LARGING IT FOR THE GRID: BIG NETWORKING FOR BIG SCIENCE

Saleem N. Bhatti <s.bhatti@cs.ucl.ac.uk> and Peter Clarke <clarke@hep.ucl.ac.uk> (on behalf of MB-NG and UKLIGHT)
Networked Systems Research Group, UCL, http://www.netsys.ucl.ac.uk/

Key words to describe the work: quality of service (QoS), network control, resource management, high-speed networking.
Key Results: in progress – national and international high-speed, manageable, flexible, QoS-controlled network connectivity.
How does the work advance the state-of-the-art?: to provide control of the network for users from their desktops whilst still allowing network administrators to operate the network to meet users' needs in a scaleable with respect to user demand.
Motivation (problems addressed): the need to enable connectivity for distributed applications, with requirements to access large amounts of data remotely (e.g. bio-informatics, high-energy physics, radio astronomy, socio-economic data-mining) or with requirements for real-time interaction (e.g. distributed simulation, distributed control, real-time remote visualisation, high-quality video and audio for remote language teaching or conferencing)

Let's go large

A researcher in the Bio-informatics unit in Glasgow would like to access the database from the Human Genome Project (HGP) at Cambridge for performing a specific task. The researcher initiates the task (e.g. interactive visualisation tool) for a specified data set. The task attempts to mount and access the data. If the data is locally cached, the copy will be used. Alternative access modes might be to automatically start a file-transfer or to remotely mount the data over storage area network (SAN) technology. In either case, for the sizes of data set being considered, high capacity connectivity will be used required. The researcher has previously joined the virtual community of HGP collaborators and was issued a certificate authenticating him. This is used by the application (after local identity verification using a thumbprint scan) to access the data set and the processing facilities. The application starts, establishes the appropriate connectivity including access to the HGP database and checks the user's credentials. The researcher is granted appropriate access rights to the database. The application locates the HGP database and establishes access (locally or remotely). The researcher can then view the data or await the completion of the analysis.

In the scenario above, there are today components to enable much of what is required to make that scenario real. However, the HGP database is currently ~0.25Pbytes of data (growing at ~1Tbyte per week). The SuperJANET4 backbone has a maximum capacity of 10Gbs. In the extreme, the complete file transfer of the HGP database will take 200,000s or ~55½ hours – and this is if a single user is given the whole of the raw backbone capacity of the current SuperJANET4 backbone. This is of course, not possible. In other words, to make that scenario (and other more complex scenarios) real, and to allow many such users for the near future, we need one very important enabling component – ***a very high capacity, dynamically controllable network infrastructure that is accessible to researchers from their desktop.***

The changing landscape in networking

Today, we reach a point where multi-gigabit wide-area connectivity exists, and access speeds are commonly multi-megabits. Desktop connectivity is both possible and affordable at 1Gb/s and before long research users will be able to use that capacity with Grid applications, for example. The recent emergence of products supporting 10Gb/s technology at three times the price of 1Gb/s access speeds has served to depress costs per network port even further. And the number of users continues to increase. We are moving towards a situation where the traffic from the access networks has the potential to swamp the core WAN links. In the past, providers have relied on over-provisioning to cope with unpredictable and changing traffic patterns, but it seems unlikely that this method of network capacity planning will be viable for the future.

Academic users are also starting to run more capacity hungry and QoS-sensitive applications for both teaching and research, for example voice and video conferencing and distributed data-processing or visualisation using large, remote computer-clusters.

As networking and networking components become more sophisticated, building, simulating, testing, managing and controlling such networks becomes increasingly difficult. The complex system components and protocols when connected together and driven by application traffic exhibit an emergent behaviour that can be hard to model and predict – the outcome of putting some network elements together produces a result that is more that just the sum of the parts. This unpredictable emergent behaviour is compounded by the application-level infrastructure, which users now favour to support virtual organisations and the formation of dynamic communities. Indeed, many of these issues (specifically relating to virtual organisations and communities, complexity in systems and applications, making systems dynamically adaptable, and building some autonomy into systems so that they can become more manageable) are also highlighted as areas for future research in a recent document from the National eScience Centre (NeSC) called "Computer Challenges to emerge from eScience"[1]. This is a great challenge.

Meeting the challenge

There are two broad areas of work, here posed as questions, on which we are currently focusing our attention:

- **control:** how do we provide design and provide network mechanisms that allow very flexible and dynamic control for a network infrastructure across multiple administrative domains?

1 http://umbriel.dcs.gla.ac.uk/NeSC/general/news/↴
uk_escience_agenda.html

- **capacity:** what happens when we need to run the network at very high data rates (many Gbps) and the users themselves have a capability of generating very high network loads (certainly 100s of Mbps and even several Gbps)?

To meet the challenge requires technology trials with real equipment and networks as well as "good old-fashioned research" to engage in a harmonious partnership. Indeed, our interactions so far involve researchers and users from computer science, electronic engineering and high-energy physics (from UK and abroad) as well as UKERNA and various equipment manufacturers and network service providers.

Control: The MB-NG project (May 2002 – April 2004)

The work currently being carried out by the MB-NG (Managed Bandwidth Next Generation) project tries to address the issues encountered when trying to control the network infrastructure. The primary objective of MB-NG is to demonstrate e2e (end-to-end) managed bandwidth services in a multi-domain environment, in the context of GRID project requirements. By "e2e" we mean that it should be possible to have enough control that (at least from site-to-site) it is possible to have managed network capacity at the granularity of individual end-systems. By "multi-domain" we mean that the manageability and control required to give such e2e capability should not be constrained by administrative boundaries in the network.

Project partners include UKERNA and Cisco as well as various universities. UKERNA hopes to use this project to find a replacement for their old pilot "managed-bandwidth" service, which relied on the use of specific underlying networking connectivity (asynchronous transfer mode – ATM). The MB-NG service will not have such restrictions, as it is based on protocols that are independent of lower-level technologies.

The QoS capability control for the MB-NG network is to be provided by a combination of *differentiated services (DIFFSERV)* and *multi-protocol label switching (MPLS)*.

DIFFSERV works by marking packets with a small bit-pattern that identifies them as belonging to a particular DIFFSERV class. Each DIFFSERV class is defined to have a particular handling when being forwarded by network routers. So, effectively, this is a class-based service that requires packets at the ingress to the network to be subject to some initial filtering or classification and then marked as belonging to a particular DIFFSERV class. Controlling which packets are marked at the ingress allows control over which applications or users, for example, receive a specific QoS from the network. The classes that are currently being planned for use with MB-NG are called *expedited forwarding (EF)* and *less-than-best-effort (LBE)*. All "normal" traffic has no marking and is referred to as *best-effort (BE)*. EF is intended for use by high-priority traffic flows, but specifically for flows with requirements of low loss and low delay. EF flows do not necessarily have to require large data rates. Applications that might use EF are real-time video streams or high-priority file transfers. LBE is intended for use by long-lived flows that may be considered as very-low priority in the presence of other traffic. LBE is also referred to as a "scavenger service". This means that when there is no other traffic, an LBE traffic flow can use as much of the network capacity as it needs. If other flows appear, either BE or EF, then the LBE traffic is permitted to use less of the available network capacity. LBE is intended for such flows as very large, but low-priority, file transfers or applications requiring long-lived, "trickle feed" updates, such as data-caches or replicated data-bases.

MPLS will be used in the core of the network to provide high-performance and manageable provisioning. [It is also possible to use MPLS for the formation of virtual private networks (VPNs) to support virtual organisations.]

The three main MB-NG sites – UCL, Manchester and RAL – will be connected via the SuperJANET4 development network. The site connectivity will be 2.5Gbps.

The MB-NG network is currently being built and one of the first activities will be looking at the performance of the network. Accurate monitoring and performance evaluation, especially stress-tests, are challenging when working at such speeds so we have adapted some traditional techniques and developed some new ones in order to measure performance.

Capacity: The UKLIGHT initiative

Although the MB-NG network runs at quite high speeds (2.5Gbps), future GRID users may require much higher data rates, as our scenario at the beginning of this paper illustrates. To achieve higher data rates, we need to use optical networking. Many of the members of MB-NG are also part of another initiative looking at the provisioning of extremely high-capacity connectivity at the international level as well as nationally. The initiative is currently named *UKLIGHT* and we are in the process of formulating a scientific case for the provision of a large optical networking infrastructure for the purposes of networked systems research. The provision of UKLIGHT will allow us to interconnect with similar facilities that already exist in Europe and the US. Also, it will provide a learning experience for UKERNA to plan for future HE network provisioning. Of course, UKLIGHT is intended to provide an excellent platform for very applied and forward looking research into optical transmission systems, very high-speed networking, highly configurable networks, high-speed QoS control, network security systems and adaptable protocols and applications.

Acknowledgements and further information

Some of the text in this paper is taken from a document produced by UKLIGHT. There are many people contributing to this work, throughout the UK, in the projects:

MB-NG – http://www.mb-ng.net/

UKLIGHT – http://www.cs.ucl.ac.uk/research/uklight/

Other related projects are:

DataTAG – http://www.datatag.org/

EU-DataGrid – http://www.eu-datagrid.org/

6NET – http://www.6net.org/