PANORAMA: Predictive Modeling and Diagnostic Monitoring of Extreme Science Workflows

Ewa Deelman, Christopher Carothers, Anirban Mandal, Brian Tierney, Jeffrey Vetter, Ilya Baldin, Mark Blanco
Rafael Ferreira da Silva, Mariam Kiran, Vickie Lynch, Shirley Moore, Paul Ruth

USC Viterbi School of Engineering Information Sciences Institute · renci · Oak Ridge National Laboratory · Rensselaer · Berkeley Lab

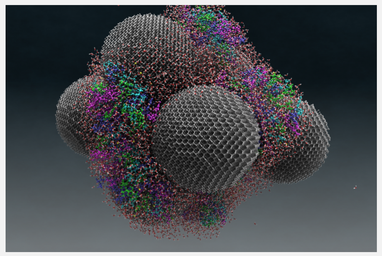
PANORAMA
Overview of the Project Description

The Panorama project aims to further the understanding of the behavior of scientific workflows as they are executing in heterogeneous environments. Panaroma's approach to modeling and diagnosing the runtime performance of complex scientific workflows is to integrate extreme-scale systems testbed experimentation, structured analytical modeling and parallel systems simulation into a comprehensive workflow framework that can characterize the **end-to-end workflow performance** on today's and future generation architectures, which can be used to improve the overall workflow performance and reliability. The Panaroma architecture includes the individual framework components: the Aspen analytical application modeling software, the ROSS simulation framework, the Pegasus workflow management system, and how they are used to model the behavior of DOE-relevant applications. By having a coupled model of the application and execution environment, decisions can be made about resource provisioning, application task scheduling, data management within the application, etc. Our approach for correlating the real time application and infrastructure monitoring data can be used to verify application behavior, perform anomaly detection and diagnosis, and support adaptivity during workflow execution.

IMPACT ON DOE SCIENCE
Diamonds that deliver!

Panorama enabled cutting-edge domain science research and development that has the potential to solve some of the challenges associated with **drug discovery and delivery**:

- The motions of a tRNA (or transfer RNA) model system can be enhanced when coupled with nanodiamonds, or diamond nanoparticles approximately 5 to 10 nanometers in size

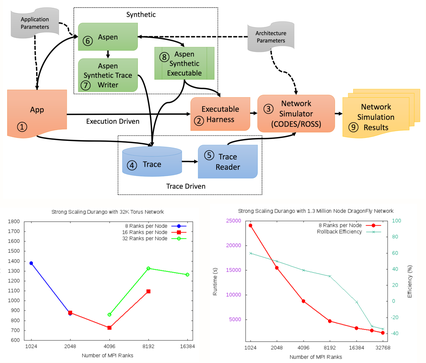


Water is seen as small red and white molecules on large nanodiamonds spheres. The colored tRNA can be seen on the nanodiamond surface. Image :Michael Mattheson, ORNL (https://www.ornl.gov/news/diamonds-deliver).

- We have developed an SNS Pegasus workflow to confirm that nanodiamonds enhance the dynamics of tRNA when in the presence of water. The workflow calculates the epsilon which best matches experimental data. These calculations used almost **400,000 CPU hours on a Cray XE6at NERSC**.
- The workflow runs NAMD parallel simulations, which varies the epsilon between -0.01 and -0.19 for each temperature specified (it requires 800 cores: equilibrium runs take ~1.5hs and production runs 12-16hs). AMBER's cpptraj removes global translation and rotation, and SASSENA calculates neutron scattering intensities from the trajectories (400 cores, 3-6hs). This workflow was used to computer 4 temperatures between 260K and 300K, which generated **~3TB of data**.

MODELING AND SIMULATION
Scalable Workload Generation for Application Performance Modeling and Simulation

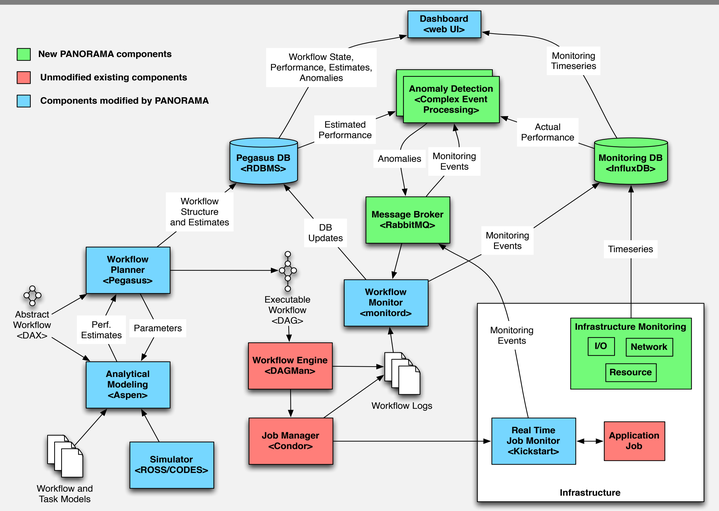
- We have created a new technique for generating scalable workloads from real applications, and implemented a prototype, called Durango, using a performance modeling toolkit.
- We demonstrate the efficacy of Durango's direct integration approach, which links Aspen into CODES as part of the running network simulation model. Here, Aspen generates the application-level computation timing events, which in turn drives the start of a network communication phase.



Durango in direct integration mode with 32K node torus network and Aspen compute node generator for 1K to 16K MPI ranks.

Durango in direct integration mode with 1.3M node dragonfly network and Aspen compute node generator for 1K to 32K MPI ranks

SYSTEM DESIGN
Workflow System, Infrastructure Monitoring, Analytical Modeling, Simulation



- Pegasus interfaces with Aspen to estimate resource requirements of individual workflow tasks as well as the entire workflow
- Aspen interfaces with ROSS/CODES to simulate network behavior not easily modeled using analytical techniques
- Workflow and infrastructure monitoring data is stored in InfluxDB and Pegasus DB
- Anomaly detection process monitors data stream and generates anomaly notifications, which are displayed in the web dashboard

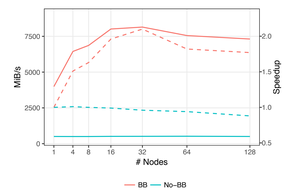
BURST BUFFERS
On the use of Burst Buffers for Accelerating Scientific Workflow Executions

Burst Buffers (BB) have emerged as a non-volatile storage solution that is positioned between the processors' memory and the PFS, buffering the large volume of data produced by the application at an higher rate than the PFS, while seamlessly draining the data to the PFS asynchronously.

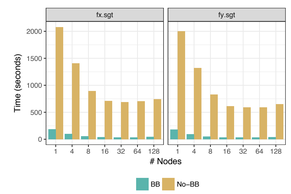
We explored the impact of Burst Buffers (BB) in scientific workflow applications. Using a software stack including Pegasus-WMS and HT-Condor, we ran a workflow on the Cori system at NERSC which included provisioning and releasing remote-shared BB nodes. Our application wrote and read about 550 GB of data.

Major Findings:

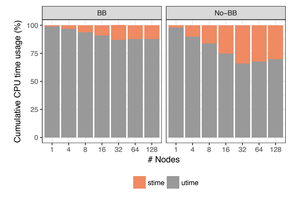
- I/O write performance was improved by a factor of 9, and I/O read performance by a factor of 15
- Performance decreased slightly at node counts above 64 (potential I/O ceiling)
- I/O performance must be balanced with parallel efficiency when using burst buffers with highly parallel applications
- I/O contention may limit the broad applicability of burst buffers for all workflow applications (e.g., in situ processing)



I/O performance estimate for read operations at the MPI-IO layer (solid lines), and average runtime speedup (dashed lines)



MPI-IO module data: Average time consumed in I/O read operations per process



Ratio between the cumulative time spent in the user (utime) and kernel (stime) spaces for different numbers of nodes

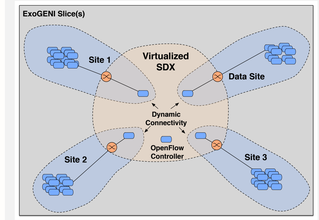
NETWORK PROVISIONING
Data Flow Prioritization for Scientific Workflows Using a Virtual SDX

We developed **mechanisms to arbitrate and prioritize data flows from competing workflows** by leveraging advanced network provisioning technologies like a **virtual Software Defined Exchange (SDX)**.

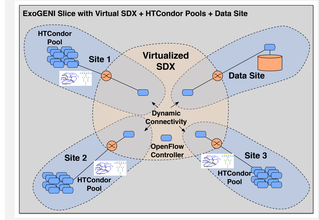
- Software Defined Exchanges (SDX) – meeting point of networks to exchange traffic, securely and with QoS, using SDN protocols
- Virtual SDX – virtual overlay acting as SDX without persistent physical location
- ExoGENI virtual SDX can modify compute, network, storage to support changing demands of SDX

Prioritized Data Flows

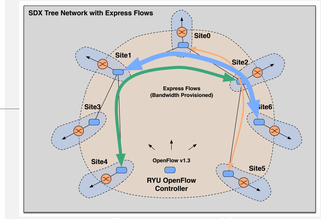
- Virtual SDX transparently arbitrates workflow data flows communicated by Pegasus



SDX: meeting point networks to exchange traffic, securely and with QoS, using SDN protocols



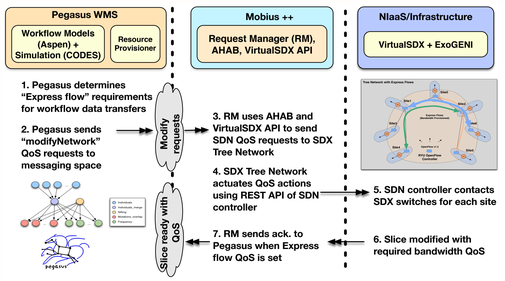
vSDX use case with HTCondor pools and Pegasus



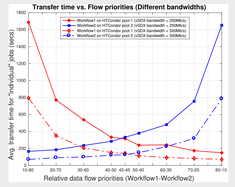
Tree network with vSDX

Flow Prioritization Use Case

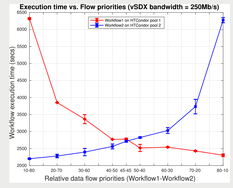
The Mobius++ framework can be used by several high-level applications to provision and adapt infrastructure based on particular requirements



Experimental Results



Effect of relative flow priorities on observed data transfer times for data-intensive workflow tasks for different vSDX provisioned bandwidths



Effect of relative flow priorities on overall workflow execution times

LEARN MORE

Panorama Website
http://sites.google.com/site/panoramaofworkflows/

Panorama is funded by the US Department of Energy under Grant #DE-SC0012636