# Lab7 for software defined infrastructure

*Integrated data and workflow management for affordable next generation sequencing (NGS)*

## Highlights

- Faster next generation sequencing (NGS) allows companies to bring drugs to the market earlier, saving them millions of dollars.

- NGS fundamentally transforms the discovery and development of drugs and the delivery of therapies for preventive and personalized medicine.

- Provides an integrated, agile and versatile sample-to-answer software platform for NGS to analyze and interpret massive datasets.

- Delivers outstanding performance with a comprehensive, powerful, and scalable compute platform.

- Offers an affordable high-performance solution compared to like-sized x86 systems with better reliability and security.

Faster and affordable NGS is fundamentally transforming the healthcare and life sciences industries. By improving time-to-market for preventive and personalized medicine, companies can save millions of dollars in drug discovery and development while delivering innovative therapies. The cost of NGS is approaching the widely anticipated 'tipping point of $1,000 per genome,[1] making the technology more cost-effective and making the promise of personalized medicine a reality.

But in order to become more mainstream, NGS technology must democratize beyond its current practitioners. Traditionally, NGS laboratories have been parts of large institutions with well-established IT infrastructures and bioinformatics teams. This is because NGS technologies are compute and data intensive—a typical sequencing run can range from a day to a week and can generate a few gigabytes to several terabytes of data. These laboratories must also meet stringent regulatory requirements and need a reliable and secure infrastructure to provide consistently faster throughput for production environments. Today, data management, analysis, and reporting in NGS are spread across a variety of software tools and applications that each addresses specific parts of the workflow. While this offers flexibility in a purely research environment, integrating these tools to work in a production environment is difficult, especially if they use different programming languages, file types, and references. As a consequence, data flow is far from streamlined and the provenance of the data—critical to regulated environments—is difficult to maintain.
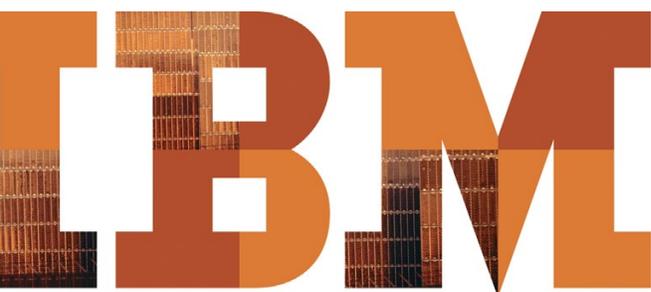
Smaller laboratories that do not have the advantage of a large IT infrastructure will therefore need an integrated hardware and software solution to take advantage of the emerging NGS opportunities.

## How Lab7 on IBM overcomes these challenges

The Lab7 Enterprise Sequencing Platform (ESP) is a complete sample-to-answer software platform for next generation sequencing. Lab7 ESP combines sample tracking and protocol management with powerful data analysis, reporting, annotation and visualization tools that previously had to be run independently. These functional components are built upon a robust workflow engine that links each component to ensure rigorous data provenance throughout the workflow process. NGS analysis is compute-intensive and requires large compute power and storage capacity to manage, analyze and visualize data. A compute architecture based on the IBM® Reference Architecture for Genomics using IBM Power Systems™, IBM Spectrum Computing and IBM Spectrum Storage™ software including IBM Spectrum LSF® and IBM Spectrum Scale™ (based on IBM GPFS™) delivers outstanding performance for such demanding requirements. The Lab7 ESP is also capable of being deployed across heterogeneous compute platforms, enabling a mix of options. For example, IBM Power Systems can use Lab7 ESP to launch jobs on x86 systems on the same network, and vice versa.

### Lab7 ESP

By combining laboratory information management system (LIMS) features with informatics features, such as an advanced production pipeline manager and reporting framework, Lab7 ESP makes it easy for NGS researchers to manage the entire sequencing workflow. The key features include:

- LIMS Lite: A tool for transparent, streamlined sample tracking and communication between lab technicians, managers, and scientists.
- Analysis Pipeline Manager: Pipeline creation and management tools to integrate seamlessly with the user's existing bioinformatics environment.

- Reporting and visualization: View reports from pipelines as they are executing, giving unprecedented access to data during the analysis phase.
- Data provenance and security: Lab7 ESP is designed to run in research, production and regulated environments. It tracks the full provenance of sequencing data from sample submission to the instrument to quality control and analytical results.

### IBM Reference Architecture for Genomics

IBM, in collaboration with key researchers and partners, created the IBM Reference Architecture for Genomics. It is an end-to-end reference architecture that defines the enterprise data management, workflow orchestration and global access capabilities across key genomics, translational and personalized medicine platforms. It supports large-scale genomics sequencing and downstream data analytics, providing: data lifecycle management to support large scale data growth; software-based abstraction layers for compute, storage, big data and cloud; and workload and workflow orchestrator for applications.

---

*"We've optimized the Lab7 ESP for IBM to provide a fully-comprehensive NGS solution that combines LIMS and informatics software capabilities with a robust compute platform that offers excellent performance, reliability, and security,"*

—Varshal K. Davé, VP Sales & Marketing, Lab7 Systems, Inc.

---
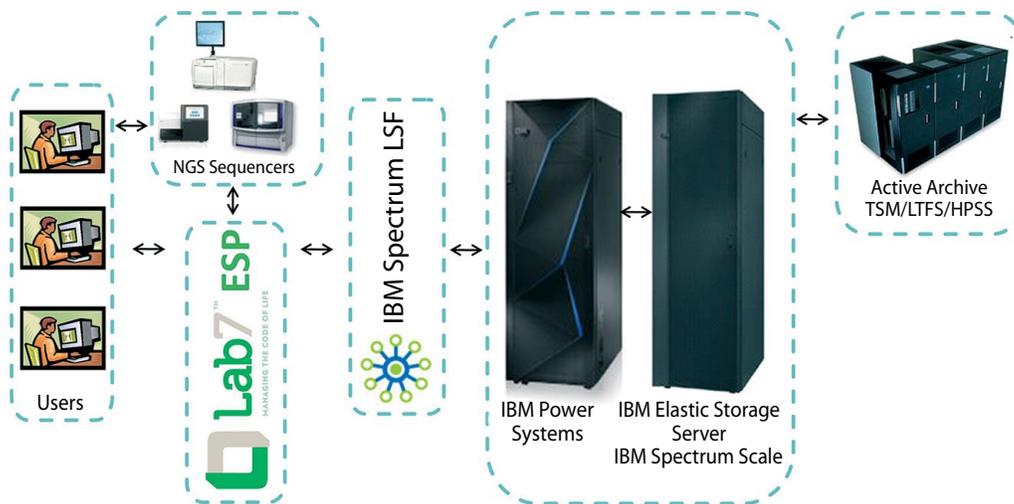
The integrated IBM-Lab7 solution architecture for NGS workflow and data management

## IBM Power Systems

IBM Power Systems deliver trusted, state of the art technology at the small, mid and enterprise computing levels and are broadly deployed in production environments worldwide. These systems compare favorably with x86 servers on cost while delivering greater performance, higher use, and superior availability. With so much consolidated compute power, Power Systems can provide outstanding performance for NGS workloads, especially for workloads that are highly parallelized or have large memory footprints. Power Systems achieve higher performance per core through:

- Massive parallelism (threads)
- Higher clock frequencies
- 8-way simultaneous multi-threading (SMT) per core
- Larger POWER® L3 on-chip cache
- PowerVM® optimization

All IBM Power® server models include reliability, availability and serviceability (RAS) features that help avoid unplanned downtime. IBM RAS engineers have optimized server design to help ensure that IBM Power servers support high levels of concurrent error detection, fault isolation, recovery and availability. IBM Power Security and Compliance (PowerSC™) enables security compliance automation and includes reporting for compliance measurement and audit. Automation capabilities include supplying prebuilt system profiles that enforce compliance to various industry regulations, such as the Health Insurance Portability and Accountability Act (HIPAA).

### IBM software defined infrastructure

IBM software defined infrastructure (SDI) offerings transforms a static IT infrastructure into an agile workload- , resource- and data-aware environment. The offerings are part of a comprehensive portfolio of workload and resource management tools – IBM Spectrum Computing and IBM Spectrum Storage – that enables any organization to deliver IT services in the most efficient way possible, optimizing resource use based on workload and data volumes and on business priorities. It is the foundation for a fully integrated software defined environment, enabling the compute, storage and networking infrastructure to adapt dynamically to changing business requirements for faster time to results and lower costs.

IBM Spectrum LSF is a powerful workload-management platform for demanding distributed and mission-critical high-performance computing environments. It provides a comprehensive set of intelligent, policy-driven scheduling features to make full use of compute infrastructure resources and to ensure optimal application performance.

For I/O intensive workloads typical in NGS, file systems and storage play a big role. IBM Spectrum Scale (based on IBM GPFS) is a scalable, high-performance data and file management solution that is proven to benefit various NGS workloads. Spectrum Scale also provides seamless capacity expansion, improved enterprise-wide efficiency, commercial-grade reliability, business continuity and the flexibility of supporting a wide variety of platforms.

### The Lab7 on IBM difference

There are several compelling reasons why life science researchers should consider using the Lab7 ESP on IBM for their NGS workloads:

- A forward-looking, end-to-end and collaborative solution for genomics research and medicine. The IBM Reference Architecture for Genomics was developed in collaboration with leading researchers and partners to address the compute, data and workflow needs in genomics research, translational medicine and personalized medicine.

- Comprehensive software platform: Lab7 ESP is the only software for NGS that combines LIMS and informatics functionalities to manage, analyze and report data accurately.
- Data provenance: Lab7 ESP maintains continuous data provenance by tracking the history of samples, analyses and results and by providing detailed audit trails.
- Sequencing platform flexibility: Lab7 ESP can manage data generated from any sequencing platform.
- Superior compute and data infrastructure: IBM Power Systems are based on powerful IBM POWER processors that deliver better performance per core compared to x86. Power Systems are widely used in supercomputing where scalability and throughput is critical. An IBM Power System solution, expertly integrated with a tuned IBM Spectrum Scale and IBM Spectrum LSF, can process massive amount of NGS data with superior performance and scalability and fully use all available resources to deliver maximum throughput. In addition, these systems possess outstanding enterprise-grade reliability and security.
- Platform flexibility: With the Lab7 ESP solution on IBM Power Systems, users can distribute their workflow across compute platforms, if needed. This flexibility makes it easy to take advantage of the performance and reliability of Power Systems without sacrificing the familiarity and functionality of existing or legacy systems.
- Total cost of ownership: IBM Power Systems are affordable compared to like-sized x86 systems. While the hardware costs are comparable, there are significant gains on software and support costs.

### Performance of the Lab7 on IBM solution

It could take days or weeks to process an end-to-end variant detection pipeline for a human genome sample, from pre-processing sequence data (fastq files) to generating filtered variant files (vcf files). With the Lab7 ESP running on IBM Power8® and IBM software defined infrastructure offerings, pipeline runtime can be significantly improved and therefore can help researchers achieve meaningful results much faster.

A dataset that contains reads from 50x coverage of the whole human genome can be processed in under 12 hours (Table 1). A150x coverage of the human whole exome can be processed in less than two hours (Table2). The following charts show the runtime of each step of a variant detection pipeline using the Lab7 ESP on the IBM Power8 solution. (The tools used in the pipeline, BWA mem, Samtools and bcftools for Power8 are available for download at www.biobuilds.org)

| Pipeline Tool | Runtime (h) |
|---|---|
| BWA | 4.09 |
| Samtools sort | 3.18 |
| Samtools index | 0.13 |
| Samtools mpileup | 4.03 |
| bcftools filter | 0.01 |
| Total | 11.44 |

*Table 1.* 50x WGS reads processed using Lab7 ESP on Power8 with 20 Cores at 3650MHz w GPFS

| Pipeline Tool | Runtime (h) |
|---|---|
| BWA | 0.37 |
| Samtools sort | 0.65 |
| Samtools index | 0.05 |
| Samtools mpileup | 0.78 |
| bcftools filter | 0.01 |
| Total | 1.86 |

*Table 2.* 150x WES reads processed using Lab7 ESP on P95a06 Power8 with 20 Cores at 3650MHz w GPFS

## Why IBM?

A compute architecture using IBM Reference Architecture for Genomics, IBM Power Systems and IBM software defined infrastructure offerings delivers outstanding performance for the large compute power and storage capacity demanded for sample tracking and protocol management and powerful data analysis, reporting, and visualization tools offered by the Lab7 Enterprise Sequencing Platform (ESP).

IBM Spectrum Computing offers a comprehensive portfolio of software defined infrastructure solutions designed to help your organization deliver IT services in the most efficient way possible, optimizing resource utilization to speed time to results and reduce costs. These offerings help maximize the potential of your infrastructure to accelerate your analytics, high-performance computing (HPC), Apache Hadoop, Spark and cloud-native applications at any scale, extract insight from your data and get higher-quality products to market faster.

Whether deployed in a data center or on the cloud, IBM Spectrum Computing solutions fuel product development, critical business decisions and breakthrough insights in financial services, manufacturing, digital media, oil and gas, life sciences, government, research and education. From designing Formula One race cars to credit risk analysis, organizations in a wide variety of industries are using IBM Spectrum Computing as a foundation for software defined infrastructure solutions for big data, analytics, HPC and cloud to improve business results.

## For more information

To learn more about Lab7 ESP, visit: www.lab7.io

To learn more about IBM Reference Architecture for Genomics, IBM Power Systems, and IBM software defined infrastructure offerings including IBM Spectrum LSF and IBM Spectrum Scale, contact your IBM representative or IBM Business Partner, or visit:

- **ibm.com**/systems/spectrum-computing/industries/ lifesciences.html
- **ibm.com**/systems/power
- **ibm.com**/spectrumcomputing

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing