Identifying the Hidden Risk of Data De-duplication:
How the HYDRAstor™ Solution Proactively Solves the Problem

October, 2006

Introduction

Data de-duplication has recently gained significant industry attention, especially regarding disk-based backup products. Data de-duplication technology (also known as commonality factoring, non-redundant storage and duplicate data reduction) identifies and stores only new, unique data. In other words, if a string of data is already stored in the system, then it is referenced by pointer rather than stored a second time.

Indeed, the benefits of de-duplication are compelling. Backup systems, by their nature, backup the same data over and over. By implementing de-duplication technology, IT organizations benefit from a lower acquisition cost, lower total cost of ownership (TCO) and the ability to cost effectively replicate data for disaster recovery purposes even in low bandwidth environments.

While first-generation disk-based backup products, such as VTLs, can improve backup operations, their high acquisition costs and TCO has limited their adoption. The inescapable reality is that these products require IT organizations to purchase a significant amount of additional expensive disk capacity. Organizations wanting to store a month’s worth of backups on a VTL need to purchase a VTL with four to five times the disk capacity of their production environment. (This ratio provides for a “grandfather-father-son” strategy with a one-month retention cycle; IT organizations that want to retain multiple months’ worth of backups on disk would require even larger amounts of disk capacity.) Reducing disk usage is where data reduction (de-duplication) technology can change the game: all the benefits of disk-based backup can be achieved with significantly less new disk capacity and related costs.

Data De-duplication Technologies

The nature of backup operations is such that data is backed up over and over, even though substantial portions of it may not have changed since the last backup. De-duplication algorithms identify this redundant data and store it only once. After the initial storage of a data string, all future backups of that string simply refer to the original string stored on disk by using a data pointer.
Over time, de-duplication technology can achieve a 20:1 data reduction, which means that only 5% of the incoming backup data is actually stored on disk. For example, with de-duplication, 100 TBs of backup images can be stored on only five TBs of physical disk. Actual reduction ratios will vary depending upon the nature of the data and the organization’s backup strategy. Organizations that perform daily full backups will achieve greater reduction ratios faster than those employing incremental backups.

The business value of data reduction technology includes the following:

- **Lower Cost of Acquisition** – Raw disk space requirement can be reduced by 80%. Products incorporating data reduction technologies can have a significantly lower price on a per-TB basis.

- **Lower Total Cost of Ownership (TCO)** – As the data reduction factor increases over time, physical disk space used will not grow as rapidly as those products that do not use data reduction technology.

- **Cost-effective Remote Replication** – Remote data replication is unrealistic in many situations using standard technology due to data volumes over constrained bandwidth. With data reduction, only unique de-duplicated data is transferred resulting in a 90-95% reduction in transferred data making replication over a WAN realistic.

While the benefits of de-duplication are clear, the “bolt-on” implementations found in first-generation products introduce significant risks. De-duplication is a CPU intensive process, as is software-based compression; the latter can reduce performance by 40%. With first-generation de-duplication products, throughput can be reduced by as much as 65%. Such a penalty largely negates the advantage of disk speeds over tape.

However, the most critical risk associated with first-generation products is the risk of lost backups and unavailable recovery. With duplicate data reduction ratios reaching 20:1 or higher, dozens to hundreds of backup images are dependent upon a single disk block for recovery. If that disk block is lost, all of the associated backup images could be rendered unrecoverable. The impact of a failed disk would be even worse and could impact the availability of all backups stored on the VTL or disk-based appliance.

Moreover, high capacity (S)ATA disk drives can take 24 hours or more to rebuild, and the rebuild process is I/O-intensive, affecting every disk in the RAID group. The substantially degraded throughput during the rebuilding process not only jeopardizes the RTO of large-scale data restoration efforts but can also lead to missed backup windows.
Does RAID provide enough protection?

Most first-generation VTLs and disk-based appliances use RAID for disk failure protection. RAID 5 uses single parity and can recover from a single disk failure with an associated storage overhead of about 20%. Nevertheless, RAID 5 has significant shortcomings. In the event of a two-disk failure within the same RAID group, the data within that group are unrecoverable. To improve data resiliency, some disk-based backup products use RAID 6 instead of RAID 5. RAID 6 uses double parity for fault protection and can recover from two drive failures in the same RAID group as a result. However, the storage overhead for RAID 6 is 35-40%, a significant increase over RAID 5.

Although RAID may provide sufficient protection for non-de-duplicated environments, it is insufficient for de-duplicated environments. In de-duplicated environments, a single disk block may contain data used by hundreds of backup images. If that block is lost, then all of the images are rendered unrecoverable. Moreover, the time to rebuild a 300 GB SATA drive in a typical storage array can be 24 hours; 500 GB drives even longer. As drive capacities increase toward a terabyte, the time to rebuild such a drive creates a statistical probability that a second or third failure will occur within the same RAID group before the rebuild completes.

In addition, RAID technology suffers severe performance degradation during a disk rebuild. Rebuilding a RAID 5/6 group is very I/O intensive, affecting every disk in the group. The substantially degraded performance during the rebuild time can jeopardize the organization’s RTO in the event of a simultaneous restore requirement. Lengthy re-build times and performance loss can also cause missed backup windows, which puts even more data at risk. Many arrays with RAID 6 do not recommend reading or writing to the RAID group during a double parity recovery because the recursive recovery process of RAID 6 will result in significant performance degradation. Interrupting the process could result in unrecoverable data.

As with RAID 5, RAID 6 systems suffer a write penalty due to the additional parity calculations. This penalty can decrease the overall performance of the VTL or disk appliance. The penalty is worse than just two parity calculations, because the double parity calculation does not easily map to standard CPU hardware and requires a table lookup.¹

The ideal solution would be one that can survive more than two disk failures, has no write performance penalty, does not experience degraded performance during a disk rebuild and uses no more disk space than RAID 5. NEC has spent the last three years designing just such a solution, and the result is HYDRAstor.

¹ H. Peter Arvin: “The mathematics of RAID 6.”
HYDRAstor: A Next-generation Disk-based Backup and Restore Solution

HYDRAstor is a next-generation disk-based data protection appliance designed from the ground up to address key backup and restore challenges without the risks and operational limitations of current VTLs and disk-as-disk appliances. HYDRAstor’s unique architecture enables high performance backup and restore scaling from 200 MB/sec to well over 20,000 MB/sec of aggregated throughput in a single instance of the solution.

HYDRAstor’s disk capacity can also be scaled easily and non-disruptively from TBs to PBs. A single system can store months or years worth of backup data for less than $0.90/GB, which is less than the cost of a typical tape system. The scalability and affordability of the system is due to HYDRAstor’s unique grid-base storage architecture and its patent-pending DataRedux™ and Distributed Resilient Data™ (DRD) technology. Unlike first-generation VTLs and disk appliances, HYDRAstor is the only solution that fully addresses IT’s requirement for increased data resiliency in a de-duplicated environment.

Grid-based Storage Architecture

HYDRAstor’s grid-based storage architecture differentiates it from all other disk-based backup systems. The architecture combines HYDRAstor’s intelligent OS with best-of-breed industry standard servers and disk to deliver unmatched data resiliency and virtually unlimited performance and capacity scalability. Unlike first-generation VTLs and disk-as-disk backup appliances that put de-duplicated data at risk, HYDRAstor introduces advanced data resiliency through patent-pending DRD technology.

Delivered as a turnkey solution, HYDRAstor’s innovative grid-based architecture eliminates the operational limitations of current products by allowing users to apply additional resources wherever they are needed. Moreover, these resources can be added without labor-intensive configuration, tuning or intervention as HYDRAstor is a fully self-managed, self-tuning, and self-healing system. Performance scalability and storage capacity may be increased independently based on business needs, and the addition of one component improves performance or capacity across the rest of the system. HYDRAstor accomplishes this scalability through the implementation of what is termed Accelerator Nodes and Storage Nodes.
Accelerator Nodes — Scaling Performance

HYDRAstor Accelerator Nodes (ANs) as pictured in Figure 1, are industry standard off-the-shelf servers. ANs connect to one or more backup servers over a gigabit Ethernet (GbE) network and support both CIFS and NFS. They are clustered for scalability and resiliency, and each AN can deliver more than 100 MB/sec throughput. ANs operate with a coherent distributed file system that not only spreads the workload across available processors, but also assures that no node is a single point of failure. If one or more ANs fail, the remaining ANs can assume the aggregate workload. ANs can be scaled non-disruptively to increase performance. For example, 10 ANs would provide over 1000 MB/sec of aggregated backup and restore throughput. To determine the number of ANs needed in a HYDRAstor environment, the following calculation would be used:

\[
AN\# = \frac{\text{Aggregate peak data transfer in MB/sec}}{100 \text{ MB/sec}}
\]

Storage Nodes — Scaling Capacity

Storage Nodes (SNs) as pictured in Figure 2, are also built on industry standard servers each with three TBs of storage capacity. SNs provide disk capacity for the backup images and are connected to the Accelerator Nodes via a private network managed by the HYDRAstor OS. The HYDRAstor OS virtualizes storage within and across multiple SNs to create one logical pool of capacity accessible by any AN. HYDRAstor’s innovative capacity virtualization also eliminates all provisioning tasks. With HYDRAstor, storage administrators do not need to create and size LUNs, volumes, or file systems. As SNs are added, HYDRAstor automatically load-balances existing data across all available capacity to optimize performance and utilization. HYDRAstor’s distributed, grid architecture provides these benefits without impacting the backup or restore performance.
DataRedux Technology — Storage Efficiency

HYDRAstor’s DataRedux technology, which provides data de-duplication, begins with NEC’s patent-pending “chunking” technology whereby data streams are separated into variable length “data chunks.” Hashing is then used to determine if the data chunk has or has not been stored previously. If the chunk has been previously stored, pointers are used to address the data. If the data has not been previously stored, data resiliency described later will be applied before the chunk is stored onto disk.

De-duplication efficiency is driven by the data-chunking algorithm. Products that use no chunking at all or split the data into chunks of fixed size typically experience a lower data reduction ratio. In the first case, changing one byte in a file would result in the whole file being stored again. In the second case, any chunk that follows the chunk containing the changed byte would also be stored. Instead, HYDRAstor uses both a sophisticated algorithm and variable size chunks to maximize its opportunities to find redundant data. Only the chunks that contain real changes to the data are identified as unique. These chunks then undergo physical compression before being stored to disk.

While duplicate elimination and physical compression create a performance bottleneck in first-generation array-based VTL products, HYDRAstor’s grid-based architecture enables it to distribute the load across many nodes thus avoiding performance degradation and idle nodes.

Distributed Resilient Data Technology — Recovery Assurance

To address RAID 5 and 6 limitations and their risks when used with data reduction technology, HYDRAstor introduces patent-pending Distributed Resilient Data (DRD) technology. DRD protects data from three or more disk failures without the RAID write penalty, without incurring performance degradation during a re-build, without long rebuild times, and with less storage overhead than RAID 6.

After a unique data chunk is identified by HYDRAstor’s data reduction technology, DRD breaks the chunk into a set number of data fragments. The default number of these original data fragments is nine. Depending on the desired resiliency level (the default is three), DRD then computes the same number of parity fragments to match the resiliency level, with all parity fragments based only on the original nine data fragments. (In other words, DRD computes parity fragments based solely on the data, not on parity.) Hence, unlike first-generation VTLs and appliances using RAID which must read parity, recalculate parity, and then rewrite parity, HYDRAstor does not incur this substantial “parity penalty.” DRD then intelligently distributes the 12 fragments (nine original fragments + three parity fragments) across as many SNs as possible in a configuration. If the system has less than 12 SNs, multiple fragments will be stored on the same SN. In this case, HYDRAstor will distribute these fragments on different disks. In the minimum recommended configuration of four SNs, three fragments would be written to different disks on each SN to maximize resilience.
With the default resiliency setting of three, a chunk of data can be re-created using any nine out of the total of 12 fragments. That means the system could simultaneously lose as many as three fragments from the group without jeopardizing data integrity. In a minimum HYDRAstor configuration of four SNs, this would protect all data in the event of one SN failure (five data disks) or the failure of any three disk drives across multiple SNs.

In larger HYDRAstor systems with 12 or more SNs, each SN stores no more than one of the 12 fragments of a chunk. Thus, the default resiliency can protect data from three or more SN failures. Administrators can optionally choose a higher level of resilience if data stored is deemed critical. Based on the resiliency setting, HYDRAstor will automatically calculate the appropriate number of original data fragments and create the necessary number of parity fragments. The fragments will always be distributed to achieve maximum data resiliency.

Because of this additional resilience, one might believe that HYDRAstor requires substantial storage overhead, but this is not the case. The disk space overhead for HYDRAstor’s default architecture is 3/12 or 25%, just slightly more than RAID 5 overhead (about 20%) and less than RAID 6 (around 35-40%), while providing substantially higher resilience than either: 300% more resiliency than RAID 5 and 50% more than RAID 6.

Unlike RAID 5 and 6, HYDRAstor is accessible for backup and restore operations during a disk rebuild without degraded performance. Because the fragments are distributed across multiple Storage Nodes that each have ample processing power, backup and restore operations will not experience performance degradation in the event of a rebuild. Failed components are automatically “discovered,” and the re-build is automatically initiated in the background. (see Figure 4).
No Single Point of Failure

The fixed hash tables of first-generation products typically have multiple points of failure. As they are based on one disk array, failure of any single component such as the motherboard, a RAID card, etc., will result in the failure of the whole system. The equivalent within HYDRAstor would be the failure of one AN or one SN. As discussed above, this has no impact on the accessibility of data in HYDRAstor (see Figure 4).

First-generation products typically have multiple points of failure. The fixed hash tables of first-generation products can also represent a single point of failure. If the hash tables are lost, then the data become unrecoverable. HYDRAstor has no such centralized tables; it distributes its hash tables across all Storage Nodes in such a way that even failure of multiple SNs would not result in the loss of hashing information.

Figure 4: Data Resiliency in action. In the event of catastrophic failure of a storage node, HYDRAstor immediately discovers the failure, rebuilds, and writes the data to other available storage nodes.
HYDRAstor Benefits

The unique benefits of HYDRAstor may be summarized as follows:

• **Affordable** – HYDRAstor’s safer data reduction capabilities enable disk-based backup at a price point comparable to tape.

• **Enhanced Resilience** – With patent-pending Distributed Resilient Data (DRD) technology, organizations can safely take full advantage of HYDRAstor de-duplication without the worries of data loss and unavailable recoveries. DRD storage overhead is comparable to RAID 5 but provides 300% better data resiliency.

• **Dynamic Scalability** – Accelerator Nodes (ANs) and Storage Nodes (SNs) may be added independently and non-disruptively to scale performance or capacity to meet the organization’s needs. No other product provides this simple, cost-effective infrastructure tuning ability.

• **No Performance Penalty** – Unlike RAID systems, HYDRAstor does not experience performance degradations during disk re-build operations, which can put your backup and restore operations at risk.

• **No single point of failure** – HYDRAstor is delivered as a fully redundant turnkey appliance, and it has no centralized resources that represent a single point of failure.

HYDRAstor offers the lower TCO benefits of data reduction plus increased system availability not found in first-generation VTLs and disk-as-disk appliances. Enterprise data centers will appreciate the superior capabilities of HYDRAstor when it comes to protecting their organization’s critical data.