DESIGN OF A NEW NETWORK INFRASTRUCTURE USING ROUTING CONTROL PLATFORM FOR THE UNIVERSITY OF HUDDERSFIELD CAMPUS GRID

YVONNE JAMES

A thesis submitted to the University of Huddersfield in partial fulfilment of the requirements for the degree of Masters by Research

The University of Huddersfield

Submission date as January 2013

Copyright statement

Abstract

Research across the University of Huddersfield has increased in terms of number of users utilising the cluster systems. These systems provide a valuable service allowing high speed, complex calculations and graphics rendering to be performed.

The clusters have grown out of user requirements and the availability of computing and networking equipment. In the early days very little consideration was given to the placement of equipment to fulfil these needs. As a result the speed of data transmission was slow and often resulted in bottlenecking and/or data loss through protocol timeouts. An increase in bandwidth has helped to improve this, but more importantly the allocation of space within the data centre has enabled a significant increase in bandwidth, as well as providing a direct connection to the network backbone.

As a result of our work the network topology has changed and has reduced the number of switches and routers packets traverse across the network from end user to cluster. The knock-on effect is a reduction in data transmission times, a reduction in the possibility of bottlenecking and much reduced data loss.

Table of Contents

List of Figures

Dedications and Acknowledgements

I would like to thank my supervisor Dr Violeta Holmes, for all her support throughout this year as well as her infectious optimism and belief in my abilities.

I also need to thank my dear friends in the HPC Research Office, who provide unpaid and undying support for the High Performance Computing at the University of Huddersfield. They are an incredible bunch of people who not only offer intellect and a much needed sense of humour, but also some amazing DIY skills: John Brennan, Stephen Bonner, Shuo Liang, David Gubb, Ibad Kureshi, Matthew Newall. Also my dear friend and colleague, Andrew Fear, who has provided technical insight into some often quite difficult aspects of network technologies.

Many thanks also to Dr Paul Elliott and Dr David Cooke, who have not only provided financial assistance, but continue to be strong supports of HPC at the University of Huddersfield, and without whom most of the work undertaken by the HPC Research Group, would become difficult to achieve.

A special thanks also to Joanna Radley, Darren Shaw, Paul Harriman, Sarah Gullick and Steve Bekus, who are key members of the Computer Services team and have provided incredible support for the HPC project.

I also need to thank my family, husband Ian and my son, Alex, who have shown great patience and allowed me to write this thesis in peace and quiet. Plus a mention to my two dogs for their patience in waiting for a walk when I wanted to write instead.

Finally, I would like to dedicate this work to my parents, for supporting me through this year and for instilling in me the importance of study for a better life.

List of abbreviations

BGP	Border Gateway Protocol
DHCP	Dynamic Host Configuration Protocol
DMZ	Demilitarised Zone
DNS	Domain Name Service
EIGRP	Enhanced Interior Gateway Routing Protocol
HPC	High Performance Computing
IP	Internet Protocol
NAS	Network Area Storage
NIC	Network Interface Card
OSPF	Open Shortest Path First
RFC	Request for Comment
TCP	Transport Control Protocol
UDP	User Datagram Protocol

Introduction

Since 2008 the University of Huddersfield research community has increased. There is considerable research work being undertaken in a number of areas including particle physics, fluid dynamics, 3D imagery, all of which require the use of high performance computers (HPC) to manage complex instruction sets.

As the needs of the university for high performance computing increases, the network infrastructure must be able to support this. This in turn means that the network infrastructure must be able to provide high speed data transfer rates using media that is capable of supporting large data and file sizes. The university will need to provide support to further improve the network infrastructure.

In order to inform the HPC network infrastructure design and development at the University of Huddersfield, it was necessary to analyse the current architecture and define new network design.

In the first instance a requirements analysis was carried out to determine the required network devices and needs of the end users. Once identified the roles of these devices was determined in line with design models and operation.

This work considers aspects of network design and installation in the context of deployment of high performance computing clusters in the campus grid at the University of Huddersfield. It identifies the current network infrastructure and determines changes necessary to improve the flow of data traffic across the grid for new cluster systems integration. In order to understand how best to utilise the existing network infrastructure, cluster installations at other universities are examined and compared. While this exercise will not compare like-for-like, it will provide comparative data to show utilisation, throughput and data transfer speeds, as well as the different types of network infrastructure, devices and cabling used.

In this report each Chapter deals with different aspects of the network design and implementation, identifying issues and proposing solutions in respect of the infrastructure installation. Chapter 1 outlines the existing infrastructure and issues with initial network design. Chapter 2 documents new network design and proposes various devices for installation of new clusters. In Chapter 3 a number of network protocols are evaluated along with the relationship with network devices and the issues which need to

be considered as the installation progresses. Chapter 4 considers other work in network design including some interesting work on data centre topologies. In Chapter 5 we examine the University of Huddersfield network infrastructure in more detail and consider the impact of a new design. Chapter 6 details the network devices and their modes of operation. Network performance issues are considered in Chapter 7 and the effect of poor network design on data transfer. Chapter 8 examines the issues surrounding the installation as well as those that will affect future direction and network expansion. Chapter 9 demonstrates the effectiveness of the new design and installation, showing how different network monitoring tools have been employed to obtain statistical and graphical information regarding the network performance.

Chapter 1 – Problem Outline

At the University of Huddersfield the HPC systems currently rely on the network infrastructure as provided across the campus by university Computer Technical Services. As HPC has grown out of various research projects conducted over the past four years, usage by researchers in the University has increased. Initially, there were only a handful of users and the number has substantially grown so that in 2012 there are over one hundred HPC registered users. The location of the HPC equipment initially was less important than a provision of HPC to satisfy emerging research needs. The rapid expansion of diverse research programmes conducted across the University has resulted in more reliance being placed on the HPC systems. As the HPC resources were added to the existing system, network infrastructure was not considered. This has resulted in all data being transferred across the main network and through the data centres to the HPC Resource Centre with an inadequate bandwidth.

The current campus network infrastructure allows students, staff and researchers to access the HPC and high throughput computing (HTC) clusters in the university campus grid from anywhere on the campus. There are also facilities to access the cluster services remotely, provided that end users are able to guarantee a static IP address which can then be added to the hosts file.

As a result of this rapid growth location and access to the existing HPC resources and a lack of network planning and management, the current network infrastructure poses a number of problems.

The main area of concern for this work is how to improve existing network infrastructure and achieve faster data transfer of large files between end-user and clusters across the University network. This requires the following:

- Identifying what current technology is in place,
- Understanding how this technology works including benefits and limitations,
- Considering how to best use the existing network infrastructure to support the existing HPC and HTC clusters, and
- Identifying the best possible location to install new clusters to achieve improved network performance.

In addition there are facilities to access other geographically remote systems which are part of national e-Science infrastructure. These provide an interesting case of data transfer across the internet as opposed to over the campus network, which has identified some other issues.

The aim of this study is examine the elements required to implement a new cluster within an existing network infrastructure.

The objectives were to:

- Gain an understanding of the technologies involved - hardware and software.
- Identify an effective network design to ensure best possible performance given the parameters for installation.
- Examine need for effect network management and load balancing
- Understand the relationship between the local and remote sites

1.1 – Current HPC network Infrastructure

The University of Huddersfield is located on the outskirts of the city centre. The campus extends beyond the canal and utilises the old mill buildings to house part of the School of Computing and Engineering. The campus is made up of a number of different buildings spread across a considerable distance. IT Technical Services are located in the Central Services building. Both the data centres and Canalside East, which currently houses the clusters in the HPC Resource Centre (HPC RC), are located at the opposite end of the campus. Data transferred by end users to the clusters is transmitted across the campus network infrastructure via various buildings and through numerous switches and routers. While these devices are necessary, they also provide an additional point of failure across the network. The more devices the data has to travel through, the slower the process becomes. This is shown later in an analysis of the network performance in Chapter 9.

1.2 Existing Campus Network Infrastructure

The existing network infrastructure provides network services across the Queensgate campus including connections to the Janet network for Internet access and links to other clusters through the North-West Grid (www.nw-grid.ac.uk) and STFC Daresbury Laboratory (www.stfc.ac.uk). The existing clusters reside on the university campus as shown in Figure 1. There are two proposed sites for the new cluster installation; one in the School of Computing and Engineering (Canalside East building) and the other in the data centre. An installation in the school building will result in a number of performance issues, some of which are already experienced by the existing clusters in HPCRC and are beyond the scope of this work. Because of that, it was decided to consider installing the new cluster in the data centre.

The data centre houses network equipment and computer systems that provide services across the campus and form an integral part of the university network. A large part of the work we have undertaken has been to examine an installation of the new cluster named Sol, within Computing Services' data centre, thereby bringing together the existing university network with the new cluster system.

It is important to understand the relationship between the existing clusters and Sol in terms of location and operation. The current setup allows users to access any of the clusters via Bellatrix server (see Figure 1), which is responsible for managing authorisation, authentication and accounting (AAA).

Due to the concerns outlined above, the new installation could not form a part of this setup, but could operate independently, including AAA. Jobs are assigned to the relevant server where the appropriate application is running, or to the three independent servers which operate as servers supporting virtualization (Shavla, Sargas and Spica) as shown in Figure 1. There are direct links from the clusters to the university network and to other geographically remote sites. With the exception of Bellatrix, the virtualization servers will remain resident in their current location (HPCRC). These machines are only used to spawn virtual images and allow for testing in alternative operating system environments.

Figure 1 also shows the relationship between the current high performance computing system which is located in Canalside East building, and the link to the university staff

network, which in this instance, provides a link across the campus. This topology is haphazard and means that data are passed back and forth between different systems using a 100Mb link. In the past, this low data transfer rate has caused performance issues and created bottlenecks, resulting in a data loss as well as jobs failing to run correctly on the clusters. There are also a number of switches in this topology which, when considering the routes taken by data around this segment, have the added effect of causing latency issues.

Our work aims to propose a new infrastructure which will overcome these problems by redesigning the network infrastructure to provide a faster link, reducing any latency and bottlenecking issues. This new network infrastructure will provide better access to the university network, and enable users to transfer data much faster to both the university clusters and to other remote sites. In essence, this segment is completely re-engineered to provide a more scalable and resilient network.

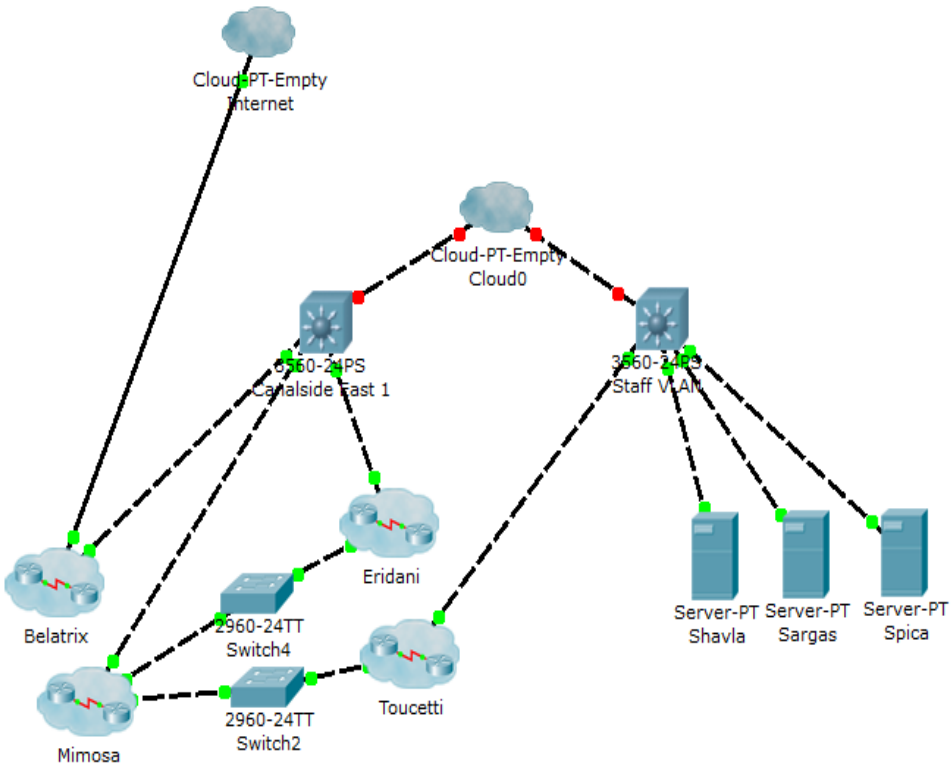


Figure 1: Cluster setup showing relationships with campus VLANs and external network connections.

Figure 2 shows the topology implemented in the data centre. This new topology offers an effective alternative. The clusters are located within the data centre, so rather than

the data being transferred around the building through several different switches, only two switches are utilised. This reduces the time taken for data to be transferred and helps to eliminate issues with poor network performance, more specifically latency and bottlenecks. These switches are further explained in Chapter 6 Network Devices.

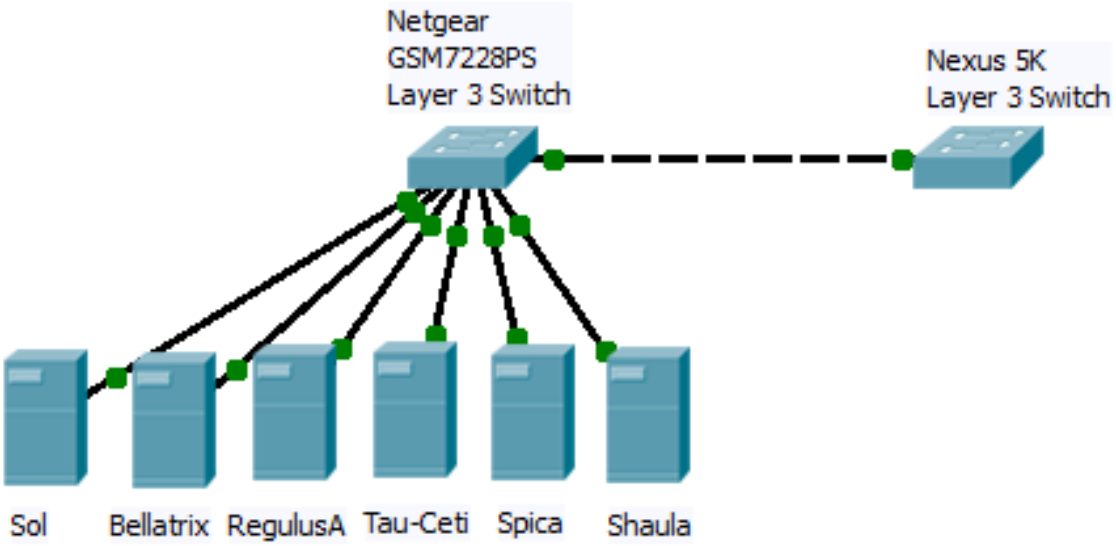


Figure 2: Proposed installation for Sol cluster

Chapter 2 – Proposal for installation of new cluster - Sol

In order to integrate the Sol cluster into the campus grid, the infrastructure needs to be examined to determine the best possible location, network devices and cabling. The servers and layer 3 switches (Netgear and Cisco) were purchased for the cluster installation. It is important to understand how these particular layer 3 switches will fit with the existing infrastructure and how they will be utilised within the cluster architecture. This section describes the operation of all the network devices involved in this installation The server setup, operating systems and application software were not considered.

2.1 Network Devices

Originally Sol cluster was to be housed and displayed as the centre piece of a new visualisation suite. Issues around green computing, power usage and speed of data transfer led to a new network design. This design ensured sufficient power to run the cluster and provide high speed network connectivity. However, this work is not concerned with green computing and power consumption, but concentrates on the network design element, including high speed data transfer rates, increased network performance and a reduction in transmission errors.

The equipment to be installed were as follows:

- Cisco Nexus 5k
- Cisco 3750
- Netgear GSM7248 layer 3 smart switch
- Nortel Baystack 5510-T layer 3 switches
- Cabling

The Nexus and Cisco 3750 are already in place, provided by the Computer Services department.

The Netgear GSM7248 is a layer 3 switch that will provide connectivity between the Cisco Nexus and the Nortel switches.

The Nortel switches connect the cluster nodes together, providing high speed interconnects for high speed data connectivity between the servers and the scratch storage. Each interconnect offers 40Gbps between two nodes.

2.1.1 Cisco Nexus 5k

This access layer switch was specifically designed for use in data centres. It provides:

- Multi-protocol support
- High performance, low latency 10Gb Ethernet using cut-through switching architecture
- Connectivity options for both Ethernet and Fibre Channel

This switch provides the high speed connection form the clusters directly into the network backbone. Latency is reduced by utilising cut-through switching architecture. Network performance is increased as a result of the 10Gb connection, which in turn means that end users are able to send large file sizes from their desktops to the cluster for processing.

2.1.2 Cisco 3750

These are stackable layer 3 switches, which offer both switching and routing technologies to forward data. These switches are located in many of the buildings around the campus and provide connectivity for incoming traffic at 10Gb and traffic internal to the building at 2Gb.

Routing data across the network uses the Open Shortest Path First protocol (OSPF), which is explained in more detail in Chapter 3 Network Protocols. This is the preferred protocol although this switch also supports Enhanced Interior Gateway Routing Protocol (EIGRP), Border Gateway Protocol (BGP) and IPv6.

The Cisco 3750 is able to offer options for Fibre Channel, Ethernet and wireless.

2.1.3 Netgear GSM7248 Layer 3 Switch

This switch provides services in relation to both layers 2 and 3 of the OSI 7 layer model. Switching happens at layer 2, while routing is done at layer 3. These switches have the ability to perform both operations as required by the needs of the network. There are two specific reasons for selecting this device. The first of these is cost. Netgear provide high specification technical devices at a reasonable cost. In order to fulfil the requirements of this network an uplink module and SPF+ module has also been purchased so that the cluster can utilise the 10Gb network backbone directly.

In respect of routing, this device is able to use OSPF and could, therefore connect straight into the university network. OSPF is discussed further in section 3.6.

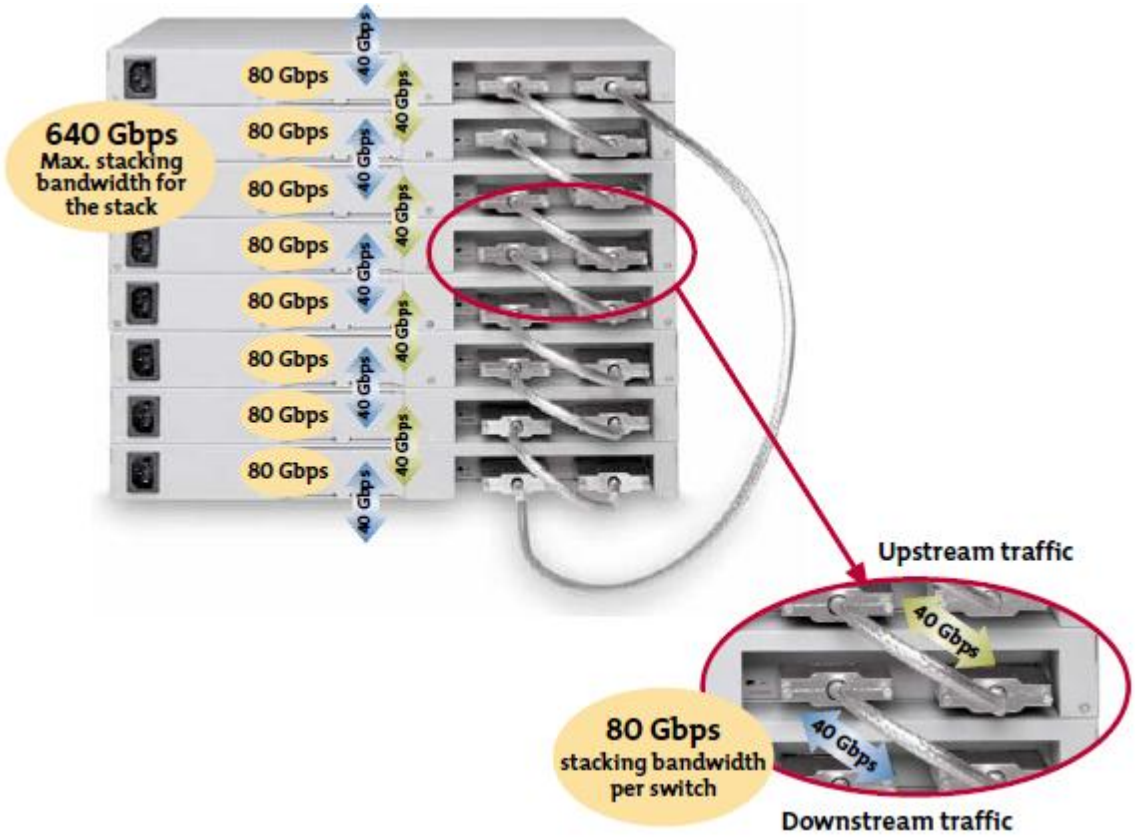


Figure 3: Interconnects providing 80Gbps stacking bandwidth per switch

The Nortel Baystack is defined in the specification as “BayStack 5510 switches are 1 rack unit high stackable 10/100/1000 Mbps Ethernet Layer 3 routing switches that are designed to provide high-density Gigabit desktop connectivity for mid-size and large enterprise customers’ wiring closets.”

These switches use stacking to provide 80Gbps traffic flow to adjacent units as shown in Figure 3. Once the cable is attached the flow of data is automatically identified by the switch and needs no configuration.

The Baystack is capable of managing 10/100/1000Mbps as well as providing 1Gb uplinks to the backbone.

The major issue with these devices is that Nortel were taken over by Avaya when the business folded. As a result Avaya consolidated their own products with Nortel’s technology and some devices were no longer manufactured. This particular layer 3 switch falls into this category. This means that there is no longer any support for Nortel

devices and parts are very difficult to find, although complete units are still available for purchase. In turn this means that the link from the Nortel switch to the Netgear switch operates at 1Gb as the Nortel cannot be upgraded.

2.1.5 Cabling

In order to connect the cluster to the main network a 10Gb fibre optic cable has been purchased. This ensures that Sol cluster maintains a high speed connection to the main network. At each end an SPF+ module is required to connect the fibre.

After the decision was made on the network devices for the installation of Sol cluster, various network protocols were examined for suitability.

Chapter 3: Network Protocols

This chapter provides an overview into major protocols that are implemented across the university network, with the exception of Enhanced Interior Gateway Routing Protocol (EIGRP). A comparison with the preferred protocol is made. In order to understand the relationship between the way data is transferred across the network and the protocols involved, it is necessary to understand the OSI 7-layer model, which is discussed below.

3.1 OSI 7 Layer Model

The Open Systems Interconnection or OSI 7 layer model is a layered framework that helps to explain communication across the network as represented in Figure 4. For the purpose of this work layers 1-3 are largely ignored. Layers 4 to 7 are responsible for network addressing and data transfer, and are key to successful data transfer. The process of encapsulation is employed to ensure the data is addressed in the appropriate manner to ensure delivery. As data is transmitted across the network it appears in different forms as identified in Figure 4.

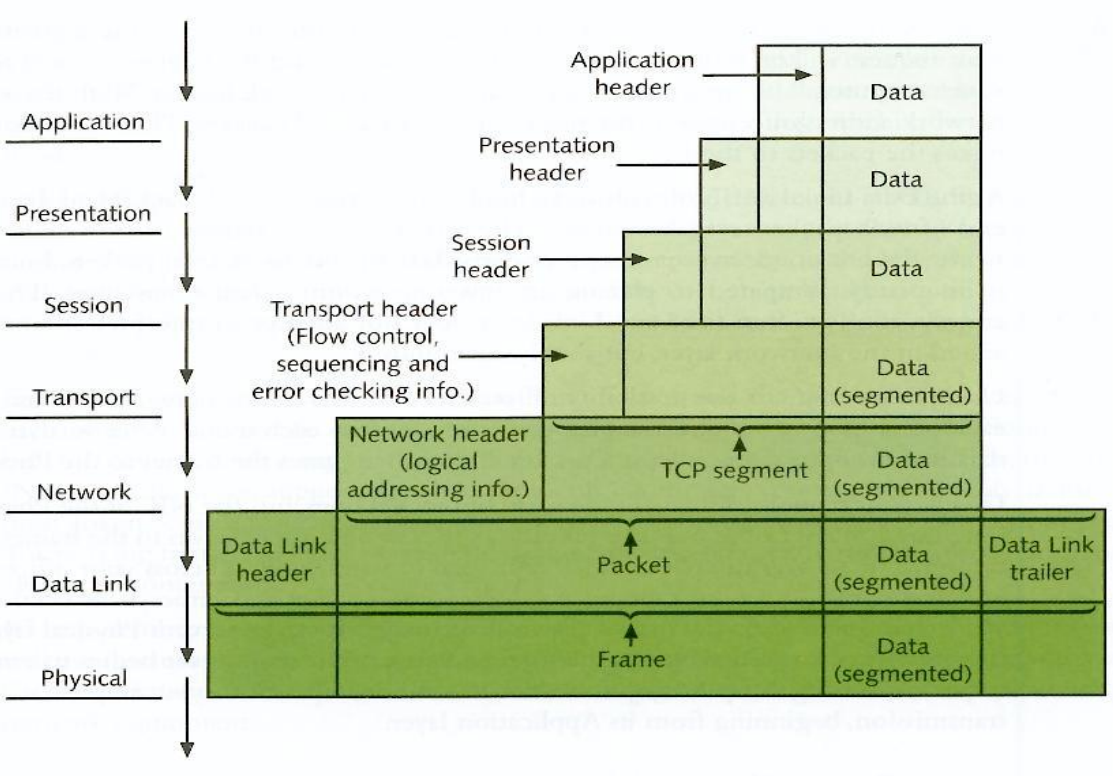


Figure 4: Encapsulation/Decapsulation process through OSI 7 layer model (Source: http://elguber.wordpress.com/2010/05/26/osi-model 2012)

The Transport, Network and Data Link layers are especially important in networking as these layers provide the mechanisms for other devices to understand where the data should be delivered to. The TCP header is added at the Transport layer and the IP addressing information is added at the Network layer. Also the data is segmented into smaller chunks known as segments if TCP is used or datagrams if UDP is used. These small chunks are vital to successful data transfer. This process is discussed later on with TCP. As the network interface card resides at the Data Link layer the MAC address information is added at this point. Figure 5 shows the addition of headers for each layer as the data travels through the OSI 7 layer model. So at layer 2 there is the addition of L2H – Layer 2 Header, plus L2F – Layer 2 Footer.

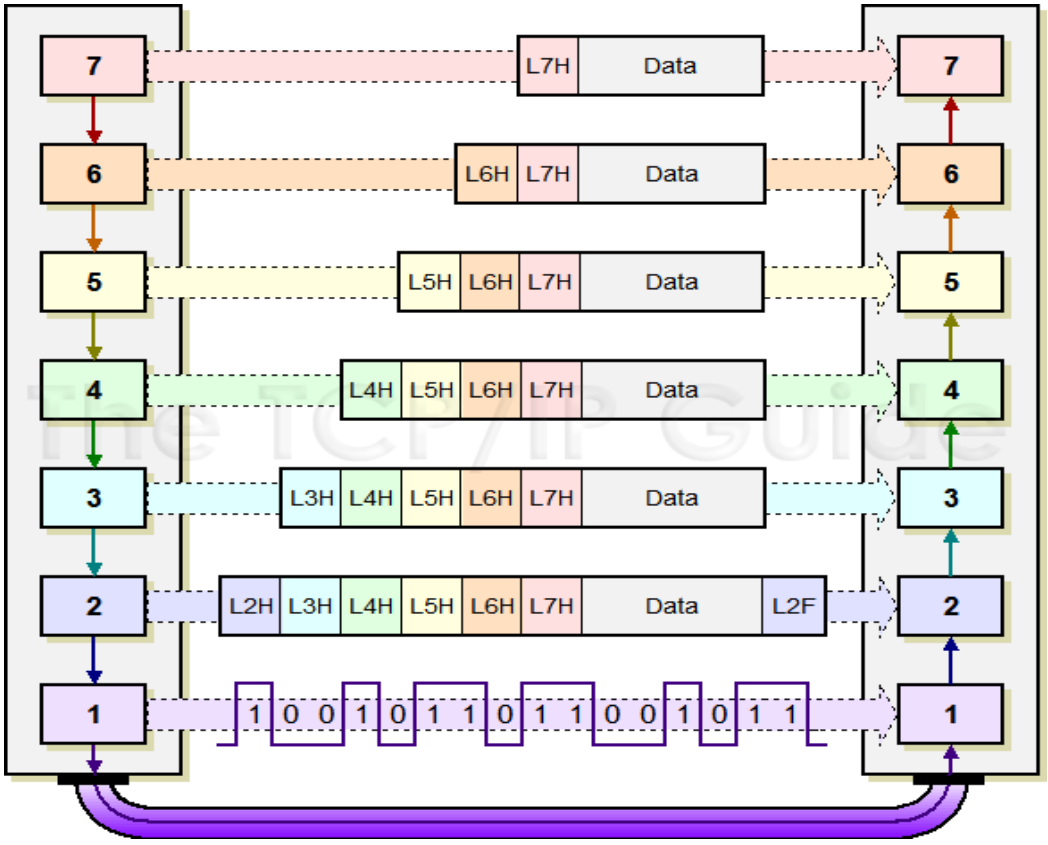


Figure 5: Encapsulation/Decapsulation process through the OSI 7 Layer model (Source: http://www.infocellar.com/networks/osi-model.htm, 2012)

3.2 TCP – Transmission Control Protocol and IP - Internet Protocol

As previously discussed TCP exists at the transport layer of the OSI 7 layer model. Specifically the transport layer provides reliable data transfer service to the upper layers of the OSI model. This reliability is managed through flow control, segmentation/desegmentation and error control. TCP is also connection orientated so the transport layer is able to track segments and retransmit any failed packets. (Dye, M.A. et al. 2008) TCP is a connection orientated protocol which uses a three-way handshake to establish a stable connection with the destination device. In order to achieve this increased functionality additional overheads are placed on the network by increasing the number of packets sent and the time taken to send them.

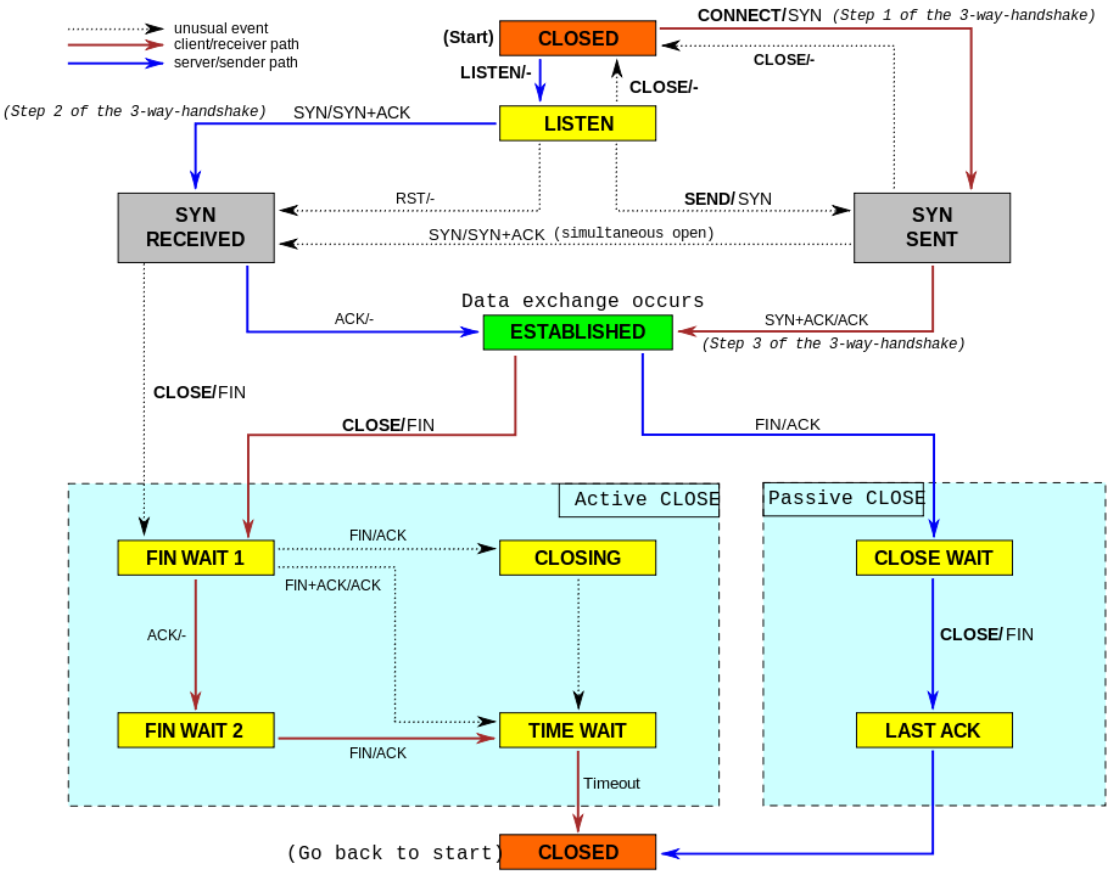


Figure 6: TCP three-way handshaking process

(Source: http://commons.wikimedia.org/wiki/File:Tcp_state_diagram_new.svg, accessed July 2012)

Essentially there are six operations that TCP provides as detailed in IETF RFC793. These are outlined below:

- Basic Data Transfer
- Reliability
- Flow Control
- Multiplexing
- Connections
- Precedence and Security

3.2.1 Basic Data Transfer

In accordance with RFC793 a major function of TCP is to provide a continuous stream of bi-directional octets. It does this by providing an end-to-end connection which establishes a reliable service for data transmission.

3.2.2 Reliability

TCP has the ability to recover from duplicated, lost or damaged data, which it does by using a handshaking process and acknowledging the arrival of packets. Damaged packets are requested to be sent again.

3.2.3 Flow Control

This is provided by using a mechanism called windowing. The window determines the allowed number of octets that the sender may transmit before the transmission is acknowledged. This process is repeated until all the data is transmitted and received successfully.

3.2.4 Multiplexing

TCP provides a set of ports used to identify which application the data is related to. Commonly used ports occupy 0-1023.

To help identify a particular connection the port number is added to the IP address creating a socket, for example 192.168.10.45:80, where 80 is the port number for HTTP. Multiple sockets can be created by one machine at any one time.

3.2.5 Connections

This relates to the previously mentioned functionality of TCP to establish and maintain a reliable connection using a combination of information – sockets, sequence numbers, and window sizes. Once the data transfer is complete the process is then able to terminate the connection thereby freeing up resources for other uses.

3.2.6 Precedence and Security

TCP is not instrumental in providing either of these functions, unless the user determines that they should be used. If not the default values for TCP are used. For example Telnet requires a lower level of security than SSH does. TCP is able to function well with both protocols and allow for the relevant level of security to be established.

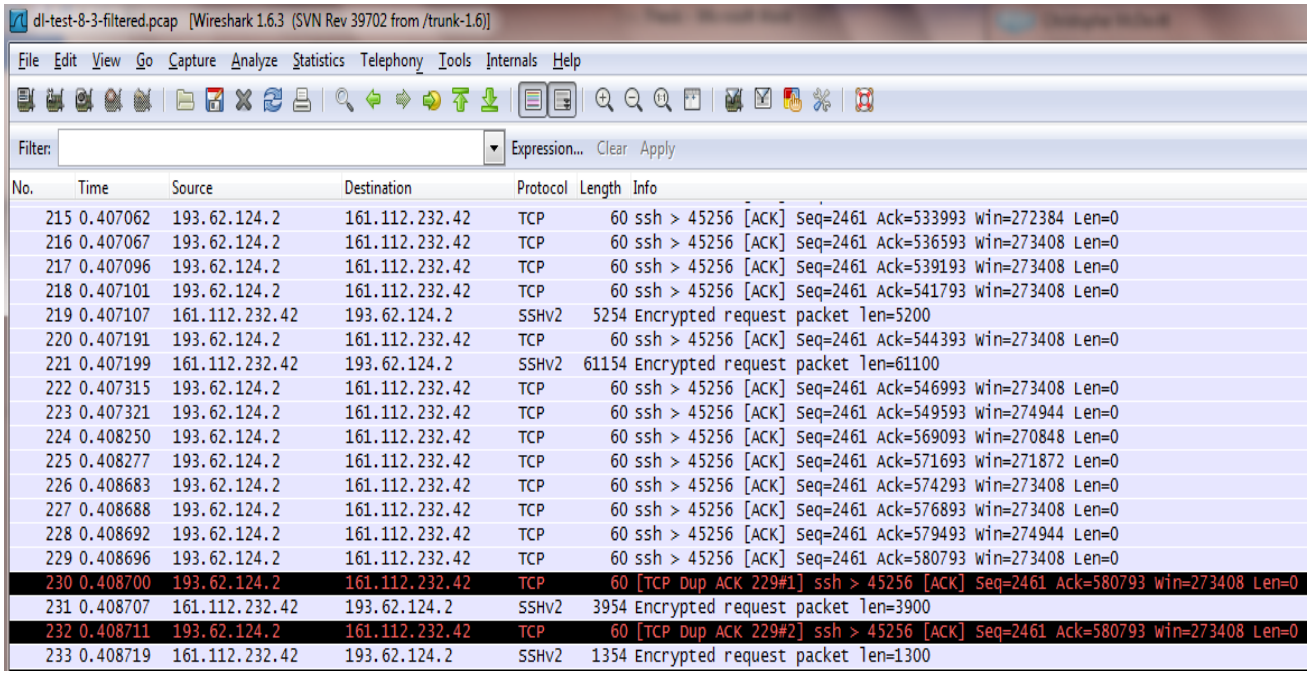


Figure 7: Wireshark (www.wireshark.org) data showing protocol exchange, windowing, sequence numbers and identifying duplicate acknowledgements (ACK).

3.3 Disadvantages of TCP

TCP has some drawbacks in that because it is connection orientated it is not suitable for the transfer of video, IPTV, VoIP, or gaming data. There are other protocols to support these media types for example UDP – User Datagram Protocol, which carries no overheads and is not connection orientated. Data sent to the clusters is requesting the use of an application which utilises TCP to ensure a reliable data transfer. Often this data is critical to the research of the end user, so any failure to transmit would cause the data set overall to become corrupt and invalid.

The reliance on the three-way handshake to establish and maintain a connection also increases the overheads on the network. Using the technique of windowing and sequence numbers, means that for every window filled with data an acknowledgement or receipt of that data transmission is sent to the sender. In turn this slows down the speed of data transfer. Network performance and speed of data transfer are sacrificed for reliability.

3.4 Internet Protocol - IP

IP is defined in RFC791. It resides at the Network layer of the OSI 7-layer model and is responsible for managing IP addressing information. IP is a routing layer datagram service which is used by all protocols within the TCP/IP suite with the exception of ARP and RARP. IP is used to route frames between hosts on the network. The header, therefore, contains routing information.

The university network is broken into segments, which can be identified by their subnet as well as name. The new cluster is identified as a subnetted class B IP address using a class C subnet mask. This means there are 256 available addresses which can be allocated to 254 hosts, saving 2 addresses for the network ID and the broadcast.

3.4.1 Advantages and Disadvantages of IP

The major disadvantage of IPv4 is that addresses have run out so there are no more addresses which can be used on the internet. To resolve this issue IPv6 has been introduced.

On an internal network this is still not an issue. Administrators find it easier to deal with IPv4 addresses as they are only 32 bit in length. Also they are numerical so are easy to

understand and to configure. By comparison IPv6 is a 128 bit hexadecimal address which is not easy to configure as a result of this complexity.

Under normal circumstances network administrators rely on DHCP services provided by operating systems to allocate IP addresses to network devices. As this new installation has a direct connection to the main network the DHCP server has been disabled in favour of using the computing services own DHCP server. In this respect the allocation of IP addresses has been further centralised, thus integrating the clusters more into the university network.

In the case of remote users the clusters are reliant on static IP addresses provided by the end user to identify computers that are external to the University network. The end user must obtain a static IP address from their ISP, which of course is an additional cost to their normal Internet charges. IP traffic is currently in a state of flux as there are some devices using IPv4 and others using IPv6 addresses. Most devices are now able to accept both IPv4 and IPv6 addresses. Once the transition is fully made and all devices are using IPv6 the input process will become a long winded and tricky one. Unlike IPv4 which is numerically based, IPv6 is a 128-bit addressing scheme which uses hexadecimal.

A move to IPv6 is likely to be facilitated by two aspects. The first of these is that all network devices are able to support IPv6. There are currently some devices in use that do not provide support for this version of the protocol. The second aspect is that as the university network grows the number of devices requiring IP support will continue to expand and the number of IPv4 addresses will be exceeded. For the most part this is easily addressed by implementing IPv6. However, there are devices which provide remote services that will need to be manually configured and this in itself will cause numerous other issues.

The implementation of IPv6 is outside the scope of this work, but offers an opportunity for further study in network design.

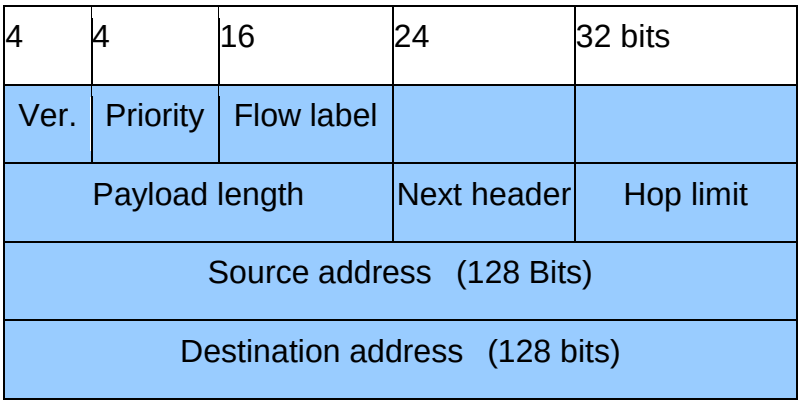


Figure 8: IPv6 Header Structure

3.6 OSPF – Open Shortest Path First

The university utilises OSPF for data communications across the network. OSPF is part of the TCP/IP suite of protocols. Teare (2010) defines OSPF as “a fairly complex protocol”. It offers complexity through functionality and the needs of the particular network. For example, there may be a requirement for some authentication to take place. OSPF is configurable to manage this process. This protocol was originally designed to work with IPv4 (RFC 2328), but has implemented changes to the structure to provide compatibility with IPv6 (RFC 2460).

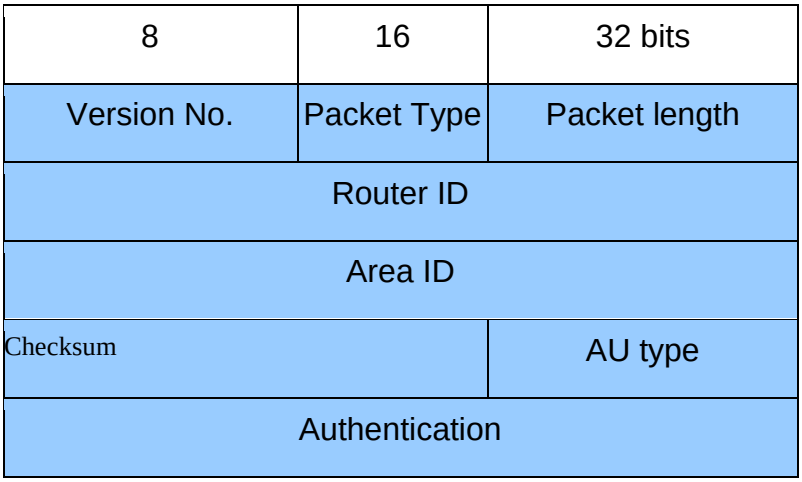


Figure 9: OSPF header structure

RFC 2328 describes OSPF as being “classified as an Interior Gateway Protocol (IGP). This means that it distributes routing information between routers belonging to a single Autonomous System.” (Moy, J. 1998) OSPF relies on sending updates about topology

changes to keep track of available routes across any network. It is based on link-state or SPF technology using Dijkstra's algorithm to calculate best paths. The algorithm specifies that a tree will be calculated of shortest paths and an attributed cost calculated from self to each. The algorithm consists of:

step 0: put (SELF, 0) on tree

step 1: look at LSP of node (N,c) just put on tree. If for any nbr K, this is best path so far to K, put (K, c+dist(N,K)) on tree, child of N, with dotted line

step 2: make dotted line with smallest cost solid, go to step 1

OSPF relies on IP addressing information to route packets across the network without adding any further layers of encapsulation. In this respect OSPF is a dynamic protocol. For example, if the router loses contact with a nieghbour router the link is invalidated within a few seconds and all paths are recalculated. This is done by sending a Hello packet every 10 seconds to the adjacent routers. If there is a response the link is labelled as active. If not the link is checked by sending for consecutive Hello packets without response, known as the dead interval. The link is then identified as being inactive and the network topology is changed instantly with a link state advertisement (LSA) to update the topology with adjacent routers. This information is maintained in the Link State Database (LSDB), which is located on the Designated Router (DR). The role of the DR is to synchronise with all other routers on the same autonomous system when any topology change happens.

OSPF understands the value of a link. 56Kbps serial link has a value of 1785, whereas a Fast Ethernet or FDDI (Fibre Distributed Data Interchange) both carry a value of 1. The latter would be selected over the former based on this value. Dijkstra's algorithm selects the best path based on the lowest number.

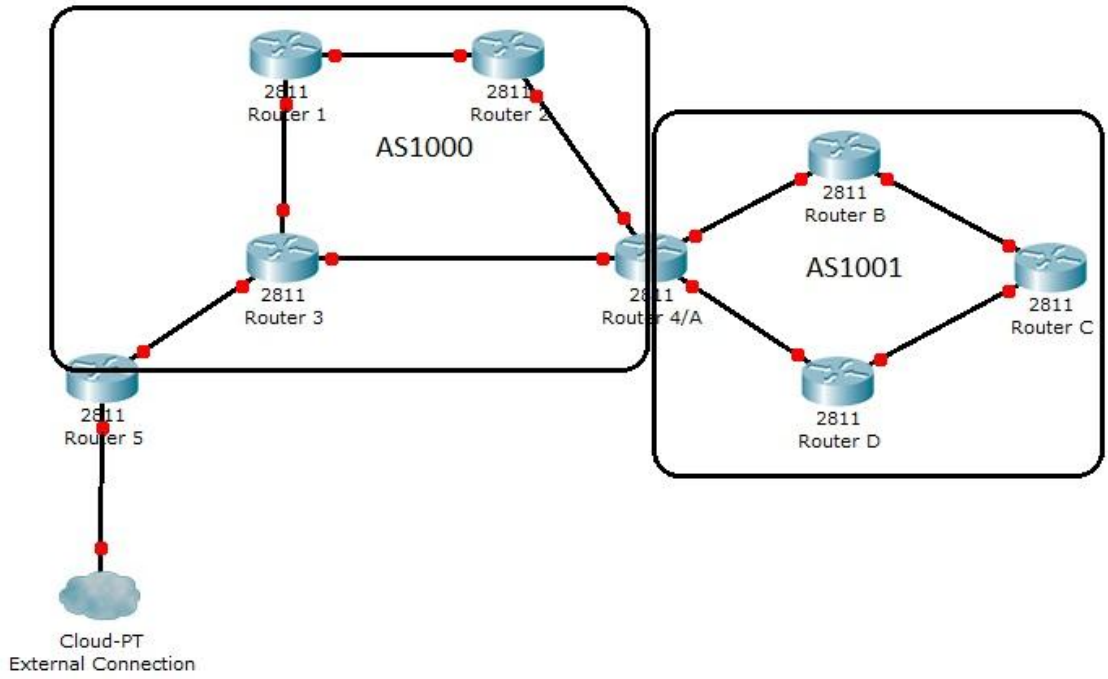


Figure 10: Routers that operate within Autonomous Systems 1000 and 1001

Routers using OSPF maintain a link-state database which shows other devices in the same autonomous system and the routes to them. Using the cost metric routes can be defined or by using equal costs on all routes the traffic can be load balanced.

OSPF offers advantages over other interior gateway protocols in that it is able to handle Classless Inter Domain Routing – CIDR and Variable Length Subnet Masks – VLSM. This allows for less data to be carried in the topology updates, thereby reducing the overhead on the network and causing very little performance degradation.

OSPF uses the Hello packet to determine other adjacent routers and networks. By doing this it is able to build a topology of the network which is then shared with other routers in the same autonomous area. The designated router, the one with the highest priority, maintains the database and sends LSAs to all other routers to ensure standardisation across the network. In addition routers are able to connect external systems and advertise these routes to the routers that need this information. Routers can join two autonomous systems together to allow packets to be exchanged between them. These routers are called Autonomous System Boundary Routers (ASBR).

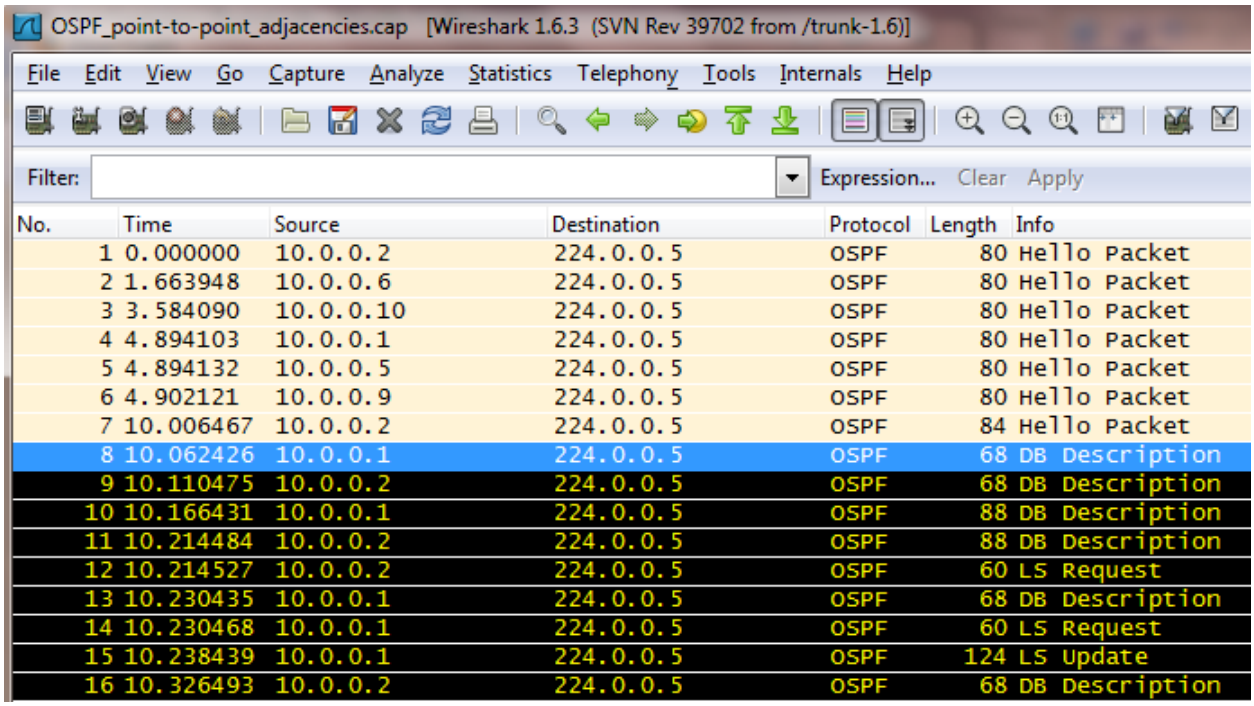


Figure 11: Wireshark capture showing discovery of adjacencies and database updates. (Source: http://packetlife.net/captures/protocol/osp, accessed July 2012)

3.6.2 Advantages and Disadvantages of OSPF

OSPF offers more advantages than it does disadvantages to its selection for use on any network. It has proved to be a stable and reliable protocol, which when set up correctly, has features that are able to reduce overheads such as routing between devices on the same autonomous system. Listed below are the main advantages and disadvantages of reselection.

One of the major advantages of OSPF is that by utilising Dijkstra's algorithm and calculating the best path, any loops that might occur in other protocols, are avoided. Based on cost metrics load balancing can be set up to ensure that all paths through the network carry an equal weight of data traffic.

OSPF is able to accommodate topology changes very quickly. Updates use only a small amount of data so there is little overhead placed on the network.

OSPF is able to communicate across other networks as well as within an autonomous system. Other OSPF routers may also be added to the existing system. As long as they

are within the same autonomous system, convergence will happen quickly and routes recalculated.

Stub area networks, those that do not propagate external advertisements, allow for internal only advertisements to be updated, thereby reducing the amount of data required for updates. In turn this means that routers do not require large amount of processing power to manage this process.

Routing updates only happen every 30 minutes, although topology changes are updated as soon as the adjacent router has identified the change.

As the network increases in size more Hello packets will be issued by participating routers. The size of the Hello packet is very small as it does not contain any routing information.

OSPF uses a multicast address in a broadcast network, thereby reducing any interference with other non-OSPF devices.

Packet authentication is supported in either clear text or cipher text with MD5 algorithm.

On the downside OSPF can be difficult to configure as complete knowledge of the autonomous systems on any network is required along with any additional features such as packet authentication.

Further OSPF cannot support unequal load balancing as it creates the metric of a route based on the bandwidth of each route. To make the load balancing viable, each route must have the same metric.

3.7 An Alternative to OSPF: Enhanced Interior Gateway Routing Protocol - EIGRP

EIGRP is the enhanced version of IGRP – Interior Gateway Routing Protocol. Unlike most other protocols EIGRP is unusual in that it is a hybrid protocol, combining characteristics of link state and distance vector routing. This protocol has been developed and implemented by Cisco Systems. Although this is a proprietary protocol other manufacturers of networking equipment have implemented it in their own layer 3 switch and router operating systems.

8	16	32 bits
Version	Opcode	Checksum
Flags		
Sequence number		
Acknowledge number		
Autonomous system number		
Type		Length

Figure 12: Format of the EIGRP header.

EIGRP has a number of advanced features that combine the behaviour of link-state and distance vector protocols. These are described below:

Routers are able to learn about the existence of other routers on directly attached networks as well as the ability to determine when these devices are no longer available. This is a dynamic process, so as the network changes the routing table is altered to reflect this.

Fast convergence: EIGRP uses Diffusing Update Algorithm (DUAL) to track routes along with composite metrics such as cost to identify the most efficient and loop-free path across the network. DUAL comprises Advertised Distances (AD) and Feasible Distances (FD). “The AD is the EIGRP metric for an EIGRP neighbour to reach a particular network.” (Teare, D. 2010). The FD is the metric for that router to reach a particular network. The FD is made up of the sum of the AD plus the metric to reach the next hop router.

Rather than sending complete topology updates as OSPF does, EIGRP sends partial updates whenever a change occurs to a link. This operation ensures that EIGRP uses less bandwidth than other protocols.

EIGRP was developed with protocol dependant modules so is able to support different protocols including IPX, IPv4 IPv6, AppleTalk.

EIGRP is able to work across WAN and LAN with little configuration.

This protocol is able to support unequal load balancing so routes across the network can have different metrics but still transfer equal amounts of data.

EIGRP uses Reliable Transport Protocol (RTP) to ensure the delivery of EIGRP packets. Multicast packets are sent to update routes. These packets contain a flag which is set to identify whether or not an acknowledgement is required by the sender.

EIGRP is also able to create summary routes which allow different networks to be summarised automatically on a router's interface. So, for example, networks 172.16.1.0 and 172.16.2.0 can be summarised into 172.16.0.0 as long as both networks are accessible via that router interface.

3.8 OSPF vs. EIGRP

This is an interesting discussion and one that does not have an outcome. The choice is purely dependant on network needs and the preferred choice of the network administrator. Both protocols offer network stability, reliability and ease of configuration. Unlike OSFP that is bound by the size of the topology with its use of autonomous systems, EIGRP is more scalable allowing more routers to be easily added, discovered and routing tables further propagated. The downside to this is that the routing table can increases in size while the topology grows infinitely, making troubleshooting more and more difficult for system administrators. If this element is taken into account OSPF scores over EIGRP.

Doyle (2007) contends that there are still issues where networks configured with EIGRP are allowed to grow exponentially stuck-in-active (SIA) conditions, in which responses to queries that are not heard within a certain time can occur. This in turn causes neighbours to be incorrectly flushed from the neighbour table, resulting in severe network destabilization and incorrect routes.

As already mentioned EIGRP is a proprietary protocol developed by Cisco Systems. By definition this protocol is included as part of the Cisco operating system on every Cisco router. The way it is configured and how it behaves cannot be altered under the terms of the licence agreement.

OSPF is an open source protocol which if freely available. Other manufacturers of networking equipment have taken advantage of this and created a version of this protocol for their particular product.

3.9 Border Gateway Protocol - BGP

This protocol is used to connect with geographically remote systems. This ensures that users are able to connect with systems located at other universities and at Daresbury Science and Technology Facilities Council (STFC) through the National e-Infrastructure Service (NES). BGP cannot be used across the campus network as its algorithm uses the identification of autonomous systems to route data across a wide area network (WAN).

BGP is a protocol that links together different autonomous systems. In the case of Figure 13 AS 6000 and AS 6500 are linked. These systems can be using any other protocol, IPv4, IPv6 or IPX internally, but BGP helps to provide data transfer across the link between them. Teare (2010) defines the main goal of BGP as providing "an interdomain routing system that guarantees the loop-free exchange of routing information between autonomous systems. BGP routers exchange information about paths to destination networks."

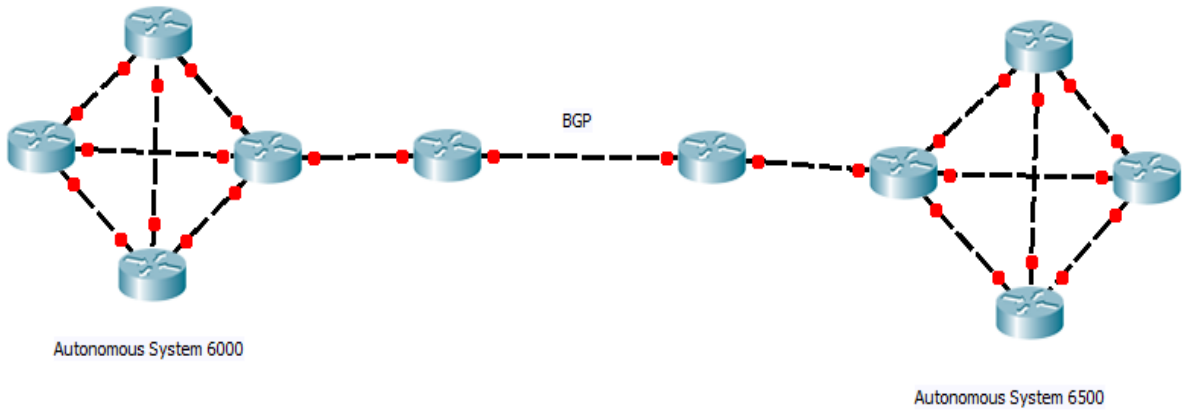


Figure 13: BGP communication between two autonomous systems

BGP is used extensively by ISPs to direct the flow of network traffic across the multitude of networks that make up the Internet. The routing table consists of directly connected, static and dynamically learned routes. BGP-4 is the only acceptable version of BGP for use on the Internet. One of the major benefits of implementing BGP is that it is able to support classless inter-domain routing (CIDR), and variable length subnet mask (VLSM). Where CIDR is used the network routes can be summarised on a core router

to around 300,000 blocks. Not using CIDR would result on over 2,000,000 routes being present in the routing table.

BGP is a policy based routing protocol and unlike other protocols, does not operate on metrics. To keep track of all the many changes that exist across the Internet, BGP reacts dynamically, advertising route changes by sending partial routing table updates.

Based on the comparison of network protocols, it was decided to use OSPF.

As far as implementing a new cluster is concerned, the major reason for choosing OSPF over EIGRP is that the University of Huddersfield campus network currently runs this protocol. This in turn means that there is technical expertise available with this protocol, which is an advantage when devices have to be configured to work with the existing network. Furthermore, it is prudent to centralise resources and improve availability of cluster systems. This helps to provide a seamless integration between the cluster systems and the university network, which is imperative for providing access to geographically remote sites and users.

Chapter 4: Literature Review on Network Design

There is a substantial amount of literature that deals with the design and implementation of computer networks from specific points of view. For example by using graphical information to determine the best topology (Sarma and Pal, 2010), geographical location (Altiparmak, Dengiz and Smith, 2003) or providing the best mechanisms for fault-tolerant networks (Szlachcic, 2006). While each of these approaches has their own merits in network design, the fundamental issue with existing networks is how to integrate new equipment into the existing infrastructure while still achieving high speed data connectivity. This becomes more important where HPC cluster implementations are concerned as file sizes are considerably larger than for other applications such as word processing or spreadsheets.

Consideration is also given to the nature of the devices used. Cisco Systems' white paper on switching technologies (Yang, Y. 2012), explains the differences between using cut-through or store and forward technologies to achieve low latency, while still allowing high throughput. This is important in allowing large files to traverse the network, while also asking questions about the difficulties integrating HPC systems into an existing infrastructure.

The issue of topology is also given considerable consideration. In many cases the integration of any High Performance Computer system with the existing systems follows a hierarchical model to ensure that devices are placed in the appropriate locations, and are able to access key services such as remote access, DNS and DHCP. From time to time, HPC systems are built from scratch so the infrastructure can be built around them. The effect of this is to create a rigid topology, which is then limited to a finite number of devices, thus making further expansion difficult. The Jellyfish topology (Singla, 2012) has been proposed as an easily expandable topology, which provides increased levels of flexibility as well as load balancing and good path redundancy. Jellyfish was compared to Fat-Tree, which determines the scalability of the network based on the number of ports available on switches, but increases rigidity across the topology.

These topologies are able to provide a stable infrastructure that allows for redundancy, a degree of flexibility and scalability. However there is little to aid the integration with an existing infrastructure, such as required by new cluster installation proposed in this work. For this reason this work concentrates on a hierarchical design model.

Chapter 5: HPC Network Design

This section considers the existing infrastructure including the location of devices and issues concerning bandwidth, before the installation of a new cluster.

5.1 Existing infrastructure and Issues

The existing infrastructure has evolved through research undertaken by the HPC Research Group. As a result of this very little consideration has been given to aspects of network design such as number of current users, projected growth of users and devices, bandwidth, number of devices spanning the network, interconnection with other internal and external services and the physical layout of existing network.

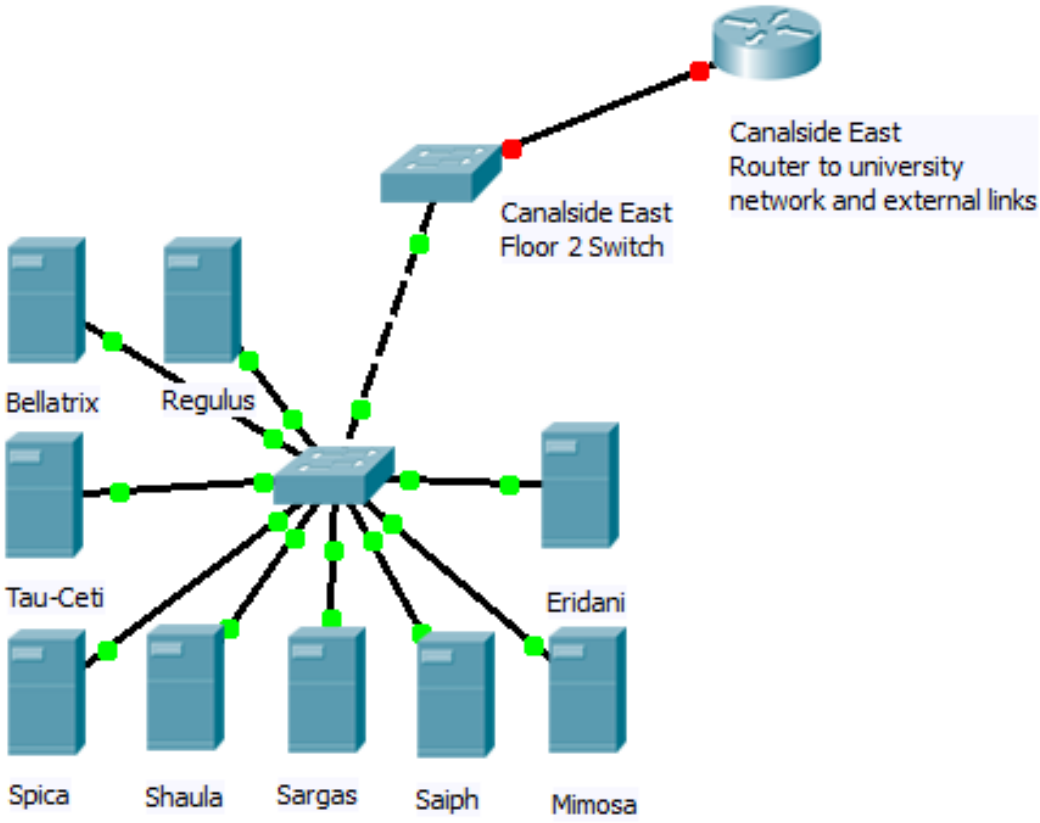


Figure 14: Original configuration of clusters and servers forming the QGG

The original configuration consisted of a number of servers and clusters, which when connected together, formed the Queensgate Grid (QGG).

Bellatrix – authentication server for internal and remote users

Regulus – North West Grid Services (NGS) authentication server for links to external systems

Tau-Ceti – AMD cluster

Spica – HPR-RC certificate authority server

Shaula – Legacy engineering flex certificate server

Sargas - Legacy engineering flex certificate server

Saiph – HPC-RC new flex license server

Mimosa – 16 TB Network Access Storage (NAS)

A 1GB switch linked them together, which then connected to a switch on floor 2 of the Canalside East building. This then connected to a router which fed back into the university network and provided links to the internet and other externally used systems.

The university was able to acquire a further cluster, which required 2 racks to house all the equipment. Due to the lack of available space in the HPC Research Centre, alternative accommodation had to be found for the new cluster. With the full agreement of the university's Computer Services, space was allocated in a data centre, which also provided air conditioning and an opportunity to link with 10GB backbone of the main university network. Besides needing more floor space than the office could offer, this new cluster also needed more bandwidth to be effectively used. This would also alleviate many of the issues around having very large file sizes traversing across a 1GB link. One of the effects of using a 1GB link has meant that some users had to split their files into smaller blocks of data and then re-assemble the block on the cluster, before submitting to the job queue.

5.2 Network Design

In conjunction with the Computer Services Network Team, a plan for implementation of the new cluster Sol was devised. As this project was integrating a new cluster into the existing infrastructure, the real design consideration was the placement of different devices to ensure that high speed connectivity could be achieved. The main focus was to provide high speed connectivity for large file sizes as well as alleviating issues such as bottlenecks and latency. Currently users are still connected using 100Mb CAT5 and CAT5e Ethernet cabling. The buildings offer a bandwidth of 1GB. The network backbone operates at 10GB. By connecting directly to the backbone, some of the issues surrounding high speed connectivity can be alleviated.

Chapter 6: Network Devices for HPC Network

The university network uses both switches and routers. This section explains how these devices operate and how they are incorporated into the new installation. In addition the layer 3 switches, used to provide high speed interconnection across the new cluster are also described.

6.1 Role of Switches in HPC Network

Essentially a switch is a multiprocessor bridge, which has the capability to forward frames “simultaneously between all pairs of its ports.” (Olifer, 2006) The name switch was derived from the switching matrix employed to connect the individual processors within the device together.

Mikalsen and Borgesen (2002) identify an important reason for the increased usage of switches being due to the “opportunities they provide for high speed networks.” Certainly this is a necessity where high performance computing is concerned, due to the size of files that are transferred across networks for processing. Switches are able to transmit millions of frames per second.

Operating at layers 2 and 3 of the OSI 7- layer model, switches are able to populate a table with MAC addresses for the attached devices. This provides a simplistic, but very effective data delivery service which conforms to the IEEE 802.1D standard and utilises the transparent bridge algorithm.

There are significant differences between switches that operate at layers 2 and 3. Layer 2 switches provide the interconnects for all other attached network devices. Layer 3 switches also perform this same activity, but in addition have routing capabilities built in. These layer 3 switches can also provide uplinks to allow a number of switches to be interconnected together, thus expanding the available ports for use by other devices.

For the planned Sol cluster implementation three different switches have been selected for their capabilities and ability to transmit at high speed: Netgear GSM7228PS, Nortel Baystack 5510-T, and Cisco Nexus 5K.

6.1.1 Switch 1 – Netgear GSM7228PS

This is a layer 3 28-port switch, offering 24 ports operating at 10/100/1000Mbps, plus 2 SPF+ ports and 2 10Gb Ethernet SPF+ interfaces to provide uplinks and stacking. One

of these SPF+ ports has a module added to allow direct connection to the university network backbone. Besides offering a range of data transmission speeds this switch is able to utilise VLANs (see Chapter 6 for more detail) and can carry voice traffic. On the routing side this switch offers VRRP and OSPF as standard protocols. OSPF fits well with the university network. This protocol is discussed in Chapter 3.

Figure 15: Netgear GSM7228PS Layer 3 Switch

6.1.2 Nortel Baystack 5510-T

There are five Nortel Baystack 5510-T switches which provide interconnectivity across the racks to enable the head nodes, servers and scratch space to communicate across the cluster. Again these operate at layers 2 and 3 of the OSI 7 layer model. As these are no longer manufactured motherboards which would allow these devices to support bandwidths over 1GB are also not available. Fortunately the interconnects are able to support transfer rates of 40GB to allow data to be transferred at high speed across the cluster. In respect of this the Netgear is able to bridge the gap between the two layer 3 switches to provide high speed data transfer across the network and into the cluster.

Figure 16: Nortel Baystack 5510-T Layer 3 Switch

6.1.3 Cisco Nexus 5K

The third device is a Cisco Nexus 5K. This is also a layer 3 switch, which performs some routing between the data centre and other buildings on campus. These switches operate using either packet switching or circuit switching. There are advantages and disadvantages to both of these modes which are discussed in section 6.1.4 and 6.1.5.

Figure 17: Cisco Nexus 5k switch with 10GB ports for fibre channel
All above switches could operate in many different modes:

6.1.4 Packet Switching

Switches using this mode have internal buffers which are used to store the packet temporarily. The switch needs to collect all parts of the packet to be able to make a decision on how it should be forwarded. This is also known as store-and-forward. In order to be able to store the packet the buffer must exceed the size of the packet.

The major disadvantage to this mode is that long queues of data can be created resulting in performance issues and placing additional stress on the switch.

6.1.5 Circuit Switching

In this mode the data is transmitted from end to end at a constant rate. In order to achieve this, a connection from source to destination must first be established. The data can then be forwarded.

6.1.6 Cut through vs. store and forward (speed vs. reliability)

Typically cut-through switches are suggested as the most appropriate devices to use for HPC environments. Cisco Systems (2008) have identified that "…cut-through switches are more appropriate for extremely demanding high performance computing (HPC) applications that require process-to-process latencies of 10 microseconds or less." This is measured using the first-in, first-out (FIFO) method as explained in section 6.1.5.

HPC systems benefit from low latency, 2 or 3 microseconds, offered by cut-through switches, which allows for high speed transfer of large files.

These switches allow data to be forwarded as soon as it arrives at the switch and the destination address has been read. A major benefit of this is that these switches operate at wire speed allowing for fast data transfer. Error checking does not take place under this switching technology, which in turn means that invalid packets are not dropped even though they are flagged. Switches using this technology do not have RAM buffers.

The alternative technology is store and forward. Using this method frames are not forwarded to the destination until the entire frame has been received by the switch. In

direct contrast to cut-through switches, store and forward have RAM buffers which store the frame until all packets have been received. The major benefit of store and forward is its ability to allow error checking to occur. Store and forward utilises CRC, cyclic redundancy check, which comprises of a mathematical formula based on the number of 1 bits in the frame to decide whether the frame has an error or not. This makes store and forward a useful addition to any environment which is supporting reliable connections and datalink protocols. Store and forward switches are measured using the last-in, first-out (LIFO) method as described in section 6.1.8

6.1.7 Alternative switching technology

There is an alternative to both the aforementioned switching technologies, fragment free. This technology encompasses elements of both cut-through and store and forward. On receipt of packets it stores the first 64 bytes of the frame. As most problems, such as collisions, occur at the first 64 bytes of the frame so this technology tries to enhance cut-through by using a small error check and then forwards the frame to the destination. While this alternative technology offers the best of both worlds there are still implications for increased latency compared to using cut-through switches.

6.1.8 Switch Latency and effect on HPC

Latency has become a major feature of high performance computing. “Ethernet switch latency is defined as the time it takes for a switch to forward a packet from its ingress port to its egress port.” (Yang, 2012) There are four measurements for latency: Last-in, first-out (LIFO), last-in, last-out (LILO), first-in, first-out (FIFO), first-in, last-out (FILO). Each method defines the way in which packets traversing between two points, are measured. The result is the latency.

Last-in, first-out: Defined in RFC 1242. The latency measurement timer is started as soon as the last bit of the packet is in the switch (T0). It stops when the first bit of the packet leaves the switch (T1).

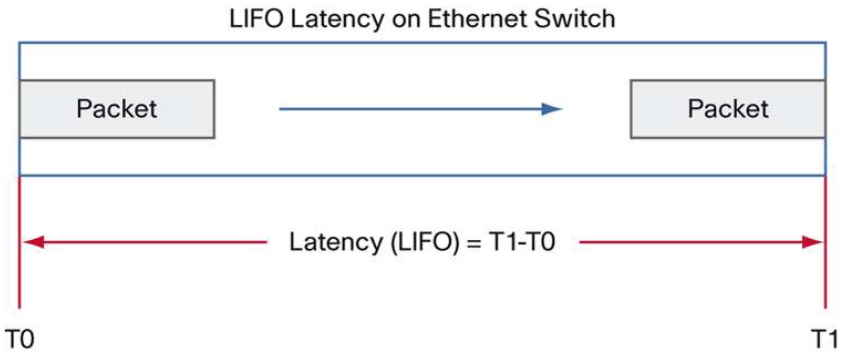


Figure 18: LIFO latency on Ethernet switch
(Source: Yang 2012)

Last-in, first-out: Defined in RFC 4689. The latency measurement timer is started as soon as the last bit of the packet enters the switch and stops when the last bit of the packet exits the switch. LIFO is measured using the following formula:

LIFO = FIFO - (Packet size in bits/Link speed)

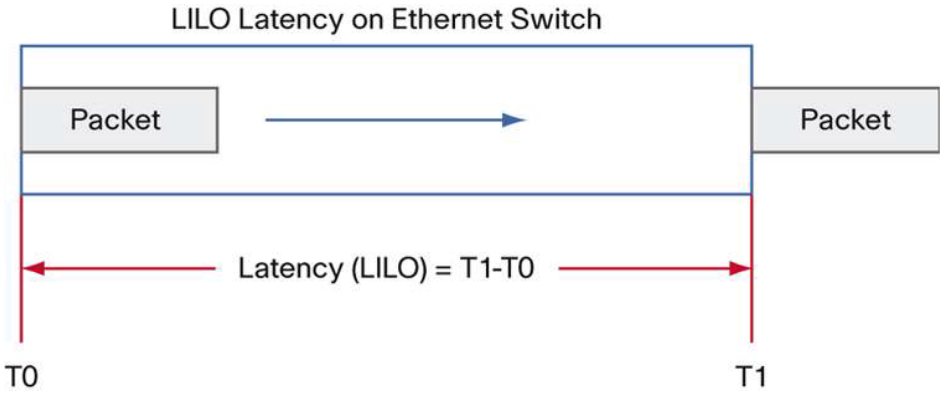


Figure 19: LILO latency on Ethernet switch

(Source: Yang 2012)

First-in, first-out: Defined in RFC 1242. The latency measurement timer is started as soon as the first bit of the packet enters the switch and stops when the first bit of the packet exits the switch. When the measurements are compared to LILO the results are the same. The timestamp is added just before the Frame Checksum Sequence (FCS) of every packet that is measured.

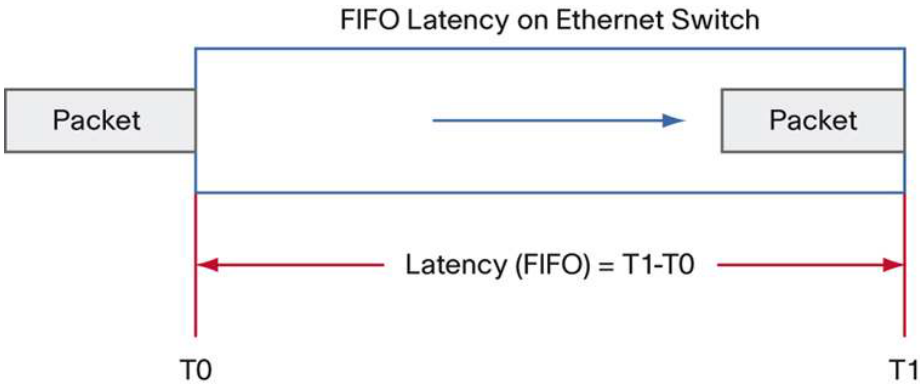


Figure 20: FIFO latency on Ethernet switch
(Source: Yang 2012)

First-in, last-out: This method is currently not defined in any RFC. The latency measurement timer is started as soon as the first bit of the packet gets in the switch and stops when the last bit of the packet exits the switch. FILO is measured using the following formula:

FILO = FIFO + (Packet size/Link speed)

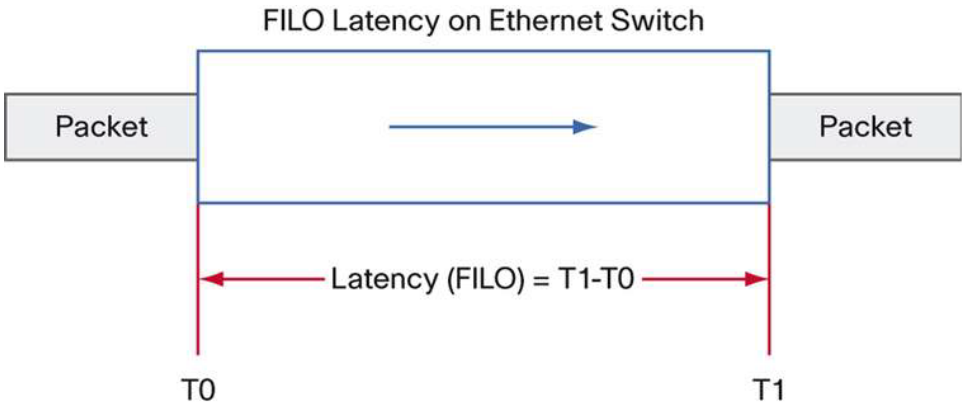


Figure 21: FILO latency on Ethernet switch
(Source: Yang 2012)

The university network is made up of a number of different Cisco switches which link every building on campus with two campus data centres, and send data both internally and externally. By utilising OSPF the quickest path across the network between routs can be established, but switches have no such mechanism. Chapter 3 explains the

mechanism switches use to forward data. Whenever a packet packed visits a network device more latency is added to the total time to traverse the network. This additional time is not seen by the end user as it is measured in milliseconds. However, a significant build-up of latency can cause performance issues if the network is not design, implemented and managed properly.

6.2 Role of Routers in HPC network

The router's role is to read the packet header to determine where the packet is destined for from the network address provided within the IP address. This then helps to determine the route for the packet to enable it to reach its destination. This might be on the local area network or across the campus area network including the backbone and any edge routers that provide connectivity to the Internet.

In the University of Huddersfield campus there are a number of buildings located over a large geographical area. To connect these together, each building has a communications room which houses some of the necessary switching and routing equipment required to connect devices located through the building. For example, the communications room in Canalside East has switches on each of the five floors, plus a router to connect the switches back to the backbone. Each router then provides a high speed connection to the next building, also known in networking terms as a hop. In Figure 22 there are three routers in each of three buildings, Business School, Canalside East and Canalside West. They are connected together using fibre optic cabling to provide a backbone that runs at 10 Gbps. Each router is 1 hop away from the next router and configuration of the protocol dictates that routes are either static or dynamic. In other words, they are either manually configured to only use that route, or are able to learn and change dynamically as the network topology alters.

From Canalside West to Data Centre 1 there are 2 hops if the data is sent from Canalside East or 3 hops if load balancing is instigated, or a route is unavailable and the data is sent via the Business School.

Each router will create a routing table, which contains network topology information, identifying the protocol in use and the known neighbour router(s).

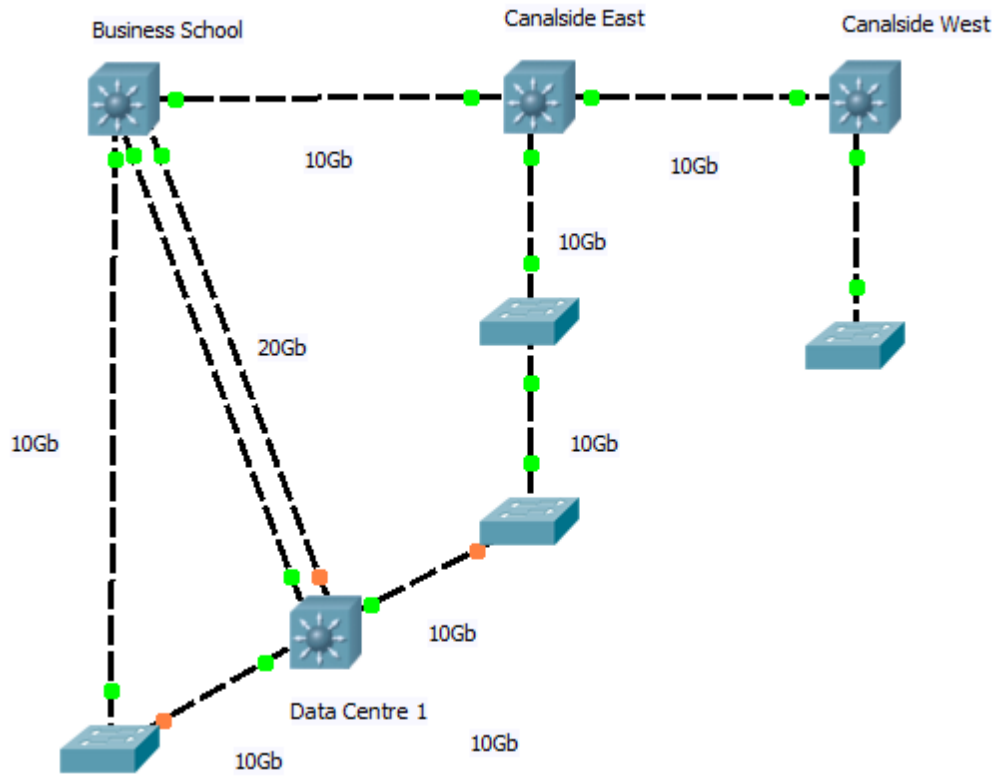


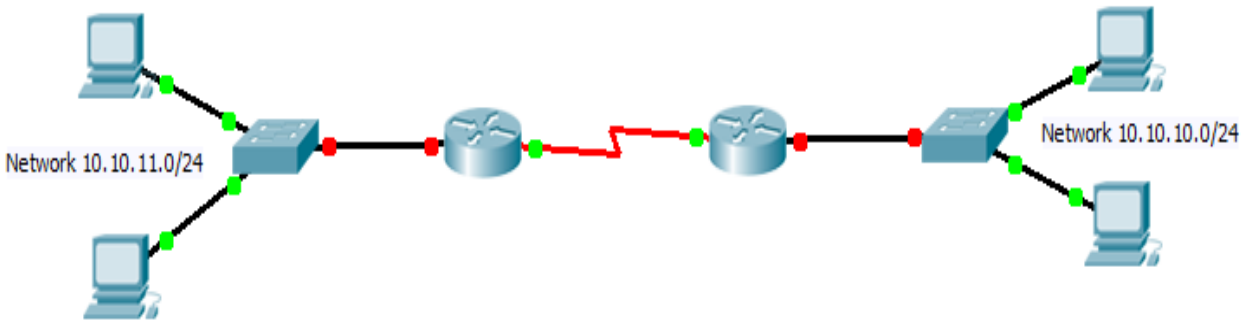
Figure 22: Example of router setup and hops across the university network

6.3 Router Control Plane

Routers were originally designed using a single centralised processor to perform routing calculations. The Edge routers, those that are located on the edge of a network and provide connectivity to external networks, use high port densities and therefore undertake more route processing. Single processor routers are not capable of handling multiple routes, computing and re-computing routes, as are required in this instance.

The role of the control plane is to propagate the routing table so that data can be transferred efficiently based on the requirements of the protocol. Routes are configured in two ways:

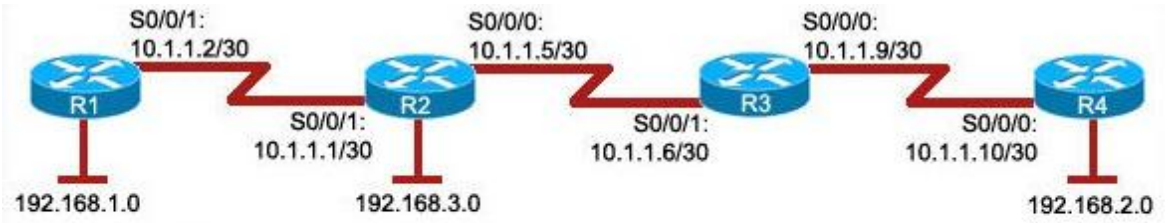
Static routes – these routes are manually added by the network administrator. They are only used on small networks where data is traversing between limited numbers of nodes, typically point-to-point where there is only one route available as the routers are directly connected. In Figure 23 the routing table shows the directly connected routes so data from the 10.10.10.0 network can be sent to the 10.10.11.0 network.




```
     10.0.0.0/24 is subnetted, 1 subnets
C       10.10.10.0 is directly connected, Serial0/3/0
```

Figure 23: Static route between two networks.

Dynamic routes - these routes are learned by the routers that make up the network and are connected together. However, dynamic routes are able to deal with much larger number of available routes across a network and are able to build a map of the network based on adjacent routers and routes to other networks via other routers.




```
R2# show ip route

   10.0.0.0/30 is subnetted, 3 subnets
C      10.1.1.0 is directly connected, Serial0/0/1
C      10.1.1.4 is directly connected, Serial0/0/0
C      192.168.3.0/24 is directly connected, FastEthernet0/0
R      10.1.1.8 [120/1] via 10.1.1.6, 1d00h, Serial0/0/0
R      192.168.1.0 [120/1] via 10.1.1.2, 1ea0h, Serial0/0/1
```

Figure 24: Dynamic routing table showing connections between different networks.

The router control plane builds the table of destination addresses so that when a packet is received by a router it can be de-capsulated to determine first of all, the IP Address and then matched against the routing table to identify the correct interface to forward the data out of. This technology is built into all routers and also into layer 3 switches. This enables the layer 3 switch to operate as a router if the network infrastructure requires. In this network the Cisco 3560 layer 3 switches perform both roles. Data leaving the building is received at the switchport. The destination is identified and the

router part of the device takes over, identifying the destination network address to forward the packet on.

The Cisco Nexus 5k located in the data centre, also has this capability. This also means that data can be passed from the cluster directly to the backbone and then routed across the network in an efficient manner.

6.4 SPF and SPF+ Modules

These modules are additional requirements for the Netgear layer 3 switch to connect to the Cisco Nexus 5k as well as Bellatrix, the authentication server and RegulusA the network area storage. Bellatrix and RegulusA both use a 10 GB SFP Ethernet module which is copper wire construction. The SFP+ module provides fibre channel connectivity from the Netgear the Nexus 5k. SFP is also used between a second Netgear switch and the university network backbone to provide access to the DMZ for remote users.

Figure 25: SFP+ module for fibre channel (left) and SFP for Gigabit Ethernet

6.5 Role of NAS

The network area storage (NAS) consists of two 16Tb servers. The role of these servers is to handle user data. As files are received, the data is stored and then synchronised on both NAS servers. The data is not kept as a backup, but is instead written into scratch space – temporary working space, processed, the results written into the scratch space again. Finally the data is returned to the user on request.

Chapter 7: QGG – Systems and Issues

As it is often the case in any network or system implementation, there are issues that occur both before and as the processes happen. This section provides an overview of a new cluster implementation and highlights the issues concerning the network installation.

7.1 Overview to QGG

In the early days of cluster development at the university, all the equipment was housed in a small HPC RC connected to the rest of the main network via a 1 GB link. There were few users and therefore not substantial amounts of data traversing this link. This meant there were few performance issues and systems were able to manage user requests for processing without any major problems. This was the initial implementation of the Queensgate Grid – QGG. Kureshi (2010) has documented the full extent of the development program including individual cluster specifications.

In the first two years, from 2008, the number of users, software applications and the demand for increased performance has increased substantially. In 2011 a new system was offered to the university which is able to meet current user requirements as well as providing high speed connectivity. As a result of this, the new system (Sol) was implemented in a data centre, which provided the high speed connectivity required.

This implementation also raised issues concerning users connecting to the system and how the network area storage (NAS) is to function. There was a further issue with the Netgear switch which was consequently resolved.

7.1.1 Connecting and Authenticating Users

Bellatrix server is responsible for the authentication of the users when they request connections to any of the other clusters in the QGG. Bellatrix is also responsible for passing the jobs to the relevant clusters for processing. When a job is submitted all the data is sent. This means that Bellatrix has to handle large files, although these are only being passed on when the job is scheduled.

From a network point of view, this raised considerable issues about the flow of data traffic. The major issue here is related to the availability and allocation of space on

campus. This meant that the clusters have evolved into any available space with little consideration for the required bandwidth and the network design.

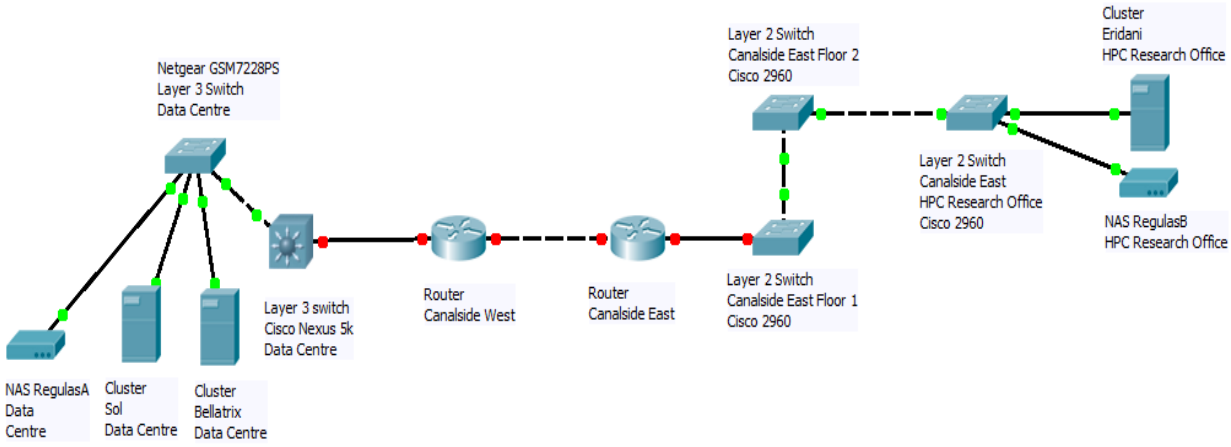


Figure 26: Network schematic showing preferred route between data centre and HPC Research Cenre.

Some servers are still located in the HPC Research Centre and this means that there is always a potential to cause a bottleneck or for users to experience latency. Although clusters receive large file sizes these are segmented into smaller packet sizes, so this situation is currently unlikely to occur. The main issue at present is one of user authentication. The following details the stages of user authentication. Figure 14 is useful as a reference to the location of the servers.

- A user sends authentication information and is authenticated by Bellatrix (Data Centre)
- The job data is sent through Bellatrix
- The job is scheduled on Eridani and sent to the queue (HPC Research Office)
- Data is passed to RegulusB (HPC Research Office)
- RegulusB synchronises with RegulusA (Data Centre)
- Job is processed
- Results are saved and the two NAS servers are synchronised again
- Results are requested by the user (anywhere on the campus)

If this route is followed, the user's data will encounter the following possible routes:

Route A - from User to data centre

- 100Mb User's PC to switch
- 2Gb Switch to router
- 10Gb Router to data centre – Bellatrix for authentication

Route B - from Bellatrix to Eridani

- 10Gb to Nexus 5k
- 10Gb to router in Canalside West
- 10Gb from router in Canalisde West to Canalside East
- 10Gb to Eridani, HPC Research Office

The same route is used to transfer the data to RegulusB.

The reverse route is used to synchronise data from RegulusB to RegulusA. This is done twice as the results are synchronised as well. This is demonstrated in Figure 27.Route A is reversed to return the results to the end user.

If the space had been available in the data centre for all the servers and clusters to be housed together, all this data that is traversing the main network could be sent through only one switch. Instead, this data is transferred over multiple devices which increases the number of times each packet is encapsulated and decapsulated. It also adds to the amount of latency, which builds cumulatively with every device visited which has also made the network design element very difficult to undertake.

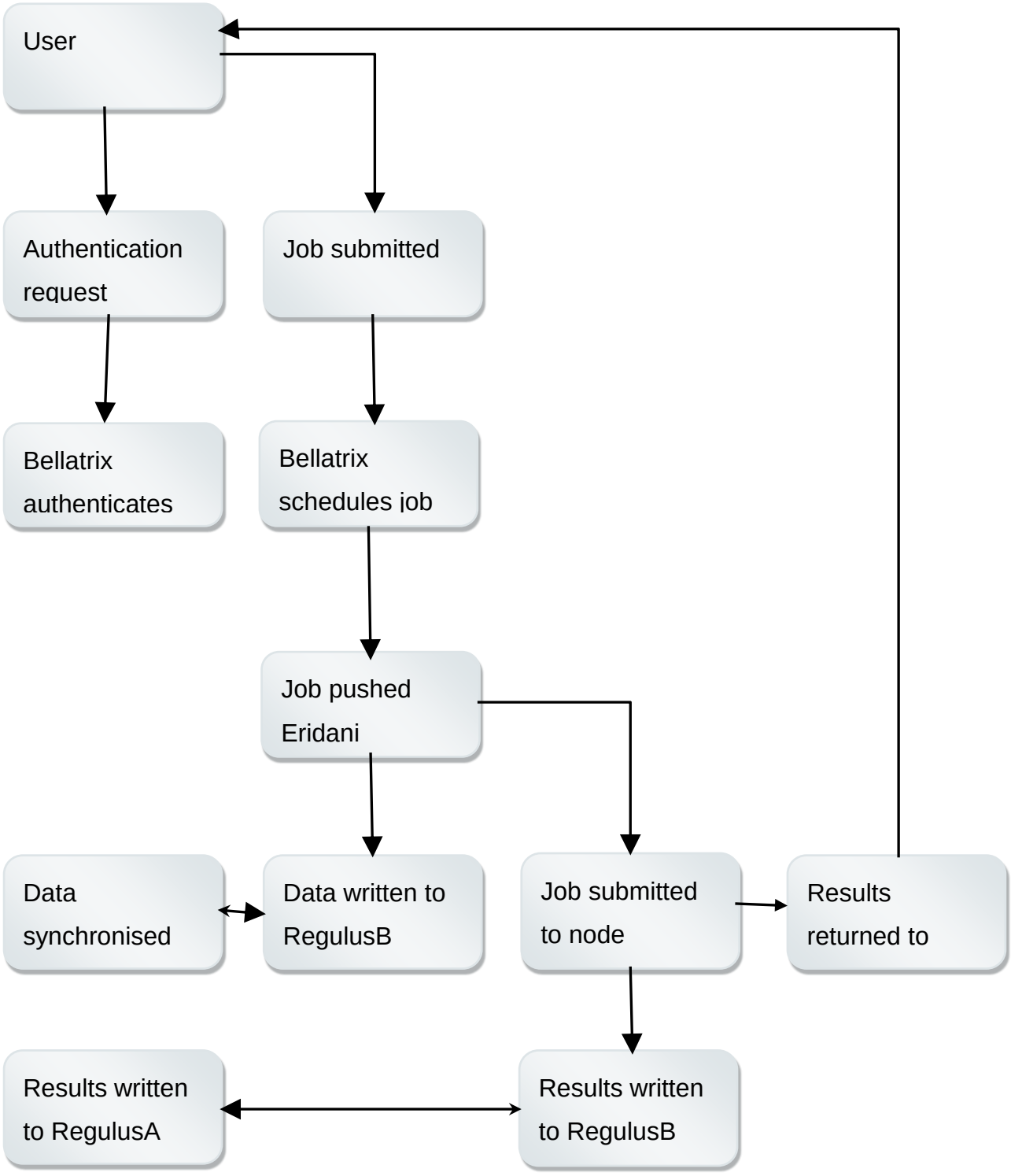


Figure 27: Flow chart for job submission and processing showing data flow across the network

Chapter 8: Sol Implementation

The devices and relevant protocols have already been discussed in earlier Chapters. IP addressing has also been explained, but not in direct relation to the clusters. The following section deals with that aspect and looks at DHCP, DNS and DMZ.

8.1 Implementation of IP Addressing

As the university network, like most networks today, uses IP protocol each device that is required to communicate on the network must have an IP address. In addition, there must be a subnet mask to define the network and the host addresses, as well as a default gateway so data can be forwarded to the next hop address – the nearest router. The Netgear layer 3 switch was assigned an IP address in a given range allocated to this project. This in turn means that traffic on this switch can be monitored and forwarded correctly with the inclusion of a default gateway address. The clusters were set to pick up DHCP so IP addresses would be allocated automatically. The major benefit being that the administrators no longer need to set each IP address manually.

Some of the servers also need to be accessed from outside of the campus, so are given DNS entries on the DNS servers to forward and reverse lookup their given hostnames to an IP address. It is more difficult to remember a list of IP addresses, but much easier to remember a uniform resource locator or URL. Users are able to submit jobs remotely in the same way as they would from any PC on campus. The external interface of Bellatrix for authentication purposes, Sol's user interface for direct submission of jobs, Spica to provide virtualised environment and Vega to offer GPU processing, all have external interfaces which are visible to remote users.

It is important to note that the clusters have two network interface cards which provide one internal link and one external link. The external links are used for remote access. The IP addresses that the external network interface cards are given are in a different range to the internal addresses. This provides an increased layer of security, which is especially useful in an environment like this where students come and go on three or four year cycle. In respect of this our work has been to ensure connectivity by correctly configuring devices to enable access to network services and resources. This has resulted in a considerable amount of network design work, which helped us to identify potential problems with the implementation.

8.2 Limitations of proposed infrastructure

While the addition of the new cluster, Sol, has proved to be very successful in terms of usage since it became available in early September. Already the CPU usage is averaging 76% and has peaked at 100% on 6 occasions. Ganglia Monitoring System is open source software designed for monitoring high performance systems. The statistical outputs allow users to see all traffic from initial data transfer to processing on each compute node. The other network monitoring system utilised is Solar Winds provided courtesy of the university Computing Services. It provides statistical and graphical information on data transfer across the network and through each device.

CPUs Total:	260
Hosts up:	65
Hosts down:	0
Current Load Avg (15, 5, 1m):	79%, 79%, 79%
Avg Utilization (last hour):	79%

Figure 28: Ganglia statistical information for Sol load 5th October 2012 at 10am. (Courtesy of High Performance Computing Research Centre)

This suggests that the new cluster with 256 cores is still not big enough to meet the demands of the university's research community. This in itself raises some issues in terms of financial input and space which could hamper the growth of the HPC services.

(Core availability across all clusters still not enough to satisfy the need of the research community – backup with Ganglia data to show 73% + usage over time and up 100% in the last few days.) We have extracted raw data in .csv format as well as statistical graphical data from Ganglia and Solar Winds to monitor the traffic flow across the network to and from the clusters.

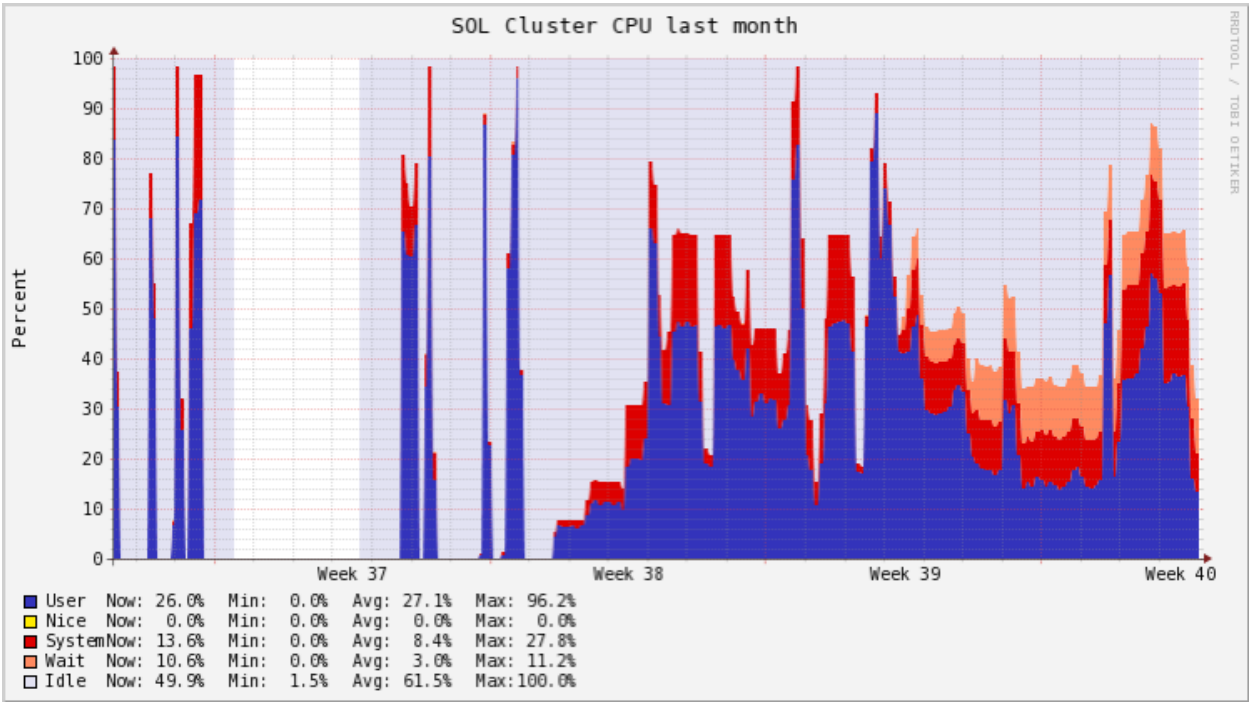


Figure 29: Ganglia data showing CPU usage for September 2012
(Courtesy of High Performance Computing Research Centre)

Due to the lack of available data centre space the clusters have been split, so that when the data is written to the scratch space and mirrored with a second network area storage server, the data has to be transferred across the network twice. As research becomes more complex and file sizes increase, this data transfer could cause performance issues across the network, creating bottlenecks. Further, until the network infrastructure is increased in bandwidth at the user's desks, the 100MB link from PC to the nearest switch will always raise cause for concern, especially as file sizes increase and the number of researchers using the HPC grows. However, this is a strategic, long-term issue which will be implemented as the desktop technologies change.

Probably the most important issue that has arisen in this installation process is the difficulty of trying to enable devices, manufactured by different vendors, to communicate over the network. The Netgear and the Cisco Nexus switches will only communicate if the Netgear has very little configuration set. The initial configuration had VLANs set so that HPC traffic and administration traffic could be separated. Plus setting up VLANs also ensures a level of security as users are not automatically given access to all VLANs, but are allocate to groups. When VLANs were setup on the Netgear switch, the data traffic greatly reduced to just a few small packets every minute. None of the

servers in the cluster could be reached using Ping. Nor could the servers make contact with the DHCP server to obtain IP addresses. Once the VLAN settings were removed the data traffic increased substantially and all devices were able to communicate effectively. Hopefully future advancements in routing and switching will see smaller, more cost effective switches made available with 10GB ports, rather than having to spend significant amounts of pounds on devices that are incompatible despite the guidance provided by the manufacturer's specification. Our work in this area highlighted this issue, which when explored further, had been experienced by other network engineers undertaking similar tasks.

8.3 Potential Solutions

There are a number of solutions that will aid the further development of the HPC systems and the research group.

- Inclusion in the university computing strategy so that the HPC can be an instrumental part of the growth and development of computing services across the campus.
- Allocate more space in the data centre to allow the clusters to grow in size as well as processing power.
- Provide dedicated, expert staff who can manage the HPC systems. Currently the systems are managed by post-graduate researchers with little experience of large scale multi-core computer systems.
- Provide a blueprint for HPC installations that sets out the budget requirements based on the right choice of devices and software.
- Increased investment from the university to allow the HPC to grow as the demands of the researchers increases. Currently the HPC Research Group relies on obtaining money from external funding bodies and the goodwill of the departments who utilise the services. This will also help to underpin the HPC blueprint for installation.

In time the university will also need to increase the amount of processing power available. This will mean replacing the current clusters with machines that offer more

cores and utilise infiniband for interconnects rather than high speed Ethernet connectivity.

Chapter 9: Network Management

Quantifying the data that is extracted from the cluster systems is a difficult task. Unlike other systems there is no consistency to the volume of data that is send over the network to the cluster. It is also very difficult to extract cluster traffic from other network traffic.

Three network monitoring applications are used to monitor the flow of data across the campus to the clusters. The first application, Solar Winds, is used through the Cisco Nexus 5K and the Netgear switch. This allows the traffic across the 10GB link to be monitored. As this is high bandwidth and job submission files are relatively small (need files sizes), the average flow is 23Mbps.

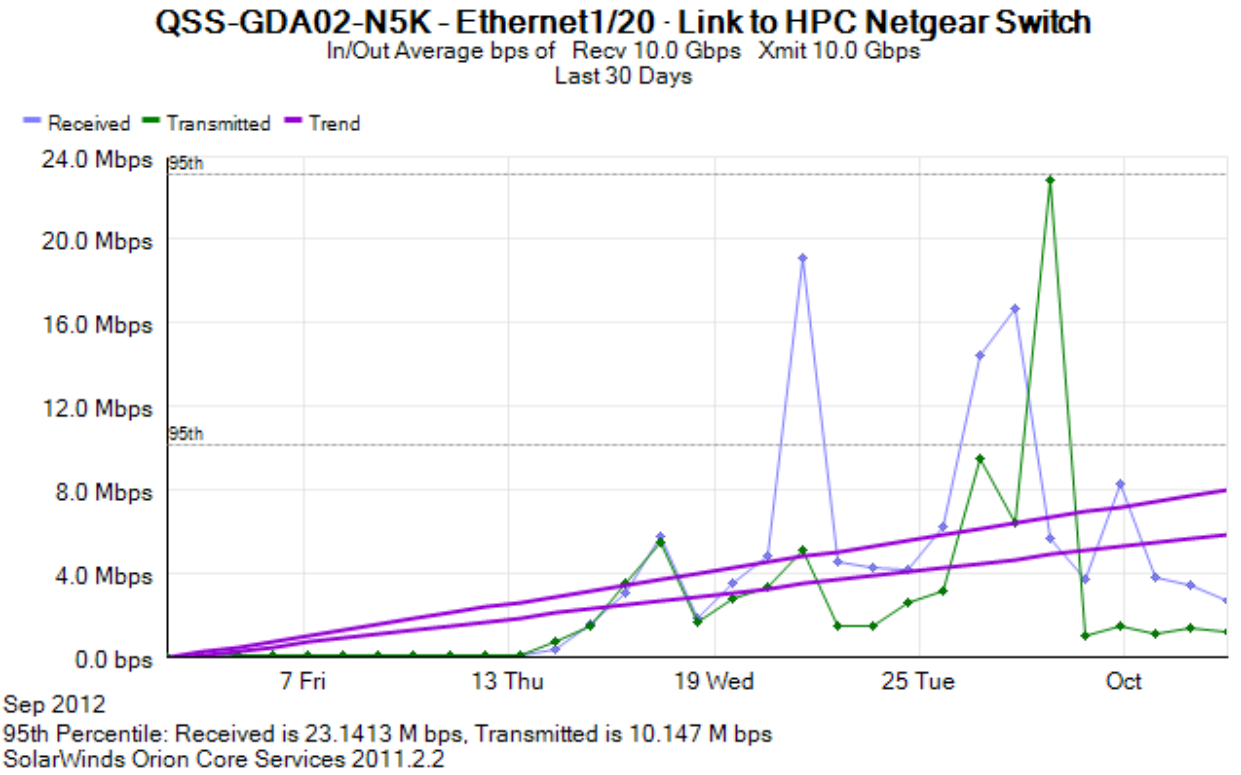


Figure 30: Data transmission rate Mbps between the Cisco Nexus and Netgear switches

(Courtesy of University of Huddersfield computing Services)

From time to time larger files are transmitted, which are shown in Figure 31. Files sizes of up to 28GB have been recorded. Over a 100MB or 1GB link submission of these files could have caused a bottleneck.

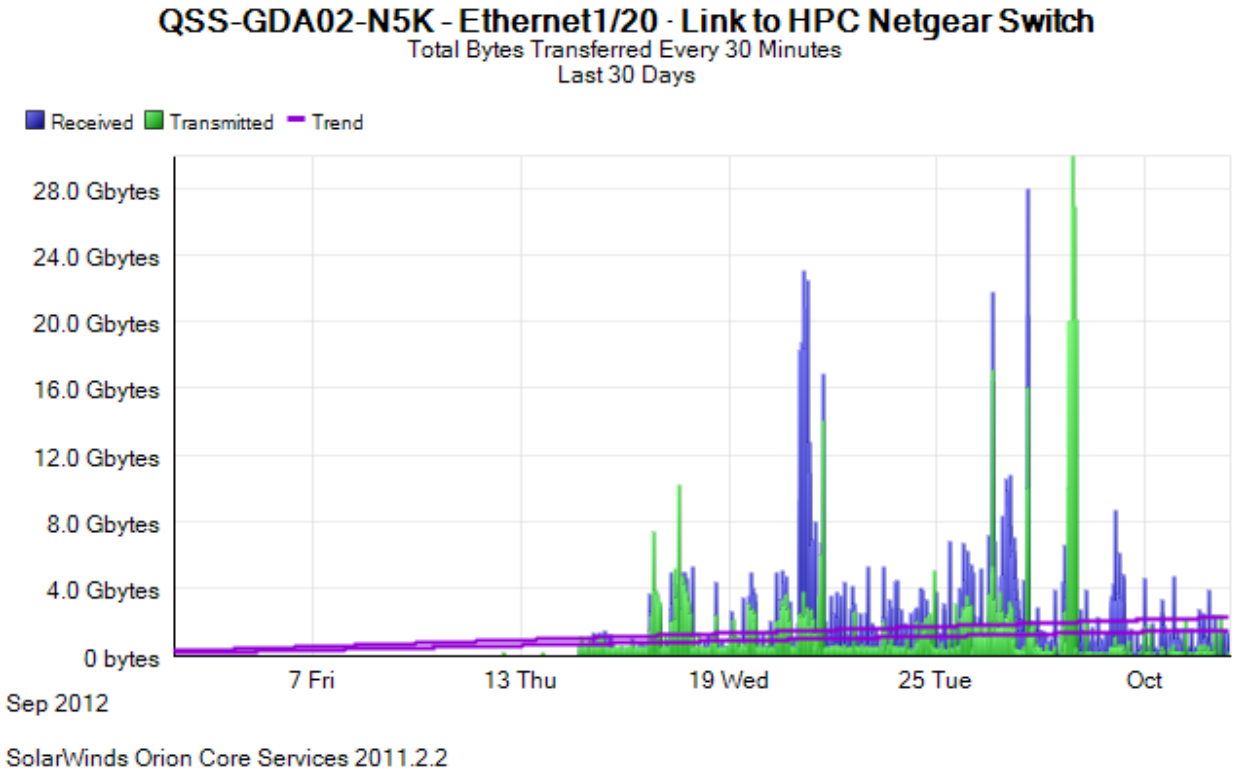


Figure 31: Flow of data between the Cisco Nexus and Netgear switches
(Courtesy of University of Huddersfield Computing Services)

This data only shows the traffic across the link between the two switches and does not account for any other network traffic.

The second application is the Netgear configuration tool that is pre-installed on the switch. This software provides useful statistics to show how much data has been transmitted, received and is broadcast traffic since the switch was turned on. This is further broken down into how much data per port so it is easy to see how much data has been transferred to Sol, or to Bellatrix for authentication.

The third application is Wireshark, a packet sniffer, which is used to help identify types of data. For example the software identifies different protocols and provides statistical as well as graphical information on how many packets are being transferred in a given time period.

In addition to these tools data transfer can also be measured using the *tracert* and *ping* utilities from the command prompt. Tracert identifies the number of hops across the network and shows the path that data is being routed along and the length of time it took to reach that router. Ping sends a very small packet of data and measures the

response time for the answer. Typically this latter utility is used to test connectivity rather than to measure time taken.

Chapter 10: Geographically Remote System Issue

Part of the new cluster implementation process was to re-establish links with geographically remote systems. This required some simple configuration of IP addresses, DNS and DMZ entries plus connectivity testing to ensure that everything was working correctly. This proved to be the case at the University of Huddersfield campus. As ICMP sends a 64 byte packet the ensuing issue was not identified until larger packets were transferred.

10.1 TCP Timeout Issue

Just before SoI was implemented, the STFC laboratory at Daresbury, undertook some upgrade work on existing clusters and network devices. However, this process highlighted an interesting issue for one of the University of Huddersfield remote users. Having connected to Sid, one of the Daresbury systems, the user transferred data for processing. The data was not able to reach its destination as TCP was timing out and then re-trying. The duplicate acknowledgements can be clearly seen in Figure 32.

No.	Time	Source	Destination	Protocol	Length	Info
51	0.061201	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=141969 Ack=36601 Win=1373 Len=0 TSval=1179106090 TSecr=340919937
52	0.061448	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=141969 Ack=39497 Win=1373 Len=0 TSval=1179106091 TSecr=340919939
53	0.061698	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=141969 Ack=42401 Win=1373 Len=0 TSval=1179106091 TSecr=340919939
54	0.061835	10.4.88.72	10.71.88.159	SSH	114	Encrypted response packet len=48
55	0.062340	10.4.88.72	10.71.88.159	SSH	15994	Encrypted response packet len=15928
56	0.062344	10.4.88.72	10.71.88.159	SSH	570	Encrypted response packet len=504
57	0.062363	10.4.88.72	10.71.88.159	SSH	386	Encrypted response packet len=320
58	0.064928	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=158769 Ack=44689 Win=1373 Len=0 TSval=1179106094 TSecr=340919942
59	0.065315	10.4.88.72	10.71.88.159	SSH	5858	Encrypted response packet len=5792
60	0.065319	10.4.88.72	10.71.88.159	SSH	1458	Encrypted response packet len=1392
61	0.066938	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=165953 Ack=46185 Win=1373 Len=0 TSval=1179106096 TSecr=340919944
62	0.067182	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=165953 Ack=49081 Win=1373 Len=0 TSval=1179106096 TSecr=340919944
63	0.067429	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=165953 Ack=51977 Win=1373 Len=0 TSval=1179106096 TSecr=340919944
64	0.067697	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=165953 Ack=54425 Win=1373 Len=0 TSval=1179106097 TSecr=340919944
65	0.067890	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=165953 Ack=57409 Win=1373 Len=0 TSval=1179106097 TSecr=340919944
66	0.068441	10.4.88.72	10.71.88.159	SSH	15994	Encrypted response packet len=15928
67	0.068450	10.4.88.72	10.71.88.159	SSH	554	Encrypted response packet len=488
68	0.068544	10.4.88.72	10.71.88.159	SSH	5858	Encrypted response packet len=5792
69	0.068548	10.4.88.72	10.71.88.159	SSH	210	Encrypted response packet len=144
70	0.071459	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=188305 Ack=60305 Win=1373 Len=0 TSval=1179106101 TSecr=340919947
71	0.071704	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=188305 Ack=63201 Win=1373 Len=0 TSval=1179106101 TSecr=340919947
72	0.071832	10.4.88.72	10.71.88.159	TCP	66	ssh > 57913 [ACK] Seq=188305 Ack=64657 Win=1373 Len=0 TSval=1179106101 TSecr=340919947

Figure 32: Wireshark data showing duplicate ACKs

(Source: www.wireshark.org)

10.1.1 Possible Causes

If a transmission is unsuccessful a re-transmission is requested. This can result in the packets being out of order. Another cause of this is that a network device has held a

packet in its buffer until such a time as it is able to transmit it. Both these instances may result in the following:

Packet 1: Sequence 10 - ACK 10
Packet 2: Sequence 20 - ACK 20
Packet 3: Sequence 30 - ACK 30
Packet 4: Sequence 10 - ACK 10

Packet 4 is actually Packet 1 and has been transmitted twice. This would result in the packet analysis showing duplicate ACKs.

TCP re-transmits using fast retransmit. Often this is triggered if there is congestion on the network, which may result in the aforementioned scenario. The fact that there is congestion can also lead to data being lost. This may occur as a result of store and forward switches being used, so data is stored in the buffer until all the data is assembled before being forwarded. An obvious alternative would be to configure switches that are able to cut-through so the data transmission speed is increased as data is passed straight through. This is discussed in more detail in Chapter 6 Role of Switches.

10.2 TCP Timeout Solution

The resulting packet analysis showed that duplicate acknowledgements were issues by TCP, indicating the initial timeout and the retry. Using the tracert utility it was found that the data had been transferred from Sid, but had not made its way out of the Daresbury network. The problem proved to be a combination of the timeout values set on new network routers and the way the switches had been configured for operation. To resolve the issue the default TCP timeout was set to 5 seconds on all new devices. This meant reconfiguring to 12 seconds to allow the data transfer to happen before TCP timed out. Connectivity testing using dummy data proved that the communications were re-established and functioned correctly. Plus the switches that could support cut-through, were reconfigured to this mode of operation allowing for much faster data transmission, without the interference of data being held in buffers. Since these changes were initiated the packet analysis clearly shows a single ACK where before there had been two or three.

Conclusion

Throughout this work, we have demonstrated that collaboration with computing services has been key to a successful installation. Without this support it would have been difficult to gain access to acquire knowledge such as the existing network infrastructure as demonstrated in Chapter 1. This has enabled us to procure the correct devices as outlined in Chapter 2, and to understand the relationship between different vendor switch operating systems.

In Chapter 3 the protocols were considered, and provided a comparison between the two most commonly used interior protocols, drawing a conclusion on the most suitable.

Chapter 4 provided an opportunity to look at data centre topologies, allowing us to consider the best network infrastructure to implement at this time, while also taking into consideration future expansion.

Chapters 5 and 6 looked at the network infrastructure in more detail and detailed the mechanisms for routing and switching data across the network.

There were a number of issues concerning the network infrastructure and the original topology design, some of which are still present. These are identified in Chapter 7 and further supported in Chapters 8 and 9, which looks in more detail at the volume of traffic.

Chapter 10 concludes with an interesting issue with a remote system which shows that new cluster and device implementation doesn't always run smoothly.

It is fair to say that the installation of Sol has been successful. The choice of OSPF as a routing protocol has ensured that alternative vendor equipment has been integrated seamlessly into the university network, despite differences with the way in which VLANs operate. However there are still issues with the data transfer across the network as a result of the clusters being divided over two locations on the campus.

In order to grow and provide an ever improving HPC system to support the work of much of the university's research community, there needs to be a strategy developed alongside Computing Services to ensure the future growth. There needs to be a blueprint for installation of HPC systems that will tie in with the strategies designed by Computing Services. This will help to ensure that the right equipment is purchased and that network services such as DHCP, can be managed centrally.

There needs to be space provided in the data centre to allow for more equipment to be installed, while ensuring that the university's green policy for computing is adhered to. Centralisation of all the HPC resources will also help to eliminate potential network performance problems, whilst also allowing for centralised network management to take place, so statistical performance data can be collected.

Increasingly more universities are consolidating their individual departmental HPC services into one centralised offering. As is the case with the University of Huddersfield, these universities have also grown their networking services in what was initially a haphazard manner, but later conforming to a hierarchical design model. HPC installations have had to adapt to the existing infrastructure rather than being the central feature of the network. This universities have approached HPC installations differently and this has indicated that a blueprint for installation and integration of HPC systems would prove beneficial.

Ghamry's work on using two hops instead of one provides an interesting opportunity to consider how the routing table operates and the possibilities for route consolidation shortest route identification and fastest routes for data transfer over multiple routers.

We have shown that the network has grown substantially and that the new cluster is already heavily utilised.

Future Work

There are a few possibilities for future work arising from this project, some encompassing more technical elements than others and requiring an understanding of algorithms, data transfer and the make-up of packets.

There are possibilities for a further cluster to be installed which again will have an impact on the existing infrastructure. However, this also provides an opportunity to implement IPv6 rather than IPv4. Clearly this will offer further research opportunities as this implementation will not be straightforward. A clear understanding of IPv6 in relation to clusters will be required especially as many IP addresses are still manually configured to allow remote access.

In Chapter 6, we gave some consideration to the router control plane (RCP), which works with the switching fabric to provide a mechanism to transfer data. There is some research which uses an overlay network to simulate different conditions within a production environment without any impact on the existing infrastructure. There is scope to use overlay networks to aid with IPv6 design and implementation.

There are a number of tools that allow data traffic to be analysed by packet type, but these do not identify whether the data is normal network traffic or cluster traffic. However it may be possible to replicate network activity and examine the results through simulation software such as NS3 (http://www.ns3.org). This would allow for real time network activity to be simulated, producing statistical and graphical information for further analysis. While it is easy to see how much network traffic is routed in the direction of any cluster it would be more useful to know which traffic is cluster only. This offers the opportunity to pursue the development of a network monitoring and identification tools that can pick out specific types of data for further analysis. There is also scope for these tools to be utilised for further investigation into duplicate TCP ACKs in respect of network congestion, device utilisation to manage the forwarding process more effectively and big data transfer.

Appendices

Appendix 1: Data Specification for Netgear GSM7228PS Layer 3 Switch

Technical Specifications	
Physical Interfaces	24/48-port Gigabit 802.3af PoE 8 fi rst ports 802.3at PoE+ (30W) 4 shared SFP (Gigabit fi ber) interfaces 2 built-in 10 Gigabit SFP+ (front) 2 additional 10 Gigabit module bays (rear) Physical stacking up to 8 switches/384 ports Stack also with non-PoE GSM73xxS series Removable power supply + RPS option Layer 2+ (Layer 3 lite – IPv4 routing) Fabric (24/48-port) 144 / 192 Gbps Performance 107.1 / 285.7 Mpps L2, L3, L4 ACL (access control lists L2, L3, L4 QoS (8 priority queues, DiffServ) IGMP snooping v2, v3 IGMP proxy, IGMP querier 8,000 MAC – 1,024 VLANs – 1,024 MC groups 64 trunks 8-port each – DHCP server/relay 224 IP routes – 128 IP interfaces
POE	All 24 Gigabit RJ45 ports are PoE IEEE® 802.3af (up to 15.4 Watts/port)
POE+	The fi rst 8 Gigabit RJ45 ports are PoE+ IEEE 802.3at (up to 30 Watts/port)
Processor / Memory	Processor: MPC8633 @ 666 MHz System memory: 256 MB (RAM) Packet buffer memory: 0.75 MB per switch Code storage (fl ash): 32 MB
Performance Summary	Switching fabric: 144 Gbps Throughput: 107.1 Mpps Forwarding mode: Store-and-forward Latency (64-byte frames, 10 to 100 Mbps): <35.2μs Latency (64-byte frames, 1 Gbps): <4.1μs Latency (64-byte frames, 10 Gbps): <2.0μs Addressing: 48-bit MAC address Address database size: 8,000 MAC addresses Number of VLANs: 1,024 (IEEE 802.1Q) Number of multicast groups fi ltered: 1,024 Number of trunks: 64 trunks, 8-port per trunk Number of hardware queues for QoS: 8 Number of static routes: 224

	Number of IP interfaces: 128 Jumbo frame support: up to 9K packet size Acoustic noise (ANSI-S10.12): 44 dB @ 25°C ambient temperature Heat dissipation: 260.49 Btu/hr Mean time between failures (MTBF): 211,069 hours (~24.1 years) @ 25 °C and98,705 hours (~11.3 years) @ 55 °C ambient temperature
L3 Services - ROUTING	L2+ static routing (Subnets, VLANs) 224 IP routes (L3-capable hardware) 128 IP interfaces
L3 Services - DHCP	DHCP server (1,024 clients) DHCP L2 relay, DHCP snooping

Appendix 2: Data Specification for Cisco Nexus 5000 Series Layer 3 Switch

Technical Specification	
Performance	Cisco Nexus 5010: Layer 2 hardware forwarding at 520 Gbps or 386.9 mpps MAC address table entries: 16,000 Low-latency cut-through design that provides predictable, consistent traffic latency regardless of packet size, traffic pattern, or enabled features on 10 Gigabit Ethernet interfaces Line-rate traffic throughput on all ports
Interfaces	Cisco Nexus 5020: 40 fixed 10 Gigabit Ethernet and FCoE ports (ports 1 to 16 are Gigabit Ethernet and 10 Gigabit Ethernet); additional interfaces through two expansion modules Cisco Nexus 5010: 20 fixed 10 Gigabit Ethernet and FCoE Ports (ports 1 to 8 are Gigabit Ethernet and 10 Gigabit Ethernet); additional interfaces through one expansion module Expansion modules: 6-port 10 Gigabit Ethernet and FCoE module 4-port 4/2/1-Gbps Fibre Channel plus 4-port 10 Gigabit Ethernet and FCoE module 8-port native 4/2/1-Gbps Fibre Channel expansion module 6-port native 8/4/2/1-Gbps Fibre Channel expansion module Extension through the Cisco Nexus 2000 Series (up to 12 fabric extenders per Cisco Nexus 5000 Series Switch)
Layer 2 Features	Layer 2 switch ports and VLAN trunks IEEE 802.1Q VLAN encapsulation Support for up to 507 VLANs per switch Rapid Per-VLAN Spanning Tree Plus (PVRST+) (IEEE 802.1w compatible) Multiple Spanning Tree Protocol (MSTP) (IEEE 802.1s): 64 instances Spanning Tree PortFast and PortFast Guard Spanning Tree UplinkFast and BackboneFast Spanning Tree Root Guard Spanning Tree Bridge Assurance Cisco vPC technology Link Aggregation Control Protocol (LACP): IEEE 802.3ad Advanced PortChannel hashing based on Layer 2, 3, and 4 information Jumbo frames on all ports (up to 9216 bytes) Pause frames (IEEE 802.3x) Storm control (unicast, multicast, and broadcast) Private VLANs Private VLAN over trunks (isolated and promiscuous)

	Private VLANs over vPC and EtherChannels VTP Client and Server (Versions 1 and 2)
QoS	Layer 2 IEEE 802.1p (CoS) 8 hardware queues per port Per-port QoS configuration CoS trust CoS marking Port-based CoS assignment Modular QoS CLI (MQC) compliance Access control list (ACL)-based QoS classification (Layers 2, 3 and 4) MQC CoS marking Per-port virtual output queuing CoS-based egress queuing Egress strict-priority queuing Egress port-based scheduling: Weighted Round-Robin (WRR)
Security	Ingress ACLs (standard and extended) on Ethernet and virtual Ethernet ports Standard and extended Layer 2 ACLs: MAC addresses, protocol type, etc. Standard and extended Layer 3 to 4 ACLs: IPv4 and v6, Internet Control Message Protocol (ICMP), TCP, User Datagram Protocol (UDP), etc. VLAN-based ACLs (VACLs) Port-based ACLs (PACLs) Named ACLs ACL logging and statistics Optimized ACL distribution ACLs on VTY Dynamic Host Configuration Protocol (DHCP) snooping with Option 82
High-Availability Features	ISSU Hot-swappable field-replaceable power supplies, fan modules, and expansion modules 1:1 power redundancy N:1 fan module redundancy
Management	Switch management using 10/100/1000-Mbps management or console ports CLI-based console to provide detailed out-of-band management In-band switch management Locator and beacon LEDs on Cisco Nexus 2000 Series vPC configuration synchronization Module preprovisioning Configuration rollback Port-based locator and beacon LEDs SSHv2

	Telnet Authentication, authorization, and accounting (AAA) AAA with RBAC RADIUS TACACS+ Syslog Embedded packet analyser SNMP v1, v2, and v3 Enhanced SNMP MIB support XML (NETCONF) support Remote monitoring (RMON) Advanced Encryption Standard (AES) for management traffic Unified username and passwords across CLI and SNMP Microsoft Challenge Handshake Authentication Protocol (MS-CHAP) Digital certificates for management between switch and RADIUS server Cisco Discovery Protocol Versions 1 and 2 RBAC Switched Port Analyzer (SPAN) on physical, PortChannel, VLAN, and Fibre Channel interfaces

Appendix 3: Data Specification for Nortel Baystack 5510-T layer 3 Switch

Technical Specifications	
Performance Specifications	160 Gbps for the switch; Up to 1,280 Gbps for the full stack Stacking bandwidth: 80 Gbps for the switch; Up to 640 Gbps for the full stack Maximum data throughput: 768 Gbps for a full stack of BayStack 5520-48T-PWR Frame forwarding rate: • 71.4 Mpps (million packets per second) for the BayStack 5520-48T-PWR • 35.7 Mpps for the BayStack 5520-24T-PWR • 571.4 Mpps for a full stack of 8 BayStack 5520-48T-PWR units
Port forwarding/filtering performance	For 10 Mbps: 14,880 pps maximum (64-byte packets) For 100 Mbps: 148,810 pps maximum For 1000 Mbps: 1,488,100 pps maximum Address database size: 48,000 entries at line rate (48,000 entries without flooding) Addressing: 48-bit MAC address Frame length: 64 to 1518 bytes (IEEE 802.1Q Untagged) 64 to 1,522 bytes (IEEE 802.1Q Tagged) Jumbo frame support: Up to 9,216 bytes Multi-Link Trunks: Up to six trunks, four members per trunk VLANs: Up to 256 port- or protocol-based per VLAN Tagging option Multiple Spanning Tree Groups: Up to eight STGs
Interface options	10BASE-T/100BASE-TX/1000BASE-T: RJ-45 (8-pin modular) connectors for Auto MDI/MDI-X interface with auto-polarity The BayStack 5510 Switches support the following SFP GBICs: 1000BASE-SX Uses short wavelength 850 nm MTRJ or LC type fiber optic connectors to connect devices over multimode (275m, 62.5um core or 550m, 50.0um core) fiber optic cable 1000BASE-LX: Uses long wavelength 1300nm duplex LC type fiber optic connector to connect devices over single mode (10km, 9um core) fiber optic cable 1000BASE-CWDM: Uses long wavelength 1470, 1490, 1510, 1530, 1550, 1570, 1590, 1610nm LC type fiber optic connector to connect devices over single mode (40km, 9um core or 70km, 9um core) fiber optic cable
Network protocol and	IEEE 802.3 10BASE-T (ISO/IEC 8802 3, Clause 14)

standards compatibility	IEEE 802.3u 100BASE-TX (ISO/IEC 8802-3, Clause 25) IEEE 802.3u Autonegotiation on Twisted Pair (ISO/IEC 8802-3, Clause 28) IEEE 802.3x (Flow Control on the Gigabit Uplink port) IEEE 802.3z 1000BASE-SX and 1000BASE-LX IEEE 802.1d MAC Bridges (ISO/IEC 10038) IEEE 802.1p (Prioritizing) IEEE 802.1Q (VLAN Tagging) IEEE 802.1D Spanning Tree Protocol IEEE 802.3ad (manual/static) IEEE 802.3ad (LACP)† IEEE 802.1s† IEEE 802.1w† IETF DiffServ
RFC support	RFC 1213 (MIB-II); RFC 1493 (Bridge MIB); RFC 2863 (Interfaces Group MIB); RFC 2665 (Ethernet MIB); RFC 2737 (Entity MIBv2); RFC 2819 (RMON MIB); RFC 1757 (RMON); RFC 1271 (RMON); RFC 1157 (SNMP); RFC 2748 (COPS); RFC 2940 (COPS Clients); RFC 3084 (COPS Provisioning); RFC 2570 (SNMPv3); RFC 2571 (SNMP Frameworks); RFC 2573 (SNMPv3 Applications); RFC 2574 (SNMPv3 USM); RFC 2575 (SNMPv3 VACM); RFC 2576 (SNMPv3); RFC 2572 (SNMP Message Processing) RFC 791 (IP); RFC 792 (ICMP); RFC 793 (TCP); RFC 783 (TFTP); RFC 826 (ARP); RFC 768 (UDP); RFC 854 (TELNET); RFC951 (Bootp); RFC 2236 (IGMPv2); RFC 1112 (IGMPv1); RFC 1945 (HTTP v1.0); RFC 2138 (RADIUS); RFC 894 (IP over Ethernet); RFC 2674 (Q MIB); RFC 1058/RFC 1723 (RIPv1/v2); RFC 2178 and RFC1583 (OSPF); RFC 2030 (SNTP (Simple NTP))

Bibliography

Dye, Mark A., McDonald, Rick, Rufi, Antoon, W. (2008) Network Fundamentals, CCNA Exploration Companion Guide, Cisco Press

Lewis, W. (2008) LAN Switching and Wireless, CCNA Exploration Companion Guide, Cisco Press.

McCabe, James D.. Network Analysis, Architecture, and Design, Third Edition. Morgan Kaufmann Publishers. © 2007. Books24x7. http://common.books24x7.com.libaccess.hud.ac.uk/toc.aspx?bookid=37217> (accessed April 17, 2012)

Kureshi, I. (2010) Establishing a University Grid for HPC Applications, Published by the University of Huddersfield

McQuerry, S. (2008) Interconnecting Cisco Network Devices, Part 2 (ICDN2), Third Edition, Cisco Press

Mikalsen, A. and Brogesen, P. (2002) Local Area Network Management, Design and Security, A Practical Approach, John Wiley & sons Ltd.

Olifer, N. and Olifer, V. (2006) Computer Networks, Principles, Technologies and Protocols for Network Design, John Wiley & sons Ltd.

Sivasubramanian, B., Froom, R. and Frahim, E. (2010) Implementing Cisco Switched Networks (SWITCH), Foundation Learning Guide, Cisco Press.

Teare, D. (2010) Implementing Cisco IP Routing (ROUTE) Foundation Learning Guide, Cisso Press.

Journals

Altiparmak, F., Dengiz, B., Smith, A. E., Optimal Design of Reliable Computer Networks: A Comparison of Metaheuristics, Journal of Heuristic, Vol. 9, No. 6, 2003-12-01, Kluwer Academic Publishers

Ausavarungnirun, Rachata, Kai-Wei Chang, Kevin, Fallin, Chris, Mutlu, Onur, Adaptive Cluster Throttling: Improving High-Load Performance in Bufferless On-Chip Networks, Computer Architecture Lab (CALCM), Carnegie Mellon University

SAFARI Technical Report No. 2011-006, September 6, 2011, Accessed 27th December 2011

Baker, Mark, Cluster Computing White Paper, Version 2.0, Date – 28th December 2000, University of Portsmouth, UK

Freire, M.M., The Differentiated Services Architecture. Encyclopaedia of Internet Technologies and Applications. IGI Global, 01 Jan 2008

Freire MM. Ethernet to the Doorstep of Metropolitan Area Networks. Encyclopaedia of Internet Technologies and Applications.: IGI Global, 01 Jan 2008

Freire MM. Quality of Service Routing. Encyclopaedia of Internet Technologies and Applications.: IGI Global, 01 Jan 2008

Ghamry, W.K.,Elsayed, M.F. and Nassar, A.M. On the Interplay of Network Structure and Routing Strategies for Performance in Scale-Free Networks, IEEE 2009

Kim, J., Chandra, A., and Weissman, J.B., Passive Network Performance Estimation for Large-Scale, Data-Intensive Computing, IEEE Transactions on Parallel and Distributed Systems, Vol. 22,No. 8, August 2011

MMAE Lima, NLS da Fonseca, Active Queue Management, Encyclopaedia of Internet Technologies and Applications, 2008 - Information Science Publishing

K.-K. Nguyen; H. Mahkoum; B. Jaumard; C. Assi; M. Lanoue; , "Toward a Distributed Control Plane Architecture for Next Generation Routers," Universal Multiservice Networks, 2007. ECUMN '07. Fourth European Conference on , vol., no., pp.173-182, Feb. 2007 doi:10.1109/ECUMN.2007.53
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4115201&isnumber=4115172, accessed September 2012

Pal, S., Sarma, S., computer Network Topology Design in Limelight of Pascal Grapsh Property, International Journal of Next-Generation Networks, 2.1 (2010) 30-35; doi:10.5121/ijngn.2010.2103

Pierre, S., Beaubrun, R. Integrating routing and survivability in fault-tolerant computer network design, Computer Communications, Volume 23, Issue 4, 15 February 2000, Pages 317-327, ISSN 0140-3664, 10.1016/S0140-3664(99)00171-1.

Rakheja, P. Kaur, P. Gupta, A. and Sharma, A. Performance Analysis of RIP, OSPF, IGRP and EIGRP Routing Protocols in a Network. International Journal of Computer Applications 48(18):6-11, June 2012. Published by Foundation of Computer Science, New York, USA

Singla, A., Chi-Yao, H. Popa, Brighten, L., Godfrey, P., Jellyfish: Networking Data Centers Randomly, Computer Science - Networking and Internet Architecture, eprint arXiv:1110.1687, 10/2011, Published by Cornell University Library

Szlachcic, E., Fault Tolerant Topological Design for Computer Networks, 2006 .International Conference on Dependability of Computer Systems, 2006 IEEE

Yang Y. (2012) Understanding Switch Latency, White Paper, Cisco

Web References

Cut-Through and Store-and-Forward Ethernet Switching for Low-Latency Environments, [Online] Available http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-465436.pdf Cut-Through and Store-and-Forward Ethernet Switching for Low-Latency Environments, accessed September 2012

File:Tcp state diagram new.svg [Online] Available http://commons.wikimedia.org/wiki/File:Tcp_state_diagram_new.svg, accessed July 2012

The Transmission Control Protocol,[Online] Available http://condor.depaul.edu/jkristof/technotes/tcp.html accessed October 2012

Ganglia Monitoring System, (2001, Jan), [Online] Available http://ganglia.sourceforge.net/ accessed September 2012

Store and Forward vs Cut-Through, [Online] Available http://www.inetdaemon.com/tutorials/networking/lan/switching/architectures/store_vs_cut.shtml, accessed September 2012 accessed July 2012

EIGRP vs OSPF, (2007 May) [Online] Available http://www.networkworld.com/community/node/16276, accessed July 2012

NS Network Simulator 2 (2009, December) [Online] Available http://www.nsnam.org/overview/what-is-ns-3, accessed January 2013

OSPF, (2009, September) [Online] Available http://packetlife.net/captures/protocol/ospf, accessed July 2012

What causes duplicate ACK records, (2011, May) [Online] Available http://serverfault.com/questions/266764/what-causes-duplicate-ack-records accessed November 2012