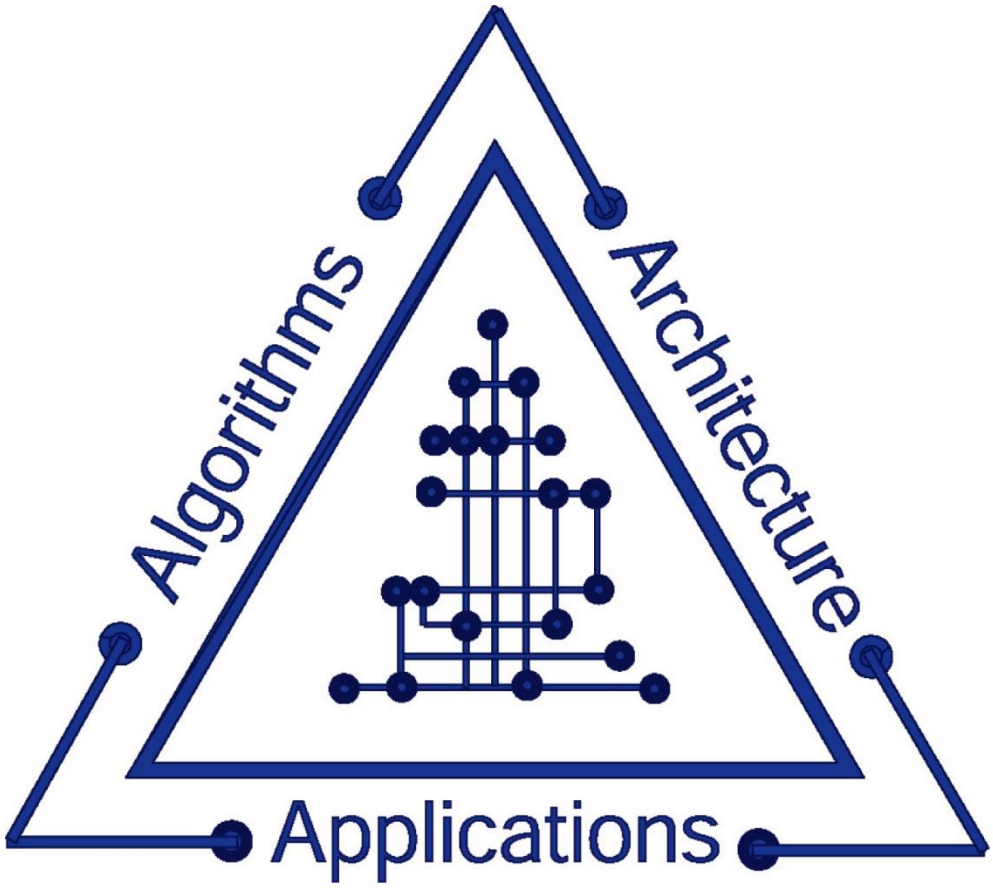
The Salishan Conference on

HIGH SPEED COMPUTING



April 24 – 27, 2017

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: Perspectives on HPC's Current *Cambrian Explosion*. The speakers have been selected to address our theme and give attendees information about the latest technologies and issues facing high performance computing (HPC). 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 and/or complete the electronic survey at the end of the conference (http://salishan.ahsc-nm.org/2017Survey.html).

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

Conference Committee: Kim Cupps and Katie Lewis, *LLNL*
Jim Ang and Ron Brightwell, *SNL*
Carolyn Connor and 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 Dee Cadena, Jan Susco 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.

Visit our website at: **http://salishan.ahsc-nm.org**

Next Conference Dates: April 23-26, 2018 April 22-25, 2019 April 27-30, 2020

Main Lodge Map

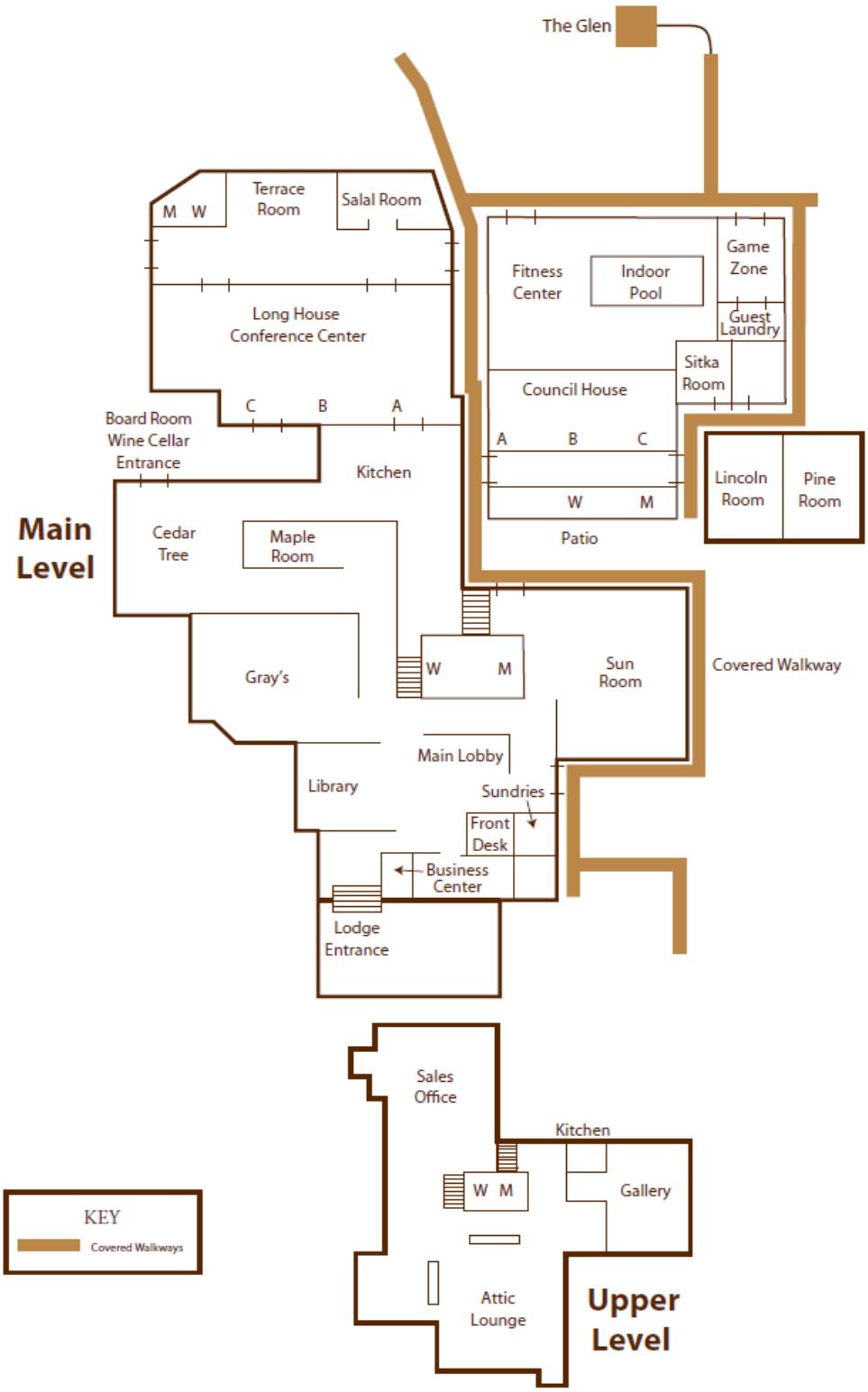


Table of Contents

Welcome and Logistics 1

Lodge Map 2

Sponsorship 5

Conference Theme 7

Conference Program

Monday: Keynote Address 11

Tuesday: Session 1: Many-Core Computing – Application Challenges and Trinity/Cori Specifics 12

Session 2: Heterogeneous Computing – Application Challenges and Sierra/Summit Specifics 13

Wednesday: Session 3: Neuro-Inspired Computing 15

Thursday: Session 4: Quantum Computing 18

Session 5: Crosscutting and Integration Topics 19

Abstracts 21

Attendees 37

Conference Notes 45

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 sessions:

Advanced Micro Devices, Inc.

ARM

Cray, Inc.

D-Wave Systems, Inc.

DDN Storage

Dell EMC

Hewlett Packard Enterprise

IBM Corporation

Intel Corporation

Micron Technology, Inc.

NVIDIA Corporation

Penguin Computing

Seagate Government Solutions

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

THIS PAGE LEFT BLANK
INTENTIONALLY

Conference Theme

Perspectives on HPC's Current *Cambrian Explosion*

The 2017 Salishan Conference on High Speed Computing will explore the explosion in diverse computing models that have arisen in recent years. The traditional evolution of HPC will be examined in sessions on many-core computing and heterogeneous node computing. We will also examine some computing models that are characterized as *Beyond Moore Computing*, but include two very different models on computing in sessions on neuro-inspired computing and quantum computing. We are presently in an era of tremendous HPC architectural diversity, analogous to the *Cambrian Period*, ~490-540 million years ago, when there was a major explosion in the diversity of life on earth. This was when simple cell-based life in the primordial soup gave way to an explosion of life forms.

Shifting to near-term history, the Salishan Conference was established in 1981 to serve as a forum for our community to share lessons learned and technology challenges for the advent of Cray vector supercomputers. This was followed by a period of relative stability in supercomputing, as Cray's custom vector supercomputers became the *de facto* standard. In the 1990s, the *Attack of the Killer Micros* hit our community and there was a period of disruption in supercomputer architectures as performance increases in Cray supercomputers stalled. Many approaches to parallel architectures were explored with DOE and DoD support, and from this architectural diversity, the ASCI program helped establish massively parallel processor systems with explicit message passing as the dominant supercomputing model for the next 15+ years. This was a golden age for performance improvement when Moore's Law and Dennard Scaling allowed for easy application performance increases from doubling of transistor counts and attendant increases in processor frequencies.

The end of Dennard Scaling has led us to multi-core processors, many-core processors and GPUs. But increases in peak performance from these processor architectures are often decoupled from performance increases for our real applications due to data movement bottlenecks. New types of data analytic applications have also arisen with very different models of computing from our traditional scientific and engineering modeling and simulation drivers. Both of these factors have given rise to the increase in HPC architectural diversity that is the focus for this year's Salishan Conference. In this conference, we want to explore the key technical issues and tradeoffs that arise among the Application, Algorithm and Architecture perspectives for this diverse collection of computing models.

Session 1: Many-Core Computing – Application Challenges and Trinity/Cori Specifics

The realities of the utilization wall combined with the emergence of stringent application constraints, particularly those linked to energy consumption, have necessitated new system architectural strategies (e.g., many-core and heterogeneous HPC systems) and real-time operational adaptability approaches. Such complex systems require new and powerful design and programming methods to ensure optimal and reliable operation. The Trinity and Cori computing systems are two examples. Both utilize the Intel Knights Landing processor, a self-hosted, many-core processor with on-package high-bandwidth memory that delivers more than 3 teraFLOPS of double-precision peak performance per single socket node, higher intra-node parallelism, and longer hardware vector lengths. While these enhanced features provide opportunity for significant performance improvements, fully capitalizing on the power of many-core architectures comes with increased burden and complexity. This session will focus discussion on the key challenges facing application developers on many-core architectures. It will examine the implications of increasing degrees of parallelism at the compute node level, as well as increasing memory level complexity (e.g., on-package Multi-channel dynamic random access memory (MCDRAM) and off-package DIMMs). The driving question is: What are the key technical challenges facing application developers, and how can these challenges be addressed?

Session 2: Heterogeneous Computing - Application Challenges and Sierra/Summit Specifics

The move to heterogeneous compute architectures is driven by the need to model complex systems at finer levels of resolution and accuracy while minimizing the total cost of ownership of the massive machines required to run these important problems. The challenge of fully realizing the processing power improvements promised by these hybrid architectures lies in our ability to successfully adapt our codes, algorithms and tools to fully exploit the advantages of the GPU without increasing data movement and transferring so much complexity to the programmer that productivity losses outstrip computing power increases. Sierra and Summit, planned to be running at LLNL and ORNL in late 2018 are DOE's first foray into computing platforms of over 100 Petaflops using a heterogeneous architecture. This session will describe specific efforts related to adapting codes and algorithms and developing tools to increase efficiency of the codes running on Sierra and Summit and of the programmers who write them. Further, this session will also describe future specialized processors being developed to continue our quest for increased simulation capability at sustainable costs. The driving questions behind this session are: What are the key problems facing developers and what is being done to address them, and what future specialized processors can help us continue to meet the demands of our most pressing problems?

Session 3: Neuro-Inspired Computing

This session will explore the theory and applications of neuro-inspired computing to solve traditionally complex problems and extend past the current limits of von Neumann architectures. Although the concept of computational models for neural networks or artificial neural networks have existed for over 70 years, the last decade has seen rapid growth in applications, from credit card fraud to image and video analysis. The use of artificial neural networks and, more generally, the use of learning algorithms is a game-changer in modern algorithms, yet harnessing this power is not always evident. New, neuro-inspired hardware, like IBM's TrueNorth chip, gives us a platform to expand even further beyond previous constraints of energy usage and parallelism on a large scale. With the expansion of interest in neuro-inspired computing beyond academia, and the development of larger-scale platforms, DOE laboratories are evaluating applications of this hardware to meet mission needs. We face many challenges in this goal, from evaluating the programmability and integration potential for these architectures to the paradigm shift of recognizing the usefulness of inexact solutions for certain applications. This session will address the following questions: What influences have advancements in machine learning had on neuro-inspired architectures? How do current neuro-inspired architectures meet or deviate from theoretical models? How can we use neuro-inspired computing to solve our challenges? How can this technology affect the way we design computational algorithms?

Session 4: Quantum Computing

Quantum computing has recently been receiving a lot of attention due to the efforts by IBM, Google, Microsoft, Intel, D-Wave, Rigetti, and various academic institutions and national laboratories to realize quantum computers and make them practical to use. Some of these efforts concern "real" quantum computers while others merely exploit quantum effects to solve classical problems. In all cases, new algorithms and programming techniques must be developed because existing ones do not port directly to these fundamentally different machines.

Naturally questions about the class of algorithms and hence type of problems that offer improvements over classic computers arise. This session will touch on the different physical implementations of quantum computers as well as the new algorithm developments needed to support these. And last, but not least, early investigations into using quantum hardware will be shown. Questions: What are the trade-offs when moving beyond traditional HPC systems to this new model of computing? Will gate-model or annealing-model dominate the future? When will quantum computing be practical for HPC? How do we move today's programmers to the new paradigm?

Session 5: Crosscutting Fragments

New adaptations and increased complexity are two important characteristics associated with the Cambrian explosion and both are key considerations related to the increase in HPC architectural diversity as well. This session will touch on these cross-cutting areas through a loosely-coupled set of topics. New adaptations in computing technology may require alternative models of computation, such as approximate and probabilistic computing approaches, in order to fully exploit new hardware capabilities. Memory technology will also need to adapt to keep pace with the advancements in processing technology. As large-scale systems evolve to be composed of more diverse hardware organisms and applications begin to explore and embrace these new computational models, complexity increases significantly, and achieving the desired levels of performance likely requires advancements in methods and tools for performance analysis. This session will address the following questions: How will applications and application development be impacted by these new models of computing and extended memory addressing capability? What algorithms are most appropriate for probabilistic and approximate computing models? Can integrated performance tools reduce the complexity of tuning and adapting codes to new node and system architectures?

Conference Program

Perspectives on HPC's Current *Cambrian Explosion*

Monday, April 24, 2017

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

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

Title: **Bits, Qubits, Neurons: Their Connections and the Near Future of Computing**

Speaker: William Camp, *Camp Research*

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

Tuesday, April 25, 2017

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

Breakfast available (Terrace)

8:30 am **Session 1: Many-Core Computing – Application Challenges and Trinity/Cori Specifics**
Chair: Carolyn Connor

Title: Intel's Many-Core Journey: Past, Present and Future
Speaker: Alan Gara, *Intel Corporation*

Title: The Impact of Increasing Memory System Diversity on Applications
Speaker: Gwen Voskuilen, *Sandia National Laboratories*

Title: Trinity Center of Excellence: "I Can't Promise to Solve All Your Problems, But I Can Promise You Won't Face Them Alone"
Speaker: Hai Ah Nam, *Los Alamos National Laboratory*

10:00 am **Break**

Refreshments available (Terrace)

10:30 am **Title: Many-Cores for the Masses – Moving the NERSC Workload to Knights Landing and Beyond**
Speaker: Jack Deslippe,
Lawrence Berkeley National Laboratory/NERSC

11:00 am **Panel Discussion**

Tuesday, April 25, 2017

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

1:30 pm **Session 2: Heterogeneous Computing – Application Challenges and Sierra/Summit Specifics**
Chair: Mark Gary

Title: Portable Performance with OpenMP in a Heterogeneous Era
Speaker: Tom Scogland,
Lawrence Livermore National Laboratory

Title: Heterogeneous Computing Challenges & Directions
Speaker: John Danskin, *NVIDIA Corporation*

Title: Principles of Porting Disorder Therapy for the HPC Clinician: An Introduction
Speaker Fernanda Foertter,
Oak Ridge National Laboratory

3:00 pm **Break**

Refreshments available (Terrace)

3:30 pm **Title: A Case for Intelligent (Co-)Design**
Speaker: Sean Treichler, *NVIDIA Corporation*

4:00 pm **Panel Discussion**

Tuesday, April 25, 2017

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

Title: **Winemaking: The Art and Science of Rotting Grapes to Perfection**

Speaker: James Osborne, *Oregon State University*

8:00 pm **Reception and Informal Discussions**

(Immediately following the Working Dinner in the Cedar Tree Room)

Wednesday, April 26, 2017

8:00 am **Introduction to Sessions**

Breakfast available (Terrace)

8:30 am **Session 3: Neuro-Inspired Computing**
Chair: Katie Lewis

Title: **The Brain's Circuits Suggest Computing with High Dimensional Vectors**
Speaker: Pentti Kanerva, *University of California, Berkeley*

Title: **TrueNorth Ecosystem**
Speaker: Dharmendra Modha, *IBM Corporation*

Title: **Exploring Neuromorphic Computing at Lawrence Livermore National Laboratory**
Speaker: Brian Van Essen,
Lawrence Livermore National Laboratory

10:00 am **Break**

Refreshments available (Terrace)

10:30 am **Title:** **Neural-Inspired Computing Algorithms and Hardware for Image Analysis and Cybersecurity Applications**
Speaker: Conrad James, *Sandia National Laboratories*

11:00 am **Panel Discussion**

Wednesday, April 26, 2017

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

1:30 pm **No Scheduled Sessions**

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

Based on feedback from survey respondents, we have decided to try something new for the 2017 Random Access session. Rather than having a "first come, first served" sign-up board, we want to make the selection of talks more interactive and reflect the interests of the whole audience.

We invite Salishan 2017 attendees who want to give a Random Access talk to sign-up online (maximum of one talk per participant) at: **http://salishan.ahsc-nm.org/2017RandomAccess.html**. Only a talk title and abstract tweet (max. 140 characters) are needed at time of sign-up. Alternatively, sign-ups can be done in person with the conference organizers. Sign-ups close Tuesday at 8:00 pm. *In fairness to at-large attendees, invited session speakers are requested to not submit proposals for Random Access talks*.

On Wednesday morning, a comprehensive list of talk titles and tweet abstracts will be published online for voting (an email will be sent to attendees with the link to vote). Every attendee will be allowed three equally-weighted votes until polls close on Wednesday at 1:00 pm. The talks with the most votes will be presented in the Random Access session starting on Wednesday at 5:00 pm. The final agenda of talks will also be posted online by Wednesday at 3:00 pm.

Wednesday, April 26, 2017

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.

Ryan Bleile, *University of Oregon, Eugene*

Nikoli Dryden, *University of Illinois at Urbana-Champaign*

Simon Garcia De Gonzalo, *University of Illinois Urbana-Champaign*

Georg Hahn, *Imperial College London*

William Killian, *University of Delaware*

Sheng Lundquist, *Portland State University*

Ciaran Ryan-Anderson, *University of New Mexico*

Felix Wang, *University of Illinois at Urbana-Champaign*

Yuliana Zamora, *University of New Mexico*

Thursday, April 27, 2017

8:00 am **Introduction to Sessions**

Breakfast available (Terrace)

8:30 am **Session 4: Quantum Computing**
Chair: Christoph Junghans

Title: **Quantum High Performance Computing**
Speaker: Matthias Troyer, *Microsoft Research*

Title: **Quantum Computing: Cladogenesis Beyond Exascale HPC**
Speaker: Andrew Landahl, *Sandia National Laboratories*

Title: **Quantum Annealing at NASA: Current Status**
Speaker: Rupak Biswas, *NASA Ames Research Center*

10:00 am **Break**

Refreshments available (Terrace)

10:30 am **Title:** **What to Do with More Than 1000 Quantum Bits**
Speaker: Scott Pakin*, Los Alamos National Laboratory*

11:00 am **Panel Discussion**

Thursday, April 27, 2017

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

1:30 pm **Session 5: Crosscutting Fragments**
Chair: Ron Brightwell

Title: **128 Bit, Exascale Memory Reference Models for Next Generation, ExaByte Capacity Physical Memory Systems**
Speaker: Steven Wallach, *Micron Technology, Inc.*

Title: **Cross-Stack Approximate Computing for Modern Applications and Future Substrates**
Speaker: Luis Ceze*, University of Washington*

Title: **Probabilistic Computing in the Post-Moore's Era**
Speaker: Laura Monroe, *Los Alamos National Laboratory*

3:00 pm **Break**

Refreshments available (Terrace)

3:30 pm **Title:** **Towards Always-On Integrated Performance Analysis Tools**
Speaker: Matthew LeGendre
Lawrence Livermore National Laboratory

4:00 pm **Panel Discussion**

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

2017 Salishan Conference on High Speed Computing

SURVEY

WE VALUE YOUR INPUT!

Thank you for participating in this year's conference.
It would be appreciated if you can take a few minutes to complete this brief online survey:

http://salishan.ahsc-nm.org/2017Survey.html

Abstracts

Keynote Address

Bits, Qubits, Neurons: Their Connections and the Near Future of Computing

William Camp
Camp Research

The world's enterprises run on bits—von Neumann/Boolean (VNB) computing. As we seek to continue the exponential growth of computing capability, we are approaching an asymptote in our race to the bottom: near atom-scale computing. In our current approach, Silicon CMOS-based computing, we are also finding it increasingly difficult to continually decrease the energy cost per computing operation. Some see this as portending the end of the run for CMOS and Silicon.

At the same time, our intellects apparently run on a fundamentally different approach to computing: some form of what we term neural computing. With the advent of relatively inexpensive computing capabilities for training them, neural networks are moving from a theoretical and intellectual challenge to a practical method to attack many problems that are resistant to VNB approaches. Indeed, there is some basis to think that neural approaches may inevitably broaden their domain, for example to overcome hard combinatorial problems.

Today, quantum logic circuitry and quantum relaxational computing are showing signs of overcoming hard issues sufficiently to allow quantum approaches to succeed where other methods fail for at least some hard problems.

It is not a reach to posit that the first two of these three approaches will play critical roles in exascale computing and beyond. While the future of quantum approaches is still highly uncertain, they remain at the forefront of new directions in computing R&D. From a physics and mathematical viewpoint, there are rather fascinating connections among classical statistical mechanics, quantum information theory and neural computing. I will at least hint at those connections in this talk.

Session 1: Many-Core Computing-Application Challenges and Trinity/Cori Specifics

Intel's Many-Core Journey: Past, Present and Future

Alan Gara
Intel Corporation

In this talk, Al Gara will discuss the challenges past, present and future in many-core processor design. He will also discuss the likely direction to reach exascale through traditional silicon technology based on balancing memory, processor and fabric and optimizing for perf/$. This optimization will include constraining the system to provide exceptional performance for both legacy workloads and workloads where extensive modification is possible to optimize for architecture improvements. The architectural trade-offs to provide such system with relatively broad applicability versus a more focused machine tailored to those willing to do significant work on their applications will also be discussed.

The Impact of Increasing Memory System Diversity on Applications

Gwen Voskuilen
Sandia National Laboratories

The simultaneous rise of many new memory technologies, none of which appear to be a drop-in replacement for DDR DRAM, is leading to increasing diversity within memory systems. While a diverse, multi-level memory system may promise higher performance, higher capacity, and lower cost than a single-level memory, realizing this promise requires intelligently managing data placement and movement among the different levels to ensure that the data use matches the characteristics of the memory level. Doing this in the face of memory technologies whose differences span a wide range of metrics (bandwidth, capacity, latency, non-volatility, etc.) and for diverse applications with distinct memory behaviors, poses a significant challenge. Adding to the challenge, different management approaches can have vastly different impacts on application performance, hardware cost, programmer effort, and system complexity.

In this talk I'll describe some recent work at Sandia to explore, in the context of Trinity, the application impact of different multi-level memory management techniques. We evaluated strategies along two axes – hardware versus software-driven management and static versus dynamic allocation. Additionally, we analyzed the memory behavior of several mini-apps from the DOE APEX benchmarks suite and looked at how the behavior can affect the viability of different management approaches.

Session 1: Many-Core Computing-Application Challenges and Trinity/Cori Specifics

Trinity Center of Excellence: "I Can't Promise to Solve All Your Problems, But I Can Promise You Won't Face Them Alone"

Hai Ah Nam
Los Alamos National Laboratory

In the days of yesteryear, yore and yonder - pre GPGPU and many-core processors like the Intel Xeon Phi Knights Landing - many application developers could achieve increased performance with minimal code development effort from the increase in processing power on the node, scale of the system and system software. However, the lesson to be learned on current generation petascale systems is, "what you put into it, is what you will get out." Application teams are being challenged to make significant investments to their codes to achieve performance by exploiting new architectural features. The Trinity system with over 9500 KNL processors and an equally sizable Intel Xeon Haswell partition, presents significant on-node challenges, including increasing parallelism to use the increased number of cores and threads, enabling (or not hindering) compiler vectorization with AVX-512 instructions, and identifying data structures that will benefit from residing in high bandwidth memory and explicitly managing the memory hierarchy. The Trinity Center of Excellence (COE) was established to provide a long-term collaboration between vendor subject matter experts and application developers to jointly tackle the challenges of efficiently using this new architecture. This presentation describes some of the co-design best practices from the Trinity COE. Focusing more human effort, with diverse views and experience, brought together through the COE is key to creating long-lasting investments in code development for current and future generation system.

Session 1: Many-Core Computing-Application Challenges and Trinity/Cori Specifics

Many-Cores for the Masses - Moving the NERSC Workload to Knights Landing and Beyond

Jack Deslippe
Lawrence Berkeley National Laboratory/NERSC

How is optimizing science applications for Knights Landing like an ant farm? You'll find out in this session. NERSC is actively deploying the Cori HPC system (acquired in collaboration with the Trinity system) with over 9,000 Knights Landing Xeon-Phi processors. Preparing user applications for this system has been one of the main focuses for the center for the past two years. This session will describe the optimization strategy developed for porting a diverse set of science codes to Knights Landing, as well as highlight progress and lessons learned from the optimization process on 20+ applications associated with the NERSC Exascale Science Application Program (NESAP). I will describe how applications use the novel features of the Knights Landing architecture, including the MCDRAM and associated NUMA/Memory/Cache Modes, the wide vector units, and the 68 cores per processor. We also discuss challenges associated with porting apps to the architecture, how applications overcome them, and how we have learned to engage the "masses" in a productive conversation about application performance.

Session 2: Heterogeneous Computing – Application Challenges and Sierra/Summit Specifics

Portable Performance with OpenMP in a Heterogeneous Era

Tom Scogland
Lawrence Livermore National Laboratory

As the diversity of our systems increases, so too does the complexity of providing programming models that efficiently abstract the wide variety of target hardware. OpenMP has grown from its shared-memory multithreaded roots to include an offload model for heterogeneous systems as well as support for discrete memory, but there remains a long way to go to provide portable performance across the vast diversity of target systems. Ensuring that data is moved efficiently is a particularly critical task. As the proliferation of memory architectures expands to include not only discrete accelerator memories, but multiple host memories with differing performance characteristics, we require more powerful abstractions not only to select the appropriate memories, but to express the structure of complex data. At the same time, application developers desire ease of use, and the ability to leverage the programmability of their target hardware and the newest features of both the base language and higher-level libraries. These conflicting goals present a significant challenge. In this talk I will describe some of the research and design efforts underway to provide developers with the expressiveness they require to handle complex problems, as well as our efforts to keep solving the simple problems as simple as possible.

Session 2: Heterogeneous Computing – Application Challenges and Sierra/Summit Specifics

Heterogeneous Computing Challenges & Directions

John Danskin
NVIDIA Corporation

The premise of heterogeneous computing is a reflection of Amdahl's law. Computation is a mix of serial and parallel sections. Serial sections should run on a fast serial, or latency optimized, processor. Parallel sections should run on an efficient parallel throughput processor. Homogeneous solutions strain to balance serial performance and efficiency. Heterogeneous solutions optimize the two regimes separately, potentially achieving better efficiency in each. Heterogeneity introduces the issues of communication, data management, and balance. Summit and Sierra address communication and data management with a high-bandwidth local interconnect supporting memory coherence and synchronized virtual memory. The balance of high performance throughput and serial computing is addressed by customizing the balance between latency optimized and throughput processors for each installation. The high speed local communication required for multi-chip heterogeneous computing introduces the opportunity for "fat nodes". Fat nodes can reduce the number of network targets, which benefits many communication patterns including all to all. Speculative possible future directions could include integration of latency optimized processing, disaggregation of latency optimized computing, and tighter integration of throughput accelerators and system networks.

Principles of Porting Disorder Therapy for the HPC Clinician: An Introduction

Fernanda Foertter
Oak Ridge National Laboratory

Porting Disorder refers to a condition characterized by delays in the development and porting of applications to newer architectures. Symptoms include difficulty with changes in architectures, lack of motivation to use modern development approaches, distortions in the physical reality of Moore's Law, and irrational beliefs in magical compilers. This talk will cover novel approaches in helping users overcome porting anxiety using the latest in pseudo-psychological methods. We will discuss how to build the ideal HPC ecosystem where users and their apps can thrive. Finally, we will cover how HPC centers can move beyond emotional arguments over deployment of heterogeneous architectures.

Session 2: Heterogeneous Computing – Application Challenges and Sierra/Summit Specifics

A Case for Intelligent (Co-)Design

Sean Treichler
NVIDIA Corporation

By geological standards, the early Cambrian period was incredibly productive. This hyper-inflationary period increased biological diversity by several orders of magnitude and saw the emergence of most of today's animal phyla. Architectures for high-performance computing are undergoing an analogous diversification today. However, by human standards the forces of random mutation and natural selection that drove the Cambrian expansion are unbearably slow and appallingly wasteful. We can't afford to spend 20 million years to get to exascale. The challenge is increased by the fact that we are interested not simply in the proliferation of high-performance computing hardware, but of larger organisms that incorporate hardware, runtime and operating systems, compilers, libraries, and application code, each filling a particular niche in the scientific computing ecosystem. The irreducible complexity of these organisms precludes most meaningful changes to one part of an organism as they are likely to cause other parts to cease functioning.

An alternative is to take the wheel ourselves, using our intelligence to design the optimal organism for each niche. While it may be tempting to simply start from scratch, the complexity of the organisms and the number of niches that must be filled makes this approach intractable.

Instead, I will argue for a co-design approach that guides evolutionary forces by focusing on the irreducibly complex mechanisms in current organisms as fulcrums for change. The effectiveness of a mechanism for a given niche (or collection of niches) can be assessed, and if a superior approach exists, the mechanism can be replaced through an intervention that changes all the impacted parts of the organism simultaneously.

I will offer several examples of this approach in action, as well suggesting a number of other areas that are ripe for this sort of intervention. And while co-design efforts are not free, I will argue that they represent a critical investment for the future of high-performance computing, and that now is the time to be making that investment.

Dinner Speaker

Winemaking: The Art and Science of Rotting Grapes to Perfection

James Osborne
Oregon State University

In some ways, the process of turning grapes into wine has changed little in thousands of years. Sugars in the grape are converted into alcohol by yeast and human civilization enjoys the resulting product. However, in many other ways winemaking in the 21st century bears little resemblance to how wine was originally made. As science has advanced, so has our understanding of the chemical and biological processes that occur during the winemaking process, and how these processes can be managed to produce certain styles and qualities in a wine.

Advances in agriculture have helped us determine what grape varieties should be grown in what geographical areas in order to produced grapes of the highest quality. Oregon is an excellent example of this progress. While it was once thought that quality wine grapes could never be grown in such a wet and cool climate, Oregon now has a thriving industry based on the production of Pinot Noir wines rivaling the finest French wines. While Oregon cannot compete on a volume base with wine regions such as California, the industry succeeds because of its focus on the production of ultra-premium quality wines. In this way research and development has been critical in the rise of the Oregon wine industry and continues to drive the industry forward.

Understanding the many factors that impact wine quality and how to optimize these is the ongoing challenge for the industry, and where research at the Oregon Wine Research Institute is focused. Faculty with expertise ranging from soil microbiology through analytical chemistry approach the research questions in a collaborative approach, where the goal is to understand the process from “Vine to Wine.”

Session 3: Neuro-Inspired Computing

The Brain's Circuits Suggest Computing with High-Dimensional Vectors

Pentti Kanerva
University of California, Berkeley

Traditional computing is deterministic. It is based on bits and assumes that the bits compute perfectly. However, (near) perfection at high speeds consumes large amounts of energy and limits the scaling-down of the underlying circuit elements. Contrast this with the brain's powers of perception and learning. They go far beyond what computers can do, are accomplished with little energy by slow and imprecise neurons wired together according to a general plan, with many details left to chance. Can we make sense of that kind of computing and build systems on the same principles?

Computing with high-dimensional vectors (hypervectors, ultrawide words) is based on the idea that the brain's powers are rooted in the mathematics of very-high-dimensional spaces – when the dimensionality is in the thousands-to-tens-of-thousands. The very size of the brain's circuits suggests very high dimensionalities

Information encoded into a hypervector is distributed equally over all vector components – the vectors are not subdivided into fields. Operations on such vectors are the key to computing with them. Traditional computers work with bits and numbers, and their most important operations are built into the hardware. These include circuits for the addition and multiplication of numbers. Similar operations are used with hypervectors: one corresponds to addition and another to multiplication. A third operation simply permutes the coordinates and a fourth computes the similarity of vectors.

It is possible to build very rich systems of computing based on these four operations. Their practical value follows from several factors: (1) simple operations yield vectors that are machine-learning friendly; (2) the operations are easily distributed and performed in parallel; (3) computation is extremely robust even with unreliable low-resolution components; and (4) the prerequisite circuits are a natural fit to next-generation nanotechnology. In the field of computing at large, high-dimensional computing combines symbolic computing and neural-net computing/deep learning, and it complements numeric computing. It has been demonstrated with semantic vectors, language identification, and EEG signal classification, among other things, and research into the theory and its application is ongoing.

Session 3: Neuro-Inspired Computing

TrueNorth Ecosystem

Dharmendra Modha
IBM Corporation

2016 was a breakthrough year for TrueNorth Ecosystem with the Misha Mahowald Prize, induction into the Computer History Museum, and the R&D Magazine Scientist of the Year. I will describe progress that includes:

- Delivery of single-chip NS1e systems to Army Research Lab
- Delivery of scale-out NS1e16 systems to Air Force and Army Research Labs
- Delivery of scale-up NS16e system to Lawrence Livermore National Lab
- End-to-end demonstration of TrueNorth with Samsung and iniLabs DVS sensors
- Development of placement algorithms for scale-up NS16e systems
- Development of convolutional networks for TrueNorth with > 6000 frames/sec/ Watt on CIFAR 100
- Real-time streaming interface to TrueNorth
- Applications by 160+ developers at 40+ universities and government agencies

Finally, I will provide a glimpse into 2017 directions.

Exploring Neuromorphic Computing at Lawrence Livermore National Laboratory

Brian Van Essen
Lawrence Livermore National Laboratory

Lawrence Livermore National Laboratory is exploring the application of Neuromorphic computing to problems of interest within the Advanced Simulation and Computing (ASC) program. This effort is part of the larger DOE Beyond Moore's Law Computing Program. In this talk, I will give an overview of several laboratory applications, describe our efforts to map them to the TrueNorth architecture, and showcase our initial results. These applications range from object recognition in cluttered images, to detecting numeric instabilities in ALE-based simulation codes, to exploring heuristic graph optimization problems.

Session 3: Neuro-Inspired Computing

Neural-Inspired Computing Algorithms and Hardware for Image Analysis and Cybersecurity Applications

Conrad D. James
Sandia National Laboratories

As traditional numerical computing has faced challenges due to the end of Dennard scaling and increasing power budgets for new generation supercomputers, researchers have turned towards alternative computing approaches to reduce power-per-computation metrics and to develop more efficient algorithms. Neural-inspired computing has matured into a field that has produced promising methodologies for solving difficult problems including pattern recognition and anomaly detection. Yet the challenges with neural-inspired computing are well-known: the need for large amounts of training data, the relative brittleness of algorithms to data variability, and the ad-hoc nature of optimization techniques. We have taken an approach to address these concerns by strengthening the connection between machine learning and neuroscience concepts – termed neural machine learning – in order to improve the resiliency and adaptability of such algorithms.

The Hardware Acceleration of Adaptive Neural Algorithms (HAANA) project at Sandia National Laboratories is developing neural machine learning algorithms and hardware for applications in image processing and cybersecurity. While machine learning methods are effective at extracting relevant features from many types of data, the effectiveness of these algorithms can diminish over time in real-world environments. Our team has generated several approaches that support continual adaptation of algorithms to changing data and that incorporate unsupervised learning to reduce the reliance on subject-matter experts to manually craft features of interest. An example is the dynamic incorporation of new processing nodes into deep learning neural networks. Neurogenesis deep learning enables trained networks to learn new data patterns with minimal loss to previously learned data patterns - a critical advancement towards generating truly adaptive algorithms that can operate in real-time.

In addition to developing algorithms, we are also designing several hardware architectures to enhance learning and performance in a suite of spiking (i.e., time-encoded) and non-spiking neural-inspired algorithms. We have designed a non-von Neumann architecture to implement

complex temporal dynamics in spiking algorithms such as liquid state machines, and this led to an improvement in classification accuracy in speech data. To accelerate training in algorithms such as deep learning, we used a resistive memory crossbar architecture to parallelize prevalent computational kernels such as vector-matrix multiplication. In order to quantitatively evaluate the impact of architecture design on algorithm performance, we constructed a multiscale simulation framework that couples circuit-level architecture characteristics with microelectronic device properties and algorithm tuning parameters. Also, included within our research is the fabrication of novel microelectronic devices that are specifically tuned for neural-inspired computing applications. Our non-volatile memory redox transistor is a device capable of fine-resolution multi-state analog switching, a feature that is beneficial for high accuracy in algorithm performance. Using these devices in a simulation of hardware acceleration, we predicted minimal degradation in algorithm performance accuracy as compared to 32-bit floating point numerical accuracy.

Session 4: Quantum Computing

Quantum High Performance Computing

Matthias Troyer
Microsoft Research

A century after the development of quantum mechanics we have now reached an exciting time where computational devices that make non-trivial use of quantum effects can be built. Quantum random number generators, analog quantum simulators and quantum annealers are already commercially available and work on quantum computers is accelerating. I will talk about the future of computing in a world with quantum computers. First demonstration devices exist and a universal quantum computer with computational powers beyond that of any imaginable classical computers seems just over the horizon and feasible within the next few years. I will explain the origin of the exceptional computational power of quantum computers, and how this can be applied to solve problems that are impossible to solve on classical computers. I will discuss the need for educating a new generation of quantum software engineers, the design of hybrid classical-quantum HPC systems, and the challenges involved in bringing quantum algorithms to bear on important real-world problems.

Quantum Computing: Cladogenesis Beyond Exascale HPC

Andrew Landahl
Sandia National Laboratories

Quantum computing promises to be the most radical innovation to HPC since Babbage's Analytical Engine. By literally harnessing new laws of computation, quantum computers transcend many challenges faced by HPC today, including data movement, memory, parallelism, and power. In this talk, I will survey the state of the art in quantum algorithms, quantum computing hardware, and quantum computing architectures. I will also highlight some of the main software challenges moving forward, including developing effective hardware benchmarks, programming paradigms, and fault-tolerant computing schemes. I will focus on universal, or general-purpose, quantum computing, not special-purpose quantum annealer.

Session 4: Quantum Computing

Quantum Annealing at NASA: Current Status

Rupak Biswas
NASA Ames Research Center

Success of NASA missions depends on solving numerous complex computing challenges, some of which are NP-hard and intractable on traditional supercomputers. Quantum computers have unique capabilities for solving such difficult problems by harnessing quantum mechanical effects such as tunneling, superposition, and entanglement. The Quantum Artificial Intelligence Laboratory (QuAIL) at NASA Ames Research Center is the space agency's primary facility for conducting research in quantum information sciences (QIS). If explored to its full potential, QIS can greatly accelerate a wide range of tasks leading to new technologies and discoveries that will significantly change the way we solve real-world problems. NASA routinely deals with hard optimization problems related to its missions in aeronautics, Earth and space sciences, and space exploration. The QuAIL team hopes to demonstrate that a quantum system will dramatically improve NASA's ability to solve these problems. Our initial target has been quantum annealing – a strategy that is widely applicable to solving optimization problems. In this talk, I will give a brief overview of the D-Wave 2X quantum annealer and present recent results from a set of NASA application areas.

What to Do with More Than 1000 Quantum Bits

Scott Pakin
Los Alamos National Laboratory

In August 2016, Los Alamos National Laboratory (LANL) took ownership of a D-Wave 2X quantum annealer. Relative to "true" quantum computers, quantum annealers are computationally feeble. However, they are vastly more scalable: While the state of the art in quantum computers is approaching only low double-digit numbers of qubits, LANL's D-Wave 2X is already operating with 1095 qubits.

In this talk, we cover some of LANL's initial investigations and experiments with its D-Wave 2X system. LANL researchers have explored the D-Wave 2X's suitability for signal processing, biological structures, graph algorithms, computer vision, and other computational domains. We discuss these efforts and their findings in an attempt to characterize what classes of problems are a good match for a quantum annealer.

Session 5: Crosscutting Fragments

128 Bit, Exascale Memory Reference Models for Next Generation, ExaByte Capacity Physical Memory Systems

Steven Wallach
Micron Technology, Inc.

As we plan for Exascale computing in the next decade, we must also plan for the mechanisms needed to reference Exabytes of physical Memory. These mechanisms may require having an address space greater than 64 bits. We will examine architectural defined memory reference extensions that include: 128 bit virtual address, global protection control, incorporation of non-volatile memory in the processor address hierarchy and PGAS memory reference mechanism.

Cross-Stack Approximate Computing for Modern Applications and Future Substrates

Luis Ceze
University of Washington

A significant proportion of computer system resources are devoted to applications that can inherently tolerate inaccuracies in their data, execution and communication. Hence, "approximate computing" is promising for performance and energy efficiency. However, taking advantage of approximate computing needs: language support to specify where and how to apply approximation; analysis and mechanisms that ensure good output quality; and hardware/system support that take advantage of approximation.

In this talk I will describe our effort on co-designing language, hardware and system support to take advantage of approximate computing across the system stack (compute, storage and communication) in a safe and efficient way. I will end with some thoughts on how effective approximate computing techniques cannot only improve computer systems in the short and medium term but can also enable the use of new substrates.

Session 5: Crosscutting Fragments

Probabilistic Computing in the Post-Moore's Era

Laura Monroe
Los Alamos National Laboratory

Probabilistic computing is a non-deterministic approach to computation, and can be done in hardware, software, or both. The non-determinism may come from hardware or software; for example, device reliability is expected to decrease due to decreased feature size, and voltage threshold techniques are on roadmaps now, which could also increase faults.

Probabilistic and approximate computing show promise for power savings, and are likely to apply to many post-Moore's Law architectures. However, this is a different way to compute and requires rethinking basic computational paradigms. If you aren't always going to get the right answer, when do you care and why? How do you know that your answer is "good enough"? In this talk, we discuss probabilistic and approximate computation, their advantages, and challenges of use that may arise.

Towards Always-On Integrated Performance Analysis Tools

Matthew LeGendre
Lawrence Livermore National Laboratory

HPC performance analysis tools have an unfortunate but well-deserved reputation for being unreliable and hard to use. Much of this comes from the complex software infrastructure behind these tools, and the difficulty in supporting it across the diverse HPC ecosystem. In this talk I will argue that performance tools should shift away from being monolithic stand-alone products, and instead become lightweight utilities that tightly integrate into and share software infrastructure with applications. I will discuss some our efforts towards producing these integrated lightweight tools and infrastructure, the new capabilities (such as always-on performance analysis) that this approach enables, and some of drawbacks.