OFA Workshop 2020 Session Abstracts

Amazon Elastic Fabric Adapter: A Year in Review

Raghunath Raja Chandrasekar, Amazon Web Services

AWS launched Elastic Fabric Adapter (EFA), a high-performance network interface for EC2 instances, in early 2019. Since the initial launch, the hardware, software stack, and eco-system have evolved considerably to better meet the performance and functional needs of HPC and ML customers running demanding applications on EC2. This talk will present an overview of the EFA architecture and the underlying Scalable Reliable Datagrams (SRD) protocol, review the features and enhancements that have been developed since the initial launch, and present some case studies with real customer workloads that leverage EFA to satisfy their demanding networking needs. The talk will also provide a brief glimpse toward the directions we are investing future development efforts in.

An FPGA platform for Reconfigurable Heterogeneous HPC and Cloud Computing

Bernard Metzler, IBM Research Zurich; Francois Abel, IBM Research Zurich; Burkhard Ringlein, IBM Research Zurich; Beat Weiss, IBM Research Zurich; Christopher Hagleitner, IBM Research Zurich

Energy and compute efficient devices such as GPUs and Field-Programmable Gate Arrays (FPGAs) are entering modern data centers (DCs) and HPC clusters to address the end of Dennard scaling and the wind down of Moore's law. While these heterogeneous devices seem to be a natural extension for current compute servers, their traditional bus-attachment limits the number of accelerators that can be deployed per server. With cloudFPGA, we introduce a new platform that decouples the reconfigurable accelerators from the CPU of the server by connecting the FPGAs directly to the DC network. This approach turns the FPGAs into a disaggregated standalone computing resources that can be deployed at large scale in DCs. Each FPGA implements an integrated 10GbE network controller interface (NIC) with TCP/IP offload engine. This deployment of network-attached FPGAs is particularly cost- and energy-efficient as the number of spread-out FPGAs becomes independent of the number of servers. During the presentation, we describe the cloudFPGA platform which integrates up to 1024 FPGAs per DC rack. We focus on the networking aspects of the architecture, the development toolchain required to seamlessly migrate today's workload, and we discuss the potential applicability of standalone network-attached FPGAs for different HPC and Cloud Computing workloads.

An Update on CXL Specification Advancements

Jim Pappas, CXL Consortium

Compute Express Link™ (CXL) is a high-speed CPU-to-Device and CPU-to-Memory interconnect designed to accelerate next-generation data center performance. This presentation will provide an update on the latest advancements in CXL specification development, its use cases and industry differentiators. CXL enables a high-speed, efficient interconnect between the CPU and platform enhancements and workload accelerators. Attendees will learn how CXL technology: Allows resource sharing for higher performance Reduces complexity and lowers overall system cost Permits users to focus on target workloads as opposed to redundant memory management Builds upon PCI Express® infrastructure Supports new use cases for caching devices and accelerators, accelerators with memory and memory buffers. The CXL Consortium has released the CXL 1.1 Specification and the next generation of the spec is currently under development. Consortium members can contribute to spec development and help shape the ecosystem

Deep Dive into the OFA's upcoming cluster

Doug Ledford, Red Hat, Inc.

As part of reimagining the OFA's interop testing plan, the OFA landed on building out a cluster, using existing software already open sourced, that can be so much more than just a twice a year interop testing program. It will be a full time available cluster with the ability for doing development work, on demand testing, logo testing, certification, etc. The cluster will be built out using a combination of software available today: Beaker (the lab controller), a collection of beaker tests contributed by Red Hat (with the intent that this collection will take on an open source life of its own with collaborative development on the upstream mailing list helping to improve the tests), and connection into gitlab runners for doing monitoring of upstream git repositories with automated continuous integration testing on those repositories. This presentation will introduce people to these technologies, what they look like, how they are used, etc.

Designing a Deep-Learning Aware MPI Library: An MVAPICH2 Approach

Hari Subramoni, The Ohio State University; Ammar Ahmad Awan, The Ohio State University; Jahanzeb Maqbool Hashmi, The Ohio State University; Ching-Hsiang Chu, The Ohio State University; Dhabaleswar Panda, The Ohio State University

Multi-core CPUs and GPUs are seeing widespread adoption in current and next-generation HPC systems. In addition to scientific applications, we are witnessing an increasing usage of Deep Learning (DL) models and applications in HPC environments. DL has taken over multiple application domains because of the rise in computational capabilities, availability of large-scale datasets, and easy to use DL frameworks like Tensorflow, Caffe and PyTorch. This has led to a fresh set of challenges that must be tackled to extract the best performance. In this work, we present advanced designs to tackle such challenges in the MVAPICH2 MPI library on the modern HPC systems. We focus on performance of Allreduce, which is essential for scalable DNN training on multiple CPU and GPU nodes. We provide an in-depth performance characterization of state-of-the-art DNNs like ResNet(s) and Inception on many-core CPUs on TACC Frontera (#5 on Top500) and Stampede2 clusters. We utilize optimized versions of DL frameworks to ensure best performance. We report up to 125x speedup for ResNet-50 on 128 nodes. On Frontera, we report 250,000 images/second for ResNet-50 training on 2,048 nodes. Similarly, our GPU-optimized Allreduce designs provide scalable training of ResNet-50 on Summit (#1 on Top500) up to 1,536 GPUs.

Distributed Asynchronous Object Storage (DAOS)

Kenneth Cain, Intel, Corp.; Johann Lombardi, Intel, Corp.; Alexander Oganezov, Intel, Corp.

The Intel team developing the Distributed Asynchronous Object Storage (DAOS) system proposes to present and engage in discussion with the OFA community of administrators, developers, and technology providers. DAOS architecture and implementation will be presented by exploring its interaction with fabric hardware/software (e.g., libfabric) and its role in delivering (with storage class persistent memory and NVMe SSDs) very high-performance scale-out object storage. Feedback from building this service using libfabric, and potential opportunities to further leverage fabric will be discussed. Areas of potential interest to this audience may include: increasing diversity of I/O patterns (e.g., large volumes of random reads/writes), checkpoint/restart snapshots,producer/consumer flows, small I/O and metadata handling using fabric and persistent memory, bulk data handling with fabric, persistent memory and SSDs. collective communication for scalability and efficient dissemination/retrieval of metadata, dynamic contraction/expansion of storage servers, data scaling for performance, data replication, erasure coding and fault domain awareness for resilience, metadata service resilience and protocols/fabric communications, user-space design for storage (PMDK, SPDK) and fabric (OFI) interactions, asynchronous data and metadata operations and progress, online data rebuild when storage node/target is lost or removed ; rebalancing when added. Integration with Lustre and unified namespace. POSIX filesystem emulation, MPI-I/O and HDF5 over DAOS.

Enhancing MPI Communication in MVAPICH2 using Hardware Tag Matching

Dhabaleswar Panda, The Ohio State University; Hari Subramoni, The Ohio State University; Mohammadreza Bayatpour, The Ohio State University

Message Passing Interface (MPI) standard uses (source rank, tag, and communicator id) to properly place the incoming data into the application receive buffer. In the state-of-the-art MPI libraries, this operation is either being performed by the main thread or a separate communication progress thread. Mellanox InfiniBand ConnectX-5 network architecture has introduced a feature to offload the Tag Matching and communication progress from host to InfiniBand network card. When a message arrives, HCA searches through the list of offloaded tags and upon finding a tag that matches with the arrived message's tag, HCA directly places the arrived message into the application buffer. In this talk, experiences with the MVAPICH2 MPI library to exploit the hardware Tag Matching feature will be presented. Various aspects and challenges of leveraging this feature in an MPI library will be discussed. Characterization of hardware Tag Matching using

different benchmarks and providing guidelines for the application developers to develop Hardware Tag Matching-aware applications to maximize their usage of this feature will be presented. Experimental results showing up to 35% improvement in point-to-point latency as well as up to 1.8X improvement in latency of collective operations on 1,024 nodes of the TACC Frontera system will be presented.

Enhancing OFI for invoking acceleration capabilities on an integrated networking/accelerator FPGA platform (COPA)

Venkata Krishnan, Intel, Corp.; Olivier Serres, Intel, Corp.; Michael Blocksome, Intel, Corp.

The OpenFabrics Alliance's Interop and Logo Program provides a valuable service to both Alliance members and the community as a whole by providing a venue for independent testing of device and software interoperability for high performance networking. Over the past year, we have launched a companion program to incorporate Linux Distro testing. In this session, we will explore several possible paths toward 'rationalizing' these two programs in such a way that additional synergy is created, with the goal of increasing the value of both programs. The session begins with a brief discussion of the path that led us to this point to be used as the jumping off point for exploring several possible paths forward. The results of this session are expected to result in specific proposals to the OFA for advancing the structure of the existing Interop program in order to increase its overall value. This session is intended to work together with a companion BoF to encourage the community to reach rapid consensus on selecting a path forward.

Gen-Z: An Open Memory Fabric for Future Data Processing Needs

Kurtis Bowman, Gen-Z Consortium

With the evolution of advanced workloads like AI and machine learning, the need for new memory system architectures that include memory expansion, memory sharing, and disaggregated memory continues to increase. Gen-Z is a new fabric architecture that utilizes memory-semantic communications to move data between memories on different components with minimal overhead. Gen-Z fabric is highly scalable and composable, allowing day-to-day reconfiguration and reallocation of resources desired by hyperscale data center operators. Gen-Z also provides advanced fabric security built around access and region keys. To enable its full functionality, Gen-Z memory fabric involves switching, routing, security, zoning, access control and fabric management. The core properties of Gen-Z fabric consist of scalable and provisioned memory infrastructure; shared memory for data processing; connectivity of processors, GPUs, accelerators, and optimized engines; next-generation DRAM, FLASH and storage class memory; and enabling persistent memory. The Gen-Z Consortium is committed to a robust, open ecosystem, and has released a suite of Gen-Z specifications to address the unique needs of the industry. The Gen-Z Fabric Management Specification will be available for public download on the Consortium's website soon. In this presentation, attendees will learn about Gen-Z Fabric Management and its numerous benefits to the high-performance fabrics industry.

HDR IB with Advanced CC and AR and HPC and AI in the Cloud

Tzahi Oved, Mellanox Technologies; Alex Rosenbaum, Mellanox Technologies; Ariel Almog, Mellanox Technologies

Efficient network BW utilization and deterministic low latency with low latency jitter is very important for many large scale HPC and AI applications. In this session we will present the latest enhancements to network congestion avoidance and congestion control using the combination of Adaptive Routing and enhanced Congestion Control. We will show the effect of enabling the two on network performance. The increased compute power using accelerators and faster processing units and the increase in AI hungur for data and compute power pushes the requirements from the interconnect beyond low latency and high BW. To achieve the required network performance and scaling efficiency, a new approach is required. In this talk we will show how in-network computing can not only accelerate the time to complete complex operation can be reduced 10x but also the processing engine in freed enabling efficient overlap between computation and communication. We will present SHARP In-network compute technology and show how it accelerates HPC and AI applications.

How do We Debug?

Tzahi Oved, Mellanox Technologies; Alex Rosenbaum, Mellanox Technologies; Ariel Almog, Mellanox Technologies

As RDMA deployment becomes wider, the challenge of keep it working, tuning its performance and find failure point is becoming more complex. In this session we will go over debug flow and will discuss the tools (both existing and futuristic) that are being used for debug and performance tuning. We will focus on additional functionality that is being

pushed by Mellanox for RDMA tool (part of iproute2 package) to enhance the visibility and debuggability of RDMA network. We will discuss the challenges of capturing offloaded RDMA traffic for debug.

Lessons Learned from Implementing the Elastic Fabric Adapter Libfabric Provider

Robert Wespetal, Amazon

This talk will discuss some of the implementation details of the Elastic Fabric Adapter (EFA) Libfabric provider and the lessons we learned implementing a performant Libfabric provider. A high level overview of the memory registration and data path components of the EFA provider will be discussed including improving memory registration costs and optimizing the data path for various message types and sizes. After discussing those components, some best practices for configuring the EFA provider will be provided. Finally, we will provide some insight on the challenges faced and approaches we have taken to reduce software overhead in this provider via shared Libfabric components like the utility code and performance related tools in Libfabric.

Libfabric Intranode Device Support

Alexia Ingerson, Intel, Corp.

With device memory usage becoming more common, ongoing work is being done to add device support to libfabric, the shared memory provider being one target area of where to add this support. The provider currently uses shared memory for local communication but needs extensions to support copying to and from device memory. This talk will provide an overview of the current design of the provider and how it will be adapted to support a variety of device types and transfers, including additions and modifications of its current protocols.

Lustre Network Multi-Rail Feature Set

Amir Shehata, Whamcloud DDN

Lustre Networking (LNet) can have multiple interfaces available for sending messages. These interfaces can be of the same or different types (IE eth, mlx, opa). It is important for Lustre to use all the interfaces to increase both its bandwidth and resiliency. To effectively use the interfaces in such a manner an LNet Multi-Rail and Health features were implemented. The Multi-Rail feature allows the use of all available interfaces irregardless of the underlying wire protocol. The Health feature adds resiliency such that the healthiest interfaces are always used. A single Lustre network must have homogeneous interfaces, example all Mellanox or all OPA. However, if two Lustre peers can be reached on both networks, there is no reason not to use both to communicate between the peers. However, if one of the networks becomes unreachable, Lustre traffic should switch to the other network without dropping messages. This is achieved by maintaining the health status of each of the interfaces and resending messages over other healthier interfaces. A set of parameters such as timeout, retry count and sensitivity along with traffic control policies provide fine-tuned traffic control to allow admins to configure Lustre in a way which fits their network.

MVAPICH Touches the Cloud: New Frontiers for MPI in High Performance Clouds

Hari Subramoni, The Ohio State University; Shulei Xu, The Ohio State University; Seyedeh Mahdieh Ghazimirsaeed, The Ohio State University; Dhabaleswar Panda, The Ohio State University

Cloud Virtual Machines (VM) and instances have become a new trend of HPC and will significantly change its future. This talk will introduce recent contributions from the MVAPICH team to provide support and new designs for high performance computing (HPC) cloud platforms, including Amazon AWS and Microsoft Azure. Amazon has recently announced a new network interface named Elastic Fabric Adapter (EFA) targeted towards tightly coupled HPC workloads. First, this talk takes a high-level view of the features, capabilities, and performance of the adapter. Next, it explores how EFA's transport models such as UD (Unreliable Datagram) and SRD (Scalable Reliable Datagram) impact the design of high-performance MPI libraries. A new zero-copy design in MVAPICH2 will be introduced as a transfer mechanism over unreliable and order-less channels. Microsoft has also recently announced Azure HB and HC series VM targeted with InfiniBand for HPC workloads. The second part of this talk will present optimizations in the MVAPICH2 library for both Azure HB and HC instances. The availability of a One click easy and quick deployment scheme, with help from Microsoft Azure team, will be presented. The talk will conclude with an in-depth performance evaluation results of the new design with multiple benchmarks.

oneAPI, oneCCL and OFI: Path to Heterogeneous Architecture Programming with Scalable Collective Communications

Sayantan Sur, Intel, Corp.

Hetergeneous architectures are becoming more prevalent as a mainstream compute resource. The challenges of having disparate programming models, tools and workflows of development is difficult especially as there are no common standards or broad community efforts to get to a solution. oneAPI is an industry initiative that Intel is driving along with the community to develop standards-based programming models along with libraries that focus on delivering cross architecture support. Within oneAPI, the oneCCL Collective Communication Library is designed to offer easy integration of high-performance collective communication patterns into popular Deep Learning Frameworks. In this talk, we will describe the how the combination of oneAPI, oneCCL and OFI provide a pathway towards portable heterogenous architecture programming with performance.

Proposed Intellectual Property Rights (IPR) Policy for the OFA

Jim Ryan, OpenFabrics Alliance

The is rewriting its Bylaws and IPR policy in separate but related projects. This session will review a set of specific IPR policies covering all of the OFA's activities, projects and related properties. The objective is to solicit feedback from attendees as to what is required so the project can reach completion.

RDMA Updates and New Capabilities

Alex Rosenbaum, Mellanox Technologies; Ariel Almog, Mellanox Technologies; Tzahi Oved, Mellanox Technologies

In this session I would like to present some new features coming up in RDMA subsystem. Such as allowing multi-process verbs collaboration based on shared PD design, new send operations and memory mapping API's. In addition, discuss some RDMA_CM updates for connection establishment rate improvement. RoCE CM MAD's and traffic QP's path alignments.

RDMA with GPU Memory via DMA-Buf

Jianxin Xiong, Intel, Corp.

Discrete GPUs have been widely used in systems for high performance data parallel computations. Scale-out configuration of such systems often include RDMA capable NICs to provide high bandwidth, low latency inter-node communication. Over the PCIe bus, the GPU appears as peer device of the NIC and extra steps are needed to set up GPU memory for RDMA operations. Proprietary solutions such as Peer-Direct from Mellanox have existed for a while for this purpose. However, direct use of GPU memory in RDMA operations (A.K.A. GPU Direct RDMA) is still unsupported by upstream RDMA drivers. Dma-buf is a standard mechanism in Linux kernel for sharing buffers for DMA access across different device drivers and subsystems. In this talk, a prototype is presented that utilizes dma-buf to enable peer-to-peer DMA between the NIC and GPU memory. The required changes in the kernel RDMA driver, user space RDMA core libraries, as well as Open Fabric Interface library (libfabric) are discussed in detail. The goal is to provide a non-proprietary approach to enable direct RDMA to/from GPU memory.

Remote persistent memory access API - the second approach

Tomasz Gromadzki, Intel, Corp.; Jan Michalski, Intel, Corp.

PMDK team started working on combining RDMA with Persistent Memory many years ago because we've recognized the potential this combination had. Our initial approach, as seen in librpmem, was very conservative and focused on one use case: synchronous replication of data for libpmemobj pools. The result is that librpmem is great at that one specific workload, but we've since discovered that the APIs exposed by the library can be limiting for other things. That's the primary reason for why we've started on the path towards a new library, partially based on librpmem's implementation, that will have a generic enough interface to be useful for a wide variety of other workloads. A new API has been defined, a prototype has been created and is now reviewing with potential users. The presentation is going to demonstrate API itself but also reveals the major facts behind key design decisions.

SparkUCX – RDMA acceleration plugin for Spark

Alex Rosenbaum, Mellanox Technologies; Ariel Almog, Mellanox Technologies; Tzahi Oved, Mellanox Technologies

Many HPC clusters are shared with Big Data crunching frameworks for better resource utilization. It makes a lot of sense to use that RDMA enabled network to accelerate the Big Data framework applications, such as Spark or Hadoop. In this session I would like to present the Java binding developer in the UCX open source project. We will review the UCX architecture and see how this new Java UCX API's allows fast and simple Java application integration into the RDMA subsystem. I will present the SparkUCX shuffle plugin we created, based on the jUCX API's and show the performance improvement it provides.

SPDK based user space NVMe over TCP transport solution

Ziye Yang, Intel, Corp.

SPDK (storage performance development kit, http://spdk.io) already provides accelerated user space NVMe over Fabric (NVMe-oF) target, which provides much better performance compared with kernel solution on RDMA transport. And it is adopted by many cloud storage service vendors in China. In November 2018, NVM express releases the new spec of TCP transport for NVMe over fabrics. In this talk, we would like to introduce the design, implementation and development plan of NVMe-oF TCP transport in SPDK. Currently, SPDK implements both TCP transport in host and target side and can be tested against Linux kernel solution with good interoperability. Besides, some experiments results will be presented to demonstrate the performance and scalability of SPDK's NVMe-oF TCP transport implementation. Moreover, we will introduce some techniques for the further performance improvement of SPDK's solution, e.g., (1) leveraging user space TCP stack (e.g., VPP/Seastar + DPDK) to replace the kernel TCP stack; (2) leveraging some features of hardware such as ADQ on Intel's E810 NIC. Compared with kernel solution, SPDK based NVMe-oF solution has much better per CPU core performance in different aspects (e.g., IOPS, latency).

Status of OpenFabrics Interfaces (OFI) Support in MPICH

Yanfei Guo, Argonne National Laboratory

This session will give the audience an update on the OFI integration in MPICH. MPICH underwent a large redesign effort (CH4) in order to better support high-level network APIs such as OFI. We will show the benefits realized with this design, as well as ongoing work to utilize more aspects of the API and underlying functionality. This talk has a special focus on how MPICH is using Libfabric for GPU support and the development updates on GPU fallback path in Libfabric.

The New OFA Testing Program

TBD

The group previously known as Interop Working Group (IWG) has been redesigned into a Testing Work Group which is going to include RDMA hardware and software testing, contribute to standardized test plans and serve the RDMA Linux kernel community by providing regular testing coverage of the RDMA stack for new kernel RCs and enabling quick bug reporting. It is also going to present opportunities to a broader audience to get hands-on experience with RDMA technologies. This presentation aims to outline the activities of the new Testing Program and establish the general expectations from its deliverables. OFA is now looking for members to join the new Testing Program and carry out its function as they strive to improve the current RDMA technologies and user experience.

Toward an Open Fabric Management Architecture

Jeff Hilland, HPE; Russ Herrell, HPE; Paul Grun, HPE

An open fabric management framework that is adaptable to existing and emerging fabrics (Gen-Z, CXL, Ethernet and others) has merit in that it preserves client software investment as fabrics change and as new fabrics emerge. Such a framework amortizes the development cost of fabric managers for various fabrics across the industry as well as lessening the development burden for emerging fabric software. This session proposes Redfish as the basis for such an open fabric management framework. Redfish is an industry standard that is emerging as a widely adopted infrastructure management interface. It is designed to be scalable, allowing it to target systems ranging in size from small clusters all the way up to the largest HPC systems. We are proposing DMTF, the OFA, Gen-Z, and other fabric standard consortia to collaborate to bring about this new, open, fabric management framework.

TriEC: An Efficient Erasure Coding NIC Offload Paradigm based on Tripartite Graph Model

Xiaoyi Lu, The Ohio State University; Haiyang Shi, The Ohio State University

Erasure Coding (EC) NIC offload is a promising technology for designing next-generation distributed storage systems. However, we find that there are at least three major limitations of current-generation EC NIC offload schemes on modern SmartNICs, which include the Bipartite graph-based EC encoding paradigm (BiEC), only supporting to offload the encode-and-send primitive, and out-of-band recovery. Thus, in this talk, we propose a new EC NIC offload paradigm based on the tripartite graph model, namely TriEC. TriEC supports both encode-and-send and receive-and-decode operations efficiently. Through theorem-based proofs, co-designs with Memcached (i.e., TriEC-Cache), and extensive experiments, we show that TriEC is correct and can deliver better performance than the state-of-the-art EC NIC offload schemes (i.e., BiEC). Benchmark evaluations demonstrate that TriEC outperforms BiEC by up to 1.82x and 2.33x for encoding and recovering, respectively. With extended YCSB workloads, TriEC reduces the average write latency by up to 23.2% and the recovery time by up to 37.8%. TriEC outperforms BiEC by 1.32x for a full-node recovery with 8 million records.

Using Libfabric for Scalable Distributed Machine Learning: Use cases, Learnings, and Best Practices

Rashika Kheria, Amazon

Amazon launched Elastic Fabric Adapter (EFA) targeting High Performance Computing (HPC) and distributed Machine Learning (ML) training workloads running on AWS EC2 servers. Distributed ML training workloads which require synchronous updating of Stochastic Gradient Decent are similar to common HPC workloads as they rely on Bulk Synchronous Parallel processing. Therefore, they also benefit from Libfabric's OS-bypass high-bandwidth low latency communication. In this talk, we will describe how the common ML training frameworks has evolved over the years, and share how Libfabric providers brought significant performance improvements compared to traditional TCP based communication. We will then discuss our learnings from deploying a high-performing libfabric provider, EFA, for a variety of ML training customers at different scales: starting from smaller 16 GPU (2 nodes) clusters to massive 2048 GPUs hosted on 256 nodes. We will describe various customers' use cases, learnings, and the gains customers have seen with EFA. Lastly, we will conclude with performance benchmarks and best practices to be followed for deploying ML training applications on EC2.

Using SPDK to Optimize Your NVMe-oF RDMA Stack

Seth Howell, Intel, Corp.; Alex Rosenbaum, Mellanox Technologies; Tzahi Oved, Mellanox Technologies

The Storage Performance Development Kit (SPDK) provides developers with everything they need to create blazing fast NVMe-oF initiators and targets. Come learn about what the SPDK open source community has created in this space - SPDK is a user-space toolkit focused on taking full advantage of today's latest hardware using state of the art programming techniques. You'll also learn about several recent innovations in the SPDK NVMe-oF RDMA transport to increase both performance and reliability including DIF/DIX support, port failover, and SEND and RECV operation batching. We'll also take a look at the usefulness of using SPDK within the context of a hardware offload application.

Visualize and Analyze your Network Activities using OSU INAM

Hari Subramoni, The Ohio State University; Pouya Kousha, The Ohio State University; Kamal Raj Ganesh, The Ohio State University; Dhabaleswar Panda, The Ohio State University

As heterogeneous computing (CPUs, GPUs etc.) and, networking (NVLinks, X-Bus, etc.) hardware continue to advance, it becomes increasingly essential and challenging to understand the interactions between High-Performance Computing (HPC) and Deep Learning applications/frameworks, the communication middleware they rely on, the underlying communication fabric these high-performance middlewares depend on, and the schedulers that manage HPC clusters. Such understanding will enable application developers/users, system administrators, and middleware developers to maximize the efficiency and performance of individual components that comprise a modern HPC system and solve different grand challenge problems. Moreover, determining the root cause of performance degradation is complex for the domain scientist. The scale of emerging HPC clusters further exacerbates the problem. These issues lead to the following broad challenge: How can we design a tool that enables an in-depth understanding of the communication traffic on the interconnect and GPU through tight integration with the MPI runtime at scale?