HITACHI
Inspire the Next

WHITE PAPER

Architecture and Concepts Guide

Hitachi Virtual Storage Platform Gxx0 and Fxx0

Hitachi Vantara
Charles Lofton
Technical Operations – Center for Performance and Tools

August 2018

HITACHI
Inspire the Next

WHITE PAPER

Contents

Notices and Disclaimer

This document has been reviewed for accuracy as of the date of initial publication. Hitachi Vantara may make improvements and/or changes in product and/or programs at any time without notice.

The performance data contained herein was obtained in a controlled isolated environment. Results obtained in other operating environments may vary significantly. While Hitachi Vantara has reviewed each item for accuracy, there is no guarantee that similar results can be obtained elsewhere.

All designs, specifications, statements, information and recommendations (collectively, "designs") in this document are presented "AS IS", with all faults. Hitachi Vantara, its affiliates and their respective suppliers disclaim all warranties, including without limitation, the warranty of merchantability, fitness for a particular purpose and non-infringement or arising from a course of dealing, usage or trade practice. In no event shall Hitachi Vantara Corporation, its affiliates or their respective suppliers be liable for any indirect, special, consequential or incidental damages, including without limitation, lost profit or loss or damage to data arising out of the use or inability to use the designs, even if Hitachi Vantara, its affiliates or their respective suppliers have been advised of the possibility of such damages.

WARNING: This document can only be used as Hitachi Vantara internal documentation for informational purposes only. This documentation is not meant to be disclosed to customers or discussed without a proper non-disclosure agreement (NDA).

About This Guide

Introduction

This document covers the hardware architecture and concepts of operations for the Hitachi Virtual Storage Platform Gxx0 and Fxx0 storage systems. This document is not intended to cover any aspects of the storage software, customer application software, customer specific environments, or features available in future releases.

Intended Audience

This document will familiarize Hitachi Vantara sales personnel, technical support staff, approved customers, and value-added resellers with the features and concepts of the VSP Gxx0 and Fxx0 family. Users who will benefit the most from this document are those who already possess an in-depth knowledge of the Hitachi Virtual Storage Platform Gx00 architecture.

This document will receive future updates to refine or expand on some discussion as the internals of the design are better understood or as upgrades are released.

Document Revisions

Revision	Date	Description
0.1	June 2018	Initial release of the document for internal review
0.2	July 2018	Initial release of the document for PM review
1.0	August 2018	Final draft

References

- *VSP Gx00 and Fx00 Architecture and Concepts Guide*
- *VSP G1500 Architecture and Concepts Guide*
- *Various Hitachi Specification and Training documents*

Contributors

The information included in this document represents the expertise, feedback, and suggestions of a number of skilled practitioners. The author would like to recognize and sincerely thank the following contributors and reviews of this document (listed alphabetically by last name):

- Wendy Bawden *(Global Field and Industry Solutions)*
- James Byun *(Center for Performance and Tools)*
- Gilbert Gerber (*Global Field and Industry Solutions)*
- Greg Loose *(Global Field and Industry Solutions)*
- Alexey Silin *(Global Field and Industry Solutions)*

System Highlights

The VSP Gxx0 storage systems are the successors to the Hitachi Virtual Storage Platform Gx00 product line and continue the transition to a single microcode base where all models, from entry level to enterprise, run the same Storage Virtualization Operating System (SVOS).

Across the board, the VSP Gxx0 family is a tremendous leap forward from the VSP Gx00 series in both performance and ease of use. The VSP G350 brings enterprise class in-system and remote replication functionality to the entry level model while the VSP G900, at the other end, provides 2.5X the cache-miss random read IOPS of VSP G800 with significantly lower latency. The controller design is a “compacted logical” implementation of the VSP G1500, and is akin to the VSP Gx00 design albeit with newer generation CPUs and highly efficient microcode optimized for low-latency performance with flash drives.

Augmenting the hybrid (spinning disk and flash media) VSP Gxx0 models are the all flash VSP Fxx0 models: VSP F350, F370, F700, and F900. These are architecturally identical to the equivalent Gxx0 models but are configured as all flash arrays (AFA). The midrange F350 and F370 can be configured with SSDs, while the higher-end F700 and F900 are available with SSDs and/or FMDs. While most aspects of the Gxx0 models are applicable to the corresponding Fxx0 models, the Fxx0 configuration rules are unique and discussed in detail separately.

These are the distinguishing features of the VSP Gxx0 family:

- The VSP Gxx0 has a single control chassis, either 2U (G350/G370) or 4U (G700/G900) in height, and one or more 19” racks. A maximum of 1,440 disks may be installed using 24 dense drive boxes (DB60) in the G900 model, but the max drive count will vary based on the array model and type of drive boxes selected. All Gxx0 models can be completely diskless, where all storage is externally attached and virtualized. Gxx0 & Fxx0 models can also have a combination of internal, and external virtualized storage.
- Each controller is packaged as a single controller board called a Controller Blade. There are two clustered Controller Blades per system.
  - At the heart of each Controller Blade is either a single or dual CPU socket, housing Intel Broadwell EP Xeon processors. Each model operates with one logical Microprocessor Unit (MPU) per controller. Each MPU logically encapsulates a single Xeon processor (G350/G370) or two Xeon processors (G700/G900).
  - Like the VSP Gx00 models, the VSP Gxx0 family does not utilize any custom ASICs. This is a significant change from the VSP G1500 and HUS VM designs. As a point of comparison, all of the functions performed by the HM ASIC in the HUS VM are now implemented using a combination of the hardware capabilities of the Intel CPUs and software ASIC emulation routines executed by the MPUs.
  - The VSP G350 and G370 Controller Blades feature an integrated, non-removable Back-end SAS controller that is connected to the controller chassis internal drive slots and connected to external disk trays with 4 x 12 Gbps SAS links. Two versions are available: the standard Controller Blade (CTLS) and the Controller Blade with encryption support (CTLSHE). The VSP G700 and G900 rely on Back-end Modules (BE Modules) that are installed in chassis I/O Module slots. These BE Modules each contain a single SAS Protocol Controller chip (standard or encrypting) that each provide 2 disk tray cable ports, each port having 4 x 12 Gbps SAS links.
  - Host connectivity (and external storage or remote copy links) for all G/Fxx0 models comes in the form of Front-end Modules (FE Modules) installed into I/O Module slots either in the controller blade (G350 and G370) or in the controller chassis (G700/G900). All ports are “universal” meaning they can be used as a target and/or initiator for host access, replication and virtualization although it is preferable to use each port for only one purpose to make monitoring and troubleshooting easier. There are two types of FE Modules available:
    - 4 x 8/16/32 Gbps Fibre Channel (FC) ports
    - 2 x 10 Gbps iSCSI ports (copper or optical)

- The FE Modules listed above as well as the BE Modules are common across the VSP Gx00 and VSP Gxx0 families.
- The VSP G900 supports an external 2U Front-end I/O Expansion Box which provides eight slots for FE Modules. It connects to the G900 via four PCIe Pass Through Modules installed into controller chassis I/O Module slots, for a net gain of four additional FE Modules.
- Each Controller Blade contains half of the system cache and connection to Cache Flash Memory (CFM) which is used for cache destage in case of power failure, and backup copies of configuration data, encryption keys, etc.
- The connectivity within each Controller Blade is provided by PCI Express 3.0 links and also Intel QuickPath Interconnect (QPI) for the G700 and G900 models which have two CPU sockets per blade. QPI enables one CPU direct access to the PCI express paths on the other CPU without requiring interrupts or communication overhead (aka transparent bridge). The VSP G350/G370 do not feature QPI because the controller blades only feature a single CPU. The external connections between the two Controller Blades are provided by PCI Express 3.0 links using non-transparent bridging.
- Like the VSP G1500, the VSP Gxx0 family uses a cache-based Shared Memory (SM) architecture, often referred to as Control Memory to indicate its system function. The master SM copy is mirrored between the two Controller Blades. Additionally, each Controller Blade has a local, non-mirrored copy of SM that is used by the MPU on that blade for accessing metadata and control tables for those volumes (LDEVs) it manages. The majority (perhaps 80%) of all SM accesses are simply reads to this local copy. Updates are written to both the local and master SM copies.
- Each MPU controls all I/O operations for a discrete group of LDEVs (LUNs when they are mapped to a host port) in the same manner that they are managed by single VSD processor boards in the VSP G1500 array. When first created, each new LDEV is automatically assigned in a round-robin fashion to one of the two MPUs. However the administrator can optionally choose the assignment when created, and the assignment can be changed while the LDEV is in service for load balancing purposes.
- Each MPU executes the Storage Virtualization Operating System (SVOS) for the following processes for those volumes (LDEVs) that it manages:
    - Target mode (Open Systems hosts)
    - External mode (Virtualization of other storage)
    - Back End mode (Operate FMD/SSD/HDD drives in the subsystem and handle parity calculation)
    - Replication Initiator mode (GAD, HUR, or TrueCopy Sync)
    - Replication Target mode
    - In-system Replication (ShadowImage or Thin Image)
    - Data Reduction (compression and deduplication)

- The VSP Gxx0 family uses the same drive boxes used by VSP Gx00 arrays. The 12 Gbps drive box variations are listed below:
    - DBS: 2U 24-slot SFF, with one row of 24 x 2.5” vertical disk drive slots
        - The G350/G370 Controller Chassis with integrated DBS is known as the CBSS
    - DBL: 2U 12-slot LFF, with three rows of 4 x 3.5” horizontal disk drive slots
        - The G350/G370 Controller Chassis with integrated DBL is known as the CBSL
    - DB60: 4U 60-slot dense LFF, with five top-loaded rows of 12 x 3.5” disk drive slots
    - DBF: 2U 12-slot FMD, with four horizontal rows of 3 x FMD slots

- Drive choices include:
  - SFF drives (DBS): 480 GB SSD (excluding G/F900), 960 GB SSD,1.9 TB SSD, 3.8 TB SSD, 7.6 TB SSD, 600 GB 10K, 1.2 TB 10K, 1.8 TB 10K, and 2.4 TB 10K
  - SFF drives in a LFF carrier (DB60 only) 1.2 TB 10K, 2.4 TB 10K
  - LFF drives (DBL, DB60): 6 TB 7.2K and 10 TB 7.2K
  - FMD DC2 drives (DBF): 3.2 TB
  - FMD HD drives (DBF): 7 TB and 14 TB

What is New in VSP Gxx0

- Controller Blades have been upgraded with Intel Xeon Broadwell EP processors with core architectural advancements and improved DDR4 memory bandwidth.
  - Maximum supported cache per Controller Blade has doubled compared to the VSP Gx00 family
- Optimized SVOS RF 8.1 includes Direct Command Transfer, which doubles random IOPS per processor core with reduced latency. Microcode improvements include:
  - Integrated front end and back end read jobs to reduce processing overhead
  - Reduced inter-controller access for faster response time
  - Streamlined transaction processing to reduce CPU time per I/O
- Ease of use has been improved in the following ways:
  - SVOS automatically calculates the required SM capacity based on total DIMM and pool capacity
  - A simple, intuitive graphical user interface for basic provisioning and monitoring, Hitachi Storage Advisor Embedded, was built into the controller. The PF_REST API was implemented in the controller instead of the Service Processor. Hi-Track now uses the external Hi-Track Monitor application on a customer's server (as many of our other products do) instead of the Hi-Track Agent running on the Service Processor. These (and other) changes make it possible for the system to be used without a Service Processor in many cases (but with some caveats).
  - For pools that use FMDs with Accelerated Compression enabled, SVOS RF will, by default, automatically create new Pool Volumes and grow the pool to take advantage of compression savings. This is documented in the Provisioning Guide. At this time the default behavior can only be changed (to manual growth) via CLI.
- VSP Gxx0 employs a single MPU per controller, vs. two MPUs per controller on VSP Gx00. The single-MPU architecture has the following advantages:
  - More efficient cache allocation when multiple cache logical partitions are used
  - Reduced LR emulator overhead because fewer commands must be assigned to a different MPU
  - Increased parallelism for LDEV-related tasks

Hitachi Virtualization (UVM: Universal Volume Manager)

The VSP Gxx0 family provides the same Hitachi Virtualization mechanism as found on the VSP G1500 and earlier generation high-end systems. Other customer storage systems (often being repurposed upon replacement by newer systems) may be attached to some of the front-end FC or iSCSI ports. These paths would then operate in External Initiator mode. The LUNs that are supplied by these external systems are accessed and managed by hosts that are attached to the same or other front-end ports (utilizing Target mode). As far as any host is concerned, all of the visible virtualized LUNs passed through the VSP Gxx0 external ports to the host target ports simply appear to be normal internal LUNs in the VSP Gxx0 array. The VSP Gxx0's bidirectional FC and iSCSI ports allow simultaneous host and external storage connectivity without the need for dedicated "ePorts" and host ports.

UVM can greatly simplify storage management, particularly when used to consolidate and virtualize arrays from multiple storage vendors. Another advantage of virtualization is the ability to non-disruptively migrate data from external to internal storage or vice versa. These LDEV migrations are able to proceed while the original source LDEV remains online to the hosts, and the VSP Gxx0 will seamlessly switch over the mapping from the source LDEV to the target LDEV when completed. No changes to host mapping are required. Virtualized (external) storage can also be used with HDP and HDT such that data can be automatically migrated between different performance tiers, whether they are internal or external. While it isn't required, it is generally recommended to dedicate higher tiers to internal storage and use external storage for lower tiers due to the added latency in the I/O path.

Hitachi Dynamic Provisioning (HDP)

Hitachi Dynamic Provisioning (HDP) provides a logical mechanism for grouping LDEVs from multiple Parity Groups into a single pool (with many such independent pools possible) that presents an automatically managed, wide striped block device to one or more hosts. An HDP pool is defined with one or more LDEVs (pool volumes) from Parity Groups with one RAID level and one drive type. HDP balances page allocation across Parity Groups, so the best practice is to use Parity Groups having approximately the same capacity. The pool mechanism creates a structure of 42 MB pool pages from each LDEV within the pool when initially created. This is similar to the use of a host-based logical volume manager (LVM) and its wide striping mechanism across all member LUNs in its "pool" with use of a (typically) large volume chunk size (usually 1 MB or greater).

Dynamic Provisioning Volumes (DP-VOLs or virtual volumes) are then created, with a user specified logical size (up to 256 TB). The host accesses the DP-VOL (or many of them – even hundreds) as if it were a normal volume (LUN) over one or more host ports. A major difference is that disk space is not physically allocated to a DP-VOL from the pool until the host has written to different parts of that DP-VOL's Logical Block Address (LBA) space. The entire logical size specified when creating that DP-VOL could eventually become fully mapped to physical space using 42 MB pool pages from every LDEV in the pool. If new LDEVs from new Parity Groups are added to the pool later on, a rebalance operation (restriping) of the currently allocated pool pages onto the new PGs is initiated automatically.

Hitachi Dynamic Tiering (HDT)

Hitachi Dynamic Tiering (HDT) is a feature that enhances the operation of an HDP pool. Previously, each HDP pool had to be created using one RAID level and one drive type. For example, a high performance pool could be set up using FMDs or SSDs in a RAID-10 (2D+2D) configuration. A standard performance pool could use 10K SAS drives and RAID-5 (such as 7D+1P). An archive performance pool could be set up with 7200 RPM SAS drives in RAID-5 or RAID-6 for near-line storage. The lower performance pools might also be created using LUNs from virtualized storage (typically with 10K or 7200 RPM SAS drives).

The HDT feature allows a single pool to contain multiple types of Parity Groups (pool volumes, using any available RAID level) and any type of drive, as well as external LUNs from virtualized storage. Up to three choices from these possible combinations are allowed per pool. An example would be FMDs (Tier 1), 10K SAS (Tier 2), and external storage using 7.2K SAS (Tier 3). Only one RAID level is normally used per Tier.

However, when a Tier is to be changed from a drive type or RAID level to another, the makeup of each of the Tiers may change temporarily for the time the migration is in progress. For example, Tier 2 may have been established using 15K SAS drives and RAID-5 (7D+1P) but it is desired to change this to SAS 10K and RAID-6 (6D+2P). The 10K SAS drives would temporarily become part of Tier 3 (which allows disparate drive types) until the migration is complete and the original 15K SAS pool volumes removed, at which point the 10K SAS drives will be moved up to Tier 2.

The original pool volume may be deleted using pool shrink, and in doing so the HDT software will relocate all allocated 42 MB pages from that pool volume to all the new ones. Once the copy is completed (may take a long time) then that pool volume (an LDEV) is removed from that pool. That LDEV can now be reused for something else (or deleted, and the drives for that Parity Group removed from the system).

HDT manages the mapping of 42 MB pool pages within these various tiers within a pool automatically. Management includes the dynamic relocation of a page based on frequency of back end disk I/O to that page. Therefore, the location of a pool page (42 MB) is managed by HDT according to host usage of that part of an individual DP-VOL’s LBA space. This feature can eliminate most user management of storage tiers within a subsystem and can maintain peak performance under dynamic conditions without user intervention, while being more granular than volume or LUN-based approaches. This mechanism functions effectively without visibility into any file system residing on the volume. The top tier is kept nearly full at all times, with stale pages being moved down a tier to make room for new high-activity pages. In addition, the I/O type to each page is tracked so that pages with the highest random read rates are given priority access to the FMD/SSD tier. HDT also provides optional per DP-VOL policies to force the min or max % of data per tier to handle exceptions when the default tiering policy may not be appropriate. HDT by itself does not provide real-time page relocation as it must wait for the end of a monitoring cycle (minimum 30 minutes) before taking action.

When Accelerated Compression is enabled on the FMD tier in a pool, HDT is able to react to changes in compression and overprovisioning level, such that it will move data down to the next tier rather than exhaust all physical FMD capacity which would lead to an error condition.

Active Flash

When enabled on an HDT pool, Active Flash allows for real-time relocation of 42 MB pages by adding a short-term access frequency counter for each page. Prompt promotions and high priority demotions can be triggered immediately without waiting for the next HDT relocation cycle to begin, enhancing HDT’s ability to respond to sudden host workload changes. Wear leveling across back-end parity groups is also performed when pages are promoted into the flash tier. Active Flash can be dynamically enabled or disabled per pool and is a non-disruptive process.

Adaptive Data Reduction

Introduced in SVOS 7.0, Adaptive Data Reduction adds controller based compression and deduplication, complementing the existing hardware compression capability of the FMD HD drives. Collectively, these data reduction technologies can increase the effective storage capacity presented by the array. The LZ4 lossless compression algorithm is used to reduce the number of physical bits needed to represent the host written data. Deduplication removes redundant copies of identical data segments and replaces them with pointers to a single instance of the data segment on disk. These capacity saving features are supported in conjunction with HDP, so that only DP-VOLs can have either compression or deduplication plus compression enabled. Deduplication without compression is not supported. If an HDP pool is comprised of POOL-VOLs all from FMD DC2 or FMD HD parity groups with Accelerated Compression enabled, SVOS will offload compression tasks to the FMDs for reduced system overhead.

When deduplication is enabled in an HDP pool, four additional Virtual Volumes called Deduplication System Data Volumes (DSD-VOLs) are automatically created in it. These are metadata volumes used by the data reduction engine for such things as storing the hash table and keeping track of deduplication status. The DSD-VOLs are 10 TB thin provisioned DP-VOL and combined on average will consume about 1% of the used data capacity of the pool. Each DP-VOL has a capacity saving attribute, for which the settings are “Disabled”, “Compression”, or “Deduplication and Compression”. DP-VOLs with either capacity saving attribute set are referred to as Data Reduction Vols (DRD-VOL). The deduplication scope is at the HDP pool level for all DRD-VOLs with the “Deduplication and Compression” attribute set. As a result, in order to support global deduplication for an entire array it must be configured with a single pool. The chunk size that the data reduction engine operates on is 8 KB, for both deduplication and compression.

In practice, the data reduction engine uses a combination of inline and post-process methods to achieve capacity saving with the minimum amount of overhead to host I/O. Normally with HDP, each DP-VOL is made up of multiple 42 MB physical pages allocated from the HDP pool. But with data reduction enabled, each DRD-VOL is made up of 42 MB virtual pages. If a virtual page has not yet been processed for data reduction, it is identified as a non-DRD virtual page and is essentially a pointer to an entire physical page in the pool. After data reduction, the virtual page is identified as a DRD virtual page and it then contains pointers to 8 KB chunks stored in different physical pages in the HDP pool.

The initial data reduction post-processing is done to non-DRD virtual pages that have not had write activity in at least five minutes. The non-DRD virtual page is processed in 8 KB chunks and compressed data is written in log-structured fashion to new locations in the pool, likely one or more new physical pages. If enabled for the DRD-VOL, deduplication is then performed on the compressed data chunks, so that duplicate chunks are invalidated (after a hash match and bit-by-bit comparison) and replaced with a pointer to the location of the physical chunk. Garbage collection is done in the background to combat fragmentation over time by coalescing the pockets of free space resulting from invalidated data chunks. Subsequent rewrites to already compressed data are then handled purely inline for best performance.

For more details on how to utilize compression and deduplication, refer to the *VSP G and F Series Provisioning Guide*.

Glossary

At this point some definitions of the various terminology used is necessary in order to make all of the following discussions easier to follow. Throughout this paper the terminology used by Hitachi Vantara (not Hitachi Ltd. in Japan) will normally be used. As a lot of storage terminology is used differently in Hitachi documentation or by users in the field, here are the definitions as used in this paper:

- **Array Group** (installable, drive feature): The term used to describe a set of at least four physical drives installed into any disk tray(s) (in any “roaming” order on VSP Gxx0). When an Array Group is formatted using a RAID level, the resulting RAID formatted entity is called a Parity Group. Although technically the term Array Group refers to a group of bare physical drives, and the term Parity Group refers to something that has been formatted as a RAID level and therefore actually has initial parity data (here we consider a RAID-10 mirror copy as parity data), be aware that this technical distinction is often lost. You will see the terms Parity Group and Array Group used interchangeably in the field.
- **Back-end Module** (**BE Module**, installable, **DKB** feature): A SAS drive controller module that plugs into a socket in the Controller Chassis and provides the eight back-end 12 Gbps SAS links via two SAS 4-Wide ports per module. There are four of these modules (two pairs) installed in a VSP G700, and 4 or 8 installed in a G900, unless it is purchased as a diskless system, which can then have extra FE Modules (more FC or iSCSI ports) installed instead of the BE Modules. Strictly speaking, the VSP G350 and G370 do not have BE Modules but integrated Back-End controllers, which are identical in function to the BE Modules but are not removable.
- **Bidirectional Port**: A port that can simultaneously operate in Target and Initiator modes. This means the port supports all four traditional attributes without requiring the user to choose one at a time:
  - Open Target (TAR)
  - Replication Target (RCU)
  - Replication Initiator (MCU)
  - External Initiator (ELUN)

- **Cache Directory**: The region reserved in cache for use by the MPUs in managing the User Data cache region. The Cache Directory size varies according to the size of the User Data cache region, which is directly affected by the size of Shared Memory.
- **CB** (**Controller Chassis**): Hitachi's name for the bare Controller Box, which can come in one of five types.
  - **CBSS**: VSP G350/G370 chassis with internal DBS
  - **CBSL**: VSP G350/G370 chassis with internal DBL
  - **CBL**: VSP G700/G900 chassis
- **CFM** (**Cache Flash Module**): SATA SSD that serves as a cache backup device in case of power loss. There are designated CFM slots into which these are installed.
- **CHB** (**Channel Blade**): Hitachi's name for the Front-end Module.
- **CHBB** (**Channel Blade Box**): Hitachi's name for the Front-end I/O Expansion Box.
- **Cluster**: One half or side of the array, consisting of a Controller Blade, its components, and the I/O Modules connected to it. Cluster 1 refers to the side containing Controller Blade 1 and Cluster 2 refers to the side containing Controller Blade 2.
- **Concatenated Parity Group**: A configuration where the VDEVs corresponding to a pair of RAID-10 (2D+2D) or RAID-5 (7D+1P) Parity Groups, or four RAID-5 (7D+1P) Parity Groups, are interleaved on a RAID stripe level on a round robin basis. A logical RAID stripe row is created as a concatenation of the individual RAID stripe rows. This has the effect of dispersing I/O activity over twice or four times the number of drives, but it does not change the number, names, or size of VDEVs, and hence it doesn't make it possible to assign larger LDEVs to them. Note that we often refer to RAID-10 (4D+4D), but this is actually two RAID-10 (2D+2D) Parity Groups interleaved together. For a more comprehensive explanation refer to *Appendix 5* of the *VSP G1500 Architecture and Concepts Guide*.
- **CTL** (**Controller Blade**): The shorthand name for the Controller Blade, not to be confused with an HUS 100 Controller which was also abbreviated CTL. The Intel Broadwell EP processors and Cache DIMMs are physically installed in the CTL.
- **DARE** (**Data-at-Rest Encryption**): Controller-based data encryption of all blocks in a Parity Group, enabled via software license key.
- **DB** (**Disk Box**): Hitachi's name for the disk enclosures.
  - **DBS**: 2U 24-slot SFF SAS box
  - **DBL**: 2U 12-slot LFF SAS box
  - **DB60**: 4U 60-slot dense LFF SAS drawer (supports SFF intermix via special drive canisters)
  - **DBF**: 2U 12-slot FMD box
- **DIMM** (**Dual Inline Memory Module**): A "stick" of RAM installed in the corresponding DIMM sockets on the Controller Blades.
- **DKB** (**Disk Blade**): Hitachi's name for the Back-end Module.
- **DKC** (**Disk Controller**): Hitachi's name for the controller unit as a whole, comprised of the Controller Chassis (CB), Controller Blades (CTL), FE and BE Modules, Power Supplies, etc. The Controller Chassis (CB) is often also referred to as the DKC.
- **DP-VOL** (configurable, **Dynamic Provisioning VOLume**): The Virtual Volume connected to an HDP pool. Some documents also refer to this as a V-VOL, not to be confused with a VMware VVol. Each DP-VOL has a user specified size between 48 MB and 256 TB in increments of one block (512 byte sector) and is built upon a set of 42 MB pages of physical storage.

- **Drive** (**Disk**): An FMD, SSD, or HDD. SATA disks are not supported in the VSP Gxx0 family.
- **DRR** (**DRR Emulator**, **Data Recovery and Reconstruction**): Virtual processors that run on the VSP Gxx0 MPUs in microcode (software) that manage RAID parity operations and drive formatting or rebuilds.
- **eLUN** (configurable, **External LUN**): An External LUN is one which is located in another storage system and managed as though it were just another internal LDEV. The external storage system is attached via two or more FC or iSCSI Ports and accessed by the host through other front-end target ports. The eLUN is used within the VSP Gxx0 as a VDEV, a logical container from which LDEVs can be carved. Individual external LDEVs may be mapped to a portion of or to the entirety of the eLUN. Usually a single external LDEV is mapped to the exact LBA range of the eLUN, and thus the eLUN can be "passed through" the VSP Gxx0 to the host.
- **FC Port**: Any of the Fibre Channel ports on a Fibre Channel FE Module. Each VSP Gxx0 family FC Port is a Bidirectional Port.
- **Feature** (package): An installable hardware option (such as an FE Module, BE Module, or Cache DIMM) that is orderable by Feature Code (P-Code). Each of the VSP Gxx0 features is a single board or module, and not a pair like some of the VSP G1000 features.
- **FMD**: The Flash Module Drive that installs in the DBF disk box. First generation FMDs come in 1.6 and 3.2 TB capacities while second generation FMD DC2 models which feature in-drive data compression come in 1.6, 3.2, and 6.4 TB capacities. Third generation FMD HD drives are high density variants of the FMD DC2, also featuring built-in compression, a lower bit cost ($/GB) and larger 7 and 14 TB capacities. The latest FMD HD drives are based on higher-density 3D NAND. Note that FMD and FMD DC2 drives, while listed with capacity in decimal TB, are actually sized in binary TiB. Converted to TB, the actual capacities for these modules would be 1.75, 3.5, and 7 TB. Refer to *Table **6*** for the exact raw size of each supported FMD drive.
- **GB (gigabyte) / GiB (gibibyte)**: The *gigabyte* (GB) is a decimal unit of measurement equal to 10^9 bytes while the *gibibyte* (GiB) is a binary unit of measurement equal to 2^{30} bytes. 1 GiB is roughly equivalent to 1.074 GB. Standard practice has been to report all drive capacities in decimal units (GB, TB) while LDEVs or LUNs are sized in binary units (MiB, GiB, TiB) despite using the incorrect nomenclature (MB, GB, TB).
- **GUM (Gateway for Unified Management)**: The embedded micro server (Linux) on each Controller Blade that runs Hitachi Storage Advisor Embedded, the REST API, the CCI CLI and provides the interface for the Hi-Track Monitor (external server).
- **HSAE (Hitachi Storage Advisor Embedded)**: The simple graphical user interface for basic provisioning and monitoring that runs on the GUM.
- **KB (kilobyte) / KiB (kibibyte)**: The *kilobyte* (KB) is a decimal unit of measurement equal to 1,000 (10^3) bytes while the *kibibyte* (KiB) is a binary unit of measurement equal to 1,024 (2^{10}) bytes. 1 KiB is equal to 1.024 KB.
- **LDEV** (configurable, **Logical DEVice**): A logical volume internal to the system that can be used to contain customer data. LDEVs are uniquely identified within the system using a six-digit identifier in the form LDKC:CU:LDEV. LDEVs are carved from a VDEV (see VDEV), and there are three types of LDEVs: internal LDEVs, external LDEVs, and DP-VOLs. LDEVs are then mapped to a host as a LUN. Note: what is called an LDEV in all Hitachi enterprise systems and the HUS VM is called an LU or LUN in Hitachi Vantara modular systems like the Hitachi Unified Storage 100 (HUS 100) family.
- **LR** (**Local Router**, **Local Router Emulator**, or **Command Transfer Circuit**): Virtual processors that run on the VSP Gxx0 MPUs in microcode (software) that facilitate the transfer of commands between FE or BE Modules and the MPUs.
- **LUN** (configurable, **Logical Unit Number**): The host-visible identifier assigned by the administrator to an existing LDEV to make it usable on a host port. An internal LUN has no actual queue depth limit (but 32 is a good rule of thumb) while an external (virtualized) eLUN has a Queue Depth limit of 2-128 (adjustable) per external path to that eLUN. In Fibre Channel, the host HBA Fibre Channel port is the initiator, and the system's virtual Fibre

Channel port (or Host Storage Domain) is the target. Thus the Logical Unit Number is the number of the logical volume within a target.

- **MB (megabyte) / MiB (mebibyte)**: The *megabyte* (MB) is a decimal unit of measurement equal to 10^6 bytes while the *mebibyte* (MiB) is a binary unit of measurement equal to 2^{20} bytes. 1 MiB is roughly equivalent to 1.049 MB.
- **MP** (**Microprocessor**): An individual MPU core, which is a single core of the Intel Broadwell EP Xeon CPU. Not to be confused with a FED MP or BED MP from USP V and earlier enterprise arrays.
- **MPU** (**Microprocessor Unit**): The multi-core logical unit that is superimposed on the physical CPU. In the case of the VSP G350, the MPU is a 6-core logical unit (6 MPs) that comprises the 6 cores on a single CPU. The MPU in a G370 encapsulates a single 10-core CPU. The G700 MPU is a 12-core logical unit (12 MPs) that comprises the 12 cores on two 6-core CPUs. The G900 MPU contains 20 MPs (two 10-core CPUs). Each MPU also contains a small amount of Package Memory (PM).
- **OPEN-V**: The name of the RAID mechanism on all VSP G series storage systems for Open (non-mainframe) hosts. Refer to *Appendix 1* of the *VSP G1500 Architecture and Concepts Guide* for more details.
- **Parity Group** (configurable, known as a **RAID Group** in AMS and HUS midrange arrays): A set of drives formatted as a single RAID level, either as RAID-10 (sometimes referred to as RAID-1+ in Hitachi Vantara documentation), RAID-5, or RAID-6. The VSP Gxx0's supported Parity Group types are RAID-10 (2D+2D), RAID-5 (3D+1P, 4D+1P, 6D+1P, or 7D+1P), and RAID-6 (6D+2P, 12D+2P, and 14D+2P). The OPEN-V RAID chunk (or stripe) size is fixed at 512 KiB. Internal LDEVs are carved from the VDEV(s) corresponding to the formatted space in a Parity Group. The maximum size of an internal LDEV is 6,442,319,360 (512-byte) blocks (sectors), or 3,298,467,512,320 bytes, which is 2.999938726 TiB. If the formatted space in a Parity Group is bigger than this value, then multiple LDEVs must be employed to use all the space on that Parity Group. Note that in earlier generations of Hitachi subsystems, the maximum VDEV size was also ~2.99 TiB, and all VDEVs other than the last in a Parity Group were maximum size. Now there is only one VDEV for each Parity Group, regardless of the size of the Parity Group. Note also that there actually is no discrete 4D+4D Parity Group type – see Concatenated Parity Group
- **PCIe** (**PCI Express**): A multi-channel serial bus connection technology that supports x1, x4, and x8 lane configurations. The resulting connection is called a "PCIe link". The VSP Gxx0 uses the x8 type in most cases. The PCIe 3.0 x8 link is capable of 8 GB/s send plus 8 GB/s receive in full duplex mode (i.e. concurrently driven in each direction). Refer to *Appendix 8* of the *VSP G1500 Architecture and Concepts Guide* for more details.
- **PDEV** (**Physical DEVice**): A physical internal drive.
- **RAID-1**: Used in some documents to describe what is usually called "RAID-10", a *stripe of mirrored pairs*. Thus when we say "RAID-1" in the context of a Hitachi VSP-family system, we mean the same thing as when we say "RAID-10" in the context of an HUS 100 modular system. Note that the alternative RAID-0+1 used by some vendors is quite different, as it is the very vulnerable *mirror of two RAID-0 stripes*, where if one disk fails, all protection is lost. In a mirror of stripes, it's not that you lose the data on a single drive failure, because after all, it's still a mirror, but in a mirror of stripes if one drive fails, the entire stripe goes down, and you are very vulnerable to a second drive failure in the other stripe. In generic RAID-10 where there is a stripe of mirrors, if two drives fail in different mirror pairs, then each mirror pair is still "alive" within the stripe and thus no data are lost. This is how the VSP Gxx0 family works.
- **SAS (Serial Attached SCSI)**: A point-to-point serial protocol for the transfer of data to and from storage devices. A SAS link or channel is an individual bidirectional path running at 6 or 12 Gbps. External connections are provided by SAS cables that package four links into a single wire bundle, sometimes referred to as a SAS wide cable. Most SAS drives are dual-ported, that is they can be attached to two SAS links at the same time, but FMDs are quad-ported, meaning they can be attached to four SAS links at the same time.
- **Shared Memory** (**SM**): Otherwise referred to as Control Memory or the Management Area, it is the region partitioned from physical cache memory that is used to manage all volume metadata and system states. In

general, Shared Memory contains all of the metadata in a storage system that is used to describe the physical configuration, track the state of all LUN data, track the status of all components, and manage all control tables that are used for I/O operations (including those of Copy Products). The overall footprint of Shared Memory in cache can range from 39.25 GB - 81.25 GB for VSP G350, to 118.25 GB – 169 GB for VSP G900.

- **SVP** (**Service Processor**): A Windows server that functions as a management interface to the VSP Gxx0 array. Hitachi Vantara offers a 1U server for running a dedicated SVP, but customers have the option of installing the Management Appliance (MApp) software on their own physical server or a Windows virtual machine. The MApp software consists of:
    - Block Element Manager (BEM) – This is the Storage Navigator software component.
    - SVP Remote Method Invocation (RMI) code – This is how the SVP interfaces with the GUM on each Controller Blade.
- **TB (terabyte) / TiB (tebibyte)**: The *terabyte* (TB) is a decimal unit of measurement equal to 10^{12} bytes while the *tebibyte* (TiB) is a binary unit of measurement equal to 2^{40} bytes. 1 TiB is roughly equivalent to 1.1 TB.
- **VDEV**: The logical storage container from which LDEVs are carved. The original VDEV in the earliest Hitachi enterprise RAID subsystems corresponded to the formatted space in a Parity Group. This type of VDEV is called an **Internal** VDEV. There are several additional types of VDEVs on the VSP Gxx0:
    - **External** storage VDEV (60 TiB max) – Maps to a LUN on an external (virtualized) system. LDEVs carved from external VDEVs are called external LDEVs. Often a single external LDEV is mapped to the entire external storage VDEV (one external LDEV of exactly the same size mapped to the LUN on the virtualized subsystem).
    - Each DP-VOL in Hitachi Dynamic Provisioning (HDP) or Hitachi Dynamic Tiering has its own underlying VDEV, as do the LDEVs used in the earlier Copy on Write (CoW) or the current Hitachi Thin Image products.
    - Each DP-VOL in Hitachi Dynamic Provisioning (HDP) with ADR enabled has two underlying VDEVs: one for uncompressed data, and a second VDEV for compressed data written in log-structured format.

Architecture Overviews by Model

This section discusses the general architecture overviews of each model. An expanded discussion of the internal elements of each model is found in the following sections. Tables 1-2 list the major differences within the VSP Gxx0 and Fxx0 families. All quantities are for the complete system and are a sum of the components found on both Controller Blades. The two Controller Blades must be configured in the same manner (host port options and cache). *Table **3*** shows the major system limits for each model in the VSP Gxx0 family with a comparison to the VSP Gx00 series models.

Model	Maximum Disks	Cache Sizes (GB)	Total Host Paths	SAS Disk Links	Internal Cache Bandwidth
G900	1,440 *1	512,1024	16 FE Modules *5, each with: 4 x 4/8/16/32 *6 Gbps FC, or 2 x 10 Gbps iSCSI	32, 64 *7	272 GB/sec. (16 DIMMs)
G700	1200 *2	512	12 FE Modules, each with: 4 x 4/8/16/32 *6 Gbps FC, or 2 x 10 Gbps iSCSI	32	240 GB/sec. (16 DIMMs)
G370	372 *3	256	4 FE Modules, each with: 4 x 4/8/16/32 *6 Gbps FC, or 2 x 10 Gbps iSCSI	16	68 GB/sec. (4 DIMMs)
G350	252 *4	128	4 FE Modules, each with: 4 x 4/8/16/32 *6 Gbps FC, or 2 x 10 Gbps iSCSI	16	60 GB/sec. (4 DIMMs)

Table 1: Comparison of the VSP Gxx0 Systems

*1 Requires 24 x DB60 (DBS max is 1,152 and DBL max is 576)

*2 Requires 20 x DB60 (DBS max is 864 and DBL max is 432)

*3 Requires CBSL and 6 x DB60 (DBS max is 288 and DBL max is 144)

*4 Requires CBSL and 4 x DB60 (DBS max is 192 and DBL max is 96)

*5 Requires use of the I/O Expansion Box with the Standard Back-end (4 BE Modules)

*6 32Gbps requires a specific SFP. 4/8/16Gbps and 8/16/32Gbps SFPs can be mixed on the same CHB.

*7 32 SAS links in the Standard Back-end (4 BE Modules), 64 SAS links in the Performance Back-end (8 BE Modules)

Note: The two Controller Blades must be symmetrically configured with respect to the FE module types and slots they are installed in.

Model	Maximum Disks	Cache Sizes (GB)	Total Host Paths	SAS Disk Links	Internal Cache Bandwidth
F900	1152 *1	512,1024	16 FE Modules *5, each with: 4 x 4/8/16/32 *6 Gbps FC, or 2 x 10 Gbps iSCSI	32, 64 *7	272 GB/sec. (16 DIMMs)
F700	864 *2	512	12 FE Modules, each with: 4 x 4/8/16/32 *6 Gbps FC, or 2 x 10 Gbps iSCSI	32	240 GB/sec. (16 DIMMs)
F370	288 *3	256	4 FE Modules, each with: 4 x 4/8/16/32 *6 Gbps FC, or 2 x 10 Gbps iSCSI	16	68 GB/sec. (4 DIMMs)
F350	192 *4	128	4 FE Modules, each with: 4 x 4/8/16/32 *6 Gbps FC, or 2 x 10 Gbps iSCSI	16	60 GB/sec. (4 DIMMs)

Table 2: Comparison of the VSP Fxx0 Block Systems

*1 Requires 48 x DBS (DBF max is 576)

*2 Requires 36 x DBS (DBF max is 432)

*3 Requires CBSS and 11 x DBS

*4 Requires CBSS and 7 x DBS

*5 Requires use of the I/O Expansion Box with the Standard Back-end (4 BE Modules)

*6 32Gbps requires a specific SFP. 4/8/16Gbps and 8/16/32Gbps SFPs can be mixed on the same CHB.

*7 32 SAS links in the Standard Back-end (4 BE Modules), 64 SAS links in the Performance Back-end (8 BE Modules)

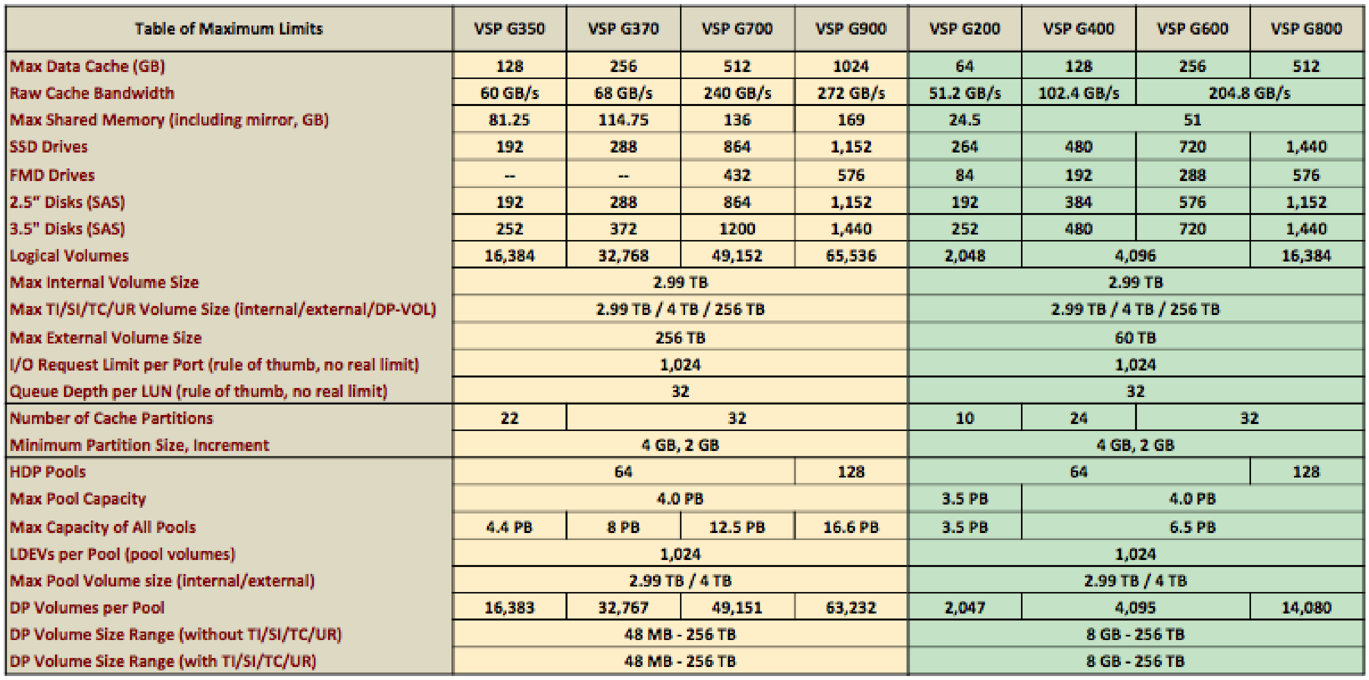
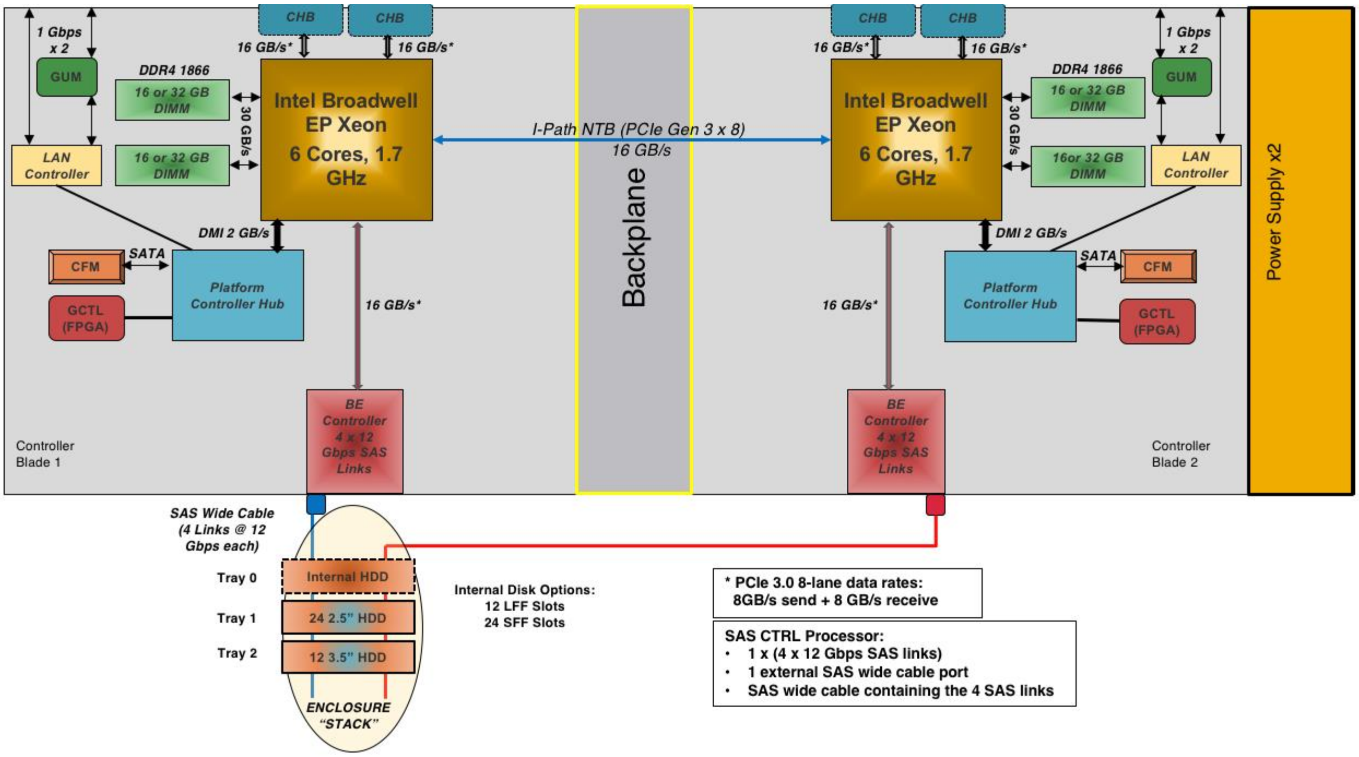


Table of Maximum Limits	VSP G350	VSP G370	VSP G700	VSP G900	VSP G200	VSP G400	VSP G600	VSP G800
Max Data Cache (GB)	128	256	512	1024	64	128	256	512
Raw Cache Bandwidth	60 GB/s	68 GB/s	240 GB/s	272 GB/s	51.2 GB/s	102.4 GB/s	204.8 GB/s	
Max Shared Memory (including mirror, GB)	81.25	114.75	136	169	24.5	51		
SSD Drives	192	288	864	1,152	264	480	720	1,440
FMD Drives	--	--	432	576	84	192	288	576
2.5" Disks (SAS)	192	288	864	1,152	192	384	576	1,152
3.5" Disks (SAS)	252	372	1200	1,440	252	480	720	1,440
Logical Volumes	16,384	32,768	49,152	65,536	2,048	4,096		16,384
Max Internal Volume Size	2.99 TB				2.99 TB			
Max TI/SI/TC/UR Volume Size (internal/external/DP-VOL)	2.99 TB / 4 TB / 256 TB				2.99 TB / 4 TB / 256 TB			
Max External Volume Size	256 TB				60 TB			
I/O Request Limit per Port (rule of thumb, no real limit)	1,024				1,024			
Queue Depth per LUN (rule of thumb, no real limit)	32				32			
Number of Cache Partitions	22	32			10	24	32	
Minimum Partition Size, Increment	4 GB, 2 GB				4 GB, 2 GB			
HDP Pools	64			128	64			128
Max Pool Capacity	4.0 PB				3.5 PB	4.0 PB		
Max Capacity of All Pools	4.4 PB	8 PB	12.5 PB	16.6 PB	3.5 PB	6.5 PB		
LDEVs per Pool (pool volumes)	1,024				1,024			
Max Pool Volume size (internal/external)	2.99 TB / 4 TB				2.99 TB / 4 TB			
DP Volumes per Pool	16,383	32,767	49,151	63,232	2,047	4,095		14,080
DP Volume Size Range (without TI/SI/TC/UR)	48 MB - 256 TB				8 GB - 256 TB			
DP Volume Size Range (with TI/SI/TC/UR)	48 MB - 256 TB				8 GB - 256 TB			

Table 3: Summary of Maximum Limits, VSP Gxx0 and VSP Gx00 Family

VSP G350



Internal Disk Options:
12 LFF Slots
24 SFF Slots

* PCIe 3.0 8-lane data rates:
8GB/s send + 8 GB/s receive

SAS CTRL Processor:
- 1 x (4 x 12 Gbps SAS links)
- 1 external SAS wide cable port
- SAS wide cable containing the 4 SAS links

Figure 1: Block Diagram Overview of the VSP G350 Controller

A fully configured VSP G350 system includes the following:

- A 2U form factor Controller Chassis (DKC) providing:
  - 1 Controller Box with either:
    - 12 x LFF drive slots (CBSL)
    - 24 x SFF drive slots (CBSS)
    - 2 x AC/DC power supplies
  - 2 Controller Blades, each with:
    - 1 x 6-core Intel Broadwell EP Xeon 1.7 GHz processor
    - 64 GB of cache (2 x 32 GB DDR4-1866 DIMMs) for a total of 128 GB of cache in the subsystem
    - 2 FE module slots supporting 1 or 2 FE modules (4 x 4/8/16 Gbps FC, 4 x 8/16/32 Gbps FC, or 2 x 10 Gbps iSCSI each)
    - 1 integrated BE controller (4 x 12 Gbps SAS links), standard or encrypting. Encryption requires a different Controller Blade (CTLSE).
    - 1 x 240GB SSD for cache backup
- Up to 7 DBL, or DBS disk boxes, supporting a maximum of:
  - 96 LFF disks (including 12 LFF disks in CBSL)
  - 192 SFF disks (including 24 SFF disks in CBSS)
- Or up to 4 DB60 dense disk boxes, supporting a maximum of:
  - 252 LFF disks (including 12 LFF disks in CBSL)
- Or an intermix of disk box types, not to exceed a disk box count of 7, where:
  - Each DBL and DBS is counted as one disk box
  - Each DB60 is counted as 2 disk boxes
- Service Processor (SVP--optional) provided by Hitachi Vantara or running on customer supplied server or virtual machine
- DKC Embedded User Interface including the Maintenance Utility, RAID Manager, the REST API, and Hitachi Storage Advisor Embedded (a new GUI for simple provisioning)
- One or more 19” standard racks (Hitachi Vantara supplied or appropriate third party)

The internal drive slots are located at the front of the Controller Chassis. The two Controller Blades are installed in slots at the rear of the Controller Chassis, with two logical boundaries called Cluster 1 (left side) and Cluster 2 (right side). The two clusters share redundant power from the two Power Supply Units (PSUs). The FE Modules are installed into slots on the Controller Blades. The BE Modules are integrated into the Controller Blades themselves. *Figure **2*** shows a rear view of the chassis.

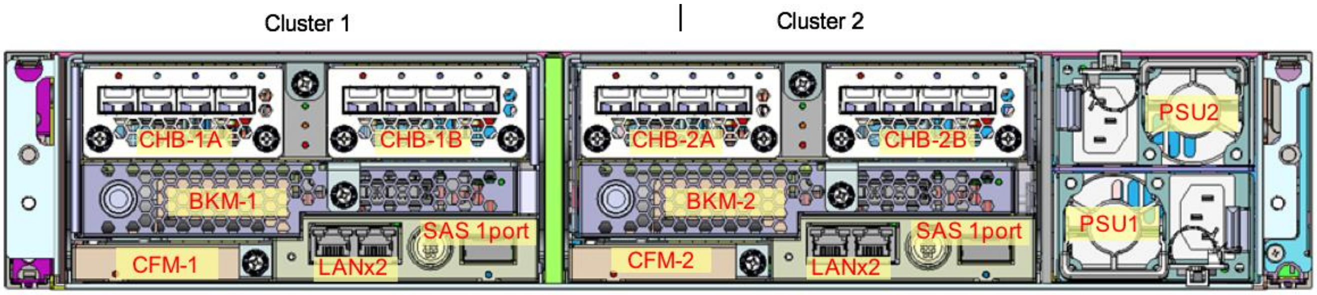


Figure 2: VSP G350 Controller Chassis (DKC) Organization

The Controller Blade slots for the FE Modules are labeled “A-B”, where the A slots represent the default features and the B slots represent the optional features that can be installed. The FE Module port numbers are either “1, 3, 5, 7” for Cluster 1 (odd) or “2, 4, 6, 8” for Cluster 2 (even). The name of a given FE Module comes from the Cluster it is installed in and the slot within the Cluster. For example, the first FE Module (slot A) in Cluster 1 is FE-1A. Likewise, the name for an individual port is the combination of the port number and the FE Module slot. For example, the last port on FE-2A is Port 8A. For FE Modules with only 2 ports, the port numbers are either “1, 3” for Cluster 1 or “2, 4” for Cluster 2.

The VSP G350 back end has a total of 8 x 12 Gbps full duplex SAS links provided by the two integrated BE controllers (one port each). The disk boxes are connected as a single enclosure “stack” with the Controller Box’s internal drives serving as the first disk box (DB-00). Up to seven additional disk boxes can be attached, numbered DB-01 to DB-07. Each half of the stack of disk boxes that share the same SAS port can be considered a “SAS boundary” akin to the power boundaries that define the two DKC clusters. Each dual ported drive is accessible via four SAS links from the BE controller in Cluster 1 and another four SAS links from the BE controller in Cluster 2.

VSP G370

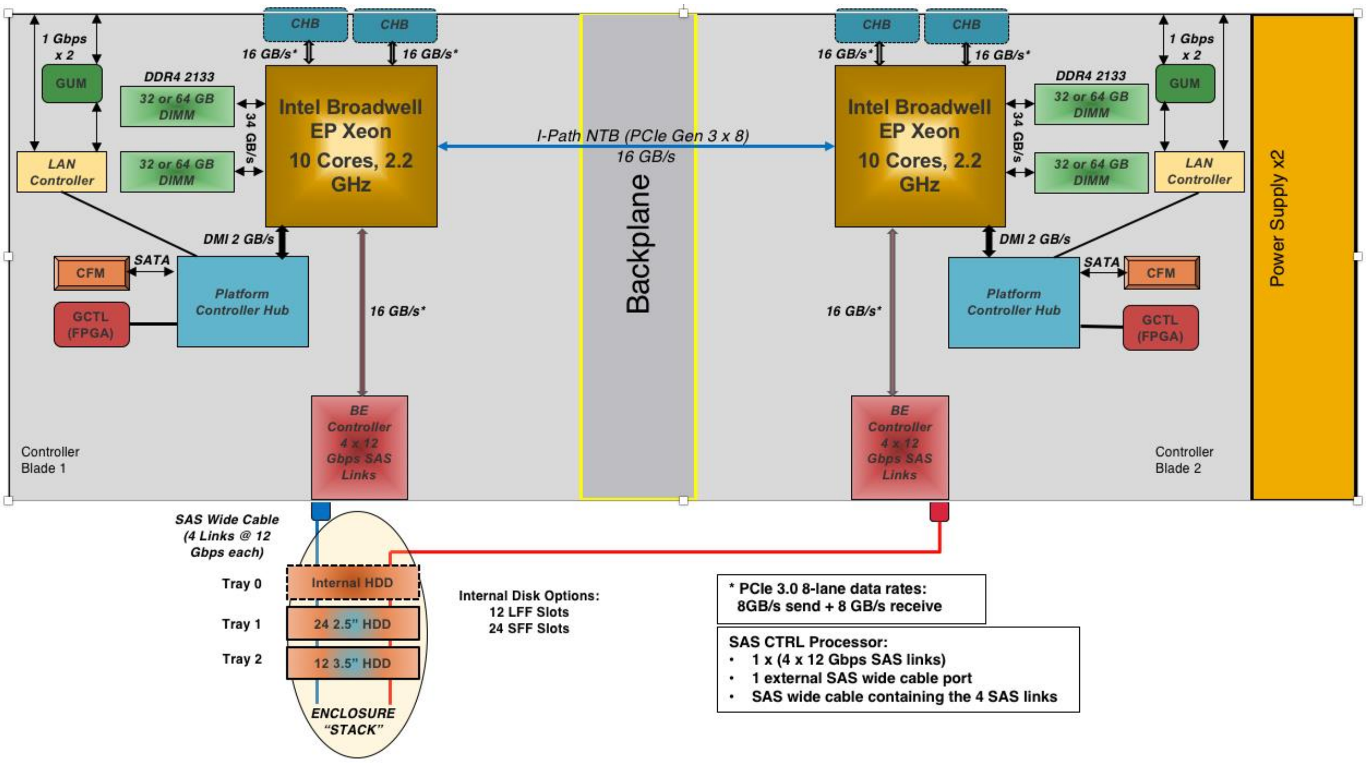


Figure 3: Block Diagram Overview of the VSP G370 Controller

A fully configured VSP G370 system includes the following:

- A 2U form factor Controller Chassis (DKC) providing:
  - 1 Controller Box with either:
    - 12 x LFF drive slots (CBSL)
    - 24 x SFF drive slots (CBSS)
    - 2 x AC/DC power supplies
  - 2 Controller Blades, each with:
    - 1 x 10-core Intel Broadwell EP Xeon 2.2 GHz processor
    - 128 GB of cache (2 x 64 GB DDR4-2133 DIMMs) for a total of 256 GB of cache in the subsystem
    - 2 FE module slots supporting 1 or 2 FE modules (4 x 4/8/16 Gbps FC, 4 x 8/16/32 Gbps FC, or 2 x 10 Gbps iSCSI each)
    - 1 integrated BE controller (4 x 12 Gbps SAS links), standard or encrypting. Encryption requires a different Controller Blade (CTLSE).
    - 1 x 240 GB SSD for cache backup
- Up to 11 DBL, or DBS disk boxes, supporting a maximum of:
  - 144 LFF disks (including 12 LFF disks in CBSL)
  - 288 SFF disks (including 24 SFF disks in CBSS)
- Or up to 6 DB60 dense disk boxes, supporting a maximum of:
  - 372 LFF disks (including 12 LFF disks in CBSL)
- Or an intermix of disk box types, not to exceed a disk box count of 11, where:
  - Each DBL and DBS is counted as one disk box
  - Each DB60 is counted as 2 disk boxes
- Service Processor (SVP--optional) provided by Hitachi Vantara or running on customer supplied server or virtual machine
- DKC Embedded User Interface including the Maintenance Utility, RAID Manager, the REST API, and Hitachi Storage Advisor Embedded (a new GUI for simple provisioning)
- One or more 19” standard racks (Hitachi Vantara supplied or appropriate third party)

The internal drive slots are located at the front of the Controller Chassis. The two Controller Blades are installed in slots at the rear of the Controller Chassis, with two logical boundaries called Cluster 1 (left side) and Cluster 2 (right side). The two clusters share redundant power from the two Power Supply Units (PSUs). The FE Modules are installed into slots on the Controller Blades. The BE Modules are integrated into the Controller Blades themselves. *Figure **4*** shows a rear view of the chassis.

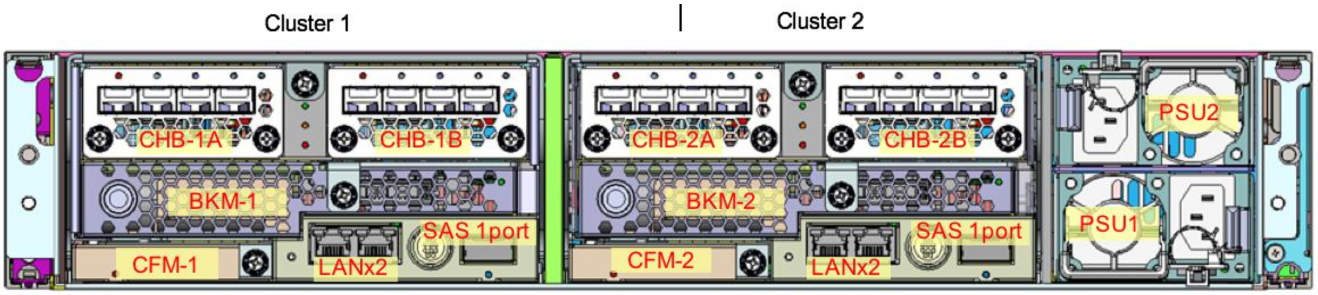


Figure 4: VSP G370 Controller Chassis (DKC) Organization

The Controller Blade slots for the FE Modules are labeled “A-B”, where the A slots represent the default features and the B slots represent the optional features that can be installed. The FE Module port numbers are either “1, 3, 5, 7” for Cluster 1 (odd) or “2, 4, 6, 8” for Cluster 2 (even). The name of a given FE Module comes from the Cluster it is installed in and the slot within the Cluster. For example, the first FE Module (slot A) in Cluster 1 is FE-1A. Likewise, the name for an individual port is the combination of the port number and the FE Module slot. For example, the last port on FE-2A is Port 8A. For FE Modules with only 2 ports, the port numbers are either “1, 3” for Cluster 1 or “2, 4” for Cluster 2.

The VSP G370 back end has a total of 8 x 12 Gbps full duplex SAS links provided by the two integrated BE controllers (one port each). The disk boxes are connected as a single enclosure “stack” with the Controller Box’s internal drives serving as the first disk box (DB-00). Up to eleven additional disk boxes can be attached, numbered DB-01 to DB-11. Each half of the stack of disk boxes that share the same SAS port can be considered a “SAS boundary” akin to the power boundaries that define the two DKC clusters. Each dual ported drive is accessible via four SAS links from the BE controller in Cluster 1 and another four SAS links from the BE controller in Cluster 2.

VSP G700

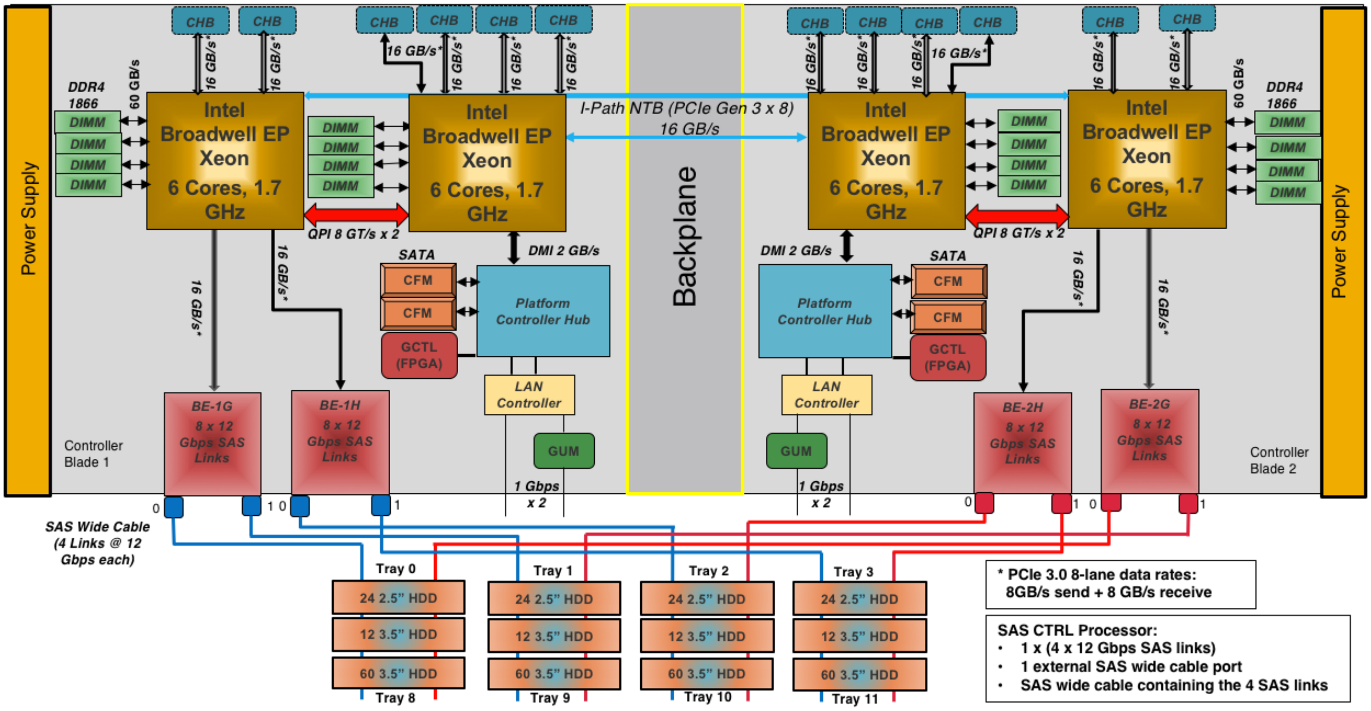

Power Supply
CHB
CHB
CHB
CHB
CHB
CHB
CHB
CHB
CHB
CHB
CHB
CHB
16 GB/s*
16 GB/s*
DDR4 1866
60 GB/s
DIMM
Intel Broadwell EP Xeon 6 Cores, 1.7 GHz
I-Path NTB (PCIe Gen 3 x 8) 16 GB/s
Backplane
QPI 8 GT/s x 2
DMI 2 GB/s
SATA
CFM
GCTL (FPGA)
Platform Controller Hub
LAN Controller
GUM
1 Gbps x 2
BE-1G 8 x 12 Gbps SAS Links
BE-1H 8 x 12 Gbps SAS Links
BE-2H 8 x 12 Gbps SAS Links
BE-2G 8 x 12 Gbps SAS Links
Controller Blade 1
Controller Blade 2
0
1
SAS Wide Cable (4 Links @ 12 Gbps each)
Tray 0
Tray 1
Tray 2
Tray 3
24 2.5" HDD
12 3.5" HDD
60 3.5" HDD
Tray 8
Tray 9
Tray 10
Tray 11
* PCIe 3.0 8-lane data rates: 8GB/s send + 8 GB/s receive
SAS CTRL Processor:
• 1 x (4 x 12 Gbps SAS links)
• 1 external SAS wide cable port
• SAS wide cable containing the 4 SAS links

Figure 5: Block Diagram Overview of the VSP G700 Controller

The VSP G700 marks the transition to a 4U form factor controller chassis having two CPU sockets per controller blade and greater than three times the cache bandwidth of the VSP G3x0 models. In addition to these performance enhancements, an additional 25 disk boxes (or 14 additional DB60s) are supported. Unlike the smaller G350/G370, Flash Module Drives as well as diskless configurations may be configured on the G700.

A fully configured VSP G700 system includes the following:

- A 4U form factor Controller Chassis (DKC) providing:
  - 2 Controller Blades, each with:
    - 2 x 6-core Intel Broadwell EP Xeon 1.7 GHz processor
    - 256 GB of cache (8 x 64 GB DDR4-1866 DIMMs) for a total of 512 GB of cache per subsystem
    - 8 I/O module slots
    - Two 240 GB SSDs for cache backup
- Choice of I/O module configurations
  - 2 to 12 FE modules (4 x 4/8/16 Gbps FC, 4 x 8/16/32 Gbps FC, or 2 x 10 Gbps iSCSI each)
  - 4 BE modules each with 8 x 12 Gbps SAS links (2 ports, 4 x 12 Gbps SAS links per port, 32 links total), standard or encrypting
    - For virtualization configurations without internal disks, the BE modules can be replaced with additional FE modules
- Up to 36 DBL, DBS, or DBF disk boxes, supporting a maximum of:
  - 432 LFF disks
  - 864 SFF disks
  - 432 FMDs
- Or up to 20 DB60 dense disk boxes, supporting a maximum of:
  - 1200 LFF disks
- Or an intermix of disk box types not to exceed a disk box count of ***36***, where:
  - Each DBL, DBS, and DBF is counted as 1 disk box
  - Each DB60 is counted as 1.8 disk boxes
- Service Processor (SVP) provided by Hitachi Vantara or running on customer supplied server or virtual machine
- DKC Embedded User Interface including the Maintenance Utility, RAID Manager, the REST API, and Hitachi Storage Advisor Embedded (a new GUI for simple provisioning)
- One or more 19” standard racks (Hitachi Vantara supplied or appropriate third party)

The Backup Fan Modules (BKMF) and CFM slots are located at the front of the Controller Chassis. The two Controller Blades are installed in slots at the rear of the Controller Chassis, with two logical boundaries called Cluster 1 (bottom) and Cluster 2 (top). The two clusters share redundant power from the two Power Supply Units (PSUs). The FE and BE Modules are installed into slots on the Controller Blades. *Figure* ***6*** shows a rear view of the G700 chassis.

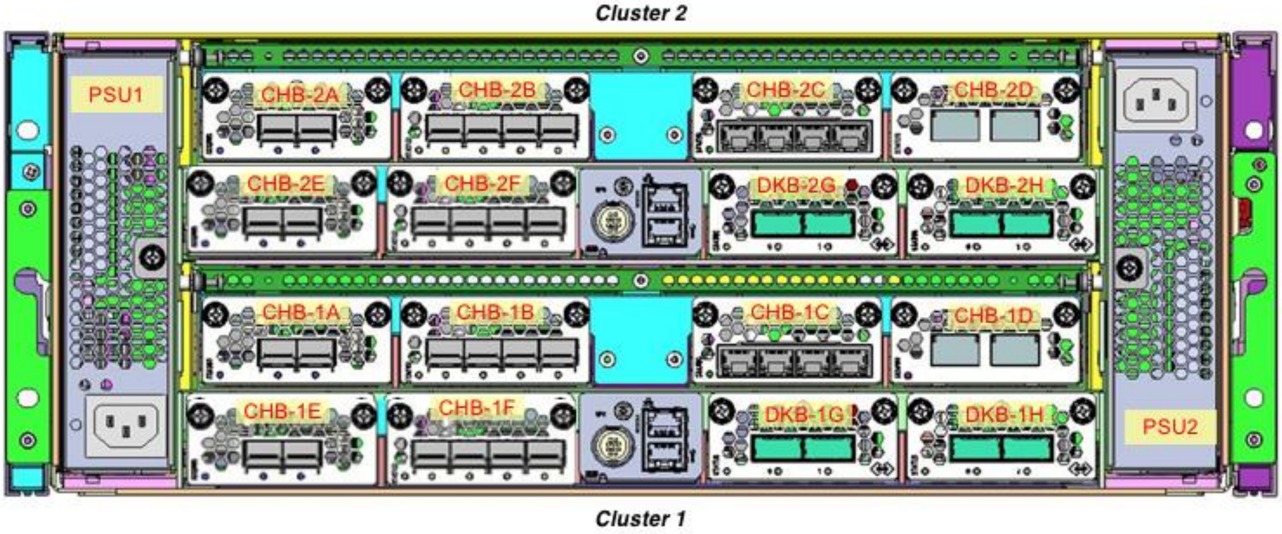


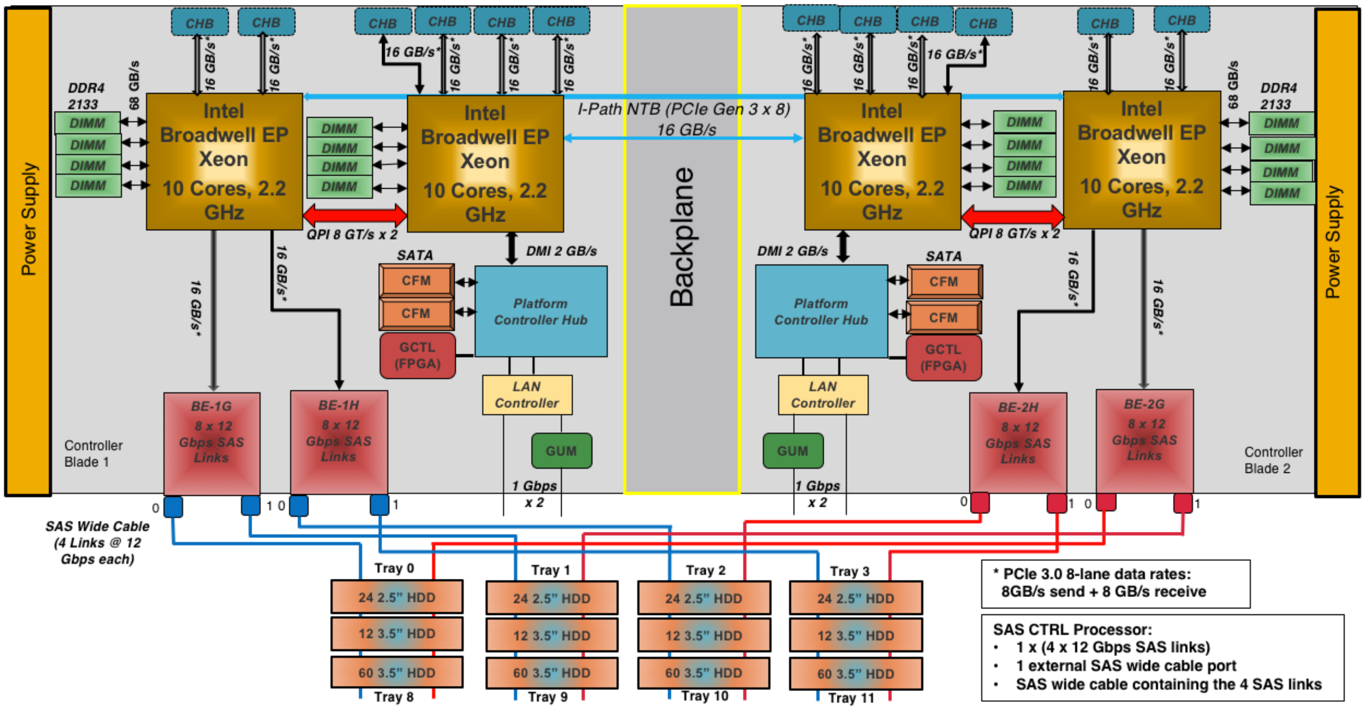
Figure 6: VSP G700 Controller Chassis (DKC) Organization

Pictured in *Figure **6*** is a G700 configuration comprised of 12 FE Modules and 4 BE Modules. The slots for the FE Modules are labeled “A-F”, where the A slots represent the default features and the B-F slots represent the optional features that can be installed. The FE Module port numbers are either “1, 3, 5, 7” for Cluster 1 (odd) or “2, 4, 6, 8” for Cluster 2 (even). The name of a given FE Module comes from the Cluster it is installed in and the slot within the Cluster. For example, the first FE Module (slot A) in Cluster 1 is FE-1A. Likewise, the name for an individual port is the combination of the port number and the FE Module slot. For example, the last port on FE-2F is Port 8F. For FE Modules with only 2 ports, the port numbers are either “1, 3” for Cluster 1 or “2, 4” for Cluster 2.

The slots for the BE Modules are labeled “G, H” and the SAS port numbers are “0, 1”. For a diskless system dedicated to virtualization use, the four BE Modules can be replaced by four more FE Modules to provide a total of 64 FC or 32 iSCSI ports.

Each G700 enclosure “stack” can support a maximum of 360 drives when used with the DB60 disk boxes. The supported number of disk boxes and drives will vary depending on the back-end configuration and the type of disk box selected. Refer to the maximum drive counts presented in *Table **4*** for the breakdown by Gxx0 model and disk box type.

VSP G900



* PCIe 3.0 8-lane data rates: 8GB/s send + 8 GB/s receive

SAS CTRL Processor:
- 1 x (4 x 12 Gbps SAS links)
- 1 external SAS wide cable port
- SAS wide cable containing the 4 SAS links

Figure 7: Block Diagram Overview of the VSP G900 Controller

The VSP G900 is a more scalable version of the G700, both in terms of number of supported drives and performance. It shares a common DKC with the G700 but utilizes different Controller Blades featuring Intel Xeon processors having four additional cores per socket, and 23% faster clock speed than the G700 CPUs. Maximum raw cache bandwidth rises to 272 GB/s with the standard 2133 GHz DDR4 DIMMs. Optionally, additional BE Modules can be installed to double back-end bandwidth and increase the number of supported DBS, DBL, or DBF disk boxes.

A fully configured VSP G900 system includes the following:

- A 4U form factor Controller Chassis (DKC) providing:
  - 2 Controller Blades, each with:
    - 2 x 10-core Intel Broadwell EP Xeon 2.2 GHz processors
    - 256 GB of cache (8 x 32 GB or 4 x 64 GB DDR4-2133 DIMMs) for a total of 512 GB of cache per subsystem, or 512 GB of cache (8 x 64 GB DDR4-2133 DIMMs) for a total of 1024 GB of cache per subsystem
    - 8 I/O module slots
    - 2 480 GB SSDs for cache backup
- Choice of I/O module configurations
  - Standard Back-end (shown in *Figure* **7**) with 4 BE modules (each with 2 ports, 4 x 12 Gbps SAS links per port, 32 links total):
    - 2 to 12 FE modules without I/O Expansion Box
    - 2 to 16 FE modules with I/O Expansion Box
  - Performance Back-end with 8 BE modules (each with 2 ports, 4 x 12 Gbps SAS links per port, 64 links total):

        - 2 to 8 FE modules without I/O Expansion Box
        - 2 to 12 FE modules with I/O Expansion Box
    - For virtualization configurations without internal disks:
        - 2 to 16 FE modules without I/O Expansion Box
        - 2 to 20 FE modules with I/O Expansion Box
- Disk box configurations with Standard Back-end (4 BE Modules)
    - Up to 24 DBL, DBS, or DBF disk boxes, supporting a maximum of:
        - 288 LFF disks
        - 576 SFF disks
        - 288 FMDs
    - Or up to 24 DB60 dense disk boxes, supporting a maximum of:
        - 1,440 LFF disks
    - Or an intermix of disk box types not to exceed a disk box count of 24, where:
        - Each disk box regardless of type (DBL, DBS, DBF, DB60) is counted as 1 disk box
- Disk box configurations with Performance Back-end (8 BE Modules)
    - Up to 48 DBL, DBS, or DBF disk boxes, supporting a maximum of:
        - 576 LFF disks
        - 1,152 SFF disks
        - 576 FMDs
    - Or up to 24 DB60 dense disk boxes, supporting a maximum of:
        - 1,440 LFF disks
    - Or an intermix of disk box types not to exceed a disk box count of 48, where:
        - Each DBL, DBS, and DBF is counted as 1 disk box
        - Each DB60 is counted as 2 disk boxes
- Service Processor (SVP) provided by Hitachi Vantara or running on customer supplied server or virtual machine
- DKC Embedded User Interface including the Maintenance Utility, RAID Manager, the REST API, and Hitachi Storage Advisor Embedded (a new GUI for simple provisioning)
- One or more 19” standard racks (Hitachi Vantara supplied or appropriate third party)

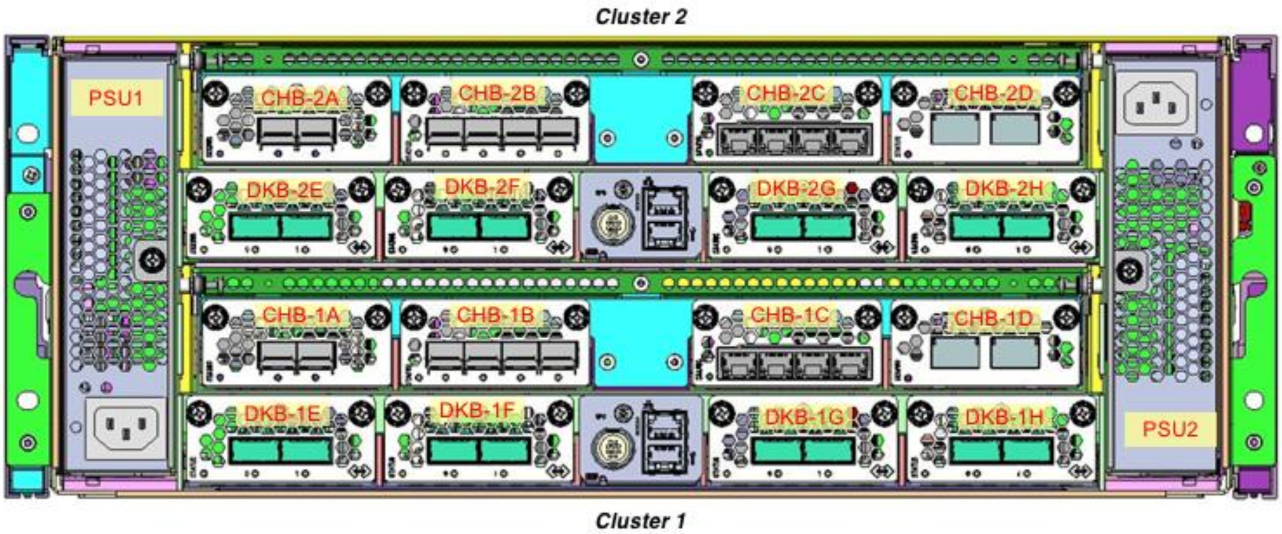


Figure 8: VSP G900 Controller Chassis (DKC) Organization

The VSP G900 offers the most flexibility when it comes to front-end and back-end configuration. The option to install 4 or 8 BE Modules can be combined with the Front-End I/O Expansion Box for a wide variety of supported I/O configurations.

Figure ***8*** shows a G900 configuration with the high performance back-end comprised of 8 BE Modules. The slots for the FE Modules are labeled “A-D” where the A slots represent the default features and the B-D slots represent the optional features that can be installed. The FE Module port numbers are either “1, 3, 5, 7” for Cluster 1 (odd) or “2, 4, 6, 8” for Cluster 2 (even). The name of a given FE Module comes from the Cluster it is installed in and the slot within the Cluster. For example, the first FE Module (slot A) in Cluster 1 is FE-1A. Likewise, the name for an individual port is the combination of the port number and the FE Module slot. For example, the last port on FE-2D is Port 8D. For FE Modules with only 2 ports, the port numbers are either “1, 3” for Cluster 1 or “2, 4” for Cluster 2.

The slots for the BE Modules are labeled “E-H” and the SAS port numbers are “0, 1”. For a diskless system dedicated to virtualization use, the eight BE Modules can be replaced by eight more FE Modules to provide a total of 64 FC or 32 iSCSI ports.

Each G900 enclosure “stack” can support a maximum of 360 drives when used with the DB60 disk boxes. The supported number of disk boxes and drives will vary depending on the back-end configuration and the type of disk box selected. Refer to the maximum drive counts presented in *Table* ***4*** for the breakdown by Gxx0 model and disk box type.

The G900 configuration with the standard back-end comprised of 4 BE Modules has the same layout as the G700 configuration shown in *Figure* ***6***.

Figure ***9*** shows a diskless configuration with the 2U Front-end I/O Expansion Box installed. PCI Express Pass-Through Modules (PTM) are installed in the “E, F” slots and connected via external cables to the PCIe Cable Connecting Packages (PCP) in the expansion box. Note that the PTMs are shown installed in the “E, F” slots but they can also be installed in the “A, B” or “C, D” slots. As shown, the slots for the FE Modules in the DKC are labeled “A-D” and “G-H”, and the slots for the FE Modules in the I/O Expansion Box are labeled “J-M”, providing a combined total of up to 80 FC or 40 iSCSI ports. Refer to the Front-end I/O Expansion Box section for more details on the hardware design and considerations when installing FE Modules in the expansion box.

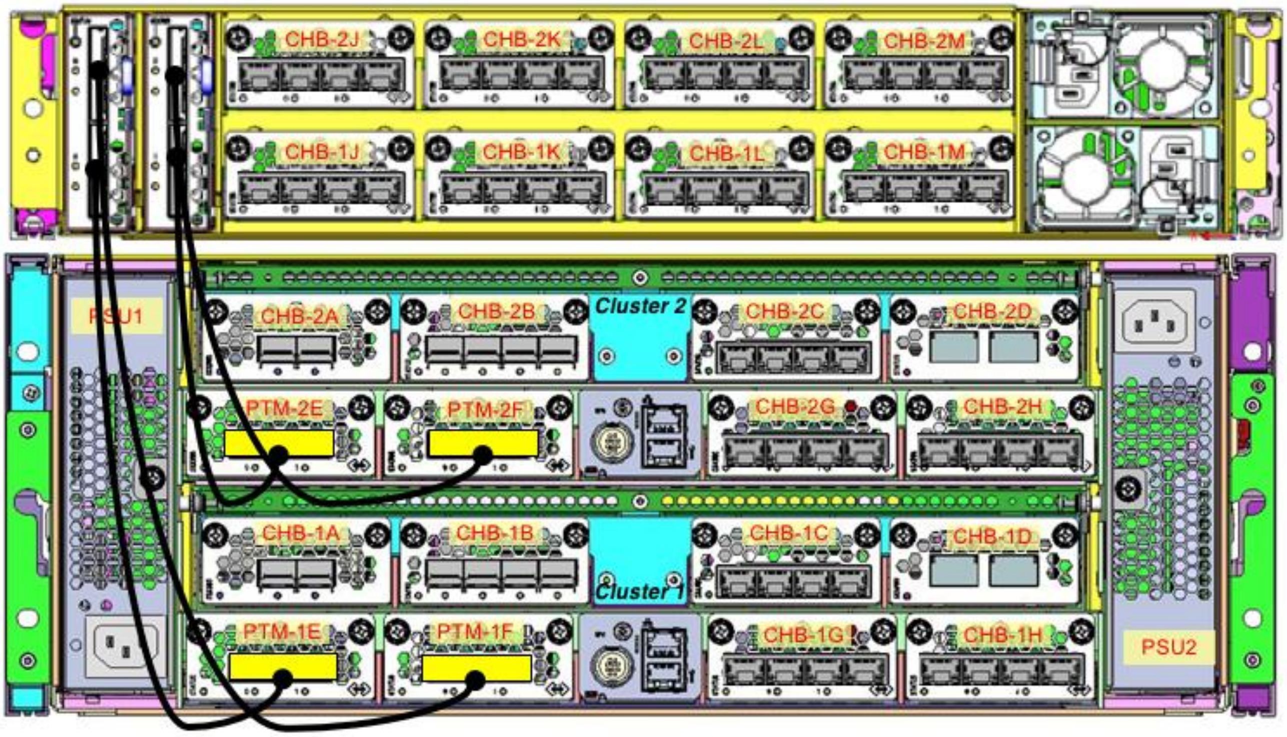


Figure 9: VSP G900 Controller Chassis (DKC) Organization with I/O Expansion Box -- Diskless

VSP F350 / F370 / F700 / F900

The VSP F350, F370, F700, and F900 are All Flash Array (AFA) variants of the corresponding G series models. Hardware and system architecture are unchanged while the number of supported system configurations is simplified.

The VSP F350 consists of:

- The same Controller Chassis and Controller Blades as VSP G350
- One Controller Box with 24 x SFF drive slots (CBSS)
- The same cache and FE Module options as the VSP G350
- One integrated BE controller per Controller Blade (4 x 12 Gbps SAS links), standard or encrypting. Encryption requires a different Controller Blade (CTLSE).
- Up to 7 DBS disk boxes, supporting a maximum of 192 SFF SSDs (including 24 SFF SSDs in CBSS)

The VSP F370 includes:

- The same Controller Chassis and Controller Blades as VSP G370
- One Controller Box with 24 x SFF drive slots (CBSS)
- The same cache and FE Module options as the VSP G370
- One integrated BE controller per Controller Blade (4 x 12 Gbps SAS links), standard or encrypting. Encryption requires a different Controller Blade (CTLSE).
- Up to 11 DBS disk boxes, supporting a maximum of 288 SFF SSDs (including 24 SFF SSDs in CBSS)

The VSP F700 consists of:

- The same Controller Chassis, Controller Blades and cache options as the VSP G700
- 2 to 12 FE modules (4 x 4/8/16 Gbps FC, 4 x 8/16/32 Gbps FC, or 2 x 10 Gbps iSCSI each)
- 4 BE modules each with 8 x 12 Gbps SAS links (2 ports, 4 x 12 Gbps SAS links per port, 32 links total), standard or encrypting
- Up to 36 DBS or DBF disk boxes, supporting a maximum of:
  - 864 SFF SSDs
  - 432 FMDs

The F900 options include:

- The same Controller Chassis, Controller Blades, and cache options as the VSP G900
- Choice of I/O module configurations
  - Standard Back-end (shown in *Figure 7*) with 4 BE modules (each with 2 ports, 4 x 12 Gbps SAS links per port, 32 links total):
    - 2 to 12 FE modules without I/O Expansion Box
    - 2 to 16 FE modules with I/O Expansion Box
  - Performance Back-end (see *Figure 8*) with 8 BE modules (each with 2 ports, 4 x 12 Gbps SAS links per port, 64 links total):
    - 2 to 8 FE modules without I/O Expansion Box
    - 2 to 12 FE modules with I/O Expansion Box
- Disk box configurations with Standard Back-end (4 BE Modules)
  - Up to 24 DBS, or DBF disk boxes, supporting a maximum of:
    - 576 SFF SSDs
    - 288 FMDs
  - Or an intermix of disk box types not to exceed a disk box count of 24, where:
    - Each disk box regardless of type (DBS, DBF) is counted as 1 disk box
- Disk box configurations with Performance Back-end (8 BE Modules)
  - Up to 48 DBS, or DBF disk boxes, supporting a maximum of:
    - 1,152 SFF SSDs
    - 576 FMDs
  - Or an intermix of disk box types not to exceed a disk box count of 48, where:
    - Each DBS and DBF is counted as 1 disk box

Blade and Module Details

Overview

This section describes the Controller Blades as well as the FE and BE Modules in detail. The VSP Gxx0 family uses three types of blades and installable modules in the controller chassis (DKC):

- Controller Blades (CTL) with the MPU logical processors, cache, data and control paths, external interface slots and system management interfaces
  - There are four types of Controller Blades (G350, G370, G700, and G900) and each will be described in detail separately
- Front-end Connectivity modules (FE) with 4 x 4/8/16 Gbps FC, 4 x 8/16/32 Gbps FC, or 2 x 10 Gbps iSCSI ports
- Back-end Drive Controller modules (BE) with 8 x 12 Gbps SAS links in two SAS wide ports
  - The G350 and G370 do not support pluggable BE modules and the integrated BE controller in each G350/G370 Controller Blade provides 4 x 12 Gbps SAS links in one SAS wide port

The two Controller Blades are the core of the system architecture, with the Intel Xeon processors organized into MPU logical units that execute the system software, manage all I/O, and emulate all the specialized functions that were done previously by custom ASICs (DCTL ASIC in HUS 100 family, HM ASIC in HUS VM, DA and DRR ASICs in VSP G1500). FE modules are plugged into slots directly on the G350 and G370 Controller Blades, while FE and BE modules are plugged into slots at the rear of the G700 and G900 DKC that are connected to an individual Controller Blade. These modules are extensions to the Controller Blade, not an independent unit like the autonomous FED and BED boards on the Grid on a VSP G1500.

The FE modules are based on the Hilda or Baker chips, which are powerful processors with significant independent functionality (described in detail later). Similarly, the BE modules are based on a powerful dual-core SAS Protocol Controller (SPC) which also functions mostly on its own. The FE and SAS processors communicate with their Controller Blade's Local Routers (LR). This is their command transfer circuit to the MPU logical processors and the method by which host I/O requests get scheduled.

Each FE or SAS processor also has several DMA channels built-in. These are used to directly access the Data Transfer Buffer (DXBF) and User Data regions of cache (described later in this paper). Access to the User Data regions first requires assignment of a cache address from the MPU that owns the LDEV in question.

Controller Blade Overview

The two Controller Blades form the core of the VSP Gxx0 system. All I/O processing, cache management, and data transfer is performed on these boards. While each Controller Blade is functionally equivalent across the entire model range, the physical boards themselves are different between the G350, G370, G700, and G900.

Controller Blade (G350)

Central to the G350 Controller Blade is a 6-core 1.7 GHz Intel Broadwell EP Xeon processor. In the VSP Gxx0 design, this CPU provides nearly all of the system functionality, from executing the system software (SVOS), providing the PCIe switch paths, and emulating the functions that were done previously by custom ASICs. The functions provided by the Intel Xeon processor include:

- Microprocessor Units (MPU)
- PCI Express Switch Paths
  - Two PCIe 3.0 x8 links for FE Modules

    - One PCIe 3.0 x8 link to the embedded Back-end SAS controller
    - One PCIe 3.0 x8 link (I-Path) to cross connect to the CPU on the other Controller Blade. The I-Path is also referred to as the Non-Transparent Bridge (NTB) path.
- Dual-channel memory controller attached to two DDR4-1866 DIMM slots, with 15 GB/s of bandwidth per channel and 30 GB/s of total cache bandwidth.
- DMA channels to allow the PCIe attached FE and BE Modules to access cache managed by the on-die memory controller, or to allow the MPUs to access the Cache Directory or Shared Memory.
- ASIC emulation is performed in microcode to provide a consistent interface for the SVOS system software. The ASIC functions that are emulated include:
    - Local Router (LR): The sole function of the LR is to transfer commands between the FE and BE Modules and MPUs.
    - Data Recovery and Reconstruction (DRR): The primary function of the DRR is to perform RAID parity operations, but it is also responsible for drive formatting and rebuilds (correction copy).
    - Direct Memory Access (DMA): This DMA function applies only when user data must be transferred between Controller Blades across an I-Path. An MPU on the source cluster manages the first step of the data transfer and an MPU on the destination cluster manages the second step of the data transfer.

The embedded Back-end SAS controller is similar to the BE Module used in the G700 and G900 models. Here, the SPCv 12G processor also provides 8 x 12 Gbps SAS links, but four of these links are connected to an external SAS 4-Wide port and the other four links are connected to the internal SAS expander that's part of the internal drive box.

Cache memory is installed into two DDR4-1866 DIMM slots on each Controller Blade, which are attached to the on-die memory controller in the Intel Xeon CPU and organized as two independent memory channels. Each channel has a peak theoretical transfer rate of 15 GB/s, so the entire cache system has a peak rating of 60 GB/s. The supported cache size for each blade is 64 GB (2 x 32 GB DIMMs).

Note that the cache memory on each of the two Controller Blades is concatenated together into one larger global cache image. For "clean" data in cache, meaning data that is a copy of what is already on disk, data that is kept in cache to serve possible future read hits, only one copy of the clean data is kept in the global space. Thus clean data is only kept on one Controller Blade's cache memory and is not mirrored across both Controller Blades. Only "dirty" data, meaning data recently written by the host that has not yet been written to disk, is duplexed with a copy of the data being retained in each of the two Controller Blades.

The rest of the auxiliary system management functions are provided via the Platform Controller Hub (PCH) chip. The PCH is connected to the Intel Xeon CPU via a DMI 2.0 connection, which is electrically comparable to a PCI Express 2.0 x4 link. The PCH itself has a SATA 6 Gbps controller built in that interfaces with a 240 GB Cache Flash Module (CFM). The CFM is a normal SATA SSD that is used for backing up the entire contents of cache in the event of a total loss of power. If there is a partial loss of power to just one cluster, this is the backup target for that cluster's cache space. In the case of a planned power off, it is the backup target for just the Shared Memory region. During a power outage, the on-blade battery power keeps the DIMMs, CFM, and Controller Blade functioning while destage occurs to the flash drive. There is generally enough battery power to support a couple such outages back-to-back without recharging.

The PCH has a PCIe connection to an FPGA (Field Programmable Gate Array) processor that is responsible for environmental monitoring and processing component failures. The FPGA relies on an environment microcontroller that has monitoring connections to each of the components on the Controller Blade (FE and BE Modules, Power Supply Units, Fans, Battery, etc.) as well as an interface to the other Controller Blade in the opposite cluster.

The PCH also has connections to the Gateway for Unified Management (GUM) and a pair of network interface controllers (LAN Controllers). The GUM is an embedded micro server that provides the management interface that the storage management software running on the SVP talks to. The LAN Controllers are what provide the public and management network ports (gigabit Ethernet) on each Controller Blade.

Controller Blade (G370)

The G370 Controller Blade is a variation of the G350 Controller Blade, featuring a 10-core 2.2 GHz Intel Broadwell EP Xeon processor. In the VSP Gxx0 design, this CPU provides nearly all of the system functionality, from executing the system software (SVOS), providing the PCIe switch paths, and emulating the functions that were done previously by custom ASICs. The functions provided by the Intel Xeon processor include:

- Microprocessor Units (MPU)
- PCI Express Switch Paths
  - Two PCIe 3.0 x8 links for FE Modules
  - One PCIe 3.0 x8 link to the embedded Back-end SAS controller
  - One PCIe 3.0 x8 link (I-Path) to cross connect to the CPU on the other Controller Blade. The I-Path is also referred to as the Non-Transparent Bridge (NTB) path.
- Dual-channel memory controller attached to two DDR4-2133 DIMM slots, with 17 GB/s of bandwidth per channel and 34 GB/s of total cache bandwidth.
- DMA channels to allow the PCIe attached FE and BE Modules to access cache managed by the on-die memory controller, or to allow the MPUs to access the Cache Directory or Shared Memory.
- ASIC emulation is performed in microcode to provide a consistent interface for the SVOS system software. The ASIC functions that are emulated include:
  - Local Router (LR): The sole function of the LR is to transfer commands between the FE and BE Modules and MPUs.
  - Data Recovery and Reconstruction (DRR): The primary function of the DRR is to perform RAID parity operations, but it is also responsible for drive formatting and rebuilds (correction copy).
  - Direct Memory Access (DMA): This DMA function applies only when user data must be transferred between Controller Blades across an I-Path. An MPU on the source cluster manages the first step of the data transfer and an MPU on the destination cluster manages the second step of the data transfer.

The embedded Back-end SAS controller is similar to the BE Module used in the G700 and G900 models. Here, the SPCv 12G processor also provides 8 x 12 Gbps SAS links, but four of these links are connected to an external SAS 4-Wide port and the other four links are connected to the internal SAS expander that's part of the internal drive box.

Cache memory is installed into two DDR4-2133 DIMM slots on each Controller Blade, which are attached to the on-die memory controller in the Intel Xeon CPU and organized as two independent memory channels. Each channel has a peak theoretical transfer rate of 17 GB/s, so the entire cache system has a peak rating of 68 GB/s. The supported cache size for each blade is 128 GB (2 x 64 GB DIMMs).

Note that the cache memory on each of the two Controller Blades is concatenated together into one larger global cache image. For "clean" data in cache, meaning data that is a copy of what is already on disk, data that is kept in cache to serve possible future read hits, only one copy of the clean data is kept in the global space. Thus clean data is only kept on one Controller Blade's cache memory and is not mirrored across both Controller Blades. Only "dirty" data, meaning data recently written by the host that has not yet been written to disk, is duplexed with a copy of the data being retained in each of the two Controller Blades.

The rest of the auxiliary system management functions are provided via the Platform Controller Hub (PCH) chip. The PCH is connected to the Intel Xeon CPU via a DMI 2.0 connection, which is electrically comparable to a PCI Express 2.0 x4 link. The PCH itself has a SATA 6 Gbps controller built in that interfaces with a 240 GB Cache

Flash Module (CFM). The CFM is a normal SATA SSD that is used for backing up the entire contents of cache in the event of a total loss of power. If there is a partial loss of power to just one cluster, this is the backup target for that cluster's cache space. In the case of a planned power off, it is the backup target for just the Shared Memory region. During a power outage, the on-blade battery power keeps the DIMMs, CFM, and Controller Blade functioning while destage occurs to the flash drive. There is generally enough battery power to support a couple such outages back-to-back without recharging.

The PCH has a PCIe connection to an FPGA (Field Programmable Gate Array) processor that is responsible for environmental monitoring and processing component failures. The FPGA relies on an environment microcontroller that has monitoring connections to each of the components on the Controller Blade (FE and BE Modules, Power Supply Units, Fans, Battery, etc.) as well as an interface to the other Controller Blade in the opposite cluster.

The PCH also has connections to the Gateway for Unified Management (GUM) and a pair of network interface controllers (LAN Controllers). The GUM is an embedded micro server that provides the management interface that the storage management software running on the SVP talks to. The LAN Controllers are what provide the public and management network ports (gigabit Ethernet) on each Controller Blade.

Controller Blade (G700)

Central to the G700 Controller Blade is a pair of 6-core 1.7 GHz Intel Broadwell EP Xeon processors. In the VSP Gxx0 design, these CPUs provide nearly all of the system functionality, from executing the system software (SVOS), providing the PCIe switch paths, and emulating the functions that were done previously by custom ASICs. The functions provided by each Intel Xeon processor include:

- Microprocessor Units (MPU)
- PCI Express Switch Paths
  - Four PCIe 3.0 x8 links for I/O Modules
    - 4 FE Modules attached to the first CPU on each Controller Blade
    - 2 FE Modules and 2 BE Modules attached to the second CPU on each Controller Blade
  - Two PCIe 3.0 x8 links (I-Paths) to cross connect to a CPU on the other Controller Blade. The I-Paths are also referred to as the Non-Transparent Bridge (NTB) paths.
- Quad-channel memory controller attached to four DDR4-1866 DIMM slots
  - 15 GB/s of bandwidth per channel, 60 GB/s of cache bandwidth per CPU, and 120 GB/s of total cache bandwidth for the Controller Blade.
- DMA channels to allow the PCIe attached FE and BE Modules to access cache managed by the on-die memory controllers, or to allow the MPUs to access the Cache Directory or Shared Memory.
- ASIC emulation is performed in microcode to provide a consistent interface for the SVOS system software. The ASIC functions that are emulated on each MPU include:
  - Local Router (LR): The sole function of the LR is to transfer commands between the FE and BE Modules and MPUs.
  - Data Recovery and Reconstruction (DRR): The primary function of the DRR is to perform RAID parity operations, but it is also responsible for drive formatting and rebuilds (correction copy).
  - Direct Memory Access (DMA): This DMA function applies only when user data must be transferred between Controller Blades across an I-Path. An MPU on the source cluster manages the first step of the data transfer and an MPU on the destination cluster manages the second step of the data transfer.

Cache memory can be installed into half (Basic features) or all (Basic + Optional features) of the DDR4-1866 DIMM slots. However, Hitachi Vantara only supports the following combination of cache size and bandwidth:

- 256 GB (8 x 32 GB DIMMs): 120 GB/s cache bandwidth

The entire cache system of the G700 with all DIMMs populated has a peak rating of 240 GB/s (120 GB/s per Controller Blade).

The rest of the auxiliary system management functions are provided via the Platform Controller Hub (PCH) chip. The PCH is connected to the Intel Xeon CPU via a DMI 2.0 connection, which is electrically comparable to a PCI Express 2.0 x4 link. The PCH itself has a SATA 6 Gbps controller built in that interfaces with two 240 GB Cache Flash Modules (CFM). The CFM is a normal SATA SSD that is used for backing up the entire contents of cache in the event of a total loss of power. If there is a partial loss of power to just one cluster, this is the backup target for that cluster's cache space. In the case of a planned power off, it is the backup target for just the Shared Memory region. During a power outage, the on-blade battery power keeps the DIMMs, CFM, and Controller Blade functioning while destage occurs to the flash drive. There is generally enough battery power to support a couple such outages back-to-back without recharging. The G700 may require one or two CFMs per Controller Blade, depending on the number and type of DIMMs installed.

The remaining PCH functions are the same as in the G370 described earlier.

Controller Blade (G900)

The G900 Controller Blade is a more powerful variant of the G700 Controller Blade, featuring a pair of 10-core 2.2 GHz Intel Broadwell EP Xeon processors. In the VSP Gxx0 design, these CPUs provide nearly all of the system functionality, from executing the system software (SVOS), providing the PCIe switch paths, and emulating the functions that were done previously by custom ASICs. The functions provided by each Intel Xeon processor include:

- Microprocessor Units (MPU)
- PCI Express Switch Paths
  - Four PCIe 3.0 x8 links for I/O Modules
    - 4 FE Modules attached to the first CPU on each Controller Blade
    - 2 FE Modules and 2 BE Modules (or 4 BE Modules) attached to the second CPU on each Controller Blade
  - Two PCIe 3.0 x8 links (I-Paths) to cross connect to a CPU on the other Controller Blade. The I-Paths are also referred to as the Non-Transparent Bridge (NTB) paths.
- Quad-channel memory controller attached to four DDR4-2133 DIMM slots
  - 17 GB/s of bandwidth per channel, 68 GB/s of cache bandwidth per CPU, and 136 GB/s of total cache bandwidth for the Controller Blade.
- DMA channels to allow the PCIe attached FE and BE Modules to access cache managed by the on-die memory controllers, or to allow the MPUs to access the Cache Directory or Shared Memory.
- ASIC emulation is performed in microcode to provide a consistent interface for the SVOS system software. The ASIC functions that are emulated on each MPU include:
  - Local Router (LR): The sole function of the LR is to transfer commands between the FE and BE Modules and MPUs.
  - Data Recovery and Reconstruction (DRR): The primary function of the DRR is to perform RAID parity operations, but it is also responsible for drive formatting and rebuilds (correction copy).
  - Direct Memory Access (DMA): This DMA function applies only when user data must be transferred between Controller Blades across an I-Path. An MPU on the source cluster manages the first step of the data transfer and an MPU on the destination cluster manages the second step of the data transfer.

Cache memory can be installed into half (Basic features) or all (Basic + Optional features) of the DDR4-2133 DIMM slots. The resulting combinations of cache sizes and bandwidth per blade are:

- 256 GB (8 x 32 GB DIMMs): 136 GB/s cache bandwidth
- 256 GB (4 x 64 GB DIMMs): 68 GB/s cache bandwidth
- 512 GB (8 x 64 GB DIMMs): 136 GB/s cache bandwidth

The entire cache system of the G900 with all DIMMs populated has a peak rating of 272 GB/s (136 GB/s per Controller Blade).

The rest of the auxiliary system management functions are provided via the Platform Controller Hub (PCH) chip. The PCH is connected to the Intel Xeon CPU via a DMI 2.0 connection, which is electrically comparable to a PCI Express 2.0 x4 link. The PCH itself has a SATA 6 Gbps controller built in that interfaces with two 240 GB Cache Flash Modules (CFM), or one or two 480 GB CFM. The CFM is a normal SATA SSD that is used for backing up the entire contents of cache in the event of a total loss of power. If there is a partial loss of power to just one cluster, this is the backup target for that cluster’s cache space. In the case of a planned power off, it is the backup target for just the Shared Memory region. During a power outage, the on-blade battery power keeps the DIMMs, CFM, and Controller Blade functioning while destage occurs to the flash drive. There is generally enough battery power to support a couple such outages back-to-back without recharging. The G900 may require one or two CFMs per Controller Blade, depending on the number and type of DIMMs installed.

The remaining PCH functions are the same as in the G370 described earlier.

LDEV Ownership by MPU

Each of the available MPUs is assigned a specific set of LDEVs to manage. Refer to *Table **3*** for the system limit on the number of LDEVs that may be established by model. The single MPU on Cluster 1 is named MPU-10. The MPU on Cluster 2 is named MPU-20. When new LDEVs are created, they are round-robin assigned across the MPUs using a pattern of MPU-10, MPU-20.

The individual LDEV associations to MPU can be looked up and manually changed either by Storage Navigator or by a script that uses CLI commands from the ***raidcom*** utility. There is no automatic load balancing mechanism to move “hot” LDEVs around between the MPUs in order to even out the processing loads. It is not necessary to keep every LDEV from the same Parity Group assigned to the same MPU, just as for VSP G1500.

An MPU will accept all I/O requests for an LDEV it owns without regard for which FE port received that host request, or which BE modules will perform the physical disk operations. Each Cluster contains a local copy of the LDEV-to-MPU mapping tables so that the LRs can look up which MPU owns which LDEV. As a rule of thumb, the MPs in an MPU should be kept below 80% busy to manage host latencies, and only 40% busy if providing for processor headroom to maintain host performance in case of a failure of another MPU. Note: higher peak utilization during batch operations is fine as long as the average utilization over the time remaining in the batch window is below 40%.

I/O Module Overview

The Front-end connectivity modules provide the fibre channel ports for connections to hosts, external storage or remote copy connections (Hitachi Universal Replicator, TrueCopy Sync, Global Active Device). These modules are installed into module slots on the VSP G350/G370 Controller Blade or slots in the rear of the G700/G900 DKC that are connected to a specific Controller Blade. Each G350 or G370 Controller Blade can have two FE modules installed. Each G700 or G900 Controller Blade can have 8 I/O modules installed, in various combinations of FE and BE modules, and the I/O Expansion Box for the G900 model. There are no newly developed I/O Modules for VSP Gxx0. Each of the modules supported on VSP Gxx0 are interchangeable with the original VSP Gx00 family.

The FE module is fairly simple, primarily having a host interface processor on the board. The slot that the FE module plugs into provides power and a PCI Express 3.0 x8 link to an Intel Broadwell EP Xeon processor on the Controller Blade. All interaction with an FE module is via the LR emulator running on the MPU in the same Controller Blade. The host interface processor has four DMA channels for moving data blocks into or out of cache.

Front-end Module: 4 x 16/32 Gbps FC Ports

This module type has four FC ports that operate at 16 Gbps (auto-negotiate down to 4 Gbps) or optionally at 32 Gbps (auto-negotiate down to 8 Gbps) with installation of 32 Gbps SFPs. The ports on a board can be an intermix of empty, 4/8/16Gbps SFP, and 8/16/32Gbps SFP. Up to 16 FC ports at 16 or 32 Gbps per G350/G370 system, up to 56 FC ports at 16 or 32 Gbps for G700 systems, (64 ports at 16 or 32 Gbps when diskless), or up to 64 FC ports at 16 or 32 Gbps for G900 systems (80 ports at 16 or 32 Gbps when diskless and using the I/O Expansion Box) are supported when this FE Module type is used.

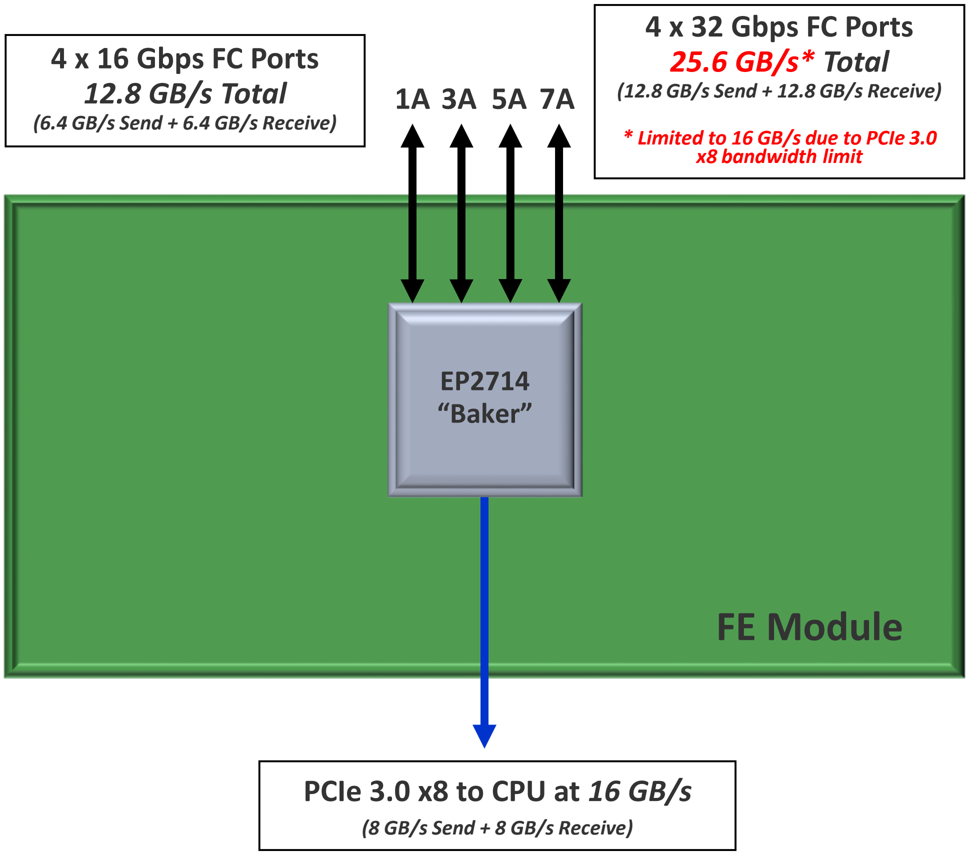

4 x 16 Gbps FC Ports
12.8 GB/s Total
(6.4 GB/s Send + 6.4 GB/s Receive)
1A 3A 5A 7A
4 x 32 Gbps FC Ports
25.6 GB/s* Total
(12.8 GB/s Send + 12.8 GB/s Receive)
* Limited to 16 GB/s due to PCIe 3.0 x8 bandwidth limit
EP2714
"Baker"
FE Module
PCIe 3.0 x8 to CPU at 16 GB/s
(8 GB/s Send + 8 GB/s Receive)

Figure 10: 4 x 16/32 Gbps FE Module

Baker Chip

The host interface processor for this module type is a "Baker" processor (single chip) and is used to bridge the Fibre Channel host connection to a usable form for internal use by the storage controller. Baker is an internal codename for the QLogic 2700 series of processors. These FE modules use the QLogic EP2714 chip. This is a Gen 6 FC only controller that can operate at 4 Gbps to 32 Gbps link speeds depending on the SFP type that is installed. This high power Baker chip provides a variety of functions, to include:

- A conversion of the Fibre Channel transport protocol to the PCIe x8 link for use by one to four controller processors:
  - SCSI initiator and target mode support
  - Complete Fibre Channel protocol sequence segmentation or reassembly

- Conversion to the PCIe link protocol

- Provides simultaneous full duplex operations of each port
- Separate processor, memory, DMA channels, and firmware for each port
- Error detection and reporting
- Packet CRC encode/decode offload engine
- Auto-sync to a 4 Gbps, 8 Gbps 16 Gbps, or 32 Gbps port speed per path (depending on the SFP installed)

The Baker processors can provide extremely high levels of performance as they are directly connected by a high-performance, PCI Express 3.0 x8 link directly to the Intel Broadwell EP Xeon processor. The EP2714 can drive all four of its 16 Gbps ports at full sequential speed but is bottlenecked by the PCIe 3.0 x8 connection to the Controller Blade when driving all four host ports at 32 Gbps. In the latter case, the maximum attainable throughput is 62.5% of the total line rate for the four 32 Gbps FC ports. For random small block loads it cannot drive all four ports at full speed (limited by handshaking overhead).

Front-end Module: 2 x 10 Gbps iSCSI Ports

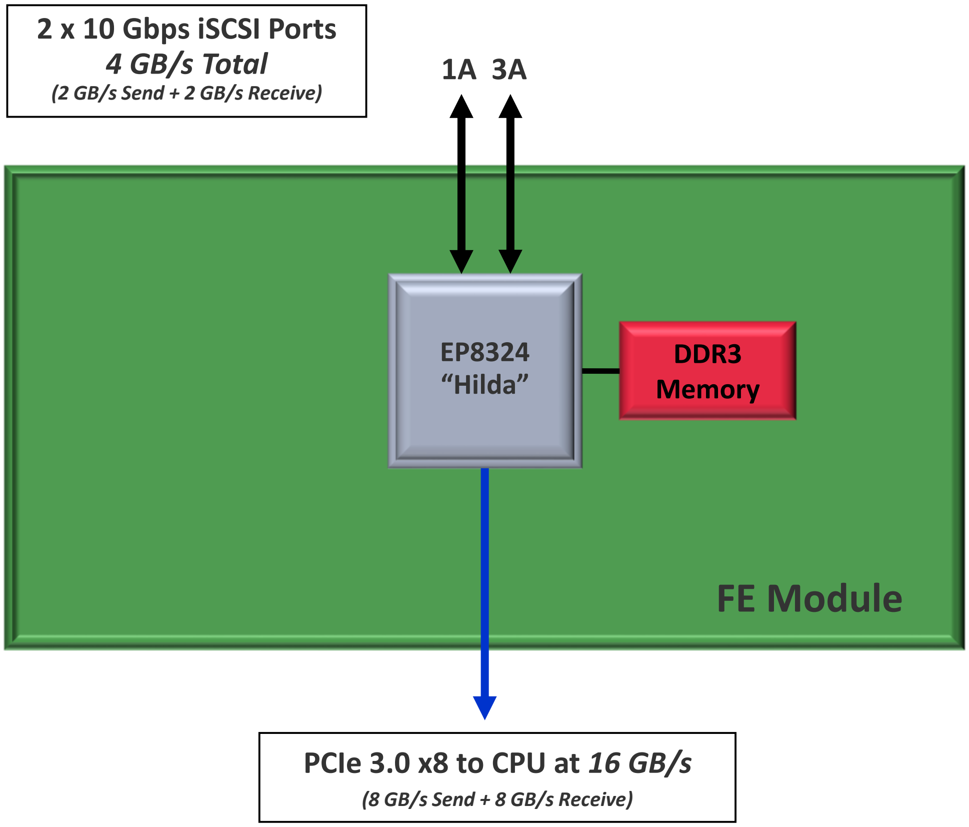

2 x 10 Gbps iSCSI Ports
4 GB/s Total
(2 GB/s Send + 2 GB/s Receive)
1A 3A
EP8324
“Hilda”
DDR3
Memory
FE Module
PCIe 3.0 x8 to CPU at 16 GB/s
(8 GB/s Send + 8 GB/s Receive)

Figure 11: 4 x 16/32 Gbps FE Module

This module type is very similar to the VSP Gx00 2 x 16 Gbps FC module, but instead provides two 10 Gbps iSCSI ports, either 10GBASE-T copper or 10 Gbps SFP+ optical. Up to 8 iSCSI ports at 10 Gbps per G350/G370 system, up to 28 iSCSI ports at 10 Gbps for G700 systems (32 ports at 10 Gbps when diskless), or up to 32 iSCSI ports at 10 Gbps for G900 systems (40 ports at 10 Gbps when diskless and using the I/O Expansion Box) is supported when this FE Module type is used.

The QLogic EP8324 processor is used here, running firmware to operate as a 10 Gbps network interface controller. In this mode, it provides a variety of functions including:

- Full hardware offload engine for IP, TCP, and UDP checksums
- Jumbo frame support (9600 bytes)

The PCB for this module includes additional ECC-protected DDR3 DRAM necessary to support the iSCSI connections and features either pluggable 10 Gbps SFP+ optical transceivers or 10GBASE-T RJ45 connectors. The optical and copper variants have separate part numbers and there is no ability to use copper transceivers to convert an optical iSCSI module into a copper one.

Back-end Module: 8 x 12 Gbps SAS Links

Overview

The Back-end drive controller modules provide the SAS links used to attach to the disks in a set of disk boxes. These modules are installed into specific slots at the rear of the G700/G900 DKC and are connected to a specific Controller Blade. There are two types of modules, with a standard version and an encryption version. Each type includes two SAS wide ports on the rear panel by which a pair of SAS wide cables to the first row of two disk boxes (DB-00 to DB-01) are connected. There are four 12 Gbps full duplex SAS links per port, with two ports per module. This module is common to the VSP Gx00 and VSP Gxx0 family of arrays.

Except for a pure virtualization (diskless) configuration, four modules are installed per VSP G700 system, two in each Controller Blade. These provide 32 x 12 Gbps full duplex SAS links (over 8 SAS wide ports) and allows the G700 system to support 1200 drives. Four or eight modules are installed per VSP G900 system, two or four modules in each Controller Blade. These provide 32 or 64 x 12 Gbps full duplex SAS links (over 8 or 16 SAS wide ports) and allows the G900 system to support 1,440 drives.

The BE module is fairly simple, primarily having a power SAS Protocol Controller chip (SPCv 12G or SPCve 12G) on the board. The enhanced SPCve 12G processor is used on the encrypting version of the BE module. All BE modules in a system must either be the standard type or encrypting type. The slot that the BE module plugs into provides power and a PCI Express 3.0 x8 link directly to the Intel Broadwell EP Xeon processor. All interaction with a BE module is via the LR emulators running on the MPUs in the same Controller Blade.

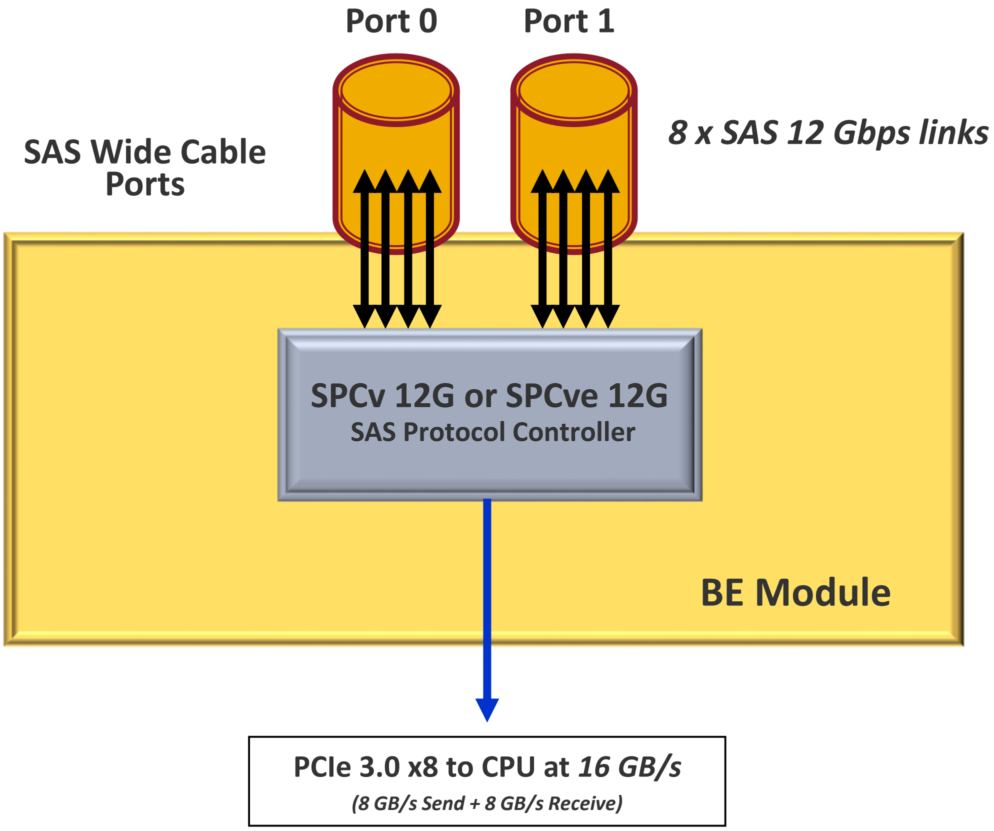


Figure 12: BE Module

The BE modules control all direct interaction to the drives (HDD, SSD, or FMD). Each BE module has a powerful dual-core SPC that executes all I/O jobs received from the MPUs (via the LR emulators running on the MPUs in that Controller Blade), managing all reading or writing to the drives. Each BE module provides two SAS wide ports on the rear panel for the cable attachment to two disk boxes. Each port is a bundle of four independent 12 Gbps SAS links. Each port controls up to 240 drives (VSP G350) or up to 360 drives (VSP G370, G700, G900) in different "stacks" of disk boxes.

SAS Protocol Controller
Each SPC is a high performance dual-core CPU that has the ability to directly control HDD, SSD, or FMD drives over its eight full-duplex 12 Gbps SAS links. The SPC has multiple DMA channels for moving data blocks in or out of cache. The SPC works in conjunction with the LR and DRR emulators running on the MPUs in the same Controller Blade.

The back-end functions managed by the SPC include the following:

- control all reads or writes to the attached drives
- execute all jobs received from an MPU processor via LR
- use internal DMA channels to move data or parity blocks in or out of cache
- track each drive's status and alert an MPU if a problem occurs or is developing
- perform all encryption/decryption of data blocks to DARE enabled Parity Groups (encrypting module only) as they pass through the SPCve from cache to disk or back

The SPC creates a direct connection with one drive per link via a disk box's (DB) internal ENC SAS switch when the SPC needs to communicate with a drive. The speed at which a link is driven is 12 Gbps, with intermixing of 6 Gbps and 12 Gbps drives supported.

Should a link fail within a port, the SPC will no longer use it, failing over to the other 3 links in that port.

Data-at-Rest Encryption Overview

The VSP Gxx0 family provides for controller-based (as opposed to drive-based) Data-at-Rest Encryption (DARE). There are two types of optional BE modules: standard and encrypting (AES256). In order to enable DARE on one or more Parity Groups, all BE modules in the system must be of the encrypting type. A license key enables the use of DARE. The Data Encryption Keys (DEK) are kept in the BE Modules and backed up to the system disk areas. The number of available encryption keys varies by model:

- VSP G350/G370: 1,024
- VSP G700/G900: 4,096

Data encryption is per drive (any types) per Parity Group, and must be enabled before any LDEVs are created. Each drive has a data encryption key. All data blocks contained in a DARE enabled Parity Group are encrypted. DARE must be disabled on a Parity Group before removing those drives from the system. Drive removal results in the data being "crypto-shredded".

Data blocks from a DARE enabled Parity group are encrypted/decrypted by the BE module's SAS SPCve controller chip as data is either written to or read from data blocks. User data is not encrypted or decrypted in cache.

Non-disruptive field upgrades of the BE modules are possible. This should be scheduled during periods of low activity to minimize any performance impact during the upgrade.

According to factory and Techops testing, the performance impact of this encryption/decryption (reads) process is minimal.

Front-end I/O Expansion Box

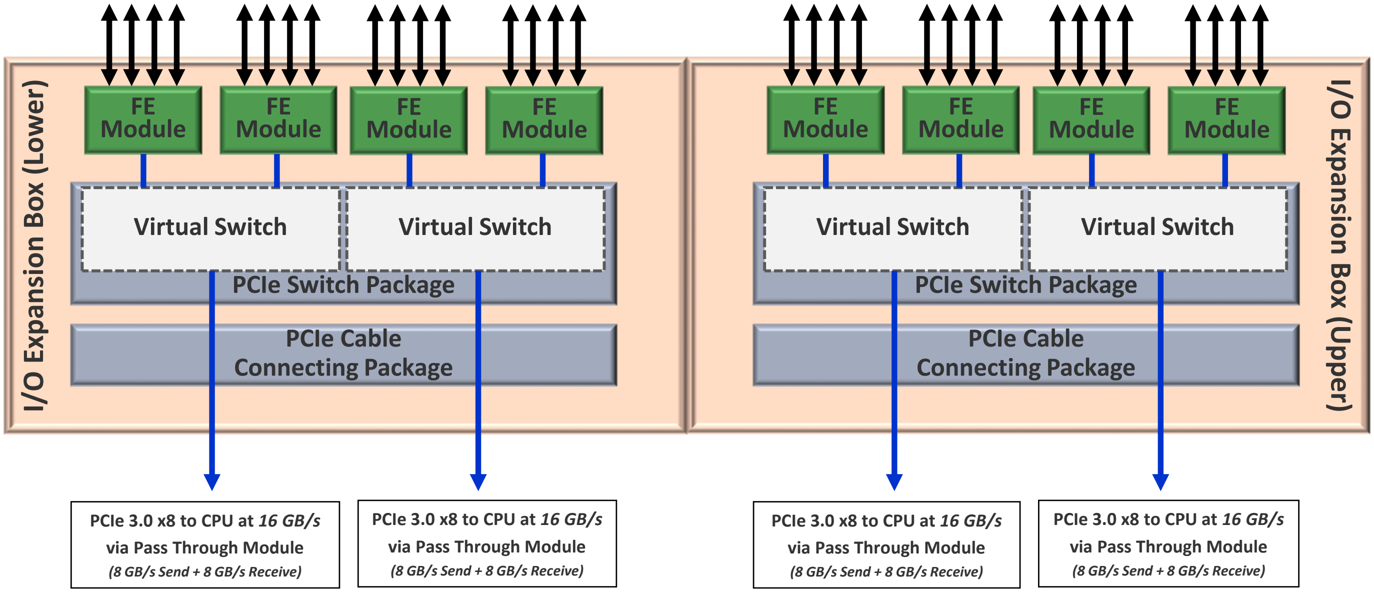

I/O Expansion Box (Lower)
FE Module
FE Module
FE Module
FE Module
Virtual Switch
Virtual Switch
PCIe Switch Package
PCIe Cable Connecting Package
FE Module
FE Module
FE Module
FE Module
Virtual Switch
Virtual Switch
PCIe Switch Package
PCIe Cable Connecting Package
I/O Expansion Box (Upper)
PCIe 3.0 x8 to CPU at *16 GB/s* via Pass Through Module *(8 GB/s Send + 8 GB/s Receive)*
PCIe 3.0 x8 to CPU at *16 GB/s* via Pass Through Module *(8 GB/s Send + 8 GB/s Receive)*
PCIe 3.0 x8 to CPU at *16 GB/s* via Pass Through Module *(8 GB/s Send + 8 GB/s Receive)*
PCIe 3.0 x8 to CPU at *16 GB/s* via Pass Through Module *(8 GB/s Send + 8 GB/s Receive)*

Figure 13: Front-end I/O Expansion Box

The Front-end I/O Expansion Box is a 2U chassis that increases the number of FE Modules that can be installed in a VSP G900 array. It houses a PCI Express switch with 8 FE Module slots and two PCIe Cable Connecting Packages (PCP), each with two external PCIe cable ports. Each PCP port is connected via external PCI Express cable to a PCIe Pass Through Module (PTM) that installs in an I/O Module slot in the G900 controller chassis. Together, 4 I/O Module slots in the DKC are used to provide 8 FE Module slots in the I/O Expansion Box, so it is in essence providing 2:1 multiplexing or oversubscription.

Each PCIe Switch Package is logically divided into two Virtual Switches, which is akin to zoning a Fibre Channel switch. Two adjacent FE Module slots share one PCIe 3.0 x8 path to the Intel Xeon CPU in the Controller Blade meaning that they share I/O path bandwidth to the rest of the array. Note that the Virtual Switch does not enforce a hard 50/50 split on PCIe path bandwidth, so one FE Module can use the entire PCIe 3.0 x8 path bandwidth if the other FE Module is idle or not installed.

Note that use of the 4 x 16/32 Gbps FE Module can lead to cases where full line rate on all ports is not achievable. For instance, with two 4 x 16 Gbps FE Modules connected to one Virtual Switch, only 62.5% of full line rate (16 GB/s vs 25.6 GB/s) is achievable. With two 4 x 32 Gbps FE Modules connected to one Virtual Switch, only 31.25% of full line rate (16 GB/s vs 51.2 GB/s) is achievable.

Drives and Drive Boxes

There are four types of optional disk boxes (DB):

- DBS: 2U 24-slot SFF SAS box
- DBL: 2U 12-slot LFF SAS box
- DB60: 4U 60-slot dense LFF SAS drawer
- DBF: 2U 12-slot FMD box

The DBS box holds up to 24 of the 2.5” SAS drives. These include all the 10K HDD and SSD options currently available

The DBL holds up to 12 of the 3.5” SAS drives. These include the 6 TB and 10 TB 7200 RPM drives.

The DB60 drawer holds up to 60 of the 6 TB or 10 TB drives or optionally the 1.2 TB 10K or 2.4 TB 10K HDD with a special DB60 LFF conversion canister.

The DBF box holds 12 Hitachi Flash Module Drives, which can be second generation FMD DC2 (3.2 TB), or fourth generation FMD 3D (7.6 TB and 15.3 TB). FMD DC2 and FMD 3D drives are nearly identical architecturally and both support in-drive data compression, but FMD 3D drives have a lower bit cost ($/GB) by virtue of their larger capacities.

Drive Box Connections to BE Modules

Figures 14-15 illustrate the connection of disk boxes to the BE Module ports. Notice that the highest performing drive boxes have been placed closest to the BE Module ports. It is best practice (but not required) that they be installed so that each “stack” has the same number of disk boxes. This does not apply to the VSP G350/G370 which both have a single stack of disk boxes, but it is best practice for the G700 and G900 models to install disk boxes to balance the back end connections. For the G700, best practice is to install disk boxes four at a time. For the G900 best practices is to install disk boxes four or eight at a time, depending on which back-end configuration is selected. The drive selection strategy when creating Parity Groups should also be considered. There are merits to each method, whether strictly linear (all drives from the same disk box), strictly dispersed (one drive per disk box), or some combination thereof. Each new row of disk boxes introduces an incremental amount of switch-to-switch latency.

The maximum number of disk boxes that can be daisy chained in each back-end stack differs by model:

- VSP G350: 8 (1 internal + 7 external)
- VSP G370: 12 (1 internal + 11 external)
- VSP G700: 12
- VSP G900: 6

For Gxx0 systems with disk boxes that span more than one rack, it is possible to separate the racks so that they aren’t adjacent to each other. This requires the use of optical SAS cables (5, 10, 30, 100m lengths) to connect the drive boxes in the physically separated racks, with the following limits on maximum SAS cable path length:

- VSP G350: 130m
- VSP G370: 150m
- VSP G700: 140m
- VSP G900: 125m

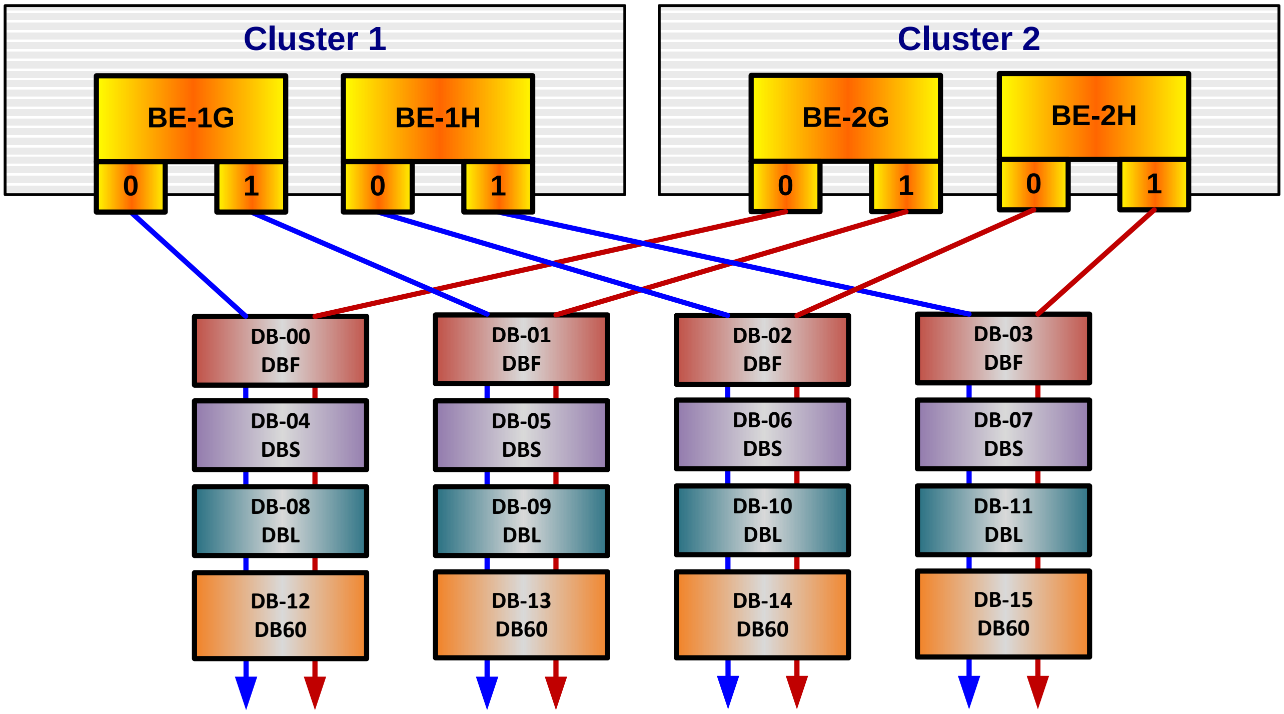


Figure 14: Map of the BE Module Ports to DBs: G700 (Or G900 with Standard Back-end)

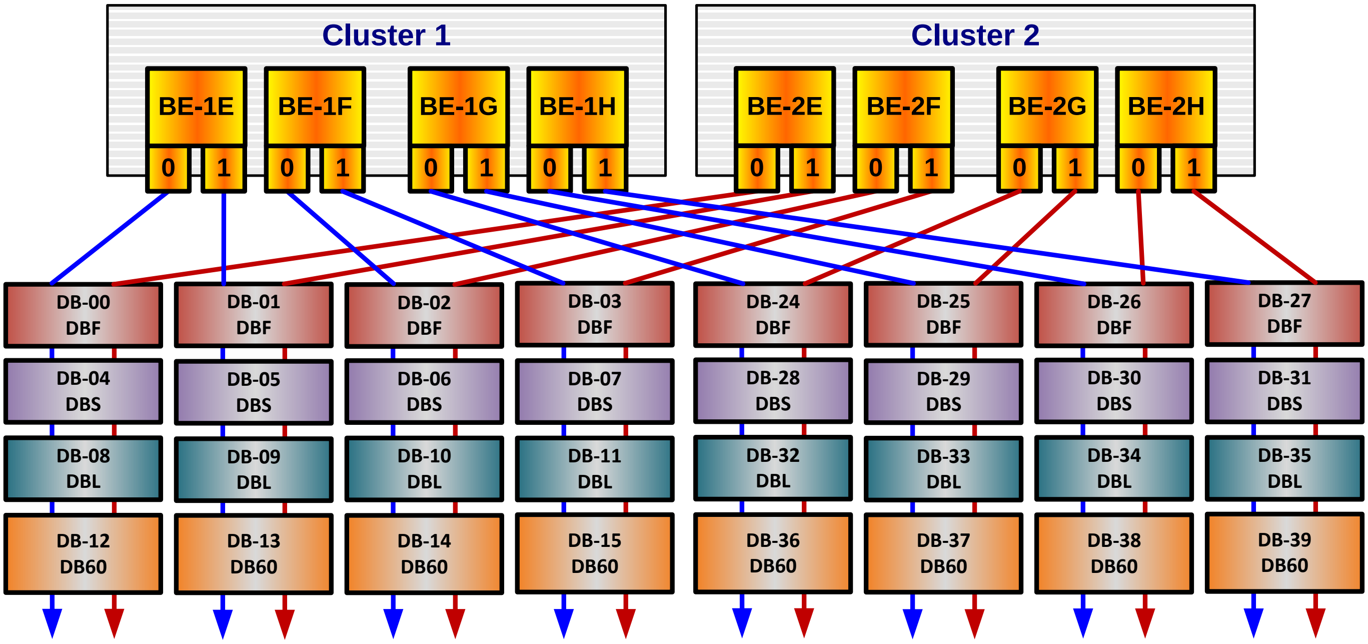


Figure 15: Map of the BE Module Ports to DBs: G900 with Performance Back-end

24-Disk SFF Tray (DBS)

*Figure **16*** depicts a simplified view of the 2U 24-disk 2.5” SAS disk box. Inside each 24-disk tray is a pair of SAS “expander” switches. These may be viewed as two 32-port SAS switches attached to the two SAS wide cables coming from the BE Module ports, or from the next drive box in the stack in the direction of the controller. The two expanders cross connect the dual-ported disks to all eight of the active 12 Gbps SAS links that pass through each box. Any of the available 2.5” HDD or SSDs may be intermixed in these trays, including a mix of 6 Gbps and 12 Gbps SAS interface drives. The drives are arranged as one row of 24 slots for drive canisters. All drives within each box are dual attached, with one full duplex drive port going to each switch.

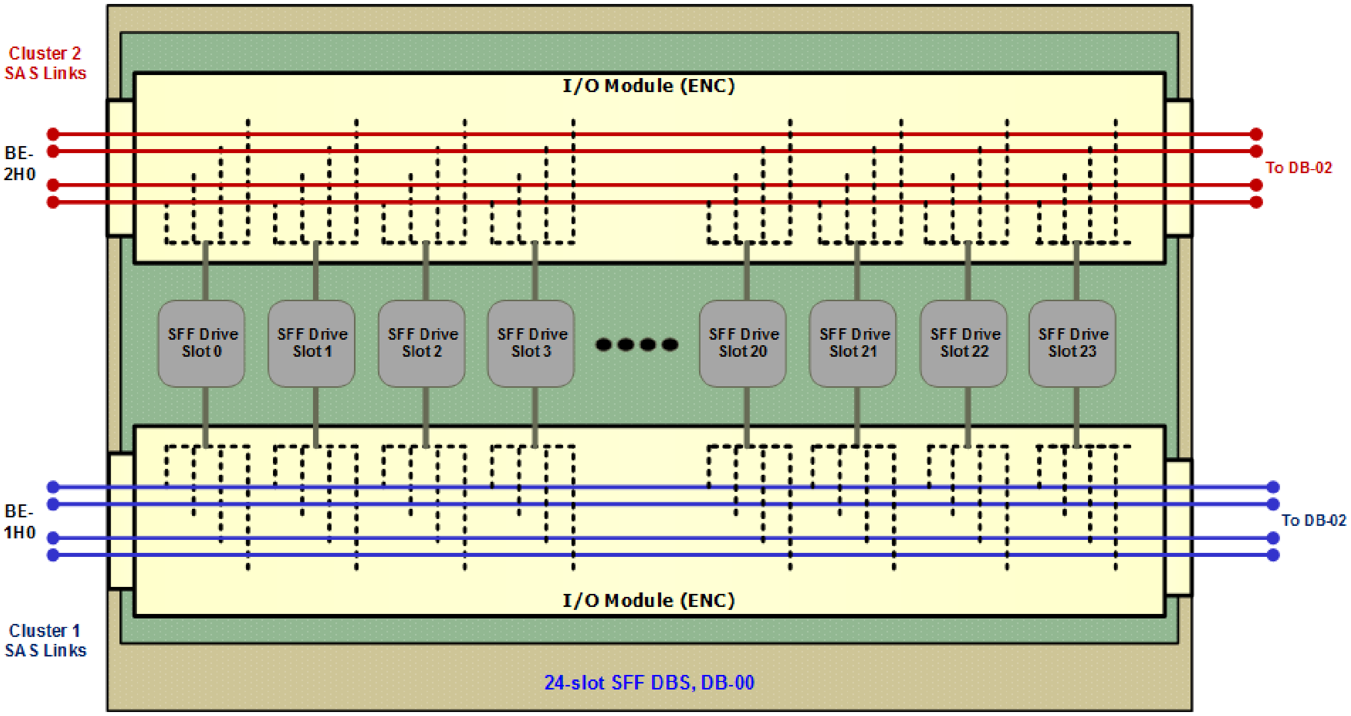


Figure 16: DBS Details

12-Disk LFF Tray (DBL)

*Figure **17*** depicts a simplified view of the 2U 12-disk 3.5" SAS disk box. Inside each 12-disk tray is a pair of SAS expander switches. These may be viewed as two 20-port SAS switches attached to the two SAS wide cables coming from the BE Module ports, or from the next drive box in the stack in the direction of the controller. The two expanders cross connect the dual-ported disks to all eight of the active 12 Gbps SAS links that pass through each box. Any of the available 3.5" HDDs may be intermixed in these trays, including a mix of 6 Gbps and 12 Gbps SAS interface drives. The drives are arranged as three rows of 4 slots each for drive canisters. All drives within each box are dual attached, with one full duplex drive port going to each switch.

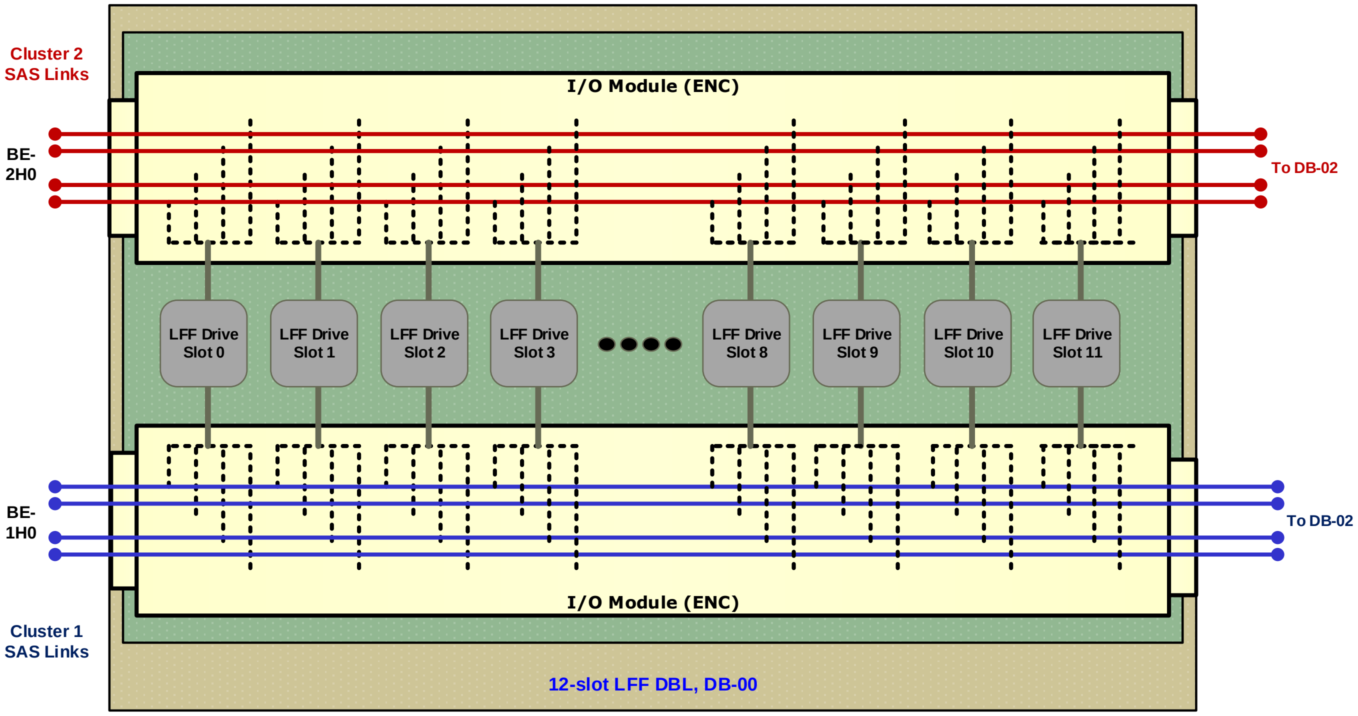


Figure 17: DBL Details

60-Disk LFF Drawer (DB60)

The high density 4U LFF disk drawer shown in *Figure **18*** has 60 3.5” vertical disk slots. Inside each 60-disk drawer is a pair of SAS expander switches. These may be viewed as two 68-port SAS switches attached to the two SAS wide cables coming from the BE Module ports, or from the next drive box in the stack in the direction of the controller. The two expanders cross connect the dual-ported disks to all eight of the active 12 Gbps SAS links that pass through each box. Any of the available 3.5” HDDs may be intermixed in these trays, including a mix of 6 Gbps and 12 Gbps SAS interface drives. In addition, there are specially ordered SFF drives (1.2 TB and 2.4 TB 10K RPM) that ship with special canisters to fit the DB60 LFF slots. This drawer operates as a single disk box, with the drives arranged as 5 rows of 12 slots each for drive canisters. All drives within each box are dual attached, with one full duplex drive port going to each switch.

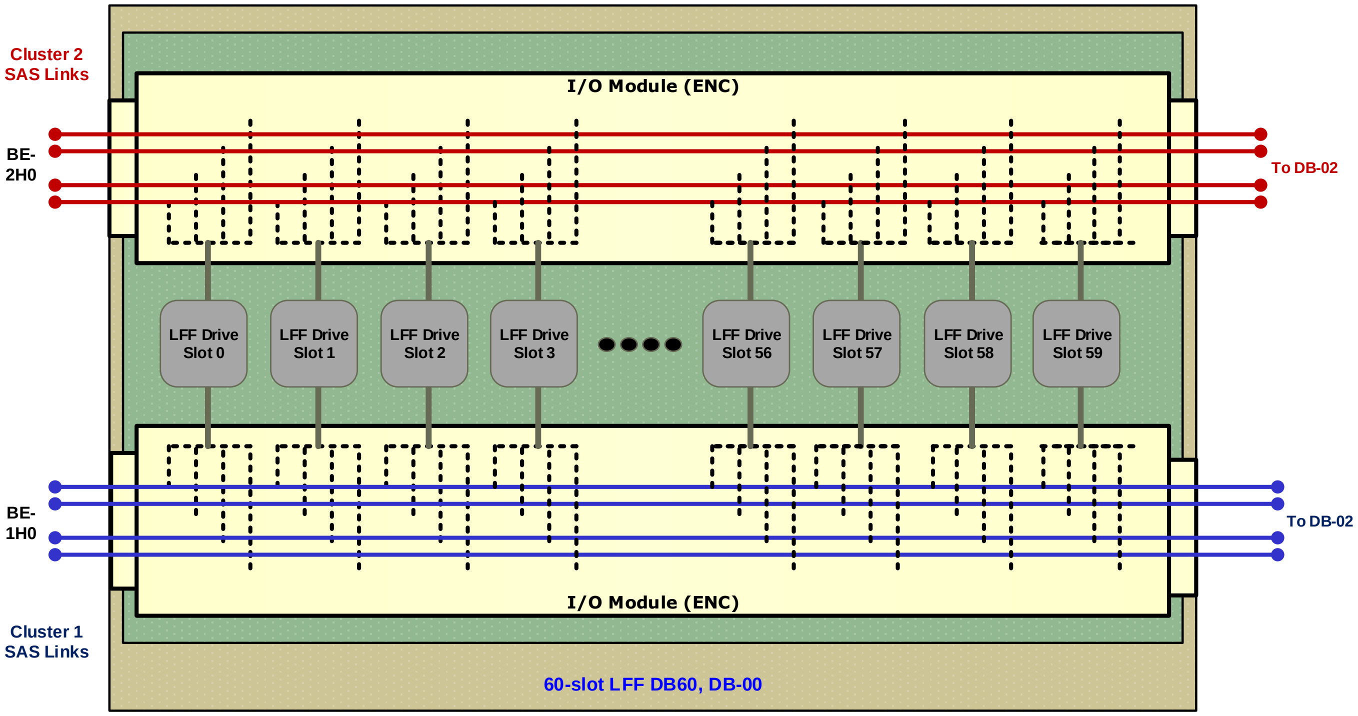


Figure 18: DB60 Details

12-Disk FMD Tray (DBF)

*Figure **19*** depicts a simplified view of the 2U 12-slot DBF box details. Inside each 12-FMD module box there are two SAS expander switches. These may be viewed as two 20-port SAS switches attached to the two SAS wide cables (8 SAS links) coming from the BE Module ports, or from the next drive box in the stack in the direction of the controller. The two expanders cross connect the Flash Module Drives (FMD DC2 or FMD 3D) to all eight of the active 12 Gbps SAS links that pass through each box. The FMDs are arranged as four rows of 3 slots each. All FMDs within each box are quad attached, with two full duplex drive ports going to each switch.

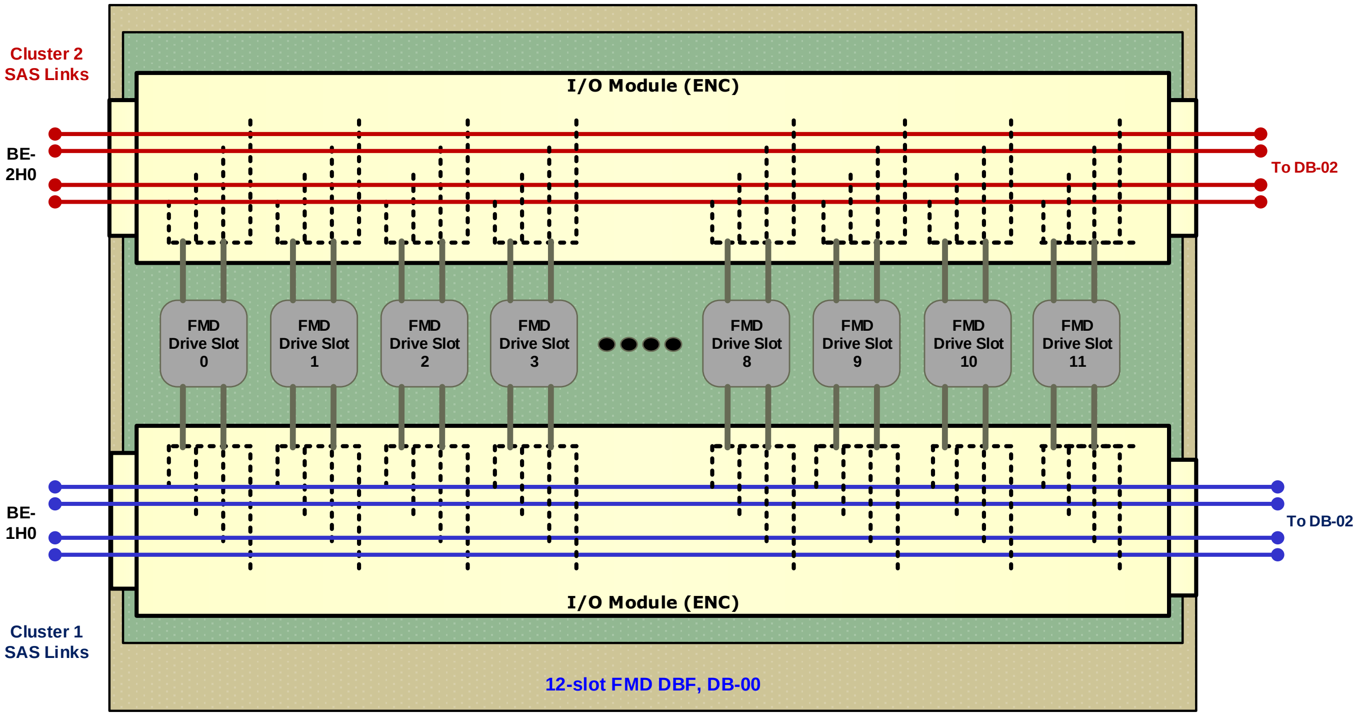


Figure 19: DBF Details

Drive Details

*Table **4*** lists the maximum drive count for each VSP Gxx0 model and drive type.

Model	SFF HDD (DBS) *5	LFF HDD (DBL)	LFF HDD (DB60)	FMD (DBF)	Spares
VSP G350	192 *1	96 *2	252 *3	n/a	16
VSP G370	288 *1	144 *2	372 *3	n/a	24
VSP G700	864	432	1200	432	48
VSP G900	1,152 *4	576 *4	1,440 *4	576 *4	64

Table 4: Maximum Drive Counts

*1: Includes 24 drives in CBSS with remainder in DBS disk boxes.

*2: Includes 12 LFF drives in CBSL with remainder in DBL disk boxes.

*3: Includes 12 LFF drives in CBSL with remainder in DB60 disk boxes.

*4: Requires the Performance Back-end (8 BE Modules).

*5: Includes SFF SSDs.

Drive Type	RPM	Form Factor	Port Speed	Advertised Size (GB)	Raw Size (GB)	Nominal Random Read IOPS
600 GB SAS	10K	SFF	6 Gbps	600	576	145
			12 Gbps			
1.2 TB SAS	10K	SFF, LFF Canister	6 Gbps	1,200	1,153	145
			12 Gbps			
2.4 TB SAS	10K	SFF, LFF Canister	12 Gbps	2,400	2,306	145
6 TB SAS	7200	LFF	12 Gbps	6,000	5,874	100
10 TB SAS	7200	LFF	12 Gbps	10,000	9,790	100

Table 5: Supported HDDs and Rule-of-Thumb IOPS Rates

Drive Type	NAND Type	Form Factor	Port Speed	Advertised Size (GB)	Raw Size (GB)
480 GB SSD	Toshiba MLC	SFF	12 Gbps	480	473
960 GB SSD	Toshiba MLC	SFF	12 Gbps	960	945
1.9 TB SSD	Toshiba MLC	SFF	12 Gbps	1,900	1,890
1.9 TB SSD	Toshiba TLC	SFF	12 Gbps	1,900	1,890
3.8 TB SSD	Toshiba MLC	SFF	12 Gbps	3,800	3,781
3.8 TB SSD	Hitachi TLC	SFF	12 Gbps	3,800	3,781
3.8 TB SSD	Toshiba TLC	SFF	12 Gbps	3,800	3,781
7.6 TB SSD	Toshiba TLC	SFF	12 Gbps	7,600	7,562
7.6 TB SSD	Hitachi TLC	SFF	12 Gbps	7,600	7,562
15 TB SSD	Toshiba TLC	SFF	12 Gbps	15,000	15,048
3.5 TB FMD DC2	Hitachi MLC	FMD	12 Gbps	3,500	3,518
7 TB FMD HD	Hitachi TLC	FMD	12 Gbps	7,000	7,037
14 TB FMD HD	Hitachi TLC	FMD	12 Gbps	14,000	14,074

Table 6: Supported Flash Media

The type and quantity of drives selected for use in the VSP Gxx0 and the RAID levels chosen for those drives will vary according to an analysis of the customer workload mix, cost limits, application performance targets and the usable protected capacity requirements. The use of 6 Gbps SAS drives does not affect the overall speed of the 12 Gbps SAS backend, as each drive will operate at its native interface speed. In general, the SAS interface is not the bottleneck for individual drive performance, so the interface speed should not be used to choose one drive type over another. Also note that all of the drives available for VSP Gxx0 now support 12 Gbps SAS connections.

Nominal IOPS are not listed for SSDs and FMDs because they are all capable of exceptionally high performance (100K+ Random Read IOPS) but require a large number of host I/O threads and correspondingly deep queue depth in the flash drive in order to achieve this.

SAS HDD

Hard Disk Drives (HDD) have long been the primary drives used in both enterprise and midrange storage systems from Hitachi. These are the workhorse drives, having both relatively high performance and good capacity. They fall in between the random performance levels of SSDs and SATA HDDs, but are much closer to the SATA end of the performance spectrum. SAS HDDs come in three rotation speeds: 15K RPM, 10K RPM, and 7.2K RPM, but only the latter two are available for the VSP Gxx0 platforms. All three have about the same sequential read throughput when comparing the same RAID level. The SAS interface speed of all of the supported HDDs is 12 Gbps, with some that also support 6 Gbps.

SAS drives are designed for high performance, having dual host ports and large caches. The dual host ports allow four concurrent interactions with the attached BE modules. Both ports can send and receive data in full duplex at the same time, communicating with the DRAM buffer in the drive. However, only one data transfer at a time can take place internally between the DRAM buffer in the drive and the serial read/write heads.

SAS drives may be low level formatted to several different native sector sizes, for instance 512, 520, 524, or 528 bytes. Hitachi uses the 520 byte sector size internally, while all host I/Os to the storage system are done using the usual 512 byte host sectors size. An 8-byte check code comprising a checksum over the host-view logical address (LDEV and Logical Block Address) and a checksum over the contents of the sector is added to each 512-byte sector received from the host to build the 520-byte internal sector that is written to disk. These check code fields are used to detect errors in virtualization mechanisms and corruption of sector contents, respectively.

SSD

Solid State Drives (SSD), as a storage device technology, has reached a level of maturity and market adoption that belies the small niche it once occupied. By replacing spinning magnetic platters and read/write heads with a non-rotating NAND flash array managed by a flash translation layer, the SSD is able to achieve very high performance (IOPS) with extremely low response times (often less than 1ms). It is these two characteristics that make them a viable choice for some workloads and environments, especially as the top tier of an HDT configuration. Price-capacity ($/GB) and price-performance ($/IOPS) are two areas where SSDs typically do not compare well to HDDs.

Small block random read workloads are where SSDs perform their best and often justify their high cost. They are far less cost effective for large block sequential transfers compared to HDDs, since the ultimate bottleneck is host or internal path bandwidth and not usually the performance of the individual storage device. The SAS interface speed on all supported SSDs is 12 Gbps.

FMD DC2

Hitachi’s custom Flash Module Drives could be described as turbo-charged SSDs. Each FMD has higher performance than an SSD due to much higher internal processing power and degree of parallelism. Inside each FMD is a 4-core ARM processor and ASIC (the Advanced Switch Flash Controller – ASFC) that controls the internal operations, the four 12 Gbps SAS ports, and the NAND flash memory chips. Hitachi FMDs have more cores and more independent parallel access paths to NAND flash chips than standard SSDs. There is a considerable amount of logic in the ASIC that manages the space and how writes are managed. This greatly aids in substantially increasing the write rate for an FMD over a standard SSD.

The FMD DC2 (Data Compression, 2nd Generation) is an evolution of the first generation FMD and features a higher performance 4-core ARM processor, four 12 Gbps SAS ports, improved performance, and native data compression. The FMD will compress written data and uncompress read data in an inline fashion. This uses the processing power within the FMD to achieve full data compression offload so that there is no MP busy overhead to the system. To utilize the saved capacity from compression, LDEVs from the parity groups with Accelerated Compression turned on must be used as pool volumes in an HDP pool. Capacity management and compression statistics are maintained at the HDP pool level.

FMD HD

The FMD HD (3D NAND) is a variant of the FMD DC2 with higher capacity and lower bit cost ($/GB) and is functionally identically to FMD DC2, including Accelerated Compression support.

Cache Memory Structure

The VSP Gxx0 family has a single physical memory system comprised of all the memory DIMMs installed on both Controller Blades. There are two regions of cache memory, also known as Cache Sides, corresponding to the set of memory DIMMs located on a given Controller Blade. For instance, Cache Side A corresponds to the set of DIMMs populated on Controller Blade 1 (Cluster 1). Likewise, Cache Side B corresponds to the set of DIMMs populated on Controller Blade 2 (Cluster 2).

Cache Memory Overview

The VSP Gxx0 uses a dynamic, quasi-partitioned cache allocation system in order to provide distinct areas of memory for each of the system software's functions. Each MPU "owns" a set of LDEVs for which it controls all I/O operations. This distribution of LDEV ownership is normally fairly uniform across the available MPUs in the system. In the VSP Gxx0 series there are 2 MPUs (one per Controller Blade) available to the system.

Cache space is managed in the same manner as on the VSP G1x00, with 64 KB cache segments mapped to an MPU for host blocks for a specific LDEV. Each MPU maintains its own Write Pending queue which tracks its current set of dirty cache segments.

The top level management of the cache system is the **Global Free List** that manages all of the 64 KB cache segments that make up the **User Data Cache** area. Each MPU also operates a Local Free List (of all of the 64 KB segments currently allocated to it from the master Global Free List) from which it privately draws and returns segments for individual I/O operations on its LDEVs. If the system determines that an MPU (based on its recent workloads) holds excess 64 KB segments, some of these 64 KB segments are pulled back to the system's Global Free List for reallocation to another MPU that needs them.

Cache Memory Areas

There are six distinct memory areas that serve different functions:

- **Local Memory** (**LM**) – Memory allocated as a workspace for each individual MP (core). There is a separate LM area on each Cache Side for the MPs on that Cluster.
- **Package Memory** (**PM**) – Memory allocated as a workspace for each MPU. There is a separate PM area on each Cache Side for the MPUs on that Cluster.
- **Data Transfer Buffer** (**DXBF**) – Memory used as a temporary staging area for data transfers between the two Controller Blades. There is a separate DXBF area on each Cache Side. The DXBF has Front-end (FE DXBF) and Back-end (BE DXBF) sub-areas that are used for staging data to be sent to a host and staging data read from disk respectively.
- **Shared Memory** (**SM**) – Control memory that contains the configuration data, metadata, and control tables necessary for system operation. There is a separate SM area on each Cache Side. The SM has mirrored and non-mirrored sub-areas. The Mirrored SM sub-area contains the global set of control memory and updates are mirrored between both Cache Sides. The Non-Mirrored SM sub-area is a reduced copy of Shared Memory used by the MPU on a given Cluster.
- **Cache Directory** (**CD**) – Lookup table that contains a mapping of LDEVs and active 64 KB cache segments organized by MPU. There is a separate CD on each Cache Side.
- **User Data Cache** (**CM**) – All space not allocated for the previous functions that is set aside for general use by the available MPUs for user data blocks. CM is interleaved across both Cache Sides and cache segment allocation is generally performed on a round-robin basis to balance data distribution. However, there are complex

algorithms that factor in the workload pattern and cache hit rate to place user data in the optimal location such that data transfer between the Controller Blades (via I-Path) is minimized.

The Gxx0 family uses the concept of a Cache Management Area, which includes all of the following: LM, PM, DXBF, SM, and CD (everything except user data cache). SVOS calculates the required Management Area capacity based on total DIMM capacity and total pool capacity, and automatically takes the necessary amount of memory from user data cache. This is a significant improvement in ease of use compared to the VSP Gx00 systems, where the user was responsible for calculating shared memory requirements based on pool capacity and program product usage.

On a planned shutdown, first all dirty data is destaged from cache Write Pending to back-end drives, and then Shared Memory is copied to the cache backup SSDs (CFM) on the Controller Blades. In the event of power loss, the entire contents of cache (including all Write Pending segments) will be copied to the CFMs on the Controller Blades.

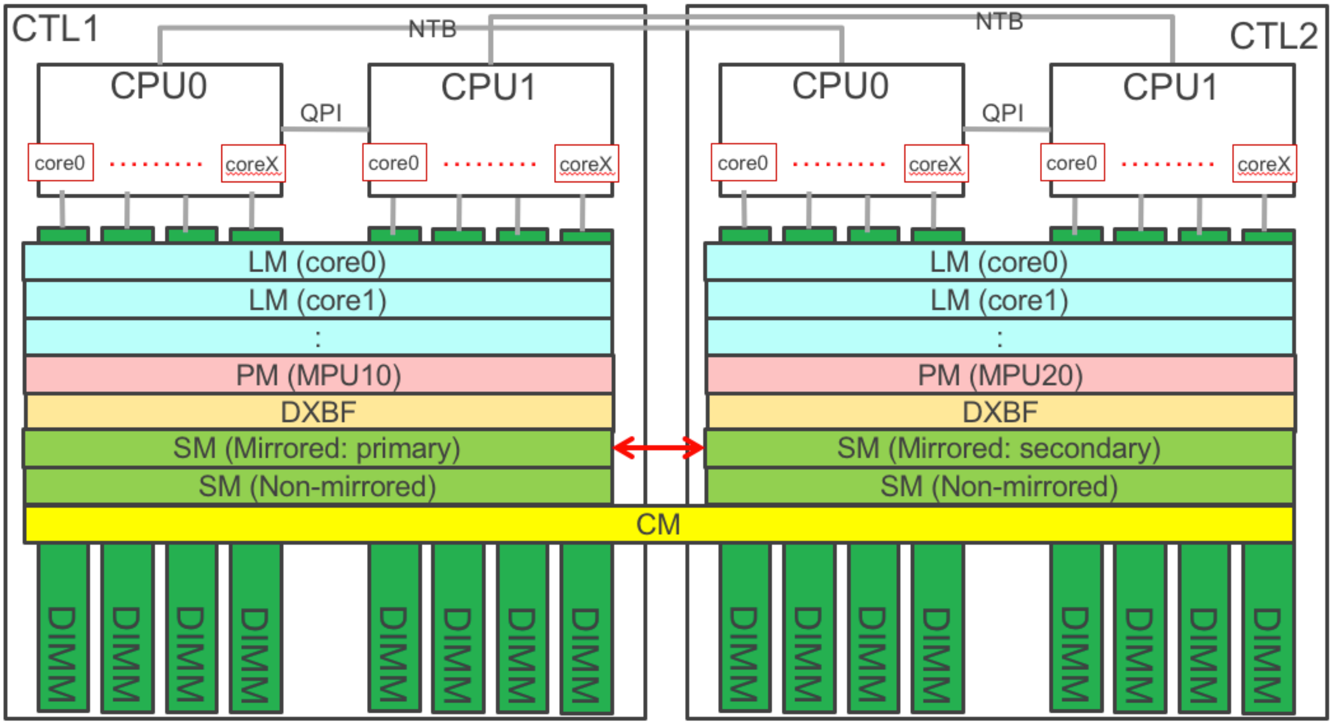


Figure 20: Cache Area Layout (VSP G700, G900)

Cache Memory Groups

The memory DIMMs on a Controller Blade are organized into groups referred to as *Cache Memory Groups* (CMG). *Figure **21*** shows CMG0 on a VSP G350/370 consists of 2 DIMMs. *Figure **22*** shows CMG0 on a VSP G700 or G900 consists of 4 DIMMs, 2 per CPU. For G700 and G900, CMG1 is an installable option and consists of an additional 4 DIMMs, 2 additional per CPU. The purple boxes represent memory DIMMs and the red lines represent DDR4 memory channels. Note that the G350 and G370 can only use two memory channels per CPU, while the G700 and G900 with the optional CMG1 can use all four memory channels from each CPU.

All DIMMs within a CMG must be the same size. In addition, CMG0 must be installed before CMG1 and the DIMM capacity must be the same across both CMGs. Recall that the two Controller Blades must be symmetrically configured, so this means that a mix of DIMM capacities is not permitted regardless of model.

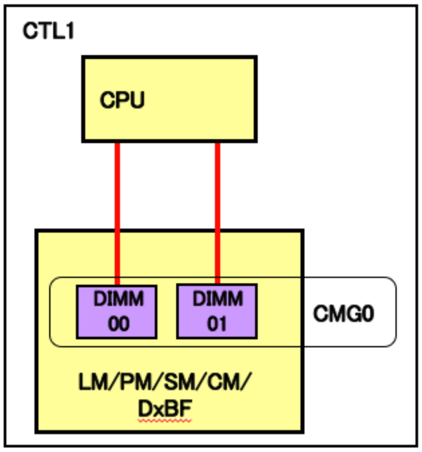


Figure 21: Cache Memory Group (VSP G350, G370)

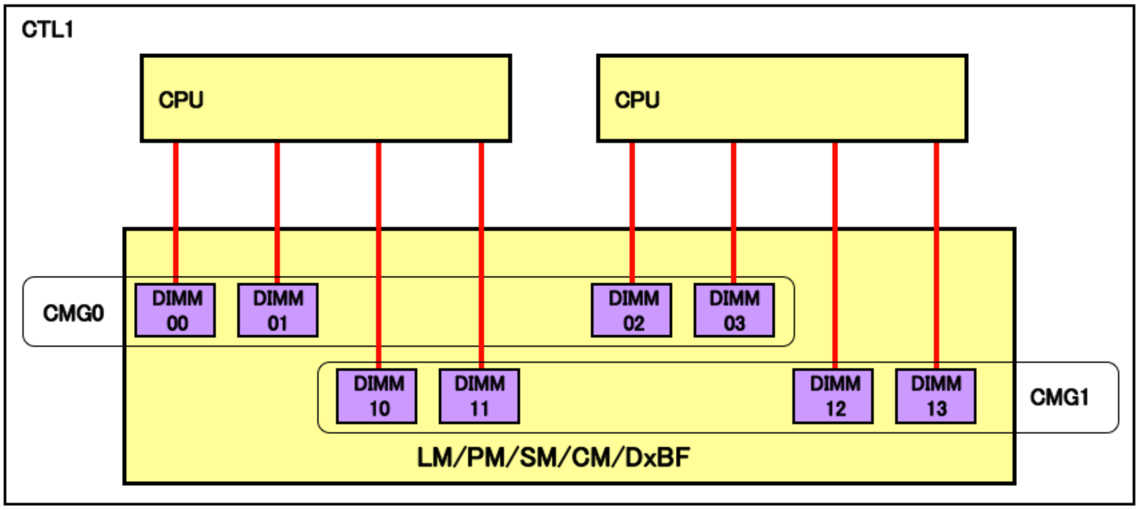


Figure 22: Cache Memory Group (VSP G700, G900)

Cache Logical Partitions

Each VSP Gxx0 system has a base or default CLPR 0 that includes all available User Data Cache space. Optional CLPRs may be established (by reducing the size of CLPR 0) to manage workloads on certain Parity Groups or DPVOLs. All Parity Groups and DPVOLs are assigned to CLPR 0 by default. CLPR 0 will use the entire User Data Cache area as a pool of 64 KB segments to be allocated to the available MPUs. The space within CLPR 0 (less a small reserve for borrowing) is divided up among the MPUs.

MPU Partitions within CLPRs

The space within each CLPR (or only CLPR 0 if just using the default) is divided up into private dynamic (not fixed in size) partitions for the MPUs. The maximum initial space per MPU partition is roughly determined by dividing the size of the CLPR by the number of MPUs (2). So if CLPR 1 is 64 GB, each MPU in the system would nominally have about 32 GB of that CLPR. This allocation can change dynamically, with the busier MPU “borrowing” cache space from the less busy MPU within the CLPR.

Creating Additional CLPRs

The **Virtual Partition Manager** (VPM) program product can be used to reduce the size of CLPR 0 and then create one or more additional CLPRs. The minimum CLPR size is 4 GB and the capacity increment is 2 GB.

When creating additional CLPRs, the original User Data Cache area is essentially being split into multiple areas of sizes specified by the user, where that space is only used for those Parity Groups or DPVOLs mapped to that CLPR. So cache usage by application or by storage tier (and external LDEVs) may be controlled in this fashion.

The space in each new CLPR created will always be shared by both MPUs, with each starting off with a fixed share of that space whether or not they have LDEVs assigned there. There is a minimum footprint per CLPR per MPU that will always remain regardless of usage. For example, if CLPR 1 is 16 GB in size but only one MPU has LDEVs assigned to this CLPR, 8 GB will initially go unused. Over time, the busy MPU will borrow cache within CLPR 1 from the idle MPU, but there will always be some minimum reserved capacity that will be untouched.

Each CLPR has its own space and queue management mechanisms. As such, each CLPR will have its own Free, Clean, and Dirty queues and each MPU maintains a separate Write Pending counter per CLPR. It is important for the storage administrator to drill down and monitor each individual MPU-CLPR Write Pending counter, as the global Write Pending counter is an average across all MPU-CLPR pairings. This can hide the fact that some MPU-CLPR pairings have hit 70% Write Pending (emergency destage and inflow control) while others are sitting at 0% Write Pending.

I/O Process Concepts

This section is based on our current understanding of the VSP Gxx0 system architecture and the manner in which it processes internal commands. As we learn more from design engineering or determine new details through our own lab tests, this section will be revised as needed.

Each Controller Blade hosts one or two Intel Broadwell EP Xeon processors. These are organized into MPU logical units (one per Controller Blade) and are where the vast majority of the activity takes place in this architecture. The MPUs are responsible for executing the system software, otherwise known as the Storage Virtualization Operating System (SVOS). This is where the basic I/O management and processing tasks are executed, along with any virtualization or replication software in use.

The Controller Blades cross-connect with a pair of external links (I-Paths) which form a pair of PCI Express Non-Transparent Bridges (NTB). The I-Paths use PCI Express 3.0 lanes provided by the embedded PCI Express switch in the Intel Xeon processors. A VSP Gxx0 I-Path is considered non-transparent because it provides a limited interface over which command messaging and user data transfer can occur. This contrasts with the transparent bridging ability of the I-Paths in the HUS VM design. There, the HM ASICs provide a direct data transfer path across the I-Paths between MAIN blades, such that user data in cache on Cluster 2 can be DMA transferred and sent via I-Path out to a Front-end port on Cluster 1. Here, user data transfers across I-Paths always occur in two steps. First, the data is DMA transferred from cache and sent across an I-Path into a Data Transfer Buffer (DXBF) on the other Cluster. Then a DMA transfer is performed on the other Cluster to move the data in the DXBF to its destination, either to an attached host via FE Module port or to Back-end disk via BE Module. Due to the extra overhead involved in these two step I-Path transfers, SVOS algorithms optimize the placement of data in cache to minimize I-Path utilization.

Within a G700 or G900 Controller Blade, the two Intel Xeon processors are connected with a QuickPath Interconnect (QPI) which is a high speed, low latency path that functions effectively as a transparent bridge. Either CPU can access the DDR4 memory channels, PCI Express paths, and FE and BE Modules physically connected to its partner CPU with virtually no latency. QPI is not used on the VSP G350/G370 Controller Blade since there is only a single Intel Xeon CPU installed.

The FE module host interface processors communicate with the LRs on the same Controller Blade. Each CPU in the system runs an ASIC emulator that executes traditional hardware ASIC functions in microcode on the CPU itself. One of these emulated functions is the Local Router function. I/O requests are placed in the LR emulator queue on the same cluster as the receiving FE module and the first available LR pulls the request off the queue then creates and hands off a “job” to the owning MPU to execute. The LR emulators have tables that contain maps of LDEV-to-MPU ownership. If the owning MPU is in the same Cluster, then this is considered a Straight I/O. If the owning MPU is in the other Cluster, the job is forwarded to that MPU via one of the I-Paths, and this is considered a Cross I/O. The owning MPU responds to the LR with a cache address (whether for a read hit, a read miss, or a write buffer). If a read miss request, the owning MPU will invoke integrated RAID configuration and single-drive access functions to communicate with one or more BE Module SPC Processors that will then perform the requested physical disk read. For writes, the MPU will spawn one or more additional jobs and dispatch these to the one or more BE Module SPC processors that will then perform the requested physical operations. For writes, the BE Modules use the DRR emulator function running on the owning MPU to generate the parity blocks required and then write these blocks to the appropriate disks. There will be two disk updates required for RAID-10 or RAID-5, and three updates for RAID-6.

Read Hits (from cache)

For read hits, the owning MPU will perform a Cache Directory lookup to determine where in cache the requested data already exists (from a recent previous read or a completed write). If the data is in cache on the same cluster as the FE Module that received the host request, the owning MPU will pass the requested cache address to the FE host interface processor (via LR). The host interface processor will then use one of its DMA channels to read that User Data cache address and transmit the requested blocks to the host.

If the data is in cache on the cluster opposite of the FE Module that received the host request, the owning MPU will spawn a data transfer job to read the data from User Data cache and copy it into the FE DXBF on the opposite Cluster via I-Path. The owning MPU then passes the addresses in the FE DXBF to the FE host interface processor (via LR). The host interface processor then uses one of its DMA channels to read the FE DXBF address and transmit the requested blocks to the host.

Read Miss (from disk)

For a read miss (data not found in cache), the host requests (via LR) that the owning MPU schedule a read operation from a back-end drive by one of the BE modules to a target User Data region cache address specified by the MPU. How the target cache location is selected is based on the I/O type and the LUN cache hit rate:

- For I/O that is judged to be Sequential in nature, the target cache location is on the same cluster as the FE Module that received the host request. Since I/O that is judged to be Sequential triggers the system to do Sequential prefetch, this reduces the amount of data that needs to be sent across the I-Paths.
- For I/O that is judged to be non-Sequential to a LUN with a low cache hit ratio, the target cache location is on the same cluster as the FE Module that received the host request. This is done to avoid incurring the overhead of transferring data across the I-Paths.
- For I/O that is judged to be non-Sequential to a LUN with a high cache hit ratio, the target cache location is determined on a round robin basis, alternating between Clusters 1 and 2. Since the likelihood of this data being read again by the host is high, this aims to balance cache allocation and read hits across the entire User Data cache space.

The appropriate BE SPC on the same cluster as the target cache location can directly push the requested data via DMA into cache. After the back end read is completed, the same process as for Read Hits occurs, with the MPU passing the requested cache address to the FE host interface processor, or first initiating a data transfer job via I-Path then passing the FE DXBF address to the FE processor. Ultimately, the FE processor uses its DMA to transfer the requested blocks and complete the host I/O.

Writes (to disk)

For every host write request, the process can be split into a Front-end Write process and a subsequent Back-end Write process. The Front-end Write process writes two copies of the data into cache, one copy per cache side, so that in the event of a single failure data will not be lost. After the Front-end Write process is complete, the host is informed that the I/O is complete. Then after that, the Back-end write process will asynchronously destage one of the copies to a back-end parity group.

The Front-end Write process begins when the host write request is received and the FE Module's host interface processor communicates the request to the first available LR on the same Cluster. The LR that picks up the request creates a Front-end write job and passes it to the owning MPU.

The MPU then does a Cache Directory lookup to see if the target 256 KiB cache "slot" is already mapped in the cache directory. A cache slot is a metadata object with four slots to plug in 64 KiB "segments". A "write hit" is when the slot is already found in the cache directory, whether or not the target 64 KiB cache "segment" is mapped within the slot. A "write miss" is when the target slot is absent from the cache directory.

For a write miss, an exclusive global lock on the cache directory is obtained in order to insert a new cache slot into the directory. A cache slot can be either "clean" meaning that it contains a copy of data that is already stored on a back end parity group, or it is "dirty", meaning that it contains data newly written by the host that has not yet been destaged to back-end drives. Then 64 KiB segments are taken from the free queue to populate the cache slot as necessary to contain the dirty data. If the slot being written to started out being clean, as it is converted to "dirty", one or more additional 64 KiB segments are obtained from the free queue to store the second copy of the data.

Once the dirty cache slot is ready to receive the data from the host, the FE host interface processor uses DMA to transfer the host data into the cache region specified by the owning MPU. This new host data is then duplexed, or mirrored, to the opposite side of cache on the other cluster, via command and data transfer across an I-Path. After the data has been written to both sides of cache, a data integrity check is executed by the MPU on the mirror side. If the data integrity check passes, the owning MPU routes a write completion message to the FE processor via LR, whereby the processor then notifies the host. If this was an overwrite of data already in cache, the new data in the FE DXBF is then copied to the old data location in User Data cache, and the FE DXBF data is discarded. This completes the Front-end Write process.

The Back-end Write process begins when the owning MPU creates a back-end job for that LDEV which it routes via LR to the BE Modules that will perform the writes (on either Cluster, depending on where the necessary BE Modules are located). The BE Module SAS SPCs will then receive the MPU commands to operate on this user data via LR, and the SPCs can directly access the data blocks in cache via their internal DMA channels.

For RAID-10 LDEVs there will be two write operations, potentially on two BE Modules, read directly from the User Data region of cache by the BE Module(s). For RAID-5 LDEVs, there will be two reads (old data, old parity) and two writes (new data and new parity), probably on two BE Modules. For RAID-6 LDEVs there will be three reads (old data, old P, old Q) and three writes (new data, new P, new Q) using two or more BE Modules. [Note: RAID-6 has one parity chunk (P) and one redundancy chunk (Q) that is calculated with a more complicated algorithm.]

In the case of a RAID-5 or RAID-6 full-stripe write, where the new data completely covers the width of a RAID stripe, the read old data and read old parity operations are skipped. For a RAID-5 full-stripe write, there will only be two logical operations (write new data and write new parity). For a RAID-6 full-stripe write, there will only be three logical operations (write new data, write new P, write new Q).

The parity generation operation must be done on the same cluster as the owning MPU. For the case of a full stripe write, this could be as simple as the DRR emulator on the owning MPU just calculating new parity on the new data. Or it could become quite complex if some of the old data or old parity reside in the opposite side of cache in the other cluster. In this case, it could require one or more data transfer jobs utilizing the I-Paths to get everything into the owning MPU's side of cache before the DRR emulator could then calculate the new parity.

Failure Modes

This section explores the different component failure scenarios and describes the impact to system functionality in each case. Note that a single component failure will never cause a VSP Gxx0 array to go offline entirely.

Controller Blade Failure

In the case of a Controller Blade failure, all the components on that blade are blocked, for instance the MPU, Cache DIMMs, FE Modules, and BE Modules. Even though half of cache becomes inaccessible, the system continues to operate in write through mode without any data duplexing. Clean data in cache already exists on disk, so there is no data loss. For dirty data in cache, a second copy exists on the other Cache Side, so there is also no loss of user data.

LDEV ownerships will automatically be taken over by the MPU on the remaining Controller Blade. Each MPU is connected to a "partner" MPU in the other Cluster via I-Path and LDEV ownership is transferred based on these partnerships:

- MPU-10 ←→ MPU-20

I/O Module Failure

FE Module
If a FE Module port fails then the individual port is blocked and there is no impact to the other ports on the same module. If the entire module fails then all the ports will become blocked. In either scenario, if multipath software is used then there will be no impact to the attached hosts.

BE Module
If a SAS port fails on a BE Module it will become blocked and the microcode will use the SAS links on the other Cluster to access drives on the affected backend path. If the entire module fails then both SAS ports are blocked and the microcode will use the SAS links on the other Cluster to access drives on the affected backend paths.

DIMM Failure

When the system is running, microcode is stored in the Local Memory (LM) area of cache on each Cache Side. Since Local Memory is interleaved across all the installed DIMMs on the Controller Blade, a DIMM failure means the microcode becomes inaccessible on that Controller Blade. So a DIMM failure leads to the entire Controller Blade becoming blocked and the net effect is the same as a Controller Blade failure.

MP Failure (some but not all cores of a CPU)

If a few but not all cores of a CPU fail, then the failed cores are blocked and there is no other impact to the system, aside from the loss of processing capability of the failed core(s). Transfer of LDEV ownership is not necessary.

MP Failure (all cores of a CPU)

If all the cores of a specific CPU fail, then all the cores of the NTB-connected partner CPU on the other controller must also be blocked. For example if all cores on CPU-10 failed, then all cores on CPU-20 would be blocked. Blockading all cores of the partner CPU is required because the partner CPU has no way to communicate over NTB with the controller that has the failed CPU cores. Transfer of LDEV ownership is not necessary but note that failure of all CPU cores on a Controller Blade having a single CPU (G350, G370), is handled the same as a Controller Blade failure.

CPU Failure

If an Intel Xeon processor fails, it causes the loss of PCIe and memory buses. Without memory buses, memory interleave for the controller cannot be properly executed, so a CPU failure requires the entire controller to be blocked, with LDEV ownership transferred to the MPU on the remaining Controller Blade.

Drive Failure

If a drive fails, the Parity Group is placed in correction access mode and data on the failed drive is read or reconstructed from one or more remaining drives in the Parity Group. All LDEVs within the Parity Group will remain online and accessible by hosts.

Parity Group Failure

Multiple drive failures beyond the resiliency of the Parity Group’s RAID level will cause the Parity Group and its LDEVs to become blocked and user data will be lost. This could occur due to simultaneous drive failures, or the failure of other drives in the Parity Group during a dynamic sparing or correction copy (rebuild) process. If the failed Parity Group contains HDP/HDT pool volumes, then those pools and all their DP-VOLs will be blocked.

While this scenario is alarming, the use of spare drives and RAID-6 makes the chance of data loss in a production environment highly unlikely.

Drive Box Failure

Each drive box contains two ENCs, or enclosure circuit boards, each of which contains a SAS Expander and two SAS wide ports. If one ENC in a drive box fails, then all the backend paths daisy chained behind that ENC will be blocked. But since each drive box will have at least one functional backend path, all of the drive boxes will remain online.

If both ENCs or both PSUs fail, then the entire drive box becomes blocked. All drive boxes daisy chained behind the failed one will also become blocked. Parity Groups with drives in the blocked trays will remain online if RAID integrity can be maintained. Parity Groups that cannot maintain RAID integrity will be blocked, as well as their LDEVs, but user data will remain intact.

On G/F700 and G/F900 with standard back end, PGs can be protected from a drive box failure by using RAID-10 2+2, or RAID 5 (3+1) or RAID 6 (6+2) with the members of the PG equally distributed across drive boxes on separate back end paths. The G/F900 with expanded back end can additionally provide this protection for all other supported RAID configurations.

Loss of Power

In the event of a planned shutdown, dirty data in cache will be written to disk and Shared Memory will be written to the CFM SSDs. But in the case of a total loss of power the entire contents of cache, including dirty data that is in the Write Pending queues, will be written to the CFM SSDs using battery power. When power is restored, the array will return to its state at the point of power loss and the onboard batteries will recharge.

Appendix A – Front-end Port Operations

I/O Request Limits, Queue Depths, and Transfer Sizes

Every Fibre Channel path between the host and the storage system has a specific maximum instantaneous request capacity, known as the ***Maximum I/O Request Limit***. This is the limit to the *aggregate* number of requests (host tags) being directed against the individual LDEVs mapped to a host port. For the VSP Gx00 and VSP Gxx0, the port limit is 1,024 tags, based on a firmware limitation of the Baker chip. (The Baker chip supports 2,048 active exchanges per port, but only half of these resources are available for target mode transactions).

In addition, each Controller Blade has a tag limit of 64k entries for all of the LDEVs managed by its MPU from all host ports in use. Note that the number of supported hosts per front-end port (FC, iSCSI) is 255.

LUN Queue Depth is the maximum number of outstanding I/O requests (tags) **per LUN** on the ***host*** port within the server. This is distinctly separate from the system's port ***Maximum I/O Request Limit***. A LUN queue depth is normally associated with random I/O workloads and high server thread counts since sequential workloads are usually operating with a very low thread count and large block size.

On the VSP Gxx0, the per-LDEV rule-of-thumb queue depth is 32 per port but can be much higher (no real limit) in the absence of other workloads on that FC port in use for that LDEV. There isn't a queue depth limit per LDEV as such, but the usable limit will depend upon that LDEV's Parity Group's ability to keep up with demand.

In the case of external (virtualized) LUNs, the individual LDEV queue depth per external path can be set from 2 to 128 in the ***Universal Volume Manager*** GUI. Increasing the default queue depth value for external LUNs from 8 up to 32 can often have a very positive effect (especially on response time) on OLTP-like workloads. The overall limits (maximum active commands per port and maximum active commands per external storage array) are not hard set values, but depend on several factors. Please refer to *Appendix A* of the *VSP G1000 External Storage Performance* paper for an in-depth explanation of how these values are calculated.

There is also a host tunable parameter at the driver level that controls the **maximum transfer size** of a single I/O operation. This is usually a per-port setting, but it also might be a per-host setting. The defaults for this can range from fairly small (like 32 KB) to something midsized (128 KB). The maximum is usually 1 MB - 16 MB. It is probably a best practice to set this to the largest setting that is supported on the host. The maximum transfer size that the VSP Gxx0 will accept on a host port is 16 MB. This value controls how application I/O requests get processed. If one used a large application block size (say 256 KB) and the port/LUN default was just 32 KB, then each request would be broken down (fragmented) into 8 * 32 KB requests. This creates additional overhead in managing the application request. The use of a large maximum transfer size such as 1 MB will often be readily apparent on the performance of the system.

External Storage Mode I/O

Each bidirectional FC or iSCSI port on the FE module can operate simultaneously in four modes: a **target** mode (accepting host requests), an **external** mode (acting as an HBA to drive I/O to an external subsystem being virtualized), as a **Replication Initiator (MCU)** mode (respond) or as a **Replication Target** (**RCU**) mode (pull).

The **external mode** is how other storage systems are attached to FE ports and virtualized. Front-end ports from the external (secondary) system are attached to some number of VSP Gxx0 FE ports as if the VSP Gxx0 was a Windows server. In essence, those front-end ports are operated as though they were a type of back-end port. LUNs from attached systems that are visible on these external ports are then remapped by the VSP Gxx0 out through other specified front-end ports that are attached to hosts. Each such external LUN will become an internal VSP Gxx0 LDEV that is managed by one of the MPUs.

As I/O requests arrive over host-attached ports on the VSP Gxx0 for an external LUN, the normal routing operations within the FE module occur. The request is managed by the owner MPU in (mostly) the Local Memory working set of the global Control Memory tables, and the data blocks go into that managing MPU's data cache region. But the request is then rerouted to one or more other front-end ports (not to BEs) that control the external paths where that LUN is located. The external system processes the request internally as though it were talking to a server instead of the VSP Gxx0.

The two **Replication MCU** and **RCU** port modes are for use with the Hitachi Universal Replicator and True Copy Sync software that connects two Hitachi enterprise systems together for replication.

Cache Mode Settings

The external port (and all LUNs present on it) may be set to either "Cache Mode ON" or "OFF" when it is configured. This controls the write behavior and the maximum transfer size. The "ON" setting has cache behavior identical to that of internal LDEVs, with immediate response to the server once the write data has been duplexed into cache, and uses up to a 256 KB transfer size when gathering write is effective. The "OFF" setting directs the managing MPU to hold off on the write completion message to the server until the write request is acknowledged by the *external* system. The "OFF" setting does not change any other cache behavior such as reads to the external LUNs, but it does tend to decrease the effective transfer size compared to "Cache Mode ON". However, this change makes a significant difference when writing to slower external storage that presents a risk for high write pending cases to develop. "Cache Mode ON" should normally not be used when high write pending rates are expected to occur in the external system. In general, the rule of thumb is to use "Cache Mode OFF".

Other External Mode Effects

Recall earlier that the Queue Depth for each external LUN can be modified to be between 2 and 128 in the Universal Volume Manager configuration screens. Increasing the default value from 8 up to 32 (probably the best choice overall), 64, or perhaps even 128 can often have a very positive effect (especially Response Time) on OLTP-like workloads. Along with this, note that the maximum overall I/O Request Limit for an external port is 384 (not 1,024). If cache mode is ON and the external subsystem cannot keep up with the host write rate, then data will start backing up in Write Pending in the VSP Gxx0. Once this reaches the limit, and there is no more available space in WP, then host I/O will be delayed until there is space in Write Pending.

The external port represents itself as a Windows server to the external system. Therefore, when configuring the *host type* on the ports on the virtualized storage system, it must be set to "Windows mode".

Appendix B – General Storage Concepts

Understand Your Customer's Environment

Before recommending a storage design for customers, it is important to know how the product will address the customer's specific business needs. The factors that must be taken into consideration include: ***Capacity, Performance, Reliability, Features and Cost*** with respect to the storage infrastructure component. The more data you have, the easier it is to architect a solution as opposed to just selling another storage unit. The types of storage configured, including the drive types, the number of Parity Groups and their RAID levels, as well as the number and type of host paths is important to the solution.

RAID Levels and Write Penalties

The VSP Gxx0 system currently supports RAID levels 5, 6, and 10. RAID-5 is the most space efficient of these three RAID levels.

RAID-5 is a group of drives (Parity Group or RAID Group) with the space of one drive used for the rotating parity chunk per RAID stripe (row of chunks across the set of drives). If using a 7D+1P configuration (7 data drives, 1 parity drive), then you get 87.5% capacity utilization for user data blocks out of that Parity Group.

RAID-6 is RAID-5 with a second parity drive for a second unique parity block. The second parity block includes all of the data chunks plus the first parity chunk for that row. This would be indicated as a 6D+2P construction (75% capacity utilization) if using 8 drives, or 14D+2P if using 16 drives (87.5% capacity utilization as with RAID-5 7D+1P). Note that in this case we use the term "parity" for the Q block in RAID-6, even though the calculation of the Q block is much more complicated mathematically.

RAID-10 is a mirroring and striping mechanism. First, individual pairs of drives are placed into a mirror state. Then two of these pairs are used in a simple RAID-0 stripe. As there are four drives in the Parity Group, this would be represented as RAID-10 (2D+2D) and have 50% capacity utilization.

The VSP Gxx0, like the VSP G1500 and earlier Hitachi enterprise subsystems, has a RAID-10 Parity Group type called 4D+4D. Although we use the term 4+4, in the VSP Gxx0, a 4+4 is actually a set of two 2+2 parity groups that are RAID stripe interleaved on the same 8 drives. Thus the maximum size LDEV that you can create on a 2+2 is the same as the maximum size LDEV that you can create on a 4+4, although on the 4+4 you can create two of them, one on each 2+2 interleave.

RAID-10 is not the same as RAID-0+1, although usage by many would lead one to think this is the case. RAID-0+1 is a RAID-0 stripe of N-disks mirrored to another RAID-0 stripe of N-disks. This would also be shown as 2D+2D for a 4-disk construction. However, if one drive fails, that RAID-0 stripe also fails, and the mirror then fails, leaving a user with a single unprotected RAID-0 group. In the case of real RAID-10, one drive of each mirror pair would have to fail before getting into this same unprotected state.

The factors in determining which RAID level to use are cost, reliability, and performance. *Table* ***B1*** shows the major benefits and disadvantage of each RAID type. Each type provides its own unique set of benefits so a clear understanding of your customer's requirements is crucial in this decision.

	RAID-10	RAID-5	RAID-6
Description	Data Striping and Mirroring	Data Striping with distributed parity	Data Striping with two distributed parities
Number of drives	4/8	4, 5, 7, 8	8, 14, 16
Benefit	Highest performance with data redundancy; higher write IOPS per Parity Group than with similar RAID-5.	The best balance of cost, reliability, and performance.	Balance of cost, with extreme emphasis on reliability
Disadvantages	Higher cost per number of physical drives	Performance penalty for high percentage of Random Writes	Performance penalty for all writes

Table B1: RAID Level Comparison

Another characteristic of RAID is the idea of "write penalty". Each type of RAID has a different back end physical drive I/O cost, determined by the mechanism of that RAID level. The table below illustrates the trade-offs between the various RAID levels for write operations. There are additional physical drive reads and writes for every application write due to the use of mirrors or XOR parity.

Note that larger drives are usually deployed with RAID-6 to protect against a second drive failure within the Parity Group during the lengthy drive rebuild of a failed drive.

SAS / SSD/FMD Drives	HDD IOPS per Host Read	HDD IOPS per Host Write
RAID-10	**1**	**2**
RAID-5	**1**	**4**
RAID-6	**1**	**6**

Table B2: RAID Write Penalties

	IOPS Consumed per host I/O request
RAID-10	**1 Data Write, 1 mirrored Data Write**
RAID-5	**2 reads (1 data, 1 parity), 2 writes (1 data, 1 parity)**
RAID-6	**3 reads (1 data, 2 parity), 3 writes (1 data, 2 parity)**

Table B3: Breakdown of RAID Level Write Costs

Parity Groups and Array Groups

On Hitachi enterprise products the terms **Parity Group** and **Array Group** are used when configuring storage. The Array Group is the set of four associated drives, while the Parity Group is the RAID level applied to one or more array groups. However, in general use in the field, the terms Array Group and Parity Group are often used interchangeably.

RAID Chunks and Stripes

The Parity Group is a logical mechanism that has two basic elements: a virtual block size from each drive (a **chunk**) and a row of chunks across the group (the RAID **stripe**). The cache slot size is 256 KiB on the VSP Gxx0 and VSP G1500, but two such contiguous cache slots per drive comprise a **512 KiB** chunk. The stripe size is the sum of the chunk sizes across a Parity Group. This only counts the “data” chunks and not any mirror or parity space. Therefore, on a RAID-6 group created as 6D+2P (eight drives), the stripe size would be 3,072 MiB (512 KiB x 6).

Note that some industry usage replaces *chunk* with “stripe size”, “stripe depth”, or “interleave factor”, and *stripe size* with “stripe width”, “row width” or “row size”.

Note that on all current RAID systems, the chunk is a primary unit of protection and layout management: either the parity or mirror mechanism. I/O is not performed on a chunk basis as is commonly thought. On Open Systems, the entire space presented by a LUN is a contiguous span of 512 byte blocks, known as the Logical Block Address range (LBA). The host application makes I/O requests using some native request size (such as a file system block size), and this is passed down to the storage as a unique I/O request. The request has the starting address (of a 512 byte block) and a length (such as the file system 8 KB block size). The storage system will locate that address within that LUN to a particular drive address, and read or write only that amount of data – not that entire chunk. Also note that this request could require two drives to satisfy if 2 KB of the block lies on one chunk and 6 KB on the next one in the stripe.

Because of the variations of file system formatting and such, there is no way to determine where a particular block may lie on the raw space presented by a volume. A file system will create a variety of metadata in a quantity and distribution pattern that is related to the size of that volume. Most file systems also typically scatter writes around within the LBA range – an outdated hold-over from long ago when file systems wanted to avoid a common problem of the appearance of bad sectors or tracks on drives. So while it may not be possible to align application block sizes with RAID chunk sizes, it is recommended practice to align disk partitions to RAID chunk boundaries where possible. In fact, most operating systems do this automatically by offsetting the first partition created on a volume by a fixed size (often 1 MiB) and rounding the requested partition size to an “even” block boundary.

The one alignment issue that should be noted is in the case of host-based Logical Volume Managers. These also have a native “stripe size” that is selectable when creating a logical volume from several physical storage LUNs. In this case, the LVM stripe size should be a multiple or divisor of the RAID chunk size due to various interactions between the LVM and the LUNs. One such example is the case of large block sequential I/O. If the LVM stripe size is equal to the RAID chunk size, then a series of requests will be issued to different LUNs for that same I/O, making the request appear to be several random I/O operations to the storage system. This can defeat the system’s sequential detect mechanisms, and turn off sequential prefetch, slowing down these types of operations.

LUNs (host volumes)

On a VSP Gxx0, when space is carved out of a Parity Group and made into a volume, it is then known as a Logical Device (LDEV). Once the LDEV is mapped to a host port for use by a server, it is known as a LUN (Logical Unit Number). On an iSCSI configuration, the LUN is identified by a target iSCSI Qualified Name (IQN). Note that general usage has turned the term of “LU” into “LUN”.

Number of LUNs per Parity Group

When configuring a VSP Gxx0, one or more LDEVs may be created per Parity Group, but the goal should be to clearly understand what percentage of that group’s overall capacity will contain active data. In the case where multiple hosts attempt to simultaneously use LUNs that share the same physical drives in an attempt to fully utilize capacity, seek and rotational latency may be a performance limiting factor. In attempting to maximize utilization, Parity Groups should contain both active and less frequently used LUNs. Note that the discussion of LUNs per Parity Group applies primarily to LDEVs addressed directly by hosts, and not those that contribute physical capacity to a pool. HDP stripes host workloads across multiple Parity Groups, which reduces the likelihood of drive bottlenecks.

Mixing Data on the Physical drives

Physical placement of data by Parity Groups (not just the LDEVs from the same Parity Group) is extremely important when the data access patterns differ. Mixing highly “write intensive” data with high “read intensive” data will cause both to suffer performance degradation. This performance degradation will be much greater when using RAID-5 or RAID-6 due to the increase in back end drive operations required for writes. Remember there are two physical I/Os required for each random write for RAID-10, four for RAID-5, and six for RAID-6. Again, the preceding caution is most applicable to LDEVs addressed directly by a host, rather than those that contribute physical capacity to a pool.

Selecting the Proper Disks

In all cases, distributing a workload across a higher number of small-capacity, high-RPM drives will provide better performance in terms of full random access. Even better results can be achieved by distributing the workload over a higher number of small LUNs where the LUNs are the only active LUNs in the Parity Group. When cached data locality is low, flash drives or multiple small-capacity high-RPM drives should be used.

One must also take into consideration the case where a system that is only partially populated with 15K RPM drives will be able to provide a much higher aggregate level of host IOPS if the same budget is applied to lower cost 10K RPM drives. For instance, if there is a 100% increase in the cost of the 15K drives, then one could install twice as many drives of the 10K variety. The individual I/O will see some increase in response time when using 10K drives, but the total IOPS available will be much higher.

Mixing I/O Profiles on the Physical drives

Mixing large-block sequential I/O with small-block random I/O on the same physical drives can result in poor performance. This applies to both read and write I/O patterns. This problem can occur at the LUN level or Parity Group level. In the first case, with a single large LUN, files with different I/O profiles will result in poor performance due to lack of sequential detect for the large block sequential I/O requests. In the second case, where multiple LUNs share a Parity Group, files having different I/O profiles will result in sequential I/O dominating the drive access time, due to pre-fetching, thereby creating high response times and low IOPS for the small-block random I/O. The resolution to these two issues is to use different LUNs and different Parity Groups for different data types and I/O profiles.

Front end Port Performance and Usage Considerations

The flexibility of the front end ports is such that several types of connections are possible. A couple of port usage points are considered:

- Port fan-in and fan-out
- I/O profile mixes on a port

Host Fan-in and Fan-out

With a SAN, the ability to share a port or ports among several host nodes is possible, as is configuring a single host node to multiple storage ports. Fan-in is one of the great promises of fibre channel based SAN technologies—the sharing of costly storage resources. Several host nodes are fanned *into* a single storage port, or a host node has fanned out to several storage ports.

Fan-in

Host Fan-in refers to the consolidation of many host nodes into one or just a few storage ports (many-to-one). Fan-in has the potential for performance issues by creating a bottleneck at the front end storage port. Having multiple hosts connected to the same storage port does work for environments that have minimal performance requirements. In designing this type of solution, it is important to understand the performance requirements of each host. If each host has either a high IOPS or throughput requirement, it is possible that a single 16 Gbps FC port will not satisfy their aggregate performance requirements.

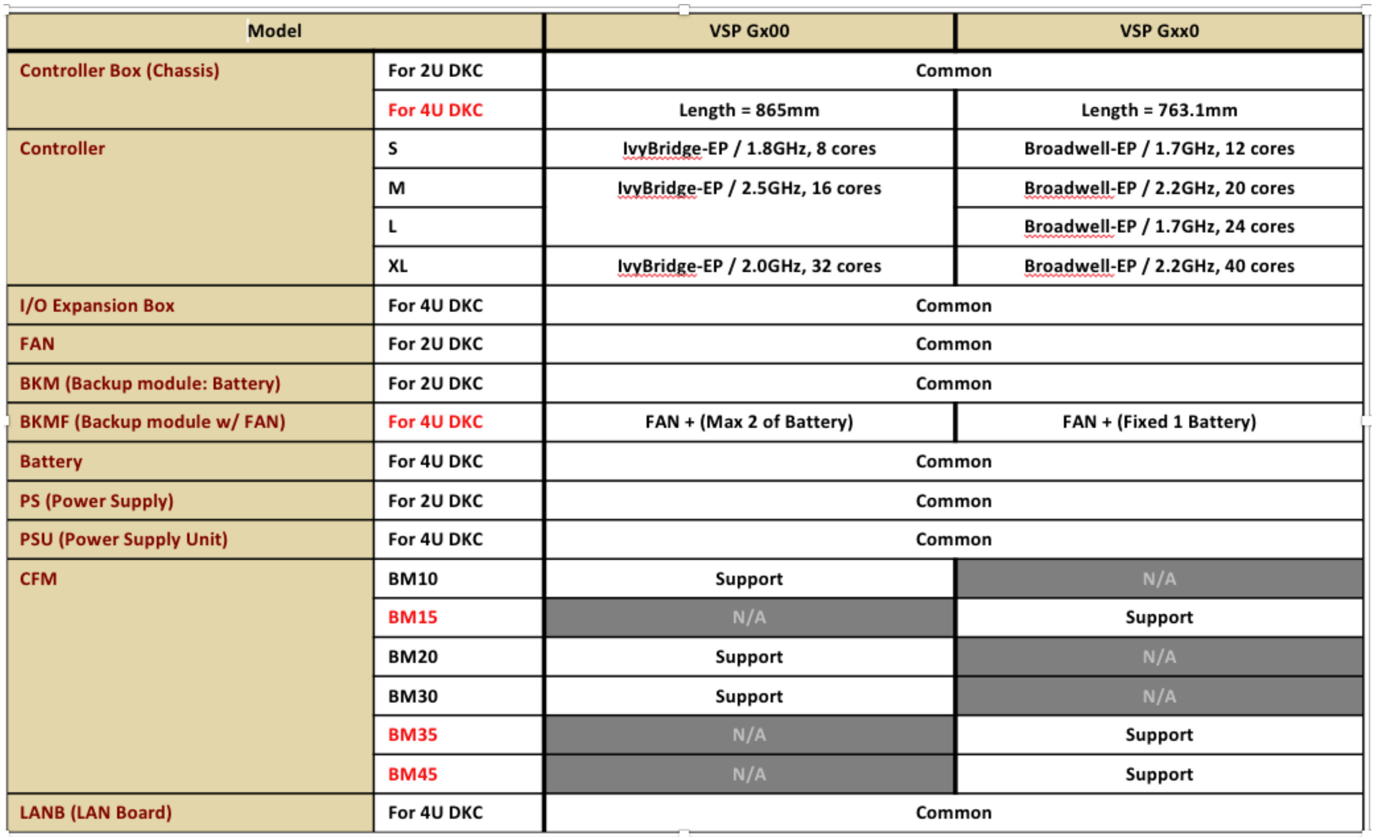
Fan-out

Fan out allows a host node to take advantage of several storage ports from a single host port (one-to-many). Fan-out has a potential performance benefit for small block random I/O workloads. This allows multiple storage ports (and their queues) to service a smaller number of host ports. Fan-out typically does not benefit environments with high throughput (MB/sec) requirements due to the transfer limits of the host bus adapters (HBAs).

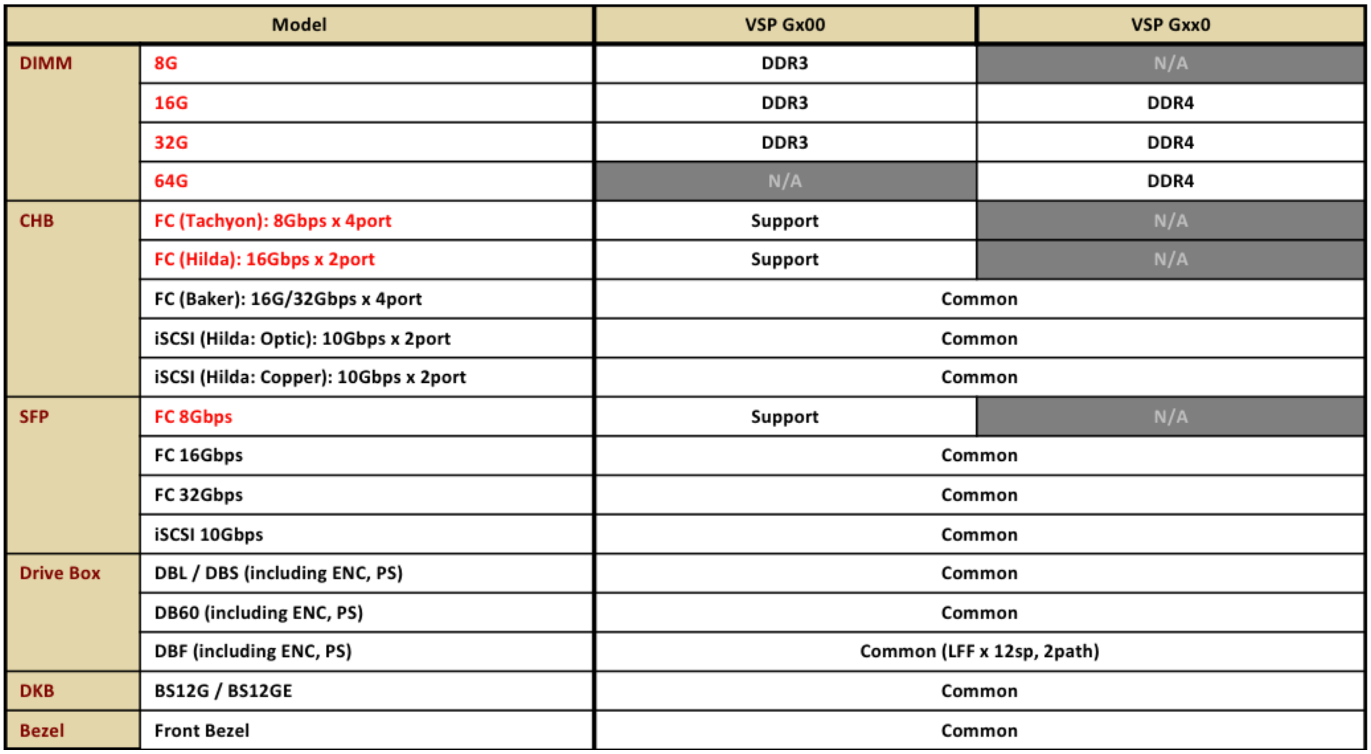
Mixing I/O Profiles on a Port

Mixing large-block sequential I/O with small-block random I/O on the same storage port can result in reduced performance. Because fibre channel is a serial protocol, only one transfer at a time can take place on an FC port. Large-block transfers keep the FC port occupied for much longer than small-block transfers do, simply because the data transfer time increases linearly with data transfer length. So for example, a 256KB transfer keeps an FC port busy for 64 times as long as a 4KB transfer. Because the FC port by default has no way to prioritize some requests over others, the large-block sequential workload will tend to dominate the port, thus increasing response times for small-block random workloads sharing the same port. But large-block sequential throughput will also tend to degrade on a port shared with a small-block random access pattern, since small-block requests can arrive with relatively high frequency and must be interleaved with the large-block transfers. This problem can be compounded when HBAs of different speeds connect to the same target port, since the effective transfer rate of the target port may be reduced during intervals when a slower HBA is transferring data. For optimal performance, assign a large block workload to its own storage port, or to a port shared only with similar workloads.

Appendix C – Common Parts List



Model		VSP Gx00	VSP Gxx0
Controller Box (Chassis)	For 2U DKC	Common	
	For 4U DKC	Length = 865mm	Length = 763.1mm
Controller	S	IvyBridge-EP / 1.8GHz, 8 cores	Broadwell-EP / 1.7GHz, 12 cores
	M	IvyBridge-EP / 2.5GHz, 16 cores	Broadwell-EP / 2.2GHz, 20 cores
	L		Broadwell-EP / 1.7GHz, 24 cores
	XL	IvyBridge-EP / 2.0GHz, 32 cores	Broadwell-EP / 2.2GHz, 40 cores
I/O Expansion Box	For 4U DKC	Common	
FAN	For 2U DKC	Common	
BKM (Backup module: Battery)	For 2U DKC	Common	
BKMF (Backup module w/ FAN)	For 4U DKC	FAN + (Max 2 of Battery)	FAN + (Fixed 1 Battery)
Battery	For 4U DKC	Common	
PS (Power Supply)	For 2U DKC	Common	
PSU (Power Supply Unit)	For 4U DKC	Common	
CFM	BM10	Support	N/A
	BM15	N/A	Support
	BM20	Support	N/A
	BM30	Support	N/A
	BM35	N/A	Support
	BM45	N/A	Support
LANB (LAN Board)	For 4U DKC	Common	



Model		VSP Gx00	VSP Gxx0
DIMM	8G	DDR3	N/A
	16G	DDR3	DDR4
	32G	DDR3	DDR4
	64G	N/A	DDR4
CHB	FC (Tachyon): 8Gbps x 4port	Support	N/A
	FC (Hilda): 16Gbps x 2port	Support	N/A
	FC (Baker): 16G/32Gbps x 4port	Common	
	iSCSI (Hilda: Optic): 10Gbps x 2port	Common	
	iSCSI (Hilda: Copper): 10Gbps x 2port	Common	
SFP	FC 8Gbps	Support	N/A
	FC 16Gbps	Common	
	FC 32Gbps	Common	
	iSCSI 10Gbps	Common	
Drive Box	DBL / DBS (including ENC, PS)	Common	
	DB60 (including ENC, PS)	Common	
	DBF (including ENC, PS)	Common (LFF x 12sp, 2path)	
DKB	BS12G / BS12GE	Common	
Bezel	Front Bezel	Common	

Hitachi Vantara

Corporate Headquarters
2845 Lafayette Street
Santa Clara, CA 95050-2639 USA
www.HitachiVantara.com | community.HitachiVantara.com

Regional Contact Information
Americas: +1 866 374 5822 or info@hitachivantara.com
Europe, Middle East and Africa: +44 (0) 1753 618000 or info.emea@hitachivantara.com
Asia Pacific: +852 3189 7900 or info.marketing.apac@hitachivantara.com

WP-xxx-x Author First Initial/Last Name Month 20xx (Note to Author: this information will be completed by Marcom during the review process.