

White Paper

HP StoreOnce: The Next Wave of Data Deduplication

By Lauren Whitehouse

November, 2011

This ESG White Paper was commissioned by Hewlett-Packard and is distributed under license from ESG.

Contents

Introduction	3
The State of Deduplication	4
Data Growth Driving Demand	4
First- Versus Next-Wave Deduplication Solutions.....	5
HP StoreOnce Deduplication	7
Federated Deduplication	7
Deduplication Optimization.....	7
Scale-Out Architecture with High Availability	8
Rapid Restore Performance.....	8
Off-Site Copies	8
Cost Implications	8
The Bigger Truth	9

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from time to time. This publication is copyrighted by The Enterprise Strategy Group, Inc. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of the Enterprise Strategy Group, Inc., is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at (508) 482-0188.

Introduction

Leveraging deduplication in backup environments yields significant advantages. The cost savings in reducing disk capacity requirements change the economics of disk-based backup. For some organizations, it allows disk-based backup—and, importantly, recovery—to be extended to additional workloads in the environment. For others, deduplication makes it possible to introduce disk-based backup where it may not have been feasible before.

Deduplication in data protection is not new; however, it is being implemented in new ways. Its availability in secondary disk storage systems was the predominant delivery vehicle just a few years ago. Today, the technology is available as an integrated feature of backup software, cloud gateway, and software-as-a-service (SaaS) solutions, delivering bandwidth savings in addition to reduced storage capacity benefits. In addition to distributing deduplication processing across multiple points in the backup data path, there are many more deduplication techniques and approaches today too. Vendors are perfecting and optimizing algorithms that identify and eliminate redundancy to meet the ever-changing requirements driven by relentless data growth and IT's desire to keep pace with the volume of data under management.

The evolution of deduplication is being provoked by user requirements, as well as improvements in IT infrastructure, including larger, faster disk drives, and APIs facilitating better integration between data protection hardware and software components. IT organizations that have or plan to implement deduplication want greater flexibility in how and where deduplication is deployed, tighter integration with the backup policy engine and backup catalog, faster performance for backup and recovery, the ability to deduplicate within and across domains to gain more efficiency. And, they want it for the lowest cost possible.

What is emerging from these new requirements is a clear delineation of first- and next-wave solutions. First-wave deduplication solutions met the early-stage circumstances but may have now stalled. Next-generation deduplication solutions are evolving at the same rate as user's maturation with and application of the technology. In the latter category is HP StoreOnce technology.

HP is introducing future-ready deduplication in its StoreOnce technology. StoreOnce is a modular deduplication algorithm that is portable and interchangeable—whether deployed in HP hardware or software—that can drive greater efficiency. StoreOnce deduplication is now implemented in both HP data protection appliances and in HP Data Protector backup/recovery software. HP's portfolio of hardware solutions includes small and midrange appliances as well as the new HP B6200 StoreOnce Backup System for enterprise data centers. The B6200 is a large-scale appliance built on scale-out architecture and offering high availability. HP claims it is the fastest solution in the market, and has twice the capacity at 20% less cost than other leading solutions. HP's software StoreOnce solution enables federated deduplication – the native movement of data between homogeneous systems in its deduplicated state. On the software side, HP Data Protector deduplicates at the media server as a software target using available disk capacity. The B6200 backs up at up to 28TB/hour, while Data Protector backs up at up to 1.8 TB/hour.

With StoreOnce, HP is addressing what ESG sees as the shortcomings of “phase one,” “point solution” deduplication offerings available today: high cost, complexity, high operational overhead, rigid solution stacks, scalability limitations, single points of failure, lack of restore performance, and heterogeneous silos of storage. HP StoreOnce deduplication simplifies the deployment of deduplication technology across the IT infrastructure. It is a portable engine that can be easily embedded in multiple infrastructure components, eliminating the complexity seen in earlier-generation deduplication.

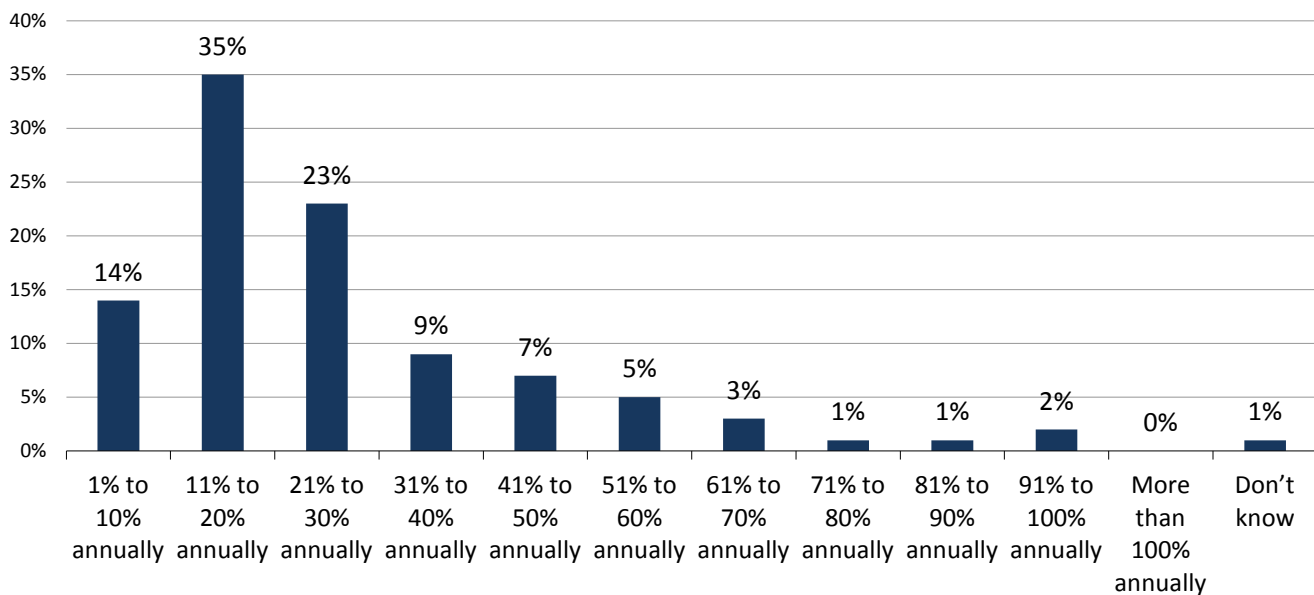
The State of Deduplication

Data Growth Driving Demand

Relentless growth in data volumes is complicit in driving demand for data reduction technologies. The data-dependent processes employed by most companies have created a deluge of data to manage and protect. As shown in Figure 1, ESG survey respondents cite annual data growth rates of various amounts. However, the majority fall into the 11% to 30% per year range.¹ With data volume doubling every few years, it's no surprise that survey respondents report "managing data growth" as a top IT priority in 2011.²

Figure 1. Annual Data Growth Rates

At approximately what rate do you believe your total volume of data is growing annually for each of the following categories? (Percent of respondents, N=510)



Source: Enterprise Strategy Group, 2010.

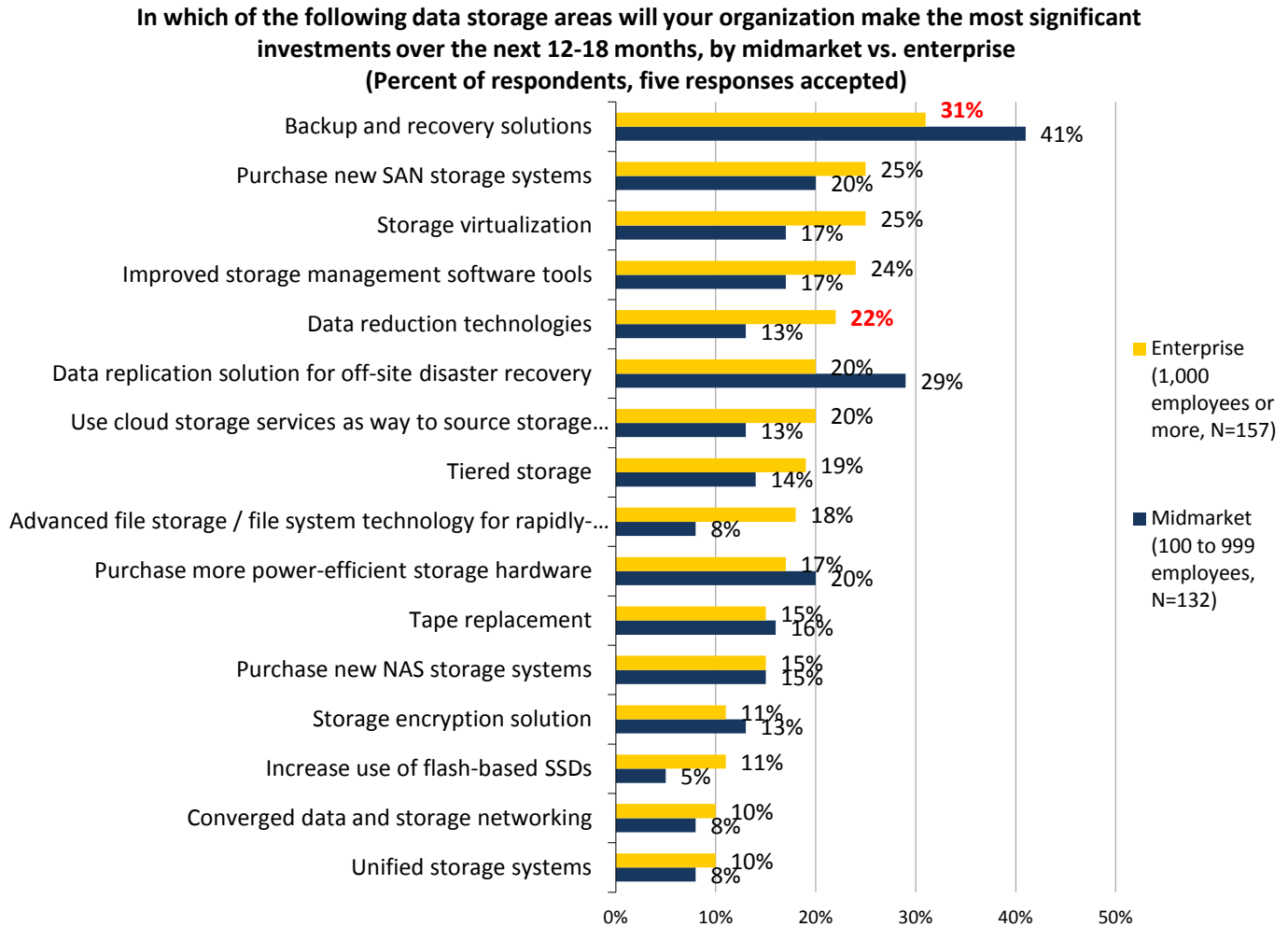
As the volume of data increases in primary storage environments, the impact on the backup storage environment is significant. This is especially true when multiple copies of production data are maintained for recovery purposes, and for extended periods of time for compliance and records retention. Addressing data growth—and its impact on backup—is a perennial pain point, as evidenced by ESG research. For 2011 data storage spending priorities, ESG survey respondents ranked "backup and recovery" and "data reduction" in the top five (see Figure 2).³

¹ Source: ESG Research Report, [2010 Data Protection Trends](#), April 2010.

² Source: ESG Research Report, [2011 IT Spending Intentions Survey](#), January 2011.

³ Ibid.

Figure 2. 2011 Data Storage Spending Priorities – by Midmarket vs. Enterprise



Source: Enterprise Strategy Group, 2011.

First- Versus Next-Wave Deduplication Solutions

To understand and appreciate what matters in deduplication implementations today, an examination of early-stage offerings and how they compare with what ESG sees as next-wave deduplication solutions is warranted:

- Deduplication packaging.** The initial wave of deduplication technology was delivered via purpose-built appliances that, except for gateway implementations, were coupled with storage. These target devices were designed to process the entire non-deduplication backup load either pre- or post-ingestion on disk. Deduplication was specifically tied to the device on which it is shipped. Alternatively, next-generation solutions are modular and portable. Deduplication is an integrated component of backup/recovery software, and/or primary and/or secondary storage systems. An advantage is the flexibility in where deduplication processing occurs, as well as a singular deduplication approach.
- Deduplication differentiation.** The pioneers delivering deduplication initially focused on when (inline or post-process deduplication approaches), where (deduplication occurring at the “source” production system or “target” storage system), and how (the specific deduplication algorithm, such as a hash calculation and index comparison versus delta byte differencing) deduplication occurred. As IT’s use of the technology matured and evolved, it became evident that there is no “right” or “wrong” way of executing deduplication. Choices for when, where, and how deduplication is performed are dependent on the type of workload being optimized and/or the specific use case, such as backup data at a remote office being transported to a

central data center. Therefore, the focus of deduplication differentiation in next-wave solutions is flexibility: the ability to customize deduplication configurations and policies with a single solution.

- **Deduplication effectiveness.** Reduction ratio denotes how much optimization is yielded from data deduplication, typically the ratio of protected capacity to the physical capacity stored. Initially, it was a top-line point of differentiation in deduplication marketing. A 10:1 ratio means that 10 times more data is protected than the physical disk space required to store it and a 20:1 ratio means that 20 times more data can be protected on disk. Factoring in data growth, retention, and assuming deduplication ratios in the 20:1 range, 2 TB of storage capacity could protect up to 40 TB of retained backup data. While factors such as backup policy (full, incremental, differential), retention settings, data type (file types that are highly compressible versus those that are not), and rate of change from one backup to the next are important factors in reduction ratio achieved, early-phase deduplication offerings focused on the nuances of their deduplication approach for differentiation, such as the block size used for redundancy comparison and the ability to vary the block size to gain greater efficiency in hash-and-compare deduplication methods. The next-wave solutions still focus on which blocks are compared. However, the domain of comparison has also expanded to improve reduction ratios (i.e., global deduplication). The emphasis on maintaining a smaller index to improve performance is another consideration.
- **Deduplication architecture.** First-to-market solutions were characterized as single node scale-up architectures (one-dimensional expansion in capacity up to a maximum threshold). The trade-off in this approach is that, depending on the solution, there was limited flexibility and scalability. Users typically had to purchase an appliance configuration for today's needs, while planning for future growth—oftentimes being forced to over-purchase. The alternative (not planning for growth) typically resulted in a “forklift” upgrade in the not-too-distant future. Next-wave deduplication solutions are characterized by scale-out architectures (the ability to expand on multiple dimensions of capacity and performance). A highly scalable system that can seamlessly grow as requirements warrant—without necessitating a disruptive upgrade or running multiple systems that result in deduplication “islands”—can offer greater efficiency.
- **Centralized versus distributed processing.** Early-phase deduplication processing occurred on a single, high-performance controller—to deliver the fastest deduplication rates. Unfortunately, the single point-of-processing can also create a bottleneck. Deduplication solutions that can distribute processing at different points in the backup data path and/or across nodes in a clustered multi-controller configuration can deliver competitive performance rates while eliminating the risk of a bottleneck or a single point of failure.
- **Risk factors.** One of the big concerns with first-phase deduplication offerings was the risk of a “false positive” (a data block deemed redundant and discarded was actually unique and now lost). Early adopters overcame that concern via better education by first-phase deduplication vendors. Today, one of the main risk factors in deduplication systems is single point-of-failure. Losing the ability to consistently access data due to hardware failure or deduplication index corruption can be addressed via redundancy (system components, power and cooling, for example). Therefore, the introduction of high availability (HA) is a crucial element for continuity. Should the primary component fail, the second will automatically assume the primary role.
- **Focus on recovery.** When it came to performance, first-wave deduplication solutions emphasized backup ingest rates (i.e., write speed) to meet backup windows. Deduplication methods that could process redundancy checks, index updates, and disk writes faster were desired. However, as users gained experience with deduplication and encountered issues when trying to reconstitute highly-fragmented data chunks and recover data, the emphasis switched to focus on the speed of “rehydration” and reading data. Recovery performance is the critical criteria for businesses trying to quantify their recovery time objective (RTO) in the event of system or site failure.
- **Off-site copies.** Maintaining an off-site copy is a component of disaster recovery best practices. The ability to make off-site copies was focused on integrated tape creation and optimized device-to-device replication in first-wave deduplication. Today's next-wave deduplication solutions are extending off-site copies to

cloud storage. Nevertheless, tape is still an integral component of backup/recovery and disaster recovery strategies for many organizations. Therefore, vendors that can make recommendations for the best overall solution—keeping both recovery and long-term retention/archiving requirements in mind—are best positioned. Further, keeping the backup catalog informed of copies made by a deduplication target device and eliminating the need to reconstitute data between storage mediums are also key.

- **CapEx versus OpEx.** Cost is typically a top purchasing consideration. In the case of first-wave deduplication adoption, cost was measured in terms of capital expenses. The investment in the deduplication system had to be justified. As the market has matured, the operational expenses of deduplication have been exposed—specifically, the costs of managing and maintaining the solution, as well as the penalties for inefficient deployments.

HP StoreOnce Deduplication

HP introduced future-ready deduplication in its StoreOnce technology in 2010. StoreOnce deduplication is now available in HP StoreOnce B6200 Backup System, and backup/recovery software, HP Data Protector.

HP StoreOnce is aligned with next-wave deduplication solutions or as HP refers to it, “Deduplication 2.0.” Deduplication 2.0 addresses many of the shortcomings of “1.0 solutions,” such as incompatible deduplication algorithms implemented in software- and hardware-based offerings, comparatively slower recovery versus backup performance, a lack of HA, and less-than-efficient methods of scaling deduplication to meet ever-expanding capacity requirements.

Federated Deduplication

HP Labs developed a deduplication engine that can be deployed across the storage infrastructure. HP’s federated deduplication leverages a common StoreOnce deduplication algorithm. It provides deployment independence for deduplication since the common deduplication engine enables the native communication and movement of data across various HP systems without undeduplicating the data. This increases efficiency, especially as data is moved between sites over a low-bandwidth connection. It also provides flexibility in deduplication strategy. Federated deduplication allows data reduction to occur in HP Data Protector software or in HP’s StoreOnce family of appliances.

Deduplication Optimization

HP created a highly-optimized deduplication approach that introduces time- and space-saving techniques. With the goal of eliminating the maximum amount of redundancy in its data inspection, while also maintaining a small index to deliver the fastest performance, the company focused on two components of its deduplication approach: an average variable chunk size of 4K, and a sparse index.

HP’s Adaptive Micro-Chunking uses variable-length data segments or chunks. The backup stream is broken down into approximately 4K variable-length segments that are examined for redundancy versus previously stored information. Smaller segments means there are more chunks and index comparisons, which also means a higher potential to locate and eliminate redundancy (and produce higher reduction ratios). Comparative solutions use block sizes that range from 8K to 32K. The tradeoff with small chunk sizes is a greater number of index look-ups—which could mean slower deduplication performance. However, HP Labs developed HP Predictive Acceleration technology to maintain performance and reduce RAM requirements. By using a subset of key values stored in memory, StoreOnce determines a small number of sequences already stored on disk that are similar to any given input sequence—what HP refers to as sparse indexing. Then each input sequence is only deduplicated against those few sequences. This minimizes disk IO and uses less disk and little memory, creating more efficiency and enabling faster ingest and, importantly, restoration of data. HP’s approach accelerates reads/writes, and delivers rapid ingest rates of up to 28 TB/hour. Predictive Acceleration has enabled HP to require up to 50% less RAM than comparable solutions.

Scale-Out Architecture with High Availability

The HP StoreOnce B6200 Backup System is a large-scale deduplication appliance built on a scale-out architecture. It can grow from a single 48 TB two-node couplet to up to 768 TBs of storage capacity based on a massive single namespace. The pay-as-you-grow model allows for seamless scalability, enabling nodes to be added to the configuration without downtime. This is in marked contrast to first-phase scale-up deduplication appliances that, after hitting the maximum capacity threshold, require a disruptive “forklift” upgrade or additional standalone appliances to expand—introducing inefficiency, downtime, and management overhead issues.

The architecture also lends itself to high availability. With single node scale-up solutions, at any point in the backup data path, the controller node itself is a single point of failure. If the node fails, then backup or recovery fails. HP StoreOnce introduced high availability features to reduce this risk. Node failure is eliminated by pairing nodes within a couplet, so the surviving node can take over if its companion node fails. HP’s Autonomic Restart allows the restart of the backup job after node failover without any human involvement in the backup application or process. StoreOnce automatically detects certain failures and adapts to unpredictable failure modes while masking the complexity from operators and users.

Rapid Restore Performance

HP’s aforementioned Predictive Acceleration speeds performance— backup and recovery performance that HP claims are the fastest in the industry. The HP StoreOnce B6200 Backup System is able to restore data at the same rate as backup processing: 28 TB/hr. HP’s approach involves large-container technology which employs superior data layout. StoreOnce avoids a high degree of fragmentation by not replacing small amounts of duplicate data with pointers to faraway places with no other related data. Data is also defragmented after deduplication. The result is that restoring data takes less time because reconstituting it does not require many slow random seeks. This approach greatly improves restore speed with only a bit more extra data stored.

Off-Site Copies

Typically, disk-based backup with deduplication is a replacement for tape-based backup. If that’s the case, then how can backup sets be moved off site for disaster recovery purposes? The HP StoreOnce B6200 Backup System offers device-to-device remote replication. While there may be an added cost for acquiring and deploying a second system at a remote location, doing so will provide a safeguard in the event that the primary site (or the backup set managed at that site) is unavailable. HP provides consolidation from multiple deduplication systems to a centrally-located deduplication system at a fan-in ratio of 384:1, which can deliver greater economies of scale for disaster recovery. Further, HP simplifies software licensing and reduces costs for customers by only charging for the replication license at the “target” site. The “source” site replication licenses are included at no charge.

Cost Implications

ESG research found that, overwhelmingly, cost is most important to IT organizations’ evaluation and selection of a deduplication solution.⁴ The initial capital investment of a solution can be justified based on the cost benefits deduplication will deliver over the life of a solution. Since deduplication identifies and eliminates redundant data, it optimizes storage capacity—effectively reducing the cost of disk by allowing more backup data to be stored on the same footprint (increasing capacity up to 10 to 30 times). Lower disk capacity requirements impact hardware acquisition expenses, power and cooling fees, and operational staff costs.

In addition to eliminating the need to over-provision on the initial purchase of an HP StoreOnce solution, HP’s new B6200 costs about 20% less than competitive offerings. Furthermore, the scale-out architecture and HA features of the HP StoreOnce B6200 Backup System contribute to reduced operational costs based on the lower administrative overhead associated with maintaining and managing the deduplication environment.

⁴ Source: ESG Research Report, [2010 Data Protection Trends](#), April 2010.

The Bigger Truth

As more organizations implement disk in the backup process, data deduplication is a fast follower. It dramatically improves the value proposition of disk-based data protection since it eliminates the redundancy typically seen in secondary storage processes. The use of deduplication will drive further backup-to-disk adoption and deliver associated backup performance and reliability benefits.

Selecting a strategy for data deduplication requires consideration of several factors so there are no surprises down the road. Having a clear understanding of how deduplication works and what capabilities are most important goes a long way toward selecting and designing a solution that delivers maximum business, operational, and financial benefits.

HP offers a more forward-focused approach to deduplication. The technology is:

- Modular, enabling it to be embedded/integrated in HP data protection hardware and software components.
- Efficient, eliminating the need to “rehydrate” data to move it between storage systems.
- Highly scalable, increasing flexibility and eliminating up-front over-purchasing .
- Faster with restores, enabling better and more predictable RTOs.
- Highly available, removing the risk of a single point-of-failure in the storage system.

These and other features of HP’s StoreOnce-enabled deduplication solutions are aligned with next-wave deduplication requirements. The architecture of HP’s solution has a significant impact on HP being able to offer the flexibility it does: federated deduplication (hardware- and/or software-based deduplication), ease of scale, and high availability. HP’s deduplication approach has implications for the backup and, more importantly, recovery performance it delivers: distributed processing; small, variable-sized chunks; and sparse indexing to deliver high redundancy matching using a small, memory-resident index.

Choosing a deduplication strategy is not a simple task. Technology maturity varies considerably and the vendor landscape is in flux. As solutions are considered, cut through the hyperbole. Test vendors’ claims. Thorough due diligence up front may stave off surprises later.



Enterprise Strategy Group | **Getting to the bigger truth.**

20 Asylum Street | Milford, MA 01757 | Tel:508.482.0188 Fax: 508.482.0218 | www.enterprisestrategygroup.com