OFA Workshop 2022 Abstracts

Trainings & Tutorials

Fabric Software Develop Platform (FSDP) "Office Hours"

Doug Ledford, Red Hat; Tatyana Nikolova, Intel Corp.; Brian Chae, Red Hat; Afom Michael, Red Hat

The FSDP office hours will have knowledgeable and experienced people on hand to guide people through the various steps necessary to make use of the FSDP cluster. This includes account creation, vpn setup, access to the builder machine, access to the beaker web interface, installation of the beaker command line tool, creation of simple automated tests, running of those tests, running of manual tests, and setting up CI pipelines for upstream repositories.

Introduction to Networking Technologies for High-Performance Computing

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

InfiniBand (IB), High-speed Ethernet (HSE), RoCE, Omni-Path, EFA, Tofu, and Slingshot technologies are generating a lot of excitement towards building next-generation High-End Computing (HEC) systems including clusters, datacenters, file systems, storage, cloud computing, and Big Data environments. This tutorial will provide an overview of these emerging technologies, their offered architectural features, their current market standing, and their suitability for designing HEC systems. It will start with a brief overview of IB, HSE, RoCE, Omni-Path, EFA, Tofu, Slingshot, and Omni-Path. In-depth overview of the architectural features of IB, HSE (including iWARP and RoCE), and Omni-Path, their similarities and differences, and the associated protocols will be presented. An overview of the emerging NVLink, NVLink2, and NVSwitch architectures will also be given. Next, an overview of the OpenFabrics stack which encapsulates IB, HSE, and RoCE (v1/v2) in a unified manner will be presented. An overview of libfabrics stack will also be provided. Hardware/software solutions and the market trends behind these networking technologies will be highlighted. Sample performance numbers of these technologies and protocols for different environments will be presented. Finally, hands-on exercises will be carried out for the attendees to gain first-hand experience of running experiments with high-performance networks.

High Performance Machine Learning, Deep Learning, and Data Science

Dhabaleswar Panda, The Ohio State University; Hari Subramoni, The Ohio State University; Arpan Jain, The Ohio State University; Aamir Shafi, The Ohio State University

Recent advances in Machine and Deep Learning (ML/DL) have led to many exciting challenges and opportunities. Modern ML/DL and Data Science frameworks including TensorFlow, PyTorch, and Dask have emerged that offer high-performance training and deployment for various types of ML models and Deep Neural Networks (DNNs). This tutorial provides an overview of recent trends in ML/DL and the role of cutting-edge hardware architectures and interconnects in moving the field forward. We will also present an overview of ML/DL frameworks with special focus on parallelization strategies for model training. We highlight new challenges and opportunities for communication runtimes to exploit

high- performance CPU/GPU architectures to efficiently support large-scale distributed training. We also highlight some of our co-design efforts to utilize MPI for large-scale DNN training on cutting-edge CPU/GPU architectures available on modern HPC clusters. The tutorial covers training traditional ML models including—K-Means, nearest neighbours—using the cuML framework accelerated using MVAPICH2-GDR. Also, the tutorial presents accelerating GPU-based Data Science applications using MPI4Dask, which is an MPI-based backend for Dask. Throughout the tutorial, we include hands-on exercises to enable attendees to gain first-hand experience of running distributed ML/DL training and Dask on a modern GPU cluster.

Technical Sessions

Accelerating MPI and Deep Learning Applications with the DPU Technology

Dhabaleswar Panda, The Ohio State University; Hari Subramoni, The Ohio State University; Arpan Jain, The Ohio State University; Nawras Ahnaasan, The Ohio State University; Tu Tran, The Ohio State University; Bharath Ramesh, The Ohio State University; Aamir Shafi, The Ohio State University

Modern Data Center Processing Units (DPU) provide programmable ARM cores in the network adapter. This technology allows part of a middleware to be executed on the DPU. Such flexibility allows overlap of computation and communication across host and DPU cores. This talk will focus on how to offload MPI non-blocking collectives into DPU to accelerate MPI applications. Examples from the MVAPICH2-DPU library together with performance evaluation results for a set of non-blocking collectives and applications (such as P3DFFT) on a 1,024 process environment with Bluefield-2 will be presented. Schemes to offload different parts of Deep Learning Frameworks with the DPU technology will also be highlighted. Performance evaluation results for offloading ResNet-50 training with CIFAR-10 dataset on Bluefield-2 adapters will be presented.

Benefits of Compute Express Link™ (CXL™) for High-Performance Computing

CXL Consortium

Compute Express Link™ (CXL™) is an open industry-standard interconnect offering coherency and memory semantics using high-bandwidth, low-latency connectivity between the host processor and devices such as accelerators, memory buffers, and smart I/O devices.

CXL technology is designed to address the growing needs of high-performance computational workloads by supporting heterogeneous processing and memory systems for applications in Artificial Intelligence, Machine Learning, communication systems, and High-Performance Computing. These applications deploy a diverse mix of scalar, vector, matrix, and spatial architectures through CPU, GPU, FPGA, smart NICs, and other accelerators. The CXL 2.0 specification introduces support for switching, memory pooling, and persistent memory – all while preserving industry investments by supporting full backward compatibility.

Based on member feedback, the CXL Consortium technical working groups published Engineering Change Notices (ECN), primarily to enhance performance, reliability, software interface, and testability while offering design simplification. The presentation will also provide a high-level sneak peek of the latest advancements in the CXL 3.0 specification development, its new use cases and industry differentiators.

Bridging RDMACM Traffic between Infiniband and ROCE

Christoph Lameter, Deutche Boerse

We developed a bridge between Infiniband and ROCE after we found that there were no solutions available out there to realize what we wanted.

We have the problem of running a large scale Infiniband installation that has aged quite a bit but also at the same time grown significantly. EUREX and various other trading venues critical to the operation of the Markets are based on the

availability of the platform. We thought it was unrealistic to think that we could switch out the technology on which our T7 system is based without significant downtime. The system is still growing organically and so we want to avoid a cut over day.

We also have expertise with Ethernet since the fan in and fan out to the T7 system is based on Ethernet connections from our customers. When we found ROCE, which allows the use of RDMA on Ethernet, we thought that to be the ideal solution because that would allow us to use Ethernet for everything.

However, how do we get this done? Ideally we would like to transfer one server at a time to the new fabric and keep the whole system operating while we slowly migrate to the new technology. For that purpose we need a bridge/gateway that enables our applications running on ROCE and Infiniband to communicate. We have developed such a solution as an open source project on Github and would like to have other join with us in our efforts to get this done.

Cloud-Native Supercomputing: Bare-Metal, Secured Supercomputing Architecture

Gilad Shainer, NVIDIA; Richard Graham, NVIDIA

High-performance computing and artificial intelligence have evolved to be the primary data processing engines for wide commercial use, hosting a variety of users and applications. While providing the highest performance, supercomputers must also offer multi-tenancy security. Therefore they need to be designed as cloud-native platforms. The key element that enables this architecture is the data processing unit (DPU). DPU is a fully integrated data-center-on-a-chip platform that can manage the data center operating system instead of the host processor, enabling security and orchestration of the supercomputer. This architecture enables supercomputing platforms to deliver bare-metal performance, while natively supporting multi-node tenant isolation. We'll introduce the new supercomputing architecture, and include applications performance results.

Communication and Computation API Composability - XPU support in OFI

Sean Hefty, Intel Corp.

Accelerators, such as GPUs and FPGAs, are frequently used to scale-UP HPC and AI applications. However, GPUs are accessed through low-level computational APIs, such as CUDA, RoCR, and oneAPI Level-0. Scale-OUT APIs, on the other hand, fall into the realm of communication APIs use to access NICs and other network hardware. This talk will discuss the impact making the libfabric communication API aware of heterogenous compute devices, such as CPUs and GPUs, collectively referred to as XPUs. It will discuss the features needed by both the communication and computation APIs in order to support high-performance applications needing both scale-up and scale-out support in a device agnostic fashion.

Designing High-Performance Alltoall Solutions for Dense GPU Systems

Hari Subramoni, The Ohio State University; Dhabaleswar Panda, The Ohio State University; Qinghua Zhou, The Ohio State University; Kawthar Shafie Khorassani, The Ohio State University; Chen-Chun Chen, The Ohio State University; Aamir Shafi, The Ohio State University

As more High-Performance Computing (HPC) and Deep Learning (DL) applications are adapting to scale using GPUs, the communication of GPU resident data is becoming vital to end-to-end application performance. Among the available MPI operations in such applications, All-to-All(v) are two of the most communication-intensive operations that become the bottleneck of efficiently scaling applications to larger GPU systems. Over the last decade, most research has focused on the optimization of large GPU-resident data transfers. However, in state-of-the-art GPU-Aware MPI libraries, MPI AlltoAll(v) communication still suffers from poor performance due to the limitation of commodity networks and data transfer patterns.

On the other hand, our research has shown that, by utilizing zero-copy load-store IPC mechanisms for multi-GPU communication within a node and GPU-based compression algorithms, we are able to significantly accelerate the

communication while concurrently reducing the volume of data transferred. In this talk, we focus on both these aspects to accelerate Alltoall communication in modern high-performance GPU-aware MPI libraries.

We demonstrate that the proposed designs achieve benefits at both microbenchmark and application levels. At the microbenchmark level, the proposed compression design can reduce the All-to-all communication latency by up to 87%. For PSDNS, a traditional HPC application, our proposed design can reduce the All-to-all communication latency and total runtime by up to 29.2% and 21.8%, respectively. For DeepSpeed, a DL optimization library, the proposed compression design reduces the MPI AlltoAll runtime by 26.4% compared to a state-of-the-art MPI library with point-to-point compression while ensuring data validation. The IPC-aware design demonstrates up to 59x better performance on

ThetaGPU for the DeepSpeed DL application. On the HPC side, with heFFTe and PSDNS, the IPC designs achieve approximately 27x and 71% better performance on ThetaGPU and Lassen, respectively.

Diving Into the New Wave of Storage Management

Richelle Ahlvers, Intel Corp.; Phil Cayton, Intel Corp.

As the NVM Express® (NVMe®) family of specifications continue to develop, the corresponding Swordfish management capabilities are evolving: the SNIA Swordfish™ specification has expanded to include full NVMe and NVMe-oF™ enablement and alignment across DMTF™, NVMe, and SNIA for NVMe and NVMe-oF use cases.

The SSM TWG and OFA™ OFMFWG are working together to bring to life an open-source OpenFabrics Management Framework, with a Redfish/Swordfish management model and interface.

If you haven't caught the new wave in storage management, it's time to dive in and catch up on the latest developments of the SNIA Swordfish specification. These include:

- Expanded support for NVMe and NVMe-oF Devices

- Managing Storage Fabrics

This presentation provides an update on the latest NVMe-oF configuration and provisioning capabilities available through Swordfish, and an overview of the most recent work adding detailed implementation requirements for specific configurations, ensuring NVMe and NVMe-oF environments can be represented entirely in Swordfish and Redfish environments.

Exploiting RDMA Mistakes in NVMe-oF Storage Applications

Konstantin Taranov, ETH Zurich; Benjamin Rothenberger, ETH Zurich; Adrian Perrig, ETH Zurich; Daniele De Sensi, ETH Zurich; Torsten Hoefler, ETH Zurich

Our work discovers vulnerabilities in the InfiniBand architecture. We show that any system that tries to make use of RDMA opens an attack surface allowing local users to bypass the security mechanisms of the operating system and its kernel. Importantly, we demonstrate how any unprivileged user can inject packets into RDMA connections created on a local network controller, even if they are created in kernel space. Therefore, any kernel application that makes use of RDMA opens an attack surface allowing the attacker to manipulate the kernel-level applications from user space by injecting RDMA requests into their connections. As an example, we show how the adversary can bypass security mechanisms of operating and file systems to directly manipulate NVMe disks at the block level without administrative privileges by manipulating NVMe-oF kernel modules.

Graphcore IPU over Fabric (IPUoF)

Wei Lin Guay, Graphcore; Dag Moxnes, Graphcore; Ville Silventoinen, Graphcore; Lars Paul Huse, Graphcore; Ola Tørudbakken, Graphcore

Graphcore Intelligence Processing Unit (IPU) provides Artificial Intelligent (AI) acceleration over the standard PCIe interface or the IPU over Fabric (IPUoF) interfaces. The key motivation behind IPUoF is to provide disaggregation of IPU processing resources pool from the host, which cannot be achieved with the standard Host IO interface. The IPUoF interface is designed to operate across various fabric technologies such as the de-facto Ethernet/RDMA over Converged Ethernet (RoCE), InfiniBand, Gen-Z or CCX. However, today it is only implemented with RoCE - thanks to the stable software stack provided by the Open Fabrics (OFED) software community.

The RoCE based IPUoF interface consists of an IPUoF-client and an IPUoF-server that form a software layer written with standard RDMA/RDMA CM verbs. The IPUoF-client may be located on any host systems (ARM, x86) running the Graphcore Poplar software stack whereas the IPUoF-server is always located on the Graphcore M2000 platform, an aarch64 based system. The Graphcore M2000 system consists of heterogeneous memory from Graphcore Gateway DDR controllers and the IPU memory (SRAM) itself. Together, IPUoF-client and IPUoF-server act as a data mover, which feeds performance-critical data asynchronously and non-timing critical data that is usually fetched in bulk. The optimization of the data retrieval is out-of-scope of this presentation. This presentation discusses the challenges and our proposed solution to support memory registration with heterogeneous memory.

HatRPC: Hint-Accelerated Apache Thrift RPC over RDMA

Xiaoyi Lu, University of California Merced

In this talk, we propose a novel hint-accelerated Remote Procedure Call (RPC) framework based on Apache Thrift over Remote Direct Memory Access (RDMA) protocols, called HatRPC. HatRPC proposes a hierarchical hint scheme towards optimizing heterogeneous RPC services and functions. The proposed hint design is composed of service-granularity and function-granularity hints for achieving varied optimization goals and reducing design space for further optimizing the underneath RDMA communication engine. We co-design a key-value store called HatKV with HatRPC and LMDB. The effectiveness and efficiency of HatRPC are validated and evaluated with our proposed Apache Thrift Benchmarks (ATB) and YCSB workloads. Performance evaluations show that the proposed HatRPC approach can deliver up to 55% performance improvement for ATB benchmarks compared with other state-of-the-art RDMA protocols. In addition, the co-designed HatKV can achieve up to 85.5% improvement for YCSB workloads.

In-Network Collective Communication Accelerations

Sean Hefty, Intel Corp.

Collective communications are critical for HPC and AI applications. As a result, several companies have introduced hardware specifically targeting the acceleration of collectives. This includes switch-based collective support, FPGAs, stand-alone accelerators, and NIC-based features. This talk will introduce collective accelerations, with a focus on switch based implementations. It will also define the libfabric architecture and software interfaces to support in-network collective accelerations, demonstrate how it maps to existing hardware, and describe possible future extensions.

Offloading Scatter-Gather via Custom Accelerators on COPA FPGA Network Platform

Shweta Jain, Intel Corp.

The Configurable Network Protocol Accelerator (COPA) framework enables FPGAs to incorporate custom inline/lookaside accelerators and attach directly to a 100 Gigabit/s Ethernet network. The hardware component of COPA provides the necessary networking/accelerator infrastructure allowing custom accelerator modules to be integrated seamlessly. Additionally, COPA software abstracts the FPGA hardware by providing support for the OpenFabrics Interfaces (OFI). The COPA OFI provider also supports an enhanced OFI interface that exposes the acceleration and networking capabilities to upper-layer middleware/applications. The performance of HPC middleware for distributed

programming models (such as MPI and OpenSHMEM) would benefit from the acceleration capabilities provided by COPA's enhanced OFI, thereby improving overall application performance.
In this presentation, we will outline the hardware architecture for enabling inline/lookaside acceleration in COPA. We will describe the COPA OFI infrastructure that will enable the use of accelerated COPA transmit commands as well as lookaside accelerator blocks to speed up widely used vector read and write operations. We leverage the COPA acceleration capability to gather sparse local data at network speeds as we perform a write operation and further offload the scattering of data to a dedicated hardware block in a remote read operation to maximize system performance. Additionally, we will also go over the different invocation models – local as well as remote (triggered acceleration). We will share OFI microbenchmark results that compare our implementation to that of a software-based approach.

OpenFabrics Management Framework (OFMF) Demonstrations

Russ Herrell, HPE; Jim Hull, IntelliProp

The OpenFabrics Alliance (OFA), together with its partners, the DMTF, SNIA, and the Gen-Z Consortium are in the process of designing and developing the OpenFabrics Management Framework (OFMF) to be used by clients to deliver security services, switch and end point inventories, route management, telemetry, performance and diagnostics, and more. A demonstration of the OFMF will use an agent to manipulate memory blocks that are connected through a Gen-Z fabric. The demonstration will show how the memory block resources are modeled into a Redfish/Swordfish representation of a computing system. A client can make simple queries and take actions in the Redfish/Swordfish domain that are then executed, physically, using a Gen-Z Agent and the Gen-Z Zephyr Fabric Manager. The purpose of the OFMF demonstration is to provide attendees with a current understanding of the extensive capabilities of the OFMF, to enable attendees to understand the benefits of the abstracted control and coordination of fabrics and resources, and to foster and enable collaborative work using the OFMF.

Omni-Path Express (OPX) Libfabric Provider: Overview & Status

Tim Thompson, Cornelis Networks; Dennis Dalessandro, Cornelis Networks; Lakshmi Pedda, Cornelis Networks

Omni-Path Express (OPX) is an alternative libfabric provider for Omni-Path 100 networks, written from the ground up. It is largely a drop-in replacement for PSM2, and has many of the same requirements as PSM2, such as the hfi1 kernel driver and the opafm service. No changes to the hfi1 module parameters or fabric configuration are needed.

OPX was originally based on the libfabric BGQ provider. The core logic for the Omni-Path Host Fabric Adapter was written from scratch to be an optimal semantic match to libfabric-enabled applications like mpich, Open MPI, and OpenSHMEM. Most of the embedded system constraints from BGQ have been maintained, giving OPX a small cache line footprint and favorable instruction count metrics when compared with PSM2 for many operations. OPX is still under active development and more enhancements are in plan.

We will present some current benchmarks and profiling information that highlights the gains of OPX over PSM2, as well as the current status of development, stability, and scale that development has reached.

PCIe® 6.0 Specification: A High-Performance I/O Interconnect for Advanced Networking Applications

Debendra Das Sharma, Intel Corp.

For the past three decades, PCI-SIG® has delivered a succession of industry-leading PCI Express® (PCIe®) specifications that remain ahead of the increasing demand for a high-bandwidth, low-latency interconnect for compute-intensive systems in diverse market segments. The PCIe 6.0 specification, released in early 2022, is the latest evolution of PCI Express technology. PCIe 6.0 architecture doubles the data rate of the PCIe 5.0 specification to 64 GT/s (up to 256 GB/s for a x16 configuration) and introduces innovative features like Pulse Amplitude Modulation with 4 levels (PAM4) signaling, low-latency Forward Error Correction (FEC) and Flit-based encoding while maintaining full backwards compatibility with previous generations.

In this session, attendees will learn the fundamentals of PCIe 6.0 technology, including power efficiency, error handling, and performance and reliability metrics. The presentation will highlight PCIe 6.0 technology benefits, use cases and the applications that will be accelerated by PCIe 6.0 technology, such as artificial intelligence/machine learning and high-performance networking. Attendees will also receive updates on the adoption of PCIe 4.0 technology in data center applications and the PCIe 5.0 Compliance Program.

Performance Characterization of MPI Libraries on Arm-based AWS HPC Cloud Instances with Elastic Fabric Adapters

Aamir Shafi, The Ohio State University; Dhabaleswar Panda, The Ohio State University; Hari Subramoni, The Ohio State University; Shulei Xu, The Ohio State University

Recent advances in the adoption of High Performance Computing (HPC) by major Cloud vendors have made efficient Virtual Machine (VM) instances more accessible. These instances are typically equipped with multi-core processors, accelerators, and efficient adapters. In this context, the emerging Arm based HPC instances are also receiving attention. Amazon Web Services (AWS) — a leading Arm-based cloud system vendor – recently announced new c6gn instances with Graviton 2 Arm CPUs and Elastic Fabric Adapters. In this talk, we characterize the performance and capability of the AWS Arm architecture. We explore optimizing the performance of our production-quality MPI library – MVAPICH2 – based on the features of Arm-based cloud systems and Scalable Reliable Datagram protocol of Elastic Fabric Adapter. We later evaluate the impact of our optimization that depicts significant improvement for MPI communication at both the benchmark and the application level. We gain up to 86% performance improvement for benchmark-level collective communication operations and up to 9% improvement for the Weather Research and Forecasting (WRF) application. We provide a comprehensive performance evaluation for several popular MPI libraries on AWS Arm-based Cloud systems with EFA support. HPC application developers and users will be able to get insights from our study to achieve better performance for their applications on Arm-based cloud systems with EFA support.

Performance Evaluation of MPI on the Slingshot Interconnect

Dhabaleswar Panda, The Ohio State University; Hari Subramoni, The Ohio State University; Kawthar Shafie Khorassani, The Ohio State University

The deployment of the HPE Slingshot interconnect technology on upcoming exascale systems such as Frontier at OLCF and El-Capitan at LLNL drives a need for a thorough analysis of the Slingshot networking ecosystem. MPI libraries have been extensively optimized over the years for an underlying Infiniband networking environment. The interconnect between nodes heavily influences performance and scalability of communication, motivating a need for a thorough evaluation of MPI-level performance on a system connected by slingshot interconnect technology. In this work, we present a detailed analysis of the performance of various MPI and communication libraries on the early access Spock system at OLCF with nodes connected by Slingshot-10. We evaluate point-to-point and collective performance using MVAPICH2, OpenMPI + UCX, Cray MPICH, and RCCL on AMD MI100 GPUs and AMD Epyc Rome CPUs.

SOFA-Storage: Creating a vendor agnostic framework to enable seamless storage offload using SmartNICs

Raphael Polig, IBM; Jonas Pfefferle, IBM; Nikolas Ioannou, IBM

Recently, SmartNIC accelerators became the state-of-the-art solution for providing storage and network virtualization in cloud environments. With more computational power and additional features these devices stretch well beyond just offloading network-related tasks. Recent generations of SmartNICs are capable of acting as storage devices and offer accelerated access to remote storage. SmartNICs offer both increased security by enforcing strong isolation from the host system, as well as increased performance through hardware offloads for storage access and encryption. Datacenter system integrators want to provide a Block-Storage-as-a-Service on top of smartNICs with minimal changes to their control plane, and to have the capability to use smartNICs from multiple vendors. However, the APIs for configuring block storage offload on smartNICs are typically different from vendor to vendor, making it difficult to integrate smartNICs in a vendor-agnostic manner. SOFA-Storage tries to address this issue by exposing a unified and remotely

accessible API from the smartNIC towards the control plane, while providing a plugin framework to translate API requests to vendor-specific operations.

We present our early experience and proof-of-concept framework that provides a vendor-agnostic smartNIC storage API, to present block storage to bare metal hosts, virtual machine hosts, and kubernetes containers.

sPIN: High-performance streaming Processing in the Network

Salvatore Di Girolamo, ETH Zurich; Daniele De Sensi, ETH Zurich; Torsten Hoefler, ETH Zurich

The capacity of offloading data and control tasks to the network is becoming increasingly important, especially if we consider the faster growth of network speed when compared to CPU frequencies. In-network compute alleviates the host CPU load by running tasks directly in the network, enabling additional computation/communication overlap and potentially improving overall application performance.

sPIN is a programming model for in-NIC compute, where users specify handler functions that are executed on the NIC, for each incoming packet belonging to a given message or flow. It enables a CUDA-like acceleration, where the NIC is equipped with lightweight processing elements that process network packets in parallel. We investigate the architectural specialties that a sPIN NIC should provide to enable high-performance, low-power, and flexible packet processing. We introduce PsPIN, a first open-source sPIN implementation, based on a multi-cluster RISC-V architecture and designed according to the identified architectural specialties. We investigate the performance of PsPIN with cycle-accurate simulations, showing that it can process data at full network speed for several use cases, introducing minimal latencies (26 ns for 64 B packets) and occupying a total area of 18.5 mm 2 (22 nm FDSOI).

Status of OpenFabrics Interfaces (OFI) Support in MPICH

Yanfei Guo, Argonne National Laboratory

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

Towards Universal Management of Fabric Interfaces

Richelle Ahlvers, Intel Corp.; Phil Cayton, Intel Corp.; Michael Aguilar, Sandia; Russ Herrell, HPE

The OpenFabrics Alliance (OFA), together with its partners, the DMTF, SNIA, and the Gen-Z Consortium are in the process of designing and developing the OpenFabrics Management Framework (OFMF) to be used by clients to deliver security services, switch and end point inventories, route management, telemetry, performance and diagnostics, and more. Currently, targeted clients include Workload Managers, MPI and SHMEM, distributed deployment services, and others. The OFMF consists of a set of common tools designed for managing and assembling, and aggregating underlying fabrics with simple abstractions. Through the use of universal set of tools, client APIs and methods can create resource/client associations, sub-fabrics construct super-fabrics, get performance information, and manipulate underlying fabrics.

Improvements in processing data are creating more architecturally complex computing systems. The OFMF provides a set of tools and methods that is targeted to boost computational performance. Each resource component, such as memory, storage, compute, and accelerators, are interconnected by high speed fabrics. With no common way of querying or manipulating high-speed fabrics and resources there is a current design limit to how resources can be applied to computations. Through the use of dynamic construction and aggregation of components within new large-scale distributed architectures, new optimized methods can be instantiated when requested by clients for optimized computing solutions for High-Performance Computing, Machine Learning, Cloud-based systems, and Enterprise

environments. The OFMF provides a logical representation of the fabric for easy client control and dynamic client manipulation.

The OFMF services are presented to all clients via Redfish API calls. The OFMF maintains the aggregate Redfish model of all fabrics it controls and all resources on those fabrics. When clients request data or request changes to model objects, the OFMF determines which Redfish objects in the model are impacted, makes the required changes to those Redfish objects in the model. Any relevant actions or requests that affect state or configuration of the fabric manager or actual fabric hardware are relayed by the OFMF to the fabric-specific provider. System Administrators, Application Designers, and System Architects can design, deploy, maintain, and use any sort of fabric-based computing system. Some simple client requests (for example, to DELETE an object) may translate to multiple changes to multiple model objects and potentially require multiple exchanges with the fabric provider. Maintaining the integrity and consistency of the aggregate Redfish model of all fabric resources is one of the primary duties and major values of the OFMF. InfiniBand, Gen-Z, Slingshot, and others types of fabrics can be managed by the OFMF.

Update on PSM3 Capabilities and Architecture

James Erwin, Intel Corp.; Todd Rimmer, Intel Corp.

Over the last year a number of new capabilities have been added to PSM3 along with some refinements to its internal architecture. This session will provide a quick overview of the latest capabilities, including enhancements to the aws-nccl-plugin to enable NCCL over OFI/PSM3, PSM3 support for TCP/IP, etc. In addition to sharing the new capabilities, the latest performance data for PSM3 will be shared.

SNIA Swordfish™: In Action Demonstration

Richelle Ahlvers, Vice Chair and Executive Committee, SNIA Board of Directors/Storage Technology Enablement Architect, Intel Corp.

Check out an interactive demonstration of the current state of Swordfish: see mockups, tools, and emulators. See how the Swordfish works in practice, interactive examples of different types of storage devices, and walk through real use cases.

Scalable Storage Management Demonstrations – SNIA Swordfish™: In Action Demonstration

Richelle Ahlvers, Vice Chair and Executive Committee, SNIA Board of Directors/Storage Technology Enablement Architect, Intel Corp.

Check out an interactive demonstration of the current state of Swordfish: see mockups, tools, and emulators. See how the Swordfish works in practice, interactive examples of different types of storage devices, and walk through real use cases.

Computational Storage Demonstrations

View three interactive demonstrations of how computational storage is successfully implemented in today's demanding storage life on the edge environments featuring high performance, retail, security, and artificial intelligence/machine learning solutions.

- **Computational Storage -- High Performance Object Storage (HPOS) with SmartSSD Demonstration**
  Mayank Saxena, Member, SNIA Computational Storage Special Interest Group/Senior Director, Engineering, Samsung

- **Computational Storage - Far Edge Retail chain and Far Edge/Near Edge Data Security Use Case Demonstrations**
  J B Baker, Member, SNIA Computational Storage Special Interest Group/Senior Director, of Product Management, ScaleFlux

- **Computational Storage - AI/ML Workloads Run in Drive and Distributed/Parallel Processing Demonstrations**
  Eli Tiomkin, Member, SNIA Computational Storage Technical Work Group/VP Business Development, NGD Systems