IPv6 Testing @ QMUL

Terry Froy
<t.froy@qmul.ac.uk>

Daniel Traynor
<d.traynor@qmul.ac.uk>

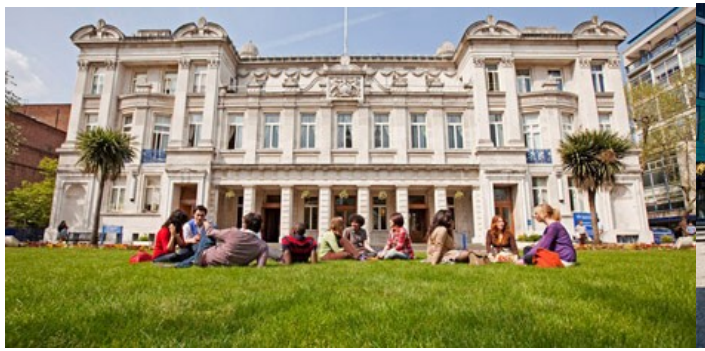
School of Physics and Astronomy
Queen Mary University of London

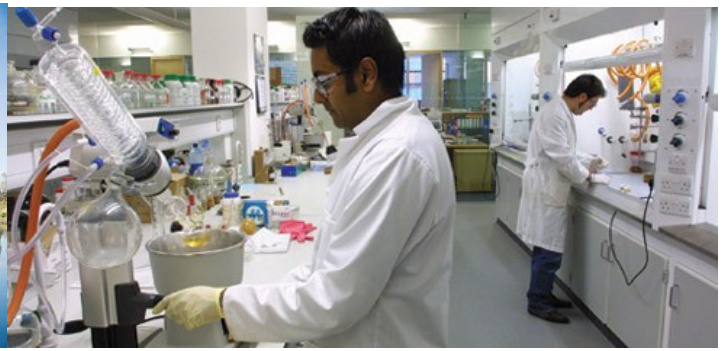
GridPP 37, Ambleside, September 2016

Queen Mary University of London

- Research-focused higher education institution.
- Four main campuses in London.
- 21 academic departments.
- 20,000+ post- and undergraduate students.
- 4,000 staff.

<http://www.qmul.ac.uk>





GridPP
UK Computing for Particle Physics

Queen Mary
University of London

State of QMUL IPv6 Deployment @ GridPP 36

- All front-end services (CEs, SEs, XRootD, etc) are dual-stacked with AAAA records published in DNS.
- Public-facing DNS still not accessible via IPv6 transport.
- All front-end services at QMUL are also operating with an MTU of 9000.
- All worker nodes are **currently** ignorant of IPv6.

State of QMUL IPv6 Deployment @ GridPP 37

- We are operating a small number of IPv6 worker nodes (WNs) on our production queues.
- 50% of these WNs are dual-stacked in the usual manner.
- 50% of these WNs are dual-stacked (we still require Legacy IP due to Lustre still having no IPv6 support) but these WNs are configured ‘a bit special’.

Regular Dual Stack Worker Nodes

- They Just Work™
- A useful step on the IPv6 journey.
- Dual stack adds complexity:
  - Two routing tables per host
  - Two firewalls per host
  - Legacy IP and IPv6 addressing for each host
- What comes after dual stack ?
  - Legacy IP is considered DEPRECATED by the IETF

‘Special’ Dual Stack Worker Nodes

- We don’t have a Legacy IP defaultroute.

```
[root@cn639 ~]# ip route ls 0.0.0.0/0
[root@cn639 ~]#
```

- … so how do we reach Legacy IP-only hosts ?
- How does a regular host do it ?
- DNS lookups first...

```
cybernoid:~ tez$ host -t a www.lancs.ac.uk
www.lancs.ac.uk has address 148.88.2.80
cybernoid:~ tez$ host -t aaaa www.lancs.ac.uk
www.lancs.ac.uk has no AAAA record
```

‘Special’ Dual Stack Worker Nodes

- www.lancs.ac.uk has a Legacy IP address but no IPv6 address.
- A regular dual-stack host would be forced to use Legacy IP.
- So, how does this work ?

```
[root@cn639 ~]# ping6 -c 4 www.lancs.ac.uk
PING www.lancs.ac.uk(www-ha.lancs.ac.uk) 56 data bytes
64 bytes from www-ha.lancs.ac.uk: icmp_seq=1 ttl=53 time=10.2 ms
64 bytes from www-ha.lancs.ac.uk: icmp_seq=2 ttl=53 time=16.8 ms
64 bytes from www-ha.lancs.ac.uk: icmp_seq=3 ttl=53 time=10.3 ms
64 bytes from www-ha.lancs.ac.uk: icmp_seq=4 ttl=53 time=10.2 ms

--- www.lancs.ac.uk ping statistics ---
4 packets transmitted, 4 received, 0% packet loss, time 3014ms
rtt min/avg/max/mdev = 10.227/11.903/16.802/2.829 ms
[root@cn639 ~]#
```

DNS64 [RFC6147] – How It Works(tm)

- The nameservers for lancs.ac.uk will only return an A record for www.lancs.ac.uk - they don't serve a AAAA record.
- Our 'special' worker nodes query DNS64-enabled resolvers (powered by the awesome open-source PowerDNS Recursor 4.x).
- A regular DNS resolver merely relays questions to servers and answers to clients (optionally caching the answers in the process).
- A DNS64-enabled resolver behaves slightly differently:
  - Client asks DNS64-enabled resolver for AAAA record for www.lancs.ac.uk.
  - DNS64-enabled resolver asks lancs.ac.uk nameserver which says 'No AAAA record for www.lancs.ac.uk'.
  - DNS64-enabled resolver asks for A record instead.
  - lancs.ac.uk nameserver responds with 148.88.2.80.

DNS64 [RFC6147] – Base 10 vs. Base 16

- Client asked for an AAAA record though.
- DNS64-enabled resolver ‘fixes’ this by converting the Legacy IP address 148.88.2.80 contained in the A record into hexadecimal:

DEC	148	88	2	80
HEX	94	58	2	50

- The DNS64-enabled resolver ‘synthesizes’ a AAAA record by appending this 32-bit hexadecimal representation of the Legacy IP address to a /96 IPv6 prefix:
  - 64:ff9b::[/96] Well-Known NAT64 Prefix [RFC6052]
  - ::9458:250 www.lancs.ac.uk
- The DNS64-enabled resolver returns an AAAA record of 64:ff9b::9458:250 to the client.

NAT64 [RFC6164] – The Next Step

- All DNS64 does is ‘spoof’ the IPv6 availability of Legacy IP-only end hosts.
- The real magic happens a bit further down the traffic path…

```
[root@cn639 ~]# ip -6 route ls | grep via

64:ff9b::/96 via 2a01:56c0:4033::6464 dev eth1  metric 1024
mtu 9000 hoplimit 4294967295

default via fe80::5:73ff:fea0:34b dev eth1  proto kernel
metric 1024  expires 1447sec mtu 9000 hoplimit 64

[root@cn639 ~]#
```

NAT64 [RFC6164] – How It Works(tm)

- We run a NAT64 implementation [JooL] on CentOS 7.2.
- The Well-Known NAT64 Prefix 64:ff9b::/96 is routed towards it; so, as in our previous example, an ICMPv6 packet is sent towards www.lancs.ac.uk.
- NAT64 sees IPv6 packet addressed to 64:ff9b::9458:250.
- An available Legacy IP address/source port (if applicable) is assigned for the ‘source’ Legacy IP information in the translated Legacy IP traffic.
- The NAT64 implementation only cares about the last 32-bits of the destination IPv6 address; it converts that back into a Legacy IP address and then uses that as the destination Legacy IP address – the original destination port (if applicable) is carried over from the original IPv6 ‘conversation’.

NAT64 [RFC6164] – Summary

- What does this provide us ?
  - Single-stack IPv6 across 95%+ of our cluster.
  - Efficient use of our scarce Legacy IP resources.
  - Complexities of dual-stack are constrained to a small number of cluster machines.
- What does this mean for IPv6-only WNs operated by others ?
  - WNs are able to freely communicate with Legacy IP-only resources (like CVMFS).
  - Sites do not need to run their own DNS64/NAT64 – ideally, the upstream NREN (i.e. JANET) should run such a service from their PoPs (Point-of-Presence).

NAT64 [RFC6164] – IP Literals

- What doesn’t work ?
  - IP literals [i.e. http://192.0.2.1/]
  - Plain ol’ FTP [Jool lacks support for RFC6384]
- Does this matter ?
  - No job failures have been observed that can be attributed to use of IP literals.
  - For everything else, there is 464XLAT [RFC6877].

What Next For QMUL ?

- NAT64 works…
  - Deploy to all other production worker nodes.
- Further testing of 464XLAT
  - Solves the IP literal use-case but adds extra layer of complexity to each WN.
- Experimentation with ‘optimal’ jumbo MTUs
  - Improving end-to-end performance for both Legacy IP and IPv6.
  - Inspired by in-depth discussion while drinking with Dr. Chris Walker (who sends his regards to everybody here!)

Relevant RFCs

- NAT64 [RFC6164]
- 464XLAT [RFC6877]
- IPv6 Path MTU Discovery [RFC1981]
- DNS64 [RFC6147]
- Discovery of the IPv6 Prefix Used for IPv6 Address Synthesis [RFC7050]
- Scenarios and Analysis for Introducing IPv6 into ISP Networks [RFC4029]
- An FTP Application Layer Gateway (ALG) for IPv6-to-IPv4 Translation [RFC6384]

Special Acknowledgements

- PowerDNS.COM BV [http://www.powerdns.com/] for their awesome PowerDNS Recursor software [http://www.powerdns.com/recursor.html]
- ITESM [http://www.itesm.mx/] and NIC Mexico [http://www.nicmexico.mx/] for their work on JooL [http://www.jool.mx/]

Questions ?