High Productivity MPI – Grid, Multi-Cluster, and Embedded System Extensions

Pirabhu Raman, Anthony Skjellum, Rossen Dimitrov, Kumaran Rajaram, and ***Puri Banglore***
Verari Systems Software, Inc.
Phone: +1-205-314-3471
E-mail Addresses: {pirabhu, tony, rossen, kums, puri}@mpi-softtech.com

High Productivity MPI is an approach to extending MPI to support multiple

- Implementations (IMPI, IMPI-2)
- Owner domains
- Architectures
- Networks
- Operating Systems
- Faults
- Interacting dynamic groups

without relying on a two-level implementation (one MPI implementation calling another). MPI implementations must be able to connect, reconnect, and work well with dynamic, intermittent resources, under the expectation that user applications will also become somewhat fault-aware in order to retain scalability.

This paper addresses the many concerns that arise in offering composable sessions in which multiple-vendor MPI's can be supported (starting from but not ending with IMPI protocol). Experiences with IMPI, and a new proposal, IMPI-2, are offered. This paper addresses specific issues about interoperating the gamut of MPI-2 services in the interoperable setting, which to our knowledge have not been addressed elsewhere.

The results of this work are open specifications, together with our own vendor-implementation of these MPI capabilities. Other open and commercial MPI's could adopt IMPI plus these other extensions in order to participate in the hierarchical, heterogeneous, grid computing settings, without mandating new MPI implementations in such settings. The authors' goals in offering these new protocols as proposals is to encourage the High Productivity Computer world to enter into significant discussions about their adoption. The goal is to offer these capabilities without mandating grid-computing infrastructure.

Specific requirements for supporting several overlapping and non-overlapping networks of varying performance (overhead, bandwidth, latency, concurrency) are discussed, in terms of progressive MPI implementations, and the joint progressive nature of compliant MPI's working together with the IMPI-2 protocol framework. Note that existing multi-cluster/grid solutions apparently cause excessive polling, and do not support the degree of scalability or appropriate intra-network performance that would otherwise pertain to correctly composed MPI implementations.

In the spirit of pursuing practical fault tolerance in this setting, the extension of checkpoint-restart facilities to High Productivity MPI is considered both from the perspective of MPI I/O on

single implementations, and in terms of the MPI I/O as extended to multiple implementations connected through an IMPI-2 protocol framework.

Other aspects of fault-tolerant MPI are beyond the scope of this paper.

The authors note that there are many thorny issues associated with supporting all the features of MPI-2 across multiple vendor implementations, and several of these issues are highlighted. A "core MPI-2" (subset of MPI-1 plus subset of MPI-2) plus some additional dynamic processing extensions are suggested as a best practice for computing in this setting.

Networks of grids (or multi-clustering) represent an interesting capability but also a challenge in terms of the publication of the entire structure (including IP addresses) globally. This work also considers the use of techniques such as NAT and port forwarding, together with gateway nodes, in order to allow for structured, manageable descriptions of hierarchical parallel resources, so that appropriate communication bandwidth remains possible between clusters, without mandating a public model for all resources involved. This has been accomplished with IMPI as is, and extensions in the IMPI-2 framework are discussed.

The paper also tries to address some of the drawbacks with existing IMPI protocols including

- Involvement of user in starting MPI jobs distributed across multiple platforms
- Global collective operations mandated by standard
- Non-portable parallel job startup mechanism

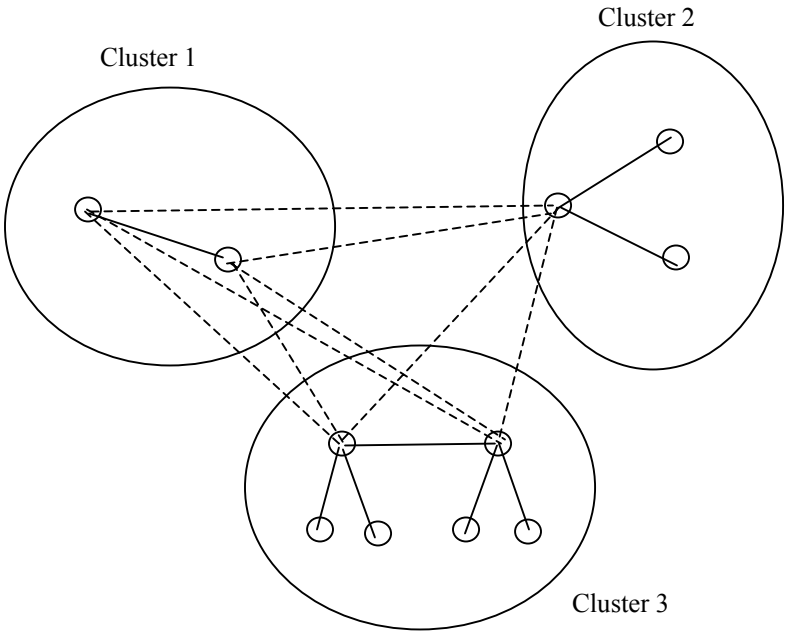

Figure 1: Connectivity Architecture of a multi-cluster, multi-implementation MPI

Figure1 shows a typical scenario where multiple-clusters (possibly under different owner domains and different internal networks) having multiple implementations of MPI. Our approach enables these multiple implementations to work seamlessly without requiring a new-layered implementations and also enables individual implementations to use proprietary and optimized communication protocols with no increased overhead. The IMPI-1 implementation, which is currently available as part of MPIPro, was supported by NIST through the Contract # 50-DKNB-1-SB082.

References:

1. MPI Forum. *MPI: A Message-Passing Interface Standard*. 1994. http://www.mpi-forum.org/docs.
2. IMPI Steering Committee. *IMPI Draft Standard*. 1999. http://www.nist.gov/impi/.
3. Gropp, W. et al. A High-Performance Portable Implementation of the Message-Passing Interface. *Journal of Parallel Computing*, 22, 1996, pp: 789-828.
4. Skjellum, A. and McMahon, T. *Interoperability of Message-Passing Interface Implementations: a Position Paper*. http://www.mpi-softtech.com/company/publications/files/interop121897.pdf