

1000 Services, 1 Year, 0 Downtime: Airbnb's Zonal Cluster Migration

Sunny Beatteay

The Team



Jordan Ly
Tech Lead



Daniel Low
Engineer



Sunny Beatteay
Engineer



Ruan Arcega
Engineer

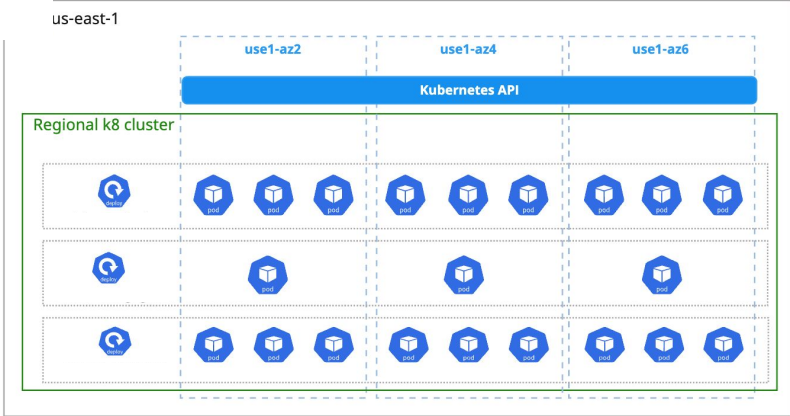


Swetha Vaidy
TPM

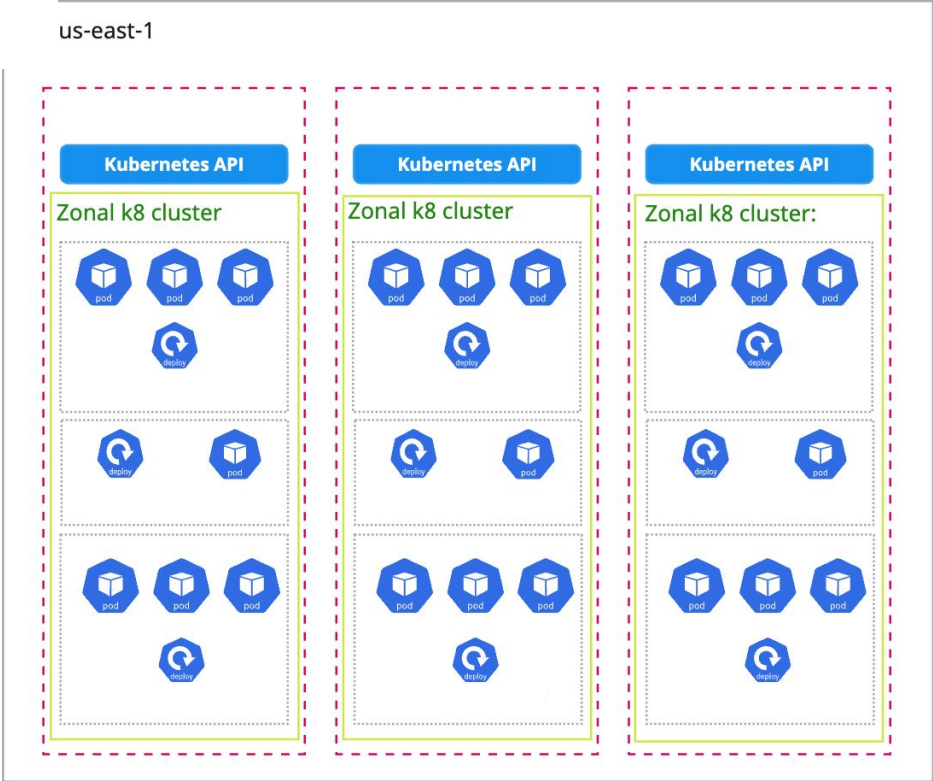


Moh Bazyar
Engineer

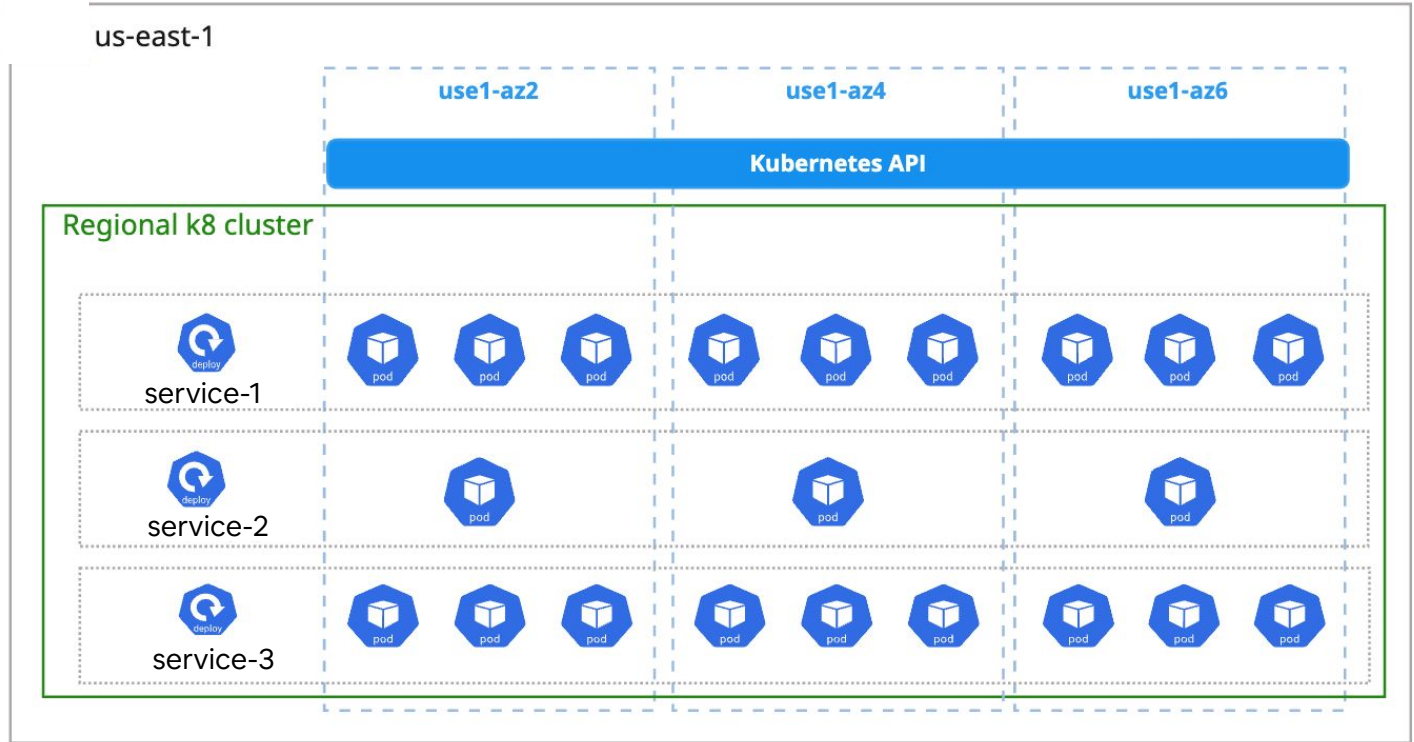
Regional vs Zonal



VS



Regional Cluster

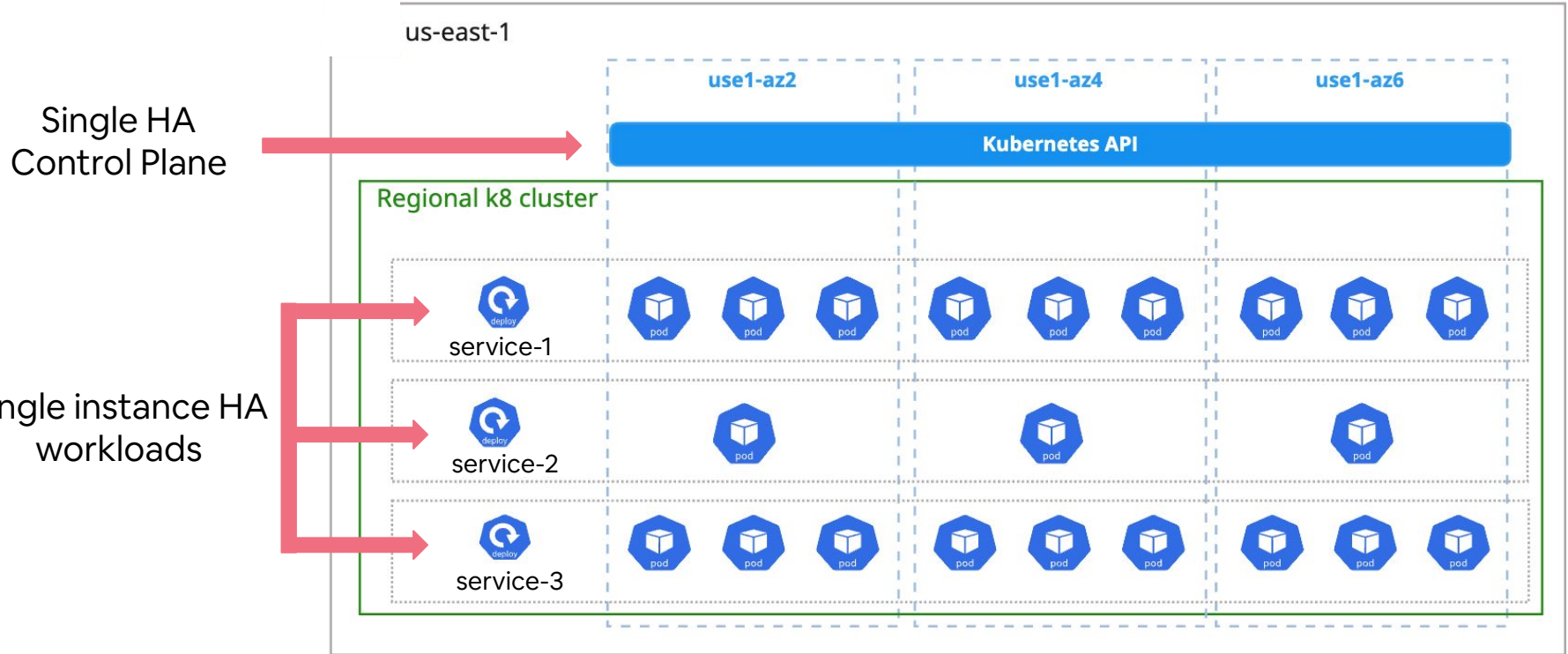


Regional Cluster

Single HA
Control Plane

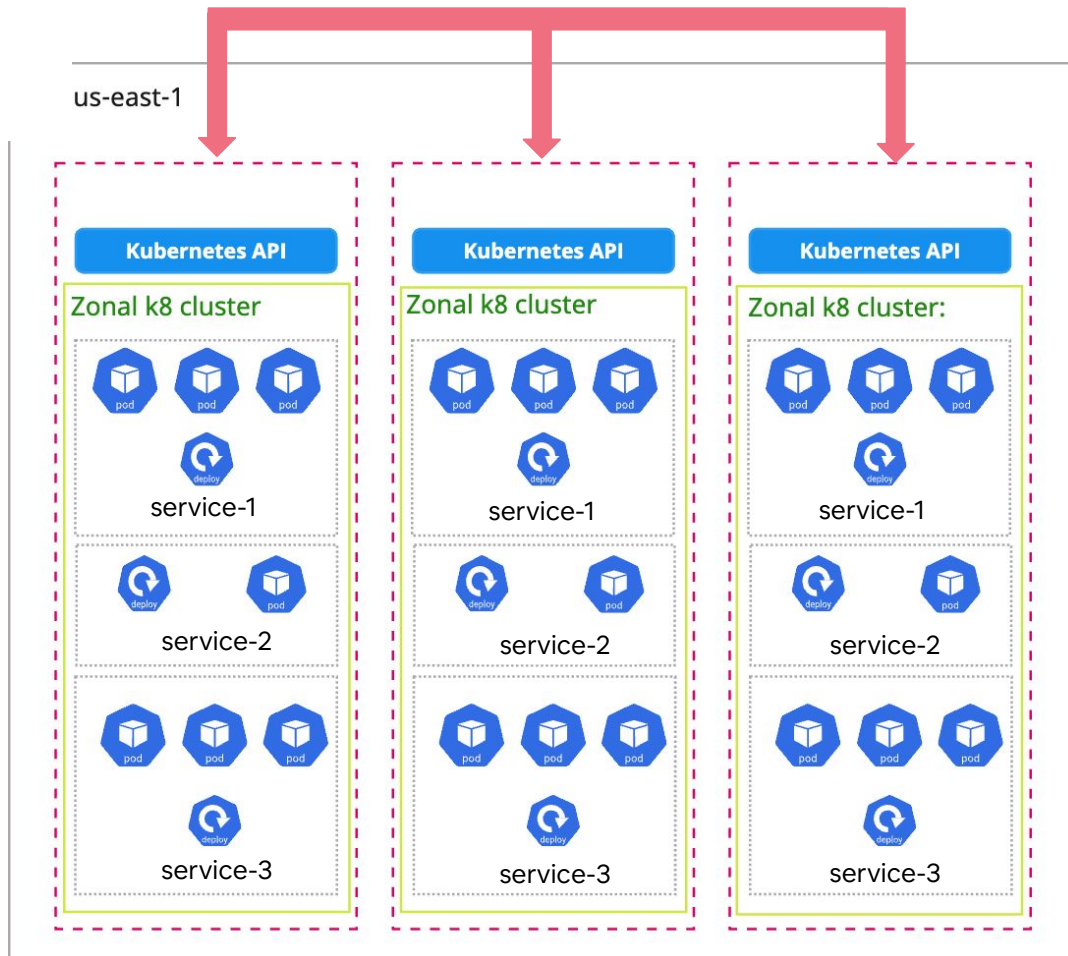


Regional Cluster



Zonal Clusters

Cluster per AZ

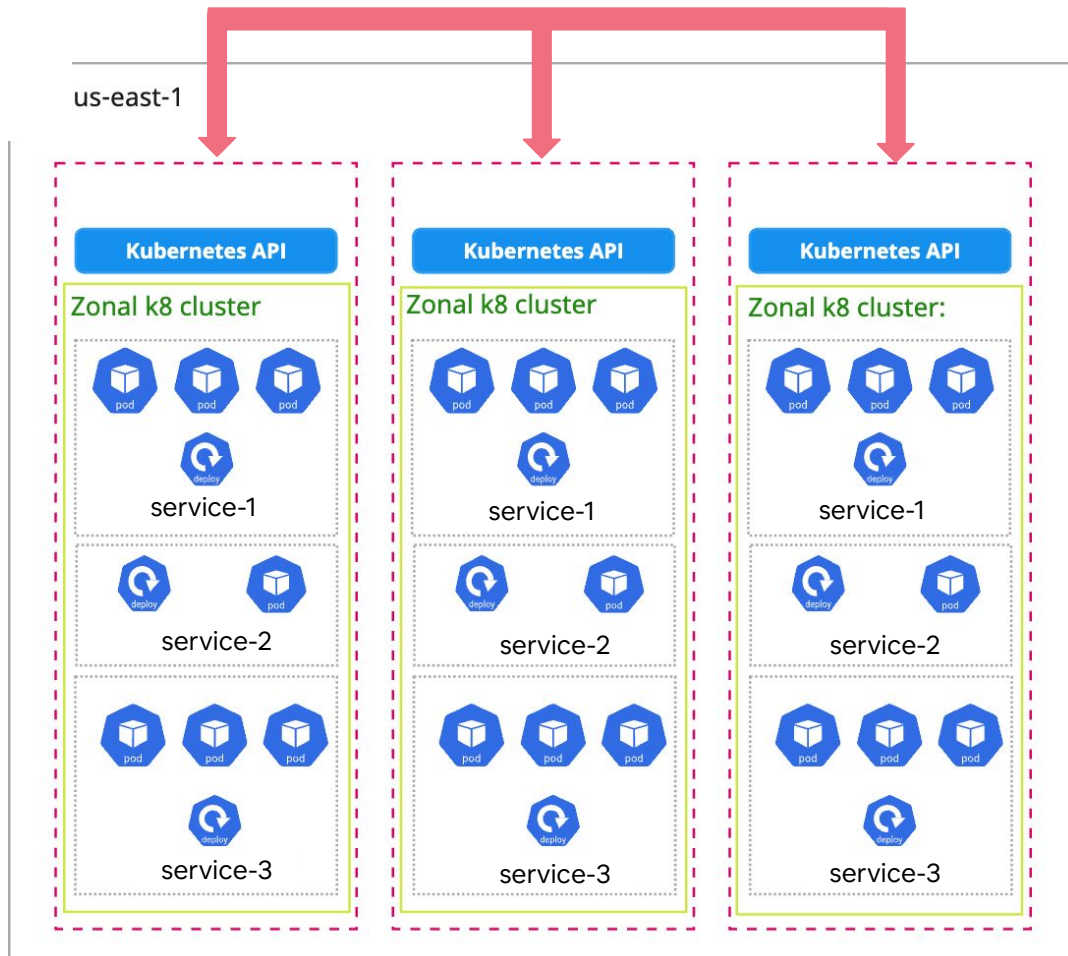


Zonal Clusters

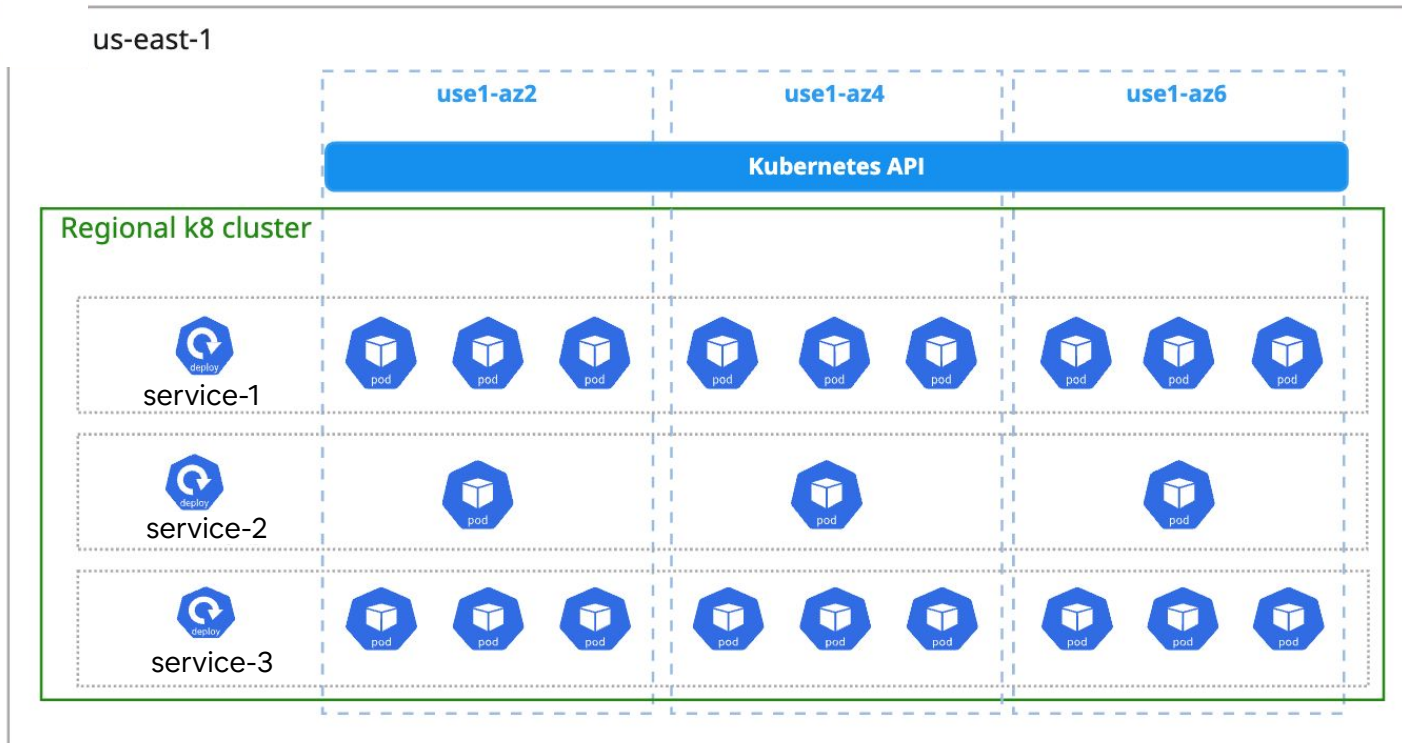
Cluster per AZ

us-east-1

Sharded
workloads

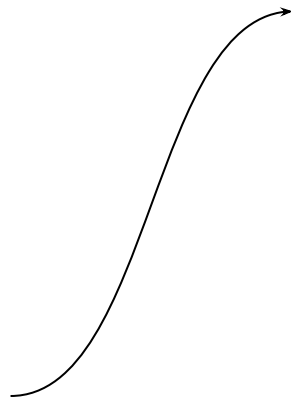


State of Airbnb Compute in 2018-2023

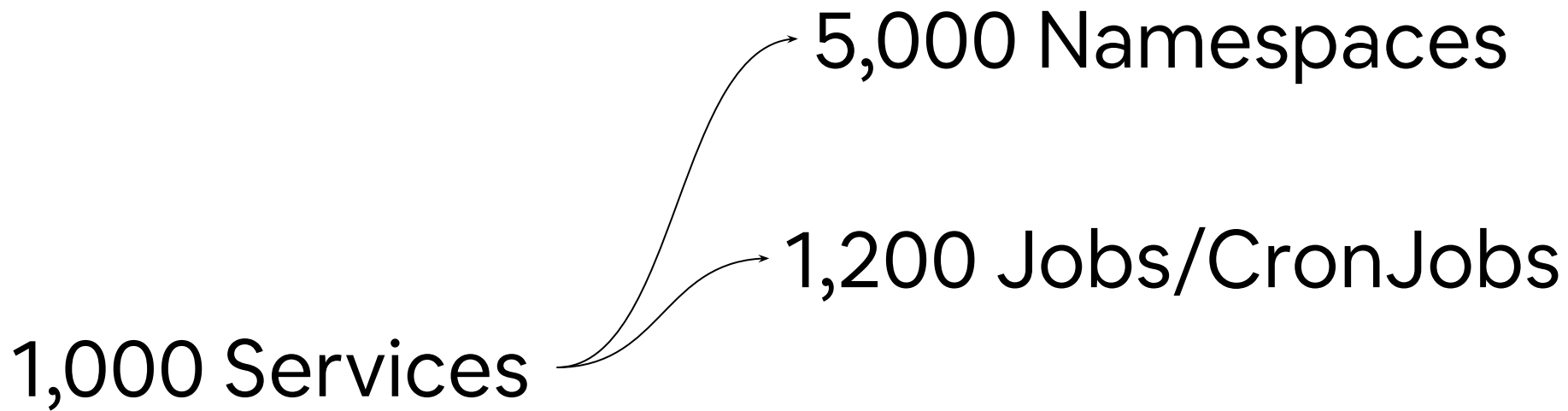


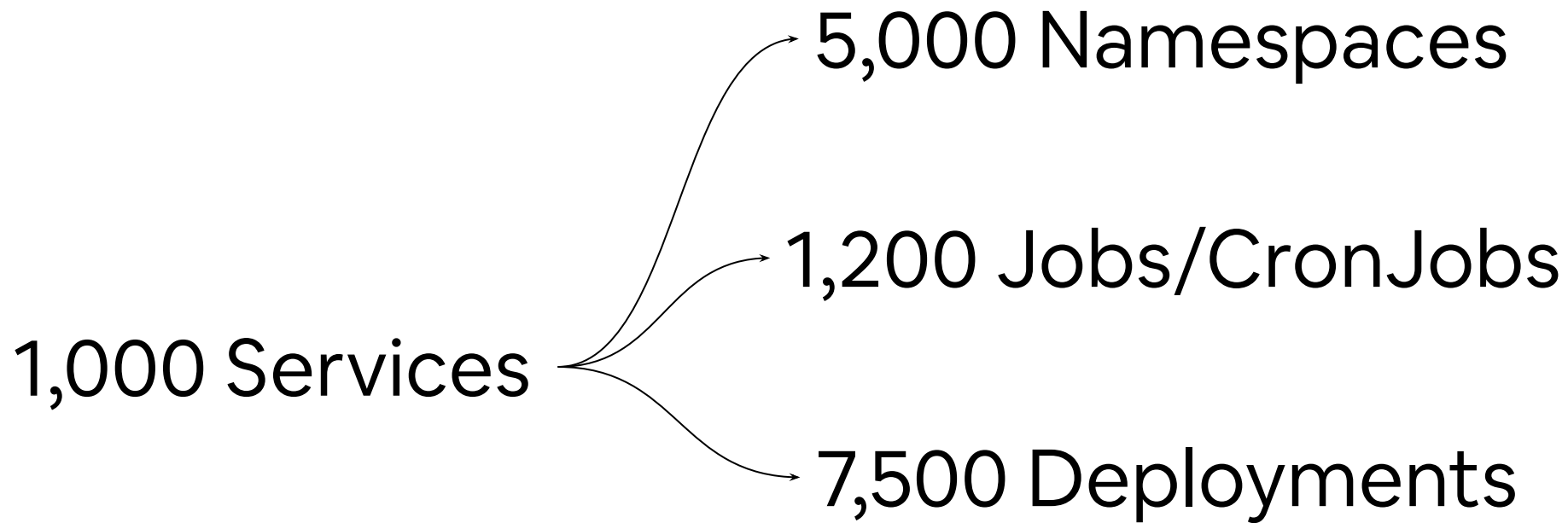
1,000 Services

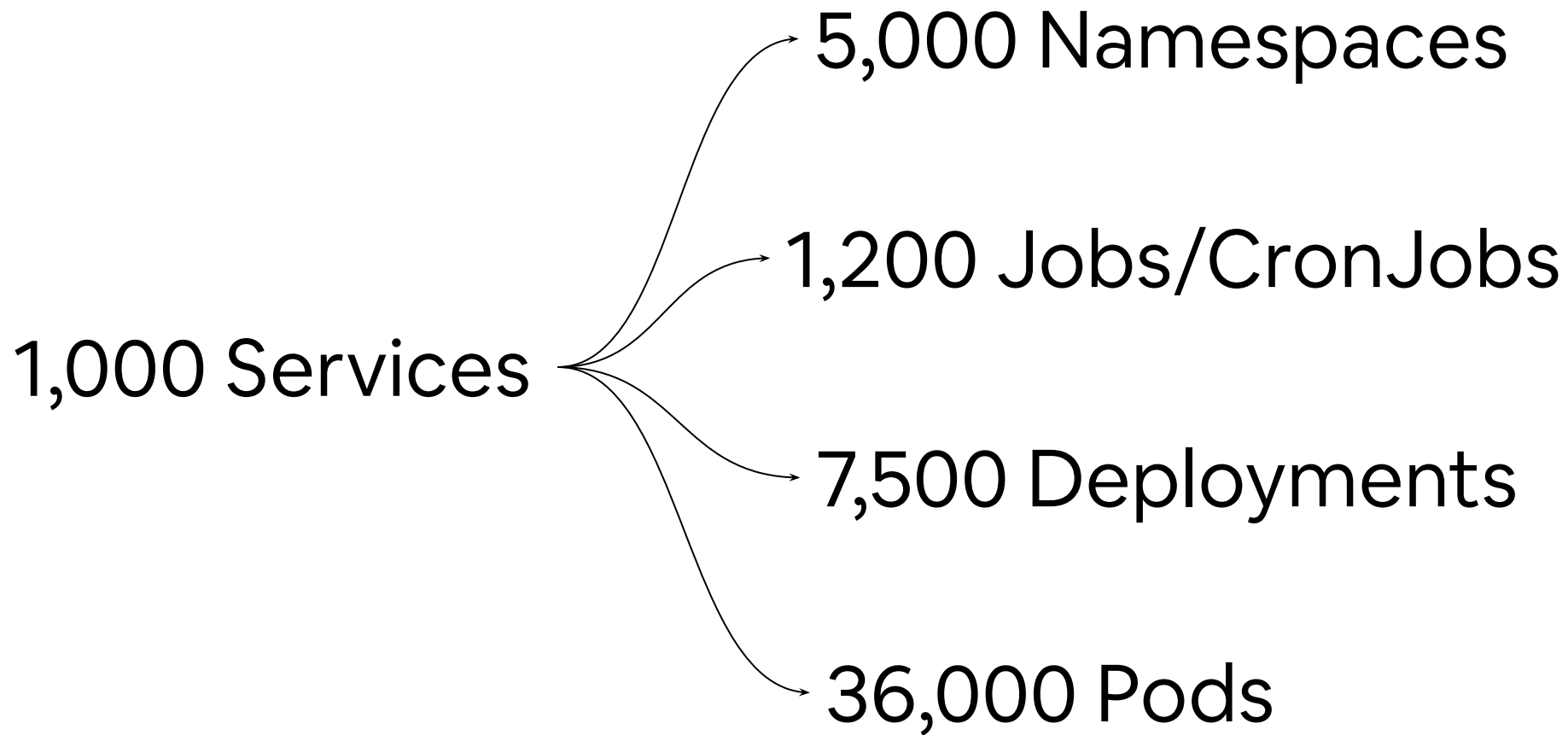
1,000 Services



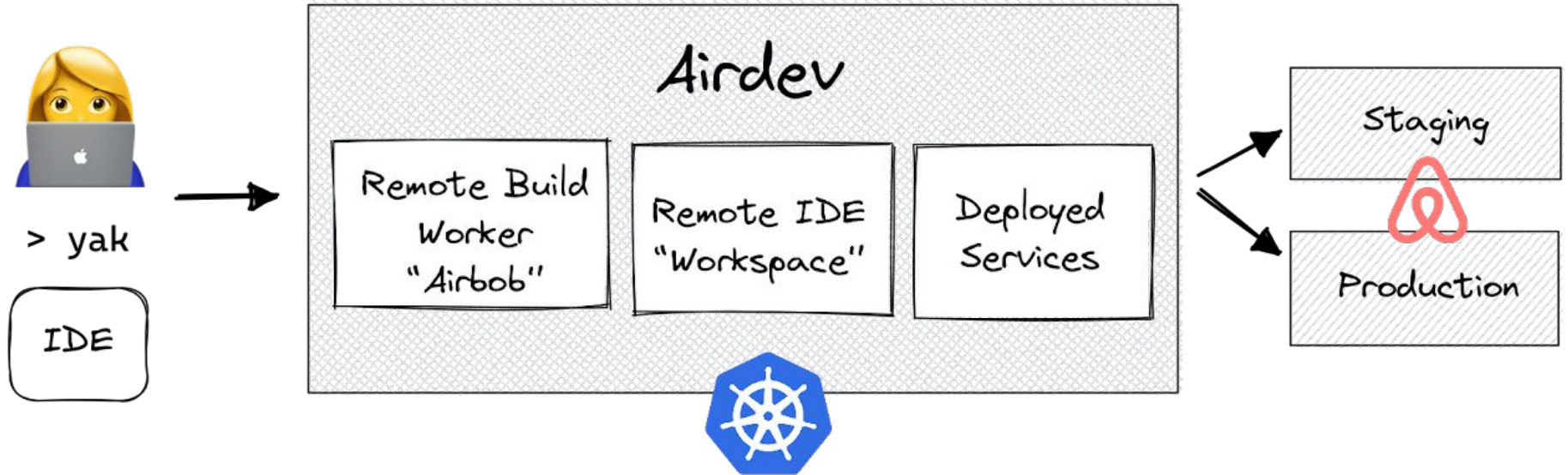
5,000 Namespaces







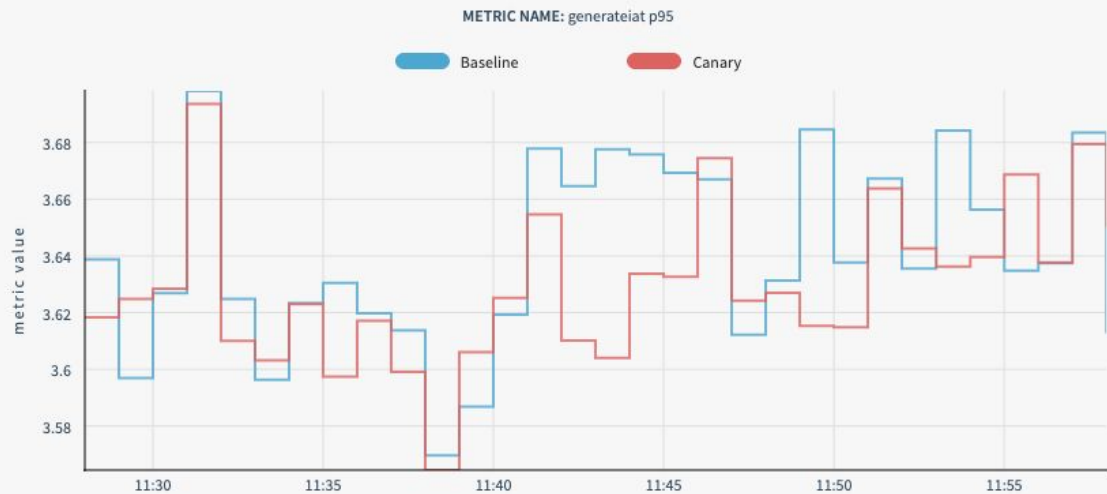
Incident Prevention



Incident Prevention

ALL	ERRORS	LATENCY	BUSINESS LOGIC	MEMORY AND CP
-----	--------	---------	----------------	---------------

Metric Name	Deviation	Result
generateiat p95	-0.3%	Pass
getauthenticateduser p95	+1.1%	Pass
service latency p95	-0.2%	Pass
service latency p99	+2.8%	Pass

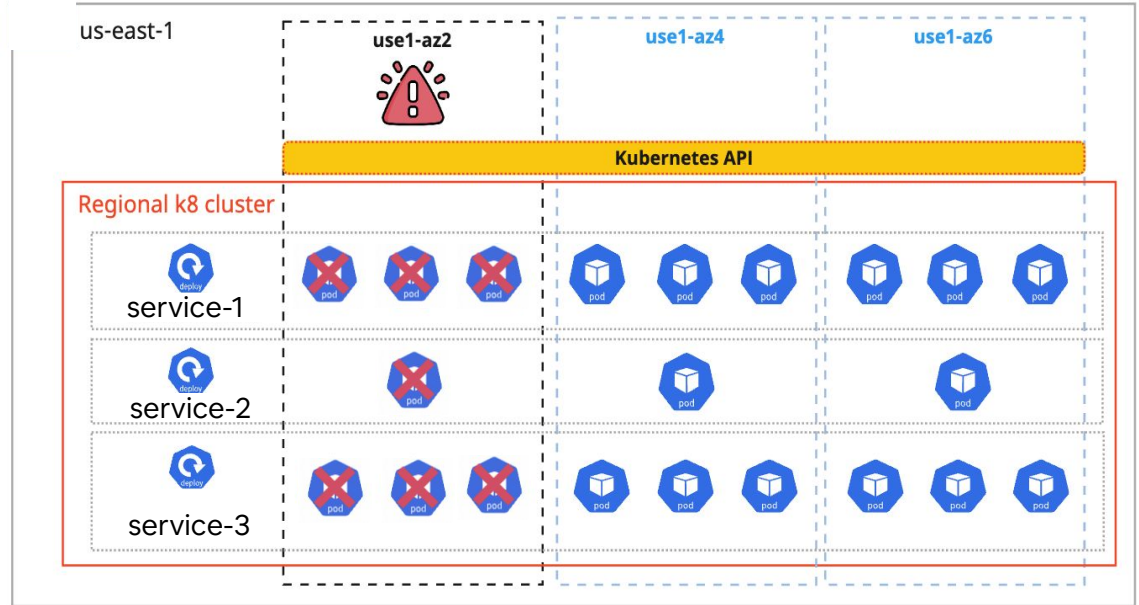


Why migrate?

Incident Recovery

Regional Clusters

Protects against AZ failures

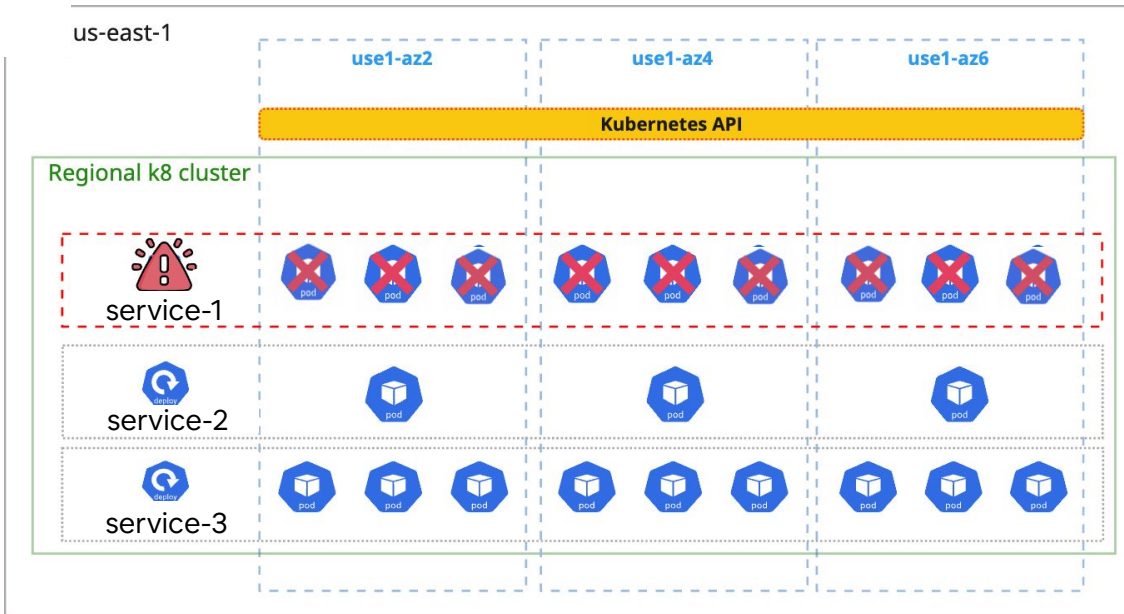


Regional Clusters

But doesn't protect us from ourselves.

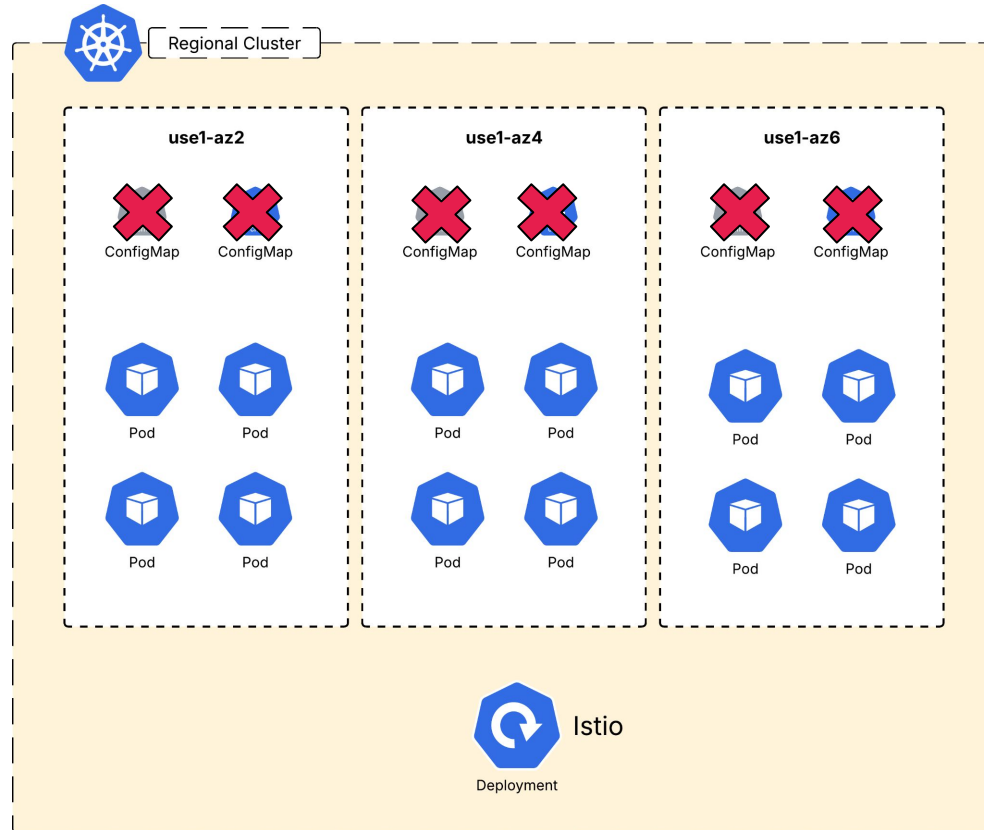
High blast radius for operations:

- App Deploys
- Infra Updates



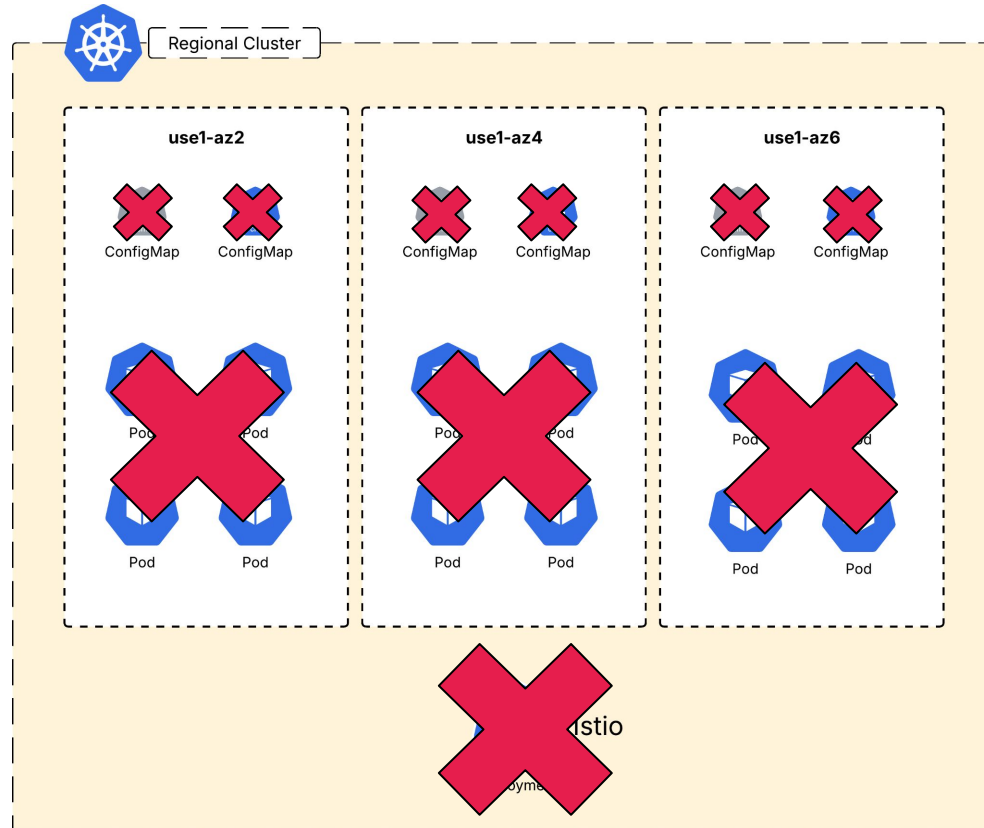
Incident Scenario: Regional vs Zonal

Incident Scenario - Regional



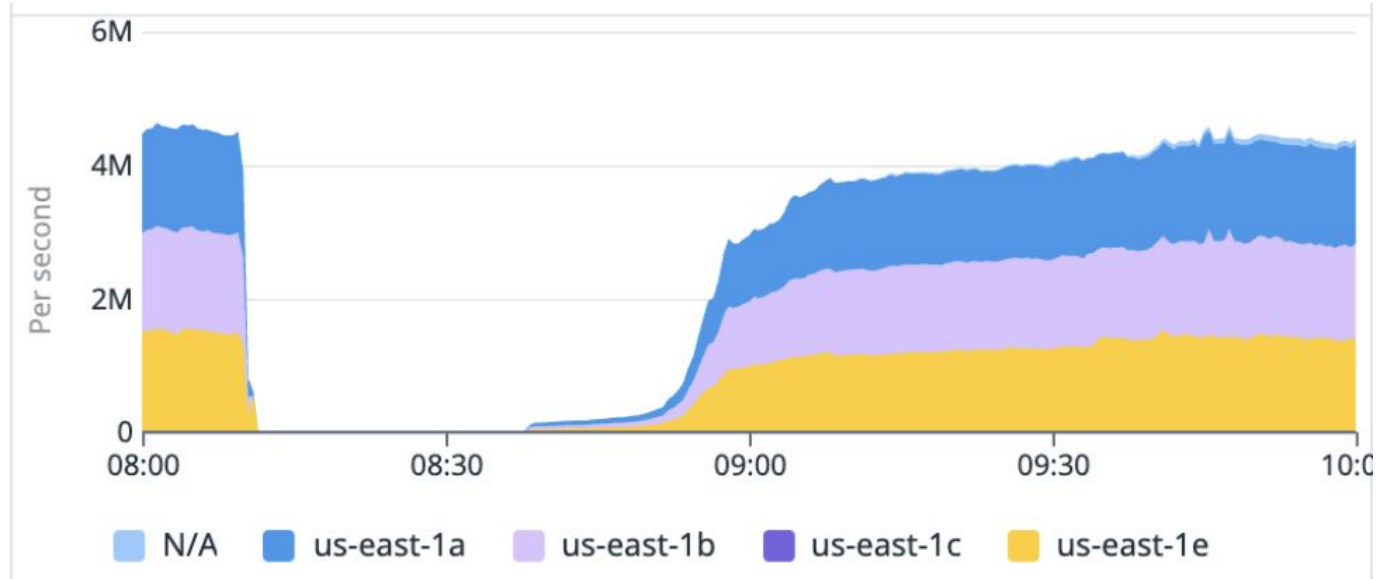
Incident Scenario - Regional

Cascading failures



Incident Scenario - Regional

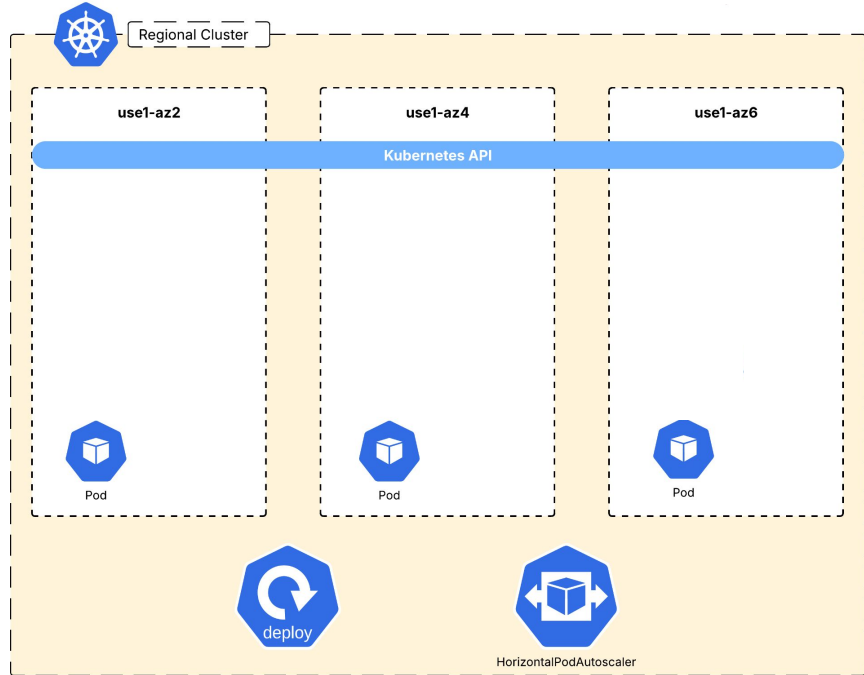
Traffic drops
by 96%



Incident Scenario - Regional

While root causing and hot fixing:

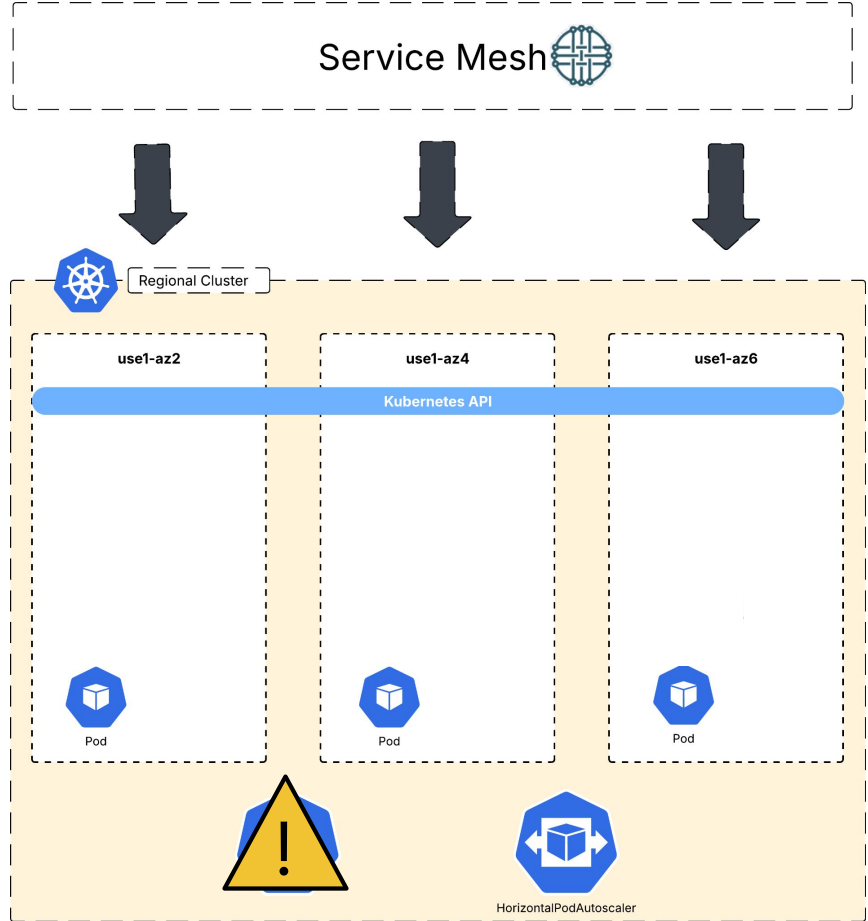
- Workloads have scaled down due to lack of traffic.



Incident Scenario - Regional

While root causing and hot fixing:

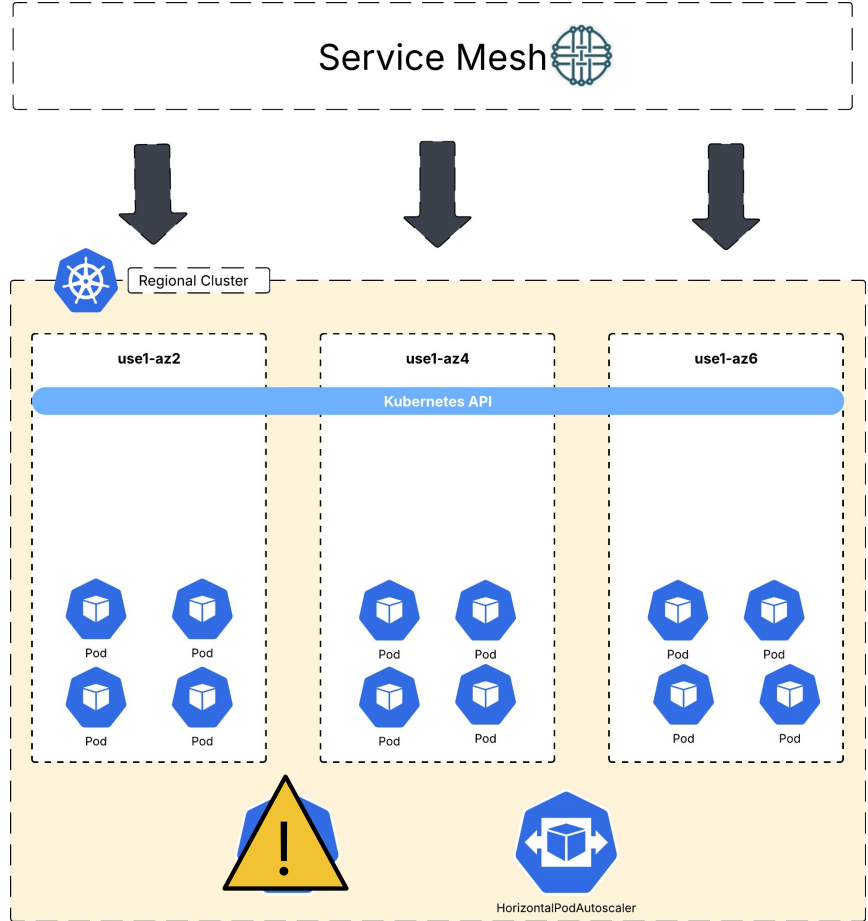
- Workloads have scaled down due to lack of traffic.
- When traffic recovers, workloads are overwhelmed.



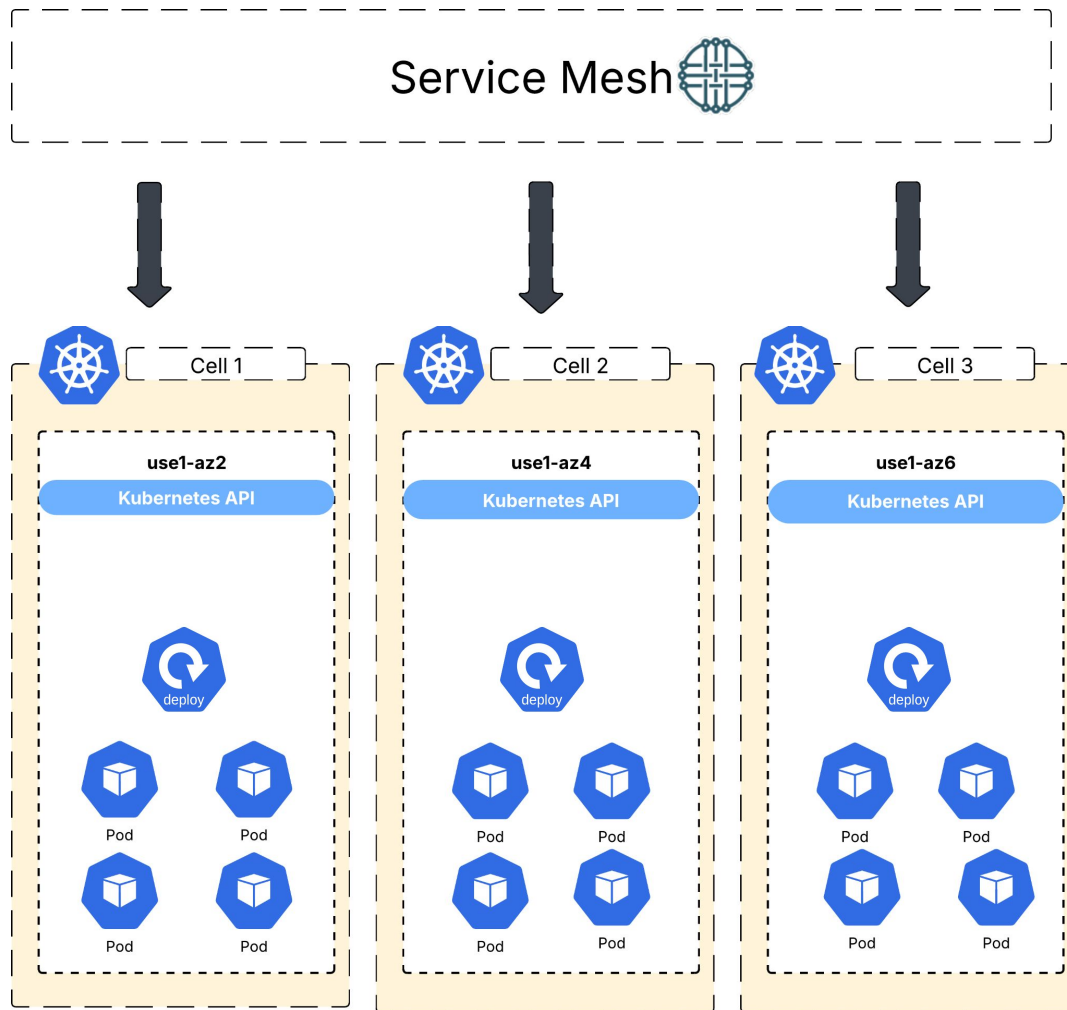
Incident Scenario - Regional

While root causing and hot fixing:

- Workloads have scaled down due to lack of traffic.
- When traffic recovers, workloads are overwhelmed.
- Backpressure as workloads scale, prolonging the issue.

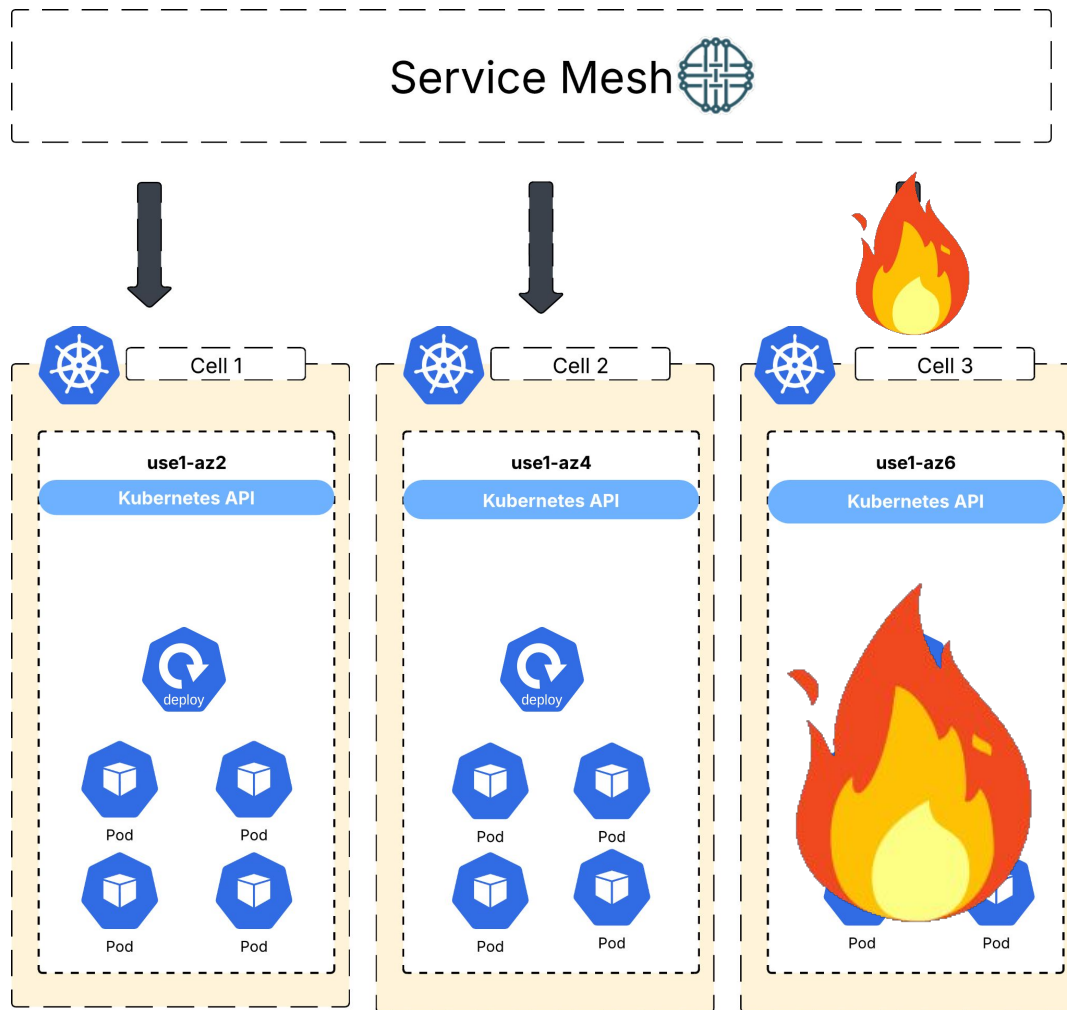


Zonal Scenario



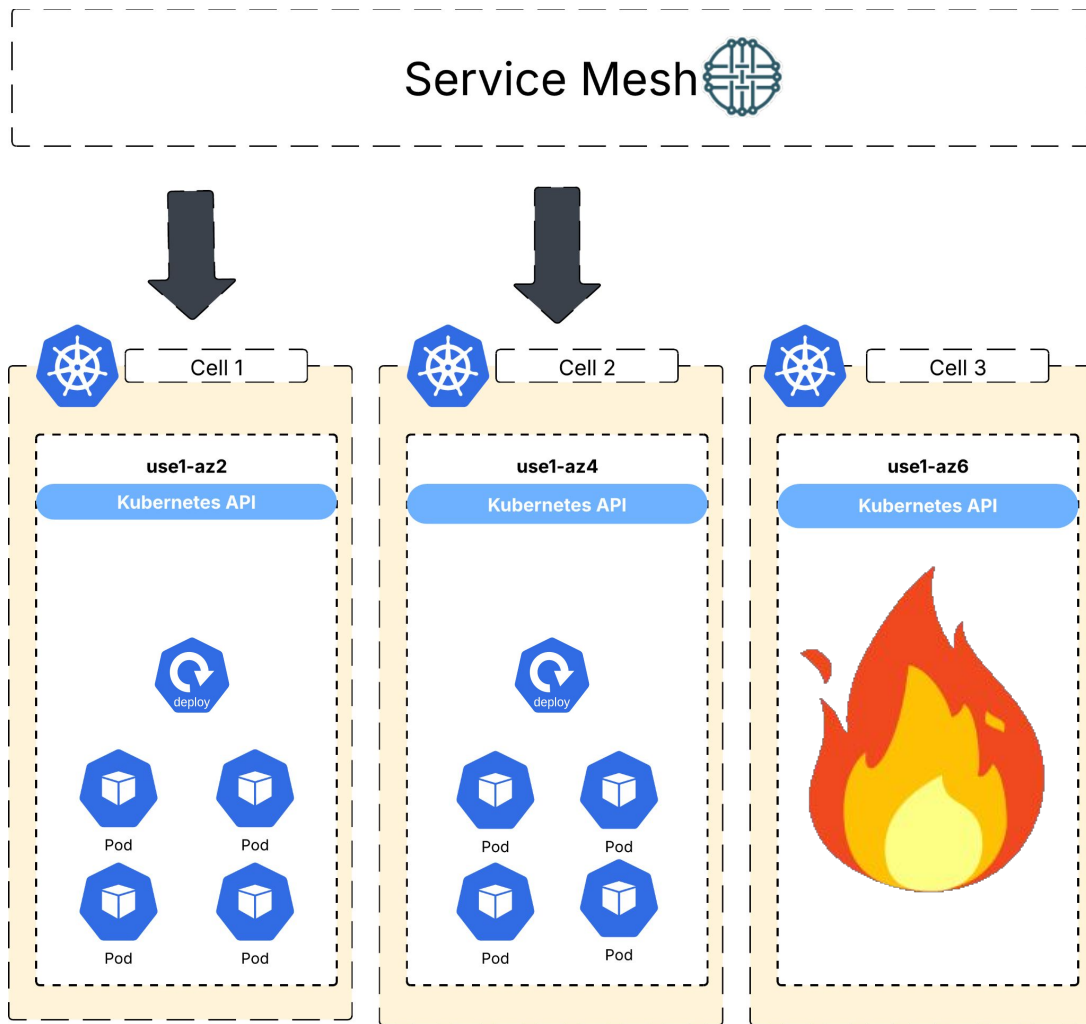
Zonal Scenario

- Outage is contained to 1/3 of region.



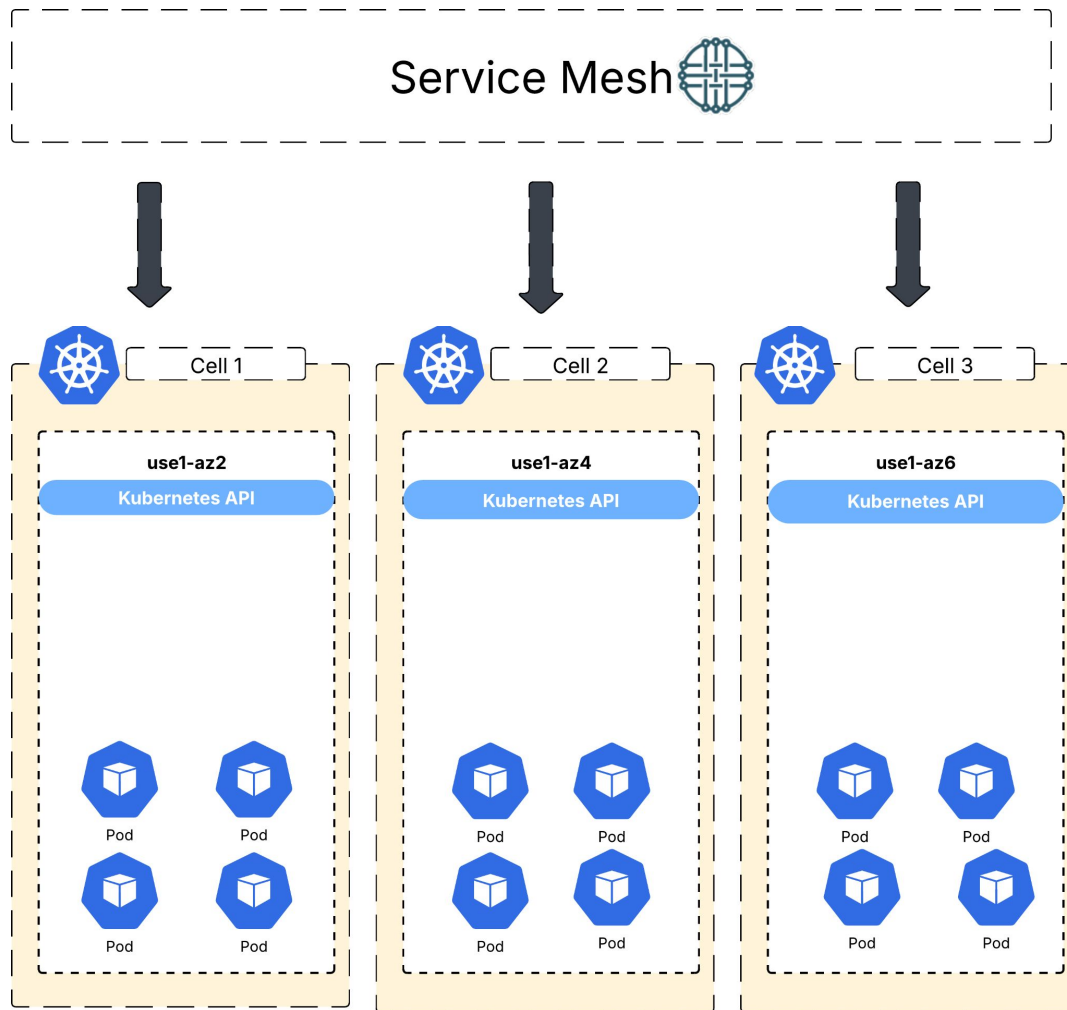
Zonal Scenario

- Outage is contained to 1/3 of region.
- Failover traffic to healthy clusters.



Zonal Scenario

- Outage is contained to $\frac{1}{3}$ of region.
- Failover traffic to healthy clusters.
- Traffic shift back after recovery.



Reduce the blast radius of changes

Reduce the blast radius of changes

Faster time-to-recovery

Proof of Concept

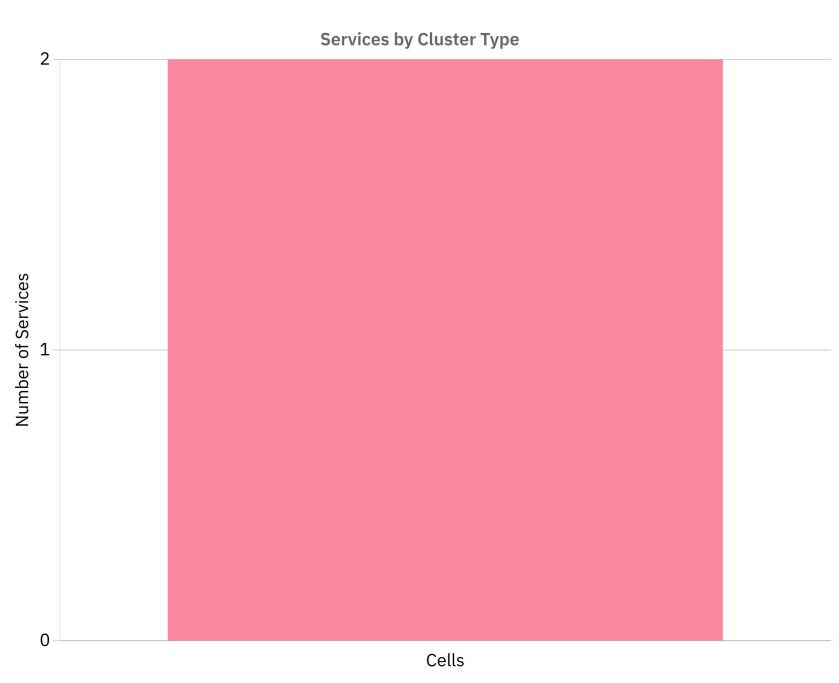
Proof of Concept

Small scale

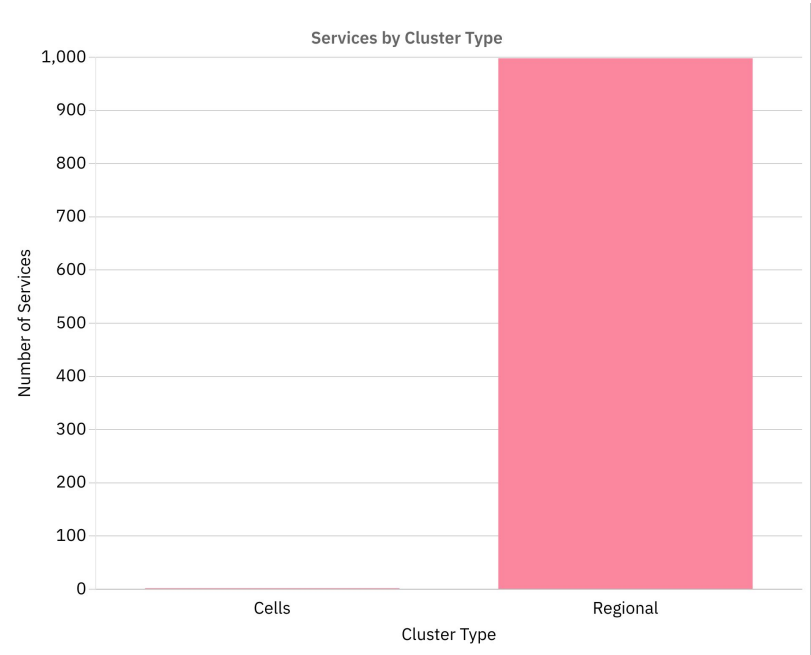
Create zonal clusters

Find volunteers to migrate

Success!



Success!
...kind of



To put into perspective

2 Migrations in 3 months

==

?

To put into perspective

2 Migrations in 3 months

==

1000 Migrations in 125 years

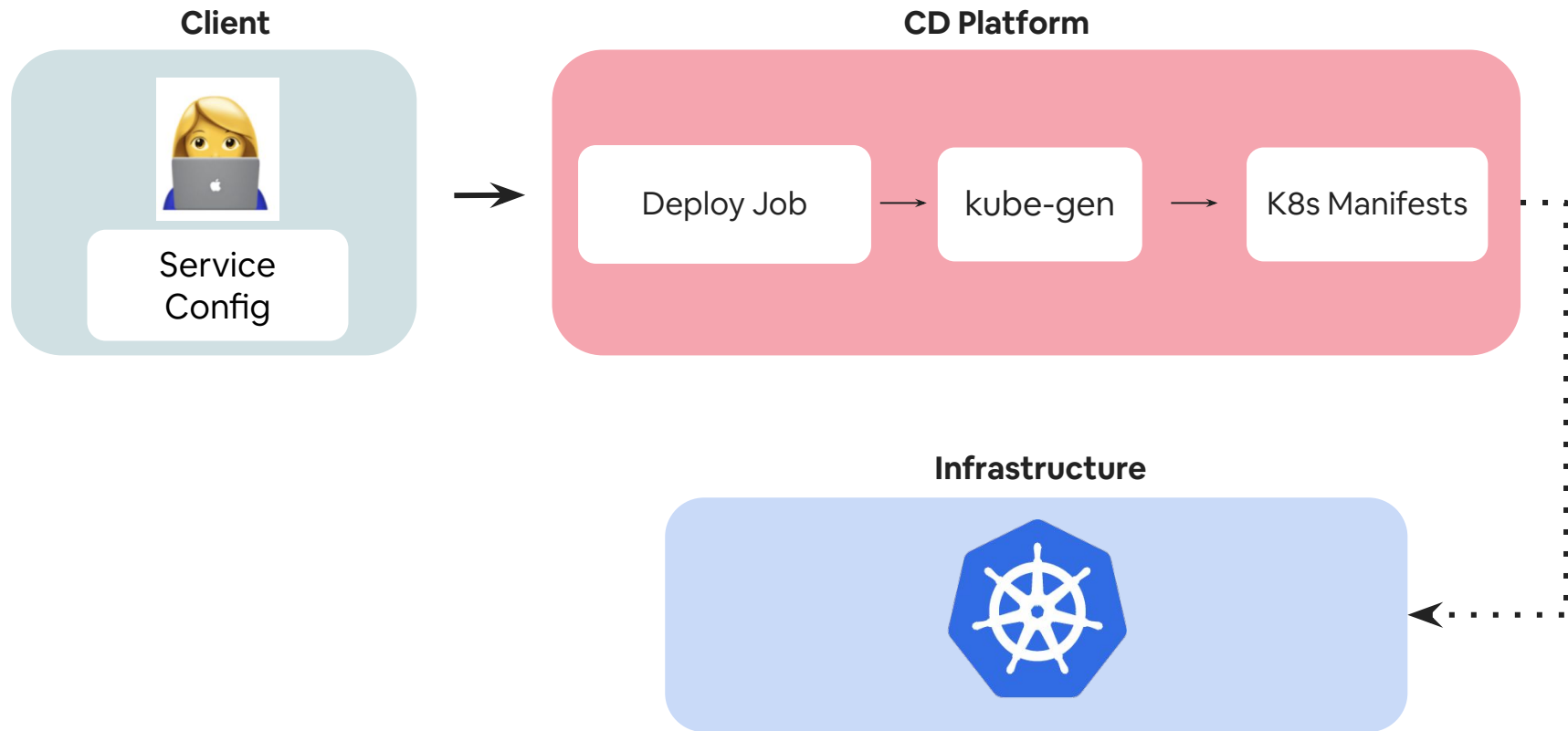
Migration Friction Points

Technical Challenges

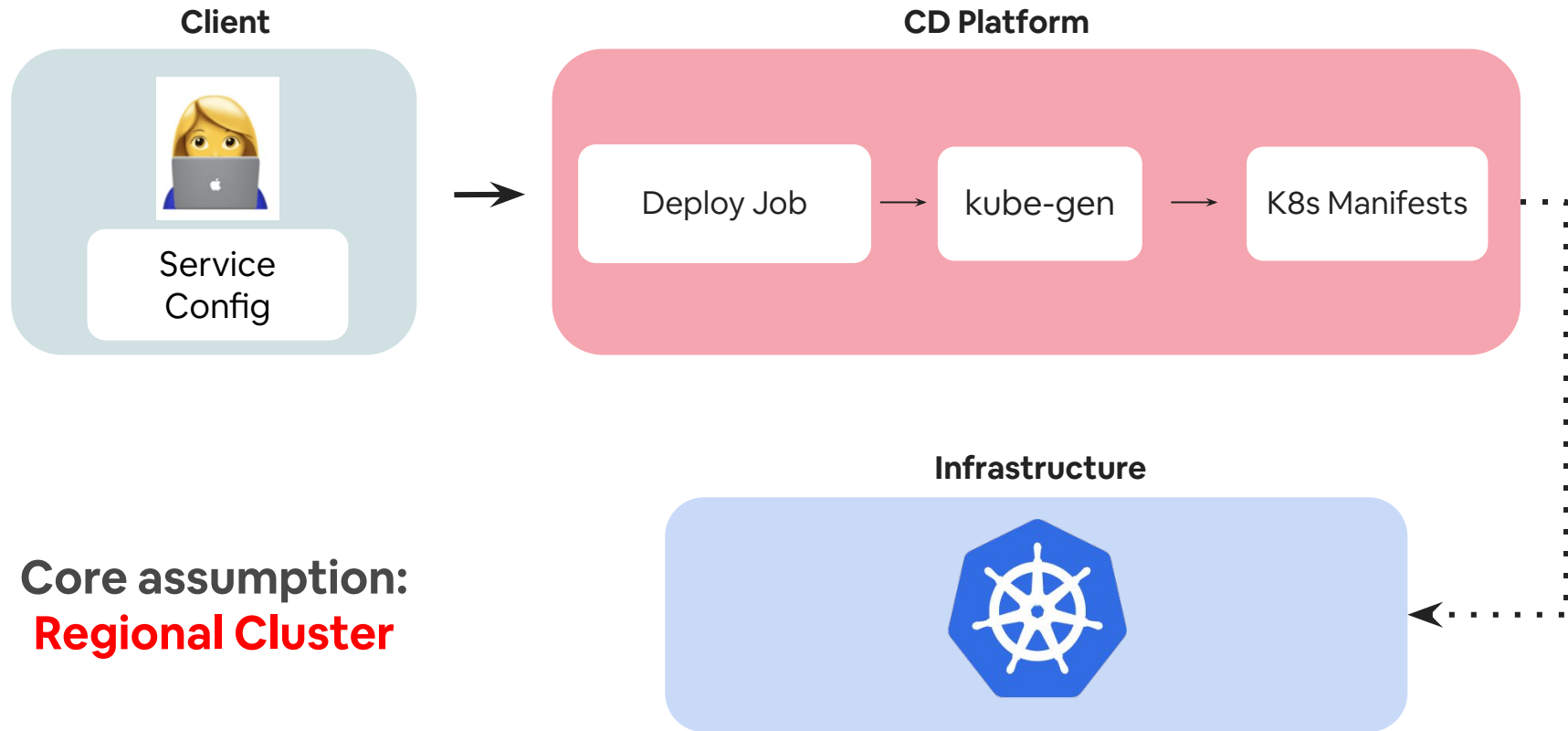
Deployment System

OneTouch

OneTouch System



OneTouch System



OneTouch

```
# kube-gen.yml
```

```
environments:
```

```
  production:
```

```
    context: cluster-1
```

```
    params:
```

```
      minReplicas: 9
```

```
      maxReplicas: 27
```

```
# cd.yml
```

```
stages:
```

```
  - name: deploy-to-prod
```

```
    type: onetouchDeploy
```

```
    parameters:
```

```
      context: cluster-1
```

```
      environment: prod
```

**OneTouch Status:
Maintenance Mode**

Long and Messy Configs

Number of lines in kube-gen.yml



```
{{ if eq .Env.Name "staging" }}  
{{ else }}  
{{ if eq .Env.Name "canary" }}  
{{ else }}  
{{ if eq .Env.Name "canary-baseline" }}  
{{ else }}  
{{ if eq .Env.Name "test-rc" }}  
{{ else }}  
{{ if eq .Env.Name "test" }}  
{{ else }}  
{{ if eq .Env.Name "test-hotfix" }}  
{{ else }}  
{{ if eq .Env.Name "test-train" }}  
{{ else }}  
{{ if eq .Env.Name "dev" }}  
{{ end }}
```

“Service configs are too hard to maintain.”

“I have no idea what half the things do and whether they matter or not.”

“App teams don’t have the tooling to debug their service issues”

Deployment System

Service Migrations

Service Migration

Step 1

```
# Prepare

# Get the current number of pods in the regional deployment.
kubectl --context "${CONTEXT}" --namespace "${NAMESPACE}" get "deployment/${APP_ENV}" -o=jsonpath='{.spec.replicas}'

# Set START_REPLICAS with the desired maximum scale during the migration
START_REPLICAS=30

kubectl --context "${CONTEXT}" --namespace "${NAMESPACE}" patch "hpa/${APP_ENV}-autoscaler" -p "{\"spec\": {\"minReplicas\": ${START_REPLICAS}}}"

# We work with the HPA for scale-in to allow easy reversibility.
STAGE_NUM=1
CUR_REPLICAS=$(( ${START_REPLICAS} - ${STAGE_NUM} * ${START_REPLICAS} / 6 ))

# Check current replicas.
echo "Scaling $(kubectl --context "${CONTEXT}" --namespace "${NAMESPACE}" get "deployment/${APP_ENV}" -o=jsonpath='{.spec.replicas}') to ${CUR_REPLICAS}"

# Teardown

kubectl --context "${CONTEXT}" --namespace "${NAMESPACE}" patch "hpa/${APP_ENV}-autoscaler" -p "{\"spec\": {\"maxReplicas\": ${CUR_REPLICAS}, \"minReplicas\": ${CUR_REPLICAS}}}"

STAGE_NUM=2
CUR_REPLICAS=$(( ${START_REPLICAS} - ${STAGE_NUM} * ${START_REPLICAS} / 6 ))

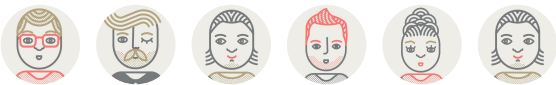
echo "Scaling $(kubectl --context "${CONTEXT}" --namespace "${NAMESPACE}" get "deployment/${APP_ENV}" -o=jsonpath='{.spec.replicas}') to ${CUR_REPLICAS}"

# Scale down
kubectl --context "${CONTEXT}" --namespace "${NAMESPACE}" patch "hpa/${APP_ENV}-autoscaler" -p "{\"spec\": {\"maxReplicas\": ${CUR_REPLICAS}, \"minReplicas\": ${CUR_REPLICAS}}}"
```

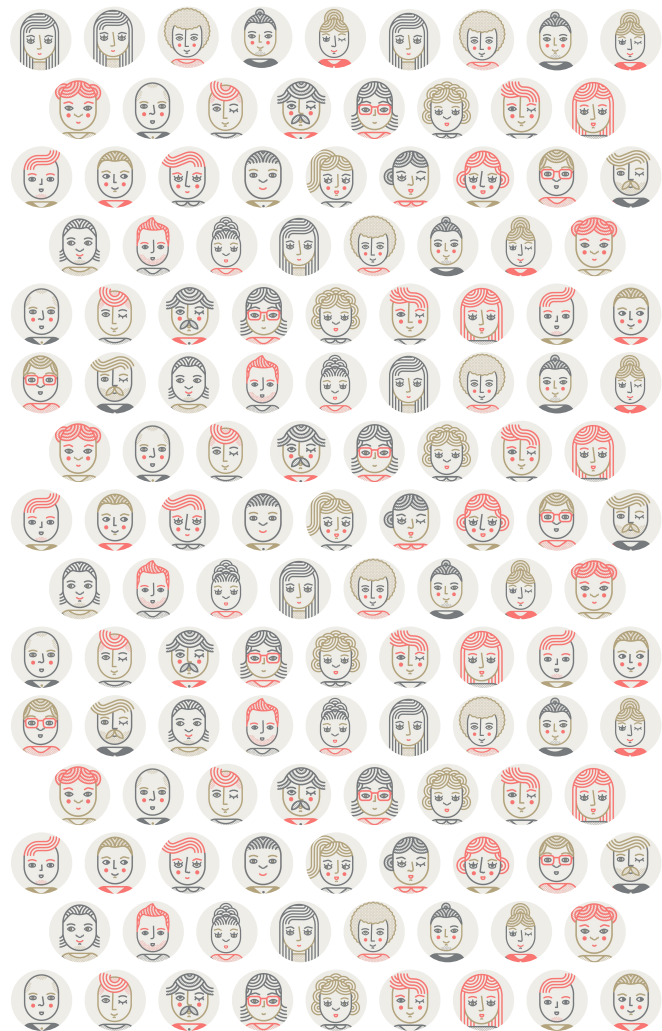
Step 2

“Please don’t break”

Organizational Challenges

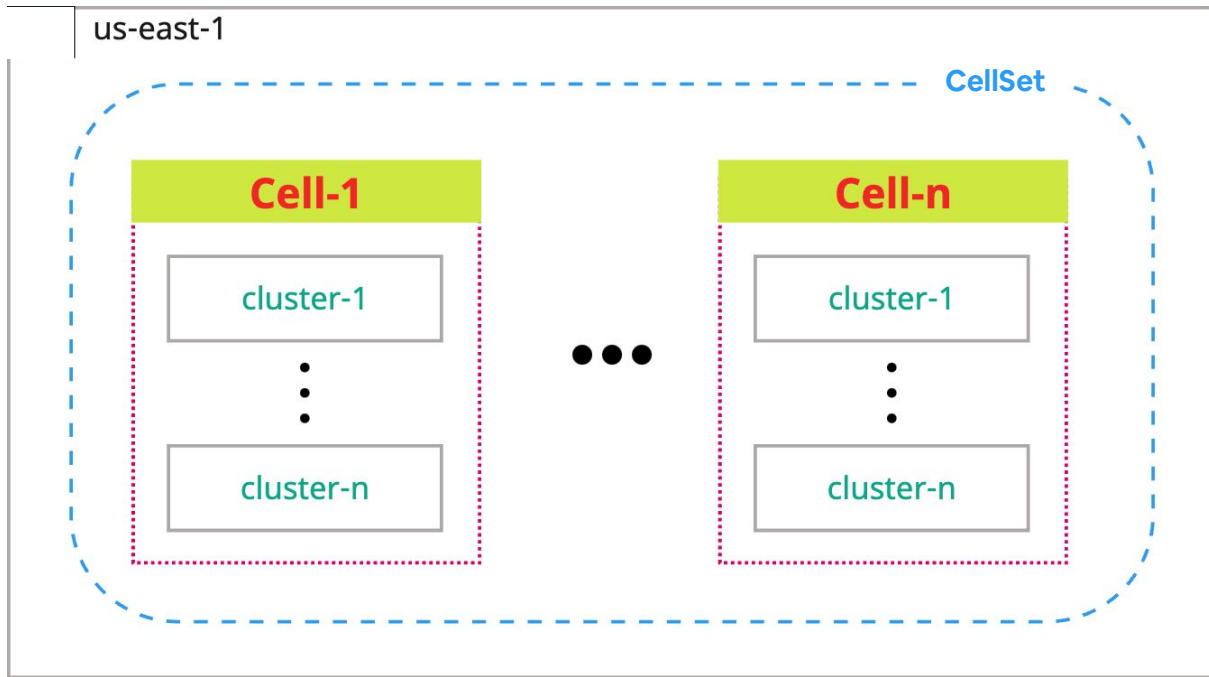


“Please trust us!”



Project Cellularization

Cells and CellSets: The Compute Abstraction



OneTouch Refactor and Improvement

OneTouch Refactor

>250

Pull Requests

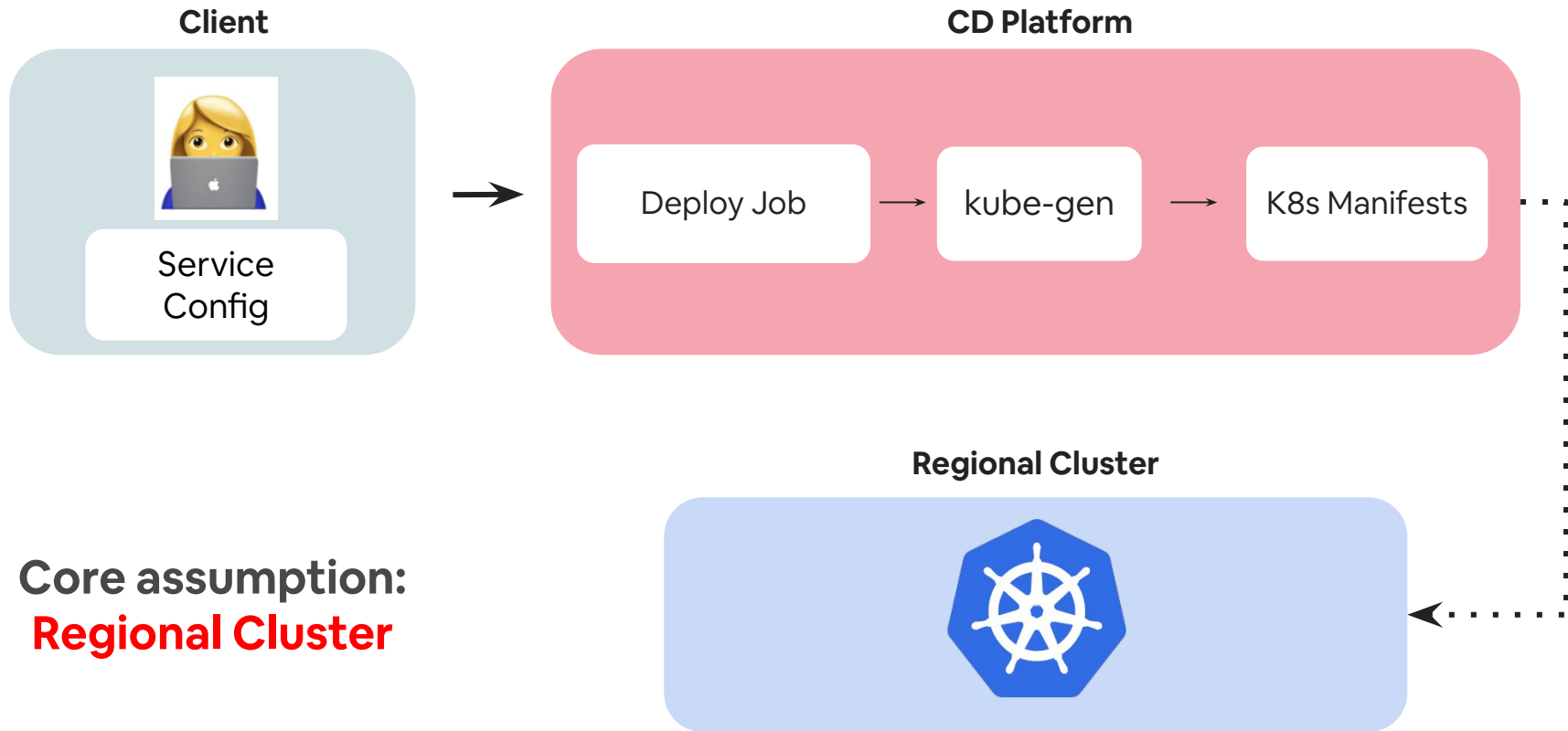
~50K

Lines of code **added**

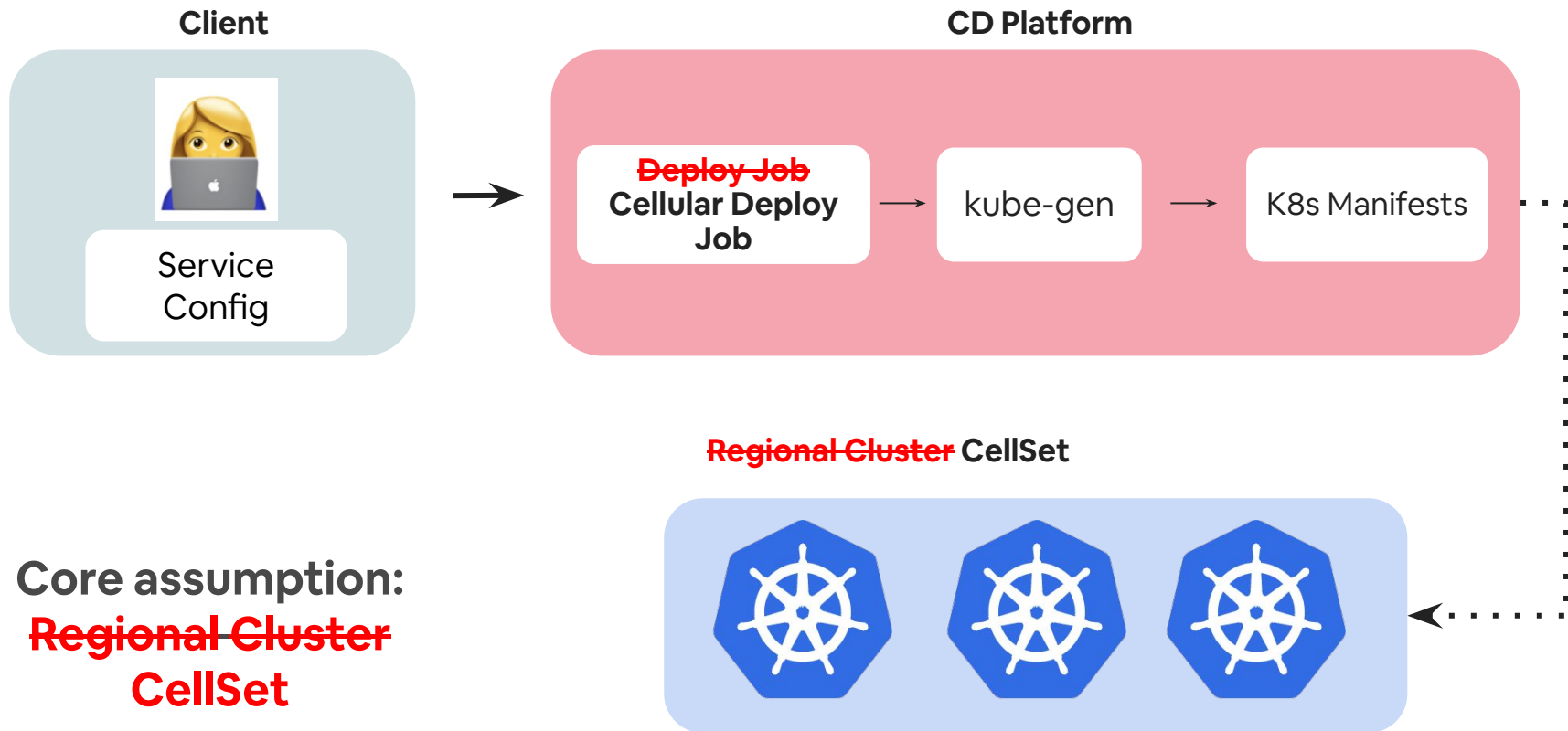
~12K

Lines of code **removed**

OneTouch System



OneTouch System



Config Updates

OneTouch Changes

```
environments:
  production:
    - context: cluster-1
    + cellset: us-east-1
    params:
      - minReplicas: 9
      - maxReplicas: 27
    + minReplicas: 3 # per cell
    + maxReplicas: 9 # per cell
```

CD Pipeline Changes

```
stages:
  - name: deploy-to-prod
    - type: onetouchDeploy
    + type: cellularDeploy
    parameters:
      - context: cluster-1
      + cellset: us-east-1
    environment: prod
```

Tooling Updates

```
$ k status --env staging
```

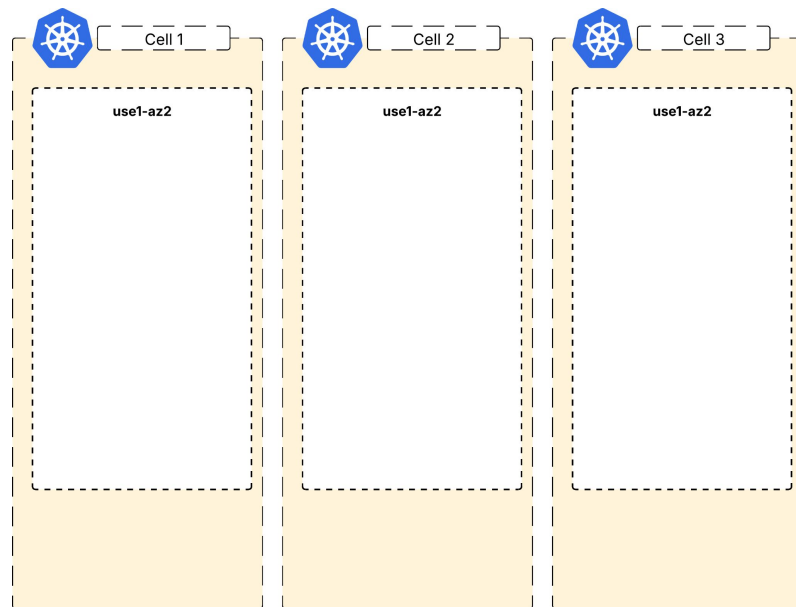
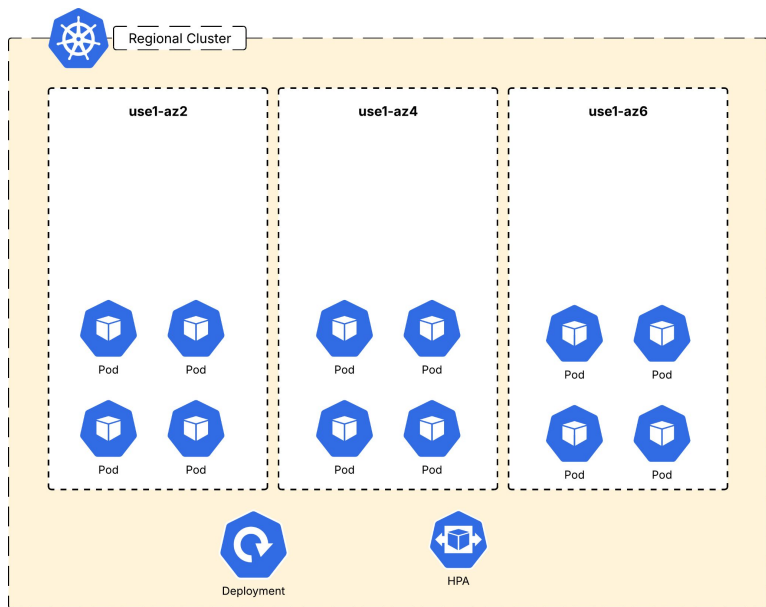
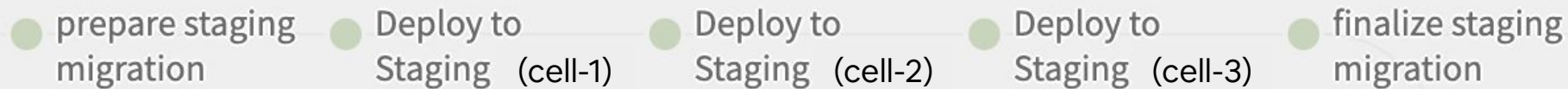
```
...
```

NAME	READY	STATUS	RESTARTS	AGE	CELL
banana-staging-5d47f94f68-6smcz	13/13	Running	0	2d5h	cell-1
banana-staging-5d47f94f68-vhkdb	13/13	Running	0	2d5h	cell-1
banana-staging-65bd5f6b9d-cnjft	13/13	Running	0	2d5h	cell-2
banana-staging-65bd5f6b9d-v2j8f	13/13	Running	0	2d5h	cell-2
banana-staging-7fgier9tud-djvi9	13/13	Running	0	2d5h	cell-3
banana-staging-7fgier9tud-wqyc1	13/13	Running	0	2d5h	cell-3

Dashboard Updates



Automated Service Migration



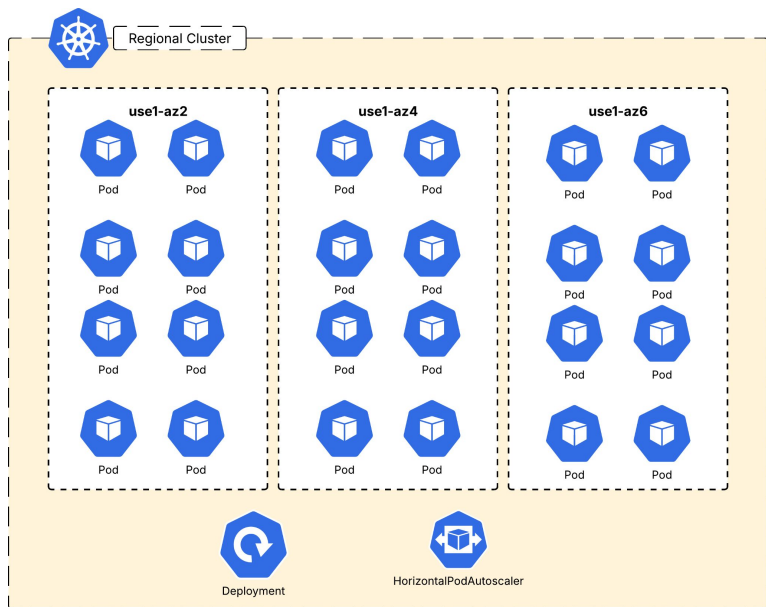
● prepare staging migration

● Deploy to Staging (cell-1)

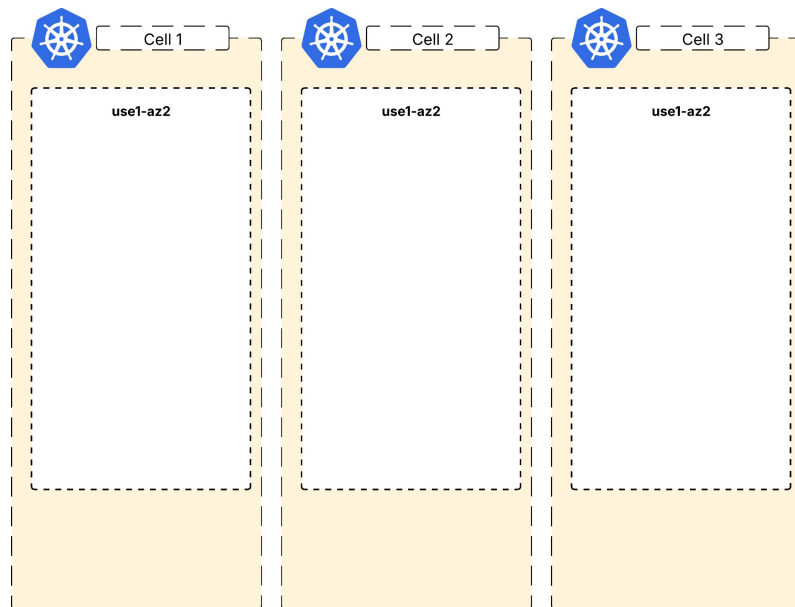
● Deploy to Staging (cell-2)

● Deploy to Staging (cell-3)

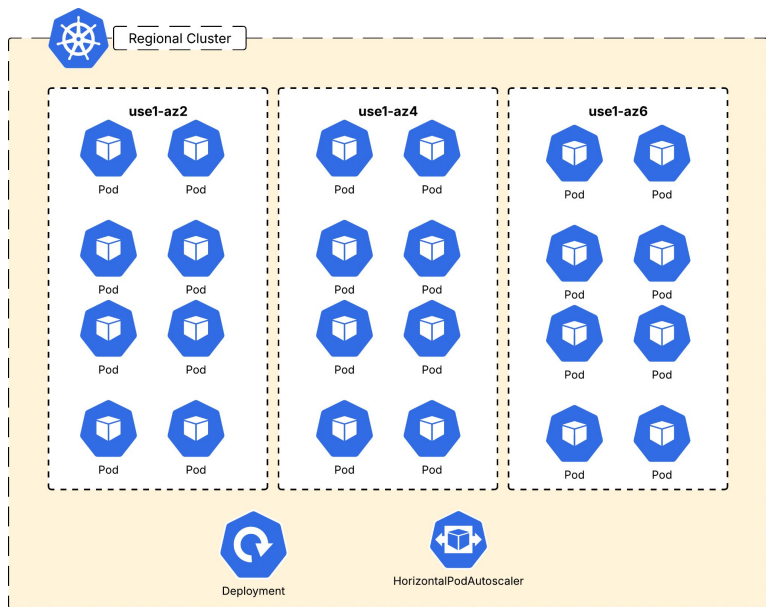
● finalize staging migration



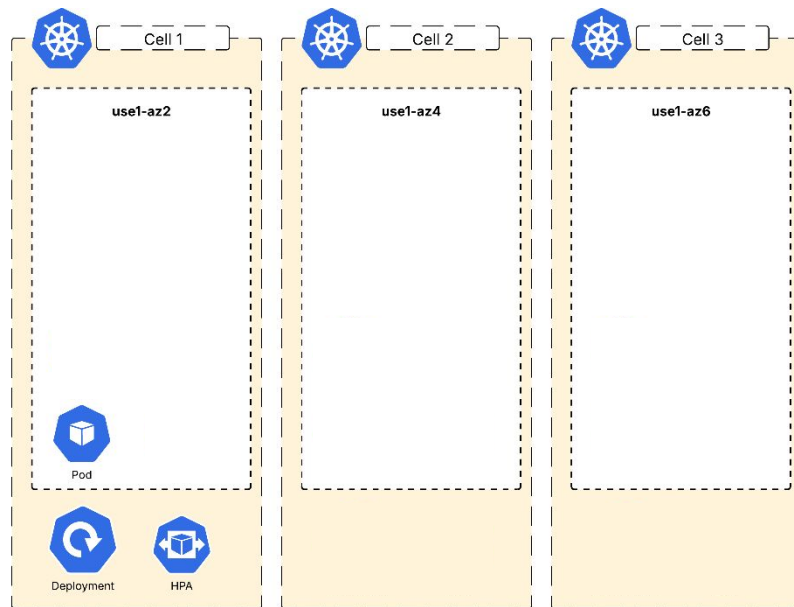
↑
Scale up regional

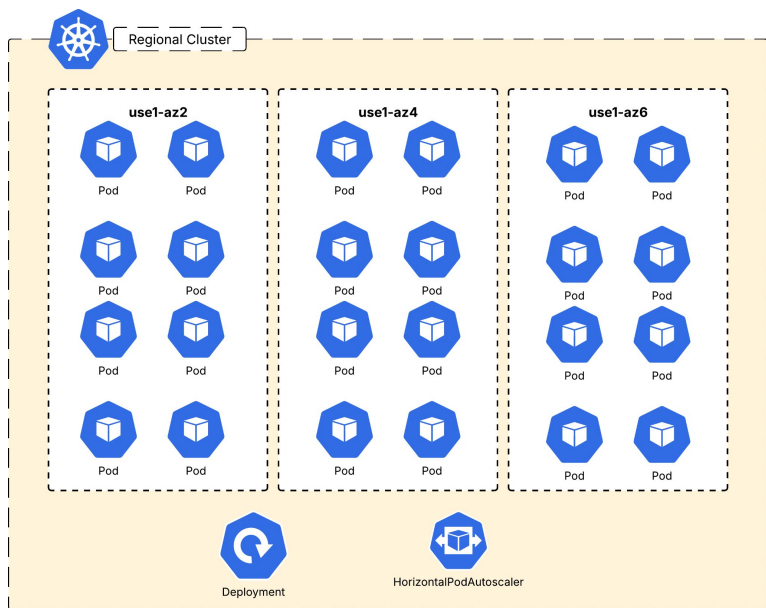
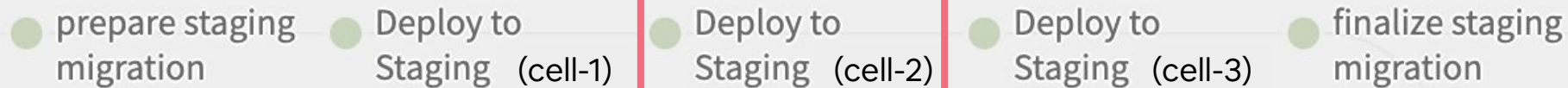


- prepare staging migration
- **Deploy to Staging (cell-1)**
- Deploy to Staging (cell-2)
- Deploy to Staging (cell-3)
- finalize staging migration

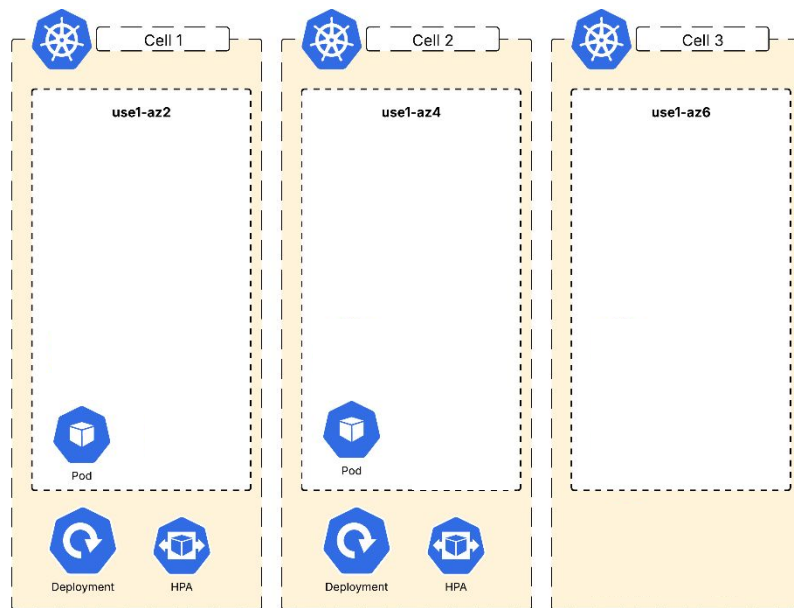


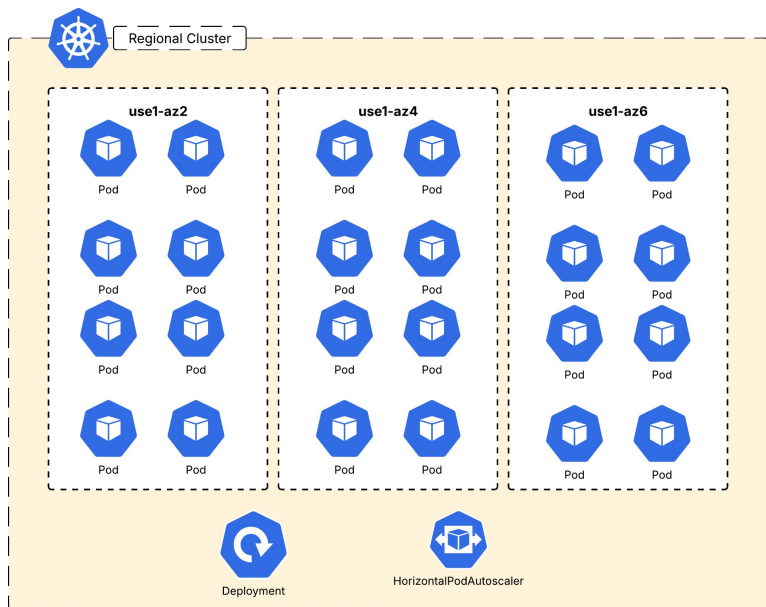
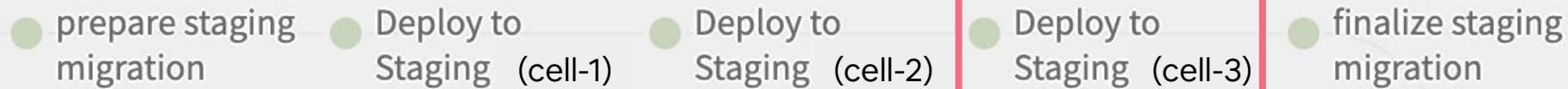
Deploy
Cell-1



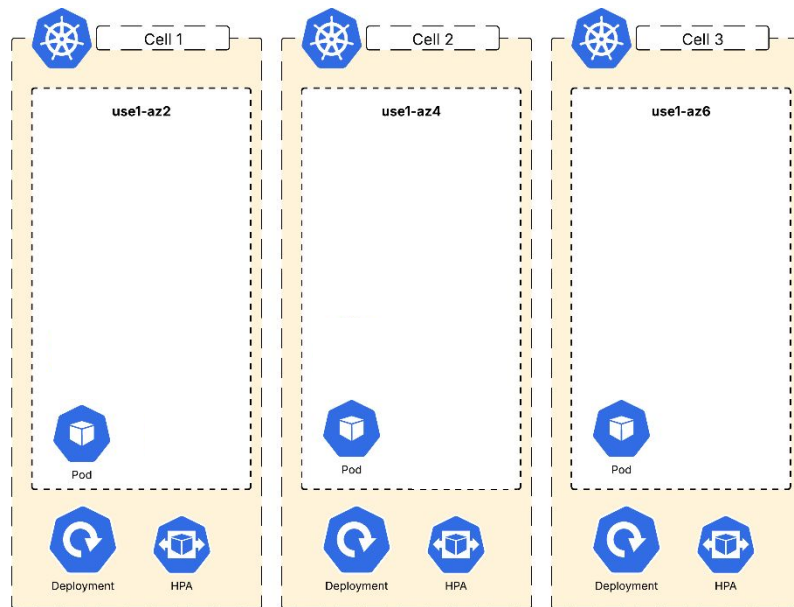


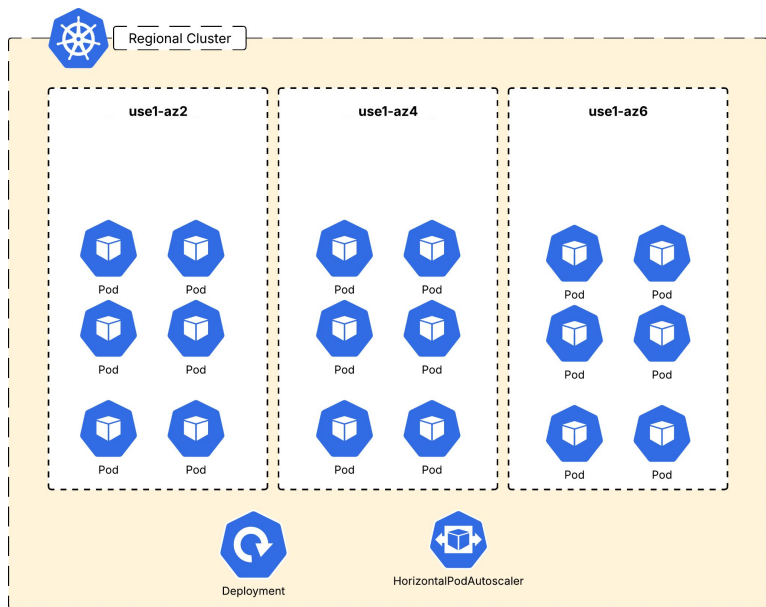
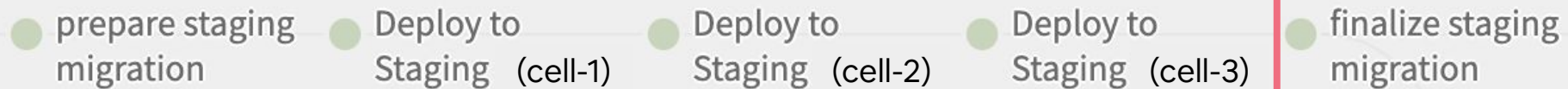
Deploy
Cell-2





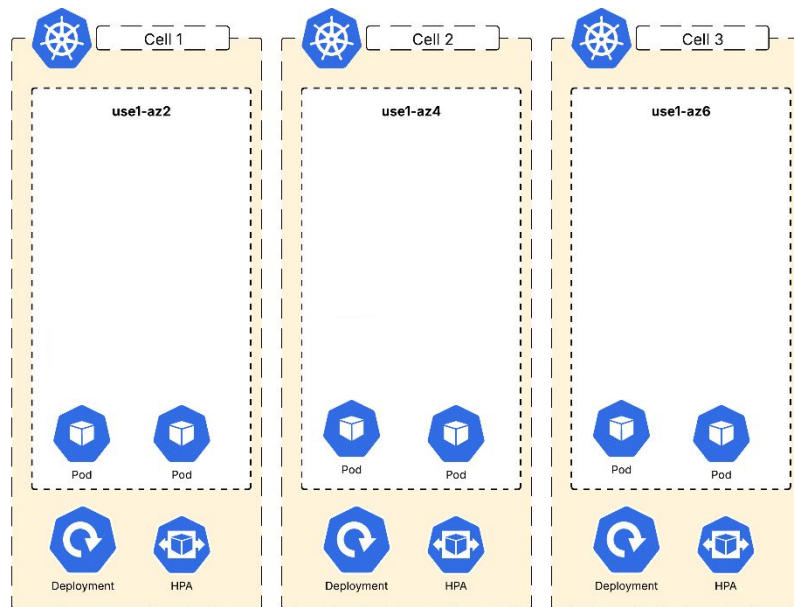
Deploy
Cell-3

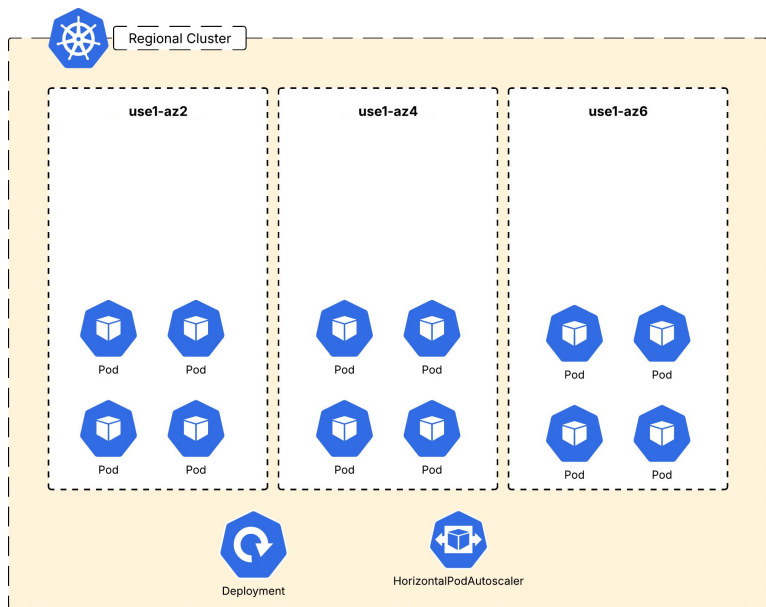
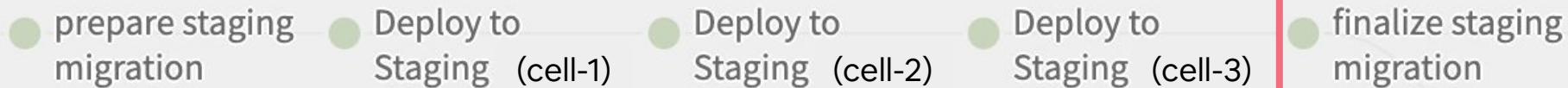




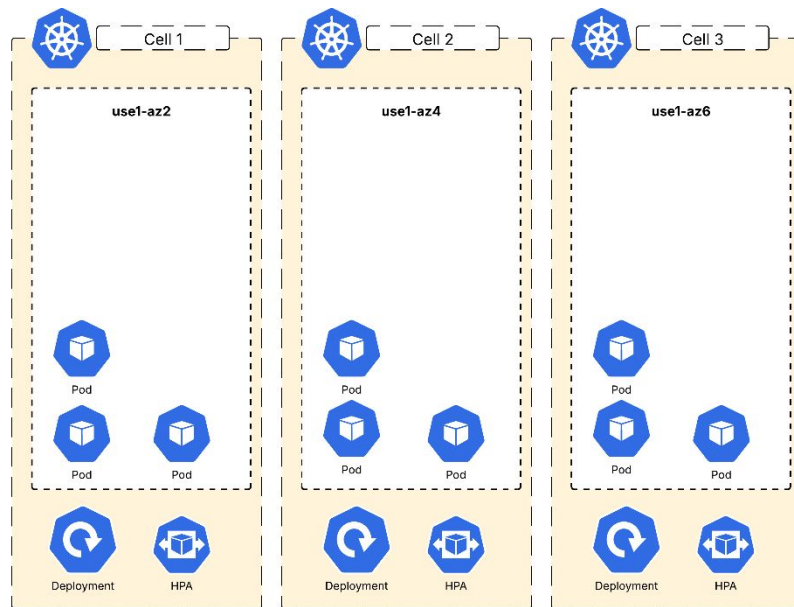
↓

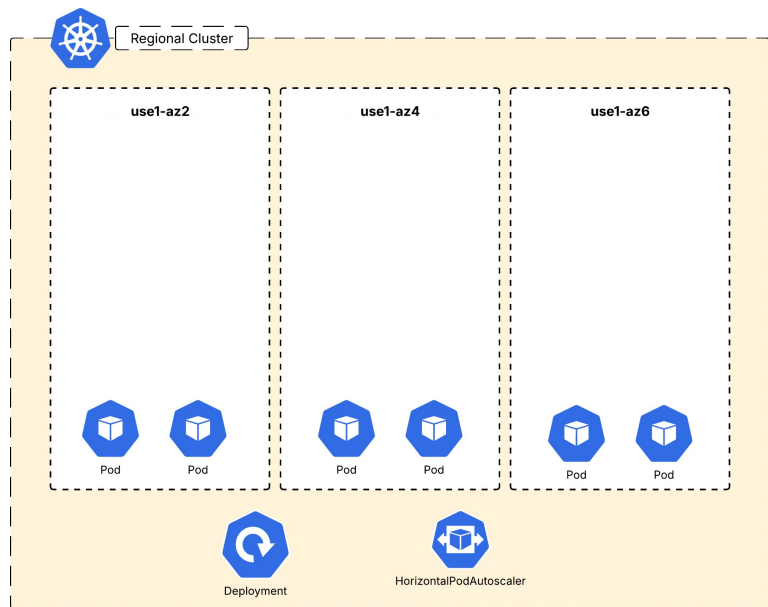
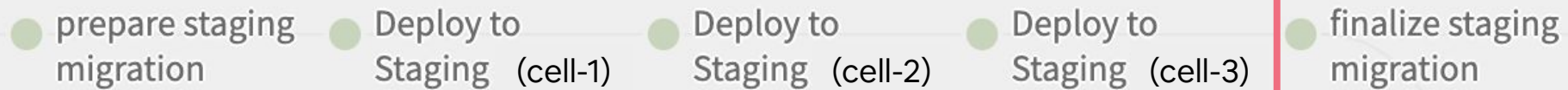
Scale
down
regional
75%





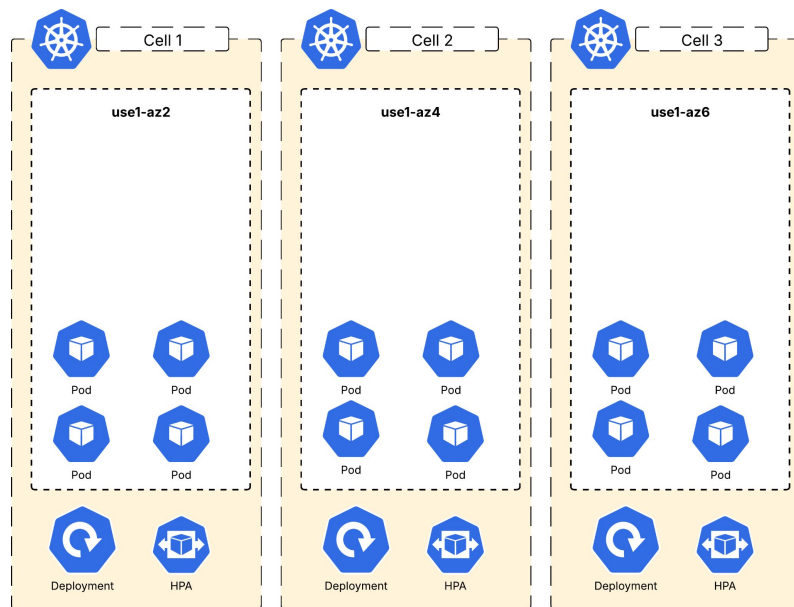
Scale
down
regional
50%

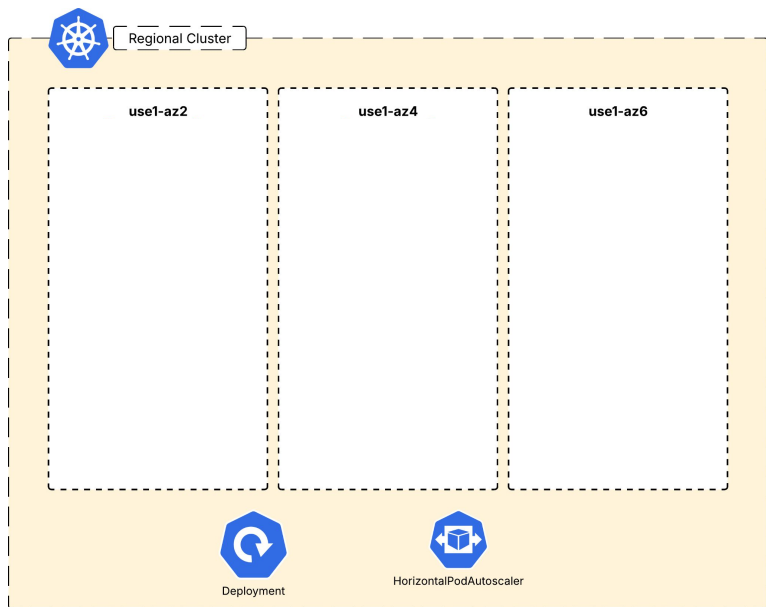
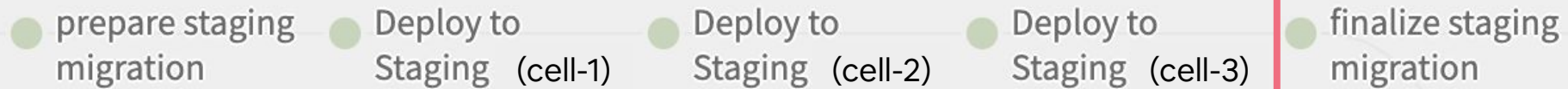




↓

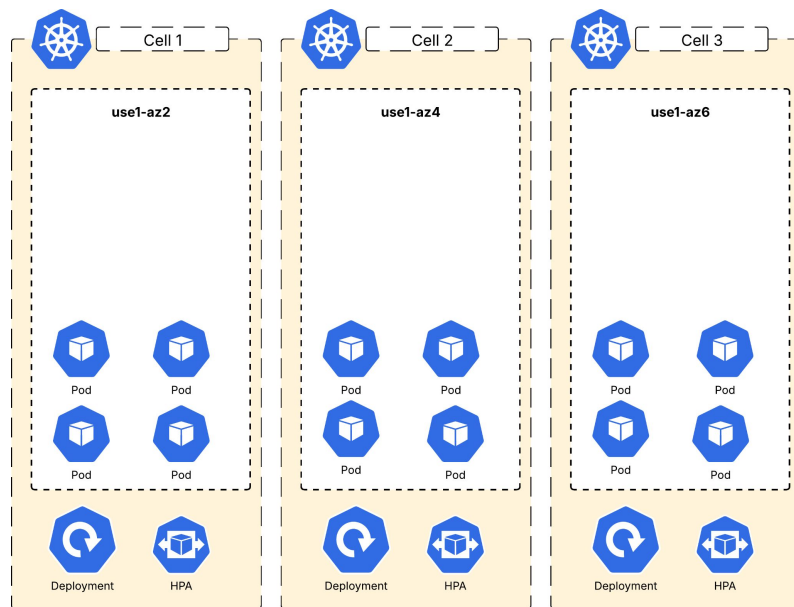
Scale
down
regional
25%



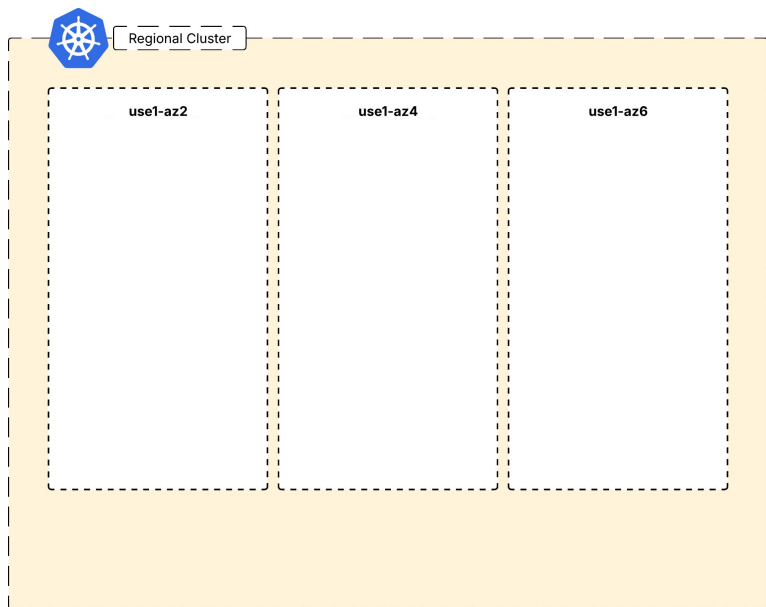


↓

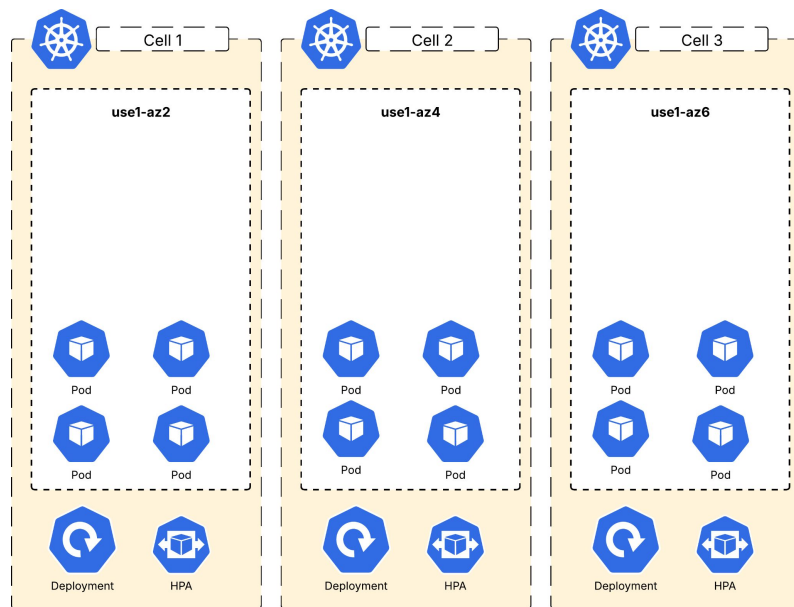
Scale
down
regional
0%



- prepare staging migration
- Deploy to Staging (cell-1)
- Deploy to Staging (cell-2)
- Deploy to Staging (cell-3)
- finalize staging migration



Cleanup



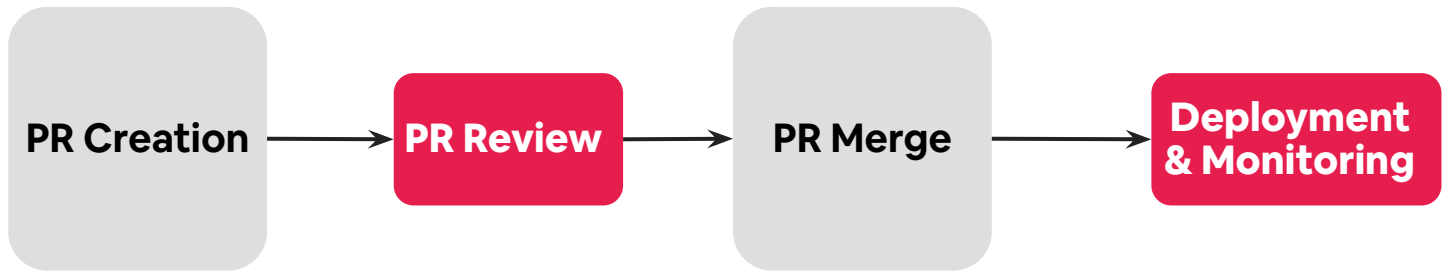
Spreading the message

Tech L&D

Cellularization

The Cornerstone to Zonal Resilience, and How it Affects You

HOST: Jordan Ly



We* create the PR to migrate environment(s) to cells and notify in shared slack channel.

Service team reviews and approves the PR.

PRs Automatically Approved and Merged After 3-Business Days

We* merge the PR after approval.

Service team deploys via regular CD deployment & monitors.

Documented everything

Cellularization

Home

High-Level Overview

[Overview & Benefits](#)

Current State

Future State

Key Differences

Migration Guide

Read This First!

What Happens During a Migration

Migration Operations Guide

Self-Service Migration Guide

Failure Scenarios & Troubleshooting

Rollback Guide

FAQ

Roadmap, Status & Documentation

Contact

2025 Schedule



Scope = Stateless OT Projects

Execute

Beginning of 2025

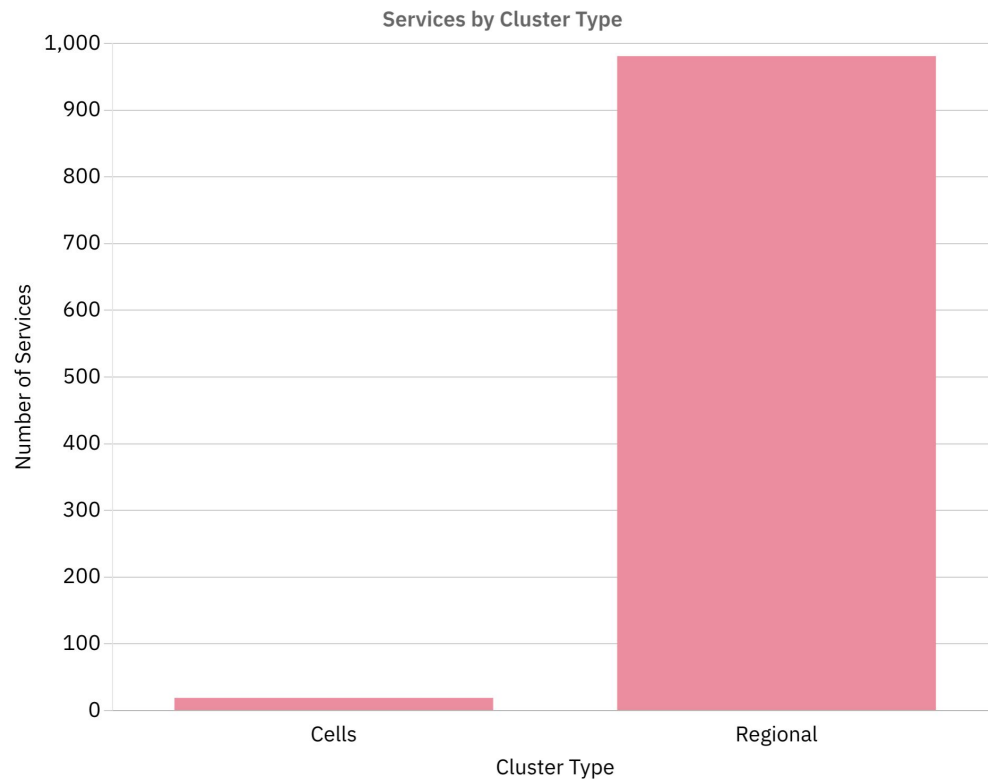
19

Services Migrated

30%

Total Compute on Cells

Beginning of 2025



Batch Sizes by Service Criticality

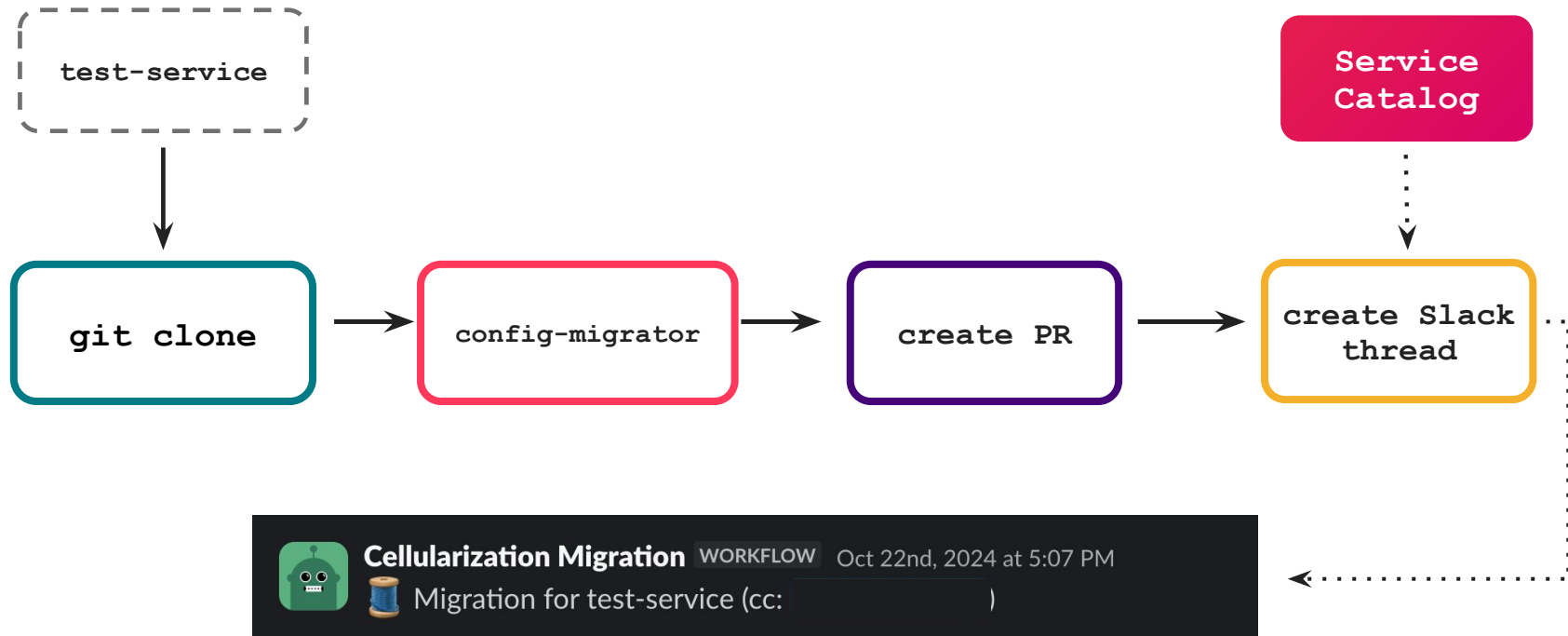
10-20

Tier-0 or Tier-1

30-40

Tier-2+

Automated Pull Requests (via Refactorator)



Initial Thread Message



Cellularization Migration

WORKFLOW

Oct 22nd, 2024 at 5:07 PM



Migration for test-service (cc:

)

Nonprod (staging, test, etc)



refactorator APP

Preparing stage: nonprod

PR: <https://git.musta.ch/airbnb/treehouse/pull/928075>

Please review the next stage of the migration and approve the PR if it looks good. If you have any questions, please reply here.

Partialprod (canaries)



refactorator APP

Preparing stage: partialprod

PR: <https://git.musta.ch/airbnb/treehouse/pull/850042>

Please review the next stage of the migration and approve the PR if it looks good. If you have any questions, please reply here.

Production



refactorator APP

Preparing stage: prod

PR: <https://git.musta.ch/airbnb/treehouse/pull/851161>

Please review the next stage of the migration and approve the PR if it looks good. If you have any questions, please reply here.

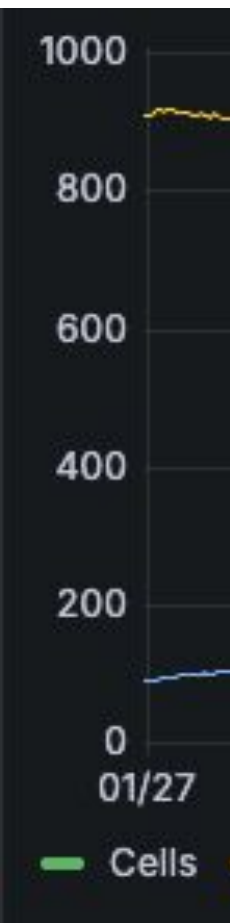
Underestimated the
support demands



Catch Up
99+ new

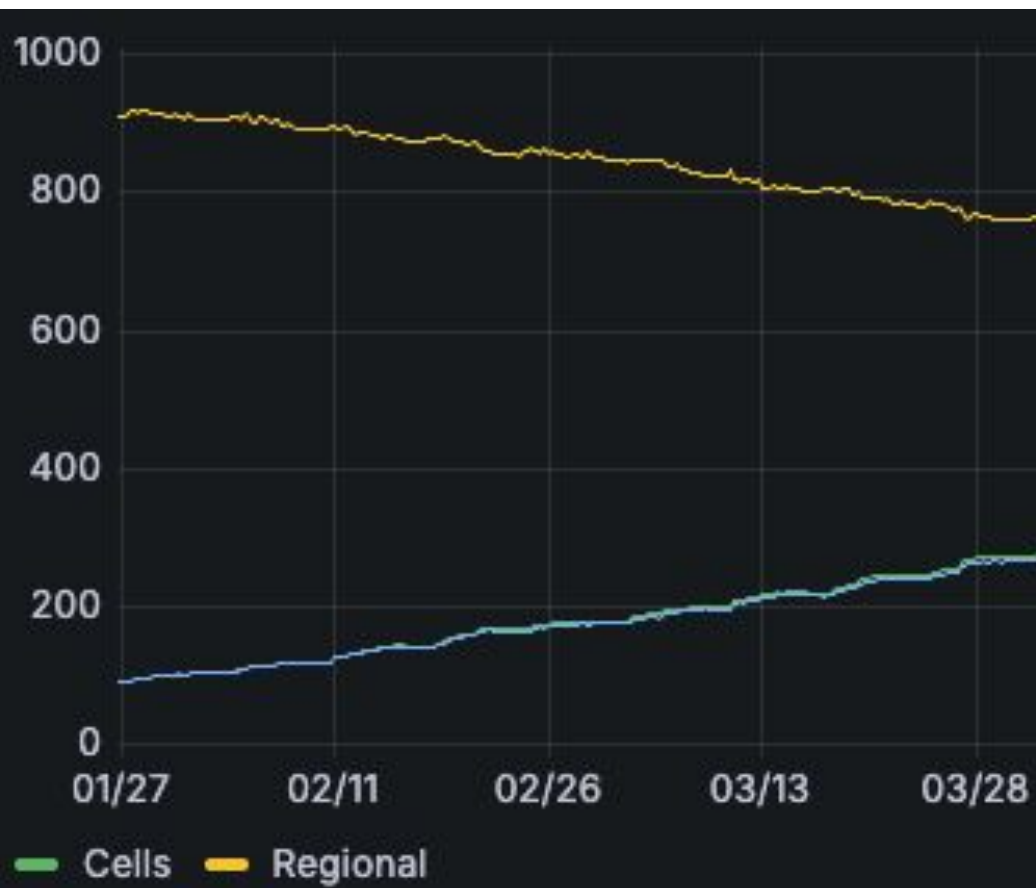


Threads
41 mentions

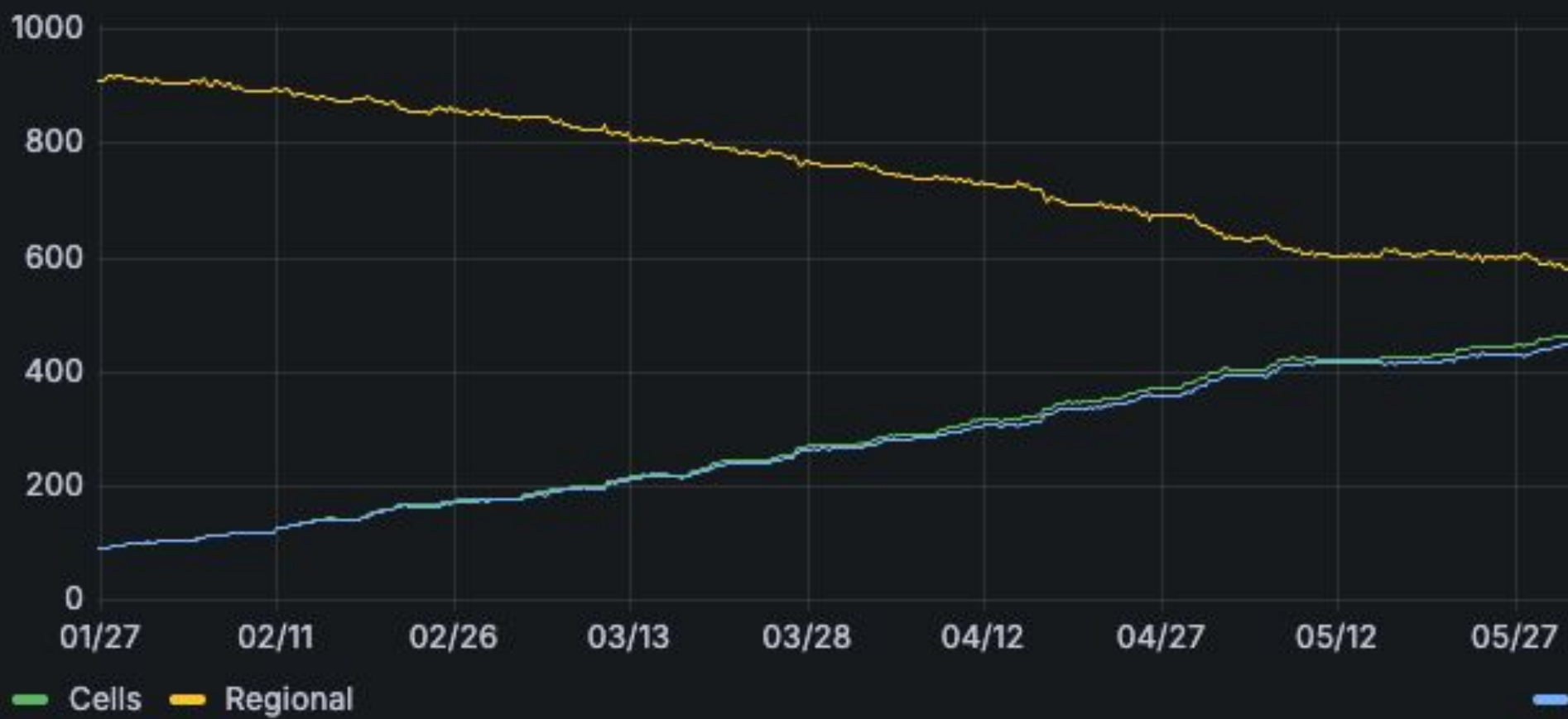


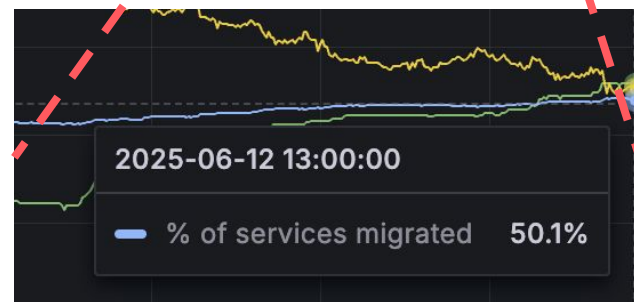
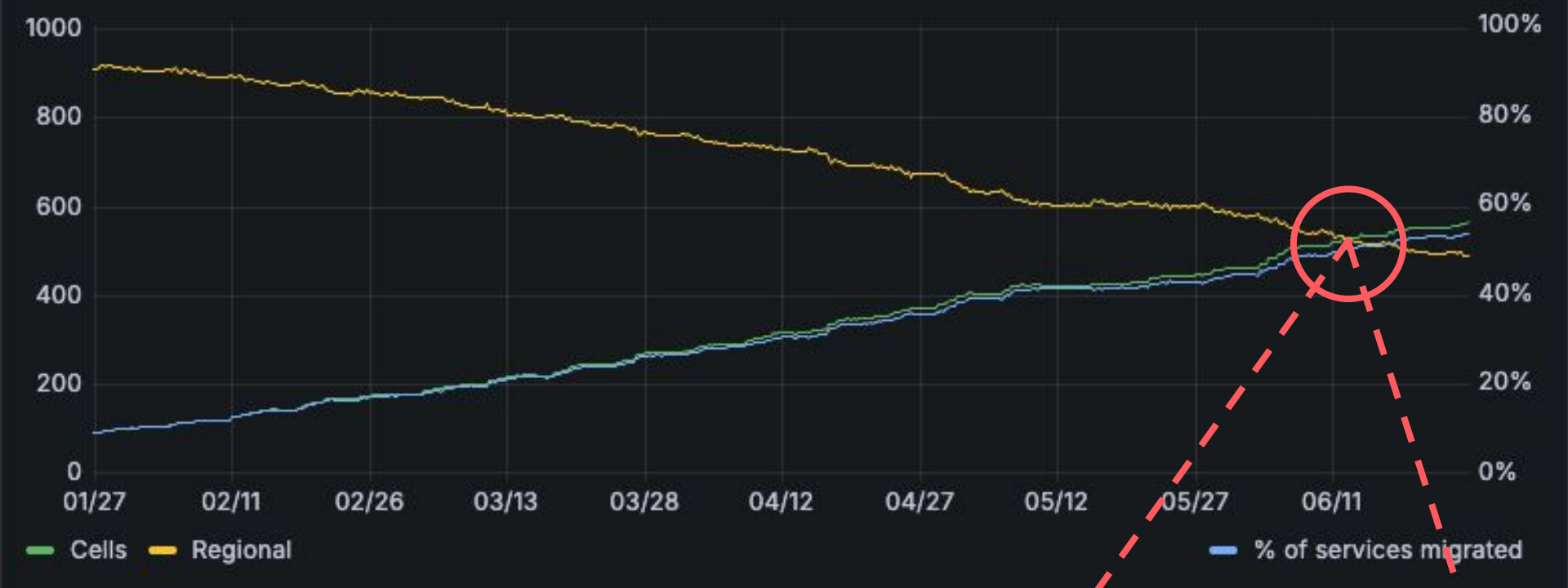
10% migrated by end of first month (Jan 2025)











Apr 10th, 2025 at 2:12 PM

Are we in the future already? LLM to do the migrations?



sunny EST

Apr 10th, 2025 at 2:14 PM

I wish lol, it's still me



3



1



1





Avg Service: Migration complete in 1-2 weeks

 **Cellularization Migration - Webhook** WORKFLOW 9:53 PM

 Migration for (cc:)

   10 replies

For Ω : Migrations could take months



Cellularization Migration (Tier 0) - Webhook

WORKFLOW

4:05 PM



Migration for

(cc:

)

293 replies



Cellularization Migration (Tier 0) - Webhook

WORKFLOW

3:23 PM



Migration for

(cc:

)

233 replies



Cellularization Migration

WORKFLOW

4:31 PM



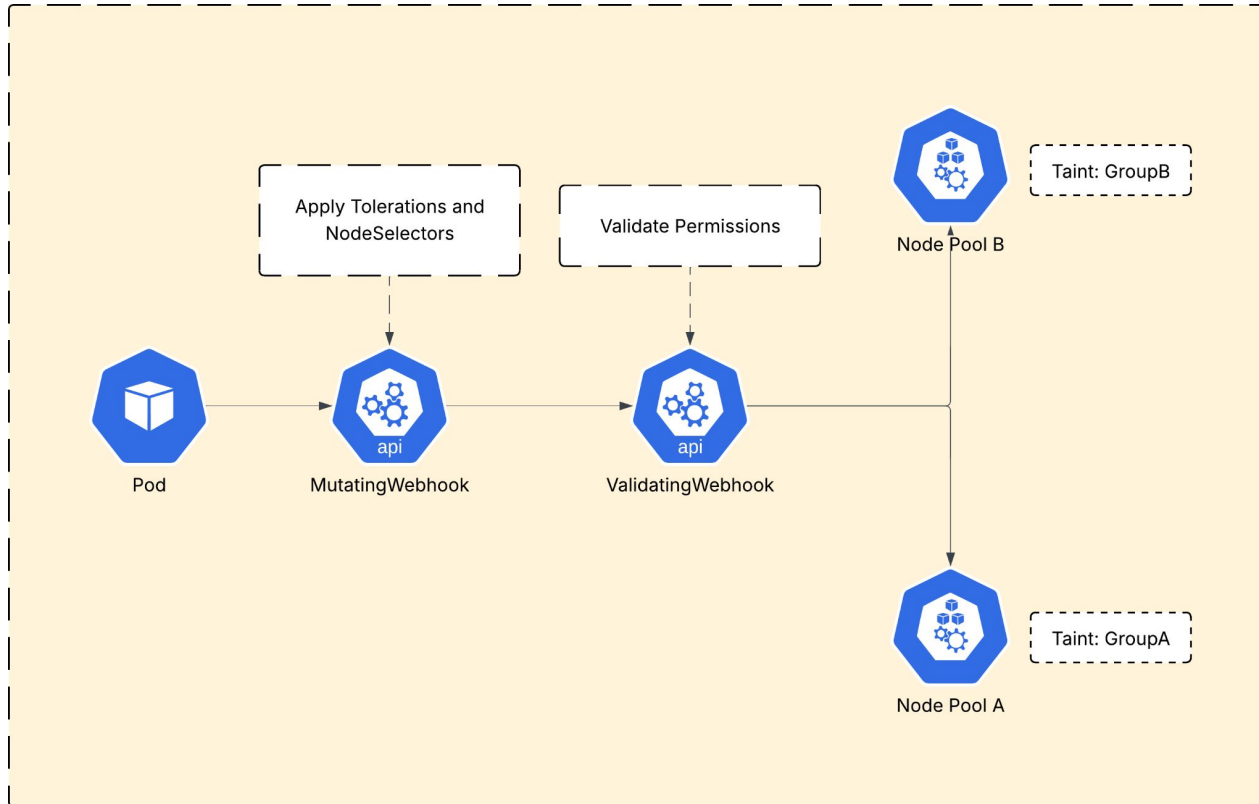
Migration for

(cc:

)

178 replies

Node Pools



Friday Night Deploys



sunny EST

Aug 15th, 2025 at 6:00 PM

heads up, Ω is **deploying to production** right now. Just figured I'd let others know

ETCD Capacity Upgrades

6:00 PM

Big thanks to the team for getting the updated etcd DB size upgraded from 8GB to 16GB! This was critical work to unblocking our progress on cellularization!



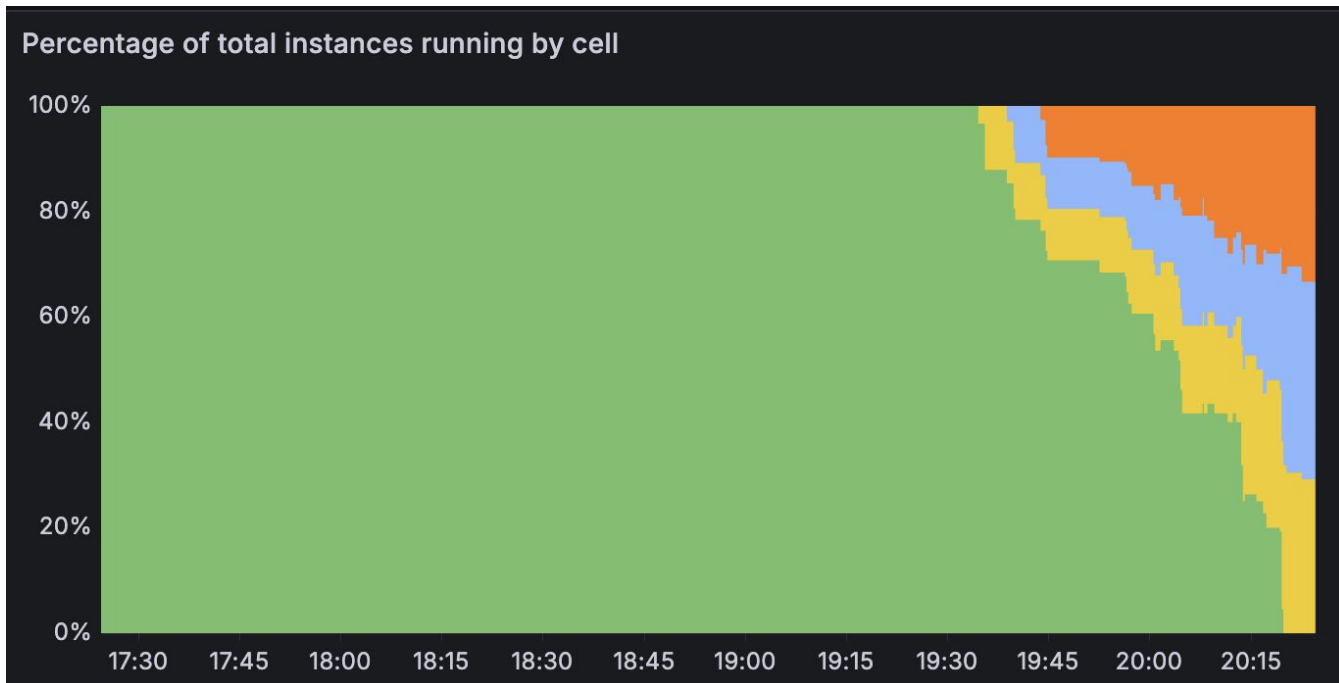
13



8



The final Ω migration





sunny EST

Ω

is officially on cells 🎉



4



2

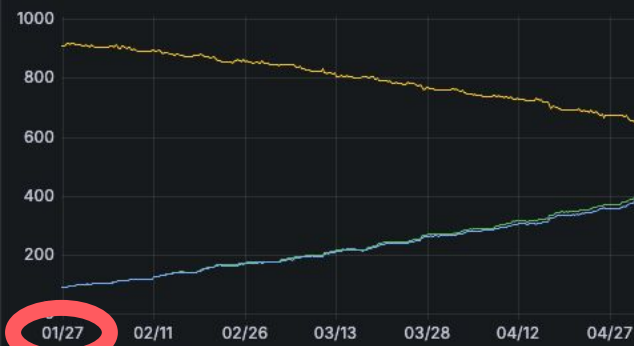


Regional
Compute
Decreased 40%



Regional cluster
deletion



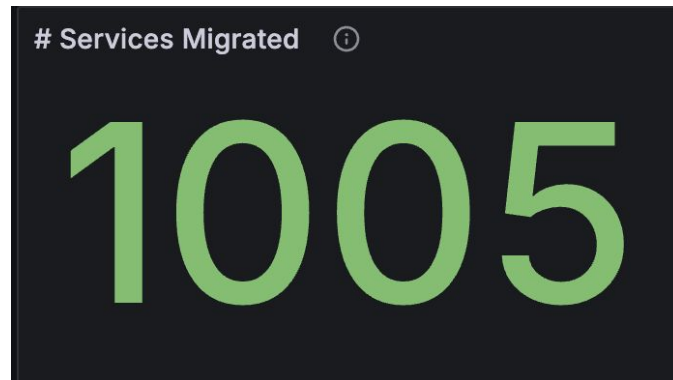


Celebrate

As of today

95%

Compute running in cells

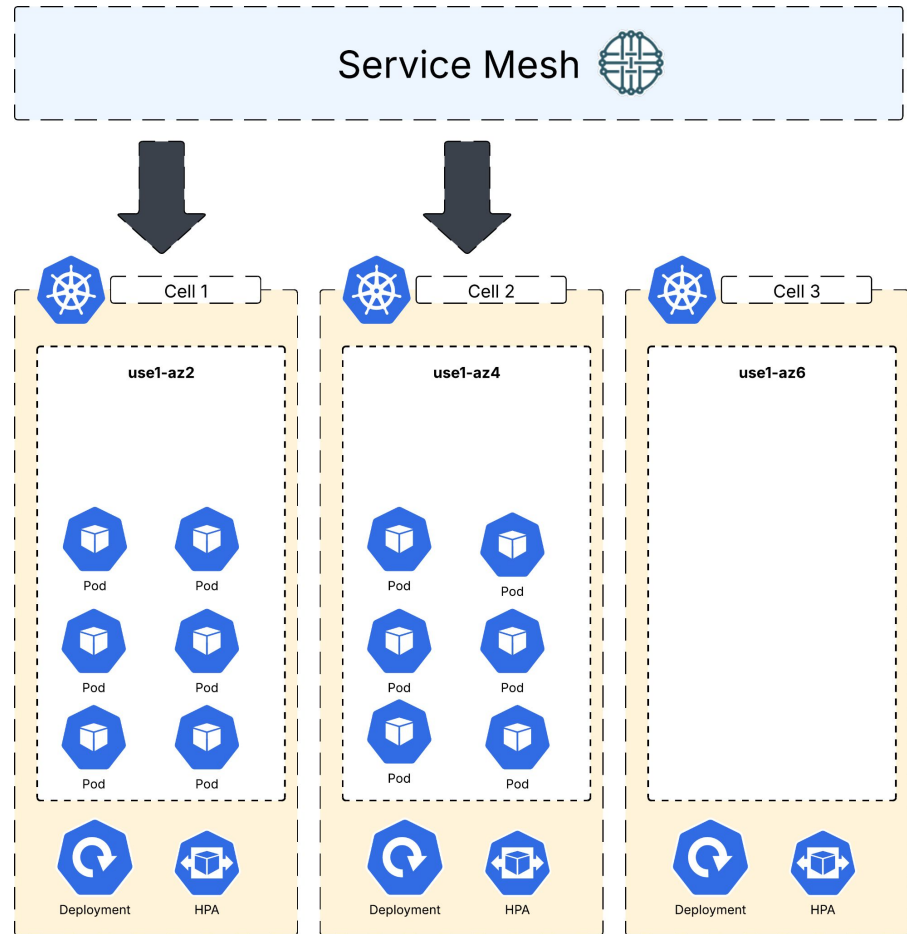


Services migrated

With 0 downtime!



Sept 2025:
First full zonal failover
in <10 mins



Still more work to do

“Service configs are too hard to maintain.”

“I have no idea what half the things do and whether they matter or not.”

“App teams don’t have the tooling to debug their service issues”

Better Debugging Tools

App: banana

[Projects](#) / [banana](#) / [production](#) / banana 

Pod Health

cell-1

cell-2

cell-3

Ready

Pending

Initializing

Terminating

Unhealthy

Cell

Cluster

Node

ReplicaSet

Deployments



Name 	Type 	App 	Git SHA 	Replicas	Status 	Strategy 	Canary Step 
banana-production	Deployment	banana	0f3fdfd	2/2	Available	RollingUpdate	-
banana-production	Deployment	banana	0f3fdfd	2/2	Available	RollingUpdate	-
banana-production	Deployment	banana	0f3fdfd	2/2	Available	RollingUpdate	-

Pods



Name 	App 	Ready	Restarts	Status 	Node 	Age 
banana-production-7554bcb798-l2mx2	banana	2/2	0	Running 	ip-100-119-5-157.ec2.internal	11h 
banana-production-7554bcb798-pfqhc	banana	2/2	0	Running 	ip-100-119-11-128.ec2.internal	4h 

Simplifying Configs

```
# Before
mainContainer:
  name: my-service
  command: [dumb-init]
  args:
    - /srv/run.sh
    - -Dfile.encoding=UTF-8
    - -Xms4g
    - -Xmx4g
    - -XX:MetaspaceSize=768m
    - --
    - server
    - /srv/my-service/config.yml
  ports:
    - containerPort: 8080
  resources:
    requests: {cpu: 4000m, memory: 12Gi}
    limits: {memory: 12Gi}
  startupProbe:
    httpGet: {path: /probe/started, port: 8080}
    failureThreshold: 60
    periodSeconds: 1
# ... readiness probe, liveness probe, fileMo
# ... components, rollout strategy, airmesh s
```

YAML Lines in Total



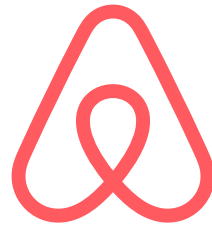
Workload Placement Controllers for PR-less Migrations



dlow 6:13 PM

does anyone run a tier3 OneTouch services that would like to try one of our fancy new EKS clusters? we'll do 100% of the work for you, you won't even need to look at a PR (edited)





Please leave
feedback