Workshop Report:
2025 International Symposium of Quantitative Codesign of Supercomputers

Version: 1.0

Acknowledgements

We would like to acknowledge the efforts of the following people who not only made our symposium possible, but worked to make our symposium outstanding: SC'25 workshop committee chair Miquel Pericàs, vice chair Danielle A. Ellsworth; the anonymous workshop proposal reviewers who provided helpful guidance; the audio-visual team that provided support for our symposium, and DOE/ASCR program manager David Rabson. Our symposium's chair received support by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under the Next Generation Scientific Software Technologies portfolio and received support by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy.

Thank You!

– Terry Jones, chair Quantitative Codesign of Supercomputers

citation: Terry Jones, James Ang, James Brandt, Michael R. Jantz, Estela Suarez. *Workshop Report: 2025 International Symposium of Quantitative Codesign of Supercomputers*. Oak Ridge National Laboratory. April 2026. ORNL/LTR-2026/4434.

TABLE OF CONTENTS

EXECUTIVE SUMMARY 3
1. BACKGROUND ON QUANTITATIVE CODESIGN 4
2. PURPOSE OF THE WORKSHOP 5
3. WORKSHOP STRUCTURE 6
3.1 Theme for 2025 7
3.2 Agenda Structure 7
3.3 A Hybrid Format: Accommodating In-Person and Remote Participation 7
4. WORKSHOP RECAP 7
4.1 Symposium Agenda 7
4.2 Terry Jones Presentation 9
4.3 Jack Dongarra Presentation 9
4.4 Sadaf Alam Presentation 10
4.5 Georg Hager Presentation 10
4.6 Rio Yokota Presentation 11
4.7 Mark Wilkinson Presentation 12
4.8 Moderated Panel Discussion 12
5. POST-WORKSHOP FINDINGS, RECOMMENDATIONS AND "NEXT STEP" STRATEGIES 24
5.1 Workshop Findings 24
5.2 Evolving SQCS 24
5.3 Next Steps 25
APPENDIX 1 – RELATED ACTIVITIES 26
APPENDIX 2 – SYMPOSIUM SPEAKER BIOGRAPHIES 27
APPENDIX 3 – ORGANIZING COMMITTEE 29
APPENDIX 4 – ATTENDEES & WORKSHOP PHOTOGRAPHS 31

Executive Summary

The Quantitative Codesign of Supercomputers symposium is an annual workshop series that aims to significantly improve the effectiveness of high-performance computing through bringing about increased understanding of current limitations and improved development processes. This symposium considers combining two methodologies—collaborative codesign and data-driven analysis—to realize the full potential of supercomputing. Quantitative codesign offers a collaborative evidence-based approach to address our existing needs and our upcoming ambitions. The theme for SQCS'25 was opportunities and challenges in Holistic Performance Engineering. We considered the question *how can quantitative codesign be applied to address integrated and comprehensive concerns surrounding supercomputer performance engineering*. This symposium was created to bring together leaders in the field to review current efforts across centers and discuss areas that show potential.

For 2025, our **speakers** included:

- Jack Dongarra – 2021 Turing Award winner and professor at both University of Tennessee and University of Bristol,
- Sadaf Alam – Director of Advanced Computing Strategy and CTO at the Bristol Centre for Supercomputing (BriCS),
- Georg Hager – head of the Research Division at Erlangen National High Performance Computing Center,
- Rio Yokota – Professor at the Supercomputing Research Center, Institute of Integrated Research, Institute of Science Tokyo,
- Mark Wilkinson – National Director of the Science and Technology Facilities Council, Distributed Research using Advanced Computing (DiRAC) HPC Facility.

The workshop finds that the present state of quantitative co-design is still nascent with plenty of divergent paths to choose from. There are a number of recommended "next steps" that should be followed to increase the usability of quantitative codesign of supercomputers. As this report details, our 2025 **findings** are:

- An emphasis on the need for more investment in research at universities and government laboratories to experiment with new computer architecture designs. 93% of the machines comprising the Top 500 list are based on commodity parts. Sadly, co-design is becoming mostly aspirational.
- The gap between processor speeds and data movement is becoming more problematic. Again, this emphasizes the need for research and experimentation to improve this situation.
- With the advent of superscalars, some see an increasing risk that scientific computing is becoming too inconsequential to vendors.
- Effective automated ways to capture operational and performance related data is an active field and AI/ML are making impressive strides. Such information can be used for many purposes included future machine planning and application performance tuning.
- The rise in importance of lower precision floating point operations due to AI gives raises the need for new languages that can more fully express precision constraints and flexibility -- perhaps using a new typing model.
- Domain specific co-designs are worth exploring.

2025 International Symposium on the Quantitative Design of Supercomputers
Held in conjunction with Supercomputing '25
St. Louis, Missouri – November 17, 2025

1. Background on Quantitative Codesign

The Quantitative Codesign of Supercomputers symposium is an annual workshop series that aims to significantly improve the effectiveness of high-performance computing through bringing about increased understanding of current limitations and improved development processes. This symposium considers combining two methodologies—collaborative codesign and data-driven analysis—to realize the full potential of supercomputing. For full potential of supercomputing, we consider everything pertaining to output production including, but not limited to, the performance of applications, system software, workflows, health of hardware. Our centers store vast sums of information, yet using this data is a demanding task. To a large extent, the difficulty in obtaining quantitative insight has to do with discovering, accessing, and analyzing the right data. Codesign also presents formidable challenges, e.g. on how to use the data collected on current systems to facilitate the (potentially very different) design of next-generation supercomputers and successfully support our upcoming environments. Quantitative codesign offers a collaborative evidence-based approach to address our existing needs and our upcoming ambitions. This symposium was created to bring together leaders in the field to review current efforts across centers and discuss areas that show potential.

Over the past decade, there has been a growing awareness of the multi-faceted benefits we can derive from data-driven strategies like Quantitative Codesign. This increasing awareness, along with improvements in Machine Learning (ML) technologies, have driven vendors, operations staff, and application developers to espouse integrating an ever-increasing level of instrumentation into their products. The time is ripe for turning this vast trove of available information and the incredible advances in analysis technologies it represents into appropriate knowledge and understanding. Doing so would create a feedback loop that could assist vendors and software developers in their desigtrjns. The recent Genesis Mission[1] underscores the need for timely access for developers of technologies, architectures, and systems to carry out the research needed to create the future computing software ecosystem, and Quantitative Codesign provides a solution to the 'access problem' of these extremely rare machines. If the future envisioned by the CSESSP report is to be realized, our software base will require significant investment in both modified and new code — an activity enormously assisted by Quantitative Codesign. There is no disagreement that more knowledge is good though there is still lack of concurrence across HPC stakeholders as to the cost/benefit tradeoff for varying fidelities of information collection and long term storage. The benefits of Quantitative Codesign will come through integrating design processes with more detailed knowledge of the interactions of the various components within the HPC ecosystem.

Quantitative Codesign is also essential for addressing challenges brought about by the recent trend of increasing heterogeneity and varied accelerators in HPC architectures. For example, many HPC machines now incorporate alternative types of memory alongside conventional DDR SDRAM. Technologies such as

[1] https://www.energy.gov/undersecretaryforscience/genesis-mission/genesis-mission

"on-package" or "die-stacked" DRAM as well as non-volatile RAMs can provide distinct advantages compared to conventional DRAM, including higher performance as well as cheaper and more energy efficient storage per byte. Each of these technologies also comes with its own limitations, such as smaller capacity or less bandwidth for reads and writes. Further complications arise because some of these new technologies can interface directly with processor caches, while others can only be accessed through peripheral devices, such as GPUs or other accelerators.

Quantitative Codesign could mitigate many of the current problems with allocating and managing such heterogeneous resources effectively. Detailed knowledge of application demands will enable architects to make better decisions about how to select and organize computing and memory hardware. This approach can also help system software, including operating systems, compilers, and runtime software, distribute the available hardware resources among applications more effectively. Codesigned system software could utilize knowledge from new data sources for better energy efficiency and workflow management. Integrating high-level profiling and analysis with low-level resource management routines will enable these systems to implement new policies that respond flexibly to changes in application demands and could potentially expose important new efficiencies on platforms with heterogeneous hardware.

2. Purpose of the Workshop

The purpose of the workshop was to build the necessary community support to build up and foster concrete implementations of quantitative codesign. As architectural options expand in type and complexity, the need for a quantitative basis to drive architectural directions becomes increasingly urgent. We do not have the primary mission to raise awareness of an individual's research; rather we wish to bring more wide-ranging interactions highlighting vision and positions and stimulating discussions.

Any shortfall in our detailed understanding of operations and performance impacts the whole spectrum of stakeholders. Whether providing hardware architectures, system software, application programming environments, or production run-time environments, having the appropriate knowledge to optimize the interaction and configuration of all of these critical components as well as the evolution of the HPC ecosystem is critical to continued growth. The rapidly changing HPC landscape demands a codesign that effectively uses the data collected on previous and current systems to facilitate the design of next-generation supercomputers and successfully support our upcoming environments. Specifically, we would like to bring increased clarity for our challenges and opportunities.

- Challenges: We have important issues to resolve, but we are not starting from scratch. HPC computing centers already collect a wealth of information on the health, usage, and efficiency of our machines, workflows and programming environments. While collection and analysis of this information has evolved and improved over the years, there are still severe gaps that have left us unable to provide the knowledge that is needed by hardware and software vendors, system operations staff, application developers, and user groups to create and operate highly efficient and secure large scale HPC systems. Would-be users of this information face difficulties in obtain insight from the collected data a timely manner, and efforts to provide both data and analysis means are currently fragmented across centers both at national and international levels. The infrastructure to collect, store, share and analyze the volumes of available information is a core capability—yet, many barriers remain due in large part to the many stakeholders and insufficient coordination, but also due to data privacy and security issues. With many new potential information sources in future systems, we must quickly identify and address critical requirements and gaps across the various stakeholders. Doing so will enable us to create collective and collaborative solutions that address both existing challenges and emerging needs and effectively support our upcoming HPC environments. The nature of this challenge suggests that it is an excellent opportunity for a codesign approach. Codesign is defined as the process of jointly designing interoperating components of a computer system—in particular: applications, algorithms, programming models, system software, as well as the hardware on which they run, and the facilities hosting them. Designing solutions based on intelligence derived from the data collection and analysis processes described above are henceforth referred to as Quantitative Codesign of Supercomputers.

Making progress at the highest end of HPC without access to the needed data can be compared to being asked to fly an airplane at night without sufficient instrumentation. Vendors are provided with example applications to target, but often lack a true understanding of where inefficiencies manifest on full scale workloads. Furthermore, computer architecture simulators face an inevitable challenge in trying to incorporate all the critical performance-killing attributes of current generation technologies and their integration: a simulation that includes all details of the architecture, from the chip micro-architecture up to infrastructure, would take forever to run. For this reason, simulations must make tradeoffs between the accuracy of their representation and the required modelling time. Hence the vendors miss opportunities for improvement. Moreover, users often only have feedback on operating efficiency at the granularity of total application execution time. Low-level interactions frequently cause substantial performance degradations that users are unable to explain. Likewise, operations staff often lack knowledge of application resource utilization and cannot diagnose the longer run times experienced by the users. In addition, operations staff cannot ensure secure operations without an understanding of normal (expected) behavior and anomalies that deviate from that. Since root causes go undiagnosed on current systems, next generation systems will also fail to address the very same problems.

- Opportunities: First and foremost, we wish to discuss the merits of a coordinated effort to bring together the helpful data from each stakeholder in the codesign space into a framework where data discovery and access is straightforward regardless of data source while respecting data privacy and security concerns. The envisioned Quantitative Codesign environment would pull together data traditionally held by disjointed communities (e.g., sysadmins, application teams, vendors, and so on) into a framework where the needed data is easily accessible. This framework would provide flexible but secure mechanisms for data providers who wish to share their data with others including application teams, vendors, facilities, operations, and system software researchers. In many cases, we seek to bring together data that is currently being produced although not generally known or utilized for a variety of reasons; in a few instances, we seek to extend and provide new data collection capabilities.

  For example, one area that is ripe for integration with Quantitative Codesign processes is the intersection of application development and run-time environments. In the past few years Continuous Integration (CI) has been widely adopted by development teams to continuously test development efforts. As part of these CI efforts, developers test across a variety of platforms on a daily basis and typically provide a pass/fail result for each. Introducing targeted run-time data collection (e.g., memory, application & hardware counters, MPI, OpenMP, GPGPU, I/O, energy consumption) and quantitative analysis into this process would enable feedback to users and identify issues within applications, compiler capabilities, runtimes, and differences across platform architectures that ultimately would drive improvements across the spectrum of stakeholders.

  Integrating Quantitative Codesign capabilities with existing design processes will enable more effective solutions across the computing stack. Information derived from monitoring and analysis would provide valuable insight for users, application developers, system architects, and facility designers as to how, and why, applications make use of the underlying system resources. Furthermore, by identifying the appropriate stakeholders and introducing them to information originating from diverse collection regimes, this symposium seeks to facilitate the discovery and sharing of potentially useful intelligence among larger teams and communities. In doing so, this approach also has the potential to spark further discussions and research on how to collect, employ and share this information more effectively. Thus, there is significant opportunity for discoveries that will not only increase application performance, but also benefit the broader HPC and scientific communities.

3. Workshop Structure

The Quantitative Codesign of Supercomputers symposium took place during the opening day of the 2024 Supercomputing conference. The workshop was was structured for hybrid-attendance with both in-person and virtual attendees and speakers. Further, the workshop was framed in the Symposium format to achieve the kind of deep interactions that lead to change within HPC. Our preference for audience interaction was in response to the state of the field (which we see as in its infancy).

3.1 Theme for 2025

The theme for SQCS'25 was opportunities and challenges in Holistic Performance Engineering. We considered the question *how can quantitative codesign be applied to address integrated and comprehensive concerns surrounding supercomputer performance engineering*. For our purposes, we defined performance engineering to be a process that ensures software applications meet performance and scalability goals. It's a continuous process that involves quantitative co-design of hardware and system software before a new machine is deployed, as well as the testing and monitoring of applications throughout the life of the machine. While we cannot take credit for the accomplishment, the February 2025 Deepseek announcement provided a concrete example of how impactful performance engineering can be. SQCS'25 considered supercomputing design and operation aspects that traditionally have suffered from insufficient interaction. How can performance engineering be moved from a collection of separable aspects into a complete and holistic approach? We considered the question from several different perspectives and apply quantitative codesign benefits toward broad-scope type objectives.

3.2 Agenda Structure

Given our desire to bring more wide-ranging interactions highlighting vision and positions and stimulating discussions, we developed a schedule designed to facilitate these interactions (see Table 1 below). In particular:

- The keynote speaker was chosen based on his long history in HPC with work that spans all areas of codesign including novel architectures, system and application software, tool development, performance diagnostics and more, in both lab and academic environments.
- Three distinguished speakers were chosen who, as an aggregate, provided codesign perspectives on the use of carefully crafted kernels to inform telemetry and monitoring, how quantitative co-design was recently used to design memory architectures for CEA, and ORNL's experience with quantitative co-design.
- A moderated discussion of audience, speakers, and panelists was included to enable both technical discussions and community-building.

3.3 A Hybrid Format: Accommodating In-Person and Remote Participation

As with our previous Symposium, the lingering effects of COVID and an increased confidence in the effectiveness of virtual participation had an impact on the format and character of the workshop. This was the fourth time for the SC series of conferences to ever have a hybrid format: SC25 supported both in person attendees at America's Center Conference Center in St. Louis, Missour and remote attendees though the revamped SC25 online platform, Zoom and Sli.do. The role of the session chair and organizer remained largely the same as in previous years with some adjustments and increased responsibilities to account for remote participation by speakers and attendees. The Quantitative Codesign of Supercomputers symposium was presented via ***Live stream sessions***. Under this format, content was recorded by AV technicians at the convention center and sent to remote participants in real time via Vimeo. Remote presenters connected via zoom. For all remote symposium presenters, we arranged for an internet assessment on the day of the symposium prior to the symposium start. This was used to ensure no fallback measures were needed. All remote speakers were able to participate as planned.

4. Workshop Recap

4.1 Symposium Agenda

Featuring 2022 Turing Award winner Jack Dongarra, our speakers included five well-known HPC leaders representing Europe, Asia and the Americas, each addressing opportunities and challenges in holistic performance engineering. Our moderated discussion with the speakers received a 50-minute block. The final schedule is presented in Table 1 below.

Table 1 – Symposium Agenda

All Times US CT	Speaker/Panelist	Abstract
9:00 to 9:10	 Terry Jones	**Opening Remarks from Workshop Chair** Welcome and workshop logistics
9:10 to 10:00	 Jack Dongarra	**Codesign of Supercomputers, Are You Kidding?** Future supercomputers must be designed not only for raw speed but also for efficiency, scalability, and fitness to scientific and AI-driven workloads. Quantitative codesign provides a systematic way to achieve this by linking workload characterization, performance modeling, and hardware–software trade-offs.This talk will outline the principles of quantitative codesign and show how metrics such as time-to-solution, energy, and numerical accuracy guide choices in processors, memory, interconnects, and software. Unfortunately, we have not had a real co-design baked into our supercomputers.
10:00 to 10:30	[break]	[break location]
10:30 to 10:45	 Sadaf Alam	**An AI-driven approach for delivering sustainable supercomputing services** Isambard-AI is a UK National AI Research Resource (AIRR) that has been formally launched as a service in July 2025. In preparation for the launch, multiple co-design efforts were undertaken to address accessibility challenges faced by diverse AI research teams and to align with performance and environmental sustainability objectives. This talk highlights the application of AI frameworks, including the Model Context Protocol (MCP), together with operational data collection methods to evaluate the effectiveness of our data-driven co-design strategies.
10:45 to 11:00	 Georg Hager	**Co-design in the HPC space with analytic resource models** In order to design hardware for a specific (class of) application or adapt applications for the hardware at hand, the starting point should always be a thorough understanding of the relevant bottlenecks of these applications on existing hardware platforms, or, more generally, their resource requirements. White- and gray-box Analytic modeling is an invaluable tool achieve this goal. Contrary to popular belief, analytic models can be partially automated and used in production and architectural exploration. In this talk, we first define white- and gray-box analytic resource modeling and show whether and how the modeling process can be automated. We then demonstrate how such models can employed in co-design efforts.
11:00 to 11:15	 Rio Yokota	**Co-Design of HPC Applications for Tensor Cores** The AI market will continue to drive the processor architecture, where we will see more lower-precision matrix multiplication engines such as Tensor Cores. HPC applications that can utilize this type of hardware effectively will see huge gains in performance, whereas those that cannot will see minimal increase in performance. Most HPC applications can be formulated in multipole ways, and the underlying algorithms can be changed. This talk will focus on one such example in GROMACS, where the conventional particle-mesh Ewald method that relies on FFT, is substituted with a fast multipole method that relies on dense matrix multiplication, thus facilitating the use of Tensor Cores.
11:15 to 11:30	 Mark Wilkinson	**Collaborative co-design at scale: the DiRAC HPC Facility experience** This presentation will focus on the collaborative, quantitative co-design approach to the deployment of large-scale computing services adopted by the STFC DiRAC HPC Facility in the UK (www.dirac.ac.uk). Over the past 15 years, successive generations of DiRAC services have demonstrated how workflow-centred co-design can maximise the scientific impact of computing investments. The co-design of DiRAC services has ranged from silicon-level to system-level, alongside extensive software development effort, and has delivered significantly increased system capabilities. Looking to the future, I will explore how co-design can be used to develop cost-effective and energy-efficient heterogeneous computing ecosystems for AI and simulation.
11:35 to 12:25	Moderated Discussion	Your opportunity for audience & panelists (our 4 invited speakers) to dig deeper
12:25 to 12:30	Terry Jones	Closing Remarks.

4.2 Terry Jones Presentation

- **Introduction to Symposium:** Terry Jones from Oak Ridge National Laboratory welcomed participants to the Fifth International Symposium on the Quantitative Co Design of Supercomputers and provided a brief introduction to the day's proceedings.
- **Co-Design Concept:** Jones explained the concept of Co-design, emphasizing the importance of both breadth and depth in making supercomputers efficient. He highlighted the involvement of various communities and the integration of different aspects such as materials, devices, architectures, and software.
- **Quantitative Approach:** Jones discussed the importance of a quantitative approach in supercomputing, advocating for data-driven decisions rather than relying on people's hunches or past experiences.
- **Agenda Overview:** Jones previewed the agenda, mentioning the featured speaker Jack Dongara and other speakers including Sadaf Alam, George Hager, Rio Yakota, and Mark Wilkinson. He also mentioned an hour for audience participation and a moderated discussion.
- **Introduction of Jack Dongara:** Jones introduced the featured speaker, Jack Dongara, highlighting his expertise in numerical algorithms, parallel computing, and his numerous accolades, including the 2021 Turing Award.

4.3 Jack Dongarra Presentation

- **Skeptical View on Co-Design:** Jack Dongarra, who recently retired from Oak Ridge National Lab, expressed skepticism about the current approach to purchasing high-performance computing systems. They highlighted the reliance on off-the-shelf components and the need for a more integrated approach to hardware, software, and applications.
- **Definition of Co-Design:** Dongarra defined co-design as the simultaneous development of hardware, software, and applications, involving architects, domain scientists, mathematicians, and computer scientists. The goal is to create optimized systems based on mutual feedback from various stakeholders
- **Historical Perspective on University Research:** Dongarra reminisced about the mid-20th century when universities were at the forefront of designing and building hardware architectures. They emphasized the need for more investment in research at universities and government laboratories to experiment with new designs.
- **Top 500 List Preview:** Dongarra provided a preview of the upcoming Top 500 list, highlighting that 93% of high-performance machines are based on commodity parts, with a significant reliance on Intel and AMD processors, NVIDIA accelerators, and Ethernet or InfiniBand interconnects.
- **Cloud Providers' Custom Hardware:** Dongarra discussed how cloud providers like Alibaba, Amazon, Google, and Microsoft are developing their own hardware for cloud services, which is not available for purchase. This approach contrasts with the reliance on commodity parts in high-performance computing.
- **Market Capitalization Comparison:** Dongarra compared the market capitalization of traditional computing companies with that of hyperscalers like Facebook, Amazon, Google, Microsoft, and Apple. They noted the significant wealth of cloud-based services and their investments in custom hardware.
- **NVIDIA's Market Influence:** Dongarra highlighted NVIDIA's market worth and its focus on AI-driven hardware. They noted that NVIDIA's hardware is being driven by the AI community, impacting the hardware used in high-performance machines.
- **Historical Benchmarking:** Dongarra shared their experience as an accidental benchmarker during their graduate studies, working on the LINPAC project. They discussed the origins of the Top 500 list and its evolution over the years.

- **Current Top 500 Machines:** Dongarra provided details about the current top machines on the Top 500 list, including the number one machine at Lawrence Livermore National Lab, which has 11 million cores and achieved 1.8 exaflops.
- **Challenges in Co-Design:** Dongarra discussed the challenges in co-design, particularly the gap between processor speeds and data movement. They emphasized the need for research and experimentation to improve this situation.
- **NVIDIA's Hardware Design:** Dongarra explained NVIDIA's approach to hardware design, focusing on tensor cores and mixed precision arithmetic. They highlighted the differences between the Hopper and Blackwell processors and their impact on performance.
- **Mixed Precision Algorithms:** Dongarra discussed the importance of mixed precision algorithms in scientific computing. They shared an example of an algorithm that leverages 32-bit arithmetic to achieve the same accuracy as 64-bit arithmetic, resulting in significant performance improvements.
- **New Machine at Argonne:** Dongarra mentioned a new machine being purchased for Argonne National Laboratory, which will be run by Oracle and use NVIDIA parts. This machine represents a shift in the traditional model of purchasing scientific computers.
- **Future of Co-Design:** Dongarra concluded by stating that co-design is mostly aspirational today, with AI workloads driving hardware design. They emphasized the need for engagement in co-design to ensure scientific computing does not become a second-class citizen.

4.4 Sadaf Alam Presentation

- **Introduction of Dr. Sadaf Alam:** Dr. Sadaf Alam was introduced as the director of Advanced Computing Strategy and CTO at the Bristol Center for Supercomputing. They have a significant background in supercomputing and AI, having previously worked at the Swiss National Supercomputing Center and Oak Ridge National Laboratory.
- **Bristol Center for Supercomputing Overview:** Alam provided an overview of the Bristol Center for Supercomputing, highlighting its recent establishment and various national and local services, including the Isambard AI national AI research resource and the Airport AI research resource portal.
- **Isambard AI Project:** Alam discussed the Isambard AI project, emphasizing its rapid development and successful launch in July 2025. The project involved collaboration from various stakeholders, including contractors and suppliers.
- **Sustainable Service Management:** Alam explained the concept of sustainable service management, focusing on empowering stakeholders and project owners. They highlighted the importance of quick access to AI resources and the collaborative nature of the project.
- **AI Research Outcomes:** Alam shared some outcomes of the AI research facilitated by Isambard AI, including projects by the UK AI Security Institute and a version of Alpha Fold for studying heart conditions in young people.
- **Future Plans for Isambard AI:** Alam outlined future plans for Isambard AI, including the development of a sovereign AI prototype with more open-source and open-standard solutions. They emphasized the importance of international collaboration in this effort.

4.5 Georg Hager Presentation

- **Introduction to Erlangen:** Georg Hager introduced Erlangen's history in parallel computing, mentioning significant projects like the Erlangen General Purpose Array, Dermu, and the Soprano project aimed at developing the first Gigaflop computer in Germany.
- **Co-Design in HPC:** Hager discussed the importance of co-design in HPC, emphasizing the need for hardware and software to work together efficiently. They highlighted the role of performance modeling in understanding hardware-software interactions.

- **Performance Modeling Types:** Hager explained different types of performance models, including white box, grey box, and black box models. They provided examples and discussed the strengths and weaknesses of each type.
- **White Box Modeling Example:** Hager shared an example of white box modeling involving the transfer of data from the Lightness Computing Centre in Munich to Garching, demonstrating the importance of latency and bandwidth in decision-making.
- **Grey Box and Black Box Modeling:** Hager elaborated on grey box and black box modeling, discussing the use of measured values in grey box models and the challenges of fitting functions to measured data in black box models.
- **Automating Performance Models:** Hager described efforts to automate performance modeling, including tools like Contrast and Osaka. They emphasized the importance of understanding data transfers and execution overlap in the cache hierarchy.
- **Challenges in Parallel Code Modeling:** Hager highlighted the complexities of modeling highly parallel codes, discussing the need for accurate execution and communication models and the phenomenon of idle waves in MPI processes.
- **Future Needs for Performance Modeling:** Hager outlined several areas for improvement in performance modeling, including better compiler-assisted code execution modeling, improved hardware metrics, more generic loop nest modeling, automatic application skeletonization, and micro benchmarking for performance and energy.

4.6 Rio Yokota Presentation

- **Introduction to Co-Designing HPC Applications:** Rio Yokota introduced the topic of co-designing HPC applications, mentioning efforts to emulate higher precision matrix multiplication using lower precision tensor cores. They highlighted two separate efforts: recovering single precision FP-32 from FP16 tensor cores and recovering double precision gem from int 8 tensor cores using the Ozaki scheme.
- **Compensated Summation Method:** Yokota explained the compensated summation method used to recover single precision FP-32 from FP16 tensor cores. They described how the matrix A is split into A and delta A, and the multiplication is done using four terms instead of one. The method involves doing the addition outside the tensor core to avoid error accumulation due to rounding modes.
- **Performance of Compensated Summation Method:** Yokota discussed the performance of the compensated summation method, highlighting that it achieves high flops even with the addition done outside the tensor core. They compared the performance of their method with Kublas SGM and noted that their method gets around 50 teraflops on a 100 and more than 30 on TF 32.
- **Ozaki Scheme:** Yokota introduced the Ozaki scheme, which uses an arbitrary number of matrices to express a high precision matrix. They explained the difference between the compensated summation approach and the Ozaki scheme, noting that the latter includes the exponent part and allows for controlling the number of slices to achieve desired accuracy.
- **Application of Ozaki Scheme:** Yokota described the application of the Ozaki scheme, using an artificial matrix with a uniform random number and a controllable exponent range to observe the relative error. They emphasized that the number of slices used affects the accuracy, with more slices resulting in lower relative error.
- **Algorithmic Changes for HPC Applications:** Yokota discussed the importance of going back to the algorithmic drawing board to change the algorithm itself for HPC applications. They provided an example of working with the Gromax team to replace the particle mesh evald with the fast multiple method, which can be modified to become a gem and achieve peak performance on tensor cores.
- **Flexibility in Algorithms and Applications:** Yokota emphasized the flexibility in algorithms and applications, noting that there is a choice in how to discretize PDEs and what basis to use. They

highlighted the importance of co-designing algorithms with hardware to exploit fine grain homogeneity and achieve better performance on modern hardware. 10:23

- **Conclusion:** Yokota concluded the presentation by reiterating the importance of co-designing algorithms with hardware to exploit tensor cores effectively. They emphasized that fine grain homogeneity is key to achieving better performance and that adaptivity can be done at the coarse grain level.

4.7 Mark Wilkinson Presentation

- **Introduction to Co-Design:** Mark Wilkinson introduced the concept of Co-design in the context of the Dirac HPC facility, emphasizing the importance of treating high-performance computers as scientific instruments designed by the users for their specific scientific needs.
- **Dirac HPC Facility Overview:** Wilkinson provided an overview of the Dirac HPC facility, which serves the theory community in SDFC, including particle physics, nuclear physics, astrophysics, cosmology, and planetary science. The facility offers compute and training resources, handling both capability runs and data-intensive calculations.
- **Challenges in Data Management:** Wilkinson highlighted the challenges faced in managing the large volumes of data generated by simulations, which can exceed 10 petabytes. He also mentioned the increasing use of AI and machine learning to enhance simulations. 2:14
- **Co-Design Process:** Wilkinson explained the Co-design process used at Dirac, which starts from the science case and involves multiple stakeholders, including the Dirac board, technical directorate, research software engineers, and industry partners. The process ensures that the community is involved at every level.
- **Dirac Systems and Their Specializations:** Wilkinson described the four systems at Dirac, each designed for specific scientific workflows: memory-intensive system at Durham for cosmology simulations, extreme scaling system at Edinburgh for particle physics, and data-intensive systems at Leicester and Cambridge for heterogeneous workflows.
- **Benefits of Co-Design:** Wilkinson emphasized the benefits of Co-design, citing examples such as the Tersa cluster in Edinburgh, which achieved significant performance improvements through collaborative efforts between software engineers, tech support teams, and industry partners.
- **Energy Efficiency Improvements:** Wilkinson discussed efforts to improve energy efficiency, including clocking down GPU systems to save energy without impacting performance and implementing energy-aware scheduling using AI-based models.
- **Conclusion on Co-Design:** Wilkinson concluded that system Co-design is worth the effort, as it involves optimizing not just the compute and applications but also the network and middleware. He stressed the importance of involving people at every step and highlighted the return on investment in terms of productivity and efficiency.

4.8 Moderated Panel Discussion

The following is a transcript automatically generated from the SC25 video recording of the symposium via Microsoft Copilot. Minor edits have been applied to remove the timestamps and add speaker names when possible.

Jim Ang

So thank you, Terry. For those of you who don't know me, I'm Jim ANG. I'm at Pacific Northwest National Lab. I did want to acknowledge the organizing committee for this workshop. So I wanted to

acknowledge and Gentilly and Jim Brand from Sandia National Labs, Michael Jantz from University of Tennessee at Knoxville and Estella Suarez from Eulish and University of Bonn. So the we've, we form the organizing committee for this workshop. And so I have the the opportunity now to moderate this panel and we'll kick things off with the questions for the panelists. If you cannot see them, I will read them for you:

> **Question 1**: As supercomputing systems are becoming more diverse with new compute and monitoring capabilities in the processing units, heterogeneous memory and storage tiers, and faster interconnect options merging, what do you see as opportunities and or challenges for collecting and leveraging quantitative data? Are there any game changing techniques or possibilities on the horizon?
>
> **Question 2**: Can we count on AI to help fill in the gaps of incomplete telemetry data and help us to achieve a better understanding of what is happening on our systems?

And any of you are free to pick up a mic and share your thoughts, your answers that one on do we need to turn on switch it on? OK, thank you.

Georg Hager

First of all, I object to the statement that we have enough telemetry data. If anything, we don't have enough telemetry data, by which I mean that if you are operating a big computer system and and allowing we like 1 1/2 orders of magnitude below like the top, but it's still a large system, right? What we see is that as soon as you procure a new system, you get new hardware, new networks, even the same hardware from a new vendor. Stuff changes so much that you spend a lot of time dealing with these variations. And as I said in my talk, the amount of data you get from different components of the system, for example, from the CPUs, what's going on within the cache hierarchy, what's going on with power consumption? Can I measure that? Can I do that on ACPU by CPU basis? That's getting worse and worse, basically. And for me, the way to solve this problem would be to have more of a standardization effort for this kind of telemetry data. And that goes from the core level to the whole system level. So on the core level and the cache hierarchy level, I'd like to have a standardized set of metrics that every CPU, every GPU has to supply. No, it's a wish list. OK, And he doesn't listen to me. But if they did, they would supply that on the node level. I would like to have fine granular measurements of power consumption, for example, so I can see what happens to the system, to the power, if some code is running that I want to analyze without actually employing a tool. If the resolution is good enough, I can just run my monitoring that's running anyway and look at the data and see, OK, this guy is not doing something wrong, OK, power consumption is not correct or something like that. And on the on the full system level, I just want to have enough data so that I can still digest it if necessary, I can boil it down with AI or some compression method. But still if the whole interface from the core to the full system level is standardized and has a proper, you know, documented resolution and and way of presenting the data, that would help a lot in centre operations. So that would be my stance on this.

Mark Wilkinson

So I think one of the things I think is most challenging at the moment is that a, the the compute side is we have a lot of information and it's relatively we, so there's systematic ways to deal with that. I think the data part is much more difficult. And so the data movement around within a cluster for a lot of our workloads, it's, it's either the the memory bandwidth or the network bandwidth or both or what are limiting things. And so it can be very challenging to dig out the the telemetry information about that and understand exactly what where the the bottlenecks are. I think another challenge is this came up actually there were several interesting talks about digital twin in the digital twins workshop yesterday morning looking at digital twins of HPC and it was HPC nodes and then clusters

and then data centres and then federations of data centres. And at each level if the the question being asked was how how far do we need to go in terms of effectively Co design using digital twins? Do we need a digital twin that also in or? Yeah, a digital twin that includes the power grid so that we know how the whole system will behave and influence the power the the power grid. So that brings in federation. Certainly I know in the US that's a lot of discussion is going on about federating the large facilities. It's again, it's something we're discussing a lot in the UK. And when you build in Co design that then it, it becomes important if you are able to move your data around the country and use different bits, different physical supercomputers to do different, different parts of your workflow. That then changes how you will approach the Co design of each one. Because in principle, they could each be much more specialized because no one has to run their whole workflow on a single system if they can move it around. So I don't know if that's answered the question, but it's, it's focused on challenges, I think.

Jim Ang

Let me give you both some feedback. Maybe one of the challenges that Jack laid out for us is that when we talked about scientific computing and simulation, we've got to think about a new Co design opportunity. So that that there there are things we should be asking for and thinking of as gaps that we're not seeing in our current commodity hardware. And as we're trying to Co design new hardware, we should be able to then fill in those gaps. We need to have a list of what what capabilities we need to be looking for. But real.

Rio Yokota

Yes, I agree with my fellow panelists that is still insufficient amount of telemetry that we are getting in particular. So I think Co design nowadays does not mean you know Co designing a chip anymore, right. We, we have very little control over what goes on at the chip level, but we do still have a lot of freedom in the, the system design, right? Like at the interconnect level or, you know, even within a node, like the memory devices that we place. And there's some new memory technology that's going to be coming And, and you know, like optic optical interconnect technology that will be changing, you know, not, not at the chip level, right, that we're no longer influencing any of the design inside the chip, but we, we do have a lot of choice on how we design the entire system there. There's no, I don't think AI lock in that is happening at the entire system level, right. There's there's a lot of different designs that are coming out. There's a lot of ASICS for inference that look very different in in, in nature. Well, some of those actually are different at the chip level too. But you know, I, I think the Co design is happening more at a coarser level and and therefore I think the metrics that we really need to focus on are at that level as well. So, you know, I know Garib is more in very fine grain measurements, but when you're working on, for example, you know, like these large scale training or inference, trying to speed that up, a lot of the profiling you're doing is at a a quite coarse level compared to like traditional HPC, you know, kernel optimizations. And you know, like so this kind of course level measurement, not not, you know, too much information is actually, you know, not very useful at this. It's very good to be able to have a bird's eye view of the entire application without going into too much detail and having the right granularity of information that you can actually, you know, use use it to design, make design choices in your algorithm or your workflow. And so I think the granularity is very different nowadays, right? If you're looking too close, you're not probably not seeing the whole picture. You're probably doing something that is, you know, not not globally the optimal thing to work on. I think it's very important to have a, a very course level view of of exactly the, the type of things we need to measure. Yes, so that that's my so,

Jim Ang

So Rio, I was, I was reminded from your talk about your history with MD Grape and and many, many years of supercomputing when you and your colleagues were presenting progress and successive generations. And I get a sense that that type of specialization is going to be in our future if we want to make progress and get above, you know, fractions of a percent of peak for scientific simulations.

I'm not talking about AII agree completely. You know, the AI community is charging off on its own with much larger budgets and much larger resources for training and for inferencing. But as members of a scientific supercomputing community, I think Jack has issued a challenge for us, which is, you know, maybe apply lessons from the AI community where they have a lot of specialization for the all the, you know, a large range of artificial neural networks and and AI focused activities. We need to take a similar magnifying glass to what are the key kernels, such as the ones you a couple that you've identified, but there are many more in the scientific computing community. And that's probably our challenge to, to, to take up the, you know, the, the mantle of and, and try and address those challenges because the commercial community is probably not going to do that. You know, if we're going to sit back and we'll just look for a commercial general purpose processing solution, that means we're accepting fractions of a percent efficiency, right. So please share your reactions to my comment.

Rio Yokota

Yeah, maybe I can go first on this one. So yeah, I think so. You mentioned AI and you know the HPC community. I, I see, you know, one big difference in sort of the AI software stack is there's enormous amount of reuse of everything that's developed, you know, at, at a lower level, basically because everyone's using, you know, the trans, it's, it's one algorithm that they're running. It's a transformer. I, I know there are other variants like state space models and like diffusion, but a majority is still Transformers, right? And everyone agrees that that that that is like 90% of the market now. And and every single thing like flash attention, you know, 123, any of those things that is highly optimized. There's like a gazillion people using that as a library, right. On the HPC side, you have very specialized algorithms per application, right? Some of whom which have only like, you know, maybe less than 100 users, right. So there's the market of the entire HPC community is smaller compared to AI, but within that community we're further divided in terms of development, right? We're not like focusing on one big HPC algorithm that everyone is running. So, so there's, there's this huge, it's a, it's a small pie that's split even further. And so I think that that's one of the peculiar, you know, differences between like HPC and AI. It's we're so divided. It won't be a large market, but please.

Mark Wilkinson

So I think that's a really interesting point. And do you have a question then, do you think we should be, should we be looking at the AI community and getting better at reusing stuff ourselves? Because I agree we have multiple codes that do almost the same thing and there's usually a good reason. But maybe at this particular junction, it's it's not the time to keep diversifying in codes if we should maybe be recycling fit bits of them.

Rio Yokota

Yeah. So that, that's a very difficult question because, you know, on the one hand, it's, it's very powerful to a physicist or, you know, a domain scientist to be able to do whatever they want in their, you know, what, whatever language they choose. And, and to be able to express, you know, with, without having to coordinate with, you know, so many, you know, like my students that work on the AI side, I have both like HPC and AI projects. And the ones that work on AI, they, they build a lot of prototypes, but integrating that into, you know, like a Megatron or core or any of those things, it takes a long time because you, you, you know, you need to issue like a pull request. NVIDIA never answers, right? They, they, they don't have like a, a streamlined way of incorporating requests from users. So it's hard to do the development at the production level because there's so many people, right, need to agree on what, what the design should be, right? Like designing certain features in Pythorch is probably a nightmare to get people to agree on what the common denominator is. But in HPC, you can just do your own thing. If people like it, they'll just pull from those features and they'll just use it. So it's a much more dynamic atmosphere. I think AI has been very, you know, it's like industrialized now. It's like in production. And so there's less experimental things happening at high quality, right? You, you have people experimenting, but it's not at the same quality as what NVIDIA

offers as products. So that there's this huge gap in research versus like industry. Whereas HPC that difference is, is much smaller, right? What we do in, in research is, is very close or even exceeds performance in, in industrial applications in many cases. So that there's also this difference, but I, I don't think it's necessarily a bad thing that HPC applications are, you know, very segmented but but very like unique and original, right.

Jim Ang
So George,

Georg Hager
yeah, I think what we're missing and I think this summarizes what you both said is the economies of scale, right? I can think about lots of interesting details of a new architecture I want to see on the ground, but if it lacks the economies of scale, I will never get it from anyone. So as much as I like the the stuff that, for example, Barcelona is doing with open chip designing the next vector computer, I like those ideas because they resemble what we had in the 80s, for example. But they, I, I doubt that it's going to take off because it will lack the economies of scale unless it's good for AI.

Jim Ang
Yeah, I, I want to come back to that. But before we do that, oh Mark,

Mark Wilkinson
well, just an example of that I think is the great board. So I, I was a user of grade 4-5 and six many years ago and they were they were fantastic. And then once commodity GPU's could do it, I think that was kind of what, what replaced them. So yeah, it's at at the time there weren't, there was no other way to do things as effectively. This was gravity and body simulations. I was using them for and you know, they're incredibly efficient and extremely good value for money at the time because it was nothing else. But then, yeah, once the commodity, commodity hardware takes over, it's that's, that's good if you can use it.

Georg Hager
But commodity hardware, I in my opinion, and I think Cos Simon has also said that like 20 years ago, the Beowulf movement has basically killed computer architecture research and high performance computing.

Jim Ang
It did.

Georg Hager
And at that time we didn't really care because there was such a large scale to commodity computing. And we saw that a large enough cluster with standard X86 CPUs did the job quite well. So who cared about processor and memory? Who cared about vector CPUs? It was good enough. OK, But now we're hitting this wall and it's not sufficient. And now we're starting to care.

Jim Ang
But Terry,

Terry Jones
Yeah, so I I noticed the discrepancy between Rio's observation of where we are and something Jack said was our goal. Looking today at what can be done, Rio is saying something along the lines of, well, we're not really able to do materials research -- whatever silicon we have, we have -- and we're not really doing anything co-designed with circuits, and so forth. Then there's George's comment just now that that maybe Beowulf has killed those days. But going back to Dongarra's charge of *let's*

get back to full co-design, you wonder, well, what would it take to do full Co design? You know, from the, from materials research on up to all the way up to the applications? Is it, is it simply a matter of scientific computing needs more funding at, at at some level or, or are, is, is there is, is it an organizational thing? Is there not really a body that that's charged with, with doing it or what? What are the impediments to seeing co-design, you know, with breadth and and in depth?

Mark Wilkinson

So I think it's a very interesting question. And so Peter Boyle that I mentioned that had done the silicon level Co design within Dirac kind of 10 or 15 years ago. So he did some work. It was around 2010 I think he took a. An algorithm, No, it wasn't 2010, it was more like 2016. Downloaded a highly optimized AI algorithm from Baidu Research and worked on it on optimizing it for GP us using all the stuff all the work that he had done on optimizing HPC workloads for GP us and got a factor 10 speed up. So I wonder if there's opportunities for us as a community to start engaging with AI tools and applying the things that we've been doing for a long time. I don't maybe they are as optimized as they can be, but maybe they're not. And if we then you can improve the AI tools, we may also learn to ways that we could start using them and then we'll have influence. I think for me, I think we have to have a, just the economics of it mean we're not going to be driving the market if we're not doing something connected with what the market is interested in at the moment. I mean on the emulation side, I think that's really interesting. But I, I wonder whether ever there's obviously a move away from double precision from some companies. I wonder when the next thing after LLMS comes and needs it's kind of doing AI for science and actually needs double precision again, whether GP us will suddenly have a lot more double precision. But I think what you're doing all all the work and emulation is important in the in the interim,

Jim Ang

either you want to respond to Terry. OK, next, next.

[unknown speaker]

Hi, Jim. Thanks for organizing this panel. Did you want to bring up the topic of Co design for medium sized customers?

Jim Ang

Please proceed.

[unknown speaker]

Well, it's some question I asked Jack and it seemed like you had something you wanted to say on the topic.

Jim Ang

I just wanted to say in response to Mark's comment, come to our panel session on Thursday, generative AI for chip design. OK, so please.

[unknown speaker]

Well, it's it's the same question. You know, Jack's talk mentioned, you know, companies with trillion dollar plus market caps can afford to sustain hundred person, 200 person ship design teams. And they they purchase so much volume that the founder is willing to basically do whatever customizations they want. But there's a large number of institutions that don't have that level of investment, but they still need the Co design. So the question is, how can they get access to these benefits without having the level of capital expenditure?

Jim Ang

Let me provide an answer directly and then I'll see if the panel wants to add 1 of the things Jack asked was, well, how do, how do we, how do we do this? Well, I'll go back to Terry's original slide. We're talking about Co design is there's a vertical component through the technology stack and a horizontal component where you talk about the different communities that are engaged.
One of the key points that Jack made was the HPC community has a long history of architecture research that we were built on that where a lot of this was R&D in the academic world. This was not about profit, this is not about product. This was about experimental architectures for supercomputing that many of them didn't didn't work. They were not commercially successful, right? But we still learned this was research, architecture research and I think we need to get back to that, that that kind of driver for computer engineers for hardware architecture research that is truly Co designed. But the, the sources of support for that will probably start with government. And I can't speak for the US government, but I can say I hope that they will recognize the value of investments in this area for, for the end of Moore's law, right? So we need to be looking beyond. There's already a lot of activity and a lot of investment in quantum. And actually, I would say quantum computing is probably a good example of full stack Co design. We're going to talk about quantum algorithms, quantum applications, the the runtime systems, the software, obviously qubits and then hybrid classical quantum architecture. So this there's a lot of R&D activity right now. And granted there are a lot of commercial players there too, but I know the Department of Energy is investing in neuromorphic computing concepts. A lot of that is it's not production, it's not a product, but, and some of those will fail, but it's research, right, that there are similar activities for other types of novel computing architectures that are well away from a product. So they're, they're places for innovation to be pursued largely in the in the interest of energy efficient computing, where if you think the only way to get to computing performance is, you know, GW power plants, then, you know, we should all change, change fields to nuclear, nuclear reactor engineering or something. But my panel, what would you like to say in response?

[unknown speaker]

Your comment about quantum computers actually fits perfectly what we said to what we said before. Quantum computing Right now, there's room for innovation. There's room for trailblazing. There's room for, you know, there's no big money yet. OK, Yeah. So let's see next year somebody invents an algorithm that can speed up, keep learning by a factor of 1000 and using a quantum computer that will change the game significantly. And that will be the point where, you know, innovative ideas become less important if they are not following the path of the money.
Yeah, that's why quantum computing right now is so interesting and very interesting open for innovation.

Mark Wilkinson

So I think another thing we need to be doing is more training of and another generation of people who are interested in this because it's something that we've found. So the, the kind of work that you were describing, I've been surprised at times how a, you know, you were, you were finding a result and then trying to understand it. And, you know, the, the whole scientific approach, which is what we should be doing, had little curiosity. There often is in communities. They'll do a profile and they'll say my code is memory bandwidth limited and that's it. And then they stop there. They don't ask why, they don't ask could I do better. Some communities like Lattice QCD take that as a challenge and then they go, well, how can I completely redevelop my algorithm in order to be able to adapt to, to new hardware? And so I think one of the things that in, in the UK we're running, and I know that there's one here as well, is cluster challenges for students to try again, to get people, young undergraduates and graduate students to really engage with. Again, that's still system level Co design, but I think the more we push in that direction and reinvigorate the, the curiosity cause there's probably computer scientists in the room. So I shouldn't be too negative, but I think in the UK we've often seen that computer science hasn't been driving Co design forward in a way that you might, we might have hoped. It's often been developing a slightly better algorithm for, for something rather than

the, the kind of things you're doing, which is really, you know what, what it should be about South to answer the person's question, I think we, there's probably, there's, there's a route, as you said, if by approaching universities or, or groups that are not the, the big companies that will help to seed more people who are able to do this. And then there'd be there'd be some momentum that would mean that more more companies, more small companies could get involved because they could leverage the teams that would then be built up in universities.

Rio Yokota

So from the viewpoint of education and the and the next generation. So when we talk about Co design now, I think we're talking about like human designers of hardware and human designers of software Co designing something. But I think the Co design in the future will be humans Co designing with AI designers, right. And, and, but I, I think there are many people that would agree, you know, AI is not quite capable of autonomously doing all of this on its own. It's, it's a good copilot, but the Co design needs to have a lot of, you know, verification right from the human side. We need to have checks in place. One anecdote of my student trying to vibe code his way through like a high performance kernel is that you know the AI eventually cheats, right? Unless you have very strict checks of what it's doing and how it's actually doing the evaluation and, and thinking that it's performing well on the benchmark, it's actually not right. It actually modified the test so that it scores well. So you, you need to have these checks in place. And I think Co design with AI is all about these checks, right? To, to have, if you have enough checks in place that you are satisfied with the sort of the, the validity of the, the outcome and that you can really say that it's improving in the metric that you've specified. And then I think you can actually offload a lot of the, the tasks to the AI and, you know, let it go. But yeah, so designing these checks, designing tests, designing, you know, like unit tests or integration tests, all of these things are now becoming increasingly important, right? Do you trust someone else to write your code? If you if you do, then that someone else might as well be AI, right? It's so being able to trust someone that's not you is one step forward, right? And that someone else not being human is yet another step. But you know, the the line between trusting another human and trusting AII think is there's a fine line, right? It's it's really blurry. If you can't trust AI, then how can you trust a human as well? Right? I I don't trust anyone else to write my code. Right. So, So yes, I think it's all about the checks.

Jim Ang

What is trust? Well, we can, we can, we can come back to that. But next, next question.

[unknown speaker]

All right, I'd like to ask one question to each of them. So it's not a question to be answered by all of them. Just targeted question for each of them. It's been associated with their talks, sure. So it doesn't matter the order. So I'll start with Rio Yokota first. (1) Have you thought about defining programming paradigm in which you would express operations, whether they are computer operations or data transfer operations and also factor the accuracy as well. So then you could perform algorithm exploration to not compromise A courtesy, prevent catastrophic cancellation for instance, or things like that. And at the same time at the end, you know, try to achieve the fastest runtime under, you know, a breadth of architectures. So then you can do that Co design of algorithm, numerical performance, all of those things. I'll put the second question for Mark Wilkinson. (2) All right, so you've done some not playing on the core frequency to try to lower power consumption and make sure that, you know, you did not compromise the runtime.
Have you thought about doing something similar on the IO side, whether it's on the networking, you know, try to downgrade, you know, PC lanes or bandwidth. And I saw that you did actually an experiment on improving the local local latency on your storage. But you could also measure the the value from a power savings perspective, for example, or using other type of storage devices to, you know, reduce runtime and reduce power consumption. And then the last question is for Georg Hager. (3) Have you done also any profiling both hardware and the software, which usually are

different tools from an IO standpoint to for example, figure out and actually do this at very large scale to verify for example, if the amount of bandwidth on the networking devices, it's actually over provisioned? Because if it is, which is what I personally believe and I, I would even make a blunt statement that we are using at very large scale, no more than say 10% of the link bandwidth Y you spend between 20 to 30% of the total computing infrastructure on something that we are not using And similar experiment and analysis on the storage as well. Thanks.

Jim Ang

OK, Rio, you want to start?

Rio Yokota

Yes. So thank you for the question. It's a very interesting one. So I, I think at the moment the languages, the, the computer languages we use they, you know, the, the only way we control precision is through like types. We have these type languages and it's somehow a very indirect way of achieving a certain accuracy that that's why people default to double. And, you know, there's really no adoptive or dynamic way of actually controlling or even, you know, like probing the variables for the dynamic range and adjusting. So yeah, I, I, I think we need, you know, with this very low precision appearing now and this very mixed precision capability, there's a huge graduation of precisions available to us now, right? It's very different from back when we had only like double and single precision. Now we have this whole spectrum of precisions that makes sense. Now it makes sense to, you know, actually think about it, think about a language that can, can take advantage of this. You know, you're exactly right. We there's no such language that has, you know, like this kind of precision as a first class citizen And, and actually it's adopted and and can, you know, achieve a given users precision while being very dynamic about the types it uses. I think this is something that is very manually done at the moment. So yes, it would be great if we have this kind of new language that can express, you know, precision as as something that is not like, you know, strongly typed or, you know, very rigid typing is. Yeah, it doesn't seem ideal anymore.

Mark Wilkinson

OK, So you asked about looking at the power of other components from just the top speed of the computer. So we haven't done that explicitly, but it's an interesting question that what I would say is the workloads, the Lattice UC workload is driving that study. It's essentially getting close to wire speed on memory bandwidth and network on with so AI think, but it would be interesting to see what we could tweak and maintain that and whether we could save energy and that's something we could. And yeah, on the IO go we didn't, but again would be interesting.

Georg Hager

OK. So your question to me was why are we building all the provision networks and provision IO subsystems and disk systems that we're only using for 10% of the time, but even that's maybe 1%. That's an extremely important question. I like it very much because it's absolutely true. And the reason why we're building these kind of subsystems is because we still have this thinking in mind for lots of the software that we are in these systems that we do bulk synchronous program, bulk synchronous execution, which means you're computing 9% of the time if you're lucky, and then 10% you're communicating. In order for those 10% not to get out of hand, we need a lot of bandwidth because everybody's communicating at the same time. If we had a more relaxed attitude towards computing, like assuming or allowing for more, less synchronized computing, that would relieve a lot of the pressure on the network and the IO substance. Now, how do you do that? You can impose that in two ways. First of all, you could go all the way, you know, and have a transformative change of what you do in terms of programming models, like you could use. You could write all your software HPS, you know, massively scalable or not debatable programming model and then you have all the software that you have to change. You could do that. You can claim that this is solution called problems, or you could go incremental. And that goes into what I was saying in my talk, that

sometimes it's possible to get a little bit more asynchronous in your computing just by trying to find ways to reduce global operations, to reduce synchronization and so on, and let the system be synchronized by itself. And it's surprising how much speed you can get just by trying to remove the barriers, not because of the synchronization points, not because you're removing the time it takes to do the synchronization, but because you're letting the MPI program desynchronize overlap communication with Bristol computation. That's the incremental part.

[unknown speaker]

OK, so sort of to cast things in a bit more negative light. In the past decade, I've been involved in several Co design efforts, one of which made it to market as part of Coral 1, Summit and Sierra. And even though it wasn't as full of virtual cycle of Co designers, we would have liked to. Because in the end, again through economies of scale, the building block of the CPU and the building of the GPU were mostly fixed. And we were trying mainly to figure out how do we make them work together. Which was still an interesting problem because AC922 were the first case of actually having coherent access between a GPU and CPU in a product. I don't unfortunately see something that sort of capable happening after what we got with MI 300 AI think that was probably the last example. Right now what I'm seeing is not only is FP64 going away as we saw, but also the fact that all of these new precisions that are coming up are essentially available only within a matrix unit. It was interesting showing the pseudo code before for your compensated scheme that there were all these promotions and emotions between FP-32 and FP16 happening because in fact the functional units for the lower precisions are not available for everything, which makes things even crazier. Suddenly I'm told you need to work in FP8 to get the great performance, but you need to be promoting and demoting for anything that's not essentially a matrix operation. In addition to that, and again pointing to the fact that yes, the first multiple method is one of the greatest discoveries of the 20th century formally at this point. But I do remember as a grad student in the mid 90s getting a talk at my department by one of the two people responsible for the method and he came in triumphantly and he said I've solved turbulence for you. Using the fast multiple method will be able to do everything. Guess what, we're in 2025. Most CFD codes are not using the fast multiple method. I don't actually see them using the fast multiple method. I don't see them using spectral elements, which again is geared towards matrix multiplication. And that was my research group's strength and my thesis advisor kept saying future is spectral elements and discontinuous galactic. They have taken more of the market, but they still haven't really gained everything. So if we look at what we would like to have and where the architectures are going, I see a very dark future. I don't see a clear way out of it other than the US government, mainly because it's the 800 pound gorilla. Well so far maybe the Chinese government can become the 800 pound gorilla. But some gorilla needs to go in and basically say you know what? We can't just go to the whims of how to get a better meme generating machine or fake video generating machine faster and the quantum stuff. The part that bothers me is that the Co design can only be in this very abstract space right now, which we have for quantum classical supercommunity. How do we bring them together? Which is extremely vague at this point because of the dearth of clear algorithms. But when it comes to the quantum stuff itself, it's not driven by Co design. It's driven by the physics that the various teams are using. I haven't seen anything other than that in the supercomputing space right now. The choice of whether we go from hex to square lattices has to do again with error mitigation in physics. I haven't seen anybody saying we looked at what fits best for the algorithms. So just to give you a sort of different perspective, and maybe I'm wrong, maybe you have a better, more positive picture. But again, I'm not seeing what I'd like to see. My hope is that maybe I, an AI researcher, comes up with sparse based algorithms. True sparse. Not the silly sparsiness that we take the two smallest values in the matrix and squash them down to zero and call that sparse matrix multiplication performance. But true sparse stuff that we do with sparse holders and say, you know what, CSR is now the new thing and let's start doing things that way and let's start concentrating around that. That may be our saving grace and also somebody caring about AI in science. I don't know if any of you have seen this, a recent paper that says FP64 is all you need,

where they discovered that issues with pins are partially happened because of the fact that people were not using FP64 in the optimizer. And guess what, in the new hardware, it won't even be there.

Jim Ang

I don't know that I had a, there was a question there, but would any of the panelists like to react,

Mark Wilkinson

Sorry, I won't reply to everything, but there were a couple of points. I think the you made an interesting point about the need to all the promotions and demotions in all the calculations. I think that's something that we don't hear much about and it it is clearly it makes everything more expensive as a student doing exactly that at the moment. And she's using, she's looking at pins and again, finding the need for higher precision. And then also the the demotion, the the cost of promoting and demoting things actually driving it. I guess. So I would, I'm going to then if I'm allowed to throw back a question and this can be to you or to to anyone. So, and I'm not an apologist for any chip manufacturer, the amount of FP 64 is only decreasing slightly. It's just not increasing. And in all the applications that I'm aware of outside of AI, no one is actually using all the available FP 64 and by, you know, our most efficient code is probably using a quarter of theoretical peak. So how serious is the problem? As serious as we think or do we just need, you know, if we can, if we could use all the available FP 64? Now obviously there's the fact that you're now buying a chip which is much more expensive and not delivering anything that you're more. But aside from that, is there actually a problem?

Rio Yokota

So I think we we don't have a a clear solution to the promotion demotion issue at the moment. Yeah, like the previous question about having a new language to deal with this might actually be, you know, make sense. At the moment we're using, you know, like AI uses Python, right, mainly to write their code. And in HPC we are using a lot of C++ or, you know, some are still Fortran. And there's a certain restriction that comes from sort of the these languages that we are using and with now with AI writing the code, I think there really needs to be a new language, right? Something that works better with AI, first of all, works better with precision. And, you know, if we are not the ones writing the code anymore, doesn't really make sense to, you know, to keep using these old programming languages. I think we, there are so many changes happening now both to the hardware and from AI that we we, it makes sense now to actually have a new, completely new way of programming, right? I don't know what form this is going to take right now. It may be just another way of abstraction sitting on top of what makes our current languages look like assembly. It may be some level of abstraction or it could be something, yeah, completely geared towards AI vibe coding that, you know, we just interpret that for our own understanding. Maybe we make it human readable, but the actual syntax is more like geared towards AI, not necessarily like computers, but more towards these LLMs or these language processing machines. Because if we can do that, then AI will become much better at coding, right? I think one of the things that's holding it back is that it's trying to program in the language that we design to, for, for our benefit, not, not for AI. It's, it's not at all designed for AI to be, to, to make it programmable for AI. And, you know, we, we just want our code to be correct and to be fast, right? And, and there could be other metrics like energy consumption and like, you know, lower memory consumption, communication avoiding. But in the end, we, we have these clear metrics that we want to design for and, and we have our, you know, scientific problems that we want to solve. And it doesn't necessarily need to be the same programming languages that we're using now. There might be something, yeah, completely new that we could use to, to get rid of the promotions and the emotions of types, right. These these things may well, they, they will have to happen in hardware if the hardware is designed that way. But it, you know, it may be exposed to us in a very different way that that is less painful for us to to deal with. Yeah. So maybe, yeah, it's time to design a new language. I'm not, I'm not proposing like let's do a new, you know, domain specific, you know, there's

thousands of DSLs at this point. I'm not, you know, proposing a new one. I'm just saying something needs to change in the yeah, programming side.

[unknown speaker]

One comment to that, I think that adding more instruction of the instruction that you already have will fix anything. I don't think that the problem is that somewhere between the byte coder and the actual execution of the code, there has to be a scientific process that sort of looks at the machine and the software target traction and sees what's going on and can scientifically judge that this is good on that. And I hate to say that the group of computer scientists and other computer scientists, but the computer science curricula are doing a bad job of computer scientists. They introduce it with coders, with complexity management, with abstraction. You know, designers. Abstraction is the holy grail of computer science.

Jim Ang

Well, let me add an additional comment. So I'm not a computer scientist, I'm a mechanical engineer, very different. But I wanted to to address the comment that, you know, hardware design and having chips is really expensive. So this might be a place where the US government is still does have a Chips and Science Act. It is still proceeding. It's still a very significant investment on part of the US government. And with my colleagues in the back of the room, we're we're tracking what's going on in, in the Chips and Science Act. And there are things going on where you could be talking about hardware prototyping that's similar to software prototyping. You know, you don't, you don't start off building an application with building in all the regression tests and all the things that you need for production software in a similar sensor. Places where you can be thinking about hardware design for test prototype where you know, the the mass set doesn't have to be built on quartz. It could be built on soda glass because you're not going to you're not going to have a production run of millions of chips. You could you could do a lot of your prototyping and not the latest technology node. There's a lot of investment in triplets, so maybe you don't have to design a new GPU and CPU and memory interfaces in an SoC. Thinking about silicon in a package system, in a package where only your specialized accelerator has to be custom designed and the rest of the the system and package can be integrated for test purposes, right? So you can cut corners. Also when it's going to be a research prototype, it's going to prove out your innovative accelerator actually works and has performance improvements. Once you have that raw data, then you can argue for, let's see, going on to NRE non recurring engineering or or taking a prototype and making it into a product. So those are things in the future that I think will will benefit from Co design for the the hardware designs. The innovations you're pursuing are have have drivers from applications and algorithms and so forth. Other question.

[unknown speaker]

It's more a comment.
You're sort of advocating for a new programming language to be used by AI or maybe people.
What we know about AI is it needs a large corpus of examples to be good. So by default, it will be horrible that the new programming language found a lot of examples. Otherwise, I like the idea. It has interesting challenges. And I'm a computer scientist, so I'll take it.

Jim Ang

OK. Thank you. I realize we're overtime now, so I'm going to turn the mic back over to Terry.

5. Post-Workshop Findings, Recommendations and "Next step" Strategies

The workshop finds that the present state of quantitative co-design is still nascent with plenty of divergent paths to choose from. There are a number of recommended "next steps" that should be followed to increase the usability of quantitative codesign of supercomputers.

5.1 Workshop Findings

- An emphasis on the need for more investment in research at universities and government laboratories to experiment with new computer architecture designs. 93% of the machines comprising the Top 500 list are based on commodity parts. Sadly, co-design is becoming mostly aspirational.
- The gap between processor speeds and data movement is becoming more problematic. Again, this emphasizes the need for research and experimentation to improve this situation.
- With the advent of superscalars, some see an increasing risk that scientific computing is becoming too inconsequential to vendors.
- Effective automated ways to capture operational and performance related data is an active field and AI/ML are making impressive strides. Such information can be used for many purposes included future machine planning and application performance tuning.
- The rise in importance of lower precision floating point operations due to AI gives raises the need for new languages that can more fully express precision constraints and flexibility -- perhaps using a new typing model.
- Domain specific co-designs are worth exploring.

5.2 Evolving SQCS

The following positive results have been achieved with from the workshop:

- A large group of high performance computing professionals came together to pursue community building
- Monitoring journals (outcome and strategy) were discussed and templates provided to guide the process of data collection and the use of these data
- Videos of the invited talks and panels were recorded by SC's Live Stream AV team
- Discussion on Vision and Possibilities of Quantitative Codesign of Supercomputers were discussed, and ideas for future work were identified
- This workshop report was written to document the results

In addition, monitoring journals (outcome and strategy) were discussed and templates provided to guidethe process of data collection and the use of this data.

The America's Center Conference Center and St. Louis receives mostly positive marks from SQCS'25. Positive aspects include:

- The audio-visual team provided our best support to date in the history of our workshop.
- The room size and location within the conference center for SQCS'25 were fine.

The following offer areas of potential improvement:

- Location: St. Louis provided relatively fewer restaurants in the convention center area when compared to past locations.

5.3 Next Steps

- Continue the website (Link)

  Provide ongoing support to Quantitative Codesign of Supercomputers website. This web presence becomes an anchor for announcements and a source to discover resources and pertinent email addresses.

- Disseminate the Workshop Report

  Providing this post-workshop report of the event will an important resource for the symposium's community building objective. The contact data of the participants interested on receiving the report have been collected and will be used to spread the report in the community

- Track Potential Mission furthering opportunities

  This follow-up activity is to ensure that a wide segment of high performance computing is monitored for events, interactions and publications for opportunities to advance high performance computing through quantitative codesign concepts.

- Advance the Quantitative Codesign agenda with a 2026 Symposium

  Finally, we are encouraged to repeat the workshop in 2026. This fourth workshop should consider *Upcoming Machines and Emerging Technologies*. The state of High-Performance Computing (HPC) is undergoing a fundamental transformation, moving beyond a sole focus on raw peak FLOPS (floating-point operations per second) to prioritizing time-to-solution, energy efficiency, and AI-simulation convergence. The landscape is defined by the rise of "post-exascale" machines, the integration of generative AI into scientific workflows, and the emergence of hybrid quantum-classical architectures. How can quantitative co-design be applied in such a swiftly moving and chaotic time? What should be our priorities as we plan for new machines? What technologies are poised to be gamechangers in the near future? We wish to consider these questions from several different perspectives and apply quantitative codesign benefits toward broad-scope type objectives.

Appendix 1 – Related Activities

Among the related activities that we wish to augment are the following:

- The Center and Application Monitoring Session held during the ECP Annual Meeting.
- The International Workshop on Monitoring and Operational Data Analytics (MODA) held with the annual ISC High Performance conference.
- The Workshop on Monitoring and Analysis for High Performance Computing Systems Plus Applications (HPCMASPA) held with the annual IEEE Cluster conference.
- The Workshop on Performance Monitoring and Analysis of Cluster Systems (PMACS) held with the annual Euro-Par conference.

Each of these related activities share an interest in the wealth of information exposed by these systems about how the system resources are being utilized. Our Symposium is unique in its emphasis on applying data to improve the codesign process. The Quantitative Codesign Symposium also has a distinguishing format and venue.

Appendix 2 – Symposium Speaker Biographies

Jack Dongarra – Leadoff Speaker and Panelist

Jack Dongarra specializes in numerical algorithms in linear algebra, parallel computing, the use of advanced computer architectures, programming methodology, and tools for parallel computers. He holds appointments at the University of Manchester and the University of Tennessee, where he founded the Innovative Computing Laboratory. He is a Fellow of the AAAS, ACM, IEEE, and SIAM; a foreign member of the British Royal Society and a member of the U.S. National Academy of Sciences and the U.S. National Academy of Engineering. He received the 2021 ACM A.M. Turing Award for his pioneering contributions to numerical algorithms and software that have driven decades of extraordinary progress in computing performance and applications.

.

Sadaf Alam – Invited Speaker and Panelist

Dr Sadaf R. Alam drives Bristol's frontier supercomputing capabilities as Director of Advanced Computing Strategy and CTO at the Bristol Centre for Supercomputing (BriCS). Since joining the University of Bristol in 2022, she has led transformation of research computing and data services, co-founding Isambard-AI, national federated supercomputing resources, and sustainable digital research infrastructure (DRI) strategies. Previously CTO at CSCS (Swiss National Supercomputing Centre), she was chief architect for multiple generations of the Piz Daint systems and the MeteoSwiss operational forecasting platforms. With a PhD in Computer Science from the University of Edinburgh, and experience at the Oak Ridge National Laboratory (ORNL), Sadaf blends deep technical expertise with visionary leadership in sovereign AI, HPC, quantum and data services architecture.

Georg Hager – Invited Speaker and Panelist

Georg Hager holds a PhD and a Habilitation degree in Computational Physics from the University of Greifswald. He heads the Research Division at Erlangen National High Performance Computing Center (NHR@FAU) and is an associate lecturer at the Institute of Physics of the University of Greifswald. Recent research includes architecture-specific optimization strategies for current microprocessors, performance engineering of scientific codes on chip and system levels, power and energy modeling of HPC workloads, and structure formation in large-scale parallel codes. He served as a PI in the ESSEX (Equipping Sparse Solver for Exascale) project within the SPPEXA DFG priority program. Georg Hager has authored and co-authored over 100 peer-reviewed publications and was instrumental in developing and refining the Execution-Cache-Memory (ECM) performance model and energy consumption models for multicore processors. Together with colleagues from FAU, HLRS Stuttgart, and TU Wien he develops and conducts successful international tutorials on node-level performance engineering and hybrid programming.

Rio Yokota – Invited Speaker and Panelist

Rio Yokota is a Professor at the Supercomputing Research Center, Institute of Integrated Research, Institute of Science Tokyo. He also leads the AI for Science Foundation Model Research Team at RIKEN Center for Computational Science. His research interests lie at the intersection of high performance computing, machine learning, and linear algebra. He has been optimizing algorithms on GPUs since 2007, and was part of a team that received the Gordon Bell prize in 2009 using the first GPU supercomputer. More recently, he has been leading distributed training efforts on Japanese supercomputers such as ABCI, TSUBAME, and Fugaku. He is the co-developer

of the Japanese LLM Swallow, and LLM-jp. He is also involved in the organization of multinational collaborations such as ADAC and TPC.

Mark Wilkinson – Invited Speaker and Panelist

Prof. Mark Wilkinson is the National Director of the STFC DiRAC HPC Facility (www.dirac.ac.uk), which provides high performance computing resources for the theoretical astrophysics, particle physics, cosmology and nuclear physics communities in the UK. Co-design of computing hardware and software is at the heart of DiRAC's design philosophy, working extensively with industry partners to deliver cost-effective deployments of highly productive, large-scale computing services. Mark is a Professor of Astrophysics at the University of Leicester, and his recent work focusses on the use of AI to enhance and accelerate simulations in astrophysics and cosmology. Mark edited the 2019 community-led white paper "UKRI National Supercomputing Roadmap 2019-30" and chaired the editorial board for the peer-reviewed "UKRI Science case for UK Supercomputing", published in 2020. He recently chaired the STFC AI Strategy Development Working Group. He currently co-Chairs the STFC Exascale Working Group and is a member of the UKRI Advisory Group for Digital Research Infrastructure (AGD).

Appendix 3 – Organizing Committee

Terry Jones – Chair – Oak Ridge National Laboratory, USA

Terry Jones is a Senior Research Staff member at Oak Ridge National Laboratory (ORNL) where he has worked since 2008 in the Computer Science and Mathematics Division (CSMD) as a Computer Scientist. Prior to that, he held a Computer Scientist position at Lawrence Livermore National Laboratory (LLNL). Terry earned a Master of Computer Science degree from Stanford University. Terry's research interests include system software for high performance computing, runtime systems and middleware, parallel and distributed architectures; performance monitoring; memory and storage systems; distributed clock synchronization, and resilience for complex distributed systems.

Estela Suarez – Co-organizer – Jülich Supercomputing Centre & University of Bonn, Germany

Dr. Estela Suarez is research group leader at the Jülich Supercomputing Centre from Forschungszentrum Jülich, which she joined in 2010. Since 2022 she is also Professor for High Performance Computing at the University of Bonn. Her research focuses on HPC system architectures and codesign. As leader of the EU-funded DEEP project series she has driven the development of the Modular Supercomputing Architecture, including hardware, software and application implementation and validation. Additionally, since 2018 she leads the codesign efforts within the European Processor Initiative. She holds a PhD in Physics from the University of Geneva (Switzerland) and a Master degree in Astrophysics from the University Complutense of Madrid (Spain).

Jim Ang – Co-Organizer & Moderator – Pacific Northwest National Laboratory

James is the Chief Scientist for Computing in the Physical and Computational Sciences Directorate at Pacific Northwest National Laboratory, where he serves as the lab lead for the DOE Office of Science (DOE/SC), Advanced Scientific Computing Research (ASCR) Program. PNNL's ASCR portfolio includes over 20 R&D projects in applied mathematics, computer science, advanced architectures, and computational modeling and simulation. His computing leadership role also intersects with foundational technology challenges associated with microelectronics and semiconductors. James helped organize the panel on co-design for beyond exascale at the DOE/SC workshop on Basic Research Needs for Microelectronics; served on the executive committee for the Semiconductor Research Corporation Decadal Plan; and was appointed by the U.S. Commerce Secretary to serve on the NIST Industrial Advisory Committee to provide input on R&D gaps for the CHIPS and Science Act. James has a BA in Physics from Grinnell College, a BS in Mechanical Engineering from the University of Illinois at Urbana-Champaign, and MS and PhD degrees in Mechanical Engineering from the University of California at Berkeley.

Jim Brandt – Co-Organizer – Sandia National Laboratories, USA

James (Jim) Brandt is a Distinguished Research Staff Member (Computer Scientist) at Sandia National Laboratories. Jim's research interest for the past two decades has been in holistic data-driven analysis of HPC eco-system resource utilization and state. He leads the development effort for Sandia's Lightweight Distributed Metric Service (LDMS) which has been in production use for a decade and installed on largescale systems across the DOE and NSF. Jim also leads SNL's AppSysFusion project, which enables run time combined application+system monitoring, through the interoperability of LDMS with other tools

including Kokkos, Darshan, and Caliper. Jim leads work in the area of application of AI/ML to modeling and optimization of application resource utilization and anomaly detection. Jim has a M.S. degree in Computer Engineering from Santa Clara University and a B.S in Physics from California State University Hayward.

Mike Jantz – Co-Organizer – The University of Tennessee, USA

Mike Jantz is an Associate Professor of Computer Science at the University of Tennessee, Knoxville. At UT, Mike leads the CORSys research group, which aims to design and build innovative system tools and techniques to achieve faster, safer, and more efficient execution on modern and emerging architectures. His group has conducted and published research on a variety of topics related to computing performance and efficiency, program profiling and analysis, runtime data management, and dynamic compilation. His work is supported by a number of government and industrial institutions, including the National Science Foundation (NSF), the U.S. Department of Energy, and Intel Corporation. In 2020, he received the NSF CAREER award for his proposal on application guided data management for complex memory systems.

Appendix 4 – Attendees & Workshop Photographs

We noted approximately 80 in-person participants with a few participants coming and going during the morning; we were unable to collect information on remote participants. We collected names and email addresses for our attendees.

Last year, our SC'24 attendance was approximately 70, and the year before that we had 64 in-person participants.

These pictures show the SQCS'25 room forma and a moment from the presentations of Mark Wilkinson, Rio Yokota, Georg Hager and Sadaf Alam..