Experts in Hybrid Cloud
experis TECHNOLOGY GROUP

Hewlett Packard Enterprise

HPE PRIMERA ARCHITECTURE

CONTENTS

MISSION CRITICAL REFINED FOR THE INTELLIGENCE ERA

HPE Primera is AI-driven storage for proven tier-0 performance and resiliency

Powered by AI, HPE Primera storage redefines mission-critical storage for tier-0 applications. Designed for NVMe and Storage Class Memory, HPE Primera delivers remarkable simplicity, app-aware resiliency for mission-critical workloads, and intelligent storage that anticipates and prevents issues across the infrastructure stack.

HPE Primera delivers on the promise of intelligent storage advanced data services and simplicity for your mission-critical applications with a services-centric OS that sets up in minutes, and upgrades seamlessly to minimize risk and be transparent to applications. All of these capabilities add up to enable HPE Primera to provide 100% availability guaranteed.[1]

This white paper describes the architectural elements of the HPE Primera 600 storage family.

HPE PRIMERA HARDWARE ARCHITECTURE

Each HPE Primera storage system features a high-speed, full-mesh passive interconnect that joins multiple controller nodes (the high-performance data movement engines of the HPE Primera architecture) to form an all-active cluster. This low-latency interconnect allows for tight coordination among the controller nodes and a simplified software model.

In every HPE Primera storage system, each controller node has at least one dedicated link to each of the other nodes that operates at 8 GiB/s in each direction. Also, each controller node may have one or more paths to hosts—either directly or over a SAN. The clustering of controller nodes enables the system to present hosts with a single, highly available, high-performance storage system. This means that servers can access volumes over any host-connected port—even if the physical storage for the data is connected to a different controller node. The extremely low-latency full-mesh backplane enables a system wide unified cache which is global, coherent and fault tolerant.

HPE Primera storage is the ideal platform for mission-critical applications, for virtualization and cloud-computing environments. The high performance and scalability of the HPE Primera architecture is well suited for large or high-growth projects, consolidation of mission-critical information, demanding performance-based applications, and data lifecycle management. High availability (HA) is also built into the HPE Primera architecture through full hardware redundancy. Controller node pairs are connected to dual-ported drive enclosures. Unlike other approaches, the system offers both hardware and software fault tolerance by running a separate instance of the HPE Primera OS on each controller node, thus facilitating the availability of your data. With this design, software and firmware issues—a significant cause of unplanned downtime in other architectures—are greatly reduced.

HPE Primera ASIC

At the heart of every HPE Primera system is the HPE Primera ASIC, which is designed and engineered for NVMe performance. There are up to four ASIC slices per node, and each ASIC is a high-performance engine, which moves data through dedicated PCIe Gen3 high-speed links to the other controller nodes over the full-mesh interconnect. An HPE Primera 600 storage system with four nodes has 16 ASICs totaling 250 GiB/s of peak interconnect bandwidth. These interconnects each have 64 hardware queues with priority control to meet the low-latency and high-concurrency demands of an NVMe-centric architecture.

Each HPE Primera ASIC, known as a slice, has a dedicated hardware offload engine to accelerate RAID parity calculations, perform inline zero detection, and calculate deduplication hashes. The ASICs also automatically calculates CRC Logical Block Guards to validate data stored on drives with no additional CPU overhead. This technology enables the Persistence Checksum feature that delivers T10-PI (Protection Information) for end-to-end data protection (against media and transmission errors) with no impact to applications or host OSs. A fourth ASIC slice is also dedicated for internode communication completing the full-mesh all-active architecture.

Full-mesh controller backplane

The HPE Primera full-mesh backplane is a passive circuit board that contains slots for up to four controller nodes. As noted earlier, each controller node slot is connected to every other controller node slot by at least one 8 GiB/s full-duplex high-speed link (16 GiB total throughput), forming a full-mesh interconnect between all controller nodes in the cluster—something that Hewlett Packard Enterprise refers to as an all-active design.

[1] 100% Availability Guarantee

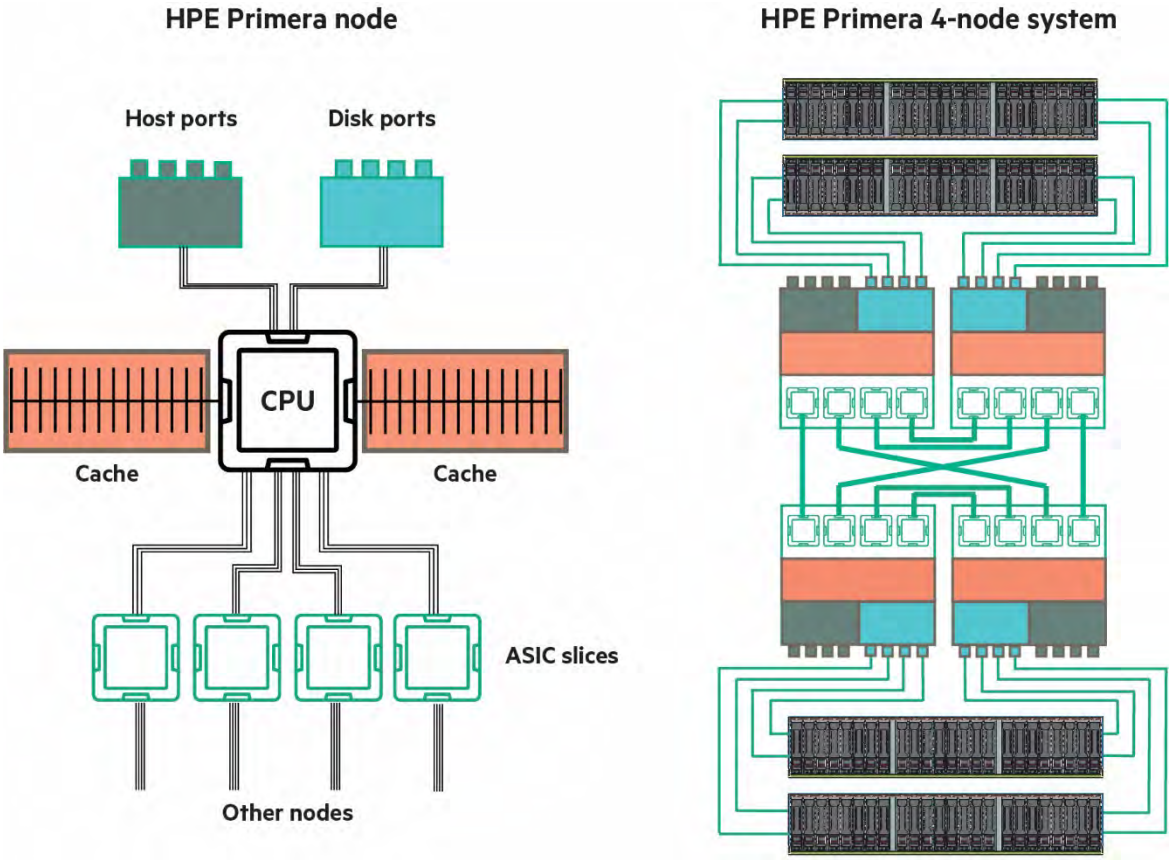

FIGURE 1. HPE Primera controller node design and the full-mesh all-active cluster architecture

These interconnects utilize a low overhead protocol that features rapid internode messaging and acknowledgment. Also, a completely separate full-mesh network of 1 Gb Ethernet links provides a redundant channel of communication for exchanging control information between the nodes.

Active/Active versus all-active

Most traditional array architectures fall into one of two categories: monolithic or modular. In a monolithic architecture, being able to start with smaller, more affordable configurations (that is, scaling down) presents challenges. Active processing elements not only have to be implemented redundantly, but they are also segmented and dedicated to distinct functions such as host management, caching, and RAID/drive management. For example, the smallest monolithic system may have a minimum of six processing elements (one for each of three functions, which are then doubled for redundancy of each function). In this design—with its emphasis on optimized internal interconnectivity—users gain Active/Active processing advantages (for example, LUNs can be coherently exported from multiple ports). However, these architectures typically involve higher costs relative to modular architectures.

In traditional modular architectures, users are able to start with smaller and more cost-efficient configurations. The number of processing elements is reduced to just two because each element is multifunction in design—handling host, cache, and drive management processes. The trade-off for this cost-effectiveness is the cost or complexity of scalability. Because only two nodes are supported in most designs, scale can only be realized by replacing nodes with more powerful node versions or by purchasing and managing more arrays. Another trade-off is that dual-node modular architectures, while providing failover capabilities, typically do not offer truly Active/Active implementations where individual LUNs can be simultaneously and coherently processed by both controllers.

The HPE Primera architecture was designed to provide cost-effective single-system scalability through a unified, multinode, clustered implementation. This architecture begins with a multifunction node design and, like a modular array, requires just two initial controller nodes for redundancy. However, unlike traditional modular arrays, enhanced direct interconnects are provided between the controllers to facilitate all-active processing. Unlike legacy Active/Active controller architectures—where each LUN (or volume) is active on only a single controller—this all-active design allows each LUN to be active on every controller in the system, thus forming a mesh. This design delivers robust, load-balanced performance and greater headroom for cost-effective scalability, overcoming the trade-offs typically associated with 2-node modular and monolithic storage arrays.

System-wide striping

The HPE Primera all-active design not only allows all volumes to be active on all controllers but also promotes system-wide striping that automatically provisions and seamlessly stripes volumes across all system resources to deliver high, predictable levels of performance. The system-wide striping of data provides high and predictable levels of service for all workload types through the massively parallel and fine-grained striping of data across all internal resources (drives, ports, cache, processors, and others). As a result, as the use of the system grows—or in the event of a component failure—service conditions remain high and predictable.

For flash-based media, fine-grained virtualization combined with system-wide striping drives uniform I/O patterns, thereby spreading wear evenly across the entire system. Should there be a media failure, system-wide sparing also helps guard against performance degradation by enabling many-to-many rebuild, resulting in faster rebuilds. Because HPE Primera storage automatically manages this system-wide load balancing, no extra time or complexity is required to maintain an efficient system.

Controller node architecture

The most important element of the HPE Primera architecture is the controller node. It is a powerful data movement engine that is designed for mixed workloads. As noted earlier, a single system, depending on the model, is modularly configured as a cluster of two or four controller nodes. This modular approach provides flexibility, cost-effective entry footprint, and affordable upgrade paths for increasing performance, capacity, and connectivity as needs change. Also, the minimum dual-controller configuration means that the system can withstand an entire controller node failure without impacting data availability. Controller nodes can be added in pairs to the cluster nondisruptively and upgraded to a more powerful controller node model, and each node is completely hot pluggable to enable online serviceability.

The controller nodes are designed to handle the concurrency demands of the NVMe era. Each HPE Primera 600 controller node can have up to 12 host ports, 8 drive enclosure ports, 40 CPU cores, and 4 HPE Primera ASICs to facilitate the massive parallelism necessary. The controller nodes are also designed with the flexibility to support the various connection technologies and topologies of today and the future where they are Fibre Channel (FC), iSCSI, or NVMeoF based.

The HPE Primera ASICs are used to perform RAID parity calculations on the cached data and the Zero Detect mechanism built into the ASICs removes streams of zeroes present in I/O prior to writing data to the back-end storage system in order to reduce capacity requirements and prolong SSD lifespan. The HPE Primera ASIC is also a crucial element of the system's ability to perform inline, block-level deduplication with Express Indexing (see the "Deduplication with Express Indexing" section for more details).

HPE PRIMERA SOFTWARE ARCHITECTURE

Services-centric OS

A unique OS in its class, the HPE Primera OS is a modular, service-centric design, which is one of the key enablers for the 100% Availability Guarantee. Features such as the I/O stack, RAID, cluster communication, HBA drivers, remote copy, and data reduction are implemented as independent services within the HPE Primera OS. This means that unlike traditional monolithic storage platforms the HPE Primera OS can be updated, upgraded, and extended without the need to reboot the controllers. This enables faster, more frequent updates that are easier to install and significantly less risky to perform than on other high-end storage systems. This radically simplified update process enables HPE Primera to provide a pipeline of innovation that can easily be tapped so that as new features become available, they can be added in minutes.

Highly virtualized

To help ensure performance and improve the utilization of physical resources, the HPE Primera OS is highly virtualized multiple layers of abstraction.

This fine-grained virtualization approach divides each physical disk into granular allocation units referred to as chunklets, each of which can be independently assigned and dynamically reassigned to different logical disks (LDs) that are used to create virtual volumes (VVs). This enhances performance for all applications as required capacity will be virtualized and striped across dozens or even hundreds of drives. It also helps eliminate stranded capacity because allocations are made in small increments from a disk to a LUN.

Multiple layers of abstraction

The first layer of abstraction employed by the HPE Primera OS divides media devices into 1 GiB chunklets to enable higher utilization and avoid stranded capacity. This fine-grained virtualization unit also enables new media technologies such as Storage Class Memory.

The second layer of abstraction takes the 1 GiB chunklets created from abstracting physical drive capacity and creates LDs striped across the system's physical drives (PDs) and implementing the RAID level. Multiple RAID sets made from chunklets of different PDs are striped together to form an LD. All chunklets belonging to a given LD will be from the same drive type. LDs can consist of all NL, FC, or SSD chunklets. Also, the first- and second-level mappings taken together serve to massively parallelize workloads across all the physical drives behind a node.

VVs are the virtual capacity representations that are ultimately exported to hosts and applications as virtual LUNs (VLUNs) over FC target ports. A single VV can be coherently exported through as few as two ports or as many as ports as desired (not fewer than two, one from each of two different nodes as a minimum).

The third layer of abstraction maps LDs to VVs with a granularity of 32 MiB or 128 MiB. With this approach, a very small portion of a VV associated with a particular LD can be quickly and nondisruptively migrated to a different LD for performance or other policy-based reasons, whereas other architectures require migration of the entire VV. This layer of abstraction also implements many high-level features such as snapshots, caching, and remote replication.

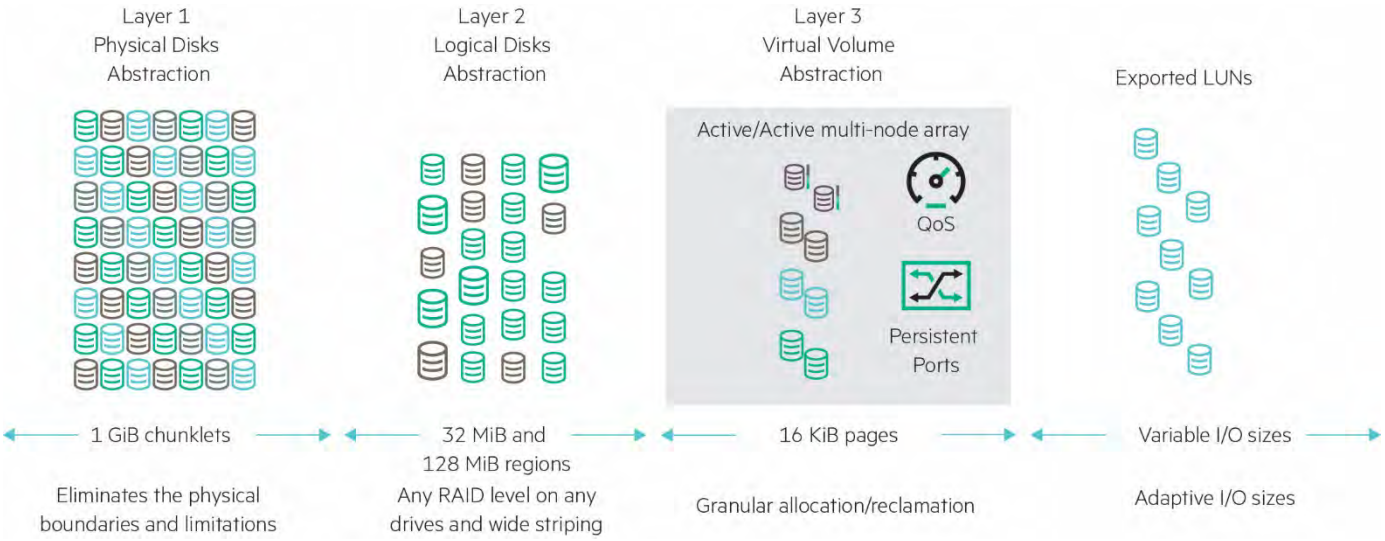

FIGURE 2. Virtualization with a tri-level mapping methodology that provides three layers of abstraction

The 3-layer abstraction implemented by the HPE Primera OS can effectively utilize any underlying media type. This means that HPE Primera storage is able to make the most efficient use of SSDs through load balancing across all drives to enable ultra-high performance and prolong flash-based media life span.

Physical disks

Every PD that is admitted into the system is divided into 1 GB chunklets. A chunklet is the most basic element of data storage of the HPE Primera system and forms the basis of the RAID sets. Depending on the sparing algorithm and system configuration, some chunklets are allocated as spares. Unlike many competitive arrays that reserve dedicated spare drives that then sit idle, system-wide sparing with HPE Primera storage means that spare chunklets are distributed across all drives. This provides additional protection and enables a balanced load that extends the SSD life span by providing even wearing. It also protects against performance degradation by enabling a **many-to-many** rebuild in the event of a failure.

Logical disks

There are two types of LDs:

- Shared data (SD) LDs provide the storage space for thin-provisioned VVs (TPVV), reduction VVs, and snapshots (virtual copies).
- Shared administration (SA) LDs provide the storage space for the metadata used for VVs and snapshots.

As mentioned earlier, RAID functionality is implemented at the LD level, with each LD mapped to chunklets in order to implement RAID 6 (multiple distributed parity, with striping).

The HPE Primera OS will automatically create LDs with the desired availability and size characteristics.

Every LD has an **owner** and a **backup** owner. Using the default layout, chunklets from any given PD are owned by a single node with the partner node as the backup owner; thus, every node creates LDs from the PDs it **owns**.

Common provisioning groups

A common provisioning group (CPG) creates a virtual pool of LDs that allows VVs to share the CPG's resources and allocate space on demand. You can create TPVVs and data reduction VVs that draw space from the same CPG LD pool.

CPGs enable fine-grained, shared access to pooled logical capacity. Instead of prededicating LDs to volumes, a CPG allows multiple volumes to share the buffer pool of LDs. For example, when a TPVV is running low on user space, the system automatically assigns more capacity to the TPVV by mapping new regions from LDs in the CPG to the TPVV. As a result, large pockets of unused but allocated space are eliminated.

The CPGs dynamically allocate storage in increments, which are determined by the number of nodes in the system and the RAID set size that has automatically been selected. This on-demand allocation unit determines the automated response to the growth demands of volumes. To grow the volume, the HPE Primera OS may expand existing LDs according to the CPGs growth increment or create additional ones. Growth is triggered when the CPG's available space falls below 85% of the growth increment value. A mechanism with warnings and limits can be configured on the array to control the growth of a CPG.

Virtual volumes

There are two kinds of VVs: base volumes and snapshot volumes. A base volume directly maps all the user-visible data, which can be considered to be the original VV and is either (a) a TPVV or reduction VV or (b) a snapshot volume created using HPE Primera Virtual Copy software. When a snapshot is first created, all of its data is mapped indirectly to the parent volume's data. When a block is written to the parent, the original block is copied from the parent to the SD space and the snapshot points to this data space instead. This methodology is known as Copy-on-Write (CoW). Similarly, when a block is written in the snapshot, the data is written in the SD space and the snapshot points to this data space.

VVs have three types of space:

- The **user space** represents the user-visible size of the VV (that is, the size of the SCSI LUN seen by a host).
- The **shared data space** is used to store the data of the VV and any modified data associated with snapshots. The granularity of data mapping is 16 KiB pages.
- The **shared admin space** is used to store the metadata (including the page tables) for VVs and snapshots.

Each of the three space types is mapped to LDs, with all of these LDs striped across all controller nodes; thus, VVs can be striped across multiple nodes for maximum load balancing and performance.

All user created VVs are TPVV. A TPVV has space for the base volume allocated from the associated CPG and snapshot space allocated from the associated snapshot CPG (if any). If data reduction is selected when creating a TPVV then common pages of data will be shared with other data reduction volumes in the CPG and the remaining data will be compressed. The data shared is determined via the inline deduplication mechanism described later in this paper. Data reduction is supported only on CPGs that use SSDs as a tier of storage. The size limit for an individual TPVV without data reduction is 64 TiB and with data reduction, it is 16 TiB.

On creation, 256 MiB per node is allocated to a VV. Storage is allocated on demand in the SD area. The SA area contains the metadata indexes that point to the user data in the SD area. Because the SA metadata needs to be accessed to locate the user data, the indexes are cached in policy memory to reduce the performance impact of the lookups.

User created VVs associated with a common CPG share the same LDs and draw space from that pool as needed, allocating space on demand in small increments for each controller node. As the volumes that draw space from the CPG require additional storage, the HPE Primera OS automatically allocates additional 256 MiB increments to the volumes.

VLUNs and LUN masking

VVs are only visible to a host once the VVs are exported as VLUNs. VVs can be exported as follows:

- To specific hosts (set of Worldwide Names [WWNs])—The VV is visible to the specified WWNs, regardless of which ports those WWNs appear on. This is a convenient way to export VVs to known hosts.
- To specific hosts on a specific port.

Optimized for NVMe and Storage Class Memory

HPE Primera all-active architecture, system-wide striping, fine-grain virtualization, advanced metadata handling, and system-wide sparing are just some of the pillars of HPE Primera architecture that deliver on the promise of NVMe and Storage Class Memory. Flash-based media can deliver many times the performance of conventional spinning HDDs, and it can do so at very low, sub-millisecond latency. However, it is important to understand that these advantages can only be realized by an architecture that has optimized its entire I/O path to be performance centric. If the storage controllers that sit between servers and back-end flash devices can't keep up with the performance of the flash drives, they become performance bottlenecks.

To work with flash-based media in the most performance-optimized manner, the HPE Primera architecture includes features designed to handle the flash media in a substantially different way than spinning media, or even SAS-attached NAND. It also exploits every possible opportunity to extend flash-based media life span by reducing factors that contribute to media wear.

- **Express Layout:** This unique technology from the HPE Primera 3-layer virtualization technology allows HPE Primera controller nodes to share access to SSDs in order to further drive efficiency. Replacing traditional layouts for flash, Express Layout allows each SSD to be actively accessed by both controllers in a node pair at the same time. This allows a node pair to use capacity from every drive to build logical capacity. For smaller configurations, like an 8-drive system, Express Layout allows the nodes to significantly reduce the overhead historically associated with parity RAID layouts and can result in more than 10% reduction in overhead in conjunction with increased performance by allowing more than one controller to deliver I/O through to the drive.
- **Adaptive Sparing:** The HPE Primera architecture extends SSD media utilization and endurance through patented Adaptive Sparing technology. HPE collaborates with SSD suppliers to release capacity typically reserved for wear management to allow HPE Primera systems access to a higher amount of drive capacity. This is achieved by reducing capacity typically reserved by media suppliers for wear management and then using that space more efficiently. At a system level, increasing usable drive capacity also helps spread writes more broadly to extend SSD endurance.
- **Cache Offload:** Cache Offload is a flash optimization that eliminates cache bottlenecks by changing the frequency at which data is offloaded from cache to flash media based on system utilization. This ensures consistently higher performance levels as the system scales performance to hundreds of thousands and even millions of IOPS. New writes coming into the array are acknowledged to the host as soon as the I/O gets written to cache in two nodes for protection. The in-cache write then gets flushed to the storage media at a rate based on cache utilization. At higher levels of utilization, HPE Primera increases the frequency at which flushes take place that allows the system to deliver consistent performance without running into cache bottlenecks, even at extreme performance levels.

HIGH AVAILABILITY

With HPE Primera storage, you can securely partition resources within a shared infrastructure in order to pool physical storage resources for lower storage costs without compromising security or performance.

The HPE Primera storage platform was built from the ground up to deliver multitenant capacity that supports massive consolidation with ultra-high performance. The multicontroller scalability and extreme flexibility built into HPE Primera storage makes deploying and maintaining separate storage silos to deliver different QoS levels outdated. To support multiple tenants and workloads, HPE Primera storage provides secure administrative segregation of users, hosts, and application data. The following sections provide insight into the architectural elements that support each of these core capabilities.

Tier-0 resiliency

HPE Primera storage is designed to support massive consolidation by supporting mixed workloads and secure administrative segregation of users, hosts, and application data. Multitenancy allows IT organizations to deliver higher performance levels, greater availability, and next-generation functionality securely to multiple user groups and applications from a single storage system.

Today's IT realities—including complex infrastructure, constant firefighting, and fragmented data silos—demand the ability to deliver predictable service levels in an inherently unpredictable world, and make system resiliency the single most important requirement. Traditionally, tier-0 storage has been characterized by hardware redundancy, advanced replication capabilities, and massive scalability in capacity and host connectivity.

Hardware and software fault tolerance, as well the ability to predictably prevent downtime and handle failures in a way that is nondisruptive to users and applications, become critical. The HPE Primera architecture allows you to consolidate with confidence and achieve higher service levels for more users and applications with less infrastructure.

Hardware and software fault tolerance

HPE Primera storage delivers tier-0 resiliency with an architecture designed to eliminate any single point of failure (hardware or software) in the system. To mitigate single points of failure at the hardware layer, the system is designed with redundant components, including redundant power domains. In fact, to raise the bar with the fault tolerance mechanism, HPE Primera 650/670 storage systems are configured with two self-encrypting boot drives that work in redundant mode.

An independent copy of HPE Primera OS runs on every controller node, so even in the smallest configuration, with two controller nodes, the system offers resiliency also for the software stack.

HPE Primera storage components such as storage nodes, disk- and host-facing host bus adapters (HBAs), power supplies, batteries, and disks all feature N+1 and in some cases N+2 redundancy so that any of these components can fail without system interruption. The only nonredundant component in the system is a 100% completely passive controller node backplane that, given its passive nature, is virtually impervious to failure.

HPE Primera storage offers up to four current load-balanced power distribution units (PDUs) per rack, which provide a minimum of two separate power feeds. The system can support up to four separate data center power feeds, providing even more power resiliency and further protection against power loss as well as brownouts.

Controller nodes in an HPE Primera storage system include redundant physical drives that contain a separate instance of the HPE Primera OS as well as space to save cached write data in the event of a power failure.

The controller nodes are each powered by two (1+1 redundant) power supplies and backed up by two batteries. Each battery has sufficient capacity to power the controller nodes long enough to flush all dirty data in cache memory into the local physical drive in the event of a complete power failure to the node. Although many architectures use battery-backed RAM as cache (to maintain the data in cache while waiting for power to be restored), these are not suitable for extended downtimes that are usually associated with natural disasters and unforeseen catastrophes.

Another common problem with many battery-powered backup systems is that it is often impossible to ensure that a battery is charged and working. To address this problem, the HPE Primera storage controller nodes are each backed by at least two batteries. Batteries are periodically tested by slightly discharging one battery while the other remains charged and ready in case a power failure occurs while the battery test is in progress. Following a power failure HPE Primera OS keeps track of battery charge levels and limits the amount of write data that can be cached based on the ability of the batteries to power the controller nodes while they are recharging following the power failure.

The HPE Primera storage power failure protection mechanisms eliminate the need for expensive batteries to power all of the system's drive chassis while dirty cache data is destaged to disks on the back end of the array. Note that, because all cached write data are mirrored to another controller node, a system-wide power failure would result in saving cached write data onto the internal drives of two nodes. This offers further protection following a power failure in the event a node in the cluster being damaged by the power failure. The second node containing the data can be used for recovery of the saved data. Because each node's dual power supplies can be connected to separate AC power cords, providing redundant AC power to the system can further reduce the possibility of an outage due to an AC power failure.

Advanced fault isolation

Advanced fault isolation and high reliability are built into the HPE Primera storage system. The drive chassis, drive magazines, and physical drives themselves all report and isolate faults. A drive failure will not result in data being unavailable.

HPE Primera storage constantly monitors drives via the controller nodes and enclosures, isolates faults to individual drives, and then **offlines** only the failed component.

Each drive enclosure has two redundant I/O modules that plug into the drive chassis midplane. The drive chassis components—power supplies, I/O modules, and drives—are serviceable online. Redundant power supply/fan assemblies hot plug into the rear of the midplane. Should the drive chassis midplane fail for any reason, partner cage or cages will continue to serve data for those volumes that were configured and managed as HA Cage volumes. If the **HA Cage** configuration is available at volume creation, the controller node automatically manages the RAID 6 data placement to accommodate the failure of an entire cage without affecting data access.

Controller node redundancy

The HPE Primera OS instance running on each of the controller nodes is both statefully managed and self-healing, providing protection across the all-active storage controller nodes should one or more processes fail and restart.

Also, controller nodes are configured in logical pairs whereby each node has a partner. The partner nodes have redundant physical connections to the subset of physical drives owned by the node pair. Within the pair, each serves as the backup node for the LDs owned by the partner node. If a controller node were to fail, data availability would be unaffected since the node's partner takes over the LDs for the failed node.

HPE Primera RAID protection

Exponential growth in SSD capacity without commensurate improvements in reliability or performance results in greater risk of data loss. For example, consider the 15.36 TB SSDs available on HPE Primera storage systems. The capacity difference alone implies that reconstruction of a failed disk upon replacement can be expected to take more than 4X longer than a 3.84 TB drive. This creates a larger window of vulnerability during which a second disk failure could cause data loss when using RAID 1 or RAID 5. RAID 6 addresses this problem by using two different parity values, which allows the data to be reconstructed, even in the event of two drive failures.

The HPE Primera RAID 6 implementation uses a forward error correction method based on Erasure Coding and offers multiple distributed parities with striping. Today, 2 parity blocks are supported with either a 4+2 (that is, 4 data blocks and 2 parity blocks), 6+2, 8+2, 10+2 configuration, and this can be extended to support 3 parity blocks in the future. In an appropriately configured HPE Primera array, all available RAID options allow HPE Primera storage to create parity sets on different drives in different drive cages with separate power domains for greater integrity protection.

Data integrity checking

Supplementing the hardware fault tolerance, all HPE Primera storage systems offer automated end-to-end error checking during the data frames' journey through the HPE Primera storage array to the disk devices to help ensure data integrity in support of tier-0 resilience. In addition to this HPE Primera ASIC comes with the Persistent Checksum feature known as T10 Data Integrity Feature (T10-DIF) that ensures end-to-end data protection, from host HBA to physical drives.

Embedded Cyclical Redundancy Checking (CRC) includes, but is not exclusive to, the following layers within all HPE Primera storage systems:

- CRC/parity checks on all internal CPU and serial buses
- Control cache ECC checks
- Data cache ECC checks
- PCIe I2C bus CRC/parity checks
- HPE Primera ASIC connection CRC/parity checks
- Protocol (FC) CRC checks at the frame level (hardware accelerated via the HPE Primera ASICs)
- Disk devices CRC checks at the block level, occurring once the data has landed and throughout the lifecycle of the data once it's stored to disk

CRC error checking is also extended to replicate data with HPE Primera Remote Copy software, which helps ensure that potential cascaded data issues do not occur. HPE Primera storage replication includes a link pre-integration test to verify the stability of Remote Copy replication links in advance for use with HPE Primera Remote Copy over an IP network (RCIP).

All drives in the HPE Primera 600 storage systems are formatted with 520-byte blocks in order to provide space to store a CRC Logical Block Guard, as defined by the T10-DIF for each block. This value is computed by the HPE Primera HBA before writing each block and is checked when a block is read. NL SAS does not support 520-byte blocks, so on Enterprise NL SAS drives, data blocks are logically grouped with an extra block to store the CRC values. The CRC Logical Block Guard used by the T10-DIF is automatically calculated by the host HBAs to validate data stored on drives without additional CPU overhead.

HPE Primera storage continuously runs a background **PD scrubber** process to scan all blocks of the physical drives in the system. This is done to detect any potential issues at the device block layer and trigger RAID rebuilds down to 512-byte granularity if necessary. This is particularly important when it comes to flash media because it allows the system to proactively detect and correct any low-level CRC and bit errors.

Furthermore, Self-Monitoring, Analysis and Reporting Technology (SMART) predictive failures mean that any disk device crossing certain SMART thresholds would cause the storage controller nodes to mark a drive as **predictive failure**, identifying it for replacement before it actually fails.

HPE Primera storage systems also issue logical error status block (LESB) alerts if a frame arriving in the storage interface has CRC errors beyond a certain threshold. This indicates that a cable or component between the host and storage device needs replacing or cleaning.

Persistent technologies

No one has time for downtime, which is why modern tier-0 resiliency requires that data access and service levels be maintained during failure recovery, maintenance, and software upgrades. Tier-0 resiliency demands that failures not only be prevented, but that the system can recover quickly in the event that something goes wrong. Not only is HPE Primera storage designed to be nondisruptively scalable and upgradable, but the system also has several advanced features to prevent unnecessary downtime and to maintain availability and performance levels during planned as well as unplanned outage events. These features are collectively known as persistent technologies.

Persistent Checksum

Persistent Checksum addresses media and transmission errors that can be caused by any component in the I/O stack from the server HBA through the SAN switches and into the HPE Primera HBAs making the data secure all the way from the hosts right to the drives and providing additional protection above CRC transmissions for FC alone. Persistent Checksum is server and application independent (it does require server HBAs that support the feature) and offers elaborate host OS support. When using unsupported HBAs, T10-DIF tags are added and verified on the array target ports, internode copies, and back-end HBAs. When using supported HBAs, T10-DIF tags are added by the host HBAs and verified throughout the HPE Primera storage system, making the data secure all the way from the hosts to the drives. Where Persistent Checksum detects media or transmission errors, graceful error recovery will take place, avoiding impact on the host application.

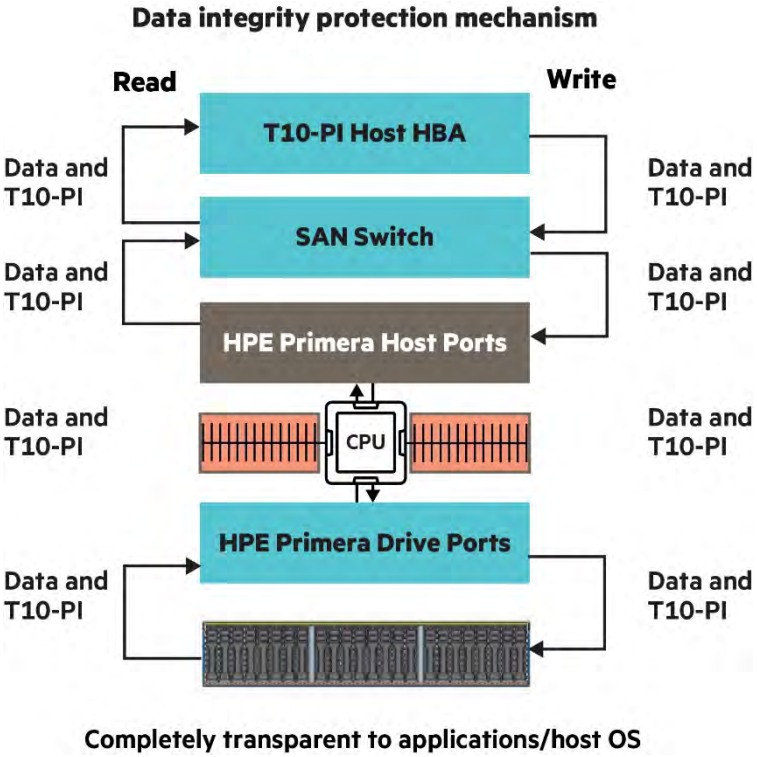

FIGURE 3. Persistent Checksum

Persistent Cache

HPE Primera Persistent Cache is a resiliency feature built into the HPE Primera OS that allows graceful handling of an unplanned controller node failure or planned maintenance of a controller node. This feature eliminates the substantial performance penalties associated with traditional modular arrays and the cache **write-through** mode they have to enter under certain conditions. HPE Primera storage can maintain high and predictable service levels even in the event of a cache or controller node failure by avoiding cache write-through mode via HPE Primera Persistent Cache.

With HPE Primera Persistent Cache, when a write I/O occurs, the node receiving the request will mirror the data to the cache of another node in the cluster. This can be any other node not just the node pair partner. In the event of controller node failure, the surviving node of a pair does not have to go into write-through mode for the LDs it owns, as it will continue to mirror the remaining cluster nodes, so data integrity is ensured in the unlikely event it should fail too.

Persistent Ports

HPE Primera Persistent Ports allow for a nondisruptive environment (from the host multipathing point of view) where host-based multipathing software is not required to maintain server connectivity in the event of a node or link outage on any SAN fabric. This applies to firmware upgrades, node failures, and node ports that are taken offline either administratively or as the result of a hardware failure in the SAN fabric that results in the storage array losing physical connectivity to the fabric.

From a host standpoint, connections to HPE Primera storage systems continue uninterrupted with all I/O being routed through a different port on the HPE Primera storage array. This helps you achieve an uninterrupted service level for applications running on HPE Primera storage systems.

The Persistent Port functionality works for FC the transport layer and provides transparent and uninterrupted failover in response to the following events:

- HPE Primera OS firmware upgrade
- HPE Primera node maintenance or failure
- HPE Primera array **loss sync** to the FC fabric
- Array host ports being taken offline administratively
- Port laser loss for any reason (applies to FC only)

HPE Primera Replication software

HPE Primera Replication software brings a rich set of features that can be used to design disaster-tolerant solutions that cost-effectively address disaster recovery challenges. It is a uniquely easy, efficient, and flexible replication technology that allows you to protect and share data from any application.

Implemented over a native IP network (through the built-in 10GbE interface available on all nodes) users may flexibly choose one of two different data replication modes—**asynchronous periodic** (for asynchronous replication) or **synchronous**—to design a solution that meets their solution requirements for recovery-point objectives (RPOs) and recovery-time objectives (RTOs).

Synchronous replication

It provides zero data loss in the event of failure for the ultimate RPO but can have an impact on host performance. As spinning media solutions measure performance in tens of milliseconds, creating an exact copy of data over an extended distance adds some latency, but it's generally acceptable in terms of meeting service-level agreements (SLAs). All-flash systems are far more sensitive to latency overheads as performance is now measured in microseconds, so any overhead measured in milliseconds can significantly increase latency. The overhead associated with replicating every write request over an IP link twice (round trip) has this impact.

Asynchronous periodic

Based on snapshots and delta resyncs, asynchronous periodic replication has minimal impact on host performance, but it does require compromise as RPOs are measured in minutes, not seconds or milliseconds. This may be suitable for many environments where RPOs in minutes are acceptable but data compliance and business requirements can often drive the need for lower RPOs. Changed data within an HPE Primera Remote Copy volume group is transferred only once—no matter how many times it may have changed—between synchronization intervals. Additionally, efficiencies in the initial copy creation of the target volumes that do not require replication of **zero** data across the replication network (regardless of target volume type, thin or reduce) result in a faster initial synchronization and better network utilization.

Privacy, security, and multitenancy

Updated security enhancements

Security concerns plague the corporate environment daily. New threats, old threats, hacks, and outright nefarious actions threaten corporate data at alarming rates. HPE Primera helps to mitigate those threats by use of upgrades and patches. HPE Primera changes the architecture by which we can push fixes for Common Vulnerabilities and Exposures (CVE) in a relatively short period due to HPE Primera OS residing in user space and not kernel space. We no longer need to wait on a new kernel build of the OS to address vulnerabilities. With HPE Primera, we just create a patch and push it down to users using HPE InfoSight, and the user can apply the patch during a scheduled maintenance window or immediately upon arrival without disruption to the user environment. This approach ensures the HPE Primera array is up to date to guard against latest known CVE.

Virtual Domains

HPE Primera Virtual Domains software is an extension of HPE Primera virtualization technologies that delivers secure segregation of virtual private arrays (VPAs) for different user groups, departments, and applications while preserving the benefits delivered by the massive parallelism architected into the HPE Primera platform. It supports HPE Primera multitenancy paradigm.

By providing secure administrative segregation of users and hosts within a consolidated, massively parallel HPE Primera storage system, HPE Primera Virtual Domains allows individual user groups and applications to affordably achieve greater storage service levels (performance, availability, and functionality).

HPE Primera Virtual Domains is completely virtual and represents no physical reservation of resources. To use HPE Primera Virtual Domains, a master administrator first creates a virtual domain, and then assigns logically defined entities to it. These include one or more host definitions based on WWN groupings, one or more provisioning policies (disk type), and one or more system administrators (who are also granted role-based privileges by the master administrator).

Depending on the level of access, users can create, export, and copy VVs. HPE Primera Virtual Domains is ideal for enterprises or service providers looking to leverage the benefits of consolidation and deploy a purpose-built infrastructure for their private or public cloud.

Data encryption

Data is perhaps the most important asset for organizations in today's digital age. Companies are looking to protect data against theft and misuse while meeting compliance requirements. The HPE Primera storage complies with the standards set forth by the National Institute of Standards and Technology (NIST) and Federal Information Processing Standard (FIPS) 140-2 and features data-at-rest (DAR) encryption that helps protect valuable data through self-encrypting drive (SED) technology. SED drives are HDDs and SSDs with a circuit (ASIC) built into the drive's controller chipset that automatically encrypts and decrypts all data being written to and read from the media.

HPE Primera storage supports full-disk encryption (FDE) based on the Advanced Encryption Standard (AES)-256 industry standard. The encryption is part of a hash code that is stored internally on physical media. All encryption and decryption is handled at the drive level and needs no other external mechanism.

Authentication keys are set by the user and can be changed at any time. The local key manager (LKM) included with the HPE Primera storage encryption license is used to manage all drive encryption keys within the array and provides a simple management interface. In the event of a drive failure or the theft of a drive, a proper key sequence needs to be entered to gain access to the data stored within the drive. When an SED drive is no longer powered on, the drive goes into a locked state and requires an authentication key to unlock the drive when power is restored. Without the key, access to the data on the SED is not possible.

HPE also offers enhanced encryption support on the HPE Primera storage systems by offering FIPS 140-2 compliant SED drives that provide the ability to use an external enterprise secure key manager (ESKM). ESKM is deployed whenever you use encrypted storage or communication methods to protect sensitive information. Herein, you store and serve keys to unlock the data stored on FIPS 140-2 compliant drives within the HPE Primera storage systems with strong access controls and security.

FIPS 140-2 compliance provides you the satisfaction of knowing that your data is securely stored on the HPE Primera array. Key management on the array, with either LKM or ESKM coupled with FIPS drives, offers you a safe environment to securely store your data.

Transport Layer Security (TLS) 1.2-only support

HPE Primera OS allows TLS 1.2-only configurations, which eliminate any potential impact of security vulnerabilities by preventing TLS 1.0/1.1 connections, which allows you, as an HPE Primera customer, to enhance their Payment Card Industry Data Security Standard (PCI DSS) 3.2 compliance strategy.

General Data Protection Regulation

The General Data Protection Regulation (GDPR) is a new European privacy law, which came into force on May 25, 2018 and significantly increased the risks for companies that fail to use and protect personal data in compliance with the law. The GDPR introduces significant monetary penalties of up to a maximum of 20 million Euros or 4% of the annual worldwide turnover of a corporate group. The GDPR requires organizations to implement appropriate technical and organizational measures to secure data and introduces new breach notification requirements.

HPE Primera storage by its inherent design and architecture with security that is built into the product will assist you in meeting your GDPR security requirements. HPE Primera security categories can be identified as the following:

- Authorization
- Authentication
- Availability
- Encryption
- Integrity
- Auditing

These categories are all fundamental security aspects by which HPE Primera continues to enhance and harden the overall product architecture. HPE Primera has already and will continue to adopt security by design into its OS, appliances, and tools, which support the array.

MAINTAINING HIGH AND PREDICTABLE PERFORMANCE LEVELS

The ability of HPE Primera storage to maintain high and predictable performance in all environments is made possible through architectural innovations that utilize all available array hardware resources at all times, thereby eliminating resource contention, supporting mixed workloads, and enhancing caching algorithms to accelerate performance and reduce latency.

Load balancing

Purpose-built for the enterprise as well as virtual and cloud data centers, the HPE Primera architecture is unlike legacy controller architectures. Its all-active system design allows each volume to be active on any controller in the system via the high-speed, full-mesh interconnection that joins multiple controller nodes to form a cache-coherent Active/Active cluster. As a result, the system delivers symmetrical load balancing and utilization of all controllers with seamless performance scalability by adding more controllers and disk drives to the system.

Priority optimization

Quality of service (QoS) is an essential component for delivering modern, highly scalable multitenant storage architectures. The use of QoS moves advanced storage systems away from the legacy approach of delivering I/O requests with **best effort** in mind and tackles the problem of **noisy neighbors** by delivering predictable tiered service levels and managing **burst I/O** regardless of other users in a shared system. Mature QoS solutions meet the requirements of controlling service metrics such as throughput, bandwidth, and latency without requiring the system administrator to manually balance physical resources. These capabilities eliminate the last barrier to consolidation by delivering assured QoS levels without having to physically partition resources or maintain discreet storage silos.

HPE Primera Priority Optimization software enables service levels for applications and workloads as business requirements dictate, enabling administrators to provision storage performance in a manner similar to provisioning storage capacity. This allows the creation of differing service levels to protect mission-critical applications in enterprise environments by assigning a minimum goal for I/O per second and bandwidth, and by assigning a latency goal so that performance for a specific tenant or application is assured. It is also possible to assign maximum performance limits on workloads with lower service-level requirements to make sure that high-priority applications receive the resources they need to meet service levels.

The Priority Optimization feature and industry-leading **latency goal** feature enables the storage administrator to set SLAs as low as 500µs for volumes residing on SSD storage. It also makes it possible to configure service-level objectives in terms of KB/s and I/O bandwidth on a VV set (VVset) or between different virtual domains. All host I/Os on the VVset are monitored and measured against the defined service-level objective. HPE Primera Priority Optimization control is implemented within the HPE Primera storage system and can be modified in real time. No host agents are required, and physical partitioning of resources within the storage array is not necessary.

Performance benefits of system-wide striping

In a traditional storage array, small volumes either suffer from poor performance by using few drives or waste expensive resources by using more drives than required for capacity in order to obtain sufficient performance. On HPE Primera storage systems, even modest-sized volumes will be widely striped using chunklets spread over multiple drives of the same type. Wide striping provides the full performance capabilities of the array (nodes, CPUs, buses, cache, disk drives) to small volumes without provisioning excess capacity and without creating hotspots on a subset of physical drives.

Additional details about striping are provided in the "Multiple layers of abstraction" section.

Sharing and offloading of cached data

Because much of the underlying data associated with snapshot volumes is physically located on the base VVs, data that is cached for the base VV can often be used to satisfy read accesses for a snapshot of that base VV.

In the event that three or more drives that underlay a RAID 6 set become temporarily unavailable—for example, if all cables to those drives are accidentally disconnected—the HPE Primera OS automatically moves any **pinned** writes in cache to dedicated Preserved Data LDs. This helps ensure that all host-acknowledged data in cache is preserved so that it can be properly restored once the destination drives come back online without compromising cache performance or capacity with respect to any other data by keeping cache tied up.

On flash-based systems, Cache Offload mitigates cache bottlenecks by automatically changing the frequency at which data is offloaded from cache to flash media. This helps ensure high performance levels consistently, as workloads are scaled to hundreds of thousands of IOPS.

Write caching

Writes to VVs are cached in a controller node, mirrored in the cache of another controller node, and then acknowledged to the host. The host, therefore, sees an effective response time that is much shorter than would be the case if a write were actually performed to the drives before being acknowledged. This is possible because the mirroring and power failure handling help ensure the integrity of cached write data.

In addition to dramatically reducing the host write response time, write caching can often benefit back-end drive performance by:

- Merging multiple writes to the same blocks so that many drive writes are eliminated
- Merging multiple small writes into single larger drive writes so that the operation is more efficient
- Merging multiple small writes to a RAID 6 LD into full-stripe writes so that it is not necessary to read the old data for the stripe from the drives
- Delaying the write operation so that it can be scheduled at a more suitable time

Capacity efficiency

Thin provisioning

HPE Primera Thin Provisioning makes storage more efficient and more compact by dedicating space on demand, allowing you to purchase only the disk capacity you actually need and only as you actually need it.

Thin Persistence is a feature that keeps VVs and read/write snapshots of VVs small by detecting pages of zeros during data transfers and not allocating space for the zeros. This feature works in real time and analyzes the data before it is written to the source VV or read/write snapshot of the VV. Freed blocks of 16 KB of contiguous space are returned to the source volume, and freed blocks of 128 MB of contiguous space are returned to the CPG for use by other volumes.

Thin copy reclamation keeps storage as lean and efficient as possible by reclaiming the unused space resulting from deleted Virtual Copy snapshots. After a snapshot is deleted, the shared space is reclaimed from the VV and returned to the CPG for reuse by other volumes. Deleted snapshot space can be reclaimed from virtual copies, physical copies, or Remote Copy volumes.

Data reduction technologies

HPE Primera Data Reduction combines deduplication and compression to help maximize space savings. With data reduction volumes, the incoming data is checked for duplicates before being compressed.

Deduplication with Express Indexing

Deduplication is a technology designed to eliminate duplicate information from being committed to disk. The HPE Primera ASIC features a dedicated, high-performance, low-latency hashing engine used for deduplication that can lead to not only massive savings over standard deployment methodologies but also a much smaller performance overhead when deduplication is enabled. Deduplication employs Express Indexing, a mechanism that provides extremely high-performance lookup tables for fast detection of duplicate write requests.

When a new write request enters cache, a hash (or fingerprint) of the data is generated in order to draw a match against other data stored on the array. Generating a hash of every data write is an extremely CPU-intensive task and many software-implemented hashing algorithms commonly found in all-flash platforms add a significant performance overhead to write performance. With HPE Primera Deduplication software, the CPU-intensive jobs of calculating hash signatures for incoming data and verifying reads are offloaded to the ASICs, freeing up processor cycles to perform other critical data services.

Once the hash has been calculated, Express Indexing performs instant metadata lookups in order to compare the signatures of the incoming request to signatures of data already stored in the array. If a match is found, the system flags the duplicate request and prevents it from being written to the back end. Instead, a pointer is added to the metadata table to reference the existing data blocks. To prevent hash collision (when two write pages have the same signature but different underlying data), HPE Primera Deduplication software leverages the controller node ASICs once again to perform a high-performance bit-to-bit comparison before any new write update is marked as a duplicate, preventing incorrect data matches. This is an important step to prevent data corruption and should be at the core of any deduplication implementation.

FIGURE 4. Deduplication—the process of removing data blocks, which are exact duplicates

This hardware-assisted approach enables an inline deduplication implementation that carries multiple benefits, including increased capacity efficiency, flash performance protection, and flash media life span extension. The combination of hardware-assisted hashing and Express Indexing is powerful and efficient.

Compression

While deduplication looks for opportunities to remove entire blocks of data by comparing them against each other, compression works by looking for opportunities to reduce the size of pages before they're written to flash. When compression and deduplication are enabled together, duplicate blocks are removed first and then the remaining data is compressed.

FIGURE 5. Compression—the process of shrinking the size of data blocks

HPE Primera implements an extremely efficient, modern compression algorithm that can deliver supreme performance for both compression and decompression tasks while yielding excellent compression savings. HPE Primera implements Express Scan, a technology that further reduces the CPU overhead associated with compression. This is achieved by inspecting blocks to ensure data is compressible, rather than wasting CPU cycles attempting to compress data identified as incompressible. Read and write performance profiles are very important with compression; write performance needs to be high to support incoming write streams of data but since writes are cached in system memory before they're committed to flash, compression is essentially an asynchronous task so doesn't impact write latency as heavily. However, reads are far more sensitive because not all reads are delivered from cache; whenever a read **miss** occurs (when a read is requested and it's not in cache), the array must read the back-end data, decompress it, and return it to the host. Performance here is the key, as latency will increase as decompression performance reduces.

Data Packing

Once data has been compressed, the result is a number of data blocks that are smaller but are also odd sizes (for example, 1.3 KiB, 4.2 KiB, 5.6 KiB.). These blocks are not only odd sizes, but they're very hard to write to flash since flash pages are fixed in size—writing these pages directly to flash will result in lost performance and reduced efficiency, neither of which are desirable. Data Packing addresses this issue by packing these odd-sized pages together into a single page before they're written to flash.

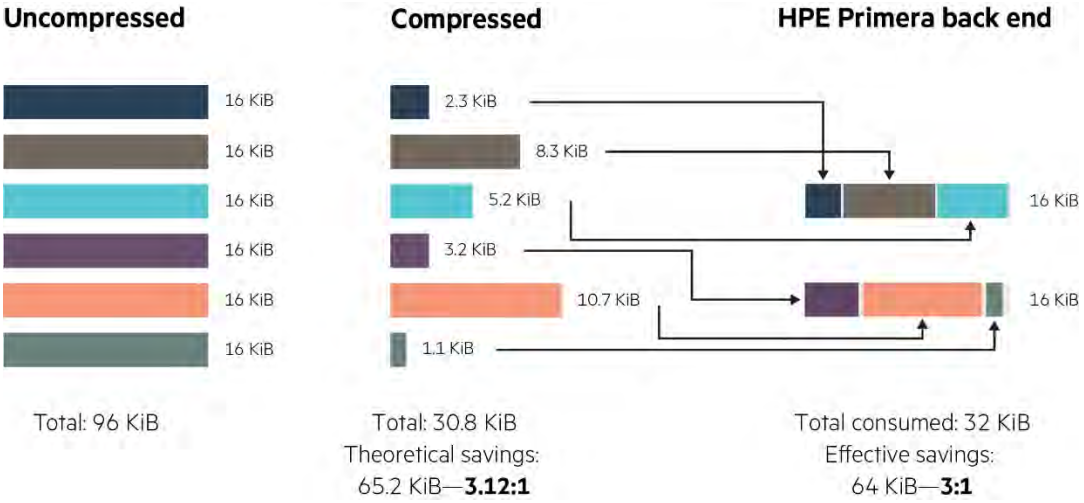

FIGURE 6. Multiple pages stored together through Data Packing

HPE Primera systems use 16 KiB physical pages to store data. When data reduction is used, HPE Primera Data Packing allows multiple compressed pages to be stored in a single 16 KiB physical page. This Data Packing technique is part of the inline process that not only optimizes physical space but also creates larger, more media efficient write sizes than other approaches, which improves both performance and media endurance. If an overwrite occurs, the system will refit the new compressed page in place if there is available space. However, if the new data will not fit into the existing page, it will be queued for packing with other new writes. HPE Primera Express Indexing technology, common to all HPE Primera thin volume types, is used to track the data within the compressed pages.

Compaction
In addition to Data Reduction, HPE Primera systems offer additional capacity-efficiency technologies that include market-leading hardware-accelerated Thin Provisioning, Thin Clones, Thin Reclaim, Virtual Copy, and other technologies. Savings from these technologies are not included in Data Reduction (or data reduction ratios) and are instead reported in the compaction ratio—a ratio that exposes the full suite of efficiency technologies. Therefore, compaction is the combination of Thin technologies and Data Reduction.

Virtual Copy
Virtual Copy is the HPE Primera snapshot implementation used to provide a point-in-time virtual copy of a VV and ensure that the original data can always be obtained should a problem occur when updating the data on a VV. Virtual Copy implements an efficient variant of CoW mechanism. For CoW, the HPE Primera OS uses a delayed copy on write (DCoW) that eliminates any performance impact to host I/O; DCoW is used for snapshots of thin-provisioned volumes. With HPE Primera DCoW, the reading of the original data, updating of the base volume with the new data, and the copy of the original data occurs in the background after the write update has been acknowledge to the host.

Virtual copies are always thin, reservationless, with only one copy of a changed block being kept. Thanks to efficient metadata handling, you can configure thousands of read-only and read-write snapshots. Flexible management allows any snapshot to be promoted without destroying any other snapshots.

Data migration
HPE Primera Peer Motion Utility (PMU) software is a nondisruptive, do-it-yourself data mobility tool for enterprise block storage that does not require any external appliance to be included in the datapath nor does it introduce any additional overhead on the host resources. Unlike traditional block migration approaches, HPE PMU allows data to be migrated between any two HPE Primera storage systems without complex planning or dependency on extra tools. The HPE PMU tool also enables nondisruptive data migration from non-HPE Primera storage systems. HPE Primera Peer Motion software leverages the same built-in technology that powers the simple and rapid inline thin conversion of inefficient fat volumes on source arrays to more efficient, higher-utilization thin volumes on the destination HPE Primera storage system.

STORAGE MANAGEMENT
HPE Primera OS helps simplify, automate, and expedite storage management by handling provisioning and change management automatically and intelligently, at a subsystem level, and without administrator intervention.

The system's user interfaces have been developed to offer automatic administration, which means that the interfaces allow an administrator to create and manage physical and logical resources without requiring any overt action. Provisioning does not require any preplanning, yet the system constructs volumes intelligently based on available resources, unlike manual provisioning approaches that require planning and the manual addition of capacity to intermediary pools.

Ease of use
HPE Primera OS reduces training and administration efforts through the simple, point-and-click HPE Primera UI, the unified HPE 3PAR SSMC application, and the scriptable HPE Primera Command Line Interface (CLI). These management options provide uncommonly rich instrumentation of all physical and logical objects for one or more storage systems, thus eliminating the need for the extra tools and consulting often required for diagnosis and troubleshooting.

Open administration support is provided via SNMP, Storage Management Initiative Specification (SMI-S), and Web Service API.

HPE Primera UI is an easy-to-use GUI for managing and servicing a single HPE Primera storage system. The HPE Primera UI software is included in each HPE Primera storage system and does not require installation on a separate server.

The HPE Primera UI delivers a simplified experience for on-premises storage infrastructure, including expanding a system and upgrading the HPE Primera OS. The initial systems installation can be performed in as little as 20 minutes and systems can be expanded nondisruptively in 10 minutes.

HPE 3PAR SSMC is a GUI that provides contemporary, browser-based interfaces for monitoring and managing multiple HPE Primera and HPE 3PAR storage systems. The software is available as a virtual appliance download from the HPE Software Depot. The software can be deployed in several supported virtual machine environments.

HPE Primera Performance Insights eliminates the guesswork and time out of diagnosing bottlenecks and optimizing application performance. With Performance Insights, you are provided automated details that show the root cause to complex anomalies with application-aware insights eliminating tuning of infrastructure with painful trial-and-error methods.

Available in HPE 3PAR SSMC, Performance Insights offers machine-learned algorithms that are trained in the cloud and deployed on-premises. This capability enables a fast time to response and extends HPE InfoSight to dark sites. Performance Insights offer the following benefits:

- Identify when performance issue is due to saturation
- Plan workloads better with knowledge of available headroom
- Identify root cause to complex anomalies with application-aware insights

HPE InfoSight deeply integrates with the HPE Primera UI to allow enhanced predictive analytics to be performed within the HPE cloud. This predicts, prevents, and resolves problems, such as part failure, data availability, or data loss issues, across the infrastructure stack and ensures optimal performance and efficient resource use. HPE InfoSight watches over the infrastructure 24x7, continuously monitoring every system across the install base, so you don't have to spend days, nights, and weekends dealing with infrastructure issues.

Within each HPE Primera storage system, there are thousands of sensors. This instrumentation effectively tracks every I/O through the system and provides statistical information, including service time, I/O size, KB/s, and IOPS for VVs, LDs, and PDs. Performance statistics such as CPU utilization, total accesses, and cache hit rate for reads and writes are also available on the controller nodes that make up the system cluster. HPE InfoSight constantly analyzes and correlates millions of these sensors every minute to arrive at various insights that are made available to the storage administrator to act upon. The repertoire of intelligence grows daily in HPE InfoSight and new signatures are continually added for swift anomaly detection, creating a supremely powerful storage administration tool.

HPE Primera Web Services API (WSAPI) is an even more powerful and flexible way to manage HPE Primera storage systems. This API enables programmatic management of HPE Primera storage systems. Using the API, the management of volumes, CPGs, and VLUNs can be automated through a series of HTTPS requests. The API consists of a server that is part of the HPE Primera OS and runs on the HPE Primera storage system itself and a definition of the operations, inputs, and outputs of the API. The software development kit (SDK) of the API includes a sample client that can be referenced for the development of customer-defined clients.

HPE Primera PowerShell Toolkit provides Microsoft® Windows Server® cmdlets for accessing HPE Primera systems. The toolkit allows PowerShell scripts to use cmdlets that issue HPE Primera CLI commands or WSAPI calls to manage the logical objects of the HPE Primera system.

OpenStack® integration enables enterprises to increase agility, speed innovation, and lower costs. HPE is committed to the OpenStack community and has been a top contributor to the advancement of the OpenStack project. HPE's contributions have focused on continuous integration and quality assurance, which support the development of a reliable and scalable cloud platform that is equipped to handle production workloads. To support the need that many larger organizations and service providers have for enterprise-class storage, HPE has developed the HPE Primera block storage drivers, which support the OpenStack technology across the FC protocol. This provides the flexibility and cost-effectiveness of a cloud-based open source platform to you with mission-critical environments and high resiliency requirements.

HPE Smart SAN

SAN plays a critical role in any data center by providing access and connectivity between storage arrays and servers via a dedicated network. FC has been the dominant storage protocol that enjoys significant SAN market share. FC is popular for storage because of its enterprise-class performance, availability, and security. FC zoning is a key feature that adds to security and better management of the SAN by providing necessary segregation and allowing controlled communication among selected devices within a large fabric. However, configuring zones is a complex, tedious, and error-prone operation in a majority of SAN installations. Thus, signifying a need for automating these operations as much as possible to avoid human errors and reduce potential SAN downtime.

HPE Smart SAN for HPE Primera comes with a set of innovative features, one of which is automated zoning to address the above issues. Also, it also supports standards-based device registrations and diagnostic data collection for better configuration, visibility, and diagnostic purposes. Automated zoning, as implemented on HPE Primera as part of HPE Smart SAN 2.0, utilizes peer zoning as defined in FC standards, thus empowering HPE Primera storage system to configure zones automatically whenever hosts are provisioned on the target side.

MULTISITE RESILIENCY

HPE Primera Peer Persistence

HPE Primera Peer Persistence software enables HPE Primera storage systems located within a metropolitan distance to act as peers to each other for delivering a high-availability, transparent failover solution for the connected VMware vSphere®, Microsoft Hyper-V, and Microsoft Windows clusters. HPE Primera Peer Persistence allows an array-level, high-availability solution between two sites or data centers where failover and failback remains completely transparent to the hosts and applications running on those hosts. Unlike traditional disaster recovery models where the hosts (and applications) must be restarted upon failover, HPE Primera Peer Persistence allows hosts to remain online serving their business applications, even when the serving of the I/O workload migrates transparently from the primary array to the secondary array, resulting in zero downtime.

In an HPE Primera Peer Persistence configuration, a host cluster would be deployed across two sites and an HPE Primera storage system would be deployed at each site. All hosts in the cluster would be connected to both of the HPE Primera storage systems. These HPE Primera systems present the same set of VVs and VLUNs with same volume WWN to the hosts in that cluster. The VVs are synchronously replicated at the block level so that each HPE Primera storage system has a synchronous copy of the volume. A given volume would be primary on a given HPE Primera storage system at any one time. Using Asymmetric Logical Unit Access (ALUA), HPE Primera Peer Persistence presents the paths from the primary array (HPE Primera storage system on which the VV is primary) as **active/optimized** and the paths from the secondary array as **standby** paths. Issuing a switchover command on the array results in the relationship of the arrays to swap, and this is reflected back to the host by swapping the state of the paths from active to standby and vice versa. Under this configuration, both HPE Primera storage systems can be actively serving I/O under normal operation (albeit on separate volumes).

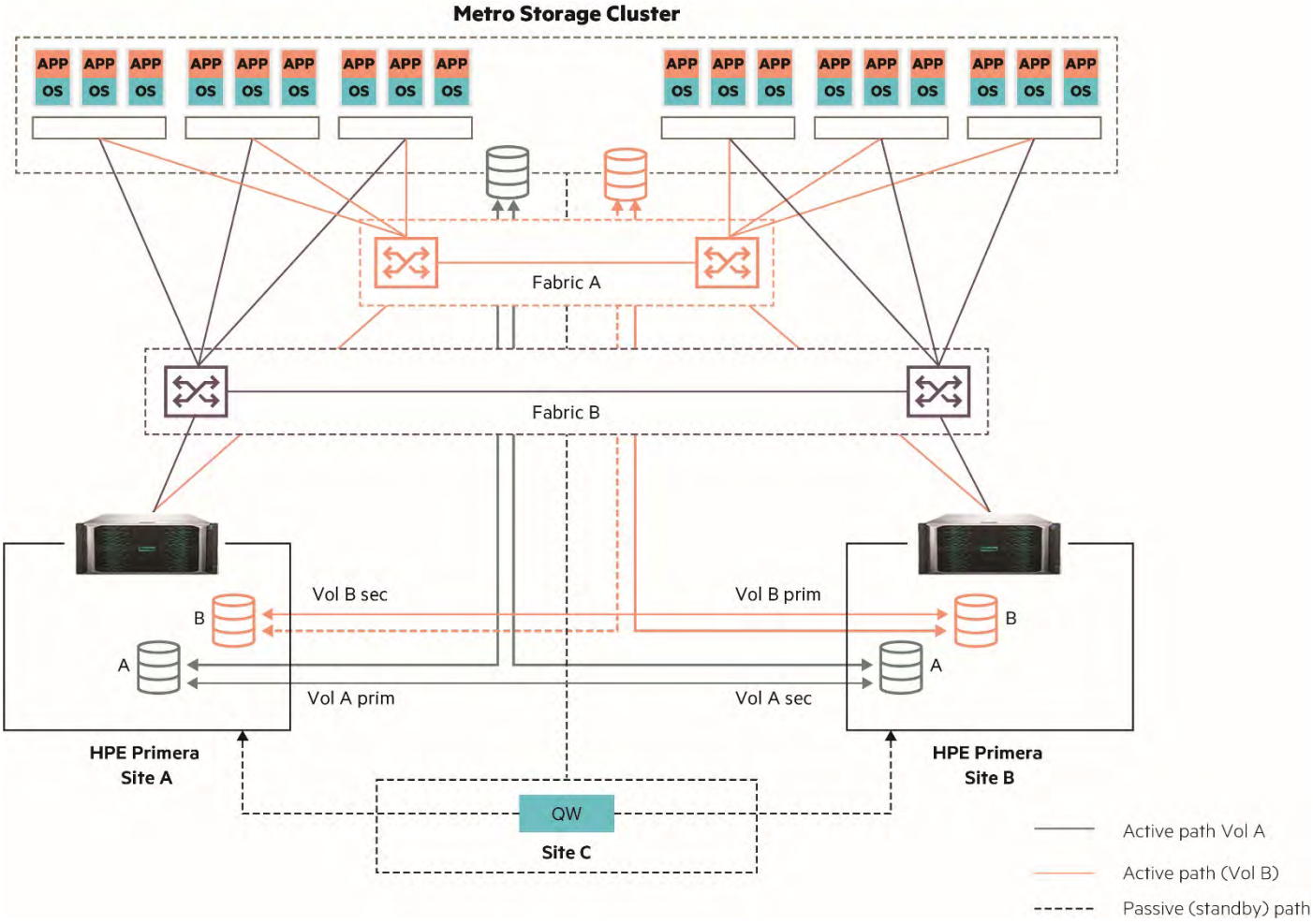

FIGURE 7. Transparent failover with HPE Primera Peer Persistence software

SIMPLIFIED SERVICEABILITY

HPE Primera systems offer remarkably simple serviceability. You can replace or upgrade HPE Primera hardware components via the HPE Primera UI. Online replacement is supported for nodes, DIMMs, boot drives, adapter cards, enclosure PCMs, node PCBMs, I/O modules, SFPs, and drives. Online upgrades are available for nodes, adapter cards, and drives.

Software serviceability is simple and quick with new releases and updates automatically downloaded from HPE InfoSight. After the download, the system notifies the administrator of the recommended action including the criticality of the update. Updates are performed online from the HPE Primera UI, with no controller node reboots, in minutes.

PROACTIVE SUPPORT

HPE support for HPE Primera storage provides a global support infrastructure that leverages advanced system and support architectures for fast, predictive response and remediation. The HPE Primera Secure Service Architecture provides secure service communication between the **HPE Primera storage systems** at your site and HPE support, enabling secure diagnostic data transmission and remote service connections. Key diagnostic information such as system health statistics, configuration data, performance data, and system events can be transferred frequently and maintained centrally on a historical basis. As a result, proactive fault detection and analysis are improved and manual intervention is kept to a bare minimum.

This implementation provides automated analysis and reporting that delivers accuracy and consistency, full system information in hand that reduces on-site dependencies, and fully scripted and tested automated point-and-click service actions that reduce human error.

HPE Primera storage systems include a built-in management console that monitors and enables remote monitoring and remote servicing of the array. This integrated storage setup minimizes the complexity of setup, installation, and usage for you.

The HPE Primera UI takes care of all service-related communications. It leverages the industry-standard HTTPS to secure and encrypt data for all inbound and outbound communications. The information collected and sent to HPE includes system status, configuration, performance metrics, environmental information, alerts, and notification debug logs. No data is sent.

The data sent is used by HPE support teams to proactively monitor the array and contact you if potential issues are discovered. You are warned proactively about potential problems before they occur. In the case of switch issues, you are advised of an issue and replacement parts are dispatched. Trained HPE service personnel can service the system at your convenience. If the management console cannot dial HPE for any reason, both the HPE Primera storage system and HPE support centers will receive alerts.

The HPE Primera UI is also used to download new patches, maintenance updates, new firmware revisions, and diagnostics. If remote access is needed for any reason, you can configure inbound secure access for OS upgrades, patches, and engineering access. If your data center does not permit **phone home** devices, then all alerts and notifications will be sent to your internal support team. You can then notify HPE support of an issue or suspected issue, either over the phone or via the web.

SUMMARY

In the intelligence era, new applications and workloads are creating a massive growth in data being created and actioned across hybrid cloud. Data is transformative only when it can be refined and accessed at the right place and at the right time, driving actionable insights into new revenue streams. However, extracting maximum value out of it is much easier said than done.

HPE Primera delivers intelligent storage with a tier-0 all-flash foundation to unlock the potential of your data. With HPE Primera, your storage is:

- **AI-driven:** HPE Primera uses advanced analytics and machine learning through HPE InfoSight to not only remove the burden of managing infrastructure but also serves as the foundation to provide context-aware intelligence about how your data should be managed.
- **Built for cloud:** HPE Primera applies intelligence to see, manage, and automate your storage no matter where your data lives. For example, powerful toolsets are available to automate and manage your HPE Primera for the cloud, DevOps, virtualization, and container environments. You can also effortlessly orchestrate intelligent, multi-tiered data protection from on-premises arrays to the public cloud—driven by policy and business need.
- **As-a-service experience:** HPE Primera delivers the flexibility to align to your specific consumption and investment needs. With as-a-service, on-premises storage solutions from HPE GreenLake, you get scalability and simplified IT operations—even operated for you by HPE—all in a pay-per-use model. Eliminate overprovisioning to save significantly on storage cost, deploy workloads when they are needed, and free up your staff to focus on core business initiatives.

HPE Primera storage does all this with a tier-0 all-flash foundation to support your mission-critical applications and beyond. Built to meet the extreme requirements of massively consolidated cloud service providers, HPE Primera enables you to confidently consolidate mixed and unpredictable workloads with ease. All HPE Primera models are built on a single flash-optimized architecture, run the exact same HPE Primera OS, and offer a common set of enterprise data services. Get ready for anything by starting small and scaling big with HPE Primera storage.

Resource

For detailed and up-to-date specifications on each of these products, see the product QuickSpecs:

- HPE Primera 600 Storage QuickSpecs

LEARN MORE AT

hpe.com/storage/hpeprimera

a50000189ENW, October 2019, Rev. 1

Hewlett Packard Enterprise