aws

COMMUNITY DAY

AHMEDABAD

aws

COMMUNITY DAY

Scaling your web applications

From 100k to 100M requests per second

aws COMMUNITY DAY

Introduction

- 10+ year into the industry
- Worked for enterprises like Oracle, Apple, Twilio
- Expert in Backend, Cloud & DevOps
- Consulted for multiple small & large scale clients
- Technical writer at Geekflare, ButterCMS, MkYong, Signoz and other popular platforms
- Presently Principal Engineer @ Twilio Segment

Agenda

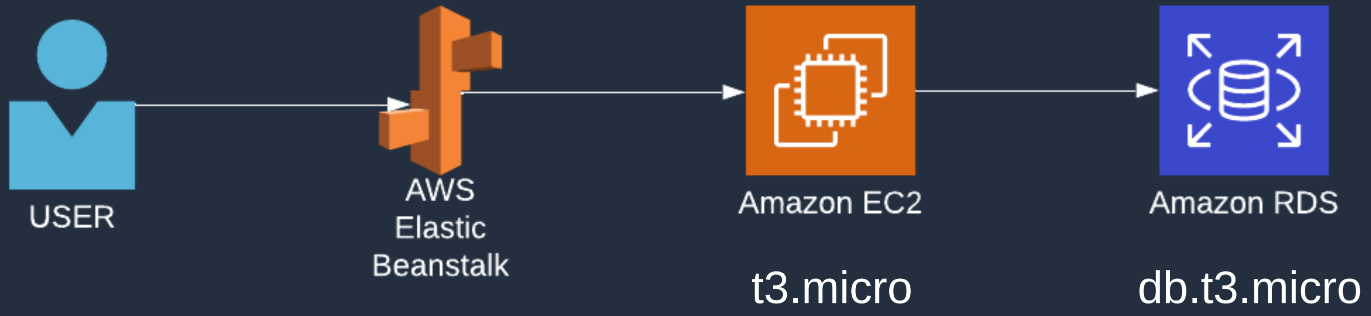
- Some facts !
- Hosting a simple web application with database in AWS
- Handling higher traffic for the web application (10k rps)
- Scaling to multiple instances (100k rps)
- Introducing High availability, auto-scaling & fault tolerance
- Going Cloud Native
- 1M rps - The Cloud reaches its limits !
- 100M rps - Architecting for massive scale

aws COMMUNITY DAY

Some facts !

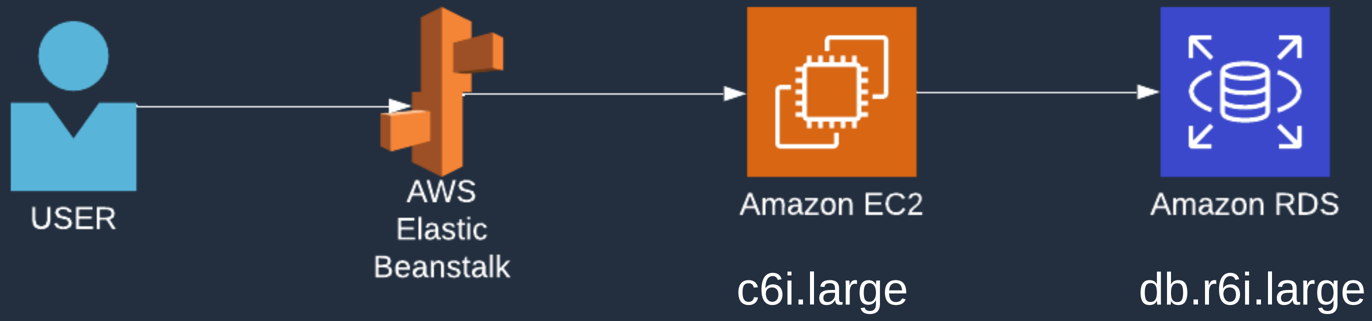
- Amazon Web Services now covers 77 Availability Zones (AZs) in 24 geographic regions across the globe !
- Clouds have limits too :)
- Highest instance memory - 24576 Gb
- Highest instance processors - 448 logical processors
- EKS cluster can handle a maximum of 13500 managed nodes per cluster !

Hosting a simple web application with database in AWS



- Beanstalk simplifies provisioning as well as deployment
- Uses EC2 as compute
- RDS instance in a small single AZ setup

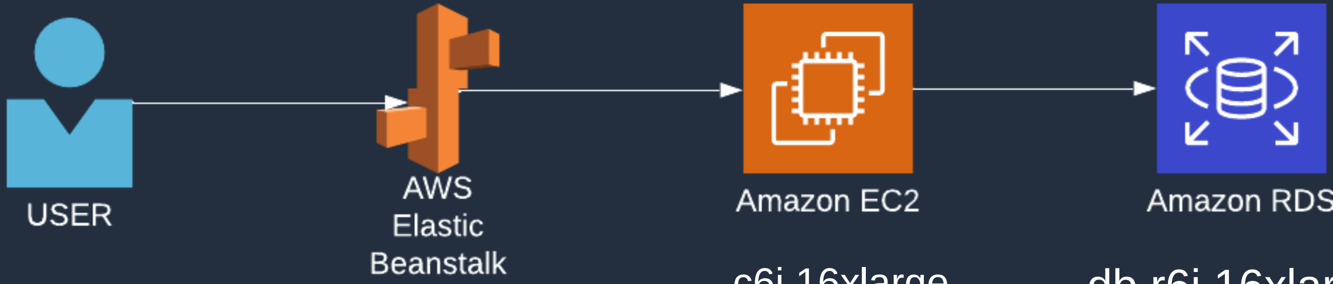
Handling higher traffic (10k rps)



- Move towards larger instance & database configuration
- Improve compute quality to provide good response time

aws COMMUNITY DAY

Scaling further - 100k rps 🚀



c6i.16xlarge

db.r6i.16xlarge

Oh ! Let me just increase the size further

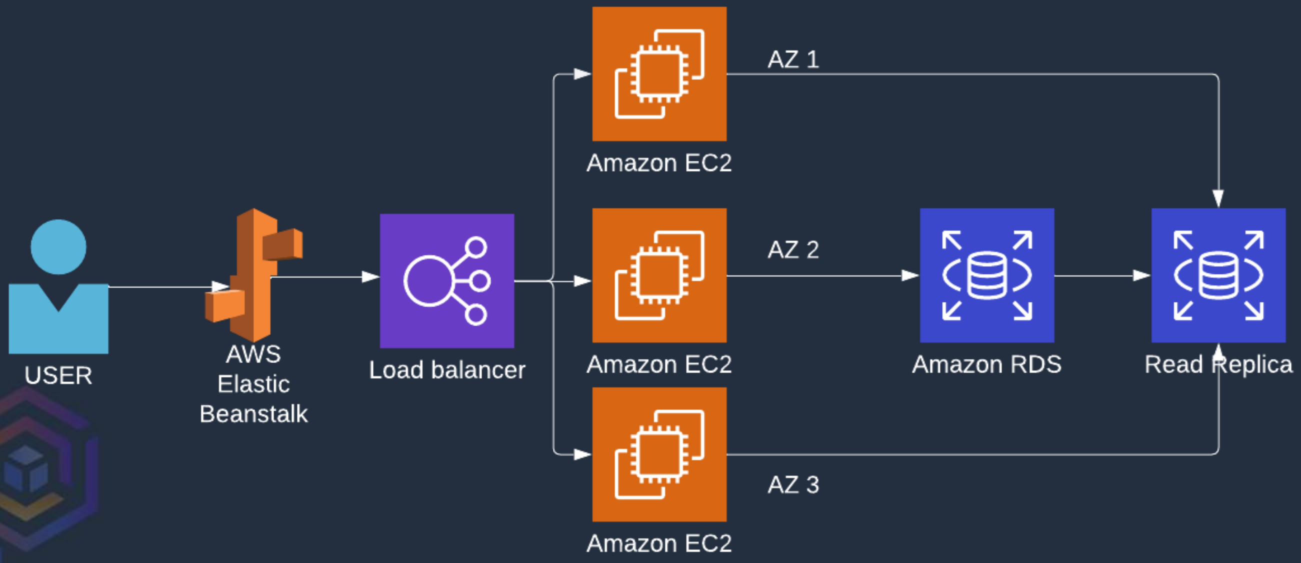
😀

How did it go?

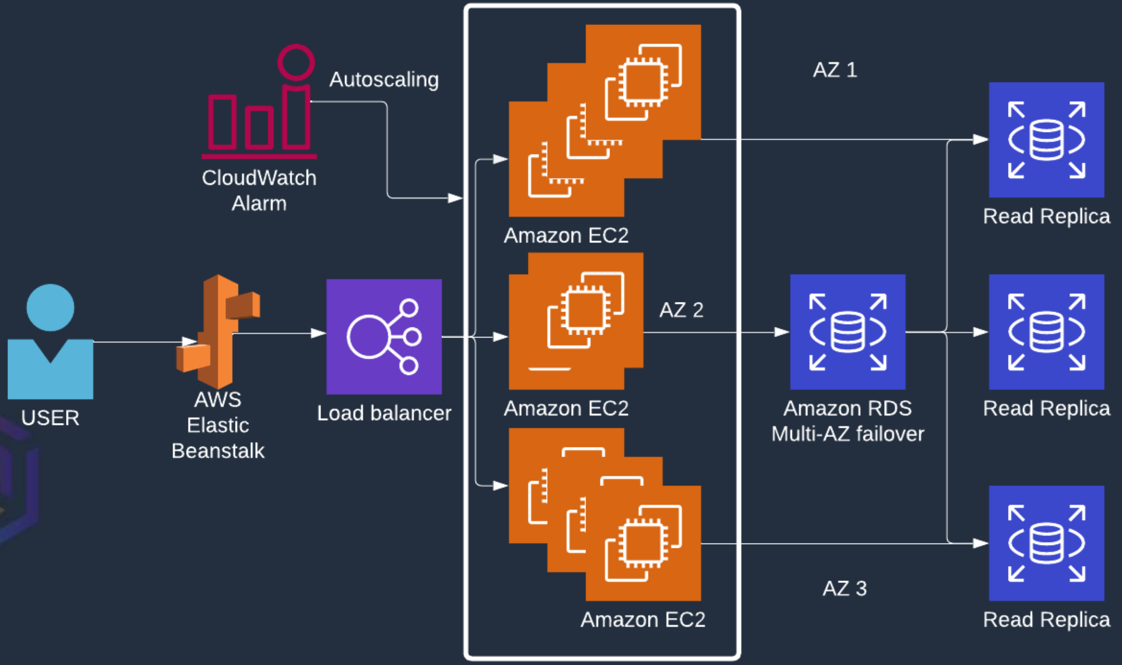
Space X starship prototype blast
(Source: https://www.ndtv.com/world-news/spacexs-starship-prototype-explodes-on-landing-after-test-launch-2336582

aws COMMUNITY DAY

Scaling to multiple instances - 100k rps 🚀



Introducing High availability, auto-scaling & resilience

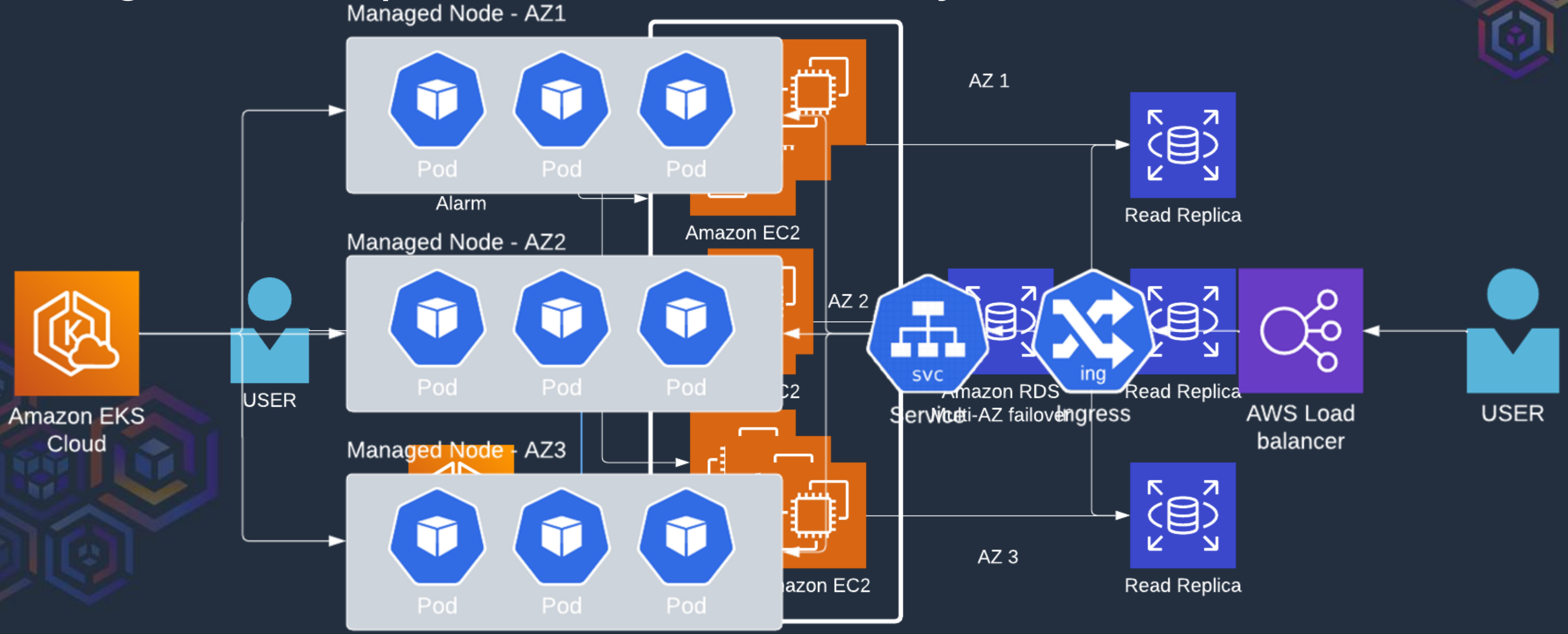


Going Cloud Native

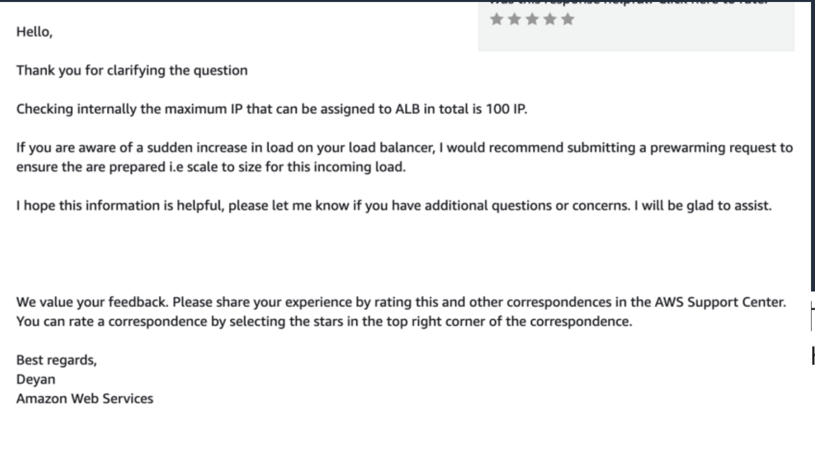
- Easier hosting of microservices
- Easy path based routing using Ingress & path based rules
- Ability to vertically and horizontally auto-scale
- Better monitoring with Operator model based observability agents
- Developer friendly application configuration - easy to define resources,scaling rules and routing rules
- Easier internal communication using services

aws COMMUNITY DAY

Scaling to 100k rps the cloud native way



1M rps - The Cloud reaches its limits !



Was this response helpful? Click here to rate:

★★★★★

Hello,

Thank you for clarifying the question

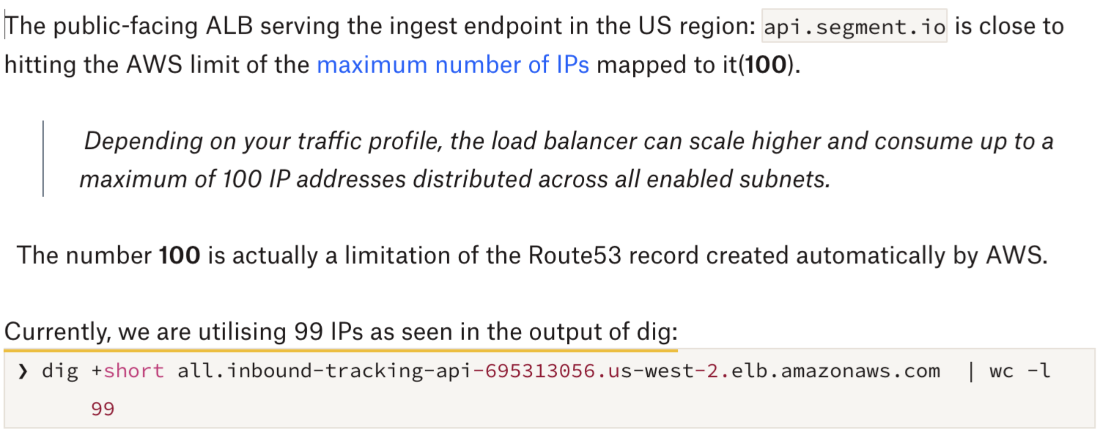
Checking internally the maximum IP that can be assigned to ALB in total is 100 IP.

If you are aware of a sudden increase in load on your load balancer, I would recommend submitting a prewarming request to ensure the are prepared i.e scale to size for this incoming load.

I hope this information is helpful, please let me know if you have additional questions or concerns. I will be glad to assist.

We value your feedback. Please share your experience by rating this and other correspondences in the AWS Support Center. You can rate a correspondence by selecting the stars in the top right corner of the correspondence.

Best regards,
Deyan
Amazon Web Services



The public-facing ALB serving the ingest endpoint in the US region: `api.segment.io` is close to hitting the AWS limit of the maximum number of IPs mapped to it(**100**).

> *Depending on your traffic profile, the load balancer can scale higher and consume up to a maximum of 100 IP addresses distributed across all enabled subnets.*

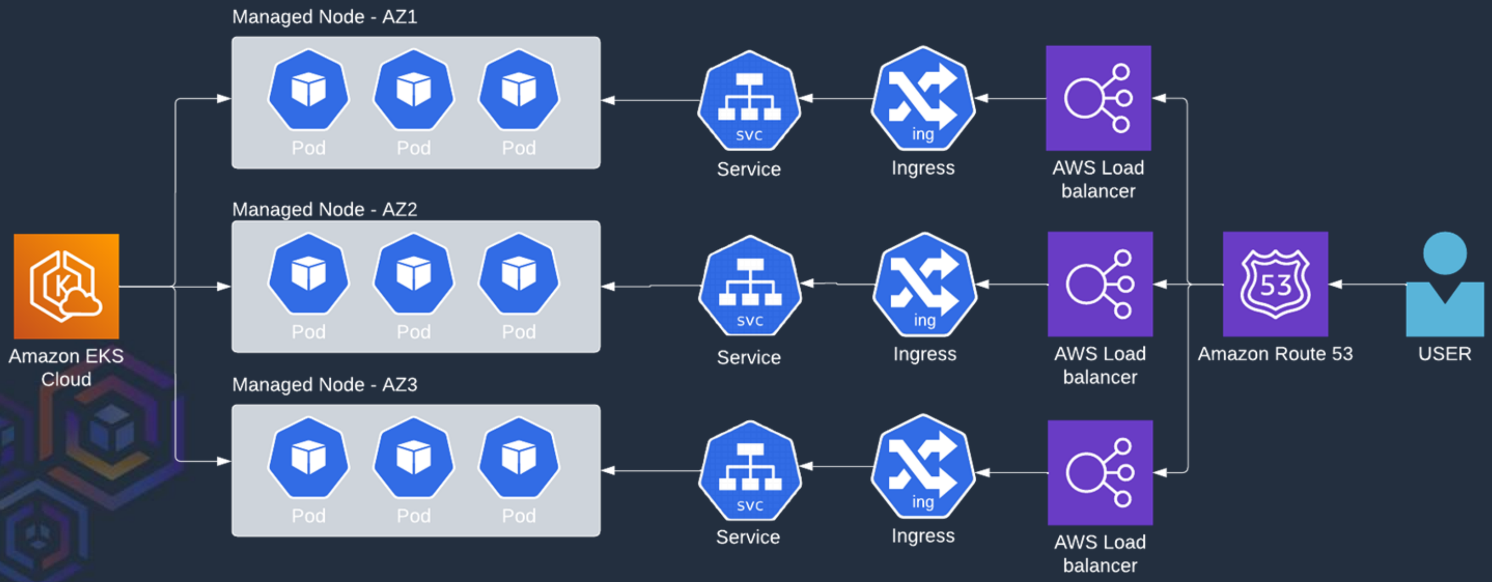
The number **100** is actually a limitation of the Route53 record created automatically by AWS.

Currently, we are utilising 99 IPs as seen in the output of dig:

```
❯ dig +short all.inbound-tracking-api-695313056.us-west-2.elb.amazonaws.com  | wc -l
      99
```

aws COMMUNITY DAY

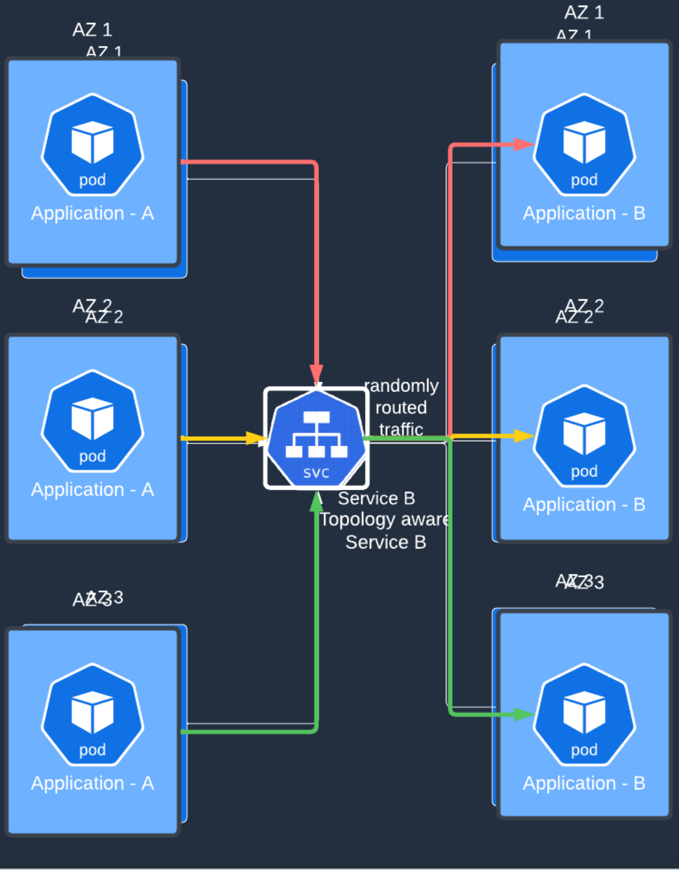
100M rps - Architecting for massive scale



aws
COMMUNITY DAY

100M rps - Cost Efficiency

- App to App communication start going cross-AZ
- Cross-AZ network costs quickly start to add up
- Maintaining all the traffic in single AZ could be a disaster too
- Topology aware services to the rescue 🚀



100M rps - Points to look out

- Cross-AZ traffic
- Fallback during a zone failure
- Health checks to auto-failover
- Spikes v/s uniform increase - handling the load
- Pre-warming load balancer and Node Pool

aws

COMMUNITY DAY