Achieving High Data Throughput in Research Networks[1]

Warren Matthews and Les Cottrell

(*Stanford Linear Accelerator Center, Stanford University, Menlo Park, CA 94025*)

Abstract: After less than a year of operation, the *BaBar* experiment at SLAC has collected almost 100 million particle collision events in a database approaching 165TB. Around 20 TB of data has been exported via the Internet to the *BaBar* regional center at IN2P3 in Lyon, France, and around 40TB of simulated data has been imported from the Lawrence Livermore National Laboratory (LLNL). *BaBar* collaborators plan to double data collection each year and export a third of the data to IN2P3. So within a few years the SLAC OC3 (155Mbps) connection will be fully utilized by file transfer to France alone. Upgrades to infrastructure is essential and detailed understanding of performance issues and the requirements for reliable high throughput transfers is critical. In this talk results from active and passive monitoring and direct measurements of throughput will be reviewed. Methods for achieving the ambitious requirements will be discussed.

Keywords: Network, Performance, Throughput, Tuning, Optimizing.

1 Introduction

The ambitious goals of particle physics research worldwide has created the need for a better understanding of the practical dynamics of network performance. It is understood that bandwidth is an essential ingredient but not a sufficient one. Methods for tuning and optimising the networks and the applications are also required. For example, out-of-the-box file transfer programs such as ftp are notorious for barely utilizing the network. Modified programs such as bbftp [1] which allow the user to modify some of the TCP/IP parameters must be used instead.

The experiments' collaborators have come to rely on the network to conduct research, and as backbone capacity increases so does their expectation. European physicists expect *BaBar* data to be available to them at the Tier-1 sites at IN2P3 in France and RAL in the UK as quickly and as seemlessly as data is available at SLAC to the US researchers. Projects such as the Particle Physics Data Grid (*PPDG*) will further raise expectations and future experiments such as those at the Large Hadron Collider (*LHC*) at CERN will challenge the networks even further.

In section 2 the historic network performance will be reviewed. The methods for improving the throughput will be detailed and measurements and observations will be discussed in sections 3 and 4. Furthermore, the complete end-to-end environment and the potential effect on perceived 'network' performance is explored in section 5. Some further effort will be detailed in section 6.

2 Past and Present Performance

Measurements from the *PingER* Network Monitoring Project [2] indicate the key components for good network performance, packet loss and round trip time (RTT), are excellent and improving on the high capacity backbone networks used by the High Energy and Nuclear Physics (HENP) community. In particular, the Energy Sciences Network (ESnet) and the Abilene Network (Internet2) are traversed when U.S. Labs and U.S. Universities communicate. Typically the packet loss between sites connected across these high capacity backbone networks is less than 0.2%.

ESnet and Internet2 have carefully engineered connections with research networks in Europe and Asia at locations in Chicago and New York. So transoceanic performance is also excellent. Recent packet loss between sites on ESnet and sites in Western Europe has typically been less than 0.3% and in the last few months more than half of the sites monitored have not suffered any packet loss.

Although these statistics show a vast improvement on performance from, say 5 years ago, the goal for data intensive science is for all site-to-site connections to maintain less than 0.1% packet loss.

If accurate estimates of packet loss rate and round trip time are available, throughput can be calculated using a simple equation [3]. However, in addition to low-impact active monitoring such as *PingER*, more abusive methods of monitoring are often utilized. The iperf tool is widely used to estimate the throughput that can be achieved on a link. Studies at SLAC have observed the throughput measured by iperf has improved over time.

3 Achieving High Data Throughput

The needs of high energy physics has resulted in these high capacity, well engineered networks. However the experience of using tools such as iperf and real user applications such as file transfer with ftp indicates that the typical TCP/IP stack available in most end-systems is incapable of utilizing the available resources. The default parameters are set expecting poor network performance, an assumption valid for many users but not for the networks traversed by HENP.

Extensive tests have been run by many researchers [4] to identify the potential of the network and determine the optimal TCP/IP parameters that the application should be tuned to. The optimum window size for a single stream is predicted using the bandwidth * RTT product. In particular studies at SLAC determined the optimal window size and number of parallel streams to use for file transfers with the bbftp program. Between SLAC and IN2P3, routed via CERN, the bottleneck bandwidth is estimated using the pipechar tool at between 80Mbps and 100Mbps. The RTT from ping measurements is about 175 ms. Thus the optimum window size is about 2.2 MBytes. This is above the maximum that Solaris by default allows. It can also be observed that an optimal throughput of between 70 and 80 Mbps appears to be achieved for a product of window size * number of streams of 2048 Bytes. As the product of window size * number of streams exceeds 2048, the throughput becomes very variable.

The highly dynamic nature of the network itself means the optimal window size is also highly dynamic. This makes it tremendously difficult to keep throughput at maximum. Projects such as web100 [5] are attempting to modify the TCP/IP stack to automatically tune the parameters to maintain the optimal setting without manual intervention. Typically multiple streams are not a network-friendly solution to throughput problems and web100 may help eliminate the practice. The project developers estimate any application is capable of achieving 90% of the available bandwidth with a modified TCP/IP stack.

4 Results

The results from testing with the iperf tool have been valuable in understanding both the optimal settings for the parameters for bbftp and the limitations of the TCP/IP stack in the operating systems.

A tuned transfer using bbftp from SLAC to IN2P3 usually maintains a total throughput of 30Mbps. So, even knowing the optimal parameters this is only around 30% of the bottleneck bandwidth. This short-fall is mostly because bbftp allows a maximum of 10 streams.

5 End-to-end Issues

The *PingER* project has been concerned with the wide-area network performance. However, it has also demonstrated that to achieve the goals of modern HENP the entire end-to-end environment must be considered. Efforts such as the Internet2 end-to-end initiative [6] are intended to gather experts in all areas to co-operate and enable the applications to fully utilize the network capabilities. SLAC has observed bottlenecks in machine cpu or io. It is also expected that soon factors such as AFS overhead will be amongst the limiting factors.

6 Future Work

At the time of writing, SLAC has just been approved for an upgrade to an OC12 (622Mbps) connection to ESnet. This means the route between SLAC and RAL will be 622Mbps from site-to-site. It is known the current link is performing well and therefore a significant increase in throughput is expected. Furthermore the connection between IN2P3 and CERN is scheduled for an upgrade to 622Mbps, it is expected the bottleneck between SLAC and IN2P3 will become STAR TAP.

Such high capacity links inevitably raise expectation and future work will not only track network performance at a finer grain but also consider the entire end-to-end nature of high performance computing. This is partcularly important for instrumenting data grids such as the Particle physics Data Grid (*PPDG*). New projects such as *AIME* will be provide the knowledge for accurate monitoring and instrumenting of the science grids.

Acknowledgments

The authors would like to thanks Doug Chang and Paola Grosso for their efforts in producing some of the results documented in this paper.

References

[1]   http://ccweb.in2p3.fr/bbftp/
[2]   http://www-iepm.slac.stanford.edu
[3]   Mathis, Semke, Mahdavi & Ott in Computer Communication Review, 27(3), July 1997.
[4]   See for example the talks in this conference from LBL and Caltech.
[5]   http://www.web100.org
[6]   http://www.internet2.org/e2eperf