Report from Dagstuhl Seminar 23171

Driving HPC Operations With Holistic Monitoring and Operational Data Analytics

Jim Brandt*[1], **Florina Ciorba**†[2], **Ann Gentile**†[3], **Michael Ott**†[4], **and Torsten Wilde**†[5]

1 **Sandia National Laboratories, US.** brandt@sandia.gov
2 **University of Basel, CH.** florina.ciorba@unibas.ch
3 **Sandia National Laboratories, US.** gentile@sandia.gov
4 **Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, DE.** ott@lrz.de
5 **Hewlett Packard Enterprise – Böblingen, DE.** wilde@hpe.com

Abstract

Advances in analytic approaches have brought the vision of efficient High Performance Computing (HPC) operations enabled by dynamic analysis driving automated feedback and adaptation within reach. Many HPC centers have started the development and deployment of frameworks to enable continuous and holistic monitoring, archiving, and analysis of performance data from their production machines and related infrastructures. The impact of such frameworks rests upon the ability to effectively analyze such data and to take action based on analysis results. Analytic techniques have been successfully developed and applied in other domains but their features may not apply directly to HPC operations data and situations. Response options are limited in HPC implementations. Leveraging, adapting, and extending analysis techniques and response options would open up new avenues for research and development of actionable analytics that can drive more intelligent operations through both manual and automated response to conditions of interest.

This Dagstuhl Seminar 23171 brought together practitioners and researchers in the areas of HPC system management and monitoring, analytics, and computer science to collaboratively work on developing community solutions for revolutionizing HPC system operations. The topics discussed in this seminar spanned use cases, data and analytic approaches required to address the use cases, use of analysis results to improve performance and operations, and research in the development and use of autonomous feedback loops.

Seminar April 23–28, 2023 – https://www.dagstuhl.de/23171
2012 ACM Subject Classification General and reference → General literature; Information systems → Information systems applications
Keywords and phrases Monitoring, Operational Data Analytics, Dagstuhl Seminar, WAFVR
Digital Object Identifier 10.4230/DagRep.13.4.98

* Editorial Assistant / Collector
† Editor / Organizer

Driving HPC Operations With Holistic Monitoring and Operational Data Analytics, *Dagstuhl Reports*, Vol. 13, Issue 4, pp. 98–120
Editors: Jim Brandt, Florina Ciorba, Ann Gentile, Michael Ott, and Torsten Wilde
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Jim Brandt (Sandia National Laboratories, US, brandt@sandia.gov)
Florina Ciorba (University of Basel, CH, florina.ciorba@unibas.ch)
Ann Gentile (Sandia National Laboratories, US, gentile@sandia.gov)
Michael Ott (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, DE, ott@lrz.de)
Torsten Wilde (Hewlett Packard Enterprise, DE, wilde@hpe.com)

The Dagstuhl Seminar 23171 (April 23–28, 2023) brought together 35 practitioners and researchers in the areas of HPC system management and monitoring, data analytics, and computer science to collaboratively work on developing community solutions for revolutionizing HPC system operations. Autonomous operations have long been the vision for efficient HPC system operations due to the size and complexity of current and evolving HPC systems and the need for pervasive, low-latency response. Autonomous operations are a complex topic encompassing monitoring, analysis, feedback, and response. The seminar goals were to make substantial progress on the technical, community, and funding challenges necessary for the community to move forward and reach this vision.

The seminar schedule comprised a mix of keynotes presentations, seed and position talks, enlisted and ad-hoc lightning talks, interleaved with plenary discussions and working group discussions, both in the seminar rooms as well as outdoors.

These program elements and the active participation of the attendees lead to many fruitful discussions on the following topics:

- Center-specific urgent use cases that drive data collection, analysis, and response requirements across the variety of institutions represented.
- Types of available data, including sources, semantics, and fidelity, to support continuous analyses.
- Requirements for actionable analytics. What is needed to convert raw data into information upon which action can be taken (e.g., confidence measures, explainability requirements not inherent in AI approaches, representation of results, latency, etc.)?
- Applicability of existing analytics and informatics approaches to the domain-specifics of HPC operations. While there are many promising ML/AI approaches in other domains (e.g., image/speech processing, autonomous vehicles), it is not yet clear how many and which of those apply to the HPC operations and research domains (e.g., the occurrence of rare fault events, discontinuity of inertia-less measurements).
- Opportunities for response involving infrastructure, hardware, system software, and applications. Identification of feedback hooks that would need to be added to existing and evolving system components (e.g., hardware, firmware, system software, application software) to support automated response.
- Exploration of formalism and architectural design patterns from the field of Self-Adaptive Systems to facilitate common, interoperable, and interchangeable design and development paths forward.

The technical presentations and engaging discussions reinforced the urgent need and desire for a community approach to advance the state and practice of HPC Monitoring and Operational Data Analytics, with the goal of revolutionizing HPC operations and research, in order to deliver efficient and sustainable HPC systems and applications.

The fundamental results of this community discussion are given in Section 10 of this report. These include assessments of the state of autonomous loops in HPC operations and assessments of challenges and opportunities. The community agreed to continue meeting, on a monthly basis and at upcoming community-relevant events. They further agreed to develop *proofs of concepts* for concrete use cases that will showcase both the need for holistic monitoring and analysis and their benefits for more efficient HPC operations. These proofs of concepts will also serve as a basis for technical design decisions and prototype solutions to be deployed in various HPC systems. The final goals of our effort are to continue to build and progress on a community collaborative *technical path* forward, a community *interaction path* forward, and a community *collaborative funding path* forward, as described in Section 11, to fulfill the vision of autonomous and efficient HPC system operations.

2 Table of Contents

Executive Summary

Jim Brandt, Florina Ciorba, Ann Gentile, Michael Ott, and Torsten Wilde 99

Keynotes

Community Readiness and Opportunities for Progress in HPC Monitoring, Analysis, Feedback, and Response
Jim Brandt 104

Visions for HPC System and Facility onitoring and Operational Data Analytics
Utz-Uwe Haus 104

Talks: Use Cases

Continuous Performance Counter Sampling
Michael Ott 104

System Monitoring from the Performance Measurement Perspective
Kevin Huck 105

HPC Operations and Monitoring
Esa Heiskanen 105

Position Talks

Workflows & Patchwork: Building a Big Picture Out of Scraps of Data
Taylor Groves 105

Challenges for Holistic Monitoring When Attempting Codesign Support
Terry Jones 106

Large-scale Vendor-Neutral Power Monitoring
Tapasya Patki 106

Towards an Efficient and Concise Characterization of Temporal I/O Behavior
Francieli Boito 106

Containerized Real Time Job Anomaly Detection
Mike Showerman 107

A Holistic View of Memory Utilization on HPC Systems: Current and Future Trends
Ivy Peng, Kathleen Shoga 107

Talks: Feedback Driven Response

Where Rubber Meets the Road: Challenges in Actionable Response to I/O Performance Data
Phil Carns 108

Driving Response to Uncertainty in Job Duration and Loss of Work
Frédéric Suter 108

Hierarchical Control for Runtime Adaptation
Valeria Cardellini 108

Laying the Foundation for Self-Organizing Systems
Ann Gentile . . . 109

Talks: Analytic and Informatics Approaches

Time Series Analysis for Performance Monitoring of HPC Systems
Abdullah Mueen . . . 109

Missing Gaps and Opportunities in HPC Operational Data Analytics
Devesh Tiwari . . . 109

Continuously Detecting Workflow Anomalies using Graph Neural Networks – Lessons Learnt
Krishnan Raghavan . . . 110

Best Practices on the Practical Use of Machine Learning in Hybrid Memory Management
Thaleia Doudali . . . 110

AI & Analytics in Service of Green Computing
Hilary Egan . . . 110

Talks: Actionable Analytics and Response

AIOPS: Artificial Intelligence for IT Operations in HPC Systems
Jeff Hanson, Torsten Wilde . . . 111

Energy Management on HPC Systems
Oriol Vidal . . . 111

Talks: Autonomous Loop Lightning Talks

Malleability, Monitoring and Modeling
Isaías A. Comprés Ureña . . . 112

Control Loops in the PaPP Project
Sven Karlsson . . . 112

Job Efficiency Reporting
Thomas Jakobsche . . . 112

Code Path and Parameter Selection for Compute Loops
Thomas Gruber . . . 113

Feedback Loop for Mitigating Assignment of Slow OSTs for File I/O
Jim Brandt . . . 113

Fundamental Results

Assessment of the state of autonomous feedback and response loops in HPC operations . . . 114

Assessment of most important challenges and opportunities . . . 114

Planned Paths Forward

Identification of a community collaborative technical path forward . . . 116

Identification of community interaction path forward . . . 117

Identification of community collaborative funding path forward . . . 117

Acknowledgments 117

Participants 119

Remote Participants 120

3 Keynotes

3.1 Community Readiness and Opportunities for Progress in HPC Monitoring, Analysis, Feedback, and Response

Jim Brandt (Sandia National Laboratories – Livermore and Albuquerque, US, brandt@sandia.gov)

This talk presents work in progress at Sandia National Laboratories. It includes a comprehensive vision on holistic monitoring and feedback as well as the mechanisms currently being used for data collection and feedback. Approaches to Machine Learning that have been explored are presented along with work in progress for turning signals from a ML engine into feedback to drive response for mitigating specific application performance degrading conditions in an HPC Lustre file system.

3.2 Visions for HPC System and Facility onitoring and Operational Data Analytics

Utz-Uwe Haus (HPE HPC/AI EMEA Research Lab – Zurich, CH, utz-uwe.haus@hpe.com)

Starting from the well-known 4x4 table of characteristics for component vs action levels for MODA (Bates et al.) we highlight the importance and tight interconnection to digital twins of data centers, in particular HPC data centers including System level twins. Recent advances in data-aware middlewares, scheduling simulation tools and high frequency monitoring are discussed as essential components in a holistic, sustainability focused approach.

4 Talks: Use Cases

4.1 Continuous Performance Counter Sampling

Michael Ott (Leibniz Supercomputing Centre – Garching, DE, ott@lrz.de)

User applications very often are black boxes for HPC operators in terms of requirements and runtime characteristics. Yet, having this information about user workloads would be beneficial for multiple purposes: procurement of new systems that better fit the user requirements, influencing scheduling decisions to optimize system throughput, or identify jobs with low performance. This talk will present ideas on continuous performance counter sampling to obtain better insight into HPC workloads without user interaction.

4.2 System Monitoring from the Performance Measurement Perspective

Kevin Huck (University of Oregon – Eugene, OR, US, khuck@cs.uoregon.edu)

From the application performance measurement perspective, limited to the user space domain, there are several motivations for runtime monitoring of the application, operating system and hardware. These motivations run the gamut between simple curiosity to identifying the causes of software failures. In this talk, we will explore these motivations and discuss the barriers to getting access to exposed as well as privileged data to contribute to performance understanding.

4.3 HPC Operations and Monitoring

Esa Heiskanen (CSC – IT Center for Science – Kajaani, FI, esa.heiskanen@csc.fi)

Installing and operating large HPC systems is complicated from an operational point of view. There are always unpredictable things that happen and those challenges and risks can be mitigated with good preparation and design of different elements. Having better visibility and predictability to system and applications helps operators to understand the workload of a system and perform preventive anomaly detection and maintenance. Digital twins are defined in 5 levels as descriptive, informative, predictive, comprehensive and autonomous. Building a digital twin of a data center and HPC system is challenging and needs co-operations with multiple parties to achieve goals that give real benefits of digital twin.

5 Position Talks

5.1 Workflows & Patchwork: Building a Big Picture Out of Scraps of Data

Taylor Groves (NERSC – Berkeley, US, tgroves@lbl.gov)

The focus of NERSC is moving from simulations limited to a single site to complex workflows that span multiple sites – incorporating both data center and HPC technologies. This patchwork of telemetry and control creates additional challenges for data analysis. In this lightning talk, I present the perspective from NERSC and highlight challenges and opportunities for future research.

5.2 Challenges for Holistic Monitoring When Attempting Codesign Support

Terry Jones (Oak Ridge National Laboratory – Knoxville, US, trjones@ornl.gov)

For many years, High Performance Computing (HPC) relied on performance improvements based chiefly on frequency scaling; during that era, performance portability was straightforward between different generations of machines. Current computer architecture trends have shifted to other strategies for improvements and exhibit rapidly increasing complexity and extreme heterogeneity to provide performance gains. Codesign activities have emerged as vital processes for mapping applications to the underlying hardware. This talk covers key challenges that arise for holistic monitoring when trying to support codesign.

5.3 Large-scale Vendor-Neutral Power Monitoring

Tapasya Patki (Lawrence Livermore National Laboratory – Livermore, US, patki1@llnl.gov)

Low-level power and performance dials vary significantly from vendor-to-vendor, reducing the portability of system software and monitoring tools. This talk presents Variorum, a library for vendor-neutral power management, and its integration with the Lightweight Distributed Metric System (LDMS).

5.4 Towards an Efficient and Concise Characterization of Temporal I/O Behavior

Francieli Boito (University of Bordeaux/Inria – Bordeux, FR, francieli.zanon-boito@inria.fr)

Existing options for I/O characterization on HPC systems include the creation of complete traces – a log of all I/O operations and their information – or of aggregated statistics, such as the total number and size of accesses. In this talk, I discuss the importance of temporal I/O behavior information and recent advances in using signal processing techniques to obtain it and represent it in a concise manner.

5.5 Containerized Real Time Job Anomaly Detection

Mike Showerman (NCSA – Urbana, US, mung@illinois.edu)

Job anomaly detection allows for identifying irregular job behavior that may affect job execution times, other user's job performance, or system throughput. Examples include load imbalances, stalled jobs, unused requested resources (like GPUs), or high shared resource usage. This talk presents the container-based job anomaly detection that was deployed on the Blue Waters system at NCSA. It allowed for analyzing thousands of jobs in real time on a 27,000 nodes HPC system.

5.6 A Holistic View of Memory Utilization on HPC Systems: Current and Future Trends

Ivy Peng (KTH Royal Institute of Technology – Stockholm, SE, ipeng@acm.org)
Kathleen Shoga (Lawrence Livermore National Laboratory, US, shoga1@llnl.gov)

The Sonar monitoring infrastructure is a central system-monitoring infrastructure deployed at Livermore Computing. Its main components include sample monitoring, data ingestion, staging, archive, and analytics. In this work, we showcase the analysis results from Sonar monitoring infrastructure that can drive the system co-design for the memory subsystem on next generation HPC systems. Memory subsystem is one crucial component of a computing system. Co-designing memory subsystems becomes increasingly challenging as workloads continue evolving on HPC facilities and new architectural options emerge. This work provides a large-scale study of memory utilization with system-level, job-level, temporal and spatial patterns on a CPU-only and a GPU-accelerated leadership supercomputer. From system-level monitoring data that spans three years, we identify a continuous increase in memory intensity in workloads over recent years. We showcase how monitoring data on production systems can reveal different hotspots in memory usage in applications on large-scale systems. We introduce two metrics, “spatial imbalance” and “temporal imbalance”, to quantify the imbalanced memory usage across compute nodes and throughout time in jobs. Finally, we identify representative temporal and spatial patterns from real jobs, providing quantitative guidance for research on efficient resource configurations and novel architectural options.

6 Talks: Feedback Driven Response

6.1 Where Rubber Meets the Road: Challenges in Actionable Response to I/O Performance Data

Phil Carns (Argonne National Laboratory – Lemont, US, carns@mcs.anl.gov)

Methods of measuring and recording performance data have been around as long as computers themselves, and the first reaction to any performance data is to imagine how to improve performance. Why isn't this more automated by now? This talk will share my experiences dealing with this problem in the arena of HPC I/O, including challenges in identifying trigger behaviors, selecting responses, and mechanisms for enacting responses. Thoughtful solutions to these problems have the potential for enormous positive impact in scientific computing.

6.2 Driving Response to Uncertainty in Job Duration and Loss of Work

Frédéric Suter (Oak Ridge National Laboratory – Knoxville, US, suterf@ornl.gov)

Uncertainty in job duration, caused by different input parameters or resource sharing can cause a loss of work if the resource reservation is not well dimensioned. To address this issue, several approaches can be followed at different times. Simulation can be used as a comprehensive twin of the job and help to improve the estimation of the job duration before its submission. Uncertainty-aware scheduling strategies can be exploited by the job and resource management system which is the only system component that has a global view of all the jobs and the resources. Finally, by running a job co-pilot that can access both the performance data exported by the application and information coming from the job scheduler, we can better react to unexpected events and prevent unwanted loss of work.

6.3 Hierarchical Control for Runtime Adaptation

Valeria Cardellini (University of Roma Tor Vergata, IT, cardellini@ing.uniroma2.it)

To tackle the fundamental challenges of performance and workload uncertainty, we can exploit a hierarchical control architecture to adapt at runtime the managed system or application. This control pattern avoids the scalability limitations of fully centralized controllers as well as the lack of coordination of fully decentralized schemes. Moreover, it allows for separation of concerns and time scales. As a case study, I present how to manage the horizontal auto-scaling of distributed applications deployed over heterogeneous computing infrastructures by means of heterogeneity-aware adaptation policies that run on a two-layered hierarchy of controllers. Application-level controllers steer the adaptation process for whole applications, aiming to guarantee user-specified Quality of Service requirements while easing the setting of control knobs that are exposed to the users. Lower-layer controllers take auto-scaling decisions for single application components using reinforcement learning techniques. To adopt the latter in an efficient and beneficial way, we can blend together models and learning to improve the learning velocity as well as function approximation techniques to reduce the memory requirements.

6.4 Laying the Foundation for Self-Organizing Systems

Ann Gentile (Sandia National Laboratories – Albuquerque, US, gentile@sandia.gov)

Main reference SAND2023-02343PE

Operational efficiency can be gained through data-driven operations, intelligent resource management, and resiliency handling, however the current data-gathering, communication, and response hooks do not exist! We propose that a fundamental redesign of system software, middleware, applications, and architectures and their interactions is needed to provide the hooks and all the components must be self-aware and self-organizing in order to make the right local and global decisions on the time-scales necessary for effective response.

7 Talks: Analytic and Informatics Approaches

7.1 Time Series Analysis for Performance Monitoring of HPC Systems

Abdullah Mueen (University of New Mexico – Albuquerque, US, mueen@unm.edu)

HPC performance monitoring tasks must respond back to the system and users in order to improve the overall performance, usability, resilience, robustness among many others. Data mining techniques can serve these needs by catering real-time knowledge extraction tools. In this talk, I show streaming data mining tools for pattern matching, clustering and anomaly detection that are data agnostic and applicable to performance monitoring data. I also discuss the potential response mechanisms that can be supported by offline knowledge extraction and online learning techniques with inputs from applications.

7.2 Missing Gaps and Opportunities in HPC Operational Data Analytics

Devesh Tiwari (Northeastern University – Boston, US, d.tiwari@northeastern.edu)

Traditionally, we have assumed that large-scale computing users are fairly boring and that their workloads often do similar things repetitively. But, now things are changing and changing fast. Our workloads and users are becoming interesting and, often, are surprising us with new trends and behavior. In this talk, I will discuss a few lessons I learned as we applied AI/ML methods to HPC resource management problems. In particular, I will discuss performance variability identification, performance auto-tuning, and open-sourcing a large HPC datasource.

7.3 Continuously Detecting Workflow Anomalies using Graph Neural Networks – Lessons Learnt

Krishnan Raghavan (Argonne National Laboratory – Lemont, US, kraghavan@anl.gov)

In this talk, I present the analysis challenges that we face when applying graph neural networks in detecting workflow anomalies. I begin this talk with an overview on graph neural networks and describe the need for them. I then discuss the challenges that are encountered when modeling the data from a workflow as graphs and illustrate the notion of imbalance and noise in this context. I then define how these challenges amplify when more and more data is collected from the system. Moreover, to correct this issue, I illustrate the need for continual learning in this problem and end the talk with the question, “how can we model and correct the challenges due to the drift in the distribution of the data?”

7.4 Best Practices on the Practical Use of Machine Learning in Hybrid Memory Management

Thaleia Doudali (IMDEA – Madrid, ES, thaleia.doudali@imdea.org)

In this talk, I will share some lessons learned on how we can build practical foundations when designing hybrid memory management systems that leverage machine learning methods to improve their effectiveness. Then, I will make a case on how the integration of visualization inside the systems software has potential to accelerate and improve resource management systems. I will end with a crazy idea on how changing data representations into image or text can unlock new opportunities for integrating cross-domain algorithms into systems solutions.

7.5 AI & Analytics in Service of Green Computing

Hilary Egan (NREL – Golden, US, hilary.egan@nrel.gov)

In this talk I present a series of case studies in attempting to use AI/Analytics to improve the sustainability in computing at NREL’s ESIF data center, and some future needs and directions for AI algorithms in this space. I begin with an case anomaly detection for advanced cooling systems study done in collaboration with HPE, showing examples on a blower door failure and scale build-up in a heat exchanger. Next I show attempts to optimize system PUE with facility set points and argue we need to integrate digital twin modeling.

I then argue a wide variety of grid-integration technologies will require sophisticated load shaping and present preliminary results in this space. I conclude with discussing future directions in AI including digital twins, modular AI, time-series foundation models.

8 Talks: Actionable Analytics and Response

8.1 AIOPS: Artificial Intelligence for IT Operations in HPC Systems

Jeff Hanson (Hewlett Packard Enterprise – Spring, US, jeff.hanson@hpe.com)
Torsten Wilde (Hewlett Packard Enterprise – Böblingen, DE, torsten.wilde@hpe.com)

Available data for HPC data centers are increasing in velocity and scale. Using traditional threshold based alerting mechanisms or visual inspection of operation graph data are limited in scope and resolution of detection. In this talk we present an implemented framework that uses ML/AI models to work on streaming data to detect anomalous behavior and reduce alarm fatigue for the system administrators. Enabling the research community to develop and implement control loops based on system information is not easy since system integrators provide their own management frameworks to keep the system in a defined and safe operating state. In the second part of this talk we will highlight the challenge using system power and energy management as an example. We will propose different approaches that would enable community developed management and control processes to work in synergy with system integrator provided mechanisms.

8.2 Energy Management on HPC Systems

Oriol Vidal (Barcelona Supercomputing Center, ES, oriol.vidal@bsc.es)

Full power management of large infrastructures requires an autonomous way to monitor system running jobs, as well as identifying those power saving opportunities. This talk will present how the EAR software fills ODA framework's gaps concerning HPC power and energy management, from basic monitoring and accounting to power saving strategies during application runtime and cluster-level power-cap.

9 Talks: Autonomous Loop Lightning Talks

9.1 Malleability, Monitoring and Modeling

Isaías A. Comprés Ureña (TU München – Garching, DE, compresu@in.tum.de)

The priority-based First-Come First-Serve (FCFS) with backfilling heuristic has been very successful in maximizing node utilization. However, since node utilization depends on system state and user provided jobs that are queued, it is not always possible to reach close to full node utilization. We propose malleability as a means to add additional opportunities for schedulers to reach near full node utilization. With malleability, nodes can be added or removed from jobs. Monitoring and modeling techniques are needed, to ensure quality decisions are made.

9.2 Control Loops in the PaPP Project

Sven Karlsson (Technical University of Denmark – Kongens Lyngby, DK, svea@dtu.dk)

Since the improvement of computer performance can no longer be satisfied by simply raising clock frequencies, architectures are evolving towards both the multiplication of processing elements and heterogeneity of their functions. The aforementioned execution platforms combined with steadily exacting non-functional performance requirements such as execution speed, timeliness, and power consumption challenges traditional design methods. The PaPP project included work on many aspects including improved resource management, adaptivity in all parts of the technology stack including the applications, and a control loop for managing non-functional performance requirements. The PaPP software infrastructure and the control loops were briefly introduced with demonstrations presented.

9.3 Job Efficiency Reporting

Thomas Jakobsche (University of Basel, CH, thomas.jakobsche@unibas.ch)

HPC systems provide high computing power to different domain scientists. Not all users possess the necessary knowledge and experience to fully utilize the capabilities of HPC systems. Certain users tend to submit inaccurate resource requests, leading to inefficient job scheduling, longer job wait times, and ultimately to wasted computing resources. HPC systems offer several tools to support users in order to improve their resource requests (e.g. SLURM's `seff` and `sacct` commands). However, these commands are missing easily digestible job statistics that can provide, for example, the longest execution time for a group of jobs. This lightning talk presented a command-line functionality based on the SLURM `sacct` command that introduces grouping and tabular-like functionality to job accounting data, that is easily available to administrators and users. This tool can support administrators and users to quickly identify groups of jobs with inaccurate resource requests, create job efficiency reports to raise awareness, and support users to improve future resource requests based on previous job submissions.

9.4 Code Path and Parameter Selection for Compute Loops

Thomas Gruber (Universität Erlangen-Nürnberg, DE, thomas.gruber@fau.de)

Applications for ODEs and PDEs often benefit from different coder variants but performance is highly dependent on the input. For testing these variants as well as hardware and runtime parameters different combinations need to be tested. By letting a genetic algorithm select testable parameter combinations, the amount of combinations can be reduced to a reasonable set. With measurements, the combinations can be graded and the best one selected for the given input. With refinements from run to run, the application can be executed with little disturbance and still find optimal execution settings. The user interacts with the library by supplying code variants through code instrumentation.

9.5 Feedback Loop for Mitigating Assignment of Slow OSTs for File I/O

Jim Brandt (Sandia National Laboratories – Livermore, and Albuquerque, US, brandt@sandia.gov)

This presentation was about an explicit use case at Sandia National Laboratories where processes of an application can be handed Lustre Object Store Targets (OSTs) with significantly poorer performance than other processes. The worst performance gates the application performance and in the presented case significantly increases the application run time (e.g., >50%). The mitigating approach being explored is creation of a feedback loop from the monitoring and analysis system to the application which could then take the action of deleting and re-creating the file it is writing to and ideally be provided with more performant OSTs.

10 Fundamental Results

Bringing together experts from different disciplines was needed to address the wide-ranging and interdependent aspects of the problem space of HPC Operations and enabled us to create a community and network around moving towards autonomous HPC system management and operation.

We discussed the fundamental requirements and gaps to enable autonomous feedback and response loops in production HPC systems. These are summarized, at a high level, below.

10.1 Assessment of the state of autonomous feedback and response loops in HPC operations

The community view was that progress on autonomous feedback and response loops has been limited. Individual participants have implemented feedback and response loops (either with or without human-in-the-loop) in order to improve energy efficiency; to improve performance of applications in the face of shared resources, such as networks and file systems; to diagnose performance and configuration issues, and to optimize competing goals of cost vs. performance. Many of these efforts addressed specific cases, specific architectures, and/or were designed to demonstrate potentials. While many of these have been successful in terms of their limited goals, they have not typically resulted in continuous production deployment to affect or improve operations. We believe that more *holistic* design and approaches to feedback and response are needed due to the interdependencies of systems, subsystems, and applications in order to fully benefit from autonomous feedback in HPC operations.

10.2 Assessment of most important challenges and opportunities

The community identified a number of major *challenges* that have impeded progress in the development of autonomous feedback and response loops in HPC operations. The HPC systems domain innately consists of complex architectures, the monitoring of which produces high-dimensional time-series data, making actionable analytics difficult. Effective feedback and response hooks are limited or require privileged access. It is difficult to build the required verification and trust necessary to deploy prototypes that alter yet improve system and application behavior on production systems. Without a clear path forward for development and deployment, many efforts have been independent, resulting in implementations that are not interoperable, are not interchangeable, are limited in scope, and are difficult to maintain.

There exists the *opportunity* to leverage the community's experience in feedback and response loops to develop infrastructure and standards for continuous loops to be deployed in production HPC systems that can both facilitate community development and affect production systems.

10.2.1 Conventions to Facilitate Autonomous Feedback and Response Loop Development

We identified the *opportunity* to leverage existing formalism, established in the fields of Self-Adaptive Systems and Autonomic Computing, i.e., the *Monitor, Analyze, Plan and Execute (MAPE), with shared Knowledge (MAPE-K)* control loop [Kephart and Chess, The Vision of Autonomic Computing, 2003], because of the similarity of concepts and interplay of *monitoring, analysis, feedback, and response* in our domain. Arbitrarily complex autonomous actions can be supported by different decentralized architectural design patterns comprising the MAPE-K components. A number of such patterns and heuristics for their application have been established. By leveraging the MAPE-K formalism, we hope to take advantage of the established design patterns and thereby facilitate the development of autonomous loops and the required loop infrastructure that would enable holistic feedback and response for HPC operations.

We discussed that several MAPE-K autonomy control loops can be defined in HPC systems, for example one in each subsystem. Each autonomy loop can be implemented as a centralized, hierarchical, or distributed loop, and can connect to other autonomy loops in the system.

Additionally, we identified the *challenge* of standardizing the interfaces between the different major control loop components, which would allow for easy interchange of components and for collaborative development. The community identified a potential *opportunity* for the development of conventions. This will be realized in a community technical path forward (Section 11.1) where development of the loop components, which include supporting analysis methodologies and decision-making logic, will be used to extract and define system software (e.g., APIs, and modular design) and data engineering (e.g., data formats and data requirements) conventions that will be adopted by the community.

10.2.2 Open data sets/challenges

An *opportunity* exists to build and leverage the external community via the application and development of analysis for our domain. The HPC systems community has substantial production telemetry datasets. While there exists some understanding of our architectures and applications, there are still substantial limitations in the description of our data and our understanding of how that data reflects performance issues and headroom opportunities. This has complicated the application of analytics in our domain. Statistical and Machine Learning algorithms may be suited to overcome these *challenges*. However, the lack of expertise in these techniques in the systems community has limited our ability to make analytic progress on our own. By building a community that includes not only HPC experts but also analysis researchers, further progress may be made.

Such collaboration and progress will be non-trivial. The processes of producing data and analyzing data are not independent. Significant exchange of knowledge will be required to ensure uniform understanding of data and expectations of metrics and features of importance. To be of use, telemetry data should include semantic labels so that not just names, but also understanding of meaning, is conveyed. Datasets need to include events of interest and with frequencies and details that meet the requirements of algorithmic application (e.g., evidence of sufficient rare events, unanticipated bias features in the data). Additionally, a domain expert's identification of expected features of significance and what they functionally relate to in the operation of an HPC system would be of immense benefit to consumers of this data. System state (idle vs. busy with well-known benchmarks/proxies/canaries running) and changes in state should be provided over the time window of the supplied data set.

The participants explored that an *opportunity* to facilitate this could be through the development of an open dataset and challenge where we would test the readiness of and refine the requirements for the monitoring data produced and collected on the systems to be suitable for analysis by AI/ML/statistical methods.

The participants also agreed that the community would benefit from a survey of existing data analysis methods and data sets with the properties described above.

11 Planned Paths Forward

At the seminar, we laid the groundwork for continuing progress as a community on three fronts: technical, interaction, and funding. These are discussed in this section.

11.1 Identification of a community collaborative technical path forward

To drive the community along a collaborative technical path forward, we discussed various use cases that are commonly encountered across HPC systems and sites. Three use cases were identified as representative of burning problems for which the community could develop proof-of-concept solutions. The development process would enable us to explore options for architectural design patterns and interfaces and their extensibility to support multiple cases. Successful development could then drive technical change in both loop development and production adoption, form a foundation for a community vision report, and serve as the basis for future funding proposals.

These use cases are:

1. Interaction of an HPC monitoring and analysis system with the scheduler on behalf of running applications in order to ensure that a reasonably progressing application is not terminated prematurely. The autonomous loop would involve continuous monitoring and analysis to project an application's execution time, which may vary due to system conditions, competing workload, or problem specifics, and feedback to extend the allowed running time. This use case is representative of instances where feedback would need to be directed to system software and middleware.
2. Interaction of an HPC monitoring and analysis system with applications in a feedback loop to inform a target application of a problem for which it could take mitigating action. This use case is related to the above case, but is representative of instances where feedback is directed to applications.
3. Detecting performance degrading misconfiguration of an application deployed on HPC resources. The autonomous loop would involve performance assessment and attribution of existing configurations to provide feedback to humans and applications of potential misconfigurations. Distinctions in this use case, as opposed to the above two, are the source of the condition of interest and the feedback direction and target.

Participants have agreed to further define the scope and parameters of these use cases and devote resources to collaboratively develop proof-of-concept-grade solutions. These solutions will inform further understanding of how to generalize approaches to solving the classes of problems represented and to extract commonalities that we will use to define conventions for development. Sites with HPC resources that would be open to deploying exploratory autonomous feedback and response loops for controlling operations and that could potentially host external developers were identified. These include potential large-scale, end-of-life, systems. These sites will be looking further into details for the necessary arrangements.

Note that changing the HPC operational model to incorporate pervasive autonomy loops is not merely a software development and data engineering challenge. Significant work will remain to explore issues that need to be overcome for system administrator and site acceptance and adoption. These include identifying and granting authority for control, establishing scope of decisions made, avoiding competing and thrashing responses, and quantifying impact and ensuring response on meaningful timescales.

11.2 Identification of community interaction path forward

Several opportunities to continue interaction and collaboration within the community and make technical progress on improving the performance of HPC applications and efficiency and sustainability of HPC systems as a community have been identified and agreed upon:

- Setting up a regular virtual meeting to continue the discussions started at Dagstuhl. Progress on the proof-of-concept autonomous control will also be coordinated at these meetings.
- Further meetings at other community-relevant events (ISC/MODA, SC/Quantitative Supercomputing Codesign, IEEE Cluster/HPCMASPA conferences, and EEHPCWG Workshop) are envisioned to drive further discussions, finalize, and disseminate the community vision report.
- A proposal for a Birds-of-a-Feather (BoF) session at SC'23 and/or a position paper to a workshop or symposium will be submitted to widen community awareness.
- The organization of a Special Issue at a leading journal will also be explored as a way to expand technical and community knowledge of the field.

11.3 Identification of community collaborative funding path forward

The community is spread all over the world, with most participants of the seminar being either US- or Europe-based. As there is interest to collaborate further on the topics discussed during the seminar, many attendees showed interest in submitting a joint proposal to fund further activities to drive the development of autonomous feedback and response loops for HPC operations forward. However, as a joint EU/US funding call that allows for project partners from both continents seems out of reach, the attendees agreed to continue with separate proposals in the US and EU that would support each other and include close collaboration between the two. Additionally, other funding instruments that would allow for bilateral collaboration have been identified.

12 Acknowledgments

The organizers would thank Schloss Dagstuhl and their staff for hosting and making this seminar possible.

Some of the contributors to this research were supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration.

A portion of this work was supported by the Scientific Discovery through Advanced Computing (SciDAC) program funded by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) under contract DE-SC0021299.

Some of the participants received funding from the "Adaptive multitier intelligent data manager for Exascale (ADMIRE)" project, funded by the European Union's Horizon 2020 JTI-EuroHPC Research and Innovation Programme (grant 956748).

Some of the participants received funding from the "DAPHNE: Integrated Data Analysis Pipelines for Large-Scale Data Management, HPC, and Machine Learning" project, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 957407.

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Participants

- Francieli Boito
INRIA – Bordeaux, FR
- Jim Brandt
Sandia National Labs – Albuquerque, US
- Valeria Cardellini
University of Rome "Tor Vergata", IT
- Philip Carns
Argonne National Laboratory, US
- Florina M. Ciorba
Universität Basel, CH
- Isaías Alberto Comprés Ureña
TU München – Garching, DE
- Thaleia Dimitra Doudali
IMDEA Software Institute – Madrid, ES
- Hilary Egan
NREL – Golden, US
- Ahmed Eleliemy
Universität Basel, CH
- Ann Gentile
Sandia National Labs – Albuquerque, US
- Taylor Groves
Lawrence Berkeley National Laboratory, US
- Thomas Gruber
Universität Erlangen-Nürnberg, DE
- Jeff Hanson
HPE – Lakewood, US
- Utz-Uwe Haus
HPE HPC/AI EMEA Research Lab – Wallisellen, CH
- Esa Heiskanen
CSC Ltd. – Kajaani, FI
- Kevin A Huck
University of Oregon – Eugene, US
- Thomas Ilsche
TU Dresden, DE
- Thomas Jakobsche
Universität Basel, CH
- Terry Jones
Oak Ridge National Laboratory, US
- Sven Karlsson
Technical University of Denmark – Lyngby, DK
- Allen D. Malony
University of Oregon – Eugene, US
- Henrique Mendonça
CSCS – Lugano, CH
- Abdullah Mueen
University of New Mexico, US
- Michael Ott
LRZ – München, DE
- Tapasya Patki
LLNL – Livermore, US
- Ivy Bo Peng
KTH Royal Institute of Technology – Stockholm, SE
- Krishnan Raghavan
Argonne National Laboratory, US
- David Schibeci
Pawsey Supercomputing Centre – Kensington, AU
- Kathleen Shoga
LLNL – Livermore, US
- Michael Showerman
University of Illinois at Urbana-Champaign, US
- Frédéric Suter
Oak Ridge National Laboratory, US
- Oriol Vidal
Barcelona Supercomputing Center, ES
- Torsten Wilde
HPE- Böblingen, DE
- Keiji Yamamoto
RIKEN – Hyogo, JP

Remote Participants

- Devesh Tiwari
Northeastern University –
Boston, US