Huawei OceanStor Dorado Storage Systems

The Huawei OceanStor Dorado V6 all flash storage family consists of the 3000, 5000, 6000, 8000, and 18000. The OceanStor Dorado line is similar to the Huawei OceanStor storage systems, with the Dorado designation signaling all flash systems. The V6 systems are the newest generation of Dorado systems, bringing performance increases and significant technology updates to the previous V3 systems.

The Dorado systems are targeted at accelerating applications such as databases, virtual desktop, and server virtualization. The Dorado (officially OceanStor Dorado) line uses Huawei SSDs attached by 12 Gb/s SAS or by PCIe using the NVMe protocol.

Dorado models feature a modular architecture with 2U or 4U controller enclosures, depending on the model, and additional device enclosures for either SAS or NVMe SSDs. For NVMe devices, Huawei uses a custom physical form to allow more SSDs compared to a standard 2.5" SAS SSDs.

In addition to custom designed SSDs, Huawei has implemented a full set of custom chips in the OceanStor Dorado V6 models. V6 models utilize a total of five specialized chips that serve a variety of purposes.

The Huawei Dorado all flash systems compete other all flash block storage systems including EMC Unity all flash models, IBM FlashSystem, NetApp All Flash FAS, HPE Primera, PureStorage FlashArray and other leading all flash systems.

Highlights

- All solid state storage systems using Huawei SSDs
- Dual active-active controllers – scalable in pairs to a maximum of 16 (32 for 18000) controllers.
- Block support with Fibre Channel, InfiniBand and iSCSI
- SAS and PCIe interface device attachment
- SCSI and NVMe protocol support for device attachment
- Remote copy (synchronous or asynchronous) and consistency group support
- Snapshot (point-in-time copy) capabilities supporting LUN snapshots
- Thin provisioning and replication
- VMware VAAI support
- Stretched Clusters
- Virtualization of 3rd party attached storage
- QoS
- Deduplication
- Compression

Overview of System

The Huawei Dorado all flash storage systems support block storage access over FC, and iSCSI. The dual controller configuration is active-active for performance and high availability. The controllers are housed in a 2U enclosure with storage devices for the 3000, 5000, and 6000 models. The 2U controller enclosures can house up to 25 SAS SSDs or, for the 5000 and 6000 models, up to 36 NVMe SSDs. The 8000 and 18000 models house two or four controllers in a 4U enclosure without storage devices.

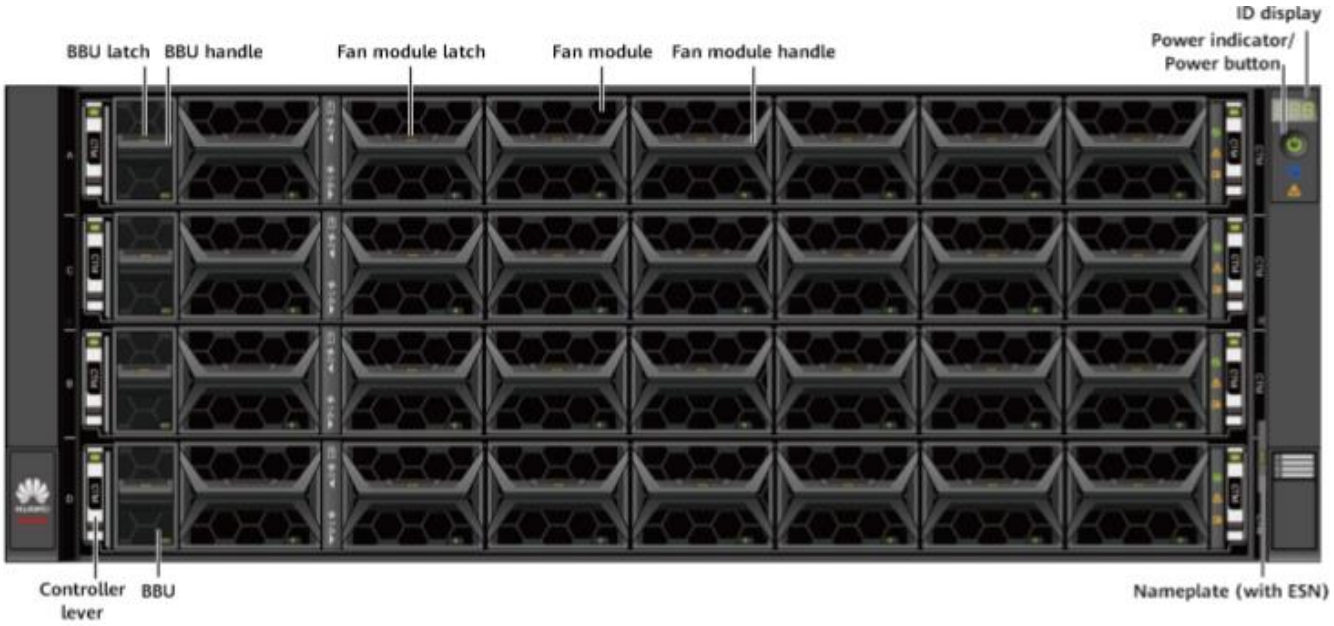

Figure 1: Huawei Dorado 4U Enclosure Front View (Source: Huawei)

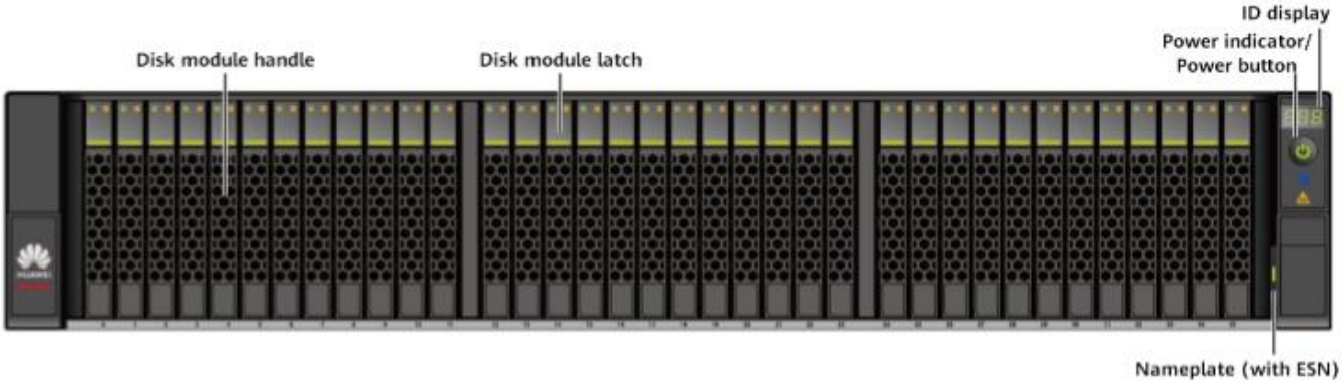

Figure 2: Huawei Dorado 2U Enclosure with 36 NVMe SSDs (Source: Huawei)

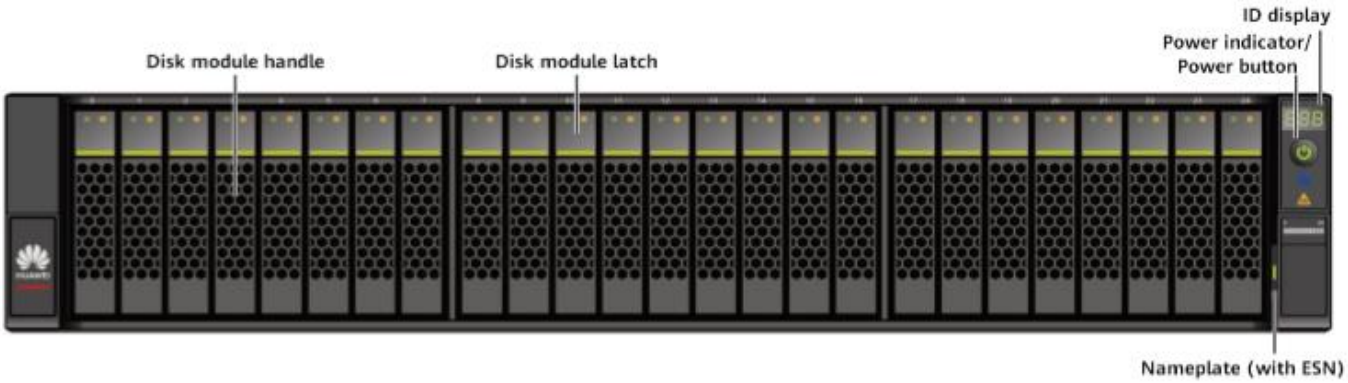

Figure 3: Huawei Dorado 2U Enclosure with 24 SAS SSDs (Source: Huawei)

Huawei Dorado Hardware Architecture

The Huawei Dorado V6 systems received significant hardware changes compared to the previous V3 systems. The hardware architecture varies slightly between the high-end, midrange, and entry level Dorado models.

The high end 18000 and 8000 models utilize 4U independent controller enclosures, capable of holding up to four controllers. Each enclosure contains 28 interface module slots which are shared amongst the controllers and two power planes. The system is designed to be resilient and is capable of handling a failure of three out of the four controllers. Each controller supports up to four of Huawei's custom designed Kunpeng 920 processors.

The high-end models feature front end connectivity support for FC, FC-NVMe, and ethernet and back end connectivity via 12 Gb/s SAS or 100 Gb/s RDMA depending on the device enclosures used. The systems' scale out interface modules are 2 port 100 Gb/s RDMA interface modules.

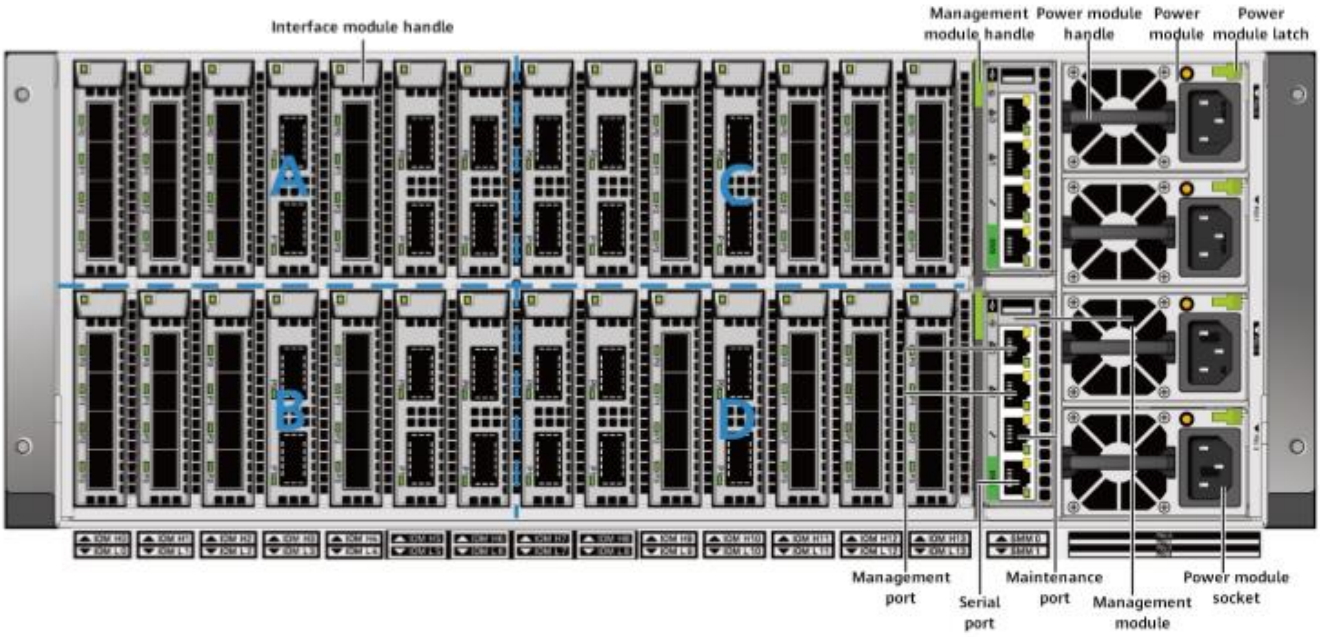

Figure 4: Huawei Dorado 4U Enclosure Rear View (Source: Huawei)

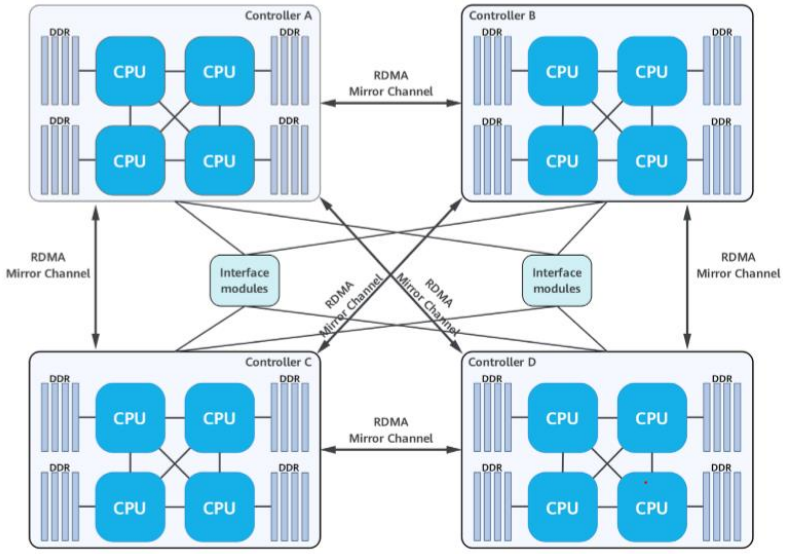

Figure 5: Huawei Dorado High End System Logical Diagram (Source: Huawei)

Midrange Dorado V6 systems, the 5000 and 6000, utilize a 2U enclosure consisting of two controllers and SSD's. The enclosures can hold 25 2.5" SAS SSDs or 36 NVMe SSDs that are custom designed by Huawei to provide this higher density. The midrange model controllers feature two Kunpeng 920 processors each.

Similar to the high-end enclosures, the midrange enclosures support host connection via FC, FC-NVMe, and ethernet, along with 12 Gb/s SAS or 100Gb/s RDMA back end connectivity. The midrange models use 4 port 25 Gb/s RDMA module interfaces for scale out interfaces.

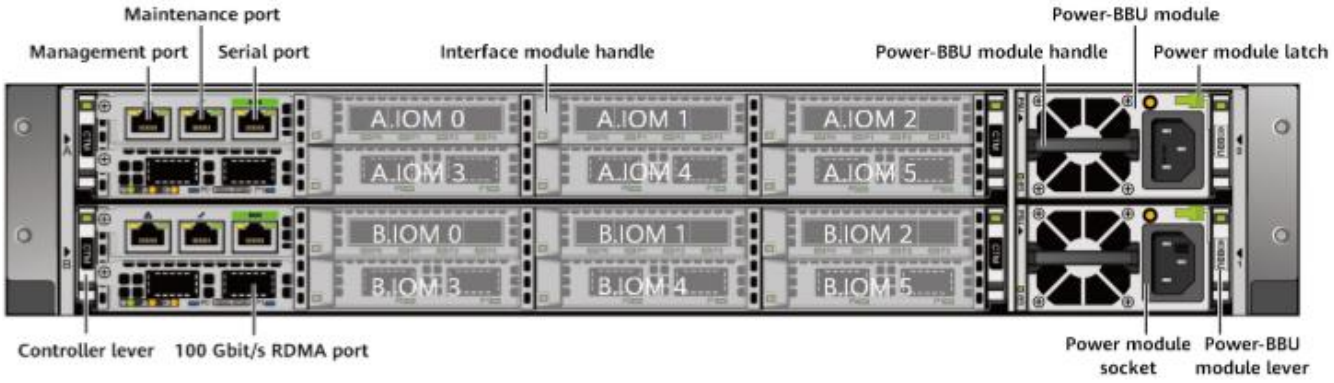

Figure 6: Huawei Dorado 2U Enclosure Rear View (Source: Huawei)

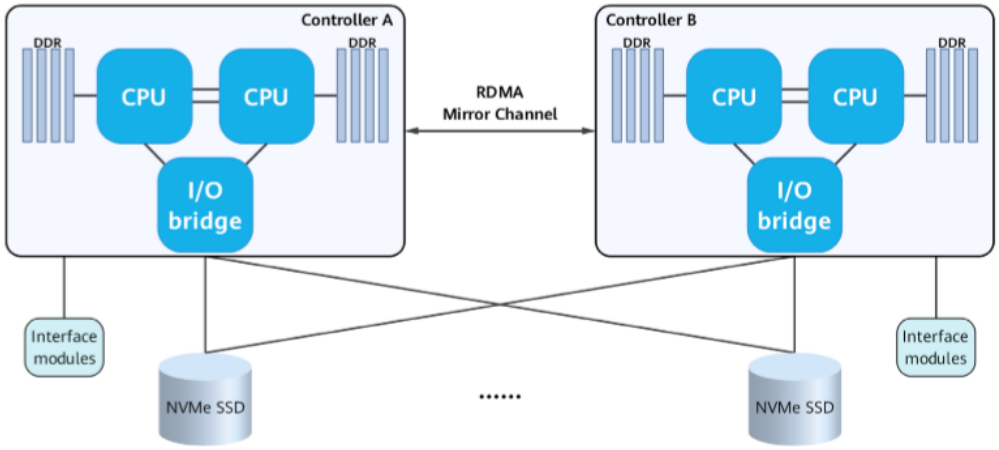

Figure 7: Huawei Dorado Midrange Logical Architecture (Source: Huawei)

The Huawei Oceanstor Dorado entry level system, the 3000 model, supports a 2U controller enclosure capable of holding two controllers and up to 25 SAS SSDs. This model does not support NVMe devices. Connectivity options are similar to the midrange models with the exception that only 12 Gb/s SAS connectivity is supported for the back end. The 3000 model controllers feature one Kunpeng 920 processor each.

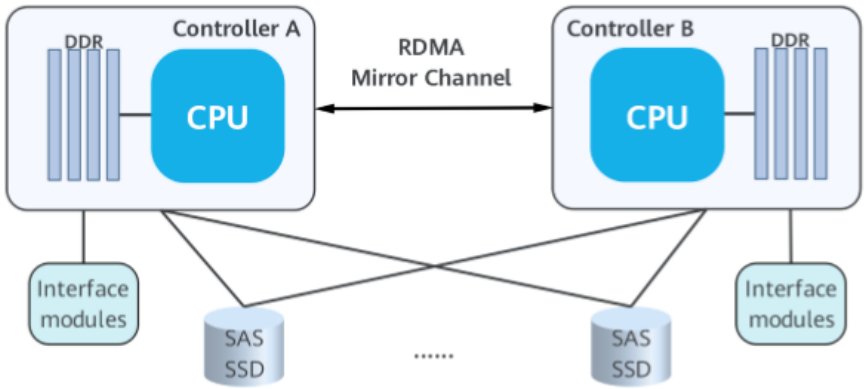

Figure 8: Huawei Dorado Entry Level Logical Architecture (Source: Huawei)

Huawei OceanStor Dorado V6 5000, 6000, 8000, and 18000 models support three options for device enclosures: a standard SAS SSD enclosure, a smart SAS SSD enclosure, and a smart NVMe SSD enclosure. The entry level 3000 model only supports the standard SAS SSD enclosure, which connects via 12 Gb/s SAS. The standard SAS SSD enclosure is 2U and supports up to 25 SAS SSDs.

The "smart" designation of the additional two device enclosure options refers to the addition of a Kunpeng 920 processor and DDR memory included to offload some of the computation from the controllers. The smart SAS and smart NVMe device controllers connect via 100 Gb/s RDMA and support up to 25 SAS SSDs or 36 NVMe SSDs respectively.

Model Comparison

Table 1 lists some of the specifications and characteristics of the Huawei Dorado system.

Model / Capability	3000 V6	5000 V6	6000 V6	8000 V6	1800 V6
Max Controllers	16	16	16	16	32
No. of Devices	1,200	1,600	2,400	3,200	6,400
Max Capacity - raw	36.8 PB	49.2 PB	73.7 PB	98.3 PB	196.61 PB
SAS Devices Supported	SAS SSD: 960 GB / 1.92 TB / 3.84 TB / 7.68 TB / 15.36 TB / 30.72	SAS SSD: 960 GB / 1.92 TB / 3.84 TB / 7.68 TB / 15.36 TB / 30.72 TB	SSD: 960 GB / 1.92 TB / 3.84 TB / 7.68 TB / 15.36 TB / 30.72 TB	SSD: 960 GB / 1.92 TB / 3.84 TB / 7.68 TB / 15.36 TB / 30.72 TB	SSD: 960 GB / 1.92 TB / 3.84 TB / 7.68 TB / 15.36 TB / 30.72 TB

NVMe Devices Supported	None	NVMe SSD: / 1.92 TB / 3.84 TB / 7.68 TB / 15.36 TB	NVMe SSD: / 1.92 TB / 3.84 TB / 7.68 TB / 15.36 TB	NVMe SSD: / 1.92 TB / 3.84 TB / 7.68 TB / 15.36 TB	NVMe SSD: / 1.92 TB / 3.84 TB / 7.68 TB / 15.36 TB
Front End Connectivity	8/16/32 Gb/s FC/NVMe, 10/25/40/100 GigE	8/16/32 Gb/s FC/NVMe, 10/25/40/100 GigE	8/16/32 Gb/s FC/NVMe, 10/25/40/100 GigE	8/16/32 Gb/s FC/NVMe, 10/25/40/100 GigE	8/16/32 Gb/s FC/NVMe, 10/25/40/100 GigE
Back End Connectivity	12 Gb/s SAS	12 Gb/s SAS or 100 GB/s RDMA	12 Gb/s SAS or 100 GB/s RDMA	12 Gb/s SAS or 100 GB/s RDMA	12 Gb/s SAS or 100 GB/s RDMA
Cache	192 – 1,536 GB	256 GB – 4 TB	1 TB - 8 TB	512 GB – 16 TB	512 GB – 32 TB

Table 1: Huawei Dorado V6 Overview

RAID Level Support

The Huawei Dorado system supports multiple RAID levels, including RAID 5 (striped parity group), RAID 6 (dual parity), RAID10, and RAID TP (triple parity).

Evaluator Group Comments: An important distinction is how RAID levels are constructed and utilized. The Dorado systems have a storage pooling implementation added as another layer on top of the RAID groups. This allows allocation of chunks from devices with dynamic allocation.

Volume Configuration

The system software divides physical devices into logical chunks of up to 64KB for allocation from the underlying RAID groups. The stripe size is based on the size of the RAID group. The chunk size can be specified on a per LUN basis by the administrator to match the application usage in allocation.

Huawei Custom Chips

The Huawei OceanStor Dorado V6 models contain a number of custom designed chips to perform different functions. These chips were custom designed by Huawei and include a front-end interface chip (Hi1822), the Kunpeng 920 chip, an Ascend AI chip (Ascend 310), an SSD controller chip, and a baseboard management controller (BMC) chip (Hi1710). These chips are all designed to improve performance and efficiency in their respective areas.

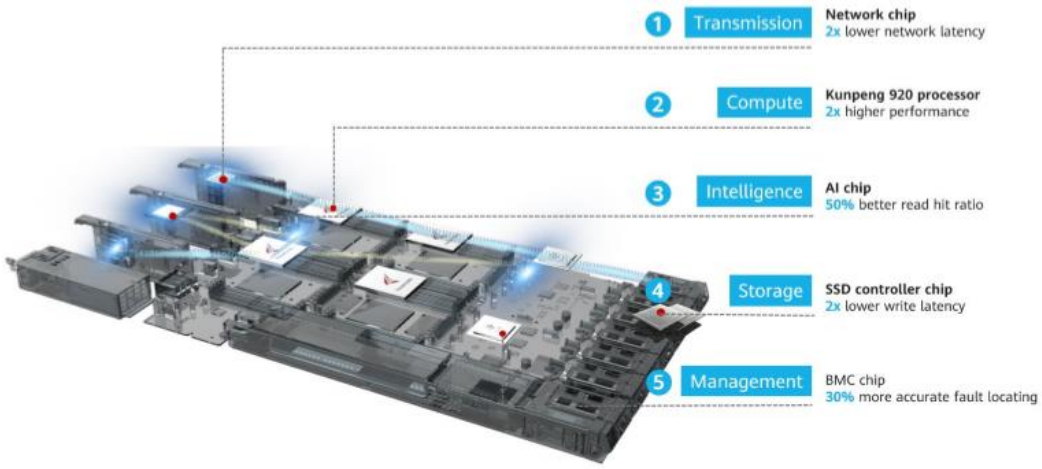

Figure 9: Huawei Custom Chips (Source: Huawei)

Connectivity Options

Dorado systems provide Fibre Channel, InfiniBand and iSCSI connectivity options.

Below is the maximum number of front end ports supported per controller for each model:

- 3000 V6: 40 ports
- 5000 V6: 48 ports
- 6000 V6: 56 ports
- 8000 V6: 104 ports
- 18000 V6: 104 ports

FC Connectivity

The V6 models support 8, 16, or 32 Gbit/s FC or FC-NVMe .

InfiniBand

The InfiniBand interface supports QDR – 56 Gb/s connectivity.

iSCSI Connectivity

10/20/40/100 GigE Ethernet is supported. Each port provides TCP/IP offloading in hardware termed TOE.

Dedicated Replication Ports

No dedicated ports are available for remote replication between systems, supporting both FC and iSCSI over Ethernet physical connections.

Huawei OceanStor Software Overview

The Huawei OceanStor operating system for the Dorado systems is the generation from the "T" series of block storage systems. The OceanStor operating system, along with associated value added software support a variety of data protection and management features.

Each controller in the storage system runs its own independent copy of the operating system. However, the OceanStor DeviceManager manages the controllers as a single system and provides all fault tolerance features utilizing built in hardware redundancy.

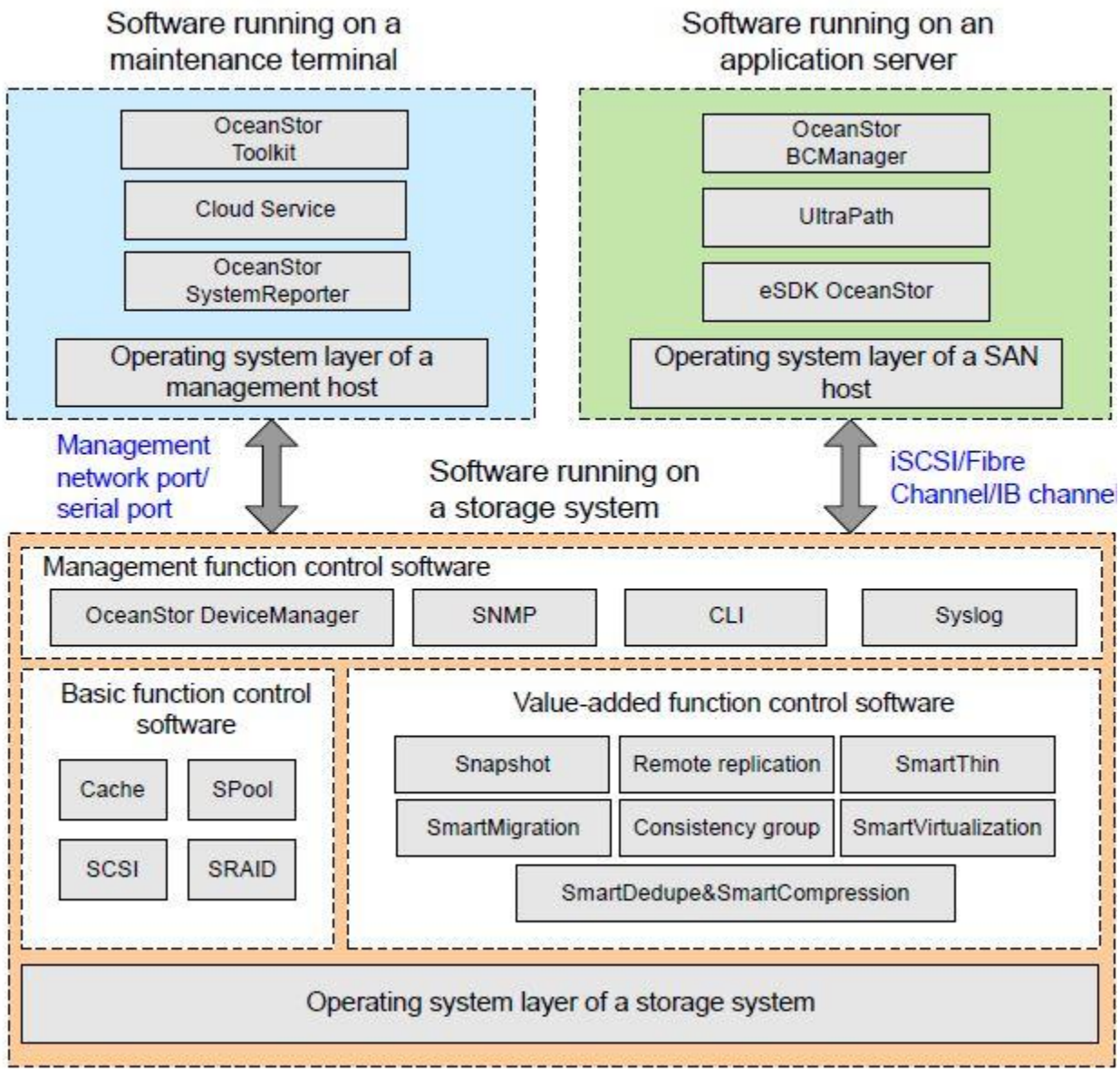

Figure 2: Huawei OceanStor Dorado Software Architecture (source: Huawei)

OceanStor Operating System Storage Pooling

RAID 2.0+ - Storage Pooling and Device Virtualization

Physical devices are divided into logical chunks of up to 64KB from which RAID groups are created, which may then be allocated as a volume or LUN or filesystem. As a result, data is spread across more physical devices, with fewer limitations than with RAID groups comprised of entire disk drives where it is possible to perform wide striping, but only with large amounts of total storage. Huawei calls this level of storage pooling and virtualization RAID 2.0+.

A chunk is called a "grain" by Huawei and is user settable on a per-filesystem or per LUN (thin-provisioned LUN) basis. The user setting can range from 4KB to 64KB is usually set based on the application usages such as file system allocations.

Storage Pooling	Support
Max LUN's / System	65,536
Max LUN's / port	255
Max number of storage pools	128
Max number of LUNs in a storage pool	8,192
Max LUN size	256 TB

Table 2: OceanStor LUN Support

SmartThin - Thin Provisioning

Volumes with capacity provisioned as needed are called thinly provisioned. Huawei uses the term SmartThin for the OceanStor implementation of thin provisioning. The RAID 2.0+ storage pooling implementation allows chunks to be allocation for capacity when needed for thin provisioned volumes. The allocation has a minimum size, which is 64KB regardless of the chunk size. Because the allocation of space is dynamic – when space is required – mapping tables are used for directing access to the blocks in a volume.

Space reclamation for allocated space that is no longer in use (deleted files or volumes for example) is supported with two methods. One is using the SCSI UNMAP command where the application, file system, or operating system notify the storage system to unmap specific blocks which are then returned to the storage pool as free space. The other method is for the OceanStor system to detect all-zero writes (zero-blocks) and unmaps those blocks from the mapping tables and return completion notification to the host.

SmartVirtualization – External Storage System Virtualization

In addition to virtualizing the underlying physical storage, OceanStor can also virtualize attached third-party storage systems, which is similar to other solutions such as the IBM Storwize V7000 and Hitachi VSP. Third party (or Huawei) storage systems are attached to Fibre Channel interfaces of the controllers, which become initiators rather than targets. This does reduce the number of available host connections.

An externally attached system is configured with LUNs managed by the OceanStor system. These LUNs are called eDevLUNs to distinguish them from local system LUNs. The maximum number of external LUNs supported is 1,024 with the 6800 V3. Up to 256 external storage systems may be virtualized. There is a maximum of eight paths to each external LUN. External LUNs and local LUNs have the same properties so other features such as replication, tiering, snapshots, etc. all work the same.

SmartMigration – Migration of LUNs

LUNs can be migrated between physical devices without interruption of access. The LUN migration defines source and targets that can be local logical devices or external LUNs from a virtualized storage system. When the migration completes, all accesses will be transferred to the new physical location of the LUN automatically. The migration between virtualized storage systems and local storage can be for various reasons including draining data from the external system, converting thickly provisioned volumes to thinly provisioned volumes, applying data reduction to a LUN (compression and deduplication), and for performance quality of service reasons. Capacity balancing and load balancing also use SmartMigration as the means to move data between physical devices.

Reliability changes by migrating LUNs to different RAID protection levels are also accomplished using SmartMigration.

Performance-Focused Software Features

OceanStor has a number of features that have direct effect on the system performance. Some are basic to the system while others are configurable or add-on software with additional licensing charges. The hardware implementations with direct impact on performance were discussed in the hardware section and include:

- Processors – number and type
- Memory used for caching
- PCIe switch for controller interconnect
- Interfaces – front (Host) end and back end connections

The scale-out architecture is also a major contributor for performance both in the ability to scale performance as more controllers are added to the configuration and in the ability to minimize performance impact during failure situations or system updates.

The additional features that can improve performance are described in the remainder of this section.

SmartQoS - Quality of Service

Resources in the storage system can be controlled to provide a quality of service for defined LUNs using the feature SmartQoS. SmartQoS controls I/O resource by managing priorities that are defined by the user for the selected LUNs. Three areas of controls are defined:

- I/O priority scheduling – allocation of system resources is given priority based on settings of high, medium, and low.
- I/O traffic control – user defined IOPs or bandwidth goals will cause some LUN operations to be restricted so that QoS controlled LUNs can meet their goals.
- I/O performance assurance – minimum performance settings can be set where other operations can be delayed until the selected minimums are met.

Resources managed by QoS include front-end interface, processor execution, cache, and device access.

Data Protection

HyperSnap - Snapshot

A snapshot is a virtual copy of a LUN or filesystem created with a mapping table and a copy-on-write implementation. Snapshot copies are instantly available.

HyperMirror - Remote Replication

Synchronous and asynchronous remote replication to a remote OceanStor system is performance with the replication software. Remote replication creates a mirror on a volume by volume basis. Periodic asynchronous (change data only) is not currently supported. ReplicationDirector manages the remote replication function.

Evaluator Group Comment: Setting up a remote replicated environment is very complicated and needs to be well-understood to gauge the ability to protect from disasters. The information provided publicly is limited so either additional information must be obtained or the implementation must be done by the services group with the knowledge required.

Consistency Groups

Consistency groups are supported with Asynchronous and Synchronous replication, which ensures that all data on the marked logical volumes are consistent to a specific point in time by managing the replication operation in order defined batches. Consistency groups are defined in ReplicationDirector and based on volumes and not the host, which the volumes are mapped to, providing consistency across multiple hosts and applications. Both synchronous and asynchronous replication supports consistency groups, which enables applications that utilize multiple volumes the ability to maintain application consistency when replicating between systems.

A summary of the data protection products is provided in the following table.

	Snapshot	Remote Copy
Copy Type	Copy on Write	Sync and Async
Architecture	Bit Map	I/O copy
Max # Copies	256 total	32 replication pairs
Re-sync	No	Yes (w/ Async)
Potential Data Exposure	Exposure to time activated	None with Sync, time between re-sync for Async
Availability for Restore	Available for restore immediate, or able to replace volume after copy complete	Sync: immediate w/ failover
Disaster Recovery, Fail-over Capability	Yes Able to replace volume after command completes	Nearly instantaneous for sync and async, auto-failover is supported
Default Action	Copy when command issued	Depends upon mode
Additional Capacity	Variable	100%
Access to Copy	R/W after snap	Only after failover
Access to Source	R/W after complete	Full access

Table 2: OceanStor Dorado Data Protection Options Comparison

Integration with Application or System Software

OceanStor and VMware vSphere

The OceanStor has a plug-in for vSphere vCenter to provide an integrated view of VMs and associated storage resources. From the vSphere console, an administrator can identify:

- Mapping of VMs to physical and logical storage resources including LUNs and volumes.
- LUN properties (thin vs. fat, etc.)
- Device types (SSD, FC, SATA disk)
- RAID groups and RAID levels

VAAI Support

Included since vSphere 4.1 are the following storage-related APIs (VAAI), which OceanStor supports:

Full Copy - leverages existing SCSI commands available in the SCSI command set and sent from the ESXi host to the storage array to clone, snapshot, and migrate data within an array using array-based software. These functions are normally used when creating new VMs and will be particularly useful in Virtual Desktop Infrastructure (VDI) deployments. No ESXi host server cycles are consumed in the performance of these copy functions, enhancing overall VMware performance and reducing the time required to create new VMs and virtual desktops when host-based copy functions are used for these purposes. Only written data is copies to improve the efficiency.

Hardware-Assisted Locking - supports block-level granularity when protecting VMFS metadata. Previously, SCSI reservations were used at the LUN level to satisfy the same requirement.

Block Zeroing - speeds-up the standard disk initialization process in a VMware environment. This function is particularly useful in reducing the time required to create virtual disks in eager-zero thick (EZT) format.

OceanStor and Microsoft System Center

The OceanStor has a plug-in for Microsoft System Center for managing the virtual environment for the OceanStor storage resources.

Advanced Features

SmartDedupe & SmartCompression - Data Reduction

Data reduction is implemented with deduplication and compression. Data deduplication is done for file systems and thin LUNs. Because of the variable chunk size used when the thin LUNs or file systems were created, deduplication uses the granularity of the volume or file system. The chunks have a digital fingerprint calculated and the result is checked against existing fingerprints for duplicate data. The user can also enable a byte-by-byte comparison in case there is a concern about the integrity of digital fingerprints.

After deduplication, compression is performed on the chunks of data that are unique. Multiple blocks in successive logical block addresses are concatenated to improve the effectiveness of the compression algorithm if the block size is less than 32KB. Blocks greater than 32KB are compressed individually. If the pre-check on compression shows the effectiveness is below a defined threshold, data is not compressed in order to speed the retrieval.

Hardware add-in cards for the controllers are used to perform the deduplication and compression as an inline operation.

Reliability, Availability, Serviceability Features

The following summarizes the RAS capabilities:

- Redundant hot-plug cooling fans, batteries, and power supplies are standard on all models
- RAID protection of device storage
- Online serviceable device expansion modules
- Redundant, hot-pluggable controller canisters can be scaled non-disruptively
- Remote service management processor provides diagnostic reports to service center and notifies technical support through its internal online support service in the event of a failure. Information provided includes the part(s) required and system location.
- End-to-end integrity checking

Error Checking and Correction (ECC)

Data integrity is provided through end-to-end ECC calculations on each host read or write operation throughout the entire I/O path.

Data Scan

OceanStor provides continuous scan of data to repair bad sectors. There are no administrator settings with respect to bad sector repair.

Redundant Power System

The power system is based on redundant power supplies in the modular enclosures with dual line cord inputs. All power system components are replaceable non-disruptively.

Fault Isolation

Fault isolation and reliability features are included with elements in the controller and device enclosures, which all report and isolate faults.

Controller Redundancy

Controllers are configured in pairs whereby each controller node has an active partner. The two partner controllers have redundant physical connections to the subset of devices owned by the node pair, mirror their write cache to each other, and serve as the backup node for the logical disks owned by the partner node.

Upon the failure of a controller node, the node failover recovery process automatically flushes the dirty write cache to disk, transfers ownership for the volumes owned by the failed node to its partner node, and puts all logical disks owned by the remaining partner node in write-thru (non-cached) mode.

Concurrent Maintenance

The modular architecture permits modules to be taken out of service while others take over the workload. Any one portion of a redundant pair may be taken out of service and or replaced without interruption to service. That is, each controller in a controller pair may be replaced at a time, along with each drive cage in an HA configuration may also be replaced without disruption of operations. Similarly, batteries, power supplies and drives may also be replaced.

Software Updates

The embedded software also known as the Operating System may be upgraded on the controllers non-disruptively. Each controller runs its own copy of the Operating System.

Remote Support

Remote support is provided on OceanStor systems with the management unit by use of SMTP to send messages to support personnel. This provides remote monitoring of system conditions, and provides error alerts. There are no separate fees for remote support. However, users must have a valid support contract.

Service Contracts

A variety of service contracts are available, including no contract, 1 year, 2 year and 3 year contracts, with 24X7 coverage, 9X5, and response times of either 2 hours, 4 hours or next business day.

Management

OceanStor has various functions of management software that execute on a separate server or computing system (called a maintenance terminal), on an application server, and on a management module in the storage controller enclosure.

Management Module

With the Management Module in the storage controller enclosure, OceanStor provides several no-charge options for management including:

- GUI – DeviceManager executes on the Management Module and provides configuration, management, and maintenance action.
- CLI for scripted or remote administration through the management network port or the serial port on the storage system
- SNMP support for management and configuration through third party software. SNMP traps can be used for reporting of conditions.
- Syslog is a data capture and reporting function in the Management Module that sends alerts, alarms, and triggered events to user selected targets over protocols such as SMTP.

OceanStor DeviceManager

The OceanStor DeviceManager is the element manager presenting a GUI interface for managing the OceanStor system. Configuration, management and maintenance are performed through the DeviceManager. Wizards to configure the OceanStor system are invoked through the DeviceManager. Access to the DeviceManager is through a standard browser.

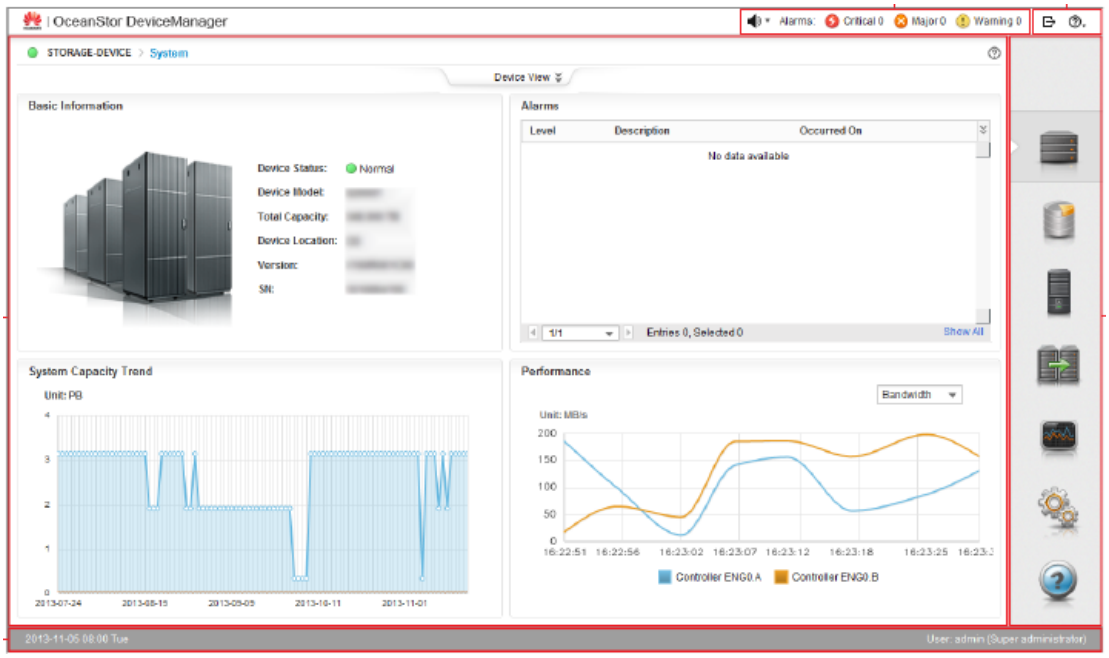

Figure 3: OceanStor Device Manager (source: Huawei)

Maintenance Terminal

The "maintenance terminal" has management software for OceanStor that includes:

- OceanStor Toolkit – for service personnel to deploy, maintain, and update systems
- OceanStor SystemReporter – performance and capacity reporting software
- OceanStor Cloud Service – remote maintenance software for monitoring, alerting, and status interrogation

Application Server

Management software with specific purpose executes on an application server:

- Replication Director – manages the remote replication, snapshot, LUN copy, and clone functions.
- UltraPath – is multipath driver software for optimizing accessing the OceanStor system.
- eSDK OceanStor – are software tools for integrating with OceanStor. Included are plug-ins for VMware vCenter and Microsoft System Center.

Performance

Huawei regularly posts performance data for storage systems. The SPC website **www.storageperformance.org** should be checked for the latest information.

The latest vendor reported numbers provided by Huawei claim a maximum of 20 million IOPs with 0.1 ms latency.

Evaluator Group Comments

Strengths:

The Huawei OceanStor Dorado V6 systems are all flash block storage systems that build on the achievements of Huawei's other OceanStor models and the previous V3 Dorado systems, but add significant technological advancements. The systems feature five different custom designed chips that increase the performance and efficiency of the storage system. In addition, all but the entry level model support end-to-end NVMe, which has become a competitive feature amongst all flash block storage systems. The scalability of systems, along with different model options gives the Dorado V6 systems flexibility to meet a wide variety of storage needs.

The Huawei Dorado V6 systems are also well equipped with a number of advanced features such as data reduction (via compression and deduplication), storage virtualization, remote replication, and stretched clusters. The Dorado V6 systems will likely meet many of the advanced feature, performance, and reliability requirements of enterprises.

Perceived Challenges:

Huawei is virtually unknown in North America, which is a very large market that also influences overseas subsidiaries. Launching into this should be expected in order to broaden the opportunities.

Another potential concern is the large amount of custom hardware included in the systems. While the chips used are designed to improve performance and efficiency, custom hardware such as this can be inflexible with regards to future changes and has the potential to hamper future innovation.

More detailed information is available at **www.evaluatorgroup.com**