Optimised MPI for HPEC applications

Gérard Cristau

Thales Computers
3100 Spring Forest Road
Raleigh, NC 27616
Tel: (919) 231-8000
Fax: (919) 231-8001

Gerard.Cristau@thalescomputers.fr
(France)

We propose an implementation of the MPI standard, suitable for HPEC applications.

The HPEC requirements addressed by the proposed implementation include:
- support of "zero copy" transport mechanisms
- memory management
- handling heterogeneous communication topologies

Zero copy communications are key for HPEC, in order to allocate memory bandwith to actual communications rather than to data transfers between buffers. The semantics and implementation issues associated with the mapping of MPI primitives on zero copy mechanisms such as VIA [1] or RDMA [2] have been discussed [3, 4, 5] and the feasibility and performance benefits have been demonstrated on the basis of popular MPI implementations such as LAM [6] or MPICH [7].

Optimised, industrial MPI implementations based on zero copy transport should be provided on the high-performance, low latency media suitable for HPEC systems. The implementation described in this paper uses a serial RapidIO network across PowerPC computing nodes. The applicability to Infiniband or Gigabit ethernet is also discussed.

Zero copy communications implies constraints on memory allocation and buffer management; memory buffer management is neither covered nor precluded by the MPI standard. Defining an API allowing the needed level of memory buffer management, while preserving compatibility with the MPI standard is one of the issues addressed by this implementation.

The paper finally proposes a mechanism, compatible with the MPI standard, to describe the communication topology of a HPEC software, such as a dataflow, signal processing application. Such an application is described as a set of tasks interconnected by virtual links. The application graph, and its mapping on a possibly heterogeneous computing and communication infrastructure, is defined through an independent configuration file, allowing to modify the application deployment without recompilation. The translation of the application graph in terms of MPI groups and communicators is carried out at run-time, and is transparent to the user.

References

[1] http://www.viarch.org

[2] http://www.rdmaconsortium.org

[3] Jiuxing Liu, Jiesheng Wu, Sushmitha P. Kini, Pete Wyckoff
High performance RDMA-based MPI implementation over InfiniBand
Proceedings of the 17th annual international conference on Supercomputing
San Francisco, CA, USA, 2003

[4] R. Dimitrov and A. Skjellum.
An Efficient MPI Implementation for Virtual Interface (VI)
Architecture-Enabled Cluster Computing. http://www.mpi-softtech.com/publications/, 1998.

[5] Olivier Aumage, Guillaume Mercier.
MPICH/MADIII: a Cluster of Clusters Enabled MPI Implementation.
In 3rd {IEEE/ACM} International Symposium on Cluster Computing and the Grid
Tokyo, Japan, May 2003.

[6] Olivier Aumage, Guillaume Mercier, and Raymond Namyst.
MPICH/Madeleine: a true multi-protocol MPI for high-performance networks.
In Proc. 15th International Parallel and Distributed Processing Symposium (IPDPS 2001)
San Francisco, April 2001.