**IBM Systems**
Solution Brief

# Accelerating discovery at a lower cost in genomics medicine

*Deeper, faster insights with composable building blocks based on IBM Spectrum Scale*

## Highlights

- Store, access, secure, manage, share and analyze significant amounts of genomics data that emerge constantly, creating an expanding data ecosystem
- Leverage an integrated solution for genomics based on composable infrastructure that disaggregates the underlying compute, storage and network resources
- Allow clinical and research organizations to analyze massive amounts of genomics data quickly and to identify new patterns and relationships
- Enable IT architects and IT administrators to easily design, install and manage deployment in a timely manner without being overwhelmed

Advancing the science of medicine by targeting a disease more precisely with treatment that is specific to each patient relies on access to that patient's genomics information and the ability to process massive amounts of genomics data quickly. According to survey results published in the *NEJM Catalyst*, a publication of the New England Journal of Medicine, 40 percent of respondents said genomics data will become one of the most useful data sources in five years, up from just 17 percent today.[1]

As genomics data becomes a critical source for precision medicine, it is expected to create an expanding data ecosystem. This means that hospitals, genome centers, medical research centers and other clinical institutes need to explore new approaches that will help them unleash the real value of significant amounts of genomics data.

A key takeaway of a new cognitive healthcare study conducted by HIMSS Analytics, was that data management is the clear top area of investment, with 32 percent of the participants marking it as the Number 1 priority.[2]

Healthcare and life sciences organizations that are running data-intensive genomics workloads on an IT infrastructure that lacks scalability, flexibility, performance, management and cognitive capabilities will need to modernize and transform their infrastructure to support current and future requirements.

IBM offers an integrated solution for genomics based on composable infrastructure—a model where compute, storage and network are treated as well-defined services. This solution enables administrators to build an IT environment in a way that disaggregates the underlying compute, storage, and network resources.

A composable, building-block-based solution for genomics such as the IBM offering addresses the most complex aspect of data management and allows organizations to store, access, manage and share huge volumes of genomic sequencing data. In addition, it addresses workload management challenges to enable IT to build, refine, submit and orchestrate computational jobs with maximum resource utilization.

These disaggregated building blocks provide the required granularity for enhanced flexibility of the infrastructure, allowing information to be sliced, diced, expanded and contracted based on the actual need. The solution enables clinical and research organizations to analyze massive amounts of genomics data quickly and to identify new patterns and relationships, helping accelerate discoveries, treatments and insights.

## Composable infrastructure-based genomics solution from IBM

IT administrators, physicians, data scientists, researchers, bioinformaticians and other professionals involved in the genomics workflow need the right foundation to achieve their objectives efficiently while improving patient care. By engineering using design-thinking methods, IBM has developed a consumable, easy-to-deploy solution for genomics workloads based on the principles of composable infrastructure. It includes three disaggregated building blocks:

- Storage services
- Compute services
- Network services

Composable infrastructure helps researchers deploy the resources they need for data-intensive genomics research.

This design creates a single, cohesive infrastructure for genomics that, because it is composable, can scale each individual resource independently.

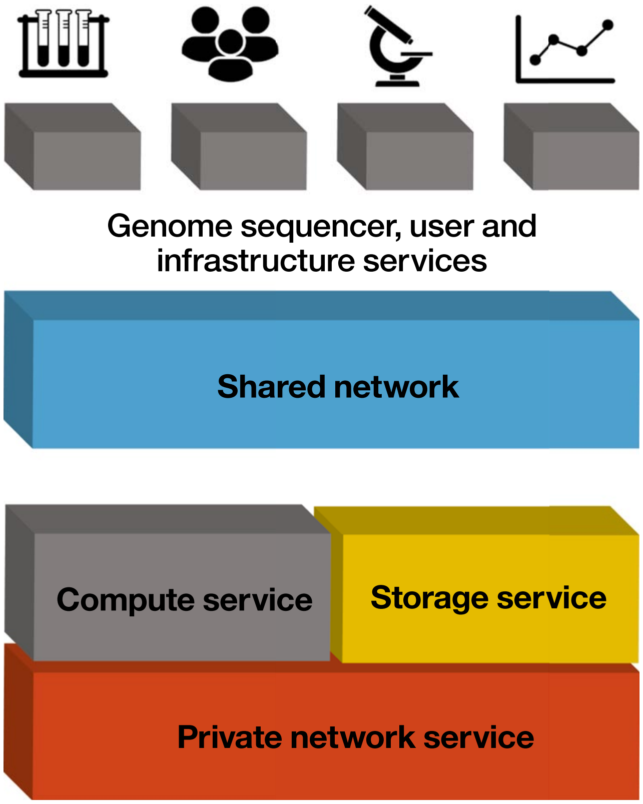
Using IBM software-defined infrastructure (SDI), the solution provides healthcare and life sciences leaders features that include:

- Fully qualified and tested stack for next-generation sequencing (NGS) workloads with crisp configuration templates including tuning and optimization of the blocks to efficiently run complete workflows from the Broad Institute Genome Analysis Toolkit (GATK) and Burrows-Wheeler Aligner (BWA).
- Easy and rapid assembly of tested composable infrastructure to create a scalable, high-performance computing (HPC)-like cluster to analyze genomics data.
- Flexible deployment models derived from the composable design, allowing simple and easy scalability of the storage, compute and network building elements. These models include best practices for implementing genomics clusters in different configuration sizes based on actual customer needs.
- Agnostic compute platforms with the support for IBM® Power® and/or x86 systems.
- Innovative graphical user interface for submitting and managing compute jobs and for viewing cluster status and utilization.
- User nodes to access interactive applications.

The IBM solution can enable data scientists to efficiently deliver results to physicians, meet physicians' on-demand requests for analysis and improve the underlying infrastructure to meet core objectives. Using this solution, IT architects and IT administrators can easily architect, install and manage deployment in a timely manner without being overwhelmed.

## IBM Spectrum Scale: Composable storage building block for genomics

The IBM solution for genomics offers a well-tested, policy-driven, storage building block based on IBM Spectrum Scale™ designed for high-performance and large-scale workloads with integrated data analytics.



Composable building block for genomics.

### Understanding the need

The size, variety and unstructured nature of genomic sequencing workloads create demanding storage requirements. The NGS pipeline requires the underlying storage to be scalable in terms of capacity and performance. The workflow involves data being ingested from sequencers, processing of raw data, genome alignment and variant detection to support analytics and provide efficient collaboration among researchers, scientist and physicians.

This workflow poses challenges of input/output performance, data management, multi-protocol support and big-data integration for the storage system. At the same time, it also must address storage cost, build an effective long-term retention strategy, and ensure security safeguards to protect the genomics data.

Genomics clusters are also very dynamic and vary in size. These characteristics mean storage systems must be able to fit like building blocks within the composable infrastructure. It is therefore important to have the right kind of storage architecture for this demanding workload. Traditional scale-out network-attached storage (NAS) architectures are not ideal for genomics workloads as they were not built for these new workloads.

Traditional NAS devices suffer from various limitations, including the inability to understand the performance characteristics and scalability requirements of the different stages in the genomic sequencing workflow. This shortcoming leads to non-optimized, inefficient storage deployments—which implicitly leads to the suboptimal utilization of compute resources when servers are idle and waiting for arrival of data that is scheduled to be analyzed.

## IBM Spectrum Scale for Genomics

Genomic sequencing is a technical computing workload that can benefit from the practices of HPC where scalability and performance are among the key challenges to storage systems. IBM Spectrum Scale is a leading software-defined storage system specifically designed to excel with such workloads due to its scalability and performance characteristics. Additionally, IBM Spectrum Scale enables flexible deployment options for genomics workloads allowing organizations to start small and scale out quickly based on needs.

IBM Spectrum Scale is available as a tightly coupled offering called IBM Elastic Storage™ Server based on IBM POWER® server technology. IBM Elastic Storage Server is a high-capacity, high-performance, ready-to-use storage solution for managing data at scale with the distinctive abilities to perform archiving and analytics on data in-place. Hence, IBM Spectrum Scale and IBM Elastic Storage Server are the foundational building blocks for data storage and data management of life sciences workloads—namely, genomic sequencing.

## Key features

Key features offered by IBM Spectrum Scale and IBM Elastic Storage Server for genomics include:

- Scale-Out File System with POSIX interface to run performance-savvy workflow of genome variant alignment and detection
- Integrated network file system (NFS), server message block (SMB) and object (SWIFT/S3) support for ingest from sequencers as well as collaboration and data sharing with scientists and physicians
- Single Global Namespace to help eliminate the silo of medical data (such as electronic medical records, imaging, genomics and clinical data), offering centralized integrated analytics
- Economic efficiencies resulting from placing the right kind of data at the right tier (from flash to tape or cloud object storage) via automated information lifecycle management policies, leading to a reduction of storage costs by up to 90 percent.[3]
- Scientific collaboration across geographies with support of wide area network caching technology such as active file management
- In-place Hadoop-based analytics
- Long-term retention of genome data by integrating with tape or IBM Cloud object storage

- Integrated data protection with secure data at rest and secure data in motion across all protocols
- Support for both x86 as well as POWER-based compute systems
- End-to-end checksum to help ensure data integrity all the way from application to disk
- Quota management for user and project groups
- Snapshots for user and project groups

IBM Spectrum Scale and IBM Elastic Storage Server can provide the simplified management, high availability, economics and enhanced performance with scalability needed to manage life sciences workflows. These solutions support today's needs and future growth, making them ideal storage building blocks for genomics.

## IBM Spectrum Computing: Efficient building block for genomics

Optimization of workload orchestration and collaboration, along with greater resource allocation, are critical capabilities that must be addressed in the compute layer to manage genomic sequencing workflow efficiently. It is vital for compute solutions to be provisioned with the right size and ability to grow or shrink as needed. Compute disaggregation from storage should allow a genomic deployment to plan and deploy compute independent of storage. A requirement for change in storage infrastructure should not impact the compute part of the genomic deployment. Thus, the need is to define the compute building blocks as independent of storage building blocks, so they can be rapidly scaled up or down based on actual needs.

IBM Spectrum™ Computing is SDI based on advanced software designed to consolidate and transform static, siloed systems into a dynamic, integrated and intelligent policy-driven infrastructure. IBM Spectrum Computing empowers healthcare and life sciences organizations to efficiently allocate resources and scale workloads, improve IT agility, lower time to results and help reduce costs.

IBM Spectrum LSF and IBM Spectrum Conductor with Spark can manage workloads for demanding, distributed and mission-critical HPC environments such as genomics data analysis. For example, in an IBM Spectrum LSF-enabled environment, multiple instances of GATK for processing configurations run in parallel, managed by IBM Spectrum LSF, leading to an optimization of compute resources and ensuring optimal application performance.

IBM Spectrum Computing paired with IBM Spectrum Scale provides a compelling compute building block for genomic sequencing tasks. The IBM solution for genomics offers a well-defined and tested compute and storage building block that can be rapidly assembled and scaled in both directions based on workload needs.

Organizations can build the compute building block on an IBM Power or x86 architecture using the IBM solution for genomics.

## Network building block for genomics

Networking forms an integral part of the overall genomics solution. The IBM solution for genomics includes a well-defined and fully tested network building block along with storage and compute blocks, making the overall solution a composable infrastructure for genomics data.

## Blueprints and runbooks

Also available are fully tested and documented blueprints and runbooks. These can help organizations deploy proven solutions for genomics faster based on IBM best practices from across the globe.

The blueprints consist of a composable solution with detailed architecture definitions for storage, compute and networking services for genomics. These definitions enable solution architects to benefit from tried-and-tested deployments, helping them quickly plan and architect an end-to-end infrastructure deployment.

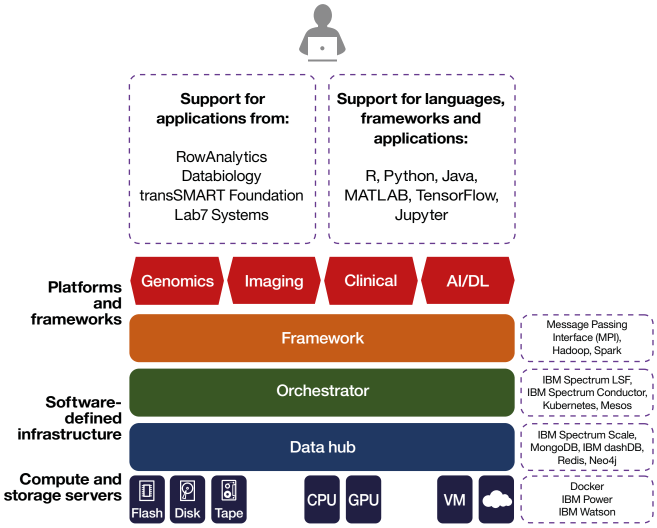
The blueprints and fully tested runbooks provide capabilities required, including ready-to-use configuration and tuning templates, for the different building blocks (compute, network and storage). The combination of these capabilities enables simpler deployment and enlarges the level of assurance for genomics workload performance. The ready-to-consume blueprints and runbooks outline scale-in or scale-out compute for network and storage building blocks that are independently based on the needs of genomics deployments. The solution is designed to be elastic in nature. The disaggregation of building blocks allows IT administrators to easily and optimally configure the solution with maximum flexibility.

Blueprints and runbooks help to build a solution that serves different needs of the technical, functional and clinical leaders involved in genomics workflow. They serve as catalysts for achieving a more optimized and effective outcome.

## IBM Reference Architecture for Genomics: A holistic solution

The IBM reference architecture for high-performance analytics in healthcare and life sciences[4] is an end-to-end architecture designed to address the common requirements of organizations pursuing genomics, personalized medicine and other big-data initiatives in biomedical research. IBM Spectrum Scale and IBM Elastic Storage Server are central to this reference architecture and are well-integrated with other ecosystem partners such as IBM Spectrum LSF, IBM Spectrum Conductor with SPARK, Hortonworks for Hadoop-based analytics and IBM Watson™.

**Reference architecture for high-performance data analytics**


Support for applications from:
RowAnalytics
Databiology
transSMART Foundation
Lab7 Systems
Support for languages, frameworks and applications:
R, Python, Java, MATLAB, TensorFlow, Jupyter
Platforms and frameworks
Genomics
Imaging
Clinical
AI/DL
Framework
Message Passing Interface (MPI), Hadoop, Spark
Software-defined infrastructure
Orchestrator
IBM Spectrum LSF, IBM Spectrum Conductor, Kubernetes, Mesos
Data hub
IBM Spectrum Scale, MongoDB, IBM dashDB, Redis, Neo4j
Compute and storage servers
Flash
Disk
Tape
CPU
GPU
VM
Docker
IBM Power
IBM Watson

## Proven solutions at leading institutions

The success of clinical and research organizations depends on the ability of IT systems to fulfill a common set of requirements. Leading institutions that have based their genomics workload on IBM technologies are benefitting from optimized performance, cost and time for final analysis:

Sidra Medical and Research Center is advancing Qatar's biomedical research capabilities with a unified storage and compute infrastructure that now serves as a national resource for researchers and scientists and as the high-performance data analytics platform for the Qatar Genome Programme (QGP).

Icahn School of Medicine, Mt. Sinai is accelerating genetics research and medicine 500 percent with IBM SDI. Core dumps fell to zero over a one-year period while overall scalability increased to 500,000 jobs per queue.

Louisiana State University is increasing scale and speed for Hadoop-enabled genome analysis, enabling the development of new tools for sequencing and analysis. The converged, high-performance analytics solution that is based on the IBM Reference Architecture for Genomics helps to achieve three times the performance with one third the compute hardware.

Lab7 designed a complete solution for the management and analysis of genomic-scale data workloads and lab operations that provides a balanced mix of simplicity, speed to value, cost and performance. Its Genomic Cloud can support up to 500 compute nodes, with more than 10 petabytes of storage, helping achieve faster results and optimized resource usage for HPC applications.

## Why IBM?

IBM is a leading provider of flexible and scalable high-performance storage and SDI. The SDI architecture optimizes data for cost, performance, collaboration and compliance—on-premises and in the cloud—to ensure the highest levels of data availability and reliability.

IBM Storage and SDI solutions offer an efficient foundation to deliver faster time-to-results for compute and data-intensive workloads, enabling organizations to cost-effectively accelerate discoveries to support precision medicine.

## For more information

To learn more about IBM Spectrum Scale and IBM Elastic Storage Server, contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/us-en/marketplace/scale-out-file-and-object-storage and **ibm.com**/us-en/marketplace/ibm-elastic-storage-server

For more information please review:
**IBM Redpaper: IBM Spectrum Scale Best Practices for Genomics Medicine Workloads:**
http://www.redbooks.ibm.com/abstracts/redp5479.html

**IBM Reference Architecture for Genomics: Speed, Scale, Smarts:**
http://www.redbooks.ibm.com/abstracts/redp5210.html?Open

**IBM Reference Architecture for Genomics, Power Systems Edition:**
http://www.redbooks.ibm.com/abstracts/sg248279.html?Open

**Performance optimization of Broad Institute GATK Best Practices on IBM reference architecture for healthcare and life sciences:**
https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=TSW03540USEN

**IBM reference architecture for high performance analytics in healthcare and life science:**
https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=TSW03536USEN&

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

IBM Systems
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
January 2018

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions. It is the user's responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs.

All client examples cited or described are presented as illustrations of the manner in which some clients have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions. Contact IBM to see what we can do for you.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Actual available storage capacity may be reported for both uncompressed and compressed data and will vary and may be less than stated.

♻ Please Recycle

[1] Amy Compton-Phillips, "What Data Can Really Do for Health Care," *NEJM Catalyst*, NEJM Group, Massachusetts Medical Society, March 2017: http://join.catalyst.nejm.org/hubfs/Insights%20Council%20Monthly%20-%20Files/Insights%20Council%20March%202017%20Report%20What%20Data%20Can%20Really%20Do%20for%20Health%20Care.pdf?utm_campaign=Insights_Council_Monthly_March2017&utm_source=hs_automation&ut

[2] Read the HIMSS Analytics study here: https://www-01.ibm.com/marketing/iwm/dre/signup?source=urx-20814&S_PKG=ov60762

[3] IBM lab measurements, August 2015

[4] "IBM reference architecture for high-performance analytics in healthcare and life science," *IBM Corp.*, November 1017. https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=TSW03536USEN&