Intel 10GBASE-T in TACC Dynamic Hadoop Environment

Benefits of 10 GbE in rapidly changing Hadoop environments

Testing done by the Texas Advanced Computing Center (TACC) at The University of Texas at Austin with Intel Ethernet X540 Converged Networking Adapters highlights the benefits of Intel X540 10 GBASE-T Converged Networking Adapters in their project-driven dynamic Hadoop environment.

Overview
As we continue to generate substantial volumes of data at an accelerated pace, companies have been looking for ways to extract additional value from the large amounts of data they are collecting. Gartner estimates that organizations will spend $28 billion in 2012 and $34 billion in 2013 in information technology to handle big data. i

Hadoop® has become one of the fundamental tools used to process and manage the unstructured and structured data being generated. Hadoop provides a scalable, open-source platform to process data in a distributed manner. Designed to work using commodity hardware, Hadoop implementations typically use 1 GbE (Gigabit Ethernet) interconnects.

Due to advances in 10GBASE-T technology, faster interconnects are now more affordable. This gives organizations the opportunity of deploying a cost-effective, high performance Hadoop cluster based on 10 GbE interconnects. As part of this paper, we have partnered with the Texas Advanced Computing Center (TACC) to examine how their unique implementation of Hadoop can benefit from Intel’s cutting-edge 10GBASE-T CNA: the Intel Ethernet Converged Networking Adapter X540.

Hadoop @ TACC: Enabling Research with Hadoop
The Data Mining & Statistics Group (DMS) at TACC focuses on helping researchers to meet their data analysis needs and facilitating the data-driven research in diverse scientific domains through collaboration and consulting support. To better assist with the increasing demands of “Big data”, the DMS group initiated the “Hadoop on Longhorn” project in 2010. The project enables users to create and construct dynamic Apache™ Hadoop clusters on demand.

In recent years, there has been an increasing interest in running Hadoop clusters and analysis programs in the “cloud”. This implies starting a Hadoop cluster with a remote shared infrastructure to conduct data analysis tasks on demand. The most common example of cloud computing is Amazon’s EC2 web service that allows developers to run virtual Hadoop clusters, paying only for the computation they use. In this model, a user first requests a set of computing nodes from a remote system. Then an instance of a Hadoop cluster is started directly or through loading prebuilt virtual machine images to allow the user to carry out their data analysis tasks.

There are several advantages to running Hadoop in the cloud. First, users do not need to maintain a physical cluster, instead only paying for their computing time that is commonly calculated by the number of CPU hours used. Consequently, the operational cost to the user is very low. Secondly, the user can easily increase or decrease the size of the cluster based on computing needs and the capacity of the remote infrastructure. Thirdly, the centralized infrastructure can consist of the high-end hardware such as high bandwidth inter-connections, powerful CPUs and a large amount of memory, all of which are prohibitively expensive to typical users. Lastly, a centralized infrastructure can be reused and shared by many users to maximize the hardware utilizations and facilitate collaboration while using a similar development environment.

About TACC
The Texas Advanced Computing Center (TACC) at The University of Texas at Austin is one of 11 centers across the country providing leading computing resources to the national research community through the National Science Foundation XSEDE project. Its mission is to enable discoveries that advance science and society through the application of advanced computing technologies. With more than 110 staff and students, TACC operates several of the most powerful supercomputers and visualization systems in the world, and the network and data storage infrastructure to support them.ii
A challenge to apply this model for running a Hadoop cluster is the data transfer cost in and out of the dynamically constructed cluster. For an on-demand Hadoop cluster, data to be processed must first be imported into the cluster. Because processing large amounts of data is central to the typical Hadoop program, the cost of importing those data into a cluster is extremely relevant to overall performance. “Therefore, high bandwidth connections are especially appreciated in such dynamic Hadoop environments,” says Dr. Wei Jia Xu, manager of Data Mining & Statistics group at TACC.

**Building Dynamic Hadoop Clusters**

One of the objectives of this group is to build and support an infrastructure to create dynamic Hadoop clusters for researchers’ use. When a researcher submits a Hadoop job, nodes within TACC’s Longhorn cluster are allocated to a custom Hadoop cluster with the appropriate number of DataNodes based on the specific job requirements. Once the job completes, the nodes can be allocated to a different Hadoop cluster if needed.

The group also focuses on analyzing additional problems encountered when creating dynamic clusters. When a new Hadoop cluster is being setup, the HDFS data previously stored in the DataNodes’ drives is erased and HDFS is reconfigured. Since new Hadoop clusters are setup regularly, data is loaded in and out of HDFS more frequently than in long-term clusters. This is one of the most time-consuming steps within the cluster setup process. As part of this paper, we examined how 10 GbE would reduce the time to load data into HDFS.

**Writing, Running, and Tuning Hadoop Jobs**

Another focus for this group is to collaborate with researchers to parallelize code using Hadoop. This includes working with researchers to write new jobs, help them run existing code, and tune existing code to run faster using TACC’s infrastructure. In this paper, we compared the performance of 10 GbE and 1 GbE when running an existing Hadoop job.

**Hadoop and 10 GbE**

Due to technical advances in design and manufacturing, today’s servers can process and access data faster than ever before. For example, the Intel Xeon® E5 series processor can handle 671 tasks in parallel. The high computational capacity means you can process more data, drive higher network throughput, and take full advantage of a 10 GbE network. At the same time, Hadoop workloads can generate large volumes of data that needs to be transferred across nodes, especially during the shuffle, sort, and reduce phases.

Intel Ethernet 10 GbE Server Adapters provide a robust, forward-looking basis for network connectivity in a Hadoop implementation by:

- Supporting data-intensive applications with controllers that are designed in conjunction with the latest Intel server platforms for highly optimized performance

- Counting on mature, stable drivers for Intel Ethernet that help ensure trouble-free, dependable operation on Linux and Windows operating systems.

One of the main reasons organizations are not using 10 GbE is because of the costs associated with this technology. 10GBASE-T makes it possible to significantly reduce the cost of a 10 GbE deployment. Intel Ethernet 10GBASE-T Controller (X540) brings the following additional benefits:

- Reduced deployment costs to bring 10GbE to the broad market through integrated 10GbE, lower cable costs, and flexible cable lengths up to 100m over Cat 6a.

- Simplified transition to 10GbE from backward compatibility with the switching and cabling infrastructures of existing 1 GbE networks.

- Reduced complexity through industry-standard twisted-pair copper cabling and support for advanced I/O virtualization and unified networking.

**Tests and Results**

In order to examine how 10 GbE can improve the performance of TACC’s Hadoop usage models, we ran tests representing TACC’s dual focus on Hadoop. This was achieved by running Hadoop benchmarks and using 10 GbE on a genome resequencing research problem. Note that all tests were run by TACC staff using Hadoop clusters with four, eight, or twelve nodes.

**Loading Data into Dynamic Hadoop Clusters**

TACC identified loading data into HDFS when a new cluster is configured as a pain point they looked to relieve using 10 GbE. We identified two tests that would help us examine how 10 GbE helps with this issue.

The first test we selected is TestDFSIO. This is a popular benchmark for HDFS performance and is part of the Apache Hadoop distribution. It measures throughput, average I/O rate and run time when writing and reading to HDFS.

The second test selected was developed internally by TACC to measure the time taken to copy files to each DataNode.

**Test DFSIO**

For this test, 10 to 500 files of size 1000 MB or 10,000 MB were written and read in HDFS. The replication factor used was 3. The figure below shows the write performance for a 12 node cluster. The average throughput was 3.99 times that of 1 GbE and average I/O rate was 4.77 times that of 1 GbE I/O. Overall, this resulted in a reduction of run time to 39% of the original 1 GbE run time.

The read performance was significantly improved when using 10 GbE across all three metrics. The improvement was more pronounced as the num-
ber of nodes was increased. The adjacent figure shows the results for a 12 node cluster. The average throughput was 8.44 times that of 1 GbE and average I/O rate was 4.91 times that of 1 GbE I/O. Overall, this resulted in a reduction of run time to 36% of the original 1 GbE run time.

File Transfer
In this test, up to 16 files of size 10 MB through 1 GB were loaded into each DataNode. The disk, network, and checksum times are measured during the test.

In both the 1 GbE and 10 GbE runs, the disk and checksum times are very similar, with up to a 10% decrease when using 10 GbE. As shown below, there is a significant difference in network time. Using 10 GbE results in a 77% reduction in network time over 1 GbE, leading to an overall time reduction of 25%.

Genome Resequencing with Crossbow
One of the workloads commonly run on TACC’s Hadoop cluster is that of genome resequencing analysis using Crossbow. Genome sequencing is a process through which the complete DNA sequence of an organism’s genome is determined. Once this process has been completed, it is then possible to perform comparative sequencing, or resequencing, to identify polymorphisms, mutations, and structural variations between organisms.

Crossbow is a software tool developed at Johns Hopkins University for whole genome resequencing analysis. Crossbow uses Bowtie, a short read aligner and SOAPsnp for genotyping. The short read aligner arranges the sequences of DNA, RNA, or protein to identify regions of similarity between the sequences. Next, SOAPsnp is used to identify single-nucleotide polymorphisms (SNPs), which are variations between the two sequences.

Crossbow builds upon Hadoop to perform genome resequencing and alignment analysis. The insight behind this is that alignment and SNP calling can be framed as a series of Map, Sort and Reduce steps. The Map step is short read alignment, the Sort step bins and sorts alignments according to the genomic position aligned to, and the Reduce step calls SNPs for a given partition. The figure below provides more details about the Crossbow workflow.
Crossbow results with E-coli full manifest example

Two workloads were used to test Crossbow performance on a 1GbE network vs. the Intel 10GbE network; a subset of several sequenced genomes of E-coli\(^\text{vi}\) and a mouse genome. The total run times, averaged over 5 runs per configuration, are shown in the adjacent figures. Although the difference is not large, some performance benefit can be seen from the faster 10GbE interconnect. It is hypothesized that these results indicate that the Hadoop workload is bottlenecked by other limiting factors such as CPU speed and disk I/O, and does not take full advantage of the increased network speed.

Conclusion

The testing clearly confirms that 10 GbE dramatically helps TACC improve the performance of loading data into their dynamic Hadoop environment. Performance on a common TACC Hadoop workload was also examined and 10GbE was shown to provide a small improvement over 1GbE. The tests were run on enterprise-class SAS hard disk drives—clearly a performance bottleneck. It seems logical to expect that the performance benefit of running Hadoop jobs on 10 GbE will only increase given improvements in disk I/O speed as storage media moves to solid state drives or tiered storage systems with a flash tier.

TACC’s need for faster dynamic creation of Hadoop clusters was met by the performance benefit and simplicity of Intel 10 GBASE-T CNAs. By using 10 GbE, TACC was able to relieve their data loading bottleneck with a network time reduction of 77% over 1 GbE.

Crossbow results with mouse genome workload

Test Environment

Servers
- Dell PowerEdge R710
- 2 Intel Xeon E5540 quad-core processors
- 48 GB RAM
- 7 146 GB (15,000 RPM, 6 Gbps) SAS drives

Network Adapters
- 1 Intel Ethernet 10 Gigabit Converged Network Adapter (10 GBASE-T) per server
- 1 NetXtreme II BCM5709 Gigabit Ethernet per server

System Software
- Operating System: CentOS 5.3
- Hadoop: Apache™ Hadoop 1.0
- Java: Java 1.6.0

Switches
- Arista 7050T 48-port 1 GbE/10 GbE switch
- Dell Power Connect 6248 1 GbE switch

---

\(^\text{i}\) http://www.gartner.com/it/page.jsp?id=2200815
\(^\text{ii}\) http://www.tacc.utexas.edu/about
\(^\text{iv}\) http://bowtie-bio.sourceforge.net/crossbow/index.shtml
\(^\text{v}\) http://schatzlab.cshl.edu/publications/posters/SC09-Crossbow.pdf

Intel, the Intel logo, and Xeon are trademarks of Intel Corporation in the U.S. and other countries.
*Other names and brands may be claimed as the property of others
Copyright © 2012, 2013 Intel Corporation. All rights reserved.

4