ac

Self-Protecting Storage™

Simplifying Your Data Storage Infrastructure

Dave Therrien
CTO and Founder
ExaGrid Systems Inc.

Overview

Today, storing, protecting and managing your corporate business data is a complex, costly, error prone, and time consuming endeavor. This paper details the challenges that IT departments face in delivering reliable data storage services to their business users and applications. Then, a collection of emerging data storage and data management technology will be reviewed – technology that is instrumental in addressing the requirements of an ideal storage system. Finally, an innovative **Self-Protecting Storage™** system that leverages these emerging technologies will be described.

Top 8 Data Storage Challenges

IT departments have always been challenged to deliver increasing amounts of reliable, available data storage capacity. Today, for each and every data storage system that must be deployed, it takes a creative mix of multiple independent hardware storage components, software licenses and management tools to meet the budgetary limitations as well as the availability requirements of business units' applications.

Data storage and storage management wasn't always this fractured and complicated. In the days of mainframe storage, disk and tape subsystems were highly integrated with data management software that guaranteed reliable, available data storage. With open systems, this integrated data storage and data management environment was replaced with dozens of independent, often incompatible hardware components, software packages and management tools. The storage administrator today must be an expert at properly installing, configuring, monitoring and maintaining dozens of disparate vendors' products. They must be able to blend their selected mix of data storage, data protection and data management products into a workable, reliable storage system.

Here's a list of **8 specific data storage challenges** that IT administrators are facing today:

1. **Primary Storage Management** – Every time a filesystem runs out of available storage capacity, IT administrators must get involved in manually creating additional space. "Out of Space" capacity management remains a complex, manually intensive, error prone process that requires

immediate attention from IT administrators, regardless of the hour of the day.

2. **Primary Storage Utilization -** IT administrators must often make decisions on the specific placement of data across multiple storage subsystems and technologies with little to no information on the criticality of data to the business.
3. **Traditional Backup -** Backup software and backup systems are complex to install, configure and manage on a daily basis. Manual backup operations are time-consuming and error-prone. Costly backup infrastructures (servers, backup software, tape libraries, drives, and media) are complex, difficult to maintain, and must be expanded each year to accommodate the increasing amount of data that has to be backed up every weekend on dozens of new backup tapes.
4. **Traditional Restore –** Customers have reported tape-based restore success rates as low as 70%. Anything less than 100% restore success rates have never been acceptable to business users, but today, when regulatory auditors visit to request data that must be restored from tape, it's not good enough to restore just MOST of the data.
5. **Traditional Archiving –** When data is archived to tape, it's deleted from the servers to free up space for new data. This archive/delete model frustrates end-users and causes applications to fail when the "wrong" data is archived and deleted from servers. In addition, data on older generation tapes need to be regularly upgraded to newer tape media technology as older tape drives become obsolete. This is prohibitively expensive to remedy for companies with hundreds or thousands of older generation tapes.
6. **"Tapes in Trucks" for Disaster Recovery** - Most customers don't build out a disaster recovery site until a site disaster occurs, so when disaster strikes, it takes weeks just to purchase, deploy, and configure the recovery systems, storage, software and networking assets. Once this infrastructure is in place and operational, the process of restoring data from tapes that are brought in from an offsite tape storage facility begins. It could easily take days, and sometimes weeks to completely and successfully recover all of the data.
7. **Over-replication** - For each megabyte of user/application data that is created and modified, today's isolated data management products can easily create 10 to 20 megabytes of data in various replicated forms. RAID, snapshots, onsite backups, offsite backups, archive tapes, HSM tapes and offsite replication copies all contribute to the over-replication problem.

8. **In Tapes, we don't trust -** Storage administrators have lost their trust in successfully restoring data from backup magnetic tapes. And as these tapes age or are reused, they are less likely to reliably retain their data.

With so many data storage and data management issues, there needs to be some new thinking about how data storage systems should be architected in the future.

Emerging Storage Technologies

Fortunately, university and industry research in the field of data storage scalability, reliability, and availability is being adapted to address the issues that storage administrators are facing today. Many of these emerging storage technologies cannot just be layered onto today's existing data storage systems and data protection tools. It will take a different kind of storage architecture to bring these features and associated benefits to commercial use. Here are **10 emerging storage technologies** that will bring about dramatic improvements in reducing the cost and complexity of data storage and data management solutions while increasing the overall availability of data.

1. **Integrated Data Management** – Today's isolated data management products (backup, archiving, HSM, and replication) will be replaced by efficient, space conserving integrated data management products. These will maximize availability of business data and consume the minimal amount of storage capacity. These integrated data management products will be driven by a single, simple protection policy that replaces today's multiple, independent data protection management interfaces.
2. **Redundant Array of Inexpensive Servers (RAIS)** – This new storage architecture will replace today's independent primary storage and data protection hardware and software products with low-cost compute/network/storage server "bricks". Each brick will operate independently to reliably store and protect its own data and all bricks will work with each other to ensure overall availability across multiple inter-networked data centers.
3. **Location Independent Storage** – This is a key technology to providing high availability systems that can survive server, network or power failures. When a single "brick" fails, data can be re-replicated to surviving bricks without a loss of access by clients and applications.
4. **Grid computing** systems today are delivering distributed, scalable, high-performance, high-availability, processing power to business applications with heavy computational demands. Grid computers allow compute resources to be discovered automatically, to be serviced and upgraded

non-disruptively and to be shared by multiple applications to increase overall resource utilization rates. Now imagine a grid computer that is designed solely for providing primary storage, onsite and offsite backups, fast disaster recovery, long-term data preservation, and tiered storage along with all of the additional benefits of grid computing. And every time more storage capacity is added to the system, more processing power is added to deliver unprecedented scalability.

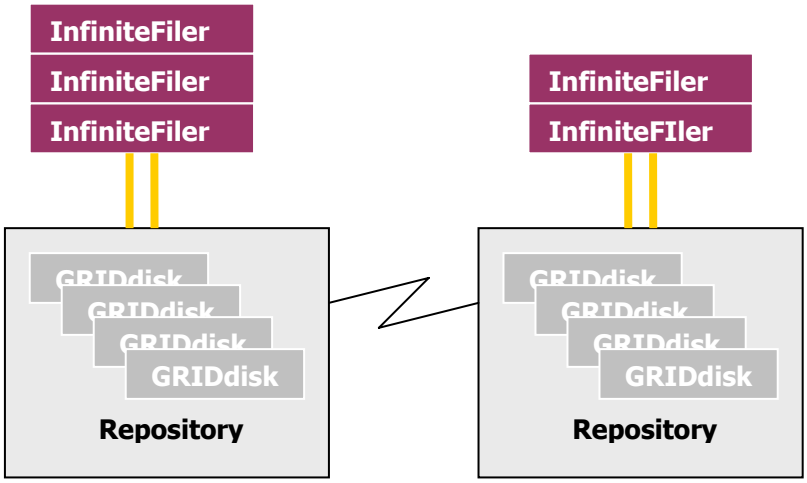
5. **Reverse Delta Compression** – This delivers constant "full-backup" grade restore performance from an "incremental-only" continual backup model. This technology consumes 40x to 1000x less capacity (and operator time) than today's backup processes and software.
6. **Content Naming –** Hashing codes like SHA-1 and MD5 provide a number of useful functions for underlying data management products. They can be used to uniquely name files, to automatically place files at specific destination nodes based on their content name, and to perform integrity checking and correction of all data on a continual basis.
7. **Version Chains** – These represent a concise packaging of the complete time-based lifecycle of a file from creation, through each modification to deletion. This space-efficient storage mechanism replaces wasteful weekly full tape backups where more than 90% of the backup content is unchanged from week to week, but still gets written to another set of backup tapes anyway.
8. **Delta-based WAN transfers** – In order to replace today's "tapes in trucks" model of offsite storage / disaster recovery management with MAN or WAN distributed disk-based repositories, smart "difference only" inter-site replication is a requirement since these MAN/WAN links have limited bandwidth compared to LAN bandwidth.
9. **Data integrity "scrubbers"** – Tens to hundreds of terabytes of backup and archive data can be checked AND corrected on a daily basis by "scrubbers" that run in parallel across tens to hundreds independent, intelligent storage bricks. This helps in providing ultra-reliable data restores compared with today's lower restore success rates.
10. **Hierarchical Storage Management (HSM) –** What if 90% of your company's inactive data could be automatically migrated to a lower cost tier of disk storage – from high performance FibreChannel or SCSI disk storage to lower cost SATA disk storage ? HSM is a technology that's been around since the heyday of the mainframe. The benefits of HSM are numerous.
    - HSM eliminates the manual processes associated with file systems filling up – imagine eliminating the manual processes of LUN allocation, volume expansion and filesystem expansion activities.

- With HSM, all files remain "visible" to applications, whereas with archiving, inactive files are deleted from servers once they are committed to tape.
- HSM knows which files are inactive, so it takes the operator guesswork out of determining which files to migrate to lower cost storage and which files should remain on high performance storage.
- HSM can actually help to accelerate disaster recovery times by allowing users to access their data as soon as the much smaller pointer files have been re-established.

Storage product vendors that can incorporate these emerging storage technologies into their storage strategy will more effectively be able to address the requirements of the ideal storage system. One such system is explained below.

Self-Protecting Storage™

The emerging storage technologies outlined above can be leveraged to create a Self-Protecting Storage™ system that integrates primary storage with complete onsite and offsite backup, automated data migration, site disaster recovery, and long-term data preservation.



A Self-Protecting Storage™ system provides the following features:

- **InfiniteFilers provide distributed NAS primary storage** for clients and applications that generate file data. **HSM** technology allows these NAS servers to effectively never run out of storage capacity.

- **Repositories** represent a virtual pool of scalable, disk-based storage capacity for storing InfiniteFiler backup data and for maintaining inactive InfiniteFiler data in a lower-cost tier of storage. Each repositories is comprised of 2 or more GRIDdisks that act like a distributed grid computer to deliver incrementally scalable, pay-as-you-grow, storage capacity. **GRID computing and RAIS** technologies allow the GRIDdisks of these repositories to be auto-discovered, and auto-configured in order to eliminate tedious manual storage allocation tasks.
- Data is **transparently migrated** between high performance InfiniteFiler disk storage and lower-cost Repository disk storage based on client or application access patterns. **HSM** incorporated into the InfiniteFilers provides this automated migration capability.
- Fast, continual, unattended, incremental-only **onsite & offsite backups. Version chains, incremental-only backups, delta compression and delta-based WAN compression** all contribute to delivering dramatic improvements in backup capacity consumption as well as backup execution times.
- **Fast and ultra-reliable restores** of any version of a file, or any directory of files from a previous point in time.
- Fast two-phase **site disaster recovery. HSM** and **delta-based WAN compression** help to reduce site disaster recovery time by 30x.
- **Self-healing** mechanism that checks and corrects its data continually. **RAIS technology, GRID-computing technology, data integrity scrubbers, content naming,** and **location independent storage** all contribute to a self-healing architecture.

The benefits of **Self-Protecting Storage™** include:

- **Reduced capital equipment costs**
  - Leverages low-cost commodity PC servers, high performance SCSI and low-cost SATA disk storage, gigabit Ethernet
  - Inactive data is automatically placed at a lower cost tier of storage
  - No tapes to purchase, no backup software, backup servers, tape library units or tape drives for backup
  - No more replication software and replication servers to purchase
  - No more archiving software and archive servers to purchase
- **Reduced operational costs**
  - An integrated approach to data storage and data protection reduces the complexity, eliminates the over-replication of data and increases the availability of corporate data.

- Simple capacity expansion – just add disks to the network – no primary storage LUN allocation, volume management or filesystem management
- No tapes to purchase, load and unload, ship offsite and request onsite
- No tape backup process hassles (tapes stuck in drives, no available tapes in the tape pool, …)
- No weekends tied up monitoring and managing backup job failures
- Simple point and click disaster recovery – completed within hours
- Self-correcting data delivers ultra-reliable restores
- No more having to guess about what data to archive from high performance storage to lower cost storage with fully automated data migration
- No more having to hunt down old tapes from an archive – all data is accessible all of the time.

- **Reduced service costs**
  - No more storing tapes offsite at a tape storage vault

If you're interested in learning more about how a Self-Protecting Storage™ system can reduce your storage cost and complexity while increasing your applications' data availability, please contact us at www.exagrid.com.

ExaGrid Systems Inc., 2000 West Park Drive, Westboro, MA 01581
Telephone : 508-898-2872 x226 Email : marketing@exagrid.com