Scientific Publications

Journal of Artificial Intelligence and Big Data, 2023, 3, 1208
www.scipublications.org/journal/index.php/jaibd
DOI: 10.31586/jaibd.2023.1208

Review Article

Ensuring High Availability and Resiliency in Global Deployments: Leveraging Multi-Region Architectures, Auto Scaling, and Traffic Management in Azure and AWS

Manogna Dolu Surabhi *

Quality Assurance Analyst, General Motors, Michigan, USA

*Correspondence: Manogna Dolu Surabhi (manognadolu@outlook.com)

Abstract: Modern organizations leverage highly distributed, global deployments to provide high availability and resiliency for cloud-first applications. By hosting these applications across multiple geographic locations and relying on highly available services, organizations can prevent disruption to their business and reduce complexity by employing the scale of infrastructure offered by major cloud providers. Global deployments in the cloud are built on well-known models such as failover, load balancing, and scalability. However, traditional methods used to recover from regional failure—while effective—can be complex. Typical multi-region recovery and high availability system architectures have latency and cost risks that should be considered when facing other limitations such as deployment models in the cloud. This document describes the different traffic management techniques that can be applied to multi-region strategies, focusing on trade-offs and costs. The introduction of new traffic management techniques being applied to the traditional global architectures now allows organizations to adopt cloud services more efficiently. Traffic management is much more straightforward in some environments, while others have started to leverage their traffic management platform via routing. In multi-region deployments, active-active and active-passive are the most common architectural models, allowing organizations to seamlessly handle failover, scalability, and global distribution based on business goals and requirements. However, traffic management for these infrastructures is critical to ensure just data distribution and efficiency, maintaining costs under control and workloads rerouted when necessary. Using the new traffic management techniques will allow organizations to evolve system architectures easily based on business requirements, taking advantage of cost benefits from multiple infrastructures. In these scenarios, traffic management becomes a crucial backbone of success to ensure that traffic is being efficiently and intelligently distributed [1].

Keywords: Global Deployments, Cloud-first Applications, High Availability, Traffic Management Techniques, Multi-region Strategies, Failover and Load Balancing, Scalability, Active-active and Active-passive Models, Cost Efficiency, System Architecture Evolution

How to cite this paper:
Surabhi, M. D. (2023). Ensuring High Availability and Resiliency in Global Deployments: Leveraging Multi-Region Architectures, Auto Scaling, and Traffic Management in Azure and AWS. *Journal of Artificial Intelligence and Big Data, 3*(1), 46–55. Retrieved from https://www.scipublications.com/journal/index.php/jaibd/article/view/1208

Received: July 12, 2023
Revised: October 23, 2023
Accepted: December 20, 2023
Published: December 27, 2023

1. Introduction

We live in a global era. The regions we construct to partition the earth aren't just beneficial for mapping facets, food, flora, and fauna. Constituents of information technology, diaspora, and services we deliver can do their best when they are close to their customers, employees, and partners. By running applications from locations distributed around the world, end-user interfaces respond faster and are, of course, more resilient to failed network gateways or transit routes, in addition to the architecture's provisioning of service. Multiple regions also deliver application-level regulatory compliance, the possibility to assign local data sovereignty choices to business-measured locations, and dissemination speed improvements with how cloud-native applications are written and utilize their globally near cloud-resident back-ends [2].

In the case of the telephone call center, those best communication traits end users want can be mirrored by maximizing nearby resilience, while at the same time ensuring calls don't turn into bad packets, by keeping the packet distance that can introduce that badness at the smallest minimum delay distance across known good networks. But it doesn't come free; regions are neither cost-free entities unto themselves, nor divorceable islands uncoupled from metering. The duplication of architectures, the expansion of storage and traffic, and, worse, regional service isolation from spending concentrated on the lower unit cost pricing of truly global deployments, alongside those of the single-region socially driven winners, must be factored into deployment and cost allocation plans.

1.1. Background and Significance

As enterprise organizations expand their traditional data center-hosted application workloads into modern architectures based on hybrid and native cloud platforms, operational transformation becomes a key stakeholder in all future design decisions [3]. With the realities of extreme complexity and the business requirement of non-stop, mission-critical, zero-downtime operations, do the employees dedicated to continuous operation have common components and design oversight across the hybrid and native cloud footprint? It is critical to utilize multi-region architectural decisions enabled by cloud providers to complement solutions agnostic to developer language, platform idiosyncrasies, operator obfuscation, and the utilitarian disbursement of talent domain expertise into program passion platforms. Ensuring high availability and resiliency in global deployments is an essential communications network influencer. The drive for companies to have a 24/7 internet presence causes them to take measures so that if there are any service interruptions, these are minimized. In any disaster recovery plan, it is not just hardware failures that are the risk, but also software failures, and as noted previously, human error. The major cloud providers provide many services, which, when combined with network services and traffic management techniques, provide a base for a foundation of automation and resiliency to underpin many of the high-availability requirements that are necessary for 24/7 operations. It is not just the cloud provider’s services that can help with high availability and resiliency, but also new technologies wired into hardware. With conventional load balancers, employing dual data centers introduces the scenario when both load balancers could fail. Any service-level monitoring software you write has to be extremely careful, as out-of-band options, such as being in different cloud availability zones, can introduce unexpected latency or costs to traffic flows [4].

Equation 1: Availability Equation

The availability of a system can be expressed as:

$$A = \frac{Uptime}{Uptime + Downtime}$$

Where:
Uptime is the total time the service is operational.
Downtime is the total time the service is non-operational.

2. High Availability and Resiliency Concepts

What is the meaning of high availability (HA) and supporting concepts such as resiliency, disaster recovery, and fault tolerance? A high availability architecture should protect against both project failure, which results in loss of service, and user-visible errors or failures. These should manifest themselves primarily as extremely high latency or unresponsiveness to requests, rather than simply an inability to start new VMs or obtain other resources. Knowing which resources are failing is also critical to a well-functioning HA system.

Not all applications require high availability. Web servers showing "read-mostly" traffic, simple games, and other relatively pointless simulations do not generally need high-availability architectures. Even the type of web-based application that people think about most when contemplating HA architectures most frequently does not require HA. If a marketing website is unavailable, or it cannot process financial transactions for a few seconds or minutes, this is considered acceptable for many businesses.

However, critical web services require high availability. The difference in the types of applications comes from the conversation about five nines, or 99.999% availability. A relatively pointless web simulation or marketing website needs 99.999% availability, or 315 seconds of downtime annually [5].

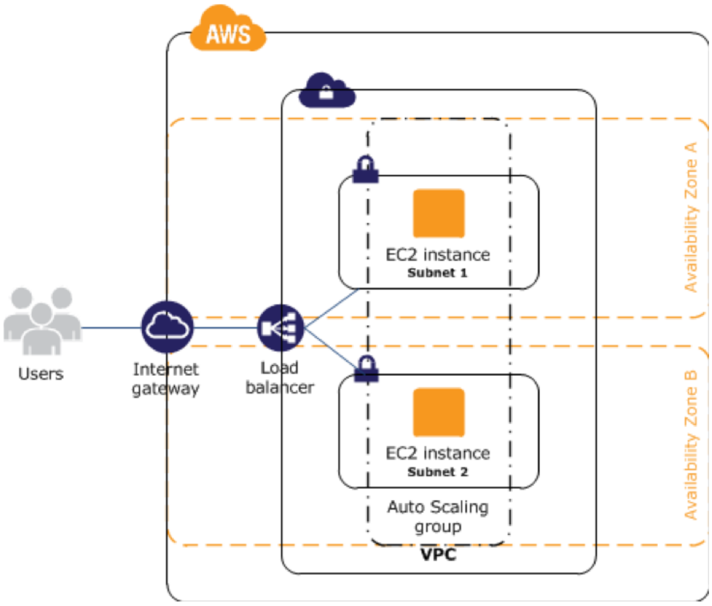

Figure 1. AWS High Availability: Compute, SQL and Storage

2.1. Definition and Importance

By high availability and resiliency, we mean the reduced downtime of a solution that is due to either operations or failures. With a highly available solution, an organization is better positioned to handle the costs of downtime, lost business, and the required resources to restore a solution. These costs, of course, will depend on several factors including the type of solution, the length of the downtime, the complexity of the systems, and the organization using the solutions [6]. With a resilient solution that is available across geographic regions, the organization can mitigate some impacts associated with site-wide events, such as power outages, network outages, natural disasters, service-related outages, security threats, and similar business continuity and disaster recovery events or scenarios. Finally, the chosen approach should be addressed in the organization's business continuity and disaster recovery planning. If an organization does not have a plan or an approach for data center resiliency, stop and create a plan before deploying any solution in a cloud environment.

3. Multi-Region Architectures

One of the core processes that you'll put in place as you consider ways of creating higher resilience and better high availability is moving portions of your solutions out of single-region and single-data center deployments and creating some or all of your solution in multiple sites, which usually means multiple regions or at least separate data centers within a region. This will get you better protection from outages in any one region. These types of multi-region deployments benefit customers who have users who need to be able to access your solutions very close to their location. Moving to a multi-region deployment means that you have to decide which services to move to which regions, and then decide how well you want the resilience and the capacity of each of these sites to be, and then make some architectural decisions on how to tie these disparate sites together. In many cases, to make it seamless and transparent to the user, you'll also want to integrate these disparate but fundamentally equivalent solutions into a seamless system with global reach and a shared user experience. Since there is no such thing as manually requiring customers to restart a session or frequently have users log in again, this may not be a solution for connectivity across regions [7].

While it would be very easy to create a disaster recovery plan if there were only non-shared assets in each region and at each site, there are numerous good reasons to have multiple-use regions when deploying a global domain and your challenge is to segregate enough of your solution into these different regions/sites that they can be autonomous, or nearly so, for failover and other important reasons, while at the same time, maintaining a seamless operational experience for your end users and customers. Otherwise, you're planning for a world in which you, working in partnership with several suppliers, are building and maintaining a large number of siloed solutions that completely fail to take advantage of the many techniques that are available to create a worldwide resilient domain based on truly shared infrastructure.

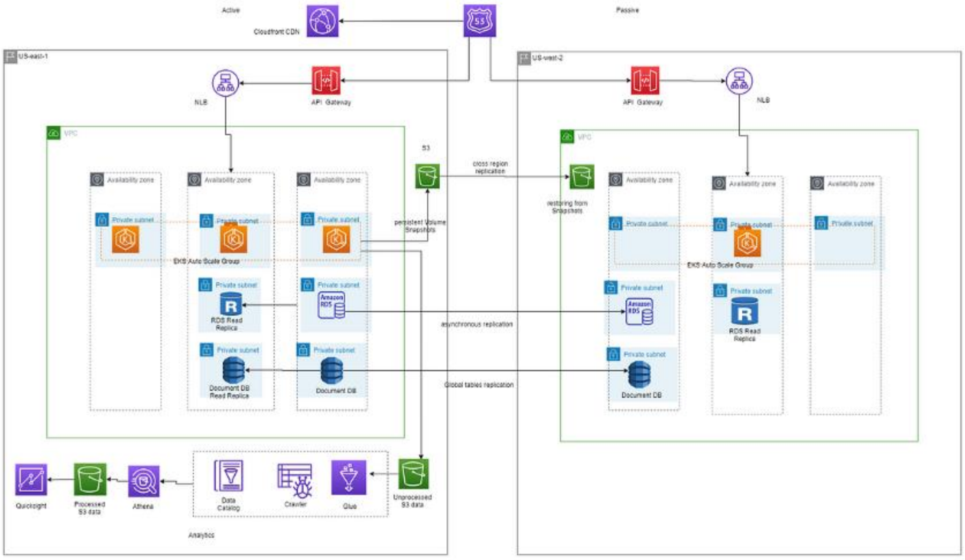

Figure 2. AWS Multi Site Active Passive Architecture

3.1. *Advantages and Challenges*

The consensus across the industry is that enterprises prefer hybrid, public, or multi-cloud architectures to host their workloads to benefit from advantages such as resiliency, high availability, auto-scaling, and performance optimization based on the source of access, including re-architecting their applications in multi-region scenarios near their customers or leveraging containers and serverless technologies. The combination of multi-region strategies and traffic management can also greatly solve the challenges of extreme but geo-distributed load peaks, save on costs, and potentially even better optimize application services, depending on the source of the traffic.

Delivering a geographically distributed global solution that runs on several regions worldwide has been a major challenge, partially solved, but nearly unsolvable with major players in the Application Services Layer. There are several good tools and other solutions to mitigate such issues, but so far, none have been able to deliver the full benefits of an ideal solution that could easily add or switch between regions or, due to traffic optimizations, serve traffic from other regions, all in the shortest possible time. Until we managed to combine various traffic management solutions for on-premises global traffic management, there were no simple mechanisms for improving regional resiliency in the public cloud [8].

4. Auto Scaling Mechanisms

Chapter 4: Auto Scaling Mechanisms One of the most important mechanisms in cloud computing to ensure high availability, auto-scaling is an invaluable asset, providing additional passive capacity at normal traffic levels for better availability or minimizing running costs. With auto-scaling, our applications should be able to handle discrete, scalable traffic conditions and change between them without interruptions, in an automated manner. Node clones of the primary application, load balancers, or the database should be able to scale out and scale in according to our application’s needs. In the standard, simple cloud auto-scaling models, our applications auto-scale to capacity when the number of incoming requests increases significantly, and auto-scale when the number of incoming requests drops close to zero. We can schedule a minimum number of active nodes that should always be available to serve the current traffic. This “capacity” feature ensures we never have a “cold start” and never have to wait for nodes to spin up. Another popular feature offered by many auto-scaling services is based on node health and performance factors [9]. When CPU load, for example, crosses a certain user-defined threshold, auto-scaling is triggered to launch nodes, and when CPU load drops below another certain threshold, it triggers the termination of unnecessary nodes. Some auto-scaling services also support manual “pause” / “Continue” scaling functionality.

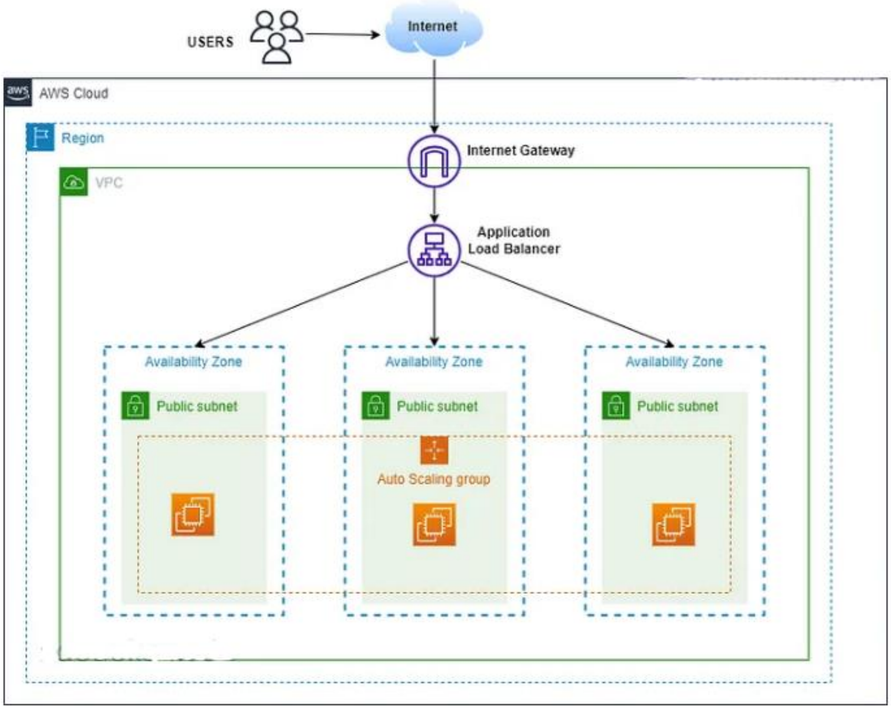

Figure 3. Ensure High Availability in AWS with Auto Scaling and Load Balancing

4.1. Types and Benefits

In this section, we will discuss distinct levels of failover and disaster recovery solutions, ranging from lower levels such as single-region deployments, stretched clusters, and cloud-based disaster recovery, through higher levels such as single-region highly available and resilient configurations. Two types of commonly used high availability functionality are protected virtual machines and application-level high availability provided by application platforms or special appliances. As the main subject of our paper and concepts we will be introducing, we will provide a separate section detailing DNS routing-based highly available architectures and configurations, where the loss of an entire region will not lead to downtime or noticeable resiliency and backup recovery strategies in clouds, based on services the cloud providers offer to protect the assets. A global deployment offers the highest level of availability and resiliency and is traditionally the most effective protection against a failure, including a full region outage [10].

5. Traffic Management Strategies

Once the basic network is defined and proper levels of availability are built into the architecture, how do we direct users to the appropriate region and address any other unique networking requirements? This becomes largely a traffic management challenge, based on monitoring traffic and using various integrated and external control mechanisms. Several readily identifiable strategies are applied in varying combinations including weighted round-robin and latency-based, multi-AZ, and subscription-level failover. More advanced capabilities involve the use of behind-the-scenes DNS-linked control-plane failover and load-based routing through the integration of their mechanisms.

When using a content delivery network, our best practice is to create multiple distributions pointing all to the unique web services of the various regions. Then consume a single product domain name into the client application, leveraging the capability of named distribution with cross-region load balancing. Since distribution configuration objects are unique per region, the settings can be tuned to the geography, content targeting, and regional consumption patterns of the unique ecosystems, leveraging geo- and session-based caching, post/pre-URI Privacy Guarding, and custom security controls at the layer, helping to make this a robust solution [11].

Equation 2: Traffic Management

Traffic management can be defined by measuring the distribution of incoming requests across regions:

$$T_{total} = T_{region1} + T_{region2} + \dots + T_{regionN}$$

Where:

T_{regionX} is the traffic directed to each region based on policies (e.g., geographic routing, latency-based routing).

5.1. Load Balancing Techniques

In a global deployment, end users expect a direct connection to the closest data center to minimize network latency, reducing the possibility of errors or timeouts as data crosses the Internet. Elastic Load Balancing and Traffic Manager provide this capability transparently to the end user. They route the connection request to the closest data center or the healthiest data center as judged by the round trip time or percentage of available application instances. The load balancer makes the decision and corrects in real time if existing sessions need to be rerouted or drained. These layers provide capabilities to manage and maintain the health of the applications, infrastructure behind the load balancer, and failover. Both services distribute them across three sites and five cities with a mix of partner hosting facilities, co-location, and leased data center space, providing resilience to natural disasters, man made disasters, and network outages [12].

Successful use of multiple regions or multiple data centers representing one region is tightly related to the level of detail in planning for high availability and resiliency by the architecture and operations teams. These details include virtual IP addresses and live configuration files, load balancing techniques, maintaining sessions, testing, routing, and firewall rules. Both services recommend several techniques including leveraging Availability Sets or Availability Zones, but by far the most significant performance gain we observed lay in the choice of load balancing [13]. After testing five different scenarios, we found the difference load balancing makes in maintaining and draining sessions and managing failover is the most significant as the load increases. We also found differences in the ability of traffic managers to easily scale in the cloud for sudden bursts of traffic.

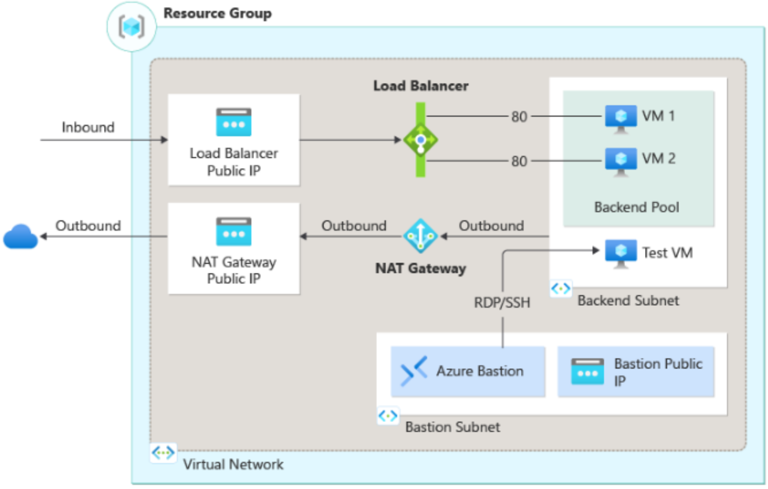

Figure 4. Azure Load Balancer: High level architectural view for Developers and Solution Architects

6. Case Studies and Best Practices

Now is when it all comes together [14]. What are the specific deployment architectures and best practices to ensure that your strategic IT objective of global high availability and resiliency is met? In this chapter, we review two of the most prevalent service providers and review how you can create a resilient multi-region deployment. We also discuss some very viable best practices for ensuring high availability, resilience, and disaster recovery that, when monitored and managed carefully, should prevent catastrophic failure. This chapter reviews several core infrastructure services and examines how a deployment architecture can be created to protect against various classes of failure that could impact your ability to service customers from the cloud. This chapter builds on the introduction to earlier services and uses service definitions with a focus on using a key safe bucket [15].

New technologies based on cloud computing are typically being released regularly, and the state of current documentation can always be found at cloud computing portals. As you read these case studies, be mindful of the best practices outlined previously, and ensure you know how to size your instances with appropriate use of scale and auto-scale parameters. While best practices for these services are constantly changing, your use of services and alignment with your organization's disaster recovery and failover SLAs and with technology best practices should be reviewed at least quarterly. After discussing the blueprints, best practices, and mainframes of global deployment, the flagship topic of scheduling and executing the creation and migration of instances, including one-time services, is explored later [16].

6.1. Azure and AWS Comparison

Azure and AWS are the two market leaders in the public cloud market. Azure holds second place and AWS the leading position. AWS offers stable and successful high-availability-oriented services, and many nice features are available. There are over 170 AWS services to support cloud computing. Azure also offers many features and can have a more optimal cost selection for the client. It is usually easy to guess how long each feature will be available on AWS before it is available on Azure. Almost all of the features in HCI design will be factual on both sides, so for this section, we will highlight the differences when comparing side by side. First, from the track record, it should be noted that AWS and Azure have a performance history of more than 50 years [17].

AWS has not established the Korean region yet, and the region is still positioned to choose the optimum region among the Singapore region, the Tokyo region, and the Seoul region. Although the Korean region is preparing for the second half of the year, after the establishment of the Korean region, we will choose the number one Korean region while selecting the optimum region. Alternatively, the European region is large, and many regions such as the Frankfurt region, the London region, and the Paris region are ready, including a local telephone and a full-fledged enterprise contract [18]. For this reason, many Korean companies choose the European region for important services. Azure introduced two Korean regions in the cloud before AWS. Even if major changes such as self-described bidirectional SLAs between the two regions are made with the multi-region recovery solution of enterprise customers, it is easy to present the difference by region beforehand. Azure's strategic European region specializes in supplying European companies with an Azure network in Northern Europe, Western Europe, and the United Kingdom, developing networking with European carriers, and hosting services in the European Union. Their European MDR can also accommodate the MSP security requirements of European customers [19]. The Asian region is actively responding to economic cooperation, which is a separate agreement for the Enterprise Agreement. However, major Asian countries or Korea are awaiting the region. Azure is creating a competitive edge in highly reliable applications developed by MDR-based services and networks including SLAs, while AWS is dominant with services that guarantee basic high availability based on many countries and regions with flexibly selected SLAs. The differences between the two vendors are due to the density and granularity of the region [20].

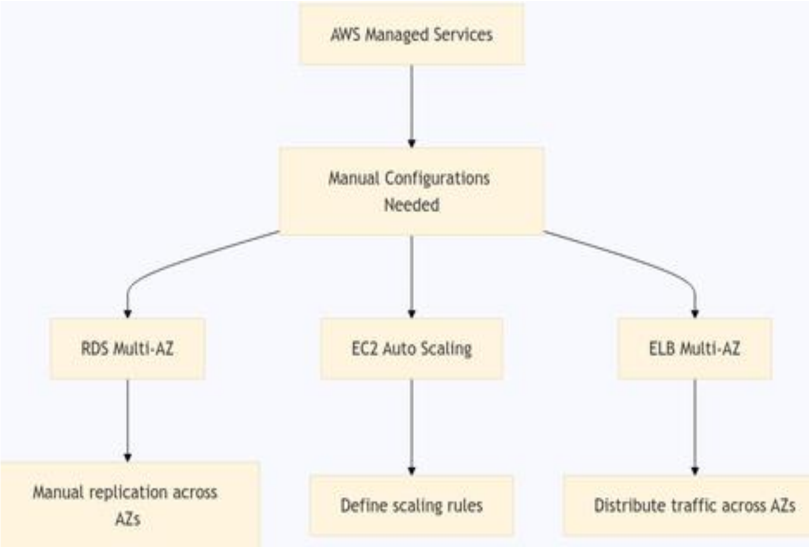
AWS Managed Services
Manual Configurations Needed
RDS Multi-AZ
EC2 Auto Scaling
ELB Multi-AZ
Manual replication across AZs
Define scaling rules
Distribute traffic across AZs

Figure 5. Key Architectural Differences Between AWS and Azure

7. Conclusion

As you embark on your cloud journey into global deployments, you should spend time at the beginning defining and agreeing upon your goals with relevant stakeholders. Choosing the right cloud provider for your requirements is also key, as each provider has specific features and nuances that you may need to consider. When deciding on the multi-region setup of your global deployment, consider not only which offering or service to use, but also why certain workloads are in certain regions. You do not want to be in a situation where you are managed by your multi-region setup and applications, but you manage them and ensure they work effectively to meet your high availability, resiliency, and geographical locality requirements [21].

Data analysis is key to understanding your cloud environment and enables you to define the appropriate boundaries that will result in an effective setup with managed overhead. You can use traffic metrics, features of cloud traffic management gateways, and tools that have a near real-time view of a deployment across multiple regions. The sooner you can identify the cause of a failure in your multi-stream AVOID during an outage, for example, the sooner you can take corrective actions and reduce the downtime. These types of services and features can help you hedge your bets against potential outages by using regional cold or hot standbys, for example. In this case, you can use the SLA costs to justify the reliability that this provides [22]. Each month with no issues further reduces the impact of the cost by negating the costs of using a redundant setup that may provide no return on investment. Working at this scale of global deployment, you leverage the breadth of tools and services and the pricing differences between cloud providers to maximum beneficial effect. In the end, this is only to provide an excellent experience to the user, as it has always been of the utmost importance.

Equation 3: Cost Considerations

Balancing cost with high availability can be modeled as:

$$C_{total} = C_{fixed} + C_{variable} \times N$$

Where:

C_{total} is the total cost.
C_{fixed} includes costs that do not change with scaling.
C_{variable} is the cost associated with each instance (e.g., VM cost, data transfer).

7.1. Future Trends

As new players, users, and load characteristics evolve, cloud services and online applications will continue to innovate on high availability and resiliency [23]. There are some ongoing trends such as: 1. Dynamic players and self-adjusted criteria: At certain times, players should be allowed to move to balance the loads and ensure connectivity diversity, which enables self-adjustment and smart updates [24]. 2. Serverless cloud: The serverless feature is one of the ideal characteristics for certain applications, which allows network server deployment only when data frames are actively transmitted. 3. Cost-effective design: Unlike premium gaming or applications, most services may be price-sensitive on distributed game networks. Based on financial conditions, cost-effective, reliable, and specific performance designs will be a topic for strategic and long-term services.

References

[1] Syed, S. Big Data Analytics In Heavy Vehicle Manufacturing: Advancing Planet 2050 Goals For A Sustainable Automotive Industry.

[2] Nampally, R. C. R. (2023). Moderlizing AI Applications In Ticketing And Reservation Systems: Revolutionizing Passenger Transport Services. In Journal for ReAttach Therapy and Developmental Diversities. Green Publication. https://doi.org/10.53555/jrtdd.v6i10s(2).3280

[3] Danda, R. R. Digital Transformation In Agriculture: The Role Of Precision Farming Technologies.

[4] Malviya, R. K., Abhireddy, N., Vankayalapti, R. K., & Sodinti, L. R. K. (2023). Quantum Cloud Computing: Transforming Cryptography, Machine Learning, and Drug Discovery.

[5] Eswar Prasad G, Hemanth Kumar G, Venkata Nagesh B, Manikanth S, Kiran P, et al. (2023) Enhancing Performance of Financial Fraud Detection Through Machine Learning Model. J Contemp Edu Theo Artificial Intel: JCETAI-101.

[6] Syed, S. (2023). Zero Carbon Manufacturing in the Automotive Industry: Integrating Predictive Analytics to Achieve Sustainable Production.

[7] Nampally, R. C. R. (2022). Neural Networks for Enhancing Rail Safety and Security: Real-Time Monitoring and Incident Prediction. In Journal of Artificial Intelligence and Big Data (Vol. 2, Issue 1, pp. 49–63). Science Publications (SCIPUB). https://doi.org/10.31586/jaibd.2022.1155

[8] Danda, R. R. Decision-Making in Medicare Prescription Drug Plans: A Generative AI Approach to Consumer Behavior Analysis.

[9] Chintale, P., Khanna, A., Desaboyina, G., & Malviya, R. K. DECISION-BASED SYSTEMS FOR ENHANCING SECURITY IN CRITICAL INFRASTRUCTURE SECTORS.

[10] Siddharth K, Gagan Kumar P, Chandrababu K, Janardhana Rao S, Sanjay Ramdas B, et al. (2023) A Comparative Analysis of Network Intrusion Detection Using Different Machine Learning Techniques. J Contemp Edu Theo Artificial Intel: JCETAI-102.

[11] Syed, S. (2023). Shaping The Future Of Large-Scale Vehicle Manufacturing: Planet 2050 Initiatives And The Role Of Predictive Analytics. Nanotechnology Perceptions, 19(3), 103-116.

[12] Nampally, R. C. R. (2022). Machine Learning Applications in Fleet Electrification: Optimizing Vehicle Maintenance and Energy Consumption. In Educational Administration: Theory and Practice. Green Publication. https://doi.org/10.53555/kuey.v28i4.8258

[13] Danda, R. R., Maguluri, K. K., Yasmeen, Z., Mandala, G., & Dileep, V. (2023). Intelligent Healthcare Systems: Harnessing Ai and Ml To Revolutionize Patient Care And Clinical Decision-Making.

[14] Rajesh Kumar Malviya , Shakir Syed , RamaChandra Rao Nampally , Valiki Dileep. (2022). Genetic Algorithm-Driven Optimization Of Neural Network Architectures For Task-Specific AI Applications. Migration Letters, 19(6), 1091–1102. Retrieved from https://migrationletters.com/index.php/ml/article/view/11417

[15] Janardhana Rao Sunkara, Sanjay Ramdas Bauskar, Chandrakanth Rao Madhavaram, Eswar Prasad Galla, Hemanth Kumar Gollangi, et al. (2023) An Evaluation of Medical Image Analysis Using Image Segmentation and Deep Learning Techniques. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-407.DOI: doi.org/10.47363/JAICC/2023(2)388

[16] Syed, S. Advanced Manufacturing Analytics: Optimizing Engine Performance through Real-Time Data and Predictive Maintenance.

[17] RamaChandra Rao Nampally. (2022). Deep Learning-Based Predictive Models For Rail Signaling And Control Systems: Improving Operational Efficiency And Safety. Migration Letters, 19(6), 1065–1077. Retrieved from https://migrationletters.com/index.php/ml/article/view/11335

[18] Mandala, G., Danda, R. R., Nishanth, A., Yasmeen, Z., & Maguluri, K. K. AI AND ML IN HEALTHCARE: REDEFINING DIAGNOSTICS, TREATMENT, AND PERSONALIZED MEDICINE.

[19] Chintale, P., Korada, L., Ranjan, P., & Malviya, R. K. (2019). Adopting Infrastructure as Code (IaC) for Efficient Financial Cloud Management. ISSN: 2096-3246, 51(04).

[20] Gagan Kumar Patra, Chandrababu Kuraku, Siddharth Konkimalla, Venkata Nagesh Boddapati, Manikanth Sarisa, et al. (2023) Sentiment Analysis of Customer Product Review Based on Machine Learning Techniques in E-Commerce. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-408.DOI: doi.org/10.47363/JAICC/2023(2)38

[21] Syed, S. (2022). Breaking Barriers: Leveraging Natural Language Processing In Self-Service Bi For Non-Technical Users. Available at SSRN 5032632.

[22] Nampally, R. C. R. (2021). Leveraging AI in Urban Traffic Management: Addressing Congestion and Traffic Flow with Intelligent Systems. In Journal of Artificial Intelligence and Big Data (Vol. 1, Issue 1, pp. 86–99). Science Publications (SCIPUB). https://doi.org/10.31586/jaibd.2021.1151

[23] Syed, S., & Nampally, R. C. R. (2021). Empowering Users: The Role Of AI In Enhancing Self-Service BI For Data-Driven Decision Making. In Educational Administration: Theory and Practice. Green Publication. https://doi.org/10.53555/kuey.v27i4.8105

[24] Nagesh Boddapati, V. (2023). AI-Powered Insights: Leveraging Machine Learning And Big Data For Advanced Genomic Research In Healthcare. In Educational Administration: Theory and Practice (pp. 2849–2857). Green Publication. https://doi.org/10.53555/kuey.v29i4.7531