Technical white paper

Integrity Plus for HP StoreOnce Deduplication

Table of contents

Introduction

In any storage system it is essential to ensure that the integrity of the data stored is maintained so data can be recovered exactly as it was written. HP StoreOnce appliances have been designed with the necessary technology that delivers this essential high degree of data protection.

HP has unique technology that protects data throughout its lifecycle when stored on the HP StoreOnce appliance. This paper will discuss the methods used at various stages to provide this high degree of data integrity.

HP Integrity Plus for StoreOnce Deduplication leads the way in protecting data.

Executive summary

Data protection is essential to any IT organization. Regular backups and deduplication technology has enabled more data to be kept on disk based storage and seamlessly moved offsite. Not only are regular backups essential but the integrity of the data is equally vital. Storage administrators live in fear of corrupted backups only discovered on restore at a later date. HP StoreOnce technology has inbuilt protection that not only checks data at many stages in the process and when recovered but also continually checks the data at rest correcting errors if necessary. This paper will describe how the HP StoreOnce architecture takes care of data from backup server to disk spindle and back again.

HP StoreOnce—key features and benefits

HP StoreOnce deduplication, store more data on disk
HP StoreOnce deduplication reduces the disk space required to store backup data sets without impacting backup performance. Retaining more backup data on disk longer enables greater data accessibility for rapid restore of lost or corrupt files and reduces downtime.

Deduplication ratios are strongly influenced by two factors—data change rate and backup data retention periods. Low data change rates and data retained for longer periods of time yield higher deduplication ratios.

Optimized replication
HP StoreOnce deduplication is the technology enabler for HP StoreOnce replication, which allows fully automated replication without rehydration. Optimized replication can use low-bandwidth WAN links to a disaster recovery (DR) site. This is also a cost-effective DR solution for data centers and remote office/branch offices (ROBO).

Rapid restore of data for dependable, worry-free data protection
HP StoreOnce offers immediate access to backups for rapid restores. HP StoreOnce deduplication allows more data to be stored closer to the data center for longer periods of time, which offers immediate access for rapid restores.

Automate, simplify, and improve the backup process
HP StoreOnce automates the backup processes, allowing reduced time spent managing data protection. Implementing hands-free, unattended daily backup is especially valuable for environments with limited IT resources, such as remote or branch offices.

HP StoreOnce Catalyst
HP StoreOnce Catalyst technology allows backup applications to communicate directly with the StoreOnce appliance. This allows advanced features such as deduplication at the backup server for bandwidth-optimized backups and replication to one or more StoreOnce appliances without the overhead of rehydration. StoreOnce Catalyst is supported with HP Data Protector, Symantec NetBackup with OST, Backup Exec, Oracle RMAN, and BridgeHead Healthcare Software.

Data security
HP StoreOnce systems have optional built-in data encryption and a secure erase.

Technology overview

Deduplication works by examining the data stream as it arrives at the storage appliance, checking for small blocks[1] of data that are identical and eliminating redundant copies. If duplicate data is found, a pointer is established to the original set of data as opposed to actually storing the duplicate blocks, removing, or "deduplicating" the redundant data. The key here is that the data deduplication is being done at the block[2] level to remove far more redundant data than deduplication done at the file level where only duplicate files are removed. Data compression is used by HP StoreOnce prior to storing data. Data compression works at a byte level eliminating repetitive sequences of data up to around 2 KB.

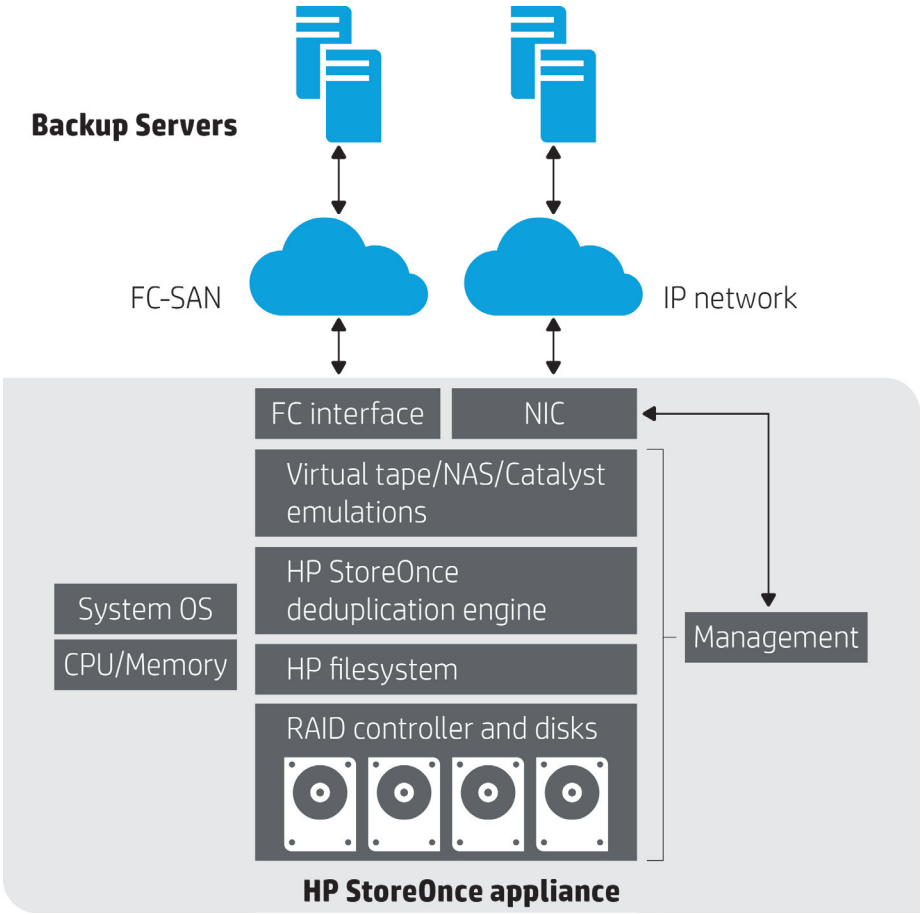
[1, 2] "Block" is sometimes referred to as "segment" in other deduplication technology.

Data deduplication is especially powerful when it is applied to backup, since most backup data sets have a great deal of redundancy. The amount of redundancy will depend on the type of data being backed up, the backup methodology, and the length of time the data is retained.

HP StoreOnce provides virtual tape (VT), NAS, or StoreOnce Catalyst target devices for data protection applications. Interfaces can be via a network connection or Fibre Channel (FC). Figure 1 shows the basic components of the StoreOnce appliance. The actual storage medium is hard disk and these are arranged in a RAID 6 configuration with an enterprise-class HP-designed RAID controller. Data is written across all disks in the RAID. RAID 6 prevents data loss in the case of two hard disk failures. RAID disks in current StoreOnce appliances are either 2 TB or 4 TB serial-attached SCSI (SAS) disk drives.

HP StoreOnce deduplication is also used to move backups to other HP StoreOnce appliances in a bandwidth efficient manner. This enables customers to move backups to another physical location often using a WAN connection with no human intervention. In the event of a total site loss, the data is still secure at the disaster recovery site and systems can be quickly restored.

Figure 1. Basic HP StoreOnce appliance components



StoreOnce data integrity in the deduplication process

For data integrity in a storage system such as HP StoreOnce, many steps are taken to help ensure that data is not corrupted as part of the process. There are also different strategies for maintaining and correcting data errors within the disk subsystem and in the deduplication process. These processes all need to be fast as data can be moving in or out of the StoreOnce appliance at high rates. Storage systems must also protect against power failure as data may be “in transit” somewhere in the system.

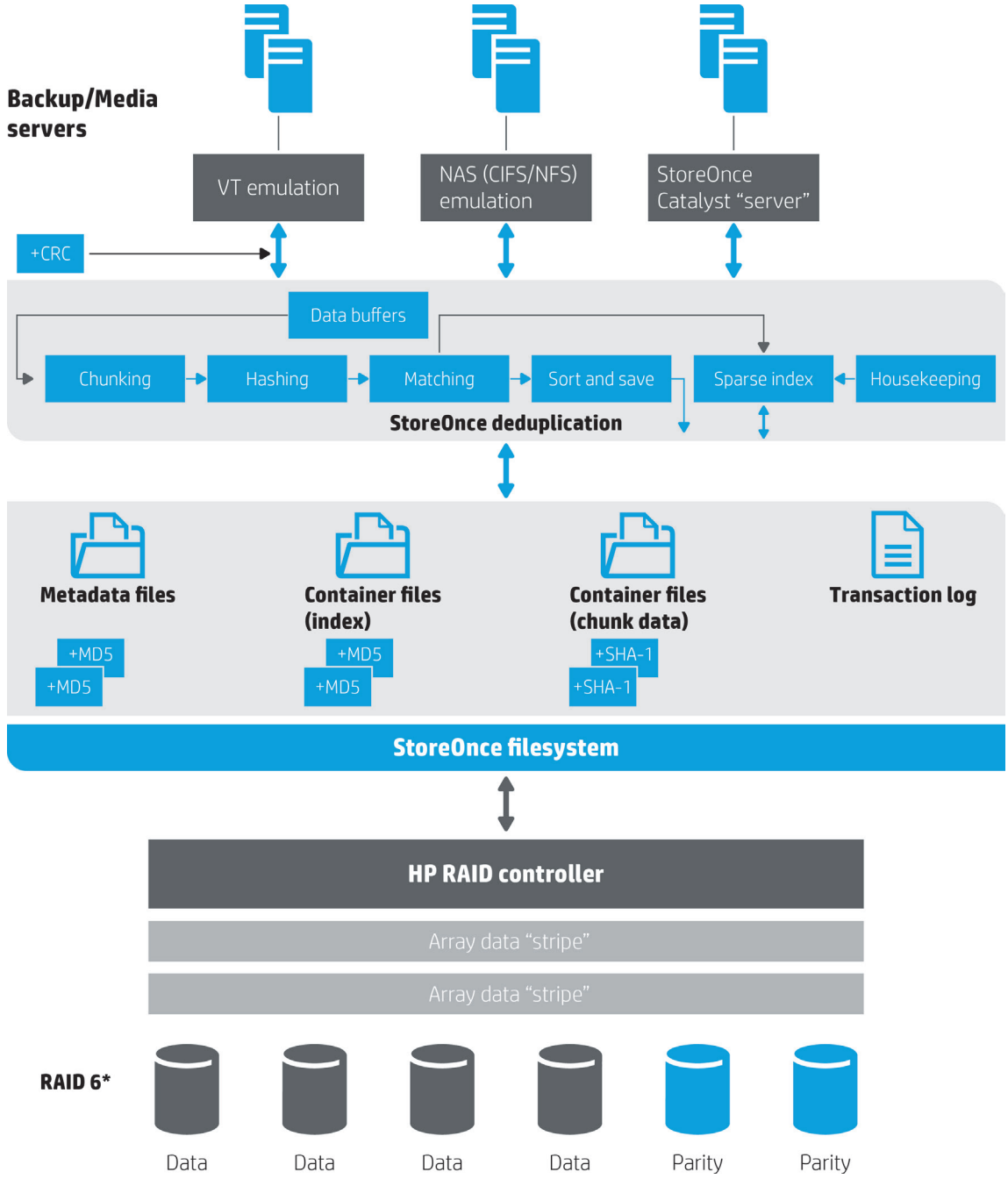
HP StoreOnce systems have a complete end-to-end verification process, which is necessary for storage devices designed for data protection. Data is often stored on StoreOnce appliances for long periods of time and customers need to have confidence that the data is stored correctly in the first place and checked at periodic intervals. When data is restored it is checked again for errors. The Integrity Plus technology in HP StoreOnce appliances is 100 percent HP intellectual property, much of it originating in HP Labs and developed by HP Storage R&D.

The end-to-end process of error checking starts with the backup server, which will send or receive data or commands using Fibre Channel or a network (IP) connection. Both Fibre Channel protocol and TCP/IP have systems that check data integrity of transfers between the host and the StoreOnce appliance. These error protection methods are well documented industry-standard protocols and will not be covered in this paper. Figure 2 shows the checks at various stages of the process from the interface connection, target device emulation, deduplication "engine," and storage system.

Let's consider the lifecycle of some data and its transformation from ingest, deduplication, long-term storage, and retrieval.

Data is written via the selected interface, which appears to the backup server as a virtual tape drive, NAS share, or HP StoreOnce Catalyst target. Critical to data integrity checking is that once the relevant acknowledgment is sent to the host server, the data can be retrieved following any power fail or system crash.

Figure 2. HP StoreOnce deduplication and storage process



* The entry level HP StoreOnce 2700 has only four disks in total and uses a RAID 5 configuration.

For every block of virtual tape data, a cyclical redundant checksum (CRC) is generated and stored with the data on disk. When the data is read back, a CRC is re-computed and compared with the original, and any discrepancy reports an SCSI check condition. NAS and Catalyst emulations do not apply a CRC at this stage in StoreOnce systems because of the fundamental differences in their protocol from serial devices, e.g., NAS share data may be modified at the request of the host server. However, both NAS and StoreOnce Catalyst have rigorous checks applied at different stages while en route to the StoreOnce appliance.

Note
HP StoreOnce 2700 systems are RAID 5.

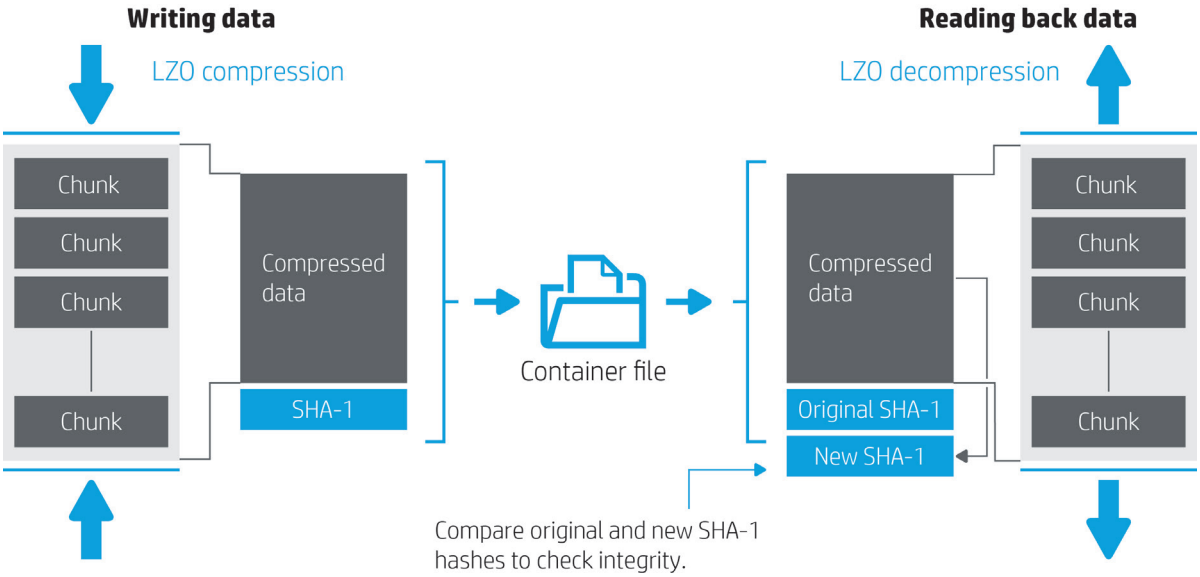
Data is now in transit and held in the main memory (RAM) of the StoreOnce system. The RAM of course has error detection and correction systems of its own. The deduplication process consists of dividing the data stream into "chunks" (sometimes called segments in other deduplication methods), which are of variable length and on average 4 KB in size. This process uses the HP Labs patented Two Thresholds Two Divisors (TTTD) technology to make intelligent decisions on how to divide the data stream. The chunks have an electronic signature applied in the form of a SHA-1 hash. Chunk hash codes are then matched against previous data received and, if completely new, are stored in an internal StoreOnce structure known as a container. As part of the matching process, a sparse index (HP patented) is used to quickly make decisions as to which previous containers are best for matching. The hash codes and container identifiers are stored separately as "metadata," which is essentially a "recipe" file for reconstructing or rehydrating the data. Chunk data is compressed before storage. Any identical chunks are not stored and an index count is simply incremented in the container index. The container index file contains a list of hash codes, a pointer to their location in the container files, and the compressed length. This data will enable chunk data to be recovered for rehydration.

As this process is working, a transaction log is being maintained. Depending on instructions received via the host connection, this log and critical information are "flushed" to disk. A good example would be the receipt of a write file mark instruction by a virtual tape device. This is important in the case of a power fail, which is covered later. It means that the system will recover all committed data, and uncommitted changes are reversed on appliance recovery. This is not unlike the behavior of a real tape. The sparse index, although not essential for data recovery, is also stored to disk as it grows after a set number of entries.

At this stage StoreOnce Integrity Plus adds a critical set of checksums. In order to maintain system performance, both SHA-1 and MD5 hashes are used to protect data stored in the container files and the metadata file. A SHA-1 hash is a unique 160-bit value for a given set of data. MD5 is a unique 128-bit value. For efficiency of data compression and filesystem performance, multiple chunks are compressed in a single operation and stored with each write to the container file. With each set of chunks, a SHA-1 hash is added following the compression process, which is checked when data is read back (figure 3). HP StoreOnce technology uses the SHA-1 hash because evaluation by HP Labs proved it to be faster than using a CRC check.

See glossary for more detailed description of the hashing algorithms.

Figure 3. Adding a SHA-1 checksum to chunk data



The index data and metadata files are much smaller than the container files so an MD5 hash provides adequate protection. HP has also chosen a data compression algorithm that has a safety mechanism to detect errors on de-compressing the data.

Note that once data is "committed," i.e., the host backup server has confirmation of a successful "write," StoreOnce makes sure that the data is successfully stored in the RAID subsystem.

When a host backup server reads back data, hash codes are re-computed from the retrieved data and checked against the hash that was stored.

In the case of virtual tape data, the additional CRC check is used.

So, at this stage HP StoreOnce has successfully stored the data and added several mechanisms to detect possible errors. The RAID controller now continually checks the integrity of the data stored and corrects any errors individual disk drives may cause as a background process. So, at this stage the data has been successfully stored on disk and has checksums in place for the restore process.

Housekeeping

Any deduplication technology requires a process of "cleaning" (HP StoreOnce systems call this "housekeeping") in order to reduce fragmentation and reclaim free space. This process runs in the background but HP customers have the option to select "blackout" periods usually at peak periods where it does not run.

If, for example, a virtual tape is overwritten (standard practice in a grandfather–father–son tape management cycle) that tape will be known to the HP StoreOnce system as a list of hash codes and container locations. If deduplication is working well then many of the chunks in the containers will be referenced by other tapes that are required. So, although that virtual tape may be overwritten, the old references remain and the index counts need to be adjusted. So, if this tape contained a particular chunk "xyz," which was used by three other virtual tapes, the index count would be four. It now needs to be reduced to three. StoreOnce of course cannot remove the chunk because the index count is greater than zero. After more overwrites of other media, this index count reaches zero and therefore the chunk is redundant.

However, it still occupies space in the container. After regular use, the containers become fragmented (lots of redundant chunks). The housekeeping jobs perform container compaction to remove this fragmentation. HP StoreOnce housekeeping is normally scheduled to avoid times when the appliance is busy (e.g., backups or restores in progress). There is an additional integrity checking process embedded in the housekeeping process. When container compaction takes place the SHA-1 hashes are checked and re-computed. This is yet another point at which the data is checked—the efficient SHA-1 algorithm enabling low extra loading.

RAID subsystem and StoreOnce filesystem

Storage administrators live in fear of corrupted data. We have seen above that the StoreOnce deduplication process takes great care to check for errors and protection of data that is in midflight within the appliance. HP StoreOnce Integrity Plus also extends to the filesystem and RAID subsystem where data resides for longer periods. Here, HP StoreOnce has the ability to not only detect errors but also correct and rebuild stored data.

HP StoreOnce uses a proprietary filesystem to store the user data. This has been designed for efficiency and use in the multi-node systems where a common filesystem is shared between two nodes that are running the StoreOnce system. It also allows for automatic load balancing over time. Following an unexpected system shutdown (e.g., power failure), it is necessary to check the consistency of the filesystem. The normal procedure is to run a filesystem consistency check (commonly known as "fsck"). This can be time consuming and has to be performed offline for standard filesystems. The custom HP filesystem can be checked online (online filesystem consistency check) and is multi-threaded for additional speed (Note: multi-threaded processes make good use of modern multi-core processors).

This is an important consideration where storage systems have usable capacities in the order of hundreds of terabytes. Not only that, the HP unique multi-node architecture has more processors than single-node competitor systems, which shares out this recovery process over multiple servers.

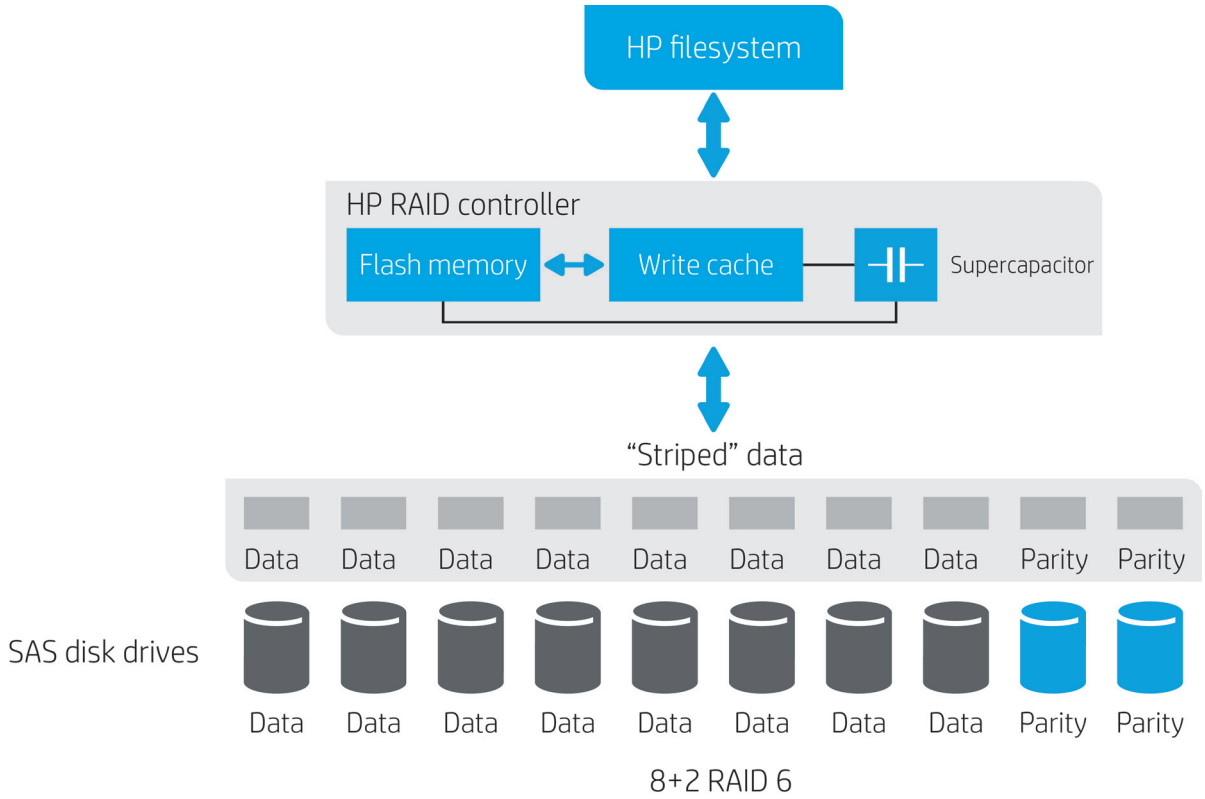
HP StoreOnce systems use proprietary HP hardware RAID controllers designed with specific data integrity protection technology. RAID is required because disk drives can produce uncorrectable errors at a rate of 1 in 1015 bits. RAID technology uses additional parity disks that can be used to correct errors. Apart from the entry-level HP StoreOnce 2700 system, which uses RAID 5, all other StoreOnce systems use RAID 6 disk storage systems (Note: The StoreOnce 2700 system has only four disk drives, which cannot be used in a RAID 6 configuration). RAID protects against failures of a whole disk and individual uncorrectable errors by writing data across multiple disk drive in "stripes" with one (RAID 5) or two (RAID 6) parity disks that can be used to reconstruct data. RAID 6 has the advantage that data integrity is preserved even if any two disks fail in the RAID (parity or data in any combination). RAID 5 can only survive one disk failure. Figure 4 shows a sample StoreOnce RAID. Often when referring to RAID, the data and parity disk quantities are specified—e.g., 8+1 in the diagram refers to eight data disks and two parity disks.

Protection against power fail

For improved performance, the HP-designed RAID controller has a 1 GB write cache memory module with a flash (non-volatile) memory backup.

Once again, protection is essential against power fail because blocks of data may be in cache and have been modified by filesystem writes but not written to disk (known as "dirty"). These could, in the case of HP StoreOnce, perhaps be new chunks added to the container that is in cache. If power is lost, the controller will copy the cache into a flash (non-volatile) memory. A supercapacitor is used to provide a small power reserve in order that data can be transferred from cache to flash memory. Flash memory requires no power to retain data. On restart from power failure, the flash memory data is returned to cache and then the "dirty" block is rewritten to the disk. A supercapacitor is a bit like a small battery providing a reserve of DC power. This is a superior solution to older technology that uses a rechargeable battery to maintain the cache memory backup system. However, battery power can only be maintained for a finite time and also rechargeable batteries deteriorate over time.

Figure 4. HP StoreOnce RAID



HP RAID controller continuous surface analysis

In addition to normal data reads and writes to the RAID subsystem, HP's RAID controller performs surface analysis in the background for ongoing data integrity checks. This is sometimes referred to as "scrubbing." This activity is performed independent of normal I/O activity as a background process by the RAID controller.

In the previous paragraph, it was described how data is written across all disks in the RAID with parity data written across two disks in a RAID 6 configuration. Each stripe is read in turn and the parity information re-calculated. The parity information is then compared with the retrieved parity data, which reveals any errors in the data or the parity information.

In this case the "stripe" is rewritten as a whole. This is performed at a logical volume level and is always secondary to normal read/write activity. The system is designed to scan all the data in each logical volume every 24 hours. Information is passed back to the StoreOnce system and the software monitors error rates and can trigger service requests for disk drive replacement. Any disk in the RAID can of course be replaced online.

In this way, HP StoreOnce helps ensure that disk errors are corrected on a regular basis.

It should be noted that individual drives have error correction mechanisms that run continually. However, increasing error rates although correctable are monitored. This is known as predictive maintenance and StoreOnce systems will generate alerts and flag drives for replacement during scheduled maintenance periods.

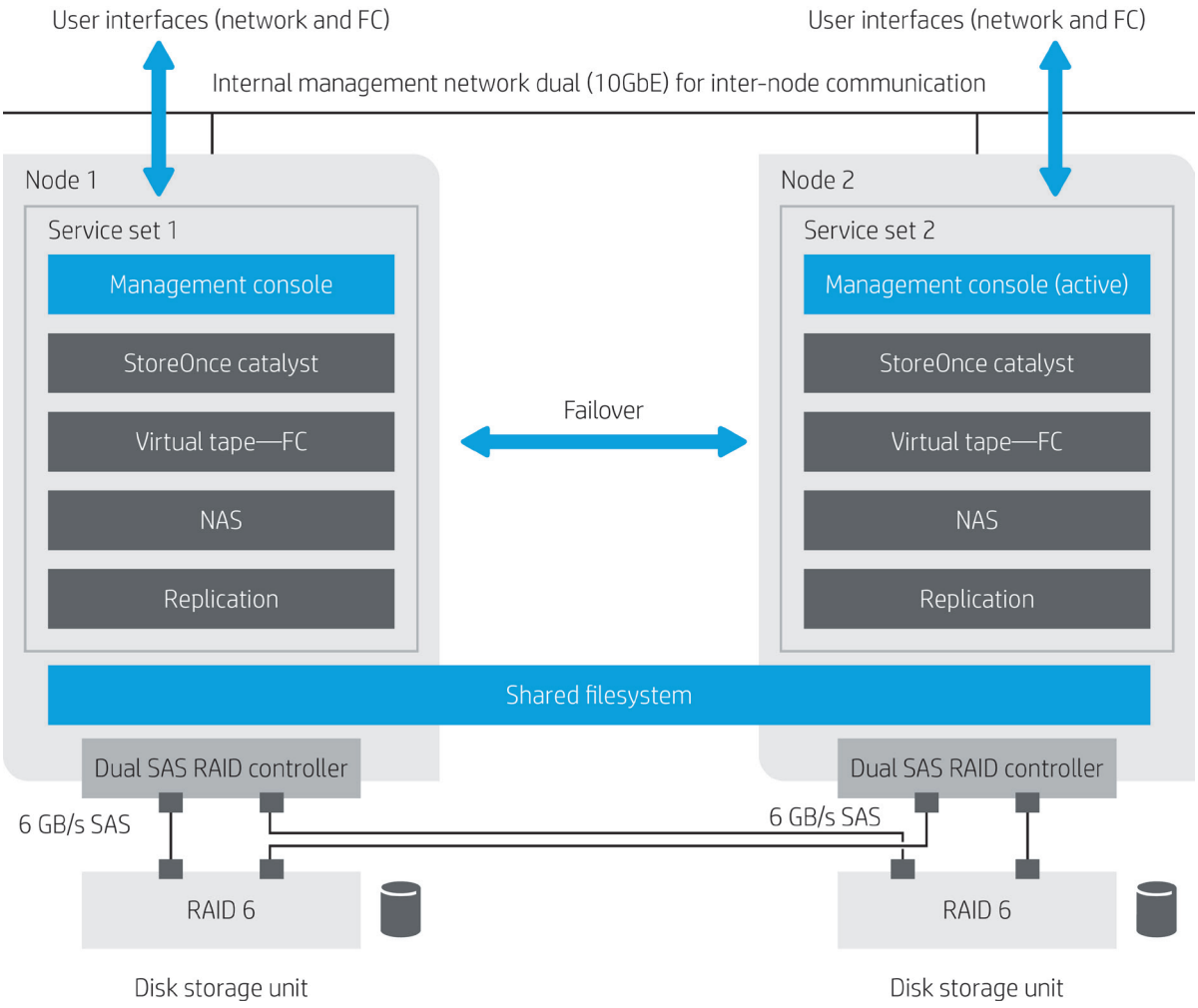
Hot spares

The larger HP StoreOnce models (HP StoreOnce 4900 and HP StoreOnce 6500) have "hot spare" disk drives. In the case of drive failure, the RAID controller will initialize the hot spare drive as part of the RAID while normal activity continues. The defective drive can be replaced under a routine maintenance activity without compromising the resilience of the system.

HP RAID controllers in the multi-node systems

HP StoreOnce 6500 multi-node systems have a filesystem common to both nodes in a system (see figure 5). The RAID controllers have dual SAS ports and their dual connections to each disk storage unit ["just a bunch of disks" (JBOD)] via 6 Gb/s SAS. This is necessary for high availability. The RAID controllers in this case both maintain data describing which blocks in the cache are flagged as "dirty" (this means they have been modified but not actually written to physical disk). It is necessary to maintain this "mirror" of the write back cache for consistency following power failure or when one node's hardware fails. This is a unique HP feature where the controllers operate on a master-slave mode for high availability, removing single points of failure, and enabling data integrity in the rare event of total power loss. StoreOnce 6500 systems have dual power systems that, in modern data center use, can be connected to independent power sources.

Figure 5. Dual RAID controllers in HP StoreOnce 6500 system



StoreOnce replication

HP StoreOnce systems may be configured to replicate backup via LAN or WAN to other StoreOnce systems. In this case, two or more (in the case of StoreOnce Catalyst) have a network connection and, using standard TCP/IP protocols, data is sent between systems. StoreOnce replication uses the fact that with deduplicated data stores only new data "chunks" need to be sent over the WAN or LAN after the initial synchronization. Often called bandwidth optimization, it saves valuable bandwidth when replicating data. HP StoreOnce extensively checksums each data frame with an MD5 checksum. If an error is detected, the frame is re-transmitted. Over WAN links this is especially important as the error checking is less thorough than that used on local area networks.

Additional StoreOnce architectural advantages

The HP multi-node technology available in the StoreOnce 6500 offers additional failure protection with increased redundancy and "autonomic failover."

HP ProLiant servers used throughout the range have sophisticated monitoring systems that are used by HP StoreOnce technology to report problems before they become critical.

Conclusion

Using a combination of methods, including proven HP hardware platforms and intensive lab testing, HP StoreOnce systems can give customers the confidence that their data can be fully protected throughout its lifecycle. Additional checking methods help ensure that corrupt data is seldom restored to host servers. HP StoreOnce offers many unique features such as enterprise-class HP-designed RAID controllers and HP Smart Memory, to keep data safe and secure.

We can see that HP StoreOnce technology uses a multi-layered approach to data integrity—checking the data constantly from ingest, disk storage, and restore. This is essential for both detection and correction of data errors. It is also essential that the system is designed to recover properly from power fail. Using the latest server and RAID technology, together with sophisticated HP StoreOnce deduplication, your data could not be in safer hands.

Glossary

SHA-1 hash function
Hashing is a cryptographic function that produces a unique output of 20 bytes for any string of data. It is called a one-way function as data cannot be reconstructed from the original hash code. It produces a unique digital signature for sets of data and is extremely efficient, thus well suited to the HP StoreOnce appliance. The chance of a hash "collision" is extremely rare (this is when the same hash code is produced from different data). The function can accommodate inputs of up to 264–1 bits, hence is extremely resistant to "collisions." By checking the original hash code stored with the data with the hash code computed on the data when read back, the integrity is fully verified.

MD5 hash function
This is similar to the SHA-1 hash but produces a 128-bit output and is less collision resistant. However, it is faster and given the low data volume it will be protecting in StoreOnce (metadata and indices) the extra space used by SHA-1 is not required.

Hash "collision"
This is defined as what happens if two different data inputs produce the same hash code. Technically, this is mathematically possible but extremely unlikely. However, to put things in proportion, a hash collision is much less likely than a multiple lottery win. For example, a typical national lottery requires the prediction of six numbers from 50 with the odds of winning at 14 million to one. A SHA-1 hash collision has odds of one in 263 (nine million million million to one).

CRC
CRC stands for cyclical redundancy check. It is a fixed byte output similar to the hash function. However, CRCs are reversible—this means CRCs can be used to rebuild data. CRC checking, however, is less efficient in terms of performance, hence the use of SHA-1 or MD5 hash codes. In the case of StoreOnce VT use, the CRC is re-computed from each data block read, and this is compared with the original CRC computed before the data is written.

Supercapacitor
Supercapacitors are used for high density energy storage and to bridge the gap between batteries and capacitors. They have large values (up to 2000F) and can be charged or discharged more times than traditional batteries. They are ideal for maintaining power for short periods. Used in the HP RAID controllers, they keep the cache memory RAM intact long enough to copy to non-volatile flash memory.

StoreOnce Catalyst

HP StoreOnce Catalyst brings the HP StoreOnce vision of a single, integrated enterprise-wide deduplication algorithm a step closer. It allows the seamless movement of deduplicated data across the enterprise to other HP StoreOnce Catalyst systems without rehydration. HP StoreOnce Catalyst uses a standard network connection to a backup server, but has the ability to communicate with a "client" application programming interface (API) integral to backup applications. StoreOnce Catalyst can work with Symantec's OST plug-ins for Backup Exec and NetBackup and of course HP Data Protector. Catalyst has the unique ability to perform some of the deduplication workload on the backup server making communication more bandwidth efficient and faster. StoreOnce Catalyst can also enable Oracle users to back up directly to a Catalyst store increasing deduplication efficiency. Because Catalyst understands commands from the backup software, it can move data without rehydration to other StoreOnce systems and also remove expired data automatically. Of course HP did not leave out data integrity—comprehensive error checking enables Catalyst to operate over WAN and LAN connections, even for international links. Now, data can really be moved offshore with no human intervention.

TTTD

This HP patented technology is used to divide the ingested data stream into "chunks" of an average size of 4 KB. TTTD stands for "Two Thresholds Two Divisors" and makes intelligent decisions on where to set the data "chunk" boundaries. This is very important for good deduplication efficiencies because only some parts of a backup set will change between subsequent backups. TTTD results in finer grain chunks for better performance.

Learn more at

hp.com/go/storeonce

Sign up for updates

hp.com/go/getupdated

Share with colleagues

Rate this document

4AA5-1935ENW, April 2014