THE NEW CHALLENGES FOR SCIENTIFIC RESEARCH WORKFLOWS

Michael Feldman

White paper

November 2014

MARKET DYNAMICS

Computation has become the third leg of science, and this is nowhere more apparent than in life science. Its use in bioscience research and development is spread across the pharmaceutical industry, medical research, agriculture, and environmental engineering. Over the last two decades especially, genomics and other bioinformatics technologies have advanced tremendously. These have matured to the point of commercial viability, driving applications in drug discovery, medical treatments, and agricultural production. Spending in computational hardware and software for commercial bioscience applications is expected to top $730 million in 2014[1]. At the same time, university-based life science research using advanced computational methods continues to grow, thanks to a steady increase in bioinformatics research and the declining cost of high performance computing.

The Workflow Challenge

The improvement in gene sequencing technology has created a flood of data that requires analysis by high performance computing (HPC) systems. Scientists at hospitals, universities, pharmaceutical firms, and other life science-based organizations are employing these technologies to identify hereditary diseases, develop more refined cancer treatments based on cell genetics, and advance plant and animal breeding programs. Likewise, drug companies and non-profit research organizations are using molecular modeling applications to identify new drug candidates for a wide range of diseases and medical conditions. In addition, agricultural firms are using advanced analytics to optimized crop production and harvesting.

These applications rely on compute-intensive and data-intensive software running on HPC systems to provide the needed performance. For both commercial and non-commercial environments, application throughput is a critical factor. Although the price-performance profile of HPC continues, it still represents a significant investment for organizations. As a result, workload management is a critical component of these research pipelines.

[1] High Performance Computing Forecast for 2011 through 2015: Economic Sectors and Vertical Markets, Intersect360 Research, June 2011

Furthermore, as researchers look for additional compute capacity and flexibility, more of their application workloads are finding a home on remote systems and cloud-hosted clusters. Although still a small part of the total HPC market (about 1 percent), the use of compute clouds, both private and public, is growing rapidly (18.6 percent CAGR)[2]. And because many bioscience applications are highly parallel and loosely coupled, they are especially suitable for these environments. Cloud bursting, in particular, can offer cost-effective flexibility for research efforts that have widely varying needs for computation over extended periods of time.

OPPORTUNITY FOR ADAPTIVE COMPUTING

Adaptive Computing, with its focus of intelligent workload management, has offered advanced resource management tools to HPC users for more than 10 years. According to our latest HPC Site Census Survey, the company is the top supplier of job management solutions, with 44 percent of sites citing Adaptive as a supplier[3]. Likewise, Moab was the most oft-mentioned middleware package by those same respondents.

The company's Moab-branded products, which can operate across traditional clusters, private clouds, and public clouds, have been designed to provide automated job scheduling and resource management for many types of workloads, including the growing array of research workloads found in hospitals, drug companies, and universities. Moab is especially adept at maximizing throughput in environments where compute resources limit the amount of computation that can be accomplished. Further, Moab technology offers a great deal of flexibility and control in where and how workloads are executed.

Moab Brings Big Workflow to Workload Management

Adaptive has continually refined its product set, most recently to incorporate the notion of "Big Workflow," a term coined by the company that provides a model for integrating data-intensive and compute-intensive workflows under a single workload management scheme[4]. Essentially, the model draws together HPC simulations with big data analytics workloads so that customers can manage them in a unified manner.

---

[2] Worldwide High Performance Computing, 2013 Total Market Model and 2014–18 Forecast, Intersect360 Research, June 2014
[3] HPC User Site Census: Middleware, Intersect360 Research, April 2014
[4] Big Workflow: More than Just Intelligent Workload Management for Big Data, Intersect360 Research, February 2014

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

In the latest rendition of the Moab HPC Suite (Moab 8.1) and the Moab Cloud Suite, Big Workflow services deliver dynamic scheduling, provisioning and management of multiple applications across HPC, traditional cloud and big data environments. By doing so, Moab is able to automate much of the workflow management, which otherwise would need to be performed manually with scripts and administrator intervention. For researchers, this increases application throughput and streamlines the complexities of job management, which, in turn, shortens the time to discovery.

As a consequence, universities, drug companies, and other research organizations can maximize their investments in HPC spending. Hospitals, for example, tend to invest in a central HPC cluster, to run workloads submitted by different departments working on different aspects of a disease – whole genome analysis, protein analysis, cancer DNA analysis, etc. HPC budgets are shared by the departments, but it's in everyone's interest that the central resource be used to maximum capacity, whether or not an individual department is using its proportional share at any given moment.

Moab is able to maximize cluster utilization in these situations by using the concept of compute "condos." Condos represent portions of a cluster owned by a department that are put into a global resource pool and borrowed by other departments when they are not being used. A job running in a condo may be pre-empted according to priority policies set up by the hospital so that individual departments are guaranteed access to their hardware when needed.

In commercial businesses, a more siloed approach to resource sharing is common, where individual groups buy and operate their own HPC clusters. For example, at any given time, a pharmaceutical company could be developing several drugs, which often have their own computational resources committed to each of them. Although less inherently flexible than a central resource approach, Moab can still maximize usage across multiple clusters if they are connected to each other and have the requisite policies in place that enable inter-cluster resource sharing.

Cloud bursting is another approach that, as mentioned previously, offers flexibility for situations where the need for research computation can fluctuate over time. University research, in particular, operates in a highly changeable environment, where, for instance, in the US, seasonal NSF funding and research work lead to irregular demands for computation. This is encouraging some research groups to invest in small clusters (say, 8 to 16 nodes) and devote the remaining system budget to cloud cycles. Moab, with its support of OpenStack, a popular open source operating system for managing public and private clouds, is able to leverage cloud infrastructure in such environments.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

Bursting also allows hospitals to share internal private clouds between different domains. For example, a compute cloud that provides general hospital administration may be underutilized at times, and thus available for, say, genome analysis work. Privacy and data integrity are protected as Moab handles the shredding of the OpenStack instances before resources transition from one application to another. Although, the underlying infrastructure might not be ideal for such performance-demanding applications, the advantages of greater overall throughput outweigh any loss in compute efficiency.

Underlying cloud support and resource management capabilities in general, is Moab's ability to provide data staging. Data staging manages file transfers in such a way as to maximize system utilization and minimize resource inefficiency. Essentially, a job's required data is sent to the location where the execution will occur, so that it is available when the job is scheduled to begin.

For research workloads with large data sets, such as genomic data, such a capability can be crucial in optimizing performance within a cluster, between multiple clusters, or out to a cloud. Genomic data, in particular, is often shared across multiple applications, so intelligent workflow configuration can minimize data transfer overhead significantly.

INTERSECT360 RESEARCH ANALYSIS

Computational research, especially in the life science arena, is undergoing fundamental change. The ever-declining cost of gene sequencing and HPC hardware is driving new advancements in bioinformatics, from personal genomics, and cancer treatments, to drug discovery and proteomics. As a result the data loads and workflow complexities are increasing, which means workload managers are critical assets to these research efforts.

In additions, because R&D is not a profit center, it tends to operate under strict budgetary constraints, which means there is a great deal of pressure to get the most out of capital and operating expenditures. Again, workload managers are often the key here, to maximize the utilization of the computational hardware, and in the case of the cloud computing option, also offering additional flexibility and resource efficiencies.

With its Moab HPC Suite and Moab Cloud Suite, Adaptive Computing offers an advanced solution for research application workloads across many different types of research environments. Hospitals, pharmaceutical companies, universities, and other research organizations encompass a range of computational and data workflows, as well as infrastructure preferences. In each case, Moab's offers an array of features that offers flexible, automated workload management for these organizations.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com