Network Expectations for Data Intensive Science

Introduction

This guide is aimed at researchers who are running or planning to run data intensive science applications over the Janet network.

As it becomes more common for scientific (and other) research to be conducted between multiple collaborators and sites, there's a growing need for ever-larger data sets to be transferred between sites, over our national research and education network, Janet.

The guide seeks to help you, as a researcher, set realistic expectations for what's achievable over the network, and to get a better understanding of the factors that might affect the performance of the networked data-intensive applications you may need to use.

Much of the content of this document may equally apply to other applications, e.g. those supporting remote delivery of teaching material. The guide may also be of interest to network managers or operators at campus networks, in helping them understand their researchers' perspectives.

The problem

Researchers in a variety of disciplines are running ever more data-intensive applications which allow them to capture, visualise, analyse and archive their data. While previously these tasks might all have been performed within a single lab, department or local data centre, increasingly data is being exchanged between collaborating universities, or being sent to remote compute facilities (be that an academic HPC facility, or a commercial cloud provider such as Amazon, Microsoft or Google); in either case, the capability of the data network becomes an important factor. As an example. an HPC facility pushing data to a remote archiving site might currently look to achieve throughput in the region of 3-4Gbit/s.

New types of networked scientific equipment are also being deployed, capable of generating a very high volume and variety of research data. Such equipment will not necessarily have a complementary bespoke compute capability on-site, so the data will often need to be transferred to a remote compute and storage location. An example, it might be necessary to push data captured from a modern electron microscope at around 10Gbit/s to a remote compute facility, in order to get a visualisation of results within a minute or less. At an extreme scale, the newest square kilometre array (SKA) equipment can easily produce 100Gbit/s+ of data, though deployment is still away off.

While the specific requirements your application will have to run effectively over the network may vary, it's likely that without due attention being given in advance, the network can become a limiting factor on your ambition.

It's important to remember that your campus or research organisation will need to determine how best to support your applications, both now and into the future, bearing in mind it will also have to support the more routine network traffic generated by staff and students, such as web, email, and video streaming, and to do so in an appropriately secure way. It's thus recommended that you should seek to establish and maintain regular dialogue with your local computing service staff, if you haven't done so already. It's quite possible, and we've seen a number of examples already, that universities will have multiple data-intensive science research groups at the same site, and thus the local computing service will need to determine how to support each of them effectively, alongside the other day-to-day network traffic.

Articulating your network requirement

As a researcher, your view of the problem you are trying to solve, or the thesis you are testing, is likely to map to a methodology or workflow you have defined, but it's perhaps less likely that you will have an understanding of the demands put on network infrastructures, locally and beyond, by that workflow. You may, for example, be fetching remote data to process locally, sending data to a remote computation facility from where results may be returned to you, or archiving data to longer-term storage.

Ideally, all researchers would be able to articulate their network requirements, but in practice of course that's not a reasonable expectation. While some communities, such as the particle physics (GridPP) community, have been successfully transferring traffic at rates of 10Gbit/s and above for a number of years, and thus have experience in articulating their network requirements and performing "future looks", other emerging communities may understandably lack such knowledge.

In principle, the questions are not complex. How much data is being generated or transferred? Over what period of time? If data is being sent to a remote compute facility, to be returned for visualisation, what is an acceptable turnaround time, and how bound is that process by the network, as opposed to the remote compute, or the local rendering of the result? In general, a ball park idea is sufficient, to understand at least the order of magnitude of throughput required, but the more accurate the estimate, the better. We'll look at theoretical throughput, and factors affecting it, in more detail later in this guide.

The good news is that some communities, especially GridPP, have already demonstrated the ability to achieve significant throughput, e.g. Imperial College have recently had GridPP data flows approaching 40Gbit/s, which they have also demonstrated running over both IPv4 and IPv6. While your requirements may be relatively modest compared to this, achievements at higher data rates help establish a knowledgebase within Jisc and the community for wider adoption of such capabilities, and help ensure there is a path to move Janet-connected sites towards 100Gbit/s capability as and when needed in the coming years.

Janet capacity

Speaking of Janet, it's useful for researchers to have some understanding of the capacity of the Janet network to support traffic flows, even though the most likely bottlenecks will be at the edges, i.e. at the campus border to Janet (the capacity of your site's connection to Janet), and within the internal campus networks, depending on their architectures and link speeds.

The Janet core network is designed to have sufficient latent capacity for the immediate future, and utilisation of the infrastructure is under regular review by Jisc. To give an idea of the backbone capacity, most of the links between the Janet core backbone routers are being upgraded to 600Gbit/s by 2018. External links to other academic networks, and to commercial providers, are also reviewed regularly to ensure they are provisioned to have enough headroom for the traffic being sent over them. For example, as more Janet sourced traffic goes to Google, Amazon or Microsoft cloud services, Janet's connectivity to those providers is upgraded appropriately. While Jisc can't control what happens within those clouds, it will provide optimal connectivity to them for you, through Janet.

It is worth noting that Janet has not deployed any generic quality of service mechanisms, so there is no differential treatment of traffic across Janet's IP network.

However, Jisc does offer the Netpath service[1], formerly referred to as the Lightpath service, which can support both dedicated capacity between sites where required, and/or an extended layer 2 service (Layer 2 VPN, or L2VPN). If setting up a dedicated Netpath service of a certain capacity, you should note that your throughput is then of course limited to that dedicated capacity. That is one reason that Jisc generally encourages campuses and their research communities to try to make use of their regular Janet IP service before looking to arrange any configured Netpath service, not least because Netpath also has a cost to implement.

Campuses connect to the Janet core network through the regional access networks, which with the latest instance of Janet, Janet6, are now managed directly by Jisc; you thus have end-to-end network support across Janet from Jisc, and at a service level through the Janet Network Operations Centre (NOC).

[1] https://www.jisc.ac.uk/netpath

Your campus connectivity

Your campus will have network connectivity to Janet of a certain capacity. It may also have a resilient link, for use in the event of the primary link failing. An example for a typical university might be that it has dual 10Gbit/s links, but connectivity varies from site to site, so your own campus may have a different capacity.

As mentioned above, your campus connectivity will serve a variety of day-to-day purposes – teaching, research, administration, student hall networks, etc. Many of these include mission critical functions which depend on the network being available.

There is thus a challenge for your computing service should you expect to run an application at (say) 5Gbit/s, if the campus has a 10Gbit/s Janet connection, but already has day-to-day traffic peaking at 5Gbit/s or above. While your computing service might plan its capacity requirement based on an organic growth model for traffic, just one new data-intensive scientific application might present a significant step change in those requirements.

Most campus networks will have network security measures in place, specifically firewalls and intrusion detection systems, that are required to enforce an appropriate security policy on traffic entering or leaving the campus. While such devices may perform adequately for routine traffic, but they may not be capable of supporting high throughput, data-intensive science applications. There are examples of Janet sites where research traffic was constrained by such devices to a few hundred Mbit/s, but which then obtained an order of magnitude better performance when the science traffic was routed such that it did not pass through the main campus firewall. That doesn't mean to say there should be no network security for data-intensive applications, but that policy needs to be applied more efficiently for such traffic.

If your application has significant requirements above and beyond your campus's regular network capability, your computing service team may need to consider what steps to take to support you. This might include:

- Contacting Jisc, via your organisation's Account Manager, or via the Janet NOC/Service Desk, to discuss the increased requirements; this might involve a conversation about upgrading the campus connectivity, e.g. from 10Gbit/s to 20Gbit/s;

- Contacting Jisc to explore the use of the Netpath service to support the application;

- Considering adjusting the local campus network architecture to better support your requirement; this might involve routing traffic to your systems more efficiently, perhaps not via the main campus firewall, and instead using other more application-specific security measures.

There is a growing number of examples of universities where local network engineering techniques have been applied to help support data-intensive applications, so there is also a growing level of expertise in the community in this area. Jisc will be helping to promote relevant best practices; an example of this was the campus network engineering workshop held in 2016[2], where the "Science DMZ" model was discussed.

Some campuses might be tempted to (unofficially) use their resilient link for "bulk" science data. While this is one way of avoiding having your science data not interfere with the regular campus traffic, if there is an outage of the primary link then the campus may have difficulty re-routing the regular traffic down the resilient link if that link is already carrying a significant volume of science traffic.

In principle, it's better to have a conversation with Jisc about upgrading the main link capacity, while also arranging an appropriate level of resilience. Similarly, there should be no need for your application traffic to be rate-limited (capped to a specific throughput) to avoid it contending with your campus's regular traffic; again while rate-limiting may be a necessary interim measure as a new science application is first deployed, longer term it would be preferable to have that capacity upgrade conversation.

[2] https://www.jisc.ac.uk/events/campus-network-engineering-for-data-intensive-science-workshop-19-oct-2016

Note that Jisc monitors the utilisation of all Janet campus connections, but can only initiate discussion with campuses about capacity upgrades on the basis of predictable, organic traffic growth; new step changes through new data-intensive science applications being deployed cannot be predicted that way.

Exploring theoretical network throughput

To understand your application's network requirement, for which the most important property is likely to be raw throughput, the most obvious approach is to determine the size of the data set to be transferred, and the time in which that needs to happen. Where data sets are collected or generated in advance, their size can obviously be inspected to allow an estimate to be made. But where data is being streamed live from a piece of equipment, that may not be so easy to do, but one might expect, for example. that the encoding mechanism is documented, so that an idea of at least an average bit rate is available.

You should remember when calculating data rate estimates that file sizes are typically expressed in bytes, e.g. 1TB, while network links are typically expressed in bits per second, e.g. 10Gbit/s, so in calculating the throughput required you need to remember to convert each byte to 8 bits, or vice versa.

As an example, a 100GB data set would be 800,000,000,000 bits. To transfer that data set in one day, you would need to achieve a throughput of 800,000,000,000 / (24 x 60 x 60) = 9.2 Mbit/s.

Alternatively, if you were looking to send 1PB of data over a 100Gbps link, it would take 80,000 seconds (8,000,000,000,000,000 bits in 100,000,000,000 seconds), which is just under a day (0.93 days).

The following table, taken from a publication by ESnet[3], shows the *theoretical* throughput required to transfer a given size of data set in a range of example time periods.

	1 Min	5 Mins	20 Mins	1 Hour	8 Hours	1 Day	7 Day	30 Days
10 PB	1,333Tbps	266.7Tbps	66.7Tbps	22.2Tbps	2.78Tbps	926Gbps	132Gbps	30.9Gbps
1 PB	133.3Tbps	26.7Tbps	6.67Tbps	2.2Tbps	278Gbps	92.6Gbps	13.2Gbps	3.09Gbps
100 TB	13.3Tbps	2.67Tbps	667Gbps	222Gbps	27.8Gbps	9.26Gbps	1.32Gbps	309Mbps
10 TB	1.33Tbps	266.7Gbps	66.7Gbps	22.2Gbps	2.78Gbps	926Mbps	132Mbps	30.9Mbps
1 TB	133.3Gbps	26.67Gbps	6.67Gbps	2.22Gbps	278Mbps	92.6Mbps	13.2Mbps	3.09Mbps
100 GB	13.3Gbps	2.67Gbps	667Mbps	222Mbps	27.8Mbps	9.26Mbps	1.32Mbps	309Kbps
10 GB	1.33Gbps	266.7Mbps	66.7Mbps	22.2Mbps	2.78Mbps	926Kbps	132Kbps	30.9Kbps
1 GB	133.3Mbps	26.7Mbps	6.67Mbps	2.22Mbps	278Kbps	92.6Kbps	13.2Kbps	3.09Kbps
100 MB	13.3Mbps	2.67Mbps	667Kbps	222Kbps	27.8Kbps	9.26Kbps	1.32Kbps	0.31Kbps

Thus, in principle, if you need to move 100GB in 20 minutes, you will need at least a 1Gbit/s capacity, end to end. Or, if you have a 10Gbit/s link, you can in principle move 100TB in a day (at a rate of 9.26Gbit/s).

While Terabit networking is not with us in production yet, it will come, and it's important that Jisc is able to plan for that capacity on its backbone, to enable sites to increase their own capacity over time. Currently, very few Janet sites have 100Gbit/s connectivity, but Jisc is starting to see some more requests for such capacity, and it is thus useful for campuses to perform "future looks" on what their what their future requirements might be, so they can work with Jisc to plan future Janet capacity upgrades in a timely fashion, based on their genuine research needs. It may be worthwhile to ask your computing service about such planning, to ensure your requirements are being included; it's likely they'll be more than happy to discuss this with you.

Factors affecting general throughput

The theoretical throughput described above is generally challenging to achieve, for a variety of reasons. There are a number of factors that may limit your applications' performance. These include:

[3] http://fasterdata.es.net/home/requirements-and-expectations

- Competing traffic on the links over which you're sending the data;
- Network devices, especially firewalls, that are not capable of supporting the necessary throughput;
- Limitations in the end systems to source or sink the data from/to disk;
- Limitations in the tools or applications used to transfer the data;
- The nature of the data; a large number of very small files may be less efficient to transfer.

These are just examples of the end-to-end performance challenges you may face. It's important to remember that while the network plays an important part, there are many other nuanced considerations in play.

Where there is competing traffic, it is possible that network congestion will occur, at which point packets of data in excess of the capacity will be dropped. It is important to understand the impact of such packet loss on throughput; indeed an apparently quite minor loss rate can have a surprisingly significant effect on performance.

That impact will depend on whether the application uses TCP or UDP, which are the two dominant transport protocols used on the Internet today, both of which run over IP. TCP provides a connection-oriented service to the application and is "friendly" in the face of congestion, in that TCP adjusts its sending rate based on perceived packet loss; it will initially ramp up the rate, but drop it back down when congestion is detected, then again ramping up until another congestion event happens. Thus multiple TCP applications on a path will over time have equitable use of that path. In contrast, UDP is connectionless and leaves the issue of dropped data retransmission to the application; if using a constant bit-rate it is not at all considerate of other applications on the same path.

TCP application throughput

While it's not reasonable to expect researchers to understand the fine details of network transmission theory, there are a couple of formulae that can be useful in helping you set and understand realistic expectations.

Firstly, the Mathis equation allows you to predict the maximum practical throughput for an application using a conventional version of TCP, based on the packet size (in TCP speak, the maximum segment size, or MSS), the round trip time (RTT) to the destination and back, and the probability of any given packet being dropped (as a probability, p). A simplified version of this formula is:

rate <= (MSS / RTT) * (1 / sqrt(p))

So, for a typical MSS of 1460 (for Ethernet, with IPv4), a RTT of 20ms, and a loss rate of 1e06 % (i.e 0.0001%), the result is 584Mbit/s. But if the loss rises to just 0.1% you drop to a maximum of 18.5Mbit/s. Thus it should be apparent that even a very low loss can have a significant effect on performance, especially for higher RTT paths, and that ideally you need to eliminate such loss as far as possible.

You can try different values for yourself with the online calculator at the SWITCH site[4].

You can also get a somewhat rough idea of RTT and loss to a destination you're interested in by using the *ping* tool, e.g.

```
$ ping www.jisc.ac.uk
PING www.jisc.ac.uk.cdn.cloudflare.net (104.20.27.251) 56(84) bytes of data.
64 bytes from 104.20.27.251: icmp_seq=1 ttl=51 time=4.54 ms
64 bytes from 104.20.27.251: icmp_seq=2 ttl=51 time=4.60 ms
64 bytes from 104.20.27.251: icmp_seq=3 ttl=51 time=4.43 ms
64 bytes from 104.20.27.251: icmp_seq=4 ttl=51 time=4.49 ms
64 bytes from 104.20.27.251: icmp_seq=5 ttl=51 time=4.47 ms
64 bytes from 104.20.27.251: icmp_seq=6 ttl=51 time=4.59 ms
64 bytes from 104.20.27.251: icmp_seq=7 ttl=51 time=4.42 ms
64 bytes from 104.20.27.251: icmp_seq=8 ttl=51 time=4.40 ms
64 bytes from 104.20.27.251: icmp_seq=9 ttl=51 time=4.49 ms
64 bytes from 104.20.27.251: icmp_seq=10 ttl=51 time=4.44 ms
```

[4] https://www.switch.ch/network/tools/tcp_throughput/

```
64 bytes from 104.20.27.251: icmp_seq=11 ttl=51 time=4.41 ms
64 bytes from 104.20.27.251: icmp_seq=12 ttl=51 time=4.45 ms

--- www.jisc.ac.uk.cdn.cloudflare.net ping statistics ---
12 packets transmitted, 12 received, 0% packet loss, time 11005ms
rtt min/avg/max/mdev = 4.407/4.481/4.606/0.108 ms
```

The above example shows an average 4.481ms RTT, with no loss, but is obviously a very small sample, and not representative of, especially, loss over time. To get a much better, and more accurate (over time) measurement of RTT and loss, you should consider running a more advanced network performance measurement tool that records measurements continuously, such as perfSONAR; more on this below.

Throughput will also depend on how well the sending system's network stack is able to fill the path between the sender and receiver. This is particularly important on longer RTT links, because transmission takes time (the speed of light is only so fast) so you need to have as many packets in flight concurrently as possible; think of it loosely as cars flowing along a motorway, the more cars moving, the higher the overall throughput.

This issue is generally expressed through something called the Bandwidth Delay Product (or BDP), which is the product of the link capacity and the RTT. For example, a 1Gbit/s link with a 60ms RTT has a BDP of 7.5MB. In order to fill such a link with data, the sending system will need a 7.5MB TCP send buffer, such that at any one time it can have 7.5MB of data in transit to the receiver, and be able to hold that data in memory to be able to resend it in the event of a loss being detected. A not uncommon problem here is that many operating systems have much smaller default buffer sizes, e.g. a 64KB send window is not uncommon. Unfortunately, the throughput will then be limited; in the case of a 1Gbps link with a 60ms RTT to just 8.74Mbit/s. However, for a nearby destination, with a RTT of perhaps of just 2ms, a 64KB window can deliver up to 262Mbit/s.

Tuning the buffers / window sizes on systems is more important where the BDP is higher, e.g. if you are collaborating with sites in the US, where the RTT may be around 70-150ms, but it's useful that you're aware of the general issue. You can also explore BDP and buffer settings at the SWITCH site.

There is an interesting development with Google's recently published TCP-BBR variant[5], which can reportedly achieve higher throughput in the face of packet loss up to 15%. It still behaves in a "friendly" way, but is smarter about its pacing of sending data. We'll be looking to explore TCP-BBR further at Jisc; there is an available implementation for Linux.

UDP application throughput

UDP-based applications will typically send data at a constant bit-rate, but might not back off in the event of congestion; it's important to understand whether they do (based on application measurements) to know the impact on your other (TCP) traffic on your network.

There are UDP-based data transfer applications available, such as UDT and Aspera. The latter is a commercial product, which is in use by some Janet-connected research organisations.

Choice of data transfer application

There is a wide range of data transfer tools, from the simplest ftp, scp or HTTP-based copy tools, to more advanced tools such as GridFTP. The GridFTP application uses TCP, but will parallelise the transfer, such that if four TCP flows are being used, and one experiences loss, the other flows do not back off; this makes GridFTP more efficient in the face of occasional loss.

The File Transfer Service (FTS) is an application that can manage large volume data transfers. The use of FTS3 for data archiving transfers between Durham and RAL was recently documented[6], including the details of certificate deployment and monitoring of the transfers.

[5] http://queue.acm.org/detail.cfm?id=3022184

[6] http://astro.dur.ac.uk/~dph0elh/documentation/transfer-data-to-ral-v1.4.pdf

What application throughout is possible over Janet today?

Jisc aims to support a wide variety of networking requirements for its Janet-connected sites, both by ensuring the Janet network is appropriately provisioned, and working with sites to help you make optimal use of your connection.

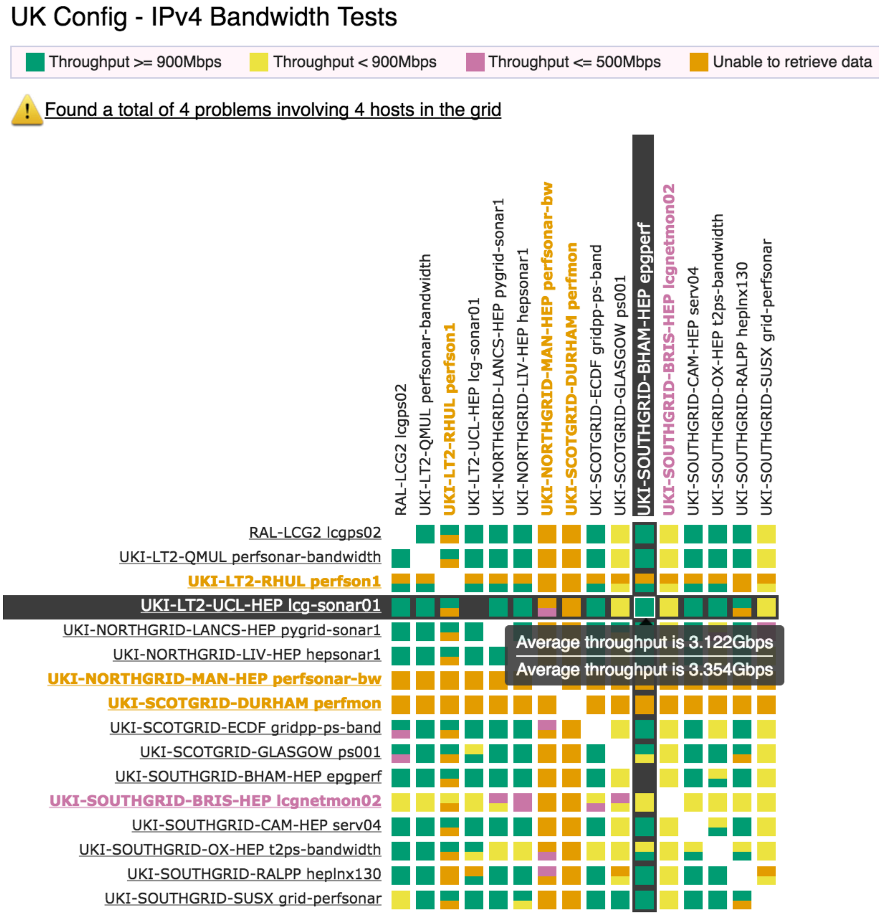
It's interesting, and challenging, to push the "high water mark" for networked applications. The 40Gbit/s being achieved by Imperial College for its GridPP traffic is currently (we believe) the highest throughput to/from a Janet site. While this is partly achieved by GridPP-specific engineering (using OPN / LHCONE links), it also shows that the data transfer nodes are capable of sourcing and sinking the data at that rate.

Other examples exist, e.g., there is a Janet site achieving 6Gbit/s to a collaborator in Spain, through the European GEANT academic backbone, and the Spanish equivalent of Janet, RedIRIS.

Performance measurement

One approach to performance measurement is to record the details of transfers, and the data transfer rate observed, within the application. FTS, as mentioned above, does this, for example.

It is also desirable to deploy a network performance measurement tool that can record network characteristics such as throughput, loss and latency over time. One such tool is perfSONAR[7], which is an open source toolkit that provides active measurement of network characteristics between two or more systems running it. While perfSONAR can be used to test network characteristics over time between just two sites, it can also be configured as a "mesh" that can monitor and summarise performance between multiple sites in a research community, through an intuitive web interface, as illustrated below for a UK GridPP dashboard.



UK Config - IPv4 Bandwidth Tests

🟩 Throughput >= 900Mbps 🟨 Throughput < 900Mbps 🟪 Throughput <= 500Mbps 🟧 Unable to retrieve data

⚠ Found a total of 4 problems involving 4 hosts in the grid

Average throughput is 3.122Gbps
Average throughput is 3.354Gbps

RAL-LCG2 lcgps02
UKI-LT2-QMUL perfsonar-bandwidth
UKI-LT2-RHUL perfson1
UKI-LT2-UCL-HEP lcg-sonar01
UKI-NORTHGRID-LANCS-HEP pygrid-sonar1
UKI-NORTHGRID-LIV-HEP hepsonar1
UKI-NORTHGRID-MAN-HEP perfsonar-bw
UKI-SCOTGRID-DURHAM perfmon
UKI-SCOTGRID-ECDF gridpp-ps-band
UKI-SCOTGRID-GLASGOW ps001
UKI-SOUTHGRID-BHAM-HEP epgperf
UKI-SOUTHGRID-BRIS-HEP lcgnetmon02
UKI-SOUTHGRID-CAM-HEP serv04
UKI-SOUTHGRID-OX-HEP t2ps-bandwidth
UKI-SOUTHGRID-RALPP heplnx130
UKI-SOUTHGRID-SUSX grid-perfsonar

[7] http://www.perfsonar.net/

The perfSONAR dashboard shows rows and columns for each site, with colour-coded boxes in the matrix indicating the throughput being achieved, green being throughout above 900Mbps, and purple being throughput below 500Mbps (these thresholds can be changed if desired). perfSONAR can also display throughput, or other characteristics over time, by clicking on one of the boxes in the matrix.

There is a separate guide, "The case for perfSONAR deployment in supporting of networking for data-intensive research", which can be found online at the Janet End-to-End Performance Initiative community area (listed below); you may wish to point your computing service at this document to encourage them to deploy it.

Jisc has a perfSONAR test node deployed in London with a 10Gbit/s interface. Organisations wishing to run tests against this node are welcome to contact us at Jisc (see the details below). We are also working to produce a small node perfSONAR build, that can drive tests of up to 1Gbit/s on low-cost hardware (under £250); again, if organisations are interested in this, please contact us.

Expectations for cloud applications

Jisc is exploring, and seeking to optimise, network performance characteristics between Janet and various cloud computing providers, such as AWS and Azure. This work is in its early stages; for example, we have anecdotal reports from researchers of them achieving 2-3Gbit/s to AWS. There is also information about potential throughput for AWS at https://aws.amazon.com/ec2/instance-types/; the performance listed depends on the 'size' of the resource purchased.

Contacts

Your first port of call for support for network performance matters should generally be your local campus computing service; they in turn can refer questions to the Janet NOC, or they may contact their assigned Jisc Account Manager.

The Janet End-to-End Performance Initiative can be reached via Tim Chown at tim.chown@jisc.ac.uk. We're more than happy to discuss any end-to-end application performance topics with you.

Further Reading

Further information can be found at:

- Janet End-to-End Performance Initiative project web site
  https://www.jisc.ac.uk/rd/projects/janet-end-to-end-performance-initiative

- Jisc End-to-End Performance Initiative community area
  https://community.jisc.ac.uk/groups/janet-end-end-performance-initiative

- JISCmail End-to-End Performance Initiative mail list (open to all)
  https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=E2EPI

- ESnet Fasterdata site, for various guidance
  http://fasterdata.es.net/

Dr Tim Chown
Network Development Manger
Jisc, Lumen House, Library Avenue, Harwell Oxford, Didcot. OX11 0SG
E: tim.chown@jisc.ac.uk
T: +44 1235 822106

04 April 2017
Document version 1.0