High Performance Networks for Data Intensive Science

Craig Tomkow

January 29, 2017

Athabasca University

COMP601 - Survey of Computing and Information Systems

Richard Huntrods

Abstract

Data intensive science that relies on grid computing requires high performance networks to transfer petabytes of data. A typical enterprise network cannot provide the necessary performance for transferring petabytes of data over long distances. There are various data intensive scientific disciplines, such as climate modeling and high energy particle physics, that rely on high throughput data transfers. Without modification, high bandwidth, low latency networks suffer from poor network performance and slow data transfers. Custom data transfer applications are required to achieve efficient TCP operation through parallelism and operating system modification. 100 gbps wide area networks have been deployed across countries to provide the necessary bandwidth required. Finally, network architecture patterns has been created for optimal local area network deployment. By carefully engineering all the various parts, a high performance network can be realized.

1. Introduction

Geographically distributed computing (grid computing) used in various research fields rely on computer networks for data transfers between geographically separate compute clusters. For data intensive research, high performance networks are needed to efficiently transfer large amounts of data between collaborating organizations and compute resources. More and more research fields are relying on data intensive computing which in turn creates an increasing demand for high performance networking. These various domains range from high energy particle physics to climate modeling. To solve the issue of achieving highly efficient data transfers, a sophisticated approach is required. Attempting to naively deploy a commodity network stack will result in poor data transfer speeds for large data sets. Many enhanced data transfer applications have been devised such as gridFTP [9] and MMCFTP [13]. Building upon the success of gridFTP, research has resulted in novel enhancements such as hardware accelerated gridFTP [11]. Meanwhile, other research has investigated abandoning the transmission control protocol (TCP) altogether as the main transport protocol and utilizing the user datagram protocol (UDP) with application level reliable delivery [5]. Remembering that overall network performance relies on end to end high performance, advanced 100 gbps wide area networks (WAN) such as the Internet2 [12] and ESnet [1,6] have been deployed. Additionally, optimal local area network (LAN) design, such as the science DMZ [4], has been embraced. Ultimately, data transfer performance is affected by every part of the network stack and necessitates a well thought out and sophisticated implementation at all layers from the applications and hardware to the WAN and LAN.

2. Background

Data intensive research relies heavily on grid computing to perform the necessary computations at scale. A grid is the idea of having access to a vast range of specialized compute, storage, and network resources that are geographically distributed to achieve large scale computing. This differs from cluster computing where all compute resources are localized or cloud computing which is a commodity resource that cannot accommodate various unique requirements such as needing high performance computing (HPC) [8]. Access to a wide range of organizations compute power is enabled through a standardized user and infrastructure level middleware (enabling resource brokering, accounting, authentication, data transfer) that sits on top of the low level infrastructure that is unique to each organization. This grid infrastructure addresses the two main issues in geographically diverse computing which is coordination and distribution [2].

The grid architecture relies on data transfers as a core service to effectively move data sets between the various geographically dispersed resources. Many grid architectures and associated middleware standardize on the widely deployed and accepted TCP/IP network stack for interoperable data transmission. The transmission control protocol (TCP) was designed for achieving reliable delivery of data; performance was not the main concern. Furthermore, due to TCPs sliding window design which determines the amount of data to be transmitted, or ‘in flight’, performance drops as bandwidth and latency increase. The theoretical maximum amount of data in transit at once is determined by the bandwidth-delay product (BDP). The Mathis Equation [10] predicts the performance of TCP given a certain percentage of data loss. Given that wide area network connectivity between organizations can be limited [7], the need for maximizing throughput is important. And even moreso, data intensive research fields that produce petabytes of data critically rely on high performance networks that achieve maximum throughput.

3. High Performance Networks

1. Specialized vs General Purpose

The network requirements for efficient transfer of large data sets go beyond a typical enterprise network. The various factors that need to be accounted for are many. Eliminating packet loss along the path the traffic takes is the single most important goal when deploying a network for science research. Put more simply, “if the science applications cannot achieve adequate performance, the science mission of the infrastructure has failed” [4]. An organization's typical network cannot achieve this mission due to various factors such as: inadequate WAN bandwidth, packet loss in the WAN or LAN, poor application performance, and unnecessary potential failure domains. In addition, typical network monitoring tools can be inadequate in detecting minor network issues that can have a major impact on end to end performance. It is necessary to take extra care in designing the network and not be tempted to rely on existing general network engineering best practices if they work against achieving high network performance.

2. Requirements for Data Intensive Science

Having established some concerns with utilizing commodity enterprise networks for high performance networking, a wise question to ask is when is high performance necessary? It is trivial to state that high performing networks are needed in data intensive scientific research. However, having more specific statements regarding the special needs that require such networks are required. There are two key elements that determine the need for this scale of networking. The area of research must generate huge amounts of data (petabytes) and also must either rely on grid computing for distributing workloads or the need to collaborate with other institutions. Given these special needs that demand an exceptional network, it is best to look at some examples of current fields that make use of these networks. Climate modeling research is a field that generates massive data sets that need to be shared between collaborators and institutions. Petabytes of data is generated in climate modeling research. This data is typically processed on supercomputers located at offsite locations such as the National Energy Research Scientific Computing Center (NERSC) which is located in Berkeley, California [3]. Therefore, it is necessary to utilize a high performance research network to transfer the massive amounts of data. Another field that requires high performing networks is in high energy particle physics. Scientific research that is performed at sites such as the Large Hadron Collider (LHC) are steadily increasing the size of data sets generated. For example, the CMS heavy ion experiments being conducted at the LHC are producing 1-2 petabytes of data a year [14]. This incredible generation of data is increasing the need for 100 gbps networks and beyond. Due to the LHC being located in Europe, the need for high performing networks that span international distances is a necessity, not a luxury. As ubiquitous, distributed, large scale computing is enabling big data within more and more research fields, the need for high performing networks will continue to grow.

3. High Bandwidth Wide Area Networks

At the core of computing is invariably the hardware which dictates, at an individual level, the total performance that can be achieved. With computer networks, the maximum bandwidth is limited by the hardware that exists to transmit the data at a certain line rate. Given the need for large amounts of bandwidth, there has been significant development of high capacity switching, routing, and optical hardware. Currently 100 gbps hardware is being being produced and deployed with even larger bandwidths being realized at 400 gigabit and terabit speeds. The Internet2 is an initiative that was started in 1996 to create a new type of wide area network; a high performance network to serves advanced research and spur the standardization of future technologies [12]. The Internet2 has deployed 100 gbps hardware which meets the demand for data intensive science. In addition, the Energy Science Network (ESnet) deployed a 100 gbps network in cooperation with Internet2 to enable data intensive research [1,6]. ESnet's 100 gbps network connects to Europe and is heavily utilized by researchers at the Large Hadron Collider. ESnet carries on average 20 petabytes of monthly data while estimating the need to transfer over 100 petabytes a month by 2016 [6]. Aside from these mainly North American networks, there exists many other international networks: the NORDUnet that interconnects various Nordic countries, SLIX in Singapore, GÉANT in Europe, and many more. All these networks share the requirement for 100 gbps hardware to enable large data transfers.

4. TCP Efficiency

Special software is needed to initiate an efficient data transfer over a high bandwidth, high latency network. The de facto software platform in grid computing is gridFTP. The software package implements various features to realize high throughput. The main method that is commonly used is to create many parallel TCP connections for data transmission. Research has shown [5,10] that single TCP streams over high bandwidth, high latency links perform poorly. Therefore, the natural solution was to pipeline many TCP streams to achieve high throughput. Other research has explored utilizing UDP data transfers while relying on application level reliable delivery [5]. GridFTP has been expanded to allow the option of using UDP as another solution to addressing the throughput problem. Furthermore, the underlying operating system that contains the TCP/IP stack also needs to be considered. Since TCP using windowing, the operating system negotiates various settings during the TCP connection establishment such as window size and scaling. In addition, the operating system defines the amount of buffering needed for a network connection. These details are critical in determining the efficiency of TCP. GridFTP is able to automatically tune these system details to ensure optimal TCP performance. Finally, there has been unique research done to enhance GridFTPs efficiency through hardware assisted NIC offloading [11]. Ultimately, the network protocols, operating system network stack, and user level applications all play an important part in enabling a high performance network.

5. Local Area Network Architecture

To achieve a high performance network you need connectivity to the appropriate wide area network, use the appropriate data transfer applications, and ensure the operating systems are properly tuned. Given that these independent parts need to be carefully engineered to achieve the desired results; there was an opportunity to create a network design pattern to outline the requirements for a high performance network. The Science DMZ [4], created by E. Dart et al., defines a best practice architecture for building a local area network that is optimized for high throughput networking. There are four key designs that are outlined that aid in constructing a best practice network: physical location, dedicated systems, performance monitoring, and security. The physical location recommendations are meant to separate the high performance network from the organization's enterprise network as well as reducing complexity [4]. Isolation from the general network is necessary in reducing complexity as well as minimizing the failure domain. The recommendations for dedicated systems refer to purpose built compute, storage, applications, and network hardware for optimizing data transfer. It is important to virtually eliminate packet loss. Therefore, details such as consistent bandwidth sizing throughout the network is important. When mismatched ingress and egress interfaces exist, packet buffering is necessary which almost always causes packet loss during high utilization. The third design recommendation emphasizes the need for performance monitoring in detecting soft errors within a network [4]. Issues such as silently failing hardware must be detectable as even small packet loss causes big performance issues on high bandwidth, high latency networks. Finally the correct security solutions should be adopted. It is recommended that traditional firewalls are unnecessary on a DMZ network. This is mainly due to the network being segregated from the general purpose network. The correct use of ACL's (that are line rate hardware processed) can then be applied to only the scientific traffic and applications [4]. The Science DMZ architecture provides a robust framework for building out a highly efficient network infrastructure.

4. Conclusion

To achieve a high performance network capable of efficiently transferring petabytes of data, many considerations must be made. A dedicated high performance network architecture must be considered as typical enterprise networks are insufficient. Access to a high bandwidth (100 gbps) wide area network is important to provide a maximum data transfer speed. The data transfer applications must be carefully considered and deployed. Without efficient TCP performance, high throughput cannot be achieved. Additionally, the local area network must be properly deployed to ensure minimal packet loss and minimize failure domains. Given the increasing dependence on distributed computing and the consistent growth in research data, high performance networks will continue to be necessary. Therefore, it is more important than ever to understand the current problems facing high performance networking and ensuring proper solutions are deployed.

References

[1] J. Bashor, (2012, November 13). Department of Energy's ESnet rolls out world's fastest science network [Online]. Available: http://cs.lbl.gov/news-media/news/2012/100-gigabit-network/. [Accessed: 28-Jan-2017].

[2] F. Berman et al., “The Grid: past, present, future,” in *Grid Computing - Making the Global Infrastructure a Reality*, Chichester, England: John Wiley & Sons Ltd., 2003, ch 1, pp. 9-47.

[3] D. Bernholdt et al., “The Earth System Grid: Supporting the Next Generation of Climate Modeling Research,” *Proc. of the IEEE*, vol. 93, no. 3, pp. 485-495, Mar., 2005.

[4] E. Dart et al., “The Science DMZ: A network design pattern for data-intensive science,” *Sci. Programming,* vol. 22, no. 2, pp. 173-185, 2014.

[5] P. Dickens et al., “High Performance Wide Area Data Transfers Over High Performance Networks,” in International Parallel and Distributed Processing Symposium, 2002 © IEEE. doi: 10.1109/IPDPS.2002.1016675.

[6] ESnet. The Network [Online]. Available: https://es.net/engineering-services/the-network/. [Accessed: 28-Jan-2017].

[7] F. Feldhaus et al*., “*State-of-the-Art Technologies for Large-Scale Computing*,”* in *Large-Scale Computing Techniques For Complex System Simulations,* W. Dubitzky et al., 1st ed. New Jersey: John Wiley & Sons, Inc., 2012, ch. 1, pp. 1-17.

[8] I. Foster et al., “Cloud Computing and Grid Computing 360-Degree Compared,” in *IEEE Grid Computing Environments*, TX, 2008, pp. 1-10.

[9] F. Lyon, “Globus GridFTP: what’s new in 2007,” in *GridNets ‘07*, Brussels, 2007, no. 19.

[10] M. Mathis et al., “The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm,” *ACM SIGCOMM Comput. Commun. Review*, vol. 27, no. 3, pp. 67-82, Jul., 1997.

[11] M. Rashti et al., “Long-haul secure data transfer using hardware-assisted GridFTP,” *Future Generations Comput. Syst.*, vol 56, pp 265-276, 2016.

[12] L. Saunders-McMaster, “Internet 2: an overview of the next generation of the Internet,” *Comput. in Libraries*, vol. 17, no. 3, pp. 57, Mar., 1997.

[13] K. Yamanaka et al., “Long distance fast data transfer experiments for the ITER Remote Experiment,” *Fusion Eng. and Design*, vol. 112, pp. 1063-1067, 2016.

[14] J. Zurawski. (2013, August 22). High Energy Physics and Nuclear Physics Network Requirements [Online]. Available: https://www.es.net/assets/Papers-and-Publications/HEP-NP-Net-Req-2013-Final-Report.pdf.