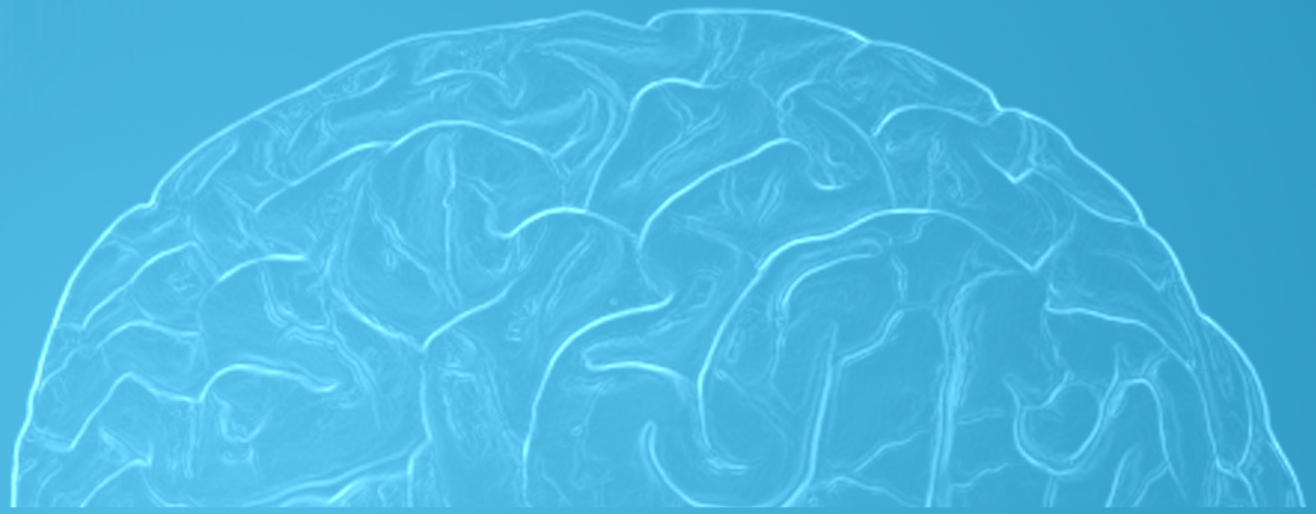
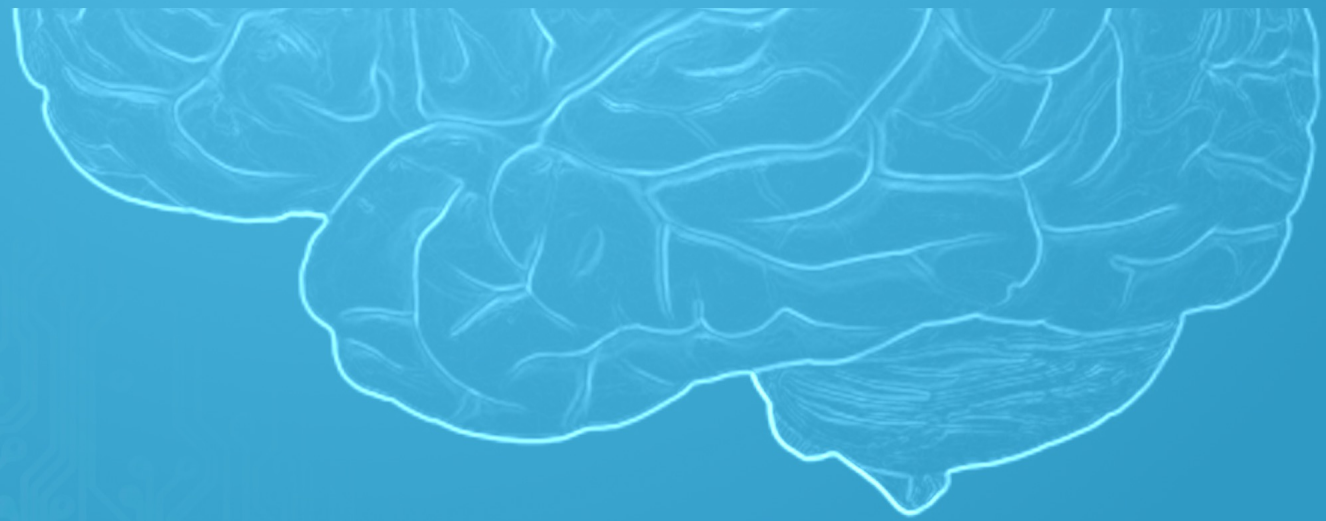


Sisteme Tolerante la Defecte
Experiență practică

Alexandra Tudor
SRE la Adobe Systems Romania



Lect. Dr. Ing. Cristian Chilipirea
cristian.chilipirea@mta.ro

Despre mine - Alexandra Tudor

- Universitatea Politehnica, Bucuresti
  - Facultatea de Automatica si Calculatoare, sectia Calculatoare
  - 2012 - 2016

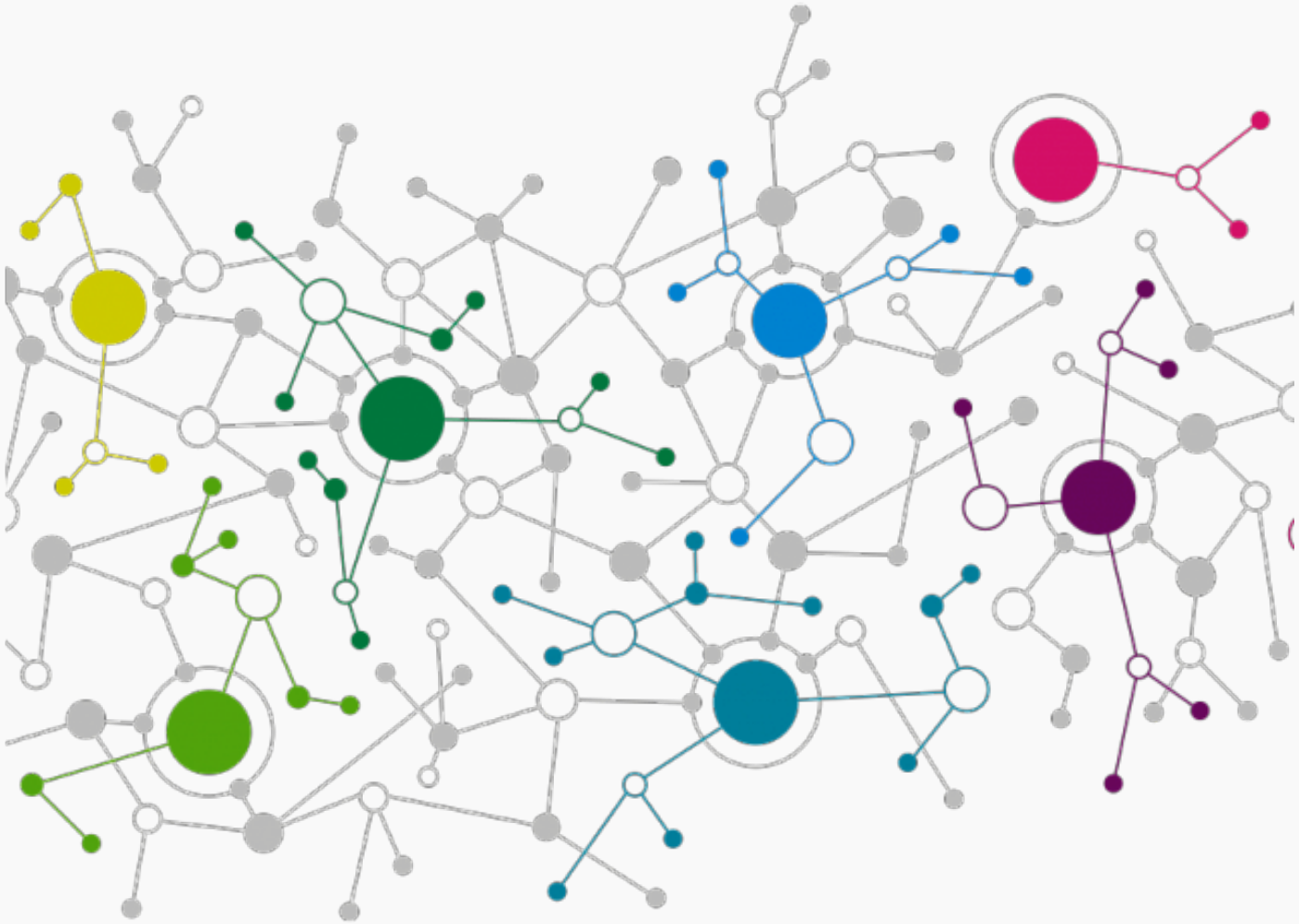
- Adobe Systems Romania, Bucuresti
  - Site Reliability Engineer
  - 2017 - Prezent

Adobe Systems

Creative Cloud	Document Cloud	Experience Cloud
Creativity and Design	PDF and electronic signatures	Marketing and commerce

https://www.adobe.com/ro/

Experience Cloud



Availiability

Latency

Performance

Security

Maintainability

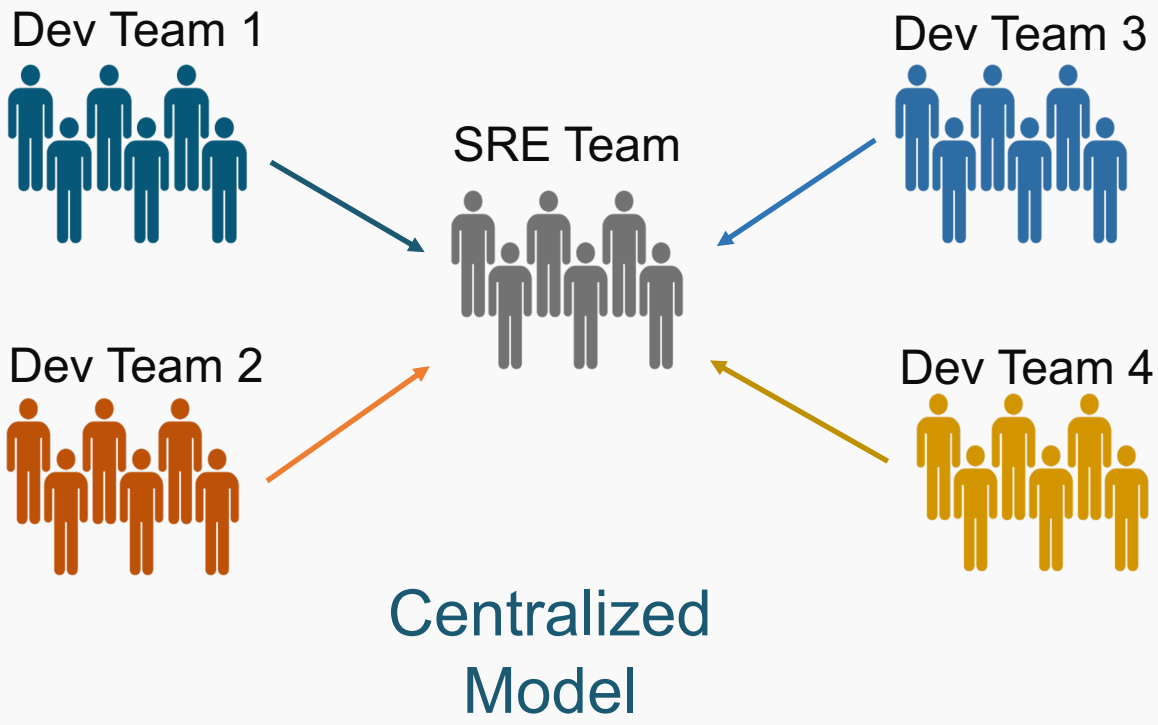
Site Reliability Engineer

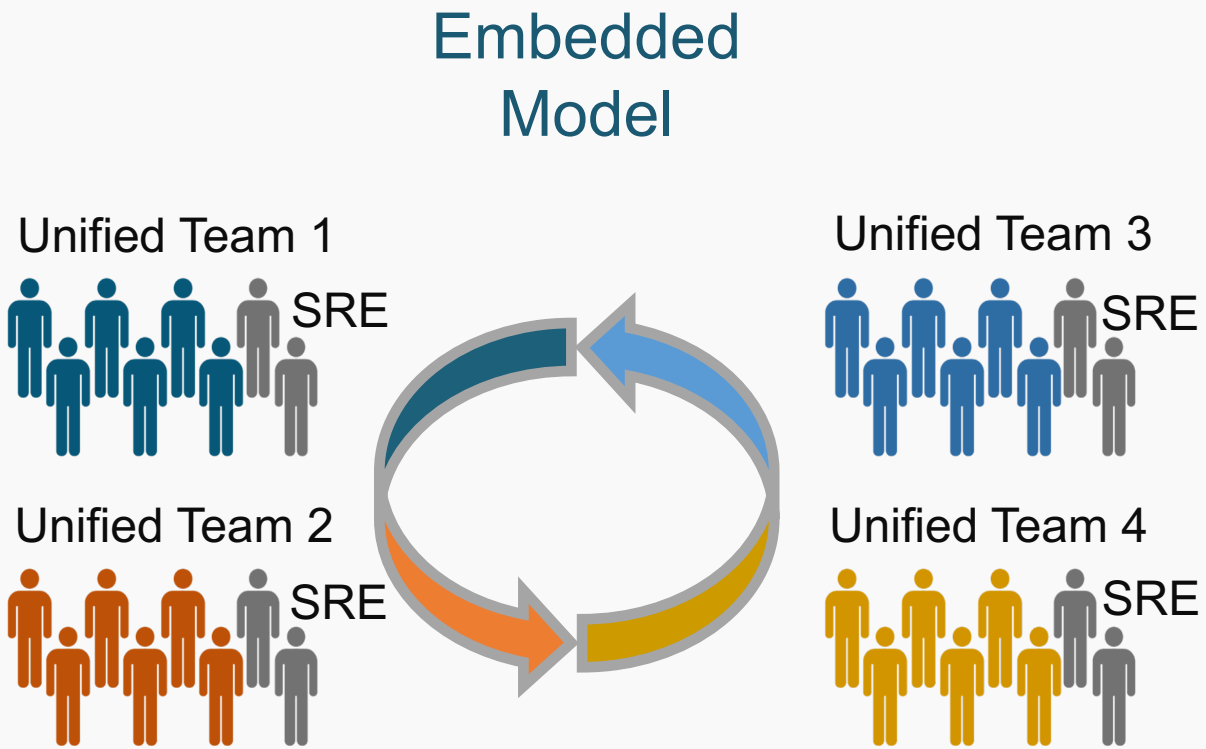
- SRE concept born at Google in 2003
- Software Engineer Team
- Make Google's already large-scale sites more reliable, efficient, and scalable
- Later adopted by **Amazon**, **Netflix**… **Adobe**

"Site reliability engineers create a bridge between development and operations by applying a software engineering mindset to system administration topics."

- Books - https://sre.google/books/

SRE Teams


Dev Team 1
Dev Team 3
SRE Team
Dev Team 2
Dev Team 4
Centralized Model


Embedded Model
Unified Team 1
SRE
Unified Team 3
SRE
Unified Team 2
SRE
Unified Team 4
SRE

Embedded SRE Responsibilities

- Monitoring
- Defining SLIs / SLTs / SLAs
- Alerting
- On Call
- Capacity Planning
- Operational Readiness
- Change Management

Monitoring

- Number of incoming/outgoing requests
- Error count
- Request latency
- Database query latency
- Queue lag
- File descriptors
- Threads
- CPU/Memory/Disk usage
- …

Prometheus

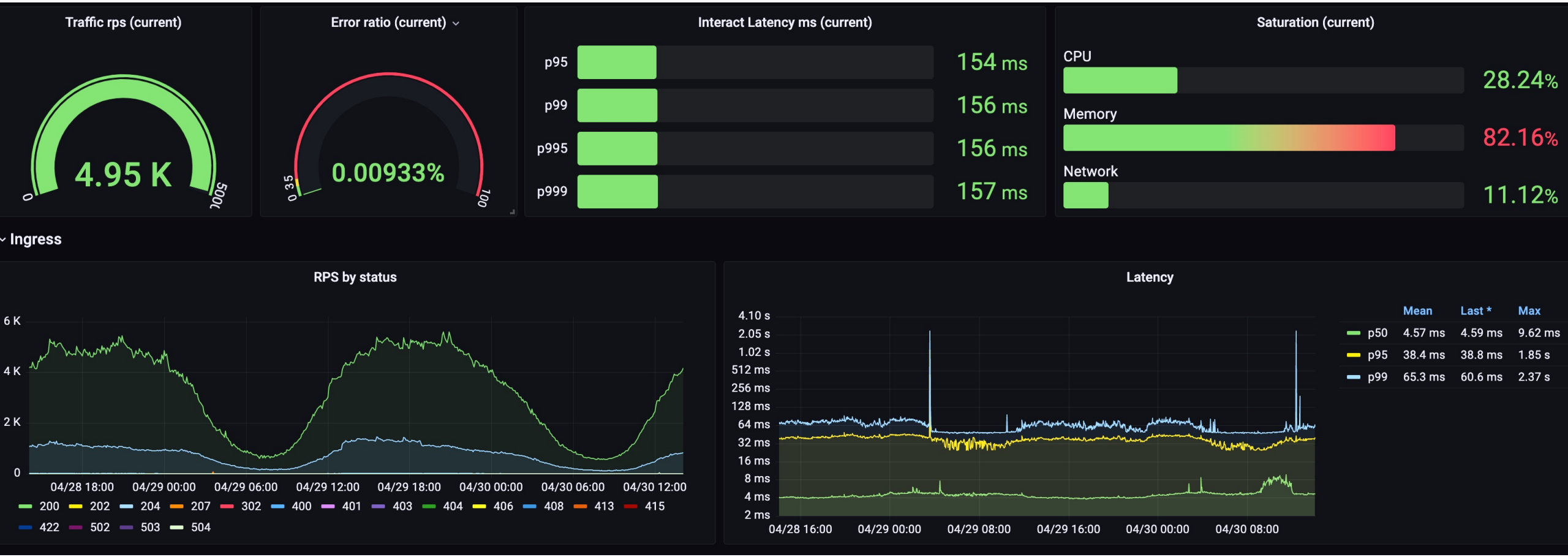
Grafana

Monitoring

Too many metrics?

- Google 4 golden signals
  - Latency, Traffic, Errors, Saturation
- RED Method
  - Request Rate, Error, Duration

Monitoring Example



Why Monitoring

- Alerting
- Debugging
- Scaling (Autoscaling)
  - CPU, Memory
  - Queue Lag
- Analyzing long term trends (High tide, Low tide)
- …

SLIs drive **SLTs** which inform **SLAs**

SLI

Service Level Indicator

Request latency

Failures per second

95th percentile latency of homepage requests over past 5 minutes < 300 ms

SLT or SLO

Service Level Target

Binding target for a collection of SLIs

95th percentile homepage SLI will succeed 99.9% over trailing year

SLA

Service Level Agreement

Business agreement between Adobe and the customer based on the defined SLTs

Service credits if 95th percentile homepage SLI succeeds less than 99.5% over trailing month

Error Budget

SLI

95th percentile latency of homepage requests over past minutes < 300 ms

SLT

95th percentile homepage SLI will succeed **99.9%** over trailing month

SLA

Service credits if 95th percentile homepage SLI succeeds less than **99.5%** over trailing month

Error Budget Example

100% - rps{http_response_code=~"5[0-9]+"} / rps * 100

Alerting

- Based on SLTs

```
- alert: "Availability"
  annotations:
    message: 'Availability measured at the load balancer is {{"{{$value}}"}}.'
    runbook_url: https://.../developer/runbooks?id=availability_sop
  expr: 'availability{env="prod"} < 0.99'
  for: 5m
  labels:
    severity: CRITICAL
```

- Alert classification
  - **Critical – Incident – Phone Call**
  - **Warning - SMS**
  - **Info - email**

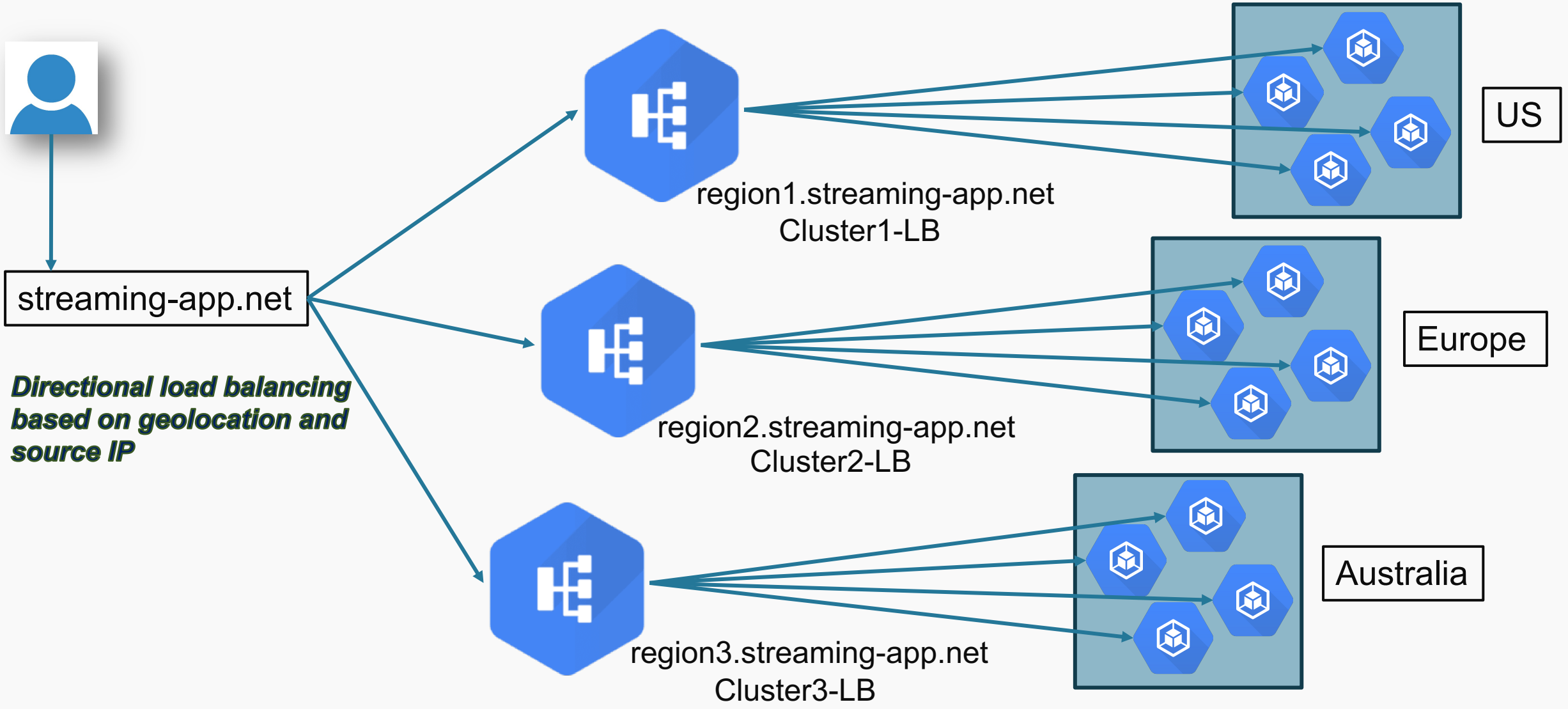
On Call

- On-Call model
  - Introducing the **On-Call** & **On-Duty** Engineers
  - MTTD, MTTR, MTBF

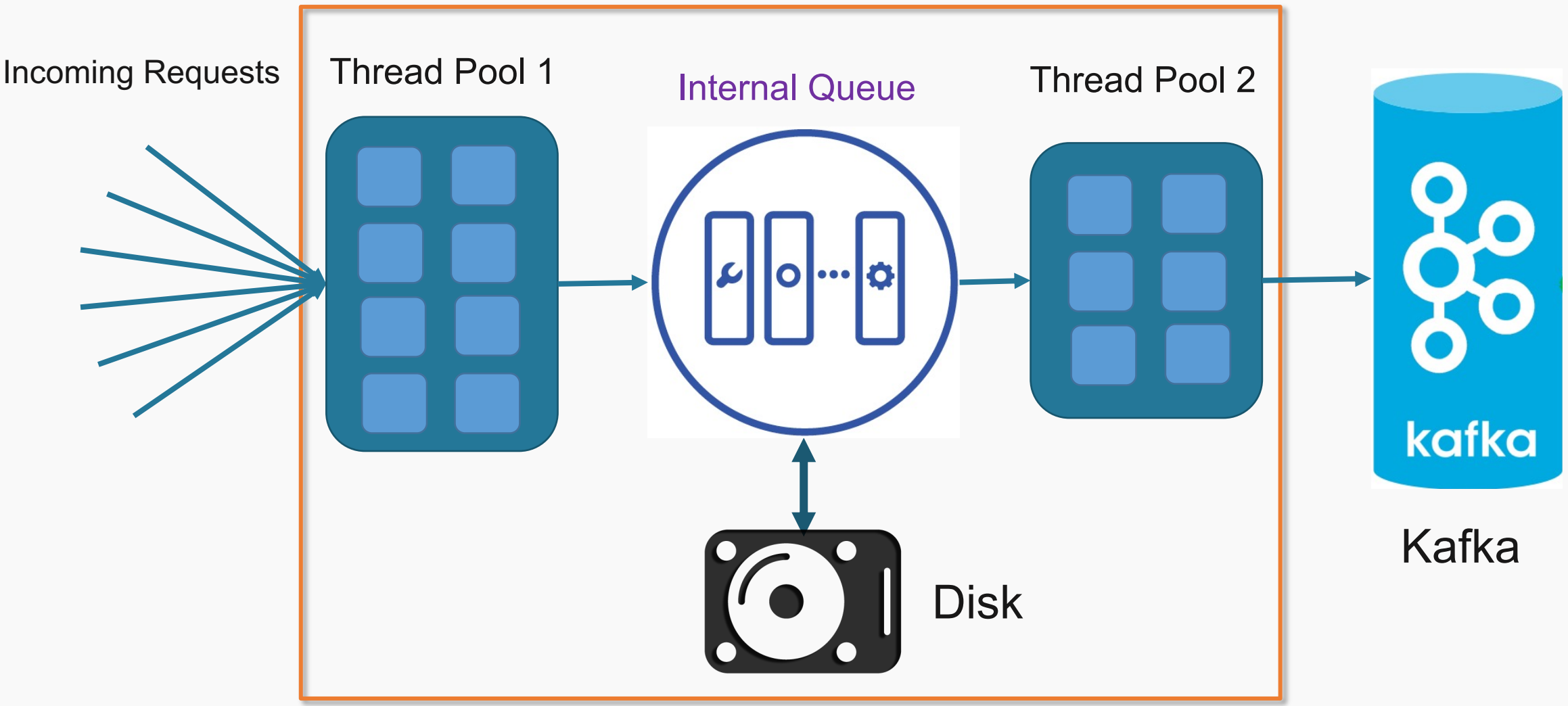
- Alerts and SOPs (Standard Operating Procedures)

- Postmortem process
  - Action Items

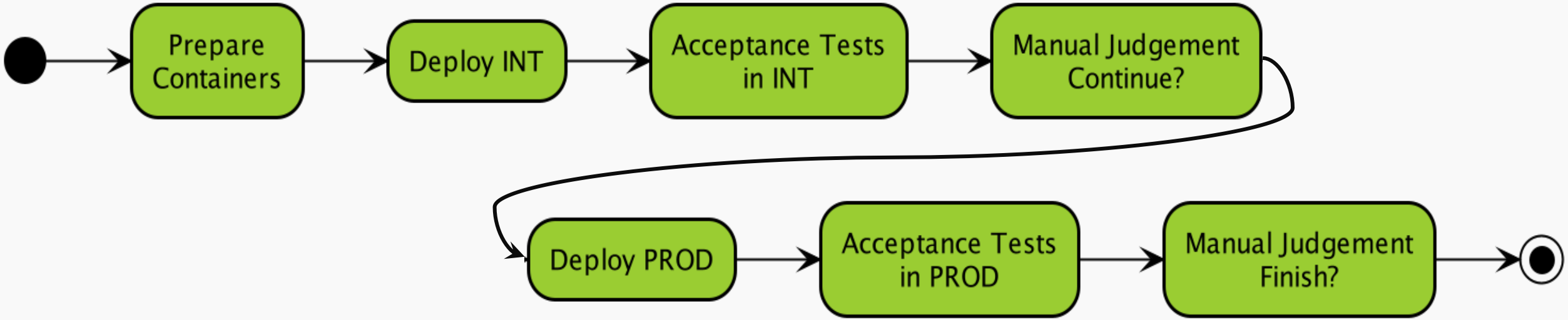
Streaming App – high level overview



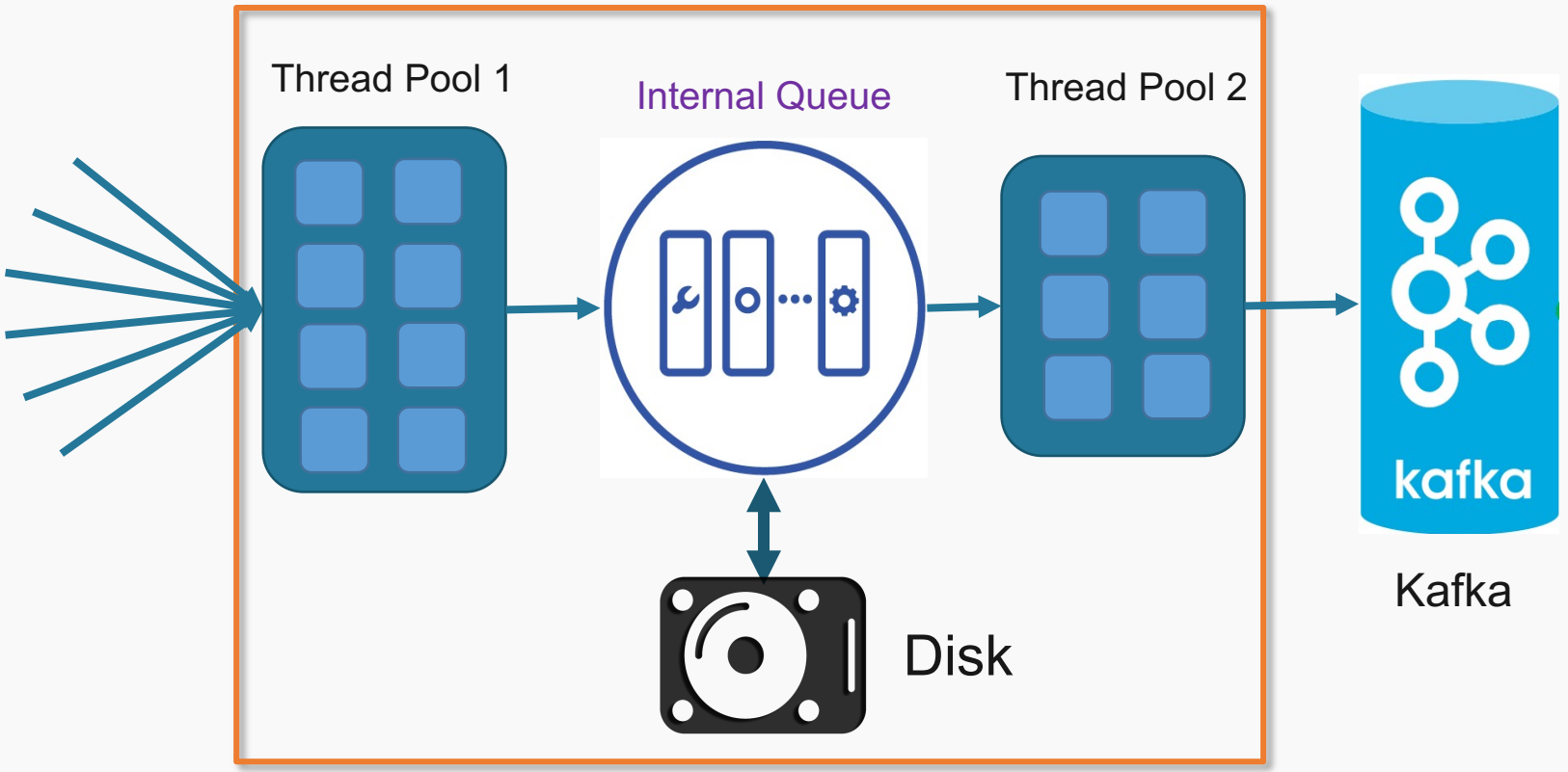
Streaming App – zooming in



Streaming App – Release Day



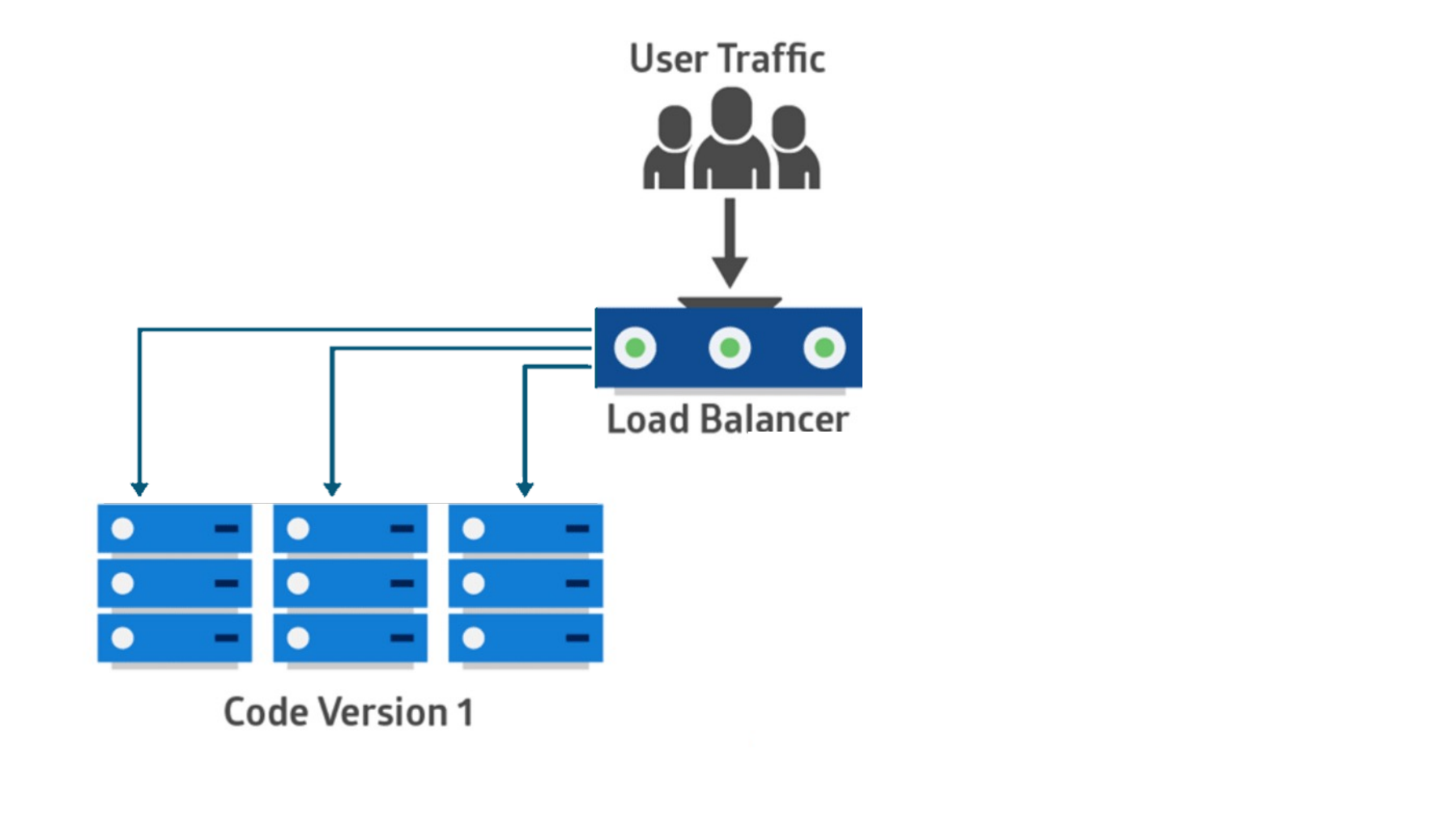
Streaming App – Troubleshooting



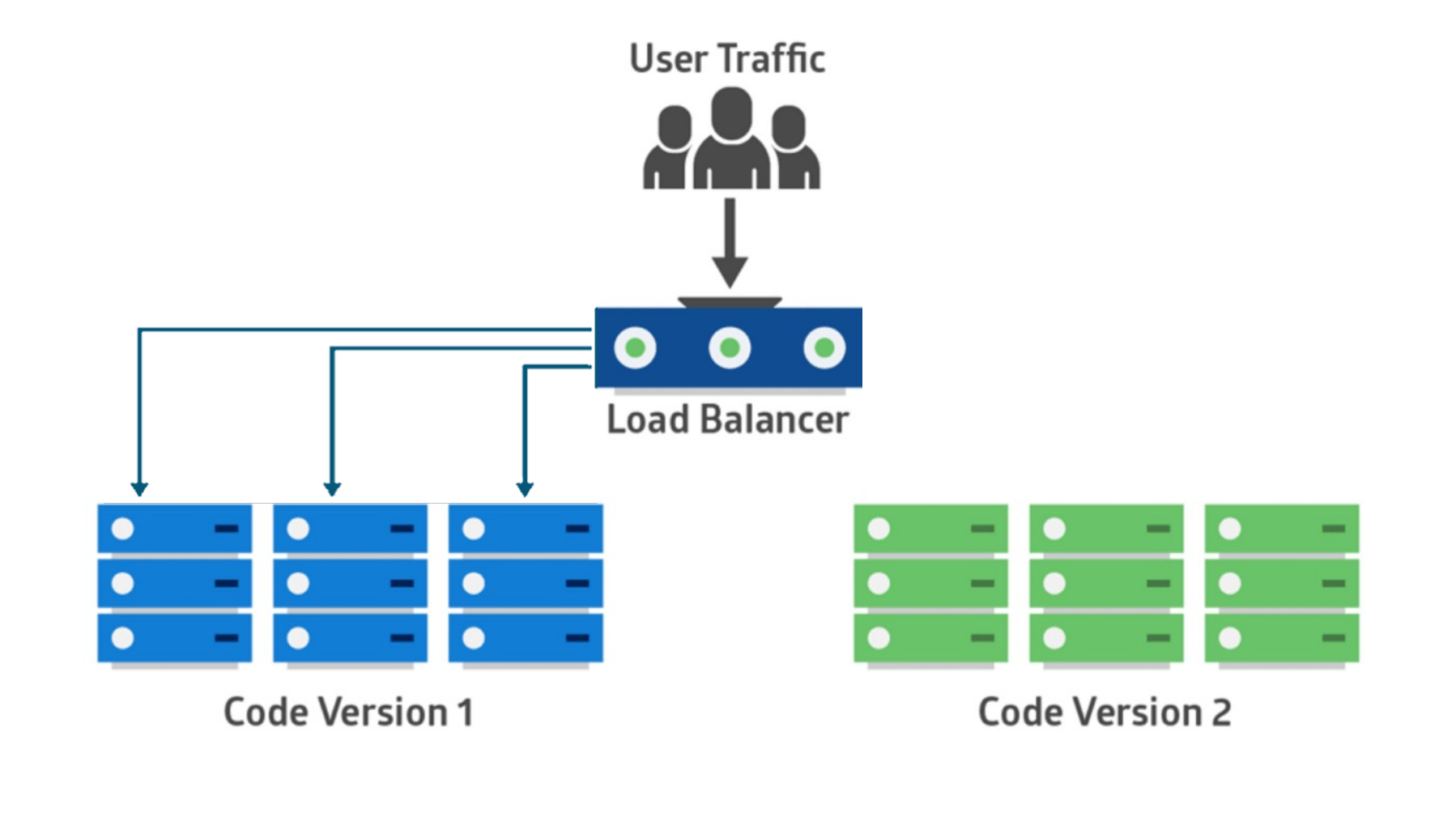
Alert: Publish to Kafka Lag in Europe Region

Number of messages published to Kafka / number of messages published to Internal Queue * 100 < 80 over 5 min

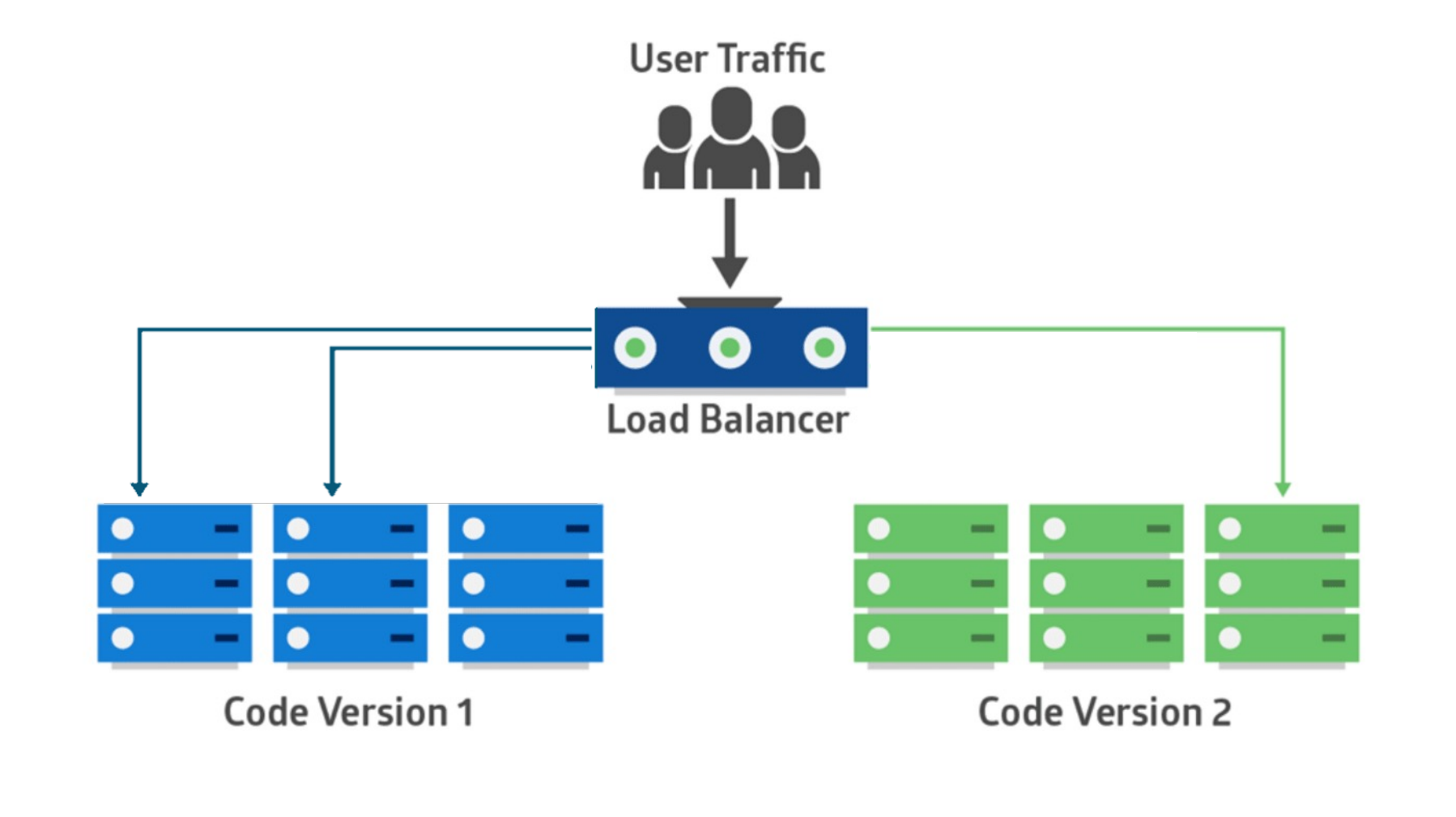
Streaming App – Blue Green Deployment



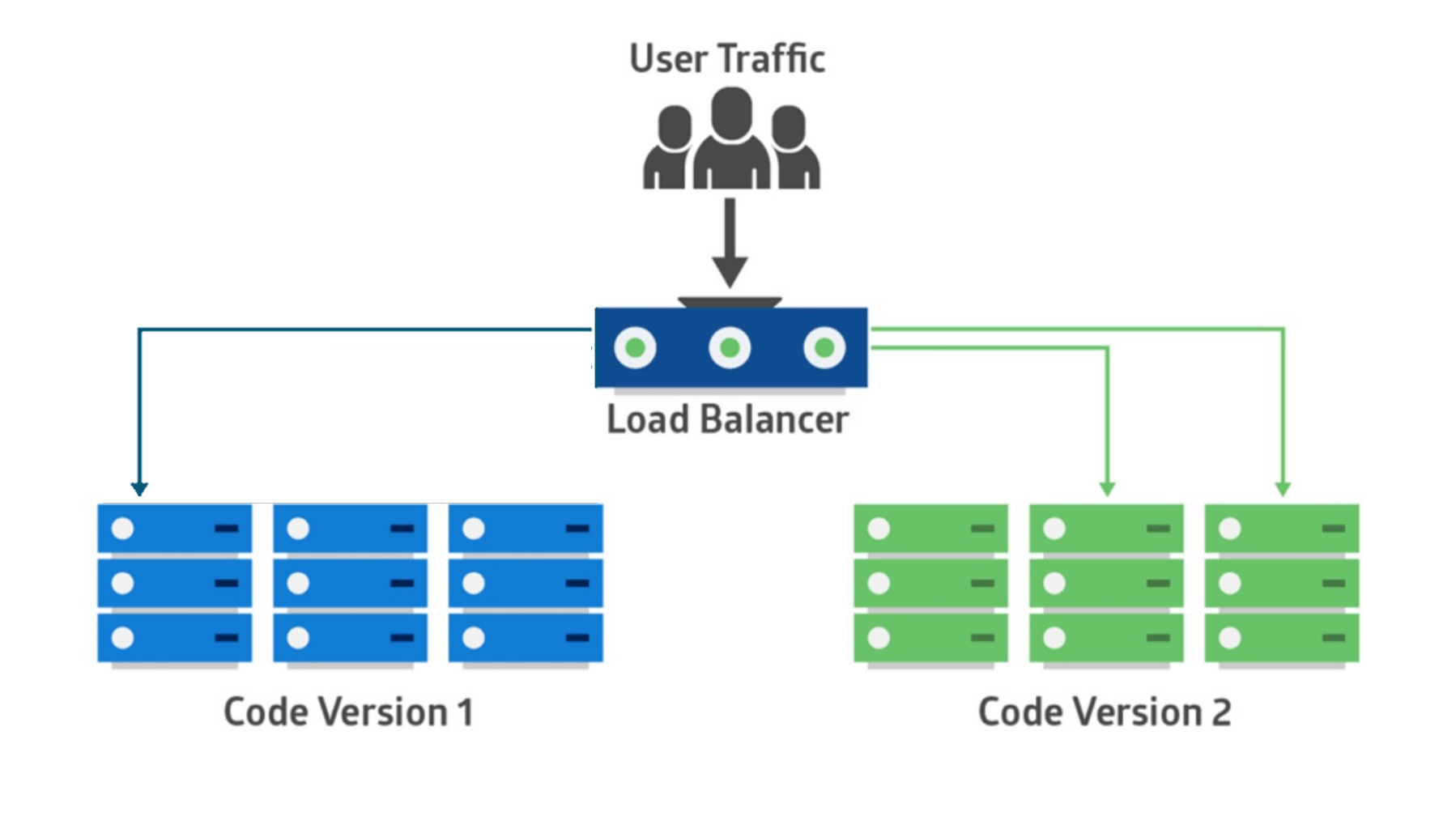
Streaming App – Blue Green Deployment



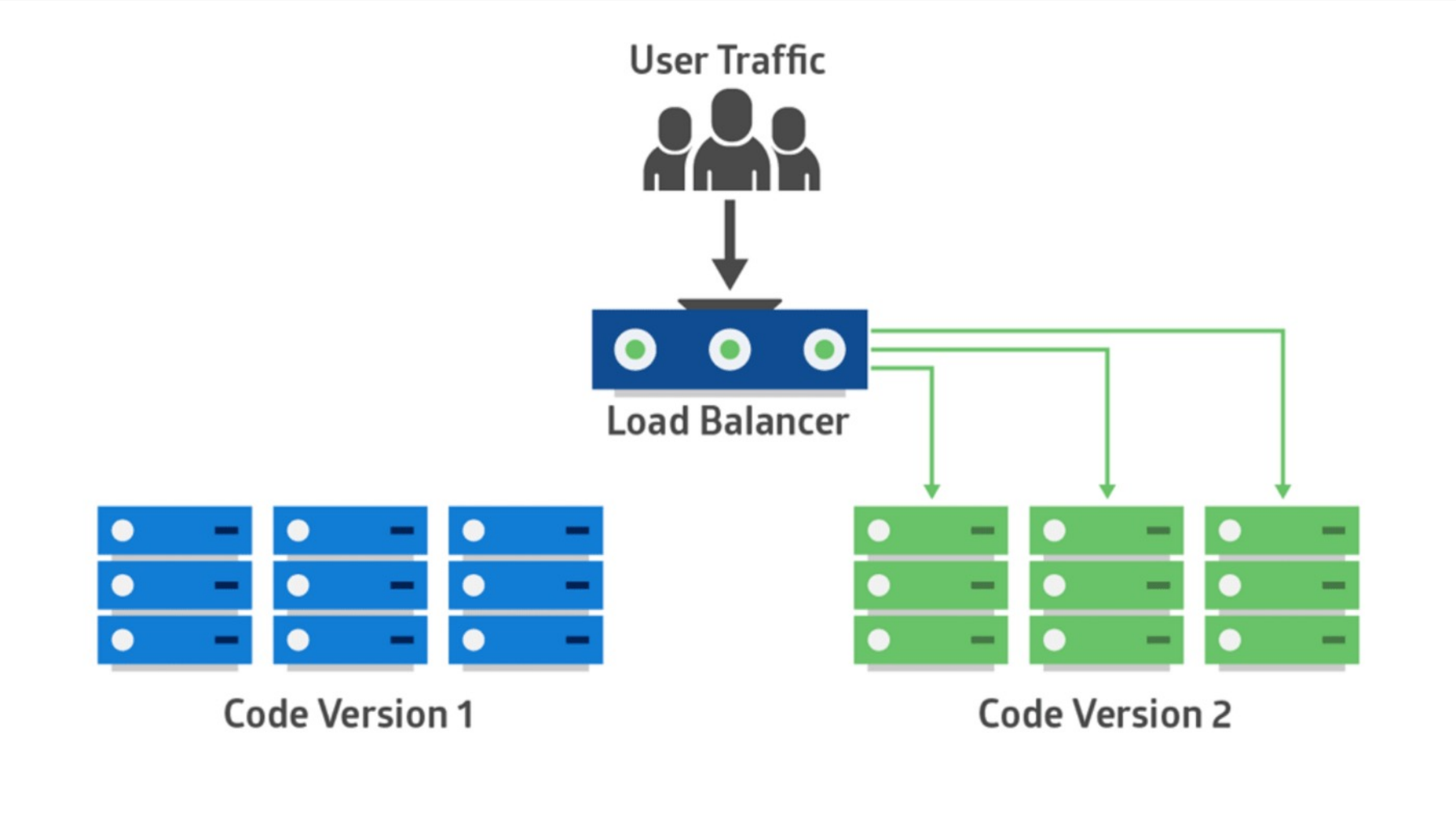
Streaming App – Blue Green Deployment



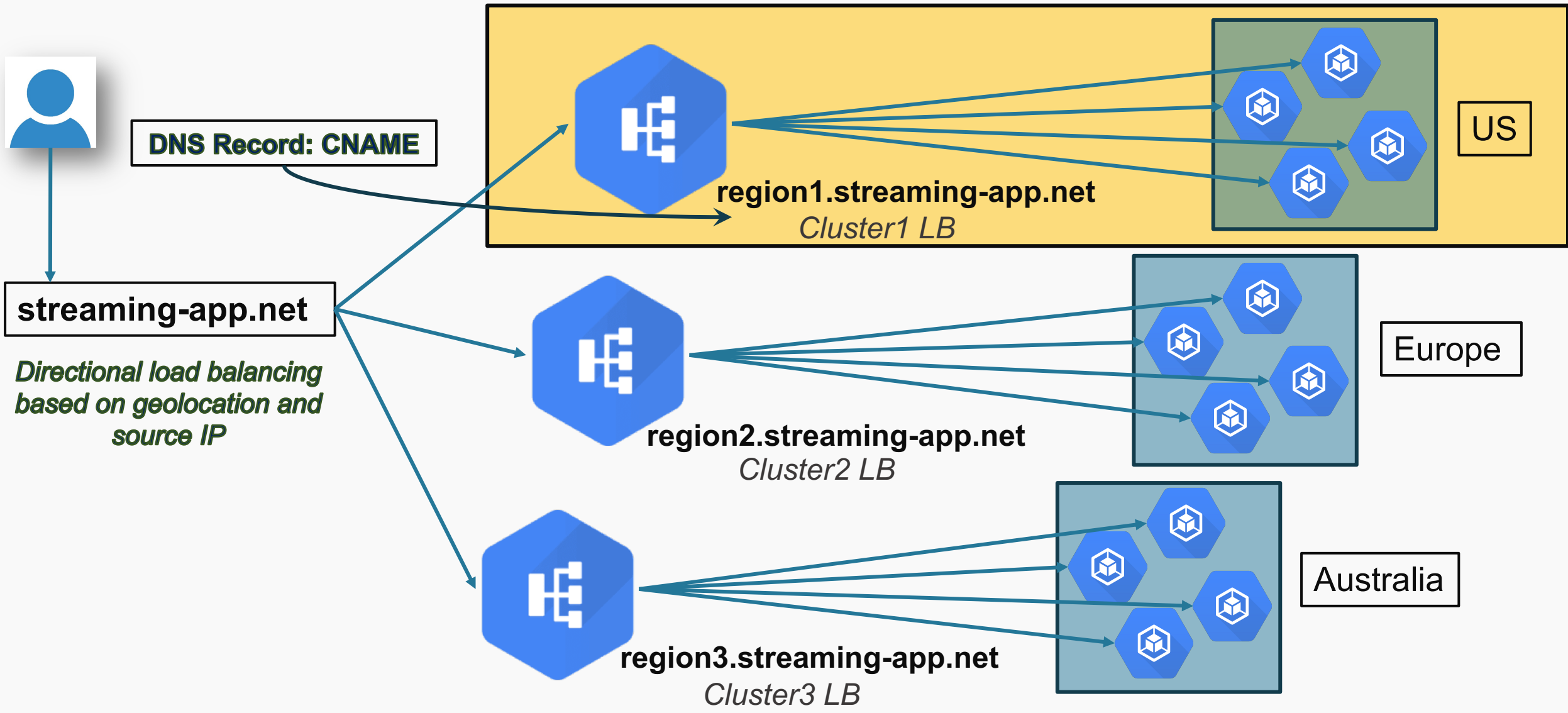
Streaming App – Blue Green Deployment



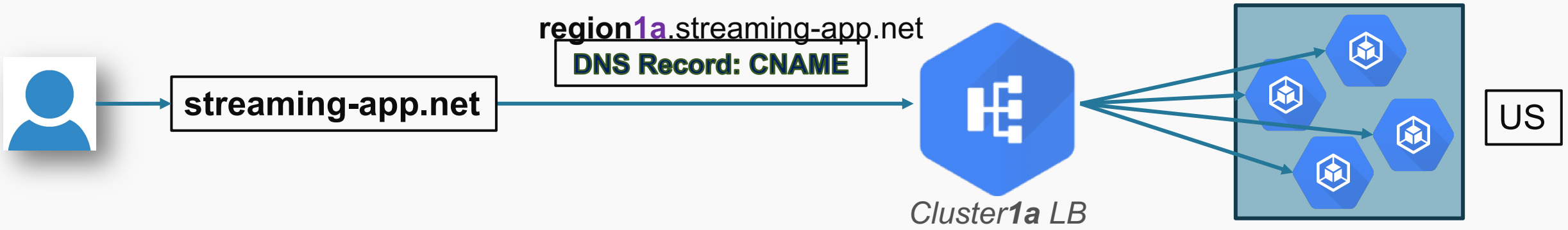
Streaming App – Blue Green Deployment



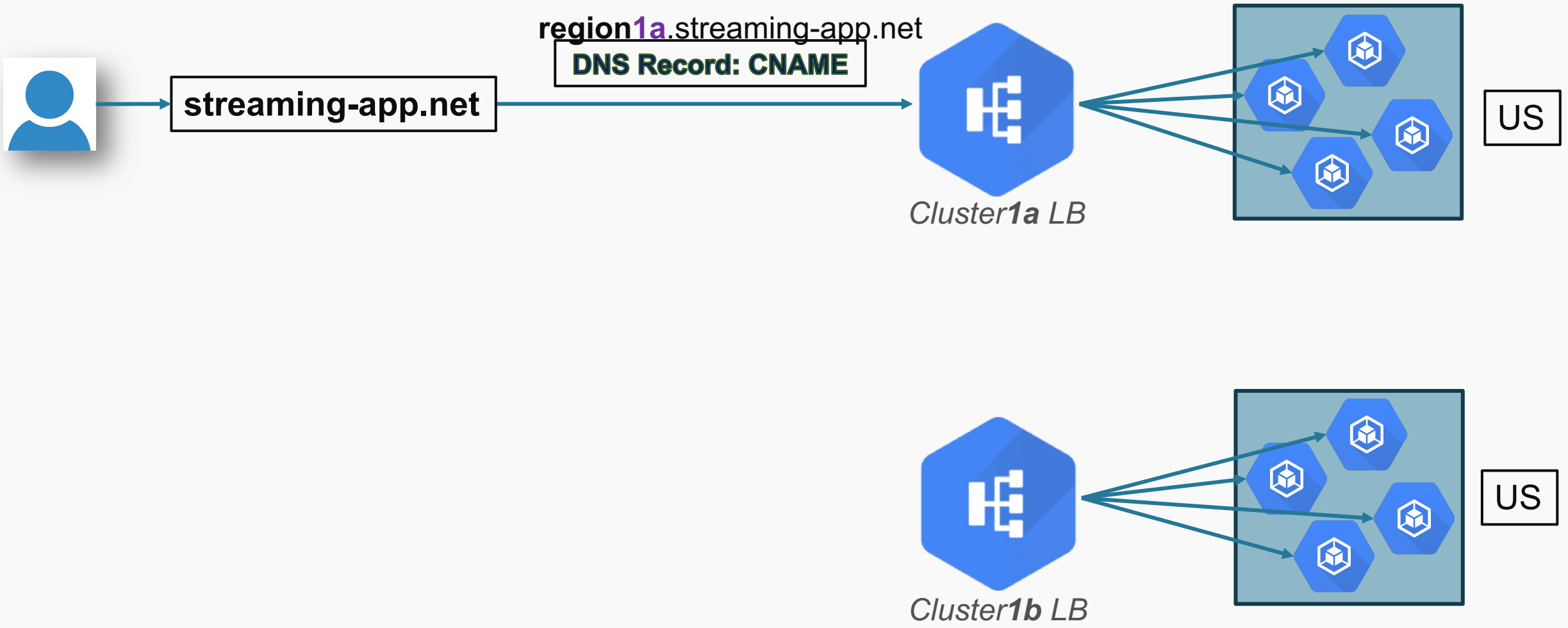
Service migration – slowed down due to DNS caches



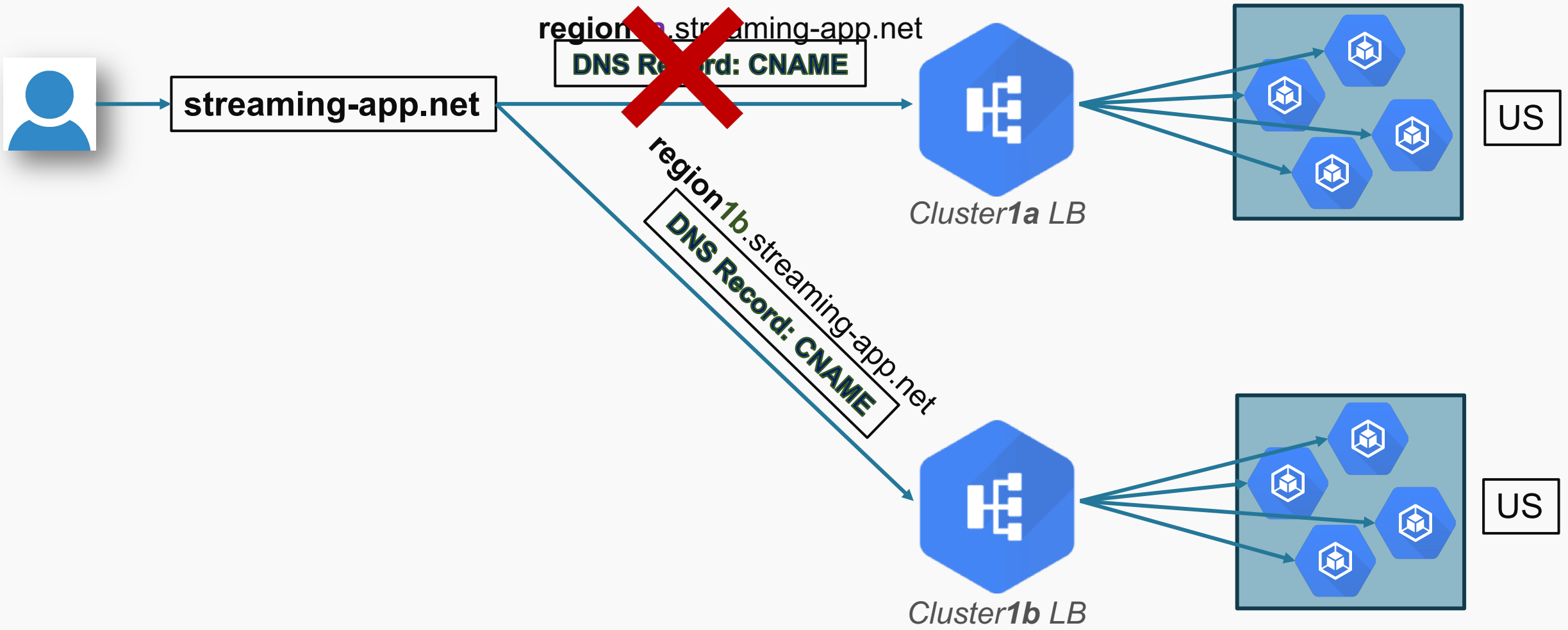
Service migration – slowed down due to DNS caches



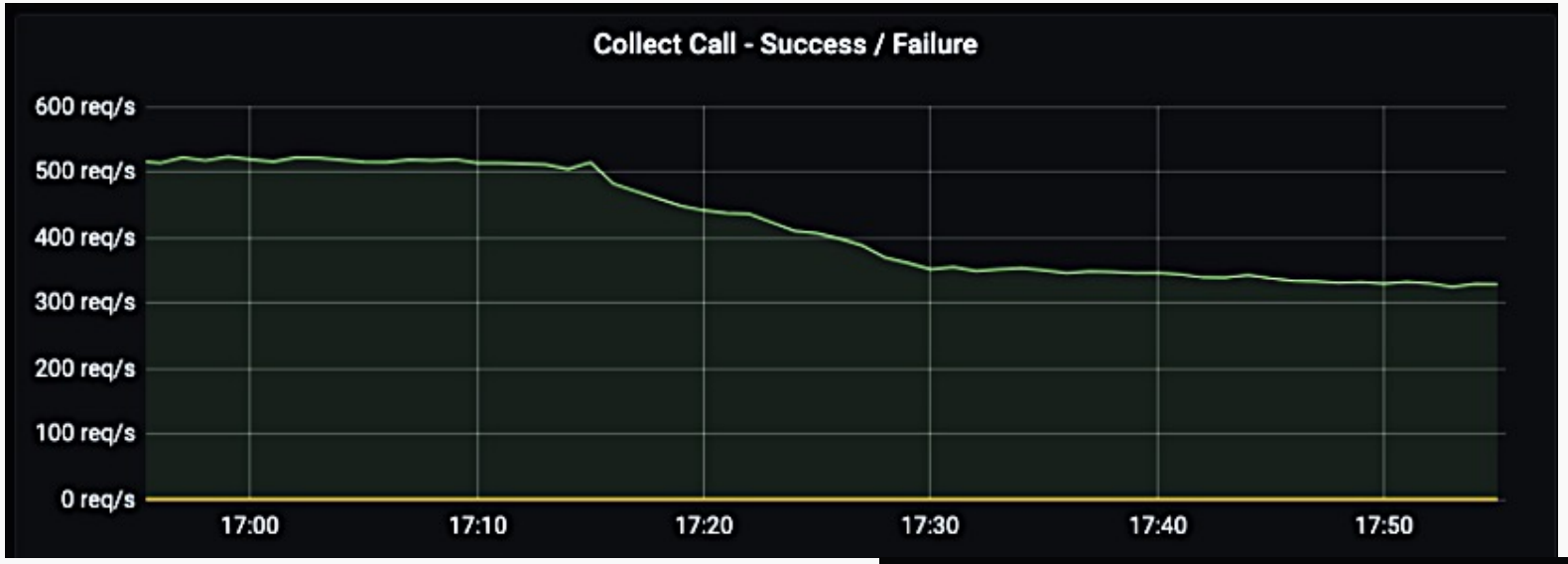
Service migration – slowed down due to DNS caches



Service migration – slowed down due to DNS caches



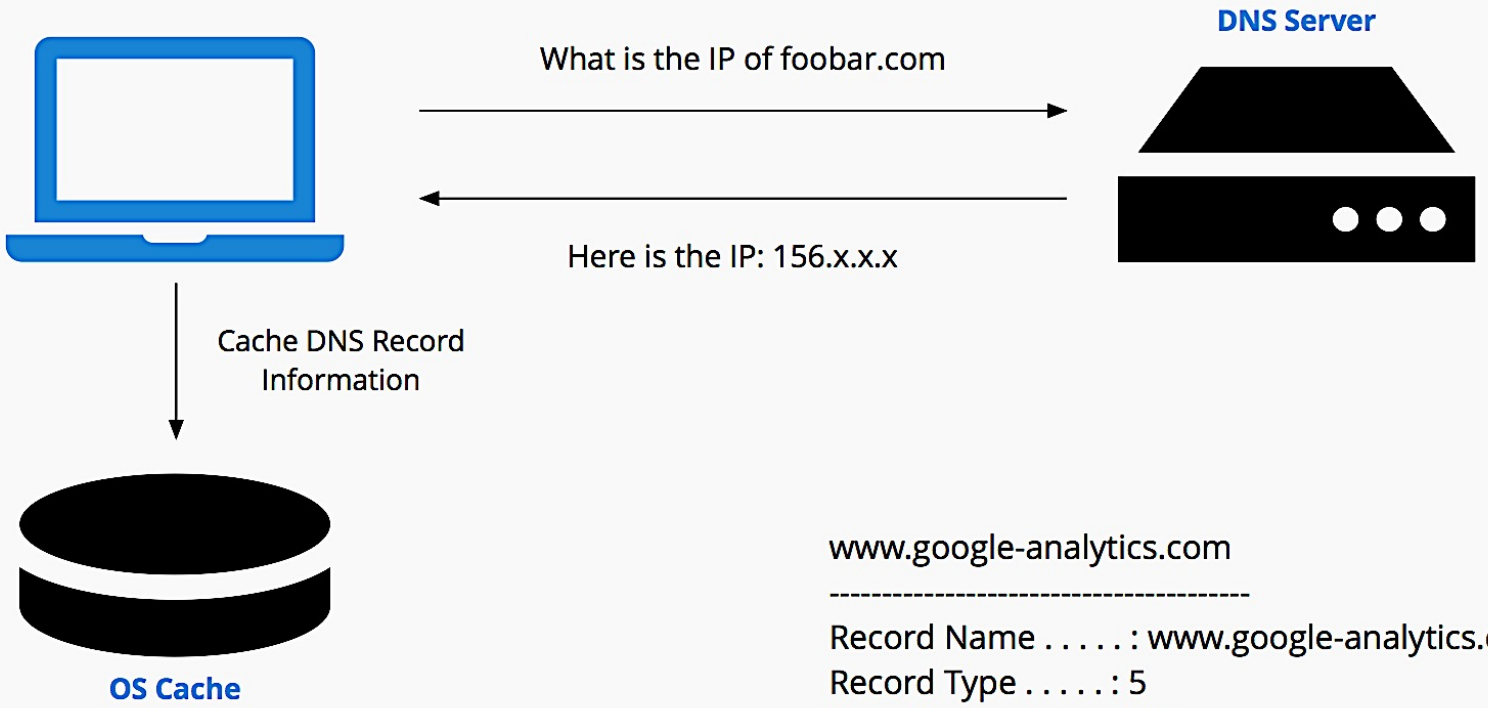
Service migration – slowed down due to DNS caches



Contrary to the expectation, not all requests were immediately routed to the new deployment

What about a revert at this point?

Service migration – slowed down due to DNS caches



DNS Cache

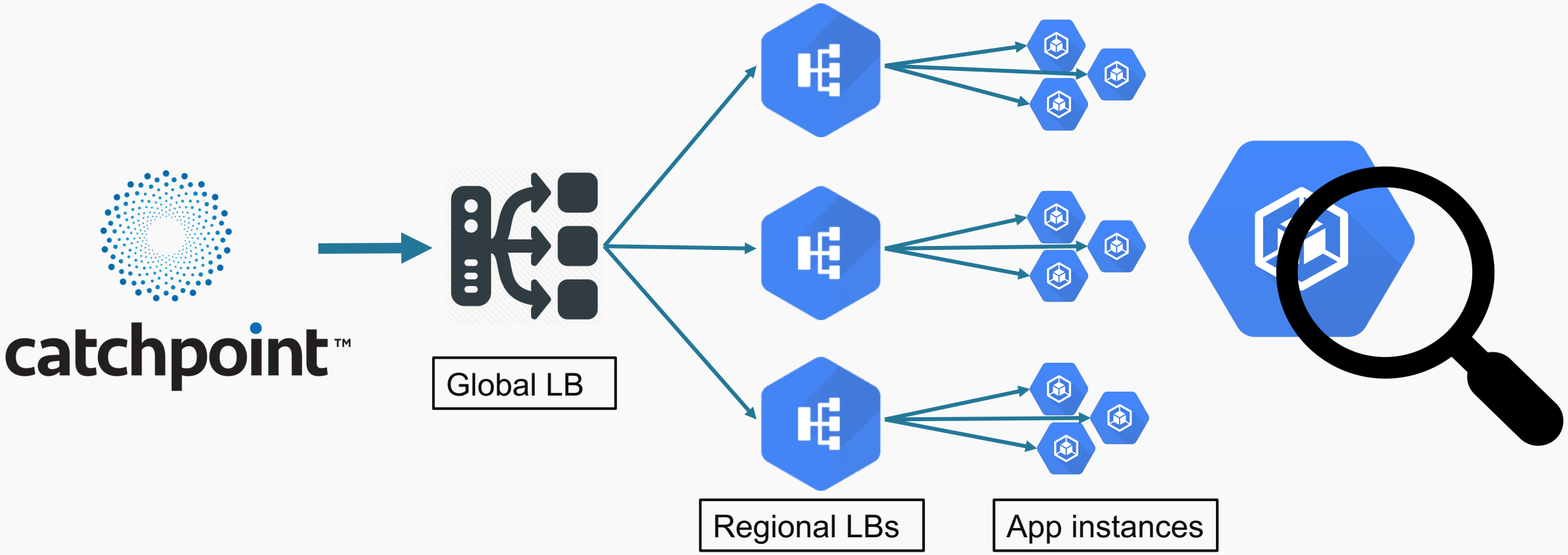
www.google-analytics.com

Record Name . . . . . : www.google-analytics.com
Record Type . . . . . : 5
Time To Live . . . . : 104
Data Length . . . . . : 4
Section . . . . . : Answer
CNAME Record . . . . : www-google-analytics.l.google.com

Blackbox monitoring

Why/when do we need blackbox monitoring?

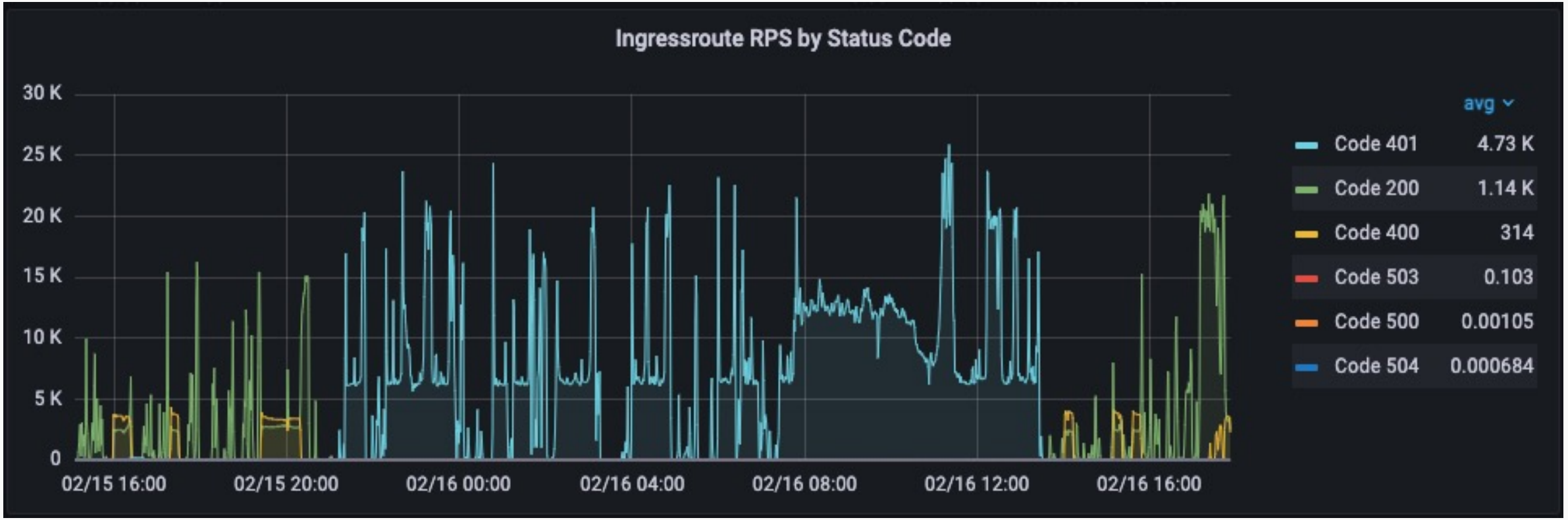
- The application works very well even if it doesn't receive any requests, or even better



A user error is not always user's fault

Sometimes we need to monitor the 4xx responses too

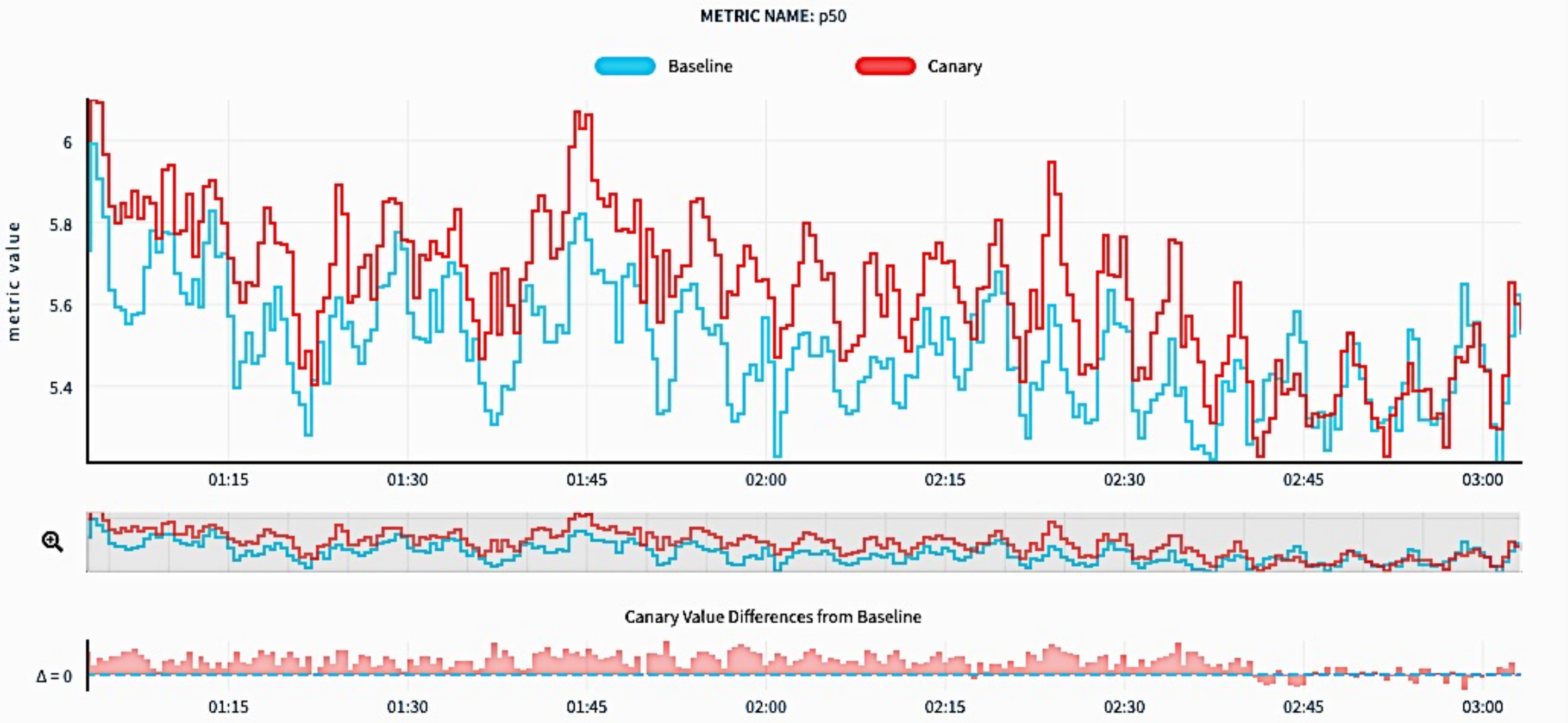
- 401 Https Status Code - Unauthorized



Canary Deployment

- Deploy 2 new pods:
  - **Baseline** (current version)
  - **Canary** (new version)
- Analyze the differences between the key metrics over a longer period of time – 2-3 hours
- Decide whether to go forward and do a full deploy or stop the process and fix the issues causing the degradation

Canary Analysis



Adobe Experience Platform - Technology Landscape