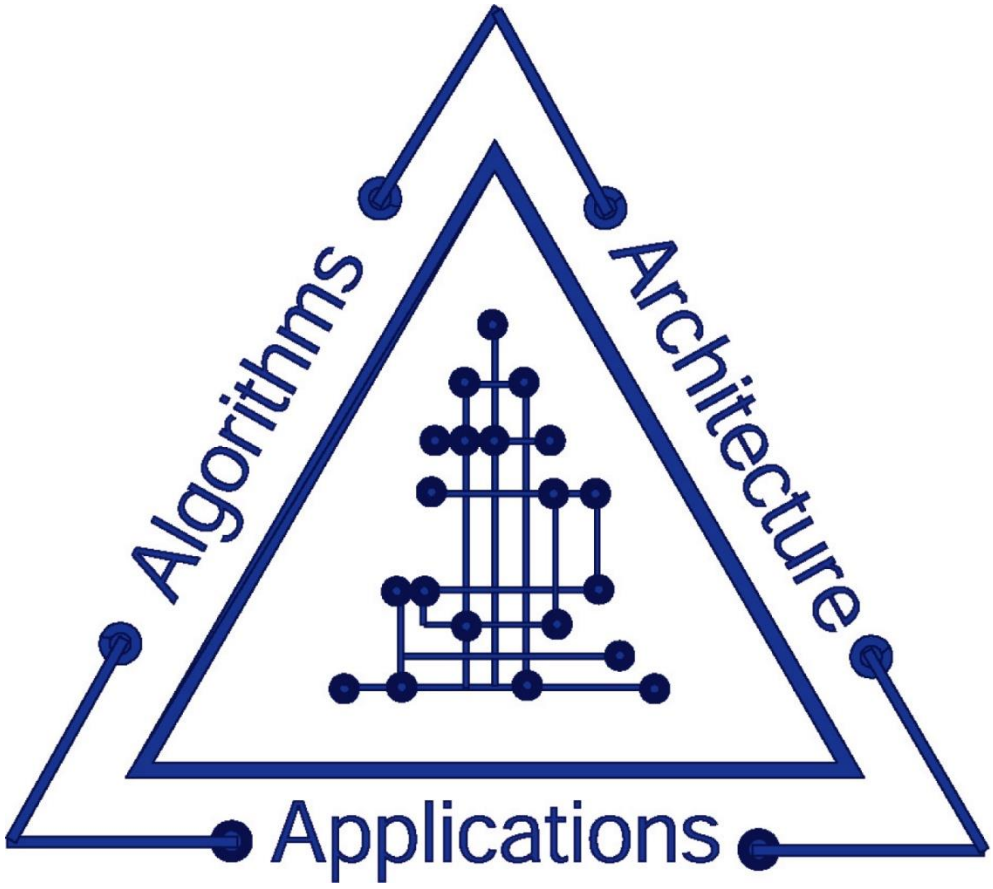
The Salishan Conference on

HIGH-SPEED COMPUTING



April 25 – 28, 2016

Salishan Lodge
Gleneden Beach, Oregon

Welcome

The Association for High Speed Computing welcomes you to the Salishan Conference on High-Speed Computing. This conference was founded in 1981 gathering experts in computer architectures, languages, and algorithms to improve communication, develop collaborations, solve problems of mutual interest, and provide effective leadership in the field of high-speed computing. Attendance at the conference is by invitation only; we limit attendance to about 170 of the world's brightest people. Participants are from national laboratories, academia, government, and private industry. We keep the conference small to preserve the level of interaction and discussion among the attendees.

The conference agenda and selection of participants has been designed to focus discussion on technical issues of relevance to our conference theme: Data Movement for Computing at Scale. The speakers have been selected to address our theme and give attendees information about the latest technologies and issues facing high performance computing. The evening sessions are structured to encourage informal discussions and networking among all participants.

If you have any comments or suggestions for future topics and/or speakers, we encourage you to speak to any of the conference committee members.

We hope you find this conference stimulating, challenging, and also relaxing—enjoy!

Conference Committee

Kim Cupps & Bert Still, *LLNL*
Jim Ang & Ron Brightwell, *SNL*
Carolyn Connor & Christoph Junghans, *LANL*

Logistics

Conference sessions and the Random Access session will be held in the Long House. Lunches and the working dinner will be held in the Council House.

For administrative support, please speak to Jan Susco, Dee Cadena or Gloria Montoya-Rivera, located in the registration area (Salal Room). If you have specific questions regarding audiovisual equipment or network connectivity, please seek out administrative support.

Next Conference Dates:	April 24-27, 2017
	April 23-26, 2018
	April 22-25, 2019

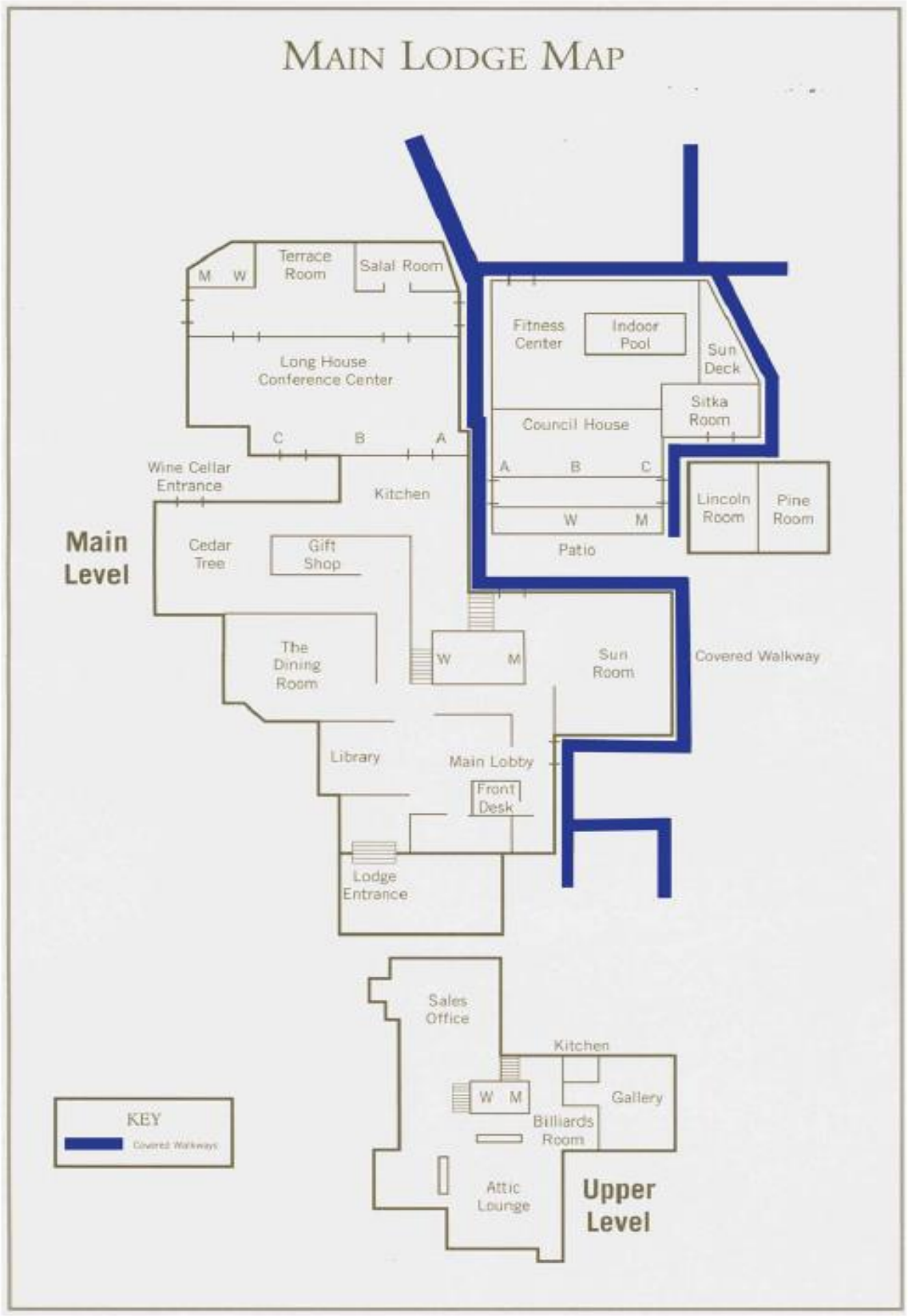

MAIN LODGE MAP
Terrace Room
Salal Room
M W
Long House Conference Center
C B A
Wine Cellar Entrance
Kitchen
Fitness Center
Indoor Pool
Sun Deck
Sitka Room
Council House
A B C
W M
Patio
Lincoln Room
Pine Room
Main Level
Cedar Tree
Gift Shop
The Dining Room
W M
Sun Room
Covered Walkway
Library
Main Lobby
Front Desk
Lodge Entrance
Sales Office
Kitchen
W M
Gallery
Billiards Room
Attic Lounge
Upper Level
KEY
Covered Walkways

Table of Contents

THIS PAGE LEFT BLANK
INTENTIONALLY

Sponsorship

The Salishan Conference on High-Speed Computing is administered, hosted, and managed by The Association for High Speed Computing (AHSC). Additional sponsorship for the evening portions of our program is provided by the corporations listed here.

One of the highlights of the conference are the informal discussions held each evening. These sessions help us to go beyond the formal presentations to exchange ideas, solve problems and develop friendships.

This year the following companies are helping to sponsor the evening informal discussions sessions:

Advanced Micro Devices, Inc.

ARM

Cray, Inc.

D-Wave Systems, Inc.

DDN Storage

EMC

Hewlett Packard Enterprise

IBM Corporation

Intel Corporation

Micron Technology, Inc.

NVIDIA Corporation

PGI Compilers & Tools

Seagate Government Solutions

Silicon Graphics International, Inc.

We would like to express our gratitude to these companies for their generous support

THIS PAGE LEFT BLANK
INTENTIONALLY

Conference Theme

Data Movement for Computing at Scale

The Salishan Conference on High-Speed Computing, was founded to pursue collaborative discussions among national labs, government, academia and industry in the areas of algorithms, architectures, and languages, and to maximize the nation's return on investment in high performance computing (HPC). In particular, balancing the performance and cost of scientific simulations on current and future platforms has been a primary objective. Optimizing data movement for speed, cost, or space has been a compelling research topic for many years. This year's conference examines the decadal challenges we face in the context of data movement for computing at scale.

As HPC approaches exascale, there are new and changing pressures and drivers affecting data movement, including but not limited to the explosion of cores, increase in heterogeneity, increasingly complex memory hierarchies, and power management concerns. Careful workflow analysis informs data movement decisions at every level of the system. Data movement must be reviewed from multiple perspectives within the overall context of an integrated environment. During this year's sessions, participants will examine data movement challenges and opportunities from the hardware, system software, applications, storage, data analysis, workflow, and visualization perspectives. Invited talks will focus on recent research in areas that are particularly important to facilitate data movement optimized for computing at scale. The main conference goal, as always, is to provide ample forums for discussion among participants, and to provide feedback, discuss issues, develop collaborations, and recommend solutions.

Session 1: Hardware Architecture Data Movement Capabilities

From a hardware perspective, the balance of data movement capabilities has a primary impact on both realized application performance and energy to solution. For the HPC community, concepts of balance include both on-node data movement up/down the local memory hierarchy and off-node data movement across the interconnection network fabric. New architectural capabilities such as heterogeneous multicore and many-core node architectures and multilevel main memory with stacked DRAM, conventional dynamic random access memory (DRAM), and non-volatile random access memory (NVRAM) complicate the analysis of system balance. From an interconnection network perspective, the system includes compute nodes that may support in-situ visualization/analysis, input/output (I/O) storage nodes and perhaps new node categories, NVRAM nodes for burst-buffer functions, in-transit data analysis and visualization, all adding further complexity to the analysis of system-level data movement balance. This session examines data movement hardware capabilities at both the local node architecture level and the global system architecture level. Key questions include: what data movement capabilities have the greatest potential for synergy between HPC and high-performance data analysis (HPDA)? What hardware technology demonstrators are needed in the next few years to create an on-ramp for high-performance, energy-efficient, resilient data movement capabilities as candidate technologies for integration into the future Department of Energy (DOE) exascale platforms?

Session 2: System Software Data Movement Capabilities

As new capabilities are introduced into memory, compute, and interconnect hardware, system software must continue to evolve to manage an increasing amount of complexity in HPC systems. New and alternative approaches to programming HPC systems and developing scalable applications bring about new requirements and challenges for operating and runtime systems and the programming interfaces between layers of the software stack involved in moving data. The desire to support a broader set of applications, including integrated applications composed of many components and more sophisticated application workflows, also creates challenges for system software. The need to reduce the cost of data movement within endpoints as well as across all of the endpoints in the system motivates the need for more sophisticated resource allocation and management strategies. This session examines the data movement challenge from a system software perspective from both the node- and system-level perspective. Key topic areas for discussion include: programming interfaces for data movement, resource management strategies for reducing data movement, requirements and challenges introduced by alternative programming models and runtime systems, and software support for implicit and explicit communication mechanisms.

Session 3: Data Movement from the Applications Perspective

This session will illustrate the problems and solutions of data movement for computing at scale from the applications side. With the growing size of machines and increasing levels of memory hierarchies, efficient data movement will become a greater burden for the application programmer, and certainly new algorithms and programming techniques are required. In this session, the following questions should be attacked: How does data gets moved through different cache levels and to the memory while the application processes? How can applications be written to minimize unnecessary data movement? How can this data management be abstracted away from the application designer using portable frameworks? Can the data movement be modeled using analytical or semi-analytic methods at small scale to predict results at a larger scale? How can the above optimizations be done in a portable way?

Session 4: Data Movement from the Data Analysis, Workflow, and Visualization Perspectives

In the search for high performance with improved energy efficiency, large-scale systems over the next decade increasingly will continue to utilize alternatives to traditional disk storage. Technologies such as NVRAM or solid-state disk can be placed on the compute node, or close-by, to provide higher bandwidth and faster access but at the expense of reduced capacity. One such use case is the burst-buffer concept that will be present on flagship machines delivered to several DOE and DOE/NNSA laboratories over the next four years. These technologies present opportunities and challenges for data analysis, workflow, and visualization. The key will be effectively handling the large amounts of data generated by a simulation. Key questions include: how might the workflow change to achieve better data efficiency? How well can in-situ analysis and visualization techniques provide equivalent understanding to today's traditional methods? Can partial analysis techniques reduce the data set sufficiently to fit in a burst buffer while retaining enough features to enable data exploration? What might an efficient framework look like for supporting this tiered analysis? Could data analytics techniques (from the "big data" space) be useful?

Session 5: Input/Output, File Systems, and Data Storage Data Movement Challenges

This session examines the decadal challenges facing data storage, file systems, and I/O in the context of data movement for computing at scale, including but not limited to new and changing pressures and drivers, both technical and economic. In concert, the session explores game-changing ideas and technology and/or radical changes in perspective or approach to meet the identified challenges. Key questions include: what are the consequences in an ecosystem in which the economics of storage/data movement/data analysis are fundamentally changing and where the simple model of memory/disk/tape is insufficient to meet minimum technical requirements for computing at scale? What are the complexities and challenges and potential solutions that arise at the interface where parallel data movement meets the wide area network? Are opportunities hidden in the melding of cloud- and HPC-based technologies and approaches? Is there a role for peer-to-peer storage solutions?

THIS PAGE LEFT BLANK
INTENTIONALLY

Conference Program

Data Movement for Computing at Scale

Monday, April 25, 2016

4:30–7:00 pm **Registration (Salal Room)**

6:00 pm **Welcome/Keynote Address**

Title: **Supercomputers and Superintelligence**

Speaker: Horst Simon, *Lawrence Berkeley National Laboratory*

7:00 pm **Reception and Informal Discussions (Immediately following the Keynote in the Council House)**

Tuesday, April 26, 2016

8:00 am **Registration Opens (Salal Room)**

Breakfast available (Terrace)

8:30 am **Session 1: Hardware Architecture Data Movement Capabilities**

Title: **Fabric Data Movement: An Integrated Hardware/Software Challenge**
Speaker: Keith Underwood, *Intel Corporation*

Title: **Exascale Starts with the Memory System**
Speaker: Bruce Jacob, *University of Maryland*

Title: **Is Memory Scaling Really Over? Myths, Facts and a Path Forward**
Speaker: Hillery Hunter, *IBM Corporation*

10:00 am **Break**

Refreshments available (Terrace)

10:30 am **Title:** **Extreme Scale Computing with Optically Enabled Data Movement**
Speaker: Keren Bergman, *Columbia University*

11:00 am **Panel Discussion**

Tuesday, April 28, 2016

12:00 pm **Lunch (Council House)**

1:30 pm **Session 2: System Software Data Movement Capabilities**

Title: **Rearchitecting Systems Software for Memory Heterogeneity and Scale World**

Speaker: Ada Gavrilovska, *Georgia Institute of Technology*

Title: **Data Movement with MPI in a Many-Threaded**

Speaker: Ryan Grant, *Sandia National Laboratories*

Title: **Active RDMA – New Tricks for an Old Dog**

Speaker: Torsten Hoefler, *ETHZ*

3:00 pm **Break**

Refreshments available (Terrace)

3:30 pm **Title:** **Got Burst Buffer. Now What?**

Speaker: Adam Moody, *Lawrence Livermore National Laboratory*

4:00 pm **Panel Discussion**

Tuesday, April 26, 2016

6:00 pm **Working Dinner/Speaker (Council House)**

Title: **How Facebook Uses Advanced Interconnect Technology to Make the World More Open and Connected**

Speaker: Katharine Schmidtke, *Facebook*

8:00 pm **Reception and Informal Discussions (Immediately following the Working Dinner in the Cedar Tree Room)**

Wednesday, April 29, 2016

8:00 am **Introduction to Sessions**

Breakfast available (Terrace)

8:30 am **Session 3: Data Movement from the Applications Perspective**

Title: **An Early Start with Multi-Level Memory – An Opportunity for Performance or Just the Next Chapter in Programmer Hell?**

Speaker: Simon Hammond, *Sandia National Laboratories*

Title: **Integrated Modeling for Rapid Assessment and Performance Prediction of HPC Applications**

Speaker*:* Jason Liu, *Florida International University*

Title: **Data Movement Challenges at NTS: A Paradigm for HPC Applications I/O**

Speaker: Mark Miller, *Lawrence Livermore National Laboratory*

10:00 am **Break**

Refreshments available (Terrace)

10:30 am **Title:** **Data Movement from the Applications Perspective**

Speaker: Sanjay Padhi, *Amazon*

11:00 am **Panel Discussion**

Wednesday, April 27, 2016

12:00 pm **Lunch on your own**

1:30 pm **No Scheduled Session**

5:00 pm **Random Access (Long House)**

The Random Access session consists of timely communications from participants in areas of interest to the Conference. Presentations are strictly limited to 10 minutes. A sign-up board is provided in the registration lobby.

8:00 pm **Reception and Informal Discussions (Council House)**

Student Poster Session (Council House)

This conference selects and hosts students from various universities, inviting them to present posters and discuss their research with our Salishan participants. All conference attendees are encouraged to visit with this year's students:

Lee Savoie, *University of Arizona*

Stephen Herbein, *University of Delaware*

Alfredo Giminez, *University of California, Davis*

Taylor Groves, *University of New Mexico*

Shang Li, *University of Maryland*

Jacob Hemstad, *University of Minnesota*

Sergio Pino, *University of Delaware*

Jesus Pulido, *University of California, Davis*

Divya Banesh, *University of California, Davis*

Thursday, April 28, 2016

8:00 am **Introduction to Sessions**

Breakfast available (Terrace)

8:30 am **Session 4: Data Movement from Data Analysis, Workflow, and Visualization Perspectives**

Title: **Workflow Analysis – An Approach to Characterize Application and System Needs**
Speaker: David Montoya, *Los Alamos National Laboratory*

Title: **Open, Reproducible HPC Data Analysis and Visualization: Minimizing Data Movement**
Speaker: Marcus Hanwell, *Kitware*

Title: **What HPC Can Learn from the Cloud Approach to Big Data**
Speaker: Dale Southard, *NVIDIA*

10:00 am **Break**

Refreshments available (Terrace)

10:30 am **Title:** **Extreme Data Management Analysis and Visualization for Exascale Supercomputers**
Speaker: Valerio Pascucci, *University of Utah*

11:00 am **Panel Discussion**

Thursday, April 28, 2016

12:00 pm **Lunch (Council House)**

1:30 pm **Session 5: Input/Output, File Systems, and Data Storage Data Movement Challenges**

Title: **HPC Storage and IO Trends and Workflows**
Speaker: Gary Grider, *Los Alamos National Laboratory*

Title: **Accelerating Science with the NERSC Burst Buffer Early User Program**
Speaker: Deborah Bard, *NERSC*

Title: **From File System to Services: Changing the Data Management Model in HPC**
Speaker: Rob Ross, *Argonne National Laboratory*

3:00 pm **Break**

Refreshments available (Terrace)

3:30 pm **Title:** **Breaking Free from Globally Serializable OLTP Embedded in Parallel File Systems**
Speaker: Garth Gibson, *Carnegie Mellon University*

4:00 pm **Panel Discussion**

5:00 pm **Reception and Informal Discussions (Council House)**

Abstracts

Keynote Address

Supercomputers and Superintelligence

Horst Simon, *Lawrence Berkeley National Laboratory*

In recent years the idea of emerging superintelligence has been discussed widely by popular media, and many experts voiced grave warnings about its possible consequences. This talk will use an analysis of progress in supercomputer performance to examine the gap between current technology and reaching the capabilities of the human brain. In spite of good progress in high performance computing (HPC) and techniques such as machine learning, this gap is still very large. I will then explore two related topics through a discussion of recent examples: what can we learn from the brain and apply to HPC, e.g., through recent efforts in neuromorphic computing? And how much progress have we made in modeling brain function? The talk will be concluded with my perspective on the true dangers of superintelligence, and on our ability to ever build self-aware or sentient computers.

Session 1: Hardware Architecture Data Movement Capabilities

Fabric Data Movement: An Integrated Hardware/Software Challenge

Keith Underwood, *Intel Corporation*

Too often, hardware and software is decomposed into a distinct set of layers that are treated orthogonally. This locks the system into a vicious cycle, where applications optimize to existing hardware/software stacks, and hardware optimizes to slowly changing applications. In 2007, we began a research effort around Portals 4 to co-design the network API and hardware interfaces to meet the needs of HPC software. The effort evolved into a collaboration between Intel and Sandia National and eventually lead to a hardware architecture definition. Our efforts focused on creating API constructs that enable efficient messaging layer software while being implementable in hardware. MPI, SHMEM, and UPC were all prototyped in the hardware architecture phase of the program and led to substantial API and hardware architecture changes. The first portion of this talk will focus on lessons learned in the co-design process, and then provide an update on how those learnings have been contributed to the open community effort to develop a new network API.

Fundamental challenges still face both HPC and datacenter networks. These include aspects such as power management, the mapping of application topologies to network topologies, and congestion management. Many proposals today focus on inferring the user's intent: based on what can be observed by the hardware, adjustments are made to the hardware behavior. This process can be made more efficient by designing the full network stack along with the applications to better communicate how the application will leverage the fabric. We will explore a few network challenges, why it is difficult to infer user behavior for these cases, and how the application may be able to help improve system performance by providing additional information.

Finally, this talk will explore how some of the challenges facing large scale fabric have symmetries between the datacenter and HPC environments. We will provide a brief survey of some approaches to common problems and seek to provide insight into the question: why are the solutions different?

Session 1: Hardware Architecture Data Movement Capabilities

Exascale Starts with the Memory System

Bruce Jacob, *University of Maryland*

Energy and power costs are the primary reasons that our system-wide execution throughput (OPS) is no better than it currently is … while one could staple together 100 Oak Ridge Titans to create an exascale system, not many could afford to pay the resulting electric bill. And yet, a significant amount of research is still focused on increasing processor performance, rather than decreasing power and energy-to-solution. Modern high-performance systems are not throughput-bound; they are power-bound. Processing is so cheap it is effectively free; shaving power and energy costs at every opportunity, and at every level of the system, is the trick.

Similarly, memory and communication are the primary reasons that our time-to-solution is no better than it currently is … the memory system is slow; the communication overhead is high; and yet a significant amount of research is still focused on increasing processor performance, rather than decreasing (the cost of) data movement. Modern high-performance systems are not compute-bound; they are data-bound. ALUs are so cheap that some propose to put them out in the memory … processing is free; getting the right data to the right place, cheaply, is the trick.

Is Memory Scaling Really Over? Myths, Facts, and a Path Forward

Hillery Hunter, *IBM Corporation*

In recent years, pundits have repeatedly predicted doom for memory technologies -- first for NOR Flash, then NAND Flash, and finally DRAM. While NOR has indeed hit a scaling wall, NAND and DRAM seem to be going strong, but exascale demands are looming and it is key that the systems community drive scalability innovations for these technologies. This talk will explore some of the hype and realities behind memory scaling, and discuss memory system architecture which may alleviate the memory scaling wall and put the industry on a more sustainable path.

Session 1: Hardware Architecture Data Movement Capabilities

Extreme Scale Computing with Optically Enabled Data Movement

Keren Bergman, *Columbia University*

Performance scalability of next generation computing systems is becoming increasingly constrained by limitations in memory access, power dissipation and chip packaging. The processor-memory communication bottleneck, a major challenge in current multicore processors due to limited pin-out and power budget, becomes a detrimental scaling barrier to data-intensive computing. These challenges have emerged as some of the key hardware barriers to realizing the required memory bandwidths and system wide data movement. Recent manufacturing advances in silicon photonic interconnect and switching technologies are providing the infrastructure for developing energy-efficient high-bandwidth optical interconnection networks. Importantly, the insertion of photonics into next-generation computing systems is not a one-to-one replacement. This talk examines the design and potential impact of photonic-enabled architectures for creating new classes of future extreme scale computing.

Session 2: System Software Data Movement Capabilities

Rearchitecting Systems Software for Memory Heterogeneity and Scale

Ada Gavrilovska, *Georgia Institute of Technology*

Next generation exascale machines will include significantly larger amounts of memory, greater heterogeneity in the performance, persistence or sharing properties of the memory components they encompass, and increase in the relative cost and complexity of the data paths in the resulting memory topology. This poses several challenges to the systems software stacks managing these memory-centric platform designs. First, hardware advances in novel memory technologies shift the data access bottlenecks into the software stack. Second, current systems software lacks capabilities to bridge the multi-dimensional non-uniformity in the memory subsystem to the dynamic nature of the workloads it must support. In addition, current memory management solutions have limited ability to explicitly reason about the costs and tradeoffs associated with data movement operations, leading to limited efficiency of their interconnect use. To address these problems, next generation systems software stacks require new data structures, abstractions and mechanisms in order to enable new levels of efficiency in the data placement, movement, and transformation decisions that govern the underlying memory use. In this talk, I will present our approach to rearchitecting systems software and services in response to both node-level and system-wide memory heterogeneity and scale, particularly concerning the presence of non-volatile memories, and will demonstrate the resulting performance and efficiency gains using several scientific and data-intensive workloads.

Data Movement with MPI in a Many-Threaded World

Ryan Grant, *Sandia National Laboratories*

MPI is one of the most successful HPC data movement methods in scientific computing today. Until recently MPI implementations have primarily focused on a many-process model and have not provided high-performance support for a many-threaded model. Several changes are on the horizon for MPI in order to support foreseen explosion in the number of threads that will be used in future generation systems. These approaches can consist of allowing individual threads to become MPI endpoints, or allowing for better thread sharing of the existing "processes as endpoints" MPI model. The requirements for increasing levels of concurrency within MPI implementations themselves in response to the use of multiple threads are challenging. This talk will discuss the possible solutions to handling multi-thread concurrency in MPI, detailing the strengths and weaknesses of the current body of proposals.

Session 2: System Software Data Movement Capabilities

Active RDMA – New Tricks for an Old Dog

Torsten Hoefler, *ETHZ*

Remote memory access or partitioned global address space programming have been around for more than a decade. Their original idea was to allow put/get access to remote memory to enable a programming model similar to shared memory but with two explicit levels of locality. Remote direct memory access (RDMA) hardware enabled the basic put/get mechanisms at very high speeds. Will show with three examples how simple additional hardware functions can improve performance of various application classes significantly. For example, to implement a producer/consumer pattern, put/get is not sufficient but needs to be extended with a consistent notification mechanism. Furthermore, some simple but data-intensive computations can be moved to the data and can be processed by a handler on data access. Last but not least, one could envision global transactions similar to transactional memory which can accelerate irregular applications using optimistic concurrency control. After outlining these techniques and results, we want to stir the discussion to design and implement such features in hardware RDMA

Got Burst Buffer. Now What?

Adam Moody, *Lawrence Livermore National Laboratory*

Some supercomputer systems already provide fast, distributed storage between the compute nodes and the parallel file system, and more of these systems are coming in the near future. In this talk, I will discuss our early experience with and our future plans for this new capability. So far, we've found value in checkpoint / restart, extended memory, and improved application start up times, and we are optimistically researching potential gains in machine learning and new user-level, distributed file systems.

Dinner Speaker

How Facebook Uses Advanced Interconnect Technology to Make the World More Open and Connected

Katharine Schmidtke, *Facebook*

Facebook is growing quickly and now has more than 1 billion daily active users. Presenting the most relevant and delightful content to you requires terabits per second of data to be moved between hundreds of thousands of processors within a data center. Data movement for computing at this scale requires carefully selecting the latest technology innovations and optimizing performance for the lowest possible power consumption and cost. To do this, Facebook data centers are migrating to 100G optical interconnects while redefining the requirements originally set for the telecommunication industry.

Session 3: Data Movement from the Applications Perspective

An Early Start with Multi-Level Memory – An Opportunity for Performance or Just the Next Chapter In Programmer Hell?

Simon Hammond, *Sandia National Laboratories*

Application developers want it all - high performance, low power, large capacity memories with infinite reliability. Not much to ask for. The reality is that limitations on cost, power and fundamental hardware design make this extremely challenging. So far, the closest the industry can get to these expectations is the provision of multiple memories in a single node design with each having different performance characteristics. Such hardware designs are a radical departure from the last two decades of computing where developers have grown used to the expectation of a large, uniform performance memory space. It is no surprise then that the new approach to designing hardware is posing some complex challenges to existing algorithms.

In this talk I will describe some research efforts underway at Sandia to analyze application behavior and to port algorithms to new, multiple memory space hardware designs. We assess the success of these efforts as performance improvements against running either entirely in the slower memory since this houses the vast majority of problem state. Our experimentation utilizes simulation, emulation and in some cases, new prototype hardware systems.

Integrated Modeling for Rapid Assessment and Performance Prediction of HPC Applications

Jason Liu, *Florida International University*

Interconnection network is a critical component of high-performance computing architectures. For many scientific applications, the communication complexity poses a serious concern as it may hinder the scaling properties of these applications on novel high-end computing architectures. It is apparent that a scalable, efficient, and accurate model is essential for performance evaluation and architecture/application co-design. This talk will be divided into two parts. We will first take an introspective look into the current approaches on modeling the applications, the interconnection networks, and other important architectural elements for capturing data movement in computing at scale. We will then focus on the ongoing effort in creating a performance prediction framework that relies on rapid modeling at appropriate abstraction levels and using advanced parallel discrete-event simulation techniques to capture both architectural and algorithmic details of computation-physics code running at scale on high-performance computing platforms.

Session 3: Data Movement from the Application Perspective

Data Movement Challenges at NTS: A Paradigm for HPC Application I/O

Mark Miller, *Lawrence Livermore National Laboratory*

At the Nevada Test Site (NTS), tens to hundreds of megabytes of data from a nuclear explosion would be collected in a handful of micro-seconds. That's the equivalent of hundreds of terabytes/sec, enough to be the envy of any of today's top 500 machines. And, the DOE was doing this as far back as the 1950s. Looking back at how we met data movement challenges in real tests sheds light on possible ways to tackle those same challenges in the age of virtual testing and exascale computing. Not surprisingly, as HPC applications march ever closer to faithful, predictive simulations of real world experiments so too are data analysis tools and techniques beginning to look more and more like their real world instrumentation and methodological counterparts used to collect data from real experiments. This talk will give a historical perspective of data movement challenges at NTS and use the analogy to suggest a new paradigm for I/O in extreme scale computing.

Data Movement from the Applications Perspective

Sanjay Padhi, *Amazon*

In this modern era, where the underlying patterns of computation and data movement in applications are very different from those of conventional high-performance computation, data movement is highly correlated to the processing/consumption layer and thus the applications associated with it. In this talk based on the applications, data movements starting from continuous (Amazon Kinesis Streams, Amazon Kinesis Analytics, etc.) to live data in global storage elements (Amazon S3 based applications) modes will be outlined. In most applications datacentric processing is favored. Examples from multi-national distributed petascale computing at the Large Hadron Collider (LHC) to earth observation data, like Landsat 8, NEXRAD, etc. will be discussed, where not only the processing can be triggered based on an event (AWS Lambda), but also the Amazon SNS mechanism to broadcast the arrival of new data can be effective. Future prospects and implications of data sharing between systems and applications with elastic file systems on our national supercomputing centers will also be presented.

Session 4: Data Movement from the Data Analysis, Workflow, and Visualization Perspective

Workflow Analysis – An Approach to Characterize Application and System Needs

David Montoya, *Los Alamos National Laboratory*

Workflow has always been used to describe jobs and applications progressing and interacting with systems. As we move toward exascale, applications and systems are becoming more tightly integrated. The application stack and system environment can be seen as workflows built on workflows. We need a descriptive model to focus analysis and provide for comparison and contrast that begins to look like a map and a language for communication. With the diversity and complexity of exascale efforts, having a map and common language is a welcome base to build on.

In this talk I will describe an effort at LANL where we started by developing a workflow taxonomy with layers describing the application stack, how we have used it for initial assessment data usage patterns, and the potential to further describe lower layers that integrate with mapping of system layers of workflow. As this evolves it can develop into a communication layer to bridge the many focus areas we have across exascale efforts.

Session 4: Data Movement from the Data Analysis, Workflow, and Visualization Perspectives

Open, Reproducible HPC Data Analysis and Visualization: Minimizing Data Movement

Marcus Hanwell, *Kitware*

New challenges are emerging as supercomputer architectures become more diverse, and complex. The addition of GPGPU, many-core CPUs, burst buffers and in-situ analysis/visualization lead to the increased need for closer integration of the data analysis and visualization pipeline with simulation codes. Computational power is outstripping I/O bandwidth as we move towards exascale computing, and the importance of *in situ* processing coupled with strategies for performing processing in burst buffers is more pronounced. The ideal solutions will minimize data movement, both on-node and internode, providing opportunities to process at the point of generation, or in burst buffers before data is flushed to disk.

At Kitware we are working on a number of highly scalable, open source HPC solutions for data analysis and visualization. Well known projects such as ParaView and Catalyst offer solutions for post-processing, and *in situ* visualization. VTK-m offers a platform where scientists can develop a computational kernel once, and the framework will deploy this as a TBB, OpenMP or CUDA kernel. As we move forward it is critical that simulation developers engage with visualization and data analysis teams as they develop codes for next generation architectures in order to fully reap the rewards of these systems, apportioning funding to these efforts and more deeply engaging as data formats, movement, and simulation algorithms are designed, developed and deployed.

Session 4: Data Movement from the Data Analysis, Workflow, and Visualization Perspectives

What HPC Can Learn from the Cloud Approach to Big Data

Dale Southard, *NVIDIA*

Though high performance computing and cloud computing share similarities in hardware and scale, their approaches to computing and I/O have been diverging at an increasing rate over the last decade. Examining the root causes behind this divergence may help shape the design of future I/O, data reduction, and visualization solutions for HPC.

Extreme Data Management Analysis and Visualization for Exascale Supercomputers

Valerio Pascucci, *University of Utah*

Effective use of data management techniques for analysis and visualization of massive scientific data is a crucial ingredient for the success of any supercomputing center and cyberinfrastructure for data-intensive scientific investigation. In the progress towards exascale computing, the data movement challenges have fostered innovation leading to complex streaming workflows that take advantage of any data processing opportunity arising while the data is in motion.

In this talk I will present a number of techniques developed at the Center for Extreme Data Management Analysis and Visualization (CEDMAV) that allow to build a scalable data movement infrastructure for fast I/O while organizing the data in a way that makes it immediately accessible for analytics and visualization. In addition, I will present a topological analytics framework that allows processing data in-situ and achieve massive data reductions while maintaining the ability to explore the full parameter space for feature selection.

Overall, this leads to a flexible data streaming workflow that allows working with massive simulation models without compromising the interactive nature of the exploratory process that is characteristic of the most effective data analytics and visualization environment.

Session 5: Input/Output, File Systems, and Data Storage Data Movement Challenges

HPC Storage and IO Trends and Workflows

Gary Grider, *Los Alamos National Laboratory*

The Trends in computer memory/storage technology are in flux perhaps more so now than in the last two decades. Economic analysis of HPC storage hierarchies has led to new tiers of storage being added to the next fleet of supercomputers including Burst Buffers or In-System Solid State Storage and Campaign Storage. This talk will cover the background that brought us these new storage tiers and postulate what the economic crystal ball looks like for the coming decade. Further it will suggest methods of leveraging HPC workflow studies to inform the continued evolution of the HPC storage hierarchy.

Accelerating Science with the NERSC Burst Buffer Early User Program

Deborah Bard, *NERSC*

NVRAM-based Burst Buffers are an important part of the emerging HPC storage landscape. The National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory recently installed one of the first Burst Buffer systems as part of its new Cori supercomputer, collaborating with Cray on the development of the DataWarp software. NERSC has a diverse user base comprised of over 6500 users in 750 different projects spanning a wide variety of scientific applications, including climate modeling, combustion, fusion, astrophysics, computational biology, and many more. The potential applications of the Burst Buffer at NERSC are therefore also considerable and diverse.

I will discuss the Burst Buffer Early User Program at NERSC, which selected a number of research projects to gain early access to the Burst Buffer and exercise its different capabilities to enable new scientific advancements. I will present details of the program, in-depth performance results and lessons-learnt from highlighted projects.

Session 5: Input/Output, File Systems, and Data Storage Data Movement Challenges

From File Systems to Services: Changing the Data Management Model in HPC

Rob Ross, *Argonne National Laboratory*

HPC applications are composed from software components that provide only the communication, concurrency, and synchronization needed for the task at hand. In contrast, parallel file system are kernel resident, fully consistent services with semantic obligations developed on single core machines 50 years ago; parallel file systems are old-fashioned system services forced to scale as fast as the HPC system. Rather than the monolithic storage services seen today, we envision an ecosystem of services being composed to meet the specific needs of science activities at extreme scale.

In fact, a nascent ecosystem of services is present today. In this talk we will discuss drivers leading to this development, some examples in existence today, and some steps we can take to accelerate the rate at which these services are developed and mature to meet application needs.

Breaking Free from Globally Serializable OLTP Embedded in Parallel File Systems

Garth Gibson, *Carnegie Mellon University*

Decades ago the HPC community wisely decided to base massively parallel coordination on user-level libraries and middleware, asking the underlying operating system to become more stable and efficient within the scope of a single machine, but not become an ever more parallel distributed operating system. A key exception to the banishing of distributed systems from the operating system kernel was the parallel file system. File system access is still mostly mediated by each machine's operating system, forcing the parallel file system to be a kernel-level distributed system scaling as fast as the overall cluster. A decade ago HPC file systems did at least decouple data movement from namespace metadata processing, delaying the file server bottleneck for years, but at its core, every namespace metadata operation is handled the same way an online transactional database system would — a globally serialized order is determined on the fly for all concurrent operations. In our recent SC14 best paper, we parallelized file server functions to offer another two orders of magnitude scaling for file system namespace metadata processing, but we did not remove the global serialization of concurrent operations. In this talk we will propose the total elimination of the traditional file system server and an escape from on-the-fly globally serialized namespace metadata operations, promising perhaps many more orders of magnitude on the scaling of data movement in the HPC storage systems.

Conference Notes

Conference Notes