Designing Primary Storage to Ease the Backup Burden

Prepared by: George Crump, Lead Analyst
February 2015

When IT planners map out their primary storage architectures they typically focus on how well the system will perform, how far it will scale and how reliable it will be. Data protection, that process that guards against corruption or system failure in primary storage or even a site disaster, is too often a secondary consideration, and often made by someone else. But what if the primary storage system could be designed to protect itself from these occurrences? Would that make it possible to simplify or even eliminate the data protection process altogether?

Ignoring costs for a moment, the technology required for a primary storage system to protect itself from the corruption or infrastructure failure is available from several vendors today. And thanks to initiatives like software-defined storage (SDS), we can finally address the cost issues as well.

A Protected Primary Storage Architecture

A design like this would leverage the use of snapshot technologies to rapidly provide point-in-time copies of data, addressing the data corruption and accidental deletion problems. Increasingly, modern storage software is able to create snapshots instantly and to keep hundreds or thousands of snapshot iterations.

Server protection is also straightforward if that system is on a shared storage network because the data is accessible from multiple servers. Certainly, virtualization via VMware and/or Hyper-V helps here. In the case of a VM failure a new virtual machine can be quickly created, and in the event of a virtual host failure, all the VMs can be evacuated to an alternate host in the virtual cluster.

However, primary storage systems that can eliminate or at least lighten the load on data protection process typically fall short if there's a storage system failure. They do have protection from a drive failure via mirroring or some form of parity based protection like RAID or Erasure Coding, but these technologies can be ineffective when the storage system sustains a multiple-drive failure. Unfortunately, storage system failure is one of the scenarios in which the data protection process is counted on the most.

However, drive failures are not the only things that can bring a storage system down; network connectivity (Fibre Channel or Ethernet adaptors) can fail, controller boards can fail, and power supplies can fail as well. For this reason most primary storage manufacturers design in redundancy for these components, but the potential for failure still exists.

While the chance of these situations actually occurring is slight the impact is severe enough that they need to be protected against. Most times this is the justification for a separate data protection process and the reason, other than cost, that most data centers don't count on primary storage to do the job.

But again, setting cost aside for the moment, if a second storage system is bought and used in tandem with the primary storage system the data center is protected from the impact of these potential failures. Using a snapshot replication process the primary storage system could continually replicate data to secondary system.

Instead of running separate snapshot and replication tracking processes, snapshot replication merely copies the changed data between snapshots to the secondary storage system. The downside is that it's not continuous, meaning there is a gap between the primary and secondary copy. For data centers that have a narrower recovery point objective applications could be designed to write to both systems simultaneously. In either case, on the second system a separate set of snapshots would be taken and kept. This provides point-in-time granularity on the second system in case the first system fails.

The Benefits of a Second Local System

The benefits of a second local primary storage system that's synchronized with the first are numerous and go well beyond the above worst-case situations. For example one of the challenges that many storage systems have is that performance drops while its parity-based data protection scheme goes through the rebuild process after a drive failure. With a second system the critical workloads could easily and safely be supported while this rebuild occurs, ensuring users' performance levels are maintained. This also protects against subsequent drive failures, something that's more likely to occur under the stress of the current drive rebuild process.

The second system wouldn't need to sit idle either, half the workload could be loaded on the secondary system and then cross-replicated to the original system. In effect both systems would be acting as primary and secondary and the workload would be distributed for better performance when both systems were in a known good state.

Alternatively, the secondary system could be used just for redundancy and its data could be snapshotted, mounted and used for test-development or potentially analytics processing. If the IT planner still wanted a dedicated data protection process, the second system could be the one that's backed up, freeing the original system from the complexity of backup agents and allowing it to focus on providing high performance to the applications it supports.

DR - The Final Leg of the Stool

The final piece to consider is protection against a site disaster where, for one reason or another, the primary data center is inaccessible. To some extent, the same process used to keep the second on-site system up to date, snapshot replication, already addresses the DR requirement. Since most DR sites are located a considerable distance from the primary data center this replication will certainly be asynchronous, which means more latency between the primary and DR copy than exists between the primary and a local secondary storage system described above. Efficient use of bandwidth and potentially the use of WAN acceleration tools should be key considerations.

The Problems with Primary Storage Replacing Data Protection

There are significant challenges to this design, otherwise it would be the de facto standard in the data center. First, of course, is price. The cost of three primary storage systems (two in the main data center and a third off-site for DR) could be substantial. Second, is the challenge of system management. For simplicity, and in many cases for the replication/snapshot strategy to work, all storage systems must be from the same vendor, even though the secondary systems are often lower-cost models. Finally, there is the challenge of snapshot management. While modern storage systems can support many more snapshots than their predecessors could, few provide adequate tools to wade through those snapshots and find the exact piece of data that's needed.

The Software-Defined Storage Solution

SDS has the potential to resolve all of these challenges and enable a data protection strategy where primary storage assumes most of the responsibility for its own protection. First, implementing the above design may be better handled by an SDS solution that is software-only, meaning it doesn't come bundled with a vendor's storage hardware. This flexibility allows the IT planner to drive down cost by leveraging more commodity storage solutions and do so safely, since there are three levels of redundancy in the design.

A key to the primary storage protection design is that the first storage system can manage all the various types of I/O that may be thrown at it. A software-only SDS solution allows the user to add components based on performance or capacity needs, but manage them from a single interface. More importantly, the snapshot and replication technology that is so important to this strategy is consistent with a software-only SDS solution that allows for a mixture of hardware to be used but the all important snapshot and replication software to be the same. This keeps acquisition and operational costs low. Finally, software-only SDS solutions seem to be better equipped to support the newer, advanced data services that are essential to the job of primary storage self-protection, like improved snapshot queries and cataloging.

The Role of Backup

What is the role of backup if the data center moves to a design where primary storage protects itself? First, they might actually convert the entire backup process to an archive only mode, where data is moved to very high capacity, “tertiary” storage like large, object based disk archives or tape. Many of these systems now present their stores as file systems to which data can be copied over using application-specific tools, or even simple copy commands.

However, most data centers would continue to use backup, for some of their critical datasets. There is a level of confidence in knowing that the most important data is on an entirely different storage medium and that it got there by a dedicated process. The key difference though, is that the pressure placed on the backup process is greatly reduced and the importance of 100% success 100% of the time is eliminated.

Conclusion

The underlying technology (snapshots and replication) that allows the IT planner to design a primary storage system that protects itself has been around for over a decade. The compelling event that makes its use practical is software-defined storage because of its promise to lower hardware costs, ease storage management and incorporate common software between different hardware platforms. This combined with a more universal need, regardless of business size, for very narrow recovery time objectives and recovery point objectives makes this approach practical for today’s data center.