OPENFABRICS ALLIANCE

14th Annual Workshop Abstracts

April 9-13, 2018
Embassy Suites Boulder
Boulder, CO

Keynote

Software Foundation for High-Performance Fabrics in the Cloud

Bill Magro, Intel Corporation

Artificial Intelligence and High Performance Data Analytics workloads in the cloud are being fed by a deluge of data emanating from the Internet-connected population of people and things. Autonomous driving will deliver even more data to the cloud. Finally, High-Performance computing use is growing in the cloud. More and more cloud workloads now demand high performance from their fabrics, and these workloads are also imposing new requirements beyond those driven by traditional simulation and modeling.

This talk highlights the broadening role of OpenFabrics, in general, and the Open Fabrics Interface, in particular, to rise to the challenge of meeting the emerging requirements and become the software foundation for high-performance cloud fabrics.

Birds of a Feather

ISO C++ Standardization of Fabrics

Chris Taylor, U.S. Department of Defense

The Open Fabrics Working Group organically formed an effort to pursue the development of a fabrics extension to the ISO C++ Networking Technical Specification. The session will review the standardization effort's goals and current progress. An introduction and overview of the current C++ Networking Technical Specification will be presented. The ISO C++ SG14 (a C++ working group) communicated to the Open Fabrics Working Group that a library implementation is critical to the ISO C++ standardization process. Requirements for the library implementation will be identified. The fabric user community will have opportunities at the session to provide immediate feedback and input to the standardization process.

NVMEoFabric Testing

Paul Bowden, Intel Corporation

Discussion on adding NVMEoFabric testing to OFA-IWG. Looking for feedback on how we can accelerate solutions into the market beyond the existing NVMEoF testing. Focus is at the Fabric level and addressing how to reliably test and deploy industry solutions, reducing the time to market and testing the features which end customers require.

OFA OS Distro Testing

Paul Bowden, Intel Corporation

Discussion on future directions of OFA pre-release OS Distro Fabric testing. Quick review on current strategy. Looking for participation and feedback on how to extend the program to improve OS and end user value: including (but not limited to) SW application testing enhancements, end user coverage areas of interest, and additional test suites. Also discussion HW configuration changes of interest in the future.

SNIA NVM Programming Model

Andy Rudoff, SNIA/Intel Corporation

In this presentation, Andy will report on the latest developments around the SNIA NVM Programming Model, especially focusing on Persistent Memory Programming.The model has continued to evolve but also some new, interesting challenges have come up. Andy will summarize the current support for Persistent Memory programming in operating systems and libraries, and give us a peek at upcoming work to address the challenges.

Technical Sessions

Accelerating Ceph with RDMA and NVMe-oF

Haodong Tang, Intel Corporation

Efficient network messenger is critical for today’s scale-out storage systems. Ceph is one of the most popular distributed storage system providing a scalable and reliable object, block and file storage services. As the explosive growth of Big Data continues, there’re strong demands leveraging Ceph build high performance & ultra-low latency storage solution in the cloud and bigdata environment. The traditional TCP/IP cannot satisfy this requirement, but Remote Direct Memory Access (RDMA) can.

In this session, we’ll present the challenges in today’s distributed storage system posed by network messenger with the profiling results of Ceph All Flash Array system showing the networking already become the bottleneck and introduce how we achieved 8% performance benefit with Ethernet RDMA protocol iWARP. We’ll first present the design of integrating iWARP to Ceph networking module together with performance characterization results with iWARP enabled IO intensive workload. The send part, we will explore the proof-of-concept solution of Ceph on NVMe over iWARP to build high-performance and high-density storage solution. Finally, we will showcase how these solutions can improve OSD scalability, and what’s the next optimization opportunities based on current analysis.

Amazon and Libfabric: A Case Study in Flexible HPC Infrastructure

Brian Barrett, Amazon

Amazon Web Service’s EC2 Cloud Computing infrastructure allows users to dynamically build a variety of compute environments. Continual improvements in compute performance, available accelerators, and network performance have led to EC2 being an attractive platform for many HPC use cases. As network performance becomes a larger bottleneck in application performance, AWS is investing in improving HPC network performance. Our initial investment focused on improving performance in open source MPI implementations, with positive results. Recently, however, we have pivoted to focusing on using libfabric to improve point to point performance. Libfabric provides a number of features that make it ideal for Amazon's development: changes in libfabric apply to the majority of MPI implementations, libfabric's interface allows customers to experiment with programming interfaces other than MPI, and, most importantly, the hardware agnostic interface of libfabric allows Amazon room to innovate across an ever-evolving set of hardware capabilities. We'll talk about our current experiences in getting started with libfabric, capabilities we'd like to add in 2017, and how we think about HPC networking in the Cloud.

API to Expose Verbs Object Counters to Application

Alex Rosenbaum, Mellanox Technologies

Until today, each application had to spend CPU cycles to accumulate the amount of data transfers. Now, an application that creates a QP can request HW assistance to count the amount of traffic it sent or received.

There are many IB spec counter events which the application is not even exposed to, like rnr-retry attempts. Through this new verbs counters API an application will be able to collect real time statistics and can even change its behavior due to value it reads from the different counters.
We will present application use cases, the new libibverbs API, and different objects that can be queried.

Building a High-Performance Storage Controller over Ethernet RDMA

Subhojit Roy, IBM

This paper describes various design and implementation changes in Ethernet block storage applications to take advantage of Ethernet RDMA for building high performance Enterprise storage controllers that works on both RoCE and iWARP. Existing Data centers predominantly deploy Fibre Channel as a storage interconnect, mainly due to performance and its suitability to storage requirements. However with advent of RDMA technologies (iWARP and RoCE) over High speed ethernet, Ethernet as a storage interconnect is gaining ground. This paper talk about various design and implementation changes in storage application to take advantage of low latency and Hight IOPS RDMA connectivity e.g Lock Less IO Pool, Pre-allocating CQ and QP memory

This paper also talks about various challenges while building such storage applications and other practical considerations, given that storage applications are quite different from standard applications running on server environments.

Building Efficient Clouds for HPC, Big Data, and Neuroscience Applications over SR-IOV-enabled InfiniBand Clusters

Xiaoyi Lu, The Ohio State University

Single Root I/O Virtualization (SR-IOV) technology has been steadily gaining momentum for high-performance interconnects such as InfiniBand. SR-IOV can deliver near-native performance but lacks locality-aware communication support. This talk presents an efficient approach to building HPC clouds based on MVAPICH2 and RDMA-Hadoop with SR-IOV. We discuss high-performance designs of the virtual machine and container aware MVAPICH2 library over SR-IOV enabled HPC Clouds. This talk will also present a high-performance virtual machine migration framework for MPI applications on SR-IOV enabled InfiniBand clouds. We will also discuss how to leverage the high-performance networking features (e.g., RDMA, SR-IOV) on cloud environments to accelerate data processing through RDMA-Hadoop package, which is publicly available from http://hibd.cse.ohio-state.edu/. To show the performance benefits of our proposed designs, we have co-designed a scalable and distributed tool with

MVAPICH2 for statistical evaluation of brain connectomes in the Neuroscience domain, which can run on top of container-based cloud environments with natively utilizing RDMA interconnects and delivering near-native performance. This tool is publicly available from http://neurohpc.cse.ohio-state.edu/.

Comprehensive, Synchronous, High Frequency Measurement of InfiniBand Networks in Production HPC Systems

Michael Aguilar, Sandia National Laboratories

In this presentation, we will show InfiniBand performance information gathered from a large Sandia HPC system, Skybridge. We will show detection of network hot spots that may affect data exchanges for tightly coupled parallel threads. We will quantify the overhead cost (application impact) when data is being collected.

At Sandia Labs, we are continuing to develop an InfiniBand fabric switch port sampler that can used to gather remote data from InfiniBand switches. Using coordinated InfiniBand switch and HCA port samplers, a real-time snapshot of InfiniBand traffic can be retrieved from the fabric on a large-scale HPC computing platform. Due to the time-stamped and light-weight data retrieval with LDMS, production job runs can be instrumented to provide research data that can be used to specify computing platforms with improved data performance.

Our implementation of synchronous monitoring of large-scale HPC systems provides insights into how to improve computing performance. Our sampler takes advantage of the OpenFabrics software stack for metric gathering. The OFED stack supports a common inter-operable software stack that provides the inherent ability to gather traffic metrics from selected connection points within a network fabric. We use OFED MAD and UMAD to collect the remote switch port traffic metrics.

Designing Scalable, High-Performance Communication Runtimes for HPC and Deep Learning: The MVAPICH2 Approach

Hari Subramoni, The Ohio State University

The current wave of advances in processor, interconnect, and storage technologies are driving HPC towards Exascale. On the other hand, the emergence of Deep Learning (DL) has led to many exciting challenges and opportunities. These trends have led to many new challenges in designing next-generation high-performance communication runtimes for parallel programming models like MPI as well as Deep Learning. The MVAPICH2 software libraries have been enabling HPC clusters during the last 16 years to extract performance, scalability, and fault-tolerance. We will present the approach being taken by the MVAPICH2 project including support for new verbs-level capabilities (DC, UMR, ODP, offload), SHArP, tight integration with NVIDIA GPUs (with GPUDirect RDMA and GPUDirect Async), optimized support for Intel KNL and IBM OpenPower, supporting and co-designing Deep Learning frameworks (Caffe and CNTK), and designs leading to reduced energy consumption. We will also highlight how the capabilities of InfiniBand Network Analysis and Monitoring (INAM) can be used together with the new MPI_T feature to analyze and introspect performance of an MPI program on InfiniBand clusters. We will also present features and plans of the MVAPICH2 project to provide support for the next-generation HPC and DL environments using the OpenFabrics ecosystem.

DLoBD: An Emerging Paradigm of Deep Learning over Big Data Stacks on RDMA-enabled Clusters

Xiaoyi Lu, The Ohio State University

Modern HPC clusters are having many advanced features, such as multi-/many-core architectures, RDMA-enabled interconnects, SSD-based storage devices, burst-buffers and parallel filesystems. Current generation Big Data processing middleware (such as Hadoop and Spark) have not fully exploited the benefits of the advanced features on modern HPC clusters. This talk will present RDMA-based designs using OpenFabrics Verbs to accelerate multiple components of Hadoop, Spark, and Memcached. An overview of the associated RDMA-enabled software libraries (being designed and publicly distributed as a part of the HiBD project, http://hibd.cse.ohio-state.edu) for Apache Hadoop (integrated and plug-ins for Apache, HDP, and Cloudera distributions), Apache Spark and Memcached will be presented. In addition, Deep Learning over Big Data (DLoBD) is becoming one of the most important research paradigms to mine value from the massive amount of gathered data. Many emerging deep learning frameworks start running over Big Data stacks, such as Hadoop and Spark. This talk will present a systematic characterization methodology and extensive performance evaluations on four representative DLoBD stacks (i.e., CaffeOnSpark, TensorFlowOnSpark, MMLSpark, and BigDL) to expose the interesting trends regarding performance, scalability, accuracy, and resource utilization. Finally, we will present some in-depth case studies to further accelerate deep learning workloads on Spark framework.

Dynamically-Connected Transport

Tzahi Oved, Mellanox Technologies

Dynamically-Connected (DC) transport is a combination of features from the existing UD and RC transports: DC has the ability to send every message to a different destination, like UD does, and is also a reliable transport - supporting RDMA and Atomic operations as RC does. The crux of the transport is dynamically connecting and disconnecting on-the-fly in hardware when changing destinations. As a result, a DC endpoint may communicate with any peer, providing the full RC feature set, and maintain a fixed memory footprint regardless of the size of the network. In this talk, we present the unique characteristics of this new transport, and show how it could be leveraged to reach peek all-to-all communication performance. We will review the DC transport objects and their semantics, the Linux upstream DC API and its usage.

Ethernet over Infiniband

Evgenii Smirnov, ProfitBricks GmbH

In this work we present our solution for providing virtual Ethernet networks over Infiniband. The solution is implemented as a Linux kernel driver represented as a standard network interface with all inherent benefits, such as bridging, vlan and standard tooling support like ip tool, ethtool, etc. It allows up to 5*10^8 virtual networks separated on the InfiniBand layer, supports checksum and segmentation offload on mlx4, and does not require specific IB hardware (e.g. BridgeX). It is an equivalent of Omni-Path VNIC for InfiniBand, and is similar to EoIB concept presented by Ali Ayoub at OFA-2013

Exploring Behaviors in a Production HPC InfiniBand Network

Serge Polevitzky, Sandia National Laboratories

Sandia’s Lightweight Distributed Metric Service (LDMS) is provided in DoE’s Tri-Lab Operating System Stack (TOSS) distributions. TOSS is a RHEL-based Operating System.

LDMS provides new insights into HPC behavior and performance. This presentation will focus on data that we have recently collected and analyzed from one of Sandia's HPC clusters. While LDMS can capture a wide spectrum of HPC performance and behavior data, this presentation will focus on data gathered and analyzed from parts of a Sandia HPC QDR InfiniBand (IB) fabric.

Recent investigations of LDMS data has allowed us to capture high-resolution sampling of parts of an IB fabric here at Sandia. Time series charts of the data will be used to show the difference between what we assumed we knew and what was actually happening.

The goal is to develop insight using the high-resolution LDMS data in order to be able to anticipate bad behavior in the fabric, especially anticipating or preventing destructive congestion conditions. Hints will be provided as to possible "tipping point" Host Channel Adapter (HCA) values to monitor for potential fabric slowdowns or meltdowns. In closing, a suggested list of "items we think we have learned from this analysis" will be presented.

High-Performance Big Data Analytics with RDMA over NVM and NVMe-SSD

Xiaoyi Lu, The Ohio State University

The convergence of Big Data and HPC has been pushing the innovation of accelerating Big Data analytics and management on modern HPC clusters. Recent studies have shown that the performance of Apache Hadoop, Spark, and Memcached can be significantly improved by leveraging the high-performance networking technologies, such as Remote Direct Memory Access (RDMA). Most of these studies are based on `DRAM+RDMA' schemes. On the other hand, Non-Volatile Memory (NVM)and NVMe-SSD technologies can support RDMA access with low-latency, high-throughput, and persistence on HPC clusters. NVMs and NVMe-SSDs provide the opportunity to build novel high-performance and QoS-aware communication and I/O subsystems for data-intensive applications. In this talk, we propose new communication and I/O schemes for these data analytics stacks, which are designed with RDMA over NVM and NVMe-SSD. Our studies show that the proposed designs can significantly improve the communication, I/O, and application performance for Big Data analytics and management middleware, such as Hadoop, Spark, Memcached, etc. In addition, we will also discuss how to design QoS-aware schemes in these frameworks with NVMe-SSD.

HotPlug Support of RDMA devices During VM Migration

Alex Rosenbaum, Mellanox Technologies

HotPlug of IB devices support is required in order to allow IB resource serving and/or transferring among virtual machines and containers. We've extended libibverbs to report and properly handle an IB device that is been removed from the system or when a new IB device is introduced.

In the talk we'll explain how device hotplug is handled in the IB core, how libibverbs handles the hotplug events, and also how an user-space application should handle them. We'll discuss legacy application behavior and what new application logic can gain from this new support. We present details about a real deployment application which already today is based on the libibverbs hotplug support in order to properly handle migration and service down time periods in a VM environment.

Intel® Omni-path and NVIDIA GPU Support

Ravindra Babu Ganapathi, Intel Corporation

Intel® Omni-Path Architecture (OPA) maximizes support for heterogeneous clusters consisting of Intel®

Xeons, Intel® Xeon PHI and GPU. This session discusses the Intel® Omni-Path support for GPU and associated performance optimizations. For maximizing MPI application performance we address two key areas of performance 1) Message transfers from GPU to GPU over Intel Omni-Path using GPUDirect RDMA technology to optimize the HFI<->GPU communications 2) Providing best possible affinity between GPU and OPA device where multiple OPA devices are connected to the node.

Journey to Verbs 2.0

Michael Ruhl, Intel Corporation

A discussion on the lessons learned while adapting Intel HFI1 character driver device to the new Verbs 2.0 Ioctl interface. Topics will include discussions on compiler issues because of the defined macro language used to describe the interface objects, understanding the data structure overrides (offsetof usage, etc), data structure usage (IDR, PTR, etc) nuances.

Moving Forward with Fabric Interfaces

Sean Hefty, Intel Corporation

This talk will examine the impact of libfabric on application development, plus work being done within libfabric and the OFIWG to improve the fabric software ecosystem. It will analyze how select applications are using libfabric features, new features under development, and areas of exploration. The discussion will include enhancements to the libfabric core to benefit developers as well as applications. Finally, it will touch on the standardization process for adding fabric services directly to the C++ language.

Multi-process Sharing of RDMA Resources

Alex Rosenbaum, Mellanox Technologies

Sharing of IB device, PD, QP, CQ, MR between processes is not possible today. The only exception is XRC QP. There are few inter-process sharing models for kernel resources but most of these are not applicable for RDMA resources. In this talk we would like to present few approaches to solve this. The solutions can be based on: fork and join model, shared memory model, and opening a shared handle per object model. Each of these designs has advantages and limitations. Each of these RDMA resource sharing models can fit different application and solutions requirements. We would like to present the motivation for exposing shared RDMA resources, discuss the few approaches and collect feedback from the audience regarding future directions.

New Types of Memory, Their Support in Linux and How to Use Them via RDMA

Christopher Lameter, Jump Trading LLC

Recently new types of memory have shown up like HBM (High Bandwidth Memory), Optane, 3DXpoint, NVDIMM, NVME and various “nonvolatile” types memory. This talk gives a brief rundown on what is available and gives some example on how the vendors enable the actual use of this memory in the operating system (f.e. DAX and filesystems) and then show how an application would make use of this memory. In particular then we will be looking at what considerations are important for the use of RDMA to those memory devices.

Non-Contiguous Memory Registration

Tzahi Oved, Mellanox Technologies

Memory registration enables contiguous memory regions to be accessed with RDMA.

In this talk, we show how this could be extended beyond access rights, for describing complex memory layouts. Many HPC applications receive regular structured data, such as a column of a matrix. In this case, the application would typically receive a chunk of data and scatter it by the CPU, or use multiple RDMA writes to transfer each element in-place. Both options introduce significant overhead. By using a memory region that specifies strided access, this overhead could be completely eliminated: the initiator posts a single RDMA write and the target HCA scatters each element into place.

Similarly, standard memory regions cannot describe non-contiguous memory allocations, forcing applications to generate remote keys for each buffer. However, by allowing a non-contiguous memory region to span multiple address ranges, an application may scatter remote data with a single remote key. Using non-contiguous memory registration, such memory layouts may be created, accessed, and invalidated using efficient, non-privileged, user-level interfaces.

NVMe-over-Fabrics Target Offload

Oren Duer, 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 newer 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. This session will describe how the configuration and operation of such feature should be done using verbs.

NVMf-based Integration of Non-volatile Memory in a Distributed System - Lessons Learned

Bernard Metzler, IBM Zurich Research

The commodization of both high performance, multi-gigabit networking and fast non-volatile storage technologies motivates its combined deployment as a high performance, distributed storage resource in todays Big Data processing frameworks. By example of the Crail project, which recently reached Apache Incubator status, we discuss several aspects of NVM integration with a high performance network infrastructure. Crail defines a new I/O architecture for the Apache data processing ecosystem. Serving as a distributed, fast storage media for ephemeral data, an NVMf based Crail data store can substitute large amounts of costly DRAM resources, while even susbtantially exceeding baseline performance of a legacy Apache Spark deployment. We will discuss in some detail our approach of integrating NVMe and 3D XPoint technologies using the NVMf protocol, covering several solution pathes we took. This includes using the SPDK/DPDK framework, a Java implementation of the NVMf client protocol, and utilizing NVMf target offloading.

On-device Memory Usage for RDMA Transactions

Alex Rosenbaum, Mellanox Technologies

Modern HCA and NIC can expose parts of their device memory for RDMA application to take advantage. This device memory is typically closer to the network, compared to tradition host memory, which resides past the PCI bus.

Application utilizing this device memory allow the HCA to handle RDMA transactions coming from the network and terminated on the device itself instead of an additional PCI bus transaction to the host memory. Such RDMA transactions complete much fast, with lower latency and provide high transaction rates for the HCA, application and cluster in general.

In this talk we will present the device memory concept and different types of device memories that can be allocated. We will describe the verbs interface which allows to allocate, register and use the device memory capabilities. We will discuss several application use cases which benefit from the device memory in order to reduce latency and improve overall application and cluster performance. “

OpenFabrics Verbs (OFV) multi-vendor support

Brian Hausauer, Intel Corporation

Over the years, the OpenFabrics Verbs (OFV) userspace and kernel APIs have grown considerably in size and complexity, supporting many new features, some of which are not widely supported. We survey the OFV calls and parameters to determine which are supported by multiple RNIC vendors. We promote multi-vendor OFV to end users and application developers to maximize compatibility and interoperability. We survey existing user and kernel applications to see which already conform to multi-vendor OFV. We suggest the community of end users, application developers, and RNIC vendors come together to maintain and expand the definition of multi-vendor OFV.

OpenSHMEM and OFI: Better Together

James Dinan, Intel Corporation

The Open Fabrics Interfaces libfabric API has been shown to provide a solid foundation for PGAS programming models. Recently, the OpenSHMEM community ratified the OpenSHMEM 1.4 specification, which introduces threading support along with a novel communication management API called contexts. OpenSHMEM contexts enable threading integration, performance isolation, and sophisticated communication management at the application level. As a result, they present fabric interfaces with new multithreaded usage models and resource management challenges that are uncommon among conventional HPC communication middlewares. In this talk, I will detail these and other features introduced by OpenSHMEM 1.4 and describe our experiences implementing and tuning them for OFI libfabric.

Overview of Persistent Memory Programming

Andy Rudoff, Intel Corporation

This talk covers the emerging Persistent Memory products, expected capabilities and benefits, and how developers will write programs to use it. The fundamental programming model, as defined by the SNIA NVM Programming Technical Workgroup, will be presented, along with use case examples, an overview of persistent memory libraries, and a few small code examples.

In additional to providing an overview of the current state of persistent memory, this talk provides the foundation for Tom Talpey's talk on remote access to persistent memory.

Persistent Memory over Fabrics - Beyond High Availability

Paul Grun, Cray Inc.

Persistent Memory is taking the data center by storm. But so far, much of the focus has been on local persistent memory. Extending Persistent Memory over a Fabric (PMoF) may cause a profound re-think in the way that applications share and store information.

Much of the focus on PMoF to date has been on high availability use cases. These important use cases focus on replicating cached data across a fabric to a remote PM device. This session suggests several additional emerging use cases that may significantly accelerate the adoption of remote persistent memory; this session will describe some of those use cases. The objective for this session is to expand our thinking about how remote persistent memory can be applied to enable us to tackle and solve problems that we have not been able to previously easily address. Along the way, some suggestions will be offered for possible improvements to the fabric, and how a consumer accesses the fabric, to take better advantage of remote persistent memory.

Persistent Memory Programming – The Remote Access Perspective

Tom Talpey, Microsoft

Persistent Memory access from remote contexts, for example via RDMA, is a key application of the new ultra-low-latency storage technology. Building on the concepts in Andy Rudoff's Persistent Memory Programming talk, this segment explores how the programming interfaces may support remote access. With RDMA protocol extension, a highly compelling application environment common to local and remote access can be implemented with minimal implications. The PMEM interface, Verbs, RDMA libraries, and RDMA protocol issues are explored in the remote-access context.

Proactive Identification and Remediation of HPC Network Subsystem Failures

Susan Coulter, Los Alamos National Laboratory

Network resiliency is a critical problem for the High Performance Computing (HPC) community. As HPC technology continues to gain complexity, the use and reliance on multiple types of network technologies is becoming more prevalent. At LANL these integrated networks are referred to as the IO subsystem. This subsystem consists of a high speed intra-cluster fabric, which communicates with one or more external network technologies via gateway nodes. These gateway nodes manage connectivity to external file systems and other necessary resources. Faults that occur within this subsystem can be intermittent, hard to detect, and often result in failures that do not easily indicate root cause. This presentation discusses a monitoring and detection solution that has greatly reduced the number and range of failures resulting from faults within this critical IO subsystem.

RDMAtool and Resource Tracking in RDMA subsystem

Leon Romanovsky, Mellanox Technologies

In latest kernel releases, the RDMA susbsystem was extended to provide greater debugging visibility to the users, developers and administrators. Such extension was implemented in two layers: kernel and user space application. Kernel layer provides fully featured resource (like QP, CM_ID, MR e.t.c) tracking abilities (PIDs, QP numbers, states e.t.c) together with powerful netlink interface to expose the state to the user space with proper PID namespace separation between objects. The RDMAtool is supplementary user space tool to query such information. The output includes human readable format as long as JSON output for embedding into third party applications. It provides various filtering ability to simplify the output. Fully upstreamed solution, without hidden parts and it comes as part of iproute2 package which is already available in your favorite distribution. In this talk, we will present live demonstration of the tool, will talk about current state and will present the future road map for the RDMAtool and resource tracking.

RoCE Containers - Status Update

Parav Pandit, Mellanox Technologies

Using RDMA in containerized environment is secure manner is desired. RoCE needs to operate and honor net namespace other than default init_net. This paper focuses on recent and upcoming enhancements for functionality and security for RoCE. Various modules of IB stack including connection manager, user verbs, core, statistics, resource tracking, device discovery and visibility to applications, net device migration across namespaces at minimum are the key areas to address for supporting RoCE devices in container environment.

SELinux Support in hfi1 and psm2

Dennis Dalessandro, Intel Corporation

Security Enhanced Linux, or SELinux is a combination of kernel support and user space utility. Recently SELinux support was introduced for InfiniBand. The SELinux feature is based off of PKey and Subnet ID and supports verbs usage today. However Omni-Path devices support not only verbs but also the high performance kernel bypass psm2 library.

This talk will explore our design and implementation of SELinux support
for psm2 through the hfi1 driver. We will look at the differences between SELinux for verbs and psm2. We will also look at the opportunities for code reuse, particularly the driver details of how psm2 uses the kernel.

State of the Alliance

Susan Coulter, Los Alamos National Laboratory

This talk will summarize the work of the OFA over the last year and will set the stage for the OFA's work going forward.

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.

T10-DIF Offload

Oren Duer, Mellanox Technologies

T10-DIF is a standard that defines how to protect the integrity of storage data blocks. Every storage block of is proceeded by a Data Integrity Field (DIF). This field contains CRC of the preceding block, the

LBA (block number within the storage device) and an application tag. Normally the DIF will be saved in the storage device along with the data block itself, so that in the future it will be used to verify the data integrity.

Modern storage systems and adapters allow creating, verifying and stripping those DIFs while reading and writing data to the storage device, as requested by the user and supported by the OS. The T10-DIF offload RDMA feature brings this capability to the RDMA based storage protocols. Using this feature, RDMA based protocols can request the RDMA device to generate, strip and/or verify DIF while sending or receiving a message. DIF operation is configured in a new Signature Memory-Region. Every memory access using this MR (local or remote) results in DIF operation done on the data as it moves between wire and memory. This session will describe how the configuration and operation of this feature should be done using verbs API.

The Storage Performance Development Kit and NVMe-oF

Paul Luse, Intel Corporation

The Storage Performance Development Kit (SPDK) is an open source set of tools and libraries for writing high performance, scalable, user-mode storage applications. It achieves high performance by moving all of the necessary drivers into users pace and operating in a polled mode instead of relying on interrupts. The SPDK Project released an open source NVMe over Fabrics (NVMe-oF) software target in conjunction with release of the NVMe-oF specification in 2016. This target has continued to evolve to ensure linear scaling with the addition of CPU cores, NICs, and NVMe devices, all while maintaining the low latency characteristics of RDMA. This talk will provide an overview of the SPDK project, what it’s components are, who is using it and why, with a specific focus on the NVMe-OF software target.

Using Fabrics to Accelerate Artificial Intelligence Deep Learning

Ira Weiny, Intel Corporation

Artificial Intelligence deep learning is rapidly growing in accuracy and hence value. A key challenge in deep learning is training time. Recent developments have demonstrated how clusters and fabrics can greatly reduce the time to train.

This session will briefly review the nature and evolution of deep learning. The approaches to parallelizing training and the communication data patterns observed will be covered. The resulting advantages and performance of using fabrics to accelerate training will then be presented.

Using Open Fabric Interface in Intel® MPI Library

Michael Chuvelev, Intel Corporation

In the beginning we briefly present history of OFI involvement into Intel MPI Library from a fabric of choice to the only fabric used in the upcoming product. Intel MPI Library 2019 architecture reflects the increasing scope and robustness of OFI as well as the OFI-based MPICH-CH4 development.
We talk about OFI features that are used in Intel MPI 2019 such as Scalable Endpoints and demonstrate simplicity and performance efficiency of the MPI code based on top of these, then elaborate on supported interconnects and OSes, and emphasize much better ability of Intel MPI to embrace new interconnects support provided that the corresponding OFI provider with certain capabilities is implemented.

Finally, we discuss potential areas of OFI development that might be useful for Intel MPI, such as collective operations that could abstract different vendor-specific semantics (that is, offload, triggered etc) and provide ultimate performance on specific platforms.

Windows RDMA (Almost) Everywhere

Omar Cardona, Microsoft

Windows has made significant investments in RDMA enablement everywhere (Almost). Window Server has evolved from Native NIC RDMA support, to multiple SW NICs on a PF, and finally full RDMA capable IOV-VFs. All of which encompass multiple modes of Teaming and Failover independent of iWARP, RoCEv1/2, and Converged Fabrics. Windows Clients now support RDMA providing significant performance improvements Workstation environments.

This talk will focus on the design, development, and challenges with virtualizing and securing RDMA for On-Prem Hyperconverged Storage, to multi-tenant secure cloud deployments. Challenges and tradeoffs for Teaming and High-Availability solutions, necessary HW accelerations, and future directions to complement RDMA in VMs.

NOTE: The workshop agenda and list of abstracts are tentitive and subject to change.