IBM Spectrum Scale Blueprints

IBM Spectrum Scale Blueprint for Genomics Medicine Workloads

Spectrum Scale User Group Meeting @ Client Insight UK (CIUK)
Dec 12th, 2017

Joanna Wong, Kevin Gildea, Kumaran Rajaram, Luis Bolinches, Monica Lemay, Piyush Chaudhary, Sandeep Ramesh, Ulf Troppens
Speaker: Ulf Troppens

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

Background

General

- Not tight to Spectrum Scale 5.0
- We are using Spectrum Scale 5.0 as shipping vehicle to promote the blueprint
- The initial version is based on ESS 5.2 and Spectrum Scale 4.2.3.4 and 4.2.3.5.

Timeline

- 4/2017 Design Thinking Workshop
- 7/2017 Started to establish team
- 10/2017 Redbook residency
- 12/2017 Published 1st draft of Redbook

Core Team

- Joanna Wong (HPC Architect, Client Centers)
- Kevin Gildea (DE)
- Kumaran Rajaram (Real Fast)
- Luis Bolinches (Lab Based Services)
- Monica Lemay (Real World)
- Piyush Chaudhary (HPDA)
- Sandeep Ramesh (Client Enablement)
- Ted Hoover (Sponsor Manager)
- Ulf Troppens (Client Enablement)

Extended Team

- Interlock with worldwide Health Care and Life Science sales team
- Interlock with worldwide Health Care and Life Science marketing team

Enablement Activities

IBM Spectrum Scale

Webinar 1 – Accelerating Discoveries with IBM Spectrum Scale for Genomic Medicine Workload
Date: 16th Oct 2017 (Replay available)
Invited Audience: LBS, Pre-Sales, BP, Support Team
https://w3-03.ibm.com/sales/support/ShowDoc.wss?docid=SGDM575772V10098F26

Webinar 2 – Deep Dive – Spectrum Scale Blueprint for Genomic Medicine Workload
Date: 28th Nov 2018 (Replay will be available)
Invited Audience: LBS, Pre-Sales, BP
https://w3-03.ibm.com/sales/support/ShowDoc.wss?docid=SGDQ343191R30079T88

Solution Brief
Date: Dec 2017

Redpaper: Spectrum Scale Best Practices for Genomic Medicine Workload
Date: Dec 10th, 2017
http://www.redbooks.ibm.com/abstracts/redp5479.html

WarRoom
Date: Dec 2017

IBM Spectrum Scale

Spectrum Scale Best Practices Guide for Genomic Medicine Workload 1.0 (Solution Overview)

Dec 4th, 2017

Summary

- There are successful Spectrum Scale based deployments for storing and analyzing huge amounts of genomic data that enable customers to get results more quickly.
- The Spectrum Scale Blueprint for Genomic Medicine Workload compiles best practices that enable IT architects to create a solution architecture for genomics medicine that meet the customer’s performance requirements.
- The Spectrum Scale Blueprint for Genomic Medicine Workload describes expertly engineered building blocks that can be composed to meet customers varying performance and functional needs.
- The Spectrum Scale Blueprint for Genomic Medicine Workload provides an approach to integrate selected building blocks into the customer’s already existing infrastructure to protect already made investments.

Outline

- ***Market Opportunity***
- Composable Solution Architecture
- Driven by Design Thinking
- Driven by Agile Development
- Blueprint Capabilities
- Example Configuration

Discussion Points

Towards personal treatments

- Costs for genomics sequencing are going down
- Genome sequencing is arriving hospitals for translational medicine
- Single cell sequencing

Data management challenge

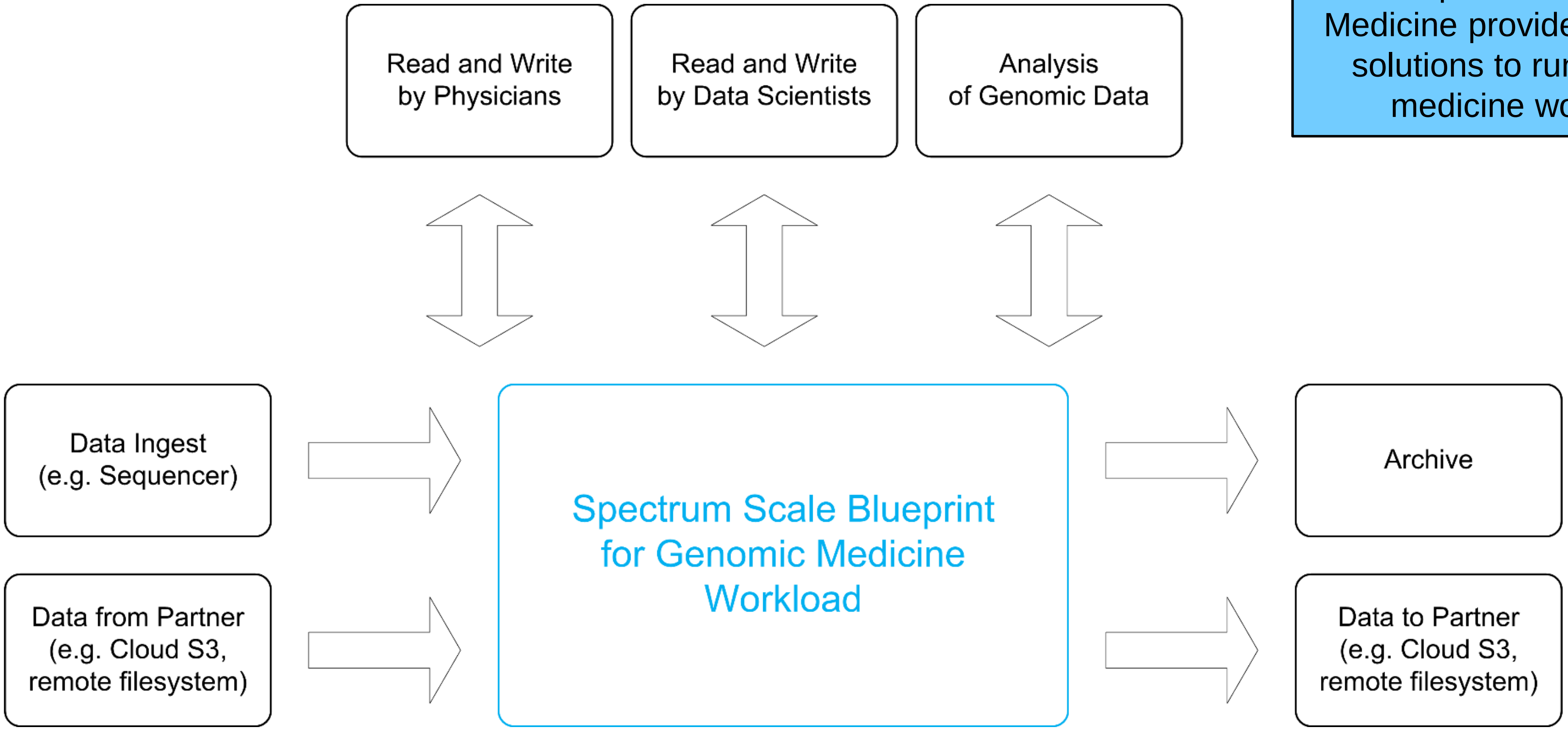
- Single sample has a size of 100GB-250GB
- Data is acquired outside the data center
- Distributed teams, global collaborations
- Legal obligations and best practices for research require to archive data for at least ten years
- Small customers quickly grow into the double digit PB range for both active and archived data

Date analysis challenges

- Complex tools and workflows
- Workflow for single sample runs several 10 hours
- A few predominant applications like GATK
- Broad ecosystem of hundreds of applications
- Fist use case that requires sizeable IT infrastructure
- End users are more IT agnostic than in other scientific fields

System Context

The Spectrum Scale Blueprint for Genomic Medicine provides pretested solutions to run genomic medicine workload.



Outline

- Market Opportunity
- ***Composable Solution Architecture***
- Driven by Design Thinking
- Driven by Agile Development
- Blueprint Capabilities
- Example Configuration

Why Blueprints?

(1) Spectrum Scale is a flexible Swiss army knife which can be tweaked to support a broad range of workloads and applications.	(2) There are successful deployments of Spectrum Scale to support new workloads such OpenStack, Hadoop/Spark and file-based workflows.
(3) Spectrum Scale beginners are overwhelmed and overtaxed by broad range of configuration and deployment options.	(4) Blueprints enable IT architects and IT specialists to design, deploy and operate Spectrum Scale based solutions using expertly engineered building blocks.

Composable Infrastructure

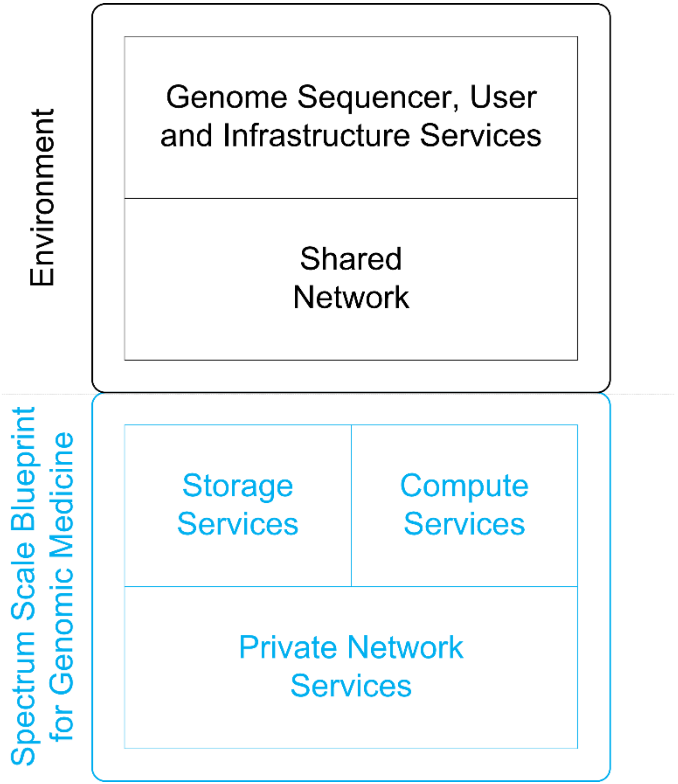
General

- Composable solutions are built in a way that disaggregates the underlying building blocks viz. compute, storage, and network services.
- These disaggregated services provide the required granularity allowing the infrastructure that can be sliced, diced, expanded and contracted at will and based on the actual need.
- It facilitates ease in deployment with well defined configuration and tuning templates per building block.

Spectrum Scale Blueprint for Genomic Medicine Workload

- Compiles best practices that enable IT architects to create a solution architecture for genomics medicine that meet the customer’s performance requirements.
- Describes expertly engineered building blocks that can be integrated into an end-to-end solution that meets customers varying performance and functional needs.
- Provides an approach to integrate selected building blocks into the customer’s existing infrastructure to protect already made investments.

Composable Building Blocks



Shared Network

- **High-speed NFS and SMB Data Access,** connected to shared campus network.
- **User Login** to submit and manage batch jobs and to access interactive applications.

Compute Services

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

Storage Services

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

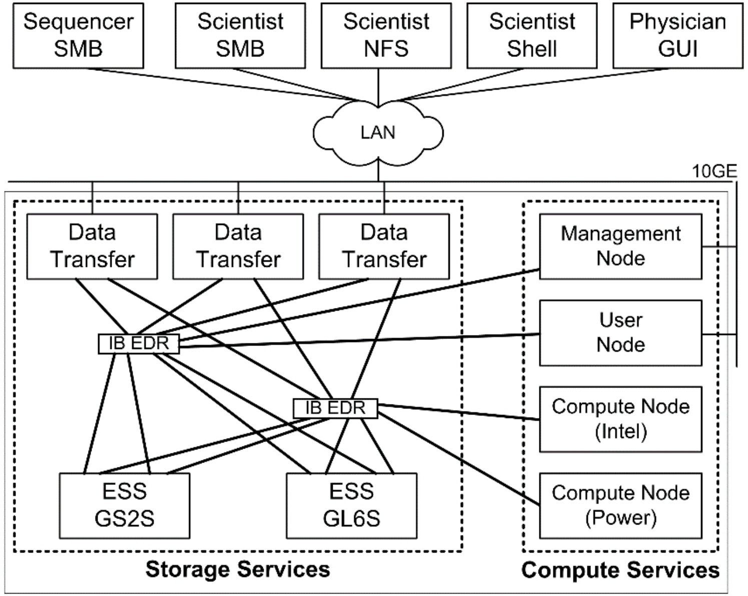
Private Network Services

- **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.

Best Practices Guides

Spectrum Scale Best Practices Guide for Genomics Medicine Workload

1) Solution Overview
2) Best Practices for Compute Services
3) Best Practices for Storage Services
4) Best Practices for Private Network Services



Spectrum Scale Blueprint for Genomic Medicine

➔ Best Practices Guides include one specific example environment.

Outline

- Market Opportunity
- Composable Solution Architecture
- ***Driven by Design Thinking***
- Driven by Agile Development
- Blueprint Capabilities
- Example Configuration

Hills

1 Art, the seller, can give the winning proposal(*) to a client for genomic medicine workload that gives the data scientist faster time to results compared to competition. (*) Metric + win/loss reports

2 Aidan, the IT architect, can create a Spectrum Scale based solution architecture that meets genomic medicine workload performance requirements without consulting from IBM R&D.

3 Aidan, the IT architect, can integrate a Spectrum Scale based solution that meets their genomic workload performance requirements into an existing infrastructure without consulting from IBM R&D, without service disruption to the data scientist.

Chris

Data Scientist

- Chris has to deliver analysed results to the hospital physicians.
- Chris has to run her workload, when the physicians want it, in a timely manner.
- Chris doesn't want to worry about knowing IT.
- Chris is frustrated because it takes too long to get the system implemented.
- Chris has to know too much about IT.
- Chris needs IBM to come in to help.
- Chris just wants to do her job.

➔ Chris needs a way to run the workload to get results to make a treatment result when needed/quickly.

➔ Chris needs to get insights without knowing about IT.

Art

IBM/Partner Seller

- Art is responsible for proposing a solution to Aidan, the IT architect, that satisfies Aidan's needs and meets Aidan’s budget.
- Art feels the opportunity is too high touch, takes too long, takes too much of his time.
- Art just wants to sell a box.
- Art has to understand technical requirements to create one off solutions, and takes too much time.
- Art is not aware of the complexities and sells ESS as an appliance without understanding the overall solution.

➔ Art needs to manage his sales pipeline, so he can sell a lot, e.g. pick the right opportunity.

➔ Art needs to create winning proposals and close deals quickly.

Aidan

IT Architect

- Aidan is responsible for delivering the architecture that will run Chris', the data scientists, workload and deliver Chris workload in a timely manner.
- Aidan is overwhelmed by the complexity.
- Aidan is uncertain that it will work.
- Aidan needs to have to call development or research for advice. Aidan just wants development or research to validate her assumptions, but she feels like they are guessing.

→ Aidan needs to plan and architect an end-to-end solution optimized to run genomic medicine workload.

→ Aidan needs certainty on cost and performance to run the workload within budget.

Kevin

IT Admin

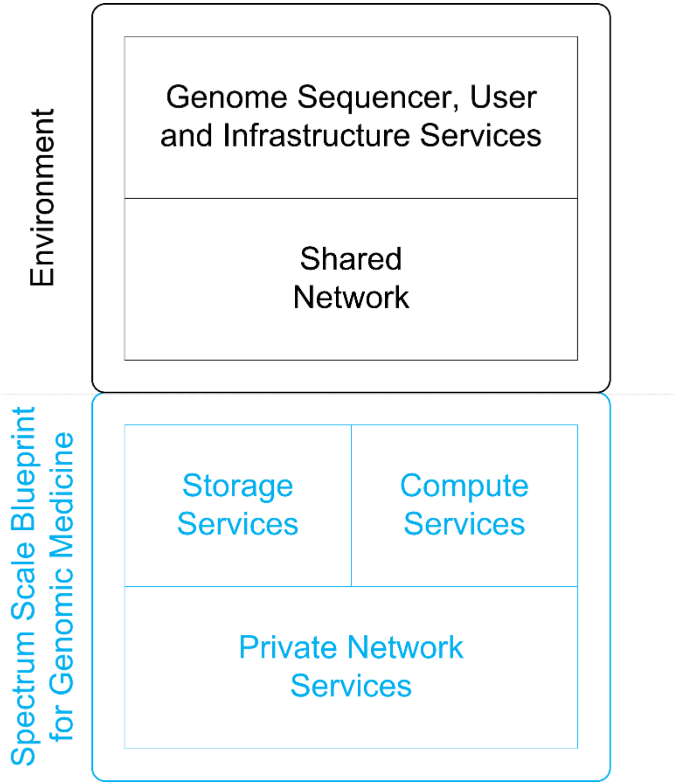
- Kevin is responsible for implementing the architecture defined by Aidan, the IT architect.
- Kevin is overwhelmed by the complexity.
- Kevin doesn't know the right way to implement it.
- Kevin feels like he's first and the only one doing it.
- Kevin opens PMRs to ask IBM what to do.
- Kevin has no idea how to monitor, tune or troubleshoot.

→ Kevin needs to install and configure the planned architecture to meet the workload demand.

→ Kevin needs to operate the solution and continue to deliver the SLA to keep stakeholders happy.

What	How
Improve healthcare patient outcomes.	Speed up multiple workloads and workflow performance. Make workload available.
Clients realize value from the purchase.	It works quickly.
Development reduces time in CritSit, planning, and support.	Repeatable engagement process resulting in less issues.
Sell what we have quickly (IBM and BP).	Create blueprints/reference architectures that show the value.
Articulate the competitive advantages and compare to competition.	Provide market collateral backed by the blueprint.
Attack a growing genomics medicine market.	Provide a competitive workload optimized solution.
Prerequisites and technical architecture are well understood.	Document best practices (service).
Customers have realistic and confident expectations.	Clearly communicate solution capabilities and skills required for success.
Sellers can identify and progress sales opportunities.	Understand how and what to position and what not.
Understand how to predict and tune performance.	Create best practices based on real workload and models on realistic hardware.
Simpler faster and more predictable deployment.	Simplified offering.

Composable Building Blocks



The Spectrum Scale Blueprint for Genomic Medicine Workload consists of expertly engineered, composable building blocks which include:

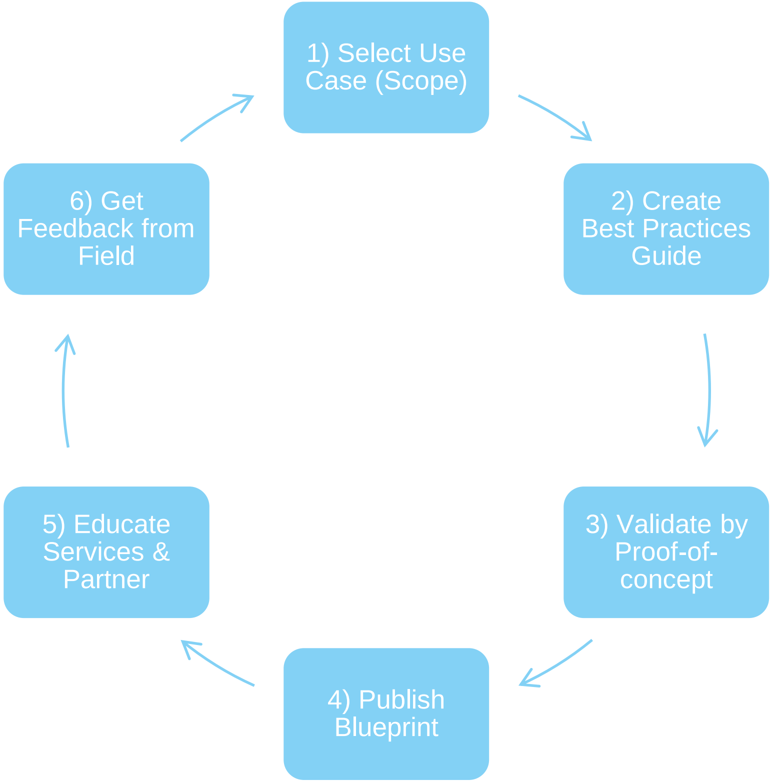
- **best practices guides** for architecture and configuration settings,
- **runbooks** which describes how to install, configure, monitor and upgrade example configurations,
- **sizing guidelines** which help to define a solution which meets the customers performance requirements,
- **deployment workshop** with client to customize solution to client's specific needs,
- **sales material** which enable seller to identify opportunities and create winning proposals,
- a **war room** where sellers and architects get easy access to subject matter experts.

Outline

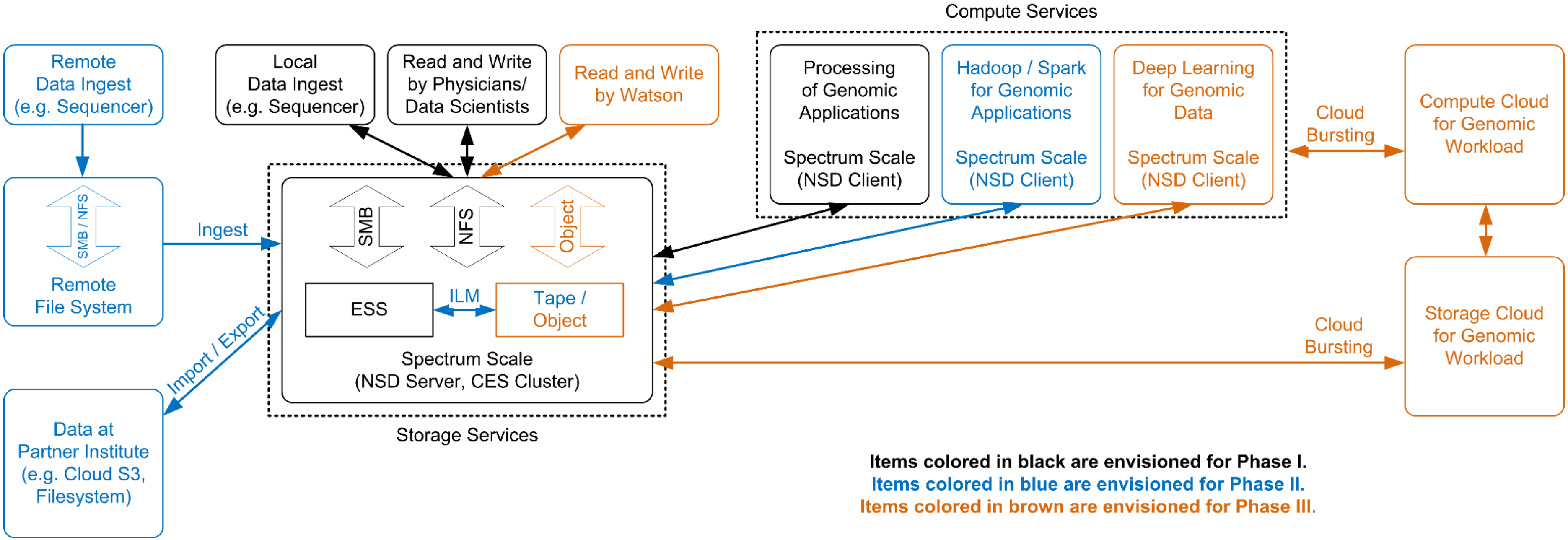
- Market Opportunity
- Composable Solution Architecture
- Driven by Design Thinking
- **Driven by Agile Development**
- Blueprint Capabilities
- Example Configuration

Staged Approach

- Blueprint for Genomic Medicine is an iterative approach.
- We want to get something out quickly, get feedback from the field and then refine.
- Three major phases
    I. Deliver Minimal Viable Product (MVP)
    II. Leverage existing Spectrum Scale features
    III. Enable hybrid cloud
- Each phase will have multiple iterations (see circle on the right)



Staged Approach



Items colored in black are envisioned for Phase I.
Items colored in blue are envisioned for Phase II.
Items colored in brown are envisioned for Phase III.

Outline

- Market Opportunity
- Composable Solution Architecture
- Driven by Design Thinking
- Driven by Agile Development
- ***Blueprint Capabilities***
- Example Configuration

Capabilities – Blueprint V1.0 – 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 Health Care and Life Science team.
➔ Blueprint capabilities are written in a product neutral language to emphasize end user requirement.

Capabilities – Blueprint V1.0 – 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 and for **sharing 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 Health Care and Life Science team.
➔ Blueprint capabilities are written in a product neutral language to emphasize end user requirement.

Capabilities – Blueprint V1.0 – 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 Health Care and Life Science team.
➔ Blueprint capabilities are written in a product neutral language to emphasize end user requirement.

Outline

- Market Opportunity
- Composable Solution Architecture
- Driven by Design Thinking
- Driven by Agile Development
- Blueprint Capabilities
- ***Example Configuration***

Example Configuration - Components

The Storage Cluster of the Example Environment consists of:

- 1x ESS Management Node (EMS) to manage the Storage Cluster,
- 1x IBM Elastic Storage Server (ESS) GS2S with SSD to store metadata,
- 1x IBM Elastic Storage Server (ESS) GL6S with NL-SAS to store genomics data,
- 3x CES Protocol Nodes for NFS and SMB to ingest and access genomics data.

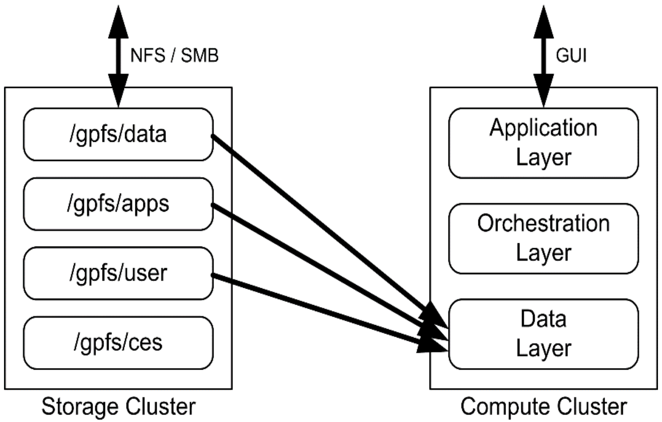
The Compute Cluster of the Example Configuration consists of:

- 2x Management Nodes to manage the Compute Cluster,
- 1x User Node (Power) for user login and to start batch jobs,
- 7x Compute Nodes (Power) to analyze genomics data,
- 2x Compute Nodes (Intel) to analyze genomics data.

The Private Network of the Example Configuration consists of:

- 2x InfiniBand EDR switches for the high-speed data network,
- 1x Gigabit Ethernet switch for the provisioning network and the service network.

Example Configuration – Logical View



All EMS, ESS and CES Protocol Nodes build a Spectrum Scale Storage Cluster.

Four Spectrum Scale filesystems are created:

- /gpfs/data to store genomic data,
- /gpfs/apps to store application binaries, scripts, configuration files and logs,
- /gpfs/user to store user data for the execution of batch jobs,
- /gpfs/ces to store metadata for the Cluster Export Services (CES).

The three CES Protocol Nodes build a CES Cluster and are configured to provide NFS and SMB services.

The /gpfs/data filesystem is exported via NFS and SMB for data ingest from devices such as genome sequencers and microscopes for access by data scientists and physicians.

All nodes of the Compute Cluster build a Spectrum Scale Compute Cluster.

The Spectrum Scale Compute Cluster imports /gpfs/data, /gpfs/apps, and /gpfs/user via Spectrum Scale multi-cluster remote cluster mount.

All compute resources provided by the Compute Nodes are managed by IBM Spectrum LSF to enable high-throughput execution of batch jobs.

Spectrum LSF provides a Workload Management GUI to submit and manage batch jobs to analyze genomics data.

➔ See the Best Practices Guide of each service for details.

IBM Spectrum Scale

Spectrum Scale Best Practices Guide for Genomic Medicine Workload 1.0 (Compute Services)

Dec 5th, 2017

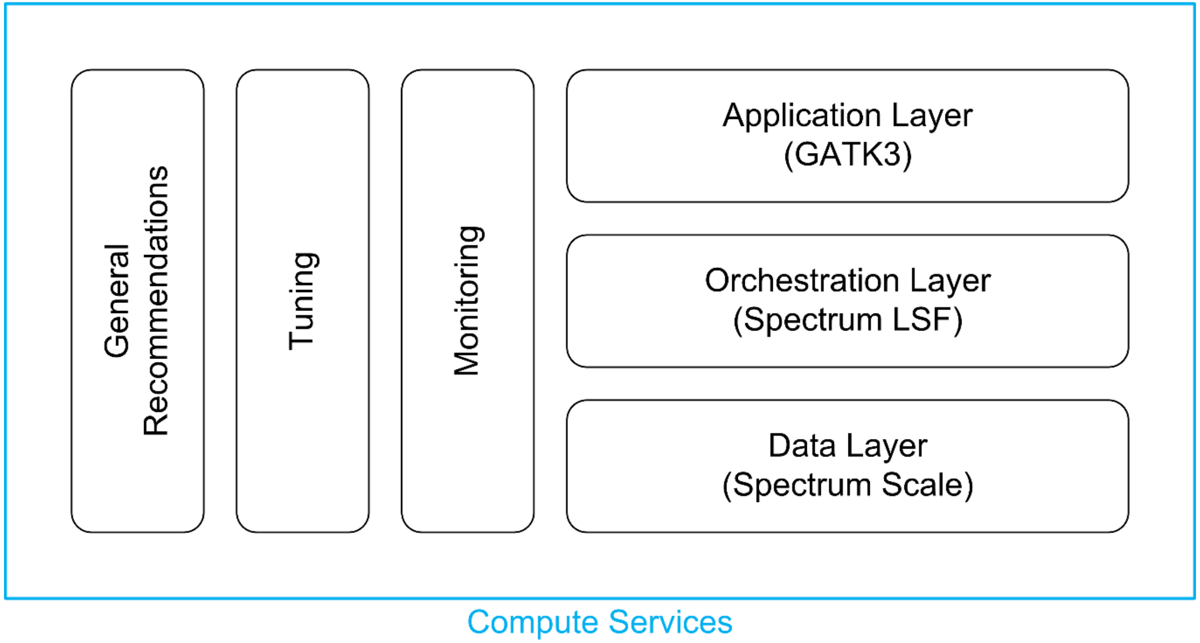
Summary

- The Spectrum Scale Blueprint for Genomic Medicine Workload describes Compute Services, Storage Services and Private Network Services. The next charts describe the Best Practices for Compute Services.
- The Spectrum Scale Blueprint for Genomic Medicine Workload is optimized for the “Broad Institute GATK Best Practices on IBM reference architecture”. Though, most of the recommendations are generic and apply to other workloads.
- The Spectrum Scale Blueprint for Genomic Medicine Workload uses IBM Spectrum LSF as workload scheduler. Though, most of the recommendations are generic and apply to other schedulers.
- Contact the Genomics War Room for help with different applications, different server architectures, and different schedulers.

Outline

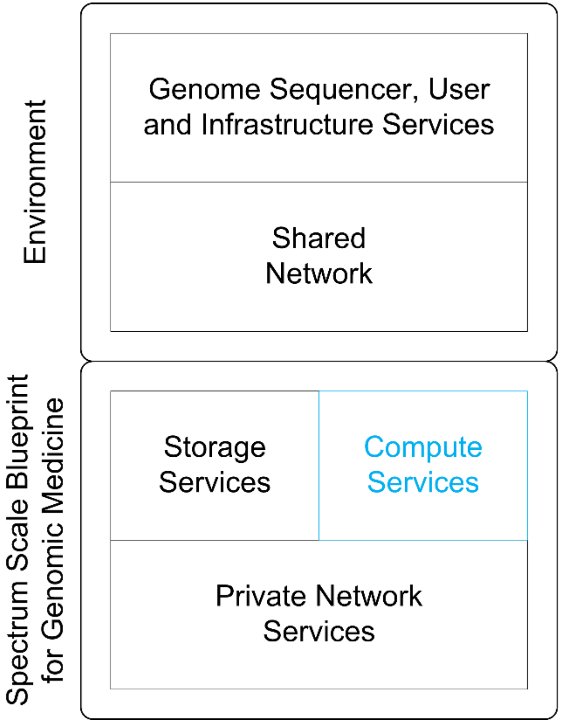
1. ***Composable building blocks***
2. Building block details

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.***

Compute Services – Capabilities



- 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

Compute Services – Solution Elements

Capability	Provided by
User GUI for physician/data scientist to submit and manage batch jobs	IBM Spectrum LSF – Application Center
User GUI for physician/data scientist to create and manage custom workflows	IBM Spectrum LSF – Process Manager
Workload Management GUI for IT administrator to view cluster status and utilization	IBM Spectrum LSF – Application Center
A Workload Scheduler enables high-throughput execution of batch jobs	IBM Spectrum LSF
Tuning Recommendations following the “Broad Institute GATK Best Practices on IBM reference architecture”	Spectrum Scale Best Practices Guide for Genomic Medicine Workload (This Guide)
Compute Nodes: Power and/or x86-64 as user nodes and for batch processing	IBM Spectrum LSF
User Nodes for physician/data scientist to log on and access the resources	IBM Spectrum LSF
Secure high-speed internal access to files stored on Storage Cluster	IBM Spectrum Scale – Remote Cluster Mount

Example Configuration

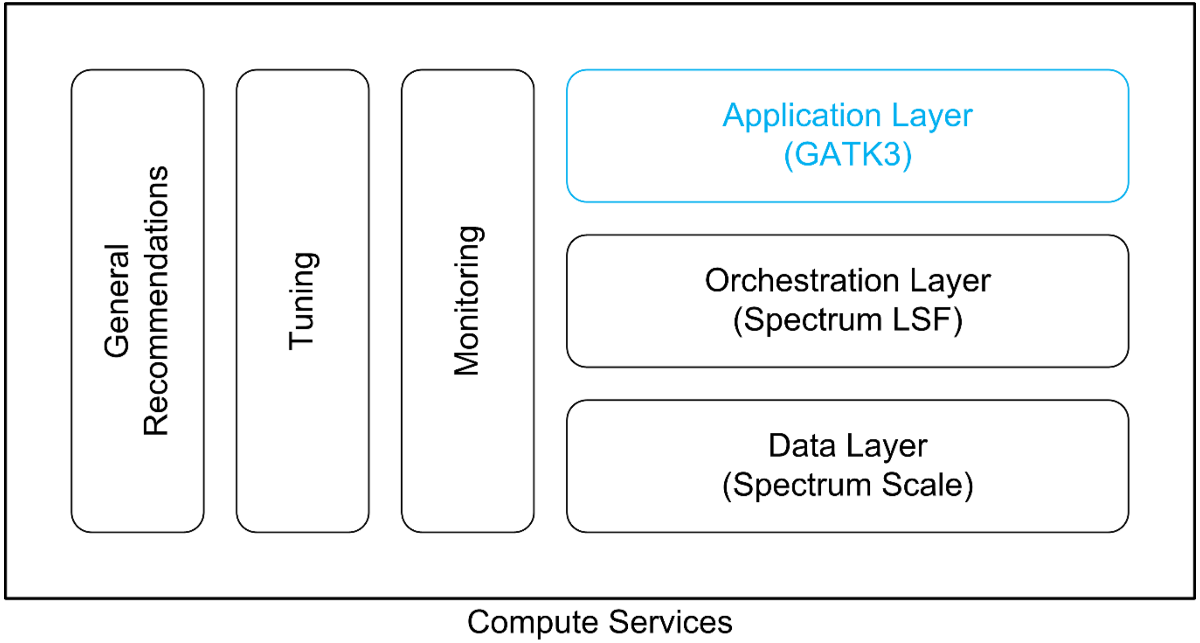
- In the following we describe the design decision for a Compute Cluster that comprises:
  - 2x Management Nodes
  - 1x User Node (Power)
  - 7x Compute Nodes (Power)
    - 4x nodes with 256GB RAM (default node)
    - 2x nodes with 512GB RAM (field feedback suggest to investigate 512GB RAM nodes)
    - 2x nodes with 1TB RAM (acceleration of sam2bam)
  - 2x Compute Worker Node (Intel)
- Software Levels
  - Spectrum Scale 4.2.3.5
  - Spectrum LSF 10.1.0.3
  - RHEL 7.3 Little Endian (LE)

Outline

IBM Spectrum Scale

1. Composable building blocks
2. ***Building block details***

Application Layer



➔ ***Broad Institute GATK comprises a widely used set of applications to analyze genomic data.***

Application: GATK3

- The “Broad Institute GATK Best Practices on IBM reference architecture” comprise a set of applications within a workflow for variant discovery analysis of both germline and somatic genomes.
  - Multi-step step workflow; each step has its own set of tools.
  - Output for each step is input to the next step.

IBM Systems
Technical White Paper

May 2017

Performance optimization of Broad Institute GATK Best Practices on IBM reference architecture for healthcare and life sciences

IBM’s commitment to enhance performance

Overview

The Genome Analysis Toolkit (GATK) Best Practices from the Broad Institute [1] has been widely adopted by the genomics community to perform variant discovery analysis of next-generation sequencing (NGS) data. A 30 times coverage of the whole human genome can take days to process using GATK Best Practices pipeline [2]. This paper describes how IBM has significantly accelerated the workflow on IBM reference architecture for healthcare and life sciences. It demonstrates that the GATK workflows can take advantage of the simultaneous multithreading (SMT) feature of IBM® POWER8® by parallelization of the GATK workflow. With the optimization on IBM’s reference architecture, it takes approximately 10 hours to complete GATK Best Practice pipeline for germline variant detection with 30 times coverage of the whole human genome using the GRCH37 reference genome and approximately 13 hours using the GRCH38 reference genome. These timings represent a significant speedup compared to the published Intel® results [2].

Overview

Challenge
Customers need faster turnaround time for processing the GATK best practices pipeline from the Broad Institute.

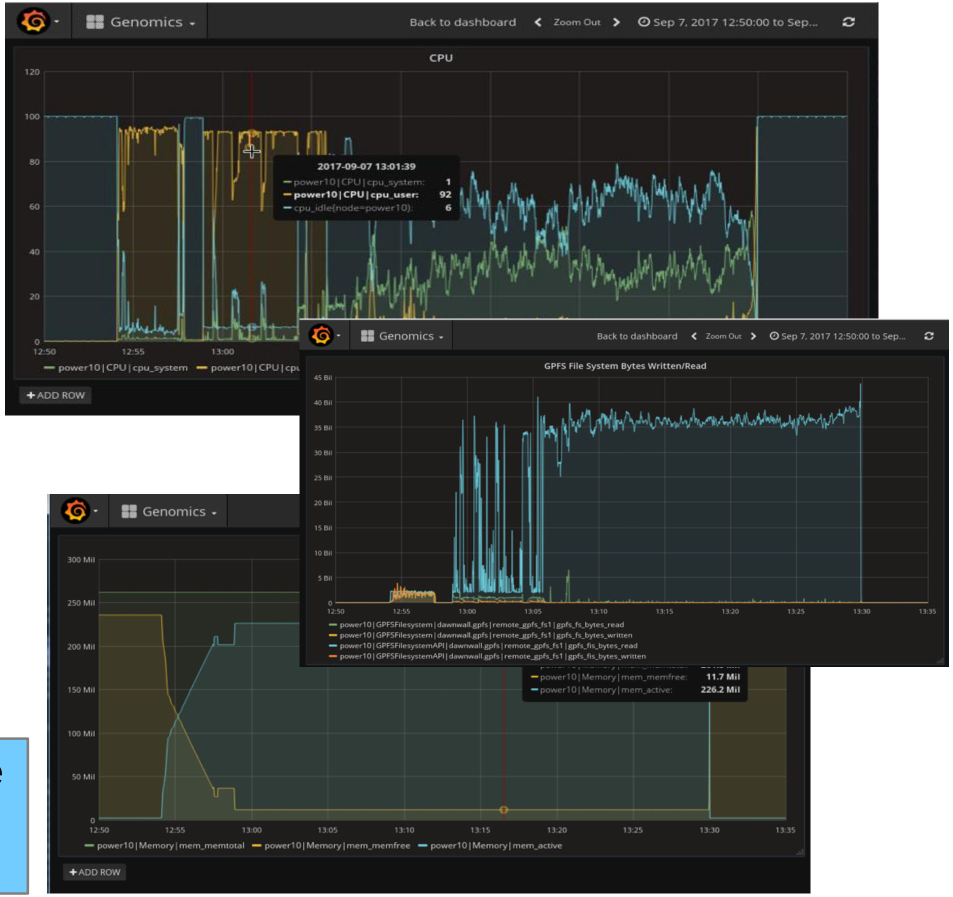
Solution
IBM has optimized performance of GATK Best Practices pipeline on the IBM POWER8 platform by taking advantage of unique features of IBM POWER8. Additionally, tools within the pipeline were parallelized for additional performance improvement. The IBM Spectrum Scale I/O performance supported the I/O acceleration to achieve superior results.

1

https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=TSW03540USEN

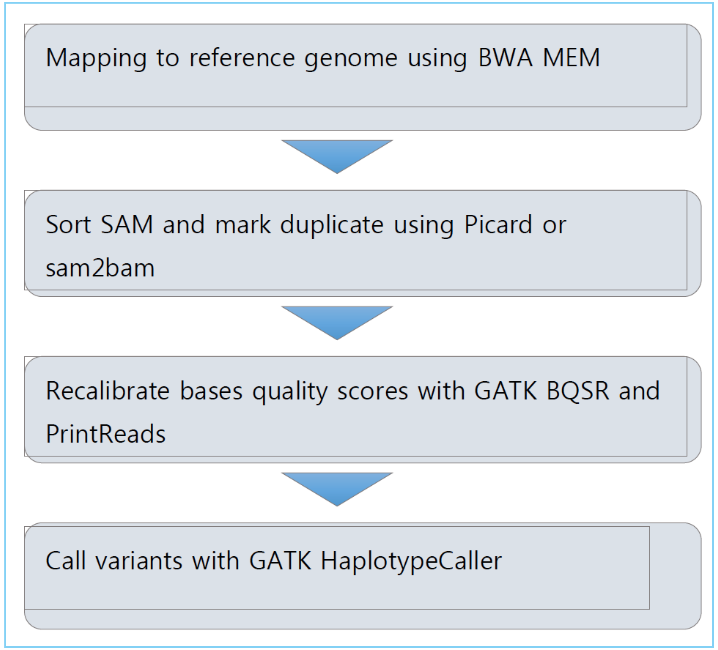
Application Profiling

- Profiling environment:
  - 1x Power8 Node (IBM 8247-22L with SMT=8) with 256GB memory to execute whole workflow.
  - 1x ESS GS4 storage based on SSD (>= 23 GB/s write bandwidth and >= 30 GB/s read bandwidth)
  - Dual rail FDR InfiniBand aggregating to ~13 GB/s
- The next charts provide a summary of the profiling. See backup section for more details.



➔ Most recommendations of this Best Practice Guide are generic and apply to other workloads and other server architectures (e.g. x86-64).

GATK Workflow – Execution Time on Profiling Environment



	Solexa WGS Broad dataset with b37 reference
BWA-Mem	303 min 47 sec
sam2bam (storage mode)	35 min 53 sec
GATK BaseRecalibrator (java setting -Xmn10g -Xms10g -Xmx10g)	87 min 21 sec
GATK PrintReads (java setting -Xmn10g -Xms10g -Xmx10g)	97 min 1 sec
GATK HaplotypeCaller (java setting -Xmn10g -Xms10g -Xmx10g)	261 min 37 sec
GATK mergeVCF (java setting -Xmn10g -Xms10g -Xmx10g)	0 min 51 sec

➔ Execution time was measured on the example configuration (see previous chart). The actual throughput or performance that any user will experience will vary depending upon many factors.

GATK Workflow – Profiling Summary

	BWA-Mem	sam2bam (storage mode)	GATK BaseRecalibrator	GATK PrintReads	GATK HaplotypeCaller	GATK mergeVCF
CPU	Intensive. Close to 100% CPU utilization	~93% (initial phase) and ~40% in later phases	~70% CPU utilization	~70% CPU utilization	~40% CPU utilization	Less than 1% CPU utilization
Memory	Low memory consumption	Higher memory consumption with ~223 GB consumed	Total of 18 x Java threads with each thread customized with 10 GB ➔ 180 GB	Total of 18 x Java threads with each thread customized with 10 GB ➔ 180 GB	Not memory intensive	Not memory intensive
File data I/O access pattern	Pattern of writes followed by reads, Predominantly sequential I/O.	Write I/O predominantly sequential I/O. Read I/O is random access in units of 512 KiB	Predominantly read intensive. Read is mix of sequential and random I/O	Mix of read and write. Write I/O is mostly 512 KiB with mix of sequential and random. Read is mostly sequential	Mix of read and write. Write I/O is mix of sequential and random. Read is mostly sequential	Mix of read and write. Read and write I/O is predominantly sequential I/O.

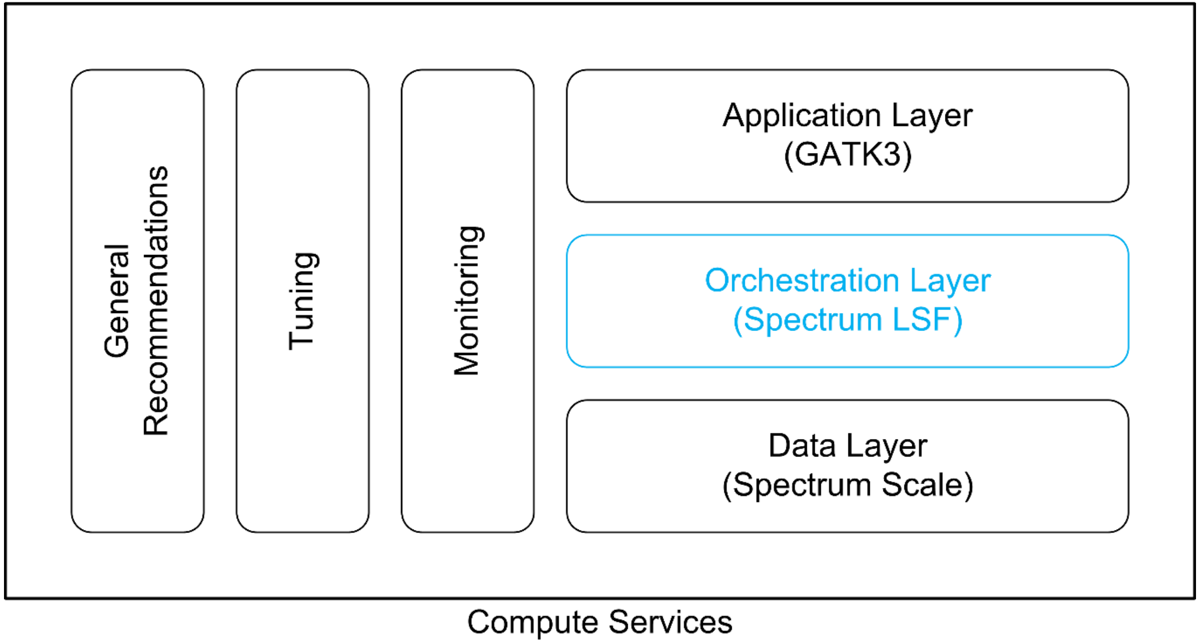
GATK Workflow – Profiling Summary (continued)

	BWA-Mem	sam2bam (storage mode)	GATK BaseRecalibrator	GATK PrintReads	GATK HaplotypeCaller	GATK mergeVCF
File I/O bandwidth	<= 200 MB/s (read and write)	Write < 2.5 GB/s. Sustained read < 300 MB/s. High degree of pagepool cache hits during reads (< 36 GB/s).	<= 100 MB/s (read and write)	Write < 150 MB/s and read < 75 MB/s.	Write < 100 MB/s and read < 100 MB/s.	Write < 1.5 GB/s and read < 2 GB/s.
File Metadata	<=2 inode updates	Initial phase <= 60 inode updates. Later phase, <=2 inode updates.	~24 file open and ~24 file closes.	~24 file open and ~24 file closes.	~20 file open and ~20 file closes.	~2 file open and ~2 file closes.
Output file(s)	Single output file (*.sam) <= 380 GB file size	Two output files. ~77 GB (.bam) and ~9 MB (.bam.bai).	Total of 52 files. 26 x ".table.log-4" files (<200 KB) and 26 x "*.table" files (< 300 KB)	Total of 78 files. 26 x ".recal_reads*.bam" files (< 15 GB), 26 x "*.bai" files (< 750 KB), and 26 x "*.recal_reads*.bam.log" files (< 200 KB)	Total of 78 files. 26 x ".raw_variants*.vcf" files (< 6 GB), 26 x "*.raw_variants*.vcf.log" files (< 400 KB), and 26 x "*.raw_variants*.vcf.idx" files (< 20 KB)	Single output file (*.raw_variants.vcf) with ~66 GiB file size

GATK Workflow – Derived Tuning Considerations

- BWA-Mem is CPU intensive. For optimal performance, execute this application on Compute Node with higher core count as well as higher clock frequency.
- sam2bam is memory intensive. For optimal performance, execute this application in memory mode on Compute Node with >= 1 TiB of memory.
- GATK is memory intensive. For optimal performance, execute this application on Compute Node with >= 512 GiB of memory.
- Separate filesystem metadata and data storage pools. Configure the data storage pool with larger Spectrum Scale filesystem block size (8 MiB).
- Configure networking for Spectrum Scale over low-latency and high throughput network interface.
- Apply all tuning described in this Best Practices Guide.

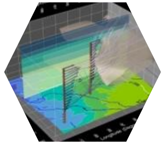
Orchestration Layer



➔ A Workload Scheduler enables high-throughput execution of batch jobs.
➔ IBM Spectrum LSF is used as example. Any other scheduler can be used.

IBM Spectrum LSF

High Performance Computing



IBM Spectrum LSF

Scalable, comprehensive workload management **accelerates throughput up to 150X** for simulation, design & research

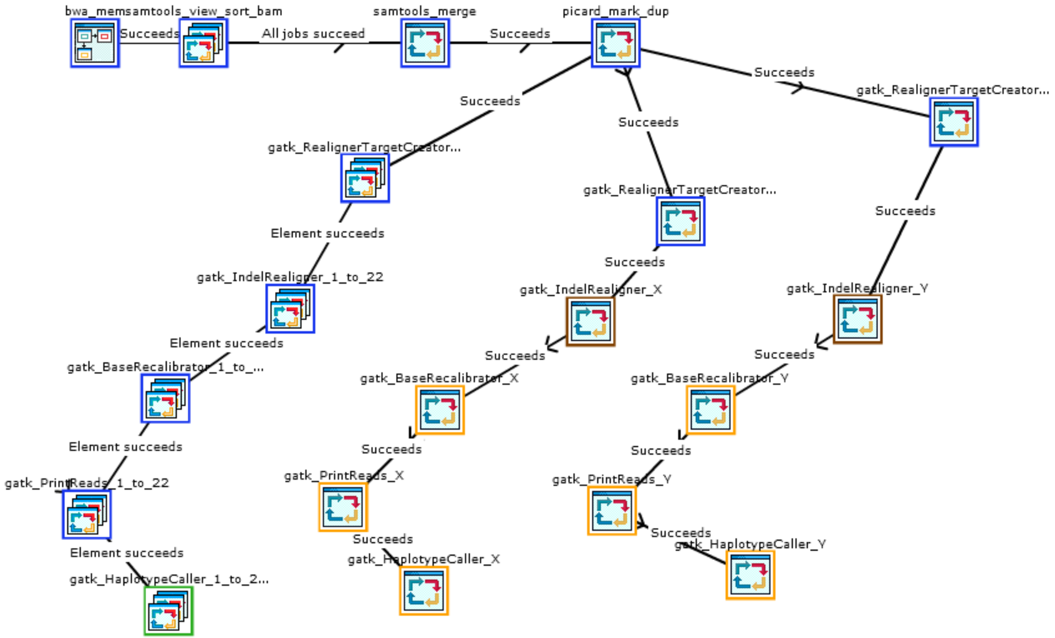
Selected add-ons to IBM Spectrum LSF:

- **IBM Spectrum LSF Application Center**
  A rich environment for building easy-to-use application-centric web interfaces, simplifying job submission, management and remote visualization. Use the web-based interface to remotely monitor jobs, access job-related data and perform basic operations
- **IBM Spectrum LSF Process Manager:**
  A powerful interface for designing complex engineering computational processes and capturing repeatable best practices that can be leveraged by other users. Integrate with IBM Spectrum LSF Application Center to create a consistent web-based environment.

➔ The Spectrum Scale Blueprint for Genomic Medicine Workload uses IBM Spectrum LSF as an example.
➔ Most of the recommendations are generic and apply to other schedulers too.

Flow Templates

- IBM Spectrum LSF Process Manager provides a GUI for creating, submitting and monitoring of complex workflows.
- IBM Spectrum LSF Process Manager refers a workflow as ‘flow’.
- IT administrators and end user (e.g. scientists) can use the GUI to create and modify flow templates.



Job Queues

- To keep the configuration simple, a single **default job queue (normal)** is recommended:
  - Different host capabilities (e.g. Power vs. x86-64, memory size, GPU available) will be specified in the lsb.hosts configuration file.
  - The flow templates specify the required host capability (in LSF jargon: resource requests).
  - Having a single job queue only moves the complexity of workflow optimization from the end user (e.g., physician) to the creator of the flow template (e.g., IT administrator).
- Depending on the customer requirements, additional job queues can be configured. Typical examples include and are not limited to a ‘high priority queue’ and an ‘admin queue’.
- Providing multiple job queues is outside the scope of the blueprint.

Filesystems

- A workload scheduler requires a shared file system (e.g., NFS, Spectrum Scale) to store binary files, configuration files and log files.
- The shared filesystem can become a performance bottleneck for Compute Clusters with a very large number of nodes and short running batch jobs.
- For Compute Clusters with up to 1,000 Compute Nodes that run genomics workload it is recommended to store all Spectrum LSF files in a Spectrum Scale filesystem. Storing those files in Spectrum Scale eliminates the need for an external NFS service. This simplifies the configuration and reduces costs.
- The Spectrum Scale filesystem for Spectrum LSF files should be separated from the application data (e.g., genomic data sets) to tune for different I/O patterns and to isolate the respective I/O loads from impacting each other.
- Writing files from different Compute Nodes into the same directory triggers underlying GPFS Token Traffic to keep the directory structure consistent across all nodes. That impacts performance. Having a dedicated sub directory per Compute Node eliminates this bottleneck.
- See the IBM Platform LSF Best Practices Guide for 'IBM Platform LSF 9.1.3 and IBM GPFS in Large Clusters' for alternative configuration options.
  https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/New%20IBM%20Platform%20LSF%20Wiki/page/LSF%20best%20practices%20&%20tips

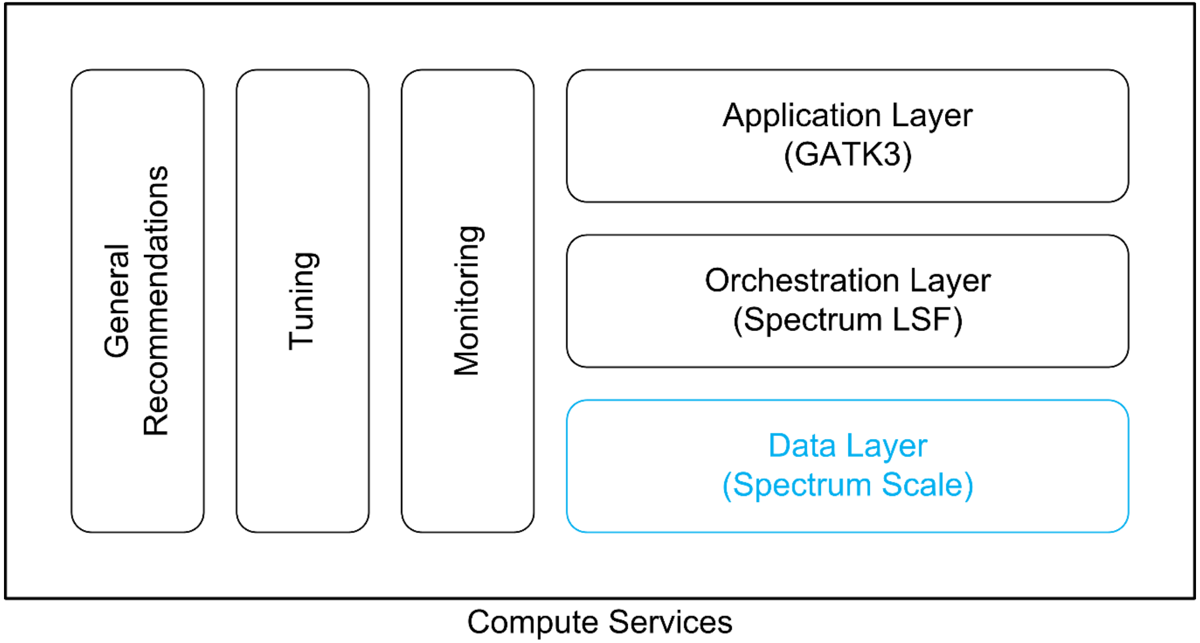
Directories and Filesets for Spectrum LSF

Purpose	Variable Name	Variable Value	Used by	Comment
Full path to the top level LSF installation directory	LSF_TOP	/gpfs/app/lsf	All nodes	-
Directory in which the job history and accounting logs are kept for each cluster	LSB_SHAREDIR	/gpfs/app/lsf/work	Master host	-
Defines the LSF system log file directory (*)	LSF_LOGDIR	/gpfs/app/lsf/log/%H	All nodes	Dedicated sub directory per host
Specifies the directory for buffering batch standard output and standard error for a job (*)	JOB_SPOOL_DIR	/gpfs/app/lsf/log/%	Execution hosts	Dedicated sub directory per host
Cluster-wide current working directory (CWD) for the job	DEFAULT_JOB_CWD	/gpfs/app/lsf/cwd/%H	Execution hosts	Dedicated sub directory per host
Specifies the path and directory for temporary LSF internal files (*)	LSF_TMPDIR	/gpfs/app/lsf/tmp/%H	Execution hosts	Dedicated sub directory per host

(*) Spectrum Scale independent filesets will be configured for /gpfs/app/lsf/log, /gpfs/app/lsf/spool, and /gpfs/app/lsf/tmp.

➔ Having dedicated sub directories per Compute Node eliminates potential GPFS Token Traffic.
➔ Spectrum Scale independent filesets enable automated data management on the Storage Cluster.

Data Layer

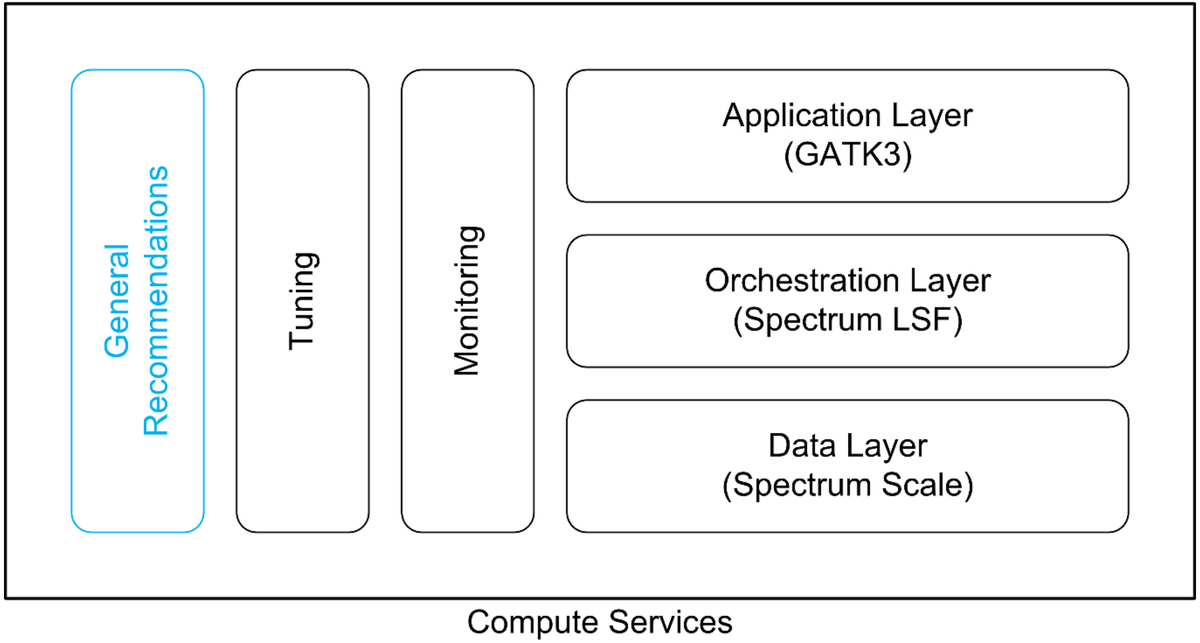


➔ ***Spectrum Scale provides secure and high-speed access to files that are stored and managed on the Storage Cluster.***

Spectrum Scale

- All Compute Nodes build a Spectrum Scale cluster.
  - Each Compute Cluster Node will be configured as Spectrum Scale Node.
- All Compute Nodes will be configured to start Spectrum Scale on boot.
  - *mmchconfig autoload=yes –N <compute_node_class>*
- The Compute Cluster will import the following Spectrum Scale Filesystems from the Storage Cluster via GPFS multi-cluster remote cluster mount:
  - /gpfs/data → genomic data and analysis results
  - /gpfs/app → application binaries, configuration files and log files
  - /gpfs/user → user data for execution of batch jobs (optional)
- All Spectrum Scale Filesystems will be configured and managed on the Storage Cluster.
  - See the Best Practices Guide for Storage Services for details.
- All Compute Cluster nodes and all Storage Cluster nodes need to be connected via a high-speed, low-latency network.
  - See the Best Practices Guide for Private Network Services for details.

General Configuration Recommendations



➔ ***Best practices increase the operational efficiency for managing the whole compute infrastructure.***

Node Designation

	Compute Node Type	Memory	End User Login	Spectrum Scale Node	Spectrum Scale Quorum	Spectrum Scale Manager	Spectrum Scale Admin	Spectrum Scale GUI	Spectrum LSF Node Type	Spectrum LSF AC	Spectrum LSF PM
Power 1	Management (Primary)	256GB	No	X	X	X	X	X	Master	(X)	(X)
Power 2	Management (Standby)	256GB	No	X	X	X	X	X	Master (Stand-by)	X	X
Power 3	User Login	256GB	Yes	X	X				Submission		
Power 4	Worker	256GB	No	X					Execution		
Power 5	Worker	256GB	No	X					Execution		
Power 6	Worker	256GB	No	X					Execution		
Power 7	Worker	512GB	No	X					Execution		
Power 8	Worker	512GB	No	X					Execution		
Power 9	Worker	1024GB	No	X					Execution		
Power 10	Worker	1024GB	No	X					Execution		
Intel 1	Worker	256GB	No	X					Execution		
Intel 2	Worker	256GB	No	X					Execution		

Node Types

Management Nodes

- Runs all services to dispatch and manage batch jobs
    - Scheduler
    - GUI to submit and manage batch jobs
    - GUI to create and manage custom workflows
    - Workload Management GUI to view cluster status and utilization
- Login restricted to administrative users
- Most stable nodes and therefore good candidates to run additional infrastructure services

User Login Node

- User login to compile applications, submit jobs and flows via command line interface (CLI)
- Stable nodes and therefore reasonable candidates to run additional infrastructure services

Worker Node

- Execute batch jobs as dispatched by the scheduler
- Login restricted to administrative users
- Can get unstable when end users experiment with new applications.

Node Designation – Spectrum Scale

Quorum Nodes

- The general recommendation is to define three or five quorum nodes, but there is no single correct answer how many **Quorum Nodes** should be configured.
- The Spectrum Scale Nodes which assume the role of a **Quorum Node** needs to be on reliable nodes, as much as possible.
- Each **Quorum Node** should have independent failure domain to avoid single point of failure, e.g. different power circuit, different rack, different network switch.
- Each **Quorum Node** will automatically become a **Config Server**.

Manager Nodes

- Spectrum Scale has a capability to define which nodes can assume the role of a **Manager Node**.
- Spectrum Scale will automatically assign the following roles to the available Manager Nodes:
  Cluster Manager, Filesystem Manager, Token Manager.

Node Designation – Spectrum Scale

Admin Nodes

- Spectrum Scale **Admin Nodes** are responsible for issuing any and all Spectrum Scale administrative commands.
- Spectrum Scale commands maintain the appropriate environment across all nodes in the cluster.
- The **Admin Nodes** have similar requirements as the Management Nodes: password less root ssh and scp to all other Spectrum Scale Nodes, access restricted to administrative users only.
- For redundancy, it is best, if possible to have at least two Spectrum Scale Nodes that are **Admin Nodes**.

GUI Nodes

- The Spectrum Scale **GUI Nodes** are always Admin Nodes.
- The GUI does not allow root login. Only an admin login exists.
- The GUI subsystem passes commands as root to the other Spectrum Scale Nodes of the cluster.
- Most, but not all, Spectrum Scale functions can be run from the GUI, so occasionally, some commands require root login for CLI access.
- All GUI Nodes run a performance monitoring collection daemon that is used by the GUI to report cluster health and performance.

Node Designation – Spectrum LSF

What node types does Spectrum LSF support?

- **Master host:** LSF server host that acts as the overall coordinator for the cluster, doing all job scheduling and dispatch.
- **Server host:** A host that submits and runs jobs.
- **Client host:** A host that only submits jobs and tasks.
- **Execution host:** A host that runs jobs and tasks.
- **Submission host:** A host from which jobs and tasks are submitted.

→ To keep the configuration simple, we use Master hosts, Submission Hosts and Execution hosts only.

Spectrum Scale Master Host

- LSF allows to configure multiple master host candidates.
- There is only one concurrent active master node. LSF has built-in failover, in case current master node fails.

Recommendation for blueprint

- Configure first Management Node as **LSF Master host**.
- Configure second Management Node as **LSF Master host candidate**.
- Configure all other nodes as **LSF Execution host**.

Node Designation – Spectrum LSF Application Center

General Considerations

- IBM Platform Application Center (PAC) was renamed to IBM Spectrum LSF Application Center.
- IBM Spectrum LSF Application Center is an add-on to IBM Spectrum LSF that provides a WebUI for jobs submission, job monitoring and basic LSF Cluster management.
- IBM Spectrum LSF Application Center has no built-in capabilities for fail-over to stand-by server.

Recommendation for Blueprint

- Configure second Management Node as active **LSF Application Center Server**.
- Configure first Management Node as stand-by **LSF Application Center Server**.
  - The binaries and configuration files are installed. They need to be started manually if the active fails.

Node Designation – Spectrum LSF Process Manager

General Considerations

- IBM Platform Process Manager (PPM) was renamed to IBM Spectrum LSF Process Manager.
- IBM Spectrum LSF Process Manager is an add-on to IBM Spectrum LSF that provides a WebUI for flow creation and flow management.
- IBM Spectrum LSF Process Manager has no built-in capabilities for fail-over to stand-by server.

Recommendation for Blueprint

- Configure second Management Node as active **LSF Process Manager Server**.
- Configure first Management Node as stand-by **LSF Process Manager Server**.
  - The binaries and configuration files are installed. They need to be started manually if the active fails.

External Dependencies

Spectrum Scale depends on highly available Name Resolution Services (**DNS**) for name resolution and reverse name resolution.

- **Each Compute Node** needs to connect to the customer provided **DNS** service.
- In most cases the customer already has such a service. Otherwise such a service must be configured.

Spectrum Scale depends on Time Services (**NTP**) for time synchronization:

- **Each Compute Node** needs to connect to the customer provided **NTP** service.
- In most cases the customer already has such a service. Otherwise such a service must be configured.

Certain user and administrative commands depend on proper **ID Mapping**:

- **Each Compute Node** needs to connect to the customer provided **ID Mapping** service.
- In most cases the customer already has such a service. Otherwise such a service must be configured.

It is best practice that each Compute Node contacts those customer provided infrastructure services via the Compute Cluster Management Nodes.

- Some customers prefer that the Compute Cluster Management Nodes runs an instance of each service and connect it to the respective customer provided service.
- Some customers prefer to route respective network traffic via the Compute Cluster Management Nodes to the external server.
- The blueprint supports both approaches.

Server Deployment and Management

- Customers typically have an infrastructure to install and manage servers to:
  - Automatically install and configure the operating system,
  - Automatically monitor and report hardware failures.
- There is a broad variety of tools available and used by customers.
- In most cases the customer already has such a service. Otherwise such a service must be configured.
- The runbooks illustrate one example for automated installation and configuration of compute sever.

Miscellaneous

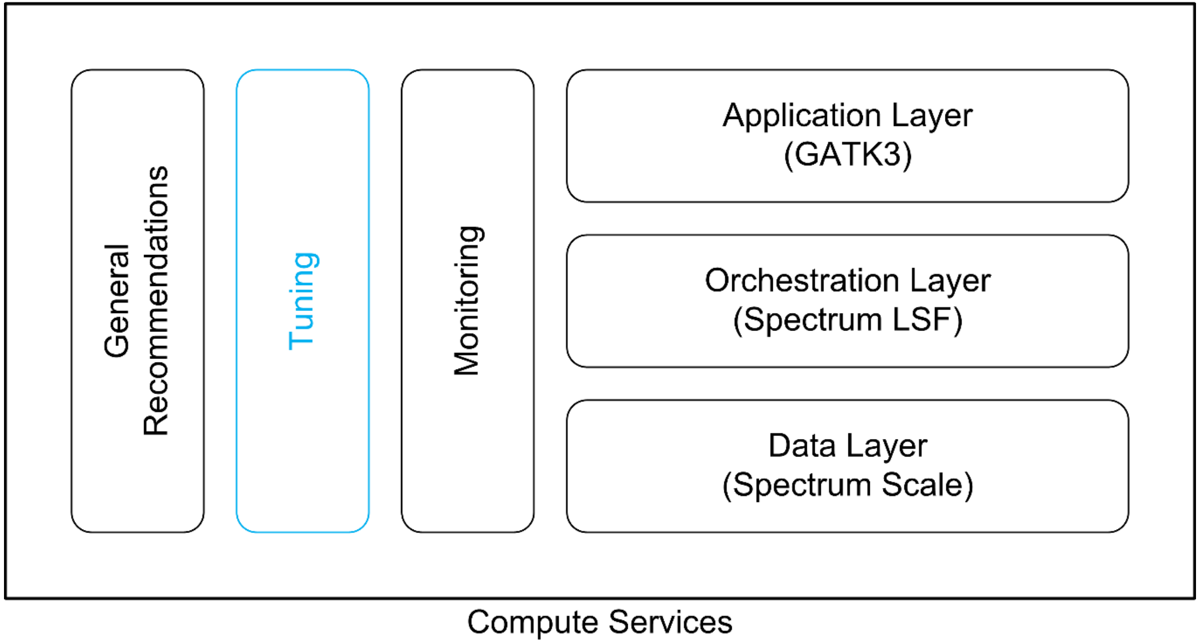
General Considerations – Spectrum Scale

- All nodes of same cluster need to be able to communicate to each other.
    - See Best Practices Guide for Private Network Services for details.
- All nodes of compute cluster needs to be able to communicate to all nodes of the storage cluster.
    - See Best Practices Guide for Private Network Services for details.
- All nodes used for administering Spectrum Scale must be able to do ssh and scp on any other node in the cluster as user root without the use of a password.
- Sudo wrappers will not used in order to use the same approach as for the Storage Services.

General Considerations – Spectrum LSF

- Spectrum LSF spawns dispatched jobs with the UID and GID of the user who submitted the job.
- All LSF hosts need to be configured with user and group information of lsfadmin and all LSF user from external authentication source (e.g. AD or LDAP).
- User and group information for LSF and for SMB and NFS (Storage Services) needs to be consistent.

Tuning



➔ The tuning recommendations are optimized for the "Broad Institute GATK Best Practices on IBM reference architecture". Though, most settings are generic.

Tuning Guidelines – Compute Nodes

OS tunable

- ulimit (Include the following in /etc/security/limits.conf)

  ** soft memlock unlimited*

  ** hard memlock unlimited*

  ** soft nofile 16384*

  ** hard nofile 16384*

  [detailed output in the notes]

- tuned configuration
  - /etc/tuned/active_profile is set to "throughput-performance"
  - /usr/lib/tuned/throughput-performance/tuned.conf

    [cpu]

    governor=performance

    energy_perf_bias=performance

    min_perf_pct=100

    [detailed output in the notes]

➔ This set of tunables is best practice for Spectrum Scale and needs to be applied for genomic workload.

Tuning Guidelines – Compute Nodes

Network tunable

- On Mellanox Adapters,
  apply Mellanox OFED Tunings
  https://community.mellanox.com/docs/DOC-2489

- /etc/sysctl.conf

  net.ipv4.tcp_timestamps=0
  net.ipv4.tcp_sack=0
  net.core.netdev_max_backlog=250000
  net.core.rmem_max=16777216
  net.core.wmem_max=16777216
  net.core.rmem_default=16777216
  net.core.wmem_default=16777216
  net.core.optmem_max=16777216
  net.ipv4.tcp_rmem=4096 87380 16777216
  net.ipv4.tcp_wmem=4096 65536 16777216
  net.ipv4.tcp_low_latency=1
  net.ipv4.tcp_adv_win_scale=2
  net.ipv4.tcp_window_scaling=1
  net.core.somaxconn = 8192
  vm.min_free_kbytes = 512000
  kernel.sysrq = 1
  kernel.shmmax = 137438953472

➔ This set of tunables is best practice for Spectrum Scale and needs to be applied for genomic workload.

Tuning Guidelines – Compute Nodes

Spectrum Scale tunables (Compute nodes will be based on version 4.2.3.5 or later PTF)

- Since the storage backend is ESS, apply gssClientConfig.sh (Node-ems:Dir-/usr/lpp/mmfs/samples/gss) on the compute_node_class with pagepool set to 16 GiB
  gssClientConfig.sh -P 16384 <compute_node_class>
- On InfiniBand networking, enable GPFS verbsRdma and verbsPorts to the correct IB HCA/ports
  mmchconfig maxFilesToCache=32K –N <compute_node_class>
  mmchconfig maxMBpS=20000 –N <compute_node_class>
  mmchconfig socketMaxListenConnections=8192 –N <compute_node_class>
  mmchconfig envVar="MLX4_USE_MUTEX=1 MLX5_SHUT_UP_BF=1 MLX5_USE_MUTEX=1“ –N <compute_node_class>
- On all Compute Nodes, increase the file size to cache to improve the performance of certain Genomics applications (e.g. bcl2fastq)
  mmchconfig seqDiscardThreshold=8M –N <compute_node_class>
  mmchconfig writebehindThreshold=8M –N <compute_node_class>
  Average file size for bcl2fastq is 3-7 MB. Setting the thresholds to 8MB improves the application performance owing to file-data caching.

➔ This set of tunables is optimized for IBM Elastic Storage Server (ESS) and Broad Institute GATK3.

Tuning Guidelines – Compute Nodes

blue ~ tuning applied for genomic workload
grey ~ default settings for ESS client node

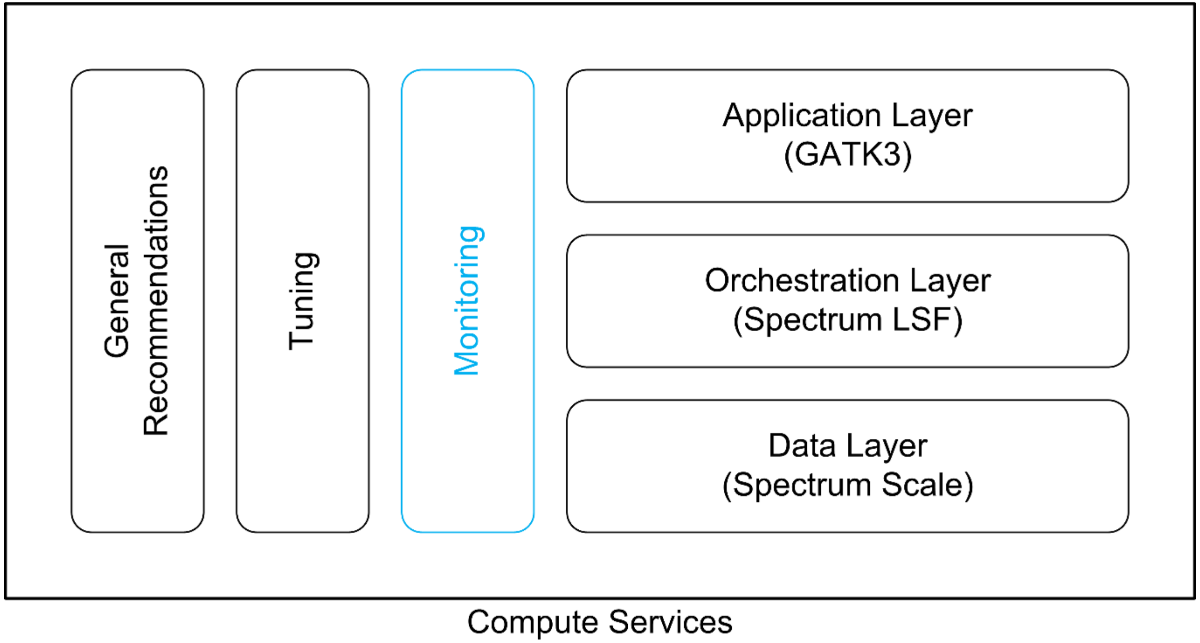
Snip of mmlsconfig:

[compute]
pagepool 16384M
numaMemoryInterleave yes
maxFilesToCache 32k
maxStatCache 0
maxMBpS 20000
workerThreads 1024
ioHistorySize 4k
verbsRdma enable
verbsRdmaSend yes
verbsRdmasPerConnection 256
verbsSendBufferMemoryMB 1024
ignorePrefetchLUNCount yes
scatterBufferSize 256k
nsdClientCksumTypeLocal ck64
nsdClientCksumTypeRemote ck64
socketMaxListenConnections 8192
envVar MLX4_USE_MUTEX=1 MLX5_SHUT_UP_BF=1 MLX5_USE_MUTEX=1
verbsPorts <active_verbs_ports>
seqDiscardThreshhold 8M
writebehindThreshhold 8M

[common]
cipherList AUTHONLY
adminMode central

➔ This set of tunables is optimized for IBM Elastic Storage Server (ESS) and Broad Institute GATK3.

Management of Compute Services



➔ *User GUI 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*

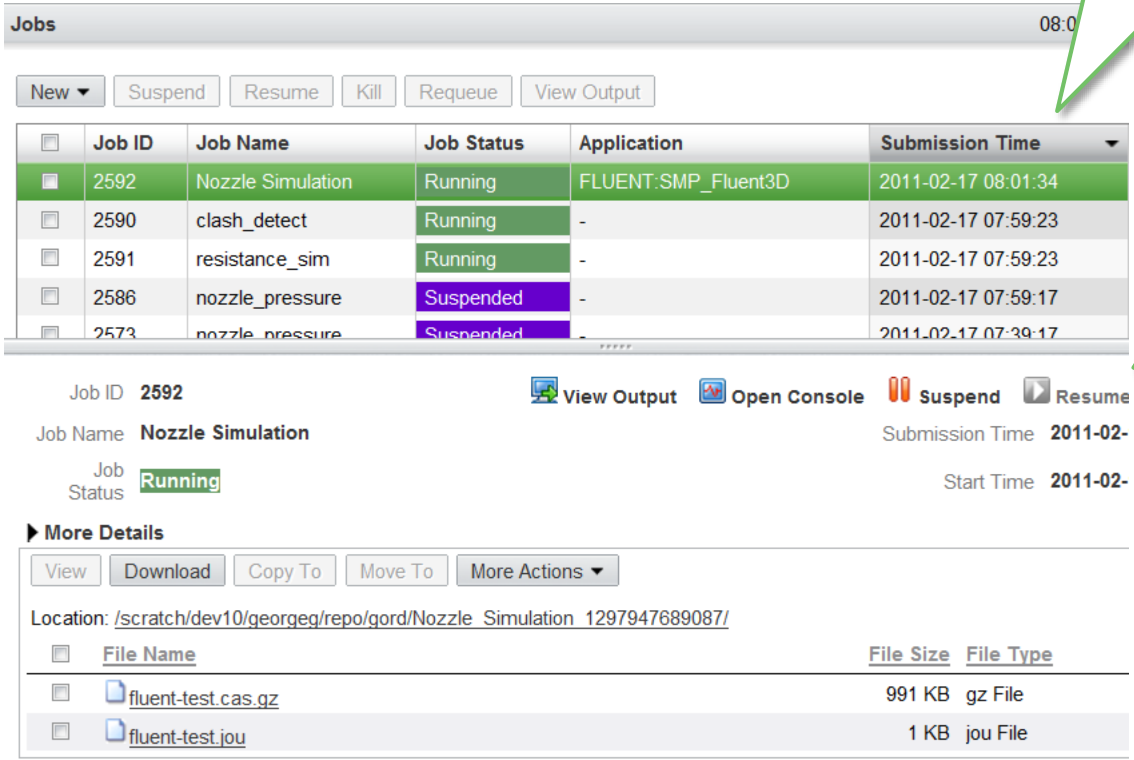
IBM Spectrum LSF Application Center

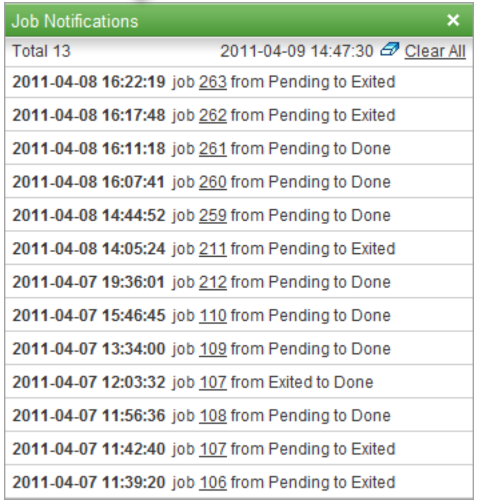
Monitor and Manage Jobs

Users can monitor and manage jobs from any device with a browser.

Upload local data or access sharable server side repositories and improve collaboration.

Proactive notification of job status changes makes application users more efficient.



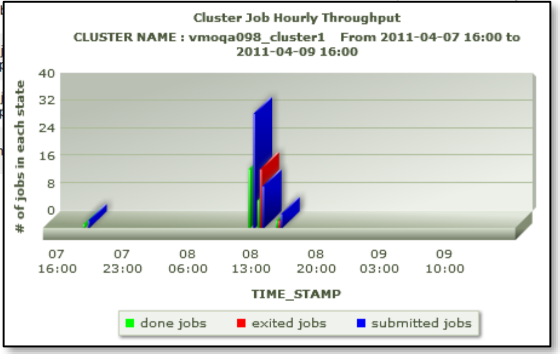


IBM Spectrum LSF Application Center

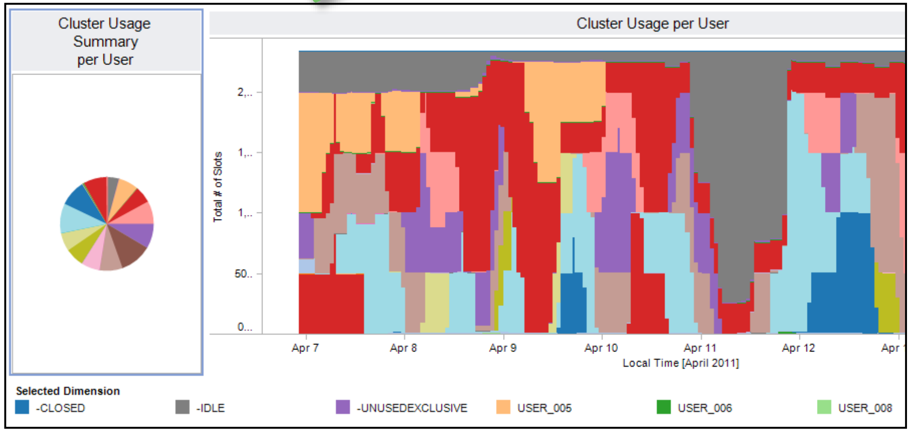
Integrated Reporting

Extensive library of built-in, relevant reports related to resource usage and jobs.

Report	Summary	Category
Host Resource Usage	Resource usage trends for selected hosts.	LSF
Active Job States Statistics by Queue	Number of active jobs in each active job state in a selected queue	LSF
License Usage	The license usage under License Scheduler. You can only produce this report if you use LSF License Scheduler.	LSF License Scheduler
Cluster Availability - LSF	LSF host availability in a LSF cluster.	LSF
Cluster Job Hourly Throughput	Number of submitted, exited, and done jobs in a cluster.	LSF
Cluster Job Slot Utilization	Job slot utilization levels in your cluster.	LSF
Job Slot Usage by Application Tag	Job slots used b	
Jobs Forwarded to Other Clusters	The number of j produce this rep	
Jobs Received from Other Clusters	The number of j produce this rep	
Performance Metrics	Internal perform	



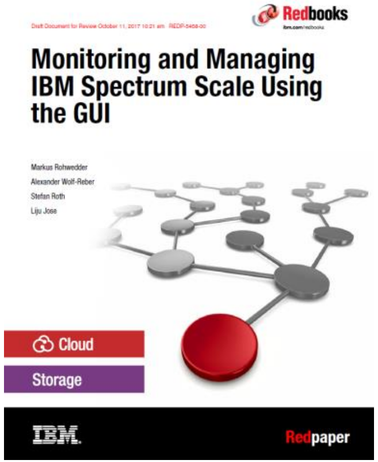
Access reporting and analysis functions directly through IBM Spectrum LSF Application Center.



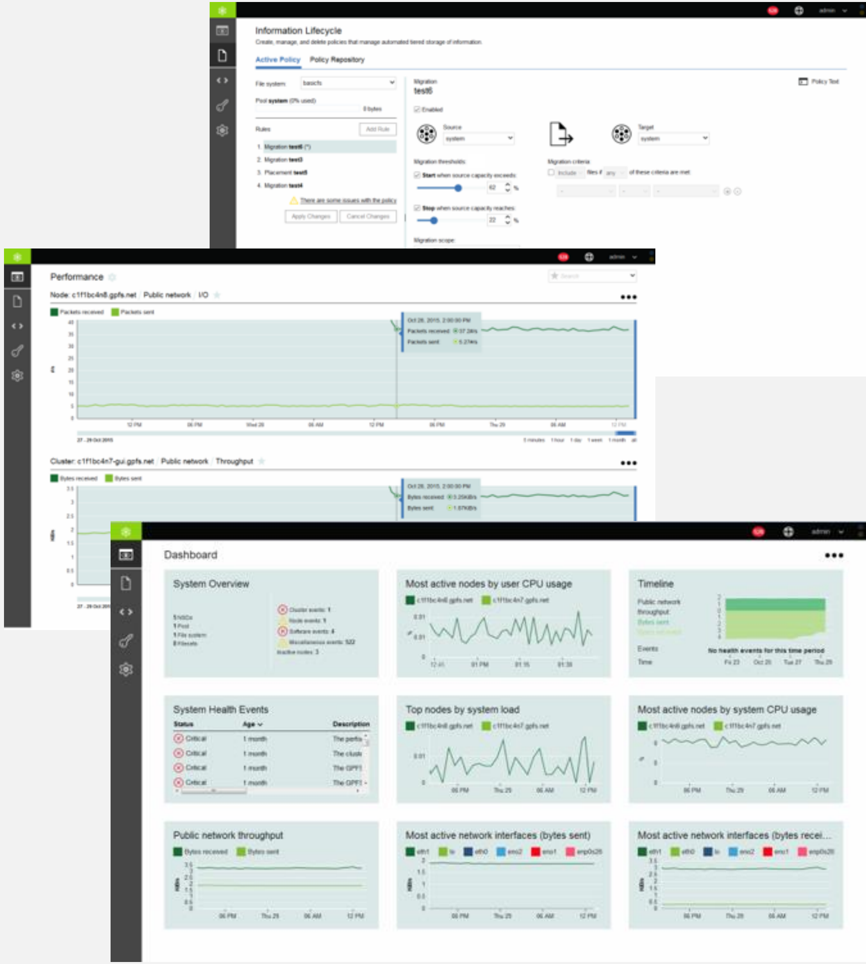
Spectrum Scale GUI

- Reduces administration overhead
  - Graphical User Interface for common tasks
  - Guided interfaces for common tasks
  - Supports Spectrum Scale and ESS
- See Redpapers for Monitoring Best Practices

http://www.redbooks.ibm.com/abstracts/redp5418.html

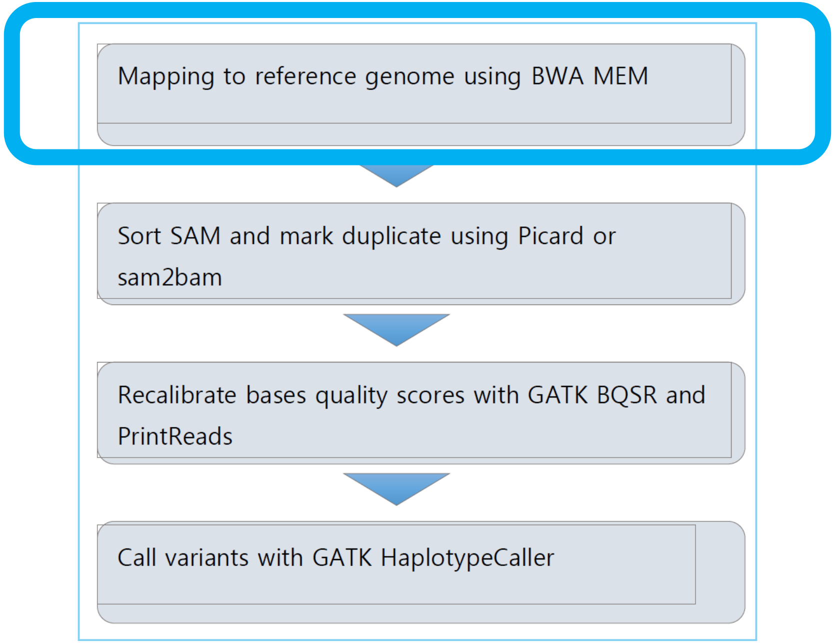


http://www.redbooks.ibm.com/redpieces/abstracts/redp5458.html



BACKUP

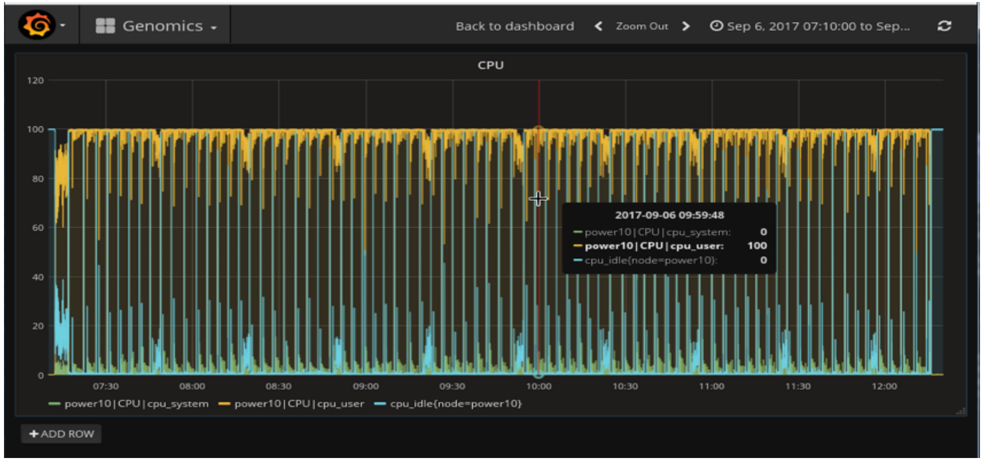
Application Workflow



Tool	BWA
Version	0.7.15-0 used for profiling
Source	https://biobuilds.org/tools-in-biobuilds/biobuilds-2016-11/

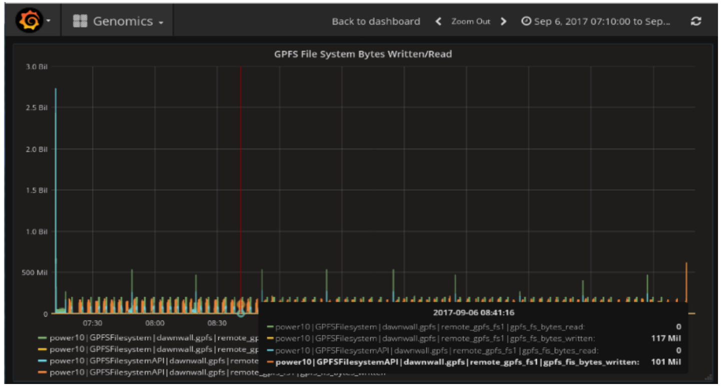
Application Profiling – BWA Mem

- Is CPU intensive (%user consuming close to 100%). Faster CPU can improve the overall runtime (Number of threads launched = 160 <=== ncpu=20; ${bwa_dir}/bwa mem -t $((${ncpu} * 8)).
- Not memory intensive.
- The I/O pattern: a pattern of writes followed by reads. Average bandwidth for write and read is within around 200 MB/s. Write is sequential I/O (WritebehindWorkerThread) and Read is sequential I/O (PrefetchWorkerThreads). The fs block size is 16 MiB and we see "dump iohist" nSec is 32768 sectors.

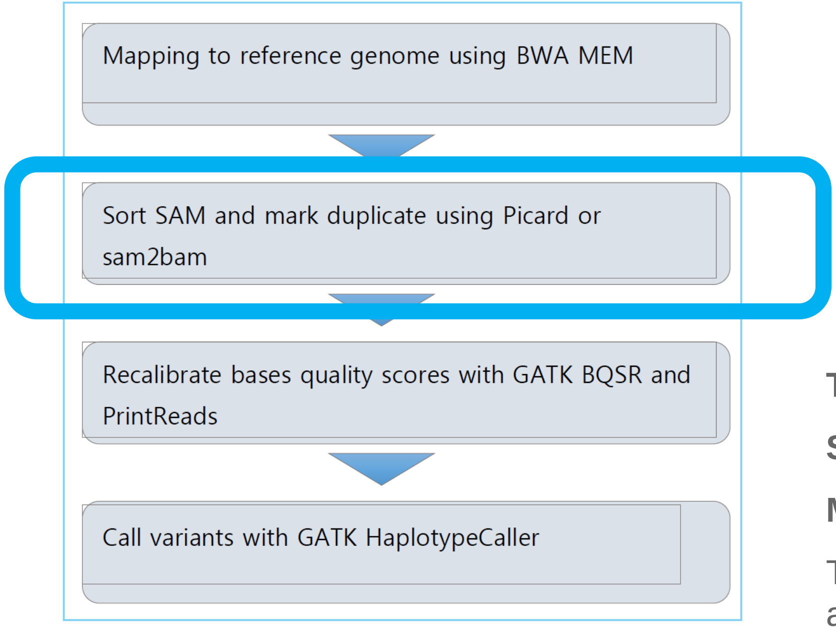


Input file format:
.fastq or .fq

Output file format:
.bwa.sam



Application Workflow



Tool	sam2bam
Version	1.2-157 used for profiling
Source	https://github.com/OpenPOWER-HCLS/sam-to-bam

Two modes supported

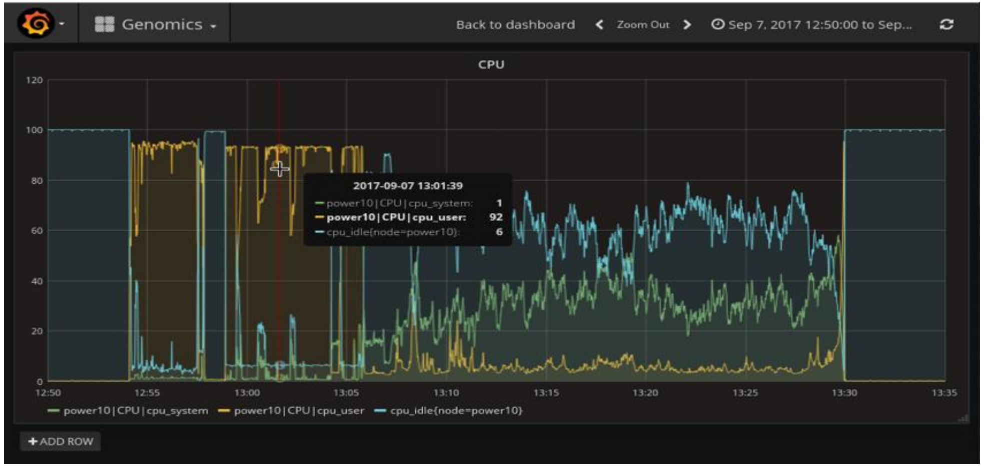
Storage Mode – Only if 1 TB of memory is not availble

Memory Mode – Default

The POWER8 processor makes use of a large number of on- and off-chip memory caches to reduce memory latency and generate very high bandwidth for memory and system I/O.

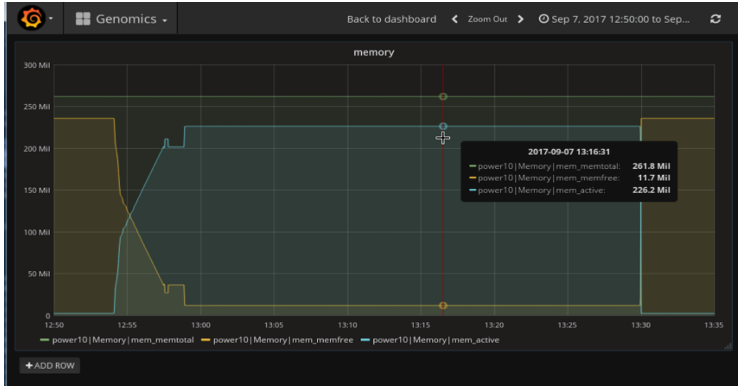
Application Profiling – Sam2Bam (Storage Mode)

- Consumes ~93% CPU in the initial phase (~10 minutes) and then around 40% CPU in the later phase.
- Is memory intensive even in storage mode. The sustained memory consumption of sam2bam in storage mode is around 223 GiB.
- The I/O pattern in the initial phase (~5 minutes) was write I/O. In the later phase it was predominantly read I/O.
- The gpfs_fis_bytes_read (~ 36 GB/s) is significantly higher compared to gpfs_fs_bytes_read (~300 MB/s). The average read bandwidth of this workload is ~300 MB/s. The sustained I/O capabilities from this node is ~12 GB/s. The high gpfs_fis_bytes_read indicates sam2bam read I/O benefitting from pagepool cache hits (~16 GiB pagepool). The application read I/O is random access in units of 512 KiB.

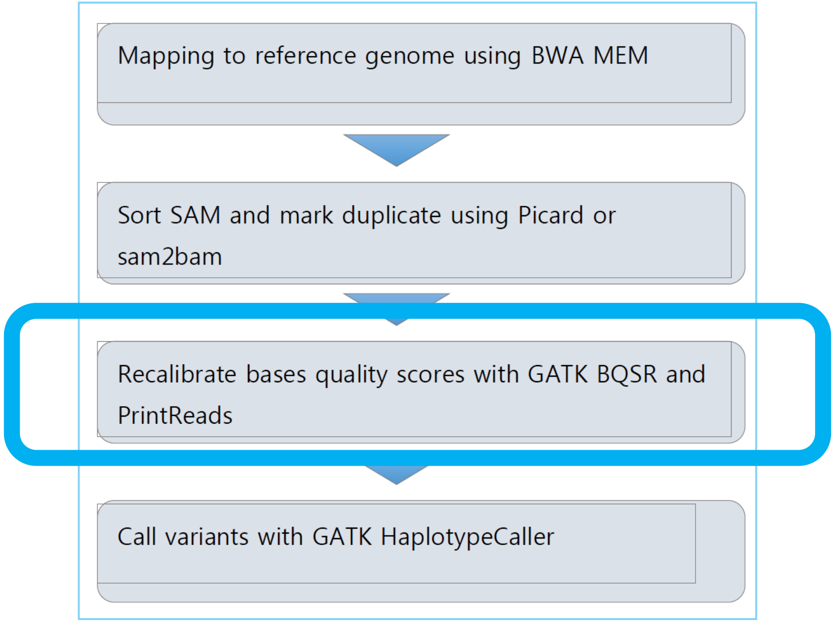


Input file format:
.bwa.sam

Output file format:
.md.bam



Application workflow

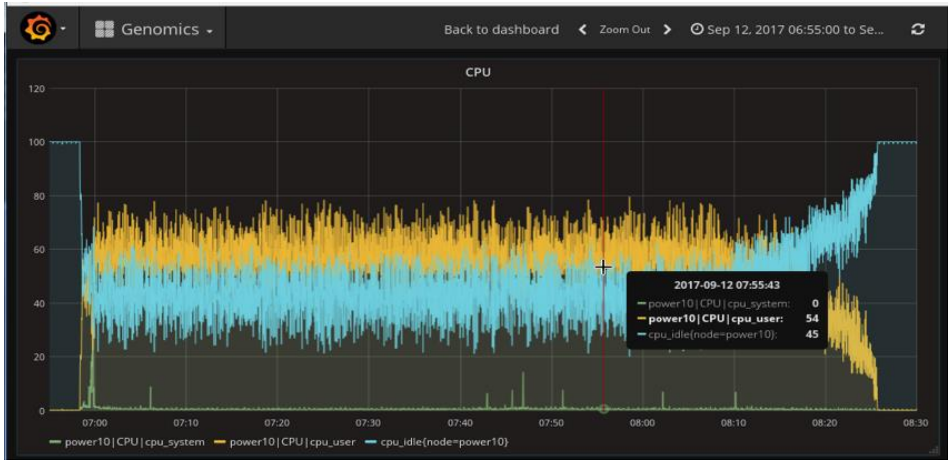


Tool	GATK (*)
Version	3.7-0 used for profiling
Source	https://software.broadinstitute.org/gatk/download/

(*) GATK archive versions are located at: https://software.broadinstitute.org/gatk/download/archive

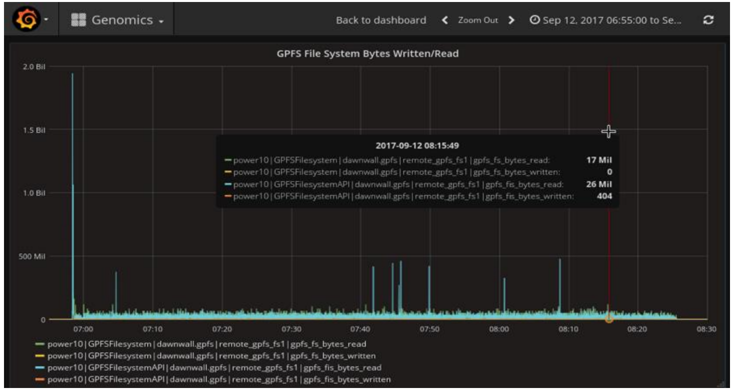
Application Profiling – GATK BQSR

- Consumes around 70% of CPU.
- GATK BaseRecalibrator is memory intensive. There are total of 18 x Java Threads. The memory for each Java thread was reduced to 10G (-Xmn10g -Xms10g -Xmx10g), so that aggregate memory consumption of GATK-BaseRecalibration Java component was 180G to fit within the node's memory capability.
- I/O pattern, this workload is predominantly read intensive. Average bandwidth for write and read is within 100 MB/s. Most of the read I/O size is in unit of file-system block-size (16 MiB) with mix of sequential and random I/O.



Input file format: .md.bam

Output file format: multiple .recal_reads<number>.table

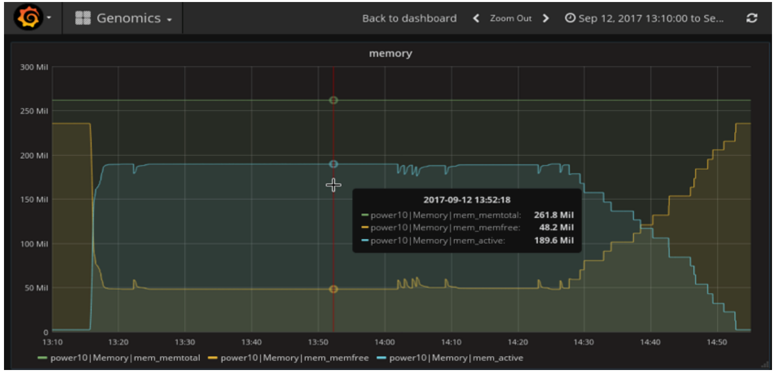


Application Profiling – GATK PrintRead

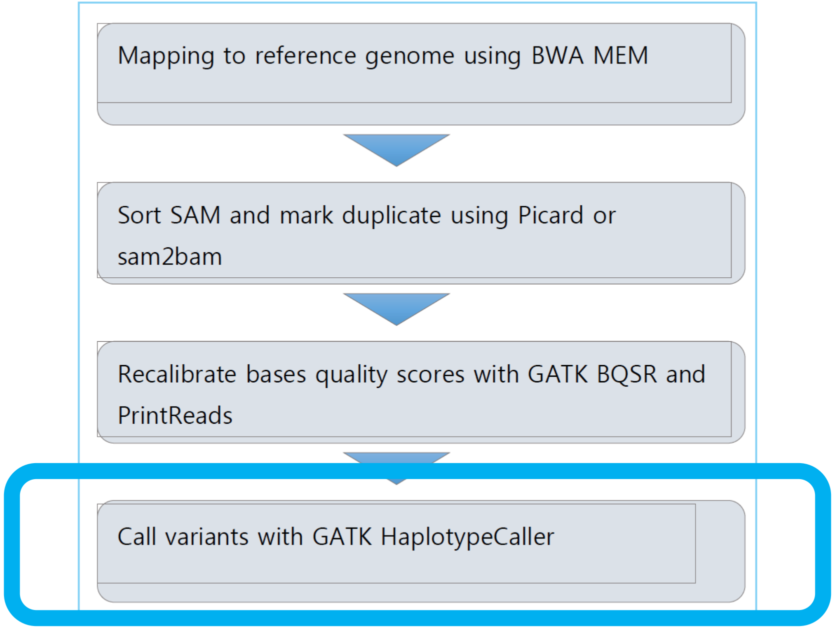
- Consumes around 70% of CPU.
- Is memory intensive. There are total of 18 x Java Threads. The memory for each Java thread was reduced to 10G (-Xmn10g -Xms10g -Xmx10g), so that aggregate memory consumption of GATK-PrintRead Java component was 180G to fit within the node's memory capability.
- I/O pattern, this workload has mix of read and write. Average bandwidth for write is within 150 MB/s. Average bandwidth for read is within 75 MB/s. The write I/O size is varied but mostly above 512 KiB with mix of sequential and random I/O. The read I/O size is mostly sequential I/O in units of FS block-size (16 MiB).

Input file format: multiple .recal_reads<number>.table

Output file format: multiple .recal_reads<number>.bam



Application Workflow



Tool	GATK (*)
Version	3.7-0 used for profiling
Source	https://software.broadinstitute.org/gatk/download/

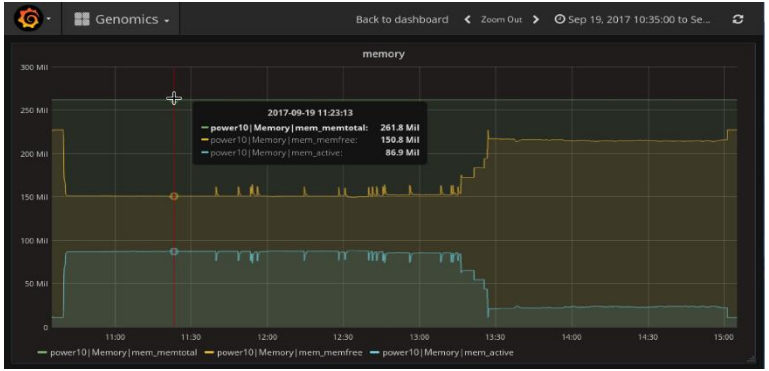
(*) GATK archive versions are located at: https://software.broadinstitute.org/gatk/download/archive

Application Profiling – GATK HaplotypeCaller

- Consumes around 40% of CPU.
- Is not memory intensive.
- In terms of I/O pattern, this workload has mix of read and write. Average bandwidth for write is within 100 MB/s. Average bandwidth for read is within 100 MB/s. The write I/O size is varied with mix of sequential and random I/O. The read I/O size is mostly sequential I/O in units of FS block-size (16 MiB).

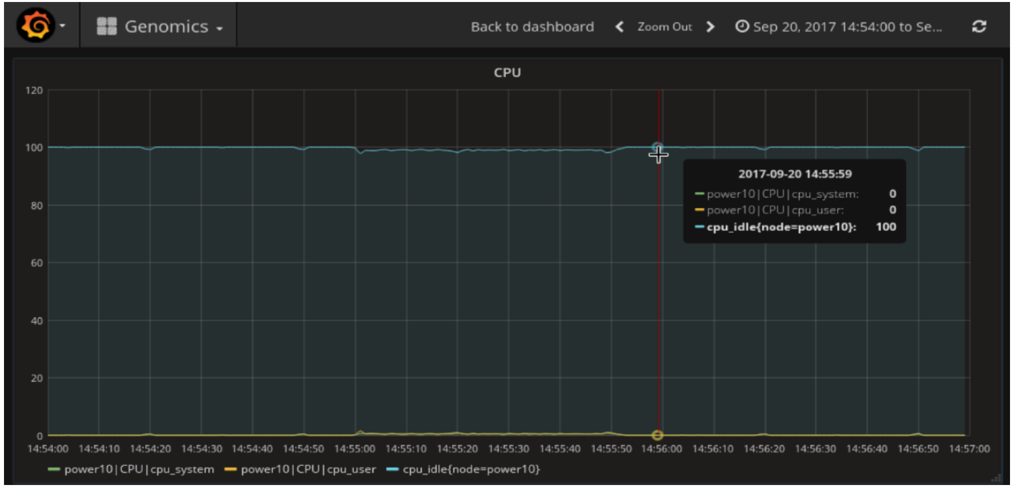
Input file format:
multiple
.recal_reads<number>.bam

Output file format:
Multiple
.raw_variants<number>.vcf



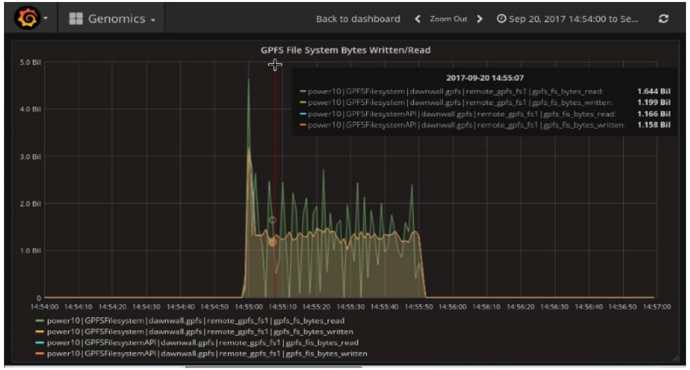
Application Profiling – GATK MergeVCF

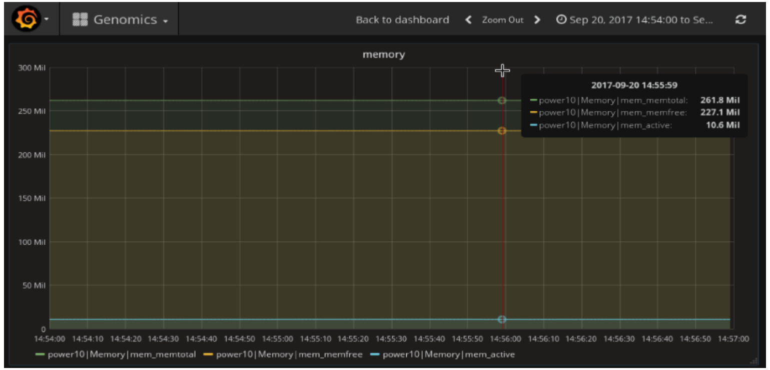
- Not CPU intensive.
- Is not memory intensive.
- I/O pattern, this workload has mix of read and write. Average bandwidth for write is within 1.5 GB/s. Average bandwidth for read is within 2 GB/s. The read I/O size is mostly sequential I/O in units of FS block-size (16 MiB). The write I/O size is mostly sequential I/O in units of FS block-size (16 MiB).



Input file format: multiple .raw_variants<number>.vcf

Output file format: Single .raw_variants.vcf file





IBM Spectrum Scale

Spectrum Scale Best Practices Guide for Genomic Medicine Workload 1.0 (Storage Services)

Dec 4st, 2017

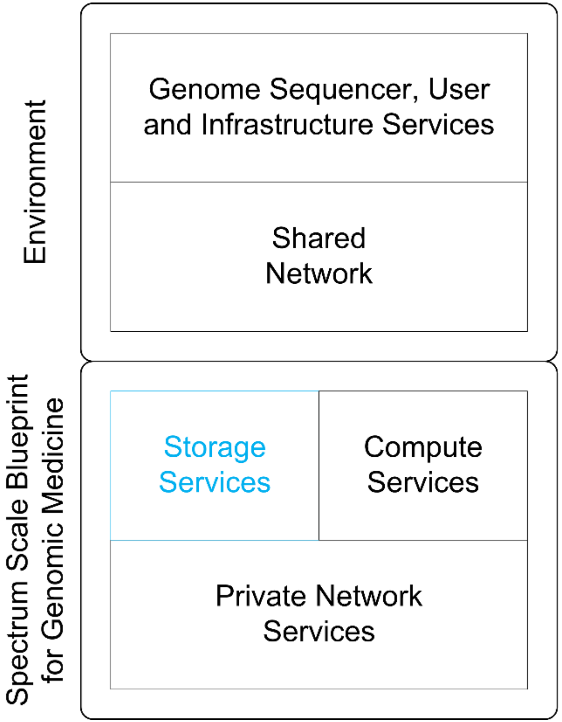
Summary

- The Spectrum Scale Blueprint for Genomic Medicine Workload describes Compute Services, Storage Services and Private Network Services. The next charts describe the Best Practices for Storage Services.
- The Spectrum Scale Blueprint for Genomic Medicine Workload is optimized for the “Broad Institute GATK Best Practices on IBM reference architecture”. Though, most of the recommendations are generic and apply to other workloads.
- Contact the Genomics War Room for help with different applications.

Outline

1. ***Composable building blocks***
2. Building block details

Storage Services – Capabilities



- 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, microscopes, etc., for access by data scientists/physicians and for **sharing 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 from a few 100 TB to Tens of PB of genomics 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

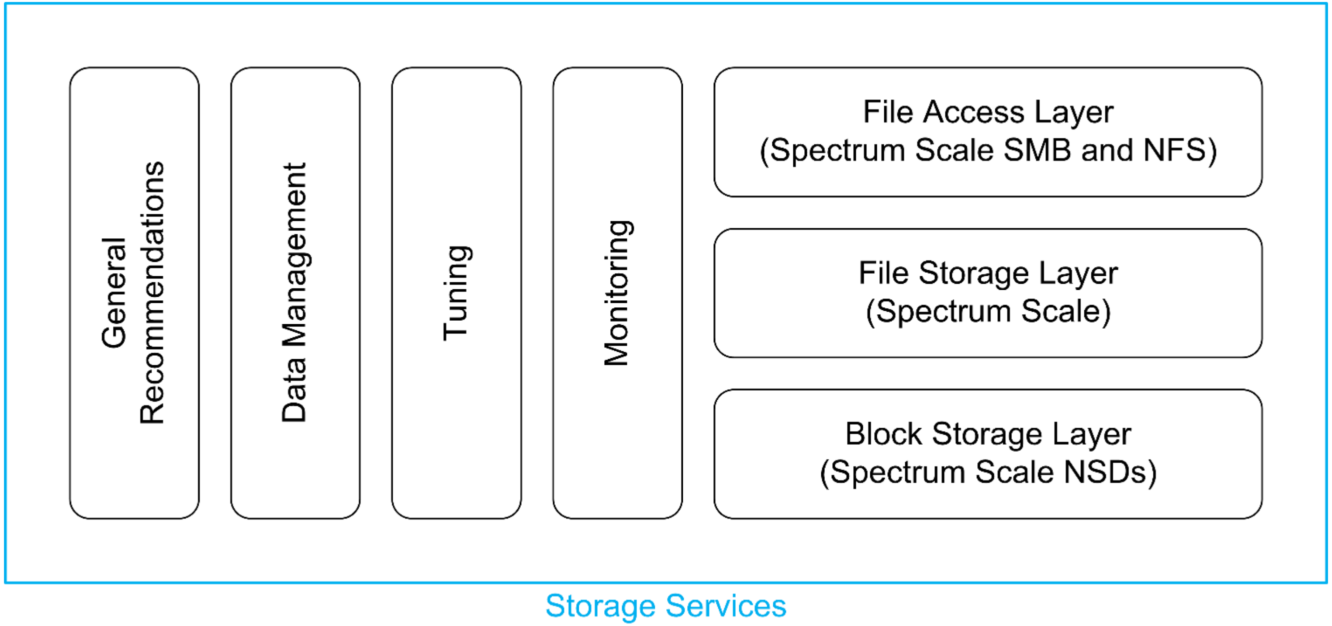
Storage Services – Solution Elements

Capability	Provided by
Scale-out architecture that is capable to store data from a few 100 TB to Tens of PB of file data	IBM Spectrum Scale
Data transfer nodes for secure high-speed external access via NFS and SMB to ingest data, user access and sharing	IBM Spectrum Scale – Cluster Export Services (CES)
Secure high-speed internal access for analysis on Compute Cluster	IBM Spectrum Scale – Remote Cluster Mount
End-to-end checksum to ensure the data integrity all the way from the application to the disks	IBM Elastic Storage Server (ESS)
Data Management GUI to configure and monitor storage resources	IBM Spectrum Scale – GUI
Optional professional services ranging from management of daily operation to consultancy for major configuration changes	IBM Lab Based Services

Example Configuration

- In the following we describe the design decision for a Storage Cluster that comprises:
  - 1x ESS Management Node (EMS)
  - 1x IBM Elastic Storage Server (ESS) GS2S with SSD
  - 1x IBM Elastic Storage Server (ESS) GL6S with NL-SAS
  - 3x CES Protocol Nodes for NFS and SMB
- Software Levels
  - ESS 5.2.0 (includes Spectrum Scale 4.2.3.4)
  - Spectrum Scale 4.2.3.4 also on CES nodes
  - RHEL 7.3 Little Endian (LE)

Storage Services – Composable Building Blocks

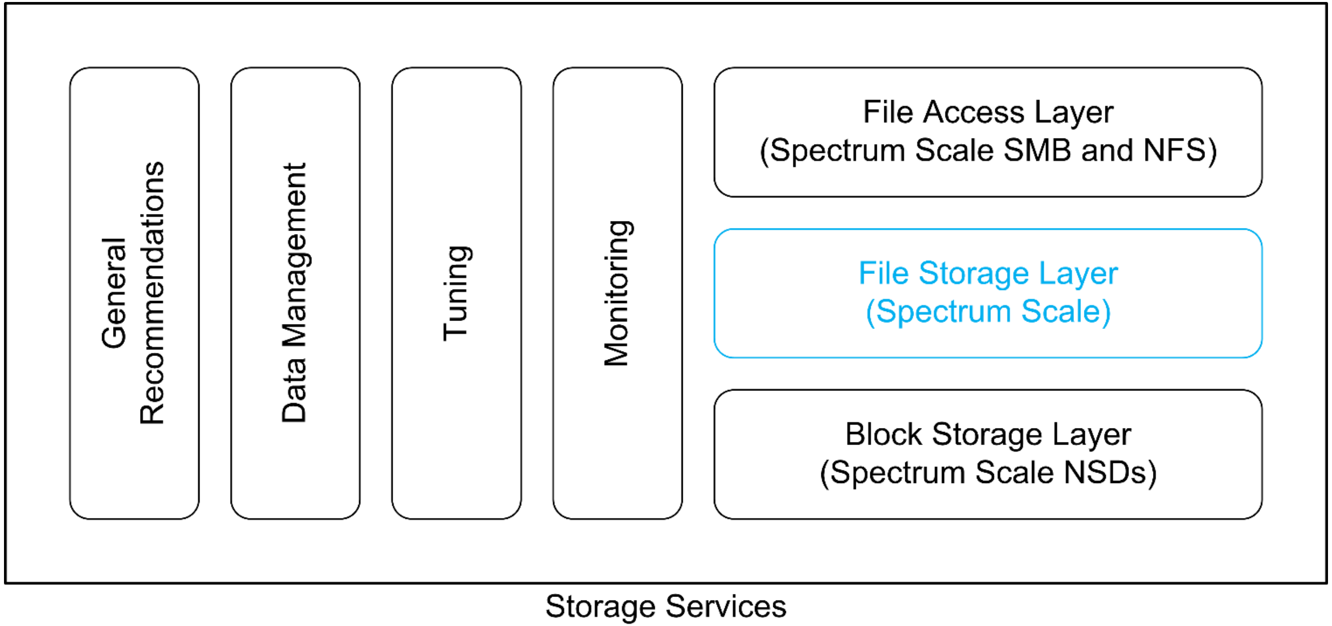


➔ ***A set of expertly engineered building blocks enable IT architects to compose solutions that meet customers varying performance and functional needs.***

Outline

1. Composable building blocks
2. ***Building block details***

Spectrum Scale Filesystems



➔ IBM Spectrum Scale is a parallel, scale-out filesystem that is capable to store data from a few TiBs to 100s of PiB of file data in one filesystem.

Spectrum Scale Filesystems – General Guidelines

How many **Spectrum Scale Filesystems** to configure? Do I need more then one?

- Multiple filesystems increase **administrative overhead**.
- Spectrum Scale **stripes data** across all available resources. Multiple Filesystems might **isolate resources**. For most workloads it is better, if Spectrum Scale can stripe across all available resources.
- Always think if **a new filesystem c**an be replaced by a **Spectrum Scale Fileset** of an already existing **Spectrum Scale Filesystems**.

Reasons for having more then one **Spectrum Scale Filesystem**:

- Each Spectrum Scale Filesystem can be configured with one Block Size only. Sometimes it is required to configure filesystems with **different Block Sizes** to tune performance for different workloads.
- Multiple Spectrum Scale Filesystems increase **resiliency**. Maintenance tasks such as fsck run longer on large file systems which implies longer downtime. If there is more than one filesystem, then data ingest and analysis can continue on remaining filesystems.
- **Quotas** may have an impact to write workload. With quota enabled the clients and the quota manager must communicate to ensure the quota is within the hard limits. This might degrade write performance with extreme workloads, for instance when in large clusters many nodes write huge amounts of data to many disks at the same time.
- **Snapshots** can be created on the fileset level, though snapshots flush the data of the whole filesystem. The snapshot does not proceed, until all the data from all NSD clients are flushed.
- Filesystems are the granularity for **exposing data to remote clusters**.

Spectrum Scale Filesystems – General Guidelines

Choosing the filesystem **Block Size** impacts space allocation and performance.

Allocation

- Spectrum Scale can store files smaller than ~3.5KiB (without extended attributes) in an inode, if the inode size is configured to 4KiB. For those files the filesystem block size does not matter.
- All other files are stored in filesystem blocks or filesystem subblocks. The minimal allocatable space is a subblock.
- Spectrum Scale up to 4.2.3 can divide a filesystem block in up to 32 subblocks of equal size. Storage capacity is wasted, if the file size is not a multiple of the subblock size.
- Note: Spectrum Scale 5.0 will introduce a capability to allow more than 32 subblocks. This allows a filesystem block to store more than 32 small files (e.g. 8KiB) within one filesystem block (default block size is 4MiB). Therefore this recommendation will change with Spectrum Scale 5.0.

Performance

- ESS systems offer overall better performance (considering client IO, rebuilds, etc) when configured with a filesystem block size of at least 4MiB for data storage pools.
- For ESS bases deployments, 4MiB block size is the best choice for very small and mixed file workloads, while still giving a lot of sequential performance for larger files in the same filesystem.
- If you use ESS and know your exact workload (I/O pattern), a file system block size that matches the workload I/O size may offer better client I/O performance, as long as the block size is greater than or equal to 4MiB.

Spectrum Scale Filesystems – General Guidelines

Block Allocation Map

- Spectrum Scale supports two different methods to allocated space: cluster and scatter.
- Scatter method provides more consistent file system performance on large clusters and large filesystems by averaging out performance variations due to block location.
- Scatter method is appropriate in most cases and is the default for GPFS clusters with more than eight nodes or file systems with more than eight disks.

Log File Size

- The Log File Size specifies the size of the internal log files.
- An increased Log File Size is useful for file systems that have a large amount of metadata activity, such as creating and deleting many small files or performing extensive block allocation and deallocation of large files, typical with Genomics application I/O workload.

Replication

- Spectrum Scale supports the replication of data and/or metadata on the filesystem level.
- Enabled replication of data and/or metadata reduces the overall usable capacity.
- Enabled replication protects against Spectrum Scale NSD or underlying block storage errors.
- In odd situations a whole Spectrum Scale NSD can get lost.

  → This should not happen, but sometimes things happen that should not happen.
- Loosing a Spectrum Scale NSD that stores metadata implies the loss of the whole filesystem.
  - The restore of a peta-scale filesystem can take very loooong.
- It is recommended to replicate at least the metadata to increase the resilience of the filesystem.

Spectrum Scale Filesystems – General Guidelines

Number of Spectrum Scale Nodes

- Certain internal data structures of a Spectrum Scale Filesystem are optimized for the number of nodes where the filesystem will be mounted.
- The number of nodes includes nodes in the local Spectrum Scale Cluster as well as nodes of all remote Spectrum Scale Clusters where the filesystem is mounted with multi-cluster remote cluster mount.
- The ‘-n <number>’ option of the mmcrfs command gives Spectrum Scale a hint to optimize these data structures for the given number of nodes.
- The data structures will be initialized during the creation of a filesystem.
- This value can be changed later on, but a migration of files to a new storage pool is required to make the value effective.
- When creating a new filesystem is better to overestimate the number of nodes by a factor of two than to making it to small. For instance, when you plan to create a Spectrum Cluster with 80 nodes, then it is reasonable to specify:
‘mmcrfs … -n 128 …’.

ACLs

- Spectrum Scale supports POSIX ACLs and NFSv4 ACLs.
- Spectrum Scale Cluster Export Services (CES) require to configure the respective Spectrum Scale filesystem with NFSv4 ACLs.
- Configure all Spectrum Scale filesystems with the same ACL type to keep ACL management consistent across all filesystems.

Spectrum Scale Filesystems – Guidelines for Genomic Workload

There should be up to four Spectrum Scale filesystems in the Genomics Blueprint.

- /gpfs/data → genomic data and analysis results
- /gpfs/app → application binaries, configuration files and log files
- /gpfs/user → user data for execution of batch jobs (optional)
- /gpfs/ces → helper filesystem for Cluster Export Services (CES) to provide NFS and SMB

They are required to optimize performance of each by

- Isolating the respective I/O load from impacting each other.
- Setting specific blocksize, relatime, and replication based on each file-system's function.
  - */gpfs/data: -j scatter, -B 8MiB, -n <customer_specific>, --metadata-block-size 1MiB , -L 32 MiB, -S relatime*
  - */gpfs/app: -j scatter, -B 4MiB, -n <customer_specific>, --metadata-block-size 1MiB, -L 32 MiB*
  - */gpfs/user: -j scatter, -B 4MiB, -n <customer_specific>, --metadata-block-size 1MiB, -L 32 MiB (optional)*
  - */gpfs/ces: -j scatter, -B 1MiB, -n <customer_specific>*
- Note: Relatime reduces meta data traffic.
- Note: Separating the file-system metadata and data storage pools also enhances performance. The system pool is comprised of metadataOnly NSDs from the ESS GS2S and data is comprised of dataOnly NSDs from the ESS GL6S.

Spectrum Scale Filesystems – Guidelines for Genomic Workload

Considerations for user filesystem

- Execution of batch jobs on the Compute Cluster require a shared home directory. This genomics blueprint suggests two possible methods:
    1. Utilize existing shared home export from existing NFS server to mount it onto the compute cluster nodes to avoid data silos
    2. Create separate Spectrum Scale filesystem (/gpfs/user) on Storage Cluster to avoid dependency to external NFS service

Considerations for /gpfs/user (optional filesystem)

- Many customers have the requirement for hourly snapshots of user data.
    - Isolate hourly snapshots from other Spectrum Scale filesystems
- Do not export /gpfs/user via CES
    - External NFS or SMB access might impact performance of running batch jobs

Spectrum Scale Filesystems – Guidelines for Genomic Workload

Name	/gpfs/data
Purpose	Store genomic data and analysis result
Why separate filesystem?	This filesystem 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 Spectrum Scale multi-cluster remote cluster mount)
Exported via CES	Yes (SMB and NFS)
Number of Nodes	Customer specific (see guidelines on the previous charts)

Spectrum Scale Filesystems – Guidelines for Genomic Workload

Name	/gpfs/app
Purpose	Stores all applications binaries, scheduler binaries, configuration files and log files needed on the compute nodes
Why separate filesystem?	Maintenance on /gpfs/data (e.g. file system check) must not impact availability of applications on compute nodes
Size	Depends on customer requirements. Rule of thumb: ~50TiB at least
Metadata	1 MiB block size on SSD, no replication
Data	4 MiB block size on NL-SAS, no replication, System Pool only
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	Root fileset only
Relatime	Use default
Quota	Enable quota (-Q yes) (avoids remount when we enable quota later)
Exported to Compute Cluster	Yes (via Spectrum Scale multi-cluster remote cluster mount)
Exported via CES	No
Number of Nodes	Customer specific (see guidelines on the previous charts)

Spectrum Scale Filesystems – Guidelines for Genomic Workload

Name	/gpfs/user
Purpose	User data for execution of batch jobs (optional filesystem)
Why separate filesystem?	Isolate activity from other Separate Scale filesystems
Size	Depends on customer requirements. Rule of thumb: ~50GiB per user
Metadata	1 MiB block size on SSD, no replication
Data	4 MiB block size on NL-SAS, no replication, System Pool only
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	Root fileset only
Relatime	Use default
Quota	Enable quota (-Q yes) (avoids remount when we enable quota later)
Exported to Compute Cluster	Yes (via Spectrum Scale multi-cluster remote cluster mount)
Exported via CES	No
Number of Nodes	Customer specific (see guidelines on the previous charts)

Spectrum Scale Filesystems – Guidelines for Genomic Workload

Name	/gpfs/ces
Purpose	Metadata for Cluster Export Services (CES)
Why separate filesystem?	Isolation from all other filesystems to increase resiliency of NFS and SMB
Size	64 GiB
Metadata + Data	1 MiB block size on SSD, System Pool only
Log File Size	32 MiB (-L 32M)
Block Allocation Map	Scatter
Replication	Replicate data and metadata (-M 2 -R 2 -m 2 -r 2)
ACL Type	NFSv4 only
Filesets	Root fileset only
Relatime	Use default
Quota	No
Exported to Compute Cluster	No
Exported via CES	No
Number of Nodes	Customer specific, typically 32 or 64

Spectrum Scale Filesets – General Considerations

What is a fileset?

- A fileset is a sub-tree of a file system namespace that provides a means of partitioning the filesystem to allow administrative operations. From a user point-of-view, a fileset looks like a directory.
- There are two types of filesets: An independent fileset has its own inode space. A dependent fileset shares its inode space with an associated independent fileset. A filesystem can have up to 1,000 independent filesets and up to 10,000 dependent filesets. Some data management functions have a dependency to the fileset type.

When to use filesets?

- Consider dependent filesets to use advanced placement policies and ILM tiering.
- Consider independent filesets to use project level quotas, snapshot and AFM in addition to advanced placement policies and ILM tiering. But keep the limit of 1,000 independent filesets in mind.

How to design filesets?

- Filesets are not required, but filesets are a great tool to effectively automate data and capacity management.
- Filesets need to be configured right from the beginning. Introducing filesets or changing fileset boundaries later might trigger expensive copy or move operations.
- Configure at least one independent fileset for each filesystem, to allow to configure data management later, even if it is not required data at the beginning.
- The fileset design is customer specific and depends on how the customer organizes data. This can be a complex task which needs some experience. Consider to procure professional services.

Again: always think if a fileset can replace a filesystem.

Spectrum Scale Filesets and Directories – Guidelines for Genomic Workload

/gpfs

- Directory under Linux root filesystem ("/")

/gpfs/data

- Spectrum Scale File System under /gpfs

/gpfs/data/project1, /gpfs/data/project2, /gpfs/data/project3, ...

- Use independent filesets under /gpfs/data, if you do not hit the 1,000 fileset limit
- Customer may want to choose different naming convention

/gpfs/app

- Spectrum Scale File System under /gpfs
- Directory structure for workload scheduler needs special consideration.
  → See Reference Guide for Compute Services for best practices

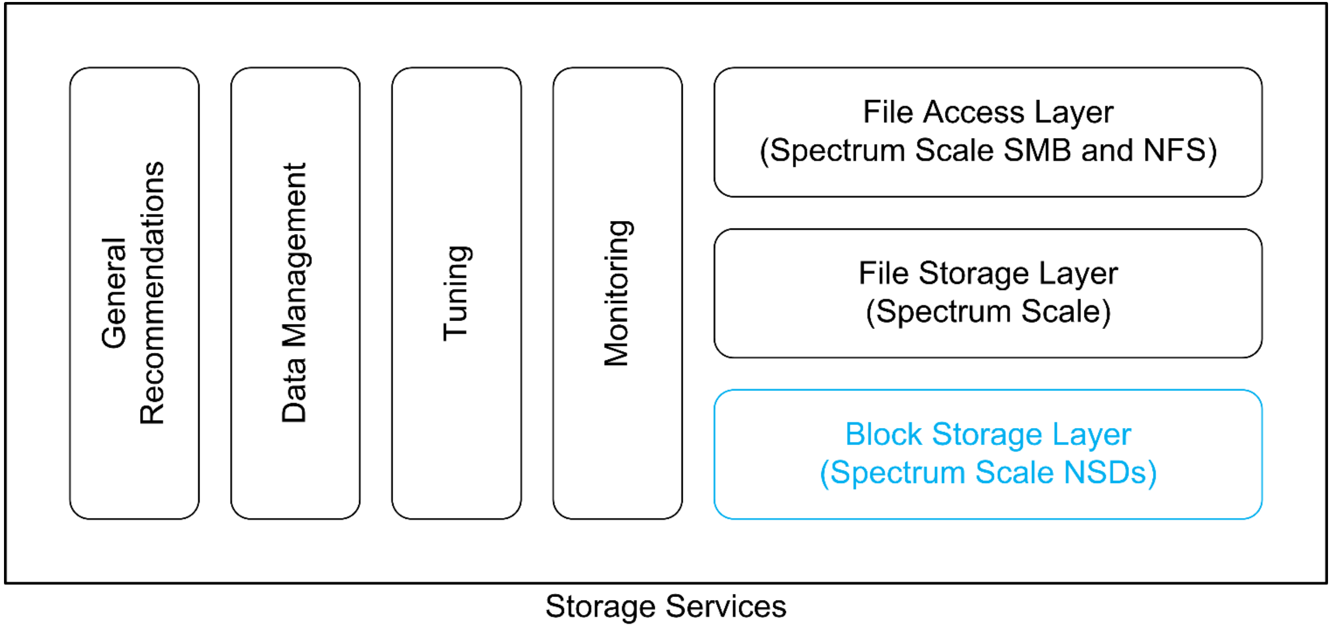
/gpfs/user

- Optional Spectrum Scale File System under /gpfs/

/gpfs/ces

- Spectrum Scale File System under /gpfs/

Spectrum Scale NSDs



→ *ESS NSDs provide end-to-end checksum to ensure the data integrity all the way from the applications to the disks.*

Spectrum Scale RAID – General Guidelines

Spectrum Scale RAID (a.k.a. GPFS Native RAID, GNR)

- IBM Elastic Storage Server (ESS) includes Spectrum Scale RAID
- Spectrum Scale RAID makes ESS an excellent choice for performance and resiliency
- Fast disk rebuilds: Disks rebuild in minutes vs hours/days of traditional RAID 5 and RAID 6.
- End-to-end data integrity: Spectrum Scale RAID maintains checksum of data blocks from the client to the blocks on the disk and validates at every point, thus eliminating the chances of silent data corruption or data loss.
- Higher storage resiliency: The erasure coding is with up to three parity blocks and can survive three disk failures with only 27% overhead in capacity compared to 200% overhead with three-way replication. It uses fault domains to layout disks in such a way that it can survive entire disk shelf (enclosure) failures. It also uses a disk hospital to pro-actively identify sick drives (disks with bad sectors or media errors) and either a) replace the disk or b) fix any bad data from parity.

General Considerations

- Separate metadata and data to enable fast metadata access and updates
- Metadata: On SSDs with 4-way replication.
    - Protects against three disk failures and enclosure failure.
- Data: On NL-SAS disks with 8+3P erasure encoding.
    - Protects ESS GL6S against three disk failures and enclosure failure.

Spectrum Scale RAID – Guidelines for Genomic Workload

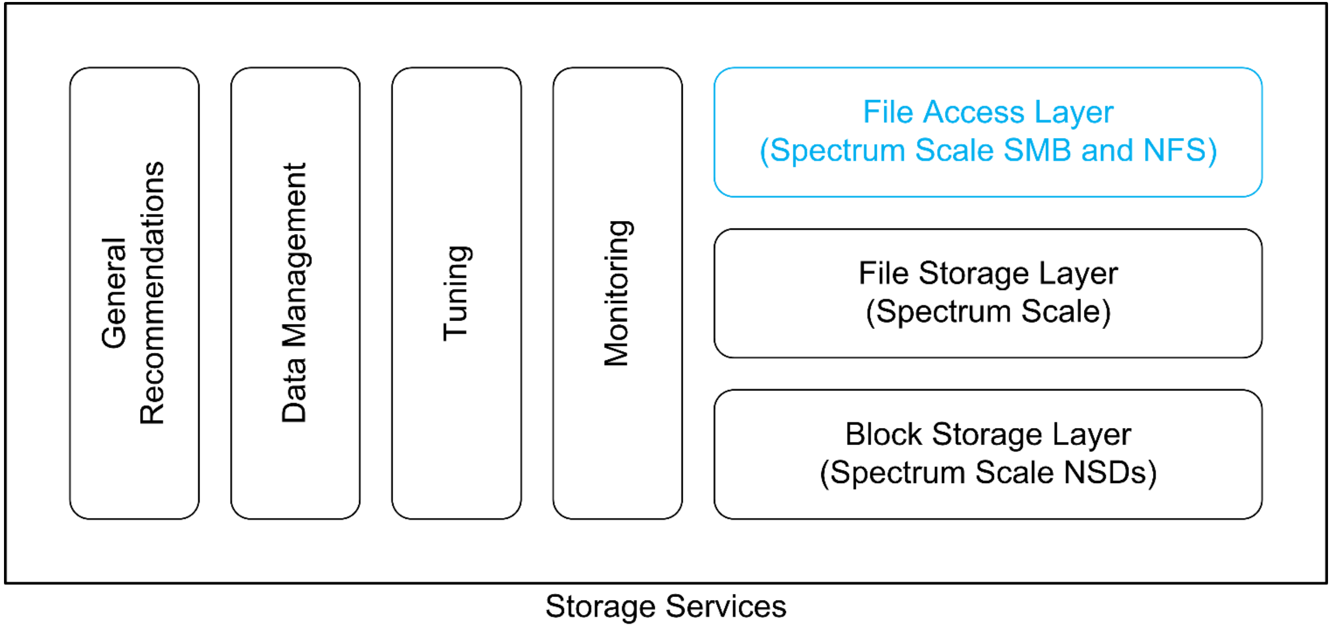
Recommendation for example configuration:

Name	Metadata	Data
/gpfs/data	4W on SSD	8+3P on NL-SAS
/gpfs/app	4W on SSD	8+3P on NL-SAS
/gpfs/user	4W on SSD	8+3P on NL-SAS
/gpfs/ces	4W on SSD	8+3P on NL-SAS

Notes

- ESS GS2S provides SSDs for metadata
- ESS GL6S provides NL-SAS for data

File Access – NFS and SMB



➔ *Spectrum Scale Cluster Export Services (CES) enable secure high-speed external access via NFS and SMB to ingest data from genomic sequencers, microscopes, etc., for access by data scientists/physicians and sharing across sites and institutions.*

File Access – NFS and SMB

What access protocols to use?

- Spectrum Scale Cluster Export Services (CES) provide built-in support for NFS and SMB.
- Access of devices such as sequencers and microscopes is determined by the interfaces that the devices provide. Most devices provide a capability to write acquired data to an SMB or NFS share.
- Field experiences shows that SMB provides effective access for laptops and workstations running Windows, Linux and macOS.

What do I need to consider to configure NFS and SMB?

- Spectrum Scale Cluster Export Services (CES) depend on an external authentication and ID mapping source for user identification and user authentication such as LDAP or Active Directory.
  - See Spectrum Scale Knowledge Center for supported authentication methods and other planning tips: https://www.ibm.com/support/knowledgecenter/STXKQY_4.2.3/com.ibm.spectrum.scale.v4r23.doc/bl1in_PlanningForProtocols.htm
- Spectrum Scale Cluster Export Services (CES) depend on an external network to connect external devices and users.
  - See Spectrum Scale Knowledge Center for external network configuration:
    https://www.ibm.com/support/knowledgecenter/en/STXKQY_4.2.3/com.ibm.spectrum.scale.v4r23.doc/bl1adv_cesnetworkconfig.htm
    https://www.ibm.com/support/knowledgecenter/en/STXKQY_4.2.3/com.ibm.spectrum.scale.v4r23.doc/bl1ins_planningsmb.htm
    https://www.ibm.com/support/knowledgecenter/en/STXKQY_4.2.3/com.ibm.spectrum.scale.v4r23.doc/bl1ins_deployingprotocolstasks.htm
  - Load balancing for protocol services is outside the scope of this version of the blueprint.

File Access – NFS and SMB

What directories to export via CES?

/gpfs/data

- Devices such as sequencers and microscopes typically support data acquisition via SMB and/or NFS.
- IT is general best practice to connect workstations and laptops of end user like data scientists via SMB.
- Physicians typically access results via a download via portal. This is outside the scope of the blueprint.

/gpfs/app

- Not exported via CES to avoid potential performance impact of running batch jobs by concurrent NFS or SMB access.

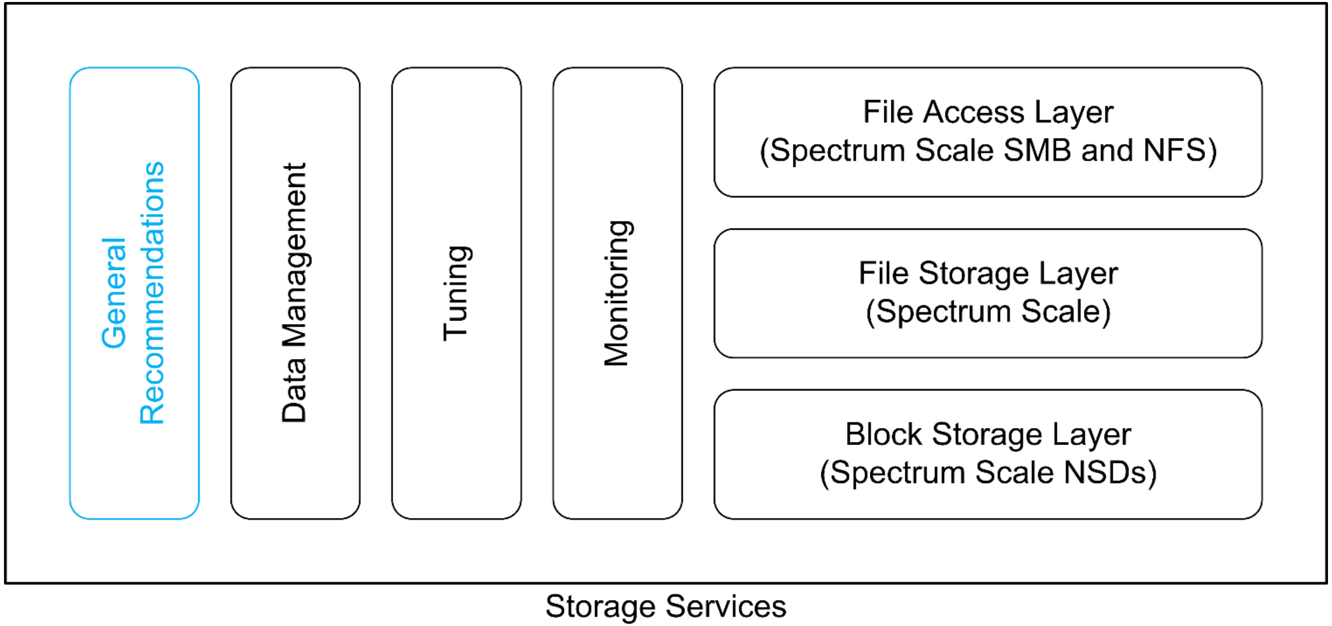
/gpfs/user

- Not exported via CES to avoid potential performance impact of running batch jobs by concurrent NFS or SMB access.

/gpfs/ces

- Not exported via CES, because that is an internal filesystem for CES metadata only.

General Configuration Recommendations



➔ *Best practices increase operational efficiency for managing the whole storage infrastructure.*

Node Designation – Example Configuration

	Node Type	Memory	Spectrum Scale Node	Spectrum Scale Quorum	Spectrum Scale Manager	Spectrum Scale Admin Node	Spectrum Scale Contact Node	Spectrum Scale GUI
ESS EMS	ESS Mgmt	32 GB (*)	X			X		X
ESS GS2S I/O 1	ESS I/O	256 GB (**)	X				X	
ESS GS2S I/O 2	ESS I/O	256 GB (**)	X				X	
ESS GL6S I/O 1	ESS I/O	256 GB (**)	X				X	
ESS GL6S I/O 2	ESS I/O	256 GB (**)	X				X	
CES Protocol 1	CES	128 GB	X	X	X	X		
CES Protocol 2	CES	128 GB	X	X	X	X		
CES Protocol 3	CES	128 GB	X	X	X	X		

(*) ESS EMS Nodes are always configured with 32GB memory.
(**) ESS I/O Nodes are always configured with 256GB memory.

Node Designation – Spectrum Scale

Quorum Nodes

- The general recommendation is to define three or five quorum nodes, but there is no single correct answer how many **Quorum Nodes** should be configured.
- The Spectrum Scale Nodes which assume the role of a **Quorum Node** needs to be on reliable nodes, as much as possible.
- Each **Quorum Node** should have independent failure domain to avoid single point of failure, e.g. different power circuit, different rack, different network switch.
- **ESS I/O Nodes** must not be configured as **Quorum Node**.
- Each **Quorum Node** will automatically become a **Config Server**.

Manager Nodes

- Spectrum Scale has a capability to define which nodes can assume the role of a **Manager Node**.
- Spectrum Scale will automatically assign the following roles to the available Manager Nodes:
  Cluster Manager, Filesystem Manager, Token Manager.
- **ESS I/O Nodes** must not be configured as **Manager Node**.

Contact Nodes for Multi-Cluster Remote Cluster Mount

- Contact Nodes are required on Spectrum Scale clusters that export Spectrum Scale filesystems to other Spectrum Scale Cluster via multi-cluster remote cluster mount.
- The contact nodes can be identified through either their hostnames or IP addresses.

Node Designation – Spectrum Scale

Admin Nodes

- Spectrum Scale **Admin Nodes** are responsible for issuing any and all Spectrum Scale administrative commands.
- Spectrum Scale commands maintain the appropriate environment across all nodes in the cluster.
- The **Admin Nodes** have similar requirements as the Management Nodes: password less root ssh and scp to all other Spectrum Scale Nodes, access restricted to administrative users only.
- For redundancy, it is best, if possible to have at least two Spectrum Scale Nodes that are **Admin Nodes**.

GUI Nodes

- The Spectrum Scale **GUI Nodes** are always Admin Nodes.
- The GUI does not allow root login. Only an admin login exists.
- The GUI subsystem passes commands as root to the other Spectrum Scale Nodes of the cluster.
- Most, but not all, Spectrum Scale functions can be run from the GUI, so occasionally, some commands require root login for CLI access.
- All GUI Nodes run a performance monitoring collection daemon that is used by the GUI to report cluster health and performance.
- ESS allows to configure only the EMS Node as GUI Node.

Miscellaneous – Spectrum Scale

Autoload

- The autoload option determines whether Spectrum Scale will be started automatically when a node is booted. This is a node setting.
- Please note that autoload is different to the automount option of a Spectrum Scale filesystem. The automount option indicates whether a filesystem is mounted automatically on all nodes.
- It is best practice that all quorum nodes and manager nodes are configure with autoload=yes. This increases the resiliency of the Spectrum Scale cluster.
- It is best practice that all NSD client nodes are configure with autoload=yes. This simplifies the management of the Spectrum Scale cluster.
- ESS I/O nodes and EMS nodes will be configured with autoload=no. This is the default setting configured by the ESS install scripts.

General Considerations

- All nodes of same cluster need to be able to communicate to each other.
  - Configure GPFS daemon/network communication over high-speed network. On Infiniband networking, enable GPFS configuration parameter ‘verbRdma=enable’.
  - See best practices guide for Network Services for details.
- All nodes used for administering Spectrum Scale must be able to do ssh and scp on any other node in the cluster as user root without the use of a password.
- Sudo wrappers cannot be used, because deployment toolkits for CES and ESS do not support it yet.

External Dependencies

Spectrum Scale depends on high-available Name Resolution Services (**DNS**) for name resolution and reverse name resolution.

- **Each Spectrum Scale Node** needs to connect to the **DNS** service running on the EMS.
- The **DNS** services running on the EMS needs to connect to the customer provided **DNS** service.

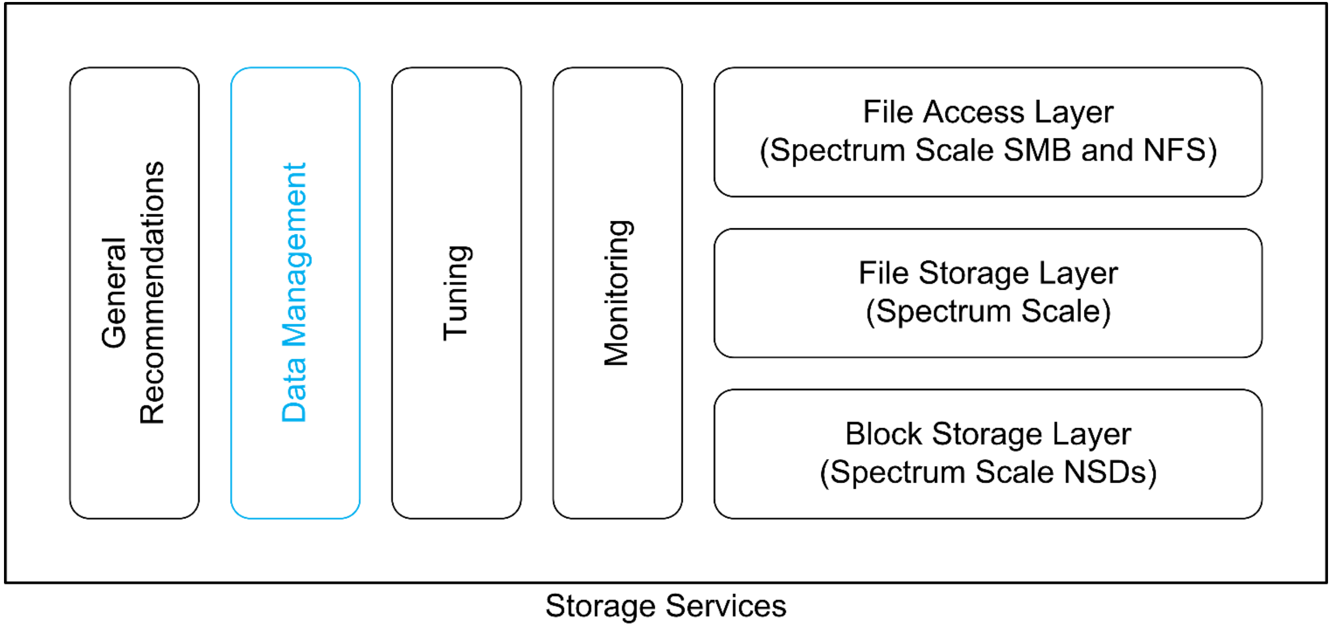
Spectrum Scale depends on Time Services (**NTP**) for time synchronization:

- The time of all **Spectrum Scale Nodes** needs to be synchronized.
- In addition, the **CES Nodes** need to be synchronized with protocol clients and authentication services.
- **Each Spectrum Scale Node** needs to connect to the **NTP** service running on the EMS.
- The **NTP** services running on the EMS needs to connect to the customer provided **NTP** service.

Spectrum Scale Admin Nodes need to map UID and GIDs to user and group names and vice versa.

- **Each Spectrum Scale Admin Node** needs to connect to the customer provided ID Mapping service.
- Best practice is to configure all Spectrum Scale Nodes with ID Mapping to keep configuration of all nodes the same.

Automated Data Management

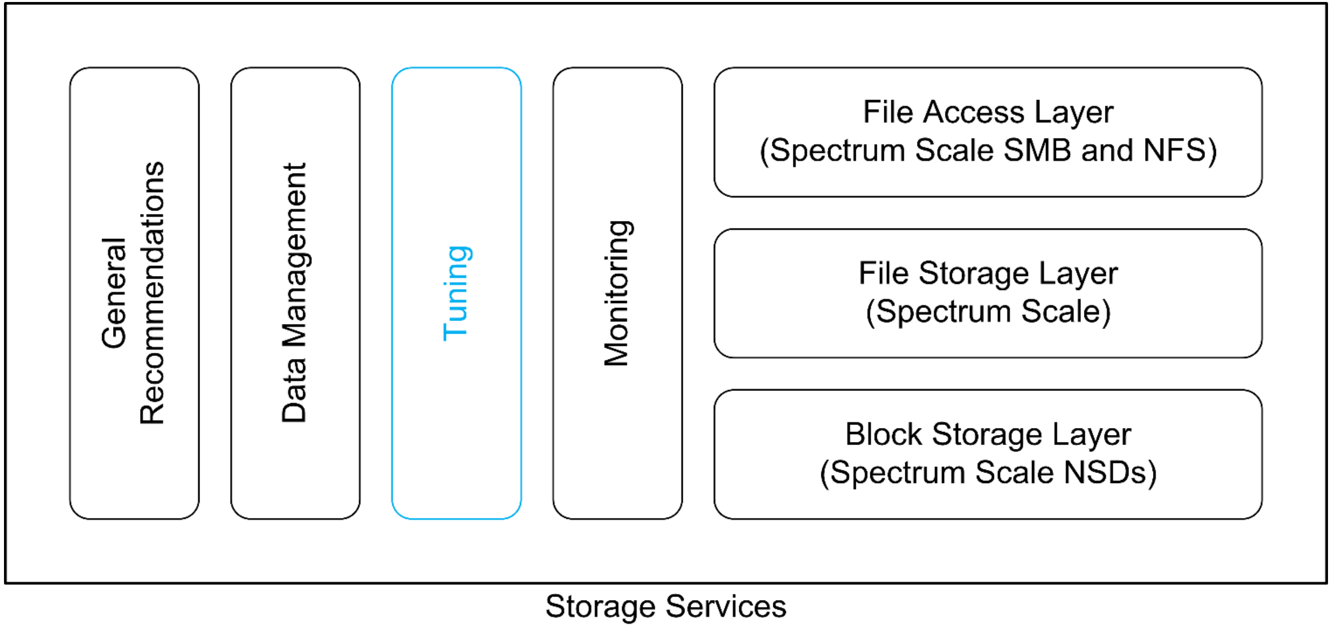


➔ ***Spectrum Scale's built-in data management capabilities increase operational efficiency for managing and storing huge amounts of data.***

Automated Data Management

- Separate Data and Metadata to enable fast metadata access and updates
  - System Pool for metadata only on SSD
  - Data Pool for data on NL-SAS
- More automated data management capabilities will be added to a future update of the blueprint

Tuning



➔ The tuning recommendations are optimized for the "Broad Institute GATK Best Practices on IBM reference architecture" and IBM Elastic Storage Server (ESS). Though, most settings are generic.

Tuning Guidelines – ESS IO Nodes

- Genomics Blueprint 1.0 will be based on ESS 5.2 (GPFS 4.2.3.4) on the ESS IO nodes and Spectrum Scale 4.2.3.4 (or higher) on the Protocol and Compute nodes.
- Install ESS recommended firmware and software packages
  - See Runbooks for Genomics Medicine Workload for details.
- Create the GPFS Storage cluster using the ESS scripts
  https://www.ibm.com/support/knowledgecenter/en/SSYSP8_5.2.0/sts52_welcome.html
  https://www.ibm.com/support/knowledgecenter/SSYSP8_5.2.0/ess_qdg.pdf?view=kc
- /etc/sysctl.conf
  - *net.core.somaxconn = 8192*
- GPFS configuration
  - *mmchconfig socketMaxListenConnections=8192 –N <essio_node_class>*
  - *mmchconfig envVar="MLX4_USE_MUTEX=1 MLX5_SHUT_UP_BF=1 MLX5_USE_MUTEX=1" –N <essio_node_class>*

➔ This set of tunables is best practice for IBM Elastic Storage Server and needs to be applied for genomic workload.

Tuning Guidelines – ESS I/O Nodes

blue ~ tuning applied for genomic workload
grey ~ default settings for ESS I/O node

Snip of mmlsconfig:

[ESS I/O Nodes]
nsdRAIDBufferPoolSizePct 80
maxBufferDescs 2m
nsdRAIDTracks 128k
nsdRAIDSmallBufferSize 256k
nsdMaxWorkerThreads 3k
nsdMinWorkerThreads 3k
nsdRAIDSmallThreadRatio 2
nsdRAIDThreadsPerQueue 16
nsdRAIDEventLogToConsole all
nsdRAIDFastWriteFSDataLimit 256k
nsdRAIDFastWriteFSMetadataLimit 1M
nsdRAIDReconstructAggressiveness 1
nsdRAIDFlusherBuffersLowWatermarkPct 20
nsdRAIDFlusherBuffersLimitPct 80
nsdRAIDFlusherTracksLowWatermarkPct 20
nsdRAIDFlusherTracksLimitPct 80
nsdRAIDFlusherFWLogHighWatermarkMB 1000
nsdRAIDFlusherFWLogLimitMB 5000
nsdRAIDFlusherThreadsLowWatermark 1
nsdRAIDFlusherThreadsHighWatermark 512
nsdRAIDBlockDeviceMaxSectorsKB 8192
nsdRAIDBlockDeviceNrRequests 32
nsdRAIDBlockDeviceQueueDepth 16
nsdRAIDBlockDeviceScheduler deadline
nsdRAIDMaxTransientStale2FT 1
nsdRAIDMaxTransientStale3FT 1
nsdMultiQueue 512
nspdQueues 64
numaMemoryInterleave yes
maxFilesToCache 128k
maxMBpS 16000
workerThreads 1024
ioHistorySize 64k
verbsRdma enable
verbsRdmaSend yes
verbsRdmasPerConnection 128
verbsSendBufferMemoryMB 1024
scatterBufferSize 256K
nsdClientCksumTypeLocal ck64
socketMaxListenConnections 8192
envVar MLX4_USE_MUTEX=1
MLX5_SHUT_UP_BF=1 MLX5_USE_MUTEX=1
maxStatCache 0
pagepool <60% of memory>
verbsPorts <active_verbs_ports>

➔ This set of tunables is best practice for IBM Elastic Storage Server and needs to be applied for genomic workload.

Tuning Guidelines – Protocol Nodes

OS tunable

- ulimit (Include the following in /etc/security/limits.conf)

  ** soft memlock unlimited*

  ** hard memlock unlimited*

  ** soft nofile 16384*

  ** hard nofile 16384*

  [detailed output in the notes]

- tuned configuration
  - /etc/tuned/active_profile is set to “throughput-performance”
  - /usr/lib/tuned/throughput-performance/tuned.conf

    [cpu]

    governor=performance

    energy_perf_bias=performance

    min_perf_pct=100

    [detailed output in the notes]

➔ This set of tunables is best practice for Spectrum Scale CES and needs to be applied for genomic workload.

Tuning Guidelines – Protocol Nodes

Network tunable

- On Mellanox Adapters,
  apply Mellanox OFED Tunings
  https://community.mellanox.com/docs/DOC-2489
- Connect the protocol 10GigE/40GigE interface (for shared protocol access) to high-speed network port in the Ethernet switches.

- /etc/sysctl.conf

  net.ipv4.tcp_timestamps=0
  net.ipv4.tcp_sack=0
  net.core.netdev_max_backlog=250000
  net.core.rmem_max=16777216
  net.core.wmem_max=16777216
  net.core.rmem_default=16777216
  net.core.wmem_default=16777216
  net.core.optmem_max=16777216
  net.ipv4.tcp_rmem=4096 87380 16777216
  net.ipv4.tcp_wmem=4096 65536 16777216
  net.ipv4.tcp_low_latency=1
  net.ipv4.tcp_adv_win_scale=2
  net.ipv4.tcp_window_scaling=1
  net.core.somaxconn = 8192
  vm.min_free_kbytes = 512000
  kernel.sysrq = 1
  kernel.shmmax = 137438953472

➔ This set of tunables is best practice for Spectrum Scale CES and needs to be applied for genomic workload.

Tuning Guidelines – Protocol Nodes

Spectrum Scale tunables (Compute nodes will be based on version 4.2.3.4 or later PTF)

- Since the storage backend is ESS, apply gssClientConfig.sh (Node-ems:Dir-/usr/lpp/mmfs/samples/gss) on the protocol_node_Nodeclass with pagepool set to 32GiB
  gssClientConfig.sh -P 32768 <protocol_node_class>
- On InfiniBand networking, enable GPFS verbsRdma and verbsPorts to the correct IB HCA/ports
  mmchconfig maxFilesToCache=2M –N <protocol_node_class>
  mmchconfig maxMBpS=20000 –N <protocol_node_class>
  mmchconfig socketMaxListenConnections=8192 –N <protocol_node_class>
  mmchconfig envVar="MLX4_USE_MUTEX=1 MLX5_SHUT_UP_BF=1 MLX5_USE_MUTEX=1“ –N <protocol_node_class>
- GPFS pagepool was increased to 32GiB so that NFS/SMB server can benefit from GPFS caching.
- Increase of maxFilesToCache is a general best practice for protocol nodes to cache the file inodes for recently used files that have been closed and thereby improve the NFS and SMB performance.

➔ This set of tunables is optimized for IBM Elastic Storage Server (ESS) and Spectrum Scale CES and needs to be applied for genomic workload.

Tuning Guidelines – Protocol Nodes

blue ~ tuning applied for genomic workload
grey ~ default settings for ESS client node

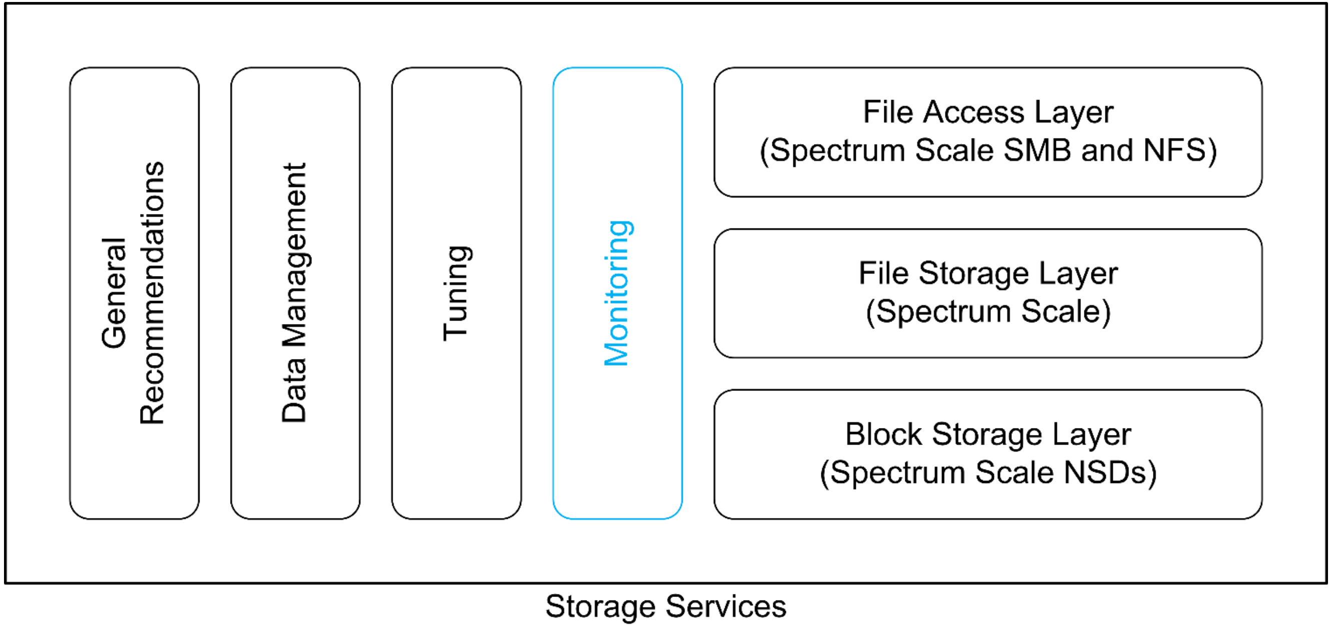
Snip of mmlsconfig:

[protocol]
pagepool 32768M
numaMemoryInterleave yes
maxFilesToCache 2M
maxStatCache 0
maxMBpS 20000
workerThreads 1024
ioHistorySize 4k
verbsRdma enable
verbsRdmaSend yes
verbsRdmasPerConnection 256
verbsSendBufferMemoryMB 1024
ignorePrefetchLUNCount yes
scatterBufferSize 256k
nsdClientCksumTypeLocal ck64
nsdClientCksumTypeRemote ck64
socketMaxListenConnections 8192
envVar MLX4_USE_MUTEX=1 MLX5_SHUT_UP_BF=1 MLX5_USE_MUTEX=1
verbsPorts <active_verbs_ports>

[common]
cipherList AUTHONLY
adminMode central

➔ This set of tunables is optimized for IBM Elastic Storage Server (ESS) and Spectrum Scale CES and needs to be applied for genomic workload.

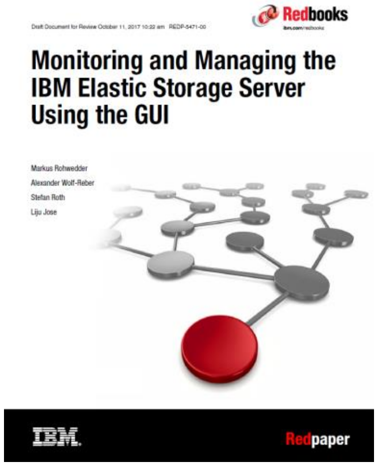
Management of Storage Services



➔ A Data Management GUI enables efficient configuration and monitoring of the storage resources.

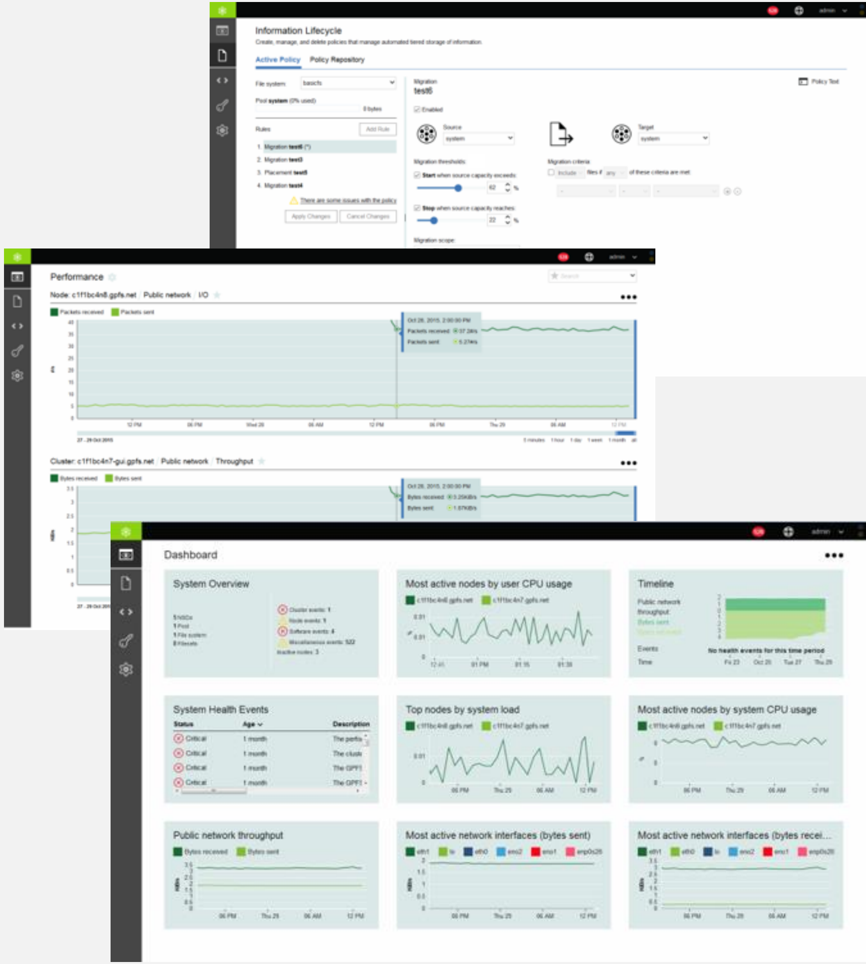
Spectrum Scale GUI

- Reduces administration overhead
  - Graphical User Interface for common tasks
  - Guided interfaces for common tasks
  - Supports Spectrum Scale and ESS
- See Redpapers for Monitoring Best Practices



http://www.redbooks.ibm.com/abstracts/redp5418.html

http://www.redbooks.ibm.com/redpieces/abstracts/redp5471.html



IBM Spectrum Scale

Spectrum Scale Best Practices Guide for Genomic Medicine Workload 1.0 (Private Network Services)

Dec 5th, 2017 – v3

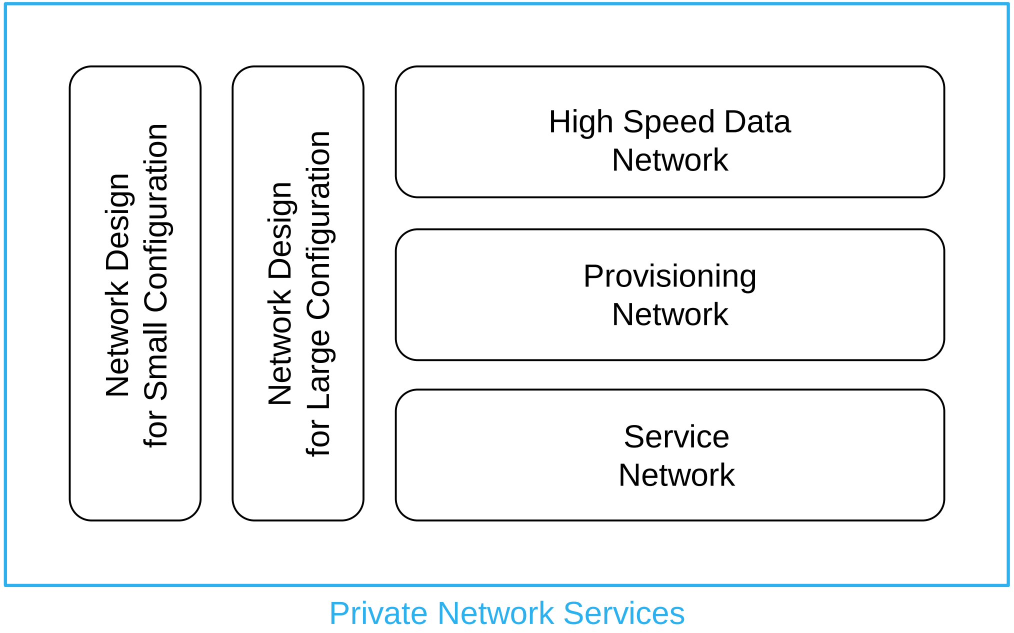
Summary

- The Spectrum Scale Blueprint for Genomic Medicine Workload describes Compute Services, Storage Services and Private Network Services. This section describes the Best Practices for Private Network Services.
- The Spectrum Scale Blueprint for Genomic Medicine Workload is optimized for the "Broad Institute GATK Best Practices on IBM reference architecture". Though, most of the recommendations apply to other workloads.
- The Spectrum Scale Blueprint for Genomic Medicine Workload is based on InfiniBand.
- Contact the Genomics War Room for help with questions on using different network technologies.

Outline

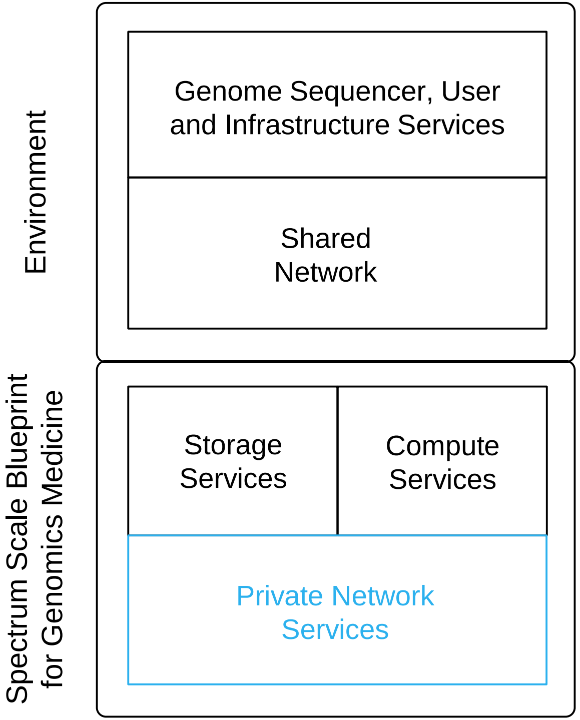
1. ***Composable building blocks***
2. Building block details

Network Services – Composable Building Blocks



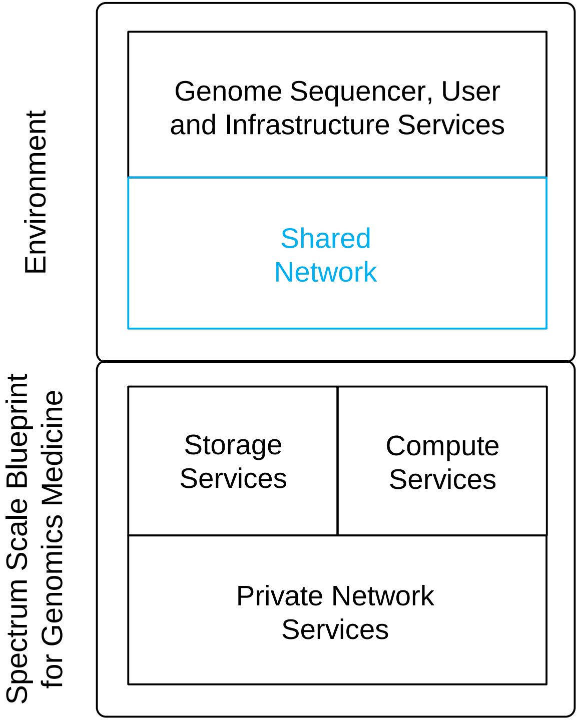
➔ A set of expertly engineered building blocks enable IT architects to compose solutions that meet customers varying performance and functional needs.

Private Network Services – Capabilities



- To integrate the Compute Services and 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 connected to the network with at least two links for high availability.
    - **Compute Nodes** are connected to the network with one port or with two ports if you want high availability.
  - **Provisioning Networks** for provisioning and in-band **management** of the storage and compute components and for **administrative login**.
  - **Service Networks** for out-band management and monitoring of all solution components.
  - A **Scalable Design** that can **start small starter** and grow to a large configuration that consists of **hundreds of compute nodes** and **tens of PB of storage**.

Shared Network

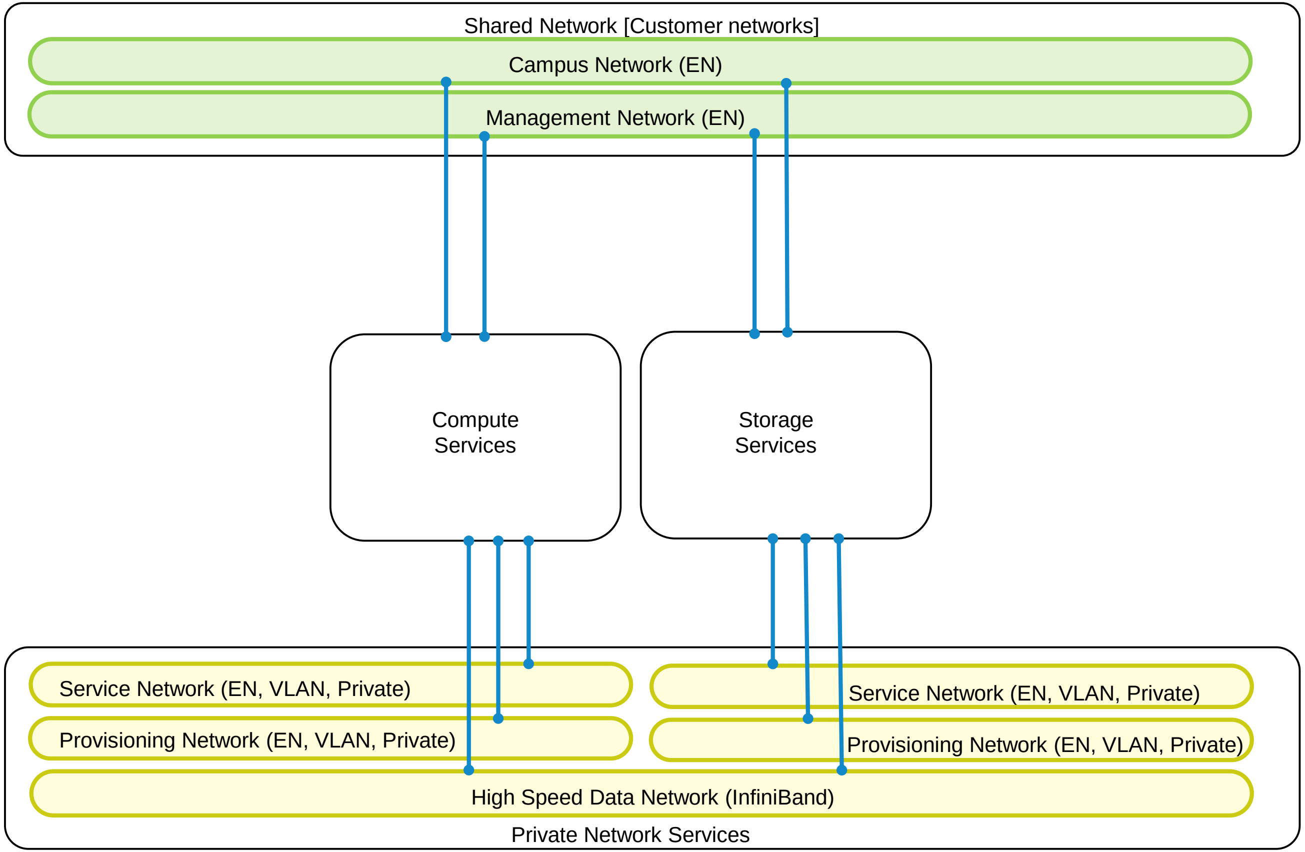


- The **Shared Network** connects the components of the Spectrum Scale Blueprint with the customers environment and services.
- It is customers responsibility to provide the Shared Network.
- The Shared Network typically includes a **Campus Network** and a **Management Network.**
  - The Campus Network is usually a public network that is externally visible from the cluster. It is the primary path for users to access the system. Users access the **Workload Management GUI** over the campus network. The campus network is also the default path for movement of data into and out of the system via NFS and SMB provided by the **Storage Services**.
  - The management network is used by **administrators** or other **privileged users** to access elements that are not intended to be accessible to users. The management network is also used to connect to **infrastructure services** like **NTP**, **DNS, and authentication service**.
- Some customers deploy separate campus and management networks whereas some customers combine the two. This blueprint can support either environment.

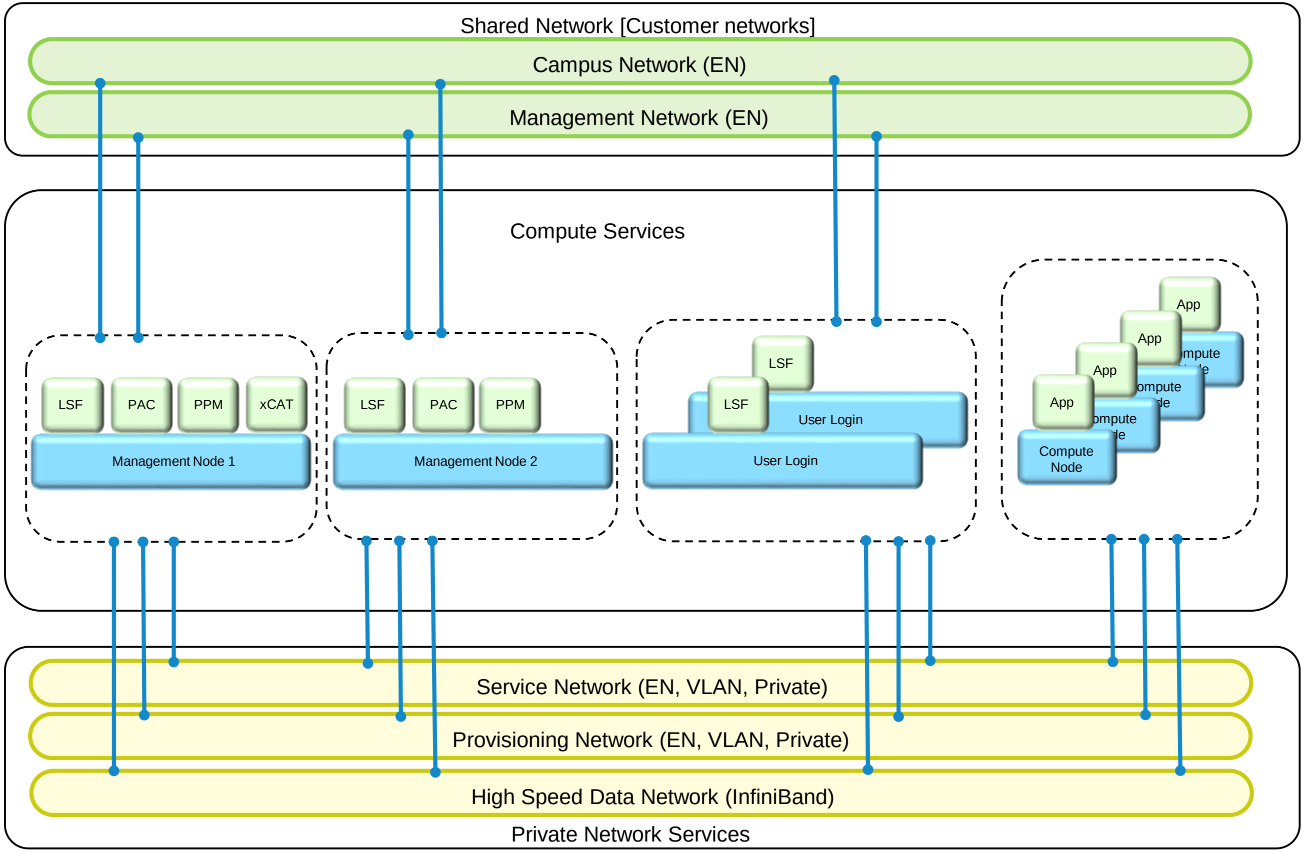
Private Network Services – Solution Elements

Capability	Provided by
A high speed data network for application communication and data access	InfiniBand
Provisioning networks for provisioning and in band management of the storage and compute components	1Gb Ethernet
Service networks for out-of-band management and monitoring of the solution components	1Gb Ethernet
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 multiple storage building blocks	Ready-to-use network layouts

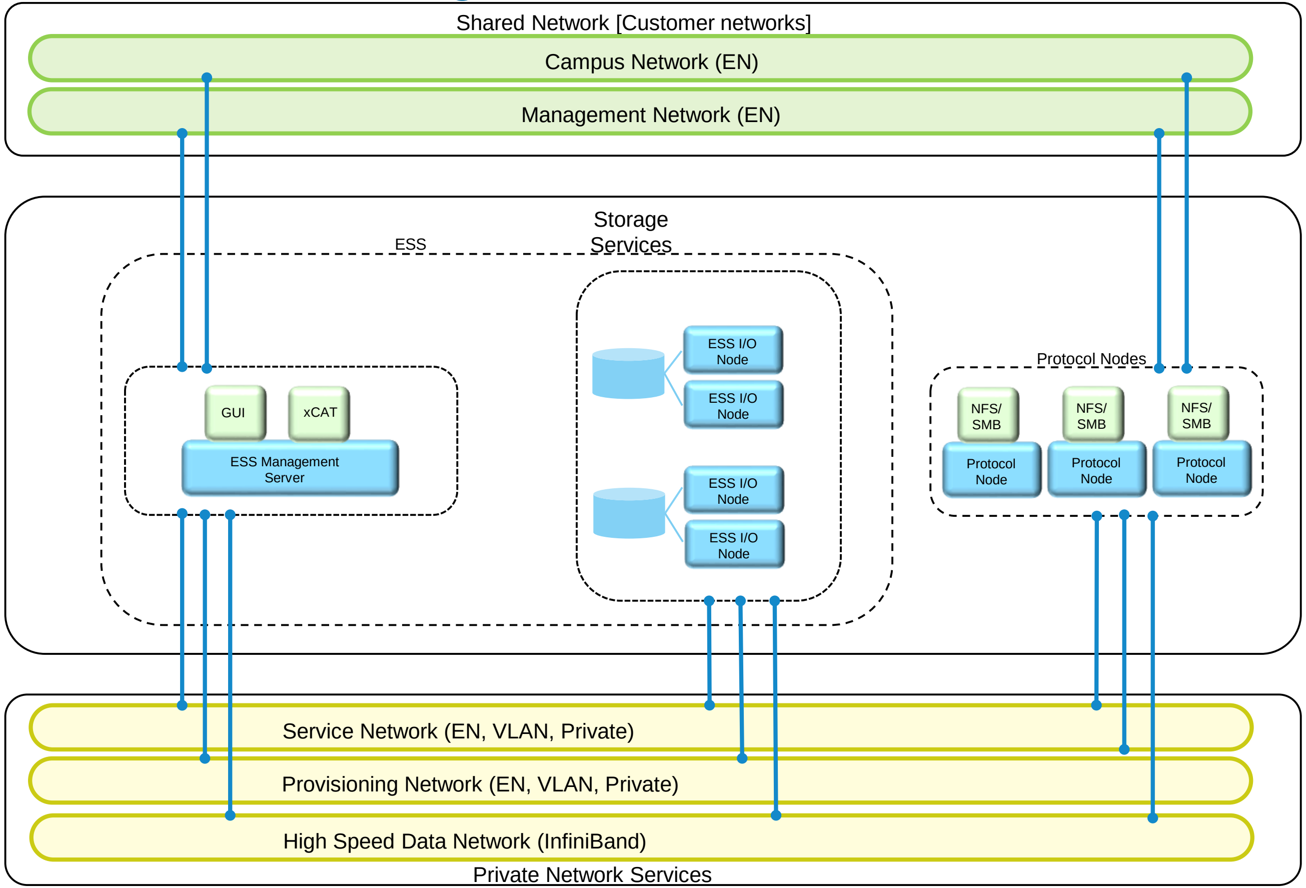
Network Services – Overview



Network Services – Compute Services



Network Services – Storage Services

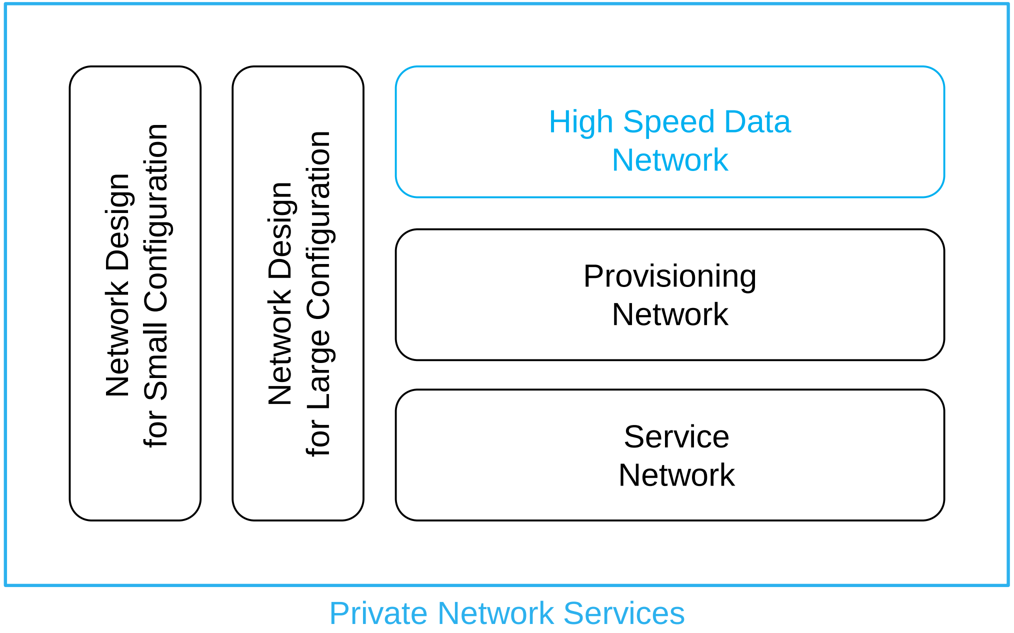


Outline

IBM Spectrum Scale

1. Composable building blocks
2. ***Building block details***

High Speed Data Network



➔ ***A dedicated and stable High Speed Data Network connects all Spectrum Scale nodes to provide high speed data access.***

High Speed Data Network – General Guidelines

Spectrum Scale requires two networks:

- Spectrum Scale daemon network
- Spectrum Scale admin network

What is the **Spectrum Scale daemon network**?

- The Spectrum Scale daemon network is used for communication between the mmfsd daemon of all nodes.
- The Spectrum Scale daemon network requires TCP/IP.
- In addition to TCP/P, Spectrum Scale can be optionally configured to use RDMA for daemon communication. TCP/IP is still required, if RDMA is enabled for daemon communication.
- The performance of Spectrum Scale depends on the bandwidth, latency and reliability of the Spectrum Scale daemon network.

What is the **Spectrum Scale admin network**?

- The Spectrum Scale admin network is used for the execution of administrative commands.
- The Spectrum Scale admin network requires TCP/IP.
- The Spectrum Scale admin network can be
  - the same network as the Spectrum Scale daemon network OR
  - a different network than the Spectrum Scale daemon network.
- The reliability of Spectrum Scale depends on the Spectrum Scale admin network.

High Speed Data Network – General Guidelines

Spectrum Scale is a clustered filesystem that **depends on a high performance, low latency and stable network**:

- The Spectrum Scale mmfsd daemon runs on each node which participates in a Spectrum Scale cluster.
- The mmfsd daemons of all cluster nodes need to communicate with each other to maintain a global cluster state which includes distributed file and directory locks and a distributed cache. This requires low latency RPC communication and high throughput daemon communication between all Spectrum Scale Nodes.
- Non-blocking network fabrics meet Spectrum Scale's network requirements. Non-blocking network fabric means that the throughput between two nodes is not constrained by inter switch links.

It is a best practice to connect all Spectrum Scale nodes via a **dedicated private high speed data network** for Spectrum Scale management traffic and Spectrum Scale data transfer:

- The private network is **not connected to external networks** such as the data center network or the internet.
- Experience in the field has proven that using the existing data center network can be problematic since most shared networks are not designed for high-throughput and low latency I/O. Other activity on the shared network can cause Spectrum Scale to degrade (e.g. node failures, long running commands).
- Experience in the field has proven that running this network over shared infrastructure can be problematic. Features like VLAN and Quality of Service on shared links need to be configured carefully to support all protocols and ports used by Spectrum Scale.
- Spectrum Scale nodes can be connected to multiple networks to connect them to other servers and services.

High Speed Data Network – Guidelines for Genomic Workload

- It is recommended to use Mellanox InfiniBand EDR (100GBit/s) switches for the High Speed Data Network.
- Storage Cluster
  - All Storage Cluster nodes are connected for high availability and non blocking
  - All Storage Cluster nodes are connected with InfiniBand EDR (100GBit/s)
  - ESS GS2S I/O nodes: four InfiniBand EDR links per node
  - ESS GL6S I/O nodes: six InfiniBand EDR inks per node
  - ESS Management Node (EMS): two InfiniBand EDR links per node
  - CES Protocol nodes: two InfiniBand EDR links per node
- Compute Cluster
  - It is sufficient to connect Compute Cluster nodes with Infiniband FDR (56GBit/s)
  - Cluster Management nodes are connected for high availability (at least two links)
  - Worker Nodes can be connected with one InfiniBand link to reduce cost or with two InfiniBand links for high availability
  - Cluster Management nodes: two InfiniBand EDR or FDR ports per node
  - Compute nodes: one or two InfiniBand EDR or FDR ports per node
- The Spectrum Scale daemon network is provided by InfiniBand
  - IPoIB is enabled to provide TCP/IP
  - Bonding (active/passive) is enabled on nodes that have more than one InfiniBand link
  - RDMA will be enabled on all nodes
- The Spectrum Scale admin network will use TCP/IP over the same IPoIB network.

High Speed Data Network – Miscellaneous

Fabric Management

- Each InfiniBand fabric requires a subnet manager.
- See the Network Designs for details.

Monitoring

- Detailed monitoring of InfiniBand networks requires Mellanox Unified Fabric Management (UFM) software.
- UFM requires a software license and an x86-64 server.
- UFM is outside the scope of this blueprint.

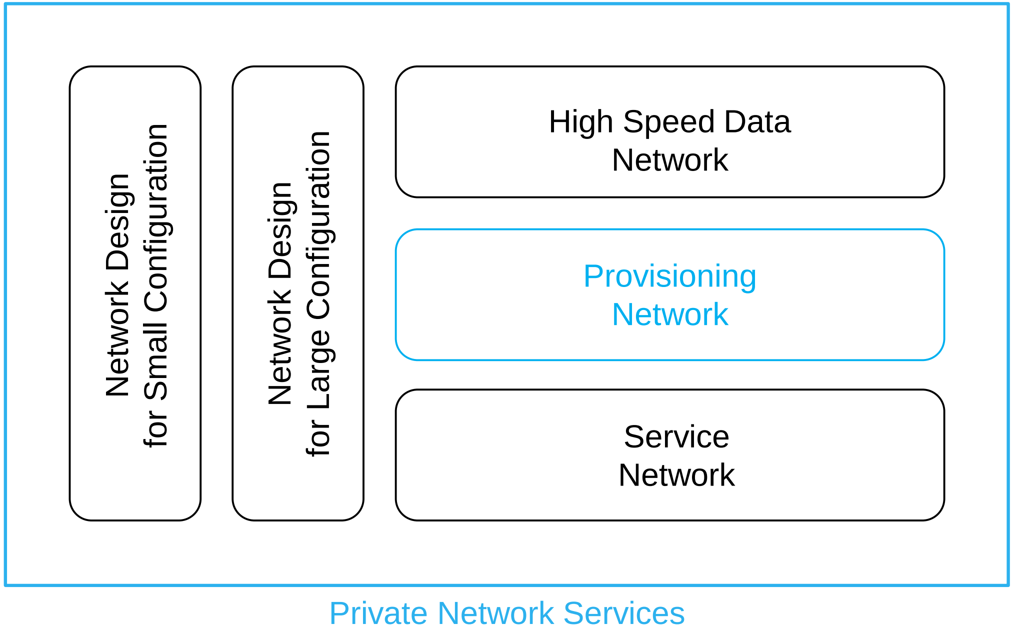
Application Traffic

- Some parallel applications require a high-speed network for inter process communication (e.g. MPI).
- Most genomic applications are single node jobs.
- Spectrum Scale Daemon Network and application traffic (e.g. MPI) share the same InfiniBand network. A misbehaved application can impact the stability and performance of Spectrum Scale. Administrators should monitor for such applications and limit their impact.

IP Addresses

- Spectrum Scale depends on Static IP Addresses
- The Storage Services provide static IP Addresses for all storage nodes.
- It is customer responsibility to provide static IP Addresses for all compute nodes.

Provisioning Network



➔ *The Provisioning Networks enable provisioning and in-band management of the storage and compute components.*

Provisioning Network

General Remarks

- The Provisioning Network is also known as the xCAT network.
- The Provisioning Network is a private network that is used by the cluster manager (e.g. xCAT) to provision the compute and storage components of the solution and subsequently manage and monitor those components.
- There are separate Provisioning Networks for Storage Services and Compute Services.

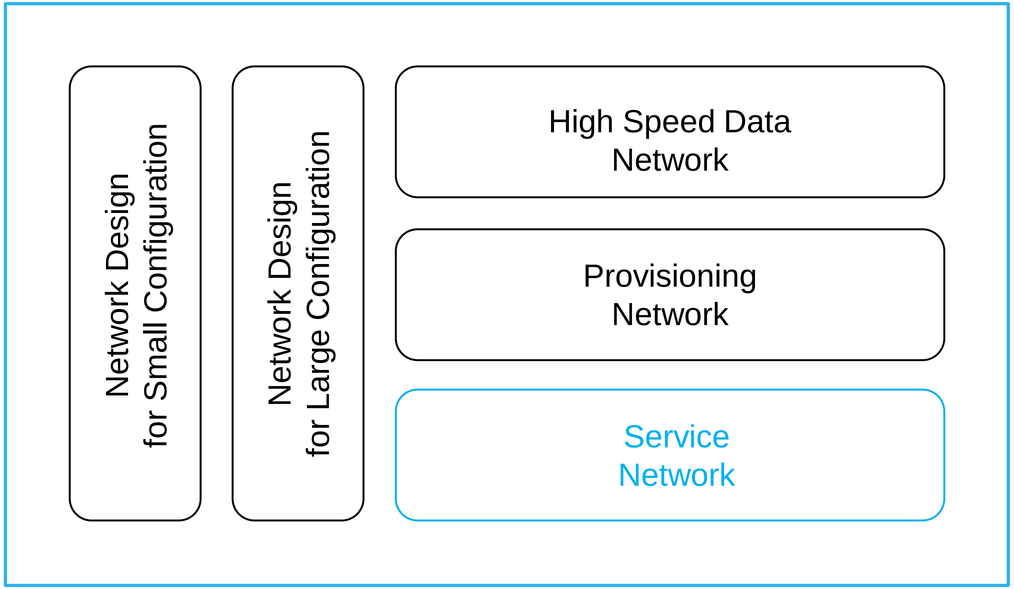
Storage Services

- The Storage Cluster requires a dedicated private Provisioning Network.
- All nodes in the Storage Cluster need a single connection to the Provisioning Network.
- DHCP is used to assign static IP addresses for all interfaces on the Provisioning Network.
- SNMP monitoring of the Provisioning Network components is out of scope for this blueprint.
- HA for the Provisioning Network is out of scope for this blueprint.
- IPv6 is disabled on the Provisioning Network interfaces on all nodes.

Compute Services

- Compute Nodes cannot be connected to the Provisioning Network for the Storage Services. Each Provisioning Network includes its own DHCP server and DHCP does not allow to have two DHCP server on the same network.
- In most cases the customer already have a Provisioning Network in their data center. Otherwise a Provisioning Network for the Compute Nodes must be configured.

Service Network



Private Network Services

➔ ***The Service Networks enable out-of-band management and monitoring of all solution components***

Service Network

General Remarks

- The Service Network is typically a private Ethernet network that is used to access the management processors of the servers within the system.
- A management processor can be an FSP (typical for Power Systems servers) or a BMC (typical for OpenPower and x86-64 servers).
- A cluster manager can use a protocol like IPMI to do hardware discovery, power control, and out of band management and monitoring of the solution components.
- There are separate Service Networks for Storage Services and Compute Services.

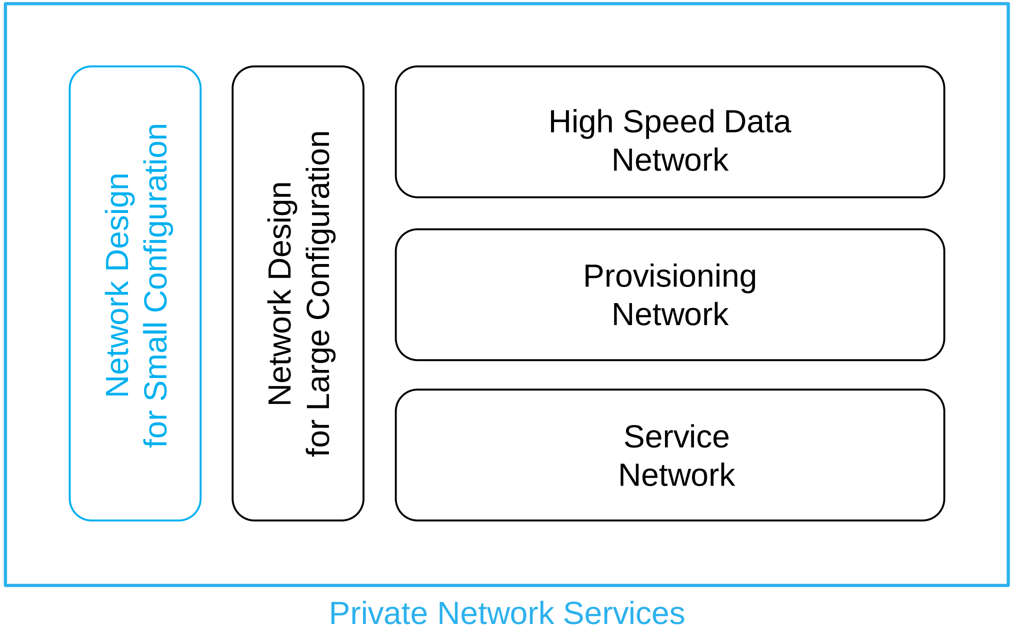
Storage Services

- The Storage Cluster requires a dedicated private Service Network.
- The ESS Management Server (EMS) needs a single connection to the Service Network.
- The FSP port of each storage node (ESS, CES) needs a single connection to the Service Network. The FSP port of the EMS is optionally connected to a customer provided Service Network.
- DHCP is used to assign dynamic IP addresses to the FSP ports that are part of the Service Network.
- SNMP monitoring of the Service Network components is out of scope for this blueprint.
- HA for the Service Network is out of scope for this blueprint.

Compute Services

- Compute Nodes cannot be connected to the Service Network for the Storage Services. Each Service Network includes its own DHCP server and DHCP does not allow to have two DHCP server on the same network.
- In most cases the customer already have a Service Network in their data center. Otherwise a Service Network for the Compute Nodes must be configured.

Network Design for Small Configuration



➔ The small configuration is more easy to use,
but is limited by the numbers of ports of a single InfiniBand switch.

Network Design for Small Configuration

The High Speed Data Network for the Storage Services and the Compute Services comprises:

- A pair of IB EDR (8828-E36) switches to support redundant data network.
- The InfiniBand subnet manager will be configured on the CES Protocol Nodes.
  - See next charts for details.
- The switches can be ordered with an ESS.

The Provisioning Network for the Storage Services that comprises:

- A 1Gb Ethernet (8831-S52) switch that is shared with the Service Network.
- Use untagged VLAN to separate Provisioning Networks and Service Networks.
- The switch is configured with spanning tree disabled.
- The switch can be ordered with an ESS.

The Service Network for the Storage Services that comprises:

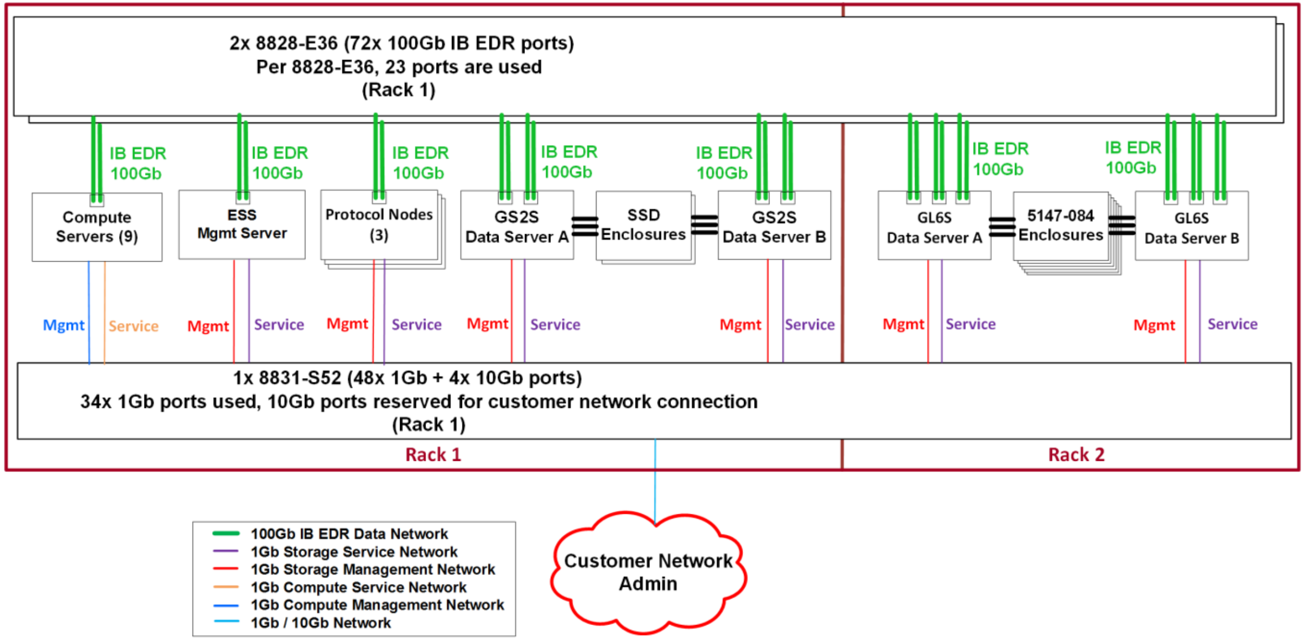
- A 1Gb Ethernet (8831-S52) switch that is shared with the Provisioning Network.

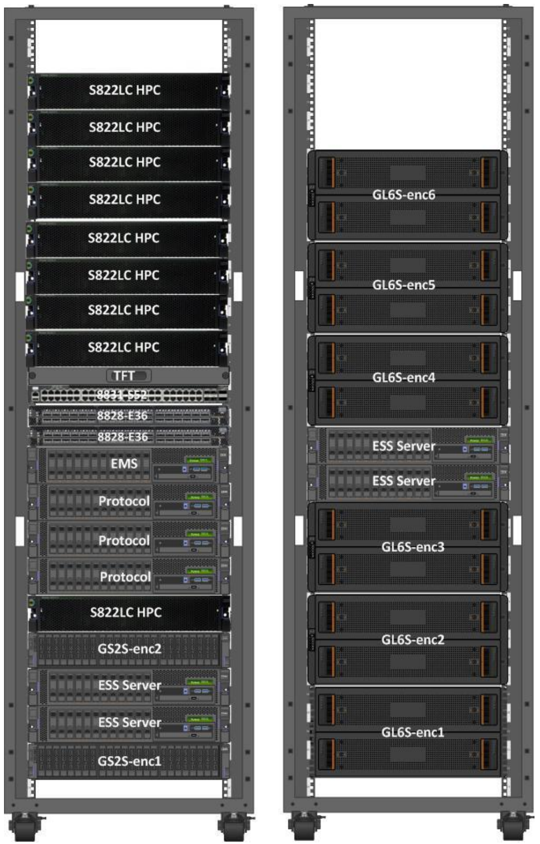
➔ Use the Small Configuration only, if you do not plan to grow the storage nodes and the compute nodes to exceed the number of available InfiniBand switch ports.

Example Configuration – Small for High Availability

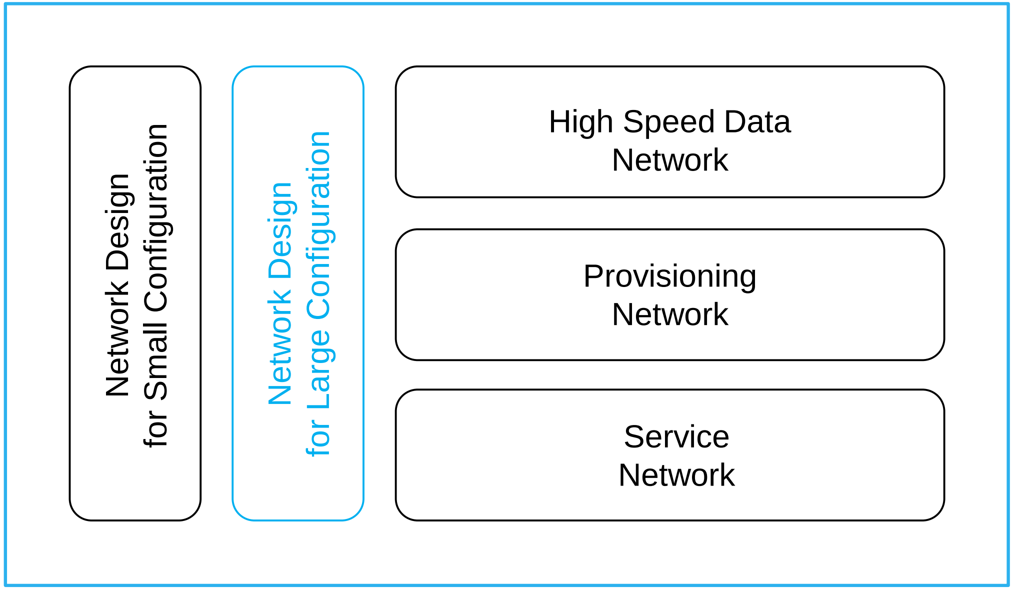
- Each 8828-E36 InfiniBand switch has 36 ports
  - 14 ports are used by Storage Nodes
  - There are no inter switch links
  - The remaining 22 ports can be used for the Compute Cluster
  - This configuration supports up to 20 User Login Nodes and Worker Nodes.
- The Storage Cluster requires 28 InfiniBand switch ports, 14 ports in each switch
  - 2x ports for ESS EMS
  - 8x ports for ESS GL2S
  - 12x ports for ESS GL6s
  - 6x ports for Protocol Nodes
- Each Compute Node is connected to both InfiniBand switches
  - 2x ports for each Compute Cluster Management Node (1x port per switch)
  - 2x ports for each User Login Node (1x port per switch)
  - 2x ports for each Worker Node (1x port per switch)
- The InfiniBand subnet manager will run on the first two Protocol Nodes.

Example Configuration – Small for High Availability





Network Design for Large Configuration



Private Network Services

➔ The large configuration scales up to hundreds of compute nodes and multiple storage building blocks.

Network Design for Large Configuration

- The Network Design for the Large Configuration will be added to a future version of the blueprint.

➔ Use the Large Configuration right from the beginning, if you plan to grow the storage nodes and the compute nodes to exceed the number of available InfiniBand switch ports of the Small Configuration.

➔ Contact the Genomics War Room for help with the Large Network Configuration.

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.