White Paper

The SGI® MPI Strategy

Options and Optimizations

Table of Contents

A number of different MPI implementations are available for SGI® Altix® servers. The SGI Message Passing Toolkit (MPT) is an MPI implementation optimized for SGI hardware and fully supported by SGI. Intel MPI, Scali MPI Connect, Voltaire MPI, and Microsoft MPI provide additional supported commercial implementations, while open-source options include Open MPI, MVAPICH, and MVAPICH-2. Choosing the best MPI solution for your needs requires careful consideration of a number of factors such as performance, support, and portability issues. This paper explores the available options in terms of a set of defined selection criteria to assist users in making an informed choice.

1.0 Introduction

The Message-Passing Interface (MPI) is a library specification enabling communication between the nodes of a compute cluster when running a parallel application. While the interface definition is standard, there are various commercial and Open Source implementations of MPI available.

To provide SGI users with the best possible MPI options on its InfiniBandTM based platforms—the Altix XE and Altix ICE—SGI supports a full range of commercial and Open Source MPI solutions, including the SGI Message Passing Toolkit (MPT). Given the diversity of options, the question obviously arises, "which implementation of MPI is best?"

Not surprisingly, the answer to this question depends on your requirements. Factors to consider may include which SGI platform you have, the other platforms on which you want your application to run, which interconnects you use, etc.

This paper explores the various MPI libraries available for SGI hardware in terms of a range of possible selection criteria to give you the information to make an informed choice. Relative performance numbers on a standard benchmark are provided.

MPI Version	Measured Latency (microseconds)	
	8 byte message	256 KB message
MPT	3.4	220
MVAPICH	3.3	225
MVAPICH2	3.3	213
Open MPI	3.5	317
Intel MPI	4.5	213

Table 2. Measured small and large message latency for several MPI implementations [using the OSU MPI latency benchmark.

2.0 MPI Selection Criteria

There are a number of possible factors which may contribute to your decision when selecting an MPI library. This section describes the factors that have come up most often in customer discussions. In later sections, we present each MPI library in terms of these critieria. It's up to you to decide how to weight the various factors.

2.1 Performance

For many, performance will be the single most important factor in MPI selection. To better understand the relative performance of each option, SGI undertook a performance study using the Ohio State University MPI latency benchmark (http://mvapich.cse.ohio-state.edu/benchmarks/).

A number of MPI implementations were tested on an identical hardware configuration consisting of Altix ICE systems with 2.33 GHz Xeon processors and an InfiniBand interconnect. Tests were performed with both small (8 byte) and large (256Kbyte) messages. Each test was performed 1000 times and results were averaged.

MPI Library	SGI Support			Strengths
	Altix 4X	AltixXE	AltixICE	
Commercially Supported Options				
SGI MPT	✔	✔	✔	Optimized performance on SGI hardware and supported by SGI
Intel MPI	✔	✔	✔	Multi-platform, multi-interconnect, ISV preferred
Scali MPI Connect	✔	✔	✔	Multi-platform, multi-interconnect, ISV preferred
Voltaire MPI		✔		Multi-platform, multi-interconnect, ISV preferred
Microsoft MPI		✔	✔	
Open Source Options				
MVAPICH		Note	✔	Source can be modified, research orgs may prefer
MVAPICH2		✔	✔	Source can be modified, research orgs may prefer
Open MPI		✔	✔	Source can be modified, research orgs may prefer

Table 1. MPI Implementations available for SGI Altix platforms.

As you can see, with two notable exceptions, there is not much separating the performance of the various MPI libraries. Intel MPI is about 29% slower for small messages, while Open MPI is 45% slower with large messages. These are obviously significant differences and would be expected to effect the performance of applications that rely on small or large messages, respectively. We did not explore the possible reasons for these results. The other implementations performed near the limits of the underlying InfiniBand interconnect on these tests.

With the exception of these two outlying results, this small study shows no performance-based reason to choose one implementation over another, with results varying by no more than 6%.

2.2 Open Source vs. Commercial

You may already have a pretty good idea whether you will choose an Open Source MPI library or a commercial library. In general terms, independent software vendors (ISVs) lean toward commercial versions of MPI, while research organizations and in-house developers tend toward Open Source. MPT is also targeted for in-house developers creating applications for the SGI platform. Additional factors to consider are:

- Implementation quality (bug frequency, features, etc.)
- Frequency of releases/updates
- Ability to get timely support
- Ability to make local modifications and fixes to the source code of the MPI implementation

2.3 Support

The ability to obtain support for a particular MPI library including technical assistance, bug fixes, enhancements, etc. may be an important consideration.

With commercially supplied MPI libraries, you get support direct from the vendor. For instance, SGI provides full technical and engineering support for MPT. When choosing a commercial option, you may want to assess the quality of support available from your supplier and/or vendor.

With Open Source MPI options you have three possible sources of support:

1. Self-support. You can view and modify the library code as necessary.
2. Community support. The group supplying the library may provide some limited support and periodic updates with bugfixes, etc. The Open Source community may also provide assistance.
3. Limited vendor support. Vendors that supply Open Source MPI libraries may provide limited technical support. For instance, SGI makes several Open Source options (currently MVAPICH-2 and Open MPI) available via Supportfolio. SGI Technical Support will take support calls on these options, but SGI Engineering does not provide engineering support.

2.4 Portability

Another important question to ask is how portable does your MPI application need to be? There are three factors to consider when assessing portability:

1. Hardware Coverage. Does the option support all your hardware?
2. OS Coverage. Does the option support all the operating systems you need to run the application on?
3. Interconnect Coverage. Will the option run on the interconnect(s) you will use?
4. Binary compatibility. Since MPI is a standard API, all MPI implementations provide application source code compatibility. Some implementations such as MPT, Intel MPI, and HP MPI provide binary compatibility across a range of hardware, OS, and interconnect variations.

2.5 MPI-2 Feature Support

The MPI specification has been created and extended over time by the MPI Forum, which defines and maintains the MPI standard. The first version of MPI, MPI-1 was first presented in 1994. Since that time, it has been widely used and accepted, and many existing applications comply with the most recent update to MPI-1, the MPI-1.2 specification which was released in 2003.

Because the MPI 1.2 specification met most of the needs of application programmers and the large size of the MPI 2 standard, adoption of MPI-2 by MPI implementers has been relatively slow. However, certain features of MPI-2 such as scalable file I/O, dynamic process management, and collective communication are highly desirable and frequently requested.

MPI-2 adds a number of capabilities on top of MP1-1.1 while retaining compatibility with it. The major extensions to MPI-1 are:

- Process creation and management
- One-sided communications
- Extended collective operations
- External interfaces to define new nonblocking operations
- More flexible I/O options

Some MPI implementations adhere to much of the MPI-2 specification (see www.mpi-forum.org/docs/mpi2-report.pdf for the full specification) while others have adopted important features. If you need a certain feature from MPI-2, you must be careful to make sure that the implementation you choose supports that feature.

3.0 Commercial MPI Libraries

3.1 SGI Message Passing Toolkit

MPT, which has been optimized to run on all SGI hardware platforms and is fully supported by SGI, has proven itself in the field over years of use on Altix systems running Intel® Itanium® Processors. MPT is now available for use with Altix XE and Altix ICE x86-64 systems running InfiniBand platforms. Because MPT offers optimized performance, comprehensive technical support, binary compatibility, and support for key MPI-2 features it may be the best commercial option for those who are primarily interested in running applications on SGI systems.

3.2 Intel MPI Library

The Intel MPI library is a widely-used commercial implementation of MPI-2 with the broadest hardware, OS, and interconnect support of the commercial MPI options. It may be a good option if you require full MPI-2 compatibility, or need to run applications across a variety of hardware. However, SGI testing revealed possible performance limitations with small message sizes relative to other implementations (see table 2). This factor should be considered carefully by users with applications that rely on small messages.

SGI Message Passing Toolkit (MPT)	
Performance	Consistent performance for both small and large messages.
Open Source or Commercial?	Commercial library
Support	Full technical and engineering support from SGI
Portability	
– Hardware Support	SGI Altix hardware based on Intel® Xeon® and Intel® Itanium processors
– OS Support	SUSE Linux Enterprise Server and Red Hat Enterprise Linux
– Interconnect Support	GBE, NUMAlink, InfiniBand, shared memory, sockets
– Binary Compatibility	All SGI platforms
MPI-2 Support	Added support for MPI I/O, MPI Thread Safety, MPI Process Spawn, One-sided communication, Fortran 90 and C++ language bindings, MPI Ports

Table 3. SGI Message Passing Toolkit (MPT)

Intel MPI Library	
Performance	Lower observed performance for small messages in SGI tests
Open Source or Commercial?	Commercial library
Support	Full support from Intel
Portability	
– Hardware Support	IA-32, Intel® 64, or IA-64 architecture using Intel® Pentium® 4, Intel® Xeon® processor, Intel® Itanium processor family and compatible platforms
– OS Support	Microsoft Windows* Compute Cluster Server 2003 (Intel® 64 architecture only); Red Hat Enterprise Linux* 3.0, 4.0, or 5.0; SUSE* Linux Enterprise Server 9 or 10; SUSE Linux 9.0 thru 10.0 (all except Intel® 64 architecture starts at 9.1); HaanSoft Linux 2006 Server*; Miracle Linux* 4.0; Red Flag* DC Server 5.0; Asianux* Linux 2.0; Fedora Core 4, 5, or 6 (IA-32 and Intel 64 architectures only); TurboLinux*10 (IA-32 and Intel® 64 architecture); Mandriva/Mandrake* 10.1 (IA-32 architecture only); SGI* ProPack 4.0 (IA-64 architecture only) or 5.0 (IA-64 and Intel 64 architectures)
– Interconnect Support	TCP/IP, Myrinet, InfiniBand, Quadrics, Shared Memory, and others
– Binary compatibility	Yes
MPI-2 Support	Implements the full MPI-2 specification.

Table 4. Intel MPI Library

3.3 Scali MPI Connect

Scali MPI Connect is available for sale from SGI and specifically tuned for HPC performance. Although the latency performance of Scali MPI Connect was not assessed in the testing described in this paper, Scali does provide links to a number of application-specific benchmarks on the MPI Connect webpage (http://www.scali.com/content/view/35/).

3.4 Voltaire MPI

Voltaire offers a complete family of InfiniBand interconnect solutions. Voltaire MPI is provided as part of Voltaire's InfiniBand software stack. Because the SGI Altix XE uses Voltaire InfiniBand technology, Voltaire MPI can be used with Altix XE (or other clusters that use Voltaire InfiniBand.) Voltaire MPI is based on MVAPICH-0.94 with features taken from version 0.98. It supports local in-memory communication, and provides support for C and Fortran applications.

Scali MPI Connect	
Performance	Not measured in SGI tests.
Open Source or Commercial?	Commercial
Support	Vendor
Portability	
– Hardware Support	Intel® IA32; Intel Itanium® IA64; Intel EM64T; AMD Opteron/AMD64
– OS Support	Red Hat RHEL 3,4,5; Novell SLES 9,10; Sun Solaris 10
– Interconnect Support	TCP/IP; Full Ethernet including 10 Gigabit; SCI; Myrinet®; InfiniBand®; InfiniPath™
– Binary Compatibility	Yes
MPI-2 Support	MPI-1.2 compliant with MPI-2 multi-thread safety

Table 5. Scali MPI Connect

Voltaire MPI	
Performance	Not measured in SGI Test
Open Source or Commercial?	Commercial
Support	Vendor
Portability	
– Hardware Support	x86_64 (AMD & EM64t); x86; ia64; ppc64
– OS Support	RHEL 4 UP 5; RHEL 5; Suse SLES 10 sp 1
– Interconnect Support	InfiniBand
– Binary Compatibility	No
MPI-2 Support	None

Table 6. Voltaire MPI

3.5 Microsoft MPI

Microsoft MPI can be used to run MPI applications on Altix XE, or other servers when they are configured with Windows Server 2003. Microsoft MPI (MS-MPI) is a version of the Argonne National Labs Open Source MPI2 implementation that is widely used by existing HPC clusters. MS-MPI is compatible with the MPICH2 Reference Implementation and other MPI implementations

4.0 Open Source MPI Libraries

In addition to the available commercial MPI implementations, a variety of Open Source implementations are available. MVAPICH-2 and Open MPI are available via SGI Supportfolio (support.sgi.com) for download and use on SGI systems. MVAPICH and other Open Source options will also work on SGI platforms, but are not available through Supportfolio.

MVAPICH is an MPI-1 implementation based on MPICH (Argonne National Laboratory) and MVICH (Lawrence Berkeley Laboratory). MVAPICH was implemented and is available from the Network-Based Computing Laboratory at Ohio State University (http://mvapich.cse.ohio-state.edu/index.shtml).

MVAPICH2 is an MPI-2 implementation that also comes from Ohio State. It is based on MPICH2 and MVICH.

Open MPI is an MPI-2 implementation developed and maintained by a consortium of academic, research, and industry partners (www.openmpi.org).

Of the three options, Open MPI showed markedly longer latencies for large message transfers in SGI testing (see Table 2) and should therefore be avoided for use with any applications that will rely on large messages. MVAPICH and MVAPICH2 demonstrated similar performance and both were consistent with the best results seen. Either is a good option depending on whether you want MPI-1 or MPI-2.

Microsoft MPI	
Performance	Not tested.
Open Source or Commercial?	Commercial
Support	Microsoft
Portability	
– Hardware Support	AMD Opteron; AMD Athlon 64; Intel Xeon with Intel EM64T; Intel Pentium with Intel EM64T
– OS Support	Windows Server 2003
– Interconnect Support	Gigabit Ethernet, InfiniBand, or any network that provides a WinSock Direct–enabled driver
– Binary Compatibility	No
MPI-2 Support	MPI-2 compliant

Table 7. Microsoft MPI

MVAPICH	
Performance	Comparable to other implementations
Open Source or Commercial?	Open Source
Support	OSU Network-Based Computing Laboratory, SGI courtesy support
Portability	
– Hardware Support	EM64T, Opteron, IA-32, IBM PPC and Mac G5
– OS Support	Linux, Solaris, and Mac OSX
– Interconnect Support	TCP/IP, InfiniBand, shared memory
– Binary Compatibility	No
MPI-2 Support	None

Table 8. MVAPICH

MVAPICH2	
Performance	Comparable to other implementations
Open Source or Commercial?	Open Source
Support	OSU Network-Based Computing Laboratory, courtesy support from SGI
Portability	
– Hardware Support	EM64T, Opteron, IA-32, IBM PPC and Mac G5
– OS Support	Linux, Solaris, and Mac OSX
– Interconnect Support	InfiniBand, TCP/IP, iWARP,
– Binary Compatibility	No
MPI-2 Support	MPI-2 compliant

Table 9. MVAPICH2

Open MPI	
Performance	Poor performance with large messages on SGI Testing.
Open Source or Commercial?	Open Source
Support	Open MPI mailing lists
Portability	
– Hardware Support	All common platforms supported by the following OS's.
– OS Support	Linux, OS X, Solaris
– Interconnect Support	TCP/IP over Ethernet, Shared memory, Myrinet / GM, Myrinet / MX, Infiniband / OpenIB, Infiniband / mVAPI, Portals
– Binary Compatibility	Yes
MPI-2 Support	MPI-2 Implementation, supports everything except one-sided operations API which will be released with MPI v1.1.

Table 10. Open MPI

5.0 Recommendation

Based on the information provided and the requirements of the installation the reader should select the best MPI for their application.

Corporate Office
1140 E. Arques Avenue
Sunnyvale, CA 94085
(650) 960-1980
www.sgi.com

North America +1 800.800.7441
Latin America +55 11.5185.2860
Europe +44 118.912.7500
Japan +81 3.5488.1811
Asia Pacific +61 2.9448.1463

4105 [04.22.2008] **J15374**