Extreme Scripting and Other Adventures in Data-Intensive Computing

Ian Foster, Allan Espinosa, Ioan Raicu, Mike Wilde, and Zhao Zhang
(Argonne National Labs)

Abstract

Data analysis in many scientific laboratories is performed via a mix of standalone analysis programs, often written in languages such as Matlab or R, and shell scripts, used to coordinate multiple invocations of these programs. These programs and scripts all run against a shared file system that is used to store both experimental data and computational results.

While superficially messy, the flexibility and simplicity of this approach makes it highly popular and surprisingly effective. However, continued exponential growth in data volumes is leading to a crisis of sorts in many laboratories. Workstations and file servers, even local clusters and storage arrays, are no longer adequate. Users also struggle with the logistical challenges of managing growing numbers of files and computational tasks. In other words, they face the need to engage in data-intensive computing.

We describe the Swift project, an approach to this problem that seeks not to replace the scripting approach but to scale it, from the desktop to larger clusters and ultimately to supercomputers. Motivated by applications in the physical, biological, and social sciences, we have developed methods that allow for the specification of parallel scripts that operate on large amounts of data, and the efficient and reliable execution of those scripts on different computing systems. A particular focus of this work is on methods for implementing, in an efficient and scalable manner, the Posix file system semantics that underpin scripting applications. These methods have allowed us to run applications unchanged on workstations, clusters, infrastructure as a service ("cloud") systems, and supercomputers, and to scale applications from a single workstation to a 160,000-core supercomputer.

Swift is one of a variety of projects in the Computation Institute that seek individually and collectively to develop and apply software architectures and methods for data-intensive computing. Our investigations seek to treat data management and analysis as an end-to-end problem. Because interesting data often has its origins in multiple organizations, a full treatment must encompass not only data analysis but also issues of data discovery, access, and integration. Depending on context, data-intensive applications may have to compute on data at its source, move data to computing, operate on streaming data, or adopt some hybrid of these and other approaches.

Thus, our projects span a wide range, from software technologies (e.g., Swift, the Nimbus infrastructure as a service system, the GridFTP and DataKoa data movement and management systems, the Globus tools for service oriented science, the PVFS parallel file system) to application-oriented projects (e.g., text analysis in the biological sciences, metagenomic analysis, image analysis in neuroscience, information integration for health care applications, management of experimental data from X-ray sources, diffusion tensor imaging for computer aided diagnosis), and the creation and operation of national-scale infrastructures, including the Earth System Grid (ESG), cancer Biomedical Informatics Grid (caBIG), Biomedical Informatics Research Network (BIRN), TeraGrid, and Open Science Grid (OSG).

For more information, please see www.ci.uchicago/swift.