AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

CMP325-R2

Scaling your compute proactively with predictive scaling

Scott Horsfield
WW SA Leader, EC2 Scalability
Amazon Web Services

Ankur Sethi
Sr. Product Manager, EC2
Amazon Web Services

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

A quick introduction to predictive scaling

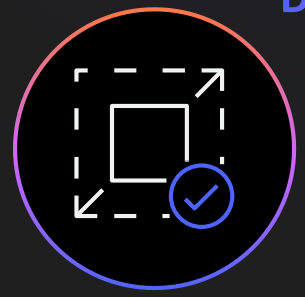
A demonstration of predictive scaling in action

Open discussion of common use cases

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling policies explained

Dynamic scaling



Simple/step scaling

Monitors metrics and adds/removes instances according to steps defined by the customers

- Manually calculate capacity
- Reactive in nature



Target tracking

Thermostat-like control mechanism that automatically adds or removes instances to maintain metrics at a customer-defined target

- Automated
- Reactive in nature

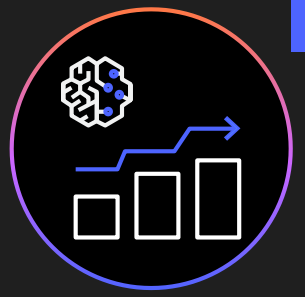


Scheduled scaling

Launch/terminate instances as defined by customer on a schedule

- Manually calculate capacity
- Proactive in nature

NEW



Predictive scaling

Proactively launch capacity based on historic trends

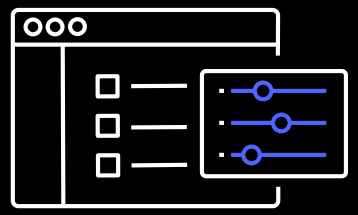
- Automated
- Proactive in nature

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Introducing predictive scaling

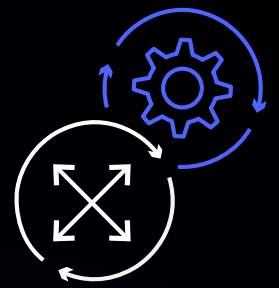
A new Amazon EC2 Auto Scaling policy that proactively launches EC2 instances in anticipation of predicted demand



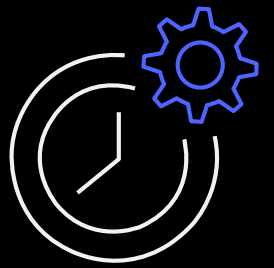
Proactively manages capacity



Reduces overprovisioning



Works within defined boundaries



Adjusts automatically

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Common use cases



Recurring load spikes

Scales proactively instead of reactively ahead of recurring spikes in load



Alternative to manual scaling

Learns and adjusts over time so you don’t need to adjust manual or scheduled scaling processes

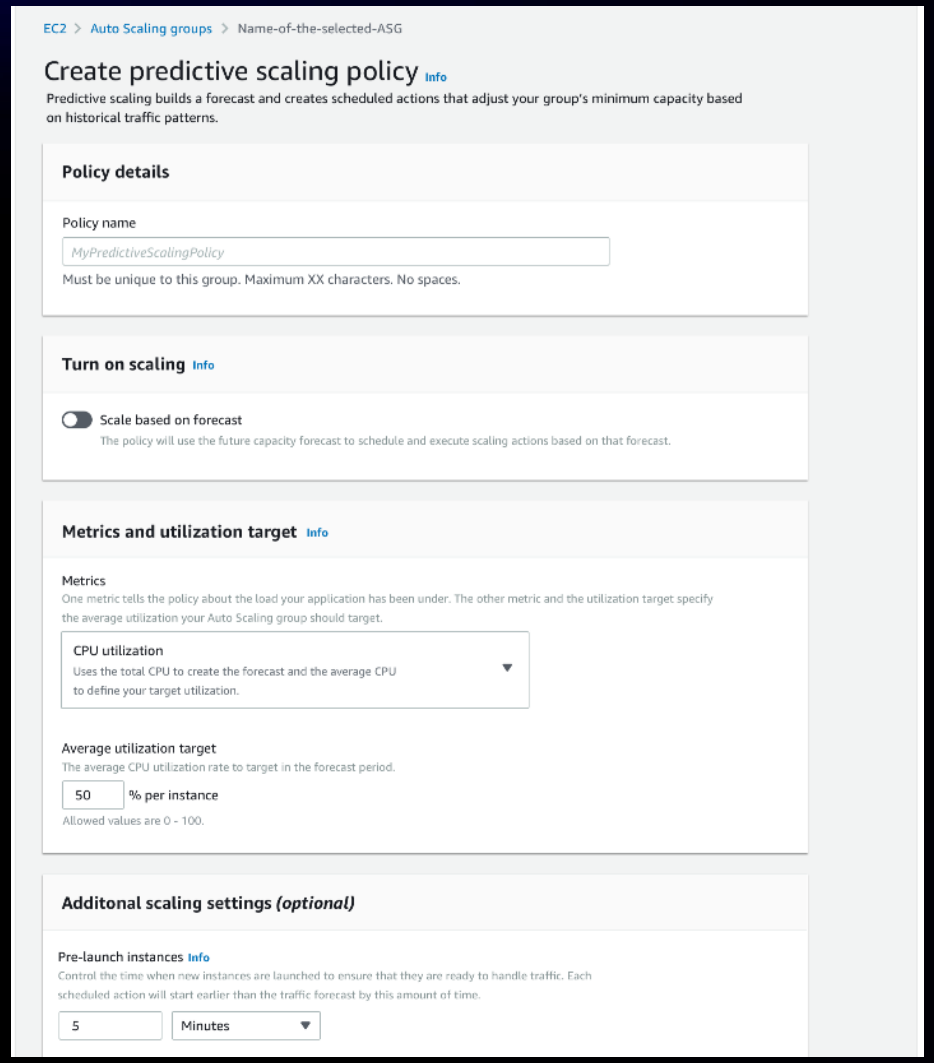


Long initialization time

Can launch capacity ahead of time, allowing instances to complete initialization activities

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Create a policy



Choose metrics

Supports 4 predefined metrics: CPU utilization, network in and out, ALB request count per target group. Also supports custom metrics.

Set target metric values

For example, 50% CPU utilization average per instance

Configure additional settings

Pre-launch for applications with long initialization time; buffer max capacity to handle spikes above defined maximum

Validate before scaling

Validate in forecast only mode, switch to forecast and scale when ready for capacity to be automatically launched

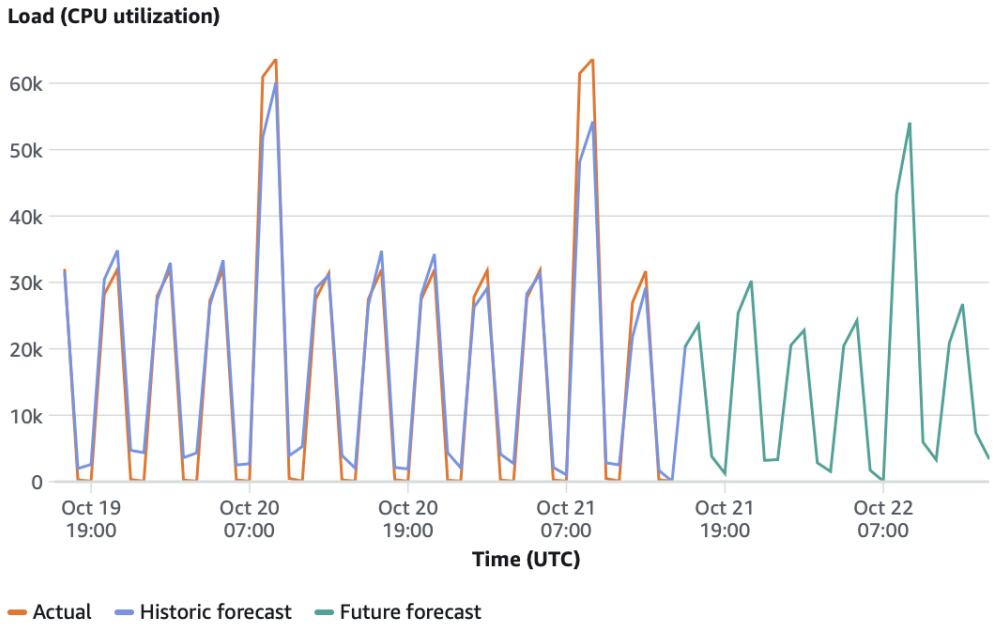
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

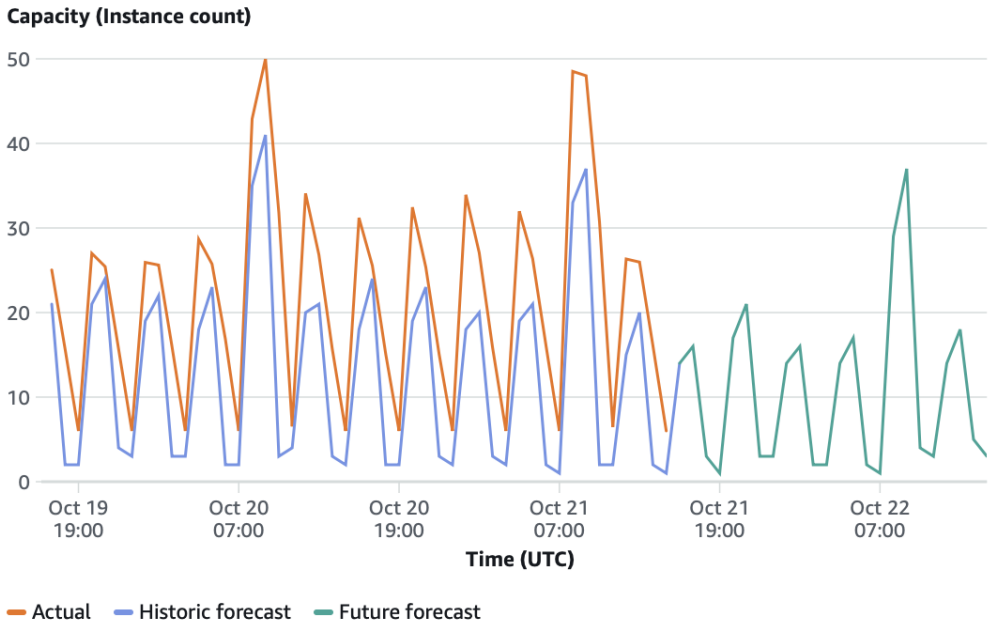
Review policy

PredictiveScalingCPU

Scaling	Metric pair	Scaling metric	Pre-launch instances
Scale based on forecast	CPU utilization	-	5 minutes
	Load metric	**Target utilization**	**Override maximum capacity**
	-	25	No

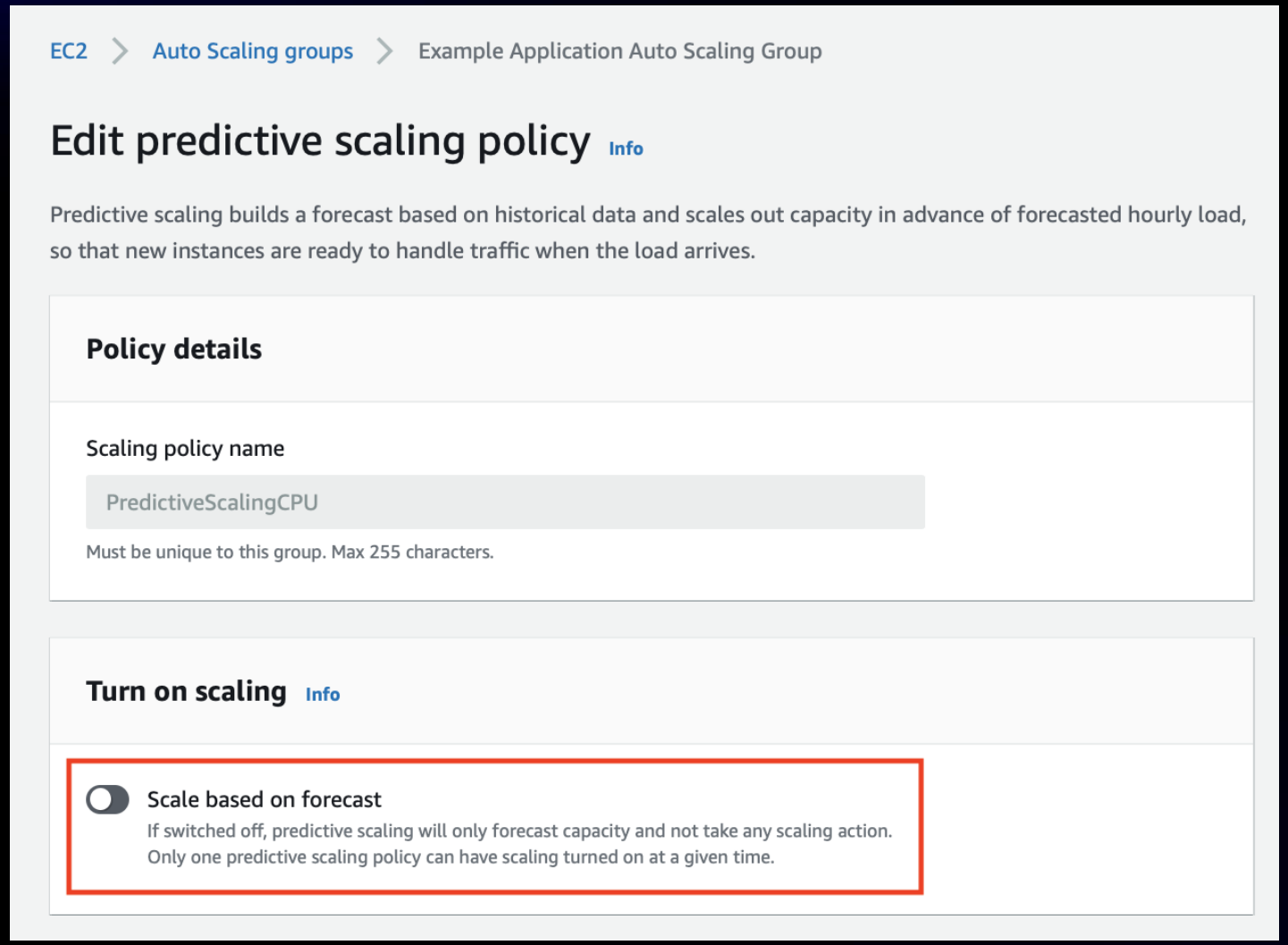
2d | 1w | 2w | 4w | 6w | 8w





© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Enable scaling



Turn on scaling

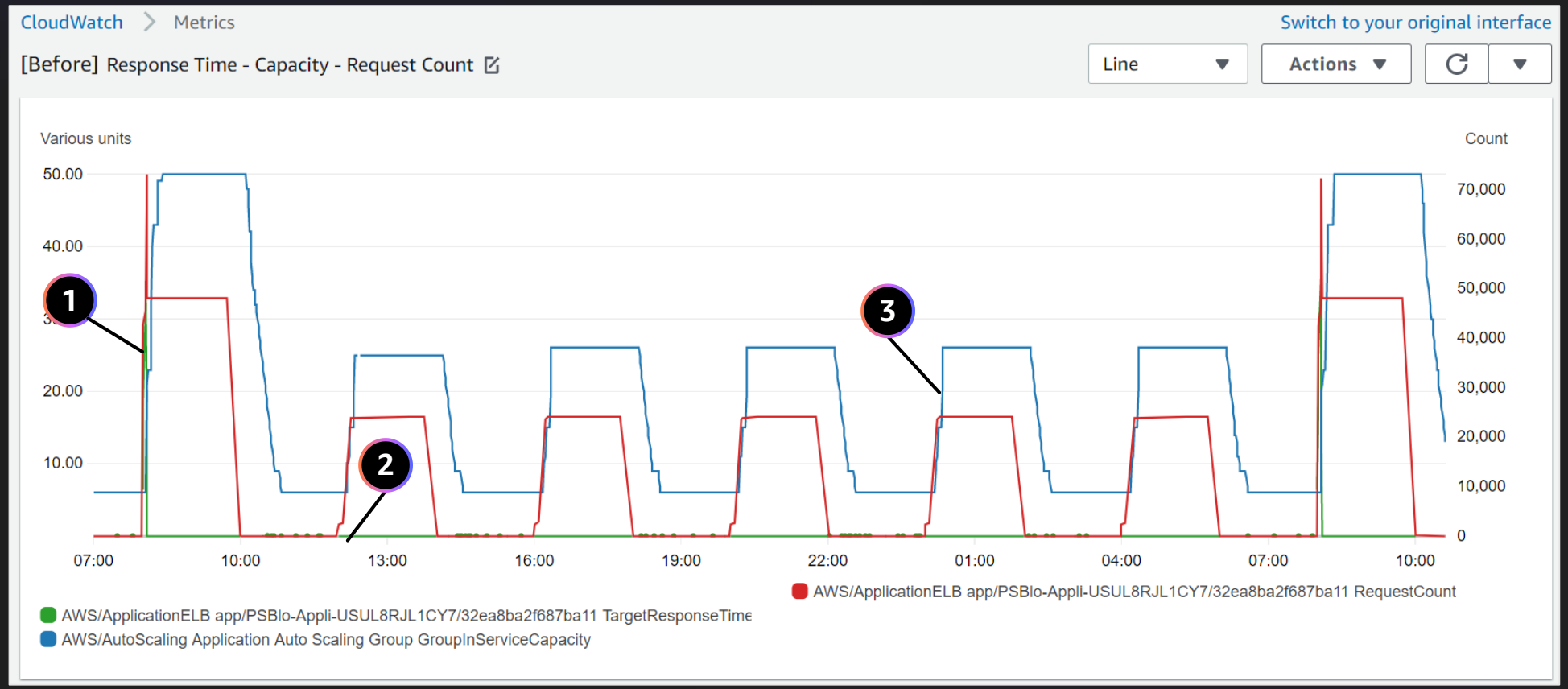
Edit your predictive scaling policy, switch the toggle, and your predictive scaling policy will now manage capacity proactively

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Predictive scaling in action

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Huge spike in response time when demand changes rapidly

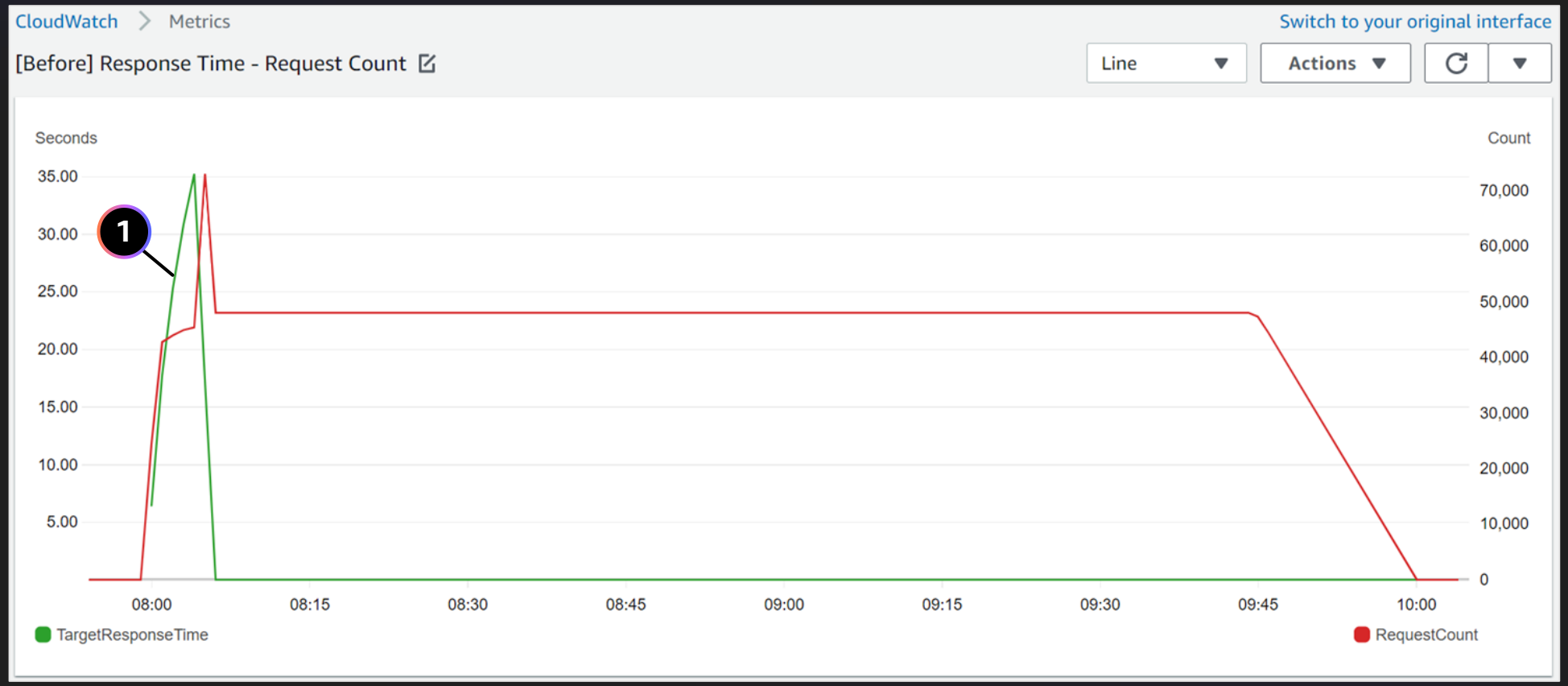


1. Response time spikes when demand changes rapidly
2. No such spike observed for gradual demand change
3. Dynamic scaling adds capacity after the request count starts to increase; it works well for gradual spikes but not for such steep spikes

- ALB response time
- ALB request count
- Number of instances

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Huge spike in response time when demand changes rapidly



1 Response time spikes to 35 seconds while capacity is getting added by dynamic scaling

ALB response time

ALB request count

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Reviewing predictive scaling forecasts



1. Actual versus predicted load; load is represented as the sum of CPU utilization of each instance in the ASG

2. Actual versus predicted capacity for each hour; if switched to forecast and scale, the predicted capacity will be launched at the beginning of the hour and act as the baseline

- Actual
- Historic forecasts
- Future forecasts

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

No latency spikes after applying predictive scaling

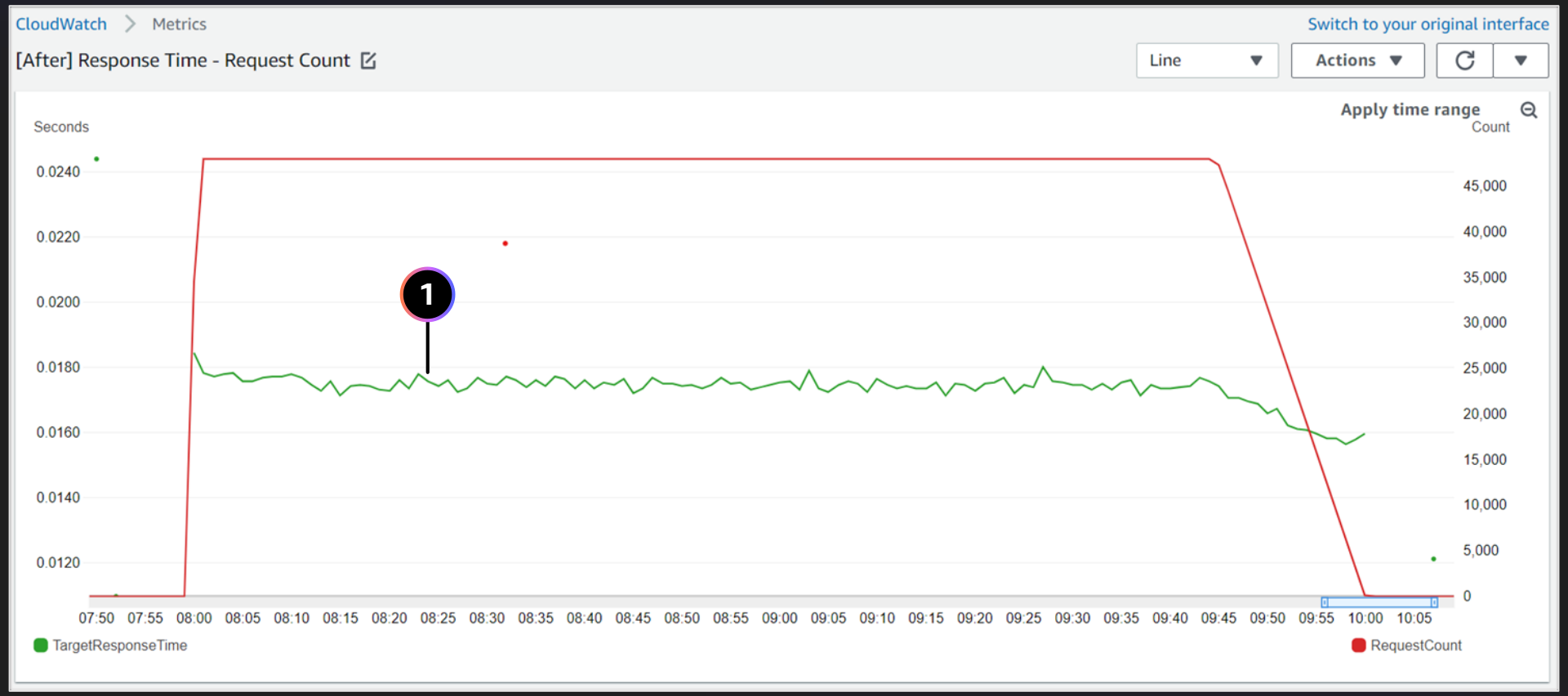


1. No more spikes in response time metric
2. Predictive scaling adds instances at the beginning of each hour
3. Dynamic scaling further adds capacity according to requirement

- ALB response time
- ALB request count
- Number of instances

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Applying predictive scaling in forecast and scale mode



1 Response time stays below 0.02 seconds; without predictive scaling, we observed response time spiking to significantly higher values while demand ramps up

ALB response time

ALB request count

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Predictive scaling resources

Blogs

https://aws.amazon.com/blogs/compute/introducing-native-support-for-predictive-scaling-with-amazon-ec2-auto-scaling/

https://aws.amazon.com/blogs/compute/retaining-metrics-across-blue-green-deployment-for-predictive-scaling/

Documentation

https://docs.aws.amazon.com/autoscaling/ec2/userguide/ec2-auto-scaling-predictive-scaling.html

Sample code

https://github.com/aws-samples/amazon-ec2-auto-scaling-group-examples

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Ankur Sethi
sethiank@amazon.com

Scott Horsfield
shhorsfi@amazon.com

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.