Genomics Deployments:
Getting It Right with Software Defined Infrastructure

March 26, 2018

Spectrum Scale User's Group - Singapore

J.D. Zeeman
Worldwide Business Development Executive
Software Defined Storage Solutions
IBM Systems

Disclaimer

- IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM’s sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

- Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

High Performance Data & AI for Healthcare & Life Sciences

IBM Blueprint, Architecture and Platform for Cognitive Infrastructure

World of Expertise

60%
Exogenous Factors

30%
Genomics Factors

10%
Clinical Factors

IoT & RWE

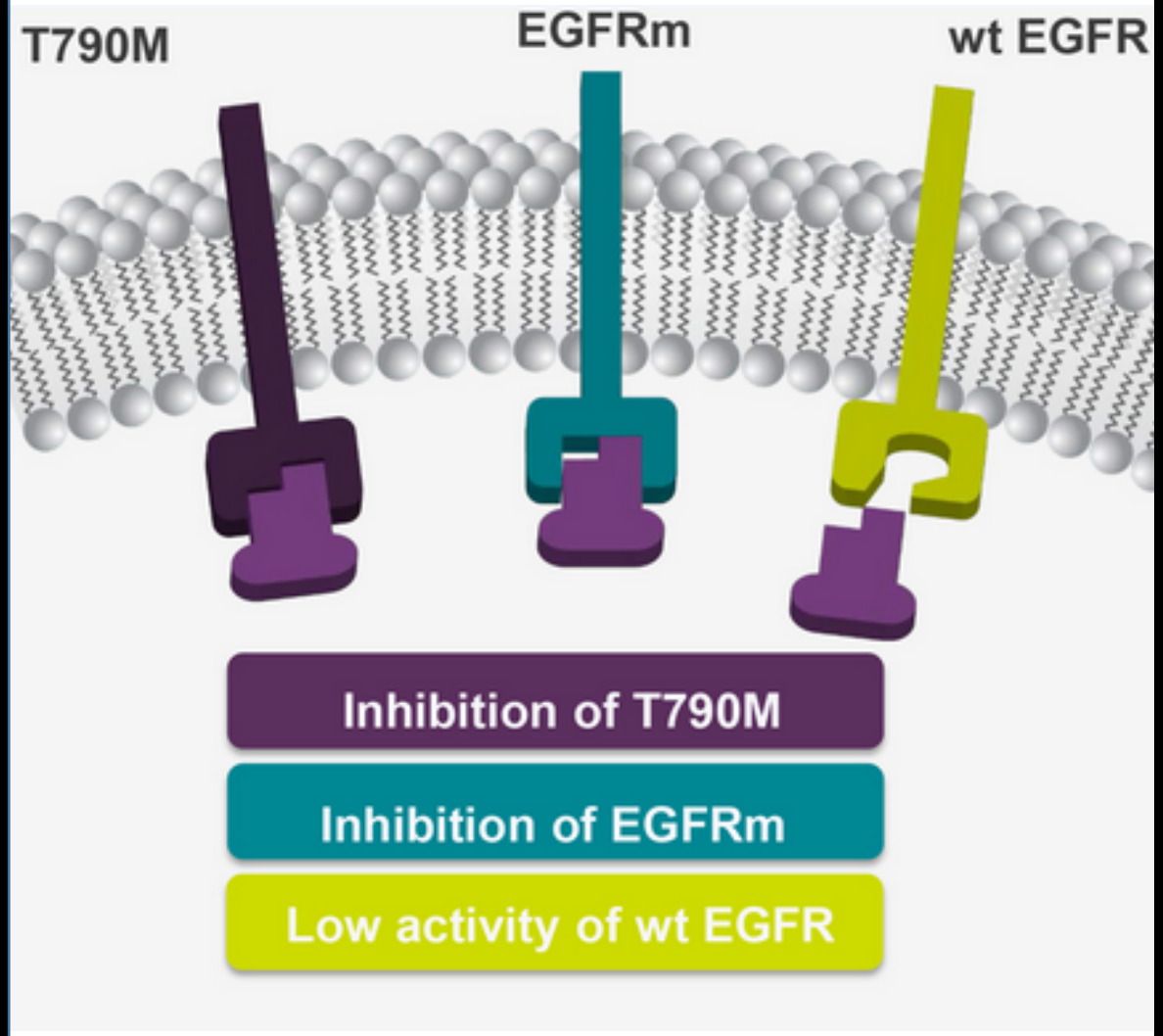
Genomics

Medical Imaging

Clinical

Precision Medicine

AZD9291



Example: Targeting mutation in EGFR receptor that can cause lung cancer

2.5 Years
from start of clinical trial to FDA approval (Nov 2015)

TAGRISSO™ osimertinib
AstraZeneca

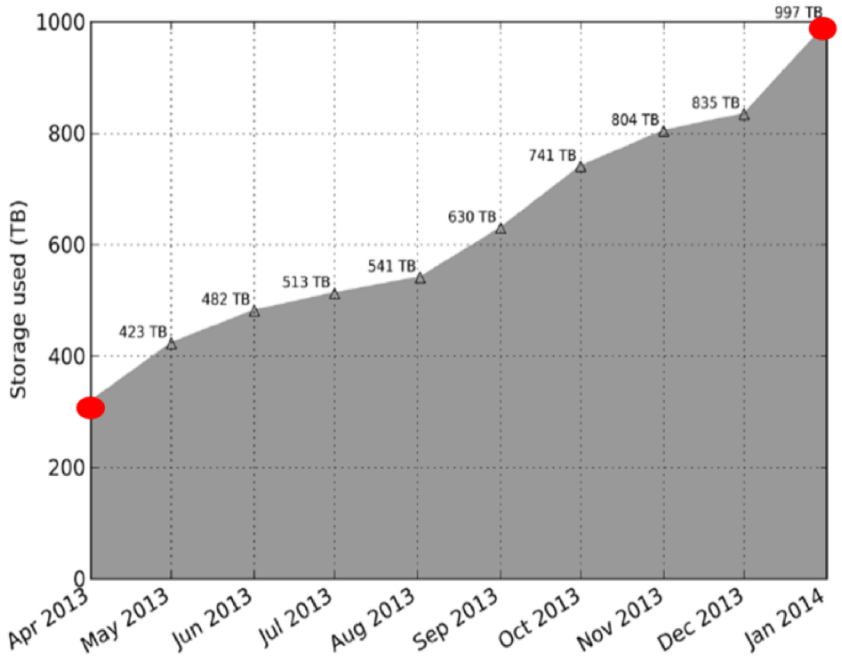
8 Hours
from tissue isolation to sequencing test results

Roche

World of Expertise

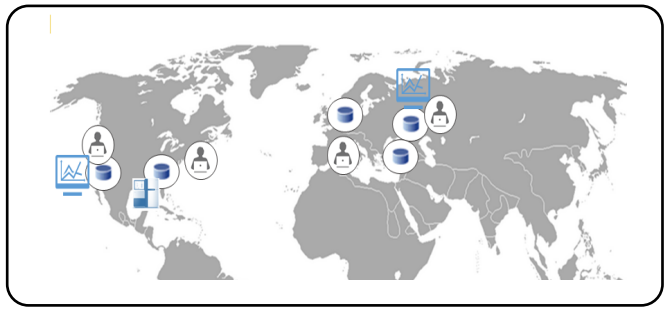
Ocean of Data

- **Byte:** 1 Grain of Rice
- **Terabyte:** 2 Container Ships
- **Petabyte:** Blankets Manhattan
- **Exabyte:** Blankets US West Coast States
- **Zettabyte:** Fills Pacific Ocean
- **Yottabyte:** AN EARTH SIZE BALL OF RICEç

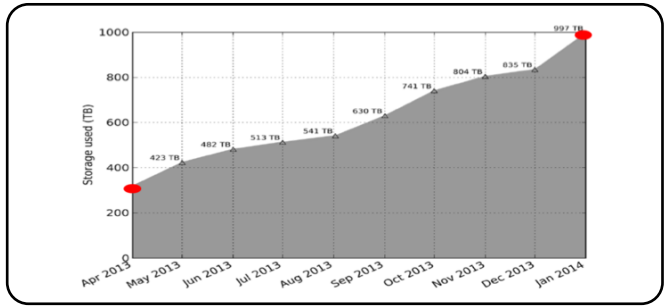


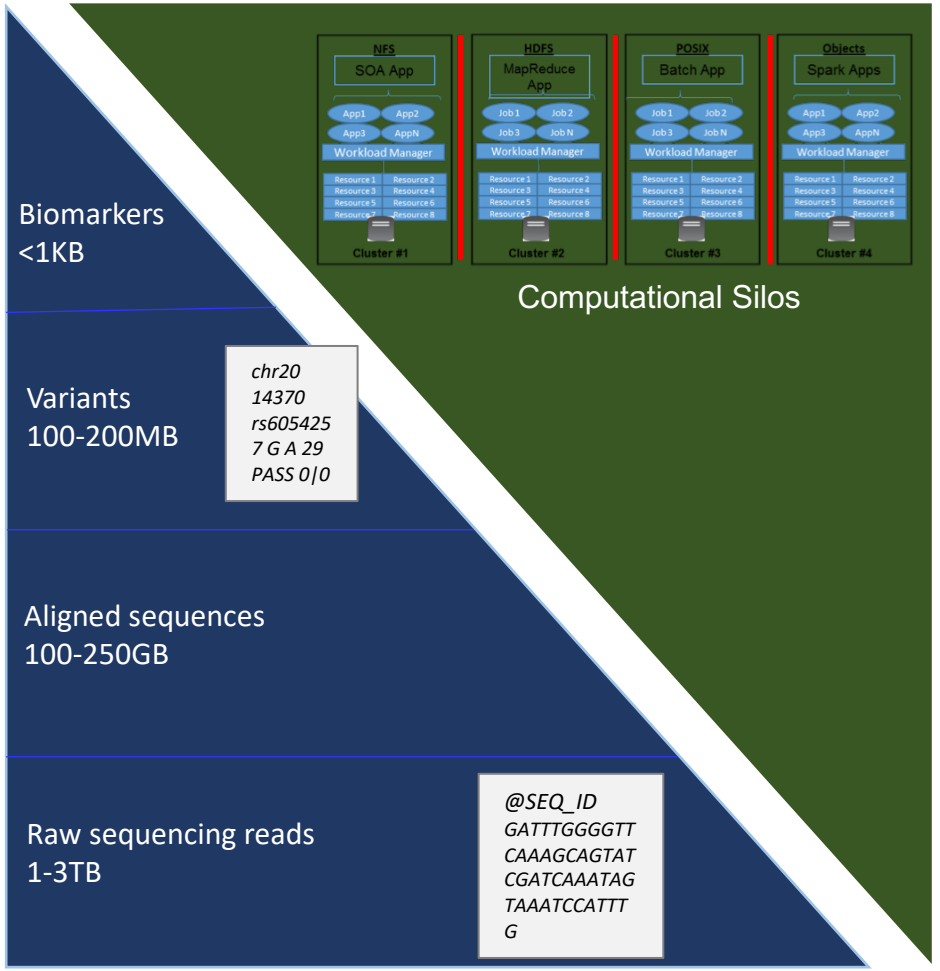
Genomics Today

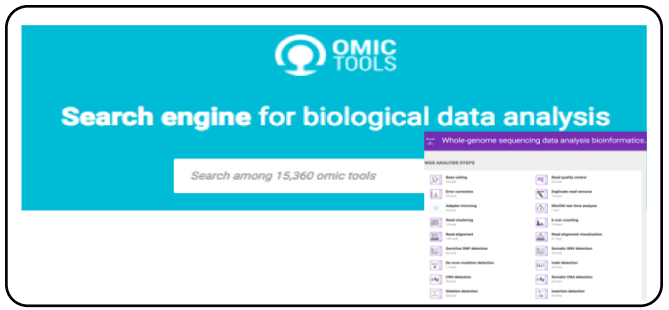
Distributed Data



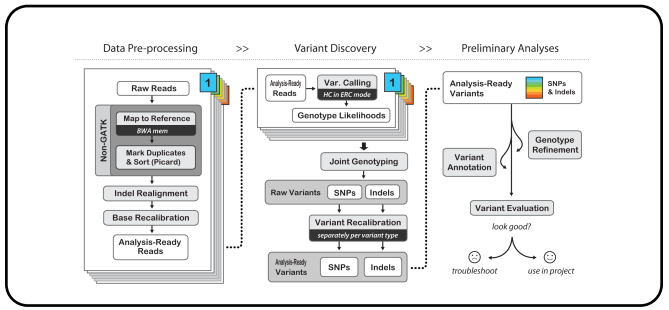
Fast Growing Data





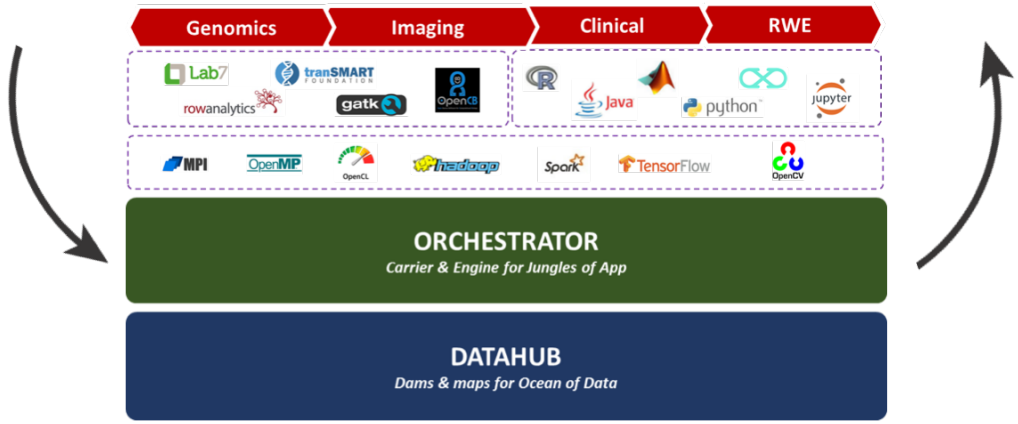


Sophisticated Tools and Workflows

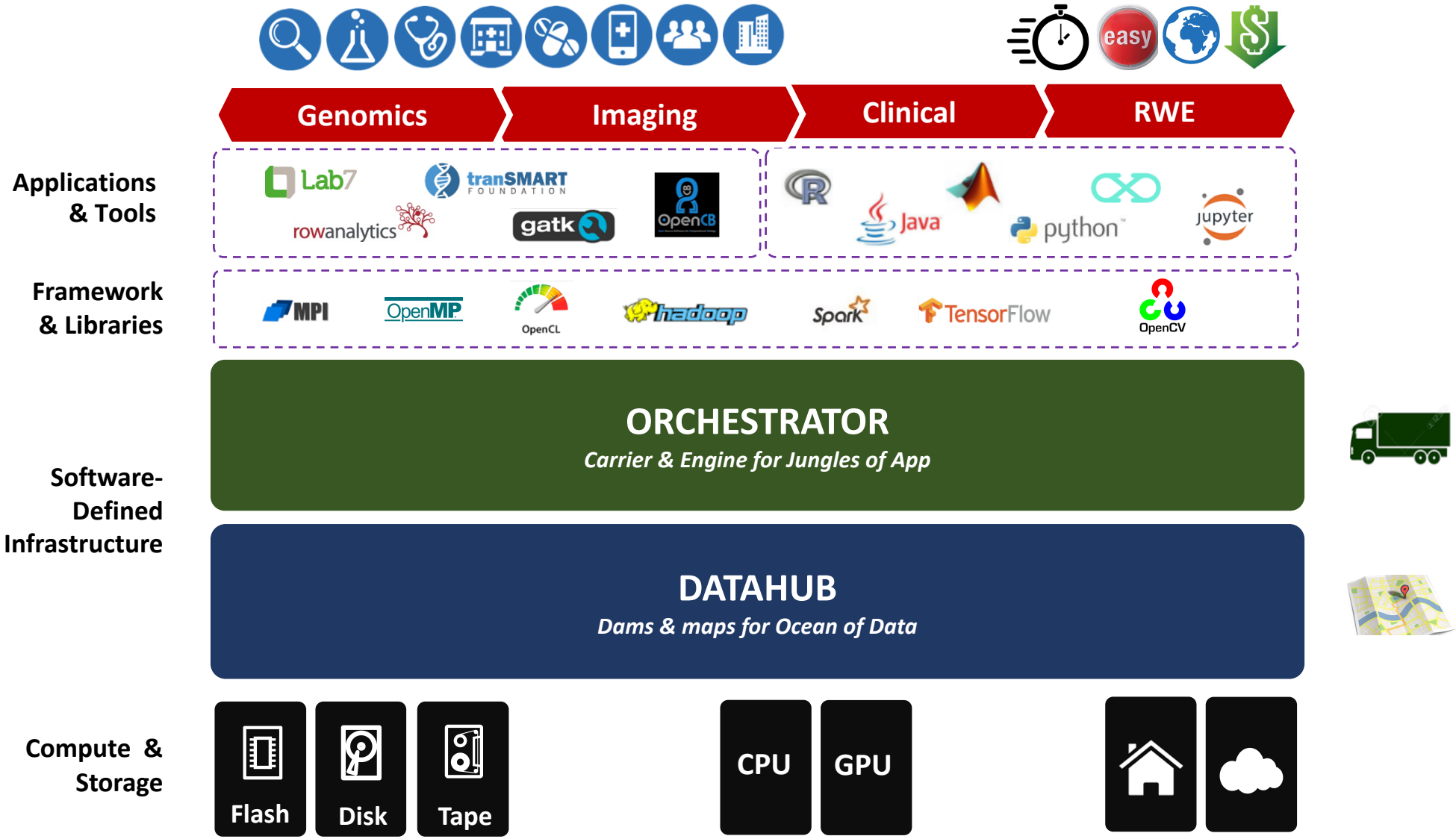


IBM Genomics Blueprint

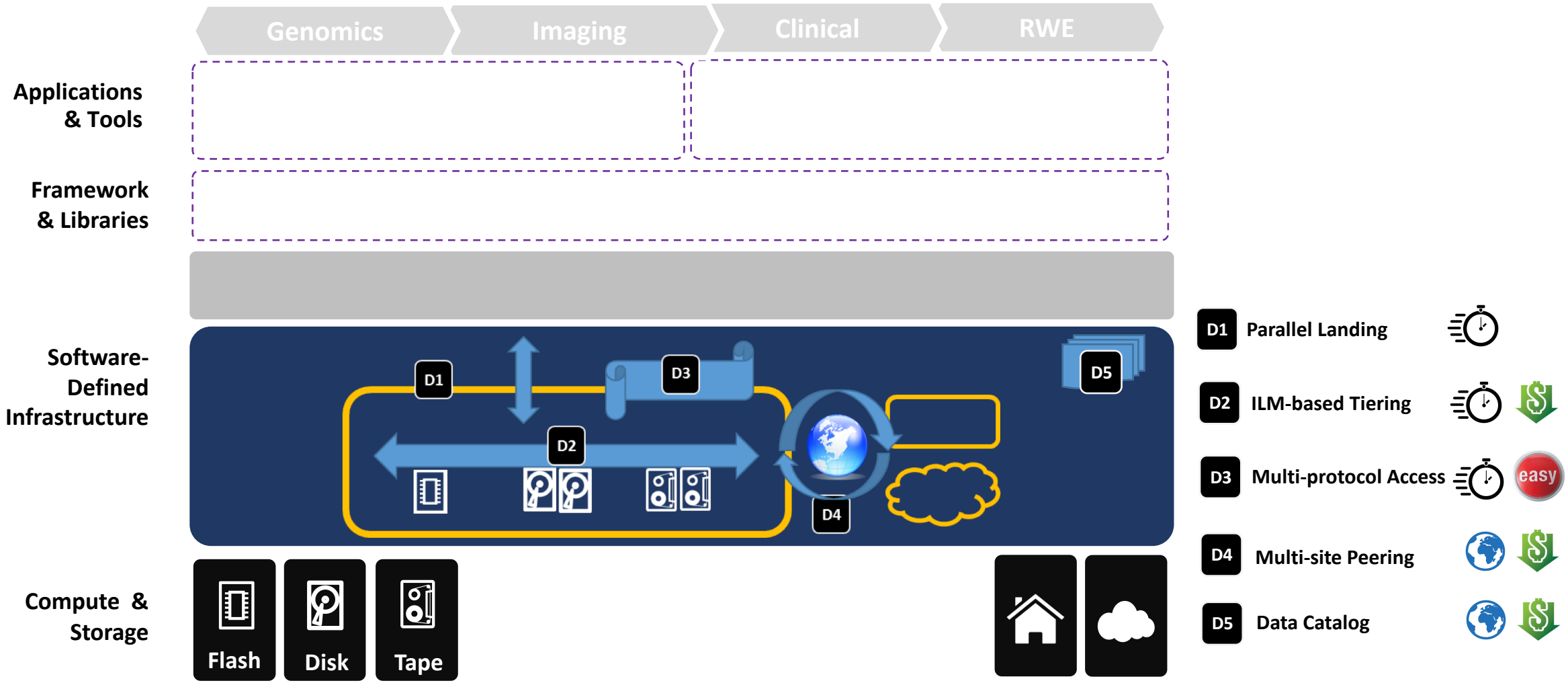
Data Scientists | Medical Researchers | **Clinicians** | **Radiologists** | **Pharma R&D** | Developers | Engineers | Managers & Administrators



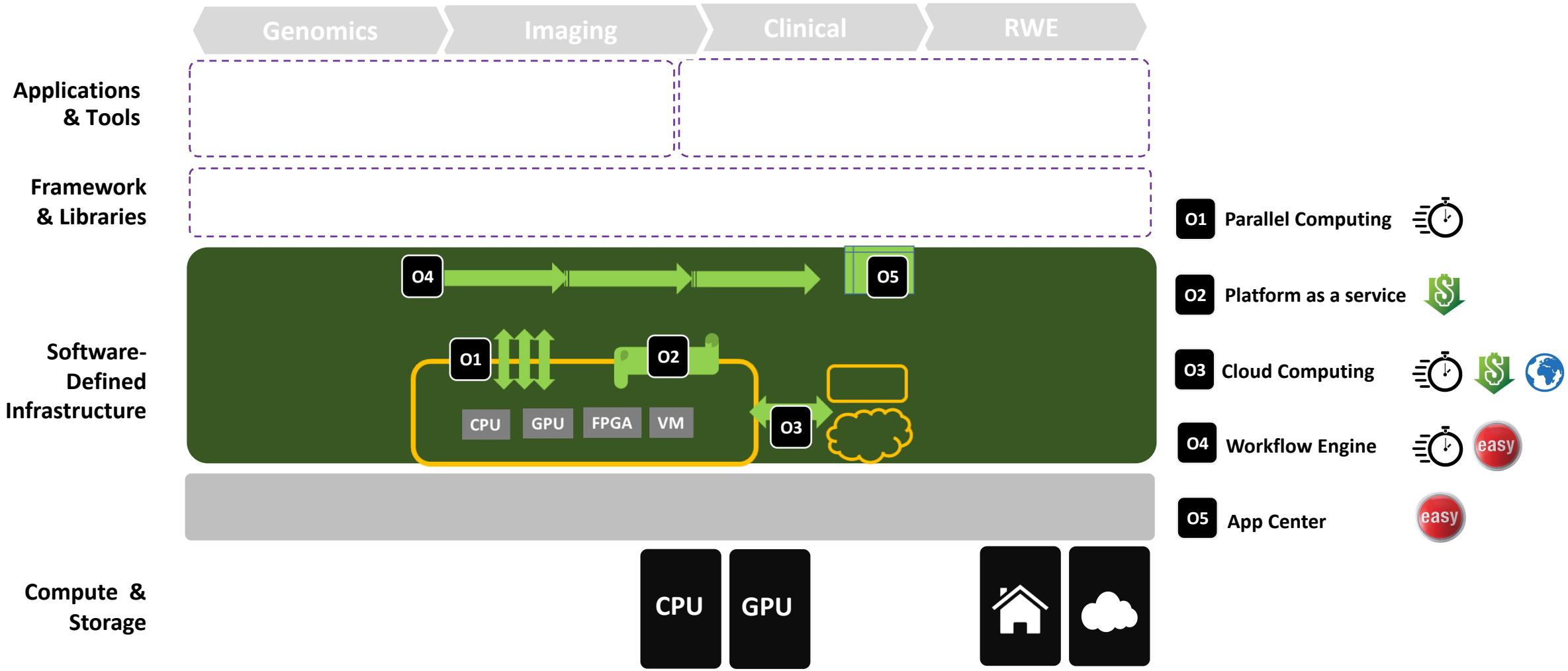
High Performance Data & AI (HPDA) Architecture



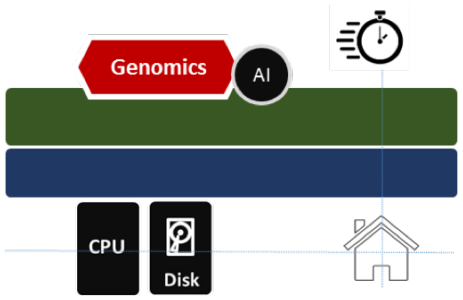
HPDA DATAHUB Overview



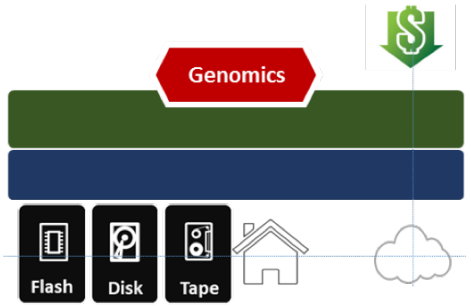
HPDA ORCHESTRATOR Overview



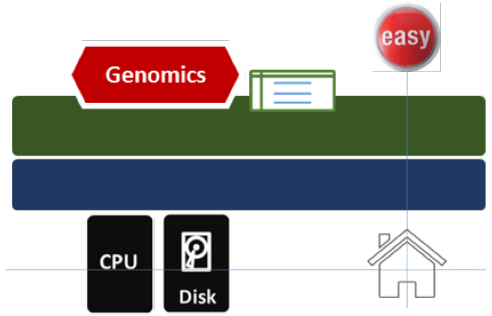
HPDA Genomics Representative Use Cases



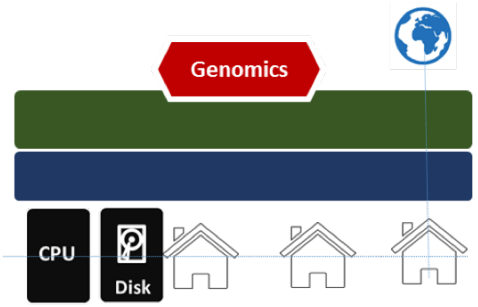
High Speed / High Performance
Reduce Time to Results by 10X



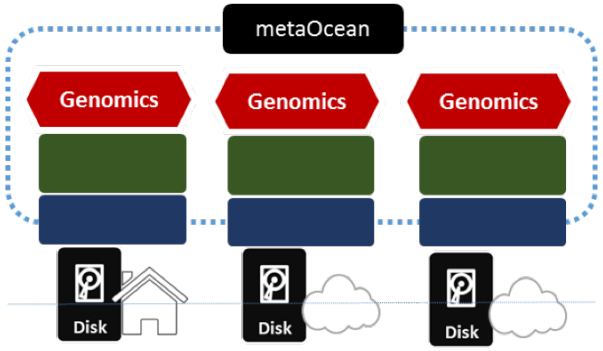
Lower Cost
Reduce Cost by 10X



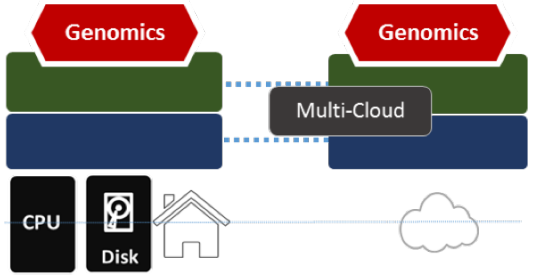
Ease of Use
Enable an App Store



Facilitate Collaboration
Share Data Globally



Harness Meta Data
Generate a Catalog



Enable Multi-Cloud
Burst to Public Cloud

Success Stories

Reduced time to completion for long running jobs while increasing resource utilization

"Analyzing hundreds of samples in parallel on a regular basis requires a robust HPC system to handle the load properly. From our experience, IBM systems has proven to be reliable in helping us address this technical requirement."

Dr. Mohamed-Ramzi Temanni, Manager, Bioinformatics Technical Group at Sidra Medical and Research Center

More than 3x performance using 1/3 the nodes

"Delta will enable quantitative analysis and interpretation of large biological genomics data generated at LSU".

Gus Kousoulas, associate vice president for research and economic development, Louisiana State University

96% reduction in the runtime of a standard genome analysis pipeline

"With IBM Cloud, and in particular its high-performance compute infrastructure and services, we found the ideal platform for building a comprehensive cloud solution for genomics".

Christopher Mueller, Ph.D., President and Chief Technology Officer of Lab7 Systems

Accelerating genetics research and medicine 500% with IBM SDI

As a result of replacing its open-source workload manager—which crashed on a monthly basis—with IBM Spectrum LSF to improve both reliability and scalability, the team has seen core dumps fall to zero over a one-year period while overall scalability increased to 500,000 jobs per queue.

Icahn School of Medicine, Mount Sinai

Composable Infrastructure for Genomics Workload

IBM Spectrum Scale Best Practices for Genomics Medicine Workload

Overview

Spectrum Scale: Solution Brief

IBM is helping life science companies across the globe to accelerate research and drug development by providing infrastructure to store, share and manage huge amounts of genomics data and to analyze it quickly.

Deeper, Faster insights with compostable building blocks based in IBM Spectrum Scale

Gives a quick overview of the solution its advantages and references.

Download from:
https://ibm.co/2uhCvuM

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

IBM

Highlights

- Manage the expanding genomics data ecosystem to store, access, secure, manage, share and analyze significant amounts of data
- 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 architect, 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]

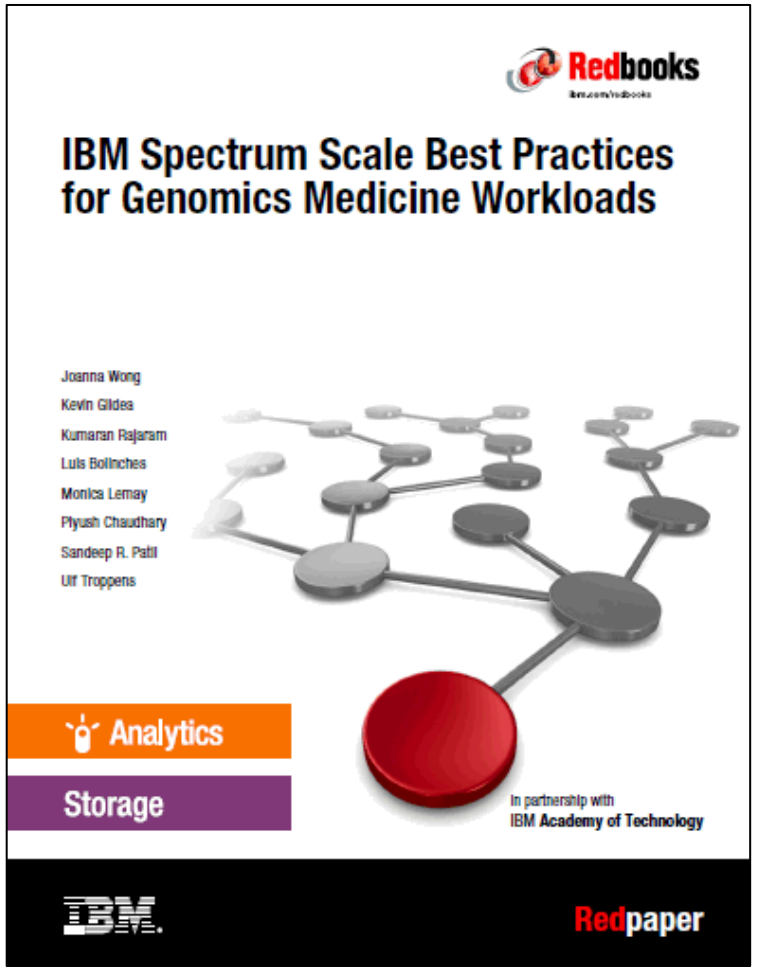
As genomics data becomes a critical source for precision medicine, it is expected to create an expanding data ecosystem. This means that hospitals, genome centers, medical research centers and other clinical institutes need to explore new approaches that will help them unleash the real value of significant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

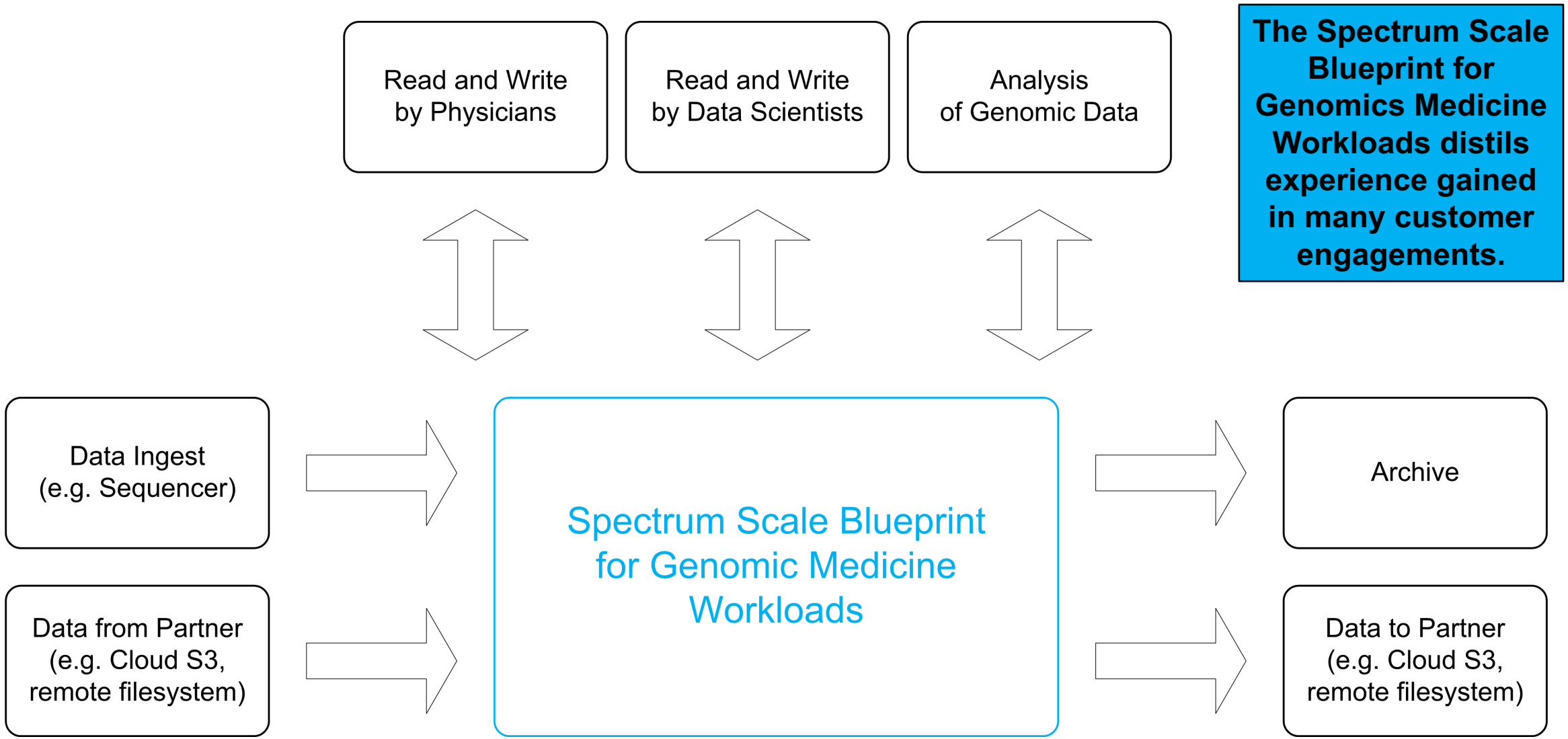
Additional Detail



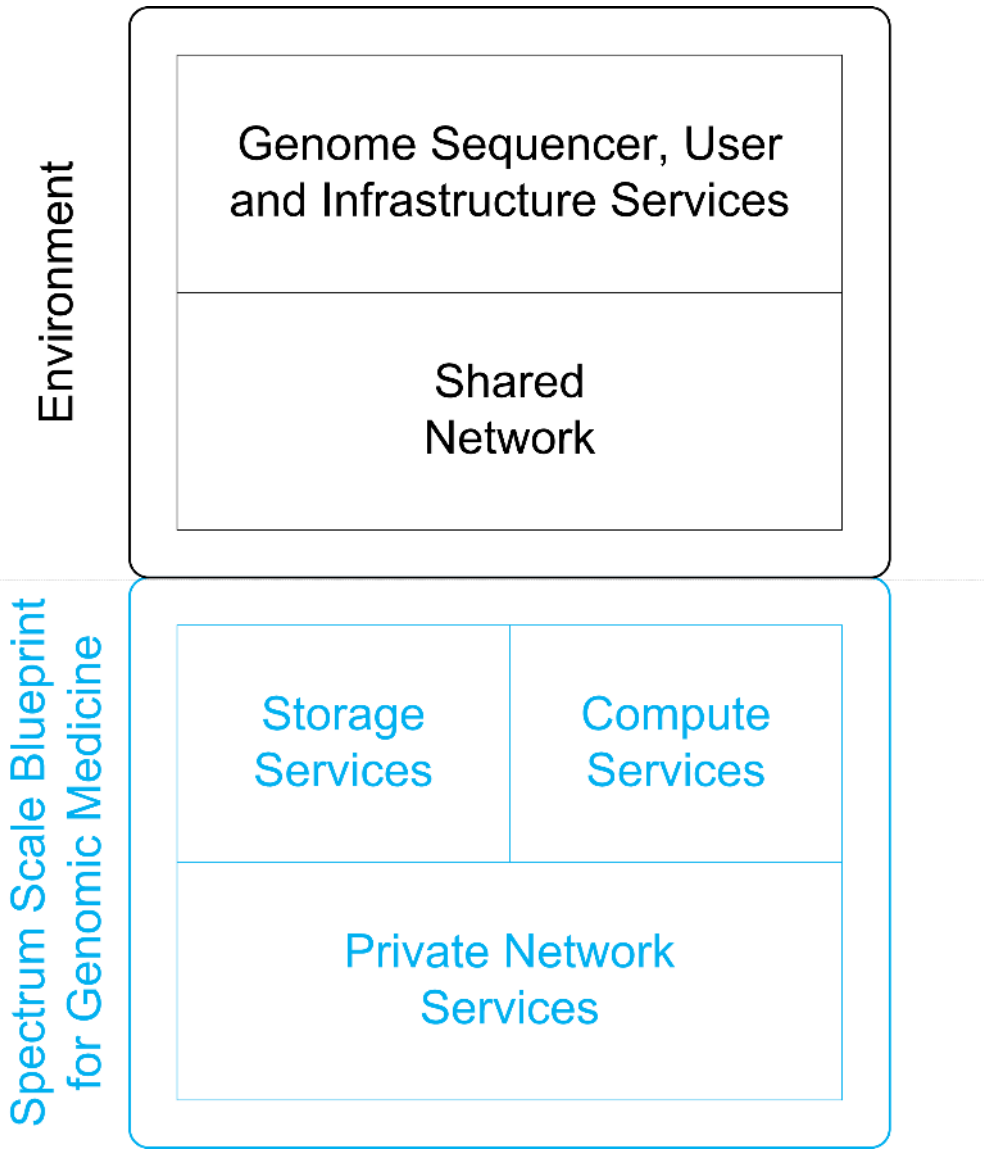
- IBM distilled the experience gained in the many customer engagements in the IBM Spectrum Scale Best Practices for Genomics Medicine Workloads.
- The best practices guide provides composable infrastructure based on expertly engineered building blocks that enable IT architects to customize deployments for varying functional and performance needs.
- The modular approach allows to integrate selected building blocks into the customer's already existing infrastructure to protect already made investments.

http://www.redbooks.ibm.com/abstracts/redp5479.html?Open

Overall Context



Composable Building Blocks



Compute Services

- Scale-able **Compute Cluster** to analyze genomics data.

Storage Services

- Scale-able **Storage Cluster** to store, manage and access genomics data.

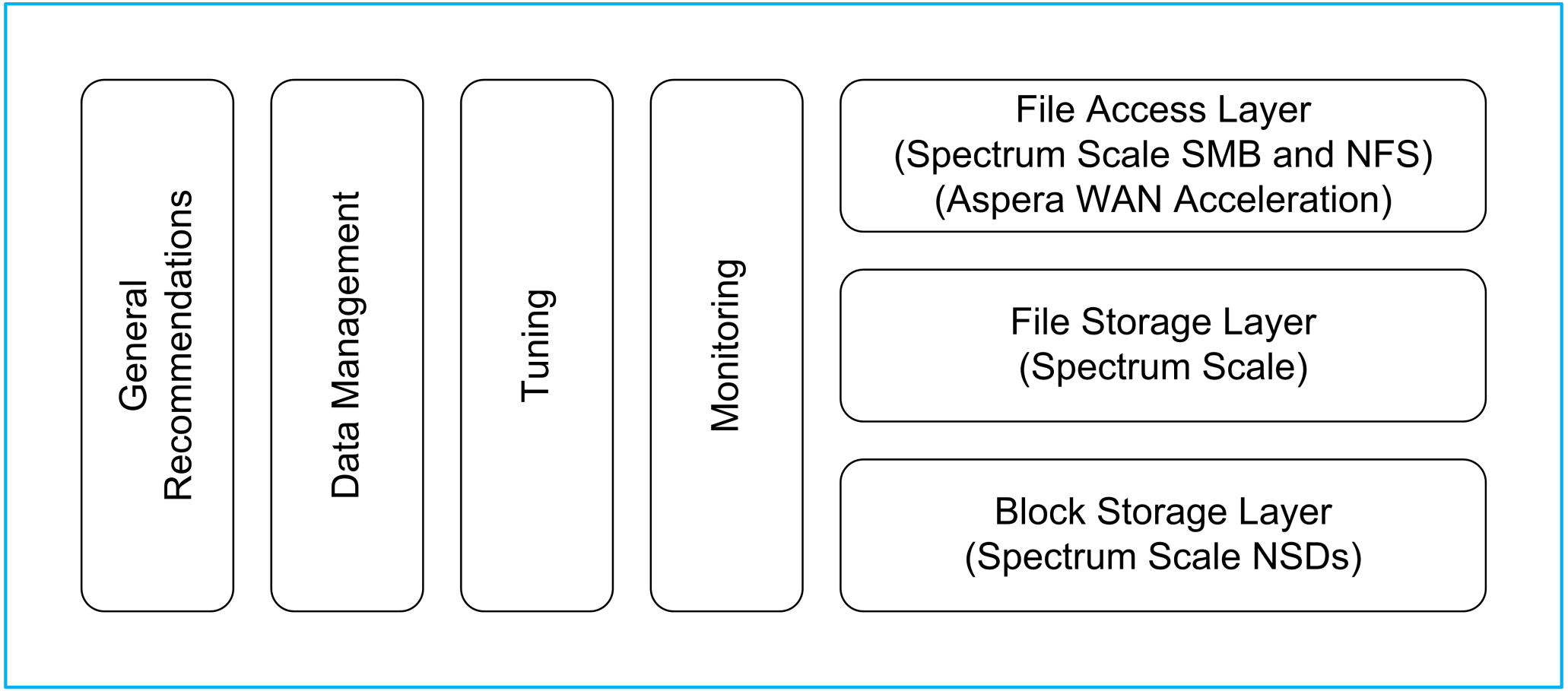
Private Network Services

- **High-speed Data Network,** not connected to data center network.
- **Provisioning Network** and **Service Network** for administrative login and hardware services, optionally connected to shared campus network.

Interfaces with Shared Network

- **User Login** to submit and manage batch jobs and to access interactive applications.
- **High-speed NFS and SMB Data Access,** connected to shared campus network.
- WAN Optimization for fast and secure remote access to enable **collaboration across sites and institutions**.

Storage Services - Composable Building Blocks



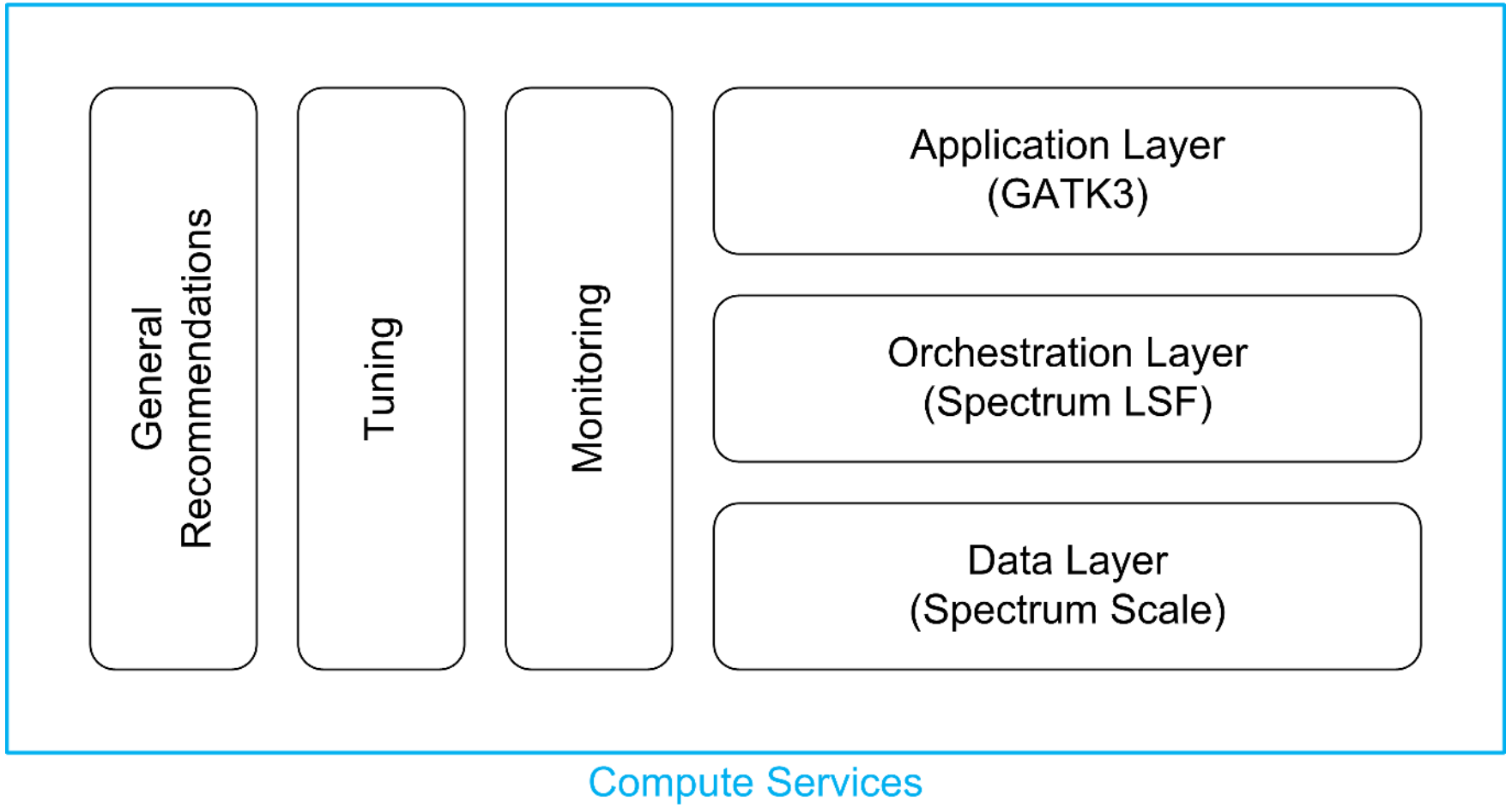
➔ ***A set of expertly engineered building blocks enable IT architects to compose solutions that meet customers varying performance and functional needs.***

Capabilities - Blueprint V1.1 - Storage Services

- To enable **access to genomics data** the **Storage Cluster** provides:
  - **Data Transfer Nodes** for secure **high-speed external access via NFS and SMB** to ingest data from genomic sequencers, microscope, etc., for access by data scientists/physicians
  - **WAN Acceleration** for **collaboration across sites and institutions**
  - Secure **high-speed internal access** for analysis on Compute Cluster
- To **effectively store and manage genomics data** the **Storage Cluster** provides:
  - **Scale-out architecture** that is capable to store data from a few 100 TB to Tens of PB of file data
  - **End-to-end checksum** to ensure the data integrity all the way from the application to the disks
  - **Quota Management** for user and project groups (future)
  - **Snapshots** for user and project groups (future)
  - **Integrated Back-up** and **Fast Restore** of PBs of data (future)
  - **Data Management GUI** to configure and monitor storage resources
  - Optional **Professional Services** ranging from management of daily operation to consultancy for major configuration changes

➔ Blueprint capabilities have been reviewed with and prioritized by IBM Healthcare and Life Science team.
➔ Blueprint capabilities are written in a product neutral language to emphasize end user requirement.

Compute Services - Composable Building Blocks



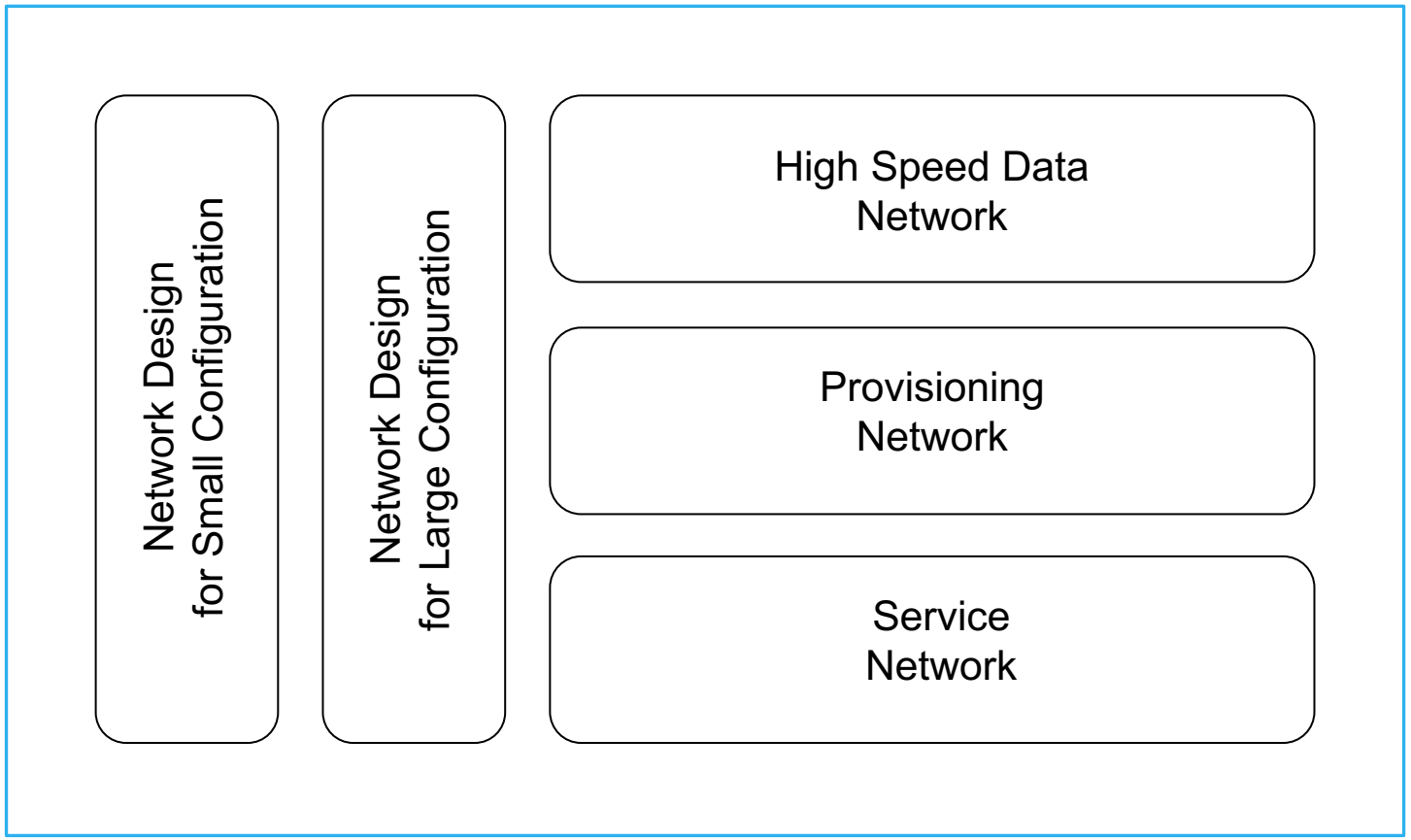
➔ ***A set of expertly engineered building blocks enable IT architects to compose solutions that meet customers varying performance and functional needs.***

Capabilities - Blueprint V1.1 - Compute Services

- To enable the **analysis of genomics data** the **Compute Cluster** provides:
  - **User GUI** for physician/data scientist to submit and manage batch jobs and to create and manage custom workflows
  - **Workload Management GUI** for IT administrator to view cluster status and utilization
  - Secure **high-speed access** to files stored on Storage Cluster
- Scaling
  - A **Workload Scheduler** enables high-throughput execution of batch jobs
- Performance
  - **Tuning Recommendations** supporting the “Broad Institute GATK Best Practices on IBM reference architecture”
- Node Types
  - **Power and/or x86-64 Nodes** for batch processing and for interactive login to access the resources

➔ Blueprint capabilities have been reviewed with and prioritized by IBM Healthcare and Life Science team.
➔ Blueprint capabilities are written in a product neutral language to emphasize end user requirement.

Network Services - Composable Building Blocks



➔ ***A set of expertly engineered building blocks enable IT architects to compose solutions that meet customers varying performance and functional needs.***

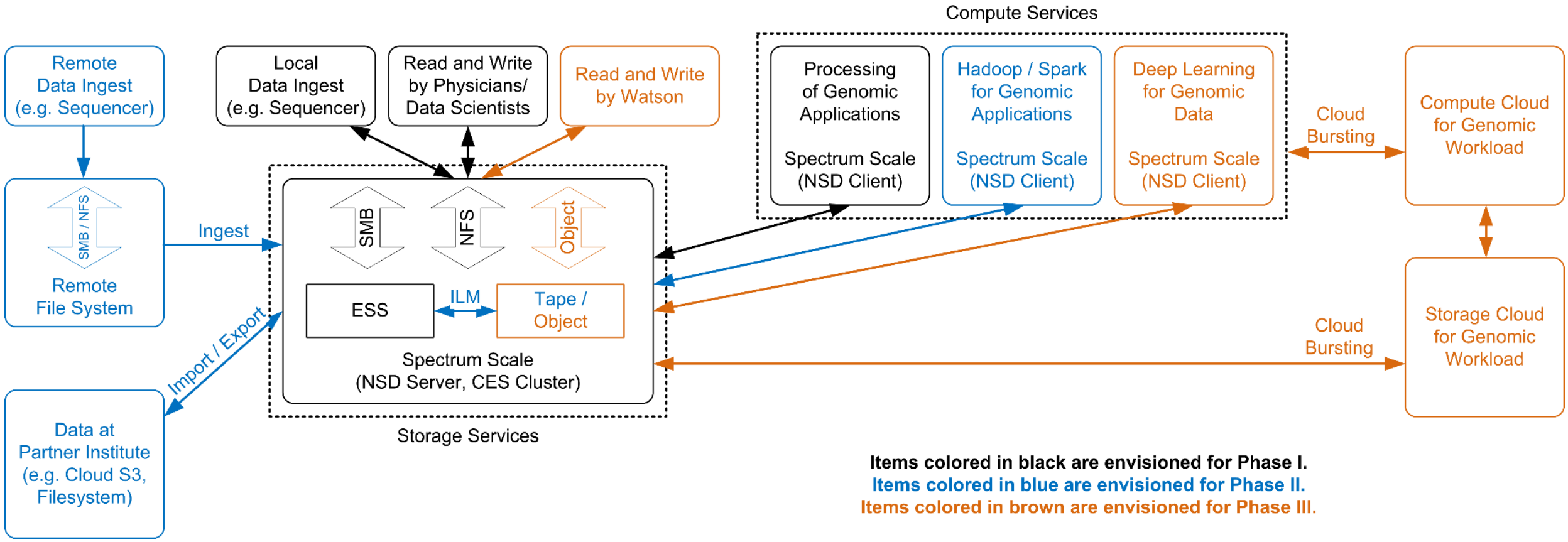
Capabilities - Blueprint V1.1 - Private Network Services

- To integrate all components of the Compute Services and all components of the Storage Service into an **IT Infrastructure Solution for Genomics Workload** the **Private Network** provides:
  - A **High-Speed Data Network** for **fast and secure access to genomics data**:
    - **Storage Nodes** are configured with high availability by default (at least two links).
    - **Compute Nodes** are optionally configured with high availability (one or two links).
  - A **Provisioning Network** for provisioning and in-band **management** of the storage and compute components and for **administrative login**.
  - A **Service Network** for out-band management and monitoring of all solution components.
  - A **Scalable Design** that can start from a **small starter configuration** and grow to a large configuration that consists of **hundreds of compute nodes** and **tens PB of storage**.

➔ Blueprint capabilities have been reviewed with and prioritized by IBM Healthcare and Life Science team.
➔ Blueprint capabilities are written in a product neutral language to emphasize end user requirement.

IBM is Utilizing a Staged Approach

IBM Genomics Blueprint

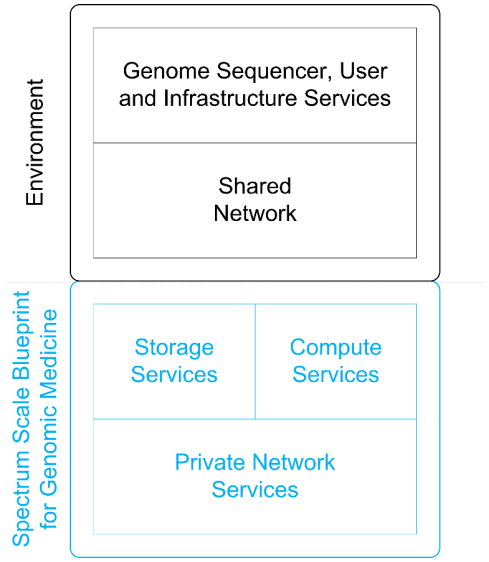


IBM is Providing Go Forward Support

The IBM Blueprint for Genomics Medicine Workload consists of expertly engineered, composable building blocks which include:

- **Best practices guides** for architecture and configuration settings
- **Runbooks** which describes how to install, configure, monitor and upgrade example configurations
- **Sizing guidelines** which help to define a solution which meets the customers performance requirements
- **Deployment workshop** available to clients to customize solution to client specific requirements

IBM Spectrum Scale File Systems – Guidelines for Genomics Workload	
Name	/gpfs/data
Purpose	Store genomics data and analysis result
Why separate file system?	This file system is the workhorse to store most of the data
Size	Depends on customer requirements: Few TiB up to Hundreds of PiB
Metadata	1 MiB block size on SSD
Data	8 MiB block size on NL-SAS
Log File Size	32 MiB (-L 32M)
Block Allocation Map	Scatter
Replication	Replicate metadata only (-M 2 -R 2 -m 2 -r 1)
ACL Type	NFSv4 only
Filesets	Multiple independent filesets (details follow later)
Relatime	Suppress the periodic updating of the value of atime (-S relatime)
Quota	Enable quota (-Q yes) (avoids remount when we enable quota later)
Exported to Compute Cluster	Yes (via IBM Spectrum Scale multi-cluster remote cluster mount)
Exported via CES	Yes (SMB and NFS)
Number of Nodes	Customer specific



IBM is enabling Client Value

- Effectively storing, securing, managing, sharing and analyzing the emerging “data tsunami”
- Successfully supporting an expanding data ecosystem of frameworks and applications
- Allowing distributed clinical and research professionals to analyze massive amounts of genomics data with speed, low cost and ease of use
- Assisting IT architects and IT administrators to more easily design, install and manage deployment with speed, low cost and ease of use
- Providing robustness and flexibility via a Software Defined Infrastructure to fulfil both current and future requirements

Legal notices

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectually property rights, may be used instead.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER OR IMPLIED. IBM LY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. IBM makes no representations or warranties, ed or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 1 0504- 785
U.S.A.

Information and trademarks

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.

Special notices

This document was developed for IBM offerings in the United States as of the date of publication. IBM may not make these offerings available in other countries, and the information is subject to change without notice. Consult your local IBM business contact for information on the IBM offerings available in your area.

Information in this document concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The information contained in this document has not been submitted to any formal IBM test and is provided "AS IS" with no warranties or guarantees either expressed or implied.

All examples cited or described in this document are presented as illustrations of the manner in which some IBM products can be used and the results that may be achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions.

IBM Global Financing offerings are provided through IBM Credit Corporation in the United States and other IBM subsidiaries and divisions worldwide to qualified commercial and government clients. Rates are based on a client's credit rating, financing terms, offering type, equipment type and options, and may vary by country. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice.

IBM is not responsible for printing errors in this document that result in pricing or information inaccuracies.

All prices shown are IBM's United States suggested list prices and are subject to change without notice; reseller prices may vary.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Users of this document should verify the applicable data for their specific environment.

Thank You.

IBM Storage & SDI