EVERYWHERE & NOWHERE: ENVISIONING A COMPUTING CONTINUUM FOR SCIENCE

A PREPRINT

🆔 **Manish Parashar**[*]
Scientific Computing and Imaging (SCI) Institute,
University of Utah,
Salt Lake City, UT, 84112, USA
manish.parashar@utah.edu

June 10, 2024

ABSTRACT

Emerging data-driven scientific workflows are seeking to leverage distributed data sources to understand end-to-end phenomena, drive experimentation, and facilitate important decision-making. Despite the exponential growth of available digital data sources at the edge, and the ubiquity of non trivial computational power for processing this data, realizing such science workflows remains challenging. This paper explores a computing continuum that is everywhere and nowhere – one spanning resources at the edges, in the core and in between, and providing abstractions that can be harnessed to support science. It also introduces recent research in programming abstractions that can express what data should be processed and when and where it should be processed, and autonomic middleware services that automate the discovery of resources and the orchestration of computations across these resources. [2]

1 Everywhere & Nowhere: The Emerging Computing Continuum

Disruptive innovations coupled with technological advances across the computing stack are resulting in dramatic changes in all aspects of the scientific computing/high-performance computing landscape (HPC). These changes include the emergence of novel processor and system architectures influenced by the end of Dennard scaling and the slowing of Moore's law, and driven by in large part by the transcendence of AI and the dominance of the hyperscalers Reed et al. [2022]; pervasive availability of non trivial (and growing) computing capacities at the edge, within the networks and along the data-path, and extreme capabilities at high-performance computing and cloud data centers; rapidly growing data sources and data volumes and rates; and increasing network bandwidths and in-network services.

These changes are coupled with a shift in the workloads that are driving the innovations. These workloads are increasingly data-driven, leverage artificial intelligence/machine learning (AI/ML) techniques, and have changing and flexible notions of precision (i.e., exploring 8 and 16 bit, in addition to the typical 64 bit). The supporting software stacks for resource access and application development, deployment, and execution are similarly evolving to include, for example, as-a-service, serverless, and containerized approaches that are typically found in enterprise systems.

Perhaps the most significant change is in the value structures, i.e., what users value most. Maximizing performance was traditionally valued above all else, but we are increasingly seeing that other aspects, such as ease of access and use,

---

[*]*Manish Parashar* is Director of the Scientific Computing and Imaging (SCI) Institute, Chair in Computational Science and Engineering, and Presidential Professor, Kalhert School of Computing at the University of Utah. Manish is the founding chair of the IEEE Technical Community on High Performance Computing (TCHPC), and is Fellow of AAAS, ACM, and IEEE/IEEE Computer Society. For more information, please visit http://manishparashar.org.

[2]This paper is based on the author's IEEE Sidney Fernbach award presentation at SC23, The International Conference for High Performance Computing, Networking Storage and Analysis, Denver, CO, USA, November 2023. (https://sc23.supercomputing.org/)

time to science, energy/environmental impact, are becoming more important to users, and users are willing to give up some performance in order to gain in these dimensions.

Clay Christensen, in his book *Innovators Dilemma Christensen [1997]*, explored the impacts of such changes in value structures on innovation and technology. He noted that, when what the customers value changes, it creates opportunities for disruptive innovations allowing for the introduction of new technologies and approaches. We may be experiencing this in HPC, where existing performance levels, on average, meet user needs, and the differentiators are along other attributes such as usability, accessibility, robustness, environmental impacts, etc. At the same time, there are new applications, such as AI/ML-based workflows, that are driving alternate solutions across the stack. Looking at the current HPC landscape, it is clear that what applications are driving it, and how these applications use it, are evolving very quickly.

The result of these technological (and socio-technical) innovations and disruptions is a rapidly emerging, connected, and seamlessly accessible continuum of computational (computing, data, communication) capabilities Balouek-Thomert et al. [2019]. Furthermore, a vision of HPC enabled by this continuum is one that is *everywhere and nowhere*.

We are already seeing the pervasive integration and availability of non trivial computing capabilities everywhere, in our automobiles, in devices that we carry, in appliances that we use, etc. HPC is integrating itself in some form or another into all aspects of our lives, and these HPC capabilities will only increase as technology advances – i.e., *HPC is everywhere*.

At the same time, how we interact with and use HPC is also changing. Traditionally, using HPC has been an involved process – we develop specialized codes, which are compiled on a special system, and then queued to run on the HPC system – i.e., a very deliberate process often involving multiple systems. This process is in contrast with current trends across enterprise software stacks and the abstractions they provide to the user, e.g., those based on services, containers, serverless deployments, and notebook/gateway access. As the HPC community increasingly embraces these abstractions, we are seeing a move to a more transparent way of using HPC that is more seamlessly integrated into our workflows and our lives: i.e., *HPC will be everywhere and nowhere.*

2   Harnessing the Computing Continuum for Science

There exists a natural continuum across the evolving science ecosystem spanning large-scale instruments, experimental facilities, observatories, and sensor networks, all streaming data; high-speed networks and network services; and a range of computing capabilities along the continuum, from edge, to in-network, to large-scale data centers (see Figure 1).

This continuum is also spurring a natural evolution in the types of application workflows that are being developed. These workflows combine sensing and streaming data (e.g., from observatories and experimental facilities) with simulations and data-driven modeling and actuation, to understand, analyze, predict and actuate.

One class of such applications workflows that is enabled by the continuum and is being increasing deployed is end-to-end experiment management, where streaming data from an experiment or instrument is analyzed and modeled, and the result of the modeling is used to control, manage, and/or optimize the experiment. One example here is an instrumented oil-field workflow. The goal of this workflow is to use streaming data from an actual oilfield along with subsurface flow simulations and ML-based optimizers, to manage oil production, reduce environment impacts, etc. Bangerth et al. [2005]. Another example is from a fusion workflow, and implements a diagnostics system within the continuum that enables early prediction of anomalies while the tokamak experimental facility operates, which is important as these anomalies can damage the instrument[3].

Another class of applications that can be viewed as a natural extension, is digital twins for large-scale complex system. These systems are digital representations of actual real-world physical systems and can serve as vehicles for understanding, managing, optimizing, protecting, etc., the physical systems [4]. These applications highlight the need for combining real-time data acquisition with large-scale modelling, both data-driven and mathematical, and possible actuation, and the computing continuum will play a large role in making these systems a reality.

---

[3]WDMAPPâĂŞTheFirstSimulationSoftwareinFusionHistorytoCoupleTokamakCoretoEdgePhysics, https://www.exascaleproject.org/highlight/wdmapp-the-first-simulation-software-in-fusion-history-to-couple-tokamak-core-to-edg

[4]National   Academies   of   Sciences,   Engineering,   and   Medicine – Foundational   Research   Gaps and   Future   Directions   for   Digital   Twins,   https://nap.nationalacademies.org/catalog/26894/foundational-research-gaps-and-future-directions-for-digital-twins.
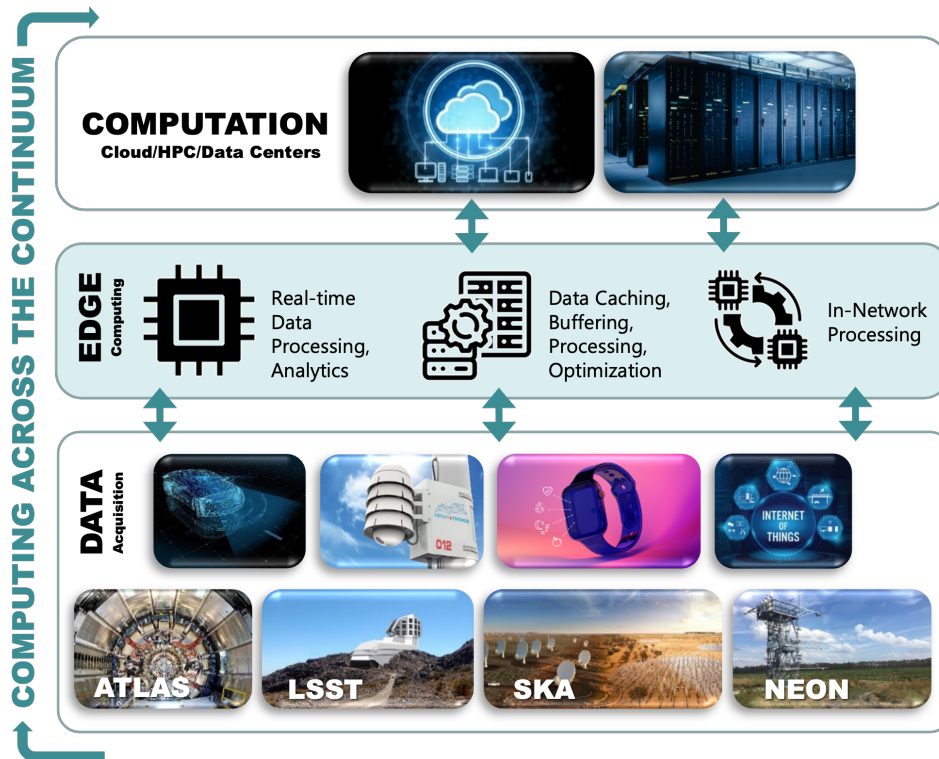
Figure 1: A computing continuum across the evolving science ecosystem spanning large-scale instruments, experimental facilities, observatories, and sensor networks, all streaming data; high-speed networks and advanced network services; and a range of computing capabilities and capacities along the continuum, from edge, to in-network, to large-scale data centers.

3 Urgent Computing and the Computing Continuum

An important class of applications enabled by the computing continuum is urgent computing. Urgent computing can be defined as computing under strict time and quality constraints to support decision-making with the desired confidence, and within a defined time interval Balouek-Thomert et al. [2020]. The goal is to leverage data and computations to support decision-making during an emergency. Figure 2 illustrated the urgent computing workflow, which uses the computing continuum to process data from a range of data sources along with other resources and services along the continuum to detect events, develop a response, and trigger actions.

Recent conversations about the importance and potential of urgent computing were triggered by experiences with the COVID19 HPC Consortium Brase et al. [2022] during the recent pandemic. The consortium was a remarkable international partnership across government, academia, and industry aimed at providing computing resources for pandemic-related research. It brought together an international group of over 40 resource providers and supported over 100 research projects, leading to important research outcomes spanning drug design, treatments, vaccine research, logistics, etc.

Although the consortium was very effective in highlighting the tremendous potential of computing and data in dealing with emergencies such as pandemics, extreme weather events, and wildfires, it also highlighted many infrastructure, partnership, training, and policy gaps that prevent these resources and efforts such as the consortium from achieving the desired impacts in an urgent situation, and the need for thinking differently about HPC, both technically and operationally.

A more immediate (and local) use case that is driving our current work in urgent computing is the impact of wildfires on air quality, and specifically, the impact of wildfires in California on air quality in the Salt Lake Valley in Utah. While the impact of wildfires on air quality is a broader issue, there are significant implications for the Salt Lake Valley (in Utah, USA) due to periodic temperature inversions, which trap cold air underneath a layer of warm air. Such inversions act like a "lid" on the Salt Lake Valley, causing particulate pollution to double quickly.
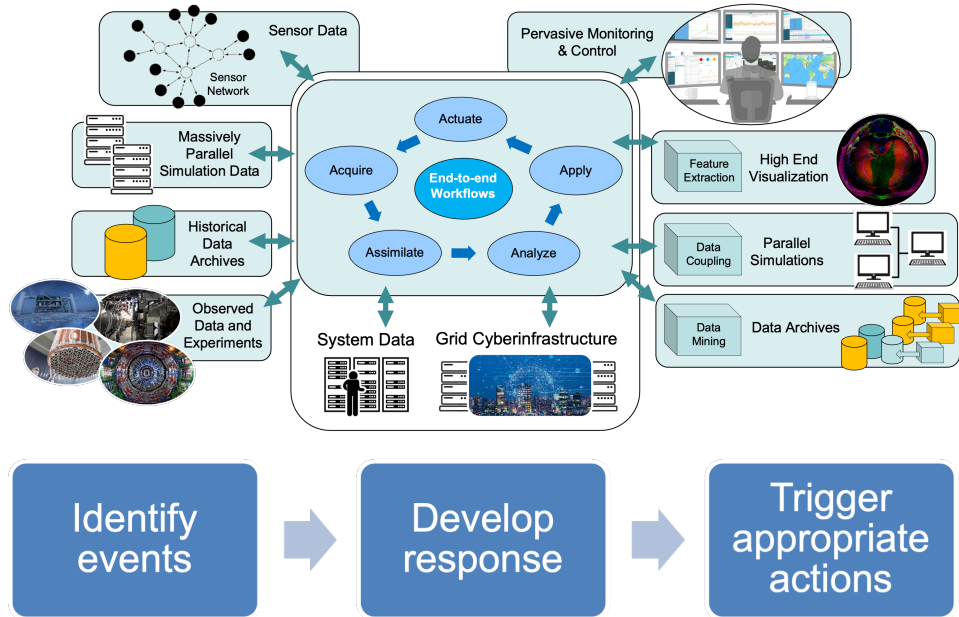
Figure 2: The urgent computing workflow uses the computing continuum to process data from a range of data sources, along with other resources and services along the continuum, to detect events, develop a response, and trigger actions.

Recent studies have noted that air pollution is a leading health risk factor globally and has a wide range of health implications, including on physical, respiratory, and mental health. Understanding and predicting these impacts requires combining a range of real-time data feeds about the fires, smoke, wind/weather conditions, and integrating atmospheric models, fire models, air quality models, etc. Furthermore, any action must include an understanding of the layout of the region, population distributions and demographics, road and traffic conditions, and other similar aspects. Note that air quality can deteriorate to severely unhealthy levels very quickly, making it an urgent challenge, and its impacts are not uniform, typically impacting under-resourced and under-developed neighborhoods the hardest.

Another example urgent computing use case driving our research is *Early Earthquake Warning*. In this use case, we used machine learning to analyze streaming 3-D time-series data along the continuum, at the edge and in-transit within the network Fauvel et al. [2020]. The goal here is to deliver alerts before the ground motion reaches sensitive areas.

Technically, ground motion in an earthquake is caused by two types of waves, Primary waves (or P-waves) and Secondary waves (or S-waves). P-waves move longitudinally along the Earth's crust and are around 1.7x faster than S-waves, which move in the transverse direction along Earth's interior. It is the S-waves that are responsible for severe damage. The goal of our research is to detect P-waves as quickly as possible, and it is important to do this as the data streams in. It is also important to combine 3D time series data from seismometers and GPS sensors to detect a range of earthquakes. Seismometers can effectively detect medium ($5 \leq$ magnitude < 6, Richter scale) earthquakes but are not as effective for large earthquakes ($6 \leq$ magnitude, Richter scale), for which GPS sensors are more effective. In this project, we leveraged the computing continuum to integrate and analyze these data streams at the edge and in transit and combined with modeling results coming from the core to predict the magnitude of seismic events, thereby allowing for timely alerts.

3.1   Research Challenges

There are many underlying research challenges in making urgent computing and the underlying data-driven workflows and the computing continuum a reality. Some key research questions include: (1) How do you drive computation through data? (2) How do you ensure security, privacy and trust in all aspects of the infrastructure? (3) How do you accommodate uncertainties in data and computation? (4) How do you build applications and manage workflows so that they can adapt to increase their value? (5) How do you continuously optimize workflow execution in a dynamic data-driven environment? (6) How do you develop robust system infrastructure and services to support dynamic execution? (7) How do you incorporate appropriate utility models, market models, social/trust models, etc.? and (8) How do you formulate the necessary policy and governance structures to manage operation?

Complementing fundamental advances in addressing these research questions is translational research that is critically important in achieving the urgent computing vision. Translational computer science researchAbramson and Parashar [2019][5] is the bi-directional integration and interplay between foundational research and the delivery and deployment of its outcomes. It aims to closely couple cycles of innovation between computer science and other disciplines to significantly accelerate the transformative impact of computer science.

3.2 Recent Research Efforts

Our recent research addresses some of the underlying research issues that were listed above. These include:

How do you drive computation through data?   Urgent workflows are triggered by attributes and/or content of data streams, and the data attributes/content determines what, when, and where execute computations. For example, the online analysis of 3D time series data from seismometers and GPS sensors triggers earthquake detection workflows on edge and/or cloud/HPC resources. The R-Pulsar programming system Renart et al. [2019] leverages the Associative Rendezvous (AR) interaction model to allow users to programmatically define data-driven workflows executing across the computing continuum as reactive behaviors based on the content of streaming data. It provides abstractions to express workflow topologies that are triggered based on the availability of resources and/or data as well as data values, statistical trends over time/spatial windows, etc., i.e., data streams are evaluated at runtime to decide when, how, and where to process their data.

How do you discover and aggregate resources based on current needs?   Discovering, federating, and utilizing (computing, data, etc.) resources along the computing continuum in an online manner are essential to urgent computing, including adapting the federation and discovering and aggregating new resources as the application needs evolve and/or new resources become available. Our research has leveraged constraints-based autonomic federation to realize a dynamic software-define system that can support urgent workflows AbdelBaky et al. [2018]. We have also developed recommendation systems for scientific data to support intelligent data discovery and delivery Parashar and Altintas [2023]. Specifically, this research 1) uses user query analysis techniques that model access patterns and associated localities and affinities; 2) optimizes data caching, data prefetching, and data steaming mechanisms to support optimized push-based data delivery; and 3) develops a data recommendation framework based on the collaborative knowledge-aware graph-attention network (CKAT) recommendation model.

How do you manage execution (and QoS) in a dynamic environment?   The highly dynamic requirements of urgent application workflows coupled with inherent dynamism of the computing continuum warrant autonomic runtimes that can effectively manage and optimize execution. Our research is developing autonomic runtime services for data and workflow management as part of the Virtual Data Collaboratory and National Data Platform Parashar and Altintas [2023] projects. Specfically, these services manage workflow scheduling and execution across the computing continuum based on user-defined policies and constraints to implement necessary tradeoffs.

4 Conclusion: A Call to Action

We are witnessing urgent events with increasing frequency and increasing impacts. The HPC community has an opportunity to collectively leverage its expertise and the computing continuum to make a difference. One recent example is Hurricane Ottis. When hurricane Ottis recently made landfall in Acapulco, Mexico, as the first Category 5 storm ever to hit the Pacific Coast of North or South America, no formal hurricane warning had been issued. In fact, 16 hours before landfall, the National Hurricane Center still forecast only a Category 1 hurricane. Within hours, it grew into a record-breaking Category 5 hurricane with winds of 165 miles per hour. Ottis is probably the most expensive hurricane in Mexican history. Experts analyzing the event and its progression have have noted that real-time sensor data along with the ability to integrate other parameters (e.g., water surface temperature and salinity) into the model could have been more effective in predicting this hurricane and its strength. Urgent computing approaches that leverage the computing continuum can have a tremendous impact in such events, in how we predict, detect, and manage our response, as well as how can we learn and adapt and evolve to do better.

One initiative that is focused on establishing policies, structures, and mechanism for leveraging the computing continuum for support urgent applications is the US Whitehouse Office of Science and Technology Policy (OSTP) led *National Strategic Computing Reserve (NSCR)*[6]Friedlander and Parashar [2021]. NSCR envisions an advanced computing

---

[5]Translational Computer Science, `https://translational-cs.org/`.

[6]The report, "National Strategic Computing Reserve: A Blueprint," available at `https://www.nitrd.gov/national-strategic-computing-reserve-blueprint/`, outlines a Federal proposal for setting up a National Strategic Com-

Figure 3: A call to action: We are witnessing urgent events with increasing frequency and increasing impacts, and the computing continuum along with urgent computing has the potential to help us understand, manage, and mitigate the impacts of these event. The HPC community has an opportunity to collectively leverage its expertise and the computing continuum to make a difference.

cyberinfrastructure as a strategic national asset that can be mobilized during an emergency response. Its goal is to ensure the availability of a ready *reserve* of resources (computing, data, software, services) and expertise that can be leveraged nimbly in times of urgent need, and to establish policies, processes, and agreements to enable effective resource mobilization and coordinate across agencies, stakeholder communities, and other national reserves. The NSCR vision is analogous to the roles of the US Civil Reserve Air Fleet and the United States Merchant Marine (among others) that can be called upon to assist the military in a crisis. Although this initiative is in its early stages, it can potentially transform how we effectively leverage the computing continuum to detect and respond to national and global emergencies.

5   Acknowledgements

References

Daniel Reed, Dennis Gannon, and Jack Dongarra. Reinventing high performance computing: Challenges and opportunities, 2022.

Clayton M. Christensen. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business School Press, USA, 1997. ISBN 0875845851.

Daniel Balouek-Thomert, Eduard Gibert Renart, Ali Reza Zamani, Anthony Simonet, and Manish Parashar. Towards a computing continuum: Enabling edge-to-cloud integration for data-driven workflows. *The International Journal of High Performance Computing Applications*, 33(6):1159–1174, 2019.

Wolfgang Bangerth, Hector Klie, Vincent Matossian, Manish Parashar, and Mary F Wheeler. An autonomic reservoir framework for the stochastic optimization of well placement. *Cluster Computing*, 8:255–269, 2005.

Daniel Balouek-Thomert, Ivan Rodero, and Manish Parashar. Harnessing the computing continuum for urgent science. *ACM SIGMETRICS Performance Evaluation Review*, 48(2):41–46, 2020.

Jim Brase, Nancy Campbell, Barbara Helland, Thuc Hoang, Manish Parashar, Michael Rosenfield, James Sexton, and John Towns. The covid-19 high-performance computing consortium. *Computing in Science & Engineering*, 24(1): 78–85, 2022.

puting Reserve (NSCR) that can be called on in times of national crisis to rapidly activate a multi-sector advanced computing reserve infrastructure that can speed solutions, and it defines a blueprint for operational and coordination structures that will support an NSCR implementation.

Kevin Fauvel, Daniel Balouek-Thomert, Diego Melgar, Pedro Silva, Anthony Simonet, Gabriel Antoniu, Alexandru Costan, Véronique Masson, Manish Parashar, Ivan Rodero, et al. A distributed multi-sensor machine learning approach to earthquake early warning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 403–411, 2020.

David Abramson and Manish Parashar. Translational research in computer science. *Computer*, 52(9):16–23, 2019.

Eduard Gibert Renart, Daniel Balouek-Thomert, and Manish Parashar. An edge-based framework for enabling data-driven pipelines for iot systems. In *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 885–894. IEEE, 2019.

Moustafa AbdelBaky, Javier Diaz-Montes, and Manish Parashar. Software-defined environments for science and engineering. *The International Journal of High Performance Computing Applications*, 32(1):104–122, 2018.

Manish Parashar and Ilkay Altintas. Toward democratizing access to science data: Introducing the national data platform. In *2023 IEEE 19th International Conference on e-Science (e-Science)*, pages 1–4. IEEE, 2023.

Amy Friedlander and Manish Parashar. The u.s. needs a national strategic computing reserve, 2021.