GridPP
UK Computing for Particle Physics

GridPP Project Management Board

LHC Network Forward Look 2017

Document identifier:	**GridPP-PMB-179-NetworkFL-2017**
Date:	**September/2017**
Version:	
Document status:	**Draft**
Author	**P.Clarke, D.Rand, D.Colling, R.Jones, A.McNab, A.Sansum**

Introduction

This document is a forward look for UK networking capacity required for LHC operations. This is prepared mid 2017 as the LHC enters the third year of Run-2.

The scope of this report includes performance and capacity for

- The UK Tier-1 to CERN and other Tier-1s via the LHCOPN
- The UK Tier-1 to Janet
- Tier-2 networking via Janet

The future LHC schedule is:

- 2017: Year 3 of Run-2
- 2018: Final year of Run-2 and beginning of long shutdown 2 (LS2)
- 2019: LS2

Summary

The LHC is now just over half way through its Run-2, which started in 2015 and will finish mid 2018. To date, the LHC has been very successful, leading to larger data rates/volumes than expected. In addition, several experiments are moving further towards "move data to compute" models, or accessing data over the network. Therefore, we expect the network requirement to increase.

- We expect that the Tier-0 ⇔ RAL Tier-1 network requirement will rise to approximately 40 Gb/s within the next 2 years, with at least 20 Gb/s fall back redundancy. Tier-0 ⇔ Tier-1 data will continue to be transferred via the LHCOPN for the foreseeable future.

- The Tier-1 connection to Janet should be served by the on-going upgrade plans for the RAL campus, which targets 60 Gb/s and even 100 Gb/s later. This is within the planning scope of STFC-RAL and its connection onto Janet.

- In the next 2 years, the larger Tier-2 sites (Glasgow, Manchester, Lancaster, QMUL, Imperial, RAL) are likely to need at least ~ 20 Gb/s connections for Tier-2 traffic alone and some may even require 40 Gb/s (ICSTM already has 30 Gb/s).

- We expect the medium and smaller GridPP sites to require 5-10 Gb/s for Tier-2 traffic, though there are some "medium" sites that currently have a high network I//O rate and a connection bandwidth similar to the larger sites.

- We note that ATLAS and CMS plans envisage major Tier-2 sites connected at 40-100 Gb/s in the next few years. This is already being implemented in the US and there are plans in Germany and Italy. GridPP will keep this under review and advise Janet appropriately. This would be limited to the large sites listed above.

- The UK still sees no imperative to use LHCONE at present, but continues to work to join in order to understand any associated issues.

Experiment Statements

The experiment forecasts with respect to current data rates for 2017-2019 are:

- ATLAS is in the process of constructing a more detailed network model. However, the large-scale features are already clear: Over the next 2(4) years ATLAS expect a growth of processing and analysis capacity of a factor 1.5(2) (roughly 20% per year). These processes will either fetch data from storage and/or produce new datasets for storage. However, the architecture of the distributed system is changing, with an increased concentration of storage at fewer sites. Consequently, there will be an additional increase in the bandwidth requirements, with larger sites needing to serve remote CPUs.

  The overall computing model in terms of formats, versions and selections has recently been revised and will be stable for another 4-5 years. This means that on the 2-year timescale, that the large ("nucleus") ATLAS sites (QMUL, Manchester, Lancaster and Glasgow) with need at least ~ 20Gb/s for the Tier-2 traffic alone (and hence a greater capacity at the campus site overall). It is possible that some of these sites may require 40 Gb/s on this timescale. The UK needs to monitor the rate of growth of the network bandwidth at the larger ATLAS sites.

  For the smaller lower storage ("satellite") sites, operational considerations require the site to have a demonstrated connection to a nucleus site of greater than 800Mbp/s. In the UK, the satellite sites are actually relatively large and should be able to provide connections to ~4 UK nucleus sites, with a weak scaling with the processing capacity available for ATLAS. Therefore, a bandwidth of 2-4Gbp/s at the smaller sites should be sufficient with headroom for other traffic.

- CMS has been successfully making use of remote data access using AAA ('Any data Anytime Anywhere') for 3-4 years and it now forms a significant part of the computing model. For example, CMS jobs run opportunistically at four UK ATLAS sites: QMUL, Glasgow, Oxford and RHUL, and read their CMS data remotely from storage at CMS sites.

  In the future CMS is looking at 'data lakes' where sufficient copies of data are placed in a given region to support the use of that data in that region. Regions will be chosen to make sense in network connectivity terms; these could be a country or even a continent. Such data lakes will require 'data hosting sites' which will be mainly Tier-2s, but the Tier-1s will also act as data sources.

  CMS will require similar capacity to ATLAS at its main sites (ICSTM, Brunel, RAL), i.e. at least 20 Gb/s for most of the UK Tier-2 traffic alone in the next 2 years. The exception to this is the large Tier-2 site at ICSTM which already uses up to 30 Gb/s of a 40 Gb/s connection, and further upgrades toward 100 Gb/s may be indicted in the future. The large USA Tier-2s are already at 100GB/s and there are plans in Germany and Italy to similarly upgrade. The UK needs to monitor this situation.

- LHCb primarily uses Tier-1 sites and selected Tier-2 sites for its data processing. For Tier-2 sites the data will be down/up-loaded from/to Tier-1 storage which in the past led to only minimal increase in transfer bandwidths. The LHCb trigger rate in Run2 has increased to 12.5 kHz (from 5 kHz). Simulation jobs will run on any Tier level and for sites without storage the output data will be uploaded to a topologically close storage site. User analysis jobs requiring input data are executed on sites with storage (Tier-1, Tier-2-D) aiming to retrieve the local replica for processing and only in case of non-availability will fall-back on a remote storage. The overall network use for the remainder of Run2 and LS2 will therefore stay approximately at the same level. LHCb does not intent to move data around significantly, and in any case, it is only a small part of the overall UK LHC resource usage. Thus, LHCb will have modest requirements of bandwidth at Tier-2 sites.

In summary, it is clear that the larger Tier-2 sites (Glasgow, Manchester, Lancaster, QMUL, ICSTM, RAL) are set to require at least ~ 20 Gb/s connections for Tier-2 traffic alone within the next 2 years, and some may even require 40 Gb/s. The medium and smaller sites will continue to require between 5-10 Gb/s. GridPP needs to keep an eye on the move to higher bandwidth connections to Tier-2 sites elsewhere and advise Janet as appropriate.

UK Tier-1 to OPN

The LHC Optical Private Network (LHCOPN) connects the WLCG Tier-1 centres together through a star shaped network centred at CERN. Until recently the operational configuration was a pair of diversely routed 10Gb/s circuits between RAL and CERN, operated in resilient, active/passive mode giving 10Gb/s capacity with the second circuit providing resilience in the event of a failure.

Performance:
Following the start of LHC's running in May 2016, network load on the Tier-1's external links grew substantially beyond that experienced in the previous year(s). In particular, the 10Gb/s primary OPN link from CERN to RAL began to saturate regularly and individual file transfer rates began to suffer substantially. By July 2016 the link was flat-topping for weeks at a time. Following discussion with CERN the unused backup link was brought into play in August 2016 as a second primary link and traffic load balanced over both links, much improving the situation.

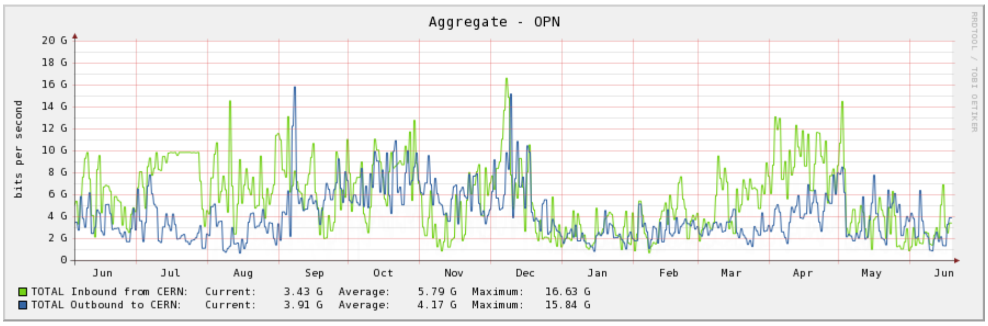


Figure 1: Aggregate Tier-1 OPN Traffic June 2016 to June 2017: The bandwidth was increased after the flat-topping seen in July.

Whilst average network rates have been usually well within the 20Gb/s maximum capacity available, peak rates since August 2016 (Figures 1 and 2) are frequently close to the maximum available bandwidth of 20Gb/s.

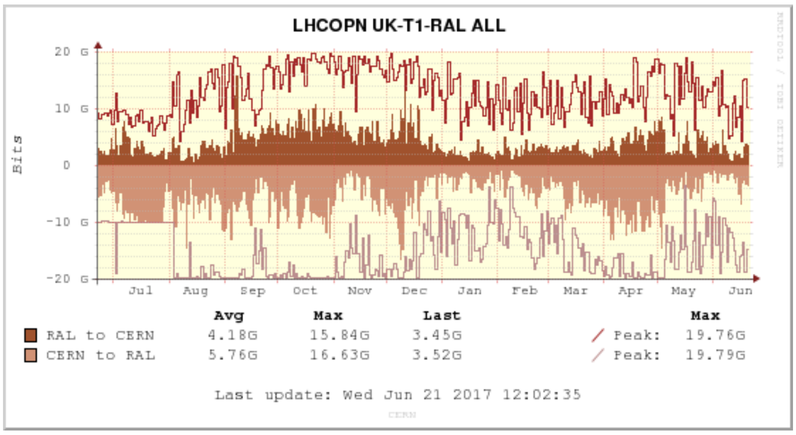


Figure 2: LHCOPN Peak Rates (line) between CERN and RAL. Upper half of plot RAL-CERN, lower half CERN-RAL (average rates - solid)

While bringing the second link into operation as a primary increased bandwidth, in the event of a failure (scheduled or unscheduled) capacity immediately reverts to 10Gb/s having an immediate impact on transfer rates and consequently service operation. Demand is expected to continue to rise as the run proceeds and to address continued rising demand and to improve the ability to handle a link failure, a third 10Gb/s link has been provided by Janet (within the GridPP5 planning cost envelope) and this is expected to be brought into service in June 2017 providing a total of 30Gb/s.

David Britton 5/9/2017 10:28
Comment [1]: Update?

A continuing complication is that there exist failure modes (and scheduled interventions) that can reduce available bandwidth to 10Gb/s. As load continues to climb, the operational impact (even if temporary) of reducing bandwidth by two thirds may prove to be unacceptable and it may be necessary to increase bandwidth on the secondary path to match that on the primary.

Forward look

1. **We expect that the Tier-0 ⬄ Tier-1 network requirement will rise to approximately 40 Gb/s within the next 2 years, with at least 20 Gb/s fall back redundancy.**
2. **Tier-0 ⬄ Tier-1 data will continue to be transferred via the LHCOPN for the foreseeable future.**

UK Tier-1 Janet

The RAL site is resiliently connected at 2 x 40Gb/s active/standby directly to the Janet Core, distributed at RAL across two geographically distinct endpoints located in the Atlas building and in building R89. The RAL network infrastructure is constantly monitored and reviewed by the STFC Network Technical Design Authority (TDA) that has the authority to recommend specific improvements through the STFC ICT governance process.

The Tier-1 connects to the RAL core and Janet routers by two routes, both resilient 40Gb/s connections:

- Control traffic and batch worker node traffic is routed through the main Tier-1 router, via the site firewall to Janet.
- Data traffic from the Tier-1 storage systems bypasses the main firewall (for performance reasons) but instead routes through a router shared with the LHCOPN network.

Performance:

Until October 2016 traffic between the Tier-1 and Janet via the bypass route used a single 10 Gb/s link that bypassed the firewall, however this link became a significant bottleneck through the summer of 2016 and in early October 2016 one of the Tier-1 routers was replaced, allowing a direct 40Gb/s link to the edge of site.

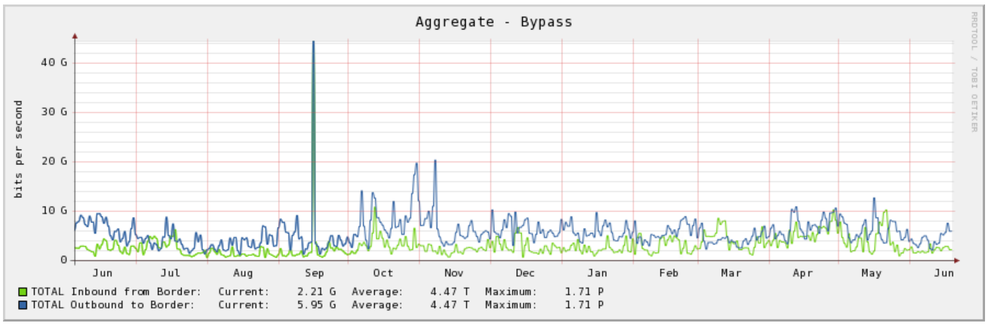


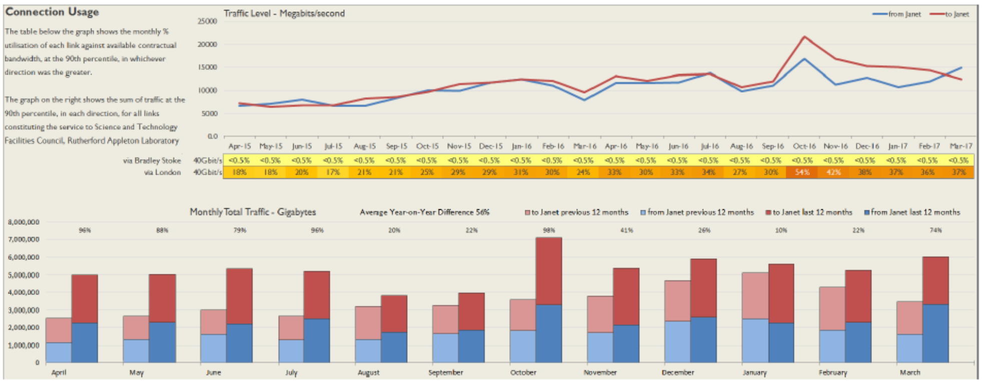
Figure 3: Tier-1 Traffic to Janet via RAL Science DMZ (biggest peak being an artefact of monitoring system) June 2016 to June 2017

Following the update, in early October 2016, Tier-1 traffic flows peaked dramatically, spiking at approximately 20Gb/s although subsequently they have fallen back to similar levels as seen at the start of last year’s run.

The Tier-1's Janet connection is the usual data route for traffic to the Tier-2s, primarily in the UK but also potentially globally. Traffic growth through the year (apart from the October spike) has been relatively modest, however it remains hard to predict how demand might change through the remainder of Run 2. There are several factors to consider:

- As the experiments reduce the number of replicas globally (in response to resource constraints) there is a growing requirement for Tier-2s to either access the data remotely or cache temporarily and flush the caches frequently.
- Ongoing enhancements in the Tier-1 storage system (deployment of the Echo service) may raise I/O performance, making transfers from RAL more attractive.
- Removal of I/O bottlenecks from the batch farm may become necessary, routing traffic via the bypass in order to raise batch job efficiency.

At the RAL site routers, traffic from the Tier-1 converges with traffic from the Tier-2, JASMIN and other data producing and consuming services. Traffic to/from Janet continues to grow from the RAL site (*Figure 4*) and while average rates remain around 15Gb/s, peak rates can be much higher. STFC is investigating options for further upgrades to the RAL site link, initially to 60Gb/s and eventually 100Gb/s however timing of such an upgrade will depend on financial constraints.



Connection Usage

The table below the graph shows the monthly % utilisation of each link against available contractual bandwidth, at the 90th percentile, in whichever direction was the greater.

The graph on the right shows the sum of traffic at the 90th percentile, in each direction, for all links constituting the service to Science and Technology Facilities Council, Rutherford Appleton Laboratory

		Apr-15	May-15	Jun-15	Jul-15	Aug-15	Sep-15	Oct-15	Nov-15	Dec-15	Jan-16	Feb-16	Mar-16	Apr-16	May-16	Jun-16	Jul-16	Aug-16	Sep-16	Oct-16	Nov-16	Dec-16	Jan-17	Feb-17	Mar-17
via Bradley Stoke	40Gbit/s	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%
via London	40Gbit/s	18%	18%	20%	17%	21%	21%	25%	29%	29%	31%	30%	24%	33%	30%	33%	34%	27%	30%	54%	42%	38%	37%	36%	37%

Figure 4: RAL Site Traffic April 2015 - March 2017

Forward look

- **The Tier-1 connection to Janet should be served by the on-going upgrade planning for the RAL campus Janet connection, which is looking at 60 Gb/s and even 100 Gb/s later**

Tier-2

The Table 1 below summarises, for each Tier-2, the Institute connection to Janet, and the Tier-2 site connection itself. Comments have been solicited from all site admins and are also shown in the table. Current Tier-2 connections range between 4-20 Gb/s.

Forward look

- In the next two years, the larger Tier-2 sites (Glasgow, Manchester, Lancaster, QMUL, ICSTM, RAL) are set to require at least ~ 20 Gb/s connections for Tier-2 traffic alone within the next 2 years, and some may even require 40 Gb/s (ICSTM already has 30 Gb/s).

- We expect the medium and smaller sites to require between 5-10 Gb/s for Tier-2 traffic. Note there are some “medium” sites that actually have a high network I//O rate and connection similar to larger sites.

- ATLAS and CMS planning calls for some major Tier-2 sites to require 40-100 Gb/s. GridPP does not call for this at this stage but will keep it under review.

The UK sees no imperative to use LHCONE in order to meet the requirements of WLCG at present. However, these capabilities are very likely to become essential in order to meet the requirements of new communities and GridPP sites continue to work to join LHCONE in order to understand any associated issues.

	Janet Connection	Tier-2 Connection	Forward look	Comment from Site Admin (including any routing and firewall bypass arrangements, DTZs …etc)
UKI-LT2-Brunel	4x10 Gb/s	2x10 Gb/s	Could be upgraded to 30 Gb/s for Tier-2 if required	Fabric, switches and firewall are ready for 80 Gb/s.
UKI-LT2-IC-HEP	2 X (2 X10) = 40 Gb/s	Shared but can can use up to 35 Gb/s of the 40.		
UKI-LT2-QMUL	2 x 20 Gb/s	20 Gb/s dedicated, but shared in case of failure of college link.	We have indicated to central IT that we would like this upgraded to ~80Gb/s for LHC Run3	College is part of a project to develop improved data transfers with RAL for non GRIDPP traffic.
UKI-LT2-RHUL	2 X 10 Gb/s	We currently have our own dedicated 10 Gb/s with an additional 10 Gb/s link for passive failover.	Capacity may be doubled.	Firewall protecting our NAT'd worker nodes. Public network interfaces such as storage bypass this firewall.
UKI-LT2-UCL	2 x 10 Gb/s	10 Gb/s to the UCL core		
UKI-NORTHGRID-LANCS-HEP	2x10 Gb/s	1 x10 Gb/s dedicated during normal operation (it is University backup)	The University is looking at going to 2x(2x10Gb) links, so we would get the 2x10Gb backup link to ourselves in normal operation.	Talk of putting the grid site in a simple border firewall pass through (the motivation being as much to reduce stress on the firewall).
UKI-NORTHGRID-LIV-HEP	2 X 10Gb/s	2x20 Gb/s then shared onto 2 x 10	Hope to improve bandwidth towards10 Gb/s through development work and collaborations with central services.	Realistic T2 throughput is 4 to 5 Gb/s. By-pass arrangements. There is no University firewall bypass in-place at present, but our traffic is subjected to less filtering than normal.
UKI-NORTHGRID-MAN-HEP	2 x 10 Gb/s	1 x 10 Gb/s dedicated direct to NNW	Intend to enable 2 x 10Gb/s for production	
UKI-NORTHGRID-SHEF-HEP	1 x 10Gbps link for LHC / Grid / HPC	1 x 10Gbps uplink.	No bandwidth upgrade is planned for LHC / Grid / HPC activities.	
UKI-SCOTGRID-DURHAM	1 x 10 Gb/s	1 x 4 Gb/s shared	Internal requirement for at least a 10Gb/s. Discussions are ongoing	Janet connection speeds are being discussed with DiRAC and Durham University. All traffic is outside of the University firewalling as it was impacting performance.
UKI-SCOTGRID-ECDF	2X10 Gb/s	2 X10 Gb/s at ACF / shared	Plans to increase internal network to 4x10Gb/s. Possible increase	Ongoing studies into the use of our existing connection and impact of traffic from national machines (notable RDF) before

			of Janet connection to at least 4x!0Gb/s	any formal request for increase
UKI-SCOTGRID-GLASGOW	2X 10 Gb/s	10 Gb/s shared	Maintain 10 Gbs/s link potential upgrade with new DC	Site moving to a new Data Center in 2018 where new hardware may allow upgrade. Current traffic bypasses firewalls with a minimal set of ACLs to restrict certain protocols.
UKI-SOUTHGRID-BHAM-HEP	1 x 10 Gb/s	1 x 10 Gb/s dedicated bypass to SJ6	2x10 Gb/s almost in place.	
UKI-SOUTHGRID-BRIS-HEP	3 X 10 Gb/s (2 back up)	1 Gb/s shared, others on 5 Gb/s shared (throttled)	University planning to upgrade to 20 Gb/s with T2 using more than the current 5 Gb/s .	
UKI-SOUTHGRID-CAM-HEP	2 x 20 Gb/s (@2x10)	10 Gb/s shared	Campus backbone to be upgraded to 3x40Gb/s this calendar year. Janet connection will upgrade when required	We have a bypass in the campus IDS/IPS to avoid limiting the throughput on individual transfers. Even with planned campus backbone upgrades an individual transfer will be limited to 10Gbps.
EFDA-JET	2 X 1Gb/s	0.1 Gb/s		
UKI-SOUTHGRID-OX-HEP	4 x 10 Gb/s	1 x 10 Gb/s dedicated		University has 20Gb/s (Actually 2*10) with 20Gb/s failover. Grid site has to share the normal routing so in effect is sharing a 10Gb/s link for any single transfer.
UKI-SOUTHGRID-RALPP	Covered by RAL network description.	2x40Gb/s Connection to RAL Core/Tier-1 2x10Gb/s Firewall bypass connection (active/passive pair)	Upgrade firewall bypass to 2x2x10Gb/s or 2x40Gb/s Looking at joining LHCONE with RAL Network team and Tier-1	Direct firewall bypass for a /25 subset of the /22 Grid VLAN.
UKI-SOUTHGRID-SUSSEX	1 x 10 Gb/s	5 Gb/s Shared use of 10 Gb/s		

Table 1: Network connectivity to Janet and forward look from each Tier-2 site.

IPv6 readiness

The IPv6 status of each Tier-2 is shown in the table below.

	IPv6 Ready	Description/Comments
UKI-RAL-Tier1	Yes	Tier-1 is running IPv6 on production network Services already in dual-stack mode: squid, CVMFS Stratum1, CMS xroot redirector Services to go dual stack by end of 2017: FTS, Echo, Frontier, GOCDB. Other services early 2018 All hosts dual-stack by default by April 2018 Tier-1 IPv6 data to Janet via bypass at 10Gbs Tier-1 IPv6 data to CERN T0 via OPN using standard link Tier-1 other IPv6 to site via 10Gbs (dedicated failover pair) Tier-1 other IPv6 to Janet via firewall
UKI-LT2-Brunel	Yes	Brunel has been on dual stack for nearly 4 years.
UKI-LT2-IC-HEP	Yes	IC is fully IPv6 dual stack.
UKI-LT2-QMUL	Yes	IPv6 connectivity to all outward facing GRID services (SE, GRIDFTP, Webdav, xrootd, CEs, SQUIDs...) . Plan to extend to worker nodes in the future
UKI-LT2-RHUL	No	
UKI-LT2-UCL		
UKI-NORTHGRID-LANCS-HEP		
UKI-NORTHGRID-LIV-HEP	Ongoing	Capability established, tranche of addresses provided, technically tested. However, link very slow. Cluster room refurb is our priority until Sept; after that, we'll look for bandwidth improvements, and a gradual migration to IPv6.
UKI-NORTHGRID-MAN-HEP		
UKI-NORTHGRID-SHEF-HEP	No	The University of Sheffield does not currently run IPv6 on its Internet links due to hardware/software limitations on current border routers, which are unlikely to be replaced in 2018. If IPv6 is requirement for GridPP, then a separate, IPv6-only link could possibly be sourced from Janet (at additional cost), but this would also need separate firewalling hardware, (again, at extra cost). If IPv6 is requirement, but high-bandwidth isn't, the University could ask Janet to provide IPv6 link over a single 1Gbps link as was done previously. Firewalling this 1Gbps link would still need consideration.
UKI-SCOTGRID-DURHAM	Partial	IPv6 is available and usable, waiting for IPv6 reverse DNS support from the University for full implementation.
UKI-SCOTGRID-ECDF	No	
UKI-SCOTGRID-GLASGOW	Has an IPv6 allocation	Site has an IPv6 allocation, at present does not route IPv6 to production cluster. Rollout of IPv6 is slated to take place with move to Data Centre in 2018
UKI-SOUTHGRID-BHAM-HEP	No	
UKI-SOUTHGRID-BRIS-HEP	We believe campus is IPv6 enabled.	UKI-SOUTHGRID-BRISTOL-HEP lcgce01, & lcgse01 & its I/O server are in IPv6 production as far as we think / know. We're quite sure about the Storage. We *think* the CE is enabled & configured correctly for IPv6.
UKI-SOUTHGRID-CAM-HEP	Yes	Campus has been ready for several years. We have a /64 allocated. Grid site shares the local network with group resources (ie. Linux, Windows and MAC) and so there is some concern about what will happen when IPv6 is turned on.
EFDA-JET		
UKI-SOUTHGRID-	No	Some test systems with limited bandwidth. Awaiting University

OX-HEP		upgrade.
UKI-SOUTHGRID-RALPP	No	IPv6 has now reached the routers above the Tier 2 router. We have requested an IPv6 allocation and developed an addressing plan. We will meet with networking soon to develop a plan to roll it out on the Tier 2.
UKI-SOUTHGRID-SUSSEX		

Table 2 : IPv6 status of each Tier-2 site.