Experimental Analysis of 10Gbps Transfers Over Physical and Emulated Dedicated Connections

Bradley W. Settlemyer, Nageswara S. V. Rao, Stephen W. Poole, Stephen W. Hodson, Susan E. Hicks, Paul M. Newman
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831
Email: {settlemyerbw,raons,spoole,hodsonsw,hicksse,newmanpm}@ornl.gov

Abstract—Long haul data transfers require the optimization and balancing of performances of host and storage systems as well as network transport. An assessment of such transport methods requires a systematic generation of throughput profiles from measurements collected over different system parameters and connection lengths. We describe host and storage systems to support wide-area I/O transfers at 10 Gbps, and present measurements of memory and disk transfer throughputs over suites of physical and emulated connections of several thousands of miles. The physical connections are limited by the infrastructure and incur significant costs. The emulated connections could be of arbitrary lengths at significantly lower costs but only approximate the physical connections. We present a differential regression method to estimate the differences between the performance profiles of physical and emulated connections, and then to estimate “physical” profiles from emulated measurements. We present a systematic analysis of wide-area memory and disk transfer throughput measurements, and establish that robust estimates of physical profiles can be generated using much less expensive emulated connections.

Index Terms—I/O throughput measurements, disk and host systems, performance analysis, differential and segmented regression, wide-area connections.

I. INTRODUCTION

There has been an increasing number of applications that produce large data sets, in the range of petabytes to exabytes, that are required to be transported over dedicated wide-area networks. For example, large simulation data sets produced by science applications on a supercomputer may be archived at a remote storage site over dedicated network connections. The needed high-performance data transfers require significant optimization and performance balancing of the host and storage systems as well as edge and long haul network connections. This task requires selecting optimal configurations and parameters of the devices and associated software to match the connection length; this is a particularly challenging task for connections of thousands of miles. Such optimization must be guided by a rigorous study of the throughput profiles using measurements over a wide range of system parameters and connection lengths. This task is complicated by the cost and complexity of the needed infrastructure of wide-area connections and powerful servers and disk systems.

Typically, production networks and testbeds are utilized to provide physical test connections, which are limited to lengths feasible within the infrastructure. Furthermore, such measurements over long haul, high-speed connections require an expensive, leading-edge infrastructure. Also, significant expertise and time are required to setup, monitor and collect the needed measurements. Network emulators, such as Anue 10GigE devices from Anue Systems, Inc. [1] offer a practical alternative: they emulate connections of arbitrary lengths at lower costs, without requiring the operational networking expertise. However, the emulated measurements only approximate those on physical connections, and their closeness to “real measurements” must be carefully assessed.

In this paper, our focus is on the relative performance of physical and emulated connections. We present memory and disk transfer throughput measurements collected over wide-area 10 Gbps connections provided by UltraScience Net (USN) [2] and emulated using Anue devices. We consider memory transfers using iperf and disk transfers using XDD: (i) iperf offers support for parallel streams and variable TCP window sizes, and (ii) XDD offers disk-aware I/O with configurable number of threads, and supports variable TCP window sizes [3]. Our overall approach, however, is more generic and is applicable to: (a) network tests involving simulated, emulated and physical connections; (b) network measurements using tools such as nuttcp; and (c) throughput measurements using file transfer tools such as GridFTP [4] and BBcp [5].

We present a method to estimate the differences between the profiles generated from measurements collected over physical and emulated connections, by specializing the differential regression method [6]. We show that this method can be used to: (a) generate robust approximations to physical profiles using measurements from emulated connections; and (b) derive estimates of throughputs at connection lengths not realizable on physical connections by using emulation. In particular, we demonstrate that emulated connections can be used to estimate throughputs at connection lengths not feasible over USN, with interpolation error below 4%. These results enable a quick but robust testing of data transfer methods using emulated connections, which typically takes few hours compared to a day or more over physical connections. Also, by anchoring the “emulation” profiles on the physical measurements, testing can be continued using emulation even after the physical infrastructure becomes unavailable subsequently.
The problem of efficiently moving data sets over wide-area networks has been studied by several research teams, using BBcp [5], GridFTP [4] and XDD [3]. A comparison of these three methods has been presented in [3]. Efficient use of both dedicated and shared 10 Gbps Ethernet network links has been studied by several research teams. Marian et al. [7], examined the congestion algorithm performance of TCP flows over high latency dedicated 10 Gbps connections, and found that congestion control algorithms such as HTCP and CUBIC provide high levels of performance even with multiple competing flows. Wu et al. [8], and Kumazoe et al. [9], studied the impacts of congestion control on shared 10 Gbps links. There have also been a number of efforts to systematically generate the performance profiles of wide-area data transfers for InfiniBand [10] and TCP over plain and encrypted connections [11]. Physical connections used in such studies incur significant cost, and our results illustrate the cost-benefit trade-offs of utilizing less expensive emulations.

This paper is organized as follows. In Section II, we describe our infrastructure for wide-area memory and disk transfers. We then discuss memory and disk transfer throughput measurements in Section III. We describe the differential regression method in Section IV. We describe the measurements and analysis results based on the differential regression method in Section V.

II. WIDE-AREA I/O CONFIGURATION

The components of an end-to-end disk transfer configuration are: storage devices, data transfer hosts, local area and wide-area connections as illustrated in Figure 1. High performance data transfers over a wide-area connection require optimizing the performance of all these components and their compositions to avoid “impedance mismatches”. Large data transfers typically require dedicated data transfer hosts with high performance Network Interface Cards (NICs) to access the network connections and Host Bus Adapters (HBA) to access the storage systems. The interconnection network provides high-speed, circuit-switched connection that allows dedicated connections between the sites. In addition, the software components include I/O modules for disk access and TCP/IP stack for network transport. The data transfers are handled by the application software running on the hosts which accesses the storage systems and long haul network connections.

1) Storage Infrastructure: Our storage system consists of six Infortrend EonStor S16F-R1430 disk arrays, each equipped with a controller and 8 Hitachi DeskStar E7K500 disks. The controllers are configured to provide RAID level 5 in a 7+1 configuration with 64 Kilo Byte (KB) stripe size. Each controller is connected to Storage Area Network (SAN) with a single 4 Gbps Fibre Channel connection. The storage network fabric is a Brocade Silkworm 4100 switch.

2) Host Infrastructure: Our host systems are commodity serves with dual Quad-Core AMS Opteron processors, running Fedora 13 Linux OS. They have 32 Giga Bytes (GB) of main memory, a Myricom Myri-10G Dual-Protocol NICs, and two dual port QLogic ISP2432-based 4 Gbps Fibre Channel HBAs. The 10 Gbps NICs are connected to the long haul connections, and each host has three Fibre Channel connections into the Fibre Channel switch. The host systems run Linux kernel version 2.6.33.3-85 with CUBIC TCP congestion control. To aggregate the three storage units into a single file system per host we constructed a single volume group striped across each physical volume with a 64 KB stripe size. We then constructed a local XFS file system on each host using the default file system parameters with the exception that we used 4 KB block sizes.

![Figure 1. Components of end-to-end data transfer.](image1)

![Figure 2. USN consists of dual 10 Gbps lambdas from Oak Ridge to Chicago to Seattle to Sunnyvale.](image2)

<table>
<thead>
<tr>
<th>connection length (miles)</th>
<th>0.2</th>
<th>1,400</th>
<th>6,600</th>
<th>8,600</th>
</tr>
</thead>
<tbody>
<tr>
<td>rtt (ms)</td>
<td>0.29</td>
<td>26.9</td>
<td>134</td>
<td>171</td>
</tr>
</tbody>
</table>

(a) USN physical connections

<table>
<thead>
<tr>
<th>Anue rtt (ms)</th>
<th>100</th>
<th>150</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>estimated conn. length (miles)</td>
<td>5,029</td>
<td>7,544</td>
<td>10,038</td>
</tr>
</tbody>
</table>

(b) Anue emulated connections

Table I

**Connection lengths and RTTs of USN and Anue connections.**
device accesses is achieved by XDD through direct I/O to storage devices, file pre-allocation support, configuration of I/O buffer sizes and TCP window sizes. For a file transfer, XDD spawns a single Target Thread that opens the file, and then spawns a user-specified number of QThreads. If the XDD is invoked as a file transfer destination, the QThreads will each bind to a network port and wait for a connection from the source-side process. After opening the file in read-only mode, the source XDD spawns the specified number of QThreads, each of which connects to one of the listening threads.

4) Network Connections: USN is a wide-area network testbed that provides suites of OC192 and 10GigE WAN/LAN-PHY connections of several thousands of miles [2]. USN consists of two parallel 10 Gbps connections with co-location sites at Oak Ridge, Chicago, Seattle and Sunnyvale as shown in Figure 2. We utilize 0.2, 1,400, 6,600 and 8,600 mile wide-area USN connections. The hosts are connected to the edge Fujitsu XG2000 10GigE switch which in turn is connected USN. In addition, Anue device is connected to Fujitsu switch to support connection emulations. Virtual Local Area Networks (VLAN) on Fujitsu are used to separate the traffic of the hosts and also to switch them between USN and Anue connections. Anue device is utilized to emulate 10GigE connections by emulating a specified delay. The emulated connections are specified by their latencies. We utilized round trip times (rtt) of USN connections shown in Table 1 to set up the corresponding emulated connections. In addition, we emulated connections with 100, 150 and 200 ms rtt, which are not feasible over USN infrastructure; the first two values are smaller than largest possible rtt on USN and the third one is larger, corresponding to 10,058 miles.

III. THROUGHPUT MEASUREMENTS

We first consider memory transfers between the hosts to establish the baseline for the best possible disk transfer throughputs, and then consider disk-to-disk transfers. In keeping with the convention, network and disk throughputs are expressed as bits per second (bps) and Bytes per second (Bps), respectively.

1) Memory Transfer Throughput: Throughputs of memory-to-memory transfers $T_M^P(x, n, w)$ are measured using iperf by varying the number of parallel streams, $n = 1, 6, 8, 12, 16, 24, 32, 48, 64, 96$, and also varying TCP window size $w = 1, 2, 4, 5, 8, 10, 16, 128$ MB, over a connection of length $x$ and type $A = \mathcal{E}, \mathcal{P}$, where $\mathcal{P}$ and $\mathcal{E}$ correspond to physical and emulated connections, respectively. To simplify the discussion on throughput profiles, we suppress the operands $x, n$, and $w$ when they are assigned fixed values and are clear from the context. We utilize $x = 0.2, 1, 400, 6, 600, 8, 600$ miles for USN connections and the corresponding emulated connections. The throughput measurements for USN and emulated connections are shown in Figures 3 and 4, respectively. In each figure: (i) top plot contains $T_M^A(x, w); n = 1$ and $T_M^A(x, w); n = 96$, which show the worst and best window profiles, respectively, and (ii) bottom plot contains $T_M^A(x, n); w = 1$ MB and $T_M^A(x, n); w = 16$ MB, which show the worst and best parallel stream profiles, respectively. In both cases, achieving

---

**Figure 3.** Physical Connections: Top and bottom plots, $T_M^P(x, w)$ and $T_M^P(x, n)$, show the effects of buffer size and parallel streams, respectively.

3) Data Movement Software: XDD software enables the movement of large data sets with the capability to drive disk arrays at their maximum bandwidth. Our enhanced version of XDD [3] provides options for configuring device access schemes, numbers of threads, and I/O scheduling policies. Data transfer is accomplished by executing matched pairs of XDD instances at source and destination hosts. Efficient I/O
high throughputs at connection lengths of thousands of miles requires utilizing large window sizes \( w = 10, 16 \) MB, and utilizing \( n = 25 \) or more parallel streams. Both \( w \) and \( n \) are free parameters in that they can be chosen at the hosts, whereas the connection length \( x \) is determined by the locations of data transfer sites. For fixed values of \( w \) and \( n \), \( T_M^x(x) \) is a monotonically decreasing function of \( x \), which shows that values of \( w \) and \( n \) chosen for longer connections would suffice for shorter ones as well. Furthermore, this monotonicity is a critical property for establishing the effectiveness of the differential regression method in the next section.

5) Disk Transfer Throughput: Throughputs of disk-to-disk transfers \( T_D^A(x, t) \) are measured using XDD while varying the number of I/O threads, \( t = 1, 4, 8, 16, 32 \) for transfers between single hosts, and \( t = 1, 4, 8, 16, 32, 48, 64 \) for transfers between pairs of hosts over a connection of length \( x \) and type \( A \). Figure 5 shows the end-to-end file transfer throughputs, denoted by \( T_D^P(x, t) \) and \( T_D^E(x, t) \), respectively, for physical and emulated connections, using a 200 GB file copied between single hosts using XDD with 32 MB request size. The wide-area USN connection has the capacity of 9.6 Gbps, which corresponds to 1200 MBps of data throughput. Our disk arrays are capable of throughput below 1000 MBps as shown in the measurements, and hence we conducted additional tests by utilizing pairs of hosts at each end of the connection. The end-to-end file transfer throughputs using physical and emulated connections, are denoted by \( T_D^{P}(x, t) \) and \( T_D^{E}(x, t) \), respectively, using pairs of hosts at each connection endpoint to transfer a 200 GB file. We utilized the multi-host and multi-NIC capabilities of XDD for these measurements, and the connection capacity was saturated by the addition of two flows at endpoints. The number of threads \( t \) is a free parameter to be chosen for the hosts, whereas the connection length \( x \) is specified by the locations of transfer sites. While \( T_D^{A}(x, t) \) is a monotone function of \( t \), \( T_D^{B}(x, t) \) is not particularly for USN, and hence a suitable \( t \) value must be chosen around \( t = 32 \). However, in both cases for a fixed value of \( t \), throughput is a monotone function of \( x \), and again this property will be utilized in the statistical estimation of the differential regression.

IV. DIFFERENTIAL REGRESSION

The concept of differential regression [6] is to capture the differences between the measurements collected under two different modalities, namely physical or emulated connections. Consider the measurements of previous section represented generically by \( T^D_i(x) \), \( A = P, E \), \( B = M, D, DD \) as a function of connection length \( X \) with all other parameter being fixed. Let \( Y \) represent the measurement corresponding to throughput \( T^D_i(x) \). Consider a function \( f : \mathbb{R}^m \to \mathbb{R}^d \) such that \( f(X_i) \) is used as an estimate of throughput measurement \( Y_i \) over a connection of length \( X_i \) of modality \( M_i \). We are given two data sets of two modalities \( M_i \) and \( M_j \),

\[
D_i : \quad (X_{i,1}, Y_{i,1}), (X_{i,2}, Y_{i,2}), \ldots, (X_{i,n_i}, Y_{i,n_i})
\]

\[
D_j : \quad (X_{j,1}, Y_{j,1}), (X_{j,2}, Y_{j,2}), \ldots, (X_{j,n_j}, Y_{j,n_j})
\]

Figure 4. Emulated Connections: Top and bottom plots, \( T_M^x(x, w) \) and \( T_M^x(x, n) \), show the effects of buffer size and parallel streams, respectively.

which are distributed according to \( P_{X_i,Y_i} \) and \( P_{X_j,Y_j} \), respectively. The expected error \( I_{M_i}(f) \) and empirical error \( \hat{I}_{M_i}(f) \) based on the data set \( D_i \), of \( f \) as defined follows, respectively:

\[
I_{M_i}(f) = \int (f(X_i) - Y_i)^2 dP_{X_i,Y_i}
\]

\[
\hat{I}_{M_i}(f) = \frac{1}{n_i} \sum_{i=1}^{n_i} (f(X_i) - Y_i)^2
\]
We define the differential regression by the triple \( f_{\Delta_{i,j}} = \left( f_{D_i}, f_{D_j}, f_{D_j} - f_{D_i} \right) \), such that the regressions \( f_{D_i} \) and \( f_{D_j} \) fit the individual data sets \( D_i \) and \( D_j \), respectively, and the \( f_{D_j} - f_{D_i} \) fits their differences. The expected error of \( f_{\Delta_{i,j}} \) is defined as [6]

\[
\hat{I}_{M}(f) = \frac{1}{n_i} \sum_{k=1}^{n_i} \left( f(X_{i,k}) - Y_{i,k} \right)^2.
\]

The best expected differential regression \( f_{\Delta_{i,j}}^* \) minimizes the expected error

\[
C \left( f_{\Delta_{i,j}} \right) = 2I_{M}(f_{D_i}) + 2I_{M}(f_{D_j}).
\]

Detailed knowledge of the underlying distributions is needed to compute \( f_{\Delta_{i,j}}^* \), which are very complex for memory and disk transfer throughputs since they depend on a number of disparate components. Instead, we have access to the measurements, \( D_i \) and \( D_j \), which enable us to compute an approximation \( \hat{f}_{\Delta_{i,j}} \) to \( f_{\Delta_{i,j}}^* \) with performance guarantees [12]:

\[
P \left[ C \left( \hat{f}_{\Delta_{i,j}} \right) - C \left( f_{\Delta_{i,j}}^* \right) > \epsilon \right] < \delta
\]

where \( \epsilon > 0 \) and \( 0 < \delta < 1 \), and \( \hat{f}_{D_i} \) and \( \hat{f}_{D_j} \) are chosen from a class of monotone functions. This guarantees that “error” of \( \hat{f}_{\Delta_{i,j}} \), which is computed based entirely on \( D_i \) and \( D_j \), is within \( \epsilon \) of optimal error (of \( f_{\Delta_{i,j}}^* \)) with an arbitrary high probability \( 1 - \delta \), irrespective of the underlying distributions.

Let \( \hat{f}_B^A(.) \) be the regression estimate of \( T_B^A(.) \), \( A = \mathcal{P}, \mathcal{E}, \mathcal{B} = M, D, DD \) computed based entirely on measurements. These estimated regressions can be utilized as follows:

(a) **Throughput Interpolation/Extrapolation:** We utilize \( \hat{f}_B^P(x) \) to estimate the throughput of a physical connection of length \( x \), which may not be feasible on the physical infrastructure. If \( x \) is far from the test connection lengths, the quality of \( \hat{f}_B^P(x) \) could be low, which can be improved using emulation as described in (b).

(b) **Physical Throughputs from Emulations:** Consider that we computed \( \hat{f}_B^P \) by collecting physical measurements, and \( \hat{f}_B^{E \otimes P} = \hat{f}_B^P - \hat{f}_B^E \) by emulating several connection lengths. Then we estimate the throughput of physical connection of length \( x_P \) by: (a) measuring throughput \( Y_{E,B} \) of emulated connection of length \( x_P \), and, (b) computing the estimate of physical throughput as

\[
\hat{Y}_{P,B} = Y_{E,B} + \hat{f}_B^{E \otimes P}(x_P),
\]

where the second “correction” term is the differential regression that accounts for the difference between emulated and physical connections. This estimate is derived by emulation of connection length \( x_P \) and the precomputed differential regression, and it does not require physical measurements. Such estimation can be carried out even if the physical infrastructure is no longer available.

### V. Measurements and Analysis

We collected memory and disk transfer throughput measurements on both physical and emulated connections of lengths \( \mathcal{L}_{PE} = \{0.2, 1400, 6600, 8600\} \) miles corresponding to rtt\(\mathcal{R}_{PE} = \{0.2, 26.9, 134, 171\} \) milliseconds. In addition, we collected measurements on emulated connections with rtt\(\mathcal{R}_E = \{100, 150, 200\} \) milliseconds, which correspond to connection lengths shown in Table 1.

1) **Differential Regression:** For lengths \( \mathcal{L}_{PE} \) feasible both on physical and emulated connections, we compute the differential regressions \( \hat{f}_B^E, \hat{f}_B^P, \hat{f}_B^{E \otimes P} = \hat{f}_B^P - \hat{f}_B^E, = -\hat{f}_B^{P \otimes E} \) for \( B = M, D, DD \). Specific values are shown in Table II for parameter values that yielded high throughputs over all connection lengths \( \mathcal{L}_{PE} \); we use \( w = 16 \) MB and \( n = 48 \) for memory transfers, and \( t_D = 8 \) and \( t_{DD} = 64 \), for single host and double host disk transfers, respectively. As indicated in top row of Table II the estimated throughputs are consistently higher at all connection lengths. For disk transfers, differential regression terms are both positive and negative.

2) **Interpolation and Extrapolation:** We use the differential regression function to estimate throughput for 6600 mile (134 ms) connection by interpolating the measurements from 1400 and 8600 mile connections for both physical and emulated connections. For parameters \( n = 48, w = 16, t_D = 8 \) and \( t_{DD} = 64 \), we compare these interpolated values with measurements as shown in Table III. In addition to physical and emulated connections, we also estimated the physical connection throughput using the emulated measurements and differential regression; these values are shown in rows marked as anue-usn in Table III. Overall, the estimation error for physical connections is within 4% of the measured throughput.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( 0.2 )</th>
<th>( 1400 )</th>
<th>( 6600 )</th>
<th>( 8600 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{f}_B^E(x)/T_B^E(x) )</td>
<td>5.14%</td>
<td>5.80%</td>
<td>10.99%</td>
<td>14.98%</td>
</tr>
<tr>
<td>( T_B^E(x) ) - Gbps</td>
<td>9.73</td>
<td>9.65</td>
<td>8.83</td>
<td>8.81</td>
</tr>
<tr>
<td>( \hat{f}_B^P(x)/T_B^P(x) )</td>
<td>28.03%</td>
<td>-2.82%</td>
<td>-2.26%</td>
<td>-0.91%</td>
</tr>
<tr>
<td>( T_B^P(x) ) - MB/s</td>
<td>829.47</td>
<td>644.59</td>
<td>670.57</td>
<td>640.98</td>
</tr>
<tr>
<td>( \hat{f}<em>B^{E \otimes P}(x)/T</em>{BD}(x) )</td>
<td>7.37%</td>
<td>-2.19%</td>
<td>0.56%</td>
<td>-0.99%</td>
</tr>
<tr>
<td>( T_{BD}(x) ) - MB/s</td>
<td>1233.98</td>
<td>899.91</td>
<td>715.89</td>
<td>684.06</td>
</tr>
</tbody>
</table>

Table II

**DIFFERENTIAL REGRESSIONS FOR**

**\( n = 48, w = 16, t_D = 8 \) AND \( t_{DD} = 64 \).**
both for memory and disk transfers, and is below 6% for emulated connections.

For emulated connections we utilize measurements from \( \mathcal{P} \) to estimate throughputs for connections corresponding to \( \mathcal{E} \) and compare them to measurements in Table IV. Interpolation is used for throughputs of 100 ms and 150 ms connections, extrapolation is used for 200 ms connection. While the errors for memory transfers are below 1% and for double host memory transfers are below 6%, the errors for single disk transfers are relatively higher, but below 10%.

3) Estimation of Physical Profiles: We now utilize the measurements from emulated connections corresponding to \( \mathcal{E} \) to estimate the corresponding physical profiles by applying the differential regression computed based on measurements at \( \mathcal{P} \). The actual errors of these estimates cannot be computed since they are not realizable on USN. However, these profiles are qualitatively quite similar to the physical profiles. And combined with interpolation error less than 4% and analytical performance guarantees described in the previous section, they represent robust approximations to physical profiles.

**VI. CONCLUSIONS**

An assessment of wide-area data transfer solutions requires a systematic generation of throughput profiles from measurements collected over different system parameters and connection lengths. We presented measurements of memory and disk transfer throughputs over suites of physical and emulated connections of several thousands of miles. We presented a differential regression method to estimate the differences between the performance profiles of these two connection types, and then to estimate “physical” profiles from emulated measurements. We illustrated that robust physical throughput estimates can be generated using much less expensive emulation infrastructure, particularly for connection lengths not feasible over physical infrastructure.

It would be of future interest to test the proposed methods for data transfer at 40 and 100 Gbps. It would be of interest to study extensions of the differential regression method to include performance profiles that do not satisfy the monotonicity conditions and to include network simulations.

**VII. ACKNOWLEDGMENTS**

This work is carried out at the Extreme Scale Systems Center, sponsored by U. S. Department of Defense, and is also funded by the Mathematics of Complex, Distributed, Interconnected Systems Program of U. S. Department of Energy, at Oak Ridge National Laboratory managed by UT-Battelle, LLC under contract No. DE-AC05-00OR22725.

**REFERENCES**