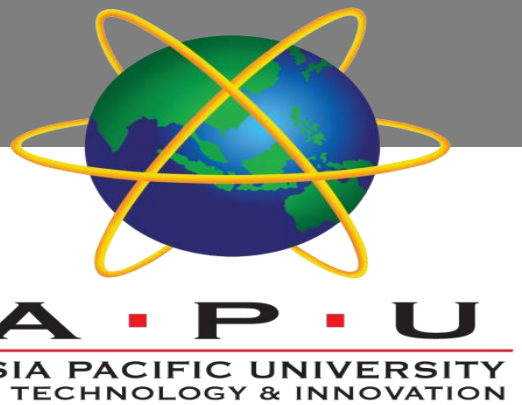
Data Protection and Management



Data Backup -Data deduplication-Replication-Archiving and Data migration

Learning Outcome

Upon completion of this lecture, you should be able to:

- Deploy data backup and data recovery strategy in an organization
- Describe in detail the data deduplication components
- Describe in detail data replication processes
- Differentiate between data backup, data deduplication and data replication

Why Do We Need Data Backup?

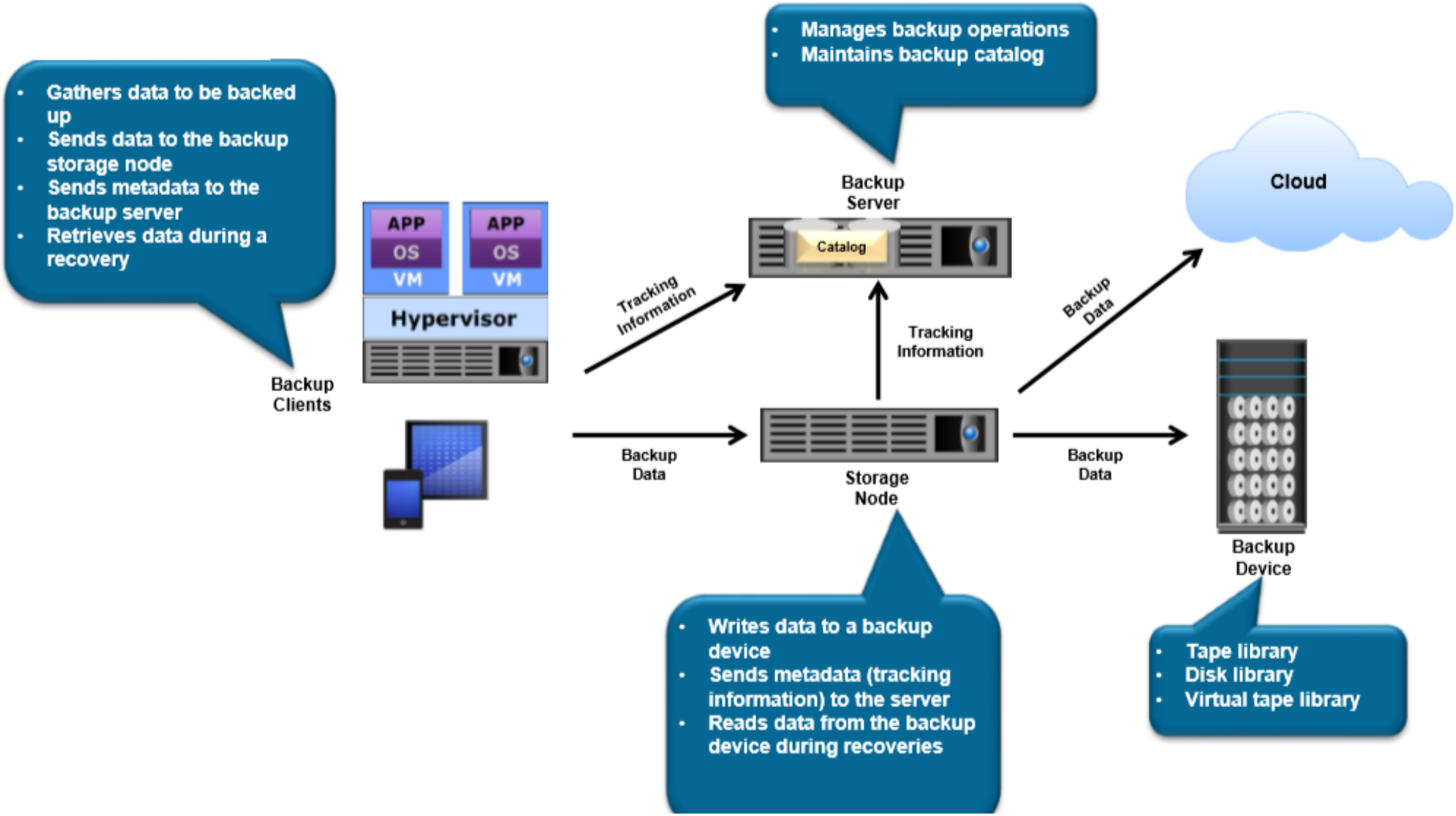
To recover the lost or corrupted data for smooth functioning of business operations

To meet the demanding SLAs

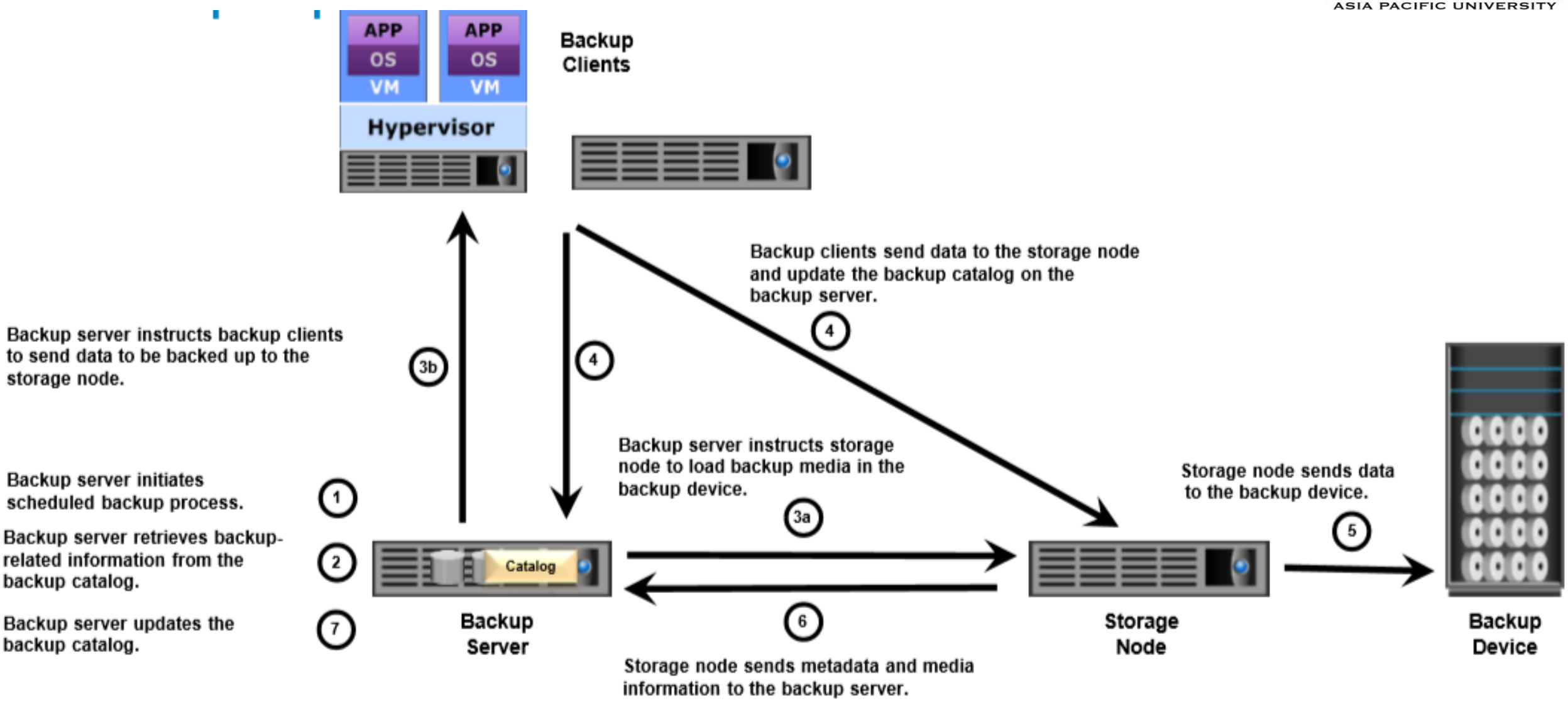
To comply with regulatory requirements

To avoid financial and business loss

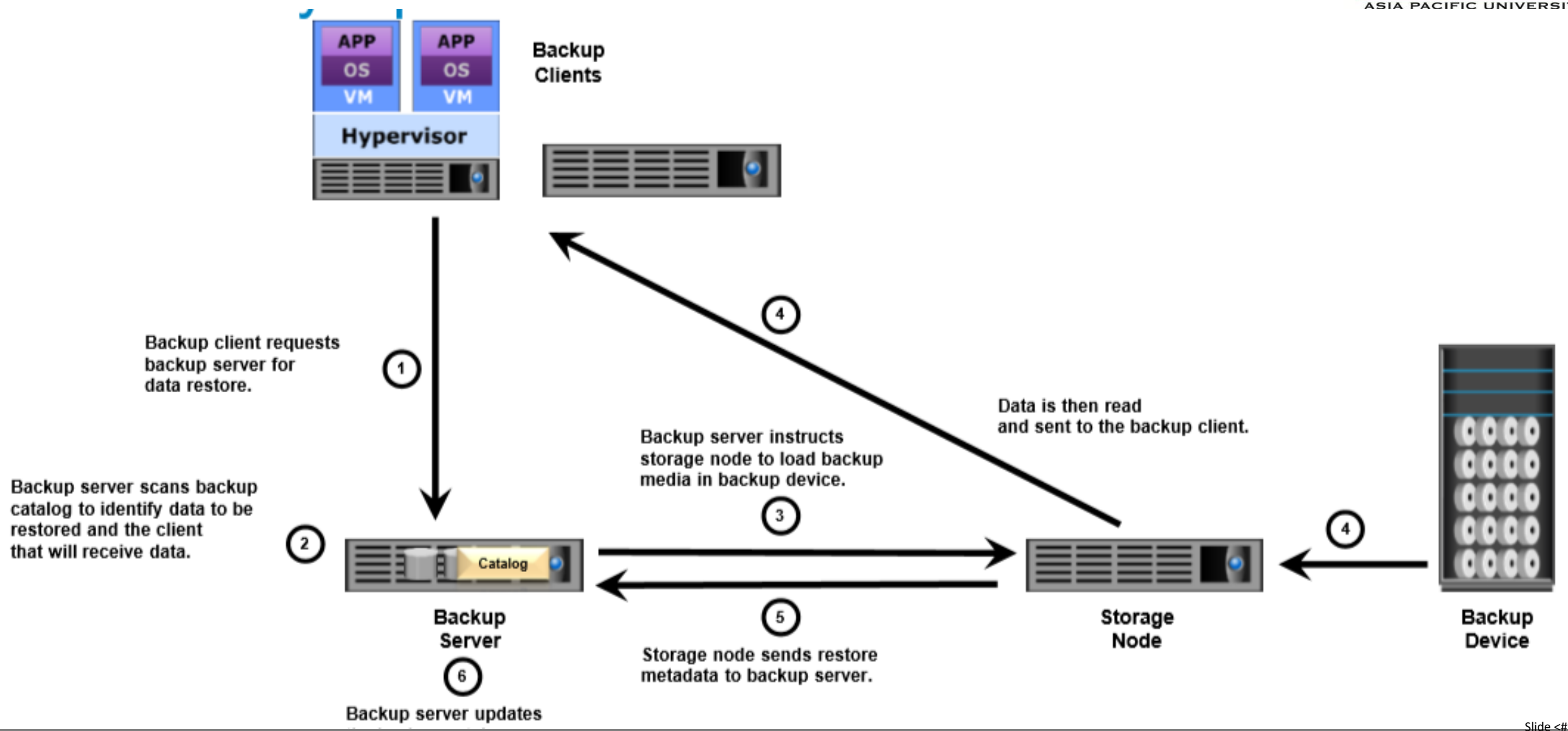
Backup Architecture



Backup Operations



Recovery Operations



Backup Category

Backup granularity depends on business needs and the required RTO/RPO.

Based on the granularity, backups can be categorized as

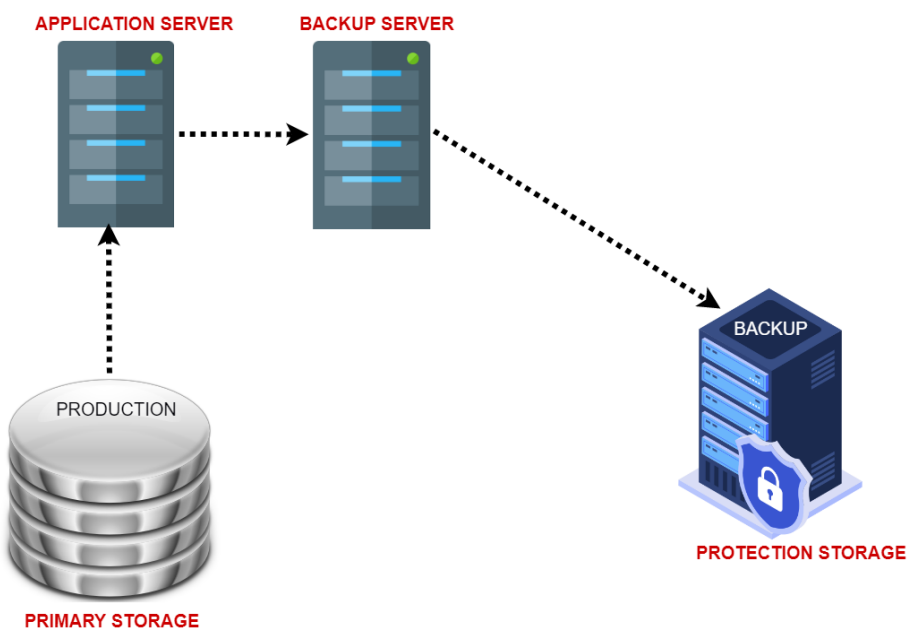
- full,
- incremental,
- cumulative (or differential),

Most organizations use a combination of these backup types to meet their backup and recovery requirements.

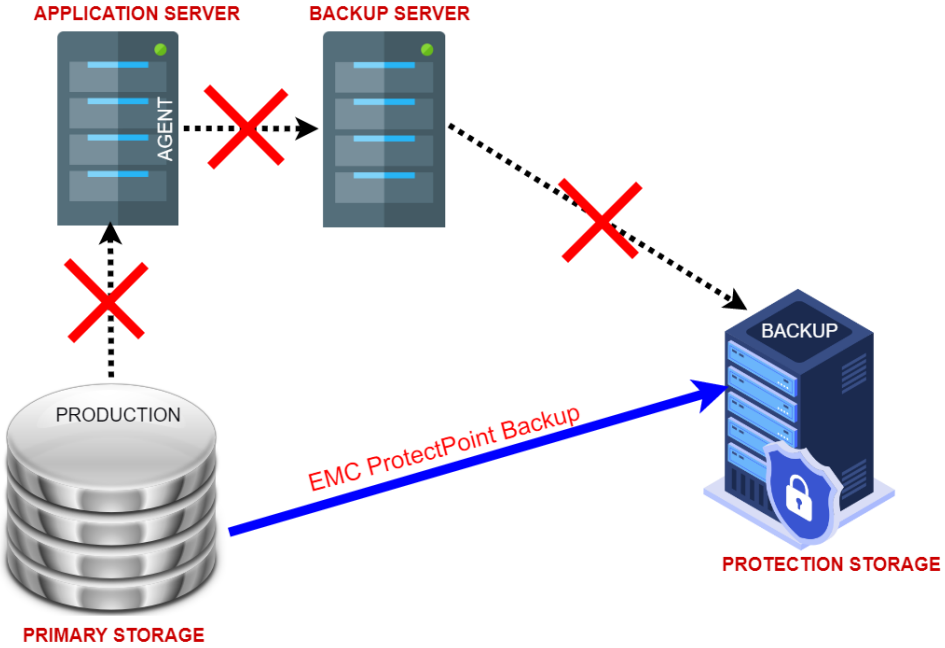
EMC NetWorker and EMC ProtectPoint

NetWorker	ProtectPoint
• Software that centralizes, automates, and accelerates data backup and recovery • Supports multiplexing • Supports source-based and target-based deduplication capabilities by integrating with EMC Avamar and EMC Data Domain respectively	• ProtectPoint backs up data directly from primary storage (EMC VMAX) to Data Domain system • Eliminates the backup impact on the application server • Leverages primary storage change block tracking technology

Traditional Backup VS EMC ProtectPoint Backup



VS



Traditional Backup

ProtectPoint Backup

VMware vSphere Data Protection Advanced

vSphere Data Protection Advanced

- Backup and recovery solution designed for vSphere environments and supported by EMC backup products
- Provides agentless, image-level backups to disk as well as guest-level, application-consistent protection
- Supports network-efficient, encrypted replication to replicate backups to one or more DR sites

VMWare VSpere

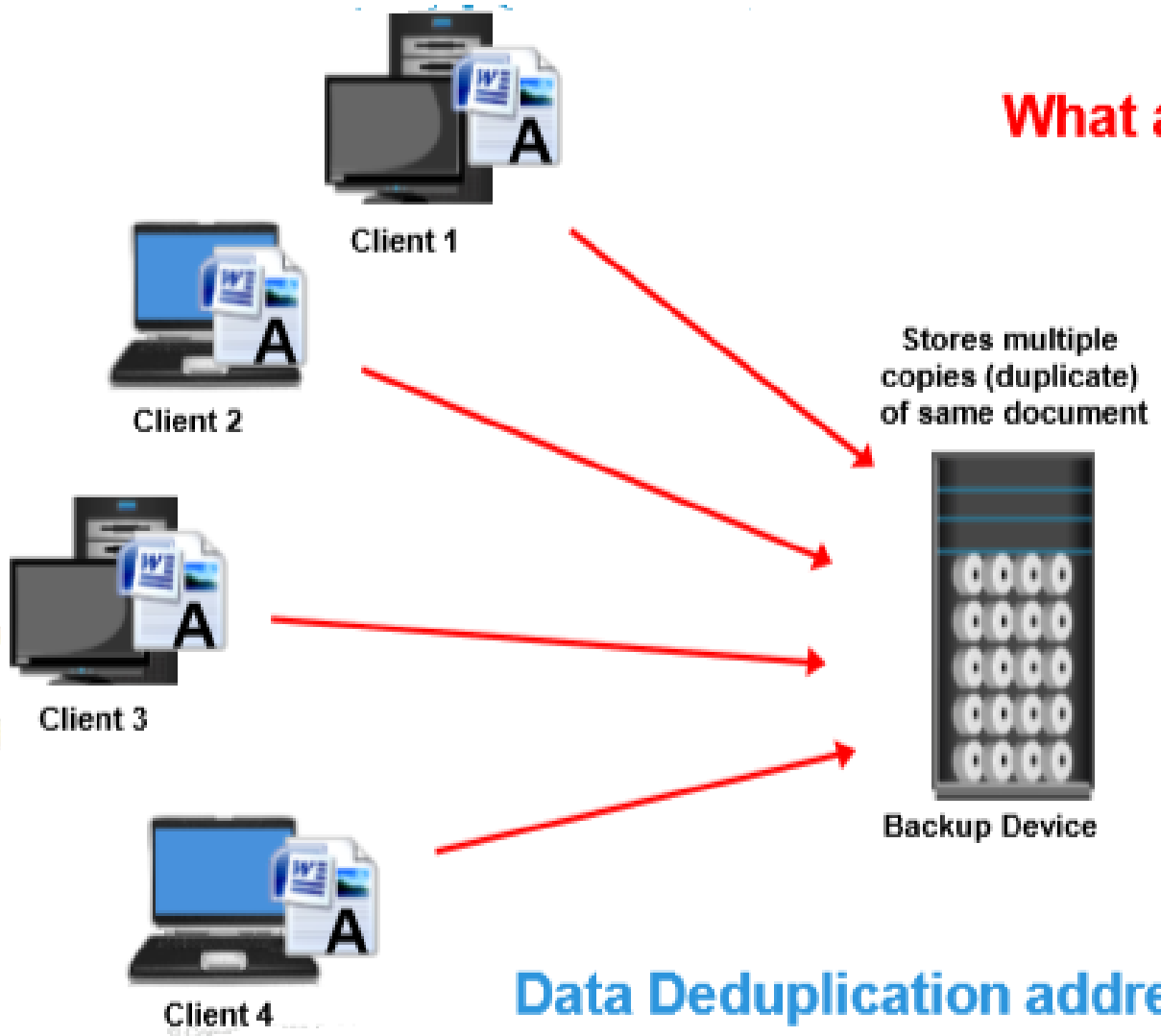
- VMware vSphere for Beginners – YouTube
- https://www.youtube.com/watch?v=3OvrKZYnzjM

Quick Review

- Using Box-and line diagram design;
  - a basic data backup architecture
  - a basic data restore architecture
- What is different between Traditional Backup and ProtectPoint Backup?
- Why it is said that ProtectPoint Backup is superior than Traditional Backup?

Deduplication

Why Do We Need Data Deduplication?

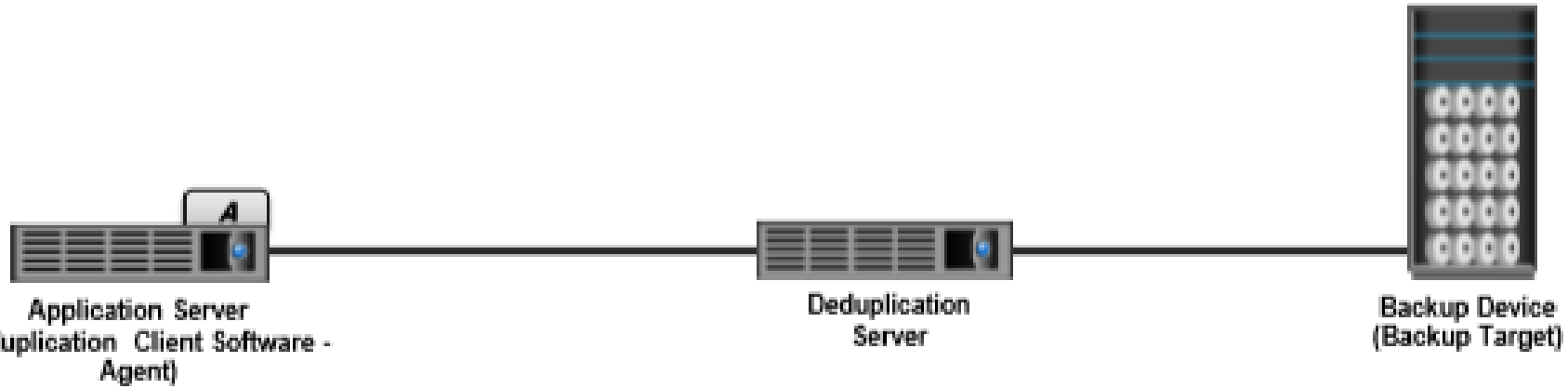


What are the challenges of duplicate data in a data center?

- Difficult to protect the data within the budget
- Impacts the backup window
- Increases the network bandwidth

Data Deduplication addresses these challenges

Key Data Deduplication Components



Deduplication Client Software (Agent)

- Deduplication client software is installed on the application server to perform backup and deduplication

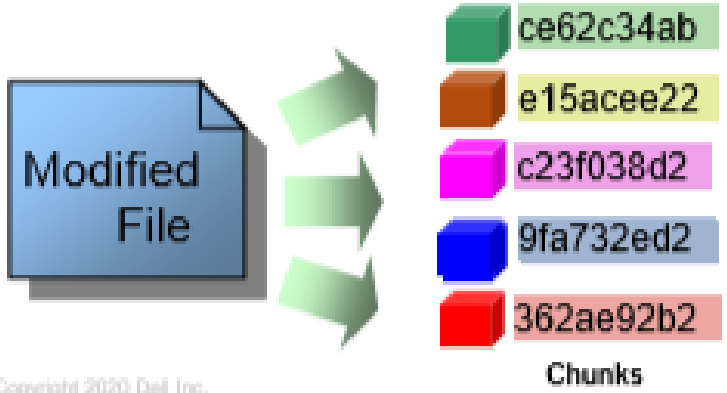
Deduplication Server

- Maintains the index of the deduplicated data
- Deduplication server software can be installed on a general purpose server that accesses the backup target available in the environment
- In some implementation, the deduplication server along with backup target comes as an appliance.

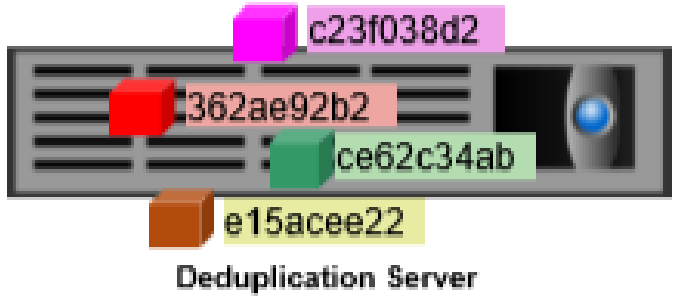
Data Deduplication and Backup Process

1. Client agent checks the file system and determines if a file has been backed up before

2. Modified files are broken into chunks and hashed. Hashes are compared to hash cache



3. Hashes are checked against server



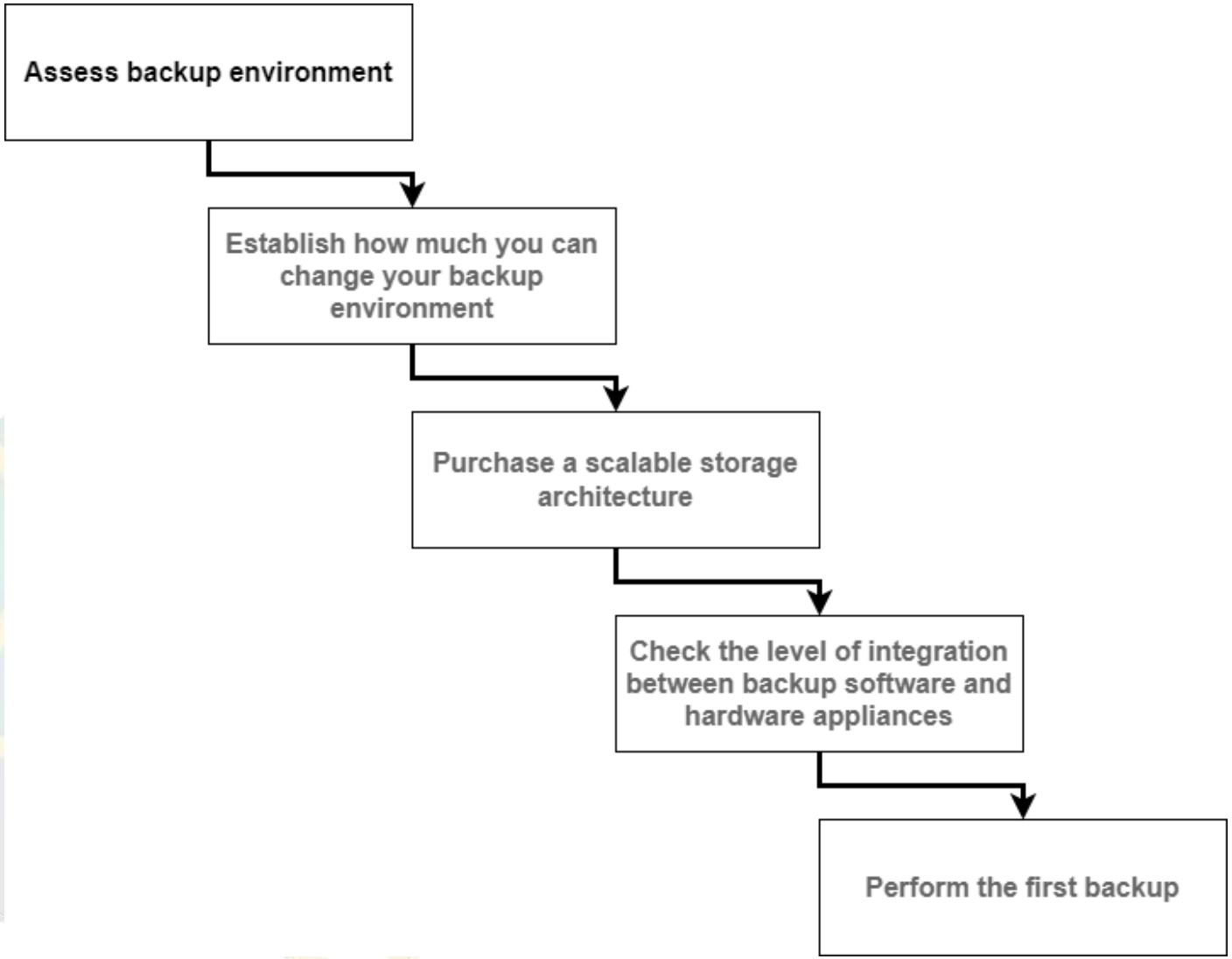
4. Only unique data are backed up on the server

Deduplication Ratio

$$\text{Deduplication Ratio} = \frac{\text{Total data before reduction}}{\text{Total data after reduction}}$$

Factors affecting deduplication ratio	Description
Retention period	The longer the data retention period, the greater is the chance of identical data existence in the backup
Frequency of full backup	The more frequently the full backups are conducted, the greater is the advantage of deduplication
Change rate	The fewer the changes to the content between backups, the greater is the efficiency of deduplication
Data type	The more unique the data, the less intrinsic duplication exists.
Deduplication method	Variable-length, sub-file deduplication discover the highest amount of deduplication across an organization

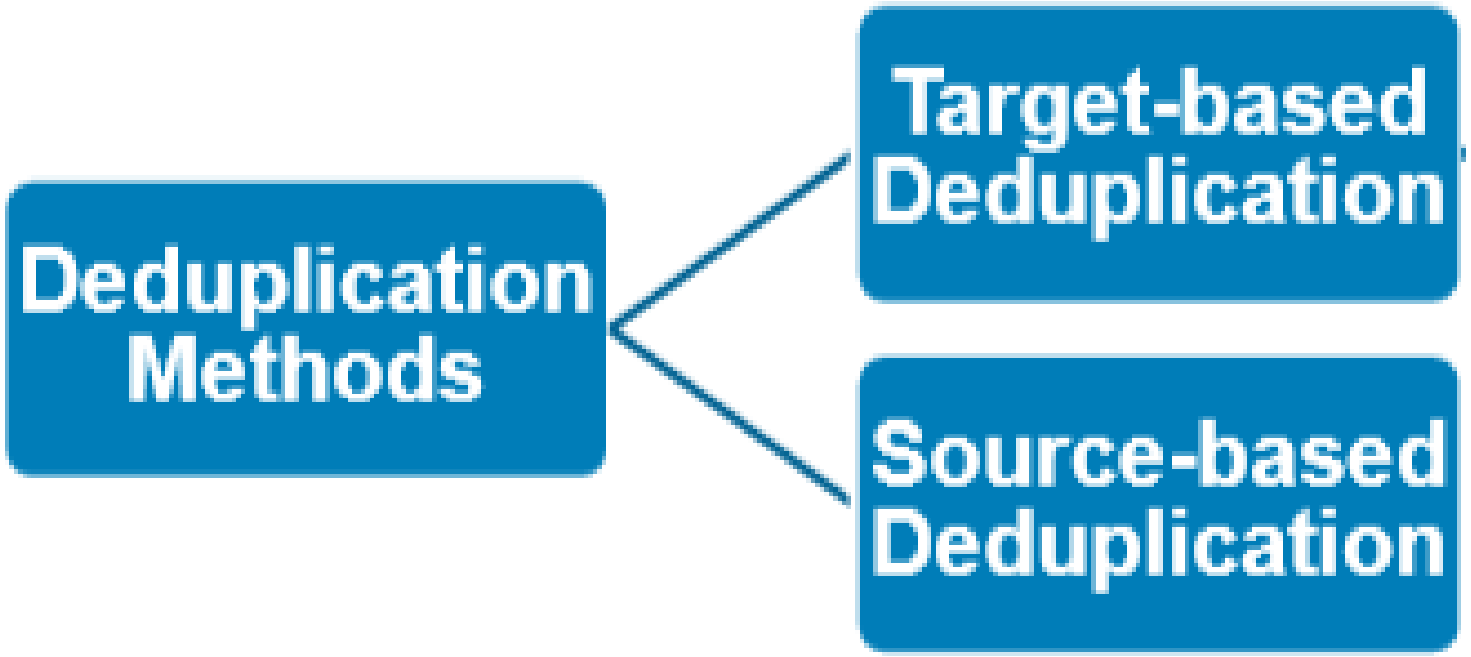
Guideline to implement data deduplication



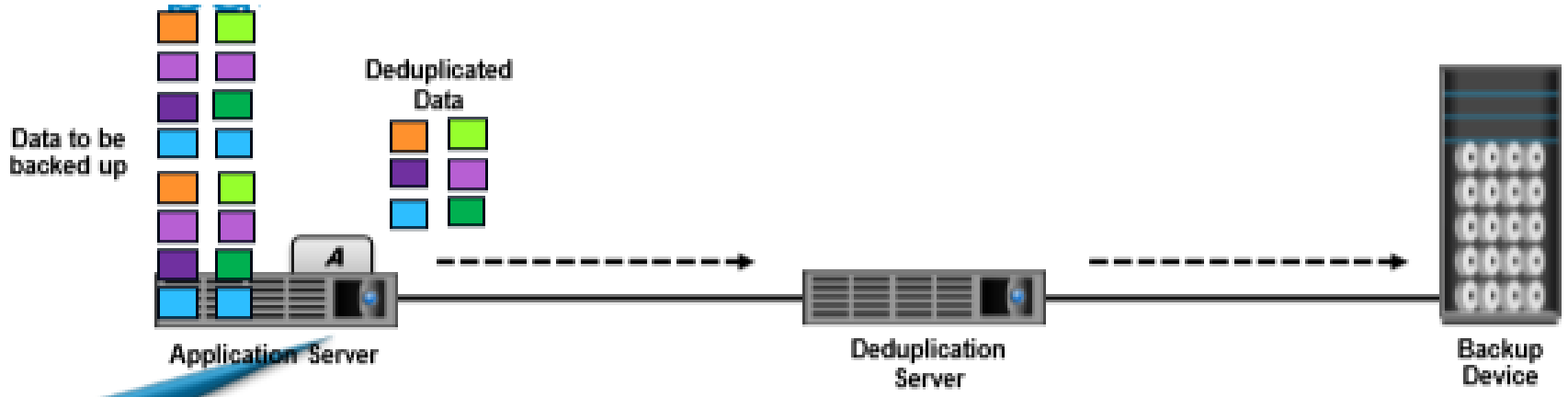
Deduplication Benefits

Benefits	Description
Reduces infrastructure costs	By eliminating redundant data, less space is required to store the backup data
Enables longer retention periods	Reduces the amount of redundant content in the daily backup, and hence, users can extend their retention policies
Reduces backup window	Less data to be backed up, which reduces backup window
Reduces network bandwidth requirement	Eliminating the redundant data reduces the amount of data to be sent over the network

Deduplication Methods



Source-based Deduplication

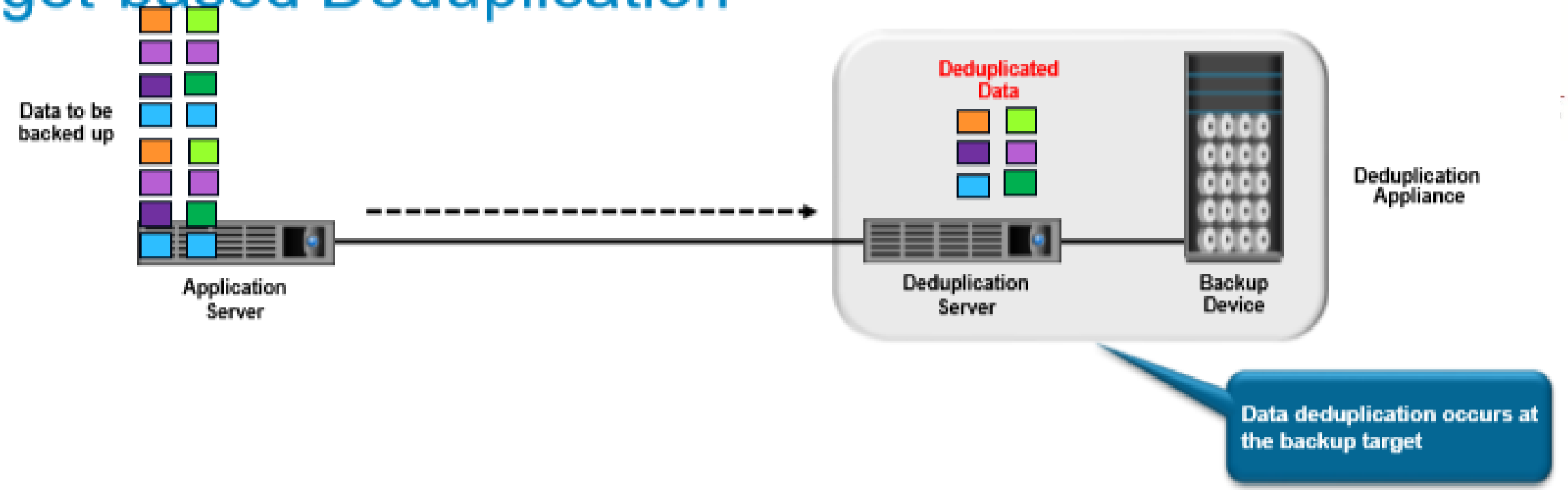


A
Deduplication Agent

Source-based Deduplication

- Data is deduplicated at the source (backup client)
- Backup client sends only new, unique segments across the network
- Suitable for environment where storage and network is a constraint
- Requires a change in the backup software if this option is not supported by the existing backup software
- Consumes CPU cycles on client and may impact the application performance
- Recommended for remote office branch office environment for performing centralized backup

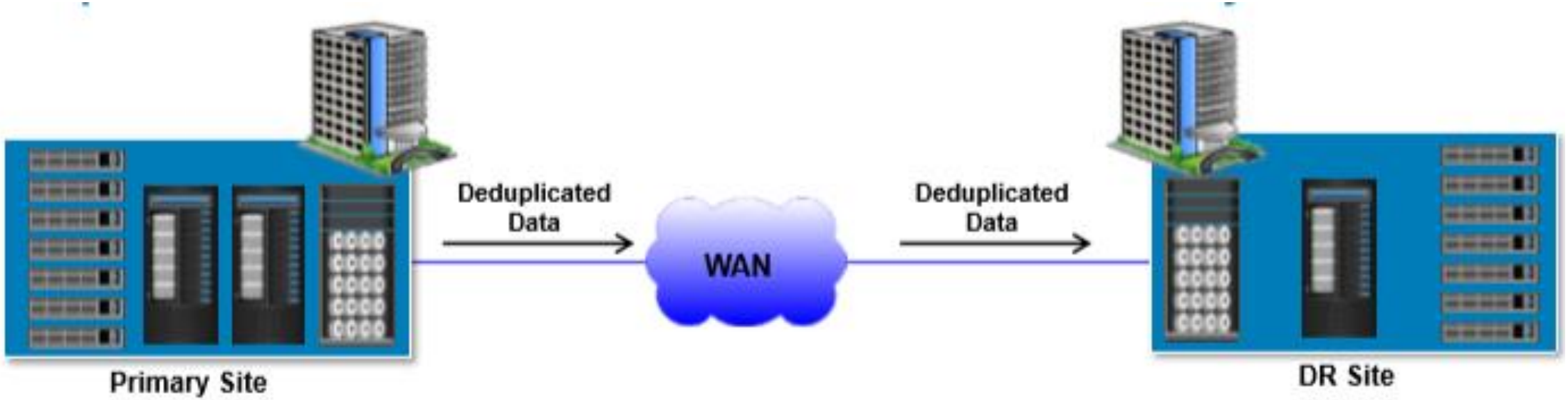
Target-based Deduplication



Target-based Deduplication

- Data is deduplicated at the target
- Supports current backup environment and no operational changes are required
- Client is not affected since deduplication process takes place at target
- Requires sufficient network bandwidth to send data across LAN or WAN during the backup
- Data is deduplicated at the backup device, either inline or post-process

Deduplication Use Case: Disaster Recovery



- Deduplication significantly reduces the network bandwidth to transfer the data from the primary site to the remote site (DR site or Cloud) for DR purpose
- Deduplication also reduces the storage requirement at the remote site

EMC Avamar and EMC Data Domain

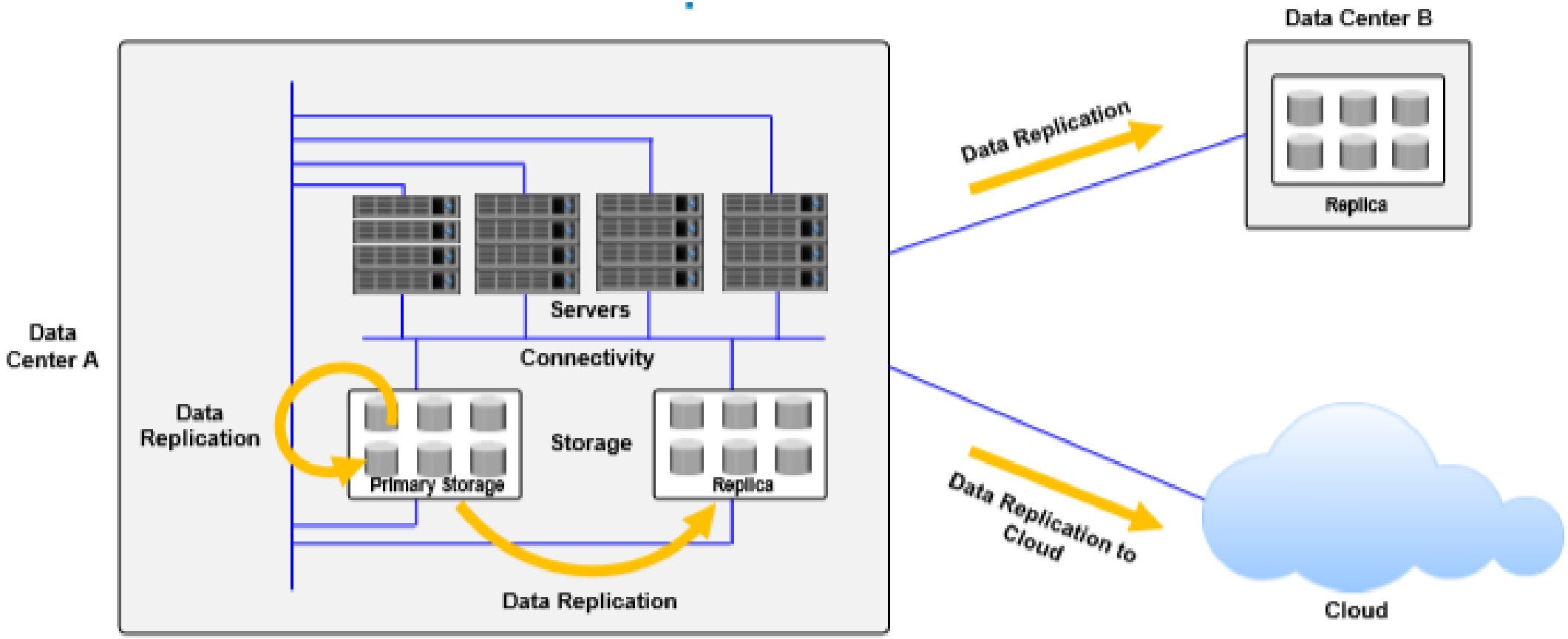
Avamar	Data Domain
• Disk-based backup and recovery solution that provides inherent source-based deduplication • Avamar provides a variety of options for backup, including guest OS-level backup and image-level backup • Data is encrypted and deduplicated to secure and minimize the network bandwidth consumption	• Data Domain is a target-based data deduplication solution • Data Domain Boost software increases the backup performance by distributing parts of deduplication process to the backup server • Provides secure multi-tenancy • Supports backup and archive in a single system

Quick Review

- What are the disadvantages of data deduplication?
- Explain the differences between target-based deduplication and source-based deduplication

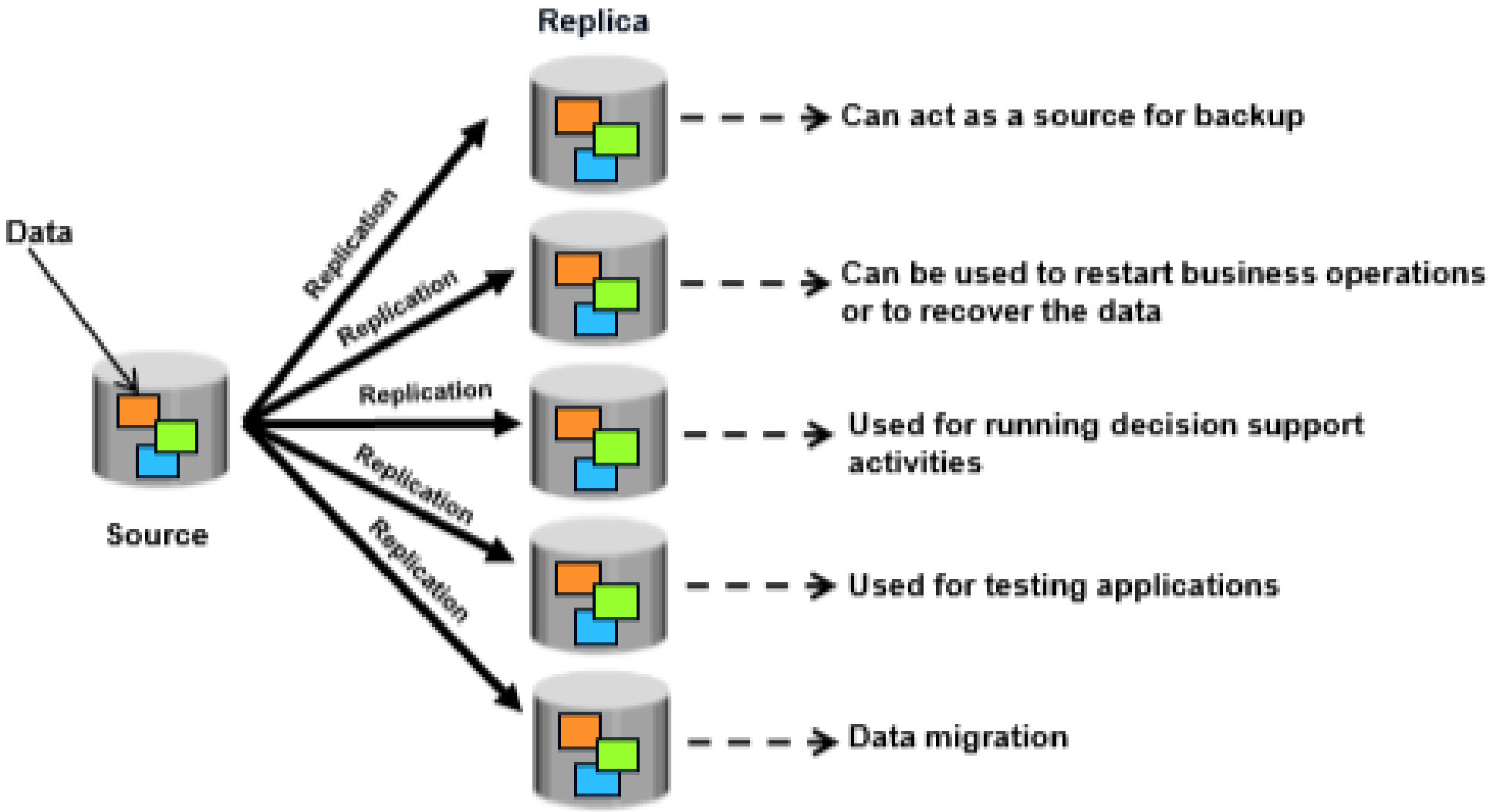
Replication

Introduction to Data Replication



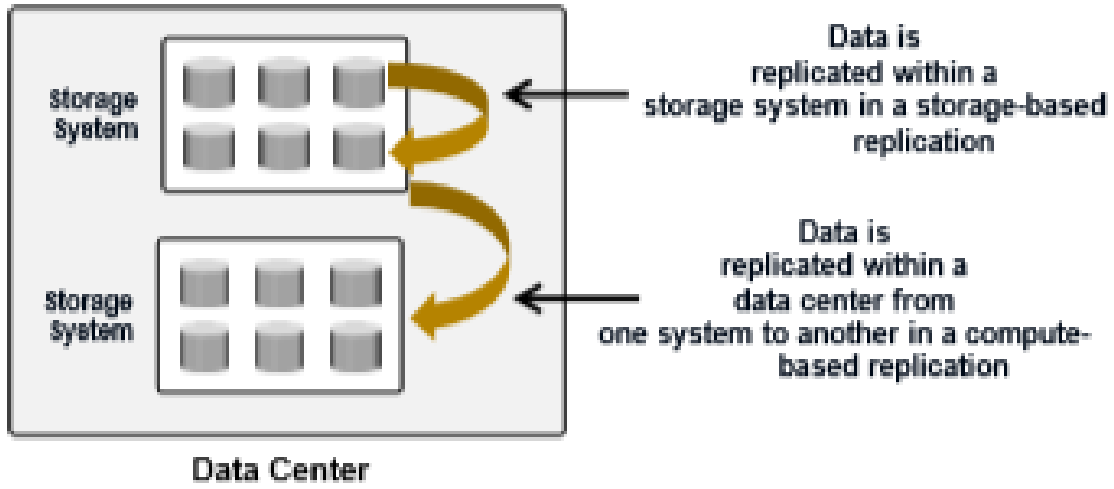
- Process of creating an exact copy (replica) of the data to ensure business continuity in the event of a local outage or disaster
- Replicas are used to restore and restart operations if data loss occurs

Primary Uses of Replicas



Types of Replication

Local Replication



Local Replication

- Refers to replicating data within the same location
  - Within a data center in compute-based replication
  - Within a storage system in storage system-based replication
- Typically used for operational restore of data in the event of data loss
- Can be implemented at compute, storage, and network

Remote Replication

Remote Replication

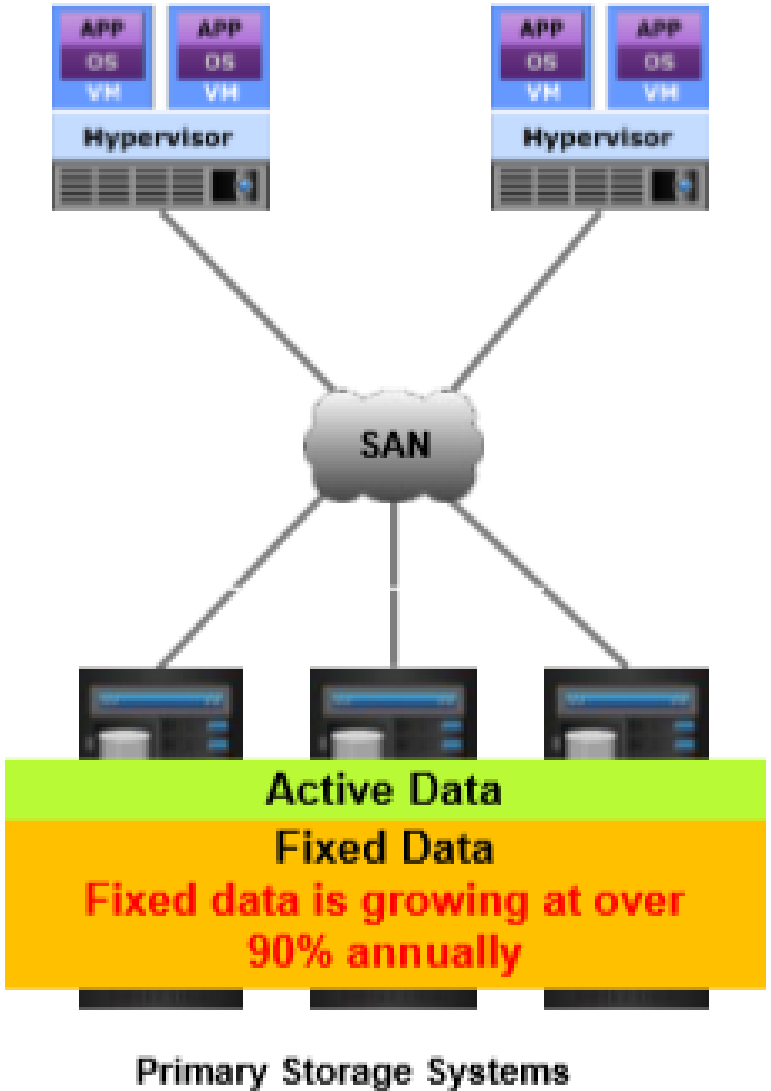
- Refers to replicating data to remote locations (locations can be geographically dispersed)
- Data can be synchronously or asynchronously replicated
- Helps to mitigate the risks associated with regional outages
- Enables organizations to replicate the data to cloud for DR purpose
- Can be implemented at compute, storage, and network

Quick Review

- What is the importance of recoverability and consistency in local replication?
- Describe the uses of a local replica in various business operations.
- What are the considerations for performing backup from a local replica?

Data Archiving

Why Do We Need Data Archiving?



What are the challenges of keeping fixed data in primary storage?

- Increasing consumption of expensive primary storage
- High performance storage for less frequently accessed data
- Risk of compliance breach
- Increased data backup window and cost

Data archiving addresses these challenges

Data Archiving and Its Benefits

Data archiving moves fixed data that is no longer actively accessed to a separate low cost archive storage system for long term retention and future reference

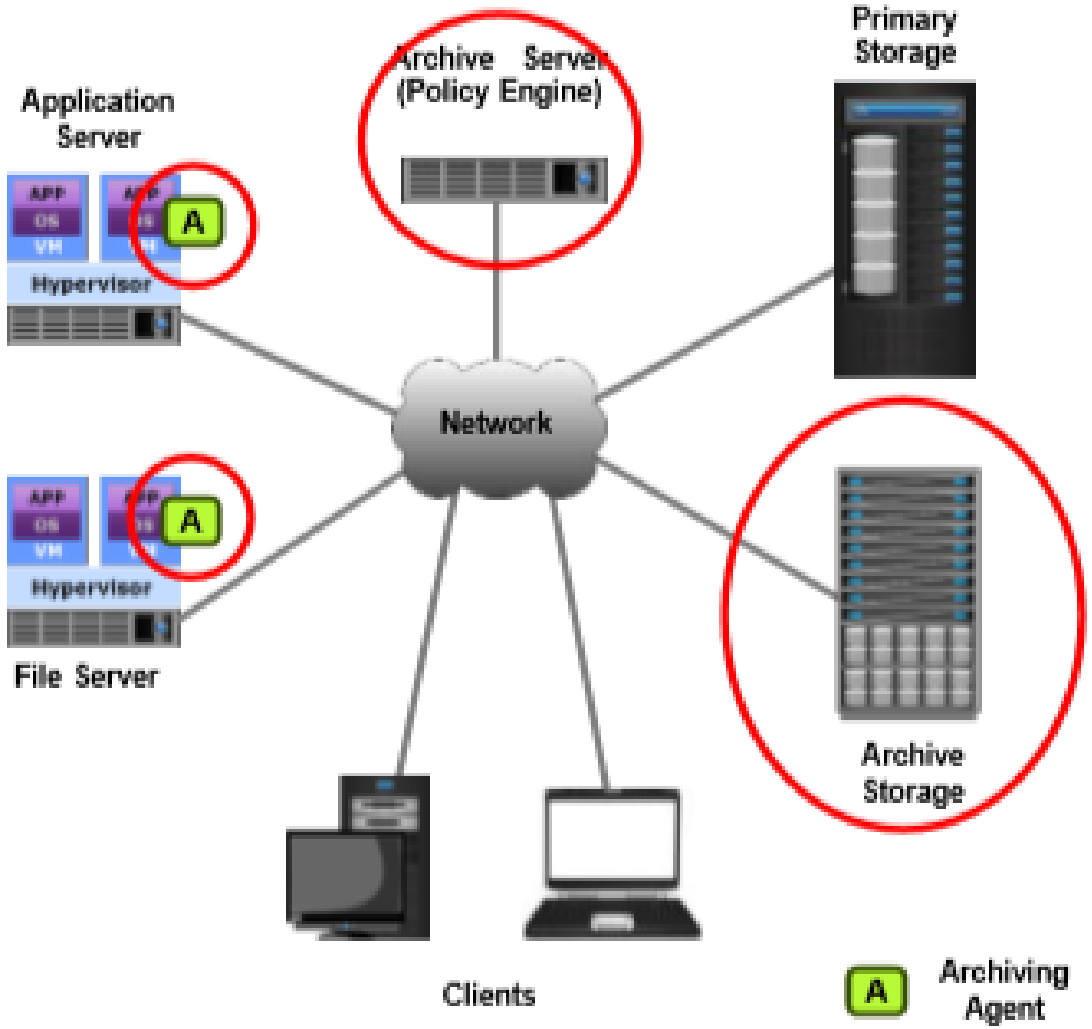
- Saves primary storage capacity
- Data archiving moves fixed data that is no longer actively accessed to a separate low cost archive storage system for long term retention and future reference
- Moves less frequently accessed data to lower cost archive storage
- Reduces backup window and backup storage cost
- Preserves data for future reference and adherence to regulatory compliance

Backup vs. Archiving

Backup vs. Archiving

Data Backup	Data Archiving
Secondary copy of data	Primary copy of data
Used for data recovery operations	Available for data retrieval
Primary objective – operational recovery and disaster recovery	Primary objective – compliance adherence and lower cost
Typically short-term (weeks or months) retention	Long-term (months, years, or decades) retention

Archiving Architecture



Archiving agent, installed on the application and file servers, scans files and archives them based on archiving policy

Archive server:

- Enables administrators to configure the policies for archiving data
- Maintains an index of archived files for search and retrieval operations

Archive storage stores fixed data

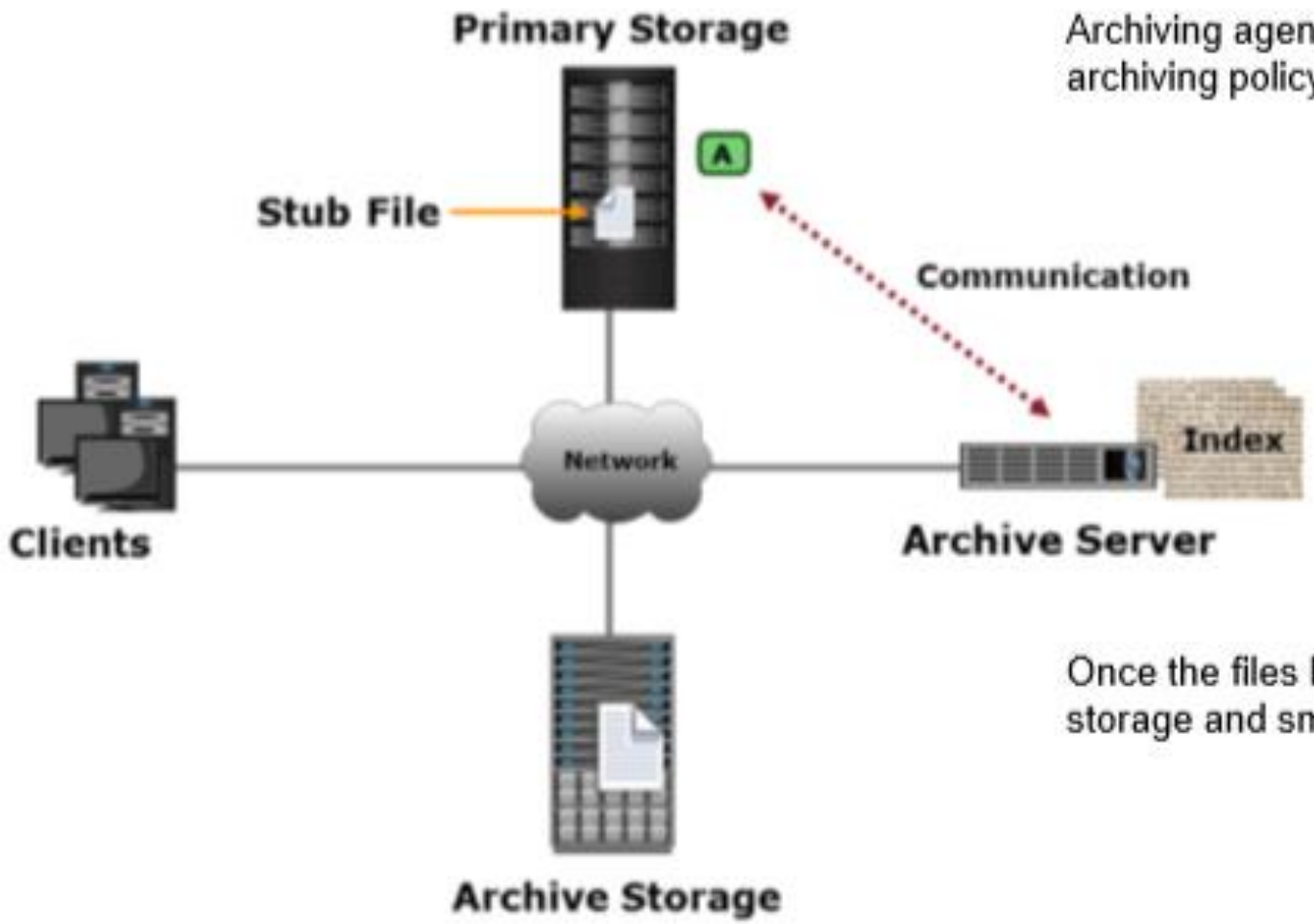
Examples of Data Archiving Regulations

SEC Rule 17a-4
Rule for data retention, indexing, and accessibility for companies which deal in the trade or brokering of financial securities.

Sarbanes-Oxley Act
Rule that protects investors by improving the accuracy and reliability of corporate disclosures. The rule applies to all public companies and accounting firms.

Health Insurance Portability and Accountability Act (HIPAA)
Rule that establishes national standards for health care industry. It provides guidelines for protection and retention of patient records.

Data Archiving Operation

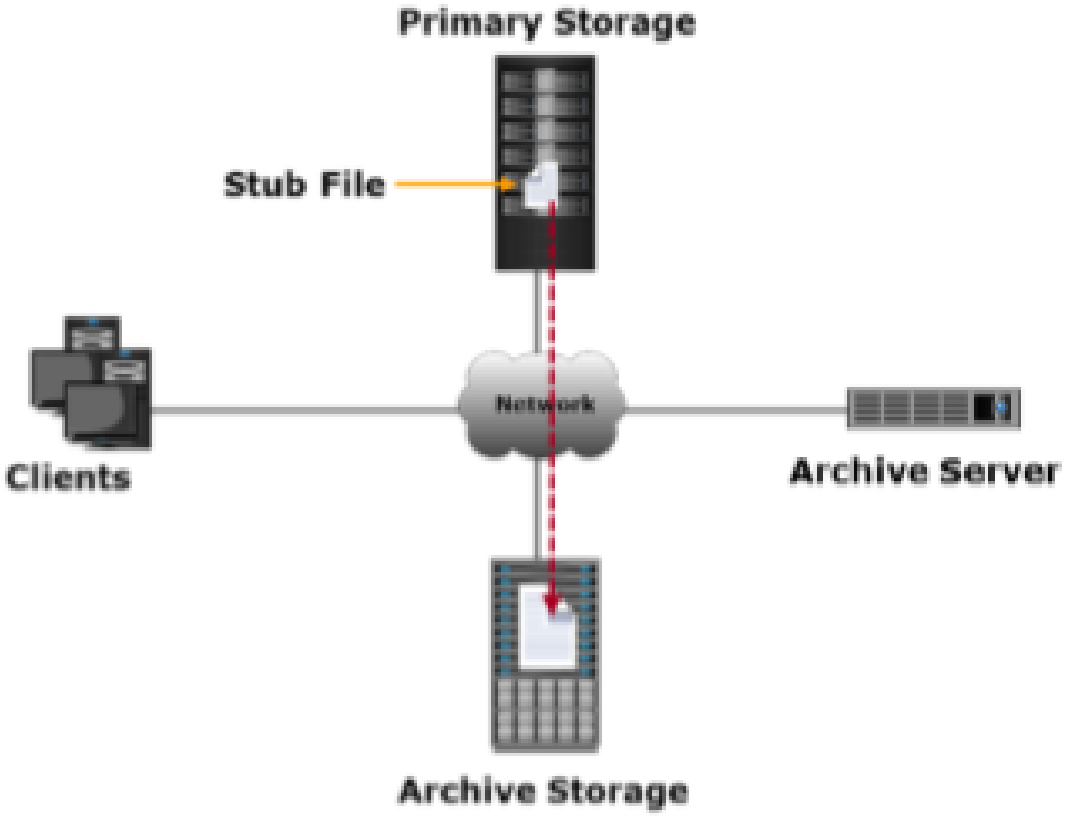


Archiving agent scans primary storage to find files that meet the archiving policy. The archive server indexes the files.

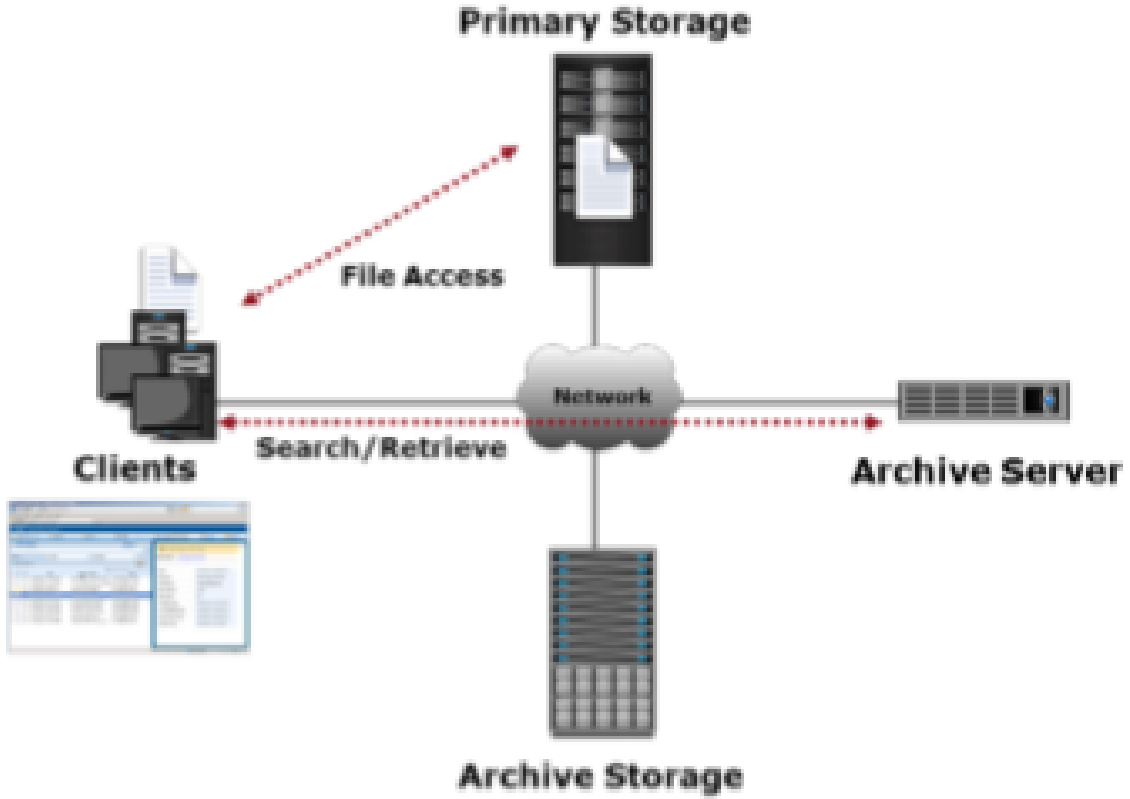
Once the files have been indexed, they are moved to archive storage and small stub files are left on the primary storage.

Data Retrieval Operation

When a client attempts to access a file, the stub file is used to retrieve the file from archive storage.



By utilizing the index for archived files, users may also search and retrieve files.



Why Data Migration?

Data center maintenance without downtime

Disaster avoidance

Technology refresh

Data center migration or consolidation

Workload balancing across data centers

Data Migration Techniques

SAN-based Migration

- Storage system to storage system direct data migration
- Storage system to storage system data migration through intermediary virtualization appliance

NAS-based Migration

- NAS to NAS direct data migration
- NAS to NAS data migration through intermediary compute system
- NAS to NAS data migration using virtualization appliance

Host-based Migration

- Host-based migration tool
- Hypervisor-based migration
  - VM live migration
  - VM storage migration

Application Migration

- Migration of application from one environment to another

Q & A