An IPv6 Update

Tim Chown, Jisc (tim.chown@jisc.ac.uk)

HEPSYSMAN, University of Oxford, 22 June 2023

Agenda

An update on IPv6

- I gave a tutorial on IPv6 at HEPSYSMAN back in 2017
- But what's important to consider today?
  - Reasons to deploy (if you haven't yet…☺)
  - IPv6 addressing and configuration
  - IPv6 thinking: operational, deployment and security perspectives
  - New protocol developments (spoiler: actually very few)

  - Jumbo frames (yes, this is IPv4 too!)

Jisc

A quick Janet reminder

Jisc manages and operates Janet for the UK R&E community

The network has supported IPv6 natively for 20 years

All our external peerings are now dual stack (IPv4 and IPv6)

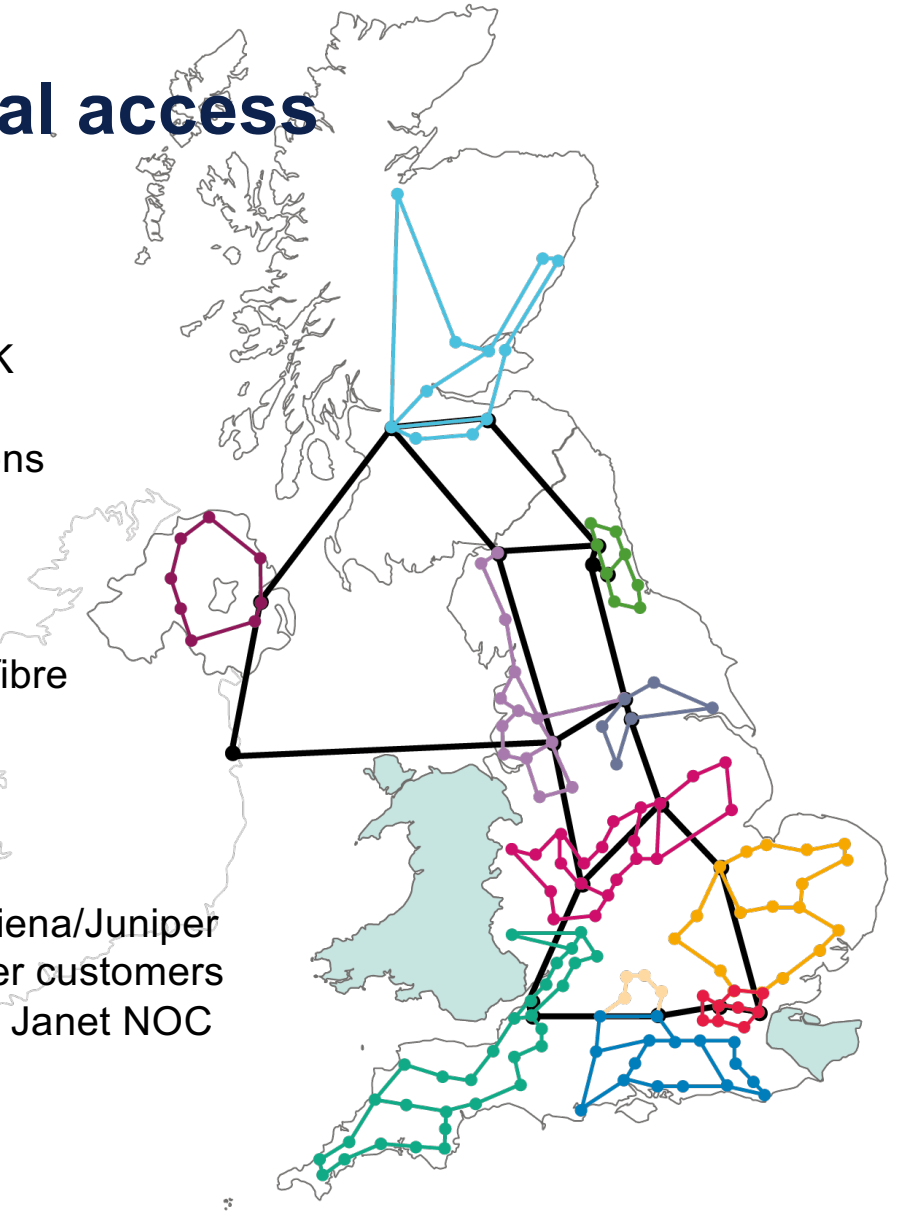
Janet backbone and regional access infrastructure

- Janet backbone
- Scotland
- North West
- Yorkshire
- Northern Ireland
- North East
- Midlands
- East
- South West
- Thames
- South
- London
- Public sector networks

Jisc is the ISP for UK HE/FE, and many research organisations like STFC

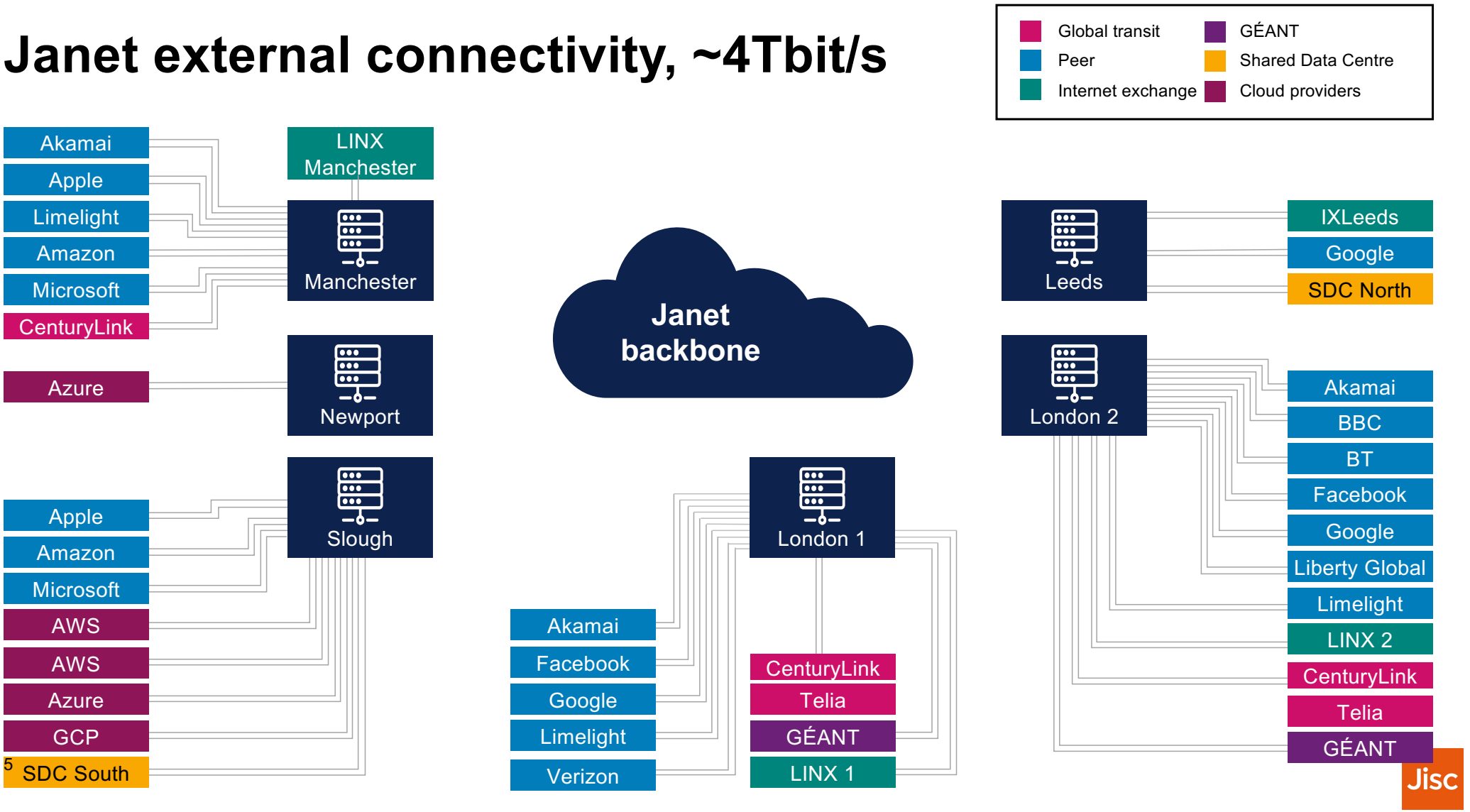
800G in main core
Around 9,000km of fibre
~1,000 customers
~1,500 connections

Network is largely Ciena/Juniper
~430 managed router customers
~700 devices run by Janet NOC



Jisc

Janet external connectivity, ~4Tbit/s



Rationale for IPv6?

Why deploying IPv6 is important

It's primarily about address space

- Sounds obvious, but worth emphasising
- The key difference is **128-bit addressing**
  - Enough globally unique address space to support future growth and innovation
- Ensures you can uniquely address all devices in your infrastructure
  - e.g., GridPP wants to directly address all worker nodes
- There's no new, unused IPv4 available; Jisc has no significant reserves
  - For more IPv4 addresses you need to go to the IPv4 broker market
  - Currently $40-$50 per address, see https://ipv4.global/
  - Or about $3M for a /16 of IPv4

Jisc

And other reasons still apply

Including…

- Supporting teaching and research
- Ensuring robust access to your public-facing services for IPv6-only client devices
- Removing NAT and private addressing complexity from network operations and management
- Minimising dependency on an ever-more fragile IPv4 network (witness CG NAT, etc)
- Security in an ‘IPv4 only’ network; IPv6 is supported on common platforms and on by default
- Enabling innovation at the edge
- Scalability for campuses of the future; IoT is increasingly using IPv6 (e.g., Matter)
- Being ready for IPv6-only applications and communities (the WLCG direction of travel)
- New capabilities in IPv6 such as enhanced packet marking
- …

Jisc

Why not deploy?

Do you have specific concerns?

- We can discuss later, or feel free to ask now ☺
- It's quite useful to look at other concerns, whether real or FUD, at https://ipv6bingo.com/

Jisc

Important recent developments

Many positive examples

- These help counter the reasons not to do it

- US government issued new mandate to deploy IPv6-only by 2025
  - OMB-21-07 - important because vendors want the US government business
  - https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-07.pdf

- Much improved support in cloud and container platforms, e.g.:
  - Kubernetes support – see https://kubernetes.io/docs/concepts/services-networking/dual-stack/ (v1.20 onwards)
  - AWS – see https://www.ipv6.org.uk/2022/10/13/ipv6-council-annual-meeting-2022/

Jisc

Deployment status and measurement

How much of the total network traffic is IPv6?

- There are various measures out there for IPv6 deployment, including summaries at:
  - http://www.worldipv6launch.org/measurements/
  - https://labs.ripe.net/Members/mirjam/content-ipv6-measurement-compilation
- Overall, Internet user traffic is around 40-45% IPv6, with the UK similar
  - We will hit 50% by the end of 2025 if linear growth continues
- IPv6 adoption varies by sector: residential, mobile, enterprise
- Janet IPv6 traffic sits lower at around 10-15%
  - BUT the HEP sites are a large part of that (at least on the GridPP systems)
- Generally, R&E is well behind commercial adoption
  - Probably due to no perceived **NEED** for IPv6, today

Jisc

IPv6 at Jisc

We are having a renewed push!

- Janet has been dual-stack for around 20 years
- Most of our network and security-related services support IPv6
  - NTP, DNS, eduroam peerings, etc
- Providing advice and guidance for members
  - “Eating our own dogfood”
- We are having a renewed push internally to ensure IPv6 is supported in our broader Jisc service portfolio
  - Setting IPv6 in procurement via our ITQ template
  - Including IPv6 in our PLM service transition gate
  - So in principle all new services will support IPv6

Jisc

IPv6 addressing

Things to consider

IPv6 Addressing

What changes with IPv6?

- IPv6 addresses are 128 bits
- The notation used is eight sets of four hexadecimal charaacters:
  - e.g., **2001:db8:e380:d0:920a:2380:b230:1106**
- IPv6 also has different scopes of addresses
  - Link-local (under fe80::/10), Unique Local Addresses (ULAs, under fc00::/7), and globally unique addresses (GUAs)
  - **Multi-addressing is the norm** – more later
- IPv6 address allocation and assignment policies are determined by the Regional Internet Registries (RIRs), and their members, which for us is RIPE
  - See https://www.ripe.net/publications/docs/ripe-738

Jisc

Jisc as your ISP

Connecting to Janet with IPv6

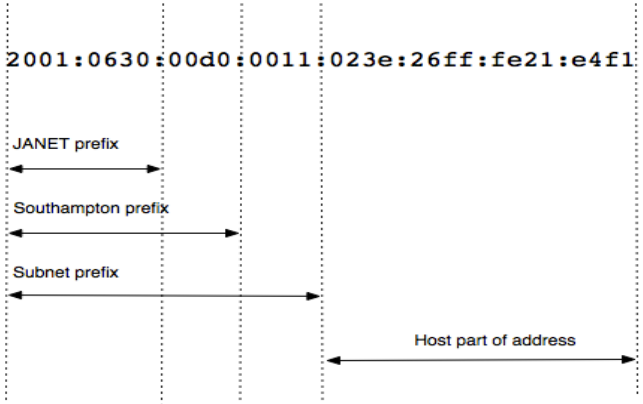
- The Janet backbone is dual-stack throughput
- IPv6 is available natively as part of the Janet IP Connection service
  - You just need to ask the Jisc Service Desk to turn IPv6 on
  - https://www.jisc.ac.uk/janet-ip-connection
- Jisc as an ISP is also a Local Internet Registry (LIR)
- As an LIR, Jisc obtained the prefix 2001:630::/32 from the RIPE NCC in 1999 (was a /35 back then, now the default is /32 with a /29 reserved)
  - Jisc can provide you an IPv6 prefix for your organisation on request
  - **Default** prefix assignment to a Janet-connected site is a **/48**

Jisc

Address assignments

From LIRs to end sites

- A typical prefix breakdown for a university site might be:



- NB: the ISP and site prefix lengths here are just defaults
- A Janet site may thus have up to 2^16 host subnets
- Host subnets are 64 bits due to the way autoconfiguration was defined

Jisc

What if I want more than a /48?

Getting bigger prefixes

- You can work with Jisc to document any rationale or need for a larger address space from our allocation, and Jisc can provide advice on that (we have to justify the use to the RIPE NCC)
- Or you can apply for LIR status yourself, as a small number of universities have done (QMUL for example) and get a /32 that way. The cost is currently low thousands of Euros each year.
- There is some internal discussion at Jisc and with the RIPE NCC as to whether we might assign larger blocks by default

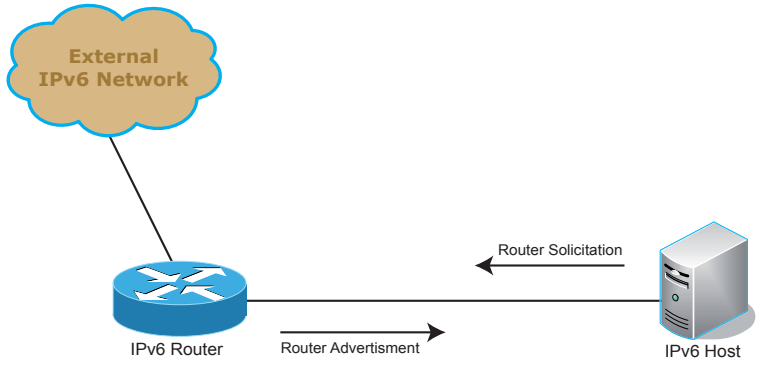
- **Jisc will route your traffic whichever approach you choose**

Jisc

Configuring addresses

Manual, SLAAC or DHCPv6

- You can do manual or automated configuration as per IPv4 if you wish
- IPv6 introduces **stateless address autoconfiguration (SLAAC)**
  - Address is then built from the prefix in Router Advertisements (RAs, 64 bits) and a host identifier (64 bits, ideally using RFC 7217)
  - SLAAC implicitly means all host subnets are /64 in size
  - RFC 8106 allows DNS configuration via RAs
- Or you can use DHCPv6 (RFC 8415)
  - Works quite similarly to IPv4
  - BUT not supported by Google on Android
- Most campuses are using SLAAC
- SLAAC includes privacy addresses; more addresses!



Jisc

Address planning

How you use your IPv6 assignment

- A similar process to IPv4
- Subnetting will typically follow geographic or administrative boundaries
- Lots of ‘clever’ things you can do because the bits are there
  - Nice guide written by SURFnet over 10 years ago
  - https://www.ripe.net/support/training/material/IPv6-for-LIRs-Training-Course/Preparing-an-IPv6-Addressing-Plan.pdf
- As mentioned before, all IPv6 host subnets are /64 due to SLAAC, whether you use SLAAC or DHCPv6
- You will probably have congruent IPv4 and IPv6 prefixes
  - The rationale for subnetting is the same either way
  - Your IPv4 subnets may vary in size over time (/24, /27, /22,…)

Jisc

Address planning (2)

Other considerations

- How big do you wish to make your layer 2 subnets?
  - If you go very big (esp. a flat layer 2 campus!) then there will be a lot of 'chatter' traffic from IPv4 ARP, IPv6 ND, etc
  - At the other extreme, RFC 8273 describes using a prefix per host for IPv6
- Consider whether all devices in a subnet will be configured the same way
  - RAs provide the same information to all hosts
- Will you use ULAs?
  - It may be tempting for systems you know do not communicate off site
  - But I don't know of any Janet sites doing so (not that ULAs would be seen…)
- Use /127 for point-to-point as per RFC 6164 (but you can reserve a /64)

IPv6 thinking

How do you need to adjust your thinking?

What makes IPv6 different?

- The larger addresses and their format
  - Need to consider all the places literal addresses appear – see RFC 5952
- Multi-addressing is normal
  - IPv6 LL, IPv6 GUA, IPv4 (private or global), IPv6 privacy addresses
  - Therefore address accountability is different
  - Need to consider address selection – defined in RFC 6724 (prefer IPv6, prefer matching scope, longest prefix match, …)
- Fragmentation is only done by end hosts, not along the path
  - Thus need PMTUD to work; RFC 4890 advises on ICMPv6 filtering

A summary

	IPv4	IPv6
Address length	32 bits	128 bits
Prefix length	Varies, typically /24	Always /64 in host subnets
Address configuration	DHCPv4	Stateless Autoconfiguration DHCPv6
Addresses used	Private **or** Global	Link-local **and** Global
Address resolution	ARP	Neighbour Solicitation / Advertisement
Host Path MTU Discovery	Optional	Required
Fragmentation	By hosts or routers	Only by hosts
Private addressing	RFC 1918	Unique Local Addresses (ULAs) (not designed for use with NAT)

Operational considerations

Examples

- What level of host/user accountability do you need?
  - Will you need new tools to handle the impact of multi-addressing?
- What about NAT?
  - Do you use it for IPv4? If so, for what reason?
  - It's possible in IPv6 (at least as NPTv6) but not recommended
- Do you want to run jumbo frames, usually 9000 MTU?
  - Janet supports it, and there is a clear performance advantage
  - If so, it's needed end-to-end, and you'll need to ensure PMTUD works
  - (More on this in a bit)

Dual-stack considerations

Running both protocols

- The common deployment model for campuses is dual stack
- Can then use IPv4 to IPv4-only systems, and IPv6 to IPv6-only
- Application can pick either protocol for a dual stack destination
- There's "Happy Eyeballs" for browsers (RFC 8305, was 6555)
  - Try IPv4 and IPv6 in parallel, pick the one that connects first
- Other applications should failover quickly if one protocol fails
- Application protocol selection has proven a challenge for GridPP
  - i.e., finding out **why IPv4 is used when IPv6 is available**
  - e.g., a Java preference was not set to IPv6 for dCache

Jisc

What about DNS?

It's the same, but different

- Delegation happens as per IPv4 for forward and reverse DNS
  - Use AAAA records for IPv6 where you use A records for IPv4
  - Reverse DNS for IPv6 is nibble-based under ip6.arpa
    - So would be under *8.b.d.0.1.0.0.2.ip6.arpa* for *2001:db8::/32*
- All common IPAM systems should support IPv6
  - Infoblox is quite a popular solution used in R&E sites
- Don't add a host's AAAA entry to DNS until all services on it support IPv6 else there will be nothing listening for the connection

- Jisc offers a primary and secondary DNS service, and a network resolver service with RPZ built-in for malware protection

Jisc

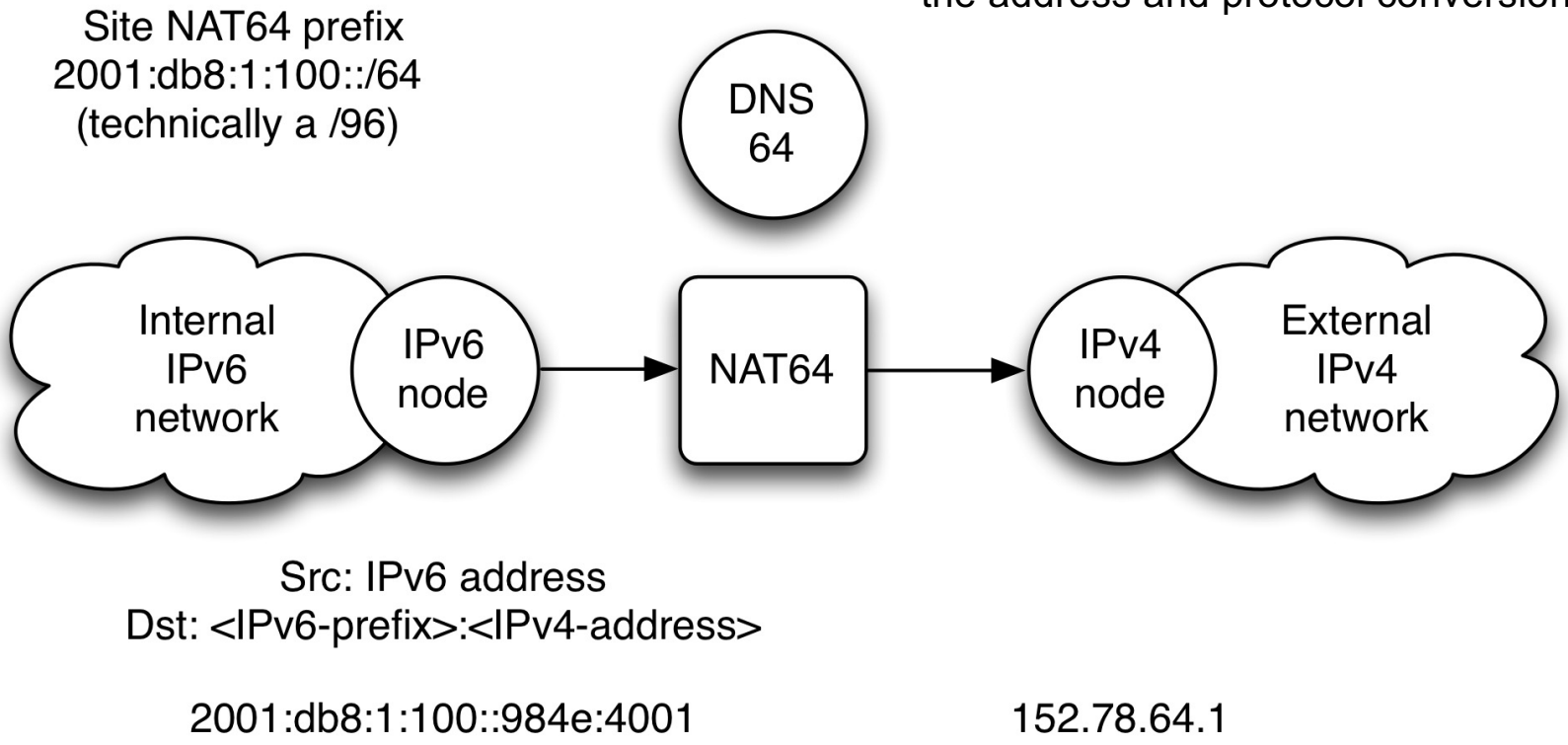
What about transition tools

The focus of ‘transition’ has shifted

- Originally multiple **tunnelling** methods were defined to support connecting IPv6 ‘islands’ over IPv4 paths, for hosts and routers
- With native IPv6 now pervasive on the R&E backbones, this need has largely gone away
  - Tunnel brokers can still be handy, e.g., *tunnelbroker.net*
  - A dual-stack VPN is a form of tunnel broker
  - Tunnelling IPv4 in IPv6 (as a service) is happening
- With a growing interest in running IPv6-only, the focus has shifted to **translation**, specifically for IPv6-only to IPv4-only content
  - The usual approach is NAT64/DNS64 and 464XLAT (for literals)
  - Also some use of MAP-T (RFC 7599)

NAT64 / DNS64 example

The ‘trick’ is used of DNS64, which returns a ‘fake’ IPv6 address formed by appending the IPv4 address to the site’s chosen NAT64 prefix, and the IPv6-only client sends to that address, with the NAT64 box doing the address and protocol conversion (out and back)



Security

Need to implement your policy for both IP versions

- Leaving IPv6 ‘open’ is not a great idea
- Dual-stack means you need to support / manage both protocols
  - Look to use tools that manage objects consistently, e.g. by using dual-stack firewall objects not separate IPv4 and IPv6 objects
- Focus where possible on feature equivalence
  - RA Guard is as important as DHCPv4 Guard, ARP vs DAD spoofing
- But recognise new threats
  - e.g., abuse of IPv6 extension headers
- And differences
  - Classic IP scanning is less feasible (see RFC7707)
- And IPv6 security matters in “IPv4 only” networks (see RFC 7123)

Jisc

(Unix) Troubleshooting

Not that different, for example:

- Tools like *ping* and *traceroute* work the same way
- *Netstat* can provide information for either protocol
- *Wireshark* and *tcpdump* work as expected
  - Useful for looking at ND traffic for example, like RAs
- You can check an ND cache like an ARP cache
- Bad ICMPv6 filtering might impact connectivity / PMTUD
- There are some other handy web-based tools
  - e.g., https://test-ipv6.com/, https://ipv6-test.com/, https://ip6.nl/

Jisc

Deployment – no Big Bang needed

You don’t need to do everything on day one

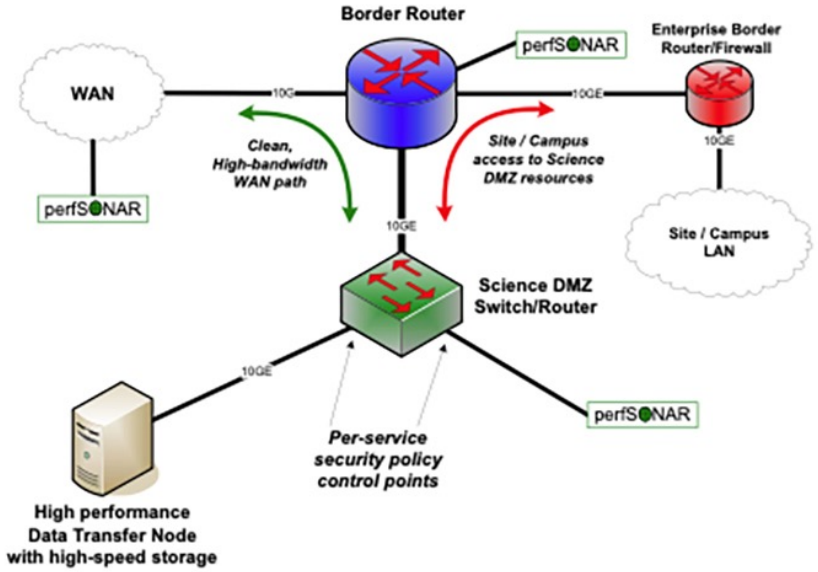
- R&E deployments to date have initially focused in one of the following areas:
  - Public facing services, particularly web, but also email and DNS
  - Campus WiFi (eduroam)
  - Computer science department, labs, etc. (teaching and research)
  - Computing service
  - **Science DMZ (e.g., GridPP systems within a site)**
- You still need to get IPv6 running in your core and edge network, but a focused initial deployment can be a more manageable project
- You can also learn from an initial smaller deployment to enhance a future wider deployment

Jisc

Aside: Science DMZ

Handling science and business traffic

- ESnet documented "**Science DMZ**" principles ~10 years ago
  - https://fasterdata.es.net/science-dmz/
- Key design elements:
  - Local network architecture to differentiate large science flows
  - Well-tuned data transfer nodes (DTNs)
  - Performant data transfer tools (**FTS**, Globus, etc)
  - Persistent monitoring of network characteristics (**perfSONAR**)
- Avoid the large flows traversing the main campus firewall
  - Apply security policy efficiently, save costs on the stateful DPI firewall capacity
- GridPP has of course evolved the same principles in parallel over time



Jisc

Procurement

Be sure to procure IPv6-capable products

- As a general guide, tenders should require feature parity, or at least a statement / roadmap from a vendor on IPv6 capability
  - Even if you don't plan to turn IPv6 on yet on the product

- There is RIPE guidance available for different types of equipment
  - RIPE772: https://www.ripe.net/publications/docs/ripe-772

- There's also the IPv6 Forum's IPv6 Ready Logo programme
  - https://www.ipv6ready.org/
  - (but not all vendors have put equipment forward for this)

Jisc

When can I remove IPv4?

Simplifying your IPv6 deployment longer term

- The prudent initial deployment approach is currently dual-stack
- But the question soon becomes “where can I remove IPv4?”
  - Internal management network? (Facebook is 100% IPv6 internally)
  - On your site WiFi?
  - For research communities? The WLCG is heading this way
- You will likely need to deploy tools that support IPv6-only devices accessing IPv4-only content, such as NAT64/DNS64/464XLAT
  - But will all your applications work when using such tools?
  - The more you support IPv6 in applications, the less you need translation

Jisc

New IPv6 protocol developments

New IPv6 protocol developments

In summary, very little

- The core spec was hardened with RFC 8200 in July 2017
- Current topics in the IETF?
  - IPv6 segment routing
  - Updating RFC6724 address selection for ULAs
  - Some new tools that use IPv6 EHs (Alt-Mark, minMTU, …)
  - Nothing earth-shattering
- We will present the WLCG IPv6 packet marking work as an informational draft at IETF 117 in July
  - Glasgow, Lancaster, Brunel testing, using XRootD

Jumbo frames – 9000 MTU

Not new, but an opportunity

(Apologies to those who were at the RNE call on Jumbo frames, slides here are re-used)

Jumbo frames

Overview…

- Periodic interest from communities to make use of jumbo frames
- Higher link capacities coming – larger frames intuitively make sense
- WLCG community made a proposal in 2018 but no formal adoption
- Other projects have recommended their use (e.g., AENEAS for SKA)
- Experimental evidence that larger MTU improves performance
- But also some concerns
- Should Janet-connected HEP sites make more use of them?

What do we mean by a “jumbo” frame?

Context given in a WLCG meeting in October 2018

- Maximum Transmission Unit (MTU) – “*largest layer 3 (IP) data unit that can be communicated in a single network transaction*”
- “Jumbo frame” is ethernet frame with **(IP) payload > 1500 bytes**
- Goal is “*end-sites to be able to set their NIC MTU=9000 and have those packets be able to traverse the intervening networks without fragmentation*”
  - Implicit choice of 9000 MTU for **hosts**
- Note there may be framing bytes added by network operators
  - MPLS adds an additional 8 bytes, VXLAN adds 50 bytes
  - So backbones will need to see larger frames

Current status on Janet

What can Janet sites do now?

- The Janet network will carry jumbo frames where sites have a standard Janet IP connection and use 9000 MTU on their LANs
- QMUL and RALPP are using jumbo frames
  - Our NOC can look for packets > 1500 bytes on the backbone
- Jisc's network test facilities at Slough and London use jumbo frames
  - So we can test with perfSONAR, for example
- However, where a L2VPN Netpath is in use on Janet, the Janet MPLS implementation may limit the MTU to less than 9000 (or require fragmentation)

Jisc

IPv4 and IPv6 have relevant differences

Important to note

- IPv4 allows fragmentation along a path
  - A router may fragment, and the receiving host reassemble
  - Either of those may cause performance issues
  - Can use the Linux option net.ipv4.tcp_mtu_probing=1 as recommended at https://fasterdata.es.net/host-tuning/linux/

- IPv6 only allows fragmentation at hosts
  - Thus PMTUD for IPv6 must work, especially Packet Too Big messages – don't blindly filter all ICMPv6 – see RFC4890

Jisc

What is the potential benefit of jumbo frames?

In principle, higher throughput

- Fewer packets to process means less load on CPU
  - Link capacities are rising, CPU speeds less so
- Larger frames means faster ramp up / recovery for most TCP algorithms after a congestion event
- The TCP calculator provided by SWITCH gives a theoretical (Mathis) perspective:
  - https://www.switch.ch/network/tools/tcp_throughput
  - Plug in MSS, RTT and estimated loss rate
  - Rate <= MSS/RTT * 1/(sqrt(loss))
- But experimental evidence is always good to see…

Jisc

Experimental results example

SURF to Jisc London – 9000 MTU

- Interval Throughput Retransmits Current Window

0.0 - 1.0	24.38 Gbps	0	90.12 MBytes
1.0 - 2.0	27.57 Gbps	0	90.73 MBytes
2.0 - 3.0	22.58 Gbps	0	90.73 MBytes
3.0 - 4.0	25.98 Gbps	0	90.73 MBytes
4.0 - 5.0	23.03 Gbps	0	90.73 MBytes
5.0 - 6.0	22.75 Gbps	0	90.73 MBytes
6.0 - 7.0	22.41 Gbps	0	90.73 MBytes
7.0 - 8.0	21.82 Gbps	0	90.73 MBytes
8.0 - 9.0	21.93 Gbps	0	90.73 MBytes
9.0 - 10.0	20.06 Gbps	0	90.73 MBytes

- Summary

Interval	Throughput	Retransmits	Receiver Throughput
0.0 - 10.0	**23.25 Gbps**	0	23.23 Gbps

SURF to Jisc London – 1500 MTU

- Interval Throughput Retransmits Current Window

0.0 - 1.0	8.91 Gbps	14145	10.54 MBytes
1.0 - 2.0	8.57 Gbps	0	10.68 MBytes
2.0 - 3.0	8.54 Gbps	263	5.41 MBytes
3.0 - 4.0	4.41 Gbps	0	5.55 MBytes
4.0 - 5.0	4.52 Gbps	0	5.69 MBytes
5.0 - 6.0	4.62 Gbps	0	5.86 MBytes
6.0 - 7.0	4.82 Gbps	0	6.18 MBytes
7.0 - 8.0	5.14 Gbps	0	6.65 MBytes
8.0 - 9.0	5.56 Gbps	0	7.25 MBytes
9.0 - 10.0	6.13 Gbps	0	8.02 MBytes

- Summary

Interval	Throughput	Retransmits	Receiver Throughput
0.0 - 10.0	**6.12 Gbps**	14408	6.09 Gbps

Raul ran these tests in February 2023 from a perfSONAR server in SURF to a Jisc server in London. The results were similar for 12 x 10 second tests and 60 x 1 second tests. Also, we noted that in both cases the 1500 MTU tests had several instances of retransmissions throughout the test, which may have affected the window size and thus performance.

Jisc

Concerns?

What potential concerns are there?

- The router and all hosts on a LAN should run the same MTU
  - Although some mixed mode cases might work
  - But is that a problem? Put non-jumbo hosts in another LAN?
- PMTUD messages may be blocked, and data transfers thus fail
  - For IPv6, RFC4890 should be followed – don't drop ICMPv6 PTB!
- NREN backbones may not have enough overhead
- Fragmented (IPv4) packets may be dropped by security policy
- Others…?

Finding more IPv6 information

IPv6 resources

Some useful links

- Jisc material:
  - IPv6 service page: https://www.jisc.ac.uk/ipv6
  - Advice and guidance page: https://www.jisc.ac.uk/guides/how-to-begin-an-ipv6-deployment
  - Janet IPv6 Technical Guide: https://repository.jisc.ac.uk/8349/1/janet-ipv6-technical-guide.pdf
  - Training: https://www.jisc.ac.uk/training/ipv6-fundamentals
  - And we're happy to spend time chatting with you, just ask
  - JiscMail ipv6-users@jiscmail.ac.uk list, subscribe at https://jiscmail.ac.uk/IPv6-users

- Other sources:
  - UK IPv6 Council: https://www.ipv6.org.uk/ - recent talks on GridPP, AWS support, Enterprises
  - UKNOF: https://www.uknof.org.uk/ - many talks on IPv6
  - Procurement : https://www.ripe.net/publications/docs/ripe-772 (updated - was RIPE 554)

Jisc

Questions?

A plug to end with…

Research Network Engineering call on Globus and the RFI

- By Silvia Ramos of the Rosalind Franklin Institute
- Friday 23 June at 2pm
- Good opportunity to learn about the work of the RFI and how they use Globus for data transfers
- See the community page at https://beta.jisc.ac.uk/get-involved/research-network-engineering-rne-community-group
  - Register by Zoom
- Also feel free to join the RNE list
  - https://jiscmail.ac.uk/RNE
  - (event details are in the email in the archive there too…)

Jisc

Contact:

Tim Chown

tim.chown@jisc.ac.uk

4 Portwall Lane,
Bristol, BS1 6NB

customerservices@jisc.ac.uk

jisc.ac.uk