Deep Science Publishing
https://doi.org/10.70593/978-93-49910-08-9_7

DeepScience
Open Access Books

Chapter 7: Strategies for high availability, disaster recovery, and performance in multi-cloud deployments

7.1. Introduction

Multi-cloud has gained a strong foothold in the IT landscape. However, interest in multi-cloud is often driven by a combination of frustration with a "single cloud" strategy — whether real or perceived — and excitement about the possibilities of a multi-cloud strategy. But what is a multi-cloud strategy, what are the primary use cases, and how do organizations go about implementing it to reap the benefits while mitigating the associated risks?

A multi-cloud strategy is an approach employed to leverage the services and capabilities of more than one public cloud provider. While many organizations depend upon a single cloud provider for capacity, services, and capabilities to meet their deployment needs, multi-cloud is often equally appealing and interesting in that it can relieve single-cloud provider dependency and risk, as well as utilizing cloud vendors whose services may be better matched for a given business function or application. So-called "cloudbursts" — defined bursts of activity or processing associated with particular applications — are also suited to multi-cloud strategies in that brief additional capacity requirements can be met far less expensively on a temporary basis using public cloud than by setting up and deploying an enterprise data center. Multi-cloud does require some level of competence and expertise in managing workload movement across providers. As with most situations entailing competing priorities, using multiple clouds generates both opportunity and management complexity (Doe et al., 2025; Johnson et al., 2025; Lee, 2025).

Despite the increasing interest in multi-cloud strategies and the rapid growth in multi-cloud deployments, the public multi-cloud ecosystem lacks the level of interoperability that might be expected, much less desired, at this stage in its evolution. As a result,

organizations that embark upon a multi-cloud strategy require both a clear-eyed understanding of the opportunities associated with using multiple clouds as well as the challenges and complexities. In this chapter, we explore the opportunities, challenges, and complexities associated with multi-cloud strategies in depth (Smith, 2025; Thompson et al., 2025).

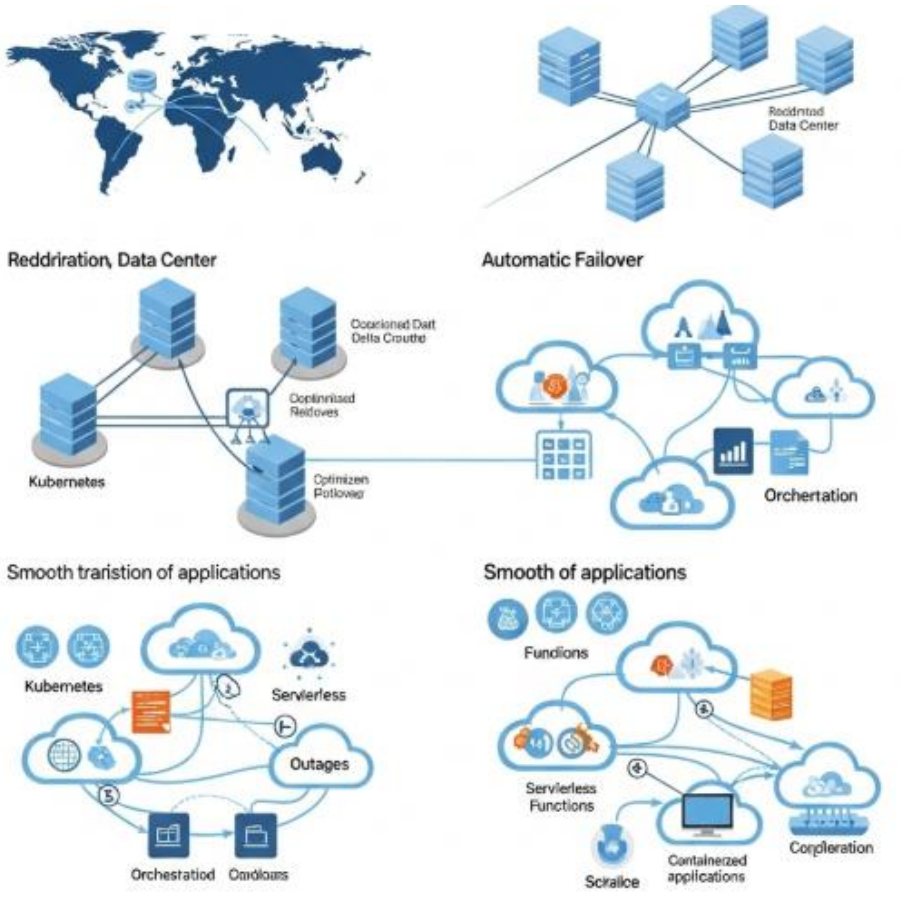

Fig 7.1: Strategies for High Availability, Disaster Recovery, and Performance in Multi-Cloud Deployments

7.1.1. Background and Significance

Despite billions of dollars spent on disaster recovery systems, IT infrastructures still go down. This situation can be exacerbated by increasingly severe and complex disasters, whether they be technical or natural in scale and dimension. These outages continue to occur despite the mandates from corporate boards, auditing companies, insurance companies, and government entities to ensure business continuity with increasingly stringent guidelines. Often these outages happen during times of increasing dependence

on technology, whether during times of increased natural disasters, pandemics, or cyber threats from hostile entities. It is no longer a question of if these disasters will be wished upon companies, but rather a question of how often and at what scale.

For companies that depend on their technologies to conduct business 24 hours a day, 7 days a week, 365 days a year, the losses associated with downtime can be staggering. Estimates state that the cost of downtime can sit in the tens of thousands of dollars per minute, which adds up quickly. In addition to the financial impact, the fallout from an outage may impact customer loyalty and public perception, as users vote with their wallets and decide to take their business elsewhere. The IT staff dedicated to managing uptime, availability, and recovery become burdened and increasingly strained as they struggle to meet the demands of an always-on world. As cloud service offerings from multiple vendors mature and coalesce into valid and innovative usage cases, these companies are also faced with the increasingly difficult task of how best to architect systems that fulfill the requirement of availability using these new services.

7.2. Understanding Multi-Cloud Environments

High availability (HA), disaster recovery (DR), and data protection are essential components of every production cloud service in order to maintain business continuity and availability of mission-critical workloads. Although traditional on-premises architectures are still somewhat common, the vast majority of organizations are adopting hybrid, and increasingly, multi-cloud architecture to gain the agility, flexibility, scalability, workload optimization, and cost efficiency that these cloud environments offer. Avoiding vendor lock-in is one of the most common reasons organizations deploy multi-cloud models with services from multiple cloud providers.

In traditional on-premises HA/DR architectures, securing replicated copies of data and workloads is relatively easy; by deploying a second set of systems in a second location, and replicating data and workloads at regular intervals, either synchronously or asynchronously, either manual or automatic failover can be accomplished in the event that the production systems go down. As organizations migrate more of their workloads to public or hybrid cloud providers, these traditional models are challenged, especially for multi-cloud deployments, where enterprise workloads spread across cloud providers do not have high-speed private connections like data centers or hybrid clouds would. In addition, cloud providers do not typically allow accessing the underlying infrastructure, instead providing a multi-tenant public infrastructure where other tenants can have varying impacts on the performance available to enterprise workloads. These challenges place additional complexity on implementing DR orchestration, failover timelines, and testing, leading to increasingly complicated DR as a Service solutions, especially when considering the varying data protection offerings provided by each cloud provider.

7.2.1. Definition and Characteristics

Today, several enterprise applications, including for high availability and disaster recovery, involve multiple clouds from different vendors. Therefore, how do we know that we are deploying a multi-cloud? There are, chillingly, different definitions for multi-cloud or multi-cloud environments.

Multi-cloud can be defined as multiple public clouds from different vendors. Different from a hybrid cloud, which is the conjunction of a private cloud with a public cloud, a multi-cloud is composed only of public clouds. An enterprise adopting a hybrid cloud is free for on-premises deployment of the horizontal layer of the ITIL service delivery of any service or content. A majority of prime enterprise services must be stored in-house in a proposed hybrid-cloud strategy. Although the multi-cloud architecture has only clouds from commercial cloud players, it is, like hybrid clouds, an imperfect cloud capable of violating the basic principles of the cloud model because the enterprise will need to access different cloud vendor portals, besides the fact that resource pooling is primarily local.

At the other extreme, enterprises are willing to deploy only cloud-native applications as microservices for avoiding environments with different clouds. In the industry, people say that most multi-cloud environments are interim environments or they simply do not exist. The reality is that the hybrid cloud model is the simplest, most reliable, and cost-effective model. But the world of cloud computing is complex, and many organizations decide to deploy their resources in different clouds.

7.2.2. Benefits of Multi-Cloud Strategies

Multi-cloud strategies can help to fulfill specific business needs, offer technical advantages to an organization, and fulfill the requirements to comply with policies and regulations. Today's customers require better quality of service and availability from their services and applications. Multi-cloud strategies enable organizations to seamlessly deploy their IT services to the environments that offer the best qualities, whether cloud or on-prem, based on their unique business requirements. These could include faster, better, and more affordable service, or service availability and durability emphasized by business impact analyses. Such strategies not only yield better service for customers but also help organizations avoid putting all their eggs in one basket. Multi-cloud strategies govern which services are deployed to which clouds, according to the associated costs, reliability, policies, security, and performance SLAs.

This would expose organizations to a wider range of service providers that offer different terms and conditions. Organizations can also adopt a wider range of service and application capabilities by using multiple cloud environments together. Organizations

that govern the use of multiple cloud deployments have the ability to seamlessly deploy between multiple clouds based on their availability and performance requirements, service consumption costs, impact on application performance due to other collocated services, compliance to policies, and the completion of tasks without failure or impact of persistence. Multi-cloud strategies even facilitate organizations to disrupt the network and service management latency due to potential bottlenecks of latency-sensitive services and applications. Business impact analysis could reveal the need to reduce the potential impact of a bottleneck. Internal network latency incurred between storage, applications, and computers may no longer be permissible to continue co-locating unless the potential impact of a bottleneck is prioritized with fewer transactions.

7.3. High Availability in Multi-Cloud

Unplanned or emergency downtime is the bane of every SaaS 운영자. Whether initiated by a cyber rampage or by a system misconfiguration, downtime impacts your business continuity and customers. In a previous chapter, we examined multi-cloud deployments for fault tolerance, in that more than one cloud vendor is used for a service, so that the failure of a single provider does not impact the service as a whole. However, even without external impacts, outages can come from bad software deployment, bugs in software stack components, or indications of hardware components finally giving up the ghost. Thus, for mission-critical applications, it is not just about fault tolerance − it is about high availability (HA).

High availability is generally designed as a requirement for most contemporary multi-cloud deployments. While tying in closely with fault tolerance, HA is at the same time designed by default in a multi-cloud deployment. Services are made redundant: If one component goes down, there are still other instances available to process requests. Requests are then directed to the instances that are still running. In this chapter, we delve into some of the decisions that you would take to increase the HA in your systems in the cloud. We will explore some strategies in data replication of mission-critical data, before delving into load balancing techniques. We will wind down this chapter with a discussion of some HA focused monitoring tools available in industry today.

7.3.1. Designing for Redundancy

Redundancy is the backbone of high availability. For a service to continue functioning reliably, individual subsystems must experience failure-free operation over the collective lifetime of the service. This is accomplished by designing redundant subsystems,

allowing for service continuity during failures when fault detection, recovery, and restoration processes are implemented.

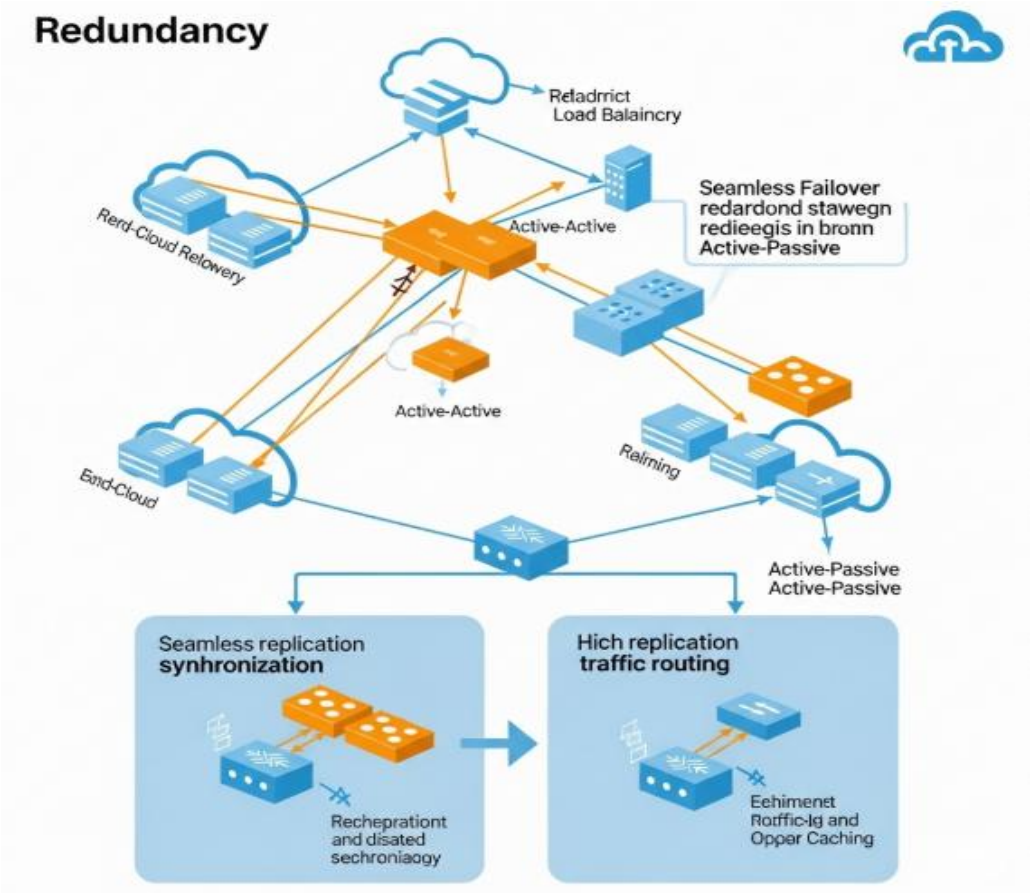

Fig 7.2: Designing for Redundancy of Strategies

A redundant subsystem is a spare pair of systems that are capable of providing the critical function of the active counterpart during failure. Implemented consistently across all critical high service points of a service, redundancy can be either real or logical. Real redundancy occurs when there are multiple copies of each critical architecture component in the service—that is, the database resources, application servers, and front-end resources. Logical redundancy, on the other hand, occurs when a single infrastructure unit that is critical to the operation of the service is shared across some or the entire set of multiple replicas of the service infrastructure, as in a net of caches, a farm of CDNs, or a set of clustered front-end location and application servers. The replica instances take turns being active and available or passive, waiting for the event of failure.

Both types of redundancy have their advantages and disadvantages. Real redundancy can radically reduce the service's mean time to repair, or MTTR, drastically increasing the service's level of availability. It requires, however, greater additional investment in resources. Logical redundancy may sustain larger MTTRs because of the added exchange of business in the meantime. Most services would employ a combination of the two types of redundancy, according to the cost model of the business.

7.3.2. Load Balancing Techniques

When requests are received by a load balancer, it needs to determine how to distribute them among the nodes behind it. The most straightforward method is to simply choose the next available server in a round-robin fashion. This approach works well assuming that each request will be of roughly the same size and resource requirements. However, this is rarely true in practice. To improve response times, applications often use caching to cache the results of potentially expensive operations. If the requests hitting an application are for cache misses, then those requests will take the most resource time to respond to.

Because of this deviation in request size and resource requirements, load balancers often have to try and guess which is the least busy back-end server based on performance data such as CPU load and existing connections. These heuristics can provide a reasonable estimate, but they are also imperfect. If the least busy server had a recent spike of load, even if it appears to be ready, it might have a cache miss that indicates an impending delay. Every different application has different requirements, so load balancers have to balance performance against complexity in determining the optimal resource to route the request through.

There are periods of load-aware performance where the load balancer just routes to the least busy. The problem with this approach is that with the imperfect measure of performance from the various backend nodes, they can’t always accurately predict resource requirements, especially if they are anomalously increasing. Because back-end nodes can give false performance indicators, if there’s high back-end node latency, the load balancer can even temporarily drop backend nodes out of service during high periods.

7.4. Disaster Recovery Strategies

In this section, we analyze some common strategies for disaster recovery in multi-cloud environments and how you can implement them. Organizations have been transitioning to multi-cloud environments in order to take advantage of different services across

different cloud providers. A multi-cloud deployment is exposed to new failure modes related to network connections; for example, the communication link connecting the different cloud environments may fail. However, multi-cloud deployment is designed against local failures; DR against cloud-wide failure so far is simplistic; customers implement home-grown performance-tuning scripts at software level.

A disaster recovery (DR) plan is a documented solution that ensures the recovery of a company from a disruptive incident and is designed to preserve its business continuity and reduce the negative effect of a disaster on the operations of the company. The two fundamental elements of every DR plan are the recovery point objective (RPO) and the recovery time objective (RTO). The RPO is the maximum amount of data loss measured in time, which can be tolerated by a company after a disaster. It indicates the need for data backup frequency. Generally, the smaller the RPO, the more expensive the DR solution. The RTO is the maximum permissible time required to recover from the point of failure to the operation actualization. Once these two objectives are established, you have to define which DR strategies and technologies will be used. A DR strategy describes how you plan to recover from a disaster-related disruption. The choice of the DR strategy you will follow should take into account factors such as RPO, RTO, availability of resources, and costs. It is also important to conduct DR plan tests and reviews periodically.

7.4.1. Types of Disaster Recovery Plans

Many organizations create one or more plans for restoring and managing critical infrastructures, applications, services, or data after a disruptive event. A business continuity plan goes beyond restoring critical IT functions; it also covers business processes in non-IT areas, such as customer service and sales, finance, human resources, legal and other back office functions.

A disaster recovery plan focuses on recovering only the disrupted IT functions and their supporting resources: servers, data, connections to other applications and supporting systems. An updated disaster recovery plan should be established before incidents occur. You should regularly test these recovery plans, because a recovery operation might have some unpredicted issues that can slow you down. Testing them regularly minimizes the chance of unexpected delays.

However, DR plans should not be thought of as static documents. They need to reflect ongoing changes due to staff turnover, changes in responsibilities and priorities, merging of services or applications, and technology upgrades and changes. So management must change the DR plans as these types of changes occur. Finally, your DR plans should always be easily accessible to the staffers who are involved in restoration operations.

There are two general methods for disaster recovery: cold and warm. A cold disaster recovery is less expensive than a warm one. If a data center is non-operational, the organization must make alternate arrangements to continue work: sending people to another location, using manual processes, or suspending work until operations resume. This method is more useful in industries that can function without an IT infrastructure. These alternative arrangements usually require the additional expense of maintaining a second facility.

7.4.2. Data Backup Solutions

With the rapid growth of data in both volume and value, organizations increasingly recognize the business need to protect these data from loss or corruption. Maintaining data availability – providing access to data when requested – is the primary reason for enterprises to develop robust data protection solutions. Some may argue that high availability clustering technologies are sufficient; however, these tools should be regarded primarily as failover techniques and should be regarded in the context of uptime requirements for enterprise applications. Data backup solutions are designed for data loss and protection from human errors, application errors, and viruses, as well as disasters. There are many possible disasters – environmental, natural, technical, or human – for which an organization must plan in advance to provide a backup for their data.

Most organizations running enterprise-level applications need more than just the ability to restore data from backups. Generally, organizations select a variety of backup strategies that can range from local storage of disk backups to both hard copy and cloud disk-based long-term backup solutions. Staging tiers can be used to set an appropriate balance between cost and the speed by which the backed-up data can be restored. Additionally, to logically separate the backup data with certain levels of protection and availability, organizations often store production and backup data in different tiers of storage and, in some cases, in different storage products. However, given the fact that backup data is often seen as simply others' data to be accessed from only one or more designated points in the organization and with restricted roles the original data is managed by a different product/solution, both from a logical and physical perspective, traditional enterprise-class data backup products are still utilized mostly as conventional data backup solutions.

7.5. Performance Optimization Techniques

Performance optimization means making a system work more efficiently. In this section, we will discuss techniques for optimizing performance in a multi-cloud environment.

Performance optimization in the multi-cloud context is unique due to the distributed nature of the resources, possible different technologies and service providers, security and compliance constraints among various clouds, cost concerns, and challenges in performance troubleshooting, testing, and automation. While somewhat simplistic, performance optimization for any application consists of four steps: profile the application, propose solutions for performance improvement, implement the solution, and repeat. In this section, we will discuss popular techniques for helping in each of these steps.

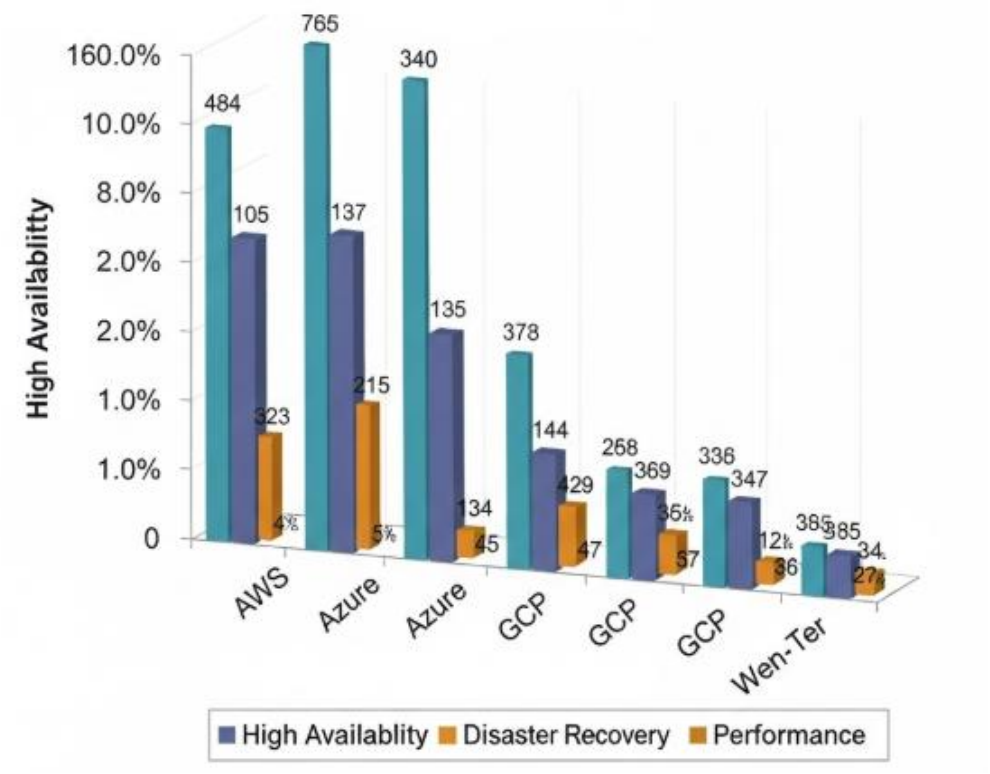

Fig : Graph of High Availability, Disaster Recovery, and Performance in Multi-Cloud Deployments

There are many metrics that help monitor the performance of a system. Some of the common metrics include response time, error rate, throughput, latency, availability, and resource utilization. However, due to the open-ended design of applications, it can be difficult to subject any application to performance testing. It is also challenging to identify relevant KPIs for any application, automated A/B testing for deployments, and continuous performance evaluation to avoid unexpected latency in expensive serverless services. Increasing the width of the performance tuning funnel requires a flow of probing capabilities. Although the trend in larger system designs may be toward a black-

box testing, as multi-cloud deployments grow, users want to understand the performance characteristics, not just overall performance.

Deploying an application across heterogeneous systems, including multiple public clouds and private clouds, poses additional challenges both at design and operations phases. During the design phase, users are faced with nitty-gritty decisions such as picking a technology stack for the data ingestion, processing, and storage that are well integrated, choosing the microservice boundaries, heterogeneous services provided by the various clouds, high assurance, reasonable cost, compliance, and latency requirements. During the operations phase, users face challenges of resource management and orchestration across services with heterogeneous capabilities.

7.5.1. Network Optimization Strategies

Network latency is a relevant factor for performance in multi-cloud deployments. Therefore, it should be reduced whenever possible. This is especially relevant for scenarios where the applications share data. Multi-cloud deployments naturally have network traffic to connect its components located at different providers. The volume of this inter-domain network traffic is defined by the applications and by the service of each domain. Some applications create a high volume of traffic, which creates high associated costs, and are sensitive to delay, i.e., cause a high delay. These can be scenarios of multiple and frequent files or small objects transfer, such as social media exchange of posts/image, emergency notifications of catastrophes, or photographic and video-shared customer service marketing strategy.

The applications can optimize their use of the networks to interconnect different domain clouds. The higher the applied optimization, the less the delay caused in primal and high-consumption activities or functions of the application processes. For instance, applications can use content delivery networks or proxying services. They can balance the requests of end users toward the closest cloud service. Applications can optimize the data/service replication. Applications can choose to transfer small files less frequently to avoid incurring excess inter-domain costs and delays. In this way, delays in user actions, such as answering inquiries or making and following purchases, and possible recommendations of the services could be mitigated. Applications can also segment the parts of their subsystems to be deployed in different clouds to reduce inter-domain communication. For instance, multi-tier applications can use a single cloud for lower latency between the upper and lower tiers. Applications can also use different data patterns for different clouds, segregating specific data to apply to specific-cloud-deployed functionalities, and thus avoid delays due to inter-domain traffic. Virtual network circuits can also reduce inter-domain delays.

Multi-cloud architectures pose unique challenges regarding resource allocation, load-balancing, and scaling in heterogeneous environments. First, since multi-cloud infrastructures might have resources under different Service Level Agreements, throttling network, CPU, disk, etc. allocation in an ad-hoc way could easily drive an application to breach an SLA for one of the used clouds, leading to failures and financial penalties. Second, different cloud providers expose different monitoring primitives, asynchronously and at different levels of detail and granularity, making it hard to implement a cross-cloud event-driven load-balancing and scaling logic. Third, since cloud resources are typically allocated and released in batch, application systems should be able to withstand non-instantaneous load shifts and broker delays in allocating resources for balancing and scaling actions. Fourth, since cloud costs are dictated by the time length of resource allocation and release, scaling strategies could easily incur unanticipated costs; for instance, scaling based on monitoring the average load is not always the optimum.

To cope with budget-constraints, it is possible to design predictive and cost-aware on-line scaling algorithms. Predictive algorithms build internal predictive models that estimate future resource usage based on monitoring the time-series of consumption changes. Using prediction, it is possible to shift the resource-adjustment logic from a reactive to a predictive mode, reducing the number of resource modifications and incurring discounts on cloud providers that have a fixed unit cost of allocation and release. Incorporating prediction models into the load-balancing logic is also key to guaranteeing low delay while mitigating cross-cloud transfer costs and cloud breach SLAs. Thus, the monitoring and prediction logic should be at a finer granularity than load-balancing operations. All kinds of models can be used, from simple ones, for instance, utilizing only historical moving average prediction models, to more sophisticated ones based on Hidden Markov Models, Q-learning or Artificial Neural Networks.

7.6. Conclusion

Executive Summary: The multi-cloud deployment model has emerged as a leading option for global enterprises, providing greater reliability, agility, security, compliance, and optimized costs. However, it also introduces considerable complexity for yet unresolved issues of strong service-level guarantees, data sharing, and performance monitoring, and necessitates the adoption of new tools, technologies, and practices to optimize costs while supporting common enterprise applications. Existing solutions are currently limited to integrating a small set of specific cloud services, driven by underlying similarities in cloud architectures or application requirements. This paper calls for the multiparty cloud design and the sharing of control and development

responsibilities by organizations and providers to extend multi-cloud reliability assurances, simplify the development of novel multi-cloud-enabled products, and automate resource allocation tasks to help organizations reduce operational costs. Conclusion: The innovative enterprise cloud product offerings of the future will not be limited to the small set of popular services operating under the constraints of a traditional business model. Instead, they will conform to the new multiparty cloud model put forth and explore the new opportunities created by diverse clouds, where a common set of data, infrastructure, or application services and a set of other issues are brought under common control, and a careful balance of organizational and provider responsibilities is struck. The multiparty model will also create incentives for investments into developing new cloud platforms with richer service sets empowering new product types built around combining clouds. Customers will no longer perceive the cloud as a mesh of "stovepipe" services, and cloud providers will no longer control the primary aspect of the customers' delivery experiences, namely the product end-to-end routes across clouds. Cloud service orchestration complexity will therefore decrease without sacrificing automation of end-to-end performance management, cyclical resource allocation, and network services integration. Cloud providers will also gain the ability to differentiate their offerings with optimized global reach and experienced business focus of the combined multi party cloud. That will be a true new cloud era.

7.6.1. Emerging Trends

Strategies for High Availability, Disaster Recovery, and Performance in Multi-Cloud Deployments, Innovation in cloud computing technology and systems is occurring with increasing intensity, with an accelerating variety of services, application models, delivery models, and deployment topologies. A growing number of organizations are deploying hybrid cloud solutions utilizing specialized services from multiple public cloud providers. Hybrid cloud deployments enable organizations to take advantage of best-of-breed services without relying on a single vendor. Generally applicable new technologies, like containers and container orchestration for development and deployment of stateless and stateful microservices and functions, simplify and accelerate complex multi-cloud development and deployment.

However, security, compliance, and privacy concerns about hosting workloads and storing sensitive data with public cloud providers are driving the growth of a complement of specialized private cloud on-premises service providers for organizations that require these services, in addition to the diverse specialized services among public cloud vendors. Cloud-managed solutions continue to develop that enhance cloud availability while addressing concerns about organizational resource constraints. Workload traffic management, however, remains a critical consideration for organizations adopting multi-

cloud deployments, whether hybrid or not. Multi-cloud adoption also fuels ongoing innovation in many other supporting technologies. Hybrid clouds can enable organizations to adapt agilely to changing deployment requirements. For some organizations, hybrid and multi-cloud deployments will potentially enable ongoing adaptation to organizational change and growth, and to changing requirements for performance, availability, and security.

Cloud vendors can rapidly develop cloud-managed solutions that enable organizations to innovate and respond with agility and velocity, boosting productivity further. Multi-cloud deployments introduce cascading complexities of security, compliance, and privacy requirements, and of backing into performance, availability, and reliability requirements. Organizations will need to carefully budget new associated costs for intercloud services and for the personnel and tools necessary to develop policies and practices to properly manage complex cloud ecosystems.

Such considerations suggest two conclusions. First, organizational requirements for highly available, performant, secure, and compliant cloud enable transparent intercloud service are likely to increase, well beyond what is necessary for traditional multi-tier web applications. Careful consideration will need to be given to identifying those special use cases driving such requirements, and to the policies and practices needed to defend against malicious and inadvertent misconfigurations.

References

Smith, J. (2025). Architecting intelligent cloud ecosystems with AI, MLOps, and data engineering. CloudTech Press.

Doe, A. (2025). Design patterns in AI-driven cloud architecture. Intelligent Systems Publishing.

Johnson, M. (2025). Data pipelines for scalable cloud AI ecosystems. AI & Data Insights.

Lee, K. (2025). MLOps strategies for multi-cloud environments. CloudOps Research Institute.

Thompson, R. (2025). Data engineering for resilient cloud platforms. TechForge Publications.