The MVAPICH Project: Evolution and Sustainability of an Open Source Production Quality MPI Library for HPC*

Dhabaleswar K (DK) Panda[1], Karen Tomko[2], Karl Schulz[3] and Amitava Majumdar[4]

[1] *Department of Computer Science and Engineering, The Ohio State University*
{panda}@cse.ohio-state.edu

[2] *Ohio Supercomputing Center, Columbus, Ohio*
{ktomko}@osc.edu

[3] *Texas Advanced Computing Center, Austin, Texas*
{karl}@tacc.utexas.edu

[4] *San Diego Supercomputing Center, San Diego, California*
{majumdar}@sdsc.edu

***Abstract*—The MVAPICH project, based at The Ohio State University, has produced open-source, high performance, and production-ready MPI libraries for over a decade. It presents a good example of how academic research projects can be sustainable while producing quality software that benefits academia and industry alike. In this paper, we present details of the MVAPICH project including its development cycle, support structure, policies, use in production clusters, and impact on the HPC community.**

I. Overview of the MVAPICH Project

The MVAPICH (for MPI-1) and MVAPICH2 (for MPI-2 and MPI-3) open-source libraries [1] have been designed and developed during the last 12 years to take advantage of modern InfiniBand, 10-40GigE/ iWARP, and emerging RDMA over Converged Ethernet (RoCE) networking technology for HPC clusters. The OSU group has taken a lead role in designing these libraries with contributions from many other organizations and users worldwide. The work has been done through funding from the NSF, the DOE, and other companies.

A. Evolution

The MVAPICH Project started in 2001 with the introduction of InfiniBand open-standard networking technology. The MVAPICH open source MPI library with support for MPI-1 features and OpenMP was introduced to the HPC community at Supercomputing 2002. MVAPICH was updated to conform with the MPI-2 standard in 2004 with the release of the MVAPICH2 software stack. Since then, MVAPICH-2 has evolved to offer support to newer MPI standards such as MPI 2.1, 2.2, and 3.0. With the increasing adaptation of the MVAPICH2 library, the MVAPICH library, which only supported the MPI-1 standard, was phased out during 2009-2010 and encountered its End of Life (EOL) during June 2010. An enhanced version of the MVAPICH2 library, MVAPICH2-X, has been available since September 2012 as a technology preview to support hybrid MPI+PGAS programming models on modern HPC clusters. MVAPICH2-X has support for: 1) MPI only (with OpenMP), 2) UPC only, 3) OpenSHMEM only, and 4) MPI + UPC, MPI + OpenSHMEM, MPI + UPC + OpenSHMEM. The MVAPICH team also designed a comprehensive Micro-benchmark suite, called OSU Micro-benchmark suite (OMB) [2], that supports benchmarking for various MPI, CUDA, OpenSHMEM, and UPC features. Figure 1 shows the evolution of the various sub-projects of the MVAPICH team over time.

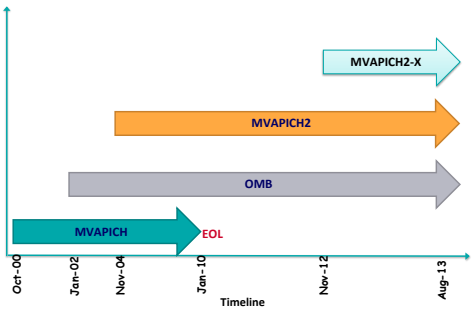

Figure 1. Evolution of the MVAPICH Project

B. MVAPICH Today

Currently, the MVAPICH, MVAPICH2, and MVAPICH2-X libraries are being used by more than 2,070 organizations worldwide (in 70 countries). A list of these organizations, registered in a voluntary manner at the MVAPICH project site, can be obtained by visiting http://mvapich.cse.ohio-state.edu/current_users/. As of September 2013, there have been more than 183,000 downloads of these libraries from the OSU web site alone. Figure 2 shows the growth of the number of downloads from the OSU site over time. These libraries are also available with the widely used OpenFabrics software stack [3], popular Linux distributions (such as Red Hat and SUSE), and several server and networking vendors. These libraries are also deployed on many production InfiniBand clusters that are on the TOP500 list. The June 2013 list includes the following systems: the 6th ranked Stampede system at TACC with 462,462 cores, the 19th ranked Pleiades system at NASA with 125,980 cores and the 21st ranked Tsubame2 system at Tokyo Institute of Technology with 73,278 cores.

II. The MVAPICH Research and Development Process

The typical release cycle of MVAPICH2 MPI library is shown in Figure 3. First, research challenges are identified and corresponding research is conducted. These research results are presented in various conferences, workshops, and journals. The best designs from these research results are then incorporated into the codebase.

*This research is supported in part by National Science Foundation grants #OCI-0926691, #OCI-1148371 and #CCF-1213084.

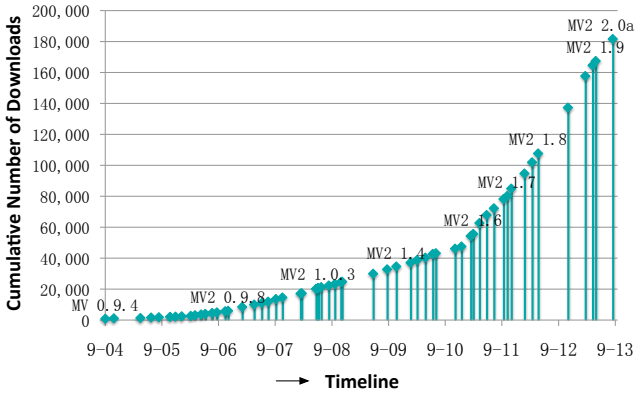

Figure 2. Download Statistics of the MVAPICH Project over Time

The MVAPICH/MVAPICH2 codebase incorporates many novel MPI-level designs (for performance, scalability, and fault-tolerance) carried out by the MVAPICH team during the last several years. These designs include: UD, XRC, UD-RC/XRC hybrid, SRQ, RDMA-fastpath, multirail, optimized intra-node communication, multi-core-aware collectives, topology-aware collectives, offloaded non-blocking collectives, optimized one-sided communication, quality of service (QoS), fault-tolerance, GPU-to-GPU communication [4], [5], optimizations for cluster with MICs [6], [7] etc. [8], [9]. In addition to the open-source code base, all of these designs have been published and disseminated by the MVAPICH team. Many of these designs have been adopted and incorporated into multiple MPI stacks (open-source and proprietary). Typically, new features incorporated into the MVAPICH/MVAPICH2 stack appear in other MPI stacks during the next 6-12 months period. In addition to the MPI-level internal designs, as indicated earlier, the MVAPICH team also introduced a comprehensive set of Benchmarks in 2002 to evaluate the performance of various MPI libraries. These benchmarks (called OSU Micro-Benchmarks, OMB) [2] have been widely adopted by HPC vendors and users during the last several years to compare the performance of different MPI libraries and HPC systems.

The collaboration between the MVAPICH team and supercomputing centers like OSC, TACC, and SDSC has enabled the co-design of MVAPICH2 MPI library, system software, and end-applications. It has also facilitated easy deployment of latest MVAPICH/MVAPICH2 library on latest XSEDE systems. Such a collaboration enables efficient design and development of system-level and library-level features. It also allows the immediate demonstration of their benefits in end-applications, thus guiding other application developers. For example, effort involving the MVAPICH team and SDSC has enabled efficient implementation of MPI one-sided semantics in MVAPICH2. It also made the redesign of AWP-ODC, a widely used seismic modeling code, demonstrating optimal application level overlap using MPI one-sided semantics possible. This effort has culminated in the application being selected as a Gordon Bell finalist at SC 2010 [10], [11]. Similarly, a collaborative effort involving the MVAPICH team and computer scientists at TACC has led to the development of new system software that enables users of InfiniBand based HPC systems to have access to underlying network topology information. This effort also resulted in publications that were accepted as Best Paper and Best Student Paper Finalists at SC 2012 [12]. Currently, three collaborative NSF-funded projects are on-going between OSU, OSC, TACC, and SDSC to design and develop various new features for MVAPICH2 and MVAPICH2-X libraries.

III. Testing, Supporting and Maintaining MVAPICH

As InfiniBand, iWARP, and RoCE technologies have matured, expectations from the community have gone up significantly with respect to the stability and the performance of the code-base. In the MVAPICH project, software libraries are tested rigorously using the in-house unit-testing framework. Using this framework, different code paths and parameters are tested and covered. The test framework consists of various tests such as MPI micro benchmarks (including the popular OSU Micro-benchmarks [2]), IMB test suite, Intel test suite, MPICH2 test suite, application kernels, and other HPC applications. Using these, the MPI library is tested and evaluated on multiple, state-of-the-art, multi-core computing platforms, network adapters, and accelerators. Each development patch is tested on these platforms. In addition to the functionality tests, performance regression tests are also done between releases and major patches to make sure performance is not degraded. Further, major bug-fixes and patches are tested on large-scale remote clusters.

The support for MVAPICH project is quite commendable. Special mailing lists are kept for MVAPICH2 release announcements and for discussing MVAPICH2 usage issues. On top of this, comprehensive user-guides are maintained from MVAPICH website, corresponding to each release version. In addition to these, performance numbers on various platforms and network adapters for major releases are linked to the project website. All of this information is publicly available.

IV. The Role of Open Source Communities and Feedback

The MVAPICH project has two open mailing lists: 1) mvapich@cse.ohio-state.edu, consisting of more than 3,100 registered users to receive announcements about periodic MVAPICH2 releases and 2) mvapich-discuss@cse.ohio-state.edu, with more than 550 registered users that provides an open forum to discuss all technical aspects of the MVAPICH2 library including performance, scalability, bug-fixes, compilations, designs, tuning, etc. The discussions

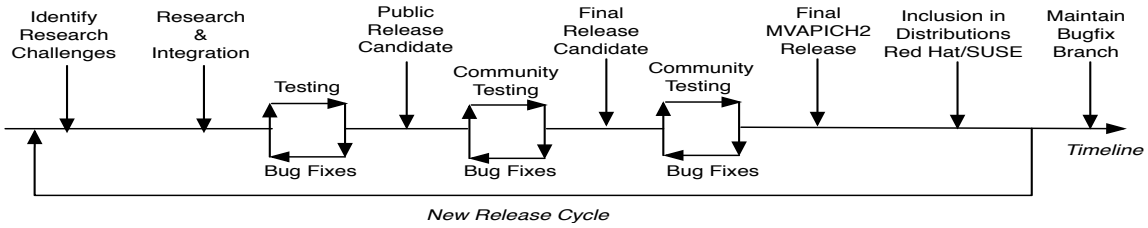

Figure 3. MVAPICH2 Research, Development and Release Process

on these mailing lists are also archived and are searchable, thus allowing a broader community to benefit from previous answers and suggestions. For example, there are often discussions on the mvapich-discuss public mailing list about installation issues, configuration and tunable parameters, performance issues, etc. Through discussions on the mailing-lists, the MVAPICH team and users help each other to resolve installation and configuration issues as well as getting optimal performance for their runtime environments and application needs. Discussions on design methodologies and their impacts on applications are also carried out on the existing mailing lists.

Over the years, the feedback received from end-users and HPC sites has become a major source of valuable community input for the MVAPICH project that is used to further enhance the software stack. For example, based on feedback from users, the OSU team designed more scalable methods and environmental control for MPI job-startup and tear-down procedures. Furthermore, external users have not only discovered bugs and made feature requests, but in many cases, they have contributed patches with fixes and enhancements, which helps to improve the quality of the library and increase portability. The requirements from collaborators, vendors, and users with large deployments (such as the recent Stampede deployment) are taken into consideration through mailing list discussion and even audio conferences.

V. Policy Issues in MVAPICH

The MVAPICH2 MPI library is made freely available to the public under open-source BSD licensing. The software sharing and dissemination plan for the library are guided by the following rules:

1) Freely available to all institutions and MPI communities
2) Availability with open-source BSD licensing permitting dissemination and commercialization
3) Engagement with MPI communities for continuous improvements based on feedback and integration of patches and enhancements to subsequent releases
4) Dissemination of the associated benchmarks and performance results

Consequently, the MVAPICH2 project is carried out via open-source activities and conforms to NSF policies on the dissemination and sharing of research results.

Through the BSD licensing mechanism, the MVAPICH2 library is also available from the software stacks of multiple vendors and Linux distributions (such as Red Hat and SUSE). In addition to the open-source code base, all designs incorporated in this software stack have been published in international journals and presented at international conferences and workshops.

VI. The Use of MVAPICH in HPC Community

The MVAPICH2 open-source library, based on many of the publications and research designs by the OSU team, has been a significant component in the InfiniBand ecosystem helping contribute to the rapid growth and adoption of InfiniBand in the HPC community. As introduced in Section I, the library is used by a large number of organizations worldwide. Not only is it available directly through the OSU website, but it is also embedded within the software stacks of many server and networking vendors. The library is empowering many production InfiniBand clusters in the TOP500 list. These clusters and other systems in the NSF XSEDE environment are being used by a large number of scientists and engineers daily.

As highlighted in Section I, the MVAPICH project has been widely adopted within the HPC community. MVAPICH2 is installed on many of the large supercomputing systems with InfiniBand. Each of these systems are typically used by many users. For example, after five years of formal production, the TACC Ranger system that recently retired in February 2013 hosted researchers from over 350 institutions in support of 2,244 research projects in which more than 4,000 users completed 2.69 million jobs, consuming 2.1 billion processor core hours. The vast majority of this usage was by users leveraging the MVAPICH project and the multiple enhancements that were introduced into the library during the system's lifespan. Worldwide, it is estimated that the MVAPICH and MVAPICH2 libraries have combined to benefit hundreds of thousands of HEC users. Many systems on the NSF XSEDE program use MVAPICH2 as the default MPI library indicating extensive use by the NSF research and education community. These include Gordon and Trestles at SDSC, Stampede and Lonestar at TACC, Keeneland at NICS, etc. Note that while the maintenance and installation of the library is handled by professional staff on

these high-end production resources, the MVAPICH2 MPI library can also be installed and used by a non-root user on any Linux cluster with InfiniBand. The voluntary registration available on the website clearly shows the wider acceptance of MVAPICH2. The open source nature of the MVAPICH2 MPI library has enabled several research efforts spanning MPI runtime-level designs, tools development, application development, scalability studies, and more. Work utilizing MVAPICH2 is published and presented constantly at many of the major conferences in the HPC domain. MVAPICH2 is the first MPI library to support MPI communication from GPU memory. It simplifies the porting of MPI applications to GPU clusters and is widely used on GPU clusters with InfiniBand, including Keeneland at NICS. Support for emerging Many-Integrated Core (MIC) is currently being worked out and will be available to the community in the near future.

VII. Broader Impact of MVAPICH on the Science Community

Through co-designing a range of scientific applications, the research related to the MVAPICH project becomes a collaborative and synergistic activity between computer scientists with different expertise (communication libraries, compilers, and applications). The investigators are actively involved in designing communication libraries, compilers support, software environments, and applications for next-generation petascale and exascale systems. The researchers are also involved with the MPI forum where discussions are currently taking place to define the next-generation MPI standard. Some of them are consulting members of the International Open Fabrics Alliance (OFA) [3], responsible for driving next generation networking and I/O technologies for clusters, data-centers, and clouds. The MVAPICH User Group meeting conducted in August 2013 allowed users to learn more about MVAPICH projects, features, and also get a hands-on session with MVAPICH developers. Furthermore, MVAPICH features and results are presented in conferences such as Supercomputing (SC), International Supercomputing (ISC), and HPC Advisory Council meetings.

The students and post-doctoral researchers involved in MVAPICH2 project are trained in designing and developing runtime environments for next generation architectures combined with networking, compiler support, systems software, and applications in an integrated manner. Some of the proposed research and development directions have been used as projects in graduate classes. The research part and the designs in MVAPICH project have been published in world-class ranked conferences and journals.

VIII. Conclusions

In this paper, we presented evolution of the popular MVAPICH project and its high performance MPI and PGAS libraries. We discussed several aspects of the project including the research and development process, release cycle, role of open source communities, use on HPC clusters, and impact on the community. Through this, we provided an example for how academic projects can be sustained while providing practical solutions that are useful for the HPC community. The past activities have been supported through funding from the NSF, the DOE and various companies. The MVAPICH team is looking forward to sustaining this momentum for designing MPI and MPI+PGAS libraries and runtimes for the emerging Exascale systems.

References

[1] Network-Based Computing Laboratory, "MVAPICH: MPI over InfiniBand, 10GigE/iWARP and RoCE," http://mvapich.cse.ohio-state.edu/.

[2] OSU Micro-benchmarks, http://mvapich.cse.ohio-state.edu/benchmarks/.

[3] "OpenFabrics Alliance," http://www.openfabrics.org/.

[4] H. Wang, S. Potluri, M. Luo, A. Singh, S. Sur, and D. K. Panda, "MVAPICH2-GPU: Optimized GPU to GPU Communication for InfiniBand Clusters," in *Int'l Supercomputing Conference (ISC)*, 2011.

[5] S. Potluri, K. Hamidouche, A. Venkatesh, D. Bureddy and D. K. Panda, "Efficient Inter-node MPI Communication using GPUDirect RDMA for InfiniBand Clusters with NVIDIA GPUs." in *International Conference on Parallel Processing (ICPP13)*, October 2013.

[6] S. Potluri, D. Bureddy, K. Hamidouche, A. Venkatesh, K. Kandalla, H. Subramoni and D. K. Panda, "MVAPICH-PRISM: A Proxy-based Communication Framework using InfiniBand and SCIF for Intel MIC Clusters," in *International Conference dor High Performance Computing, Networking, Storage and Analysis (SC13)*.

[7] K. Hamidouche, S. Potluri, H. Subramoni, K. Kandalla, and D. K. Panda, " MIC-RO: Enabling Efficient Remote Offload on Heterogeneous Many Integrated Core (MIC) Clusters with InfiniBand," in *International Conference on Supercomputing (ICS)*, 2013.

[8] Network-Based Computing Laboratory, "Papers, Technical Reports, M.S Thesis and Ph.D Dissertations," http://nowlab.cse.ohio-state.edu/publications/.

[9] The MVAPICH Team, "Publications," https://mvapich.cse.ohio-state.edu/publications/.

[10] Y. Cui and K.B. Olsen and T. H. Jordan and K. Lee and J. Zhou and P. Small and D. Roten and G. Ely and D.K. Panda and A. Chourasia and J. Levesque and S. M. Day and P. Maechling, "Scalable Earthquake Simulation on Petascale Supercomputers," in *SuperComputing (SC)*, Nov 2010.

[11] S. Potluri, P. Lai, K. Tomko, S. Sur, Y. Cui, M. Tatineni, K. Schulz, W. Barth, A. Majumdar, and D. K. Panda, "Quantifying Performance Benefits of Overlap using MPI-2 in a Seismic Modeling Application," in *24th International Conference on Supercomputing (ICS)*, 2010.

[12] H. Subramoni, S. Potluri, K. Kandalla, B. Barth, J. Vienne, J. Keasler, K. Tomko, K. Schulz, A.Moody, and D. K. Panda, "Design of a Scalable InfiniBand Topology Service to Enable Network-Topology-Aware Placement of Processes," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2012.