POSTMATES

Oct, 2019, v1.2

Невыносимая легкость масштабирования в Kubernetes

Ivan Glushkov
Staff Software Engineer

Content

- Autoscaling Quick Demos
- HPA configuration
- HPA Problems

- Configurable HPA solution
- Watermark Podautoscaler solution
- Vanilla Kubernetes Solution
- Tips and Tricks

http://bit.ly/devoops2019-chpa

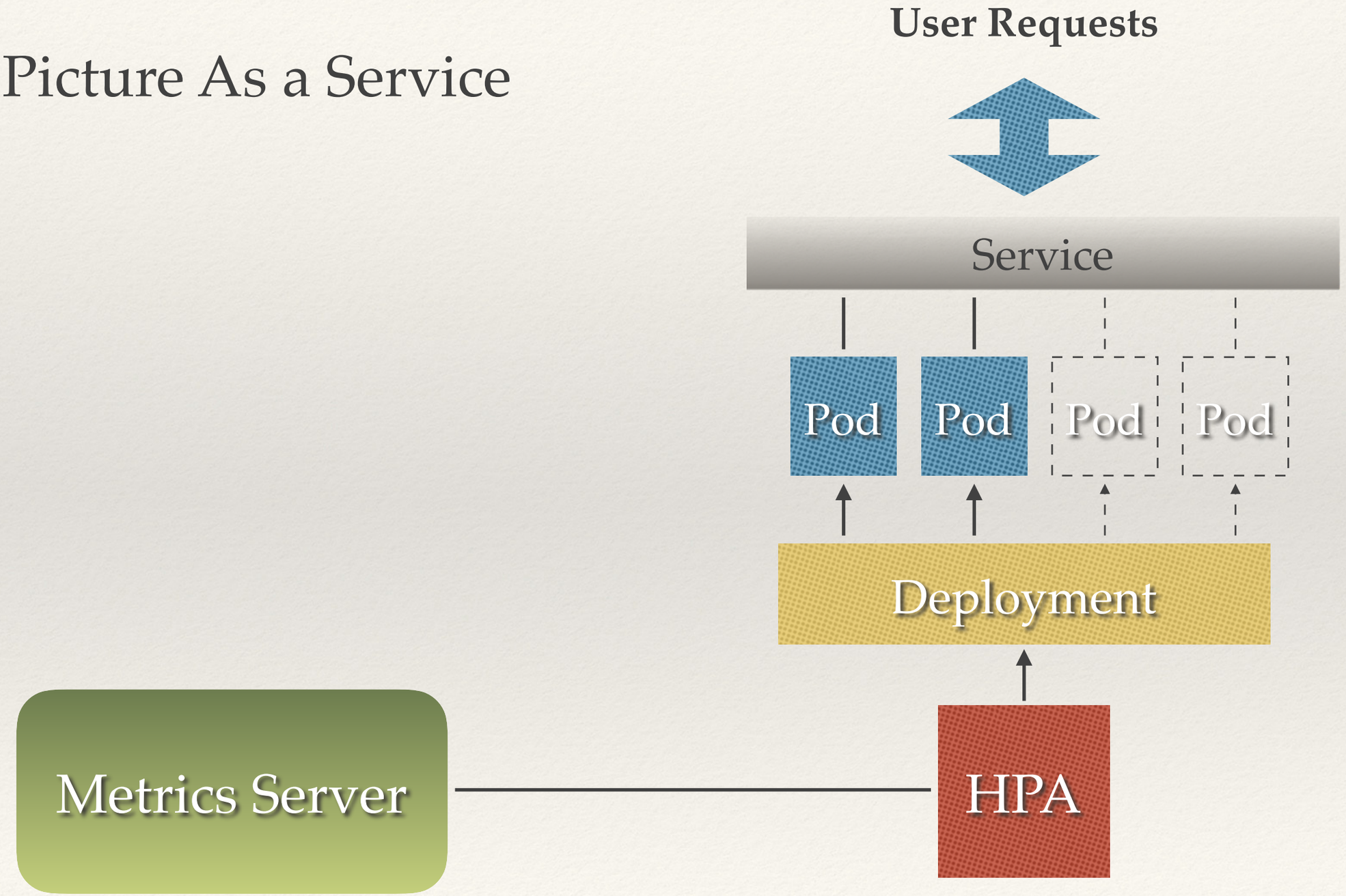
Kubernetes

- Large Community:
  - A lot of features => reduced costs
  - Good documentation => reduced costs
  - No need to teach newcomers => reduced costs
- Autoscaler => reduced costs
- Used by large players => quality (?)

Cat Picture as a Service

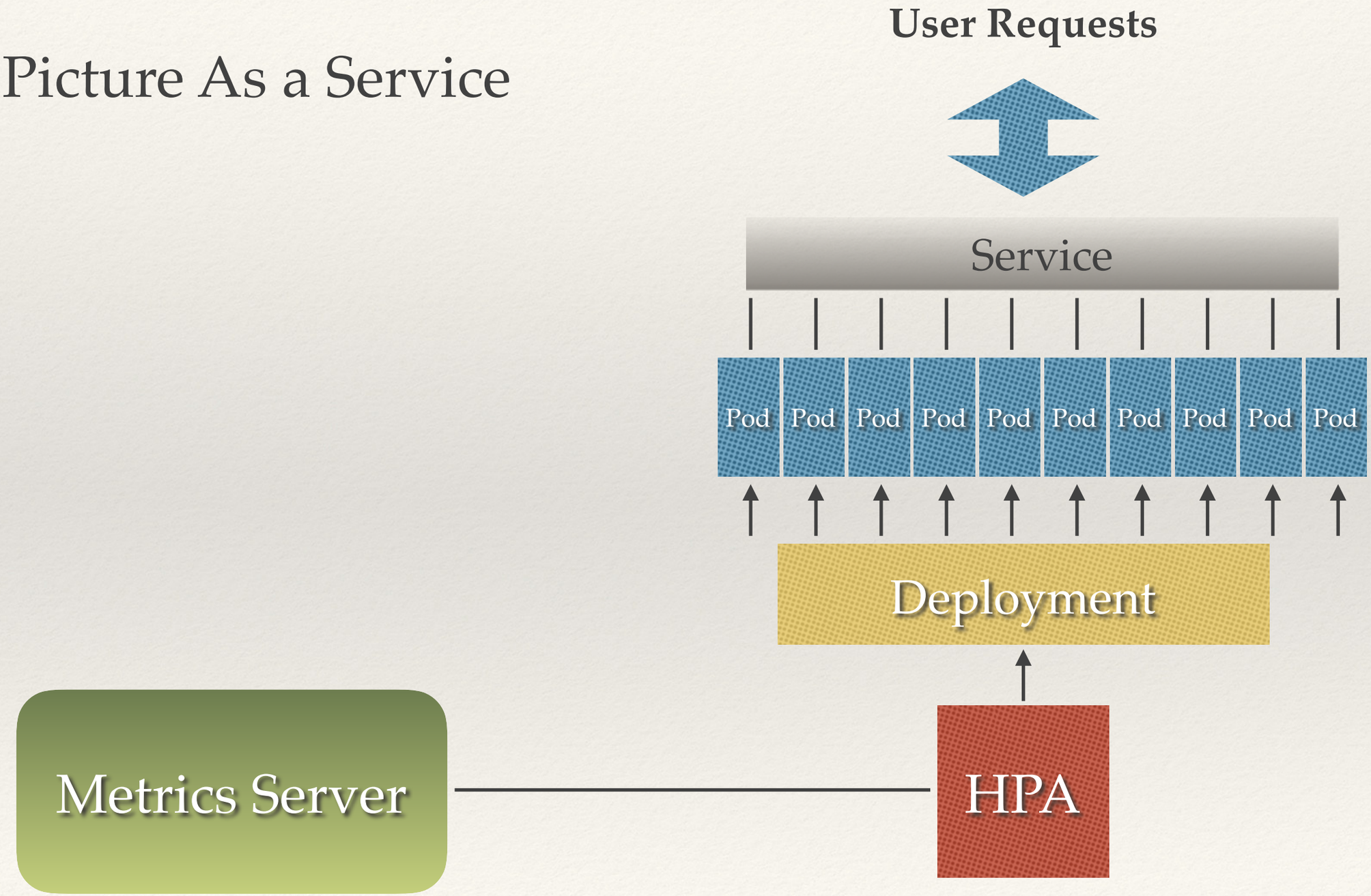
Autoscaler quick demo

Cat Picture As a Service

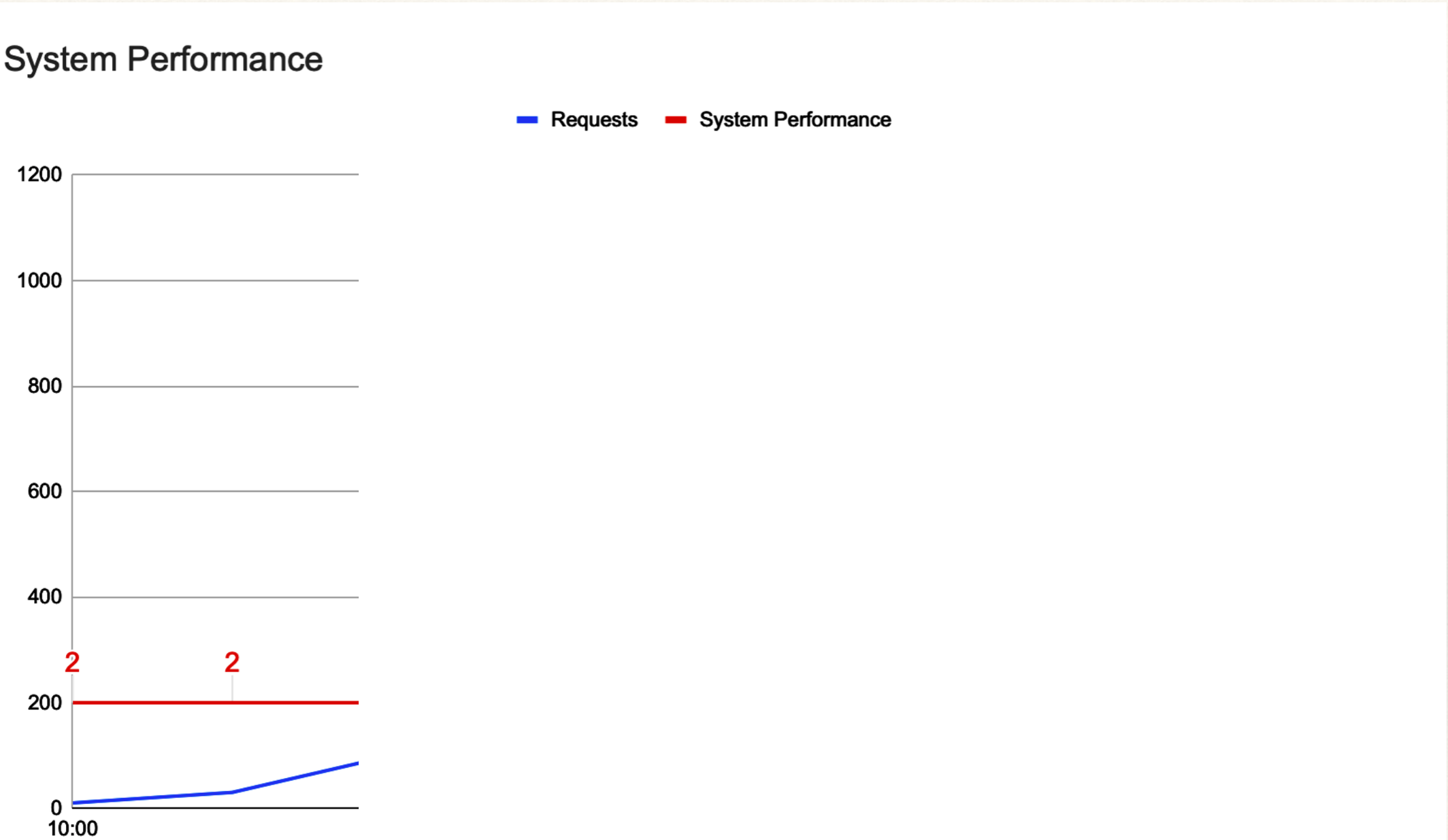


Autoscaler quick demo

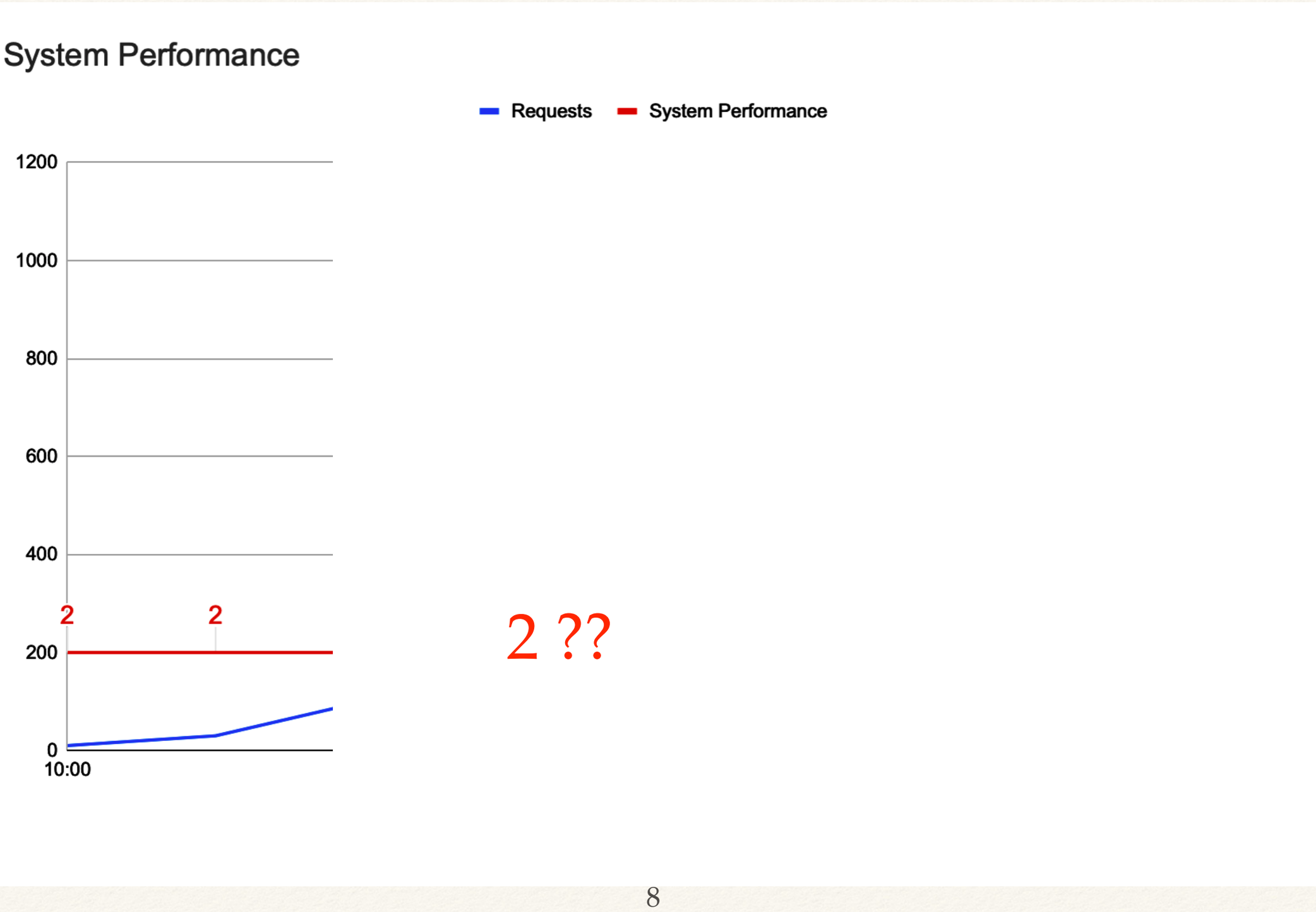
Cat Picture As a Service



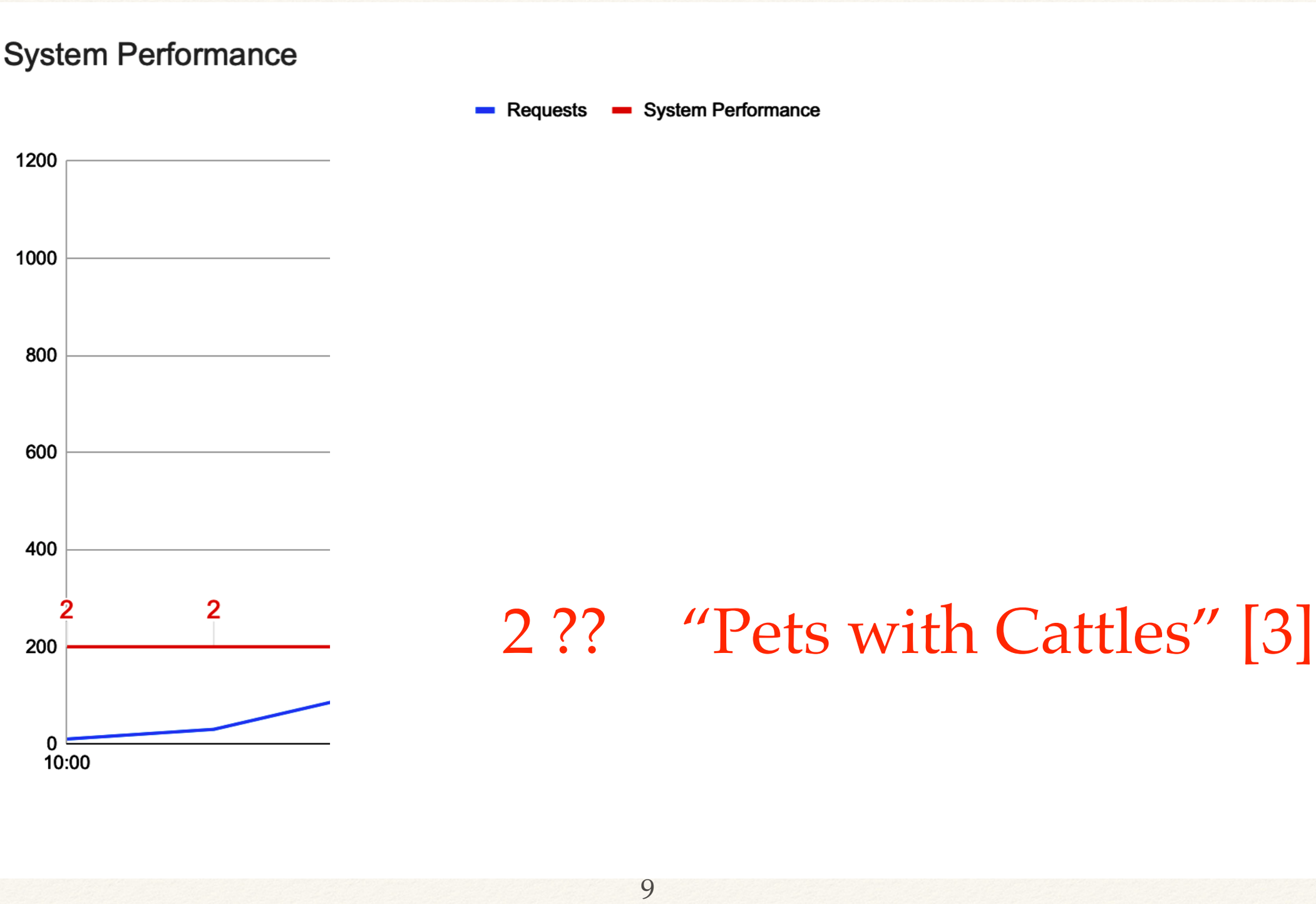
Autoscaler quick demo 1



Autoscaler quick demo 1

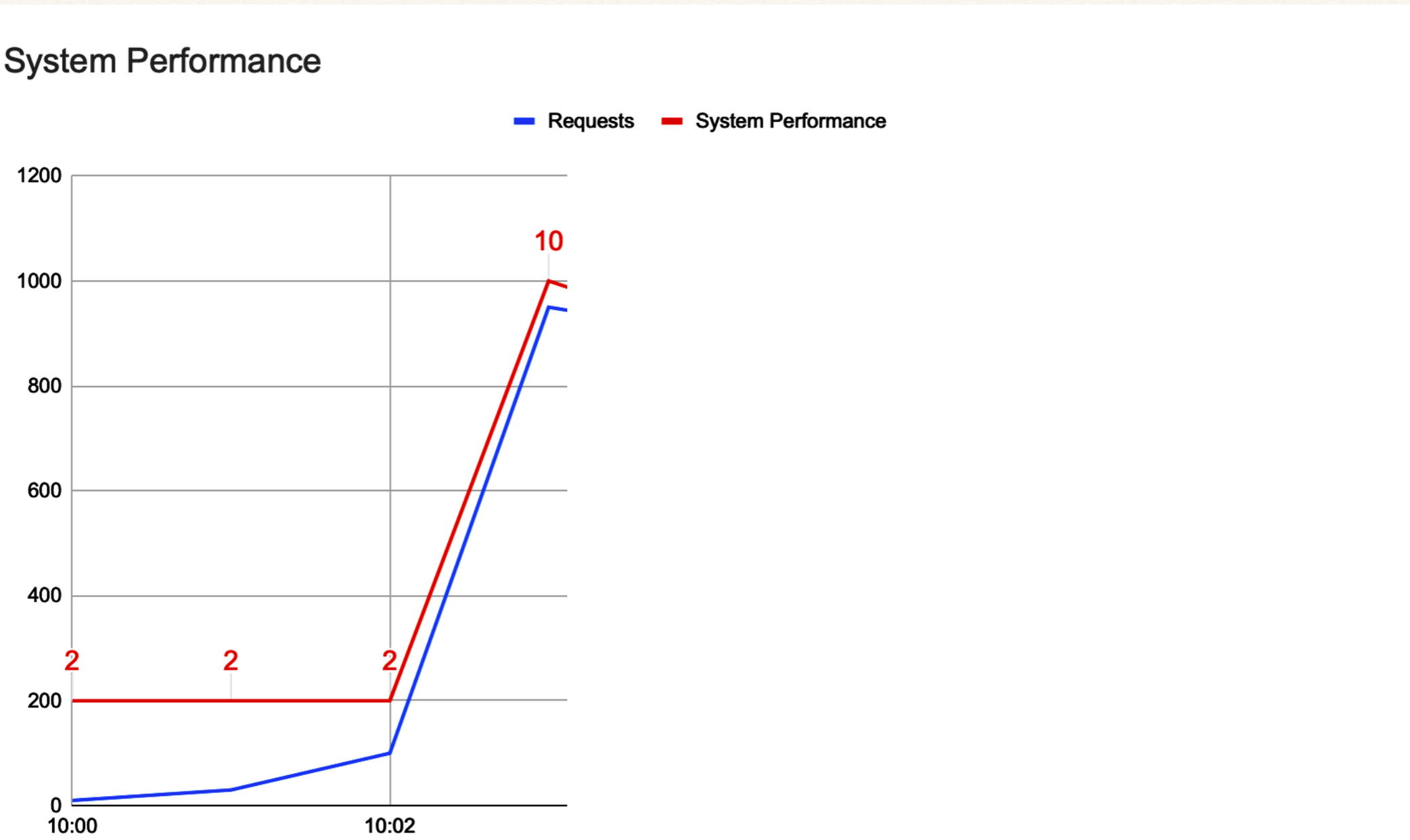


Autoscaler quick demo 1

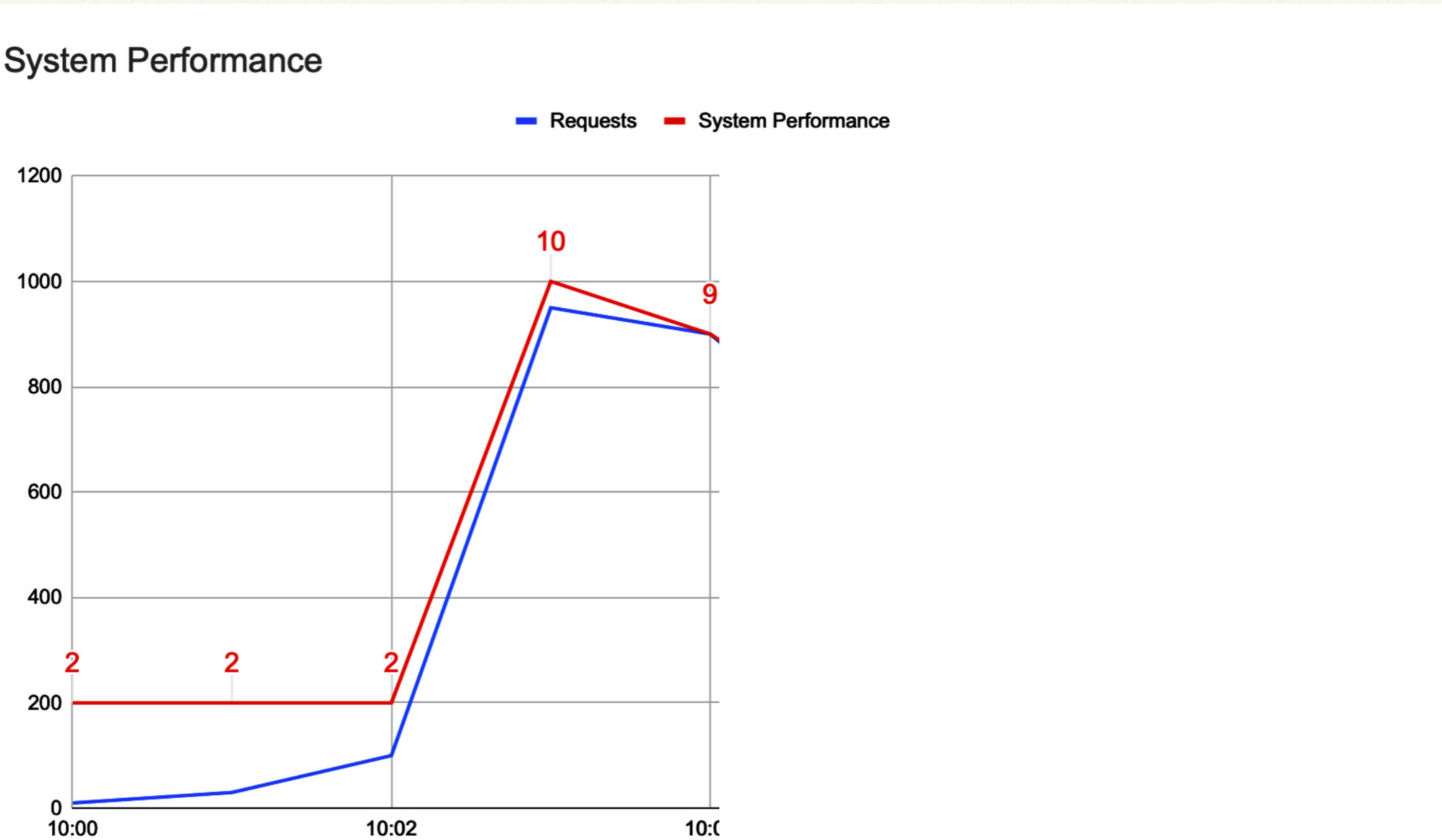


2 ?? “Pets with Cattles” [3]

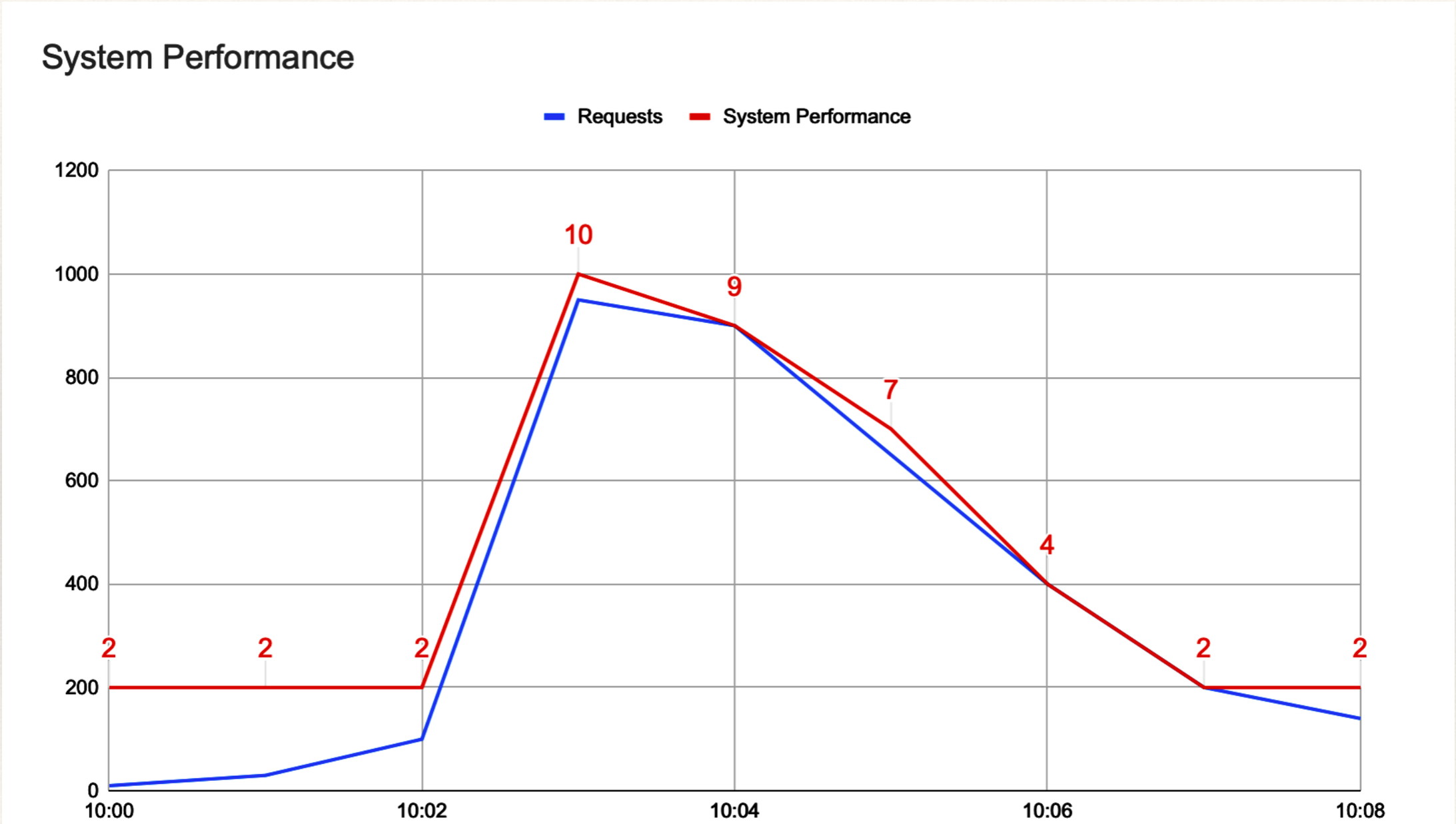
Autoscaler quick demo 1



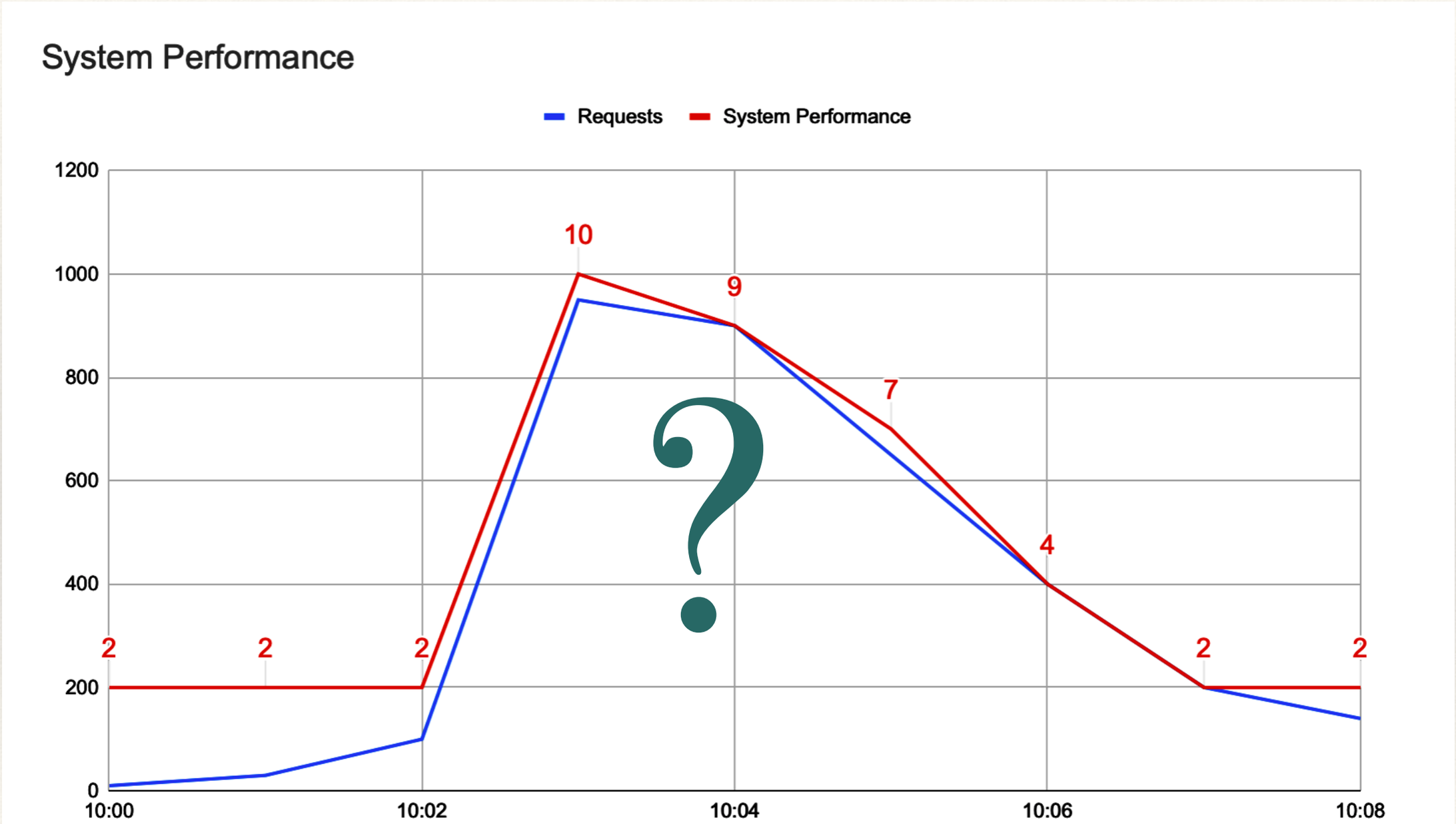
Autoscaler quick demo 1



Autoscaler quick demo 1

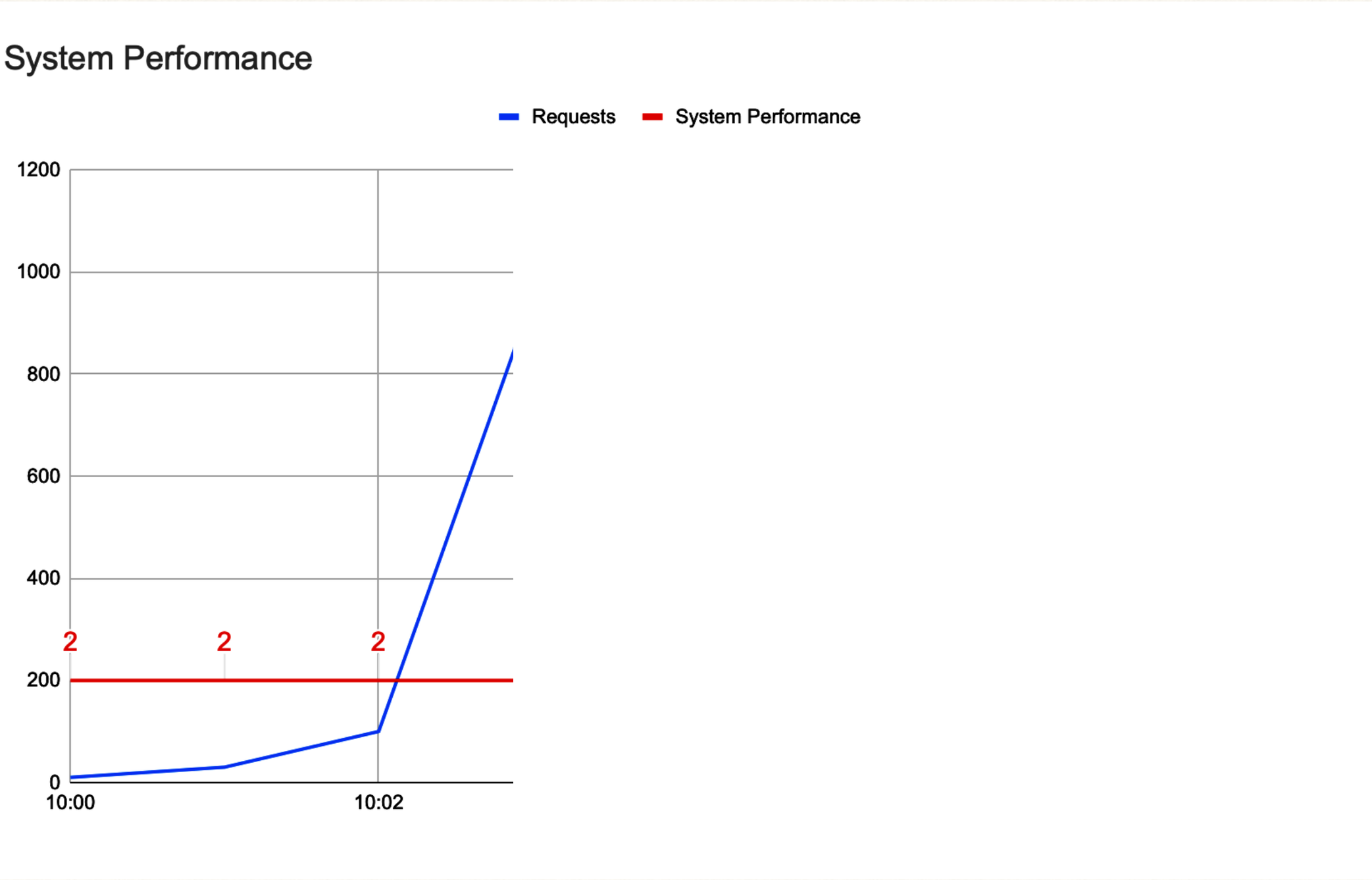


Autoscaler quick demo 1

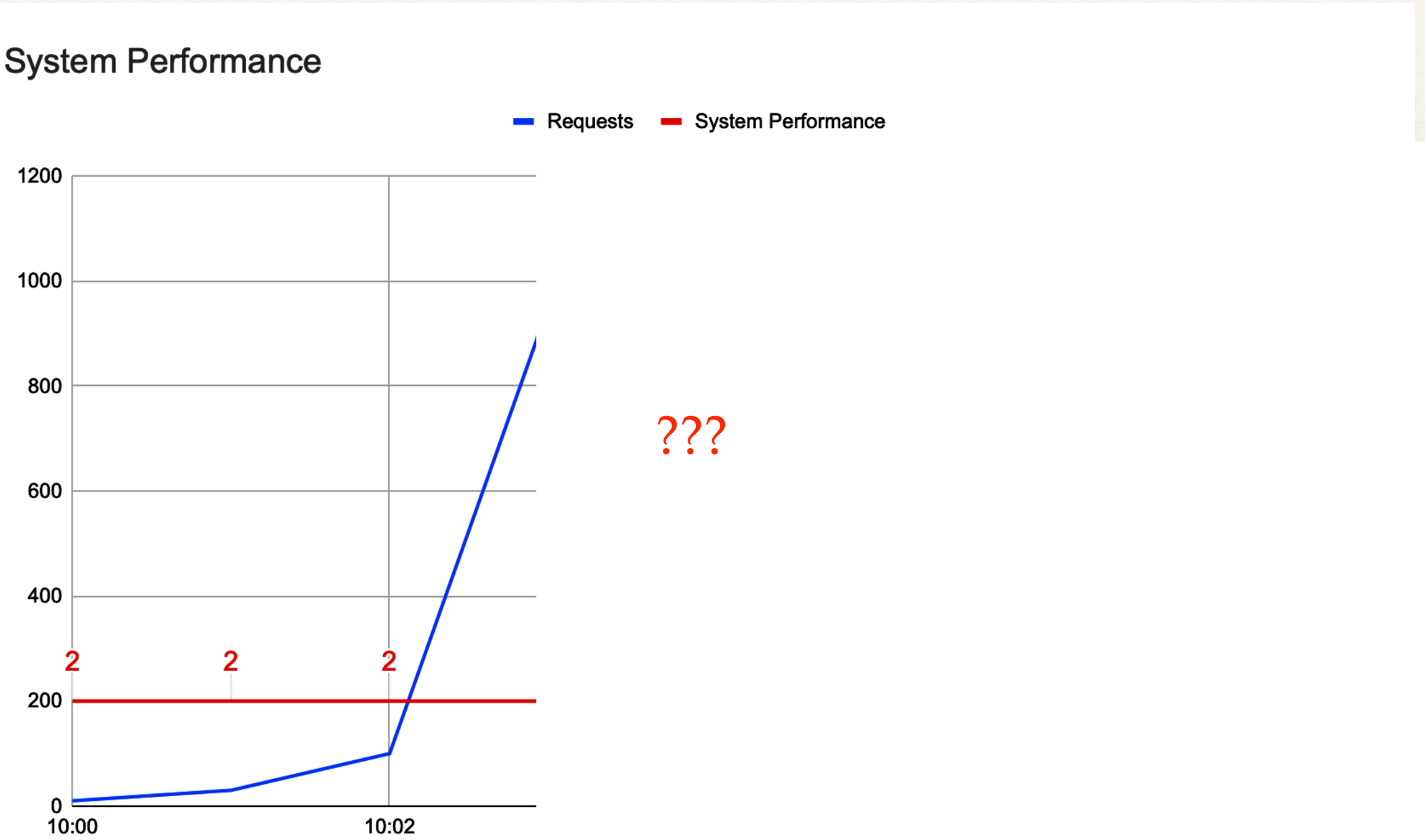


Nope

Autoscaler quick demo 2

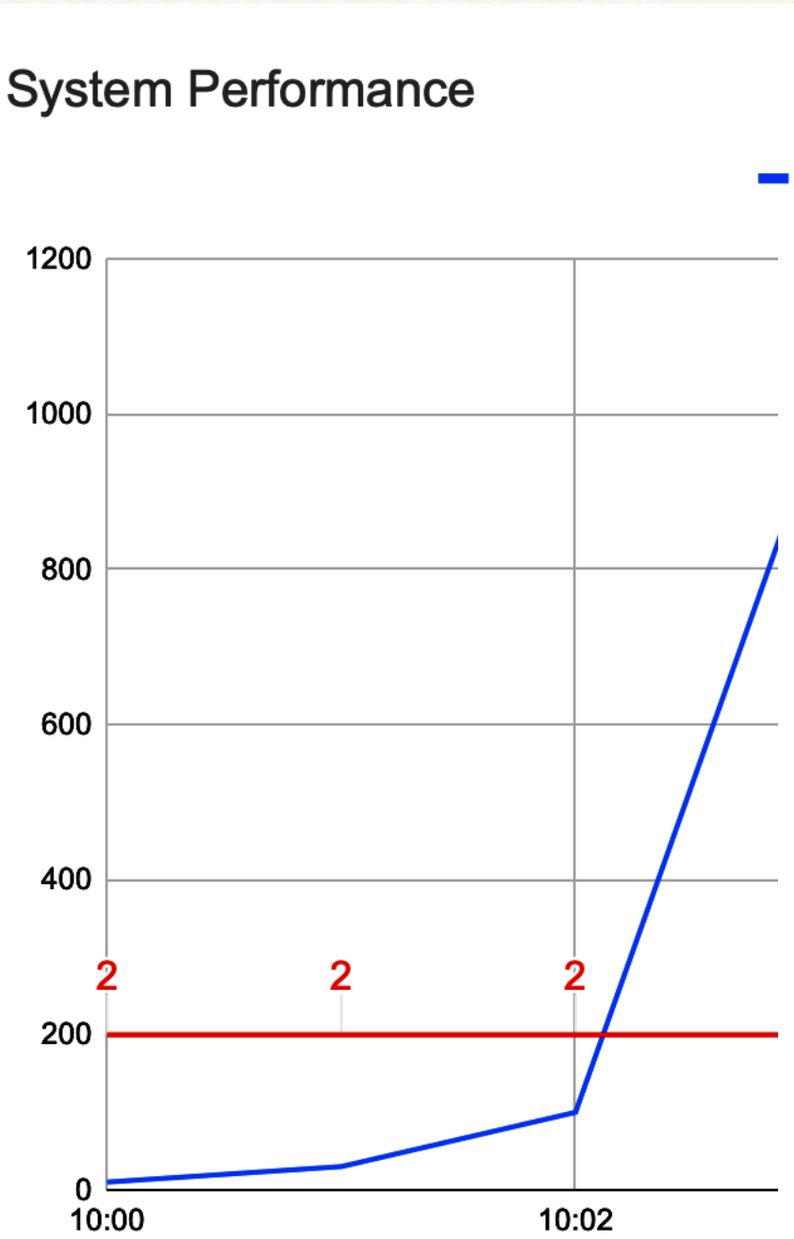


Autoscaler quick demo 2



???

Autoscaler quick demo 2

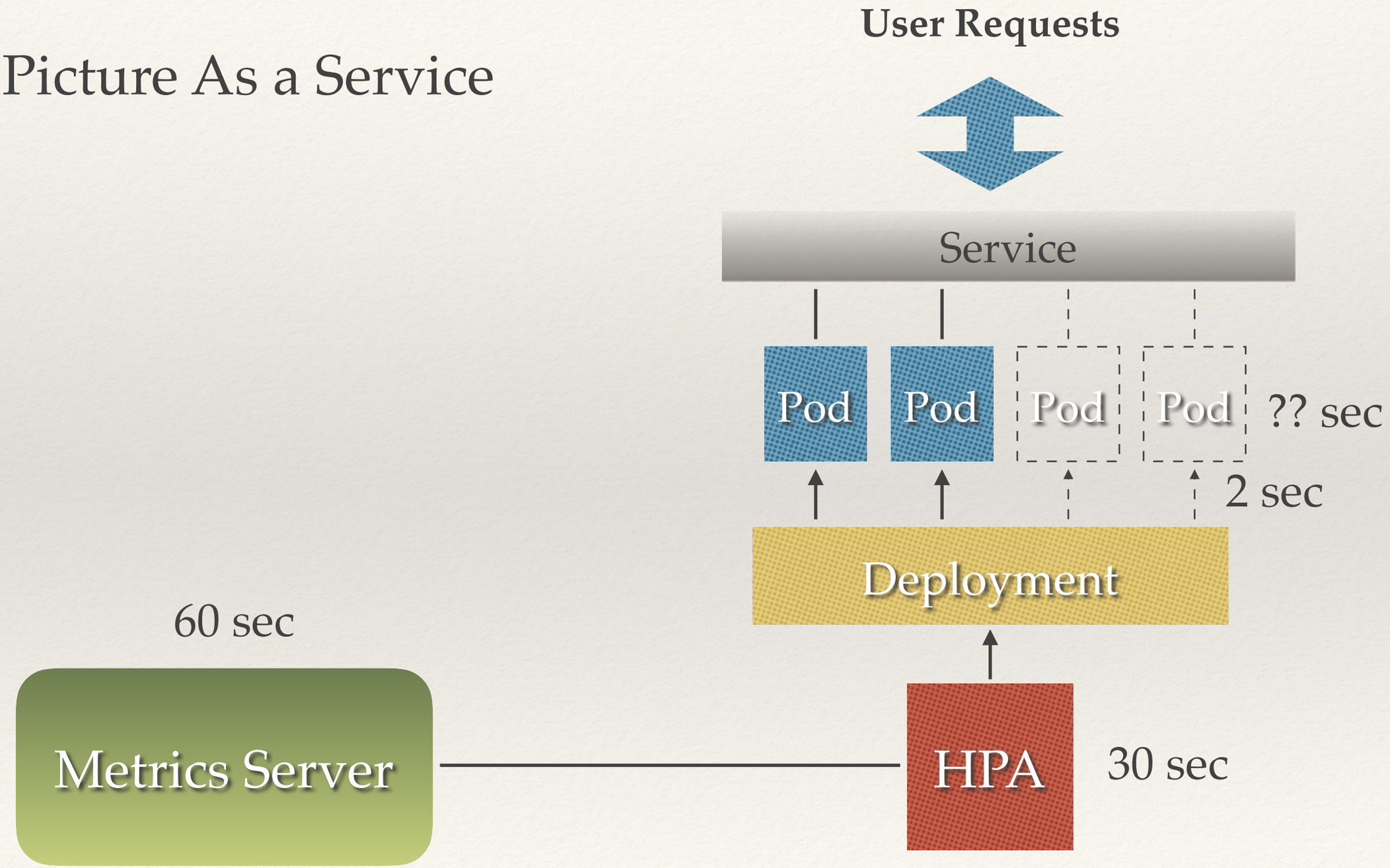


???

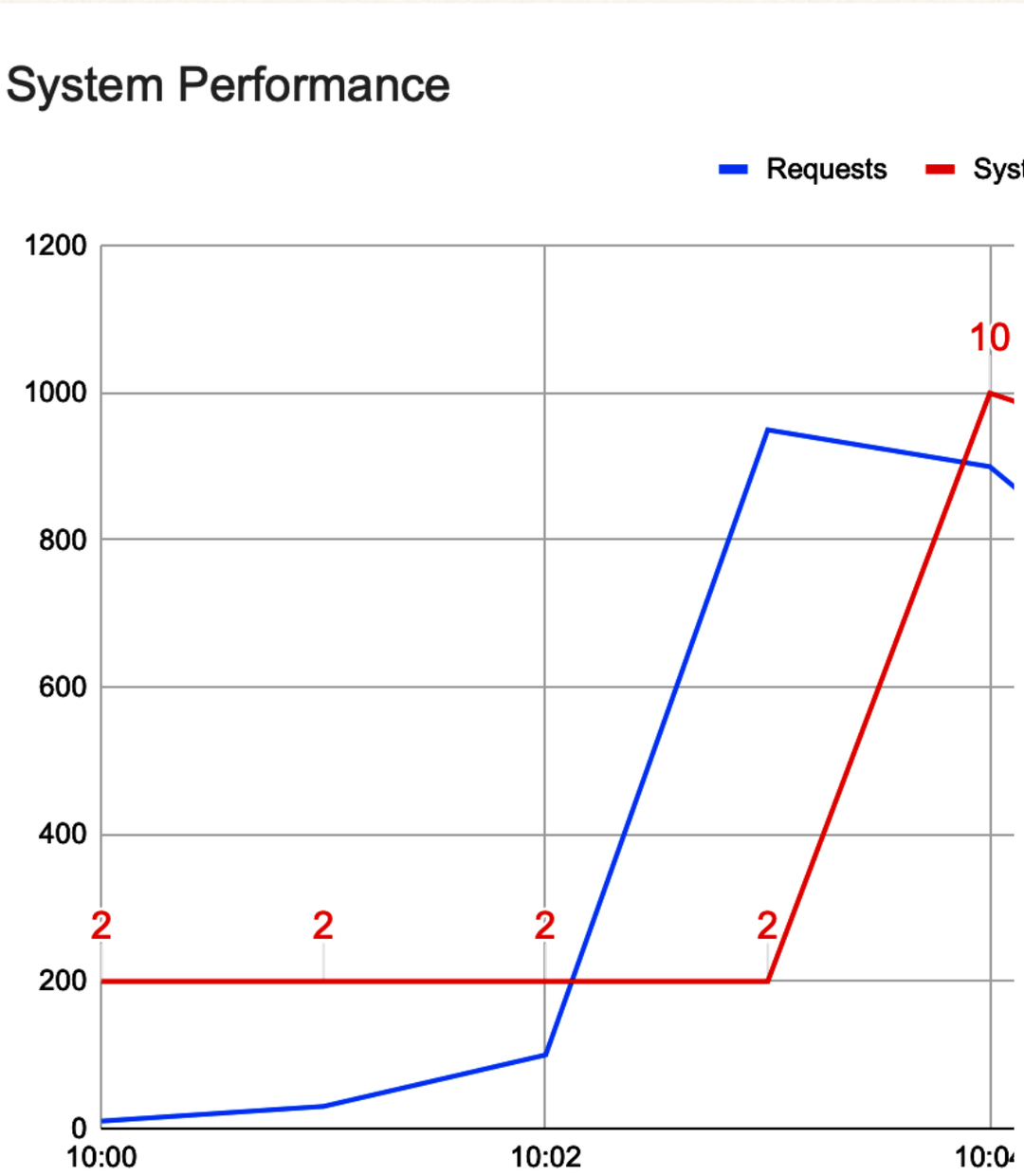
- up to 90sec to create new pod
- up to 12min to create new pod on new AWS instance
- “Delay to start new pod” [4]

Autoscaler quick demo

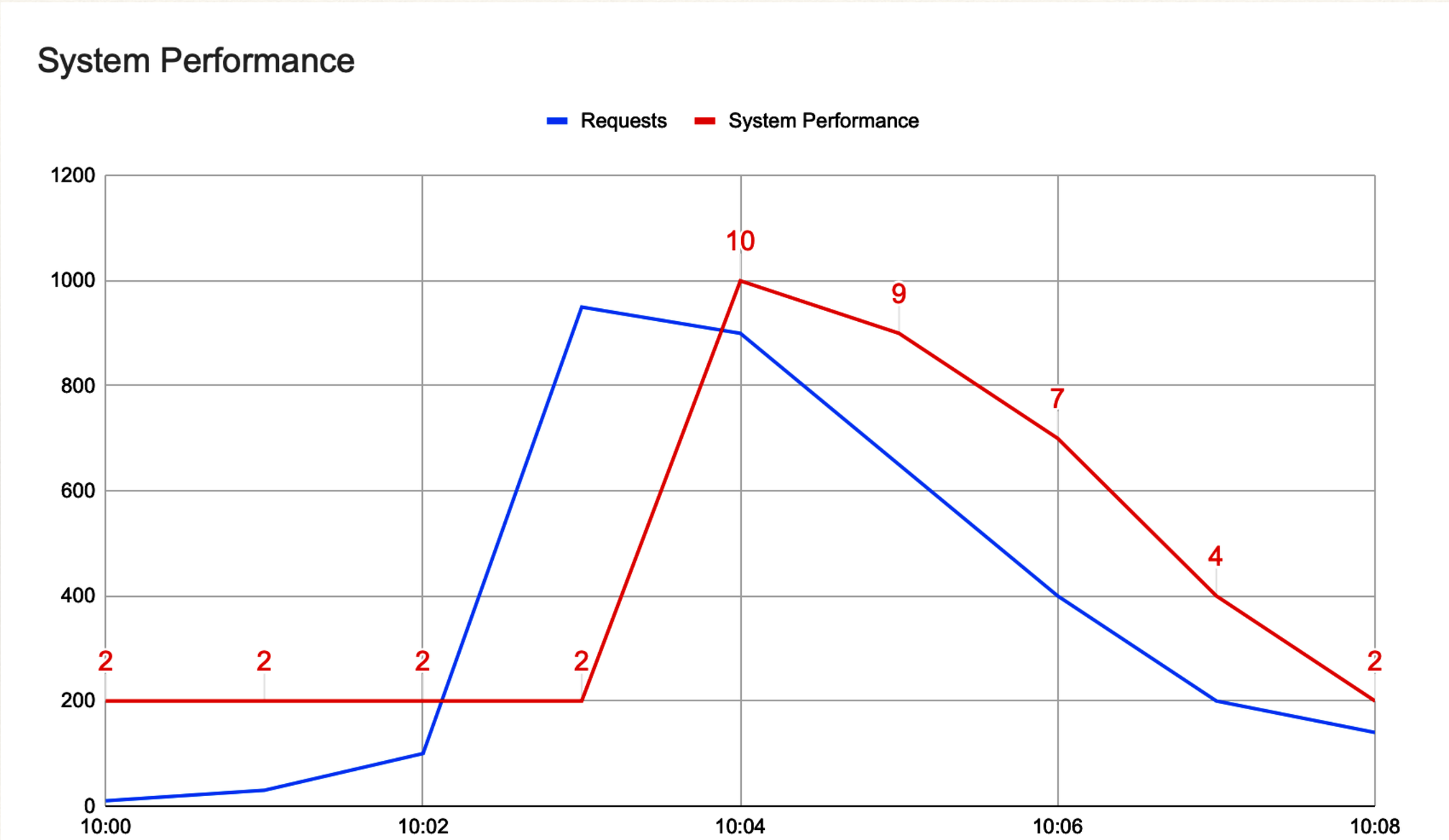
Cat Picture As a Service



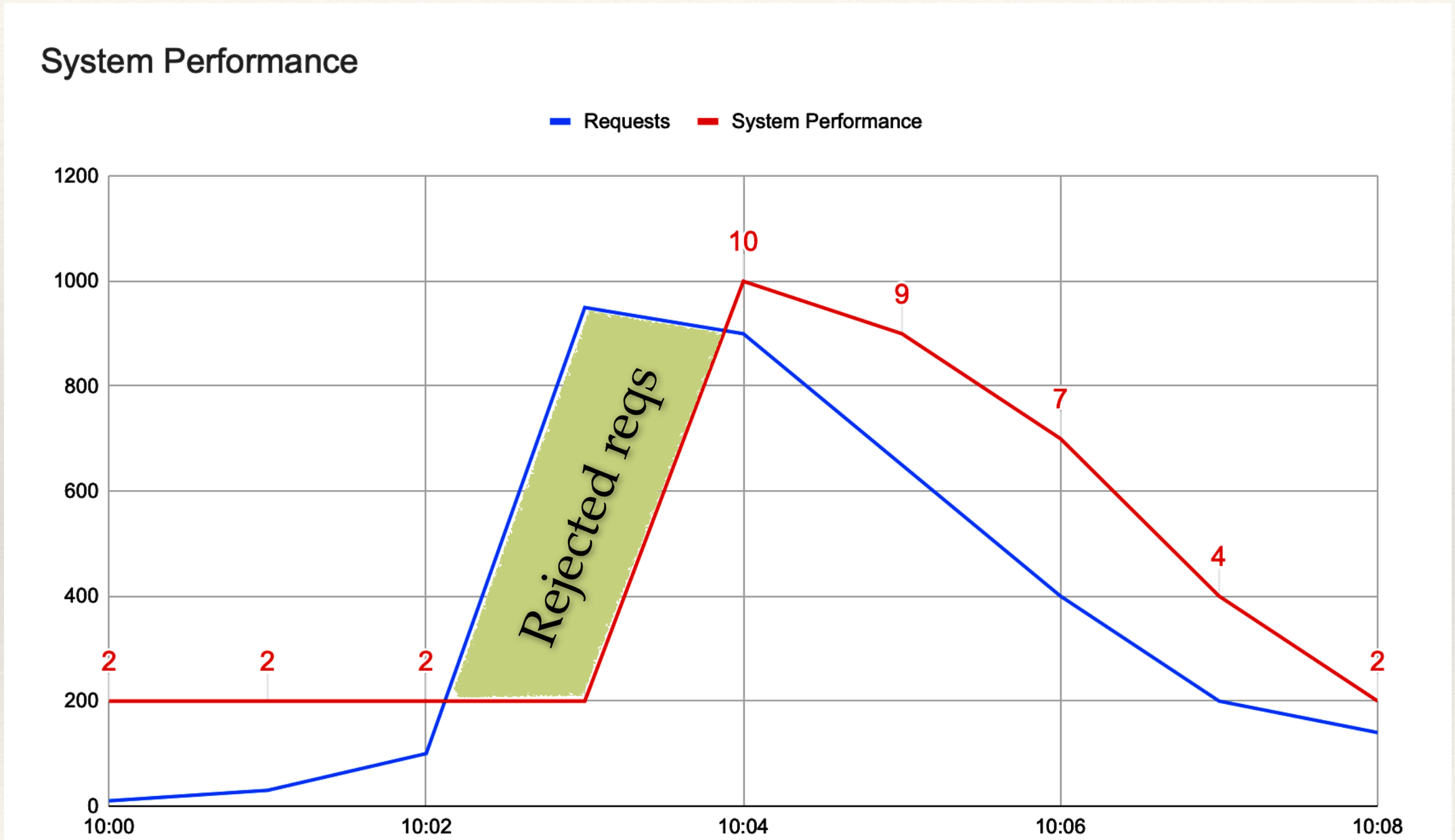
Autoscaler quick demo 2



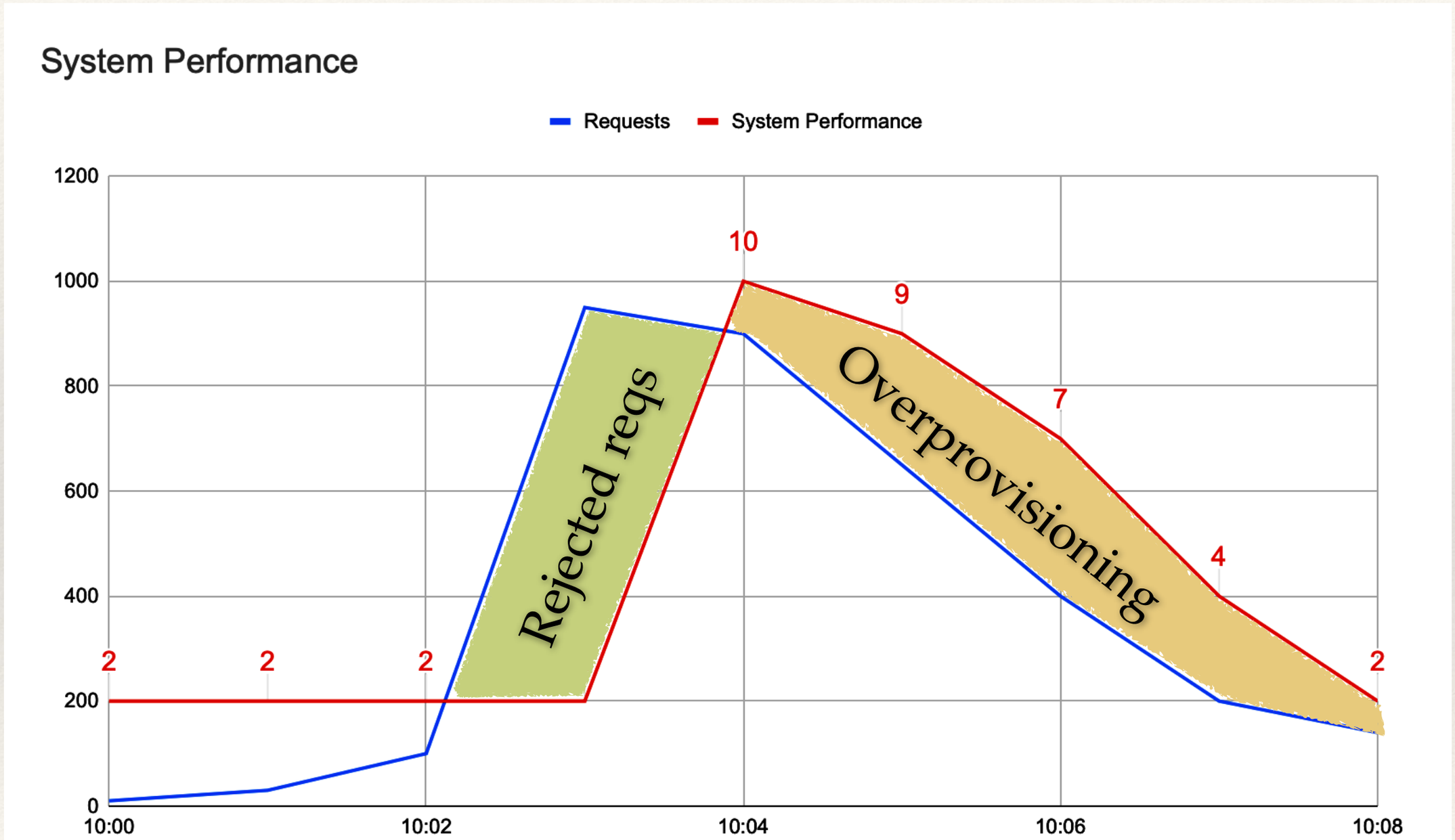
Autoscaler quick demo 2



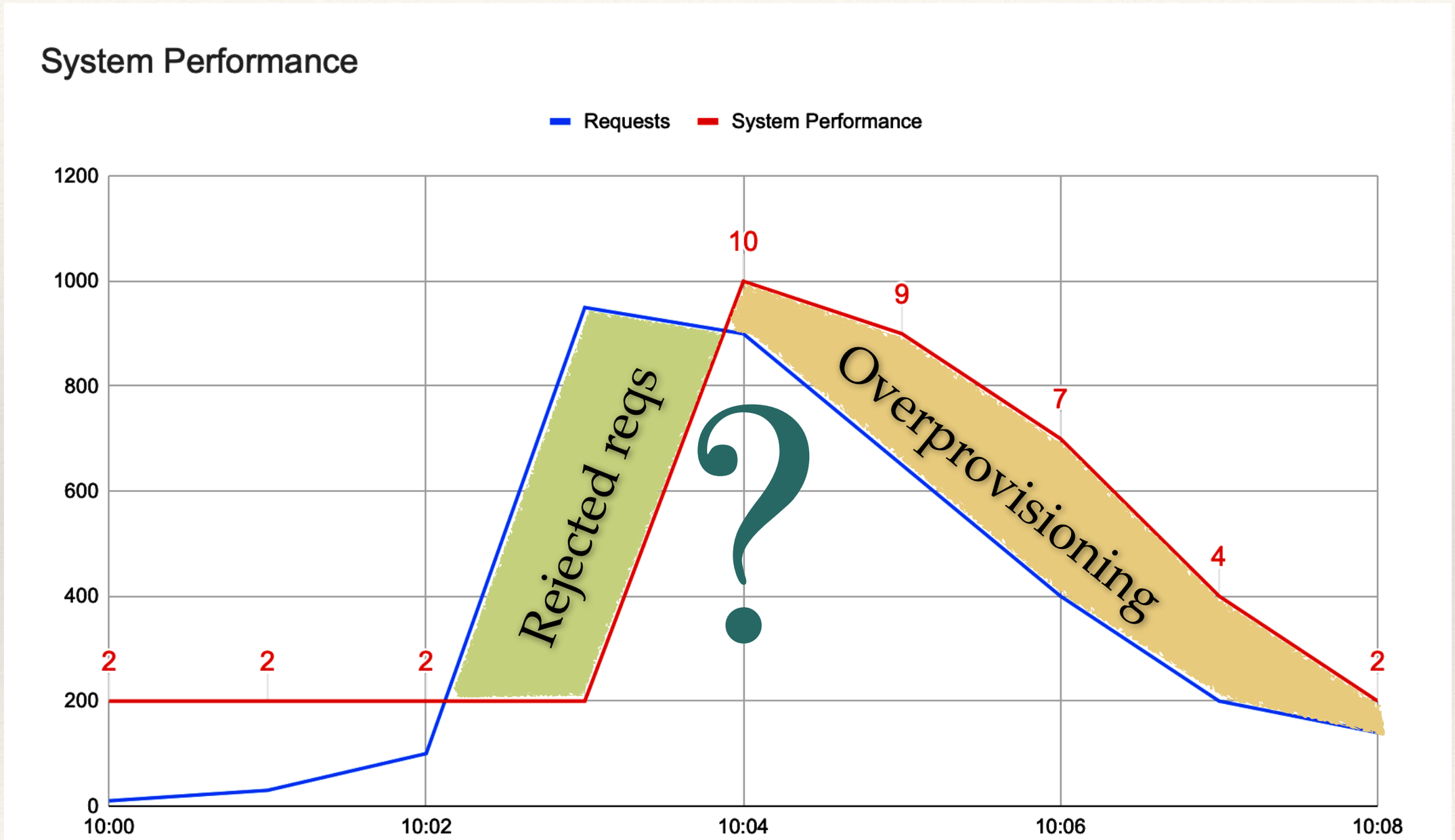
Autoscaler quick demo 2



Autoscaler quick demo 2

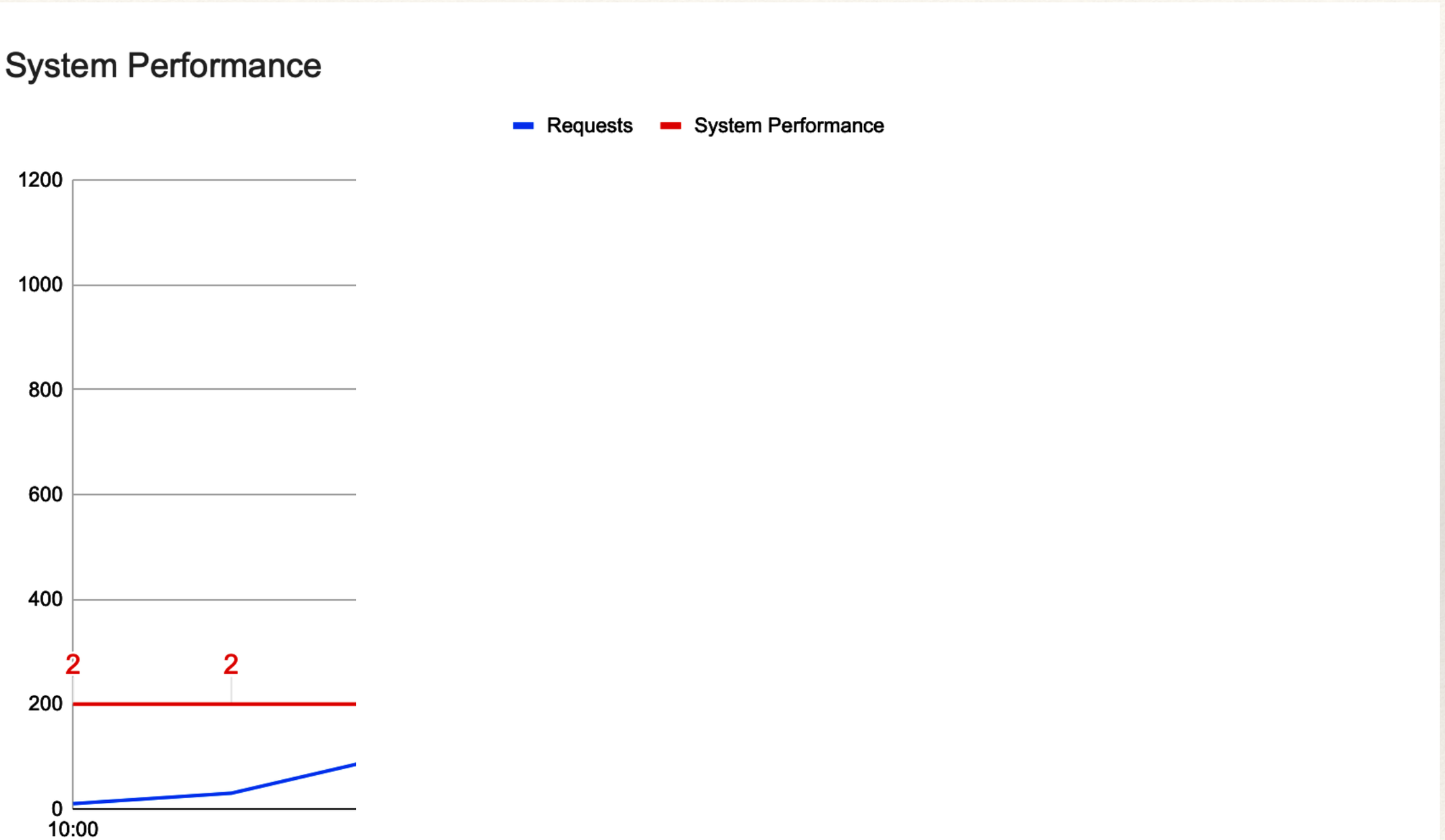


Autoscaler quick demo 2

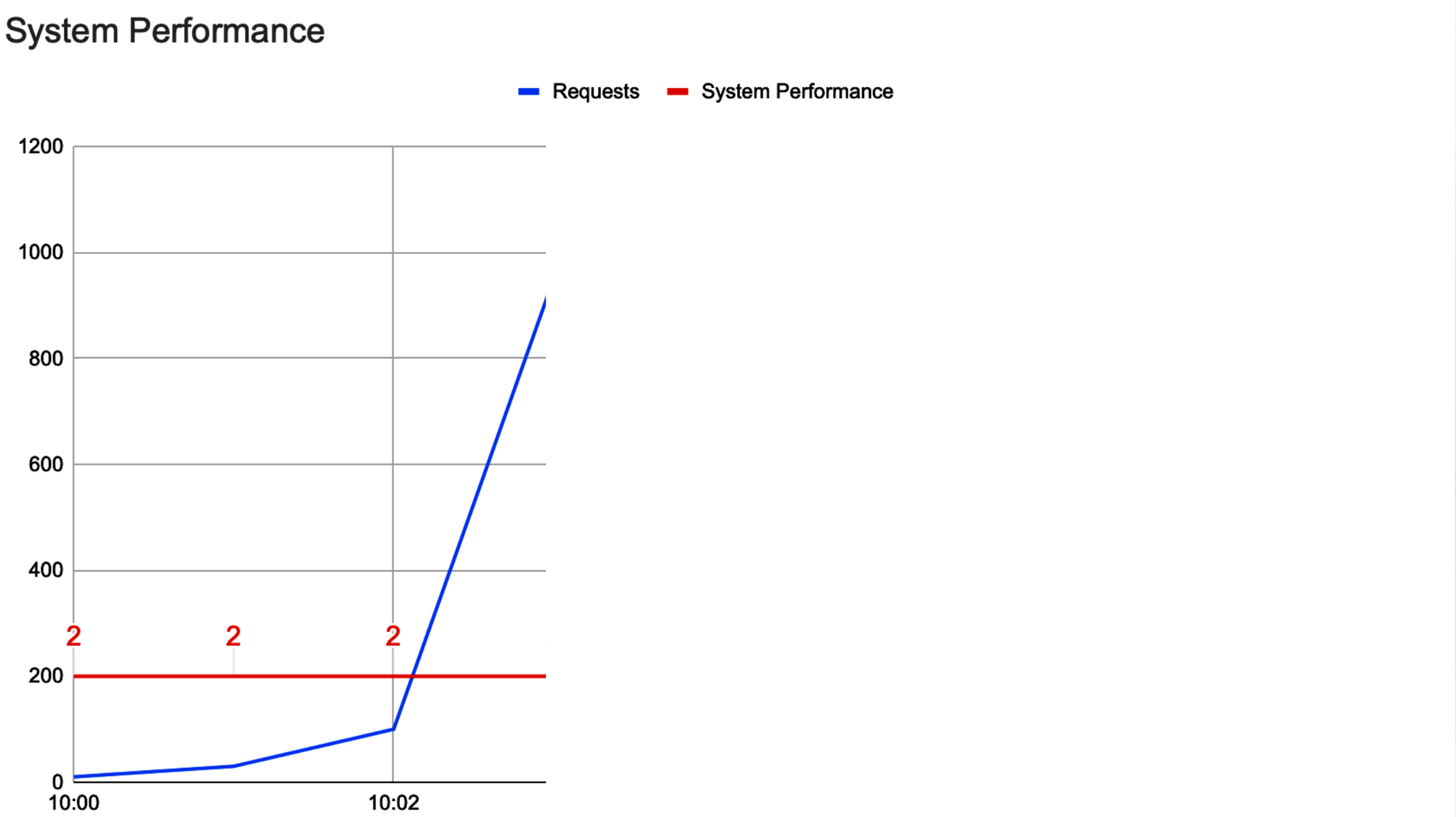


Nope, worse

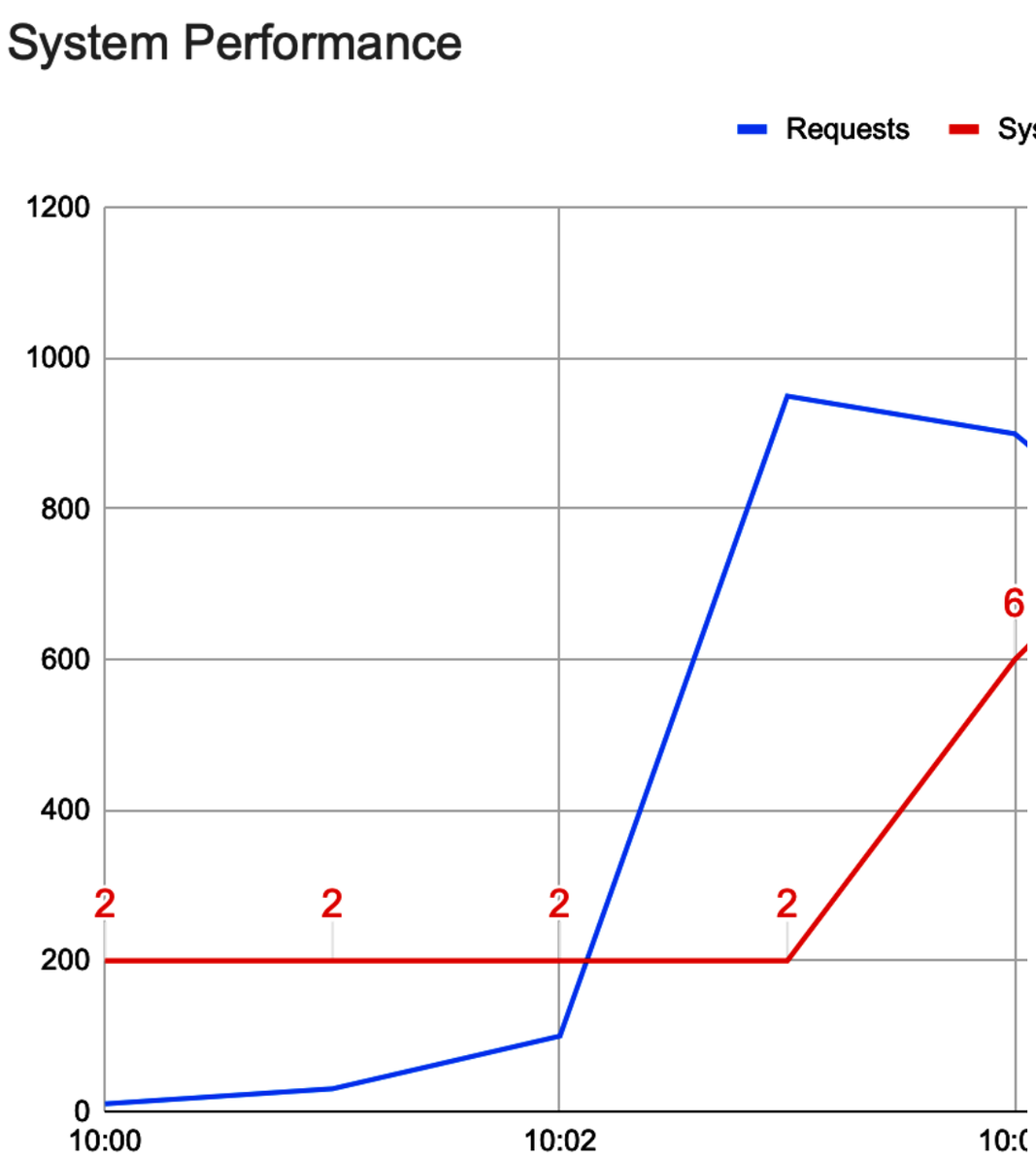
Autoscaler quick demo 3



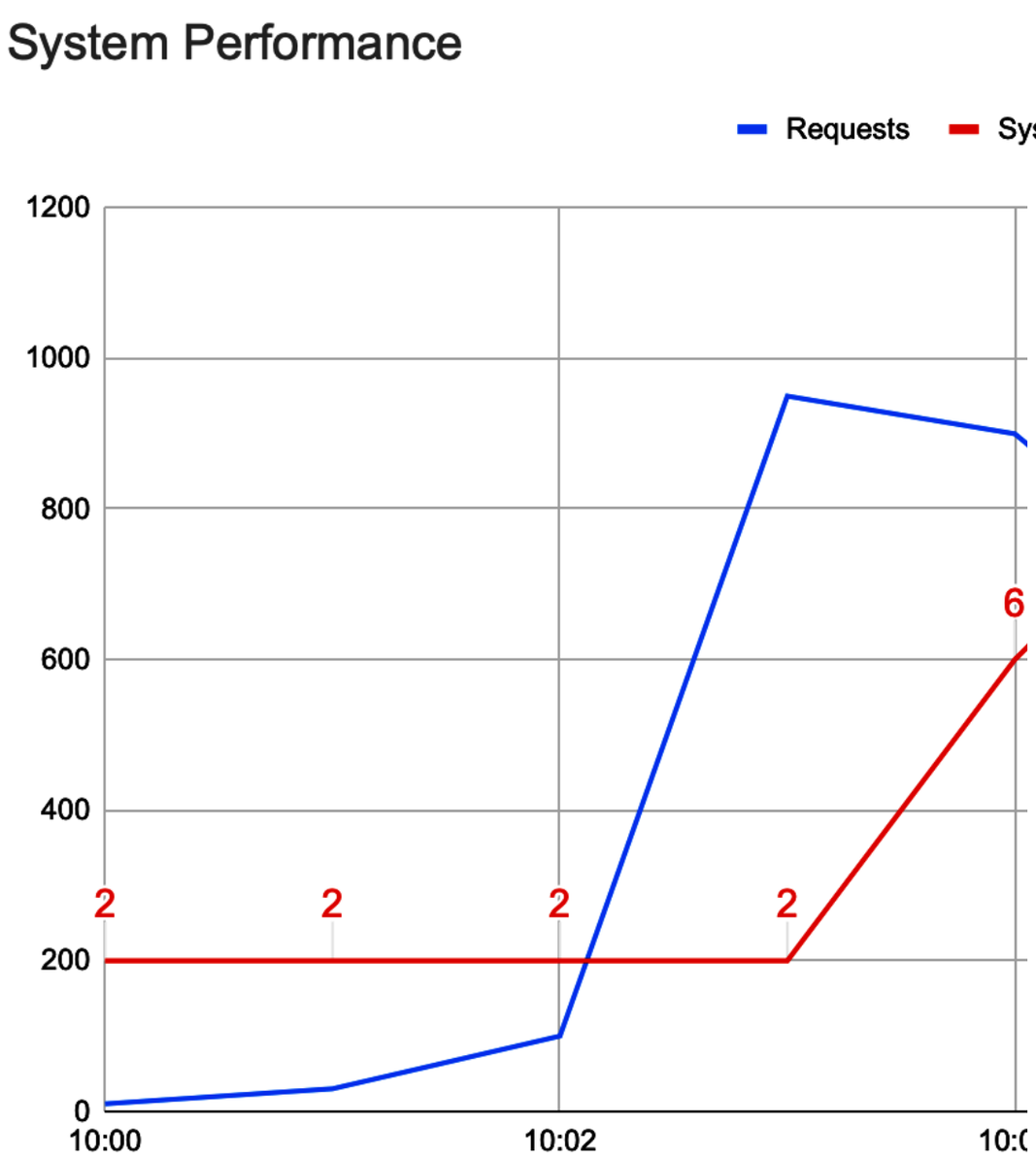
Autoscaler quick demo 3



Autoscaler quick demo 3

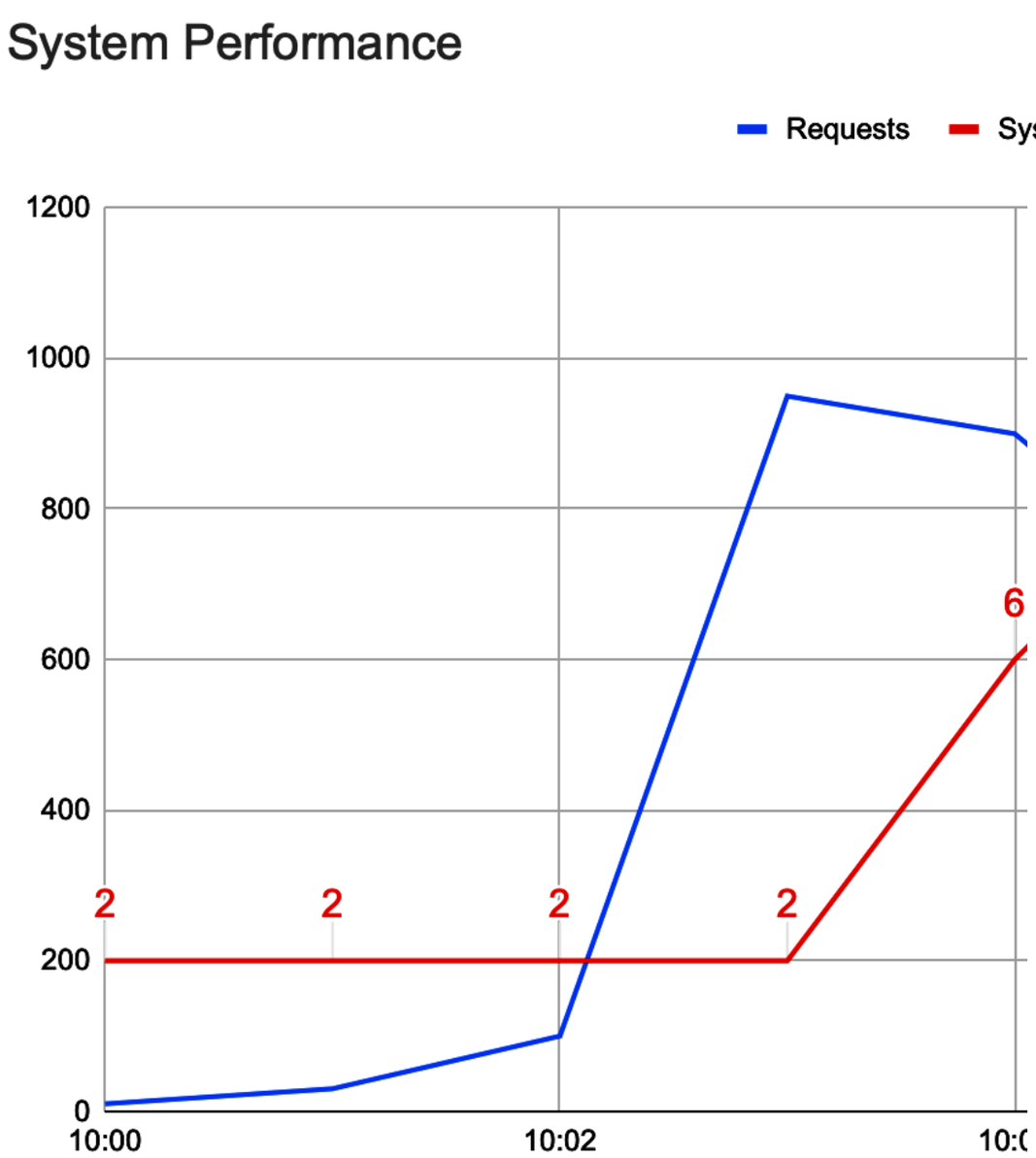


Autoscaler quick demo 3



Why 6 ??

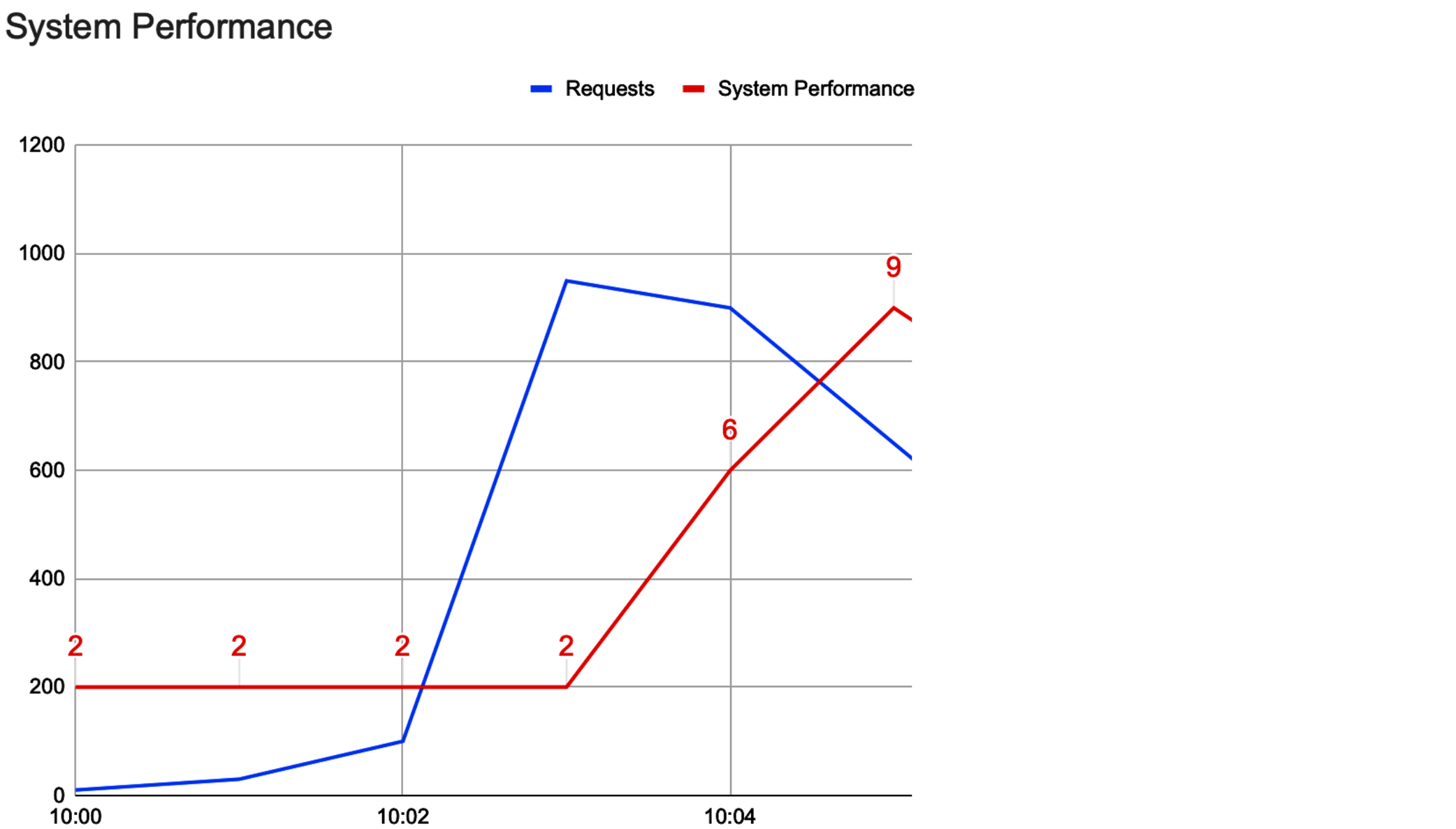
Autoscaler quick demo 3



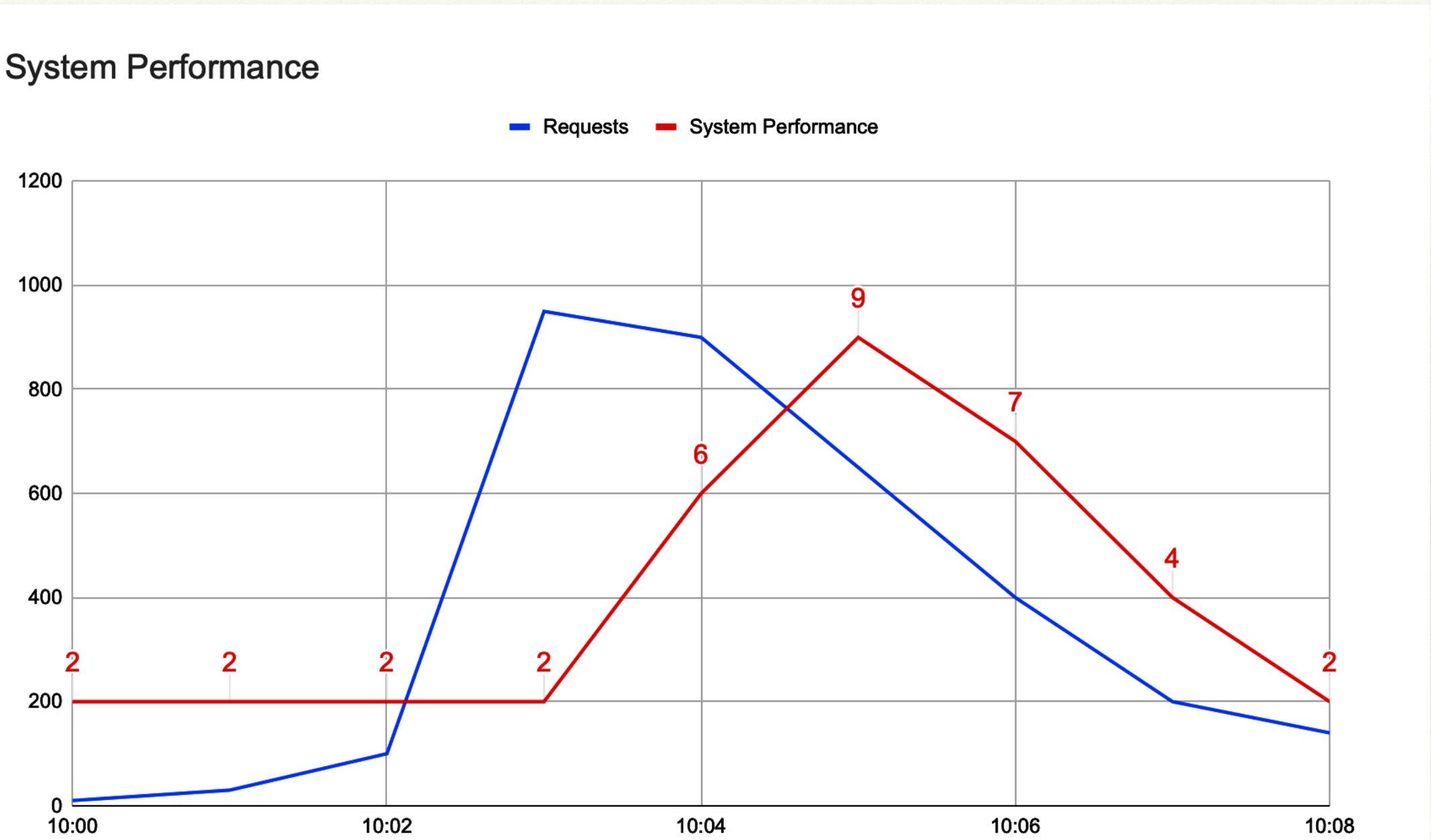
Why 6 ??

- new1 = oldReplicas + 4
- new2 = oldReplicas * 2
- newReplicas = max(new1, new2)

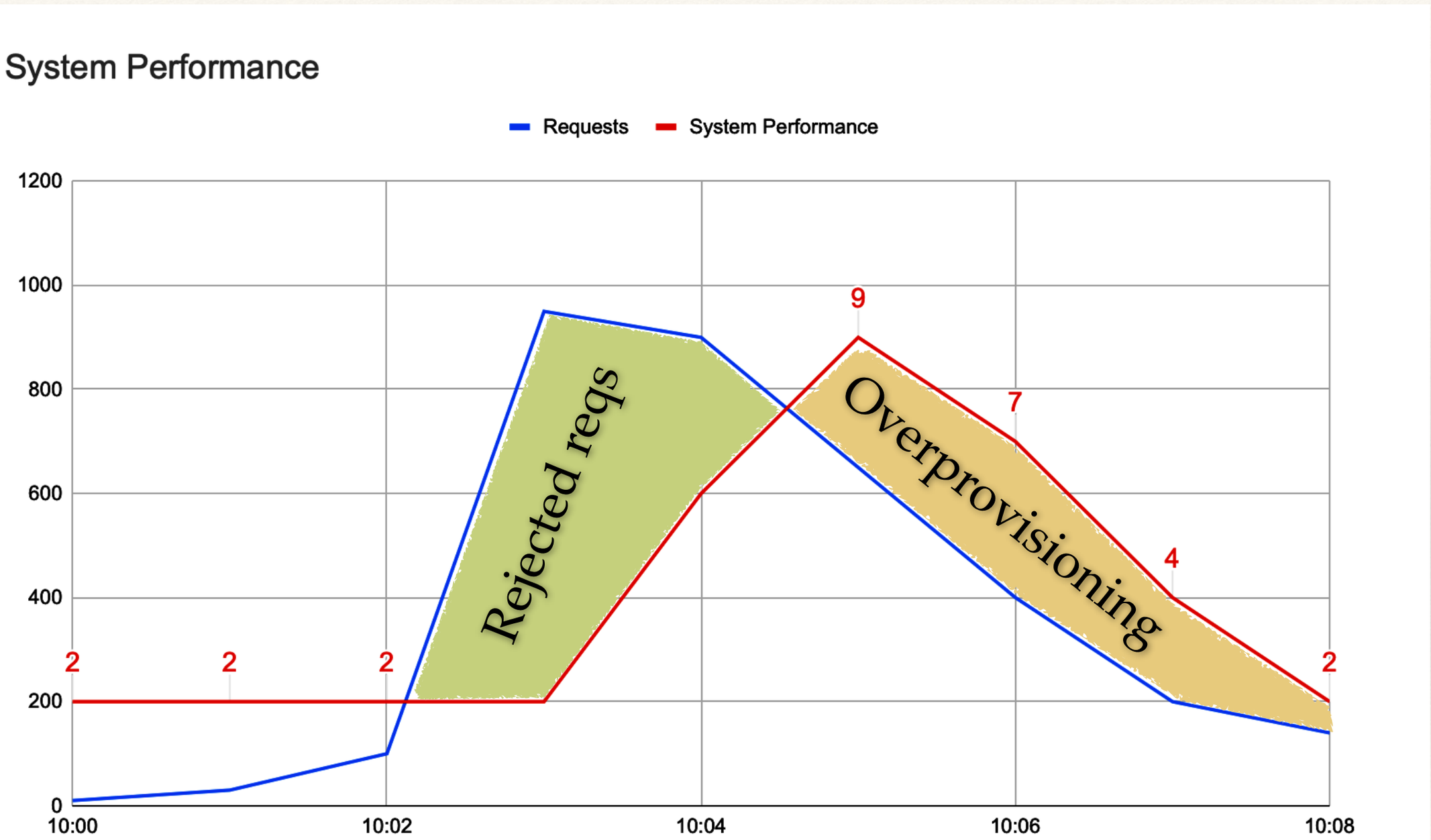
Autoscaler quick demo 3



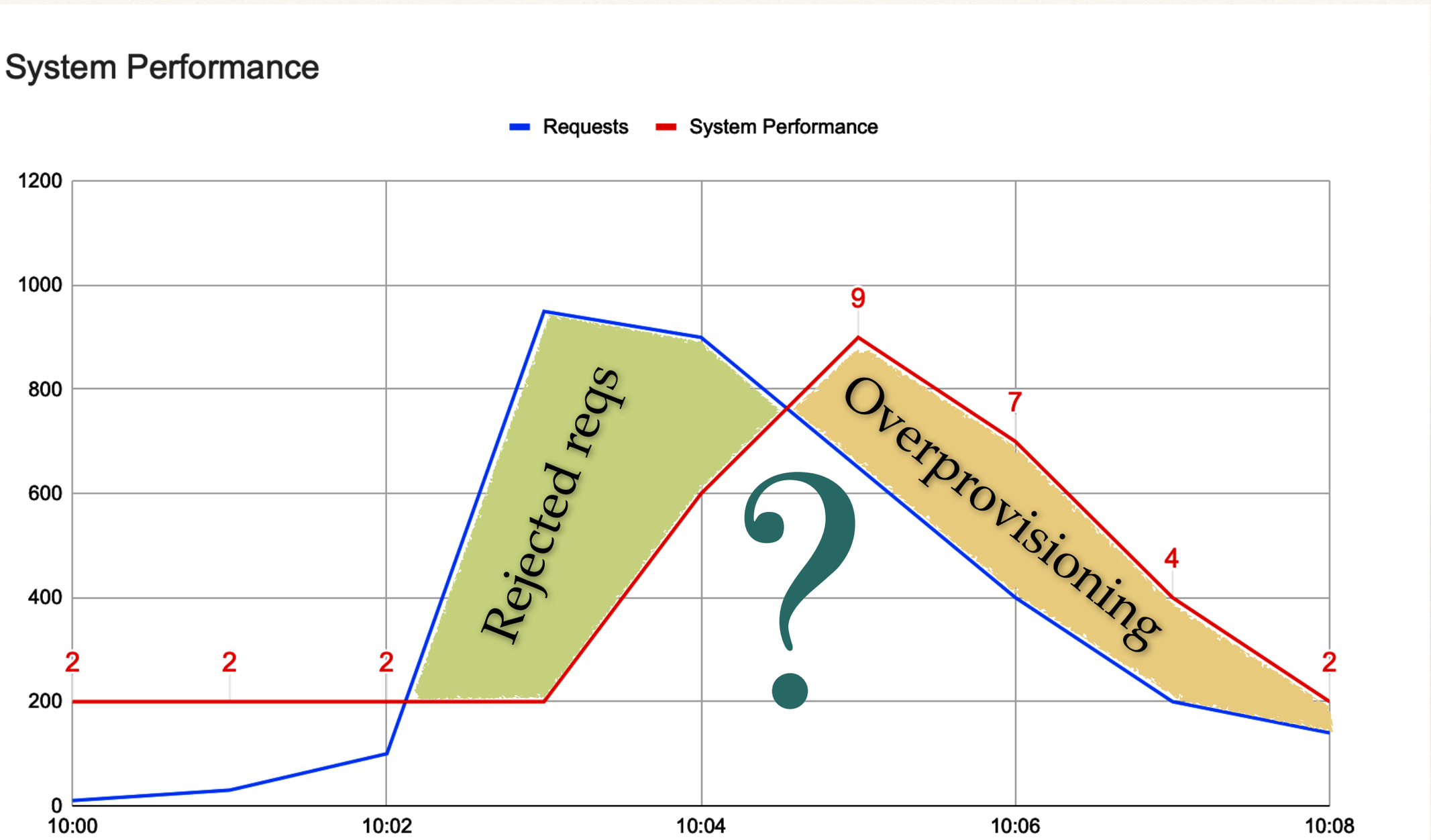
Autoscaler quick demo 3



Autoscaler quick demo 3



Autoscaler quick demo 3



Almost

HPA under the microscope

- HPA Configuration:
  - maxReplicas
  - minReplicas
  - metrics

- Hardcoded Constants:
  - scaleUpLimitFactor = 2
  - scaleUpLimitMinimum = 4

- Cluster Configuration
  - stabilizationWindow = 300
  - tolerance = 0.1

HPA under the microscope

- HPA Configuration:
  - maxReplicas
  - minReplicas
  - metrics

cluster size

scale reason

- Hardcoded Constants:
  - scaleUpLimitFactor = 2
  - scaleUpLimitMinimum = 4

- Cluster Configuration
  - stabilizationWindow = 300
  - tolerance = 0.1

HPA under the microscope

- HPA Configuration:
  - maxReplicas
  - minReplicas
  - metrics

cluster size

scale reason

- Hardcoded Constants:
  - scaleUpLimitFactor = 2 — max(10*2, 10+4) -> 20
  - scaleUpLimitMinimum = 4 — max(1*2, 1+4) -> 5

- Cluster Configuration
  - stabilizationWindow = 300
  - tolerance = 0.1

HPA under the microscope

- HPA Configuration:
  - maxReplicas
  - minReplicas
  - metrics

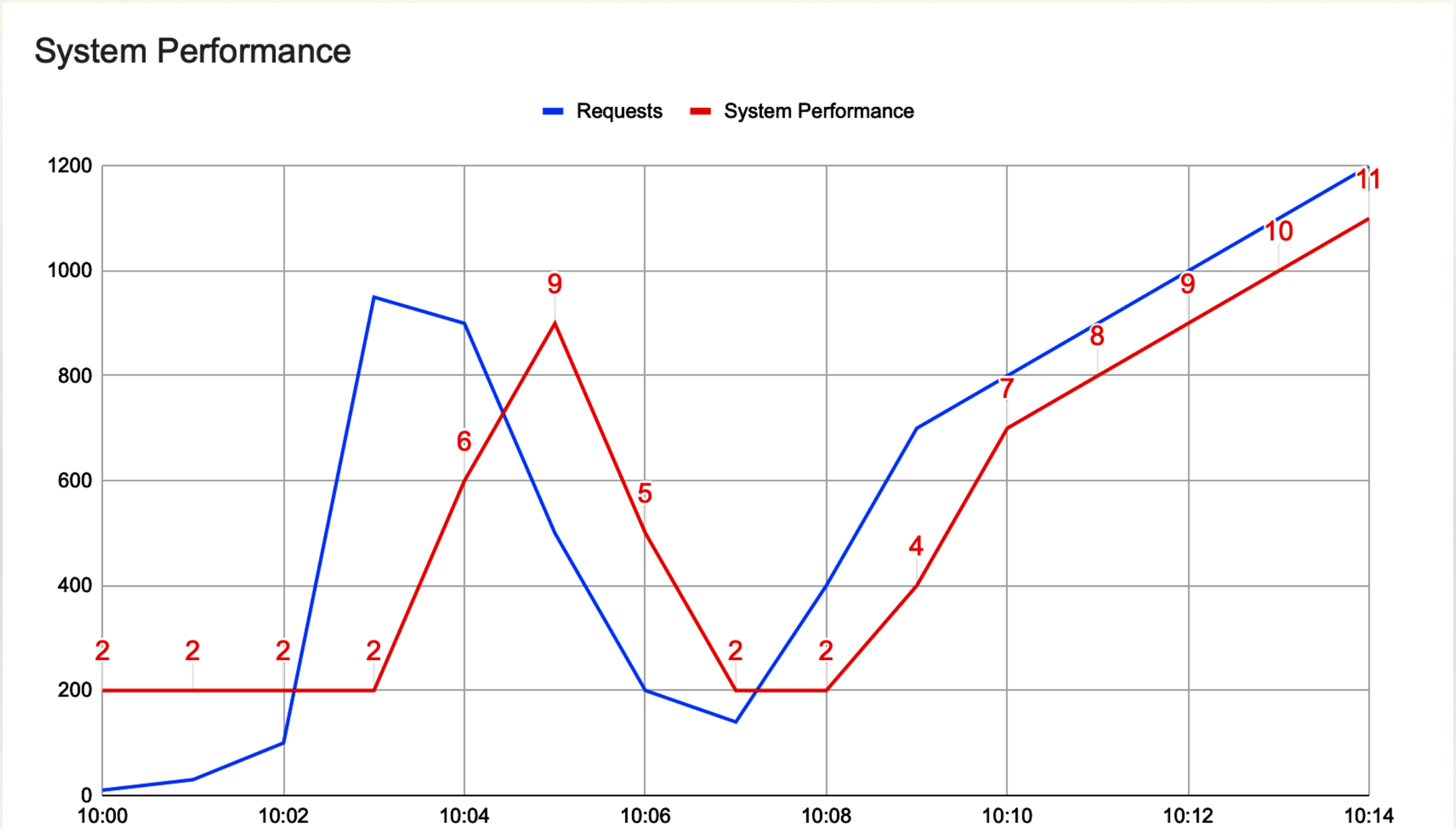
  → **cluster size**

  → **scale reason**

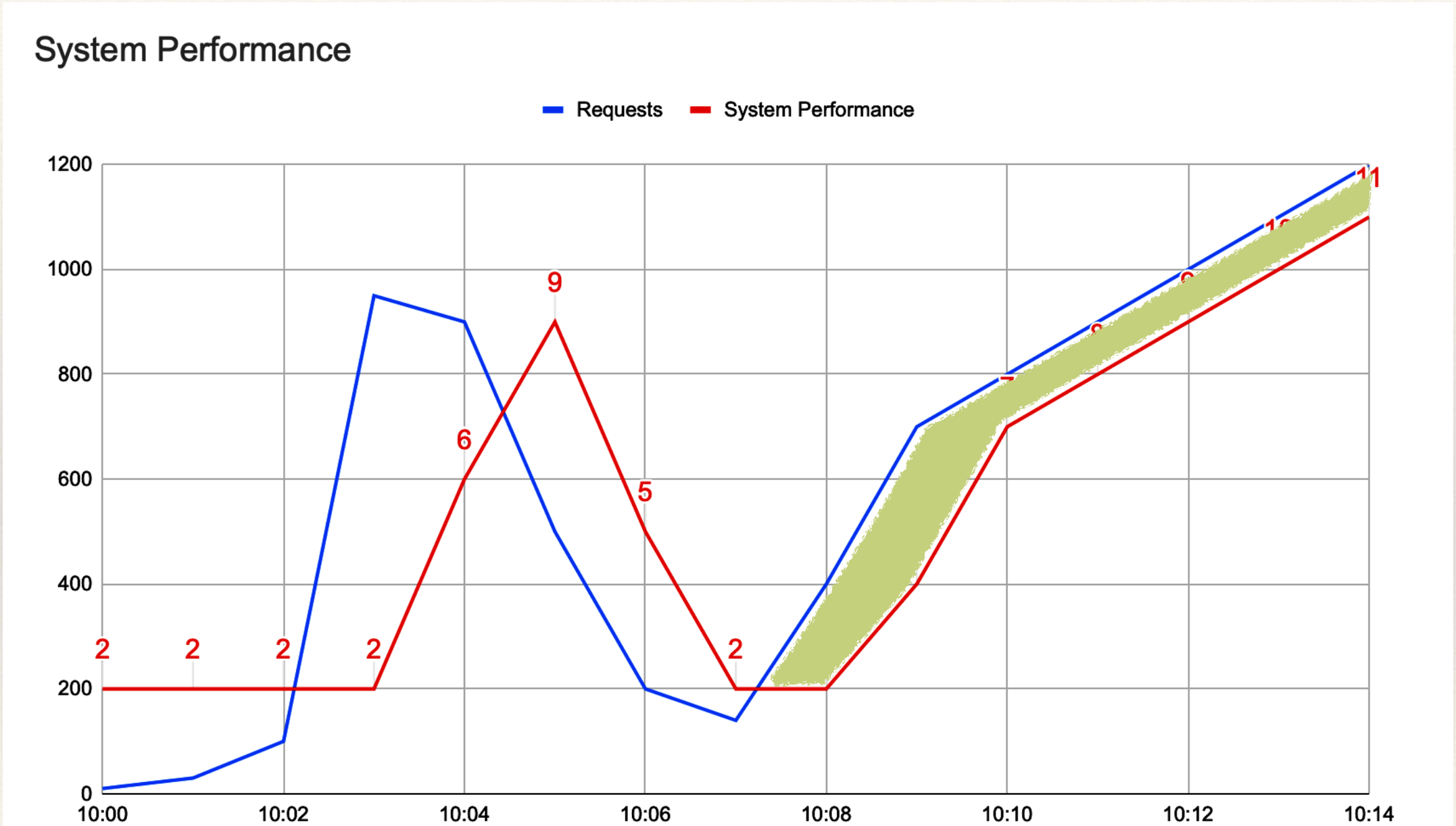
- Hardcoded Constants:
  - scaleUpLimitFactor = 2 — max(10*2, 10+4) -> 20
  - scaleUpLimitMinimum = 4 — max(1*2, 1+4) -> 5

- Cluster Configuration
  - stabilizationWindow = 300 → **stabilize scale down**
  - tolerance = 0.1 → **scale tolerance**

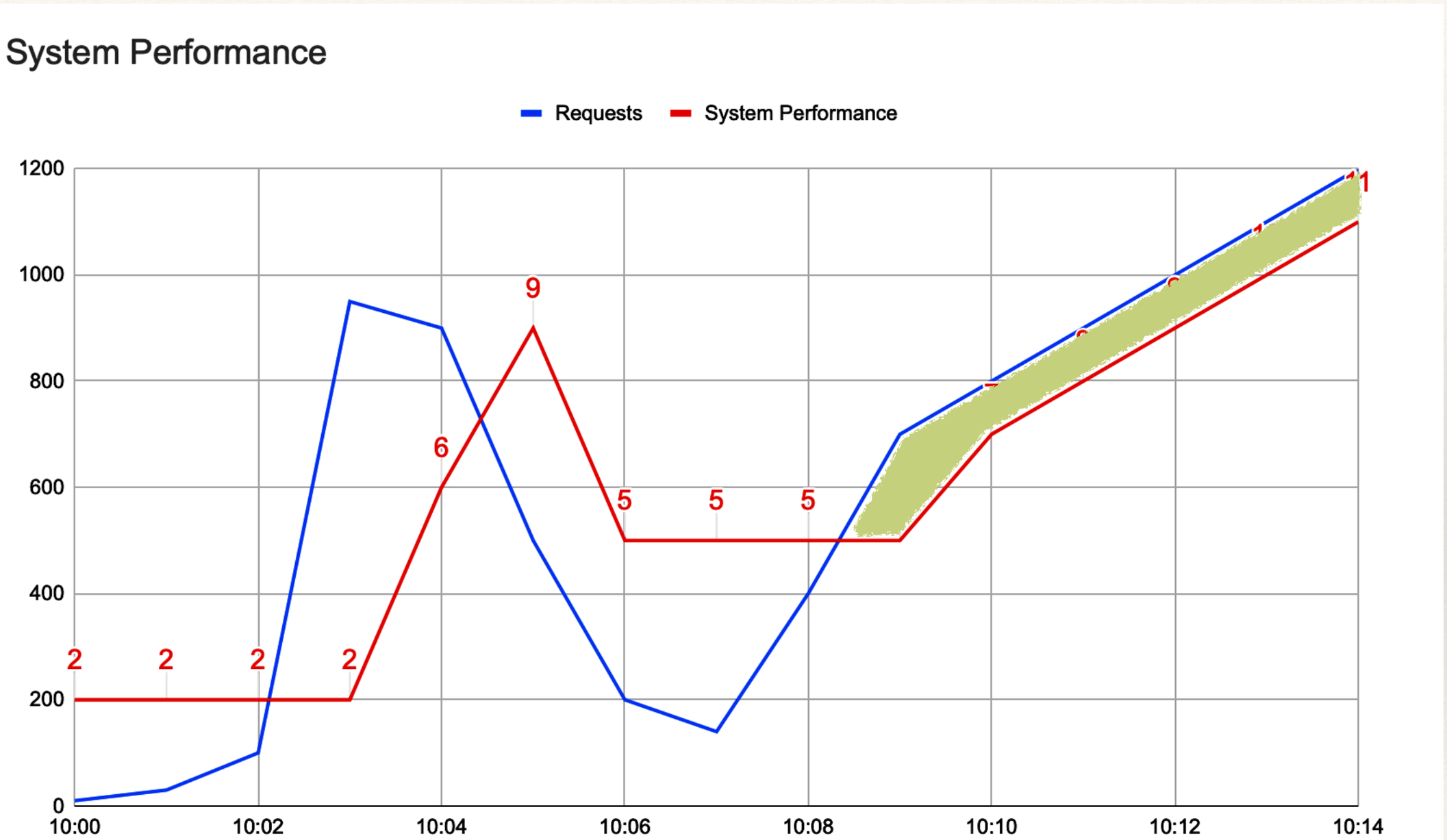
Problem: Second Spike



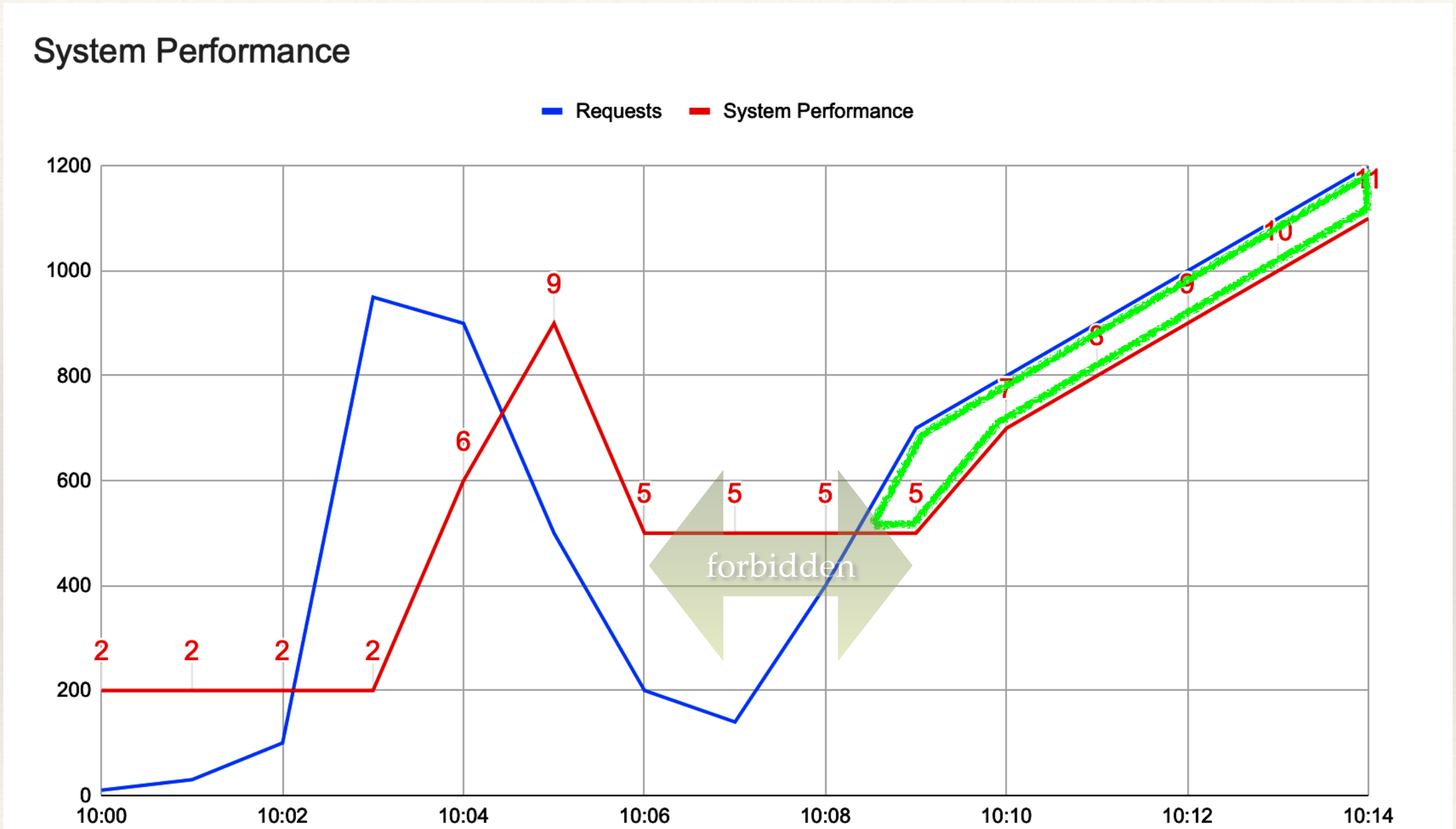
Problem: Second Spike



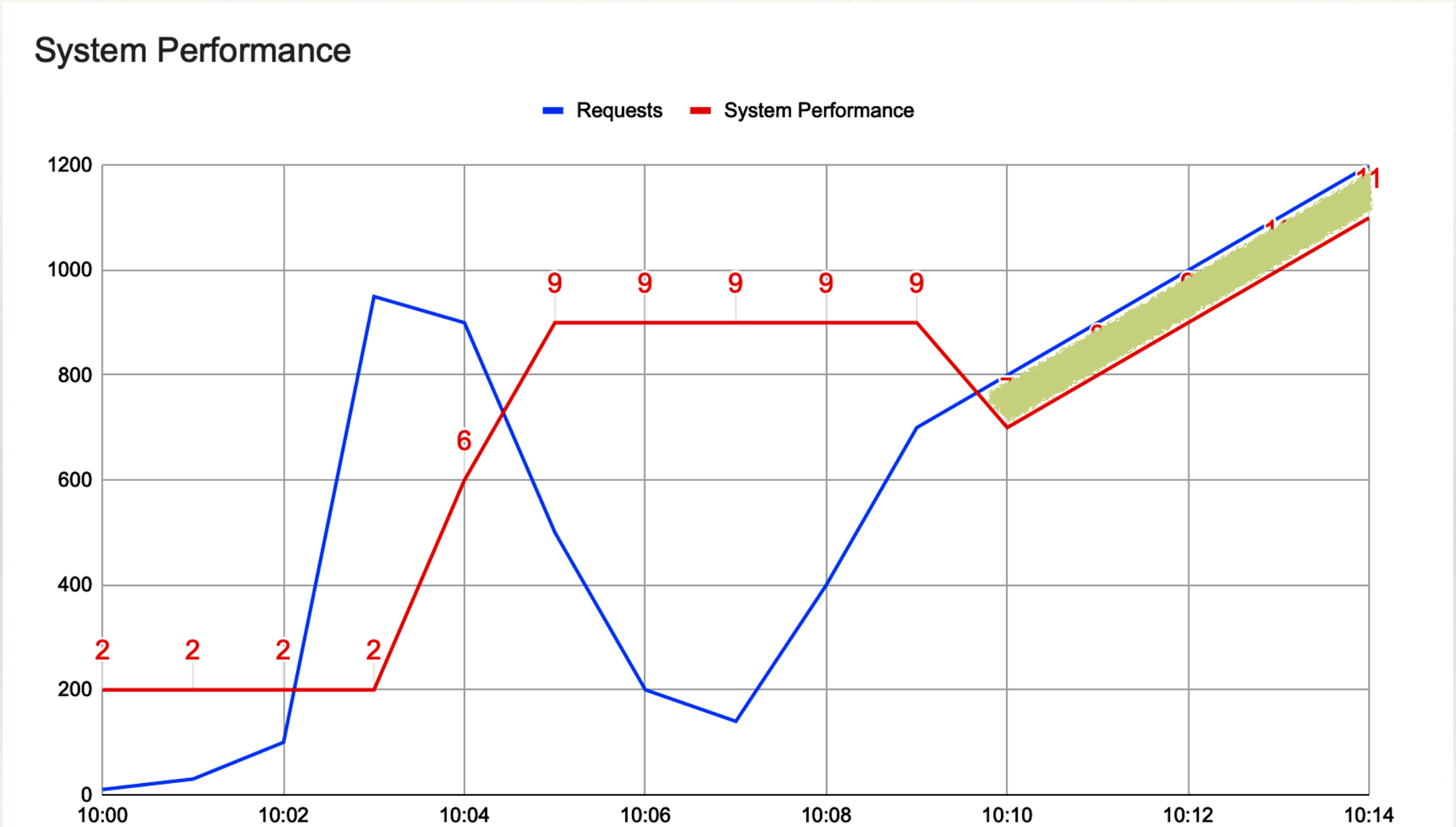
Autoscaler: Forbidden Window



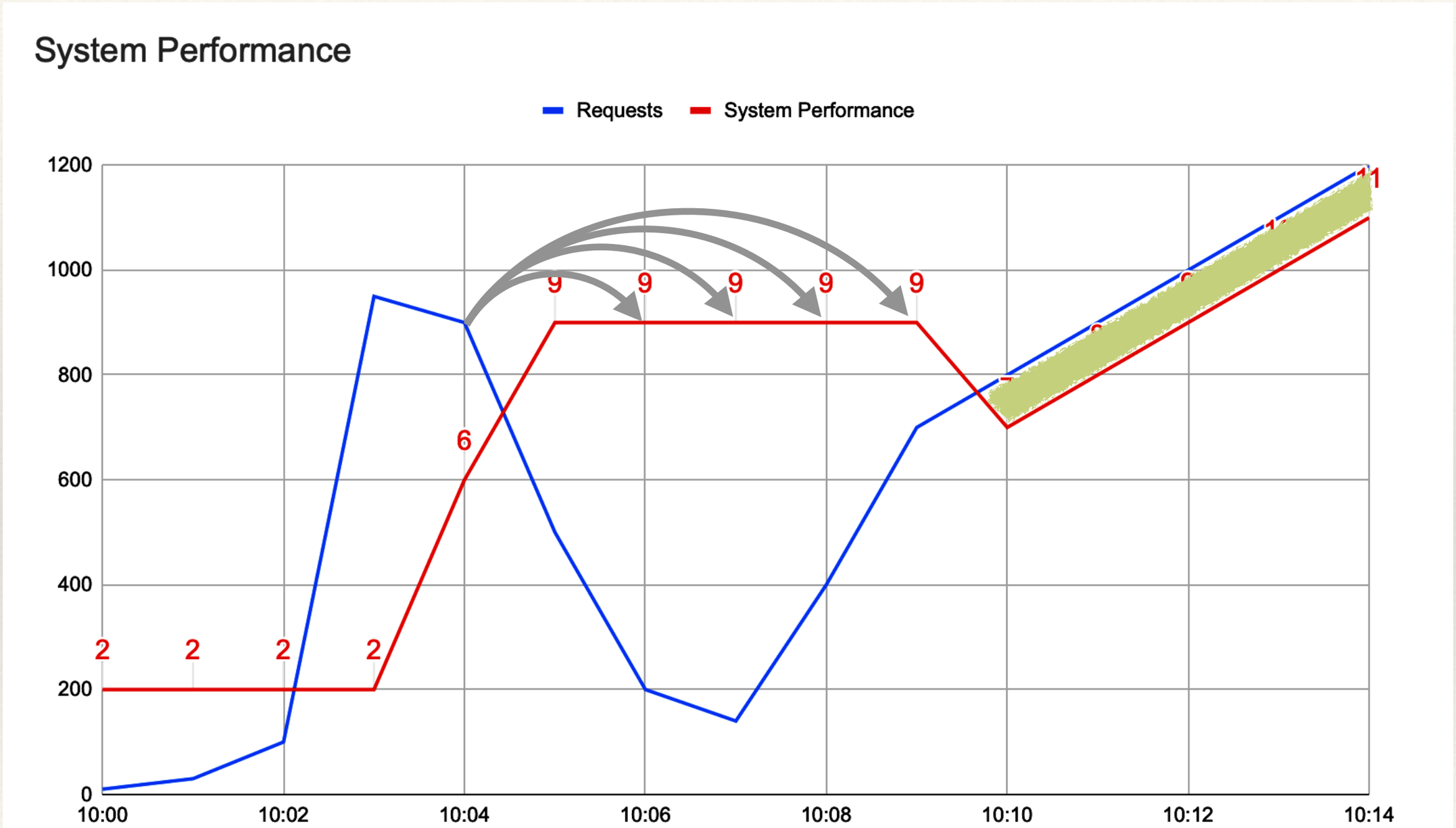
Autoscaler: Forbidden Window



Autoscaler: Stabilization



Autoscaler: Stabilization



HPA Problems

- Cannot scale slower / faster
- Cannot delay the next scale
- Cannot control the stabilization window
- All important configurations are hardcoded / cluster-wide

HPA under the microscope

- HPA Configuration:
  - maxReplicas
  - minReplicas
  - metrics

  → **cluster size**

  → **scale reason**

- Hardcoded Constants:
  - scaleUpLimitFactor = 2 — max(10*2, 10+4) -> 20
  - scaleUpLimitMinimum = 4 — max(1*2, 1+4) -> 5

- Cluster Configuration
  - stabilizationWindow = 300 → **stabilize scale down**
  - tolerance = 0.1 → **scale tolerance**

Per-HPA Configuration

- Critical Web Requests
- Ideal configuration:
  - Scale Up:
    - Infinite Velocity
  - Scale Down:
    - Long Stabilization

- Critical Data Events Processing
- Ideal configuration:
  - Scale Up:
    - Infinite Velocity
  - Scale Down:
    - No Stabilization

Per-HPA Configuration

- Critical Web Requests
- Ideal configuration:
  - Scale Up:
    - Infinite Velocity
  - Scale Down:
    - Long Stabilization

- Critical Data Events Processing
- Ideal configuration:
  - Scale Up:
    - Infinite Velocity
  - Scale Down:
    - No Stabilization

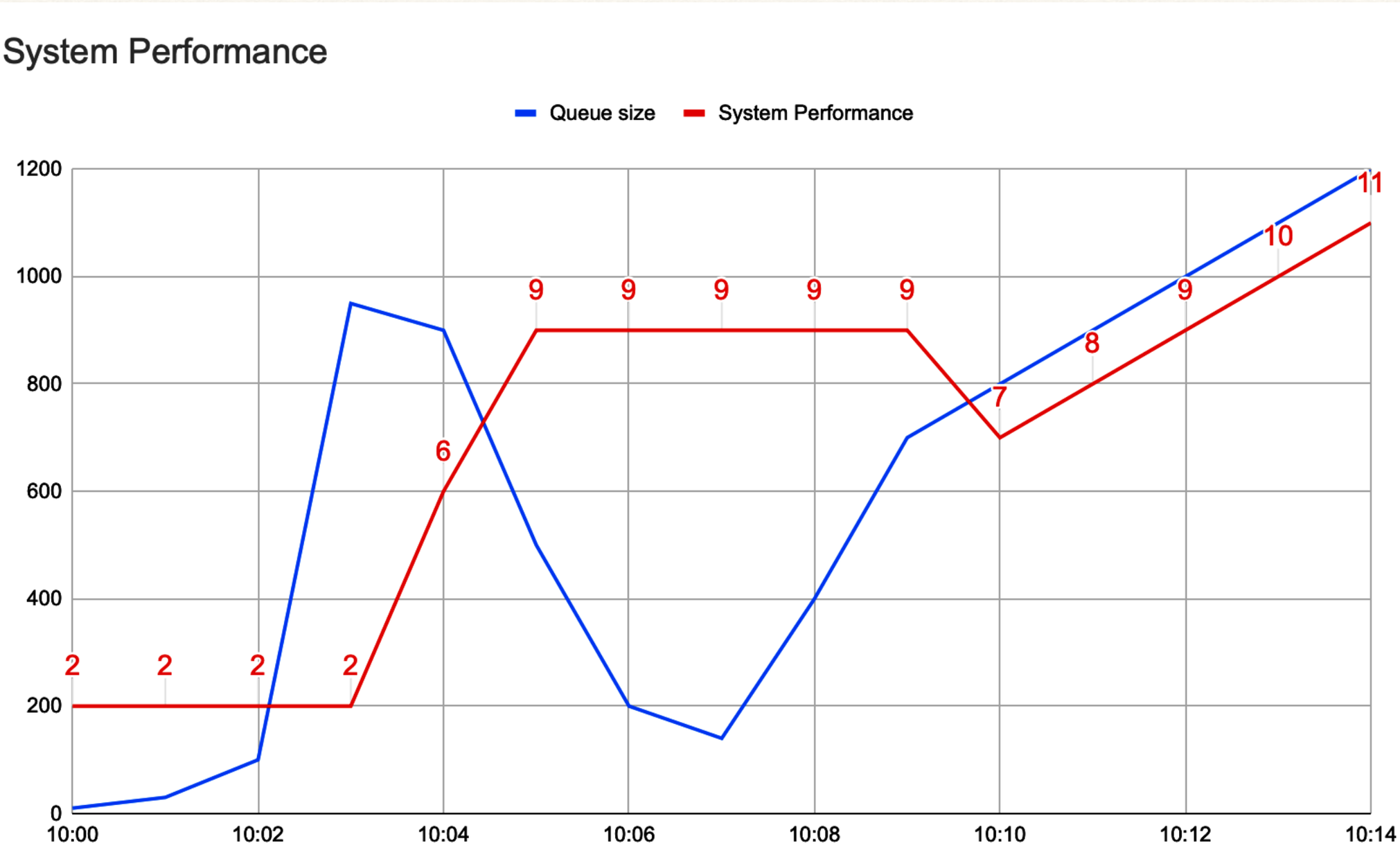
Per-HPA Configuration

- Web Requests
- Ideal configuration:
  - Scale Up:
    - Infinite Velocity
  - Scale Down:
    - Long Stabilization

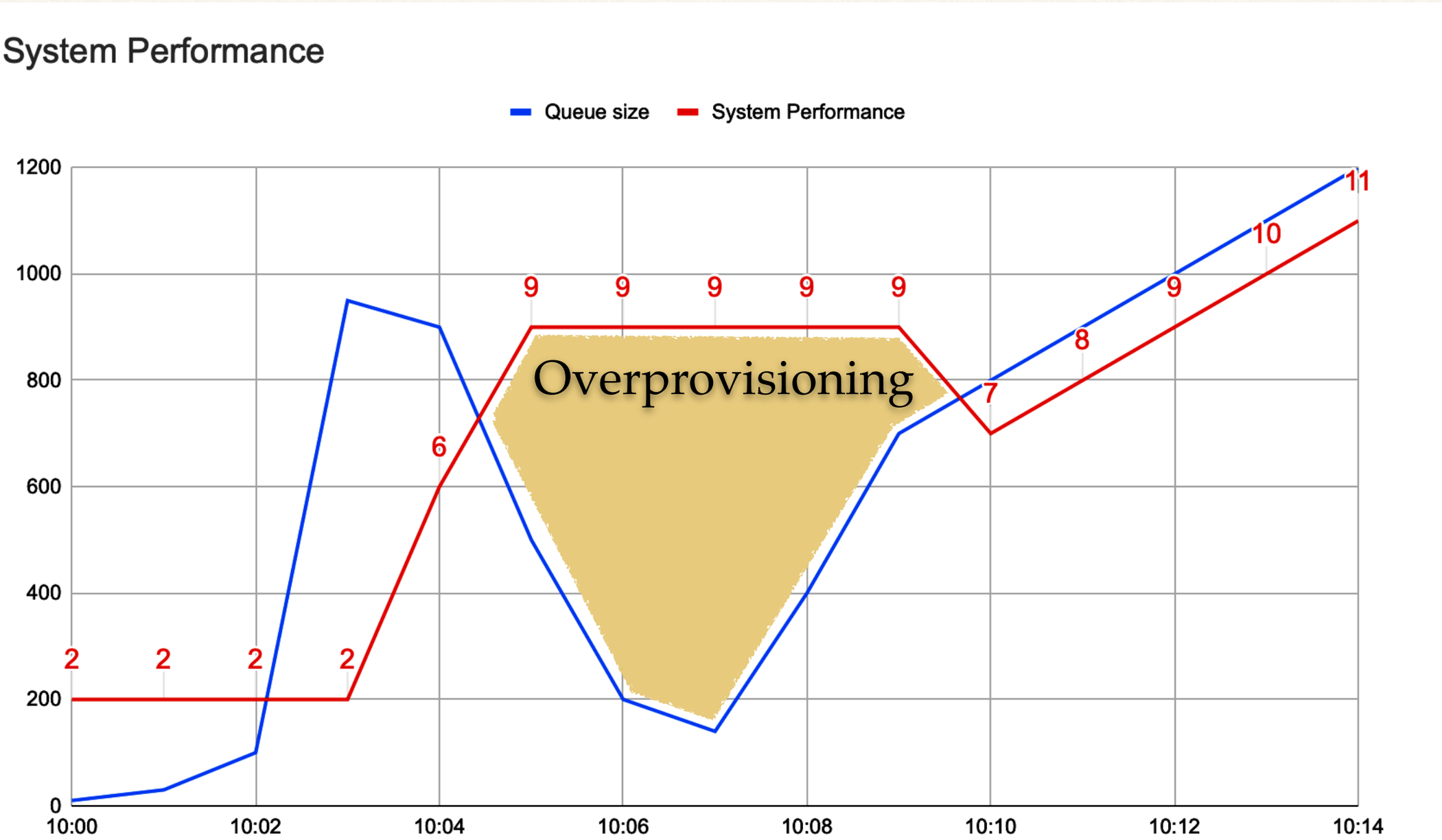
- Data Events Processing
- Ideal configuration:
  - Scale Up:
    - Infinite Velocity
  - Scale Down:
    - No Stabilization

Impossible with cluster-wide configuration

Data Event Processing



Data Event Processing



Solutions

Solutions

- Postmates: Configurable HPA [2]
- Datadog: Watermark Podautoscaler [5]
- Vanilla Kubernetes Solution (hopefully, in k8s-1.17)

Configurable HPA

- CRD + Controller
- Configuration (per deployment):
  - ScaleUpLimitFactor
  - ScaleUpLimitMinimum

  - UpscaleForbiddenWindowSeconds
  - DownscaleForbiddenWindowSeconds.

Configurable HPA

- CRD + Controller
- Configuration (per deployment):

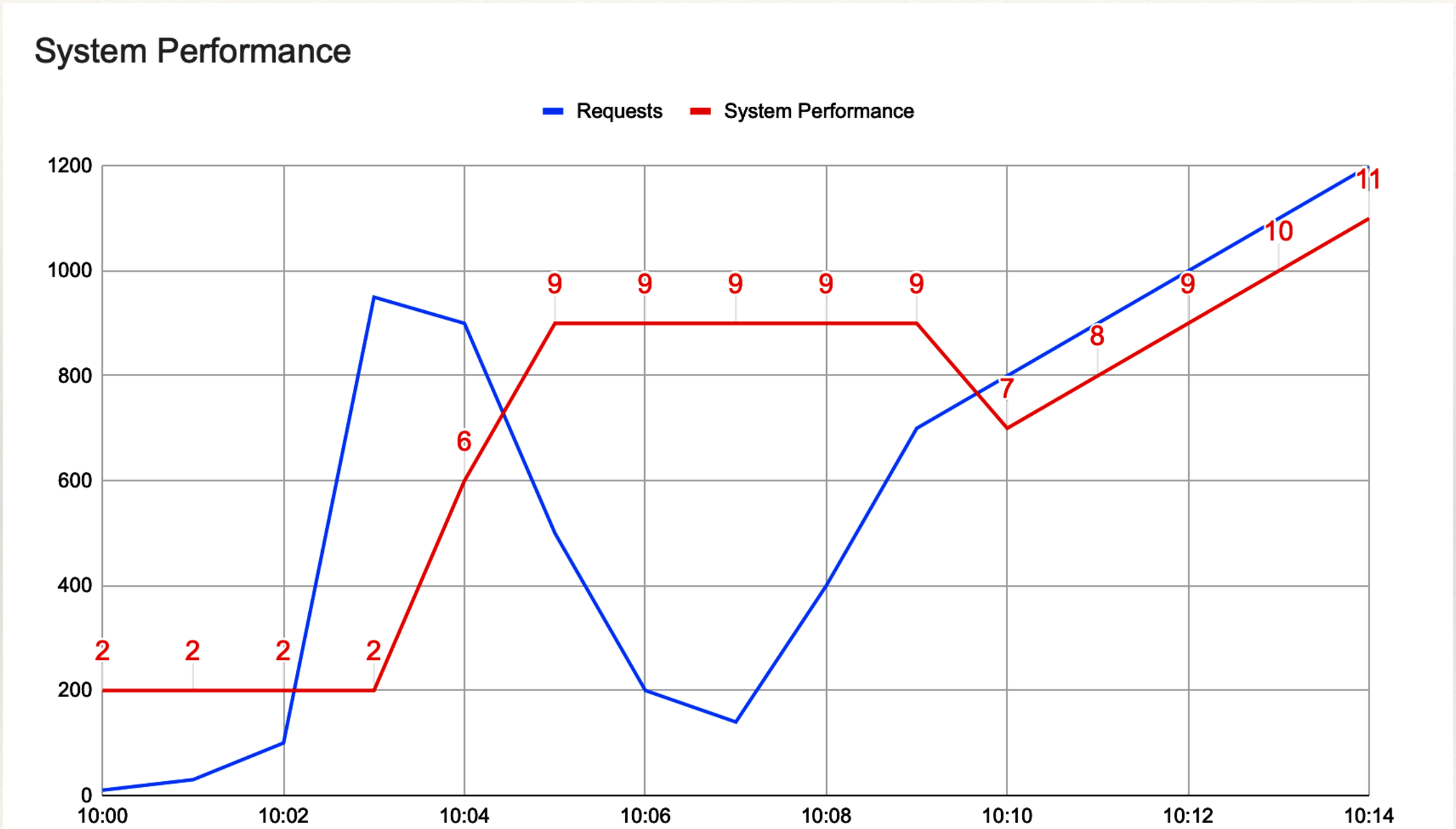
- ScaleUpLimitFactor
- ScaleUpLimitMinimum

How fast to scale pods

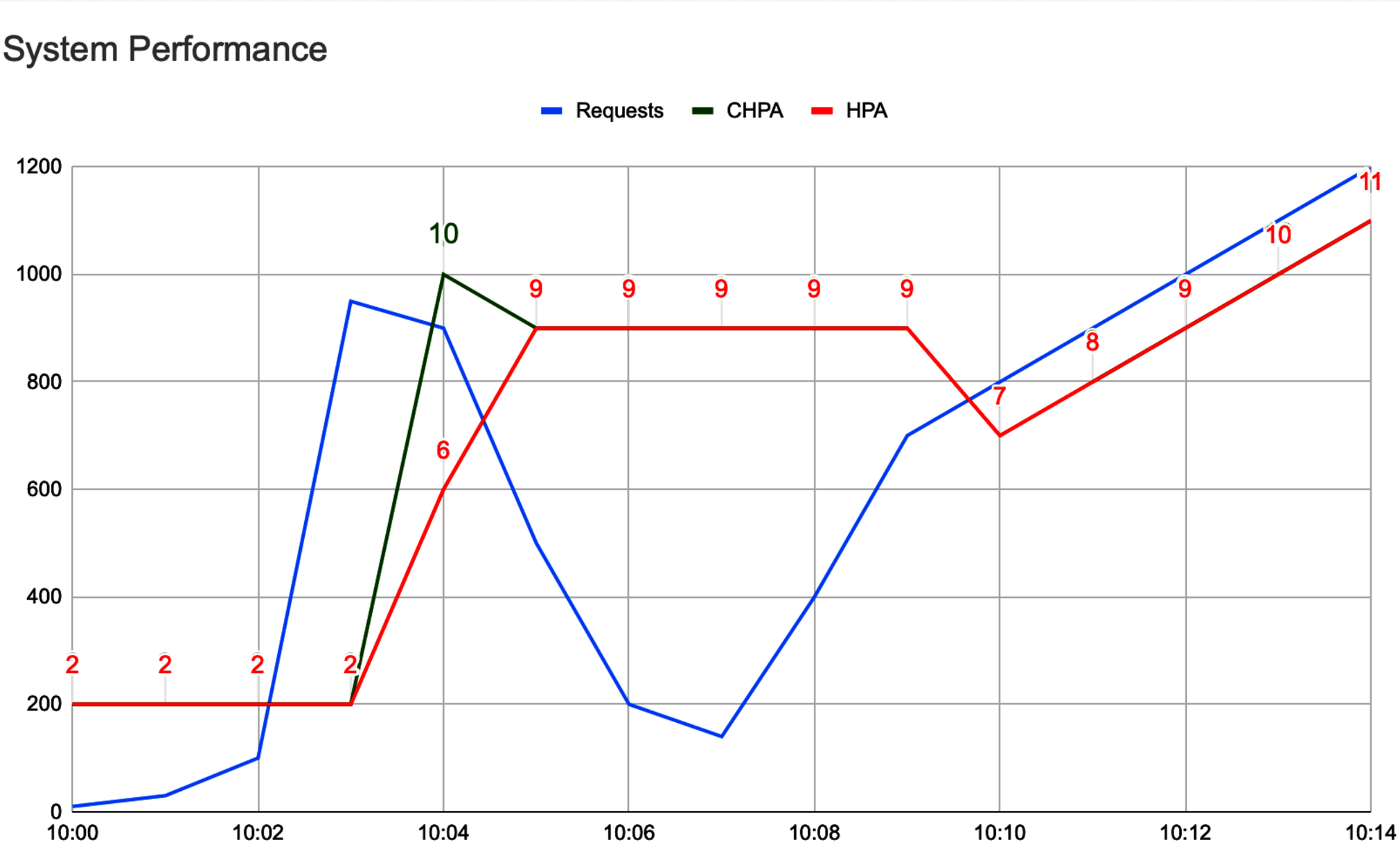
- UpscaleForbiddenWindowSeconds
- DownscaleForbiddenWindowSeconds.

How long to wait until the next scale

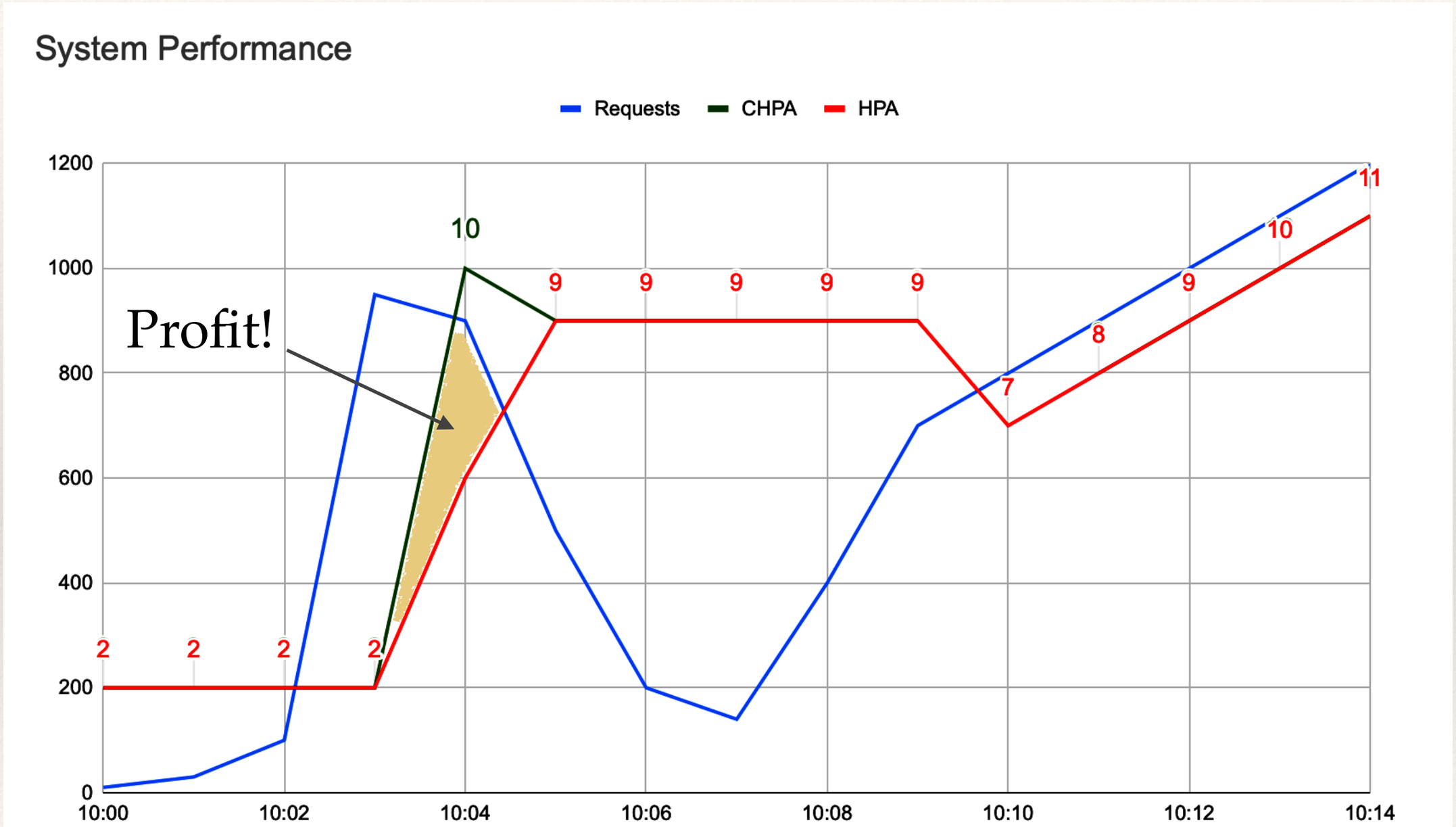
CHPA Results: Web Requests



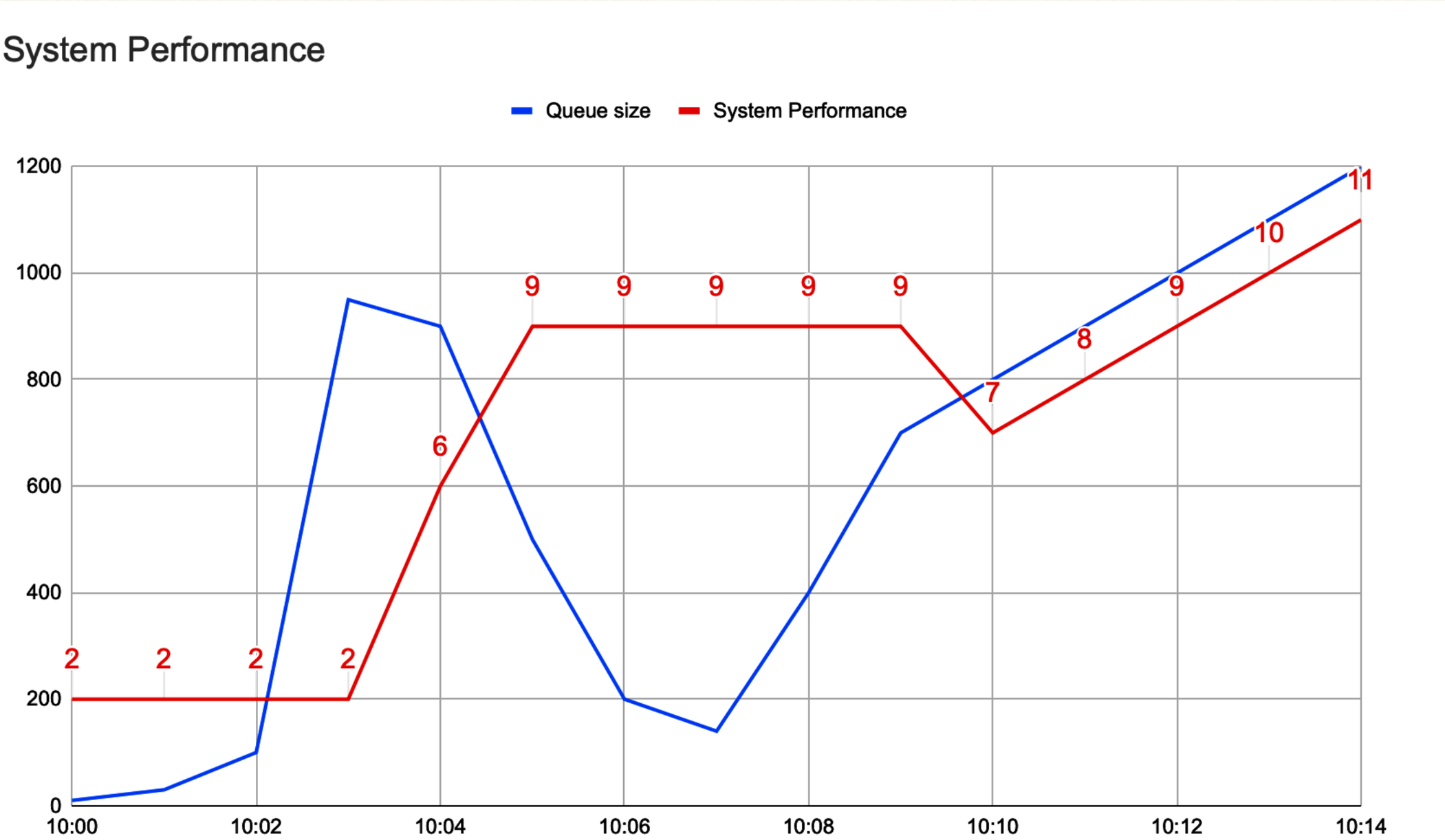
CHPA Results: Web Requests



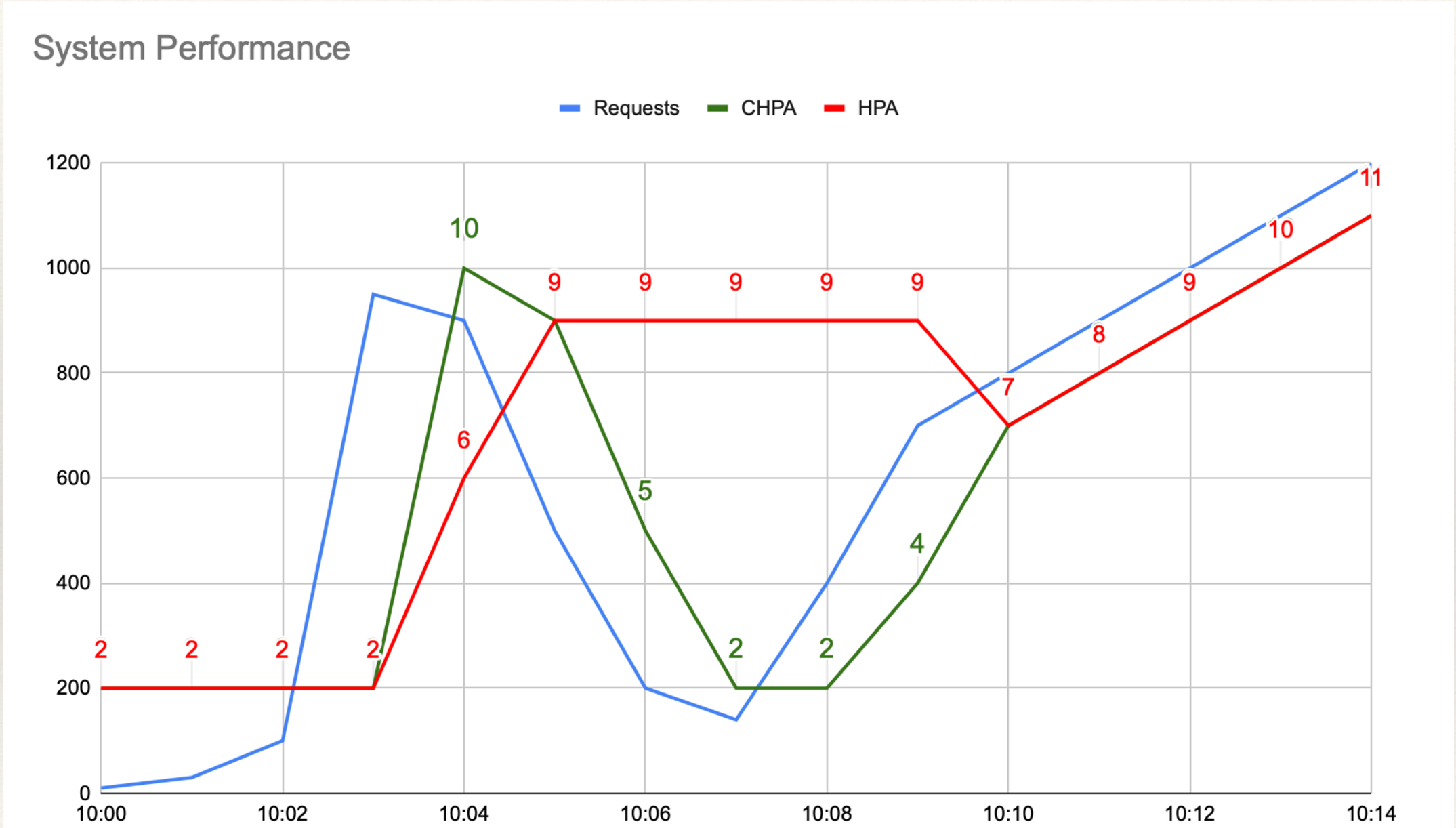
CHPA Results: Web Requests



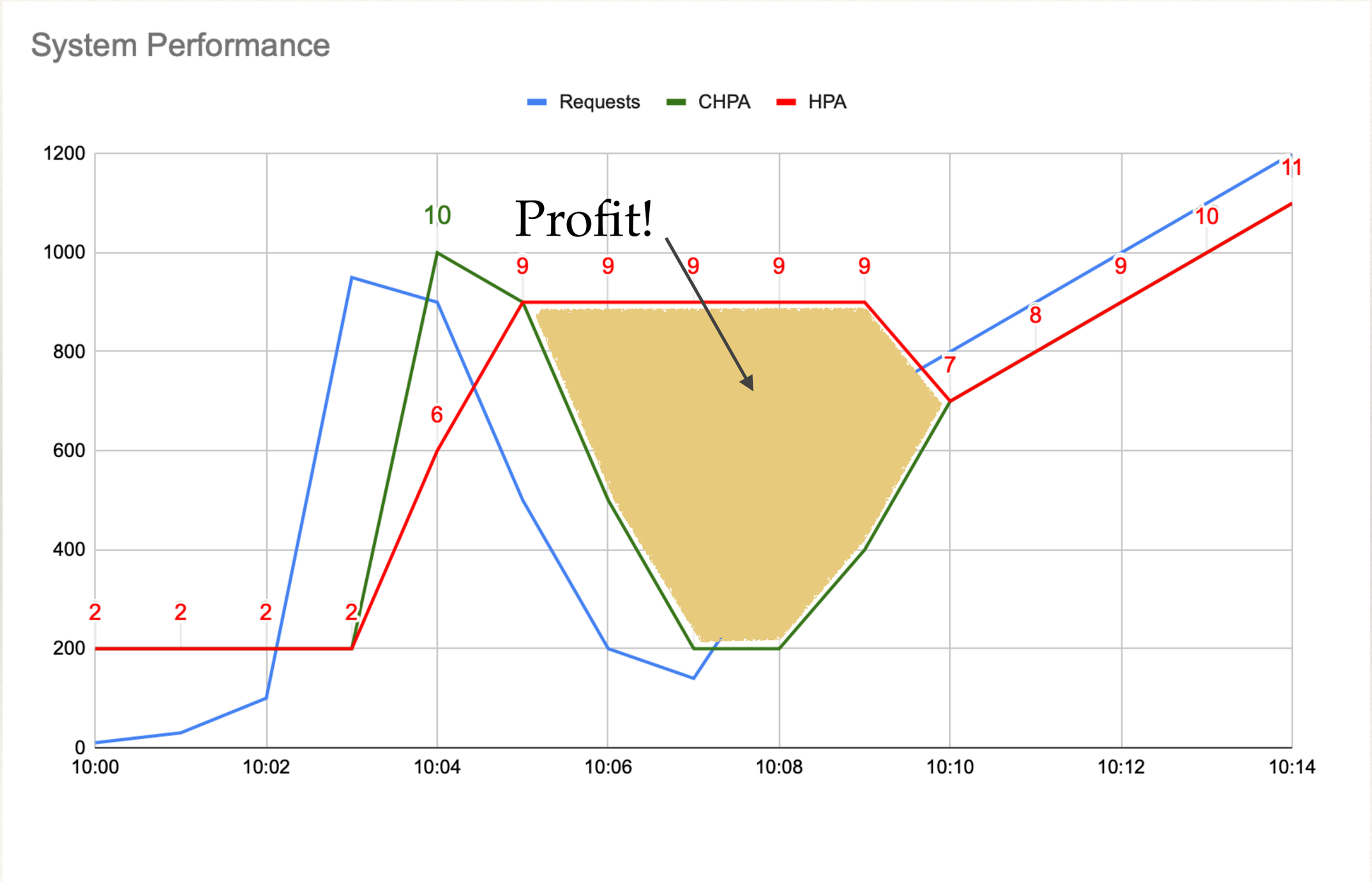
CHPA Results: Data Events



CHPA Results: Data Events



CHPA Results: Data Events



Watermark Podautoscaler

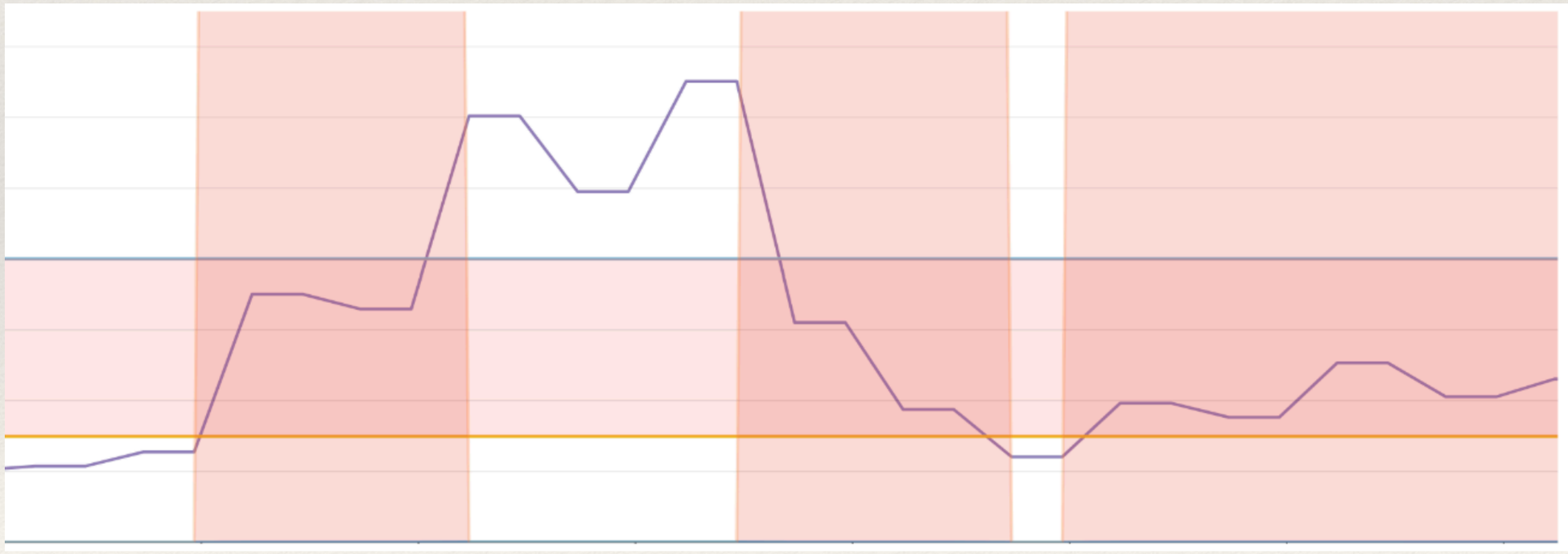
- The same as CHPA
- Tolerance -> Low & High Watermarks
  (Ratio) (CurrentMetricValue)

Ratio = currentMetricValue / desiredMetricValue

DesiredReplicas = ceil[currentReplicas * Ratio]

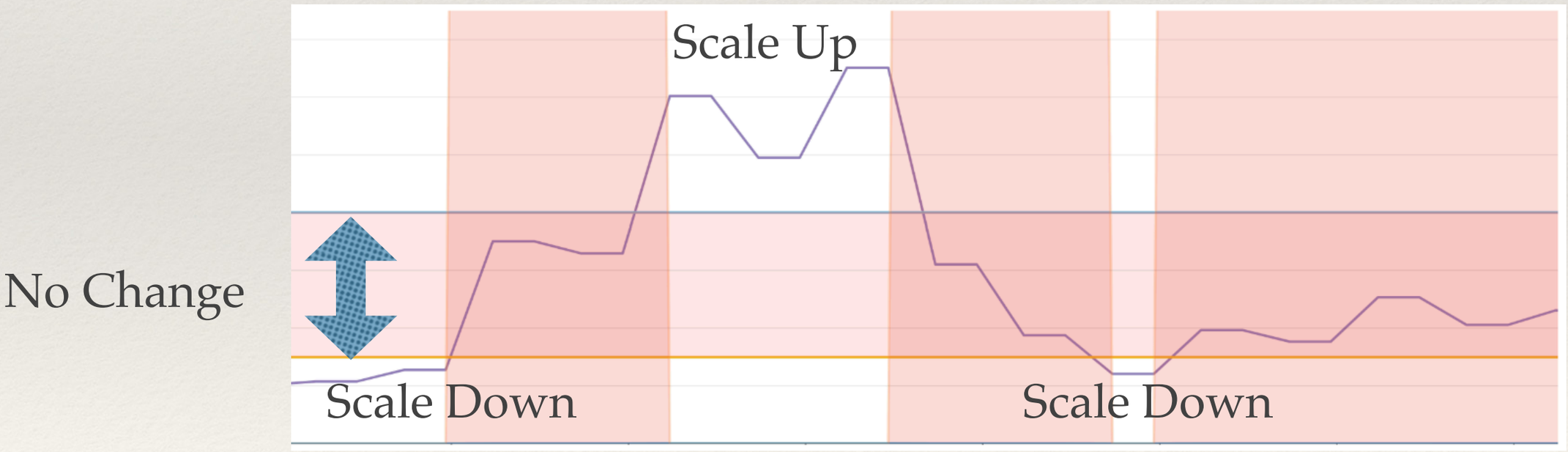
Watermark Podautoscaler

- The same as CHPA
- Tolerance -> Low & High Watermarks



Watermark Podautoscaler

- The same as CHPA
- Tolerance -> Low & High Watermarks



k8s Solution

k8s Solution

- RTFM
- KEP Approve
- API Change Approve
- PR Approve
- Documentation Approve

k8s Solution: RTFM

- https://github.com/kubernetes/community
  - contributors/devel/README.md
  - contributors/devel/development.md
  - contributors/guide/README.md
  - contributors/guide/pull-requests.md
  - contributors/guide/github-workflow.md
  - sig-list.md

k8s Solution: KEP Approve

- https://github.com/kubernetes/enhancements
- KEP = RFC ~ PEP ~ JEP …
- SIG
- https://github.com/kubernetes/enhancements/blob/master/keps/sig-autoscaling/20190307-configurable-scale-velocity-for-hpa.md

k8s Solution: API Change Approve

- Need to configure:
  - Scale velocity
    - Stabilization window
    - For both direction
- Reasonable defaults
- General and extendible API

k8s Solution: API Change Approve

- scaleUp:
  - pods: 4
  - percent: 100
  - periodSeconds: 60
  - stabilization: 0

k8s Solution: API Change Approve

- scaleUp:
  - pods: 4
  - percent: 100
  - periodSeconds: 60
  - stabilization: 0

Backward Compatibility ?

Default Values ?

Predictability ?

Extendibility ?

k8s Solution: API Change Approve

- scaleUp:
  - pods: 5
  - percent: 50
  - periodSeconds: 120
  - stabilization: 60

v1 autoscaling: since Apr 2016

v2beta autoscaling: since Aug 2017

Backward Compatibility ?

- scaleUp:
  - pods: 4
  - percent: 100
  - periodSeconds: 60
  - stabilization: 0

k8s Solution: API Change Approve

- scaleUp:
  - pods: 4
  - percent: 100
  - periodSeconds: 60
  - stabilization: 0

- scaleUp:
  - pods: 4

"percent" ??

"stabilization" ??

Default Values ?

- scaleUp:
  - periodSeconds: 30

"percent" ??

"pods" ??

k8s Solution: API Change Approve

- scaleUp:
  - pods: 4
  - percent: 100
  - periodSeconds: 60
  - stabilization: 0

- scaleUp:
  - pods: 4
  - percent: 100

What is used???
max change?
max value?

Predictability ?

- scaleDown:
  - pods: 4
  - percent: 100

What is used???
max change?
max value?

k8s Solution: API Change Approve

- scaleUp:
  - pods: 4
  - percent: 100
  - periodSeconds: 60
  - stabilization: 0

Another limiting factor ?

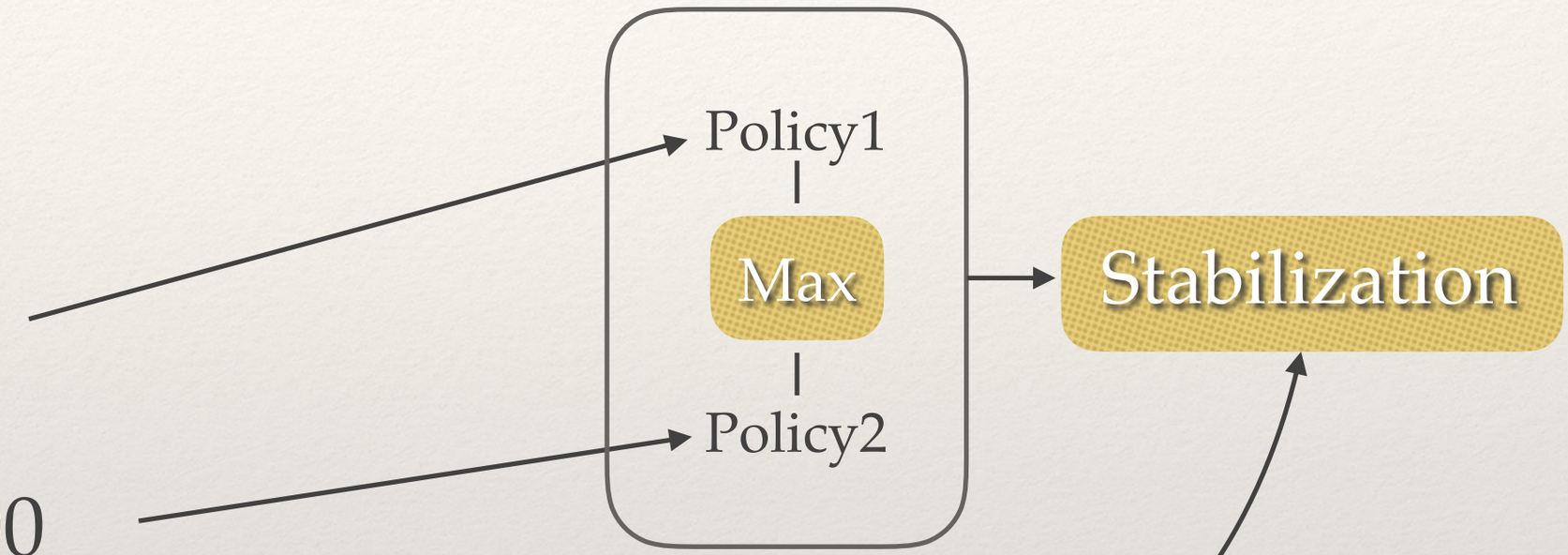
MovingAverage?

Another abstraction layer ?

Extendibility ?

k8s Solution: API Change Approve

- scaleUp:
  - pods: 4
  - percent: 100
  - periodSeconds: 60
  - stabilization: 0



k8s Solution: API Change Approve

- scaleUp:
  - pods: 4
  - percent: 100
  - periodSeconds: 60
  - stabilization: 0

- scaleUp:
  - StabilizationWindowSeconds: 0
  - SelectPolicy: Max
  - Policies:
    - type: Pods
      value: 4
      periodSeconds: 60
    - type: Percent
      value: 100
      periodSeconds:60

k8s Solution: API Change Approve

- Versioning
  - v1 must work
  - v2beta1, v2beta2 must work
  - Event-based system => any object should be convertible to/from v1 version
- Code Generation
- Difference: int32 vs *int32

K8s Solution: PR Approve

- Functionality
  - New control flow in case of large change
- Tests
  - Flaky tests (the majority of all the issues)
  - E2e tests
- Graduation Criteria

k8s Solution: Best Practices

- Prepare for spam on your GitHub account email
- Start with two separate PRs:
  - API change
  - Functionality change
- Do not use docker for build & test (16GB is not enough)
- API Change PR is hard to approve

HPA is simple?

- Difficult to test issues with concurrency. Majority of issues for sig:autoscaling are about flaky tests.
- Do not scale based on queue size.
- Can’t get metrics for CPU usage -> HPA will not work
https://github.com/kubernetes/kubernetes/issues/30471
- Rolling Upgrade with Surge => plenty of new pods
https://github.com/kubernetes/kubernetes/issues/72775
- Allow HPA to scale to 0 pods
https://github.com/kubernetes/kubernetes/issues/69687

1. Contacts. @gliush https://devzen.ru
2. “Configurable HPA by Postmates”. https://blog.postmates.com/configurable-horizontal-pod-autoscaler-81f48779abfc
3. “Pets vs Cattles”. https://medium.com/@Joachim8675309/devops-concepts-pets-vs-cattle-2380b5aab313
4. “Delay to start new pod”. https://dzone.com/articles/kubernetes-autoscaling-101-cluster-autoscaler-hori-1
5. “Datadog watermarks autoscaler” https://github.com/DataDog/watermarkpodautoscaler
6. Configurable HPA KEP. https://github.com/kubernetes/enhancements/blob/master/keps/sig-autoscaling/20190307-configurable-scale-velocity-for-hpa.md
7. Kubernetes repository https://github.com/kubernetes/community/
8. Kubernetes Community Content https://github.com/kubernetes/community/
9. Configurable HPA: API Change PR: https://github.com/kubernetes/kubernetes/pull/74525
10. Configurable HPA: Functionality PR: https://github.com/gliush/kubernetes/pull/2