Birla Institute of Technology & Science, Pilani Work Integrated Learning Program 2024-2025

Submitted by:
Name: Mukund Vishwas Chavan
Course: Data Storage Technology and Networks
Student ID: 01011

PETA-SCALE DISTRIBUTED UNIFIED STORAGE SOLUTION DESIGN

TABLE OF CONTENTS

Chapter 1: Executive Summary and Requirements Analysis

1.1 Project Overview

This project involves developing a Peta-scale Distributed Unified Storage System designed to handle satellite imagery, derivative geographic data, and processed metadata. The main goal is to provide continuous 24/7 worldwide accessibility with uniform system efficiency while ensuring data integrity and coherence across every access point.

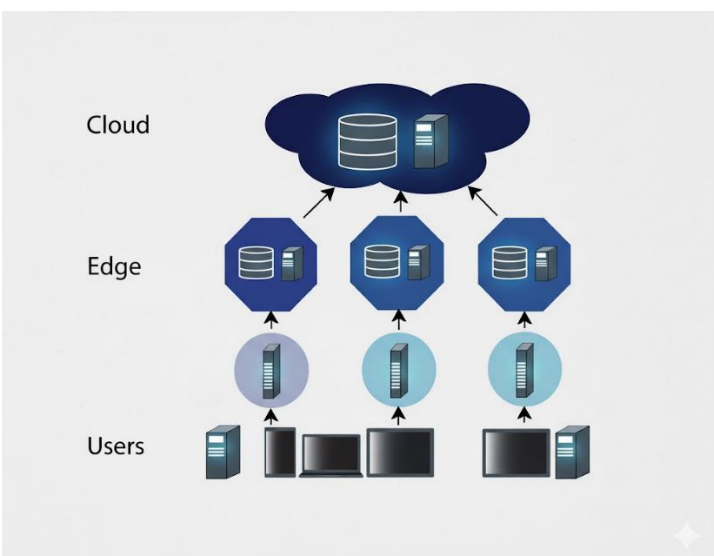
1.2 Key Architectural Drivers

- **Scale:** System capacity should reach petabytes, with potential scalability up to the exabyte range.
- **Availability:** Must provide uninterrupted 24/7 global accessibility and disaster recovery capability up to the last recorded checkpoint.
- **Performance:** Maintain consistent low-latency performance across global regions through WAN optimization.
- **Consistency:** Data must remain coherent and free of duplication across all distributed storage nodes.
- **Budget:** The infrastructure must support cost-efficient deployment using tiered storage and resource abstraction mechanisms.

Chapter 2: System Architecture and Component Design

2.1 Core Architectural Model: Geo-Distributed Software-Defined Storage (SDS)

The proposed framework utilizes a globally distributed hybrid storage architecture controlled by a Software-Defined Storage layer. This approach separates the control plane (for orchestration and management) from the data plane (handling physical storage), allowing independent scalability of each layer.



2.2 Component Breakdown and Functionality

2.2.1 Data Storage Tiers

Component	Data Type	Physical Storage	Role and Function
Tier 1: Hot Metadata	Indexes, landmark updates, current processing tasks	NVMe / High-Speed SSDs (Local to Data Centres)	Provides low-latency and high-IOPs operations with strong transactional reliability (ACID).
Tier 2: Warm Object	Raw Images (last 90 days), active map grids, processed artifact lists	Mid-range SSD/HDD hybrid (SAS/SATA)	Offers a balanced ratio of capacity and throughput, forming the working dataset for analytics.
Tier 3: Cold Archive	Historical images, complete daily backups	High-density HDDs / Tape gateways	Lowest cost per GB, designed for long-term archival, bulk sequential reads, and durability.

2.2.2 Service Layers

- **Global Load Balancer (DNS/Anycast):** Routes client requests to the nearest active data center, ensuring consistent global response times.
- **Storage Access Gateway:** Handles protocol conversion (e.g., S3 ↔ DFS), initial authentication, and directs the request to the appropriate storage tier.
- **Global Metadata Service (GMS):** A fault-tolerant cluster (based on Raft/Paxos) that maintains metadata, replication mapping, and de-duplication indexes—central to ensuring data coherence.

2.3 Access Protocols and Justification

Data Type	Access Protocol	Justification
Raw Satellite Images (WORM)	RESTful API (S3-compatible) over HTTP/S	Ideal for large-scale object storage, providing high throughput and seamless integration with cloud and CDN systems.
Processed Indexes / Metadata	POSIX / NFSv4 via DFS Layer	Required for applications that demand file-level consistency, locking, and frequent small updates.
Data Processing Applications	FUSE / Client Library	Enables low-overhead, high-performance direct access to distributed data storage fabrics.

Chapter 3: Storage Planning and Provisioning

3.1 Capacity Planning and Sizing

The solution targets peta-scale storage, ensuring redundancy and disaster recovery across three global data centers (DCs).

- **Local Replication Factor:** To protect against node or disk failures within a DC.
- **Geo-Replication Factor:** Maintains complete data copies across all three DCs.
- **Total Storage Requirement:** Accounts for both active replication and backup allocations.

3.2 Backup and Disaster Recovery Allocation

Daily full backups are maintained for seven days using content-aware de-duplication to reduce redundancy. This ensures compliance with the “previous day recovery” requirement while optimizing storage efficiency.

3.3 Budget Optimization (Tiering and Thin Provisioning)

- **Tier 1 (SSD/NVMe):** Allocated for critical metadata (~10% of total data).
- **Tier 3 (HDD):** Stores long-term historical data (~60% of capacity).
- **Thin Provisioning:** Initially, only partial storage is deployed, expanding as usage grows to minimize upfront costs (CAPEX).

Chapter 4: Data Management — Consistency, Replication, and De-duplication

4.1 Consistency Model

The system ensures global data reliability through a combination of strong local consistency and global quorum consensus.

- **Local Consistency:** Managed by Raft/Paxos-based distributed algorithms; a write is committed only after a majority of local nodes acknowledge it.
- **Global Quorum (W+R > N):** Ensures overlap between read and write operations across three replicas, maintaining “read-your-writes” integrity worldwide.

4.2 Replication Scheme

Local Replication: Synchronous triple replication within each DC for resilience.
Geo-Replication: Asynchronous updates between global DCs ensure near-real-time synchronization and disaster recovery capability.

Cloning: Maintains identical datasets across all sites to enhance performance and eliminate latency from intercontinental data transfers.

4.3 De-duplication Scheme

A **content-aware, variable-block post-process de-duplication** system is implemented:

- **Primary Enforcement Point:** During initial data ingestion, where only unique data blocks are stored.
- **Backup Enforcement Point:** Applied to historical backups, significantly reducing backup size and improving cost efficiency.

Chapter 5: Storage Virtualization and Abstraction

5.1 Scope

Storage virtualization plays a crucial role in simplifying the management of geo-distributed and heterogeneous storage hardware (such as NVMe, SSD, and HDD). It masks underlying hardware complexities and presents a single, cohesive interface to users and applications.

5.2 Global Namespace

Applications view a uniform directory (e.g., /satellite/regionX/imageY.tiff) irrespective of physical location. The SDS software dynamically maps to the closest consistent data replica.

5.3 Automated Tiering and Provisioning

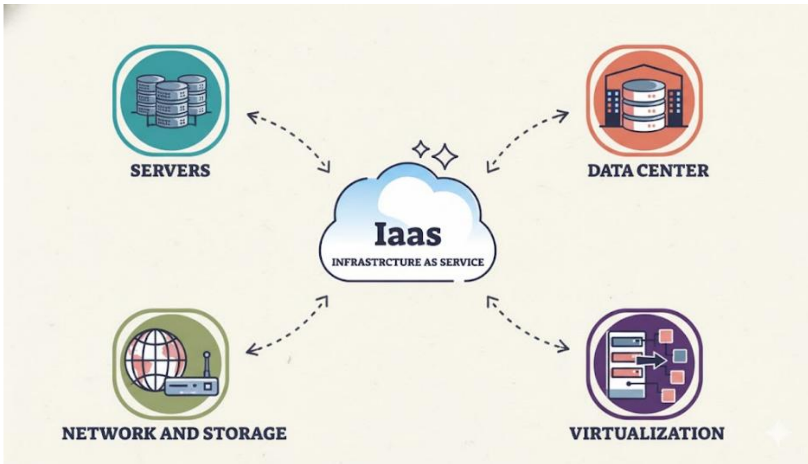
The system automatically relocates frequently accessed (“hot”) data to faster tiers and archives infrequently used (“cold”) data to cost-efficient media, maintaining optimal performance and cost balance.

Automated Tiering: The virtualization layer continuously analyzes access frequency and dynamically relocates data. Frequently accessed (“hot”) metadata is migrated to high-speed Tier 1 (NVMe), while infrequently accessed (“cold”) data is shifted to Tier 3 (HDD). This mechanism maintains performance for active datasets and optimizes cost for archival data, ensuring balanced resource utilization.

Thin Provisioning: In this approach, physical storage space is allocated only when data is truly written. This minimizes over-provisioning and capital expenditure by allowing storage capacity to scale gradually based on actual usage requirements.

5.4 Disaster Recovery and Failover

In case of data center failure, the system auto-remaps namespace references to the nearest active site, enabling uninterrupted access and transparent recovery.



Chapter 6: Conclusion and Future Scalability

6.1 Summary of Solution Benefits

Requirement	Solution Benefit	Key Technology Used
Peta-scale Capacity	Cost-efficient large-scale storage through tiered commodity hardware	SDS pooling, high-density HDDs
24/7 Global Access	Low-latency, uninterrupted access from nearest replica	Geo-distributed active-active setup
Consistency	Guaranteed coherence and data reliability globally	Quorum-based consistency (W+R>N), Raft/Paxos
Disaster Recovery	Real-time recovery far beyond daily checkpoint	Asynchronous multi-site replication, auto failover

6.2 Future Scalability Roadmap

- **Horizontal Scaling:** Add new nodes or data centers seamlessly as capacity or performance demands grow.
- **Adoption of New Media:** Future-proofing for advanced storage technologies like QLC NAND or next-gen NVMe.
- **Erasure Coding Transition:** Transitioning archival tiers to erasure-coded formats (e.g., 8+4, 10+2) for efficient exabyte-scale expansion with minimal redundancy overhead.