OFA Workshop 2019 Session Abstracts

Administration, Management and Deployment

Debug and Diagnostics

Jason Gunthorpe, Mellanox Technologies

Every technology requires tools for debug and diagnostic and RDMA is no different. As deployment becomes wider and as technology is being used by various users with various levels of expertise, the need for easy to manage debug and diagnostic tools becomes more essential. Many tools are helping to debug RDMA. Some tools are not “solely” for RDMA debug. Such tools are ethtool, ip link (iproute2 package) for checking and configuring ethernet and IP layer and tcp dump for capturing traffic. Some tools are RDMA specific. The new evolving rdmatool provides socket monitoring, RDMA counters and is a home for other RDMA configuration and monitoring. In addition, traffic generators for RDMA allow testing of utilization and latency. On top of that end to end monitoring helps to monitor and sanitized network issues. In this talk I’ll review the tools above and will focus on RDMA related tools, current state and going forward

High Speed Network Monitoring Enhancements

Brett Holman, Los Alamos National Laboratory

Los Alamos National Laboratory's in-house high speed network monitoring software was developed to provide information about fabric issues and useful information about fabric utilization. The software has successfully met these objectives in the past, however new interest in increasing its capabilities has risen from multiple teams at LANL. The goal of this project is to increase the frequency of gathering fabric performance counters such that more granular understanding of network utilization is possible. Prior to modification, the execution frequency of the software could be modified with the configuration settings. However, the run time of the monitoring code was greater than our objective sample rate for our largest systems, so gathering fabric utilization data at the desired frequency was not possible. To implement the enhancements, the code was modified so that fabric counter gathering was isolated to a separate process with an improved algorithm.

HPC Networking in the Real World

Jesse Martinez, *Los Alamos National Laboratory*

High speed networking has become extremely important in the world of High Performance Computing. As parallel processing capabilities increases and storage solutions increase in capacity, the network must be designed and implemented in a way to keep up with these trends. Los Alamos National Laboratory (LANL) has a very diverse use of high speed fabrics within its environment, from the compute clusters, to the storage solutions. This keynote/introduction session to the Sys Admin theme at the workshop will focus on how LANL has made use of these diverse fabrics to optimize and simplify the notion of data movement and communication to obtain these results for scientists solving real world problems.

NVMe over Fabrics Scale-Out Management

Phil Cayton, Intel Corp.

Los Alamos National Laboratory's in-house high speed network monitoring software was developed to provide information about fabric issues and useful information about fabric utilization. The software has successfully met these

objectives in the past, however new interest in increasing its capabilities has risen from multiple teams at LANL. The goal of this project is to increase the frequency of gathering fabric performance counters such that more granular understanding of network utilization is possible. Prior to modification, the execution frequency of the software could be modified with the configuration settings. However, the run time of the monitoring code was greater than our objective sample rate for our largest systems, so gathering fabric utilization data at the desired frequency was not possible. To implement the enhancements, the code was modified so that fabric counter gathering was isolated to a separate process with an improved algorithm.

Visualize and Analyze Your Network Activities Using OSU INAM

Hari Subramoni, The Ohio State University

OSU INAM monitors InfiniBand clusters consisting of several thousands of nodes in real time by querying various subnet management entities in the network. It is also capable of interacting with the MVAPICH2-X software stack to gain insights into the communication pattern of the application and classify the data transferred into Point-to-Point, Collective and Remote Memory Access (RMA). OSU INAM can also remotely monitor several parameters of MPI processes such as CPU/Memory utilization, intra- and inter-node communication buffer utilization etc in conjunction with MVAPICH2-X. OSU INAM provides the flexibility to analyze and profile collected data at process-level, node-level, job-level, and network-level as specified by the user. In this demo, we demonstrate how users can take advantage of the various features of INAM to analyze and visualize the communication happening in the network in conjunction with data obtained from the MPI library. We will, for instance, demonstrate how INAM can 1) filter the traffic flowing on a link on a per job or per process basis in conjunction with MVAPICH2-X, 2) analyze and visualize the traffic in a live or historical fashion at various user-specified granularity, and 3) identify the various entities that utilize a given network link.

Birds of a Feather

How to Drive the Interop and Logo Program into the Future?

Paul Grun, Cray

This BoF is planned as a follow-up to the single session titled "Driving the Interop Program into the Future". In the great spirit of BoFs, this one is designed to encourage an open discussion on the appropriate path forward for these valuable programs, particularly in light of the newly emerging Distro testing program, now underway. Topics to be discussed include the appropriate objectives for such programs, how we can find synergy among the various components of the program, appropriate vendors to implement those components, community feedback on the overall value of the program and methods to further increase its utility and value both to Alliance members, Interop Program participants, and the community at large.

Libfabric and HPC Runtime Systems

Yanfei Guo, Argonne National Lab

Libfabric is an important tool for building portable, parallel programming model runtime libraries for different high-performance networks. Implementations of MPI and OpenSHMEM have utilized Libfabric for several years. This BoF session will provide a forum for the parallel programming model runtime system developers from vendors and the community to discuss the experiences and lessons learned in adopting Libfabric in their runtime libraries. The discussion will include demonstrating the current status of using Libfabric and future plans. The speakers will also discuss the advantages and disadvantages of Libfabric from a runtime library perspective. The BoF session also provides a chance for the developers to compare notes and discuss what changes in Libfabric can be more beneficial to the programming model runtime development. Although it will be the first time this BoF session being held at the OpenFabrics Alliance Workshop, we have successfully held BoF sessions at SC with a similar organization for MPICH. Currently, we have speakers from Argonne National Laboratory, Intel, Los Alamos National Laboratory and Ohio State University planning to addend, with more pending confirmation. The speakers will represent some of the most popular HPC runtime libraries including MPICH derivatives, OpenMPI derivatives, and multiple OpenSHMEM implementations.

Interoperability

Driving the Interop Program into the Future

Paul Grun, Cray; Tatyana Nikolova, 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.

Fabrics and Technology

Amazon Elastic Adapter: Anatomy, Capabilities, and the Road Ahead

Raghu Raja, Amazon

Elastic Fabric Adapter (EFA) is the recently announced HPC networking offering from Amazon for EC2 instances. It allows applications such as MPI to communicate using the Scalable Reliable Datagram (SRD) protocol that provides connectionless and unordered messaging services directly in userspace, bypassing both the operating system kernel and the Virtual Machine hypervisor. This talk presents the designs, capabilities, and an early performance characterization of the userspace and kernel components of the EFA software stack. This includes the open-source EFA libfabric provider, the generic RDM-over-RDM (RxR) utility provider that extends the capabilities of EFA, and the device driver itself. The talk will also discuss some of Amazon's recent contributions to libfabric core and future plans.

Faster Fabrics Running Against Limits of the Operating System, the Processor and the I/O Bus

Christopher Lameter, Jump Trading LLC

In 2017 we got 100G fabrics, in 2018 200G fabrics and in 2019 it looks like 400G technology may be seeing a considerable amount of adoption. These bandwidth compete with and sometimes are higher than the internal bus speeds of the servers that are connected using these fabrics. A worry is that this trend is continuing with Terabit speed fabrics in 2022. One wonders what are the implications of this for high speed fabrics? Numerous companies have started to work on projects that remedy the situation by for example having active NICs that can do partial processing on their own, by establishing ways via other busses to devices so that full performance is possible, by sharing a NIC between multiple servers and so on. This means that there is the danger of a blooming field of new proprietary technologies and extensions to RDMA technology developing in the coming years. I think we need to consider these developments and work on improving fabrics and the associated APIs so that ways to access these features become possible using vendor neutral APIs. It needs to be possible to code in a portable way and not to a vendor specific one.

NVMe over Fabrics Offload

Tzahi Oved, Mellanox Technologies

NVMe is a standard that defines how to access a solid-state storage device over PCI in a very efficient way. It defines how to create and use multiple submission and completion queues between software and the device over which storage operations are carried and completed. NVMe-over-Fabric is a standard that maps NVMe to RDMA in order to allow remote access to storage devices over an RDMA fabric using the same NVMe language. Since NVMe queues look and act very much like RDMA queues, it is a natural application to bridge between the two. In fact, couple of software packages today implement an NVMe-over-Fabric to local NVMe target. The NVMe-oF Target Offload feature is such an implementation that is done in hardware. A supporting RDMA device is configured with the details of the queues of an NVMe device. An incoming client RDMA connection (QP) is then bound to those NVMe queues. From that point on, every IO request arriving over the network from the client is submitted to the respective NVMe queue without any software intervention using PCI peer-to-peer access.

To HDR and Beyond

Ariel Almog, Mellanox Technologies

Recently, deployment of 50 Gbps per lane (HDR) speed started and 100 Gbps per lane (EDR) which is a future technology is around the corner. These technologies exposing various new physical interfaces for copper and optical interfaces and type of transceiver like SFP-DD. Supporting these speeds also toughen the task to get low BER (Bit Error Rate) through FEC (Forward Error Correction) algorithm. The high bandwidth might cause the NIC PCIe interface to become a bottle neck as PCIe gen3 can handle up to single 100 Gbps interface over 16 lanes and PCIe gen4 can handle up to single 200 Gbps interface over 16 lanes. In addition, since the host might have dual CPU sockets, Socket direct technology, provides direct PCIe access to dual CPU sockets, eliminates the need for network traffic to go over the inter-process bus and allows better utilization of PCIe, thus optimizing overall system performance.

Realworld High Performance Networking Advantages, Comparison and Challenges in the Financial Markets

Sampath Tilakumara, Millennium IT Software (Private) Limited / LSEG Technology

Several different high performance network technologies have emerged to support financial market system requirements. However, the hardware platforms that traditionally analyzed may have been out-of-date. Latest research work indicating that Onload and Offload performance can transform with the rapid development of the computer architecture and computer hardware components. It might be required to re-analysis the performance of Onload and Offload categories on the current infrastructure platforms. The comparison required common tools to analyze between HPN variants. Both, single and multi-session simulation tools are useful to mimic the realworld application behaviors. This enables the predictions of real application behaviors and room for further optimizations. Similarly, common benchmarking processes may provide data points to decide between IB, RoCE, Onload and Ethernet. Further, a common baseline testing process enables to verify the performance of deployed environments and other factors such as failover functionality. Still there are many challenges to HPN transport mechanisms to compete with common features provided by Ethernet transport. This includes, network level HBA failover mechanisms for IB, Teaming Support for IB, Active-Active bonding support for IB, time synchronization between IB only hosts, IB packet capturing and monitoring at nano-second precision, reliable multicast and exponential multi-session latency problem of Onload techniques.

Libfabric Design and Implementation

An Overview of OFI Utility Providers: Implementation and Uses

Alexia Ingerson, Intel Corp.

The OFI libfabrics library includes various utility providers that can be layered over core providers to provide extended functionality and communication. The utility providers include 1) RxM – RDM over MSG; 2) RxD – RDM over DGRAM; 3) smr – intra-node communication through shared memory; 4) hook – intercept of provider calls. The goal of this presentation is to give an overview of these utility providers, outlining how they deal with issues such as addressing, memory registration, windowing, RMA and atomics, header optimization, scaling, and optimized protocols, and to explain how they can be used to enhance other providers

Enabling Applications to Exploit SmartNICs and FPGAs

Sean Hefty, Intel Corp.; Venkata Krishnan, Intel Corp.

Advances in Smart NIC/FPGA with integrated network interface allow acceleration of application-specific computation to be performed alongside communication. This communication works in a synergistic manner with various acceleration models that include inline, lookaside or remotely triggered ones. Bringing this technology to the HPC ecosystem for deployment on next-generation Exascale class systems however requires exposing these capabilities to applications in terms that are familiar to software developers. In this regard, the lack of a standardized software interface that applications can use is an impediment to the deployment of Smart NIC/FPGA in Exascale platforms. We propose extensions to OFI to expose these capabilities. This would improve the performance of middleware based on this interface. And in turn, this will indirectly benefit applications that use that middleware without requiring any application changes. Participants will learn about the potential for Smart NIC/FPGA application acceleration and will have the

opportunity to contribute application expertise and domain knowledge to a discussion of how Smart NIC/FPGA acceleration technology can bring individual applications into the Exascale era.

Experiences with Libfabrics

Harold Cook, Lightfleet

The Open Fabrics Alliance developed and supports the libfabric interface to provide a high-performance, scalable, application centric, extensible interface for the OFI stack that has as a goal to be hardware agnostic. Over the last several months, we have been developing to the libfabric interface and encountered a number of issues that we would like to share with the community. By no means is this presentation a condemnation of libfabric, its developers or design. Rather, it is to share experiences with the intent of improving libfabric.

Network Consumers and Runtime

Accelerating TensorFlow with RDMA for High-Performance Deep Learning

Xiaoyi Lu, The Ohio State University

Google's TensorFlow is one of the most popular Deep Learning (DL) frameworks. In distributed TensorFlow, gradient updates are a critical step governing the total model training time. These updates incur a massive volume of data transfer over the network. In this talk, we first present a thorough analysis of the communication patterns in distributed TensorFlow. Then, we propose a unified way of achieving high performance through enhancing the gRPC runtime with Remote Direct Memory Access (RDMA) technology on InfiniBand and RoCE. Through our proposed RDMA-gRPC design, TensorFlow only needs to run over the gRPC channel and gets the optimal performance. Our design includes advanced features such as Message Pipelining, Message Coalescing, Zero-Copy Transmission etc.
The performance evaluations show that our proposed design can significantly speedup gRPC throughput by up to 2.6x compared to the default gRPC design. By integrating our RDMA-gRPC with TensorFlow, we are able to achieve up to 56% performance improvement for TensorFlow training with CNN models.

Designing High Performance MPI Collectives in MVAPICH2 for HPC and Deep Learning

Hari Subramoni, The Ohio State University

The performance of collective communication operations (such as broadcast, reduce, all-reduce, and all-to-all) is critical to obtain applications-level performance and scalability for HPC and Deep Learning. Multi-/Many-core architectures are seeing widespread adoption in current and next-generation supercomputing systems due to their power/performance ratio. However, this increased density of the compute nodes and the performance characteristics of the new architecture bring in a new set of challenges in designing next-generation collective communication operations. In this talk, we present some of the advanced designs to tackle such challenges in the MVAPICH2 MPI library on modern multi-/many-core systems with high-performance interconnects like InfiniBand and Omni-Path. These designs have already been presented at prestigious conferences and have been incorporated into the MVAPICH2 libraries. In particular, we will present the following designs: a) contention-aware, kernel-assisted designs for large-message intra-node collectives (Cluster'16), b) integrated collective designs with SHARP (SC'17), c) designs for scalable reduction operations for deep learning (PPoPP'16), and d) shared-address space (XPMEM)-based scalable collectives (IPDPS'18). Benefits of these designs for a range of HPC and DL applications on various platforms will be presented.

Evaluation of Hardware-Based MPI Acceleration on Astra

Michael Aguilar, Sandia National Laboratories

As High Performance Computing marches towards the Exascale era, there is an inevitable increase in cores and sockets within a given compute node. This results in an increase in MPI endpoints and a decrease in memory per core, and with unexpected messages, smaller buffer sizes which must accommodate short MPI messages. Furthermore, the amount of virtual memory available to each core is reduced, as well, for zero-copy MPI operations, and application placement can also impact potential processing time for many MPI operations. To address these issues, NIC-based processing for MPI collectives can help provide increased operational bandwidth and help to reduce memory consumption. Mellanox has recently released the Scalable Hierarchical Aggregation and Reduction Protocol (SHArP) as a potential solution within InfiniBand NICs, however a detailed investigation is still necessary. This talk will discuss the effort at Sandia National Laboratories to evaluate NIC-based processing in SHArP on the Astra ARM64 HPC system. With thousands of nodes and

hundreds of thousands of cores on Astra, we effectively investigate the utility and detailed mechanisms within SHArP, at a sufficiently large scale. Initial results indicate SHArP, with OpenMPI, improves Allreduce collectives by over 50 percent over standard MPI collective mechanisms.

HPNL: A High-Performance, Light-Weight Network Library for Big Data Application

Jian Zhang, Intel, Corp.

Nowadays data is growing at a faster rate than ever before and it presents new challenges for large-scale Big Data analytics. Spark is expected to achieve high throughput & ultra-low latency for different workload. However, previous studies showed it can be improved by using RDMA networking. New emerging persistent memory technologies like DCPMM can offer persistency with memory-like speed, combining RDMA and persistent memory create tremendous opportunities for Spark Shuffle acceleration. We present high-performance network library (HPNL), a light weight network library built on Libfabric for big data application. It provides protocol-independent networking framework, C/JAVA API and high-level abstraction to let developer easily replace other TCP/IP based network library, like ASIO or Netty, without knowing the low-level details of RDMA programming model. We will showcase the benchmark result compared with other network libraries. One design principle of HPNL is to ease storage and network stack integration and supports API to access remote persistent memory. We will also present a new Spark Shuffle Manager based on HPNL, which leveraging non-volatile persistent memory as shuffle storage and RDMA for network transmission. Our evaluation shows that this approach significantly improves the Spark end-to-end job execution time by up to 10x.

In-Network Computing

Tomislav Janjusic, Ph.D, Mellanox Technologies

HPC applications, as well as advances in Artificial Intelligence and specifically deep learning applications, require extreme computing capabilities. These may be provided by highly parallel processors (like GP GPUs and AI accelerators) interconnected to form large clusters. Enabling efficient parallel computers require not only high bandwidth and low latency connectivity between the processing engines but also rethinking of the boundaries between system components, such as software and hardware, processing engines and network components considering, to produce more capable co-designed systems. This talk will briefly describe distributed AI applications and bottlenecks that limit their scalability, and how in-network computing enables removing these bottlenecks. The focus will be on the use of in-network data aggregation capabilities, their application to deep learning and show real world deep learning application performance results.

NCCL and Libfabric: High-Performance Networking for Machine Learning

Brian Barrett, Amazon

NCCL is a GPU-oriented collective communication library developed by NVIDIA to accelerate deep learning frameworks such as Caffe, MxNet, and TensorFlow. NCCL is topology aware, taking advantage of on-node networks as well as multiple internode network interfaces in a single node. NCCL 2 was recently made available under a BSD license on GitHub and includes provisions for adding support for net network stacks. In the fall of 2018, AWS open sourced a Libfabric driver for NCCL (https://github.com/aws/aws-ofi-nccl). This talk examines the design choices for mapping NCCL communication semantics on Libfabric, presents paths forward for supporting GPUDirect with Libfabric, and includes a discussion on how to grow the development community of the Libfabric driver for NCCL.

Remote Persistent Memory

RDMA Persistent Memory Extensions

Tom Talpey, Microsoft

RDMA protocols are being extended to support Remote Persistent Memory. Active work in the IBTA and IETF are focused on InfiniBand/RoCE and iWARP, respectively, and both standards are converging on a common set of extended operations. This talk will focus on the model and specific proposed extensions, and will outline how the OFA may plan its work to address upper layer interfaces.

A Data Store for Resource Efficient Serverless Computing

Bernard Metzler, IBM

Serverless computing is a cloud-computing execution model where the provider dynamically manages the allocation of computing resources. As a cloud service, it is becoming increasingly popular due to its high elasticity and fine-grain billing. While serverless platforms were originally developed for web microservices, their elasticity advantages make them appealing for a wider range of applications such as interactive analytics and machine learning. Unfortunately, the expensive handling of ephemeral data and task state, as well as task scheduling overheads are currently preventing the applicability of serverless to complex and data intensive workloads. While increasing task scheduling efficiency is another active research field, in our talk we focus on improved data handling, proposing an efficient and flexible storage service for serverless. To balance cost, performance and flexibility, this data store builds upon remote accessible non-volatile memory. Using NVMeF as data access method, we integrate both flash and 3D Xpoint storage media. We exemplify the deployment of our data store with its prototype integration into the Apache OpenWhisk serverless framework. We argue, that such a data store - combined with efficient task scheduling - may help to extend the applicability of serverless computing, while preserving original flexibility and easy deployment advantages.

Characteristics of Remote Persistent Memory – Performance, Capacity, or Locality. Which One(s)?

Paul Grun, Cray

Persistent Memory exhibits several interesting characteristics including persistence, capacity and others. These (sometimes)competing characteristics may require system and server architects to make tradeoffs in system architecture. A sometimes overlooked tradeoff is in the locality of the persistent memory, i.e. locally-attached persistent memory versus remote(or fabric-attached) persistent memory. In this session, we explore some of those tradeoffs and take an early look at the emerging use cases for Remote Persistent Memory and how those may impact network architecture and API design.

Scalable, Resilient, and Distributed Key-Value, Store-Based Data Management over RDMA Networks

Xiaoyi Lu, The Ohio State University

Over the recent years, distributed key-value stores (e.g., Memcached, Redis) have been extensively used for designing scalable industry data management solutions. Being an essential component, the functions and performance of a distributed key-value store play a vital role in achieving high-speed data management and processing. In this talk, we present a 'holistic approach' to designing high-performance, scalable, and resilient key-value storage systems for HPC clusters, that encompasses RDMA-enabled networking and high-speed NVMs, to maximize end-to-end performance while ensuring server scalability, resilience, and persistence. In our work, we propose designs in non-blocking API semantics that can truly leverage the one-sided semantics of RDMA, while conforming its data movement semantics to those in general in-memory and hybrid key-value stores. We also propose fast online EC-aware designs for KV stores with RDMA and evaluate opportunities for leveraging it for both online and offline workloads. Then, we explore RDMA communication engine designs for emerging byte-addressable non-volatile memory (NVRAM) technologies, that can be leveraged agnostic to the underlying persistent key-value store architecture to maximize end-to-end performance. We demonstrate the applicability of our designs using online (e.g, YCSB) and offline (e.g, burst-buffer over Lustre for Hadoop I/O) workloads on multiple real-world HPC clusters.