Multiplexing Endpoints of HCA for Scaling MPI Applications: Design and Performance Evaluation with uDAPL

Jasjit Singh, Yogeshwar Sonawane
Hardware Technology Development Group
Centre for Development of Advanced Computing (C-DAC)
Pune, India
e-mail: {sjasjit, yogeshwars}@cdac.in

Abstract—With an ever increasing demand for computing power, number of nodes to be deployed in a cluster based supercomputer is increasing. Limited hardware resources such as Endpoints (equivalent to Queue Pairs) on a Host Channel Adapter (HCA) of a high speed interconnect limit the scalability of a parallel application based on MPI that sets up reliable connections between every process pair using endpoints, prior to communication.

In this paper, we propose a novel approach of multiplexing hardware endpoints (hweps) to extend scalability. (a) We discuss critical design issues with the multiplexing technique that differentiates a hwep from its software counterpart (swep) and enables sharing of hwep by multiple sweps. (b) We introduce the concept of Virtual Identifier (VID) which ensures that the connection between hardware endpoints is strictly one-to-one. (c) We also present static mapping scheme that offsets the overheads incurred due to multiplexing.

User Direct Access Programming Library (uDAPL) defines a single set of APIs for all RDMA capable transports. We have incorporated the proposed multiplexing technique as a part of uDAPL implementation. Using this approach, we are able to scale MPI applications beyond the limit imposed by HCA and with no visible performance degradation.