothing has changed.
- Silence
- July 20: “Engineers might have found a solution for this and are double checking before implementing it”.
- July 26: a new problem occurs, suddenly ICMP probes show loss as well
  - Quickly solved by cleaning a bad fiber patch
  - Multipathing is changed to prefer a single path
  - Some router is equipped with experimental fix by the vendor, **packets are no longer dropped**
- August 9: Tinet reports problem permanently fixed (after inquiry by me)

During the investigation, I made a wild guess what the problem could be

- 35072 – 35327 is 0x8900 – 0x89ff in hex
- 0x89 = 137, hm....
- Port 137 is the well-known NetBIOS port, blocked almost everywhere
- My guess: an ACL was supposed to filter destination port 137 at the border of Tinet but did in fact filter on the high-order byte of the source port

I asked, but they never told me what the problem really was.