F5 White Paper

F5 ARX and Data Domain Deduplication Storage

Tiering Best Practices

Storage tiering helps businesses cut costs by automatically migrating aged or reduced-value files to lower-cost storage. Combining file virtualization with deduplication storage provides a simple, fast, and economical path to effective storage tiering. This paper reviews best practices for integrating Data Domain and F5 solutions for long-term archiving and data protection.

Contents

Introduction

Enterprises of all sizes are experiencing an explosion in the amount of file data under management. Users and applications are generating new content at a faster rate, while mandating online access long after the value of the content has diminished. With increasing budget pressures and IT workloads, organizations are looking for new ways to reduce the long-term costs of storing their business data. Storage tiering with deduplication storage offers a compelling solution.

Automated storage tiering with F5 ARX gives organizations the flexibility to optimize their file storage infrastructure by automatically moving file data to the appropriate storage tier as its value changes over time. Through virtualization, ARX moves files in a manner that is transparent to users and applications, maintaining online access regardless of file location. This flexibility enables organizations to take advantage of the wide array of storage technologies available today.

Data deduplication is one such storage technology that can dramatically reduce storage costs. By searching data streams for multiple instances (duplication) of the same data, then replacing the duplicate data with pointers to one reference instance of the data, deduplication can significantly reduce the amount of storage capacity required. Most vendors originally applied deduplication technology as a disk-based backup technology, but some have expanded its use to provide cost-effective, long-term file storage. In particular, organizations are deploying Data Domain deduplication storage to dramatically reduce both the capital and operating costs of storing file data for extended periods of time, while protecting data integrity and enabling rapid disaster recovery.

This paper explains how organizations can leverage deduplication storage in a tiered infrastructure to provide a simple, cost-effective solution to the problem of data growth. The first section discusses key considerations when implementing storage tiering using F5 ARX solutions. Topics include an overview of automated storage tiering, as well as considerations and best practices in designing a multi-tiered environment. Following that is a discussion of deduplication storage and how Data Domain systems can be best used in a combined solution to maximize the benefits of a tiered storage strategy.

Architecture and Technology

Tiered Storage Overview

In its most basic form, tiered storage is simply a storage environment comprised of two or more different types of storage. Each type of storage is optimized for a specific characteristic, such as performance, availability, or cost. Storage tiering matches files with the most appropriate type of storage, depending on what level of performance, availability and cost is desired.

While there are no restrictions in the number of tiers that an organization can employ, most can meet their storage management goals with just two storage tiers. Figure 1 shows an example of a two-tiered storage environment. This environment optimizes primary storage for performance and availability and secondary storage for capacity and cost.

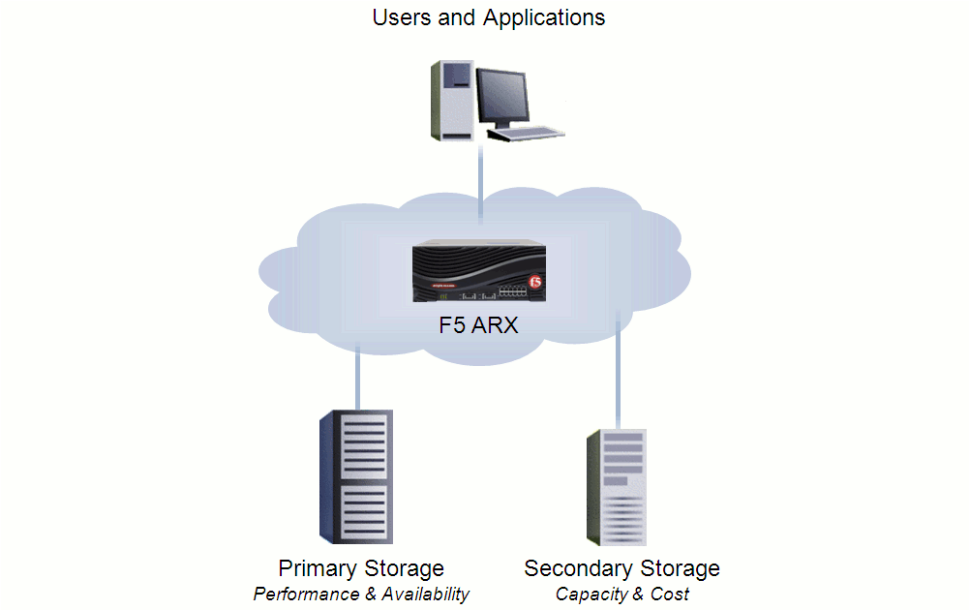

Figure 1. Tiered storage environment with F5 ARX

F5 ARX Architecture and Configurations

F5 ARX provides an inline, policy-based solution to automate storage tiering. ARX devices are logically situated between clients and file storage and act as a proxy for all file accesses. Because it is an inline tiering device, ARX can dynamically respond to changing storage requirements. The solution has two layers – presentation and automation.

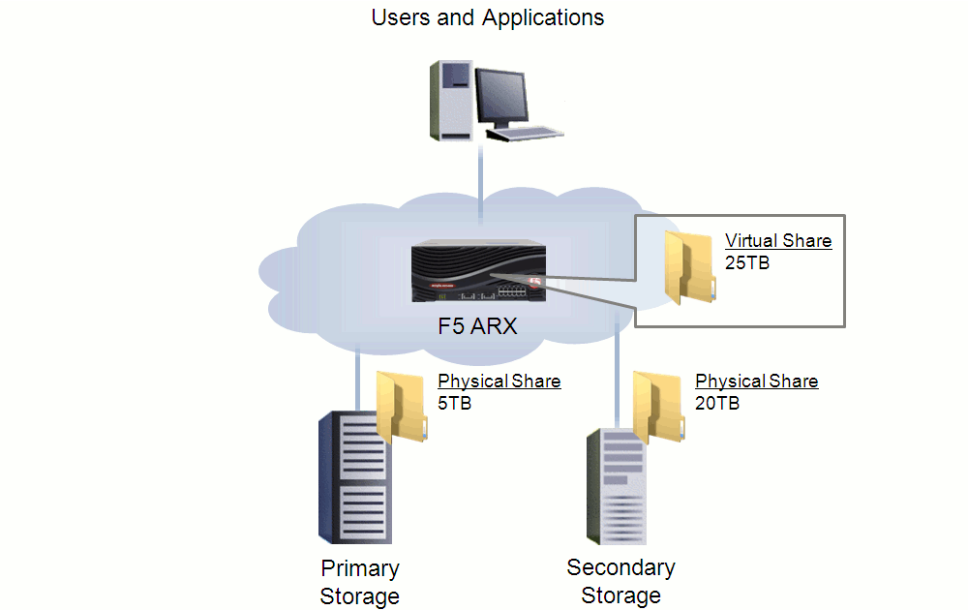

Figure 2. User presentation with F5 ARX

- **Presentation.** What is commonly referred to as file virtualization, the presentation layer allows enterprises to build storage environments comprised of different storage devices, platforms, vendors and technologies. It decouples the logical access to a file from its physical location on back-end storage.

  Through the presentation layer, the ARX device federates multiple physical storage resources into a single virtual pool. Figure 2 shows a virtual share that is comprised of capacity from two separate storage resources, a primary storage unit and a secondary storage unit, such as a Data Domain deduplication storage system. Clients see a logical view of one virtual share instead of two physical shares. Because clients perform logical access to files through the virtual share, administrators can move files between the two physical shares without impacting client access. The logical file mapping does not change, so file migrations do not require any client reconfiguration.

- **Automation.** The automation layer provides intelligence to ensure that all files are located on the appropriate storage tier. ARX devices employ data management policies that determine not only when and where to move existing files but also where to place newly created files. Administrators create policies based on a number of considerations, including a file's age, type, location, etc. The automation engine consists of several components, including filesets and place rules.

  A fileset tells the ARX to what collection of data it should apply a policy. It defines a group of files based on a shared attribute or collection of attributes. Filesets can be created based on any portion of a file's name, including its path or extension, size, or age based on last accessed or last modified date. Filesets can also be combined in unions or intersections to create more sophisticated policies. For example, a policy can target MP3 files, files that have not been modified in the last 30 days, or all MP3 files that have not been modified in the last 30 days.

  A place rule tells the ARX where to place a file matched by a fileset. Place rules can target a specific share or a collection of shares as either the source or the target for moving data.

ARX solutions are available in four platforms (Table 1). All ARX platforms provide these data management functions, but differ by scale and performance.

Platform	Ethernet Ports	Maximum Throughput	Maximum Users	Height (Rack Units)
ARX500	2x GbE	800Mbps	600	1U
ARX2000	12x GbE	4Gbps	6,000	2U
ARX4000	12x GbE or 2x 10GbE	12Gbps	12,000	4U

Table 1: F5 ARX Available Configurations

(For full specifications, see product datasheets)

Data Domain Architecture and Configurations

Data Domain inline deduplication storage systems can be used in conjunction with ARX devices as a consolidated secondary storage tier for backup, archive and disaster recovery to improve storage efficiency across the distributed enterprise. These systems dramatically reduce the amount of disk storage needed to retain and protect enterprise data. Typically, organizations realize an average of 10-30x data reduction for backup and archive data sets simply by identifying redundant files and data before they are written to disk. Data can then be efficiently stored, replicated and

retrieved over existing networks for streamlined disaster recovery and consolidated tape operations. Data Domain storage systems can be used to cost effectively store petabytes of data.

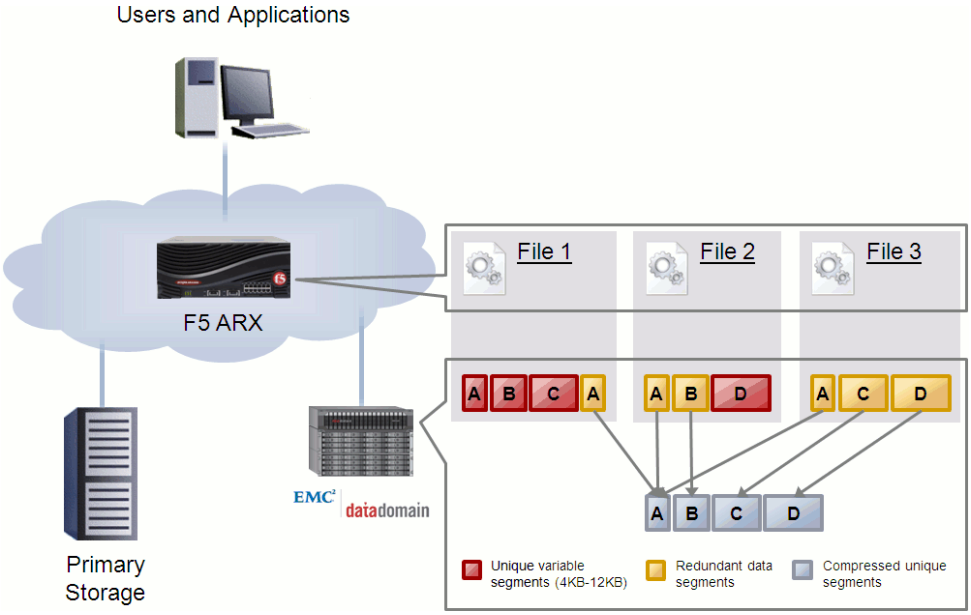

Figure 3. How Data Domain Inline Deduplication Works

Data Domain inline deduplication breaks the incoming data stream into variable-length segments, uniquely identifies each one, and then compares the segments to previously stored data (Figure 3). If the segment is unique, it is stored on disk along with the metadata. If the segment is a duplicate of what has already been stored, only the metadata with a reference to the existing segment is stored. Data Domain's Data Invulnerability Architecture lays out the industry's best defense against data integrity issues, by providing unprecedented levels of data protection, data verification, and self-healing capabilities that are unavailable in conventional disk or tape systems. Finally, in support of disaster recovery, Data Domain Replicator software transfers only the deduplicated and compressed unique changes across any IP network, requiring a tiny fraction of the bandwidth, time and cost compared to traditional replication methods.

By consolidating backup and archive data to a common disk-based target, you can avoid creating disparate islands of storage, reduce costs, and simplify data management. A single Data Domain deduplication storage system can be used for protection of enterprise applications, archiving, and online reference storage. The deduplication benefits are then shared across the aggregate data set, as are the

unique system resiliency, replication and disaster recovery capabilities of Data Domain deduplication storage. With the reduction in needed storage capacity, many organizations realize associated savings in power, cooling and space within the data center.

Data Domain solutions are available in several platforms (Table 2). All Data Domain systems provide the same data deduplication functionality, but vary in terms of performance and capacity.

Platform	Speed	Logical Capacity	Raw Capacity	Usable Capacity
DD140	450 GB/hr	17-43 TB	1.5 TB	.86 TB
DD610	675 GB/hr	75-195 TB	Up to 6 TB	Up to 3.98 TB
DD630	1.1 TB/hr	165-420 TB	Up to 12 TB	Up to 8.4 TB
DD660	2 TB/hr	.520-1.31 PB	Up to 36 TB	Up to 26.1 TB
DD690/g	2.7 TB/hr	.710-1.7 PB	Up to 48 TB	Up to 35.3 TB
DD880	5.4 TB/hr	1.4-3.5 PB	Up to 96 TB	Up to 71 TB
DDX Array	86.4 TB/hr	22.6-56.7 PB	Up to 1.5 PB	Up to 1.13 PB

Table 2: Data Domain Appliance Available Configurations
(For full specifications, see product datasheets)

Reference Architecture

Figure 4 shows a reference architecture for deploying Data Domain systems with F5 ARX in a tiered storage implementation. The primary goal of the reference architecture is to reduce storage costs through data deduplication. To that end, it proposes a two-tiered storage infrastructure, with Data Domain systems providing the secondary storage.

The secondary goal of the reference architecture is to reduce backup costs while maintaining existing levels of data protection. To this end, it employs a tiered data protection strategy, using the same Data Domain system for the primary storage backup target. All data is replicated to a second Data Domain system to provide redundancy for both secondary storage and backups of primary storage for disaster recovery.

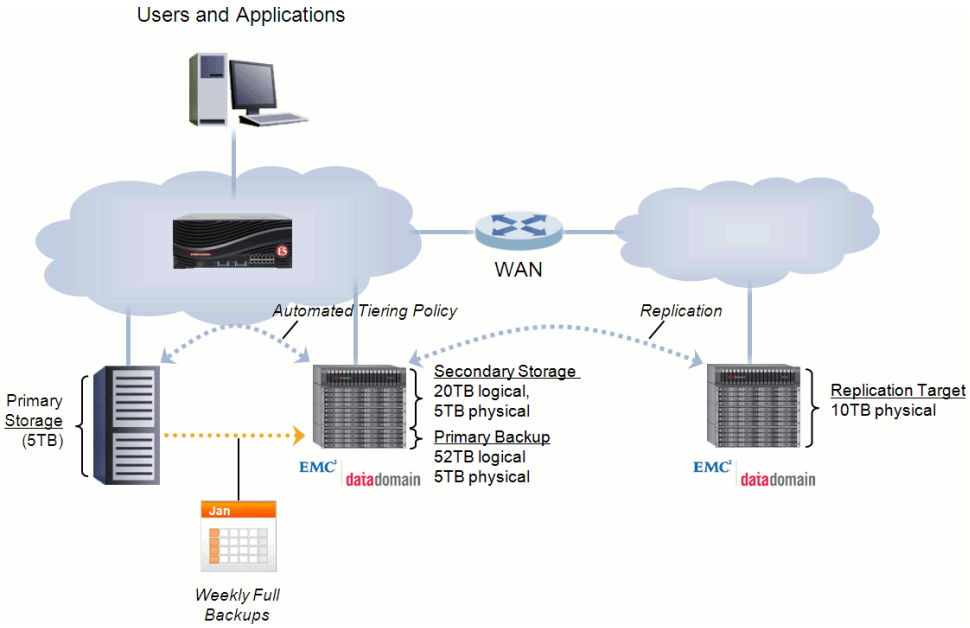

Figure 4. Reference architecture for F5 ARX and Data Domain systems

Some considerations when planning a deployment:

- **Amount of data.** This reference architecture provides an example using an environment with 25TB of data. Organizations should consider the amount of data they have under management and adjust the recommendations accordingly, but the analysis still holds for greater or lesser data volumes.

- **Primary storage.** The reference architecture assumes that the organization has existing storage capacity that is well suited – demonstrating high performance and availability characteristics – for primary storage purposes. In this example, the organization requires 5TB of primary storage.

- **Sizing secondary storage.** The reference architecture assumes that an organization is looking to augment its existing primary storage with a Data Domain tier in order to reduce its storage costs.

  The ARX will automatically move inactive or less critical data to the Data Domain system based on defined tiering policies. For a typical enterprise, this will represent approximately 80% of the total amount of file data under management. The exact percentage will vary from organization to organization and will depend on the parameters of the tiering policies.

  The Data Domain system should have sufficient useable capacity to store all inactive data (80% of data) after accounting for expected deduplication savings.

The expected deduplication will vary between different types of data. For reference, migrating user home directories to a Data Domain tier often results in between 2x and 6x deduplication. In this example, the organization would need 20TB of logical capacity, or 5TB of useable capacity (assuming 4x deduplication) on the Data Domain system.

- **Sizing primary storage backups.** Active data (20% of data) residing on primary storage can be backed up to the Data Domain system. Deduplication savings in the backup use case are normally much higher than for file storage, given the more redundant nature of backup data. On average, organizations typically see savings of 10-30x. The actual results will vary depending on a variety of factors, including the backup policy, type (full, incremental, etc.), frequency, rate of change, and retention.

  In the reference architecture, backing up 5TB of primary data will typically require ~6.5TB of logical capacity for each week of retention. This adds up to ~52TB of logical capacity, or ~5TB of physical capacity (assuming 10x deduplication), in order to store eight weeks of backups for data on primary storage.

- **Sizing the replication target system.** The first Data Domain system contains not only inactive data, but the backups of active data on primary storage. To minimize the risk of data loss from disasters or human error in the data center or remote offices, it is recommended that organizations replicate between the first Data Domain system and a second remote system for redundancy. The two Data Domain systems should have the same useable capacity, 10TB in this example.

  While outside the capacity calculations of this reference architecture, some organizations may also utilize the main system as a replication target for other Data Domain systems for disaster recovery. For example, they may replicate data from branch offices to a centralized data center and vice versa to minimize the risk of data loss. In this case, the Data Domain system should have sufficient additional useable capacity to store any replicated data that is expected.

- **Tape backup.** The reference architecture ensures that all primary and secondary data, as well as primary backup data, is stored in two separate locations. While not necessary, tape backups can be performed against either Data Domain system if required, e.g. for archival purposes.

- **Growth.** The reference architecture accounts for existing data requirements and does not account for future growth. Organizations should consider their expected growth rate for file data, as well as their preferred purchase cycle for disk capacity, and adjust the recommendations accordingly.

Design Considerations

Success Metrics

When properly implemented, tiered storage with deduplication provides users and applications with the level of performance and availability they require, at the lowest overall cost. Every organization will need to determine the best balance of these characteristics for their own environment to meet their business needs. However, some metrics of a successful tiered storage implementation include:

- **Reduction in storage costs.** The primary goal for implementing storage tiering is to reduce storage costs. By moving the majority of data to deduplication storage, savings can be realized in several areas, including reduced capacity costs, power consumption, floor space and cooling. The exact savings will be determined by a number of factors, including the amount of data that can be moved to lower cost storage and the degree of duplication in an organization's data. In addition, the degree to which storage for backup, archive and replicated data can be consolidated will enhance the cost savings.

- **Reduction in backup and recovery times.** Backup and recovery policies should reflect the business value of the data being protected. Storage tiering can help organizations prioritize data backup and recovery by separating active and inactive data, which can then be backed up at different intervals and with different retention policies. A successful tiering policy should strive to minimize the amount of time required to backup data that is actively changing in order to reduce the impact of those backups on the business. Similarly, it should also strive to minimize the amount of time required to recover data, with priority given to that which is the most active, in order to get users and applications up and running as quickly as possible if and when needed.

- **Reduction in backup media consumption.** Tiered storage with deduplication enables organizations to utilize backup media more efficiently without any loss in the level of data protection. By separating active and inactive data, organizations can apply different backup policies to each class of data, such as weekly full backups for active data and monthly or quarterly backups for inactive data. Using Data Domain deduplication storage as the target for these backups complements this approach by further minimizing the amount of disk and bandwidth required to store and replicate these backups.

- **Impact on application performance.** While the primary goal for implementing storage tiering is to reduce storage capacity costs, productivity of users and applications should not be adversely impacted. A successful tiering policy should strive to ensure that all data is placed on storage with the appropriate level of performance. For data that is business critical or frequently accessed, that may mean higher performance storage devices. Data that is non-critical or infrequently accessed may be more appropriately placed on lower-cost storage devices. For data requiring archiving or DR protection through replication, the tiering policy should target secondary storage supporting such capabilities.

Determining Business Value

The premise of storage tiering is that not all data is created equal. Different files will have varying degrees of value to the business, and that value changes over time. ARX can employ a number of methods to determine the business value of each file based on its attributes.

- **Age.** Users and applications often create data for short-term consumption, but which IT must retain for an indefinite period of time. A policy that determines value based on file age assumes that files which have not been accessed or modified recently will likely continue to be accessed infrequently, or not at all.
- **Type.** File type can often indicate business value, relative to other file types. Some types of files are inherently more or less valuable than others. For example, file types known to be created by a mission critical application are likely to be more important than a user JPG file.
- **Name.** Organizations often apply a naming schema that indicates the importance of the file, based on application, project, owner, etc. In some situations, certain applications, projects or users can be easily identified to be more important than others.
- **Location.** The location of a file can sometimes indicate business value, relative to other locations. For example, applications typically store their files within a specific and easily-identifiable directory structure. An organization can prioritize between different applications based on the locations of their data.

Selecting Storage Types

There is a wide variety of storage systems, platforms, vendors and technologies available today. Enterprises should consider the following storage characteristics when selecting the mix of storage types that satisfies their business requirements.

- **Cost.** Businesses today generate data faster than ever and retain it for longer periods of time. In contrast, many enterprises are trying to reign in media (tape and disk) and operational costs, which strain IT storage budgets. Businesses can choose from a range of technologies at different price points, including Fibre Channel (FC) drives, serial ATA (SATA) drives, solid state drives (SSD), and deduplication storage. The cost of storage capacity for these disk technologies will differ on a per TB basis. Tape has historically been another option, offering lower media costs but higher management and other operational costs, especially if fast and reliable restores or disaster recovery readiness are required. Inline deduplication storage is gaining in usage due to its ability to offer a reduction in data storage needs for archiving by 80% or more on average, and a rich feature set that ensures maximum data integrity.

- **Performance.** Many factors may influence storage performance, including the storage platform architecture, throughput or I/O capabilities of the specific storage device, type of storage media, disk size and speed, and network bandwidth. While some applications require extremely high performance, many do not – the most cost-effective architecture places data on storage that provides the appropriate level of performance so organizations only pay for the performance level they need.

- **Availability.** The cost of downtime typically drives requirements on storage availability. For example, customer-facing or revenue-generating applications may require very highly available storage, whereas user MP3 libraries generally do not. Many factors may influence storage availability, including the storage platform architecture and product maturity. Disk-based systems have huge advantages over tape storage in terms of availability of data, especially if a restore or disaster recovery is required. Restoring or recovering from tape can be slow (low Recovery Time Objective (RTO)) and prone to failure/data loss (bad Recovery Point Objective (RPO)), degrading availability. Compared to tape, inline deduplication storage can slash backup times for superior RTO and RPO and fast time to DR through real-time replication. This ensures maximum system availability.

- **Green attributes.** With energy costs rising and pressure continuing for data centers to control footprint growth, many organizations are looking to aggressively manage their consumption of data center power, cooling, and space. Higher-performance storage solutions carry a heavy power, cooling, and space penalty, so their use should be managed efficiently and balanced against more green storage alternatives, such as inline deduplication storage systems, when possible.

Getting Started with Tiering

Defining Filesets

When defining a fileset, organizations should consider the attributes that can be used to determine value in their business environment:

- **Age.** A fileset based on age should group files based on how quickly their business relevance decays for that organization. For example, log files are often frequently accessed for a short window after creation and then rarely after that window expires. A fileset that recognizes this may separate files into two or more buckets:
    - Files that have been modified in the last 30 days
    - Files that have not been modified in over 30 days
- **Filename.** A fileset based on filename will group files by matching any portion of a file's name. This can be used for different purposes.

  A fileset based on type will group files according to their extension, which is contained in the filename. This policy should consider if there are certain file types that inherently have more or less value relative to other types. For example, MP3 files generally have low value compared to other file types. A fileset that recognizes this may identify all files with "*.MP3" in their name.

  A fileset based on location will group files according to their path, which is contained in the filename. This policy should consider if there are certain locations that have inherently more or less value relative to other locations. For example, an engineering organization may keep all of the source files for each product release in a separate directory. They would likely aim to keep files related to the current release on high-performance storage to reduce build times, but would rarely need to access past releases.
- **Size.** A fileset based on size will group files based on the assumption that certain file types that have more or less value relative to other file types can be identified based on the size of file. For example, a business that streams very large video files may want to always keep very large files on high-performance storage. A fileset that recognizes this may separate files into two or more buckets:

  - Files that are smaller than 1GB in size
  - Files that are larger than 1GB in size

- **Duplication.** While a fileset cannot explicitly target the degree of duplication, an additional consideration specific to a Data Domain environment is how to optimize filesets for deduplication storage. The amount of cost savings that organizations realize from Data Domain systems is largely based on the level of duplication present in their data. Therefore, these organizations should design filesets that take the level of duplication into account in order to maximize cost savings.

  One example of such a fileset identifies files in NetBackup with the BKF extension. As with any data backups, there may be a large degree of redundancy in any collection of BKF files, based on the scheduling and retention policies governing that backup process. By creating a policy that identifies BKF files and places them on the Data Domain system, an organization could significantly reduce the amount of capacity consumed by these files.

Types of Policies

There are two types of policies used to tier file data between different storage tiers, scheduled policies and file placement policies. Both achieve different goals and many organizations use a combination of the two.

- **Scheduled policies.** Scheduled policies are the most common type of tiering policy. A scheduled policy matches attributes of existing files against the criteria for each storage tier, and migrates any relevant files to the appropriate tier at a regularly scheduled interval.

  Scheduled policies (Figure 5) are useful to match dynamic file attributes with the most appropriate storage. For example, the most prevalent form of tiering is one based on file age. A scheduled policy will watch files as they age and migrate them to different storage tiers when appropriate. Files that have not been recently modified are migrated to secondary storage. If the files are ever modified again, they can be migrated back to primary storage.

  Scheduled policies can be performed on filesets based on age, filename or size.

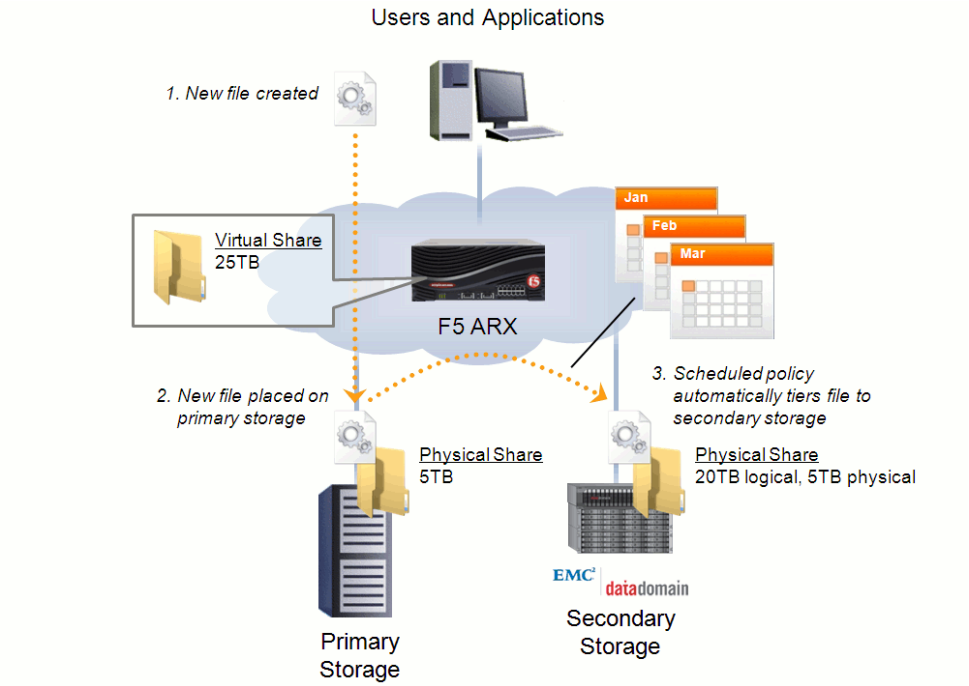

Figure 5. Example of file tiering with a scheduled policy

- **File placement policies.** A file placement policy matches attributes of files against the criteria for each storage tier as they are created. Files are then placed on the most appropriate storage tier from initial creation, rather than having to be moved at a later date.

  File placement policies (Figure 6) are useful to match static file attributes with the most appropriate storage. For example, an organization may always want to place specific file types, such as MP3 files, on a lower cost tier. A file placement policy will catch files as they are created and automatically place them on the appropriate storage tier.

  Note that a file placement policy will take precedence over a scheduled policy. Files that are placed on a specific storage tier by a file placement policy will not be moved to a different tier in the future by a scheduled policy.

  File placement policies can only be performed on filesets based on filename or size.

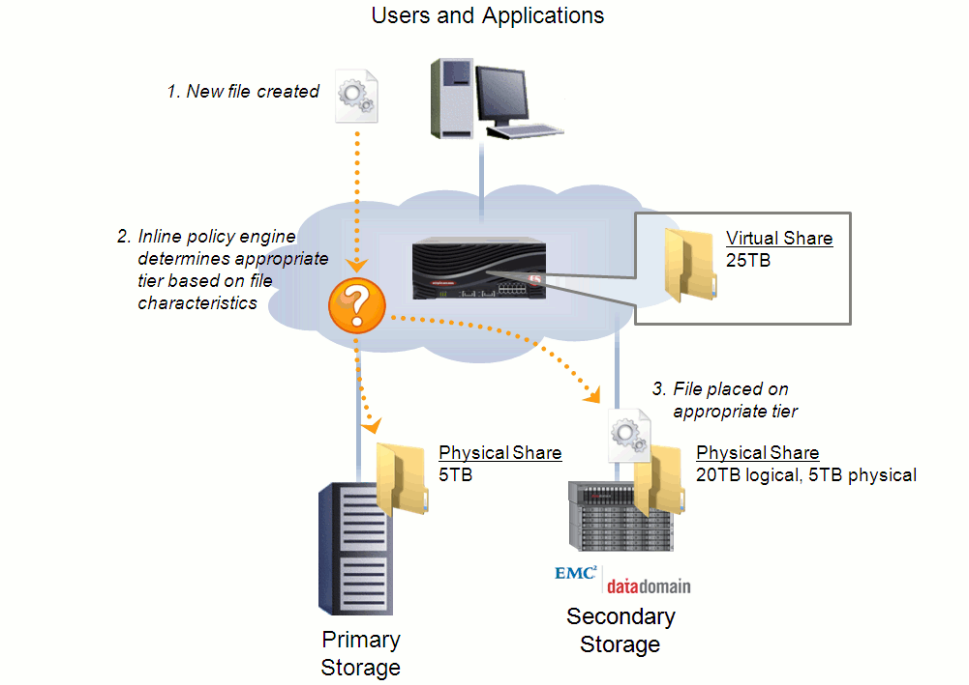

Figure 6. Example of new file creation with a file placement policy

Creating Tiering Policies

Every enterprise faces a different set of circumstances that will affect how they implement storage tiering to meet their business goals. However, here are a few best practices to consider when designing deploying automated storage tiering.

Running "What If" Scenarios

Before executing a policy for the first time, administrators should make sure that the results will be as expected. ARX can execute policies in a tentative mode to show the outcome of a particular policy in the storage environment. Administrators can then further explore how a proposed policy will affect backup, archiving and replication.

Starting with a Scheduled Policy

A scheduled tiering policy based on file age is the simplest to conceptualize. It is based on the assumption that files become less relevant as they age and therefore can be moved to lower-performance, lower-cost, storage without adversely impacting users or applications. This scheduled policy can be further tuned based on file type, location, etc.

In most organizations, 70-90% of all data under management is older than 30 days. This means that the large majority of data in a typical single-tiered environment is residing on high-performance storage but rarely accessed or modified. As a result, implementing a simple age-based tiering policy will often accomplish many of the goals of storage tiering by simply moving all inactive data to a long-term archival tier, such as deduplication storage.

Considerations. There are several considerations when designing a scheduled age-based tiering policy:

- **Delineation of active vs. inactive data.** In determining the age ranges constituting active and inactive data, organizations need to consider their operational access patterns. For example, an engineering organization may need to keep all files related to the current product release on high-performance storage. If the product development cycle is half a year, then data over half a year old pertaining to that release would be considered inactive data. However, this range may differ for the organization's user files. Users often create documents for immediate consumption and may rarely access those files after a few weeks. In that case, data older than 30 days could be considered inactive data.

Figure 7 shows a sample breakdown of file data by age for a mid-size technology enterprise, as described in the *Data Manager Sample Report*[1]. It highlights the amount of capacity that older files can consume: only 6% of data has been modified in the last 30 days and 31% in the last half year. While the breakdown will vary between organizations, the relative percentages of each age range are typical for most enterprises.

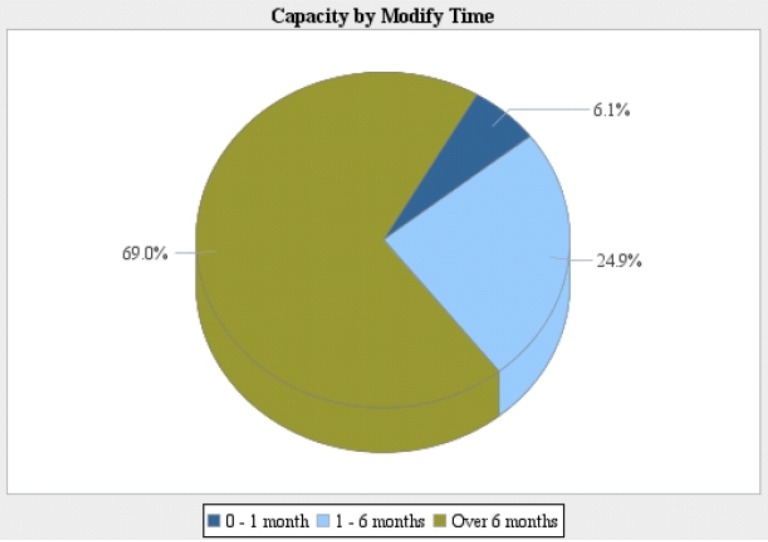

Figure 7. Sample breakdown of file data by last modified date

In order to determine the best delineation of active vs. inactive data, enterprises should start with a conservative approach – e.g., designating files more than 180 days old to be inactive. Once they are more comfortable with the impact of storage tiering in their environment, they can tune the policy to be more aggressive. Some organizations only keep files less than 30 days old on high-performance storage.

- **Last modified vs. last accessed.** Files have two attributes that an ARX can use in an age-based tiering policy – the last accessed date and the last modified date. At first glance, the last accessed date may appear to be better suited for tiering, as it indicates how recently a file has been either read or written to, as opposed to the last modified date, which only indicates the last write. However, many organizations create tiering policies based on the last modified date. This is because non-critical applications, such as anti-virus, often access all files on a

[1] Data Manager File System Inventory Summary Report, excerpted 6/1/09, http://www.f5.com/pdf/products/data-manager-sample-report.pdf

regular basis. Because this type of access updates the last accessed date, that attribute would no longer accurately represent the business value of the file.

- **Scheduling.** While ARX can run policies at very short intervals, most organizations schedule policies to run every 30, 60 or 90 days. When determining the most appropriate schedule on which to run a tiering policy, administrators should consider several factors, including the amount of data and the delineation of active versus inactive data.
- **Data backup.** Scheduling is often tied to the organization's backup schedule. For example, they may want to place all files older than 90 days on a lower cost tier and perform full backups on a monthly basis. In this case, the tiering policy should also be scheduled to run on a monthly basis and immediately precede the full backup. For more information on how to optimize scheduled policies for data backup, see *Optimizing Data Backups*, below.

Creating a File Placement Policy

A file placement policy can be employed to supplement age-based tiering and provide a more sophisticated tiering implementation. It is based on the assumption that some filesets have a permanent value (or lack thereof) to the business and should be always placed on a specific tier of storage.

An organization may have identified certain file types, such as business application files, that it must always retain on a high-performance tier. Conversely, it may also have certain file types, such as MP3 files, that it never wants to store on a high-performance tier. A type-based policy makes it very easy to ensure that these types of decisions are always enforced. Figure 8 below shows a sample breakdown of file data by type in the user home directories of a mid-size technology enterprise, as described in the *Data Manager Sample Report*.

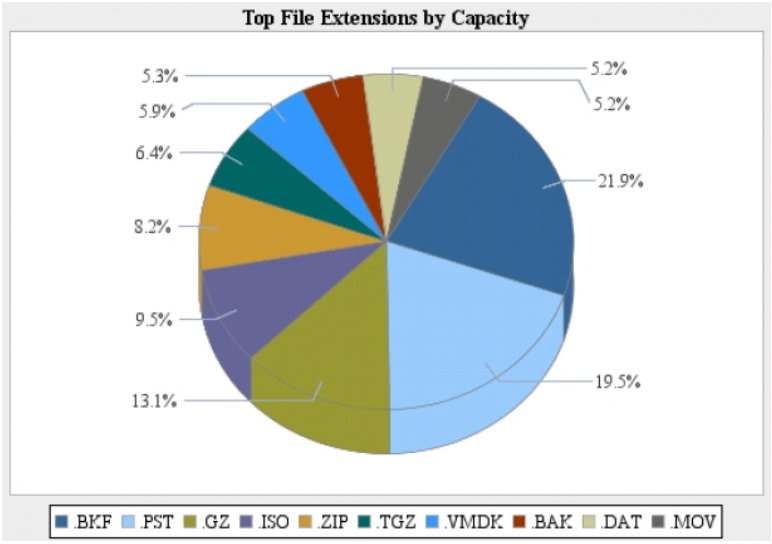

Figure 8. Sample breakdown of file data by type

Considerations. There are several considerations when designing a scheduled age-based tiering policy:

- **Precedence.** A tiering policy based on file type is a file placement policy that will always take precedence over the primary age-based policy. Therefore, organizations should ensure that files that are specifically identified by a file placement policy have a business value, relative to other file types, that does not change over time.

- **Data Backup.** File placement policies are often employed to optimize backups. Organizations often identify specific filesets that they wish to backup on a different schedule from other data.

  One example of this is Microsoft personal folder PST files. PST files can consume significant amounts of capacity in user home directories. They are important to retain but rarely change. By creating a file placement policy to always place PST files on a secondary tier, organizations can apply a separate backup policy to back up PST files less often and save a significant amount of backup media.

Optimizing Data Backups

Up to this point, this paper has examined general guidelines for file virtualization and storage tiering. This section will apply those guidelines to a specific example: data backups. While most of the sections below speak to generic recommendations and benefits with regards to backup, the *Backup with the Reference Architecture* section addresses best practices and benefits specific to the reference architecture described earlier.

Virtual vs. Physical Shares

In a virtualized file storage environment, organizations have the choice of performing data backups of virtual or physical shares. However, most organizations typically continue to perform backups of physical shares.

Benefits. While backing up a single virtual share may seem to be easier than backing up multiple physical shares, it overlooks powerful benefits that storage tiering policies provide. Many organizations design tiering policies to maximize backup benefits as well as storage cost savings. They can apply different backup policies to each storage tier to reduce the amount of redundant data being backed up. The frequency and retention of backups applied to each tier reflects not only the business value of the data on that tier, but also the characteristics of that data.

Considerations. There are several considerations related to designing data backup policies as part of a tiered storage implementation:

- **Active data.** See *Optimizing for Active Data*, below.
- **File type.** See *Optimizing for File Types*, below.
- **Reference architecture.** See *Backup with the Reference Architecture*, below.

Optimizing Backup for Active Data

Storage tiering can dramatically reduce the amount of time enterprises spend backing up and recovering data, as well as the amount of backup media required. By delineating between active and inactive data, organizations can reduce the amount of redundant data being backed up on a regular basis.

Benefits. The amount of benefit an organization will see will depend on a number of factors, including the age characteristics of their data and the breakdown in age ranges established by policy. However, organizations will typically see a significant reduction in backup times and backup media required.

Figure 9 below provides a simple yet powerful example of how an age-based tiering policy affects backups. Here a tiering policy separates active and inactive files between two or more physical shares based on the last modified date. Files that have been modified in the last 90 days are placed on Share A. Files that have not been modified in the last 90 days are placed on Share B.

This organization will continue to perform a full backup of Share A on a weekly basis. However, it can now perform a full backup of Share B on a monthly or quarterly basis without any loss in the level of data protection. This reduces the amount of backup media consumed, which can be calculated from the amount of data stored on Share B and the number of copies that must be retained. In addition, this greatly reduces the time required for weekly full backups since the backup agent only needs to scan the data residing on Share A.

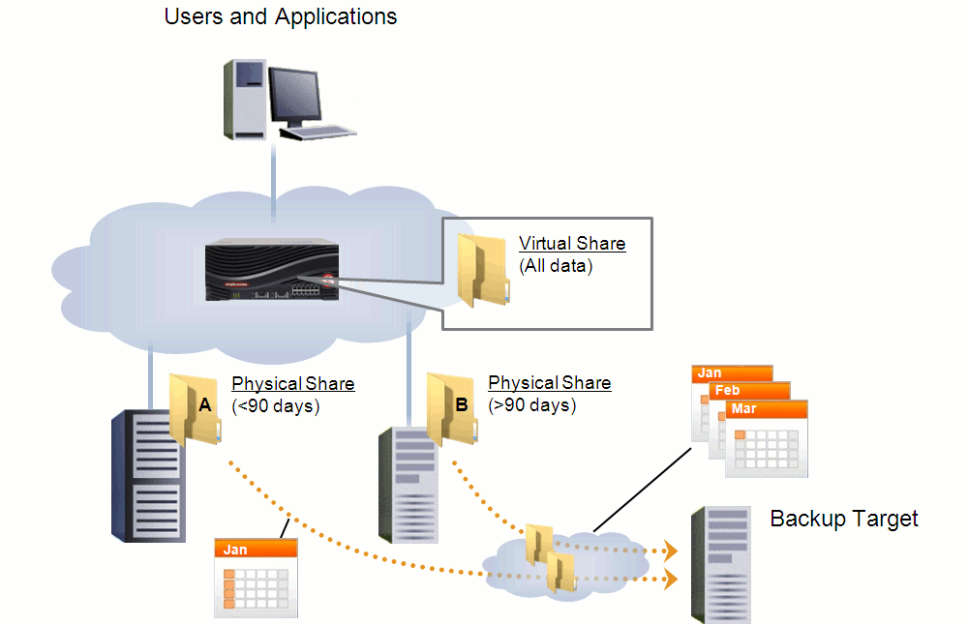

Figure 9. Optimizing backups by file age

Optimizing backups for active data also provides significant benefit for data recovery. In the event of a disaster, organizations can prioritize the data to be restored in order to get the business up and running in the shortest time. They can quickly restore

shares known to contain active data immediately, leaving shares known to contain inactive data until a later time.

Considerations. There are several considerations related to designing data backup policies as part of a tiered storage implementation:

- **Backup schedule.** An age-based tiering policy migrates files to a lower tier based on their age. For example, a tiering policy may move all files that have not been modified in the last 90 days to a long-term storage tier. In that case, then all files on that tier are guaranteed to not have been changed in the last 90 days. Therefore, it is appropriate to not back up inactive data on that tier as frequently as for active data. Organizations typically choose to forgo weekly full backups on a long-term storage tier. However, depending on the tiering policy, it may be safe to forgo even monthly or quarterly backups.
- **Reference architecture.** See *Backup with the Reference Architecture*, below.

Optimizing Backup for File Types

Organizations can also often reduce the amount of redundant data being backed up by delineating between different file types when creating tiering policies. In fact, type-based policies are often implemented with data backup in mind.

Benefits. The amount of benefit an organization will see will depend on a number of factors, including the amount of capacity and the number of files of the relevant file type. However, organizations will typically see a significant reduction in backup times and backup media consumed.

Figure 10 below provides a simple yet powerful example of how a type-based tiering policy affects backups. Here a tiering policy delineates PST files from other file types. By default, all files will be placed on Share A. However, here all PST files are placed on separate Share B.

This organization will continue to perform a full backup of Share A on a weekly basis. However, it can now perform a full backup of Share B less frequently. While the exact interval between backups will vary between organizations, this reduces the amount of backup media consumed, which can be calculated from the amount of data stored on Share B and the number of copies that must be retained. In addition, this greatly reduces the time required for weekly full backups since the backup agent only needs to scan the data residing on Share A.

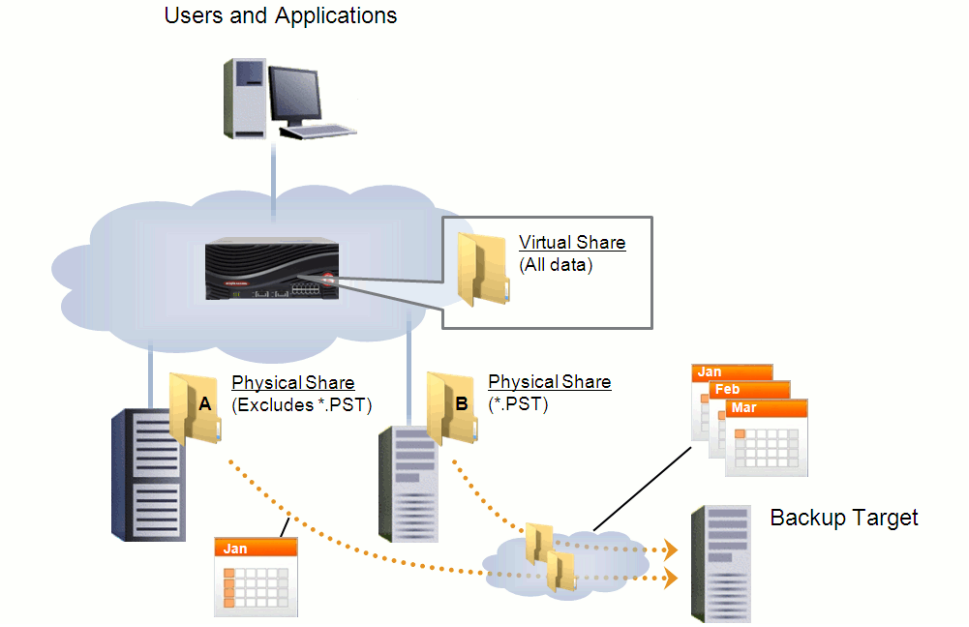

Figure 10. Optimizing backups by file type

Considerations. For additional considerations regarding the reference architecture, please see *Backup with the Reference Architecture*, below.

Backup with the Reference Architecture

The reference architecture shown in Figure 11 takes backup optimization a step further. Using this architecture, organizations of all sizes can reduce or eliminate any tape media consumption.

For backing up primary storage, the reference architecture replaces all tape media with a Data Domain system. Daily backups of active data are made to deduplication storage, which can eliminate much or all of the tape media consumption in a typical enterprise, and minimizes the amount of disk required to store weeks of these backups for recovery. Primary backup data is then replicated to a second, off-site Data Domain system for redundancy; because only deduplicated data traverses the WAN, the amount of bandwidth required is also minimized.

The reference architecture does not require backups for inactive data. As the Data Domain system already hosts secondary storage capacity, using it as a target for monthly or quarterly full backups of the same data is redundant. Instead, this inactive data is also replicated to a second, off-site Data Domain system for redundancy. Snapshots can be created and retained on a scheduled basis on either Data Domain

system to create recovery points to protect against scenarios such as accidental deletion or corruption.

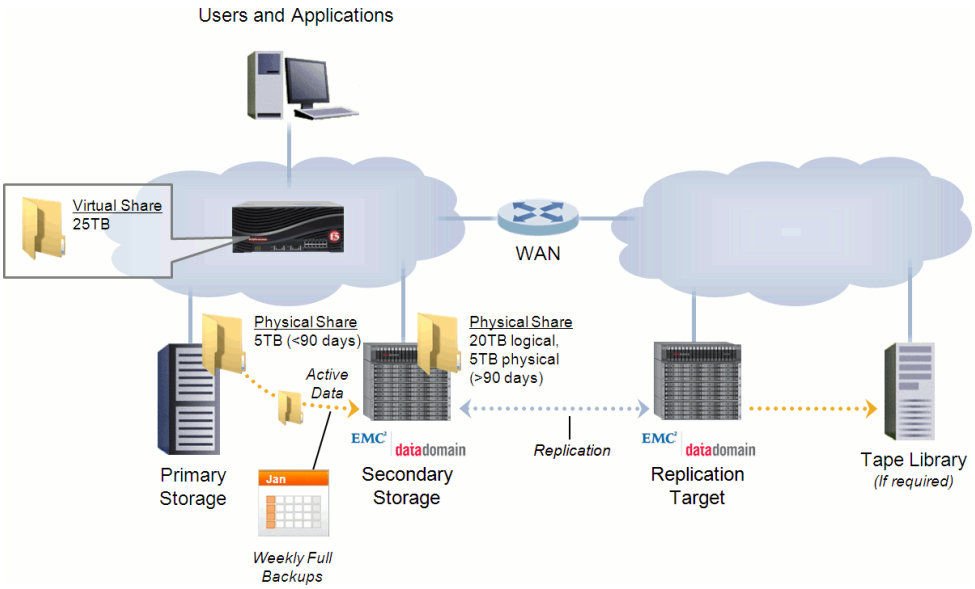

Figure 11. Optimizing backup in the reference architecture

Benefits. This backup strategy leverages systems already in place for primary and secondary storage and does not require using any tape media. For many organizations, this will provide adequate data protection as it ensures that all active and inactive data, as well as backups of active data, is stored in two separate locations. This approach facilitates consistent recovery points across the tiered storage environment.

Considerations. Some organizations are required to perform tape backups of certain data for regulatory or compliance reasons. Larger enterprises, or those mandating tape backups with offsite storage, may choose to perform full backups of the second Data Domain system on a monthly or quarterly basis. Because this system contains both inactive data and backups of active data, performing full backups of this system will ensure that all data under management can be placed on tape.

Conclusion

To maximize the efficiency of your IT infrastructure, it is critical to optimize the management and storage of data. Because different storage types have different characteristics, administrators must balance performance, availability and cost when selecting the appropriate storage for their environment. Two key criteria that can affect this optimization are where data is stored and in what form it is stored.

F5 ARX enables organizations to optimize their environment. ARX automatically and transparently moves file data between high-performance and low-costs storage tiers to take advantage of different performance, availability and cost characteristics of available storage systems.

Data Domain complements storage tiering with deduplication storage. Once files have been allocated to the proper tier, how those files are stored and data integrity protected further determines the efficacy of the IT infrastructure. Data Domain inline deduplication storage provides an average of 10-30x reduction in backup data storage needs while enabling rapid offsite replication via low-cost WAN to minimize time to DR.

Combining the use of ARX with Data Domain systems provides organizations with a simple, cost-effective and easy-to-use solution to realize compelling improvements in IT performance, efficiency, and disaster readiness, while seeing dramatic reductions in operational and storage costs.

Next Steps

For more information, contact your local F5 or EMC Backup and Recovery Systems sales representative and visit us online at:

http://www.f5.com/solutions/technology-alliances/infrastructure/data-domain.html

http://www.datadomain.com/solutions/f5.html

Data Domain, Inc.

2421 Mission College Blvd.
Santa Clara, CA 95054
+1-408-980-4800
(866) WE-DDUPE
sales@datadomain.com
www.datadomain.com

F5 Networks, Inc.
Corporate Headquarters
401 Elliott Avenue West
Seattle, WA 98119
(888) 88BIGIP Toll-free
+1-206-272-5555 Voice
+1-206-272-5556 Fax
www.f5.com

F5 Networks
Asia-Pacific
+65-6533-6103 Voice
+65-6533-6106 Fax
Info.asia@f5.com

F5 Networks Ltd.
Europe/Middle-East/Africa
+44 (0) 1932 582 000 Voice
+44 (0) 1932 582 001 Fax
emeainfo@f5.com

F5 Networks
Japan K.K.
+81-3-5114-3200 Voice
+81-3-5114-3201 Fax
info@f5networks.co.jp