WHITE PAPER

L7INFORMATICS

the L7Informatics high-performance genomic cloud powered by IBM:

Going Beyond Commodity Cloud Services for Data-Intensive Genomic Applications

As large genomic labs become even more data intensive and collaborative efforts become more geographically diverse, utilizing the cloud for analysis and storage of data is becoming advantageous over using on-premises compute resources. Most organizations who opt out of on-premises computing resources are using solutions based on cloud services from one of the big three providers – Amazon, Microsoft, or Google. With the size of genomic datasets, the long-term costs associated with commodity cloud solutions are simply not sustainable. While they are technically capable of performing the requisite tasks, their infrastructures are not optimized to efficiently handle genomic-scale workloads. Essentially, you can put together an intricate data management solution on top of Amazon, but ultimately you are still paying the regular commodity cloud cost structure for data uploads and downloads, processing, and storage, and it's just not practical.

To address these issues and develop a solution suitable for genomic-scale data, L7 Informatics took lessons learned from our years of experience building high performance computing systems for genomics and partnered with IBM to apply that experience to the cloud. This whitepaper explores how our solution uses the IBM® Bluemix® infrastructure cloud to provision optimized resources for storage, computing, and networking components that are designed from the start to efficiently manage and handle genomic-scale data.

COMMODITY CLOUD-BASED SOLUTIONS: NOT IDEAL FOR LARGE-SCALE GENOMIC RESEARCH

Given the availability of existing commodity cloud infrastructures, it was an easy path for companies focused on genomic data, such as Core Informatics, Illumina/Genologics, Seven Bridges, and DNAnexus, to use these services as the backbone of their systems. While these solutions are technically capable of processing genomic-scale data, and have a lot of tools available to enable these tasks, they were designed for less data-intensive, mass-market consumer tasks such as file and photograph storage, video sharing/streaming, database hosting, and web application deployments. Furthermore, the underlying cost structure of Amazon and other similar clouds are still in place regardless of what analytical tools companies and research centers put on top of those platforms, which results in high costs for data transfer, computing, and most importantly, storage.

ENTERPRISE-AS-A-SERVICE: A SOLUTION BUILT FOR LARGE-SCALE GENOMIC RESEARCH

The L7 High-Performance Genomic Cloud Powered by IBM is a complete "Enterprise-as-a-Service" solution that extends beyond traditional cloud and SaaS systems to include full support for all aspects of laboratory informatics. The complete software stack provides analysis, a laboratory information management system (LIMS), and infrastructure capabilities by default and can be easily extended by users to include third-party software and hardware solutions as needed. Highlights of the Genomic Cloud infrastructure include the following:

Robust purpose-built hardware

- Tightly coupled IBM Spectrum ScaleTM (GPFS) storage, high-memory compute nodes, and a fast internal network designed for high-throughput genomics
- Support for up to 500 compute nodes and more than 10 Petabytes of storage
- The ability to configure it to support remote, on premises, or hybrid cloud models in isolated single-tenant or multi-tenant instances
- Designed from the ground up for scientific computing

Secure

- Virtual network infrastructure with a dedicated VPN, VLAN, and single sign-on support
- Secure access with no additional setup or costs

Comprehensive/complete software stack that streamlines deployment and operations

- A complete supported software stack, including BioBuilds, L7 Enterprise Science Platform™ (ESP), and IBM Spectrum LSF® scheduler, is ready to run, with no additional setup costs
- Everything can be in place and operational within two weeks, while equivalent commodity cloud functionality could take up to six months to fully deploy comparable wfunctionality

L7 ESP automates analysis and reporting

- ESP orchestrates all data movement and analysis in the Genomic Cloud
- Sample and data tracking, protocol and analysis pipeline management, and powerful reporting tools offer a complete enterprise-scale data management platform
- By automating common tasks using pipelines, Docker containers, and other execution engines, you can free-up valuable bioinformatics and IT resources, leaving more time for science

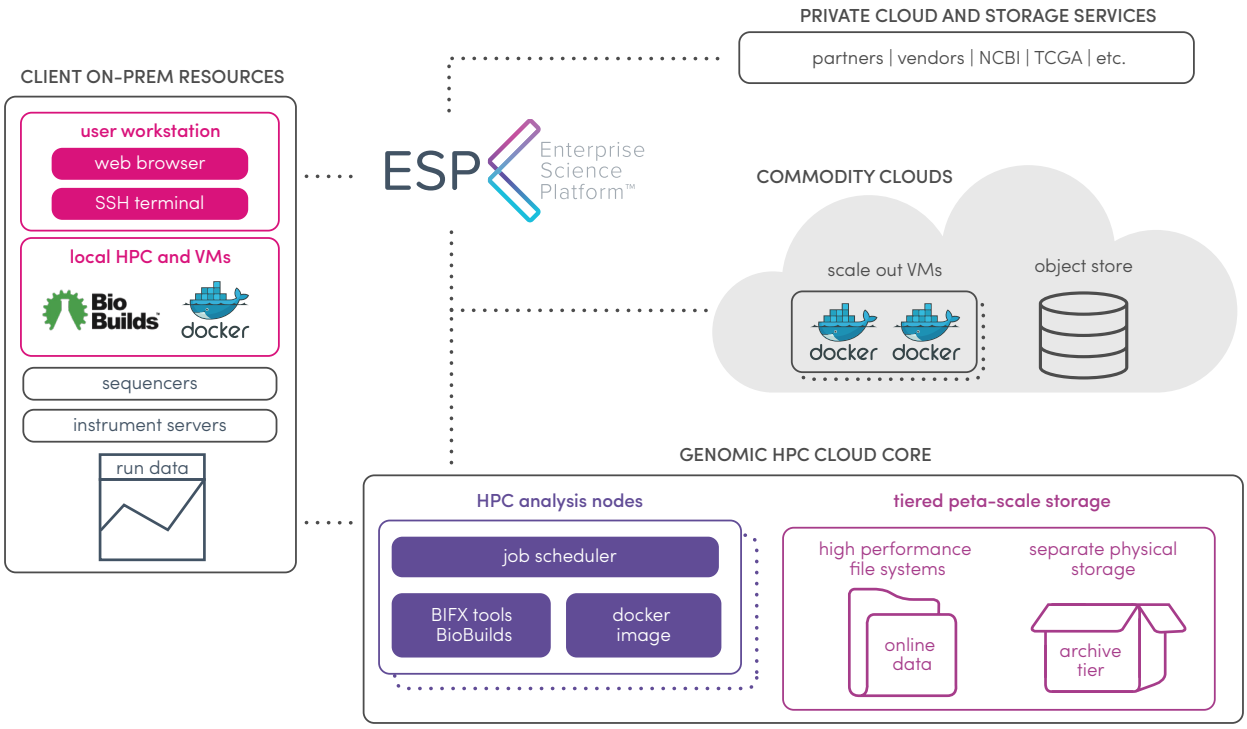

Figure 1. Genomic Cloud Architecture: Based on a readily-deployable HPC Cloud Core supporting up to 500 compute nodes and over 10 Petabytes of storage, the Genomic Cloud can be configured to support remote, on-premises, or hybrid cloud models in isolated single-tenant or multi-tenant instances. L7's and IBM's software stacks provide a complete set genomic and HPC tools.

DEPLOYMENT STRATEGIES: LEVERAGING IBM BLUEMIX INFRASTRUCTURE'S GLOBAL FOOTPRINT

The global reach of IBM Bluemix data centers means that the Genomic Cloud can be configured practically anywhere. By controlling all aspects of the hardware, L7 and IBM can support a wide range of deployment scenarios from standard multi-tenant cloud instances, to fully isolated instances that meet strict regulatory guidelines, to hybrid on premise/cloud deployments. Hybrid deployments leverage IBM Spectrum Scale storage to provide automated tiering and a single namespace filesystem, significantly simplifying file management across locations. By utilizing HPC standards, L7 Systems and IBM can provide full support for users that extends into bioinformatics and optimization services for custom computational workloads.

THE BENEFITS OF USING THE LAB7 AND IBM GENOMIC CLOUD ARCHITECTURE

When putting together this architecture, our goal was to provide the most efficient cloud resource for data processing that is set up using the mold for dedicated on-premises computing resources. Thus, we created a straightforward cost structure where clients are just billed for the infrastructure and standard support, which is designed to minimize these costs.

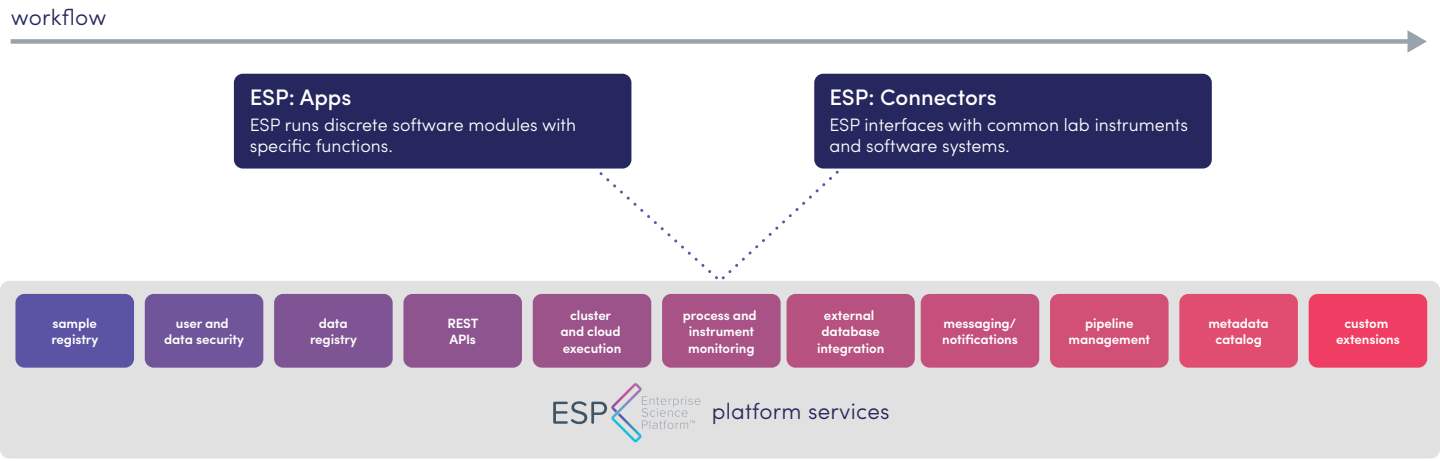

Figure 2. Enterprise Science Platfrom is an Operating System for the Lab: ESP manages data through all aspects of data intensive genomic experimental workflows. The core of the platform, the L7 Resource Manager, maintains provenance of all data that is registered with the system throughout the entirety of the laboratory workflow. Application layers, such as the LIMS and Analysis Pipeline Manager, and tools such as the Reports Engine address discrete segments of workflow. The ESP is cabable of interacting with major lab hardware and external software tools, whitle the platform is customizable to user requirements and HPC environments, including the L7 High-Performance Genomic Cloud.

Additionally, the infrastructure is optimized for genomic workflows, which makes the costs much more predictable, and the solution can scale from a single client to multiple clients.

We take Enterprise-as-a-Service seriously. Most cloud vendors sell you virtual resources and leave you to figure out what to do. Our cloud is different. Our support and dev-ops teams are always available to help with software, hardware, and even bioinformatics questions. We also have a global reach. Our extensive network of data centers allows us to place infrastructure all over the world. Complex combinations of on-premises systems and public/private cloud instances are within reach to meet your deployment requirements.

As a result of this comprehensive re-thinking of how scientific computing should occur in the cloud, we've demonstrated a reduction in the overall turn-around time of data analysis. In a recent case study, 50 paired-end whole human genomes at 10x coverage collected in the 1,000 Genomes Project were analyzed using a standard BWA/GATK Haplotyping pipeline on a commodity cloud and on the Genomic Cloud. Using basic parallelization, the 50 samples we analyzed on the Genomic Cloud in 3.3 days instead of five days on the commodity cloud, a 33 percent time savings. Further optimization yielded a 6.25-hour turnaround time, which is a 95 percent improvement in performance.

WHITE PAPER

MORE EFFICIENTLY USE THE CLOUD FOR GENOMIC-SCALE DATA WITH L7 AND IBM

As we discussed, the cloud is rapidly becoming a popular choice for the management and analysis of genomic-scale data for many highly valid reasons. However, the current collection of available options only includes cloud infrastructures that were originally designed for other purposes, and as a result, do not include optimal networking and storage capabilities nor do they maintain cost structures that can be prohibitive to their own long-term sustainability.

To address the concerns associated with commodity clouds, we developed the L7 High Performance Genomic Cloud. Our solution ultimately saves money and improves the turnaround time for results, and is the only option developed specifically for genomic-scale data and workloads, making it the clear choice for cloud-based genomic-scale data applications.

Ready for
IBM Cloud

To learn more about how L7 Systems and IBM are collaborating to optimize cloud performance for genomic-scale workflows, visit the L7 High Performance Genomic Cloud page at www.L7Informatics.com or contact L7 Systems to discuss how the Genomic Cloud can help you

L7INFORMATICS

1219 West 6th Street | Austin, TX 78703
888.461.5227 | L7informatics.com | info@L7informatics.com