Building HPC Clusters with 10Gigabit Ethernet

Introduction

High-Performance Computing (HPC) is proliferating across a wide range of industry sectors today. Traditional consumers in the HPC space have been industrial and government organizations conducting research in computational fluid dynamics, geology and aerospace. Recent arrivals in this consumer space include large enterprises in financial services, data warehousing and data mining.

It is clear that such applications will benefit from ever-higher computing engines applied to their workloads. What is often overlooked is the role of the interconnecting network in increasing throughput and reducing latency. Today, a majority of the High-Performance Computing applications use 1GE networks and latency-sensitive applications utilize InfiniBand (IB) fabrics.

This document presents data comparing the performance of two well-known HPC applications and the impact of 1GE, 10GE and InfiniBand on their performance in a benchmarking environment.

Comparing 1GE, 10G Ethernet and InfiniBand for Throughput and Latency

Networking technology can be developed to optimize certain parameters such as bandwidth, latency or quality of service classification. If these values are stretched to their limits, the results may meet the requirements but will result in a network that is difficult to setup and maintain. This leaves adopters of networks with a choice between performance optimization for a single use case and using proven and widely adopted technology that covers a majority of their use cases but not all of them. 1GE, 10GE and IB technologies are compared using well-known HPC application benchmarks and the relative merits of each are presented in terms of manageability and ease-of-use.

Figure 1 represents the System Configuration used to derive the benchmark results. The test-bed and methodology were implemented in collaboration with a premier systems vendor in the HPC community.

Processor	2 x Intel Xeon Harpertown 3.00 GHz 2x6MB L2 cache
Memory	16 x 2 GB DDR2 FBDIMM 667 MHz
Disk Controller	2 x SAS Disk Drives 146 GB 10000 RPM in RAID 1 6 x SAS Disk Drives 146 GB 10000 RPM in RAID 5
Interconnect	• 2 x Broadcom Corporation NetXtreme II BCM5708 GigE •ConnectX Infiniband 4x DDR Adapter, or •Chelsio S320 E 10 Gigabit Ethernet Adapter, or •Mellanox 10 Gigabit Ethernet
Firmware	System BIOS 1.00
Operating System	Red Hat Enterprise Linux Server 5.3 Kernel 2.6.28.9-smp
Software	Intel Fortran Compiler 11.0.074 / Intel C/C++ Compiler 11.0.074 Intel MKL 10.1.1.019 Intel MPI 3.2.011 OFED 1.3.1

Figure 1: System Configuration

Testing has been carried out using FLUENT v6 and v12 (a CFD code) and LS-DYNA MPI v3.2.1 (a nonlinear crash analysis code). Both are common in the HPC world and give a good indication of the overall system performance - FLUENT for computational fluid dynamics and LS-DYNA for simulating complex, non-linear phenomena. At the server-level, Fluent stresses CPU and memory pretty heavily, while LS-DYNA is mostly CPU-bound. Both scale pretty well on high MPI-task counts. The HPC community uses Fluent typically in the range of 16- to 64-way. LS-DYNA is typically used 8-way to 64-way today.

DYNA Results:

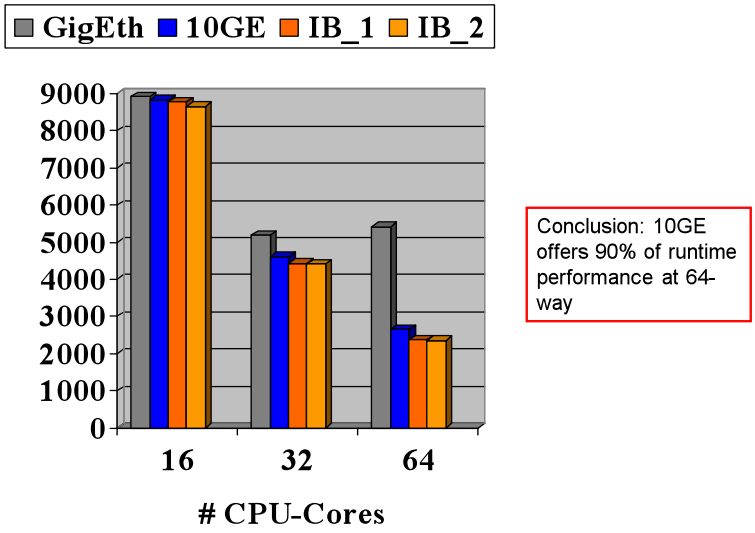

Figure 2: LS-DYNA Results – Wall-clock Runtime in sec. Model: "Car to Car"

FLUENT Test Results:

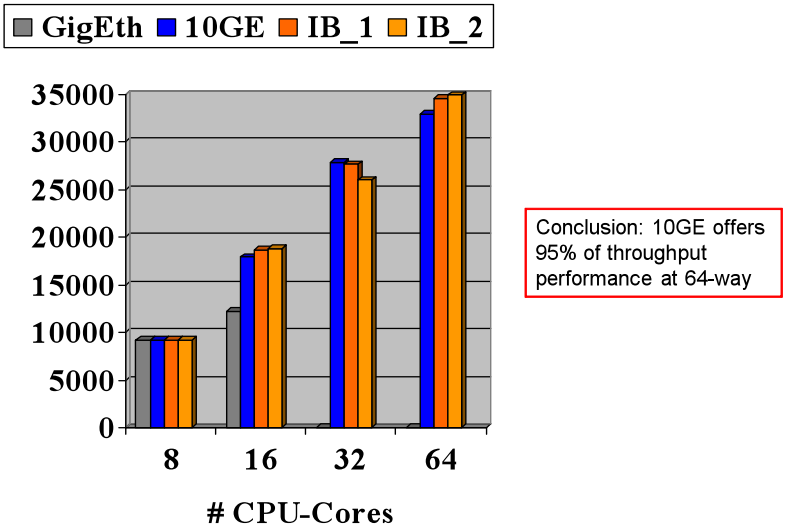

Figure 3: FLUENT Results - Fluent's Throughput Measure. Model: "S3"

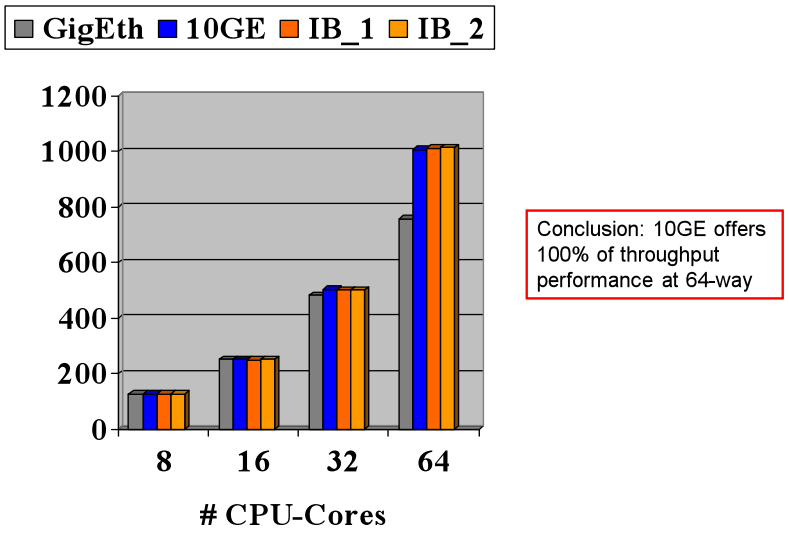

Figure 4: FLUENT Results - Fluent's Throughput Measure. Model: "L3"

Interpreting the results

The crucial finding from the test results above is the benefit that 10GE provides over 1GE in most HPC applications. As documented, 1GE networks cannot deliver the needed latency or throughput as the number of CPU cores increases. 10GE on the other hand delivers scalable performance across both the metrics, demonstrating its value while retaining the ease-of-use of an Ethernet network.

Application performance for 10GE versus 4x InfiniBand shows a 10% advantage to IB for most cases. This shows that IB's lead has narrowed considerably with the advent of 10GE.

With ease-of-management, Ethernet has the distinct advantage of bringing tools like ping and traceroute to an installation. This bypasses the need to re-train staff and ensures a faster deployment cycle for 10GE in HPC environments.

Summary

A small class of HPC applications is extremely sensitive to latency. These applications continue to be best served by the InfiniBand fabric. A majority of HPC applications on the other hand are currently running over 1GE fabrics. The reason being, Ethernet networks are easier to configure, manage and troubleshoot. In large measure this has to do with the availability of a wide variety of tools and trained network personnel well versed in IP and Ethernet technology.

The test results presented below show that the bulk of the performance gains for an HPC application accrue by upgrading the underlying 1GE network to 10GE. The resulting performance gains are several times over that obtained using the 1GE and approach within 10% of the performance derived from hard-to-manage IB networks. The 1GE networks for HPC applications are built using Ethernet switches developed for general-purpose enterprise applications. These products do not deliver the latency and non-blocking bandwidth that newer 10GE switches can provide.