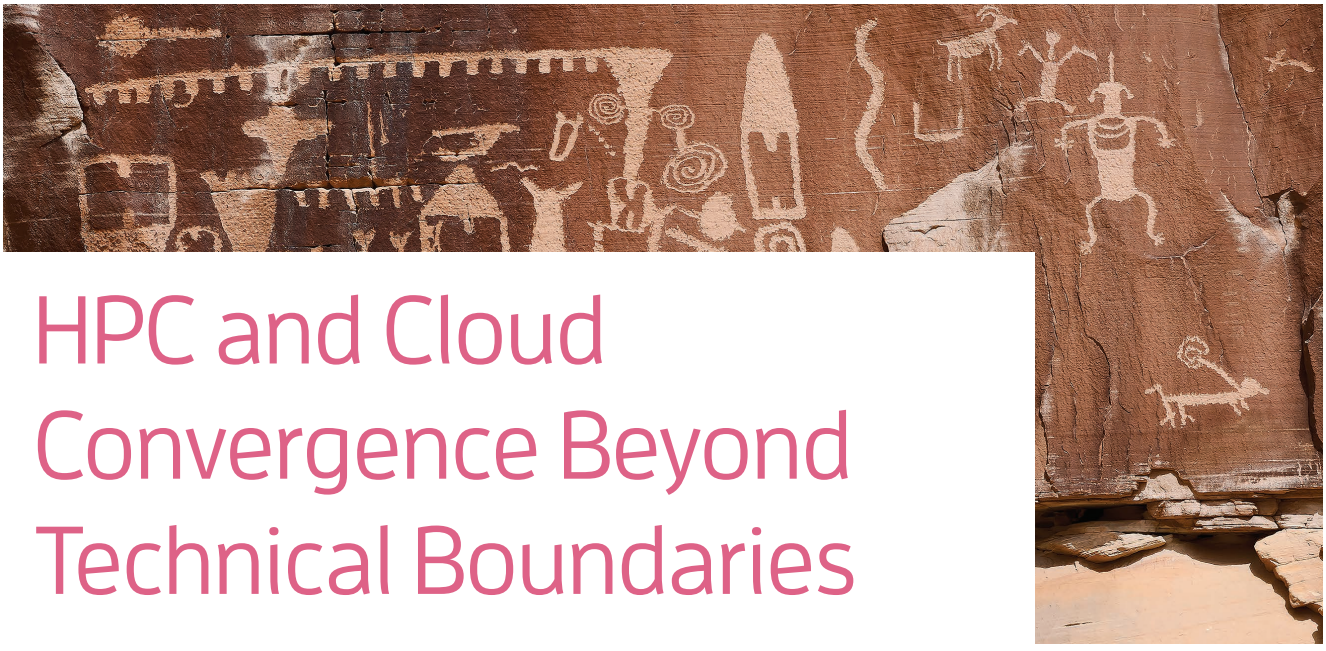

# HPC and Cloud Convergence Beyond Technical Boundaries

## Strategies for Economic Sustainability, Standardization, and Data Accessibility

**Michela Taufer**, University of Tennessee Knoxville
**Daniel Milroy** and **Todd Gamblin**, Lawrence Livermore National Laboratory
**Andrew Jones**, Microsoft
**Bill Magro**, Google LLC
**Heidi Poxon**, Amazon Web Services
**Seetharami Seelam**, IBM TJ Watson Research Center

*At the IEEE/ACM International Conference for High-Performance Computing, Networking, Storage, and Analysis (SC23), held in Denver, experts discussed the convergence of high-performance computing and cloud computing. They explored how this integration could address current scientific computing limitations, enhance computational capabilities, and foster global collaboration while focusing on economic, security, technical, and community challenges and opportunities.*

*Digital Object Identifier 10.1109/MC.2024.3387013*
*Date of current version: 3 June 2024*

At the IEEE/ACM International Conference for High Performance Computing, Networking, Storage, and Analysis (SC23), which took place in Denver, CO, USA, from 12 to 17 November 2023, a panel of experts from academia, national labs, and the technology sector convened to discuss the implications of the convergence of high-performance computing (HPC) and cloud computing for scientific computing. The discussion revolved around integrating HPC's computational performance with the elasticity of cloud technologies, focusing on the economic, security, technical, and community challenges and opportunities this convergence presents.

The panel discussed the need for economically viable solutions, enhanced security measures, and the development of an efficient software ecosystem to support this integration. The panel emphasized the importance of partnerships between HPC centers and cloud providers to combine the strengths of both domains and advance scientific research. The panel also highlighted data management issues, high-speed transfers, improved storage infrastructures, and the potential for cloud-based data repositories to foster collaboration. The panel envisioned a future where the integrated use of HPC and cloud computing not only addresses current limitations but also enables significant scientific breakthroughs driven by improved computational capabilities, global collaboration, and a flexible infrastructure tailored to the needs of data-intensive research.

In the evolving landscape of computing technologies, against a backdrop of rapid technological advancements and shifting economic dynamics, the necessity for a holistic convergence of HPC and cloud computing (integrating the performance-driven computing of HPC systems with the resource-rich environment of the cloud) is becoming increasingly urgent. This roundtable delves into the complexities and opportunities presented by the convergence of these two powerful computing paradigms.

The panel argued for the importance of adopting a collaborative strategy incorporating the two domains of HPC and cloud computing. By capitalizing on the strengths of each domain (for example, HPC's efficient resource management and low latency and cloud computing's scalability, elasticity, and automation), a vision for a more robust and versatile hybrid cloud built on HPC and cloud convergence begins to take shape. Initiatives like the Flux project[1,2] at the Lawrence Livermore National Laboratory exemplify innovative strategies that can foster seamless interoperability between HPC and cloud environments.

HPC and cloud convergence is not just about combining technologies. It is about creating a dynamic platform that can integrate multiple resources and services in complex workflows to propel scientific research and computing capabilities. However, achieving this vision requires developing accurate economic models to ensure long-term economic viability, addressing security concerns, and developing an adaptable and efficient software stack to bridge these two worlds.

At SC23, the panel "HPC and Cloud Converged Computing: Merging Infrastructures and Communities" discussed the rapid growth of integrated HPC and cloud technologies. The discussion highlighted the immense potential for this convergence to advance scientific discovery. The panel covered the multiple challenges along this path, including the sustainability of economic models, the intricacies of ensuring robust security measures, and the imperative need for a versatile and efficient software stack. The dialog underscored the critical role of strategic partnerships between HPC entities and cloud service providers, suggesting that such collaborations are crucial to leveraging these technologies' component strengths for the greater good of scientific progress. In navigating these complexities, the panelists at SC23 emphasized the necessity of innovative approaches to improve cost management and ensure economic sustainability, achieve standardization and interoperability, and guarantee data access and use. See "Main Concepts" for an explanation of the main concepts essential to this discussion.

> It is crucial for the HPC community to cultivate collaborative research and development partnerships with the cloud computing sector.

## RESPONSES FROM THE PANEL

**QUESTION 1:** *What critical components are needed for true HPC and cloud convergence?*

**DANIEL MILROY:** Cloud computing is expected to dominate the computing industry in the near future. It is a key driver of innovation, potentially transforming scientific workflows and HPC. To harness these benefits, it is crucial for the HPC community to cultivate

collaborative research and development partnerships with the cloud computing sector.

**HEIDI POXON:** Adding to Dan's point, cloud services' dynamic cost-effectiveness can significantly reduce costs for projects with fluctuating computational needs. Transparent billing practices and tools for managing expenses are crucial to leveraging cloud elasticity effectively.

**ANDREW JONES:** It is also essential to be clear about what we mean by "HPC" and "cloud." For example, HPC can be done in traditional on-premises (on-prem) environments but can also be done in cloud environments. Thus, HPC and the cloud are not really competitors; instead, they are complementary technologies. We should optimize how we unlock economic or scientific value by focusing on the capabilities needed to solve scientific problems rather than starting with labels. This approach requires us to reevaluate the economic models underpinning HPC and cloud computing, aiming for a model that benefits the specific context and the broader scientific community.

**SEETHARAMI SEELAM:** Moving to standardization and interoperability as a second component of this convergence, the future of HPC and cloud computing lies in a unified software environment. Open standards and application programming interfaces (APIs) ensure system compatibility and facilitate seamless workload migration between on-prem HPC systems and cloud platforms.

**BILL MAGRO:** Indeed, a community-driven approach to standardization, where contributions from all stakeholders are valued, is essential. This collaborative effort can make interoperability a reality, enhancing the flexibility and efficiency of computational research.

## MAIN CONCEPTS

The main concepts essential for understanding this discussion are as follows:

- *HPC and cloud convergence:* HPC and cloud convergence combines HPC with cloud computing infrastructure into a seamless best-of-both-worlds environment. Convergence leverages the strengths of both HPC (for example, efficient resource management and low latency) and cloud computing [for example, elasticity, automation, and application programming interface (API)-based management] to enhance scientific and research computing capabilities in hybrid platforms.
- *Hybrid HPC-cloud platforms:* Hybrid HPC-cloud platforms incorporate the strengths of HPC and cloud computing into multifaceted computing models to advance scientific research.
- *Portability and sustainability:* Portability and sustainability refer to the ability to run computational workflows across different computing environments without significant modifications and to do so with long-term resource use stability.
- *Trustworthiness and reproducibility:* Trustworthiness and reproducibility focus on the reliability of computational results, especially in a scientific context. They emphasize replicating experiments and trusting experimental outcomes, which are crucial in scientific research using cloud and HPC infrastructures.
- *Usability and efficiency:* Usability and efficiency in workflows address the need for computational workflows to be user-friendly and efficient, regardless of whether they run on HPC systems or in a cloud environment. They include the development of interfaces and software layers that facilitate the execution of complex workflows.
- *Cloud adaptability:* Cloud adaptability for HPC supports applying cloud computing approaches and technologies (for example, Kubernetes and containerization) to traditional HPC environments. It involves integrating cloud-like capabilities into HPC systems to enhance flexibility and accessibility.
- *Economic viability:* Economic viability in HPC and cloud use addresses the cost and value implications of using cloud services for HPC tasks. It involves considering cloud services' long-term sustainability, affordability, and benefits for large-scale scientific computing.
- *Multitenancy and security:* Multitenancy and security in HPC and the cloud refer to the ability of multiple users or organizations to securely share computing resources, whether in a cloud environment or an HPC system, without compromising data integrity or performance.
- *Data management and accessibility:* Data management and accessibility involve handling large datasets efficiently in both HPC and cloud environments and ensuring that data are easily accessible and transferable between these systems. This is particularly important for scientific collaborations and open science initiatives.

**MICHELA TAUFER:** Let's not forget the elephant in the room: the data, their movement, and access. For robust data management and accessibility, we need frameworks that prioritize accessibility, portability, and security. Community-agreed standards will ensure that data generated on HPC and cloud platforms can be easily shared and reused, facilitating cross-domain scientific discovery.

**TODD GAMBLIN:** Building on Michela's point, making it possible for developers to use advanced data analytics and artificial intelligence (AI) tools in conjunction with traditional HPC is critical. We need a model where we can easily spin up infrastructure as a service and other services alongside HPC jobs. Currently, researchers struggle to deploy these tools in HPC environments.

**QUESTION 2:** *As we navigate the complexities of merging HPC and cloud computing, cost management and economic viability emerge as significant barriers. Can we elaborate on how we can overcome such barriers?*

**GAMBLIN:** Developing cost-effective hybrid models is critical for medium- to large-scale HPC centers. On-prem HPC centers are much more cost-effective, such as flop-for-flop versus cloud. But increasingly, users want to take advantage of the many valuable services that clouds offer. Developing hybrid scheduling and deployment models is essential to leverage both on-prem HPC resources and cloud services. This approach would allow users to deploy certain complex distributed applications in the cloud while still leveraging their on-prem HPC resources for compute-intensive workloads. The cloud is also suitable for sporadic surges in demand or, for smaller centers, for large-scale runs that exceed local capabilities. Systems like Flux that enable heterogeneous workloads and hybrid scheduling are promising for allowing this model to evolve, and we're hoping that they continue to evolve to let users exploit the advantages of both on-prem and cloud HPC.

**MILROY:** The ability to run applications efficiently in as many environments as possible will become central to cost management. It is crucial to enable HPC-like applications to run on cloud-like systems and instantiate and manage cloud software on HPC infrastructure. Workflow portability will become a key cost control component as HPC is increasingly run on public clouds. Facilitating workflow portability between clouds and HPC clusters with low-performance overheads from the software environment is likewise essential. Developing effective cost models will help users understand where, when, and how to run their applications efficiently. It is necessary to build ecosystems that support hybrid execution models for on-prem HPC resources and public cloud services. The ecosystems must allow workflows to deploy and run within public clouds in much the same way as they run on-prem HPC systems.

**POXON:** Another key in collaborating between cloud providers and HPC centers is to forge transparent predictable pricing models. This transparency is crucial for users, especially in the research sector, to manage their budgets effectively. We advocate for a scenario where costs and performance are clear from the outset, enabling researchers to plan their expenses without fear of unexpected bills. It is about making cloud computing a reliable tool for the scientific community.

**SEELAM:** I echo Heidi. Transparent pricing strategies are also critical; consistent performance and pricing models enable users to manage their financial planning effectively. Cloud providers and HPC centers should work collaboratively to develop transparent, predictable, straightforward, and reliable pricing models. Such models are intended to give users well-defined cost expectations, thereby streamlining the budgeting process. To achieve this, cloud providers must commit to full disclosure regarding their pricing schemes. They should also consider creating customized pricing options well suited to the fiscal realities of research and educational institutions. By doing so, they will provide a service that offers predictable costs and performance, particularly for cloud-based AI systems, ensuring that users can allocate their budgets with greater security and predictability.

**TAUFER:** At the core of any collaborative success, there is the need for the community to focus on breaking down the financial barriers that limit access to cloud computing for research and academic purposes. By securing financial support through subsidies or grants and establishing community clouds tailored to the needs of the scientific community, we can significantly lower the cost of accessing high-powered computational resources. Under shared investments and resources, these community clouds reduce individual costs, fostering an environment of collaboration and shared discovery. It is about creating a sustainable model that supports innovation across the board.

**MAGRO:** Michela is correct; financial incentives and community clouds provide a comprehensive strategy to empower researchers. To promote the advancement of research, government entities and funding organizations must provide financial support through subsidies, grants, or specialized pricing plans for cloud computing. These incentives are tailored to meet research and academic institutions' unique needs, facilitating more equitable access to cloud technology. Establishing community clouds offers a collective solution in tandem with these financial measures. These clouds can cater specifically to the needs of the research and academic community. By pooling resources and sharing investments, such community clouds can deliver specialized computing resources at a significantly reduced

cost. This arrangement amplifies the benefits of cloud computing for individual institutions and fosters a collaborative environment that can lead to greater scientific discoveries and innovation. Reducing cost barriers and creating collaborative platforms ensure that researchers can access the necessary computational tools, accelerating research and development without the typical financial constraints.

**QUESTION 3:** *Dealing with the complexities of HPC and cloud technologies presents significant challenges. How can we address standardization and interoperability issues to ensure smooth workflow transitions from HPC and cloud computing to hybrid solutions?*

**TAUFER:** To streamline the integration of HPC and cloud environments, the development of standardized APIs and uniform software stacks is critical. Such standardization can significantly enhance interoperability, making it easier for researchers to utilize a mix of HPC and cloud resources. Establishing these common APIs and cohesive software frameworks requires a sustained effort from stakeholders across the spectrum: cloud service providers, HPC administrators, and research community members. The goal should be to reach a consensus on a set of standards that can be widely adopted, ensuring compatibility and ease of use across various platforms. Researchers can seamlessly migrate their work between HPC systems and cloud services by having universally recognized interfaces and software, focusing on scientific inquiries rather than technical integration challenges.

**JONES:** I agree with Michela that standardization, interoperability, and portability help deliver value for everyone. A significant barrier has been the lack of standardization and interoperability between HPC and cloud environments. With a wide range of computing platforms, even just on-prem systems, each with unique software stacks, operating models, and interfaces, creating a seamless integration between HPC systems and cloud services is challenging. This lack of standardization complicates the transfer of workflows and data between systems, leading to inefficiencies and limiting the potential of collaborative research. Researchers often face difficulties adapting their workflows to different environments, which can hinder the adoption of cloud computing in HPC-intensive research fields. The problem is further compounded by the rapid evolution of technologies, making it hard for standardization efforts to keep pace. However, we must be careful not to standardize to the lowest common denominator and preserve the ability to differentiate when that adds value.

**MILROY:** Containerization technologies can address portability challenges by packaging applications with their dependencies, ensuring consistent functionality across diverse computing environments. Orchestration takes this a step further by offering declarative lifecycle management for applications and workflows. Container orchestration frameworks natively support cloud-like capabilities such as elasticity, resiliency, and automation. The vast and rapidly increasing popularity of container orchestration will make it a means of portability between HPC systems and cloud platforms. Integrating containerization, virtualization, and orchestration into HPC is crucial for overcoming interoperability issues. Bringing HPC techniques for sophisticated scheduling and latency reduction into cloud environments will increase workflow performance portability. Such technologies simplify complex workflow management and boost computational research efficiency by minimizing the technical complexities of running applications across different infrastructures.

**POXON:** One of the values of HPC is our diverse set of software stacks, interfaces, and models. Our ability to work within diverse environments is part of a cycle that fosters engineering research and pushes technology innovation that, in turn, enables new scientific discoveries. That said, there needs to be a balance between user productivity, interoperability, innovation, and diverse environments. Dan provided a great example of how to enable interoperability through workload containerization. In addition to efforts like this, it is also essential to use community-led open source collaborations and standardization initiatives to manage unnecessary divergence. Encouraging open source partnerships where researchers, cloud providers, and HPC centers contribute can foster standardization where needed. These partnerships bring a broader set of experiences to encourage practices and interfaces that reduce the complexities of moving workflows between HPC centers and the cloud.

**SEELAM:** I propose a targeted strategy to overcome standardization and interoperability issues between HPC and cloud technologies. Focusing on innovations like IBM Vela,[3] we see a fusion of HPC's power with cloud flexibility and scalability. By adopting software-defined networking and storage, we facilitate a seamless integration of HPC and the cloud, ensuring that our hybrid systems achieve both supercomputing performance and the cloud's agility. The move toward a hybrid cloud is essential for better resource access. I encourage the HPC community to embrace the cloud's API-centric and software-defined models, aligning our computing resources more closely with research demands. Kubernetes and containers play a crucial role in creating a cohesive platform across various cloud services, enhancing both interoperability and flexibility. Additionally, I emphasize the need for educating both administrators and users on the advantages of hybrid environments to foster more adaptable and efficient practices.

**MAGRO:** In addressing the challenges of blending HPC and cloud technologies, I actually believe in the concept of confluence over convergence. You see, HPC is fundamental to our activities, whether it is scientific research or engineering work. On the other hand, the cloud provides a venue for these activities, offering a different set of capabilities and resources. So, it is not about merging HPC and the cloud into one but understanding how they can coexist and complement each other to enhance our computing strategies. Take, for instance, the rapid advancements in AI, which demand an exponential increase in computational power. This need pushes both the HPC and cloud realms to evolve swiftly, with cloud infrastructures like Google's Tensor Processing Unit (TPU) pods[4] showcasing that supercomputing capabilities are feasible in cloud environments without a distinction from traditional HPC setups. The real question then becomes not whether to choose HPC or the cloud but how to utilize both environments optimally. It is about determining where a specific task can be executed most efficiently and effectively. This decision-making process is crucial as we navigate the balance between the capacity and capability of our computing systems, especially given the cloud's vast resource pool.

Furthermore, when discussing the cost-effectiveness of cloud solutions for HPC workloads, it's essential to consider the value delivered against the price paid. Cloud services are designed for reliability and scalability, which might come at a higher cost than running dedicated HPC systems. However, the cloud's evolving nature, especially in its alignment with HPC usage patterns, promises a future where affordability and accessibility for HPC in the cloud are significantly improved.

**GAMBLIN:** At Livermore, we do not want to be tied to any one environment for HPC workloads as our applications need to be able to run our mission simulation codes easily on current and future platforms. At the same time, we want to take advantage of unique capabilities, like AI services and automation, offered by clouds. For example, we might want to host codes as a service for customers across the National Nuclear Security Administration, or we may want to leverage digital twins to couple simulations more closely to production facilities. We also recognize that developers' first exposure to HPC will increasingly likely be in a cloud environment. We want our software to be relevant in all of these scenarios. The first step to ensuring that our software can run just as easily on-prem as it can in the cloud is ensuring that our software deployment tools can work across the two worlds. To address this, we are collaborating with major industry players like AWS, Google, and Intel to build and test our software to work seamlessly in both settings.

We have built a large-scale cyberinfrastructures (CI) system with the Spack package manager,[5] and we are striving to ensure that it works just as well for on-prem HPC deployments as it does for cloud HPC deployments, especially containerized ones. Along with this, we've started a continuous benchmarking effort that aims to regularly and automatically ensure that the performance of binaries built with Spack is up to par. In the case of CI, we've benefited from some of the security advantages of cloud environments over traditional HPC. Strict virtualization and multitenancy models allow us to test code from contributors on GitHub, something we cannot do as easily inside the perimeter of an HPC center. Ultimately, we would like to see both sides converge in this area, with on-prem HPC developing some of the virtualization capabilities of the cloud-based software, hardware, and network performance for HPC.

**QUESTION 4:** *Considering the challenges of HPC and cloud convergence, from security and storage capacity to the seamless transfer and accessibility of large datasets, how can we develop a standardized approach to data management that ensures secure, efficient, and universally accessible data on the hybrid cloud and converged endearments?*

**MILROY:** Bringing HPC and cloud computing together leads to challenges linked to data management and accessibility. In scientific research, especially where large datasets are common, storing, accessing, and processing data efficiently is crucial. The challenges become more pronounced when integrating HPC and cloud environments as each system may have different data storage, management capabilities, and protocols. Furthermore, data storage costs, input–output operations, and transfers must be analyzed and carefully considered.

**GAMBLIN:** I think there are two advantages that HPC people should consider when it comes to cloud datasets. One is accessibility, as Dan mentioned. Throwing a dataset in object storage and exposing it for others' access is straightforward. The ability of the cloud to scale and cache (with, for example, content delivery networks) these datasets makes them easily available to anyone around the world. The second big advantage is that clouds can seamlessly couple (possibly tiered) services with data. You can run an indexing or query service in front of a dataset in an object store, depending on the needs of your application. You can run a query service to allow anyone to interface the data easily, and the queries are distributed over lots of computing resources. Neither of these things are easy in HPC centers because they are not (yet?) built to support persistent services, and storage options tend to be limited to large-scale parallel file systems. Part of this is that HPC centers just are not used to exposing things as services; it has traditionally been discouraged even to have a service listening on a port open at an HPC center. This limits the services available to those provisioned by system administrators, while in the cloud, users are empowered to deploy

the services that meet the needs of their application.

For convergence, we need a hybrid way to either leverage on-prem datasets with cloud services or make HPC centers more cloudy, allowing users to provision virtual machines and persistent storage services. This will require redesigning HPC centers and teaching the user base how to use these new capabilities.

**TAUFER:** I agree that the solution extends beyond just infrastructure enhancements. Unified data management systems, collaborative frameworks, and comprehensive training are essential for enhancing data accessibility and management across both HPC and cloud platforms. Unified data management systems should provide a consistent and user-friendly data storage, retrieval, and sharing interface, regardless of the underlying computing environment. By ensuring seamless integration with different platforms, these systems address the challenges of managing the vast amounts of data generated in scientific research, enabling efficient data handling and easy accessibility. A critical aspect of these systems is their ability to maintain functionality across various environments, simplifying the process for users to access and manage data.

The implementation of collaborative frameworks is also crucial in this context. These frameworks are designed to standardize data-sharing practices, including establishing common data formats, metadata standards, and sharing protocols. Such standardization facilitates easier access and utilization of data across different platforms, thereby addressing accessibility challenges. In addition to technical solutions, providing adequate training and support to researchers is vital. This involves educating them on the best practices in data management, including the effective use of data management tools and technologies, and imparting knowledge about the complexities of data security and compliance in a hybrid HPC-cloud environment. By combining unified data management systems, collaborative frameworks, and comprehensive training for researchers, a more efficient and accessible data management landscape can be created, catering to the diverse needs of the scientific community using HPC and cloud platforms in concert.

**MAGRO:** Michela raises a crucial aspect of training and collaboration. We must create central cloud data storage hubs that serve a wide research community. Cloud-based data repositories, envisioned as central hubs, should manage the substantial data volumes that HPC and AI applications generate. These repositories should significantly enhance collaborative research by providing unified data storage and access to a hybrid HPC-cloud platform. They should be characterized by scalable storage solutions, complemented by advanced features such as data versioning, replication, and automated backups, ensuring the integrity and availability of scientific data. In addition to offering robust storage capabilities, these cloud-based repositories should prioritize data security and compliance. Implementing stringent security measures is crucial to safeguard sensitive data. This includes integrating sophisticated encryption techniques, secure data transfer channels, and effective access control mechanisms. Furthermore, ensuring compliance with data protection regulations, such as the General Data Protection Regulation for personal data, is fundamental to these protocols. Though cloud-based repositories integrate security and compliance, the scientific community can benefit from a secure and collaborative environment. The overarching approach streamlines data management and fosters a culture of shared knowledge and resources, vital for advancing research in today's data-driven landscape.

**POXON:** Bill's proposal for cloud-based repositories is compelling, especially for fostering collaboration. To support the ability to collaborate and manage data in the most flexible way, however, there is also a need for a variety of data location, migration, and access solutions. These solutions must provide seamless integration with existing HPC infrastructures.

**SEELAM:** The emphasis on infrastructure is crucial, but let's not overlook the human element. Training and support for researchers are fundamental to navigating this hybrid landscape. A unified system is only as effective as the people who use it. As we move forward, the focus should be on creating an ecosystem that supports both technological advancements and the community it serves. Central hubs and cloud-based repositories must be designed with the end user in mind, ensuring ease of use and accessibility.

**QUESTION 5:** *To end this panel, given the multifaceted discussion on the convergence of HPC and cloud computing, let's imagine how this convergence is shaping the future of scientific computing.*

**MILROY:** The convergence of HPC and cloud computing marks a transformative era for scientific computing, presenting a blend of opportunities and challenges. While it promises unparalleled computational power and flexibility, it also demands reimagined data management, resources, and security. The evolving landscape necessitates robust and scalable infrastructure and sophisticated software systems capable of bridging the gap between traditional HPC environments and the dynamic nature of cloud platforms. As we navigate this transition, the focus must remain on ensuring that the scientific community has the tools and resources to harness the full potential of these combined technologies.

**GAMBLIN:** HPC and cloud convergence has the potential to democratize access to data and to computation and to broaden the reach of scientists by

allowing many more users to access and build on their work. There are still challenges around integration, including data transfer, costs, and bringing HPC users up to speed on deploying distributed services. There is only so much any one scientist can learn, and I do not think that we know yet how best to structure development teams around hybrid science. There is a lot of capability in the cloud, but there will also be a lot of adjustment required for the HPC community to use it.

**TAUFER:** This convergence is fundamentally altering the landscape of scientific computing, particularly fostering collaboration and democratizing access to computational resources. However, it also underscores the importance of comprehensive research training and support. As the demands on infrastructure and software systems evolve, so too must our approach to empowering the scientific community. By providing the necessary tools and education, we can unlock the full potential of HPC and cloud computing convergence for scientific discovery.

**MAGRO:** Integrating HPC with cloud computing introduces a new paradigm for scientific computing, where cloud-based data repositories become central to collaborative research. These repositories not only address the immediate challenges of data management and accessibility but also pave the way for innovative research methodologies. As we look to the future, the key will be ensuring that these systems are secure and capable of meeting the increasingly sophisticated demands of the scientific community. This convergence is not just about overcoming technical challenges; it is about creating an ecosystem that accelerates the pace of innovation in scientific computing.

**POXON:** The convergence of HPC and cloud computing is reshaping the economic landscape and pace of innovation in scientific computing. It offers unprecedented opportunities for cost-effectively scaling computational resources and access to a wider variety of technology and collaboration opportunities. This convergence also introduces challenges, particularly around optimizing costs while ensuring the availability and performance of compute, network, and storage resources. The future of scientific computing will increasingly rely on innovative financial models and scalable infrastructures that support the fluctuating demands of research projects.

**JONES:** The merging of HPC and cloud computing is not just a technological evolution; it's a strategic opportunity for the future of scientific computing. This convergence offers substantial potential for enhancing computational capabilities and fostering global collaboration. However, ensuring seamless integration between vastly different computing paradigms remains a challenge. The future demands a holistic approach that considers the immediate technical and security needs and the long-term adaptability of infrastructure and software systems. Developing strategic frameworks that leverage the strengths of HPC and cloud computing and prepare us for future challenges is crucial as we look ahead. Emphasizing agility, scalability, collaboration, and a clear understanding of value will be key to unlocking the full potential of this convergence, ensuring that the scientific community remains at the forefront of innovation, taking advantage of the driven innovation and investments of the hyperscalers.

The convergence of HPC and cloud computing is reshaping the economic landscape and pace of innovation in scientific computing.

As we move forward, we should focus on developing scalable and flexible solutions that support scientific datasets' growing complexity and size and deliver efficiency and agility to help maximize researcher productivity.

**SEELAM:** As we witness the convergence of HPC and cloud computing, the aspect that stands out most prominently is the evolving technical landscape, especially concerning security. This convergence is not just shaping the future; it demands reevaluating how we approach the security of our computational infrastructures. Integrating HPC with cloud environments brings forth unique opportunities for advancing scientific computing but also introduces complex challenges in maintaining data integrity and security across diverse and distributed systems. The key to navigating this future lies in developing robust and adaptable security protocols that protect sensitive data across both HPC and cloud platforms. Moreover, this convergence necessitates a continuous evolution of our infrastructure and software systems to meet the ever-increasing demands of scientific research, ensuring that security remains a cornerstone of technological advancement.

## ATTENDEES' RESPONSE

Attendees at the panel also highlighted several key aspects of integrating HPC and cloud technologies, focusing on the unique challenges and opportunities each presents. Attendees discussed balancing HPC's capabilities with the cloud's scalability and flexibility, emphasizing the importance of addressing security, engaging in strategic planning, finding cost-effective solutions, and developing unified software stacks to support a wide range of workloads in a hybrid computing environment. They stressed the need to understand the total cost of ownership when comparing cloud and HPC solutions, considering both immediate and long-term

infrastructure and operational expenses. Concerns about data security in the cloud emerged, especially after incidents where misconfigurations exposed sensitive information.[6,7] Ensuring compliance with zero-trust mandates for data security and user access became a vital topic.

The discussion also covered the potential for HPC to adopt more cloud-like features to enhance service quality, focusing on the need for software stack convergence over hardware differences. Attendees critiqued the inefficiency of cloud-native scheduling solutions compared to traditional HPC schedulers and called for a unified approach to effectively serve both HPC and cloud workloads. The dialogue touched on shifts toward higher reliability engineering, automation, and API services inspired by the site reliability engineering movement,[8] outlining the need for HPC environments to adopt cloud practices such as automated provisioning to improve user experience and operational efficiency, aligning with the growing preference for simple cloud-like access to HPC resources without extensive IT support. The discussion also pointed out the irreversible nature of transitioning to the cloud, along with the long-term costs and potential price setting by dominant cloud providers, emphasizing strategic planning for scalability and financial sustainability, highlighting the increasing demand for user-friendly cloud-like infrastructure within traditional data centers to meet user expectations for seamless and efficient computing experiences.

In conclusion, the ongoing convergence of HPC and cloud computing is pivotal for future technological progress. It fosters enhanced computational power and flexibility while also addressing significant challenges such as data management, security, and resource integration. This fusion is expected to drive substantial scientific advancements, leveraging the strengths of both domains to achieve groundbreaking research outcomes.

## ACKNOWLEDGMENT

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-JRNL-862477-DRAFT) and was supported by the LLNL-LDRD Program under Project 22-ERD-041. The support of NSF Grants 2138811, 2103845, 2334945, 2138296, and 2331152 is also acknowledged.

## REFERENCES

1. D. H. Ahn et al., "Flux: Overcoming scheduling challenges for exascale workflows," *Future Gener. Comput. Syst.*, vol. 110, pp. 202–213, Sep. 2020, doi: 10.1016/j.future.2020.04.006.
2. T. Patki et al., "Fluxion: A scalable graph-based resource model for HPC scheduling challenges," in *Proc. Workshops Int. Conf. High Perform. Comput., Netw., Storage, Anal. (SC-W)*, New York, NY, USA: Association for Computing Machinery, pp. 2077–2088, doi: 10.1145/3624062.3624286.
3. "Why we built an AI supercomputer in the cloud Introducing Vela, IBM's first AI-optimized, cloud-native supercomputer." IBM. Accessed: Feb. 7, 2023. [Online]. Available: https://research.ibm.com/blog/AI-supercomputer-Vela-GPU-cluster
4. "Training on TPU pods." Google. Accessed: Apr. 24, 2024. [Online]. Available: https://cloud.google.com/tpu/docs/training-on-tpu-pods
5. "Spack: A flexible package manager supporting multiple versions, configurations, platforms, and compilers." Spack. Accessed: 2024. [Online]. Available: https://spack.io/
6. J. Li, W. Xiao, and C. Zhang, "Data security crisis in universities: Identification of key factors affecting data breach incidents," *Humanities Social Sci. Commun.*, vol. 10, no. 1, pp. 270, 2023, doi: 10.1057/s41599-023-01757-0.
7. A. W. Malik, D. S. Bhatti, T. J. Park, H. U. Ishtiaq, J. C. Ryou, and K. I. Kim, "Cloud digital forensics: Beyond tools, techniques, and challenges," *Sensors (Basel)*, vol. 24, no. 2, Jan. 10, 2024, Art. no. 433, doi: 10.3390/s24020433.
8. The United States Research Software Engineer Association. Accessed: 2021–2024. [Online]. Available: https://us-rse.org/

**MICHELA TAUFER** is an ACM Distinguished Scientist and Dongarra Professor in High-Performance Computing at the University of Tennessee Knoxville, Knoxville, TN 37996 USA. Contact her at taufer@acm.org.

**DANIEL MILROY** is a computer scientist at Lawrence Livermore National Laboratory, Livermore, CA 94550 USA. Contact him at milroy1@llnl.gov.

**TODD GAMBLIN** is a Distinguished Member of the technical staff at Lawrence Livermore National Laboratory, Livermore, CA 94550 USA. Contact him at gamblin2@llnl.gov.

**ANDREW JONES** is a leader at Microsoft Azure, Microsoft, Redmond, WA 98052 USA. Contact him at Andrew.M.Jones@microsoft.com.

**BILL MAGRO** is a chief technologist for HPC at Google LLC, Mountain View, CA 94043 USA. Contact him at wmagro@google.com.

**HEIDI POXON** is a principal HPC technologist at Amazon Web Services, Seattle, WA 98109 USA. Contact her at hpoxon@amazon.com.

**SEETHARAMI SEELAM** is a principal research staff member and technical lead at IBM T. J. Watson Research Center, Armonk, NY 10504 USA. Contact him at sseelam@us.ibm.com.