12th Annual Workshop Abstracts

April 4-8, 2016

Marriott Hotel & Conference Center
Monterey, CA

Communications Middleware

Designing MPI and PGAS Libraries for Exascale Systems: The MVAPICH2 Approach

Dhabaleswar Panda, The Ohio State University

The MVAPICH2 software libraries have been enabling many HPC clusters during the last 14 years to extract performance, scalability and fault-tolerance using OpenFabrics verbs. As the HPC field is moving to Exascale, many new challenges are emerging to design the next generation MPI, PGAS and Hybrid MPI+PGAS libraries with capabilities to scale to millions of processors while taking advantages of the latest trends in accelerator/co-processor technologies and the features of the OpenFabrics Verbs. In this talk, we will present the approach being taken by the MVAPICH2 project including support for new verbs-level capabilities (DC, UMR, ODP, and offload), PGAS (OpenSHMEM, UPC, CAF and UPC++), Hybrid MPI+PGAS models, tight-integration with NVIDIA GPUs (with GPUDirect RDMA) and Intel MIC, and designs leading to reduced energy consumption. We will also highlight a co-design approach where the capabilities of InfiniBand Network Analysis and Monitoring (INAM) can be used together with the new MPI-T capabilities of the MPI standard to analyze and introspect performance of an MPI program on an InfiniBand cluster and tune it further. We will also present upcoming plans of the MVAPICH2 project to provide support for emerging technologies: OmniPath, KNL and OpenPower.

GASNet: Global Address Space Networking

Paul H. Hargrove, Lawrence Berkeley National Laboratory

This talk will describe the current status of the GASNet (Global Address Space Networking) library as it relates to Open Fabrics APIs: both Verbs and OFI. Partitioned Global Address Space (PGAS) languages and libraries provide an alternative to MPI for HPC application development in which shared-memory style programs can execute in a distributed memory environment, such as a cluster connected with InfiniBand. GASNet is the most widely used networking library for building PGAS runtimes, including implementations Unified Parallel C (UPC), Co-Array Fortran (CAF), Chapel and OpenSHMEM.

High-Performance MPI Library with SR-IOV and Slurm for Virtualized InfiniBand Clusters

Dhabaleswar Panda, The Ohio State University; Xiaoyi Lu, The Ohio State University

Significant growth has been witnessed during the last few years in HPC clusters with multi-/many-core processors, accelerators, and high-performance interconnects (such as InfiniBand, iWARP and RoCE). To alleviate the cost burden, sharing HPC cluster resources to end users through virtualization is becoming more and more attractive. The recently introduced Single Root I/O Virtualization (SR-IOV) technique for InfiniBand and High Speed Ethernet provides native I/O virtualization capabilities and is changing the landscape of HPC virtualization. However, SR-IOV lacks locality-aware communication support, which leads to performance overheads for inter-VM communication within the same host. In this context, another novel feature, Inter-VM Shared Memory (IVShmem), can support shared memory backed intra-node-inter-VM communication. In this talk, we will first present our recent studies done on MVAPICH2-Virt MPI library over Virtualized SR-IOV-enabled InfiniBand clusters, which can fully take advantage of SR-IOV and IVShmem to deliver near-native performance for HPC applications. In the second part, we will present a framework for extending Slurm with virtualization-oriented capabilities, such as dynamical virtual machine creation with SR-IOV and IVShmem resources, to effectively run MPI jobs over virtualized InfiniBand clusters. Finally, we will share our experiences of these designs running on the Chameleon Cloud testbed.

Introducing Sandia OpenSHMEM with Multi-fabric Support Using libfabric

Kayla Seager, Intel; Jim Ryan, Intel

Sandia OpenSHMEM is a high-performance implementation of the OpenSHMEM standard that was designed for extreme efficiency on HPC fabrics. The Open Fabrics Interfaces Working group (OFIWG) is developing Libfabric with the goal of enabling a tight semantic map between applications and underlying fabric services. OFIWG considered OpenSHMEM as one of the primary targets when designing its interfaces. In this talk, we present an analysis of Sandia OpenSHMEM over OFI focusing on the semantic match between Libfabric and the underlying SHMEM implementation. We also demonstrate the techniques that we used to develop a portable Libfabric SHMEM implementation that runs over multiple HPC fabrics.

OpenMPI and Recent Trends in Network APIs

Howard Pritchard, LANS

As an open source project, Open MPI has long served as a proving ground for new approaches to MPI implementations. In the area of network APIs, the project has seen the recent introduction of message layer components based on both the OFIWG libfabric and Open UCX. An overview of how these network APIs are currently being used in Open MPI will be presented. Lessons learned in incorporating them into the Open MPI software stack will also be discussed, along with relevant performance data.

Status of OFI Support in MPICH

Ken Raffenetti, Argonne National Laboratory

MPICH is a high performance and widely portable implementation of the Message Passing Interface (MPI) standard. MPICH and its derivatives power many of the fastest supercomputers in the world. MPICH supports multiple network APIs by way of an abstract network module ("netmod") interface. In this session, I will give an update on OFI netmod support in the current stable MPICH 3.2.x series. Additionally, I will give a preview of the MPICH version 3.3 design (scheduled for release mid-2017), which will significantly expand the netmod interface in order to better exploit high-level network APIs such as OFI.

12th Annual OFA Workshop Abstracts

Communications Middleware

UPC++: PGAS Programming with C++

Yili Zheng, Lawrence Berkeley National Laboratory

UPC++ is a parallel programming extension for developing C++ applications with the partitioned global address space (PGAS) model. UPC++ has three main objectives: 1) Provide an object-oriented PGAS programming model in the context of the popular C++ language; 2) Add useful parallel programming idioms unavailable in Unified Parallel C (UPC), such as asynchronous remote function invocation and multidimensional arrays, to support complex scientific applications; 3) Offer an easy on-ramp to PGAS programming through interoperability with other existing parallel programming systems (e.g., MPI, OpenMP, CUDA).

UPC++ has demonstrated excellent performance and scalability with applications and benchmarks such as global seismic tomography, Hartree-Fock, BoxLib AMR framework and more. In this talk, we will give an overview of UPC++ and discuss the opportunities and challenges of leveraging modern network features.

Distributed Applications

Challenges to Large-Scale Data Transfers for Global Distributed Processing and Storage

Linden Mercer, NRL; Steve Poole, DOD

Results from NRL's SC15 SCinet Network Research Exhibition "Dynamic Remote I/O" demonstration will be presented and discussed in this session, with particular attention to the 100G RoCE network used for this demonstration which leveraged resources from ESNet's 100G network testbed, the StarLight Software Defined eXchange (SDX) in Chicago, and from DREN/CenturyLink networks as well as 100G switches from Dell, Corsa, and Brocade. NRL demonstrated large-scale remote secure data access between distant operating locations based on a dynamic pipelined distributed processing framework while leveraging software defined networking architecture (using our own controller and the open source OpenDaylight controller, 100G MACsec, and paying particular attention to flow control options in each network element). Live production quality 4K video workflows were processed across a nationally distributed set of storage and computing resources - relevant to emerging data processing challenges. The Dynamic Remote I/O strategy allows data processing "on the fly" as data begins to arrive at each compute location rather than waiting for bulk transfers to complete. We will discuss our vision and strategy for extending data center network performance across the WAN, leveraging SDN and a number of other mechanisms for obtaining highly efficient large-scale data delivery without costly retransmissions or data starvation. We will present some simulation results for extending Ethernet flow control over WAN distances.

Fabrics and Topologies for Directly Attached Parallel File Systems

Susan Coulter, Los Alamos National Laboratory

InfiniBand fabrics supporting directly attached storage systems are designed to handle unique traffic patterns, and they contain different stress points than other fabrics. These SAN fabrics are often expected to be extensible in order to allow for expansion of existing file systems and addition of new file systems. The character and lifetime of these fabrics is distinct from those of internal compute fabrics, or multi-purpose fabrics. This presentation will cover the approach to InfiniBand SAN design and deployment as experienced by the High Performance Computing effort at Los Alamos National Laboratory.

Lustre Network Multi-Rail

Amir Shehata, Intel

Today, Lustre Network (LNet) supports different fabric types. However it supports only one Network Interface per network type per node. This restricts the IO bandwidth available to a node. This is a networking bottleneck for big Lustre client nodes with large CPU count. In particular, there are Lustre installations where a few big clients are much larger than the other client nodes or the MDS or OSS nodes. Typically these systems will use InfiniBand for the LNet network. One approach then is to install additional HCAs in the system and play tricks with additional networks and/or routing to balance data streams across the extra interfaces. However, this results in a complicated configuration that is hard to maintain.

The Multi-Rail solution is intended to address the issues discussed above by allowing multiple Network Interfaces (NI) to be configured on the same network. Outgoing messages can then be sent over any available NI, as long as the peer is reachable through that NI. Incoming messages can arrive on any of these NIs. Peers know about these NIs either by being configured statically or through a dynamic discovery algorithm. This allows LNet to distribute traffic over multiple Network Interfaces increasing performance and stability. A set of other features are also being implemented to allow finer control over traffic distribution.

The Multi-Rail development is currently underway as a joint effort between Intel and SGI. Patches are being submitted to a multi-rail public branch off the Lustre master repository: http://review.whamcloud.com/#/q/status:open+project:fs/lustre-release+branch:multi-rail.

Emerging Technologies

Extensions for NVM Remote Access

Tom Talpey, Microsoft

Persistent Memory storage (PM and NVM) is becoming available in computing platforms, and PM remote access via RDMA is a compelling goal. Remote durability guarantees are required, and approaches are being pursued in several venues, including SNIA and other standards organizations. Extensions to the RDMA protocols and verbs are especially compelling. This presentation will explore emerging methods to achieve remote persistent memory access, including transparent (local-only) and explicit (requiring protocol extension), and certain benefits of each.

SNIA NVM Programming Model Update

Doug Voigt, HP Enterprise

The SNIA NVM Programming Model describes behaviors that applications can expect from non-volatile memory systems. Several related white papers are nearing completion in the areas of atomicity and remote access. This session provides a status update on the programming model primarily focused on the new requirements and issues exposed in these areas. While remote durability is the most fundamental of these, recovery from failure and the related notion of consistency points have also emerged as areas for potential innovation.

12th Annual OFA Workshop Abstracts

Emerging Technologies

Software Defined Storage Over RDMA for Enterprise Data Centre of the Future

Tej Parkash, IBM; Subhojit Roy, IBM; Rahul Fiske, IBM

Storage SANs are currently dominated by FC (Fibre Channel) technologies due to its effectiveness in providing high bandwidth, low latency, high throughput. However with the advent and popularity of low latency all-flash arrays and need of cloud centric data centers which are standardizing on Ethernet, iSER is getting more visibility as next generation high speed interconnect. Since iSER uses standard Ethernet interface it is also suitable to new data center paradigms that use Hyper-Converged Infrastructure. ISER has scope to be better in the areas of security (CHAP, IPSec), performance and cost effectiveness compare to any other technologies. Latest technology like SDN and network virtualization are targeting Ethernet as the infrastructure of choice for optimization, hence any Ethernet based technology like iSER will be benefitted from such technologies before they benefic technologies like FC.

Most storage vendors are actively working on new RDMA based networking technology plans to support iSER e.g. iWARP from Chelsio and Intel and RoCE from Mellanox and Emulex are jumping onto the iSER boat to provide RDMA based storage capability.

This paper will talk about why and how iSER is the most compelling next generation data centre technology of choice over existing protocols, primarily from the perspective of performance, cost, ease of administration, security and benefits from latest advances in networking technology like SDN and network virtualization.

The Future of RDMA and Storage: NVMe over Fabrics and RDMA to Persistent Memory

Stephen Bates, Microsemi

In this paper we present work done in the Linux kernel to facilitate p2p transfers between NVMe devices and RDMA capable NICs running protocols like InfiniBand, RoCE and iWARP.

We present experimental results using 40Gbps iWARP and 56Gbps InfiniBand and 100Gbps RoCE RNICs that show how the latency associated with remote transfer of data can be reduced whilst also offloading the CPU allowing it to focus on other tasks.

We show how this work can act as a precursor for the NVMe over Fabrics work currently being standardized. We also show how the Controller Memory Buffer (CMB) feature introduced in NVMe 1.2 be utilized in a novel fashion to aid this work. We also discuss how we might consider extending the RDMA verbs-set to add the concept of persistent writes that add a degree of reliability and robustness to the current RDMA write verbs.

Management, Monitoring, Configuration

Intel Omni-Path Fabric® Management and Tools Features

Todd Rimmer, Intel; Ira Weiny, Intel

The Intel Omni-Path Fabric® includes a number of hardware and software features to make fabric monitoring, management and diagnosis easier. This session will provide a brief overview of the management software architecture and key features.

Monitoring High Speed Network Fabrics: Experiences and Needs

Jim Brandt, SNL; Ann Gentile, SNL

The High Speed Network (HSN) of any High Performance Compute (HPC) platform is a critical shared resource. Key to efficient operation is an understanding of how HSN resources are being utilized by applications, where, when, and for how long they become oversubscribed, and the resulting impact on running applications. Obtaining synchronous high fidelity information about bandwidth and congestion can enable evaluation of their effects on applications run-times. Additionally, such information made available for appropriate run-time analysis could drive more optimal scheduling and resource allocation and more efficient platform operation.

This presentation will describe Sandia National Laboratories' (SNL) current approaches and planned future work with high fidelity, synchronous monitoring, and analysis of a variety of HSN technologies including InfiniBand, Intel OmniPath, Cray Gemini, and Cray Aries. This presentation will address existing and desired switch instrumentation and APIs to expose it along with impediments and limitations associated with monitoring these various technologies. Example analysis and visualization results will be presented. The goals of this presentation are to: 1) present SNL experiences, lessons learned, and path forward in monitoring and analysis of HSN characteristics and 2) generate constructive conversation/collaboration to drive more efficient and useful network fabrics and monitoring techniques in the future.

Topologies and Routing in the Real World

Susan Coulter, Los Alamos National Laboratory; Jesse Martinez, Los Alamos National Laboratory

As with all sophisticated and multifaceted technologies - designing, deploying and maintaining high-speed networks and topologies in a production environment and/or at larger scales can be unwieldy and surprising in their behavior. This presentation will illustrate that fact via a case study from an actual fabric deployed at Los Alamos National Laboratory.

Using High Performance Network Interconnects in Dynamic Environments

Evangelos Tasoulas, Simula Research Laboratory; Tor Skeie, Simula Research Laboratory

High performance lossless network interconnects are traditionally associated with static environments such as HPC clusters. Due to the significant network reconfiguration costs, changes in these kind of systems are rare and usually happen only when faults occur or new hardware is added. In contrary, modern multi-tenant virtualized cloud infrastructures are very dynamic by nature, and the usage of the resources is unpredictable, thus, the need to reconfigure becomes a requisite. Over the last years the OFA community has shown the potential of using high performance networks (InfiniBand) to boost the performance of virtualized cloud environments, however, the network reconfiguration challenges still continue to exist.

In this session we present the work we have been doing on InfiniBand subnet management and routing, in the context of dynamic cloud environments. This work includes, but not limited to, techniques in order to provide better management scalability when virtual machines are live migrating, tenant network isolation in multi-tenant environments, and fast performance-driven network reconfiguration.

Network APIs and Software

Effective Coding with OFED APIs to Gain Maximum Network Performance with IB

Adhiraj Joshi, Veritas Technologies

OFED-APIs provide a way to use RDMA functionality and hardware. Using OFED-API optimally and intelligently is the key to performance. It requires a transition from traditional “Ethernet”- way of thinking to “RDMA” thinking. We will explain some of the ways to effectively code with OFED. In our “Software-defined storage” product, we have used some tricks and each of them led us to progressive stages of improvement. As a final result, the performance of the product increased multifold.

For example, (A) we can leverage hardware ACKs to get rid of packet tracking. This eliminates the need of queues and hence the locking. Other improvements include (B) tuning the OFED-APIs to utilize multiple IRQ-lines to achieve receive side parallelism and (C) optimally using 32-bit immediate data to increase the IO rate.

With graphs/tables for statistics, we will describe the critical high performance networking issues encountered and how the intelligent use of RDMA technology helped us overcome them. Ours is a kernel driver, but the same tricks are applicable for any userland application using OFED APIs, as the APIs are very similar and provide same functionality.

Experiences in Writing OFED Software for a New InfiniBand HCA

Knut Omang, Oracle

This talk will present experiences, challenges and opportunities as lead developer in initiating and developing OFED stack support (kernel and user space driver) for Oracles InfiniBand HCA integrated in the new SPARC Sonoma SoC CPU. In addition to the physical HCA function SR/IOV is supported with vHCAs visible to the interconnect as connected to virtual switches. Individual driver instances for the vHCAs maintains page tables set up for the HCAs MMU for memory accessible from the HCA. The HCA is designed to scale to a large number of QPs.

For minimal overhead and maximal flexibility, administrative operations such as memory invalidations also use an asynchronous work request model similar to normal InfiniBand traffic.

Extended Ethernet Verbs

TBA

Advanced network adaptors support richer and richer stateless HW offloads including checksum calculations for L3 and L4 headers, Large Send Offload (LSO) to perform TCP segmentation on transmit, Large Receive Offload (LRO) to coalesce TCP segments on receive, Transmit Side Scaling (TSS) to utilize multiple send queues, and Receive Side Scaling (RSS) to distribute receive traffic across multiple cores.

Additionally, in cloud environments it is popular to use tunneling and overlay networks such as VXLAN, Geneve, NVGRE, and other encapsulation formats. Consequently, advanced NICs add stateless offloads to support encapsulated Ethernet frames, whereby the aforementioned offloads are available for the inner packet as well. Finally, network adapters present capture capabilities that allows for efficient sniffing and network traffic analysis.

The above offloads are mainly available for Kernel network drivers and kernel mode stacks. However, the Verbs API allows both user- and kernel-mode applications/ULPs direct access to these capabilities in the form of Raw Ethernet QPs. This API is widely-used in OS-bypass applications such as DPDK and user-level TCP stacks, e.g., VMA.

In this talk we present Verbs API extensions to support all the above stateless offloads for both Ethernet and IPoIB applications, whether in user-space or the kernel. The API proposes new Verbs objects that that separate transport contexts from memory descriptor queues. This API is directly aligned with multi-queue Ethernet networking concepts and associated offloads and sniffing capabilities, and allows for efficient, natural usage of the HW resources.

Kernel Verb API Updates

TBA

Latest upstream kernel releases introduced several RDMA verbs API improvements, which significantly simplify core Verbs programming. The first improvement was in kernel memory registration APIs. The RDMA stack suffered from multiple alternative memory registration methods (e.g., physical MRs, memory windows, FMRs, and FRWR). Supporting all those methods became cumbersome for both ULPs and providers. Thus, the community agreed to converge on a single memory registration interface - FRWR. Moreover, the FRWR interface was updated such that the RDMA core performs most of the work while providing a simple API for ULPs based on the common kernel scatterlist data structure. Finally, the new interface allows for arbitrary SG lists for supporting devices.

Another development is the CQ abstraction API. When it comes to handling the RDMA completion queue, all kernel ULPs use almost identical logic. The new CQ API moves the CQ handling to the RDMA core and leaves only minimal settings to ULPs in the form of a completion handler. The resulting RDMA core implementation is efficient, handles all corner cases, and leaves little or no room for ULP developers to make mistakes.

A third update is an API for draining QPs. Draining all pending WRs posted on a QP before tearing it down is common practice, which prevents frequent use-after-free conditions. Previously, each ULP took a slightly different approach to implement this sort of functionality. A new interface was suggested for providing a centralized implementation through a single function call. Additional improvements, currently under review, will also be covered.

kfabric: Pathfinding new OFI Kernel Mode APIs for Data Storage / Data Access

Scott Atchley, ORNL/OpenFabrics Alliance; Stan Smith, Intel/OFA; Paul Grun, Cray, Inc./OpenFabrics Alliance

The goal of the OpenFabrics Interfaces project is to develop APIs that are carefully matched to the needs of the consumer of network services. The Data Storage / Data Access (DS/DA) working group has been exploring opportunities to extend the OFI project to include a focus on network APIs for both kernel mode and user mode for data storage and data access. This session describes the pathfinding work currently underway in the DS/DA work group to define and develop APIs for kernel mode storage access.

12th Annual OFA Workshop Abstracts

Network APIs and Software

Mainstreaming RDMA API Changes for RH 7.2

Christoph Lameter, Gentwo

We have been working with Redhat and Mellanox over the last few months to work on upstreaming new features for the RDMA stack. This involves work upstream as well as work to perform proper integration into the Redhat distribution. This covers features like:

- Multicast sendonly join across InfiniBand gateways
- Unidirectional multicast streams
- Timestamping and accurate clock synchronization
- Diagnostic counters at various levels
- Maturing the ConnectX4 driver
- OS noise measures for the RDMA drivers and subsystem
- Multicast loopback filtering
- Raw ethernet packet send and receive

New Features and Capabilities in PSM2

Todd Rimmer, Intel; Ira Weiny, Intel; Russell McGuire, Intel

A new enhanced implementation of PSM (Performance Scaled Messaging) was open sourced mid-2015. A focus area for PSM2 is extreme scalability and consistent performance at scale. This session will provide an overview of the new APIs, features and capabilities provided by PSM2.

NVMe Over Fabrics Review and Update

Phil Cayton, Intel; Dave Minturn, Intel

NVMe over Fabrics is an enhancement / expansion of the NVMe specification to support remote access to NVMe storage resources via RDMA. This presentation will review the major components of the NVMe over Fabrics specification, describe the host and target stack, discuss the plan for Linux open source drivers, and preview some performance characteristics. The discussion will also touch on possible deployment models.

OFI: APIs for Data Storage/Data Access in a Non-Volatile World

Paul Grun, Cray Inc.

The Data Storage/Data Access work group (part of the broader OFI project to develop transport-neutral APIs) has been considering the role of a possible new API in accessing Non-Volatile Memory (NVM), to include both local and remote versions. Of particular interest is the use of byte-addressable NVM, and particularly remote byte-addressable NVM. In this session, we will describe the exploratory work that has been done resulting in a small number of system models that have been developed in order to describe the methods that could be used by consumers to access these devices.

The emerging world of NVM is complex, involving all layers of the network stack from the NVM device itself all the way up to the consumer of non-volatile storage or memory services. This session will be of interest anyone working in this field at any layer in the stack, with a special emphasis on engaging those who consume non-volatile storage or memory services.

The outcome of this session will influence the work currently underway in the Data Storage/Data Access group as part of the OFA's OpenFabrics Interfaces (OFI) project.

Past, Present, and Future of OpenFabrics Interfaces

Sean Hefty, OpenFabrics Alliance

The first release of OpenFabrics Interfaces (OFI) software, libfabric, occurred in January of 2015. Since then, the number of fabrics and applications supported by OFI has increased, with considerable industry momentum building behind it. This talk discusses the current state of OFI, then speaks to the application requirements driving it forward. It will detail the fabrics supported by libfabric, identify applications which have ported to it, and outline future enhancements to the libfabric interfaces and architecture.

RDMA Reset Support

TBA

The ability to reset IO devices is important to system stability. Devices may malfunction due to PCI errors, internal errors, and other unexpected circumstances. In these cases, it is important to reset device state to regain operational state. For network devices, specifically, resets are crucial to maintain host connectivity.

Resetting RDMA devices is far from trivial. Kernel ULPs hold direct references to HW resources, and user-space applications are given direct access to HW. While the RDMA stack defines asynchronous errors to notify about device failure, applications cannot be trusted to release their HW references in time.

We present a device-independent framework for supporting RDMA resets. The framework provides graceful teardown for kernel ULPs and isolates failed devices from user-space applications that still hold references to the device by redirecting these references to a SW "zombie" device. As a byproduct, RDMA drivers can now be unloaded regardless of running applications.

RDMA SELinux Support

TBA

SELinux is enforces Mandatory Access Control in Linux. SELinux restrictions are encoded into a security policy. It restricts users and processes to only the resources they need to perform their work, and cannot be overridden by system users regardless of their privileges. SELinux today covers standard TCP/IP networking, controlling which traffic flows and network interfaces a given process is allowed to access.

This session explores how SELinux may be extended to support RDMA, which often bypasses the only source of trust – the Linux kernel – while sending and receiving traffic. We map SELinux mechanisms to the RDMA communication model, and show how concrete isolation guarantees can be established by the administrator by associating InfiniBand Partitions with SELinux security tags, and controlling SMI access permissions. All relevant RDMA user-kernel interfaces are protected by suitable SELinux hooks.

Finally, we provide guidelines for managing SELinux RDMA policies. We detail recommended host security policies for both compute and SM hosts, and discuss deployment considerations.

Network Deployment

A Database Guy's Journey Into RDMA Multicast

Markus Dreseler, Hasso Plattner Institute at the University of Potsdam; Christian Tinnefeld, SAP Labs

Modern distributed database systems heavily rely on the underlying networking infrastructure. In order to minimize the impact of network operations on query processing performance, high-throughput and low-latency network technologies such as InfiniBand and RDMA are being utilized. The resulting implications on expensive data processing operations such as joins are well studied.

With a constantly growing number of nodes per distributed database system, pieces of information have to be shared with multiple nodes and group communication operations become more relevant. Although such operations are well known and frequently used in the high-performance computing community (mainly due to MPI), there is little traction in the database (and big data) community. It also seems as if there is little information available that quantifies the benefits from hardware-assisted multicasts, which in turn makes it more difficult to justify the engineering efforts and monetary investment. This talk focuses on using RDMA multicast in the context of distributed database operations and covers the following three aspects:

- Which operations in a distributed database system can benefit from RDMA multicast?
- Which frameworks and tools are available for a database systems engineer to make use of RDMA multicast?
- Presentation of experimental results that quantify a) the benefits of RDMA multicast (in comparison to uni- cast/broadcast) and b) the associated costs (such as establishing a multicast group).

Accelerating Big Data Processing (Hadoop, Spark and Memcached) on Modern HPC Clusters

Dhabaleswar Panda, The Ohio State University

Modern HPC clusters are having many advanced features, such as multi-/many-core architectures, high-performance RDMA-enabled interconnects, SSD-based storage devices, burst-buffers and parallel file systems. However, current generation Big Data processing middleware (such as Hadoop, Spark, and Memcached) have not fully exploited the benefits of the advanced features on modern HPC clusters. This talk will present RDMA-based designs using OpenFabrics Verbs and heterogeneous storage architectures to accelerate multiple components of Hadoop (HDFS, MapReduce, RPC, and HBase), 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 and HDP distributions), Apache Spark and Memcached will be presented. The talk will also address the need for designing benchmarks using a multi-layered and systematic approach, which can be used to evaluate the performance of these Big Data processing middleware.

Experiences with Large-scale Multi-subnet InfiniBand Fabrics

David Southwell, Obsidian Strategics Inc.

Multi-subnet InfiniBand fabrics offer scalability, administrative segmentation, performance and fault isolation properties that are especially attractive when deploying at extreme scale, across multiple organisations and/or across distance. This session describes the hardware and software configurations that delivered a 7 subnet fabric across the globe at SC15 – the InfiniBand routers (Crossbow) and a new fabric manager architecture (BGFC). Particular attention will be given to topology considerations and a novel routing algorithm that eliminates the possibility of credit loops. The InfiniCortex infrastructure as deployed at SC15 will be described in detail.

InfiniBand as Core Network in an Exchange Application

Ralph Barth, Deutsche Börse AG; Joachim Stenzel, Deutsche Börse AG

Group Deutsche Boerse is a global financial service organization covering the entire value chain from trading, market data, clearing, settlement to custody. While reliability has been a fundamental requirement for exchanges since the introduction of electronic trading systems in the 1990s, since about 10 years also low and predictable latency of the entire system has become a major design objective. Both issues have been important architecture considerations, when Deutsche Boerse started to develop an entirely new derivatives trading system T7 for its options market in the US (ISE) in 2008. As the best fit at the time a combination of InfiniBand with IBM(R) WebSphere(R) MQ Low Latency Messaging (WLLM) as the messaging solution was determined. Since then the same system has been adopted for EUREX, one of the largest derivatives exchanges in the world, and is now also extended to cover cash markets. The session will present the design of the application and its interdependence with the combination of InfiniBand and WLLM. Also practical experiences with InfiniBand in the last couple of years will be reflected upon.

Multicast Use in the Financial Industry

Christoph Lameter, Gentwo

Multicast is used extensively in the financial industry to distribute event information about changes on the various markets as well as news item. Plus lot of internal communications are also dependent on rapid event propagation. Multicast can be used in creative ways to come up with autodiscovery and automatically reconfiguring cloud services. In some way Multicast allows adding node independent intelligence to a network. In some sense one can facilitate the process of talking to "communities" of servers and easily implement fallback as well as dynamically extend such an architecture. This talk gives an overview on how multicast is being used in the financial industry, presents the various challenges of using the RDMA API and InfiniBand for those purposes and also shows some creative uses of multicast for scaling cloud services.

Preparing LHCb Event Building at 4 TB/s

Sébastien Valat, CERN

In 2020, the LHC at CERN will be upgraded to use higher luminosity, which implies more data for the physics experiments. For the LHCb experiment, it will turned in a total throughput of 4 TB/s going out of the detector to be filtered by a farm of roughly 3500 nodes before long term storage. In order to achieve this requirement, we are currently evaluating available 100 Gb/s network fabrics to build a 500 nodes cluster to read and aggregate the data incoming from the 10000 optical fibers going out of the detector. Considering all the IO boards this event building part of our dataflow will end-up with a total of up to 400 Gb/s of data going out/in for each node. Our current benchmark (DAQPIPE) run on top of MPI, libfabric and TCP APIs. This presentation will cover our first user experience and some performance results with libfabric on top of Intel Omni-Path and Mellanox InfiniBand networks. As our readout software needs to run 24/7 we also want to look on failure recovery possibilities over those networks and APIs.

It will be interesting to critique our approach with experts. As we also have open questions about failure recovery, it will be interesting for us to discuss some of them before we start working on this topic.

12th Annual OFA Workshop Abstracts

Networking Technologies

Creating a Common Software Verbs Implementation

Dennis Dalessandro, Intel

The hfi1 driver is Intel's latest high performance networking driver which was recently added to the Linux kernel. In addition to supporting the new Omni-Path Architecture hardware, it acts as an Open Fabrics Verbs provider via on-load techniques similar to ipath and qib.

To promote code reuse and reduce maintenance costs we have created a generic software verbs implementation, which can be shared by multiple lower level drivers. This implementation developed in the open, and with input from the community, provides a common code base without compromising the high level of performance and scalability of the existing implementations. A common implementation not only improves things from a source standpoint, it also allows other drivers to take advantage of the verbs performance improvements, which we have developed for hfi1. We expect this work to provide an excellent foundation for drivers in the Linux RDMA community, which use a software verbs implementation.

This talk will explore the motivation, goals, and architecture of our common verbs library. We will also look at the challenges faced during development and upstream code acceptance.

Evolution of PCI Express as the Ubiquitous I/O Interconnect Technology

Debendra Das Sharma, PCI-SIG

The PCI architecture has continuously evolved to be the cornerstone for I/O connectivity in computing, communication and storage platforms for more than two decades as the attach point for storage, networking, and a wide range of external I/O connectors. It enables a power efficient, high bandwidth, and low latency interconnect between CPUs, host controllers, memory and devices.

This session delves into the latest PCI Express® (PCIe®) innovations in hardware, software, and electromechanical form-factors. PCIe 4.0, the fourth generation signaling technology at 16 GT/s, doubles per-pin bandwidth in a power-efficient manner while maintaining full backwards compatibility with the prior three generations. Architected re-timers form an essential ingredient for long reach channels in PCIe 4.0 architecture.

PCIe continues to provide architectural improvements to reduce system latencies and support low cost flexible platform integration. The specification is evolving to accommodate lower power (both active and idle), software and protocol enhancements for power-efficient performance, I/O virtualization, and SoC efficiency. A robust compliance program across different form-factors ensures a plug-and-play open standard with seamless interoperability across a wide range of platforms using common components. These mechanisms scale from large platforms with fabrics interconnecting hundreds of thousands of discrete components across multiple subsystems, to highly integrated system-on-chip (SoC) implementations.

Experiences Implementing a Paravirtual RDMA Device

Aditya Sarwade, VMWare; Adit Ranadive, VMWare; Jorgen Hansen, VMWare; Bhavesh Davda, VMWare; George Zhang, VMWare; Shelley Gong, VMWare

VMware is working on a paravirtual RDMA device allowing VMs in a cluster to communicate through Remote Direct Memory Access (RDMA) with latencies that are within a few microseconds of physical hardware and bandwidth close to that of physical hardware.

The paravirtual RDMA device achieves this by allowing physical RDMA devices to be shared among VMs with low overhead while maintaining the consolidation and isolation afforded by virtualization. It is exposed to applications through the industry standard OpenFabrics Enterprise Distribution (OFED) RDMA software stack. This allows a range of applications to be deployed immediately on VMs.

In this session, we provide an overview of the design of the paravirtual device and discuss the major challenges in keeping the virtualization overhead low while supporting advanced virtualization features such as snapshots and live migration. In addition to this, we present performance measurements of a working prototype.

The session will in particular seek to engage the audience in discussions on: 1) acceptable trade-offs for RDMA users to gain the benefits of virtualization, and 2) how to better support virtualization of RDMA through hardware support.

Extending RDMA to Handle Additional Fabrics While Maintaining Application Compatibility

Ira Weiny, Intel

The approach used to add basic Intel Omni-Path Fabric® support to the RDMA kernel stack will be reviewed. The session will focus on the ways the Omni-Path Architecture is designed to be compatible with existing RDMA applications. In addition, coming modifications will be discussed which support the future performance and scalability of Omni-Path while continuing to support legacy RDMA applications. Specific topics include, enhanced MTU support, per VL MTU support, enhanced QoS support, and scalable path information.

FlashNet: A Unified High-Performance IO Stack

Animesh Trivedi, IBM Zurich Research; Nikolas Ioannou, IBM Zurich Research; Bernard Metzler, IBM Zurich Research

In response to performance improvements of modern network and storage devices, many efforts have been put in place to increase the efficiency of end-host IO stacks. However, these efforts exclusively either target the network or the storage stack but not the combination of both. In this session, we present FlashNet, a software IO stack that unifies high-performance network properties with flash storage access and management, thus eliminating excessive application and OS involvement from IO processing. FlashNet builds upon RDMA principles and abstractions and provides direct, asynchronous, byte-addressable access to a remote flash device.

The FlashNet stack consists of an RDMA controller, a flash controller, and a file system which are designed together to make RDMA network operations fast and amenable to the flash storage. Our software prototype implementation of the unified FlashNet stack delivers up-to 50% better performance than traditional solutions for accessing data stored on remote flash devices.

Networking Technologies

InfiniBand Routers Premier

TBA

InfiniBand has gone a long way in providing efficient large-scale high performance connectivity. InfiniBand subnets have shown to scale to tens of thousands of nodes, both in raw capacity and in management. As demand for computing capacity increases, future clusters sizes might exceed the number of addressable endpoints in a single IB subnet (around 40K nodes). To accommodate such clusters, a routing layer with the same latencies and bandwidth characteristics as switches is required.

In addition, as data center deployments evolve, it becomes beneficial to consolidate resources across multiple clusters. For example, several compute clusters might require access to a common storage infrastructure. Routers can enable such connectivity while reducing management complexity and isolating intra-subnet faults. The bandwidth capacity to storage may be provisioned as needed.

In this session, we will review InfiniBand routing operation and how it can be used in the future. Specifically, we will cover topology considerations, subnet management issues, name resolution and addressing, and potential implications for the host software stack and applications.

InfiniBand Virtualization: IBTA Update

TBA

Virtualized workloads are prominent in today's data center and cloud environments. Recently, there has been increasing demand to deploy InfiniBand in the data center. Use cases range from using InfiniBand as the physical network foundation for Infrastructure as a service (IaaS) deployments, to enabling HPC Platform as a Service (PaaS) in virtual MPI clusters.

The IBTA Management Working Group has been chartered to lead the standardization for InfiniBand Virtualization support. Virtualization allows a single Channel Adapter to support multiple transport endpoints, which allow different software entities to interact independently with the fabric.

This session will provide an update on the ongoing work on virtualization in the IBTA. We will cover the concepts on InfiniBand Virtualization, and its manifestation in the host software stack, subnet management, and monitoring tools.

Intel Omni-Path Fabric® Open Source Overview

Todd Rimmer, Intel; Ira Weiny, Intel

An overview of the recently upstreamed Intel Omni-Path Fabric software will be presented. Omni-Path builds on Open Fabrics to permit application compatibility while adding significant new features and capabilities. This session will review the Omni-Path software architecture and briefly discuss a few of the key components including Omni-Path driver, PSM, libfabric support, Omni-Path fabric management and tools.

OpenPOWER Based Open Hybrid Computing: Building the Ecosystem for a Flexible Heterogeneous Compute Architecture

Bernard Metzler, IBM Zurich Research

The architecture of future high performance compute systems will dictate the efficient management of an increasingly parallel and heterogeneous compute infrastructure. For POWER based systems, this technology trend led to the formation of the OpenPower Foundation, an open technical community aiming at building and successfully deploying an open ecosystem for tight integration of heterogeneous compute resources such as CPU, networking adapters, and accelerators including GPU's, FPGA's and DSP's.

We will explain strategy and road map for the POWER CPU family and its evolving high performance interconnects for efficient accelerator and networking integration, such as NVLink technology and the Coherent Accelerator Processor Interface (CAPI). Aiming at an open ecosystem, it is essential to expose those heterogeneous resources at well-defined open programming interfaces. We will therefore discuss implications for those future interfaces and in particular look at GPUDirect RDMA and extensions to the OFA RDMA Verbs interface for better GPU/CPU integration and efficient synchronization with network IO.

RDMA and User-space Ethernet Bonding

TBA

RDMA Bonding is a new scheme for aggregating network interfaces that support RDMA and user-space (OS-bypass) Ethernet. The new bonding driver introduces a single logical RDMA device that aggregates multiple interfaces. This allows Verbs applications to use a single logical device transparently and enjoy networking high availability and fail over, load balancing, and NUMA locality according to the bonding mode and underlying hardware support.

Verbs resources created through the bond device may generate traffic through multiple physical ports according to the network adapter support. Such resources will be created from underlying slave RDMA devices according to load balancing policies. Using shared hardware resources like Completion Queue (CQ) or Shared Receive Queue (SRQ) with object sourced from different IB Device slaves may be supported in case all slaves are members of the same adapter.

The OS Bypass bonding driver works similarly and in conjunction with the Linux standard bonding driver. The latter continues to support standard network interface aggregation, optionally controlling the physical link layer and running aggregation protocols like LACP for 802.3ad. Configuration and management of the RDMA bonding driver will be added to netlink and sysfs similarly to Linux bonding.

Networking Technologies

RDMA Containers Update

TBA

Container technology provide improved isolation capabilities and fine grained resource control for Linux applications and services, and has recently risen as a light weight alternative to machine virtualization. Although there has been progress in supporting RDMA applications inside containers, much work is still needed. Basic support for using RDMA CM inside a network namespace has been accepted upstream, as well as an implementation for InfiniBand. However, RoCE and iWARP support for network namespaces is still lacking, and there is not yet a mechanism for controlling RDMA resource usage inside a container.

In this talk we wish to expand on the missing pieces, and present our work on RoCE network namespace support. In addition, recent work on an rdma cgroup provides a way to set hard limits on the number of RDMA resources a container may use, allowing the administrator to allocate per-device QPs, SRQs or CQs to different containers.

With these technologies in place users will be able to enjoy the performance characteristics of RDMA with the flexibility and efficiency of containers.

Software-defined Networking on InfiniBand Fabrics

Ariel Cohen, Oracle

A design for virtual Ethernet networks over IB is described. The virtual Ethernet networks are implemented as overlays on the IB network. They are managed in a flexible manner using software. Virtual networks can be created, removed, and assigned to servers dynamically using this software. A virtual network can exist entirely on the IB fabric, or it can have an uplink connecting it to physical Ethernet using a gateway. The virtual networks are represented on the servers by virtual network interfaces which can be used with para-virtualized I/O, SR-IOV, and non-virtualized I/O.

This technology has many uses: communication between applications which are not IB-aware, communication between IB-connected servers and Ethernet-connected servers, and multi-tenancy for cloud environments. It can be used in conjunction with OpenStack, such as for tenant networks. This will also be covered in this session. The Oracle Private Cloud Appliance uses this virtual networking technology, and this will be described as well. In addition, a network services solution using this technology will be discussed.

OpenFabrics Alliance Business

Compliance and Interoperability in an Application-Centric World

Paul Grun, OpenFabrics Alliance; Robert Russell, University of New Hampshire Interoperability Lab; Paul Bowden, OpenFabrics Alliance

Classically, "compliance" implies conformance to a standard, while "interoperability" refers to the ability of devices to interoperate. In a world where the design of network APIs is being driven by the consumers of those APIs, the distinction between interoperability and compliance is becoming blurred. For example, the present OFA Interop program focuses on assuring interoperability among hardware components including switches, cables, and device adapters, and assuring that these hardware components interoperate with OpenFabrics Software (OFS). The OFI project, on the other hand is designing an architecture that is independent of the hardware details, and hence has less dependence on hardware interoperability.

Given the recent trend in the OpenFabrics community to place a greater emphasis on the needs of the consumers of high performance networks, the question arises: What service can the OFA provide in the area of compliance and interoperability to support these new APIs while encouraging their adoption, both by vendors of networking products and by the consumers of those products?

In this session, we describe this conundrum in more detail and present some new thinking about ways that the OFA can offer a valuable service, whether it is compliance or interoperability testing, to both the providers of network solutions and importantly to the consumers of network services.

National Labs Forum: Update

Paul Grun, Cray Inc.; Jim Ryan, Intel

The national labs are an important constituent of the OpenFabrics Community. The first National Labs Forum was held at last year's workshop. This session will address work being done in the Forum.

12th Annual OFA Workshop Agenda

Schedule Flow

	Monday	Tuesday	Wednesday	Thursday	Friday
	4-Apr	5-Apr	6-Apr	7-Apr	8-Apr
8:00	DS/DA Working Group Face-to-Face	Welcome Session	Network Technology	Network Technology	Network Technology
8:15					
8:30		Network Technology			
8:45					
9:00					
9:15					
9:30			Compliance and Interop in App-centric World	National Lab Forum	Emerging Technologies
9:45					
10:00		Break	Break	Break	Break
10:15					
10:30		Network APIs and Software	Network APIs and Software	Network APIs and Software	Network APIs and Software
10:45					
11:00					
11:15					
11:30					
11:45					
12:00	Lunch	Lunch	Lunch	Lunch	Closing Remarks @ 12Noon
12:15					
12:30					
12:45					
1:00	OFI WG Face-to-Face Joint meeting with DS/DA	Management, Monitoring, and Config	Management, Monitoring, and Config	Comm Middleware	
1:15					
1:30					
1:45					
2:00		Comm Middleware	Comm Middleware		
2:15					
2:30				EQG update	
2:45					
3:00		Break	Break	Break	
3:15					
3:30		Network Deployment	Network Deployment	Distributed Applications	
3:45					
4:00					
4:15					
4:30					
4:45					
5:00	Registration Opens 5 - 7:30PM	Emerging Technologies	Emerging Technologies	Emerging Technologies	
5:15					
5:30		Break	Break		
5:45					
6:00		System Admin BoF	OFA All-hands Meeting		
6:15					
6:30					
6:45					
7:00					
7:15					
7:30	Keynote 7:30 - 8:30PM				
7:45					
8:00					
8:15					
8:30	Reception 8:30 - 10PM				
8:45					
9:00					
9:15					
9:30					
9:45					
10:00					