

Diligent HyperFactor: Redefining De-Duplication

August 2006



In the past 24 months, we have seen disk-based data protection and archiving become an essential component of most enterprises. With massive 100 TB+ disk-based repositories becoming increasingly common on near-term enterprise roadmaps, finding ways to optimize those disk capacities is now a top of mind concern. Even with widely available cost-effective disk, our research finds that capacities still account for over 50% of most data storage budgets, with disk-based secondary storage very commonly exceeding 50% year-over-year growth rates. Aggressively managing these capacities to keep them “lean and mean” is absolutely critical. The good news is that new technologies exist that can radically slim down the physical storage requirements for secondary disk storage. Taneja Group refers to these technologies as Capacity Optimization (CO) technologies. Every enterprise now needs a CO strategy, but beware, for not all solutions are created equal. In our opinion, one company on the very cutting edge of this CO wave is Diligent Technologies. Specifically, we call attention to Diligent’s HyperFactor technology. It stands out as extremely scalable, efficient, high-performance data de-duplication software that merits a deeper understanding. As the “secret sauce” in Diligent’s ProtecTier VTL solution, HyperFactor has achieved efficiencies that represent a true leap over 1st generation optimization approaches. Beyond the mind-bending 25:1 reduction ratio, the efficiency of this technology’s indexing architecture (and therefore, its performance) is second to none. In this technology brief, we will examine what IT teams should be looking for in a capacity optimization solution, and then share our perspective on the Diligent HyperFactor technology. This is a core new technology and we’re very excited by what it can mean for storage ROI, data protection performance, and reliability.

Why Optimize Storage Capacity?

Somewhat ironically, the latest growth boom in the storage industry is all about *reducing capacities*. Specifically, this innovation boom is about software technologies that promise to radically reduce the amount of physical storage capacity required to store a given amount of information. In fact, some of the most exciting companies in storage today are laser-focused on tackling precisely this challenge. Taneja Group has categorized all of these offerings under the umbrella of Capacity Optimization (CO) technologies.

This entire CO category has gone from marginal to essential in the past 24 months. Why the intense interest in optimizing storage capacities? Two words: *economics and manageability*. IT teams need to reduce the amount they are spending to store a given terabyte of information, *and* they need to find ways to manage the insanely high growth rates common across the enterprise today. Note that the biggest culprits of capacity consumption have been data protection and archiving environments (collectively, “secondary storage”). For that reason, significant vendor R&D attention has

T E C H N O L O G Y B R I E F

been focused on finding ways to optimize those capacities in order to improve storage ROI and ease of management for large repositories.

Indeed, these CO technologies have already become very powerful. As we will explore below, the most advanced offerings can achieve reduction ratios of 25:1 or more on disk-based storage platforms, scaling into the petabyte range for single solutions, all with no impact to the pre-existing data protection workflow, or additional management requirements. With capabilities like this available, the question rightfully becomes: Why *wouldn't* you optimize your secondary storage capacities?

The Rise of De-Duplication

Behind all advanced capacity optimization initiatives is some manner of de-duplication technology. At the highest level, all de-duplication technologies provide an intelligent means of organizing data being stored such that redundant data elements need not be stored more than once. We call this process “factoring”. To achieve efficient data factoring, some manner of indexing capability will maintain a running tally of all data being stored in a repository. The index is akin to a “table of contents” of the entire data repository, keeping track of where unique data elements reside.

The existence of this index, or “table of contents”, then frees up the repository to only store single instances of each data element. Different de-duplication approaches will support varying levels of granularity for these data elements. In general, the more

fine-grained the granularity, the more efficient the entire repository can become. All of this requires an ongoing, dynamic dialog between the repository and the index, as every new piece of data created immediately changes what will need to be stored in the future.

From this high-level description, it should already be obvious that the power of any given de-duplication technology rests squarely on two things: the *efficiency* of its indexing, and the *granularity* of its factoring. We will now explore these issues in more depth, taking a deeper look at the requirements of de-duplication technologies.

What Matters in De-Duplication

For IT teams that are moving into evaluation mode on their CO strategy for secondary storage, we believe it is important to establish one set of “apples to apples” requirements for comparing the various technologies they may review. To that end, we have identified the following as critical data de-duplication technology requirements for the enterprise data center:

Requirement: Efficient Indexing

One of the top requirements for any de-duplication technology is efficient index architecture. In fact, many of the other requirements listed below are directly or indirectly related to the index architecture. As discussed above, the index is at the heart of how effective the de-duplication technology can be both today and over time. This efficiency can be measured in several ways. First, the index should be as lightweight as possible as a percentage of the

T E C H N O L O G Y B R I E F

repository itself. Generally speaking, the less capacity required by an index, the higher its overall performance will be, especially at higher scales. Second, the index should be able to support very fine-grained factoring of data elements across *all* data types in the entire repository. The best solutions will have very few or zero limitations in block size sensitivities or file types.

Requirement: Scalable Support

We are regularly told by end users that they end up deploying more capacity on their D2D and disk archives than they originally anticipated. This trend tends to continue throughout the long-term deployment lifecycle of the repository, often equating to 3x to 7x the primary storage capacity. Because of such high-scale requirements in D2D backup and archive, the de-duplication approach itself must demonstrably support scalability into repository sizes significantly higher than the IT team's present environment. This means that the factoring approach of the de-duplication technique must be capable of supporting very high scale environments without performance degradation. Customers should demand that vendors provide proof points of true single repository scalability for any de-duplication technology. Vendor claims of a great reduction ratio are meaningless if it will only support your next 12 to 24 months of growth!

Requirement: High Performance

When looking at any D2D solution that leverages CO technology, performance needs to be keenly assessed. Sometimes, it can be difficult to determine how a particular D2D technology's de-duplication approach will impact the performance of backup activities

or archiving. Central to the overall performance will be the index architecture. Since most de-duplication approaches require IO hits to the index residing on disk, they can suffer performance degradation, particularly at higher scale. Again, ask the vendor to provide proof points for performance support claims.

Requirement: Open Data Support

A CO strategy in D2D needs to support as wide a swath of data types as possible, in order to make the overall investment and management value of the solution worthwhile. Accordingly, we recommend that any de-duplication technique pursued in the context of D2D backup or archiving be open to as many data types as possible. Prospective customers should query their vendors to understand if there are any application or file type limitations inherent in their particular approach. Does this then require further integration? If so, this may have a material impact on the overall reduction ratios that can be achieved, again, depending on workload.

Requirement: Non-Disruption

While the benefits of deploying a D2D backup or archive solution are now obvious, finding ways to integrate it *seamlessly* with a pre-existing data protection infrastructure can be a challenge. This need for "non-disruption" applies to both the deployment integration and ongoing management of the solution. Specifically, with regards to the de-duplication technology, "non-disruption" means that the CO approach should not degrade the pre-existing data protection operations at all. Ideally, the IT team should not even know that data de-duplication is

T E C H N O L O G Y B R I E F

happening, either from a methodological or performance impact perspective.

Bearing these requirements in mind, we will now turn our attention to a de-duplication technology that we believe satisfies these criteria, and therefore merits deeper consideration by the end user community: Diligent's HyperFactor software.

Meet Diligent HyperFactor

This brief focuses on Diligent Technologies' HyperFactor de-duplication technology because we believe it represents the state of the art in the category. Quite frankly, we believe this offering encapsulates where CO will be heading in the coming years. As we will explore below, there are many differentiators at play in HyperFactor, all of which add up to a markedly unique de-duplication approach.

HyperFactor is currently seeing duty at the core of Diligent Technologies ProtecTier Data Protection platform. In this role, it enables enterprises to transform their traditional schedule-based backup environment, by greatly reducing the amount of data it takes to store backups on an open systems disk-based platform.

Typically, HyperFactor can reduce capacities anywhere from 10:1 to 25:1 or more, depending on workload type and retention periods. When this optimization efficiency is coupled with its scalability, it translates into some boggling figures: HyperFactor can scale to support 25 petabytes of managed data, which it would optimize down to a 1 petabyte physical storage repository, all managed

under a single index. How is this possible? Because of Diligent's R&D break-throughs in indexing design. HyperFactor's index can map an entire data repository at a 250,000:1 ratio, which is an unprecedented level of efficiency.

It should be understood that HyperFactor is a 100% software-based technology. HyperFactor can be easily deployed on any standard Linux server; there is no proprietary hardware schema or appliance investment required. Perhaps most critical of all, the entire HyperFactor index can reside on this same Linux server. This means that the entire index for HyperFactor will always be running from the RAM in the host server. From this key fact alone, an entire range of benefits are possible, as we will see.

Note that because of this software-based architectural freedom and compact deployment in server + RAM, HyperFactor is highly extensible. Over time, we believe it will likely begin to find its way into a range of applications beyond the ProtecTier offering itself. However, for the purposes of this technology brief and our current exploration of HyperFactor, we will take its role enabling Diligent's ProtecTier offering as our focus.

The following four steps represent the interaction of HyperFactor with a typical ProtecTier VTL data protection workflow as it comes into a disk-based data repository.

Step 1: Index Scanning

HyperFactor deploys directly behind the ProtecTier VTL interface. True to the virtual tape usage model, ProtecTier itself is viewed as a backup target by most leading tape-

T E C H N O L O G Y B R I E F

based backup applications (e.g. ProtecTIER is certified with Symantec NetBackup, IBM Tivoli Storage Manager, EMC Legato Networker, and several others.) As backup data streams into the ProtecTier interface, the HyperFactor search engine scans the data. HyperFactor's algorithms are unique here in that it is initially looking for *similarities* to data already in the repository. It is not immediately looking to identify redundant block sets, as is the case in competing de-duplication approaches. Once HyperFactor has identified candidate data that is similar to what has already been stored, it moves to the next step: comparison.

Step 2: Byte-Level Comparison

With a working pool of data element similarities held in the RAM-based index, HyperFactor then reads directly from the repository to conduct byte-level comparisons of similar stored data versus the new incoming data. This tells the index what constitutes bona-fide unique data versus what is already stored in the repository. This process ensures 100% data integrity.

Step 3: Store New Data Deltas

Once all new byte-level data deltas have been identified, just those *new* data elements that were identified in the step above are stored in the repository. This is the stage of factoring at which the actual data reduction process takes place and we begin to see the difference between information stored versus physical capacities required. Applied across terabytes of information, it is easy to see how this compounds into a significant reduction ratio.

Step 4: Update to the Index

Finally, the index itself is updated with the latest iterations of new data in the repository. The entire index is now completely up-to-date with its "Table of Contents" of all data in the repository.

And here is the dramatic clincher: Unlike other de-duplication processes, these 4 steps listed above all transpire in-line at 200 MB/S, per node, with all index interactions taking place within the RAM of the server hosting HyperFactor. The HyperFactor index itself is so capacity efficient that in even the largest 1 petabyte deployment, it will not exceed 4 GB in size. That is still small enough to be held completely in RAM for an entire 1PB repository. To the best of our knowledge, no other player in the industry can make such a claim regarding index efficiency.

Data Restores

A frequently asked question at this point is, "how is an object reconstructed during a restore process?" Diligent has a unique approach that both improves the speed of restores and greatly reduces risk. The index itself is not used during the restore process. Rather, ProtecTIER maintains metadata for each object, tracking all pointers to existing data elements within the repository. When an object is requested by the backup application during a restore (the backup app sees this as a restore from tape), ProtecTIER follows a "scatter-gather" methodology. The metadata that describes the object is invoked to collect all data elements that are referenced by the pointers. In real-time the object is rebuilt and delivered back to the requesting media server.

T E C H N O L O G Y B R I E F

The restore process is actually 10-15% faster than the backup process.

HyperFactor Differentiators

With this basic understanding of how HyperFactor works in a ProtecTier deployment, we can turn our attention to some of the key architectural differences that we believe make this technology noteworthy in the CO space.

Differentiator: Software-Based

A significant advantage enjoyed by HyperFactor is that it is a 100% software-based solution. This has implications for deployment flexibility, scalability, and integration with the existing environment. Most de-duplication technologies today marry the software functionality with a customized server+storage infrastructure. We believe that HyperFactor's software-centric design center will provide it with significant flexibility in enterprise deployments.

Differentiator: In-line Performance

HyperFactor can factor data at an amazingly fast rate because it avoids the unintelligent, random IO disk hits that plague many de-duplication approaches. Specifically, HyperFactor can deliver 200 MB/S in-line throughput, per node. This is a performance threshold that places it well above the requirements of most data protection environments. In other words, it will not become a bottleneck to the normal D2D backup or archival operations. For this reason alone, we believe larger enterprises

will be giving Diligent a hard look on the VTL front in 2006-2007.

Beyond the performance that ProtecTier enables, the fact that it performs the de-duplication function in-line will be a critical differentiator for many customers. Because of the complexity of the de-duplication operation, some vendors conduct this as a secondary process, following the backup operation. However, in some enterprise usage cases this may pose certain limitations. Specifically, some de-duplication post processes can elongate the total time until the data on the VTL is "at rest." This means that post processing might collide with other activities within the domain of the backup application (e.g. vaulting to tape), thereby creating resource contention and degraded performance within the VTL environment.

Differentiator: High Scalability

As we have reiterated at several points in the brief already: The architecture of the index *matters*. The fact that HyperFactor is (1) amazingly lean -so lean that it can reside 100% in server-based RAM for even a 1PB repository- and (2) it can provide for very fine-grained byte-level delta identification, all adds up to a superior approach to factoring data in a D2D environment. Especially for enterprises with larger repository requirements, this kind of design advantage means not having to worry about hitting a management, performance, or scaling wall as higher level capacities are crossed in the 100+ terabyte range. Knowing that the IT team can reliably build a single repository at the 1PB level is a non-trivial breakthrough that we suspect many enterprises will be eager to take advantage of.

T E C H N O L O G Y B R I E F

Differentiator: Open Data Access

One of the elegant aspects of the HyperFactor software design is that because it operates at the byte-level to identify commonalities, it is totally open to all data types and formats. As a component in the ProtecTier VTL offering, HyperFactor works equally effectively with databases, email, or file data streams. The same cannot be said for many de-duplication technologies that leverage various approaches in the hashing or content-aware space. Additionally, it is worth noting that HyperFactor's factoring capabilities can be tuned and optimized for specific workloads, thereby ensuring that reduction ratios remain high across workloads, and at scale.

Differentiator: Totally Non-Disruptive

The last key differentiator that we believe merits discussion is the aggregate level of non-disruption that HyperFactor brings to the environment. We know that end users are increasingly sensitive to any disruptions to pre-existing methodologies, and are now savvy about performance disruptions resulting from CO approaches that exceed their limits. HyperFactor's lightweight index will not impede performance; it can adjust to variable workloads, and it deploys seamlessly on industry standard hardware. This is a very powerful combination. We expect that the bar for non-disruptive deployments in the CO field will only become higher as larger capacities and more strategic assets make their way onto disk platforms in coming years. HyperFactor strikes us as well positioned to pass scrutiny on this front.

Taneja Group Opinion

In the world of capacity optimization, it is clear that simply having a solution is no longer good enough. The quality of the factoring capability, the architecture of the index, as well as its performance and scalability, all add up to the total impact value of the solution. IT teams must choose carefully now to ensure that they build on the most advantageous platform possible.

In our review of Diligent's HyperFactor software, we come away very impressed with what the company has achieved. They have managed to create a lightweight, scalable, high-performance index that delivers reduction ratios at the very cutting edge of the industry, without impacting the pre-existing environment. This is clearly built to the strictest enterprise specifications of large enterprises, yet remains flexible enough that we believe even smaller D2D repository requirements will find it valuable and cost-effective.

The ROI implications of deploying a technology like HyperFactor are significant. When companies can reduce their schedule-based backup environments and their inherently inefficient "over-copying" on a 10:1 to 25:1 scale, we are talking about real hard cost savings. Based on customers with whom we have spoken, the ROI narrative to support these kinds of investments is a very straightforward calculation, typically supportable on hard cost capacity savings alone. And when soft savings via management gains are added, we see that this kind of ROI calculation typically becomes a deal-maker. It is a clear enough

T E C H N O L O G Y B R I E F

economic case that we believe no serious IT shop can ignore the fiscal returns that come from a properly deployed CO strategy for secondary storage. Period.

We are excited by what HyperFactor means for Diligent's prospects, as well. We expect to see the company leverage this very key intellectual property to secure a range of OEM and partnership deals in coming quarters. With the inherent extensibility in the HyperFactor software code-base, this could easily find its way into a range of offerings beyond ProtectTier for VTL, taking

the entire Diligent platform into the core of disk archival, as well.

All enterprises with significant data protection requirements need to be looking seriously at how to control their capacities. Advanced de-duplication technologies like HyperFactor are now the best weapons in the arsenal. Without question, enterprises evaluating their options for capacity optimization in D2D need to understand this very critical technology from Diligent Technologies.

***NOTICE:** The information and product recommendations made by the TANEJA GROUP are based upon public information and sources and may also include personal opinions both of the TANEJA GROUP and others, all of which we believe to be accurate and reliable. However, as market conditions change and not within our control, the information and recommendations are made without warranty of any kind. All product names used and mentioned herein are the trademarks of their respective owners. The TANEJA GROUP, Inc. assumes no responsibility or liability for any damages whatsoever (including incidental, consequential or otherwise), caused by your use of, or reliance upon, the information and recommendations presented herein, nor for any inadvertent errors which may appear in this document.*