

The Impact of InfiniBand Architecture on CPU Utilization

On-load/PSM Configuration Provides Nearly 400 Percent Greater Performance

Key Findings

- InfiniBand[®] was originally designed for data center applications and the architecture has been evolved by QLogic to meet the needs of the High Performance Computing marketplace
- InfiniBand offerings that use host adapters with offload processors are based on the original InfiniBand specification, and therefore are not as efficient as on-load InfiniBand architectures that were built specifically for High Performance Computing and higher-performing, denser core count processors
- Tests show that an on-load InfiniBand architecture provides better application performance at lower CPU utilization than an offload InfiniBand architecture

Executive Summary

High Performance Computing applications require tuned and efficient performance from the CPU, interconnect, MPI, and communication libraries to achieve optimal performance. Conventional wisdom would have you believe that Host Channel Adapters with offload (processing power on the card) capabilities would require less CPU utilization for running the communication library, allowing more CPU cycles for applications. Moreover, Host Channel Adapters using an on-load method and running the communication library on the host CPU

would require higher CPU cycles. These conventional views, however, are often times not correct. There are two very important factors to consider: the design and impact of the communication library on MPI application communications and the impact of faster and higher core count processors versus the offload processing capabilities of Host Channel Adapters.

Introduction

Today, there are two different InfiniBand architectures that are available to run MPI compute applications. In the first option, the Host Channel Adapter uses an offload processing approach and a communication interface for the MPI called Verbs. This method is used by Mellanox® with its Connect-X2 and Connect-X3 Host Channel Adapters. The second option uses an on-load architecture with a communication library called Performance Scaled Messaging (PSM). QLogic's TrueScale™ Host Channel Adapters take advantage of this type of architecture.

In the early 2000s, Verbs was defined as part of the original InfiniBand specification. InfiniBand was originally designed as a replacement for Ethernet and Fibre Channel for the enterprise data center. It was created to fit an I/O processing paradigm where the primary performance metric was millions of I/Os (Figure 1).

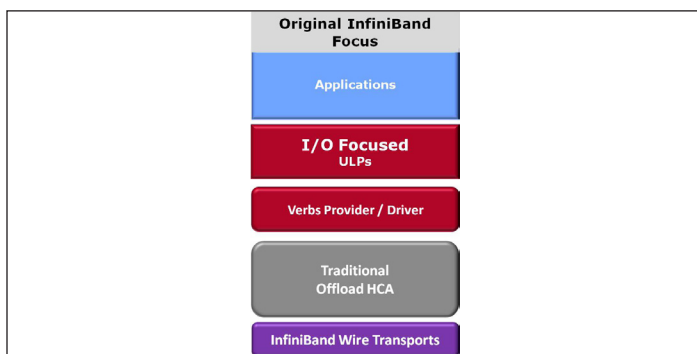


Figure 1. Original InfiniBand Focus

Verbs is the interface between the Upper Layer Protocols (ULPs) and the Host Channel Adapter. As a result of its design, a connection-oriented approach was needed to handle the data center I/O features of the interconnects and

protocols. The heavyweight design required a Host Channel Adapter with offload processing capabilities to effectively run. This is the architecture that Mellanox adopted for its ConnectX® Host Channel Adapters.

By the mid-2000s, InfiniBand had secured a niche in the High Performance Computing marketplace because of its extremely low latency and high bandwidth. However, High Performance Computing has a completely different performance paradigm built around the Message Passing Interface (MPI) protocol (Figure 2).

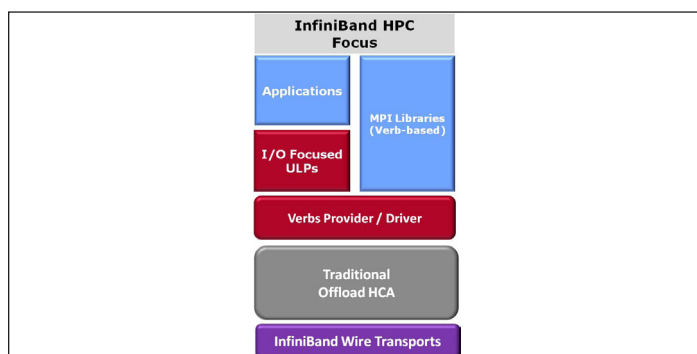


Figure 2. InfiniBand High Performance Computing Focus

MPI requires a significantly different communications architecture than what InfiniBand was originally designed for in the early 2000s. Its communication requirement is for tens of millions of relatively small messages per second. The Verbs interface, which was originally specified to allow for efficient handling of I/O requests in a data center environment, was retrofitted for the execution of the required MPI libraries. Unfortunately, due to the poor semantic match between MPI's message passing requirements and the structure of the Verbs implementation, a heavyweight protocol must be traversed to handle each message. As a result, this places a significant

burden on the host CPU and severely limits network performance, especially as a cluster is scaled.

The alternate communications interface to Verbs, PSM, was designed in the mid-2000s when it was clear that InfiniBand had found its niche in the High Performance Computing marketplace. It is implemented as a user space Linux® library with an API designed specifically for MPI. PSM's design resulted from an analysis of various MPI channel interfaces, and it was architected to perfectly match the needs of these interfaces (Figure 3).

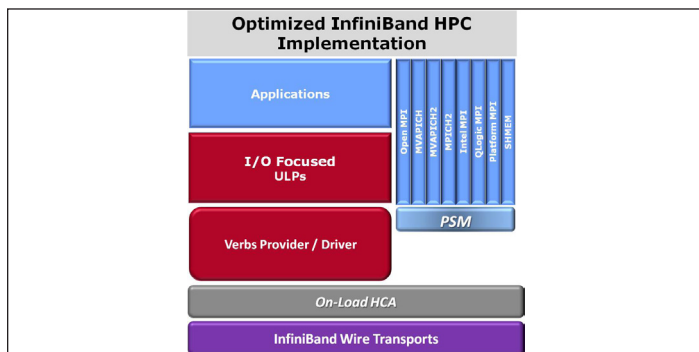


Figure 3. Optimized InfiniBand Implementation

An extremely lightweight library that facilitates MPI communications, PSM is an excellent semantic match for supporting the requirements encountered in High Performance Computing environments. Its benefits include the following:

- Extremely high MPI message rate for small messages
- Very low end-to-end latency, even at scale
- Collectives performance that scales and maintains low latency

PSM divides the responsibilities between the host driver and the Host Channel Adapter differently than Verbs-based implementations. In the PSM implementation, the host driver directly executes the InfiniBand transport layer, entirely eliminating both the heavyweight Verbs interface on the host and any transport-layer bottlenecks in the Host Channel Adapter offload processor. This makes PSM, with its on-load approach, well-suited to take advantage of today's high-performance, dense multi-core processors. QLogic's TrueScale host architecture is based on the on-load with PSM approach.

Testing was performed to determine which architecture, offload/Verbs or on-load/PSM, uses less CPU cycles for communications, thereby leaving more cycles available to the MPI application. The following comparisons are based on ANSYS® FLUENT®, which is one of the most popular computational fluid dynamics (CFD) applications in the market. The FLUENT tests were run on an Intel-based server cluster, which consists of 32 servers and one NFS server node. Each server has dual quad-core

Intel® Xeon® 5670 “Westmere” 2.93GHz processors and 24GB of memory. The Intel-based server cluster has a total of 384 cores. The cluster was run once with an offload/Verbs communications interconnect, and then the test was repeated with an on-load/PSM configuration. Platform MPI was used with the MPI stats option enabled to collect the statistics for communications and application CPU utilization.

The first ANSYS FLUENT test was with the Eddy 417K cell model. The model is relatively small, but it best shows the performance capabilities of the interconnects. The model is broken into a relatively small number of cells per core (1070 cell/core). Each step of the simulation is quickly processed by the CPU and then the results are communicated to all of the other cores in the cluster. As such, the cell model is highly sensitive to communication architecture and performance.

The results of the Eddy 417K test show that the performance of the on-load/PSM is significantly better than that of the offload/Verbs communications. The on-load/PSM configuration performed 366 percent better, which was a result of having more CPU cycles available for the application versus communications (Figure 4 and Figure 5). The MPI communication is more efficient, providing almost five times more cycles than the offload/Verbs interconnect.

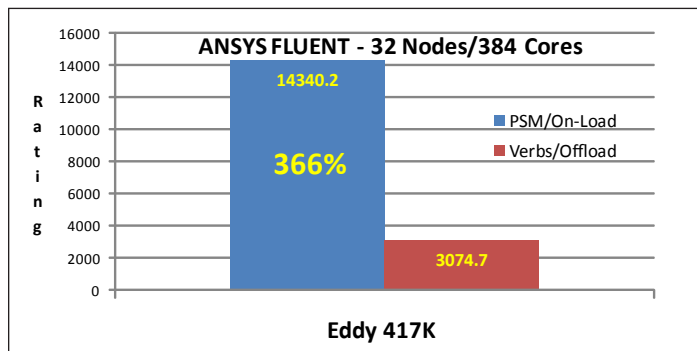


Figure 4. ANSYS FLUENT Eddy 417K Results

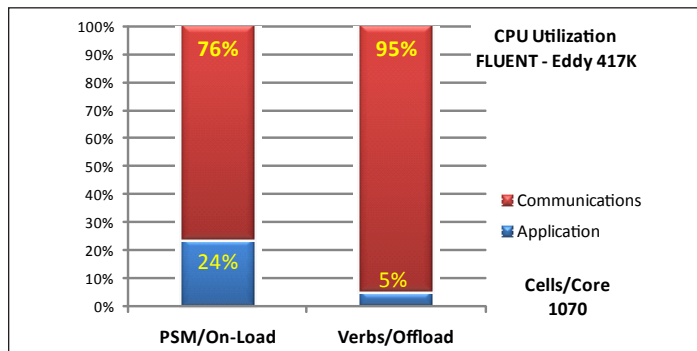


Figure 5. Eddy 417K CPU Utilization

The next test was the ANSYS Fluent Truck 111M cell model (Figure 6).

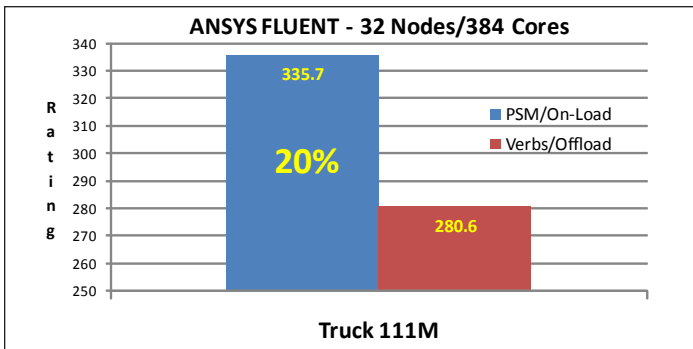


Figure 6. ANSYS FLUENT Truck 111M Cell Model

This is a much larger model, so there are more cells per core (285K cells/core) being processed for each step of the simulation. In this case, on-load/PSM performed 20 percent faster than offload/Verbs. The CPU utilization chart shows that PSM is again more communications efficient, thereby providing more CPU cycles to the application (Figure 7).

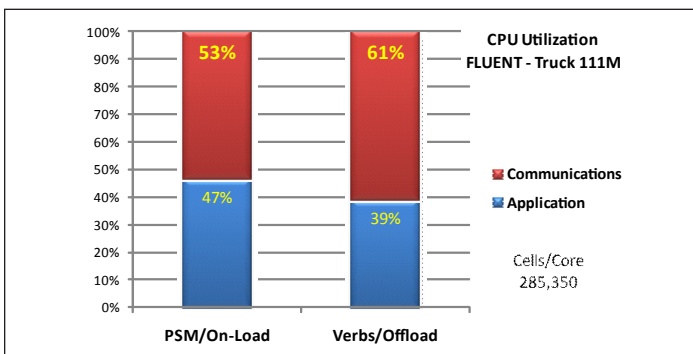


Figure 7. Truck 111M Model CPU Utilization

Conclusion

In summary, TrueScale on-load/PSM architecture with its specifically tuned MPI communications design is more efficient and higher performing than the offload/Verbs design with its retrofitted High Performance Computing/MPI implementation.

Disclaimer

Reasonable efforts have been made to ensure the validity and accuracy of these performance tests. QLogic Corporation is not liable for any error in this published white paper or the results thereof. Variation in results may be a result of change in configuration or in the environment. QLogic specifically disclaims any warranty, expressed or implied, relating to the test results and their accuracy, analysis, completeness or quality.



Corporate Headquarters QLogic Corporation 26650 Aliso Viejo Parkway Aliso Viejo, CA 92656 949-389-6000 www.qlogic.com

International Offices UK | Ireland | Germany | France | India | Japan | China | Hong Kong | Singapore | Taiwan

© 2011 QLogic Corporation. Specifications are subject to change without notice. All rights reserved worldwide. QLogic, the QLogic logo, and TrueScale are trademarks or registered trademarks of QLogic Corporation. ANSYS and FLUENT are registered trademarks of Ansys, Inc. InfiniBand is a registered trademark and service mark of the InfiniBand Trade Association. Intel and Xeon are registered trademarks of Intel Corporation. Linux is a registered trademark of Linus Torvalds. Mellanox and ConnectX are registered trademarks of Mellanox Technologies, Inc. All other brand and product names are trademarks or registered trademarks of their respective owners. Information supplied by QLogic Corporation is believed to be accurate and reliable. QLogic Corporation assumes no responsibility for any errors in this brochure. QLogic Corporation reserves the right, without notice, to make changes in product design or specifications.