IBM System Storage

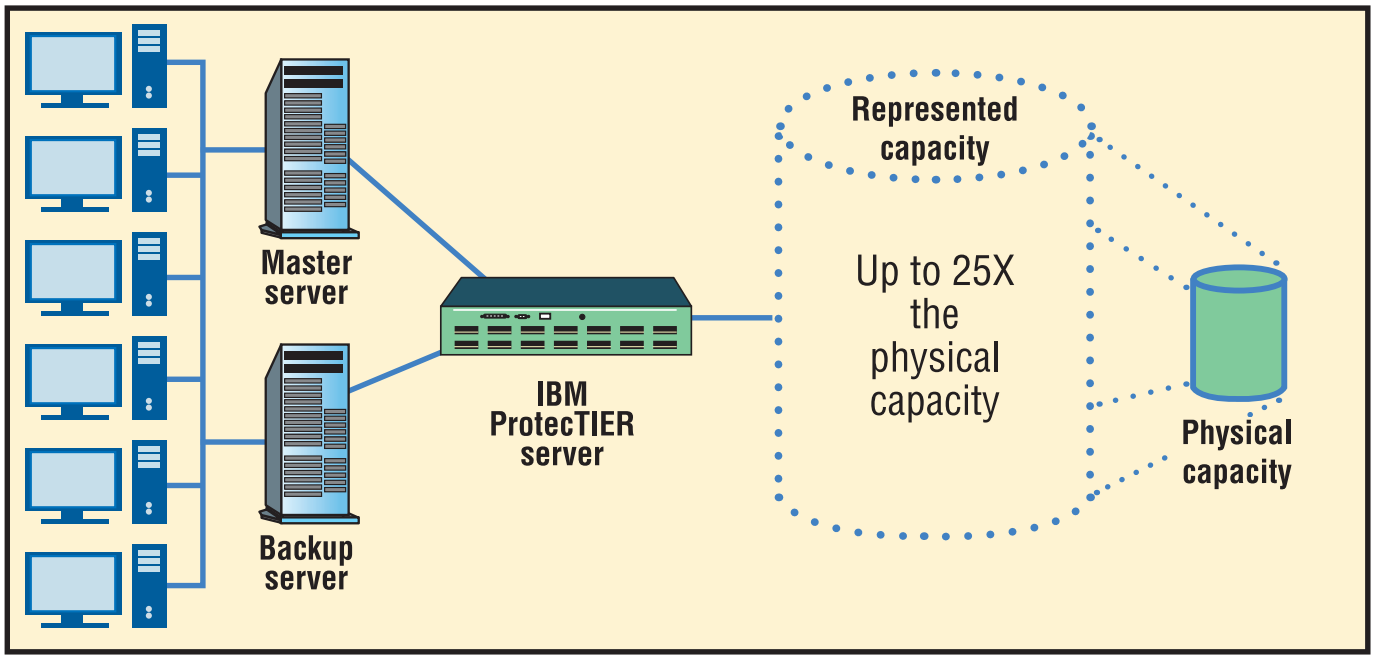
Use IBM data de-duplication to back up more data with less disk space

Until now, the only way to maintain high-level data protection for growing amounts of enterprise data was to add more disk space to the data protection infrastructure—an approach which can quickly become cost-prohibitive as data volumes continue to grow, while capital budgets for infrastructure do not. Regardless of cost, however, enterprises still have to be able to reliably back up data and perform back up and restore operations without disrupting business. Ultimately, the challenge is to find ways to do more with less infrastructure.

Data de-duplication solutions from IBM employ an advanced form of data compression that identifies and eliminates redundant data across the data landscape, making it possible to significantly reduce the amount of data that needs to be protected. This in turn dramatically increases the effective capacity of existing disk storage, so that far less physical disk is required to protect the data. Beyond the direct resource savings associated with needing less disk space—which can be in the hundreds of thousands or millions of dollars—the benefits of data de-duplication include:

- *Greater productivity that comes from being able to perform more frequent backups with the same amount of physical disk.*
- *Increased efficiency because of the greater likelihood that data will be able to be restored from disk rather than a slower medium.*
- *Reduced energy consumption that results from reducing the amount of disk in operation.*

Increasing capacity through IBM data de-duplication technology



Like other types of data compression, data de-duplication uses algorithms to compare data and eliminate duplicated or redundant data. Unlike other methods of compression, however, which work with small amounts of data at the file level, data de-duplication can be applied at the subfile level to reduce data size across a much broader landscape.

There are a number of data de-duplication solutions available today, and they vary greatly in how they are implemented, how they work, what issues they address, and the risks and rewards that accompany them. The key to finding the right solution is to knowledgeably evaluate the choices underlying technology and capabilities based on specific needs.

IBM: The right choice for enterprise-class performance

IBM System Storage™ includes data de-duplication solutions designed to meet the disk-based data protection needs of the enterprise data center while enabling significant infrastructure cost reductions, specifically the IBM System Storage TS7650G ProtecTIER™ De-duplication Gateway. The following characteristics make IBM the ideal choice for enterprise-level solutions.

Non-disruptive deployment and operations

IBM TS7650G ProtecTIER De-duplication Gateway solutions deliver high-performance de-duplication for virtual tape libraries (VTLs), which may be particularly well suited to enterprise organizations because they make it possible to leverage existing backup applications and processes. And, in fact, according to research by Enterprise Strategy Group,[1] enterprise-level organizations are likely to adopt VTLs as a means of implementing disk in the data protection infrastructure. VTL solutions are non-disruptive because they take a target-side approach to data de-duplication, meaning that the de-duplication takes place after data is processed by backup software rather than at the protected machine. IBM solutions in particular are designed for easy integration into existing data center environments, requiring no changes to backup policies, practices, or procedures that are already in place.

The IBM solution architecture is also designed to be non-disruptive in the sense that solution operations are not likely to disrupt production activities due to downtime or availability issues. IBM TS7650G uses an in-line de-duplication approach in which data is de-duplicated in real time, so that it is already de-duplicated when it is written to disk, thus reducing the risk of downtime. This is in contrast to the post-process de-duplication approach that some solutions use, in which backup images are written to disk before de-duplication, making downtime *more* likely. This is why it can be especially important to use an in-line approach such as IBM's in enterprise-class environments, where there is likely to be little tolerance for downtime.

Storage capacity demands reduced by up to 25X

IBM's patent-pending HyperFactor technology uses a pattern algorithm that can reduce the amount of space required for storage in the backup environment by up to a factor of 25, based on evidence from existing implementations. The capacity expansion that results from data de-duplication is often expressed as a ratio, essentially the ratio of nominal data to the physical storage used. A 10:1 ratio, for example, means that 10 times more nominal data is being managed than the physical space required to store. Capacity savings of 18:1 and greater have been reported from data de-duplication—up to 25:1 in the case of IBM solutions.

An important point to keep in mind in considering different data de-duplication solutions is that the reported ratio of nominal data to storage used can vary tremendously among solutions, growing to 30:1 or more. And while it may seem that a higher stated ratio would mean a superior solution, this is not necessarily true. One reason is that the realized de-duplication ratio depends heavily on variables such as the data retention period, the data change rate, and the backup practice. For example, the number of days that data is retained has a direct impact on the factoring ratio. Another reason a higher ratio is not necessarily a better one is that de-duplication ratios can be calculated in different ways. If, for example, the calculation ignores the disk overhead required for the system, that will artificially inflate the ratio. So will focusing on just the de-duplication ratio of a given data stream. This is why it may be that a ratio of 500:1 does not necessarily offer better de-duplication than a ratio of 20:1.

Highly scalable, high-performance solution

The capacity savings achieved by IBM solutions can be attributed to sustained high performance, which is in turn attributable to their granularity and scalability.

- Granularity *refers to the size of the chunks of data that are being examined for redundancy. The smaller the chunks, the more of them can be compared. IBM TS7650G solutions find and eliminate data at a very fine grain—capturing small data matches to the 2K size, which enables up to 25X de-duplication in a typical data protection environment.*
- Scalability *has to do with a solution's sustainable throughput. In IBM solutions, a cluster topology enables sustainable throughput over 900 MB, regardless of repository size. The result is enterprise-class performance to meet the most demanding data center requirements. IBM solutions are scalable to up to 1 PB of physical storage (over 25 PB of user data), which enables easy scaling of both performance and capacity.*

Increased capacity with less risk to data integrity

With business needs and regulations driving requirements for long-term, on-site, disk-based data retention, an effective solution at the enterprise level must allow for repositories of hundreds of terabytes on a per-system basis. Support for anything less than 20 TB will result in the need to manage more and more independent islands of storage. This is not an issue with IBM TS7650G solutions, since they provide up to 1 PB of storage capacity for each storage system.

In high-capacity environments, some data de-duplication solutions pose a risk to data integrity due to the method by which they find and process data redundancy. For example, solutions that use hashing algorithms for de-duplication present the risk of data loss due to hash-collision. While this risk may be statistically low, it may well be realized in very large environments. And when it is, there is no way to know about it until the data needs to be retrieved/restored. For this reason, IBM has chosen a pattern recognition algorithm for finding and processing data redundancy; this type of algorithm does not carry the risk of data loss in large environments that a hashing algorithm does.

Multiple configuration options to meet different needs

IBM offers flexible disk-based storage options in multiple configurations that can be optimized for performance and high availability, and to meet specific disk storage needs. For example, in environments requiring higher availability, and/or higher performance, a clustered configuration can be deployed to provide hardware redundancy in the event of a node failure, enabling the continuation of backup and restore operations. Beyond redundancy, a cluster configuration doubles the inline de-duplication performance throughput while keeping one repository that is accessible through any one of the two clustered nodes.

A mature product with a record of proven performance

Given the newness and complexity of data de-duplication, choosing a proven product is important. This can be evaluated based on factors such as time of production deployment and number of customers in production. IBM ProtecTIER De-duplication solutions have been in place in Fortune 500 data centers since 2005 and are today deployed by companies worldwide.

Feature—case study

Case Study: KBR

KBR, a US-based global engineering and construction company with 70,000 employees worldwide, needed to replace its existing tape backup infrastructure with a new solution. The solution had to enable KBR to back up 100 TB of production data on a weekly basis at four sites.

Solution infrastructure requirements

The sites' existing backup infrastructure was based on TSM and LTO-1 and LTO-2 tape technology. The new infrastructure was originally to be based on global standards that included LTO-3 tape as the default standard for backup media. Instead, however, it was determined that for greater cost-efficiency, a new architecture would be used that could exploit multiple types of backup media including disk. In this architecture:

- *Physical tape would be used for "fast" clients (in some cases, as SAN Media Servers), while virtual tape on disk would be the target for large numbers of slow streams.*
- *Data de-duplication would be exploited to allow cost-effective storage of all Primary Virtual Tape backup images for the full retention period.*
- *NetBackup Vaulting would be used to provide transparent duplication of all physical and virtual backup images to physical images.*

Criteria for solution selection

The company evaluated several VTL solutions based primarily on the following criteria.

- *In-line data de-duplication: To minimize any impact of data de-duplication on existing operations, KBR wanted a solution that used in-line rather than post-process data de-duplication. With in-line data de-duplication, vaulting to physical media could be done immediately upon completion of backup, with no post-processing required.*
- *Performance: KBR needed a solution with the adaptability to handle both relatively slow streams of backup data and streams driving tape at 80 MB/sec or more.*
- *Non-disruptive deployment: The solution would have to be capable of being deployed with existing hardware vendors and systems and integrated with existing disk management software.*
- *Non-disruptive operations: The solution had to deliver high availability and stability with minimal risk of downtime.*

- *Scalability: The solution had to scale in capacity up to 1 PB of storage to meet both the company's needs today (100 TB of data) and its plans for the future. The immediate goal was to have 4 GB describe 1 PB of backup images and find similarity for de-duplication.*
- *Open Store VTL: Open store was critical to enable all media management to stay under NetBackup control and to ensure that there would be no issues with moving backup images between the virtual and physical media.*

Benefits of the IBM solution

KBR selected an IBM System Storage ProtecTIER De-duplication solution, implemented as a software solution on existing KBR dual quad-core server hardware, switches, and HBAs. The solution uses in-line data de-duplication, is scalable to 1 PB of data, and provides the required adaptability to handle both a large number of slower streams and streams driving tape at 80 MB/sec or more. KBR chose a specific HA (high availability) solution, with a standby ProtecTIER node configured to allow fast fail-over, helping to minimize the risk of downtime in the KBR's business-critical data environment. System stability in production has proven robust since KBR's deployment of the solution in early 2007. This HA configuration is now achievable in an active- dual ProtecTIER server mode with the available Cluster option of the new IBM TS7650G Gateway.

Using a VTL solution with in-line data de-duplication has benefited KBR by providing the company with a new type of backup media whose unique attributes complement physical tape to provide a superior solution to the originally planned physical tape-only solution. The IBM deployment has enabled the company to retain a year's worth of data on disk, significantly improving its restore times over physical tape-only backup.

For more information

To learn more about how IBM data de-duplication solutions can help your organization significantly reduce disk requirements for data storage, contact your IBM representative or IBM Business Partner, or visit **ibm.com**/storage

IBM Systems and Technology Group
Route 100
Somers, NY 10589
U.S.A.

Produced in the United States of America
August 2008

This document could include technical inaccuracies or typographical errors. IBM may make changes, improvements or alterations to the products, programs and services described in this document, including termination of such products, programs and services, at any time and without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. The information contained in this document is current as of the initial date of publication only and is subject to change without notice. IBM shall have no responsibility to update such information.

IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein. Performance data for IBM and non-IBM products and services contained in this document was derived under specific operating and environmental conditions. The actual results obtained by any party implementing such products or services will depend on a large number of factors specific to such party's operating environment and may vary significantly. IBM makes no representation that these results can be expected or obtained in any implementation of any such products or services.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS-IS" WITHOUT ANY WARRANTY, EITHER EXPRESSED OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided.

References in this document to IBM products, programs or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM program or product in this document is not intended to state or imply that only that program may be used. Any functionally equivalent program or product that does not infringe IBM's intellectual property rights may be used instead. It is the user's responsibility to evaluate and verify the operation of any non-IBM product, program or service.

Disclaimer: The customer is responsible for ensuring compliance with legal requirements. It is the customer's sole responsibility to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the reader may have to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer is in compliance with any law or regulation.

[1] Enterprise Strategy Group, "Considerations for Enterprise-Scale VTLs," Data Protection Brief, March 2008

Recyclable, please recycle

TSB03010-USEN-00