

Diploma thesis

Shared Memory Support for InfiniBand™ MPICH2-Device

Motivation

A System Area Network (SAN) possesses a small latency period, a high bandwidth and a lightweight protocol. All of these properties fulfils InfiniBand. InfiniBand, as the latest technology for interconnecting computational nodes and I/O nodes, gains more and more importance in high performance computing. This shows the increasing number of InfiniBand systems that were built or will be planned. For example the Terascale Cluster of the Virginia Tech University with 1100 G5-Power-Macs, with 2-GHz-Dualprocessor and an InfiniBand communication network achieves more than 10 TFLOPS.

A Symetric Multiprocessor (SMP) has many advantages and is popular with Cluster as a system of many SMP machines, with 2 or more processors. But the use of SMP Cluster is still depending on the type of Cluster and the applications for the Cluster. Cluster, consisting of small SMP's, show advantages of Cluster and Multiprocessors. Also it is a simple strategy for implementing a scalable and high performance computing facility. It behaves much like a distributed memory multicomputer except that each node actually has multiple CPUs sharing a common memory. We can reach good improvements in performance for many problems with relative small expenses for hardware and sustainable effort in the system software. Multithreaded applications profit by that particularly and the communication within SMP's is carried out via shared memory variables and between SMP's by message passing.

MPICH is the most used open source implementation of MPI. Message Passing Interface (MPI) is the standard for programming parallel computers using the computational model message-passing. MPICH2 is an all-new implementation of MPI and supports both MPI standards, MPI-1 and MPI-2. It is still in development and the major part of MPI-1 and many MPI-2 routines have been implemented. Furthermore channel devices for InfiniBand and Shared Memory were implemented.

So it is attempted to join the attributes of InfiniBand with the advantages of SMP multi-processor systems and to implement this in a channel device for MPICH2.

Task formulation

A disadvantage of the existing device¹ for InfiniBand is that processors don't know if they are on one Motherboard and consequently within an SMP. Each communication operation with data of a process always happens over the InfiniBand device. Thereby an overhead arises by the network controller and by the InfiniBand protocol.

The aim is to decide if a processor is within an SMP. If it is within an SMP, it will suppose to communicate over a Shared Memory device and to bypass InfiniBand. It is possible to use the advantages of Shared Memory (bandwidth of system bus, double latency period of system bus) and the result would be a possible increase in performance.

Now the task is to improve the existing device for Infiniband with Shared Memory functionality.

¹ This device was created by René Grabner and Frank Mietke in their diploma thesis: MPICH2-Device for InfiniBand™

Following things are to consider:

- creating the Shared Memory device for the existing InfiniBand device
- the MPI-API and all communication operations are supposed to remain unchanged
- nodes shouldn't realize over which communication network the operations are executed, i.e. transparency for the applications
- building an appropriate "decision-maker" that detects which communication network is to take for the communication operation
- detecting if the remote processor is within an SMP and consequently Shared Memory can be used or if it is on another SMP, InfiniBand has to be used
- choosing the right device with an obvious identification of the processors
- measuring the performance and testing the built device

Author : Marco Steiger
marco.steiger@s1998.tu-chemnitz.de
Matrikel-Nr.: 21833

Supervising professor: Prof. Dr.-Ing. W. Rehm

Supervisor: Dipl.-Inf. Frank Mietke

Begin: 01.12.2003

End: 31.05.2004