X-SRQ - Improving Scalability and Performance of Multi-core InfiniBand Clusters

Galen M. Shipman\textsuperscript{1}, Stephen Poole\textsuperscript{1}, Pavel Shamis\textsuperscript{2}, and Ishai Rabinovitz\textsuperscript{2}

\textsuperscript{1} Oak Ridge National Laboratory*, Oak Ridge, TN, USA
\{gshipman,spoole\}@ornl.gov
\textsuperscript{2} Mellanox Technologies, Yokneam, Israel
\{pasha,ishai\}@mellanox.co.il

Abstract. To improve the scalability of InfiniBand on large scale clusters Open MPI introduced a protocol known as B-SRQ [2]. This protocol was shown to provide much better memory utilization of send and receive buffers for a wide variety of benchmarks and real-world applications.

Unfortunately B-SRQ increases the number of connections between communicating peers. While addressing one scalability problem of InfiniBand the protocol introduced another. To alleviate the connection scalability problem of the B-SRQ protocol a small enhancement to the reliable connection transport was requested which would allow multiple shared receive queues to be attached to a single reliable connection. This modified reliable connection transport is now known as the extended reliable connection transport.

X-SRQ is a new transport protocol in Open MPI based on B-SRQ which takes advantage of this improvement in connection scalability. This paper introduces the X-SRQ protocol and details the significantly improved scalability of the protocol over B-SRQ and its reduction of the memory footprint of connection state by as much as 2 orders of magnitude on large scale multi-core systems. In addition to improving scalability, performance of latency-sensitive collective operations are improved by up to 38% while significantly decreasing the variability of results. A detailed analysis of the improved memory scalability as well as the improved performance are discussed.

1 Introduction

The widespread availability of commodity multi-core CPUs from both Intel and AMD is changing the landscape of near-commodity clusters. Compute nodes with 8 cores (2 quad core CPUs) and even 16 cores (4 quad core CPUs) are becoming more common and 8 or more cores in a single socket are expected in the next 12-18 months. A number of these multi-core clusters are connected with InfiniBand (IB), thereby increasing the need to examine the scalability of MPI in such environments.

\* Research sponsored by the Mathematical, Information, and Computational Sciences Division, Office of Advanced Scientific Computing Research, U.S. Department of Energy, under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC.
Open MPI supports the IB interconnect using the reliable connected (RC) transport layer. RC in IB currently requires a connection to be established between each communicating pair of processes and consumes one page (commonly 4KB) of system memory for each connection. Multi-core systems increase the number of dedicated processes per node and therefore increase the number of connections per node. This additional memory consumed on the node may be substantial in a large scale multi-core system. Furthermore, maintaining a fixed amount of memory per core is becoming increasingly difficult as memory prices remain high relative to the falling price of a CPU core. Pressure on memory will increase as applications are migrated to multi-core machines.

This paper describes Open MPI’s use of the extended reliable connection (XRC) which alleviates some of the memory pressure in multi-core environments. In addition to reducing overall memory consumption in Open MPI, the use of XRC in conjunction with B-SRQ improves performance. This conjunction will be referred to as X-SRQ throughout this paper.

The rest of this paper is organized as follows. Section 2 provides a brief discussion of previous work in this area as well as an overview of the XRC architecture. Section 3 describes the new protocol, including necessary modifications to our on-demand connection wire-up scheme. Section 4 describes the test platform followed by performance analysis of the results. Section 5 summarizes relevant results and concludes with a discussion of areas of possible future work.

2 Background

The InfiniBand specification details 5 transport layers:

1) **Reliable Connection (RC):** connection-oriented and acknowledged
2) **Reliable Datagram (RD):** multiplexed and acknowledged
3) **Unreliable Connection (UC):** connection-oriented and unacknowledged
4) **Unreliable Datagram (UD):** connectionless and unacknowledged
5) **Raw Datagram:** connectionless and unacknowledged

RC provides a connection-oriented transport between two queue pairs (QPs). Work requests posted to a QP are implicitly addressed to the remote peer. The scalability limitations of connection-oriented transports are well known requiring \( \binom{N}{2} \) connections for \( N \) peers.

Fig. 1 illustrates two nodes, each with two cores connected via RC. In this example each core is running a single process and is connected to each of the remote processes. If we assume that shared memory is used for intra-node MPI communication then the total number of QPs is 4 per node.

RD allows using a single QP to send and receive from any other addressable RD QP. RD was designed to provide a number of desirable scalability features but in practice RD has proven difficult to implement with no manufacturer currently supporting this transport layer.

While some are examining the use of UD to enhance scalability, the additional costs of user-level reliability and implementation complexity are still being examined.