

White Paper

Optimizing File Services with Deduplication and Virtualization

Focus on F5 ARX and EMC Data Domain

By Terri McClure and Brian Babineau

July, 2010

This ESG White Paper was commissioned by F5 and EMC and is distributed under license from ESG.

Contents

Data Centers Drive Towards More Responsive and Efficient IT	3
Simplifying Administration and Reducing Cost with the ARX/Data Domain Solution.....	4
Putting the Solution to Work.....	4
Leveraging File Virtualization and Deduplication to Reduce Data Protection Costs.....	4
About Clackamas County.....	5
Automated Replication and Deduplication Reduce Costs <i>and</i> Improve DR Compliance.....	7
Automated Tiering puts Data Precisely Where it Belongs	8
Taking a Broader Market View	9
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.

Data Centers Drive Towards More Responsive and Efficient IT

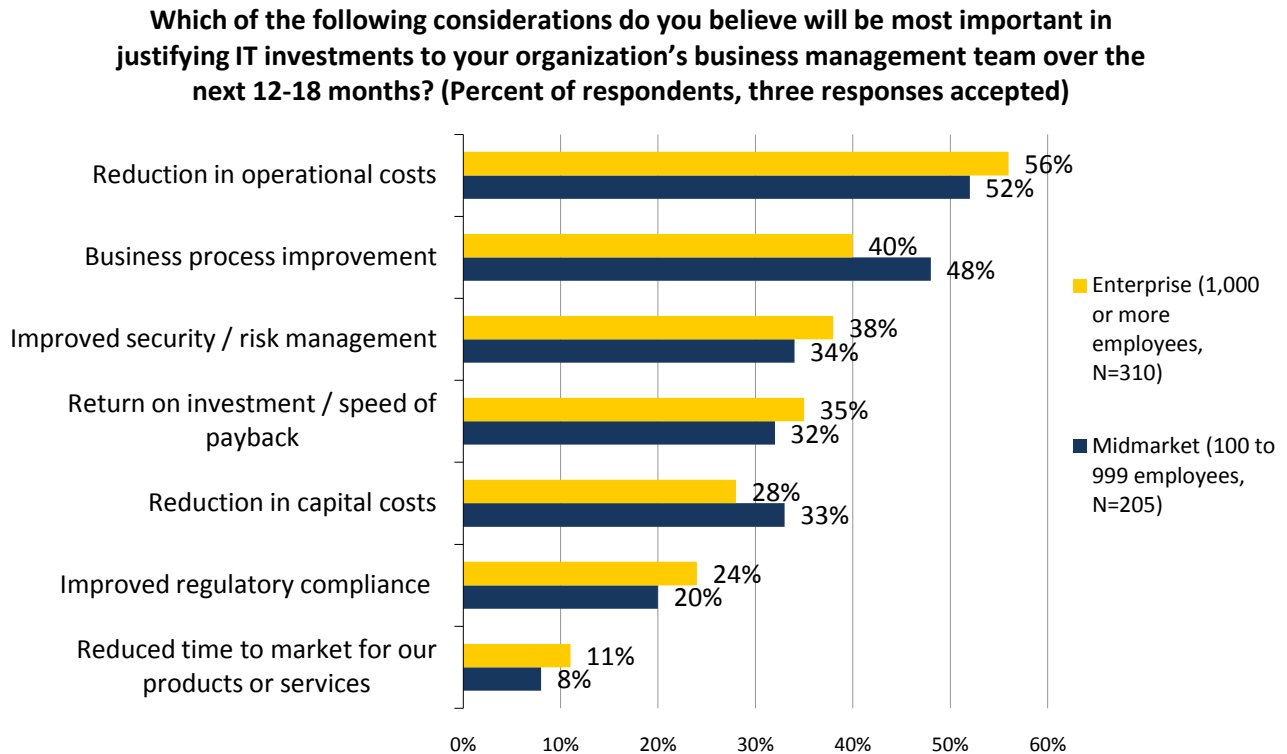
There is a tremendous amount of discussion in the media about virtualized IT and converged infrastructure. Organizations are in the process of transforming their data centers into environments that are able to better handle rapidly changing business requirements and unanticipated business needs. In order to meet these objectives, businesses are eliminating monolithic legacy data centers and infrastructures and replacing them with consolidated, highly virtualized, flexible environments that reduce costs and increase business responsiveness.

The data storage domain is the “low hanging fruit” when it comes to optimizing the IT environment, especially for fast growing file-based data which is expected to make up the bulk of data growth over the coming years. Bound by the limits of traditional network attached storage (NAS) storage systems, IT has been dealing with poor NAS utilization rates, system sprawl, and complex and wasteful file storage infrastructures. Users find themselves coping with five, six, or even seven or more copies of data to perform test and development work and meet service level agreements for data availability and recovery.

Companies like [F5](#) and [EMC Data Domain](#) are leading the way, providing solutions that provide elasticity and efficiency in the storage infrastructure and remove many of the management headaches we’ve come to know, helping users start down the path towards virtual IT.

Combining solutions from F5 and EMC enables IT to strip costs out of the data storage infrastructure by reducing the storage footprint and allowing IT to seamlessly migrate data to the appropriate storage tier, balancing price, performance, and availability. This strategy can help take significant operational costs out of the storage environment and it is all about operational costs today. Respondents to ESG’s 2010 IT spending survey ranked reduction in operational costs as the top justification criteria for IT spend for the next 18 months,¹ especially for enterprise IT which has a large legacy infrastructure to manage.

Figure 1. Reducing Operational Cost is a Key IT Spending Justification



Source: Enterprise Strategy Group, 2010.

¹ Source: ESG Research Report, [2010 IT Spending Intentions Survey](#), January 2010.

2010 is the second year in a row that operational cost reduction came out as the top justification for IT spending. Reducing capital costs was ranked as second most important in 2009, but dropped to fifth in 2010 as users look to improve business processes, security, and ROI. Spending seems to be recovering from the crush of 2009 and ESG research indicates it is doing so with an eye towards driving greater IT efficiency.

The balance of this paper will examine how the joint ARX/Data Domain solution can help IT seamlessly achieve greater flexibility and efficiency for unstructured file-based data. It will also take a look at how one user, Clackamas County, Oregon, is realizing the benefits of the solution.

Simplifying Administration and Reducing Cost with the ARX/Data Domain Solution

F5's ARX offers a robust set of data management policies that can move file data between storage tiers based on age, type, and size, amongst other criteria. ARX performs migrations transparently; there is no impact to users thanks to a global namespace that abstracts the file's network name from the physical file location. Customers can use the ARX to migrate files that are subject to record retention requirements, non-transactional or unchanging in nature, or simply old but still containing useful data to the most appropriate storage system. This allows customers to store data seamlessly on the most cost-effective storage media while still meeting access requirements for users.

EMC Data Domain deduplication storage deduplicates data inline before it is written to disk. Customers can move or copy data to Data Domain systems using standard NAS file system interfaces (NFS or CIFS), via a virtual tape library (VTL) interface using Fibre Channel (FC), or with Data Domain Boost software. Data Domain Boost provides tighter integration with backup management software solutions, such as EMC NetWorker (available 2H 2010), or Symantec NetBackup or BackupExec. As a result, many organizations have used Data Domain deduplication storage as a disk-based data protection solution because backups typically involve a significant amount of duplicate information—all of which take up expensive storage capacity. Data Domain systems have also been deployed as archive solutions to reduce the storage requirements for information being retained in an archive for information governance, electronic discovery, and business reference purposes. In addition to performing deduplication to eliminate redundancies, Data Domain systems also compress the data as it is stored, further reducing data footprint and storage requirements and allowing for network efficient replication across sites. The same Data Domain system can be used for both backup and archive use cases.

Using ARX with Data Domain deduplication storage enables the ARX solution to move files from a primary storage system to a Data Domain solution environment, resulting in savings on storage and associated costs related to power, cooling, and floor space. The Data Domain NAS file system interface can be identified and managed by the F5 ARX global namespace. As information is moved into a Data Domain system, it is deduplicated and compressed automatically. When an employee accesses a file, F5 ARX retrieves it from the Data Domain system as if it were in its original location, transparently to the user.

Putting the Solution to Work

Leveraging File Virtualization and Deduplication to Reduce Data Protection Costs

Ever-rising data volumes impact data protection strategies. ESG's recent data protection survey² found that 47% of midmarket respondents now report more than 10 TB of total data volume and 49% of enterprise organizations report more than 100 TB. Of the total data volume, an average of nearly 26% consists of on-site secondary storage and another 21% makes up off-site secondary storage. With reported annual growth in the double digits, the sheer volume of data presents ongoing challenges to IT organizations and impacts data protection strategies. That's the situation the team at Clackamas County found itself in.

² Source: ESG Research Report, [2010 Data Protection Trends](#), April 2010. All subsequent statistics are from this report unless otherwise noted.

A major driver of new capacity and data management requirements was Clackamas County's disaster recovery (DR) initiative, which requires replication of important county data to a secondary site. In addition to doubling storage capacity requirements, the DR initiative's aggressive recovery point objectives (RPOs) and high availability requirements were a strain on the team's data management and administration processes.

Numerous new regulatory and compliance challenges began to appear as well. Health, tax, and law enforcement agencies have implemented new policies, rules, and regulatory requirements for storing, protecting, retaining, and accessing specific types of information. Requirements differ from department to department, adding complication to the data management challenges of the county's small staff of systems administrators.

Finding the environment more and more difficult to manage, the Technology Services team began formulating a strategy to align data storage and management capabilities with the county's growing requirements. It was clear that the largely static existing infrastructure was not up to the task, but a forklift overhaul was out of the question. The strategy would have to meet mounting challenges in the face of a tough economy that was freezing budgets across county departments. Like nearly every IT group in today's economic environment, the Clackamas County team would be asked to do more with less.

The challenges facing Clackamas County are closely aligned with what ESG has seen from an overall business perspective in state and local government. As shown in Figure 2,³ state and local government spending will either stay flat or decrease for more than two-thirds of the state and local government organizations surveyed.

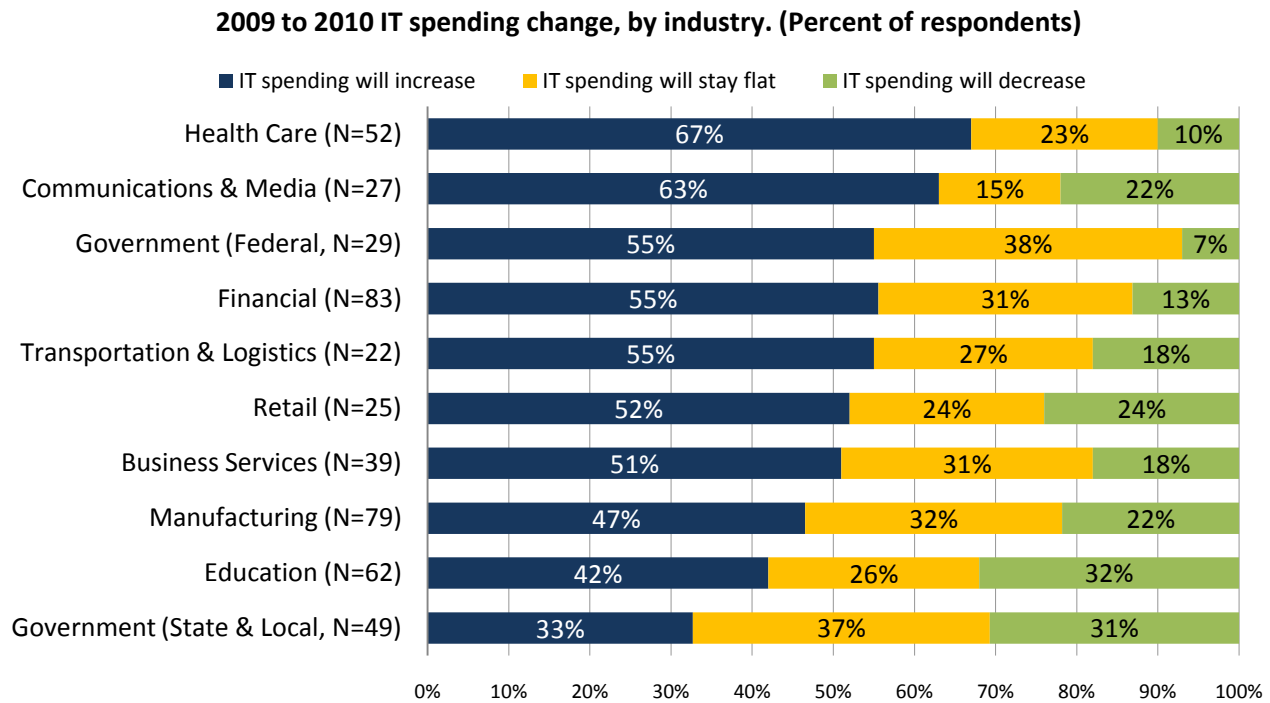
About Clackamas County

Clackamas County in the Portland, Oregon metropolitan area is home to 15 cities in its more than 1,800 square miles and boasts a population of nearly 380,000. The county government's Technology Services department, like IT shops across the country, has been asked to meet relentlessly growing service requirements without commensurate increases in budget or staff. Four system administrators manage the county's two data centers in Oregon City and have watched the county's data storage capacity requirements grow from 4 TB in 2005 to around 60 TB today.

The Technology Services team began implementation of a new storage strategy in 2008. Utilizing a joint solution from F5 and EMC, the team implemented a storage infrastructure strategy integrating file virtualization, tiered storage, data mobility and data deduplication. The new strategy produced a more fluid, scalable, and highly flexible data center. The solution has brought significant cost savings, higher service levels, and easier data management to the county in addition to providing a framework to accommodate future growth and modernization.

³ Source: ESG Research Report, [2010 IT Spending Intentions Survey](#), January 2010.

Figure 2. IT Spending Intentions, by Industry



Source: Enterprise Strategy Group, 2010.

The team knew that any solution would need to quickly expand capacity at minimal cost in order to drive more value from existing infrastructure investments. Managing this rapidly growing and increasingly complex environment without adding staff would require more efficient processes, automated tools, and continuous visibility into the entire infrastructure to strengthen future planning decisions.

The Technology Services team knew that the county’s business requirements demanded a more fluid IT environment. Centralization, virtualization, and pooled storage could optimize the use of existing resources while providing the flexibility needed to respond quickly and effectively to continued growth and rapid change. The first steps would be crucial. The team needed a solution that could tackle its most immediate challenges while enabling plans to create a more fully virtualized, dynamic infrastructure over time.

Clackamas County selected a joint solution from F5 and EMC to anchor its infrastructure modernization efforts. The solution’s architecture includes a pair of clustered F5 ARX1000 devices used to virtualize the file server environment and components of a high-performance iSCSI SAN, creating a pool of storage nodes with a single, global namespace. F5’s Data Manager software provides advanced storage administration, automation, and reporting capabilities. Two Data Domain DD500 Series systems—one in each of Clackamas County’s two data centers—power the lower layers of the architecture, providing data deduplication, replication, and archival storage.

“We were dealing with pretty significant data growth, aging file servers, disruptions to users whenever we had to do migration, and facilitating DR/high availability without a DR budget.”

— Christopher Fricke,
Senior IT Administrator
Clackamas County

The solution has been in place for about a year and the team is in the process of moving more and more of the county’s information into the new environment. The rollout strategy is to first demonstrate the value of the solution on select portions of the infrastructure and then, as departments and end-users become more comfortable with the centralized IT service approach and with trading their islands of control for more flexibility and enhanced performance, additional components of the infrastructure will be brought into the virtualized environment.

The results so far are positive and significant. The Technology Services group is confident that the F5/EMC solution lays a solid foundation for the future. File servers in the new environment no longer exist as departmental storage “silos.” The ARX system pools all storage assets in the environment under a single, global namespace to effectively make every gigabyte of unused capacity on every file server in the system available to any user or application on the network. The county has already reclaimed 10 TB of previously “stranded” capacity on its primary iSCSI SAN, equating to approximately \$100,000 in deferred capital, maintenance, and management costs. All of the 10 TB reclaimed were due to the EMC/F5 deployment. Post deduplication less than 2 TB of actual storage capacity is consumed on the Data Domain systems. Since deduplication ratios improve over time as data is retained longer, the team expects their deduplication ratios to increase.

The abstraction layer allows administrators to move files (or entire file systems) between storage nodes without disruption to applications or users. Changes to the production environment—such as decommissioning a server, upgrading software, or powering down equipment for maintenance—can now be made with no downtime or interruptions to business operations.

Automated Replication and Deduplication Reduce Costs *and* Improve DR Compliance

In Clackamas County’s configuration, one Data Domain storage system resides in each of the county’s two data centers, providing deduplication, storage capacity, and data replication in the new environment. The system in the primary data center is used as a backup target for the entire ARX environment while archival data is replicated to the Data Domain system in the secondary data center.

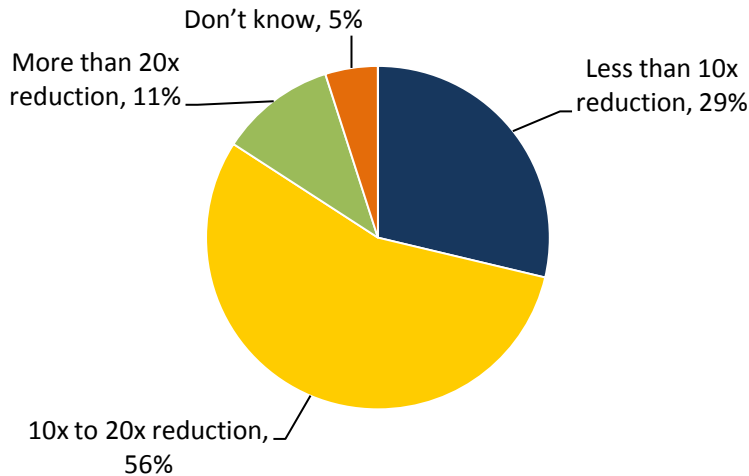
The ARX system enabled the team to develop and implement a storage policy that automatically places files in one of three distinct storage tiers based on file age and usage activity. Files under 60 days old reside on Windows file servers on the high performance iSCSI SAN; the second tier is used for files from 61-120 days old, utilizing more cost-effective storage nodes including SATA drives. The policy moves all files older than 120 days onto tier 3 for archiving. To reduce the archive footprint, data moving into the archive tier is first deduplicated on the Data Domain device in the primary data center.

The Data Domain appliances dramatically reduce storage capacity requirements on tiers 2 and 3 of the ARX environment. While Clackamas County’s DD500 Appliance Series systems are each configured to provide 5 TB of raw capacity (which is not yet fully utilized), they can scale up to 23.5 TB of raw and 980 TB of logical capacity. The variable-length-segment deduplication algorithm used in Data Domain systems is particularly well suited to the ARX file virtualization environment as it can detect, index, and remove redundant data patterns within files in a very granular fashion.

ESG research indicates that users can see significant data reduction through deployment of deduplication solutions, with 56% of users surveyed indicating a 10-20x reduction in the amount of data, significantly reducing the overall data storage footprint.

Figure 3. Data Deduplication Significantly Reduces Data to Manage

On average, what degree of capacity reduction has your organization experienced by using data deduplication technology? (Percent of respondents, N=140)



Source: Enterprise Strategy Group, 2010.

The county’s DR mandate was a driving force, leading the Technology Services team to the Data Domain portion of the joint solution. The DR plan requires all data to be replicated between the two sites and mandates recovery from any point in time. This amounts to a doubling of the county’s capacity requirements and of the resources required to manage the data and DR processes. The Data Domain systems are key enablers of the DR plan on all fronts; while deduplication greatly reduces the amount of data to be replicated, the systems replicate between themselves over low-bandwidth connections automatically, freeing administrators to focus on other tasks.

Automated Tiering puts Data Precisely Where it Belongs

Advances in storage technology over the last decade have prompted a shift in how enterprises view their storage infrastructures. While pure data volume and raw capacity are still fundamental concerns, IT shops know that all storage is not, in fact, created equal. For the Clackamas County team, the F5 Data Manager report illustrated how to better align storage allocation to support business needs.

The system automatically moves files to the appropriate tier as their statuses change; if a user or application modifies a file older than 120 days, for instance, it is moved to tier 1. Beyond its obvious efficiencies, the tiered storage and file classification model is paying dividends with the county’s backup processes as well. Classification of files according to age and modification date provides a way to perform nightly backups more selectively. If a file already exists on the backup target, backing it up again only adds value if it has changed since the last window. In the tiered environment, backups create fewer files, less data, and reduced network traffic. Clackamas County is only just beginning to realize the benefits of using Data Domain deduplication storage as a backup target and expects to reap even more benefits over time; while most of Clackamas County’s backup data still goes to tape, the team is increasingly using the Data Domain systems in the primary data center as a backup disk target.

“Over 90+% of that data sits unchanged forever, so finding an economical solution for it like the compression with the Data Domain through the virtualization and file tiering of the ARX is just fantastic.”

— Christopher Fricke,
Senior IT Administrator
at Clackamas County

Significantly, nearly all of the capacity growth in the county’s new environment is occurring on the lower (and lower cost) tiers of the storage infrastructure and that growth rate is significantly slower than if data were stored in a non-deduplicated state. Not only does the tiering scheme reduce the number of files competing for space on tier 1, but

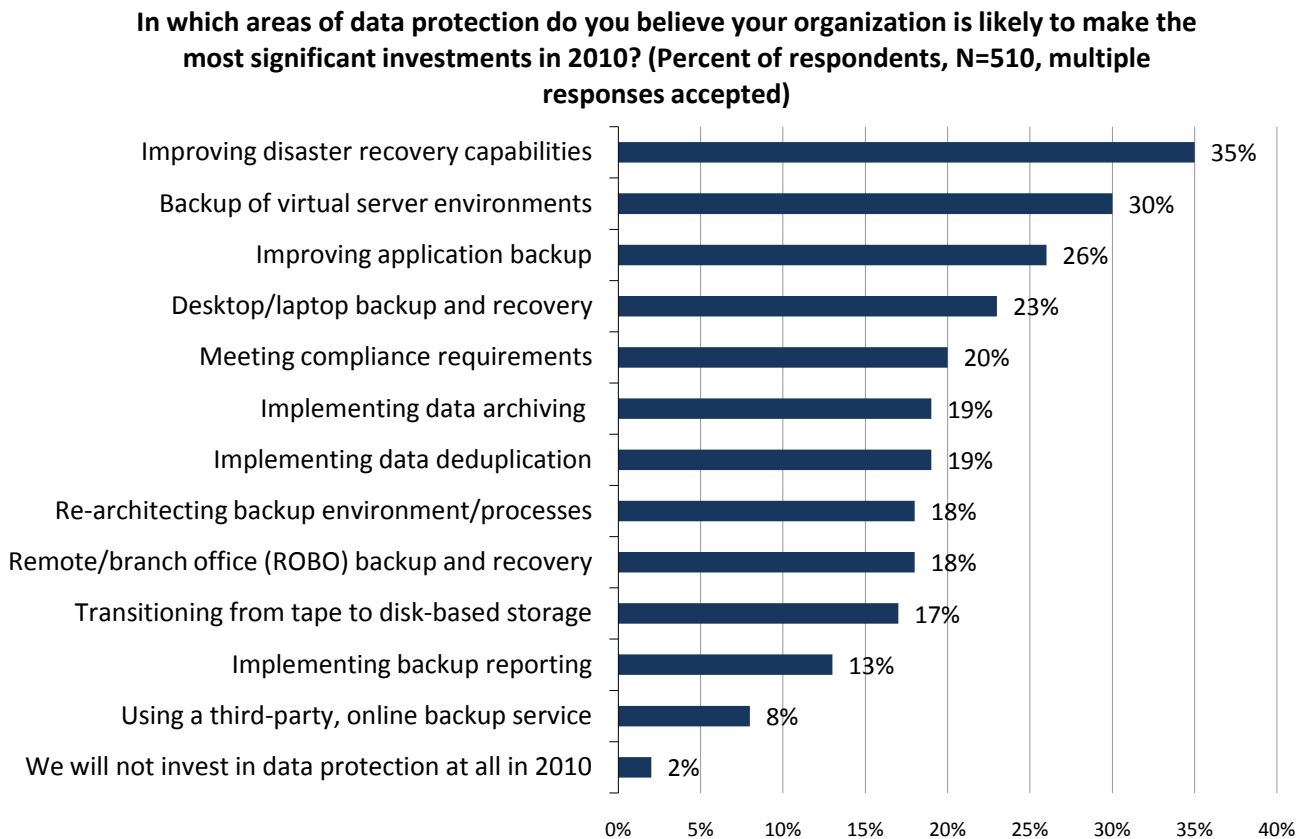
so far any new capacity requirements on the high performance SAN have been met by newly-available capacity recaptured through ARX file virtualization. Using solutions from F5 and EMC, the county was able to accommodate a year-on-year doubling of their data with zero increase in their storage budget.

The Clackamas County team counts automated storage tiering as the feature of the joint solution that has most transformed their enterprise and it is the root of much of the solution’s ROI.

Taking a Broader Market View

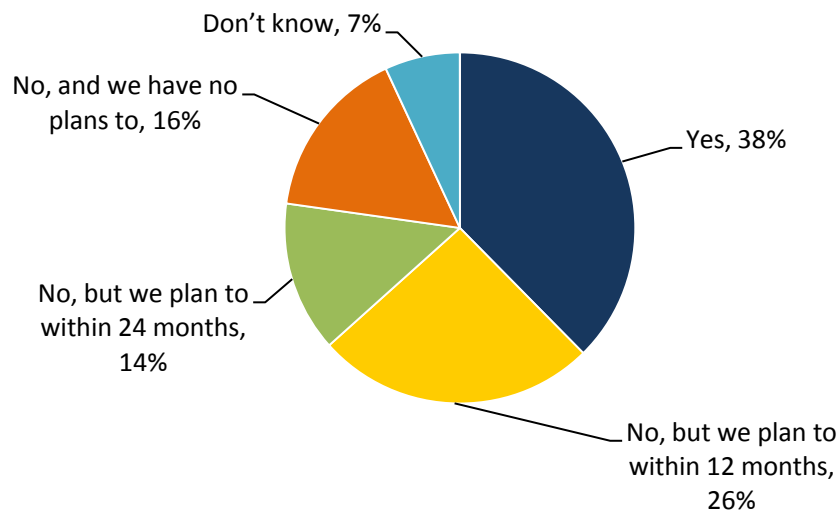
The challenges faced by the team at Clackamas County are not unique. Findings from ESG’s 2010 data protection survey indicate that improving disaster recovery (DR) is the top data protection spending initiative in 2010 (see Figure 4). In fact, this year’s top data protection investments are themed around mitigating risk and improving backup and recovery, much like what Clackamas County was facing as it modernized its infrastructure. ESG research found that of 35% of respondents cited improving DR, 20% also noted the need to meet compliance requirements, and 19% cited implementing data archiving, highlighting IT’s desire to minimize risk.

Figure 4. Plans for Data Protection Spending in 2010



Source: Enterprise Strategy Group, 2010.

ESG research also indicates that data deduplication is becoming a mainstream technology, having witnessed a large jump in deduplication adoption between its 2008 survey and today. Adoption in 2008 was 13%, whereas this year’s survey found 38% of respondents currently using deduplication.

*Figure 5. Users Turn to Data Deduplication***Is your organization currently using any data deduplication solutions? (Percent of respondents, N=369)**

Source: Enterprise Strategy Group, 2010.

Considering IT's current focus on operational cost reduction, these findings are not surprising. The legacy of the financial meltdown of 2008/2009, for IT at least, is a new willingness to spend capital dollars to save long term on operational dollars. Consider how a combination of file virtualization and data deduplication can help reach that end.

- **Transition to a flexible, services-oriented storage architecture.** The first step in transforming the storage architecture from fixed to fluid is breaking down the barriers between systems, which is essentially what F5 ARX does for the file storage environment. Once barriers are eliminated, IT can begin to take a services-oriented approach to storage as opposed to the reactive firefighting mode it has been in for so long. Data can be seamlessly moved and storage capacity deployed (or redeployed) where and when needed to support operations.
- **More efficient data protection.** F5 ARX moves inactive data to Data Domain systems, reducing the amount of content to back up from the primary file server. The results are far shorter backup and recovery times as well as significantly lower backup-related capital expenditure. Additional backup optimizations accrue from the fact that tier 2 is deduplicated.
- **Disaster recovery support.** Once on the Data Domain system, information can be replicated to another Data Domain solution for disaster recovery purposes. Since the system has already deduplicated and compressed the data, economies related to both reduced bandwidth and reduced remote storage capacity required for disaster recovery are realized.
- **Lower storage capital expenditures.** Using F5 ARX to move files to a Data Domain solution frees up capacity on primary file servers for newly created content. Between 70% and 80% of files created are persistent—they stop changing within their first 90 days, but still consume expensive tier 1 storage capacity. Moving persistent data to an archive tier frees up a significant chunk of this storage, which allows customers to delay purchase of primary or tier 1 file server capacity.
- **Lower storage operating costs.** By transparently moving data between storage tiers with F5 ARX, users can continue accessing data without disruption—there is no IT intervention required. Also, since data is deduplicated when it is moved to Data Domain deduplication storage, companies do not have to keep buying additional storage systems to save more information online for longer periods of time, which translates directly to greater power, cooling, and space efficiencies.

- **Seamless file archiving processes.** Many companies deploy tiered storage to facilitate archiving where data has to remain online and accessible for compliance, electronic discovery, and business reference reasons. Customers can configure F5 products with policies to automatically identify archive file candidates and move them to a Data Domain solution. For those organizations that need to comply with record retention regulations, like Clackamas County, they can leverage Data Domain Retention Lock software, which prevents data from being modified or deleted for the assigned retention period. In addition, EMC Data Domain Encryption software protects data in the event of theft or loss of disks or systems.

Clearly, the combination of file virtualization and deduplication is a powerful one-two punch for driving IT operational efficiency.

The Bigger Truth

Clackamas County's strategic approach and decisive response to its data storage and management challenges signaled a paradigm shift in how the team thinks about its infrastructure. Implementing the joint solution from F5 and Data Domain allowed the team to not only tackle its most immediate issues, but to visualize how a truly fluid, service-based IT environment might work for the enterprise. The team understood that to succeed more broadly, it first had to prove the joint solution's value for a significant portion of the enterprise.

The Technology Services team demonstrated that a more fluid environment featuring file virtualization, tiered storage, data mobility, deduplication, and automated replication drove significant cost savings, higher productivity levels, improved performance, and strengthened compliance. In short, the joint solution enabled Clackamas County to do more with less, *enabling the enterprise to accommodate 100% year over year data growth with zero growth in storage budget.*

While Clackamas County is only one story, it is indicative of the IT trend ESG has seen and validated through research. We are at the beginning of a journey, balancing the rigid and inflexible architectures of the past with dynamic services-oriented capabilities. Capabilities that are desperately needed to reduce waste are only just beginning to emerge. The tipping point isn't here yet, but it is becoming visible on the horizon. With solutions like those from F5 and EMC, users can start to address the management challenges associated with an overwhelming amount of unstructured data and optimize their file storage environments.



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