

White Paper

Permabit Albireo: Primary Data Deduplication That's Primarily About Massive Business Value

By Mark Peters and Steve Duplessie

August, 2011

This ESG White Paper was commissioned by Permabit and is distributed under license from ESG.

Contents

| | |
|--|----|
| The Market and Opportunity..... | 3 |
| Introduction | 3 |
| Market dynamics..... | 3 |
| Scale of the Opportunity | 4 |
| Data Efficiency is a Necessity Not a Luxury..... | 5 |
| What’s Needed and What Albireo Has..... | 6 |
| Specific Product Needs..... | 6 |
| Albireo: Delivering on the Requirements | 8 |
| Proving the Capabilities & Value | 10 |
| The Bigger Truth | 12 |

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.

The Market and Opportunity

Introduction

There is a major, growing, and well-known issue in IT—data growth is accelerating, and it is doing so at a rate that defies our budgetary ability to deal with it using standard methods and tools. And yet a technology that can address the issue is actually available today – data deduplication. While it is widely used today in backup and archive, there are enormous efficiencies that could be generated if we didn’t store lots of data in the first place....if we deduplicated primary data before it ever hit a spinning track or a solid state cell. It’s a “light-bulb” moment – and when it takes hold it will forever change the economics of storage. [Permabit’s Albireo](#) product can do just that— flexibly, economically, without performance impact and for every tier of storage.

Market dynamics

This need for a significant change—simultaneously addressing a huge industry challenge *and* presenting a great business opportunity—was discussed in depth in a recent ESG paper¹. Permabit is clearly a company that understands the issue and its Albireo Data Optimization Software is a tool that can meet every market need – it scales, it is compact and resource efficient, it delivers superior deduplication rates and it actually improves storage performance. While reading the full paper gives the most complete understanding, some of the main arguments are noted here:

1. While specific technologies can be seen as IT game-changers, the truth is that better—*way* better— economics are invariably the root value driver. Technologies may enable technical change....but it is their financial impacts that create true market shifts and value creation.
2. One of the biggest challenges facing the IT industry today is that storage demand continues to grow both rapidly and—more importantly—at a rate that exceeds its relative price decline. It is not a sustainable model: data growth trends are driving the need for another game-changing event. Data growth is outstripping IT CAPEX budgets, available IT space, management capabilities, and OPEX thresholds, and the pressures to manage more data in cloud environments only makes the challenge more acute. We are approaching—some users are even at—a breaking point.
3. Since the price of storage isn’t declining sufficiently fast, the only other realistic option is to reduce the amount of data that actually gets stored. The necessary sea-change for storage is for deduplication to be broadly applied to primary storage. It represents the level of change in both CAPEX and OPEX that produces the necessary “slap in the face” economic improvement to make IT users demand it from their vendors.
4. ESG research shows that organizations cite cost reduction as an important means of justifying IT – including storage – investments (see Table 1).² As the graphic also indicates, the research has, for the last few years, shown a greater emphasis on bringing down OPEX as opposed to CAPEX, no doubt as a realization that the lifetime costs of IT (which for storage would include space, power, and management, etc.) can be much greater than the initial purchase price. Data reduction can drop OPEX and CAPEX dramatically.

Table 1. Most Important Considerations for Justifying 2011 IT Investments, 2009 vs. 2010 vs. 2011

| Which of the following considerations do you believe will be most important in justifying IT investments to your organization’s business management team over the next 12-18 months? | | | |
|---|-------------------------|-------------------------|-------------------------|
| | 2009 (N=492) | 2010 (N=515) | 2011 (N=611) |
| Reduction in operational expenditures | 62% | 54% | 43% |
| Reduction in capital expenditures | 37% | 30% | 24% |

Source: Enterprise Strategy Group, 2011.

¹ ESG Market Report: [How Economics Alter the Storage Landscape](#), May 2011. Some of the sentences in this section are rephrased, others are taken directly from the report.

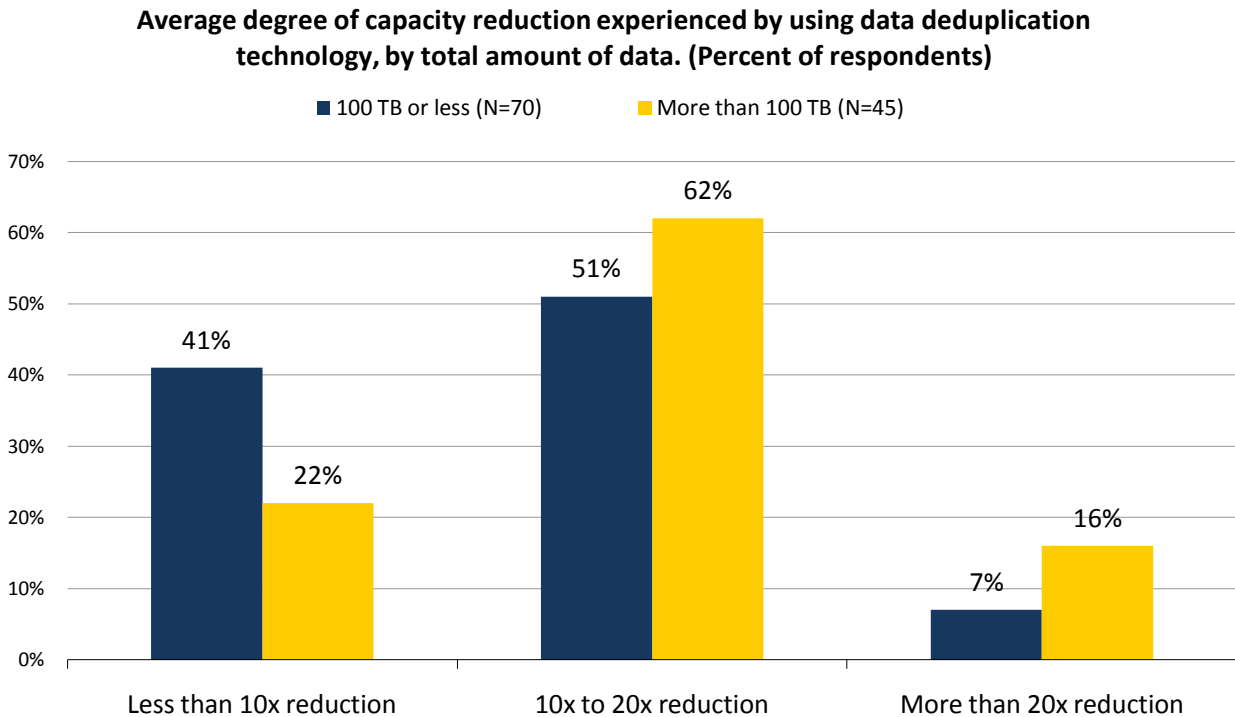
² Source: ESG Research Report, [2011 IT Spending Intentions Survey](#), January 2011.

5. Although the storage industry has done many clever things to address the escalating cost of storing data—such as dynamic tiering, thin provisioning, and the like—such methods provide only incremental cost improvements. Indeed, without a sea-change technology, the ever-widening gap between IT demands and constrained budgets means that keeping data will become essentially unaffordable and users will gradually and increasingly have to lose or discard it.
6. History shows us that the only vendors to successfully alter the status quo (and thereby steal appreciable share from the incumbents) are those that offer solutions that are 10X or more “better” than the competition. ESG refers to this as the ‘Law of 10X’. You can argue the degree of improvement—it may be 12X or 17X or something more—but the point is that it is an immediate and extreme change that results in dramatically improved financial value that cannot be ignored.
7. First-mover advantage is available to any company that understands all this and makes the land-grab while other players are talking about science projects and roadmaps or simply trying to maintain the status quo.

Scale of the Opportunity

1. A range of industry projections show annual data creation being in the tens of zettabytes over the next decade and that typically only 20% - 30% of that data is unique; the impact of data optimization could eliminate literally tens of zettabytes of duplicate data. It’s a huge number, with a huge potential impact, and it is a huge market opportunity. To give real scale to the Albireo opportunity, Figure 1 shows the real world impact that deduplication can have on the amount of data actually stored. It shows that deduplication is definitely very advantageous when it comes to data capacity reduction, and especially so among organizations with more data. In organizations with more than 100TB of data, 76% of the respondents reported capacity reductions of 10X or more.

Figure 1. Capacity Reduction from Data Deduplication Technology, by Total Amount of Data



Source: Enterprise Strategy Group, 2011.

Other factors—key current market adoption waves— accelerate the need and the opportunity for a smart, simple and efficient way to ensure dramatically less data is stored and as a result dramatically less storage space consumed:

- **Cloud:** As well as the obvious capacity requirements, cloud implementations invariably highlight cost efficient solutions. This will be a fiercely competitive business and only the most cost-efficient providers will survive and deduplication will most likely become a ‘table stakes’ technology.
- **Virtualization:** To date, less than 50% of server workloads have been virtualized. Deduplication all but eliminates the ‘storage bloat’ caused by virtualization and associated operational bottlenecks (i.e. boot storms) thus helping virtualization to realize on its game-changing promise.
- **Solid State:** Deduplication applied to data on solid state – whether persistent or as cache – has the same effect as anywhere else in the storage hierarchy, which is to dramatically reduce the effective cost and improve reliability. With primary data deduplication, SSD’s will be delivered at 10% of their current effective price! Think of what that will do to accelerate their adoption and to drive volumes up and real costs down.
- **Big Data:** Big Data workflows often collect one-time readings and measurements (for instance, seismic data, genomic sequencing and agricultural monitoring) and combine this irreplaceable data with analytics. Because the data is critical, it is replicated multiple times. Deduplication makes the multiple copies essentially free – yet maintains data safety.

The market opportunity is huge (billions of dollars), and - let’s face it - not every storage company will catch the ‘primary deduplication train’. It is likely that one or more of the biggest market incumbents—ones that look invincible today—will get hurt (in terms of sales, market share, revenue and valuation) unless they move very quickly for implementation and leadership, while one company will likely gain first mover advantage and gain the lion’s share of the game changing economic benefit.

Data Efficiency is a Necessity Not a Luxury

There is not a single data storage forecast that predicts anything but long term, continuing, dramatic growth; and this fact puts “data efficiency” into the “must have” category. It is not a mere feature, but is instead emerging as a prerequisite core element for successful IT – as important as data safety, data security and performance. It’s as simple as that. Surprisingly few vendors have the necessary competency given this market outlook, as this is a tough technology nut to crack. In this area Permabit is in a ‘class of one’ with its combination of performance, scale, deduplication effectiveness and resource efficiency – when you look at the available options from other vendors, all are limited in some, or many, respects. While Permabit is beginning to promote “Dedupe Everywhere” (this is Albireo’s flexible ability to be deployed throughout the IT infrastructure and will be covered more in the next section) the current storage optimization alternatives—from such industry stalwarts as Dell, HP, IBM, NetApp, Microsoft, and EMC—are often just for backup, or very constrained in scale, or even still only in development.

There can be little doubt that data deduplication is one of **the** key technology and business shifts in storage, up in the ranks like the Winchester drive, networked storage and RAID. It’s not just about storing more and making storage more flexible – it is about storing less and making storage far more economical. Permabit does this with no negative trade-offs and creates massive downstream value in terms of less physical storage capacity being needed – a.k.a delivering a dramatically better bang for your storage buck – from end-to-end (from SSDs to replication, archive and backups) as well as with less management and overall improved performance. In an IT world where “doing more with less” has moved from a marketing slogan to a crucial survival mantra, squeezing orders of magnitude more from every dollar speaks loudly. Think of a conceptually similar play as Data Domain, only this time it is applied to overall IT efficiency and not just backup; this time it applies to all storage, and so the impacts could be significantly bigger.

Permabit was established in 2000 by MIT engineers and has a seasoned core competency in deduplication, initially developed for its Enterprise Archive and Cloud Storage solutions. More recently, seeing the massive market potential for primary, secondary and backup optimization, and realizing that its implementation is notably more

flexible, scalable, higher-performing and systems-agnostic than any other optimization offering, it built a deduplication solution into an SDK (Software Development Kit) called Albireo. Albireo utilizes optimized deduplication algorithms and is architected for implementation across platforms, up and down storage hierarchies, within applications and even in OSs.

What's Needed and What Albireo Has

Specific Product Needs:

While the compelling factor for the adoption of Albireo is the financial one – delivering against the “Law of 10” with a 10X+ financial impact – there are still a whole bunch of “table-stakes” operational issues to also be addressed by any IT product. Some of the key “basic” requirements are:

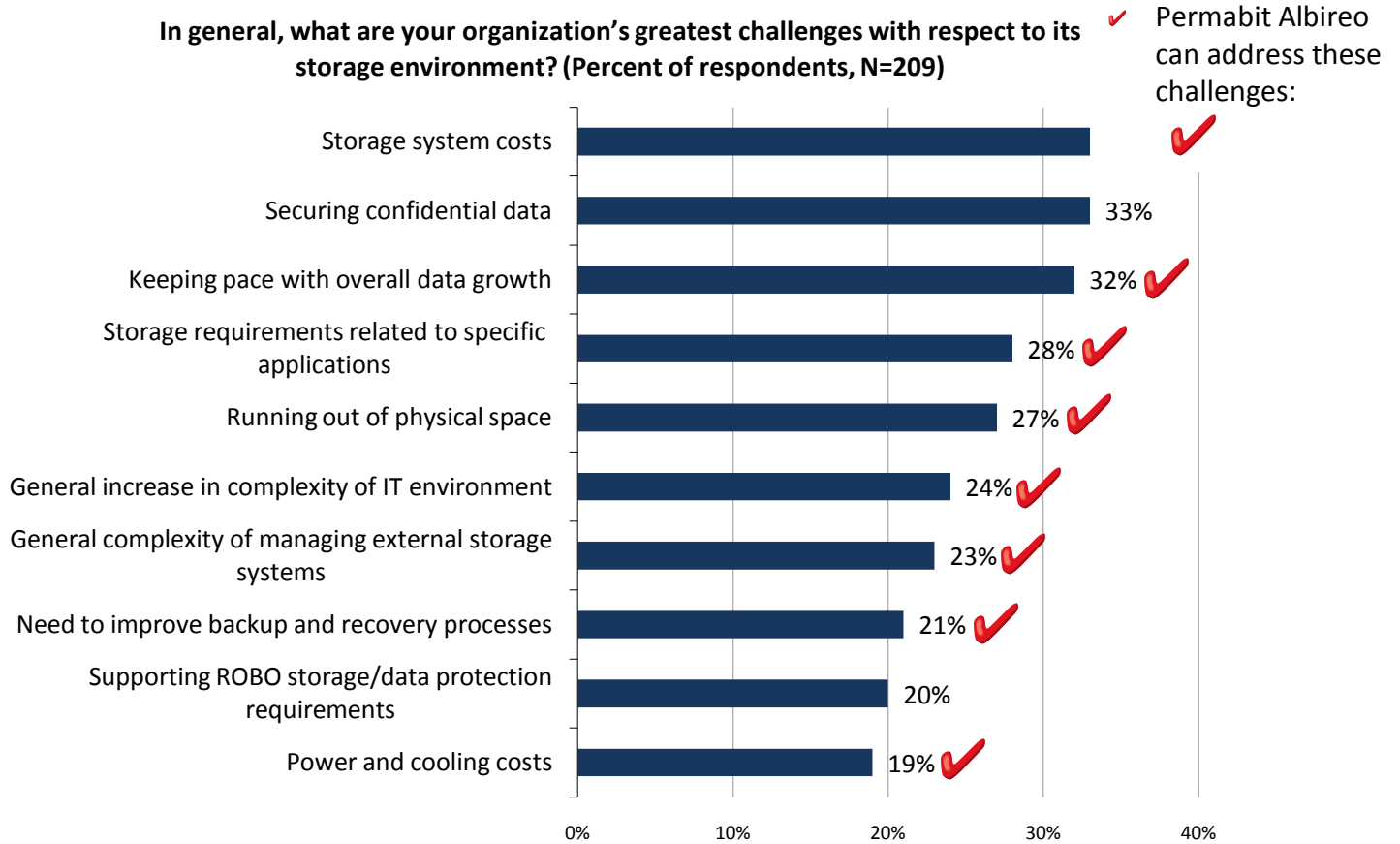
- Scalability to support data growth. Given current and projected data growth this is a critical component. Data stores are going to be multiple petabytes in size and data reduction technology that cannot scale to such levels is nothing but a dead end.
- Performance to meet primary storage needs. Primary data storage performance cannot, under any circumstances, be impacted by data reduction. Vendors have built their storage differentiation around performance, data I/O and overall throughput - these cannot be impacted.
- Efficiency to deliver the expected (or better!) data reduction. Efficiency in resources (memory & processor) utilization is key to broad adoption across existing storage platforms. Deduplication efficacy with sub-file and content aware capabilities is a direct result of the ability to index huge amounts of data (hundreds of billions of data chunks) to insure that most duplicates are found and, as a result, the most storage space is saved.
- Data safety to protect data. Any data reduction technology cannot, by its implementation, put data at risk. There cannot be single points of failure typical of appliances in the data stream.
- Transparency so as to preclude sacrificing system or application performance, while still retaining other existing differentiating functionality. Storage vendors have spent millions of dollars in developing their differentiation in areas such as performance and data protection. Any data reduction technology must be transparent and enable the vendor differentiation to be maintained.

Beyond the crucial basics, other “really good to have if at all possible” aspects appear. For instance, support for all types of storage would be a significant “plus”: primary, secondary, backup, archive, file, block, SSD, cloud, and whatever else there is. And of course that flexibility would be better still if it stretched to file, block, and object—maybe it could be flexible to be deployed inline or post-process, and so on. The more broadly and transparently the deduplication can be applied, then the more its downstream value appreciates.

Having multiple copies of data—at any point in the data hierarchy or lifecycle, and except for intentional replication—is definitely sub-optimal. Put more bluntly, it is a waste of money. The money wasted is not just the CAPEX and OPEX of the storage itself (with all its additional copies) but is further compounded by additional investments in compute power and network bandwidth as well—all of which of course also require physical space, management, and power and cooling. Preventing unnecessary data being stored in the first place therefore creates a “virtuous cycle” of benefit that goes to the heart of addressing the core challenges users have with their storage environments. According to the ESG research³ shown in Figure 2, Permabit’s Albireo is well aligned to address eight of the top ten challenges users (whether enterprise or mid-range) have with their storage environments.

³ Source: *ESG Research Storage Study*, March 2010.

Figure 2. Greatest Challenges with Storage Environments



Source: Enterprise Strategy Group, 2011.

Albireo: Delivering on the Requirements

Permabit Albireo is a genuine industry game-changer. It meets and passes the “so what?” test...is a better, more efficient, high-performing, flexible and complete implementation than any others have... and will gain attention from both the OEM community and end-users by being able to deliver against the “Law of 10X.”

What is it? Albireo is a proven data deduplication engine that provides unified data deduplication advisory services, or, as the vendor puts it: “Permabit Albireo is the industry’s first purpose built data deduplication software designed to meet the needs of storage and platform OEMs who wish to expand their existing storage solutions without negatively impacting existing differentiating capabilities or overall performance. Albireo delivers deduplication at the sub-file level and so can be flexibly integrated into existing or next generation OEM storage and/or platform technologies with the ability to be deployed as an inline, parallel or post process solution”.⁴ Obviously this paper is not designed to be a datasheet and will concentrate on analysis rather than pure description. For a complete overview of the product we suggest referencing the company’s own white paper⁵.

Architecturally it has a number of excellent features:

- Industry leading IO (performance and throughput) to perform at the high throughput levels that primary storage demands; checking for duplicates usually takes less than a few microseconds. With this performance, Permabit meets the needs of the ‘worst case use case’ for deduplication.
- Massive – essentially unlimited – scalability up into the multiple petabytes range meets current and projected future growth demands enabling deduplication on the largest data stores to provide long term and effective data reduction and cost savings.
- Extreme resource-efficiency (in terms of compute and management demands) which will flexibly deliver deduplication to existing data storage platforms no matter what configuration they have. As part of this resource efficiency, it precludes the negative system performance impact that is such an issue with lesser deduplication approaches.
- Guaranteed data safety and integrity obviates the data risk imposed by appliance-based (single point of failure) implementations. Permabit’s reduced data risk is enabled by being completely out of the read path.
- Supports sub-file, content aware segmentation to deliver extremely efficient data reduction that provides the most effective and space efficient deduplication available.
- Scalability is not only in terms of capacity; it can deploy from a single storage controller to a cluster of storage controllers or a cluster of deduplication capabilities that communicate with a storage system over an industry standard Ethernet interface - enabling nearly limitless scalability.

Albireo Can Apply “End-to-End and Everywhere”

One of the most impressive parts of Albireo is the sheer range, flexibility and extensibility of the product. It is not only a comprehensive solution, but it is also highly scalable to multiple petabytes and it has the performance necessary to meet the rigors of primary storage. These attributes make Albireo a highly differentiated solution in the marketplace (where other technologies might only address backup, or have scale limitations, or only deal with storage) and its applicability and capabilities are broad from any viewpoint:

1. **Storage media:** Flexibility to deploy on any media type – from SSD to SATA.
2. **Topographies:** Albireo comprehensively addresses the portfolio-wide needs of major IT (block, file and unified architectures) suppliers. This flexibility eliminates the need for overlapping product line specific investment and leverages R&D investment.

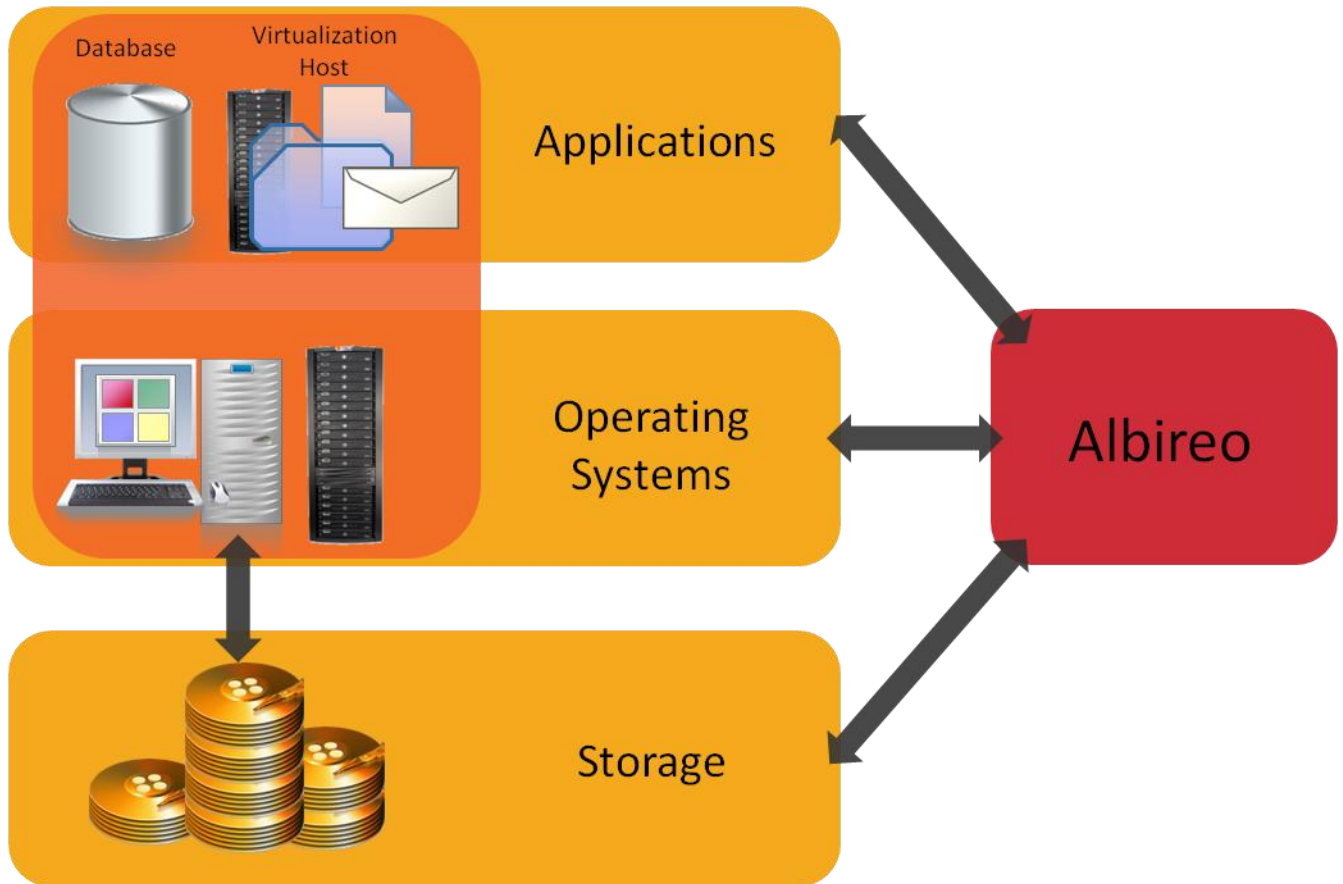
⁴ Permabit website – www.Permabit.com – 07/05/11

⁵ Permabit White Paper; *Permabit Albireo High Performance Data Optimization Software*, May 2011

3. **Storage protocols:** iSCSI, FC, CIFS, NFS providing data reduction across all types of protocols
4. **Targets everywhere:** Can fit a variety of storage, OS's, applications, databases and middleware requirements up and down the stack
5. **Implementations:** Users and/or OEMS can choose to deploy inline, parallel or post-processing in combination or uniquely to optimize resource utilization and efficiency
6. **Data types:** Suitable for primary, secondary, back up and archive data for end-to-end data optimization
7. **Locations:** Data centers, remote/branch offices, or clouds, to optimize data storage and network utilization
8. **Expected growth:** Scales to multiple petabytes to meet current and projected market needs
9. **Growth markets:** Big Data, virtualization and cloud storage are all impacted as a result of high data growth. Data optimization enables these to be affordable and viable because of the significant impact it has on effective cost of storage
10. **Growth Media:** SSD technology adoption is rapidly accelerating and data optimization enables the effective cost of SSD to compete with spinning disk on a \$/GB basis
11. **Growth Architectures:** Flexible and scalable to fit scale out architectures of multiple PB's.

Graphically the capabilities can be summarized as shown in Figure 3:

Figure 3. Albireo "End-to-End and Everywhere"



Source: Enterprise Strategy Group, 2011.

The “end-to-end” (across the storage infrastructure) and “everywhere” (throughout the applications and OS stack) “universal deduplication” brings economic value to end-users and OEMs alike. Deduplication provides an “efficiency multiplier effect” as data moves through its lifecycle from applications and databases across the storage tiers and into archive and backup. Costs are reduced initially and the space savings are maintained since the data is never rehydrated as it moves throughout the storage tiers.

End-users will clearly benefit from lower CAPEX and OPEX needs (which are multiplied by having just one integrated deduplication tool to manage). OEMs providing this added value to their users drive down effective costs, space requirements and operational needs, hence delivering competitive advantage that will be highly likely to generate rapid competitive share and revenue gains for the prime and early movers.

Proving the Capabilities & Value

Clearly the Albireo story is a very strong one but how real is it? After all, the IT market has been littered over the decades with stupendous “PowerPoint value” – claims and abilities that either never materialize, or that only deliver a fraction of the promised value. While Permabit can point to its excellent thought leadership, it is also fortunate that it can tangibly prove Albireo—albeit the product itself is just a year old—from four credible angles.

Actual OEM’s and Users: Albireo’s capabilities are actually in use, both directly (via active OEM agreements and end-user installations) and also via the genetic code pool that has led to Albireo (via 100’s of Permabit Enterprise Archive and Cloud Storage offerings).

ESG Lab Validation: ESG Lab first evaluated and tested Albireo in late 2009, and – as this paper is being published – is in the midst of further testing, from which early results have been used to update the quotes below. ESG published its original Lab Validation Report⁶ on Albireo, which

“confirmed that the Albireo deduplication advisory services work as advertised. *The capacity of data sets was reduced between 33% and 97%, for real world applications including office productivity files, virtual server images and email backups.* The patented deduplication lookup and indexing algorithm was incredibly fast (over 400 GB/sec) and extremely resource efficient (under 0.1 bytes of memory per index entry). ESG Lab is confident that the flexibility provided by the Albireo SDK is unique. It confirmed that Albireo can be used with file or block storage; at the object or sub-file level; can support an inline or post-processing model with minimal performance and resource impact. Running over a grid, it can be used to create a global pool of deduplication with predictably scalable performance and rock solid reliability. Last, but not least, the Permabit Albireo SDK was designed with quick and easy integration in mind.”

Market Value Suitability: The market research shown previously dealt with the general business initiatives and challenges of storage environments. Figure 4 turns to consider a more specific storage aspect *within* the parameters of server virtualization environments: what can be seen is the huge relevance of such things as primary deduplication to the continued success and adoption of server virtualization⁷ (something at the forefront for many of the OEMs that Permabit is focused on to license Albireo).

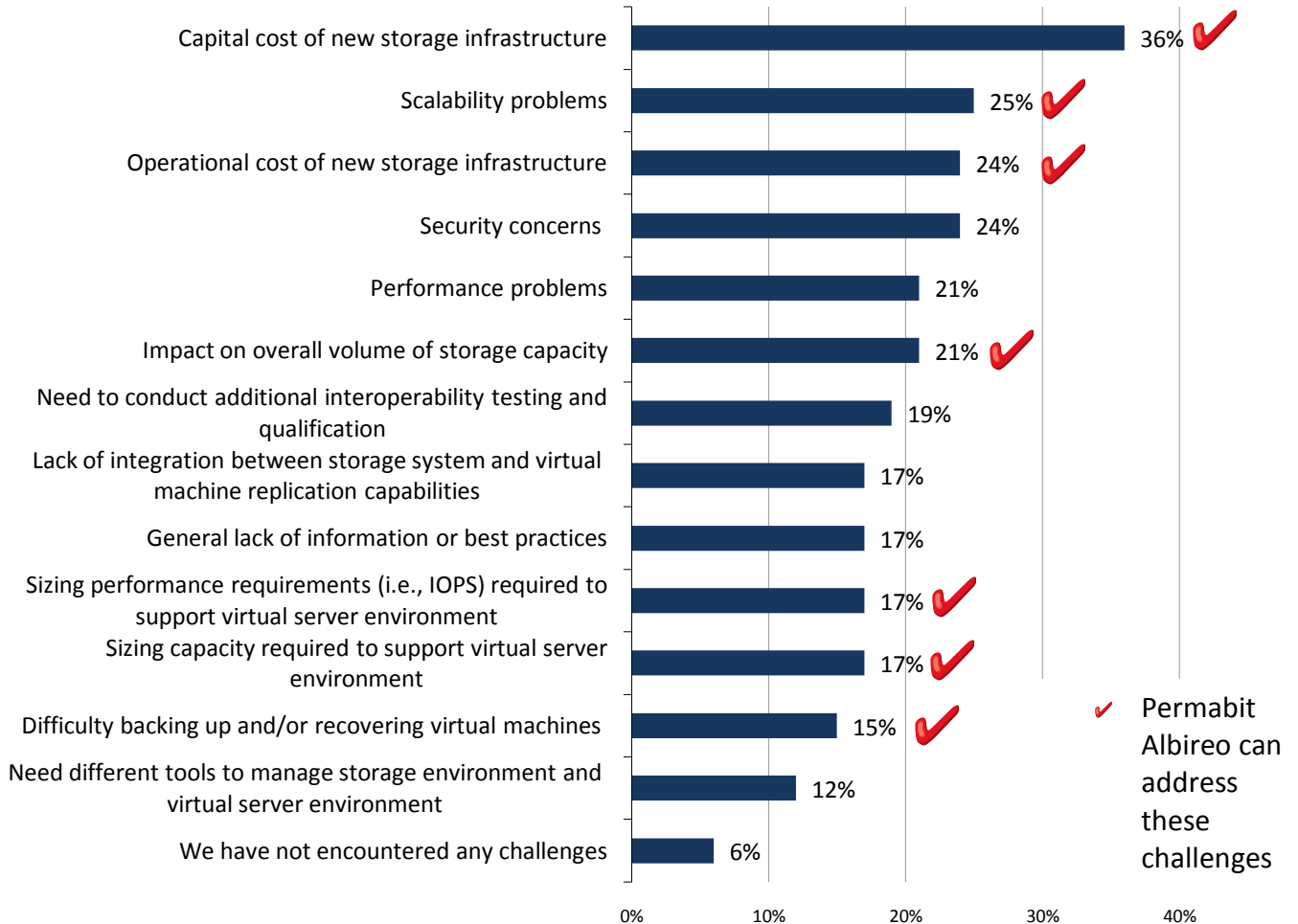
It can be seen that many of the storage challenges *specific to server virtualization* environments are ones that can be mitigated by data optimization of the sort that Albireo provides. Given that storage is all-too-often becoming an operational and financial anchor restraining the progress of server virtualization, this alone should capture a lot of interest for Permabit, which can be seen to be able to address 7 of the top 12 challenges that respondents report having in this area.

⁶ ESG Lab Validation Report: Permabit Albireo: Empowering Unified Deduplication, ESG Lab November 2009. This section of the paper is edited/derived from the Report

⁷ Source: ESG Research Report, [The Evolution of Server Virtualization](#), November 2010.

Figure 4. Storage Infrastructure Challenges Related to Server Virtualization Usage

**From a storage infrastructure perspective, which of the following would you consider to be significant challenges related to your organization’s server virtualization usage?
(Percent of respondents, N=190, multiple responses accepted)**



Source: Enterprise Strategy Group, 2011.

Patent Portfolio: Permabit has built a strong IP portfolio, which is proof of its thought leadership and development capabilities, as well as crucial to its value (both *to* users and *of* itself), and also a significant barrier-to-entry for would-be followers. The company has been granted a total of 26 patents (of which 23 are issued) covering diverse areas in data protection and archive with many more filings pending. Amongst the most recent patents awarded (Spring 2011) was one for “data deduplication for secure data sharing in the cloud”, which demonstrated Permabit’s determination to deliver on its “end-to-end and everywhere” market vision. The company’s growing IP portfolio includes patents in the areas of hash-based deduplication for scalable file and object data storage, encrypted deduplication, memory based snapshots, and many other features of Albireo.

The Bigger Truth

Deduplication, as long as it is done economically and efficiently, has to be about as close to a ‘no brainer’ as one can get in IT. Why waste money storing the same thing multiple times?! In many other walks of life such profligacy would be viewed as job-threatening, if not criminal. The idea that we can address this problem is definitely a “penny-dropping” realization. Permabit got the name “Albireo” from a double star that appears as one. And primary deduplication is an example of another vision anomaly—that of a “scotoma,” which describes something to which you were effectively blind before it was pointed out to you....which you now see all the time. Writing less data so as to store less data—what a concept!

With Albireo, Permabit has a supremely effective and flexible tool—one that can apply deduplication end-to-end and everywhere. It can better enable the Cloud....and can just as easily better enable the use and adoption of solid state storage. Given that the main aim of the product is OEM adoption, it’s understandable that the end-user marketing volume hasn’t been that high. However, we would encourage Permabit to shout louder where and when it can about this all-encompassing approach to deduplication that Albireo represents. After all, it would be a strange end-user who would not want her systems and storage vendors to offer this capability!

Let’s close where we started: the reason this all matters so much has to do with money, not technology. It is the result of the huge economic imperative generated by the dramatic and growing mismatch between storage demand and [affordable] supply. With no magic end in sight to that general mismatch, the only way to make the math work is to find a way to store less data...eschewing cutbacks in applications or support levels, the only realistic option is therefore to address the storage volume before it is stored. This sounds like a riddle, but it is one that a data optimization approach such as Albireo can solve. Adoption of primary deduplication is inevitable, and the opportunity for the vendors that can move fast to integrate it is tremendous—not just because early movers tend to get the main share of any market, but also because the market opportunity is expanding exponentially (thanks to Big Data, Virtualization, the Cloud, and other factors). The current data storage model is simply unsustainable, ordaining that we find a way to reduce the escalating demand for extra storage capacity and while a sea-change of this nature can seem daunting to some vendors, it should be embraced so as to avert a tsunami of alternative issues.



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