IBM Systems
Technical White Paper

November 2017

A reference architecture for high performance analytics in healthcare and life science

Accelerate personalized healthcare and other biomedical workloads using a cost-effective, high-performance infrastructure for big data analytics

Overview

Challenge
Healthcare and life science organizations worldwide must manage, access, store, share, and analyze big data within the constraints of their IT budgets.

Solution
IBM's reference architecture for healthcare and life sciences defines a platform for delivering the highest levels of performance for big data workloads, while also lowering the total cost of IT ownership.

Advancements in high-throughput molecular profiling techniques and high-performance computing systems have ushered in a new era of personalized healthcare where the treatment and prevention of a disease can be tailored to the unique molecular profiles, behavioral characteristics, and environmental exposures of individual patients. Discovering personalized therapies appropriate for different patient cohorts—and delivering such treatment plans in a clinical setting—requires technical computing platforms that have the ability to analyze patient characteristics and accurately predict clinical outcomes in a timely manner. For many biomedical research and clinical organizations, large-scale initiatives in personalized healthcare are technically daunting for the following reasons:

Big data: Biomedical data required to support personalized healthcare initiatives are large, varied, and often unstructured; moreover, the amount of data collected as part of these initiatives usually grows exponentially and must be archived for extended periods of time—sometimes for decades. Common data sources for personalized healthcare applications include whole genome sequences; biomedical imaging from clinical and laboratory instrumentation; electronic medical records systems; physiological sensors or wearables; and collections of curated scientific and clinical literature. Organizations often lack the storage capacity to keep up with rapidly expanding data volumes.

Data silos: Personalized healthcare requires the aggregation of information that can provide a full view of the biological traits, behaviors, and environmental exposures of each patient. However, data for a single patient is usually captured and scattered across heterogeneous storage silos within health systems and biomedical research organizations. Before disparate data sets can be analyzed, they must be integrated into a common database—a process which is often experienced as manual, painstaking, and time-consuming.

Compute- and data-intensive workloads: Analytics workloads can be extremely compute- and data-intensive. Common examples include I/O-intensive analysis pipelines that transform raw next generation sequencing data into genomic variant files; deep learning techniques for discovering patterns within complex biomedical data sets; and large-scale data mining of clinical and scientific documents. Such workloads might take hours, or even days, to complete on existing technical computing platforms.

Evolving applications and frameworks: Personalized healthcare initiatives must often support hundreds of different applications at any given time, including those related to medical informatics, genomics, image analysis, and deep learning. Such applications are often built on frameworks and databases that are continually evolving, including Spark, Hadoop, TensorFlow, Caffe, Docker, MongoDB, HBase, and a variety of graph databases. Biomedical research organizations often have difficulty supporting multiple versions of these application frameworks and databases, which are proliferating and frequently evolving—sometimes two or more times per year.

Collaboration: Data sharing across institutions—and often across geographic boundaries—is a growing necessity in the study of rare diseases and complex disease mechanisms. International scientific consortia consisting of academic, commercial, non-profit, and government entities are rapidly emerging for sharing biomedical data and related analytics. Collaborating partners often lack solutions for sharing their data sets rapidly and cost effectively without compromising protected health information and intellectual property rights.

Many organizations worldwide are finding it difficult to overcome these challenges, especially within the constraints of their IT budgets. Clinical and scientific research data must be accessed, stored, analyzed, shared, and archived in a time-efficient and cost-effective manner; but for many healthcare organizations, biomedical research institutions, and pharmaceutical companies today, data are collected in such large volumes that these organizations can no longer process, properly store, or transmit these data over regular communication lines in a timely manner. For many organizations, compute and storage silos are proliferating across clinical and research groups, as analysts collect increasing volumes of data and use those data in complex analytical workloads. To move data across long physical distances, organizations often resort to disk drive and shipping companies to transfer raw data to external computing centers for processing and storage, thus hindering speedy access and data analysis.

System components

Workload and data management middleware

- IBM® Spectrum Scale™
- IBM Spectrum LSF®
- IBM Spectrum Conductor™ with Spark

Compute and storage hardware

- IBM Power Systems™
- IBM Elastic Storage Server

Network

- IBM Aspera®

Cloud environments

- IBM SoftLayer®
- IBM Cloud Object Storage

Key platform capabilities

- **Scale** to support exponentially growing volumes of big data
- **Flexibility** to support evolving analytics applications built on Spark, Hadoop, Docker, and other frameworks
- **Simplified integration** of biomedical data across storage silos
- Storage, management, and analysis of **unstructured data**
- **Metadata** capture and storage for searchability, repeatability, and auditability of data and workflows
- **Easy collaborations** across geographic boundaries
- **Security** for protected health information and protection of intellectual property rights
- Easy and cost-effective **IT administration**

In order to overcome the technical challenges industry practitioners face in the era of personalized healthcare, IBM Systems has created a reference architecture for healthcare and life sciences. This reference architecture, which is built on IBM's history of delivering best practices in high-performance computing (HPC), can make it possible for healthcare and life science organizations to easily scale compute and storage resources as demand grows, and to support the wide range of development frameworks and applications required for industry innovation—all without unnecessary re-investments in technology. A description of the IBM reference architecture is provided in the next section.

IBM reference architecture: A diverse computing platform built on a common infrastructure

Whether they are university researchers publishing journal articles looking to win their next round of grant funding; scientists in commercial R&D organizations progressing potential drugs through clinical trials; or physicians in hospitals delivering treatments that will give their patients the best clinical outcomes, key stakeholders in personalized healthcare initiatives will depend on a reliable, flexible computing platform that meets their diverse application needs. Through the team's work with many international clients and healthcare industry business partners, IBM Systems Group has learned that organizations pursuing genomics, personalized healthcare, and other big data initiatives in biomedical research will need high-performance systems with key platform capabilities.

The reference architecture for healthcare and life sciences (as shown in Figure 1) was designed by IBM Systems to address this set of common requirements. It reflects the current evolution in HPC, where technical computing systems need to address the batch workloads of traditional HPC, as well as long-running analytics involving big data. The ability to take compute and storage components that are running HPC algorithm codes, and then dynamically provision them to handle other types of analytics is a more cost-effective and more easily managed alternative to maintaining two distinct systems. In the era of diverse computational workloads, diverse infrastructure needs, and diverse versions of application frameworks, it is important to avoid creating siloed compute clusters that typically underutilize IT resources and result in poor IT cost containment. The ability to use fewer compute resources using intelligent policy-based workload and data management tools is a key aspect of supporting a dynamic research environment.

With data volumes growing so rapidly, storage systems must be scalable from a capacity and performance perspective. Storage systems must also be designed such that the same storage media can support different access methods in a reliable, yet flexible manner. For example, a common storage server that can support parallel I/O access methods must also effectively handle unstructured data access from platforms such as Hadoop. In addition, technical computing environments that serve these ever-changing workload requirements should be easy to manage. Simple and centrally managed administration of the storage systems not only allow researchers to be more productive, but can also lower IT and related costs across the enterprise.

Scaling high-performance technical platforms to support growing data volumes and diverse applications—while continuing to accelerate workloads and minimize IT costs—requires a flexible yet tightly coordinated framework for data access, compute, and storage.

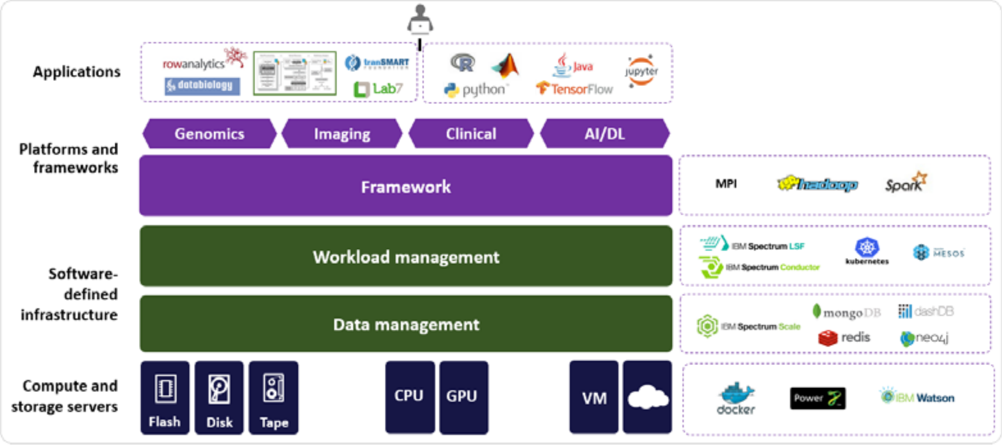

Figure 1. IBM reference architecture for healthcare and life sciences

Foundational elements: Spectrum Computing, Spectrum Scale, and Power Systems

The IBM reference architecture (in Figure 1) reflects the work that is continually underway within IBM Systems to integrate elements of IBM's compute and storage portfolio such that they deliver high levels of performance for big data, while also lowering the total cost of IT ownership. The top two layers represent the applications, databases, and frameworks that researchers and clinicians are utilizing (*purple boxes*); the bottom layer represents virtual or physical high-performance compute and storage servers where data are processed and stored (*blue boxes*); and the two middle layers (*green boxes*) contain software that enables a diverse set of biomedical applications to run on shared compute and storage servers efficiently and cost effectively.

Of the two middle layers, the workload management layer makes it possible to distribute thousands of computational workflows, in parallel, across enterprise compute servers in a way that maximizes utilization of those servers. The data management layer enables low-latency data access; convergence of heterogeneous data silos; metadata collection; and automated information lifecycle management. It gives organizations the ability to rapidly scale compute and storage capacity upwards or shrink capacity downward as workloads demand.

Workload management: IBM Spectrum Computing

The workload management layer dynamically and elastically allocates computational tasks across compute servers in a manner that is transparent to the user. It consists of multiple coherent workflow schedulers that are coordinated to place diverse compute jobs on local and remote clusters in an efficient and cost-effective way. IBM's resource-aware and policy-based schedulers include industry-leading Spectrum LSF (for HPC batch workloads), Spectrum Conductor with Spark (for Spark workloads), and IBM Spectrum Symphony™ (for Hadoop MapReduce and near real-time workloads). As shown in Figure 2, these schedulers are tightly integrated: that is, if one type of workload is using only a few resources in a cluster, then the other workload types can fully use the remaining resources in that cluster. The flexibility and elasticity of server utilization across these schedulers eliminate the need for IT organizations to provide dedicated clusters for each of the workload types. When serving a multitenant environment, these schedulers protect individual tenants by way of secure isolation and role-based access control. They make it possible for workloads to be distributed seamlessly across multiple physical and cloud environments (see Figure 3), and they support the distribution of workloads deployed in Docker and other container technologies.

IBM Spectrum LSF: Spectrum LSF is a highly scalable and reliable resource-aware workload management platform that supports demanding, distributed, and mission-critical HPC environments. It has been selected as the preferred workload management system by large genome analysis organizations for its ability to routinely orchestrate hundreds of thousands of jobs that are submitted in batch, and for its ability to readily scale with growing user demand. Clients worldwide are using technical computing environments supported by LSF to run hundreds of genomic workloads, including Burrows-Wheeler Aligner (BWA), SAMtools, Picard, GATK, Isaac, CASAVA, and other frequently used pipelines for genomic analysis.

Additional capabilities for managing and monitoring workloads are offered by the following companion software:

- **IBM Spectrum LSF Process Manager**: This user application simplifies the design and automation of complex computational workflows, and allows those workflows to be shared with collaborators in a secure environment. IBM clients interested in personalized healthcare research have invested in the LSF Process Manager to simplify the process of writing genomic workflow scripts that transform raw data from next generation sequencers (in the FASTQ format) into variant files (for example, VCF, SNV, CV) for downstream analysis. The Spectrum LSF Process Manager makes it possible for bioinformaticians to share their workflows with selected users who may or may not have formal experience in writing scripts, while also helping to maintain strict version control of those scripts.

- **IBM Spectrum LSF Application Center**: This application provides users with a simple and intuitive web-based interface for accessing, submitting, managing and monitoring workflows created with Spectrum LSF Process Manager. IBM Spectrum LSF Application Center can also serve as a multi-tenant enterprise portal by which users and external collaborators can submit, run, and monitor analysis pipelines under fine-grained role-based access controls.

- **IBM Spectrum LSF RTM**: This operational dashboard provides IT administrators with comprehensive workload monitoring, reporting, and management tools. IT administrators can track workload metrics related to cluster utilization and efficiency with high granularity—from the level of the cluster all the way down to specific users, user groups, or workload types.

- **Spectrum LSF Data Manager**: This capability enables administrators to write policies that automatically transfer data in an LSF-managed cluster (on-premises or on the cloud) to a cache residing close to the execution compute cluster. Such policies can automate the movement of data to remote locations, thereby improving data throughput and reducing compute cycle times.

IBM Spectrum Scale: Key features

- **Extreme scalability** to billions of petabytes and hundreds of GBps throughput
- **Modular expansion of storage capacity** with minimal service disruption
- **Low-latency data access** with best-in-class performance for I/O-intensive workloads
- **Single global namespace** for files and directories across heterogeneous storage clusters
- **Active file management** (AFM) caching technology to speed up data sharing among remote collaborators
- Policy-driven **information lifecycle management (ILM)** across tiers of storage media
- **Metadata** identification, collection, and search
- Policy-based **file encryption**
- **Easy administration** from a single pane of glass

IBM Spectrum Conductor with Spark: Apache Spark is an application framework that is commonly used by individuals conducting computational research in personalized healthcare and other areas of biomedical science. High-performance technical platforms supporting traditional HPC batch workloads must also support a growing number of Spark workloads. IBM Spectrum Conductor with Spark is an enterprise-grade, multitenant solution for Apache Spark that enables simultaneous support of multiple instances of Spark, and eliminates resource silos that would otherwise be tied to separate Spark implementations. By providing advanced resource sharing of all Spark workloads and different versions of Spark on a single compute platform, Spectrum Conductor allows organizations to increase utilization of existing compute servers and thereby improve IT cost containment. Moreover, it facilitates management of the workloads through a user-friendly interface.

IBM Spectrum Symphony. While not yet widely used in technical computing environments for healthcare applications or biomedical research (as is IBM Spectrum LSF), IBM Spectrum Symphony is a highly scalable enterprise grid manager that can schedule jobs at a latency on the order of milliseconds, and thereby support near-real-time analytics. IBM Spectrum Symphony Advanced Edition also includes an Apache Hadoop-compatible MapReduce implementation that has been demonstrated in an audited benchmark to deliver, on average, four times the performance of open source Hadoop.

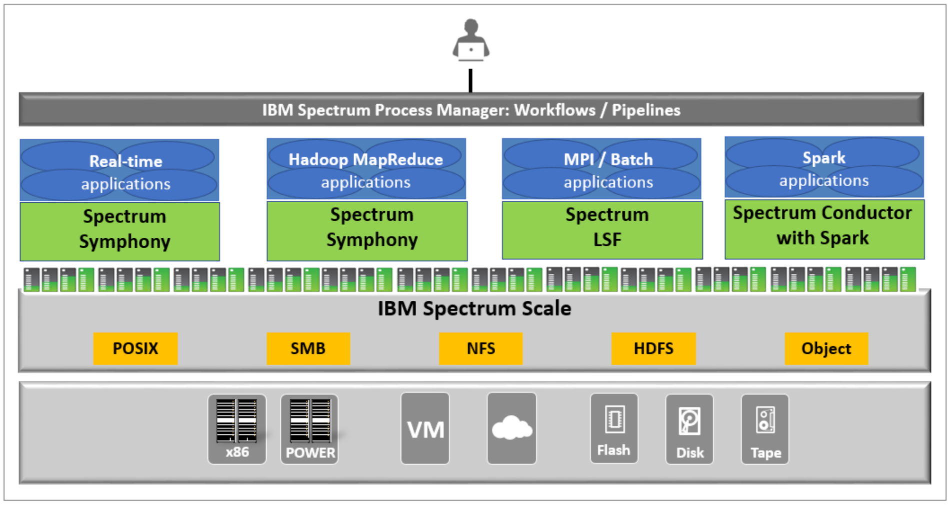

Figure 2. Workload and data management with IBM Spectrum Computing and Spectrum Scale

Data management: IBM Spectrum Scale

In the IBM reference architecture, data management for computational workloads is enabled by IBM Spectrum Scale, which was formerly known as the IBM General Parallel File System (IBM GPFS™). Spectrum Scale is a proven, scalable, high-performance data and file management solution providing world-class storage management with extreme scalability. Leading genomics centers, medical institutions, and pharmaceutical companies worldwide are already investing in Spectrum Scale to store, archive, process, and manage a vast amount of structured and unstructured data, including genomic sequences, biomedical images, and electronic medical records. Spectrum Scale is well-suited for managing biomedical data and related analytics because it addresses the following challenges:

- Rapid growth of data volumes
- I/O-intensive workloads, which is often required in the analysis of raw genomic sequences
- Heterogeneous file and object storage clusters using different operating systems, storage access protocols, storage media, and storage hardware
- Data sharing across globally distributed projects
- Metadata identification, collection, and search required for scientific and clinical repeatability, validation, and long-term archive
- Requirements for secure storage and secure deletion of data containing sensitive information

Moreover, Spectrum Scale enables policy-driven Information Lifecycle Management (ILM) across multiple storage tiers built on flash, disk, and tape media across local or remote locations. Automated policies make it possible for administrators to define where, when, and on what media data (or metadata) will be stored to maximize workload performance and minimize overall storage costs. For example, low-latency flash systems might provide the best price performance for small volume and highly utilized data in I/O-intensive workloads; by contrast, LTFS tape offers the best price performance for large volume and less utilized data sets that are ready for long-term archive.

IBM Elastic Storage Server (ESS) is a storage implementation of IBM Spectrum Scale software that has been integrated with IBM POWER8® processor-based servers and IBM disk arrays. ESS is suitable to meet the demands of big data workloads in biomedical research as it offers all the benefits of Spectrum Scale. The multithreading, large-memory bandwidth, and large cache size offered by IBM POWER8 processors enhance data throughput for I/O-intensive workloads. In addition, software declustered Redundant Array of Independent Disks (RAID) protection schemes that are installed with ESS significantly reduce critical rebuild times during moments of drive failure.

Hybrid cloud architecture built on IBM Spectrum Computing and IBM Spectrum Scale

As healthcare and life science organizations seek lower capital expenses, easier IT management, and capabilities for better data sharing and external collaboration, many of these organizations are considering high-performance cloud resources to support at least a portion of their workload. The workload and data management layers in the IBM reference architecture make it possible for IBM Systems to implement enterprise architectures for big data workloads in hybrid cloud environments. The tight integration that exists between IBM Spectrum Computing and IBM Spectrum Scale, as illustrated in Figure 3, supports the movement of cloud-enabled applications seamlessly between traditional on-premises servers and off-premises private or public clouds. IT administrators can write policies that define the relative utilization of the physical and virtual environments in a way that best meets their organization's business, technical, regulatory, and financial requirements. For clinicians and biomedical scientists focused on running analysis applications, the movement of workloads across these environments is completely transparent and creates an uninterrupted user experience.

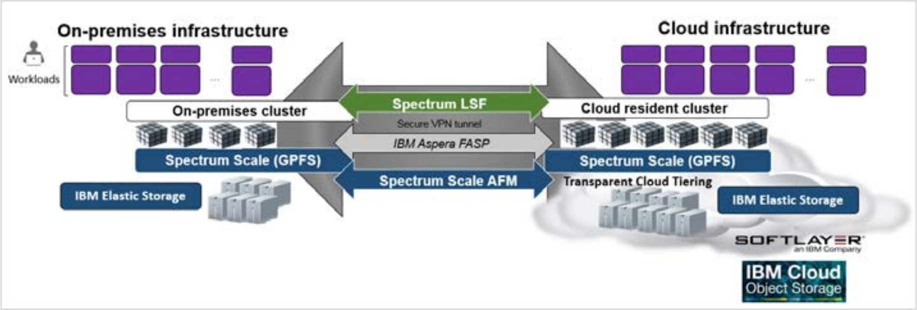

Figure 3. A hybrid cloud solution

Figure 3 depicts additional IBM offerings that can assist in the storage, movement, and management of biomedical data. The IBM SoftLayer cloud offers a high-performance environment built on Spectrum Computing and Spectrum Scale software.

IBM Cloud Object Storage, an offering now available through the acquisition of Cleversafe, provides organizations with a highly secure and massively scalable object-based storage platform. Cloud Object Storage can be deployed on premise or in a hosted IBM SoftLayer cloud environment.

IBM Aspera Fast Adaptive Secure Protocol (IBM FASP®) technology accelerates the transfer of large files and data sets—whether structured or unstructured—over an existing wide area network (WAN) infrastructure. Aspera speeds up data movement—even in remote locations hampered by poorly performing networks—with predictable, reliable, and secure delivery regardless of file size, transfer distance, and network conditions. IBM Aspera is currently used within the healthcare and life science industry to transfer raw human genome sequences (roughly 200 GB per genome) and other large files within a geographically dispersed community.

Accelerated computing with IBM POWER8 processor-based systems

IBM Systems is committed to providing superior performance on the most challenging computational workloads. The IBM strategy for achieving this goal is based on the observation that next-generation HPC systems will need to support data-intensive workloads in addition to compute-intensive workloads. As requirements for traditional HPC and newer big data analytics converge, system throughput depends not only on I/O performance improvements at the level of central processing units (CPUs), but also depends on minimizing data movement within the architecture. It also depends on tightening the integration of CPU with hardware accelerators such as graphics processing units (GPUs) and Field Programmable Gate Arrays (FPGAs), which are often used to dramatically speed up user applications.

OpenPOWER Foundation: In 2013, IBM began open source licensing of products related to the IBM Power Architecture® in order to create an alternative to proprietary solutions and to spur innovation in computing technology. IBM initiated the creation of the OpenPOWER Foundation—a technical community of more than 130 commercial and academic organizations collaborating on the development of IBM POWER® processor-based solutions that better meet specific business needs. Innovations arising from OpenPOWER collaborations have included custom systems for workload acceleration using GPU, FPGA, and advanced I/O.

OpenPOWER processors: OpenPOWER systems are being designed to support compute- and data-intensive workloads, such as machine learning and deep neural networks. The IBM POWER8 processor has simultaneous multithreading (SMT), which allows up to eight hardware threads from a single physical core. It also employs the most advanced memory subsystem available to achieve leading edge performance—making use of a large number of on- and off-chip memory caches. This processor design reduces memory latency and generates very high bandwidths for memory and system I/O.

Recent advancements: The following features have been designed to augment the performance of POWER8 systems:

- **Coherent Accelerator Processor Interface (CAPI**), which allows FPGAs and other accelerators plugged into a PCIe slot to directly access the processor bus at low latency.
- **NVLink**, a high-speed processor interconnect that can be used to connect GPUs to either CPUs or GPUs (Figure 4). NVLink, which was developed in partnership with NVIDIA, eliminates the PCI bottleneck, thereby providing five times or greater performance than PCIe. NVLink enables the logical integration of multiple GPUs and of CPU and GPU cores. Using these connections, each GPU has a direct-paged access to both the memory of the host processor and the memory on the sibling GPU. This model in GPU computing significantly decreases the programming complexity associated with both basic GPU computing and multi-GPU computing.

Applications: POWER8 processor-based servers support standard Linux® distributions, making it easy to port existing codes to the platform. Preferred applications in the healthcare and life sciences sector—such as GATK, BWA, SAMtools, BLAST, MuTect2, and tranSMART—have already been enabled and optimized on IBM PowerLinux ™. Bioinformatics specialists within IBM Systems, along with IBM Business Partners in the healthcare and life science industry, are actively engaged in porting and optimizing the performance of additional biomedical research codes on POWER. More than one hundred open source applications are enabled on POWER8.

Within research fields relevant to Personalized Healthcare such as genomics and bioinformatics, adoption of GPU-enabled workloads has been slow; however, these workloads appear to be gaining traction. An increasing number of organizations conducting biomedical research are applying deep learning techniques to uncover predictive patterns within very large sets of complex, often unstructured data such as biomedical images and time-varying physiological signals. In support of such workloads, IBM and NVIDIA are collaborating on PowerAI, a new deep learning toolkit. PowerAI is an easy-to-deploy deep learning platform that delivers popular deep learning frameworks—including Caffe, Torch, and Theano—within the IBM Power Architecture. IBM has optimized each of these deep learning software distributions to take advantage of the high bandwidth offered by the IBM POWER8 processor and NVIDIA NVLink interconnect. The toolkit also uses NVIDIA GPUDL libraries including cuDNN, cuBLAS and NCCL to deliver multi-GPU acceleration on IBM servers.

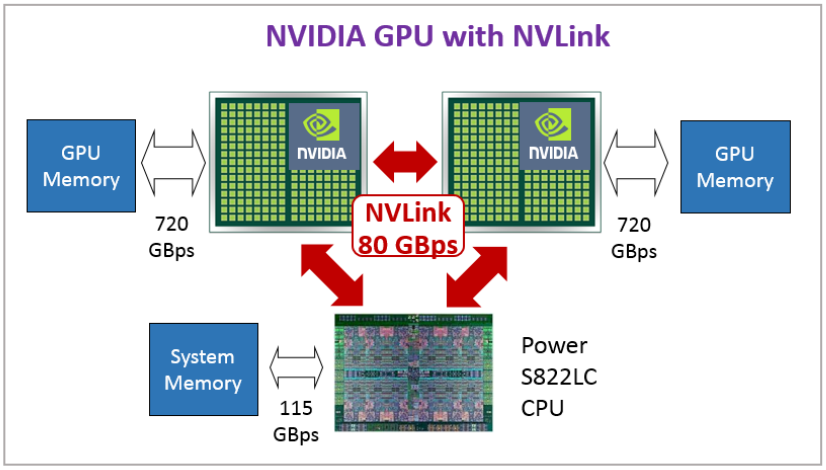

Figure 4. POWER8 with NVIDIA NVLink

An ecosystem of business partners in the healthcare and life science industry

Healthcare and life science industry subject matter experts—both within and outside of IBM—contribute to the enablement of computational algorithms, user applications, and data repositories that clinicians and research scientists might choose to run on IBM Systems infrastructure (in Figure 1; refer to the purple boxes in the IBM reference architecture). To meet full system requirements for customers pursuing personalized healthcare initiatives, IBM Systems has already formed business and development partnerships with healthcare and life science industry leaders, and some of them are detailed in this section.

IBM Watson: The IBM Watson™ division is dedicated to the development of cognitive systems that can discover repeatable and predictive patterns in even the most complex data—for example, natural language, environmental sounds, and visual images. Those patterns can then be used to impact lives in a positive way.

The sub-division, IBM Watson Health™, is responsible for developing healthcare and biomedical research applications that apply the core Watson capabilities of natural language processing and machine learning to the analysis of rich sources of personal health data including electronic medical records (EMR) and biomedical images.

Databiology: Databiology for Enterprise (DBE) is a scalable pan-omics information management platform that makes possible the secure, centralized management of omics data and analysis across globally-distributed projects. DBE provides data query, retrieval, and visualization tools to simplify the user experience. In support of compliance requirements for full data provenance and reproducible science, DBE provides automated data tracking and the ability to reprocess any analytics workload using original workload run data, software, and parameters. Databiology has integrated DBE with IBM Power Systems, Spectrum LSF, and Spectrum Scale to enhance workload, resource, and data lifecycle management in the cloud, on- and off-premises, and in hybrid models. DBE is also integrated with IBM Aspera software for secure, high-speed transfer of genomic data sets across the globe; and with IBM POWER computing environments. Refer to www.databiology.com

Lab7: The Enterprise Science Platform (ESP) from Lab7 gives small-to-midsize laboratory environments a flexible, scalable, centralized, and user-defined data management platform for continuously tracking laboratory samples, capturing data provenance, processing data, producing reports, and managing workflows. Lab7 has enabled ESP on IBM Power Systems, and also manages and continually updates the BioBuilds web service (refer to www.biobuilds.org), which provides turn-key deployment of and community support for open-source bioinformatics tools on IBM POWER. Tools supported on IBM POWER include multiple versions of Bowtie, BWA, Picard, Short Oligonucleotide Analysis Package (SOAP), Isaac, PLINK, Bioconductor, and many others. Refer to www.lab7.io

RowAnalytics. Synomics Studio from RowAnalytics provides an ultra-fast, highly scalable alternative to traditional genome-wide association studies. It enables large-scale data sets containing genomic and clinical data to be analyzed for associations between high-order genomic variant combinations and clinical outcomes. Synomics employs massively parallelized algorithms distributed across multiple GPU compute devices, and requires large amounts of very dense compute power. It has been shown to deliver superior performance on a compute architecture based on the IBM reference architecture for healthcare and life sciences, which includes IBM Power Systems, IBM Spectrum LSF, and IBM Spectrum Scale. Refer to www.rowanalytics.com

tranSMART: TranSMART is an open-source data warehouse and knowledge management system for integrating, accessing, analyzing, and sharing clinical, genomic, and gene expression data on large patient populations. Originally, based on the data model for National Institutes of Health (NIH)-sponsored and internationally recognized Informatics for Integrating Biology and the Bedside (i2b2); refer to www.i2b2.org, tranSMART has been used widely within the pharmaceutical industry and in large-scale public and private translational research initiatives. Since 2015, IBM Systems Group has been collaborating with the Data Science Institute of Imperial College in London to investigate the application of new compute and storage architectures that can improve the performance and scalability of the tranSMART platform. The application of tranSMART on IBM POWER8 with Elastic Storage Server has thus far yielded clear performance improvements relative to Intel®-based systems during data ingestion and analysis. Refer to www.transmartfoundation.org

WhamTech. SmartData Fabric (SDF) from WhamTech is a distributed data management layer that integrates into existing IT infrastructures and can support highly secure data virtualization, integration, federation, and analytics across heterogeneous data silos common to healthcare organizations. Regardless of the type, format, quality or structure of source data—whether they include big data, NoSQL databases, relational databases, files, office documents, email, or Internet of Things (IoT) devices—SDF users can access indexed source data without copying or moving that data from one location to another. Queries of data can be performed against indexed data wherever they reside using high-performance parallel processing with support for extreme scalability. WhamTech is engaged in a number of healthcare projects in the US, UK, and Australia, and is currently enabling their software on IBM POWER8 Systems and the IBM reference architecture for healthcare and life sciences. Refer to www.whamtech.com

Summary

The IBM reference architecture for healthcare and life sciences consists of key infrastructure components from IBM's high-performance compute and storage portfolio, and it supports an expanding ecosystem of leading industry partners. The reference architecture defines a highly flexible, scalable, and cost-effective platform for accessing, managing, storing, sharing, integrating, and analyzing big data within the constraints of limited IT budgets. IT organizations can use the reference architecture as a high-level guide for overcoming data management challenges and processing bottlenecks frequently encountered in personalized healthcare initiatives and other compute- and data-intensive biomedical workloads.

Get more information

To learn more about high-performance compute and storage offerings that comprise the IBM reference architecture for healthcare and life sciences, contact your IBM representative or IBM Business Partner, or visit the following websites:

- IBM Spectrum Computing
  http://www.ibm.com/systems/spectrum-computing/
- IBM Spectrum Scale
  http://www.ibm.com/systems/storage/spectrum/scale/
- IBM Power Systems
  http://www.ibm.com/systems/power/
- IBM SoftLayer
  http://www.softlayer.com/
- IBM Cloud Object Storage
  https://www.ibm.com/cloud-computing/products/storage/object-storage/cloud/
- IBM Aspera
  https://www.ibm.com/software/info/aspera/

About the authors

Jane Yu, MD, PhD is the worldwide industry architect for healthcare and life sciences within the IBM Systems Group. With more than 25 years of experience spanning clinical medicine, biomedical research, advanced analytics, systems engineering, enterprise IT, and management consulting, Dr. Yu works closely with IBM technical specialists, industry partners, and international clients to deliver best-in-class, high-performance compute and storage solutions for healthcare and biomedical research applications. She specializes in the support of organizations pursuing innovative programs in personalized healthcare and translational medicine.

Kathy Tzeng, PhD is the worldwide technical lead for life science and genomics solutions within the IBM Systems Group. Since joining IBM in 2001, Dr. Tzeng has been working with technical teams across IBM, industry partners, and open source communities on the enablement and performance optimization of life science applications on IBM solutions. She has published patents, IBM Redbooks®, technical papers, book chapters, and peer-reviewed journals.

Janis Landry-Lane is the worldwide sales executive for healthcare and life sciences within the IBM Systems Group. With more than 20 years of experience in high-performance computing, her work has spanned numerous research and development clients. Janis works closely with a multi-disciplinary team of internal IBM and external partners to address customer technical requirements for high-performance IT systems. Her team delivers scalable, extensible solutions for high-performance storage and compute—solutions which are designed to grow within on-premises environments and integrate with cloud environments, as needed.

-

IBM Systems
3039 Cornwallis Road
RTP, NC 27709

Produced in the United States of America

References in the publication to IBM products or services do not imply that IBM intends to make them available in all countries in the IBM operates.

♲ Please recycle