MODULE – 4

INTRODUCTION TO BUSINESS CONTINUITY

Chapter 9: INTRODUCTION TO BUSINESS CONTINUITY

Business Continuity (BC):

Business continuity (BC) is an integrated and enterprise wide process that includes all activities (internal and external to IT) that a business must perform to mitigate the impact of planned and unplanned downtime.

BC entails preparing for, responding to, and recovering from a system outage that adversely affects business operations. It involves proactive measures, such as business impact analysis, risk assessments, deployment of BC technology solutions (backup and replication), and reactive measures, such as disaster recovery and restart, to be invoked in the event of a failure.

The goal of a BC solution is to ensure the **“information availability”** required to conduct vital business operations.

9.1 Information Availability:

Information availability (IA) refers to the ability of the infrastructure to function according to business expectations during its specified time of operation. Information availability ensures that people (employees, customers, suppliers, and partners) can access information whenever they need it. Information availability can be defined in terms of:

1. Reliability,
2. Accessibility
3. Timeliness.

1. **Reliability:** This reflects a component’s ability to function without failure, under stated conditions, for a specified amount of time.
2. **Accessibility:** This is the state within which the required information is accessible at the right place, to the right user. The period of time during which the system is in an accessible state is termed **system uptime;** when it is not accessible it is termed **system**

downtime.

3. **Timeliness:** Defines the exact moment or the time window (a particular time of the day, week, month, and/or year as specified) during which information must be accessible. For example, if online access to an application is required between 8:00 am and 10:00 pm each day, any disruptions to data availability outside of this time slot are not considered to affect timeliness.

9.1.1 Causes of Information Unavailability

Various planned and unplanned incidents result in data unavailability.

- **Planned outages** include installation/integration/maintenance of new hardware, software upgrades or patches, taking backups, application and data restores, facility operations (renovation and construction), and refresh/migration of the testing to the production environment.
- **Unplanned outages** include failure caused by database corruption, component failure, and human errors.
- **Disasters (natural or man-made)** such as flood, fire, earthquake, and contamination are another type of incident that may cause data unavailability.

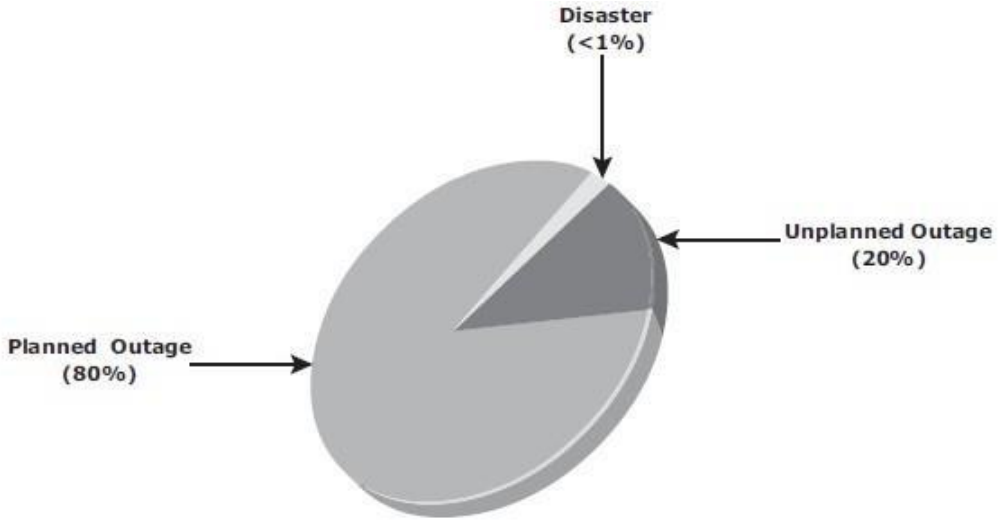


Fig 9.1: Disruptors of Information Availability

As illustrated in Fig 9.1 above, the majority of outages are planned. Planned outages are expected and scheduled, but still cause data to be unavailable.

9.1.2 Consequences of Downtime

- Information unavailability or downtime results in loss of productivity, loss of revenue, poor financial performance, and damage to reputation.
- Loss of productivity includes reduced output per unit of labor, equipment, and capital.
- Loss of revenue includes direct loss, compensatory payments, future revenue loss, billing loss, and investment loss.
- Poor financial performance affects revenue recognition, cash flow, discounts, payment guarantees, credit rating, and stock price.
- Damages to reputations may result in a loss of confidence or credibility with customers, suppliers, financial markets, banks, and business partners.
- An important metric, *average cost of downtime per hour*, provides a key estimate in determining the appropriate BC solutions. It is calculated as follows:

  Average cost of downtime per hour = average productivity loss per hour + average revenue loss per hour

  Where:

  Productivity loss per hour = (total salaries and benefits of all employees per week) /(average number of working hours per week)

  Average revenue loss per hour = (total revenue of an organization per week) /(average number of hours per week that an organization is open for business)

9.1.3 Measuring Information Availability

- Information availability (IA) relies on the availability of physical and virtual components of a data center. Failure of these components might disrupt IA. A failure is the termination of a component's capability to perform a required function. The component's capability can be restored by performing an external corrective action, such as a manual reboot, a repair, or replacement of the failed component(s).
- Proactive risk analysis performed as part of the BC planning process considers the component failure rate and average repair time, which are measured by MTBF and MTTR:

→ **Mean Time Between Failure (MTBF):** It is the average time available for a system or component to perform its normal operations between failures.

→ **Mean Time To Repair (MTTR):** It is the average time required to repair a failed component. MTTR includes the total time required to do the following activities: Detect the fault, mobilize the maintenance team, diagnose the fault, obtain the spare parts, repair, test, and restore the data.

Fig 9.2 illustrates the various information availability metrics that represent system uptime and downtime.

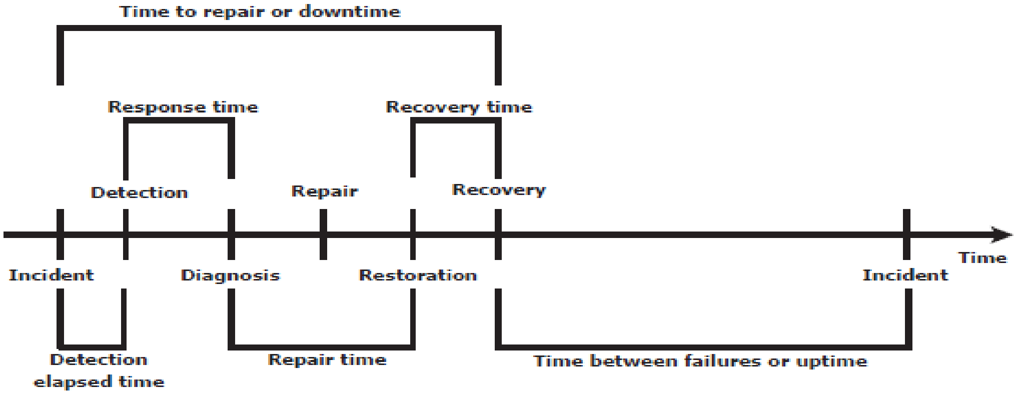


Fig 9.2: Information availability metrics

IA is the time period that a system is in a condition to perform its intended function upon demand. It can be expressed in terms of system uptime and downtime and measured as the amount or percentage of system uptime:

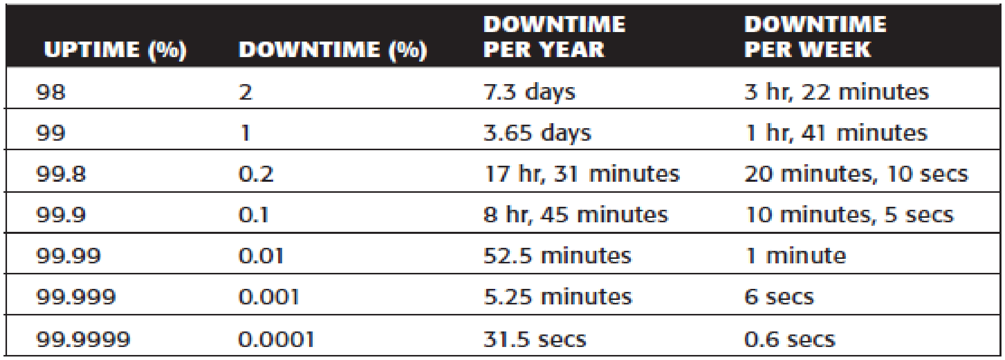
IA = system uptime / (system uptime + system downtime)

In terms of MTBF and MTTR, IA could also be expressed as

IA = MTBF / (MTBF + MTTR)

Uptime per year is based on the exact timeliness requirements of the service, this calculation leads to the number of “9s” representation for availability metrics.

Table 3-1 lists the approximate amount of downtime allowed for a service to achieve certain levels of 9s availability. For example, a service that is said to be “five 9s available” is available for 99.999 percent of the scheduled time in a year (24 × 365).



UPTIME (%)	DOWNTIME (%)	DOWNTIME PER YEAR	DOWNTIME PER WEEK
98	2	7.3 days	3 hr, 22 minutes
99	1	3.65 days	1 hr, 41 minutes
99.8	0.2	17 hr, 31 minutes	20 minutes, 10 secs
99.9	0.1	8 hr, 45 minutes	10 minutes, 5 secs
99.99	0.01	52.5 minutes	1 minute
99.999	0.001	5.25 minutes	6 secs
99.9999	0.0001	31.5 secs	0.6 secs

Table 9-1: Availability percentage and Allowable downtime

9.2 BC Terminology

This section defines common terms related to BC operations which are used in this module to explain advanced concepts:

- **Disaster recovery:** This is the coordinated process of restoring systems, data, and the infrastructure required to support key ongoing business operations in the event of a disaster.

- It is the process of restoring a previous copy of the data and applying logs or other necessary processes to that copy to bring it to a known point of consistency. Once all recoveries are completed, the data is validated to ensure that it is correct.
- **Disaster restart:** This is the process of restarting business operations with mirrored consistent copies of data and applications.
- **Recovery-Point Objective (RPO):** This is the point in time to which systems and data must be recovered after an outage. It defines the amount of data loss that a business can endure. A large RPO signifies high tolerance to information loss in a business. Based on the RPO, organizations plan for the minimum frequency with which a backup or replica must be made. For example, if the RPO is six hours, backups or replicas must be made at least once in 6 hours. Fig 9.3 (a) shows various RPOs and their corresponding ideal recovery strategies. An organization can plan for an appropriate BC technology solution on the basis of the RPO it sets. For example:

  → **RPO of 24 hours:** This ensures that backups are created on an offsite tape drive every midnight. The corresponding recovery strategy is to restore data from the set of last backup tapes.

→ **RPO of 1 hour:** Shipping database logs to the remote site every hour. The corresponding recovery strategy is to recover the database at the point of the last log shipment.

→ **RPO in the order of minutes:** Mirroring data asynchronously to a remote site

→ **Near zero RPO:** This mirrors mission-critical data synchronously to a remote site.

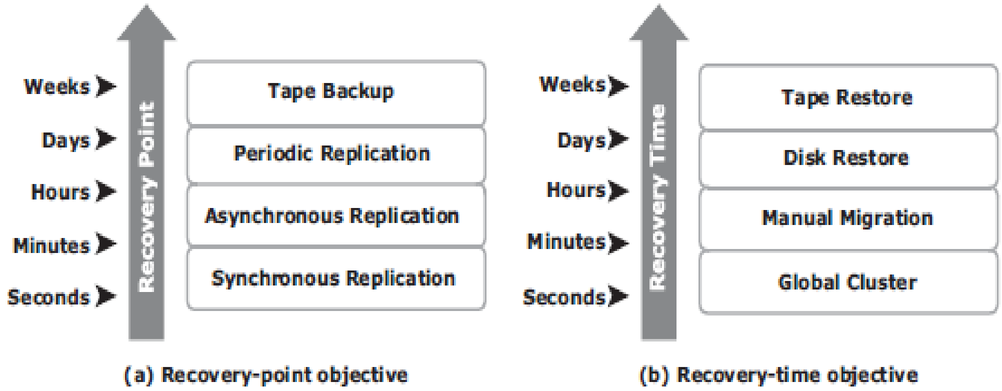


Fig 9.3: Strategies to meet RPO and RTO targets

- **Recovery-Time Objective (RTO):** The time within which systems and applications must be recovered after an outage. It defines the amount of downtime that a business can endure and survive. Businesses can optimize disaster recovery plans after defining the RTO for a given system. For example, if the RTO is two hours, then use a disk backup because it enables a faster restore than a tape backup. However, for an RTO of one week, tape backup will likely meet requirements. Some examples of RTOs and the recovery strategies to ensure data availability are listed below (refer to Fig 9.3 (b)):

  → **RTO of 72 hours:** Restore from backup tapes at a cold site.

  → **RTO of 12 hours:** Restore from tapes at a hot site.

  → **RTO of few hours:** Use a data vault to a hot site.

  → **RTO of a few seconds:** Cluster production servers with bidirectional mirroring, enabling the applications to run at both sites simultaneously.

- **Data vault:** A repository at a remote site where data can be periodically or continuously copied (either to tape drives or disks) so that there is always a copy at another site

- **Hot site:** A site where an enterprise’s operations can be moved in the event of disaster. It is a site with the required hardware, operating system, application, and network support to perform business operations, where the equipment is available and running at all times.

- **Cold site:** A site where an enterprise’s operations can be moved in the event of disaster, with minimum IT infrastructure and environmental facilities in place, but not activated.

- **Server Clustering:** A group of servers and other necessary resources coupled to operate as a single system. Clusters can ensure high availability and load balancing. Typically, in failover clusters, one server runs an application and updates the data, and another server is kept as standby to take over completely, as required. In more sophisticated clusters, multiple servers may access data, and typically one server is kept as standby. Server clustering provides load balancing by distributing the application load evenly among multiple servers within the cluster.

9.3 BC Planning Life Cycle

BC planning must follow a disciplined approach like any other planning process. Organizations today dedicate specialized resources to develop and maintain BC plans. From the conceptualization to the realization of the BC plan, a life cycle of activities can be defined for the BC process.

The BC planning lifecycle includes five stages shown below (Fig 9.4):

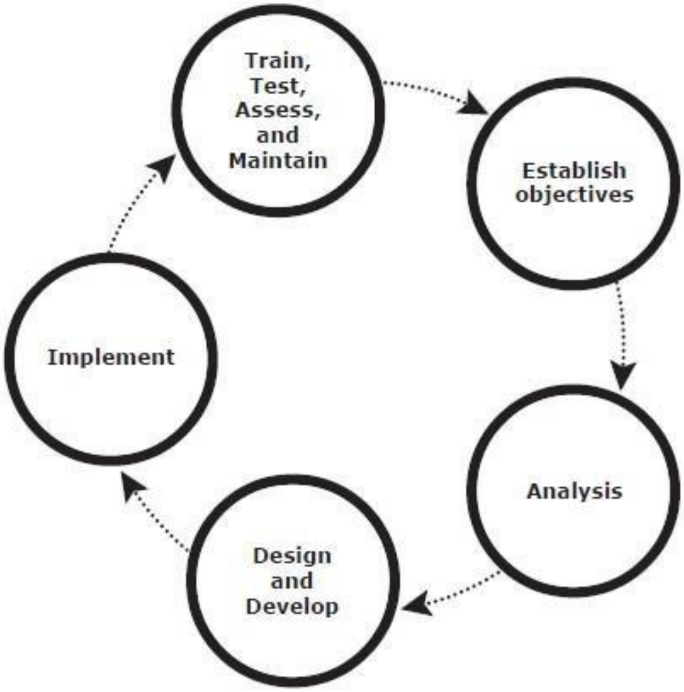


Fig 9.4: BC Planning Lifecycle

Several activities are performed at each stage of the BC planning lifecycle, including the following key activities:

1. **Establishing objectives**

→ Determine BC requirements.

→ Estimate the scope and budget to achieve requirements.

→ Select a BC team by considering subject matter experts from all areas of the business, whether internal or external.

→ Create BC policies.

2. Analyzing

→ Collect information on data profiles, business processes, infrastructure support, dependencies, and frequency of using business infrastructure.

→ Identify critical business needs and assign recovery priorities.

→ Create a risk analysis for critical areas and mitigation strategies.

→ Conduct a Business Impact Analysis (BIA).

→ Create a cost and benefit analysis based on the consequences of data unavailability.

3. Designing and developing

→ Define the team structure and assign individual roles and responsibilities. For example, different teams are formed for activities such as emergency response, damage assessment, and infrastructure and application recovery.

→ Design data protection strategies and develop infrastructure.

→ Develop contingency scenarios.

→ Develop emergency response procedures.

→ Detail recovery and restart procedures.

4. Implementing

→ Implement risk management and mitigation procedures that include backup, replication, and management of resources.

→ Prepare the disaster recovery sites that can be utilized if a disaster affects the primary data center.

→ Implement redundancy for every resource in a data center to avoid single points of failure.

5. Training, testing, assessing, and maintaining

→ Train the employees who are responsible for backup and replication of business-critical data on a regular basis or whenever there is a modification in the BC plan

→ Train employees on emergency response procedures when disasters are declared.

→ Train the recovery team on recovery procedures based on contingency scenarios.

→ Perform damage assessment processes and review recovery plans.

→ Test the BC plan regularly to evaluate its performance and identify its limitations.

→ Assess the performance reports and identify limitations.

→ Update the BC plans and recovery/restart procedures to reflect regular changes within the data center.

9.4 Failure Analysis

9.4.1 Single Point of Failure

- A **single point of failure** refers to the failure of a component that can terminate the availability of the entire system or IT service.
- Fig 9.5 depicts a system setup in which an application, running on a VM, provides an interface to the client and performs I/O operations.
- The client is connected to the server through an IP network, the server is connected to the storage array through a FC connection, an HBA installed at the server sends or receives data to and from a storage array, and an FC switch connects the HBA to the storage port
- In a setup where **each component must function as required to ensure data availability**, the failure of a single physical or virtual component causes the failure of the entire data center or an application, resulting in disruption of business operations.
- In this example, failure of a hypervisor can affect all the running VMs and the virtual network, which are hosted on it.
- The can be several similar single points of failure identified in this example. A VM, a hypervisor, an HBA/NIC on the server, the physical server, the IP network, the FC switch, the storage array ports, or even the storage array could be a potential single point of failure. To avoid single points of failure, it is essential to implement a fault-tolerant mechanism.

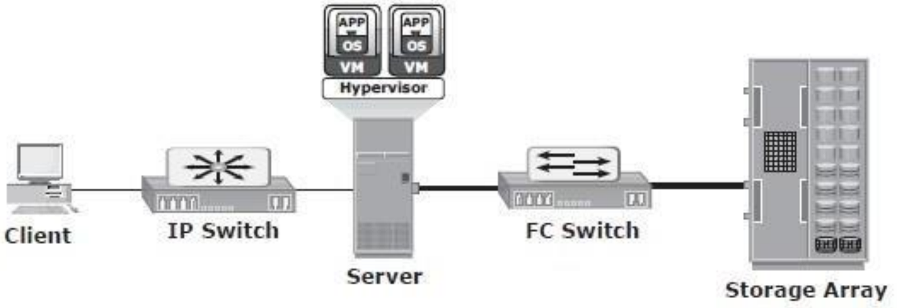


Fig 9.5: Single Point of Failure

9.4.2 Resolving Single Points of Failure

- To mitigate a single point of failure, systems are designed with redundancy, such that the system will fail only if all the components in the redundancy group fail. This ensures that the failure of a single component does not affect data availability.
- Data centers follow stringent guidelines to implement fault tolerance for uninterrupted information availability. Careful analysis is performed to eliminate every single point of failure.
- The example shown in Fig 9.6 represents all enhancements of the system shown in Fig 9.5 in the infrastructure to mitigate single points of failure:
  - Configuration of redundant HBAs at a server to mitigate single HBA failure
  - Configuration of NIC (network interface card) teaming at a server allows protection against single physical NIC failure. It allows grouping of two or more physical NICs and treating them as a single logical device. NIC teaming eliminates the single point of failure associated with a single physical NIC.
  - Configuration of redundant switches to account for a switch failure
  - Configuration of multiple storage array ports to mitigate a port failure
  - RAID and hot spare configuration to ensure continuous operation in the event of disk failure
  - Implementation of a redundant storage array at a remote site to mitigate local site failure
  - Implementing server (or compute) clustering, a fault-tolerance mechanism whereby two or more servers in a cluster access the same set of data volumes. Clustered servers exchange a heartbeat to inform each other about their health. If one of the servers or hypervisors fails, the other server or hypervisor can take up the workload.
  - Implementing a VM Fault Tolerance mechanism ensures BC in the event of a server failure. This technique creates duplicate copies of each VM on another server so that when a VM failure is detected, the duplicate VM can be used for failover. The two VMs are kept in synchronization with each other in order to perform successful failover.

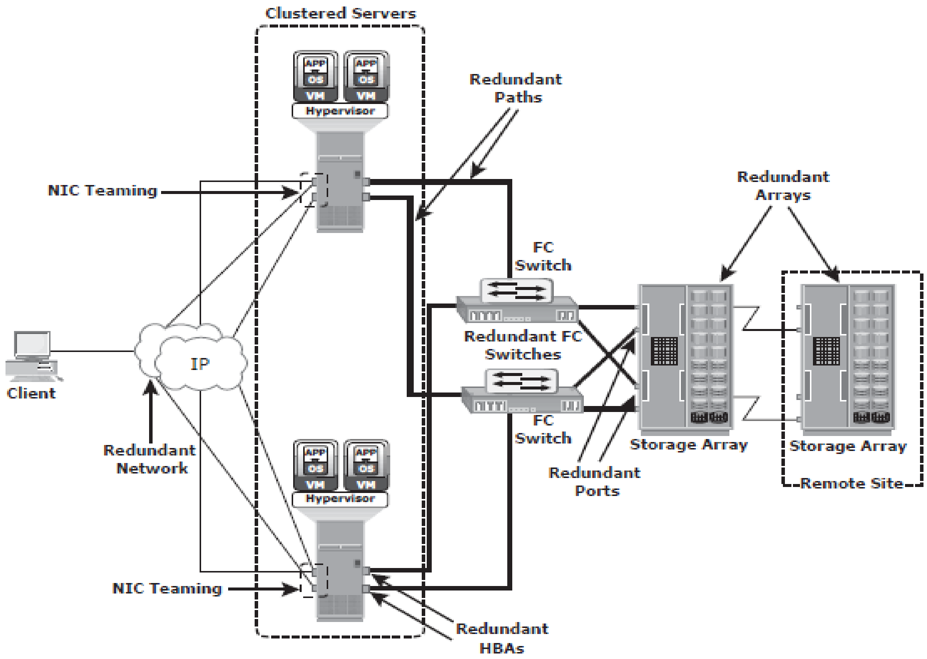


Fig 9.6: Resolving single points of failure

9.4.3 Multipathing Software

- Configuration of multiple paths increases the data availability through path failover. If servers are configured with one I/O path to the data there will be no access to the data if that path fails. Redundant paths eliminate the path to become single points of failure.
- Multiple paths to data also improve I/O performance through load sharing and maximize server, storage, and data path utilization.
- In practice, merely configuring multiple paths does not serve the purpose. Even with multiple paths, if one path fails, I/O will not reroute unless the system recognizes that it has an alternate path.
- Multipathing software provides the functionality to recognize and utilize alternate I/O path to data. Multipathing software also manages the load balancing by distributing I/Os to all available, active paths.
- In a virtual environment, multipathing is enabled either by using the hypervisor's built-in capability or by running a third-party software module, added to the hypervisor.

9.5 **Business Impact Analysis**

- A business impact analysis (BIA) identifies which business units, operations, and processes are essential to the survival of the business.

- It evaluates the financial, operational, and service impacts of a disruption to essential business processes. Selected functional areas are evaluated to determine resilience of the infrastructure to support information availability.

- The BIA process leads to a report detailing the incidents and their impact over business functions. The impact may be specified in terms of money or in terms of time. Based on the potential impacts associated with downtime, businesses can prioritize and implement countermeasures to mitigate the likelihood of such disruptions.

- These are detailed in the BC plan. A BIA includes the following set of tasks:
- Determine the business areas.
- For each business area, identify the key business processes critical to its operation.
- Determine the attributes of the business process in terms of applications, databases, and hardware and software requirements.
- Estimate the costs of failure for each business process.
- Calculate the maximum tolerable outage and define RTO and RPO for each business process.
- Establish the minimum resources required for the operation of business processes.
- Determine recovery strategies and the cost for implementing them.
- Optimize the backup and business recovery strategy based on business priorities.
- Analyze the current state of BC readiness and optimize future BC planning.

9.6 BC Technology Solutions

After analyzing the business impact of an outage, designing appropriate solutions to recover from a failure is the next important activity. One or more copies of the original data are maintained using any of the following strategies, so that data can be recovered and business operations can be restarted using an alternate copy:

1. **Backup:** Data backup is a predominant method of ensuring data availability. The frequency of backup is determined based on RPO, RTO, and the frequency of data changes.
2. **Local replication:** Data can be replicated to a separate location within the same storage array. The replica is used independently for other business operations. Replicas can also be used for restoring operations if data corruption occurs.
3. **Remote replication:** Data in a storage array can be replicated to another storage array located at a remote site. If the storage array is lost due to a disaster, business operations can be started from the remote storage array.

CHAPTER 10: Backup and Archive

- **Data Backup** is a copy of production data, created and retained for the sole purpose of recovering lost or corrupted data.
- Evaluating the various backup methods along with their recovery considerations and retention requirements is an essential step to implement a successful backup and recovery solution.
- Organizations generate and maintain large volumes of data, and most of the data is fixed content. This fixed content is rarely accessed after a period of time. Still, this data needs to be retained for several years to meet regulatory compliance.
- **Data archiving** is the process of moving data that is no longer actively used, from primary storage to a low-cost secondary storage. This data is retained in the secondary storage for a long term to meet regulatory requirements. This reduces the amount of data to be backed up and the time required to back up the data.

10.1Backup Purpose

Backups are performed to serve three purposes: ***disaster recovery, operational recovery, and archival***. These are discussed in the following sections.

10.1.1 Disaster Recovery

- Backups are 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 RPO and RTO requirements, organizations use different backup strategies for disaster recovery.
- When a tape-based backup method is used as a disaster recovery strategy, the backup tape media is shipped and stored at an offsite location. These tapes can be recalled for restoration at the disaster recovery site.
- Organizations with stringent RPO and RTO requirements use remote replication technology to replicate data to a disaster recovery site. Organizations can bring production systems online in a relatively short period of time if a disaster occurs.

10.1.2 Operational Recovery

- Data in the production environment changes with every business transaction and operation.
- Operational recovery is the use of backups to restore data if data loss or logical

corruption occurs during routine processing.

- For example, it is common for a user to accidentally delete an important email or for a file to become corrupted, which can be restored from operational backup.

10.1.3 Archival

- Backups are also performed to address archival requirements.
- 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.

Apart from addressing disaster recovery, archival, and operational requirements, backups serve as a protection against data loss due to physical damage of a storage device, software failures, or virus attacks. Backups can also be used to protect against accidents such as a deletion or intentional data destruction.

10.2Backup Considerations

- 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.
- RPO refers to the point in time to which data must be recovered, and the point in time from which to restart business operations. This specifies the time interval between two backups. In other words, the RPO determines backup frequency.
- The backup media type or backup target is another consideration, that is driven by RTO and impacts the data recovery time. The time-consuming operation of starting and stopping in a tape-based system affects the backup performance, especially while backing up a large number of small files.
- Organizations must also consider the granularity of backups, explained later in section “10.3 Backup Granularity.” 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 file size and number of files also influence 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 section “10.11 Data Deduplication for Backup”) are widely used in the backup environment because these

technologies save space on the media.

10.3Backup Granularity

- Backup granularity depends on business needs and the required RTO/RPO. Based on the granularity, backups can be categorized as full, incremental and cumulative (differential).
- Most organizations use a combination of these three backup types to meet their backup and recovery requirements. Figure 10-1 shows the different backup granularity levels.

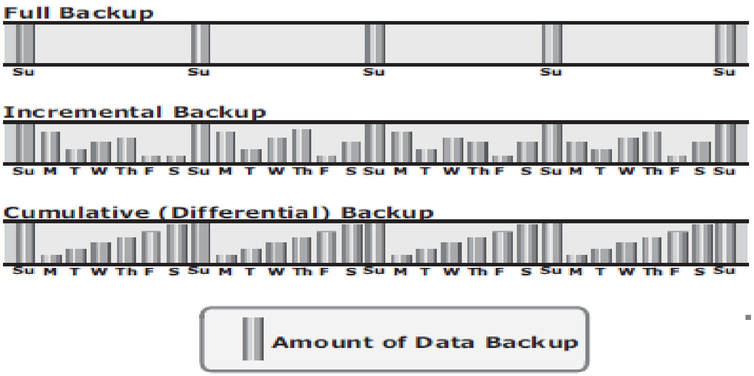


Figure 10-1: Backup granularity levels

- 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.
- 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.
- Restore operations vary with the granularity of the backup. A full backup provides a single repository from which the data can be easily restored. The process of restoration from an incremental backup requires the last full backup and all the incremental backups available until the point of restoration. A restore from a cumulative backup requires the last full backup and the most recent cumulative backup.
- Figure 10-2 shows an example of restoring data from incremental backup.

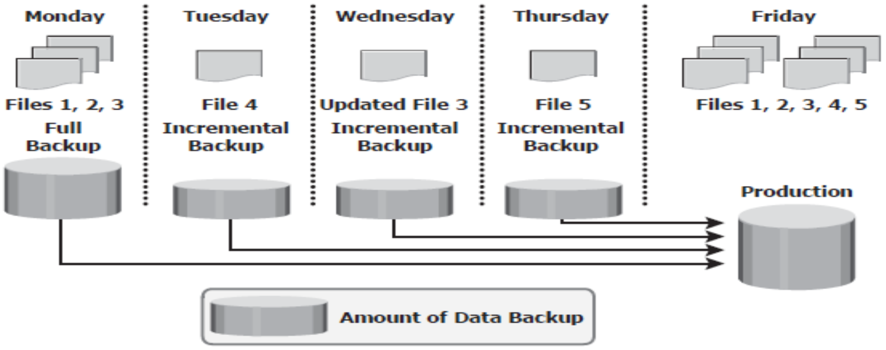


Figure 10-2: Restoring from an incremental backup

In this 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 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.

- Figure 10-3 shows an example of restoring data from cumulative backup

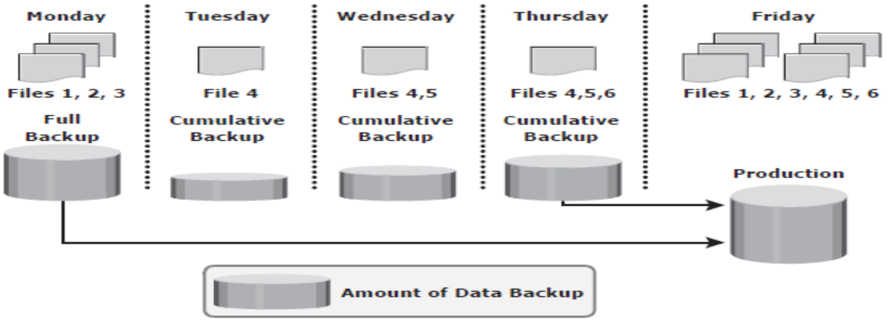


Figure 10-3: Restoring a cumulative backup

In this 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

10.4Recovery Considerations

- The retention period is a key consideration for recovery. The retention period for a backup is derived from an RPO.
- For example, users of an application might request to restore the application data from its backup copy, which was created a month ago. This determines the retention period for the backup. Therefore, the minimum retention period of this application data is one month.
- If the recovery point is older than the retention period, it might not be possible to recover all the data required for the requested recovery point. Long retention periods can be defined for all backups, making it possible to meet any RPO within the defined retention periods.
- RTO relates to the time taken by the recovery process. To meet the defined RTO, the business may choose the appropriate backup granularity to minimize recovery time.
- In a backup environment, RTO influences the type of backup media that should be used. For example, a restore from tapes takes longer to complete than a restore from disks.

10.5Backup Methods

- **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 an *online backup***.**
- In a **cold backup**, the application is not active or shutdown during the backup process and is also called as *offline backup.*
- The hot backup of online production data becomes more challenging because data is actively used and changed.
- An open file is locked by the operating system and is not backed up during the backup process. In such situations, an *open file agent* is required to back up the open file.
- In database environments, the use of open file agents is not enough, because the agent should also support a consistent backup of all the database components.
- For example, a database is composed of many files of varying sizes occupying several file systems. To ensure a consistent database backup, all files need to be backed up in the same state. That does not necessarily mean that all files need to be backed up at the same time, but they all must be synchronized so that the database can be restored with consistency.
- 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.
- Hot backup is used in situations where it is not possible to shut down the database. This is facilitated by database backup agents that can perform a backup while the database is active. The disadvantage associated with a hot backup is that the agents usually affect overall application performance.
- A **point-in-time (PIT)** copy method is deployed in environments where the impact of downtime from a cold backup or the performance resulting from a hot backup is unacceptable. The PIT copy is created from the production volume and used as the source for the backup. This reduces the impact on the production volume.
- Certain attributes and properties attached to a file, such as permissions, owner, and other metadata, also need to be backed up. These attributes are as important as the data itself and must be backed up for consistency.
- Backup of boot sector and partition layout information is also critical for successful recovery.
- In a disaster recovery environment, **bare-metal recovery (BMR)** refers to a backup in which all metadata, system information, 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 can recover a server onto dissimilar hardware.

10.6 Backup Architecture

- A backup system commonly uses the client-server architecture with a backup server and multiple backup clients. Figure 10-4 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 to the storage node as the media server because it manages the storage device.

- Backup software provides reporting capabilities based on the backup catalog and the log files.

- These reports include information, such as the amount of data backed up, the number of completed and incomplete backups, and the types of errors that might have occurred. Reports can be customized depending on the specific backup software used.

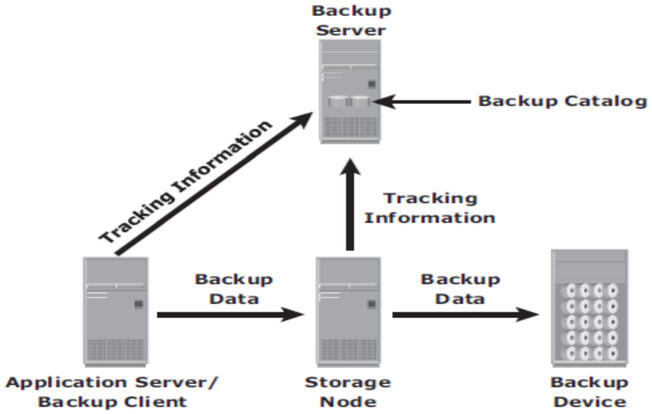


Figure 10-4: Backup architecture

10.6Backup and Restore Operations

- 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 11:00 p.m. every day.
- The backup server coordinates the backup process with all the components in a backup environment (see Figure 10-5).

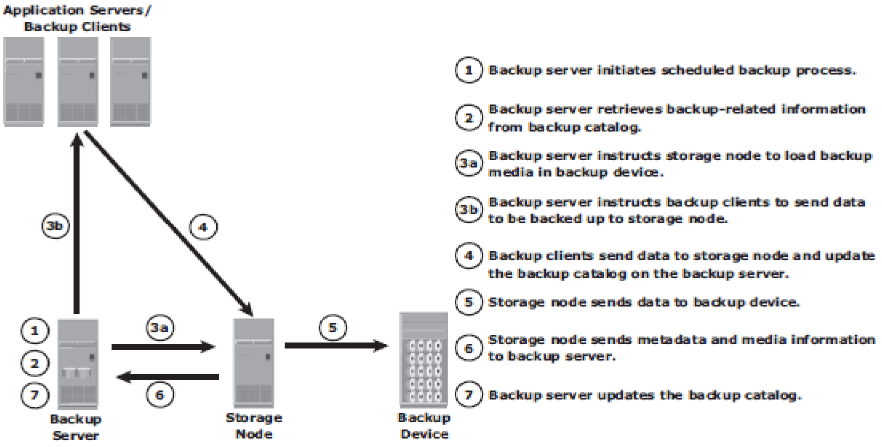
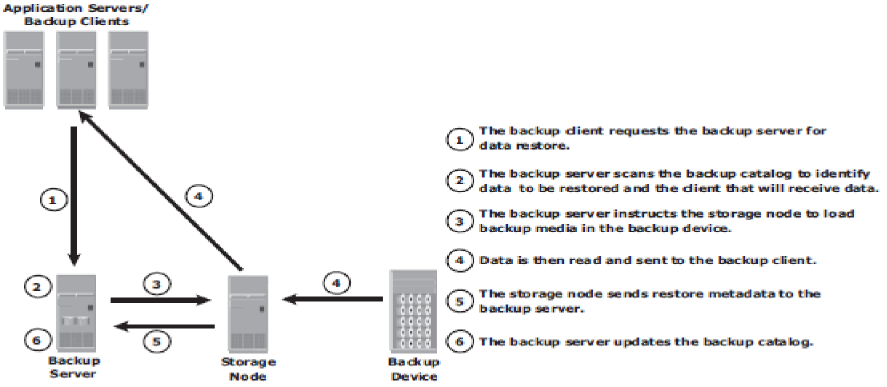

Application Servers/ Backup Clients
3b
4
1
2
7
Backup Server
3a
6
Storage Node
5
Backup Device
1 Backup server initiates scheduled backup process.
2 Backup server retrieves backup-related information from backup catalog.
3a Backup server instructs storage node to load backup media in backup device.
3b Backup server instructs backup clients to send data to be backed up to storage node.
4 Backup clients send data to storage node and update the backup catalog on the backup server.
5 Storage node sends data to backup device.
6 Storage node sends metadata and media information to backup server.
7 Backup server updates the backup catalog.

Figure 10-5: Backup operation

- 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 10-6 shows a restore operation.



1 The backup client requests the backup server for data restore.

2 The backup server scans the backup catalog to identify data to be restored and the client that will receive data.

3 The backup server instructs the storage node to load backup media in the backup device.

4 Data is then read and sent to the backup client.

5 The storage node sends restore metadata to the backup server.

6 The backup server updates the backup catalog.

Figure 10-6: 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.

10.8Backup Topologies

- Three basic topologies are used in a backup environment:

1. Direct attached backup
2. LAN based backup, and
3. SAN based backup.

- A **mixed topology** is also used by combining LAN based and SAN based topologies.
- In a **direct-attached backup**, a 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.
- The example shown in Fig 10-7 device is directly attached and dedicated to the backup client. As the environment grows, however, there will be a need for central management of all backup devices and to share the resources to optimize costs. An appropriate solution is 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.

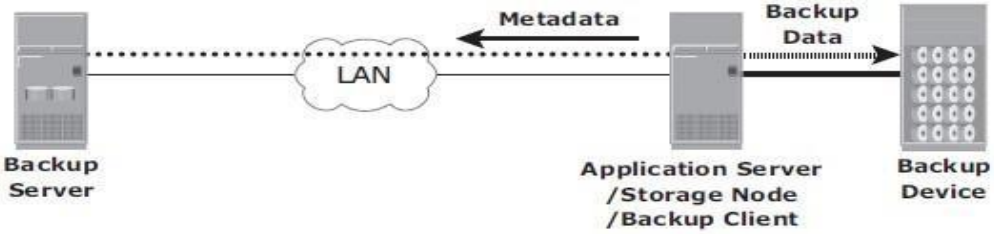


Fig 10-7: Direct-attached backup topology

- In **LAN-based backup**, the clients, backup server, storage node, and backup device are connected to the LAN (see Fig 10-8). The data to be backed up is transferred from the backup client (source), to the backup device (destination) over the LAN, which may 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.

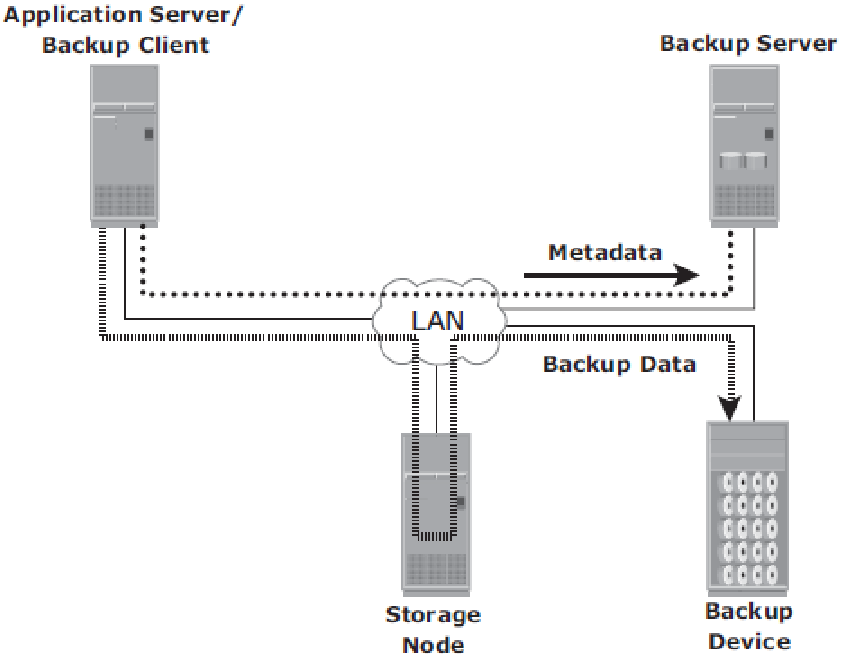


Fig 10-8: LAN-based backup topology

- The **SAN-based backup** is also known as the *LAN-free backup*. Fig 3.9 illustrates a SAN-based backup. The SAN-based backup topology is the most appropriate solution when a backup device needs to be shared among the clients. In this case the backup device and clients are attached to the SAN.
- In the example from Fig 10-9, a client sends the data to be backed up to the backupdevice 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.

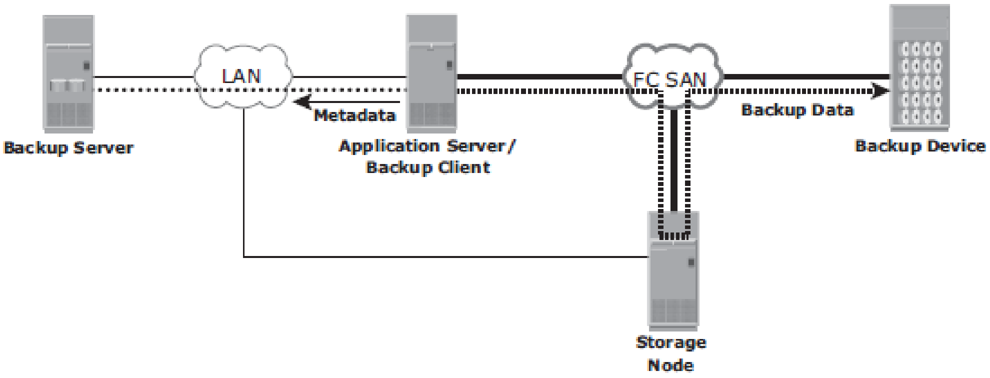


Fig 10-9: SAN-based backup topology

- 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, as shown in Fig 10-10. This topology might be implemented for several reasons, including cost,server location, reduction in administrative overhead, and performance considerations.

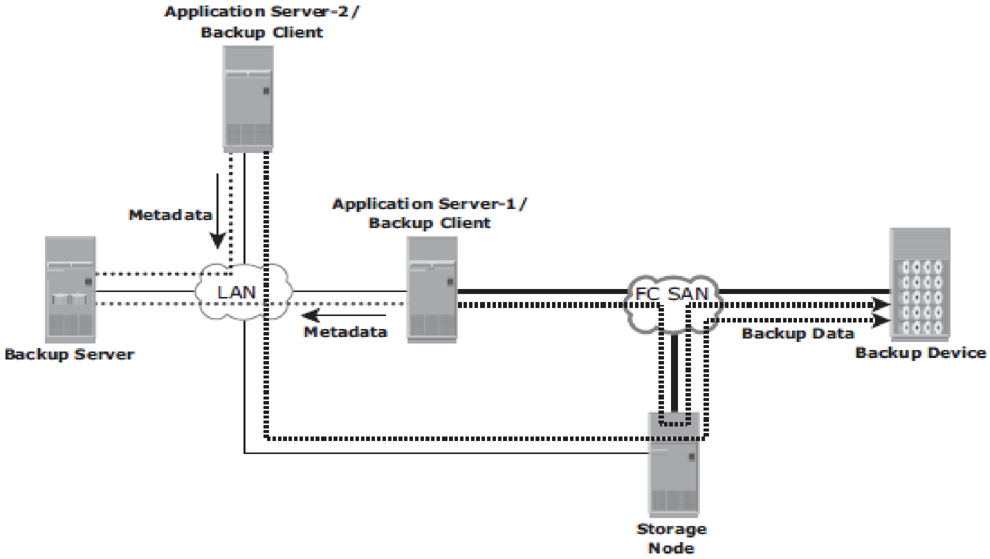


Fig 10-10: Mixed backup topology

10.9Backup in NAS Environments

- 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.

10.9.1Server-Based and Serverless Backup

- 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.

- Figure 10-11 illustrates server-based backup in the NAS environment.

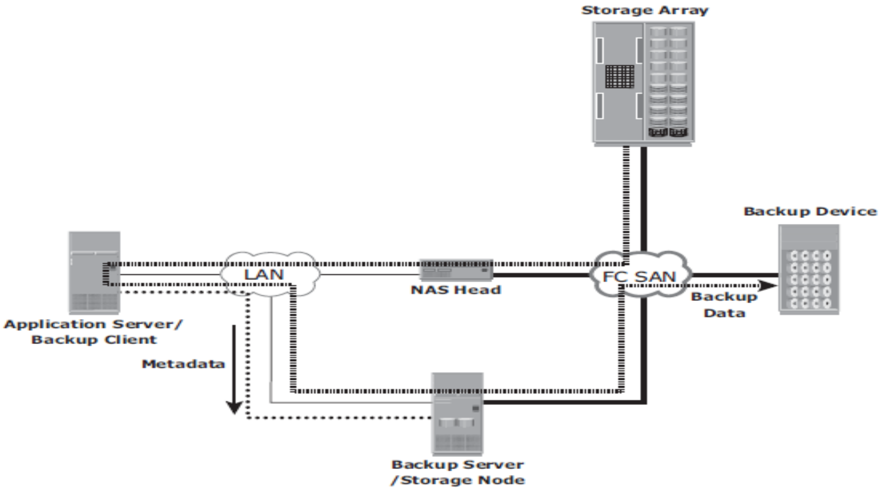


Figure 10-11: Server-based backup in a NAS environment

- 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.
- Figure 10-12 illustrates serverless backup in the NAS environment.
- 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.

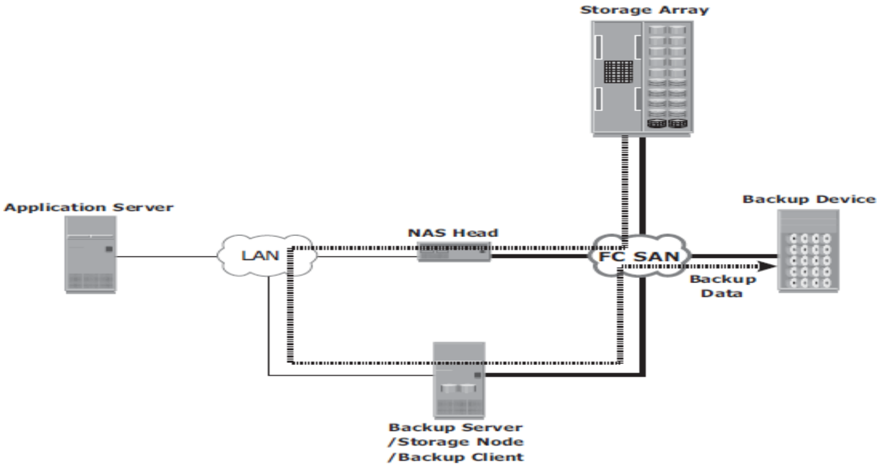


Figure 10-12: Serverless backup in a NAS environment

10.9.2 NDMP-Based Backup

- 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.
- 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 10-13 illustrates a 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.

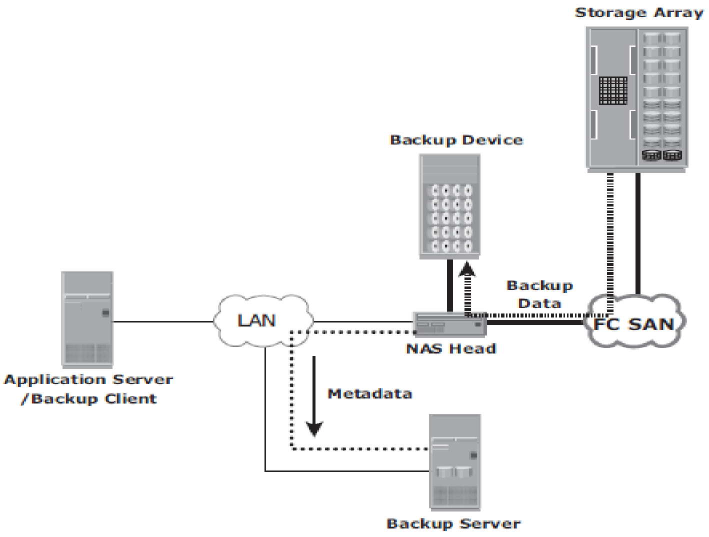


Figure 10-13: NDMP 2-way in a NAS environment

- In the NDMP *3-way method*, 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 10-14 shows a 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.

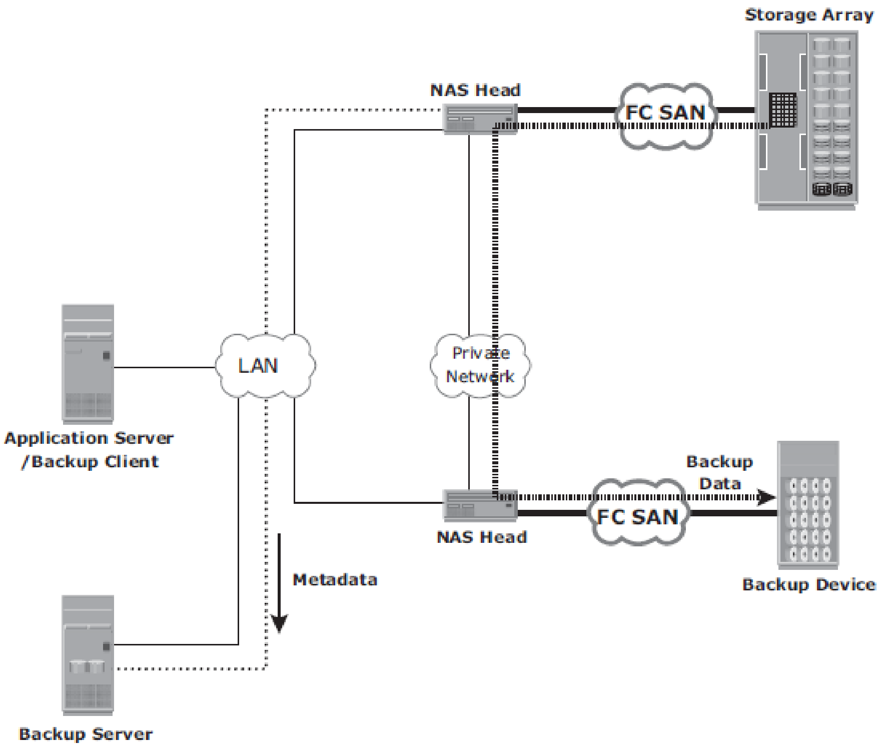


Figure 10-14: NDMP 3-way in a NAS environment