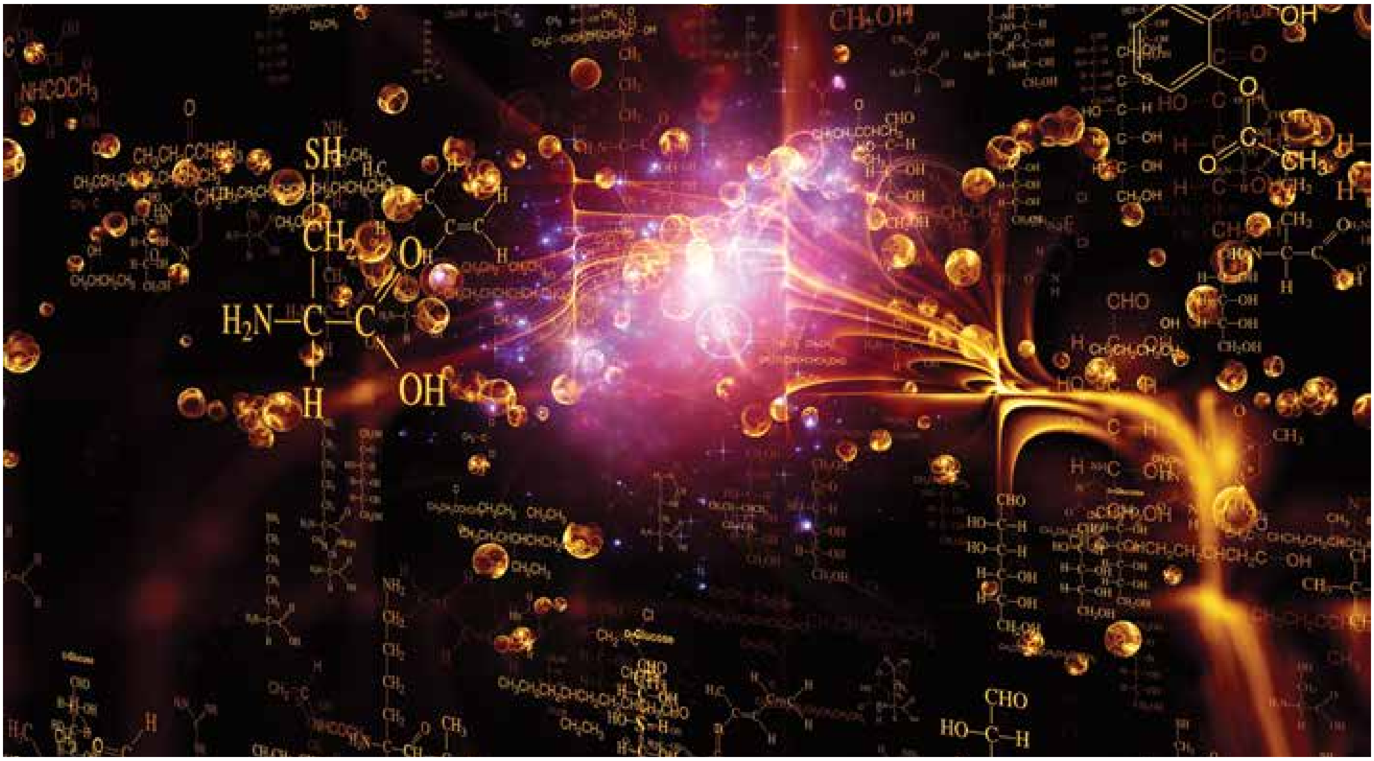
IBMSystems MAGAZINE



Speed, Scale, SMARTS

IBM reference architecture for genomics brings power to research

Much has been made of IBM Watson* technology when it comes to healthcare—and rightfully so—but Watson isn't IBM's only contribution to healthcare and life sciences industries. IBM Systems and Technology Group has developed a software-defined, data-centric and application-ready reference architecture, which has been further optimized on IBM POWER* technology and Elastic Storage platforms to bring spee

Frank Lee, Ph.D., is the lead architect of Genomic Medicine at IBM.

Genomic medicine promises to revolutionize medical research and clinical care. By investigating the human genome in the context of biological pathways and environmental factors, it's now possible for genomic scientists and clinicians to identify individuals at risk of disease, provide early diagnoses based on biomarkers and recommend effective treatments.

The success of new technology and research methods came with a big cost—the field of genomics is caught in a flood of data as huge amounts of information are churned out from next-generation sequencers. That data must then be stored, analyzed, shared and archived. Many genomics, cancer and pharmaceutical research institutions are generating so much data that they can no longer be processed in a timely manner or even transmitted over regular communication lines. Often they're shipping raw data to external computing centers for processing and storage, creating an obstacle for speedy access and analysis of data. In addition to scale and speed, it's also important for information to be linked based on a data model or taxonomy, or curated with machine or human knowledge. This smart data can then be factored into the equation when dealing with genomic, phenotypic and environmental data on a common analytical platform.

To address the needs of scale, speed and smarts for genomic medicine, IBM created an end-

Reposted with permission from *IBM Systems Magazine, Power Systems edition*

to-end reference architecture that defines the enterprise capability of data management, workflow orchestration and global access across key platforms for genomics, translational and personalized medicine. Based on this architecture, IBM has successfully built data-centric, software-defined and application-ready platforms in support of large-scale genomics sequencing and downstream data analytics. Close to a dozen of the world's leading healthcare and life science organizations have adopted IBM reference architecture for their integrated research computing infrastructure.

Data Management With Datahub

Data management is critical for genomics because of the volume, velocity and variety of data. Genomic data volume is surging as the cost of sequencing drops precipitously. The I/O throughput on a genomic system can be extremely demanding due to data volume and the large number of file and directory objects. Many data formats also exist, with varying degrees of lifecycle management requirements ranging from transient files in scratch space to variant calling files that must perpetually remain online.

Leveraging IBM Elastic Storage, the Datahub data-management layer defines an enterprise capability to meet these challenges based on its high-performance, scalable and extensible architecture.

First developed and optimized as a high-performance computing (HPC) file system, Elastic Storage manages to serve large volumes at a high bandwidth and in parallel to all of the compute nodes in the computing system(s). As genomic pipelines can consist of hundreds of applications engaged in concurrent data processing on large numbers of files, this capability is critical in feeding data to the computational genomics workflow.

A system pool built on SSD and flash disk with high-IOPS capability can be dedicated to store metadata for files/directories and, in some cases, small files directly. This drastically improves file-system performance and responsiveness to metadata-heavy operations such as listing all files in any given directory.

As a file system with a connector to MapReduce, Datahub can also serve MapReduce/big data jobs on the same set as compute nodes, eliminating the need for and complexity of Hadoop Distributed File System.

The policy-based data lifecycle management capability lets Datahub move data from one storage pool to the others, maximizing I/O performance and storage utilization while minimizing operational cost. These storage pools can range from the high-I/O flash disk to high-capacity storage appliance (GPFS* Storage Server) to low-cost tape media (through integration with a tape management solution).

The increasingly distributed nature of genomics infrastructure also requires data management on a much larger and global scale. Not only must data be moved or shared across different sites, its movement or sharing must also be coordinated with computational workload and workflow. To achieve this, Datahub leverages a key function of Elastic Storage called Advanced File Management (AFM). It enables Elastic Storage to extend the global name space to multiple sites, letting them share a common metadata catalog and a cache copy of home data for a remote client site to access locally. For example, a genomic center can own, operate and version control all reference databases or data sets while the partnering sites can access the reference data set through AFM. When the centralized copy of the database is updated, so are the cache copies of the other sites.

With Datahub, a systemwide metadata engine can also be built to index and search all of the genomic and clinical data, enabling powerful downstream analytics and translational research.

Infrastructure for genomic medicine must be:

Data-centric
meeting the challenge of explosive growth of genomics and clinical data with data management capabilities

Software-defined
defining the architecture based on software-based abstraction layers for computation, storage, big data and cloud

Application-ready
integrating a multitude of applications with a workload and workflow orchestrator

Workflow Management With Orchestrator

The workflow for genomics is complex, yet important. The number of genomic applications is growing and represents varying degrees of maturity and types of programming models—many are single-threaded or embarrassingly parallel while a few others are multithreaded or message passing interface (MPI)-enabled (mpiBLAST), yet all applications must work in concert or tandem in a high-throughput and high-performance mode to generate final results.

Through Orchestrator, the reference architect defines the capability to orchestrate applications and workflow. A unique combination of workload manager (Platform* LSF*) and workflow engine (Platform Process Manager) links and coordinates a spectrum of computational and analytical jobs into fully automated pipelines that can be easily built, customized, shared and run on a common platform. This provides the necessary abstraction of applications from the underlying infrastructure, such as an HPC cluster with graphical unit processor or a big data cluster in the cloud.

Orchestrator distinguishes itself from scripted workflow tools with its capability to handle complex workflow dynamically—individual workloads or jobs can be defined through a user-friendly interface, incorporating variables, parameters and data definitions into a standard template. The workload manager will transparently handle the submission, placement, monitoring and completion of each job. Workflow engine connects jobs in linear progression, conditional branches or loops based on user-defined criteria and requirements for completion and advancement.

To maximize the throughput of the workflow for genomics sequencing analysis, a special type of workload can be defined by using job arrays so data can be split and processed by many jobs in parallel.

In another innovative use case for genomics processing, multiple subflows can be defined as a parallel pipeline for variant analysis following the alignment of the genome. The results from each subflow are then merged into a single output and provide the analyst with a comparative view of multiple tools or settings.

The workflow can also be designed as a module and embedded into larger workflows as a dynamic building block. Not only will this approach enable efficient building and reuse of the pipelines, it will also encourage collaborative sharing of the genomic pipelines among a group of users or within larger scientific communities.

As more institutions are deploying hybrid cloud solutions with distributed resources, Orchestrator can coordinate the distribution of workloads based on data localities, predefined policies, thresholds and real-time input of resource availabilities. For example, a workflow can be designed for processing genomic raw data closer to sequencers, and followed by sequence alignment and assembly using MapReduce framework on a remote big data cluster. In another use case, a workflow can be designed to launch a proxy event moving data from a satellite system to a central HPC cluster when the genomic processing reaches 50 percent of the completion rate. The computation and data movement can happen concurrently to save time and cost.

Manage Global Access With AppCenter

In the reference architecture, IBM also defines a capability

IBM Solution for Genomic Workflow Acceleration—Power Systems Edition

In deep collaboration with IBM clients and across multiple disciplines, IBM Solution for Genomic Workflow Acceleration—Power Systems* Edition was designed by bioinformatics experts and technologists to accelerate the end-to-end analysis, storage and sharing of genomics data and hardened with widely used bioinformatics applications and pipelines. Designed for ease-of-use, this high-throughput platform delivers fast turnaround for a variety of workflows including genomic medicine, agrigenomics, biofuels, forensics and drug development.

The task of discovering subtle patterns and variations in molecular biology, in many fields of study ending in -omics, such as genomics or proteomics, represents a grand compute and storage challenge. This reference platform tackles these challenges by seamlessly integrating Next Generation Sequencing systems with a scalable and high-performance compute infrastructure. Upon completion of sequencing, raw data will be written directly into high-performance Elastic Storage (Datahub) for ready access by hybrid compute infrastructure. The computational and analytic process then can be designed into simplified and automated workflows, capturing and protecting repeatable best practices (Orchestrator). Workflow steps and dependencies are managed within an intuitive GUI, providing a fast-path to automating lengthy, repetitive tasks highly prone to human error. These adaptable workflows become self-documenting, so the process updates can be made and tracked easily. This platform also features a library of sample workflows based on best practices to accelerate deployment, fully exploit available compute and storage resources, and reduce time to discovery. These sample and other user-defined workflows can be made available for collaboration through an enterprise portal (AppCenter).

This integration of high-performance storage, workflow orchestration and portal access is available today as part of the offering with guided steps for the implementation and deployment. This predefined platform was designed to manage hybrid environments, including Power* and x86-based workloads, and can also be easily adapted to support large scale and cloud based deployments.

For more information, visit www.ibm.com/systems/power/solutions/industry/healthcare.html called AppCenter as the user interface into the genomics platforms. The goal is to provide an enterprise portal with role-based access and security control, yet allow researchers and clinicians easy access to data and workflow tools.

Built with Platform Application Center, AppCenter also has advanced logging capabilities for tracking activities including jobs, workflow provenance and data access. This is a critical feature in the event of reporting, analyzing performance or rerunning the analysis with prior settings.

To harvest knowledge and information from users and enable sharing, AppCenter functions as a catalog of pre-built and pre-tested workflow definitions and application templates so users can easily launch them directly from the portal after uploading the data. The portal can also be configured with a metadata index and search engine, enabling users to browse, query or search pre-indexed data or reference documents.

Extensible Reference Architecture

IBM reference architecture extends beyond genomics into translational and personalized medicine platforms. These platforms also leverage the Datahub, Orchestrator and AppCenter as common enterprise capabilities so that data, workloads and access points can be converged or linked together into an enterprise architecture for genomic medicine.

Check out the iPad and Android apps for additional content

IBM reference architecture can be implemented on a range of computational resources such as the recently announced POWER8* system. The integrated solution, IBM Solution for Genomic Workflow Acceleration—Power Systems Edition, bundled together Datahub, Orchestrator and AppCenter as well as workflow templates preloaded into the AppCenter as working example. The system is easy to install and use, providing a powerful platform for genomic data analysis and management.