Reliable Performance for Streaming Analysis Workflows

Kerstin Kleese van Dam (BNL), Darren Kerbyson (PNNL), Ilkay Altintas (SDSC), Kevin Barker (PNNL), Nathan Tallent (PNNL), Eric Stephan (PNNL), Jian Yin (PNNL)

Workflow systems are in wide use in the scientific community today, facilitating complex computational and analytical processes. Their increasing popularity is particularly visible at workflow sharing sites such as MyExperiment [1] or Galaxy [2-4]. High-performance computing (HPC) users also are looking toward workflow solutions to manage their complex pre- and post-processing needs. This trend likely will continue with the advent of exascale architectures, which will require extreme-scale collaborations between applications running on an exascale system and community data and knowledge repositories needed for their validation and steering [5, 6]. A new emerging use for workflows is the in-situ / streaming, often adaptive analysis of large scale simulation runs and as well as the need to analyze and interpret experimental results [10], in both cases to steer the scientific work and optimize the scientific outcome. In particular in this last case the reliably performance of the workflow is absolutely key to its usefulness.

One such use case is the steering of high end electron microscopy experiments at the Center for Functional Nanomaterials (CFN) at Brookhaven National Laboratory. In Transmission Electron Microscopy (TEM), a beam of electrons is transmitted through an ultra- thin specimen, interacting with the specimen as it passes through. These experiments can generate atomic resolution diffraction patterns, images and spectra under wide ranging environmental conditions. In-situ observations with these instruments, were physical, chemical or biological processes and phenomena are observed as they evolve, generate from 10GB-10's of TB (e.g. at BNL) of data per experiment (and getting larger) at rates ranging from 100 images/sec for basic instruments to 1600 images/sec for state of the art systems. To optimize the scientific outcome of such experiments it is essential to analyze and interpret the results as they are emerging. Infrastructures such as the Analysis in Motion framework [9] developed by PNNL can provide the necessary analytical frameworks, if they can deliver reliable performance.

However, it is becoming more difficult to design large-scale workflows for scientific computing that reliably deliver optimal performance, especially in situations where time-critical decisions must be made or computing resources are limited. Workflows, frequently composite applications built from loosely coupled parts, are designed to execute on a loosely connected set of distributed and heterogeneous computational resources. Each computational resource may have vastly diverse capabilities, ranging from sensors to high-performance clusters. Each workflow task may be designed for a different programming model and implemented in a different language, and most communicate via files sent over general purpose networks. As a result of this complex software and execution space, large-scale scientific workflows exhibit extreme performance variability. Going forward, it is critically important to have a clear understanding of the factors that influence workflow performance and sources for the potential variability in their execution to improve designs in advance and enable further optimization of workflow performance at runtime.

The DOE ASCR funded Integrated End-to-End Performance Prediction and Diagnosis for Extreme Scientific Workflows (IPPD) project is addressing this important issue. IPPD is developing and integrated approach to the modeling of extreme scale scientific workflows, bringing together modeling, simulation and empirical approaches. Its goal is to provide scientists wit the tools to:

- Explore in Advance - Design space exploration and sensitivity analysis
- Optimize at run-time - Guide execution based on dynamic behavior

To optimize workflow performance, we first and foremost need to understand the different sources of workflow performance variability and under performance. Our key interest is to identify patterns across different workflow classes that cause these issues, particularly in extreme-scale environments, to allow us to develop early-warning systems and optimization strategies both for workflow design and at runtime.

Empirical Studies - To gain a quantitative understanding of workflow performance variability and its sources we need to capture empirical information about classes of workflows and the behavior of the surrounding system execution environment. Traditionally, provenance has been largely focused on capturing workflow event history and tracing the data lineage from workflow results. On IPPD the scope has been expanded to collect metrics to gain greater insights into impacts of external factors influencing workflow behavior in clusters and in distributed workflow environments. The metrics are collected from literally hundreds of physical sensors (for instrumented clusters) or more commonly existing off the performance modeling tools that can be streamed varuomg rates into indexed time-series databases to help frame a composite picture of the ecosystem influencing workflow behavior. The IPPD provenance management solution (ProvEn) offers a horizontally scalable and load-balanced hybrid database solution for both the traditional semantic provenance describing the workflow history and the metrics streams describing environmental factors. As the situation warrants, metrics collection can take in a distributed fashion, collected at different rates, and older metrics can either be pushed to cold storage or metrics can be collected on a rolling window of time Our focus is to make these empirical measurements available for analytics and modeling teams in either ad-hoc or post-mortem fashion to as a means for workflows to adapt to changing conditions, or dynamically recover from soft/hard error conditions.[7]

Performance Modeling – A key capability necessary for the efficient execution of complex workflows is performance prediction. Because large-scale workflows are composed of a number of disparate parts, each with varying performance constraints and requirements, performance modeling is required at both the component and overall workflow levels. Because of this, a number of techniques and tools are required, ranging from low-level architectural performance modeling to workflow scale intelligent scheduling. These models must capture the behavior of each workload component, parameterize that behavior in terms of execution platform capabilities as well as input data characteristics, and most importantly, quantify the impact of contention for shared resources (e.g., file systems and networks) on task and workflow performance. To this end, the IPPD project is utilizing analytical modeling methods and tools, empirical measurements, and advanced optimization techniques analogous to Unit Commitment approaches used in power grid allocation strategies. Our goal is to improve overall workflow performance by using quantitative performance prediction to optimize the scheduling of workflow tasks in order to reduce contention on shared resources and improve utilization of distributed computing resources. [8]

For many scientific workflows in distributed environments, I/O is often the bottleneck. Many factors such as background disk loads and inter-connection networks from different sites can affect data transfer speeds and these factors can change dynamically. Furthermore, computation cannot commence prior to whole files being transferred preventing the overlap of computation and I/O, also increasing user waiting time. Our approach is to introduce an integrated and adaptive I/O layer at user-level to enable efficient pipelining of I/O and computation while dynamically adjusting data stripping across different geographically distributed sites. This layer intercepts I/O related requests and initiates data transfers from remote sites. Computation is resumed as soon as the transfer of the requested data is completed. Additionally, dynamically striping the data transfer from different remote sites is possible. To account for the impact of background load on disk and background network traffic on throughput for different sites, the system dynamically adjusts which sites are used and the proportions of data are retrieved from each site.

References

[1] Heinis, T.; Alonso, G. (2008). "Efficient lineage tracking for scientific workflows." In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (pp. 1007-1018). DOI: 10.1145/1376616.1376716.

[2] Anand, M. K.; Bowers, S.; McPhillips, T.; Ludäscher, B. (2009). "Efficient provenance storage over nested data collections." In Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (pp. 958-969). DOI: 10.1145/1516360.1516470.

[3] Anand, M. K.; Bowers, S.; Ludäscher, B. (2010). "Techniques for efficiently querying scientific workflow provenance graphs." In Proceedings of the 13th International Conference on Extending Database Technology (Vol. 10, pp. 287-298). DOI: 10.1145/1739041.1739078.

[4] Ikeda, R.; Park, H.; Widom, J. (2011). "Provenance for generalized map and reduce workflows." In Online Proceedings, Fifth Biennial Conference on Innovative Data Systems Research (CIDR) (pp. 273-283).

[5] Park, H.; Ikeda, R.; Widom, J. (2011). "RAMP: A System for Capturing and Tracing Provenance in MapReduce Workflows." In Proceedings of the VLDB Endowment, 4(12), 1351-1354.

[6] Ogasawara, E.; Dias, J.; Oliveira, D.; Porto, F.; Valduriez, P.; Mattoso, M. (2011). "An algebraic approach for data-centric scientific workflows." In Proceedings of the VLDB Endowment, 4(12), 1328-1339.

[7] Kleese van Dam, K., Stephan, E., Raju, B., Altintas, I., Elsethagen, T., Krishnamoorthy, S. November 2015. Enabling Structured Exploration of W orkflow Performance V ariability in Extreme-scale Environments. In proceedings MTAGS15, collocated with SC15

[8] Mahantesh Halappanavar, Malachi Schram, Luis de la Torre, Kevin Barker, Nathan R. Tallent, and Darren J. Kerbyson. 2015. Towards efficient scheduling of data intensive high energy physics workflows. In *Proceedings of the 10th Workshop on Workflows in Support of Large-Scale Science* (WORKS '15). ACM, New York, NY, USA, , Article 3 , 9 pages.

[9] Kerstin Kleese van Dam · Ryan LaMothe · Poorva Sharma · Dimitri Zarzhitsky · Abhinav Vishnu · Eric Stephan · Will Smith · Todd Elsethagen · Mathew Thomas**.** Building the Analysis in Motion Infrastructure. Report number: PNNL-24340, Affiliation: Pacific Northwest National Laboratory, JUNE 2015. DOI: 10.13140/RG.2.1.1701.1368

[10] Mathew Thomas · Kerstin Kleese-van Dam · Matthew J. Marshall · Andrew Kuprat · James Carson · Carina Lansing · Zoe Guillen · Erin Miller · Ingela Lanekoff · Julia Laskin. Towards Adaptive, Streaming Analysis of X-ray Tomography Data. ARTICLE *in* SYNCHROTRON RADIATION NEWS 28(2) · FEBRUARY 2015