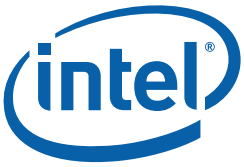

Intel® MPI Library Extensions to PMI Protocol

Intel ® MPI Library extensions to PMI protocol

This document describes modifications required to introduce into the third party process manager to make it work with Intel® MPI 3.0 tasks.

Contents

Revision History

Document Number	Revision Number	Description	Revision Date
	01	Initial release. Created by Dmitry Ezhov	09/27/2006
	02	New look and feel. Edited by Andrey Derbunovich and Julia Gasenina	10/05/2006
	03	Edited by Julia Gasenina for style and Intel guidelines for external distribution compliance	10/06/2006

Legal Notices

The Intel(R) MPI Library is based on MPICH2* from Argonne National Laboratory* (ANL).

MPICH2 Copyright Notice

+ 2002 University of Chicago

This software was authored by: Argonne National Laboratory Group

W. Gropp: (630) 252-4318; FAX: (630) 252-5986; e-mail: gropp@mcs.anl.gov

E. Lusk: (630) 252-7852; FAX: (630) 252-5986; e-mail: lusk@mcs.anl.gov

Mathematics and Computer Science Division

Argonne National Laboratory, Argonne IL 60439

http://www.anl.gov/

MPICH2 information can be obtained from http://www-unix.mcs.anl.gov/mpi/mpich2/index.htm.

The Intel(R) MPI Library is also based in part on RDMA drivers from MVAPICH2* from Ohio State University. These drivers support RDMA-capable network fabrics via the DAPL* API from the DAT Collaborative.

The MVAPICH, MIBAPICH, and MVAPICH2 softwares are developed by The Ohio State University's Network-Based Computing Laboratory (NBCL) headed by Professor Dhabaleswar K. (DK) Panda.

Copyright (c) 2002-2003, The Ohio State University. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

* Redistributions of source code must retain the above copyright notice, this list of conditions and the disclaimer below.

* Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

* Neither the name of The Ohio State University nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

* Modifications to the above softwares MUST be provided to the Network Based Computing Laboratory at The Ohio State University in electronic form by sending an e-mail to mvapich_request AT cis.ohio-state.edu, and The Ohio State University may make such modifications publicly available under the terms of this license.

* If you are presenting or publishing any results/performance numbers, demonstrating any system, or having any press release based on systems running MVAPICH, MIBAPICH, or MVAPICH2 and/or their modified versions, you MUST acknowledge the contributions by the Network-Based Computing Laboratory at The Ohio State University headed by Prof. Dhabaleswar K. (DK) Panda and their relevant publication (s) in your presentation/ publication/demo/press release.

THIS SOFTWARE IS PROVIDED BY THE OHIO STATE UNIVERSITY AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE OHIO STATE UNIVERSITY OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

MVAPICH2 information can be obtained from

http://nowlab.cis.ohio-state.edu/projects/mpi-iba.

The Intel(R) MPI Library is also based in part on the DAPL API implementation from the IBAL* InfiniBand(*) Access Layer SourceForge* project.

"This software program is available to you under a choice of one of two licenses. You may choose to be licensed under either the GNU General Public License (GPL) Version 2, June 1991, available at http://www.fsf.org/copyleft/gpl.html, or the Intel BSD + Patent License, the text of which follows:

"Recipient" has requested a license and Intel Corporation ("Intel") is willing to grant a license for the software entitled InfiniBand(tm) System Software (the "Software") being provided by Intel Corporation.

The following definitions apply to this License: "Licensed Patents" means patent claims licensable by Intel Corporation which are necessarily infringed by the use or sale of the Software alone or when combined with the operating system referred to below.

"Recipient" means the party to whom Intel delivers this Software.

"Licensee" means Recipient and those third parties that receive a license to any operating system available under the GNU Public License version 2.0 or later.

Copyright (c) 1996-2003 Intel Corporation. All rights reserved.

The license is provided to Recipient and Recipient's Licensees under the following terms.

Introduction

Intel® MPI Library [References] is a multifabric MPI-2 implementation based on MPICH2 and MVAPICH2. Intel MPI enables you to select any of the available MPI devices at runtime. It uses facilities provided by the Process Manager (PM) via the Process Management Interface (PMI).

This document specifies Intel extensions to the PMI protocol originally developed by Argonne National Laboratory. These extensions help to reduce startup overhead for MPI applications. The document helps developers of PMI-capable process managers to understand required implementation in order to make sure that MPI applications linked to the Intel MPI Library work properly in their environment.

1 Extensions to PMI protocol

The process launcher *mpiexec* communicates with the Process Manager (PM) to start the processes on the nodes of a parallel machine. It propagates the startup data to the nodes through an out-of-band communication mechanism provided through PMI. The created processes use this out-of-band communication mechanism to exchange the information before setting up the MPI communication.

Intel MPI Library v3.0 introduces extensions to the original PMI protocol. PM compatible to Intel MPI Library 3.0 provides handles to new PMI commands and sends a corresponding response into the process as described below.

At the initialization stage each process sends the `get_ranks2hosts` request to the process manager in the common used form that is `cmd=get_ranks2hosts`.

After receiving this message the PM gathers necessary information about the set of processes.

CAUTION: The list of hosts should contain unique (non-recurring) names only.

In response to the `get_ranks2hosts` command the PM sends two messages. The messages are text strings with the following formats:

```
put_ranks2hosts <msglen> <num_of_hosts>
```

```
<hnlen> <hostname> <rank1,rank2,…,rankN,>
```

These messages are described in the following sections.

First message

This message provides preliminary information about coming data.

Format

```
put_ranks2hosts <msglen> <num_of_hosts>
```

Description

`put_ranks2hosts`	**The keyword**
`<msglen>`	**The number of characters in the next message + 1**
`<num_of_hosts>`	**The total number of the non-recurring host names that are in the set**

Second message

This message contains actual information about process placement. The message is a set of sequences divided by a space.

Format

```
<hnlen> <hostname> <rank1,rank2,…,rankN,>
```

Description

`<hnlen>`	**The number of characters in the next `<hostname>` field**
`<hostname>`	**The node name**
`<rank1,rank2,…,rankN,>`	**A comma separated list of ranks executed on the `<hostname>` node; if the list is the last in the response message then it must be followed a blank space**

NOTE: Both messages must have the End of Line (EOL) `'\n'` character at the end.

The following example illustrates a real response. For example, for an MPI application with eight processes, started on two nodes *node001-node002*, the messages are as follows:

```
put_ranks2hosts 40 2

7 node001 1,3,5,7, 7 node002 0,2,4,6,
```

2 References

For more information on the concepts discussed in this document, consult the following sources:

- Intel's MPI website, http://www.intel.com/go/mpi contains complete descriptions of Intel MPI as well as case studies and reference manuals.
- The Message Passing Interface Forum: MPI-2: Extensions to the Message Passing Interface. July 18, 1997.