

Why Deduplication Technology Is Causing a Paradigm Shift in Storage Tiering

A TIP Research Paper

THE INFO PRO
The Voice of the Customer.



datadomain

TheInfoPro (TIP) Research Paper delivers findings on over 289 in-depth interviews with Storage professionals at large enterprises, most of them among the Fortune 1000. A new TIP Storage Study is released every six months. The following research paper is based on findings from studies conducted since December 2004.

Massive Data Growth

Standard IT practice calls for keeping enough backups to recover from the last couple months of data change in case of human error, a virus, rippling errors in a database, or complete system failure. As a result, recovery storage can consume five to ten times more capacity than the primary storage it is protecting. Data growth means more capacity required to support multiple storage tiers – increasing management, cost, and complexity.

Issues With Tape

For more than 40 years, tape has been the only cost-effective option for storing massive amounts of backup and archive data. Experts say the odds of recovery from a given tape backup are about 90%. The intense physical logistics of the process, lack of reliability, vast amounts of media that need to be purchased over and over, and the drain on IT staff are all contributing factors that make tape a liability.

Backup storage and data movement should be simple, safe, automated, and online. It should use existing IT infrastructure and standard systems, software, and networks. In other words, it should be like every other part of the IT plan.

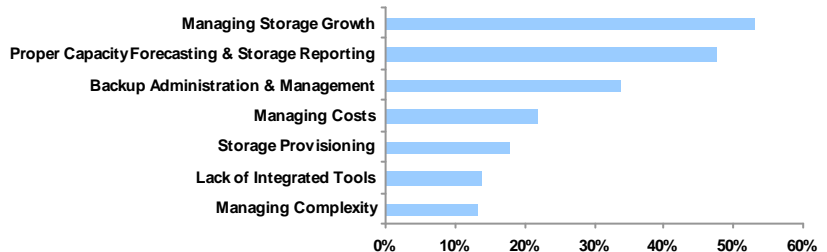
Minimizing Tape

By reducing the amount of storage required, deduplication enables disk to be a cost-effective alternative to tape. Data is available online and onsite for longer periods, and restores become fast and reliable. Storing only unique data on disk also means that data can be replicated to remote sites for network-efficient DR and consolidated tape operations.

Top Storage Team Challenges

In TheInfoPro's research, Storage professionals have consistently named managing storage growth, proper capacity reporting and planning, and backup management as the top pains facing their Storage organizations. The use of a storage target that identifies redundancy in incoming data streams and only saves those segments identified as unique, thereby conserving space, would clearly help with respect to the #1 pain, managing storage growth. In addition, since deduplication is the process of examining data for patterns, then identifying and storing only the unique data, it can help with respect to capacity planning – a key activity in capacity planning is the discovery of usage patterns. And finally, archiving duplicated content reduces the backup load, easing the strain of backup management.

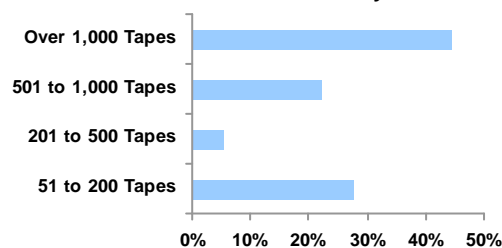
Chart 1: Top Storage Pain Points*



"Tapeless and deduplication technology are top areas for improvement. We are changing our dependency on tape for recovery to archiving for recovery."
 – Storage pro at a \$40B+ Financial Services firm

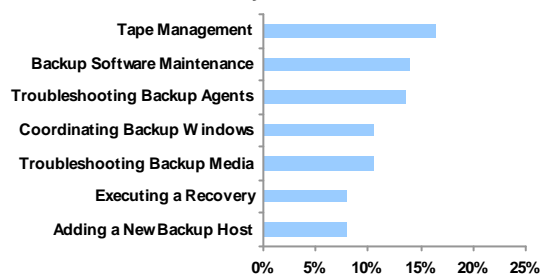
"We back up 100 to 150 TB a night, and we think deduplication will solve this pain." – Storage pro at a \$30B-\$40B Financial Services firm

Chart 2: Number of Tapes to Do a Complete Single Data Center Recovery*



Deduplication means longer onsite retention, and thus less reliance on tape. According to TheInfoPro's Wave 10 Storage research, almost 50% of Fortune 1000 Storage organizations require more than 1,000 tapes for a single data center recovery. More than 15% of a Backup team's time is focused on tape management, while close to 60% of Storage teams are backing up over 450 TB of content per month. As the pain of backup administration continues to escalate, deduplication technology maintains the promise of minimizing this pain by reducing necessary hardware, complexity, and manual effort.

Chart 3: Recovery Staff Time Allocation*



*Note difference in the different charts' scales

Why Deduplicate Data?

Eliminating redundant data can significantly shrink storage requirements and improve bandwidth efficiency. Enterprises typically store many versions of the same information. In the context of backup and nearline data, there is a great deal of duplicate data. The same data keeps getting stored over and over again, consuming a lot of unnecessary storage space (disk or tape), electricity (to power and cool the disk or tape drives), and bandwidth (for replication). This creates a chain of cost and resource inefficiencies within the organization. In addition, as data retention increases to satisfy regulatory and legal discovery mandates, the situation is exacerbated. Deduplication lowers storage costs since fewer disks are needed, and shortens backup / recovery times since there can be far less data to transfer.

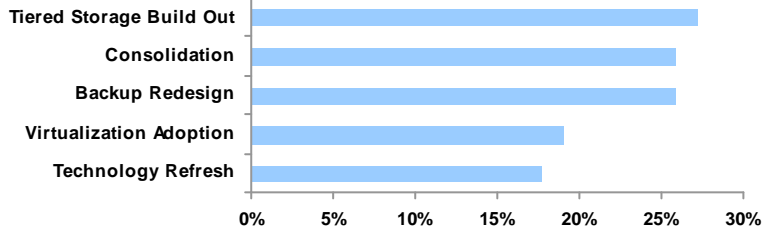
Deduplication Effects and Replication

The effect deduplication has on replication and disaster recovery windows and tape consolidation efforts can be profound. Deduplication means a lot less data needs transmission to keep the DR site up to date, so much less expensive WAN links may be used. For remote offices, reliance on tapes and physical tape transportation can be eliminated altogether.

Replication is fast since there is less data to send – only unique new backup or archive data is replicated between sites. For the most efficient time-to-DR, inline deduplication and replication of deduplicated data will yield the most aggressive and efficient results. In an inline deduplication approach, replication happens during the backup, significantly improving the time by which there is a complete restore point at the DR site.

Storage Initiatives

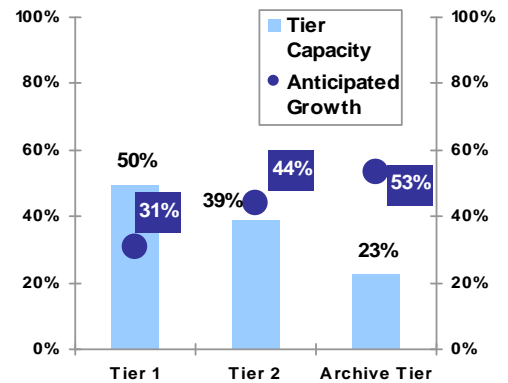
Chart 4: Top Storage Initiatives*



Deduplication and Tiering

One of the challenging aspects of deduplication is determining where to deploy the technology. Some end users talk about initially deploying out-of-band or archiving applications, while at the same time, others are quite excited about in-band or online deployments. The trend among the Fortune 1000 is to apply deduplication first to the archive. Not surprisingly, these archive deduplication tiers are projected to be the fastest-growing tiers for 2008. This growth helps contribute to the popularity of tiered storage build-out, the top Storage initiative.

Chart 5: Tier 1, Tier 2, and Archive Tier Capacity and Future Growth*



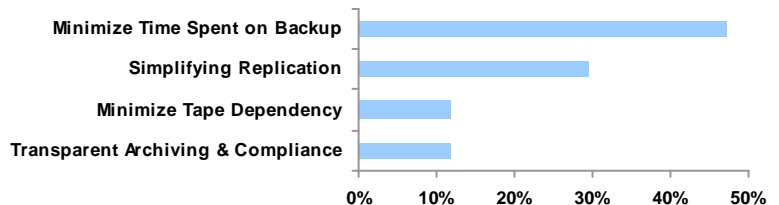
"We are very interested in deduplication. It will provide a new tier for recoverability." – Storage pro at a \$1B-\$5B Consumer Goods / Retail firm

Backup Redesign

The promise of new systems that actively identify and help manage storage growth, the possibility of minimizing the ever-increasing tape management nightmare, and the expansion of archiving and replication protection are key reasons why Storage teams cited backup redesign as a top Storage initiative.

When end users describe their backup redesign goals and motivations, they talk about consolidation and the desire to minimize their dependence on tape, while looking to simplify replication and reduce the necessary level of effort expended on backup. Deduplication technology fits squarely with these goals.

Chart 6: Backup Redesign Goals*



*Note difference in the different charts' scales

Deduplication Is a Storage Fundamental

The proliferation and preservation of many versions and copies of data propel much of the tremendous data growth most companies are experiencing. IT administrators are left to deal with the consequences. Because deduplication addresses one of the key elements of data growth, it should be at the heart of any data management strategy; it should be baked into the fundamental design of the system. Storage systems vendors who treat deduplication merely as a feature will check off a box on a feature list, but may not, in practice, deliver the benefits deduplication promises.

Data Domain Deduplication Storage

Data Domain has made it very easy by creating a fast, application-independent storage system (attachable as a file server over Ethernet, OST, or as a VTL over Fibre Channel). No client software or other configuration is required. As a result, Data Domain deduplication is transparent to the backup and recovery process. Data Domain systems can easily be used with various data movers and workloads, including non-backup data like email archives, reference data, and engineering revision libraries. More flexibility means that more consolidation is possible using less physical infrastructure, since all the redundancy across each of these data types is being eliminated by the same deduplication process. Deduplication effectiveness will of course be influenced by a number of factors, including how long the data is retained, how quickly it is changing between backup or archive events, and the data or application type.

State of Current Deduplication Environments

Over the last two years, deduplication has started to solidify its role in the data center. As mentioned on the previous pages, deduplication deployments have been targeted for email and semi-structured content with a high probability of duplication. In their product evaluations, end users have mentioned that products that sustain the highest duplication effectiveness are the most valued, as noted in Chart 9. But this does not mean that introducing deduplication can continue without any consideration for the impact on backup windows, recovery windows, and backup software integration, all of which follow closely behind deduplication compression / compaction effectiveness as the most important deduplication functionalities.

For the Storage teams that deployed deduplication technology in 2007, the average repository (compressed) is roughly 20 TB, and has an average effectiveness of 20:1. This represents about 400 TB of content, making the ROI and TCO justification pretty simple – so simple, in fact, that end users are starting to demand deduplication technology in file systems, document managers, email software, block storage arrays, and NAS. The span of 2008 and 2009 will clearly be an interesting time frame, one which will put a greater emphasis on storage arrays with intelligence, in addition to high capacity capability.

Chart 7: Where Deduplication Technology Should Reside – F1000, Midsize Enterprise, Europe Sample

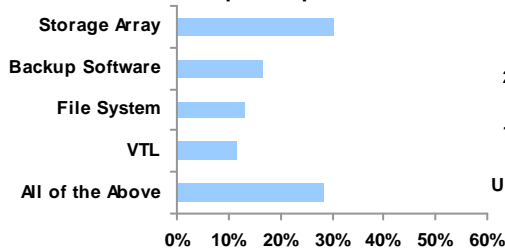


Chart 8: Size of Deduplication Repository (in TB) – F1000, Midsize Enterprise, Europe Sample

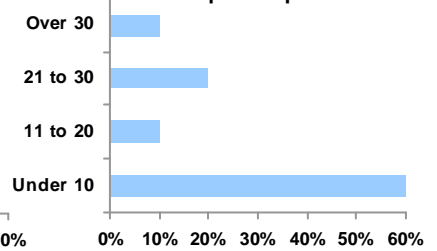
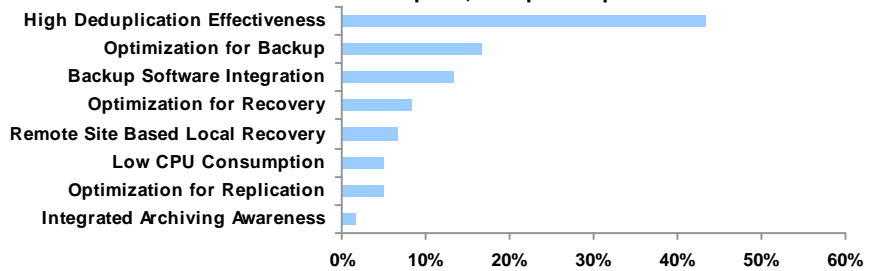


Chart 9: Most Important Deduplication Functionality – F1000, Midsize Enterprise, Europe Sample



“High deduplication is important since we want to remove tape. So the higher deduplication effectiveness, the better the ability to back up across a wide area network. We can get backup to a site where the data does not reside.”
 – Storage pro at \$5-\$10B Industrial / Manufacturing firm.

“Deduplication will be imperative for us. We have deployed it in our legacy products. I think the challenges are where the data is lost or corrupted. Throughput, maintaining bandwidth for backups, has been an issue. In the legacy space, scalability has not been an issue. The challenge is where we are going to do dedup. We need a hands-off solution for remote locations. We will need to do dedup on the server.” – Storage pro at a \$40B+ Telecom & Technology firm

Customer Benefits of Data Domain Deduplication Systems

High Capacity, High Throughput, and Green – A 16-controller DDX, using the DD580 controller, provides up to 12.8 TB/hour throughput and 8-20 PB of capacity, depending on backup policy and data change rate. With internal storage, it uses as little as 1.1 watts/TB of power and as little as 9U of a 19" rack space per petabyte.

Cost-effective Retention and Recovery – 20x-50x data reduction means more data can be stored onsite, increasing retention periods and improving data recoverability – providing disk storage at the price of tape.

Flexible DR Configuration – The DDX Series complements all Data Domain appliances, acting as a hub for recovery images vaulted efficiently from up to 320 smaller sites running Data Domain for DR and tape consolidation.

Ultimate Data Integrity – Data Domain's Data Invulnerability Architecture provides the best defense against data integrity issues. Continuous recovery verification, along with extra levels of data protection such as dual disk parity RAID (RAID 6), continuously detects and protects against data integrity issues during storage of backup data and throughout the lifecycle of the backup data.

Easy Integration Into Existing Environment – Data Domain systems work with all leading backup and archiving software, and easily integrate into the backup environment using NAS, OST, and / or virtual tape interfaces without any infrastructure change.

Field-proven – Data Domain has more deduplication customers than all other vendors combined. References are available for many applications, industries, and geographies.

Deduplication Technology Adoption Patterns

According to the most recent research, 15% of Storage organizations have deduplication technology already deployed, and 59% of Storage organizations are planning on deploying deduplication technology by the end of 2008. Additionally, of the 15% that deployed the technology in 2007, over half plan on expanding the deployment in 2008. Email archiving systems, disk-to-disk (D2D) snapshots, department file servers, and applications with high levels of duplicated content are the initial targets.

Chart 10: Deduplication Planned Adoption, Wave 8 through Wave 10

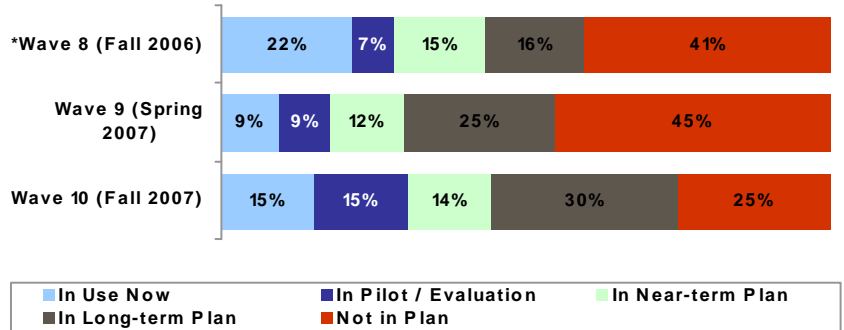
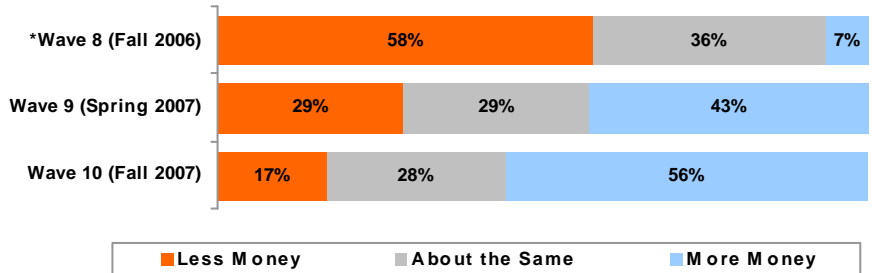


Chart 11: Deduplication In Use Spending Forecasts



Deduplication has sustained the #1 position on the TIP Storage Backup and Recovery Technology Heat Index® for two consecutive waves of research, with no signs of cooling off. This ongoing popularity shows a pattern similar to that of D2D adoption in 2003, where D2D maintained a top Heat Index position for four consecutive waves. Furthermore, deduplication is modernizing D2D with replication intelligence.

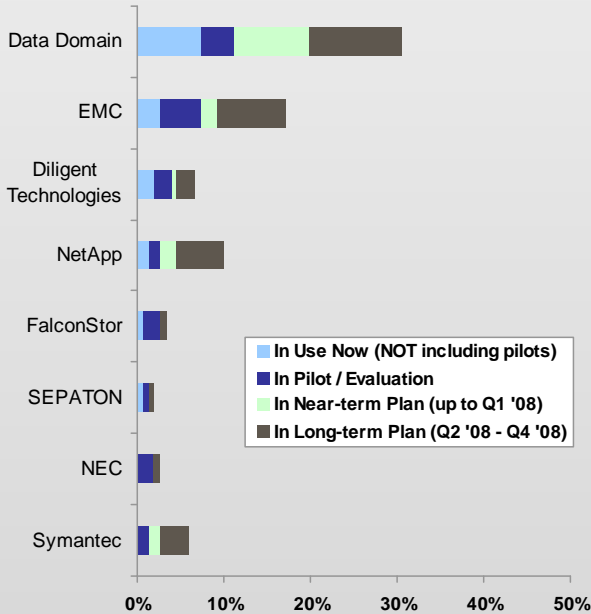
Chart 12: Deduplication Heat Index Growth

Technology	Wave 8 Rank	Wave 9 Rank	Wave 10 Rank
Deduplication*	14	1	1
Virtual Tape Library (VTL) for Open Systems	2	3	2
4 Gbps Fibre Channel	1	2	3
Remote Block Mirroring / Wide Area Replication (Async)	7	14	4

*Technology was previously categorized as De-Duplication / Capacity Optimized Storage / Single Backup Instance Store

Deduplication With Data Domain

Chart 13: Deduplication Roadmap Vendors, Wave 10



TIPNetwork Quotes on Data Domain

"Strengths are in the compression and the reliability has been great. They have exceeded on what they have promised. Great technical innovation."

– Storage pro at a \$5B-\$10B Industrial / Manufacturing firm

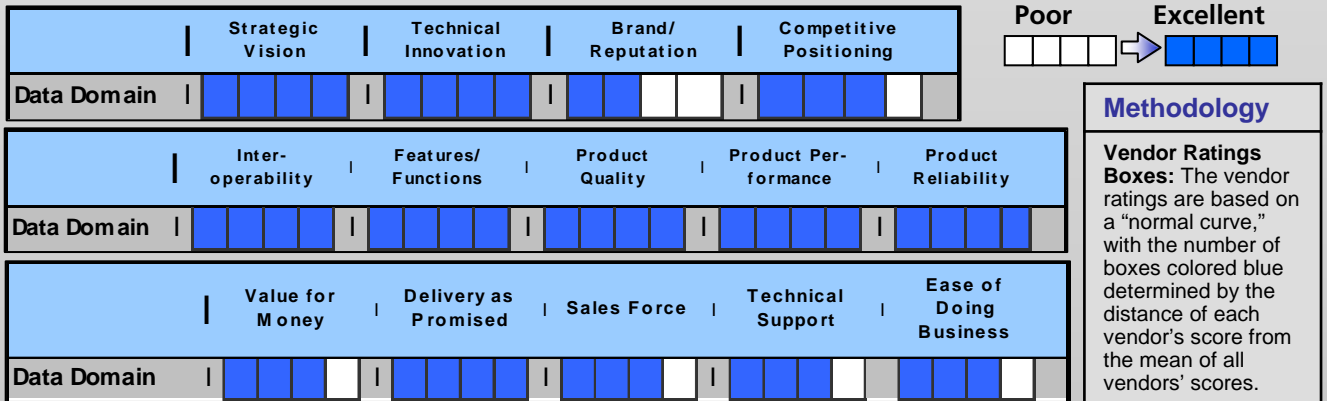
"This is the strongest dedup vendor on the market. They are the only vendor that could substantiate their performance claims. A lot of vendors talked the game, but did not have the product to back it."

– Storage pro at a \$5B-\$10B Industrial / Manufacturing firm

"This is one of those products that matches the glossy marketing materials. It is very fast and reliable, and does everything it says it does."

– Storage pro at a \$1B-\$5B Telecom & Technology firm

Chart 14: Data Domain Ratings



What Are Best Practices in Choosing a Deduplication Solution?

- Ensure ease of integration to existing environment.
- Get industry-specific customer references.
- Pilot the product / technology.
- Understand the vendor's roadmap.

Data Domain is the leading provider of deduplication storage systems for disk backup and network-based disaster recovery. Over 1,500 companies worldwide have purchased Data Domain's storage systems to reduce costs and simplify data management. Data Domain delivers the performance, reliability, and scalability to address the data protection needs of enterprises from the data center core to the remote offices. Data Domain products integrate into existing customer infrastructures and are compatible with leading enterprise backup and archiving software.