Workshop Report:
2023 International Symposium of Quantitative Codesign of Supercomputers

Version: 1.1

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'23 workshop committee chair Alfredo Goldman, vice chair Dejan Milojicic; the anonymous workshop proposal reviewers who provided helpful guidance; SC'23 conference chair Dorian Arnold; our web designer Douglas Edwardson; and the audio-visual team that provided support for our symposium.

Thank You!

– *Terry Jones,* chair Quantitative Codesign of Supercomputers

TABLE OF CONTENTS

2023 International Symposium on the Quantitative Design of Supercomputers
Held in conjunction with Supercomputing '23
Denver, CO – November 12, 2023

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 designs. The recent National Strategic Computing Initiative Update Report has recommended that we promote 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

"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 mircro-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 2023 Supercomputing conference. Due to COVID conditions at the time, the workshop was held in hybrid model with both in-person and virtual attendees and speakers. 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 Agenda

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 speakers were chosen who, as an aggregate, provided codesign perspectives on common misunderstandings of what an ISA actually provides, ways in which AI can be employed in this field, and experiences in bringing about a holistic monitoring system at a major supercomputer center.
- A collection of position papers from an international collection of experts with diverse backgrounds in codesign, HPC system software and middleware research, center wide monitoring and operational aspects, bringing HPC products to market, and application / libraries.
- A moderated discussion of audience, speakers, and panelists was included to enable both technical discussions and community-building.

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

As with our previous Symposium, the COVID pandemic had an impact on the format and character of the workshop. This was the second time for the SC series of conferences to ever have a hybrid format: SC22 supported both in person attendees at the Dallas Convention Center in Dallas, Texas and remote attendees though the revamped SC22 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 (see Figure 1). 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.

The Symposium's program committee considered SC'22's Live Stream hybrid format a "mixed-bag". Last year, SC'21 utilized a new *Hubb virtual interface* to provide coordination of in-person presentations working in concert with virtual zoom interface; last year's overall SC'21 experience was positive given Hubb was unfamiliar. For SC'22, it was decided to emphasize the in-person (in Dallas) experience. The new system was sufficiently complicated to require a fair amount of AV knowledge and/or training. Unfortunately for our symposium (which was scheduled on Sunday morning, the first timeslot of the whole conference), our AV support person was unprepared and we had significant AV issues during the first half of our symposium: (1) remote people coming in through the SC'22 website did not have audio; (2) our two remote speakers had difficulties starting their zoom presentation; (3) our in-person audience experienced a 20 minute delay while the AV person tried to figure out the AV set-up; (4) we frequently had serious feedback issues during the course of the day. It is our belief that these problems were largely a result of insufficient training for the AV person, and that the same technology could provide a positive experience next year in SC'23.

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 9:40	 Estela Suarez	**Co-design at system and component level: examples from the DEEP and EPI projects.** Optimizing the configuration of HPC systems to meet user requirements requires in-depth knowledge of application profiles and hardware limitations. A quantitative methodology, founded on the development of application-based benchmarks executed on representative hardware platforms, simulators, and models, proves invaluable for this objective. Benchmarking sheds light on how application codes and their core components perform on HPC systems, enabling the identification of performance bottlenecks and opportunities for enhancement on the software side. Moreover, it aids in understanding the impact of specific hardware features on application performance, whether positive or negative. During this presentation, we will share our experiences with benchmark-driven co-design approaches, derived from two European projects, namely DEEP and EPI.
9:40 to 10:00	 Wes Brewer	**Towards the Development of a Comprehensive Digital Twin of an Exascale Supercomputer** Over the past year, we have embarked upon an ambitious initiative to develop a comprehensive digital twin of the Frontier supercomputer. This twin includes: 3D asset modeling with virtual and augmented reality capabilities, telemetry data assimilation, AI/ML integration, simulations, and reinforcement learning for optimization. Key simulations under development include: (1) a transient simulation of the thermo-fluid cooling system from cooling tower to cold plate, (2) a rectifier loss model predicting heat generation and rectification losses, (3) a job scheduling simulator, and (4) a parallel discrete-event simulator to study network congestion. This digital twin offers insights into operational strategies, "what-if" scenarios, as well as elucidates complex, cross-disciplinary transient behaviors; it also serves as a design tool for future system prototyping. Built on an open software stack, we are in active discussions with a number of other supercomputer centers who have expressed interest in collaborating for future development.
10:00 to 10:30	[break]	[break location]
10:30 to 10:50	 Phil Carns	**Enabling Codesign in the Software Tools Ecosystem Project (STEP)** Software tools are crucial for understanding and optimizing the performance and behavior of scientific applications. To do this job effectively, they must operate at the crossroads of applications, system software, facility operations, and platform technologies. These constraints present a daunting set of challenges, but also a rich opportunity for codesign. This talk will provide an overview of the nascent Software Tools Ecosystem Project (https://www.ascr-step.org) with a particular focus on how it plans to apply the principles of codesign to address critical cross-cutting challenges.
10:50 to 11:40	Panel:	Scott Atchley (ORNL), Jim Ang (PNNL), Dave Hart (NCAR), James Lujan (LANL), Nick Wright (NERSC) Our panelists respond to workshop charge questions
11:40 to 12:25	Moderated Discussion	**Your opportunity for audience & panelists to dig deeper** .
12:25 to 12:30	Terry Jones	**Closing Remarks**.

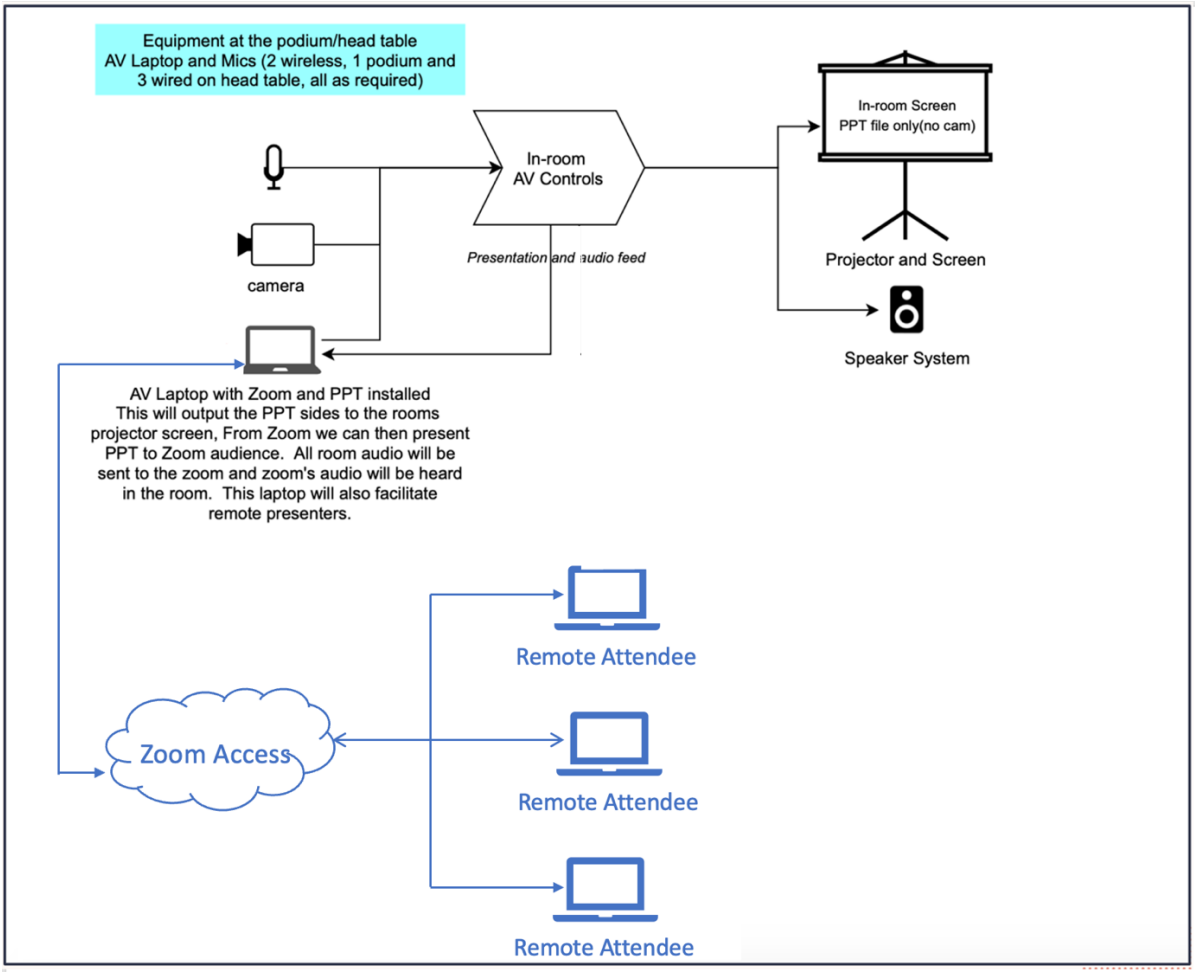


Figure 1 – Logistical Setup of Live Stream Format Used By Symposium.

4. Workshop Outcomes

4.1 HPC Contributions

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 guide the process of data collection and the use of this data.

4.2 Workshop Findings

The workshop panel and ensuing audience interaction produced a very interesting discussion. Each noted panelist was asked to comment on 5 charge questions:

- Considering the range of strategies and approaches that are available for doing data collection and quantitative analyses during the design process, are our supercomputing centers currently doing a good job of leveraging Quantitative Codesign (QC) techniques to meet computational requirements and goals?
    - Panelist A: No, it is still in early stages. We still rely on too many rules-of-thumb (e.g., checkpoint 50% of memory).
    - Panelist B: Yes, within the confines of our current procurement environment (limited vendors, limited components to what's off-the-shelf, tightly constrained budget).
    - Panelist C: Yes, QC plays a crucial role in helping understand our application behavior and performance space and helping to communicate that to the community to influence design parameters and measure success against designs that benefit those parameters. The strategies and approaches remain consistent, however, newer tools (both analysis and ways to ABSTRACT our instruction and data workflows) are needed to further the co-design Within the NNSA, there's a variety of missions and associated application types. This leads to the need for a variety of different computer architectures. One way to think about this is to map usage out on two axis: the complexity of data flow, and the complexity of control flow. AI might be medium data flow complexity and low control flow complexity. Other styles might have increases in either axes. NNSA cares about tailored architectures efficient for target applications -- this is the thrust for our codesign. For instance, we pushed memory bandwidth in the "Crossroads" system.
    - Panelist D: Reasonable Job. We are doing a reasonable job for designing machines. For applications - less so. Prioritize based upon decisions that will have greatest impact on the users and/or cost. Could always do more - volume and variety of data is the biggest issue. The workload is one of the first factors to consider in these discussions. Our institution has over ~9000 annual users from ~800 institutions. With each computer architecture generation we're seeing Increasing Need for Quantitative Co-Desig; More sources of data available. For Cori (NERSC-8), we used QC to decide how big the Intel Haswell partition should be compared to the Intel Knights Landing partion, what should the bisection bandwidth be; what should the storage capacity and bandwidth be. This was repeated in 2021 for Perlmutter. With Perlmutter, we started think much more about codesign in the software space: we had to decide which parts of OpenMP to prioritize. We used QC to do this. In the future, we will need to codesign the building and the power management with the machine because if you design everything for the gpu TDP -- you're over-engineering. We are coupling more and more to experimental and observational facilities.
    - Panelist E: No. If we're buying equipment and components are sitting unused, we haven't done a good job. At the end of Moore's Law, we no longer have this nice hierarchy of technology stack or software stack. We have to think much more about how a given technology has collateral impacts for other technologies up and down the stack. A circular diagram is much more accurate than an old-fashioned hierarchical view. The need for domain specific accelerators is increasing.

- What factors do you see as most important for enabling effective QC where it is currently being deployed?
    - Panelist A: Asking the right questions, can the data be collected, collecting the data, and analyzing the data. The first two are the hard part.
    - Panelist B: Understanding the applications that will dominate the system's workload. That includes the type of application, but also how are they planning to scale that work.
    - Panelist C: LAN is actively using QC in a variety of areas in both application performance improvements based on upcoming technologies (AMD, Nvidia, Intel), but also actively working with industry to influence future design that ultimately help increase performance efficiency of mission applications. The key factor for enabling effective QC is continuing to show success. Other teams will engage more based on the success of a couple of key application teams.
    - Panelist E: Energy efficient computing and cyber-security are a tremendous opportunity.
- What sort of changes do you think would be most impactful for improving the effectiveness and/or applicability of QC for supercomputing applications and systems?
    - Panelist A: Application profiles - Where are the bottlenecks? What are the resources used? Unused?
    - Panelist B: For organizations that don't have the 'deep pockets' to be involved in the design phase – we can interact with customizations for off-the-shelf components. At the same time, for this to be effective we would need more support from the vendors (e.g., here's all the options, here's what they mean, here's how they get put together). We also need improved capabilities to report and understand our power consumption.
    - Panelist C: We continue to improve the understanding of our application space, and we are still improving and finding new ways to communicate that to industry. We need to continue on this path, but ultimately it is a trade space of market drivers, funding, priorities within mission space.
    - Panelist D: Extremely easy to collect data. NERSC has PB of system logs, storage system logs, power measurements, performance data... Generating actionable outcomes is HARD. Requires domain AND data analytics expertise to produce. Turning this sea of data into an actionable outcome is actually quite hard -- we need to do this in a much more rigorous and scientific way with better ways of analyzing the data. I'd encourage centers to save as much data as they can -- vendors really love data.
    - Panelist E: Taking advantage of the US Chips act is an incredible opportunity. Energy efficient computing and cyber-security are a tremendous opportunity.
- Can you think of a specific example of an application, workload, and/or scenario where you feel that QC is currently being underutilized? In cases where QC techniques are limited or unused, what obstacles, factors, and/or trends are impeding the successful deployment of QC?
    - Panelist A: For example, how much interconnect bandwidth is needed for a given workload? Do tools exist to quantify usage? Can tools determine if the app scales linearly? Tools need to access data from resource-related software. Does the software expose what is needed? Can it be collected without introducing jitter to the system? How much data will be generated?

How much bandwidth is required for this data? Where will it be stored? How long will it be stored?

- Panelist B: QC is underutilized in power consumption monitoring and design processes. Getting better interfaces for getting power is important.
- Panelist C: An example - System management/operations - with more and more of this transitioning to open source, we need to engage more QC practices on not just determining what solutions to utilize but also how to influence future designs that will be positively impactful to large-scale (5k+ node) systems.

- Many QC strategies record and store detailed information about application behavior and usage patterns. What sort of obligations do tools that monitor application behavior have with respect to data privacy? What sort of privacy rights should applications teams have for data that is generated while their applications are running?
  - Panelist A: It depends on who has access. OLC policy is that all data (e.g., username/ids, project account, job id, node names) are anonymized before sharing externally. Internal users, however, do see raw data. Many of our analyses exclude staff accounts because our usage is different than production users. We stress test system (e.g., processors, interconnect, storage) in ways that production applications do not. Other analyses look to see if there are different usage patterns between different science domains (e.g., biology, materials science, astrophysics, CFD), so we need to see the actual project IDs.
  - Panelist B: We are a publicly funded institution doing open science. If we're publishing data about usage, we need to get consent. Need to comply with HIPPA, ITAR and so forth. I looked up cloud services to see what they said about this, and Google said very prominently that their service data is not sold to third parties.
  - Panelist C: No comment!

Recommended Reading:

- DOE-SC Basic Research Needs for Microelectronics: Oct 23-25, 2018. https://science.osti.gov/-/media/bes/pdf/reports/2019/BRN_Microelectronics_rpt.pdf
- PCAST Report: Revitalizing the U.S. Semiconductor Ecosystem https://www.whitehouse.gov/wp-content/uploads/2022/09/PCAST_Semiconductors-Report_Sep2022.pdf
- Government Role: Crossing the Valley of Death. https://www.nist.gov/chips/vision-and-strategy-national-semiconductor-technology-center:

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

There are a number of recommended "next steps" that should be followed to increase the usability of quantitative codesign of supercomputers.

5.1 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.

5.2 Disseminate the Workshop Report

Providing this post-workshop report of the event will be 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

5.3 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.

5.4 Advance the Quantitative Codesign agenda with a 2024 Symposium

Finally, we are encouraged to repeat the workshop in 2024. This fourth workshop should consider ...

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 – Speaker Biographies

Terry Jones

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

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).

Wes Brewer

Dr. Wesley Brewer is a Senior Research Scientist in the Analytics & AI Methods at Scale (AAIMS) group at the National Center for Computational Sciences (NCCS) at the Oak Ridge National Laboratory (ORNL). Prior to joining ORNL, he was a Computational Scientist for the Department of Defense High Performance Computing Modernization Program (HPCMP). He holds a PhD in Computational Engineering from MSU, an MS in Ocean Engineering from MIT, and BS in Engineering Science & Mechanics from UTK. Dr. Brewer has considerable experience over the past decade in machine learning for scientific workflows at scale on high performance computers, primarily for applications in computational fluid dynamics. He also has professional experience in numerical weather simulation, computational genetics, natural language processing, cloud computing, and scientific workflows. Most recently, he is interested in AI for Science (AI4S) applications at scale to enable scientific discovery.

Phil Carns

Philip Carns is a Computer Scientist in the Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL, USA. He is also an adjunct Associate Professor of electrical and computer engineering at Clemson University, Clemson, SC, USA, and a Fellow of the Northwestern-Argonne Institute for Science and Engineering. His research interests include characterization, modeling, and development of storage systems for data-intensive scientific computing. Dr. Carns received a Ph.D. degree in computer engineering from Clemson University in 2005.

Jim Ang – Panelist

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.

Scott Atchley – Panelist

Scott Atchley is a Distinguished R&D Staff Member and Chief Technology Officer with the Oak Ridge National Laboratory's National Center for Computational Science. He is the Systems Architecture team lead within the Technology Integration Group. Scott and his team focus on understanding technology trends and application needs to guide future system procurements. Scott has been heavily involved in DOE's Exascale Computing Initiative and Project. Scott served as the DOE Technical Representative for AMD's FastForward-2 Node architecture program and for AMD's PathForward program. Within the Oak Ridge Leadership Computing Facility (OLCF), Scott served as the Technical Project Officer for Frontier, OLCF's fifth leadership system. He currently serves as TPO and System Architect for OLCF's upcoming sixth leadership system.

Dave Hart – Panelist

Dave is director of CISL's Research Support Division, which provides a range of IT infrastructure and support services for CISL's HPC and data service environments. Dave manages the allocations processes for CISL's HPC resources and also has responsibility for and coordinates a range of other cross-divisional activities within CISL, including reporting and communications. He is also the co-PI for the Allocations Service award for the NSF's ACCESS program, and previously served as the director of XSEDE's Resource Allocations Service. His professional and research interests include metrics for measuring the performance and impact of cyberinfrastructure systems and activities. Before joining CISL in 2010, Dave worked for 15 years at the San Diego Supercomputer Center in a variety of leadership positions. He has an M.S. in Computer Science from Carnegie Mellon University and a Masters of Mass Communication from the University of Georgia..

Jim Lujan – Panelist

In 1984, while working on his Computer Science and Mathematics degree at New Mexico State University, Mr. Lujan supported the first parallel I/O and Fortran libraries on the early Cray systems at Los Alamos National Laboratory. He then transitioned to operating system development and support for the first version of Unix on Cray systems, and ultimately to cluster integration and project management with the deployment of the first large-scale cluster for use in the NNSA's Advanced Simulation and Computing program. Currently, he is involved in the project management, acquisition, and technical integration of current and future large-scale supercomputing systems and their associated R&D. He serves as the ASC Program/Project Director for the High Performance Computing division at Los Alamos National Laboratory. His project portfolio includes the Trinity, Crossroads, Venado, and ATS-5 projects.

Nick Wright – Panelist

Nicholas J. Wright is the chief architect and the advanced technologies group lead at the National Energy Research Scientific Computing (NERSC) center. Most recently, he led the effort to optimize the architecture of the Perlmutter machine, the first NERSC platform designed to meet needs of both large scale simulation and data analysis from experimental facilities. His research interests are in performance analysis of HPC applications and architectures and he has published more than 40 papers in these areas. Nicholas has a Ph.D. from the University of Durham in computational Chemistry and has been with NERSC since 2009.

Jim Brandt – Moderator

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 – Moderator

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 3 – Organizing Committee and Program Committee

Workshop Organizing Committee

- Terry Jones - Oak Ridge National Laboratory, USA
- Estela Suarez- Jülich Supercomputing Centre & University of Bonn, Germany
- Ann Gentile - Sandia National Laboratories, USA
- Michael Jantz - the University of Tennessee, USA

Workshop Program Committee

- Jim Brandt - Sandia National Laboratories, USA
- Florina Ciorba - University of Basel, Switzerland
- Hal Finkel - US DOE office of Advanced Scientific Computing Research, USA
- Lin Gan - National Supercomputing Center, Wuxi, China
- Maya Gokhale - Lawrence Livermore National Laboratory, USA
- Thomas Gruber - Friedrich-Alexander-University Erlangen-Nuernberg, Germany
- Oscar Hernandez - nVidia, USA
- Jesus Labarta - Barcelona Supercomputing Center, Barcelona, Spain
- Hatem Ltaief, King Abdullah University of Science and Technology (KAUST), Saudi Arabia
- Yutong Lu - Director of National Supercomputing Center in Guangzhou, China
- Esteban Meneses - Costa Rica National High Technology Center, Costa Rica
- Bernd Mohr - Jülich Supercomputing Centre, Germany
- David Montoya - Trenza, USA
- Dirk Pleiter - KTH Royal Institute of Technology, Sweden
- Mitsuhisa Sato - Riken, Japan
- Martin Schulz - Technical University of Munich, Germany

Appendix 4 – Attendees & Workshop Photographs

We noted 64 in-person participants; we were unable to collect information on remote participants.

Last year, our SC'22 attendance was 61 in-person participants and 19 remote participants.

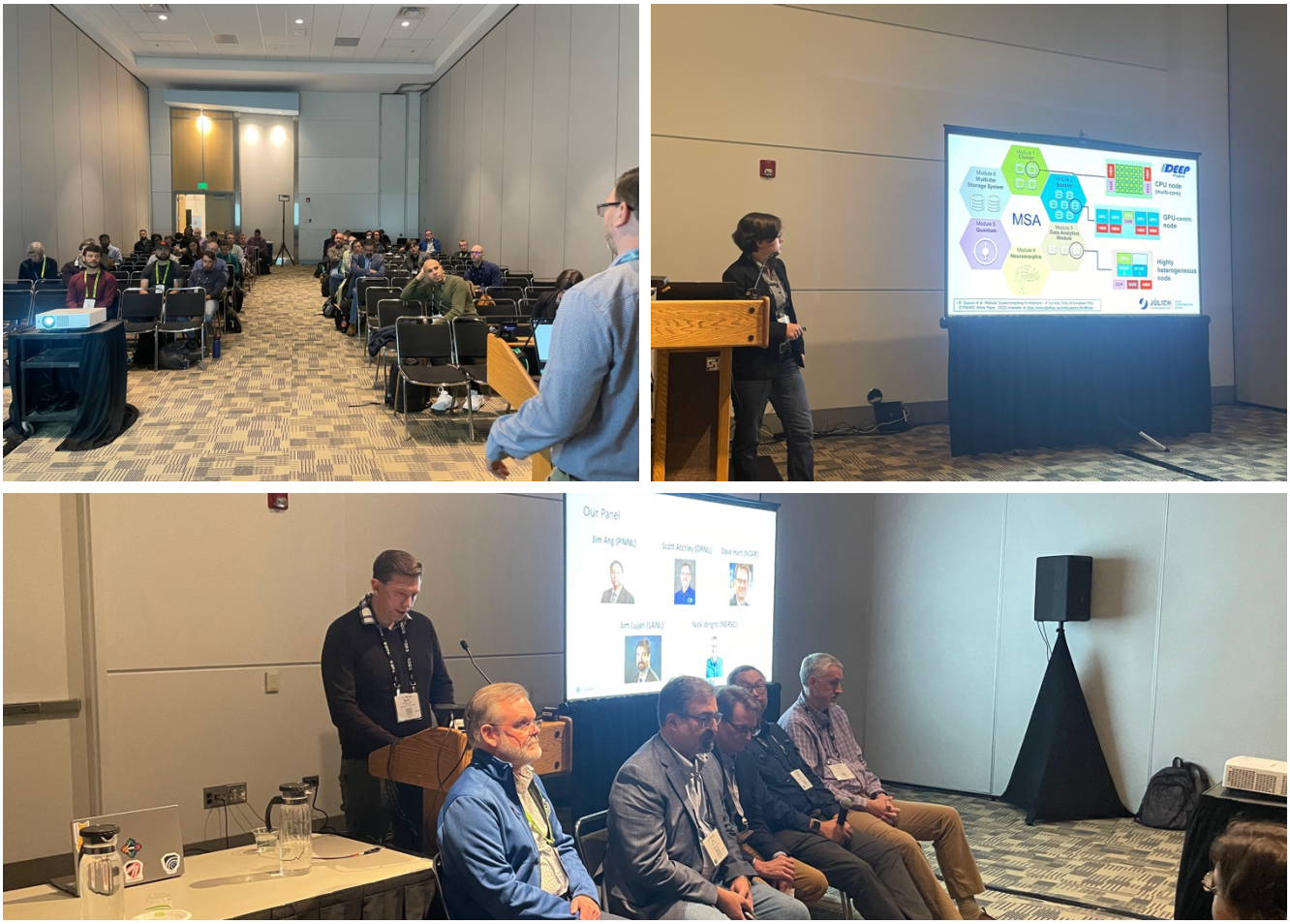
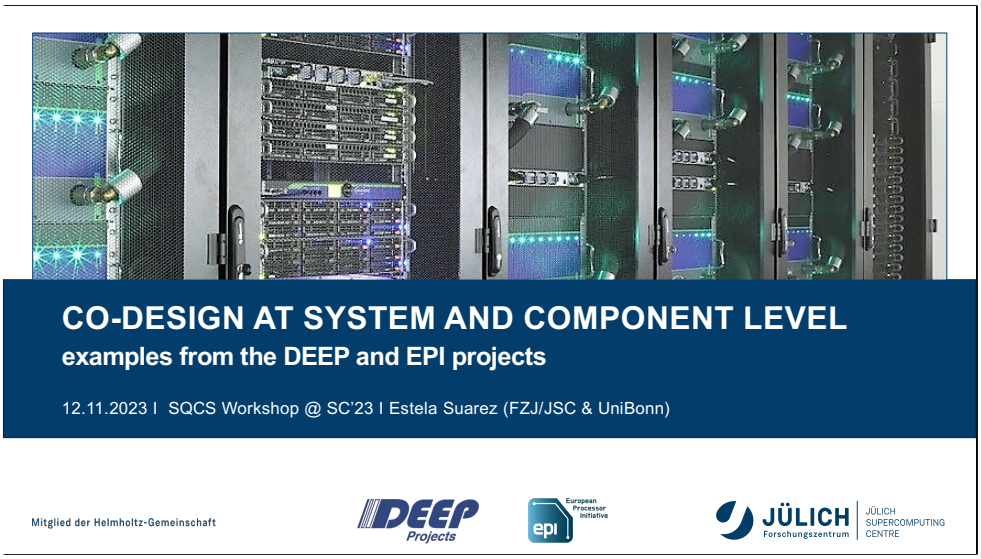


Figure 2 *Uppe-Left: Phil Carns addresses the audience during SC'23; Upper-Right: Estela Suarez presents our lead presentation; Lower-Panorama: Mike Jantz introduces our panelists: Scott Atchley, Jim Lujan, Dave Hart, Jim Ang, Nick Wright.*

Appendix 5 – Invited Presentations

Lead Presentation: Estela Suarez (1 of 15)



- 25 min + 5min Q&A
- Optimizing the configuration of HPC systems to meet user requirements requires in-depth knowledge of application profiles and hardware limitations. A quantitative methodology, founded on the development of application-based benchmarks executed on representative hardware platforms, simulators, and models, proves invaluable for this objective. Benchmarking sheds light on how application codes and their core components perform on HPC systems, enabling the identification of performance bottlenecks and opportunities for enhancement on the software side. Moreover, it aids in understanding the impact of specific hardware features on application performance, whether positive or negative.
- During this presentation, we will share our experiences with benchmark-driven co-design approaches, derived from two European projects, namely DEEP and EPI. In DEEP, our focus was primarily on the system level, while in EPI the target were on processor and core-level aspects. We will describe the differences between both approaches and the challenges that we found

1

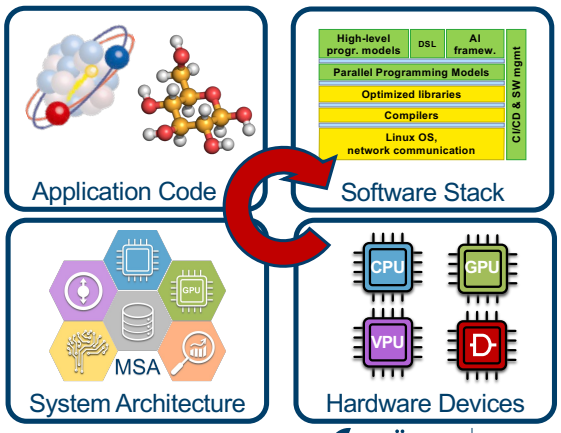
Lead Presentation: Estela Suarez (2 of 15)

OUTLINE

- **Co-design and Benchmarking**
- **Experiences**
  - DEEP Projects → System Level
  - EPI → Processor Level
- **Lessons Learned**
- **Summary**

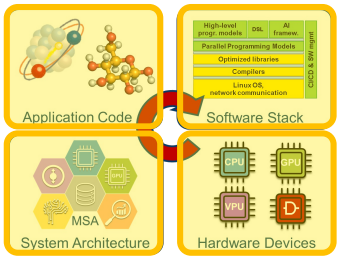
Co-design – my personal definition

- Study **interaction** between
  - **application** code,
  - system **software**,
  - **hardware** components,
  - and system **architecture**
- to **find** the **modifications** at each of those four levels
- **that bring overall best**
  - **performance** and
  - **energy efficiency**



Lead Presentation: Estela Suarez (3 of 15)

Role of Benchmarking in Co-design

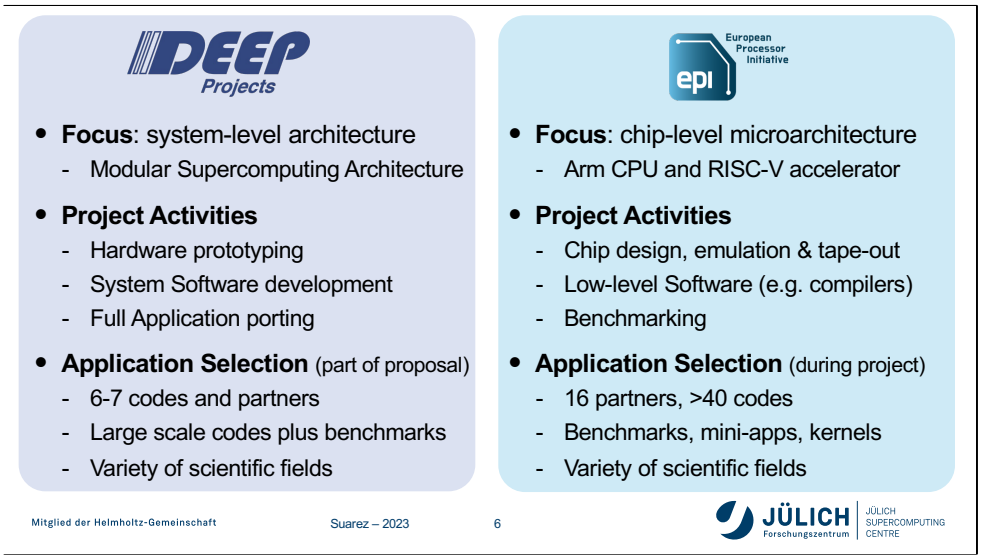


- **Characterize applications** through representative
  - Synthetic benchmarks
  - Mini-applications
  - Large scale use cases
- **Evaluate** different **software** versions/options
- **Compare** different **hardware** devices
  - Run benchmarks on hardware prototypes and systems
  - Model/simulate different architectural features
- **Determine best combination of resources** for given workload mix
  - Diverse application portfolio (not only an individual use-case)

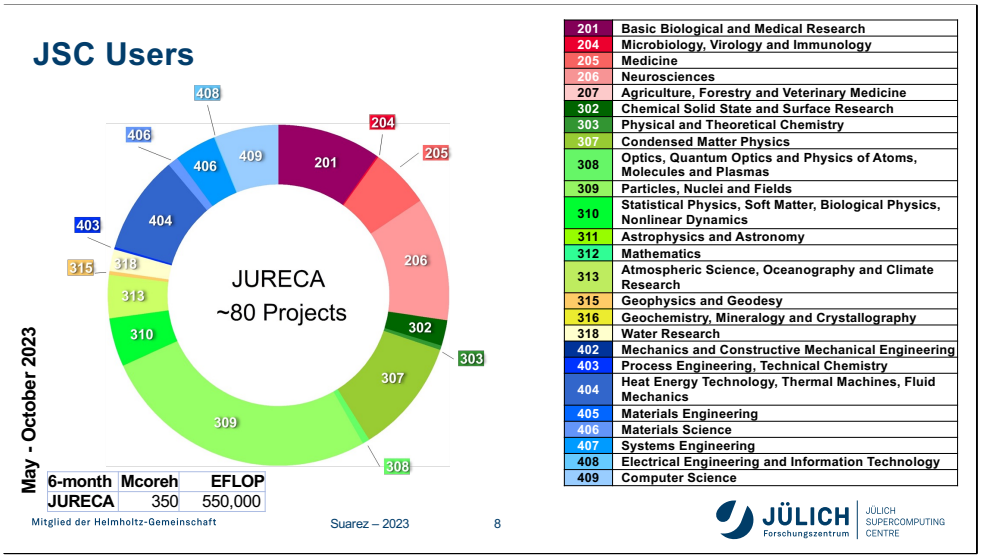
OUTLINE

- **Co-design and Benchmarking**
- **Experiences**
  - DEEP Projects → System Level
  - EPI → Processor Level
- **Lessons Learned**
- **Summary**

Lead Presentation: Estela Suarez (4 of 15)

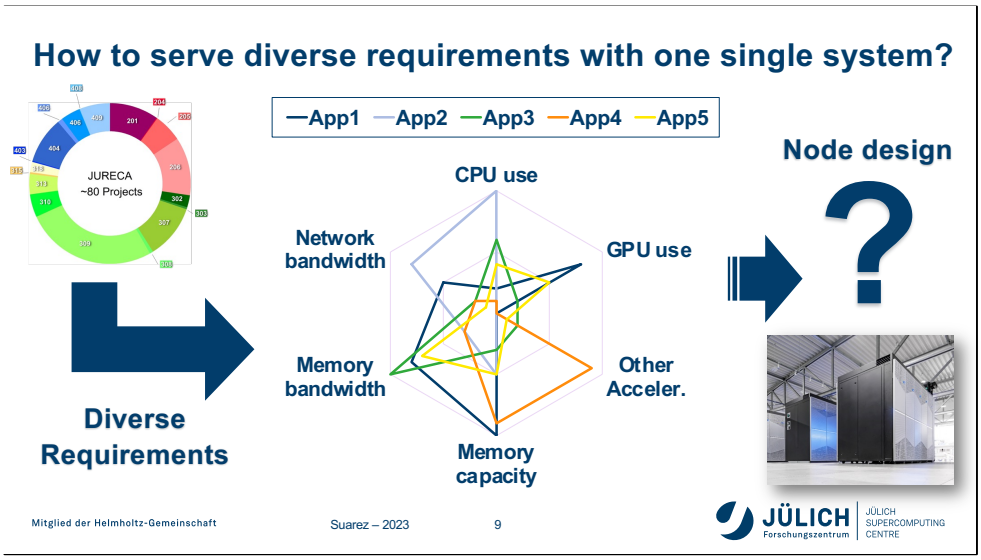


Lead Presentation: Estela Suarez (5 of 15)

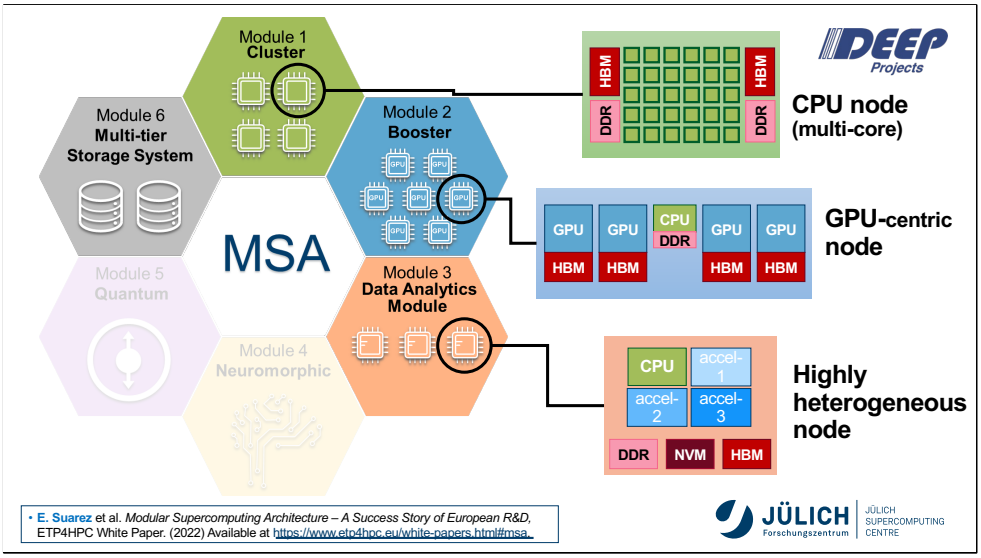


Compute time per ½ year (each period)

- **JUWELS**:
  - JUWELS Cluster CPU: 425 Mcoreh (~133.000 EFLOP)
  - JUWELS Cluster GPU: 8 Mcoreh (~ 24.000 EFLOP)
  - JUWELS Booster : 160 Mcoreh (~962.000 EFLOP)
- **JURECA**
  - JURECA Cluster CPU: 260 Mcoreh (~ 34.000 EFLOP)
  - JURECA Cluster GPU: 90 Mcoreh (~210.000 EFLOP)



Lead Presentation: Estela Suarez (6 of 15)

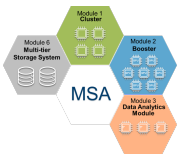


Co-Design of a Hardware Prototype

DEEP Projects

Fixed parameters

- System **architecture**: MSA
- **Design targets**:
  - **Cluster**: highest Byte/Flop ratio
  - **Booster**: highest energy efficiency
  - **DAM(*)**: highest flexibility & memory
- Installation **time**: 2020
- **Budget**: ~3.5 MEuro
- **Providers**:
  - Integration: Megware
  - Interconnect: EXTOLL

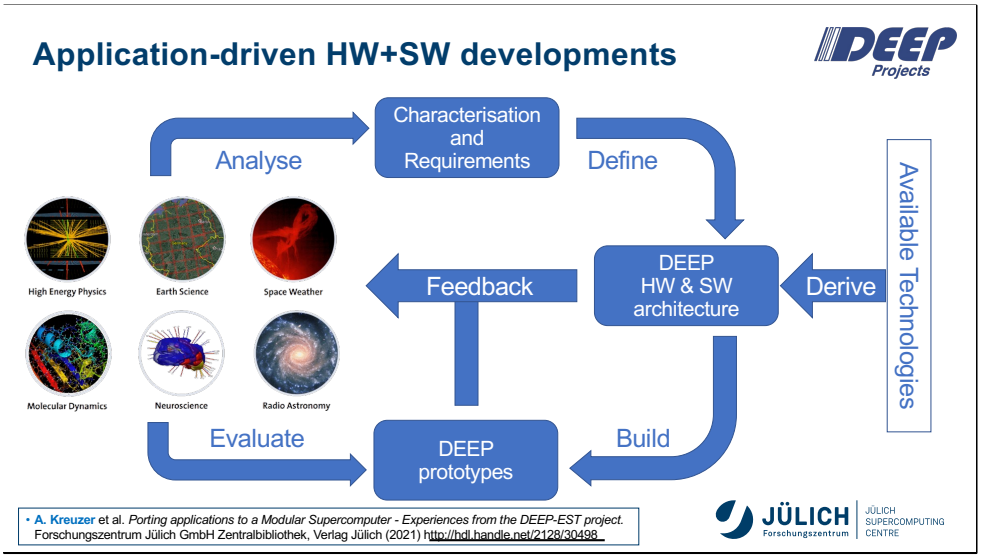


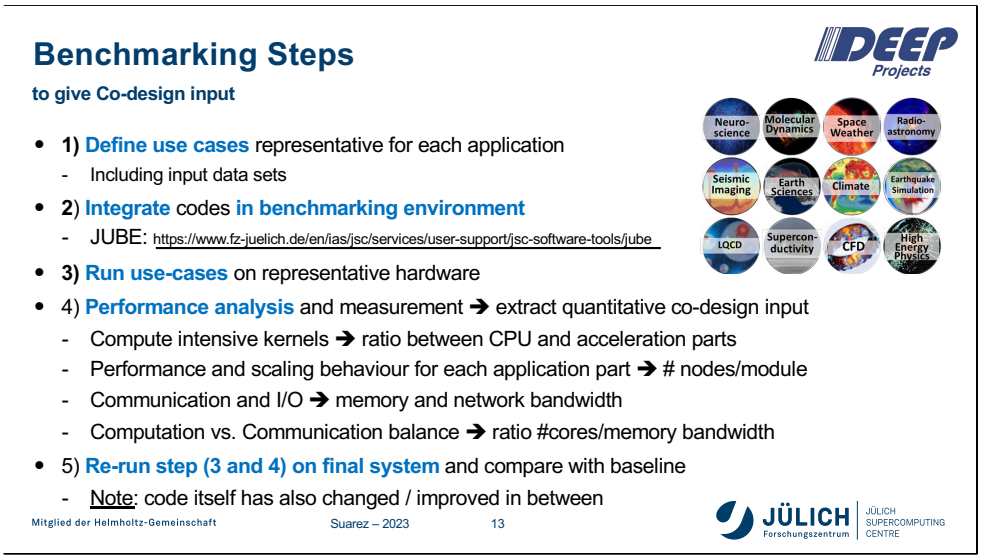
(*)DAM = Data Analytics Module

Design choices

- **Node number** in each module
  - Relative size of modules
- **Node design**
  - **Cluster**: CPU type and SKU (#cores, DDR size, etc.)
  - **Booster**: CPU type and accelerator (type and #)
  - **Data Analytics Module**: CPU and accelerator type(s)

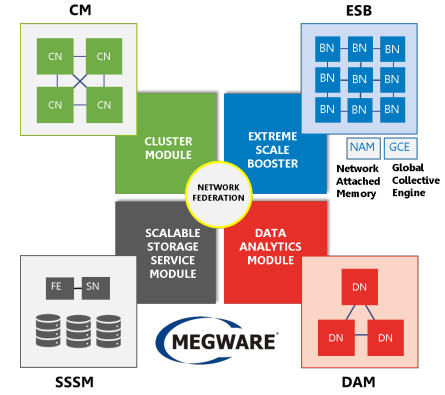
Lead Presentation: Estela Suarez (7 of 15)



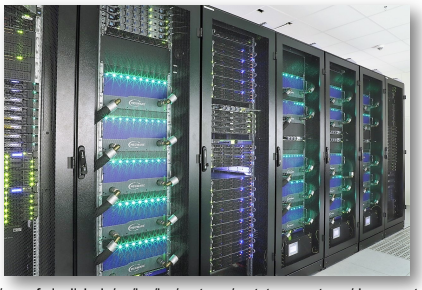


Lead Presentation: Estela Suarez (8 of 15)

DEEP-EST prototype



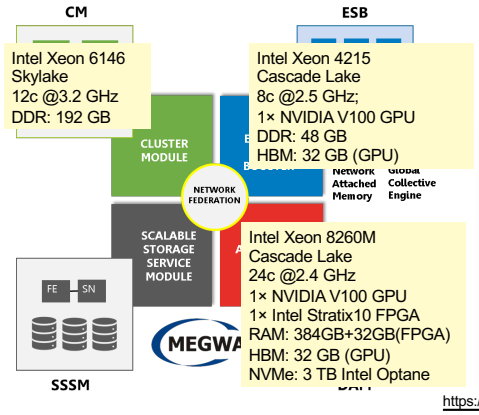
DEEP-EST Prototype
55 Cluster + 75 Booster + 16 Data Analytics
100 Gbit Extoll + InfiniBand + Eth
800 TFlop/s



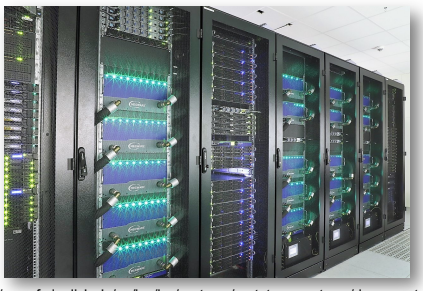
Source: FZJ

https://www.fz-juelich.de/en/ias/jsc/systems/prototype-systems/deep_system

DEEP-EST prototype



DEEP-EST Prototype
55 Cluster + 75 Booster + 16 Data Analytics
100 Gbit Extoll + InfiniBand + Eth
800 TFlop/s



Source: FZJ

https://www.fz-juelich.de/en/ias/jsc/systems/prototype-systems/deep_system

Lead Presentation: Estela Suarez (9 of 15)

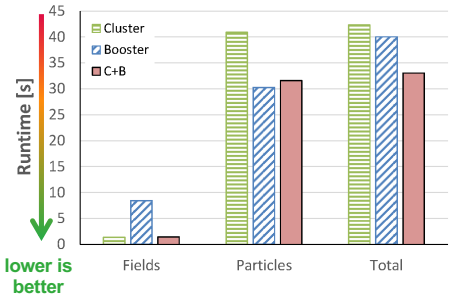
Example: xPic (Space Weather Simulation)

DEEP Projects

- **Field solver**: 6× faster on Cluster
- **Particle solver**: 1.35 × faster on Booster
- **Overall performance gain:**

1× node	**28% × gain** compared to Cluster alone **21% × gain** compared to Booster alone
8× nodes	**38% × gain** compared to Cluster only **34% × gain** compared to Booster only

– 3%-4% overhead per solver for C+B communication (point to point)



#cells per node	4096
#particles per cell	2048
Compilation flags	-openmp, -mavx (Cluster) -xMIC-AVX512 (Booster)

KATHOLIEKE UNIVERSITEIT LEUVEN xPic

A. Kreuzer et al. "*Application Performance on a Cluster-Booster System*", 2018 IEEE IPDPS Workshops (IPDPSW), Vancouver, Canada, p 69 - 78 (2018) [10.1109/IPDPSW.2018.00019]

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

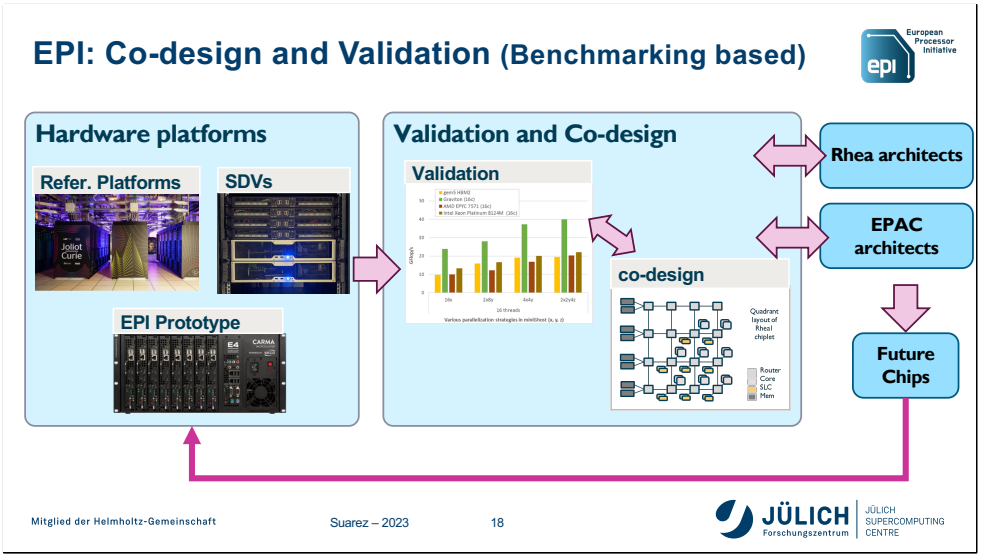
OUTLINE

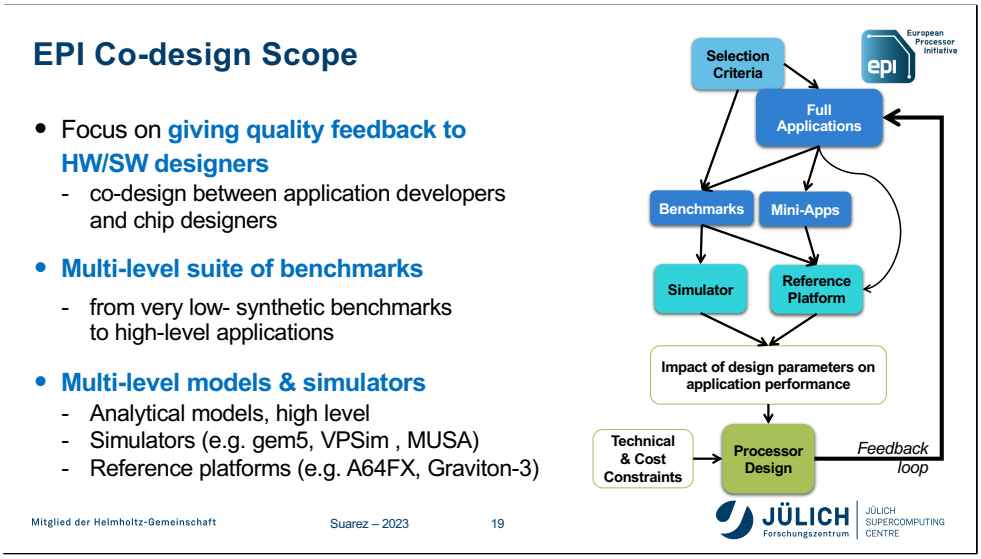
- **Co-design and Benchmarking**
- **Experiences**
  - DEEP Projects → System Level
  - EPI → Processor Level
- **Lessons Learned**
- **Summary**

Mitglied der Helmholtz-Gemeinschaft

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

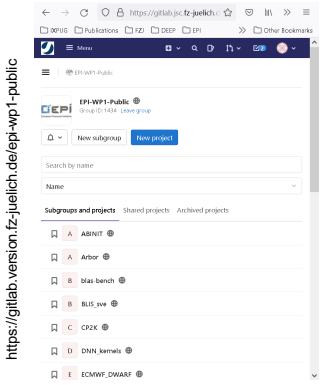
Lead Presentation: Estela Suarez (10 of 15)





Lead Presentation: Estela Suarez (11 of 15)

EPI Benchmark Suite



>40 codes, in the fields:

- Biophysics
- Biology/Medicine
- Earth Sciences/Climate
- HEP & Fusion
- Material Sciences
- CFD
- Hydrodynamics
- PDE
- Image / Media
- Automotive
- Cryptography
- HPDA
- Machine Learning
- Deep learning
- Cloud
- Data Base
- Reference benchmarks: (HPL, HPCG, Stream, DGEMM…)

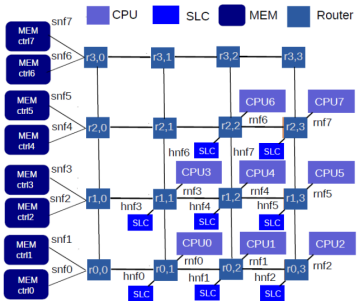
Mitglied der Helmholtz-Gemeinschaft Suarez – 2023 20

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

EPI CPU Simulation

- **Goal**
  - Understand impact of architectural parameters on application performance
- **Simulations of chip microarchitecture**
  - Detailed representation of chip elements (CPU, caches, network-on-chip, memory hierarchy)
  - Capability to change features
- **JSC contributions**
  - Develop models (gem5) that accurately represent the EPI Rhea platform (Arm-based CPU)
  - Analyse design trade-offs with benchmarks
  - Give feedback to chip developers

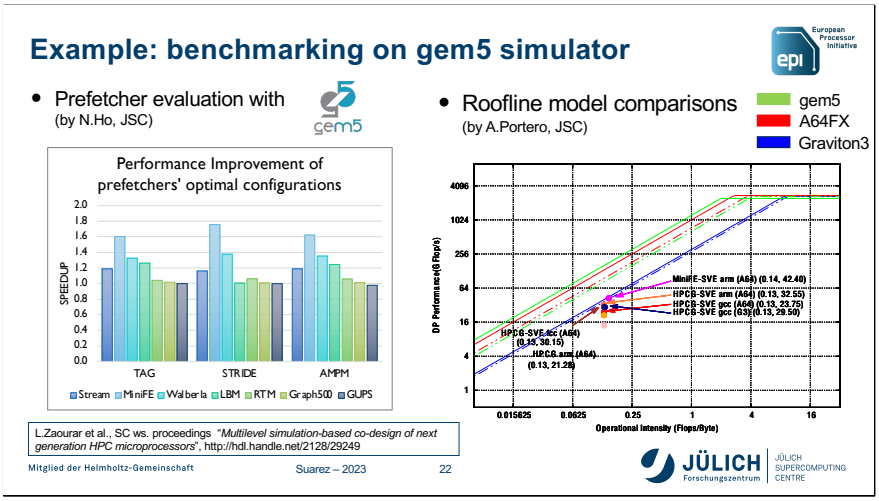
RHEA quadrant representation in: gem5



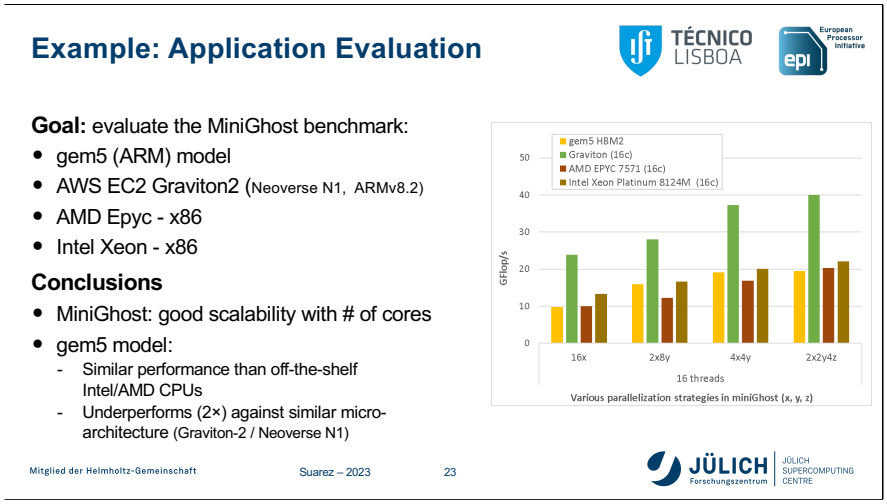
Mitglied der Helmholtz-Gemeinschaft Suarez – 2023 21

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

Lead Presentation: Estela Suarez (12 of 15)



- **OPIS:** Open Processor for Inception Systems
  - Based on tuned gem5 models (available for consortium)
  - Extracted information from the V1 reference board
  - "Best" approximation



The plots present the obtained performance (GFLOPS) in the tested platforms, by considering different parallelization strategies for the MiniGhost benchmark. The benchmark parallelization is tridimensional, allowing it to be parallelized in any direction with any number of threads. The parallelization strategy is indicated in the x-axis according to the nomenclature AxByCz, being A, B and C the amount of parallelization in each direction. As an example, 2x2y2z configuration defines a 2-by-2-by-2 parallelization grid, dividing the problem into 8 execution threads.

Lead Presentation: Estela Suarez (13 of 15)

OUTLINE

- **Co-design and Benchmarking**
- **Experiences**
  - DEEP Projects → System Level
  - EPI → Processor Level
- **Lessons Learned**
- **Summary**

Mitglied der Helmholtz-Gemeinschaft

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

Lessons Learned

Challenges:

- **Technical / Practical**
  - Hard to extract quantitative co-design input
    - Even harder for full workload mixes
  - Lack of clear baseline reference
    - codes, system-SW and -HW evolve simultaneously
  - Hard to pin-point & quantify co-design effect
    - Design decisions strongly cost-driven
    - Limited time-frame to apply co-design input
- **Strategic / Logistic / Organisational**
  - Application developers are rewarded for scientific runs (not for benchmarking or co-design)
  - Some details protected by commercial IP

Opportunities:

- **Technical / Practical**
  - Potential for optimisations in performance, energy efficiency, and scientific throughput
  - Tailor system to application portfolio
  - Enable own approaches to system architecture
  - Learn and understand each others language (from application to hardware design)
- **Strategic / Logistic / Organisational**
  - Real impact on product development roadmap
  - Real impact on application porting and performance improvements
  - Target open source simulation framework, with open benchmark suite incl. workload mixes

Mitglied der Helmholtz-Gemeinschaft Suarez – 2023 25

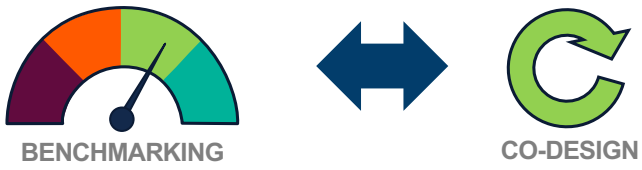
JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

Lead Presentation: Estela Suarez (14 of 15)

OUTLINE

- **Co-design and Benchmarking**
- **Experiences**
  - DEEP Projects → System Level
  - EPI → Processor Level
- **Lessons Learned**
- **Summary**

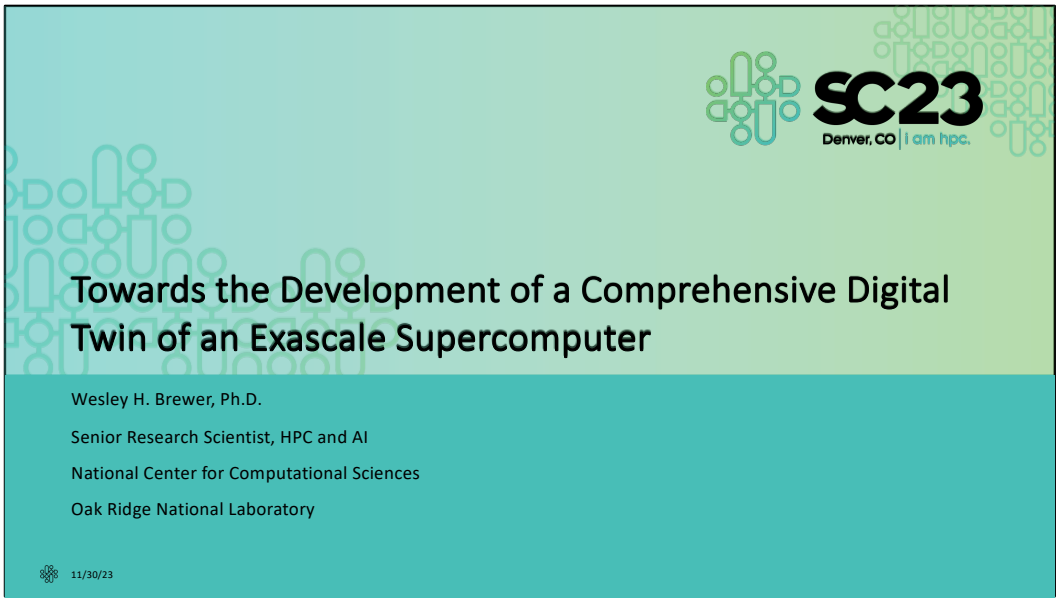
SUMMARY



- **Benchmarking is a critical tool for Co-Design** (both at system and component level)
- **Challenges** on extracting **quantitative requirements** from applications and pin-point the **impact of individual inputs** on the final design
- **Opportunities** for performance and energy efficiency **improvements**, if we invest on and **apply** a **systematic, data-driven, community effort**

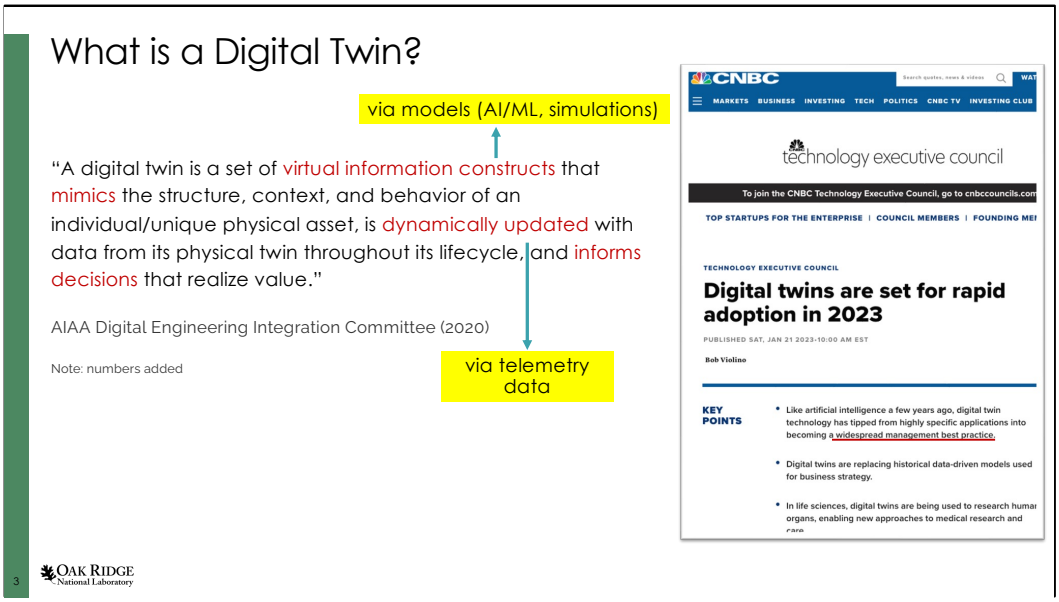
Lead Presentation: Estela Suarez (15 of 15)

Presentation: Wes Brewer (1 of 13)



23 and growing!

Presentation: Wes Brewer (2 of 13)



* this definition was the highest-voted definition from a crowd-sourced effort by the digital twin subcommittee, within the Digital Engineering Integration Committee

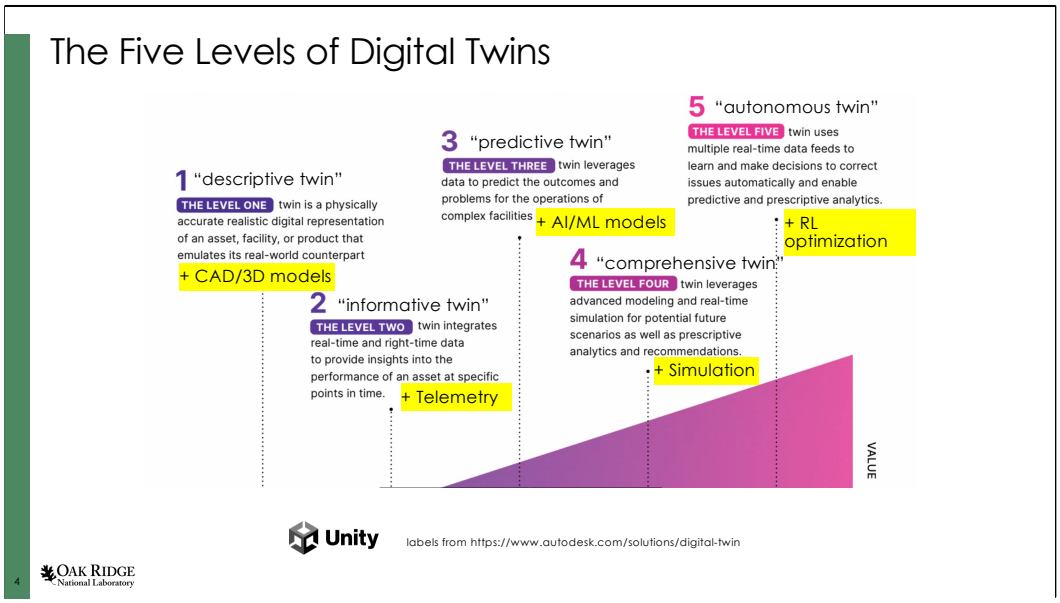
History:

- The most famous example of a digital twin is the one used to support the Apollo 13 mission.
- It consisted of 15 simulators that were used to train astronauts and mission controllers in every aspect of the mission, including multiple failure scenarios.
- NASA developed the first digital twin of apollo 13 – which was used to get three astronauts home safely.
- https://blogs.sw.siemens.com/simcenter/apollo-13-the-first-digital-twin/

"…all digital twins simulate the operation of a physical system and provide insight throughout the life of the facility. It consists of four basic elements:

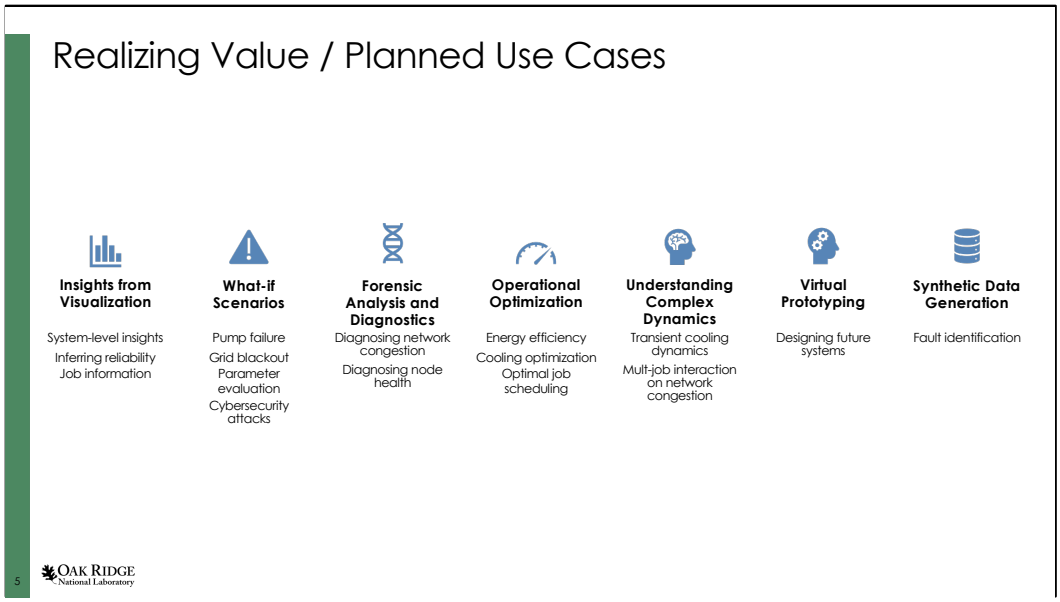
1. a mathematical model describing the system (simulation)

Presentation: Wes Brewer (3 of 13)

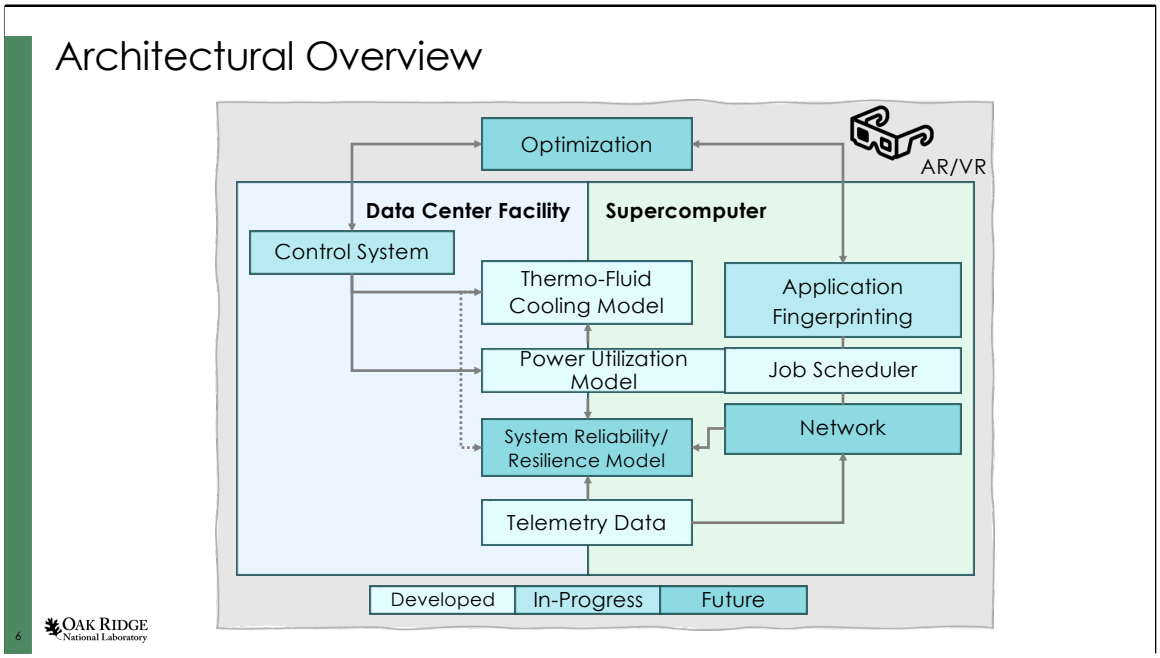


The higher the levels, the more value the DT realizes.

L1 = essentially just looking at the CAD data in a 3D VR environment
L2 = adding telemetry
L3 = adding some data-driven AI/ML models
L4 = adding simulation
L5 = adding reinforcement learning

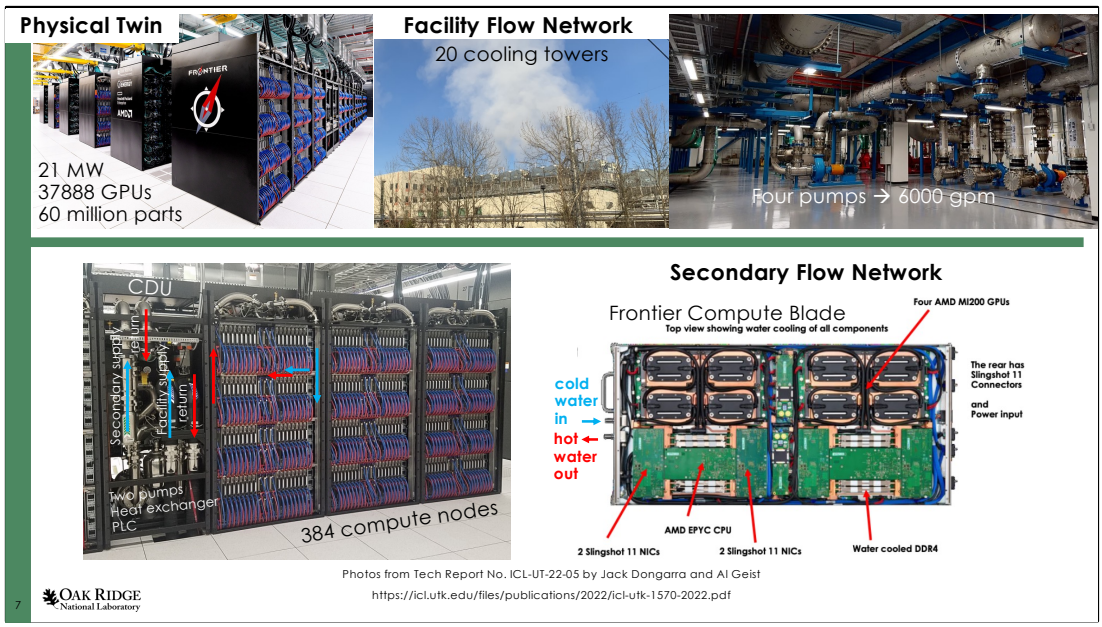


Presentation: Wes Brewer (4 of 13)



- DPC = Differentiable predictive control
- Cybersecurity – lives in a secure enclave and is also a tool to study security of the system
- Ultimate vision is to virtually submit a job, power ramps up, cooling reacts, etc.
- Collaborative tools
- Compute Resource Utilization → Slurm Simulator
- https://unity.com/solutions/digital-twins

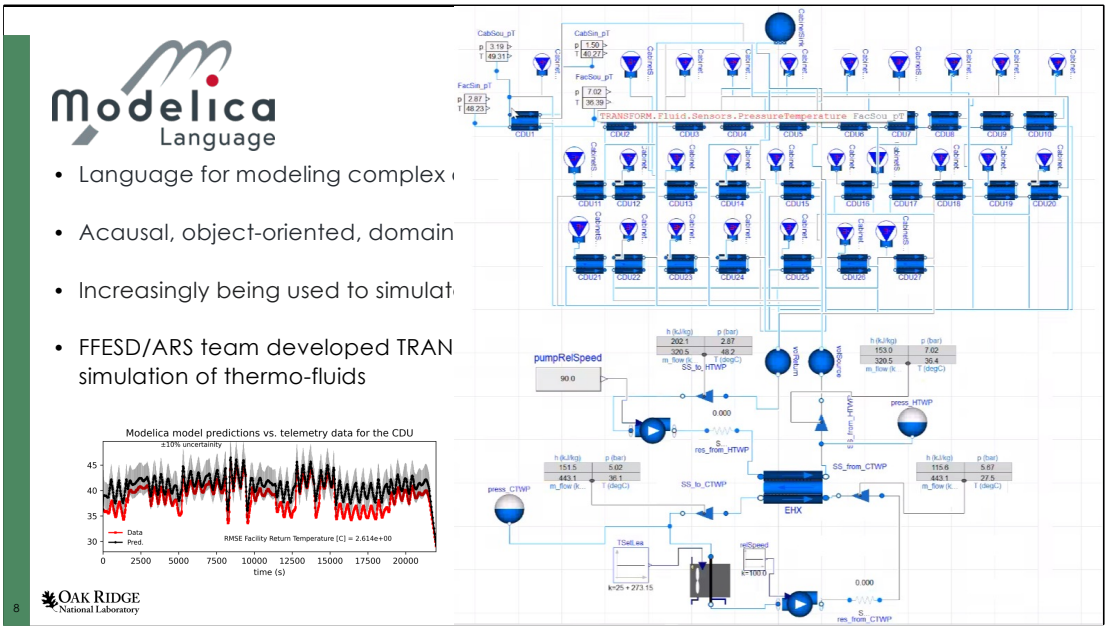
Presentation: Wes Brewer (5 of 13)



When we strip it down, there is still a lot of complexity to deal with.

- Frontier is the first exascale supercomputer, which was benchmarked at 1.1 exaflops and is currently the #1 system on the Top500 list and the second most energy efficient system on the Green500 list. Normal power range is about 21 MW, with power peaking at about 28 MW.
- It has about 38000 GPUs and contains 60 million parts.
- It is cooled with 20 cooling towers, has four main pumps on a facility flow network pumping water at 6000 gpm, which is used to cool a secondary flow network consisting of 25 cooling distribution units, which each cool 384 compute nodes via cold plates. So, you can see the flow loops here. Each blade has two compute nodes and the water flows in each node and sequentially cools eight GPUs, two CPUs, and the DDR4 memory.

Presentation: Wes Brewer (6 of 13)



FFESD = fusion and fission energy science directorate
TRANSFORM = Transient Simulation Framework of Reconfigurable Models
https://www.ornl.gov/publication/nuclear-thermal-propulsion-dynamic-modeling-modelica

* Wetter and Haugstetter (2006) – See https://w.wiki/6Kku

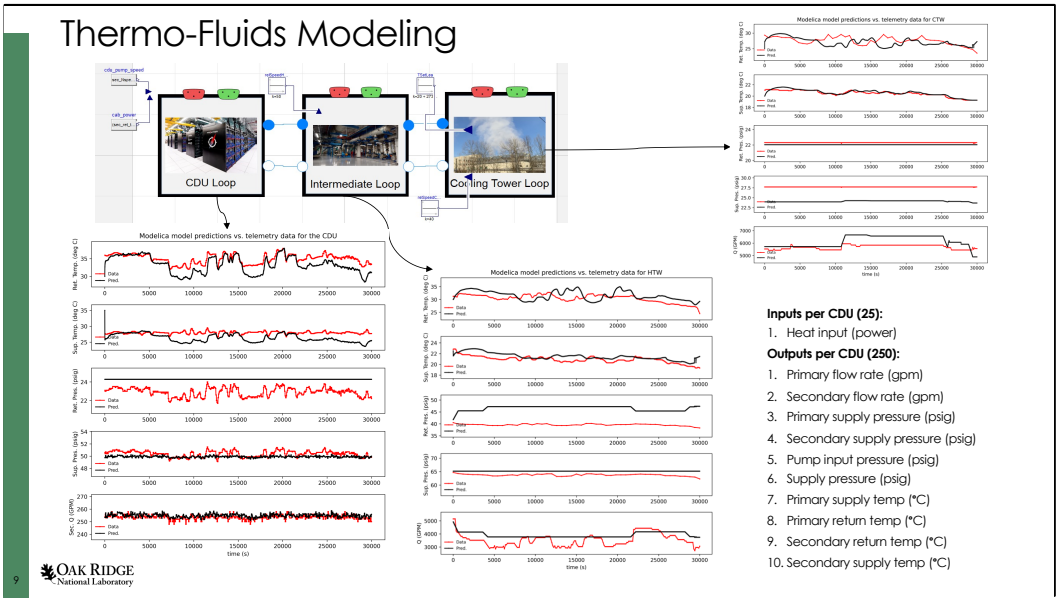
Fu et al. (2018) added Data Center Package to the Modelica Buildings Library

Used in FFESD and EESD directorates.

* domain neutral means not tied to any particular scientific or application domain

EESD = energy and environmental sciences directorate
https://www.ornl.gov/content/simulation-based-study-different-control-strategies-variable-speed-pump-distributed-ground

Presentation: Wes Brewer (7 of 13)



Take this CDU as example… important thing is transient behavior.

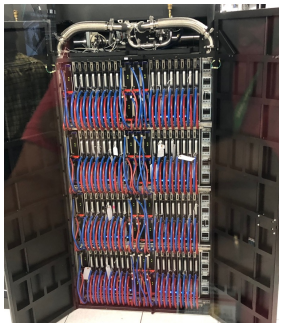
Make them understand why transient behavior is important factor – using Modelica we are able to model some of that.

Presentation: Wes Brewer (8 of 13)

The power supply system for Frontier

- 74 cabinets
- **35.5 MW maximum power coversion**
- 372 m² footprint
- 269 Tons (593 klbs)

- **480 kW power supply per cabinet**
- 32 x 15 kW power AFE rectifiers
- 128 SIVOC (**S**uper **I**ntermediate **VO**ltage **C**onverters)
- 64 nodes in the cabinet
- All liquid-cooled



- 32 rectifiers per cabinet
- 3 phase Δ 480 line-line AC input
- Power factor correction
- 380 V_{DC} output + AUX: 12 V_{DC}/1.8 A
- **Two rectifiers to be mounted side by side**

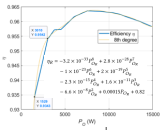
- 64 blades per cabinet
- 7.5 kW blades
- 2 nodes per blade
- **3.75 kW per node**

OAK RIDGE National Laboratory

10

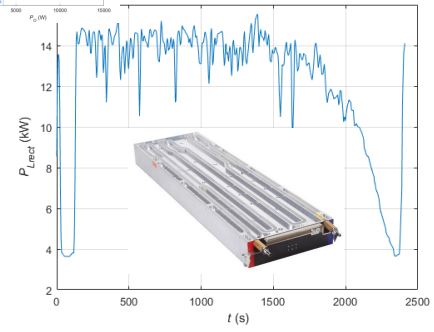
10

Rectifier and SIVOC loss models



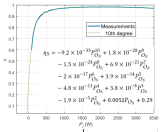
Rectifier

32 RRECTIFIER = ONE CABINET



Per cabinet: 14 kW

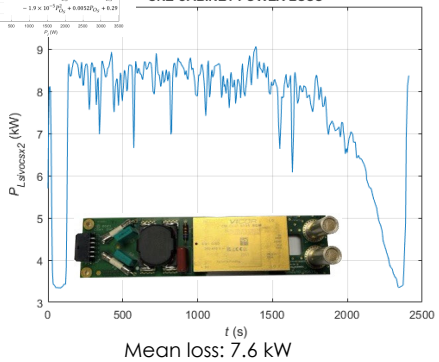
74 cabinets: 1MW



SIVOC

super-intermediate voltage converter

ONE CABINET POWER LOSS



Mean loss: 7.6 kW

Scaled up to 74 cabinets: 560 kW

OAK RIDGE National Laboratory

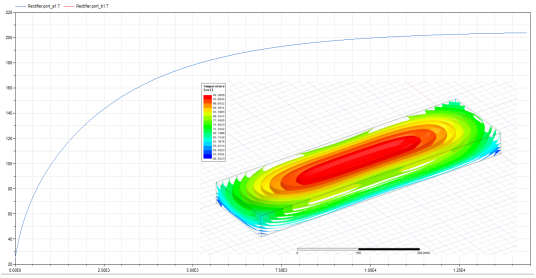
11

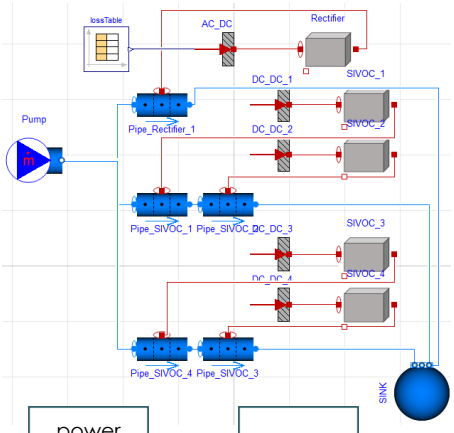
11

Presentation: Wes Brewer (9 of 13)

Power conversion blade thermal simulations

Dymola Transform package and Ansys IcePack





power load model → cooling model

OAK RIDGE National Laboratory

12

12

JOB SCHEDULER + POWER CONSUMPTION MODEL

- Schedule policy: FCFS or Backfill (SJF)
- Second-level tick count
- Each job defined by CPU/GPU utilization
- Infer power consumption from utilization
- At each time step add up power on all nodes
- Apply losses due to rectification and voltage conversion
- At job end release nodes and set power to idle
- Rack-level power are inputs to cooling model

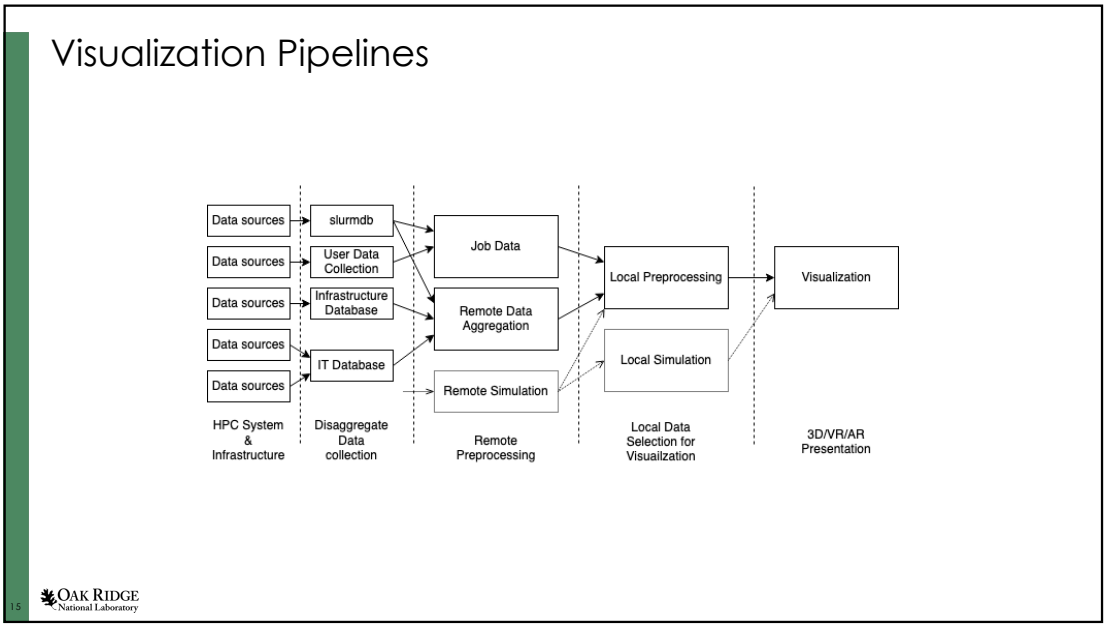
OAK RIDGE National Laboratory

13

13

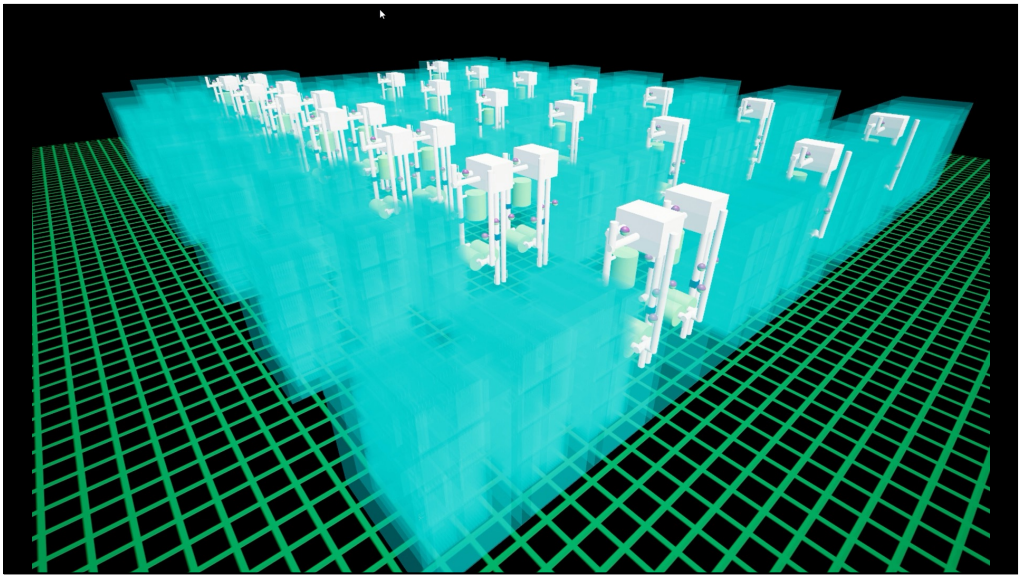
Presentation: Wes Brewer (10 of 13)

14

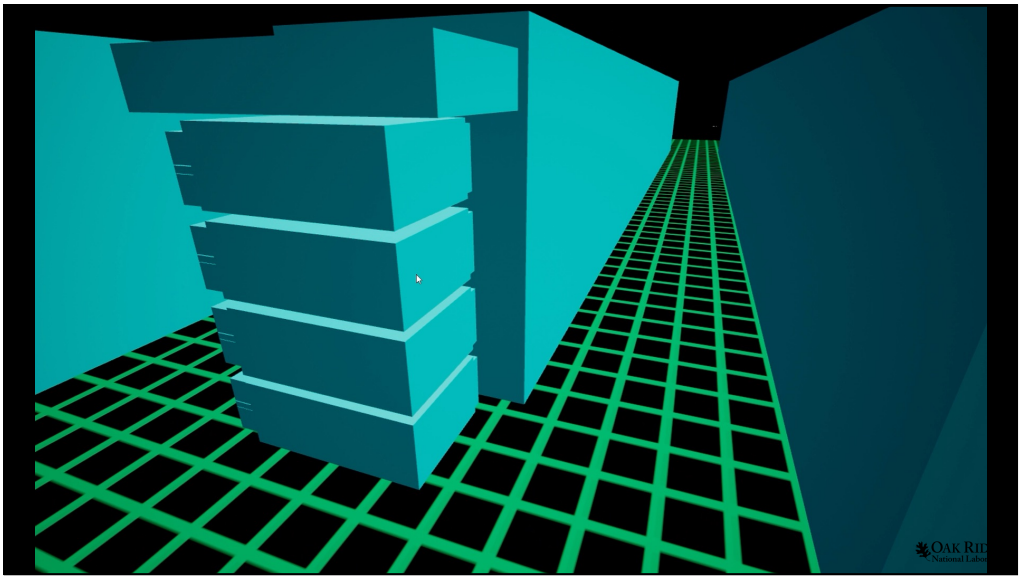


15

Presentation: Wes Brewer (11 of 13)



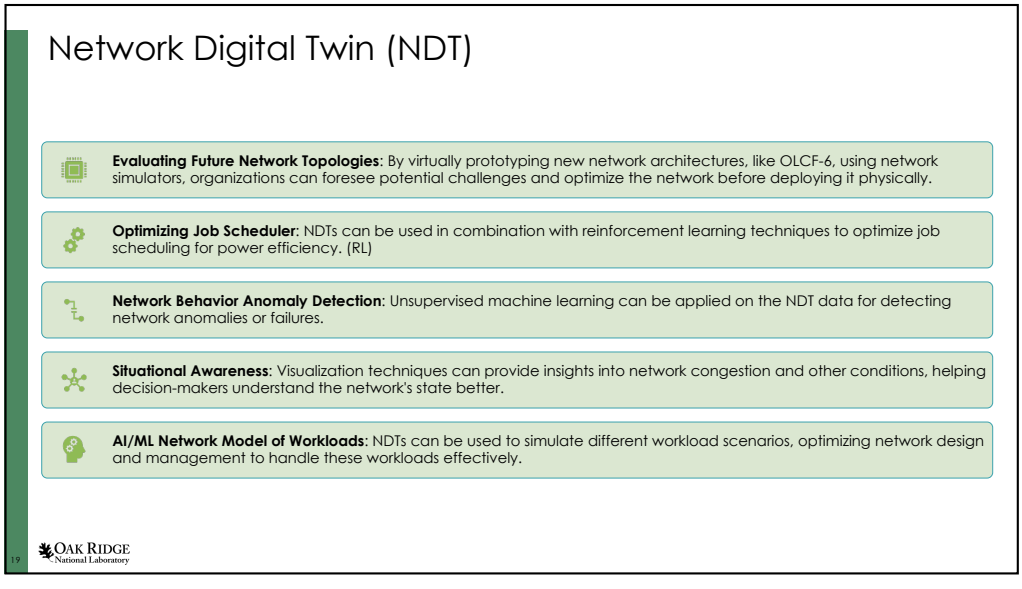
16



17

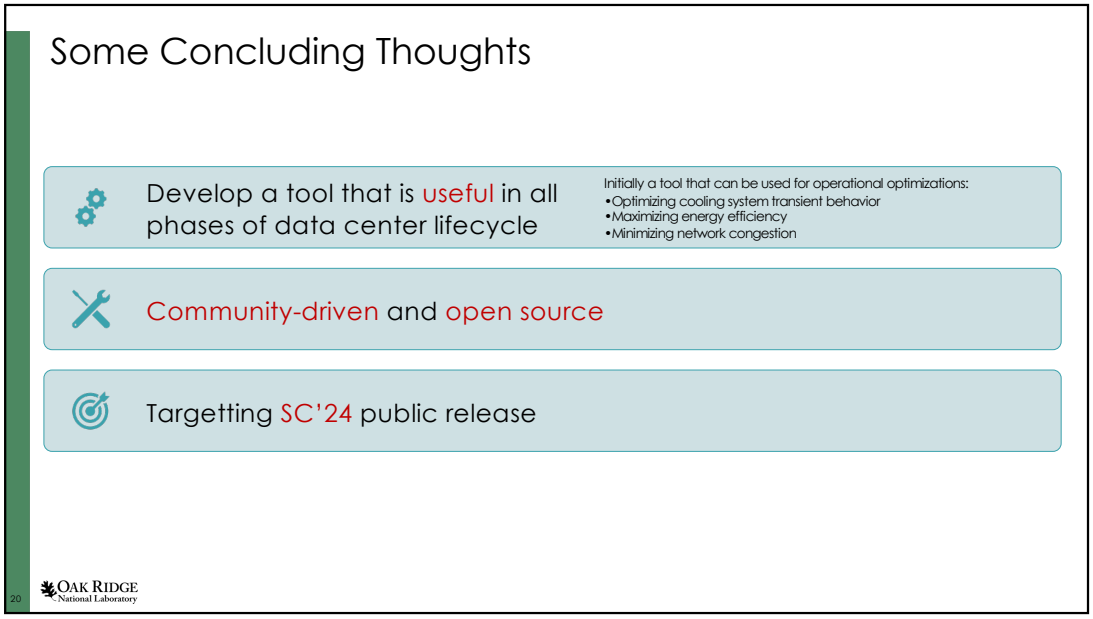
Presentation: Wes Brewer (12 of 13)

18

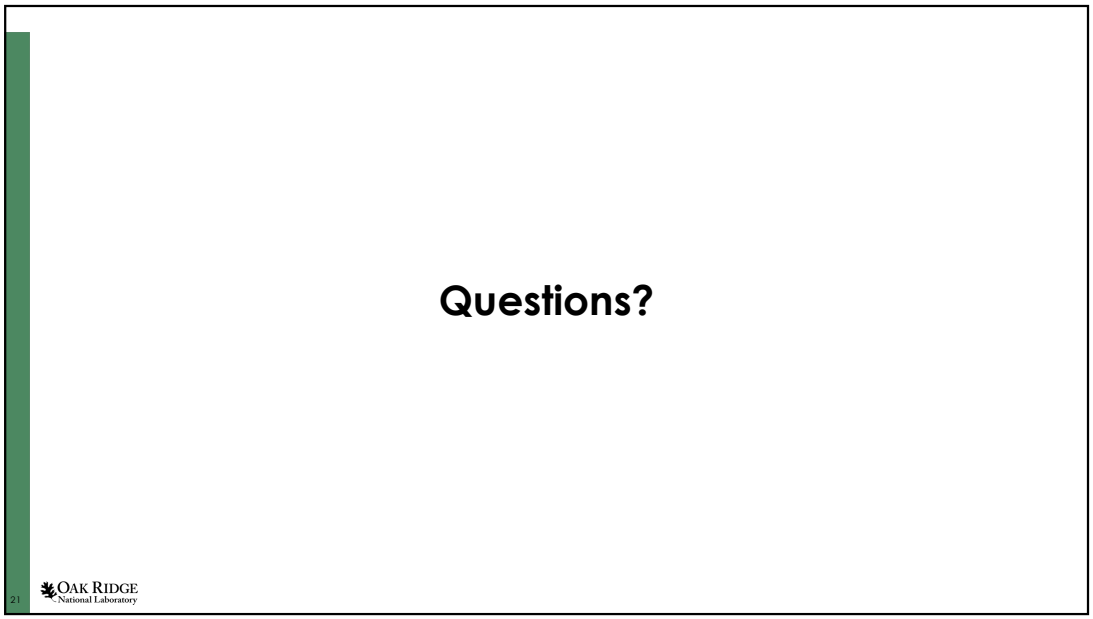


19

Presentation: Wes Brewer (13 of 13)

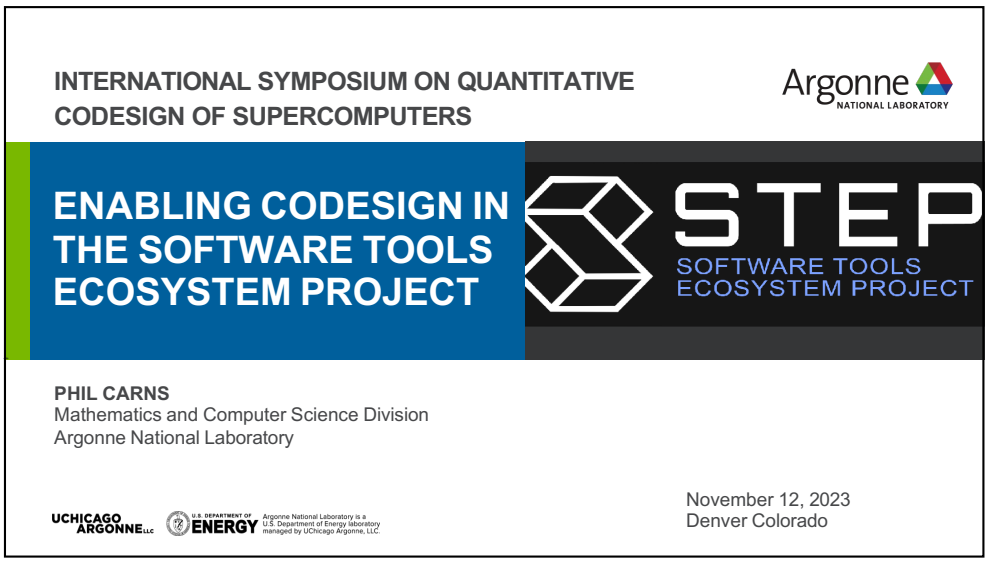


20

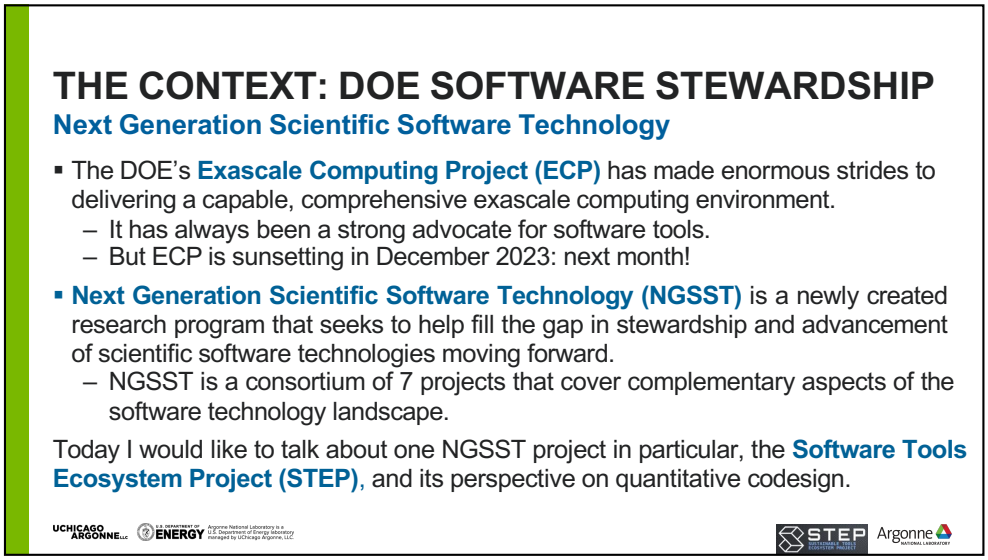


21

Presentation: Phil Carns (1 of 7)



1



2

Presentation: Phil Carns (2 of 7)

THE CONTEXT: DOE SOFTWARE STEWARDSHIP

Next Generation Scientific Software Technology

- The DOE's **Exascale Computing Project (ECP)** has made enormous strides to delivering a capable, comprehensive exascale computing environment.
  - It has always been a strong advocate for software tools.
  - But ECP is sunsetting in December 2023: next month!
- **Next Generation Scientific Software Technology (NGSST)** is a newly created research program that seeks to help fill the gap in stewardship and advancement of scientific software technologies moving forward.
  - NGSST is a consortium of 7 projects that cover complementary aspects of the software technology landscape.

Today I would like to talk about one NGSST project in particular, the **Software Tools Ecosystem Project (STEP)**, and its perspective on quantitative codesign.

UCHICAGO ARGONNE, LLC U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

STEP Argonne NATIONAL LABORATORY

2

PI: Terry Jones, ORNL
Deputy PI: Phil Carns, ANL

THE SOFTWARE TOOLS ECOSYSTEM PROJECT

Definition and objectives

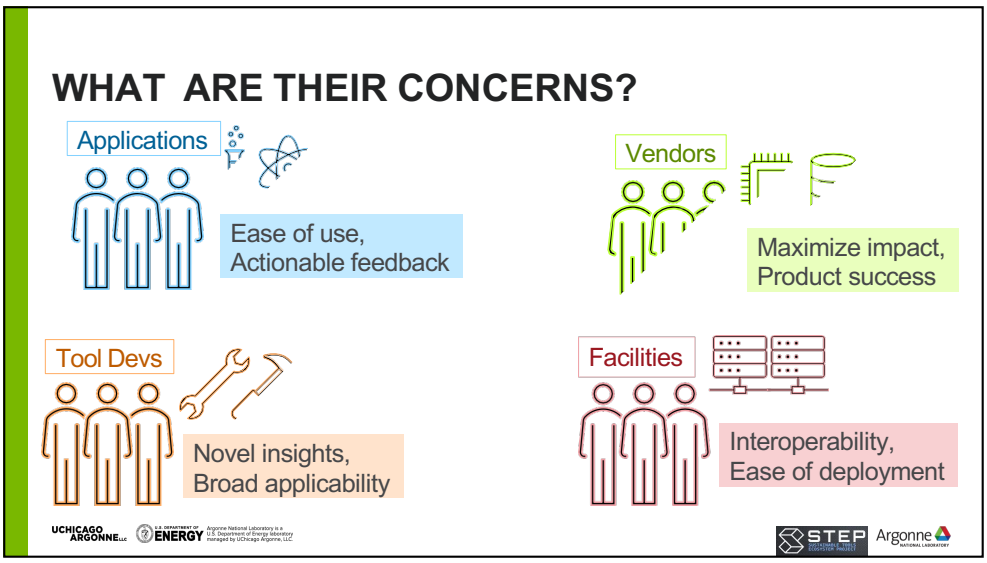
- Scope: Tools and supporting software for monitoring, analysis, and diagnosis of performance and behavior of codes on advanced computing systems.
  - Examples: application profilers, tracing tools, system monitors, etc.
- Ecosystem: The broader encompassing collection of stakeholders, platform dependencies, and interactions that influence those tools.
- STEP intends to not only support individual tools but to gather stakeholders to address cross-cutting issues.
- We kicked off this effort with a series of three town hall meetings over the summer of 2023.
- STEP was recently selected for CY 2024 funding as part of the NGSST program.

UCHICAGO ARGONNE, LLC U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

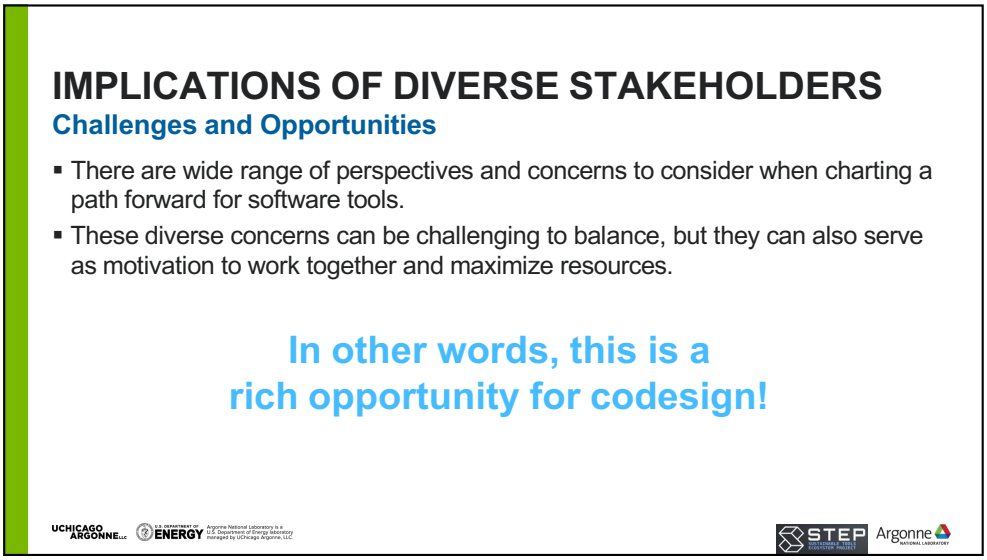
STEP Argonne NATIONAL LABORATORY

3

Presentation: Phil Carns (3 of 7)



4

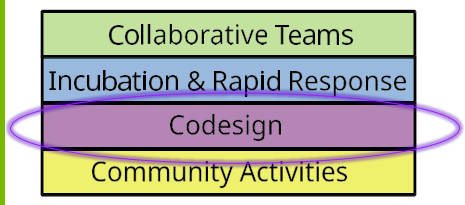


5

Presentation: Phil Carns (4 of 7)

ADDRESSING CODESIGN IN STEP

- STEP's year one activities are focused on stewarding an initial portfolio of critical tools with a limited budget.
- STEP's long-term aspirational vision is to expand into a holistic collection of ecosystem initiatives.



Codesign is one of the key elements that the community identified as critical to ensuring sustainability in the software tools ecosystem.

6

SOFTWARE TOOL CODESIGN OPPORTUNITIES

Example 1

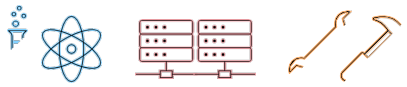
- **Vendors/developers: Codesign of product APIs**
  - Proactive coordination of product and tooling delivery
  - Improved documentation
  - Ability to correlate performance and debugging events
  - While still protecting vendor intellectual property!

7

Presentation: Phil Carns (5 of 7)

SOFTWARE TOOL CODESIGN OPPORTUNITIES

Example 2



- **Applications/facilities/developers: Codesign of mini-app suites**
  - Create a proving ground for software tool functionality
  - Demonstrate ability to quickly identify known problems
  - Demonstrate easy-to-understand, actionable outcomes
  - Ensure relevance to facility and application needs

UCHICAGO ARGONNE LLC U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

STEP Argonne NATIONAL LABORATORY

8

SOFTWARE TOOL CODESIGN OPPORTUNITIES

Example 3

- **Developers: Codesign of portability layers and shared abstractions**
  - Identify common capabilities across tools
  - Design portability layers that address those capabilities
  - Minimize duplicate development effort
  - Provide vectors that enable enhancement of multiple tools simultaneously

UCHICAGO ARGONNE LLC U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

STEP Argonne NATIONAL LABORATORY

9

Presentation: Phil Carns (6 of 7)

SOFTWARE TOOL CODESIGN OPPORTUNITIES

Example 4

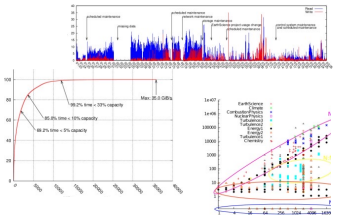
- **Facilities/developers: Codesign of data formats and UIs**
  - Improve interoperability
    - Chain multiple tools together into a diagnostic workflow
    - Enable integration of multiple data sources in a single coherent view
    - Leverage shared analysis frameworks
  - Improve platform integration
    - Simplify user experience when using multiple tools

UCHICAGO ARGONNE LLC U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

STEP Argonne NATIONAL LABORATORY

10

BUT IS IT QUANTITATIVE?



- The HPC community is accustomed to thinking of software tools as a *source* of quantitative data.
- But can we leverage quantitative data *about the tools themselves* to help drive codesign in their development?
- **Tool usage data**:
  - What tools are being used in practice?
  - What application domains do they address?
  - What fraction of a platform workload do they cover?
- **Longitudinal performance data**:
  - What problems can we diagnose in retrospect, and what can the tools community do to be more proactive in the future?

UCHICAGO ARGONNE LLC U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

STEP Argonne NATIONAL LABORATORY

11

Presentation: Phil Carns (7 of 7)

CONCLUDING REMARKS

- Quantitative codesign can have a multiplicative effect on productivity by:
  - Leveraging empirical evidence
  - Accounting for diverse perspectives
  - Helping to reinforce design decisions
  - Engaging multiple parties in sustainable solutions
- What the tools community can do to help support this approach:
  - Architect tools and infrastructure to accumulate tool usage statistics and longitudinal performance/behavioral data
  - Look for opportunities to engage stakeholders in design activities!

UCHICAGO ARGONNE LLC U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC. STEP Argonne NATIONAL LABORATORY

12

THANK YOU!

THIS WORK WAS SUPPORTED BY THE U.S. DEPARTMENT OF ENERGY, OFFICE OF SCIENCE, ADVANCED SCIENTIFIC COMPUTING RESEARCH, UNDER CONTRACT DE-AC02-06CH11357.

UCHICAGO ARGONNE LLC U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC. Argonne NATIONAL LABORATORY

13