SCALABLE OBSERVATION, ANALYSIS, AND TUNING FOR PARALLEL PORTABILITY IN HPC

by

CHAD WOOD

A DISSERTATION

Presented to the Department of Computer and Information Science
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

March 2022

DISSERTATION APPROVAL PAGE

Student: Chad Wood

Title: Scalable Observation, Analysis, and Tuning for Parallel Portability in HPC

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Computer and Information Science by:

Allen Malony	Chair
Hank Childs	Core Member
Boyana Norris	Core Member
Stephanie Majewski	Institutional Representative

and

Krista Chronister	Vice Provost for Graduate Studies

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded March 2022

DISSERTATION ABSTRACT

Chad Wood

Doctor of Philosophy

Department of Computer and Information Science

March 2022

Title: Scalable Observation, Analysis, and Tuning for Parallel Portability in HPC

It is desirable for general productivity that high-performance computing applications be portable to new architectures, or can be optimized for new workflows and input types, without the need for costly code interventions or algorithmic re-writes. Parallel portability programming models provide the potential for high performance and productivity, however they come with a multitude of runtime parameters that can have significant impact on execution performance. Selecting the optimal set of parameters, so that HPC applications perform well in different system environments and on different input data sets, is not trivial.

This dissertation maps out a vision for addressing this parallel portability challenge, and then demonstrates this plan through an effective combination of observability, analysis, and in situ machine learning techniques. A platform for general-purpose observation in HPC contexts is investigated, along with support for its use in human-in-the-loop performance understanding and analysis. The dissertation culminates in a demonstration of lessons learned in order to provide automated tuning of HPC applications utilizing parallel portability frameworks. This dissertation includes previously published and co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Chad Wood

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
Texas Christian University, Fort Worth, TX, USA

DEGREES AWARDED:

Master of Science, Computer and Information Science, 2018, University of Oregon
Bachelor of Arts, Philosophy, 2014, Texas Christian University

AREAS OF SPECIAL INTEREST:

High-Performance Computing, Optimization, Observability, Online Tuning, Systems Software, Programming Tools, Scientific Workflows, Event Systems, Extreme-Scale Software Engineering, In Situ (Online) Programming Models

PROFESSIONAL EXPERIENCE:

Graduate Employee (2021): Teaching Assistant for three lab sections of "Computer Organization," University of Oregon, Eugene, OR
Graduate Employee (2020-2021): Teaching Assistant for two sections of "C/C++ and Unix Systems," University of Oregon, Eugene, OR
Research Internship (2020): Designed next generation of the Apollo distributed tuning framework using in situ (online) techniques to automatically discover and apply optimized tuning choices at runtime. Developed initial implementation to enhance OpenMP and CUDA performance for applications using the RAJA and Kokkos performance portability frameworks. Lawrence Livermore National Laboratory, Livermore, CA

Research Internship (2017-2018): Integrated online monitoring with Ascent and Conduit visualization pipeline for understanding complex scientific workflow performance data via projection over simulation geometry at runtime. Lawrence Livermore National Laboratory, Livermore, CA

Research Internship (2016): Assisted with the design of ScrubJay semantic annotation and query framework to enable deep learning on performance data, recursively derived from sources with independent units and epochs. Lawrence Livermore National Laboratory, Livermore, CA

Teaching Assistant (2012-2014): Computer Science Department Teaching Assistant for five sections of "Introduction to Computer Science." Texas Christian University, Fort Worth, TX

Military Service (2005-2010): Promoted to Non-Commissioned Officer rank of Sergeant in Delta Co 2nd Battalion, 75th Ranger Regiment. Engaged in three deployments to combat in support of Operation Iraqi Freedom, participating in hundreds of airborne, helicopter, and ground assault missions, concluding with an Honorable discharge from the Armed Forces of the United States. 2/75 Ranger Regiment, Fort Lewis, WA

Vice President of Technology (2004-2005): VP of Technology and lead developer of integrated operations, billing, auditing, and compliance software. Responsible to understand and enforce compliance with local, state, and federal legal requirements; vendor and product-specific policies; industry best-practices; and service quality, business ethics, and patient privacy standards. Developed a fully-integrated range of custom business intelligence modules to support operations and accounting tailored to payors including Medicare, Florida Medicaid, numerous HMO and PPO carriers, and direct fee-for-service patients. All-Med Services, Inc. Miami, FL

Developer, Systems Analyst (2003-2004): Developed medical equipment provider management software covering areas of patient intake; clinical assessment; billing and receivables including electronic claims files; ANSI, JCAHO, and HIPPA compliance. Omni-Medical Management Systems, Inc., Richmond, VA

GRANTS, AWARDS AND HONORS:

University of Oregon: Voted Best TA by the Computer and Information Science Department for the 2020-2021 academic year.

Texas Christian University: Chancellor's Leadership Program, Katheryne McDorman Honors Scholar, Dean's Scholarship, Wetzler Award for Outstanding Achievement in Philosophy

United States Army: Army Commendation Medal (2nd Award), Army Achievement Medal (2nd Award), Army Good Conduct Medal, National Defense Service Medal, Global War on Terrorism Service Medal, Iraq Campaign Medal w/Campaign Star, Army Service Ribbon, Parachustist Badge, Combat Infantryman Badge, Expert Infantryman Badge, granted and held TS/SCI clearance in good standing for the duration of need.

PUBLICATIONS:

Wood, C., Georgakoudis, G., Beckingsale, D., Poliakoff, D., Gimenez, A., Huck, K., Malony, A., & Gamblin, T. (2021). Artemis: Automatic Runtime Tuning of Parallel Execution Parameters Using Machine Learning. *In International Conference on High Performance Computing* (ISC21) (pp. 453-472). Springer.

Wood, C., Larsen, M., Gimenez, A., Huck, K., Harrison, C., Gamblin, T., & Malony, A. (2017). Projecting Performance Data Over Simulation Geometry Using SOSflow and ALPINE. *In Programming and Performance Visualization Tools* (VPA17) (pp. 201-218). Springer.

Wood, C., Sane, S., Ellsworth, D., Gimenez, A., Huck, K., Gamblin, T., & Malony, A. (2016). A Scalable Observation System for Introspection and In Situ Analytics. *In 2016 5th workshop on extreme-scale programming tools* (ESPT16) (pp. 42-49). IEEE.

ACKNOWLEDGEMENTS

We will encounter adversity many times during this life. Large challenges, small ones. Some slow, quiet, or lonely difficulties; other times we find ourselves overwhelmed in situations urgent, embarrassing, or scary. How we individually respond to these experiences is important, but what so often will be decisive is how our communities, familial, professional, and social, respond to us encountering that adversity. I am grateful for the many people in each of my communities who have been there for me over the years. You are too many to name, but I can say with sincerity that the best parts of me were put in place and later protected by those who I do mention here, and also by others who I haven't. I've not forgotten, and I am very grateful.

If you are going through a difficult time, remember that no feelings are final, and better days will come. Your circumstances will change, as will your thoughts, and you will not always feel the way you do in the middle of some troubles. Your life is yet undiscovered, and worth experiencing to the fullest. You have a bright future. I hope the reader might make some time after disengaging this text to seek out and remind a few good people from your communities how much they matter. Sharing a kind word and an ember of courage with a friend may mean more to you both than you think.

I am grateful for my mother and father for the gift of life, for nurturing, empowering, correcting, and lifting me up, as I've so often needed them to do. I grew up with the blessing of my mother's presence, her delightful intelligence so generously shared, her love, wisdom, grace, and wit. There has been no finer role model for being a man than my father, who has been a poet, a scholar, a successful

entrepreneur, a clear-eyed and faithful visionary, whose great force of character is paired with a matching tenderness. They both operate with complete confidence in the deep enduring value of others around them. I am inspired by their zeal for life and the joy they find in their work. My brothers are just as astonishing and awesome, each of them. Mark, John, and Karl: I love you endlessly.

Allen Malony, my research supervisor, has done more than nearly anyone to change my life for the better. Everything I have achieved here I owe to his outstanding leadership and support, and his long-suffering faith in me. Over these several years working together, he has never once doubted me or given up on me, even during my lowest moments. Hank Childs and Boyana Norris also have provided so much needed encouragement and good counsel over the years. It feels special to be able to say, "I could not have done this without my committee."

I owe a world of gratitude to the legendary Todd Gamblin, for his generosity, encouragement, patience, understanding, research opportunities, Python advice, and all the excellent people he brought along to befriend and teach me, who have done so much to elevate my work.

For all their inspiration and effort, I must thank a few more of my academic friends and collaborators: Kevin Huck, Giorgis Georgakoudis, Matt Larsen, Stephanie Brink, David Poliakoff, Alfredo Gimenez, David Beckingsale, Daniel Ellsworth, Sudhanshu Sane, Jacob Lambert, and so many more.

Deep thanks to my Oregon crew: Mason, Wade, Alex, Tyler, Hugh, Jake, Audrey, Chris, Josh, Jason, Joe, and of course Cooper and Annie. Much love to Blake cat, the sweetest little lad, who is napping on my desk next to the keyboard as I type this; and to babboo, from the beginning.

To my mother, for teaching me to read.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

GENERAL INTRODUCTION

1.1 Investigatory Approach

What ultimately motivates this work is the desire to enable and improve *parallel portability* for HPC software. By increasing the observability of systems, providing mechanisms for information collection, sharing, analysis, and online feedback, and ultimately by embedding machine learning into this infrastructure, we have provided both a plan for achieving this goal generally, and an example of this plan in action with concrete real-world success.

This dissertation is divided thematically into two halves, the first two chapters covering the topic of *scalable observation* in depth, and the third and fourth chapters investigating support for *human analysis of codes at scale, and online automated performance tuning*, respectively.

These four chapters provide mutually-reinforcing coverage of relevant history and of our own studies investigating these four central research questions:

- **RQ1:** What are the essential components of a practical in situ system for online observation, analysis, and feedback?
- **RQ2:** Can online observation with in situ methods provide benefits to application users and developers?
- **RQ3:** Is it feasible to conduct machine learning in situ in order to derive performance benefits without a human in the loop?
- **RQ4:** Can systems be made both observable and responsive to tuning choices without costly code interventions or algorithm rewrites?

Considering the basic preconditions for achieving parallel portability through performance understanding and online adaptation, it became clear that any capable system would need to have regular access to current and contextualized performance information in order to make intelligent decisions and observe the effects of those choices. We would need to start with the topic of observability.

We began by asking a pair of hypothetical questions: *"If we were able to efficiently and quickly observe any bit of information in a system, what kind of decisions would we be able to make? Would this capability change the way we designed HPC software?"* Imagining some of the possible answers, a more basic question stood in the way, *"Has this capability existed before, and if so, what were the limitations?"*

The journey to answering our research questions must then begin with a deep dive into observability, the nature of the information that is to be observed, and what design features and trade/offs that have emerged through the history of online monitoring for HPC.

1.2 Preliminaries

The general theme of this dissertation is that of gaining insights that facilitate greater *productivity* in a high-performance computing context. This is *why* we are interested in online monitoring, analysis, and feedback systems. We will be considering both low-level and high-level aspects of insight and productivity. Note that these terms are intentionally used loosely and relatively in this document, merely to lend a rough sense of scope. The term *low-level* is taken to mean closer to the machine or software engineering. We use *high-level* to indicate something is closer to the application behavior, the science purpose of an application, or the goals of human users or managers.

Insights might be gained by investigating something as low-level as hardware counters and source code performance hotspots, or as high-level as application data dependency graphs, simulation state, or human-in-the-loop evaluation of scientific visualizations. Productivity also refers to a plurality of possible goals. It can indicate low-level enhancements to the use of network resources, application runtimes, communication slack, power utilization, machine temperatures, etc. Productivity can just as well mean making improvements to the correctness of scientific results, the quality and timeliness of reports and images, the speed at which software can be developed and debugged, or the portability of source codes and optimizations between various machines.

High-performance computing (**HPC**) refers to a specialized branch of computing traditionally used to tackle problems too large to be effectively solved using commodity computational resources. HPC architectures often couple powerful integrated compute nodes together using a high-speed interconnect. The HPC systems we are concerned with in our area of research almost exclusively run a variant of the POSIX-compliant Linux operating systems. The operating system of each compute node runs various services and specialized hardware drivers that allow applications to take advantage the resources which are distributed across several nodes. Such services usually include:

- Networking and shared memory region APIs: Allows for applications, libraries, and services to communicate with each other. This communication can take place within a single physical resources, or between processes running on different devices.

- Batch Manager (ex. *IBM®Job Step Manager* [1] or the *Slurm Workload Manager* [2]): Queues, allocates, launches, and manages user's jobs in a

batch scheduling environment, breaking apart a parallel task into ranks and establishing a shared runtime environment that may span one or more nodes.

- Network Filesystem (ex. *Lustre* [3] or *IBM®'s General Parallel Filesystem* [4]): Provides a coherent filesystem view across many nodes in parallel, where reads and write to the filesystem from multiple ranks are eventually synchronized and available to all nodes within a parallel job's allocation.

- Message Passing Interface (*MPI* [5]): Allows ranks of applications to communicate amongst themselves and coordinate their activity via point-to-point messaging and safe synchronous collective data operations.

We will look at how these common HPC software resources, among others, can be exploited for our monitoring, analysis, and feedback purposes.

This work explores an intersection of three different topics: *monitoring*, *analysis*, and *feedback*. We use the term feedback to imply interacting with applications and execution environments, based on analysis of monitored information, potentially within the same job being monitored.

Modern HPC has introduced extreme scale parallelism, large and complex codes, interactivity between coupled software components, and an unprecedented velocity of data creation, consumption, and displacement. The introduction of these changes has given rise to new computational models, performance paradigms, design challenges, and research possibilities. These recent developments in HPC have both *created new roles for* and also *expanded the prior roles* of monitoring, analysis and feedback.

It is important to understanding the structure of this effort that we are ultimately building towards the current state of the art where these three topics are able (and desired) to be integrated. At times we will discuss monitoring, analysis, or feedback as a standalone topic, but will attempt to explain why the coverage is only partial in a those specific moments as a way of giving insight into the computing landscape at the time of that prior work. Always bear in mind that we are building to what exists in the present features of HPC research in this area.

1.3 Observability

Before something can be monitored, analyzed, or utilized online, it needs to be *observable*.

Computation involves applying operations to a set of data inputs in order to transform that data according to those operation's stable rules, resulting in reliable and reproducible output. The output produced by a piece of software can be used to validate limited but crucial properties of that software, such as its mathematical precision or the correctness of the computed results compared to a trusted independent measure.

But what of the behavior of the software itself?

By the time an application has completed its work and generated its output, information about the execution of an application that is not observed and stored is lost. Observations such as the basic behavior of the software and the efficiency of its algorithms, and the interactions of its internal components and external execution environment. Information relevant to the performance of an application may include a variety of data sources, both within and external to the application. In order to make informed decisions regarding the behavior of an application, this behavior needs to be observed, annotated, and stored for later use.

Observability is a critical first step into online monitoring, but it is worth noting that points where something is made available for monitoring are often also points were feedback from analysis can be applied.

The depth and significance of observation will vary based on the method, completeness, and invasiveness of the techniques employed. Observability can be achieved or enhanced by a variety of techniques, principally including:

- Application Source Instrumentation
- Shared Library, Runtime, or Service Instrumentation
- Sampling and Tracing
- Probing and Inferences from Indirect Sources

We will now discuss each of the preceding techniques in turn.

1.3.1 Application Source Instrumentation.

Instructions to capture observations, compiled directly into the executable code of an application.

Software source code can be instrumented to self-report its progress from state to state. Such instrumentation takes the form of function calls (or macros) that are embedded in-line between normal application code. Once instrumentation is in place, it is encountered and evaluated during the normal course of application execution. This placement can be done by hand, embedding direct calls to some annotation API, or it can be done programmatically by an automatic code instrumentation tool. Hand-instrumentation is more invasive and labor-intensive than tool-based instrumentation, requiring developer time and expertise. In exchange for the extra work to emplace and maintain it, there are some added

benefits to using hand-instrumentation over tool-based solutions. By selectively instrumenting specific code regions, a developer can minimize the cost of observing a piece of software. Since no code can be observed without the computer doing a little bit of extra work, doing too much of this extra work will mask off the underlying application behaviors of interest. A developer can use their judgement to skip the observation of areas that are not of interest, or that are executed so frequently that the overhead of making observations would dominate any application performance that could be observed.

Hand-instrumentation is also able to introduce *high-level annotations* to the observed low-level execution features. High-level annotations are essentially labels which identify the nature or purpose of the region of code being observed. They allow for that data to be quickly individuated from other observations, to allow for efficient categorization and analysis (to be discussed in later sections). Developers do not always know what regions of their code are important, or the code that is having the most significant impacts on the applications' behavior will change as the codebase evolves or new inputs are fed into the program. To remedy this, it is useful to have a variety of mechanisms available to observe and explore the performance of an application. Common instrumentation interfaces [6] being embedded in codes show promise in this regard. They provide a point at which many different performance tools can be attached and activated to provide observation of the software, without needing to edit code or recompile applications. When not in use, these instrumentation interfaces would not impose any significant overhead. It will be interesting to see whether this idea gains broad support going forward.

Because instrumentation involves inserting extra instructions into code, regardless of the kind of instrumentation that is in place, it is sometimes desirable to temporarily disable it. Instrumentation is typically disabled when code moves from being actively developed and optimized into a "production" scenario where maximum efficiency is desired and introspection of application behavior is less important. To this end, it is important for instrumentation to have an "off switch" of some kind. One option is to excise the instrumentation from the application code at compile time, so that the source code remains instrumented, but those blocks of instructions are skipped over by the compiler and do not appear in the application binary at all. Recompilation can be costly, but will yield the most efficient application binary. Another option is to disable code with a setting that can be checked by a program in execution. This means leaving the instrumentation in the code, but while the code is running, whenever some instrumentation is encountered, first have it check to see if it has been disabled, and then if so skip over the block of instrumentation code and resume normal execution. With this method, it is important to be able to "do nothing, quickly", so that the impact of the extant (disabled) instrumentation is minimized. This would be an appropriate method to use if the instrumentation were not directly embedded in an application code, but emplaced in the code of a shared library or service that an application makes use of. When the application is recompiled, that library or service might not be, so runtime enabling or disabling of instrumentation is required if a system is to be observable in this way. We will now discuss that scenario.

1.3.2 Shared Library, Runtime, or Service Instrumentation.

Making observations within the code which is executed when an application makes API calls to an external library or service, or when

a runtime platform is evaluating collections of instructions (code or queries).

HPC software is often built up out of multiple libraries being interacted with by the core logic of an application. When an application is launched by the operating system, any libraries that it has been linked into will also be loaded into memory and initialized. `C` and `C++`, for example, offer standard libraries which provide many features essential to applications written in those languages, from collections of optimized data structures to network and multithreading routines. Higher-level libraries exist to provide domain-specific features, such as *SAMRAI: Structured Adaptive Mesh Refinement Application Infrastructure* [7], an implementation of optimized data structures and algorithms of general utility for adaptive mesh refinement codes, common in physics simulations. In the absence of direct application source code instrumentation, shared libraries can be good targets for observing application behavior. Through calls into the API of that library, execution will pass into the code compiled within it. Any instrumentation within will then be executed.

As will be discussed in later sections, especially "Exposing Data" (§ 1.4.4), the information that is generated or observed by instrumentation needs to either be made available for use, or stored to be used elsewhere or at another time. Control flow through the execution of a program binary is typically fixed at compile time, where the operating system will establish the basic execution environment and protected memory, initialize the stack, and begin execution at the designated starting function of a program, `main` in the `C` family of languages, for example. This process does not automatically provide hooks for tools to be initialized or optional accessory services to be started. In the case that memory needs to

be allocated for storage, or services need to be invoked that can capture and operate on monitored information, the shared library instrumentation path offers another useful engineering options: *static singletons*. This refers to a `C++` language convention where code objects can be can be marked as `static` and be executed at program initialization, and through clever means cause the initialization of a class that uses the *singleton* design pattern, where only one instance of an object is allowed to exist. This combination of techniques allows a shared library to execute some initialization routines at the beginning of a program, merely by being linked into the program and loaded when that program starts.

Two common ways to instrument libraries by adding code are to pre-load a surrogate (or *wrapper*) library, or provide customized header files that implement some instrumentation. Wrapper library instrumentation can also be achieved without needing to recompile an application. `LD_PRELOAD` is a special environment variable supported by the Linux operating system. When paths to shared libraries are set into that variable, those libraries will also be loaded by the operating system when an application is launched. This can be used to to flexibly provide instrumentation around existing libraries without needing to access their source or recompile them. The wrapper library should expose all of the same function signatures to the invoking application, such that normal API calls to the shared library will instead invoke the same function in the wrapper. On its first invocation, the wrapper can then manually load the normal shared library and populate a table of function pointers all of the normally-exposed functions within. As the normal library's functions are called and return back, the wrapper is able to track these timings and perform any other desired instrumentation or monitoring-layer interactions desired.

If recompilation of the application is not a burden, customized header files are also an option, and can impose a slightly lower performance impact as less runtime activity is required to resolve API calls. Customized header files will require the calling application to be recompiled, and its source code or build scripts updated to point to that custom header file. Within the header file, functions can be implemented instead of merely defined, and these functions can embed instrumentation around calls to the normal library. The header file technique usually requires an API to be expressed in two layers, with a public-facing API wrapping calls to an internal implementation API. This avoids the problem of *namespace collision*, what occurs when an object has two different definitions within the same callable scope, and the compiler does not know which of the two is being referenced as it attempts to build or link the software.

A successful example of header file instrumentation in the real world is provided by the MPI codebase. MPI's public API calls all begin with `MPI_` and every such call jumps into a tiny wrapper function that immediately and only calls its implementation function, which is prefixed `PMPI_` and which provides the actual implementation code. Developers can add code to the wrapper functions in that header file, as a way of intercepting calls to the MPI routines.

Both of these technique are able to facilitate a variety of advanced interactivity, such as making adjustments to the parameters being passed through the wrapper into the normal library, or changing the behavior of the normal library based on some performance observations or goals.

1.3.3 Runtimes and Services. Many HPC applications take advantage of standardized libraries and packages designed to grant traditionally-

engineered software access to the unique advantages enabled by HPC hardware, without requiring wholesale re-writes.

One such library is OpenMP [8], which presents a standard for annotating, or "decorating", the parallel regions of a block of code, and then compiler extensions which can intelligently and safely adapt the code according to those notes for it to be automatically parallelized. The primary mechanism for parallelizing codes that OpenMP uses is the spawning of multiple threads, distribution of data between those threads, and the gathering of the results produced in parallel back into a unified memory location for processing by the serial portions of the program. In addition to the injection of inline codes, OpenMP provided a flyweight runtime within the process, to manage the creation and destruction of threads, or teams of threads, achieving safe management of the memory regions those threads were operating over.

Automatic code generation, especially in this case where it profoundly altered the characteristics of the code's execution, introduced some complexity to the various source-instrumentation-based means of observing codes, though techniques were developed [9] to address this. Iterating over the years, as more modern generations of the OpenMP frameworks were designed, a tools interface named OMPT [10] was added to OpenMP to provide an organized and flexible means of interacting with the OpenMP runtime and observing the application, providing hooks into the normal semantics of the program as well as events unique to the internal activity of the OpenMP runtime. One especially useful feature of the OMPT interface is that tools can be enabled or disabled at runtime, not requiring an application to be recompiled from source. For the many HPC

applications which make use of OpenMP, this interface can be a useful source of information for online monitoring frameworks.

Elaborated further under "Distributed Computing" (§ 1.6.1.1), the Message Passing Interface (MPI) runtime can be an indispensible resource when monitoring HPC applications. In addition to posessing a number of valuable datum related to the execution of a single distributed task, the runtime is also potentially managing many other processes distributed across the machine concurrently. MPI is aware not only of some performance measures of the application, but of its own configuration and performance. With higher-level permissions and a common observational infrastructure, it is possible to observe complex interactions between parallel jobs of parallel processes in situ and online [11], observations which by necessity require shared service-level instrumentation and online monitoring.

Task-based runtimes or applications written using *partitioned global address space* (PGAS) languages [12] [13], like HPX [14] or Charm++ [15], or even distributed workflow managers such as Swift/T [16], can make it difficult to cleanly separate out the workings of the runtime service layer from the program that the service layer is facilitating the execution of. That is, a program can be broken up into so many different asynchronous parts that traditional monitoring patterns do not effectively capture coherent or developer-relevant performance data. The ratio of monitoring overhead to the overall productive work performed by the application can quickly become undesirable, especially if tasks are dispatched and retired at a very fine-grain, and have short lifespans. Tools such as the Autonomic Performance Environment for Exascale (APEX) [17] have been developed specifically to address [18] many of these challenges, but it remains an open area of research.

Another way to observe processes in vivo is by stepping outside of their execution environment entirely, and then turning around to look back in. This is most often observed in cases of commercial "cloud computing", where a system image is hosted by a virtual machine hypervisor, and applications are run within that virtualized environment. Through extensions to the hypervisor agent, such as with the ongoing work with Xen introspection extensions [19], the performance, progress, or various other information can be observed from outside the runtime instance with only relatively small increases in overhead compared to running unobserved within the virtual machine. These increases in overhead would be proportional to the overhead of monitoring the same application running natively on physical hardware.

1.3.4 Sampling and Tracing.

Exploiting binary formats, memory layout conventions, and explicit operating system APIs to apply instrumentation to a compiled application without modifying its source code.

Sampling means inspecting the state of an application and reading the available performance counters provided by the operating system. It is usually performed at some regular interval of time, so inferences can be made about the activity that transpired between those intervals, and the impact those activities had on the sampled parameters. Sampling is by far the most efficient method for gathering observations to use when monitoring an HPC system, and as such is favored for online monitoring systems. Sampling can be done directly by a tool, by making calls to the Linux operating system's `perf_events` API, reading counters from the `/proc/stats` virtual filesystem furnished by the operating system kernel, or

through registering counters of interest and making inquiry into a pre-packaged introspection tool like PAPI [20].

Tracing deviates from sampling in that *every single action* an application takes has the opportunity to be counted as a significant event and measured, though each action might not be of interest. In order to achieve the extremely fine-grained analysis afforded by tracing, many additional instructions are inserted around those of the application, able to capture and count the application's instructions and follow the control flow's branching paths through the application logic based on the input data and the evolving results of computation at runtime.

Traces often have orders of magnitude higher overhead to gather than performance measures arrived at through sampling. Hand-annotated source code has the added benefit (and developer overhead) of an expert identifying the significant regions of an application, so that uninteresting information does not need to be collected or analyzed. This lowers the overhead of performing a trace, in both time spent gathering measurements, and by reducing the space required to store any performance measurements. It also allows for code regions to be intelligently named for quick identification, for cases where a person is utilizing a monitoring system, and such insights can be exploited for code tuning, etc.

Tracing can be performed over specific domains of application events, such as tracing only the I/O of an application, the loading and storing of regions of system memory, or just generally serarching for latency [21]. One could choose to trace only the interactions between an application and the operating system kernel, or even to trace only the activity within the kernel.

While outside the scope of this survey, note that there are many types and implementations of tracers [22] available to trace both kernel and user-space

activity, including ftrace [23] [24], perf [25], LTTng [26] [27], and some commercial offerings such as Intel PT ("Processor Trace"), etc. Much of the lower level tracing infrastructure, such as the Linux `perf_events` subsystem, is available to be used in other user-space tools like Valgrind [28] or PAPI [20] to provide aspects of their overall performance information set, including data associated with branches and traces.

1.3.5 Probing and Inference from Indirect Sources.

Combining multiple external sources and epochs of information to form intuitions about the behavior of a system and its components.

There are a variety of questions an interested party may wish for their monitoring system to answer that, while requiring online monitoring, are not well-suited to the mechanics, scope, or frequency of events which are revealed by directly observing a single application, or even a single instance of a complex workflow, as the source of information. Some examples:

- On average, how long are jobs waiting in dispatch queues before being launched, including as ratios of their actual and requested runtime?
- What portion of a job's occupancy is spent waiting on shared resources to become available (i.e. physical tape archives of large data sets that need to be fetched and brought online by an automated robot)?
- How much do the power requirements of the entire facility deviate through the day, and is there a correlation with specific jobs, or machine workloads, or the exterior environment's temperature and humidity?
- How often do the processors on the nodes slow their clock rate in order to stay within their configured thermal envelope?

- What portion of the energy budget of the total machine (and its enclave within the broader HPC facility) is spent on controlling temperature, vs. on providing compute capability?
- Which applications, and at what allocation sizes, result in the greatest amount of contention for shared resources like the network interconnect?
- When job occupancy is high and network congestion is low, but CPU or GPU utilization is also low, what are the jobs that are running at that time, to inspect for some bottleneck which is preventing codes from fully exploiting the available hardware?
- How much do identical measurements vary across nodes, and how much do identical measurements vary for each node across time?
- Are there any deviations from normal performance measures that can be accurate predictors of pending hardware failure?
- How often are job walltime limits reached, and how often do jobs terminate (successfully) without using some significant portion of the walltime that they had requested?
- What are the most used system libraries, compiler versions, and versions of applications?
- What are the most frequent causes of a program being terminated by the operating system?
- What are the least-utilized components of the total machine architecture?

These are just a handful of such questions, by no means a comprehensive list. What may stand out in that list is the frequency with which the word "job" appears.

Often some measure of interest will not be observable without increasing the sample size beyond the one application or workflow that a user may have enabled instrumentation for. Observations of multiple programs and also observations of sources outside of the scope of applications are needed to answer most such questions. Some of these observations can be accessed within its context using open-source toolkits or system APIs, while other data points may get emitted from a vendor's proprietary drivers, and one must write tools to get access to and appropriately contextualize this information. Moreover, all information needs to be gathered continually, online, and retained over various epochs, in order to be interrogated later on to yield answers not anticipated by the developer who originally made some information observable in the first place.

Looking at online monitoring for HPC from a holistic perspective like this allows for interesting questions to be asked and answered, but the necessary software and sensor infrastructure gets complicated, invasive, and expensive, very quickly.

Here we can see yet again that monitoring systems serve a variety of purposes, and so their deployment and use will have a diversity of motivations. An application developer is unlikely to be personally concerned with the thermal consequences of using a high-speed solver library that activates additional circuitry and causes more heat to be dumped by a compute core over the duration of their job. The types of jobs which coorellate with an increased load on the cooling infrastructure, and the peaks and valleys and averages of such thermal readings,

likely *will* be of interest to someone who is tasked with managing a machine, budgeting for power, or maximizing the longevity of machine parts.

Gathering and making use of these data sets means taking on a wide array of engineering and design challenges, many of which are discussed in the next section and later areas. One such challenge has to do with the diversity of epochs and frequencies of measurements, and the need to capture and compose information efficiently. Thermal readings and power settings can be measured from a compute core in tiny fractions of a second, whereas some facility-wide sensors may have significant hystersis in reporting and only be updated every several minutes. This means that short-lived events can be more difficult to make accurate judgements about, for example. When composed against and considering their influence on the longer timelines described by coarse-grained measurements like power draw readings for a row of server racks, average ambient air temperature around a row of servers, or the power draw of the HVAC system responsible for cooling the entire building, etc., such short-lived events are difficult to render judgements about. They are also typically not able to be efficiently stored in any detail over many jobs or longer periods of time, to allow for sophisticated meta-analysis, though there are some serious efforts to do just this, such as the Sonar [29] [30] project at Lawrence Livermore National Laboratory.

When integrating observations made at different system layers and produced by different development teams, the semantics of what is being reported can vary widely, and must be carefully considered when composing data. One sensor might be reporting *absolute temperature* in Kelvin, and another may be reporting the *delta between two temperatures* in Farenheit. One must record units of measure at some datum's origin, or have a brokered ingestion of information into the

monitoring system, such that various sources are processed by bespoke aggregation functions to be made available as normalized statistical metrics.

The complexities inherent to online monitoring systems quickly become apparent, especially *as the monitoring need grows beyond a single point of measurement or is desired to fulfill more than a single purpose.* From this understanding, it becomes relevant to more deeply explore the topic of capturing and using information.

1.4 Capturing and Using Data

Once an event or some state in an HPC system has been observed, it must be represented in a stable format to be useful. Our practical research interest is in the type of data that can be accumulated or streamed through algorithms to discover and react to trends and patterns. This sort of data can usually be stored for reference or data-mining as a member of a set of data that can span multiple scales or epochs, being combinable to reveal facts beyond what is locally available during the immediate execution of an isolated process.

1.4.1 Overview.

> *Representing, disclosing, aggregating, storing, and accessing observed facts about processes, configurations, input data, activity, and the HPC execution environment.*

This section focuses on the mechanics of making and using data out of something that has previously been rendered observable, in one or several of the ways outlined in the prior section. In order to adequately characterize the sort of data we are interested in, something will need to be said about each of the aspects listed here:

- Representation and Meaning

– Patterns Within HPC

– Exposing or Exporting Data

– Introspection, Opacity, and Interface Standardization

Still, it is worth pointing out that not all observations have the same complexity or purpose. For example, some observations do not need to be retained or even exported from a process to be useful. Those observations may be temporarily fixed and used to inform a process-local or immediate decision-making process, and then discarded or overwritten. Such transient observations still must be encoded and accessible in a coherent format to be utilized – even by logic within the same process. A discussion of the fullest life-cycle of data sets from observations will also serve to inform the treatment of data sets with more limited purposes and characterization requirements.

So what do we do, once something is observable?

1.4.2 Representation and Meaning. It is important in all journeys to start off in the direction of one's goal. Any eventual application of observations will be counting on the observations being correct, consistent, and precise. Further, information must include not only the measurements, but some standard notion of interpreting the measurements. Mistakes or omissions here in this fundamental consideration can invalidate or undermine the entire purpose of monitoring HPC phenomena.

An engineering specification that only included the numeric component of measurements for its dimensions might give a clue about the proportions of the design in reference to itself, but would not be helpful to understand its overall scale in relationship to its environment or other engineered objects. This would

lead to the design object being difficult or impossible to accurately reproduce, or for people unfamiliar with it to have useful intuitions about its purpose or place simply by looking at that partially-annotated specification. Perhaps the simplest notion of a standard for interpreting measurements is the expression of *units of measure*, noting what "`1`" means in terms of units of length, volume, temperature, chronological element, etc. Once the standard for one unit is expressed, all measurements of that type can be scaled off of that unit. Time can pass in seconds, or in milliseconds, or in days, or even be denoted by abstract and unscaled CPU "ticks" within the context of a single architecture.

Units of measure can themselves be complex entities. Knowing that some numerical representation refers to a temperature in Celsius may only tell you half the story. A measurement may be referring to:

- an observation of an event or state at one point in time
- rate of change between two points in time
- result of a function relating multiple observations across time or domains

When considering an application for performance data like *constructing performance models* using machine learning methods, it is worth noting that some forms of machine learning are designed to function well over completely opaque or unannotated data sets, such as deep learning using neural networks. Typically the overall data set this type of learning is applied to has at least been pre-filtered and organized into a regular structure by some domain expert to contain distinctions of likely relevance presented in a consistent layout, to allow for the learning algorithm to recognize and adapt to some notion of concepts or categories within this unannotated data. There are always trade/offs to be made

regarding the selection of machine learning algorithms, such as speed, overhead, accuracy, timeliness, consistency of input data layout, and the amount of data needed to make good decisions. For now let us assume that information is needed for purposes beyond training deep learning models, and so correct annotations will have importance across a variety of purposes, and look at what is entailed by that idea.

1.4.2.1 Encoding the Data and Metadata. The simplest things can go unnoticed but be deeply important. One of these is the way in which information is encoded. In addition to storing a value for the measurement of a temperature, and having some way of knowing it refers to a change in temperature for some epoch of time, it matters how that floating point value is encoded. For example, floating point values can be stored in the condensed IEEE 754 formatting, where there are special meanings for subsequences of bits in the byte words of a 16, 32, or 64-bit encoding. This format strikes a balanced trade/off between storage and representational accuracy, and is how most floating point numbers are stored and operated over from the perspective of a CPU. If the number were to be pulled up in an ASCII text editor and reviewed by a human, it is unlikely that they would be able to determine the precise floating point value with their manual review. Numerical values can be projected out into a character string, which is much easier for a human to understand, but consumes much more storage space, and cannot be processed for mathematical operations by a CPU without converting back into the IEEE 754 encoding, potentially decaying the accuracy in the process.

In addition to the *encoding* of observations, the formatting of *multiple observations* bundled together is a significant factor in that information's openness to exploitation. Opaque file formats, or undocumented network protocols,

bearing messages or observations, can be difficult or impossible to exploit, if the information has not been orchestrated into some consistent arrangement. There do exist remedies for this, with formatting standards like CSV, YAML, XML, or JSON. These standards make no assumption about the meaning or semantics of the data they contain, but they do impose rules on how data generally will be encoded, so that at a minimum the raw values can be parsed from the collection into its individual components through consistent mechanisms.

There are higher-level standards which emerge from more fundamental encoding standards, such as the Open Trace Format (OTF) [31] [32] [33], which is purpose-built to store the performance measurements of HPC applications.

1.4.2.2 Encoding the Expertise. There are many kinds of expertise in the HPC field. For our purposes, we will focus on three:

- Users
- Developers
- Optimizers

The users of HPC applications, especially in the scientific community, such as the Dept. of Energy (DOE), are often domain experts. Users will have a deep understanding of the purpose of an application, what it is that the software system is helping them to explore, understand, or control. A user can also be thought of as a stakeholder, or someone who approves funding for projects, manages a budget that covers a machine, an entire premise, or who needs to make decisions balancing the purpose of the software with the overhead and mechanisms of building and maintaining that software. Developers need to be experts in the mechanics of software architecture generally, from design to

deployment to long-term software integrations and standards. They understand the process of designing codes, connecting components together, interpreting compiler error messages, etc. Developers are sometimes also domain experts, and users of codes, and they are usually motivated to write code that runs reasonably optimally, though it is not as incentivized as correctness. Optimizers are developers whose role is less about building software to meet the needs of a user, but about maintaining the effectiveness of codes over time. This means porting codes to new architectures, tuning adjustable parameters to best exploit the hardware to achieve the computational task.

No person can perfectly prognosticate about future architectural evolution and its specific optimizations for any given algorithm, so there is always a role for personnel who specialize in the tasks of porting and tuning codes. Generally, it is a fuzzy distinction, but it can be said that application developers are primarily rated on their application running correctly, consistently, and are not primarily tasked with maximizing the performance of codes, where the role of an optimizer of codes is to facilitate maximum performance, as well as code lifespan through portability to novel architectures.

The roles of user, developer, and optimizer will each have intersecting but distinct domains of expertise. What people consider important, when it comes to observations made about HPC systems and software, will be strongly influenced by a person's various responsibilities and their areas of expertise. Some examples of motives:

Users or system stakeholders might be more interested in minimizing the time their jobs sit in batch queues, or in the failure rate of parts, or in network congestion or other metrics related to shared resources. They may want to know

things like machine temperature, or what versions of codes are being run the most. *Application developers* might be interested in using local and remote system state or application progress to make better decisions about task assignments or dynamic simulation domain refinement within a simulation step. They may be interested in using in situ runtime services to couple together workflows out of legacy components that are not engineered by themselves to be coupled together asynchrously, and concepts of direct in situ monitoring at the application level come into play.

Optimizers can be interested in data at all sorts of levels of detail. They may wish to observe the frequency with which a function is called during a run, or its average evaluation time. They may wish to see how much time is spent in application logic vs. in the system libraries the application makes use of, seeking places where optimization can be found. They may need to observe the behaviors of a complex workflow in situ and at scale, to find performance bottlenecks that only emerge online, during the course of a run, and are not readily apparent through offline static analysis or manual code review. Drilling into deep and invasive observations, optimizers may need to record high-frequency samples of measurements, pathtrace data to map the flow of execution, or correlate full sets of application data with batteries of performance observations generated when those inputs were being processed. This can be especially important when porting codes to a novel architecture that may offer general compatibility with the previous, but have significantly different resources types and capacities, such that an optimizer must observe how the detailed internal components of a large complex application are occupying the machine and how [in]efficiently it is running as a whole. Discovering optimal compilation options is sometimes as significant

a contributor to performance gains as is learning the optimal runtime tuning parameters.

Differences in expertise lead to different priorities and values, and this means that every system will involve trade/offs in terms of implementation, integration, runtime overhead and application perturbation, since there is no free lunch. This justifies the important design priority for online systems that can be selectively enabled and/or invasive, and that offer some general utility across multiple domains of expertise.

1.4.2.3 Time, Change, Identity, and Consistency.

The continuous interactions of discrete elements, and the ability to reason about observations of change over time, is central to the purpose of online monitoring systems. A brief detour to discuss these concepts and terminology is warranted, given their constant presence in background of this entire area of study. It is not the purpose of this paper to give a thorough examination of these delicate and important conceptual underpinnings. Rather it can be said that since we will make use of the concepts mentioned in this section without completely justifying them, we wish to be reasonably clear about what is understood.

Time is a fundamental dimension of analysis for the study of computational performance. This is true in both obvious and subtle ways. One obvious way time in itself is a factor is in the definition of the objective function that directs performance tuning choices: When a region of code executes in less time, it could be considered to be more optimal than code which takes longer to execute. Normally the distinctions discussed here are implicit to the examination of code performance or the design of systems that monitor and evaluate it. It is worth taking a moment to pause and consider the subtler manner in which time is

fundamental to observability, because of the profound ramifications that it has on short-term measurement obligations and the feasibility of long-term objectives.

For change (or similarity) to be observed, there must be something to compare an immediate observation to. A subtle and contingent way then that time is a fundamental consideration can emerge when one reflects on the semantic nature of observed phenomena, and then also on the identity of phenomena as both a type and a token. At this point in the discussion the level of detail or relevance to performance is not significant, merely the formal precondition for consistency.

Synchronic consistency refers to the *stability of meaning* for phenomena that are fixed within a single epoch, regardless of that epoch's unit of measure. These are coocurrent entities, that is to say, distinct observational artifacts that are claims about states or events that were extant within an interval of measure. The priority of synchronic consistency is the semantic load, or the meaning, of these observations. Synchronically consistent datum that are denoted as being the same type of event would connote a consistent meaning, or use a compatible scale of measurements, etc.

Diachronic consistency on the other hand speaks to the sense that as some phenomena are observed across epochs of time, *the identity of the observed* thing, state, or event, is apparent and conserved. In order to establish this kind of consistency, some provenance needs to be included as a component of the observations. This requires the observed phenomena to endure long enough to be named uniquely and be distinguishable from other similar entities. In simple terms: The story that is being observed and recorded may be changing, but the character that this story is about is the same character at every point in the story.

Consistency is generally assumed by designers and developers when working within their own projects. That is to say, most observational systems both assume and combine synchronic and diachronic consistency, because they were built for a singular purpose or by a team driven by a shared motivation. As such, consistency is often silently imported, and exists as a mere assumption. One reason to be aware of these assumptions is that without attention, the assumptions can become false assertions, and the observations then fail to reflect the truth or support comparison with other observations thought to be of the same nature. No system of monitoring can capture every aspect of everything that is true at every time, the system itself would come to dominate its own observations and lead to a nonsensical infinite regress. Trade/offs have to be made about the amount of specificity that is tracked in a system in order to provide safeguards to ensure synchronic and diachronic consistency. These specifiers, or meta-observations, should be chosen to maximize their value in lending stability to the observed performance phenomena they are correlated with, in order to justify the overhead of capturing and retaining them.

Meta-observations can be thought of as *qualifiers* that say things about an observation, as well as the observed thing, helping to distinguish both the reference and referent. Qualifiers help to establish and refine functional categories, prevent contradictions, and give hand-holds to grasp and utilize the observations for practical purposes. These qualifiers need something stable to be tagged to, and this is the object of the consistency described in the above paragraphs. Here of course we encounter a regress of rigors, where qualifiers need something attached to in order to enunciate for that thing its stability of identity and meaning, and yet *the qualifiers themselves seem as though they would need qualifiers in order to be stable in the meaning or identity they are capable of conferring.* This regress

of observations and meta-data qualifiers is potentially infinite in theory, but in practice is rarely deeper than one or two layers of abstraction, and so does not emerge as problematic in most systems.

One way in which synchronic and diachronic consistency can be thought of is in the language of ontology, in the differentiation of *types and tokens*. There emerge many differences in the kinds of knowledge we might have, and the kinds of claims that our systems of observation might make about the observed. Types are used to enunciate *what something is*, while tokens are used to enunciate *that something is*. The distinction between knowledge-of and knowledge-that is useful for further unpacking the distinct between the synchronic and diachronic.

The establishing of *types* is a way of encoding knowledge synchronically. We might know it is true that `2 + 2 = 4`, but that does not tell us that "`4`" exists somewhere in the world, or that 4 of something exists or that something happened 4 times. For an easy to grasp example, let's think of how one knows *what a tiger is*. We can have some idea of what a tiger is, for example, being a big cat with striped fur, without there needing to be a tiger nearby to point to as a means of providing a more robust ostensive definition. In this sense, we can construct some kinds of knowledge about what a tiger is by assembling other concepts productively to establish a new *type*. This type of thing is a cat, it is a big cat, it is a striped cat, etc. Types can be as simple or as specific as needed for practical purposes.

When once we observe something, we may identify it and wish to make note of it. What we are identifying at that point is a *token* instance of that type of thing. We can say that "*A tiger is over there drinking water, on the other side of the river...*" This form of tiger-knowledge entails an existential claim, that not only does *a* tiger possibly exist, but *that* specific tiger exists, at a certain point

in time and space. Further, it should be mentioned explicitly here that one does not need to know the type of a thing in order to know that the thing exists, but in order to reason or communicate about that existential knowlege,some type will necessarily be applied to it. When this happens, this observation of an untyped thing, observers will often reach for more general types and begin an ad-hoc construction of compound type: "*Big stripey animal thing over there...*" Particular to our topic of performance observations within HPC, this automatic attribution of types and identities to "existential knowledge" can be a major factor in general limiting our ability to utilized those observations for analysis, optimization, and feedback.

Ending this pedantic excursion into the realm of the abstract, and returning to the applied topic of online monitoring for HPC, this discussion of time, types, and identities must remain incomplete. Hopefully the distinctions called out in this section will lend some clarity to the reasoning that is done elsewhere in this text, regarding the formal requirements for systems of observation. When an online monitoring system makes observations, an essential aspect of what gets observed and recorded must be these qualifiers that establish some stability in the type of thing, and in the identity of things over time.

1.4.2.4 Combination and Unit Semantics.

Domains, complexity, incompatibility, and a brief look at one approach to the challenge by the Scrubjay project.

As has been mentioned several times so far, there is a wide dissaray of data in the HPC universe. Data can represent activity or state from differing domains, sourced at different intervals, from various tools, encoded in unique formats, with varying degrees of online accessibility. Development tools like TAU [34] or HPCToolkit [35]

will describe the application domain, where other technologies like Ravel [36] [37] and Multipath Internet Protocol (MPIP) [38] can both describe and adapt activity in the networking or interconnect domain. Other information can be drawn from facility monitoring sensors, vendor introspection APIs for racks, power, and thermal management, etc.

There are various *Operational Data Analytics* (ODA) platforms that exist, some of which will be discussed later in "Monitoring for HPC: Dedicated Frameworks" (§ 1.5). For the most part, these integrated monitoring solutions are targeted to serve the needs of users from a particular domain, such as machine administrators, and do not offer value to users in other domains than that for which the ODA was not designed. Having a specific purpose, these monitoring solutions will often have built-in data processing routines, visualizations, logs, and reports which apply aggregation, transformation, and presentation of a priori designated domain-appropriate information.

It is interesting to conceive of a scenario where all monitored information could be retained and made available for many kinds of online and offline purposes, not limited in utility to a single domain, and where the set of data had not been transformed or aggregated in ways that would restrict its ability to be used to answer questions not anticipated at the time the system was deployed or metrics were gathered. While the computational and data storage resources necessary for such a system are themselves not trivial and would represent a significant investment of time and capital to field in production, another concern emerges regarding the dissaray of information. This is especially true when all of the data is not being prepackaged for an ODA infrastructure.

Solving the challenge of complex combinations of units and data semantics remains an open research area within HPC. A notworthy contribution to this topic was made by the ScrubJay [39] project. Scrubjay provided a constellation of tools to gather, store, annotate, and process queries over precisely the complex types of data we've just described. Effectively, it decouples the collection, representation, and semantics of data.

Scrubjay allows for data gathered from any source to be placed in a *wrapper* which represents it in a common format that can be transported, stored, and queried. Data is useless without meaning being ascribed to it, so Scrubjay also provides a framework for applying *semantics* to the wrapped data. These semantics are reusable, and can be applied automatically to all data arriving from various sources, after they are first annotated manually by a user with some system expertise. These stable semantics provide the basis for composable *derivation* functions, which define rules for inferring information from or computing relationships between various data sets. Because the number of derivations is potentially vast, the final contribution of Scrubjay is a *derivation engine* that navigates this space to efficiently find sequences of derivations that are appropriate to resolve queries over the wrapped and semantically-annotated data.

When queries are processed over this data, results are constructed combining both natural joins as well as *interpolative joins*, which are translations and projections of compatible kinds of data into the semantic categories or units of measure that a user has requesting in their query. When results are delivered, the rules the derivation engine used when making any interpolative joins are also presented, so that the derived results are open to verification and the resulting data set is reprodicible even if additional data or semantics are added to the

system in the future which would cause the derivation engine to resolve the same query differently. Unlike a traditional query where a user will specify tables and columns of data, and apply specific join rules and aggregation clauses, Scrubjay provides its own query format. In this novel format, a performance analyst needs only identify a set of data sources, and then an expression of the measurements of interest, specifying the dimensions of the domains, and the dimensions and units of the measurements of interest. Scrubjay then determines whether this request can be satisfied, and if so assembles the resulting data, allowing for it to be passed through additional filtering stages which facilitate classic relational database query semantics.

While the support requirements for the Scrubjay platform are non-trivial, as it relies on a dedicated cluster of HPC servers to store and process continuously streaming site-wide monitoring data, the approach and the tools provided by Scrubjay represent a meaningful step forward in this open research area.

1.4.3 Patterns Within HPC.

Qualifiers and considerations common within HPC scenarios, such as versioning, configuration of operating environments, hardware variability, communication hysterysis, and undifferentiated noise in observations, etc.

Application codes evolve over time, as well as the characteristics of the input data that codes operate over. Underlying the application's code, the operating system code, its version and configuration, the versions of system libraries and vendor drivers, the versions of linked libraries, and the general machine environment can be significant to the performance of codes in execution. Performance observations are often relevantly connected to various combinations of these factors, in addition

to direct choices made by application developers and resulting from the algorithms they implement.

Furthermore, user priority level and permissions may have a direct impact on the performance characteristics. Some users may have their codes run transparently on assets that are shared between multiple users, and experience wild performance fluctuations that are completely beyond their ability to influence. Higher-priority users on a machine may delay or even evict lower-priority users, leading to variability in observations that would require this condition to be known in order for a user or automated system to make sense of.

Observations of software in *highly-variable environments* can make it difficult to gain insights into the actual performance characteristics of the software. If the code configuration or input set is always changing, it can be hard to know the ground truth about general system performance. If the relative priority of a user or utilization of shared resources are always in flux, it can mask-off the behavior of particular versions of software. Without some "stable middle" of observations made about any given configuration, and a sequence of observations showing that a centroid of observed values has shifted in one direction of another, it is a challenge to know whether performance was gained or lost by any given change to the system configuration or an application's code.

When the density or distribution of noisey sets of observations are not regular, a case can be made for simply throwing away the irregular or outlying information and using what lies more towards the median. In such cases, using less of what is observed can actually be beneficial to the cause of general performance understanding. It is important to point out though that in order to detect that there is noise, and that the noise is irregularly dense and centered around a stable

middle, all of the observations of some epoch under consideration will have to have been made and analyzed. Just because some observation is later deemed to be the result of noise doesn't decrease the importance of it being either exported into a monitoring system, or exposed to inspection.

On that point, let's unpack what it means for something to be exposed or exported.

1.4.4 Exposing Data. Observability, even in an online sense, does not necessarily require information to be actively moving around within the operating environment or monitoring system. It may be sufficient to the needs of the online monitoring system that various components are available to be interrogated as needed, that is, that relevant performance metrics are merely *exposed* to a monitoring system.

Many sources of information in traditional HPC operating environments are regular system components that exhibit this "inspectability" property. A common way for performance data to be gathered is to inspect the statistics of a running process via operating system API calls, or by interrogating a filesystem abstraction such as `/proc/stats` which makes this data available through the form of memory-mapped files which are continuously updated with new statistics for processes.

In-memory logs and filesystem storage cannot be continuously populated by performance observations, this could consume all available resources over a long enough period of time. Because of the need to not burden the system with it own introspection, information that is made continuously available via exposure to inspectability is also often transient in nature. For example, the *history of values* in `/proc/stats` is not retained forever, it is continuously updated in place, obliterating the prior observations as new observations are made.

This update-in-place behavior imposes limitations to the types of understanding that can be gained, such as preventing the chance to identify that some average performance degradation was due to the interaction of two independent processes both simultaneously bursting with abnormal amounts of activity. Another point is that these systems usually do not retain performance measurements for processes which are no longer actively running on the machine, though they may be contributing in essential ways to the performance of processes which are still running, such as the case of complex scientific workflows that integrate the inputs and outputs of many independent processes, and where metrics are reflecting the behavior of the workflow in its aggregate performance from beginning to end.

While there may be counters and averages that track activity over arbitrary spans of time, the precise update interval, event density, or general distribution of events that are accounted for are not features that can be seen without this exposed information being retained in some way.

1.4.5 Exporting Data. Given the simple and limited nature of exposing information, in order to do more sophisticated things with data in our monitoring system we must retain it, and this will involve *exporting, or recording and migrating that information between components* of the system. This movement of information can happen in a number of ways. Information could be copied immediately, and in full, over to the receiving component. Or perhaps a lightweight reference to that information might be dispatched, taking the form of an event record or data pointer, to make some other element in the system aware that this information now exists and may be consumed.

In cases of lightweight dispatch of event records or pointers to data, other structures are implied, such as reference management, caching, or queing of records, transaction management services, etc. so that the information those records point to does not get "garbage collected" and vanish unaccountably. There are trade/offs here, like everywhere else, in the balance of retention policies for this information, and the needs of the monitoring system to not consume too large a share of the resources which are meant to be dedicated to productive computation.

Any time information is moved from one context to another, leaving the boundaries of a process, a shared library, a tool, or a machine, it can be said to be *exported*, for our purposes in describing this research area. Here are some of the primary techniques or models for exporting information:

- logging
- checkpoint
- cacheing
- polling and pulling
- broadcast or push
- hybrid push/pull
- publish/subscribe

We will now discuss each of these techniques in turn.

1.4.5.1 Logging. Generally this method involves a "fire and forget" or write-only approach to recording information into a monitoring system. Sources that generate data and spool it out into a log do not typically also read back from

that log. This allows the logging mechanism to be optimized for low-latency intake of data, such that it does not pause the work of the application any longer than necessary.

The mechanics of logging systems are able to be much simpler than some of the alternatives below, which make this a popular choice for developers who do not have sophisticated observational needs. Logs can be as simple as appending output to an asymptotically growing file containing messages for a particular session. Being relatively passive systems, logging mechanisms are typically enabled (or disabled) via the use of environment variables or command-line options to applications.

Some logging systems allow for "levels" of logging to be enabled, so for example one can see only critical messages at a certain level, or could see all available log output at a different level. This log level control allows a user to control the amount of overhead that the logging mechanism imposes at runtime.

1.4.5.2 Checkpoint. Long-running HPC applications that operate over large datasets do not typically have enough time to continuously write their intermediate results out to the stable long-term storage. This is due to the relatively slower speed of I/O that addresses the long-term storage, in contrast to the high speed system bus and volatile memory. In order to provide some safeguards against losing all progress in the event of an application crash, or to be able to rewind a simulation and advance down a different search path, applications can choose to periodically write out their data at some user-defined intervals. These snapshots of data are called checkpoints.

One technique used for the export and storage of performance observations is to embed the performance measurements alongside the application data, and

amortize the cost of measurement and I/O into the cost of creating and storing the checkpoints that the application is already producing. The Cheeta [40] codesign framework provides an excellent example of this, where an overall job management tool (Savanna) and a low-level performance monitoring tool (TAU) would emit their metadata and measurements into the streaming I/O layer (ADIOS) used by the application, embedding, contextualizing, and preserving performance observations.

This technique achieves two things. Firstly, the amount of overhead and additional I/O imposed by the performance metrics often dissapears into the large volume of work done to create and store an application checkpoint. Secondly, it provides a natural correlation between the productive work that an application performed, and the measurements of the performance metrics as it did that work. When "replaying" the checkpoint data, a developer has at the same moment a picture of the work performed, as well as the measurements of the execution environment and how efficiently the code was able to produce that work. This embedding of performance measures and metadata into an application's output, or into its checkpoint snapshots, gives a fair amount of additional provenance, including the scale of the job and the machine it was run on, so insights could be gained in future reviews by comparing similar jobs on the same or similar resources, to observe the impact of code changes on application efficiency over time.

1.4.5.3 Cacheing. In some systems, information is generated continually during an application's execution, and it is not overwritten in place but retained, and yet it is also not immediately exported fully into its final storage location. In such cases, information needs to be exported into a cache of some kind.

There are diverse reasons for caching information at various stages of a monitoring system. One simple and intuitive reason would be to avoid interrupting communication that is being done by the applications that are being monitored. Communications are typically orders of magnitude slower than computations. Overhead can be lowered and performance improved in many cases by retaining high-frequency events locally, and perhaps doing some compression, filtering, or other operations on the data, prior to its re-export and further transmittal.

Cacheing systems can range in sophistication from something as simple as a first-in-first-out (FIFO) queue, to complex event-processing layers with scriptable behaviors allowing custom logic that can react to the contents of things being stored in the cache. While it is not required, ideally all caching systems will have some mechanism in place to alert the user and perform appropriate failsafe actions in the event that the cache grows beyond some reasonable size.

A cache can be implemented as a variety of different data structures, sometimes embedded within additional data structures. A ring buffer or unbouned queue is just as valid a means for retaining cached observations as a hash table. The data structure that is utilized should be selected based on the desired use-case of the system. A good example of this can be found in the Caliper [41] performance introspection tool. Caliper uses different data storage models depending on which services a user has activated at runtime. This allows it to record information in a manner optimized for low overhead, factoring in both the type of contextualization that is needed for observations, and the granularity of observations being requested.

Nearly all online monitoring systems employ some form of cacheing or another, especially if they offer support for network communication of observed data. A monitoring system typically pools or stages information before writing it

out to disk into a log file, or sending it out over the network in a publish/subscribe system. In such cases we'd describe it as using a cache, but generally the system would be a logging system or a publish/subscribe system, as that reflects the behavior of the system as a whole.

1.4.5.4 Polling and Pulling. When monitoring information is retained, but frequency or volume of information, or the operating environment's sensitivity to overhead is high, it may make sense to use a *pull*-based model for monitoring. In this context, pulling refers to the request and receipt of information being exported from one context into another. This usually takes places via interprocess communication methods, and can be within a single computing node, or coordinated remotely across the interconnect between nodes.

Oftentimes as well, due to the simplicity of its design, a *polling* mechanism is built-in that allows for remote processes to determine whether it is time to pull information, or perhaps what information they would like to pull. In this model, the monitoring system will provide a mechanism for remote components to interrogate the sources of information to determine whether new information exists, or the transmission of that information is warranted. Polling ranges in sophistication from full complex query languages where results are computed and returned, to simple call-and-response notifications where the polling message essentially says, "I'm ready, send what you have."

What is distinctive about polling and pulling models is that communication of the exported information is directed by the receiving end, and the sending side operates passively, cacheing its information and servicing the remote requests for it.

1.4.5.5 Broadcast or Push. Both the broadcast and push models are similar in that the sender of information is in charge of the content and

frequency of what is exported. A system can be said to be broadcasting if it is indiscriminately dispatching information out to all other components of the system that are capable of receiving it, regardless of the content of the message or the capacity of the receiver to use it productively. This can be useful in cases where network interconnects or on-node IPC can very efficiently duplicate information out and provide it to multiple recipients within the same timeframe or with the same resource consumption as it would take to deliver it to a single recipient.

1.4.5.6 Hybrid Push/Pull. Push/pull systems [42] allow for disparate components to discover and engage with each other, but do not impose a particular coordination scheme or global state to be maintained. Each side of communication waits for incoming requests (or results from their previous outbound requests). Both sides are also freely permitted to fire off messages or push information out into the system at their own prerogative. This is useful for observing parallel applications that have ranks or components that operate independently of each other or that may finish out of synchronization with each other.

1.4.5.7 Publish/Subscribe. Monitoring systems that provide a publish/subscribe model are able to offer the finest-grain control over the movement of information of all the models discussed so far. These systems provide brokering services which connect receivers and senders together, and facilitate the orderly exporting of information through the system. In addition to the movement of monitoring information, these systems also must coordinate the state of the publishers and subscriber agents themselves, in order to provide notice about the availability and information sources, and the presence of information sinks to transmit to. These systems can make powerful contributions to the orderly

operation on an online monitoring system at scale. Because of their capabilities, they are can also be difficult to implement, and can require greater configuration to effectively deploy.

1.4.6 Introspection, Opacity, and Interface Standardization.

A particularly lamentable fact about extant online monitoring systems is that they generally have bespoke or opaque interfaces, protocols, and data formats. Since the beginning of the discipline of computer science, one of the running jokes amongst practicioners has been the sarcastic pronouncement, *"The great thing about standards is that there are so many of them to choose from!"* This applies to numerous subdisciplines in computing, but online monitoring no less. When one does not know that an information source exists, or when one does not know how to properly interact with it, it may as well not exist except for the overhead that is incurrent in its processing.

Many monitoring sources in HPC are produced for specific research experiments as one-off accessories, or for the utility of a single integrated workflow, or the operation of a specific physical compute resource. It can be difficult if not impossible to exploit the capabilities of these masked-off sources to make contributions to any more generalized online monitoring frameworks. Expert knowledge and specially-targeted and tailored code would need to be written in each instance of a deployed monitoring system in order to discover and then tap into these otherwise-observable subjects for online monitoring. That kind of knowledge and that amount of labor, for both initial implementation and for project maintenance, is obviously prohibitive in comparison to the commonly marginal value that can be discovered and extracted through online monitoring, as *the opportunities for large gains are assumed to be discovered and integrated*

into the layer of the project-specific internal introspection that is already in place, having the property of being opaque that is being discussed here.

It may not be the case though that the introspection capabilities internal to a particular project or vendor-specific OS and hardware management were future-proofed or able to fully capture and exploits the opportunities for optimization that are available, where being integrated into a broader or more holistic monitoring framework potentially could.

Oftentimes capabilities have been available but the desire to utilize these abilities is newly emergent, and can be hamstrung by the lack of observability or accessible control points. One solution to the challenge of introspection is the use of generic performance annotation hooks, source-level instrumentation that is disabled by default and imposes no overhead, but can be activated to yield a rich set of information at runtime, with detailed contextualization. The PerfStubs [6] project proposes an API and toolkit for this. PerfStubs is not tied to any specific tools, but provides hooks for performance monitoring tools to tap into and observe programs in execution. This means some tools can engage with it in sparse flyweight ways that avoid the overhead of something like full callpath tracing and context tracking, and take actions more suitable for always-on runtime monitoring or ocassional auto-tuning.

Understanding application-level context, such has having explicitly identified iteration boundaries, or having clearly defined divisions between communication and computation phases, allows a generic annotation framework to go beyond simple introspection tasks. Having performance-related annotations already baked into application codes can render them into both sources of information and targets of tuning within an online monitoring framework.

The more sophisticated the purpose-built or project-internal introspection system is, the more capabilities it is able to provide for configuration and efficient operation, there is an increasing likelihood that it is difficult to observe or interface with. Very simple models like logging, for example, where observable surface-area of noteworthy events are exported into plain-text log files with lightweight structure such as having tab or comma-delimited data fields, are relatively trivial to observe and integrate into a broader monitoring platform. More complex end-to-end workflow management systems with support for internal logic and dynamic behaviors and a robust online information flow implementing publish/subscribe capabilities, can from the outside be entirely opaque and mask off the events and status of internal components within the events and status of the management systems' data model and protocols.

It is unlikely that pure generality of observables will be achieved, given the multitudes of design influences and priorities that factor into fielding and operating even the simplest of HPC platforms in the modern era. Pure generality would mean that such that all observables can be heirarchically integrated from multiple perspectives, and with unrestricted visibility of any subsets of phenomena, and phenomena have stable identities and productively-combinable semantics. This would, of course, require incredible discipline and thoughtfulness in the design of the first-order lowest-level components of the system. Interfaces between ascending layers of integration would be required to conform to protocols that conserved the observability of any desired element operated over or participating in the computational task acting above it. There are many justifide reasons why data and activity live mostly in unobservable enclaves owned and arbitrarily operated over by so many processes. Not least among these reasons is efficient use of storage

resources, and a desire to maximize the proportion of computation that produces output significant to users, compared to the amount of work done to facilitate that productive work.

What we may clearly percieve here is how complicated the trade/offs become, at multiple levels of understanding. There are broad and almost existential trade/offs of purpose between the partially-overlapping motives of stakeholders, developers, users, and optimizers of HPC systems. Then there are comparitavely microscopic trade/offs with enormous downstream implications that are made at the software and hardware design and integration levels of HPC. We can see then that embedding capababilities and increasing the sophistication within one design layer, or within a single component of a larger system, can have the consequence of *decreasing* the visibility of that layer or component to outside observers. At the same time, the activity observed by that introspection system will have high enough overhead to export elsewhere that it becomes increasingly likely it will not be considered worth the percieved benefits of doing so.

1.4.7 Case Study: The CDC 6600 Mainframe. HPC, or "supercomputing", has moved through several different eras from the single-core mainframes of the 1950s, to vector machines, into the era of distributed memory, heterogeneity, and extreme scales of devices. As time and technology progressed, the innovations of prior eras were integrated into the newer designs, often combined into unified components that were then multiplied in number and interconnected to provide expanded compute capabilities. These growing numbers and increasing reliance on complex communication patterns led to cyclical renewal to the challenges of understanding and fully utilizing the available resources of those machines. In addition to increasing performance of codes, having insight into the

state and behavior of these complex machines could also increase the performance of developers and users. The easier some HPC resource is to understand, develop for, and debug, the more productive work can be achieved with it.

One of the earliest commercially available mainframe machines was known as the **CDC 6600** produced by the by the Control Data Corporation [43]. Principally designed by Seymour Cray [44], one of the legendary early innovators of HPC technology, the 6600 introduced a number of ideas which became fundamental to nearly every HPC system which followed. We mention it here not only because it was a conspicuously popular and important machine in the history of HPC development, but because it also shows two major challenges which are still with us today: Parallel complexity, and the trade off between opacity and operational efficiency.

The 6600 was designed with multiple functional units which were able to operate in parallel with each other, at the same time, reducing the gaps in productive work that are the natural result of operations stalling as data or new instructions get fetched from memory. In addition to a central processor (CP) which executed the majority of user code, there was an instruction cache put in place to facilitate pipelining, and a cohort of 10 different "peripheral processors" (PP). This queueing of instructions, and the processing of the different steps of an instruction (i.e. loading, evaluating, branching, storing, etc.) simultaneously, where new instructions can be introduced at the front end of the process as older instructions are partway through being evaluated and eventually retired, is typically referred to as *pipelineing*. Even naive implementations are capable of providing significant speedups, and these will typically be bounded by the depth of the execution pipeline and the frequency with which instructions cause unavoidable

delays in the handling of a stage of execution, preventing that stage from vacating and the preceding stages from moving forward, delays which are known as *stalls*. The 6600's inclusion of an instruction cache was an early example of this pipelining idea, which has grown into much richer and incredibly sophisticated forms in modern HPC.

The overall performance of codes running on this hardware was in large part a factor of how efficiently a developer could take advantage of the parallelism the multiple PPs offered. At the time of the 6600's development, operating systems and compilers were also (relatively) new concepts, and were not able to provide many of the modern advancements of automated optimization or parallization of code regions. The complexity of how the CP and the parallel PP components would cooperate to create a kind of pipelining, also created a novel burden for code developers targeting that platform, to design their implementations of algorithms around the optimal behaviors suitable for the 6600's specific internal coordination patterns. If a programmer did not take advantage of the parallelism on offer, the machine was hardly able to do it for them, but a higher degree of expertise and design overhead was thus introduced and imposed on HPC developers.

While the 6600 was not a true "multiprocessor" system in the modern sense of the term, it did support some forms of pure parallelism, where real work was being done concurrently, and not merely seeming concurrent through time/sharing techniques like context switching. The 10 PPs each operated independently of each other and the CP. Their primary task was to load and store information from the main memory of the 6600, freeing the CP to use its time more productively to perform complex multi-step operations on that data once it had been fetched from memory. The 0th PP was dedicated to running the operating system of the

entire mainframe, including the CP. The 9th PP was dedicated to running the user terminal, managing the display and interactivity for users running programs and evaluating the results. Activity for memory accesses, the PPs, and the CPs, was coordinated by setting values for different states into registers. Importantly, these private registers were not addressable by user code running on the machine.

Here we see one of the first instances of an intentional trade/off which remains an interesting challenge all through HPC into the modern era: *Exchanging opacity for efficiency*. Private registers effectively created a communication channel for the operating system, nascent though it was at the time, to orchestrate the behavior of the machine in support of user software demands, without needing to impose the overhead of synchronization with the specific activity of user programs. This did, however, mean that the state of the machine itself could not be easily introspected on by any software that was running on it, leading to novel challenges when searching for optimal use patterns, or debugging a code that was behaving unexpectedly. Another way in which this opacity exchange could be seen was in the physical presence of the machine itself. Unlike many of the machines which preceded it, the 6600 did not have any integrated display panel of lights to represent the values of the different registers. Normally these had been used to inspect the machine state or to perform debugging, even if only for the initial startup of the machine, after which printouts or a cathode-ray tube (CRT) display could be used to check state.

It would not have been feasible for the 6600 to offer a lightbulb-array-based "live look" into the state of the machine registers, as the physical layout of the device was too dense, and the number of registers that would need to be displayed was too great to be practical. So even this early on in the field of HPC, it was

understood that monitoring systems are not free, that there is a trade/off. By choosing to forgo a physical monitoring system for the 6600, Seymour Cray was able to lower the power requirements, reduce the operating temperature, and bring the components of the computer closer together. The physical locality of resources in HPC systems is significant, because at the cutting edge of design, having reduced wire length translates directly into performance gains. This informs the physical layout of the 6600, taking the form of a star-shape with the CP in the center of the device, to be as close as possible on average to each of the PPs and memory banks.

The 6600 is one machine, but a representative example, showing the origins to a couple of the enduring challenges for development and monitoring in HPC.

1.4.8 Observability: In Conclusion.

Now having familiarized ourselves with both practical and theoretical aspects of making online observations in an HPC environment, we are equipped to proceed into the discussion of tools and techniques with a richer understanding of the means and meaning underlying what is being discussed.

1.5 Monitoring for HPC: Dedicated Frameworks

As should be apparent from the discussion so far, the topic of online monitoring can take on many dimensions in the HPC context. Data sources may be as diverse as the version numbers of software being run, XML files emitted by proprietary commercial sensors and software that report the power of a building's HVAC systems every few minutes, or hundreds of temperature sensors scattered around the server room with data being aggregated every few seconds, to in situ (online) probes of scientific workflow components executing on massive clusters, high-resolution performance data being captured and aggregated by the millisecond.

The manner and means by which data sources, applications, tools, and system services conspire to produce the general outcome of online monitoring are as diverse as these examples. There are as many purposes for online monitoring as there are individual contributors or consumers to the monitoring infrastructure or the system that it is monitoring.

It is therefore worth making note of the basic fact: *Online monitoring for HPC is rarely the exclusive role of a single tool dedicated to monitoring a single aspect of the HPC system.* In Monitoring for HPC: General Topics (§ 1.6) we will discuss some of the common challenges of HPC that are closely related to online monitoring, analysis, and feedback, but are not necessarily centered on a particular monitoring concept or tool.

For now, in this section, we'll survey some of the past and present heavy-hitters amongst *purpose-built online monitoring systems* [45].

1.5.1 SuperMon. SuperMon [46] is a set of tools for cluster monitoring, engineered to be high-speed and to minimize overhead. Delivered during the terascale era of HPC in the early 2000's, one of SuperMon's design goals and achievements was to allow for low-impact monitoring high-frequency events, making previously invisible behaviors of the cluster open to observation.

The system operated online, and could gather data from all nodes and assemble it into a coherent single perspective of the cluster as a whole. SuperMon's developers described the state of the art as being extraordinarily primitive, being little more than shell scripts that would periodically run the *ping* command to test the responsiveness of nodes. If the ping attempt failed, or if a support notice arrived from a user saying their job failed or they could not log into a node, machine administrators would then direct their attention to manually determining

the cause of the failure. The monitoring sensors available to administrators were, in the scenario they describe, limited to the server daemon that handled user logins, and the daemon that would respond to ping commands. There were other solutions which might have worked on their newly constructed Linux-based terascale cluster, but that did not meet their design requirements for minimizing overhead and application performance perturbation. One such tool they mention is *rstatd* for "remote status" based on the SunRPC protocol. Again, it was deemed too slow, and also at 20 years old did not offer sufficient flexibility to describe the dynamism of events and hardware that were beginning to show up in HPC clusters. They considered this inadequate for terascale computing, and set about to construct their own solution.

The Supermon cluster monitoring system was built from three distinct parts:

- Linux kernel module to observe and emit performance data.
- *mon*: In situ data server to capture and cache data from the kernel module, and to service requests for that data.
- *Supermon*: To compose samples from any number of nodes into a single set of samples that represents the state of the cluster.

Supermon, like may monitoring systems, also utilized its own client-server protocol to exchange information between the three components. Its developers used a clever encoding of performance data into self-describing *s-expressions*, something like modern-day XML, but designed originally as a part of the LISP programming language in the 1950s. These recursively-defined self-describing packets in the Supermon protocol were advances in utility and flexibility over the existing RPC packets, which were strictly defined and packed into binary formats. S-expressions

could vary in size and content, and could be easily processed and composed into various representations or aggregations as desired. They found that processing packets of this nature, even in plain text, was faster than what was needed to serialize and deserialize everything into rigidly defined structures, and had the added benefit of not requiring the use of special RPC compilers or inspection tools.

This system represented a major step forward for online monitoring of HPC clusters, being faster, more efficient, more flexible, and easier to use than the immediately outdated RPC-based monitoring state of the art.

1.5.2 MonALISA. With grid computing, often teams would be using computing systems that were connected over the internet, distributed across a nation and even around the world. The number of constituent systems in a computing grid, and the extreme heterogeneity of them, posed a challenge to administrators and users who wished to be able to observe the system in aggregate. Around 1998 the MonALISA project was born, looking to provide practical solutions to this problem.

Monitoring Agents in A Large Integrated Services Architecture (MonALISA) utilized forward-deployed "station servers" positioned at each of the major grid system locales. These station servers would run a variety of agent-based services, forming a dynamic distributed services architecturem, capable of deploying, starting, stopping, discovering, and utilizing arbitrary monitoring agents online. This system was not overly concerned with maximizing throughput of monitoring data, focusing rather on flexibility and self-organizing capabilities. It was considered acceptable and also useful for an agent to capture individual or summary measures of its grid location once a minute, and to aggregate on the order of hundreds of station server's data from around the world every several

minutes. By modern standards this is not impressive, but at the time this was very useful, especially given the deep configurability and flexibility of the agent-based system. MonALISA was also not concerned about overhead and performance perturbation, since agents were running on their own server attached to the grid facility's network. In fact, much of the infrastructure of MonALISA was developed in JAVA, rather than the traditional high-performance languages like FORTRAN, C, and C++.

Much work was done in this project to facilitate the distributed nature of grid computing, or to model this monitoring solution around the features of distributed computing. Agents would place themselves in a common registry, report changes in their availability, and report what information they were able to provide. Monitoring clients could then subscribe to the information streams from those agents, and this subscription would propagate through the system, and that client would begin to receive streams of information from all active and available agents of that type. The MonALISA framework was built to be resilient to the vissictitudes of internet connectivity, and so all operations were asynchronous and all interacting components were loosely-coupled. Individual sites, or entire enclaves of sites, were capable of performing just as effectively in isolation as they would when completely joined and online together.

This project also included a client which could project the monitoring data over a global map, allowing for useful dashboard-style visualizations of a variety of topics, for example: system availability, load balance, and data link saturations. More than just collecting and presenting information, MonALISA could also be used to optimize grid-based workflows, based on the types of agents deployed and the sensitivity of an application to receiving directives and adjusting plans mid-

run. MonALISA interfaced with a variety of other on-site monitoring tools we will discuss here, such as MRTG and Ganglia. Those interfaces is where MonALISA gathered much of the actual site data that was ingested and shared by agents. While we are in this survey mostly interested in the in situ (online) monitoring that is closer to the nodes, applications, and facility sensors, the ability to step out and up another layer and provide monitoring across vast distances, uniting multiple clusters into an aggregated perspective, is a noteworthy achievement by the MonALISA team. This project can serve as an example for how to think about and even implement some of the technologies that are required to perform those tasks.

1.5.3 MRTG. The Multi Router Traffic Grapher (MRTG) [47] [48] first emerged as a single-purpose tool, designed to monitor the inbound and outbound traffic on a internet gateway router. This perl script would read the octet counters of the router every 5 minutes, and then generate a graph which could be seen by visiting a web page hosted on the same server where the script was running. After it became open source, people from all over the world began to use it and make code contributions, even porting parts of it to C for performance increases. MRTG quickly grew in sophistication, configurability, and monitoring capability.

The main bottleneck slowing MRTG's early adoption was the need for an external client library to interface with routers over the Simple Network Management Protocol (SNMP), since not all potential users had access to or the ability to build such libraries. Eventually, a perl-based SNMP implementation was integrated into MRTG and the project was then entirely self-contained and trivially easy to configure and use. Being implemented in perl, it was also automatically

portable to every platform where perl code ran, which was just about everywhere. Since it was simple, self-contained, useful for a variety of tasks, free, and open-source, by the mid 1990s MRTG had become a very popular monitoring tool within the IT world. Usability cannot be underestimated, when considering the value and effectiveness of monitoring solutions.

Another way in which MRTG facilitated usability was by embracing a functional opinion about monitoring data: that it is less important the less recent it is. This allowed for a "*lossy data storage*" paradigm in MRTG's implementation, which would allow MRTG to compress expiring data into rolling averages of the prior measurement periods, preventing server storage from filling up with old monitoring data if MRTG was left running for an arbitrary amount of time. While also offering a boon to administrators who did not need to manually flush logs or purge databases, it also dovetailed with the automatic activity graphs that MRTG produced. By default it would offer a 5-minute resolution of the last 24 hours, every 30 minutes for the last week, and average values for every 2 hours for the last month. Two years worth of history are archived, but compressed further to where entries represent the average value over two day periods. Having the monitoring data constantly flatting and coarsening like this kept the service running smoothly, and provided handy reports to summarize both immediate events in detail, and longer-term trends from an overview perspective. Though only a simple approach, it was very practical, and had the effect of making this monitoring solution useful for both system administrators, and site resource managers who needed to keep an eye on system utilization in order to make purchasing decisions about new systems or increases in networking capacity.

1.5.4 RRDTool. The same creator of MRTG also produced a toolkit for rapidly developing one's own monitoring solutions, the Round Robin Database Tool (RRDtool), which was released back in 1999. The core functionality it provides is a time-series data model and a suite of utilities for accessing the monitoring data repository. RRDTool is the central data storage solution running beneath a number of popular monitoring solutions, such as Ganglia, Cacti, Collectd, etc. One modern incarnation of this tool is the SE-RRDTool [49], which extends the core features of RRD with the ability to provide semantic enhancements, that is to say semantic annotations, to data sources. These annotations improve the ability of tools to utilize information gathered within the system, especially for automated learning systems that do not accomodate "human in the loop" expert review of monitoring data. SE-RRDTool allows for the expression of data ontologies for values that are captured in a monitoring service that utilizes it, including units, quality of service metrics, system heirarchy such as cloud entities, and other custom user-defined typings. In addition to marking up the data, it also enhances queries, allowing for semantic-based retrieval of values with a cursory support for automatically generating derived or projected values based on the semantic rules built into user-defiend ontologies.

1.5.5 Ganglia. Ganglia [50] [51] is a popular distributed monitoring solution that targets both clusters and Grid computing environments. The Grid computing is especially supported by the heirarchical design of the Ganglia data model and services. Its implementation uses XML for encoding its data, and the previously discussed RRDTool for data storage and analysis / visualization. By 2004, Ganglia was in use at over 500 compute clusters worldwide.

Ganglia is built around a monitoring daemon that uses TCP/IP multicast listen/announce protocols to monitor activity within a cluster, gathering a set of built-in metrics as well as allowing plugins to capture arbitrary user-defined metrics. It leans into the idea of federations of clusters very heavily, supporting this through the ability to pull in collections of child data sources from various clusters periodically, and aggregate this information into a unified data store.

Generally Ganglia has operating overhead below 0.1%, since it is focused on coarse-grained sampling of metrics like hardware counters, temperatures, general system activity, network traffic, etc., and does not need to engage with or interrupt application processes, and its use of RRDTool for data management means it does not need to retain large data sets indefinetly. Though it has local services that run in situ, and it aggregates its information online, its focus on collecting samples of metrics at a coarser-grain than the individual processes or components of a workflow lends it more value to system administration types than to developers or even users of HPC systems. As discussed above, Ganglia isn't intended to be used in all scenarios, and can be complemented or even potentially replaced for certain production environments or user sets by other online metric collection services such as LDMS.

1.5.6 Nagios. Nagios [52] [53] is another online monitoring tool with a strong emphasis on monitoring of network devices and their service statuses, to provide automatic notice to administrators when there are service failures or capacity is being approached. Like Ganglia and many other services described here, it has an in situ server that runs in the background local to the nodes of a cluster. This service periodically probes the state of the machine and services, and can fire off triggered behaviors depending on what is observed. Nagios offers a very

flexible plugin system, and over the years has gained hundreds of plugins and been used as the core component of many different commercial monitoring solutions, where the commercial product contributes their added value features in the form of proprietary plugins which run on the basic Nagios software stack.

There are some limitations to Nagios, including being user-unfriendly to configure (perhaps why it is often wrapped up into a commercial product), and also not having its own data storage solution built in. However, despite these limitations, it's being lightweight and, when configured, a stable and reliable system monitoring tool, and one that can be infinitely extended through plugins, Nagios endures as a commonly available monitoring tool for making observations to support the management of HPC clusters.

1.5.7 TACC stats. In 2013, Texas Advanced Computing Center (TACC) fielded a set of sweeping updates and enhancements to the monitoring solution for their Linux-based HPC clusters, though keeping the rather straightforward name for their project: *TACC stats* [54]. The central premise to TACC stats is that users and developers and adminstrators do not need to do anything in order for it to be enabled and functioning. TACC stats is constantly enabled and accumulates performance and utilization metrics for every single job that runs on the cluster. It utilizes a variety of sources for information, from the filesystem to the messaging services to the job scheduler, to operating system performance introspection APIs. All metrics gathered into TACC stats are resolved to the job and hardware device, so individual jobs and applications can be analysed separately. Many kinds of metrics are gathered by this system, from core-level CPU usage, socket-level memory usage, swapping and paging statistics, system and block device counters, interprocess communication, interconnect fabric traffic,

memory controller cache, NUMA coherence agents, and the power control units on servers. TACC stats is built to be modular, and can be extended to track arbitrary additional data points based on user interest and data availability.

TACC stats is a fine example of what can be achieved with an always-on monitoring solution. The overhead of collecting the monitioring data is simply amortized into the operational overhead of the cluster itself. Because it has records of every single job going back to 2013, long-term trends can be observed in use patterns, so stakeholders can get clear and detailed reports about how their machines are being used, and what users needs may be for the design and purchase of future resources, or the targeting of talent and funding to support the improvement of software packages which are seeing the most use. System administrators are also able to take a more proactive approach to the detection and diagnoses of hardware failures or configuration issues, since the system is continuously collecting and integrating the monitoring data, and constantly reviewing that data for anomalies or events which were able to be correlated with problems that had previously been discovered and resolved.

1.5.8 ProMon. Observing that the vast majority of performance tooling in HPC systems is targeted at heavyweight program introspection during development, the ProMon [55] system was developed and fielded in 2015 to offer another approach to online introspection in HPC. The defining design principle for ProMon captured by it's full name: *Production Monitoring*. Like TACC stats, ProMon is aligned with the vision of always-on monitoring, so that developers, administrators, and users do not need to take any additional actions in order to have access to runtime introspection data, and the potential benefits that it might enable.

ProMon's developers are motivated like many in the online monitoring community by the need for introspection into the runtime environment and into long-running jobs on HPC systems. Remarkable increases in system scale and heterogeniety, the integration of massive and complex software projects into campaigns of scientific workflows operating over in situ data stores, and the complex entailments of individual component failure or soft error accumulation over a long run, all add increasing motivation to the case for online in situ monitoring for HPC. The challenge then is to provide flexible low-overhead facilities to meet this monitoring need, without negatively impacting system stability or software usability. Only then will users and stakeholders of large and expensive HPC systems be willing to broadly introduce online monitoring to their production environment, and not only their development environment.

Since ProMon is a generic and programmable platform, it can be configured in ways which will cause large amounts of performance perturbation to applications. However, in realistic scenarios, its developers have claimed less than 1% overhead by the deployment and use of ProMon in a production environment. On the development side, the ProMon concept outlines how value can be gained by doing more heavyweight profiling of applications, which can be stored in performance databases and later integrated to enrich the more flyweight measurements taken on the production side at runtime. Given the focus on online monitoring in HPC, we will focus on the production aspects of the ProMon design. On the production side, ProMon consists of several components: Analyzer, Injector, Reporter, Parser, and FlowGrapher. Other components can be added, but these are the essential core of ProMon. The Injector inserts monitoring probes into applications using Dyninst [56] to perform binary instrumentation, using either

static or dynamic instrouemtnation. These probes collect and organize some local data and then send them over to the Analyzer using TCP or UDP protocols.

The FlowGrapher is where users of ProMon can identify parts of their applications that they are interested in monitoring, to drive the selection of targets for the insertion of probes. Work on this component is ongoing, but it is able to provide textual output identifying loops within codes which the user can then select from by a numerical identifier. The Analyzer is a robust service capable of receiving information from a variety of processes from different applications simultaneously. Implemented as a daemon server, it also integrates the streaming probe data into a data store with provenance that can be used to disambiguate similar types of data, or data from different sources that was generated in parallel. The Analyzer operates on single or dual-event types, where single events represent milestones such as the end of a simulation step, and dual events reperesent beginning and end times, or other forms of encoding events in terms of their duration of overlap with other events.

Like the SuperMon system, ProMon utilizes its own plain-text data format to exchange information in a simple to use self-describing format. ProMon's format is named the Production Monitoring Language (PML) and is compliant with the XML standard to make it very easy to parse, and open the use of countless extant libraries and commercial data processing tools. It comes bundled with a variety of tags for annotating performance events in rich ways, and these tags can be combined, embedded, or added to in order to extend the capabilities of ProMon to suit an arbitrary array of use cases.

ProMon is an actively developed project and in its design and implementation seems to be taking a very sensible and effective angle of attack on

the more difficult aspects of in situ (online) monitoring at scale and in production HPC environments.

1.5.9 SOS and SOSflow. This author's own research work falls squarely within the domain of online monitoring for HPC, the initial contribution being the Scalable Observation System (SOS) model for online characterization and analysis of HPC applications, and its reference implementation in the SOSflow [57] project. *SOS and SOSflow are covered in much greater detail in the following chapter, but it is worth providing a summary preview of that material here in context with other monitoring solutions.*

Three principles were core to the design and implementation of SOS when it was introduced in 2016: First, that an effective monitoring system needs to be deployed in situ and running online at the same time as and colocated with the subjects that it is monitoring. Secondly, the system needed to provide the ability for interactive exploration of monitoring data online, in order to support real-time analysis of metrics, as well as feedback and code-steering. Finally, the system needed to have a small footprint in terms of memory and CPU requirements, such that it did not perturb the environment that it was monitoring.

This also meant that interactions with the SOS system would need to be loosely-coupled and asynchronous, so that no steps in observing or communicating information into and out from or through the SOS system would require an application or operating environment to block and cease doing productive work. By co-locating the observation system's online processing and analysis of measurements with the workflow components, SOS could improve the fidelity of system performance data without requireing the costly delays of synchronization or congestion of shared network and filesystem resources. While SOS had various

other motivating concepts and grew to enable a wider variety of purposes than simple observation and online analysis, those are its core tenets.

SOS comprises several components:

- Information Producers: APIs for bringing information into SOS.
- Information Management: Online and optionally persistant databases and caches.
- Introspection Support: Services to provide online access to the SOS databases and high-speed caches.
- In Situ Analytics: Components to perform online analysis, including APIs to additional languages condusive to analytics, such as Python.
- Feedback System: APIs for sending information to non-SOS entities, as well as providing feedback to sources of data such that control loops can be established for purposes such as code steering.

These components work together to provide SOS's core features:

- **Online**: Observations are gathered and available at runtime to capture and exploit features that may only emerge in that complex interactive moment, and may not be discoverable during development or with offline single-component analysis.
- **Scalable**: SOS is a distributed runtime platform, and as the scale of the deployment increases, so too does the amount of available resources for the operating of SOS adjacent a running HPC application. Because SOS uses loosely-coupled asynchronous protocols for all of its interactions with

applications and within itself, communication bottlenecks can be avoided by adjusting settings to perform analysis in situ rather than migrating information online to centralized repositories which might become bottlenecks at extreme scales.

– **Global Information Space**: Information gathered from numerous sources, system layers, or actors within an execution environment, all are captured and stored within a common context, both on-node and across the entire allocation of nodes. This information is characterized by:

  * **Multiple Perspectives** - Queries over the observed data in SOS can isolate or aggregate the data in entirely arbitrary ways, so the system can service both fine-grained analysis as well as high-level dashboard views of the system state or an application's progress. Workflows or even campaigns can be observed in their entirety, and then individual components of those workflows can be selected and introspected on in greater detail.

  * **Time Alignment** - All values captured in SOS are time-stamped so that events which occured in chronological sequence but in different parts of the system can later be aligned and correlated.

  * **Reusable Collection** - Information gathered into SOS can be used for multiple purposes and be correlated in various ways without having to be gathered or transmitted multiple times.

  * **Unilateral Interactivity** - Sources and sinks of information need not coordinate with other workflow or SOS components about what to publishj, they can submit information and rely on the SOS runtime to

> decide how best to utilize it. The SOS framework will automatically migrate information where it is needed, or resolve online queries in a parallel distributed manner when that is superior to migrating all data online for central analysis. SOS is also capable of managing the retention of unused information, and allows users to control this selectively at runtime, as well.

SOSflow was implemented as a multithreaded Linux daemon and client library, both coded in the C language and designed to be nearly entirely self-contained so to be easy to integrate into existing applications, workflows, performance tools, or broader monitoring infrastructures. It is also coded at that lower level and without other runtime service dependencies to maximize its performance while minimizing its runtime footprint in the in situ environments it is distributed across. SOS runs in user-space, and is invoked at the beginning of a parallel job script, and brought down at the end of a user's job, with the option of exporting the database of observations to persistant storage for offline analysis.

Early versions of SOS were tested out to hundreds of nodes and the overhead of the system even in early development phases was typically below 2%, with the highest overhead as a percentage increase in walltime for jobs codeployed with SOS being extremely short-lived processes, where the presence of SOS increased the runtime by only 3%, likely to do with the distributed launching of the SOS runtime daemon within the user's allocation. The asynchronous and online nature of SOS, and the efficiency of its internal communication protocols, is one of its most robust aspects. While the distributed persistent data stores would sometimes increase in queue depth for transactional commits of batches of data during times of heavy traffic, queues would eventually drain out and the time

between a value being published into SOS and it being available for querying or other uses would eventually fall back down to its initial baseline. Regardless of the system load from codeployed simulation software or the volume of traffic being processed into the persistent data stores on the backplane of the SOS runtime, the velocity of data capture, the time cost of API calls made to SOS in the client library, and the RTT for probe messages between the client and the daemon, all remained constant and exteremely high in all experiments.

SOSflow is being actively developed and has found a variety of uses in different experiments and projects in the years since its initial release. One such experiment, an integration with the ALPINE (Ascent) project for online projection of performance metrics into the domain of simulation geometry, will be discussed in a later chapter, as it is a relevant example of the power of flexible online monitoring tools designed for the modern in situ HPC paradigm.

1.5.10 FogMon. While we've primarily focused on leadership-class massively parallel Linux clusters in our discussions of HPC, there is room here to talk about some rather cutting-edge and monitoring technologies that are looking ahead to possible futures and rather exotic dynamic computing topologies, though these works to have immediate significance to classical HPC concerns.

One such monitoring tool that has recently been developed is FogMon [58], a lightweight self-organizing distributued monitoring framework for Fog infrastructures. Cloud computing has introduced *utility computing* as a cost-effective way to ship software services to their final users by substantially reducing the operational effort required by service providers. Over the same epoch, the Internet of Things (IoT) has been constantly growing, from the rich compute and sensor capabilities of cellular devices, to the embedding of wifi and low-power

general computational capability in nearly any device with a power cord or a battery. The number of connected IoT devices has caused the amount of data being generated to increase explosively, though it is noteworthy that these so-called *edge devices* are often much more resource contrained than traditional HPC machines or cloud servers. Consequently, deployments of IoT applications are typically broken into two categories: IoT+Cloud where the majority of computing is offloaded to cloud services, and IoT+Edge where data is processed locally on the device, and dependence on availability of cloud resources is minimized. IoT+Cloud gains the massive compute capacity of cloud resources, but can suffer latency, network congestion, or even service unavailability, while IoT+Edge allows for immediate interactivity, but can aggressively consume limited resources such as battery life or storage space, and can impose a great deal of complexity on IoT application developers to have safe and coordinated information synchronization between local and cloud resources.

Addressing these constraints, the paradigm of *Fog computing* is beginning to gain traction, where applications are split into microservices which can be, along with the appropriate data, migrated and executed at the location or service layer where it is most appropriate to. Fog-enabled designs make it possible to reduce network traffic by processing and filtering IoT data before sending it to the cloud, and to reduce application response times by suitibly placing latency-critical services in proximity to the information consumer at the point of interactivity. In the abstract, Fog computing relies on a common orchestration layer which delivers a Monitoring, Analysis, Planning, and Execution loop that can theoretically support the dynamic, adaptive life-cycle management of multi-service data-aware Fog applications. FogMon is an actively developed research project which aims

to support that orchestration layer, with a strong emphasis on the monitoring component, and a design that takes the Fog environment with resources contraints and unstable connectivity as a first principle. In the FogMon paper cited above, the authors provide a robust technical account of their research accomplishments and experimental validation.

FogMon and projects like it have an interesting relationship with the history of online monitoring in HPC. In one sense, Fog computing is only the latest evolution of classic *Grid computing*, which also involved the loose coupling of powerful HPC resources over relatively slow or unstable Internet connections, and which also benefitted from and existence of an orchestration layer. For an example see the MonALISA project mentioned above. The self-organizing agents of MonALISA are almost entirely mappable onto the concepts inherent in the resource-aware microservices envisioned by Fog computing. Obviously there are differences, especially in terms of the complexity of the modern IoT and Cloud computing infrastructure, and in the vast asymmetry in compute capability between edge devices and Cloud servers compared to the more evenly distributed compute capability at the nodes of a Grid computing platform. Still, there is a clear line from Fog computing back to Grid computing, and perhaps developers in the Fog space would be well-served by surveying the research done in that era.

Looking ahead, the Fog computing concepts are likely to begin to show up in traditional HPC environments, especially at extreme scales. Gains made in this field, especially by the development and validation of fundamental service infrastructures designed for low-impact and extreme-scales, will have direct implication for classical HPC compute topologies. Designing systems to be resilient to component failure is an important paradigm when an individual job may be

distributed to so many hardware components in parallel that the likelihood of a component failing during a job approaches 100% for jobs of non-trivial duration. Further, in complex integrated in situ environments with many interacting parts and irregular spikes in demand for shared resources, there is much to be gained through thoughtfully engineering data processing systems with the discipline to not rely on direct and synchronous communication for productivity and progress. As the FogMon researches clearly are aware, the design and development of these orchestration layers and loosely-coupled application paradigms is extremely complex and sensitive task. There will be much more to say about these topics in the coming years, and the type of online monitoring that is required by and enabled by Fog computing is likely to be worth paying attention to for developers and researches interested in online monitong for HPC.

1.5.11 LDMS. One of the most important online monitoring frameworks for current petascale and future exascale HPC clusters is the Lightweight Distributed Metric Service (LDMS) [59] [60] [61] [62]. This service is widely deployed and in consistent use in both development and production environments. LDMS was designed to attempt to bridge the gap between coarse-grained system event monitoring, and fine-grained (function or message-level) application profiling tools. Because of the higher cost of collecting fine-grained performance profiling data, wrapping code and extracting detailed information at a high frequency, often impinging on the performance of the code being observed, profiling and application tuning have usually been deemed episodic activities and not a part of normal or production executions. This does leave the vast amount of time that applications are running on HPC clusters largely opaque to detailed introspection, including understanding codes' impact on overall system behavior

and other applications running concurrently but in different allocations. There are inherent complexities to HPC machine architectures, both in hardware and in their software. This is including the complex Cray architectures targeted by LDMS's developers, featuring deeply customized hardware and proprietary operating system extensions and closed vendor-specific drivers. For such systems, ready-made monitoring frameworks such as Ganglia (discussed below) were unable to meet even the basic coarse-grained monitoring needs which were motivating the creation of LDMS.

Sandia National Laboratory and the Open Grid Computing Group began a collaboration on a set of HPC monitoring, analysis, and feedback tools to attempt to begin to fill in this observational gap, and in 2014 began publishing on the monitoring component of that project, which is LDMS.

LDMS is a distributed data collection, transport, and storage tool that is highly configurable, consisting of samplers, aggregators, and storage components to support a variety of formats. Samplers periodically sample data according to user-defined frequencies, defining and exposing a metric set, and running independently from any other deployed samplers. Memory allocated for a metric set is overwritten by each successive sample, no history is retained within a sampler. Aggregators pull data from samplers or other aggregators, again according to a user-defined frequency. Distinct metric sets can be collected and aggregated at different frequencies, but unlike samplers the aggregators cannot be altered once set without restarting the aggregator. Because of the strict behavior constraints dealing with both memory and sampling frequency, LDMS' samplers and aggregators can be very well-optimized to collect very high volumes and velocities of information with low-latency and nearly zero impact on overall system performance. Further, due

to the engineering effort put into a low-level RDMA communication backplane for LDMS, individual aggregators are able to collect from an enormous number of distributed hosts, with initial experiments demonstrating successful aggreation of more than 15,000:1 for RDMA over Cray's Gemini transport. Storage can write to a variety of formats, including MySQL, flat files, and a proprietary structured file format called the Scalable Object Store (not to be confused with the "Scalable Observation System" mentioned above).

The base LDMS component is its multi-threaded server daemon *ldmsd* which is run in either sampler or aggregator mode, and can support the storage functionality when running in aggregator mode. The ldmsd server loads the sampler and aggregators dynamically in response to commands from the owner of the ldmsd process. All activity within ldmsd, including the activity of samplers and aggregators and storage modules, is processed by a common worker thread pool. In more recent iterations, LDMS has gained support for more sophisticated in situ processing of sample data, including the ability to apply complex operators to metric sets as they flow through various stages of aggregation, and including the ability to interact with other services or storage systems at intermediate stages of aggregation within the cluster. In order to retain its high-level of efficiency, LDMS does not support many of the dynamic interactivity features of other online monitoring solutions discussed here. It also does not support embedded or complex self-describing data types, nor the capture of arbitrary string-based values, rather LDMS samplers are only able to capture and encode numerica values in floating point representation, a strict discipline which allows for some deep optimization to its performance and to data movement.

LDMS is a rather straightforward project, employing simple designs to great effect. It does only a few things, but it does them very effectively, and is able to make larger or more sophisticated contributions through optional integrations with other projects or tools, serving either as an information source or a sink for them. This efficiency and simplicity has led to it being widely deployed in production environments, which in turn has led to it seeing a lot of activity with various tools seeking to exploit the fine-grained performance data it is capturing, or participate in dispatching information into LDMS for other tools to have access to it at runtime. LDMS will be playing a central role in online monitoring for HPC for many more years to come, as it has earned long-range funding, has deep developer buy-in, and offers multiple types of users or administrators a considerable monitoring capability and value at nearly no cost.

1.5.12 CluMon and ClOver. Based on the CluMon cluster monitoring project's plugin architecture, the Cluster Overseer (ClOver) [63] tool is designed to allow a high-level overview of the state of a cluster. ClOver came about around 2009, and utilized the Intelligent Platform Management Interface (IPMI) protocol, which by then was becoming a somewhat standard protocol for online management of large computing systems. Extending the "at a glance" monitoring overview capability of CluMon, ClOver's principle design goal was to more completely decouple the operation of the monitoring infrastructure from some of the legacy components that CluMon had employed, such as the PCP services for monitoring, to facilitate genuine extensibility and realize the flexibility of the plugin architecture model. It was also desired for ClOver to be able to provide its monitoring features to a variety of outlets, including streaming databases or web-based dashboards rather than a traditional desktop GUI client. The ClOver project

showed improved performance, flexibility, and ability to be integrated with a wider array of components from its predecessor.

Where there are monitoring needs and available developers, there will soon be a system implemented. These two projects in pairing are a good example of what can sometimes happen in the HPC research space: When one project begins to show its seams or has an unavoidable dependency that doesn't translate well into newer execution environments, rather than updating or extending the prior work it is often more fruitful to simply recreate the project anew but with new tools, techniques, and integrations. This type of perennial re-design and re-writing of projects is far from inefficient in many cases, and can lead to better tools with less baggage, and a more positive ongoing impact. Something to bear in mind considering our favorite HPC projects and tools, as the pace of innovation in the HPC world seems unlikely to slow down in the coming decades.

1.5.13 Additional Monitoring Solutions of Note. For continued exploration of dedicated monitoring solutions, the following frameworks may be of interest to the reader:

- Performance Co-Pilot (PCP) [64]
- PerSyst [65]
- LIKWID [66] [67]
- MPCDF [68]
- OpenNMS [69]
- Prometheus [70] [71] + Kubernetes [72]
- Pandora FMS [73]

- Telegraf [74] [75] + InfluxDB [76]
- Zabbix [77] [78] [79]
- collectd [80]
- Periscope [81]
- Ovis [82]
- XDMoD [83]

1.6 Monitoring for HPC: General Topics

As discussed in prior sections, online monitoring for HPC is rarely a simple as deploying a single service and satisfying a single user or stakeholder's interests. *Online monitoring for HPC represents a complex constellation of interests, tools, techniques, challenges, and possibilities.* Oftentimes what is desired from a system will require understanding and leveraging a variety of perspectives, talents, and technologies. These solutions can be esoteric and bespoke to individual HPC deployments or teams, but over the years and across many sites and projects some common themes emerge. Here we present a grab-bag of some of the more common challenges (or scenarios that offer opportunities) related to the online monitoring, analysis, and feedback dimensions of HPC.

All of this is to say that monitoring solutions are often deployed for one or two specific purposes, and so a discussion of some of those purposes is an important part of understanding the role of monitoring in HPC environments. In each case we will look at a selection of representative solutions.

1.6.1 Portability Frameworks as Monitoring Opportunities.

Several underlying principles or themes are motivating this entire area of research

into Online Monitoring, Analysis, and Feedback for HPC. One of them is productivity, and *portability is a key contributor to productivity.*

There have emerged many standards, toolkits, and techniques now to give HPC developers a consistent API to address, with consistent behavior, that will achieve the same effect portably on different systems, past, present, and as far as can be anticipated into future of HPC architectures. Portability frameworks make for good instrumentation targets for a variety of reasons, beyond their widespread adoption. Typically they have well-documented semantics, the way in which they are used is consistent across various projects, and in recent times they often offer generic plugin-like interfaces for instrumentation or tools to connect to at runtime and interact with applications. Let's take a look at several of these solutions and their relationship to monitoring in HPC.

1.6.1.1 Distributed Computing. Distributed computing refers to a program running in parallel across several logical or physical compute resources, where each of the "ranks" of the program is isolated from the other and must communicate via the computer network interfaces rather than by being able to inspect each other's memory directly. It is no simple feat to connect multiple compute resources together in a way that enables software to run in multiple locations, discover, and coordinate in parallel to solve tasks. One would need to write networking code, become aware of load-balancing concerns, learn about efficient transport algorithms, interact with low-level device drivers, and implement all of this infrastructure quickly, bug-free, with security in mind, and then maintain it along with the main HPC application one set off to implement in the first place. Coordinating multiple distributed processes and meeting all of the above criteria is a dauntingly complex task, and yet this represents a very common need within

HPC software. This has been the case for decades, and this has led in that time to the development and widespread adoption of several important solutions for distributed computing. Here are some of the mechanisms that have emerged over the years to help HPC developers meet these common challenges with a high-degree of productivity:

Message Passing Interface (MPI): Arguably one of the most influential and essential pieces of software in HPC, no conversation could be complete without discussing MPI [84]. With its roots going back into the late 1970s, MPI refers severally to its abstract model for distributed computation, a community-driven standard with official guides produced for each version, a fully-functional reference implementation of the standard (MVAPICH [85]), any standards-compliant API and library to link applications to, and a collection of runtime services deployed over a cluster in order to launch and manage the interactions of processes using MPI to send and receive messages.

In addition to MVAPICH, there are various alternative implementations such as OpenMPI [86] or more recently ExaMPI [87]. These alternative implementations are often able to provide compatibility with new or more experimental features of MPI before they make their way into the standard reference implementation. Vendor-specific implementations are also common, as this allows the vendor to optimize the MPI runtime environment to fit and exploit features of their chipsets or a cluster's interconnect technology that may be protected intellectual property. An example of this is Intel MPI, and they even issued product-specific optimized MPI libraries like DCFA-MPI [88], tailored for their Intel Xeon Phi architecture.

Three common techniques for monitoring and interacting with MPI applications are through the use of the *MPI Profiling Interface* (PMPI, and more recently QMPI), and via the *MPI Tools Information Interface* (MPI_T). PMPI works by allowing for any calls to MPI routines to be intercepted by a tool which implements a wrapper function with the same signature, and then internally calls the actual MPI routine. PMPI is rather rudimentary in its functionality, in that it uses the linking phase of compilation to embed the tool into the application, connecting the application to the tool's implementation of certain API calls, and then connecting any "uninstrumented" MPI calls directly to the MPI library. This has the benefit of being extremely efficient, as the entire interface step can be compiled out of the application, as it is for any routines which are not intercepted by a tool, in the event a tool is being included. A negative consequence of this design is that, without some careful tool-to-tool coordination, only one tool at a time can be observing MPI activity, since the linker will select only one library to link any give MPI call to, at the exclusion of alternatives. Further, swapping from one tool to another can in some cases require an application to be rebuilt entirely. Any software which latches into MPI using PMPI in order to provide extended functionality will then prevent other tools from successfully doing the same at runtime, though often in a way that does not appear to fail to the tools which are excluded, though their routines do not get called. Depending on the order in which shared libraries are loaded and resolved by the host operating system, software with multiple components making use of PMPI can have undefined behavior without ever emitting errors a priori, which is deeply undesirable.

The QMPI [89] represents the most cutting edge enhancements to the classic PMPI model, and it provides a more flexible remedy for multiple-tool integrations,

overcoming PMPI's limits while adding additional features and enhancements. QMPI allows for multiple tools to be registered, and for the wrapper routines of those tools to be executed concurrently, when calls are made to the parts of the MPI API the various tools have implemented hooks for.

Introduced in the MPI 3.1 standard[84], MPI_T provides a general purpose API and enumerated set of tags that tool writers can use to interrogate any standards-compliant MPI runtimes and get consistently formatted and representative metrics describing the system and job's configuration or activity. MPI_T also provides a standard interface for providing hints or adjusting the settings of the MPI runtime online, with varying degrees of control based on the specific type of directive given and the state of the MPI application in execution. These routines can be a great opportunity to perform direct monitoring, analysis, and automated tuning feedback, as has been done recently using a module of TAU [34] that engages with MPI_T so that TAU coordinates a parameter sweep for settings, observes and analyzes performance, and is able to optimize [90] MPI runtime settings online. MPI_T does not supplant the need for PMPI or QMPI, in fact may of the routines that MPI_T supports are implemented internally using calls to PMPI or QMPI.

1.6.2 Monitoring and Multiple Domains. Oftentimes it is beneficial, if not necessary, to combine observations of multiple layers or domains of a system in order to understand the behavior of individual applications or system components. When general end-to-end performance results for an application can be influenced by factors outside of the selection of algorithms or quality of the source code, it is especially useful to be able to observe beyond the source code measurements and application behavior. In 2011, Schultz, et al. identified and

discussed [91] three high-level domains of analysis: Hardware, Application, and Communication. These *intuitive domains* have some natural overlap, but a result of their work was observing an increase in understanding of performance data when observations from one domain could be overlayed or projected over observations made in another.

LBNL's NetLogger [92] monitoring tool from the late 1990s used source instrumentation to capture and log performance measurements, and included a suite of offline analysis scripts to assemble detailed graphs showing *flow of application data* through a process, including measurement of time data moved between processes over a network. Events within NetLogger refer to traces of the processing of individual chunks of data. Flowing from their source-annotated origin as a logical application event, timing data could be captured showing processing through the hardware stages of loading into cache, being operated on, being queued up, transmitted, and ultimately received for processing on the remote of a distributed parallel system. Tracing chunks of application data rather than logging flat application performance measurements on a per-code-block basis gave this tool the power to reveal the complex interactions and emergent processing bottlenecks which might occur, and express these observations in a way that was relevant and actionable to the application user or developer. It also allowed for the performance impacts with origins outside of the application to be revealed in terms of their influence on the specific behaviors of the application, which can be a great help when determining where to focus effort when attempting to improve application performance.

Uniting observations from sources of information across multiple domains can become a challenge in itself, with many factors impacting the feasibility of

the task and the overhead of the mechanisms engaged to provide a solution. The Scalable Observation System (SOS) [57] was introduced in 2016 with the notion of facilitating this kind of cross-domain online monitoring, engineered to be optimized for interacting with HPC systems and applications without introducing excessive overhead or blocking the progress of components which might interact with the SOS runtime. Later in 2017, SOS was used along with the ALPINE [93] in situ scientific visualization infrastructure to automatically capture and project performance data over the geometry being simulated by the application. The SOS and ALPINE integration captured a number of different performance observations, aggregated them online, and allowed for the simulation to be observed in real-time as the geometry evolved. Users could then select among the available performance measurements and have that projected out over the geometry of the simulation. This projection of hardware performance measures into the application domain allowed for an application developer to observe the performance of their code not in terms of individual code regions, but in terms of the complex behaviors that emerge dynamically as a simulation progresses.

An application algorithm might begin to drift away from optimality in certain conditions, and it is beneficial to easily identify those conditions, perhaps to then design and introduce an updated behavior in the application that can perform a test and then switch processing over to the most suitable algorithm. For example, as two elements in a system approach each other and begin to influence each other within the simulation, an approach like this could make visible as hotspots such conditions as cache misses, mapped out over the surface of those elements. Markers could be displayed over those elements in the full context of the scene, or animated as the simulation progresses, indicating such things as an increased number of

messages between the two parallel application ranks each responsible for one of the two elements.

While it is possible to discover many origins of performance issues through direct analysis of tabulated measurements, or by using traditional performance measurement tools [94], the ability to watch a simulation evolve online and immediately see the relative performance disturbances in brightly-enunciated graphical forms, paired to the phenomena which trigger the degradation of performance, makes the task of finding and fixing input-dependent issues much more straightforward.

1.6.3 Online Monitoring for Large and Complex Codes.

Tools that automatically pinpoint certain aspects of arbitrarily complex software stacks through online monitoring, facilitating discovery and correction of bugs or execution bottlenecks.

Diagnosing performance variation in an HPC environment, automatically, online, or otherwise, is a significant challenge. Experiments [95] [96] show that it *is a problem that indeed can be solved*, despite the numerous difficulties to overcome, and so the great work ever continues. There are at present no one-size-fits-all solutions, and solutions that are being designed and deployed [97] use parts of other solutions, or take inspiration from many other projects.

1.7 Concluding Remarks

Online monitoring for HPC is a vast and complex field with many different motivations and trade/offs. As long as HPC architectures are evolving, compute loads are changing, and scales are growing, there will be a need for innovative ideas and new research efforts.

CHAPTER II

A GENERAL FRAMEWORK FOR ONLINE MONITORING IN HPC

2.1 Introduction

Modern clusters for parallel computing are complex environments. High-performance applications that run on modern clusters do so often with little insight about their or the system's behavior. This is not to say that information is unavailable. After all, sophisticated parallel measurement systems can capture performance and power data for characterization, analysis, and tuning purposes, but the infrastructure for observation of these systems is not intended for general use. Rather, it is specialized for certain types of performance information and typically does not allow online processing. Other information sources of interest might include the operating system (OS), network hardware, runtime services, or the parallel application itself. Our general interest is in parallel application monitoring: the observation, introspection, and possible adaptation of an application during its execution.

Application monitoring has several requirements. It is important to have a flexible means to gather information from different sources on each node — primarily the application and system environment. Additionally, for the gathered information to be processed online, analysis will need to be enabled in situ with the application [98]. Query and control interfaces are required to facilitate an active application feedback process. The analysis performed can be used to give feedback to both the application, the operating environment, and performance tools. There exists no general purpose infrastructure that can be programmed, configured, and launched with the application to provide the integrated observation, introspection, and adaptation support required.

This chapter presents the *Scalable Observation System (SOS)* for integrated application monitoring. A working implementation of SOS is contributed as a part of this research effort, *SOSflow*. The SOSflow platform demonstrates all of the essential characteristics of the SOS model, showing the scalability and flexibility inherent to SOS with its support for observation, introspection, feedback, and control of scientific workflows. The SOS design employs a data model with distributed information management and structured query and access. A dynamic database architecture is used in SOS to support aggregation of streaming observations from multiple sources. Interfaces are provided for in situ analytics to acquire information and then send back results to application actuators and performance tools. SOS launches with the application, runs along side it, and can acquire its own resources for scalable data collection and processing.

2.1.1 Scientific Workflows. Scientific workflows feature two or more components that are coupled together, operating over shared information to produce a cumulative result. These components can be instantiated as lightweight threads belonging to a single process, or they may execute concurrently as independent processes. Components of workflows can be functionally isolated from each other or synchronously coupled and co-dependent. Some workflows can be run on a single node, while others are typically distributed across thousands of nodes. Additionally, parts of workflows may even be dynamically instantiated and terminated. The computational profile of a workflow can change between invocations or even during the course of one execution.

2.1.2 Multiple Perspectives. *Application state and events* can be sent to SOS from within the application at any point during its execution. Developers can instrument their programs to be efficiently self-reporting the data

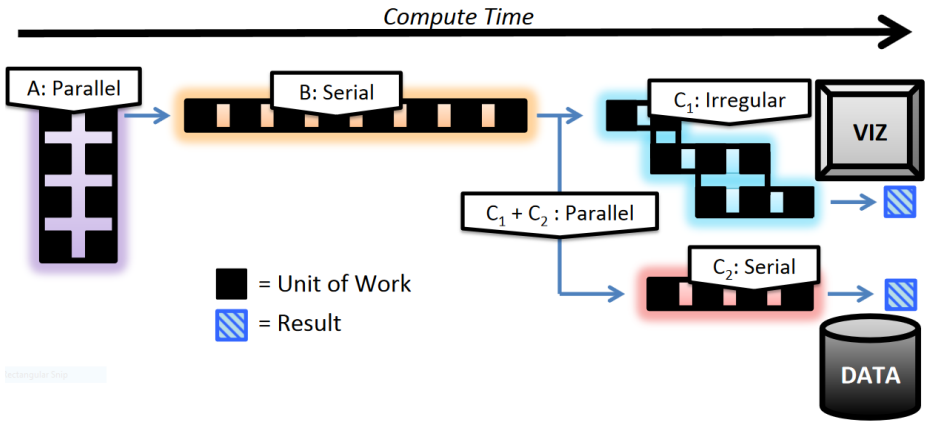


Figure 1. Applications Coupled Together Into a Workflow

that is relevant to their overall performance, such as progress through specific phases of a simulation.

Application performance can be dramatically impacted by changes in *the state of the operating environment* that is hosting it. The effects of contention for shared resources by multiple concurrent tasks can be discovered when the events of concurrent tasks are fixed into a common context for reasoning about their individual and combined performance. SOS's distributed in situ design is well-suited for capturing perspective of the global state of a machine. By co-locating the observation system with the workflow components that are observed, SOS improves the fidelity of system performance data without requiring the costly delays of synchronization or congesting the shared network and filesystem resources in use by applications.

Many existing *performance tools can provide useful observations* at runtime of applications, libraries, and the system context. The low-level timers, counters, and machine-level data points provided by specialized performance tools can be a valuable addition to the higher-level application and system data.

2.1.3 Motivation.

Observing and reasoning about the performance of workflows on *exascale computational platforms* presents new challenges. Exascale systems will be capable of more than a billion billion calculations per second, a factor of between 50 to 100 times faster than present day machines. The physical scale and complexity of exascale machines is expected to grow by similar factors as its computational speed, motivating a model that can scale to the same extent.

2.2 Related Work

Traditionally, HPC research into enhancing performance has been focused on the low-level efficiency of one application, library, or a particular machine.

Tools like TAU [99] are able to bring HPC developers a closer look into to their codes and hardware, gathering low-level performance data and aggregating it for integrated analysis after an application concludes. Low-level metrics can help identify performance bottlenecks, and are naturally suited for non-production or offline episodic performance analysis of individual workflow components. Such deep instrumentation is necessarily invasive and can dictate rather than capture the observed performance of the instrumented application when the application is running at scale or required to engage in significant amounts of interactivity. SOS provides a model that can accept low-level information such as what TAU collects, while also operating over light-weight higher-level information suitable for online operation during production runs.

Focused on the needs of large scale data centers, Monalytics [100] demonstrated the utility of combining monitoring and analytics to rapidly detect and respond to complex events. SOS takes a similar approach but adopts a general purpose data model, runtime adaptivity, application configurability, and support

for the integration of heterogenous components for such purposes as analytics or visualization.

Falcon [101] proposed a model for online monitoring and steering of large-scale parallel programs. Where Falcon depended on an application-specific monitoring system that was tightly integrated with application steering logic and data visualizations, SOS proposes a loosely-coupled infrastructure that does not limit the nature or purpose of the information it processes.

WOWMON [102] presented a solution for online monitoring and analytics of scientific workflows, but imposed several limitations and lacked generality, particularly with respect to how it interfaced with workflow components, types of data it could collect and use, and its server for data management and analytics.

Online distributed monitoring and aggregation of information is provided by the DIMVHCM [103] model, but it principally services performance understanding through visualization tools rather than the holistic workflow applications and runtime environment. DIMVHCM provides only limited support for in situ query of information.

Cluster monitoring systems like Ganglia [51] or Nagios [53] collect and process data about the performance and health of cluster-wide resources, but do not provide sufficient fidelity to capture the complex interplay between applications competing for shared resources. In contrast, the Lightweight Distributed Metric Service [59] (LDMS) captures system data continuously to obtain insight into behavioral characteristics of individual applications with respect to their resource utilization. However, neither of these frameworks can be configured with and used directly by an application. Additionally, they do not allow for richly-annotated

information to be placed into the system from multiple concurrent data sources per node.

LDMS uses a pull-based interaction model, where a daemon running on nodes will observe and store a set of values at a regular interval. SOS has a hybrid push-pull model that puts users in control of the frequency and amount of information exchanged with the runtime. Further, LDMS is currently limited to working with double-precision floating point values, while SOS allows for the collection of many kinds of information including JSON objects and "binary large object" (BLOB) data.

TACC Stats [54] facilitates high-level datacenter-wide logging, historical tracking, and exploration of execution statistics for applications. It offers only minimal runtime interactivity and programmability.

The related work mentioned here, and many other performance monitoring tools, are well-implemented, tested, maintained, and regularly used in production and for performance research studies. However, each have deficiencies that render them unsuitable for a scalable, general-purpose, online performance analysis framework.

2.3 SOS Architectural Model

Multi-component complex scientific workflows provide a focus for the general challenge of distributed online monitoring. Information from a wide variety of sources is relevent to the characterization and optimization of a workflow.

In order to gather run-time information and operate on it, SOS needs to be active in the same environment as the workflow components. This online operation is capable of collecting data from multiple sources and efficiently servicing requests for it. Information captured is distinct and tagged with metadata to enable

classification and automated reasoning. SOS aggregates necessary information together online to enable high-level reasoning over the entire monitored workflow.

2.3.1 Components of the SOS Model. The SOS Model consists of the following components:

- **Information Producers** : SOS APIs for getting information from different sources to SOS.
- **Information Management** : SOS online information databases/repositiories.
- **Introspection Support** : Online access to the information databases.
- **In Situ Analytics** : Components to perform the online analysis of the information.
- **Feedback System** : SOS APIs for sending feedback information to non-SOS entities.

2.3.2 Core Features of SOS.

- **Online** : It is necessary to obtain observations at run time to capture features of workflows that emerge from the interactions of the workflow as a whole. Relevant features will emerge given a program's interactions with its problem set, configuration parameters, and execution platform.
- **Scalable** : SOS targets running at exascale on the next generation of HPC hardware. SOS is a distributed runtime platform, with an agent present on each node, using a small fraction of the node's resources. Observation and introspection work is distributed across the observed application's resources proportionally. Performance data aggegation can run concurrently with the workflow.

Node-level SOS agents transfer information off-node using the high-performance communication infrastructure of the host cluster. SOS supports scalable numbers and topologies of physical aggregation points in order to provide timely runtime query access to the global information space.

– **Global Information Space** : Information gathered from applications, tools, and the operating system is captured and stored into a common context, both on-node and across the entire allocation of nodes. Information in this global space is characterized by —

  * **Multiple Perspectives** - The different perspectives into the performance space of the workflow can be queried to include parts of multiple perspectives, helping to contextualize what is seen from one perspective with what was happening in another.
  * **Time Alignment** - All values captured in SOS are time-stamped, so that events which occured in the same chonological sequence in different parts of the system can be aligned and correlated.
  * **Reusable Collection** - Information gathered into SOS can be used for multiple purposes and be correlated in various ways without having to be gathered multiple times.
  * **Unilateral Publish** - Sources of information need not coordinate with other workflow or SOS components about what to publish, they can submit information and rely on the SOS runtime to decide how best to utilize it. The SOS framework will automatically migrate information where it is needed for analysis while managing the retention of unused information efficiently.

2.4 Implementation

The SOSflow library and daemon codes are programmed in C99 and have minimal external dependencies:

- Message Passing Interface (MPI)
- pthreads
- SQLite

SOSflow's core routines allow it to:

- Facilitate online capture of data from many sources.
- Annotate the gathered data with context and meaning.
- Store the captured data on-node in a way that can be searched with dynamic queries in real-time as well as being suitable for aggregation and long-term archival.

SOSflow is divided into several components, central among them:

- **libsos** - Library of common routines for interacting with sosd daemons and SOS data structures.
- **sosd_listener** - Daemon process running on each node.
- **sosd_db** - Daemon process running on dedicated resources that stores data aggregated from one or more in situ daemons.
- **sosa** - Analytics framework for online query of SOS data.

2.4.1 Architecture Overview. Data in SOSflow is stored in a "publication handle" (pub) object. This object organizes all of the application context information and value-specific metadata, as well as managing the history of updates to a value pending transmission to a sosd_listener, called *value snapshots*. Every value that is passed through the SOSflow API is preserved and eventually stored in a searchable database, along with any updated metadata such as its timestamp tuples. Prior value snapshots are queued and transmitted along with the most recent update to that value.

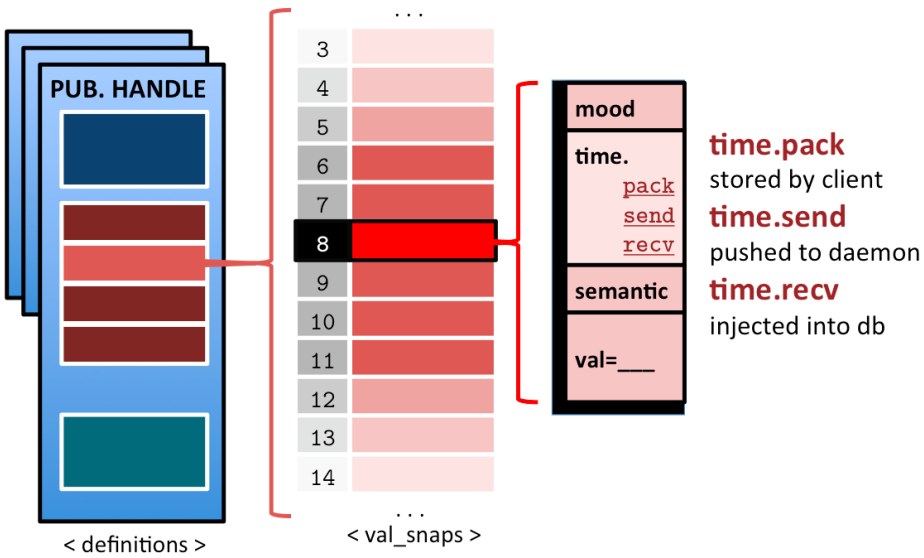


Figure 2. Complete History of Changing Values is Kept, Including Metadata

SOSflow utilizes different information transport methods and communication patterns where appropriate [42]. Communication between client applications and their on-node daemon takes place over a TCP socket connection. Messages read from the socket are immediately placed in the daemon's asynchronous queues to be processed by a worker thread. The socket is then ready for the next queued message to be received. Messages are first enqueued for storage into an on-node database. The same message is re-enqueued for transmission to an off-node

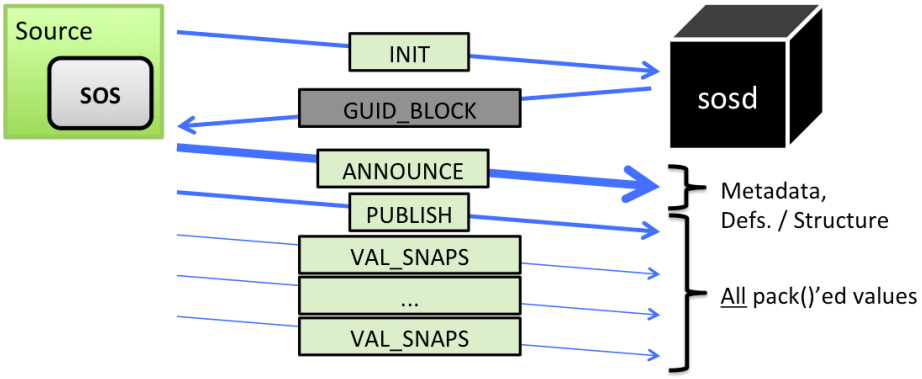


Figure 3. Client/Daemon Socket Communication Protocol

data aggregation target. The SOSflow runtime uses MPI and the high-speed interconnect network of the HPC machine when transmitting information off-node. SOSflow does not participate in the MPI communicator[s] of the applications that it is monitoring, so no special integration programming is required of application developers who already use MPI or sockets in their code.

2.4.2 Library: libsos. Applications that make direct use of SOSflow through its API are called *clients*. Clients must link in the libsos library which provides them with all of the data structures and routines essential for interacting with the SOSflow runtime platform. The library routines are thread-safe, and no process-wide state is maintained within the library, allowing application components to interact with SOSflow independent of each other.

The primary interaction between a client and SOSflow is through the pub. When a client initializes its SOSflow instance, it communicates with the daemon and obtains a set of global unique ID (GUID) tags. Clients pack values into a pub and they are automatically assigned a GUID. When the client publishes that handle, all values are transmitted to the SOSflow on-node daemon, including the complete history of each value's updates from the last to the present publish call.

All communication functions in the SOSflow client library are handled transparently. Users need only interact with a simple API to define and store values that they can then publish to the daemon as appropriate. The protocols and the codes of the client library are designed to be fast and minimize resource usage, though they will buffer values for the user if they choose to hold them and only transmit to the daemon at intervals.

Communications with the sosd_listener are always initiated by the clients, such as when they explicitly publish their pub. SOSflow clients can voluntarily spawn a light-weight background thread that periodically checks with their local daemon to see if any feedback has been sent for them. This loosely-coupled interactivity allows for run-time feedback to happen independent of an application's schedule for transmitting its information to SOSflow.

2.4.3 Daemon: sosd_listener. The sosd daemon is itself an MPI application, and it is launched as a background process in the user space at the start of a job script, before the scientific workflow begins. The daemons first go through a coordination phase where they each participate in an MPI_Allreduce() with all other daemon ranks in order to share their role (DAEMON, DB, or ANALYTICS) and the name of the host they are running on. During the coordination phase, listener daemons select the sosd_db aggregate database that they will target for automatic asynchronous transfer of the data they capture. After initialization, SOSflow does not perform any further collective communications.

2.4.4 Database: sosd_db. The open-source SQLite database engine is used by sosd_db for the on-node database. SQLite databases are persistent, lightweight, fast, and flexible, suitable to receive streams of tuple data with very

low overhead. SOSflow provides a simple API for interacting with its database to streamline access both on and off-node.

At the time of this writing, SQLite technology is also used for the aggregate databases, though work is ongoing to provide alternatives for aggregation, starting with an interface to the Cassandra database.

2.4.5 Analytics: sosa. SOSflow analytics modules are independent programs that are launched and operate alongside the SOSflow run-time. The primary role of the analytics modules is to query the database and produce functional output such as real-time visualizations of performance metrics, feedback to facilitate optimizations, or global resource bound calculation and policy enforcement. The modules can be deployed in a distributed fashion to run on the nodes where the applications are executing, or they can be deployed on dedicated resources and coupled with the aggregate databases for fast queries of the global state. Analytics modules have the ability to make use of the high-speed interconnect of the HPC machine in order to share data amongst themselves.

SOSflow provides an API for client applications to register a callback function with a named trigger handle. Those triggers can be fired off by analytics modules, and arbitrary data structures can be passed to the triggered functions. Triggers may be fired for a specific single process on one node, or for an entire node, or an entire scientific workflow. This capability facilitates the use of SOSflow as a general-purpose observation, introspection, feedback, and control platform.

2.5 Results

2.5.1 Evaluation Platform. All results were obtained by either interrogating a daemon directly through the SOS Probe Tool to inspect its state, or by running queries against the SOSflow databases.

2.5.2 Experiment Setup. The experiments performed had the following purposes:

- **Validation** : Demonstrate that the SOSflow model works for a general case.
- **Exploration** : Study the latency and overhead of SOSflow's current research implementation.

The SOSflow implementation is general-purpose and we did not need to tailor it to the deployment environment. The same SOSflow code base was used for each of the experiments. The study was conducted on three machines, the details of which are given below —

1. **ACISS** : The University of Oregon's 128-node compute cluster. Each node has 72 GB of memory and 2x Intel X5650 2.66 GHz 6-core CPUs, providing 12 cores per node. Each node is connected together with a 10GigE ethernet switch.
2. **Cori** : A Cray XC40 supercomputer at the National Energy Research Scientific Computing Center (NERSC). Nodes are equipped with 128 GB of memory and 2x Intel Xeon E5-2698v3 2.30 GHz 16-core CPUs. Cori nodes are connected by a Cray Aries network with Dragonfly topology, that has 5.625 TB/s global bandwidth.
3. **Catalyst** : A Cray CS300 supercomputer at Lawrence Livermore National Laboratory (LLNL). Each of the 324 nodes is outfitted with 128 GB of memory and 2x Intel Xeon E5-2695v2 2.40 GHz 12-core CPUs. Catalyst nodes transport data to each other using a QLogic InfiniBand QDR interconnect.

We simulated workflows using the following —

1. **LULESH with TAU** : An SOSflow-enabled branch of the Tuning and Analysis Utilities program (TAUflow) was created as a part of the SOSflow development work. On Cori, TAUflow was used to instrument the Livermore Unstructured Lagrangian Explicit Shock Hydrodynamics (LULESH) code. During the execution of LULESH, a thread in TAUflow would periodically awaken and submit all of TAU's observed performance metrics into the SOSflow system.

2. **Synthetic Workflow** : Synthetic parallel MPI applications were developed that create example workloads for the SOSflow system by publishing values through the API at configurable sizes and rates of injection.

2.5.3 Evaluation of SOS Model. This experiment was performed to validate the SOS Model and demonstrate its applicability for the general case of workflow observation. The Cori supercomputer was used to execute a LULESH + TAUflow simulation. Power and memory usage metrics were collected and stored in SOSflow for each node. During the execution of the workflow, a visualization application was launched from outside of the job allocation which connected to SOSflow's online database and was able to query and display graphs of the metrics that SOSflow had gathered.

The LULESH job was run both with and without the presence of SOSflow (all other settings being equal) in order to validate the ability of SOSflow to meet its design goals while being minimally invasive.

2.5.4 Evaluation of Latency. Experiments were performed to study the latency of data moving through SOSflow. When a value is published from a

client into SOSflow, it enters an asynchronous queue scheme for both database injection and off-node transport to an aggregation target. Latency in this context refers to the amount of time that a value spends in these queues before becoming available for query by analytics modules. To study latency we ran experiments on both ACISS and Catalyst.

Tests run on ACISS were deployed with the Torque job scheduler as MPICH2 MPI applications at scales ranging from 3 to 24 nodes, serving 10 Synthetic Workflow processes per node in all cases. The ACISS battery of runs were tuned as stress tests to ensure that the sosd daemons could operate under reasonably heavy loads. In the 24-node ACISS experiment (Figure 9), SOSflow clients published 72,000,000 double-precision floats with associated metadata during a 90 second window containing three rounds of extremely dense API calls.

Latency tests were performed on LLNL's Catalyst machine at various scales up to 128 nodes, with 8 data sources contributing concurrently on each node in each case. Catalyst's tests measured the latency introduced by sweeping across three different parameters:

- Count of unique values per publish
- Number of publish operations per iteration
- Delay between calls to the publish API

Unlike the ACISS experiments, the Catalyst tests did not attempt to flood the system with data, but rather aimed to observe how slight adjustments in size and rates of value injection would impact the latency of those values.

2.5.5 Results.

2.5.5.1 *SOS Model Validation.*

SOSflow was able to efficiently process detailed performance information from multiple sources on each node. During the LULESH run, SOSflow's online database successfully serviced queries on-line, and the results were plotted as an animated live view of the performance of the workflow. The cost of using SOSflow was calculated simply as the increase

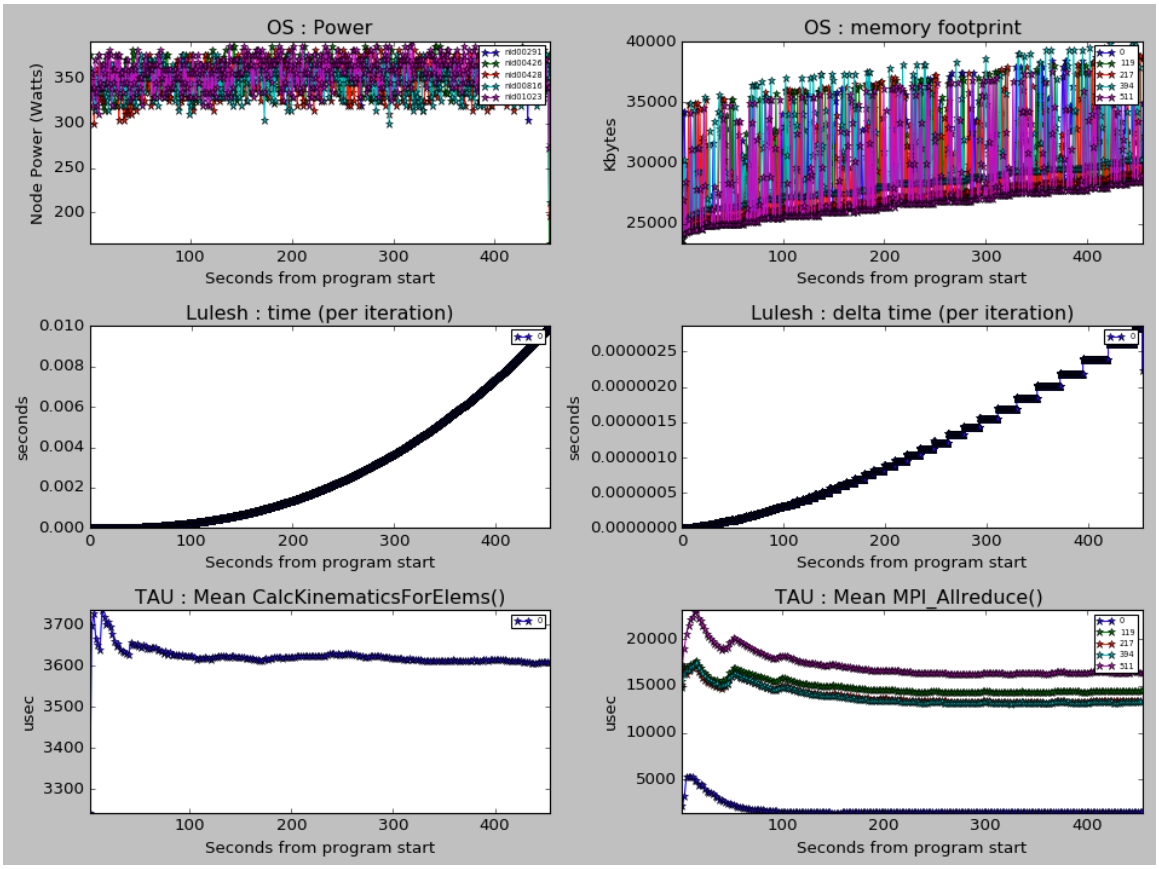


Figure 4. On-line Workflow Performance Visualization Using SOSflow on Cori. Live View of 512 Processes From Three Perspectives: OS, LULESH, TAU

in walltime for LULESH + SOSflow, expressed as a percentage of the walltime of LULESH by itself. The results of these runs are shown in Figure 5.

2.5.5.2 *Evaluation of Latency.*

The on-node (Figure 6) and aggregate (Figure 7) results from the largest 128-node runs are presented here. Results from smaller runs are omitted for space, as they show nothing new: "Time

Nodes	# Proc.	SOS	Exec Time	Iteration Count	SOS Values	SOS Cost
2	8	-	76.50	2031	126,984	3.13%
		SOS	78.90			
3	64	-	217.95	4264	2,861,617	1.84%
		SOS	221.95			
5	125	-	271.84	5382	6,956,037	1.27%
		SOS	275.29			
8	216	-	337.38	6499	14,938,528	1.18%
		SOS	341.36			
12	342	-	399.31	7624	28,336,968	2.34%
		SOS	408.65			
17	512	-	467.45	8741	48,943,001	2.00%
		SOS	476.81			

Figure 5. Percent Increase in LULESH Running Time When SOSflow is Used

in flight" queue latency at smaller scales linearly approached the injection latency figures for a single (on-node) database.

In the 128-node runs, across all configurations, the mean latency observed was 0.3 seconds (and a maximum of 0.7 seconds) for a value, and its full complement of metadata and timestamps, to migrate from one of 1,024 processes to the off-node aggregate data store, passing through multiple asynchronous queues and messaging systems on 128 nodes.

The in situ and aggregate results in Figures 6 and Figure 7 are promising, given the research version of SOSflow being profiled is not optimized. Exploring the optimal configuration and utilization of SOSflow is left to future research effort.

2.5.6 Discussion. Many of the behavioral characteristics of SOSflow are the product of its internal parameters and the configuration of its runtime deployment, rather than products of its data model and algorithms. For now, the effort was made to select reasonable default SOSflow configuration parameters

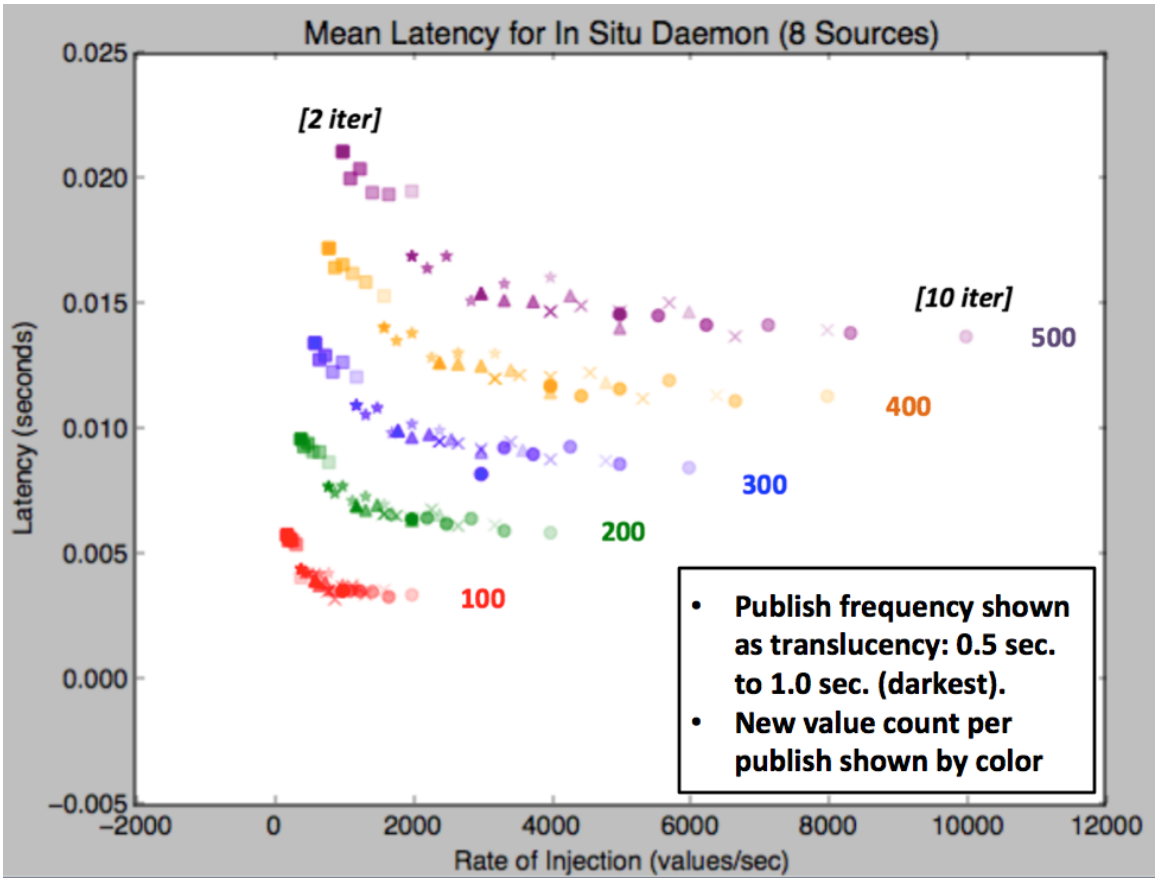


Figure 6. Average Latency for In Situ Database (128 nodes on Catalyst)

and typical/non-priviledged cluster queues and topologies. Because of the general novelty of the architecture, the results presented here could be considered the *performance baseline* for SOSflow to improve on as the research matures.

Expanding on the direct experimental results, here are some additional experiences and observations about the behavior of SOSflow: —

2.5.6.1 Aggregation Topology. The current version of SOSflow is configured at launch with a set number of aggregator databases. The validation tests on ACISS used 3 sosd_db instances to divide up the workload, while the TAUflow + LULESH experiments on Cori used a single aggregator. The parameter sweeps run on the LLNL Catalyst machine were done with four sosd_db aggregation

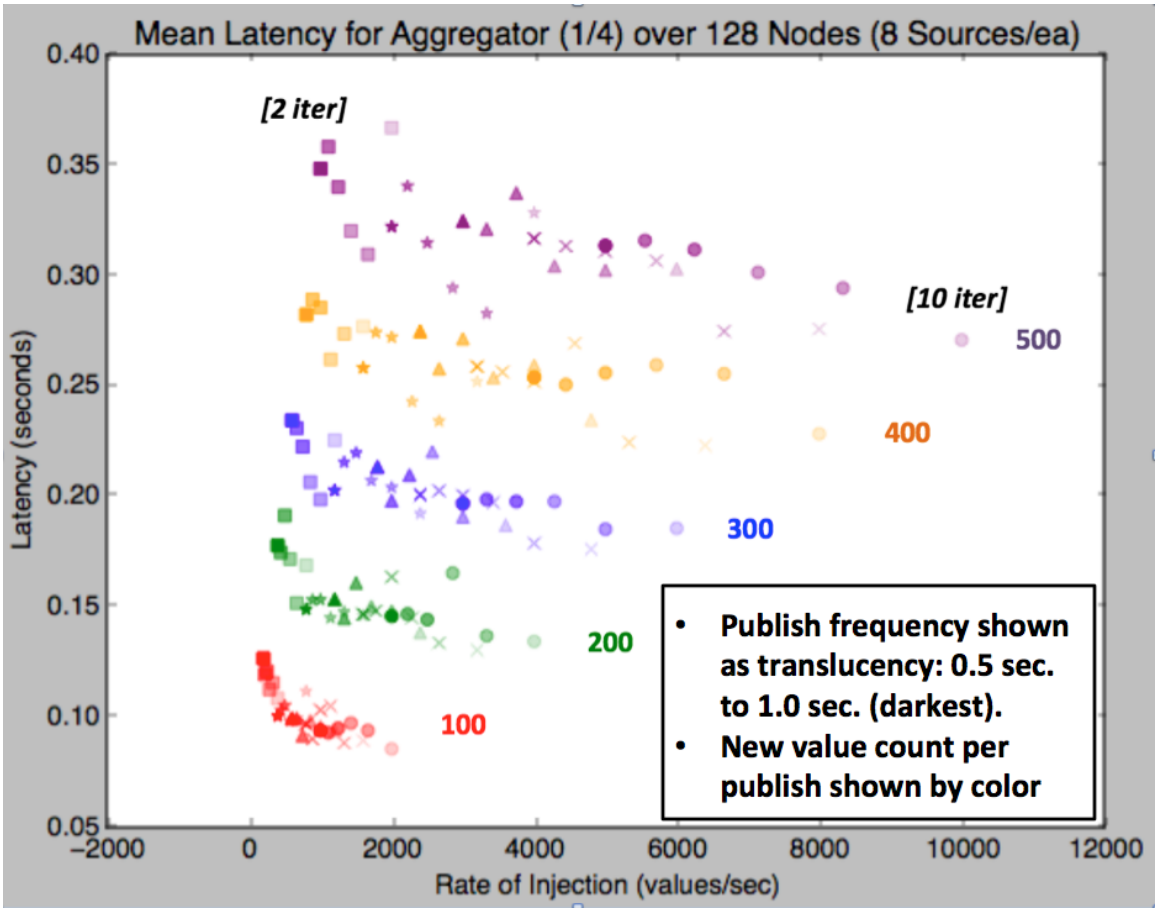


Figure 7. Average Latency for Aggregate Database (128 nodes on Catalyst)

targets at 128 nodes. Tests on ACISS and Catalyst were exploring the latency of data movement through SOSflow, and so both configurations featured dedicated nodes for sosd_db aggregators to avoid contention with other on-node work. The Cori runs captured a real application's behavior, and was primarily intended to demonstrate the fitness of SOSflow for capturing the performance of a scientific workflow along with meaningful context. Instances of aggregators can be spawned, as many as needed, in order to support the quantity of data being injected from a global perspective. All data sent to SOSflow is tagged with a GUID. This allows for shards of the global information space to be concatenated after the run concludes without collision of identities wiping out distinct references.

The data handling design trade-offs made for SOSflow do not prioritize the minimization of latency, but focus rather on gracefully handling spikes in traffic by growing (and then shrinking) the space inside the asynchronous message queues. After a value is passed to SOSflow, it is guaranteed to find its way into the queryable data stores, and there are timestamps attached to it that capture the moment it was packaged into a publication handle in the client library, the moment it was published to the daemon, and even the moment it was finally spooled out into the database. Once it is in the database, it is trivial to correlate values together based on the moment of their creation, no matter how long the value was sequestered in the asynchronous queues.

During the ACISS stress-tests, values were injected into the SOSflow system faster than they could be spooled from the queues into the database. While every value will eventually be processed and injected into the data store, some values wound up having to wait longer than others as the queue depth increased. The asynchronous queues have thread-sequential FIFO ordering, but because the MPI messages are queued up based on their arrival time, and a batch is handled completely before the next is processed, there is no real-time interleaving of database value injections, they are injected in batches. Near the bottom of the pile of MPI messages, the latency continually increases until that batch is injected. This accounts for the observed saw-tooth pattern of increasing latency seen in Figure 9, which is not seen in Figure 8.

2.5.6.2 Time Cost of Publish API. As an accessory to the study of value latency, the length of time that a client application will block inside of an SOSflow API routine was also evaluated. In situ interactions between libsos routines and the daemon are nearly constant time operations regardless of the

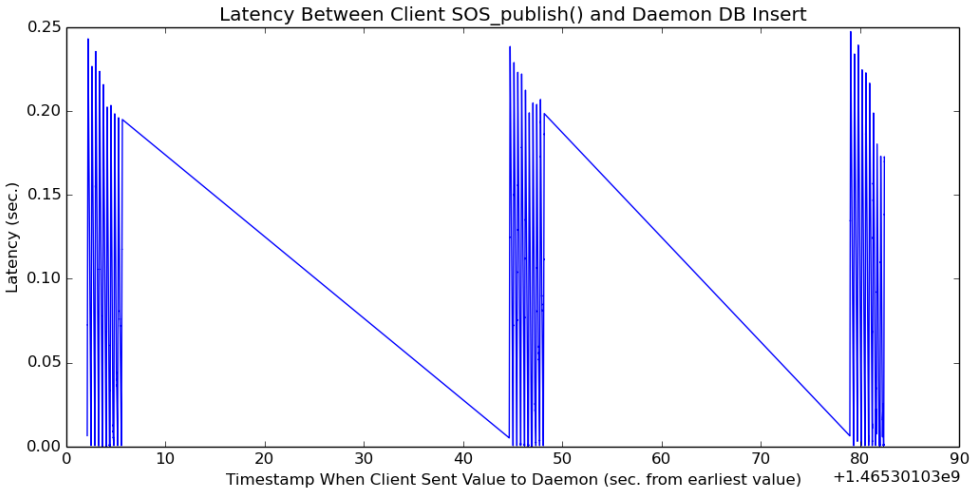


Figure 8. In Situ Latency (24 nodes on ACISS, 240 Applications)

daemon's workload. Care was taken in the daemon's programming to prioritize rate of message ingestion over immediacy of message processing so that SOSflow API calls would not incur onerous delays for application and tool developers. The constancy of message processing speed is shown in figures 10 and 11, where the round trip time (RTT) of a probe message between a client and the daemon (blue) is projected over a graph of the number of new messages arriving in a sample window (red).

This information was gathered by sending 9000+ probe messages over a 15 minute window, with a single sosd_listener rank processing an average of 724 client messages a second in total, arriving from four different processes on an 8-way Xeon node. The messages from SOS clients contained more than 14.7 GB of data, averaging to 338kB per message. Though there are a few spikes in the probe message RTT visible in Figure 10, they are likely not related to SOSflow at all, as Figure 11 reveals in detail. The RTT holds steady during low and high

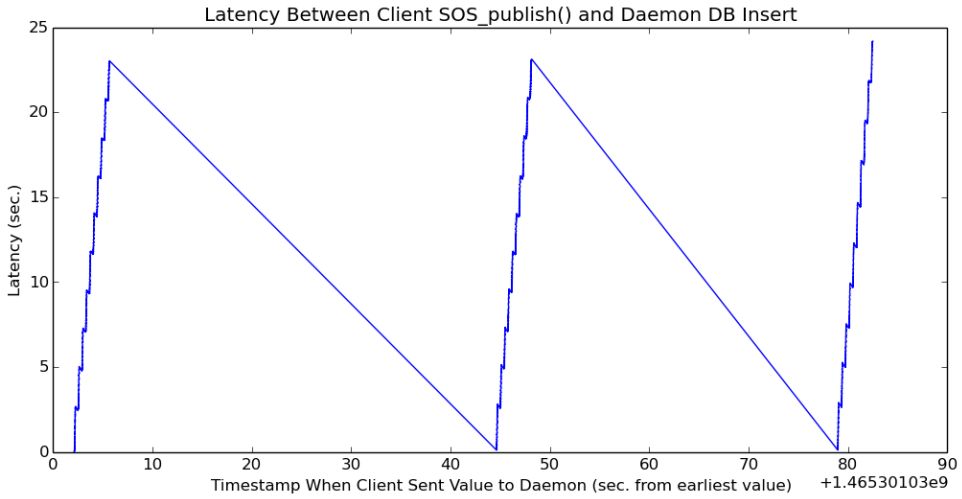


Figure 9. Aggregate Latency (24 nodes on ACISS, 240 Applications)

volume of traffic from the other in situ client processes. The mean RTT for the probe messages was 0.003 seconds, and the maximum RTT was 0.07 seconds.

These results show that the cost of making SOSflow API calls is relatively low, and holds constant under changing sosd_listener workload.

2.6 Conclusion

The SOS Model presented is online, scalable and supports a global information space. SOS enables online in situ characterization and analysis of complex high-performance computing applications. SOSflow is contributed as an implementation of SOS. SOSflow provides a flexible research platform for investigating the properties of existing and future scientific workflows, supporting both current and future scales of execution. Experimental results demonstrated that SOSflow is capable of observation, introspection, feedback and control of complex scientific workflows, and that it has desirable scaling properties.

As part of future development, we aim to continue refining and expanding the core SOSflow libraries and the SOS model. The SOSflow codes can be

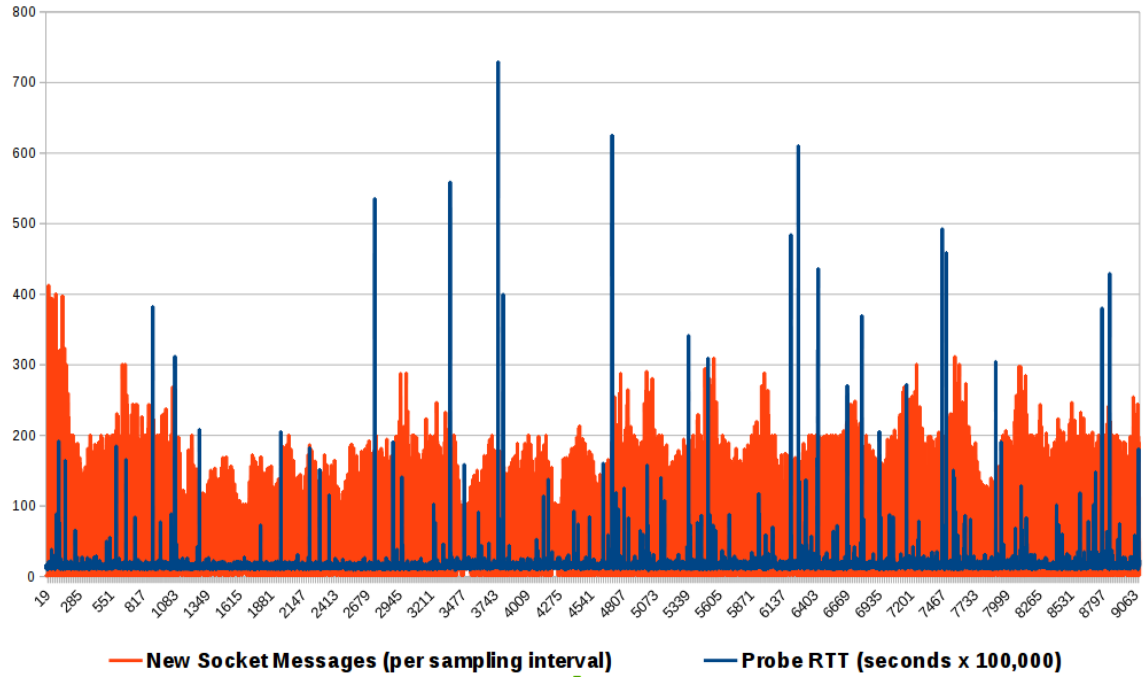


Figure 10. SOSflow Socket Communication Cost, Projected Over Message Count

optimized for memory use and data latency. Mechanisms can be added for throttling of data flow to increase reliability in resource-constrained cases. Subsequent work will map out best-fit metrics for dedicating in situ resources to monitoring platforms for the major extant and proposed compute clusters. Additionally, we plan on exploring options for deployment and integration with existing HPC monitoring and analytics codes at LLNL and other national laboratories.

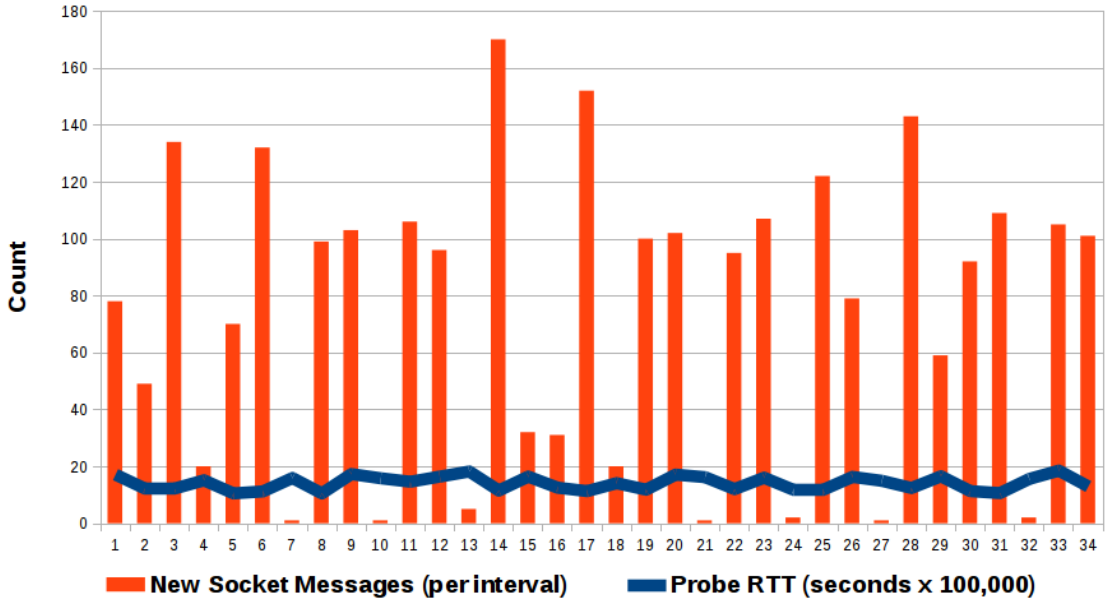


Figure 11. SOSflow Socket Communication Cost (Detail)

CHAPTER III

MULTI-DOMAIN INSIGHTS USING AN OBSERVATION SERVICE

3.1 Introduction

Projecting application and performance data onto the scientific domain allows for the behavior of a code to be perceived in terms of the organization of the work it is doing, rather than the organization of its source code. This perspective can be especially helpful [91] for domain scientists developing aspects of a simulation primarily for its scientific utility, though it can also be useful for any HPC developer engaged with the general maintenance requirements of a large and complicated codebase [104].

There have been practical challenges to providing these opportunities for insight. Extracting the spatial descriptions from an application traditionally has relied on hand-instrumenting codes to couple a simulation's geometry with some explicitly defined performance metrics. Performance tool wrappers and direct source-instrumentation need to be configurable so that users can disable their invasive presence during large production runs. Because it involves changes to the source code of an application, enabling or disabling the manual instrumentation of a code often involves full recompilation of a software stack. Insights gained by the domain projection are limited to what was selected a priori for contextualization with geometry.

Without an efficient runtime service providing an integrated context for multiple sources of performance information, it is difficult to combine performance observations across several components during a run. Further limiting the value of the entire exercise, performance data collected outside of a runtime service must wait to be correlated and projected over a simulation's geometry during

post-mortem analysis. Projections that are produced offline cannot be used for application steering, online parameter tuning, or other runtime interactions that include a human in the feedback loop. Scalability for offline projections also becomes a concern, as the potentially large amount of performance data and simulation geometry produced and operated over in a massively parallel cluster now must be integrated and rendered either from a single point or within an entirely different allocation.

The overhead of manually instrumenting large complex codes to extract meaningful geometries for use in performance analysis, combined with the limited value of offline correlation of a fixed number of metrics, naturally limited the usage of scientific domain projections for gaining HPC workflow performance insights.

3.1.1 Research Contributions. This paper describes the use of SOSflow [57] and ALPINE to overcome many prior limitations to projecting performance into the scientific domain. The methods used to produce our results can be implemented in other frameworks, though SOSflow and ALPINE, discussed in detail in later sections, are generalized and intentionally engineered to deliver solutions of the type presented here. This research effort achieved the following:

- Eliminate the need to manually capture geometry for performance data projections of ALPINE-enabled workflows
- Provide online observation of performance data projected over evolving geometries and metrics
- Facilitate interactive selection of one or many performance metrics and rendering parameters, adding dynamism to projections
- Enable simultaneous online projections from a common data source

– In situ performance visualization architecture supporting both current and future-scale systems

3.2 Related Work

Husain and Gimenez's work on Mitos [105] and MemAxes [106] is motivated similarly to ours. Mitos provides an integration API for combining information from multiple sources into a coherent memoized set for analysis and visualization, and MemAxes projects correlated information across domains to explore the origins of observed performance. SOSflow is being used in our research as an integration API, but takes a different optimization path by providing a general-purpose in situ (online) runtime.

Caliper by Boehme et al. [41] extracts performance data during execution in ways that serve a variety of uses, in much the same way our efforts here are oriented. Caliper's flexible data aggregation [107] model can be used to filter metrics in situ, allowing for tractable volumes of performance data to be made available for projections. Both ALPINE and Caliper provide direct services to users, also serving as integration points for user-configurable services at run time. Caliper is capable of deep introspection on the behavior of a program in execution, yet is able to be easily disabled for production runs that require no introspection and want to minimize instrumentation overhead. ALPINE allows for visualization filters to be compiled separately from a user's application and then introduced into, or removed from, an HPC code's visualization pipeline with a simple edit to that workflow's ALPINE configuration file. More tools like Caliper and ALPINE, featuring well-defined integration points, are essential for the wider availability of cross-domain performance understanding. SOSflow does not collect source-level performance metrics directly, but rather brings that data from tools like Caliper

into a holistic online context with information from other libraries, performance tools, and perspectives.

BoxFish [108] also demonstrated the value of visualizing projections when interpreting performance data, adding a useful hierarchical data model for combining visualizations and interacting with data.

SOSflow's flexible model for multi-source online data collection and analysis provides performance exploration opportunities using both new and existing HPC tools.

3.3 SOSflow

SOSflow provides a lightweight, scalable, and programmable framework for observation, introspection, feedback, and control of HPC applications. The Scalable Observation System (SOS) performance model used by SOSflow allows a broad set of in situ (online) capabilities including remote method invocation, data analysis, and visualization. SOSflow can couple together multiple sources of data, such as application components and operating environment measures, with multiple software libraries and performance tools. These features combined to efficiently create holistic views of workflow performance at runtime, uniting node-local and distributed resources and perspectives. SOSflow can be used for a variety of purposes:

- Aggregation of application and performance data at runtime
- Providing holistic view of multi-component distributed scientific workflows
- Coordinating in situ operations with global analytics
- Synthesizing application and system metrics with scientific data for deeper performance understanding

– Extending the functionality of existing HPC codes using in situ resources

– Resource management, load balancing, online performance tuning, etc.

To better understand the role played by SOSflow, it is useful to examine its architecture. SOSflow is composed of four major components:

– **sosd** : Daemons

– **libsos** : Client Library

– **pub/sql** : Data

– **sosa** : Analytics & Feedback

These components work together to provide extensive runtime capabilities to developers, administrators, and application end-users. SOSflow runs within a user's allocation, and does not require elevated privileges for any of its features.

3.3.1 SOSflow Daemons. Online functionality of SOSflow is enabled by the presence of a user-space daemon. This daemon operates completely independently from any applications, and does not connect into or utilize any application data channels for SOSflow communications. The SOSflow daemons are launched from within a job script, before the user's applications are initialized. These daemons discover and communicate amongst each other across node boundaries within a user's allocation. When crossing node boundaries, SOSflow uses the machine's high-speed communication fabric. Inter-node communication may use either **MPI** or **EVPath** as needed, allowing for flexibility when configuring its deployment to various HPC environments.

The traditional deployment of SOSflow will have a single daemon instance running in situ for each node that a user's applications will be executing on

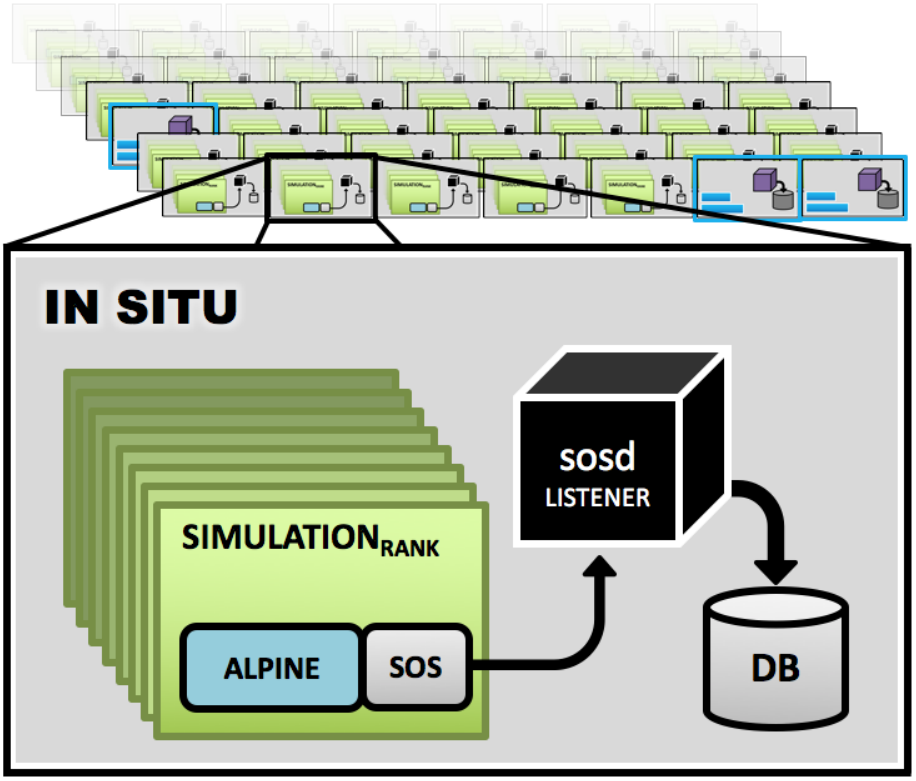


Figure 12. SOSflow's lightweight daemon runs on each node.

(Figure 12). This daemon is called the **listener**. Additional resources can be allocated in support of the SOSflow runtime as-needed to support scaling and to minimize perturbation of application performance. One or more nodes are usually added to the user's allocation to host SOSflow **aggregator** daemons that combine the information that is being collected from the in situ daemons. These aggregator daemons are useful for providing holistic unified views at runtime, especially in service to online analytics modules. Because they have more work to do than the in situ listener daemons, and also are a useful place to host analytics modules, it is advisable to place aggregation targets on their own dedicated node[s], co-located with online analytics codes.

3.3.1.1 In Situ.

Data coming from SOSflow clients moves into the in situ daemon across a light-weight local socket connection. Any software that connects in to the SOSflow runtime can be thought of as a client. Clients connect only to the daemon that is running on their same node. No client connections are made across node boundaries, and no special permissions are required to use SOSflow, as the system considers the SOSflow runtime to be merely another part of a user's workflow.

The in situ listener daemon offers the complete functionality of the SOSflow runtime, including online query and delivery of results, feedback, or application steering messages. At startup, the daemon creates an in-memory data store with a file-based mirror in a user-defined location. Listeners asynchronously store all data that they receive into this store. The file-based mirror is ideal for offline analysis and archival. The local data store can be queried and updated via the SOSflow API, with all information moving over the daemon's socket, avoiding dependence on filesystem synchronization or centralized metadata services.

Providing the full spectrum of data collected on node to clients and analytics modules on node allows for distributed online analytics processing. Analytics modules running in situ can observe a manageable data set, and then exchange small intermediate results amongst themselves in order to compute a final global view. SOSflow also supports running analytics at the aggregation points for direct query and analysis of global or enclave data, though it is potentially less scalable to perform centrally than in a distributed fashion, depending on the amount of data being processed by the system.

SOSflow's internal data processing utilizes unbounded asynchronous queues for all messaging, aggregation, and data storage. Pervasive design around

asynchronous data movement allows for the SOSflow runtime to efficiently handle requests from clients and messaging between off-node daemons without incurring synchronization delays. Asynchronous in situ design allows the SOSflow runtime to scale out beyond the practical limits imposed by globally synchronous data movement patterns.

3.3.1.2 Aggregation Targets. A global perspective on application and system performance is often useful. SOSflow automatically migrates information it is given into one or more aggregation targets. This movement of information is transparent to users of SOS, requiring no additional work on their part. Aggregation targets are fully-functional instances of the SOSflow daemon, except that their principle data sources are distributed listener daemons rather than node-local clients. The aggregated data contains identical information as the in situ data stores, it just has more of it, and it is assembled into one location. The aggregate daemons are useful for performing online analysis or information visualization that needs to include information from multiple nodes (Figure 13).

SOSflow is not a publish-subscribe system in the traditional sense, but uses a more scalable push-and-pull model. Everything sent into the system will automatically migrate to aggregation points unless it is explicitly tagged as being node-only. Requests for information from SOSflow are ad hoc and the scope of the request is constrained by the location where the request is targeted: in situ queries are resolved against the in situ database, aggregate queries are resolved against the aggregate database. If tagged node-only information is potentially useful for offline analysis or archival, the in situ data stores can be collected at the end of a job script, and their contents can be filtered for that node-only information, which can be simply concatenated together with the aggregate database[s] into a complete

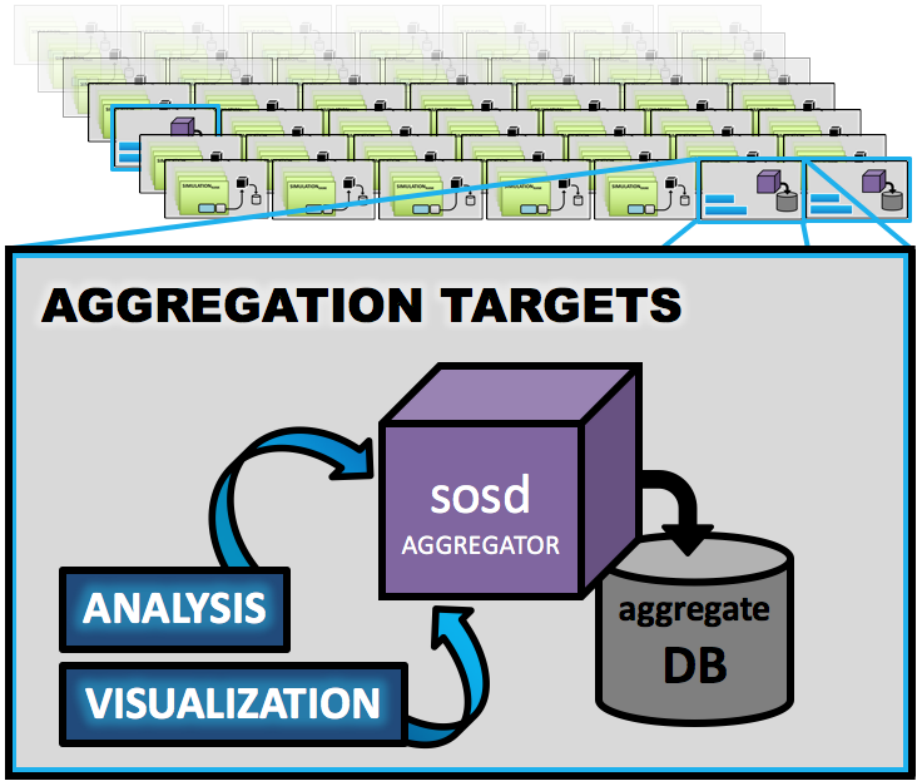


Figure 13. Co-located analysis and visualization with aggregation.

image of all data. Each value published to SOSflow is tagged with a globally unique identifier (GUID). This allows SOSflow data from multiple sources to be mixed together while preserving its provenance and preventing data duplication or namespace collision.

3.3.2 SOSflow Client Library. Clients can directly interface with the SOSflow runtime system by calling a library of functions (libsos) through a standardized API. Applications can also transparently become clients of SOS by utilizing libraries and performance tools which interact with SOSflow on their behalf. All communication between the SOSflow library and daemon are transparent to users. Users do not need to write any socket code or introduce any state or additional complexity to their own code.

Information sent through the libsos API is copied into internal data structures, and can be freed or destroyed by the user after the SOSflow API function returns. Data provided to the API is published up to the in situ daemon with an explicit API call, allowing developers to control the frequency of interactions with the runtime environment. It also allows the user to register callback functions that can be triggered and provided data by user-defined analytics function, creating an end-to-end system for both monitoring as well as feedback and control.

To maximize compatibility with extant HPC applications, the SOSflow client library is currently implemented in C99. The use of C99 allows the library to be linked in with a wide variety of HPC application codes, performance tools, and operating environments. There are various custom object types employed by the SOSflow API, and these custom types can add a layer of complexity when binding the full API to a language other than C or C++. SOSflow provides a solution to this challenge by offering a "Simple SOS" (ssos) wrapper around the full client library, exposing an API that uses no custom types. The ssos wrapper was used to build a native Python module for SOSflow. Users can directly interact with the SOSflow runtime environment from within Python scripts, acting both as a source for data, and also a consumer of online query results. HPC developers can capitalize on the ease of development provided by Python, using SOSflow to observe and react online to information from complex legacy applications and data models without requiring that those applications be redesigned to internally support online interactivity.

3.3.3 SOSflow Data. The primary concept around which SOSflow organizes information is the "publication handle" (pub). Pubs provide a private

namespace where many types and quantities of information can be stored as a key/value pair. SOSflow automatically annotates values with a variety of metadata, including a GUID, timestamps, origin application, node id, etc. This metadata is available in the persistent data store for online query and analysis. SOSflow's metadata is useful for a variety of purposes:

- Performance analysis
- Provenance of captured values for detection of source-specific patterns of behavior, failing hardware, etc.
- Interpolating values contributed from multiple source applications or nodes
- Re-examining data after it has been gathered, but organizing the data by metrics other than those originally used when it was gathered

A complete history of changes to every value is retained within the daemon's persistent data store.This allows for the changing state of an application or its environment to be explored at arbitrary points in its evolution. When a key is re-used to store some new information that has not yet been transmitted to the in situ daemon, the client library enqueues it up as a snapshot of that value, preserving all associated metadata alongside the historical value. The next time the client publishes to the daemon, current new values and all enqueued historical values are transmitted.

SOSflow is built on a model of a global information space. Aggregate data stores are guaranteed to provide eventual consistency with the data stores of the in situ daemons that are targeting them. SOSflow's use of continuous but asynchronous movement of information through the runtime system does not allow for strict quality-of-service guarantees about the timeliness of information being

available for analysis. This design constraint reflects the reality of future-scale HPC architectures and the need to eliminate dependence on synchronous behavior to correlate context. SOSflow conserves contextual metadata when values are added inside the client library. This metadata is used during aggregation and query resolution to compose the asynchronously-transported data according to its original synchronous creation. The vicissitudes of asynchronous data migration strategies at scale become entirely transparent to the user.

SOSflow does not require the use of a domain-specific language when pushing values into its API. Pubs are self-defining through use: When a new key is used to pack a value into a pub, the schema is automatically updated to reflect the name and the type of that value. When the schema of a pub changes, the changes are automatically announced to the in situ daemon the next time the client publishes data to it. Once processed and injected into SOSflow's data store, values and their metadata are accessible via standardized SQL queries. SOSflow's online resolution of SQL queries provides a high-degree of programmability and online adaptivity to users. SQL views are built into the data store that mask off the internal schemas and provide results organized intuitively for grouping by application rank, node, time series, etc.

SOSflow uses the ALPINE in situ visualization infrastructure described below to collect simulation geometry that it correlates with performance data.

3.4 ALPINE Ascent

ALPINE is a project that aims to build an in situ visualization infrastructure and analysis targeting leading edge supercomputers. ALPINE is part of the U.S. Department of Energy's Exascale Computing Project (ECP) [109], and the ALPINE effort is supported by multiple institutions. The goal of ALPINE

is two fold. First, create a hybrid-parallel library (i.e., both distributed-memory and shared-memory parallel) that can be included in other visualization tools such as ParaView [110] and VisIt [111] thus creating an ecosystem where new hybrid-parallel algorithms are easily deployed into downstream tools. Second, create a flyweight in situ infrastructure that directly leverages the hybrid-parallel library. In this work, we directly interface with the ALPINE in situ infrastructure called Ascent [93].

Ascent is the descendant of Strawman [112], and Ascent is tightly-coupled with simulations, i.e. it shares the same node resources as the simulation. While Strawman's goal was to bootstrap in situ visualization research, the ALPINE Ascent in situ infrastructure is intended for production. Ascent includes include three physics proxy-applications out of the box to immediately provide the infrastructure and algorithms a representative set of mesh data to consume. Ascent is already integrated into several physics simulations to perform traditional visualization and analysis, and we chose to embed an SOSflow client into Ascent to eliminate the need for additional manual integration of SOSflow with Ascent-equipped simulations. Ascent uses the Conduit [113] data exchange library to marshal mesh data from simulations into Ascent. Conduit provides a flexible hierarchical model for describing mesh data, using a simple set of conventions for describing meshes including structured, unstructured, and higher order element meshes [114]. Once the simulation describes the mesh data, it publishes the data into Ascent for visualization purposes. Ascent relays the mesh data to SOSflow in the manner described below. In addition to the mesh data, we can easily add performance data that is associated with each MPI rank. Coupling the performance data with the mesh geometry provides a natural way to generate an aggregate data

set to visualize the performance data mapped to the spatial region each MPI rank is responsible for.

Ascent includes Flow, a simple dataflow library based on the Python dataflow library within VisIt, to control the execution of visualization filters. The input to Flow is the simulation mesh data, and Ascent adds visualization filters (e.g., contours and thresholding) to create visualizations. Everything within Flow is a filter that can have multiple inputs and a single output of generic types. The flexibility of Flow allows for user defined filters, compiled outside of Ascent, to be easily inserted into the dataflow, and when the dataflow network executes, custom filters have access to all of the simulation mesh data published to Ascent. We leverage the flexibility of Flow to create an SOSflow filter that is inserted at runtime. The SOSflow filter uses the data published by the simulation to extract the spatial extents being operated over by each MPI rank along, with any performance data provided. Next, we publish that data to SOSflow, and then Ascent's visualization filters execute as usual.

3.5 Experiments

3.5.1 Evaluation Platform. All results were obtained by running online queries against the SOSflow runtime's aggregation targets (Figure 13) using SOSflow's built-in Python API. The results of these queries were used to create Vtk [115] geometry files. These files were used as input for the VisIt visualization tool, which we invoked from within the allocation to interactively explore the performance projections.

3.5.2 Experiment Setup. The experiments performed had the following purposes:

– **Validation** : Demonstrate the coupling of SOSflow with ALPINE and its ability to extract geometry from simulations transparently.

– **Introspection** : Examine the overhead incurred by including the SOSflow geometry extraction filter in an ALPINE Ascent visualization pipeline.

ALPINE's Ascent library was used to build a filter module outfitted with SOSflow, and this filter was used for online geometry extraction (Figure 14). ALPINE's JSON configuration file describing the connectivity of the in situ visualization pipeline was modified to insert the SOSflow-equipped geometry extraction filter. The SOSflow implementation used to conduct these experiments is general-purpose

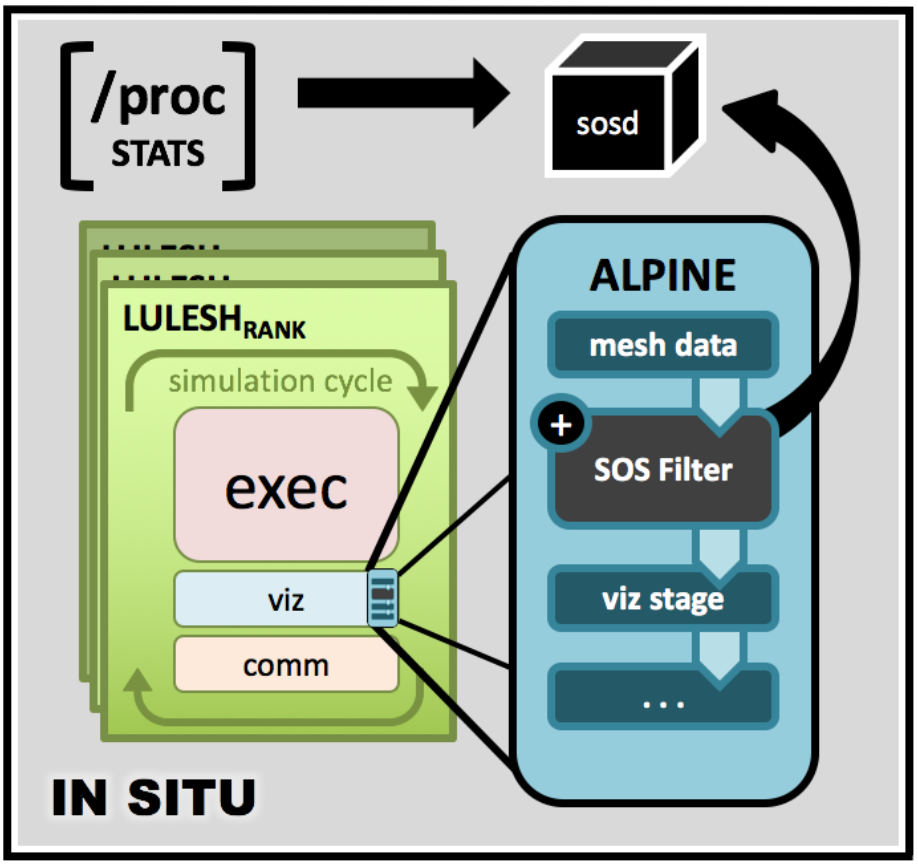


Figure 14. SOSflow collects runtime information to project over simulation geometry.

and was not tailored to the specific deployment environment or the simulations observed. The study was conducted on two machines, the details of which are included here —

1. **Quartz** : A 2,634-node Penguin supercomputer at Lawrence Livermore National Laboratory (LLNL). Intel Xeon E5-2695 processors provide 36 cores/node. Each node offers 128 GB of memory and nodes are connected via Intel OmniPath.

2. **Catalyst** : A Cray CS300 supercomputer at LLNL. Each of the 324 nodes is outfitted with 128 GB of memory and 2x Intel Xeon E5-2695v2 2.40 GHz 12-core CPUs. Catalyst nodes transport data to each other using a QLogic InfiniBand QDR interconnect.

The following simulated workflows were used —

1. **KRIPKE** [116] : A 3D deterministic neutron transport proxy application that implements a distributed-memory parallel sweep solver over a rectilinear mesh. At any given simulation cycle, there are simultaneous sweeps along a set of discrete directions to calculate angular fluxes. This results in a MPI communication pattern where ranks receive asynchronous requests from other ranks for each discrete direction.

2. **LULESH** [117] : A 3D Lagrangian shock hydrodynamics proxy application that models Sedov blast test problem over a curvilinear mesh. As the simulation progresses, hexahedral elements deform to more accurately capture the problem state.

3.5.3 Overview of Processing Steps. The SOSflow runtime provided a modular filter for the ALPINE in situ visualization framework. This

filter was enabled for the simulation workflow at runtime to allow for the capture of evolving geometric details as the simulation progressed. The SOSflow runtime daemon automatically contextualized the geometry it received alongside the changing application performance metrics. SOSflow's API for Python was used to extract both geometry information and correlated performance metrics from the SOSflow runtime. This data set was used to generate sequences of input files to the VisIt scientific data visualization tool corresponding to the cycle of a the distributed simulation.

Each input file contained the geometric extents of every simulation rank, the portion of the simulated space that each part of the application was working within. Alongside that volumetric descriptions for that cycle, SOSflow integrated attribute dictionaries of all plottable numeric values it was provided during that cycle, grouped by simulation rank. Performance metrics could then be interactively selected and combined in VisIt with customizable plots, presenting an application rank's state and activity incident to its simulation effort, projected over the relevant spatial extent.

3.5.4 Evaluation of Geometry Extraction. Our experiments were validated by comparing aggregated data to data manually captured at the source during test runs. Furthermore, geometry aggregated by ALPINE's Ascent SOSflow filter was rendered and visually compared with other visualizations of the simulation. Projections were inspected to observe the simulation's expected deforming of geometry (LULESH) or algorithm-dependent workload imbalances (KRIPKE). Performance metrics can be correlated in SQL queries to the correct geometric regions by various redundant means such as pub handle GUID, origin PID or MPI rank, simulation cycle, host node name, SOSflow publish frame, and

value creation timestamps. Aggregated performance metrics projected over the simulation regions were compared to metrics reported locally, and required to be identical for each region and simulation cycle.

3.5.5 Evaluation of Overhead. Millisecond-resolution timers were added to the per-cycle *execute* method of the SOSflow Alpine geometry extraction filter. Each rank tracked the amount of time it spent extracting its geometry, packing the geometry into an SOSflow pub handle, and transmitting it to the runtime daemon. Every cycle's individual time cost was computed and transmitted to SOSflow, as well as a running total of the time that Alpine had spent in the SOSflow filter. From a region outside the timers, the timer values were packed into the same SOSflow publication handle used for the geometric data. Timer values were transmitted at the end of the following cycle, alongside that cycle's geometry. The additional transmission cost of these two timer values once per simulation cycle had no perceivable impact on the performance they were measuring.

3.6 Results

Geometry was successfully extracted (Figures 15, 16, 17, and 18) with minimal overhead from simulations run at a variety of scales from 2 to 33 nodes. The side-by-side introspection of the behavior of KRIPKE (Figure 15) are a good example of the value this system provides to developers. The amount of work loops and the backlog of requests for computation are correlated negatively, with ranks operating in the center of the simulation space getting through less loops of work per cycle, since they are required to service data requests in more directions than the ranks simulating the corners regions. The directionality of energy waves moving through he simulated space can also be observed, with more work piling up where multiple waves are converging. A developer can quickly assess the behavior of their

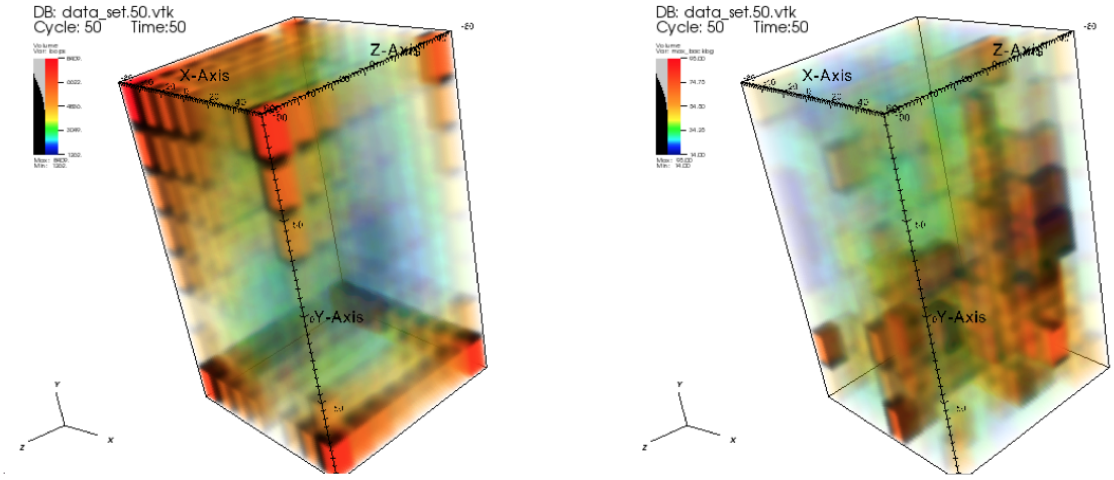


Figure 15. Loops (left) and maximum backlog (right) from one cycle of 512 KRIPKE ranks distributed to 32 nodes.

distributed algorithm by checking for hot-spots and workload imbalances in the space being simulated.

3.6.1 Geometry Extraction and Performance Data Projection. Aggregated simulation geometry was a precise match with the geometry manually recorded within applications, across all runs. After aggregation and performance data projection, geometry from all simulation ranks combined to create a contiguous space without gaps or overlapping regions, representative of the simulated space subdivided by MPI rank.

3.6.2 Overhead. The inclusion of the ALPINE Ascent filter module for SOSflow had no observable impact on overall application execution time, being significantly less than variance observed between experimental runs both with and

without the filter. The filter module is executed at the end of each simulation cycle, from the first iteration through to the simulation conclusion. Manual instrumentation was added to the SOSflow filter to measure the time spent inside the filter's *execute* method, where all simulation geometry and performance metrics were gathered for our study.

When gathering *only the simulation geometry*, filter execution never exceeded 2ms per simulation cycle. We collected performance information for our projections by reading from the /proc/[pid] files of each rank. These readings were made from within the SOSflow filter, and published to SOSflow alongside the collected geometry. Collecting 31 system metrics and application counters added additional overhead, but the filter time but did not exceed 4ms for any of the projections shown in this chapter. The filter's execution time was logged as a performance metric alongside the other in situ performance data, and is visualized for LULESH in Figure 17.

3.7 Conclusion

Services from both SOSflow and ALPINE were successfully integrated to provide a scalable in situ (online) geometry extraction and performance data projection capability.

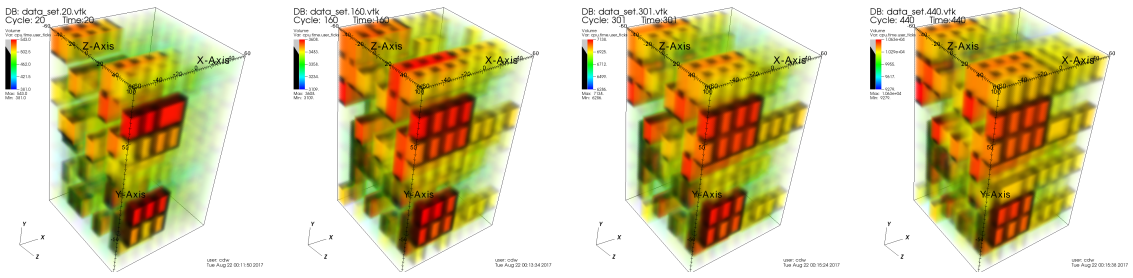


Figure 16. Cumulative user CPU ticks during 440 cycles of 512 KRIPKE ranks on 32 nodes.

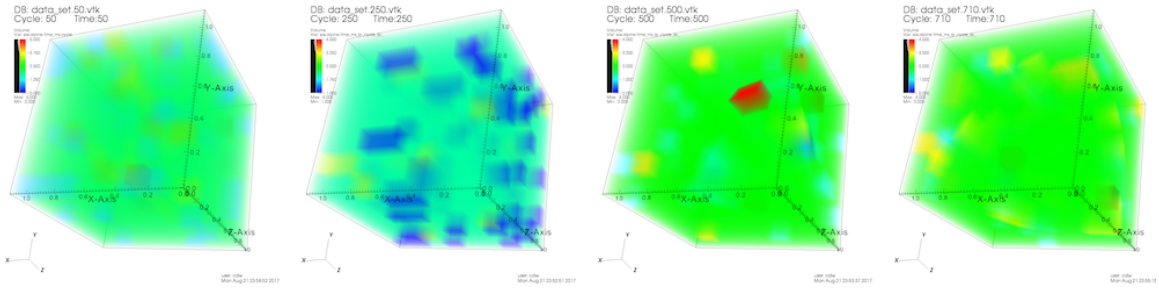


Figure 17. Filter execution (1-4ms) over 710 LULESH cycles.

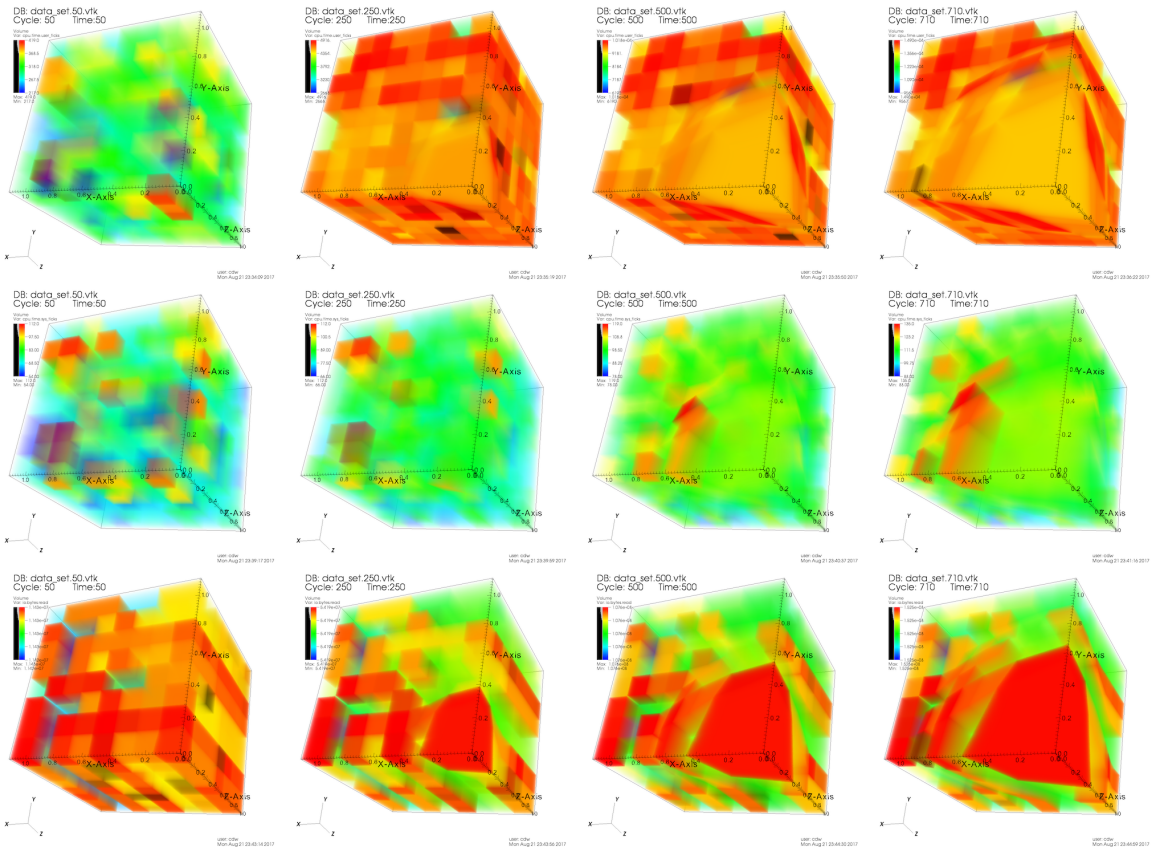


Figure 18. Many metrics can be projected from one run. Here we see (top to bottom) user CPU ticks, system CPU ticks, and bytes read during 710 cycles of 512 LULESH ranks distributed across 32 nodes.

3.7.1 Future Work. Workflows that use the ALPINE framework but have complex irregular meshes, feature overlapping "halo regions", or that operate over non-continuous regions of space within a single process, may require

additional effort to extract geometry from, depending on the organization of spatial descriptions they employ. ALPINE uses the Vtk-m [118] library for its operations over simulation mesh data. The addition of a general convex hull algorithm to Vtk-m will simplify the task of uniformly describing any spatial extent[s] being operated on by a process using ALPINE for its visualization pipeline.

The VisIt UI can be extended to support additional interactivity with the SOSflow runtime. UI elements to submit custom SQL queries to SOSflow would enhance the online data exploration utility of VisIt. SOSflow's interactive code steering mechanisms allow for feedback messages and payloads to be delivered to subscribing applications at runtime. With some basic additions to the VisIt UI, these mechanisms could be triggered by a VisIt user based on what they observe in the performance projections, sending feedback to targeted workflow components from within the VisIt UI.

While the geometry capture and performance data projection in this initial work has a scalable in situ design, the final rendering of the performance data into an image takes place on a single node. Future iterations of this performance visualization work will explore the use of in situ visualization techniques currently employed to render scientific data from simulations [119]. These emerging in situ rendering technologies will allow for live views of performance data projected over simulation geometry at the furthest extreme scales to which our simulations are being pressed.

CHAPTER IV

PARALLEL PORTABILITY WITH ONLINE MACHINE LEARNING

4.1 Introduction

HPC software can contain tens to thousands of parallel code regions, each of which may have independent performance tuning parameters. Optimal choices for these tuning parameters can be specific to a target system architecture, the set of input data to be processed, or the overall shared state of the machine during a job's execution. There are costs associated with discovering and maintaining optimal choices, in a developer's time to manually adjust settings and rebuild projects, or the compute time to explore the space of possible configurations to find optimal settings automatically.

The goal of *performance portability* in HPC is for applications to operate optimally across a range of current and future systems without the need for costly code interventions in each new deployment. Given large job scales, increasing software complexity, platform diversity, and hardware performance variability, a performance portability is a challenging problem – with the same inputs, code performance is observed to change between invocations on the same machine and, worse, can be variable even during execution.

Recent work has turned to machine learning techniques to train classification models on code and execution feature vectors that then can be used to make dynamic tuning selection for each kernel of interest [120]. For instance, the Apollo [121] work demonstrated the use of offline machine learning methods to optimize the selection of RAJA [122] kernels at runtime. The RAJA programming methodology provides abstractions that allow code regions to be implemented once but compiled for a variety of architectures, with several execution policies capable

of being selected at runtime. Apollo's offline training approach built statistical classifiers that directly selected values for tuning parameters. The classification model could then be embedded in RAJA programs to provide a dynamic, low-overhead, data-driven auto-tuning framework. The decision to do offline training was a trade-off Apollo made to avoid costly online search for autotuning.

Offline machine learning methods are not sufficient for guiding *online optimizations* that deliver general performance portability. There are several reasons for this to be the case:

1. Without knowing what the user is actually doing, combinatorial exploration of all possible settings is difficult to exhaust, even with a decent sampling strategy. A great many different models need to be represented by whatever ends up being deployed, hopefully providing optimal recommendations for every unique combination of architectures, configurations, input decks, and so on.

2. In order to cover all scenarios, the expense of training and re-training will grow. The entire campaign of parameter testing would need to be done with any new code deployment, significant modification, change in configuration, use of new input deck, or increase in job scale. Certainly, moving to a new platform or modification of an existing platform could trigger a new training study. Ideally, the testing should happen at the full scale and duration that the job was intended to be run at once its model was in use, but this is a costly proposition. Ultimately, this suggests that offline training is unable to fully capture enough for model fitness to be reliable over time.

3. Once trained offline, static models are unable to adapt to changes between application invocations or simulation steps in a workflow. Such changes can make even very good models go stale over time. Furthermore, the potential dynamic variations in the execution environment can expose gaps in the model due to the fact that they never occurred during training.

To further motivate the need for online methods, we note the paradigmatic shift in HPC underway in the move to extreme scales and cloud-based computing. Applications are increasingly being developed and deployed where it is accepted as a given that there will be dynamism in their runtime environment. Even within tightly-controlled on-site dedicated clusters, novel *in situ* resources and services are being deployed in support of classic block-synchronous applications, decreasing the emphasis on their synchronous behavior to maximally saturate available computation and I/O resources.

Our current research is motivated by the need to address tuning challenges presented by these performance complexities and realities of new *in situ* development models: the scale of jobs, asynchronous data movement, and dynamic performance characteristics of modern hardware. Instead of working against the general nature of the problem, we propose to embrace it and investigate the productive outcomes of adopting modern (online) training techniques. In the spirit of prior work, we created the *Artemis* continuous tuning framework to analyze code kernels online during application execution. Artemis trains new kernel performance models *in situ*, deploying and evaluating them at runtime, observing each model's recommendations during execution to rate its ongoing fitness.

Our primary research contributions are:

– We present Artemis, an online framework that dynamically tunes the execution of parallel regions by training optimizing models.

– We provide an implementation of a RAJA parallel execution policy that uses Artemis to optimize the execution of `forall` and `collapse` loop pattern.

– We extend Kokkos to use Artemis for tuning CUDA execution on GPUs.

– We evaluate Artemis using three HPC proxy applications: LULESH, Cleverleaf, and Kokkos Kernels SpMV. Results show that Artemis has overhead of less than 9%, and model training and evaluation overhead is in the order of hundreds of microseconds. Artemis selects the optimal policy $\tilde{8}5\%$ of the time, and can provide up to 47% speedup.

4.2 Background

Parallel programming frameworks have emerged to address the performance portability challenge by providing a "write once, run anywhere" methodology where alternate versions of a code section (called kernels) can be generated to target architectural tuning parameters. In this manner, the programming methodology decouples the specification of a kernel's parallelism from the parameters that govern policies for how to execute the work in different forms. The tuning of the policy choices and execution variants can be done without changing the high-level program.

Parallel frameworks such as RAJA[123] and Kokkos[124][125] use lightweight syntax and standard C++ features for portability and ease of integration into production applications. Related prior work on Apollo[121] focused on developing an autotuning extension for RAJA for input-dependent parameters where the best kernel execution policy depends on information known only at application runtime.

However, Apollo's methodology required executions under all runtime scenarios to create an offline static training database, leading to many of the limitations discussed in the introduction. Thus, it is interesting to pursue a new question: is it possible to train a classification model online and apply it during application execution? Of course, this question immediately raises several concerns, mainly having to do with how training data is generated, the overhead of measurement, and the complexity costs of machine learning algorithms.

4.3 Artemis: Design and Implementation

Artemis is at once a methodology for in situ, ML-based performance auto-tuning and an architecture and operational framework for its implementation. The following captures these aspects as we describe how Artemis actually works. In a nutshell, it is the observation of an application's execution of its tunable parallel code regions, extracting features and performance data with different execution policies, coupled with the training of ML models online to select optimized execution policies per-region and feature set.

Table 1. The Tuning API of Artemis.

Function	**Description**
void Artemis::init()	Initializes Artemis
Artemis::Region *Artemis::create_region(int num_policies)	Creates a tunable region and returns its pointer
void Artemis::Region::begin(vector<float> features)	Marks beginning execution of a tunable region
int Artemis::Region::getPolicyIndex()	Retuns the index of the execution policy for the region
void Artemis::Region::end()	Marks ending execution of a tunable region
void Artemis::processMeasurements()	Triggers Artemis tuning

4.3.1 Design. Without loss of generality, Artemis thinks of applications being iterative where a sequence of *steps* are conducted during which parallel regions are being executed. At the end of those steps, the application ends.

If the a parallel region is to be tuned, it must be provide the different execution policy variants it can choose between, and then Artemis must be invoked for that region. In the case of the reference implementations presented here, this can be largely automated.

The *user* of Artemis need not be thought of as the ultimate end-user of an application, but more likely the developer implementing a performance portability framework such as RAJA or Kokkos within some application. By design, our embedding of an Artemis interface into the portability framework layer enables all parallel regions of an application to be automatically decorated with the necessary Artemis API calls, and furnished with a set of common execution policies that come pre-packaged, and may be integrated into any application making use of that performance portability framework. Artemis is designed to be extensible and programmable, so expert users are always going to be able to provide their own execution policy variants, or make use of the Artemis API directly without the benefits of a performance portability layer managing it.

In the common case where an application is making use of performance portability framework as described above, all an end-user will need to do to is to select to enable Artemis functionality at build time, and then at run time they could opt to enable the Artemis tuning capabilities for any given session, which would then exploit the built-in policies that are bundled with the framework. Essentially, this is the end of involvement for the Artemis user.

Within a step, each parallel region executed is done so for a particular policy as determined by the policy model. Artemis controls how the policy model behaves. It could either be controlled to test out different policies during training, thereby allowing performance measurements to be obtained for analysis, or it could select a particular policy determined by the auto-tuned model evaluation. Each application step represents an opportunity for parallel region training or re-training. Within a step, each encounter with an Artemis-guided parallel region allows that region's model to make an optimized policy selection based on immediately-observed local features.

Artemis instruments parallel regions to collect data on their execution and tune them. Marking the beginning of region execution, the user additionally provides a set of *features* that characterize the execution and a set of execution *policies* that are selectable for the execution of this region. After the call marking the beginning of a region, the user calls the Artemis API function that returns the policy to use when executing the region. The region proceeds to execute a refactored variant of itself that corresponds to that policy selection. Finally, the instrumented region calls the Artemis API to mark the end of its execution, and Artemis makes note of the features and performance measurements. Region execution time is the primary measurement of interest, but it is possible to capture other performance data for analysis.

Artemis is implemented as a runtime library that merges with the application to provide region performance/metadata measurement/analysis, ML model training, and auto-tuning optimization. It presently targets parallel MPI programs that use RAJA or Kokkos for on-node parallelization.

4.3.2 Training and Optimization. The set of user-provided features and policies for each region are the input data to Artemis for ML training and optimization. During training, Artemis explores among the available policies and in particular measures their execution times, which is the optimization target we selected for our experimental evaluation. Artemis keeps per-region records of the feature set, policy, and measured execution time as tuples of (feature set, policy, execution time) to compile the training data and create an optimizing policy selection model. Whenever a region is executed multiple times per step, if different features are captured or policies are explored, each unique combination will have executions times recorded for use in model development.

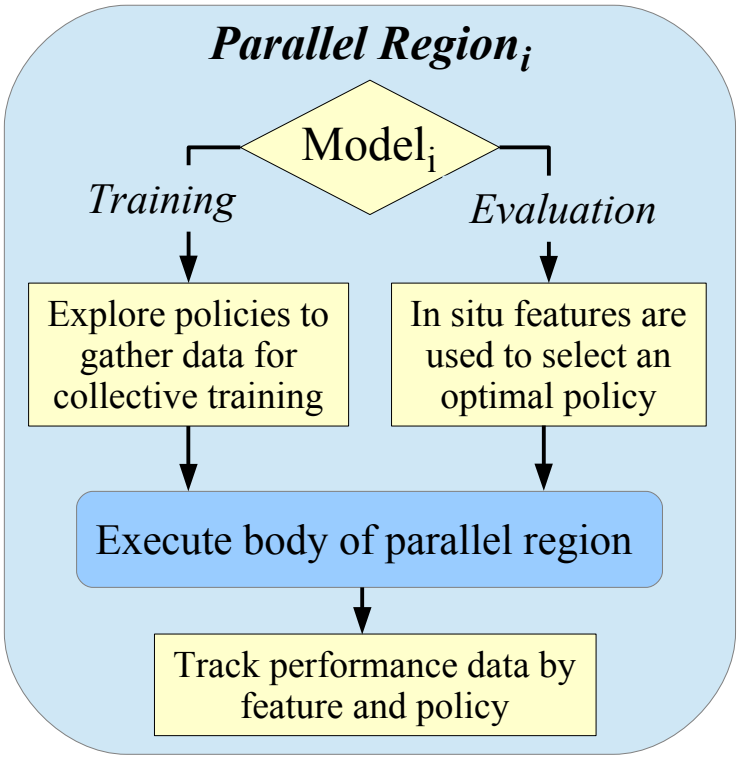


Figure 19. For each parallel code region, Artemis can explore policies to generate training data, or apply in situ observations to a previously-trained model to use the recommended policy.

By design, Artemis exposes an API call to the user to invoke optimization on-demand. Artemis expects the user to invoke the optimization API function after a sensible amount of computation has executed, permitting Artemis to have collected a representative set of measurement records. This can be different for different applications, and depends somewhat on the number of optimization points to be explored when searching the space of available policies. If models are initially trained from an inadequate set of measurements inputs, such that their fitness is insufficient to make reasonably accurate predictions of the measures for an iteration, Artemis will place the deviating regions into a training mode again to gather data on additional policies, so that future models for that region, within the run, will be more robustly informed. Programs with iterative algorithms should typically invoke optimization every time step of execution. When the user invokes the API, Artemis performs the following steps:

1. For every instrumented region it goes through the measurement records and finds the policy with the fastest measured execution for each feature set to enunciate the optimal pairs of each unique (feature set, policy) combination for this region;

2. In case of multi-process execution, Artemis communicates per-process best policy data between all executing processes to build a unified pool of these pairs and implement *collective training*,

3. From those feature set and policy pairs, it creates the training data to feed to the classification ML model, where the feature set is the feature input to the model and policy is the response;

4. Artemis feeds those data to train the ML model and derive an optimizing policy classifier for each region, that takes as input a feature set and produces as output the optimized selection policy.

When later executions of the instrumented regions query Artemis for the policy to execute, the trained model provides the optimizing policy index. Note that even after training an optimized policy selection model, Artemis continues to collect execution time data for optimized regions to monitor execution and trigger re-training, which we discuss next.

4.3.3 Validation and Retraining. As shown in Figure 20, Artemis includes a *regression* model to trigger re-training, anticipating that time-dependent or data-dependent behavior may change the execution profiles of regions, thus rendering previous optimizing models sub-optimal. Specifically, Artemis creates a regression model to predict execution time given the measurement records. The input features to train this regression model are the features set by instrumentation, including the policy selection, and the response outputs are the measured execution times.

At every invocation of the optimization API call by the user, Artemis compares the measured execution time per region, feature set, and policy to the predicted execution time provided by the regression model. When the measured time exceeds the predicted time over a threshold, Artemis discards the optimizing model and reverts the region to a training regime, trying out different execution policies on region execution to collect new data for training an optimized model. On a later invocation of the optimization API call, Artemis creates the new optimizing classification model and the new regression model for a new cycle of optimization and monitoring.

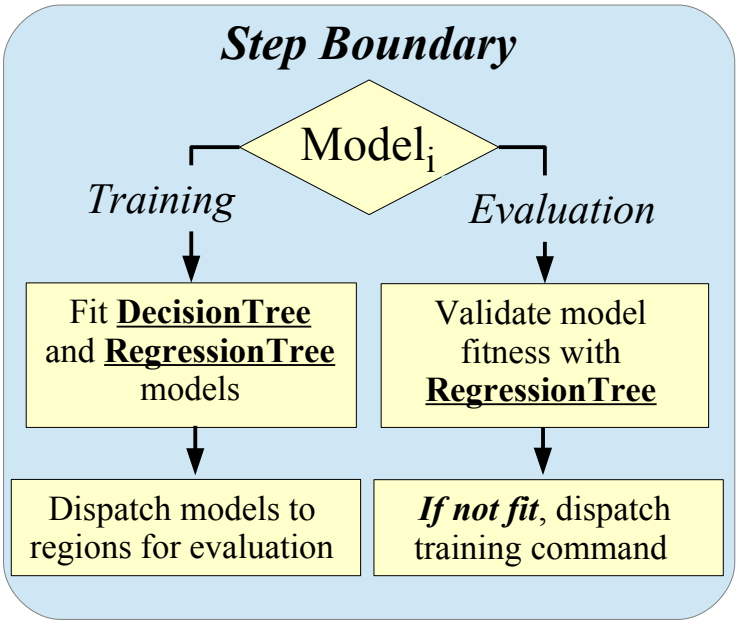


Figure 20. Artemis trains models and validates ongoing fitness.

4.3.4 Extending RAJA OpenMP execution. The RAJA [122] programming model was extended to enable Artemis optimization by defining an auto-tuned execution policy for parallel loop programming patterns implemented with OpenMP. Interestingly, much of region instrumentation is hidden by the end-user of RAJA since instrumentation happens inside the RAJA header library. The only refactoring required for a RAJA program is to make on-demand calls to the optimization API of Artemis and use the Artemis-recommended execution policy when defining parallel kernels through the RAJA templated API.

Specifically, we create an Artemis tuning policy for the `forall` programming pattern, which defines a parallel loop region, and for the `Collapse` kernel pattern, which collapses 2-level and 3-level nested to a single parallel loop, fusing the nested iteration spaces. For this implementation, we choose the `forall` and `Collapse` patterns since they are frequently used in applications. Artemis can integrate with other parallel patterns of RAJA, such as scans, OpenMP offloading, and

```
template <typename Iterable, typename Func>
RAJA_INLINE void forall_impl (artemis_exec &,
                              Iterable    &&iter,
                              Func        &&loop_body) {
static Artemis::Region *region = nullptr;
if (region == nullptr)
  region = Artemis::create_region(num_policies=2);
region->begin({ distance(begin(iter), end(iter)) });
int policy = region->getPolicyIndex();
switch(policy) {
case 0: {
  #pragma omp parallel
  { RAJA_EXTRACT_BED_IT(iter);
    #pragma omp for
    for (decltype(distance_it) i = 0; i < distance_it; ++i)
      loop_body(begin_it[i]);
  } } break;
case 1: {
    RAJA_EXTRACT_BED_IT(iter);
    for (decltype(distance_it) i = 0; i < distance_it; ++i)
      loop_body(begin_it[i]);
  } break; };
region->end();
}
```

Figure 21. Using Artemis in the RAJA forall execution pattern.

CUDA, which is work-in-progress. The Artemis policy used in our evaluation framework tunes execution by choosing between two policies: either OpenMP or sequential. The choice for those two policies is motivated by prior work [121] concluding that varying additional OpenMP parameters (number of threads, loop scheduling policy) results in sub-optimal tuning. Nevertheless, Artemis is general to tune for additional OpenMP parameters, which can be abstracted as different execution policies to input to the Artemis API. Artemis instrumentation is within the implementation of those patterns, in the RAJA header library.

Listing 21 shows a code excerpt for the instrumentation of the `forall` implementation with Artemis, redacting implementation details for RAJA closure privatization, for brevity of presentation. Note, the code for the `Collapse` kernel is similar. The `forall` implementation instruments the region execution with a call to `region->begin()` providing the number of iterations as the single feature in the feature set. For the `Collapse` implementation, the feature set consists of the iterations of all loop levels, creating a vector of features. Next, the implementation calls `region->getPolicyIndex()` which returns an index selecting the execution policy variant; 0 indicates executing with OpenMP and 1 indicates executing the region sequentially. This policy index is the input to the following `switch-case` statement that selects the execution variant. Lasty, there is a call to `region->end()` to marks the end of region execution.

This pattern of API use is general, and serves as a model for other interfaces and ports of Artemis, such as it's integration with the tuning API of the Kokkos portability framework.

4.3.5 Enhancing Kokkos CUDA execution. Besides RAJA OpenMP execution, we integrate Artemis to tune CUDA kernel execution within Kokkos [124]. Specifically, our experiment tuned parameters for the execution of an SpMV kernel computation in CUDA, including the *team size*, which is the outer level of parallelism of thread blocks, the *vector size*, which is the inner level of parallelism of numbers of threads and the *number of rows* of computation assigned to each thread.

4.3.6 Training Measurement. Initially, when Artemis first encounters an instrumented region, it deploys a *round-robin* strategy to collect training data. This strategy cycles through the set of provided policies, which

contains the OpenMP execution policy and the sequential policy in our RAJA implementation, or policies representing combinations of the various kernel launch parameters in the Kokkos integration. When searching, Artemis returns a policy index to explore a particular execution variant. In our implementation, round-robin advances the policy selection index for each region and each set of unique features independently. While searching the space of available policies, the Artemis runtime library records the unique feature set and the measured execution time for each instrumented region.

When Artemis is being used in an MPI application, it is capable of *collective training*, whereby training datasets across the processes are analyzed together.

At the end of an application step, every process issues a collective *allgather* operation to share their training datasets and gather the training datasets of every other process. Each process combines them to create a unified training dataset per region, informed by the rank-offset parallel round-robin searches, to find the best explored policy that minimizes execution time across both the local and peer training data.

4.3.7 Training Model Analysis and Optimization. Artemis processes the metrics gathered during training to construct the matrix of features to use in model construction. This includes the feature set, the performance responses, and the optimal policies. A Random Forest Classifier (RFC) model is trained per region, implemented using the OpenCV machine learning library. Artemis evaluates this RFC model in later invocations of `region->getPolicyIndex()` of a trained region, to return the optimized execution policy using as input the feature set provided in the arguments of the `region->begin( features )` call. We choose RFC modeling because it has fast

evaluation times of of $\mathcal{O}(m \log n)$ complexity for m decision trees of n depth in the forest. Fast evaluation is important for reducing the overhead during execution since `region->getPolicyIndex()` is called with every region's execution. For experimentation, we set the depth to 2 levels and the forest size to 10 trees, which has shown to be effective for optimization.

Artemis uses the same measurement data to train a per-region Random Forest Regression (RFR) model that predicts expected execution time. Artemis uses this regression model to detect time-dependent or data-dependent divergence in the execution of a region that invalidates a previously trained RFC optimizing model, indicating that re-training is needed. In the implementation, RFR models train with regression accuracy of $1e-6$, hence micro-second resolution for predicting time, and implement a forest size of 50 trees. RFR evaluation is off the critical path, hence affords the largest forest size, since it is called only on invocations of `Artemis::processMeasurements()`. For time regression analysis, Artemis compares the profiled execution time with the predicted one for all the region's feature sets. If the measured time for a feature set is greater than the predicted one given a threshold, then the model is considered *diverging*. This threshold limits re-trains due to transient perturbations when measuring execution time. We have experimentally found that this threshold value of $2\times$ filters out needless re-trains for the applications under test. Nonetheless, the threshold value is configurable and also re-training can be turned completely off, through environment variables. If the execution of an application region is pathological, such that execution time continuously diverges with the same features, then this region is ineligible for tuning and should be omitted or re-training should be turned off. This is a challenging scenario to naively automate, and future work involves exploring

strategies to effectively manage regions that do not have stable performance profiles even when features or loop inputs are held constant.

Artemis counts all diverging feature sets in a region. If they are found to be more than a threshold, more than half feature sets in a region for our implementation, Artemis deems the RFC model invalid and sets up the round-robin search strategy to re-train an optimized model for that region.

Artemis is generalized to support *heterogeneous execution*, where an application deploys to a cluster of heterogeneous machines, or for cases where a heterogeneous workload is specified on the same regions. Differences in machine architectures can be captured as a feature that describes the machine type, e.g., CPU or GPU micro-architecture. Differences in a heterogeneous workload, for the same code region, can be captured as a feature describing the condition causing it, e.g., the MPI rank or an application-designated parameter.

4.4 Experimentation Setup

The Artemis framework is intended to target environments where performance portability is important. When evaluating Artemis we want to compare its benefits to standard configurations of application and systems that they run on. On the one hand, Artemis is optimizing an application's execution on a machine from some point of reference. If that starts with an already optimized version, there is little likely to be gained. Thus, choosing a "default" version of the application with standard settings is more appropriate to gauge improvement. On the other hand, Artemis is optimizing an application across machines, where different architecture component (e.g., CPU, memory) could lead to different code variants being selected. The application code needs to be developed in such a way that making selection of those code variants is possible without completely

Table 2. Applications and Their Configurations

Application	Inputs	Nodes
LULESH	-r 100 -c 1 *or* 2 *or* 4 *or* 8 -i 100	1
Cleverleaf	Domain: (500,500), triple point calculation, 4 refinement levels, 25 timesteps, max patch size: 100×100 *or* 200×200, 400×400 *or* -1×-1(no limit)	1, 2, 4, 8
Kokkos Kernels SpMV	Domain: 100M to 600M non-zero values team size: 1-1024, vector size: 1-32 rows per thread: 1-4096	1

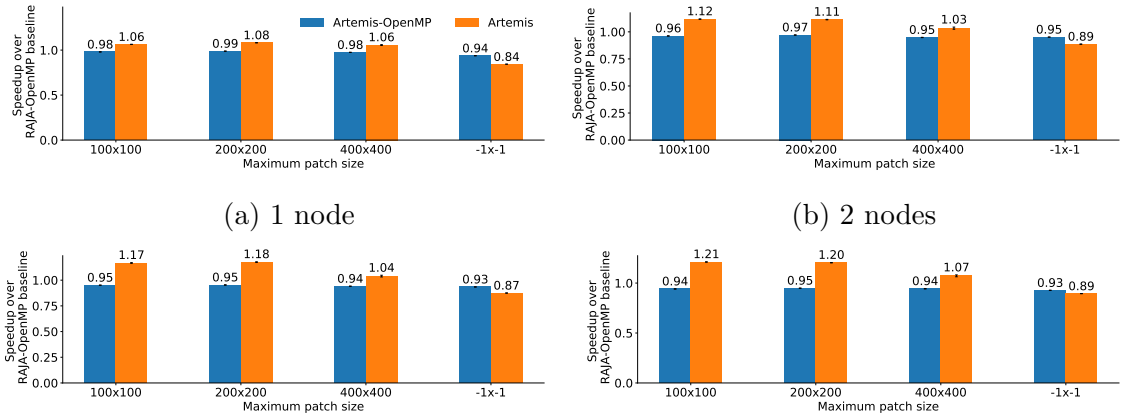
rewriting the application. This is the reason for working with RAJA and Kokkos for the experiments discussed below.

4.4.1 Comparators. The applications used in our study are developed with either RAJA or Kokkos, and we focus our attention on the parallel regions impacted by those portability frameworks. We define the *baseline* in performance comparison to be, for OpenMP, execution with the RAJA OpenMP execution policy using the same thread count for all regions, or in the CUDA case, the expert-tuned and hard-coded settings within the Kokkos Kernels suite. This is the *default mode* of executing these parallel applications. To quantify the instrumentation overhead of Artemis, we create a version of Artemis with this baseline that always selects the fixed default policy when guiding execution of a region, but does not perform any of the collection of performance measurements or online training. We call this the *Artemis-OpenMP* or *Artemis-Expert Heuristic* version. Lastly, we denote as *Artemis* the configuration where Artemis dynamically optimizes execution, using online profiling and machine learning for optimized policy selection and regression monitoring.

4.4.2 Applications. We chose three HPC proxy-applications to perform our experiments: LULESH [117, 126] and Cleverleaf [127, 128] for OpenMP, and Kokkos Kernels SpMV [129] for CUDA.

Table 2 shows details of the application inputs used and execution configurations. LULESH is configurable to create regions of different computational cost, to mimic multi-material calculation. Cleverleaf uses adaptive mesh refinement to create a range of problem subdomains, called patches, with varying computational cost. Thus, both data-dependent and input-dependent settings can create regions of different computation. Kokkos Kernels SpMV computes a sparse matrix vector product for very large matrices, allowing for a configurable count of non-zero values.

In the OpenMP codes, Artemis dynamically optimizes each parallel region by selecting OpenMP execution policies only when there is enough work to justify the overhead of parallel execution, otherwise it will elect for sequential execution. LULESH inputs create heterogeneous computation by using a large count of regions (100) that emulate different materials, changing the computational cost of various region subsets by 1, 2, 4, or 8 times the base cost – LULESH adjusts the cost of 45% of the regions to be this multiple and 5% of regions to be 10× this multiple. For Cleverleaf, heterogeneous computation is created by changing the maximum patch size permitted during refinement, ranging from from 100×100, 200×200, 400×400, up to an unlimited maximum by selecting -1×-1. The RAJA LULESH implementation does not support distributed execution with MPI, thus our experiments are single node. Cleverleaf provides support for MPI execution, so we performed experiments on multiple nodes to show Artemis's response to Cleverleaf's strong scaling properties. Kokkos Kernels SpMV experiments used



(a) 1 node

(b) 2 nodes

(c) 4 nodes

(d) 8 nodes

Figure 22. Cleverleaf, speedup of Artemis-OpenMP and Artemis over the baseline. Artemis to explore and select policies representing combinations of Kokkos settings and CUDA kernel launch parameters, across a variety of problem sizes.

4.4.3 Hardware and Software Platforms. Experiments were run on nodes featuring dual-socket Intel Xeon E5-2695v4 processors for 36 cores and 128GB of RAM per node and the TOSS3 software stack. We compiled applications and Artemis using GCC version 8.1.0 and MVAPICH2 version 2.3 for MPI support. Artemis used the OpenCV machine learning library version 4.3.0. For Kokkos CUDA we targeted the NVIDIA V100 (Volta) on an IBM Power9 architecture, using CUDA version 10.

4.4.4 Statistical Evaluation. For each OpenMP proxy application and configuration we performed 10 independent measurements. Unless otherwise noted, measurement counts the total application execution time end-to-end. Confidence intervals shown correspond to a 95% confidence level, calculated using Bootstrapping to avoid assumptions on the sampled population's distribution.

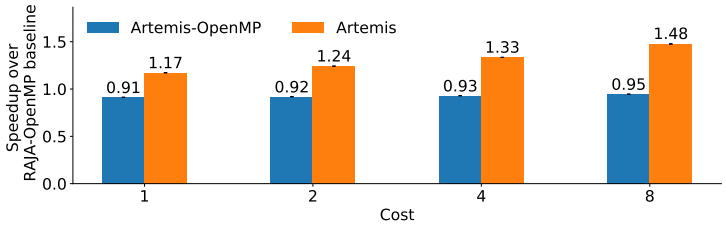


Figure 23. LULESH, speedup over the baseline of RAJA-OpenMP execution.

4.5 Evaluation

Here we provide results and detailed analysis of tuning for OpenMP with RAJA, as well as summary results from applying Artemis to tune Kokkos settings and CUDA kernel launch parameters.

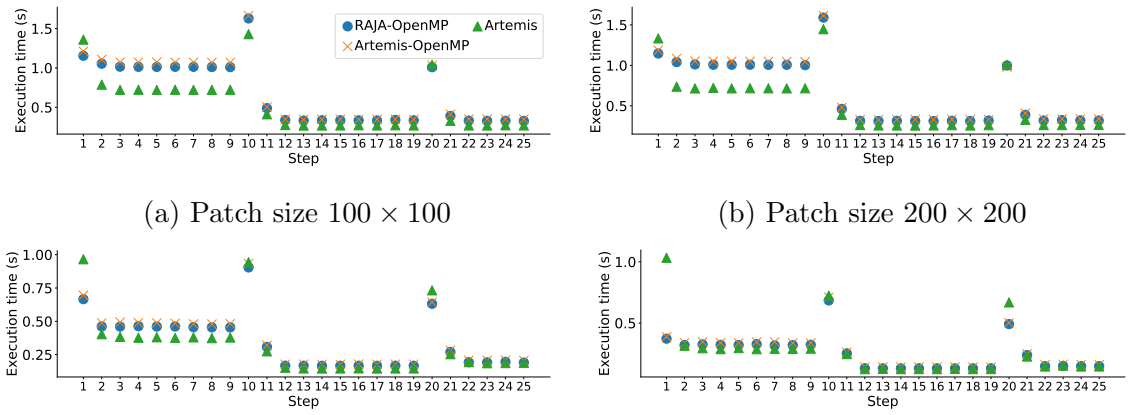
For evaluating the performance of Artemis with OpenMP, we compute the speedup over the baseline of RAJA-OpenMP execution for both Artemis-OpenMP, which always selects OpenMP execution, and the optimizing Artemis, which dynamically chooses between OpenMP or sequential execution for a region, using the machine learning methods we described. Artemis-OpenMP exposes the instrumentation overhead of Artemis, hence the expected slowdown compared to non-instrumented RAJA-OpenMP execution. Figure 22 shows results for Cleverleaf, and Figure 23 shows results for LULESH. Values on bars show the mean speedup (or slowdown) compared to RAJA-OpenMP execution.

4.5.1 Instrumentation Overhead.

Observing the slowdown of Artemis-OpenMP, the overhead of instrumentation is modest, cumulatively less than 9% across both applications and tested configurations of input and node numbers. This shows that Artemis does not overburden execution and given tuning opportunities, it should recuperate the overhead and provide speedup over non-instrumented RAJA-OpenMP execution.

4.5.2 Model Training and Evaluation Overhead. The average training time for LULESH is 310 microseconds, while for Cleverleaf is 150 microseconds, which is minimal contrasted with the timescale of execution of regions, as we show in later measurements, so Artemis recovers this overhead, effectively tuning and speeding up execution. Moreover, model training (or re-training) is infrequently done as trained models persist during execution. By contrast, model evaluation happens at every execution of a tunable region. Its overhead depends on the forest size and tree depth of the trees in the evaluated forest. Given the limits in forest size (10) and tree depth (2) set in our implementation, see section 4.3, we measure the time overhead for evaluating the maximum possible forest configuration to be less than 10 microseconds.

4.5.3 Speedup on Cleverleaf. For Cleverleaf, varying the maximum patch size changes the number and size of computational regions. A smaller size means more regions, hence more parallelism, but also finer-grain decomposition of the computation domain. So, there is greater disparity between regions that lack enough work, hence sequential policy is fastest, and regions with enough parallel work, for which OpenMP execution is fastest. Note, the special value -1×-1 means there is no maximum set and Cleverleaf by default prioritizes decomposing in larger regions. Figure 22 shows results for all node configurations, demonstrating that Artemis consistently speeds up execution for the smaller patch sizes of 100×100 and 200×200, no less than 8%, executing with one node, and up to 21%, executing on 8 nodes. For the larger patch size of 400×400, execution with Artemis is on par with RAJA-OpenMP, successfully recuperating the overhead with marginal gains, within measurement error. For the unlimited patch size of -1×-1, Artemis results in a net slowdown, also compared with Artemis-OpenMP, since



(a) Patch size 100×100

(b) Patch size 200×200

(c) Patch size 400×400

(d) Patch size -1×-1

Figure 24. Execution time per timestep for Cleverleaf on 8 nodes, varying the maximum patch size. Regridding operation performed after every 10 steps.

there is lack of optimization opportunity, and the training and monitoring overhead inflated execution time.

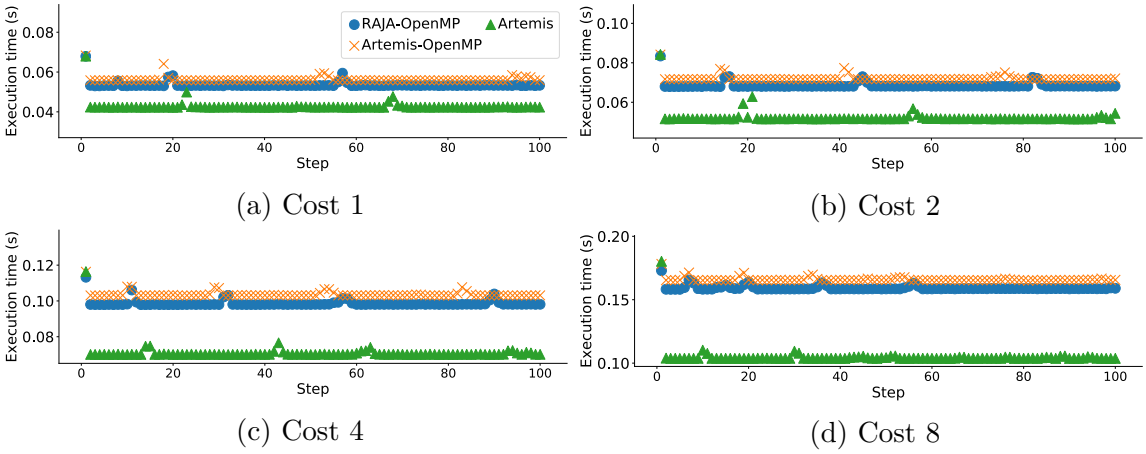
For further analysis, we show results comparing execution times per timestep for different execution modes. Figure 24 shows results when executing with 8 nodes. Results for other node counts are similar, thus we omit them for brevity. Note that Cleverleaf performs a *re-gridding* operation [130] every 10 timesteps that re-shuffles domain decomposition to reduce computation error, thus the spikes in execution time in the 10th and 20th timesteps.

Observing results, Artemis inflates execution time for the first timestep across all patch sizes, since this step includes training for bootstrapping tunable regions. For most of the rest of timesteps, Artemis reduces execution time, by as much as 40% for the least patch size of 100×100, compared to the default execution with RAJA-OpenMP. Artemis tuning potential lessens the larger the patch size, since larger regions favor OpenMP execution. Nevertheless, observing Figure 24d for the largest patch size selection, Artemis correctly selects OpenMP execution

and any performance lost is due to the initial training overhead. Notably, Cleverleaf execution with 8 nodes has second to sub-second timesteps, and Artemis is fast enough to optimize execution even at this short time scale. Expectedly, Artemis-OpenMP has slightly higher execution time per timestep compared to RAJA-OpenMP, reflecting instrumentation overhead as seen by the speedup results.

4.5.4 Effectiveness of Cleverleaf Policy Selection. Cleverleaf instantiates a multitude of regions and each region executes with multiple different feature sets, corresponding to different patch sizes from decomposing the domain and load balancing. So, to highlight Artemis effectiveness we fix the patch size to 100×100, which presents the most optimization potential, and pick one region to plot the average execution time of each feature set for the top-20 most frequently executed ones, contrasting OpenMP only execution vs. sequential execution vs. Artemis execution with dynamic policy selection. The region comprises of feature sets corresponding to 2d collapsed loops, so there are two values describing (outer,inner) loop iterations. Depending on the feature set size, OpenMP or sequential is the best. For example, feature set (3,201) executes faster with OpenMP and feature set (55, 2) executes faster sequentially. Observing execution times measured for Artemis, policy recommendations converge to the optimal policy for the majority of feature sets for which the performance difference between the sequential and OpenMP policy selection is more than 20%. Artemis selects the optimal policy in 10 of the 15 such regions.

Further, we find positive results for the accuracy of Artemis in selecting optimal policies. For the initial timestep, Artemis has low accuracy, ranging from 10% to 20%, due to training, without any discernible trend among different patch sizes. However, accuracy significantly improves after this initial, training step to a



(a) Cost 1 (b) Cost 2

(c) Cost 4 (d) Cost 8

Figure 25. Execution time per timestep for LULESH, showing different execution modes on one node, varying the cost of computational regions.

range of 85% to 95%, showing Artemis is effective in selecting the optimal policy most of the time.

4.5.5 Strong scaling with different node counts. Figures 22a–22d show results for increasing node counts. Following the discussion on smaller patch sizes that present optimization opportunities for Artemis, increasing the number of nodes also boosts the speedup achieved by Artemis. Cleverleaf distributes computational regions among different MPI ranks and executes bulk-synchronous, advancing the simulation time step after all MPI ranks have finished processing. Artemis dynamically optimizes execution per rank, thus it reduces execution time on the critical path, with multiplicative effect on the overall execution.

4.5.6 Speedup on LULESH. Figure 23 shows results for LULESH on a single node due to the limitation of the RAJA version of LULESH supporting only single node execution. For this experiment, the number of regions is kept constant (100) and the cost of computation varies between 1× (default) and 8×, as explained in section 4.4. Similarly to Cleverleaf, the instrumentation overhead

of Artemis, shown by observing the slowdown of Artemis-OpenMP, is within 9% of non-instrumented execution of RAJA-OpenMP.

Regarding speedup of Artemis, it is consistently faster than RAJA-OpenMP. Artemis improves execution time even for the default setting of cost 1× by 16%. Expectedly, increasing the cost creates more computational disparity between LULESH computational regions, thus Artemis achieves higher speedup. For the highest cost value we experiment with, a cost of 8×, Artemis achieves significant speedup of 47% over the RAJA-OpenMP baseline.

For more detailed results, Figure 25 shows execution time per timestep for all execution modes varying the cost of computational regions. Observations are similar to Cleverleaf, the first timestep under Artemis is slower due to training while the rest of the timesteps execute faster than RAJA-OpenMP. Artemis speeds up the execution of timestep up to 50% compared to RAJA-OpenMP, increasingly so as the cost input increases. Different than Cleverleaf, the resolution of the execution time of LULESH is much more fine-grain, in the range of hundreds of milliseconds. Nonetheless, Artemis effectively optimizes execution even at this time scale, showing that training effectively optimizes policy selection and overcomes any instrumentation overhead.

4.5.7 Speedup on Kokkos Kernels SpMV. Figure 26 shows the results of our integration with Kokkos, tuning the parallel team size, vector size, and number of rows assigned to each thread. The x-axis shows scaling the number of non-zero elements y-axis plots the average execution time for 1500 SpMV kernel invocations. *Expert Heuristic* is the existing, hardcoded tuning strategy set by the expert kernel developer, setting those parameters based on the input data and expert knowledge. This heuristic function settles on 1 row per thread, a vector

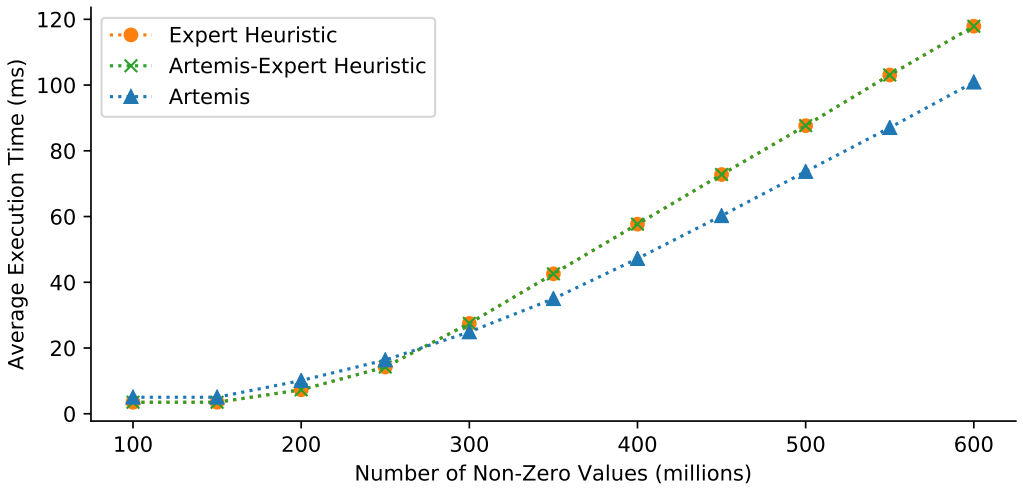


Figure 26. Artemis improves performance of the Kokkos SpMV kernel up to 16.8% compared to the hardcoded expert heuristic.

length of 2, and a team size of 256 for inputs shown. *Artemis-Expert Heuristic* exposes the instrumentation overhead of Artemis, by foregoing tuning, instead executing with the same settings of the expert heuristic. The performance of Artemis-Expert Heuristic is on par with execution of Expert Heuristic without Artemis intervening, thus instrumentation overhead is minimal. *Artemis* shows the performance improvement when tuning is enabled. Kokkos provides a range of 664 selectable policies to Artemis for tuning, with parameters team size ranging from 1–1024, vector size from 1–32, and number of rows per thread from 1–4096. Results show that Artemis succesfully navigates the tuning space, and provides increasingly faster performance as the problem size increases, for a maximum of 16.8% performance improvement on the largest input of 600M non-zero elements.

4.6 Related Work

Existing tuning frameworks are either application-specific [131] [132], programming-model-specific [133] [134], hardware-specific [135] [136], or feature the need for offline training [137] [121], and thus have limited scope. By design,

Artemis is a general framework that gives an API to tune at any of those levels, and we show its generality by integrating Artemis with the RAJA programming model, tuning a variety of HPC proxy applications and kernels. The closest to our work is the Apollo paper by Beckingsale et al.[121], with the important distinction that, rather than exhaustive offline tuning, the Artemis framework performs the search space exploration at runtime.

Empirical techniques directly measure all the possible variants and select the fastest. Established projects like the ATLAS [135] [138] and FFTW [136] libraries apply this technique with great success, but it requires the up front cost of finding the best code variant choices for each system. ATF [139] [140] presents a generic extensible framework for automated tuning, independent of programming language or domain. Oski [131] performs runtime tuning, optimizing over sparse linear algebra kernels. Orio [141] and OpenTuner [133] are able to facilitate general purpose kernel tuning using empirical techniques to select the best performing configurations for production. ActiveHarmony [142] uses parallel search strategies to perform online tuning, though sweeping large parameter spaces can take significant amounts of time.

Using some form of a model to predict the performance of the code, analytical examples make tuning decisions based on model output. Similarly to Artemis, AutoTuneTMP [134] makes use of C++ template metaprogramming to abstract-away the tuning mechanisms of kernels and facilitate performance portability. It constrains the search space for online training using parameterized kernel definitions. Unlike Artemis's use of RAJA policies that are compiled in alongside the application, AutoTuneTMP uses JIT compilation and dynamic linking at runtime to produce kernel variants, a mechanism which could impose

non-trivial overhead in a large large class of HPC codes in production settings. Mira [143] uses static performance analysis to generate and explore performance models offline. Mira's abstract performance models allow it to avoid some of the limitations to offline learning.

A statistical model is built by applying machine learning techniques, and this model is used to make tuning decisions. Sreenivasan et al. [137] demonstrated performance gains using an OpenMP autotuner framework that performs offline tuning using a random forest statistical model of the reduced search space to eliminate exhaustive tuning. HiPerBOt [144] presents an active learning framework that uses Bayesian techniques to maintain optimal outcomes while collapsing the required number of samples for learning.

Other work [145, 146, 147] has looked into auto-tuning the number of OpenMP threads in multi-program execution. Those approaches look at architectural metrics, such as Instructions-Per-Cycle and memory stalls, to dynamically throttle thread allocation when contention occurs.

4.7 Conclusion and Future Work

We have presented Artemis, a novel framework that optimizes performance by tuning an application's parallel computational regions online. Artemis provides a powerful API to integrate online tuning in existing applications, by defining tunable regions and execution variants. Artemis automatically adapts to data-dependent or time-dependent changes in execution using decision tree and regression models. We integrated Artemis with RAJA and Kokkos and evaluated online tuning performance on HPC proxy applications: Cleverleaf and LULESH, and a CUDA SpMV kernel. Results show that Artemis is up to 47% faster and its operating overhead is minimal.

Future work on the Artemis project includes:

1. Using Artemis for tuning of additional GPU-offloaded compute kernels with heterogeneous memory hierarchies.

2. Tuning additional parallel execution parameters such as loop tiling and nesting.

3. Expanding experimentation to large applications by extending the Artemis codebase and integration with RAJA, Kokkos, and lower level parallel programming models, such as OpenMP, CUDA, and HIP.

CHAPTER V

CONCLUSION

The field of HPC is wildly diverse and always in motion. There will never be a single one-size-fits-all solution to the challenges of optimally executing massively parallel software at extreme scales. But as we have shown in this overall work, hope is not lost! There are excellent opportunities to develop and field tools and service layers that will dramatically enhance both application performance, and efforts to generalize and future-proof the fitness of existing large and costly HPC software projects.

Let us revisit our four research questions now, and reflect on the answers we have found for them.

- **RQ1:** What are the essential components of a practical in situ system for online observation, analysis, and feedback?
- **RQ2:** Can online observation with in situ methods provide benefits to application users and developers?
- **RQ3:** Is it feasible to conduct machine learning in situ in order to derive performance benefits without a human in the loop?
- **RQ4:** Can systems be made both observable and responsive to tuning choices without costly code interventions or algorithm rewrites?

In Chapter I we unpacked the nature and means of accessing the information necessary to make accurate, germane, timely decisions about application performance. Along the way, various historical and current systems were described, along with discussion of their context, leading into a deeper dive

into a general model of observation which we proposed and demonstrated in Chapter II. The SOS model described there is a kind of reasonable minimum set of properties for future systems to consider. Taken together, both of these chapters serve to provide a robust answer to **RQ1** and indicate the opportunity for **RQ2** to be answered in the affirmative.

If engineering a monitoring solution for some novel HPC environment that has yet to be fielded, each of the features proposed for SOS should be considered and accounted for as early as possible, in order to maximize the flexibility and performance of that observational layer. As we showed in our results, with some thoughtful engineering, such an observational layer can exist with overhead that essentially dissolves into the noise endogenous to large-scale system performance measures. With mature engineering effort beyond the scope or length of a graduate research inquiry, these first-light results could move from being serviceable to being truly impressive. Once arbitrary information from system measures and application state can be obtained and utilized at runtime, the imagination is the limit when it comes to types of expert analysis or automated tuning one wishes to facilitate. Our next two investigations demonstrate this with compelling results.

Chapter III shows how the high-level SOS model meets real world environments and codes in practice, and how it is able to facilitate a dynamic online aggregation of general application data in addition to system performance measures, without disrupting the performance of the application or environment being observed. Further, it is able to coordinate the projection of performance measures over the application domain, including driving the online rendering of 3D images, allowing an observer to have an intuitive representation of the behavior

of their algorithms when their code is actually running on the machine. This allows us to answer **RQ2** with a decisive, "*yes.*"

Taking that feature gain in a different direction, Chapter IV dives into the solutions for both **RQ3** and **RQ4**. Setting to one side the already-demonstrated generalized observation layer, this investigation focused instead on whether or not classic machine learning techniques could be brought to bear using in situ (online) data availability, whether it could be made performant enough to not consume any gains it discovered, once the training and model processing overhead was factored in. As it turns out, the system was able to greatly exceed its overhead and both discover opportunities for, and sustainably deliver, performance gains from input-dependent and configuration-dependent deployments of previously well-optimized codes. The mechanism by which this capability was fielded provides another affirmative outcome: The parallel portability frameworks being widely adopted within the HPC community can be an extremely low-impact and flexible point of engagement for the kind of scalable observation, analysis, and tuning models we advocate for generally in this work.

This is a fast-moving and exciting area of research, with many different execution scales, data access patterns, and unique target architectures. Our intuitions, approach, and demonstrations consistently yielded positive results and should encourage further effort to be applied to the development of more robust integrated general infrastructures for in situ (online) monitoring, and the adoption of programming models like RAJA and Kokkos that support dynamic adaptivity for HPC codes. Code that can automatically and portably behave optimally provides benefits both immediately, and in the future.

REFERENCES CITED

[1] D. G. Solt, J. Hursey, A. Lauria, D. Guo, and X. Guo, "Scalable, fault-tolerant job step management for high performance systems," *IBM Journal of Research and Development*, 2019.

[2] A. B. Yoo, M. A. Jette, and M. Grondona, "Slurm: Simple linux utility for resource management," in *Workshop on Job Scheduling Strategies for Parallel Processing*. Springer, 2003, pp. 44–60.

[3] P. Braam, "The lustre storage architecture," *arXiv preprint arXiv:1903.01955*, 2019.

[4] F. B. Schmuck and R. L. Haskin, "Gpfs: A shared-disk file system for large computing clusters." in *FAST*, vol. 2, no. 19, 2002.

[5] W. Gropp, E. Lusk, N. Doss, and A. Skjellum, "A high-performance, portable implementation of the mpi message passing interface standard," *Parallel computing*, vol. 22, no. 6, pp. 789–828, 1996.

[6] D. Boehme, K. Huck, J. Madsen, and J. Weidendorfer, "The case for a common instrumentation interface for hpc codes," in *2019 IEEE/ACM International Workshop on Programming and Performance Visualization Tools (ProTools)*. IEEE, 2019, pp. 33–39.

[7] A. M. Wissink, R. D. Hornung, S. R. Kohn, S. S. Smith, and N. Elliott, "Large scale parallel structured amr calculations using the samrai framework," in *Proceedings of the 2001 ACM/IEEE conference on Supercomputing*, 2001, pp. 6–6.

[8] L. Dagum and R. Menon, "Openmp: an industry standard api for shared-memory programming," *IEEE computational science and engineering*, vol. 5, no. 1, pp. 46–55, 1998.

[9] B. Mohr, A. D. Malony, S. Shende, F. Wolf *et al.*, "Towards a performance tool interface for openmp: An approach based on directive rewriting," in *Proceedings of the Third Workshop on OpenMP (EWOMP'01)*, 2001.

[10] A. E. Eichenberger, J. Mellor-Crummey, M. Schulz, M. Wong, N. Copty, R. Dietrich, X. Liu, E. Loh, and D. Lorenz, "Ompt: An openmp tools application programming interface for performance analysis," in *International Workshop on OpenMP*. Springer, 2013, pp. 171–185.

[11] A. Bhatele, K. Mohror, S. H. Langer, and K. E. Isaacs, "There goes the neighborhood: performance degradation due to nearby jobs," in *SC'13: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. IEEE, 2013, pp. 1–12.

[12] K. Yelick, D. Bonachea, W.-Y. Chen, P. Colella, K. Datta, J. Duell, S. L. Graham, P. Hargrove, P. Hilfinger, P. Husbands *et al.*, "Productivity and performance using partitioned global address space languages," in *Proceedings of the 2007 international workshop on Parallel symbolic computation*, 2007, pp. 24–32.

[13] M. De Wael, S. Marr, B. De Fraine, T. Van Cutsem, and W. De Meuter, "Partitioned global address space languages," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, pp. 1–27, 2015.

[14] H. Kaiser, T. Heller, B. Adelstein-Lelbach, A. Serio, and D. Fey, "Hpx: A task based programming model in a global address space," in *Proceedings of the 8th International Conference on Partitioned Global Address Space Programming Models*, 2014, pp. 1–11.

[15] L. V. Kale and S. Krishnan, "Charm++ a portable concurrent object oriented system based on c++," in *Proceedings of the eighth annual conference on Object-oriented programming systems, languages, and applications*, 1993, pp. 91–108.

[16] J. M. Wozniak, T. G. Armstrong, M. Wilde, D. S. Katz, E. Lusk, and I. T. Foster, "Swift/t: Large-scale application composition via distributed-memory dataflow processing," in *2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing*. IEEE, 2013, pp. 95–102.

[17] K. A. H. PI, A. D. Malony, and M. M. A. Haque, "Apex/hpx integration specification for phylanx," 2019.

[18] M. A. H. Monil, B. Wagle, K. Huck, and H. Kaiser, "Adaptive auto-tuning in hpx using apex."

[19] B. Taubmann and H. P. Reiser, "Towards hypervisor support for enhancing the performance of virtual machine introspection," in *IFIP International Conference on Distributed Applications and Interoperable Systems*. Springer, 2020, pp. 41–54.

[20] P. J. Mucci, S. Browne, C. Deane, and G. Ho, "Papi: A portable interface to hardware performance counters," in *Proceedings of the department of defense HPCMP users group conference*, vol. 710, 1999.

[21] S. Rostedt, "Finding origins of latencies using ftrace," *Proc. RT Linux WS*, 2009.

[22] M. Gebai and M. R. Dagenais, "Survey and analysis of kernel and userspace tracers on linux: Design, implementation, and overhead," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–33, 2018.

[23] T. Bird, "Measuring function duration with ftrace," in *Proceedings of the Linux Symposium*. Citeseer, 2009, pp. 47–54.

[24] D. Fukui, M. Shimaoka, H. Mikami, D. Hillenbrand, H. Yamamoto, K. Kimura, and H. Kasahara, "Annotatable systrace: an extended linux ftrace for tracing a parallelized program," in *Proceedings of the 2nd International Workshop on Software Engineering for Parallel Systems*, 2015, pp. 21–25.

[25] A. Nagai, "Introduce new branch tracer 'perf branch'," *Linux Technology Center, Yokohama Research Lab, Hitachi Ltd., Copyright*, 2011.

[26] P.-M. Fournier, M. Desnoyers, and M. R. Dagenais, "Combined tracing of the kernel and applications with lttng," in *Proceedings of the 2009 linux symposium*. Citeseer, 2009, pp. 87–93.

[27] D. Couturier and M. R. Dagenais, "Lttng clust: a system-wide unified cpu and gpu tracing tool for opencl applications," *Advances in Software Engineering*, vol. 2015, 2015.

[28] N. Nethercote and J. Seward, "Valgrind: a framework for heavyweight dynamic binary instrumentation," *ACM Sigplan notices*, vol. 42, no. 6, pp. 89–100, 2007.

[29] S. Lammel, F. Zahn, and H. Fröning, "Sonar: Automated communication characterization for hpc applications," in *International Conference on High Performance Computing*. Springer, 2016, pp. 98–114.

[30] A. A. Gimenez and U. N. N. S. Administration, "Sonar," 11 2018. [Online]. Available: https://www.osti.gov//servlets/purl/1493001

[31] A. Knüpfer, R. Brendel, H. Brunst, H. Mix, and W. E. Nagel, "Introducing the open trace format (otf)," in *International Conference on Computational Science*. Springer, 2006, pp. 526–533.

[32] A. D. Malony and W. E. Nagel, "The open trace format (otf) and open tracing for hpc," in *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, 2006, pp. 24–es.

[33] D. Eschweiler, M. Wagner, M. Geimer, A. Knüpfer, W. E. Nagel, and F. Wolf, "Open trace format 2: The next generation of scalable trace formats and support libraries." in *PARCO*, vol. 22, 2011, pp. 481–490.

[34] S. Shende, A. Malony, G. Allen, J. Carver, S. Choi, T. Crick, and M. Crusoe, "Using tau for performance evaluation of scientific software," in *Workshop on Sustainable Software for Science: Practice and Experiences*, no. 1686, 2016.

[35] L. Adhianto, S. Banerjee, M. Fagan, M. Krentel, G. Marin, J. Mellor-Crummey, and N. R. Tallent, "Hpctoolkit: Tools for performance analysis of optimized parallel programs," *Concurrency and Computation: Practice and Experience*, vol. 22, no. 6, pp. 685–701, 2010.

[36] A. Wang, X. Mei, J. Croft, M. Caesar, and B. Godfrey, "Ravel: A database-defined network," in *Proceedings of the Symposium on SDN Research*, 2016, pp. 1–7.

[37] L. Riliskis, J. Hong, and P. Levis, "Ravel: Programming iot applications as distributed models, views, and controllers," in *Proceedings of the 2015 International Workshop on Internet of Things towards Applications*, 2015, pp. 1–6.

[38] L. Sun, G. Tian, G. Zhu, Y. Liu, H. Shi, and D. Dai, "Multipath ip routing on end devices: Motivation, design, and performance," in *2018 IFIP networking conference (IFIP networking) and workshops*. IEEE, 2018, pp. 1–9.

[39] A. Giménez, T. Gamblin, A. Bhatele, C. Wood, K. Shoga, A. Marathe, P.-T. Bremer, B. Hamann, and M. Schulz, "Scrubjay: deriving knowledge from the disarray of hpc performance data," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2017, pp. 1–12.

[40] K. Mehta, B. Allen, M. Wolf, J. Logan, E. Suchyta, J. Choi, K. Takahashi, I. Yakushin, T. Munson, I. Foster *et al.*, "A codesign framework for online data analysis and reduction," in *2019 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS)*. IEEE, 2019, pp. 11–20.

[41] D. Boehme, T. Gamblin, D. Beckingsale, P.-T. Bremer, A. Gimenez, M. LeGendre, O. Pearce, and M. Schulz, "Caliper: performance introspection for hpc software stacks," in *High Performance Computing, Networking, Storage and Analysis, SC16: International Conference for*. IEEE, 2016, pp. 550–560.

[42] O. Aaziz, J. Cook, and H. Sharifi, "Push me pull you: Integrating opposing data transport modes for efficient hpc application monitoring," in *2015 IEEE International Conference on Cluster Computing*. IEEE, 2015, pp. 674–681.

[43] J. E. Thornton, "The cdc 6600 project," *Annals of the History of Computing*, vol. 2, no. 4, pp. 338–348, 1980.

[44] C. J. Murray, *The supermen: the story of Seymour Cray and the technical wizards behind the supercomputer*. John Wiley & Sons, Inc., 1997.

[45] R. Izadpanah, B. A. Allan, D. Dechev, and J. Brandt, "Production application performance data streaming for system monitoring," *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, vol. 4, no. 2, pp. 1–25, 2019.

[46] M. J. Sottile and R. G. Minnich, "Supermon: A high-speed cluster monitoring system," in *Proceedings. IEEE International Conference on Cluster Computing*. IEEE, 2002, pp. 39–46.

[47] T. Oetiker and D. Rand, "Mrtg: The multi router traffic grapher." in *LISA*, vol. 98, 1998, pp. 141–148.

[48] T. Oetiker, "Monitoring your it gear: the mrtg story," *IT professional*, vol. 3, no. 6, pp. 44–48, 2001.

[49] S. Zhang, I.-L. Yen, and F. B. Bastani, "Toward semantic enhancement of monitoring data repository," in *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*. IEEE, 2016, pp. 140–147.

[50] F. D. Sacerdoti, M. J. Katz, M. L. Massie, and D. E. Culler, "Wide area cluster monitoring with ganglia," in *null*. IEEE, 2003, p. 289.

[51] M. L. Massie, B. N. Chun, and D. E. Culler, "The ganglia distributed monitoring system: design, implementation, and experience," *Parallel Computing*, vol. 30, no. 7, pp. 817–840, 2004.

[52] S. Mongkolluksamee, P. Pongpaibool, and C. Issariyapat, "Strengths and limitations of nagios as a network monitoring solution," in *Proceedings of the 7th International Joint Conference on Computer Science and Software Engineering (JCSSE 2010). Bangkok, Thailand*, 2010, pp. 96–101.

[53] G. Katsaros, R. Kübert, and G. Gallizo, "Building a service-oriented monitoring framework with rest and nagios," in *2011 IEEE International Conference on Services Computing*. IEEE, 2011, pp. 426–431.

[54] T. Evans, W. L. Barth, J. C. Browne, R. L. DeLeon, T. R. Furlani, S. M. Gallo, M. D. Jones, and A. K. Patra, "Comprehensive resource use monitoring for hpc systems with tacc stats," in *2014 First International Workshop on HPC User Support Tools*. IEEE, 2014, pp. 13–21.

[55] H. Sharifi, O. Aaziz, and J. Cook, "Monitoring hpc applications in the production environment," in *Proceedings of the 2nd Workshop on Parallel Programming for Analytics Applications*, 2015, pp. 39–47.

[56] W. R. Williams, X. Meng, B. Welton, and B. P. Miller, "Dyninst and mrnet: Foundational infrastructure for parallel tools," in *Tools for High Performance Computing 2015*. Springer, 2016, pp. 1–16.

[57] C. Wood, S. Sane, D. Ellsworth, A. Gimenez, K. Huck, T. Gamblin, and A. Malony, "A scalable observation system for introspection and in situ analytics," in *Proceedings of the 5th Workshop on Extreme-Scale Programming Tools*. IEEE Press, 2016, pp. 42–49.

[58] S. Forti, M. Gaglianese, and A. Brogi, "Lightweight self-organising distributed monitoring of fog infrastructures," *Future Generation Computer Systems*, vol. 114, pp. 605–618.

[59] A. Agelastos, B. Allan, J. Brandt, P. Cassella, J. Enos, J. Fullop, A. Gentile, S. Monk, N. Naksinehaboon, J. Ogden *et al.*, "The lightweight distributed metric service: a scalable infrastructure for continuous monitoring of large scale computing systems and applications," in *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2014, pp. 154–165.

[60] A. Agelastos, B. Allan, J. Brandt, A. Gentile, S. Lefantzi, S. Monk, J. Ogden, M. Rajan, and J. Stevenson, "Toward rapid understanding of production hpc applications and systems," in *2015 IEEE International Conference on Cluster Computing*. IEEE, 2015, pp. 464–473.

[61] S. Feldman, D. Zhang, D. Dechev, and J. Brandt, "Extending ldms to enable performance monitoring in multi-core applications," in *2015 IEEE International Conference on Cluster Computing*. IEEE, 2015, pp. 717–720.

[62] R. Izadpanah, N. Naksinehaboon, J. Brandt, A. Gentile, and D. Dechev, "Integrating low-latency analysis into hpc system monitoring," in *Proceedings of the 47th International Conference on Parallel Processing*, 2018, pp. 1–10.

[63] D. Montaldo, E. Mocskos, and D. F. Slezak, "Clover: Efficient monitoring of hpc clusters," 2009.

[64] (2020) Performance co-pilot: System-wide monitoring. [Online]. Available: https://pcp.io

[65] C. Guillen, W. Hesse, and M. Brehm, "The persyst monitoring tool," in *European Conference on Parallel Processing*. Springer, 2014, pp. 363–374.

[66] J. Treibig, G. Hager, and G. Wellein, "Likwid: A lightweight performance-oriented tool suite for x86 multicore environments," in *2010 39th International Conference on Parallel Processing Workshops*. IEEE, 2010, pp. 207–216.

[67] T. Röhl, J. Eitzinger, G. Hager, and G. Wellein, "Likwid monitoring stack: A flexible framework enabling job specific performance monitoring for the masses," in *2017 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2017, pp. 781–784.

[68] L. Stanisic and K. Reuter, "Mpcdf hpc performance monitoring system: Enabling insight via job-specific analysis," in *European Conference on Parallel Processing*. Springer, 2019, pp. 613–625.

[69] B. Shihada, "Conceptual & concrete architectures of open network management system (opennms)," 2002.

[70] N. Sukhija and E. Bautista, "Towards a framework for monitoring and analyzing high performance computing environments using kubernetes and prometheus," in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 2019, pp. 257–262.

[71] N. Sukhija, E. Bautista, O. James, D. Gens, S. Deng, Y. Lam, T. Quan, and B. Lalli, "Event management and monitoring framework for hpc environments using servicenow and prometheus," in *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, 2020, pp. 149–156.

[72] V. Medel, O. Rana, J. Á. Bañares, and U. Arronategui, "Modelling performance & resource management in kubernetes," in *Proceedings of the 9th International Conference on Utility and Cloud Computing*, 2016, pp. 257–262.

[73] S. Patarin and M. Makpangou, "Pandora: A flexible network monitoring platform," 1999.

[74] N. Chan, "A resource utilization analytics platform using grafana and telegraf for the savio supercluster," in *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning)*, 2019, pp. 1–6.

[75] P. Rattanatamrong, Y. Boonpalit, S. Suwanjinda, A. Mangmeesap, K. Subraties, V. Daneshmand, S. Smallen, and J. Haga, "Overhead study of telegraf as a real-time monitoring agent," in *2020 17th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, 2020, pp. 42–46.

[76] S. N. Z. Naqvi, S. Yfantidou, and E. Zimányi, "Time series databases and influxdb," *Studienarbeit, Université Libre de Bruxelles*, 2017.

[77] Zabbix distributed monitoring solution. [Online]. Available: https://www.zabbix.com

[78] E. Simmonds and J. Harrington, "Scf/fef evaluation of nagios and zabbix monitoring systems," *SCF/FEF*, pp. 1–9, 2009.

[79] T. Wang, J. Xu, W. Zhang, Z. Gu, and H. Zhong, "Self-adaptive cloud monitoring with online anomaly detection," *Future Generation Computer Systems*, vol. 80, pp. 89–101, 2018.

[80] "collectd: The system statistics collection daemon," 2020. [Online]. Available: https://collectd.org

[81] S. Benedict, V. Petkov, and M. Gerndt, "Periscope: An online-based distributed performance analysis tool," in *Tools for High Performance Computing 2009*. Springer, 2010, pp. 1–16.

[82] J. R. Mayo, F. X. Chen, P. P. Pebay, M. H. Wong, D. Thompson, A. C. Gentile, D. C. Roe, V. De Sapio, and J. M. Brandt, "Understanding large scale hpc systems through scalable monitoring and analysis." Sandia National Laboratories, Tech. Rep., 2010.

[83] J. P. White, M. Innus, R. L. Deleon, M. D. Jones, and T. R. Furlani, "Monitoring and analysis of power consumption on hpc clusters using xdmod," in *Practice and Experience in Advanced Research Computing*, 2020, pp. 112–119.

[84] "Mpi 3.1 report," Jun 2015. [Online]. Available: https://www.mpi-forum.org/docs/mpi-3.1

[85] D. K. Panda, H. Subramoni, C.-H. Chu, and M. Bayatpour, "The mvapich project: Transforming research into high-performance mpi library for hpc community," *Journal of Computational Science*, p. 101208, 2020.

[86] R. L. Graham, T. S. Woodall, and J. M. Squyres, "Open mpi: A flexible high performance mpi," in *International Conference on Parallel Processing and Applied Mathematics*. Springer, 2005, pp. 228–239.

[87] D. Schafer, I. Laguna, and K. Mohror, "Exampi: A modern design and implementation to accelerate message passing interface innovation," in *High Performance Computing: 6th Latin American Conference, CARLA 2019, Turrialba, Costa Rica, September 25–27, 2019, Revised Selected Papers*, vol. 1087. Springer Nature, 2020, p. 153.

[88] M. Si, Y. Ishikawa, and M. Tatagi, "Direct mpi library for intel xeon phi co-processors," in *2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum*. IEEE, 2013, pp. 816–824.

[89] B. Elis, D. Yang, O. Pearce, K. Mohror, and M. Schulz, "Qmpi: a next generation mpi profiling interface for modern hpc platforms," *Parallel Computing*, p. 102635, 2020.

[90] S. Ramesh, A. Mahéo, S. Shende, A. D. Malony, H. Subramoni, A. Ruhela, and D. K. D. Panda, "Mpi performance engineering with the mpi tool interface: the integration of mvapich and tau," *Parallel Computing*, vol. 77, pp. 19–37, 2018.

[91] M. Schulz, J. A. Levine, P.-T. Bremer, T. Gamblin, and V. Pascucci, "Interpreting performance data across intuitive domains," in *Parallel Processing (ICPP), 2011 International Conference on*. IEEE, 2011, pp. 206–215.

[92] B. Tierney, W. Johnston, B. Crowley, G. Hoo, C. Brooks, and D. Gunter, "The netlogger methodology for high performance distributed systems performance analysis," 12 1999.

[93] M. Larsen, J. Aherns, U. Ayachit, E. Brugger, H. Childs, B. Geveci, and C. Harrison, "The alpine in situ infrastructure: Ascending from the ashes of strawman," in *Proceedings of the In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization Workshop*, ser. ISAV2017. New York, NY, USA: ACM, 2017.

[94] C. Xie and W. Xu, "Performance visualization for tau instrumented scientific workflows," Brookhaven National Lab.(BNL), Upton, NY (United States), Tech. Rep., 2018.

[95] O. Tuncer, E. Ates, Y. Zhang, A. Turk, J. Brandt, V. J. Leung, M. Egele, and A. K. Coskun, "Diagnosing performance variations in hpc applications using machine learning," in *International Supercomputing Conference*. Springer, 2017, pp. 355–373.

[96] A. Morrow, E. Baseman, and S. Blanchard, "Ranking anomalous high performance computing sensor data using unsupervised clustering," in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2016, pp. 629–632.

[97] S. Sanchez, A. Bonnie, G. Van Heule, C. Robinson, A. DeConinck, K. Kelly, Q. Snead, and J. Brandt, "Design and implementation of a scalable hpc monitoring system," in *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2016, pp. 1721–1725.

[98] G. da Cunha Rodrigues, G. Lessa dos Santos, V. Tavares Guimaraes, L. Zambenedetti Granville, and L. M. Rockenbach Tarouco, "An architecture to evaluate scalability, adaptability and accuracy in cloud monitoring systems," in *Information Networking (ICOIN), 2014 International Conference on*. IEEE, 2014, pp. 46–51.

[99] K. A. Huck, A. D. Malony, S. Shende, and A. Morris, "Taug: Runtime global performance data access using mpi," in *Recent Advances in Parallel Virtual Machine and Message Passing Interface*. Springer, 2006, pp. 313–321.

[100] M. Kutare, G. Eisenhauer, C. Wang, K. Schwan, V. Talwar, and M. Wolf, "Monalytics: online monitoring and analytics for managing large scale data centers," in *Proceedings of the 7th international conference on Autonomic computing*. ACM, 2010, pp. 141–150.

[101] W. Gu, G. Eisenhauer, E. Kraemer, K. Schwan, J. Stasko, J. Vetter, and N. Mallavarupu, "Falcon: On-line monitoring and steering of large-scale parallel programs," in *Frontiers of Massively Parallel Computation, 1995. Proceedings. Frontiers' 95., Fifth Symposium on the*. IEEE, 1995, pp. 422–429.

[102] X. Zhang, H. Abbasi, K. Huck, and A. D. Malony, "Wowmon: A machine learning-based profiler for self-adaptive instrumentation of scientific workflows," *Procedia Computer Science*, vol. 80, pp. 1507–1518, 2016.

[103] R. K. Tesser and P. O. A. Navaux, "Dimvhcm: An on-line distributed monitoring data collection model," in *Parallel, Distributed and Network-Based Processing (PDP), 2012 20th Euromicro International Conference on*. IEEE, 2012, pp. 37–41.

[104] M. Schulz, A. Bhatele, D. Böhme, P.-T. Bremer, T. Gamblin, A. Gimenez, and K. Isaacs, "A flexible data model to support multi-domain performance analysis," in *Tools for High Performance Computing 2014*. Springer, 2015, pp. 211–229.

[105] B. Husain, A. Giménez, J. A. Levine, T. Gamblin, and P.-T. Bremer, "Relating memory performance data to application domain data using an integration api," in *Proceedings of the 2nd Workshop on Visual Performance Analysis*. ACM, 2015, p. 5.

[106] A. Giménez, T. Gamblin, I. Jusufi, A. Bhatele, M. Schulz, P.-T. Bremer, and B. Hamann, "Memaxes: visualization and analytics for characterizing complex memory performance behaviors," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 7, pp. 2180–2193, 2017.

[107] D. Böhme, D. Beckingsdale, and M. Schulz, "Flexible data aggregation for performance profiling," *IEEE Cluster*, 2017.

[108] K. E. Isaacs, A. G. Landge, T. Gamblin, P.-T. Bremer, V. Pascucci, and B. Hamann, "Exploring performance data with boxfish," in *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*. IEEE, 2012, pp. 1380–1381.

[109] P. Messina, "The exascale computing project," *Computing in Science & Engineering*, vol. 19, no. 3, pp. 63–67, 2017.

[110] J. Ahrens, B. Geveci, and C. Law, "Paraview: An end-user tool for large data visualization," *The Visualization Handbook*, vol. 717, 2005.

[111] H. Childs, E. Brugger, B. Whitlock, J. Meredith, S. Ahern, D. Pugmire, K. Biagas, M. Miller, C. Harrison, G. H. Weber, H. Krishnan, T. Fogal, A. Sanderson, C. Garth, E. W. Bethel, D. Camp, O. Rübel, M. Durant, J. M. Favre, and P. Navrátil, "VisIt: An End-User Tool For Visualizing and Analyzing Very Large Data," in *High Performance Visualization—Enabling Extreme-Scale Scientific Insight*. CRC Press/Francis–Taylor Group, Oct. 2012, pp. 357–372.

[112] M. Larsen, E. Brugger, H. Childs, J. Eliot, K. Griffin, and C. Harrison, "Strawman: A batch in situ visualization and analysis infrastructure for multi-physics simulation codes," in *Proceedings of the First Workshop on In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization*, ser. ISAV2015. New York, NY, USA: ACM, 2015, pp. 30–35. [Online]. Available: http://doi.acm.org/10.1145/2828612.2828625

[113] L. L. N. Laboratory. (2017) Conduit: Simplified data exchange for hpc simulations. [Online]. Available: https://software.llnl.gov/conduit/

[114] ——. (2017) Conduit: Simplified data exchange for hpc simulations - conduit blueprint. [Online]. Available: https://software.llnl.gov/conduit/blueprint.html

[115] W. J. Schroeder, B. Lorensen, and K. Martin, *The visualization toolkit: an object-oriented approach to 3D graphics*. Kitware, 2004.

[116] A. Kunen, T. Bailey, and P. Brown, "Kripke-a massively parallel transport mini-app," Lawrence Livermore National Laboratory (LLNL), Livermore, CA, Tech. Rep., 2015.

[117] "Hydrodynamics Challenge Problem, Lawrence Livermore National Laboratory," Lawrence Livermore National Laboratory, Tech. Rep. LLNL-TR-490254.

[118] K. Moreland, C. Sewell, W. Usher, L.-t. Lo, J. Meredith, D. Pugmire, J. Kress, H. Schroots, K.-L. Ma, H. Childs *et al.*, "Vtk-m: Accelerating the visualization toolkit for massively threaded architectures," *IEEE computer graphics and applications*, vol. 36, no. 3, pp. 48–58, 2016.

[119] M. Larsen, C. Harrison, J. Kress, D. Pugmire, J. S. Meredith, and H. Childs, "Performance modeling of in situ rendering," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE Press, 2016, p. 24.

[120] P. Balaprakash, J. Dongarra, T. Gamblin, M. Hall, J. K. Hollingsworth, B. Norris, and R. Vuduc, "Autotuning in high-performance computing applications," *Proceedings of the IEEE*, vol. 106, no. 11, pp. 2068–2083, 2018.

[121] D. A. Beckingsale, O. Pearce, I. Laguna, and T. Gamblin, "Apollo: Reusable Models for Fast, Dynamic Tuning of Input-Dependent Code," in *31st IEEE International Parallel & Distributed Processing Symposium*, 2017, pp. 307–316.

[122] D. A. Beckingsale, R. D. Hornung, T. R. W. Scogland, and A. Vargas, "Performance Portable C++ Programming with RAJA," in *Proceedings of the 24th Symposium on Principles and Practice of Parallel Programming*, 2019, pp. 455–456.

[123] R. D. Hornung and J. A. Keasler, "The RAJA Portability Layer: Overview and Status," Lawrence Livermore National Lab, Tech. Rep., 2014.

[124] H. C. Edwards, C. R. Trott, and D. Sunderland, "Kokkos: Enabling manycore performance portability through polymorphic memory access patterns," *Journal of Parallel and Distributed Computing*, vol. 74, no. 12, pp. 3202–3216, 2014.

[125] H. C. Edwards and C. R. Trott, "Kokkos: Enabling performance portability across manycore architectures," in *2013 Extreme Scaling Workshop (xsw 2013)*. IEEE, 2013, pp. 18–24.

[126] I. Karlin, J. A. Keasler, and R. Neely, "Lulesh 2.0 updates and changes," Lawrence Livermore National Laboratory, Tech. Rep. LLNL-TR-641973, August 2013.

[127] D. A. Beckingsale, "Towards scalable adaptive mesh refinement on future parallel architectures," Ph.D. dissertation, University of Warwick, 2015.

[128] D. A. Beckingsale, W. P. Gaudin, J. A. Herdman, and S. A. Jarvis, "Resident Block-Structured Adaptive Mesh Refinement on Thousands of Graphics Processing Units," in *44th International Conference on Parallel Processing*, 2015, pp. 61–70.

[129] S. Rajamanickam, "Kokkos kernels: Performance portable kernels for sparse/dense linear algebra graph and machine learning kernels." Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), Tech. Rep., 2020.

[130] D. Beckingsale, W. Gaudin, A. Herdman, and S. Jarvis, "Resident block-structured adaptive mesh refinement on thousands of graphics processing units," in *2015 44th International Conference on Parallel Processing*. IEEE, 2015, pp. 61–70.

[131] R. Vuduc, J. W. Demmel, and K. A. Yelick, "Oski: A library of automatically tuned sparse matrix kernels," in *Journal of Physics: Conference Series*, vol. 16, no. 1. IOP Publishing, 2005, p. 521.

[132] M. A. S. Bari, N. Chaimov, A. M. Malik, K. A. Huck, B. Chapman, A. D. Malony, and O. Sarood, "Arcs: Adaptive runtime configuration selection for power-constrained openmp applications," in *2016 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2016, pp. 461–470.

[133] J. Ansel, S. Kamil, K. Veeramachaneni, J. Ragan-Kelley, J. Bosboom, U.-M. O'Reilly, and S. Amarasinghe, "Opentuner: An extensible framework for program autotuning," in *Proceedings of the 23rd international conference on Parallel architectures and compilation*, 2014, pp. 303–316.

[134] D. Pfander, M. Brunn, and D. Pflüger, "Autotunetmp: Auto-tuning in c++ with runtime template metaprogramming," in *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2018, pp. 1123–1132.

[135] J. E. Baldeschwieler, R. D. Blumofe, and E. A. Brewer, "Atlas: An infrastructure for global computing," in *Proceedings of the 7th workshop on ACM SIGOPS European workshop: Systems support for worldwide applications*, 1996, pp. 165–172.

[136] M. Frigo and S. G. Johnson, "Fftw: An adaptive software architecture for the fft," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 3. IEEE, 1998, pp. 1381–1384.

[137] V. Sreenivasan, R. Javali, M. Hall, P. Balaprakash, T. R. Scogland, and B. R. de Supinski, "A Framework for Enabling OpenMP Autotuning," in *International Workshop on OpenMP*. Springer, 2019, pp. 50–60.

[138] R. C. Whaley, A. Petitet, and J. J. Dongarra, "Automated empirical optimizations of software and the atlas project," *Parallel computing*, vol. 27, no. 1-2, pp. 3–35, 2001.

[139] A. Rasch, M. Haidl, and S. Gorlatch, "Atf: A generic auto-tuning framework," in *2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 2017, pp. 64–71.

[140] A. Rasch and S. Gorlatch, "Atf: A generic directive-based auto-tuning framework," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 5, p. e4423, 2019.

[141] A. Hartono, B. Norris, and P. Sadayappan, "Annotation-based empirical performance tuning using orio," in *2009 IEEE International Symposium on Parallel & Distributed Processing*. IEEE, 2009, pp. 1–11.

[142] J. Hollingsworth and A. Tiwari, "End-to-end auto-tuning with active harmony," *Performance Tuning of Scientific Applications*, pp. 217–238, 2010.

[143] K. Meng and B. Norris, "Mira: A framework for static performance analysis," in *2017 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2017, pp. 103–113.

[144] H. Menon, A. Bhatele, and T. Gamblin, "Auto-tuning parameter choices in hpc applications using bayesian optimization," 2020.

[145] T. Creech, A. Kotha, and R. Barua, "Efficient multiprogramming for multicores with scaf," in *2013 46th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2013, pp. 334–345.

[146] T. Creech and R. Barua, "Transparently space sharing a multicore among multiple processes," *ACM Trans. Parallel Comput.*, vol. 3, no. 3, Nov. 2016. [Online]. Available: https://doi.org/10.1145/3001910

[147] G. Georgakoudis, H. Vandierendonck, P. Thoman, B. R. D. Supinski, T. Fahringer, and D. S. Nikolopoulos, "Scalo: Scalability-aware parallelism orchestration for multi-threaded workloads," *ACM Trans. Archit. Code Optim.*, vol. 14, no. 4, Dec. 2017. [Online]. Available: https://doi.org/10.1145/3158643