Traffic Pattern Segmentation for Large Scale eCommerce Applications

Title: Traffic Pattern Segmentation for Large-Scale eCommerce Applications

Author(s): Priyadarshini Jayakumar <priyakumarj@outlook.com> ; Anand Kumar Ganapathy
Anandkumar.Ganapathy2@t-mobile.com

Date: July 2023

1. Abstract

This white paper explores the implementation of traffic pattern segmentation in large-scale eCommerce applications. By leveraging segmented traffic routing within an active-active infrastructure set-up, businesses can enhance performance, reliability, and user experience. This paper delves into methodologies such as blue/green deployment and real-time A/B testing, supported by case study demonstrating successful applications of these strategies to split traffic effectively.

2. Introduction

Large-scale eCommerce platforms face significant challenges in managing and routing traffic due to high user demand and diverse behaviors specifically during events like product launches. Traditional traffic management techniques often fall short in addressing these complexities. Implementing traffic pattern segmentation, combined with segmented routing within an active-active infrastructure set-up, offers a robust solution to optimize traffic flow, ensure high availability, and enable real-time A/B testing.

There are certain pre-requisites that allow large scale companies to achieve a well-defined traffic segmentation such as setting up the required infrastructure, experimenting on various deployment strategies and having backup readily available. We will dive deep into some of these methodologies and their implementation strategies here.

3. Background

3.1 Active-Active Infrastructure

An active-active infrastructure involves deploying multiple active instances of an application across different geographic locations. This setup enhances availability, fault tolerance, and load balancing, making it ideal for large-scale eCommerce applications to be stable and reliable. By distributing traffic across multiple active nodes, the system can handle higher loads and provide redundancy in case of failures.

3.2 Blue/Green Deployment

Blue/Green deployment is a strategy where two identical environments are maintained. Traffic can be routed between either environment, allowing for seamless updates, testing, and rollback capabilities. Ideally while one of these environments is being updated with latest features, the other environment takes 100% traffic. This approach minimizes downtime and reduces the risk associated with deploying new features or updates.

4. Traffic Pattern Segmentation

Traffic pattern segmentation involves categorizing traffic into distinct patterns based on factors such as user behavior, geographic location, and time of access. This segmentation enables more precise and effective routing decisions, ensuring optimal performance and user experience.

4.1 Identifying Traffic Patterns

- **User Behavior Analysis**: Analyzing user interactions to identify patterns in browsing, purchasing, and engagement.

- **Channel-Based Segmentation:** Traffic is dynamically classified based on the origin channel—web, mobile apps, retail, or inbound. This enables channel-specific tuning of experiences and infrastructure.
- **Geographic Segmentation**: Routing users to regionally optimized environments to minimize latency, maximize responsiveness and optimize content delivery.
- **Temporal Segmentation**: Segmenting traffic based on time variables such as time of day, week, or season to manage peak loads effectively. This is a very effective during high traffic events like new product launches.

4.2 Segmented Routing

Segmented routing involves directing specific segments of a percentage of traffic to different server instances based on the identified traffic patterns. This approach ensures efficient utilization of resources and improves overall system performance.

4.2.1 Parameters of Segmented Routing

- **Traffic Distribution**: Distributing traffic across blue/green environments based on segmentation criteria.
- **Soak Period Stabilization**: Allowing the routed traffic to flow under the same specified parameters for a defined period-of-time, to study performance metrics and determine stability at full throttle.
- **Dynamic Adjustment**: Adjusting traffic routing dynamically in response to real-time conditions and performance metrics from soak period.
- **Failover Capabilities**: Ensuring seamless failover to alternate environments in case of server failures.

4.2.2 Mechanism for Segmented Routing

Segmented Routing Engine consists of

1. A traffic distribution configuration system
2. Multi Active environments or stacks that consist of the required infrastructure bundled together
3. An offline stack which is cycled through release cycle

Traffic distribution configuration can either be predictive, or a pre-defined value described in a config file. This value of the traffic distribution is determined based on previous traffic routing metrics and by a series of tests performed at peak traffic through the application.

Availability of **Multi Active environments** or stacks that consist of the required infrastructure bundled together, provides the ability to split traffic across all the available stacks carrying various features across these stacks and allowing for a distributed traffic pattern split across them. There should be a Blue/Green variant and a stand-by stack that contains the previous version of the code that is in production environments to successfully route traffic without live-traffic disruption.

A small amount of traffic can then be split across a canary test variant (green environment) with the latest feature and stable production (blue/green environment).

A specific soak-in period then allows the traffic to flow undisturbed. Traffic is monitored closely, revealing any faults. If no discrepancies are found, then the new feature is deployed across test variants which now become production variants, and entire traffic is split across them.

In case of a failure with the new feature deployment, the stand-by stack acts as a failover or a backup instance (on cloud or on prem) and can readily route all live consumer traffic.

An offline stack often maintained as a backup to the stand-by stack is cycled through the feature development cycle and maintained in offline mode always.

5. Application in Telecom eCommerce

At the heart of one of the leading Telecom tech giants in US's eCommerce innovation is Segmented Routing, a proprietary, scalable architecture that dynamically orchestrates traffic across a fleet of identical environments based on segmentation logic

The experimentations included:

- **Geographic Load Balancers** to identify user origin and distribute traffic to the nearest active region/cluster/datacenter.
- **Real-Time Customer Segmentation** to differentiate between prepaid, postpaid, enterprise, and consumer traffic for tailored experiences.
- **Feature Flag Systems with Segmented Routing**, where A/B tests of customer portals or feature releases are safely deployed to targeted segments via green environments.
- **Dynamic Resource Allocation:** Effectively utilizing resources to reduce latency and enhance the overall user experience.
- **Adaptive Real-Time Routing:** Constantly adjusting traffic flows in response to live metrics and performance insights.
- **High Availability:** An active-active configuration across environments ensuring seamless failover and zero-downtime deployments.

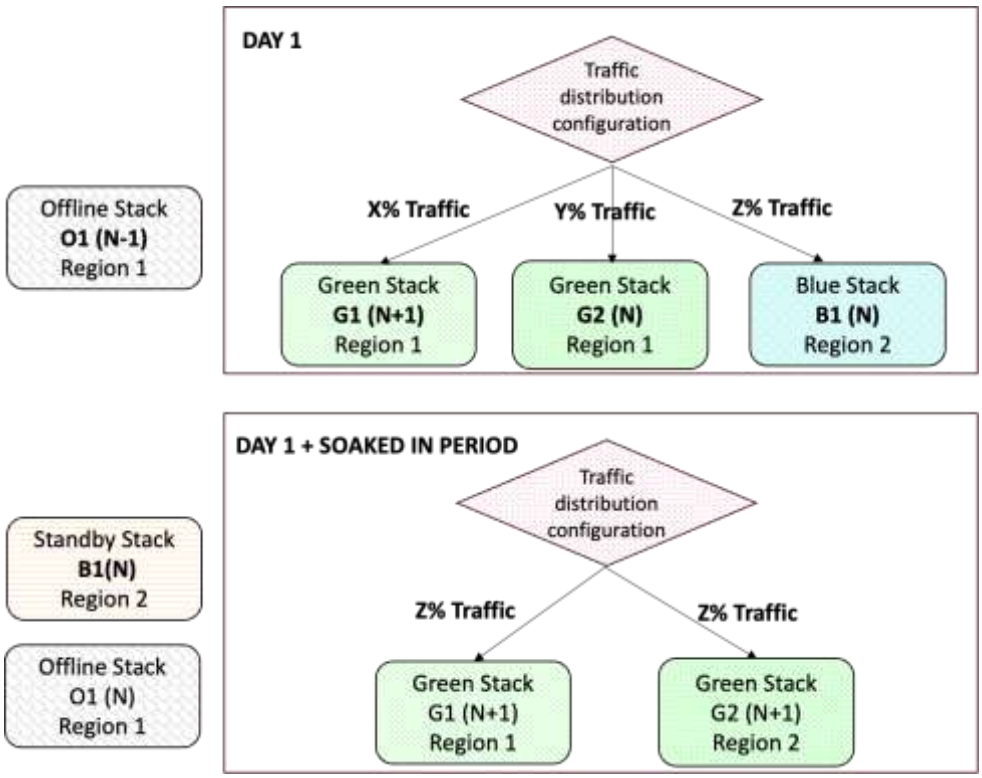

Figure 1: Example of a Segmented Routing Deployment Setup

5.1. Implementation

Implementing traffic pattern segmentation and segmented routing in eCommerce applications involves several key steps:

5.1.1 Data Collection and Analysis

Collecting and analyzing data to identify traffic patterns and segmentation criteria. This involves gathering user interaction data, geographic information, and product access patterns.

5.1.2 Infrastructure Setup

Deploying an active-active infrastructure with blue/green environments to support segmented routing. This setup requires configuring multiple identical environments and ensuring they are synchronized. Also, this setup would need access to a global data set that is maintained across these multiple identical stacks to maintain the state of data.

5.1.3 Routing Configuration

Routing configuration is critical in ensuring that segmented traffic is effectively routed to the correct environments. Companies typically use tools such as AWS Route 53**,** NGINX**,** HAProxy**,** or Istio in Kubernetes environments to implement advanced traffic routing rules. For instance, routing rules may be based on headers (e.g., user agents), geo-location, or session attributes.

- **Channel-Based:** Traffic is dynamically classified based on the origin channel—web, mobile apps, retail, or inbound. This enables channel-specific tuning of experiences and infrastructure.
- **Geographic Based:** Routing users to regionally optimized environments to minimize latency and maximize responsiveness.
- **Session Based:** Keep users in the same segment during their session to preserve state.

These routing mechanisms often include automated failover capabilities to redirect traffic in case of failures in a particular environment.

5.1.4 Monitoring and Optimization

Continuous monitoring is essential for segmented routing and traffic pattern segmentation. Effective implementations rely on:

- **Observability Pipelines** aggregate real-time telemetry, facilitating per-segment performance comparisons including Application Performance Monitoring using tools like *Datadog*, *New Relic*, *AppDynamics*.
- **Real-Time Analytics** for behavioral data segmentation.
- **Feedback Loops** from previous deployments and Customer needs feed directly into routing decisions, enhancing system reactivity. Example - Incorporating A/B test results and performance metrics at Peak into the segmentation model via automation and machine learning.

These tools allowed to optimize routing logic based on latency, error rates, and user experience data.

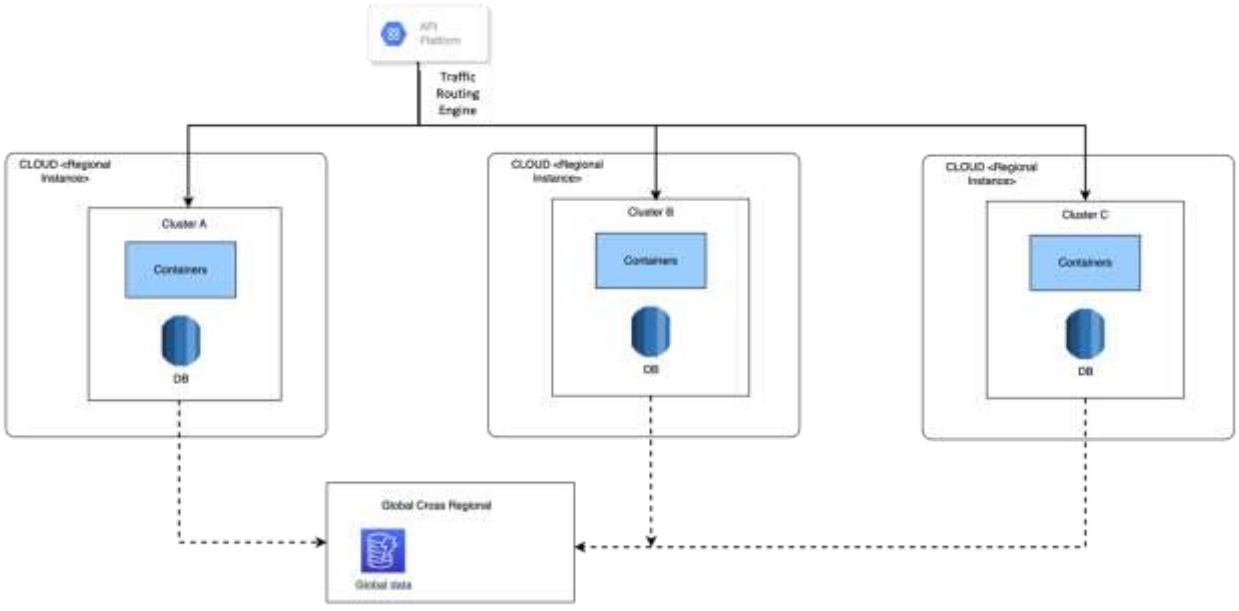

Figure 2: Example of a Segmented Traffic Routing Infrastructure Setup

6. Driving Real-Time Innovation: Segmented Routing for A/B Testing

- Through segmented routing, the telecom giant enhanced real-time A/B testing capabilities which enables to
  - **Deploy Updates Seamlessly**: Execute upgrades and introduce new features without impacting user availability.
  - **Real-Time Observability**: Quickly collect real-time performance metrics and user feedback to refine strategies.
  - **Instant Rollback Options:** Swiftly revert to stable environments, mitigating risk and minimizing disruption.

7. Optimizing E-Commerce with Segmented Traffic

By adopting segmented routing and ephemeral environments powered by Elastic Path's composable infrastructure, Telecom giant achieved remarkable outcomes:

- **100% Availability:** Ensured uninterrupted service delivery, even under high traffic conditions with zero downtime.
- **Instant Rollback**: Switching DNS or load balancer pointers aka the traffic, enables immediate reversion.
- **Enhanced Customer Engagement:** Improved experiences and increased engagement driven by precise, targeted A/B testing.

Using segmented routing of traffic, companies can:

- Collect direct user feedback segmented by demographic, device, or behavior.
- Use test outcomes to train ML models for future routing decisions (e.g., dynamic personalization).

7.1 Benefits

- **Performance Optimization**: Run performance experiments at scale (e.g., layout changes, pricing tests) and tailor routing strategies, improving load times and reducing bounce rates.
- **Scalability**: Active-active setups can absorb surges and scale globally.
- **Early Feedback Loop:** Collect direct user feedback segmented by demographic, device, or behavior.
- **Business Agility**: Rapid experimentation through segmented deployments and use test outcomes to train Machine Learning models for future routing decisions (e.g., dynamic personalization**)**

8. Conclusion

Traffic pattern segmentation, when combined with segmented routing and blue-green deployment, is a transformative strategy for large-scale eCommerce businesses. It enables personalized user experiences, real-time experimentation, and operational resilience. By leveraging advanced infrastructure and analytics, companies can dynamically respond to customer needs, ensuring performance, reliability, and innovation.

As AI and edge computing mature, the next frontier includes *predictive routing,* where traffic is preemptively segmented based on AI-driven forecasts. Large eCommerce engines like Shopify, Elastic Path and tech companies using continue to innovate towards AI-driven Predictive Routing, powered by historical usage patterns and real-time analytics.

Self-Healing systems, failure detection and auto-remediation to eliminate manual interventions and improve service level objective adherences are setting the stage for even more intelligent and proactive digital commerce platforms.

References

[1] Rahul Padhye, Anand Ganapathy, Sadique Ahmad. "Ensuring availability and integrity of a database across geographical regions". patents.google.com. https://patents.google.com/patent/US1168751982 (accessed Jun 28, 2023).

[2] G. John, A. Mui, V. Vlasceanu, M. Mansoor. "Blue/Green Deployments on AWS". docs.aws.amazon.com. https://docs.aws.amazon.com/whitepapers/latest/blue-green-deployments/blue-green-deployments.pdf (accessed Mar 23, 2023).

[3] "Optimizing E-Commerce with Customer Segmentation and Targeted Marketing - Beam Data". beamdata.ai .https://beamdata.ai/case-study/samsung-optimizing-e-commerce-with-customer-segmentation-and-targeted-marketing/ (accessed Apr 21, 2023)

[4] "Blue-Green Deployment for Zero Downtime: Seamless Ecommerce". globaltechnosol.com . https://globaltechnosol.com/case_studies/blue-green-deployment-for-zero-downtime-releases/ (accessed Apr 29, 2023)