
AWS
re:Invent

COM401

Scaling containers on AWS

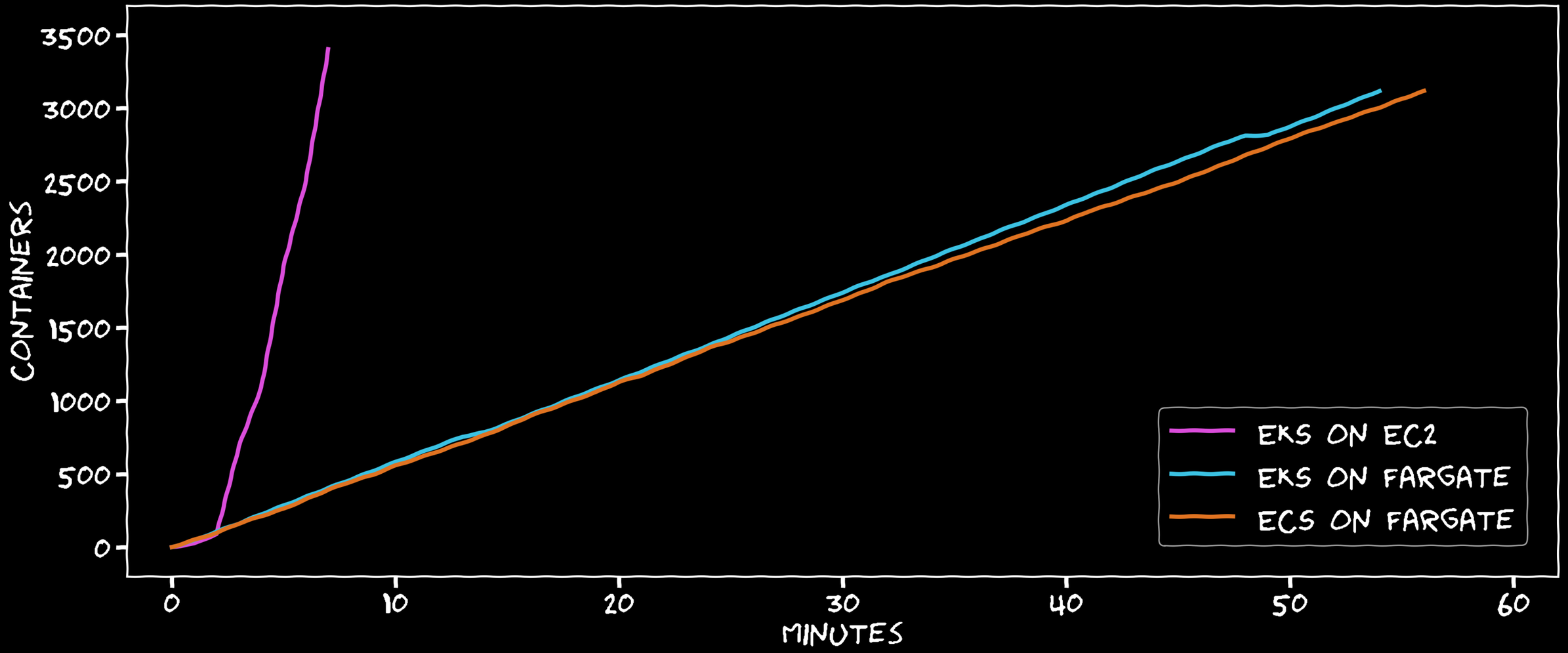
Vlad Ionescu
AWS Container Hero
Independent Consultant

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

“Scaling containers on AWS” – April 2020

https://www.vladionescu.me/posts/scaling-containers-in-aws.html



Agenda

Scaling test results – winter 2020 edition

Ideal applications for each option

Cost optimizations



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Preparation



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Preparation

- Attendees who
  - Are deeply familiar with the topic
  - Have implemented a solution on their own already
  - Are comfortable with how the technology works across multiple services, architectures, and implementations

Preparation

- Attendees who
  - Are deeply familiar with the topic
  - Have implemented a solution on their own already
  - Are comfortable with how the technology works across multiple services, architectures, and implementations
- Attendees who
  - Want to explore future developments in the technology

Preparation

T

- Application

Preparation

T

- Hello-world-style app

Preparation

T

- Hello-world-style app
  - One container

Preparation

T

- Hello-world-style app
  - One container
    - 1 vCPU
    - 2 GB

Preparation

T

- Hello-world-style app
- Spot Instances used where possible

Preparation

T

- Hello-world-style app
- Spot Instances used where possible
- Limits increased for some services

Preparation

T

- Hello-world-style app
- Spot Instances used where possible
- Limits increased for some services
- Tests performed in the U.S. East (northern Virginia)

Preparation

T

- Hello-world-style app
- Spot Instances used where possible
- Limits increased for some services
- Tests performed in the U.S. East (Northern Virginia)
- Bulk, forced scaling to 3,500 containers

Preparation

T

- Hello-world-style app
- Spot Instances used where possible
- Limits increased for some services
- Tests performed in the U.S. East (Northern Virginia)
- Bulk, forced scaling to 3,500 containers

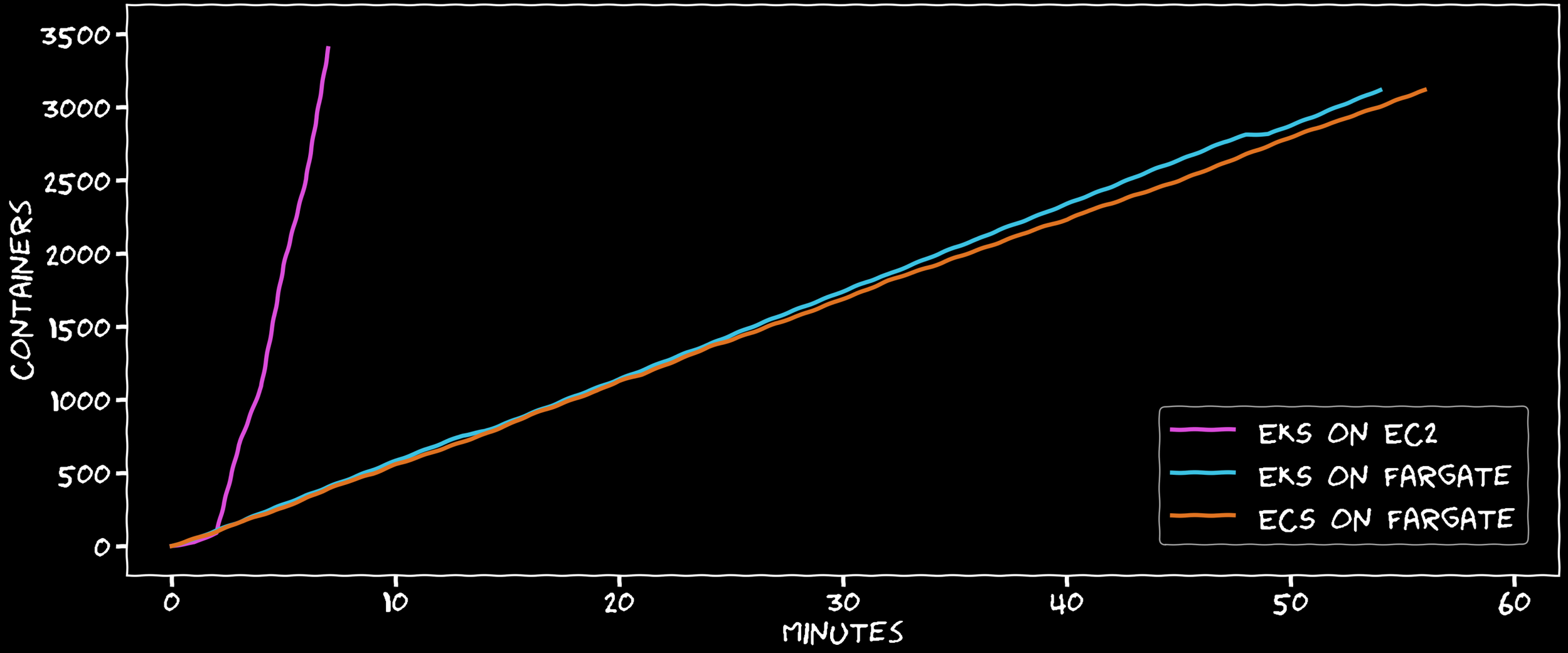
- *Extreme* test case

Results

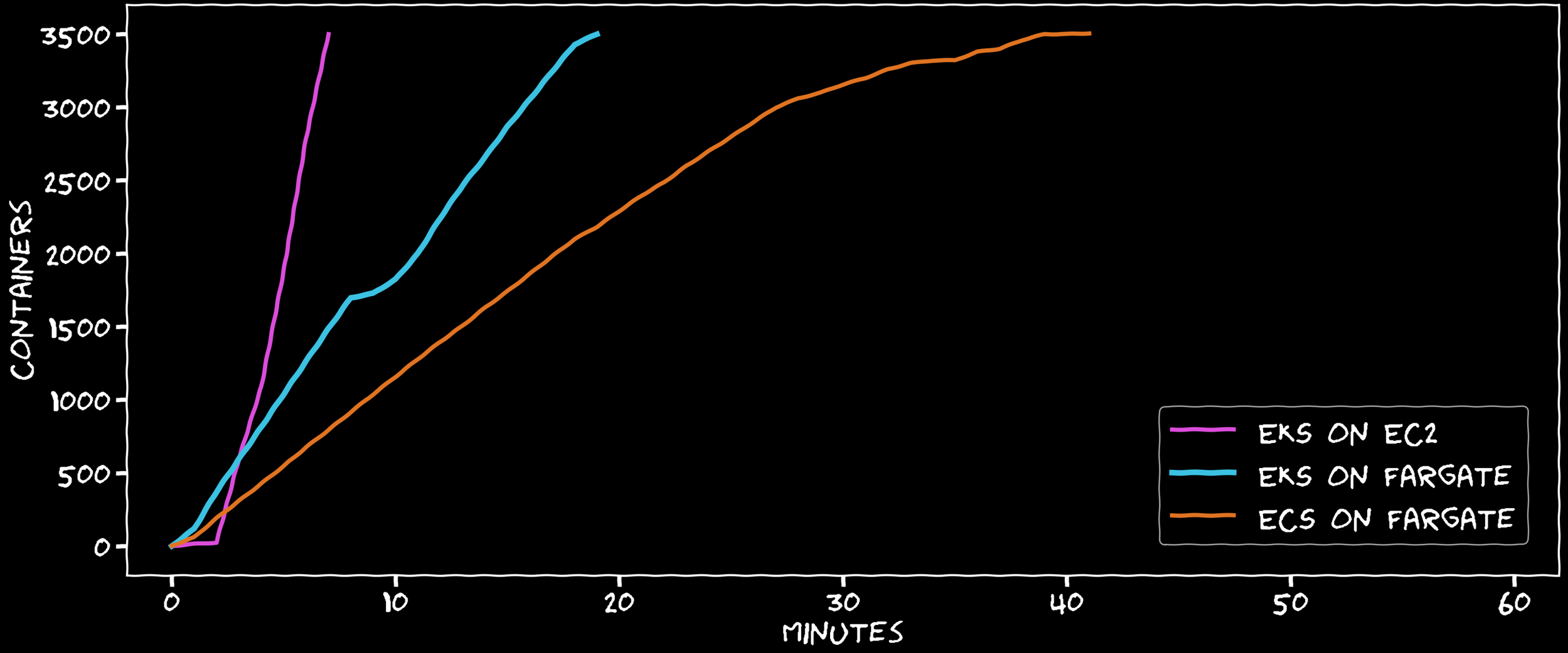


© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

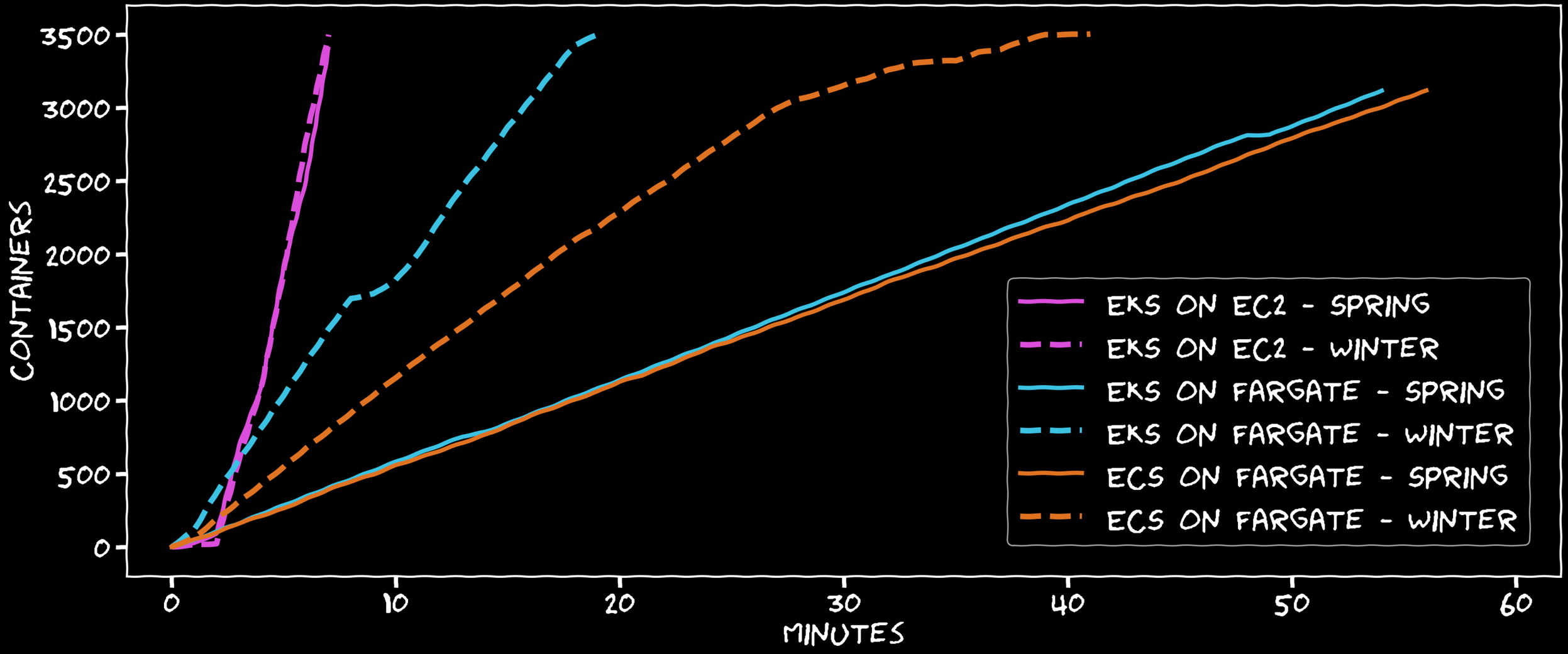
Scaling test results – spring 2020



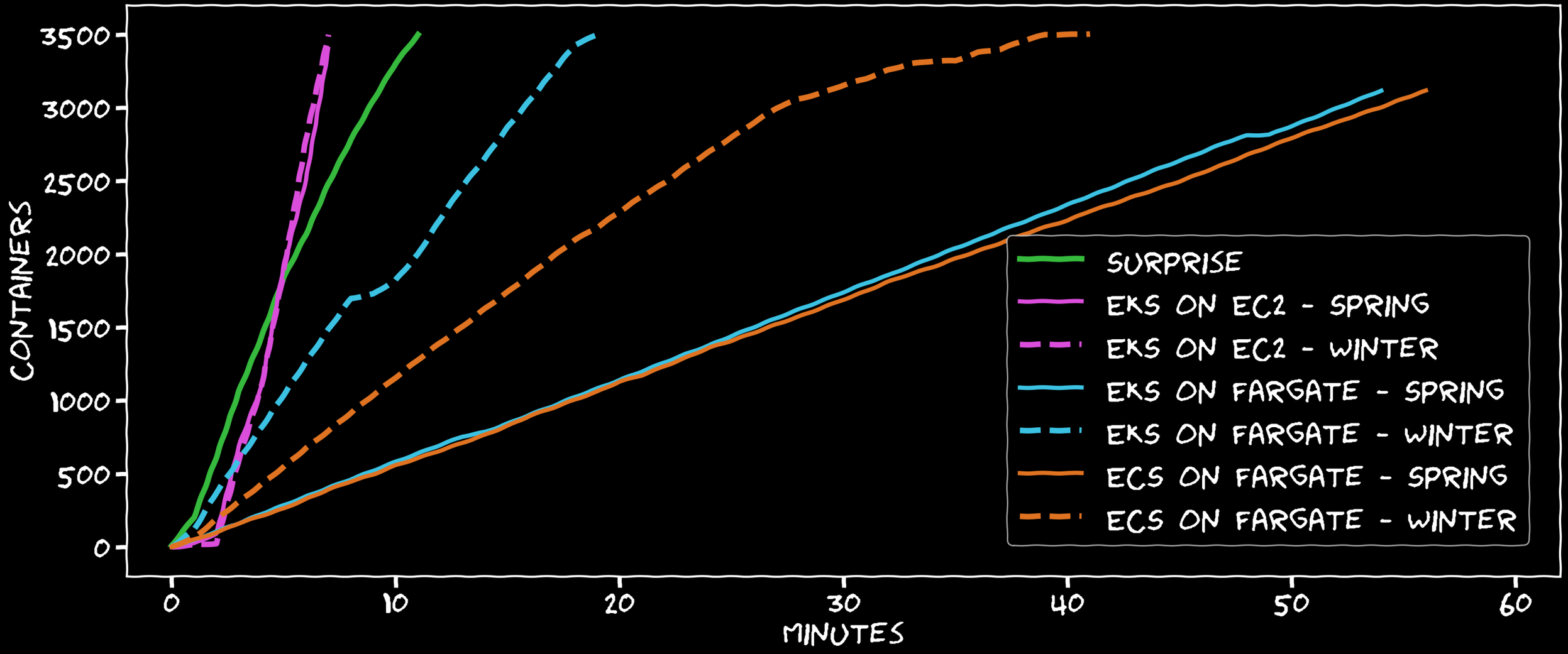
Scaling test results – winter 2020



Scaling test results



Scaling test results



Results deep dive



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 workers



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

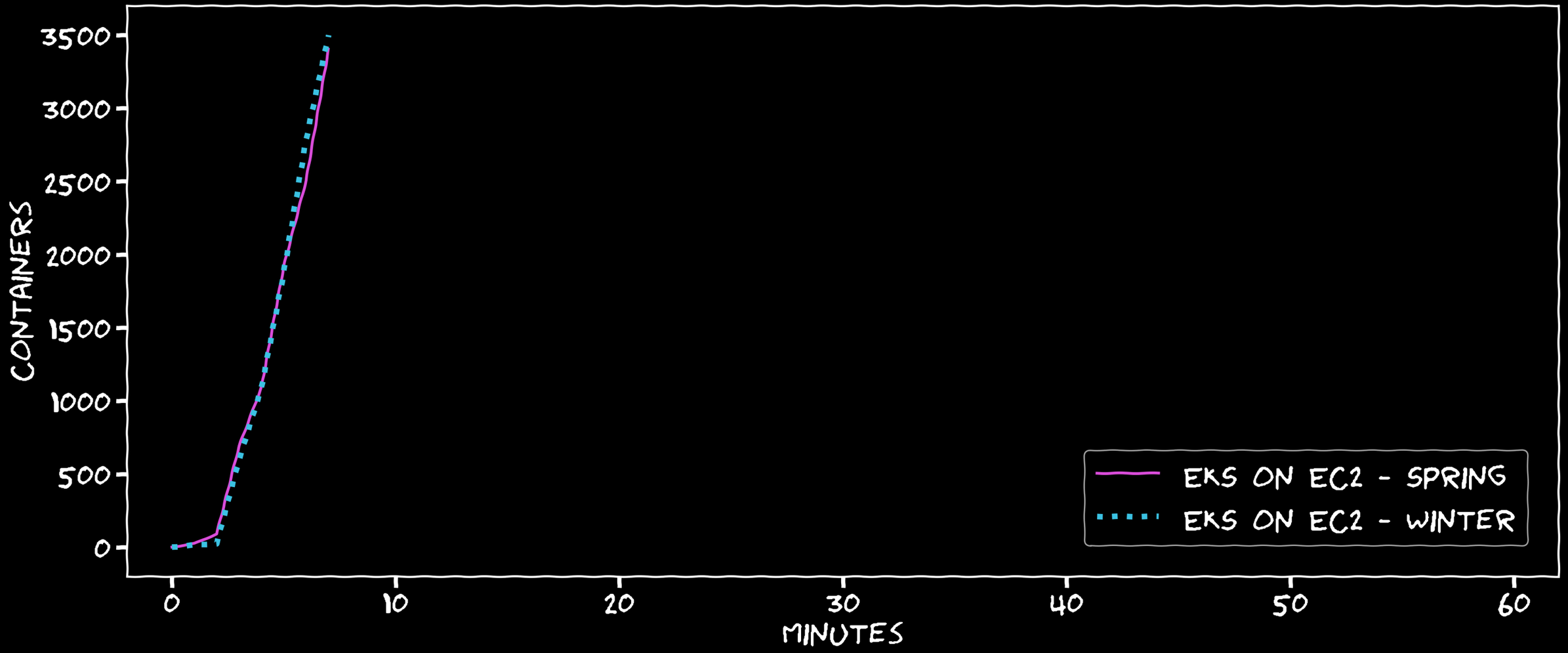
Amazon EKS on Amazon EC2



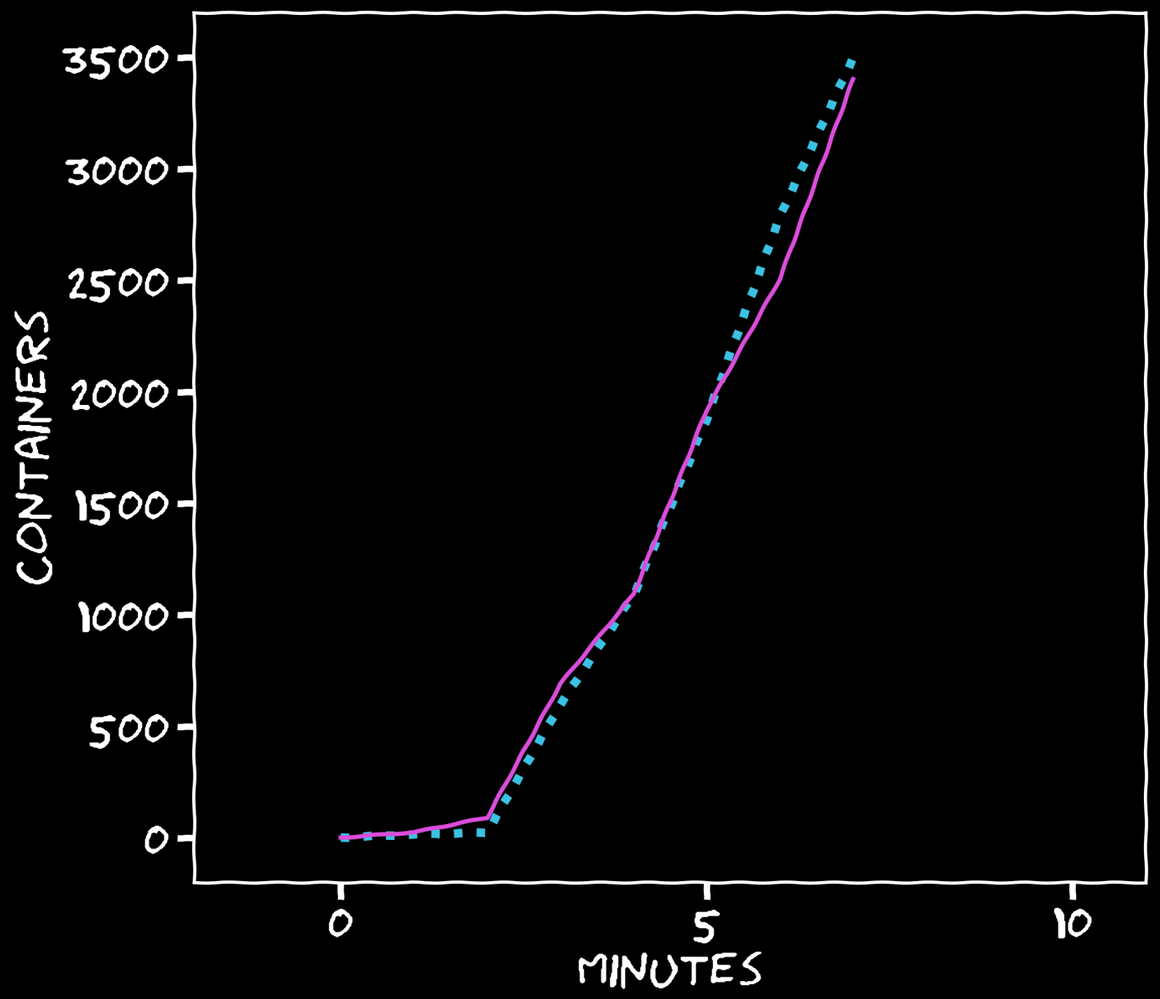
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

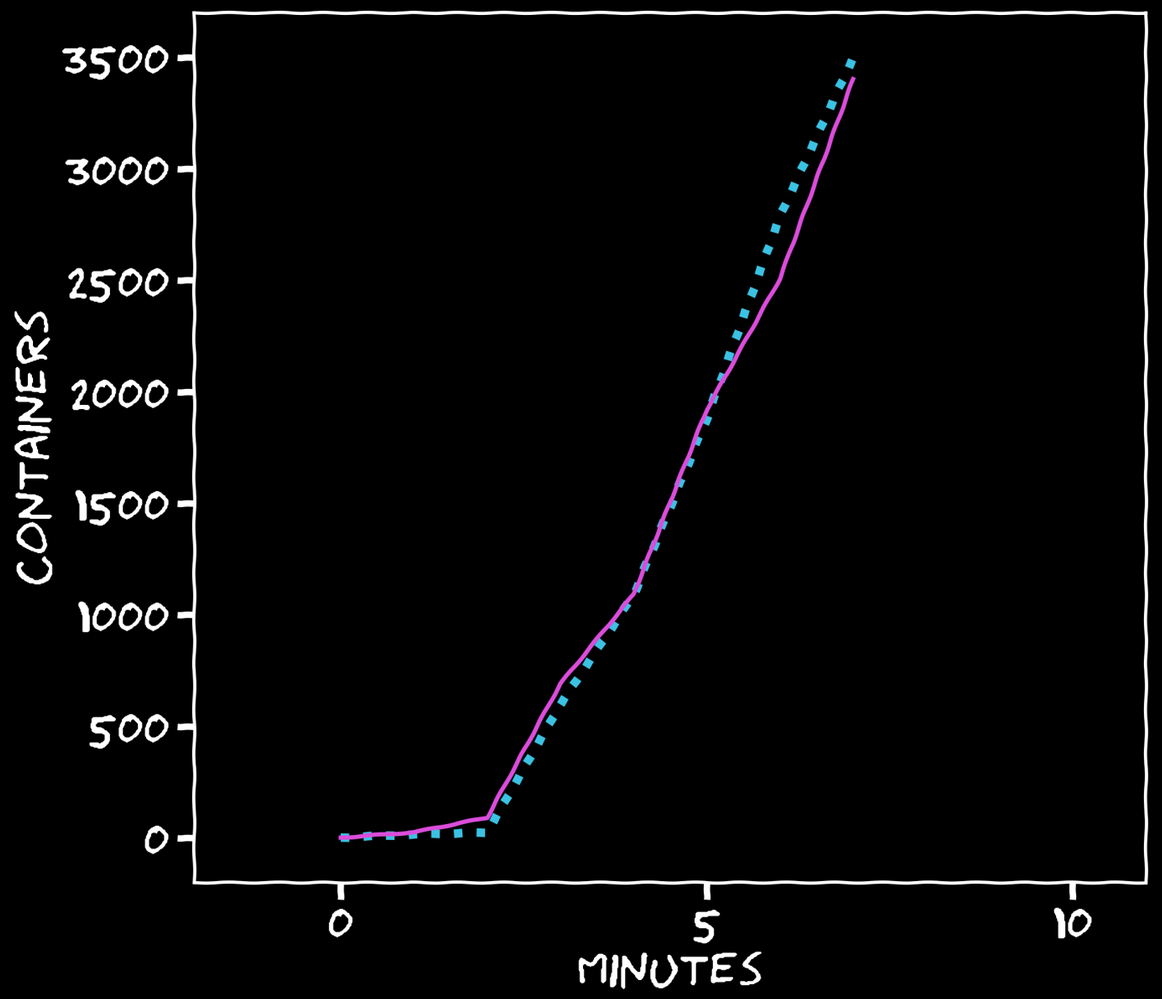
Amazon EKS on Amazon EC2



Amazon EKS on Amazon EC2

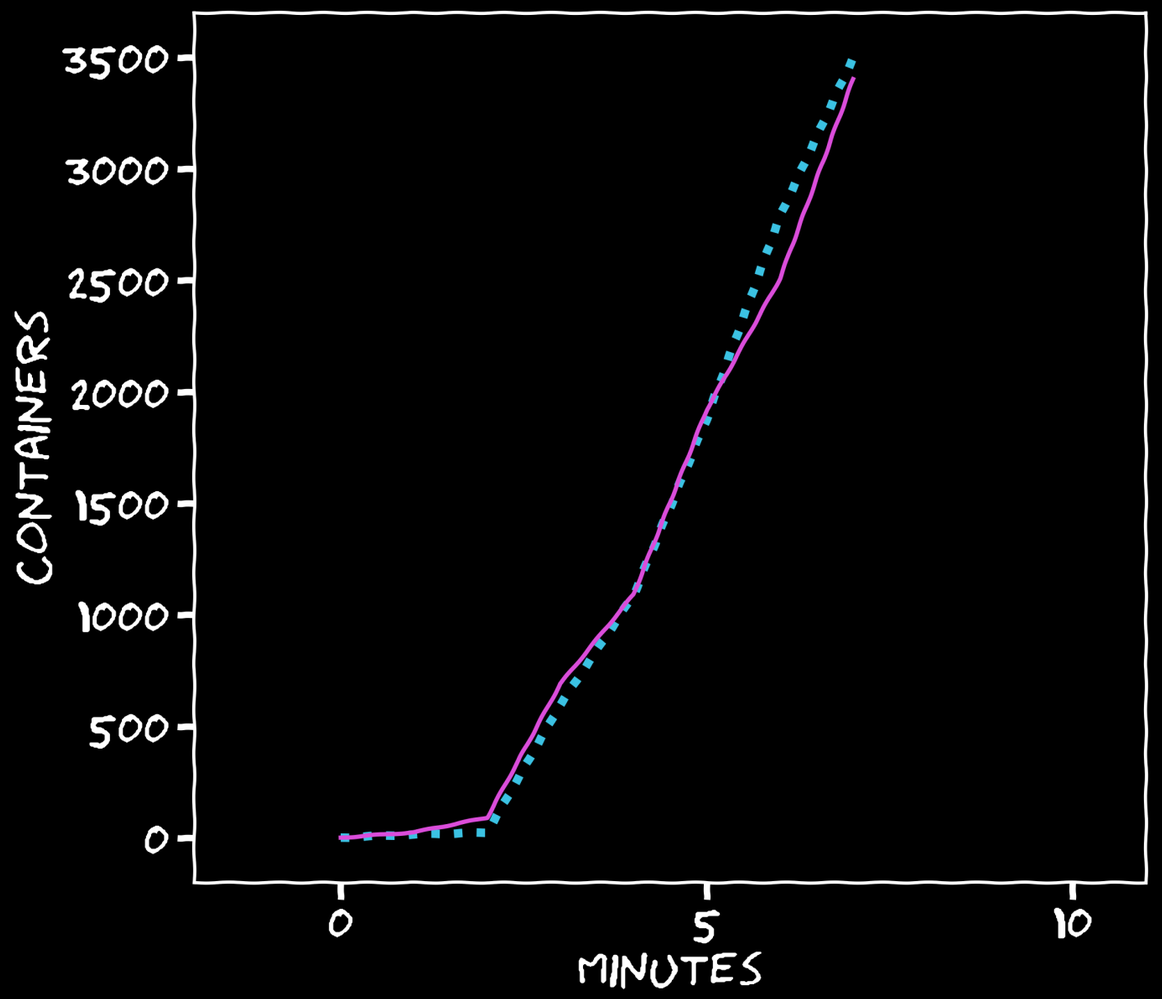


Amazon EKS on Amazon EC2



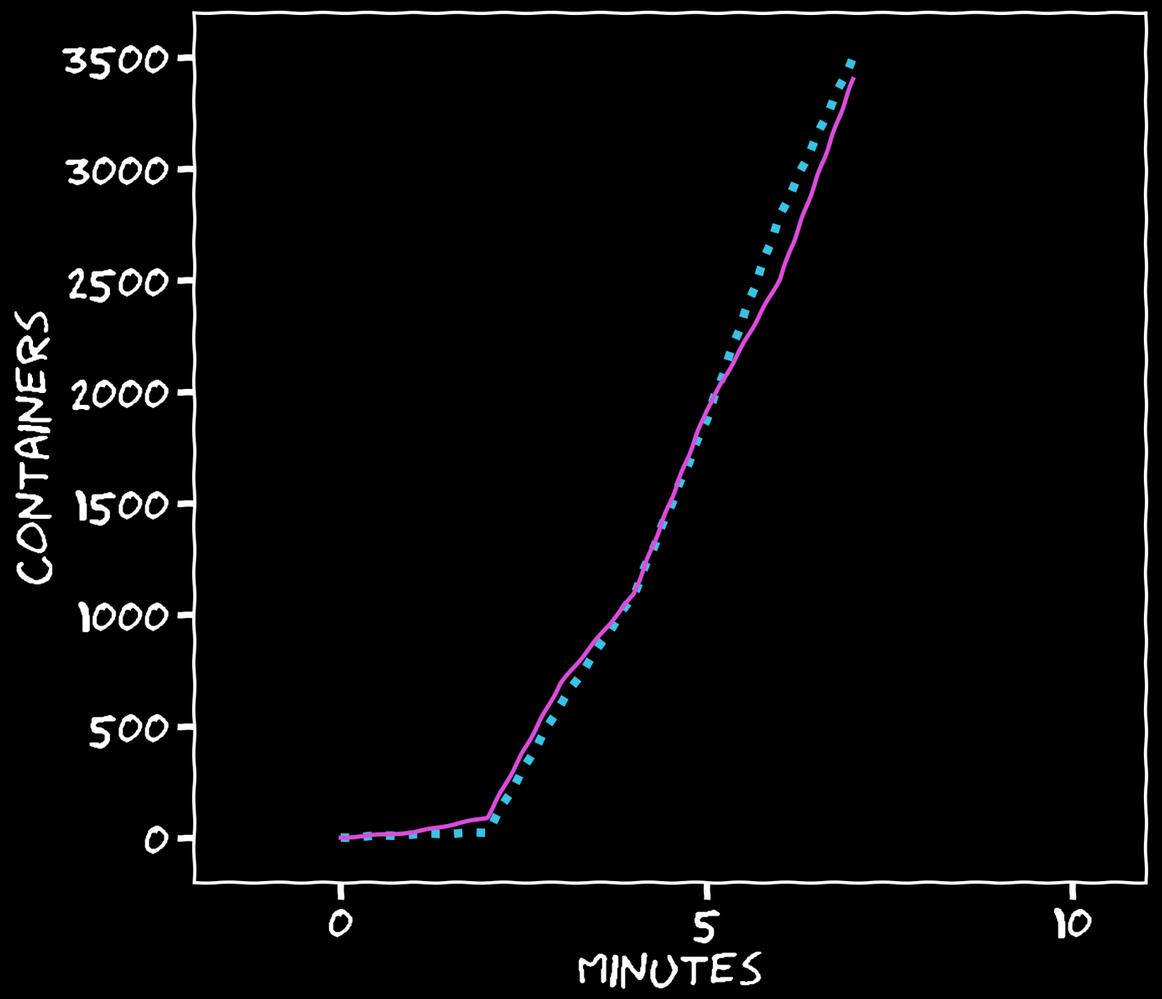
- Amazon EKS control plane prescaled

Amazon EKS on Amazon EC2



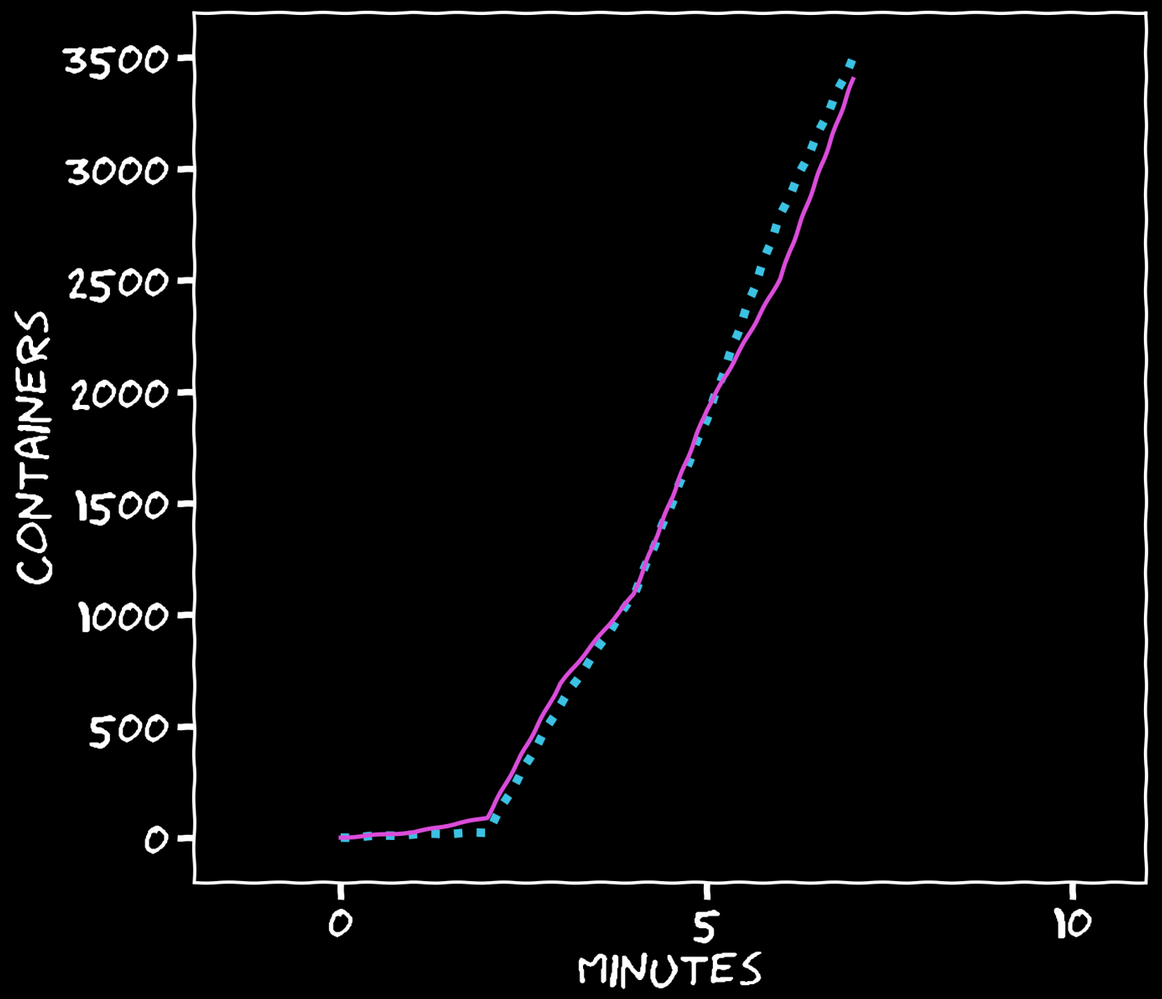
- Amazon EKS control plane prescaled
- Amazon EC2 Spot Instances

Amazon EKS on Amazon EC2



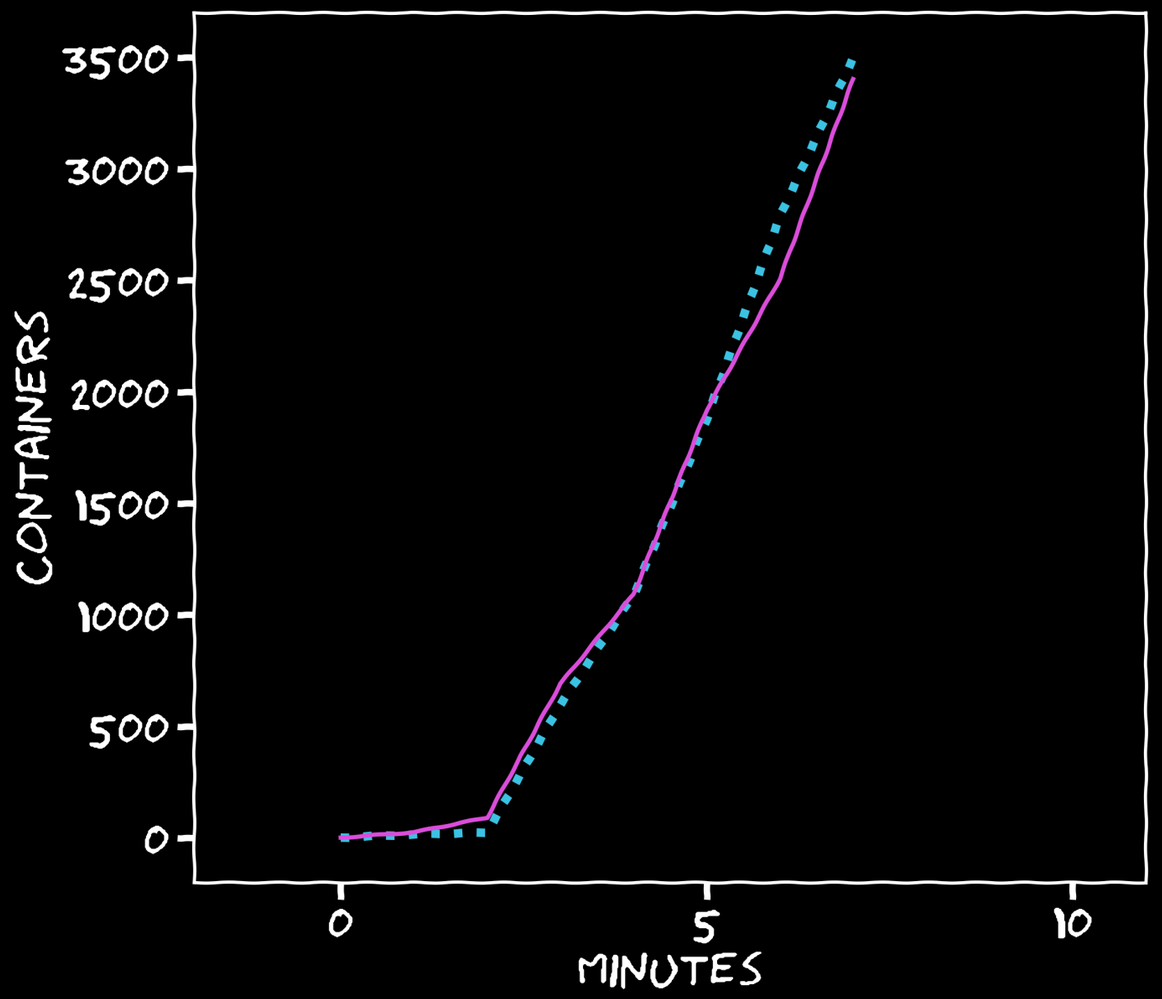
- Amazon EKS control plane prescaled
- Amazon EC2 Spot Instances
  - M5, M5N, C5, C5N, R5

Amazon EKS on Amazon EC2



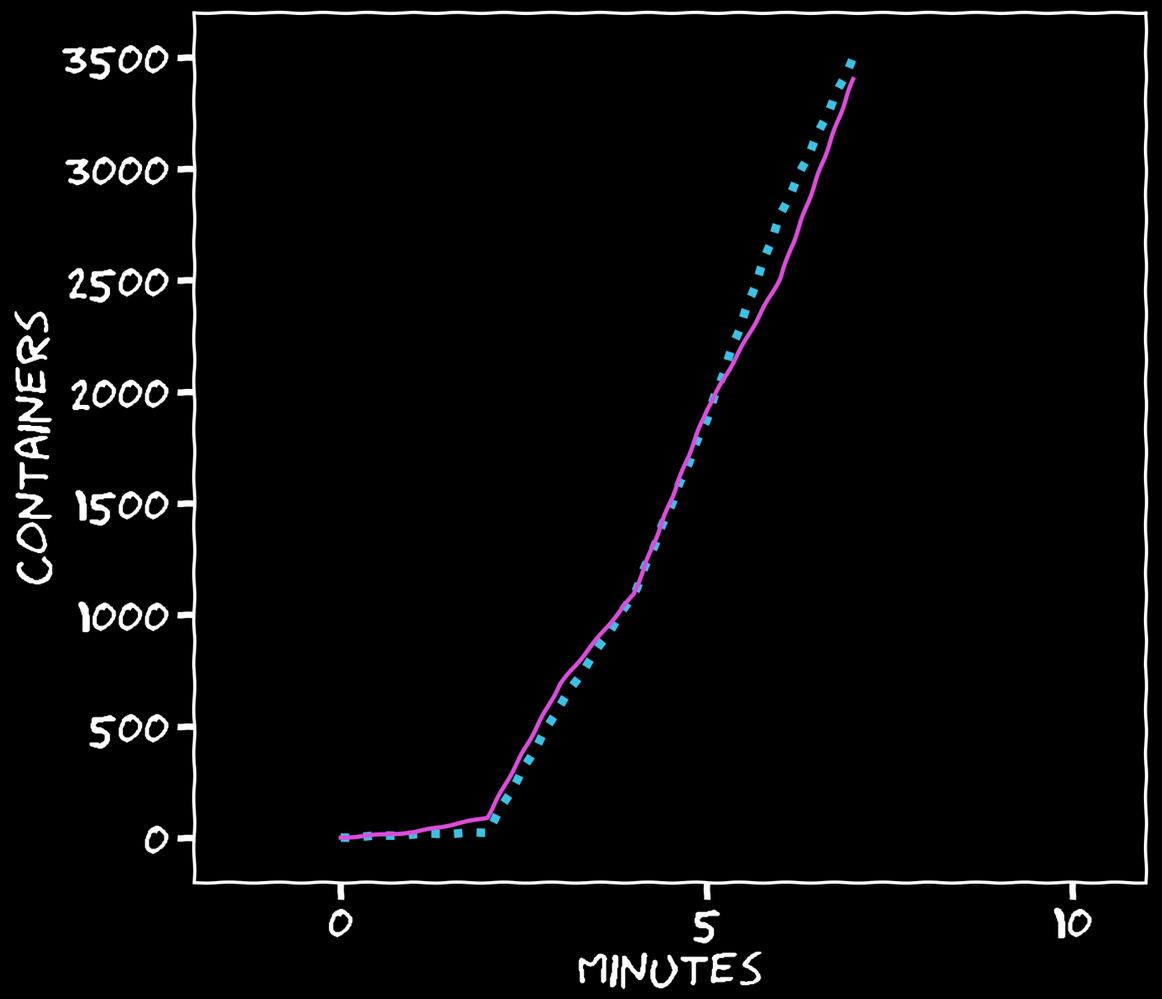
- Amazon EKS control plane prescaled
- Amazon EC2 Spot Instances
  - Auto Scaling group (ASG): from one Amazon EC2 instance scaled up to ~250 Amazon EC2 instances

Amazon EKS on Amazon EC2



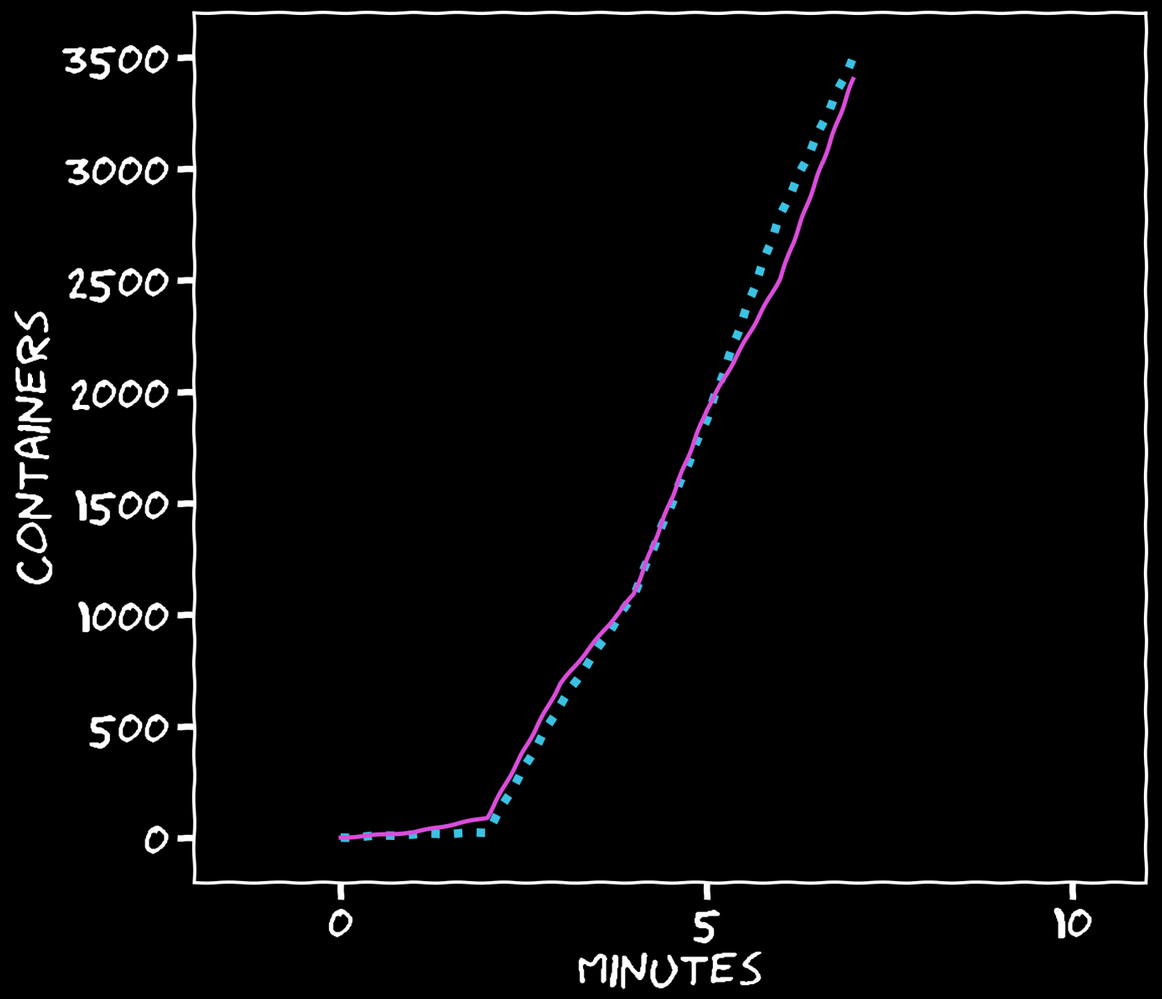
- Amazon EKS control plane prescaled
- Amazon EC2 Spot Instances
  - Auto Scaling group (ASG): from one Amazon EC2 instance scaled up to ~250 Amazon EC2 instances
  - Cluster autoscaler

Amazon EKS on Amazon EC2



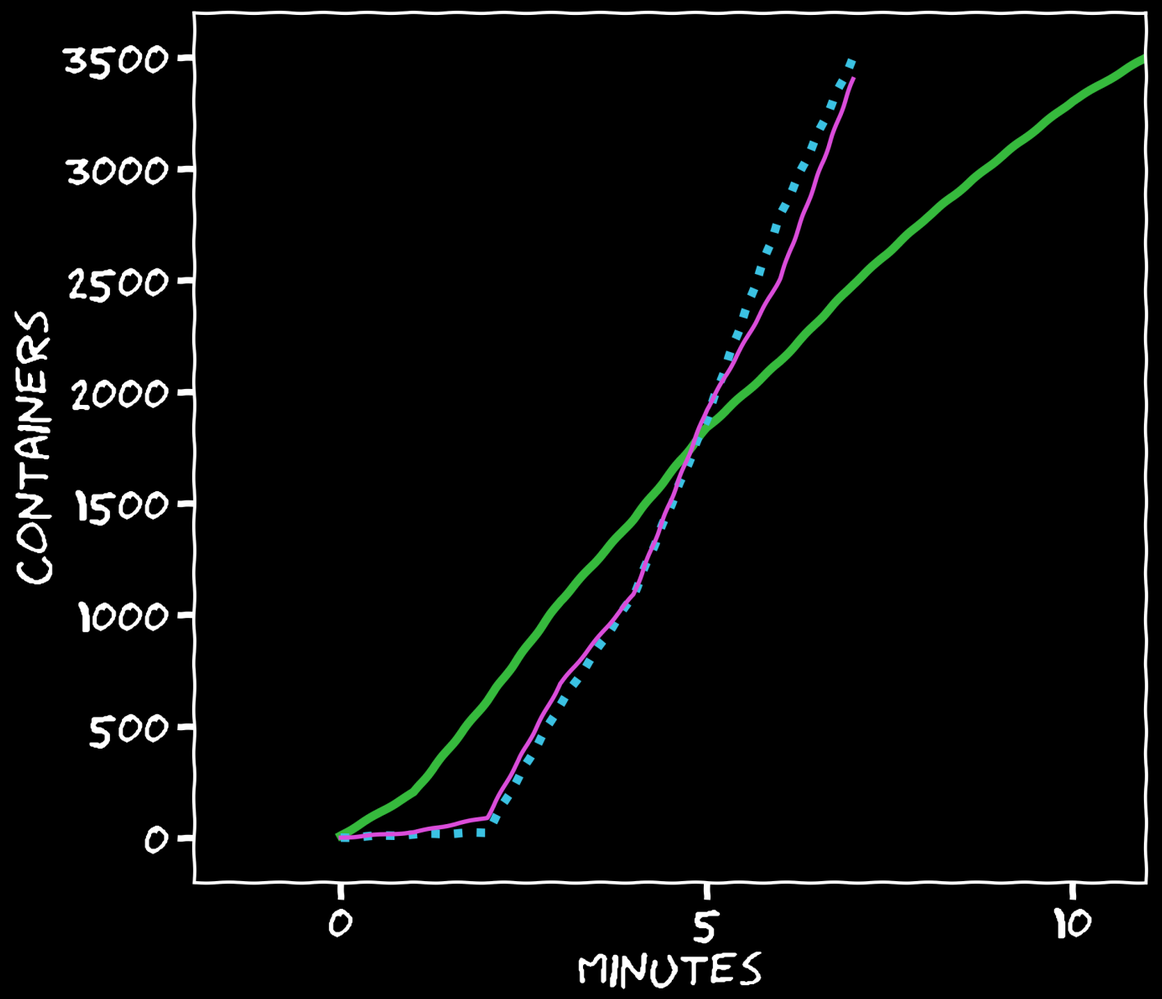
- Amazon EKS control plane prescaled
- Amazon EC2 Spot Instances
- Single deployment

Amazon EKS on Amazon EC2



- Amazon EKS control plane prescaled
- Amazon EC2 Spot Instances
- Single deployment
  - Force-scaled to 3,500 replicas

Amazon EKS on Amazon EC2



- Amazon EKS control plane prescaled
- Amazon EC2 Spot Instances
- Single deployment

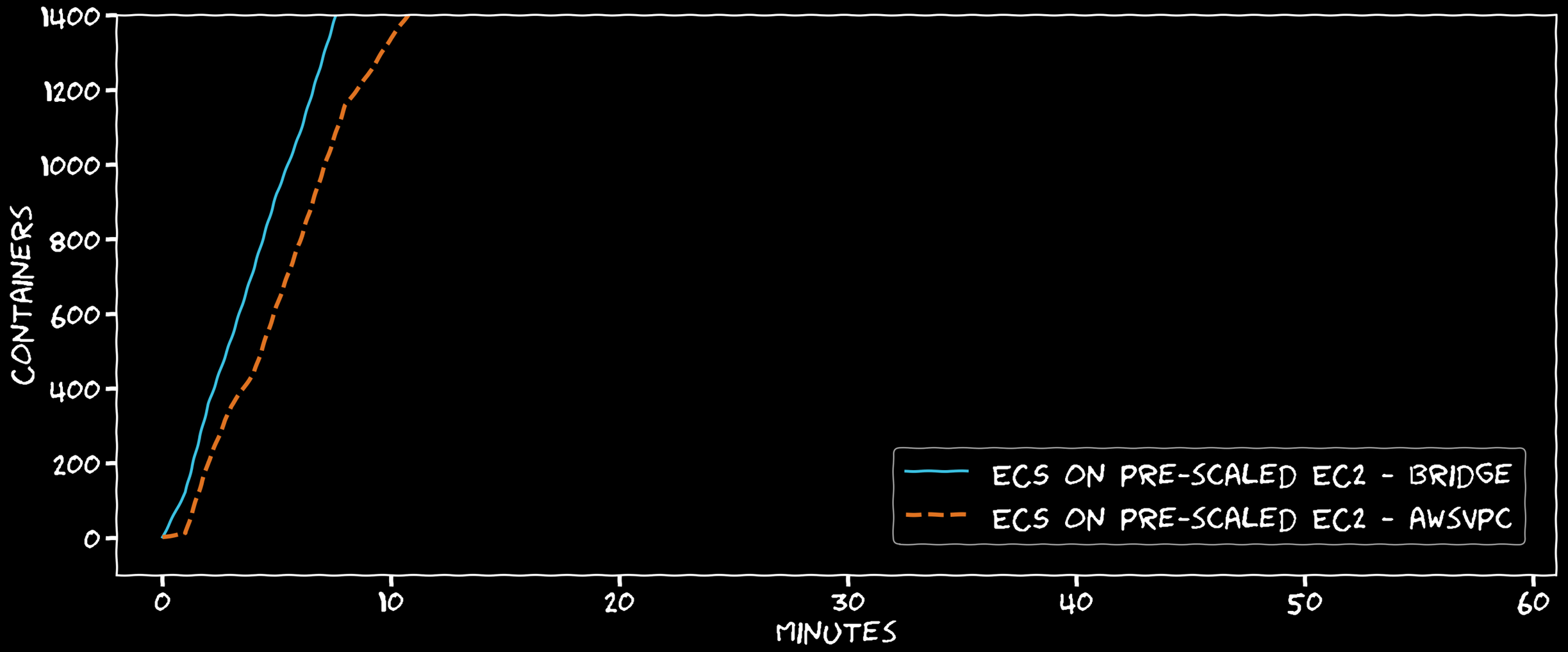
Amazon ECS on Amazon EC2



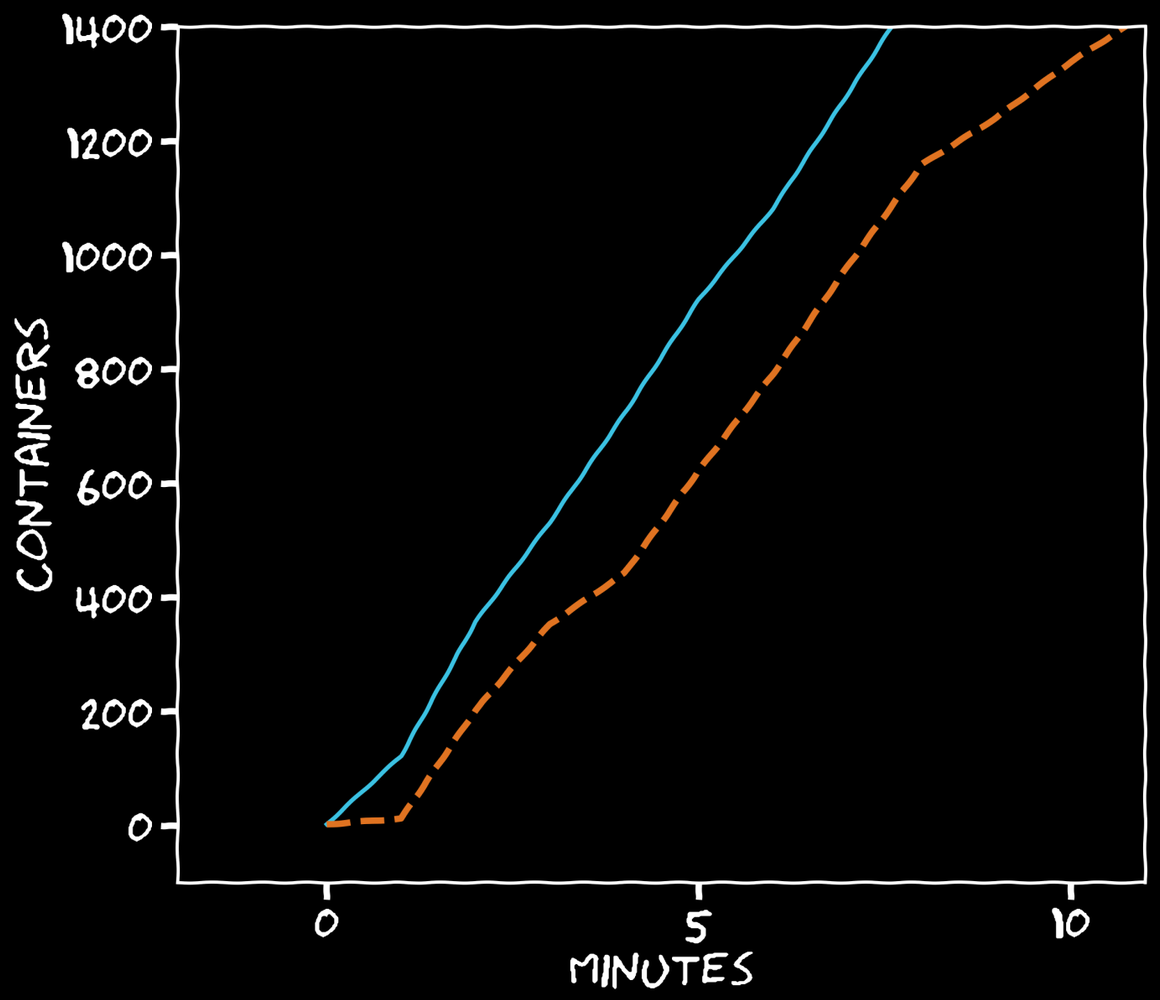
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

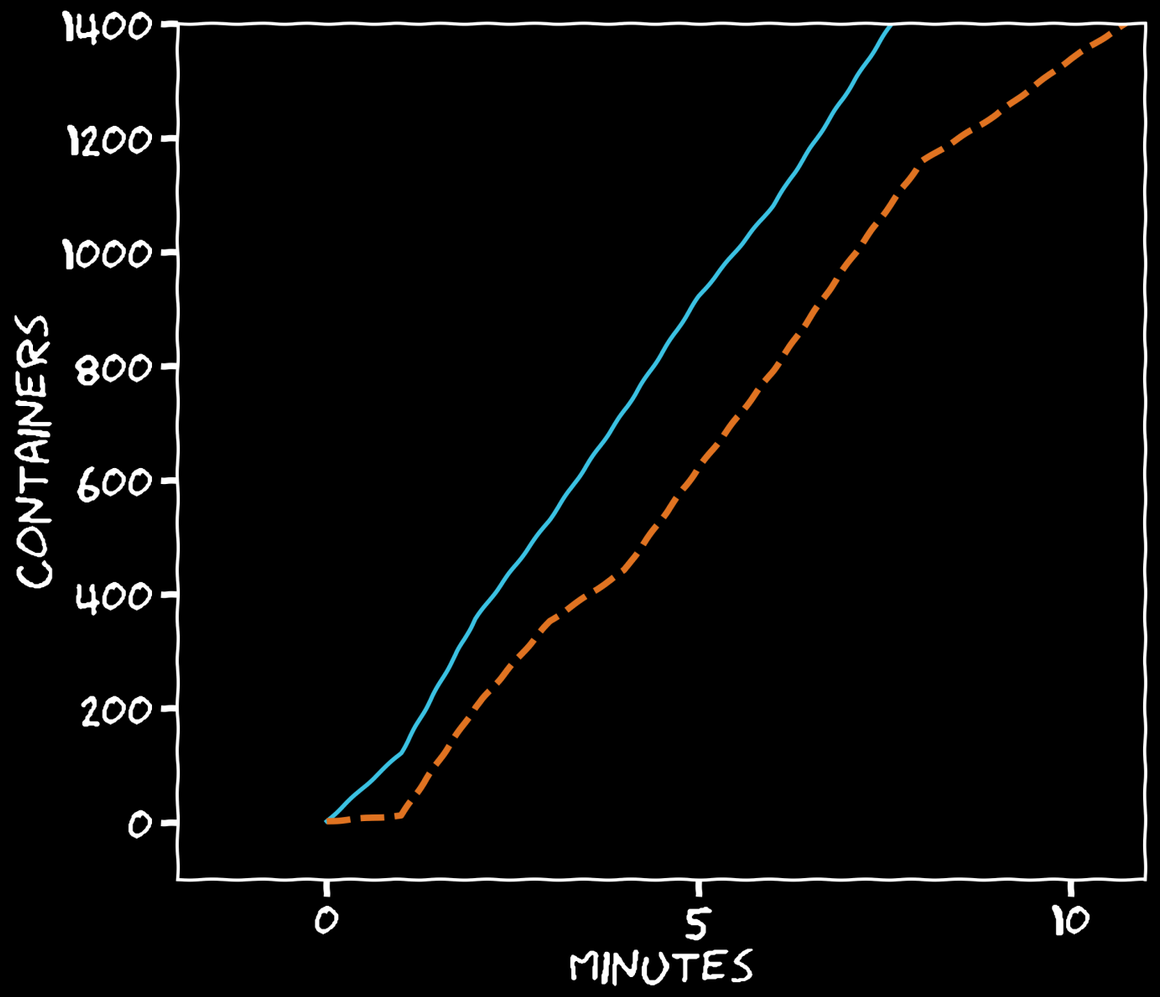
Amazon ECS on Amazon EC2



Amazon ECS on Amazon EC2

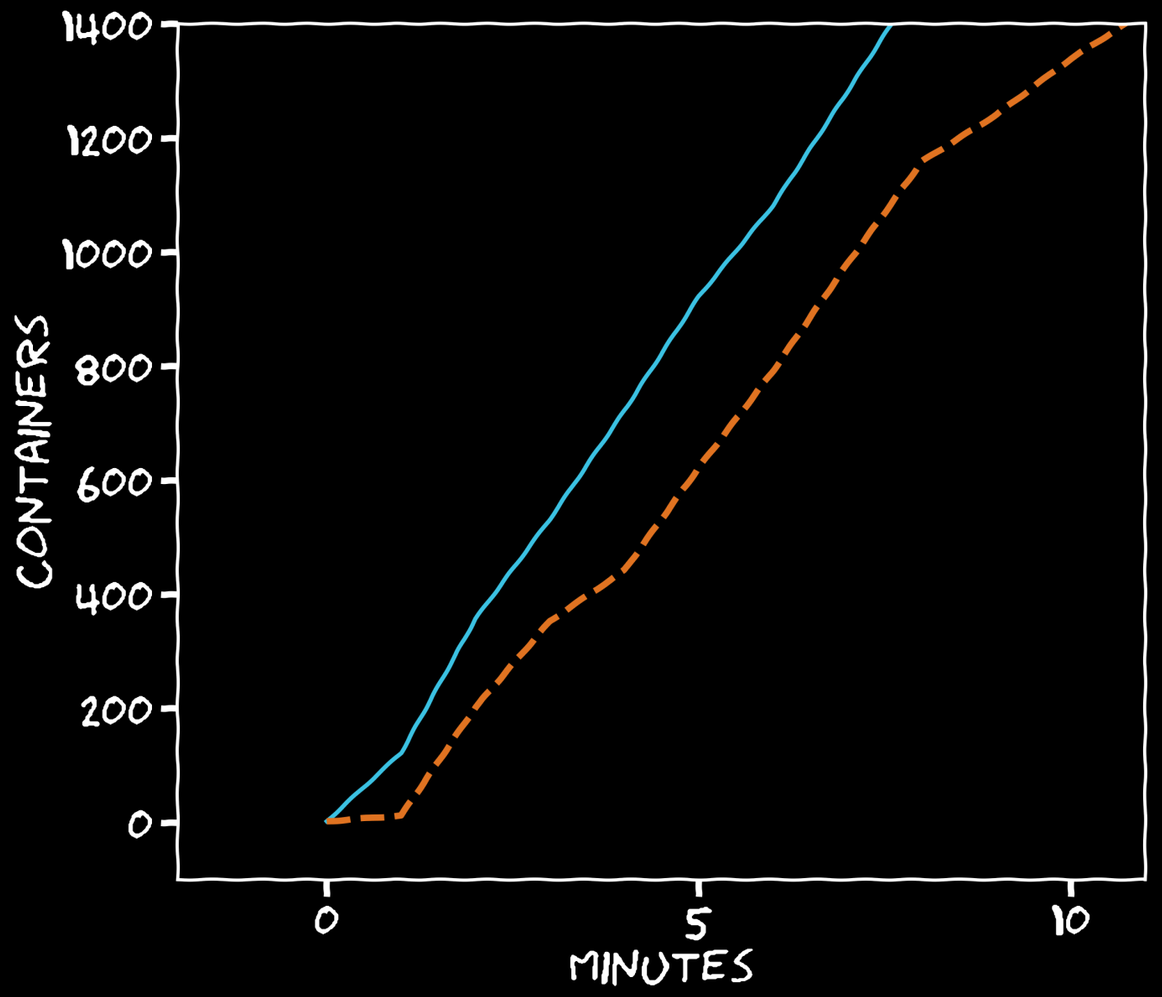


Amazon ECS on Amazon EC2



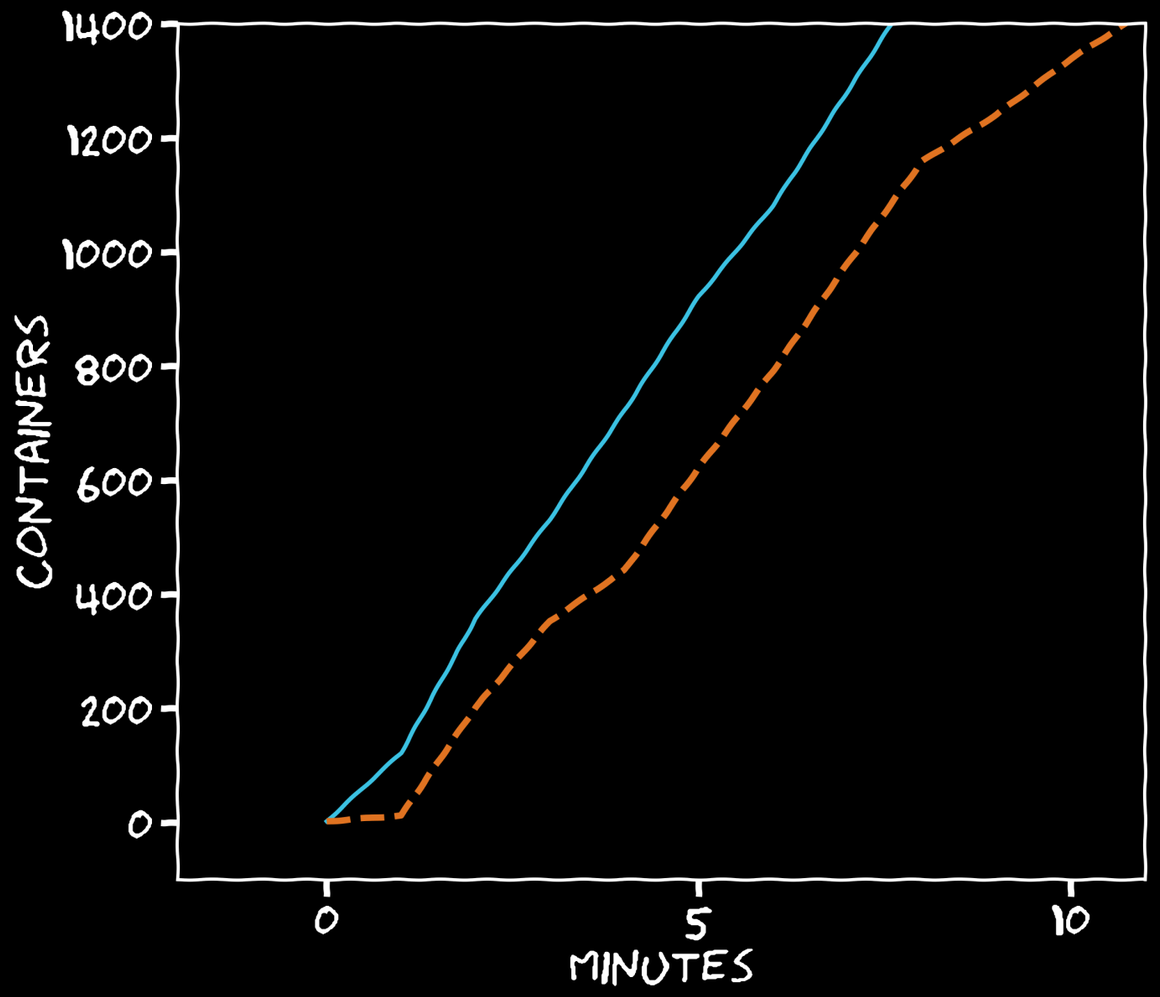
- Scaling to 1,400 containers

Amazon ECS on Amazon EC2



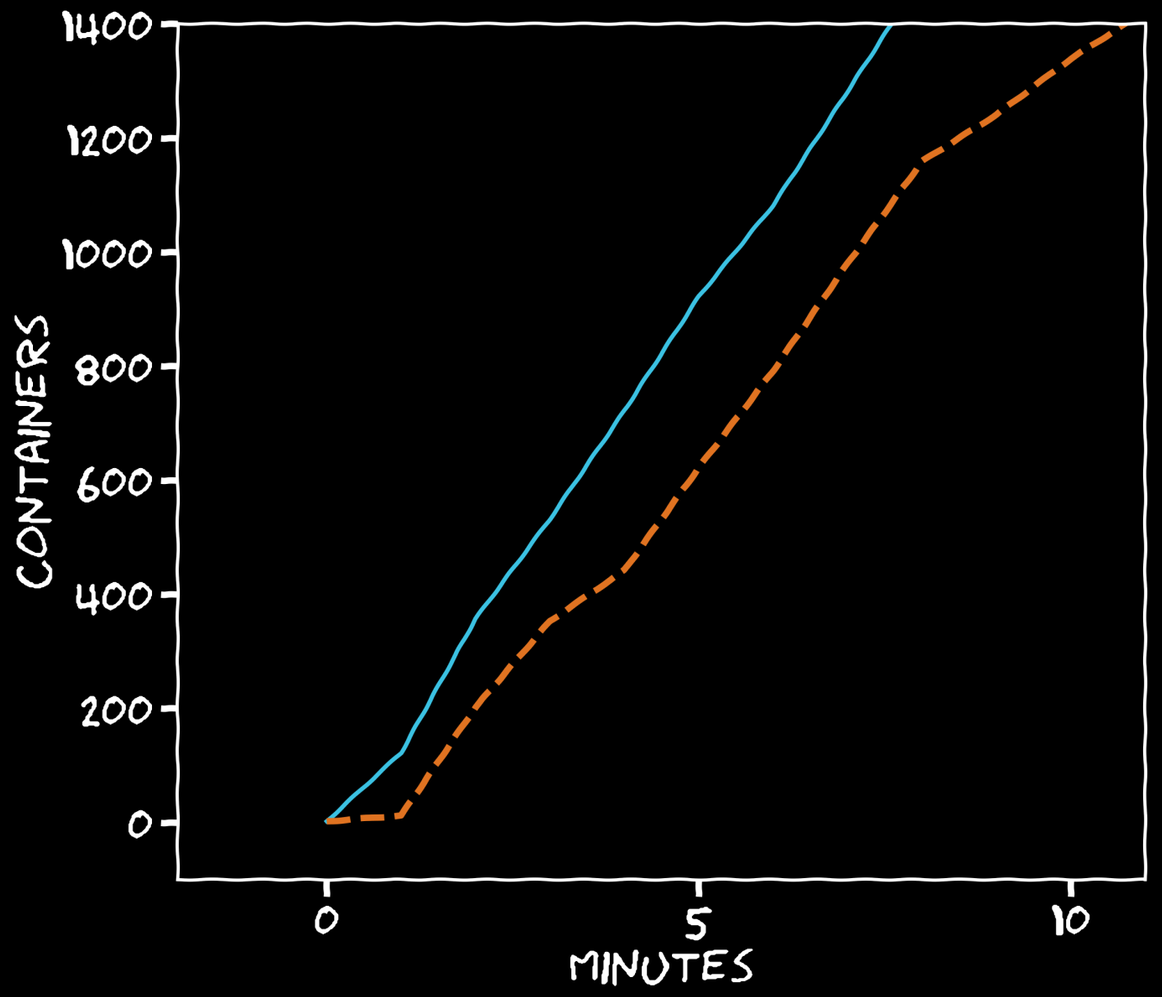
- Scaling to 1,400 containers
  - Attaching ENIs takes a tiny bit

Amazon ECS on Amazon EC2



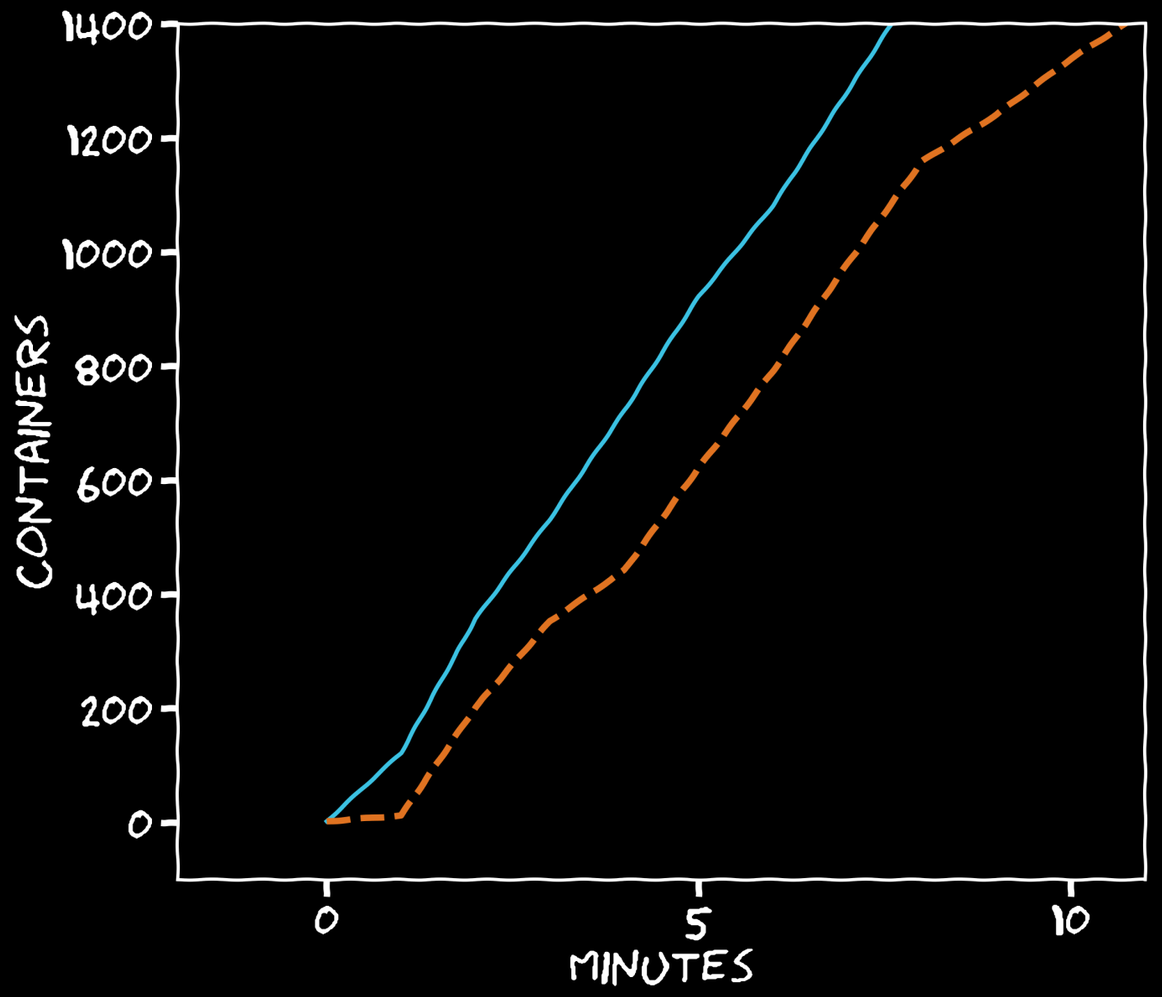
- Scaling to 1,400 containers
- Amazon EC2 Spot Instances

Amazon ECS on Amazon EC2



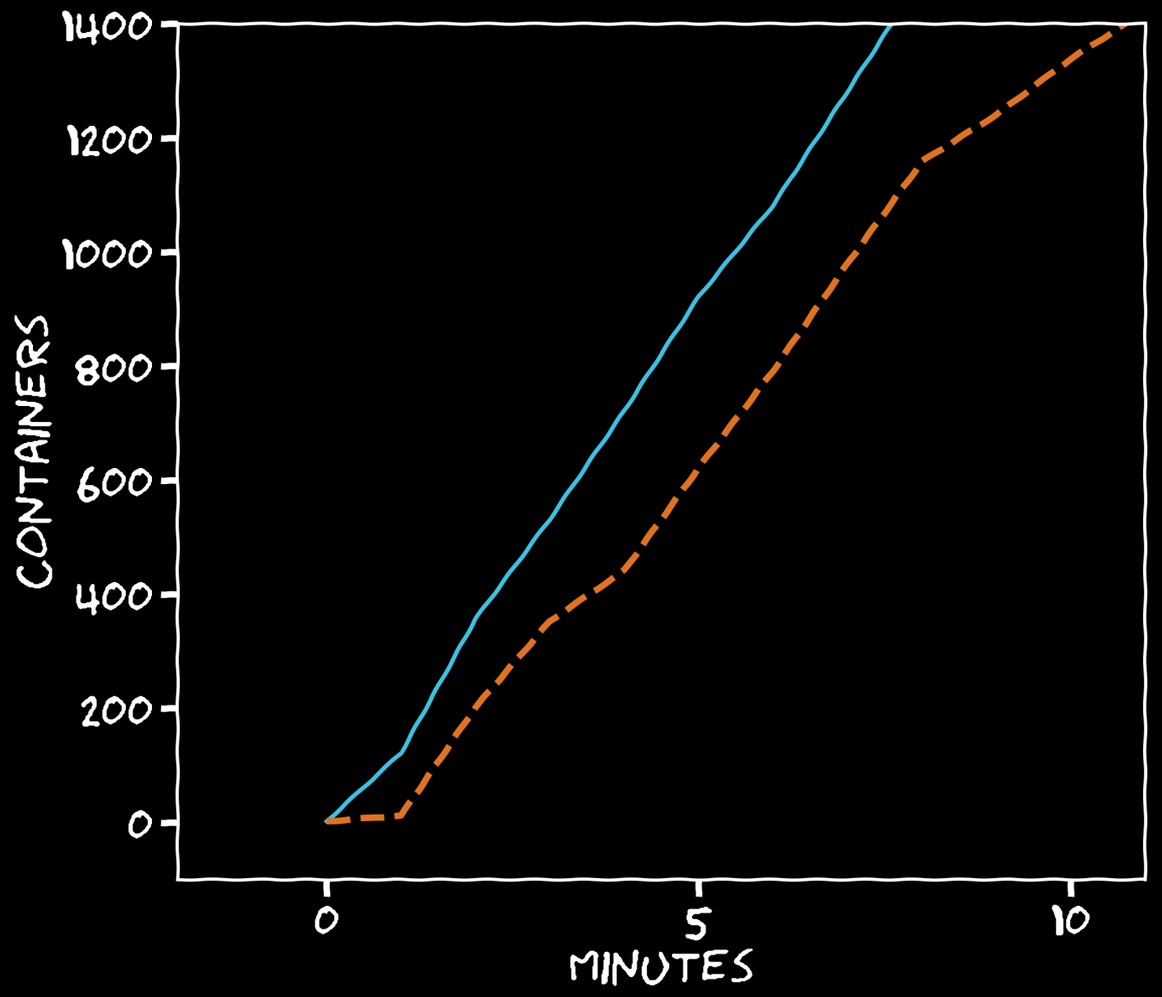
- Scaling to 1,400 containers
- Amazon EC2 Spot Instances
  - AWS Auto Scaling policies
  - Amazon ECS cluster scaling

Amazon ECS on Amazon EC2



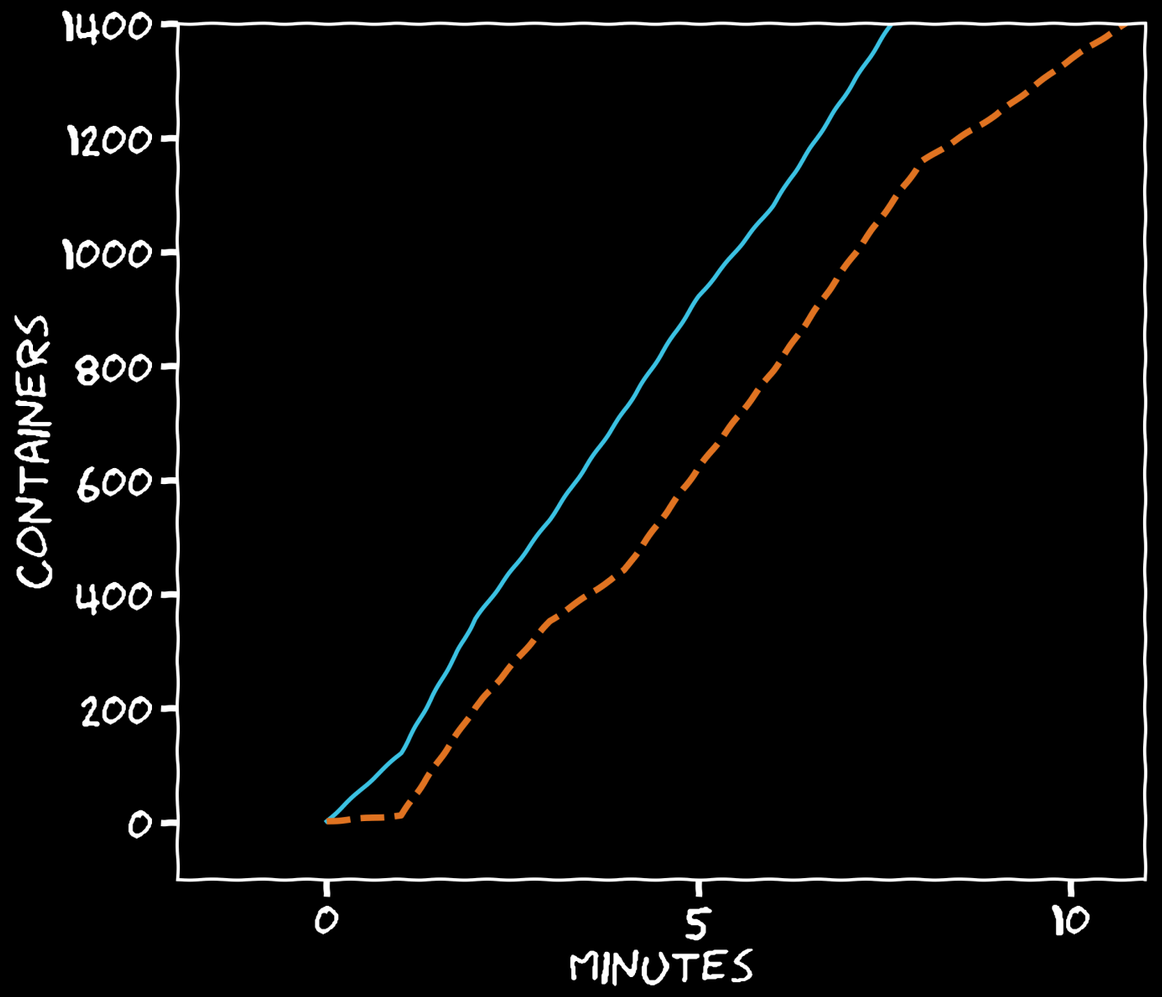
- Scaling to 1,400 containers
- Amazon EC2 Spot Instances
  - AWS Auto Scaling policies
  - Amazon ECS cluster scaling
  - Prescaled ASG
  - ASG: fixed size of ~250

Amazon ECS on Amazon EC2



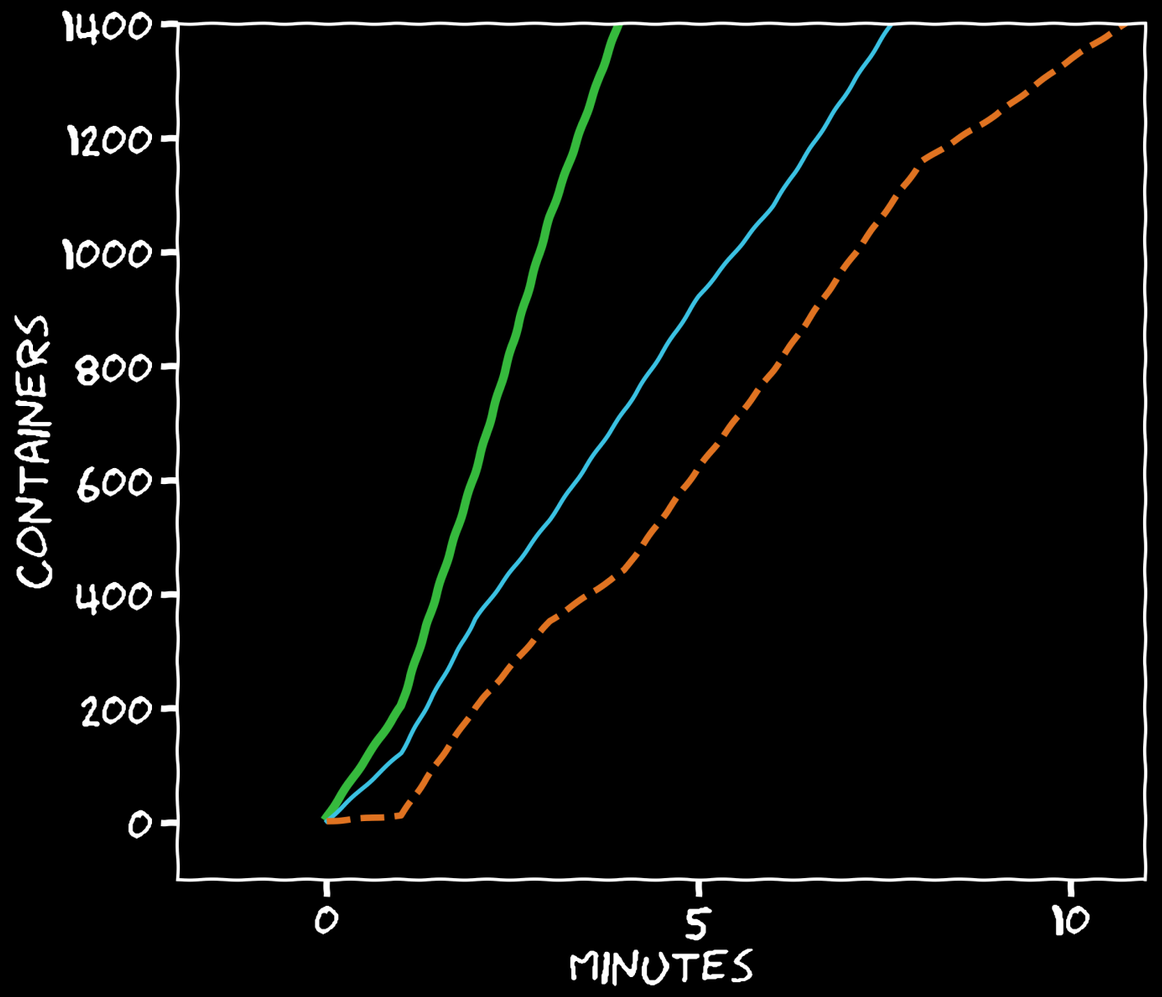
- Scaling to 1,400 containers
- Amazon EC2 Spot Instances
- Custom limits
  - 2 TPS per service

Amazon ECS on Amazon EC2



- Scaling to 1,400 containers
- Amazon EC2 Spot Instances
- Custom limits
  - 2 TPS per service ~ “prescaling” control plane

Amazon ECS on Amazon EC2



- Scaling to 1,400 containers
- Amazon EC2 Spot Instances
- Custom limits

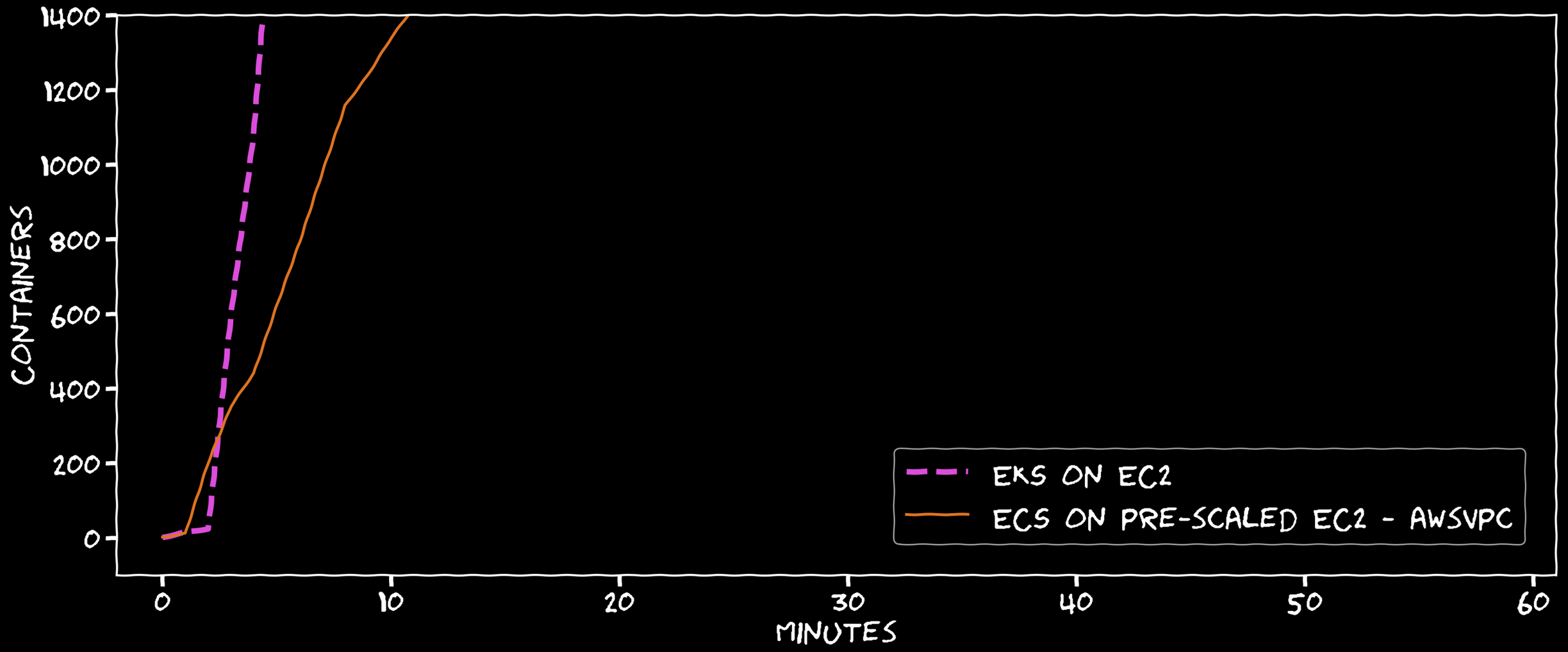
Amazon EC2 workers discussion



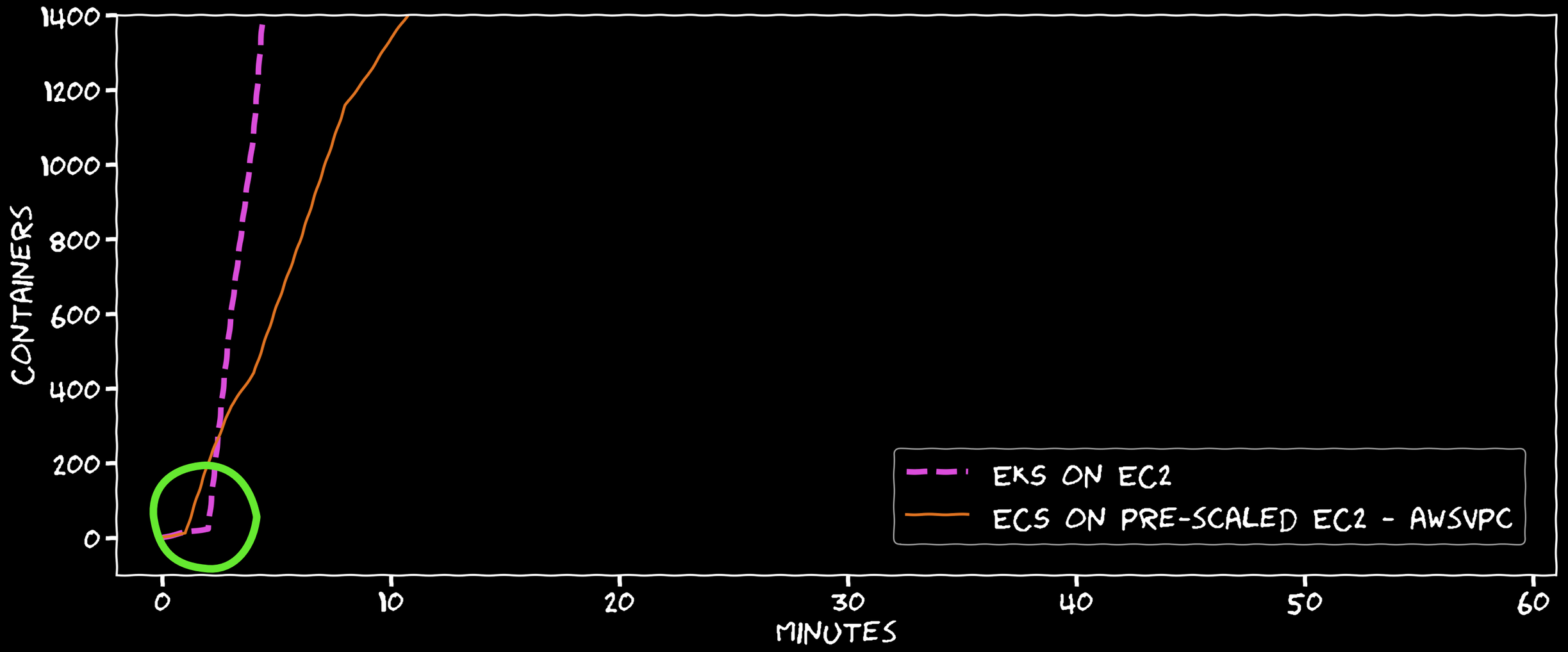
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

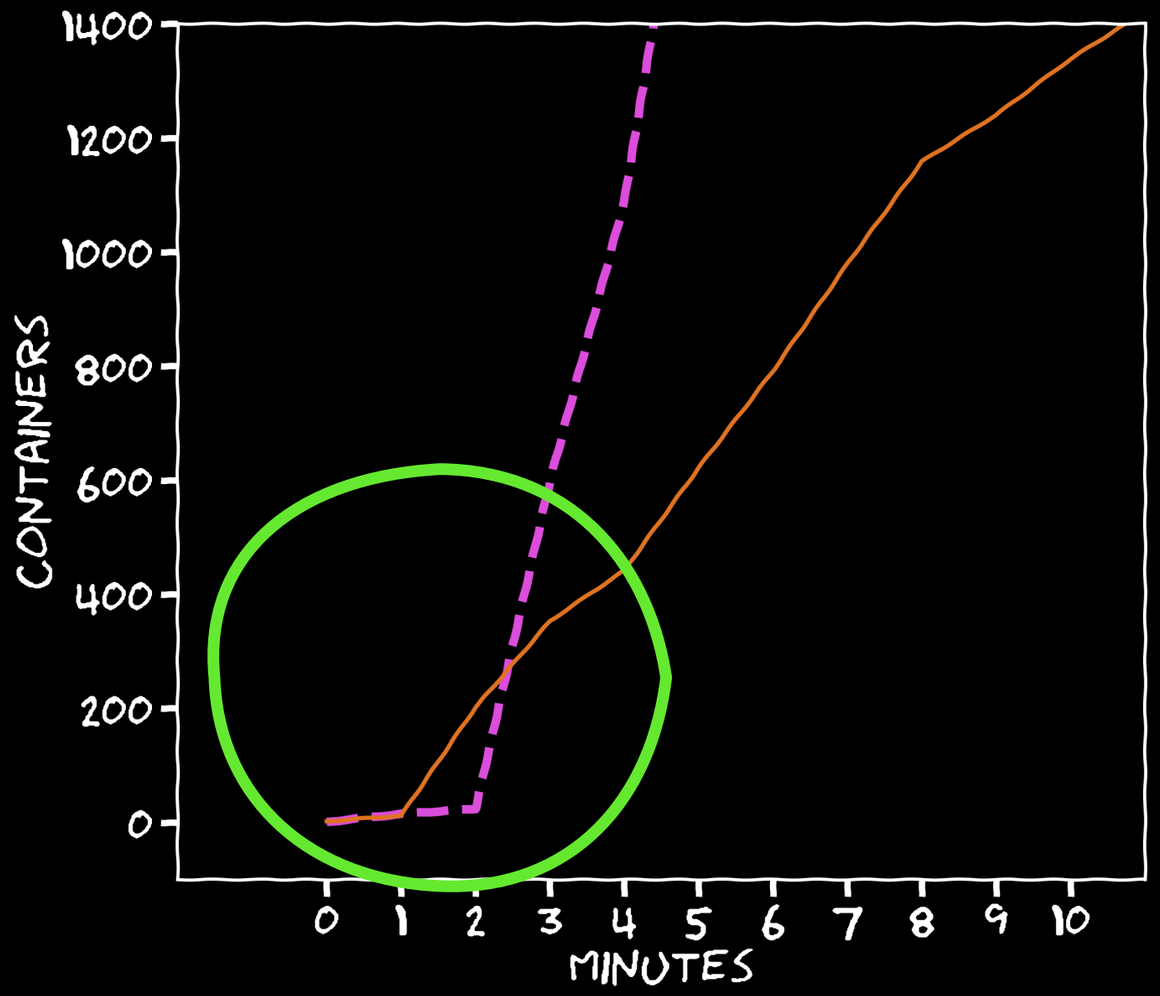
Amazon EC2 compute



Amazon EC2 compute

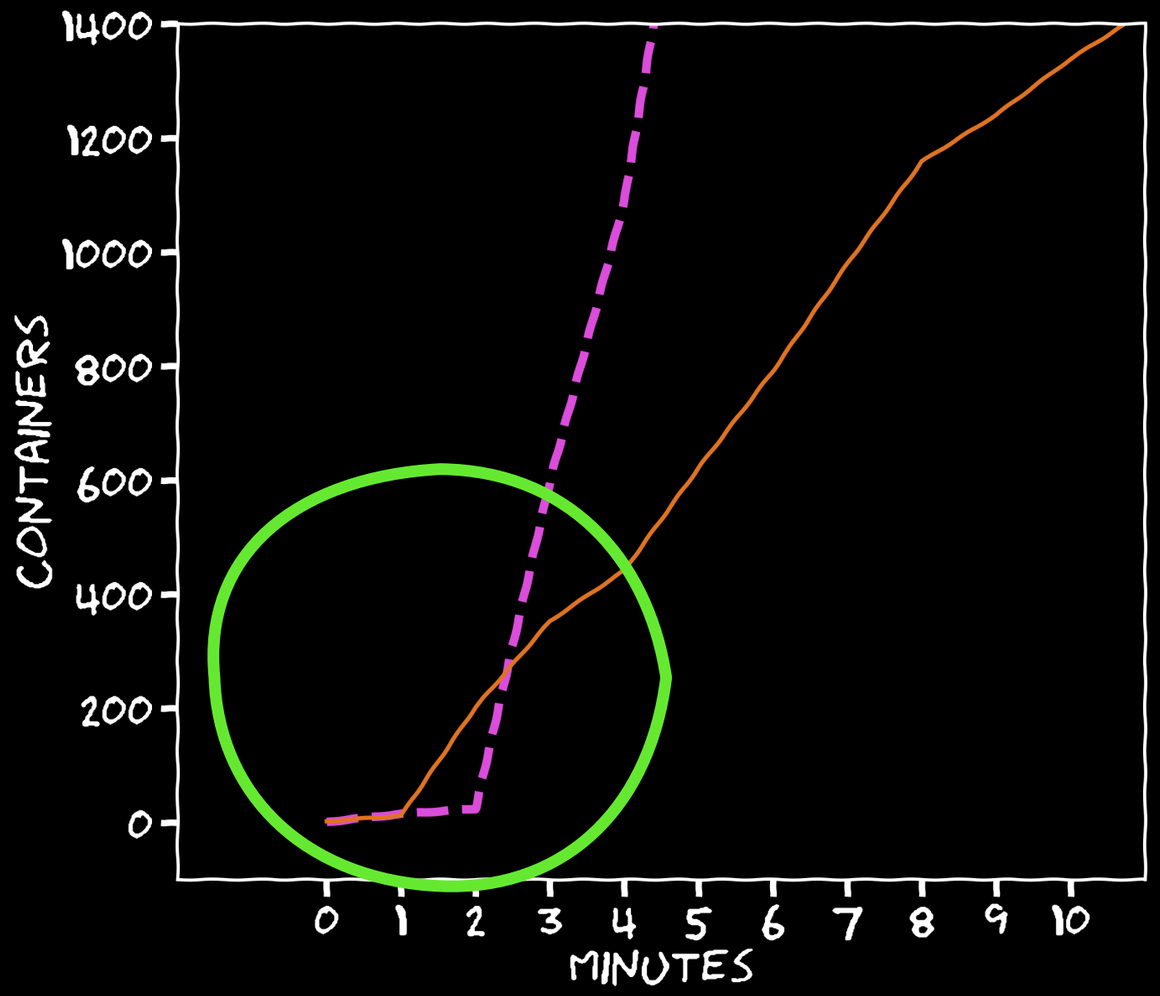


Amazon EC2 compute



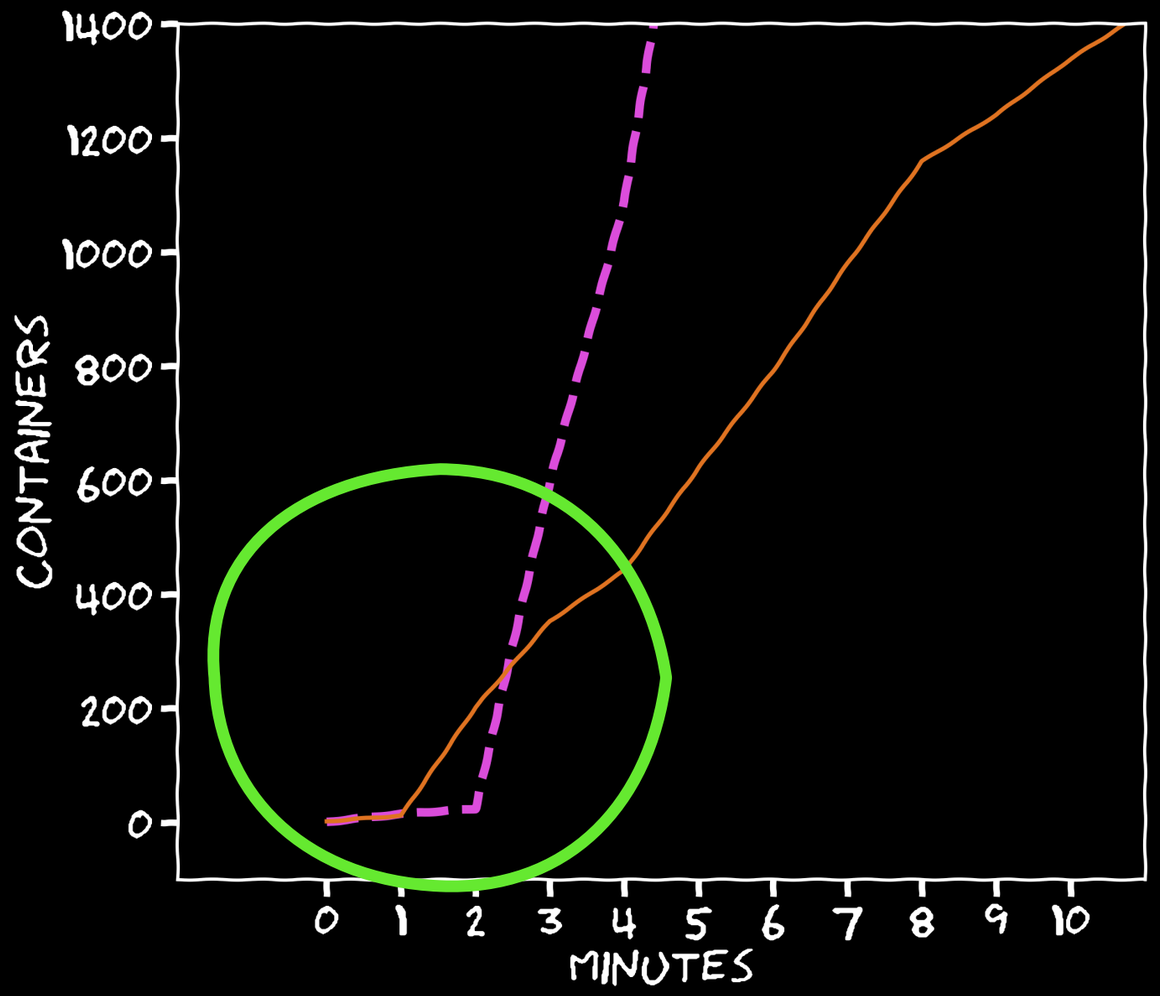
- Scaling delay

Amazon EC2 compute



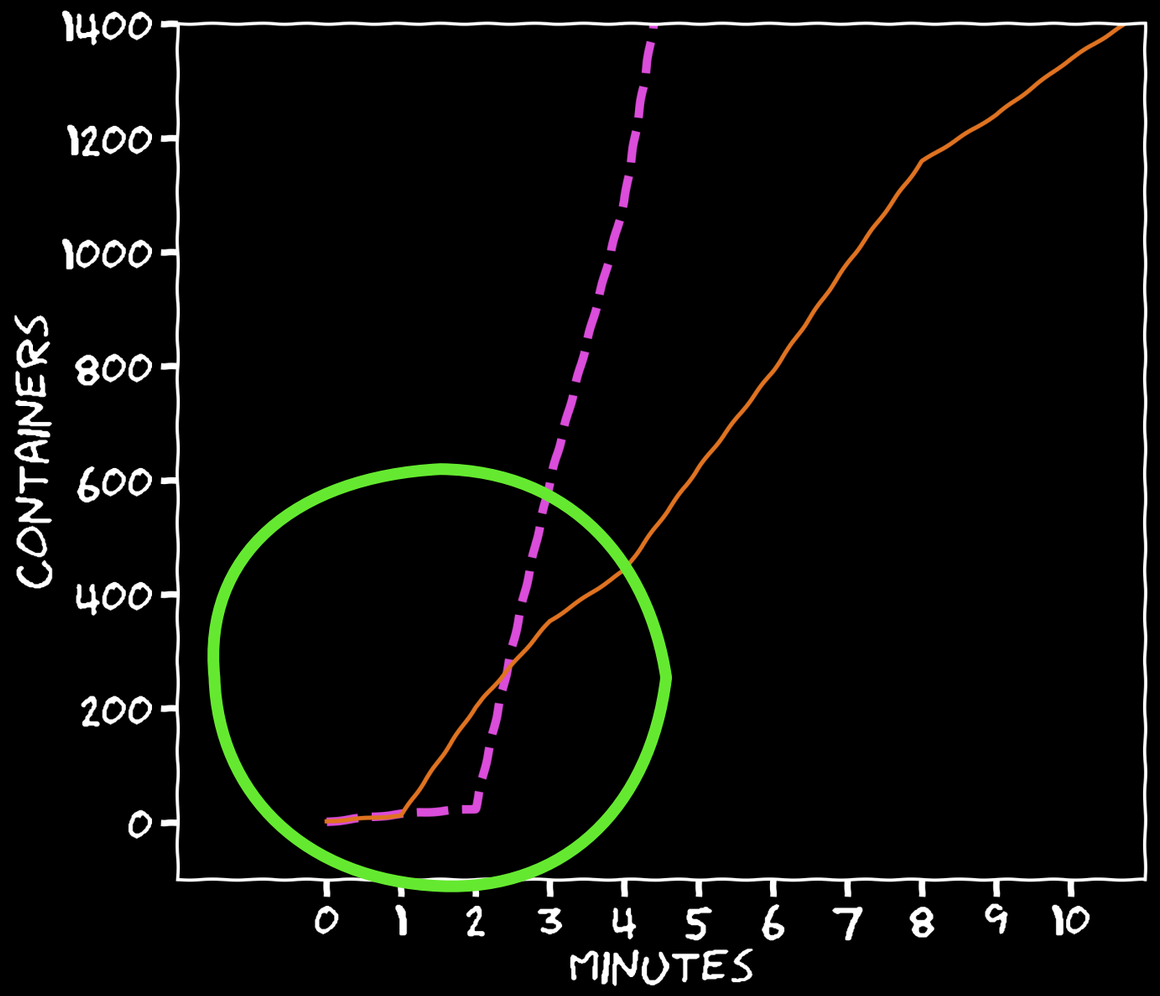
- Scaling delay
  - So, what's the impact?

Amazon EC2 compute



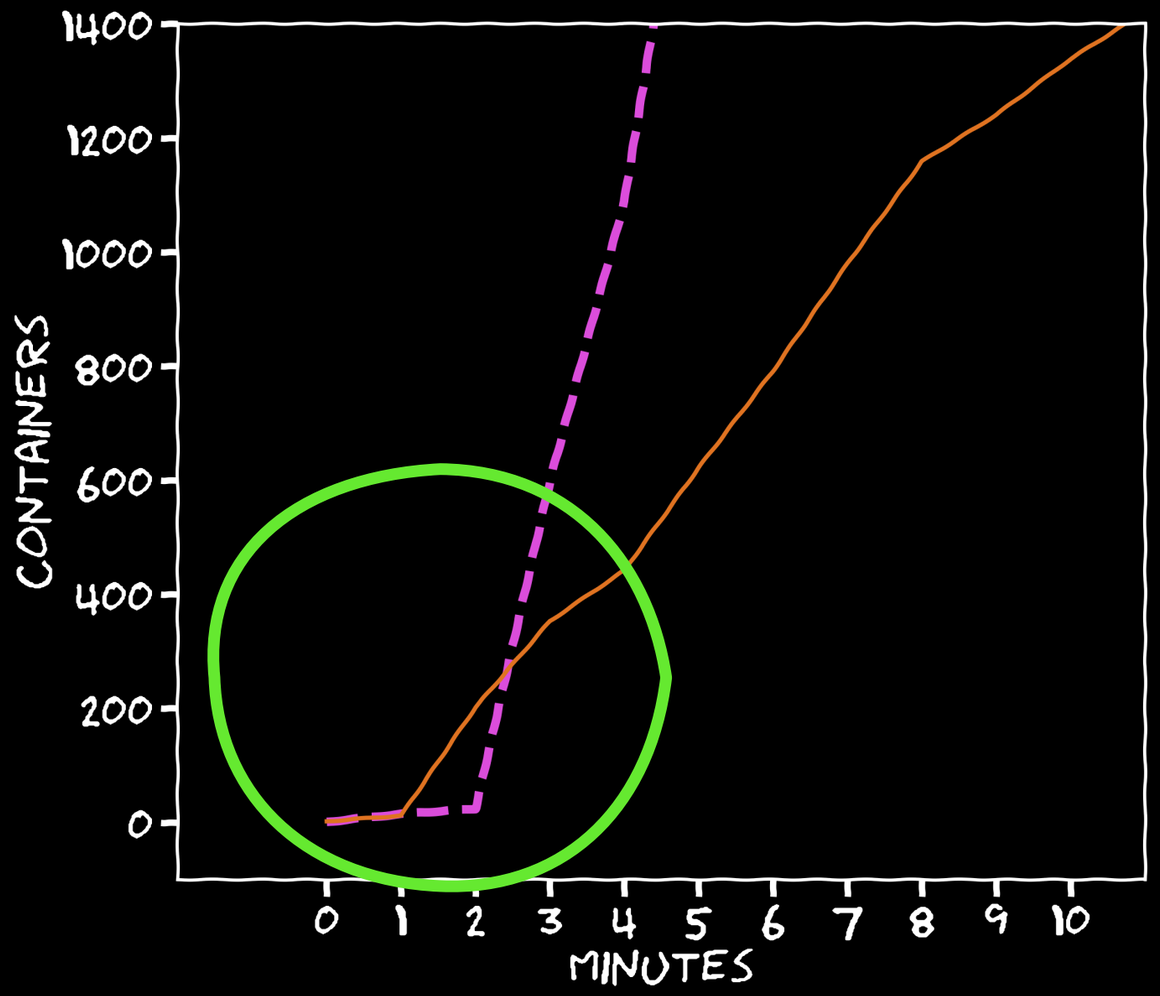
- Scaling delay
  - Web applications?

Amazon EC2 compute



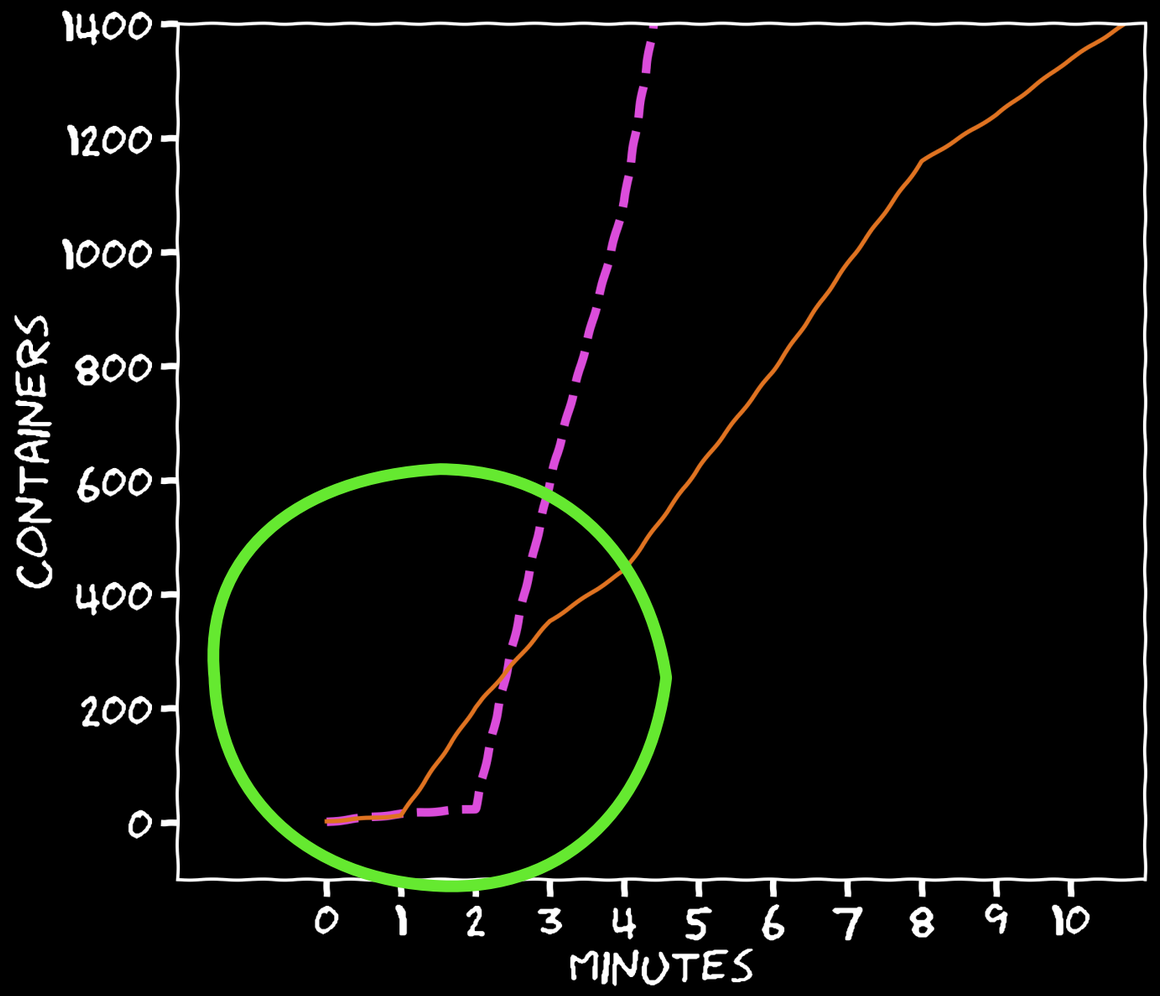
- Scaling delay
  - Web applications? X

Amazon EC2 compute



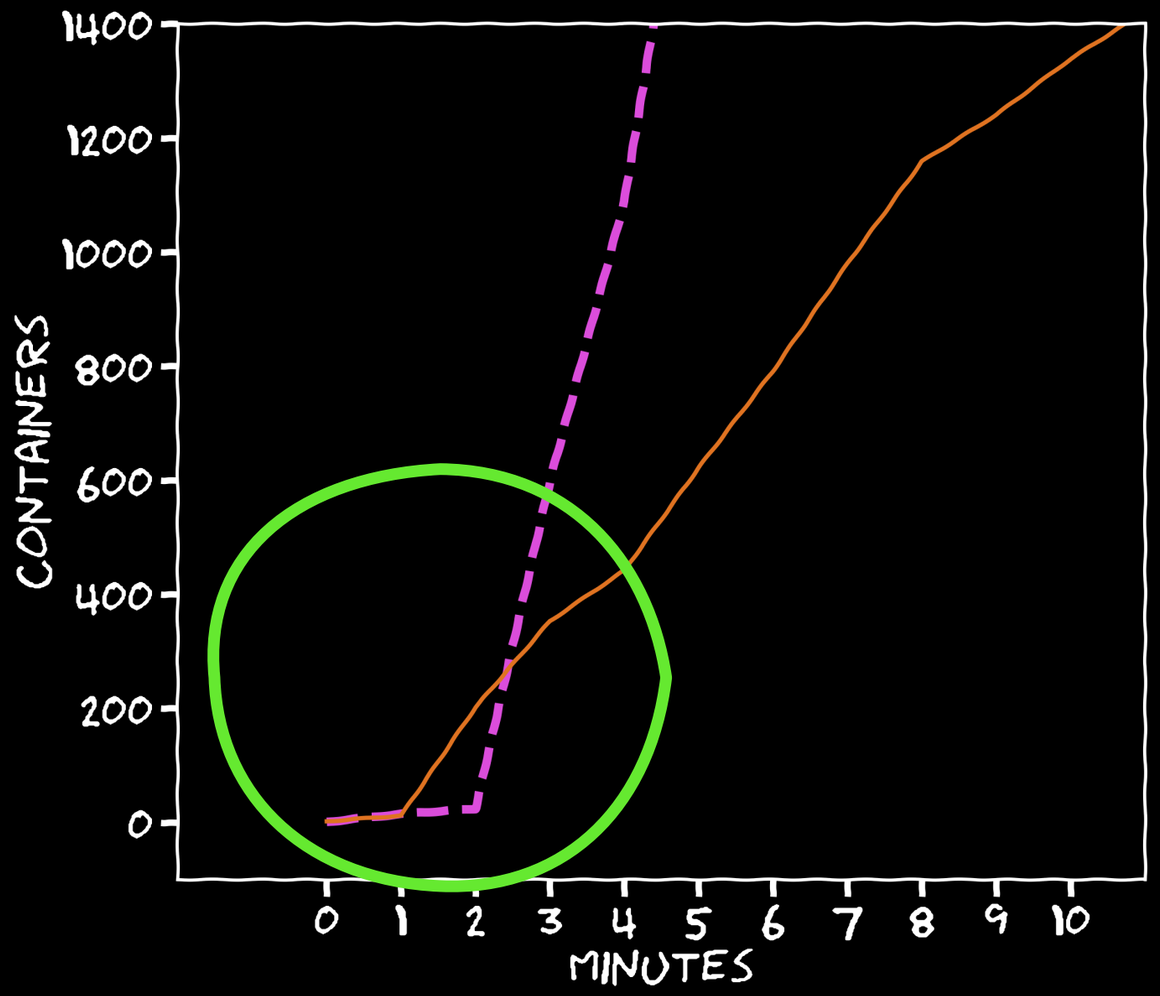
- Scaling delay
  - Web applications? X
  - User-facing applications?

Amazon EC2 compute



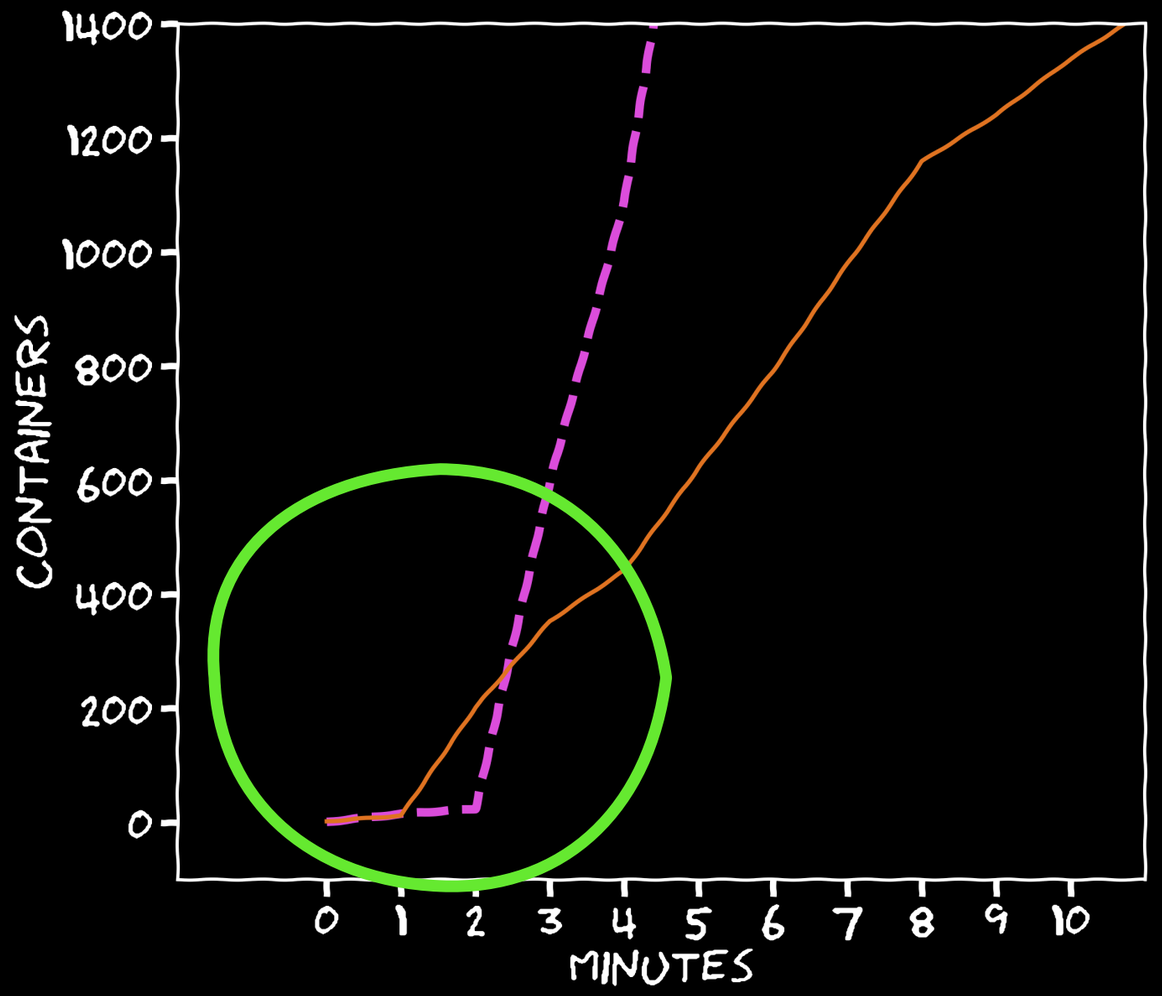
- Scaling delay
  - Web applications? X
  - User-facing applications? X

Amazon EC2 compute



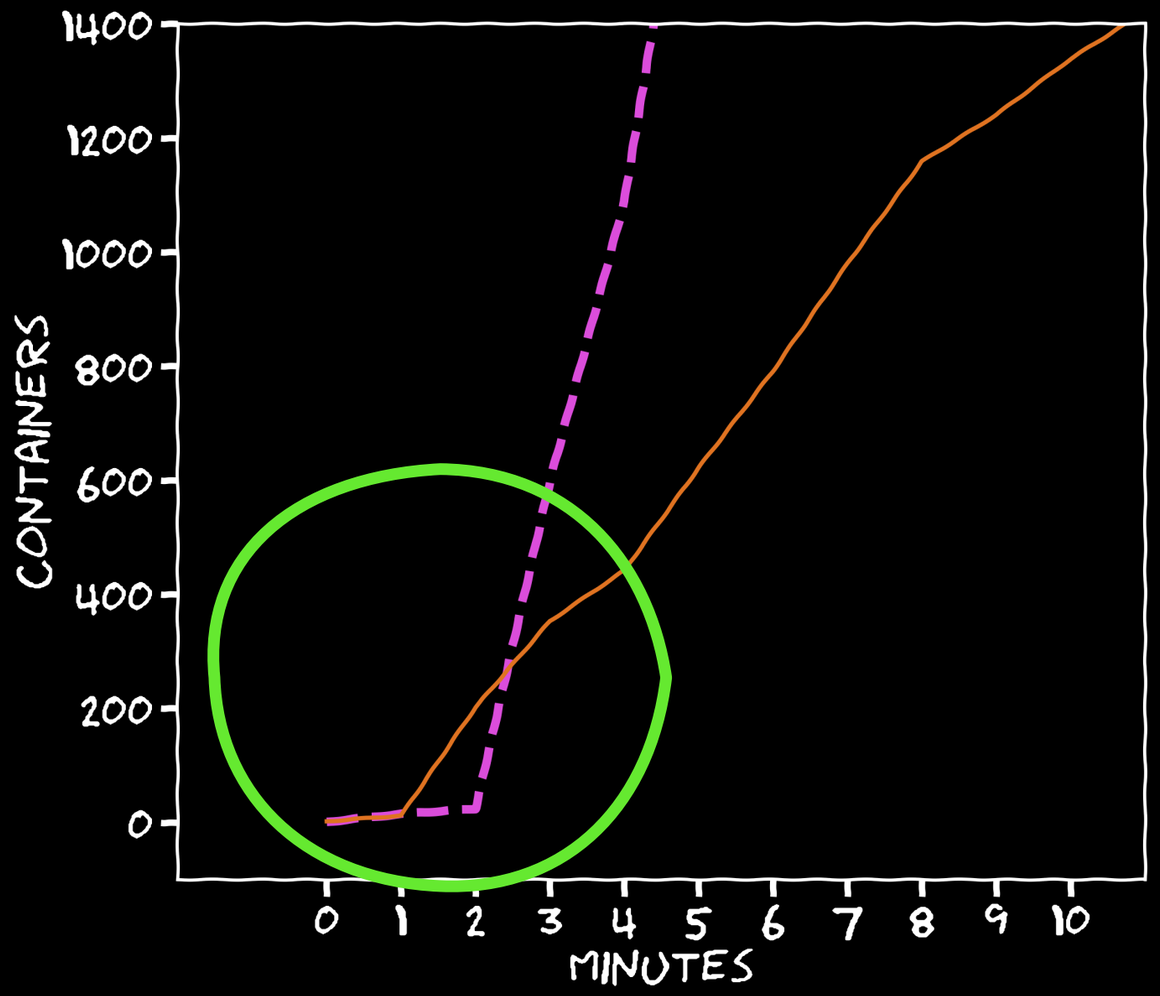
- Scaling delay
  - Web applications? X
  - User-facing applications? X
  - Latency sensitive?

Amazon EC2 compute



- Scaling delay
  - Web applications? X
  - User-facing applications? X
  - Latency sensitive? X

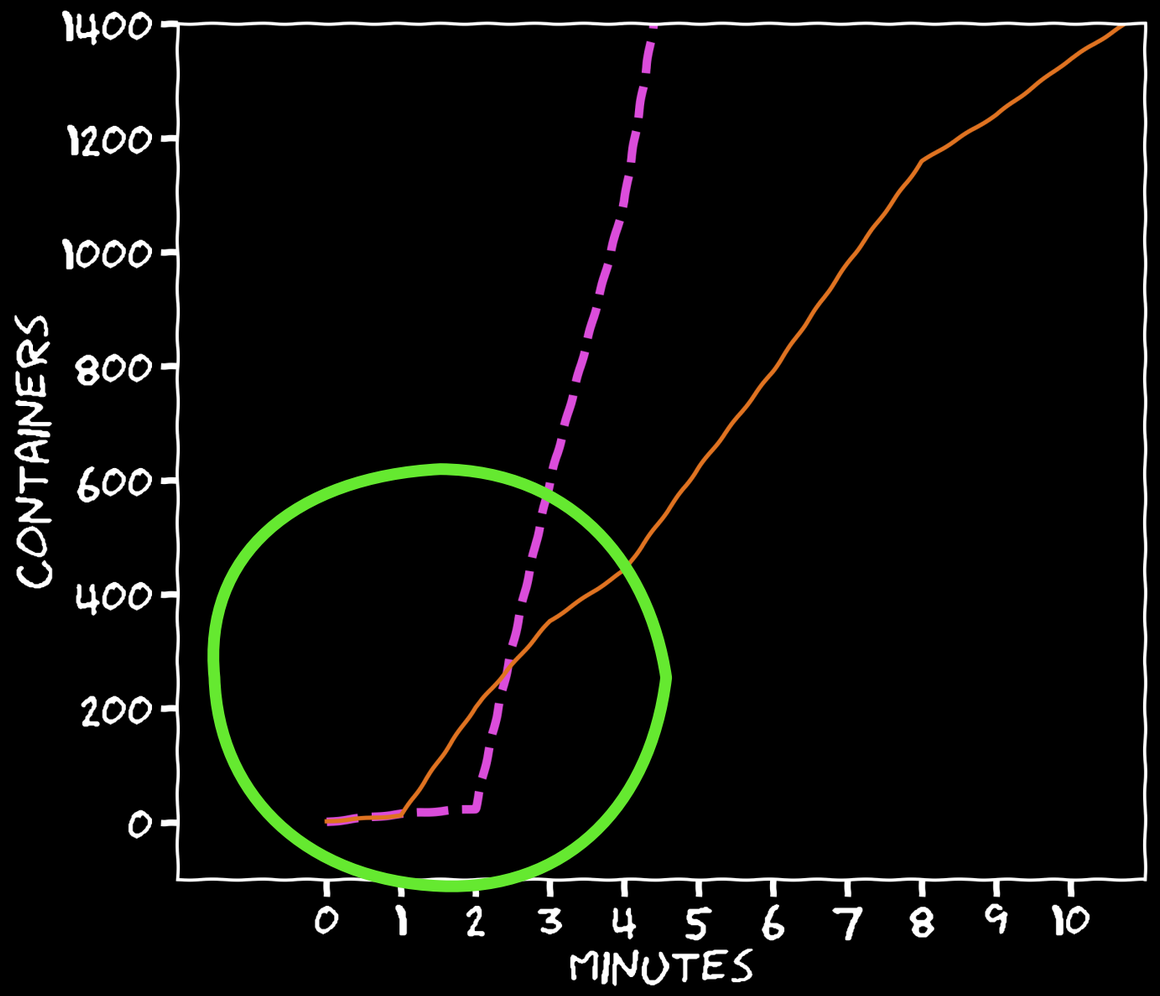
Amazon EC2 compute



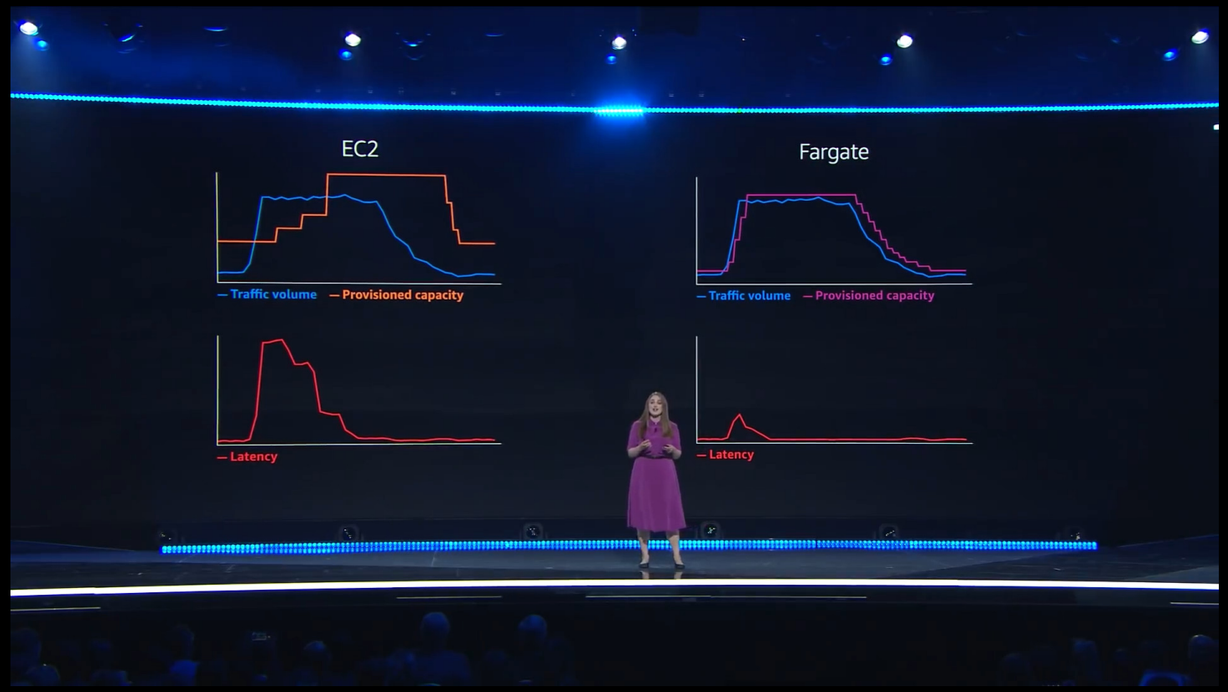
- Scaling delay



Amazon EC2 compute

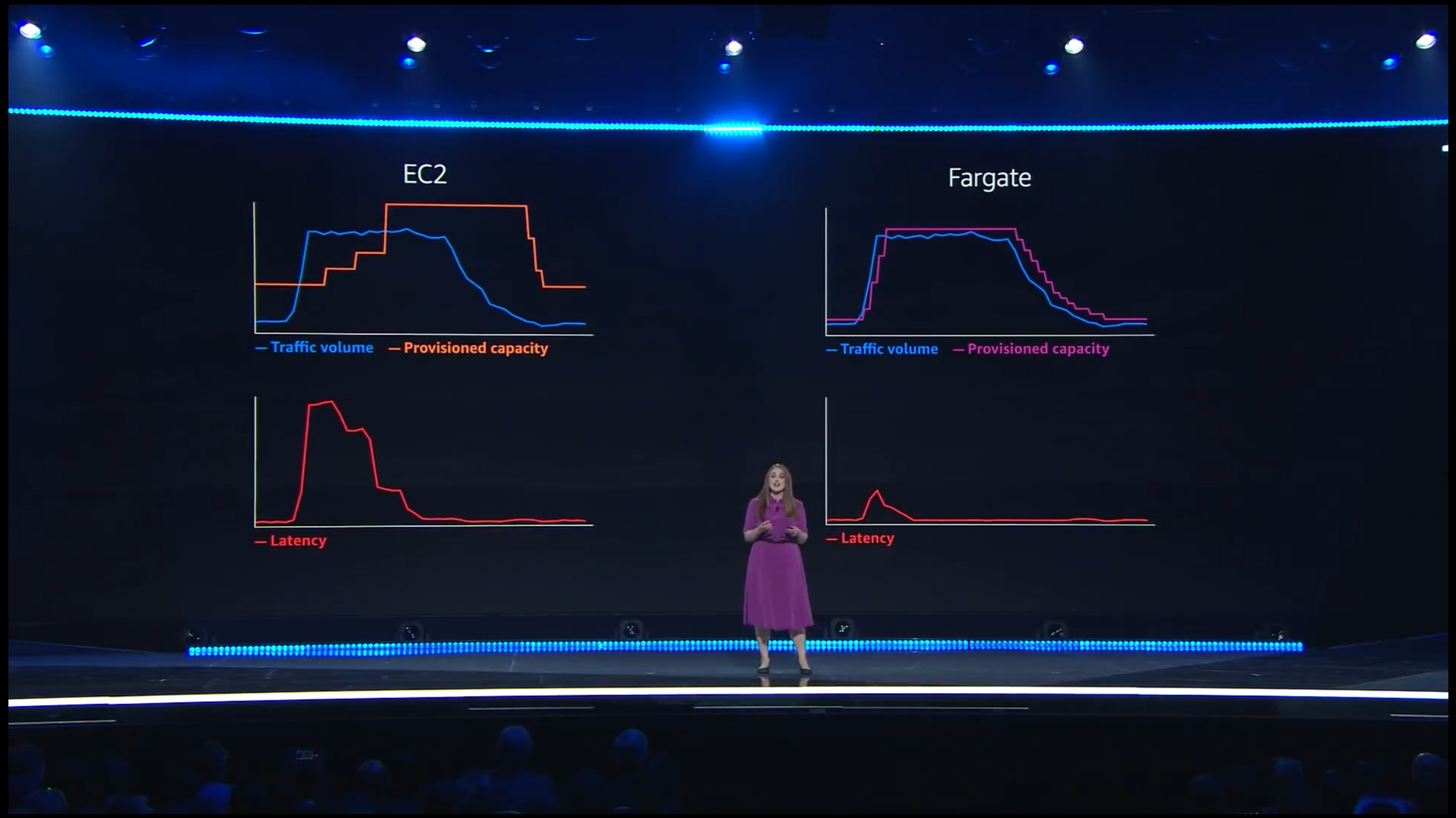


- Scaling delay

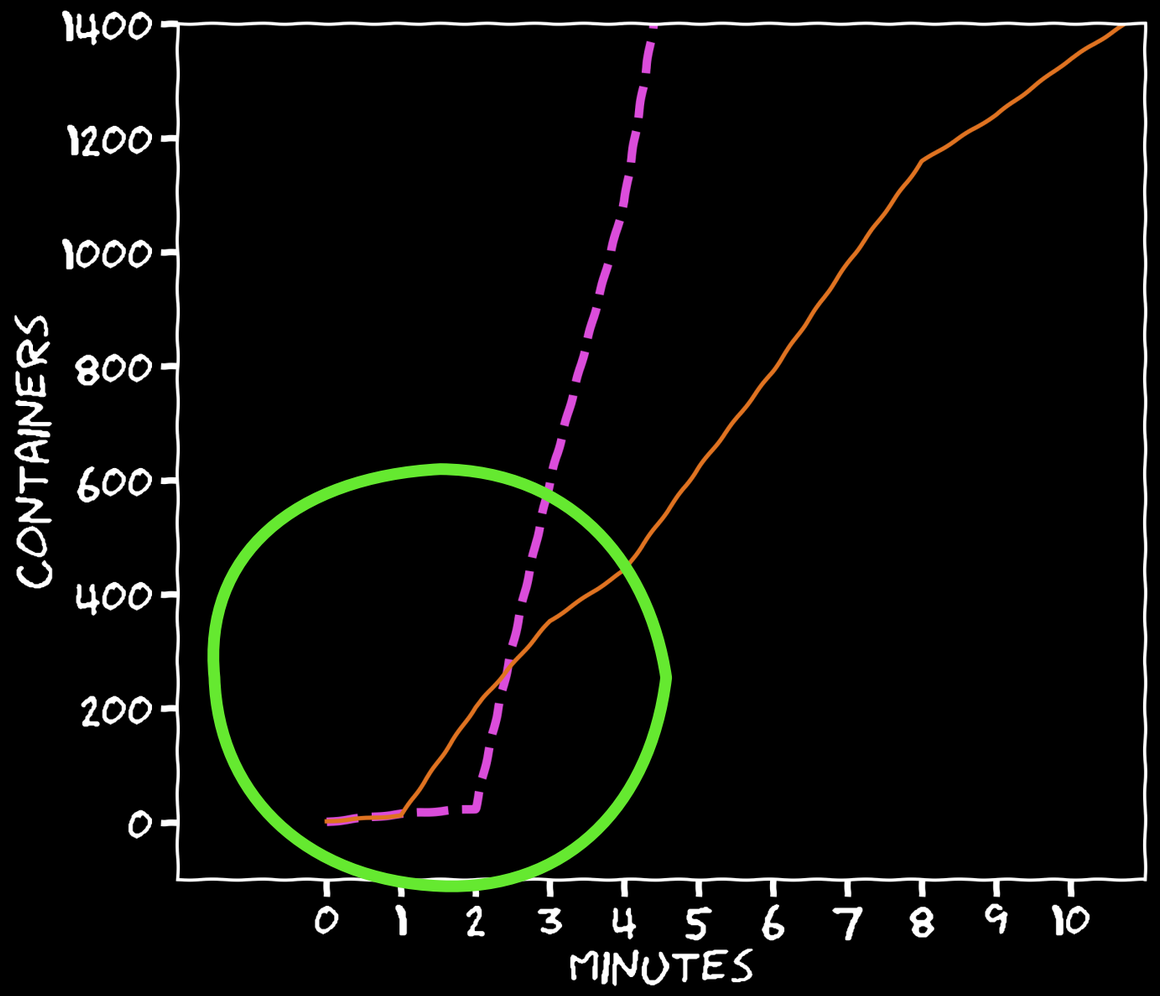


Clare Liguori at re:Invent 2019

https://youtu.be/OdzaTbaQwTg?t=1262

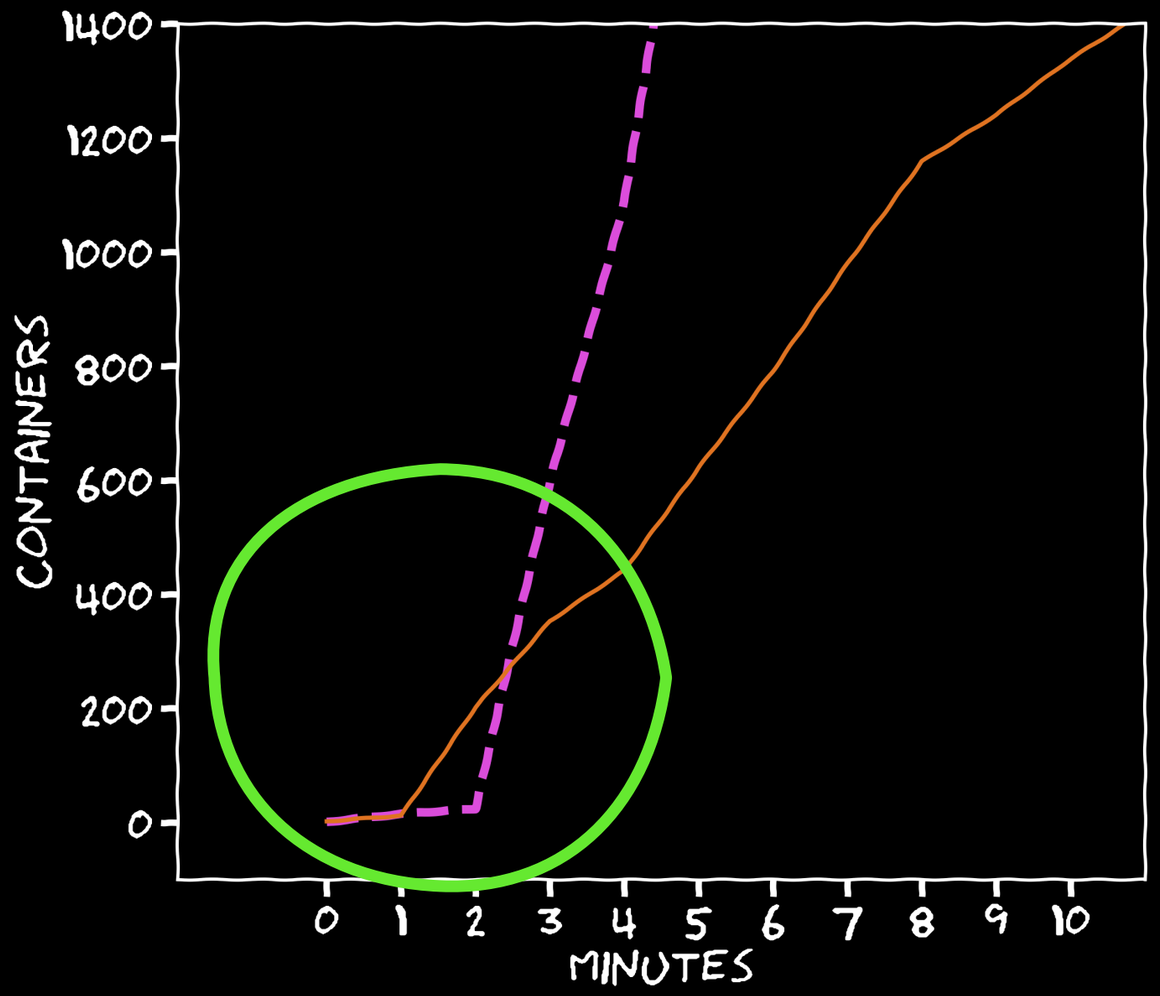


Amazon EC2 compute



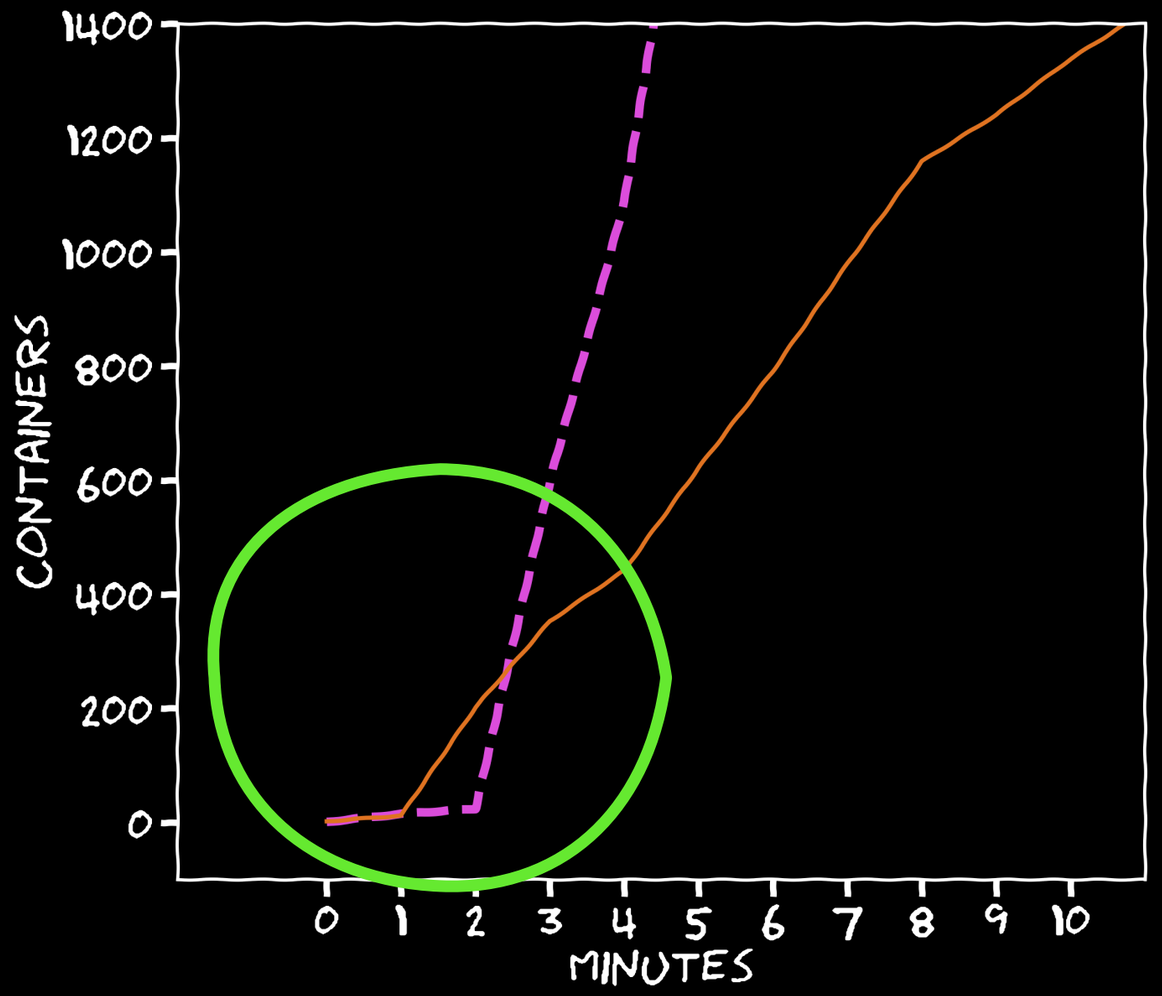
- Scaling delay
  - Web applications? X
  - User-facing applications? X
  - Latency sensitive? X
  - What is a fit?

Amazon EC2 compute



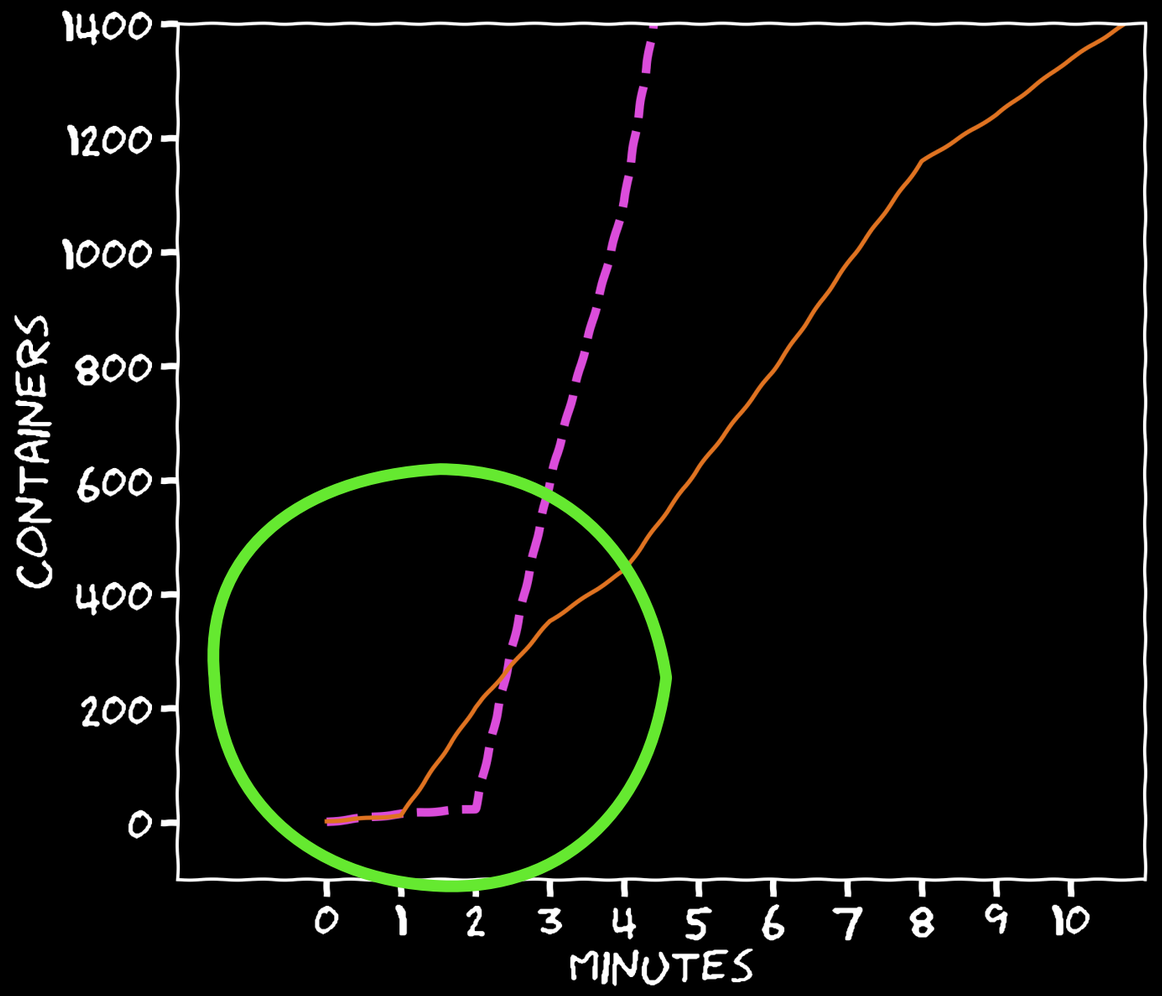
- Scaling delay
  - Web applications? ✗
  - User-facing applications? ✗
  - Latency sensitive? ✗
  - Background work?

Amazon EC2 compute



- Scaling delay
  - Web applications? ✗
  - User-facing applications? ✗
  - Latency sensitive? ✗
  - Background work? ✓

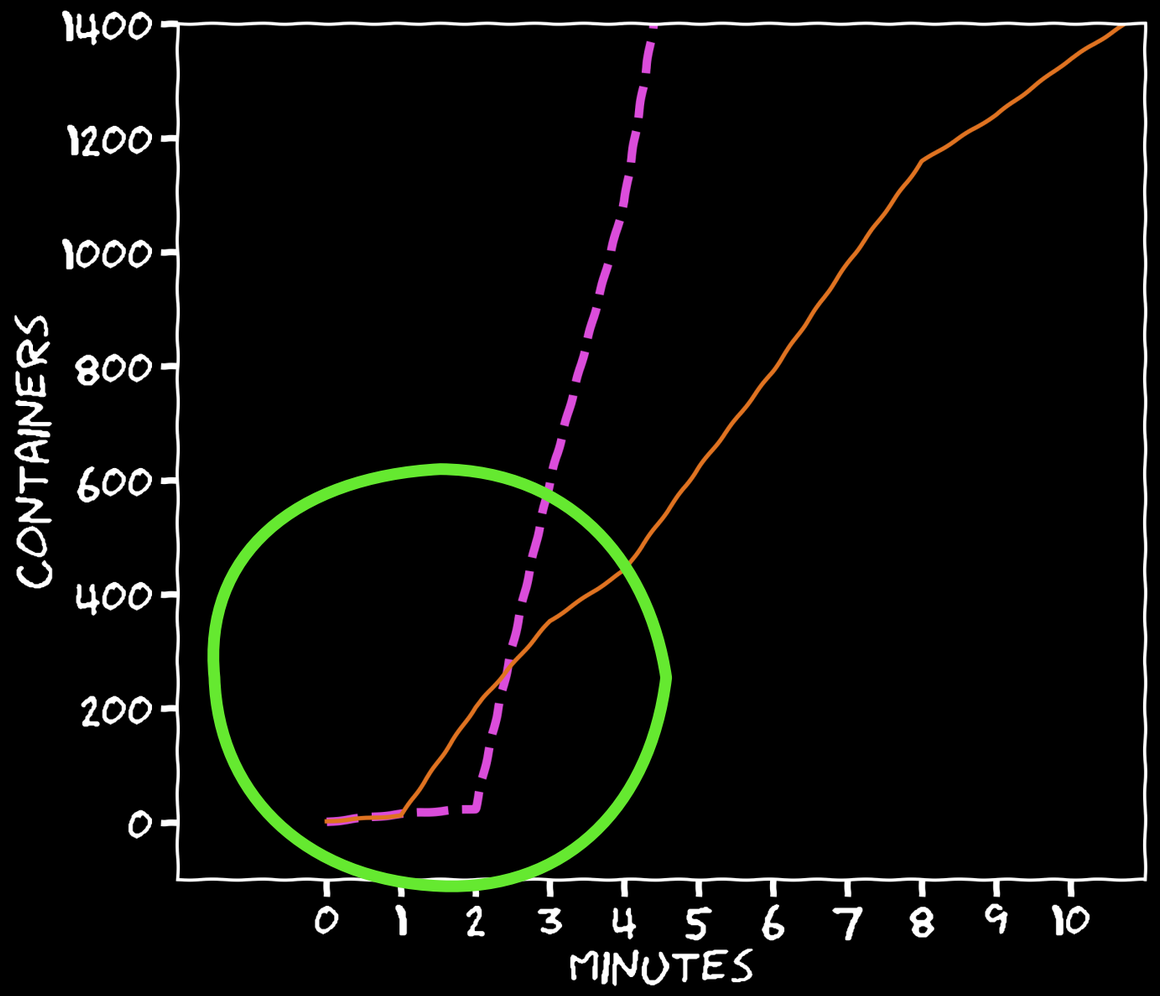
Amazon EC2 compute



- Scaling delay
  - Web applications? ✗
  - User-facing applications? ✗
  - Latency sensitive? ✗

  - Background work? ✓
  - Network-bound services?

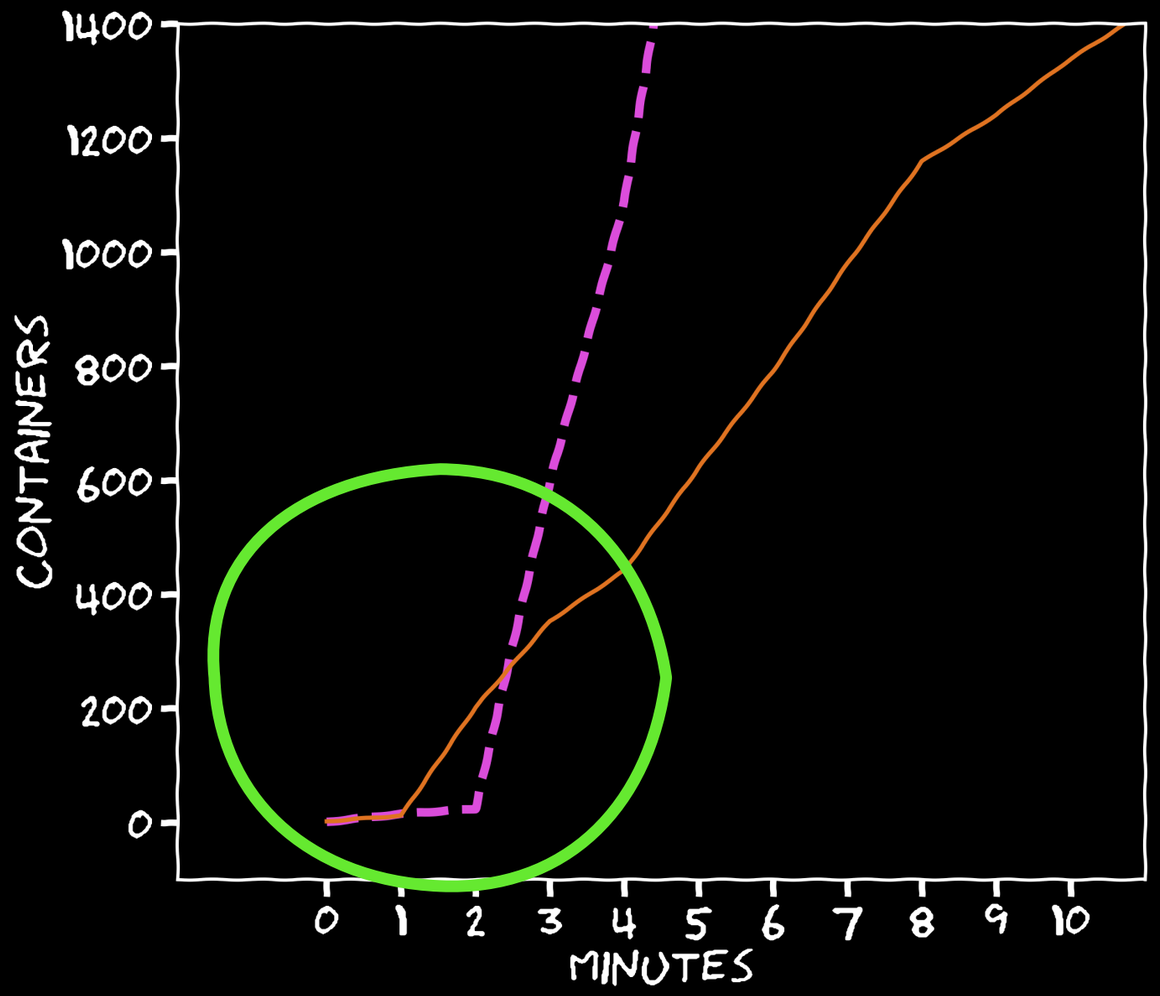
Amazon EC2 compute



- Scaling delay
  - Web applications? ✗
  - User-facing applications? ✗
  - Latency sensitive? ✗

  - Background work? ✓
  - Network-bound services? ✓

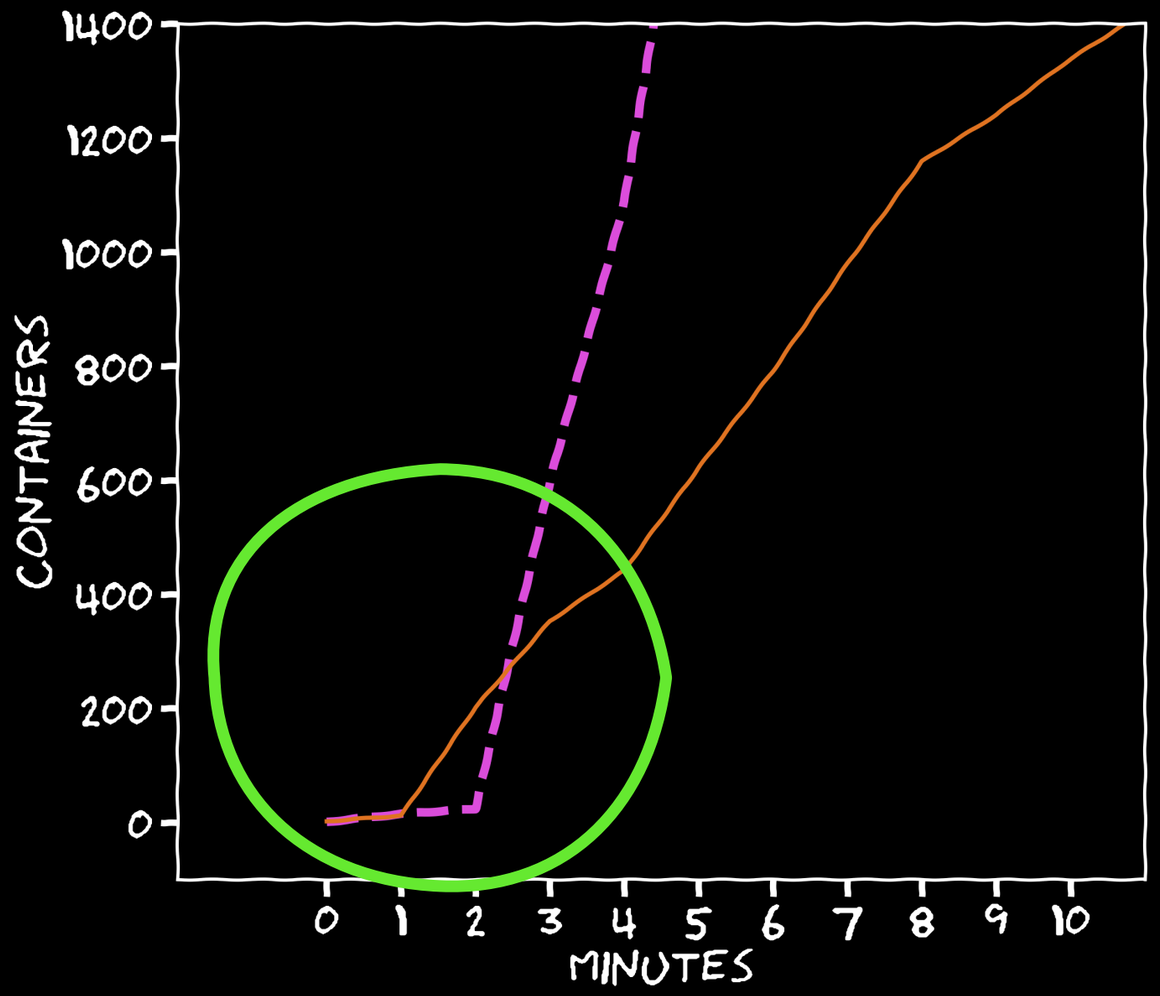
Amazon EC2 compute



- Scaling delay
  - Web applications? ✗
  - User-facing applications? ✗
  - Latency sensitive? ✗

  - Background work? ✓
  - Network-bound services? ✓
  - GPU-accelerated jobs?

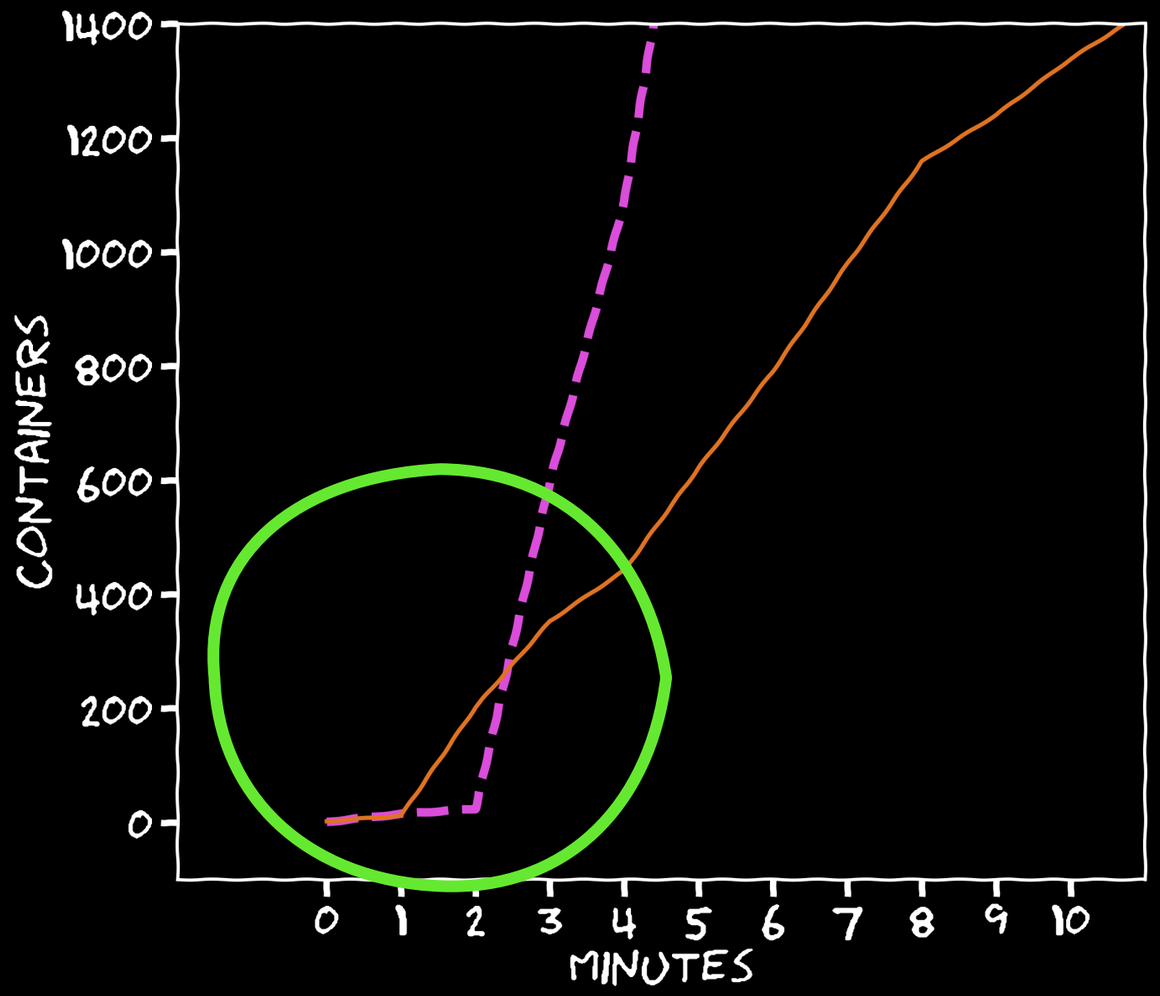
Amazon EC2 compute



- Scaling delay
  - Web applications? ✗
  - User-facing applications? ✗
  - Latency sensitive? ✗

  - Background work? ✓
  - Network-bound services? ✓
  - GPU-accelerated jobs? ✓

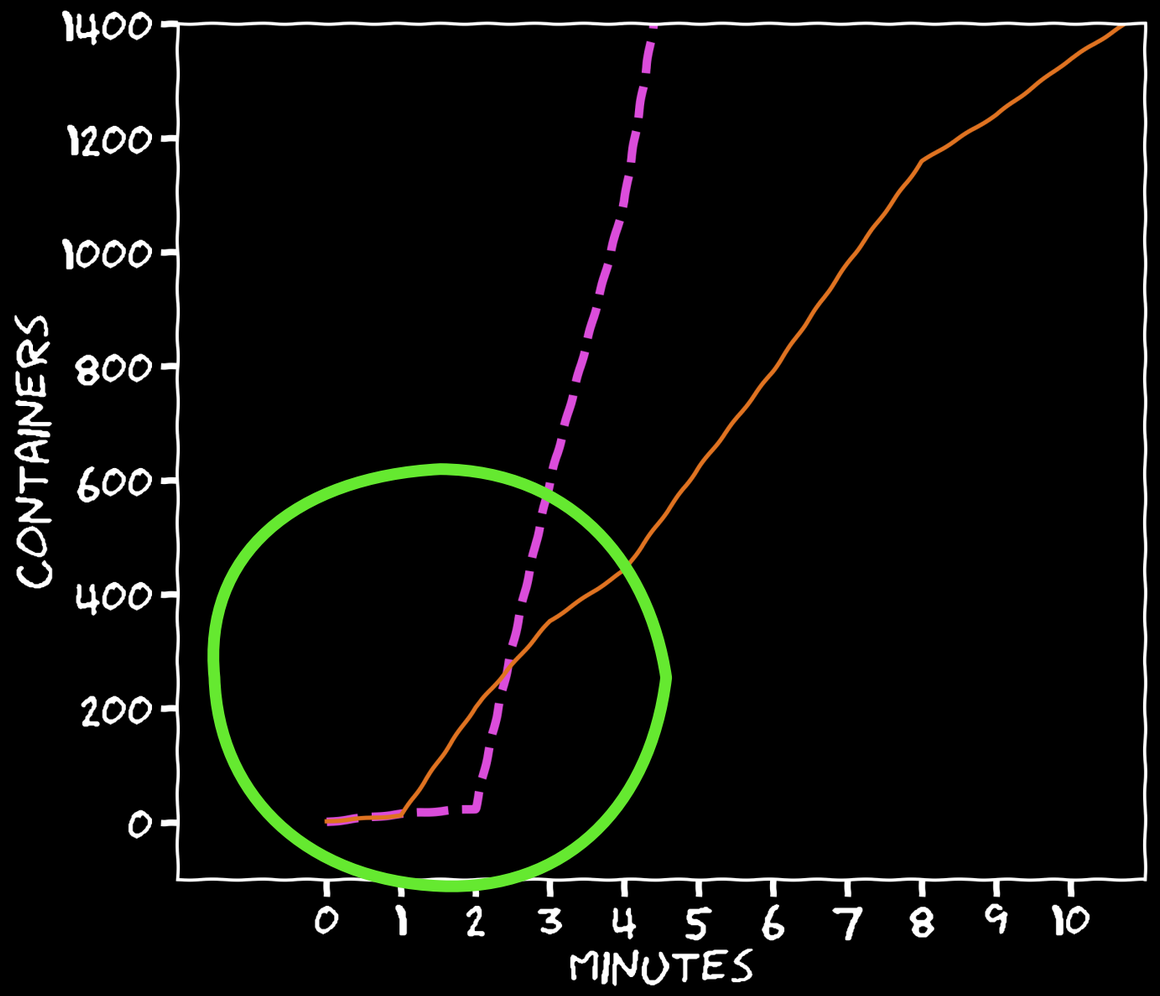
Amazon EC2 compute



- Scaling delay
  - Web applications? ✗
  - User-facing applications? ✗
  - Latency sensitive? ✗

  - Background work? ✓
  - Network-bound services? ✓
  - GPU-accelerated jobs? ✓
  - ARM workloads?

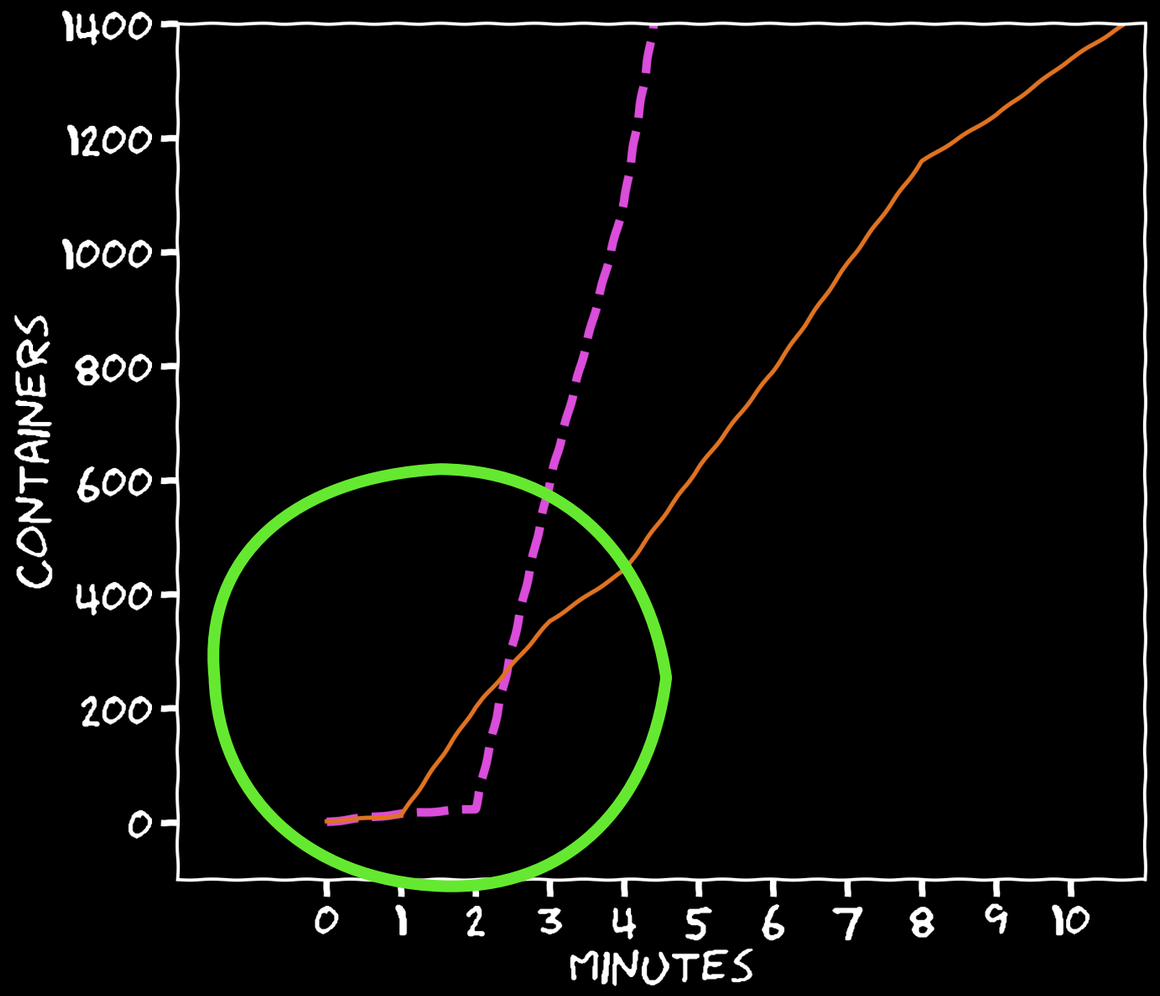
Amazon EC2 compute



- Scaling delay
  - Web applications? ✗
  - User-facing applications? ✗
  - Latency sensitive? ✗

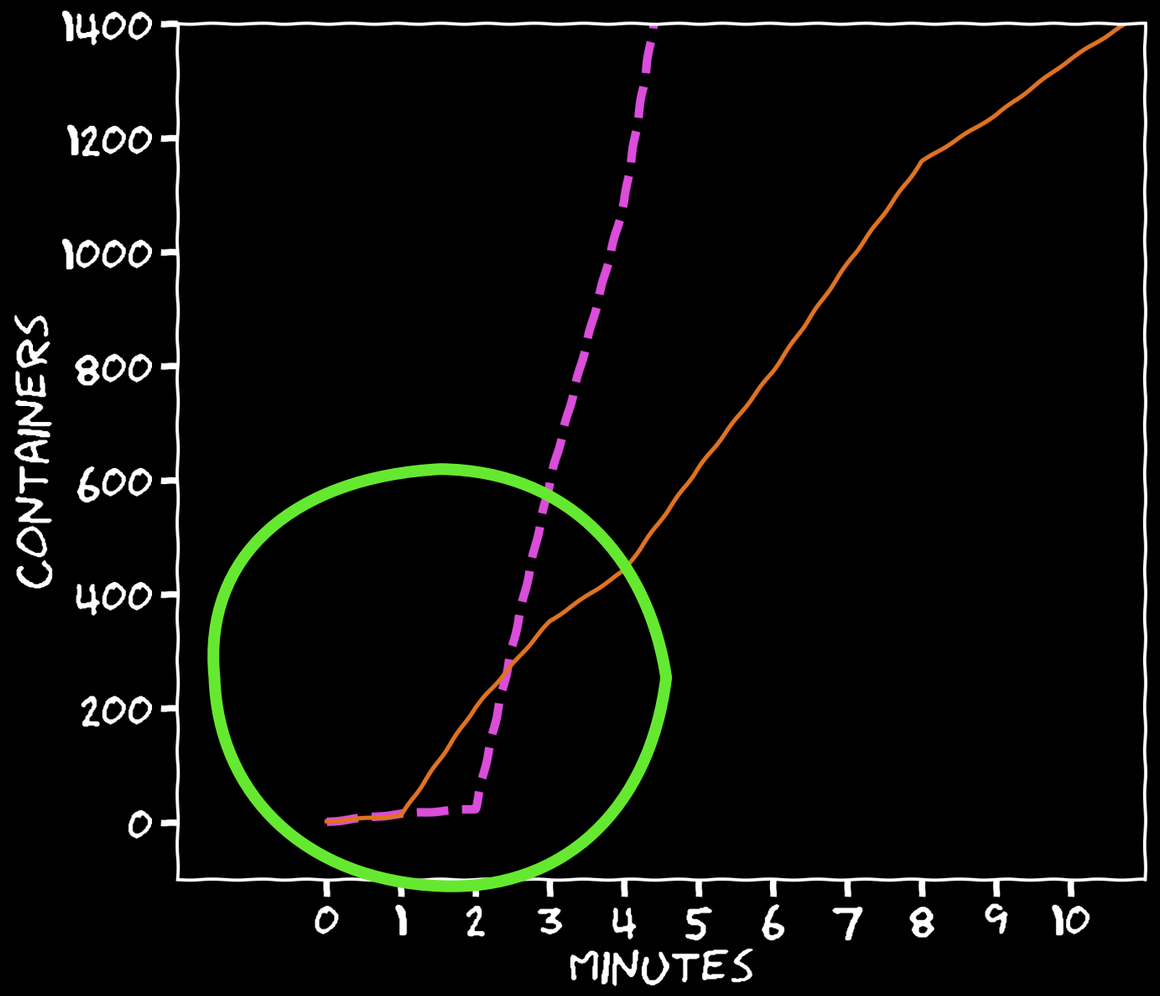
  - Background work? ✓
  - Network-bound services? ✓
  - GPU-accelerated jobs? ✓
  - ARM workloads? ✓

Amazon EC2 compute



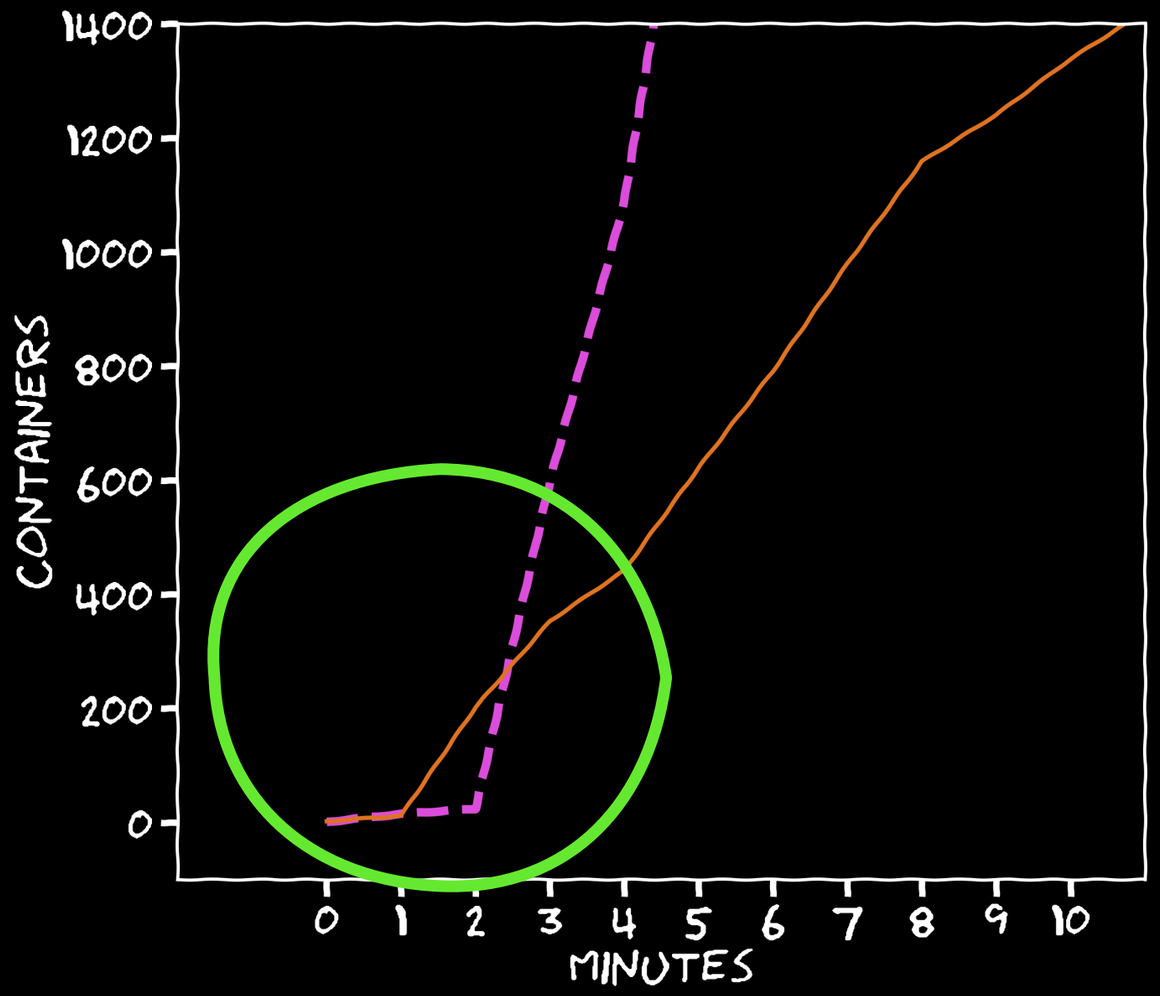
- Workarounds

Amazon EC2 compute



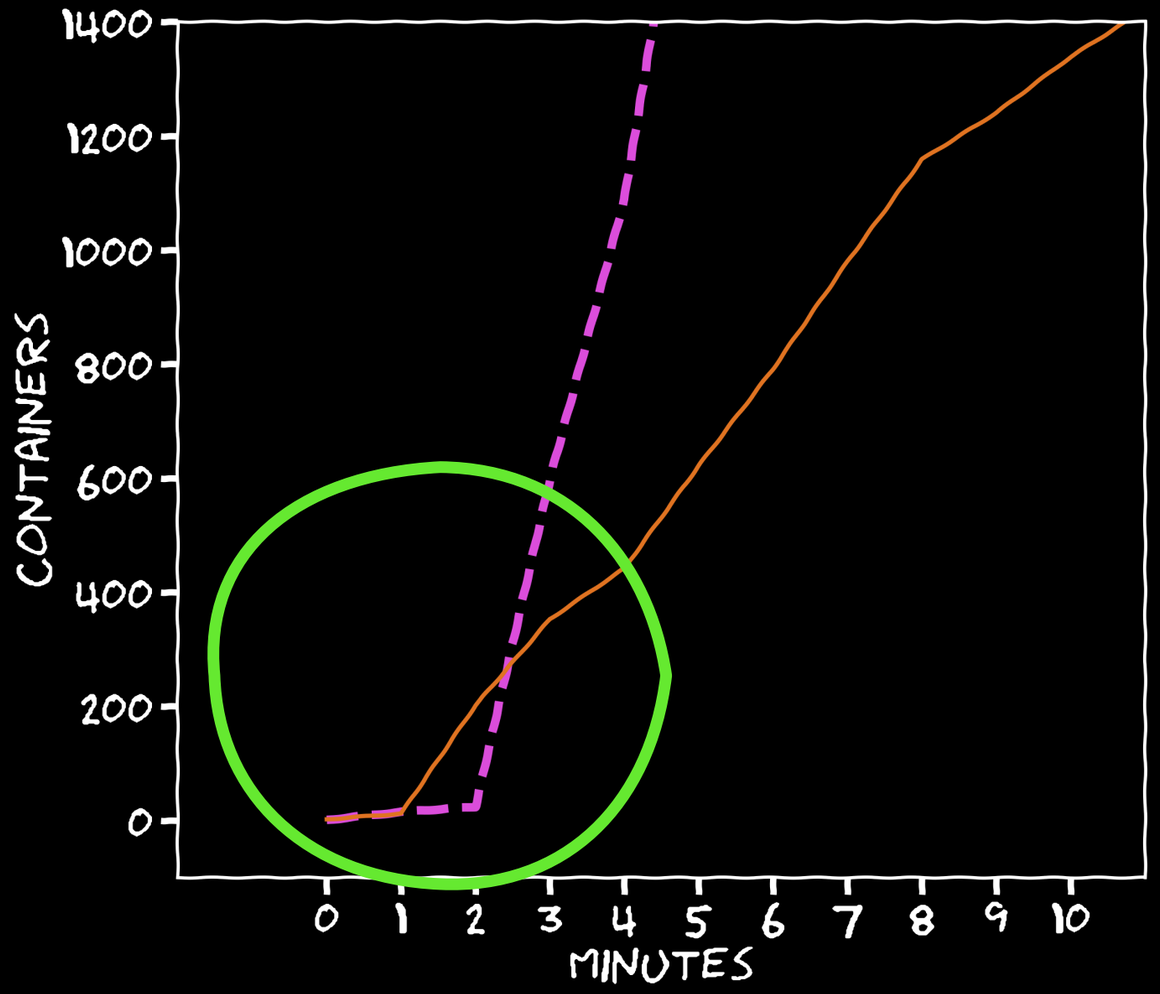
- Workarounds
  - Configure aggressive scaling

Amazon EC2 compute



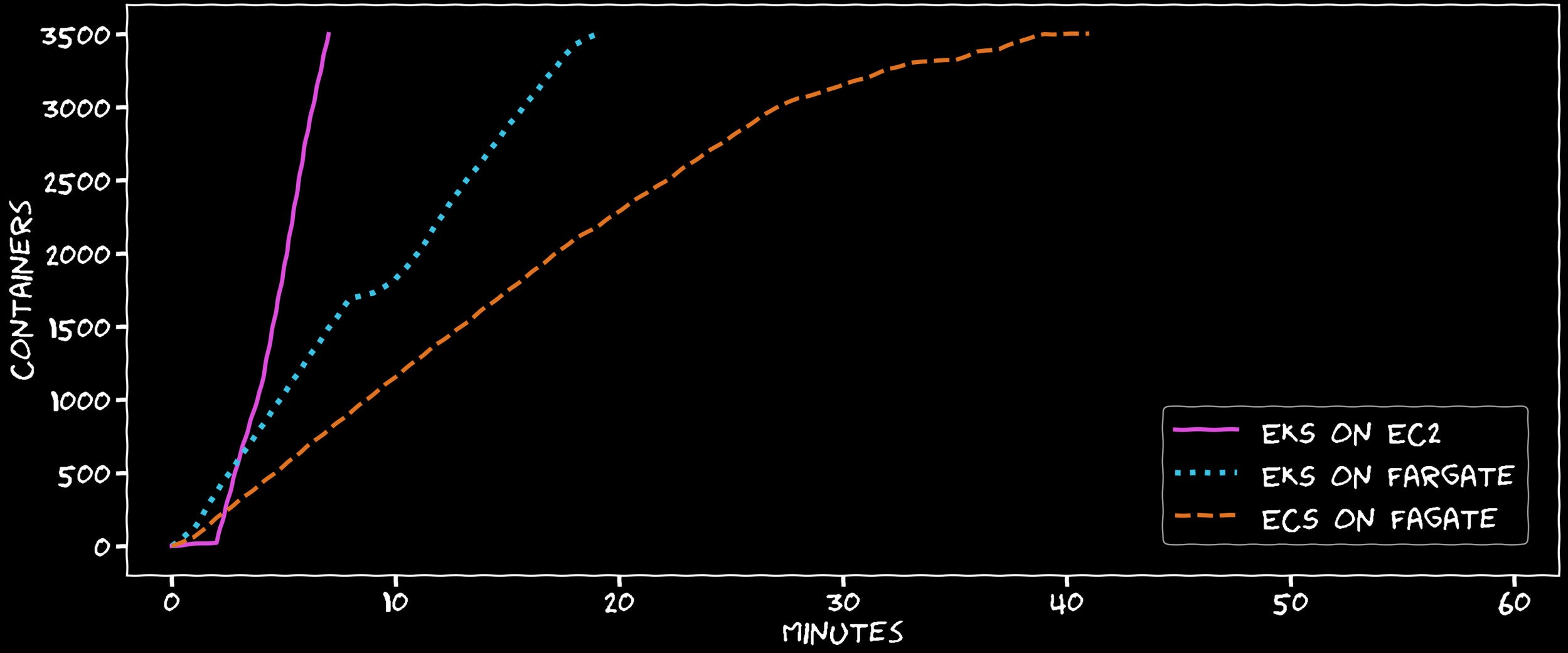
- Workarounds
  - Configure aggressive scaling
  - Prewarm and prepare capacity

Amazon EC2 compute

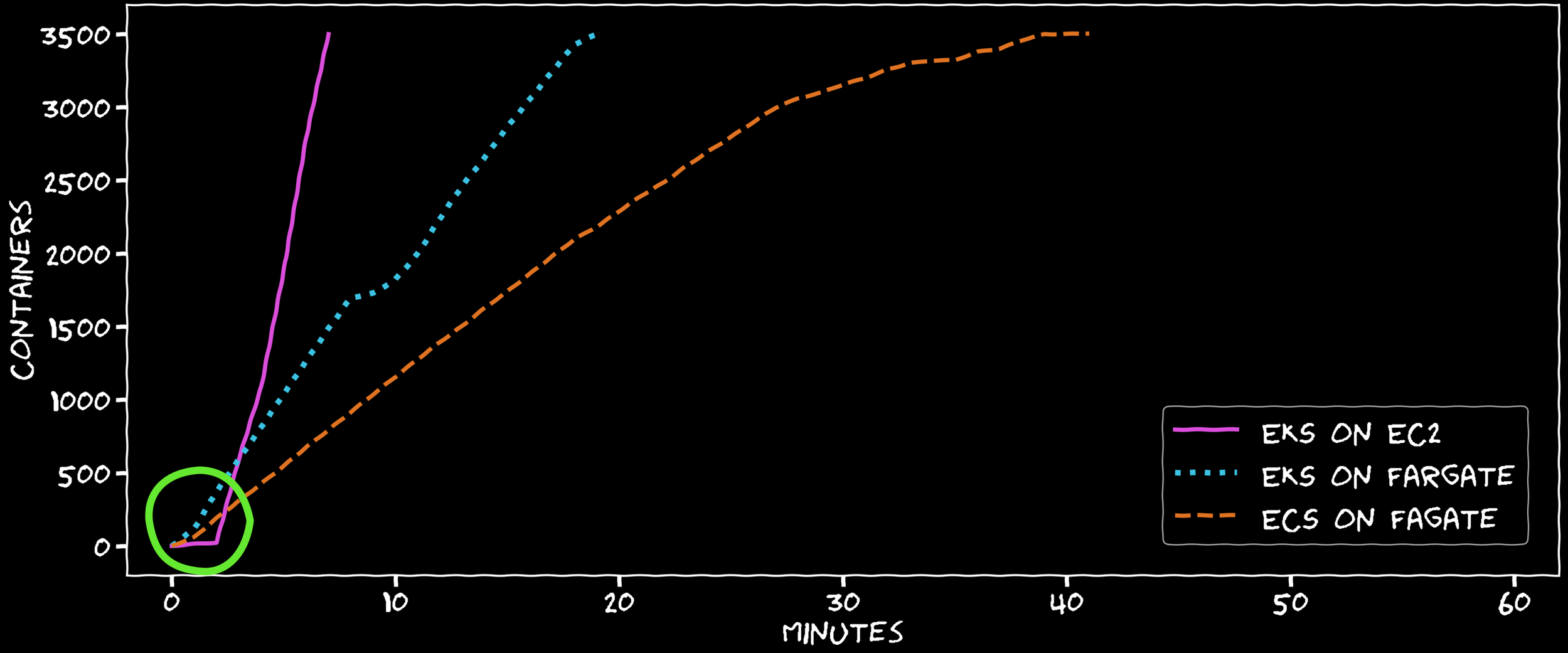


- Workarounds
  - Extra cost

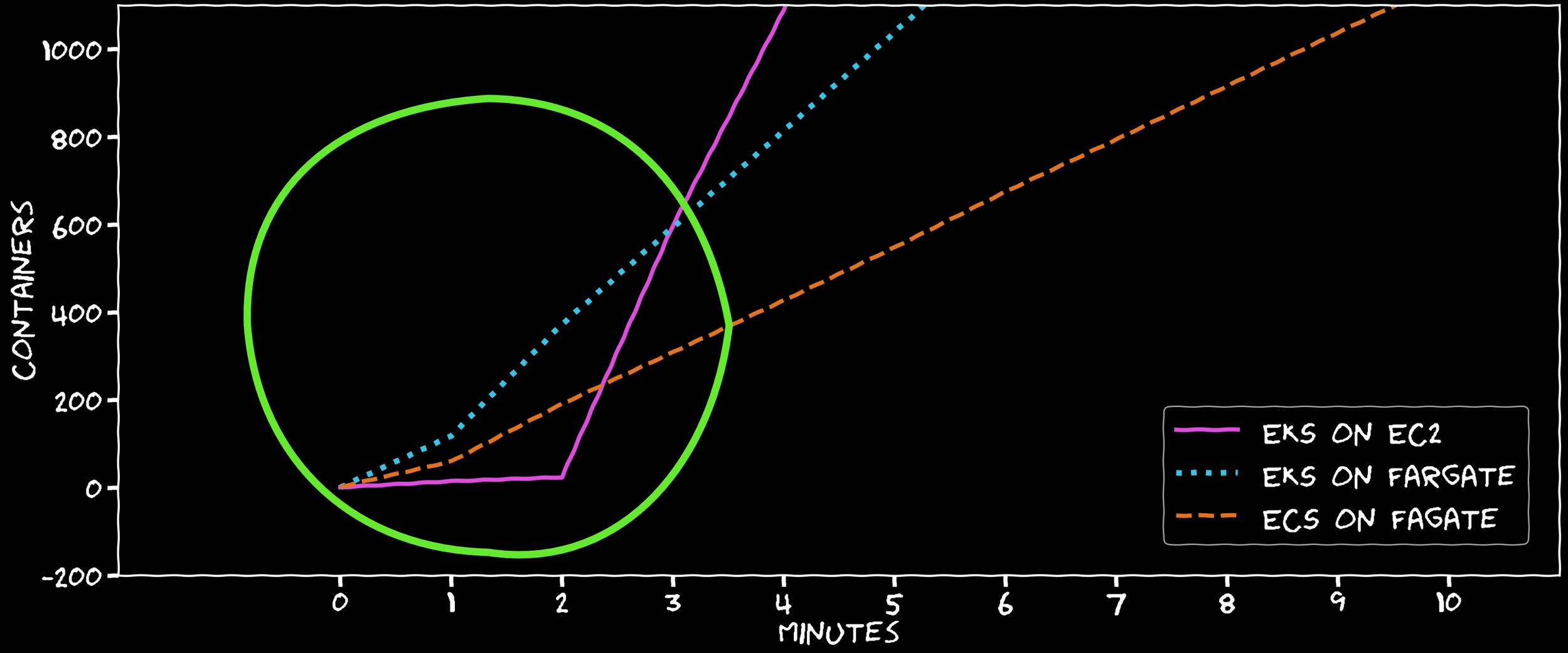
Scaling test results – winter



Scaling test results – winter



Scaling test results – winter



AWS Fargate workers



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

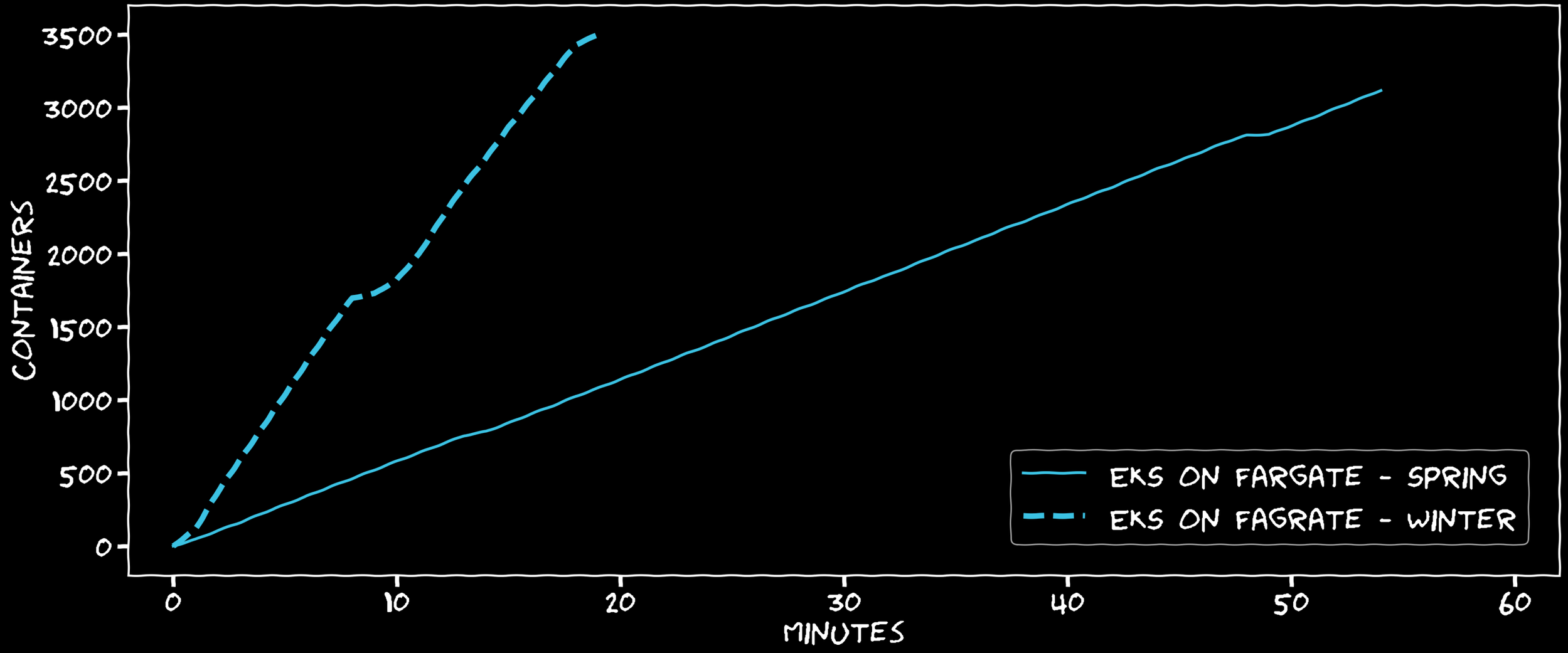
Amazon EKS on AWS Fargate



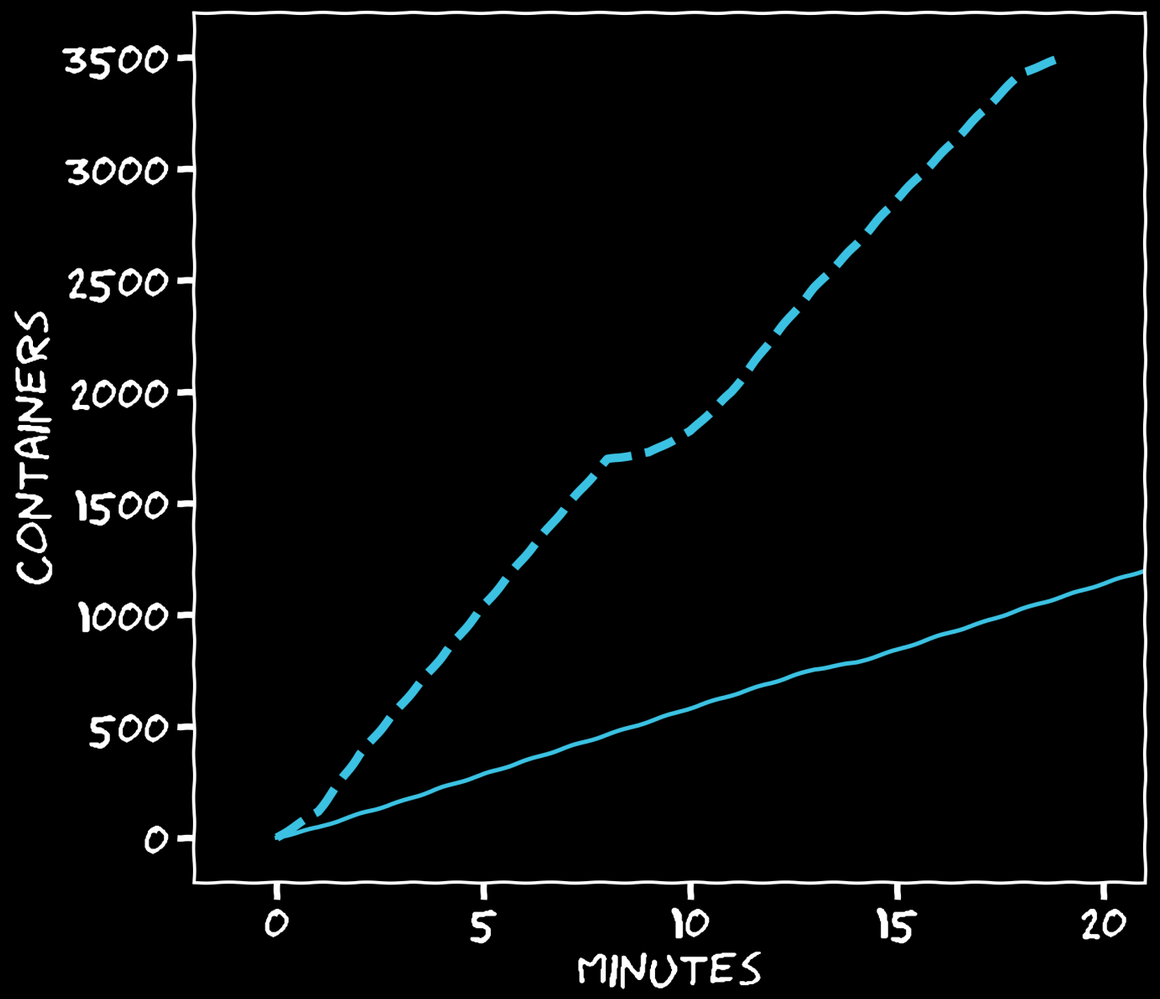
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EKS on AWS Fargate

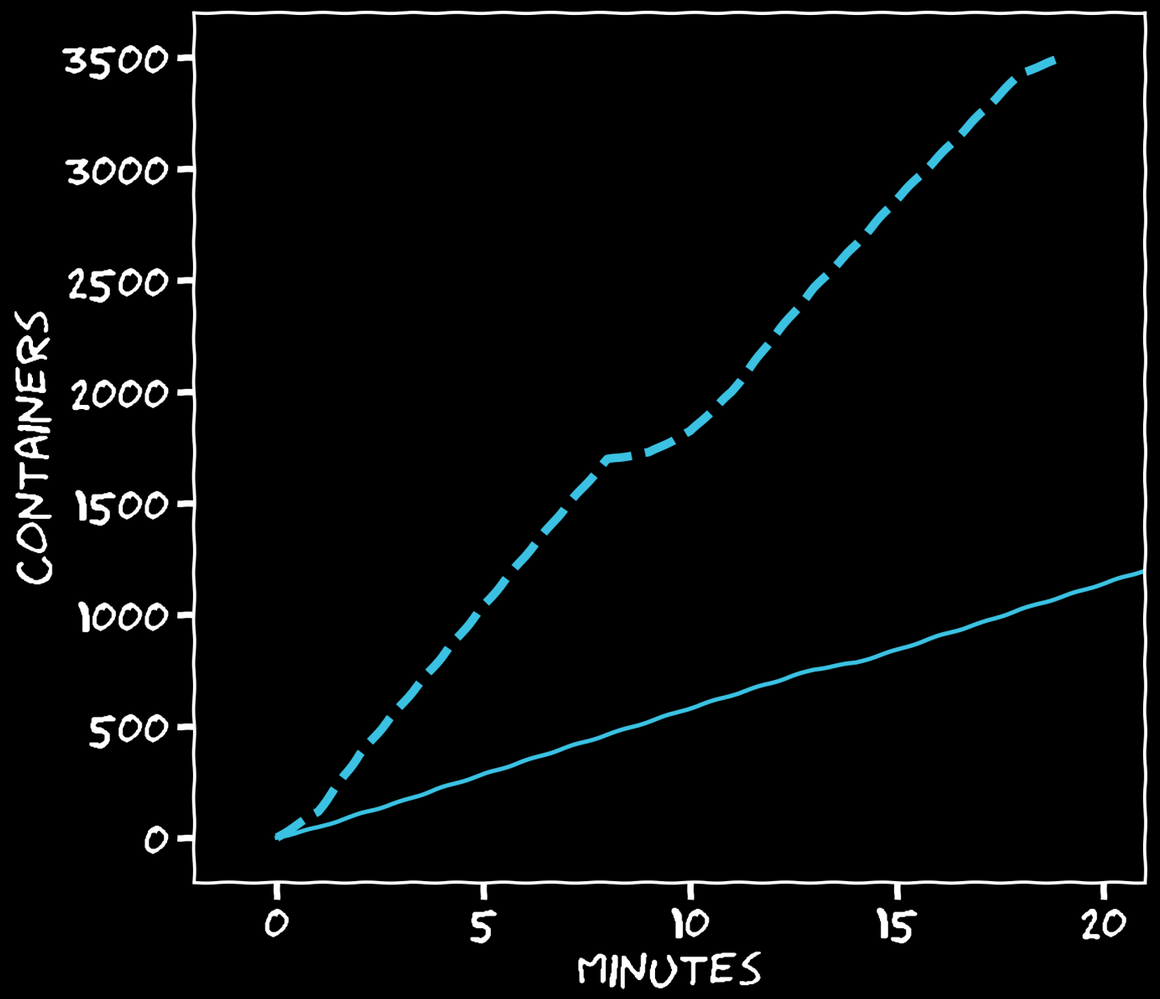


Amazon EKS on AWS Fargate



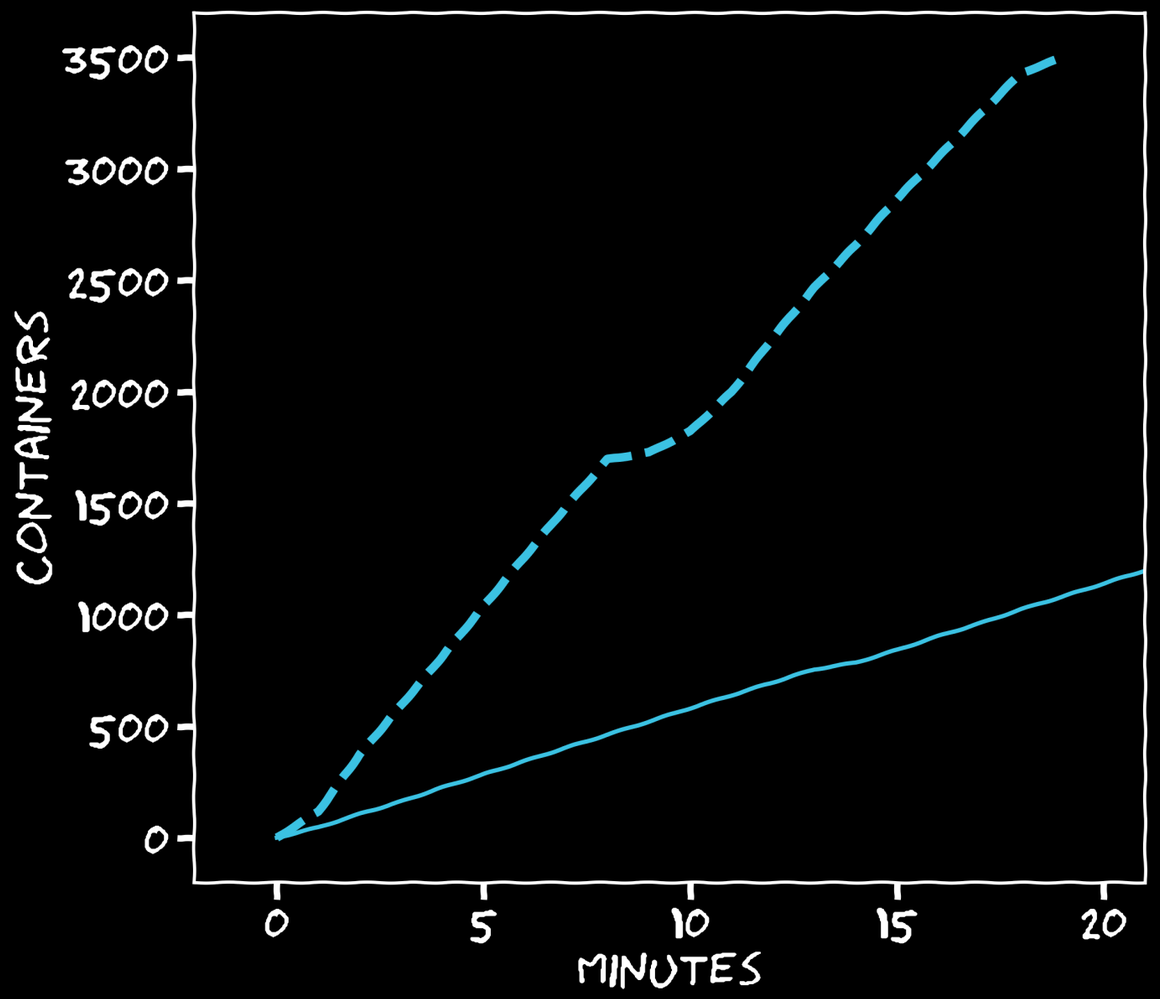
- Amazon EKS control plane prescaled

Amazon EKS on AWS Fargate



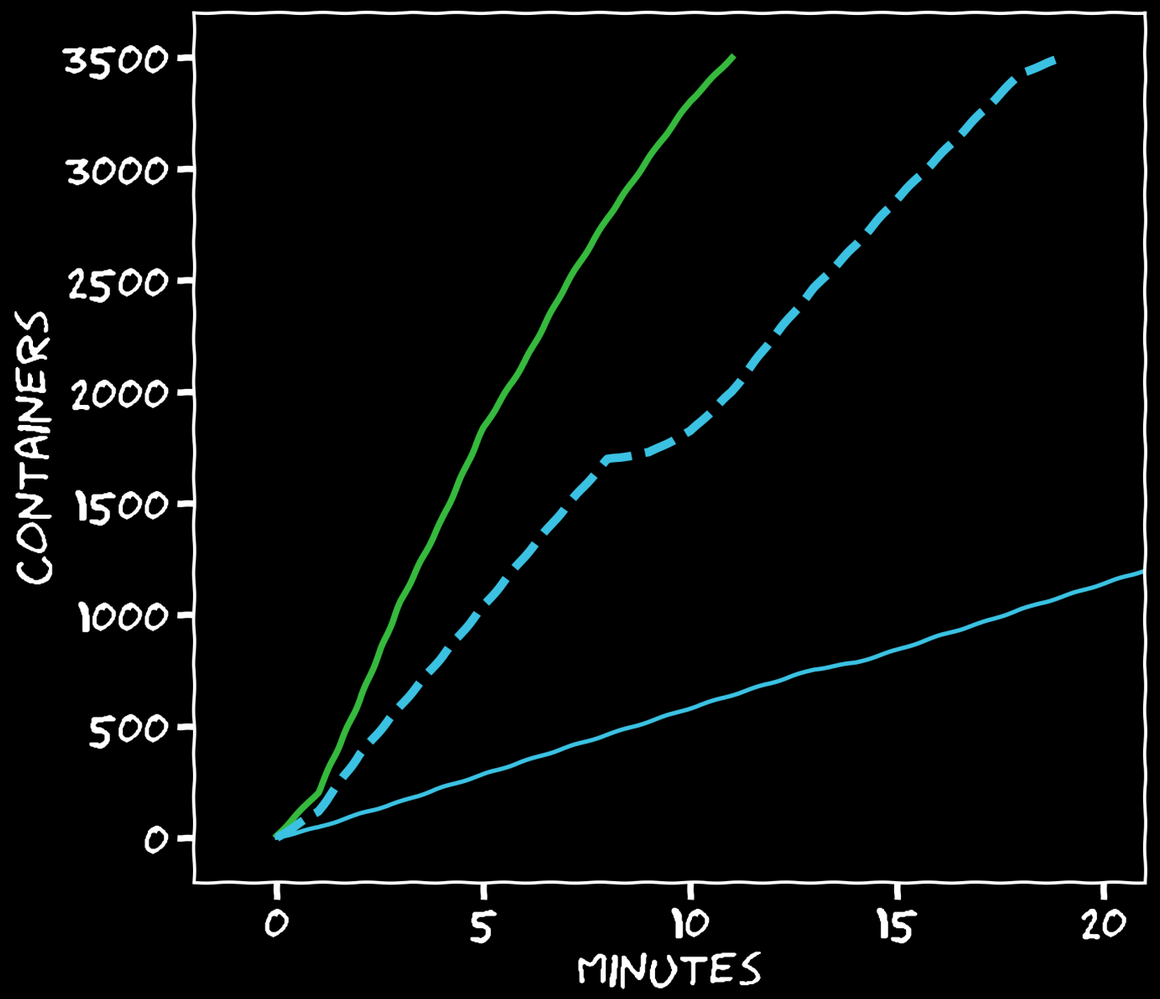
- Amazon EKS control plane prescaled
- Limit changed

Amazon EKS on AWS Fargate



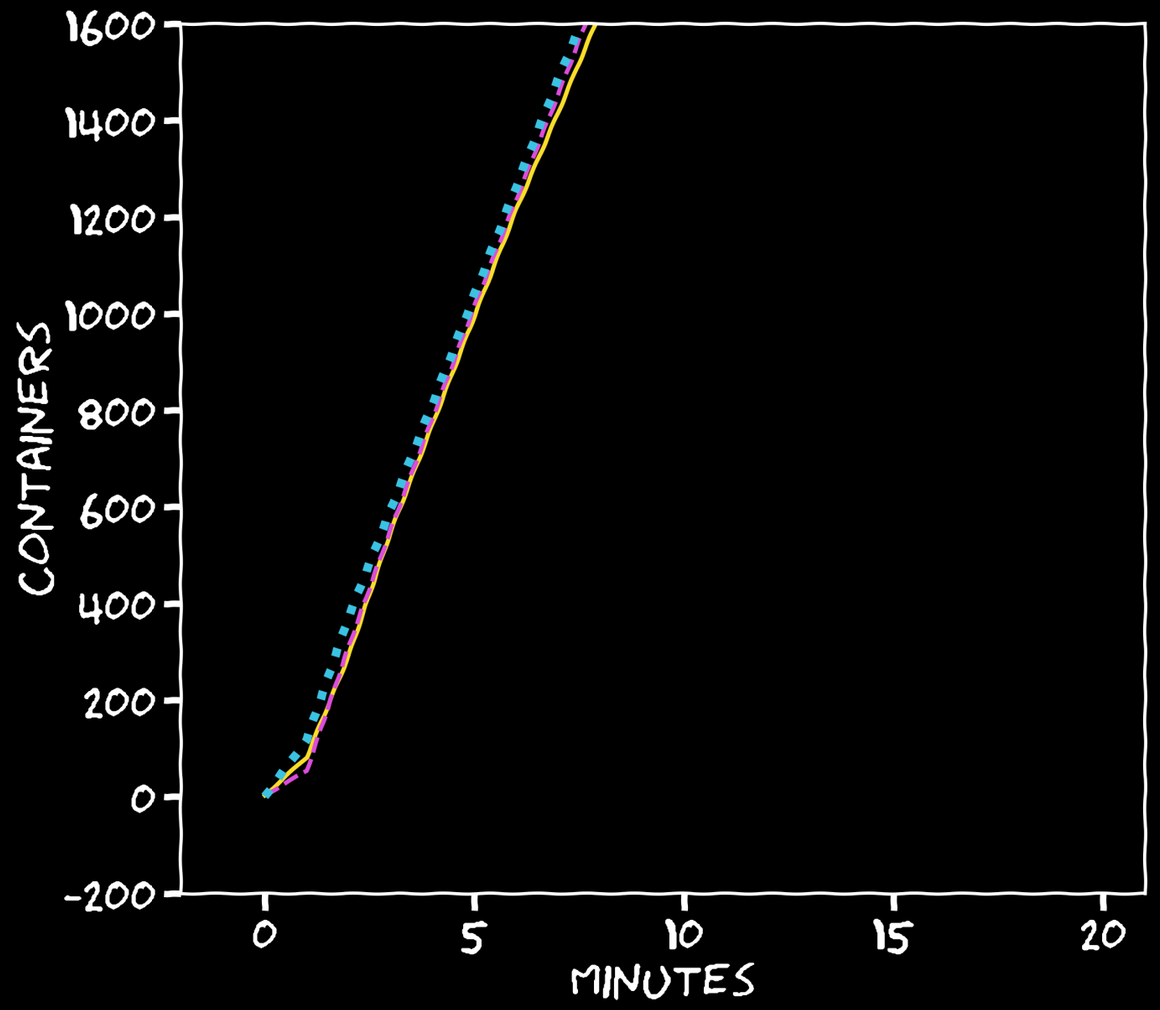
- Amazon EKS control plane prescaled
- Limit changed
  - 10 TPS per AWS account

Amazon EKS on AWS Fargate



- Amazon EKS control plane prescaled
- Limit changed
  - 10 TPS per AWS account
- Can we do better?

Amazon EKS on AWS Fargate



- Amazon EKS control plane prescaled
- Limit changed
  - 10 TPS per AWS account
- Can we do better?
  - Not right now

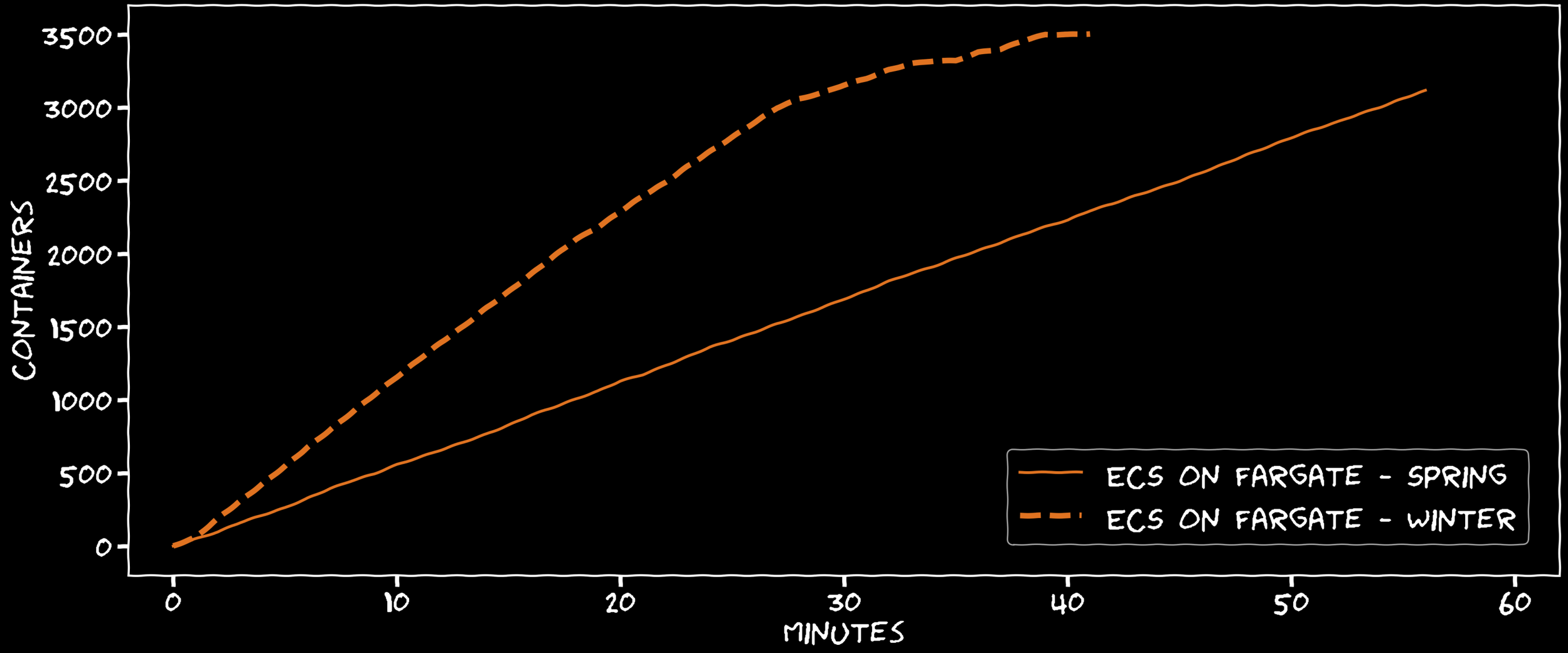
Amazon ECS on AWS Fargate



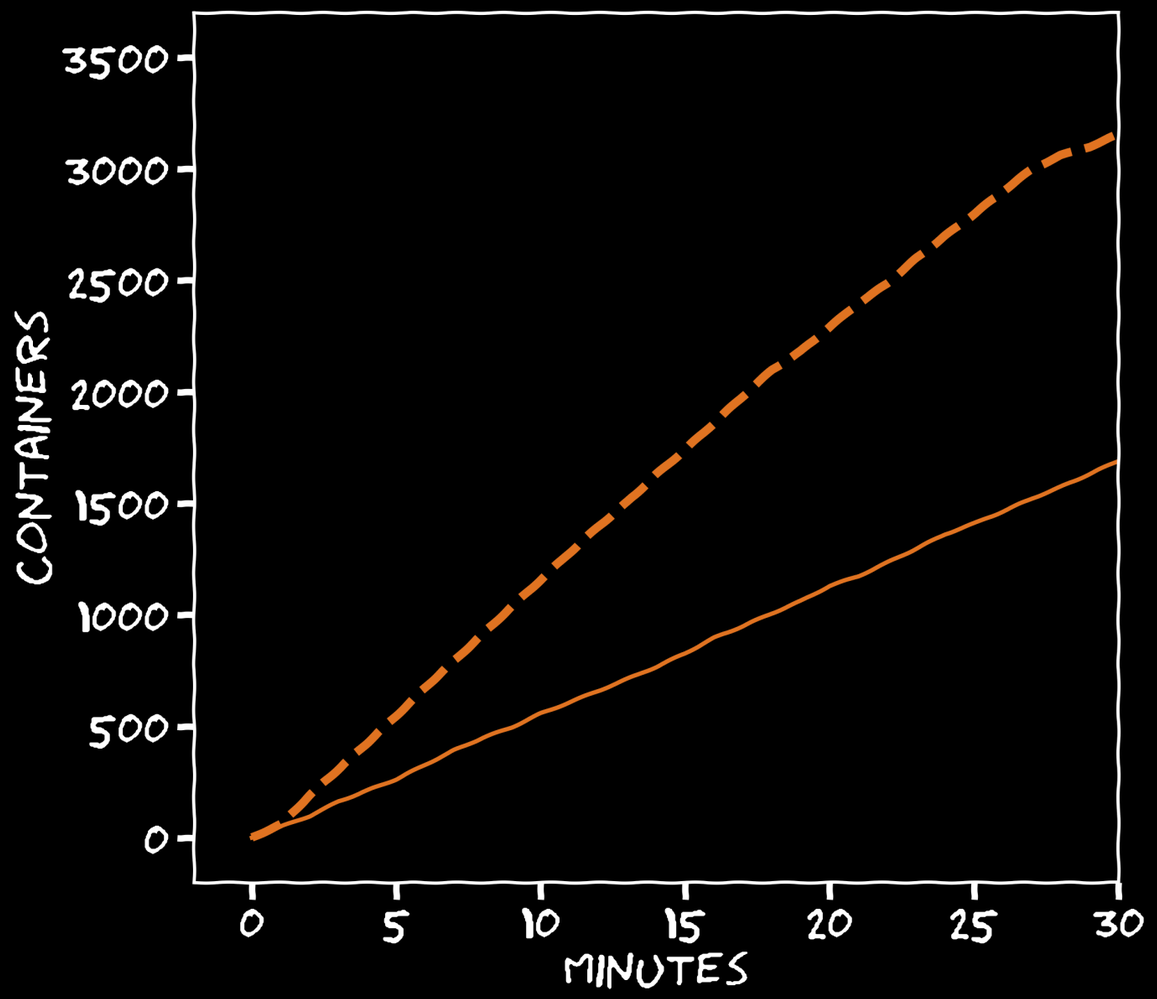
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon ECS on AWS Fargate

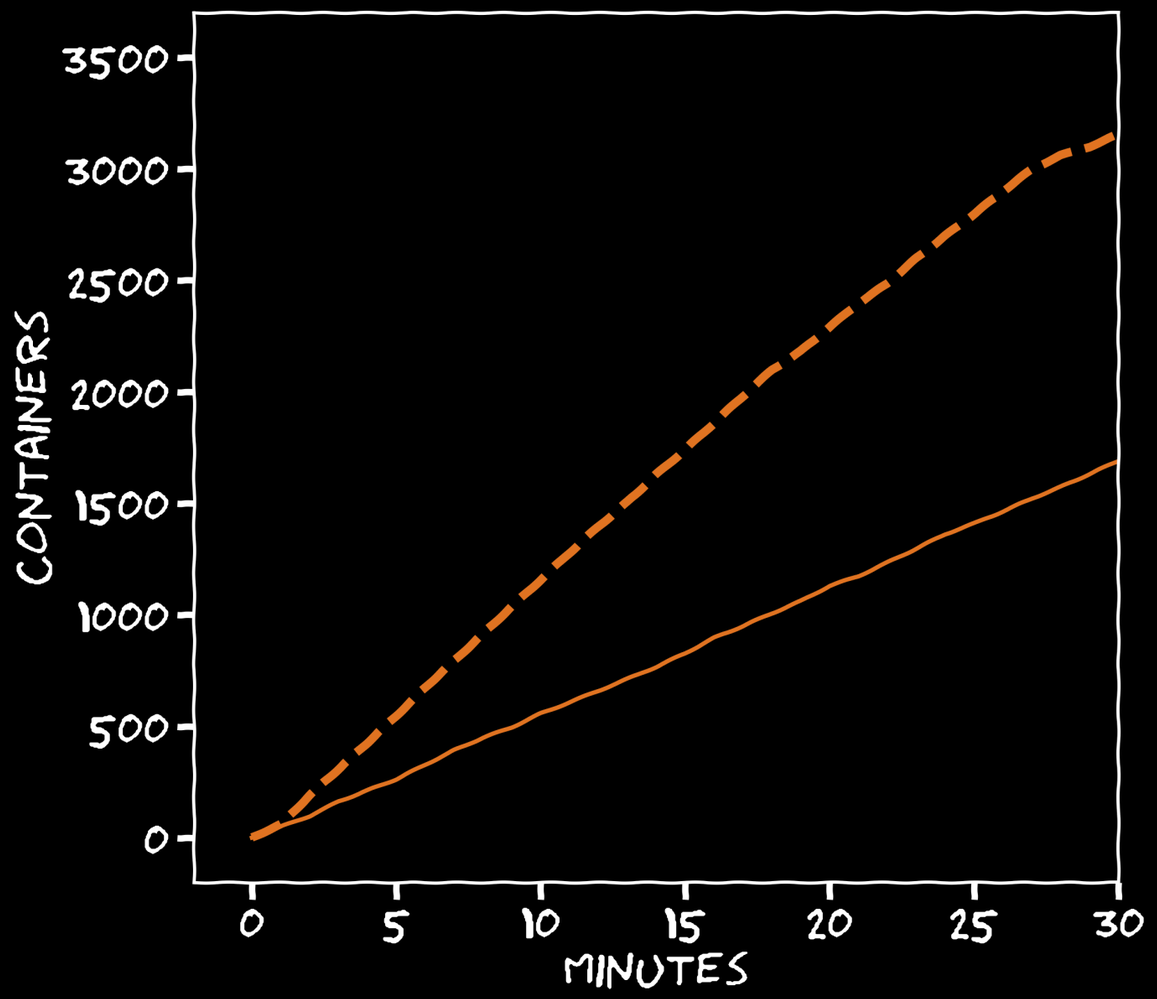


Amazon ECS on AWS Fargate



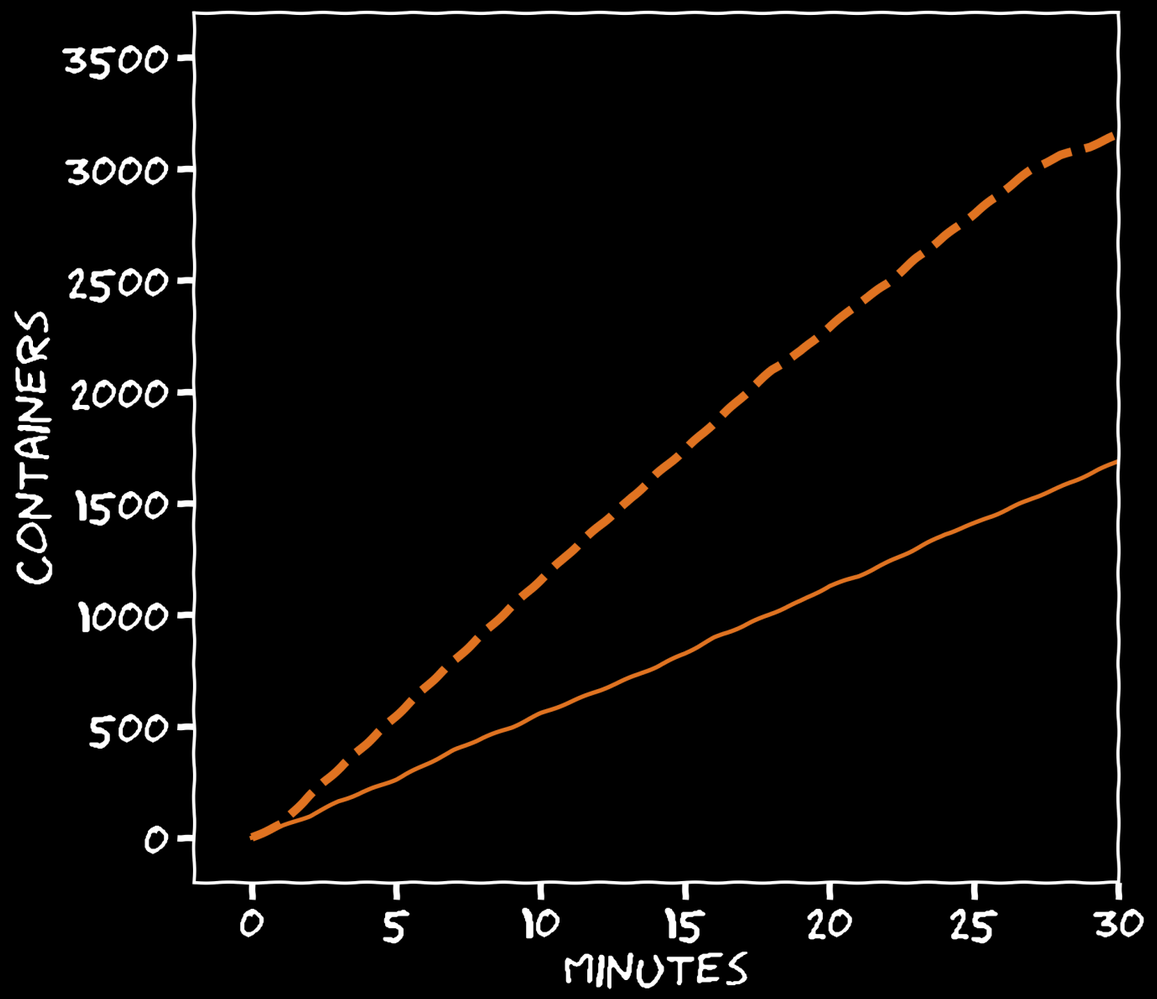
- Limits changed

Amazon ECS on AWS Fargate



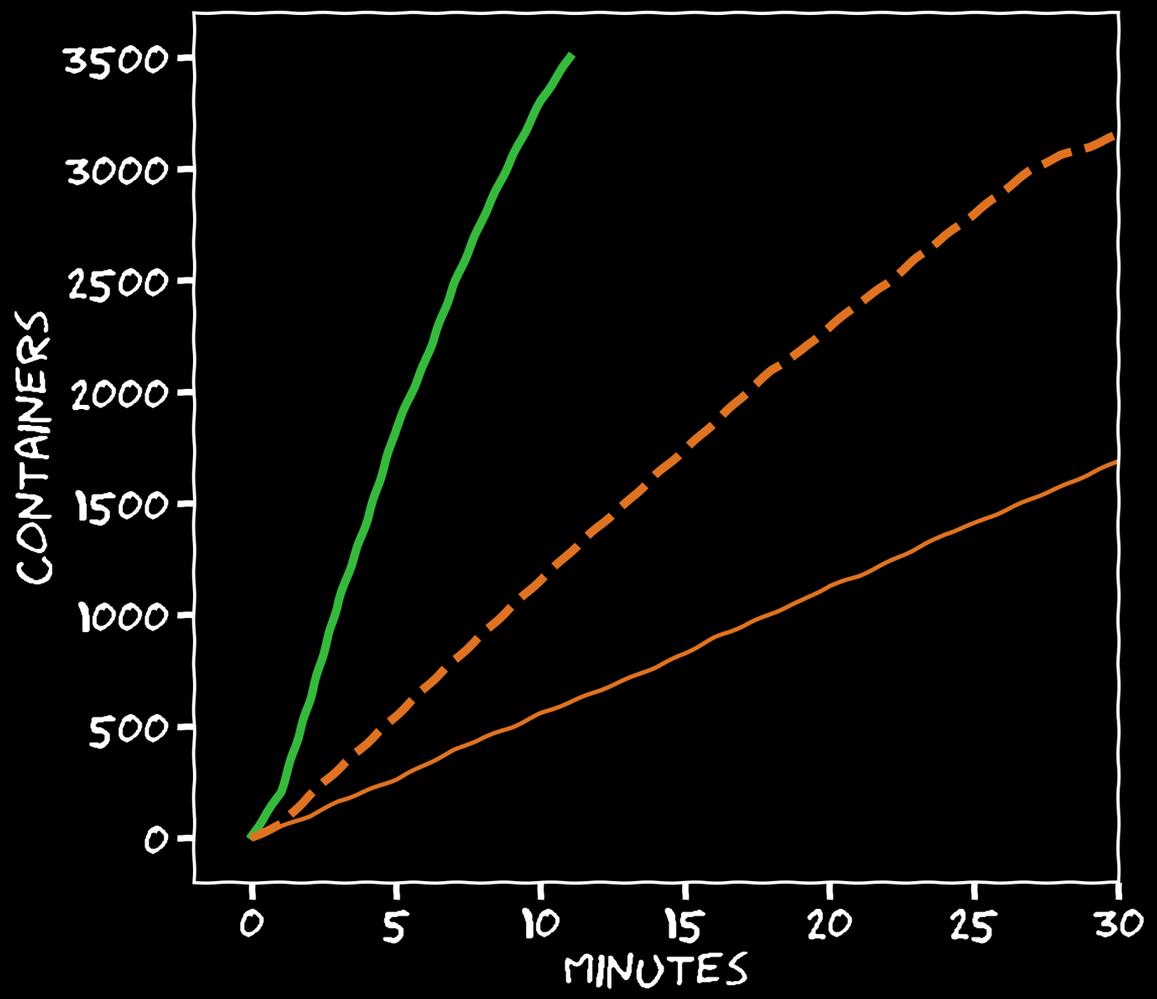
- Limits changed
  - 10 TPS per AWS account

Amazon ECS on AWS Fargate



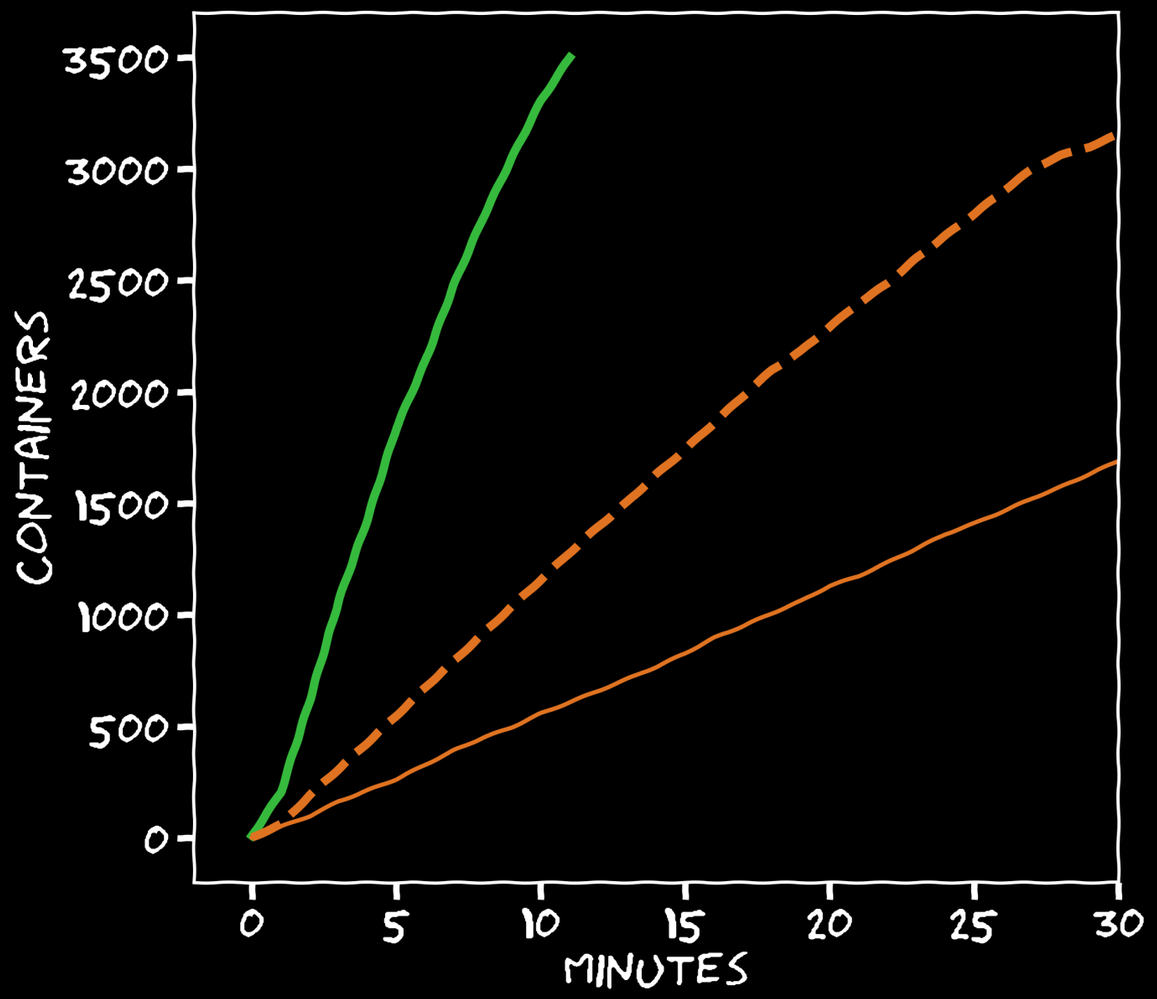
- Limits changed
  - 10 TPS per AWS account
  - 2 TPS per service

Amazon ECS on AWS Fargate



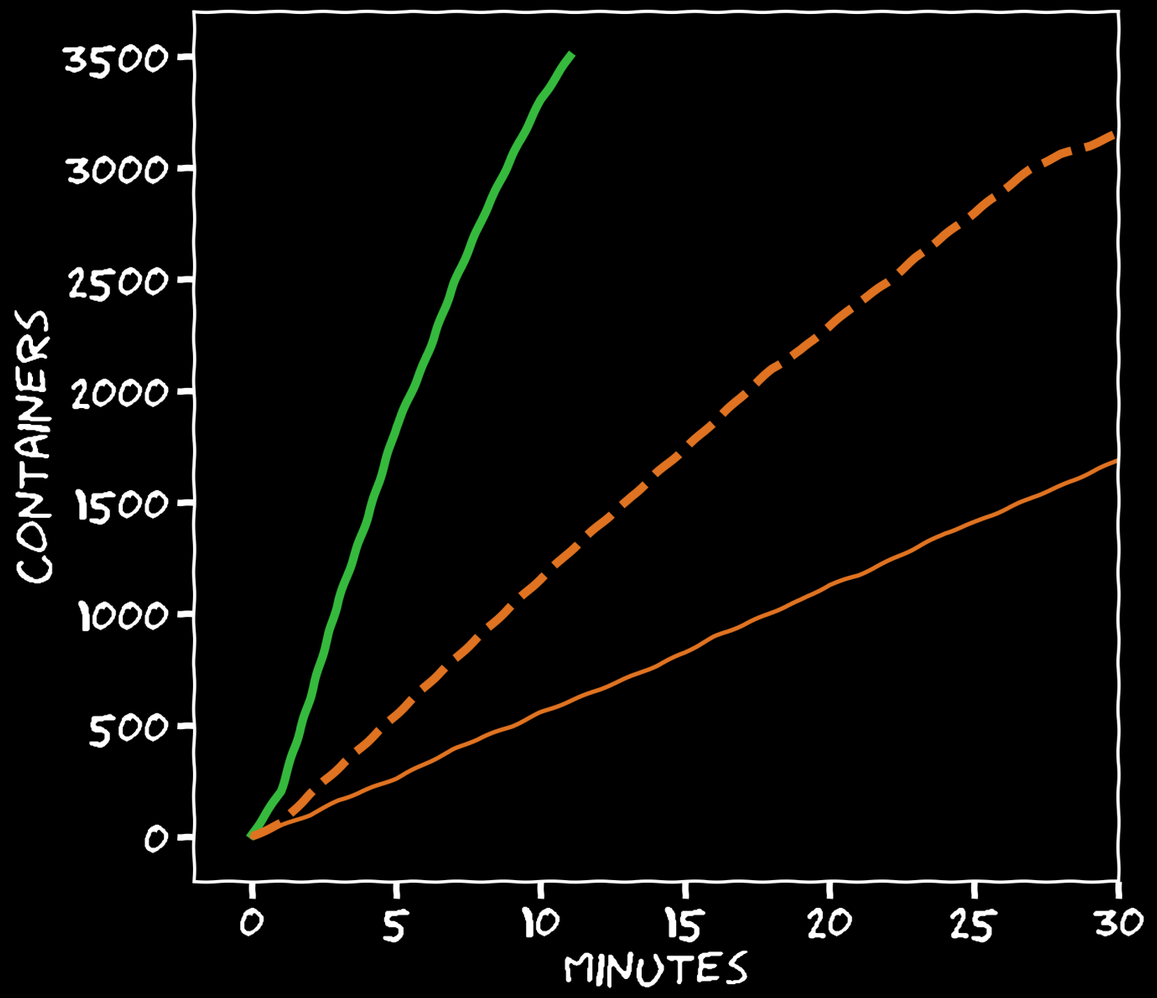
- Limits changed
  - 10 TPS per AWS account
  - 2 TPS per service
- Can we do better?

Amazon ECS on AWS Fargate



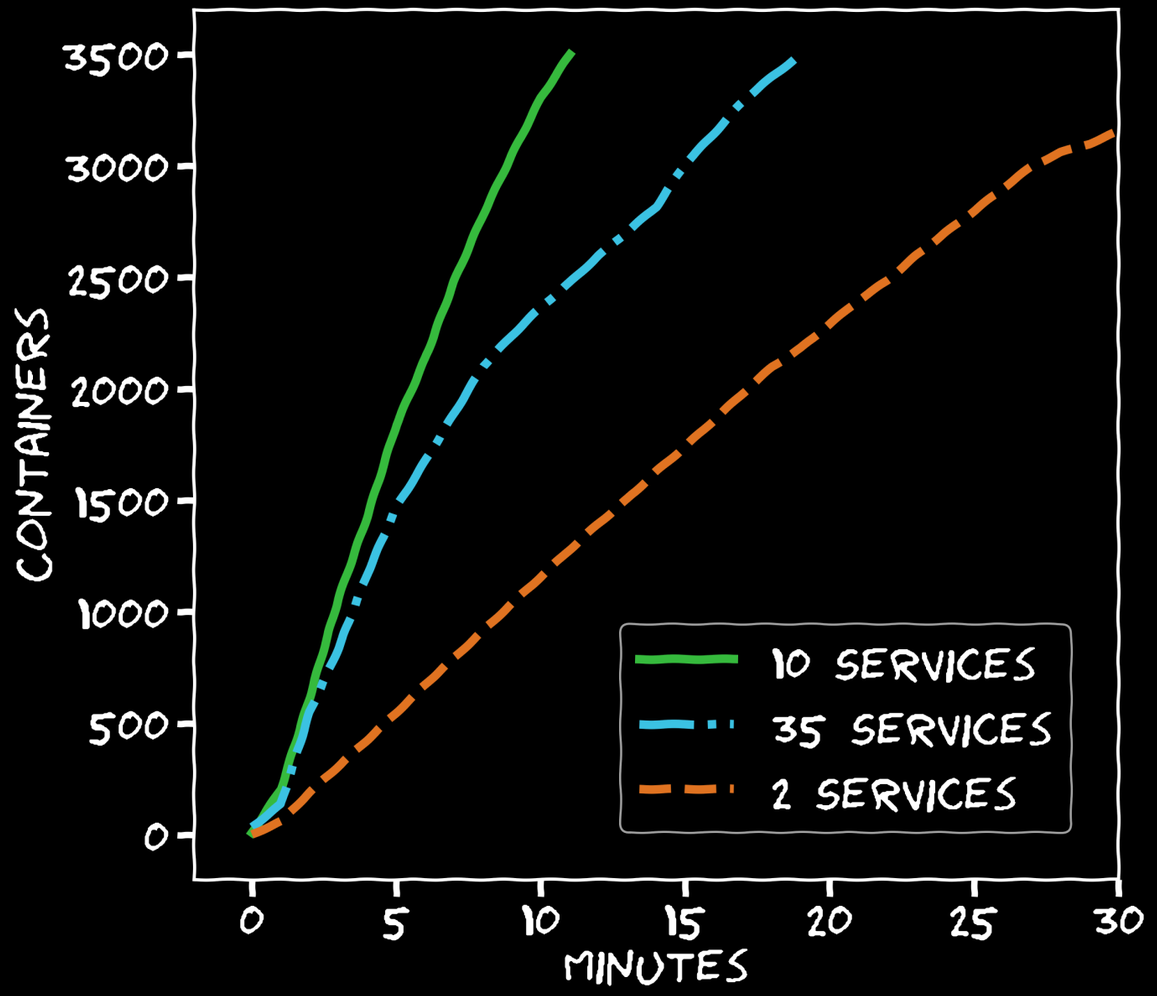
- Limits changed
  - 10 TPS per AWS account
  - 2 TPS per service
- Can we do better?
  - Yes!

Amazon ECS on AWS Fargate



- Limits changed
  - 10 TPS per AWS account
  - **2 TPS per service**
- Can we do better?
  - Yes!

Amazon ECS on AWS Fargate



- Limits changed
  - 10 TPS per AWS account
  - **2 TPS per service**
- Can we do better?
  - Yes!

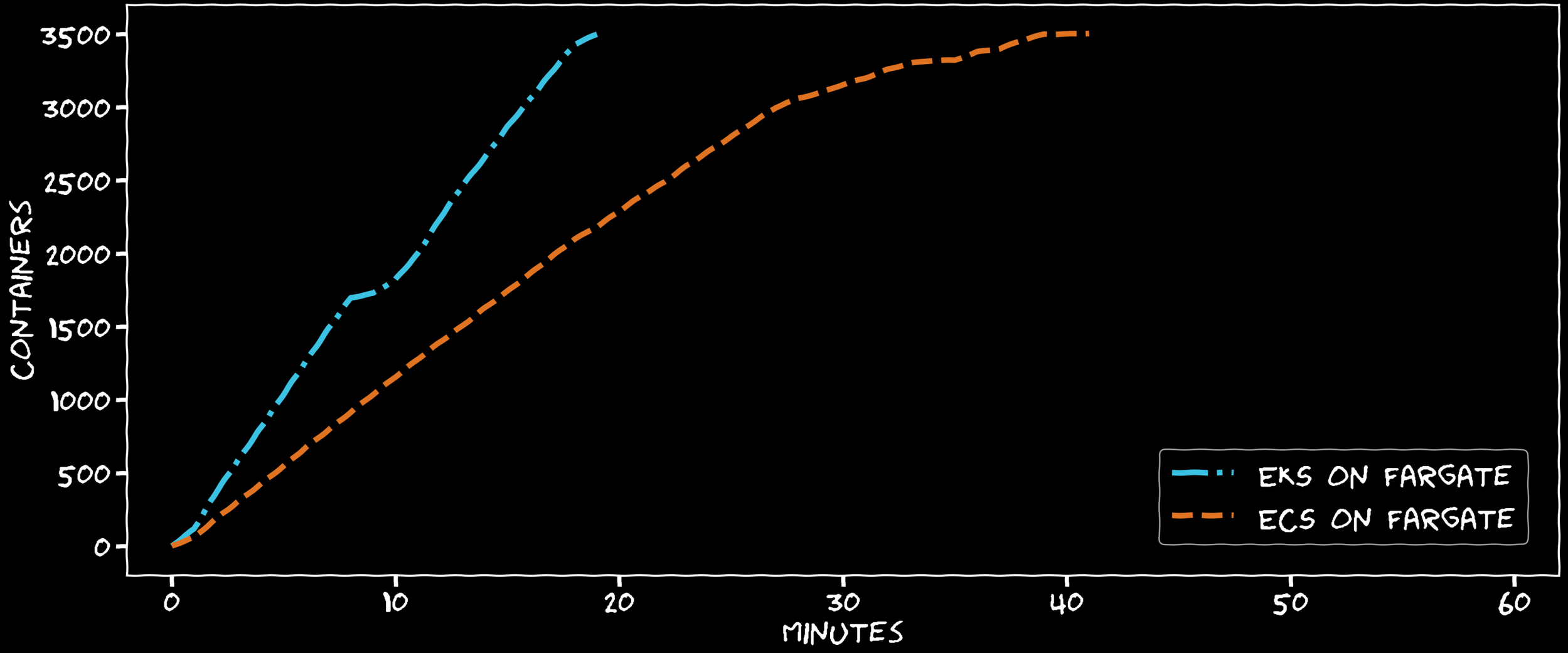
AWS Fargate compute discussion



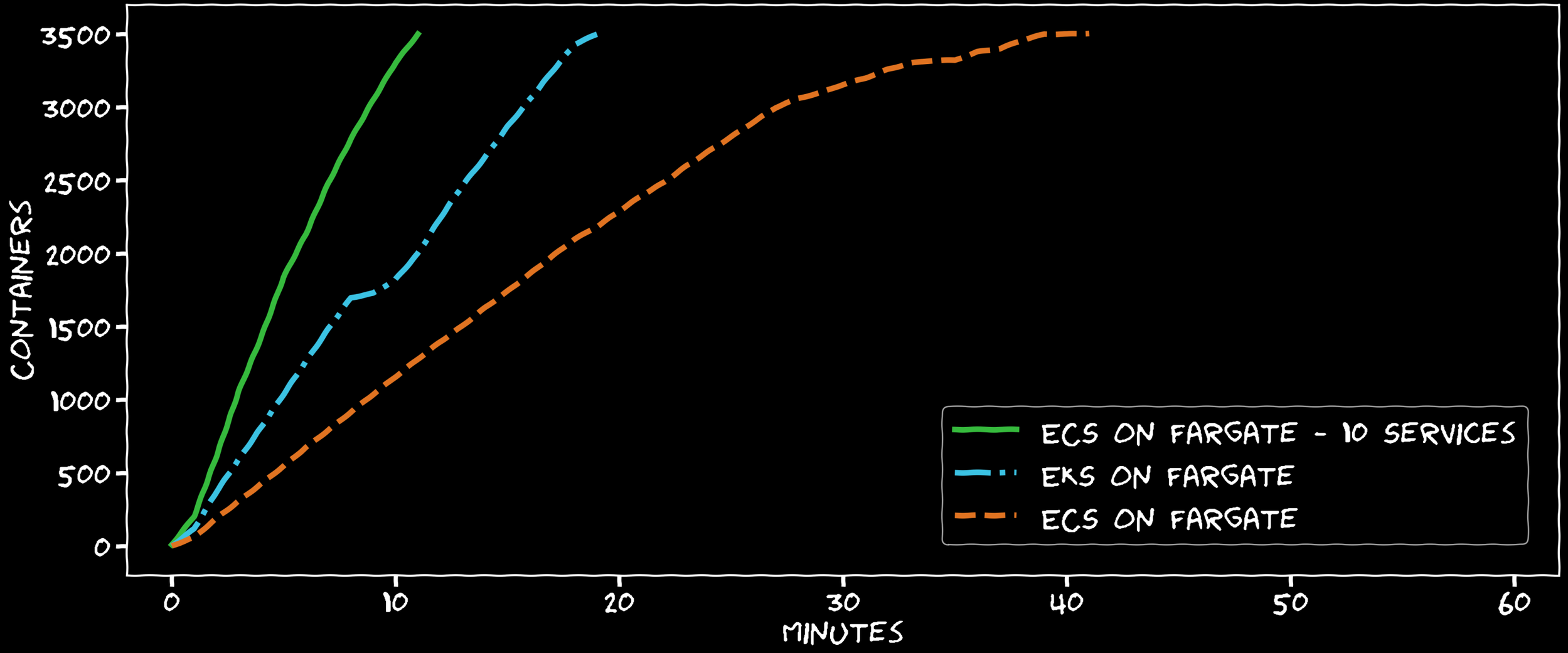
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

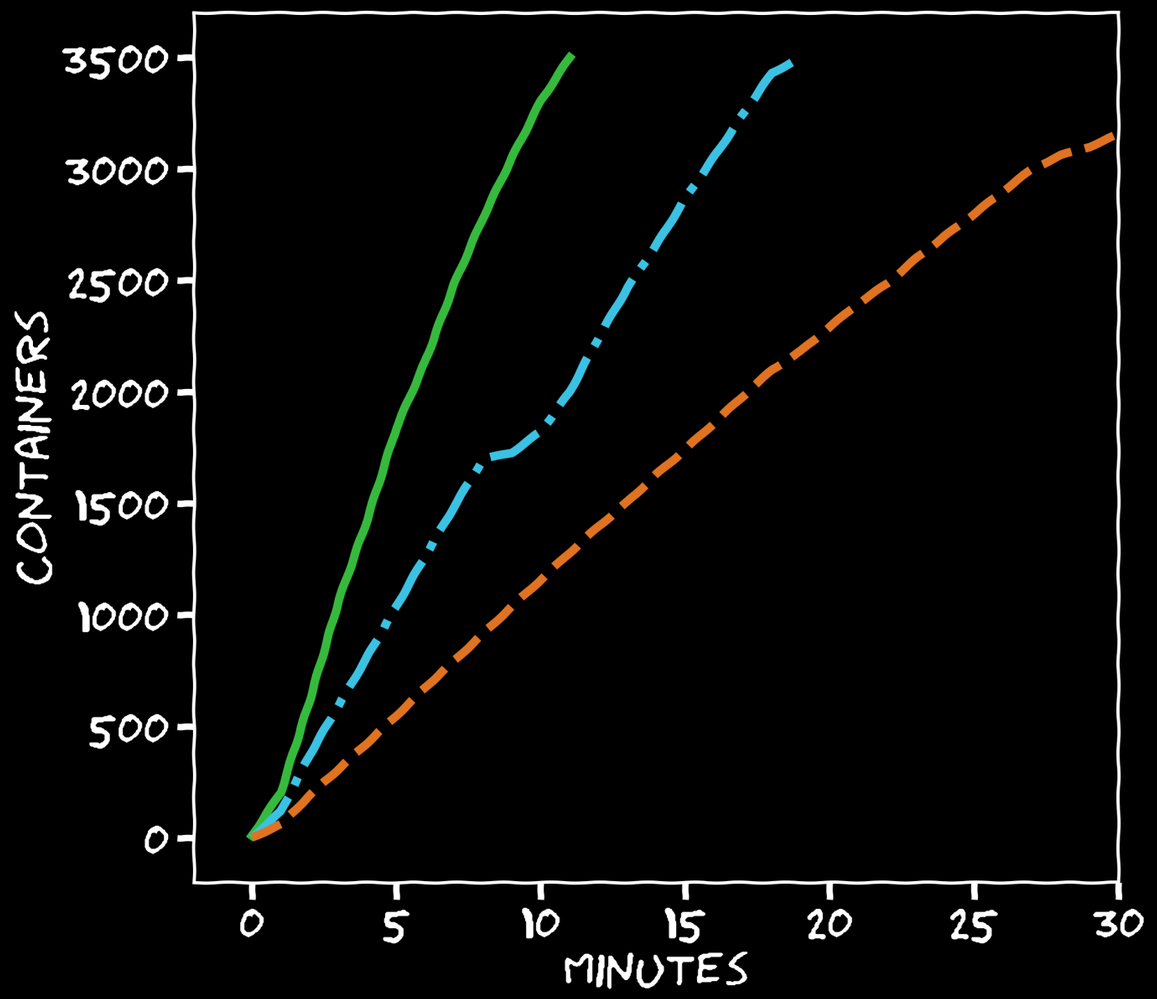
AWS Fargate compute



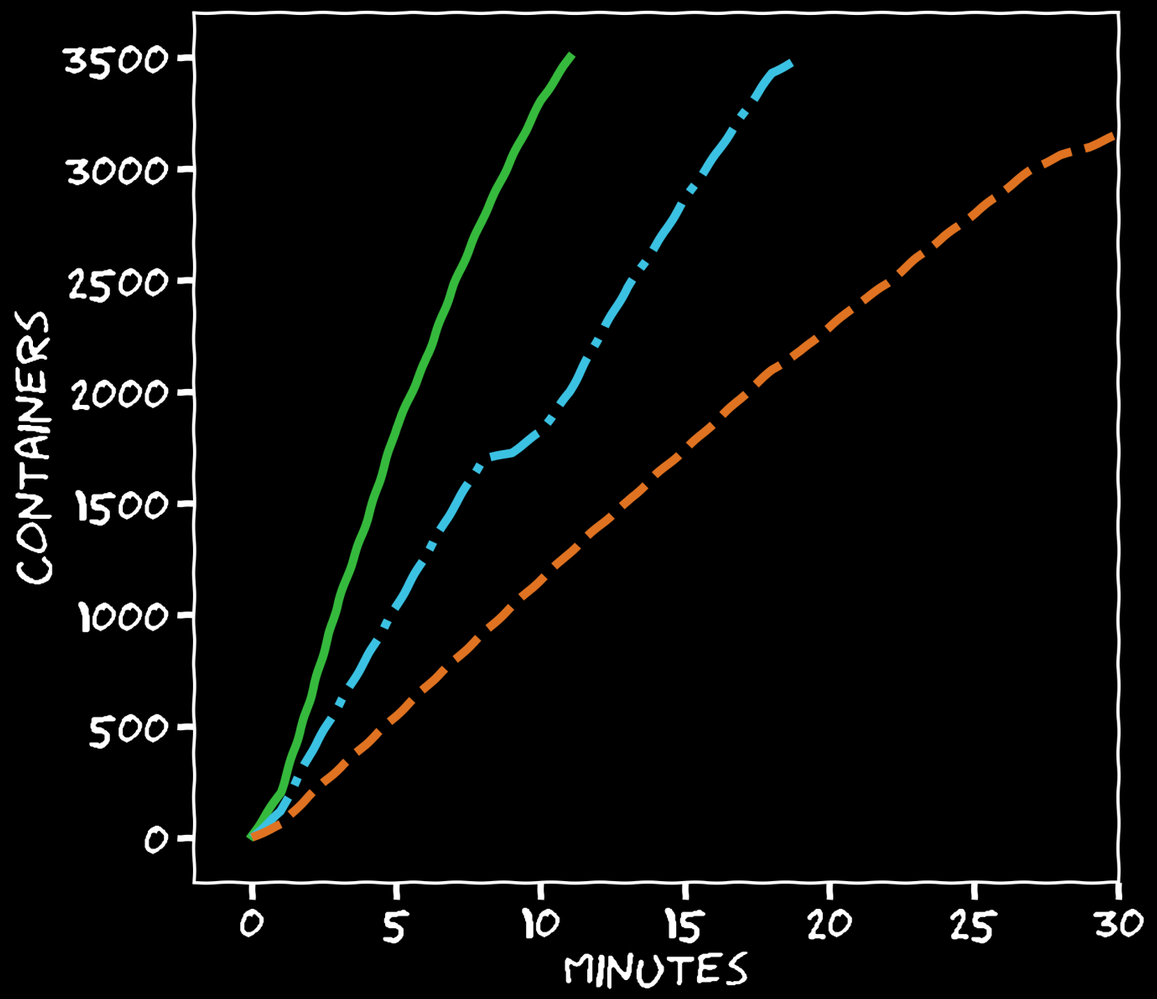
AWS Fargate compute



AWS Fargate compute

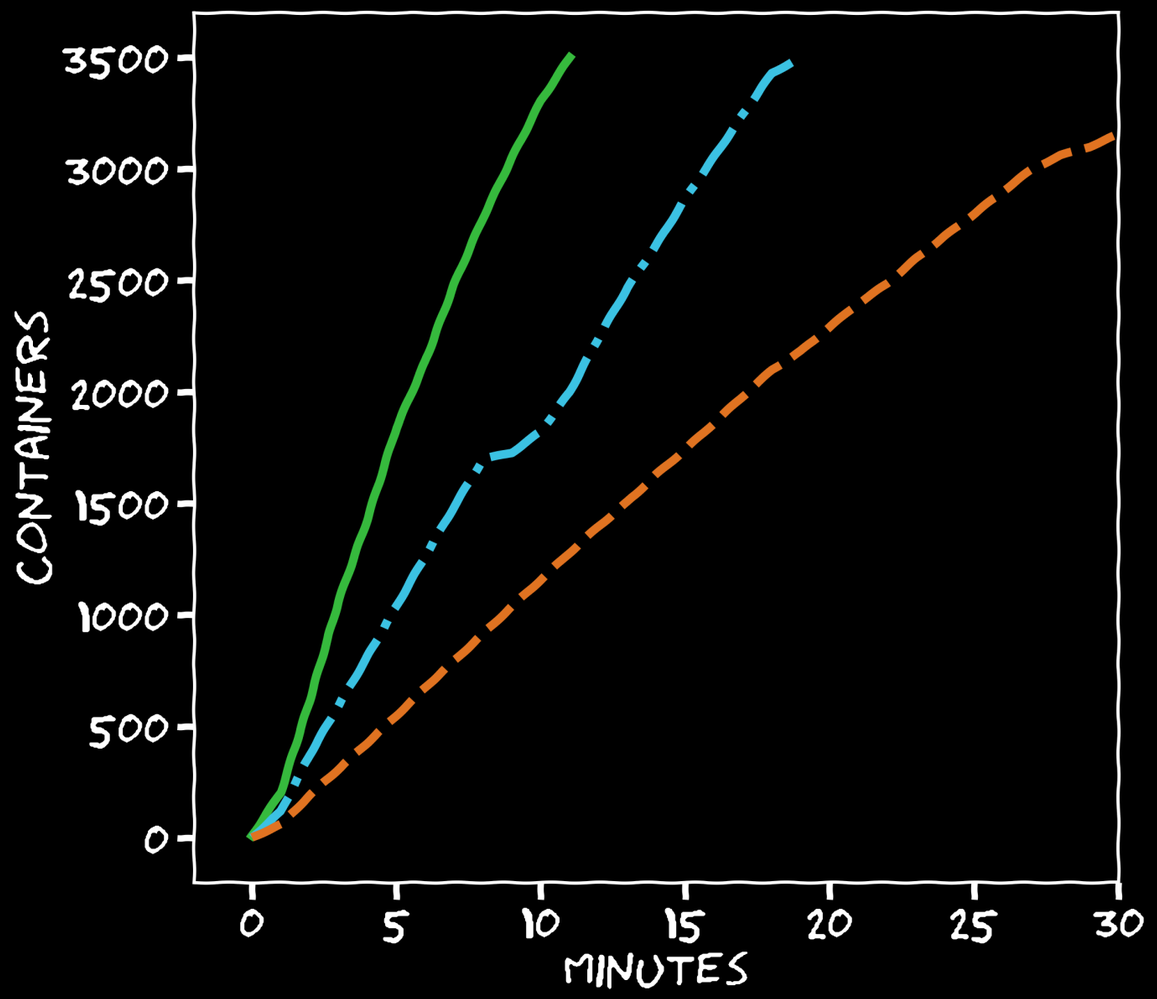


AWS Fargate compute



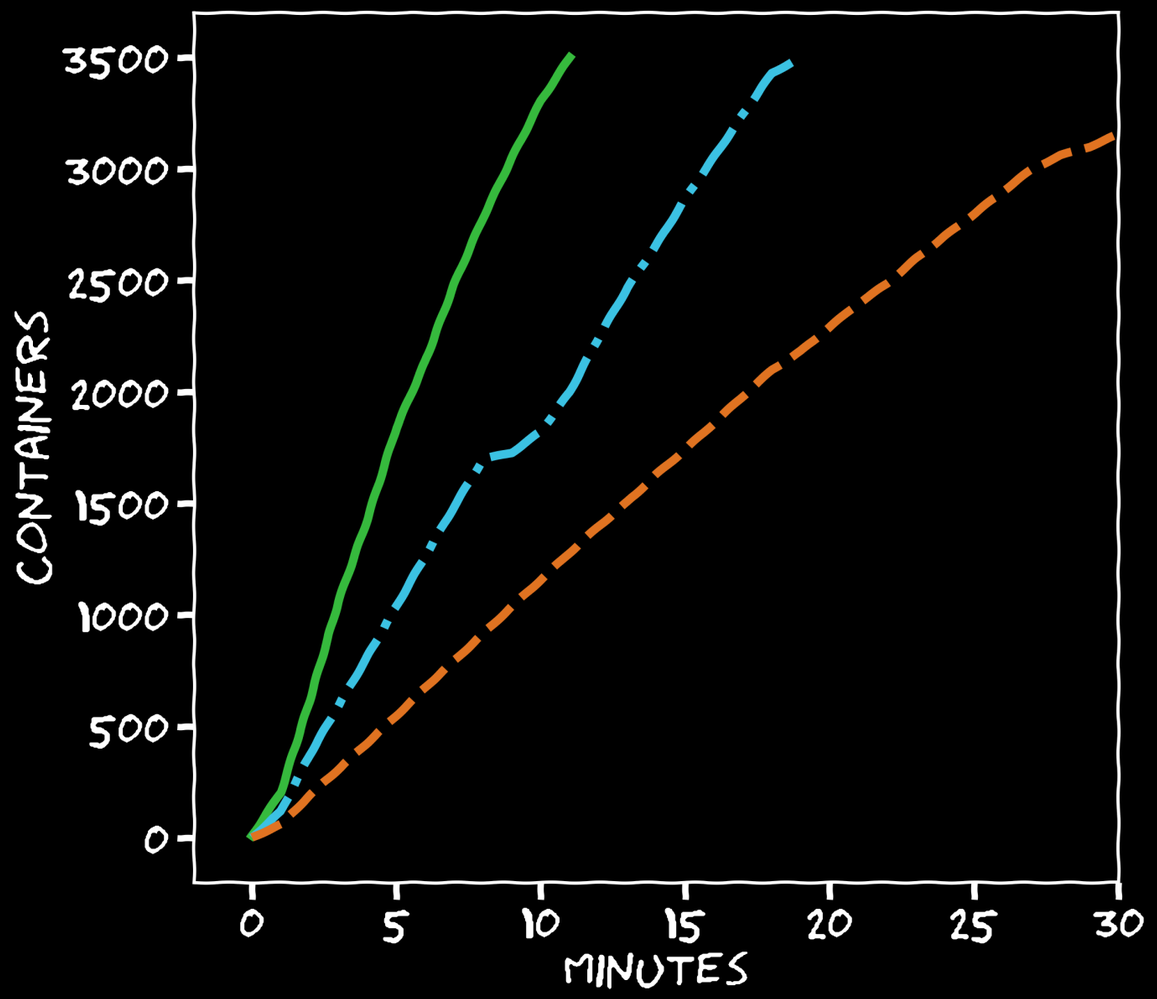
- Where would you use AWS Fargate?

AWS Fargate compute



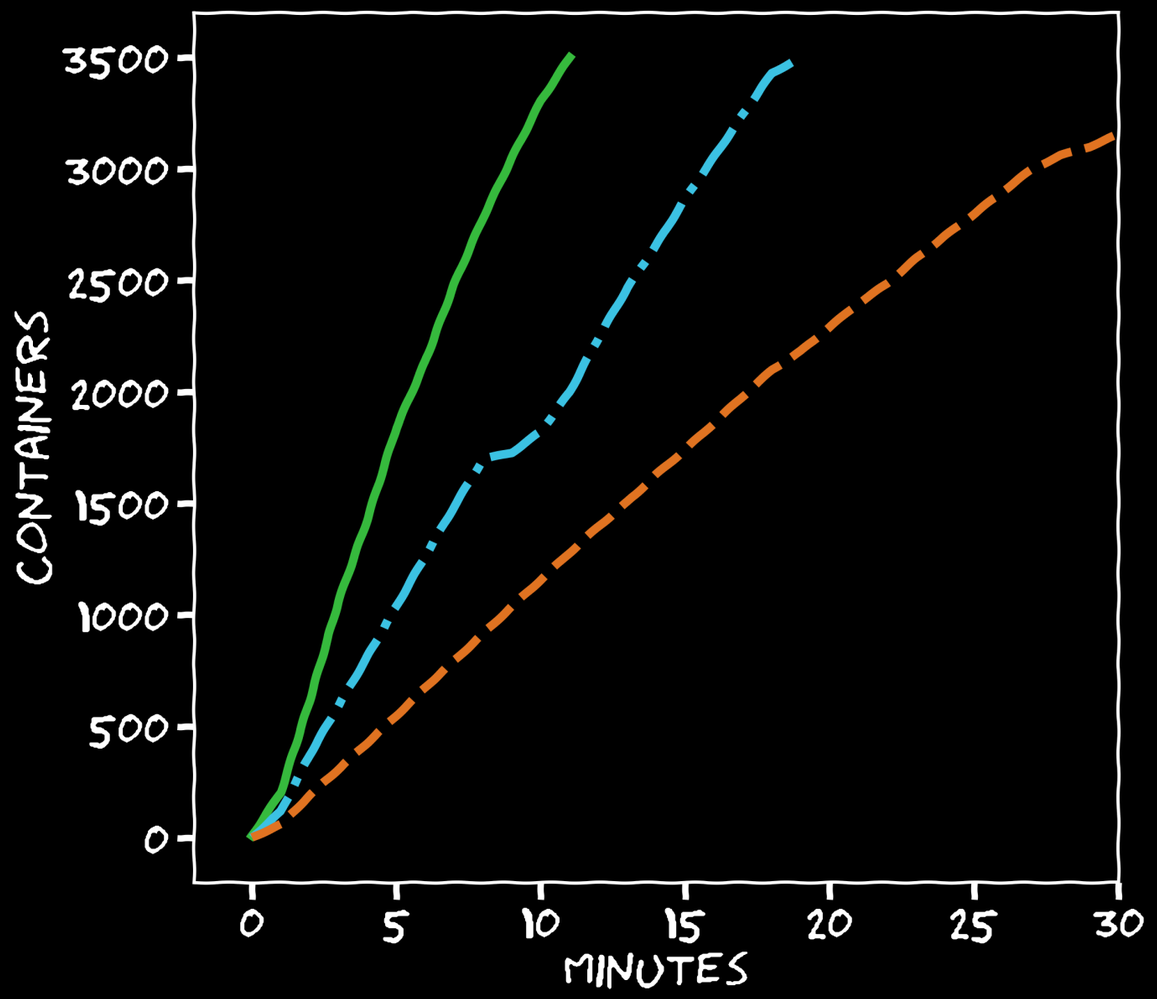
- Where would you use AWS Fargate?
  - Web applications?

AWS Fargate compute



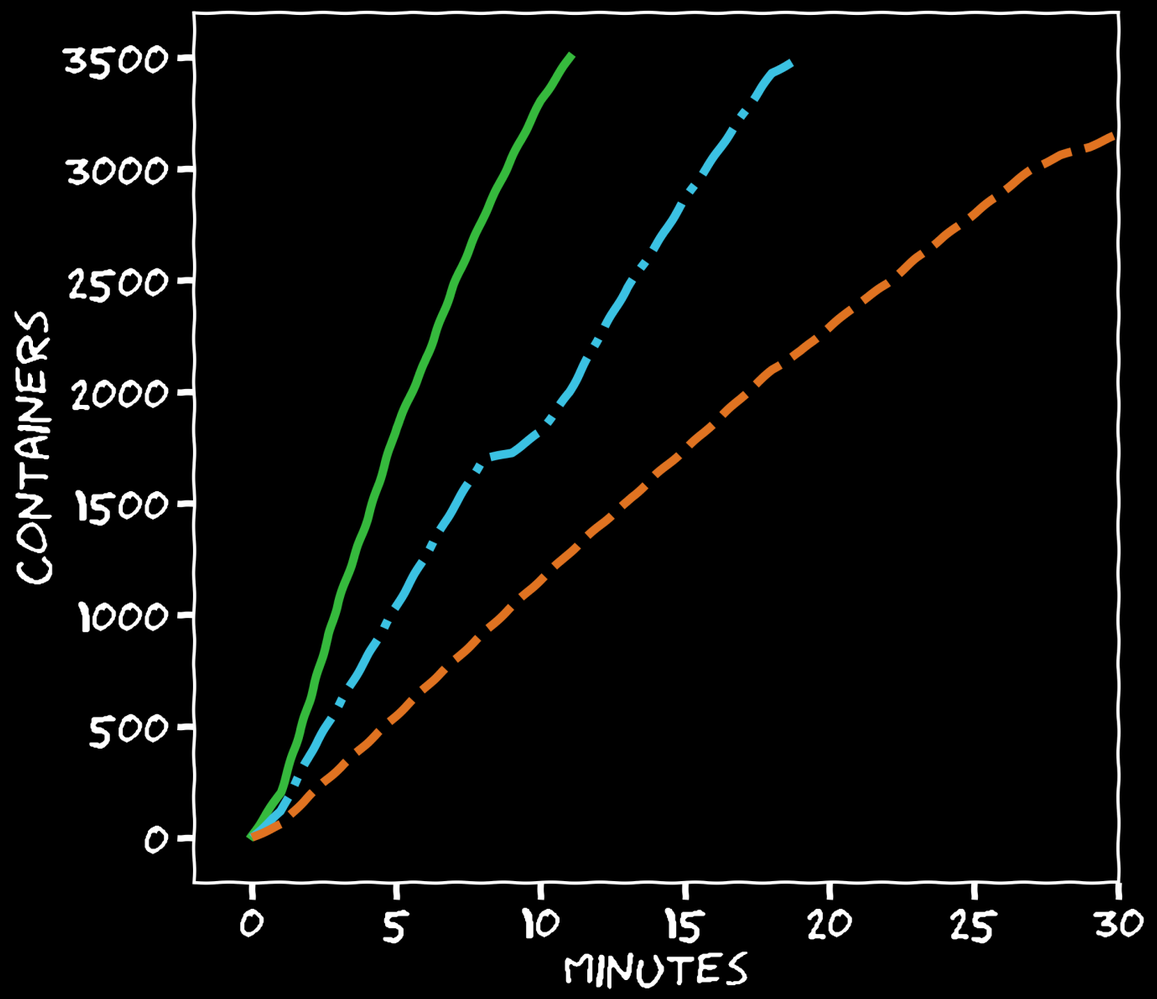
- Where would you use AWS Fargate?
  - Web applications? ✓

AWS Fargate compute



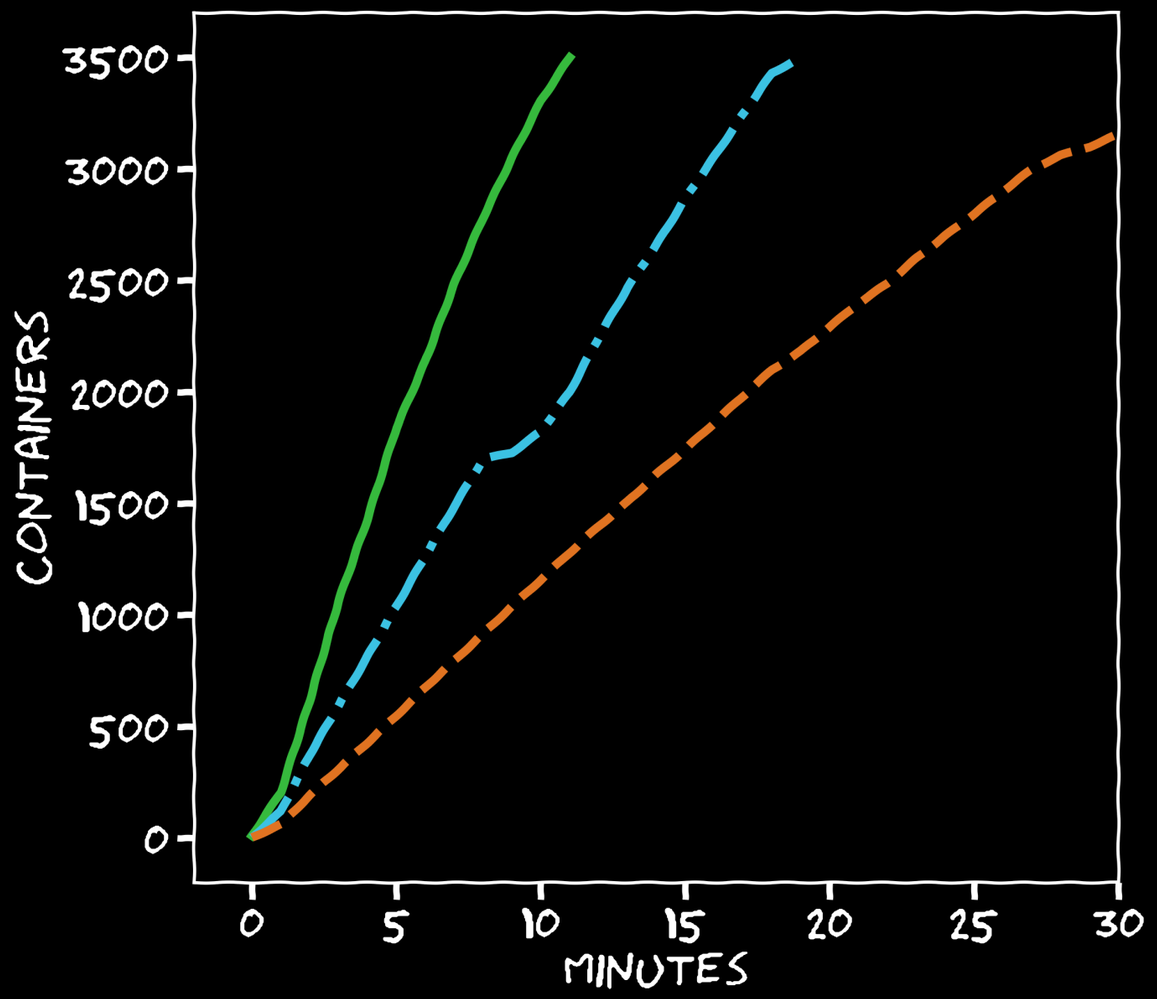
- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications?

AWS Fargate compute



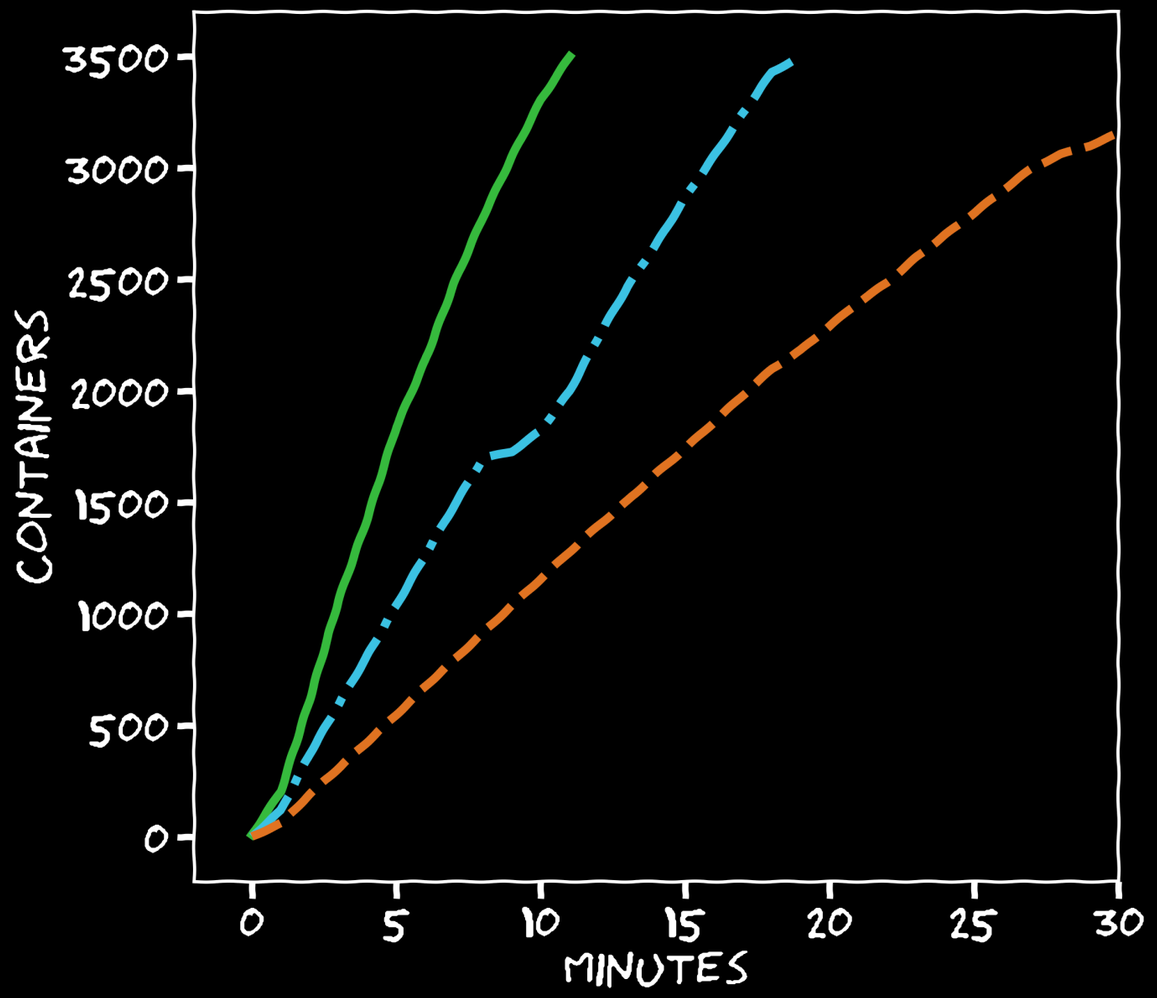
- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓

AWS Fargate compute



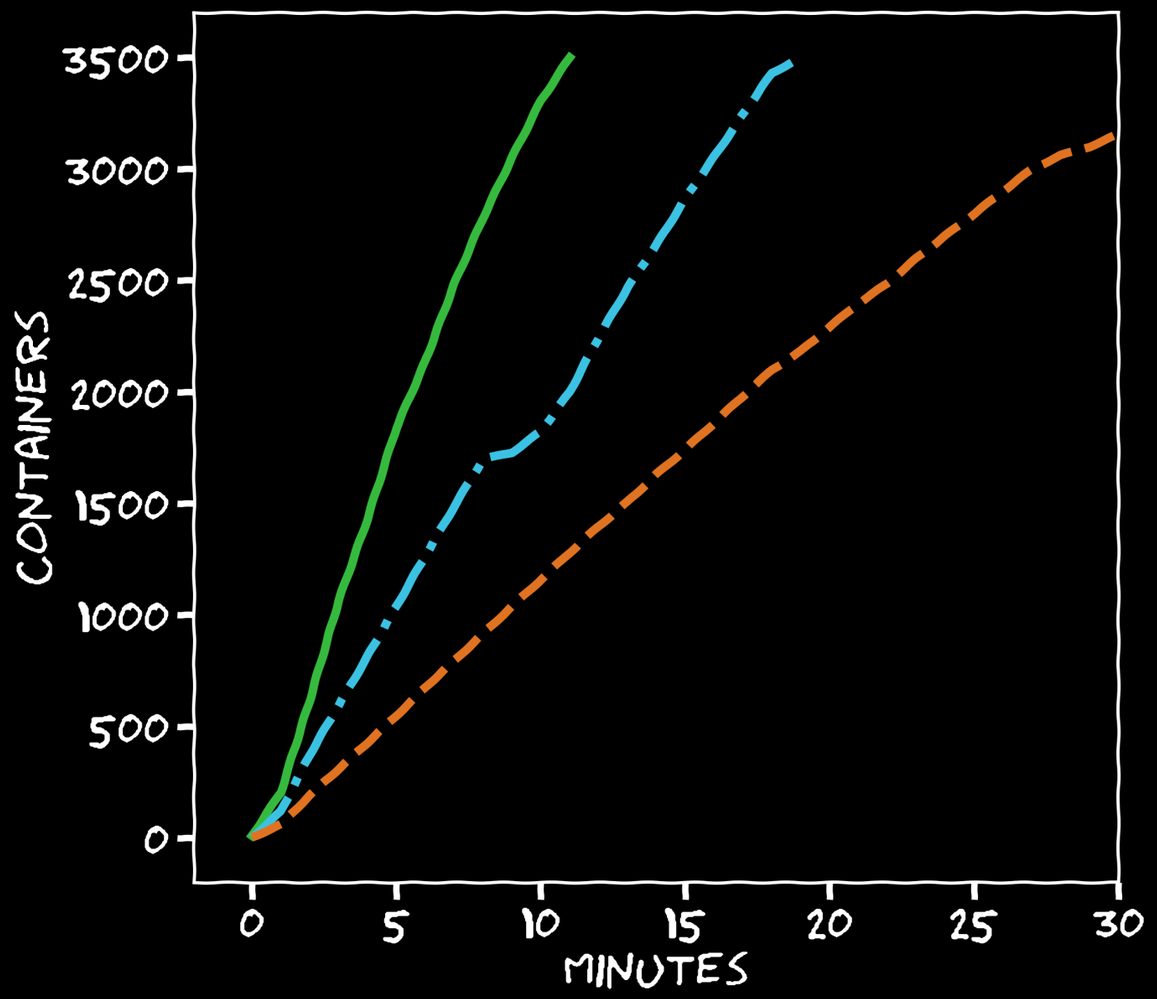
- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive?

AWS Fargate compute



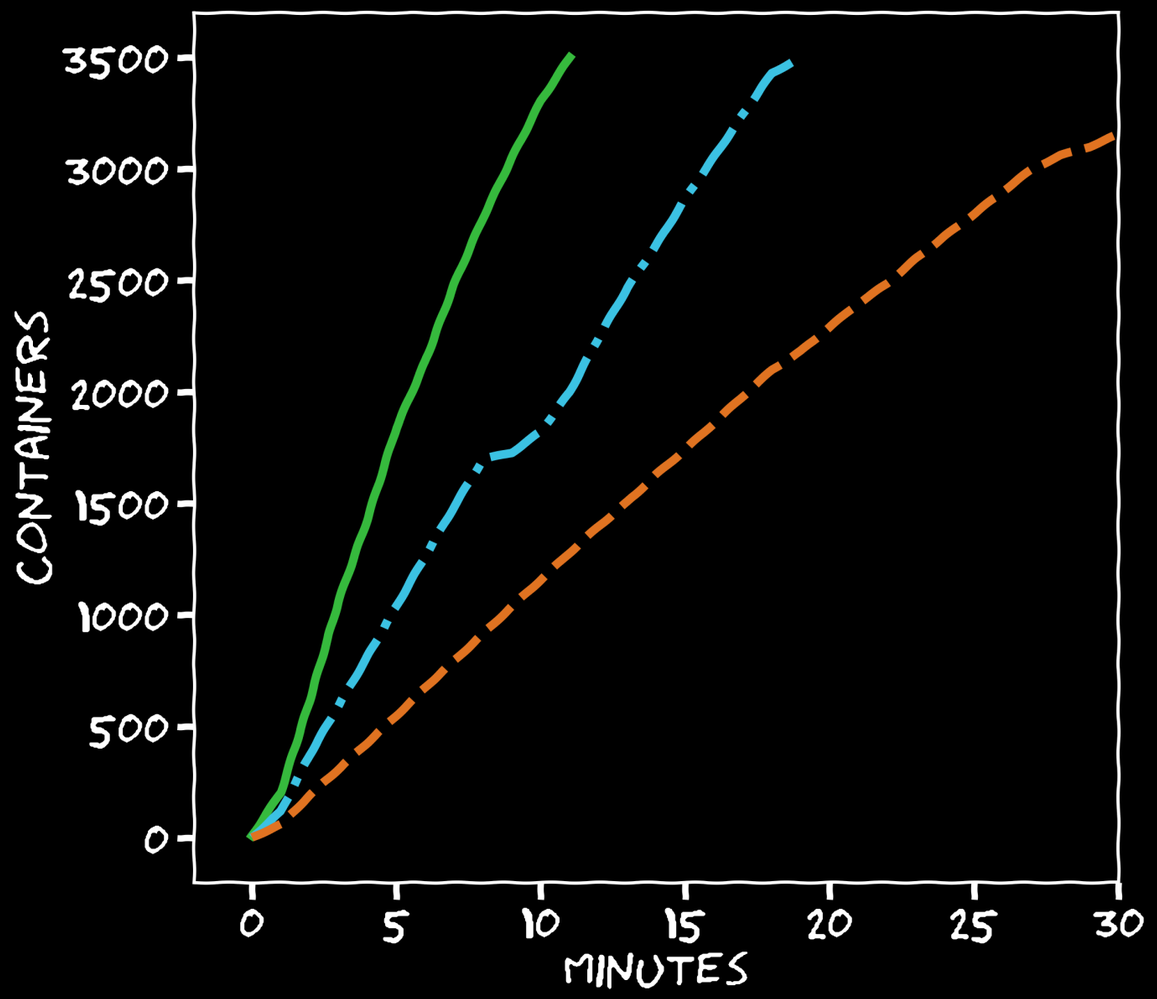
- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive? ✓

AWS Fargate compute



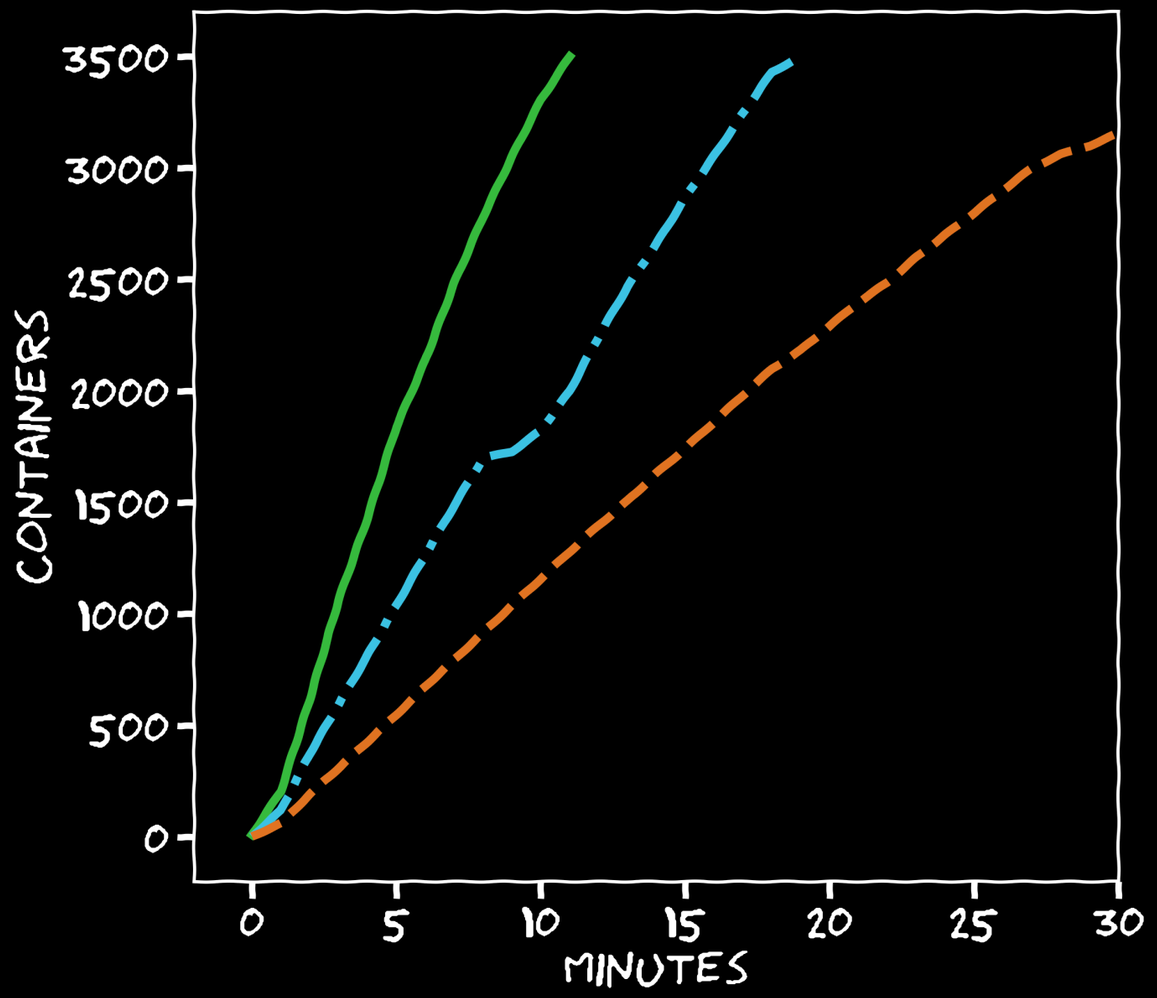
- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive? ✓
  - Background work?

AWS Fargate compute



- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive? ✓
  - Background work? ✓

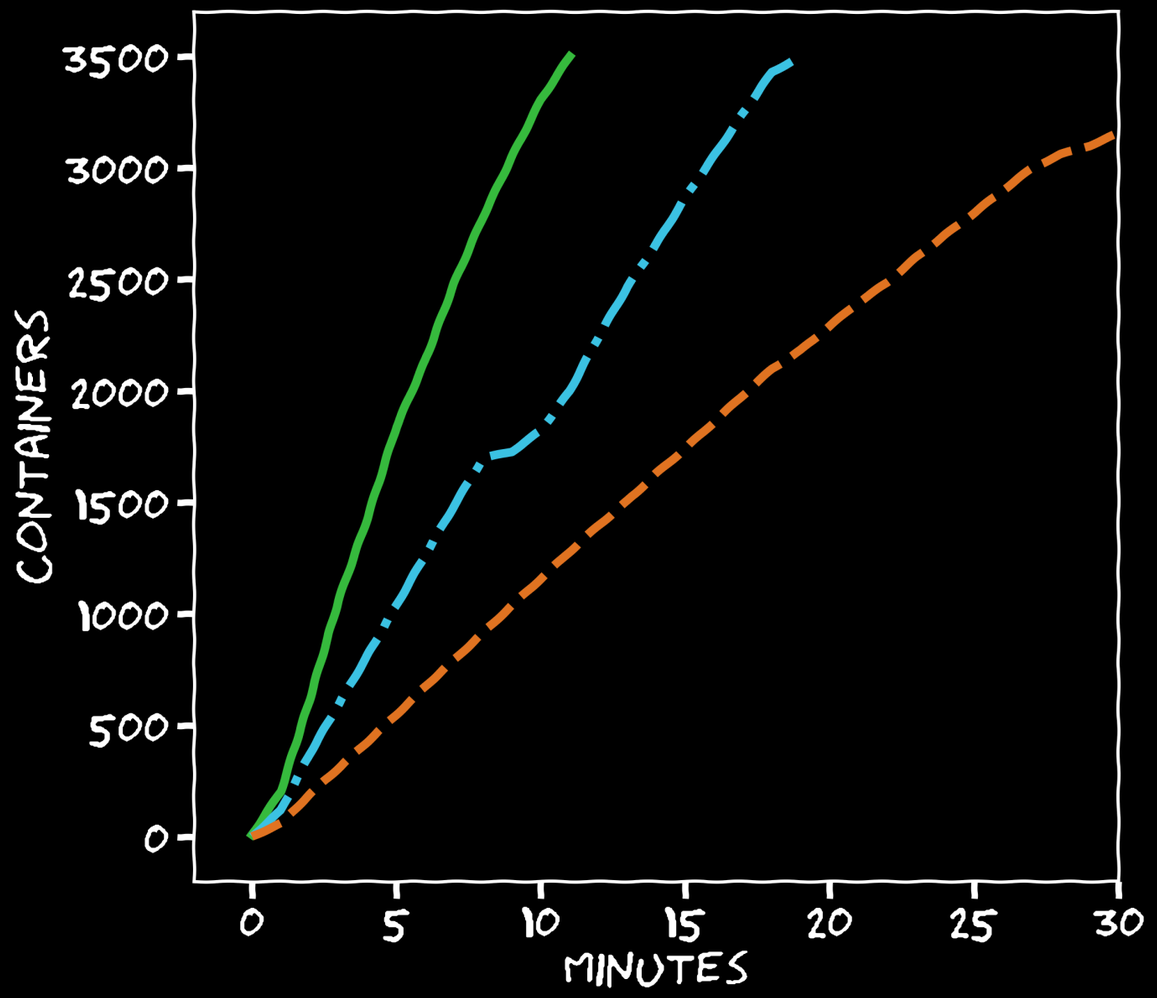
AWS Fargate compute



- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive? ✓
  - Background work? ✓

  - Network-bound services?

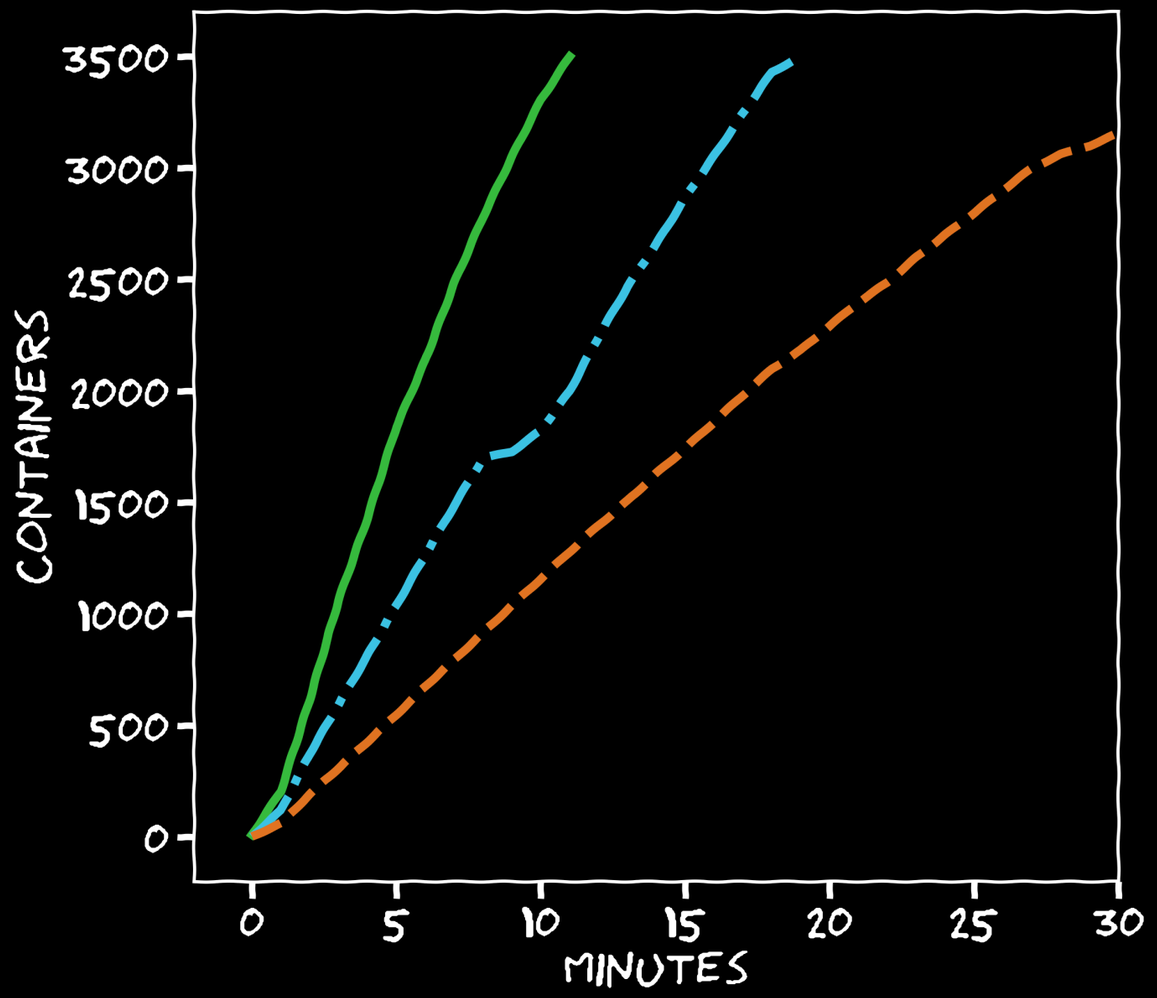
AWS Fargate compute



- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive? ✓
  - Background work? ✓

  - Network-bound services? ✗

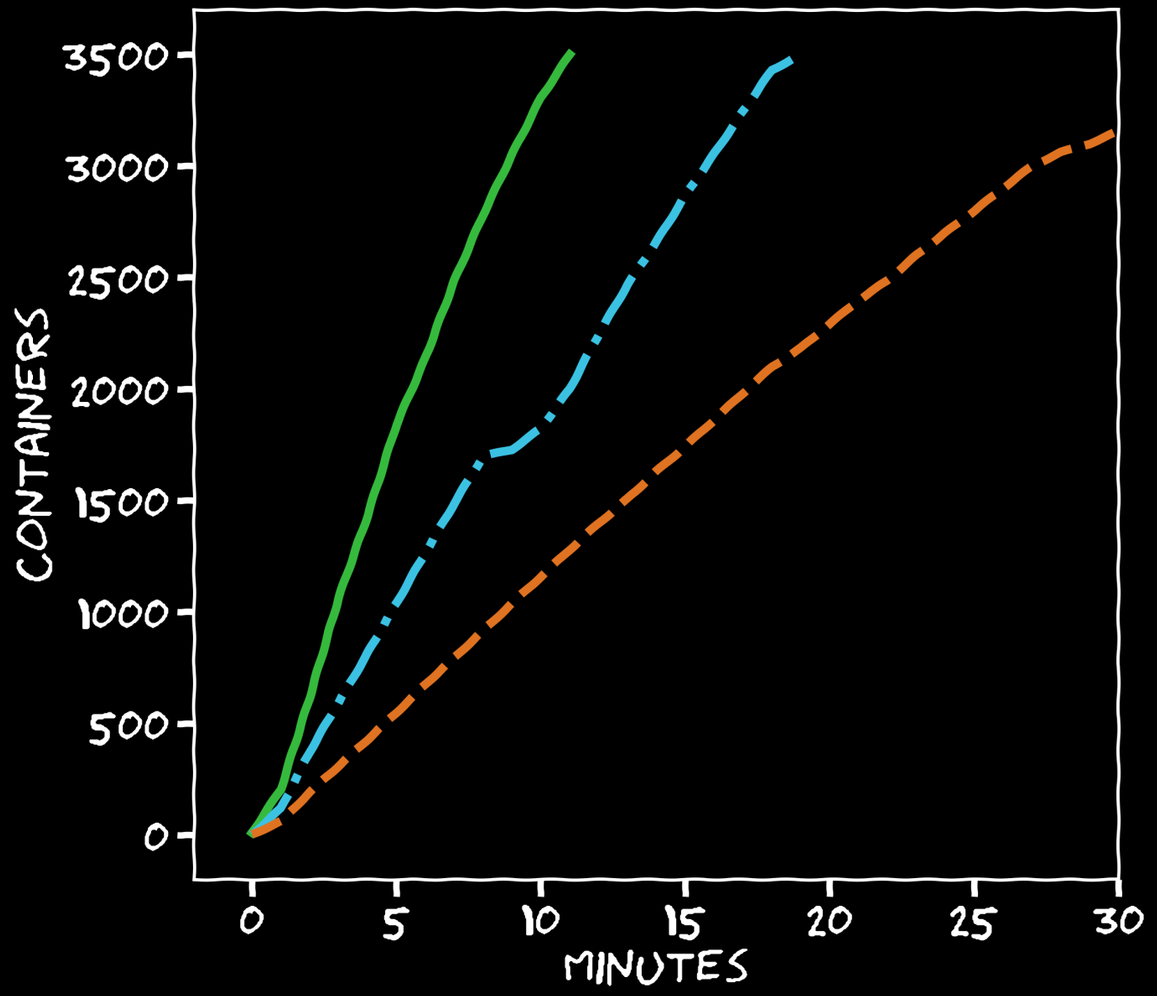
AWS Fargate compute



- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive? ✓
  - Background work? ✓

  - Network-bound services? ✗
  - GPU-accelerated jobs?

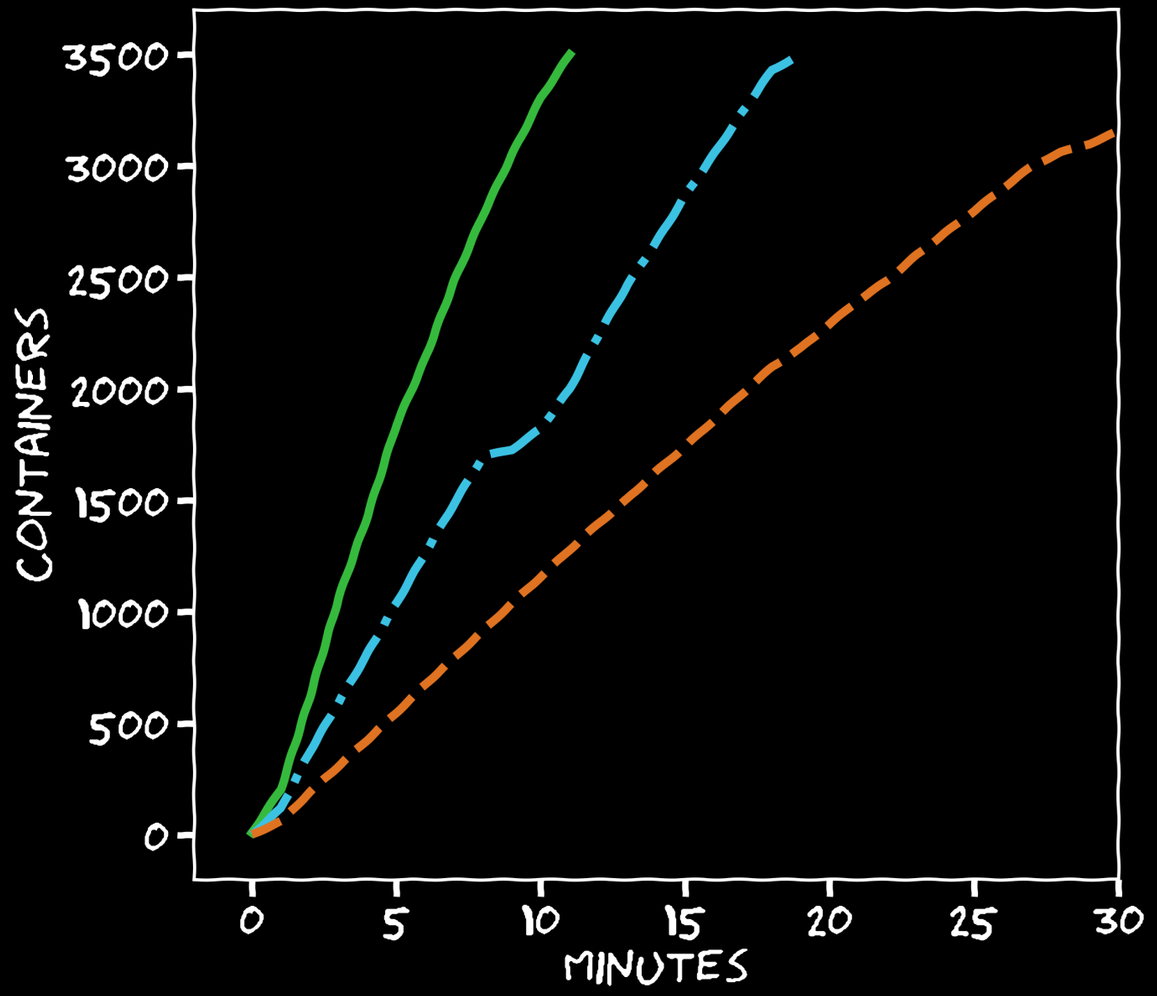
AWS Fargate compute



- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive? ✓
  - Background work? ✓

  - Network-bound services? ✗
  - GPU-accelerated jobs? ✗

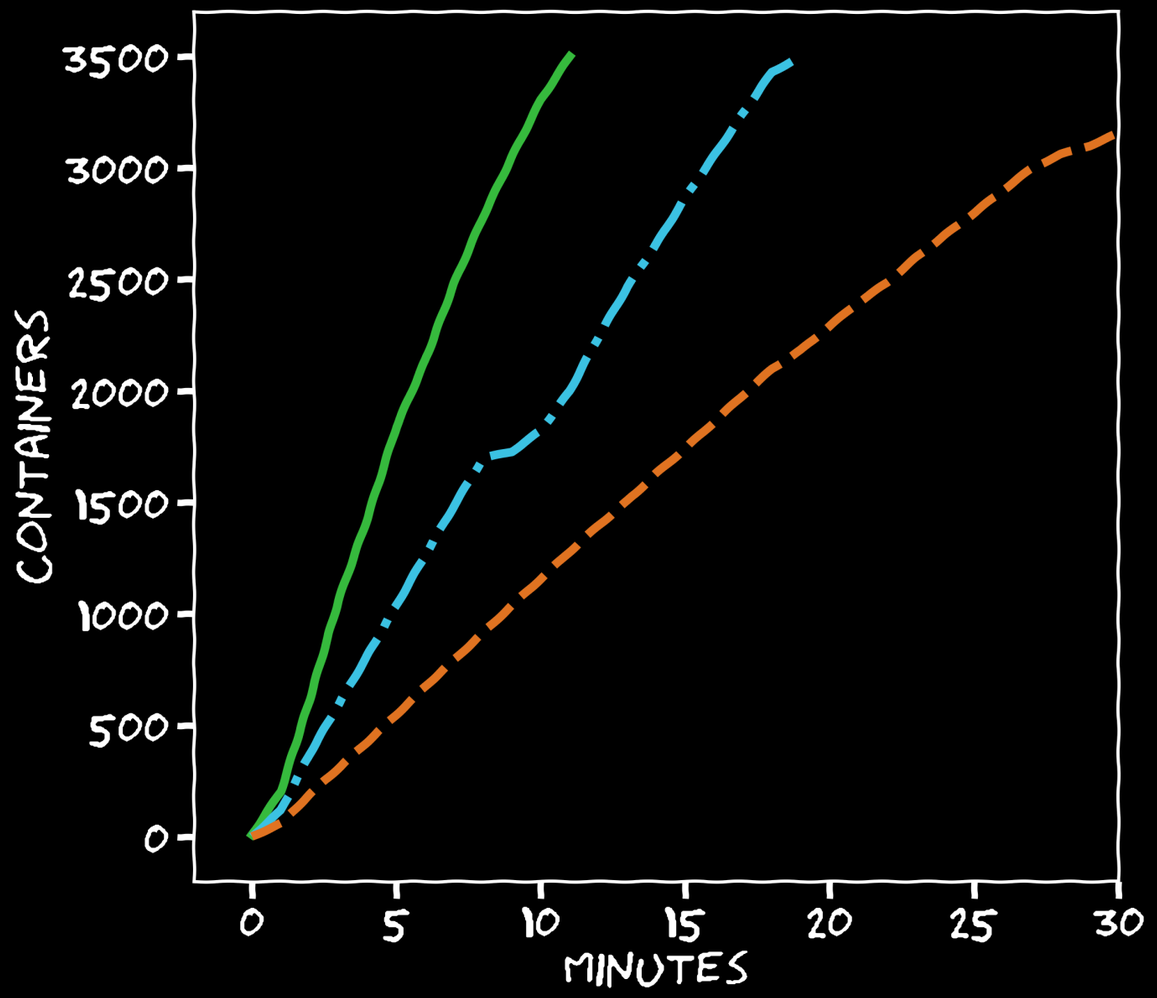
AWS Fargate compute



- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive? ✓
  - Background work? ✓

  - Network-bound services? ✗
  - GPU-accelerated jobs? ✗
  - ARM workloads?

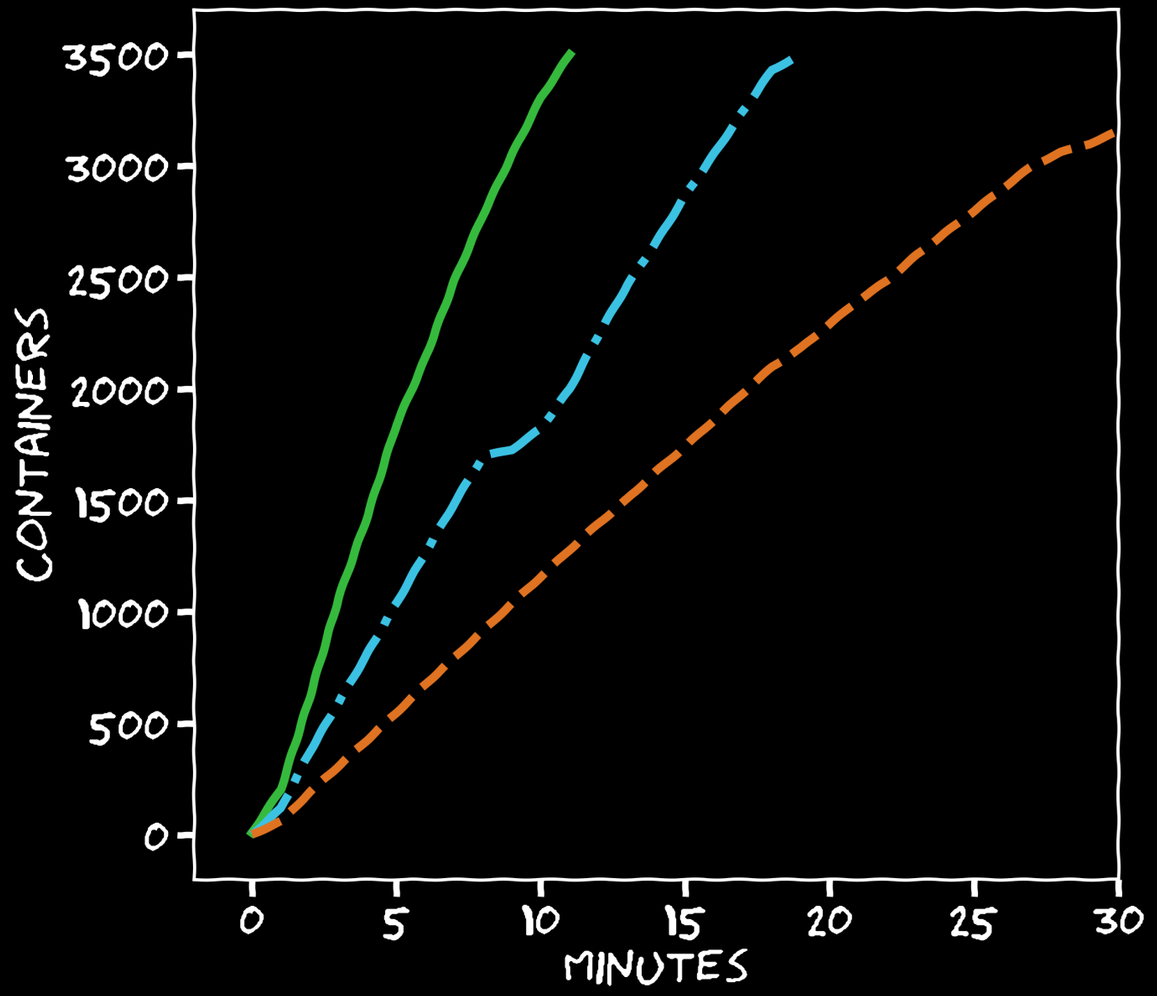
AWS Fargate compute



- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive? ✓
  - Background work? ✓

  - Network-bound services? ✗
  - GPU-accelerated jobs? ✗
  - ARM workloads? ✗

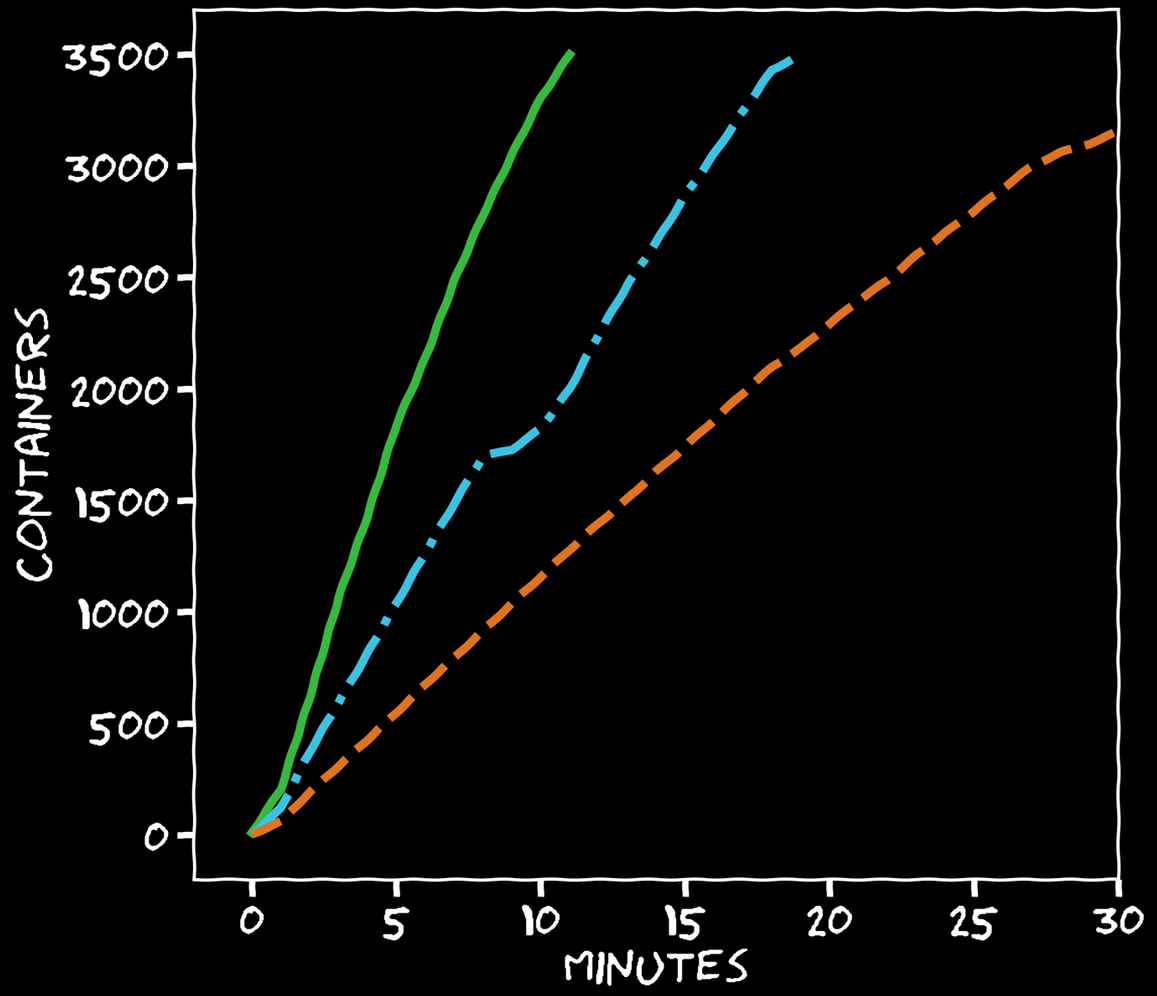
AWS Fargate compute



- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive? ✓
  - Background work? ✓

  - Network-bound services? ✗
  - GPU-accelerated jobs? ✗
  - ARM workloads? ✗
  - Windows-only apps?

AWS Fargate compute



- Where would you use AWS Fargate?
  - Web applications? ✓
  - User-facing applications? ✓
  - Latency sensitive? ✓
  - Background work? ✓

  - Network-bound services? ✗
  - GPU-accelerated jobs? ✗
  - ARM workloads? ✗
  - Windows-only apps? ✗

Cost



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cost

AMAZON

AMAZON

Cost

AMAZON

- Cluster

AMAZON

- Cluster

Cost

AMAZON

- Cluster
  - $0 per hour

AMAZON

- Cluster
  - $0.10 per hour

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers

AMAZON

- Cluster
  - $0.10 per hour
- Workers

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances

AMAZON

- Cluster
  - $0.10 per hour
- Workers
  - Amazon EC2 instances

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
    - On demand
    - Spot
    - Reserved instances
    - Savings plans

AMAZON

- Cluster
  - $0.10 per hour
- Workers
  - Amazon EC2 instances
    - On demand
    - Spot
    - Reserved instances
    - Savings plans

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances

AMAZON

- Cluster
  - $0.10 per hour
- Workers
  - Amazon EC2 instances

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
  - AWS Fargate

AMAZON

- Cluster
  - $0.10 per hour
- Workers
  - Amazon EC2 instances
  - AWS Fargate

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
  - AWS Fargate
    - On demand
    - Spot
    - Compute savings plan

AMAZON

- Cluster
  - $0.10 per hour
- Workers
  - Amazon EC2 instances
  - AWS Fargate
    - On demand
    -
    - Compute savings plan

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
  - AWS Fargate

AMAZON

- Cluster
  - $0.10 per hour
- Workers
  - Amazon EC2 instances
  - AWS Fargate
    - No Spot option

People costs!



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Opportunity costs!



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
  - AWS Fargate

AMAZON

- Cluster
  - $0.10 per hour
- Workers
  - Amazon EC2 instances
  - AWS Fargate

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
  - AWS Fargate

AMAZON

- Cluster
  - $0.10 per hour
  - Upgrades every three months
- Workers
  - Amazon EC2 instances
  - AWS Fargate

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
  - AWS Fargate

AMAZON

- Cluster
  - $0.10 per hour
  - Upgrades every three months
  - Component upgrades
- Workers
  - Amazon EC2 instances
  - AWS Fargate

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
  - AWS Fargate

AMAZON

- Cluster
  - $0.10 per hour
  - 1 to 20 engineers
- Workers
  - Amazon EC2 instances
  - AWS Fargate

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
    - AMI updates
  - AWS Fargate

AMAZON

- Cluster
  - $0.10 per hour
  - 1 to 20 engineers
- Workers
  - Amazon EC2 instances
    - AMI updates
  - AWS Fargate

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
    - AMI updates
    - AWS service updates
  - AWS Fargate

AMAZON

- Cluster
  - $0.10 per hour
  - 1 to 20 engineers
- Workers
  - Amazon EC2 instances
    - AMI updates
    - Component updates
  - AWS Fargate

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
    - One engineer
  - AWS Fargate

AMAZON

- Cluster
  - $0.10 per hour
  - 1 to 20 engineers
- Workers
  - Amazon EC2 instances
    - One engineer
  - AWS Fargate

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
    - One engineer
  - AWS Fargate
    - Version upgrades

AMAZON

- Cluster
  - $0.10 per hour
  - 1 to 20 engineers
- Workers
  - Amazon EC2 instances
    - One engineer
  - AWS Fargate
    - Version upgrades

Cost

AMAZON

- Cluster
  - $0 per hour
- Workers
  - Amazon EC2 instances
    - One engineer
  - AWS Fargate

AMAZON

- Cluster
  - $0.10 per hour
  - 1 to 20 engineers
- Workers
  - Amazon EC2 instances
    - One engineer
  - AWS Fargate

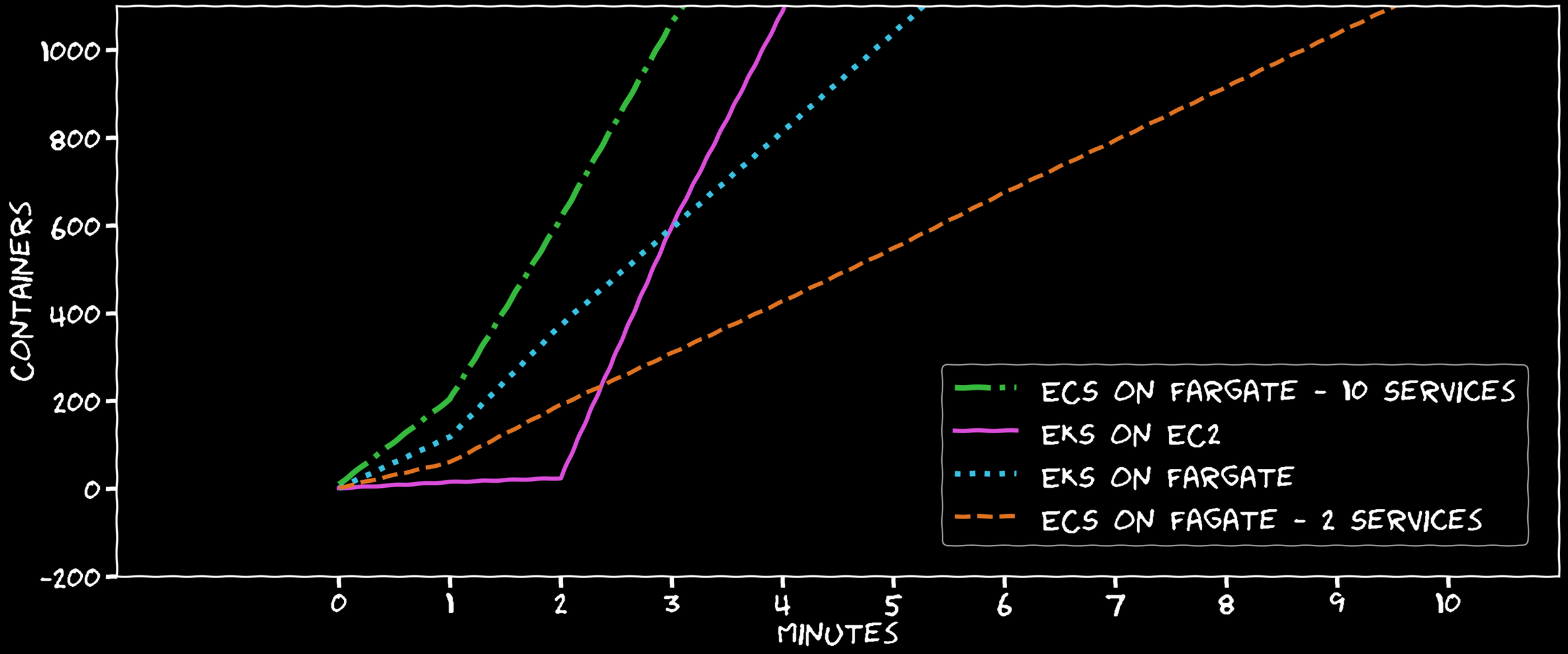
Ideal fit



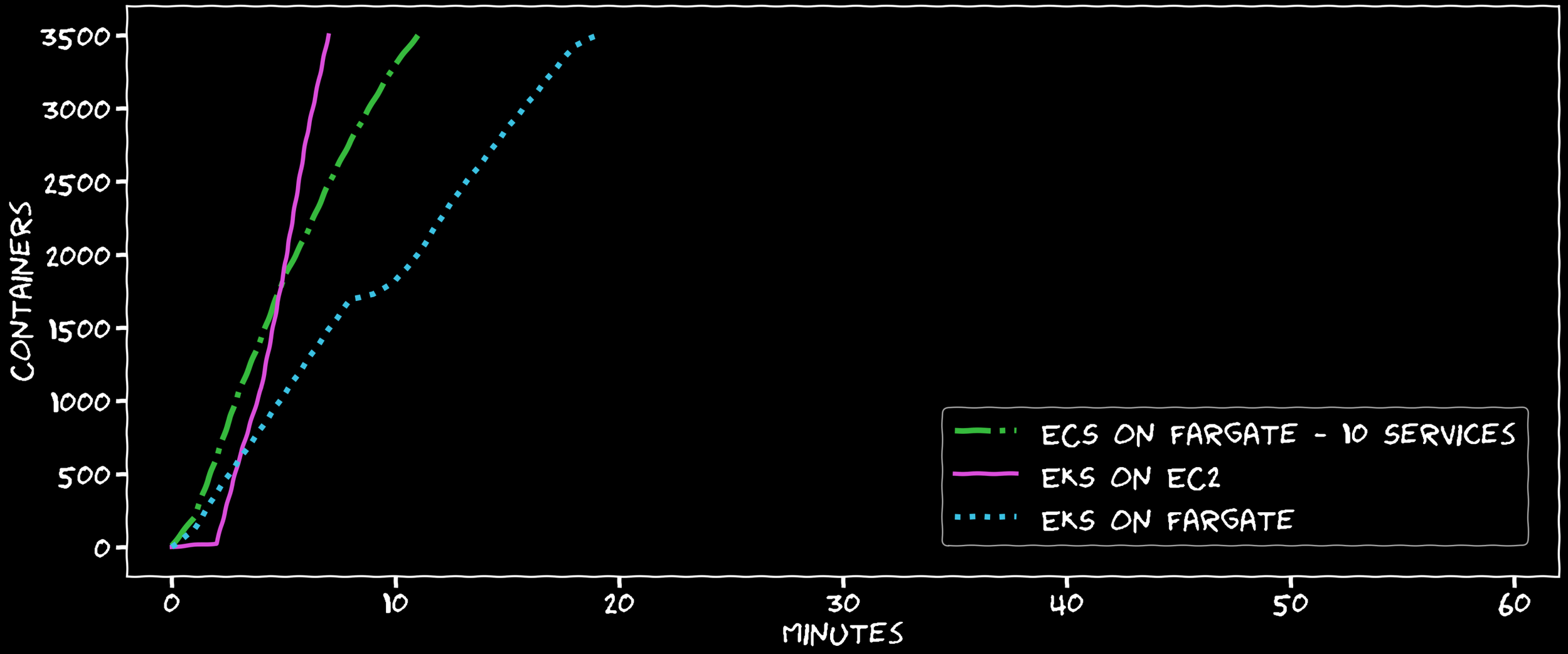
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

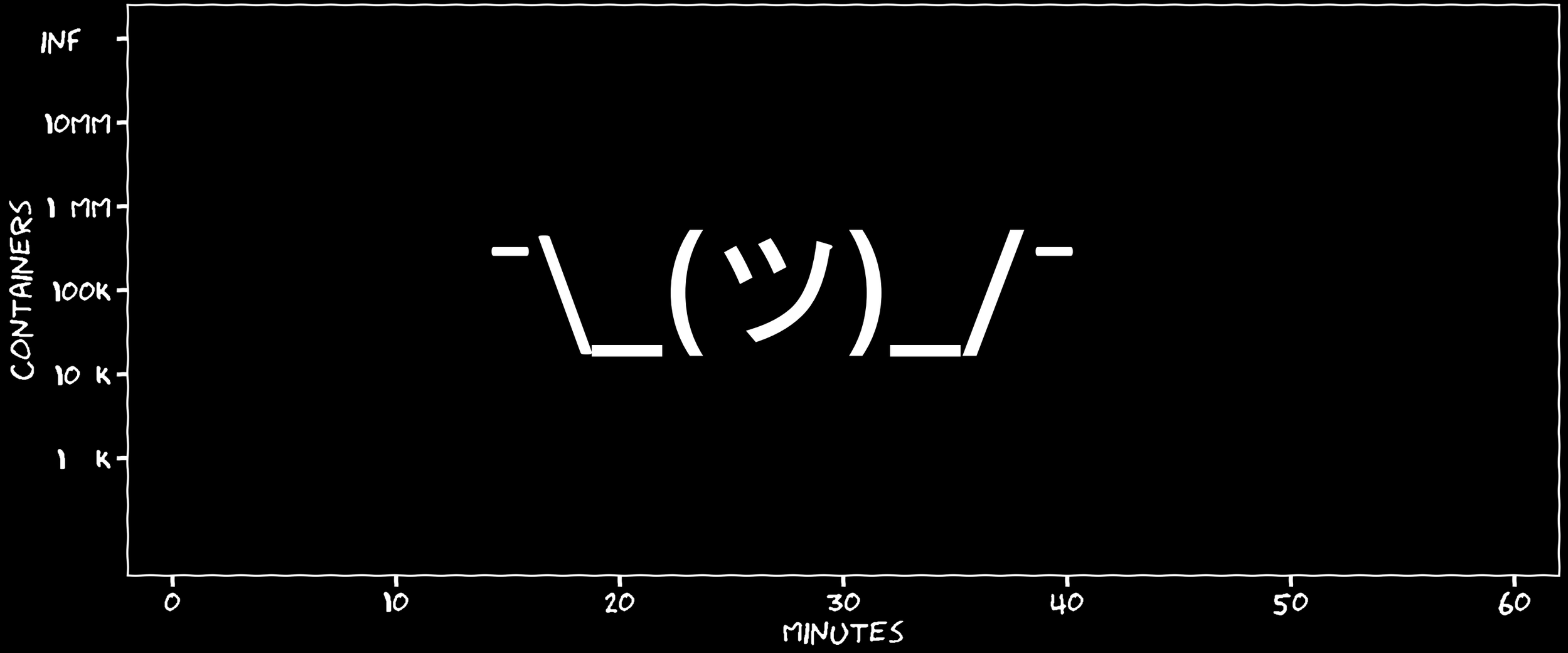
Hundreds of containers



Thousands of containers



Millions of containers



Closing thoughts



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Closing thoughts

- AWS is continually improving

Closing thoughts

- AWS is continually improving
- Amazon ECS on AWS Fargate is a great starting point

Closing thoughts

- AWS is continually improving
- Amazon ECS on AWS Fargate is a great starting point
- Kubernetes and Amazon EKS are more customizable

Closing thoughts

- AWS is continually improving
- Amazon ECS on AWS Fargate is a great starting point
- Kubernetes and Amazon EKS are more customizable

- AWS Fargate is outshining Amazon EC2 for general workflows

Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.