# White Paper: Lustre WAN over 100Gbps

*Abhinav Thota*[1]
*Robert Henschel*[1]
*Stephen Simms*[1]

[1]Indiana University Pervasive Technology Institute

February 4, 2016

Thota, A., Henschel, R., & Simms, S. "White Paper: Lustre WAN over 100Gbps" Indiana University, Bloomington, IN. PTI Technical Report PTI-TRYY-002 February 4, 2016.

INDIANA UNIVERSITY
PERVASIVE TECHNOLOGY INSTITUTE

**Table of Contents**

## 1. Introduction

This work is an international collaboration with Rheinisch-Westfälische Technische Hochschule Aachen, Germany (RWTH) and the Center of Information Services and High Performance Computing (ZIH) at Technische Universität Dresden, Germany to analyze the effect of a high-bandwidth high-latency link on the I/O patterns of scientific applications using the 100Gbps transatlantic link.

With the past testing we have done using the 100Gbps link and on local and domestic end points, we have gained insights into how best to exploit the Lustre file system over long distance, simplifying data access for a broad range of applications.

We used a Lustre file system over a wide area network (WAN) for our testing. Lustre is a distributed parallel file system and is highly scalable. More than fifty percent of the top 100 supercomputers use the Lustre file system. DC-WAN is one of the Lustre file systems at Indiana University and it has been mounted at many locations across world at various points in time, including NCSA, Pittsburgh, Mississippi State, Tucson, TACC, Dresden, and Aachen. DC-WAN is a Lustre file system optimized for mounting across long distances and this lets researchers access remote data as if that data were stored locally. This lets multiple groups, potentially from multiple organizations dispersed geographically, to work on the same data.

The transatlantic performance of DC-WAN is something we will discuss in more detail, but it will be shown that it is comparable or better than other options, especially using a dedicated 100Gbps link. The convenience of accessing the files on a central file system and having a single copy of the data instead of moving data back and forth with secure copy (SCP) or Globus is something that we will look into.

## 2. Systematic testing

We picked BLASTn, a popular application among genomics and bioinformatics researchers that uses really large but common input data sets, to do initial testing. We started off by running BLASTn on non-dedicated links to our end points in Dresden and Aachen in Germany and on Blacklight in Pittsburgh. We ran BLASTn on compute resources at all three locations mounting DC-WAN, reading data from and writing data to DC-WAN. We compared the performance to BLASTn when using a local Lustre file system and also when the data was transferred to the compute location using SCP.  This is shown in Fig 1, across the Atlantic on a non-dedicated link, DC-WAN is less than 20% slower than a local file system. It is comparable to SCP, but is more preferable given the convenience. Fluctuations in available bandwidth and latency are potential complications when using a non-dedicated link.

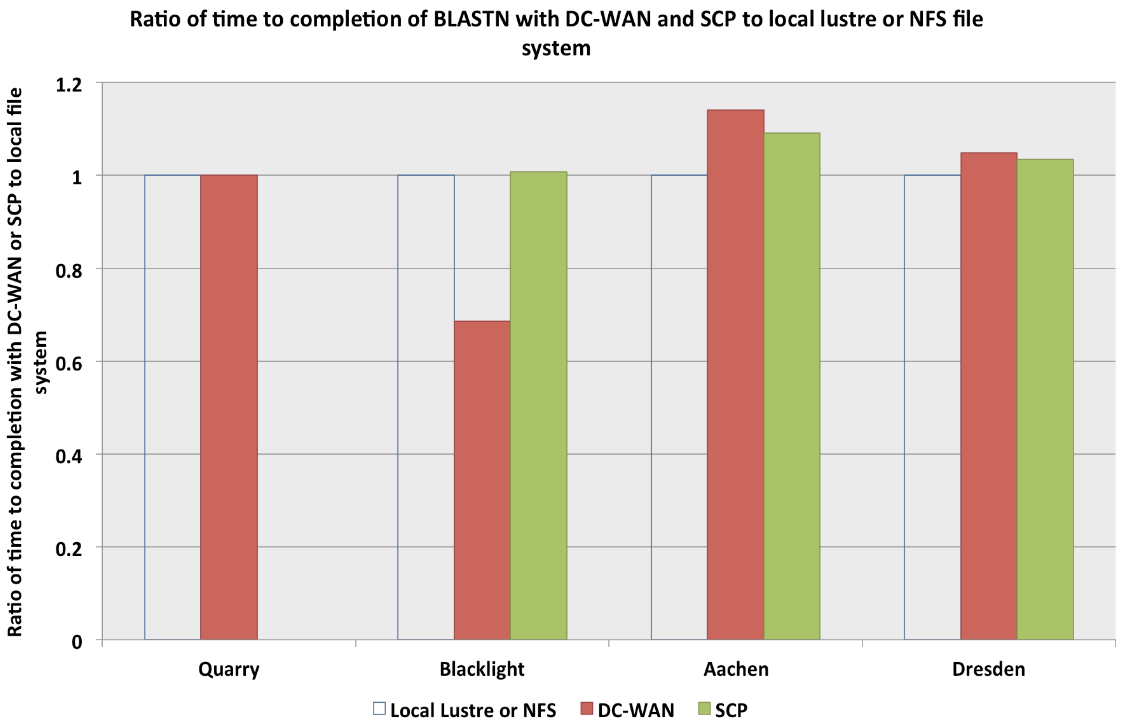

Figure 1: Across the Atlantic on a non-dedicated link, DC-WAN is less than 20% slower than a local file system. It is comparable to SCP, but is more preferable given the convenience. It can be noted that DC-WAN is faster on Blacklight, but that could be explained by faster CPUs on Blacklight.

We were able to repeat some of these tests on the dedicated 100Gbps transatlantic link. The results are shown in Fig 2 and this chart shows us clear advantages of a dedicated link. The runs on the dedicated link were 35% faster. Unfortunately, both our partners in Germany had to return their borrowed network equipment and we lost access to the Lustre endpoints at this point and we were not able to do further testing.

Given that we are not using the 100Gbps link, we decided to continue these tests locally at Indiana University with two different Lustre file systems. We already have an existing production Lustre file system and in addition to that we installed a separate Lustre file system, just for these tests. This way, we will have a system that is not affected by other users and fluctuations due to changes in usage. Given that BLASTn is a complex application and understanding and predicting it's behavior is not trivial, we worked on developing an application that reads and writes a user specified amount of data in specific intervals.

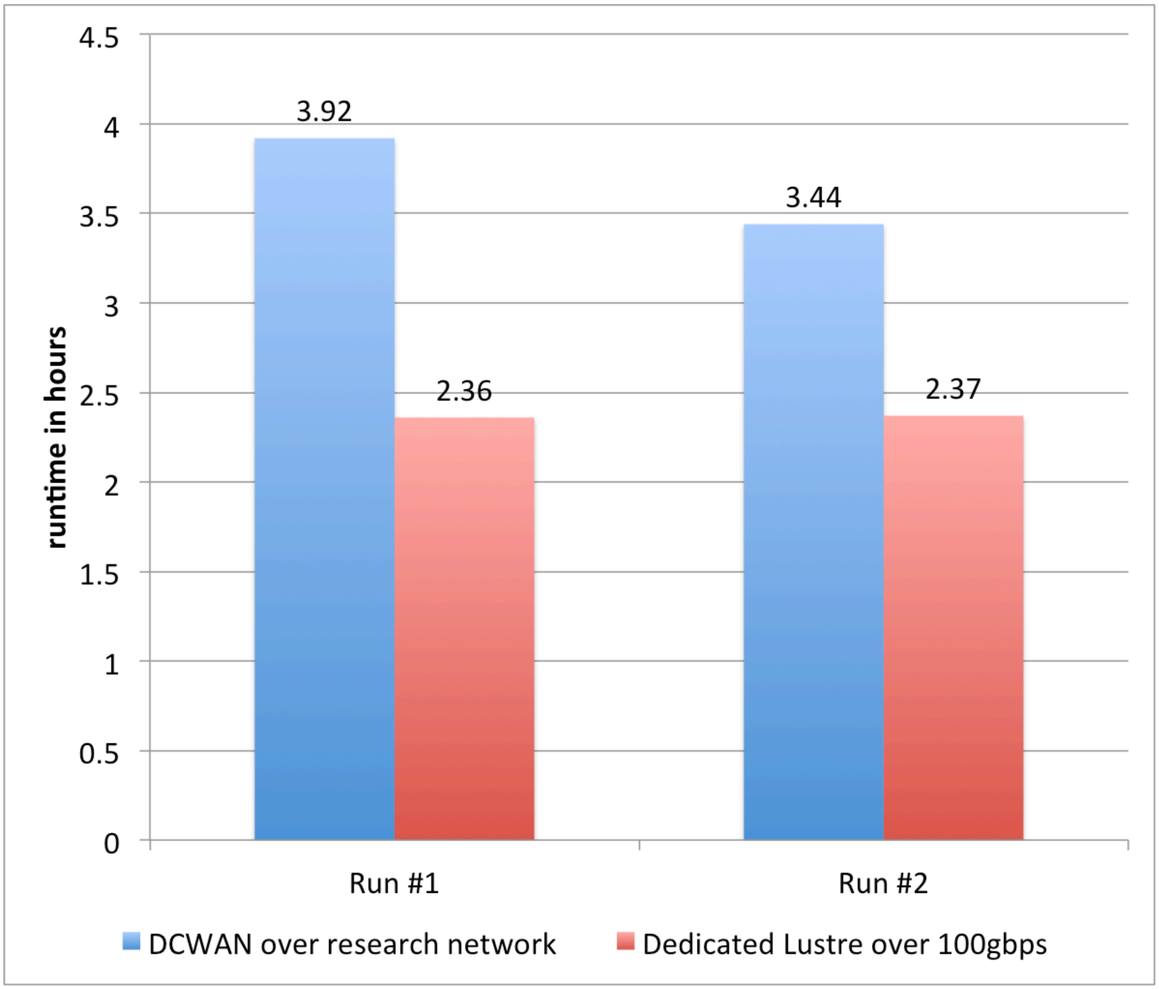

Figure 2: The blue bars show the runtimes over the regular research network using the DC-WAN production file system and the red bars show the runtimes over the 100gbps link using a dedicated lustre file system. The runs over the 100gbps link are 35% faster.

Our tests helped us in understanding the I/O characteristics of BLASTn on a Lustre file system. With the test application that we developed, we are able to say that the Lustre file system intuitively reads ahead, generally trying to be ahead of the application's read requests. This design is definitely beneficial for BLASTn and other applications that have similar I/O characteristics. BLASTn reads input sequentially.

We ran the test application with reads and writes from 1K to 100MB per second using the application we developed. We then looked at the Lustre activity on the client and the server to understand how Lustre responds to I/O requests, both big and small. We used a monitoring tool called Collectl for this purpose. We were able to verify that Lustre always reads ahead, which is really beneficial for applications that do sequential reads. This behavior can be seen in figures 3 and 4, here we ran the test application on Lustre and configured it to read 1MB from a file sequentially every 30 seconds. The reads in figure 3 do not match up with the data transmitted over the network during the same time. It can be seen that Lustre reads ahead for the first half of the run, reaches the end of the file and does not do anymore reads after that.

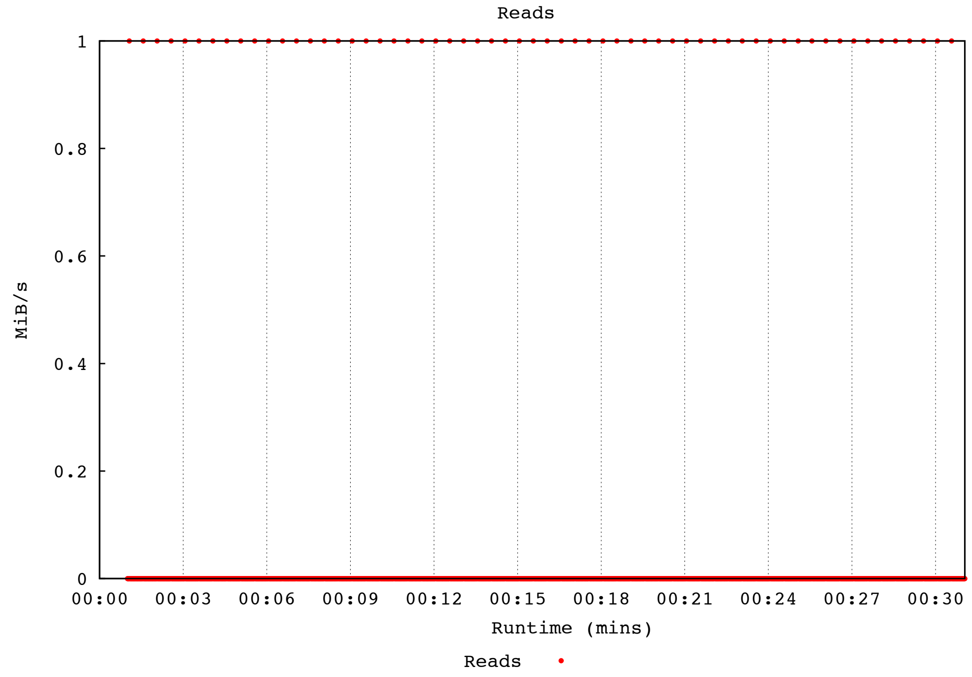

Figure 3: This chart shows the reads done by our custom application that reads 1MB sequentially every 30 seconds from a 60MB file.

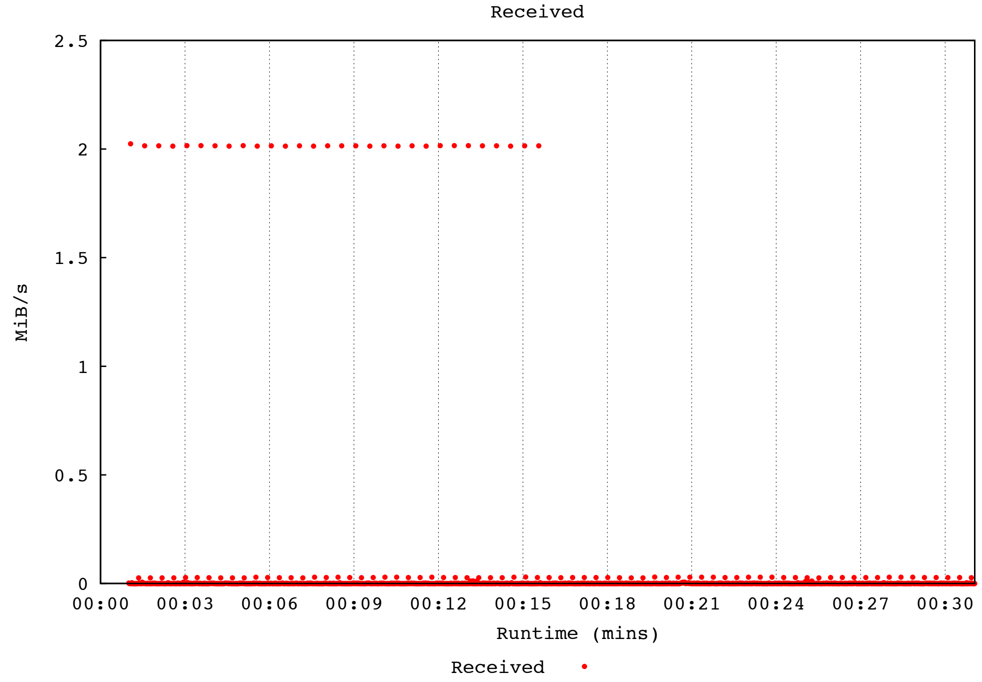

Figure 4: This chart shows the data received over the network during the custom application from Fig 3 was running. Lustre is reading ahead for the first half of the run, until end of file.

Finally, we came up with a series of formulations from our experiments comparing SCP and Lustre performance over long distance, shown in the table below.

|  | SCP v Lustre Advantage | Reason |
|---|---|---|
| Slower network | Lustre | Caching |
| Faster CPUs | Lustre | Smaller run time and a higher SCP overhead |
| Slower network and faster CPUs | Lustre | Higher SCP overhead |
| Faster network and faster CPUs | SCP | Low network overhead |
| Slower network and slower CPUs | Lustre | Run time increases and SCP overhead also increases |
| Faster network and slower CPUs | SCP | Runtime increases and low SCP overhead |