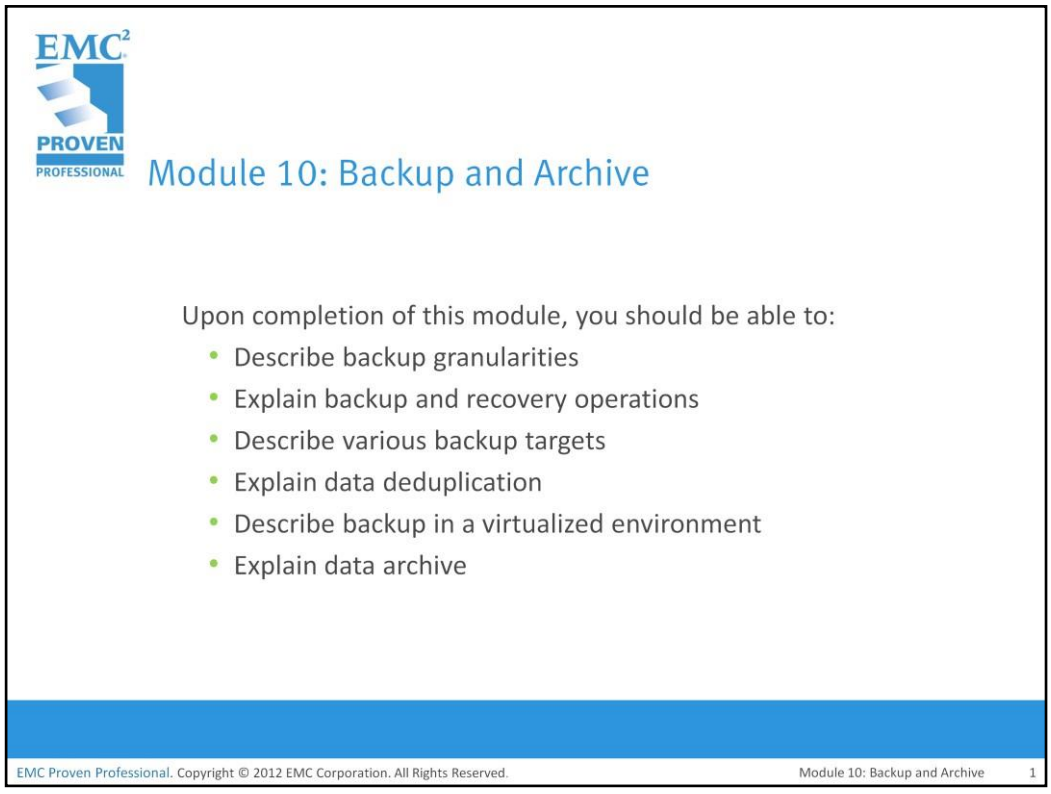
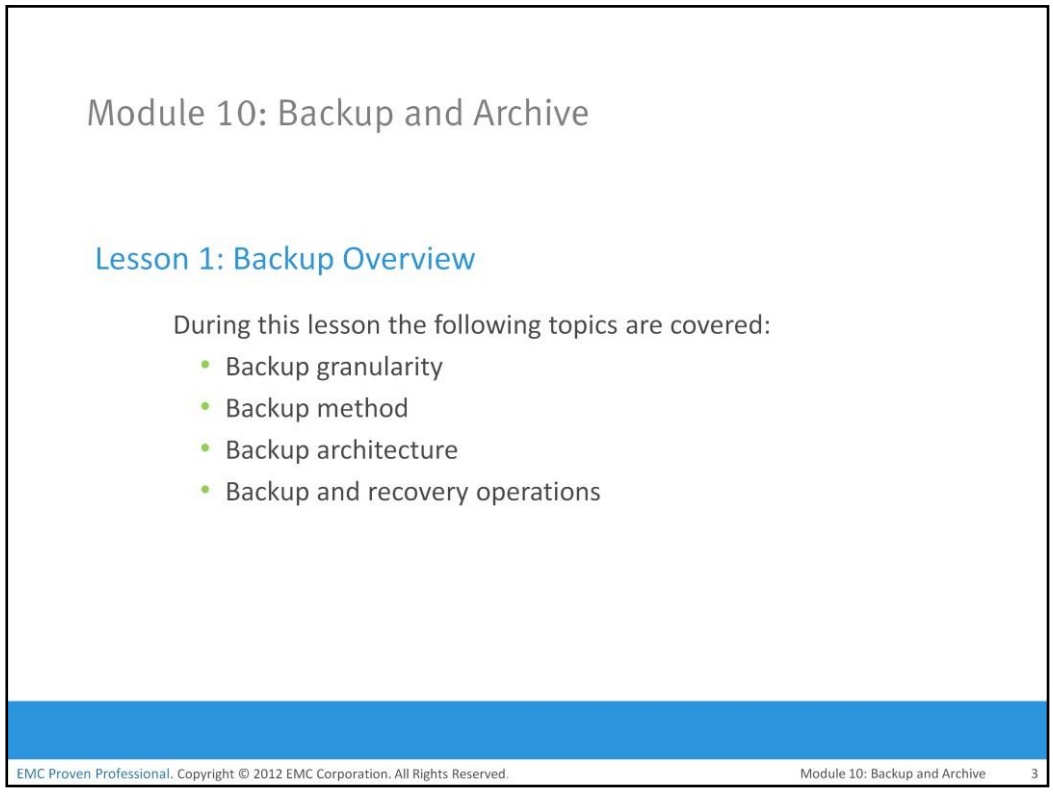
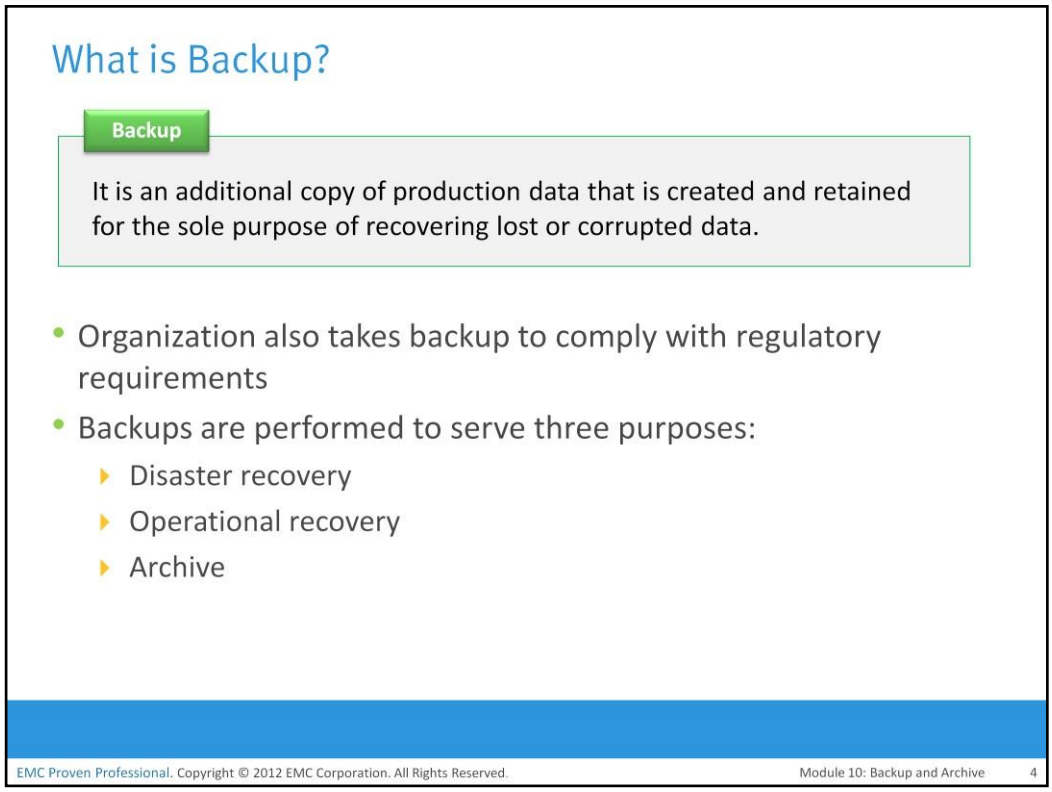
# Module – 10
# Backup and Archive




This module focuses on backup granularities and backup operations. This module also focuses on various backup targets and data deduplication. Further, this module details backup in a virtualized environment. Additionally, this module focuses on data archive.




This lesson covers various backup granularities and backup method. This lesson also covers backup architecture and operations.




A *backup* is an additional copy of production data, created and retained for the sole purpose of recovering lost or corrupted data. With growing business and regulatory demands for data storage, retention, and availability, organizations are faced with the task of backing up an ever-increasing amount of data. This task becomes more challenging with the growth of information, stagnant IT budgets, and less time for taking backups. Moreover, organizations need a quick restore of backed up data to meet business service-level agreements (SLAs).

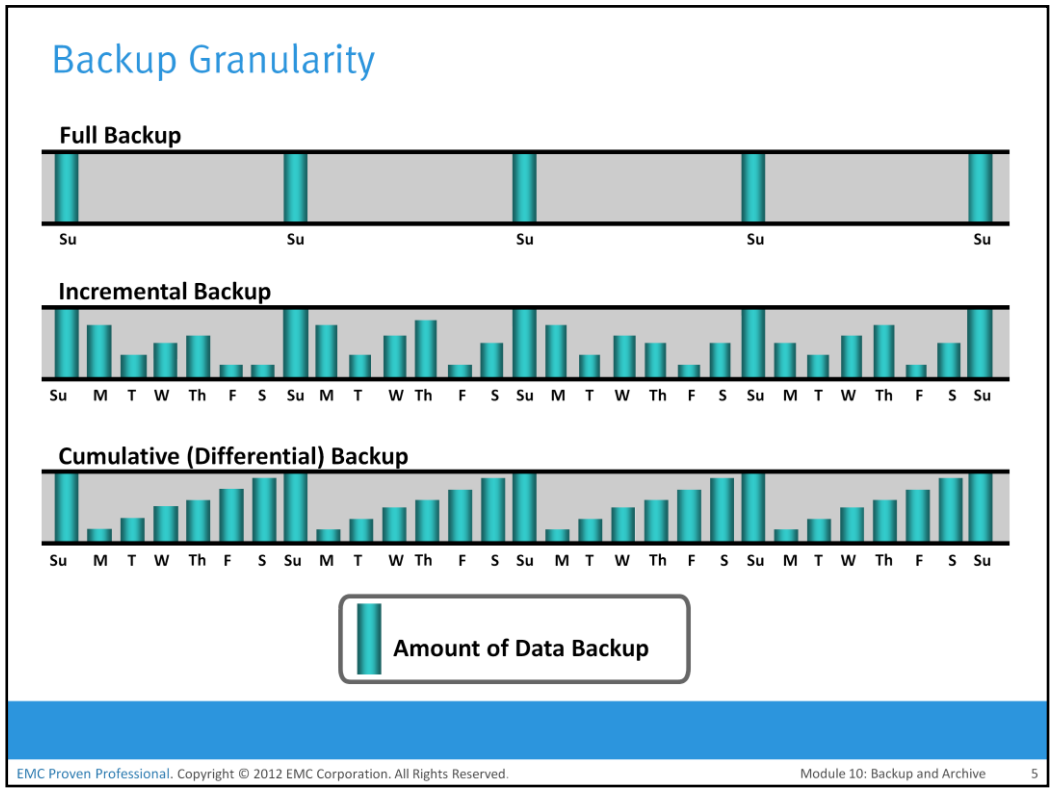
Backups are performed to serve three purposes: disaster recovery, operational recovery, and archival.

Backups can be performed to address disaster recovery needs. The backup copies are used for restoring data at an alternate site when the primary site is incapacitated due to a disaster. Based on Recovery-point objective (RPO) and Recovery-time objective (RTO) requirements, organizations use different data protection strategies for disaster recovery.

Data in the production environment changes with every business transaction and operation. Backups are used to restore data if data loss or logical corruption occurs during routine processing. The majority of restore requests in most organizations fall in this category. For example, it is common for a user to accidentally delete an important e-mail or for a file to become corrupted, which can be restored using backup data.

Backups are also performed to address archival requirements. Although content addressed storage (CAS) has emerged as the primary solution for archives (CAS is discussed in module 8), traditional backups are still used by small and medium enterprises for long-term preservation of transaction records, e-mail messages, and other business records required for regulatory compliance.

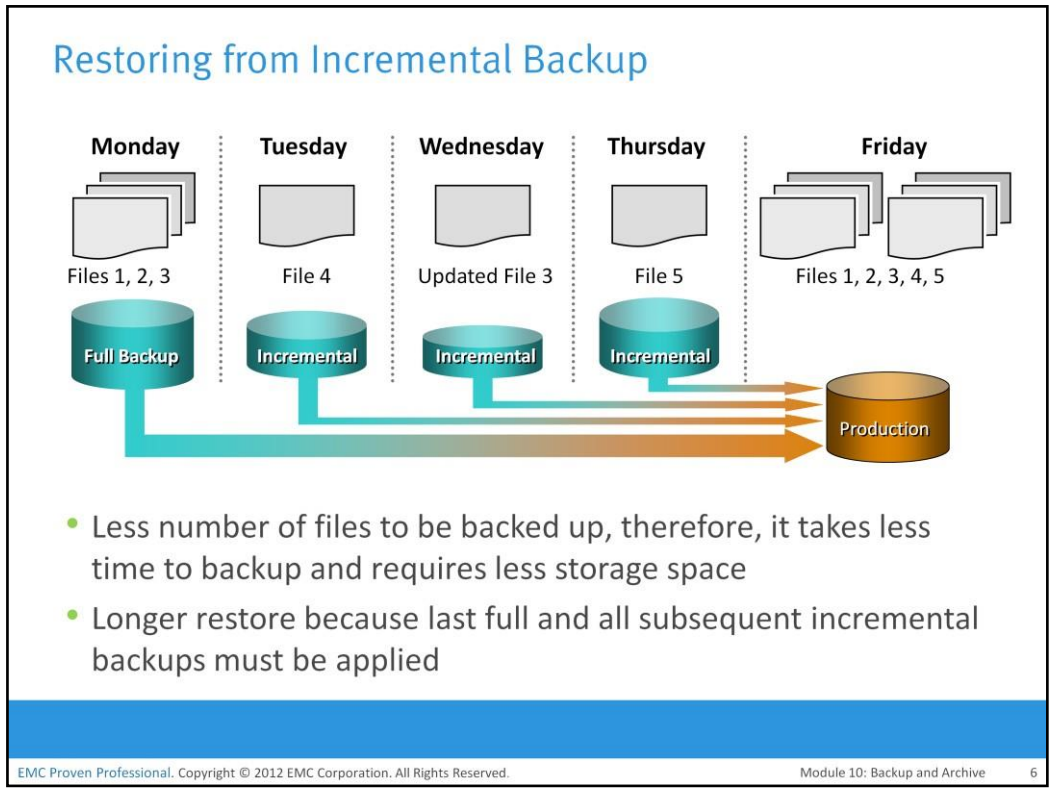
*Note:* Backup window is the period during which a source is available for performing a data backup.



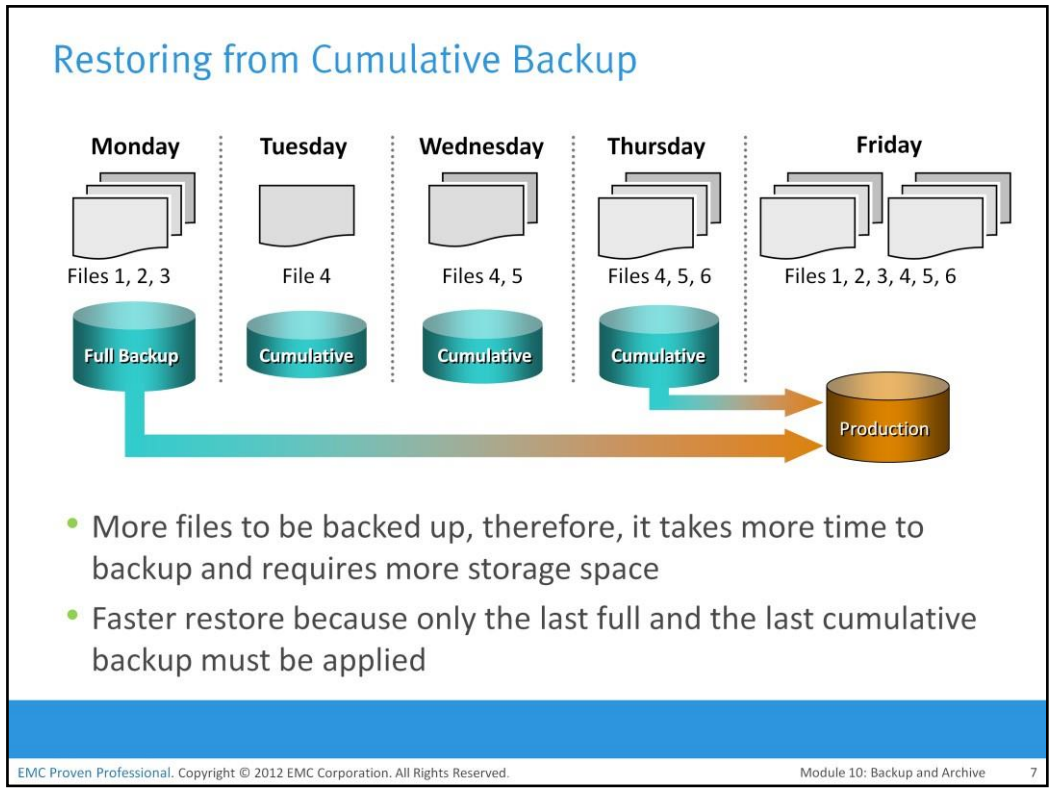

Backup granularity depends on business needs and the required RTO/RPO. Based on the granularity, backups can be categorized as full, incremental, and cumulative (or differential). Most organizations use a combination of these three backup types to meet their backup and recovery requirements. Figure on the slide depicts the different backup granularity levels.

*Full backup* is a backup of the complete data on the production volumes. A full backup copy is created by copying the data in the production volumes to a backup storage device. It provides a faster recovery but requires more storage space and also takes more time to back up. *Incremental backup* copies the data that has changed since the last full or incremental backup, whichever has occurred more recently. This is much faster than a full backup (because the volume of data backed up is restricted to the changed data only) but takes longer to restore. *Cumulative backup* copies the data that has changed since the last full backup. This method takes longer than an incremental backup but is faster to restore.

Another way to implement full backup is *synthetic* (or *constructed*) *backup*. This method is used when the production volume resources cannot be exclusively reserved for a backup process for extended periods to perform a full backup. It is usually created from the most recent full backup and all the incremental backups performed after that full backup. This backup is called *synthetic* because the backup is not created directly from production data. A synthetic full backup enables a full backup copy to be created offline without disrupting the I/O operation on the production volume. This also frees up network resources from the backup process, making them available for other production uses.



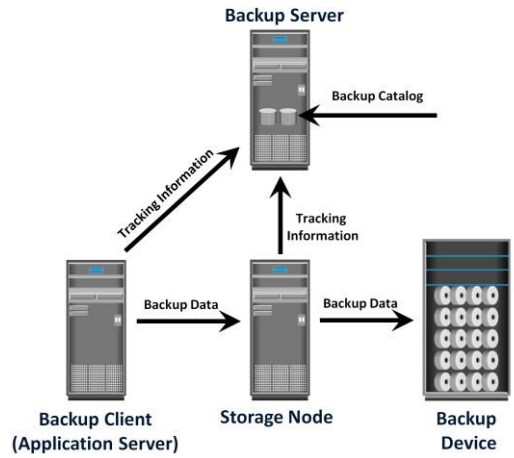

The process of restoration from an incremental backup requires the last full backup and all the incremental backups available until the point of restoration. Consider an example, a full backup is performed on Monday evening. Each day after that, an incremental backup is performed. On Tuesday, a new file (File 4 as shown in the figure) is added, and no other files have changed. Consequently, only File 4 is copied during the incremental backup performed on Tuesday evening. On Wednesday, no new files are added, but File 3 has been modified. Therefore, only the modified File 3 is copied during the incremental backup on Wednesday evening. Similarly, the incremental backup on Thursday copies only File 5. On Friday morning, there is data corruption, which requires data restoration from the backup. The first step toward data restoration is restoring all data from the full backup of Monday evening. The next step is applying the incremental backups of Tuesday, Wednesday, and Thursday. In this manner, data can be successfully recovered to its previous state, as it existed on Thursday evening.




Consider an example, a full backup of the business data is taken on Monday evening. Each day after that, a cumulative backup is taken. On Tuesday, File 4 is added and no other data is modified since the previous full backup of Monday evening. Consequently, the cumulative backup on Tuesday evening copies only File 4. On Wednesday, File 5 is added. The cumulative backup taking place on Wednesday evening copies both File 4 and File 5 because these files have been added or modified since the last full backup. Similarly, on Thursday, File 6 is added. Therefore, the cumulative backup on Thursday evening copies all three files: File 4, File 5, and File 6. On Friday morning, data corruption occurs that requires data restoration using backup copies. The first step in restoring data is to restore all the data from the full backup of Monday evening. The next step is to apply only the latest cumulative backup, which is taken on Thursday evening. In this way, the production data can be recovered faster because its needs only two copies of data—the last full backup and the latest cumulative backup.

## Backup Architecture

- Backup client
  - Gathers the data that is to be backed up and send it to storage node
- Backup server
  - Manages backup operations and maintains backup catalog
- Storage node
  - Responsible for writing data to backup device
  - Manages the backup device

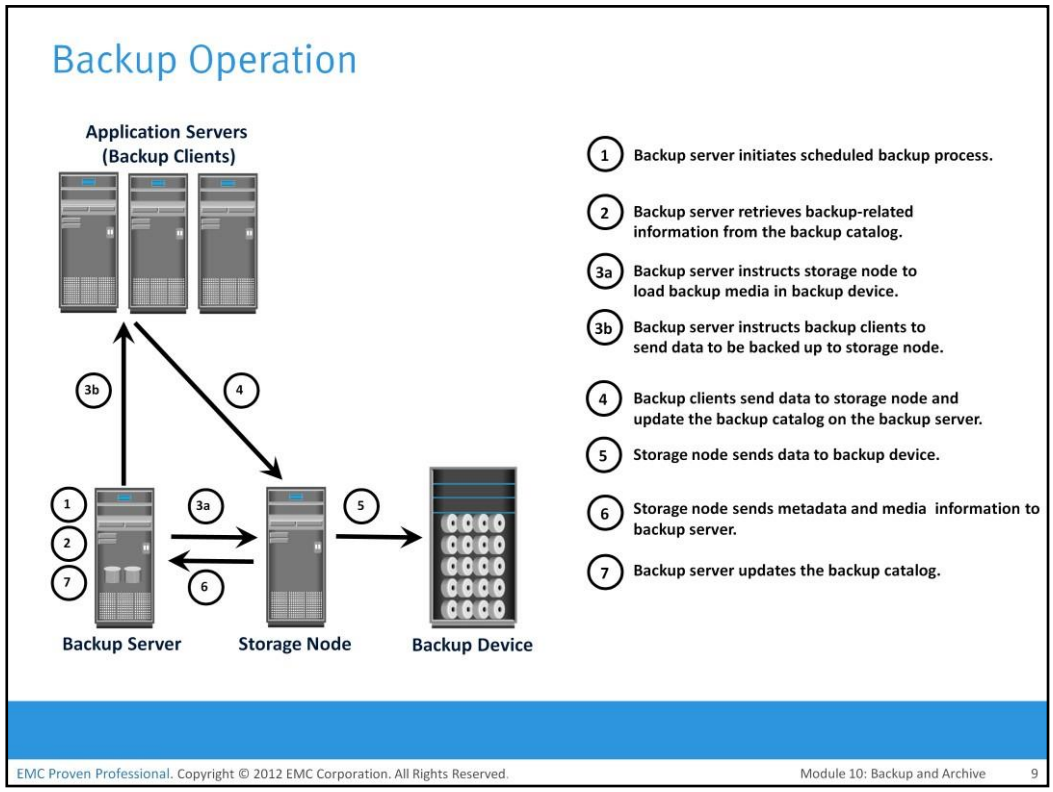






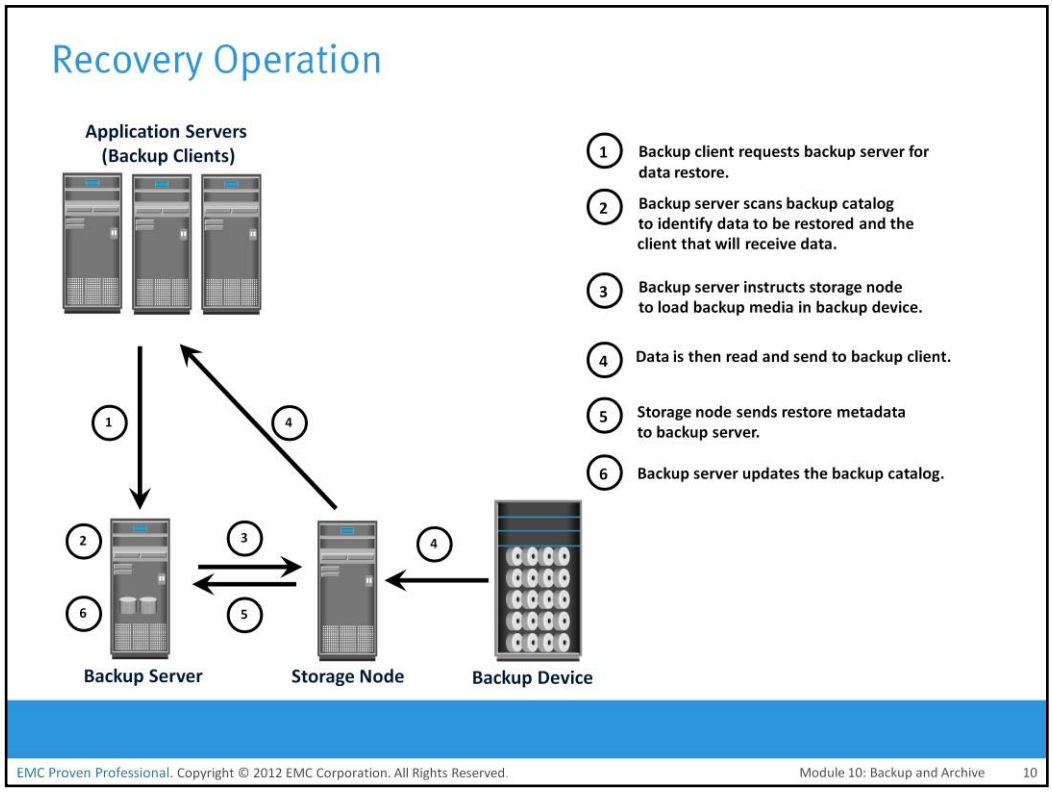
A backup system commonly uses the client-server architecture with a backup server and multiple backup clients. Figure on the slide illustrates the backup architecture. The backup server manages the backup operations and maintains the backup catalog, which contains information about the backup configuration and backup metadata. Backup configuration contains information about when to run backups, which client data to be backed up, and so on, and the backup metadata contains information about the backed up data. The role of a backup client is to gather the data that is to be backed up and send it to the storage node. It also sends the tracking information to the backup server.

The storage node is responsible for writing the data, to the backup device. In a backup environment, a *storage node* is a host that controls backup devices. The storage node also sends tracking information to the backup server. In many cases, the storage node is integrated with the backup server, and both are hosted on the same physical platform. A backup device is attached directly or through a network to the storage node's host platform. Some backup architecture refers the storage node as the *media server* because it manages the storage device.




When a backup operation is initiated, significant network communication takes place between the different components of a backup infrastructure. The backup operation is typically initiated by a server, but it can also be initiated by a client. The backup server initiates the backup process for different clients based on the backup schedule configured for them. For example, the backup for a group of clients may be scheduled to start at 3:00 a.m. every day.

The backup server coordinates the backup process with all the components in a backup environment. The backup server maintains the information about backup clients to be backed up and storage nodes to be used in a backup operation. The backup server retrieves the backup-related information from the backup catalog and, based on this information, instructs the storage node to load the appropriate backup media into the backup devices. Simultaneously, it instructs the backup clients to gather the data to be backed up and send it over the network to the assigned storage node. After the backup data is sent to the storage node, the client sends some backup metadata (the number of files, name of the files, storage node details, and so on) to the backup server. The storage node receives the client data, organizes it, and sends it to the backup device. The storage node then sends additional backup metadata (location of the data on the backup device, time of backup, and so on) to the backup server. The backup server updates the backup catalog with this information.

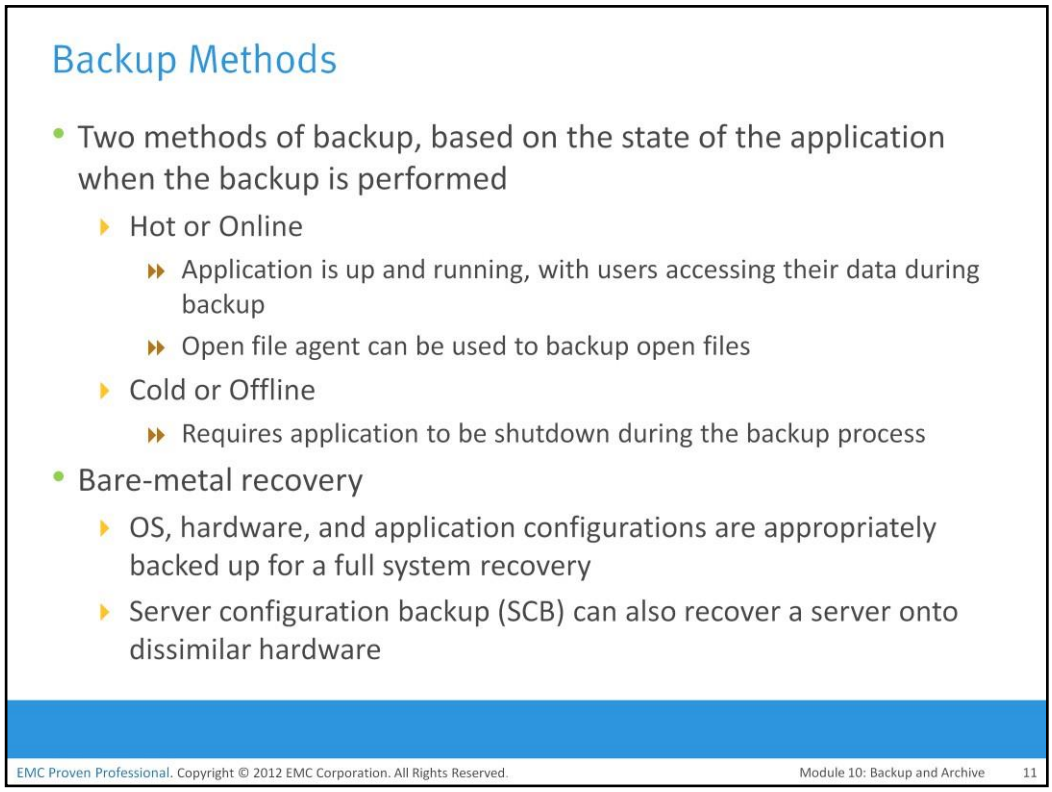



After the data is backed up, it can be restored when required. A restore process must be manually initiated from the client. Some backup software has a separate application for restore operations. These restore applications are usually accessible only to the administrators or backup operators. Figure on the slide depicts a restore operation.

Upon receiving a restore request, an administrator opens the restore application to view the list of clients that have been backed up. While selecting the client for which a restore request has been made, the administrator also needs to identify the client that will receive the restored data. Data can be restored on the same client for whom the restore request has been made or on any other client. The administrator then selects the data to be restored and the specified point in time to which the data has to be restored based on the RPO. Because all this information comes from the backup catalog, the restore application needs to communicate with the backup server.

The backup server instructs the appropriate storage node to mount the specific backup media onto the backup device. Data is then read and sent to the client that has been identified to receive the restored data.

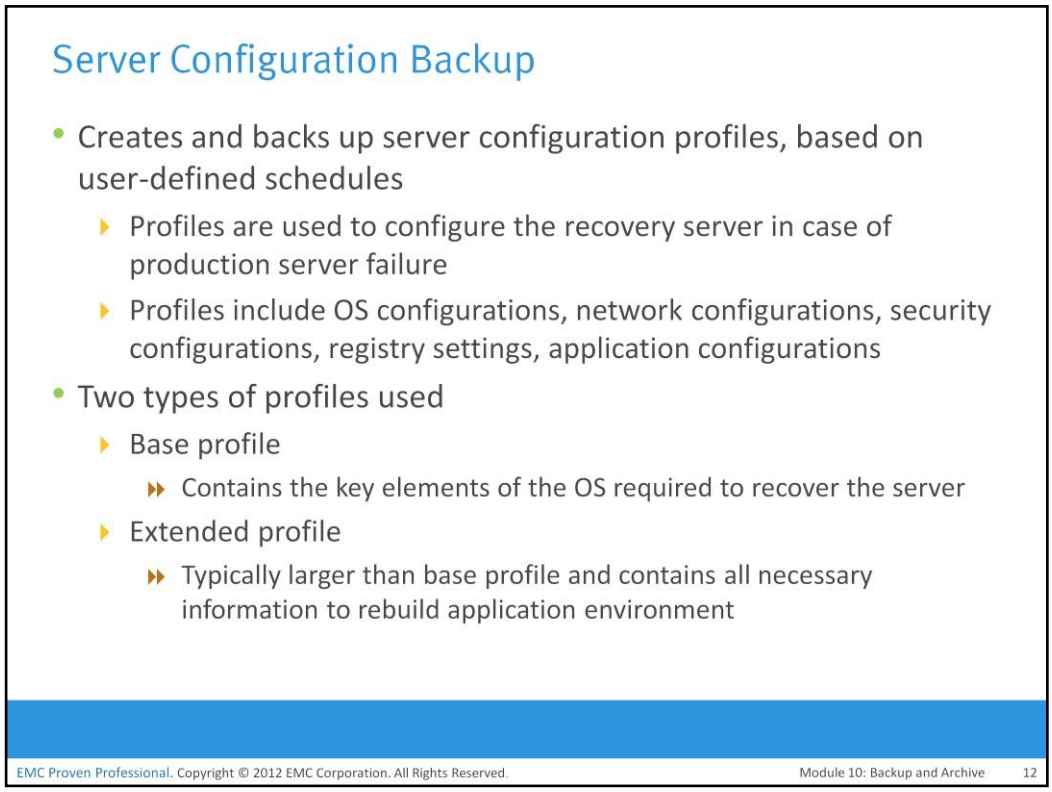
Some restorations are successfully accomplished by recovering only the requested production data. For example, the recovery process of a spreadsheet is completed when the specific file is restored. In database restorations, additional data, such as log files, must be restored along with the production data. This ensures consistency for the restored data. In these cases, the RTO is extended due to the additional steps in the restore operation.




Hot backup and cold backup are the two methods deployed for backup. They are based on the state of the application when the backup is performed. In a *hot backup*, the application is up-and-running, with users accessing their data during the backup process. This method of backup is also referred to as online backup. A *cold backup* requires the application to be shutdown during the backup process. Hence, this method is also referred to as offline backup.

The hot backup of online production data is challenging because data is actively being used and changed. If a file is open, it is normally not backed up during the backup process. In such situations, an *open file agent* is required to back up the open file. These agents interact directly with the operating system or application and enable the creation of consistent copies of open files. The disadvantage associated with a hot backup is that the agents usually affect the overall application performance. Consistent backups of databases can also be done by using a cold backup. This requires the database to remain inactive during the backup. Of course, the disadvantage of a cold backup is that the database is inaccessible to users during the backup process. All the files must be backed up in the same state for consistent backup of a database that comprises many files.

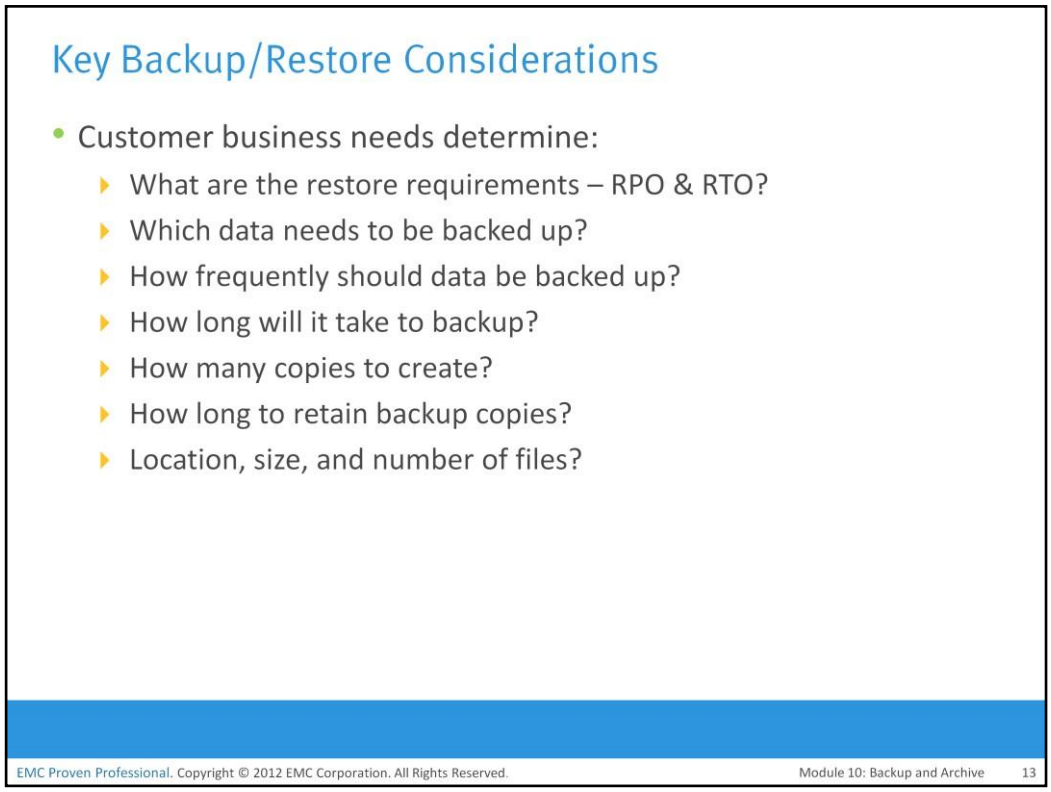
In a disaster recovery environment, *bare-metal recovery* (BMR) refers to a backup in which OS, hardware, and application configurations are appropriately backed up for a full system recovery. BMR builds the base system, which includes partitioning, the file system layout, the operating system, the applications, and all the relevant configurations. BMR recovers the base system first before starting the recovery of data files. Some BMR technologies—for example server configuration backup (SCB)—can recover a server even onto dissimilar hardware.




Most organizations spend a considerable amount of time and money protecting their application data but give less attention to protecting their server configurations. During disaster recovery, server configurations must be re-created before the application and data are accessible to the user. The process of system recovery involves reinstalling the operating system, applications, and server settings and then recovering the data. During a normal data backup operation, server configurations required for the system restore are not backed up. *Server configuration backup* (SCB) creates and backs up server configuration profiles based on user-defined schedules. The backed up profiles are used to configure the recovery server in case of production-server failure. SCB has the capability to recover a server onto dissimilar hardware.

In a server configuration backup, the process of taking a snapshot of the application server's configuration (both system and application configurations) is known as *profiling.* The profile data includes operating system configurations, network configurations, security configurations, registry settings, application configurations, and so on. Thus, profiling allows recovering the configuration of the failed system to a new server regardless of the underlying hardware.

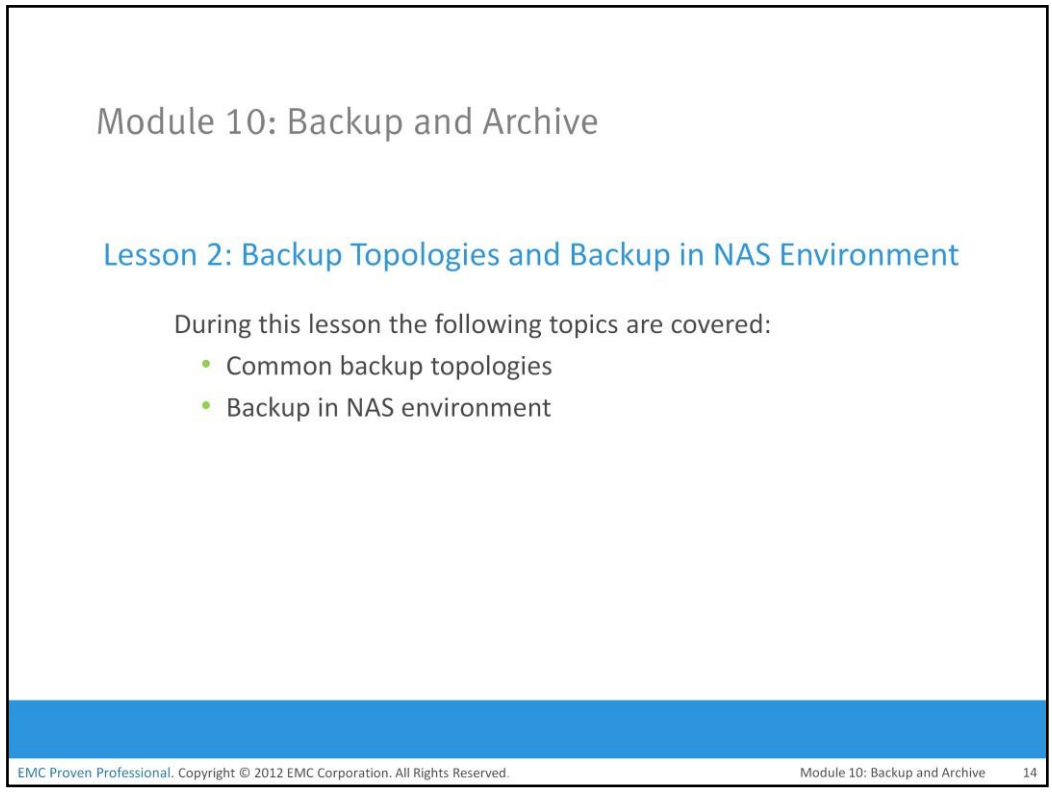
There are two types of profiles generated in the server configuration backup environment: base profile and extended profile. The base profile contains the key elements of the operating system required to recover the server. The extended profile is typically larger than the base profile and contains all the necessary information to rebuild the application environment.



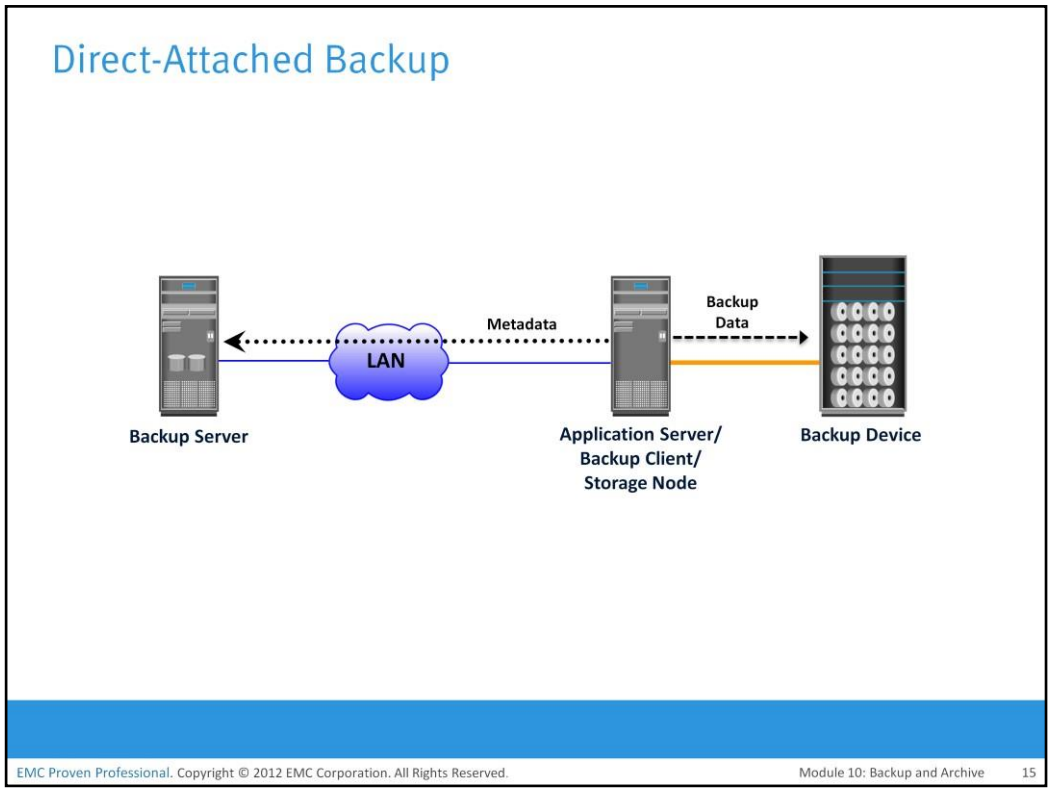

The amount of data loss and downtime that a business can endure in terms of RPO and RTO are the primary considerations in selecting and implementing a specific backup strategy. The RPO determines backup frequency. For example, if an application requires an RPO of 1 day, it would need the data to be backed up at least once every day. Another consideration is the retention period, which defines the duration for which a business needs to retain the backup copies.

The backup media type or backup target is another consideration that is driven by RTO and impacts the data recovery time. Organizations must also consider the granularity of backups. The development of a backup strategy must include a decision about the most appropriate time for performing a backup to minimize any disruption to production operations. The size, number of files and data compression should also be considered because they might affect the backup process. Backing up large-size files (for example, ten 1 MB files) takes less time, compared to backing up an equal amount of data composed of small-size files (for example, ten thousand 1 KB files). Data compression and data deduplication (discussed later in the module) are widely used in the backup environment because these technologies save space on the media.

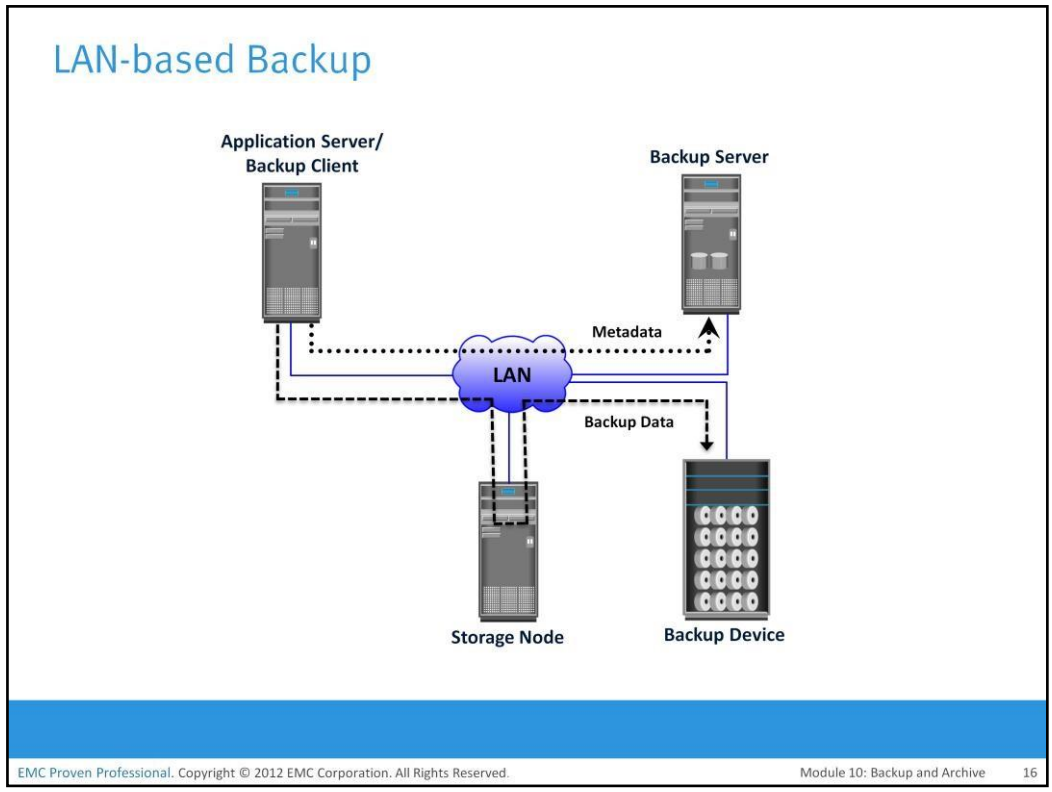
Location is an important consideration for the data to be backed up. Many organizations have dozens of heterogeneous platforms locally and remotely supporting their business. The backup process must address these sources for transactional and content integrity.



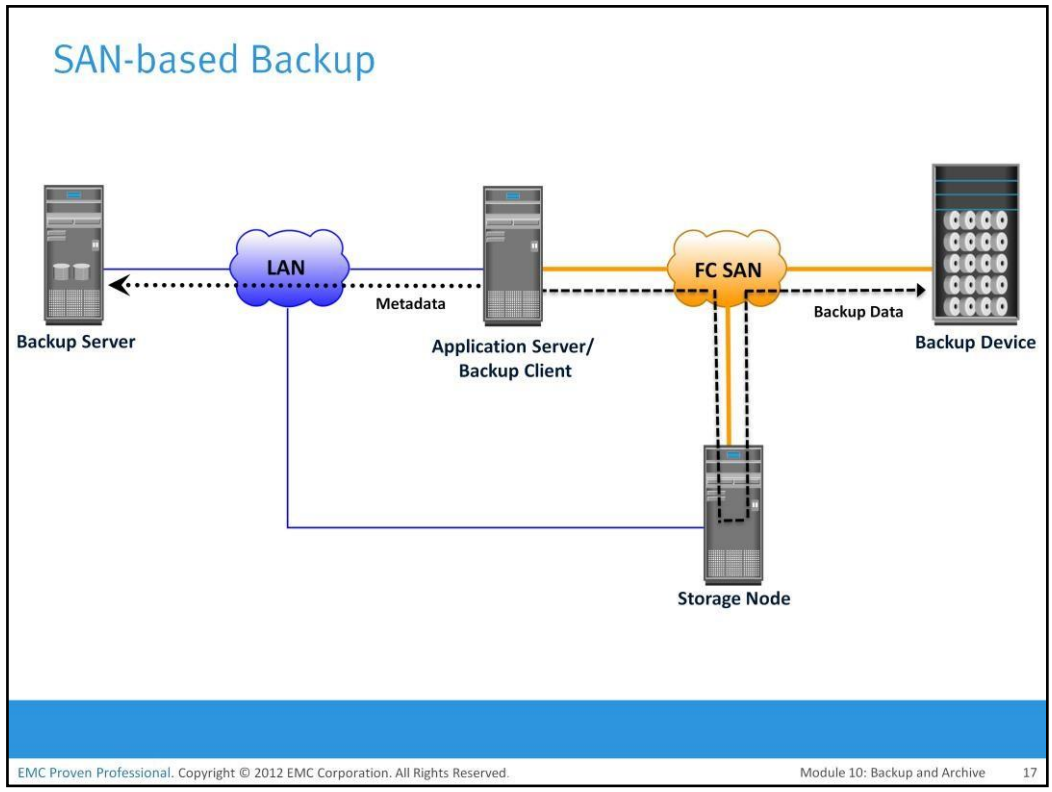

This lesson covers various backup topologies such as Direct-attached, LAN-based, SAN-based and mixed backup. This lesson also covers backup in NAS environment.




In a *direct-attached backup*, the storage node is configured on a backup client, and the backup device is attached directly to the client. Only the metadata is sent to the backup server through the LAN. This configuration frees the LAN from backup traffic. As the environment grows, there will be a need for centralized management and sharing of backup devices to optimize costs. An appropriate solution is required to share the backup devices among multiple servers. Network-based topologies (LAN-based and SAN-based) provide the solution to optimize the utilization of backup devices.

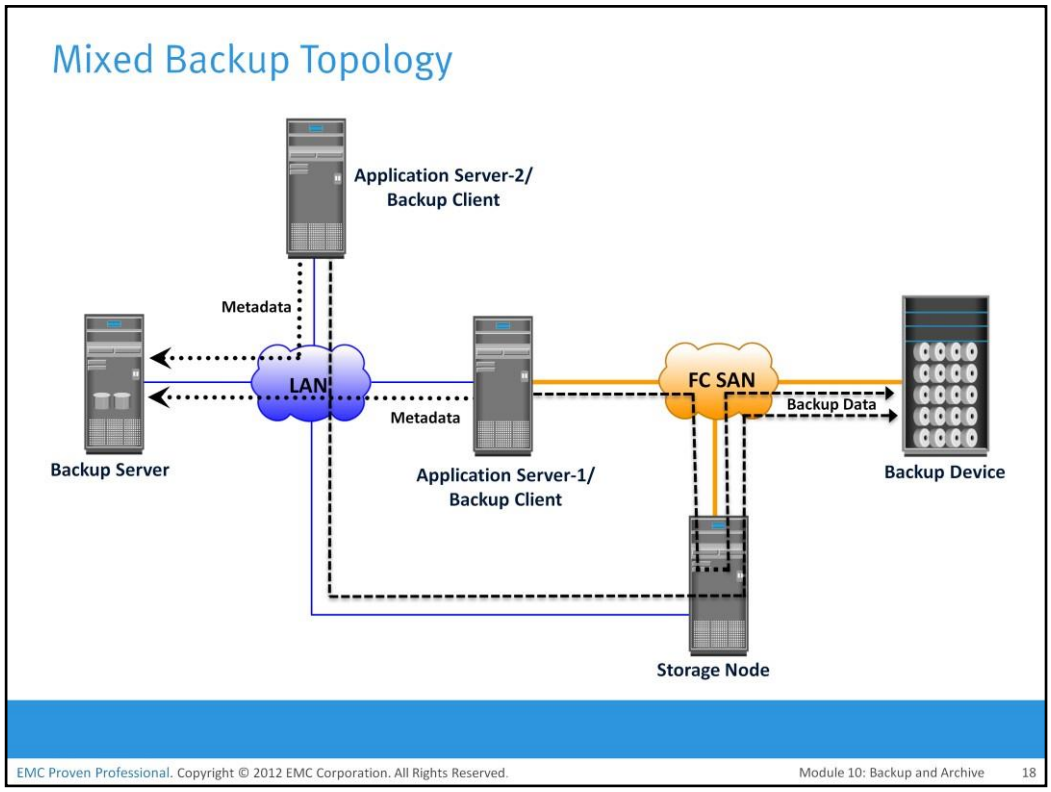



In a *LAN-based backup*, the clients, backup server, storage node, and backup device are connected to the LAN. The data to be backed up is transferred from the backup client (source) to the backup device (destination) over the LAN, which might affect network performance. This impact can be minimized by adopting a number of measures, such as configuring separate networks for backup and installing dedicated storage nodes for some application servers.

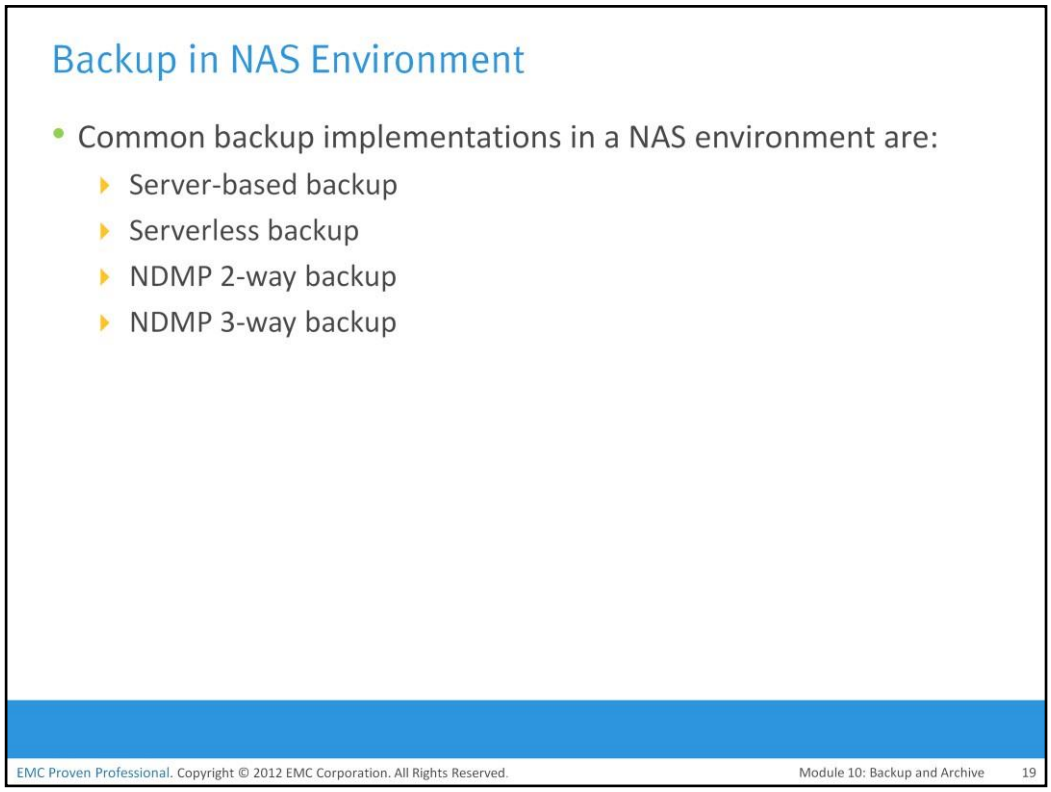



A *SAN-based backup* is also known as a *LAN-free backup*. The SAN-based backup topology is the most appropriate solution when a backup device needs to be shared among clients. In this case, the backup device and clients are attached to the SAN. In the figure shown on the slide, a client sends the data to be backed up to the backup device over the SAN. Therefore, the backup data traffic is restricted to the SAN, and only the backup metadata is transported over the LAN. The volume of metadata is insignificant when compared to the production data; the LAN performance is not degraded in this configuration.

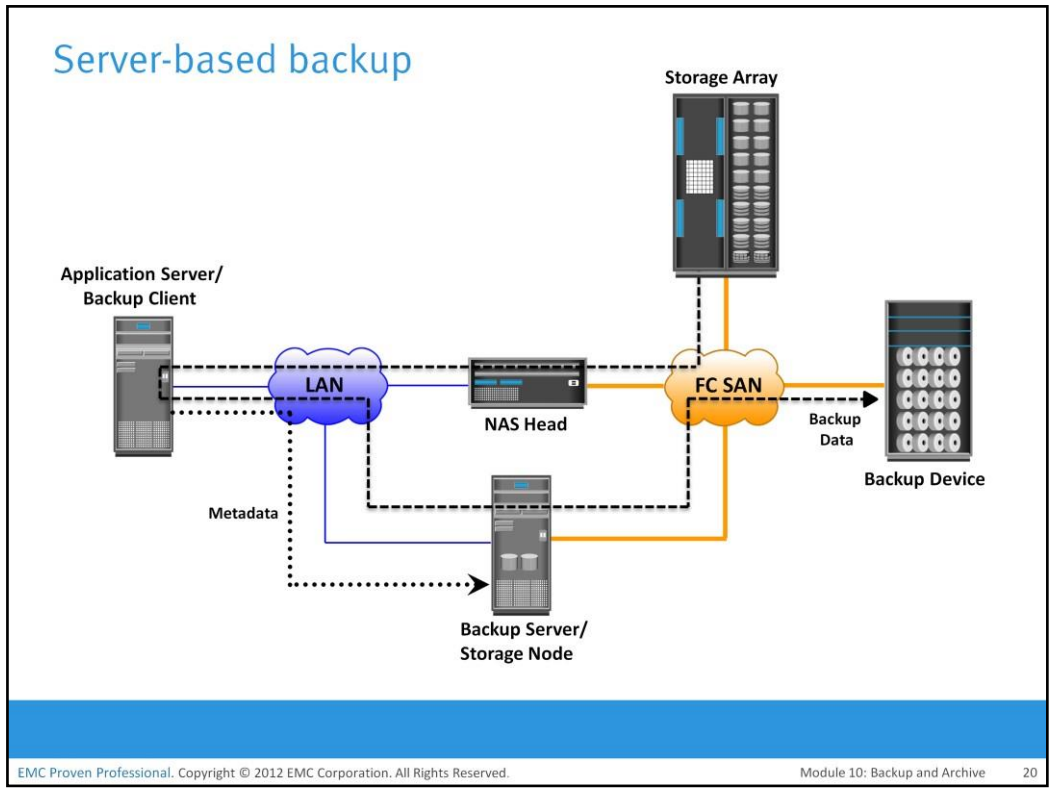
The emergence of low-cost disks as a backup medium has enabled disk arrays to be attached to the SAN and used as backup devices. A tape backup of these data backups on the disks can be created and shipped offsite for disaster recovery and long-term retention.



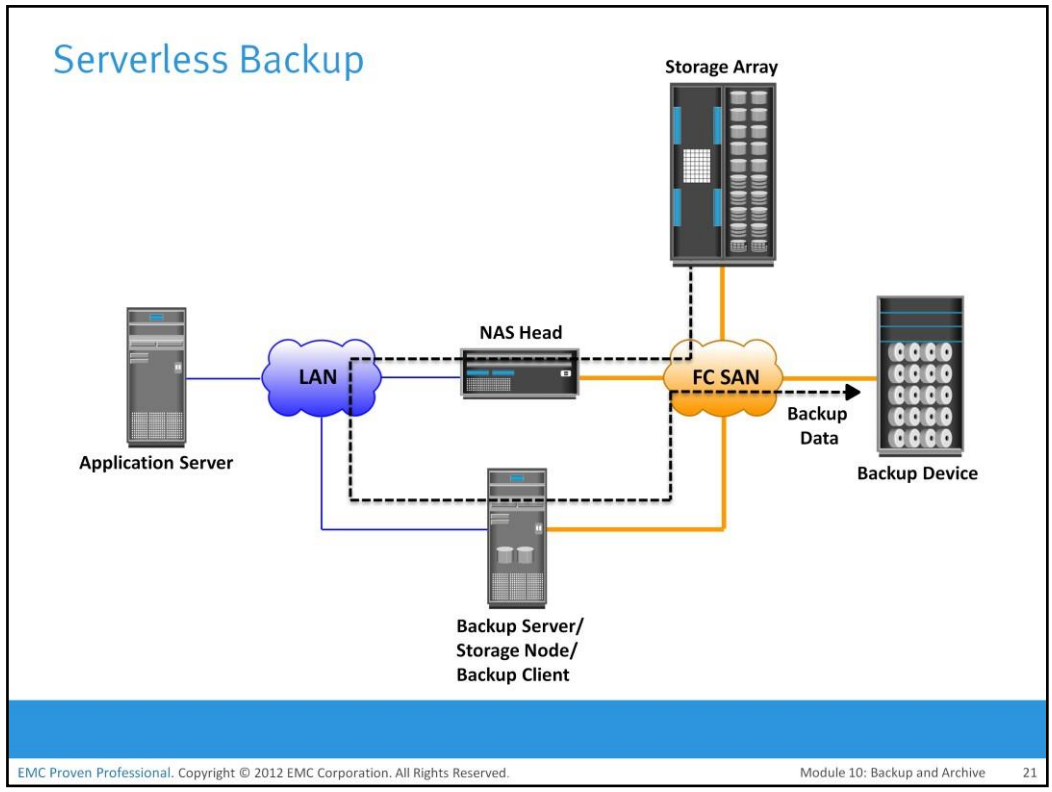

The *mixed topology* uses both the LAN-based and SAN-based topologies. This topology might be implemented for several reasons, including cost, server location, reduction in administrative overhead, and performance considerations.



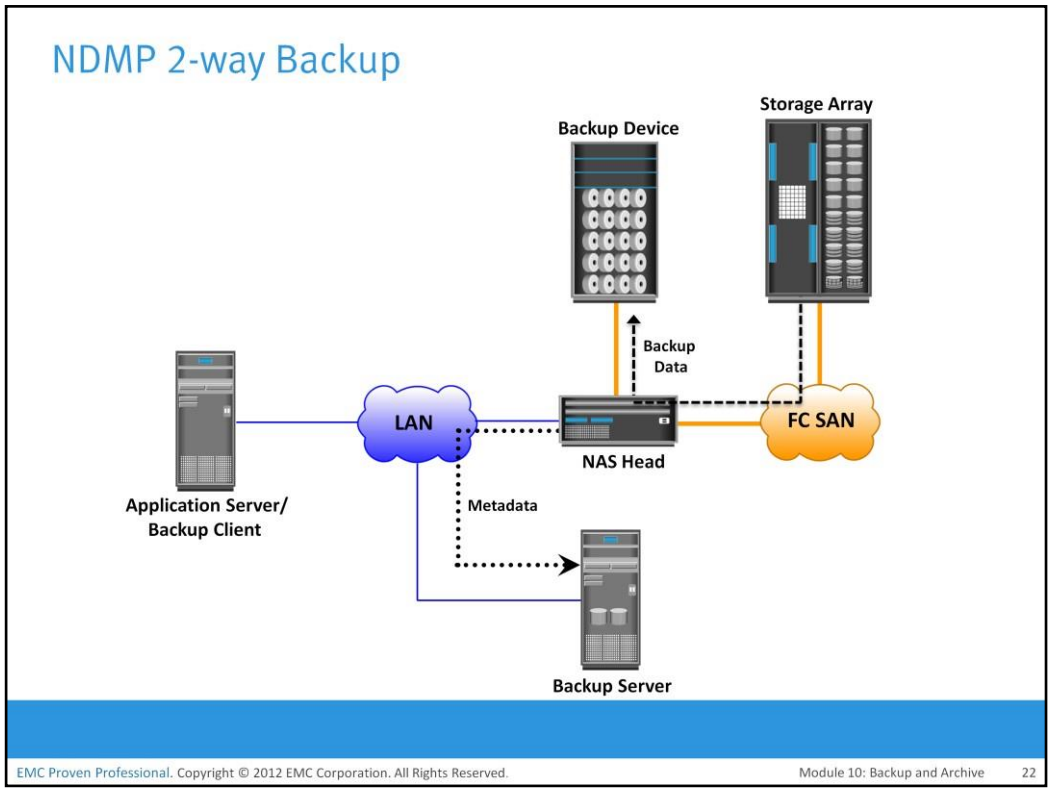

The use of a NAS head imposes a new set of considerations on the backup and recovery strategy in NAS environments. NAS heads use a proprietary operating system and file system structure that supports multiple file-sharing protocols. In the NAS environment, backups can be implemented in different ways: server-based, serverless, or using Network Data Management Protocol (NDMP). Common implementations are NDMP 2-way and NDMP 3-way.




In an *application server-based backup*, the NAS head retrieves data from a storage array over the network and transfers it to the backup client running on the application server. The backup client sends this data to the storage node, which in turn writes the data to the backup device. This results in overloading the network with the backup data and using application server resources to move the backup data.

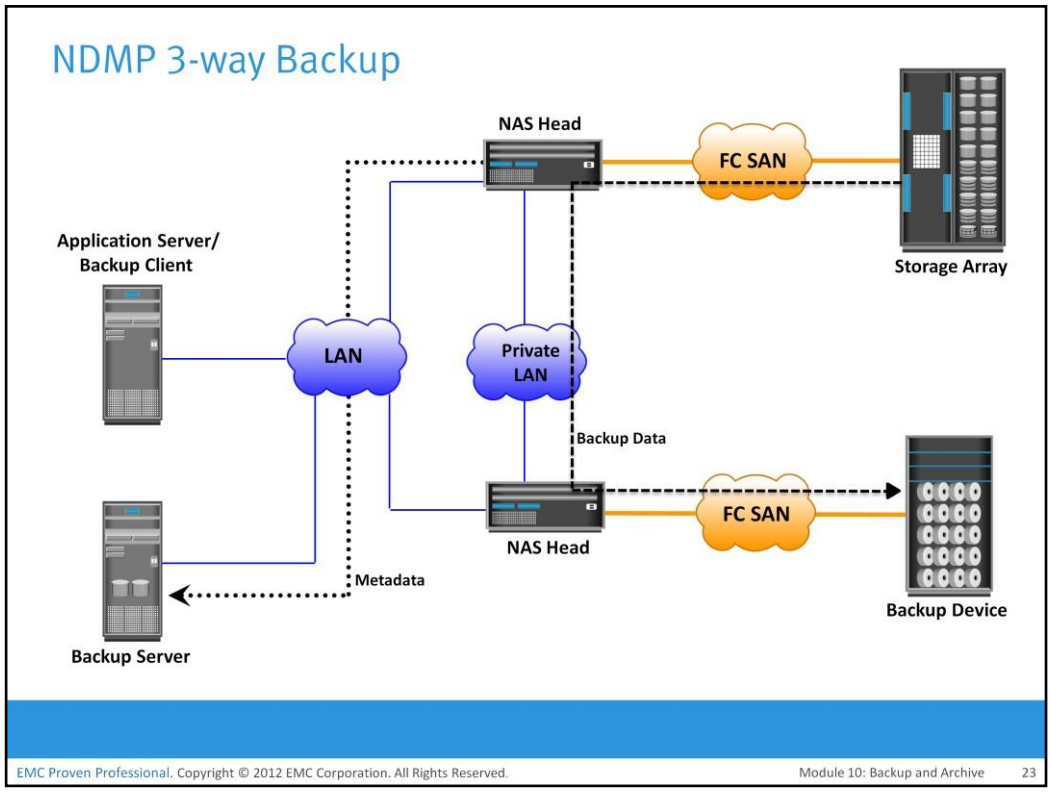



In a *serverless backup*, the network share is mounted directly on the storage node. This avoids overloading the network during the backup process and eliminates the need to use resources on the application server. In this scenario, the storage node, which is also a backup client, reads the data from the NAS head and writes it to the backup device without involving the application server. Compared to the previous solution, this eliminates one network hop.

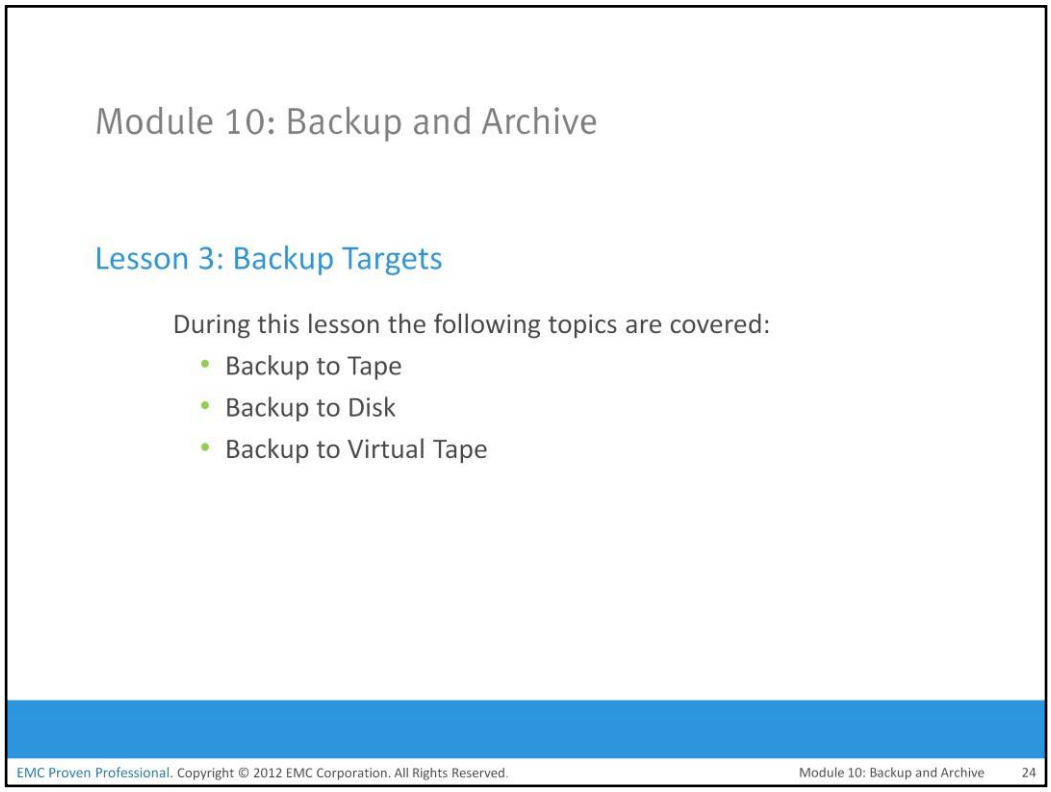



*NDMP* is an industry-standard TCP/IP-based protocol specifically designed for a backup in a NAS environment. It communicates with several elements in the backup environment (NAS head, backup devices, backup server, and so on) for data transfer and enables vendors to use a common protocol for the backup architecture. Data can be backed up using NDMP regardless of the operating system or platform. Due to its flexibility, it is no longer necessary to transport data through the application server, which reduces the load on the application server and improves the backup speed. NDMP optimizes backup and restore by leveraging the high-speed connection between the backup devices and the NAS head. In NDMP, backup data is sent directly from the NAS head to the backup device, whereas metadata is sent to the backup server.

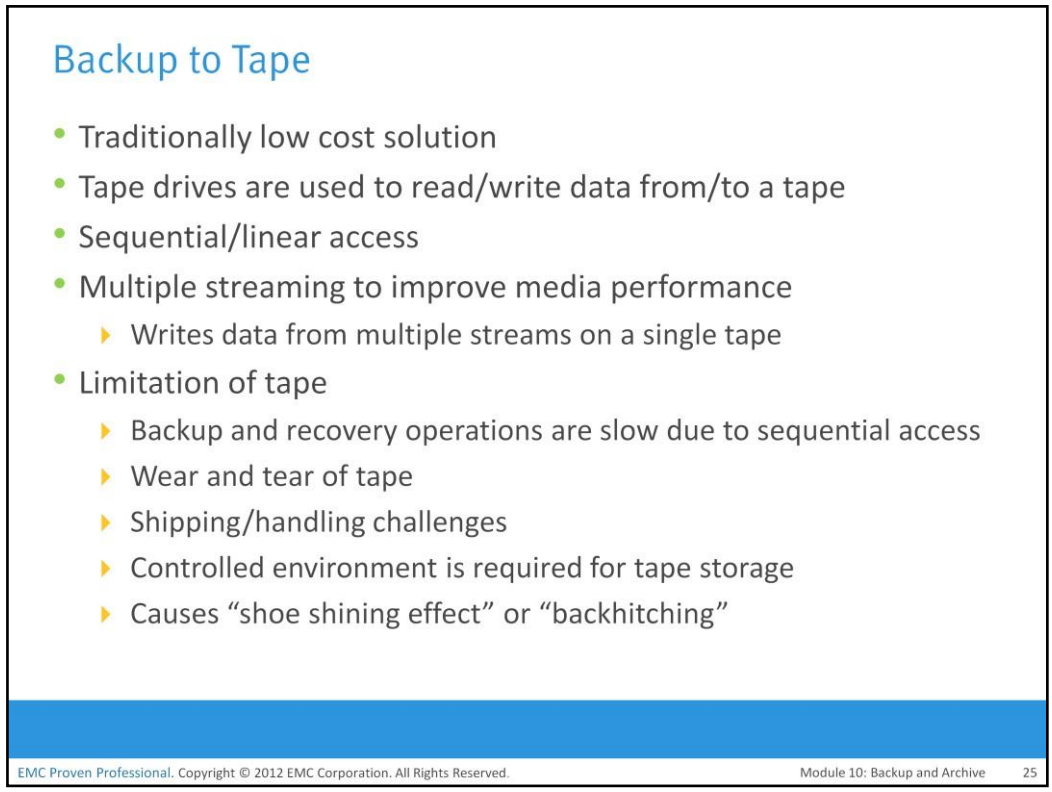
Figure on the slide illustrates backup in the NAS environment using NDMP 2-way. In this model, network traffic is minimized by isolating data movement from the NAS head to the locally attached backup device. Only metadata is transported on the network. The backup device is dedicated to the NAS device, and hence, this method does not support centralized management of all backup devices.




In the *NDMP 3-way* method, to avoid the backup data traveling on the production LAN, a separate private backup network must be established between all NAS heads and the NAS head connected to the backup device. Metadata and NDMP control data are still transferred across the public network. Figure on the slide depicts NDMP 3-way backup. An NDMP 3-way is useful when backup devices need to be shared among NAS heads. It enables the NAS head to control the backup device and share it with other NAS heads by receiving the backup data through the NDMP.



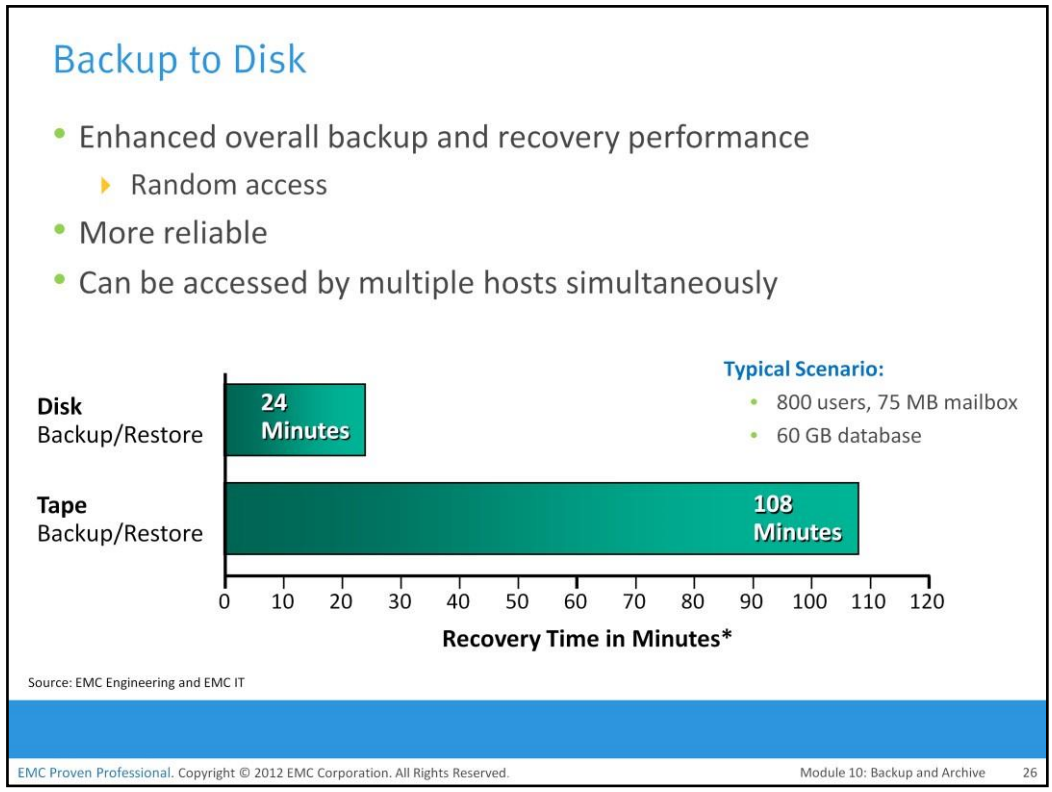

This lesson covers various backup targets such as physical tape, disk and virtual tape.




Tapes, a low-cost solution, are used extensively for backup. Tape drives are used to read/write data from/to a tape cartridge (or cassette). Tape drives are referred to as sequential, or linear, access devices because the data is written or read sequentially. A tape cartridge is composed of magnetic tapes in a plastic enclosure. Tape Mounting is the process of inserting a tape cartridge into a tape drive. The tape drive has motorized controls to move the magnetic tape around, enabling the head to read or write data.

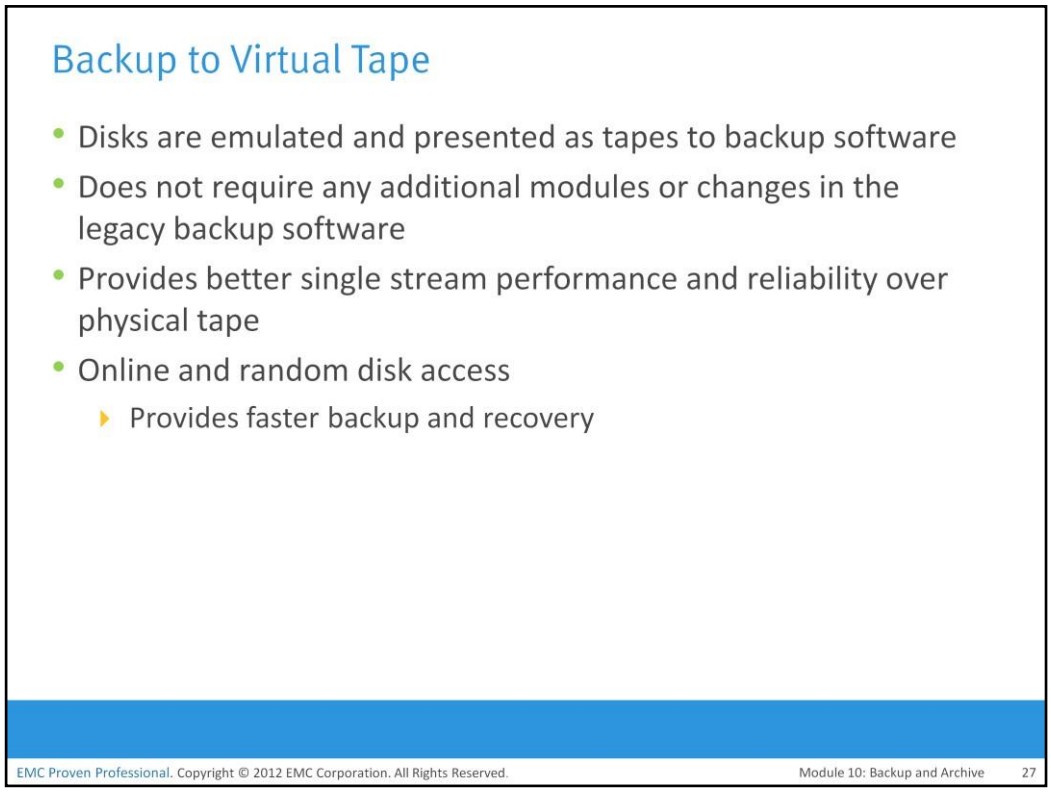
Tape drive *streaming* or *multiple streaming* writes data from multiple streams on a single tape to keep the drive busy. Multiple streaming improves media performance, but it has an associated disadvantage. The backup data is interleaved because data from multiple streams is written on it. Consequently, the data recovery time is increased because all the extra data from the other streams must be read and discarded while recovering a single stream.

Data access in a tape is sequential, which can slow backup and recovery operations. Tapes are primarily used for long-term offsite storage because of their low cost. Tapes must be stored in locations with a controlled environment to ensure preservation of the media and to prevent data corruption. Tapes are highly susceptible to wear and tear and usually have shorter shelf life. Physical transportation of the tapes to offsite locations also adds to management overhead and increases the possibility of loss of tapes during offsite shipment. Many times, even the buffering and speed adjustment features of a tape drive fail to prevent the gaps, causing the "*shoe shining effect"* or "*backhitching*." *Shoe shining* is the repeated back and forth motion a tape drive makes when there is an interruption in the backup data stream. This repeated back-and-forth motion not only causes a degradation of service, but also excessive wear and tear to tapes.

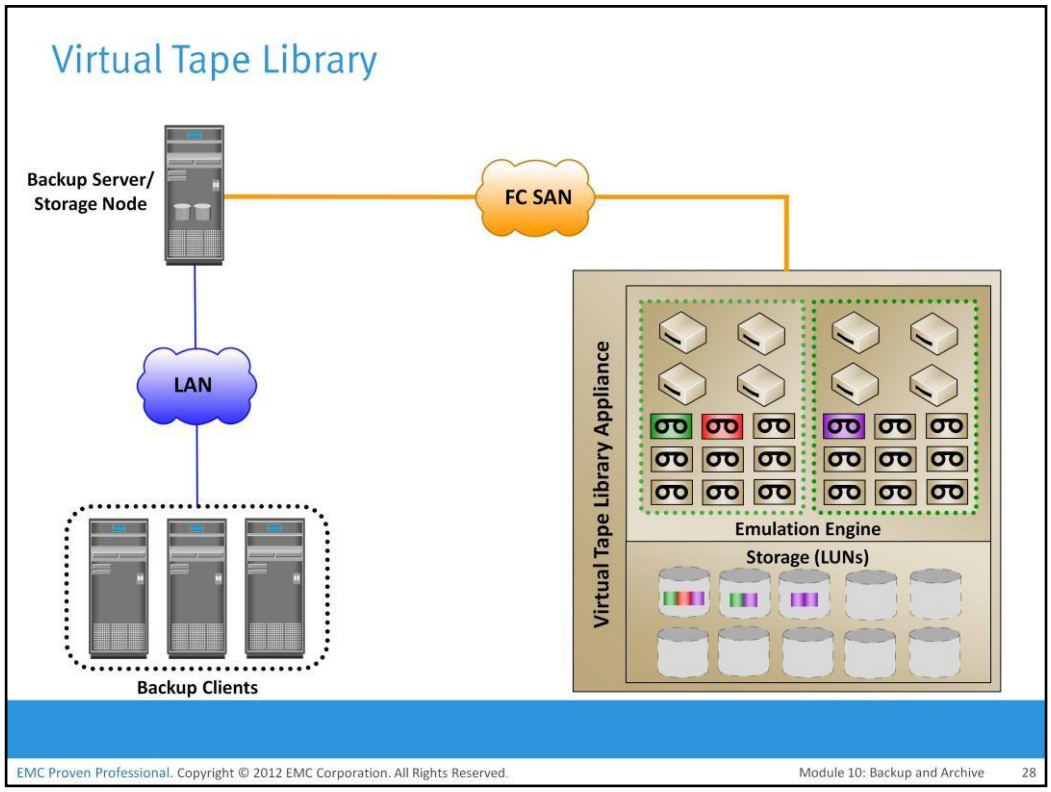



Availability of low cost disks have now replaced tapes as the primary device for storing backup data because of their performance advantages. Backup-to-disk systems offer ease of implementation, reduced TCO, and improved quality of service. Apart from performance benefits in terms of data transfer rates, disks also offer faster recovery when compared to tapes.

Backing up to disk storage systems offers clear advantages due to their inherent random access and RAID-protection capabilities. In most backup environments, backup to disk is used as a staging area where the data is copied temporarily before transferring or staging it to tapes. This enhances backup performance. Some backup products allow for backup images to remain on the disk for a period of time even after they have been staged. This enables a much faster restore. Figure on the slide illustrates a recovery scenario comparing tape versus disk in a Microsoft Exchange environment that supports 800 users with a 75 MB mailbox size and a 60 GB database. As shown in the figure, a restore from the disk took 24 minutes compared to the restore from a tape, which took 108 minutes for the same environment.



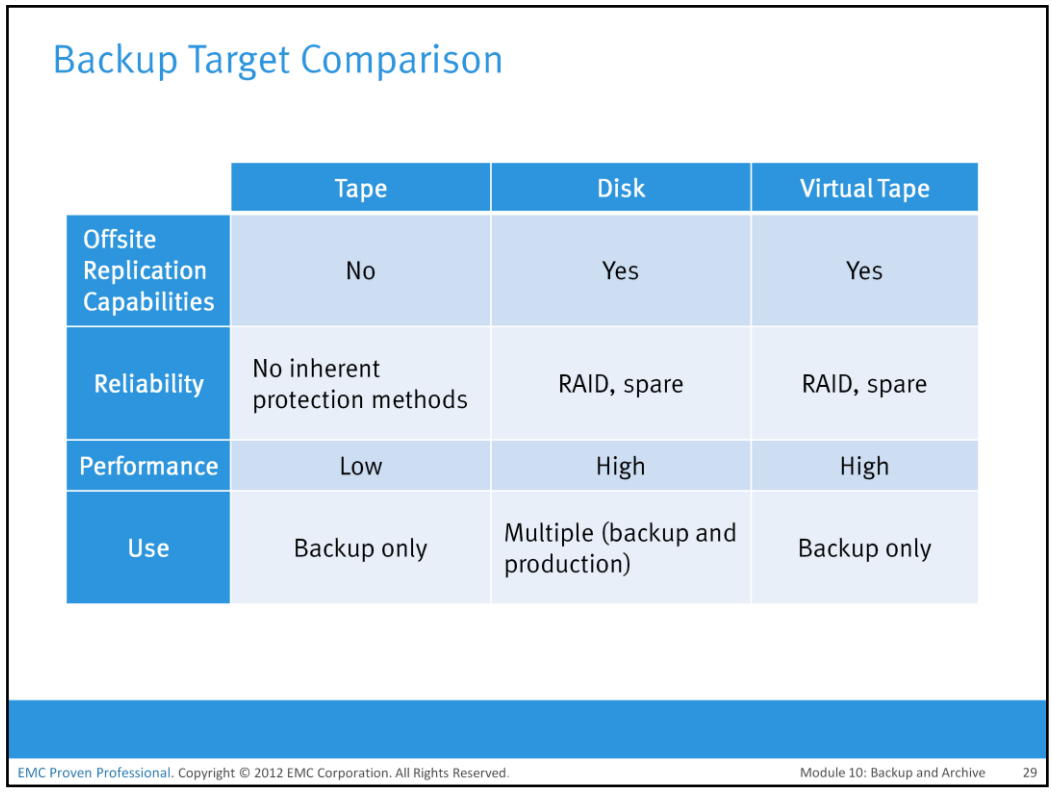

*Virtual tapes* are disk drives emulated and presented as tapes to the backup software. The key benefit of using a virtual tape is that it does not require any additional modules, configuration, or changes in the legacy backup software. This preserves the investment made in the backup software. Compared to physical tapes, virtual tapes offer better single stream performance, better reliability, and random disk access characteristics. Backup and restore operations are benefited from the disk's random access characteristics because they are online and provide faster backup and recovery. A virtual tape drive does not require the usual maintenance tasks associated with a physical tape drive, such as periodic cleaning and drive calibration. Compared to backup-to-disk devices, a virtual tape library offers easy installation and administration because it is preconfigured by the manufacturer.



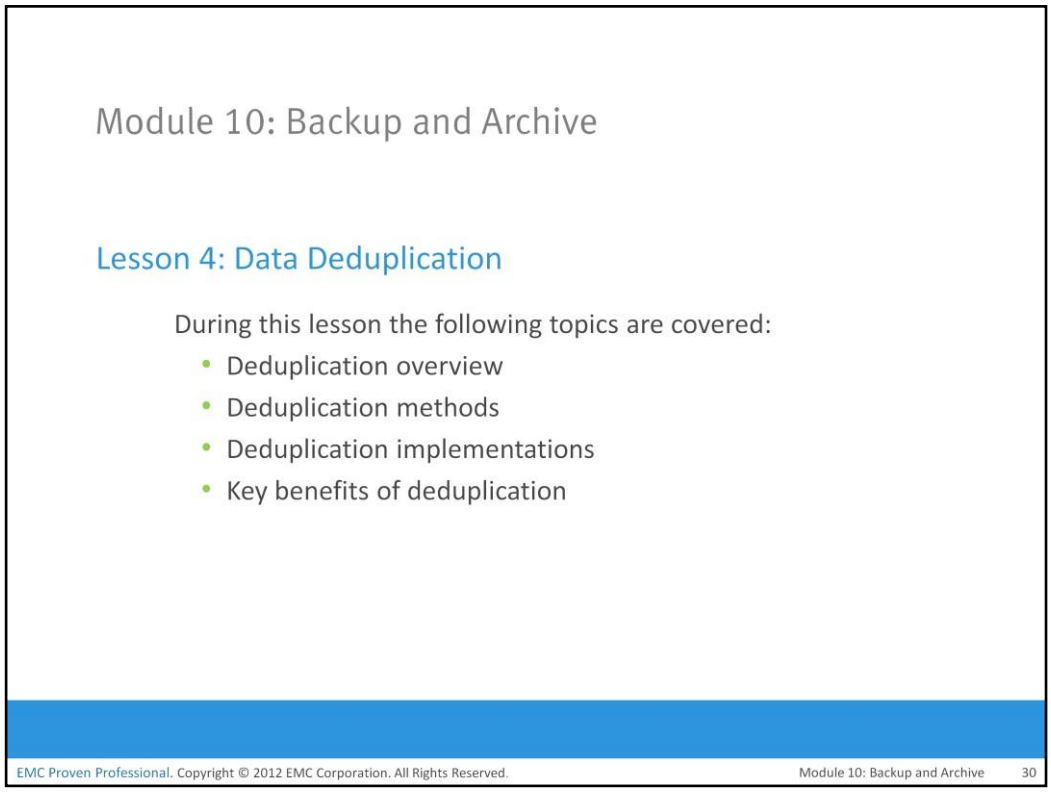

A *virtual tape library* (VTL) has the same components as that of a physical tape library, except that the majority of the components are presented as virtual resources. For the backup software, there is no difference between a physical tape library and a virtual tape library. Figure on the slide shows a virtual tape library. Virtual tape libraries use disks as backup media. Emulation software has a database with a list of virtual tapes, and each virtual tape is assigned space on a LUN. A virtual tape can span multiple LUNs if required. File system awareness is not required while backing up because the virtual tape solution typically uses raw devices.

Similar to a physical tape library, a robot mount is virtually performed when a backup process starts in a virtual tape library. However, unlike a physical tape library, where this process involves some mechanical delays, in a virtual tape library it is almost instantaneous. Even the *load to ready* time is much less than a physical tape library. After the virtual tape is mounted and the virtual tape drive is positioned, the virtual tape is ready to be used, and backup data can be written to it. In most cases, data is written to the virtual tape immediately. Unlike a physical tape library, the virtual tape library is not constrained by the sequential access and shoe shining effect. When the operation is complete, the backup software issues a rewind command. This rewind is also instantaneous. The virtual tape is then unmounted, and the virtual robotic arm is instructed to move it back to a virtual slot.

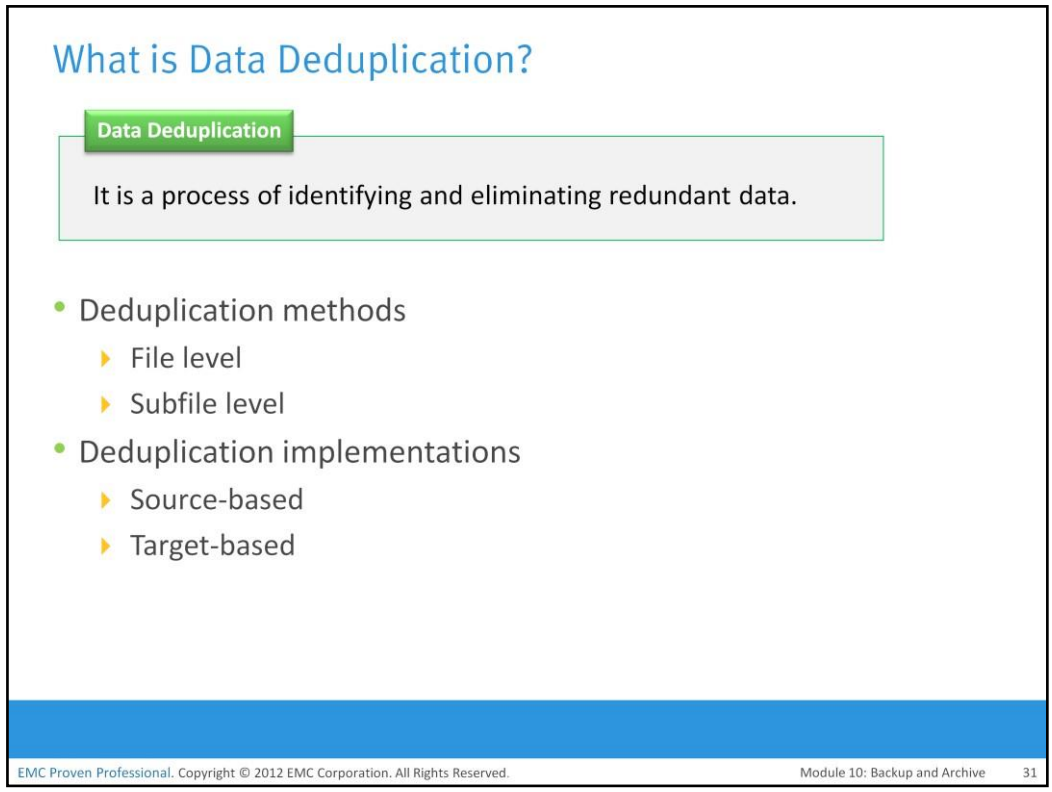
The steps to restore data are similar to those in a physical tape library, but the restore operation is nearly instantaneous. Even though virtual tapes are based on disks, which provide random access, they still emulate the tape behavior.




The table on the slide provides the comparison among various backup targets.




This lesson covers different deduplication method. This lesson also covers deduplication implementation such as source-based deduplication and target-based deduplication. Further this lesson details various key benefits of data deduplication.

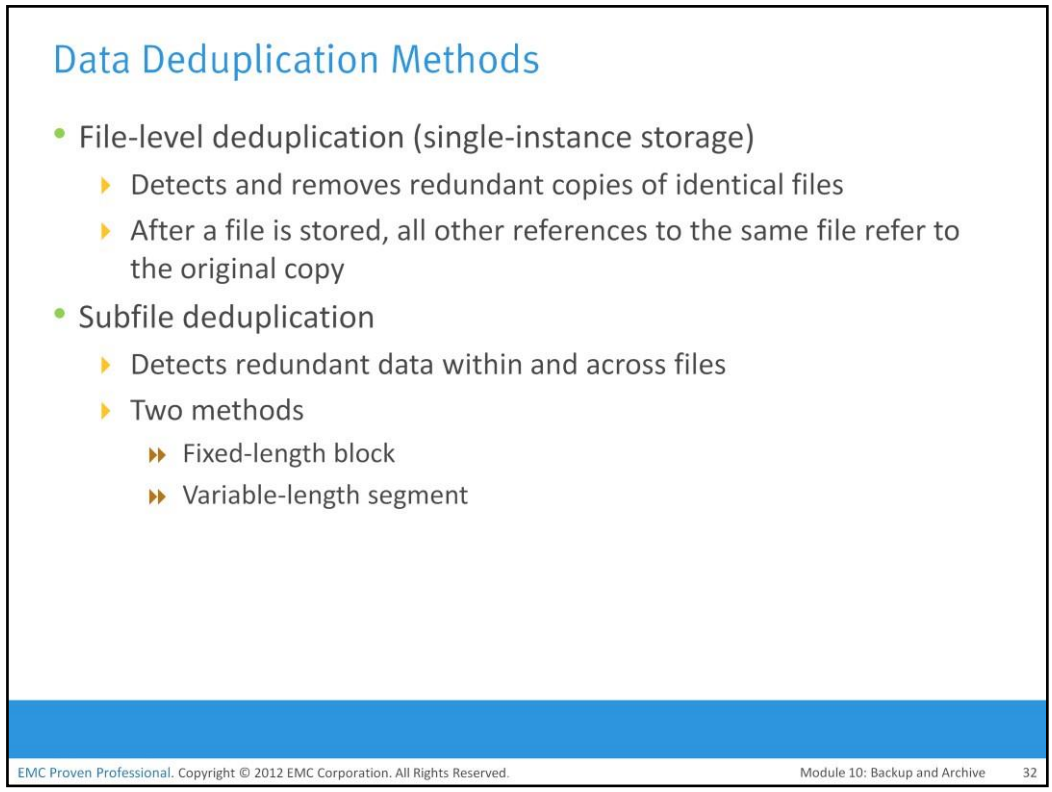



Traditional backup solutions do not provide any inherent capability to prevent duplicate data from being backed up. With the growth of information and 24x7 application availability requirements, backup windows are shrinking. Traditional backup processes back up a lot of duplicate data. Backing up duplicate data significantly increases the backup window size requirements and results in unnecessary consumption of resources, such as storage space and network bandwidth.

*Data deduplication* is the process of identifying and eliminating redundant data. When duplicate data is detected during backup, the data is discarded and only the pointer is created to refer the copy of the data that is already backed up. Data deduplication helps to reduce the storage requirement for backup, shorten the backup window, and remove the network burden. It also helps to store more backups on the disk and retain the data on the disk for a longer time.

There are two methods of data deduplication, file level and subfile level. Determining the uniqueness by implementing either method offers benefits; however, results can vary. The differences exist in the amount of data reduction each method produces and the time each approach takes to determine the unique content.

Deduplication can occur close to where the data is created, which is often referred to as “Source-based Deduplication”. It can also occur close to where the data is stored, which is commonly called “Target-based Deduplication”.

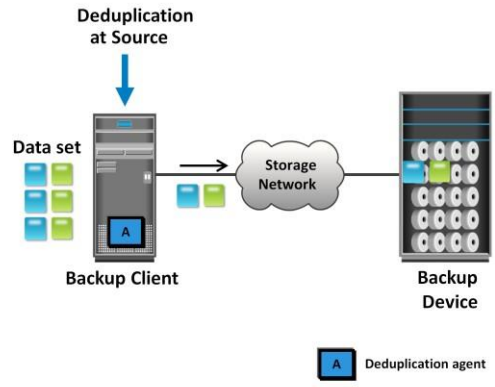



*File-level deduplication* (also called *single-instance storage*) detects and removes redundant copies of identical files. It enables storing only one copy of the file; the subsequent copies are replaced with a pointer that points to the original file. File-level deduplication is simple and fast but does not address the problem of duplicate content inside the files. For example, two 10-MB PowerPoint presentations with a difference in just the title page are not considered as duplicate files, and each file will be stored separately.

*Subfile deduplication* breaks the file into smaller chunks and then uses specialized algorithm to detect redundant data within and across the file. As a result, subfile deduplication eliminates duplicate data across files. There are two forms of subfile deduplication: fixed-length block and variable-length segment. The *fixed-length block deduplication* divides the files into fixed-length blocks and uses a hash algorithm to find the duplicate data. Although simple in design, fixed-length blocks might miss many opportunities to discover redundant data because the block boundary of similar data might be different. Consider the addition of a person's name to a document's title page. This shifts the whole document, and all the blocks appear to have changed, causing the failure of the deduplication method to detect equivalencies. In variable-length segment deduplication, if there is a change in the segment, the boundary for only that segment is adjusted, leaving the remaining segments unchanged. This method vastly improves the ability to find duplicate data segments compared to fixed-block.

## Data Deduplication Implementation – Source-based

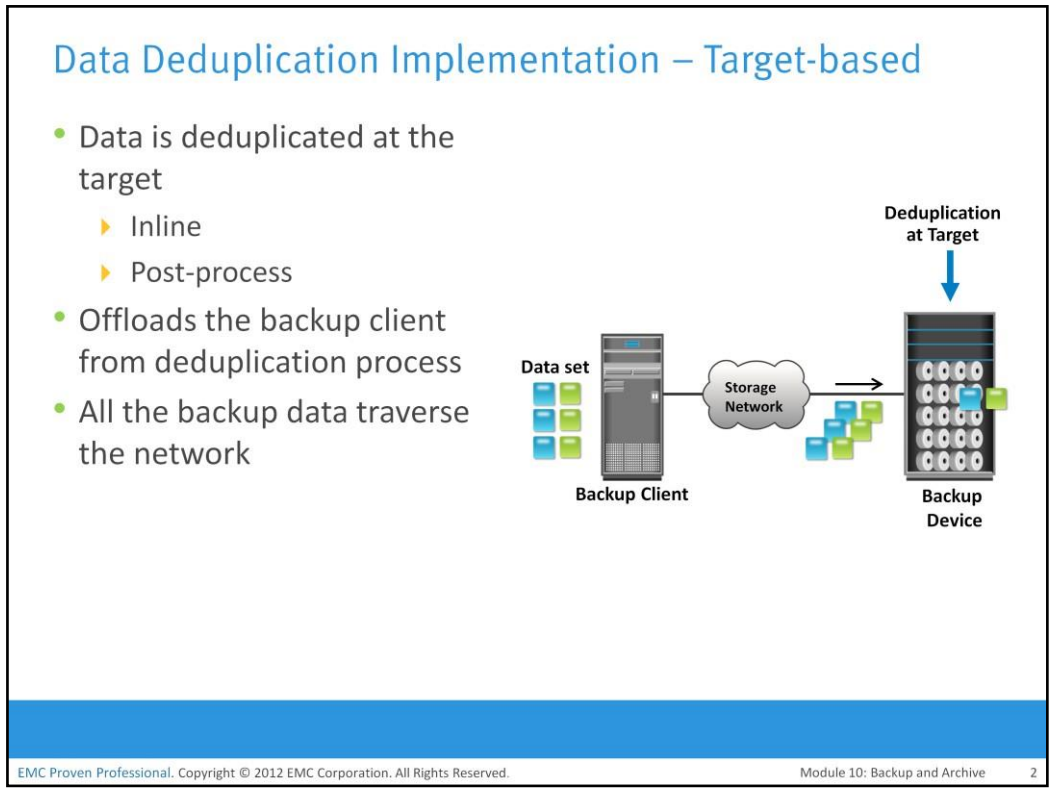
- Data is deduplicated at the source (backup client)
- Backup client sends only new, unique segments across the network
- Reduced storage capacity and network bandwidth requirements
- Increased overhead on the backup client







*Source-based data deduplication* eliminates redundant data at the source before it transmits to the backup device. Source-based data deduplication can dramatically reduce the amount of backup data sent over the network during backup processes. It provides the benefits of a shorter backup window and requires less network bandwidth. There is also a substantial reduction in the capacity required to store the backup images. Source-based deduplication increases the overhead on the backup client, which impacts the performance of the backup and application running on the client. Source-based deduplication might also require a change of backup software if it is not supported by backup software.

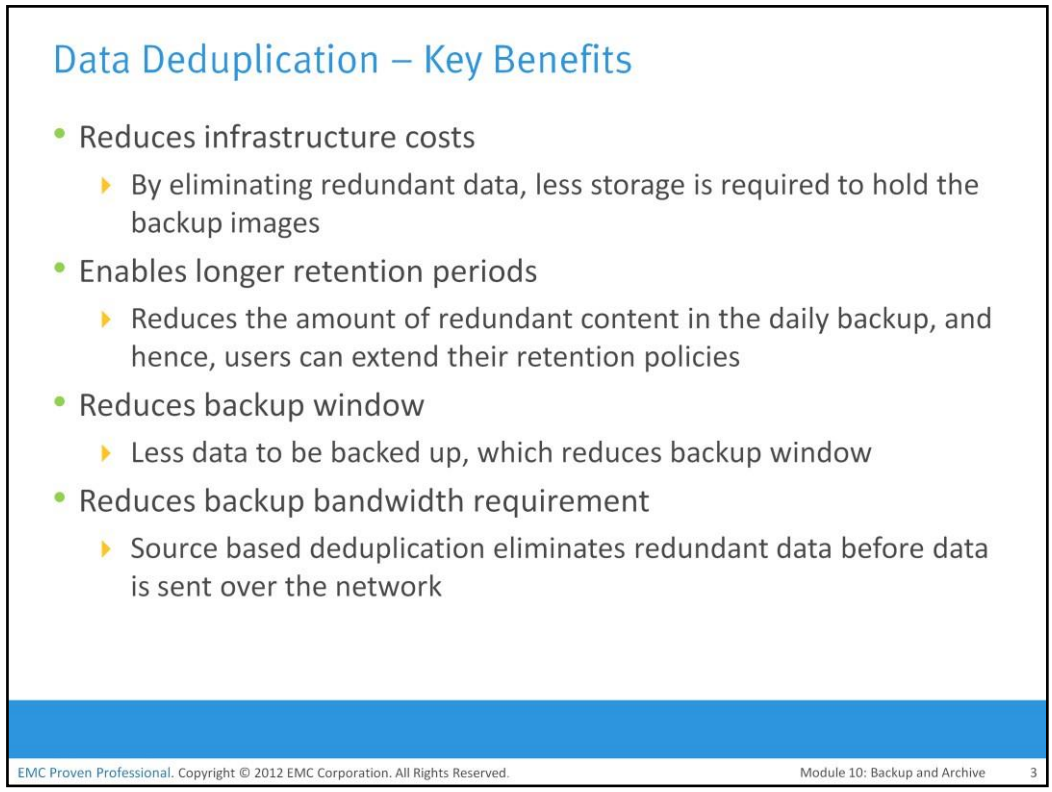



*Target-based data deduplication* is an alternative to source-based data deduplication. Target-based data deduplication occurs at the backup device, which offloads the backup client from the deduplication process. Figure on the slide illustrates target-based data deduplication. In this case, the backup client sends the data to the backup device and the data is deduplicated at the backup device, either immediately (Inline) or at a scheduled time (Post-process).

*Inline deduplication* performs deduplication on the backup data before it is stored on the backup device. Hence, this method reduces the storage capacity needed for the backup. Inline deduplication introduces overhead in the form of the time required to identify and remove duplication in the data. So, this method is best suited for an environment with a large backup window.

*Post-process deduplication* enables the backup data to be stored on the backup device first and then deduplicated later. This method is suitable for situations with tighter backup windows. However, post-process deduplication requires more storage capacity to store the backup images before they are deduplicated.

Because deduplication occurs at the target, all the backup data need to be transferred over the network, which increases network bandwidth requirements. Target-based data deduplication does not require any changes in the existing backup software.

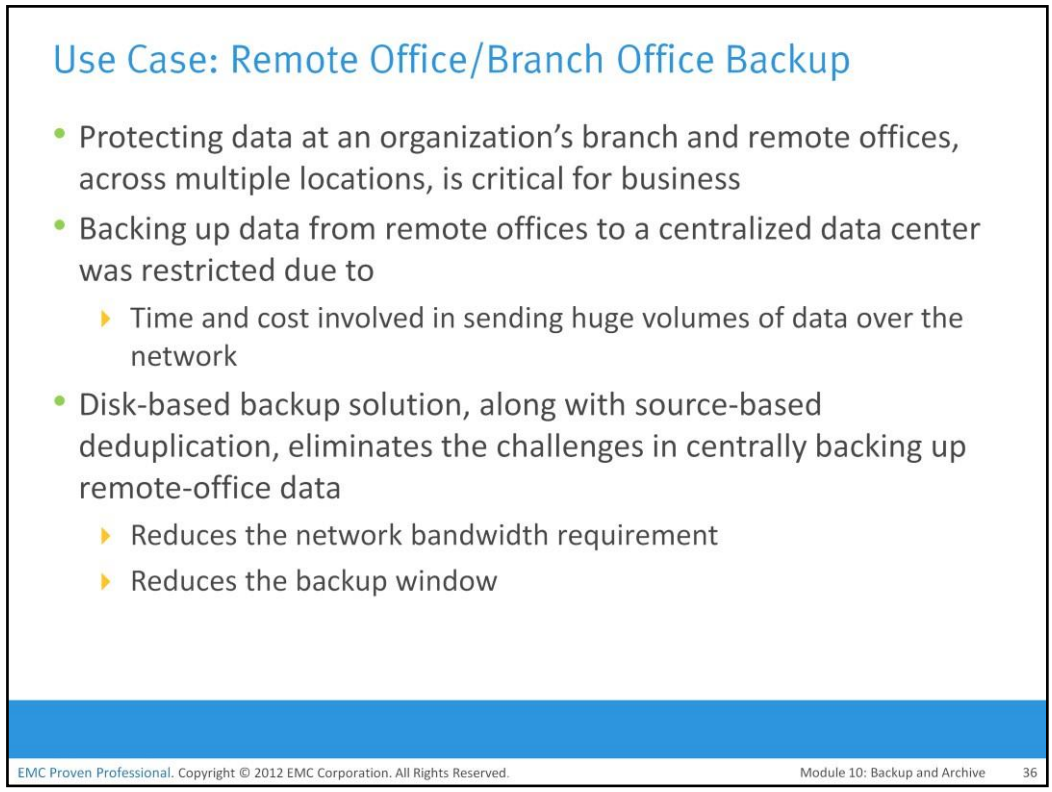



**Reduces infrastructure costs:** By eliminating redundant data from the backup, far less infrastructure is required to hold the backup images. Data deduplication directly results in reduced storage capacities to hold backup images. Smaller capacity requirements means lower acquisition costs as well as reduced power and cooling costs.

**Enables longer retention periods:** As data deduplication reduces the amount of content in the daily backup, users can extend their retention policies. This can have a significant benefit to users who currently require longer retention.

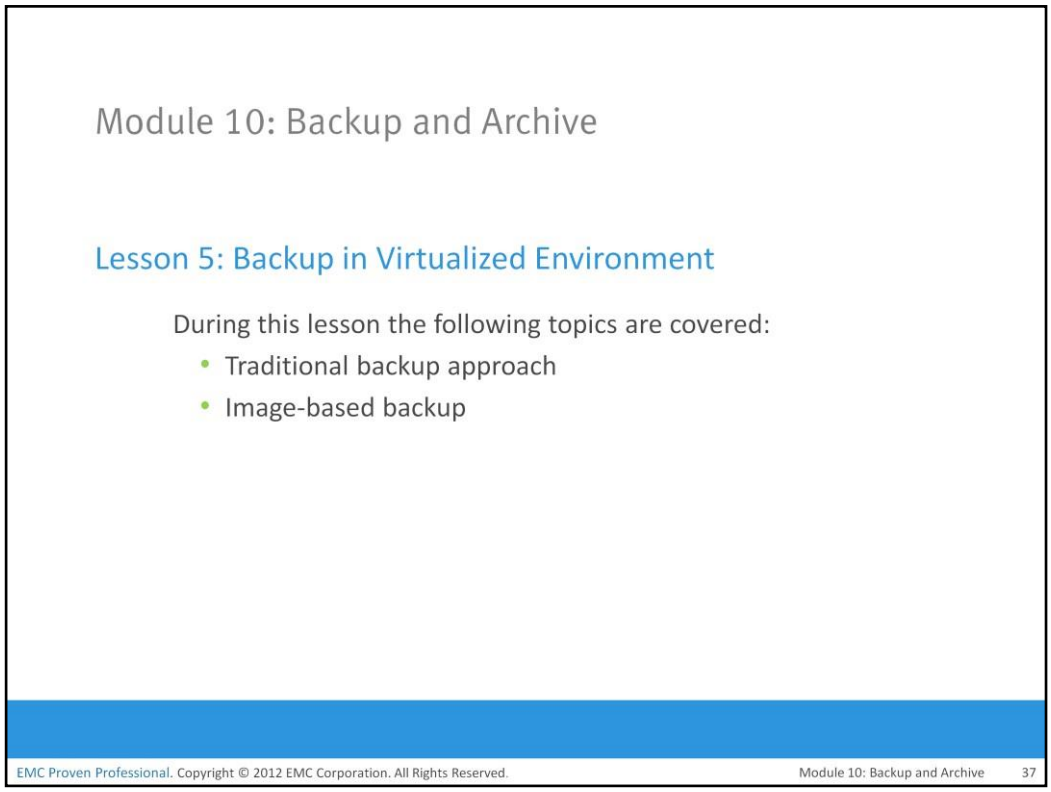
**Reduces backup window:** Data deduplication eliminates redundant content of backup data, which makes less data to be backed up and reduces backup window.

**Reduces backup bandwidth requirement:** By utilizing data deduplication at the client (source-based), redundant data is removed before the data is transferred over the network. This considerably reduces the network bandwidth required for backup.

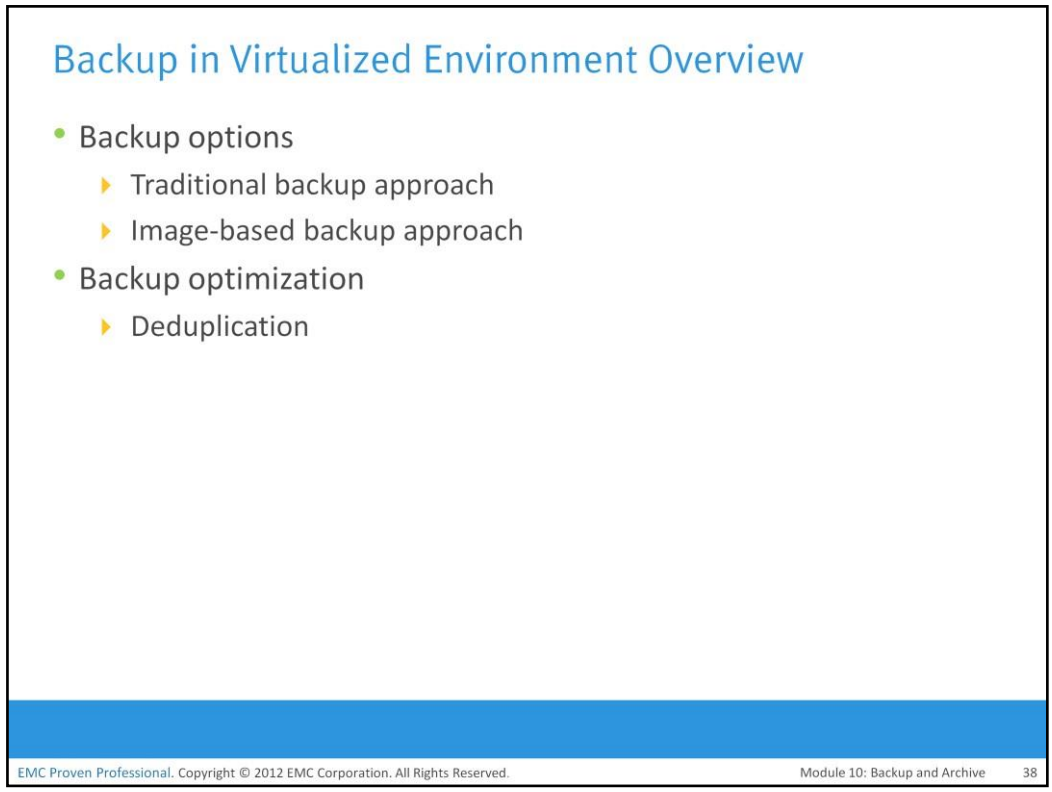



Today, businesses have their remote or branch offices spread over multiple locations. Typically, these remote offices have their local IT infrastructure. This infrastructure includes file, print, Web, or email servers, workstations, and desktops, and might also house some applications and databases. Too often, business-critical data at remote offices are inadequately protected, exposing the business to the risk of lost data and productivity. As a result, protecting the data of an organization's branch and remote offices across multiple locations is critical for business. Traditionally, remote-office data backup was done manually using tapes, which were transported to offsite locations for DR support. Some of the challenges with this approach were lack of skilled onsite technical resources to manage backups and risk of sending tapes to offsite locations, which could result in loss or theft of sensitive data. Backing up data from remote offices to a centralized data center was restricted due to the time and cost involved in sending huge volumes of data over the WAN. Therefore, organizations needed an effective solution to address the data backup and recovery challenges of remote and branch offices.

Disk-based backup solutions along with source-based deduplication eliminate the challenges associated with centrally backing up remote-office data. Deduplication considerably reduces the required network bandwidth and enables remote-office data backup using the existing network. Organizations can now centrally manage and automate remote-office backups while reducing the required backup window.



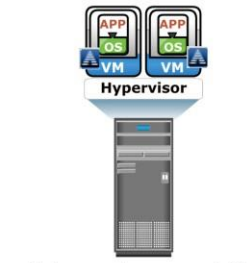

This lesson covers traditional backup approach and image-based backup in virtualized environment.



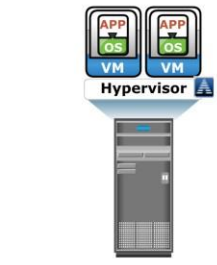

In a virtualized environment, it is imperative to back up the virtual machine data (OS, application data, and configuration) to prevent its loss or corruption due to human or technical errors. There are two approaches for performing a backup in a virtualized environment: the traditional backup approach and the image-based backup approach. Owing to the increased capacity requirements in a virtualized environment, backup optimization methods are necessary. The use of deduplication techniques significantly reduces the amount of data to be backed up in a virtualized environment. The effectiveness of deduplication is identified when VMs with similar configurations are deployed in a data center. The deduplication types and methods used in a virtualized environment are the same as in the physical environment.

## Traditional Backup Approaches

- Backup agent on VM
  - Requires installing a backup agent on each VM running on a hypervisor
  - Can only backup virtual disk data
  - Does not capture VM files such as VM swap file, configuration file
  - Challenge in VM restore
- Backup agent on Hypervisor
  - Requires installing backup agent only on hypervisor
  - Backs up all the VM files




Backup agent runs on each VM




Backup agent runs on Hypervisor

= Backup Agent




In the *traditional backup approach*, a backup agent is installed either on the virtual machine (VM) or on the hypervisor. If the backup agent is installed on a VM, the VM appears as a physical server to the agent. The backup agent installed on the VM backs up the VM data to the backup device. The agent does not capture VM files, such as the virtual BIOS file, VM swap file, logs, and configuration files. Therefore, for a VM restore, a user needs to manually re-create the VM and then restore data on to it.

If the backup agent is installed on the hypervisor, the VMs appear as a set of files to the agent. So, VM files can be backed up by performing a file system backup from a hypervisor. This approach is relatively simple because it requires having the agent just on the hypervisor instead of having on all the VMs. The traditional backup method can cause high CPU utilization on the server being backed up. In the traditional approach, the backup should be performed when the server resources are idle or during a low activity period on the network. Also consider allocating enough resources to manage the backup on each server when a large number of VMs are in the environment.

## Image-based Backup

- Creates a copy of the guest OS, its data, VM state, and configurations
  - The backup is saved as a single file – "image"
  - Mounts image on a proxy server
  - Offloads backup processing from the hypervisor
- Enables quick restoration of VM

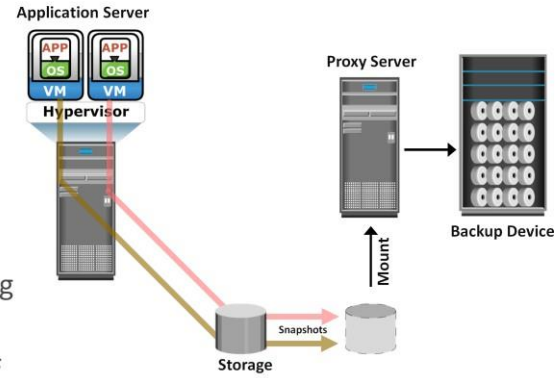
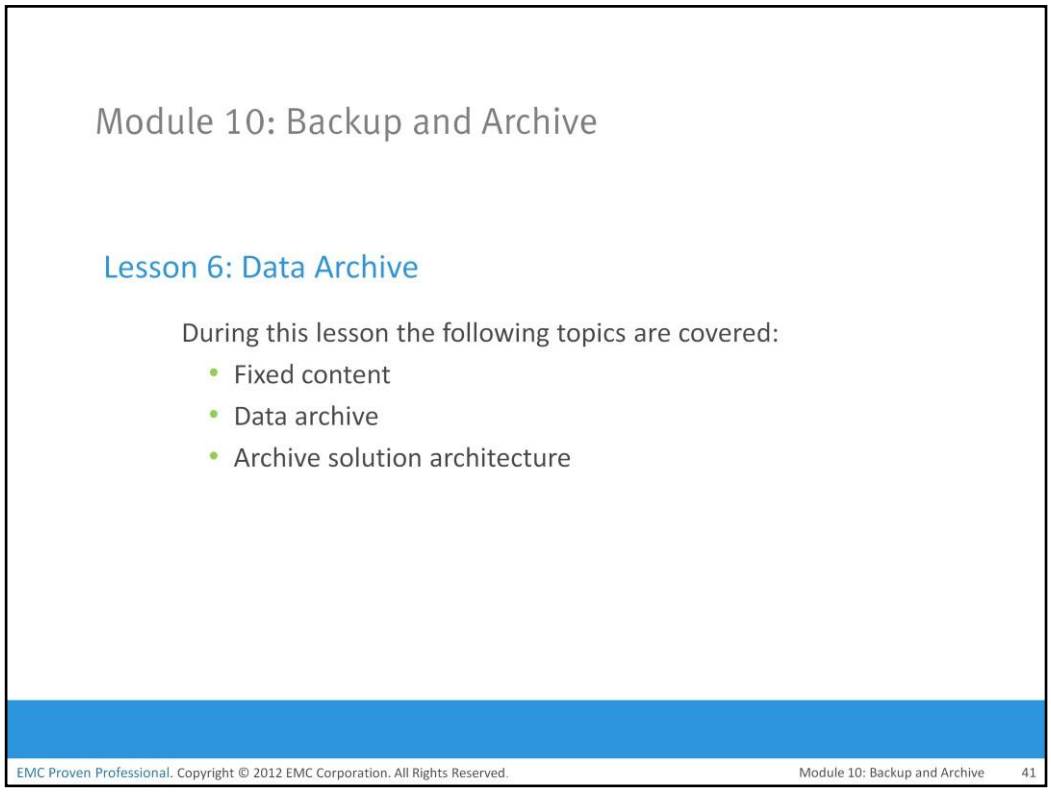


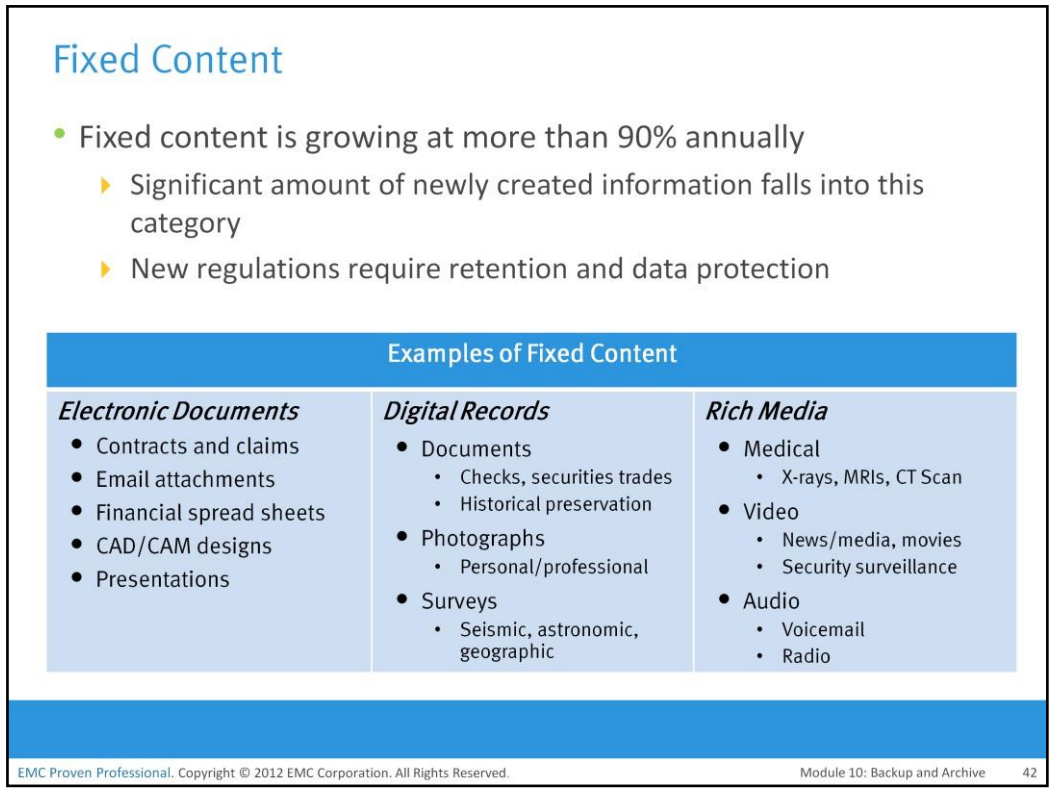




*Image-based backup* operates at the hypervisor level and essentially takes a snapshot of the VM. It creates a copy of the guest OS and all the data associated with it (snapshot of VM disk files), including the VM state and application configurations. The backup is saved as a single file called an "image" and this image is mounted on the proxy server (acts as a backup client). The backup software then backs up these image files normally. This effectively offloads the backup processing from the hypervisor and transfers the load on the proxy server, thereby reducing the impact to VMs running on the hypervisor. Image-based backup enables quick restoration of a VM.

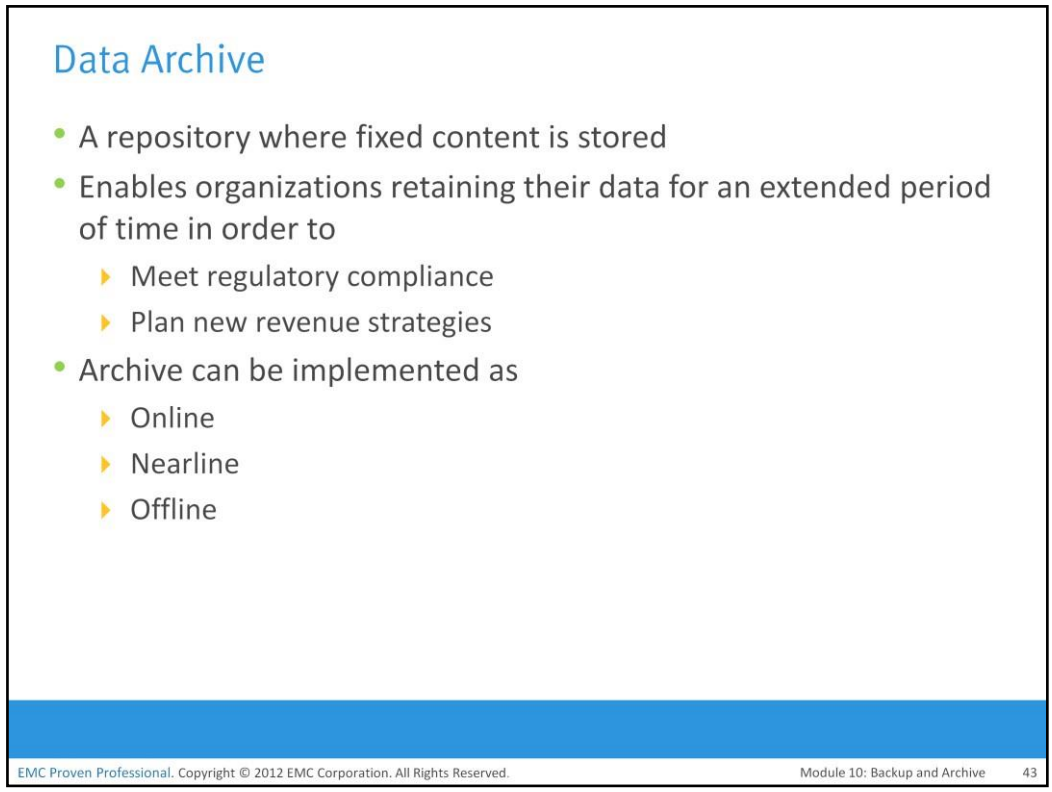



This lesson covers fixed content and challenges in storing fixed content. This lesson also focuses on data archive solution architecture.




In the life cycle of information, data is actively created, accessed, and changed. As data ages, it is less likely to be changed and eventually becomes "fixed" but continues to be accessed by applications and users. This data is called *fixed content*. All organizations may require to retain their data for an extended period of time due to government regulations and legal/contractual obligations. Organizations also make use of this fixed content to generate new revenue strategies and improve service levels.

Currently, fixed content data is the fastest growing sector of the data storage market. Assets such as X-rays, MRIs, CAD/CAM designs, surveillance video, MP3s and financial documents are just a few examples of an important class of data that is growing at over 90% annually.

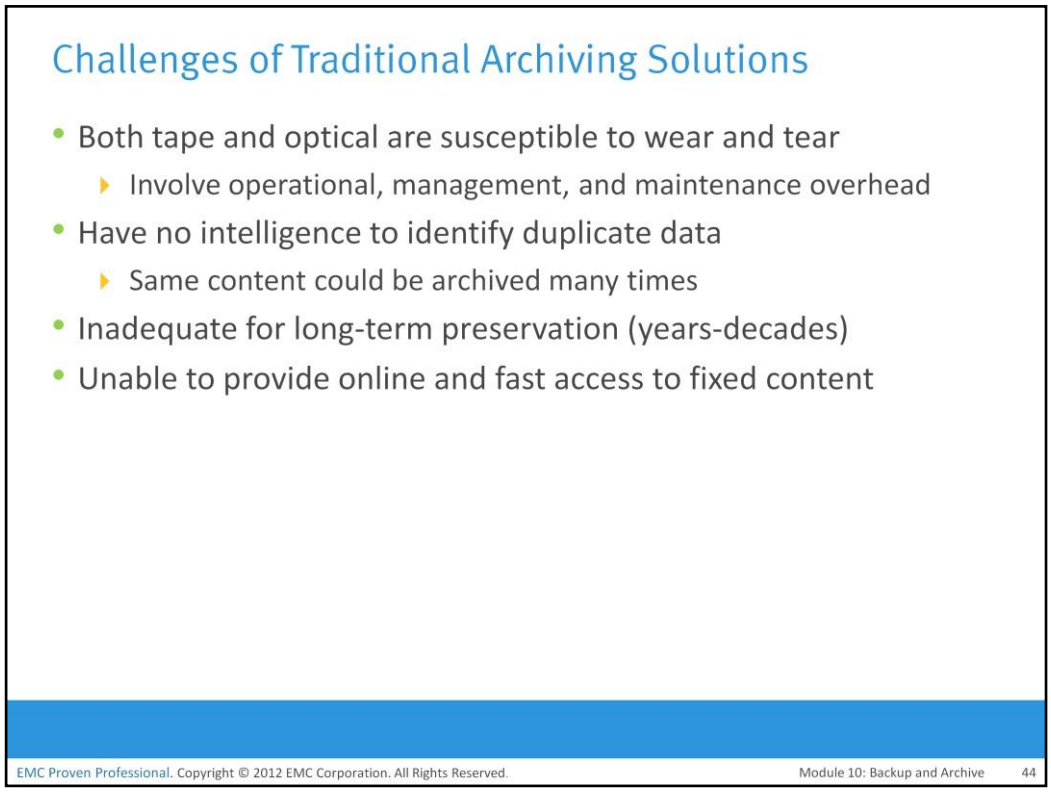



Data archive is a repository where fixed content is stored. It enables organizations retaining their data for an extended period of time in order to meet regulatory compliance and generate new revenue strategies. An archive can be implemented as an online, nearline, or offline solution:

**Online archive:** A storage device directly connected to a host that makes the data immediately accessible.

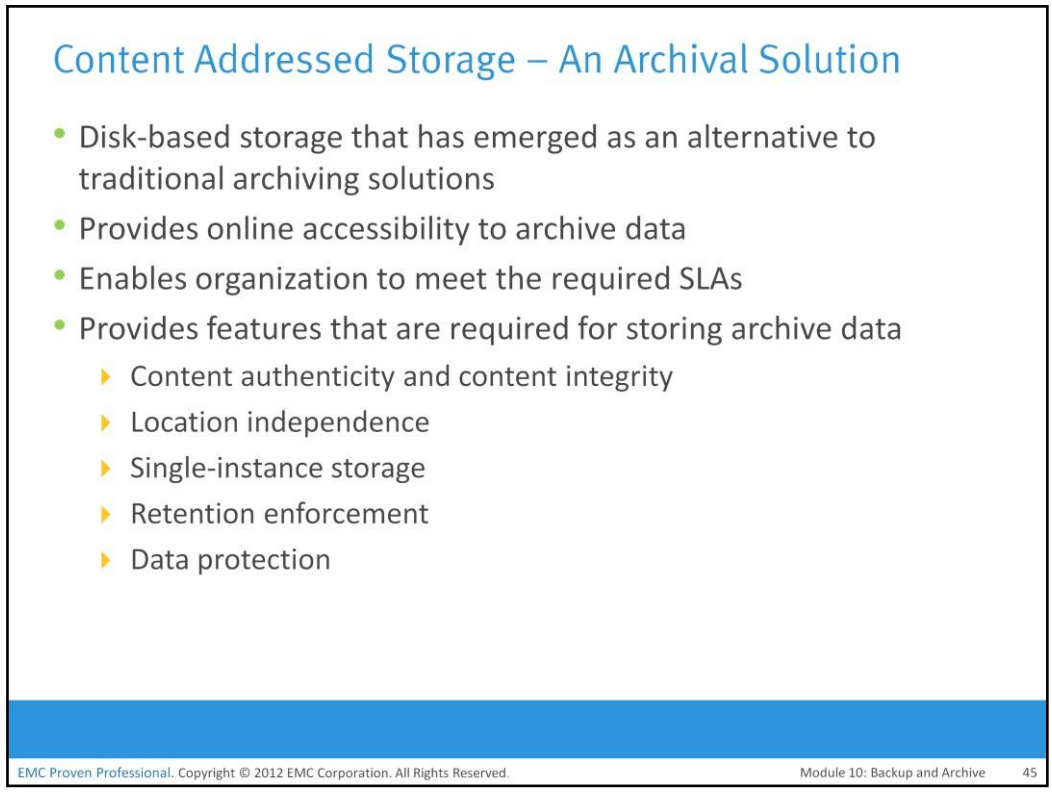
**Nearline archive:** A storage device connected to a host, but the device where the data is stored must be mounted or loaded to access the data.

**Offline archive:** A storage device that is not ready to use. Manual intervention is required to connect, mount, or load the storage device before data can be accessed.

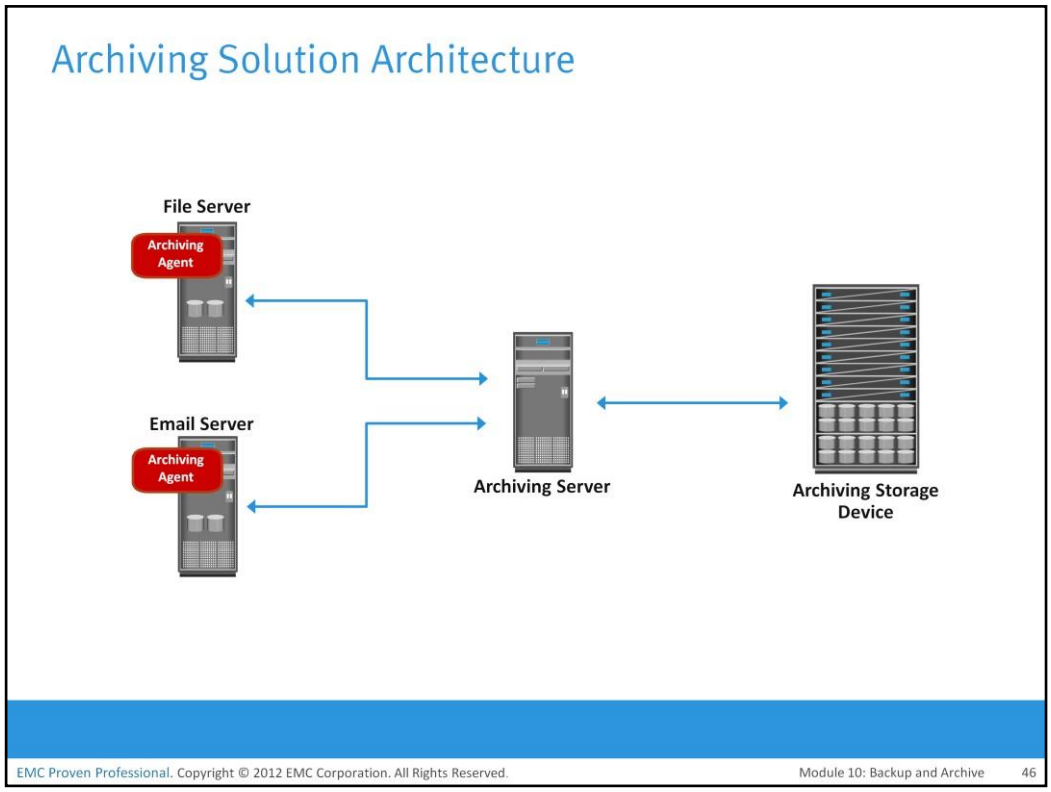



Typically, long-term preservation is required (years-decades) for fixed content and it is also important to have simultaneous, fast online access. The increase in fixed content is driven by regulatory requirements. Because of this explosive growth and changes to user requirements, traditional storage options are inadequate.

Optical media are typically *write once read many* (WORM) devices that protect the original file from being overwritten. Some tape devices also provide this functionality by implementing file-locking capabilities. Although these devices are inexpensive, they involve operational, management, and maintenance overhead. The traditional solutions using optical discs and tapes is not optimized to recognize the content, so that the same content could be stored several times. Additional costs are involved in offsite storage of media and media management. Tapes and optical media are also susceptible to wear and tear. Frequent changes in these device technologies lead to the overhead of converting the media into new formats to enable access and retrieval. Government agencies and industry regulators are establishing new laws and regulations to enforce the protection of archives from unauthorized destruction and modification. These regulations and standards have established new requirements for preserving the integrity of information in the archives. These requirements have exposed the shortcomings of the traditional tape and optical media archive solutions.




Content addressed storage (CAS) a disk based storage that has emerged as an alternative to tape and optical solutions. CAS meets the demand to improve data accessibility and to protect, dispose off, and ensure service level agreements (SLAs) for archive data. CAS is detailed in module 8.

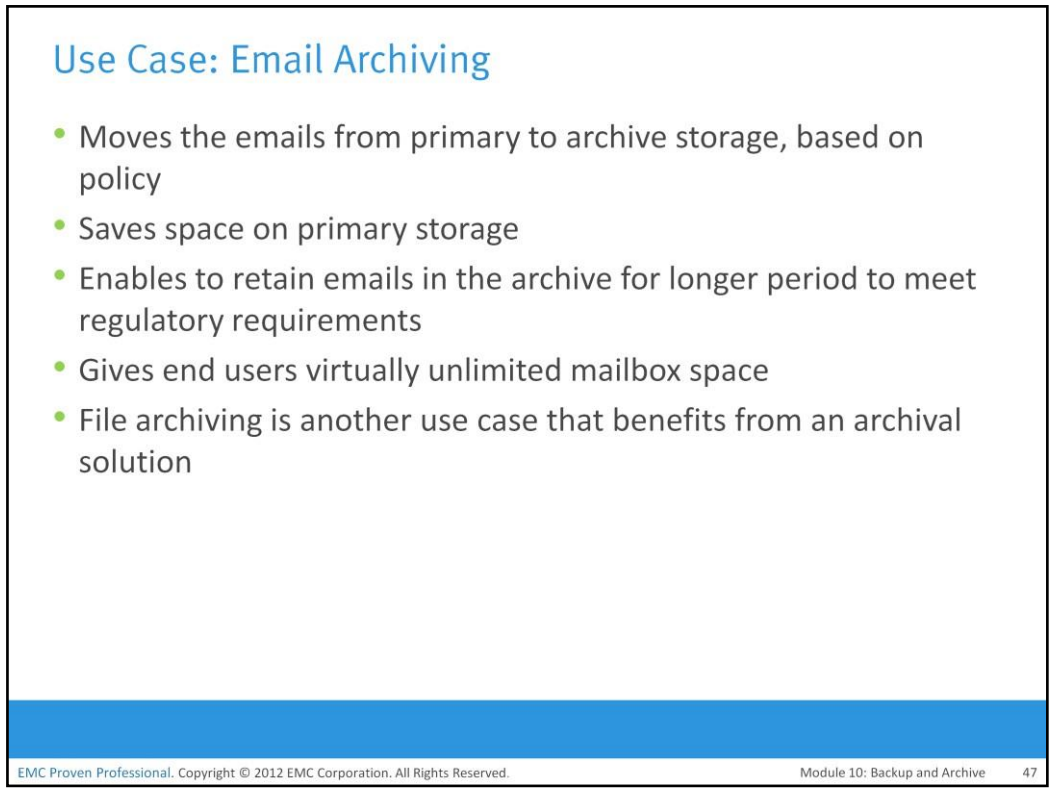



Archiving solution architecture consists of three key components; archiving agent, archiving server, and archiving storage device.

*An archiving agent* is software installed on the application server. The agent is responsible for scanning the data that can be archived based on the policy defined on the archiving server. After the data is identified for archiving, the agent sends the data to the archiving server. Then the original data on the application server is replaced with a stub file. The stub file contains the address of the archived data. The size of this file is small and significantly saves space on primary storage. This stub file is used to retrieve the file from the archive storage device.

*An archiving server* is software installed on a host that enables administrators to configure the policies for archiving data. Policies can be defined based on file size, file type, or creation/modification/access time. The archiving server receives the data to be archived from the agent and sends it to the archive storage device.

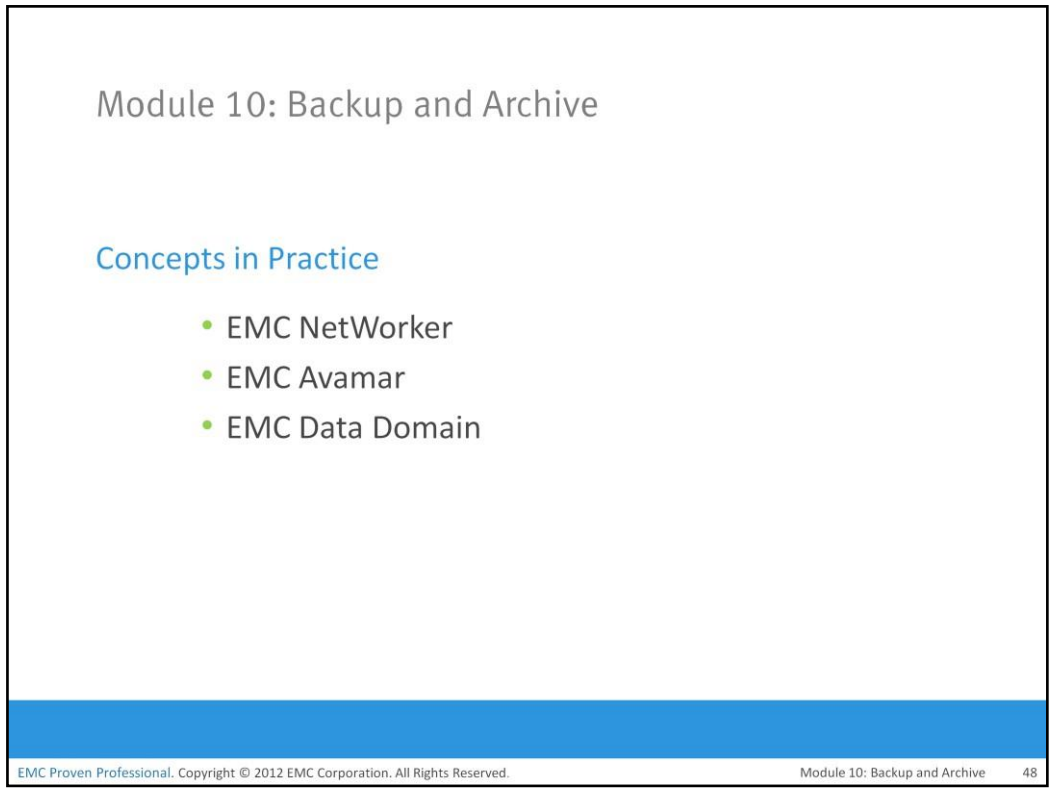
*An archive storage device* stores fixed content.



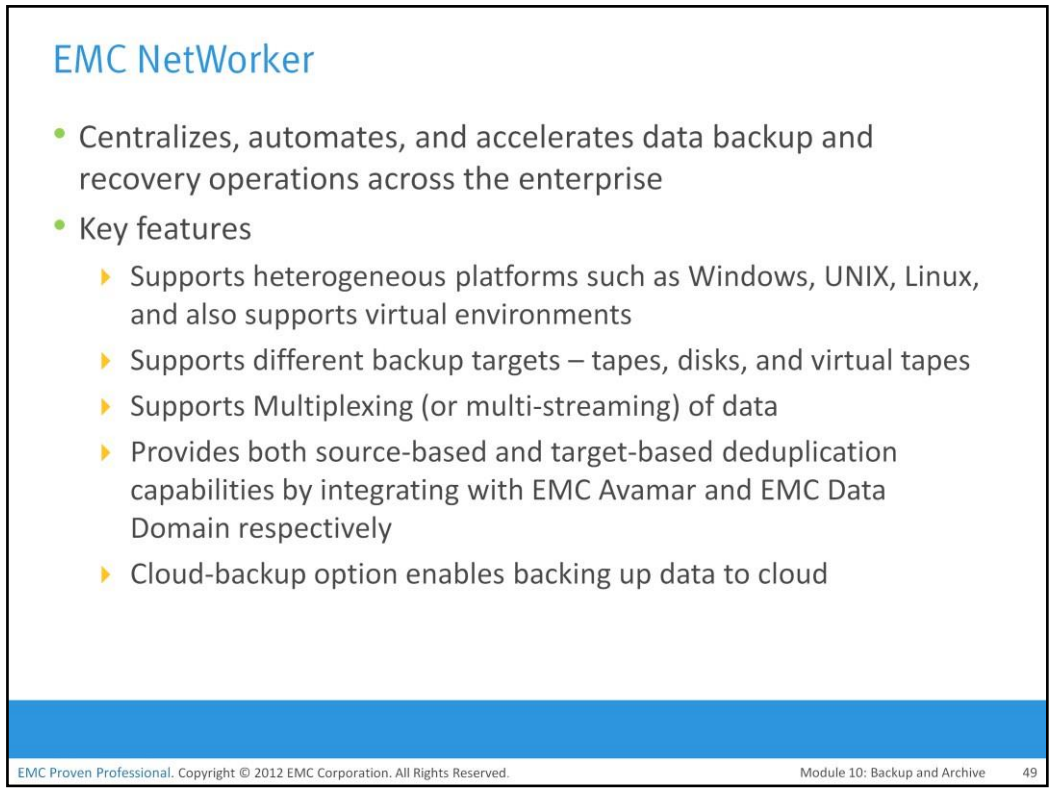

E-mail is an example of an application that benefits most by an archival solution. Typically, a system administrator configures small mailboxes that store a limited number of e-mails. This is because large mailboxes with a large number of e-mails can make management difficult, increase primary storage cost, and degrade system performance. When an e-mail server is configured with a large number of mailboxes, the system administrator typically configures a quota on each mailbox to limit its size. Configuring fixed quota on mailboxes impacts end users. A fixed quota for a mailbox forces users to delete e-mails as they approach the quota size. End users often need to access e-mails that are weeks, months, or even years old.

E-mail archiving provides an excellent solution that overcomes the preceding challenges. Archiving solutions move e-mails that have been identified as candidates for archive from primary storage to the archive storage device based on a policy—for example, “e-mails that are 90 days old should be archived.” After the e-mail is archived, it is retained for years based on the retention policy. This considerably saves space on primary storage and enables organizations to meet regulatory requirements. Implementation of an archiving solution gives end users virtually unlimited mailbox space.

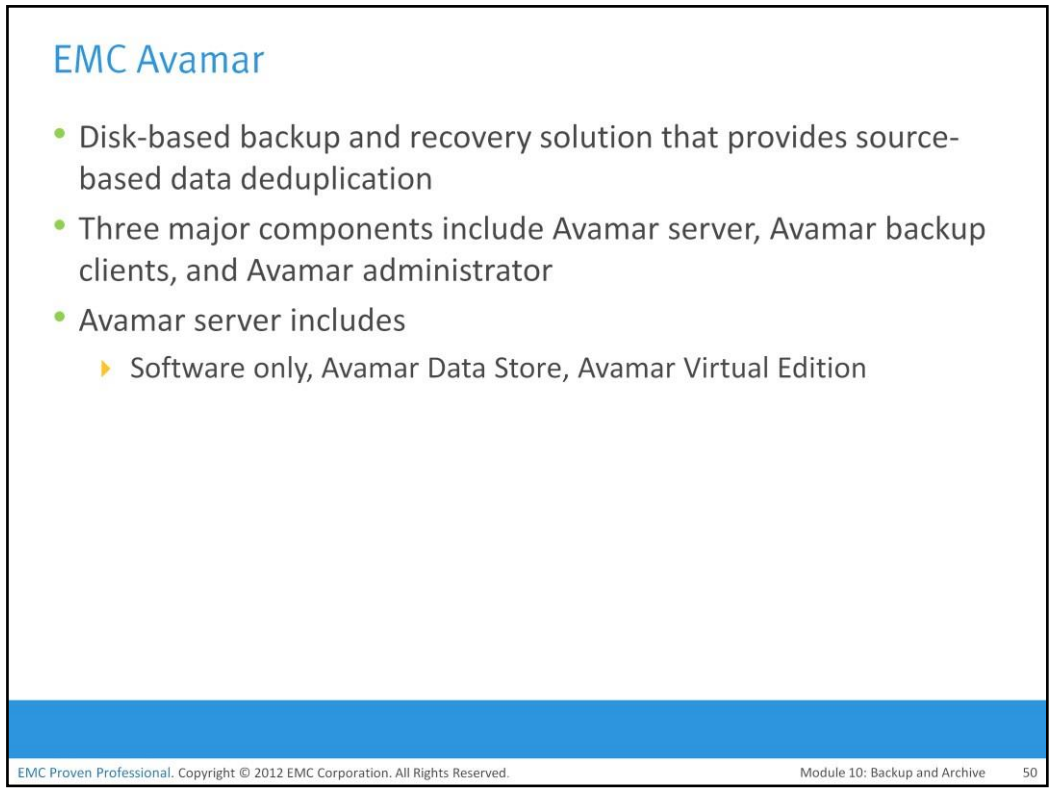
A file sharing environment is another environment that benefits from an archival solution. Typically, users store a large number of files in the shared location. Most of these files are old and rarely accessed. Administrators configure quotas on the file share that forces the users to delete these files. This impact users because they may require access to files that may be months or even years old. In some cases the user may request an increase in the size of the file share. This in turn increases the cost of primary storage. A file archiving solution archives the files based on the policy such as age of files, size of files, and so on. This considerably reduces the primary storage requirement and also enables users to retain the files in the archive for longer periods.




The concept in practice section covers various EMC backup and archive products.

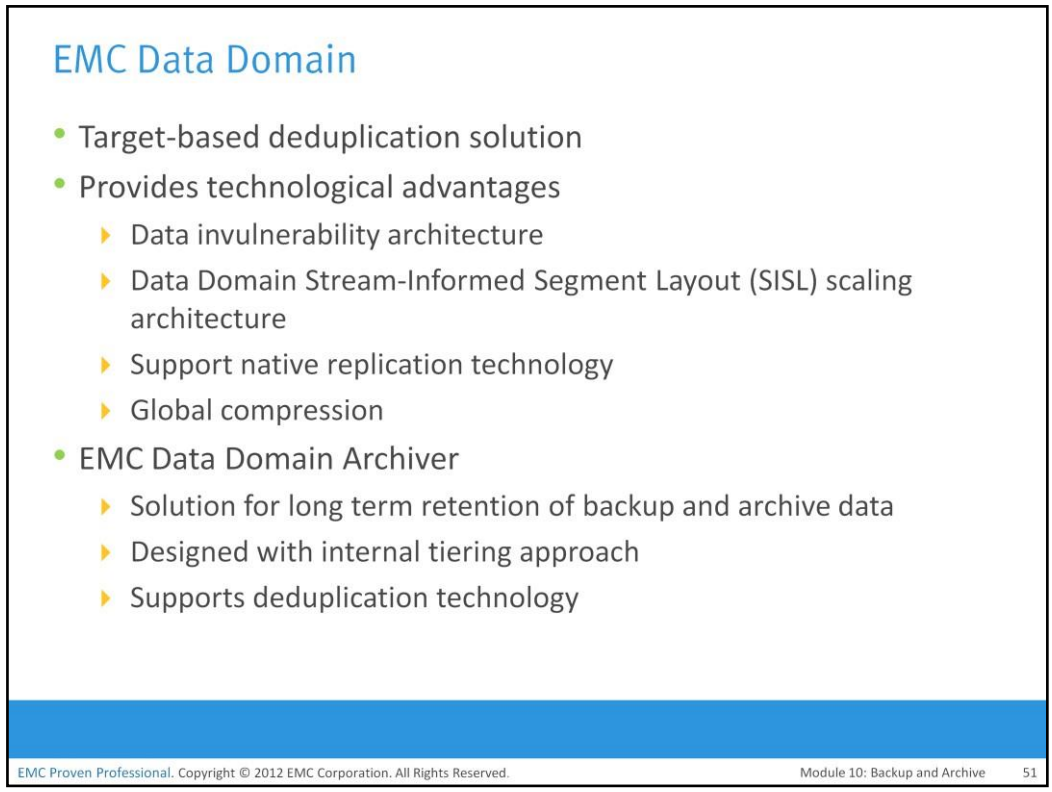



The EMC NetWorker backup and recovery software centralizes, automates, and accelerates data backup and recovery operations across the enterprise. The features of EMC NetWorker are listed on the slide.




EMC Avamar is a disk-based backup and recovery solution that provides inherent source-based data deduplication. With its unique global data deduplication feature, Avamar differs from traditional backup and recovery solutions, by identifying and storing only unique sub-file data objects. Redundant data is identified at the source, the amount of data that travels across the network is drastically reduced, and the backup storage requirement is also considerably reduced. The three major components of an Avamar system includes Avamar server, Avamar backup clients, and Avamar administrator. Avamar server provides the essential processes and services required for client access and remote system administration. The Avamar client software runs on each computer or network server that is being backed up. Avamar administrator is a user management console application that is used to remotely administer an Avamar system. The three Avamar server editions include software only, Avamar Data Store, and Avamar Virtual Edition. The features of EMC Avamar are as follows:

- **Fault tolerance:** Uses RAID, RAIN, checkpoints, and replication to provide data integrity and protection.
- **Standard IP network leveraging:** Optimizes the use of network for backup; dedicated backup networks are not required. Daily full backups are possible using the existing networks and infrastructure.
- **Scalable server architecture:** Additional storage nodes can be added non-disruptively to accommodate increased backup storage requirements.
- **Centralized management:** Enables remote management of Avamar servers from a centralized location and through the use of the Avamar Enterprise Manager and Avamar Administrator interfaces.




The EMC Data Domain deduplication storage system is a target-based data deduplication solution. Using high-speed, inline deduplication technology, the Data Domain system provides a storage footprint that is significantly smaller on an average, than that of the original data set. Data Domain systems can scale from smaller remote office appliances to large data-center systems. These systems are available as integrated appliances or as gateways that use external storage.

Data Domain deduplication storage systems provide the following unique advantages:

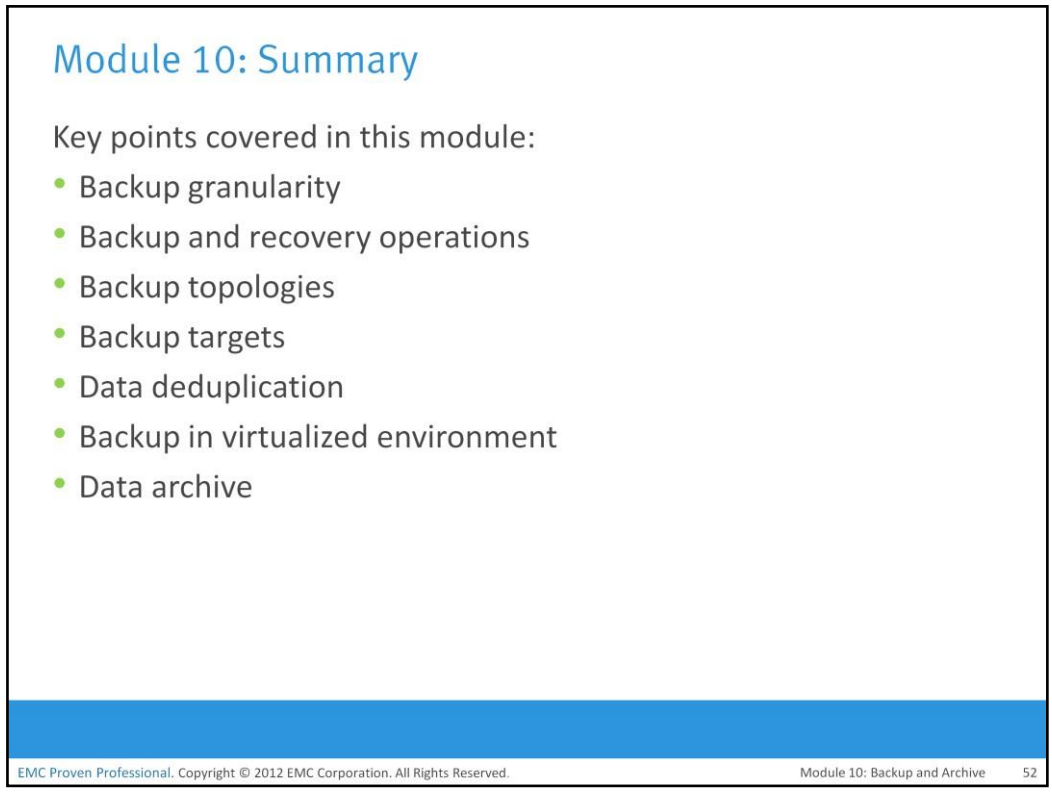
**Data invulnerability architecture:** Provides unprecedented levels of data integrity, data verification, and self-healing capabilities, such as RAID 6 protection. Continuous fault detection, healing, and write verification ensure that the backup is accurately stored, available, and recoverable.

**Data Domain SISL (Stream-Informed Segment Layout) scaling architecture:** Enables scaling of CPUs to add a direct benefit to system throughput scalability.

**Support native replication technology:** Enables automatic, secure transfer of compressed data over the wide area network (WAN) with minimum bandwidth requirement.

**Global compression:** Highly efficient deduplication and compression technology, which radically changes storage economics.

EMC Data Domain Archiver is a solution for long term retention of backup and archive data. It is designed with internal tiering approach to enable cost effective, long term retention of data on disk by implementing deduplication technology.




This module covered various backup granularities and backup operations. This module also discussed various backup topologies and backup targets. Further this module covered data deduplication and backup in virtualized environment. Additionally, this module also covered data archive in detail.

A *backup* is an additional copy of the production data, created and retained for the sole purpose of recovering lost or corrupted data. Based on the granularity, backups can be categorized as full, cumulative, and incremental.

The three basic topologies used in a backup environment are direct-attached backup, LAN-based backup, and SAN-based backup.

A wide range of technology solutions are currently available for backup targets. Tape and disk libraries are the two most commonly used backup targets. Virtual tape library (VTL) is one of the options that uses disks as backup medium. VTL emulates tapes and provides enhanced backup and recovery capabilities.

Data deduplication is the process of identifying and eliminating redundant data. When duplicate data is detected during backup, the data is discarded and only the pointer is created to refer the copy of the data which is already backed up.

In a virtualized environment, it is imperative to back up the virtual machine data (OS, application data, and configuration) to prevent its loss or corruption due to human or technical errors.

Data archive is a repository where fixed content is stored. It enables organizations to retain their data for an extended period of time, in order to meet regulatory compliance and generate new revenue strategies.

## Check Your Knowledge – 1

- Which is true about incremental backup?
  - A. Restore requires only last full and last incremental backup
  - B. Restore requires only last incremental backup
  - C. Copies the data that has changed since last full or incremental backup
  - D. Copies the data that has changed since last full backup

- What is an advantage of image-based backup over traditional backup approach in a virtualized environment?
  - A. Offloads backup processing from the hypervisor
  - B. Faster because it copies only virtual machine disk data
  - C. Faster because it copies only virtual machine configuration data
  - D. Space required is a fraction of total backup data

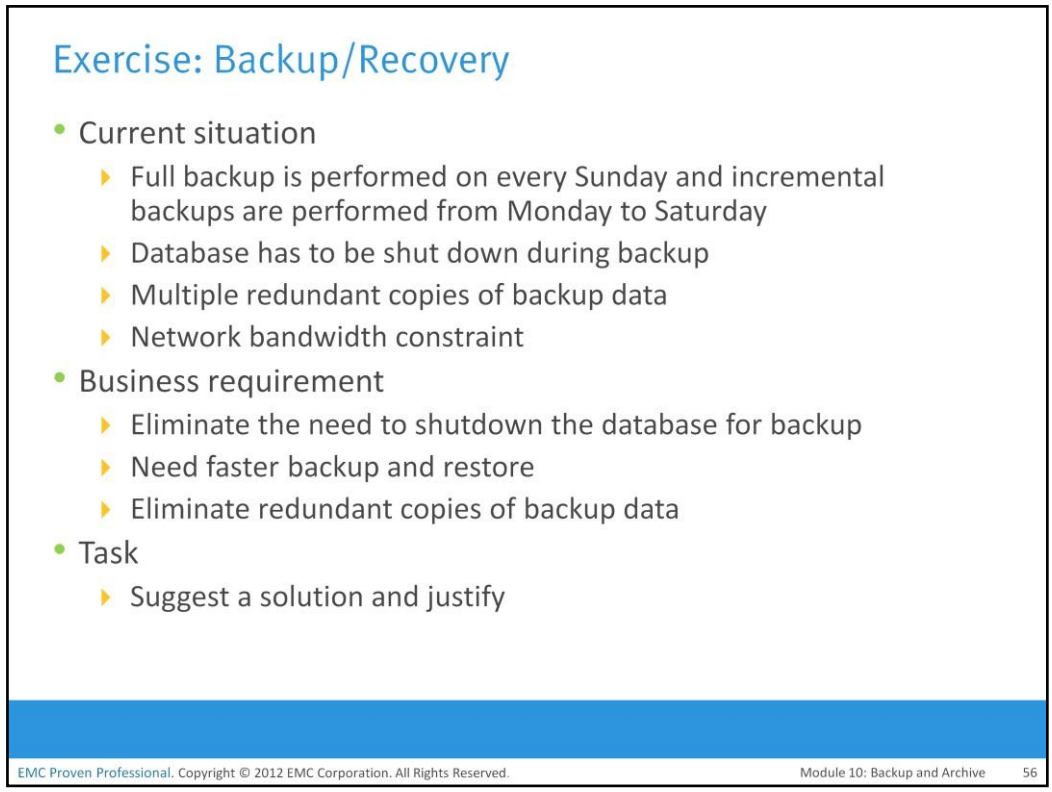
## Check Your Knowledge – 2

- Which accurately describes the role of a backup server?
  - A. Gathers the data that is to be backed up and send it to storage node
  - B. Responsible for writing the data, which client sends, to backup device
  - C. Manages the backup operation and maintains backup catalog
  - D. Controls the robotic arm in the tape library

- In backup to tape environment, what does 'shoe shining' mean?
  - A. Writing data from multiple streams on a single tape
  - B. Process of emulating disk drives and presenting as tapes to backup software
  - C. Repeated back and forth motion that a tape drive makes when there is an interruption in the backup data stream
  - D. Process of deleting redundant content in the backup data

## Check Your Knowledge – 3

- What is an advantage of source-based deduplication?
  - A. Improves the performance of backup client
  - B. Improves the performance of backup server
  - C. Reduces backup window to zero
  - D. Reduces the network bandwidth requirement for backup




**Business profile:**
An organization uses tape as their primary backup storage media for their applications.

**Current situation:**

- Full backup is performed on every Sunday and incremental on remaining days
- Their database has to be shut down during the backup process
- Multiple redundant copies of backup data
- Network bandwidth constraint

**Requirements:**

- Eliminate the need to shutdown the database for backup
- Need faster backup and restore
- Eliminate redundant copies of backup data

**Task:**
Propose a solution to address the organization's concern and justify your solution.