
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

CON407

Scaling containers from one user to millions

Abhishek Nautiyal

Senior Product Manager, ECS
AWS

Maish Saidel-Keesing

Senior Developer Advocate, Containers
AWS

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Amazon ECS overview

Scaling considerations

Troubleshooting API throttling

Performance optimization tips

Sample application

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon ECS overview

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Elastic Container Service (Amazon ECS) overview



Amazon EC2

AWS Regions, AWS Wavelength, AWS Local Zones, AWS Outposts

AWS Fargate

Serverless container compute engine

Amazon ECS Anywhere

Customer-owned container hosts

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon ECS – Unprecedented scale



Over 2.25 billion

Tasks launched every week

Tens of thousands

API requests served per second

27 AWS Regions, 6 continents

Global presence

65% of all new AWS containers customers

Use Amazon ECS

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

It’s not just horizontal scale

5.48M concurrent vCPUs

Largest single-account production workload

750/sec task placement rate

RunTask API (accounts with increased limits)

GPU and high memory workloads

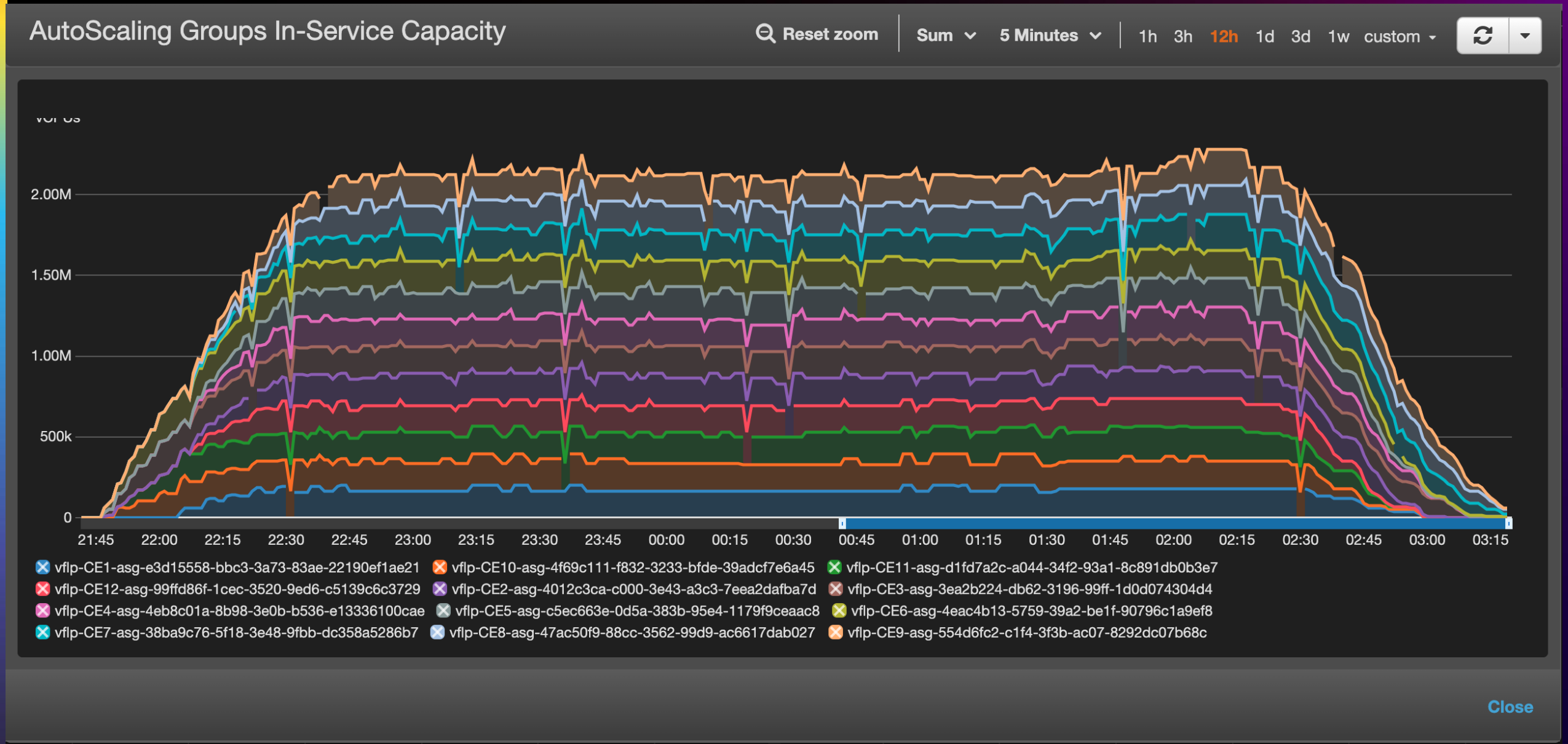
Fastest growing Amazon EC2 instance usage



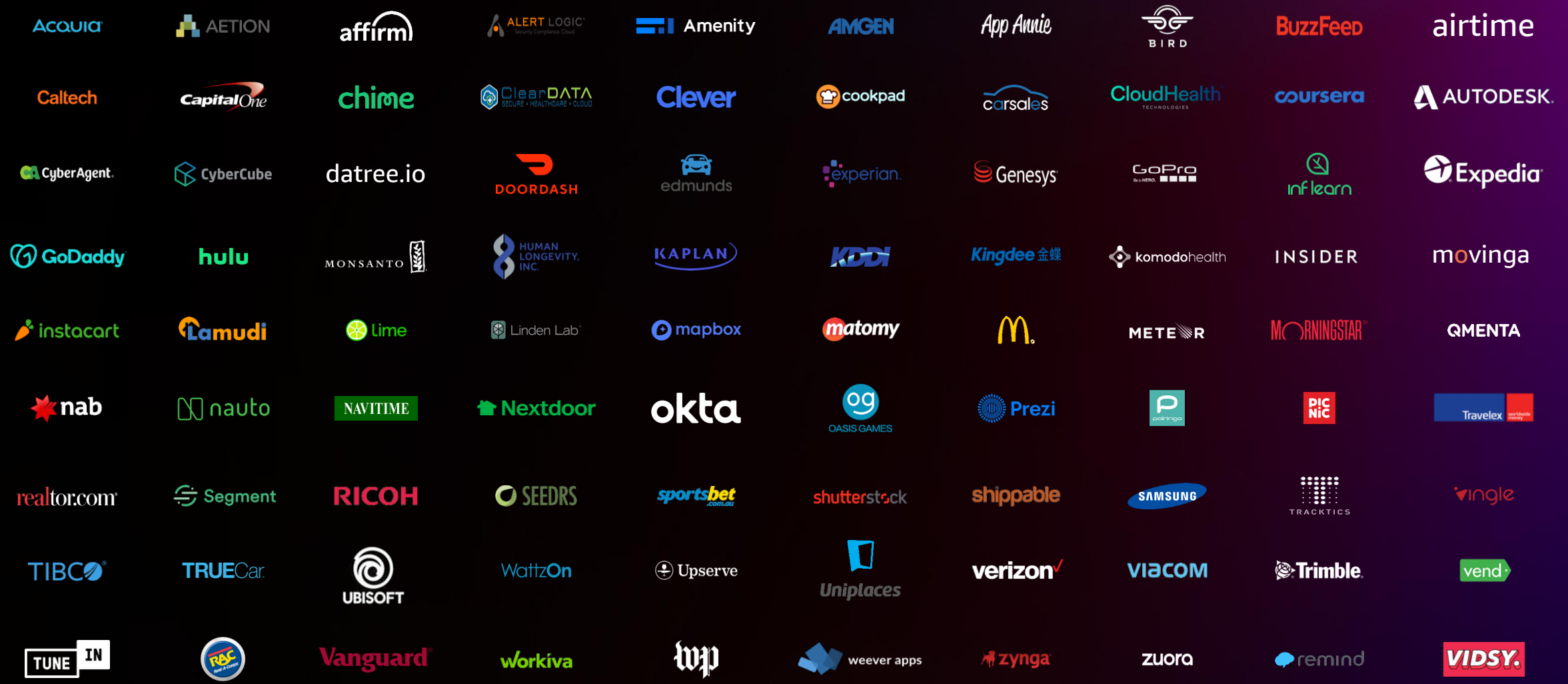
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What powerful can do



Customers love ECS for its powerful simplicity



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon services innovate with Amazon ECS


Amazon Redshift


Amazon OpenSearch Service


Amazon RDS


Amazon Transcribe


AWS Elemental MediaLive


Amazon.com recommendation


Amazon Lex


Amazon Polly


AWS Batch


Amazon SageMaker


Amazon Translate

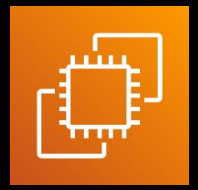
Amazon ECS delivers security, reliability, availability, and scale

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon ECS scaling and performance

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling improvements

Compute type	Task launch rate	Speed up vs. 2021	Example
 Fargate	500 tasks per min	16X	AWS Fargate on-demand, load-balanced service Launch 1,000 tasks in 4 m 12 s
 Amazon EC2	500 tasks per min	4X	Amazon EC2 capacity, with all the EC2 instances up already Launch 1,000 tasks in 4 m 30 s

Full blog: aws.amazon.com/blogs/containers/under-the-hood-amazon-elastic-container-service-and-aws-fargate-increase-task-launch-rates

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

I am building a service that needs to scale to handle millions of users during peak hours

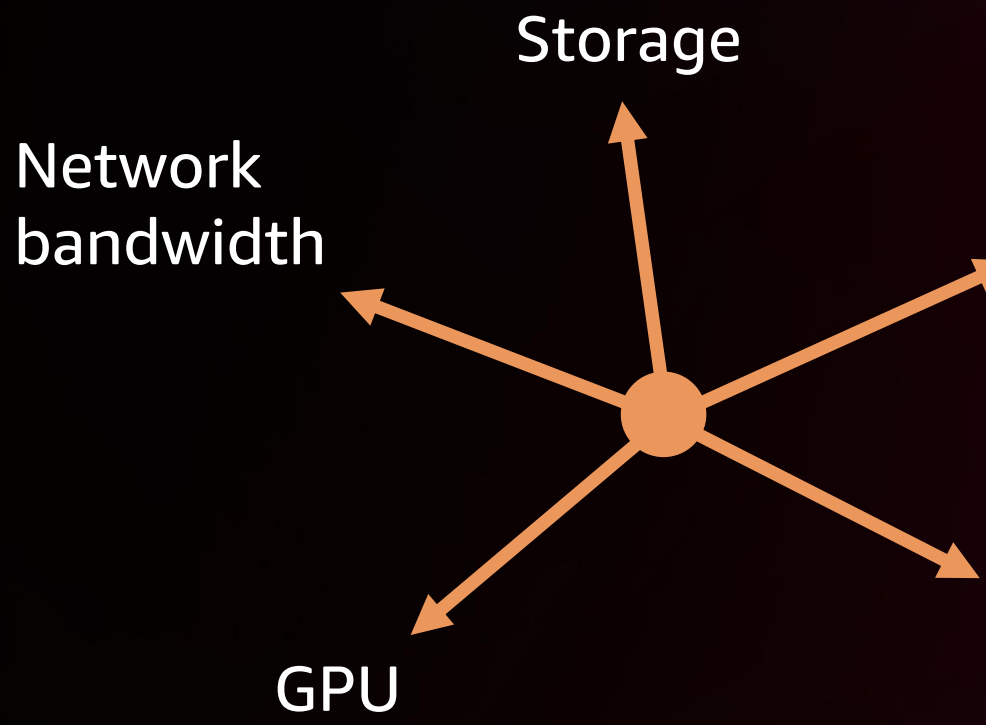
How do I build on Amazon ECS to achieve this scale?



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

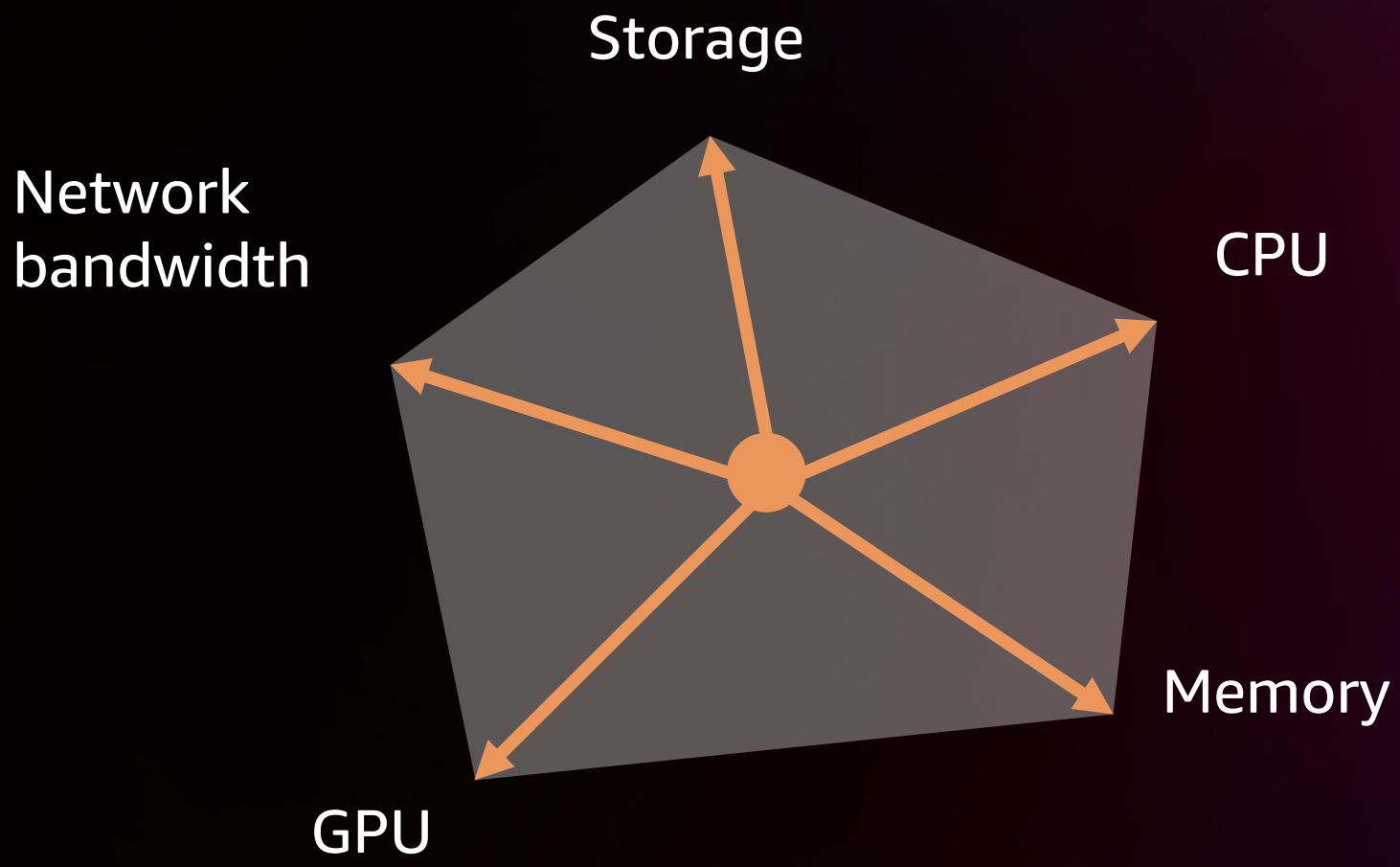
Translate to application-specific scaling units

Identify the required resources



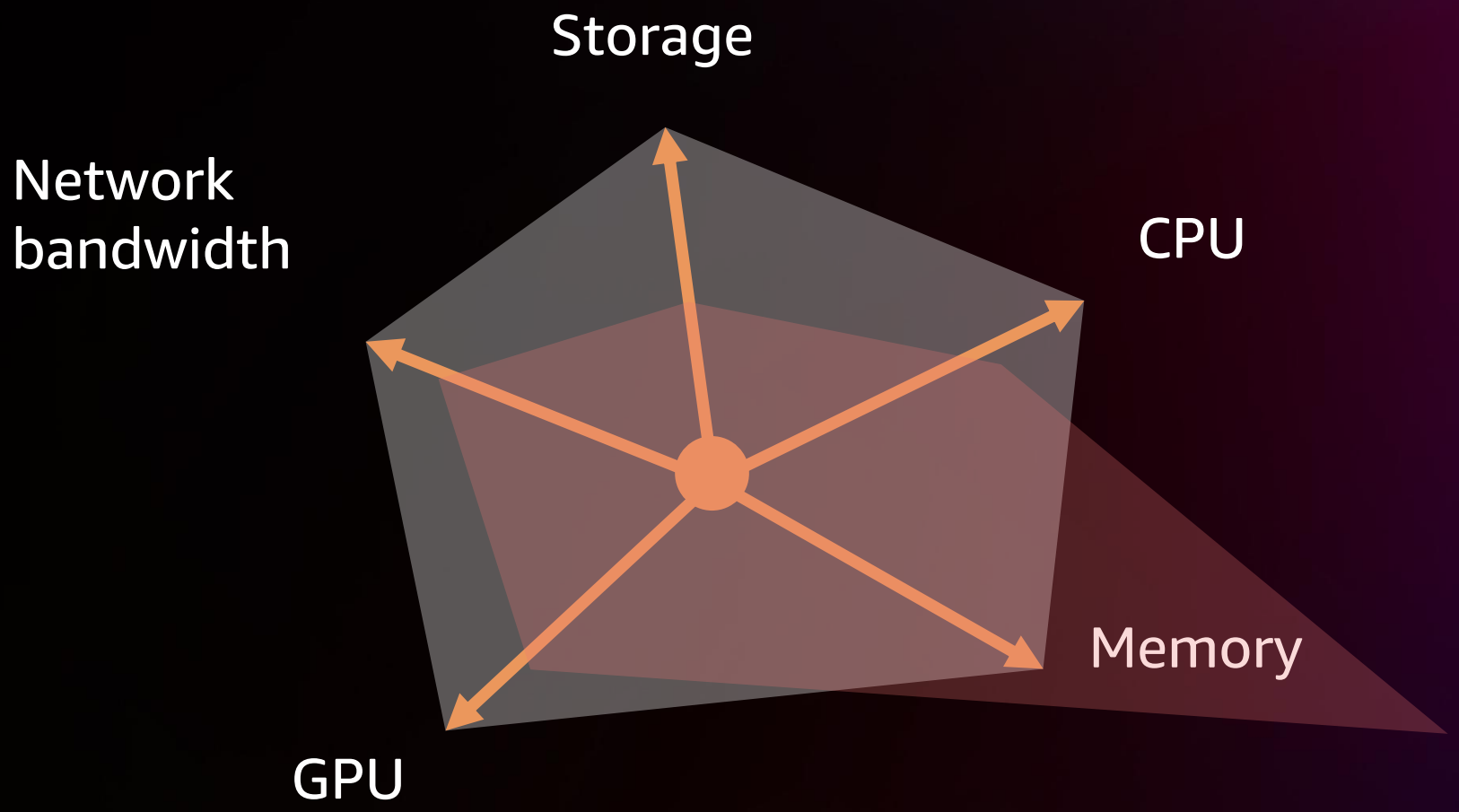
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Profile your application



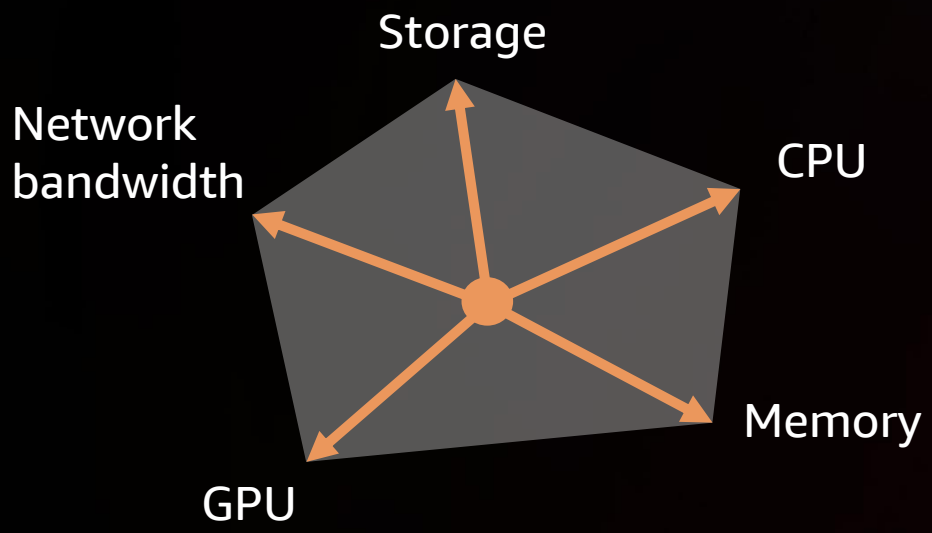
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

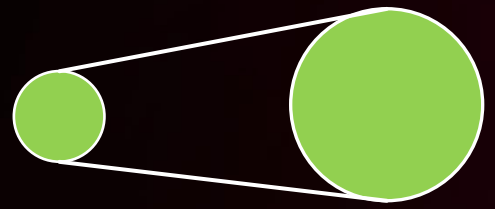
Performance degradation



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Map to Amazon ECS units







Task

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Choosing the right compute capacity

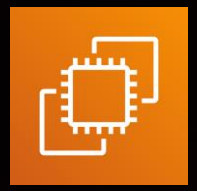
RUN ANYWHERE, BUT START WITH AWS FARGATE

Serverless



AWS Fargate

Amazon EC2 options



Amazon EC2



Spot instance

Edge and 5G



AWS Local Zones



AWS Wavelength

On premises

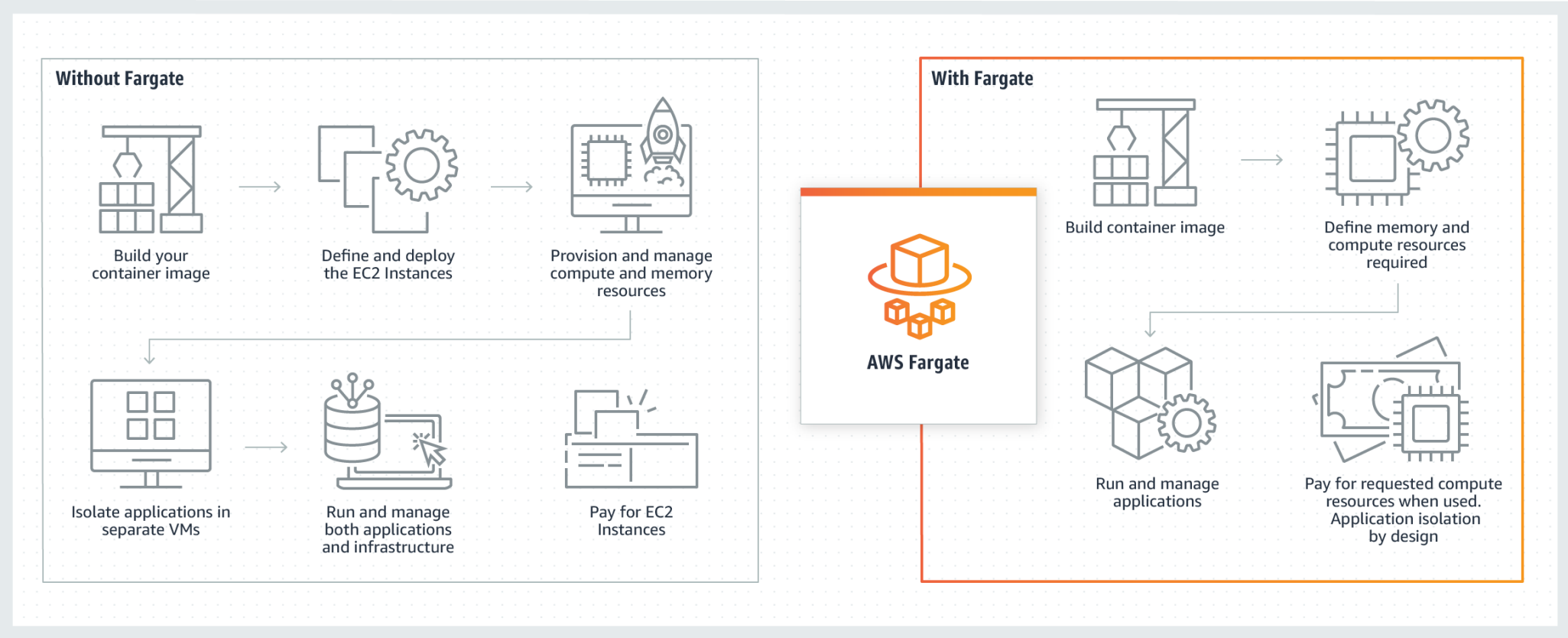


AWS Outposts



Amazon ECS Anywhere

Start with AWS Fargate



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

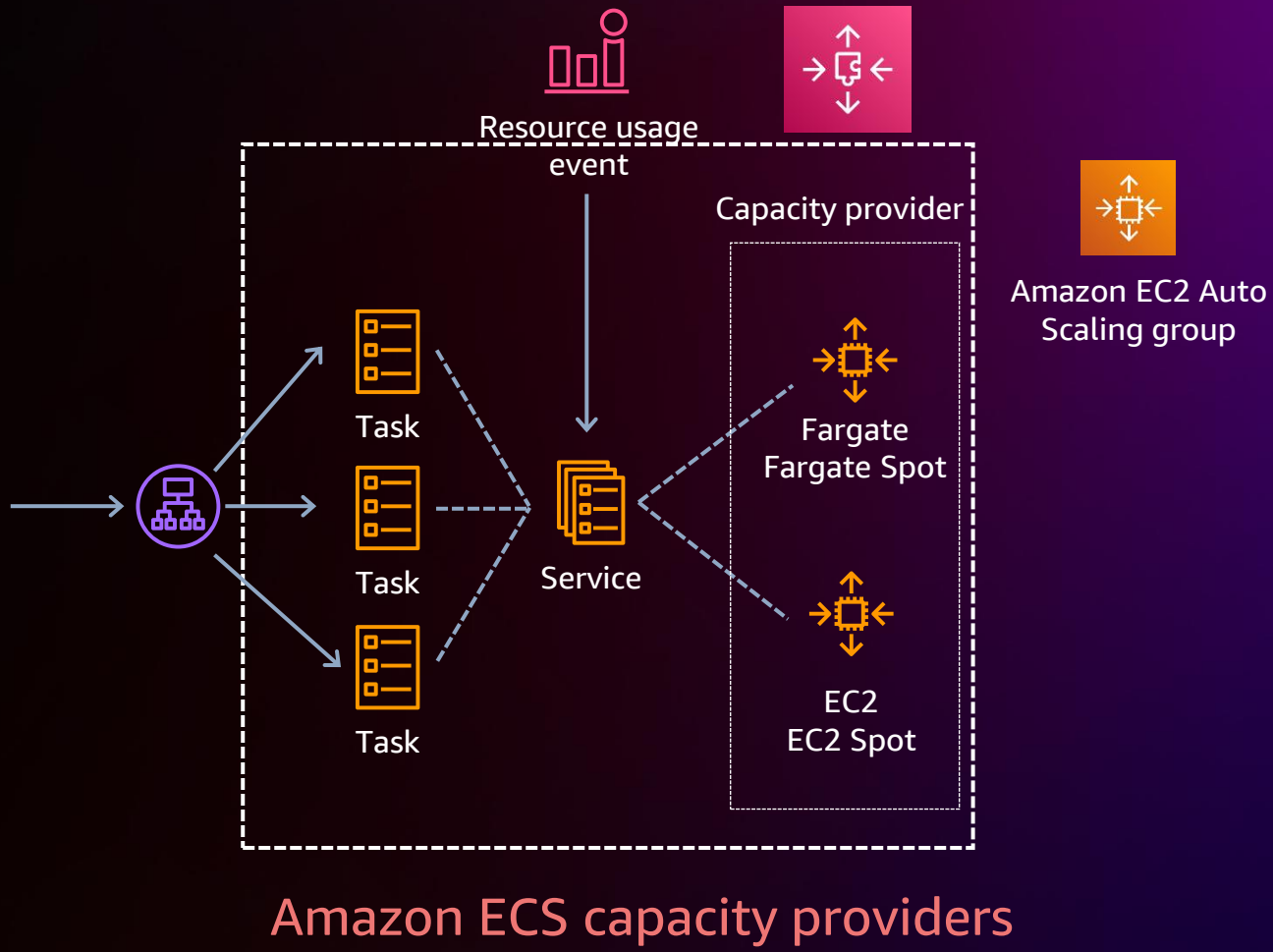
Amazon ECS capacity providers for high-performance apps

Scale application automatically as needed

Pay for what you provision

Spot is 50–70% of on-demand cost

Available with Compute Savings Plans



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ECS Scaling Considerations

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon ECS service quotas at a glance



Account quotas

Number of clusters per account



Cluster quotas

Container instances per cluster
services per cluster

Tasks in PROVISIONING state per cluster



Service quotas

Number of tasks per service
Rate of tasks launched by a service



Task quotas

16 vCPU ad 120 GiB memory for Fargate tasks

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

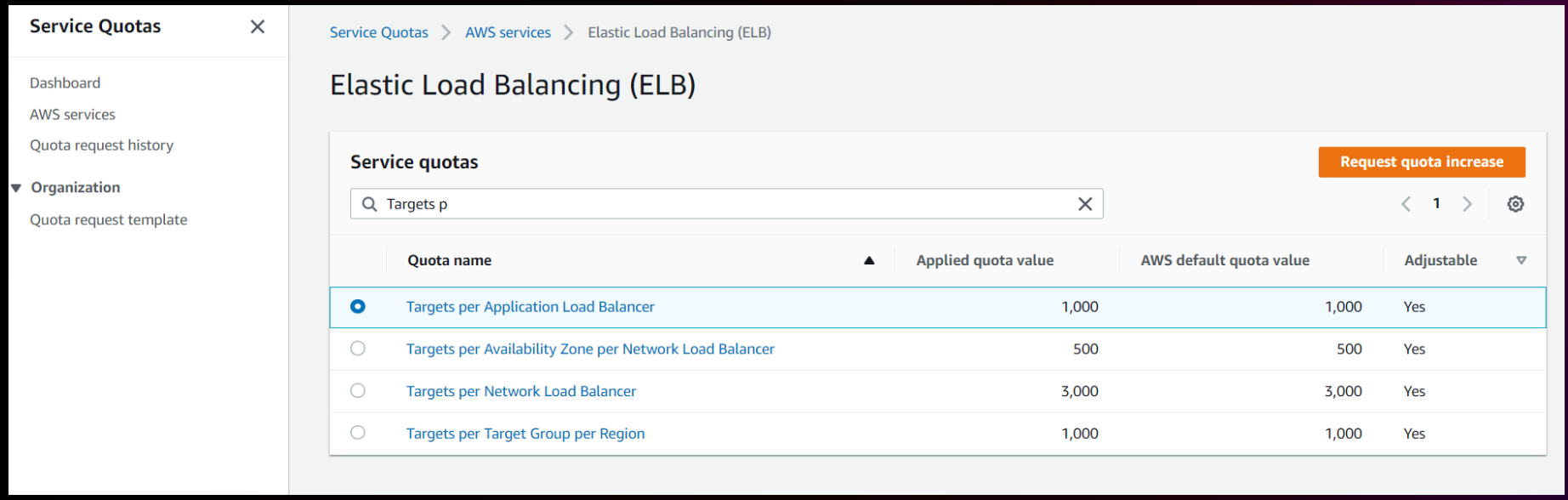
Amazon ECS architecture – service quotas

Service	Quota	Default value
Application Load Balancer (ALB)	Targets per ALB	1,000
Application Load Balancer (ALB)	Targets per target group per Region	1,000
Network Load Balancer (NLB)	Targets per AZ per NLB	500
Elastic Network Interface (ENI)	Network interfaces per Region	5,000
Elastic Network Interface (ENI)	Elastic IP addresses per Region	5
Elastic Network Interface (ENI)	Maximum network interfaces	Depends on type of instance AWSEC2/latest/UserGuide/using-eni.html#AvailableIpPerENI
AWS Cloud Map	Instances per service	1,000

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Increasing service quotas

- You can easily request increase in default quotas using service quotas console



- Contact AWS Support for recommendations if a quota is not adjustable on Console

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

API throttling

- **Synchronous**: when your application generates API requests beyond the default throttling limits, e.g., submitting too many RunTask API request in a second

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

API throttling

Asynchronous: when Amazon ECS generates API requests on your behalf, e.g., CreateNetworkInterface to attach an ENI to your tasks


```
{
   "userIdentity":{
      "arn":"arn:aws:sts::111122223333:assumed-role/AWSServiceRoleForECS/ecs-service-scheduler",
      "eventTime":"2022-03-21T08:11:24Z",
      "eventSource":"elasticloadbalancing.amazonaws.com",
      "eventName":" DescribeTargetHealth ",
      "awsRegion":"us-east-1",
      "sourceIPAddress":"ecs.amazonaws.com",
      "userAgent":"ecs.amazonaws.com",
      "errorCode":"ThrottlingException",
      "errorMessage":"Rate exceeded",
      "eventID":"0aeb38fc-229b-4912-8b0d-2e8315193e9c"
   }
}
```


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Monitoring throttling events

- Visualize your quotas and set alarms with CloudWatch
- Send throttling events with CloudTrail to CloudWatch Logs
- Parse and analyze these events using CloudWatch log Insights to identify details such as the user or IAM role initiating the API calls and the number of API calls
- Use Amazon Athena to query and analyze data using standard SQL



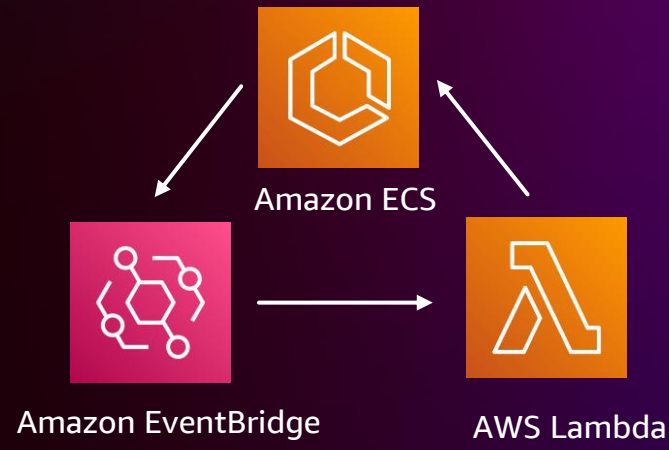

```
select eventname, errorcode,eventsource,awsregion, useragent,COUNT(*) count
FROM cloudtrail-table-name
where errorcode = 'ThrottlingException'
AND eventtime between '2022-01-14T03:00:08Z' and '2022-01-23T07:15:08Z'
group by errorcode, awsregion, eventsource, username, eventname
order by count desc;
```

Contact AWS Support to raise limits or explore alternatives if you see frequent throttling errors

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Building retry mechanisms

- Amazon ECS emits event notifications to Amazon EventBridge
- There are resource state change events and service action events including API throttling events such as ECS_OPERATION_THROTTLED and SERVICE_DISCOVERY_OPERATION_THROTTLED
- Monitor Amazon ECS service event notifications on EventBridge and build retry (with backoff) mechanisms, e.g., using a Lambda function in response to throttling error



Note: Amazon ECS services and AWS SDK automatically have inbuilt retry mechanisms

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

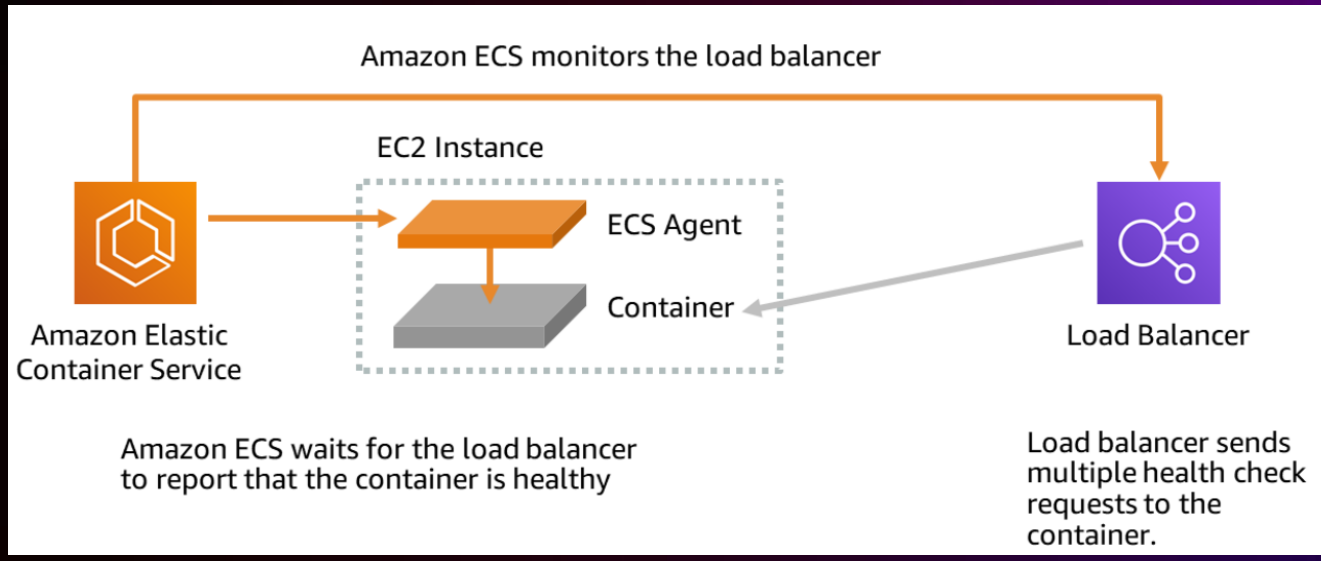
Performance optimization tips

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Faster deployments #1

Speed up load balancer health checks

- Reduce HealthCheckIntervalSeconds from default 30 s
- Reduce HealthyThresholdCount from default 5 count
- If your service starts and stabilizes under 10 s, you can reduce health check process from default 2 m 30 s (30 • 5) to 10 s by configuring 5 s and 2 s



Speed up load balancer connection draining

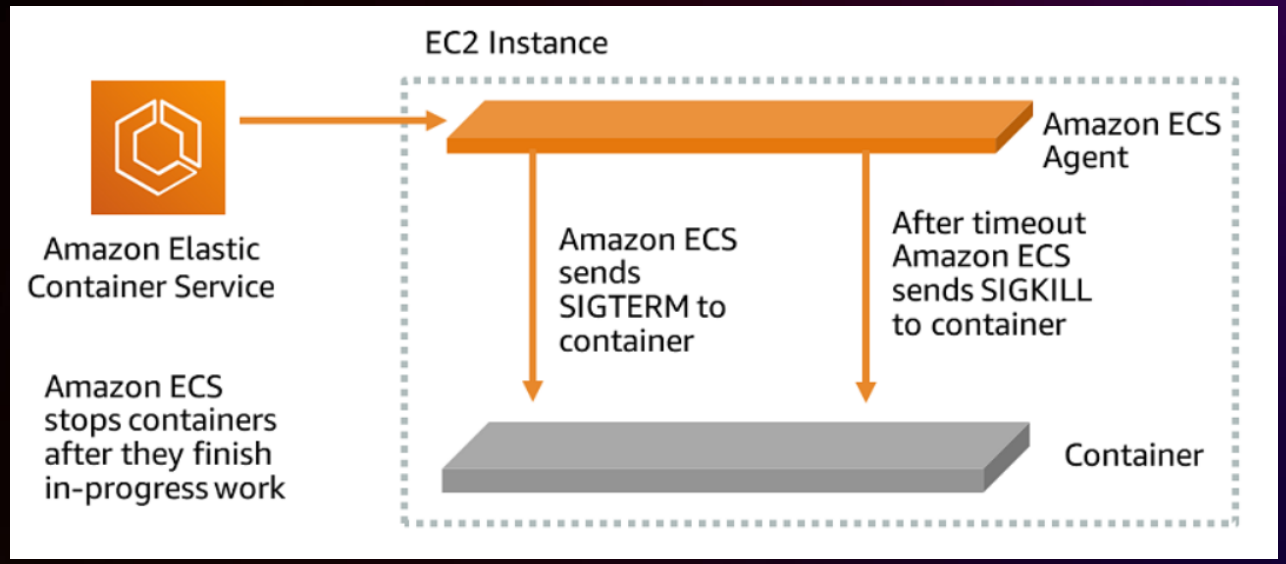
- reduce deregistration delay from the default 5 m for highly responsive services

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Faster deployments #2

Increase SIGTERM responsiveness

- reduce the default ECS_CONTAINER_STOP_TIMEOUT (30 secs) to a lower value to speed up the waiting period if your application can finish its work in less time
- For long-running tasks, you can use the task scale-in protection feature to protect them from being terminated

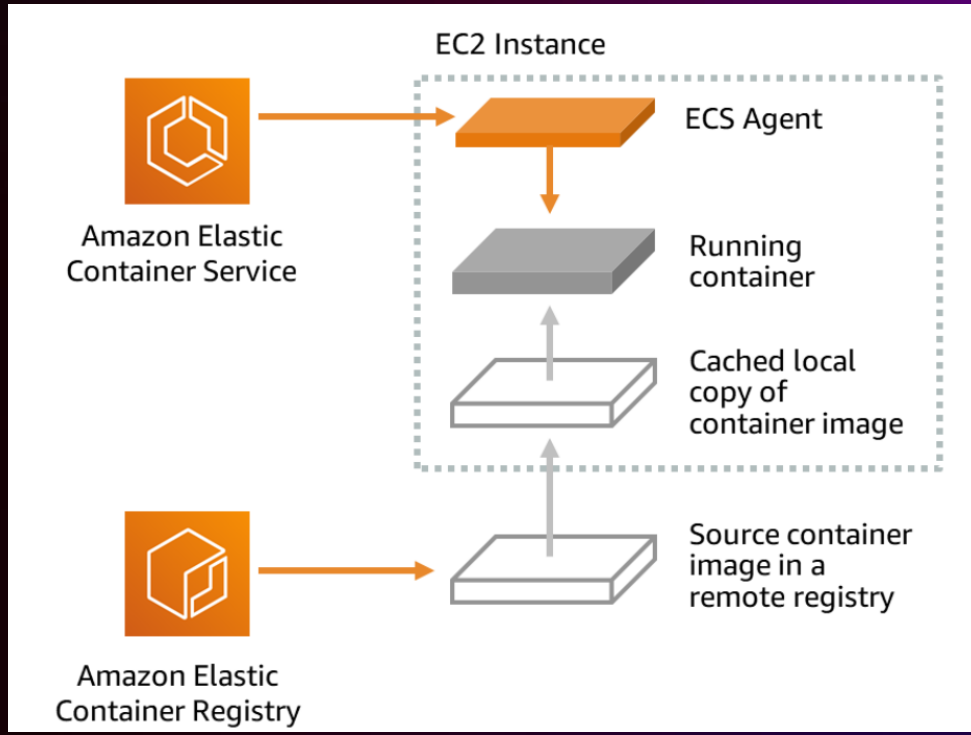


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Container caching

Container image pull behavior

- For Amazon EC2, cache container images
  - Set Amazon ECS agent parameter ECS_IMAGE_PULL_BEHAVIOR to once or prefer-cached
- For Fargate
  - Use larger task sizes with additional vCPUs to help reduce the time to extract an image during task launch
  - Use smaller base images
  - Keep the repository that stores the image in the same Region as the task

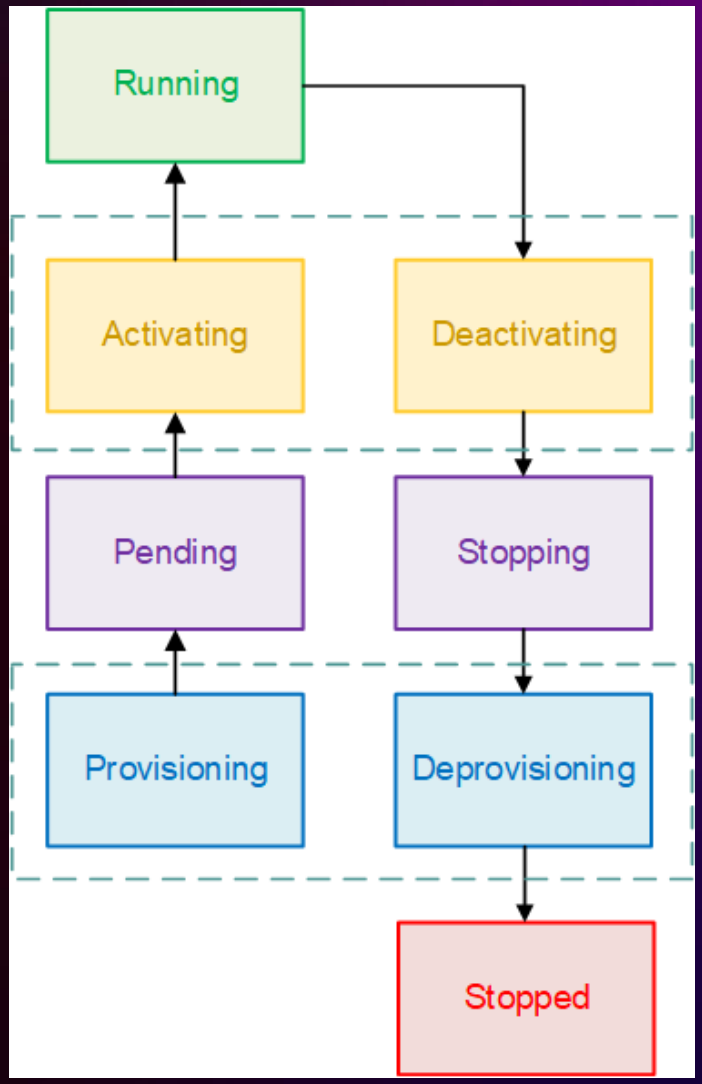


aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Speeding up task launch

Track your task lifecycle to find optimization opportunities

- Launching container image, running start-up scripts, and other configurations during application start-up can take a surprising amount of time
- Use AWS ECS Task Metadata endpoint to post metrics to track application start-up time from when container started to when your application is ready to serve traffic
- With this data, you can understand how your application contributes to total task launch time and find areas to reduce application-specific overhead and optimize your container images



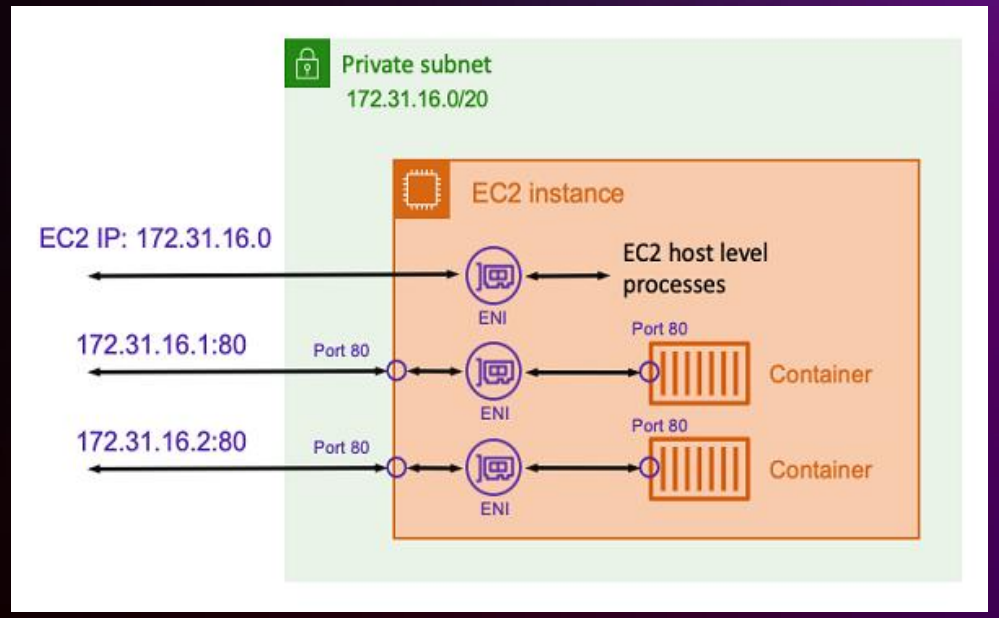
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

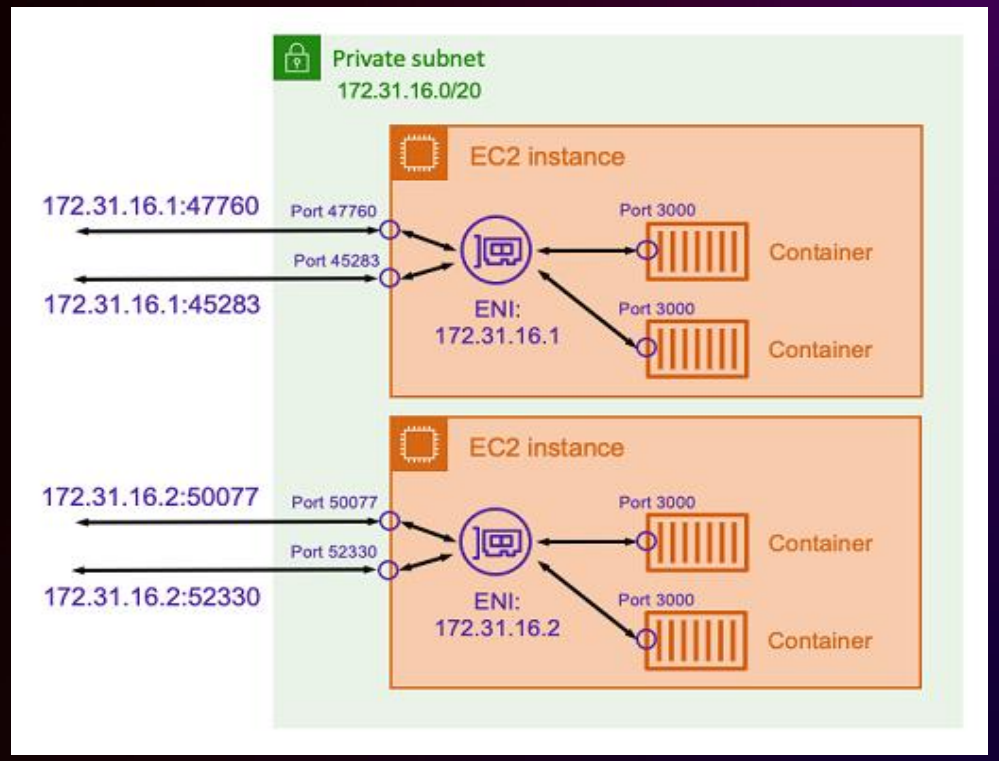
Network mode

Choose an optimal network mode

- With awsvpc mode, Amazon ECS creates a separate ENI for each task, thus each task gets its private IP address and VPC
- (+) greater flexibility to control communications between tasks at a more granular level
- (-) additional time spent in provisioning ENIs

Consider using bridge mode if speed of task launch outweighs the benefit of using awsvpc



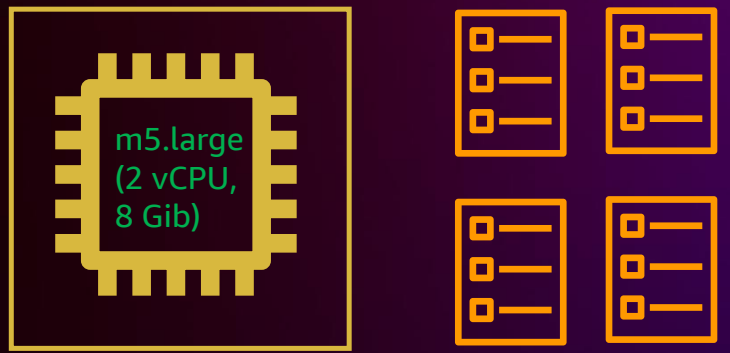


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Optimal instance type

Choose an optimal instance type when using Amazon ECS/Amazon EC2

- Calculate number of tasks that can be placed on a single instance based on resource reservation requirement (i.e., CPU, memory, ENI, GPU) of your tasks for instance sizing
- Use the binpack task placement strategy to increase task density on your Amazon EC2 instances
- When using the binpack placement strategy with the awsvpc network mode, consider using ENI trunking to increase number of tasks on each instance



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Sample application

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Application requirements

2vCPU, 2GB RAM – supports 200 customer requests

To get to a 1,000,000 requests – scale to 5,000 tasks

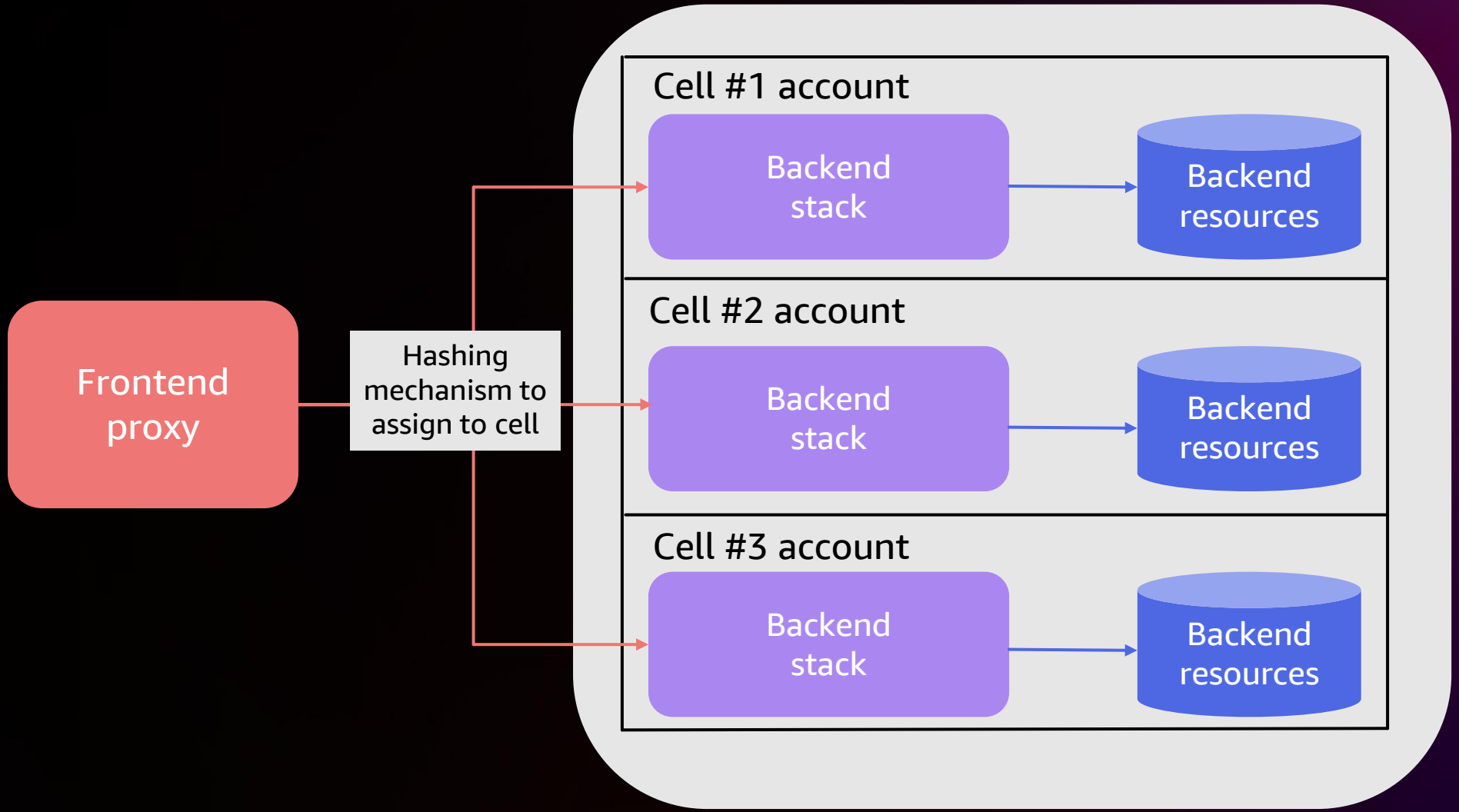
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Limits

- Tasks per service 5,000
- IP addresses /16 subnet
- vCPU limits
- Targets per ALB 1,000
- ENIs per Region 5,000
- AWS Cloud Map instances per namespace 1,000
- AWS Cloud Map instances per service 1,000

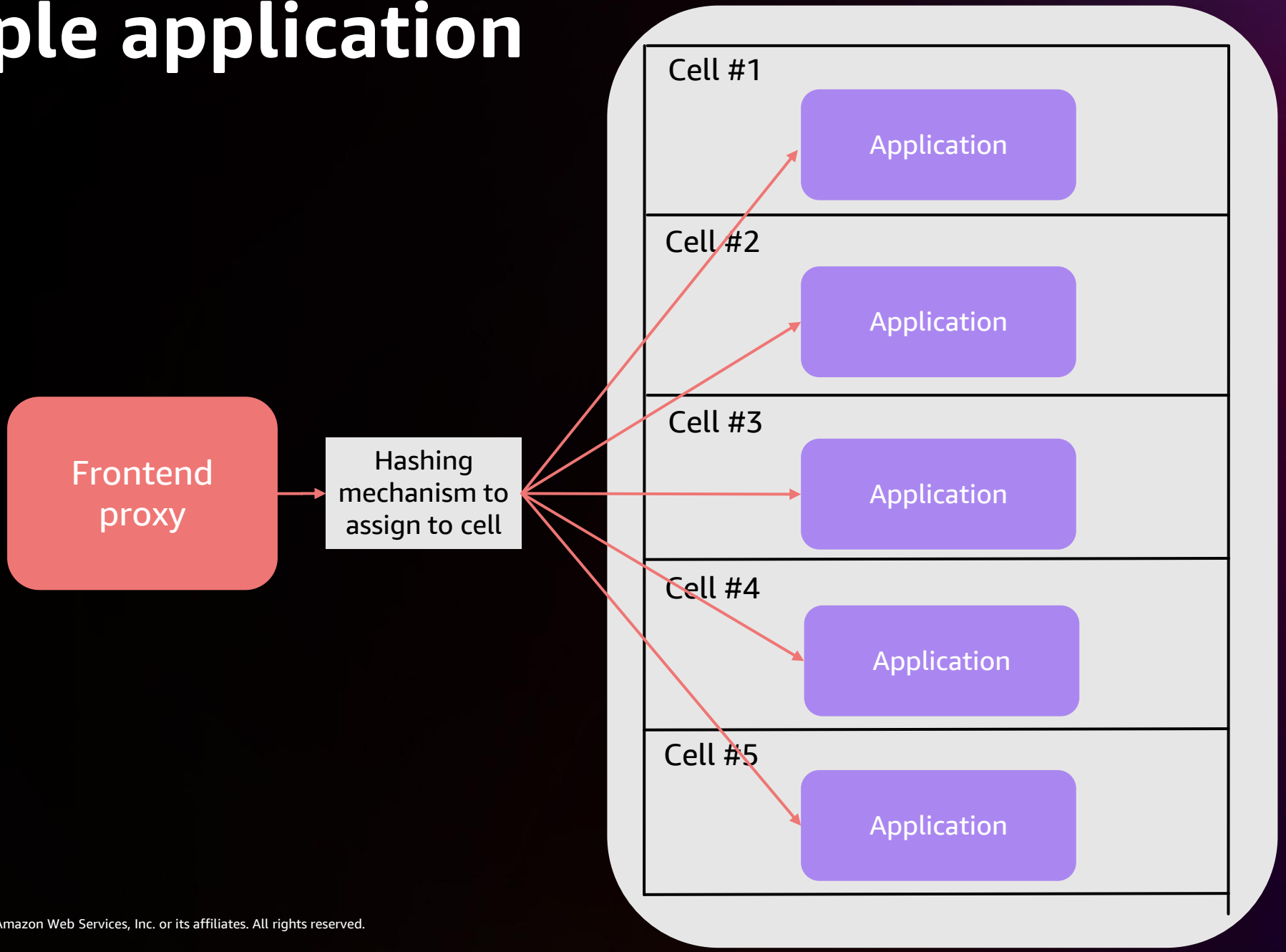
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cellular architecture



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Sample application



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Resources

- aws.amazon.com/blogs/containers/under-the-hood-amazon-elastic-container-service-and-aws-fargate-increase-task-launch-rates
- docs.aws.amazon.com/AmazonECS/latest/bestpracticesguide/intro.html
- docs.aws.amazon.com/AmazonECS/latest/developerguide/service-quotas.html
- docs.aws.amazon.com/AmazonECS/latest/APIReference/request-throttling.html
- www.vladionescu.me/posts/scaling-containers-on-aws-in-2022

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Abhishek Nautiyal
nautiya@amazon.com
@abhishekn3k

Maish Saidel-Keesing
maishsk@amazon.com
@maishsk



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.