ArgoCon

EUROPE

Argo Rollouts + HPA

ArgoCon EUROPE

1 April 2025

London, England

Anastasiia Gubska

CNCF Ambassador

SRE/DevOps

BT Group

Kapelonis Kostis

Argo Team

Developer Advocate

Octopus Deploy

QUIZ

Are you familiar with?

ArgoCon EUROPE

Argo Rollouts

2619

Advanced Kubernetes deployment strategies such as Canary and Blue-Green made easy.

Learn More

Using in Production?

Argo Rollouts

☆ 2619

Advanced Kubernetes deployment strategies such as Canary and Blue-Green made easy.

Learn More

What does this affect?

ArgoCon
EUROPE

previewReplicaCount: 5

What does this affect?

ArgoCon
EUROPE

dynamicStableScale: true

Are they the same thing?

ArgoCon
EUROPE

setWeight: 20

setCanaryScale:
weight: 20

Who will win?

Argo Rollouts

Horizontal Autoscaler

Agenda

ArgoCon
EUROPE

1. What is Argo Rollouts
2. How the Horizontal Pod Autoscaler works
3. The cost problem for Rollouts
4. Why use Argo Rollouts + HPA
5. Use cases and different scenarios with HPA

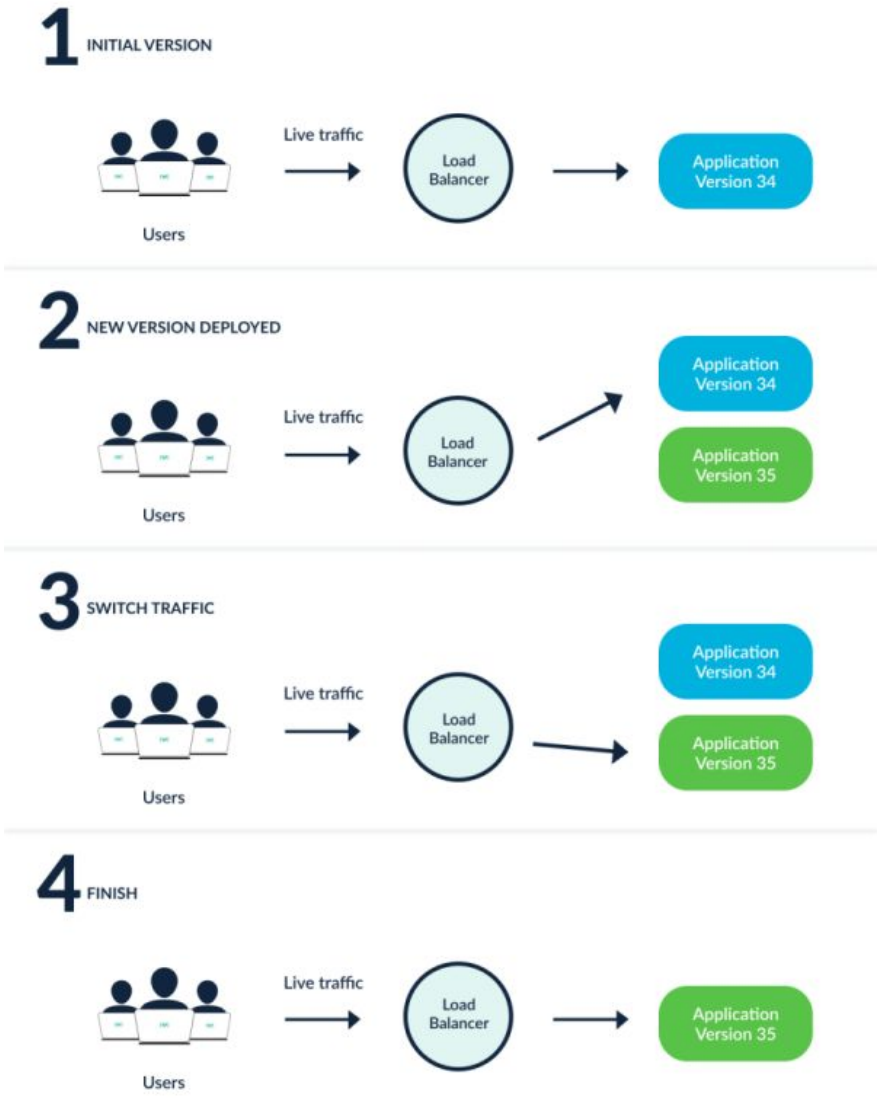
ArgoCon
EUROPE

Argo Rollouts

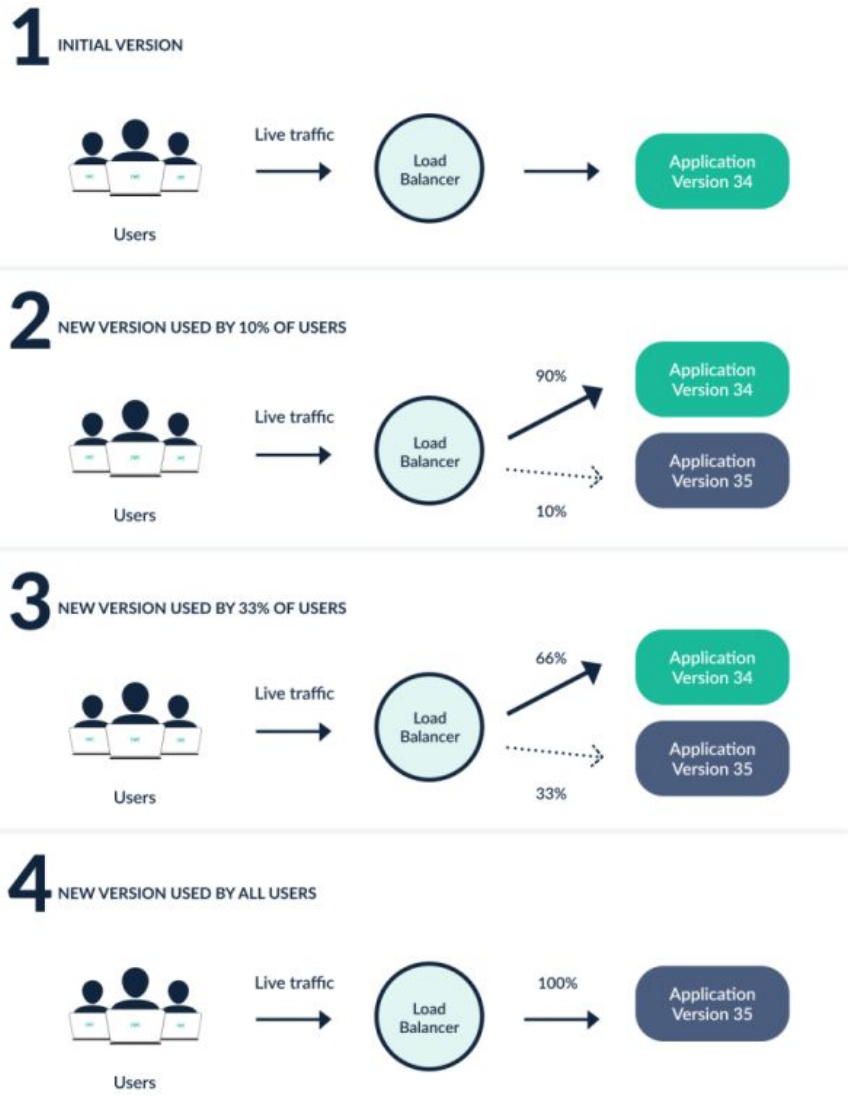
Progressive Delivery strategies

ArgoCon EUROPE

Blue/Green Deployment



Canary Release

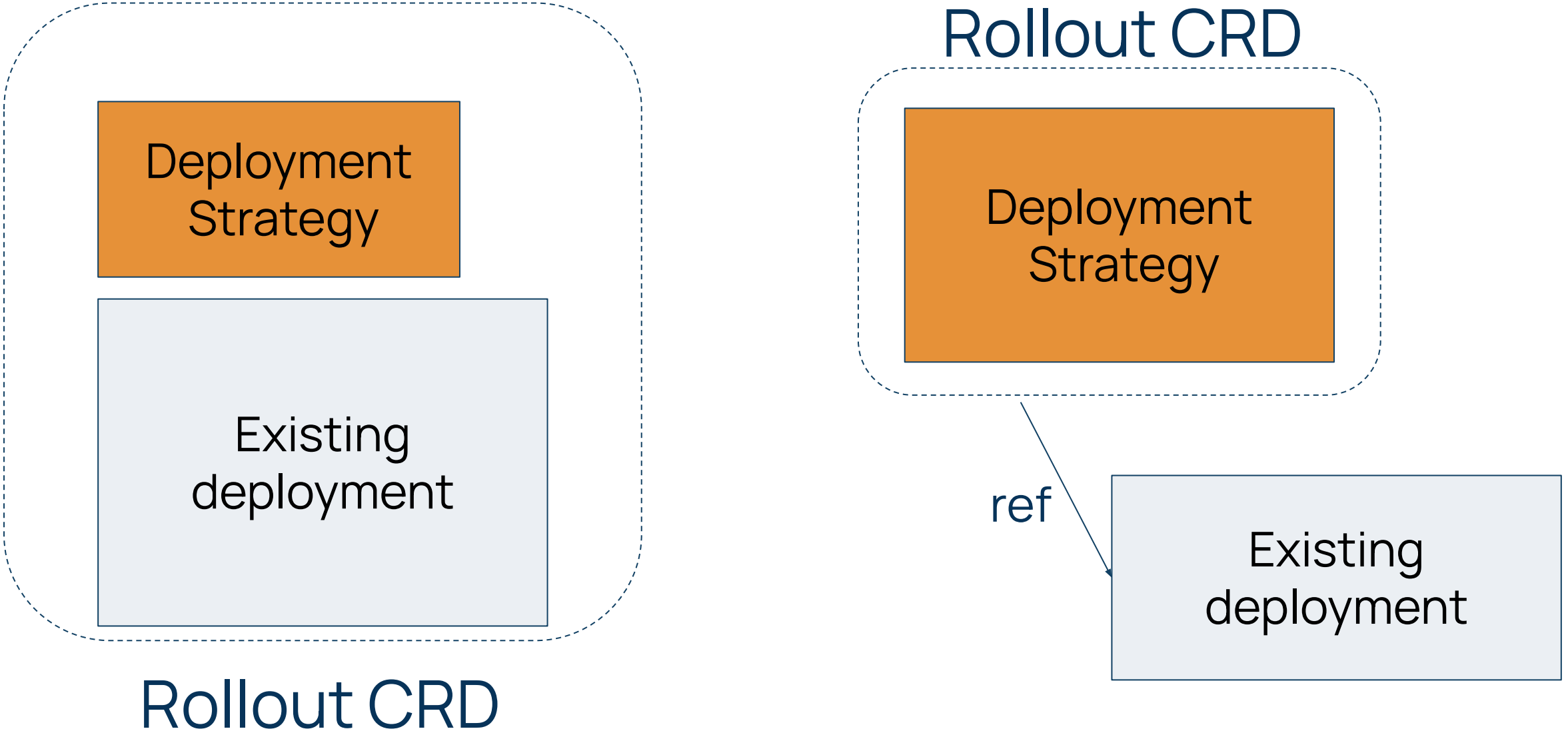


Argo Rollouts

ArgoCon
EUROPE

- Kubernetes native
- Standalone project
- Does **NOT** depend on Argo CD
- Blue/Green support
- Canary support
- A/B testing and other Experiments
- Zero downtime releases
- Automatic rollbacks based on metrics
- Installed on each deployment cluster

How Rollouts work



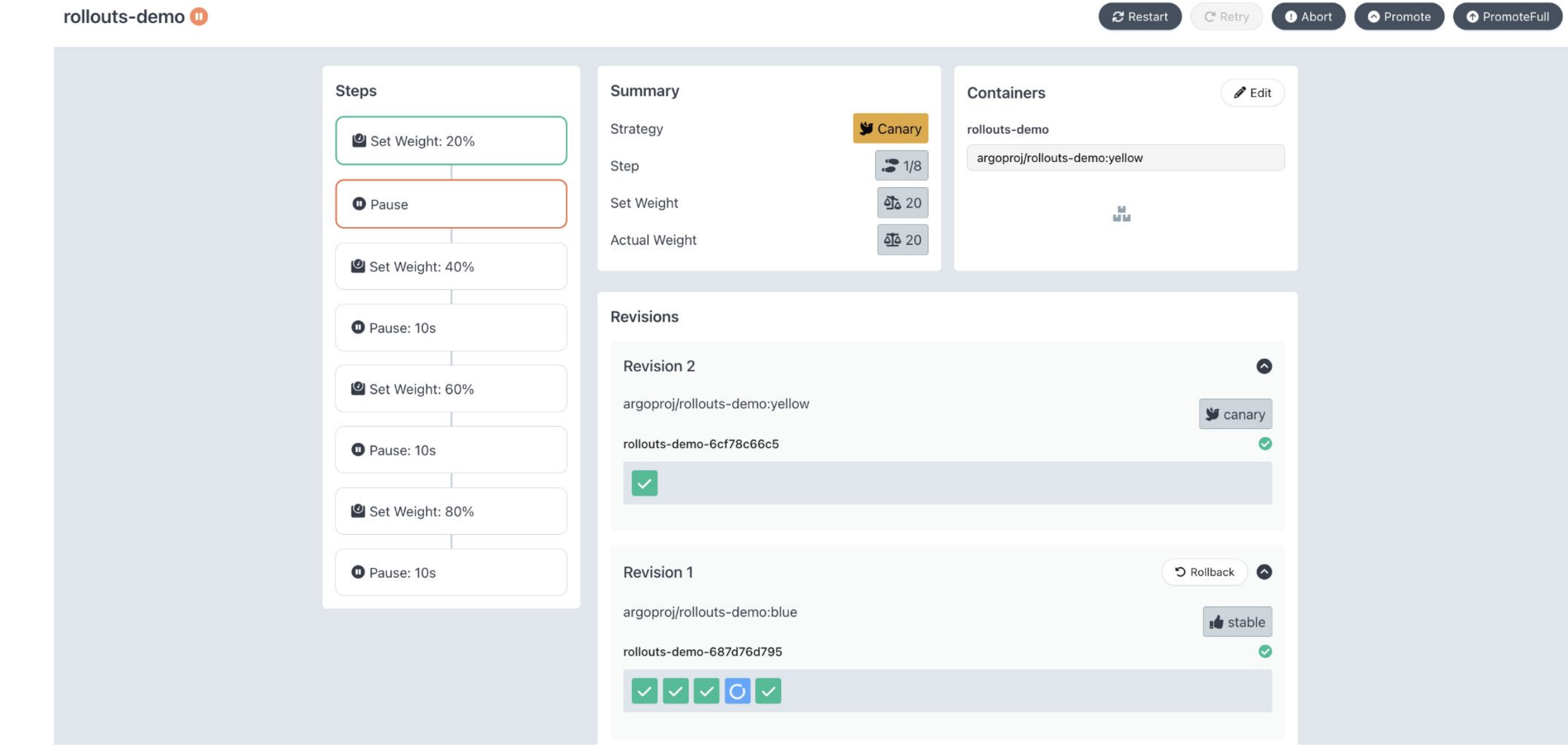
Rollout manifest


```yaml
apiVersion: argoproj.io/v1alpha1
kind: Rollout
metadata:
  name: canary-example
spec:
  replicas: 10
  revisionHistoryLimit: 2
  selector:
    matchLabels:
      app: cost-demo
  template:
    metadata:
      labels:
        app: cost-demo
    spec:
      containers:
        - name: cost-demo
          image: ghcr.io/kostis-codefresh/rollouts-autoscaling-example:v1
          ports:
            - name: http
              containerPort: 8080
              protocol: TCP
  strategy:
    canary:
      canaryService: argo-rollouts-canary-service
      stableService: argo-rollouts-stable-service
      steps:
        - setWeight: 10
        - pause:
            duration: 1h
        - setWeight: 20
        - pause: {}
```

```yaml
strategy:
  canary:
    canaryService: argo-rollouts-canary-service
    stableService: argo-rollouts-stable-service
    steps:
      - setWeight: 10
      - pause:
          duration: 1h
      - setWeight: 20
      - pause: {}
```

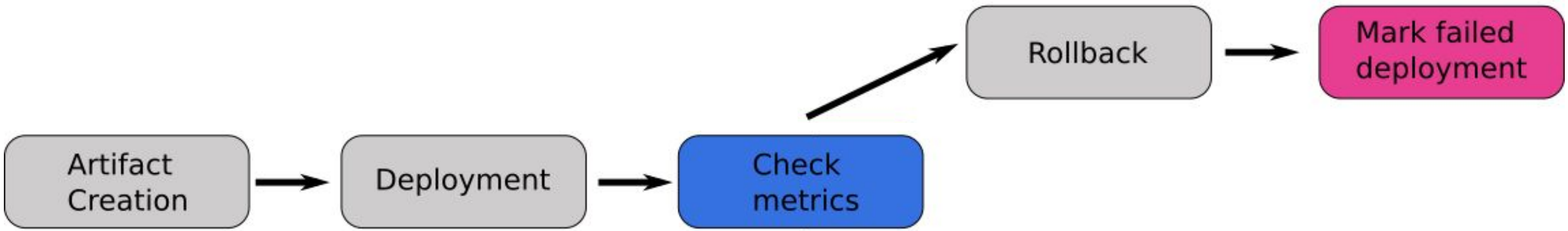
Rollout = deployment PLUS strategy

Graphical dashboard

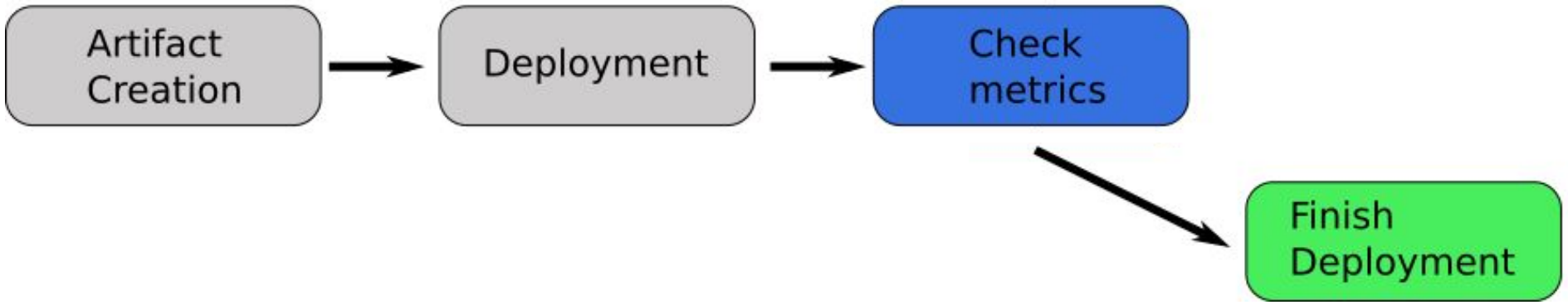


Rollback automatically

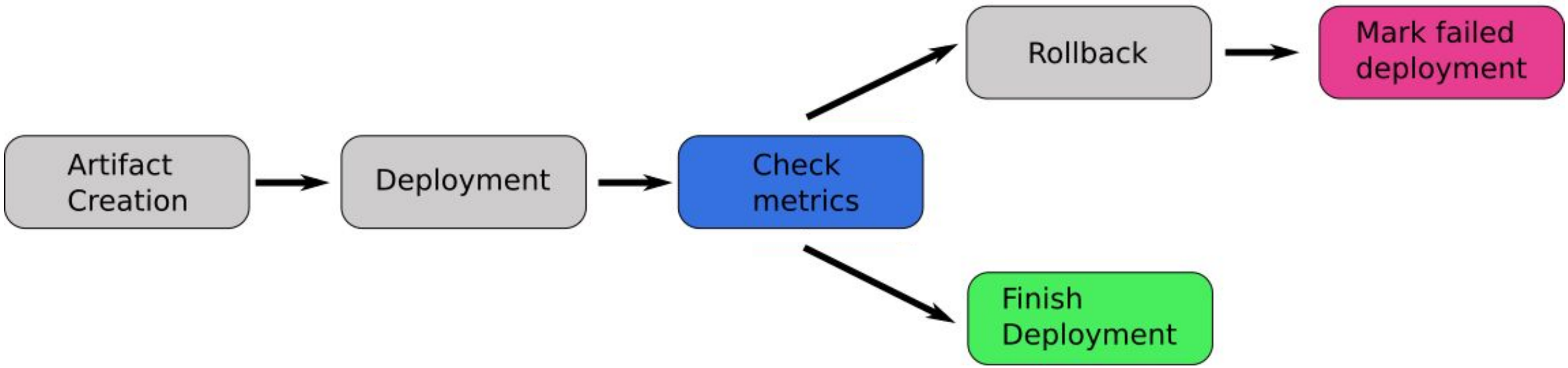
Rollback automatically



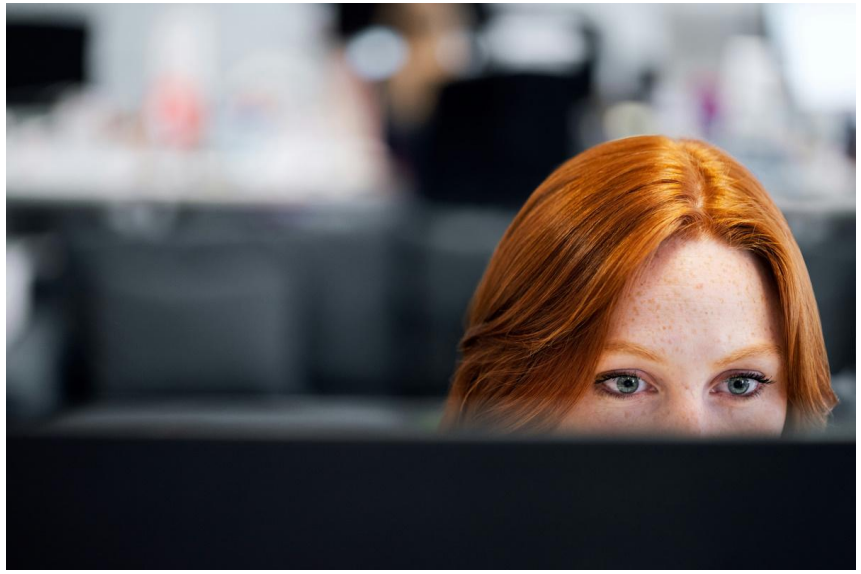
Rollback automatically



Rollback automatically



Deploy on Friday at 5:00



Deployment happens at 5.00 pm on Friday

5:15 the whole team is at the pub

ArgoCon
EUROPE

Kubernetes Autoscaling (HPA)

Kubernetes scaling options

ArgoCon EUROPE

- **Horizontal Pod Autoscaler (HPA)**
- Vertical Pod Autoscaler (VPA)
- Cluster Autoscaler
- Custom Metrics Autoscaler
- Kubernetes Event-driven Autoscaler (KEDA)

HPA manifest

- HPA monitors the metrics (like CPU, memory usage or RPS) of the pods and adjust the number of replicas to match the desired state.
- Effective for variable traffic

Before HPA

After HPA

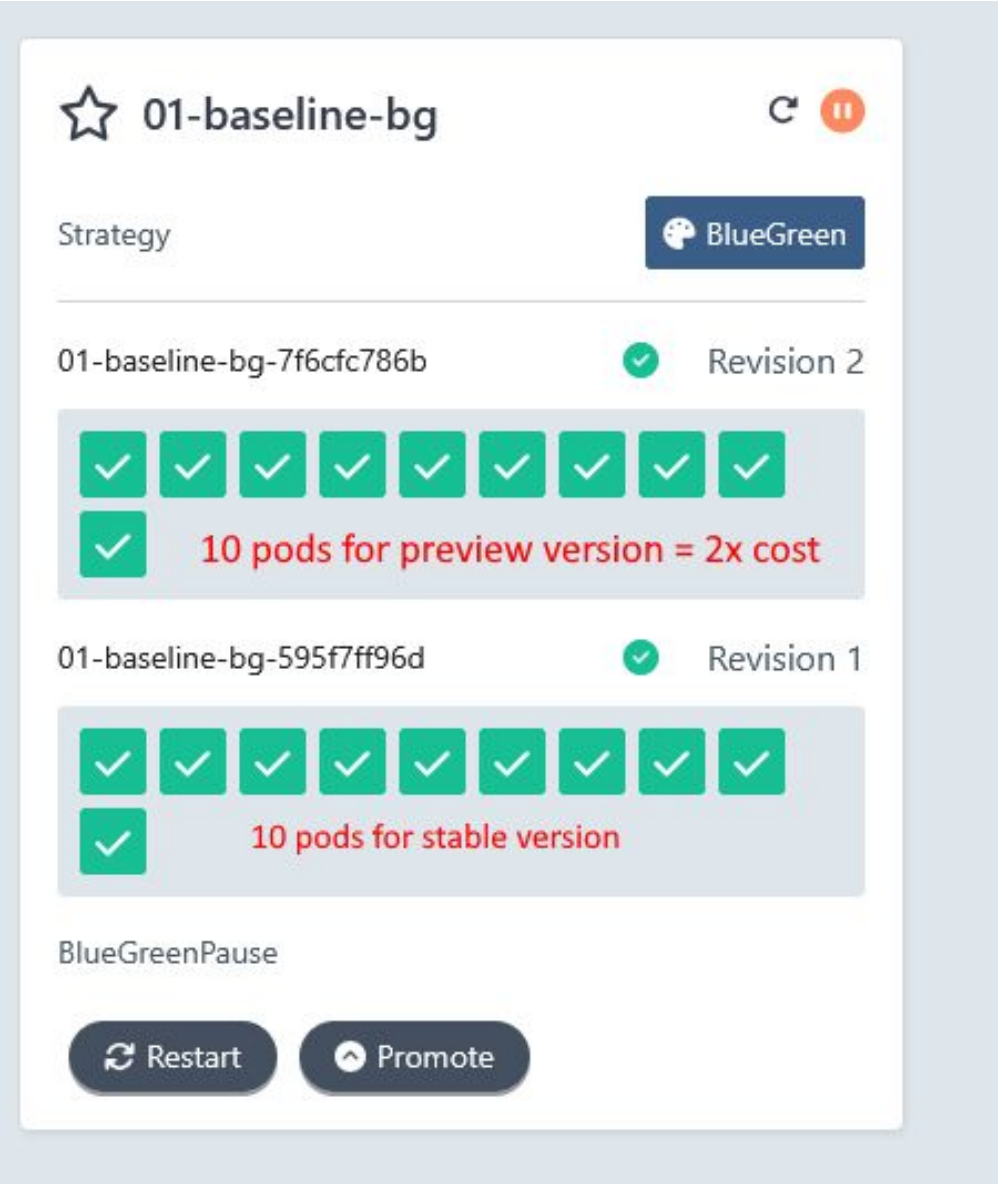
```yaml
apiVersion: autoscaling/v2
kind: HorizontalPodAutoscaler
metadata:
  name: demo-hpa
  labels:
    app: cost-demo
spec:
  minReplicas: 1
  maxReplicas: 10
  metrics:
    - type: Resource
      resource:
        name: memory
        target:
          type: AverageValue
          averageValue: 16Mi
```

ArgoCon
EUROPE

The problem of cost

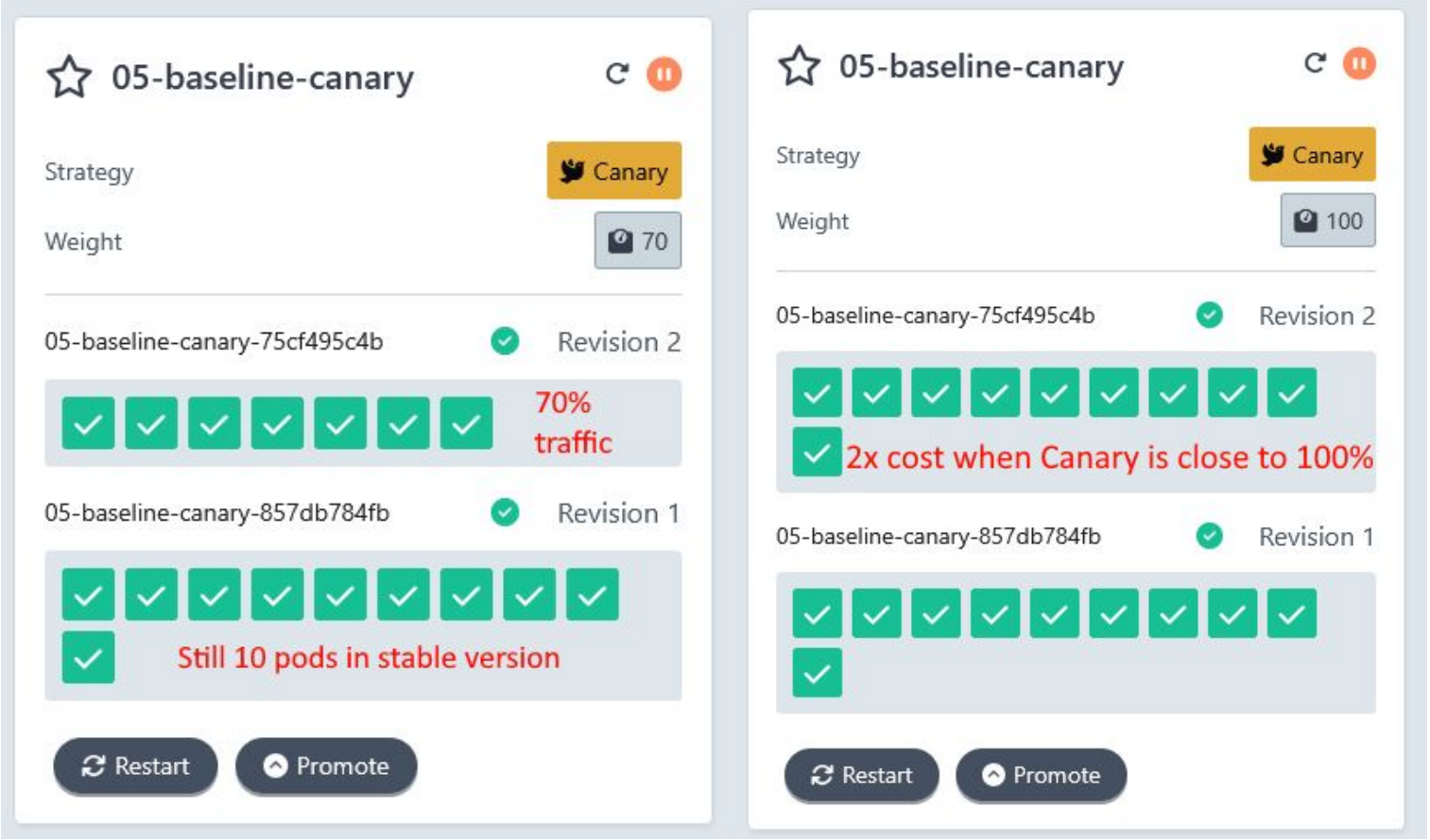
Blue/Green is 2x cost

- Launching a new color doubles the cost
- Preview version needed for smoke tests or other integration tests
- In a big company with many applications, costs quickly add up



Canary can be 2x cost

- Cost is increasing as canary is progressing
- Noticeable impact for a company with many applications



Argo Rollouts in the wild

ArgoCon EUROPE

Understand your use case

Argo Rollouts is perfect for all progressive delivery scenarios as explained in the concepts page.

You should *NOT* use Argo Rollouts for preview/ephemeral environments. For that use case check the Argo CD Pull Request generator.

The recommended way to use Argo Rollouts is for brief deployments that take 15-20 minutes or maximum 1-2 hours. If you want to run new versions for days or weeks before deciding to promote, then Argo Rollouts is probably not the best solution for you.

Keeping parallel releases for long times, complicates the deployment process a lot and opens several questions where different people have different views on how Argo Rollouts should work.

People really do want to tests canaries for days/weeks 😕

https://argo-rollouts.readthedocs.io/en/latest/best-practices/

ArgoCon
EUROPE

Autoscaling with Argo Rollouts

Argo Rollouts + HPA

ArgoCon
EUROPE

1. People want to reduce costs for their applications
2. Complete control over preview/stable pods
3. Current Argo Rollouts HPA documentation is lacking
4. Not really clear what happens when both active

Can I disable HPA while the argo rollout is progressing? #3925

Unanswered **sungho-rpls** asked this question in **Q&A**

sungho-rpls on Oct 30, 2024

HPA scales out replicas in my environment because of high CPU usage while the argo rollout progresses. High CPU usage for startup time is acceptable for me, but too many replicas by HPA cause problems with latency, resource usage, and database connections. Is it possible to disable HPA scaling when the argo rollout is progressing?

↑ 1

Shruthi 12:53 AM

Hi, I am trying to use Argo Rollouts with deployment that has HPA configured.

1. I have updated HPA's scaleTargetRef Kind to Rollout so the rollout is aware of the HPA otherwise rollout doesn't seem to be working with the HPA
2. With this setting the HPA scales both canary and stable pods resulting in doubling the number of pods, in order to avoid that added dynamicStableScale: true setting to rollout.
3. and dynamicStableScale only works when traffic routing is enabled so I created 2 services (1 for canary and 1 for stable) and updated rollout to use traffic routing
4. I set replicas to 0 in deployment and replicas to 1 in rollout and let the HPA scale the replicas based on min replica count

Can I get some help validating these steps?

David Curran Nov 6th, 2023 at 5:51 PM

https://argoproj.github.io/argo-rollouts/migrating/#reference-deployment-from-rollout

If I reference a deployment in my rollout. How does that work with HPA? I assume HPA will scale my deployment back up but not the rollout? I'd end up with double the pods running?

Jacob Hagstedt Dec 19th, 2022 at 2:36 PM

Perhaps one thing that is unclear to me is how Argo Rollouts is handling a HPA during a canary rollout?

Jacob Hagstedt Sep 21st, 2022 at 4:03 PM

`hpa` `canary strategy`

Hey! I've noticed that it is possible to use HPA's with rollouts, but how does it actually work? I have a concern about e.g. having a metric like this

```
metrics:
  - type: Resource
    resource:
      name: cpu
      target:
        type: Utilization
```

which means we want to scale based on CPU usage. But how does that work when using the canary rollout strategy?

Because we will have both 'stable' and 'canary' pods up and running. What I would like to happen is that the HPA only looks at the 'stable' pods for CPU usage. Is that the case?

Jaysheel Shah Mar 11th at 9:51 PM

Hello. I was working on implementing Argo Rollouts for my applications and I noticed an odd behavior with the HPA configuration. If a HPA config is added for an argo rollout, it will only auto-scale the pods in replica sets that are tied to the blue version (current/ latest version) or previously deployed replica sets (old blue versions). The replica set associated with the preview version (green version) will not be auto scaled. Is there any way to configure the same autoscaling behavior with the preview version as well without working with the `previewReplicaCount` each time I want to update the number of replicas? Any guidance/suggestions you may have is appreciated. Thank you.

Lukasz Boldys Apr 21st, 2023 at 4:49 PM

Hi, I tried to google for it but I couldn't find the answer. I have Rollout configured with HPA.

```
apiVersion: autoscaling/v2beta2
kind: HorizontalPodAutoscaler
metadata:
  name: app-name-hpa
spec:
  metrics:
  - type: Resource
    resource:
      name: cpu
      target:
        type: Utilization
        averageUtilization: 50
  scaleTargetRef:
    apiVersion: argoproj.io/v1alpha1
    kind: Rollout
    name: app-name
  minReplicas: 5
  maxReplicas: 5
```

And Rollout using Canary strategy. But for some reason when new Canary pods are started the HPA is not scaling down the Stable RS. So in the middle of deployment I can have 4 Canary pods and 5 Stable pods running. Only after deployment is promoted fully the Stable is being scaled down to zero (or rather, Canary becomes stable and it's 5 at that moment). Does anybody have any idea what I might have messed with the configuration?

Sam J Jan 6th at 10:39 PM

Hey all, would really appreciate some insight here. I'm trying to use the canary strategy with no traffic management. I have the following `strategy`:

```
canary:
  maxSurge: 10
  maxUnavailable: 0
  steps:
    - setWeight: 50
    - pause:
        duration: 5m
    - setWeight: 100
```

...along with an HPA with `minReplicas=10` and `maxReplicas=20`. (Note that I am not using traffic management features!)

crumb 8:09 PM

Hello, I've tried using `dynamicStableScale: true` in a canary and when I aborted the rollout it seems the stable scaled itself up to nearly double of what my HPA maximum was.

I've been having issues with rollouts scaling way up during the setWeight / pause period. Is there anything I can do to mitigate running more pods than necessary?

Several combinations

ArgoCon EUROPE

Example 01 - Base case for Blue/Green

Example 02 - Custom number for pods for preview version

Example 03 - Blue/Green with autoscaling

Example 04 - Blue/Green with autoscaling and custom number of pods

Example 05 - Base case for Canary

Example 06 - Canary without Traffic manager

Example 07 - Canary with dynamic scaling

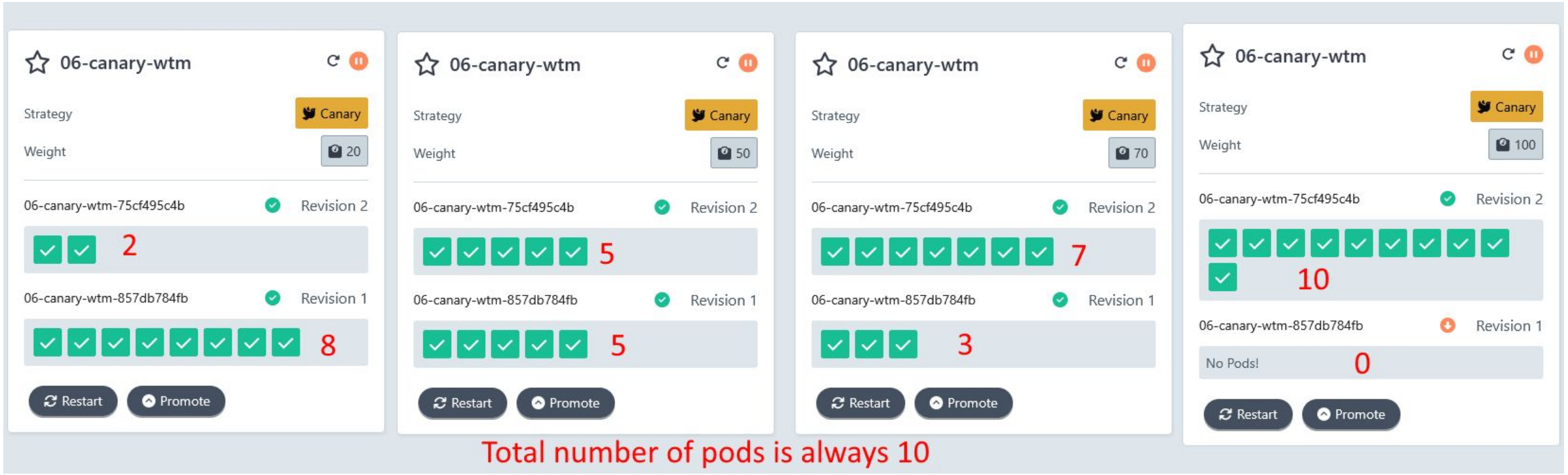
Example 08 - Canary with decoupled traffic split

Example 09 - Canary with autoscaling

Example 10 - Canary with autoscaling and decoupled traffic split

https://github.com/kostis-codefresh/rollouts-autoscaling-example

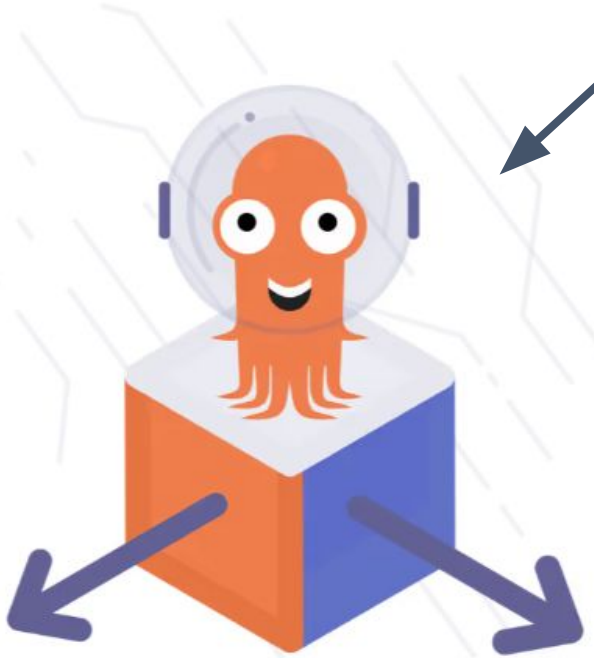
See visually what happens



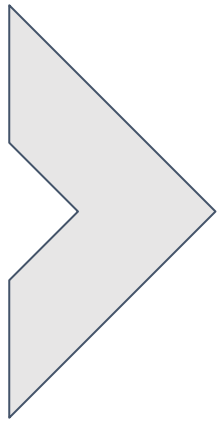
https://github.com/kostis-codefresh/rollouts-autoscaling-example

Override HPA decisions

Winner



Argo Rollouts



Horizontal Autoscaler

ArgoCon
EUROPE

Controlling costs without HPA

Pin down number of pods

ArgoCon EUROPE

Blue/Green

Use **previewReplicaCount** Property

This locks down number of pods for new color to specific number

Canaries

Use **dynamicStableScale** property

OR

Use **setCanaryScale** to specify exact number of canary pods

Specify pods for Green color

ArgoCon EUROPE

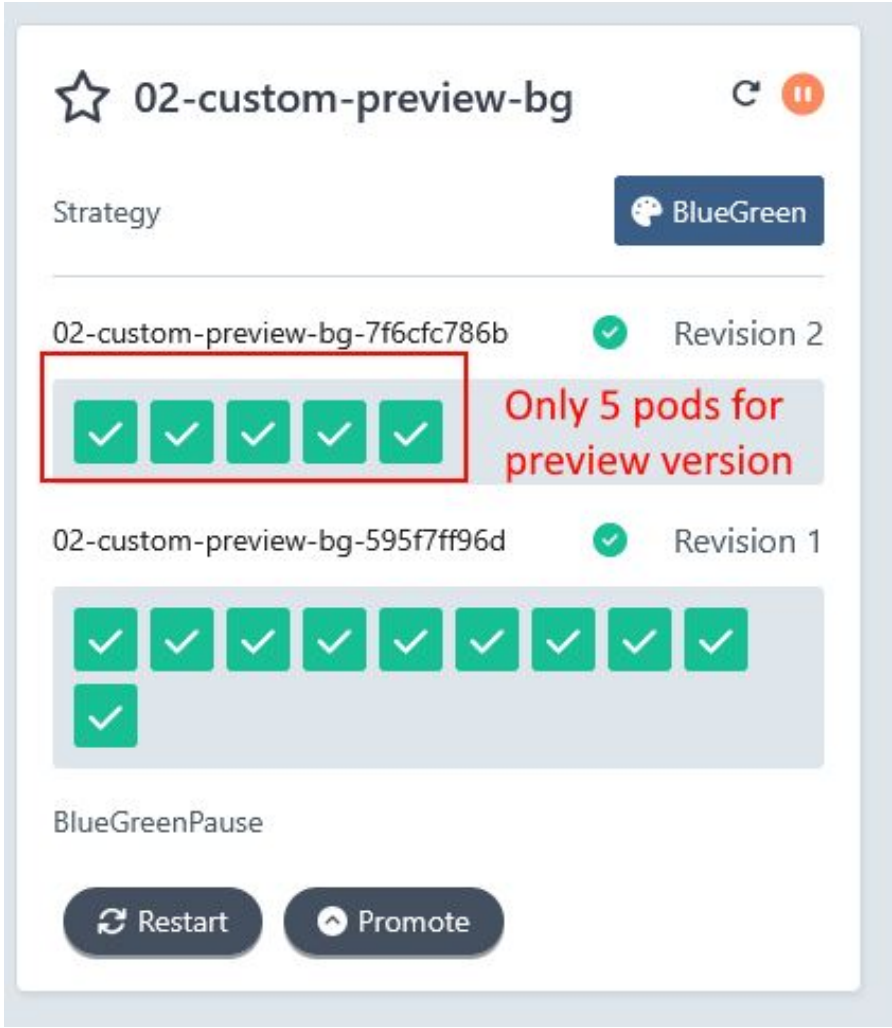
- Cut extra costs by 50% in this example
- Explicitly set the pods of new color to 5

```yaml
apiVersion: argoproj.io/v1alpha1
kind: Rollout
metadata:
  name: 02-custom-preview-bg
spec:
  replicas: 10
  strategy:
    blueGreen:
      previewService: argo-rollouts-preview-service
      activeService: argo-rollouts-stable-service
      previewReplicaCount: 5
      autoPromotionEnabled: false
  revisionHistoryLimit: 2
  selector:
```

https://github.com/kostis-codefresh/rollouts-autoscaling-example/tree/main/02-custom-preview-bg

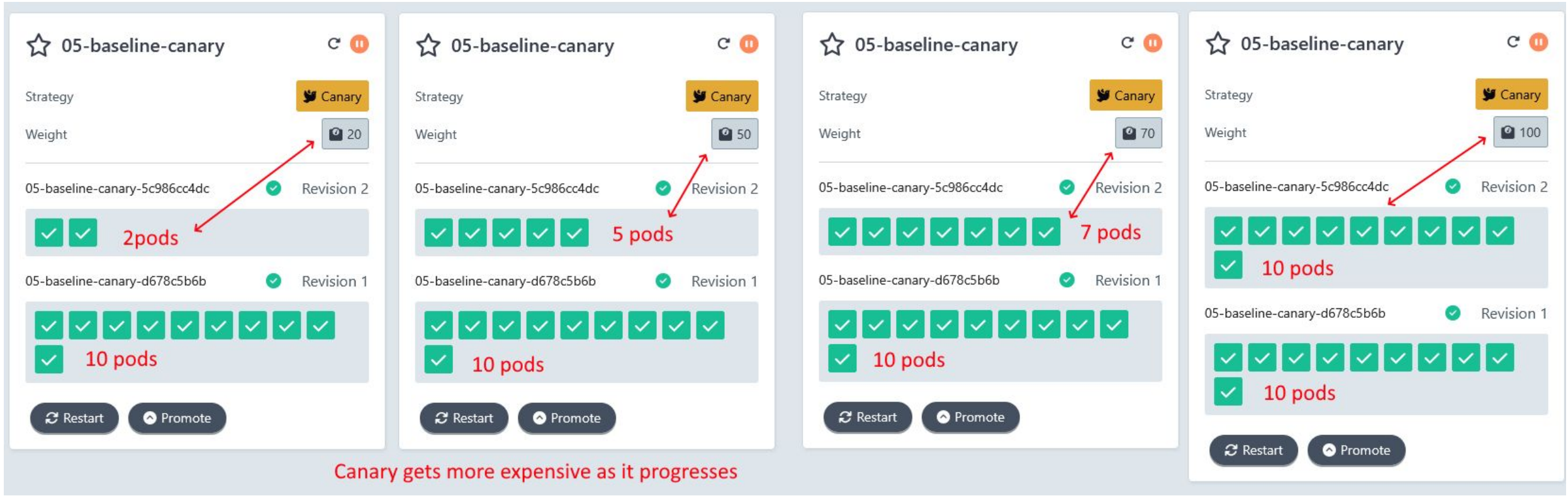
Specify pods for Green color

- Valid only while blue/green is in progress
- Once promoted preview pods will go to 10 (match stable)



https://github.com/kostis-codefresh/rollouts-autoscaling-example/tree/main/02-custom-preview-bg

Default Canary behavior



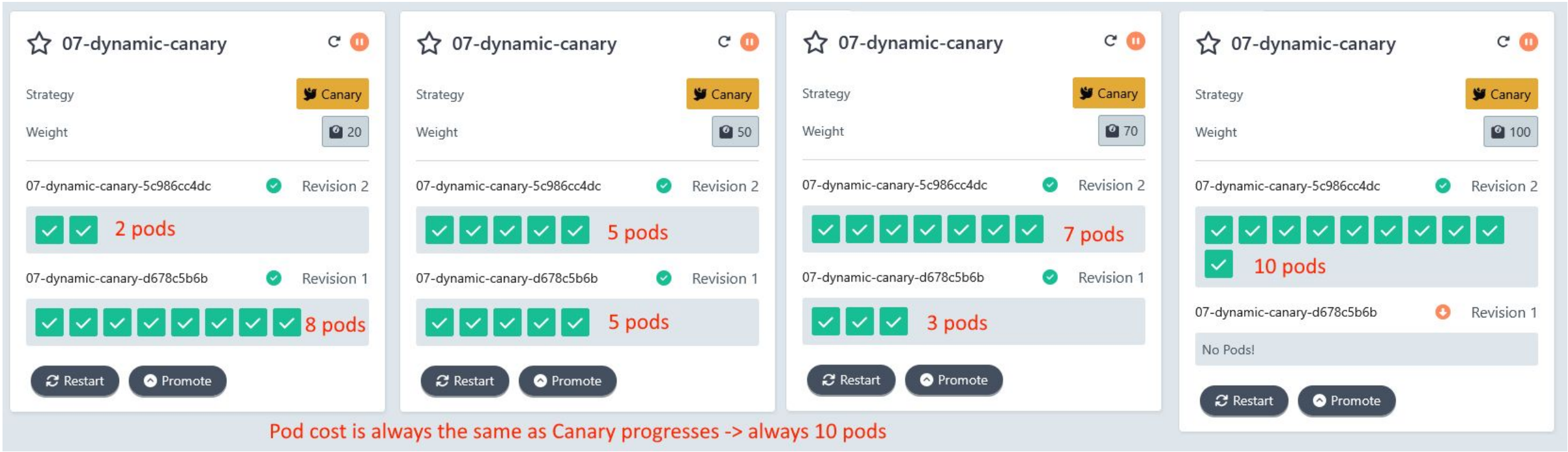
Proportional scaling

- Enabled **dynamicStableScale** property
- Stable pods will decrease as canary pods increase

```yaml
apiVersion: argoproj.io/v1alpha1
kind: Rollout
metadata:
  name: 07-dynamic-canary
spec:
  replicas: 10
  strategy:
    canary:
      canaryService: rollout-canary-preview
      stableService: rollout-canary-stable
      trafficRouting:
        traefik:
          weightedTraefikServiceName: traefik-service
      dynamicStableScale: true
      steps:
        - setWeight: 20
        - pause: {}
        - setWeight: 50
        - pause: {}
        - setWeight: 70
        - pause: {}
        - setWeight: 100
        - pause: {}
  revisionHistoryLimit: 2
```

https://github.com/kostis-codefresh/rollouts-autoscaling-example/tree/main/07-dynamic-canary

Proportional scaling



https://github.com/kostis-codefresh/rollouts-autoscaling-example/tree/main/07-dynamic-canary

ArgoCon
EUROPE

Decouple canary traffic from canary pods

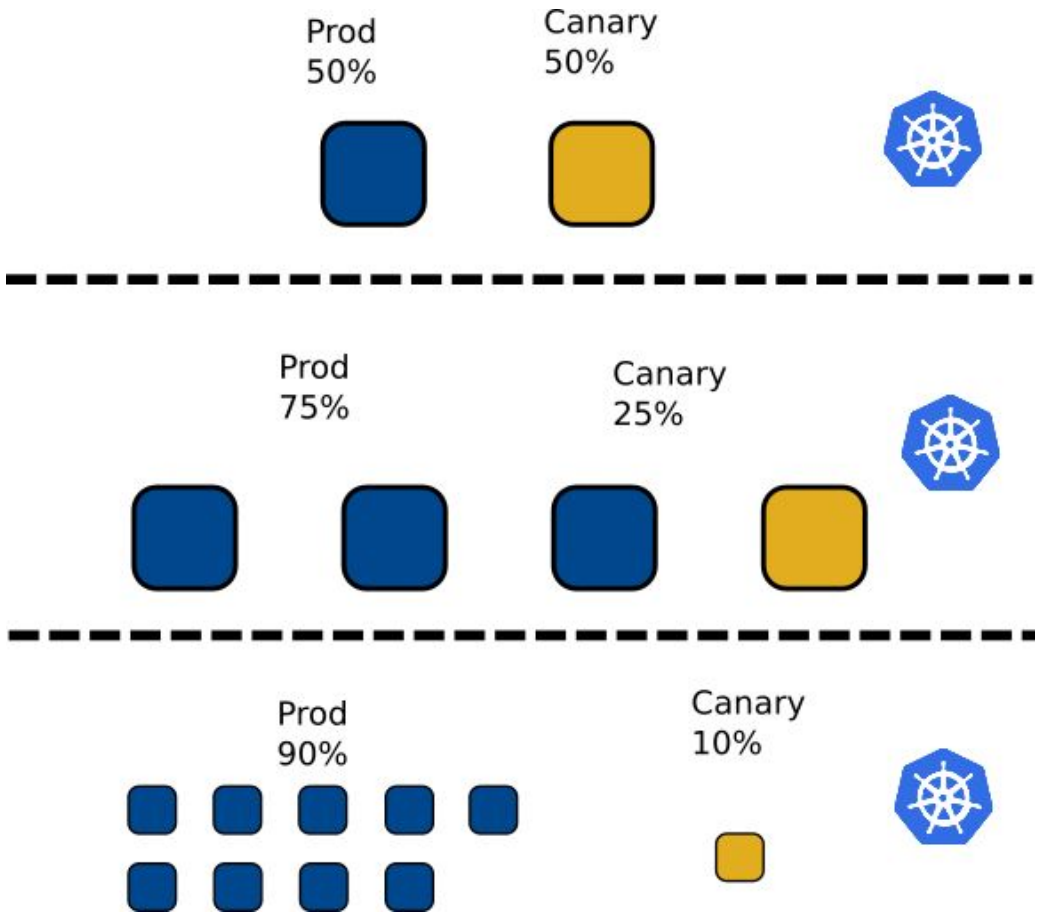
Control canary traffic

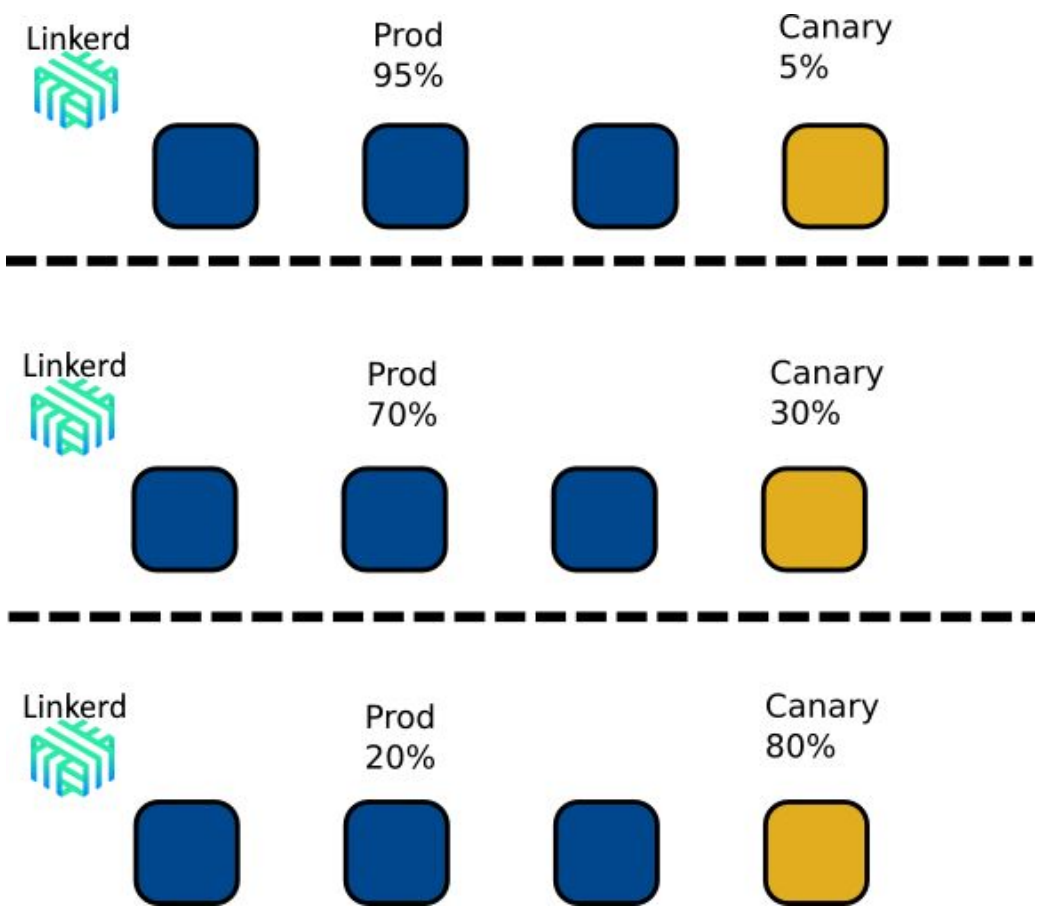
ArgoCon
EUROPE

- Default behavior -> number of canary pods get proportionate traffic
- You **CAN** change this and send **ANY** number of traffic to **ANY** number of canary pods
- Requires a traffic manager

Without/With traffic controller

ArgoCon EUROPE





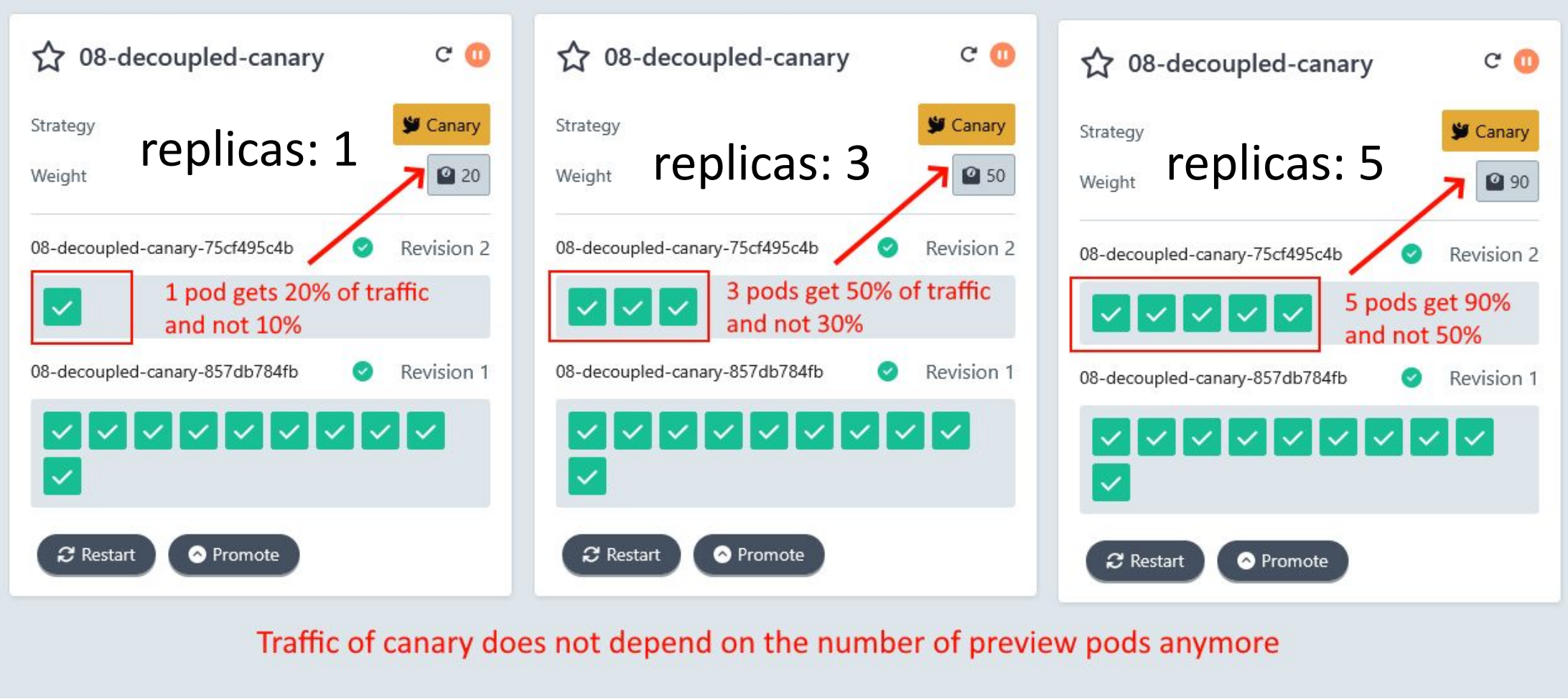
Control canary traffic

- Send 20% of canary traffic to just 1 pod
- Send 50% of canary traffic to 3 pods
- Send 90% of canary traffic to 5 pods
- Send 100% of canary traffic to 8 pods

```yaml
apiVersion: argoproj.io/v1alpha1
kind: Rollout
metadata:
  name: 08-decoupled-canary
spec:
  replicas: 10
  strategy:
    canary:
      canaryService: rollout-canary-preview
      stableService: rollout-canary-stable
      trafficRouting:
        traefik:
          weightedTraefikServiceName: traefik-service
      steps:
        - setWeight: 20
        - setCanaryScale:
            replicas: 1
        - pause: {}
        - setWeight: 50
        - setCanaryScale:
            replicas: 3
        - pause: {}
        - setWeight: 90
        - setCanaryScale:
            replicas: 5
        - pause: {}
        - setWeight: 100
        - setCanaryScale:
            replicas: 8
        - pause: {}
  revisionHistoryLimit: 2
```

https://github.com/kostis-codefresh/rollouts-autoscaling-example/tree/main/08-decoupled-canary

Control canary traffic



https://github.com/kostis-codefresh/rollouts-autoscaling-example/tree/main/08-decoupled-canary

ArgoCon
EUROPE

Employing HPA + Argo Rollouts

HPA and Argo Rollouts

Rollout CRD

```
subresources:
 status: {}
 scale:
   labelSelectorPath: .status.selector
   specReplicasPath: .spec.replicas
   statusReplicasPath: .status.replicas
```

Horizontal Autoscaler

HPA (memory based)

- Watch memory usage
- Scale 1-10 pods
- Point to a Rollout CRD

```yaml
apiVersion: autoscaling/v2
kind: HorizontalPodAutoscaler
metadata:
  name: demo-hpa
  labels:
    app: cost-demo
spec:
  maxReplicas: 10
  metrics:
    - type: Resource
      resource:
        name: memory
        target:
          type: AverageValue
          averageValue: 16Mi
  minReplicas: 1
  scaleTargetRef:
    apiVersion: argoproj.io/v1alpha1
    kind: Rollout
    name: 03-hpa-bg
```

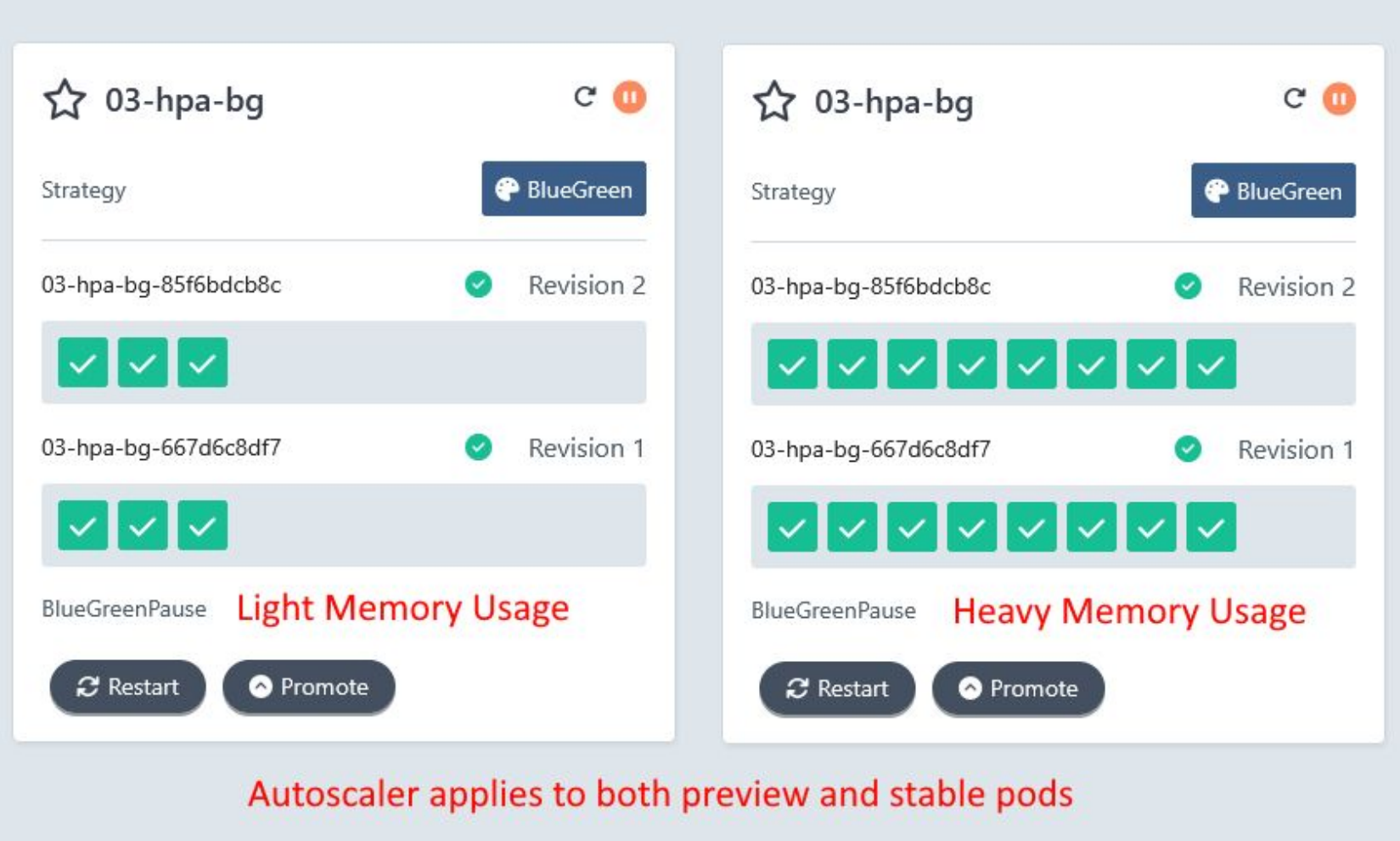
Example app

- Each HTTP call reserves 1MB memory
- /clear releases all reserved memory
- Available as container image

```go
func handler(w http.ResponseWriter, r *http.Request) {
    // Seed the random number generator with the current time
    myRandom := rand.New(rand.NewSource(int64(new(maphash.Hash).Sum64())))

    // Create a buffer of 1MB
    buffer := make([]byte, 1024*1024)

    // Fill the buffer with completely random data
    _, err := myRandom.Read(buffer)
    if err != nil {
        http.Error(w, "Failed to generate random data", http.StatusInternalServerError)
        return
    }

    // Run the SHA1 algorithm on the buffer
    hash := sha1.New()
    hash.Write(buffer)
    sha1Result := hex.EncodeToString(hash.Sum(nil))
```

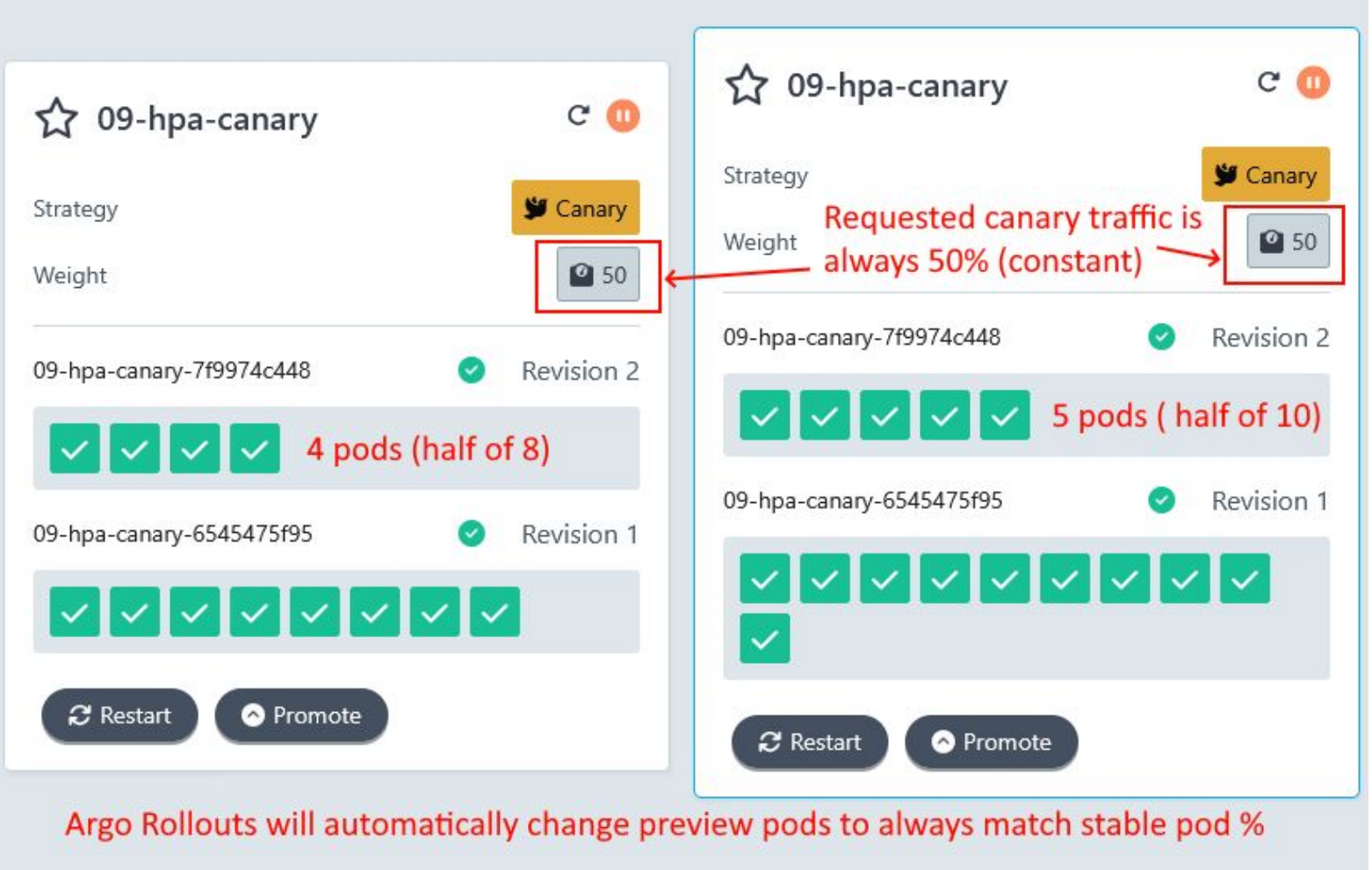
https://github.com/kostis-codefresh/rollouts-autoscaling-example/tree/main/source-code

Blue/Green + HPA



https://github.com/kostis-codefresh/rollouts-autoscaling-example/tree/main/03-hpa-bg

Canary + HPA



https://github.com/kostis-codefresh/rollouts-autoscaling-example/tree/main/09-hpa-canary

HPA versus Argo Rollouts

ArgoCon
EUROPE

Question:

- What happens if we use **previewReplicaCount** and HPA?
- What happens if we use **setCanaryScale** and HPA?

HPA versus Argo Rollouts

ArgoCon
EUROPE

Question:

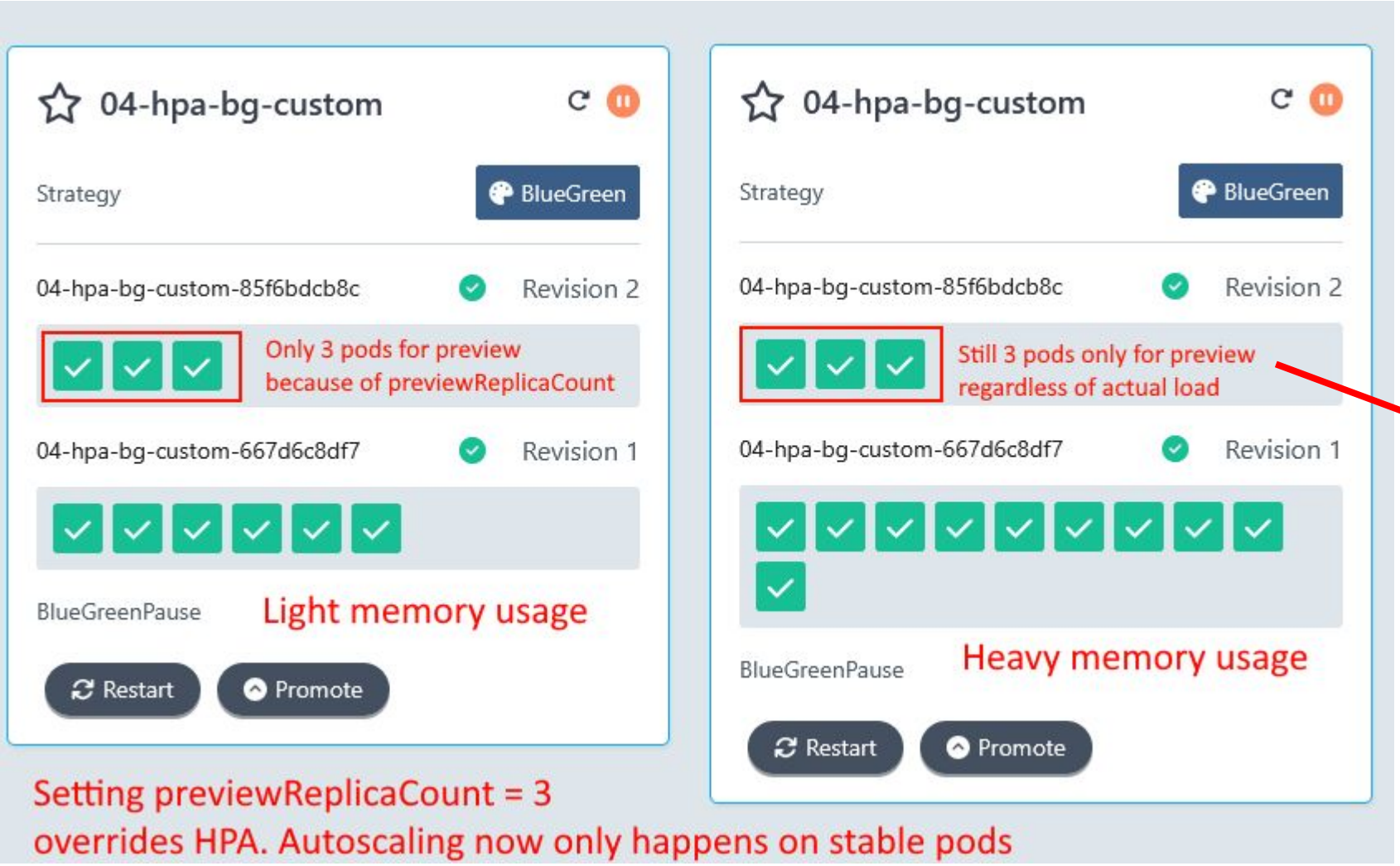
- What happens if we use **previewReplicaCount** and HPA?
- What happens if we use **setCanaryScale** and HPA?

Answer:

Argo Rollouts **overrides** HPA

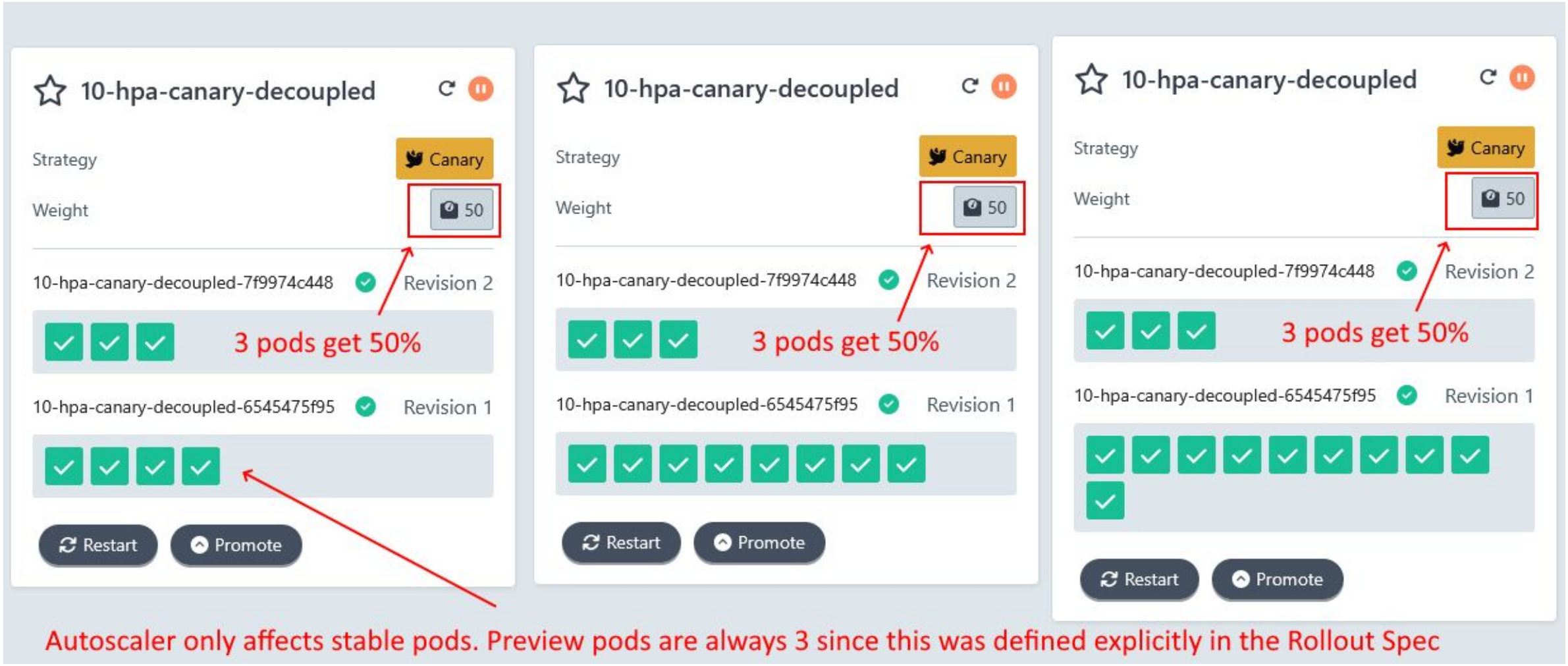
Blue/Green + HPA + Count

ArgoCon EUROPE




```yaml
apiVersion: argoproj.io/v1alpha1
kind: Rollout
metadata:
  name: 04-hpa-bg-custom
spec:
  strategy:
    blueGreen:
      previewService: argo-rollouts-preview-service
      activeService: argo-rollouts-stable-service
      previewReplicaCount: 3
      autoPromotionEnabled: false
  revisionHistoryLimit: 2
  selector:
```

Canary + HPA + custom traffic



ArgoCon
EUROPE

Conclusion

Key points

ArgoCon
EUROPE

- There are several properties you can fine tune
- Evaluate cost vs flexibility vs ease of rollback
- Using HPA is not required for simple scenarios
- When HPA and Argo Rollouts disagree,

  Argo Rollouts wins and it will override HPA

Combinations

ArgoCon EUROPE

Example	Strategy	Options	Stable pods	Preview pods
01	Blue/Green	Default options	Rollouts Spec	Equal to stable pods
02	Blue/Green	Pin preview count	Rollouts Spec	Explicit number
03	Blue/Green	HPA	Controlled by HPA	Controlled by HPA
04	Blue/Green	Pin + HPA	Controlled by HPA	Explicit number
05	Canary	Default options	Rollouts Spec (Constant)	Rollouts Spec
06	Canary	Without Traffic manager	Inverse of Preview pods	Rollouts Spec
07	Canary	Dynamic scaling	Inverse of Preview pods	Rollouts Spec
08	Canary	Decoupled traffic	Rollouts Spec	Explicit number
09	Canary	HPA	Controlled by HPA	Controlled by HPA and current canary weight
10	Canary	HPA + Decoupled	Rollouts Spec	Explicit number

https://github.com/kostis-codefresh/rollouts-autoscaling-example

ArgoCon EUROPE

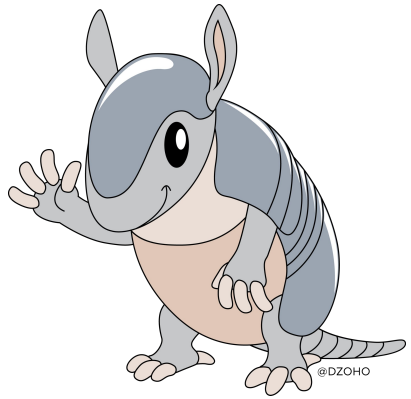
TAG Contributor Strategy

DEAF & HARD OF HEARING WORKING GROUP

Wondering what it is like to be **deaf in tech**?

Want to know what our community can do to **improve accessibility**?

Come chat with us!



ArgoCon EUROPE

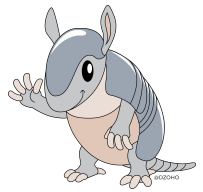
NEW

TAG Contributor Strategy
DEAF & HARD OF HEARING
WORKING GROUP

Become a Mentor for underrepresented groups in Open Source!

Passionate about fostering an inclusive open source community?

Sign Up today!



THE LINUX FOUNDATION

ArgoCon EUROPE

TAG Contributor Strategy
DEAF & HARD OF HEARING WORKING GROUP

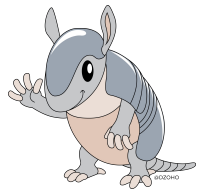
Sign Language Crash Course

Did you know? There are over 300 sign languages worldwide!

Join our course to learn a few signs in ASL (BSL available on demand).

Thu, 17:00 ~ 18:00 I DEI Community Hub

#deaf-and-hard-of-hearing (CNCF Slack)

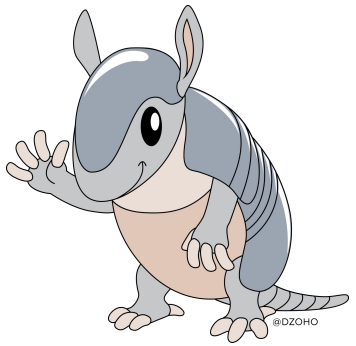


Thank you

ArgoCon EUROPE

Give us feedback! ->

- https://argoproj.github.io/rollouts/
- https://argo-rollouts.readthedocs.io/en/stable/features/hpa-support/
- https://github.com/kostis-codefresh/rollouts-autoscaling-example
- https://contribute.cncf.io/about/deaf-and-hard-of-hearing/



ArgoCon
EUROPE

Argo Rollouts + HPA

Gubska Anastasiia *(BT Group)*
Kapelonis Kostis *(Octopus Deploy)*