Panorama 360 Final Report
Funded by DOE under grant # DE-SC0012636
Ewa Deelman, USC, PI

Period of Performance: 09/01/2018 - 08/31/2021

Accomplishments

Vision : Provide a resource for the collection, analysis, and sharing of performance data about end to end scientific workflows executing on DOE facilities.

Goals: Develop a repository and associated capabilities for data collection, ingestion, and analysis for a broad class of DOE applications that span experimental and simulation science workflows. In particular, this work focuses on workflows that include experimental data generation at DOE facilities. The main activities of Panorama 360 include the development of:

1. A distributed repository that stores different types of workflow execution data (e.g., point and time series performance traces at fine- and coarse-grained levels);
2. A set of open-source data capture, curation, and publishing tools fully integrated with a state-of-the-art workflow management system that automates data ingestion to the repository and enables users to discover, query, and process data from the repository;
3. A set of analysis algorithms and machine learning based tools to perform analysis and characterization of the gathered data, which can be used to detect anomalous performance or system faults; and
4. Best practices and recommendations for workflow evaluation, analysis, execution, and architectures.

Accomplishments. During this reporting period, progress was made in multiple areas of the project:

1. **Data Capture:** We have designed and developed an architecture to collect statistics and performance data about the workflow execution in near real-time. Apart from planning and executing a workflow, this architecture uses the Pegasus WMS to orchestrate data collection into an open access repository. Our end-to-end monitoring architecture includes mechanisms to collect different classes of data sources including workflow monitoring events (Pegasus), resource utilization (Pegasus Kickstart), application-level network statistics (Globus), low level network statistics (TSTAT), and application I/O behavior (Darshan). All of the statistics and performance data are published to an AMQP endpoint so they can further be processed and ingested to the Panorama 360 data repository.

2. **Data repository:** Using state-of-the-art search and visualization technologies, we have created a repository to ingest, browse, and visualize performance data captured during the execution of workflow applications by using a flexible data format (JSON). The repository uses the ELK stack (Elasticsearch, Logstash, Kibana) to provide search, ingestion, and visualization capabilities. We capitalize on Logstash capabilities to provide custom ingestion routines, and on Kibana features to create a custom plugin for visualizing fine-grained and coarse-grained workflow performance data at workflow and job levels. The architecture and repository have been presented as a lightning talk at the WORKS 2018 workshop at Super Computing 2018, and followed up with a publication for the Future Generation Computer System (FGCS) journal.
3. **Building Models:** Using the ExoGENI testbed and Cori at NERSC, we have collected data from workflow events, resource usage, and network statistics. We have explored a number of machine learning and statistical techniques to analyze these data and the associated workflow performances. We have developed techniques for classification of normal and abnormal traffic by exploring techniques such as Principal Component Analysis (PCA), Autoencoders, Isolation Forests, and Decision and Random Forest trees. Gathering labeled data sets from simple transfer deployments, as well as from workflow executions (NERSC), we created multiple classifiers based on different techniques, which lead to better identification of abnormal transfers. Early results of applying PCA, Autoencoders, and Isolation Forest were published at the ICML 2018 workshop on machine learning for engineering systems. These results were also presented at MODSIM 2018 as a talk. Subsequently, we have improved the techniques and published the work in the Journal of Machine Learning.
4. **Analysis Algorithms:**
    a. *Wide-area network transfer anomaly using TCP data:* We collected and evaluated TCP statistics from real network transfers to understand distinguishing behaviors of how TCP behaves in certain anomalous conditions. Using a science workflow and the ExoGENI testbed, we set up real network transfers and introduced packet loss, reordering, and duplication to record 133 TCP Cubic characteristics per transfer. After experimenting with a number of unsupervised classification techniques, we used decision tree classification to extract relationships between the TCP variables across the samples. This technique led to results that improved our understanding of the underlying behavior of anomalous TCP flows. Our results have shown that while average roundtrip time (RTT) and congestion window are the most favored monitored variables, other variables such as number of retransmits and first acknowledgements recorded can provide an early indication of packet loss and reordering anomalies. Our decision tree classification was able to extract unique relationships among the

133 TCP features and also build feature relationships that can characterize TCP behaviors under abnormalities. We have conducted these experiments for a wide range of TCP variants - Cubic, Reno, Hamilton, and recently, BBR. Exploring these variants has unlocked new findings which allow us to see how different TCP variants handle packet anomalies.

b. *Workflow-level performance analysis for anomaly detection:* We used workflow-level analysis to predict overall behavior of a running workflow by clustering statistically similar workflows. We leveraged the Pegasus monitoring API and the data capture framework described above to collect the following workflow-level metrics to constitute a feature vector: percentage of successful and failed jobs in the workflow, average duration of the successful and failed jobs, and overall job success rate of the workflow. Using the k-means algorithm, we performed unsupervised clustering of this data to produce a set of clusters corresponding to different failure levels (which can be used for online classification of workflows). We conducted several experiments with different failure injection scenarios for two workflows, Montage and 1000 Genome, and performed the clustering of the collected data. Our results showed clusters for different levels of failures injected. We evaluated the quality of the clustering algorithms with standard scores like the adjusted rand scores, mutual info scores, homogeneity scores, etc. This work was published at the 2020 IEEE High Performance Extreme Computing Conference.

c. *Characterizing reliability of file transfers with various TCP variants:* We have conducted experiments with elephant and mice flows to understand how TCP behaves in each of the file scenarios in anomalous conditions. We have published this work at the INDIS workshop at SuperComputing 2019.

d. *Anomaly Detection in Scientific Workflows using End-to-End Execution Gantt Charts and Convolutional Neural Networks:* We developed a novel approach to tackle this problem by applying Convolutional Neural Network (CNN) classification methods to high-resolution visualizations that capture the end-to-end workflow execution timeline. Subtle differences in the timeline reveal information about the performance of the application and infrastructure's components. We collected 1000 traces of a scientific workflow's executions. We explored and evaluated the performance of CNNs trained from scratch and pre-trained on ImageNet. Our initial results are promising with over 90% accuracy. This work was published at the ACM Practice and Experience in Advanced Research Computing (PEARC 2021) Conference.

5. **Machine Learning Tools:** In our analysis, we experimented with a number of machine learning techniques chosen based on our data source, processing time, and the classification accuracy achieved. Starting with unsupervised deep learning approaches

such as autoencoders, we found that simpler techniques such as PCA and tree classification yielded better results (possibly due to the limited data we used). We have made these techniques available as a library and can plug-in any input data to process the classification results. We are currently improving these codes further to allow analysis of much larger and distributed data sources to build better classification models.

6. **DOE Facility Engagement:** Accessing OLCF resources with Pegasus WMS has been difficult for DOE scientists for two reasons: 1) because of the requirement to install and configure Pegasus' software stack (Pegasus and High Throughput Condor); and 2) because of the issues arising from two-factor authentication while orchestrating remote submissions. We worked on two new approaches for deploying Pegasus workflows on OLCF systems. The first approach takes advantage of the Kubernetes cluster at OLCF which allows us to run services within containers. We have created container recipes that can function as local submit nodes, enabling users to submit Pegasus workflows to Titan and Rhea (we are currently working on accessing Summit). The second approach aims to remedy the issue of remote submissions to OLCF. We are currently integrating Pegasus with the PanDA workload manager (http://news.pandawms.org/bigpanda.html) which allows users to remotely submit computational tasks to OLCF systems in a secure way. In collaboration with Brookhaven National Lab, Oak Ridge National Lab, and CERN, we are extending HTCondor's GAHP module to allow Pegasus to interact with the Panda server and submit jobs remotely to Titan and Summit in a way that is transparent to the user. The work on using a workflow submit node as a service was published at the PEARC 2020 conference.
7. **Workflow Development:** We have been collaborating with Spallation Neutron Source scientists at ORNL by developing new workflows. In particular, we evaluated the use of Dakota (https://dakota.sandia.gov/) as an optimization tool for uncertainty quantification. Based on these efforts, we extended the SNS-Nanodiamond workflow to use Dakota to optimize the force field and temperature variables during the simulations. Additionally, we have engaged in exploring new workflow concepts. Using MCViNE (http://www.mcvine.org/), we worked on a workflow to model simulation data to experimental data from the Wide Angular-Range Chopper Spectrometer (ARCS) instrument at SNS. Finally, we explored a workflow to optimize Monte-Carlo simulations of neutron scattering instruments to data collected at SNS using Javelin (https://github.com/rosswhitfield/javelin).

Opportunities for training and professional development. Mr. George Papadimitriou completed a summer internship at Oak Ridge National Laboratory (Summer 2018) under the mentorship of Dr. Jeffrey Vetter. Mr. Papadimitriou worked closely with SNS scientists developing new workflows. He also created and evaluated a local workflow execution environment located in the

infrastructure of the Future Technologies Group at ORNL. Finally, Panorama 360 sponsored Mr. Papadimitriou's attendance at the Super Computing 2018. Panorama 360 also provided Mr. Ashwin Vankatesha with the opportunity to acquire research experience in the area of Data Science and Machine Learning (Spring 2018). Mr. Vankatesha analyzed workflow execution traces in order to understand underlying execution patterns. The project also supported the work of Patrycja Krawczuk at USC. She conducted research in using machine learning techniques to conduct anomaly detection in scientific workflows.

Result dissemination to communities of interest. Within the scope of the Panorama 360 project, multiple efforts have been made to disseminate the project's outcomes to communities of interest. In addition to the journal, conference, and poster publications, we have engaged in the following activities:

- February 2018: A Pegasus WMS hands-on tutorial took place in Oak Ridge National Laboratory
  https://scitech.group/files/presentations/2018/2018-02-23-pegasus-tutorial-ornl.pdf
- June 2018: A Pegasus WMS and Panorama 360 tutorial was given under the Spallation Neutron Source - Data Science Seminar Series, targeting SNS scientists
  https://scitech.group/files/presentations/2018/2018-06-19-pegasus-seminar-ornl.pdf
- June 2018: A Pegasus WMS hands-on tutorial took place in Oak Ridge National Laboratory.
  https://scitech.group/files/presentations/2018/2018-06-21-pegasus-tutorial-ornl.pdf
- July 2018: A talk and a paper were presented at ICML workshop of machine learning in engineering systems, presenting a comparison of classification techniques.
- Aug 2018: A talk and poster were presented at MODSIM 2018, about unsupervised classification technique comparison in network transfer data.
- November 2018: A Panorama 360 presentation was given at both RENCI's and USC's exhibit booth at the Supercomputing 2019 conference.
  https://scitech.group/files/presentations/2018/2018-11-works18.pdf
- November 2018: A lightning talk was given at WORKS 2018 regarding Panorama 360's data collection architecture.
  https://scitech.group/files/presentations/2018/2018-11-works18.pdf
- March 2019: A presentation about Panorama 360 was given during the webinar organized by Eric Pouyoul under the umbrella of NGNS data collaboration initiative.
- March 2019: Pegasus Office Hours - End-to-End Workflow Monitoring
  https://scitech.group/files/presentations/2019/pegasus-office-hours-monitoring.pdf
  https://www.youtube.com/watch?v=fir9ZCxK9Gg
- May 2019: A poster was presented at the OLCF User Meeting 2019, about accessing OLCF's resources (Titan, Rhea, Summit) using the Pegasus WMS

- June 2020: A presentation about Pegasus workflows and execution environments was given at the WoWoHa workshop.

Products

Journal Publications.

- *The role of machine learning in scientific workflows*
  E. Deelman, A. Mandal, M. Jiang, R. Sakellariou, The International Journal of High Performance Computing Applications, 2019.
  https://doi.org/10.1177/1094342019852127
- *Using simple PID-inspired controllers for online resilient resource management of distributed scientific workflows*
  R. Ferreira da Silva, R. Filgueira, E. Deelman, E. Pairo-Castineira, I. M. Overton, M. Atkinson, Future Generation Computer Systems, vol. 95, pp. 615-628, 2019.
  https://doi.org/10.1016/j.future.2019.01.015
- Deelman, E., Vahi, K., Rynge, M., Mayani, R., Ferreira da Silva, R., Papadimitriou, G., & Livny, M. (2019). The Evolution of the Pegasus Workflow Management Software. Computing in Science Engineering, 21(4), 22–36. https://doi.org/10.1109/MCSE.2019.2919690
- Ferreira da Silva, R., Callaghan, S., Do, T. M. A., Papadimitriou, G., & Deelman, E. (2019). Measuring the Impact of Burst Buffers on Data-Intensive Scientific Workflows. Future Generation Computer Systems, 101, 208–220. https://doi.org/10.1016/j.future.2019.06.016
- Kiran, M., Wang, C., Papadimitriou, G., Mandal, A., & Deelman, E. (2020). Detecting Anomalous Packets in Network Transfers: Investigations using PCA, Autoencoder and Isolation Forest in TCP. Machine Learning. https://doi.org/10.1007/s10994-020-05870-y
- Ferreira da Silva, R., Casanova, H., Orgerie, A.-C., Tanaka, R., Deelman, E., & Suter, F. (2020). Characterizing, Modeling, and Accurately Simulating Power and Energy Consumption of I/O-intensive Scientific Workflows. Journal of Computational Science, 44, 101157. https://doi.org/10.1016/j.jocs.2020.101157
- Papadimitriou, G., Wang, C., Vahi, K., Ferreira da Silva, R., Mandal, A., Zhengchun, L., Mayani, R., Rynge, M., Kiran, M., Lynch, V. E., Kettimuthu, R., Deelman, E., Vetter, J. S., & Foster, I. (2021). End-to-End Online Performance Data Capture and Analysis for Scientific Workflows. Future Generation Computer Systems, 117, 387–400. https://doi.org/10.1016/j.future.2020.11.024
- Do, T. M. A., Pottier, L., Caíno-Lores, S., Ferreira da Silva, R., Cuendet, M. A., Weinstein, H., Estrada, T., Taufer, M., & Deelman, E. (2021). A Lightweight Method for Evaluating In

Situ Workflow Efficiency. Journal of Computational Science, 48, 101259. https://doi.org/10.1016/j.jocs.2020.101259

- Coleman, T., Casanova, H., Pottier, L., Kaushik, M., Deelman, E., & Ferreira da Silva, R. (2022). WfCommons: A framework for enabling scientific workflow research and development. Future Generation Computer Systems, 128, 16–27. https://doi.org/10.1016/j.future.2021.09.043

Conference Publications.

- *IoT-Hub: New IoT data-platform for Virtual Research Environments*
  R. Filgueira, R. Ferreira da Silva, E. Deelman, V. Christodoulou, and A. Krause, 10th International Workshop on Science Gateways (IWSG 2018), 2018, Edinburgh, UK, June 13-15, 2018
- *Ferreira da Silva, R., Orgerie, A.-C., Casanova, H., Tanaka, R., Deelman, E., & Suter, F. (2019). Accurately Simulating Energy Consumption of I/O-intensive Scientific Workflows. Computational Science – ICCS 2019, 138–152. https://doi.org/10.1007/978-3-030-22734-0_11*
- Papadimitriou, G., Kiran, M., Wang, C., Mandal, A., & Deelman, E. (2019). Training Classifiers to Identify TCP Signatures in Scientific Workflows. 2019 IEEE/ACM Innovating the Network for Data-Intensive Science (INDIS), 61–68. https://doi.org/10.1109/INDIS49552.2019.00012
- Wang, C., Papadimitriou, G., Kiran, M., Mandal, A., & Deelman, E. (2020). Identifying Execution Anomalies for Data Intensive Workflows Using Lightweight ML Techniques. 2020 IEEE High Performance Extreme Computing Conference (HPEC), 1–7. https://doi.org/10.1109/HPEC43674.2020.9286139
- Vahi, K., Goldstein, D., Papadimitriou, G., Nugent, P., & Deelman, E. (2020). Gearing the DECam Analysis Pipeline for Multi-Messenger Astronomy using Pegasus Workflows. Astronomical Data Analysis Software and Systems (ADASS) XXIX.
- Papadimitriou, G., Vahi, K., Kincl, J., Anantharaj, V., Deelman, E., & Wells, J. (2020). Workflow Submit Nodes as a Service on Leadership Class Systems. Proceedings of the Practice and Experience in Advanced Research Computing. https://doi.org/10.1145/3311790.3396671
- Ferreira da Silva, R., Pottier, L., Coleman, T., Deelman, E., & Casanova, H. (2020). WorkflowHub: Community Framework for Enabling Scientific Workflow Research and Development. 2020 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS), 49–56. https://doi.org/10.1109/WORKS51914.2020.00012
- Tu, H., Papadimitriou, G., Kiran, M., Wang, C., Mandal, A., Deelman, E., & Menzies, T. (2021). Mining Workflows for Anomalous Data Transfers. 2021 2021 IEEE/ACM 18th

International Conference on Mining Software Repositories (MSR) (MSR), 1–12. https://doi.org/10.1109/MSR52588.2021.00013

- Krawczuk, P., Papadimitriou, G., Tanaka, R., Do, T. M. A., Subramany, S., Nagarkar, S., Jain, A., Lam, K., Mandal, A., Pottier, L., & Deelman, E. (2021). A Performance Characterization of Scientific Machine Learning Workflows. 2021 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS), 58–65. https://doi.org/10.1109/WORKS54523.2021.00013
- Krawczuk, P., Papadimitriou, G., Nagarkar, S., Kiran, M., Mandal, A., & Deelman, E. (2021). Anomaly Detection in Scientific Workflows using End-to-End Execution Gantt Charts and Convolutional Neural Networks. Practice and Experience in Advanced Research Computing. https://doi.org/10.1145/3437359.3465597
- Do, T. M. A., Pottier Loïc, Ferreira da Silva, R., Caíno-Lores Silvina, Taufer, M., & Deelman, E. (2021). Assessing Resource Provisioning and Allocation of Ensembles of In Situ Workflows. 50th International Conference on Parallel Processing Workshop. https://doi.org/10.1145/3458744.3474051
- Casanova, H., Deelman, E., Gesing, S., Hildreth, M., Hudson, S., Koch, W., Larson, J., McDowell, M. A., Meyers, N., Navarro, J.-L., Papadimitriou, G., Tanaka, R., Taylor, I., Thain, D., Wild, S. M., Filgueira, R., & da Silva, R. F. (2021). Emerging Frameworks for Advancing Scientific Workflows Research, Development, and Education. 2021 IEEE Workshop on Workflows in Support of Large-Scale Science (WORKS), 74–80. https://doi.org/10.1109/WORKS54523.2021.00015

Posters.

- *Pegasus WMS and Panorama 360*
  G. Papadimitriou, E. Deelman, R. Ferreira da Silva, R. Filgueira, V. Lynch, A. Mandal, C. Wang, J. Vetter, USC Viterbi Annual Research Review. March 2018
  https://scitech.group/files/posters/2018/pegasus-wms-panorama360.pdf
- *PANORAMA 360: Performance Data Capture and Analysis for End-to-End Scientific Workflows*
  G. Papadimitriou, R. Ferreira da Silva, K. Vahi, R. Mayani, M. Rynge, E. Deelman, A. Mandal, C. Wang, M. Kiran, V. Lynch, J. Vetter, DOE NGNS PI Meeting. September 2018
  https://scitech.group/files/posters/2018/ngns-pi-meeting-poster-2018.pdf
- *Accessing OLCF Resources Using Pegasus WMS*
  George Papadimitriou, Pavlo Svirin, Karan Vahi, Mats Rynge, Rafael Ferreira da Silva, Jason Kincl, Vickie Lynch, Ewa Deelman, Anirban Mandal, Jeffrey Vetter, Valentine Anantharaj, Jack Wells, Alexei Klimentov, Kaushik De, OLCF User Meeting 2019. May 2019
  https://scitech.group/files/posters/2019/pegasus-oclf-user-meeting-2019.pdf
- Gearing the DECam Analysis Pipeline for Multi-Messenger Astronomy using Pegasus Workflows. K. Vahi, D. Goldstein, G. Papadimitriou, P. Nugent, and E. Deelman,

Astronomical Data Analysis Software and Systems (ADASS) XXIX, , 2019. https://scitech.group/files/posters/2019/2019-adass-pegasus-decam-poster.pdf

Websites.

- Panorama360 Website: https://panorama360.github.io
- Panorama360 GitHub: https://github.com/Panorama360

Intellectual property. *N/A*

Technologies or Techniques.

- *Pegasus monitoring daemon:* We have extended Pegasus' monitoring daemon and added the capability to publish monitoring events in JSON to AMQP endpoints.
- *Pegasus Transfer:* We have extended Pegasus transfer with Globus support and the capability to publish events/statistics from the Globus service to AMQP endpoints.
- *Pegasus Darshan:* We have created a new tool that parses Darshan's logs. This tool runs on the execution site and is triggered after an MPI run, by using a wrapper script to the job. It parses the generated Darshan logs (in binary format) and propagates a subset of the information (STDIO and POSIX-IO statistics) in JSON format to Pegasus' monitoring daemon.
- *Panorama 360 Kibana Plugin:* We have created a Kibana extension plugin, that adds a workflow performance explorer with custom visualization of the collected workflow statistics.
- *Repository - Local Deployment:* We have used container automation technologies (Docker and Docker Compose) to provide an easy way of deploying Panorama 360's repository backend (Rabbitmq, Elasticsearch, Kibana, Logstash) preconfigured and ready to collect data.
- *Workflow Traces:* A set of scripts have been developed to automatically combine performance data stored in Elasticsearch from a single workflow run into a JSON format file. This file, called workflow trace, provides the necessary metrics and workflow structure information to reproduce the workflow execution on, for example, a simulated environment. It can also be used to drive the development of synthetic workflows.
- *Machine learning library:* Classification techniques are available to understand network TSTAT data and produce classification results. The library includes PCA< autoencoder, isolation forest, decision and random forest trees.

Other products.

- **Open Access Data [USC]:** In order to make workflow traces publicly available, we provide open access to an instance of the repository containing data from workflow executions

conducted for the research paper submitted and presented as a lightning talk at the WORKS 2018 workshop. The datastore contains traces from the Nanodiamond workflow (used by the Spallation Neutron Source scientists), and the 1000-Genome workflow (a genomics workflow).

- Kibana: https://kibana.panorama.isi.edu
- Elasticsearch: https://data.panorama.isi.edu (only programmatically via api calls)

Honors and Awards

- **Best Student Paper Award at PEARC '20 Conference**
  - Papadimitriou, G., Vahi, K., Kincl, J., Anantharaj, V., Deelman, E., & Wells, J. (2020). Workflow Submit Nodes as a Service on Leadership Class Systems. Proceedings of the Practice and Experience in Advanced Research Computing. https://doi.org/10.1145/3311790.3396671
- **Editor's choice article - Future Generation Computer Systems (2021)**
  - Papadimitriou, G., Wang, C., Vahi, K., Ferreira da Silva, R., Mandal, A., Zhengchun, L., Mayani, R., Rynge, M., Kiran, M., Lynch, V. E., Kettimuthu, R., Deelman, E., Vetter, J. S., & Foster, I. (2021). End-to-End Online Performance Data Capture and Analysis for Scientific Workflows. Future Generation Computer Systems, 117, 387–400. https://doi.org/10.1016/j.future.2020.11.024
- **Best Paper Award - Future Generation Computer Systems (2021)**
  - Papadimitriou, G., Wang, C., Vahi, K., Ferreira da Silva, R., Mandal, A., Zhengchun, L., Mayani, R., Rynge, M., Kiran, M., Lynch, V. E., Kettimuthu, R., Deelman, E., Vetter, J. S., & Foster, I. (2021). End-to-End Online Performance Data Capture and Analysis for Scientific Workflows. Future Generation Computer Systems, 117, 387–400. https://doi.org/10.1016/j.future.2020.11.024

Participants and Collaborating Organizations

Participants.

- Dr. Ewa Deelman (University of Southern California)
- Dr. Jeffrey Vetter (Oakridge National Laboratory)
- Dr. Anirban Mandal (RENCI, University of North Carolina at Chapel Hill)
- Dr. Mariam Kiran (Lawrence Berkeley National Laboratory - ESnet)
- Dr. Rafael Ferreira da Silva (University of Southern California)
- Dr. Vickie Lynch (Oakridge National Laboratory)
- Dr. Cong Wang (RENCI, University of North Carolina at Chapel Hill)
- Mrs. Komal Thareja (RENCI, University of North Carolina at Chapel Hill)

- Dr. Paul Ruth (RENCI, University of North Carolina at Chapel Hill)
- Mr. Karan Vahi (University of Southern California)
- Mr. Rajiv Mayani (University of Southern California)
- Dr. Shirley Moore (Oakridge National Laboratory)
- Mr. George Papadimitriou (University of Southern California)
- Mrs. Patrycja Krawczuk (University of Southern California)
- Mr. Ashwin Vankatesha (University of Southern California)

Partners.

-- Big Panda Team, BNL: Pavlo Svirin, Alexei Klimentov, Kaushik De, integrating Pegasus with Panda
-- ZTF team: Peter Nugent, LBNL, Danny Goldstein, Caltech: developing workflows for ZFT, developing a ZFT-based benchmark
-- OLCF Team: Jack Wells, Val Anantharaj, developing a Pegasus ORNL/Summit interface

Impact

Impact on the development of the principal discipline(s) of the project. This project helps define the fundamental questions for workflow science, particularly in the area of performance modeling, anomaly detection, and adaptation.

Network and TCP algorithms: Our machine learning analysis has revealed new characteristics on how TCP variants behave under packet anomaly conditions. We studied specific network file transfers such as elephant and mice flows to validate these TCP behaviors and aim for producing innovative findings targeting the network research community. Our vision for this work may lead to new recommendations of TCP algorithm capabilities and how the maximum network bandwidth can be utilized in scientific workflow experiments to prevent anomalies that affect workflow performance.

ML for workflows: Our work has made initial inroads into a relatively unexplored area of research - use of machine learning based analysis for workflow performance management and failure detection. We have identified key workflow-level performance indicators, built automatic data collection tools to collect those relevant metrics, made them available to the community, and developed initial analysis methodologies. Our work has also advanced the state of the art in the area of anomaly detection, especially for workflow executions where several sources of

performance data, both from the infrastructure and from the application, need to be convolved using ML methodologies to detect and diagnose performance anomalies.

Impact on other disciplines. Scientific workflows provide support for advances in virtually all fields of science and engineering, many of which are key to the development and the sustainability of human societies. Our ability to execute workflows on OLCF infrastructures has the potential to impact a number of scientific disciplines, including astronomy, material science, chemistry, and others. Additionally, by identifying bottlenecks and performance issues on state-of-the-art systems, the Panorama 360 project has the potential to improve the effectiveness, and possibly the broader impact, of the execution of complex workflow applications in current and emerging platforms.

Impact on technology transfer. Open source software development for Pegasus and other critical software. The monitoring tools have been integrated into the Pegasus software stack and made available to users.