SOFTWARE

PRODUCT BRIEF

Mellanox HPC-X™ Scalable Software Toolkit

To meet the needs of scientific research and engineering simulations, supercomputers are growing at an unrelenting rate. As supercomputers increase in size from mere thousands to hundreds-of-thousands of processor cores, new performance and scalability challenges have emerged.

Mellanox HPC-X

The Mellanox HPC-X Scalable Software Toolkit is a comprehensive tool-suite for high performance computing environments providing enhancements to significantly increase the scalability and performance of message communications in the network.

HPC-X enables you to rapidly deploy and deliver maximum application performance without the complexity and costs of licensed third-party tools and libraries.

Mellanox HPC-X provides acceleration packages to accelerate both the performance and scalability of popular MPI and SHMEM/PGAS libraries. These packages, including MXM (Mellanox Messaging) which accelerates the underlying send/receive (or put/get) messages, and FCA (Fabric Collectives Accelerations) that accelerates the underlying collective operations used by the MPI/PGAS languages.

This full-featured, tested and packaged version of HPC software enables MPI, SHMEM and PGAS programming languages to scale to extremely large clusters by improving on memory and latency related efficiencies. It also assures that the communication libraries are fully optimized for the underlying Mellanox interconnect solutions. HPC-X provides full support for 3rd party interconnect solutions that are compatible to the Ethernet and InfiniBand standards.

Parallel Programming Models

There are two types of models for parallel programming. The first is the shared memory model, in which all processes interact through a globally addressable memory space. The other is a distributed memory model, in which each processor has its own memory, and interaction with another processors memory is done though message communication. The PGAS model, or Partitioned Global Address Space, uses a combination of these two methods, in which each process has access to its own private memory, and also to shared variables that make up the global memory space.

Mellanox ScalableUPC and the Berkeley UPC parallel programming language library

Unified Parallel C (UPC) is an extension of the C programming language designed for high performance computing on large-scale parallel systems. The language provides a uniform programming model for shared and distributed memory hardware. The processor memory has a single shared, partitioned address space, where variables may be directly read and written by any processor, but each variable is physically associated with a single processor. UPC uses a Single Program Multiple Data (SPMD) model of computation in which the amount of parallelism is fixed at program startup time, typically with a single thread of execution per processor.

Mellanox ScalableUPC is based on the Berkeley Unified Parallel C project. Berkeley UPC library includes an underlying communication conduit called GASNET, which works over the OpenFabrics RDMA for Linux stack (OFED™) Mellanox has

HIGHLIGHTS

BENEFITS

- Increase CPU availability and efficiency for increased application performance
- Seamless integration with Mellanox OFED 2.1 or later
- Provides the best performance with the underlying interconnect hardware
- Minimum tuning required

KEY FEATURES

- Offload collectives communication from MPI process onto Mellanox interconnect hardware
- Maximize application performance with underlying hardware architecture
- Fully optimized for Mellanox InfiniBand and VPI interconnect solutions
- Supports any interconnect solution based on Ethernet and InfiniBand standards
- Increase application scalability and resource efficiency
- Multiple transport support including RC, DC and UD
- Intra-node shared memory communication
- Receive side tag matching
- Native support for MPI-3

optimized this GASNET layer with the inclusion of their Mellanox Messaging libraries (MXM) as well as Mellanox Fabric Collective Accelerations (FCA), providing an unprecedented level of scalability for UPC programs running over InfiniBand.

Mellanox ScalableSHMEM

The SHMEM programming library is a one-side communications library that supports a unique set of parallel programming features including point-to-point and collective routines, synchronizations, atomic operations, and a shared memory paradigm used between the processes of a parallel programming application.

SHMEM (SHared MEMory), uses the PGAS model to allow processes to globally share variables by allowing each process to see the same variable name, but each process keeps its own copy of the variable. Modification to another process address space is then accomplished using put/get (or write/read) semantics.

The ability of put/get operations, or one-sided communication, is one of the major differences between SHMEM and MPI (Message Passing Interface) which only uses two-sided, send/ receive semantics.

Mellanox ScalableMPI - Message Passing Interface based on Open MPI

Message Passing Interface (MPI) is a standardized, language-independent and portable message-passing system. Open MPI is an open source implementation of MPI, the industry-standard specification for writing message-passing programs.

Mellanox ScalableMPI is a high performance implementation of Open MPI optimized to take advantage of the Mellanox acceleration capabilities. Mellanox ScalableMPI is recommended for users who need a stable, tested and packaged bundle of Open MPI.

Mellanox MXM

Mellanox Messaging Accelerator (MXM) provides enhancements to parallel communication libraries by fully utilizing the underlying networking infrastructure provided by Mellanox HCA/switch hardware. This includes a variety of enhancements that take advantage of Mellanox networking hardware including:

- Multiple transport support including RC, DC and UD
- Proper management of HCA resources and memory structures
- Efficient memory registration
- One-sided communication semantics
- Connection management
- Receive side tag matching
- Intra-node shared memory communication

These enhancements significantly increase the scalability and performance of message communications in the network, alleviating bottlenecks within the parallel communication libraries.

Mellanox FCA

FCA is a MPI-integrated software package that utilizes CORE-Direct technology for implementing the MPI collective communications. FCA can be used with all major commercial and open-source MPI solutions that exist and being used for high-performance applications. FCA with CORE-Direct technology accelerates the MPI collectives runtime, increases the CPU availability to the application and allows overlap of communications and computations with collective operations. FCA allows for efficient collectives communication flow optimized to job and topology. It also contains support to build runtime configurable hierarchical collectives (HCOL) and supports multiple optimizations within a single collective algorithm.

Mellanox IPM (Integrated Performance Monitoring)

IPM is a portable profiling infrastructure for parallel codes. It provides a low-overhead performance profile of the performance aspects and resource utilization in a parallel program for communication, computation, and IO.

IPM has extremely low overhead, is scalable and easy to use requiring no source code modification. IPM brings together several types of information important to developers and users of parallel HPC codes. IPM gathers information with minimal impact on the running code, maintaining a small fixed memory footprint and using minimal amounts of CPU.

The monitors that IPM currently integrates are:

- MPI: communication topology and statistics for each MPI call and buffer size.
- HPM: PAPI (many) or PMAPI (AIX) performance events.
- Memory: wallclock, user and system timings.
- Switch: Communication volume and packet loss.
- File I/O: Data written and read to disk

Industry standard utilities

HPC-X Scalable Toolkit also includes numerous pre-compiled HPC packages and utilities needed for fine tuning your HPC environment including numerous MPI tests, the industry standard InfiniBand verbs profiler (libibprof) and KNEM, a Linux kernel module enabling high-performance intra-node MPI communication for larger messages.

HPC-X for Higher Return-on-Investment

Mellanox InfiniBand interconnect solutions for server and storage systems provide the highest performance and efficiency for HPC applications. HPC-X helps to further accelerate the application performance, increase the CPU efficiency and future proof the system architecture.

The figures below show the benefits of HPC-X, even at lower scale, and of course, the larger the system the effects that HPC-X will demonstrate will be even greater.

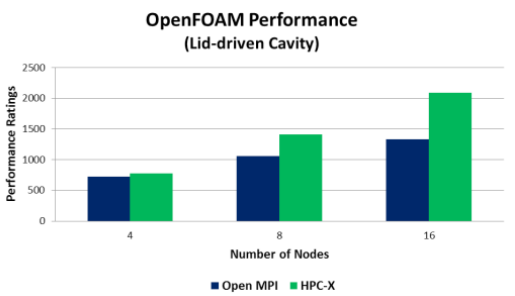

Figure 1: *Gain 58% improvement in productivity with OpenFoam with as few as 16 nodes with HPC-X*

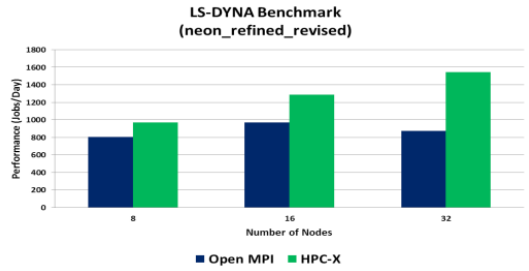

Figure 2: *HPC-X provides additional 76% productivity with LS-DYNA*

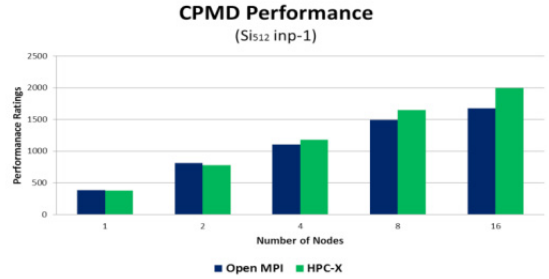

Figure 3: *Gain an additional 20% productivity with CPMD with as few as 16 nodes.*

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com

15-3937PB Rev 1.3