White Paper
March 2016

Next-Generation Data Platform for Hyperconvergence

Cisco HyperFlex™ Systems
with Intel® Xeon® Processors

Contents

Next-Generation Data Platform for Hyperconvergence

White Paper
March 2016

What You Will Learn

This document describes Cisco HyperFlex™ HX Data Platform Software, which revolutionizes data storage for hyperconverged infrastructure deployments. You'll 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 IT infrastructure.

A Platform for a New Generation of Applications and Data

Applications dictate IT architecture, and evolving requirements have resulted in an ever-changing relationship among servers, storage systems, and network fabrics. Although virtualized environments and first-generation hyperconverged systems solve some problems, they also create new infrastructure silos, fail to deliver massive scalability, and lack lifecycle management features and strong data security. Cisco HyperFlex™ Systems deliver a new generation of flexible, scalable solutions that unlock the full potential of hyperconverged solutions for a wide range of applications, workloads, and use cases.

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 Unified Computing System™ (Cisco UCS®) servers, software-defined storage with powerful new Cisco HyperFlex HX Data Platform Software, and software-defined networking (SDN) with the Cisco® unified fabric that integrates smoothly with Cisco Application Centric Infrastructure (Cisco ACI™). The result is a preintegrated cluster that is up and running in an hour or less and that scales resources independently to closely match your application resource needs (Figure 1).

Cisco HyperFlex HX Data Platform: Eliminating Storage Silos

If your IT organization uses server virtualization to consolidate physical servers, the unique data demands imposed by applications have resulted in many storage silos. A foundation of Cisco HyperFlex Systems, the Cisco HyperFlex HX Data Platform is a purpose-built, high-performance distributed file system with a wide array of enterprise-class data management services. The data platform's innovations redefine distributed storage technology, going beyond the boundaries of first-generation hyperconverged infrastructure.

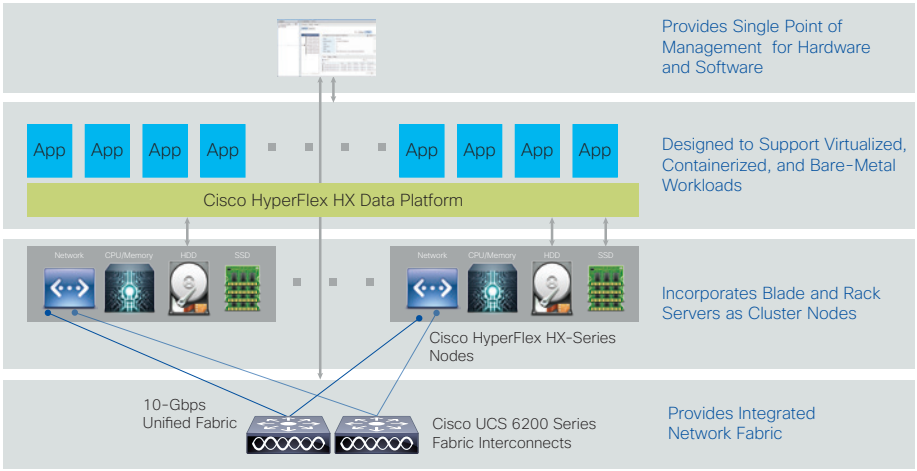

Figure 1. Cisco HyperFlex Systems Offer a New Generation of Hyperconverged Solutions with A Set of Features Only Cisco Can Deliver

The Cisco HyperFlex HX Data Platform includes:

- **Enterprise-class data management** features that are required for complete lifecycle management and enhanced data protection in distributed storage environments—including replication, deduplication, compression, thin provisioning, rapid, space-efficient clones, and snapshots
- **Simplified data management** that integrates storage functions into existing management tools, allowing instant provisioning, cloning, and snapshots of applications for dramatically simplified daily operations
- **Independent scaling** of the computing, caching, and capacity tiers, giving you the flexibility to scale the environment based on evolving business needs
- **Continuous data optimization** with inline data deduplication and compression that increases resource utilization with more headroom for data scaling
- **Dynamic data placement** in node memory, enterprise-class flash memory (on solid-state disk [SSD] drives), and persistent storage tiers (on hard-disk drives [HDDs]) to optimize performance and resiliency—and to readjust data placement as you scale your cluster
- **API-based data platform architecture** that provides data virtualization flexibility to support existing and new cloud-native data types

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 a Cisco HyperFlex HX Data Platform controller that implements the distributed file system using internal flash-based SSD drives and high-capacity HDDs to store data. The controllers communicate with each other over 10 Gigabit Ethernet to present a single pool of storage that spans the nodes in the cluster (Figure 2). 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.

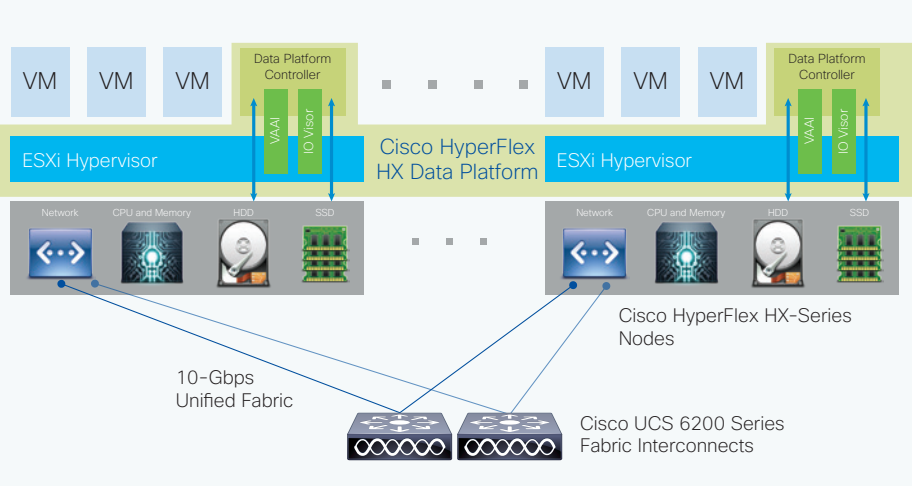

Figure 2. Distributed Cisco HyperFlex System

In the VMware vSphere environment, the controller occupies a virtual machine with a dedicated number of processor cores and amount of memory, allowing it to deliver consistent performance and not affect the performance of the other virtual machines on the cluster. The controller can access all storage without hypervisor intervention through the VMware VM_DIRECT_PATH feature. It uses the node's memory and SSD drives as part of a distributed caching layer, and it uses the node's HDDs for distributed capacity storage. The controller integrates the data platform into VMware software through the use of two preinstalled VMware ESXi vSphere Installation Bundles (VIBs):

- **IO Visor:** This VIB provides a network file system (NFS) mount point so that the ESXi hypervisor can access the virtual disk drives that are attached to individual virtual machines. From the hypervisor's perspective, it is simply attached to a network file system.
- **VMware vStorage API for Array Integration (VAAI):** This storage offload API allows vSphere to request advanced file system operations such as snapshots and cloning. The controller causes these operations to occur through manipulation of metadata rather than actual data copying, providing rapid response, and thus rapid deployment of new application environments.

How It Works

The Cisco HyperFlex 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 log-structured file system that uses a caching layer in SSD drives to accelerate read requests and write responses, and a persistence layer implemented with HDDs.

Data Distribution

Incoming data is distributed across all nodes in the cluster to optimize performance using the caching tier (Figure 3). 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

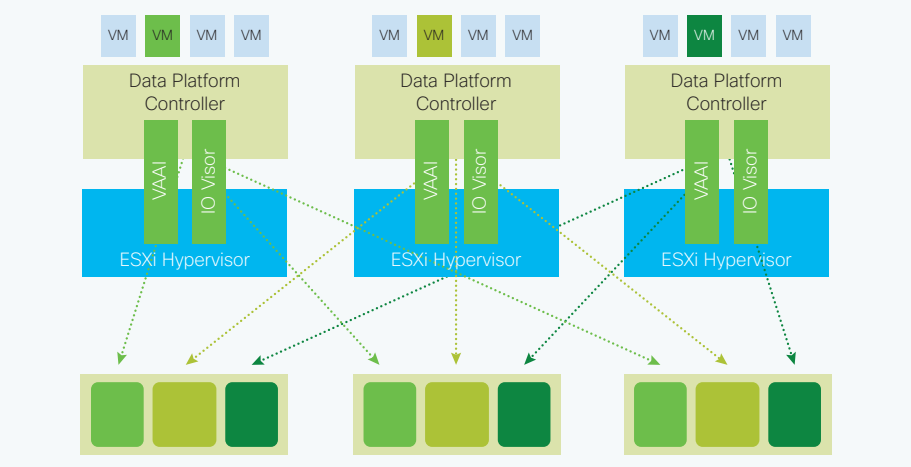

Figure 3. Data Is Striped Across Nodes in the Cluster

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 avoids both network and storage hot spots, delivers the same I/O performance regardless of virtual machine location, and gives you more flexibility in workload placement. This is in contrast to other architectures that use a locality approach that does not fully use available networking and I/O resources.

- **Data write operations:** For write operations, data is written to the local SSD cache and the replicas are written in parallel to remote SSDs before the write operation is acknowledged.
- **Data read operations:** For read operations, data that is local will usually be read directly from the local SSD. If the data is not local, the data is retrieved from a SSD on the remote node. This allows the platform to use all SSDs for read operations, reducing bottlenecks and delivering excellent performance.

When moving a virtual machine to a new location using tools such as VMware Dynamic Resource Scheduling (DRS), the Cisco HyperFlex HX Data Platform does not require data to be moved. This approach significantly reduces the impact and cost of moving virtual machines among systems.

Data Operations

The data platform implements a log-structured file system that uses a caching layer in SSD drives to accelerate read requests and write responses, and it implements a capacity layer in HDDs. Incoming data is striped across the number of nodes required to satisfy 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 drives in other nodes in the cluster. This approach reduces the likelihood of data loss due to SSD or node failures. The write operations are then destaged to inexpensive, high-density HDDs for long-term storage. By using high-performance SSD drives with low-cost, high-capacity HDDs, you can optimize the cost of storing and retrieving application data at full speed.

The log-structured file system 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 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, 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, with small amounts of data written at a time.

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 HDD storage as write cache segments are released for reuse (Figure 4).

Hot data sets—data that is frequently or recently read from the persistent tier—are cached both in SSD drives and in memory (Figure 5). Having the most frequently used data in the caching layer helps make Cisco HyperFlex Systems perform well for virtualized applications. When virtual machines modify data, the data is likely

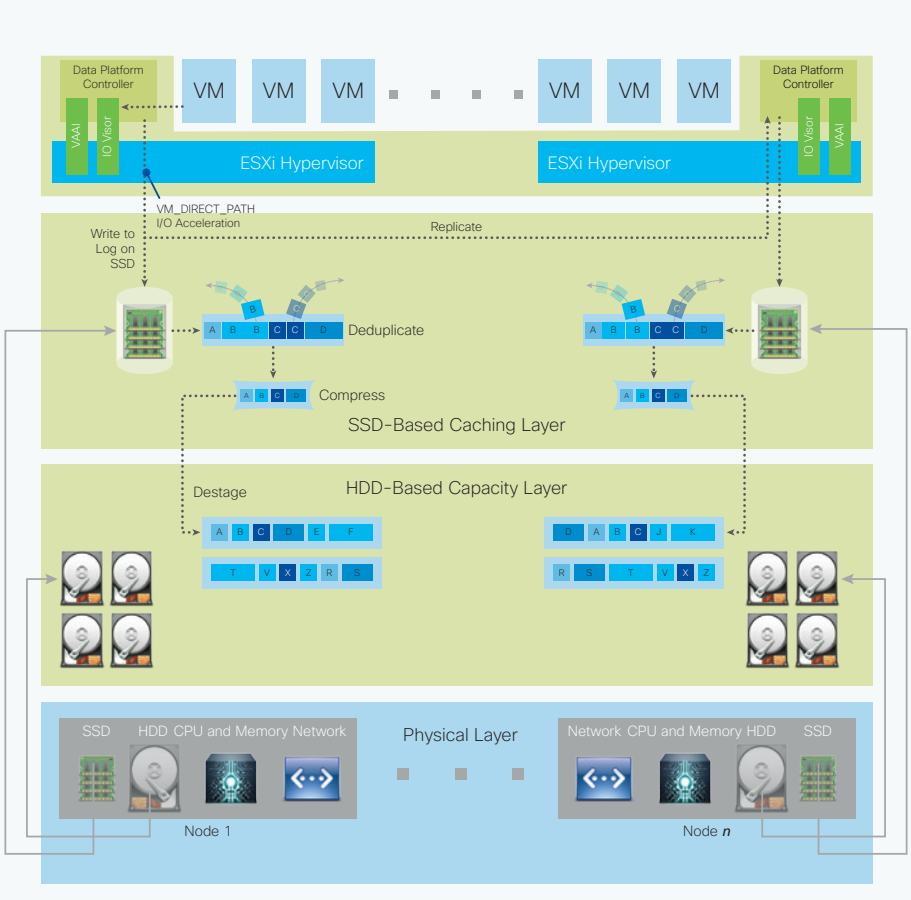

Figure 4. Data Write Operation Flow Through the Cisco HyperFlex HX Data Platform

read from the cache, so data on the spinning disk often does not need to be read and then expanded. Because the Cisco HyperFlex HX Data Platform decouples the caching tier from the persistent tier, you can independently scale I/O performance and storage capacity.

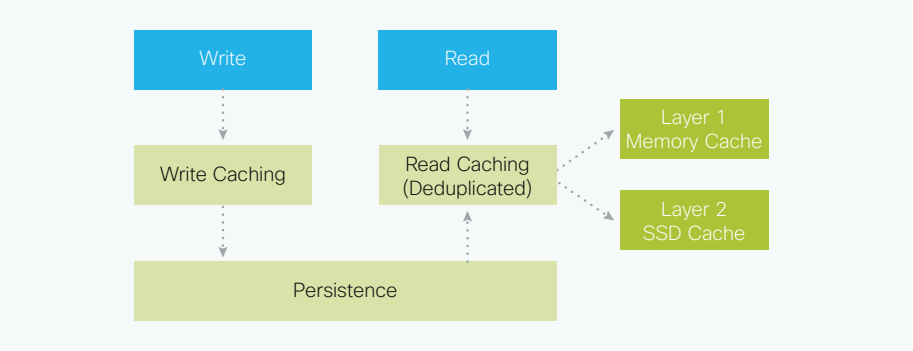

Figure 5. Decoupled Data Caching and Data Persistence

Data Optimization

The Cisco HyperFlex 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 (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.

Data Deduplication

Data deduplication is used on all storage in the cluster, including memory, SSD drives, and HDDs. Based on a patent-pending Top-K Majority algorithm, the platform uses conclusions from empirical research that show that the majority of data, when sliced into small data blocks, has significant deduplication potential based on a minority of the data blocks. By fingerprinting and indexing just these frequently used blocks, high rates of deduplication can be achieved with only a small amount of memory, which is a high-value resource in cluster nodes (Figure 6). Data is not only deduplicated in the persistence tier to save space; it remains deduplicated when it is read into the caching tier. This approach allows a larger working set to be stored in the caching tier, accelerating read performance.

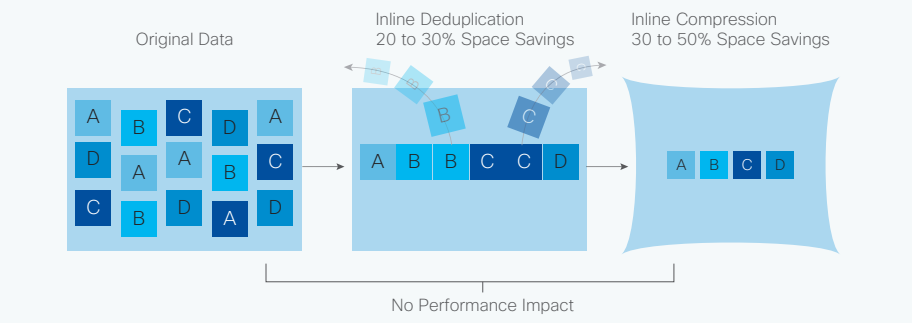

Figure 6. Cisco HyperFlex HX Data Platform Optimizes Data Storage with No Performance Impact

Inline Compression

The Cisco HyperFlex HX Data Platform uses high-performance inline compression on data sets to save disk space. 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. 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 Cisco HyperFlex 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 persistent (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, there is little or no performance impact on compression and snapshot operations or 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 7). 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.

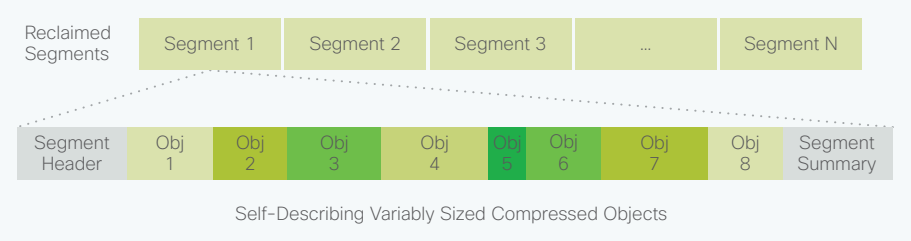

Figure 7. Cisco HyperFlex HX Data Platform's Log-Structured File System Data Layout

Data Services

The Cisco HyperFlex HX Data Platform provides a scalable implementation of space-efficient data services, including thin provisioning, space reclamation, pointer-based snapshots, 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 Cisco HyperFlex 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 Cisco HyperFlex 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.

Fast, Space-Efficient Clones

In the Cisco HyperFlex 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.

Data Availability

In the Cisco HyperFlex 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 SSD drives located in different nodes before the write operation is acknowledged to the application. This approach allows incoming writes 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 SSD drives 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 HDD or node failures. With two replicas, or a total of three data copies, the cluster can survive failure of two SSD drives, two HDDs, or two nodes without the risk of data loss. 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 HX 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 one of the reasons that the Cisco HyperFlex HX Data Platform is well suited for production applications.

Data Rebalancing

A distributed file system requires a robust data rebalancing capability. In the Cisco HyperFlex 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 persistent 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 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.

Conclusion

The Cisco HyperFlex HX Data Platform revolutionizes data storage for hyperconverged infrastructure deployments. 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. The data platform's innovations redefine distributed storage technology, providing you with the next generation of hyperconverged infrastructure.

For More Information

For more information about Cisco HyperFlex Systems, visit http://www.cisco.com/go/hyperflex.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices**.

LE-55406-00 03/16