640TB Data Warehouse Fast Track Reference Architecture for Microsoft SQL Server 2017 using ATOS BullSequana S800 and Dell EMC VMAX 250F

Configuration and performance results

Abstract

This paper describes the design principles and guidelines used to achieve an optimally balanced 640TB Data Warehouse Fast Track reference architecture for SQL Server 2017 using ATOS BullSequana S800 servers and Dell EMC™ VMAX 250F all flash storage arrays.

September 2018

Revisions

Date	Author(s)	Description	Version
20/10/2018	Roger Van Unen, Martin Halamek	Final version	1.0

Acknowledgements

Authors: Roger Van Unen ATOS, Martin Halamek Dell EMC

Special thanks to Erwan Prevost, Greig Lilienfeld and Ryno Coetzee from Dell EMC, Arnaud Aumonier, Benoit Gonsolin and Virgil Chetty from ATOS BDS and Jamie Reding and Donovan White from Microsoft

Table of Contents

Executive summary

ATOS, Dell EMC™ and Microsoft®, in cooperation, provide guidelines and principles to assist customers in designing and implementing a balanced configuration for Microsoft SQL Server® data warehouse workloads to achieve out-of-the-box scalable performance. These certified database reference architectures enable each of the components in the database stack to provide optimal throughput to match the database capabilities of the specific setup.

ATOS BullSequana S 3th generation servers, along with robust and cutting-edge Dell EMC VMAX flash array, form efficient candidates for a high performing, large data warehouse solution. The in SQL Server 2017 integrated Python and R language modules makes it possible to perform, even in real time, ML operations without moving data from the data warehouse. The implementations of Python and R are based on the open source implementations giving access to thousands of solutions for ML and statistics.

This paper describes the design principles and guidelines used to achieve an optimally balanced 640TB Data Warehouse Fast Track (DWFT) reference architecture for SQL Server 2017 using BullSequana S800 servers and Dell EMC VMAX 250F flash array. The configuration used to achieve the performance numbers for the reference configuration is presented in detail.

Introduction

Today's enterprise businesses face a constant challenge keeping pace with the huge data processing and storage requirements generated by all aspects of their business. As they face the daunting task of scaling their DBMS systems to meet short term as well as long term requirements, one of the first decisions to be made is whether to scale up (add additional resources to existing systems), or scale out (add additional separate systems).

Up until now, the choices for scaling up systems for medium and large mission critical businesses has been very limited. Not only must the system be able to scale past four sockets, it must also support a storage system that can easily scale capacity and/or performance as CPU and memory resources are increased. The ATOS BullSequana S400 and S800 with the Dell EMC VMAX flash array sets a new pace for scalability and expandability while ensuring flexibility for all transaction, analytical, and data warehouse workloads.

The ATOS BullSequana S coupled with Dell EMC VMAX storage arrays makes an ideal scale up configuration for fast growing environments. With the ability to scale up to 8 CPU sockets with 12TB of memory and virtually unlimited storage capacity in a mission critical package, Microsoft® SQL Server 2017 can meet the need for all business sizes and requirements.

Excellent manageability and support ensures that, in the unlikely event of a failure or fault, Dell EMC proactive management systems can easily rectify (automatically in some cases) or mitigate issues before they become detrimental to mission critical system uptime.

Target audience

The target audience for this paper includes database administrators, business intelligence architects, storage administrators, IT directors, and data warehousing users seeking sizing and design guidance for Business Intelligence solutions with SQL Server 2017.

Data warehouse workload challenges

Organizations use data warehouses to aggregate data collected from operational systems and elsewhere and prepare data for analysis.

A traditional data warehouse workload consists of:
• Periodic data load from operational data stores/applications.
• Complex queries run by business analysts to get insight into the data for better decision making. Such queries aggregate large amounts of data across multiple tables, often running for long durations of time while consuming significant I/O bandwidth.

To speed up query performance, the data is pre-aggregated for the efficient execution of commonly occurring query patterns. New challenges face both designers and administrators managing mission-critical data warehouses.

Data growth

As the number of IoT devices increase, the data in data warehouses is growing exponentially. In this environment, it is important to use solutions that provide high data compression without compromising query performance, while reducing storage and I/O bandwidth. Integrated Machine Learning tools like Python and R makes it possible to run ML models on data without moving the data.

Both the ATOS BullSequana S400 and the S800 with the Dell EMC VMAX are designed to grow with your needs. In the beginning of 2019, the ATOS BullSequana will be able to scale up to 32 CPU's and the Dell EMC VMAX can scale up to 4 PB.

Reducing data latency

Data latency refers to the time required to access data for analytics in a data warehouse. Data load and transformation can be a resource-intensive operation that interferes with the ongoing analytics workload. To minimize the impact on business users, the extract, transform, and load (ETL) process typically takes place during off-peak hours.

In today's global economy, however, there are no off-peak hours. Businesses are striving to reduce data latency by making data available for analytics within minutes or seconds of its arrival in operational data stores. This requires loading incremental data into the data warehouse in real time or near real time.

Faster query response

Customers require most complex analytic queries to be able to return results in seconds—to enable interactive data exploration at the speed of thought. ATOS and Dell EMC developed a solution with a large amount of memory and flash storage to address this need.

SQL Server 2017

Microsoft SQL Server 2017 has made significant improvements in data warehousing technologies and performance, including column-store features as well as many other improvements.

Column-store indices offer great advantages over traditional row stores for analytics and data warehousing queries. They are ideally suited for the star schemas, and tables with billions of rows, which are commonly seen. Among their advantages for analytics are:

- Up to 10X compression in data size: Data warehouses are very large by nature and the compression offered by column-store index technologies offers both space and cost savings as well as significantly increased performance. These benefits are possible due to the dramatically reduced I/O requirements given by the compression and coupled by the ability to only scan the specific columns required by each query. This compression also reduces the amount of memory required to hold a given number of rows from the source data warehouse.

- Additional indices: SQL Server 2017 adds the capability to add (B-Tree) indices to column store-based tables, which enables efficient single-row lookup.

In addition to these architectural features, Microsoft SQL Server 2017 has made significant improvements in optimizing the processing of queries in column-store indices in the following ways:

- Operator pushdown: Pushdown refers to moving both filter and aggregation query operations closer to the data, so that many of the filters and calculations can be done in the scan operators, dramatically reducing the volume of data that needs to be handled further on in query processing.

- Batch-mode processing: SQL Server 2017 includes enhancements in batch-mode processing that handles many rows at a time rather than serially doing calculations on each individual row. These batch operations are further optimized by leveraging Single Instruction Multiple Data (SIMD) vector processing CPU instructions in the Intel® architectures

SQL Server Data Warehouse Fast Track

The SQL Server Data Warehouse Fast Track (DWFT) program is designed to provide customers with standard and proven system architectures optimized for a range of enterprise data warehousing needs. The goal is to help enterprise customers deploy data warehouse solutions with a recommended hardware configuration appropriate for the requirements of the workload with reduced risk, cost, and complexity.

Enterprises can purchase and build on reference implementations from participating system vendors or leverage the best practice guide provided through the program. The DWFT reference architecture program is continuously being improved to incorporate new SQL Server features and customer feedback.

When enterprises use the DWFT program to set up a data warehouse built on SQL Server, they lay the foundation for a complete Data Management Platform for Analytics. They can then take advantage of newer SQL Server features, including in-memory column store technologies that improve the performance of transactional and analytics workloads, as well as the ability of SQL Server to run on either Windows or Linux.

They also gain support for both traditional structured relational data and for unstructured big data, such as Internet of Things (IoT) data stored in Hadoop, Spark, or an Azure Data Lake, all the while being able to query the data in languages such as T-SQL, Java, C/C++, C#/VB.NET, PHP, Node.js, Python and Ruby. By using PolyBase, a feature in SQL Server optimized for data warehouse workloads, enterprise customers can also merge big data into the SQL Server universe. PolyBase provides the ability to query both relational data and unstructured data, joining it together into a single result set without moving the data.

This document defines DWFT component architecture and methodology. The result is a set of SQL Server database system architectures and configurations—including software and hardware—required to achieve and maintain a set of baseline performance levels out-of-box for a range of data warehousing workloads.

Recommended reference architectures

Single server reference architecture

The following subsections describe the two different DWFT reference architectures for SQL Server 2017, comprised of ATOS BullSequana S800 servers and Dell EMC VMAX flash arrays.

Figure 1 illustrates the single server reference architecture with the major elements and Table 1 lists the component details.

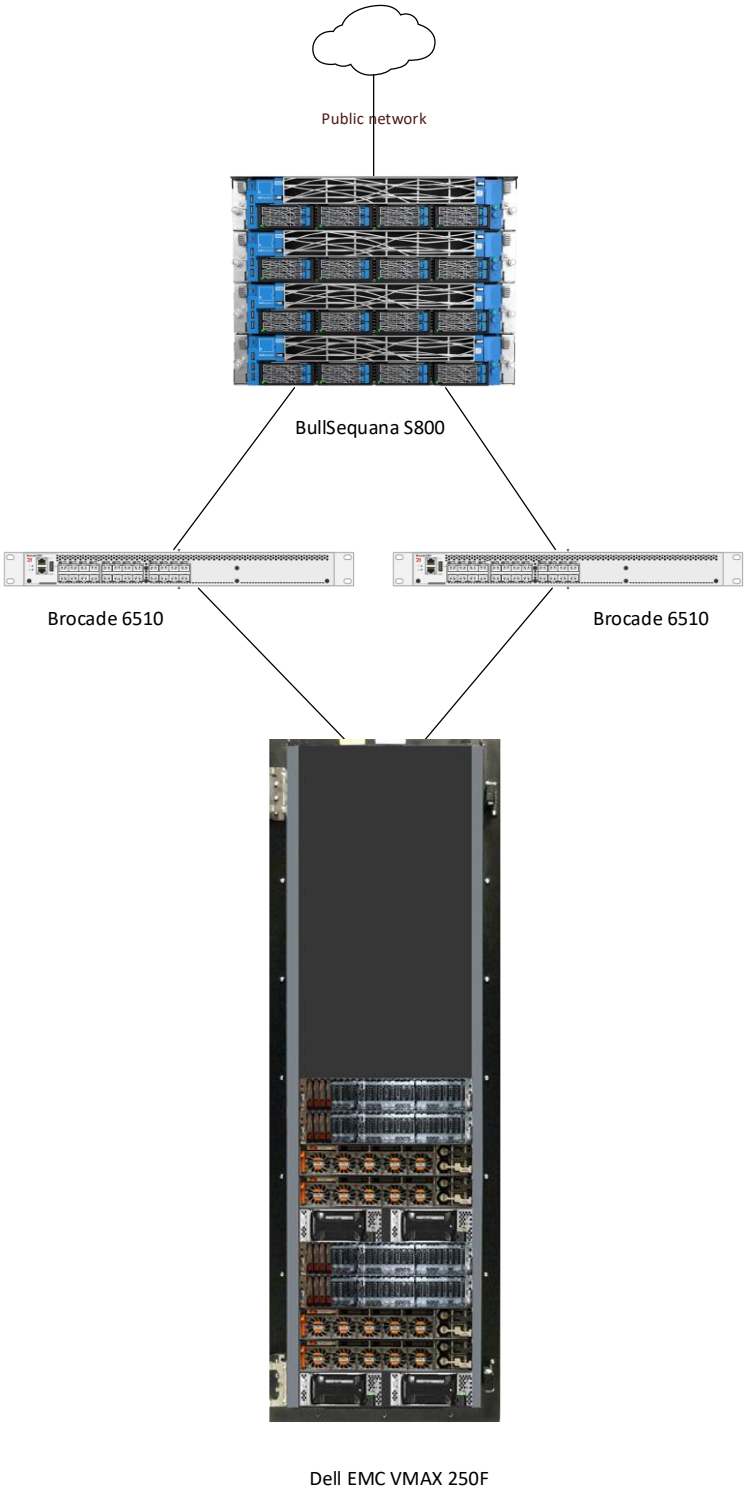

Figure 1: single server reference architecture

Table 1 Single server reference architecture

Component	Description	
Server	**BullSequana S800**	
	Processors	8x Intel® Xeon Platinum 8180M processors (2.5 MHz, 28 cores and 56 threads)
	Total cores	224
	Total Logical Processors	448
	Total RAM	12 terabytes in 128 GB DIMM's
	Host Bus Adapters	10x Emulex LPe31002-M6 adapter
	Network Adapters	8x Intel X722 10 Gbps + 8x Intel X722 SFP+
	OS disks	800 GB SSD RAID1
Software	Operating system	Windows Server 2016 Standard Edition
	Database software	SQL Server 2017 Enterprise Edition Core version
Storage	Array	2x Dell EMC VMAX 250FX VBRCK BASE 2048GB
	I/O cards	4x VMAX 250F 8MM 8 ports 16G FC
	Disk drives	64x VMAX 250 RAID5(7+1) 3840GB
	SAN switches	2x Brocade® 6510 with 16Gbps SFPs

High available reference architecture

For database high availability, Windows® failover clustering is recommended. Using Microsoft clustering services, one database server is configured as the primary (active) server and the second server is configured as the secondary (passive) server. The secondary server should have exactly the same configuration as the primary server. Since the database is only active on a single server at any point of time, the performance of the database on the primary server (active) is comparable to the single server configuration (discussed earlier).

Figure 2 illustrates the highly available reference architecture with the major elements and Table 2 lists the component details.

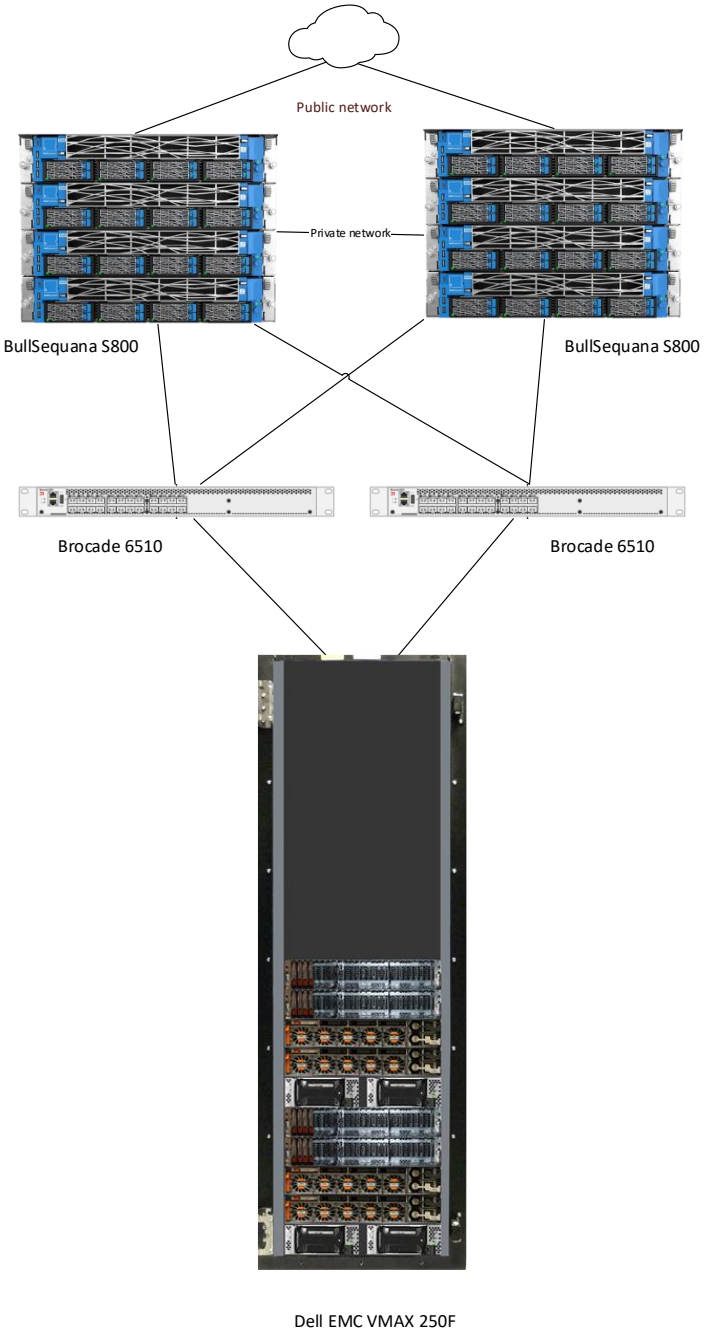

Figure 2: High available reference architecture

Table 2 High available reference architecture

Component	Description	
Server	**Two BullSequana S800 in active - passive**	
	Processors	8x Intel® Xeon Platinum 8180M processors (2.5 MHz, 28 cores and 56 threads)
	Total cores	224
	Total Logical Processors	448
	Total RAM	12 terabytes in 128 GB DIMM's
	Host Bus Adapters	10x Emulex LPe31002-M6 adapter
	Network Adapters	8x Intel X722 10 Gbps + 8x Intel X722 SFP+
	OS disks	800 GB SSD RAID1
Software	Operating system	Windows Server 2016 Standard Edition
	Database software	SQL Server 2017 Enterprise Edition Core version
Storage	Array	2x Dell EMC VMAX 250FX VBRCK BASE 2048GB
	I/O cards	4x VMAX 250F 8MM 8 ports 16G FC
	Disk drives	64x VMAX 250 RAID5(7+1) 3840GB
	SAN switches	2x Brocade® 6510 with 16Gbps SFPs

Hardware components

ATOS BullSequana S800 server

The ATOS BullSequana S800 server is a high versatile, eight-socket 8U rack server with impressive processor performance, a large memory footprint of maximum 12 terabytes, extensive I/O options and a choice of dense, high performance or low-cost, high-capacity storage, or NVMe storage or high performance NVidia Tesla GPU's.

The BullSequana S line is also available in a S200 version with 2 CPU's, a S400 with 4 CPU's (equally certified by Microsoft for SQL Server 2017 DWH Fast Track architecture), a S1600 with 16 CPU's and a S3200 with 32 CPU's.

Figure 3: BullSequana S line

For more information, see the https://atos.net/en/products/enterprise-servers/bullsequana-s product page.

Emulex LPe31002-M6 16 Gbps Host Bus Adapter

The Emulex Gen 6 (16/32G) Fibre Channel (FC) Host Bus Adapters (HBAs) by Broadcom are designed to address the demanding performance, reliability and management requirements of modern networked storage systems that utilize high performance and low latency solid state storage drives and hard disk drive arrays.

The Emulex Gen 6 FC HBAs with Dynamic Multi-core architecture offer higher performance, lower latency, enhanced diagnostics and manageability that benefit both 16GFC and 32GFC environments by applying all ASIC resources to any port that needs it. The Emulex Gen 6 HBAs delivers up to 12,800 MB/s (2 ports 32GFC, or 4 ports 16GFC, full duplex), less than half the latency, and support an industry-leading 1.6 million IOPS per adapter. The quad-port LPe32004 delivers up to 3.2 million IOPS per adapter.

For more information, see the Broadcom website:
https://www.broadcom.com/products/storage/fibre-channel-host-bus-adapters/lpe31002-m6

Brocade® 6510 switch

The Brocade 6510 switch by Broadcom is a 1U, 48-port, rack-mountable Fibre Channel switch providing up to 16Gbps of bandwidth per port. This switch enables organizations to simplify IT infrastructures, improve system performance, maximize the value of virtual server deployments, and reduce overall storage costs.

The Brocade 6510 Switch has all the capabilities you need to help improve network speed and efficiency.

- 16Gbps Fibre Channel performance — about 40 percent higher than 10GbE SANs
- Up to 48 ports that deliver 768Gbps aggregate full-duplex throughput
- Frame-based trunking at rates up to 128Gbps
- Exceptionally low power consumption (14 watts/Gbps)

For more information, see the Broadcom website: https://www.broadcom.com/products/fibre-channel-networking/switches/6510-switch

Dell EMC VMAX 250F

For enterprises that require petabyte-level scale, the VMAX All Flash is purpose-built to easily manage high-demand, heavy-transaction workloads while storing petabytes of vital data. The VMAX All Flash hardware design features the turbo-charged Dynamic Virtual Matrix Architecture that enables extreme speed and consistent sub-millisecond response time.

The VMAX All Flash architecture can scale beyond the confines of a single system footprint to deliver scalable performance where needed. It enables hundreds of multi-core Intel CPUs to be pooled and allocated on-demand to meet the performance requirements for dynamic mixed workloads. This is achieved though powerful multi-threading and the industry's first dynamic, user controlled core allocation so no workload is starved of resources. The core element of VMAX All Flash is the V-Brick. Each V-Brick has one engine, two DAEs, and usable capacity with fully redundant components.

Flash Capacity Packs are used to scale up to four petabytes. The VMAX All Flash scales by aggregating up to eight V-Bricks as a single system with fully shared connectivity, processing, and capacity resources. Each V-Brick supports up to 72 CPU cores for blazing-fast performance scaling to a maximum of 576 cores per array

Scalable Performance

• Leverage advanced multi-core/multi-threading algorithms and a flash-optimized design to meet strict SLAs for high-demand online transaction processing (OLTP), virtualized applications, and high-growth Oracle and SQL databases

• Scale out performance and scale up capacity to achieve millions of IOPS, PBs of capacity, and predictable performance (350-microsecond response time)

Mission-Critical Availability

• Mission-critical availability architecture with advanced fault isolation, robust data integrity checking and proven non-disruptive hardware and software upgrades

• Six-nines availability for 24x7 operations using SRDF software, the gold standard for multi-site remote replication and DR

Hyper Consolidation

• Achieve massive consolidation with support for mixed open, mainframe and file storage on the same system simplifying management and significantly lowering overall TCO

• Consolidate multiple concurrent workloads and multi-PBs of capacity both on premise and through tiering to cloud storage.

For more information, see the VMAX product page: https://www.dellemc.com/en-us/storage/vmax-all-flash.htm

Storage configuration

Hardware configuration of Dell EMC all flash array

Model: VMAX 250F (All Flash)
Configuration: 2 vBricks 2TB cache each
32 * 16Gbps host FC ports
64 * 3.84TB Flash Drive (Raid5-7+1), plus 2 hot spares

Front and Rear View of the system

The following pictures show the two VMAX vBricks in the rack:

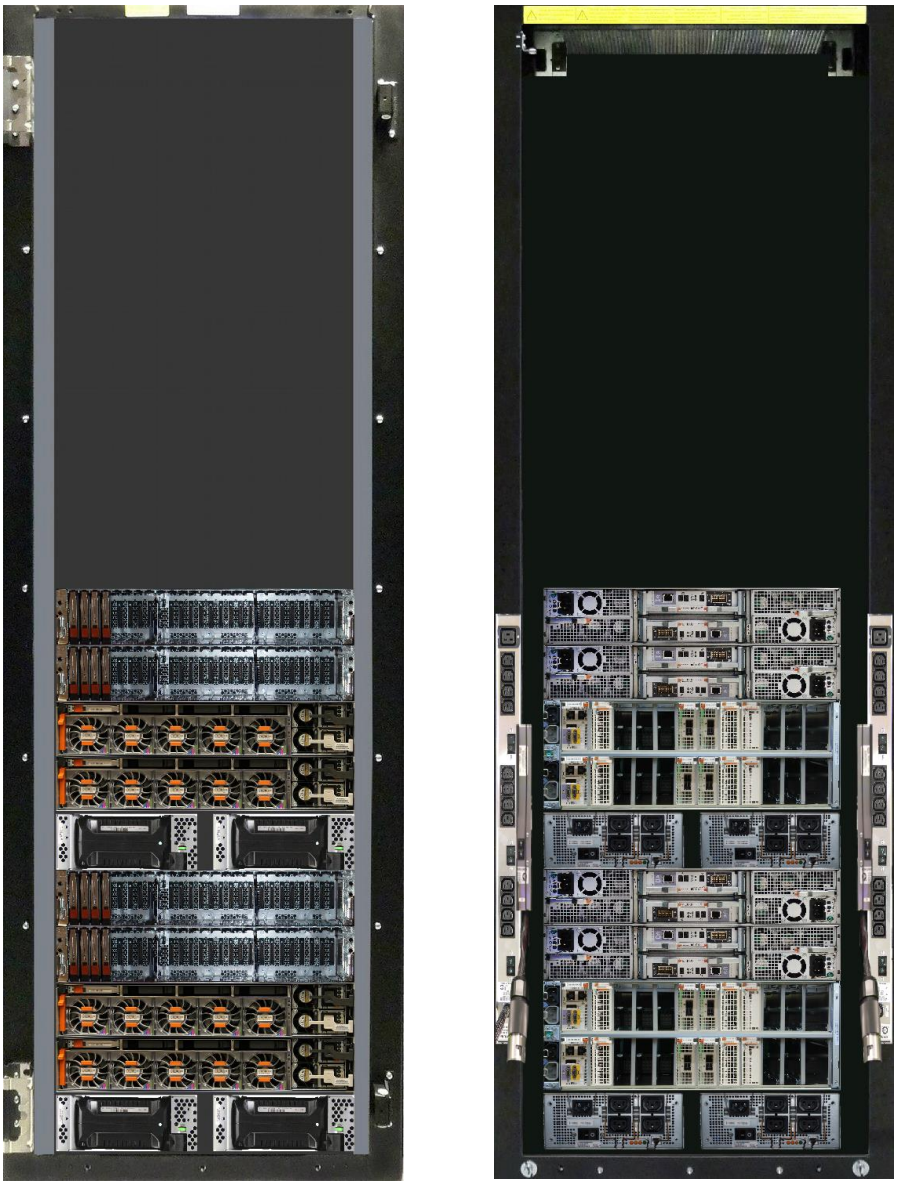

Figure 4: Front and rear system view

Engine Slots and FA Ports Layout

The following figure explains the director and the ports on the vBrick:

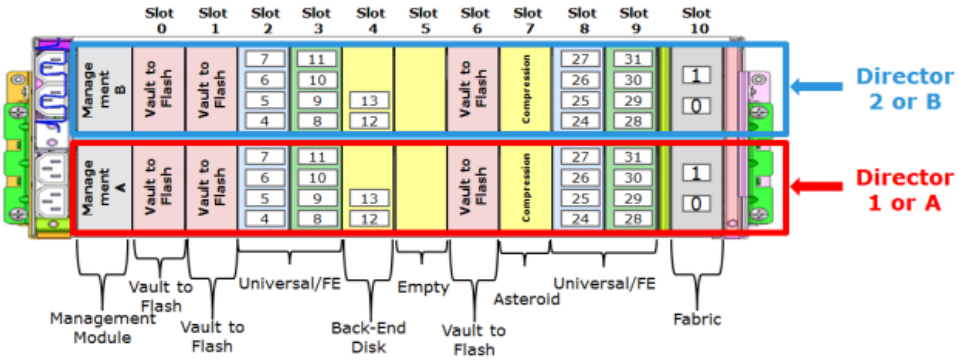

Figure 5: Engine Slots and FA ports Layout

Logical configuration of the array

The configuration is done following Dell EMC configuration and performance best practices.

Running version of HyperMaxOS: **5977.1131.1131**
Encryption: Enabled
Compression: Disabled
System Performance Profile: Baseline

Core distribution:

	ENG 1 Luna(2048GB)		ENG 2 Luna(2048GB)	
	1	2	3	4
H	NC	NC	NC	NC
G	NC	NC	NC	NC
F	NC	NC	NC	NC
E	NC	NC	NC	NC
D	FA	FA	FA	FA
C	DS	DS	DS	DS
B	EDS	EDS	EDS	EDS
A	IM	IM	IM	IM
brd	FB4	FB4	FB4	FB4

Table 3: Engine configuration

	ENG 1Luna(2048GB)		ENG 2Luna(2048GB)	
	1	2	3	4
FA	15	15	15	15
RF	0	0	0	0
EF	0	0	0	0
RE	0	0	0	0
SE	0	0	0	0
FE	0	0	0	0
DS	16	16	16	16
DX	0	0	0	0
IM	4	4	4	4
EDS	13	13	13	13
UnAlloc	0	0	0	0
Total	48	48	48	48

Table 4: Core distribution to emulations

LUN / Disk configuration and FA port assignments

All LUNs were configured as THIN LUNs in virtual pool created from 64 flash drives. Data are wide striped across all flash drives across both vBricks.

In total, there are 35 LUN's defined on the Dell EMC VMAX:

- 16 LUN's of 8 TB for the user data (128 TB usable in total)
- 16 LUN's or 2 TB for the tempdb (32 TB usable in total)
- 2 LUN's of 2 TB for the log (4 TB usable in total)

The front-end Fibre Channel (FC) ports were configured to use four fault domains in virtual port mode. Every port is one to one connected to a port on the VMAX that serves only one data LUN and one Tempdb LUN. There is a LUN for each to separate reading and writing actions on a LUN (when performing a GROUP BY or ORDER BY the data are first written in Tempdb before being used in another part of a query or present the results to the user). Each LUN is mapped to four FA ports and accessible via four paths in total. FA Port Mapping as below table.

Array	Device	Capacity (GB)	Storage Group	Notes	FA Port Mapping
250F	00024	8192.00	SG_1_DATA	DATA	FA-01D:4, FA-02D:5, FA-03D:4, FA-04D:5
250F	00025	8192.00	SG_1_DATA	DATA	FA-01D:4, FA-02D:5, FA-03D:4, FA-04D:5
250F	00026	8192.00	SG_1_DATA	DATA	FA-01D:4, FA-02D:5, FA-03D:4, FA-04D:5
250F	00027	8192.00	SG_1_DATA	DATA	FA-01D:4, FA-02D:5, FA-03D:4, FA-04D:5
250F	00028	8192.00	SG_2_DATA	DATA	FA-01D:5, FA-02D:4, FA-03D:5, FA-04D:4
250F	00029	8192.00	SG_2_DATA	DATA	FA-01D:5, FA-02D:4, FA-03D:5, FA-04D:4
250F	0002A	8192.00	SG_2_DATA	DATA	FA-01D:5, FA-02D:4, FA-03D:5, FA-04D:4
250F	0002B	8192.00	SG_2_DATA	DATA	FA-01D:5, FA-02D:4, FA-03D:5, FA-04D:4
250F	0002C	8192.00	SG_3_DATA	DATA	FA-01D:8, FA-02D:9, FA-03D:8, FA-04D:9
250F	0002D	8192.00	SG_3_DATA	DATA	FA-01D:8, FA-02D:9, FA-03D:8, FA-04D:9
250F	0002E	8192.00	SG_3_DATA	DATA	FA-01D:8, FA-02D:9, FA-03D:8, FA-04D:9
250F	0002F	8192.00	SG_3_DATA	DATA	FA-01D:8, FA-02D:9, FA-03D:8, FA-04D:9
250F	00030	8192.00	SG_4_DATA	DATA	FA-01D:9, FA-02D:8, FA-03D:9, FA-04D:8
250F	00031	8192.00	SG_4_DATA	DATA	FA-01D:9, FA-02D:8, FA-03D:9, FA-04D:8
250F	00032	8192.00	SG_4_DATA	DATA	FA-01D:9, FA-02D:8, FA-03D:9, FA-04D:8
250F	00033	8192.00	SG_4_DATA	DATA	FA-01D:9, FA-02D:8, FA-03D:9, FA-04D:8
250F	00034	2048.00	SG_1_TempDB	TemDB	FA-01D:4, FA-02D:5, FA-03D:4, FA-04D:5
250F	00035	2048.00	SG_1_TempDB	TemDB	FA-01D:4, FA-02D:5, FA-03D:4, FA-04D:5
250F	00036	2048.00	SG_1_TempDB	TemDB	FA-01D:4, FA-02D:5, FA-03D:4, FA-04D:5
250F	00037	2048.00	SG_1_TempDB	TemDB	FA-01D:4, FA-02D:5, FA-03D:4, FA-04D:5
250F	00038	2048.00	SG_2_TempDB	TemDB	FA-01D:5, FA-02D:4, FA-03D:5, FA-04D:4
250F	00039	2048.00	SG_2_TempDB	TemDB	FA-01D:5, FA-02D:4, FA-03D:5, FA-04D:4
250F	0003A	2048.00	SG_2_TempDB	TemDB	FA-01D:5, FA-02D:4, FA-03D:5, FA-04D:4
250F	0003B	2048.00	SG_2_TempDB	TemDB	FA-01D:5, FA-02D:4, FA-03D:5, FA-04D:4
250F	0003C	2048.00	SG_3_TempDB	TemDB	FA-01D:8, FA-02D:9, FA-03D:8, FA-04D:9
250F	0003D	2048.00	SG_3_TempDB	TemDB	FA-01D:8, FA-02D:9, FA-03D:8, FA-04D:9
250F	0003E	2048.00	SG_3_TempDB	TemDB	FA-01D:8, FA-02D:9, FA-03D:8, FA-04D:9
250F	0003F	2048.00	SG_3_TempDB	TemDB	FA-01D:8, FA-02D:9, FA-03D:8, FA-04D:9
250F	00040	2048.00	SG_4_TempDB	TemDB	FA-01D:9, FA-02D:8, FA-03D:9, FA-04D:8
250F	00041	2048.00	SG_4_TempDB	TemDB	FA-01D:9, FA-02D:8, FA-03D:9, FA-04D:8
250F	00042	2048.00	SG_4_TempDB	TemDB	FA-01D:9, FA-02D:8, FA-03D:9, FA-04D:8
250F	00043	2048.00	SG_4_TempDB	TemDB	FA-01D:9, FA-02D:8, FA-03D:9, FA-04D:8
250F	00044	2048.00	SG_5_LOG	LOG	FA-01D:6, FA-02D:7, FA-03D:6, FA-04D:7
250F	00045	2048.00	SG_5_LOG	LOG	FA-01D:6, FA-02D:7, FA-03D:6, FA-04D:7

Table 5: LUN mapping

VMAX storage groups (SG)

The following table defines the created storage groups. All storage group were configured with Diamond SLO (< 1ms).

Storage Group	Type	Device Count	SLO
SG_1_DATA	SG	4	Diamond
SG_1_TempDB	SG	4	Diamond
SG_2_DATA	SG	4	Diamond
SG_2_TempDB	SG	4	Diamond
SG_3_DATA	SG	4	Diamond
SG_3_TempDB	SG	4	Diamond
SG_4_DATA	SG	4	Diamond
SG_4_TempDB	SG	4	Diamond
SG_5_LOG	SG	2	Diamond

Table 6: storage groups

Read and write cache

VMAX cache is global. It is dynamically used for reads and writes. The default settings were used for system write pending limit. No cache partitions were created as there is only single application using system.

Server connectivity and multi-pathing

Connected Windows Server 2016 was using native MPIO. Alternative is to use Dell EMC multi-pathing software PowerPath.

Cabling

The hardware components where connected using the ATOS and Dell EMC best practices.

The upper ports of the Brocade switches are used to connect the ATOS BullSequana S800 server. Each even port of the Emulex HBA's was connected to the upper ports on the left of Brocade switch one and each off port of the Emulex HBA's was connected to the upper ports on the left of Brocade switch two. The bottom ports are only used to connect the FC ports of the Dell EMC VMAX.

Cabling single server configuration

The following diagram explains the cabling between the BullSequana S800 and the Dell EMC VMAX 250F. The placement of the HBA's can be different based on the placement of the HBA modules.

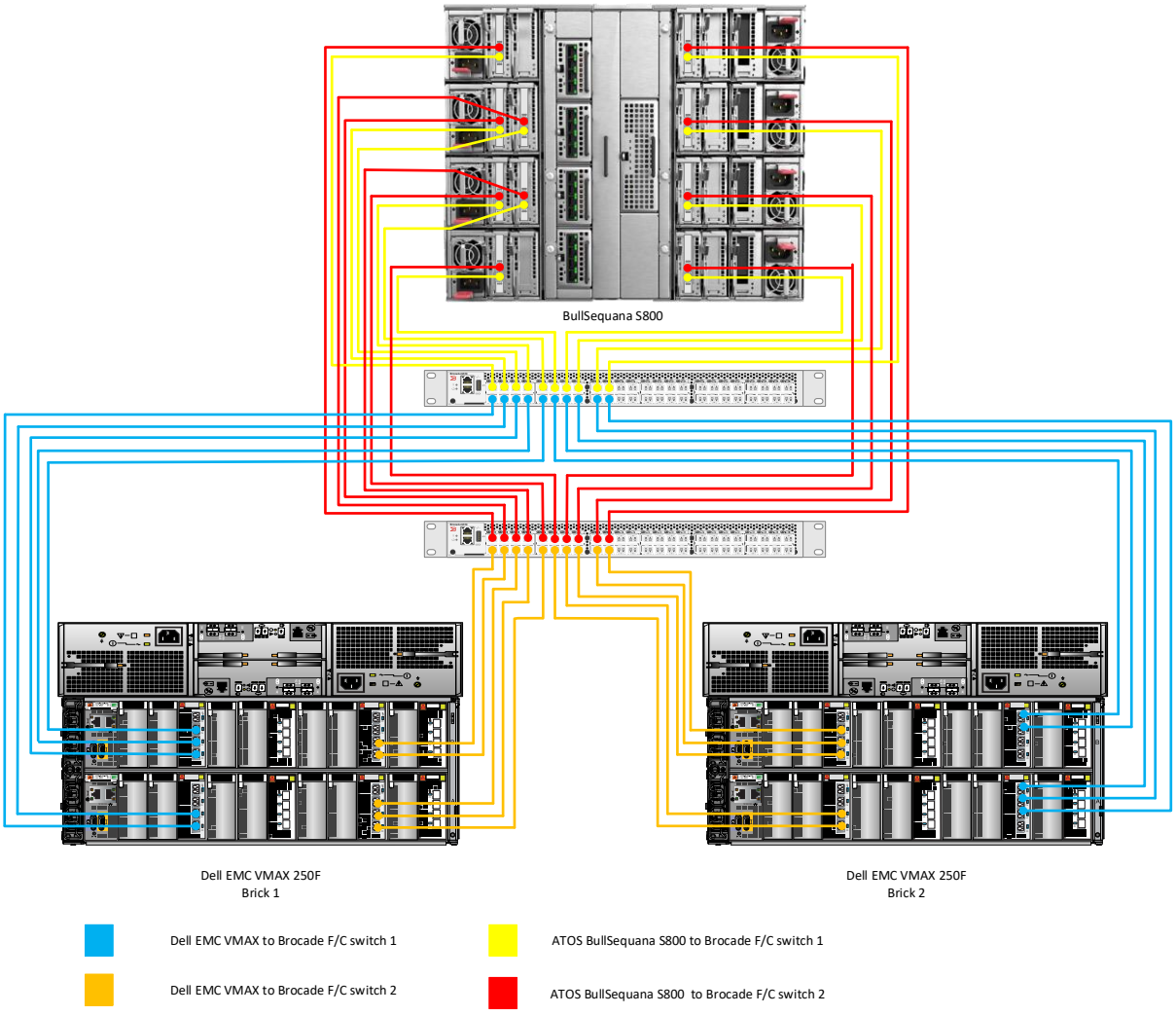

Figure 6: Single server configuration cabling diagram

Cabling high available server configuration

The following diagram explains the cabling between two BullSequana S800 servers in a high availability configuration and the Dell EMC VMAX 250F. The placement of the HBA's can be different based on the placement of the HBA modules.

To keep the diagram with all the connections readable the Brocade switches are presented twice: first only the left part and then the right part.

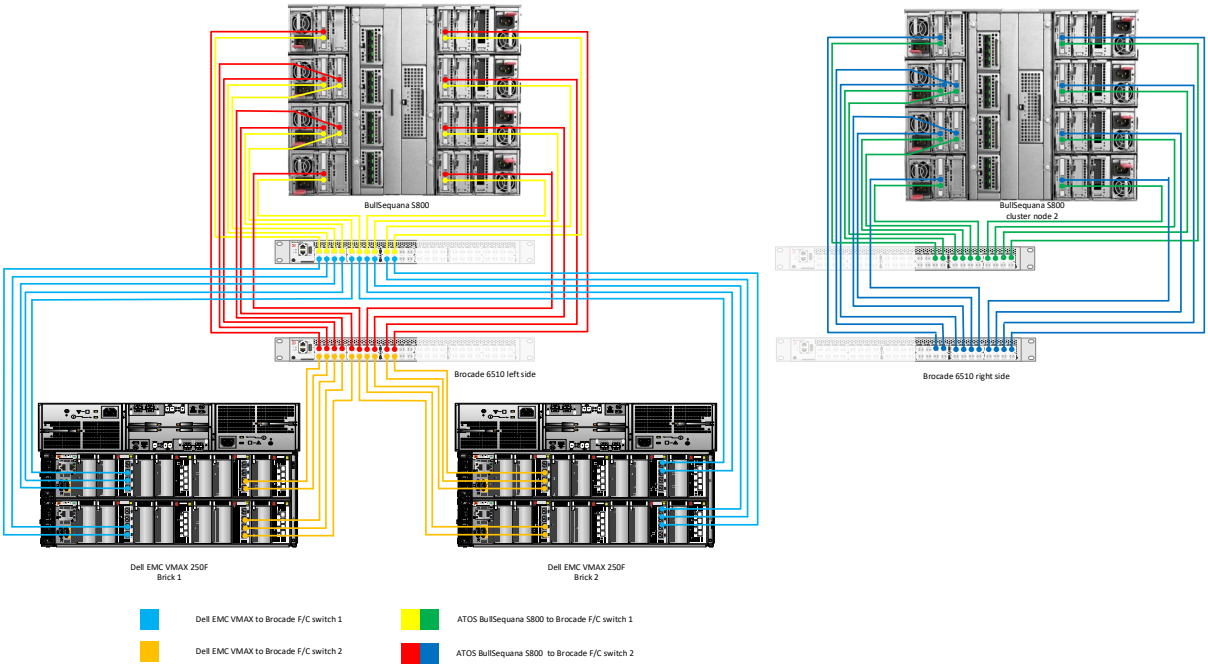

Figure 7: high available server configuration cabling diagram

BullSequana S800 server configuration

System BIOS

All options except for the BMC network configuration were left at their factory default settings. The Logical Processor option, under Processor Settings, is left at its default setting of Enabled. This enables Intel® Hyper-Threading Technology, which maximizes the number of logical processors available to SQL Server.

Emulex LPe31002-M6 16 Gbps Host Bus Adapter

All Emulex HBA's were detected by the Emulex OneCommand™ Manager installation. The firmware and driver version used during the Microsoft DWH Fast Track certification is 12.0.193.13:

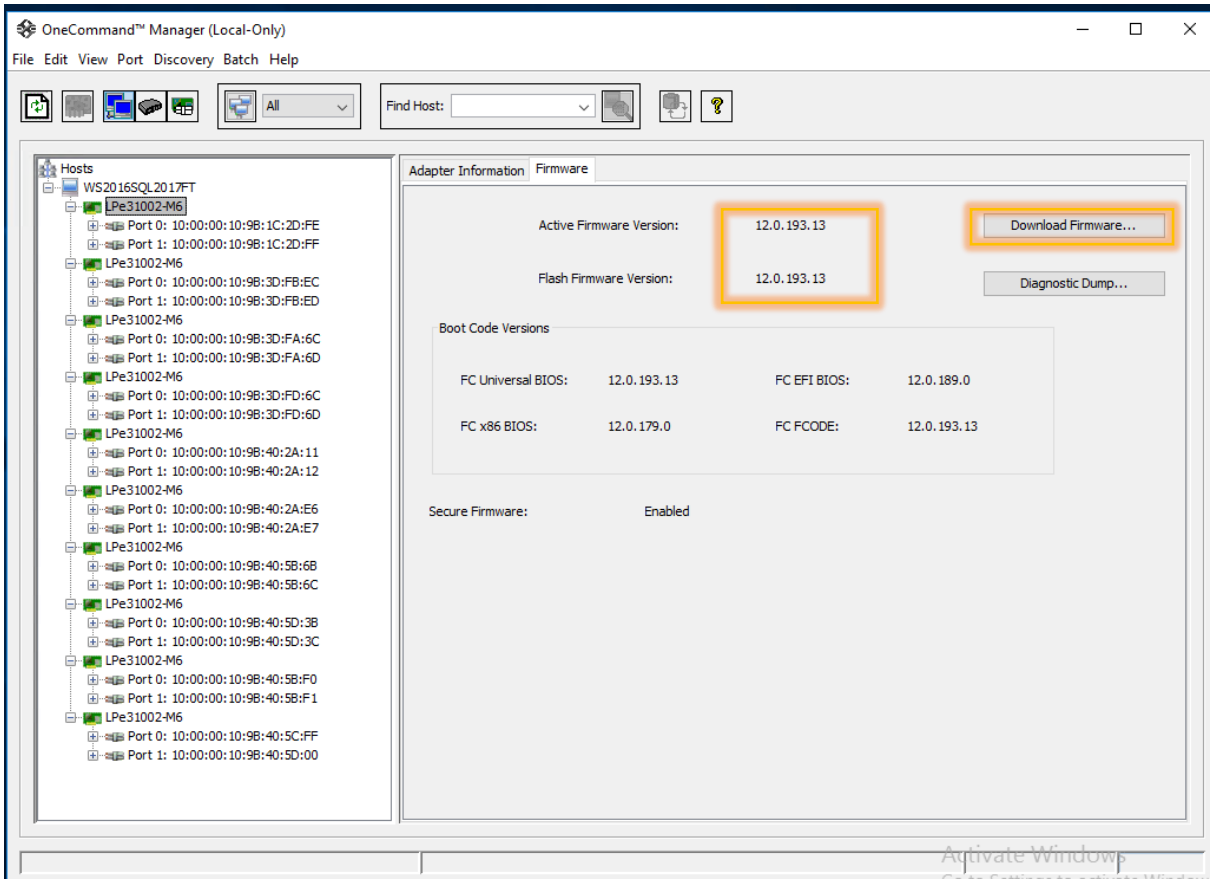

Figure 8: Emulex host bus adapter firmware version

It is important that all the HBA have the same firmware.

If this is not the case the HBA's firmware need to be upgraded to this version or the latest available for each HBA. The latest firmware can be downloaded from the Broadcom Emulex site (https://www.broadcom.com/products/storage/fibre-channel-host-bus-adapters/lpe31002-m6).

The update of the firmware can be performed using the "Download Firmware" button in the Emulex OneCommand™ Manager. Changing the driver and or the firmware could require a restart of the server.

Except the QueryDept parameter – which was changed from 32 to **64** – all the other host and HBA parameters were left at their default settings.

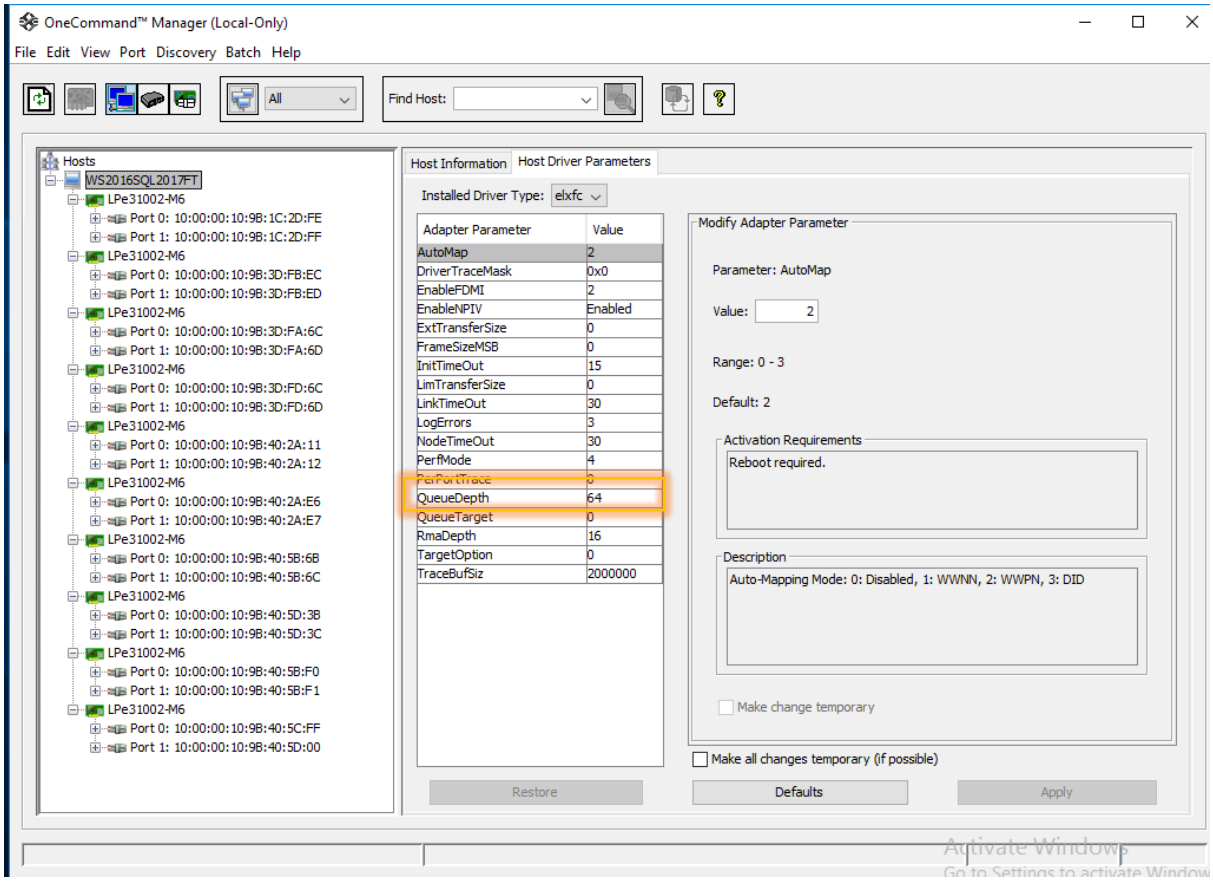

Figure 9: Emulex Host Driver Parameters

Windows Server 2016 configuration

Installation

The installation of Windows was done with the default settings. After the installation, the Windows Feature MPIO was activated as shown in figure 8.

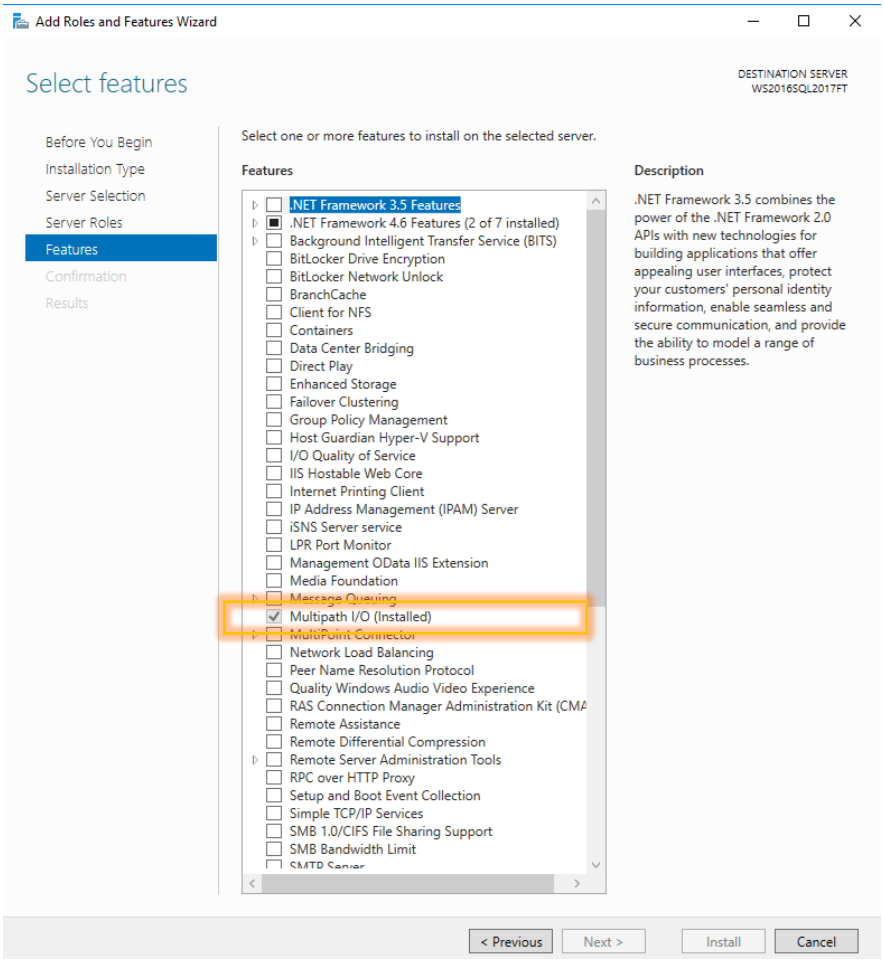

Figure 10: MPIO feature installed

After the installation of MPIO feature Windows needs to be restarted.

Drivers and packages installed

It is important to have the drivers and packages on a USB key because the network adapters are newer than Windows 2016 and the drivers are not part of the Windows Drivers Catalogue included on the installation DVD.

After the installation of Windows and the MPIO, the following drivers and packages were installed (in that order):

1. Intel chipset INF version 10.1.17711.8088_PV or later (restart required): https://downloadcenter.intel.com/download/28153/Intel-Server-Chipset-Driver-for-Windows-
2. MegaRAID Windows 2016 driver version 6.14-6.714.05.00-WHQL (restart required): https://www.broadcom.com/products/storage/raid-controllers/megaraid-sas-9361-8i
3. PROWinx64 for Intel XL7xx family version 23_2 or later: https://downloadcenter.intel.com/download/26092/Intel-Network-Adapter-Driver-for-Windows-Server-2016-?product=75021
4. OneInstall-Setep-12.0.193.18.exe or later: https://www.broadcom.com/products/storage/fibre-channel-host-bus-adapters/lpe31002-m6

After the installation of these drivers and packages, a restart of Windows might be needed.

Power plan

To maximize performance, the server was configured to use the High performance power plan as shown in Figure 9.

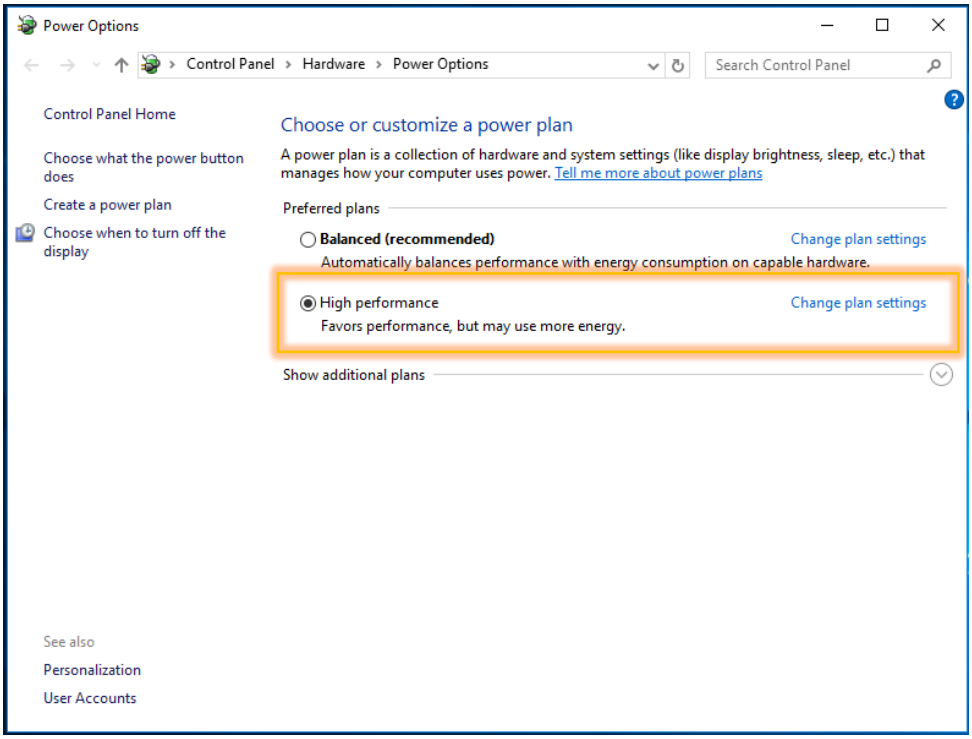

Figure 11: Windows Power Options settings

Lock pages in memory

To prevent Windows from paging SQL Server memory to disk, the Lock pages in memory option was enabled for the SQL Server service account. Remember to restart the SQL Server instance to for this setting to take effect.

For information on enabling this option, visit Enable the Lock Pages in Memory Option

Windows disks

After the zoning in the switches and the VMAX are done the LUN's show up in Windows Disk Management. After putting the disks on-line and the initialisation, all LUN's are formatted with ReFS and a block size of 64 KB:

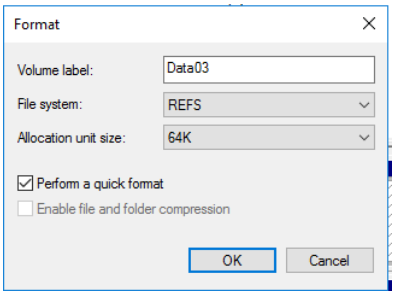

Figure 12: format disk options

If all the disks are formatted properly, Windows Disk Management shows the following list:

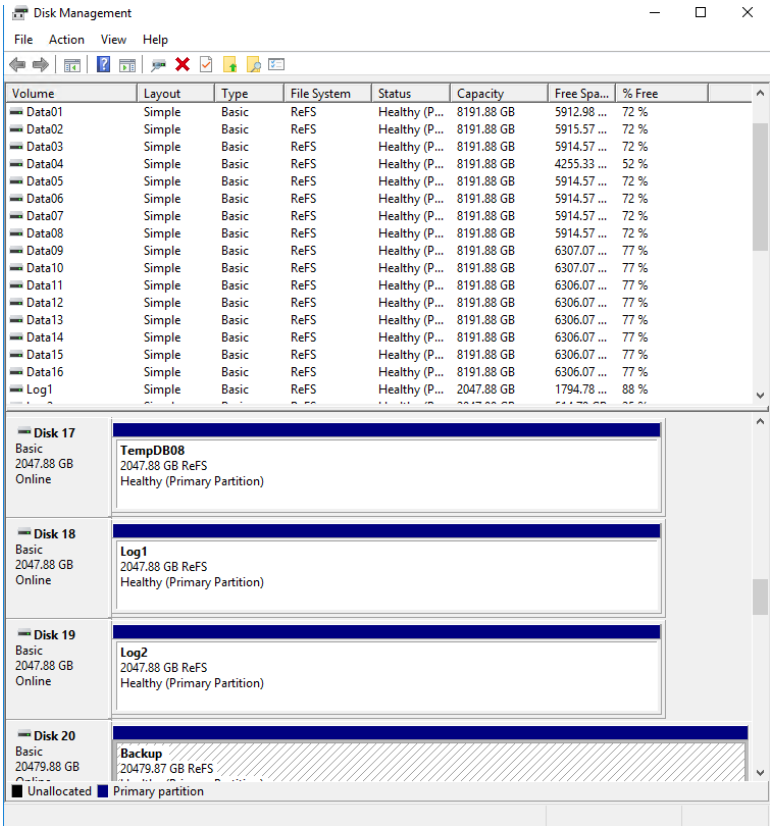

Figure 13: Disk Management

There are in total, there are 34 LUN's:

- 16 LUN's of 8 TB for the user data (128 TB raw in total)
- 16 LUN's or 2 TB for the tempdb (32 TB raw in total)
- 2 LUN's of 2 TB for the log (4 TB raw in total)

Performing a quick format of an eight terabyte LUN might take some time.

For DWFT reference architectures, ATOS and Dell EMC recommends using mount points for the volumes instead of drive letters. It is highly recommended to assign appropriate volume and mount

point names in order to simplify troubleshooting and performance analysis. Ideally, the mount point names should be assigned in a way that makes it easy to identify the VMAX volume for a given Windows volume.

The following table shows the volume labels and access paths used for the reference configuration:

VMAX LUN Alias	Windows LUN Label	Access path
Data01	Data01	C:\Storage\Data01
Data02	Data02	C:\Storage\Data02
Data03	Data03	C:\Storage\Data03
Data04	Data04	C:\Storage\Data04
Data05	Data05	C:\Storage\Data05
Data06	Data06	C:\Storage\Data06
Data07	Data07	C:\Storage\Data07
Data08	Data08	C:\Storage\Data08
Data09	Data09	C:\Storage\Data09
Data10	Data10	C:\Storage\Data10
Data11	Data11	C:\Storage\Data11
Data12	Data12	C:\Storage\Data12
Data13	Data13	C:\Storage\Data13
Data14	Data14	C:\Storage\Data14
Data15	Data15	C:\Storage\Data15
Data16	Data16	C:\Storage\Data16
Tempdb01	Tempdb01	C:\Storage\Tempdb01
Tempdb02	Tempdb02	C:\Storage\Tempdb02
Tempdb03	Tempdb03	C:\Storage\Tempdb03
Tempdb04	Tempdb04	C:\Storage\Tempdb04
Tempdb05	Tempdb05	C:\Storage\Tempdb05
Tempdb06	Tempdb06	C:\Storage\Tempdb06
Tempdb07	Tempdb07	C:\Storage\Tempdb07
Tempdb08	Tempdb08	C:\Storage\Tempdb08
Tempdb09	Tempdb09	C:\Storage\Tempdb09
Tempdb10	Tempdb10	C:\Storage\Tempdb10
Tempdb11	Tempdb11	C:\Storage\Tempdb11
Tempdb12	Tempdb12	C:\Storage\Tempdb12
Tempdb13	Tempdb13	C:\Storage\Tempdb13
Tempdb14	Tempdb14	C:\Storage\Tempdb14
Tempdb15	Tempdb15	C:\Storage\Tempdb15
Tempdb16	Tempdb16	C:\Storage\Tempdb16
Log1	Log1	C:\Storage\Log1
Log2	Log2	C:\Storage\Log2

After completing of all the LUN's the directory C:\Storage looks like this:

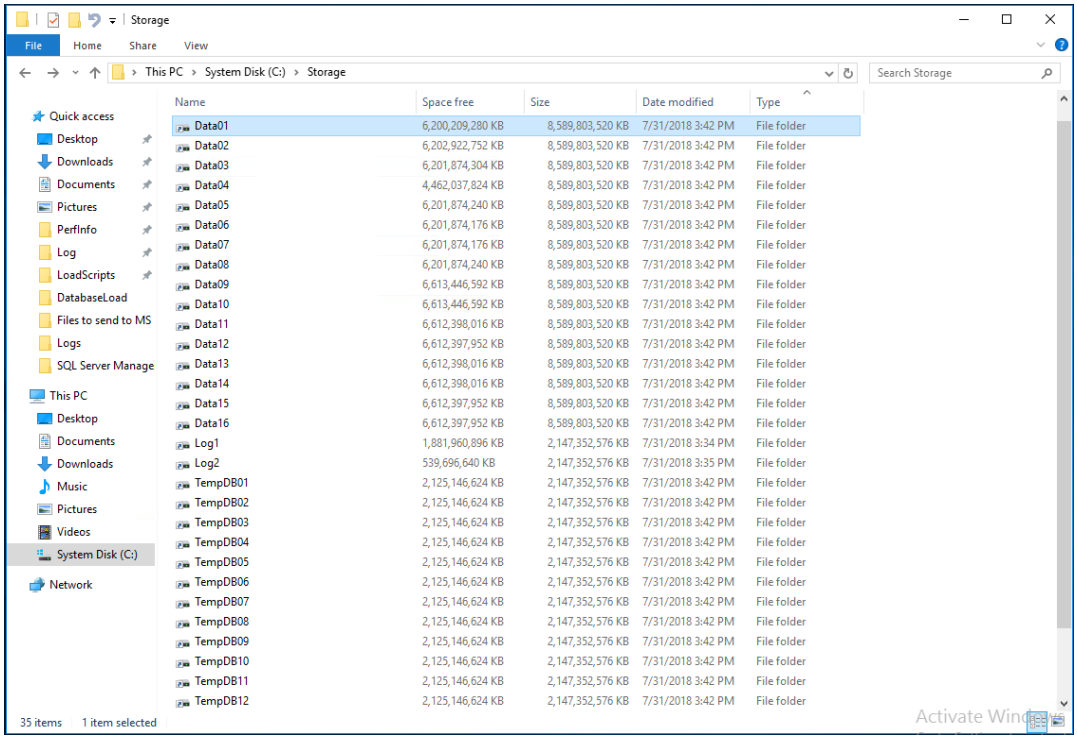

Figure 14: LUN mount points in the C:\Storage folder

MPIO

MPIO was configured using Dell EMC best practices. MPIO best practices for the VMAX array are documented in the paper, Dell EMC Host Connectivity Guide for Windows.

The MPIO policy for all volumes is set at **Least Queue Dept**, allowing a load balancing policy that sends I/O down the path with the fewest currently outstanding I/O requests.

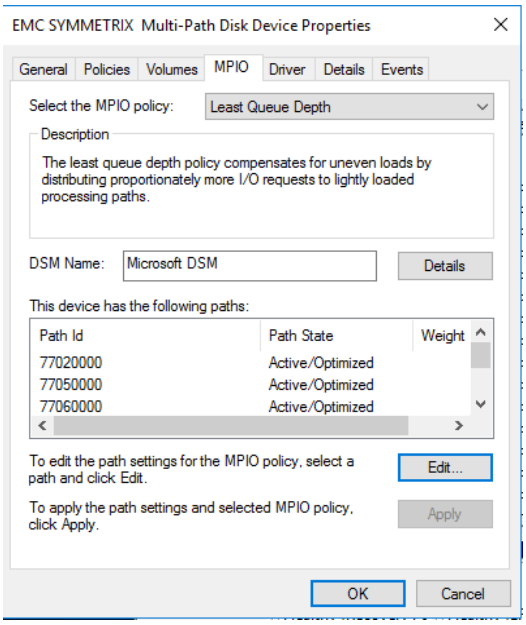

Figure 15: MPIO configuration LUN

Windows Defender configuration

Microsoft Windows Defender is a standard installed anti-virus and anti-spam component. To prevent the scanner to scan the SQL Server data and log files the following exclusions must be added to guarantee performance:

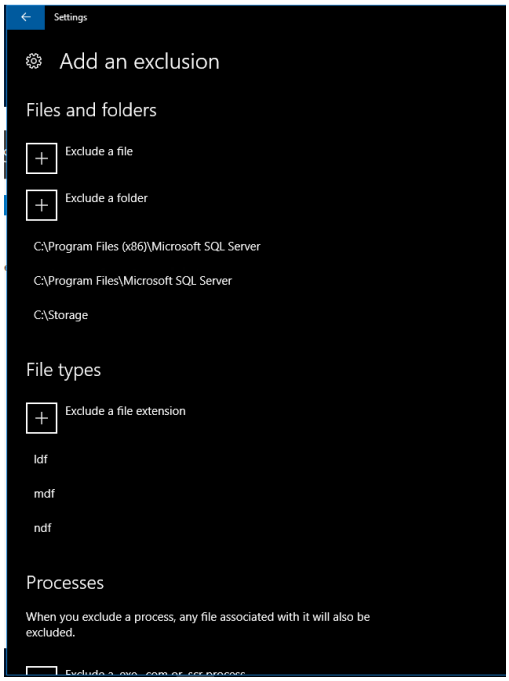

Figure 16: Windows Defender configuration single server

Exclude the following folders in a single server solution:

- C:\Program Files (x86)\Microsoft SQL Server
- C:\Program Files\Microsoft SQL Server
- C:\Storage (single server solution)

Exclude the following folders in a high availability (HA) solution (not displayed):

- C:\Program Files (x86)\Microsoft SQL Server
- C:\Program Files\Microsoft SQL Server
- C:\ClusterStorage (High Available solution)
- C:\Windows\Cluster
- Quorum drive
- MSDTC drive

Exclude the following file types:

- .ldf
- .mdf
- .ndf

These settings are also valid for other virus scanners. More information on configuring virus scanners with SQL Server is available at https://support.microsoft.com/en-us/help/309422/choosing-antivirus-software-for-computers-that-run-sql-server

SQL Server 2017 Enterprise Edition Configuration

The installation was done using mostly of the default settings. Some exceptions are listed below.

Grant perform volume maintenance task privilege

During installation of SQL Server 2017, the option to grant the SQL Server Database Engine Service the **Perform Volume Maintenance Task** privilege was selected.

Tempdb configuration

The tempdb database was configured to use sixteen data files of equal size. The data files were placed on the sixteen tempdb data volumes. The tempdb transaction log file was placed on the log2 volume. All files were expanded to the appropriate size and auto grow was enabled.

Start-up parameters for the SQL Server instance

SQL Server 2017 automatically sets the –T1117 and –T1118 trace flags so it is no longer necessary to add them to the start-up options of the instance. We have seen better performance by not using the –T834 trace flag. Microsoft recommends not to use this flag when using Clustered Column Store indexes.

The only start-up option set is the –E flag:

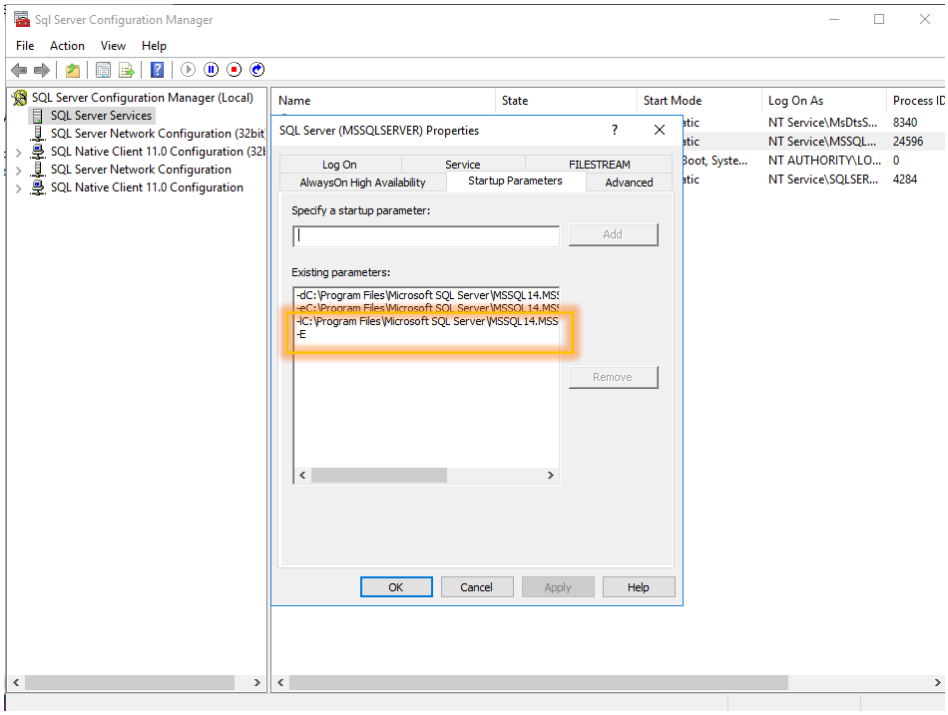

Figure 17: start-up options of the SQL Server instance

More information in the Microsoft article, DBCC TRACEON - Trace Flags and Database Engine Service Startup Options.

SQL Server maximum memory

The maximum server memory for this reference architecture should be set to 11.534.336 MB or 11.264 GB, which leaves 1.024 GB for the operating system. If additional applications share the server, adjust the amount of memory left available to the operating system accordingly.

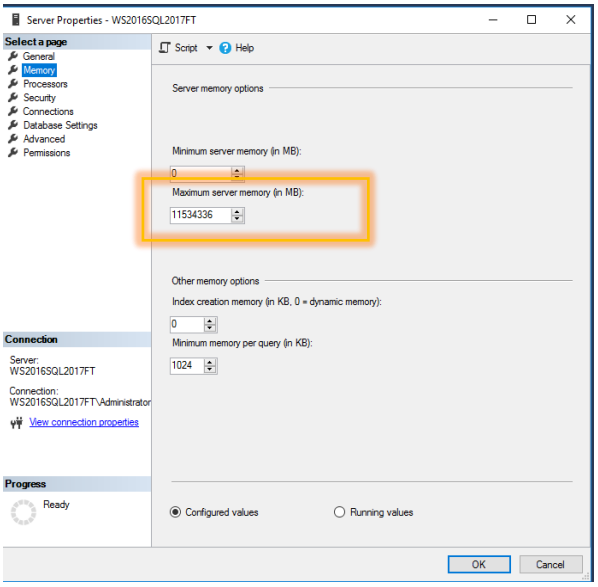

Figure 18: memory configuration of SQL Server

Maximum Degree of parallelism (MAXDOP)

The max degree of parallelism was set to 448 which correspond to the number of logical cores available in the server.

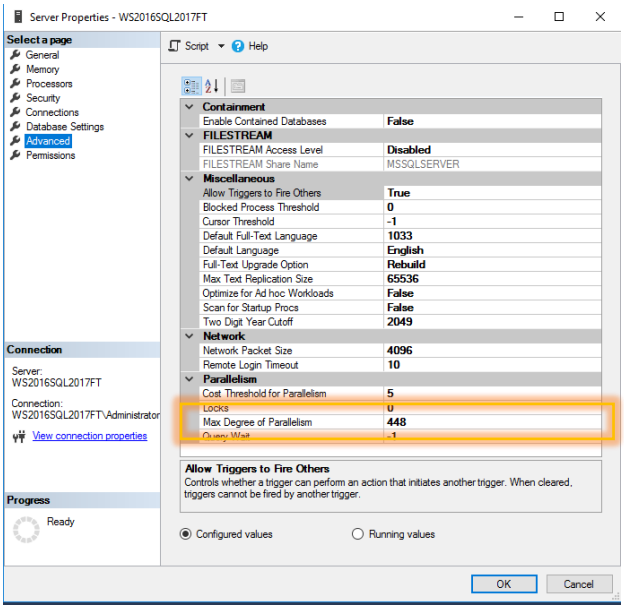

Figure 19: maximum degree of parallelism (MAXDOP)

For more information on the MAXDOP see the Microsoft article, [Configure the max degree of parallelism Server Configuration Option](#).

Resource governor

Depending on the type of workload used the most – row store (RS) or clustered column store (CS) – the settings of the Resource Governor needs to be adapted accordingly.

For the **row store** the resource governor setting used to limit the maximum memory grant is set to **12 percent**.

```
ALTER WORKLOAD GROUP [default] WITH (request_max_memory_grant_percent=12);
ALTER RESOURCE GOVERNOR RECONFIGURE;
```

For the **column store** the resource governor setting used to limit the maximum memory grant is set to **25 percent**.

```
ALTER WORKLOAD GROUP [default] WITH (request_max_memory_grant_percent=25);
ALTER RESOURCE GOVERNOR RECONFIGURE;
```

For information about the resource governor, visit the Microsoft page Resource Governor.

Database configuration

The data warehouse database was configured to use multiple file groups, each containing sixteen files distributed evenly across the sixteen data volumes. All files were allowed to grow automatically. The file groups were configured with the **AUTOGROW_ALL_FILES** option to help ensure that all files within a given file group remain the same size.

```
ALTER DATABASE <database name>
    MODIFY FILEGROUP <file group name> AUTOGROW_ALL_FILES;
```

Additional considerations for the Highly Available (HA) reference architecture

The HA reference architecture leverages Windows Failover Clustering to achieve high availability. When configuring a Windows failover cluster, there are additional storage considerations: the recommended quorum configuration is to allow all cluster nodes to have quorum votes and use a disk witness.

An additional volume needs to be created and configured as the disk witness. Dell EMC recommends using a 2GB volume for the disk witness. For more information on quorum and voting configurations in a failover cluster, see the Microsoft article, Configure and Manage a Quorum.

All volumes need to be mapped to each node of the cluster. All volumes need to be configured as a cluster resource and added to the SQL Server cluster resource group.

DWFT certification for ATOS BullSequana S800 with Dell EMC VMAX 250F

DWFT Certification #2017-18 DWFT Rev. 5.4	BullSequana S800 with Dell EMC VMAX 250F DWFT Reference Architecture		Report Date: 8/22/2018
System Provider	**System Name**	**Processor Type**	**Memory**
Atos	BullSequana S800	Intel Xeon Platinuml 8180M 2.5 GHz (8/224/448)	12 TB
Operating System		**SQL Server Edition**	
Windows Server 2016		SQL Server 2017 Enterprise Edition	
Storage Provider	**Storage Information**		
DELL EMC	64x 3.84 TB SSD for data, tempdb and log RAID 7 + 1 2x 800 GB SSD for OS (RAID 1)		

Primary Metrics			
Rated User Data Capacity[1] (TB)	Row Store Relative Throughput[2]	Column Store Relative Throughput[3]	Maximum User Data Capacity[1] (TB)
640	1,501	2,489	720

Row Store					
Relative Throughput[2]	Measured Throughput (Queries/Hr/TB)	Measured Scan Rate Physical (MB/Sec)	Measured Scan Rate Logical (MB/Sec)	Measured I/O Throughput (MB/Sec)	Measured CPU (Avg.) (%)
1,501	1,613	41,873	49,953	45,912	78

Column Store					
Relative Throughput[2]	Measured Throughput (Queries/Hr/TB)	Measured Scan Rate Physical (MB/Sec)	Measured Scan Rate Logical (MB/Sec)	Measured I/O Throughput (MB/Sec)	Measured CPU (Avg.) (%)
2,489	16,168	10,587	N/A	N/A	68

The reference configuration is a 2 socket system rated for 25TB using SQL Server 2014 and the DWFT V4 methodology

[1] Assumes a data compression ratio of 5:1

[2] Percent ratio of the throughput to the row store throughput of the reference configuration.

[3] Percent ratio of the throughput to the column store throughput of the reference configuration.

* Reported metrics are based on the qualification configuration which specifies database size and SQL Server memory.

Summary

ATOS and Dell EMC, in partnership with Microsoft, enables customers to deploy tested and validated data warehouse solutions using Data Warehouse Fast Track reference architectures for SQL Server 2017. These uniquely designed architectures ensure optimal business intelligence solutions. The end-to-end best practices and recommendations enable the customer to achieve enhanced return on investment and faster time to value with a balanced data warehouse environment that can perform better than traditional data warehouse systems.

The ATOS / Dell EMC DWFT reference architectures provide the following benefits:

- Deliver a tested and validated configuration with proven methodology and performance behaviour
- Deliver outstanding performance on the ATOS BullSequana S800 server platform with blazing processor speeds and leading-edge, all-flash-based Dell EMC VMAX 250F storage arrays
- Achieve a balanced and optimized solution at all levels of the stack by following best practices for both hardware and software components, achieving faster time to value, and lower total cost of ownership
- Avoid over provisioning of hardware resources
- Offer high availability at all levels of setup (host, switches, and storage)
- Offer a single point of contact and accountability for purchases, services, and support; SQL Server is available to purchase from Dell EMC worldwide
- Help customers avoid the pitfalls of an improperly designed and configured system
- Reduce future support costs by limiting solution re-architect efforts due to scalability challenges

This paper describes a reference architecture using an ATOS BullSequana S800 server with a Dell EMC VMAX 250F all flash storage array. By implementing Data Warehouse Fast Track for SQL Server 2017 design principles, this configuration achieved a 640TB rating.

Technical support and resources Dell EMC

Dell support is focused on meeting customer needs with proven services and Dell Tech Centre is an online technical community where IT professionals have access for Dell software, hardware, and services.

Storage Solutions Technical Documents on Dell Tech Centre provide expertise that success on Dell EMC storage platforms.

Additional resources:

- Dell EMC products: http://www.dellemc.com
- Dell SQL Server solutions: http://www.dellemc.com/sql

Bill of Materials (BOM)

Item	Description	Qty
1 X BullSequana S800 12 TB Internal disks 2.5GHz P-8180M		
PKMD300-0008	4x 2 socket server	1
FIBR008-M003	OPTICAL FIBRE OM3 MULTI-MODE (SW) LC-LC CABLE 3M	40
CBLE500-E6A03	Ethernet Cable RJ45M/RJ45M cat 6A 3m	16
BCMN100-P8180M	2 Socket 28C 2.5GHz P-8180M-205W Compute Unit	4
MEMK302-6256	256GB(2x128GB DDR4 2666 ECC RDIMM 3DS-8R)1.2V	48
BFDB100-SSD08	Front disk Blade upto 2xDSK-1X800GB 2.5" SAS SSD	2
CKTC700-08I0E	MEGARAID SAS 12Gb/s LSI9361-8i Blade	1
BFDD100-0000	Dummy Front disk Blade	30
PKHB100-FR02	Front disk RAID1 activation - 2 DISKS	1
CKTC400-E16D	16Gb/s F/C Dual HBA w/o cable-LPe31002 Blade	10
DPCI200-0000	Dummy PCI-e Blade	4
PSUP080-2000	2x Hotswap 80+ Platinum 2000W PSU	4
PKWS300-0000	Platform Check for Windows Server 2016 config	1
LABS400-NS30	Label System XAN-S30	1
PCIB600-0000	PCI-e Blade - R	4
CHSB100-0008	Compute Box for a 4x 2 socket server	1
UTRS900-0000	Resource Kit including iCare	1
BCMD100-0001	Dummy Upper Unit for Compute Unit	4
CKTC601-D0SR	10Gb/s DP Eth Adapter - CTX4 - SR Blade	2
2 X BROADCOM DS-6510-B CONNECTRIX B-SERIES SOLUTION		
DS-6510R-B	DS-6510R-B 24P/48P 16GB RTF BASE SWITCH	4
DS6510-RCKMNT	DS-6510-B RACK MOUNT KIT	4
C13-PWR-13	2 C13 PWRCORDS W/ BS546 PLUGS 250V 10A	4
DS6510-16G8PU	DS-6510B 16G 12PORT UPGRADE KIT	8
CTX-OM4-5M	OM4 50/125 MICRON OPTICAL CBL LC-LC 5M	128
W-PS-HW-001	PROSUPPORT W/NBD-HARDWARE WARRANTY	1
PSINST-ESRS	ZERO DOLLAR ESRS INSTALL	1

Item	Description	Qty
1 X VMAX 250F 200TBU		
E-ENCRYPT 250	VMAX 250 DATA AT REST ENCRYPT OS NEW TM	1
ES-CAPX	VMAX 250FX CAPACITY	188
ES-2048ADDE	VMAX 250 DELL ADD 2048GB	1
ES-VBX-2048-ADD	VMAX 250FX VBRCK ADD 2048GB	1
ES-2048BASEE	VMAX 250 DELL BASE 2048GB	1
ES-VBX-2048	VMAX 250FX VBRCK BASE 2048GB	1
JLCF33840S1	VMAX 250 3840GB FLASH SPARE	2
JLCF3384071	VMAX 250 RAID5(7+1) 3840GB	64
ESX-BEDIR	VMAX 250 DIR FX	2
ES-PWRKIT-3Y	3WYE PHASE PDU/PDP KIT	1
ES-IB	VMAX 250F FABRIC	1
ES-DE25-DIR	VMAX 250F DIRECT 25 SLT DR ENCL	4
SZID-11	SIZER ID DIGIT 11 TRACKING MODEL	100
SZID-10	SIZER ID DIGIT 10 TRACKING MODEL	99
SZID-9	SIZER ID DIGIT 9 TRACKING MODEL	1
SZID-8	SIZER ID DIGIT 8 TRACKING MODEL	99
SZID-7	SIZER ID DIGIT 7 TRACKING MODEL	5
SZID-6	SIZER ID DIGIT 6 TRACKING MODEL	1
SZID-5	SIZER ID DIGIT 5 TRACKING MODEL	9
SZID-4	SIZER ID DIGIT 4 TRACKING MODEL	6
SZID-3	SIZER ID DIGIT 3 TRACKING MODEL	4
SZID-2	SIZER ID DIGIT 2 TRACKING MODEL	1
SZID-1	SIZER ID DIGIT 1 TRACKING MODEL	1
ES-1600MOD	VMAX 250F FLASH MOD 1600	4
ES-FE80000S	VMAX 250F 8MM 16G FC	4
ES-COMPRESS	VMAX 250F HDW COMPRESSION	2
E-OPROVISION	OPROVISION FACTOR TRACKING MODEL	13
WKPROFILE-BAL	VMAX VG WORKPROFILE BALANCED	1
ES-PSNT-3Y	VMAX 250F SYS BAY1 3Y PSNT	1
ES-FORMTL	VMAX 250 FOR MTL TRK MODEL	1
ES-SYS1-3Y	VMAX 250F SYS BAY1 3Y	1
ES-SKINS	VMAX 250F SIDE PANELS	1
ES-PC3YAFLE	32A 3PHASE WYE CRD SET FLY LEAD EUROPE	1
ES-MGMT	EMBEDDED MANAGEMENT VMAX 250 TRACKING	1

Item	Description	Qty
450-001-656	VMAX ALL FLASH FX SUITE BASE=IC	1
458-001-777	SRM LIC ENABLE VMAX=IC	1
458-001-711	VMAX FLASH FX POWERPATH PRODUCT=IC	1
458-002-353	VMAX250 FX FLASH ENCRYPTION PRODUCT=IC	1
458-001-691	APPSYNC STR PK FOR VMAX FX SUITE =CC	210
458-001-516	VMAX FLASH FX SUITE OS 1TB=CC	210
450-001-207	VMAX FLASH FX SUITE ENABLER 1TB=CC	210