Deliver Hyperconvergence

With a next-generation data platform

Enterprise-grade data platform

The Cisco HyperFlex™ HX Data Platform software revolutionizes data storage for hyperconverged infrastructure deployments and makes Cisco HyperFlex Systems ready for your enterprise applications—whether they run in virtualized environments such as Microsoft Windows 2016 Hyper-V or VMware vSphere; or containerized applications using Docker and Kubernetes. Learn about the platform's architecture and software-defined storage approach and how you can use it to eliminate the storage silos that complicate your data center.

Cisco HyperFlex™ Systems with Intel® Xeon® scalable processors

Contents

Purpose-built data platform

The Cisco HyperFlex HX Data platform is a core technology of our hyperconverged solutions. It delivers enterprise-grade storage features to make next-generation hyperconvergence ready for your enterprise applications. Our Cisco Validated Designs accelerate deployment and reduce risk for applications including:

- **Microsoft Exchange**
- **Microsoft SQL Server**
- **Oracle Database**
- **SAP HANA**
- **Splunk**
- **Virtual desktop environments including Citrix and VMware Horizon**
- **Virtual server infrastructure**

The right platform for adaptive infrastructure

Evolving application requirements have resulted in an ever-changing relationship among servers, storage systems, and network fabrics. Although virtual environments and first-generation hyperconverged systems solve some problems, they fail to match the speed of your applications and provide the enterprise-grade support for your mission-critical business applications.

That's why today's new IT operating models require new IT consumption models. Engineered on Cisco Unified Computing System™ (Cisco UCS®), Cisco HyperFlex Systems unlock the full potential of hyperconverged solutions to deliver enterprise-grade agility, scalability, security, and lifecycle management capabilities you need for operational simplicity. By deploying Cisco HyperFlex Systems, you can take advantage of the pay-as-you-grow economics of the cloud with the benefits of on-premises infrastructure.

Fast and flexible hyperconverged systems

Cisco HyperFlex Systems are designed with an end-to-end software-defined infrastructure that eliminates the compromises found in first-generation products. Cisco HyperFlex Systems combine software-defined computing in the form of Cisco UCS servers, software-defined storage with the powerful Cisco HyperFlex HX Data Platform software, and software-defined networking (SDN) with the Cisco® unified fabric that integrates smoothly with the Cisco Application Centric Infrastructure (Cisco ACI™) solution. With hybrid or all-flash storage configurations, self-encrypting drive options, and a choice of management tools, Cisco HyperFlex Systems deliver a preintegrated cluster that is up and running in an hour or less. With the capability to integrate Cisco UCS servers as compute-only nodes, you can scale computing and storage resources independently to closely match your application needs (Figure 1).

Cisco HyperFlex HX Data Platform: a new level of storage optimization

The unique data demands imposed by applications, particularly those hosted in virtual machines, have resulted in many storage silos. A foundation of Cisco HyperFlex Systems, the HX Data Platform is a purpose-built, high-performance, log-structured, scale-out file system that is designed for hyperconverged environments. The data platform's innovations redefine scale-out and distributed storage technology, going beyond the boundaries of first-generation hyperconverged infrastructure, and offer a wide range of enterprise-class data management services.

The HX Data Platform includes these features:

- **Multihypervisor support** including Microsoft Windows 2016 Server Hyper-V and VMware vSphere

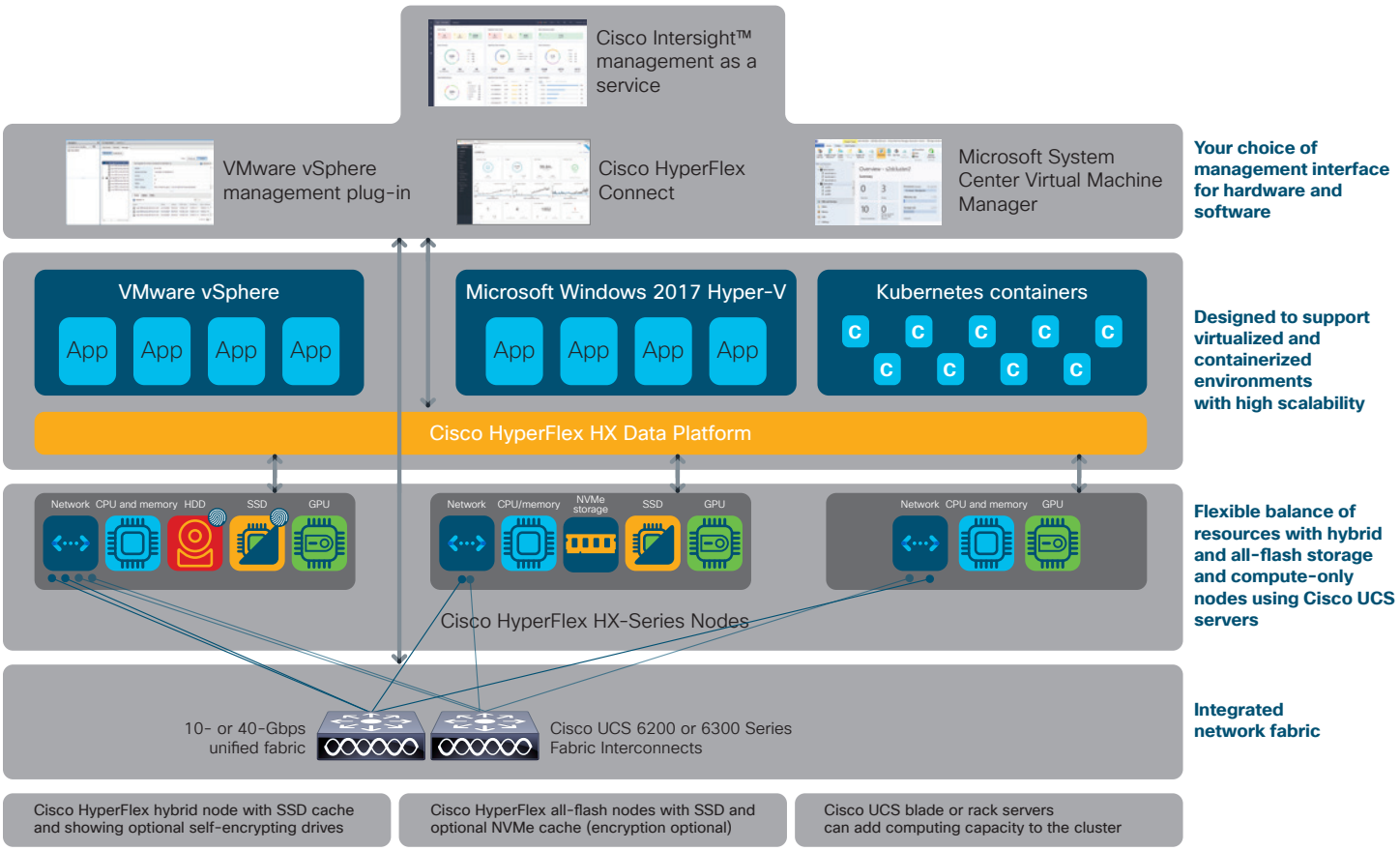


Figure 1 Cisco HyperFlex Systems offer next-generation hyperconverged solutions with a set of features only Cisco can deliver

- **Containerized application support** for persistent container data in Docker containers managed by Kubernetes.
- **Enterprise-class data management** features provide complete lifecycle management and enhanced data protection in distributed storage environments. These features include replication, always-on inline deduplication, always-on inline compression, thin provisioning, instantaneous space-efficient clones, and snapshots.
- **Support for both hybrid and all-flash models** allows you to choose the right platform configuration based on your capacity, application, performance, and budget requirements.
- **Simplified data management** integrates storage functions into existing management tools, allowing instant provisioning, cloning, and pointer-based snapshots of applications for dramatically simplified daily operations.
- **Improved control** with advanced automation and orchestration capabilities and robust reporting and analytics features delivers improved visibility and insight into IT operations.
- **Improved scalability** with logical availability zones that, when enabled, automatically partition the cluster so that it is more resilient to multiple-node failures.

What's new

Some of the most important improvements we have made in the new data platform improve scalability with built-in resiliency:

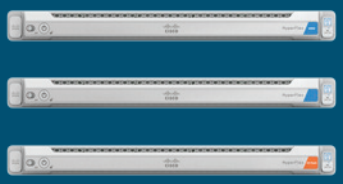
- **More nodes.** We have doubled the maximum scale to 64 nodes, with a maximum of 32 Cisco HyperFlex nodes and 32 Cisco UCS compute-only nodes.
- **More resiliency.** We have implemented logical availability zones that help you to scale wihtout compromising availability.
- **More capacity.** You can choose nodes with large-form-factor disk drives for even greater capacity. This allows your cluster to scale to higher capacity for storage-intensive applications.

- **Independent scaling** of the computing, caching, and capacity tiers gives you the flexibility to scale out the environment based on evolving business needs for predictable, pay-as-you-grow efficiency. As you add resources, data is automatically rebalanced across the cluster, without disruption, to take advantage of the new resources.
- **Continuous data optimization** with inline data deduplication and compression increases resource utilization and offers more headroom for data scaling.
- **Dynamic data placement** optimizes performance and resilience by enabling all cluster resources to participate in I/O responsiveness. Hybrid nodes use a combination of solid-state disks (SSDs) for caching and hard-disk drives (HDDs) for the capacity layer. All-flash nodes use SSD or Non-Volatile Memory Express (NVMe) storage for the caching layer and SSDs for the capacity layer. This approach helps eliminate storage hotspots and makes the performance capabilities of the cluster available to every virtual machine. If a drive fails, restoration can proceed quickly because the aggregate bandwidth of the remaining components in the cluster can be used to access data.
- **Stretch clusters** synchronously replicate data between two identical clusters to provide continuous operation even if an entire location becomes unavailable.
- **Enterprise data protection** with a highly-available, self-healing architecture supports nondisruptive, rolling upgrades and offers options for call-home and onsite support 24 hours a day, every day.
- **API-based data platform architecture** provides data virtualization flexibility to support existing and new cloud-native data types.
- Tools that simplify deployment and operations. The Cisco HyperFlex Sizer helps you to profile existing environments and estimate sizing for new deployments. Cloud-based installation through Cisco Intersight™ management as a service allows you to ship nodes to any location and install a cluster from anywhere you have a web browser.

Architecture

In Cisco HyperFlex Systems, the data platform spans three or more Cisco HyperFlex HX-Series nodes to create a highly available cluster. Each node includes an HX Data Platform controller that implements the scale-out and distributed file system using internal flash-based SSDs or a combination of flash-based SSDs and high-capacity HDDs to store data. The controllers communicate with each other over 10 or 40 Gigabit Ethernet to present a single pool of storage that spans the nodes in the cluster (Figure 2). In a Cisco HyperFlex Edge cluster, the controllers communicate with each other through an existing Gigabit Ethernet network. Nodes access data through a data layer using file, block, object, and API plug-ins. As nodes are added, the cluster scales linearly to deliver computing, storage capacity, and I/O performance.

Cisco HyperFlex nodes



Cisco HyperFlex HX220c M5 and All Flash Nodes

Cisco HyperFlex HX240c M5 and All Flash Nodes

Cisco HyperFlex Edge configuration

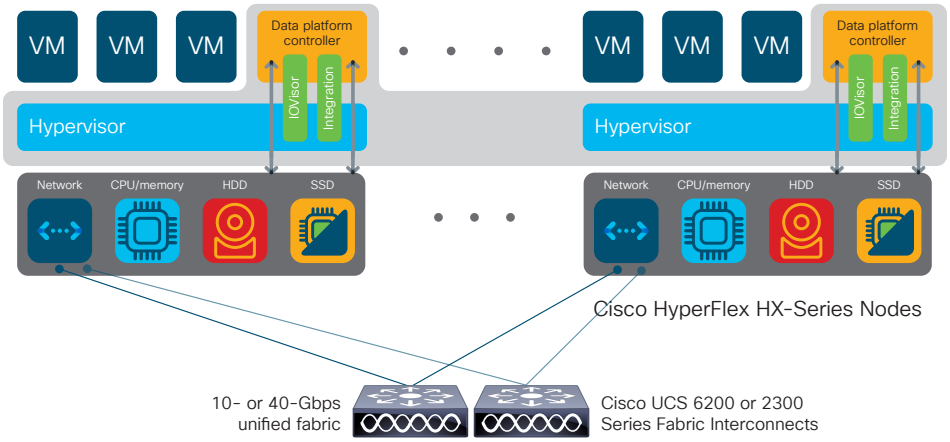


Figure 2 Distributed Cisco HyperFlex System

Modular data platform

The data platform is designed with a modular architecture so that it can easily adapt to support a broadening range of application environments and hardware platforms (Figure 3):

The core file system, cluster service, data service, and system service are designed to adapt to a rapidly evolving hardware ecosystem. This enables us to be among the first to bring new server, storage, and networking capabilities to Cisco HyperFlex Systems as they are developed.

Each hypervisor or container environment is supported by a gateway and manager module that supports the higher layers of software with storage access suited to its needs.

The data platform provides a REST API so that a wide range of management tools can interface with the data platform.

Resident on each node

The data platform controller resides in a separate virtual machine in each node, whether it is a hybrid or all-flash node or a compute-only node. The virtual machine uses dedicated CPU cores and memory so that its workload fluctuations have no impact on the applications running on the node.

The controller accesses all of the node's disk storage through hypervisor bypass mechanisms for higher performance. It uses the node's memory and SSD drives or NVMe storage as part of a distributed caching layer, and it uses the node's HDDs for distributed storage. The data platform controller interfaces with the hypervisor in two ways:

- **IOVisor:** The data platform controller intercepts all I/O requests and routes requests to the nodes responsible for storing or retrieving the blocks. The IOVisor makes the existence of the

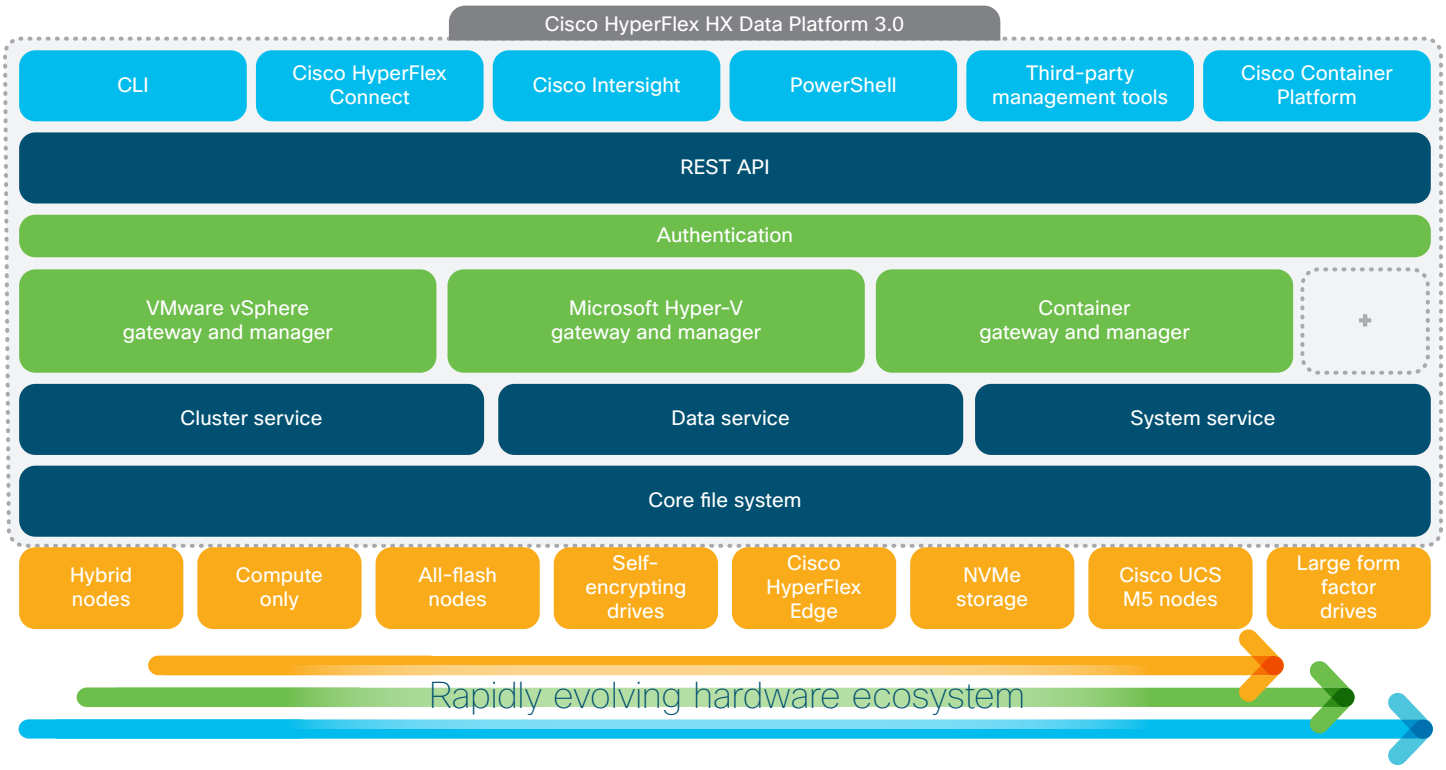


Figure 3 Cisco HyperFlex data platform modular architecture

hyperconvergence layer transparent to the hypervisor and it presents the file system or device interface to the software layer above.

- **Advanced feature integration.** A module uses the hypervisor APIs to support advanced storage system operations such as snapshots and cloning. These are accessed through the hypervisor so that the hyperconvergence layer appears just as enterprise shared storage does. The controller accelerates operations through manipulation of metadata rather than actual data copying, providing rapid response, and thus rapid deployment of new application environments.

How it works

The HX Data Platform controller handles all read and write requests for volumes that the hypervisor accesses and thus mediates all I/O from the virtual machines. (The hypervisor has a dedicated boot disk independent from the data platform.) The data platform implements a distributed, log-structured file system that always uses a caching layer in SSDs to accelerate write responses; a file system caching layer in SSDs to accelerate read requests in hybrid configurations; and a persistence layer implemented with SSDs or HDDs.

Data distribution

Incoming data is distributed across all nodes in the cluster to optimize performance using the caching layer (Figure 4). Effective data distribution is achieved by mapping incoming data to stripe units that are stored evenly across all nodes, with the number of data replicas determined by the policies you set. When an application writes data, the data is sent to the appropriate node based on the stripe unit, which includes the relevant block of information. This data distribution approach in combination with the capability to have multiple streams writing at the same time prevents both network and storage hotspots, delivers the same I/O performance regardless of virtual machine location, and gives you more flexibility in workload placement. Other architectures use a locality approach that does not make full use of available networking and I/O resources.

When you migrate a virtual machine to a new location, the HX Data Platform does not require data to be moved. This approach significantly reduces the impact and cost of moving virtual machines among systems.

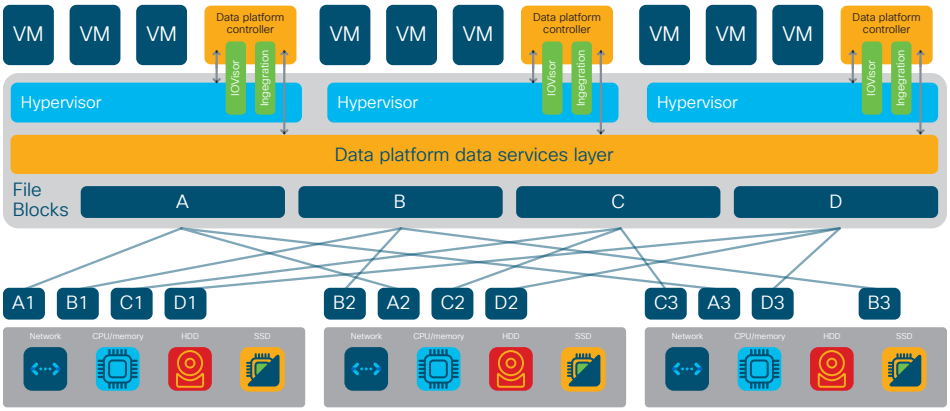


Figure 4 Data blocks are striped across nodes in the cluster

Logical availability zones

The striping of data across nodes is modified if logical availability zones are enabled. This feature automatically partitions the cluster into a set of availability zones based on the number of nodes in the cluster and the replication factor for the data.

Each availability zone has at most one copy of each block. Multiple component or node failures in a single zone can occur and make the single zone unavailable. The cluster can continue to operate as long as a single group has a copy of the data.

Without logical availability zones, a cluster with 20 nodes and a replication factor of 3 can have no more than two nodes fail without the cluster having to shut down. With logical availability zones enabled, all of the nodes in up to 2 availability zones can fail As Figure 5 illustrates, the 20-node cluster example

can have up to 8 components or nodes fail but continue to provide data availability.

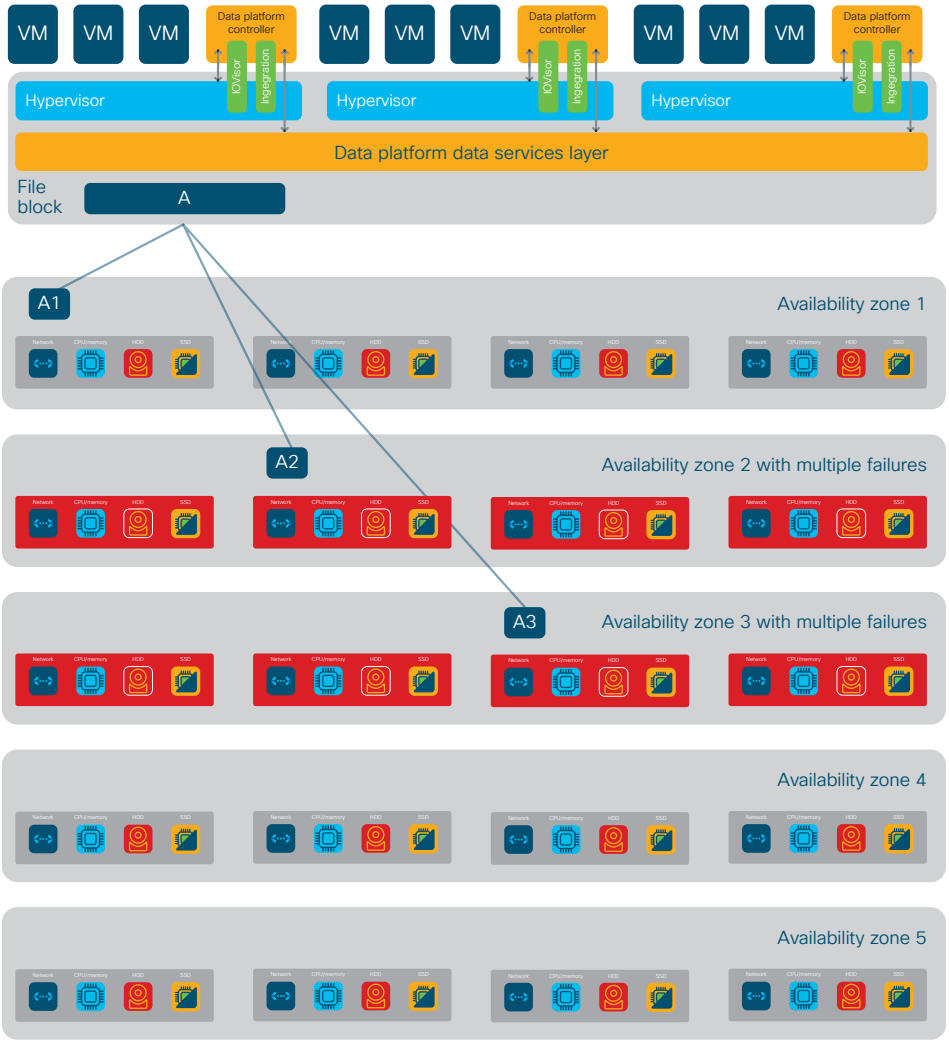


Figure 5 Logical availability zones allow the cluster to continue to provide data availability despite a larger number of node failures; in this example 8 out of 20 nodes can fail and the cluster can continue to operate

Data read and write operations

The data platform implements a distributed, log-structured file system that changes the way that it handles caching and storage capacity depending on the node configuration.

- **In a hybrid configuration,** the data platform uses a caching layer in SSDs to accelerate read requests and write responses, and it implements a capacity layer in HDDs.

- **In an all-flash-memory configuration,** the data platform uses a caching layer in SSDs to accelerate write responses, and it implements a capacity layer in SSDs. Read requests are fulfilled directly from data obtained from the SSDs in the capacity layer. A dedicated read cache is not required to accelerate read operations.

In both types of configurations, incoming data is striped across the number of nodes required to meet availability requirements: usually two or three nodes. Based on policies you set, incoming write operations are acknowledged as persistent after they are replicated to the SSD in other nodes in the cluster. This approach reduces the likelihood of data loss due to SSD or node failures. It also implements the synchronous replication that supports stretch clusters. The write operations are then destaged to SSDs in the capacity layer in all-flash configurations or to inexpensive, high-density HDDs in hybrid configurations for long-term storage. You can choose to use only SSDs to improve performance, increase density, and reduce latency or use high-performance SSDs with low-cost, high-capacity HDDs to optimize the cost of storing data.

The controller assembles blocks to be written to the cache until a configurable-sized write log is full or until workload conditions dictate that it be destaged to an SSD or a spinning disk. When existing data is (logically) overwritten, the log-structured approach simply appends a new block and updates the metadata. When the data is destaged to a HDD, the write operation consists of a single seek operation with a large amount of sequential data written. This approach improves performance significantly compared to the traditional read-modify-write model, which is characterized by numerous seek operations on HDDs, with small amounts of data written at a time. This layout also benefits SSD configurations in which seek operations are not time consuming. It reduces the write amplification levels of SSDs and the total number of writes the flash media experiences due to incoming write operations and random overwrite operations of the data.

When data is destaged to a disk in each node, the data is deduplicated and compressed. This process occurs after the write operation is acknowledged, so no performance penalty is incurred for these operations. A small deduplication block size helps increase the deduplication rate. Compression further reduces the data footprint. Data is then moved to SSD or HDD storage as write cache segments are released for reuse (Figure 6).

Hot data sets—data that is frequently or recently read from the persistence tier—are cached in memory. In hybrid configurations, hot data sets are also cached in SSDs (Figure 7). In configurations that use HDDs for persistent storage, having the most frequently used data in the caching layer helps accelerate the performance of workloads. For example, when applications and virtual machines modify data, the data is likely read from the cache, so data on the spinning disk often does not need to be read and then expanded. Because the HX Data Platform decouples the caching tier from the persistence tier, you can independently scale I/O performance and storage

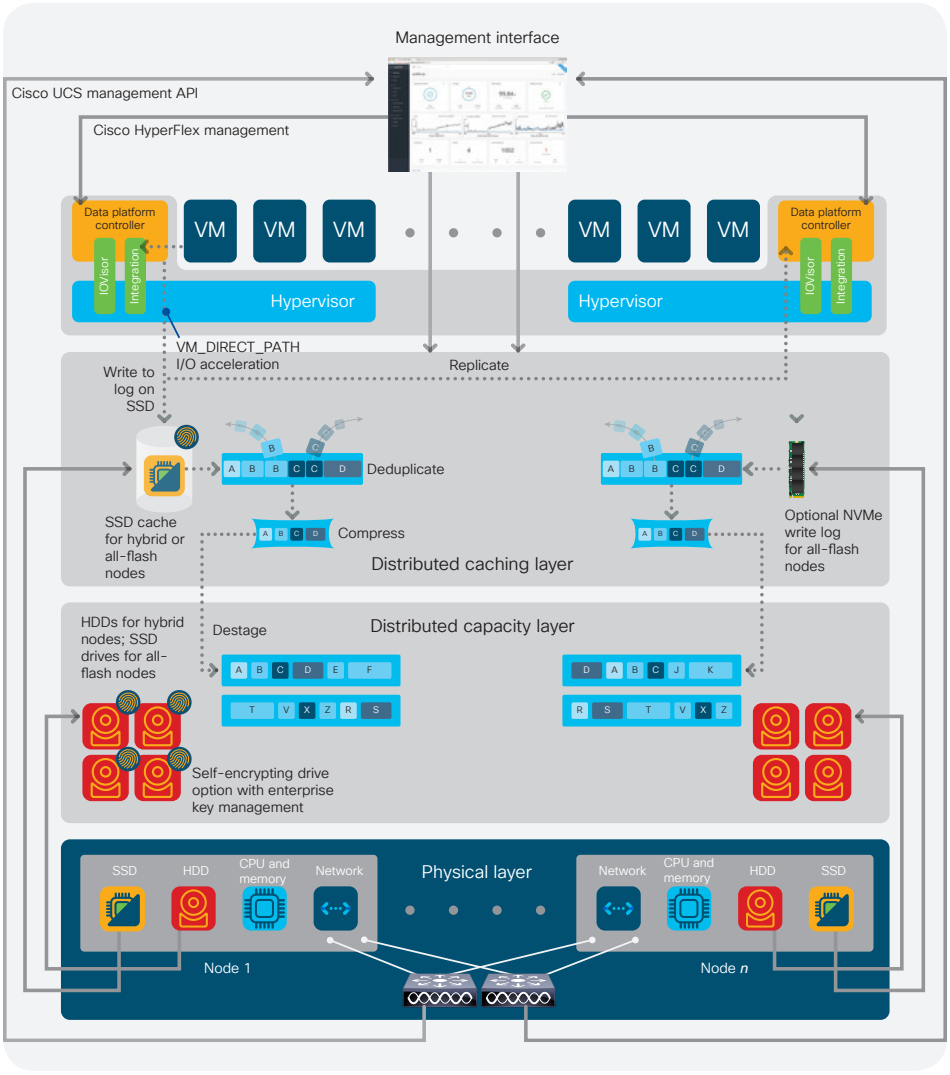


Figure 6 Data write flow through the Cisco HyperFlex HX Data Platform

capacity. All-flash configurations, however, do not use a read cache because data caching does not provide any performance benefit; the persistent data copy already resides on high-performance SSDs. In these configurations, a read cache implemented with SSDs could become a bottleneck and prevent the system from using the aggregate bandwidth of the entire set of SSDs.

Data optimization

The HX Data Platform provides finely detailed inline deduplication and variable block inline compression that is always on for objects in the cache (SSD and memory) and capacity (SSD or HDD) layers. Unlike other solutions, which require you to turn off these features to maintain performance, the deduplication and compression capabilities in the Cisco data platform are designed to sustain and enhance performance and significantly reduce physical storage capacity requirements.

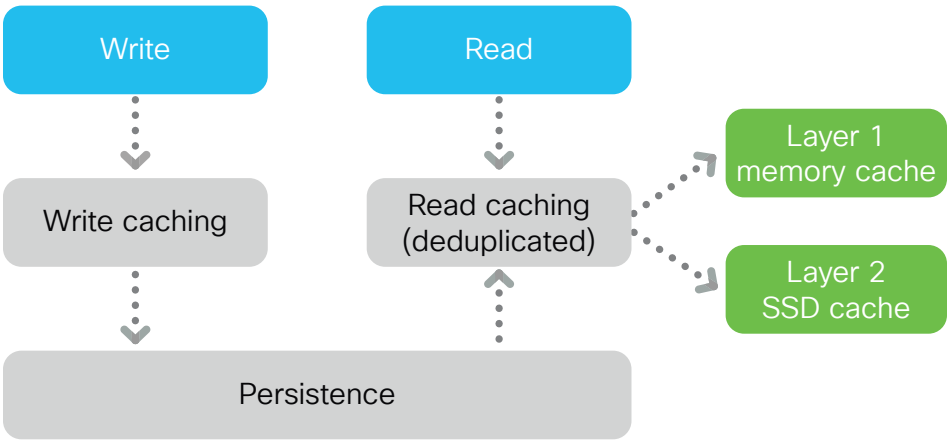


Figure 7 Decoupled data caching and data persistence

Data deduplication

Data deduplication is used on all storage in the cluster, including memory, SSDs, and HDDs (Figure 8). Data is deduplicated in the persistence tier to save space, and it remains deduplicated when it is read into the caching tier in hybrid configurations. This approach allows a larger working set to be stored in the caching tier, accelerating read performance for configurations that use slower HDDs.

Inline compression

The HX Data Platform uses high-performance inline compression on data sets to save storage capacity. Although other products offer compression capabilities, many negatively affect performance. In contrast, the Cisco data platform uses CPU-offload instructions to reduce the performance impact of compression operations. In addition, the log-structured distributed-objects layer has no effect on modifications (write operations) to previously compressed data. Instead, incoming modifications are compressed and written to a new location, and the existing (old) data is marked for deletion, unless the data needs to be retained in a snapshot.

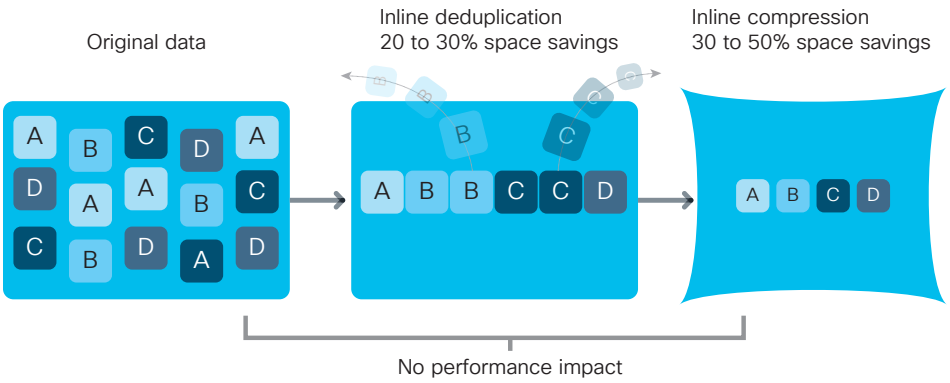


Figure 8 The data platform optimizes data storage with no performance impact

Note that data that is being modified does not need to be read prior to the write operation. This feature avoids typical read-modify-write penalties and significantly improves write performance.

Log-structured distributed objects

In the HX Data Platform, the log-structured distributed-object store layer groups and compresses data that filters through the deduplication engine into self-addressable objects. These objects are written to disk in a log-structured, sequential manner. All incoming I/O—including random I/O—is written sequentially to both the caching (SSD and memory) and persistence (SSD or HDD) tiers. The objects are distributed across all nodes in the cluster to make uniform use of storage capacity.

By using a sequential layout, the platform helps increase flash-memory endurance and makes the best use of the read and write performance characteristics of HDDs, which are well suited for sequential I/O operations. Because read-modify-write operations are not used, compression, snapshot, and cloning operations have little or no impact on overall performance.

Data blocks are compressed into objects and sequentially laid out in fixed-size segments, which in turn are sequentially laid out in a log-structured manner (Figure 9). Each compressed object in the log-structured segment is uniquely addressable using a key, with each key fingerprinted and stored with a checksum to provide high levels of data integrity. In addition, the chronological writing of objects helps the platform quickly recover from media or node failures by rewriting only the data that came into the system after it was truncated due to a failure.

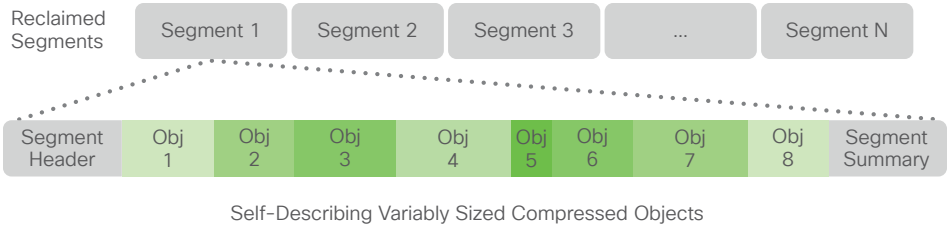


Figure 9 Log-structured file system data layout

Encryption

Using optional self-encrypting drives (SEDs), the HX Data Platform encrypts both the caching and persistence layers of the data platform. Integrated with enterprise key management software or with passphrase-protected keys, encryption of data at rest helps you comply with Health Insurance Portability and Accountability Act (HIPAA), Payment Card Industry Data Security Standard (PCI-DSS), Federal Information Security Management Act FISMA, and Sarbanes-Oxley regulations. The platform itself is hardened to Federal Information Processing Standard (FIPS) 140-1, and the encrypted drives with key management comply with the FIPS 140-2 standard.

Data services

The HX Data Platform provides a scalable implementation of space-efficient data services, including thin provisioning, pointer-based snapshots, native replication, and clones—without affecting performance.

Thin provisioning

The platform makes efficient use of storage by eliminating the need to forecast, purchase, and install disk capacity that may remain unused for a long time. Virtual data containers can present any amount of logical space to applications, whereas the amount of physical storage space that is needed is determined by the data that is written. As a result, you can expand storage on existing nodes and expand your cluster by adding more storage-intensive nodes as your business requirements dictate, eliminating the need to purchase large amounts of storage before you need it.

Snapshots

The HX Data Platform uses metadata-based, zero-copy snapshots to facilitate backup operations and remote replication: critical capabilities in enterprises that require always-on data availability. Space-efficient snapshots allow you to perform frequent online backups of data without needing to worry about the consumption of physical storage capacity. Data can be moved offline or restored from these snapshots instantaneously.

- **Fast snapshot updates:** When modified data is contained in a snapshot, it is written to a new location, and the metadata is updated, without the need for read-modify-write operations.
- **Rapid snapshot deletions:** You can quickly delete snapshots. The platform simply deletes a small amount of metadata that is located on an SSD, rather than performing a long consolidation process as needed by solutions that use a delta-disk technique.
- **Highly specific snapshots:** With the HX Data Platform, you can take snapshots on an individual file basis. In virtual environments, these files map to drives in a virtual machine. This flexible specificity allows you to apply different snapshot policies on different virtual machines.

Native replication

This feature is designed to provide policy-based remote replication for disaster recovery and virtual machine migration purposes. Through the HX Connect interface, you create replication policies that specify the repair point objective (RPO). You add virtual machines to protection groups that inherit the policies you define. Native replication can be used for planned data movement (for example migrating applications between locations) or unplanned events such as data center failures. Our native replication is designed to work with VM20/20 EZDR disaster recovery software that you can use to coordinate replication, failover, and failback activities.

Unlike enterprise shared storage systems, which replicate entire volumes, we replicate data on a per-virtual-machine basis. This way you can configure replication on a fine-grained basis so that you have remote copies of the data you care about.

The data platform coordinates the movement of data with the remote data platform, and all nodes participate in the data movement using a many-to-many connectivity model (Figure 10). This model distributes the workload across all nodes, avoiding hot spots and minimizing performance impacts. Once the first data is replicated, subsequent replication is based on data blocks changed since the last transfer. Recovery point objectives (RPOs) can be set in a range from 15 minutes to 25 hours. Configuration settings allow you to constrain bandwidth so that the remote replication does not overwhelm your wide-area network (WAN) connection.

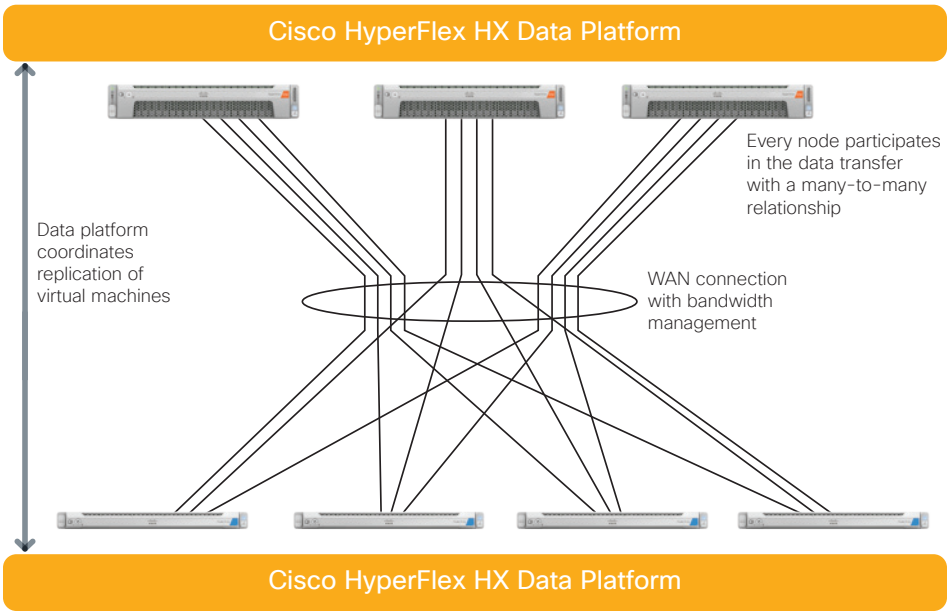


Figure 10 Native replication uses a many-to-many relationship between clusters to eliminate hot spots and spread the workload across nodes

Stretch clusters

With stretch clusters you can have two identical configurations in two locations acting as a single cluster. With synchronous replication between sites, a complete data center failure can occur and your applications can still be available with zero data loss. In other words, applications can continue running with no loss of data. The recovery time objective is only the time that it takes to recognize the failure and put a failover into effect.

Fast, space-efficient clones

In the HX Data Platform, clones are writable snapshots that can be used to rapidly provision items such as virtual desktops and applications for test and development environments. These fast, space-efficient clones rapidly replicate storage volumes so that virtual machines can be replicated through just metadata operations, with actual data copying performed only for write operations. With this approach, hundreds of clones can be created and deleted in minutes. Compared to full-copy methods, this approach can save a significant amount of time, increase IT agility, and improve IT productivity.

Clones are deduplicated when they are created. When clones start diverging from one another, data that is common between them is shared, with only unique data occupying new storage space. The deduplication engine eliminates data duplicates in the diverged clones to further reduce the clone's storage footprint. As a result, you can deploy a large number of application environments without needing to worry about storage capacity use.

Enterprise-class availability

In the HX Data Platform, the log-structured distributed-object layer replicates incoming data, improving data availability. Based on policies that you set, data that is written to the write cache is synchronously replicated to one or more SSDs located in different nodes before the write operation is acknowledged to the application. This approach allows incoming write operations to be acknowledged quickly while protecting data from SSD or node failures. If an SSD or node fails, the replica is quickly re-created on other SSDs or nodes using the available copies of the data.

The log-structured distributed-object layer also replicates data that is moved from the write cache to the capacity layer. This replicated data is likewise protected from SSD, HDD, or node failures. With two replicas, or a total of three data copies, the cluster can survive uncorrelated failures of two SSDs, two HDDs, or two nodes without the risk of data loss. Uncorrelated failures are failures that occur on different physical nodes. Failures that occur on the same node affect the same copy of data and are treated as a single failure. For example, if one disk in a node fails and subsequently another disk on the same node fails, these correlated failures count as one failure in the system. In this case, the cluster could withstand another uncorrelated failure on a different node. See the Cisco HyperFlex HX Data Platform system administrator's guide for a complete list of fault-tolerant configurations and settings.

If a problem occurs in the Cisco HyperFlex controller software, data requests from the applications residing in that node are automatically routed to other controllers in the cluster. This same capability can be used to upgrade or perform maintenance on the controller software on a rolling basis without affecting the availability of the cluster or data. This self-healing capability is

For More Information

- Cisco HyperFlex Systems
- Cisco HyperFlex Services

one of the reasons that the HX Data Platform is well suited for production applications.

Data rebalancing

A distributed file system requires a robust data-rebalancing capability. In the HX Data Platform, no overhead is associated with metadata access, and rebalancing is extremely efficient. Rebalancing is a nondisruptive online process that occurs in both the caching and persistence layers, and data is moved at a fine level of specificity to improve the use of storage capacity. The platform automatically rebalances existing data when nodes and drives are added or removed or when they fail. When a new node is added to the cluster, its capacity and performance is made available to new and existing data. The rebalancing engine distributes existing data to the new node and helps ensure that all nodes in the cluster are used uniformly from both capacity and performance perspectives. If a node fails or is removed from the cluster, the rebalancing engine rebuilds and distributes copies of the data from the failed or removed node to available nodes in the clusters.

Online upgrades

Cisco HyperFlex HX-Series nodes and the HX Data Platform support online upgrades so that you can expand and update your environment without business disruption. You can easily expand your physical resources; add processing capacity; and download and install BIOS, driver, hypervisor, firmware, and Cisco UCS Manager updates, enhancements, and bug fixes.

Conclusion

The Cisco HyperFlex HX Data Platform revolutionizes data storage for hyperconverged infrastructure deployments that support new IT consumption models. The platform's architecture and software-defined storage approach gives you a purpose-built, high-performance distributed file system with a wide array of enterprise-class data management services. With innovations that redefine distributed storage technology, the data platform gives you the hyperconverged infrastructure you need to deliver adaptive IT infrastructure.

LE-55406-03 01/18