Consistent Access to Cloud Services across Regions for Large Enterprises

Authors:

Prisha Goel

Pavvan Pradeep

Dhruv Sanjaykumar Ratanpara

Aditi Srinivas M

Shilpa S

Presenter: Prisha Goel

Prisha Goel

4th year student at People's Education Society, Bangalore, India.

Professional Experience:

-Interned at CoffeeBeans Consulting as a Backend Engineer
-Interned at Fidelity Investments as a Full Stack Engineer
-Currently interning at xto10x as a Software Developer

Other Achievements:

-Top 6 Cisco Campus Ambassadors nationwide
-Teacher's Assistant for Database Management Systems
-Event lead at Codechef, member of Association of Computing Machinery and IEEE Student Chapter

Overview of the Problem

Today, one of the critical challenges faced by large enterprises is ensuring uninterrupted service availability across multiple regions when there is a high volume of user requests. While the existing cloud providers offer high-availability services, many of them are available only in selected regions. They don’t have multi-region availability, which leads to enterprises being vulnerable to potential downtime.

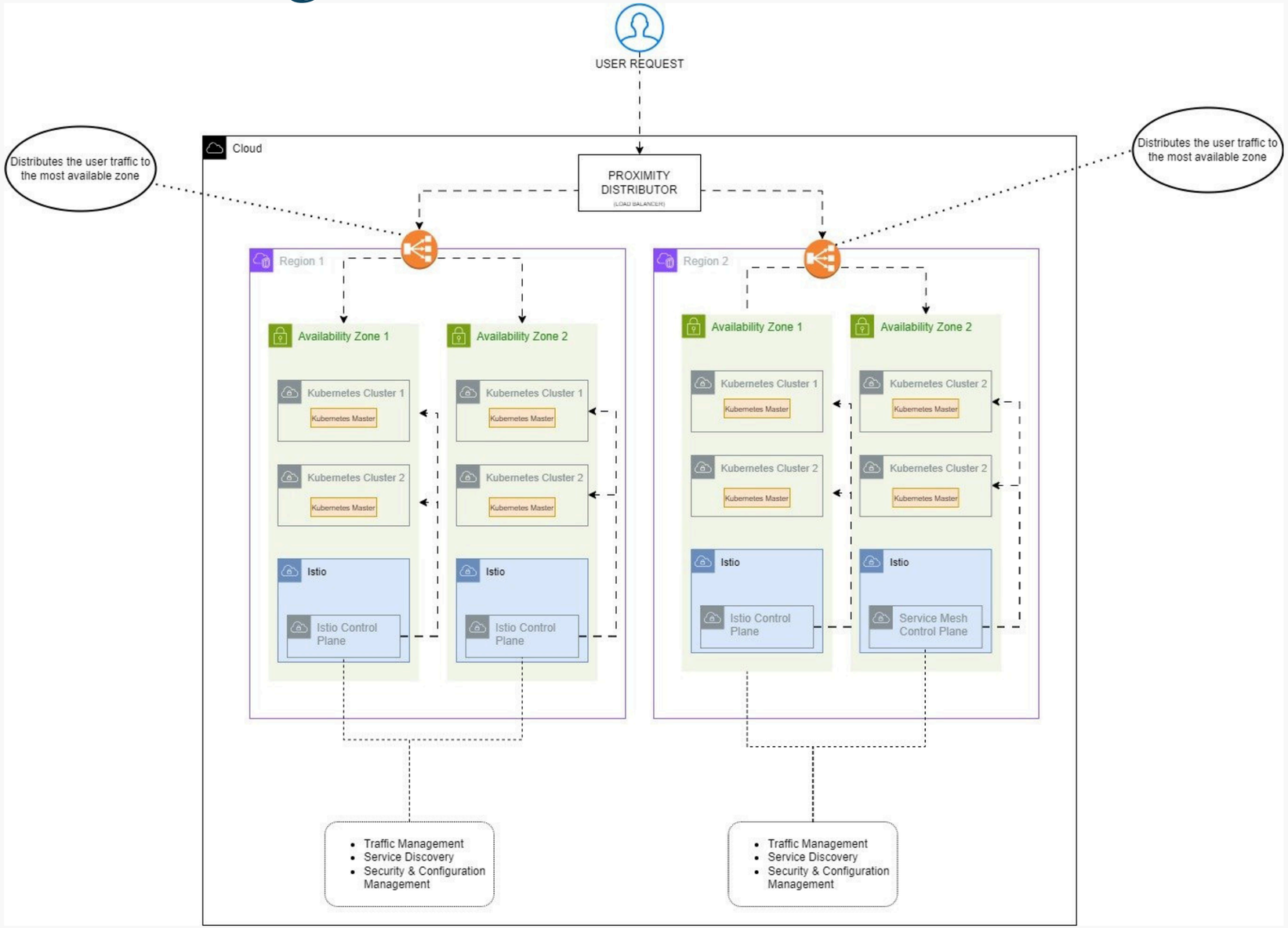
Approach

Our approach focuses on designing custom cloud platforms for large-scale enterprises to ensure continuous service availability across multiple regions. We utilize Kubernetes for container orchestration to guarantee scalability, resilience, and optimal performance. Additionally, we integrate Istio service mesh to enable secure, seamless communication, fine-grained traffic control, and comprehensive monitoring.

Architecture

- Our architecture utilizes the Istio service mesh to facilitate secure and scalable communication between microservices.
- A multi-primary Istio configuration across clusters in different regions ensures high availability and resilience.
- A custom load balancer efficiently manages routing and minimizes latency, while Istio’s failover mechanism guarantees seamless traffic management and service continuity, even in the event of zone outages.

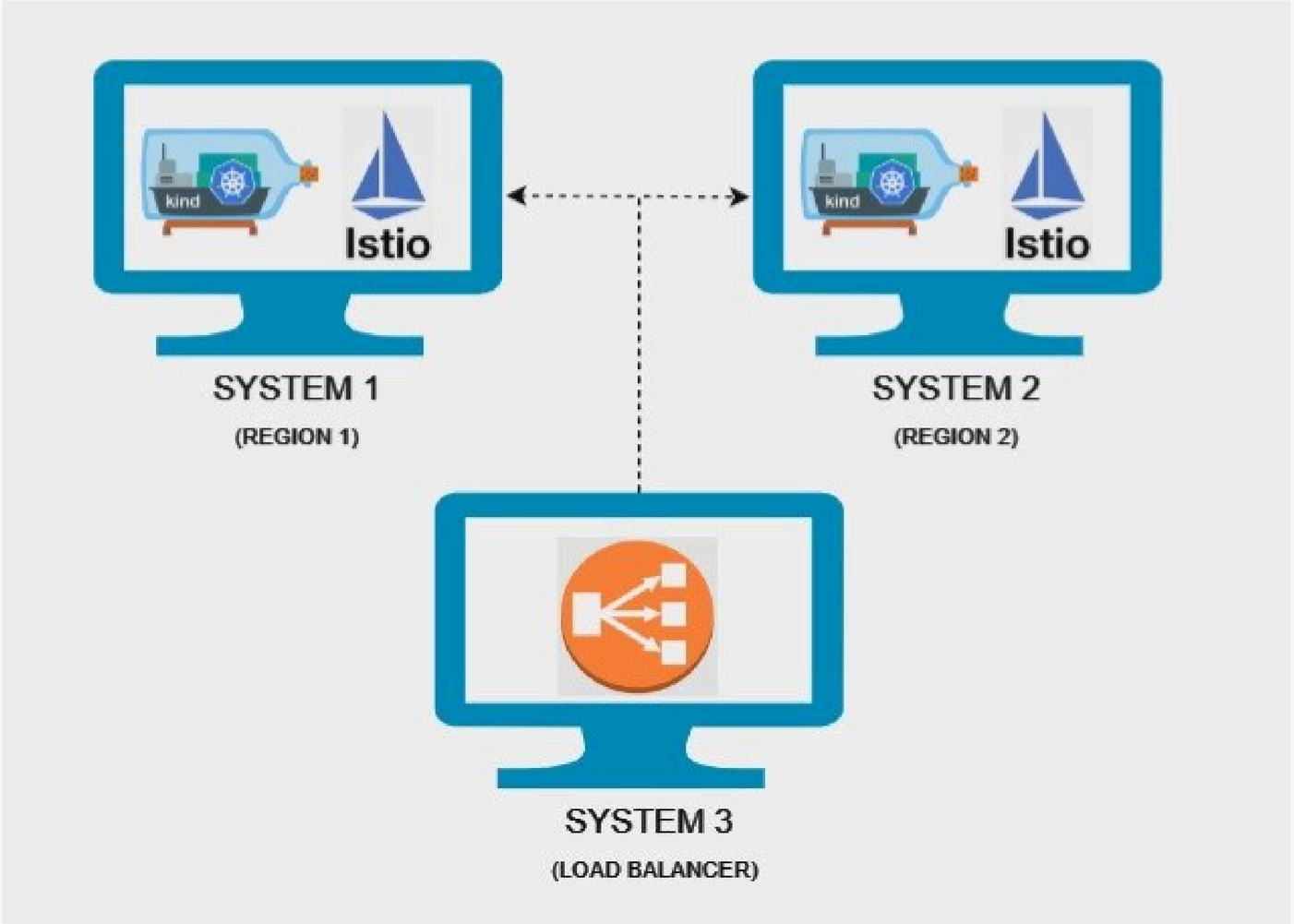
Architecture Diagram



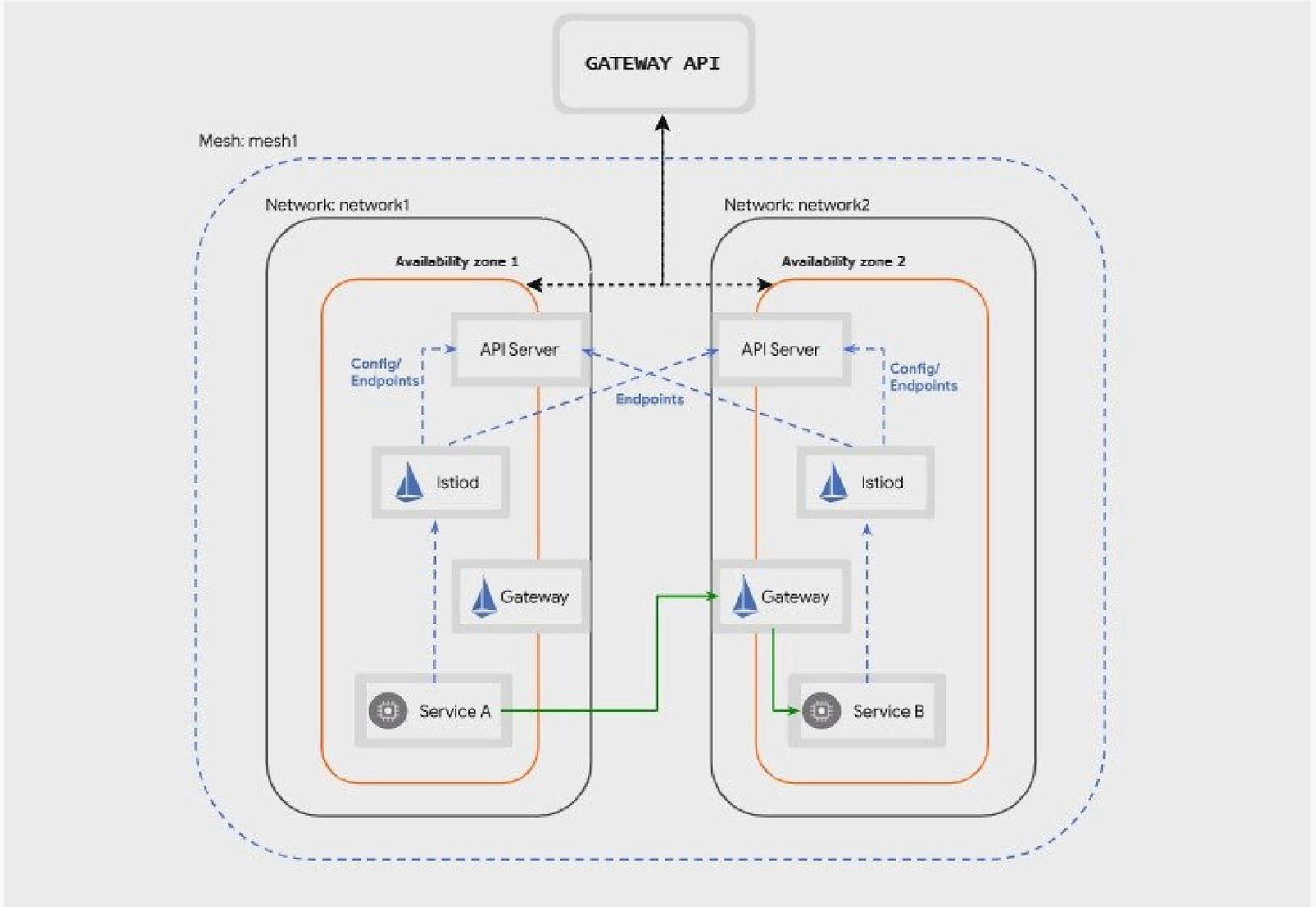
Implementation Details (Istio Integration)

- Multi-Primary Setup: Istio is integrated into the local KIND development infrastructure using a multi-primary approach, leveraging its mesh capabilities across both availability zones within a region. A multi-primary Istio configuration across clusters in different regions ensures high availability and resilience.
- The control planes are installed in each cluster and these clusters are given access to the API server of each other. Thus, there is a single service mesh (different Istio control planes), and the service discovery is seamless in the mesh. The Istio control planes don’t interfere with each other, so the failure of one Istio control plane will not affect other primary clusters.

Implementation Details



Implementation Details (Istio Integration)



Implementation Details (Istio Integration)

- Traffic Management: Istio's traffic routing features are used to distribute requests efficiently, enhancing the overall availability and scalability of the application.

Global Load Balancer Implementation

- **Location Implementation:** Retrieves user location (latitude and longitude) based on IP using the ipinfo.io API and Identifies the nearest healthy instance by calculating distances using the haversine formula, ensuring optimal service delivery.
- **Health Check Implementation:** Assesses server health by making an HTTP GET request, validating the server URL, and ensuring thread safety with mutex locks. Handles errors, non-200 status codes, and marks servers as unhealthy or healthy based on the response.

Benefits

Cost Benefits:

- **EC2 Instances:** For production grade servers across regions (eg. T3.medium or m5.large instances), assuming 10-20 instances per region for redundancy, costs could range from $1,500 - $3,000 per region, totaling $3,000 - $6,000**.**
- **EKS and Kubernetes Management:** AWS EKS pricing is $0.10 per hour per cluster plus EC2 costs for worker nodes, so a multi-region setup might cost around $1,000 - $2,000.
- **Load Balancers and Network Transfer:** With multiple load balancers and high outbound data transfer, costs might range from $500 - $2,000.

Benefits

- **Storage (eg. EBS, S3):** Persistent storage with high availability (eg. Amazon RDS or S3 for data durability) adds an additional $500 - $1,000.
- **Total Estimate:** For a mid-to-large-scale deployment, the monthly operational cost might range from $5,000 - $12,000.
- For Large Corporations and real time applications, the costs can add up to $20,000 - $50,000 monthly.
- **Since we are not dependent on AWS or any other existing cloud providers, and only take advantage of open source tools and technologies, our total cost of operation is $0.**

References

- [1] A. Malhotra, A. Elsayed, R. Torres, and S. Venkatraman, ”Evaluate
- Solutions for Achieving High Availability or Near Zero Downtime for
- Cloud Native Enterprise Applications,” in *IEEE Access*, vol. 11, pp.
- 85384-85394, 2023, doi: 10.1109/ACCESS.2023.3303430.
- [2] A. Johnson and B. Lee, ”Auto-Scaling Strategies for High Availability
- in AWS,” in *Proceedings of the International Conference on Cloud
- Computing*, 2017, pp. 112-125.
- [3] H. Lang and C. Li, ”Containerization for High Availability in Cloud
- Environments,” in *Proceedings of the International Conference on
- Cloud Computing and Big Data*, 2019, pp. 201-214.
- [4] G. Verma and R. Sushil, *Cloud Computing Implementation: Key Issues
- and Solutions*, Springer, 2015.
- [5] N. Ahmad et al., ”Strategy and Procedures for Migration to Cloud Com-
- puting,” in *Proceedings of the 2018 IEEE 5th International Conference
- on Engineering Technologies and Applied Sciences (ICETAS)*, 2018,
- pp. 1-5.
- [6] V. Mohammadian et al., ”Fault-Tolerant Load Balancing in Cloud
- Computing: A Systematic Literature Review,” in *IEEE Access*, vol.
- PP, pp. 1-1, 2021.
- [7] W. A. Aziz, ”High Availability Solution for Cloud Applications,” in
- *International Journal of Simulation: Systems, Science Technology*,
- vol. 24, 2023.
- [8] A. Berenberg and B. Calder, ”Deployment Archetypes for Cloud Appli-
- cations,” in *ACM Computing Surveys*, vol. 55, no. 3, Article 61, pp.
- 1-48, March 2023. doi: 10.1145/3498336.[9] A. Hajikhani and A. Suominen,
- ”The Interrelation of Sustainable
- Development Goals in Publications and Patents: A Machine Learning
- Approach,” *Trepo Digital Repository, Tampere University*, 2021.
- [10] M. R. Mesbahi, A. M. Rahmani, and M. Hosseinzadeh, ”Reliability
- and High Availability in Cloud Computing Environments: A Reference
- Roadmap,” in *Human-Centric Computing and Information Sciences*,
- vol. 8, no. 20, 2018. doi: 10.1186/s13673-018-0143-8.
- [11] A. Malhotra, A. Elsayed, R. Torres, and S. Venkatraman, ”Evaluate
- Solutions for Achieving High Availability or Near Zero Downtime for
- Cloud Native Enterprise Applications,” in *IEEE Access*, vol. 11, pp.
- 85384-85394, 2023, doi: 10.1109/ACCESS.2023.3303430.
- [12] S. Afzal and G. Kavitha, ”Load Balancing in Cloud Computing – A
- Hierarchical Taxonomical Classification,” in *Journal of Cloud Com-
- puting*, vol. 8, no. 22, 2019. doi: 10.1186/s13677-019-0146-7.
- [13] E. Bauer and R. Adams, ”Service Reliability and Service Availability,”
- in *Reliability and Availability of Cloud Computing*, Hoboken, NJ:
- Wiley-IEEE Press, 2012.
- [14] R. Brown and M. Davis, ”Designing Fault-Tolerant Cloud Applications:
- A Virtualization-Based Approach,” in *IEEE Transactions on Services
- Computing*, vol. 18, no. 4, pp. 567-581, 2020.

Thank you