Accelerating Time to Insight in the Exascale Ecosystem Through the Optimization of Scientific Workflows

James Ahrens, ahrens@lanl.gov, Los Alamos National Laboratory, A White Paper for the 3rd Big Data Exascale Computing Workshop, January 2015 - LA-UR-15-20354

The purpose of this white paper is to highlight software and hardware trends that will impact workflows at exascale and to describe a pathforward that harness these changes for workflow acceleration. By applying trends of automation, functional abstraction, the availability of compute power "everywhere" and accurate cost modeling to our scientific workflows, scientific productivity can be greatly improved. Our goal is to identify, automate and accelerate the manual operations that pervade our current workflows.

The first trend is the automation of all computational work in a workflow, if possible. An important realization is that manual interaction is costly in terms of performance. Applying Amdahl's law to the entire scientific workflow, the completion time, including the parallel simulation, is ultimately limited by the workflow's sequential interactive user time. Therefore all manual tasks that can be automated in a workflow should be. In a traditional workflow, visualization/analysis, debugging, data movement and scheduling require manual intervention. For visualization and debugging, the HPC community has focused on interactive approaches for the past few decades. The recent change in focus to automated *in situ* approaches is welcome. This change needs to be deeper than simply running existing interactive analysis operations in a batch-oriented manner. Specifically, batch oriented analysis operations need to focus on automatically identifying and tracking areas of interest, and then on the selecting and presenting these areas to the scientist. Furthermore, data movement and scheduling requires manual intervention due to the perceived need for the scientist to manage their storage and computing allocations. An alternative approach supports automatic directives. With these directives, scientists can state the data movement and scheduling policies they want followed without manual intervention. Due to the real time nature of business and financial computing, an automated approach to data movement and scheduling has worked well for these communities.

Efficient and effective workflow automation requires the use of the core computer science trend of functional abstraction and encapsulation. Specifically, we define a functional approach as one that defines tasks with well-defined inputs and outputs with no side effects. We need functional encapsulation at a variety computational scales: from tasks defined in a program, to the programs in a workflow, to workflows in a meta workflows such as those found in ensemble calculations and verification and validation projects. Functional abstractions enable a variety of desirable properties for workflows:

- Parallelism – Due to clear dependencies and independent units of work, parallel workflow execution at all scales is enabled.
- Resilience – Functional abstractions are without side effects and can easily be made transactional since they provide clear knowledge of when they are complete. When a workflow function at any scale fails due to software or hardware failures, these functions can be identified and rerun.
- Reproducibility – Science requires reproducibility for validation. Functional abstractions provide workflow reproducibility through a lack of side effects.

A challenge to a seamless pathforward is a lack of functional encapsulation mechanisms at different computational scales.

A third trend is the opportunity to accelerate our workflow on the multitude of processors in the HPC ecosystem. In the exascale time frame, we expect that memory, burst buffers, storage and network resources will all have processors associated with them. We envision accelerating the traditional sequential supercomputing end-to-end workflow by simultaneously executing multiple tasks at all computational scales. For example, we envision network and storage processors working in concert to prefetch data for simulation setup, floating point processors using this setup information to calculate

ensembles of simulation results, and memory-associated processors concurrently scanning these results for correlations.

The fourth trend is metered service, typically instantiated as a cost model. A cost model associated with the processing, memory, storage, and networking resources in the supercomputing ecosystem provides guidance to the scientist about the value/importance of these resources. Using this cost model, a sophisticated scheduler can automatically concurrently schedule components to optimize a workflow.

The following examples highlight the benefits of optimized workflows. Using a traditional post processing analysis workflow, simulation timesteps are written to storage for later manual interactive analysis. The cost of this traditional workflow includes the time to save full simulation timesteps for later analysis and the time/cost of the supercomputing resources needed to interactively visualize and analyze these massive timesteps. In contrast, our optimized *in situ* workflow identifies all the visualization functions and parameters that the scientist is interested in applying (Trend 2 – Functional abstraction). As the simulation is run, the Cartesian product of camera positions, functions and parameters are automatically calculated (Trend 1 - Automation). This process generates an image database that is on the order of approximately 10^6 times number of images in size. This is a significant data reduction from saving full extreme scale datasets that are 10^15 to 10^18 in size. The cost of this optimized workflow is significantly less than the traditional one. Although the optimized workflow includes the additional time to compute the visualization functions it significantly reduces storage time by only saving images for later analysis. Also no supercomputing resources are needed to interactively visualize and analyze the images. In the optimized workflow, post-processing visualization is support through exploration of the image database. Displaying images with sequenced camera positions supports interactive visualization. We can envision how to further optimize the workflow by assigning processors in an exascale storage system to concurrently compute indices that support image-based and visualization-object-based search interfaces (Trend 3 – Processors everywhere, Trend 4 – Cost Model) [1].

The next example focuses on how this approach support reasoning about optimal computing/storage data representations. The MPAS ocean simulation can save yearly, monthly, daily, hourly and every minute results. To save yearly results requires on the order of GBs whereas storing every minute requires on the order of hundreds of TBs. In the traditional workflow, the scientist makes a decision about what frequency of timesteps to save. Factors that influence their decision include the likelihood of additional analysis needed at different temporal frequencies, the availability of computing resources to regenerate timesteps and the cost, bandwidth, and accessibility of storage. In an optimized workflow, we envision a cloud-like storage system implementing automatic analysis result regeneration. We envision the system including a set of policies about the length of time a scientist is willing to wait to regenerate an analysis for the benefit of reduced storage (Trend 3 – Processors everywhere). Ultimately, we envision the system clearly elucidating the tradeoffs (i.e. how they fit in the scientist's computational resource budget), and offering suggestions on how to optimize their scientific workflow (Trend 4 – cost model). In initial experiments, we found with a cost model for compute and storage similar to Amazon Web Services it is cost effective and efficient to store daily averages. If further temporal details are needed, it can take less than a minute with an allocation of thousand cores to recalculate the intermediate time steps, automatically apply the analysis operators, and then save only the significantly reduced analysis results (Trend 1 – Automation). These examples provide a glimpse of the possibilities that result from pursuing a workflow optimization strategy for the extreme scale ecosystem.

[1] J. Ahrens, S. Jourdain, P. O'Leary, J. Patchett, D. H. Rogers, M. Petersen, "An Image-based Approach to Extreme Scale In Situ Visualization and Analysis", Supercomputing 2014, New Orleans.