REFERENCE ARCHITECTURE FOR 50-100 CONCURRENT USERS

SQREAM DB REFERENCE ARCHITECTURES

v1.2

 sqream.com

EXECUTIVE SUMMARY

This document describes the necessary hardware and software considerations for a 50-100 user installation of SQream DB GPU accelerated data warehouse.

DOCUMENT PURPOSE

The purpose of this document is to describe the SQream DB reference architecture, emphasizing the benefits to the technical audience, while providing guidance for end-users on selecting the right configuration for a SQream DB installation.

This document was written in January 2019.

TARGET AUDIENCE

This document is intended for influencers and decision makers, IT and system architects, system administrators, and experienced users who are interested in a comprehensive reference for a SQream DB installation.

SERVERS

SQream recommends rackmount servers by server manufacturers Dell, HP, Cisco, Supermicro, IBM, and others.

A typical SQream DB node includes:

- Two-socket enterprise processors, like the Intel® Xeon® Gold processor family or an IBM® Power9 processor, providing the high performance required for compute-bound database workloads. See the appendix for other suggested processors.
- NVIDIA Tesla GPU accelerators, with up to 5,120 CUDA and Tensor cores, running on PCIe or fast NVLINK busses, delivering high core count, and high-throughput performance on massive datasets
- High density chassis design, offering between 2 and 4 GPUs in a 1U, 2U, or 3U package, for best-in-class performance per cm^2

STORAGE

SQream DB relies on OS-mountable file systems. A standalone node may use mountable filesystems on simple spinning disk redundant arrays, all the way up to 24 internal SSD or NVMe drives, providing up to 40GB/s per node, or up to 10GB/s per GPU.

The increased I/O results provide a significant performance increase for I/O-bound workloads, with the flexibility to choose the desired cost-performance, by opting for various SSD types and RAID configurations.

CONTENTS

INTRODUCTION

SQream has written this document to aid in designing and deploying SQream DB-based solutions, for installations of varying sizes. This document will also help identify the hardware components required in a SQream DB solution, to simplify the procurement process.

CONSIDERATIONS

In a SQream DB installation, the storage and compute are logically separated. While they may reside on the same machine in a standalone installation, they may also reside on different nodes, providing additional flexibility and scalability.

SQream DB requires CPU, RAM, and GPU to deliver the best performance. SQream recommends about 256GB of RAM per physical GPU.

SQream DB also requires some local disk space for temporary spooling, when performing intensive larger-than-RAM operations like sorting. SQream recommends an SSD drive, in mirrored RAID 1 configuration, with about 2x the RAM size available for temporary storage. This can be shared with the operating system drive.

SQream recommends that storage in a standalone machine be configured in RAID 5 or 10, for both performance reasons and resiliency.

When using SAN or NAS devices, SQream recommends around 5GB/s of burst throughput from storage, per GPU.

SOLUTION COMPONENTS

TERMINOLOGY

Host

A physical computer or server, with a CPU, RAM, GPU, and a network interface including IP address or hostname. Hosts don't share disk or RAM with each other.

Instance

An instance of SQream DB consists of the SQream DB daemon, allocated to a specific GPU on a host. Several instances of SQream DB can be run on a host or GPU.

Node

An instance of SQream DB that is configured to run an instance of SQream DB in a cluster. To maintain database safety and high availability, at least 3 nodes are required.

Cluster

A collection of hosts with the SQream DB software installed under the same configuration. This includes the cluster manager and load balancer (see below).

Load balancer

An SQL API endpoint that allows interactive and programmatic interfaces to interact with SQream DB and run SQL statements.

Cluster manager

A SQream DB application that manages the SQream DB storage and metadata and maintains database safety and correctness.

SQream DB Storage on a scale-out NAS

A filesystem that is readable and writeable by all cluster nodes, over the network.

Figure 1 below illustrates the server and network components of a SQream DB highly available architecture.

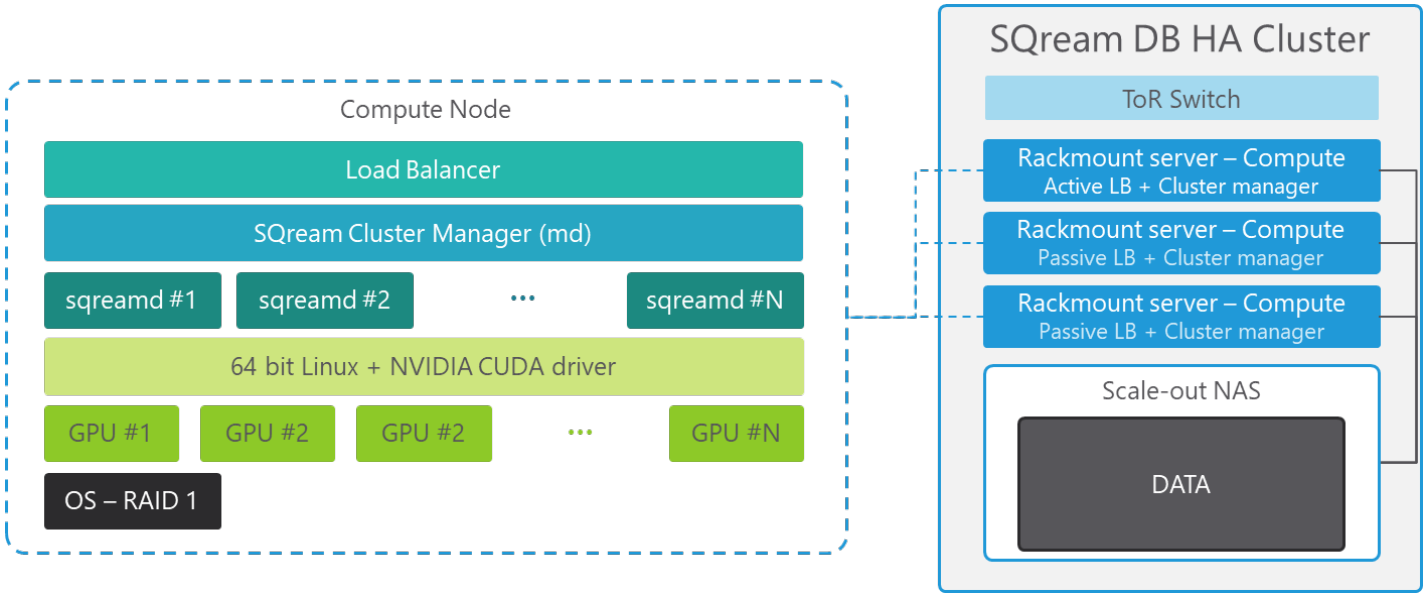

Figure 1 - SQream DB reference architecture - conceptual diagram

HIGH AVAILABILITY FEATURES

The highly available options for this reference architecture configuration are:

- SQream DB Cluster Manager (metadata server) – This piece of the architecture holds the key to the deiverse data, tables, and metadata operations of SQream DB. It also authenticates users, provides management, and coordinates transactions - synchronizing hundreds of SQream DB nodes for processing data, loading and deleting, without impacting running queries.
- Load Balancer – This element provides a floating single end-point for all connections to SQream DB. The load balancer always knows which SQream DB instances are active, and can point clients to available instances, limiting resource contention.
- Cluster utilities – At any given moment, there is one Cluster Manager and one Load Balancer active, on the same IP endpoint. SQream installations rely on floating IPs and cluster management utilities to ensure passive nodes can take over an active node, should it fail.
- Operating system reliability – For maximum reliability, the operating system disks are in RAID1 configuration, preventing failure of the system from hard disk failures.
- Data reliability – SQream recommends using high-availability NAS or SAN storage, capable of delivering 5/9 up-time and durability.
- Network reliability – This architecture relies on a ToR switch, although another switch can be added for resiliency.
- Power reliability – All SQream servers have adequate power redundancy by having two hot-swappable power supplies. SQream recommends having two PDUs in each rack.

CLUSTER DEPLOYMENT CONSIDERATIONS

Prior to designing and deploying a SQream DB cluster, there are a number of important factors to consider. This section provides a breakdown of deployment details intended to help ensure that this installation exceeds or meets the stated requirements. The rationale provided includes the necessary information for modifying configurations to suit the customer use-case scenario.

Component	Value
Compute - CPU	Balance price and performance
Compute – GPU	Balance price with performance and concurrency
Memory – GPU RAM	Balance price with concurrency and performance
Memory - RAM	Balance price and performance
Operating System	Availability, reliability, and familiarity
Storage	Balance price with capacity and performance
Network	Balance price and performance

SQREAM DB COMPUTE

SQream DB relies on CPU and GPU resources for compute.

When computation performance is a primary concern, SQream recommends a higher core-count NVIDIA Tesla V100 GPU, and a higher clock-speed/cache CPU configuration for compute nodes.

Compute – CPU

SQream DB relies on multi-core Intel ® Xeon ® Processors. SQream recommends a dual-socket machine populated with the Intel ® Xeon ® Gold 6140 2.3GHz processor, having 18C/36T.

While a higher core count may not necessarily affect query performance, more cores will enable higher concurrency and better load performance.

Compute – GPU

The NVIDIA Tesla range of high-throughput GPU accelerators provides the best performance for enterprise environments. Most cards have ECC memory, which is crucial for delivering correct results every time.

As such, SQream recommends the NVIDIA Tesla V100 32GB GPU for best performance and highest concurrent user support.

It is possible to select GPUs with less RAM, like the NVIDIA Tesla V100 16GB or P100 16GB. However, the smaller GPU RAM available will result in reduced concurrency, as the GPU RAM is used extensively in operations like JOINs, ORDER BY, GROUP BY, and all SQL transforms.

OPERATING SYSTEM

SQream DB can run on 64-bit Linux operating systems:

- Red Hat Enterprise Linux (RHEL) v6.7 and v7.5
- Amazon Linux 2017.09, 2018.03
- CentOS v6.7 and v7.5
- Ubuntu v14.04 LTS, v16.04 LTS, and V18.04 LTS
- Other Linux distributions are supported via nvidia-docker

The reference architectures listed here were tested with CentOS v7.5.

MEMORY

Use of error-correcting code memory (ECC) is a practical requirement for SQream DB and is standard on most enterprise server. SQream DB benefits from having large amounts of memory for improved performance on large 'external' operations like sorting and joining.

Although SQream DB can function with less, for full dedicated access SQream recommends a key of 256GB of RAM per GPU in the machine. Therefore, we recommend up to 1,024GB of RAM for a 4 GPU machine.

STORAGE

In a SQream DB installation, the storage and compute are logically separated.

While the storage may theoretically reside on the same machine (for example, as in Error! Reference source not found. – SQream DB standalone installation), it is recommended that they may also reside on different nodes, providing additional flexibility and scalability.

Capacity

The number of disks, and the disk capacity determine the storage capacity available for the SQream DB installation.

Redundancy

For clustered scale-out installations, SQream DB relies on scale-out NAS/SAN storage. These devices have extremely high reliability and durability, with five 9s of up-time. This ensures that blocks of data are replicated between disks, so that failure of several disks will not result in data loss or availability of the system.

SQream DB supports POSIX file-systems like EXT4, XFS, Lustre, and NFS mounts.

I/O Performance

By having more disks in a storage system, it becomes less likely that SQream DB will have multiple queries/processes accessing a given disk at the same time.

More disks result in reduced I/O queues, requests, and bottlenecks, resulting in better I/O performance.

Configuration

All SQream DB nodes are configured the same way, for redundancy and scalability.

Because storage reliability is important, SQream recommends enterprise-grade SAS SSD drives. However, as with other components – there is a tradeoff for cost/performance. When performance and reliability are important, SQream recommends SAS SSD or NVMe drives. SQream DB can also function with more cost-effective SATA drives or even large spinning-disk arrays.

NETWORK

In a standalone system, a failure of a networking link can result in loss of access to SQream DB. While no data will be lost or corrupted, this harms the system's uptime. In a clustered system, it could cause queries to fail, as connection to the cluster manager could be lost.

As such, SQream recommends using redundant ToR switches and redundant ethernet links for production environments, for additional redundancy.

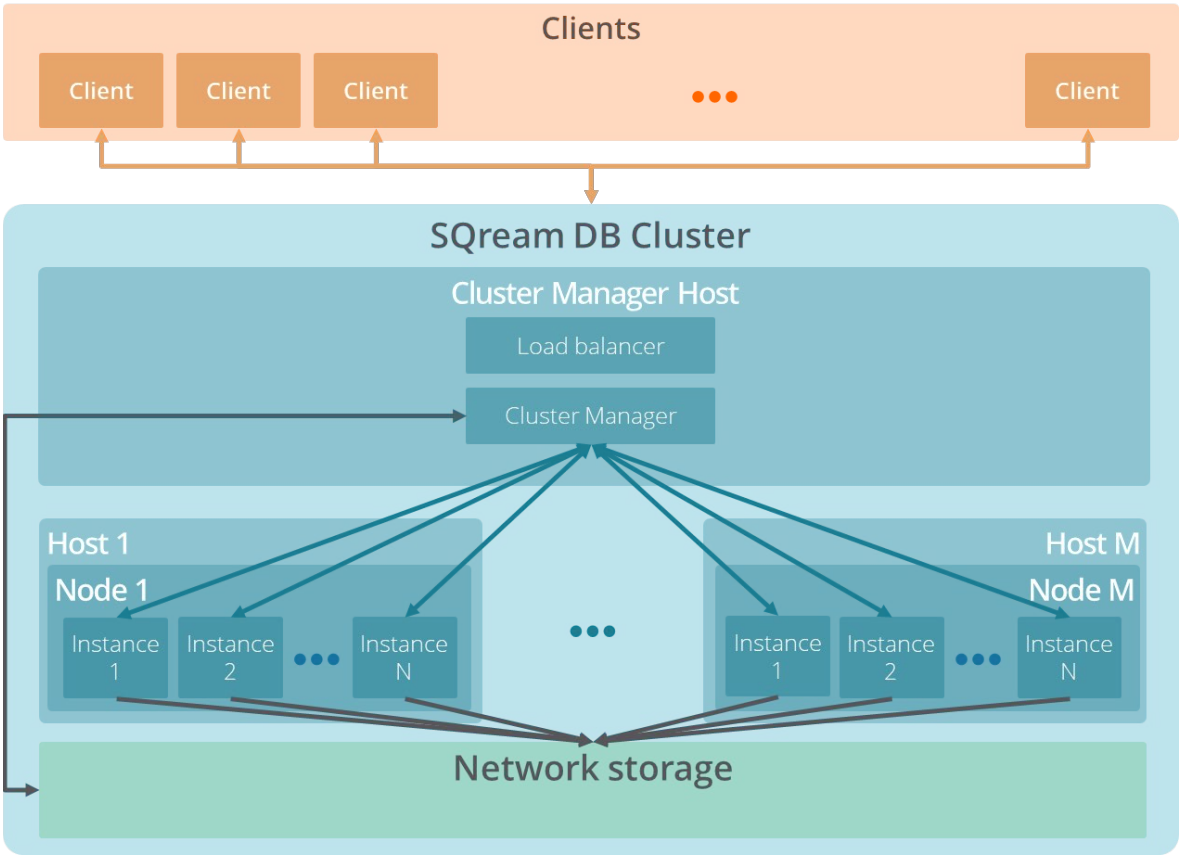

Figure 2 - SQream DB conceptual network diagram, showing active network links

Connection to BI and SQL Clients

SQream also recommends that link aggregation be configured between switches. The best way to configure link aggregation is to bond the two physical NICs on each server, each to a different ToR switch. When done properly, this technique allows the bandwidth of both links to be used. If either of the switches fail, the servers still have full network functionality.

Switch failures can be further mitigated by incorporating dual power supplies for the switches.

The bandwidth and latency provided by two bonded 10 Gigabit Ethernet (GbE) connections from SQream DB compute-nodes to SQL users is sufficient.

Connections to NAS/SAN

For connections to the NAS/SAN, it is recommended that a higher throughput connection be used.

SQream recommends redundant, bonded InfiniBand or 40GigE/100GigE uplinks to the storage core aggregation switches. However, these are dependent on the storage solution selected, as not all solutions support these uplinks.

Link aggregated (bonded) 10GigE connections are also possible for connection to the storage network, but this may result in network bottlenecks. Consult your SQream solutions architect for more information about this setup.

Cluster communication

SQream DB hosts may communicate with one another to ascertain which nodes are active, and properly balance the query workload. This network can be shared with the client network, but it is recommended that this network be separated logically with a separate VLAN.

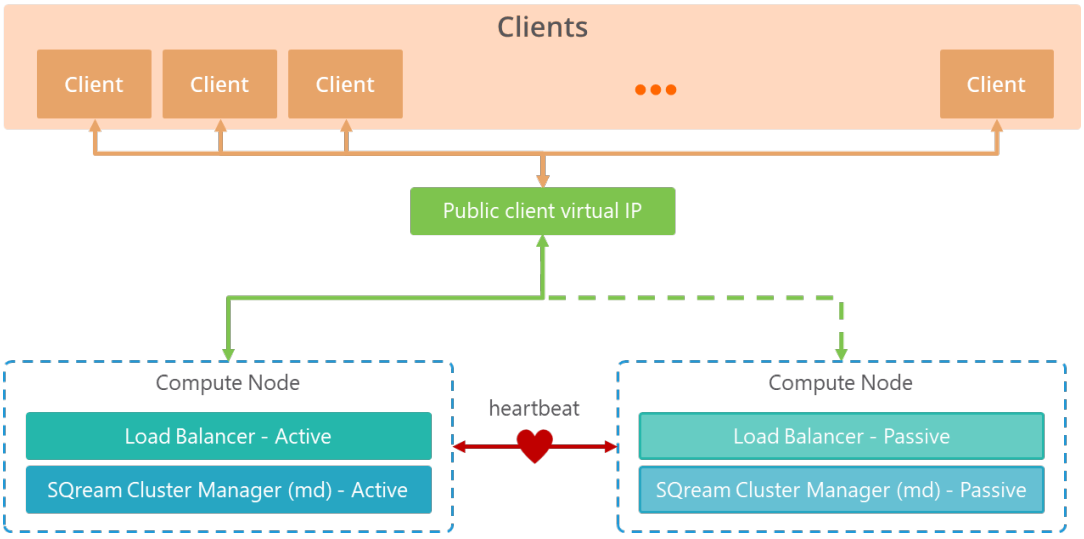

Figure 3 - Conceptual cluster communication diagram

SQream DB clustering relies on RedHat's Pacemaker clustering utilities for managing the active/passive components of the SQream DB cluster.

RedHat recommends the following ports be opened for communication between the active and any passive node in the cluster:

Service	Active on host	Default port assignment
Pacemaker daemon `pcsd`	Any host performing cluster management duties	TCP 2224
Pacemaker remote nodes `pacemaker_remoted`	Any host performing cluster management duties	TCP 3121
Quorum devices `corosync-qnetd`	Any host performing cluster management duties	TCP 5403
`corosync` multicast	Any host performing cluster management duties	UDP 5404
`corosync`	Any host performing cluster management duties	UDP 5405
DLM: `clvm` or `GFS2`	Any host performing cluster management duties	TCP 21064

It is not required that these ports be opened outside of SQream DB's host network.

ACCESS CONFIGURATION

It is important to isolate SQream DB scale-out clusters on the network so that external network traffic does not affect the performance of the cluster. Doing so also allows the cluster to be managed independently from that of its users, which ensures that only the cluster administrator can make changes to the cluster configurations. SQream recommends isolating the cluster nodes into a private subnet or VLAN.

REFERENCE ARCHITECTURE

MULTI-SERVER REFERENCE ARCHITECTURE

The multi-server architecture extends SQream DB's scalability. The design allows for additional scale-out for multi-rack clusters when needed.

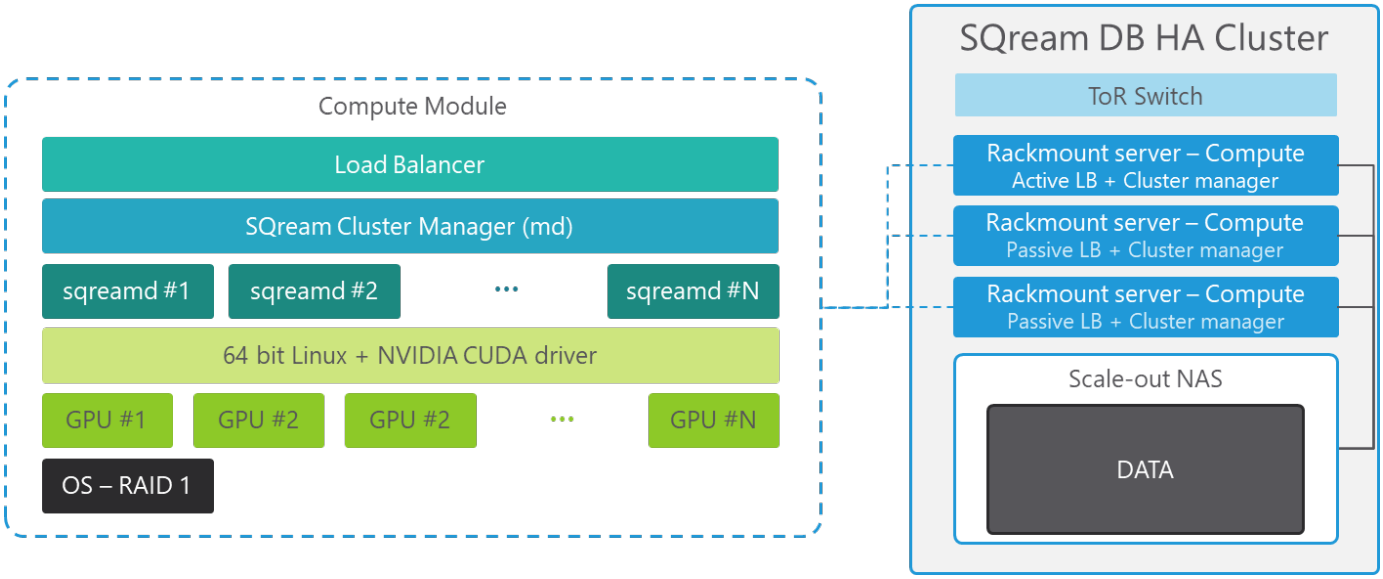

Figure 4 - SQream DB clustered installation

Rack Setup

The SQream DB HA Cluster rack contains more than one approved chassis, 2x ToR switches, and NAS/SAN storage within a 42U rack.

Network

The clustered SQream DB nodes are to be connected via a dedicated fabric to the storage appliance. SQream recommends redundant, bonded InfiniBand or 40GigE/100GigE uplinks to the storage core aggregation switches. However, these are dependent on the storage solution selected, as not all solutions support these uplinks.

A second connection fabric must be available for database end-users (BI/SQL). The bandwidth and latency provided by two bonded 10 Gigabit Ethernet (GbE) connections from SQream DB compute-nodes to SQL users is sufficient.

See the network section on page 9 above for further information about the considerations.

Software

Each SQream DB server will be configured as a cluster manager and SQream database nodes. At any given time, any SQream DB server may act as the cluster manager. This is done via clustering software installed by SQream.

Aside from the operating system, each node must contain the NVIDIA CUDA driver.

Power and Cooling

When planning scale-out SQream DB clusters, power redundancy is important. To ensure that the servers and racks have sufficient power redundancy, SQream recommends more than one PDU.

For each server, SQream recommends that each redundant power supply is connected to a different PDU.

CAPACITY AND SIZING

The capacity and sizing of a SQream DB cluster is no different than other disk-based data warehouses.

Storage sizing requires identifying current and future needs. For proper planning, it is recommended that the following issues be documented:

- Data sources
- Data frequency
- Raw size-on-disk for data sets before ingest into SQream DB
- Compressed Size-on-disk for data-sets after ingest into SQream DB
- Space for intermediate files stored on disk (spooling, temporary files)

SIZING THE SOLUTION

Calculate the storage needs:

1. Identify the size in terabytes of data – per day, per week, per month, per year. Add the ingest rate of all data sources.
2. Identify storage requirements for the short term, and long term.
3. Identify data retention decisions (size/duration).
   Which tables/data sources are required to be kept and for how long?
   Consider the maximum fill-rate and file system format space requirements on hard drives.

ANALYSIS RECOMMENDATIONS

You may have a specific requirement for usable SQream DB table capacity.

In general, consider a 10% reduction when copying data from raw sources like CSV to SQream DB's own format (uncompressed). Additional SQream DB compression will result in a factor of around 1:4-1:5 from CSV to compressed size-on-disk. Conservatively, we will use 1:3.

Example: A system with a raw disk capacity of 500 TB. Subtract 15% from the raw space for intermediate files, results in 425 TB. Assuming conservative 1:3 compression, the usable raw disk space for SQream DB to use is 1,275 TB.

Following this rule of thumb helps determine how much storage is needed for a SQream DB deployment. Compression provides additional usable space, depending on the type of compression, and data locality.

SAMPLE CONFIGURATIONS

50-USER SYSTEM CONFIGURATION

Based on our experience, most users building a database cluster initially are not aware of the eventual profile of their workload. Often, the first jobs that an organization runs with SQream DB differs from the eventual workload. It's common for SQream DB installations to take on more work as familiarity and proficiency with the system improves. Building a cluster appropriate for the workload is critical to good SQream DB performance.

For a ~50-user configuration, the number of GPUs should match the number of users. SQream DB recommends 1 Tesla V100 GPU per 2 users, for full, uninterrupted dedicated access.

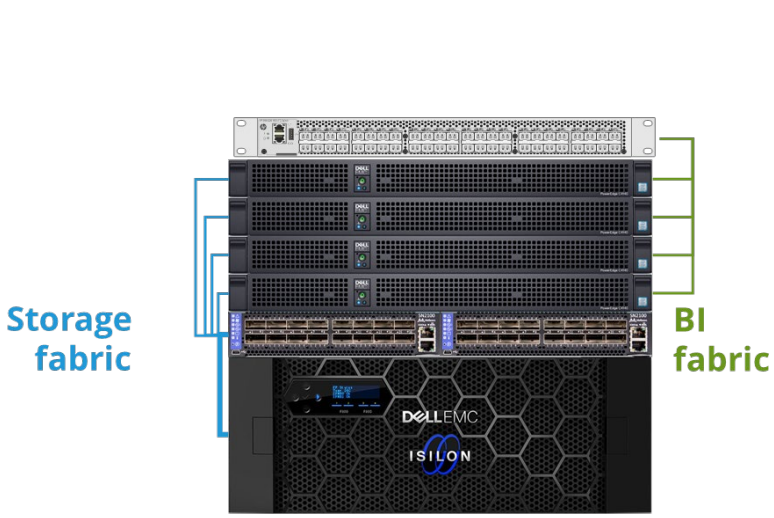

Figure 5 - SQream DB scale-up Dell-based cluster

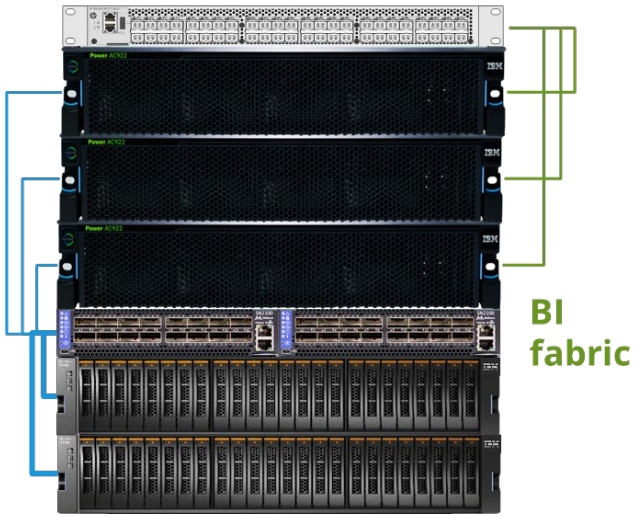

Figure 6 – SQream DB scale-up IBM Power9-based cluster

A SQream DB cluster for 50 users consists of the following components:

1. 4 high-density GPU-enabled servers, like the Dell C4140 (Configuration C) or IBM AC922 with 4x NVIDIA Tesla V100 32GB PCIe GPUs.
   Each server is equipped with dual Intel ® Xeon ® Gold 6140 CPU or IBM Power9 CPU, with 1,024GB of DDR4 RAM.
2. NAS/SAN storage, capable of delivering 1 GB/s per GPU.
   For the system above, with 4x4 NVIDIA Tesla V100 GPUs, this results in 16GB/s, over multiple bonded, 40GigE or InfiniBand links via a fabric switch.
3. ToR 10GigE ethernet switch for the BI fabric
4. 40GigE or InfiniBand switches for the storage fabric
5. At least 1 PDU

SCALING THE CONFIGURATION FOR ADDITIONAL USERS

Because SQream DB separates storage and compute, the cluster can be scaled to support additional users by joining nodes to the system. SQream DB's shared data architecture means data does not need to be redistributed. A node with SQream DB instances can join or leave the cluster at any time without affecting data availability, consistency, and durability.

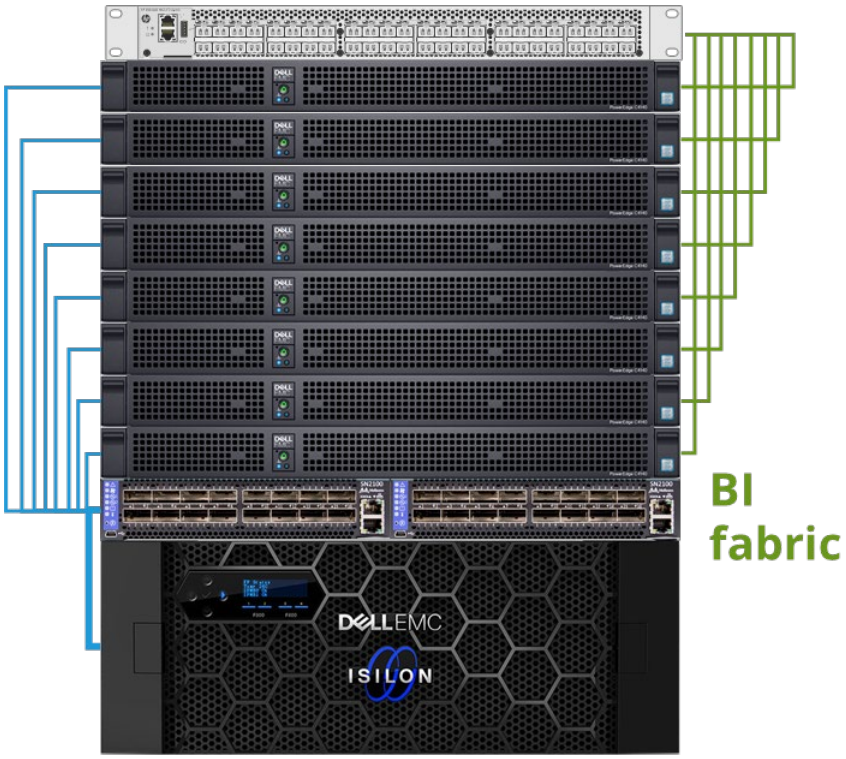

Figure 7 - SQream DB scale-up Dell-based cluster for 100 users