



# Lab Validation

## Report

### **ExaGrid Systems** Disk-based Backup with Data De-duplication

**By Claude Bouffard**  
With Brian Garrett

**March, 2008**

## Table of Contents

Table of Contents .....	i
Introduction .....	1
ESG Lab Validation .....	3
<i>Ease of Deployment</i> .....	3
<i>Performance</i> .....	5
<i>Data De-Duplication</i> .....	8
<i>Remote Replication</i> .....	10
<i>Scalability</i> .....	12
ESG Lab Validation Highlights .....	14
Issues to Consider .....	14
ESG Lab's View .....	15
Appendix .....	16

### ESG Lab Reports

The goal of ESG Lab reports is to educate IT professionals about emerging technologies and products in the storage, data management and information security industries. ESG Lab reports are not meant to replace the evaluation process that should be conducted before making purchasing decisions, but rather to provide insight into these emerging technologies. Our objective is to go over some of the more valuable feature/functions of products, show how they can be used to solve real customer problems and identify any areas needing improvement. ESG Lab's expert third-party perspective is based on our own hands-on testing as well as on interviews with customers who use these products in production environments. This ESG Lab report was sponsored by ExaGrid Systems.

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.

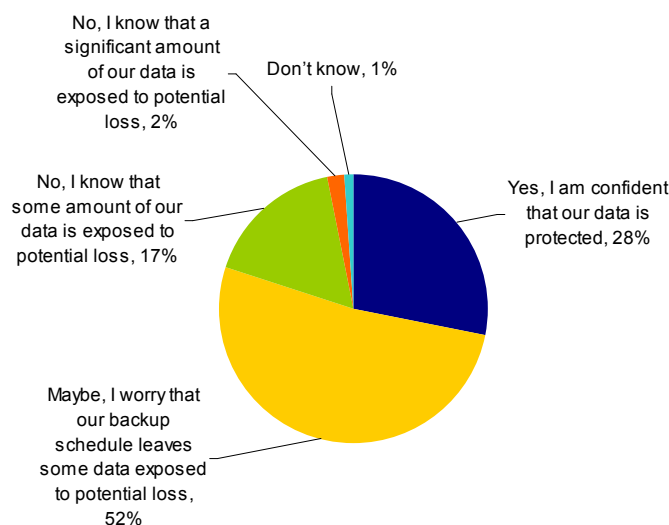
## Introduction

A growing number of organizations are embracing disk-based backup solutions to improve backup performance, eliminate tape media management issues and improve the speed and reliability of recovery operations. This ESG Lab report explores how the ExaGrid disk-based backup system uses byte level data de-duplication to provide space efficient capacity optimization, fast and reliable backup and recovery performance and WAN-optimized remote replication using a virtualized pool of network attached disk capacity that works with existing backup software, policies and procedures.

### Background

A growing number of IT managers are concerned that traditional backup methods aren't keeping up with the needs of the business. A recent ESG Research survey indicates that a majority of IT managers feel they can't provide their organizations with an adequate level of protection from data loss.<sup>1</sup> As shown in Figure 1, almost three quarters of respondents report some level of concern about their current data protection strategies. Most feel that the frequency of existing backup operations is inadequate. Slightly more than half worry that their current protection methods leave some of their data exposed to potential loss.

**FIGURE 1. DATA PROTECTION CONFIDENCE**



ESG Research also indicates that existing backup and recovery methods often fail. As a matter of fact, respondents report that 14% of backup and 17% of recovery attempts are unsuccessful. Many of these failures are due to the media management issues and human error associated with legacy tape backup methods.

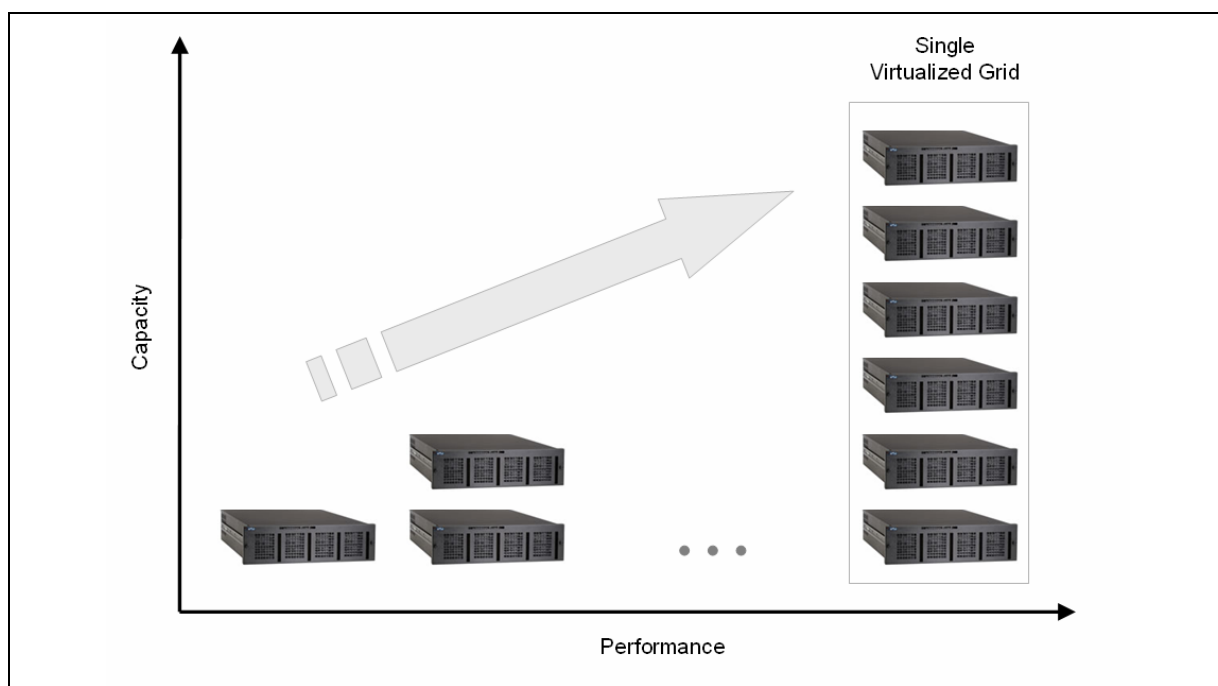
A growing number of organizations are turning to disk-based backup and recovery methods to improve the speed and reliability of backup and restore operations. While disk-based systems are inherently faster and more reliable than tape, they can be significantly more expensive. This cost concern is driving the adoption of data de-duplication technologies. Data de-duplication reduces backup to disk capacity requirements by eliminating the need to store copies of the same data over time. Eliminating extra copies reduces the capacity, and cost, of a disk-based backup and recovery solution. It can also be used to retain more backups on disk for quick and reliable recovery.

<sup>1</sup> ESG Research: *Data Protection Survey*, October, 2007; "In general, does the current frequency with which your organization backs up data provide an adequate level of protection from data loss?" (Percent of respondents, N=398)

## ExaGrid

ExaGrid offers five backup to disk models supporting from 1 to 5 TB of capacity per server. Managed from a single console, a single virtualized grid can be upgraded to support up to six servers and 30 TB of capacity per site. A 30 TB ExaGrid system can be used to store full backups for 30 TB of primary application data plus months of retention. As new servers are brought into the grid, each adds CPU horsepower, bandwidth, memory and disk drives for more speed and capacity.

**FIGURE 2. EXAGRID SCALABILITY**



ExaGrid works seamlessly with existing backup software and policies. Administrators simply point their backup software at a pool of ExaGrid disk capacity, which is presented as one or more shared file systems over an industry standard Ethernet network.

ExaGrid utilizes post-process data de-duplication, which occurs after backups complete. This background approach is designed to optimize the speed of backup and restore operations while providing capacity savings of up to 20 to 1—or more—over time. ExaGrid de-duplication can also be used to drastically reduce the amount of WAN bandwidth needed to make offsite copies of backup data

This report presents the results of ESG Lab's hands-on testing of the ExaGrid backup to disk solution with a focus on ease of deployment, performance, disk capacity optimization, WAN-efficient remote replication and scalability.

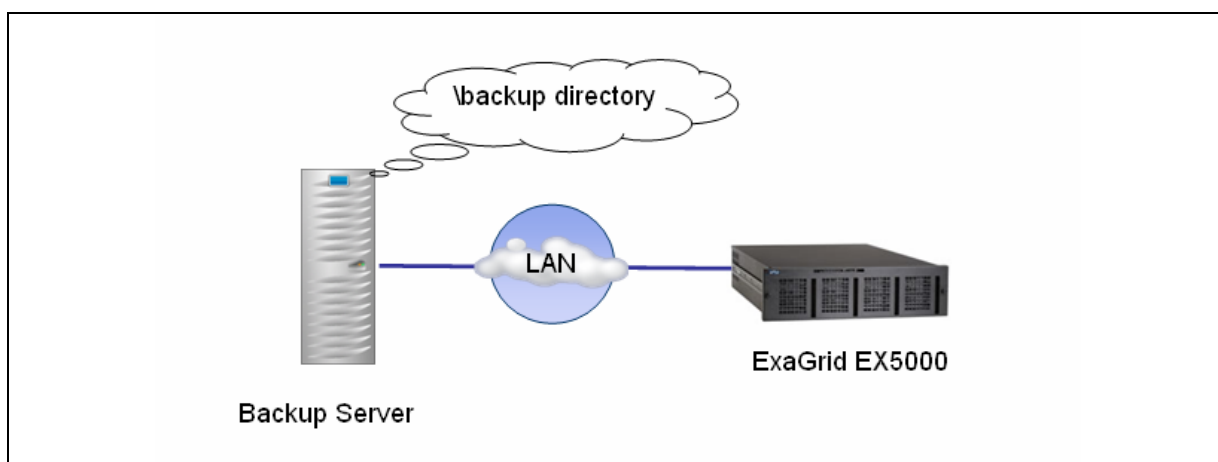
## ESG Lab Validation

ESG Lab evaluated the ExaGrid data protection solution during two days of hands-on testing at ExaGrid corporate headquarters in Westborough, Massachusetts. Testing began with an evaluation of its ease of integration with existing backup software.

### Ease of Deployment

ExaGrid presents a virtualized pool of disk capacity to backup applications as a networked attached directory (shown in Figure 3). Backup servers running Windows, Linux or UNIX operating systems access ExaGrid disk capacity using industry standard CIFS or NFS protocols. Much like a shared corporate directory that is accessed as a drive letter in a Windows environment, backup software uses network attached storage (NAS) accessed over an Ethernet LAN to access ExaGrid disk capacity.

**FIGURE 3. HOW EXAGRID WORKS WITH EXISTING BACKUP SOFTWARE AND POLICIES**



When an ExaGrid share is configured for use by a backup application, the administrator provides the name of the backup software that will be used. ExaGrid uses this information to optimize the operation and performance of backup and restore operations. This powerful capability is often referred to as content aware storage.

ExaGrid claims that systems can be configured “from box to backup in less than 20 minutes” in four easy steps:

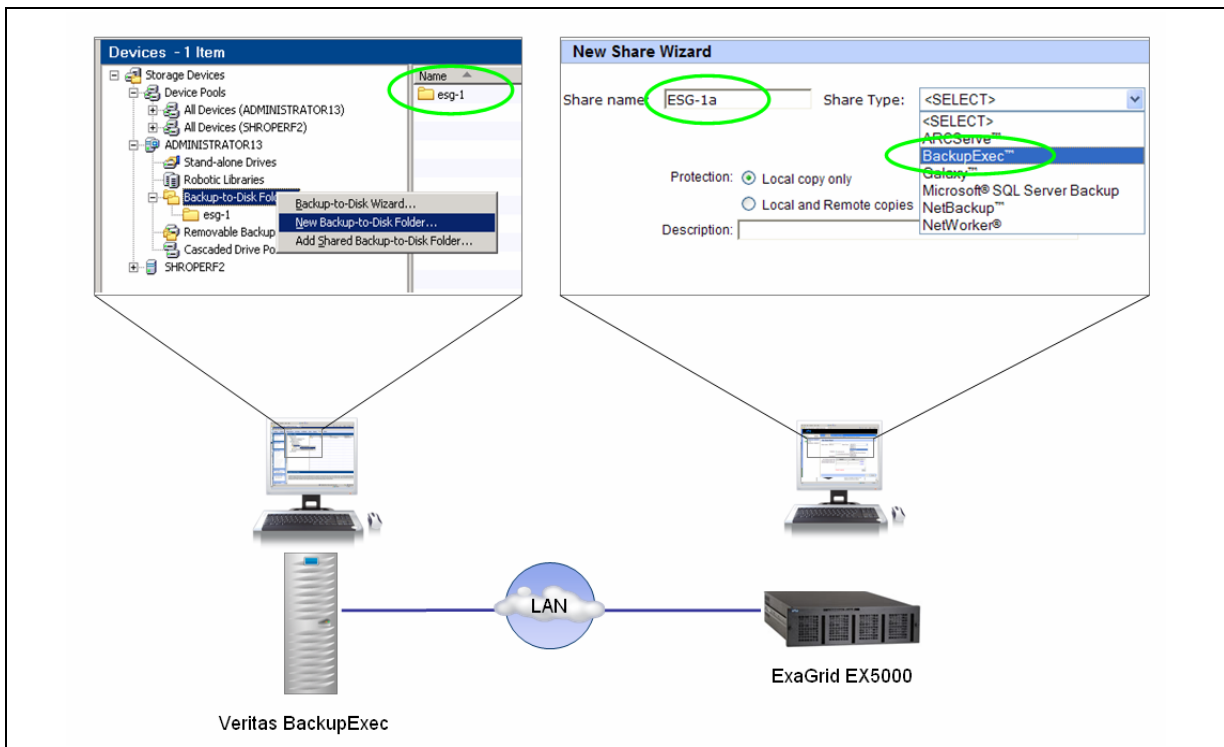
1. *Turn it on:* No loading of software needed; simply unpack it, rack it and power it on
2. *Point and click to create a share:* Shares are created using the ExaGrid console
3. *Define Shares:* Configure backup applications to use ExaGrid shares for disk-based backups
4. *Re-direct existing backup policies:* Configure existing backup policies to use an ExaGrid share

### ESG Lab Testing

ESG Lab started with a pre-racked and wired single server system. Configuration began with the assignment of basic networking settings, including an IP address for the ExaGrid server. The ExaGrid initialization wizard was accessed through a web browser. Thirteen minutes after beginning, the system was ready to be configured for the first backup job.

The web-based ExaGrid Manager console was used to create the first share (ESG-1a) as shown in Figure 5. A pulldown menu was used to select the backup application. With the share defined and ready for use, the next step was the configuration of a new backup-to-disk folder from the BackupExec console. An existing backup job was reconfigured to use the new folder and the first backup job was started.

**FIGURE 3. CONFIGURING AN EXAGRID BACKUP TO DISK FOLDER**



### Why This Matters

ESG Research indicates that integration with existing backup processes and ease of deployment are key concerns when considering a disk-based backup system. ESG Lab found that initial configuration of an ExaGrid system is simple and intuitive. Integration with existing BackupExec polices was straightforward. The first backup job was running sixteen minutes after beginning a configuration from scratch on a pre-wired system.

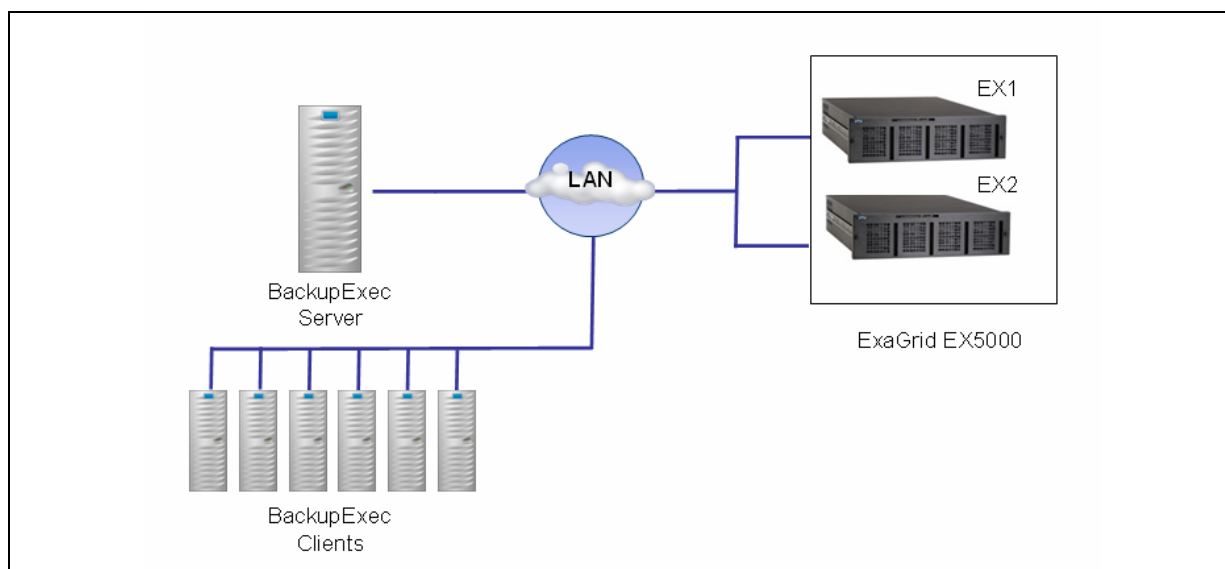
## Performance

ExaGrid systems are designed for enterprise-class scalability and performance. Post process data de-duplication ensures that backups and restores run fast. As ExaGrid servers are added to a singly managed grid, the aggregate bandwidth and performance of the system is increased.

### *ESG Lab Testing*

ESG Lab tested backup and restore performance using a single BackupExec Master server (Version 11d) and six BackupExec clients connected to a pair of ExaGrid EX5000 servers as shown in Figure 4. Three shares were created for each BackupExec client.<sup>2</sup>

**FIGURE 4. THE ESG LAB PERFORMANCE TEST BED**



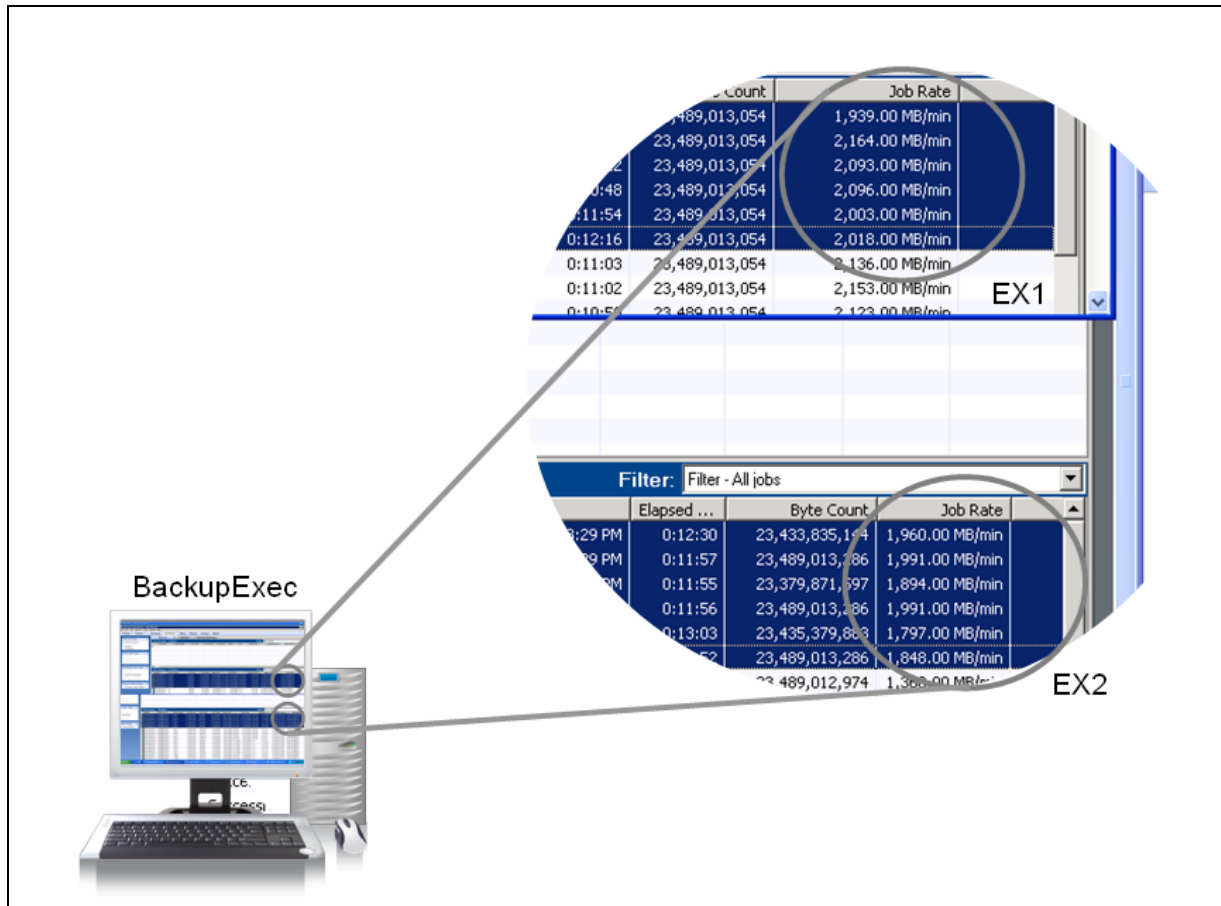
Testing began with six backup jobs running to one EX5000 server (EX1) followed by twelve backup jobs on two EX5000 servers (EX1 & EX2). Backup testing was followed by a series of restore tests using the same set of shares.

It should be noted that ESG restore tests were performed immediately after backup tests had completed. As a result, restores came from full copies of backup data that were not de-duplicated. This is a key design goal of the ExaGrid post-processing approach to data de-duplication. Backup data arriving at an ExaGrid server is retained in full form on disk to optimize the speed of restores. Over time, full form backup data is rotated, causing restore requests for data backed up days, weeks, or months ago to come from the capacity optimized de-duplication pool. Since most restore request are for data that is no more than a day or two old, this approach provides the best of both worlds—the speed and reliability of full-form disk-based images for recent restore requests and the savings of data de-duplication for older requests.

<sup>2</sup> See the appendix for additional configuration details

Performance results were collected from BackupExec job reports as shown in Figure 5. In this example, six jobs running on each BackupExec server were exercising a two server ExaGrid system (EX1 & EX2). Final performance numbers are the sum of the per backup job performance witnessed by ESG Lab.

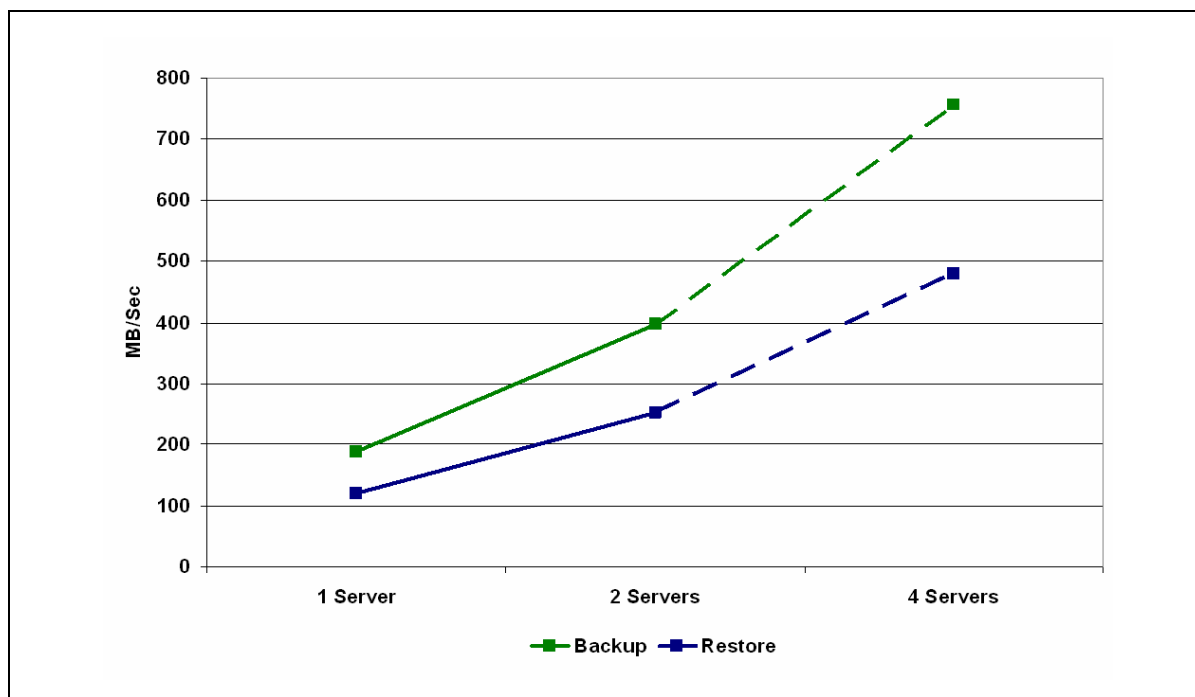
**FIGURE 5. BACKUPEXEC PERFORMANCE RESULTS**



The results of performance testing using two servers was compared to results collected previously by ExaGrid using a maximum configuration of four ExaGrid servers. An audit of the ExaGrid BackupExec logs for the two server configuration revealed that the performance recorded by ExaGrid was virtually identical to the results measured by ESG Lab.

Results are shown in Figure 6 with throughput recorded by ESG Lab shown as a solid line. The dotted line shows the audited results for a four server configuration.

**FIGURE 6. EXAGRID PERFORMANCE SCALABILITY**



#### *What the Numbers Mean*

- A single EX5000 server delivers up to 188 MB/Sec of aggregate backup throughput
- As audited by ESG Lab, performance scales up to 750 MB/sec for four servers.
- Supporting up to six controllers, an ExaGrid system delivers up to 1120 MB/sec (4TB/hr) of backup performance.
- A sustained aggregate backup performance rate of 4 TB/hr can be used to protect up to 32 TB of data in a single eight hour shift.
- Post-process de-duplication has no perceivable impact on performance

### **Why This Matters**

ESG Research indicates that performance is the top priority for data center managers considering a disk-based backup solution. ExaGrid uses a scalable post-process de-duplication approach for optimal backup and recovery performance. As an organization's data grows, the ExaGrid approach maintains day-one backup and restore performance. Using real-world data and a real-world backup application, ESG Lab has confirmed that ExaGrid can backup data at rate of 4 TB/hr for a fully configured grid.

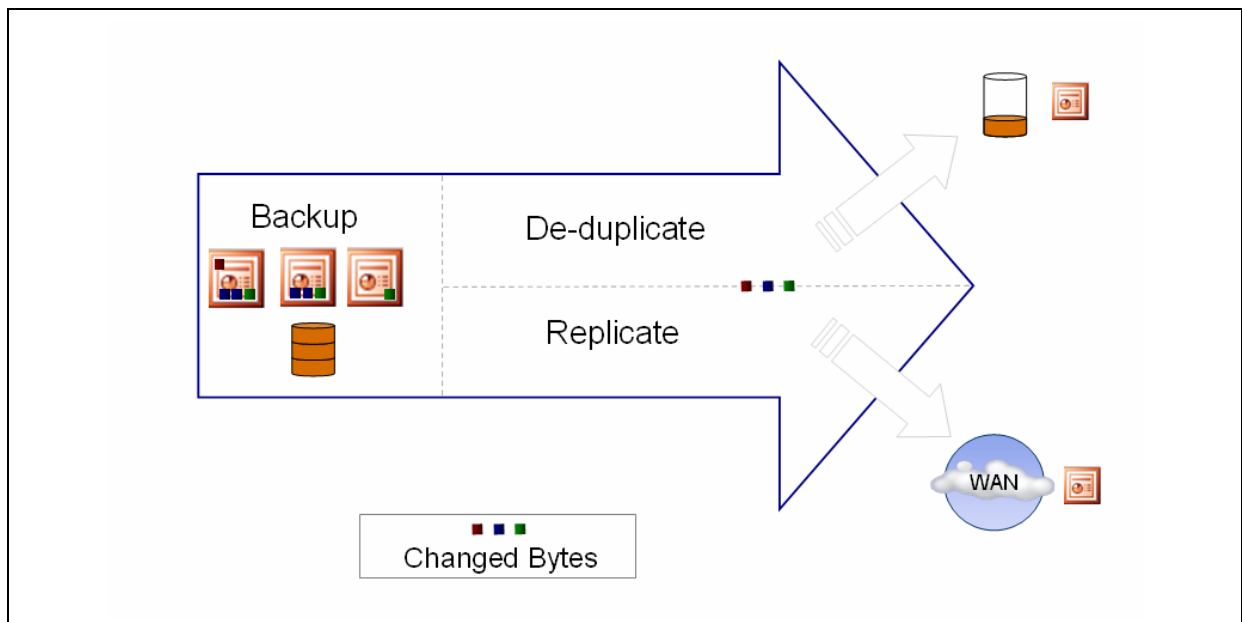
## Data De-Duplication

ExaGrid uses a byte-level post-process de-duplication approach to reduce disk capacity and optimize performance. Data de-duplication is the process of examining data to identify and eliminate redundancy. With the same data backed up over and over again, de-duplication can be used to drastically reduce backup to disk capacity. With ExaGrid byte-level data de-duplication, organizations can typically store 20 backups on disk in the space it would take to store one backup without data de-duplication.

ExaGrid byte-level data de-duplication runs as a post-processing task in parallel with remote replication as shown in Figure 7. Backup data streams into ExaGrid servers as shown on the left. In this example, subsequent backups of a Microsoft PowerPoint presentation are depicted. Byte-level modifications to the presentation that have occurred between backups jobs are shown as colored blocks. From the backup software's perspective, the PowerPoint file has changed at the time of each backup, so the entire presentation is being sent to ExaGrid.

After the backup job has completed, the de-duplication process begins. Note that ExaGrid byte-level de-duplication not only eliminates the need to store and send the PowerPoint file three times, it reduces the amount of disk capacity and WAN bandwidth required to only the changed bytes. Replication to a remote site runs in parallel with de-duplication. Because the data is de-duplicated, it uses a fraction of the WAN bandwidth normally required to make offsite copies. In this case, a single copy of the PowerPoint file, along with the information required to restore any of the last three versions, is logically stored at the remote site with only a small number of bytes sent over the WAN.

**FIGURE 7. DATA DE-DUPLICATION PROCESS**

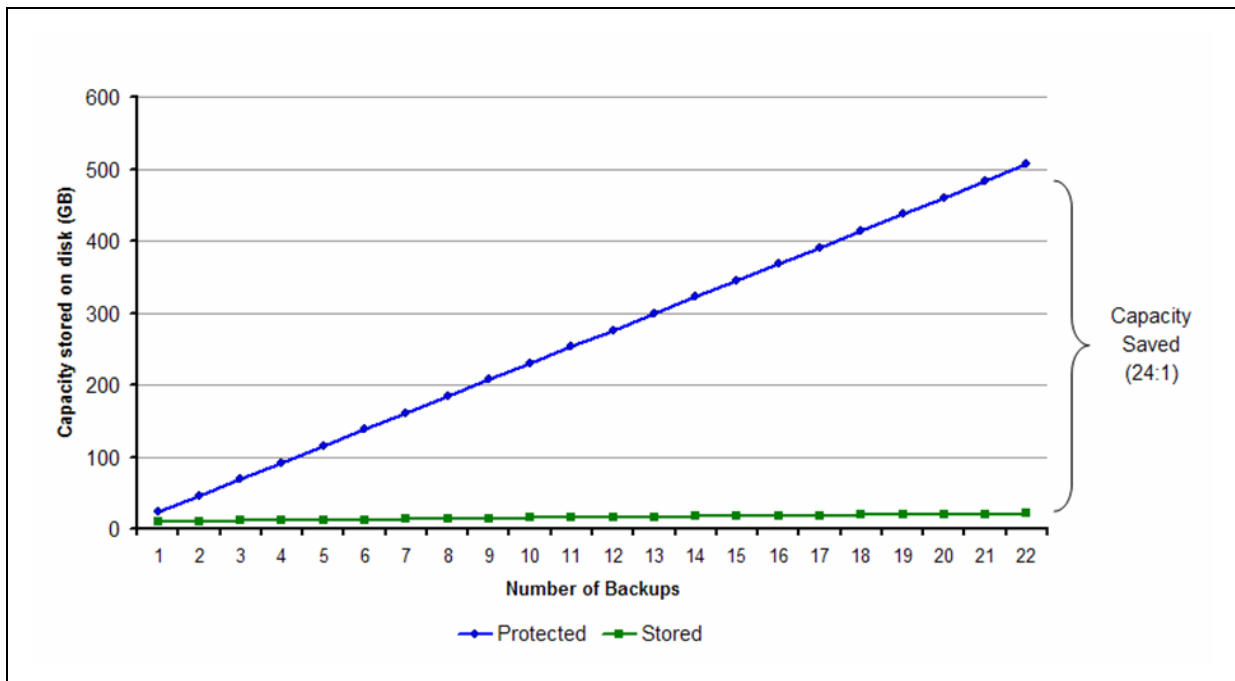


### ESG Lab Testing

ESG Lab tested ExaGrid data de-duplication using a 42 GB dataset composed of common business file types (Word, Excel, PowerPoint, etc). A series of full backups were performed using BackupExec software running on a single backup server. One percent of the data was modified to simulate a day of activity before each subsequent backup was started.

The ExaGrid console was used to monitor the de-duplication process. Immediately after the first backup had completed, ExaGrid software compression had reduced the 42 GB data set to 22.39 GB. After waiting for post-process de-duplication to complete, de-duplication had reduced required capacity to only 373 MB. This process was repeated after introducing a 1% change rate to simulate a day of activity. Results were compared against results collected by ExaGrid for a series of 22 full backups. The capacity savings over time are shown in Figure 8. In this example, ExaGrid data de-duplication reduced disk capacity by a factor of 24 to one over a simulated three week retention period.

**FIGURE 8. THE POWER OF DATA DE-DUPLICATION**



ESG Lab reviewed actual customer data de-duplication statistics harvested from more than 200 customers with over 400 ExaGrid systems installed. De-duplication rates of 34:1, 22:1, 20:1 and 12:1 were observed. ExaGrid indicates that the average de-duplication rate for their customers is 20:1 and that backups are being retained on disk for an average of 12.5 weeks.

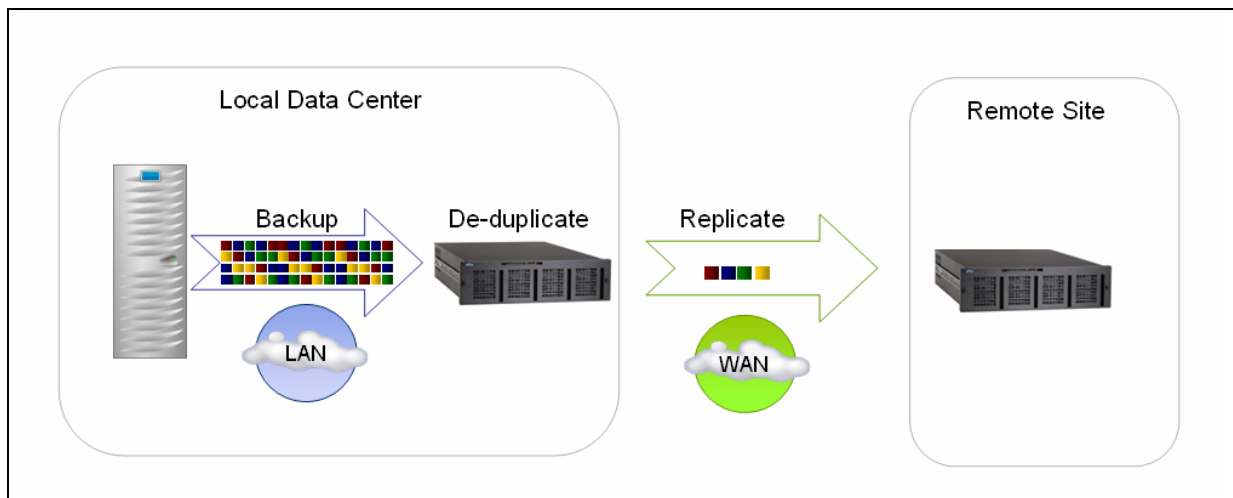
## Why This Matters

ESG Research indicates that cost is the leading obstacle to disk-based backup deployment. Data de-duplication changes the economics of backup to disk by reducing the cost of data retained on disk. ESG Lab testing and feedback from ExaGrid customers indicate that de-duplication can be used to reduce disk capacity by a factor of twenty to one or more depending on the type of data being backed up, the backup policies in use and the number of backups retained on disk.

## Remote Replication

ExaGrid's architecture has the ability to utilize a secondary ExaGrid system at a remote location for offsite protection. ExaGrid performs parallel de-duplication and replication as a post-process for optimal backup and recovery performance. As shown in Figure 9, backup data moves at full speed into an ExaGrid server at a local data center. Post-process de-duplication not only reduces disk capacity requirements, but also reduces the amount of data sent over the WAN.

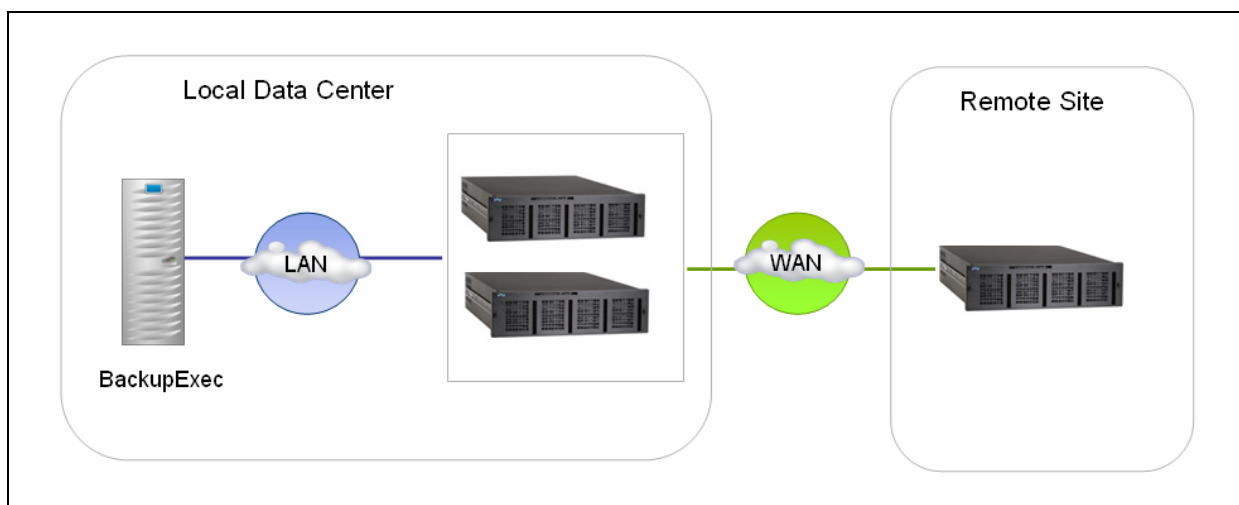
**FIGURE 9. WAN OPTIMIZED REMOTE REPLICATION**



### ESG Lab Testing

ESG Lab configured a third ExaGrid server at a simulated remote site as shown in Figure 10. A local LAN was used to emulate four T1 lines between the local data center and the remote site.

**FIGURE 10. REMOTE REPLICATION TEST BED**



Backups were conducted at the local site before and after the remote replication process had run. No performance difference was detected.

Once the data was replicated to the remote site, ESG Lab used the “Test Recover” function from the ExaGrid console to test the recoverability of the backup data replicated to the remote site. The test recover function, used by ExaGrid customers to audit their internal disaster recovery procedures, creates a remote share that is fully operational. The test recover share allows the data to be read or restored just like the share from which it originated. The test Recovery share that ESG Lab used was built from the last fully de-duplicated backup.

BackupExec was used at the remote site to perform a recovery operation using the test recover share. Comparison of the recovered data showed that the files were identical to the original data at the local site.

### **Why This Matters**

Offsite copies of backup data are needed to ensure that an organization can recover from a disaster. Organizations with large amounts of backup data can't afford to make offsite copies of backup data electronically due to the high cost of WAN bandwidth.

ESG Lab has verified that ExaGrid data de-duplication significantly reduces the amount of data that needs to be transferred over the WAN, enabling what may have otherwise been impossible or prohibitively costly. ESG Lab has also confirmed that ExaGrid remote replication has no impact on backup performance.

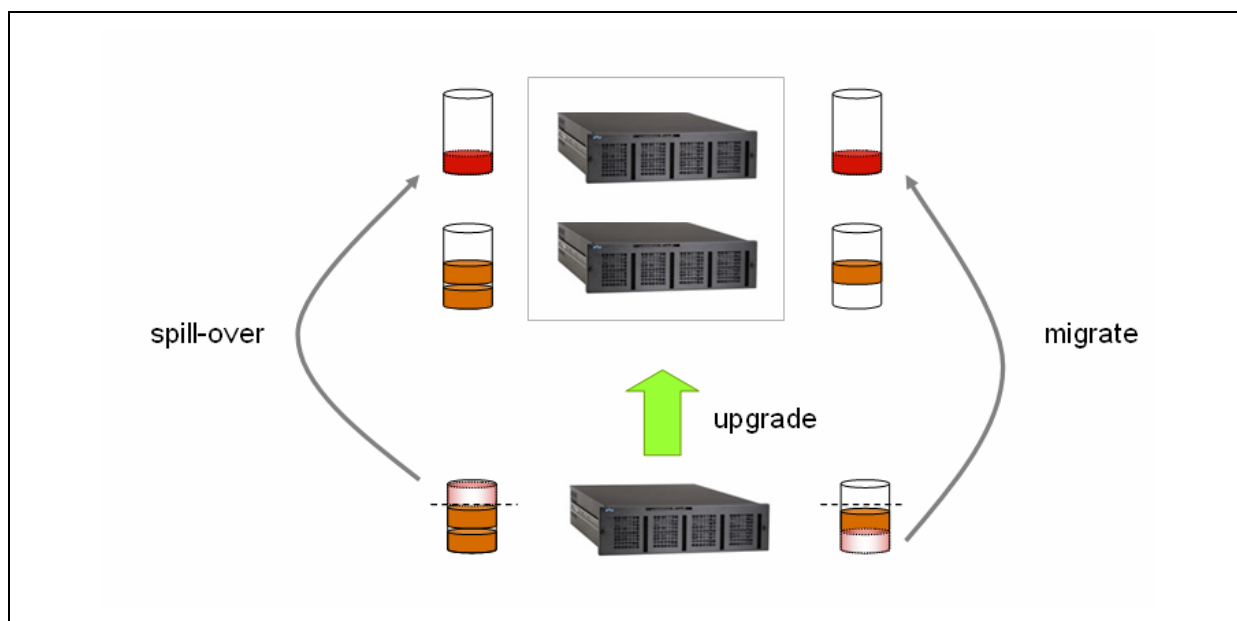
## Scalability

The scalable ExaGrid architecture provides plug and play growth as new systems are virtualized together into a single pool of backup to disk capacity. Performance scales with data growth since processing power and memory are added together with storage capacity. To help balance backup load, ExaGrid allows shares to be moved from one server to another within a grid.

Figure 11 illustrates two methods which can be used to load balance shares across all the servers in a virtualized grid. In this example, a single server system has been upgraded with the addition of a new ExaGrid server. Spill-over triggers automatically according to a user defined capacity threshold. When the threshold has been reached, share data spills over into the new ExaGrid server automatically. Manual migration is also supported for load balancing and maintenance. In this example, one share has been migrated to the new ExaGrid server.

Up to six servers per site can work together within a virtualized grid and can be managed from a single ExaGrid console. In a two-site configuration, up to 12 servers can be managed from the same console. Except for the spill-over case described above, shares are typically configured to reside within a single ExaGrid server. Shares are not striped across ExaGrid servers.

**FIGURE 11. SCALABLE UPGRADES**



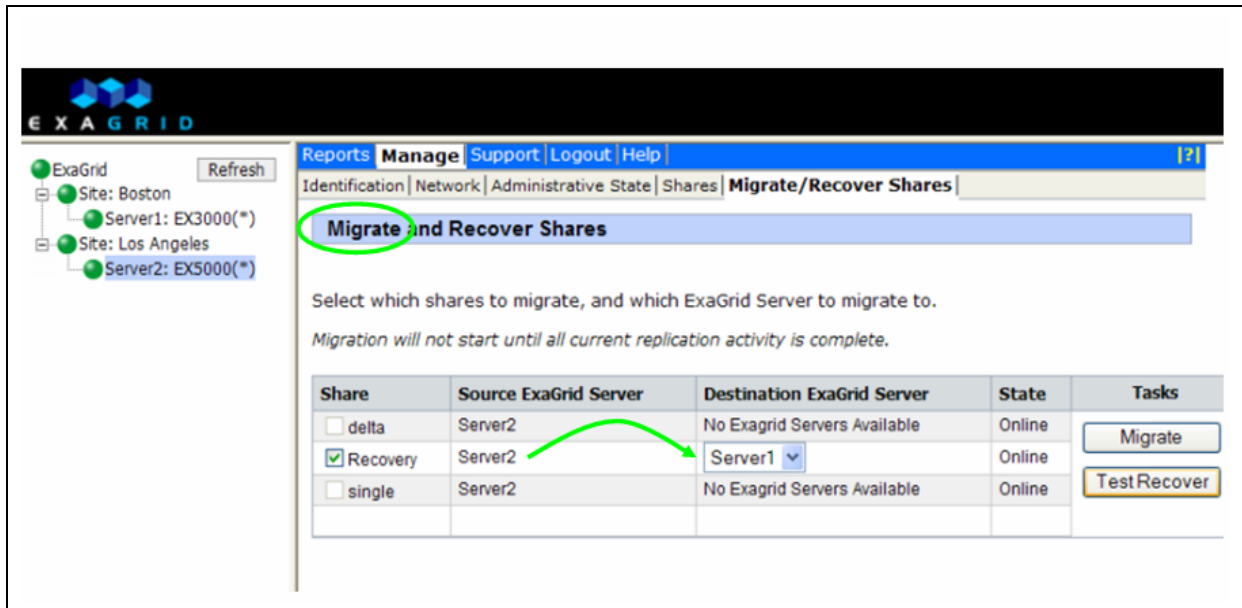
### *ESG Lab Testing*

ESG Lab upgraded from a single server to a two server configuration during the validation. The web-driven process was intuitive and straightforward.

Automatic spillover of an artificially full share was tested. Using BackupExec, backups jobs were directed to the share. ESG Lab observed that the backup job performed seamlessly and that new data had spilled over to the new server. Recoveries using share data that spanned two servers completed without error.

A manual migration of a share from ExaGrid server 2 to server 1 was initiated as shown in Figure 12. ESG Lab observed that the data residing on the share being moved was on-line and available for backup and restore operations during the migration. Data was moved over the private back-end Ethernet between ExaGrid servers.

**FIGURE 12. MIGRATING A SHARE FROM EXAGRID SERVER 2 TO SERVER 1**



## Why This Matters

Growing backup to disk requirements can pose a number of challenges for backup administrators. Cost and complexity are getting out of hand as more and more disk arrays are needed to meet growing capacity and performance demands. Capital equipment and operating expenses are rising. Management complexity is rising exponentially. More and more space, power and cooling is needed in the data center.

Supporting up to six servers in a singly managed disk-based backup system (or 12 servers across two sites), ExaGrid scales to meet a wide variety of performance and capacity requirements. ESG Lab found that adding a new server to an existing grid is straightforward. Automatic spillover and migration provides the flexibility needed to grow, maintain and tune a growing pool of virtualized capacity. Taken together with the space savings provided by de-duplication, the scalability of a singly managed ExaGrid system reduces the cost and complexity of managing growing disk-based backup requirements.

## ESG Lab Validation Highlights

- ☑ An intuitive configuration wizard was used to perform a configuration from scratch. Configuring a BackupExec disk folder to use ExaGrid disk capacity was straightforward. The first backup was running 16 minutes after starting the configuration.
- ☑ Veritas BackupExec was used to measure a sustained aggregate backup throughput rate of 188 MB/sec for a single ExaGrid server. Two ExaGrid servers were measured at 396 MB/sec. Audited results for a four server system indicates that performance scales in a near linear fashion as servers are added up to 750 MB/sec with four servers.
- ☑ A sustained backup rate of 1.12 GB/sec (4TB/hr - 6 servers) for a fully configured ExaGrid system can be used to protect up to 32 TB of data in a single eight hour shift.
- ☑ Real-world file data (docs, spreadsheets, etc) was used to measure the effect of de-duplication over time. A simulated daily change rate of 1% and a daily full backup policy with a retention period of three weeks was used to show that 506 GB of protected application data can be retained using only 21 GB of disk. This test yielded a de-duplication rate of 24 to 1.
- ☑ A simulated remote site was configured for offsite replication of backup data. ESG Lab confirmed that replication and data de-duplication tasks run in parallel as a post-process with no impact on backup performance.
- ☑ A second server was added to an existing single server ExaGrid configuration. Automatic spillover from the existing server to the new server was observed. A manual migration of an existing share was performed to show that capacity and horsepower can be reallocated on demand.

## Issues to Consider

- ☑ The amount of disk capacity and WAN bandwidth that can be saved using ExaGrid de-duplication technology depends on a number of factors including the backup policies in effect and the number of backup generations retained on disk. Take for example, a series of daily full backups which have more duplicate data than the same number of weekly full, daily incremental backups. While de-duplication rates of 50 to 1 or more are possible when retaining months of daily full backups, ExaGrid customer field data indicates that ratios of 20 to 1 are more common.
- ☑ The ExaGrid post-processing de-duplication approach automatically reserves some capacity to land and store your current backup prior to data de-duplication. While the size of the reserve is equal to the size of your full backup, ExaGrid has priced its solution to be competitive with in-line solutions that do the de-duplication in-line before data is written to disk. Doing so levels the playing field between these two types of technologies on a pure cost basis.
- ☑ If you've enabled software compression within some of your backup software policies to reduce capacity, there is no need to continue doing so with ExaGrid. This highlights a subtle advantage of the ExaGrid approach to de-duplication. ExaGrid performs industry standard compression as backup data is ingested and before it is de-duplicated. Compression running on an ExaGrid server can be used to move this valuable, but CPU intensive task from backup servers to purpose-built ExaGrid hardware.

## ESG Lab's View

Due to the extreme capacity of backup data over time, tape has been the media of choice for decades due to its low cost. Over time, organizations have learned that the low cost of tape can lead to a rise in the overall costs to the business. Managing tape is a manual and error-prone process. As the number of tapes increases, so too does the complexity and management costs associated with handling, transporting and managing media.

Disk has always been easier and more reliable than tape, but until recently its high cost has made it untouchable for most organizations. The high energy costs associated with keeping disk spinning is also a concern. Data de-duplication technology which removes duplicate data before storing it to disk changes the economics of backing up to disk. With less disk capacity required to retain weeks or months of backups, data de-duplication drastically reduces the cost of backup to disk making it an approachable option for a growing number of organizations. As a matter of fact, a recent ESG Research survey indicates that 8% of organizations have already deployed data de-duplication technology and 25% plan on doing so.<sup>3</sup>

ESG Lab has confirmed that ExaGrid backup to disk solutions combine the benefits of high density SATA drives, post-process data de-duplication and scalable grid architecture to provide a cost-effective, energy-efficient alternative to tape. ESG Lab testing has confirmed that ExaGrid can backup data at rate of 667 GB/hr per server and 4 TB/hr (with 6 servers) for a fully configured grid. Real-world data and feedback from ExaGrid customers indicates that de-duplication can be used to reduce disk capacity by a factor of twenty to one. Applying the same savings to WAN bandwidth, ESG Lab has confirmed that ExaGrid remote replication is a cost-effective alternative to tape for offsite archival.

ESG Lab believes that organizations struggling with the cost, complexity and risk associated with tape backups would be wise to consider the bottom line savings that can be achieved with ExaGrid: faster backups, quicker and more reliable restores, lower risk, lower expenses (capital and operational) and last, but not least, a greener solution with optimized power and cooling.

---

<sup>3</sup> ESG Research: *Data Protection Survey*, October, 2007; "Please rate your usage of/interest in sub-file data de-duplication technology with respect to your organization's data protection processes," (Percent of respondents, N=372)

# Appendix

## Test Configuration Details

Hardware	Software
3 ExaGrid EX5000 backup to disk servers	ExaGrid software version 3.1.1
1 Dell PE-2850 backup servers	Windows 2003, SP2
6 Dell PE-2850 client servers	Symantec BackupExec 11d
2x dual core Xeon 3.0 GHz CPU, 4 GB RAM	
Cisco ASA5005 router	



20 Asylum Street  
Milford, MA 01757  
Tel: 508-482-0188  
Fax: 508-482-0218

[www.enterprisestrategygroup.com](http://www.enterprisestrategygroup.com)