Beyond Genomics: The Role of a Software Defined Infrastructure in Healthcare

Sponsored by IBM

Srini Chari, Ph.D., MBA mailto:info@cabotpartners.com

Gordon Sissons, Cabot Partners Analyst

March, 2016

Executive Summary

Life sciences research is advancing at a rapid pace and new techniques such as next-generation sequencing (NGS), are playing a vital role in growing scientific knowledge, facilitating the development of targeted drugs, and delivering personalized healthcare. By investigating the human genome, and studying it in the context of biological pathways and environmental factors, it's now possible for scientists and clinicians to identify individuals at risk of disease, provide early diagnoses, and recommend effective treatments.

While high-performance computing (HPC) environments were previously deployed mainly in research, genomics is edging ever closer to the patient and front-line clinical care. New sequencing and analysis techniques, a greater emphasis on collaboration, and the application of new technologies like big data and cognitive computing are resulting in a re-think of how computing and storage infrastructure are deployed. While HPC has an important role to play, application requirements now extend beyond traditional HPC and including many analytic components as well.

In this paper, aimed at IT professionals and bioinformaticians, we review some of the applications of high-performance computing and analytics in healthcare, and describe the software and workloads typically deployed. We then get specific, explaining how an IBM software-defined infrastructure can provide a more capable and efficient platform for the variety of applications and frameworks being deployed in healthcare institutions today.

Population level genomics studies can involve tens of petabytes of data with storage requirements doubling every five to twelve months

The explosive growth of data

Modern biology represents an enormous data management challenge. Full genome sequences from next-generation sequencers can now provide raw data on all three billion base pairs in the human genome, representing terabytes of data that needs to be analyzed and compared to genomic databases. In cancer research, the whole genome of a tumor and a matching normal tissue sample consumes about one terabyte of data. Large population-level studies can easily involve tens of petabytes of data, with storage requirements frequently doubling every five to twelve months.

In addition to the size of the data, the sheer number of generated files presents unique challenges. In a typical genomics workflow for DNA analysis, the first step involves concatenating and de-multiplexing raw data coming from a sequencer. A DNA exome sequence may generate 20 to 40 million discrete reads per subject. One IBM customer, Mt. Sinai, has published a paper[1] sharing details of their genomics environment. At Mt. Sinai, the genomics core facility (GCF) generates approximately 6 TB of raw data per week, and a further 14 TB of data after analysis and post-processing, for a total of 20 TB of new data per week. At the time of this writing, Mt Sinai manage approximately 11 petabytes of raw data shared between cluster nodes in their HPC center, and store an additional 3 petabytes in their Hadoop-based data science environment.

In clinical environments, storage requirements also include rich datasets such as MRIs and ultrasound images and increasingly data from new sources, such as telemetry from remote heart monitors, FITBIT® data, and other sensors used together with genomic data for diagnosis. Complicating the data management challenge even more, hospitals need to retain most of this data for future comparative analysis and patient care.

A range of new applications

The NHGRI, a branch of the National Institute of Health (NIH), has defined five steps for genomic medicine[2] shown in Figure 1. These steps have become a common framework for viewing the range of applications in modern healthcare.

Understanding the genome (pictured on the left in Figure 1) requires the application of RNA and DNA sequencing methods to assemble genomes from reads, analyzing the function of human genomes, and studying variants. Translational medicine includes a variety of disciplines aimed at understanding the biology of genomes, the biology of disease, and translating learnings to help advance the science of medicine and discover new treatments. Personalized medicine involves applying advances in our understanding of the genome to individual patients for a variety of diagnostic or preventive treatments.

[1] **Mt. Sinai accelerates genetics research and medicine – Mt. Sinai paper** http://ibm.biz/BdH3W3
[2] **Source Eric D. Green, Mark S. Guyer et al. and the National Human Genome Research Institute - Nature 470, 204-213.** http://www.nature.com/nature/journal/v470/n7333/abs/nature09764.html

Choosing the right course of care involves understanding a patient's genetic make-up and a variety of external factors

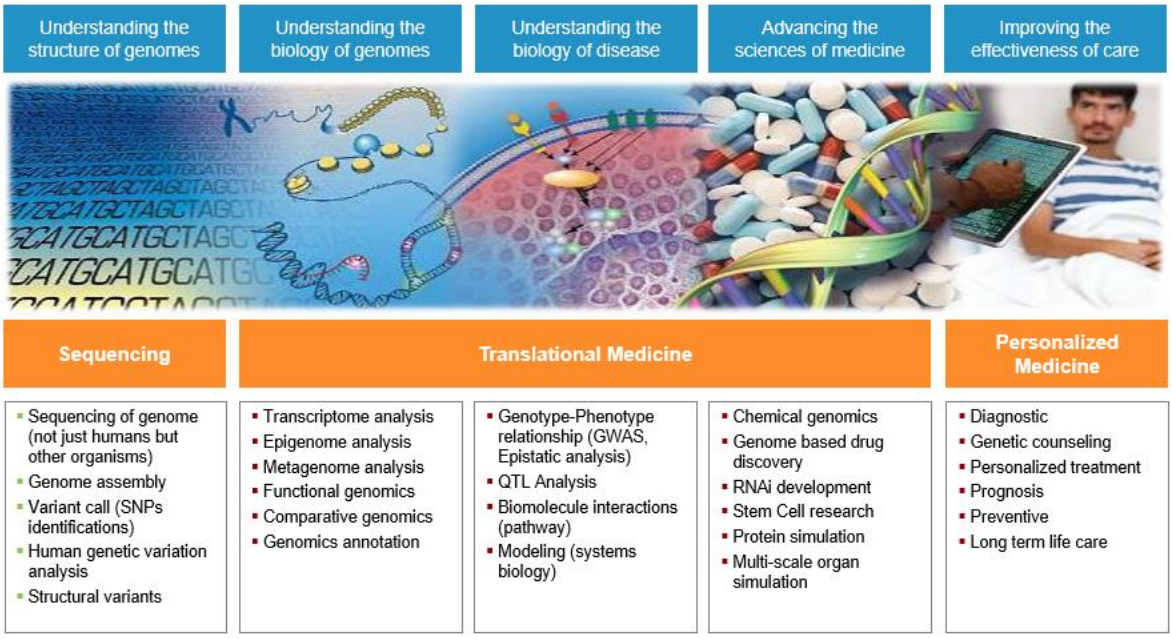


Figure 1 - The Five steps for Genomic Medicine (NHGRI)

IBM's Reference Architecture for Genomics provides a high-level blueprint for IT architects

It's easy to see how these fields of research are interconnected with clinical care. The discovery of new drugs depends on advances in the underlying science. Understanding the efficacy of treatments, and the best care for an individual patient, involves not only understanding a patient's unique genetic make-up, but a variety of external and environmental factors as well.

IBM's Reference Architecture for Genomics provides a framework for understanding the essential infrastructure requirements for supporting genomics applications.

Genomics – Applying DNA sequencing methods and bioinformatics to assemble and analyze the function and structure of genomes

Translational medicine – Improving medical outcomes by translating findings in genomics research into improved diagnostic tools, medicines, procedures and policies

Personalized medicine – A medical model involving optimized medical treatments tailored to individual patients based on genetics and other considerations

IBM's Reference Architecture for Genomics, published as an IBM Redpaper[3], discussed shortly, provides a high-level blueprint useful to IT architects interested in understanding infrastructure requirements for these environments.

[3] http://www.redbooks.ibm.com/abstracts/redp5210.html

A unique challenge in genomics is the sheer number of diverse tools

Diverse application workloads

A unique challenge with life sciences and clinical applications is the sheer diversity of tools. Many of the major platforms used for genome analysis are in fact toolkits, often comprised of dozens or even hundreds of discrete tools and components.

ABySS	BLAST (NCBI)	EMBOSS	IGV	PLINK	STAR-fusion
Accelrys	Bowtie	FASTA	Infernal	Pysam	Stratoolkit
ALLPATHS-LG	Bowtie2	FastQC	ISAAC	Python	Tabix
Appistry	BWA	FASTX-Toolkit	iRODS	RSEM	TMAP
Arpeggi	Casava	GALAXY	Lab7 ESP	Sailfish	TopHat
Bamtools	CEGMA	GATK	MAQ	SAMTools	TopHat 2
BarraCUDA	Celera_Assembler	Genestack	MIRA	SHRiMP	Trinity
Bedtools	CLCBio	GenomicConsensus	Mothur	SIFT	T-Coffee
Bfast	ClustalW	GraphViz	MUSCLE	SOAP3-DP	Variant tools
Bioconductor	Conda	HMMER	Novoalign	SOAPaligner	Velvet/Oases
BioPerl	Cufflinks	HTSeq	Numpy	SOAPdenovo	R
BioPython	DIALIGN-TX	Htslib	PICARD	SQLite	RNAStar/STAR

Figure 2 - A sampling of popular analysis tools and frameworks

How these toolkits are implemented and used can vary between facilities. There are several commonly used frameworks for accessing, orchestrating and managing tools and data. As some examples:

- The Broad Best Practices pipeline including the Genome Analysis Toolkit (GATK)[4], is a software package for the analysis of sequencing data developed by the Broad Institute. It's comprised of dozens of discrete tools[5] organized in categories like diagnostics and quality control, sequence data processing, variant discovery etc.
- Galaxy (https://galaxyproject.org/) is a web-based platform and framework for data-intensive bio-medical research developed at Penn State University and John Hopkins University. The platform is comprised of tools, data managers, custom datatypes and workflows. Participating institutions can use tools across public or private instances of Galaxy and are encouraged to develop their own tools and publish them back to the Galaxy framework.
- SOAPdenovo[6], developed at Beijing Genomics Institute, is a short-read assembly method that can build a de novo draft assembly for the human-sized genomes. The program is specially designed to assemble Illumina GA short reads. It creates new opportunities for carrying out accurate analyses of unexplored genomes in a cost effective way.
- Gene Pattern (http://genepattern.org) is another freely available computational biology package developed at the Broad Institute for the analysis of genomic data. It is essentially a scientific workflow system comprised of 220 discrete genomic analysis tools supporting data processing, gene expression analysis, proteomics and more.

While the science is certainly complicated enough, it's also easy to underestimate the challenge of supporting all of these diverse tools in a shared environment.

[4] **Broad Institute GATK - https://www.broadinstitute.org/gatk/**
[5] **https://www.broadinstitute.org/gatk/guide/tooldocs/**
[6] **SOAPdenovo - http://soap.genomics.org.cn/soapdenovo.html**

Software that manages the genomic pipeline alone is not sufficient - It needs to be supplemented with process and workload management tools

- Tools can have widely varying resource requirements in terms of CPUs, cores, memory, scratch space, and I/O requirements, and in some cases may require specialized hardware to run. As examples, NVIDIA® GPUs, Intel® Xeon Phi™ co-processors, or increasingly the use of purpose-build co-processors for life sciences such as the DRAGEN™ Bio-IT processor.

- Tools increasingly depend on underlying software frameworks. While some toolsets are self-contained, others have dependencies relying on specific MPI variants, MapReduce or HBASE implementations, or frameworks including Spark, Solr and various other middleware components.

- Tools can exhibit highly variable data access patterns. A job array comprised of thousands of concurrently executing job elements may simultaneously access a directory on a shared filesystem containing millions of files, each under 100 bytes in size. At the opposite extreme, a MapReduce application may be used to ingest, and process large, high-resolution imagery where each file is several gigabytes in size.

While the software frameworks described above can be used to orchestrate a high-level genomics pipeline, a detail sometimes missed is that these tools alone cannot solve the workload management challenge in production environments. To make an environment practical to manage, these frameworks need to be supplemented with workload scheduling tools that are both workload and resource aware and able to enforce a wide variety of policies.

As an example, when a 10,000 element job-array is 60% completed, I may wish to launch a new job to begin staging data for the next job in the pipeline sequence, or a workflow may need to make a decision at run-time based on network congestion and cluster load to determine where to execute the next step in the pipeline. Site administrators may implement a resource sharing policy, so that workloads, excluding time-critical jobs, are provided resources in proportion to departmental funding levels. Administrators often need to implement resource usage policies that vary by time of day, allowing the system to reserve resources for long-running jobs during off-peak hours. While the pipeline may be controlled using a high-level tool like Galaxy, efficient use of compute and storage resources will require sophisticated workload management tools.

IBM's Reference Architecture for Genomics

To address the needs of speed, scale and smarts in genomic medicine, IBM has created an end-to-end reference architecture that defines the most critical components for genomics computing[7] illustrated in Figure 3. The reference architecture, provided for IT practitioners, and to a lesser extent medical personnel interested in the underlying technology, provides an excellent way to view how application frameworks commonly used in personalized, translational medicine and sequencing interact with the IT infrastructure. While Figure 3 depicts Genomics tools, this is just one entry point into the reference architecture, and different workflows and tools apply in disciplines like imaging, cylometry and proteomics.

[7] http://www.redbooks.ibm.com/abstracts/redp5210.html

IBM's reference architecture defines capabilities around data management, workload orchestration and access to applications

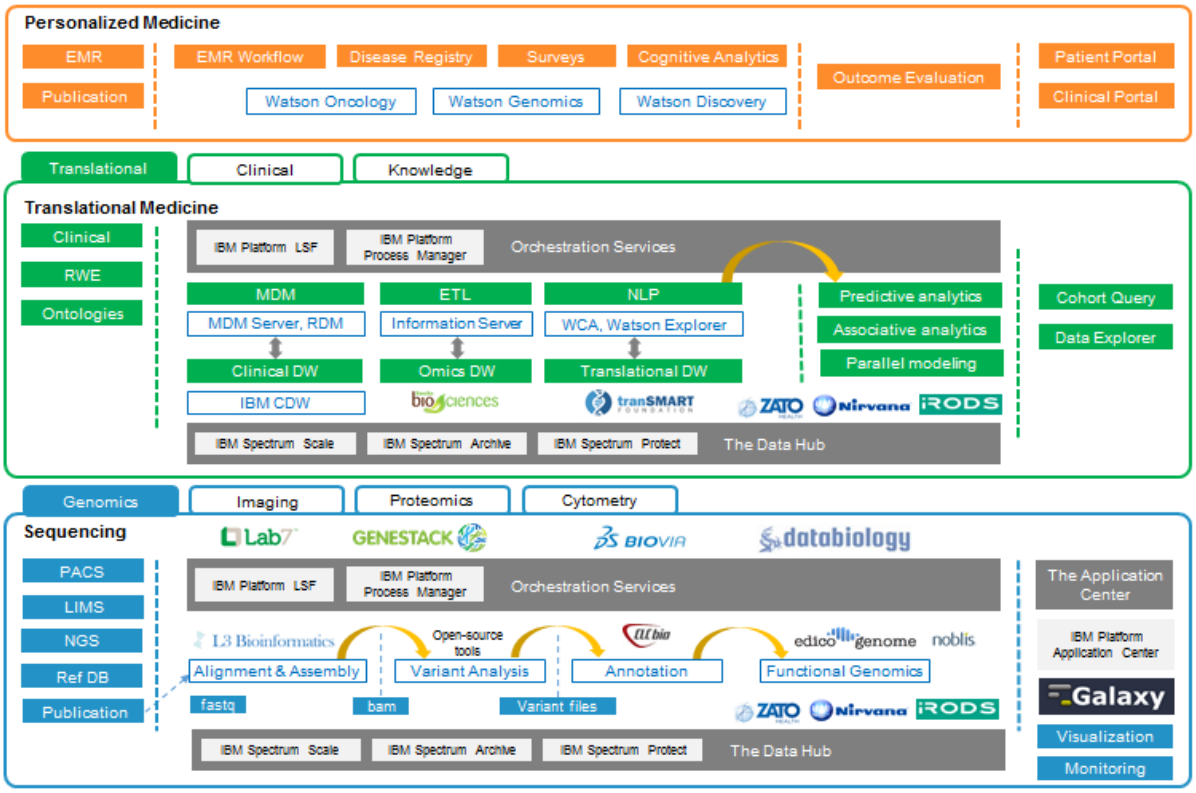


Figure 3 - IBM Reference Architecture for Genomics

IBM's reference architecture defines key capabilities around data management (referred to as *the data hub*) the need for workload orchestration (referred to as *orchestration services*), and secure enterprise access to applications (facilitated via *the application center*).

As Figure 3 illustrates, these architecture components can be shared across the different fields of medicine. Using a common infrastructure for data management, workload and resource orchestration helps organization realize economies of scale and reduce infrastructure and management costs.

The genomics reference architecture adheres to three guiding principles. It needs to be:

1. **Software-defined** – meaning the infrastructure employs software-based abstraction layers for compute, storage and cloud to help future-proof the environment as applications change and requirements inevitably grow
2. **Data-centric** – meaning that the infrastructure is designed to accommodate the explosive growth expected for genomics, imaging and clinical data, to minimize data movement, and manage this data over its life cycle
3. **Application-ready** – meaning that the software-defined environment provides support for a variety of applications including support for data management, version control, workload management, workflow orchestration and access for execution and monitoring

A software defined infrastructure helps ensure that the right workloads are run on the right infrastructure

A software-defined infrastructure for healthcare

A software-defined infrastructure helps transform static IT infrastructure into a dynamic resource, workload and data-aware environment. Application workloads are serviced automatically by the most appropriate resource running locally, in the cloud or in a hybrid cloud environment. Software Defined Environments abstract computing and storage, ensuring that application service-levels are met, and that infrastructure is used as efficiently as possible. The abstraction of compute resources in IBM's software-defined infrastructure is facilitated in large part by IBM's Platform Computing family of products, including IBM Platform LSF®, production proven running some of the world's largest clusters comprised of various physical, virtual and cloud-based assets.

Complementing the software-defined computing capabilities, IBM Spectrum Storage solutions provide a software-defined storage environment designed to accelerate data access, simplify and speed storage management, and scale with data anywhere. IBM's software-defined storage provides intelligent tiering of data and supports open APIs.

Running the right workload on the right hardware

As explained earlier, a key challenge with workloads in healthcare is the diversity of applications. Different types of applications will run better on different types of infrastructure as illustrated in Figure 4. Some applications will benefit from large-memory SMP environments, while others will run most efficiently across distributed, shared nothing environments.

Some workloads will be more cost-efficient to run on IBM OpenPOWER™ systems owing to its capacity to process more threads and its superior memory bandwidth. Other workloads may be suited to Intel-based systems. Hadoop or Spark workloads may favor one type of node in distributed environments while CPU intensive applications may favor different types of nodes. Characteristics like CPU, clock speed, OS type and version, swap space, storage characteristics or the presence of particular libraries or layered software can all be considerations in selecting the right resources. Applications with very large I/O requirements may benefit from being placed on nodes connected to a parallel file system which can provide better I/O throughput than even local disks. The point of a software-defined infrastructure is to unite all of these resources types into a shared pool, and provide mechanisms to place the right job on the right hardware at runtime.

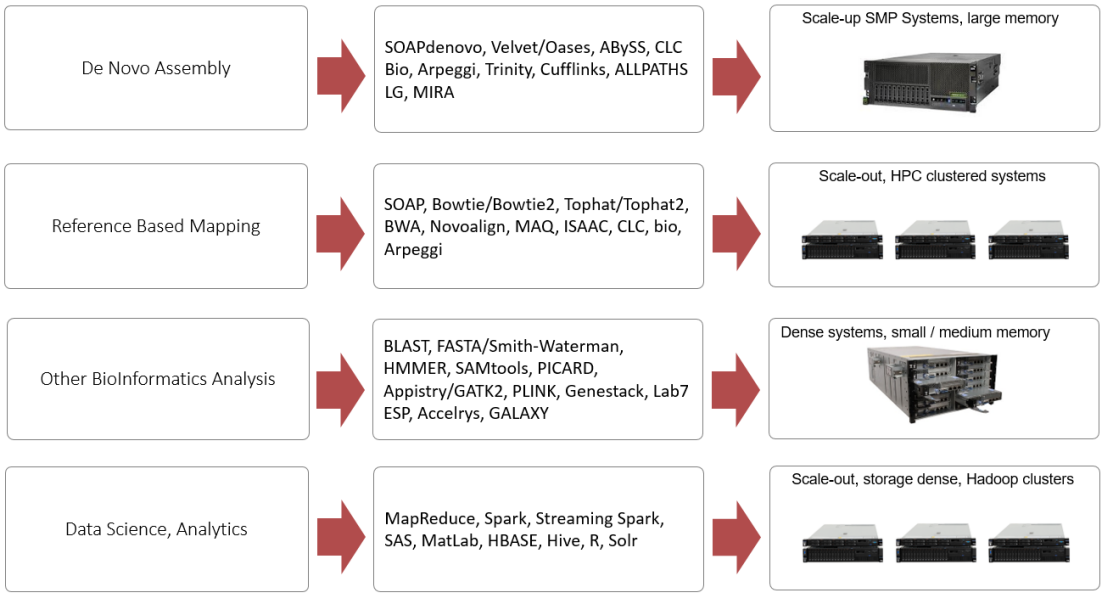


Figure 4 - Different workloads demand different types of infrastructure

Bottlenecks in scheduling can result in serious and costly underutilization of resources

The critical role of workload scheduling

Workload scheduling may sound like a mundane topic, but as clusters get large, scheduling becomes increasingly important. Readers can be forgiven if they've not given scheduling a lot of thought, so we provide a simple thought experiment to illustrate its critical role.

EXAMPLE: Understanding how scheduling affects the performance of Genomics workloads

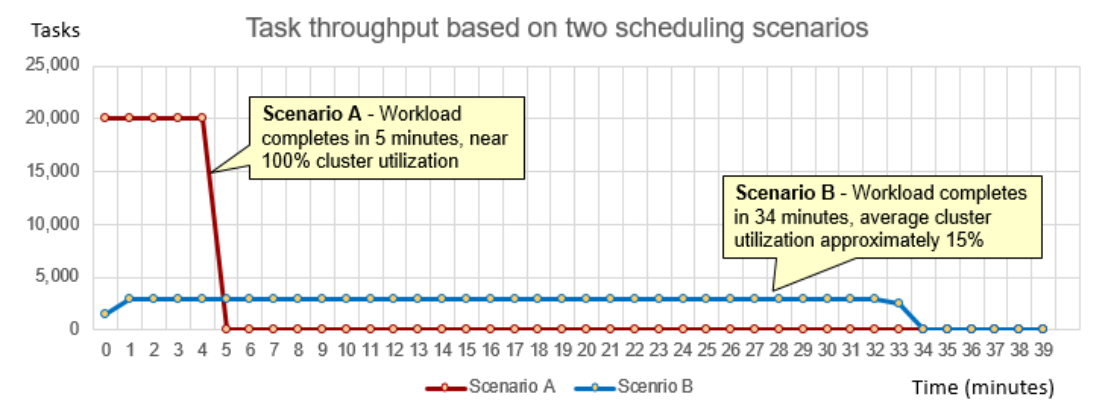
Consider a cluster comprised of 1,000 nodes, each with 20 processor cores for a total of 20,000 cores. Now consider that we wish to run a genomic analysis job comprised of 100,000 sub-elements, where each job element will execute for 60 seconds on a single core. Let's further assume that our scheduling engine can only make 50 scheduling decisions per second.

Scenario A: Assuming no scheduling bottlenecks

Assuming our cluster can be fully utilized without the scheduler introducing any delays, we have the capacity to complete 20,000 jobs per minute (all cores running a one-minute task concurrently). In theory, we should be able to complete all 100,000 jobs in five minutes. In this scenario we see a run-time of **5 minutes** for 100,000 tasks and essentially **100% cluster utilization**.

Scenario B: Factoring scheduler limitations

Recall however that our hypothetical workload manager can only dispatch 50 jobs per second. In this case, the cluster can only dispatch 50 * 60 tasks per minute or 3,000 tasks. Since tasks are completing every 60 seconds, we can never use more than 3,000 of our 20,000 available cores, limiting our utilization to approximately 15%. This means that 85% of our cluster resources sit idle due to the scheduler bottleneck. Also, because of the bottleneck, dispatching all the jobs to cluster resources takes approximately 33 minutes (100,000 tasks / 3,000 tasks per minute). This results in a run-time of approximately **34 minutes** and **15% cluster utilization** (The extra minute the result of a tail effect as previously submitted jobs finish).



Scheduling decisions can be complex, taking into account a myriad of factors such as resource requirements, dynamic loads on servers, thresholds, time windows, sharing policies and more. The example above may seem extreme, but it illustrates a real problem all too familiar to cluster administrators. Bottlenecks in production can easily result in serious and costly under-utilization as described above.

Life sciences environments supporting many users can deal with hundreds of thousands of jobs, making them particularly susceptible to these kinds of challenges. While most schedulers can keep pace with workloads on small and medium-sized clusters, keeping large clusters fully utilized is a much bigger challenge.

IBM Platform LSF has been benchmarked scheduling 5.7 million jobs per hour making it well-suited to Life Sciences workloads

One of the reasons that IBM Platform LSF is particularly well-suited to life sciences workloads is that it exhibits faster scheduling performance than competitors, even while implementing complex resource sharing and job placement policies.

A comparative study of HPC Workload Management Tools published by the Edison Group[8] determined that IBM Platform LSF exhibited significantly better performance than other schedulers evaluated. In the most recent Edison Group study, the latest version of IBM Platform LSF (9.1.3.3 at the time of this writing) was able to schedule 5.7 million jobs per hour or approximately 1,583 jobs per second. Fast scheduling throughput helps ensure that clusters can be kept close to full utilization, helping avoid infrastructure getting in the way of the work of researchers and clinicians.

Getting specific on the IT challenges

When it comes to managing complex environments, there are a variety of IT challenges where a software-defined infrastructure can help.

- **Complex, fast-evolving software frameworks** – Research and clinical environments need to support a variety of application frameworks. It would be cost-prohibitive to have dedicated hardware for each application, so having applications share resources is critical to reducing costs. IBM Platform LSF and the IBM Platform Application Service Controller can both help multiple frameworks share resources more effectively.
- **Abundant workflows, at multiple levels** – When we think of workflows in genomics, we tend to think of genomic pipelines orchestrated by toolsets like Galaxy. While there many tools that manage workflows, it is helpful to use tools that are workload and resource aware and that understand concepts like queues, application profiles, projects and user groups. IBM Platform Process Manager can augment or in some cases replace other workflow management tools, allows users to track workflows visually, and automate and manage flows and sub-flows, making them self-documenting and resilient.
- **Massive file and job counts, extreme I/O** – A single workflow may create a directory with between 0.5 to 1.2 million files, each containing just a few hundred bytes of data. These files can be organized into 10,000 or more folders, and analysis can involve 10,000 to 20,000 discrete jobs.[9] Making things even more challenging, these concurrently executing jobs will need to read and write files in a shared directory, generating in the range of 100,000 IOPS. This is just one workflow, and researchers may want to run many similar overlapping workflows in the course of a day. While Platform LSF can manage the job volume, keeping pace with I/O remains a challenge. IBM Spectrum Scale is a Software Defined Storage solution designed for this type of extreme I/O. It provides capabilities such as distributed metadata, so that I/O that would otherwise be concentrated on a single directory can be distributed across cluster nodes. It also supports the notion of sub-blocks so that small files can be stored and handled efficiently. Storage pools and storage tiering in Spectrum Scale allow I/O intensive operations to use the right type of media (in this case fast, solid-state drives) helping the storage subsystem keep pace with these massive I/O demands.
- **Multiple users and departments, wanting service level guarantees** – In a shared environment, there will often be contention for resources. For example, users in Oncology or Radiology may have analysis critical to patient care that needs to run quickly, while a research associate may want to run an analytic job for a population level study. IBM Platform LSF provides the tools necessary to support priority workloads and enforce sharing policies so that departments get their fair allocation of resources. IBM Platform LSF is unique in that it

[8] **See HPC Workload Management Tools: A Competitive Benchmark Study - https://www-01.ibm.com/marketing/iwm/iwm/web/signup.do?source=stg-web&S_PKG=ov42609**

[9] **Talk by Patricia Kovatch describing Mt Sinai Minerva systems at Usenix conference. https://www.usenix.org/sites/default/files/conference/protected-files/lisa15_slides_kovatch.pdf**

Where HIPAA compliance is an issue, IBM Spectrum Scale brings important capabilities

provides SLA-based scheduling, enabling the scheduler to consider directives around deadlines, throughput or velocity guarantees, helping ensure that business goals are achieved.

- **Privacy and the need to segment research data from clinical data** – For hospitals engaged in research, complying with HIPAA[10] privacy rules and the HITECH[11] act are important considerations. In clinical environments, a physician may find themselves having dual-roles – being responsible for a patient, while also contributing to medical research where they need to ensure that protected health information (PHI) is not inadvertently disclosed. While HIPAA compliance is a bigger topic, and IBM provides a number of supporting solutions, IBM's software-defined infrastructure components including IBM Spectrum Scale provides essential capabilities for compliance including authorization and access controls, data encryption, secure erasure, and logging and auditing facilities. IBM Spectrum Scale also supports immutability and append only controls to prevent log files from being tampered with or accidently deleted[12].
- **Insatiable demands for computing capacity** – When computing resources are shared in campus environments, demand can be very high. Individual jobs have the potential to tie up a cluster for days, so controlling usage and allocating resources fairly is critical. IBM Platform LSF has capabilities that enable administrators to provide users, groups or project teams with economic incentives to use capacity wisely, and to enforce compliance with site policies. For example, before jobs can consume resources, it can be a requirement that they are tagged with a valid project ID. Departments and projects may be provided with resource budgets. Submitting work to a low priority queue may be free, while submitting work to a high-priority queue may cost double the normal rate and have run-time limits enforced. Institutions may devise variable rate structures encouraging users to run jobs during non-peak hours. Additional incentives may be offered to encourage more precise estimates of job run-time and resource requirements. All of these capabilities play an important role in helping the shared environment run more efficiently to the benefit of all users.
- **Compliance with software licensing terms** – While many genomic tools are open-source, commercial software is used in clinical and research environments as well. This software can be expensive, and a limited number of licenses may be available depending on the tool. IBM Platform License Scheduler can help ensure that licenses are allocated optimally according to policy, and monitoring and analysis tools can help ensure that policies are having the desired result. For example, running applications using expensive license features on the fastest possible host to minimize license check-out time, or borrowing idle feature licenses from a remote cluster where licensing terms permit. License aware scheduling can help reduce contention, and avoid the need to purchase additional licenses unless absolutely necessary.
- **Dealing with diverse data formats** – As explained earlier, data formats and I/O patterns can vary depending on the nature of workloads. Also, different software frameworks may use entirely different storage technologies or access methods. For example, while traditional HPC environments have tended to use fast, parallel file systems accessed using POSIX semantics, most Hadoop environments employ HDFS (the Hadoop Distributed File System) a distributed file system written in Java. Other applications involving data sets like video or images may require an object store.

  While Hadoop is not widely used in Genomics, it is common in many analytic applications. In areas like translational medicine, it is fair to say that most institutions will find themselves needing to support all of these access methods. A benefit of IBM Spectrum Scale is that it can support all of these access methods concurrently, avoiding the need to replicate data in and out of HDFS and stage data.

[10] **HIPAA – Health Insurance Portability and Accountability Act - https://www.gpo.gov/fdsys/pkg/PLAW-104publ191/content-detail.html?cm_mc_uid=23192118585914495020065&cm_mc_sid_50200000=1454505536**
[11] **Details on HITECH are included as a part of this legislation - https://www.gpo.gov/fdsys/pkg/PLAW-111publ5/content-detail.html?cm_mc_uid=23192118585914495020065&cm_mc_sid_50200000=1454505536**
[12] **Details on Spectrum scale features for HIPAA compliance - http://www.ibm.com/developerworks/aix/library/au-gpfs/index1.html**

Tools like IBM Spectrum Archive allow for automated movement of data to lower cost media

For example, a non-Hadoop genomics application can write a large file natively to the distributed file system using POSIX methods, and a Hadoop workload can then access that same file directly using HDFS methods. Database tables can reside on Spectrum Scale file systems. Other applications can see Spectrum Scale as an object store, and access it via an easy to use RESTful API. With IBM's acquisition of Cleversafe, IBM can offer customers a range of object storage solutions. This flexibility to support multiple access methods on the same storage foundation avoids the need for redundant copies of data, and multiple siloed storage subsystems.

- **Data retention and archival** - As explained earlier, storage requirements are growing quickly for a variety of reasons. In some cases, organizations will choose to retain data simply because they expect they may need it in future, but in other cases data retention will required by regulations such as HIPAA requiring that covered entities retain data for at least six years[13]. Solutions like IBM Spectrum Archive, part of IBM's software-defined storage portfolio allow for transparent and automated movement of infrequently used data to lower cost media like tape without the need for complex or proprietary archival applications. When archived data is referenced via the file system, it can be restored automatically to near-line storage.
- **The need to share data and collaborate across centers** – Translating research into insights that contribute to patient care requires extensive collaboration and data sharing. While some datasets are small enough to be accessed through a web portal, others are enormous, and can be time consuming and costly to move between local centers and/or public and private clouds. To share large datasets, customers can take advantage of capabilities like active file management (AFM) in IBM Spectrum Scale to reliably accelerate access to large datasets across distributed environments. IBM Platform Data Manager can be used together with IBM Spectrum Scale and a variety of tools for data movement including IBM Aspera®, efficiently moving data between centers considering bandwidth, latency constraints, job deadlines, costs and a variety of other factors.

Applications beyond genomics

At the 2015 International SuperComputing Conference in Frankfurt, Germany, IBM working with Imperial College London Data Science Institute, demonstrated how large-scale genomics, coupled with analytics and cognitive computing can provide a powerful tool for physicians in the area of personalized medicine[14]. The demonstration involved tranSMART, an open-source data warehouse and knowledge management system, and IBM Watson, IBM's jeopardy winning computer (Figure 5) running on IBM OpenPOWER Systems.

tranSMART has been widely adopted by the pharmaceutical industry and public-private initiatives, for integrating, accessing, analyzing, and sharing clinical, genomic, and gene expression data on large patient populations. By facilitating discovery of associations among data items, tranSMART enables biomedical researchers to formulate and refine hypotheses on molecular diseases mechanisms. Such hypotheses can serve as the basis for further evaluation in the context of published scientific literature.
The demonstration showcased how the application of new computing technologies, OpenPOWER technology, and better storage architectures can improve the performance and usefulness of the tranSMART platform.

[13] **What you need to know for retention compliance - http://privacyguidance.com/blog/what-you-need-to-know-for-retention-compliance/**
[14] **http://openpowerfoundation.org/blogs/imperial-college-london-and-ibm-join-forces-to-accelerate-personalized-medicine-research-within-the-openpower-ecosystem/**

A demonstration involving tranSMART shows how a software defined infrastructure can improve patient care

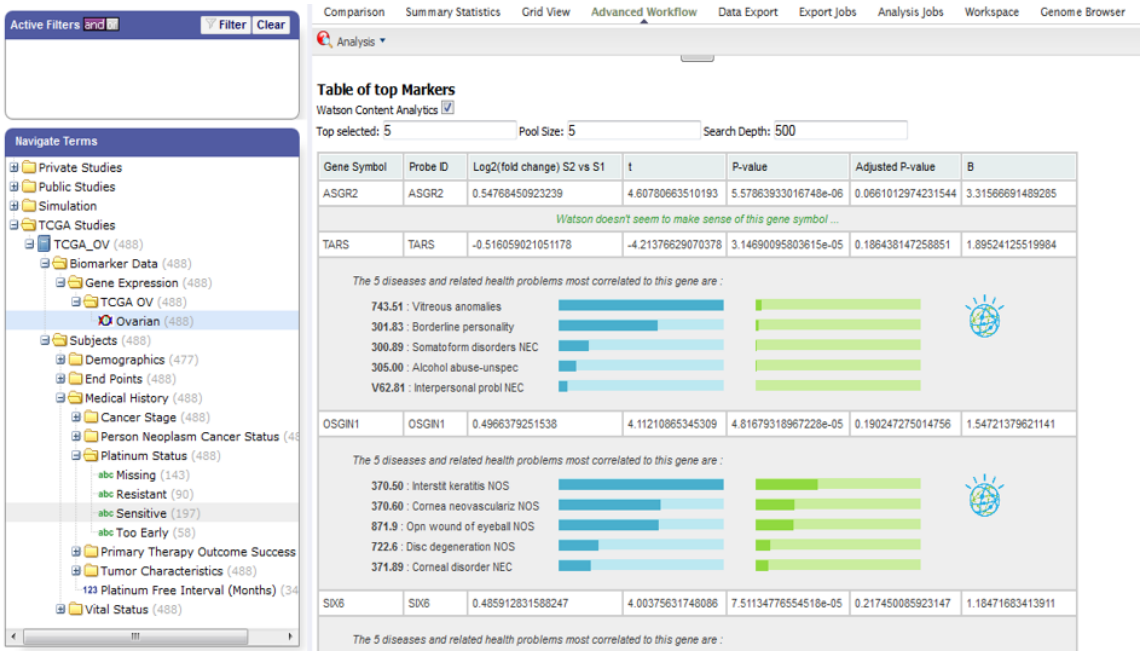


Figure 5 - IBM Watson working with tranSMART on OpenPOWER

In addition to providing a better performing infrastructure for tranSMART, the demonstration illustrated how advances in cognitive computing (IBM Watson), machine learning (ML), natural language processing (NLP), along with data federation can help support better patient care. The demonstration showed how clinicians might diagnose a patient's condition faster and more accurately based on a patient's history, population level studies, and an individual patient's symptoms and genetic characteristics.

A shared infrastructure for personalized medicine

The software environment behind the tranSMART environment described above is illustrated in Figure 6. The application architecture illustrates how a software-defined infrastructure is not only beneficial in genomics, but for applications in translational and personalized medicine as well.

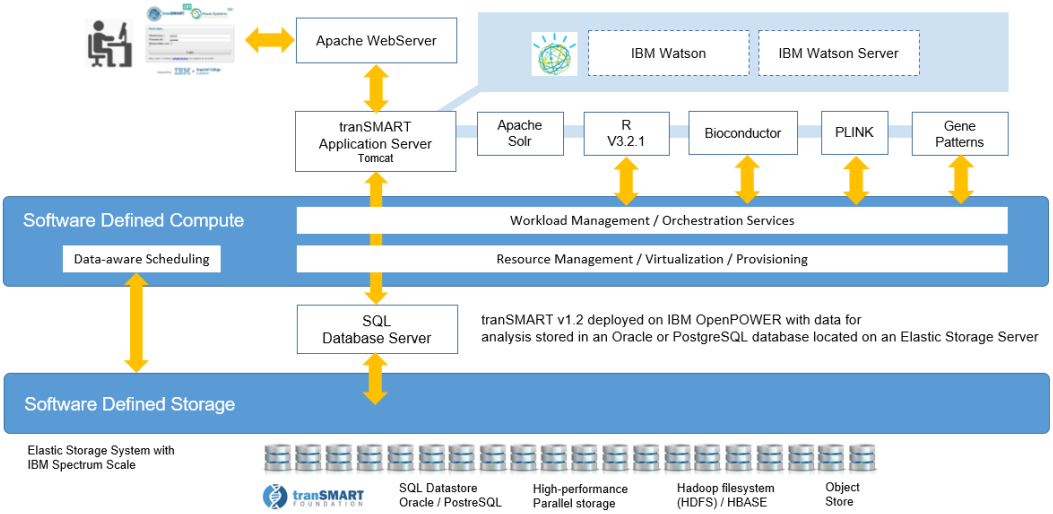


Figure 6 - The ecosystem of applications in a TranSMART deployment

Applications surrounding tranSMART all benefit from a software defined infrastructure

Looking at Figure 6, a reader could be forgiven for thinking of each of the application frameworks interacting with tranSMART as simple "black boxes", but each of these components (Solr, R, Bioconductor) are frameworks unto themselves. To scale, and run efficiently, each component runs across a distributed environment as a set of services. Each of the components can benefit from an IBM software-defined infrastructure as explained below.

- **R** is an open source statistical modeling language used in a variety of fields including statistical genetics and clinical trial design. Hundreds of R based tools for genomics and clinical trials are available as part of the Comprehensive R Archive Network (CRAN)[15]. Many of these packages are resource intensive, and parallel frameworks written in R including R BatchJobs and Rmpi are designed to utilize workload schedulers like IBM Platform LSF to run these workloads efficiently. Parallel job management features such as back-fill scheduling in IBM Platform LSF are particularly useful in sharing resources between short-duration genomics jobs, and longer-running parallel simulations using Rmpi. Platform LSF features like Advanced Reservations and SLA scheduling are essential in ensuring that large-parallel jobs running on busy, shared computing do not interfere with workloads supporting important clinical applications.

- **Apache Solr** is a distributed framework (based on Apache Lucene) that is used for fast, distributed searches of large datasets, including genomic data. To be efficient sharing resources, distributed frameworks like MapReduce, Spark, Solr, HBASE and even IBM Platform LSF all need to co-exist on the same infrastructure and share resources according to policy. Depending on site requirements, IBM Platform Application Service Controller or the IBM Platform Conductor for Spark can help these different software frameworks coexist and share a common pool of resources.

- **BioConductor,** available from http://bioconductor.org, provides tools for the analysis and comprehension of high-throughput genomic data. BioConductor is based on the open-source R language described above. In large environments, BioConductor jobs are normally submitted using the R language batch facility (R CMD BATCH) or via R Studio supported by IBM Platform LSF or other workload managers. Some steps in BioConductor workflows (such as the h5vc tool used to tally nucleotides using HDF5) benefit greatly from parallelism, and BatchJobs in concert with IBM Platform LSF facilitates parallel execution needed to speed analysis.[16]

- **PLINK** is an open-source whole genome association analysis toolkit designed to perform a range of large-scale analysis. It supports high-performance linear algebra math libraries (often deployed in HPC environments) to speed analysis. PLINK (and many similar tools) are used with IBM Platform LSF to speed analysis and share resources among applications according to policy.[17] PLINK and similar workloads often have short run-times, so scheduling efficiency and latency is essential to delivering good application performance.

- **Gene Pattern,** from http://genepattern.org, is a freely available computational biology package developed at the Broad Institute for the analysis of genomic data. It is essentially a scientific workflow system comprised of 220 discrete genomic analysis tools supporting data processing, gene expression analysis, proteomics and more. To manage these workflows, often comprised of thousands of discrete jobs, Gene Pattern supports an integration with IBM Platform LSF[18] to ensure that individual workloads run reliably on appropriate computing resources and that infrastructure is optimally used and shared with other competing software frameworks.

As we've shown above, essentially every component of the application ecosystem around tranSMART, benefits from a software-defined infrastructure. While the nature of the workloads differs, the same scalable, virtualized, distributed software infrastructure used in genomics research is important for applications in translational and personalized medicine as well.

[15] **A list of R projects specific to statistical genetics - https://cran.rstudio.com/web/views/Genetics.html)**
[16] **Details on how h5vc parallelizes tallies using tallyRangesBatch function to interface with R BatchJobs and IBM Platform LSF**
[17] **Examples of using IBM Platform LSF running PLINK and other tools**
[18] **Broad Institute GenePattern – IBM Platform LSF integration http://www.broadinstitute.org/cancer/software/genepattern/administrators-guide#using-a-queuing-system**

Summary

The opportunities in modern healthcare are immense. Rapid advances in genomics, and the ability to exploit these advances with new computing technologies and analytic methods promise to revolutionize how we treat disease.

Important to realizing this vision are more sophisticated, software-defined infrastructures able to support the full spectrum of scientific and clinical applications on-premises or in the cloud.

Software-defined infrastructures offer opportunities to realize many advantages including:

- Running research and clinical workloads faster, reducing cycle times, and providing better diagnostics and analysis in support of patient care
- Using resources more efficiently, enabling institutions to do more with less, reduce management costs, and enjoy a better return on assets
- Improved operational agility, allowing institutions to deploy promising new software frameworks more quickly, and better collaborate with others without infrastructure getting in the way

Combining often discrete HPC, data warehouse, data science and analytic environments into a shared environment represents an opportunity to not only increase capacity, but to reduce costs. A more efficient software-defined infrastructure for healthcare can help improve collaboration, reduce cycle times, improve analytic capabilities, and ultimately improve patient care - results that are important to everyone.

Cabot Partners is a collaborative consultancy and an independent IT analyst firm. We specialize in advising technology companies and their clients on how to build and grow a customer base, how to achieve desired revenue and profitability results, and how to make effective use of emerging technologies including HPC, Cloud Computing, and Analytics. To find out more, please go to www.cabotpartners.com.