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

CON308

Reduce your operational and infrastructure costs with Amazon ECS

Vibhav Agarwal

Senior Product Manager
Amazon Web Services

Francis Pereira

VP, Infrastructure Engineering
CleverTap

aws

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

Preamble

Help you build and operate efficient containerized workloads that achieve business outcomes while allowing you to maximize return on investment

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

By running workloads efficiently on Amazon ECS, customers have saved

30%

vs. On-Demand EC2 costs

73%

vs. Lambda costs

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

Agenda

Overview of Amazon ECS

Optimizing your workload on Amazon ECS

- Optimize more
- Operate less
- Choose wisely

CleverTap’s journey to cost optimization with Amazon ECS

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

Amazon Elastic Container Service (Amazon ECS) overview



Amazon EC2

Regions, AWS Wavelength, 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 performance and scale overview



Over 2.25 billion

Tasks launched every week

Tens of thousands

API requests served per second

30 AWS Regions, 6 continents

Global presence

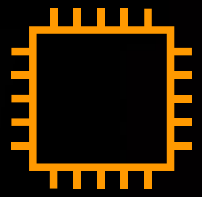
65% of all new AWS containers customers

Use Amazon ECS

aws

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

Fargate: A serverless compute engine for Amazon ECS



Host operating system

No more hours per month maintaining AMIs



Managed components

Amazon ECS agent, Amazon EFS storage drivers, logging "just work"



Local storage

20 GiB of ephemeral storage available to applications, configurable to 200 GiB



Hands-off patching

Infrastructure is always patched

aws

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

The Amazon ECS value proposition is the AWS value proposition

Scale and stability



+

Simplification and management

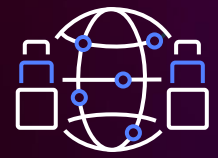


=

Cost savings



Higher availability



Faster time to production



aws

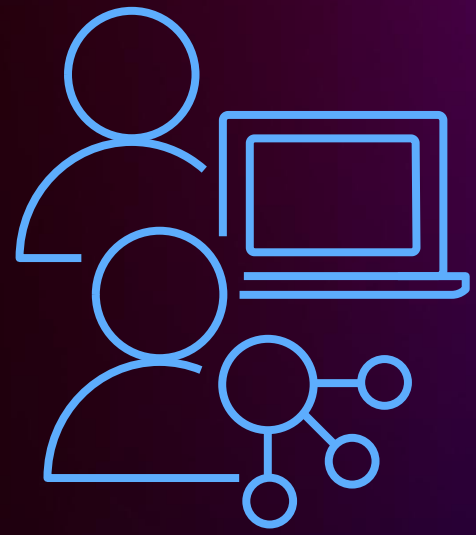
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon ECS: Your platform team

Opinionated and managed experiences
to reduce the cloud learning curve

Improve time to production

Reduce operations overhead

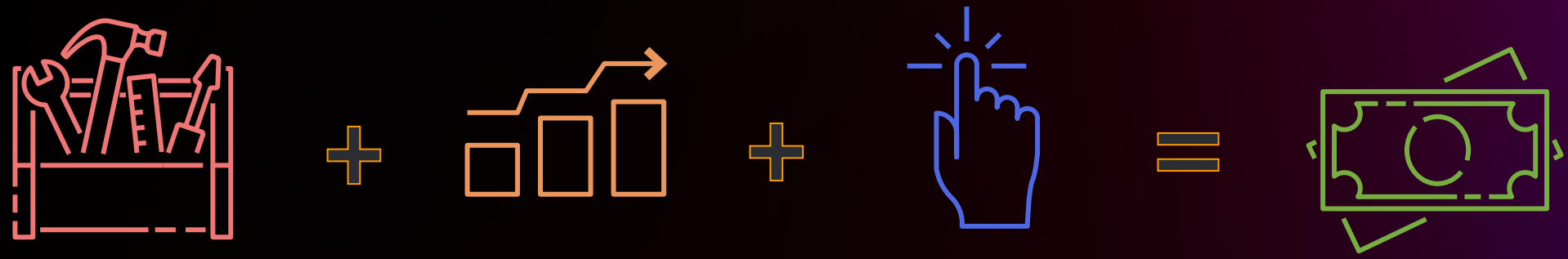


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Optimizing your container workloads with Amazon ECS

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

3 key principles to reduce costs



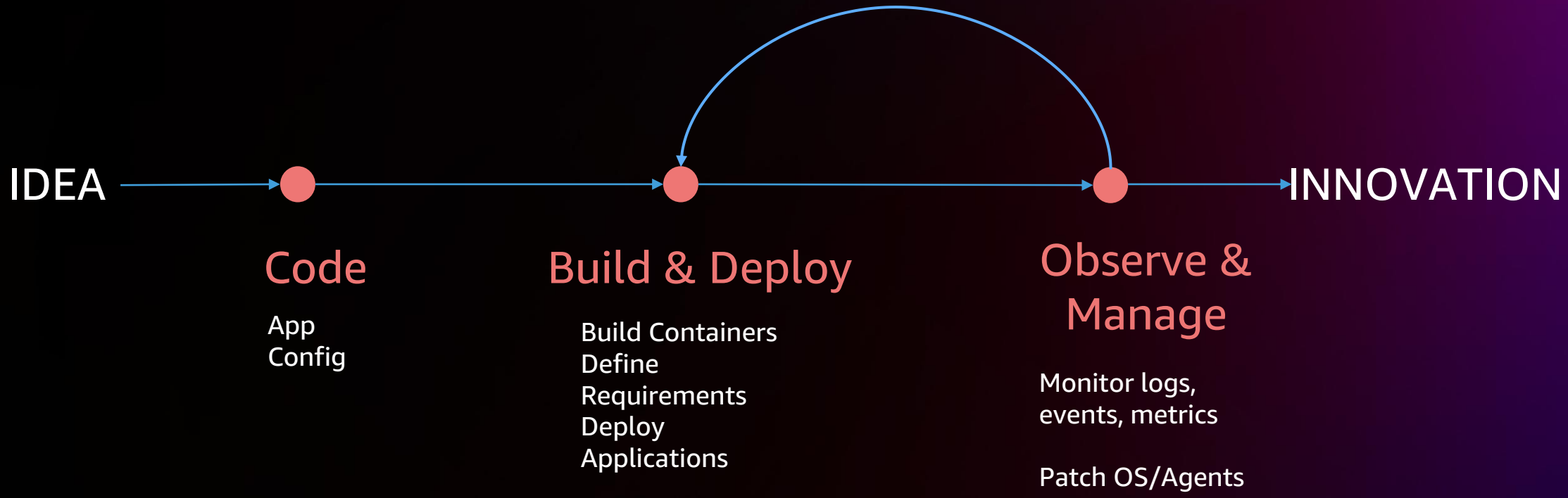
Operate Less + Utilize More + Choose Wisely =

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

Operate less

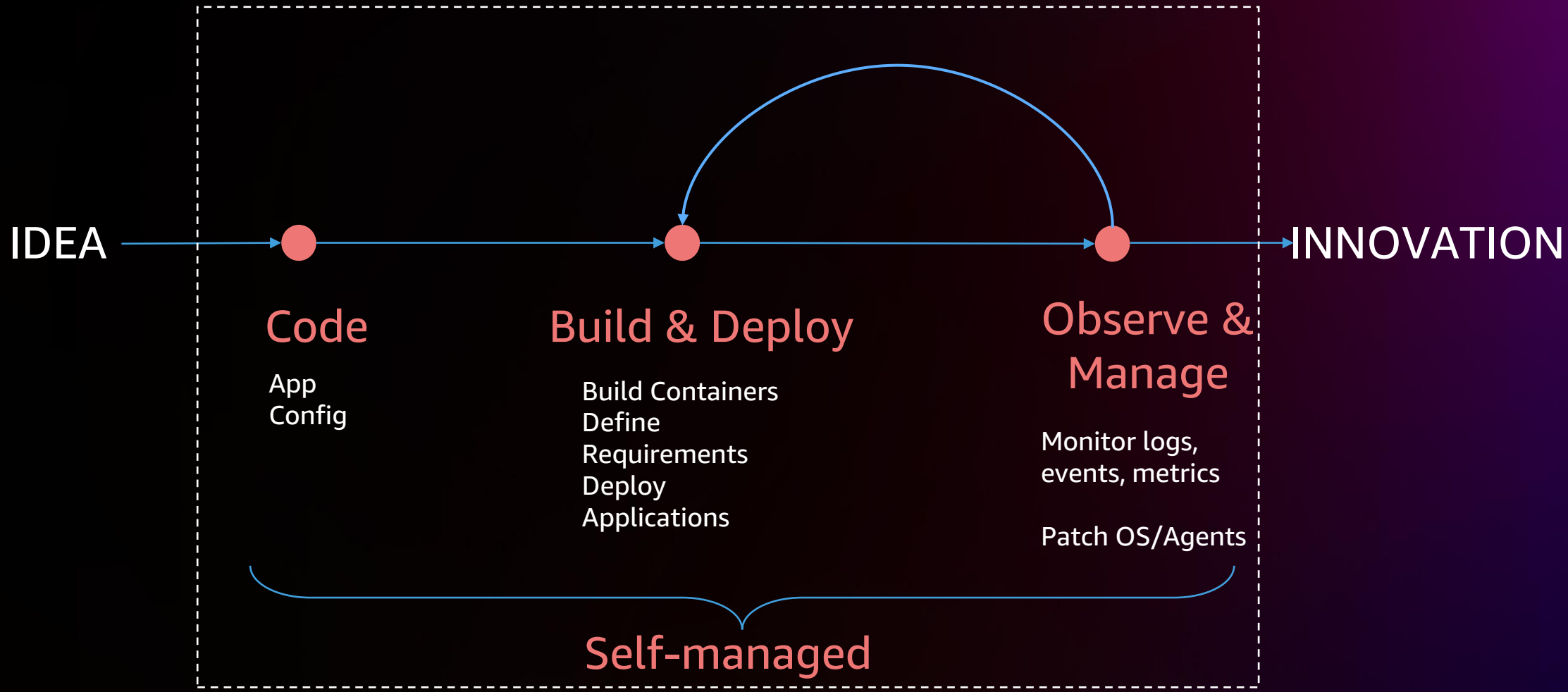
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Container workflow overview



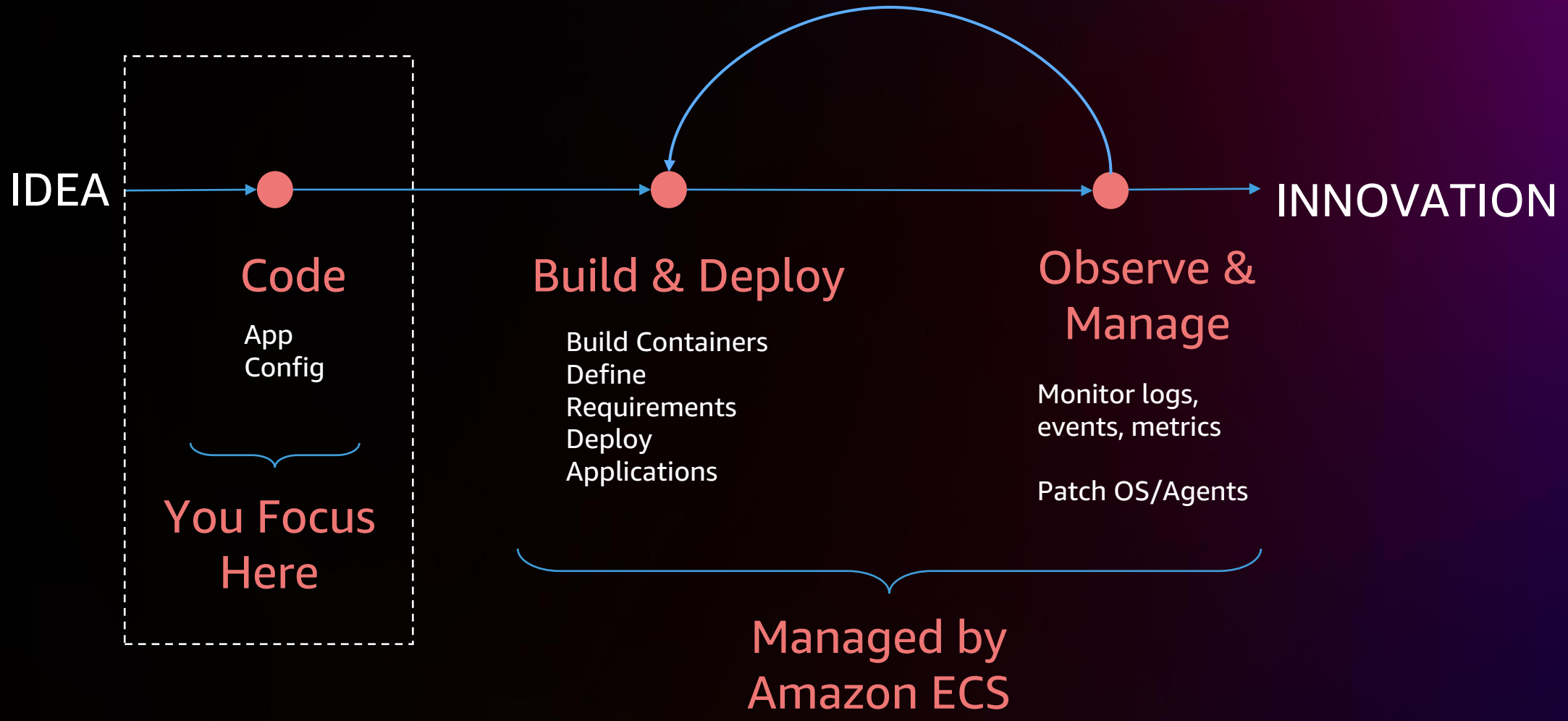
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

With most orchestrators you need to manage all of this and the orchestrator itself



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

Amazon ECS + Fargate enables you to focus on what matters



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

Amazon ECS is a serverless container orchestrator



Managed by AWS, no cluster versions, no data store to maintain or scale, free to use with Amazon EC2 or Fargate



Works with different “compute” types, including Fargate

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

Fargate is a serverless compute engine for containers



Managed by AWS, no AMIs to maintain, no Amazon EC2 instances to provision, scale or manage, pay for what you need

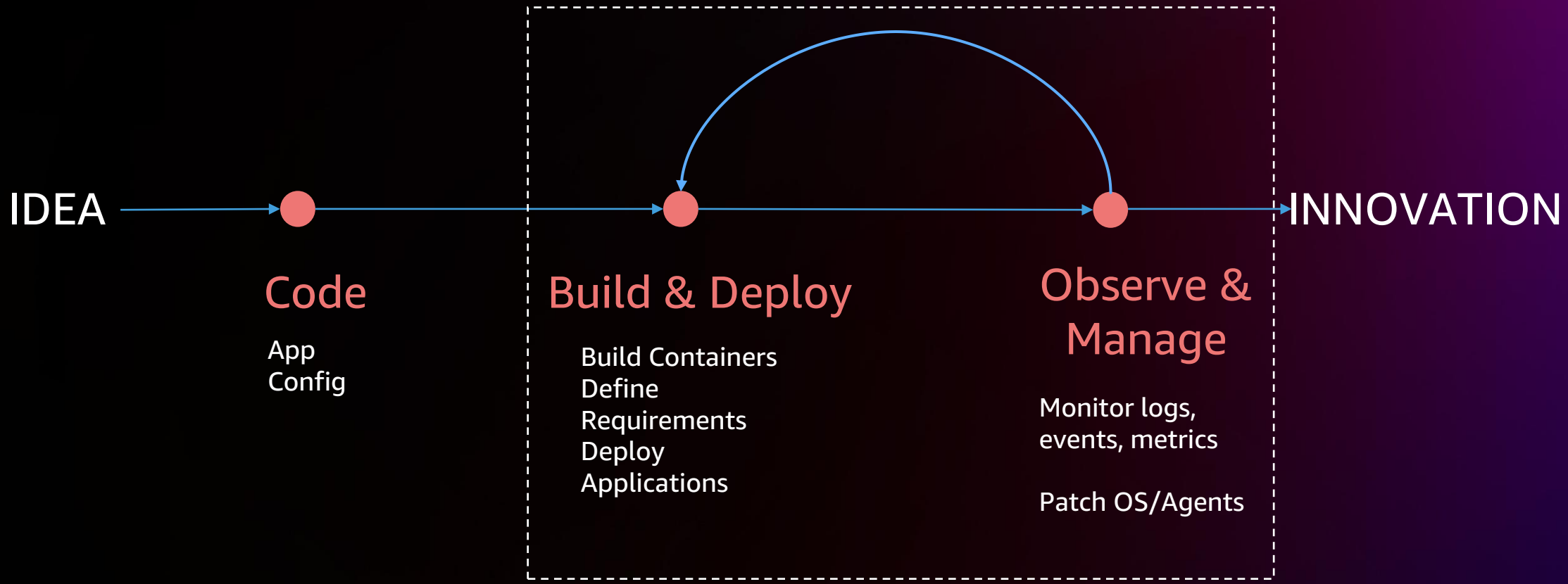


Fargate is where your containers will run. You don't interact with it directly

aws

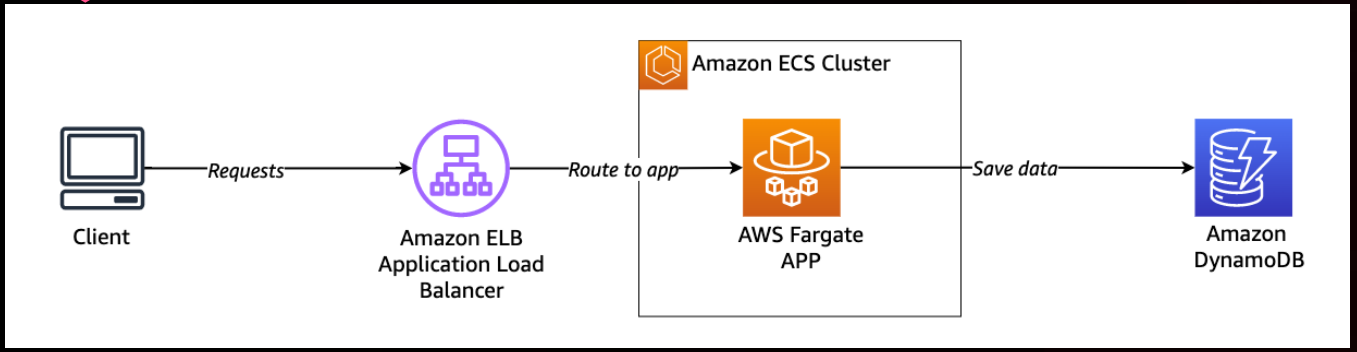
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Let’s go back and see how Amazon ECS manages all of this



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

Application-first interfaces to develop on Amazon ECS



1. Amazon ECS Blueprints

Terraform templates to deploy apps in minutes

2. Amazon ECS CDK extensions

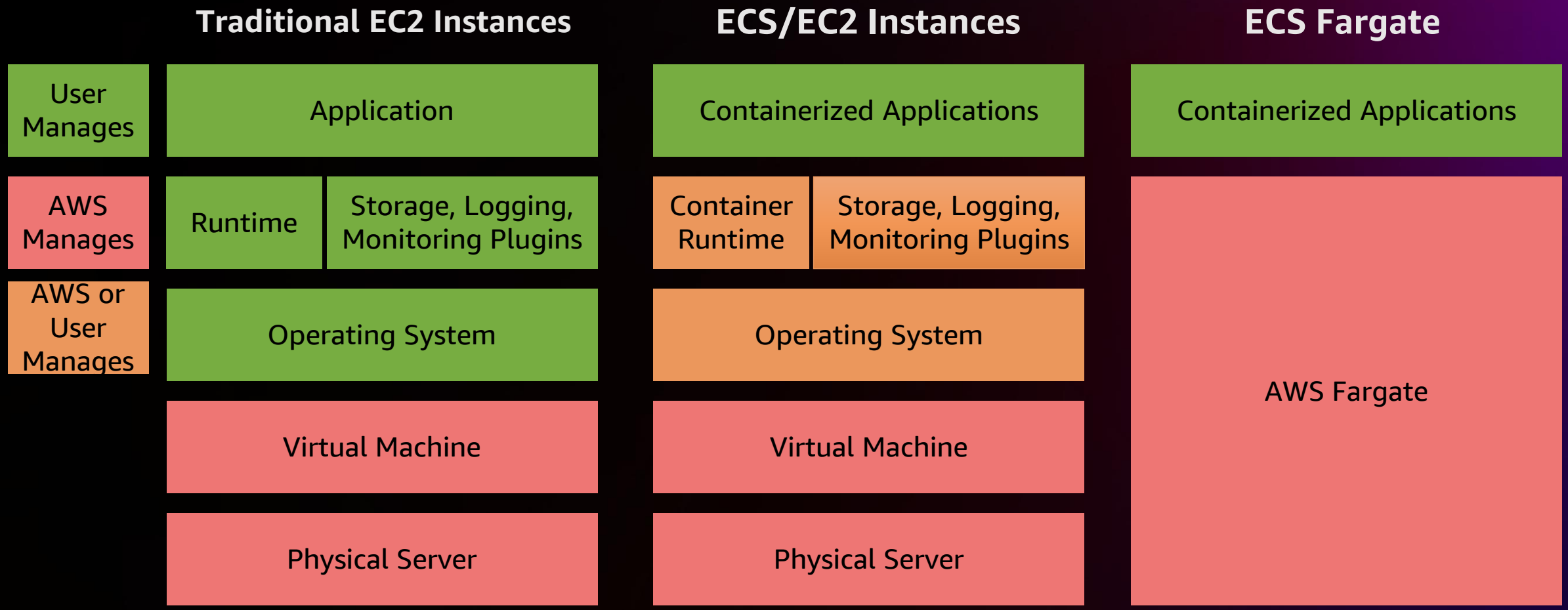
Program architecture patterns in your choice language

3. AWS Copilot CLI

Configuration-ready experience for common cloud architectures.
Preconfigured continuous delivery pipeline with auto scaling, metrics, alarms, and logs

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

Amazon ECS helps reduce your share of responsibility for infrastructure management

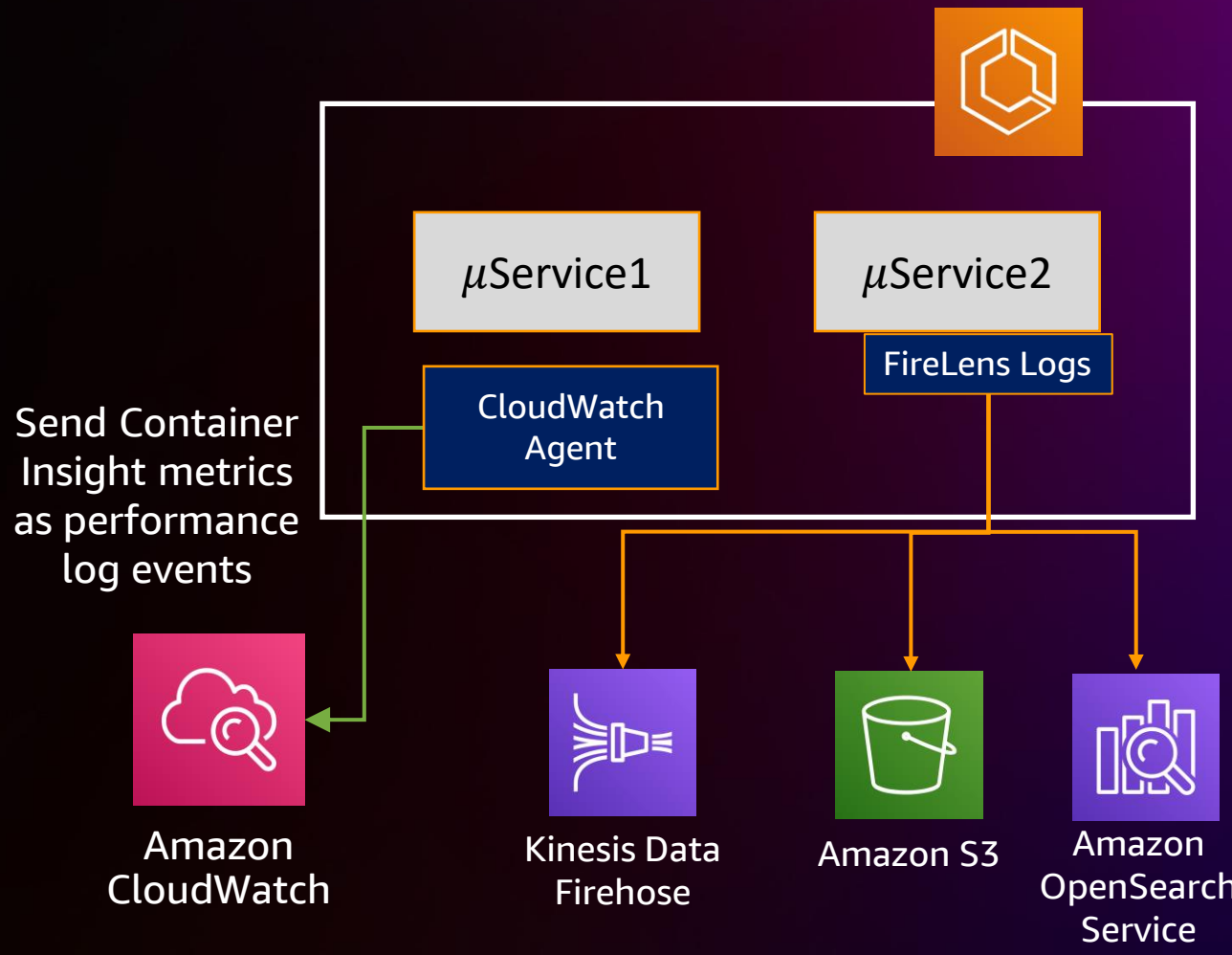


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon ECS has observability built-in

Collect, aggregate, and summarize metrics and logs with just a click with Amazon CloudWatch Container Insights

FireLens for Amazon ECS enables applications to send logs to many destinations by using the awsfirelens driver



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

But observability isn’t just about metrics and logs, you also need to monitor costs

Monitoring costs for containers is hard

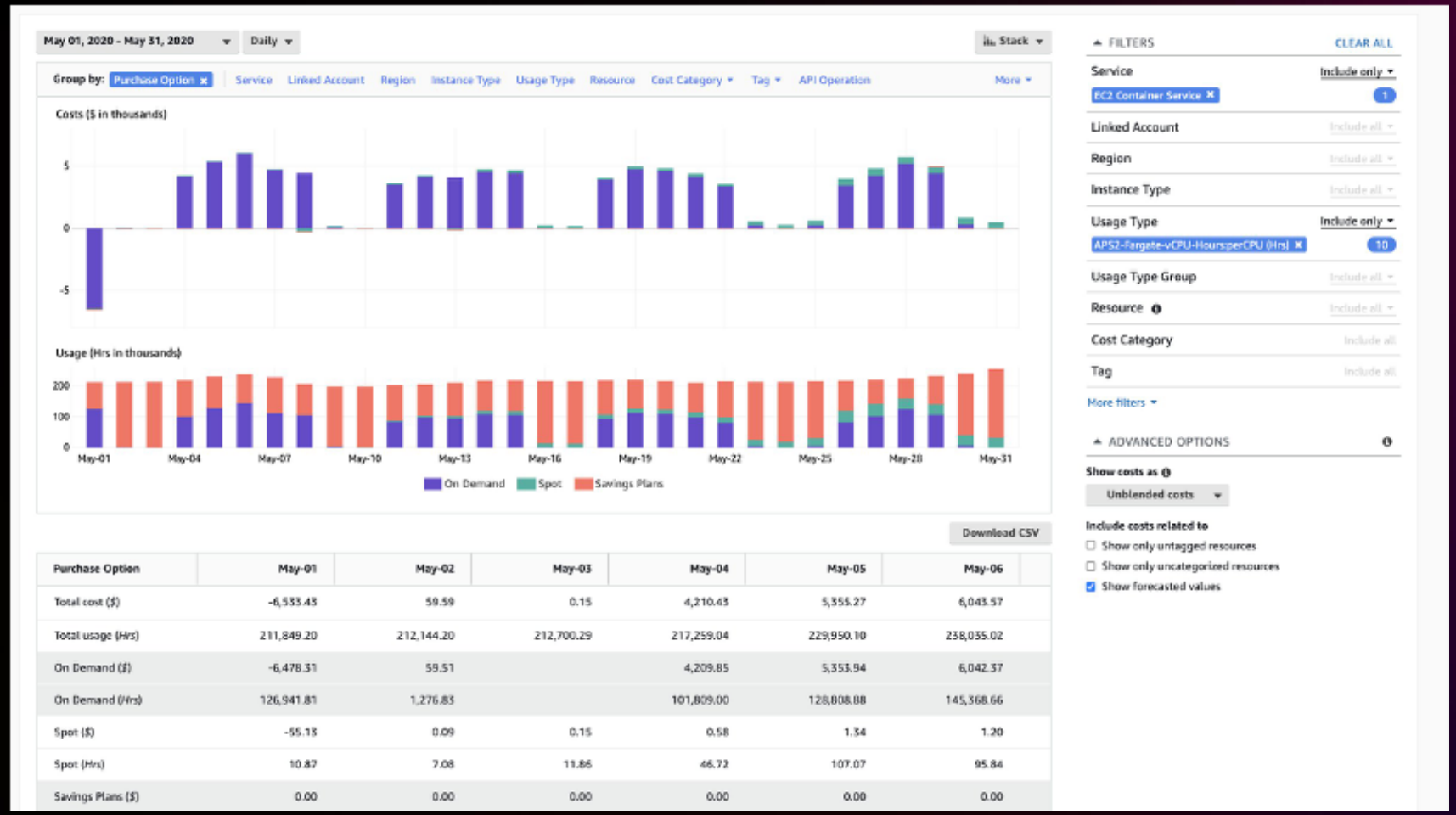
Need to distribute cost for shared, multi-tenant compute resources with high mutation rates



aws

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

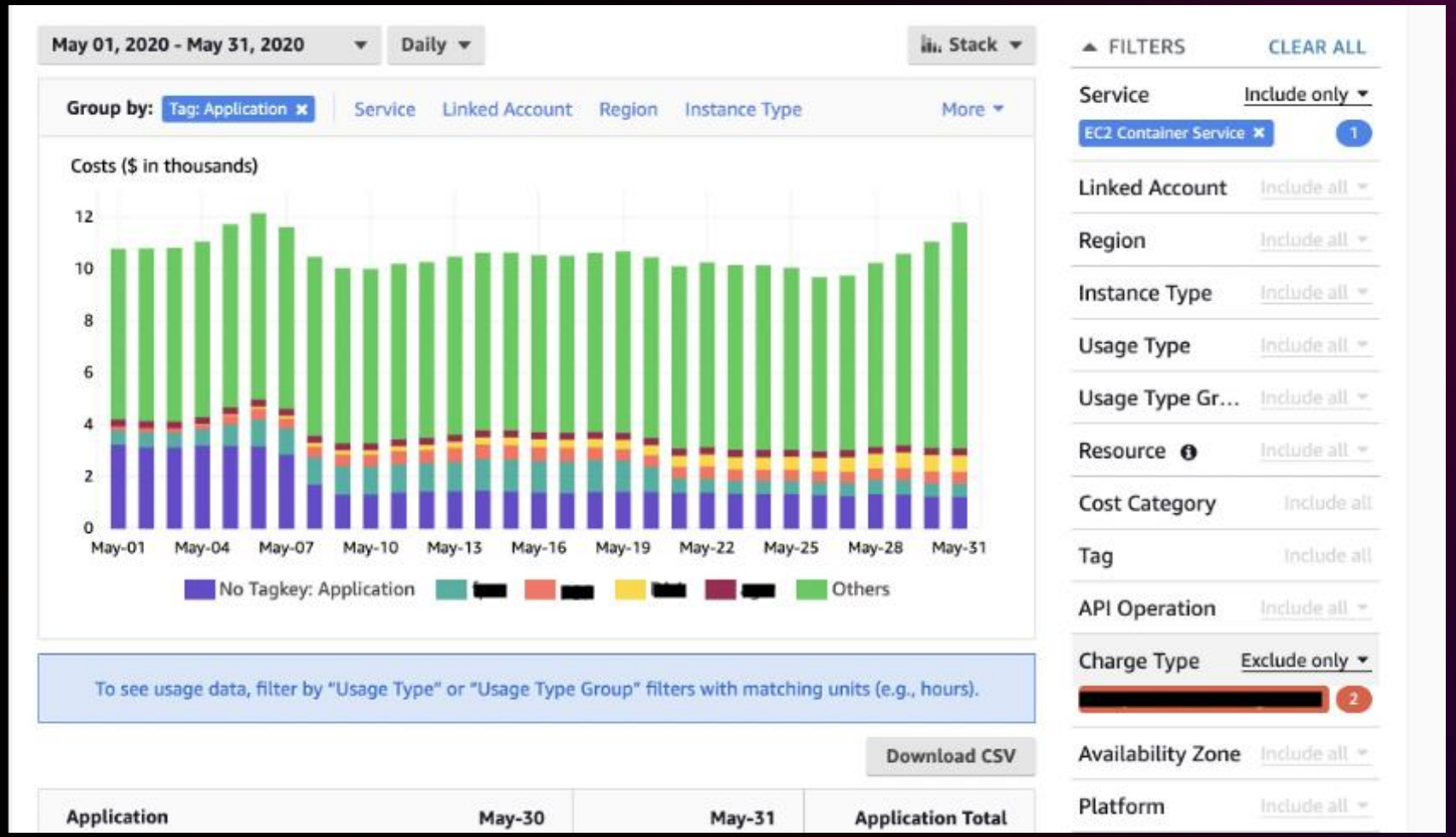
Amazon ECS + Fargate provides task-level cost visibility



Purchase Option	May-01	May-02	May-03	May-04	May-05	May-06
Total cost ($)	-6,533.43	59.59	0.15	4,210.43	5,355.27	6,043.57
Total usage (Hrs)	211,849.20	212,144.20	212,700.29	217,259.04	229,950.10	238,035.02
On Demand ($)	-6,478.31	59.51		4,209.85	5,353.94	6,042.37
On Demand (Hrs)	126,941.81	1,276.83		101,809.00	128,808.88	145,368.66
Spot ($)	-55.13	0.09	0.15	0.58	1.34	1.20
Spot (Hrs)	10.87	7.08	11.86	46.72	107.07	95.84
Savings Plans ($)	0.00	0.00	0.00	0.00	0.00	0.00

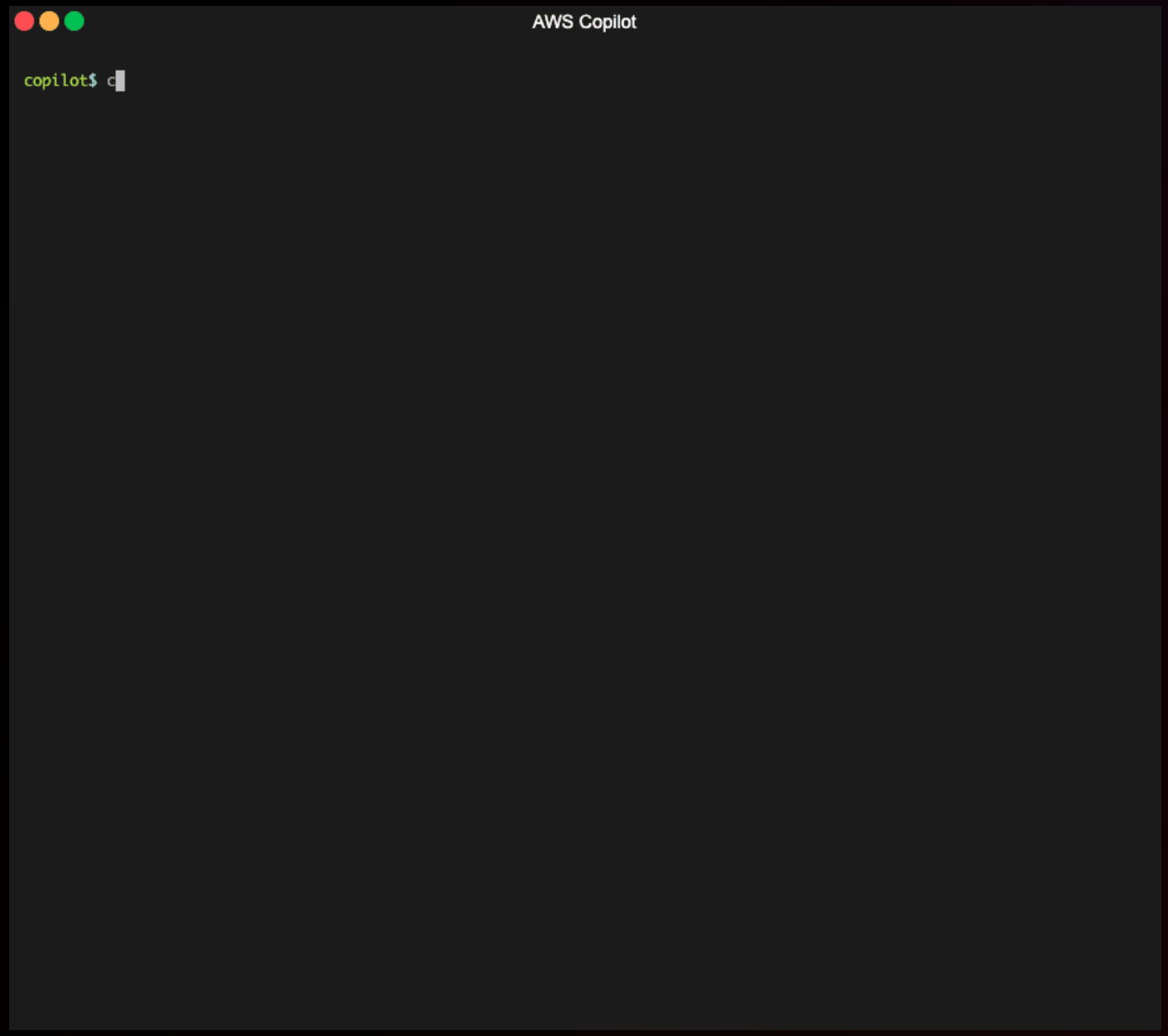
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

View cost per application with tags



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

Well architected from day-1 with AWS Copilot



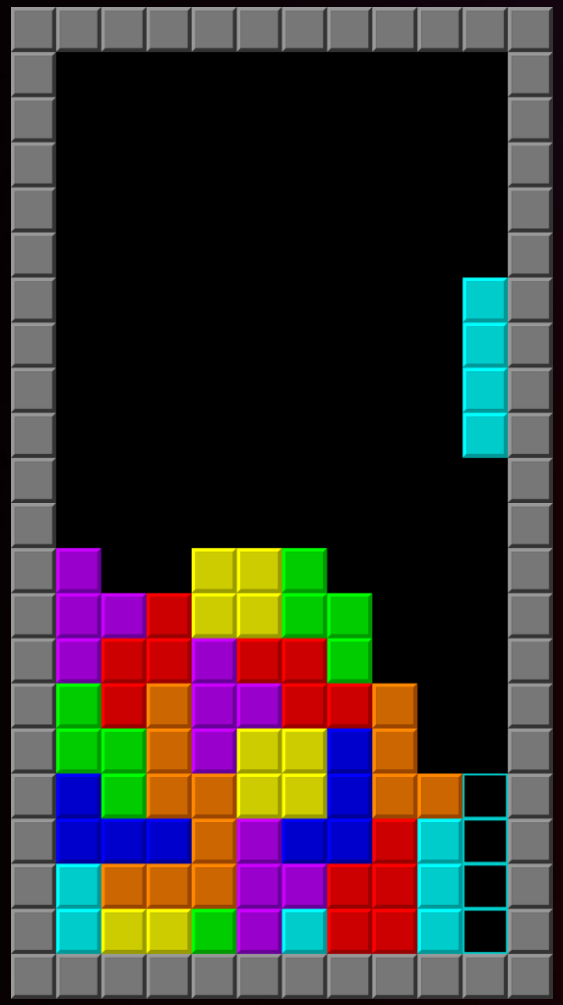
Well-architected infra setup

Continuous deployment pipelines

Effective troubleshooting and operations

Optimize more

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



The best way to optimize costs is to improve efficiency

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

Amazon ECS terminology



An Amazon ECS cluster is a logical namespace – nothing more!



Tasks are the basic unit

Similar to Kubernetes pods



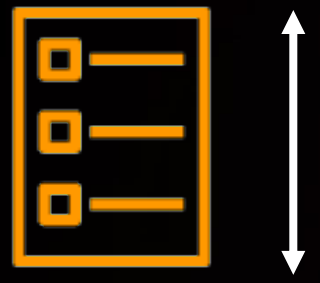
Service can have many tasks

Auto-replaces failed tasks

aws

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

3 key dimensions to improve utilization



Size of task

Vertical Scale



Number of tasks in service

Horizontal Scale

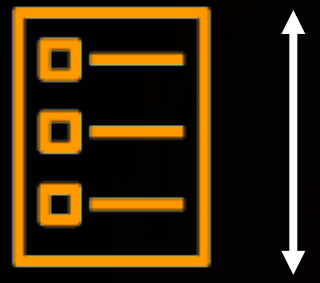


Compute capacity in cluster

Fleet Utilization

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

Let’s start at the task level



Size of task

Vertical Scale



Number of tasks in service

Horizontal Scale

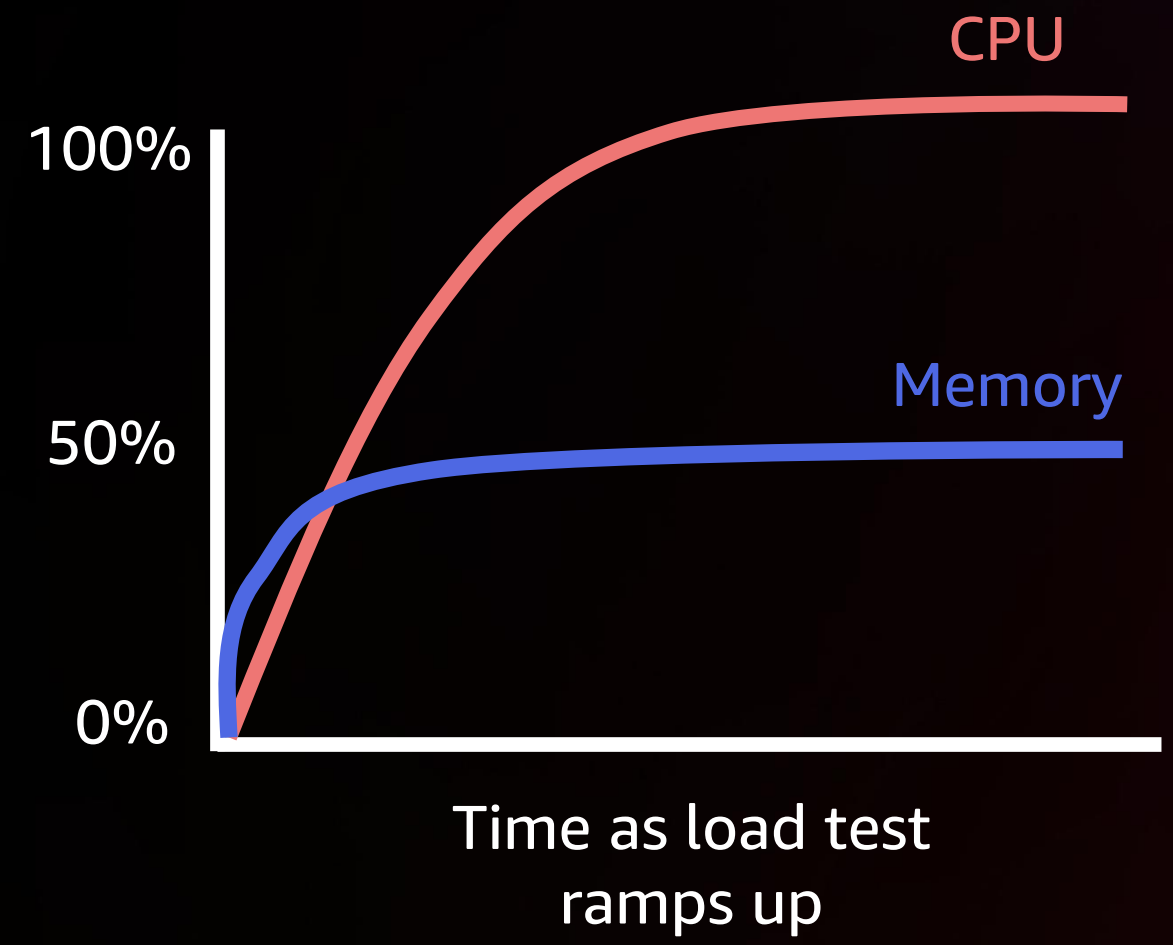


Compute capacity in cluster

Fleet Utilization

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

An example scenario



Workload is primarily CPU bound, as it runs out of CPU first

Consider adding more CPU, but perhaps reducing the amount of reserved memory

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

Right-sizing tasks

Sizing is important. Use historical metrics to identify the resources needed per-workload

Measure and adjust. Sizing on day 1 may not be exactly what an application needs – adjust over time to find the best fit for performance and price

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

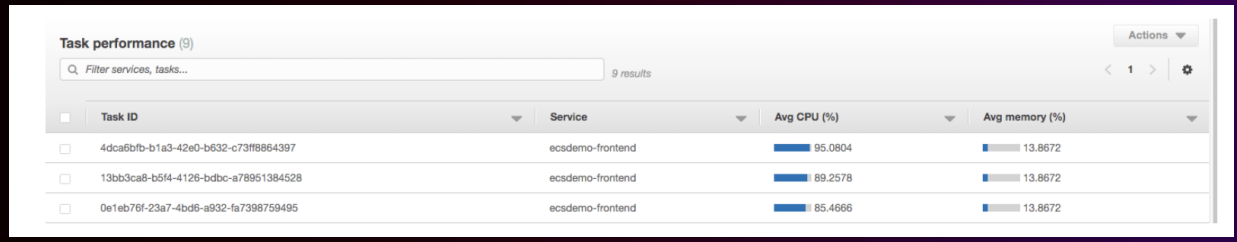
Monitor task utilization with Amazon CloudWatch Container Insights

Amazon ECS integrates with CloudWatch to provide task-level utilization metrics



CloudWatch

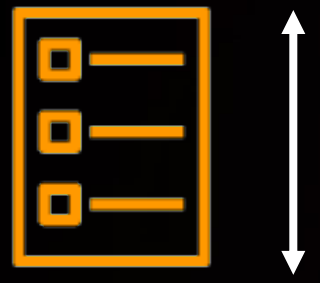
Monitor CPU/memory utilization to identify opportunities to reduce size



Don’t just look at the **average** – peak utilization might be different

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

Now, let’s talk about optimizing tasks in your service based on traffic



Size of task

Vertical Scale



Number of tasks in service

Horizontal Scale

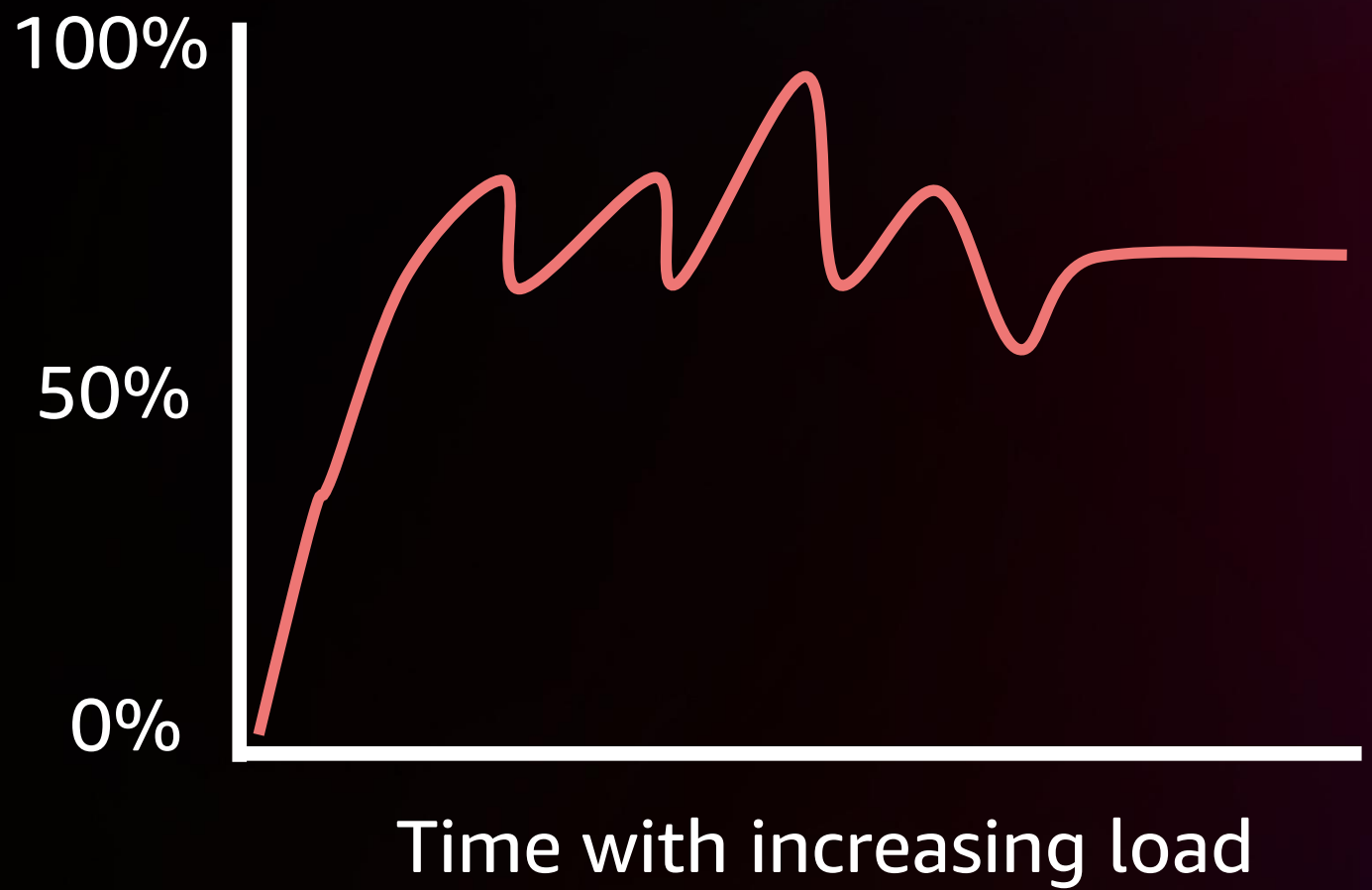


Compute capacity in cluster

Fleet Utilization

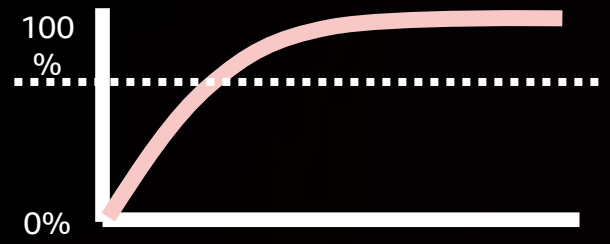
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Traffic served by your application can vary a lot



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

Amazon ECS integrates with AWS Application Auto Scaling to auto scale your service horizontally



Aggregate metrics from the service

Application Auto Scaling

Auto scaling policy is triggered

Amazon ECS

Amazon ECS launches more tasks or stops tasks according to the number of tasks that the policy wants

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

You can auto scale your Amazon ECS service using a variety of metrics

Configure scaling policies for your ECS service, using the right metric for your application



Default Amazon ECS metrics

CPU, memory utilization



Predefined metrics

For example, ALB average latency, Amazon SQS queue depth



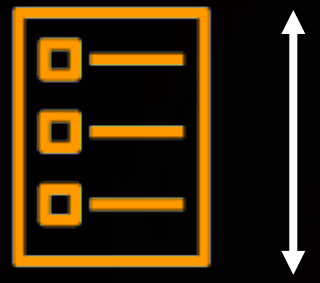
Custom metrics

For example, Amazon SQS backlog/task

aws

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

Finally, let’s talk about utilization at the cluster level



Size of task

Vertical scale



Number of tasks in service

Horizontal scale



Compute capacity in cluster

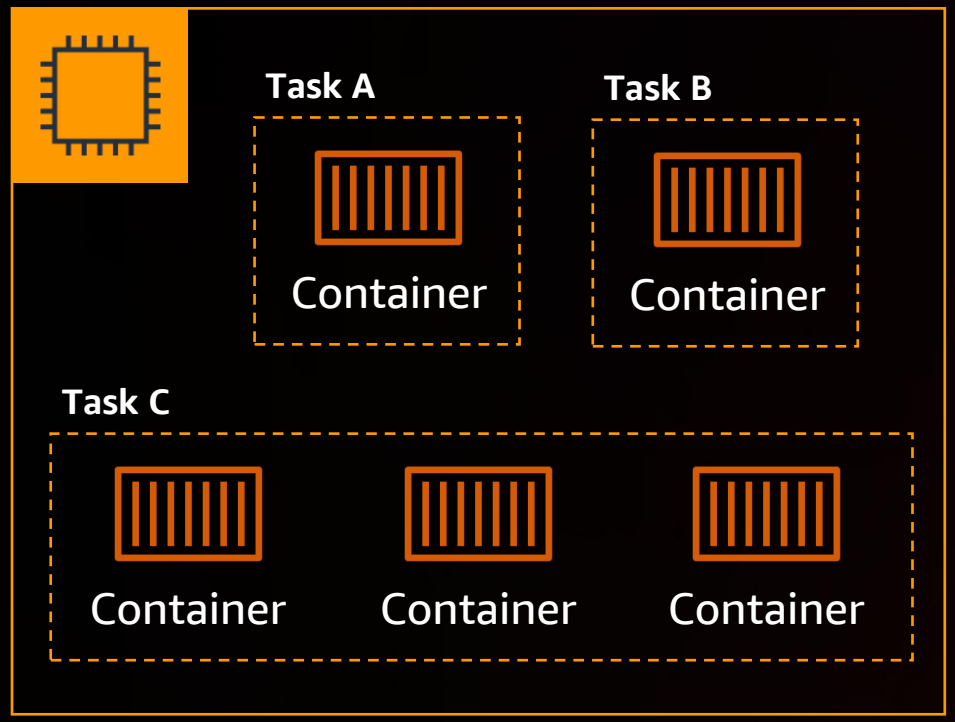
Fleet utilization

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

Efficiently setting the size of a cluster with many tasks on many instances is hard

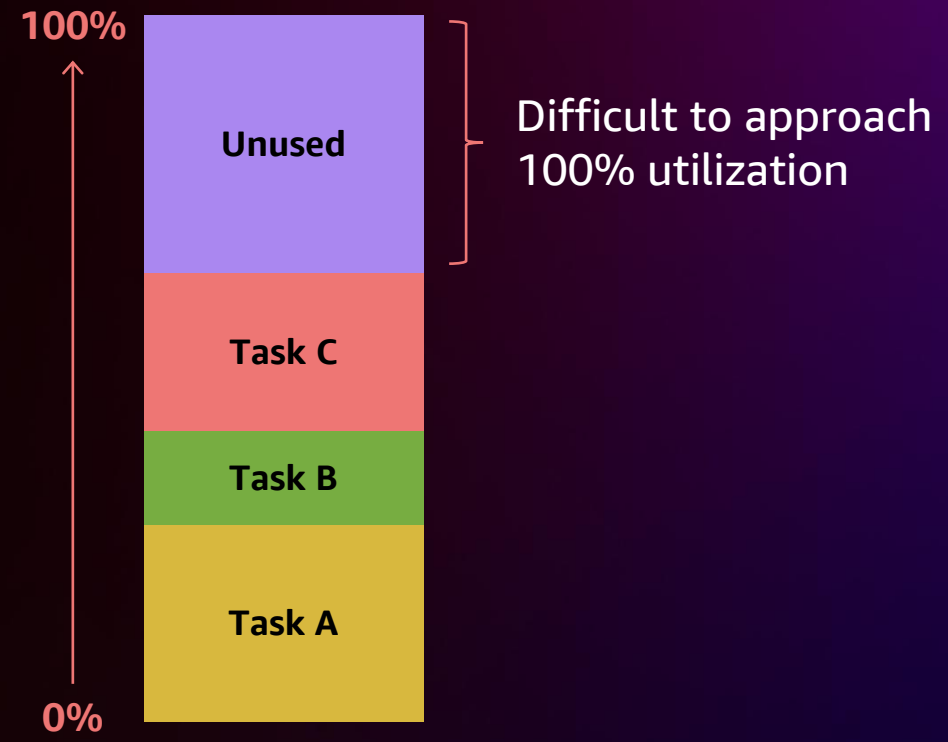
EC2 instance

Difficult to achieve balanced number of containers and resource capacity



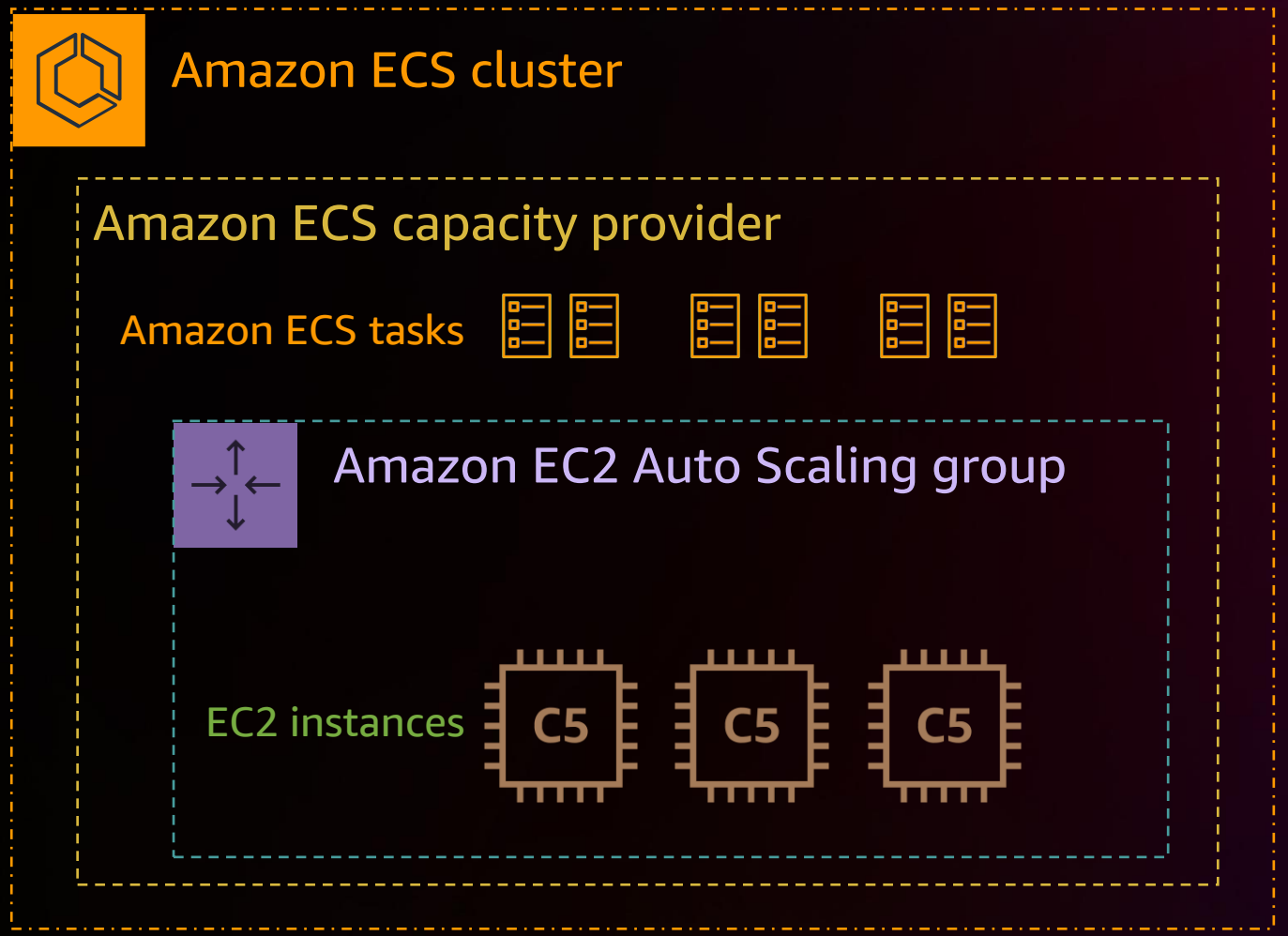
Resource utilization

Often paying for large amounts of unused EC2 resources across multiple cluster nodes



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

Amazon ECS capacity providers helps infrastructure respond to application

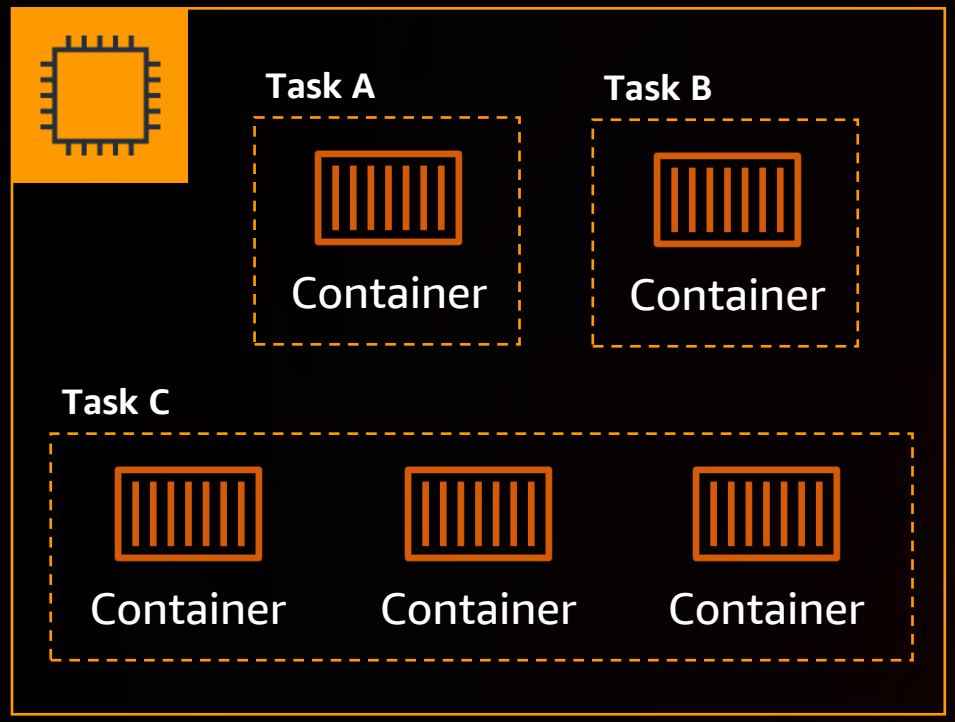


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

But what if you didn't need to think about Amazon EC2 utilization at all?

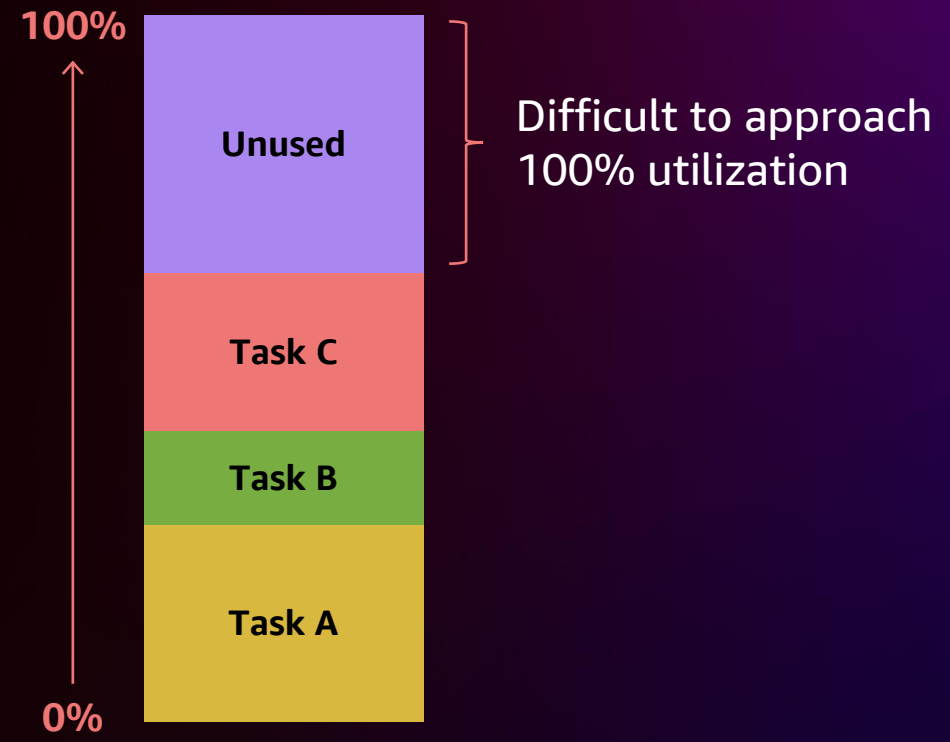
EC2 node

Difficult to achieve balanced number of containers and resource capacity



Resource utilization

Often paying for large amounts of unused EC2 resources across multiple cluster nodes

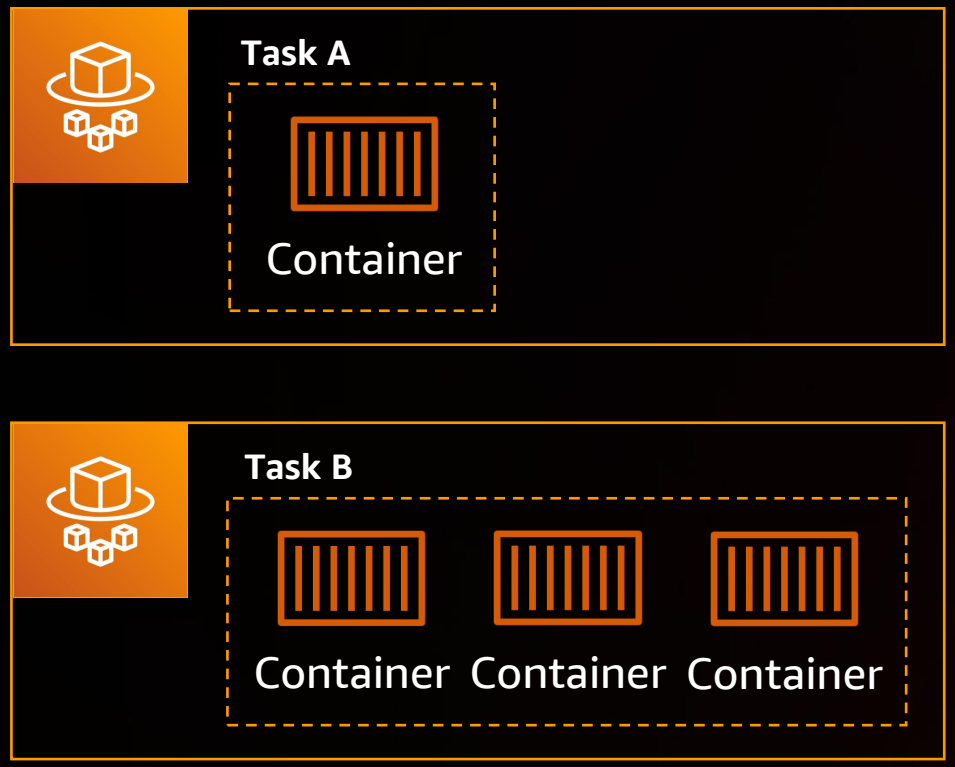


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Pay only for compute you use with Fargate

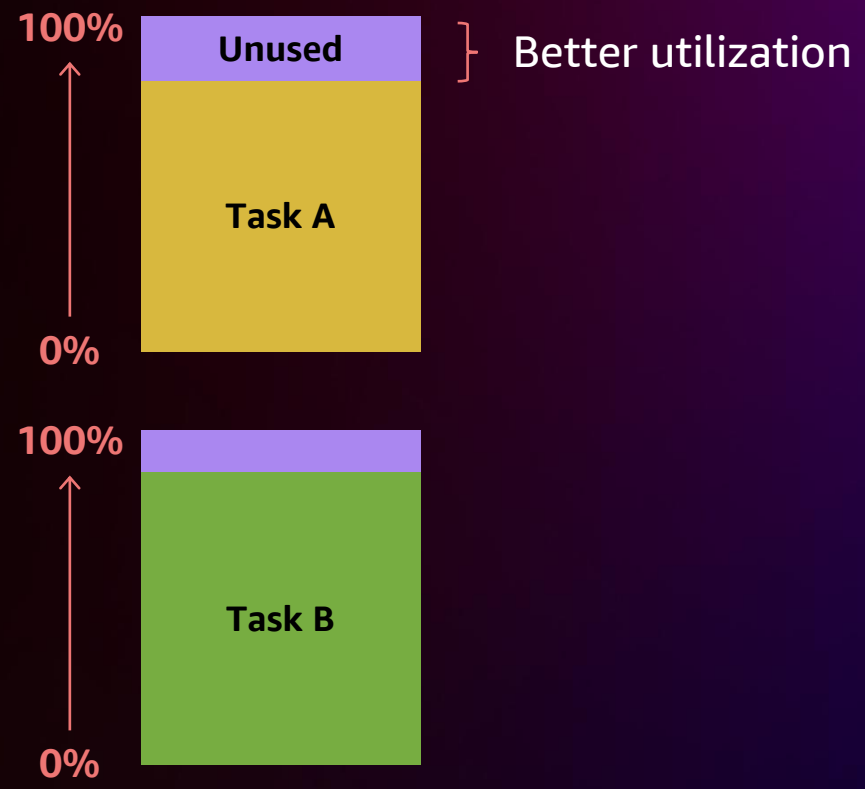
Fargate task

Size each task based on required memory and vCPU resources



Resource utilization

Set more granular control over resources with Fargate, minimizing wasted capacity



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

Optimizing costs with Fargate



- Challenge and solution Edmunds.com increased availability and reliability for its end users by providing fault-tolerant applications and cut costs on compute power, all while saving its team time by not having to manage Amazon EC2 instances

- Reduced costs 25–30% compared to Amazon EC2 On-Demand costs
- Cut billing by $100,000 from October 2019 to August 2020
- Runs 80% of customer-facing applications on AWS Fargate Spot
- Increased availability to 99.999%
- Releases 20–30 daily services compared to 10–20
- Auto scales in seconds rather than minutes

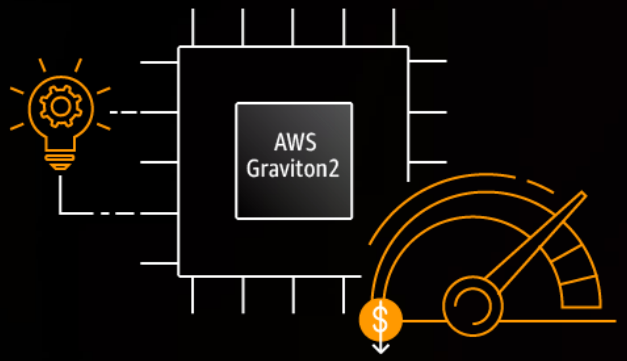
aws

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

Choose wisely

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

3 key infrastructure choices to reduce costs







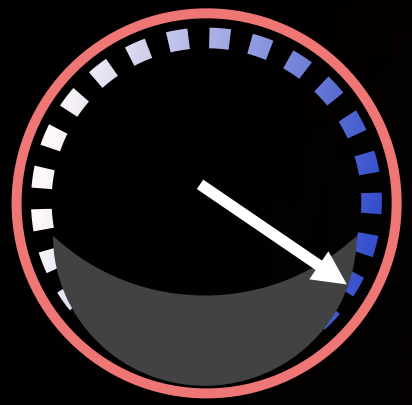
Hardware Type

Capacity Type

Commitments & Reservations

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

Reduce compute cost and improve performance with AWS Graviton



Highest performance
in their instance families



20% lower cost
versus same-sized comparable fifth-generation instances



Up to 40% better price-performance
versus comparable x86-based instances

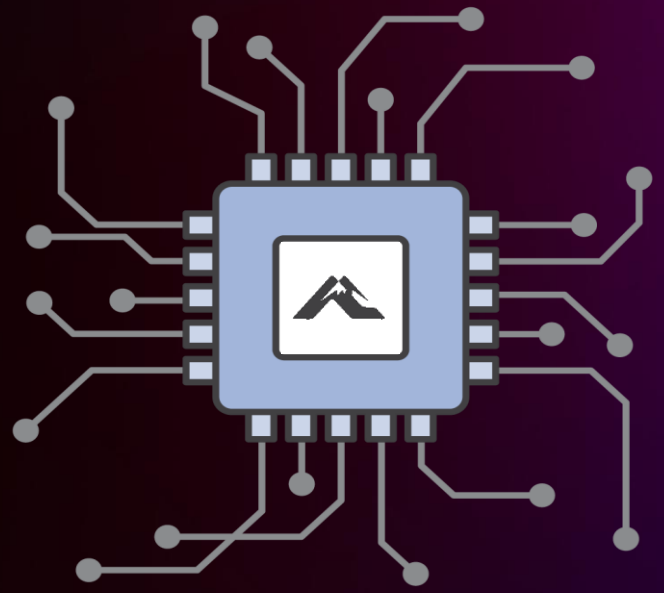
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon ECS provides full support for AWS Graviton

Fargate supports AWS Graviton2

ECS-optimized Amazon Linux 2 (arm64) AMI for EC2

Full support for multi-arch images



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

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

Use AWS Graviton with Amazon ECS for a variety of workloads

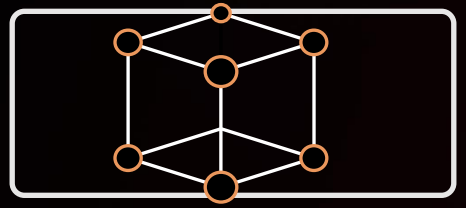
Web and gaming servers

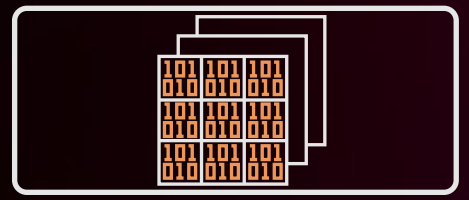




Open-source databases

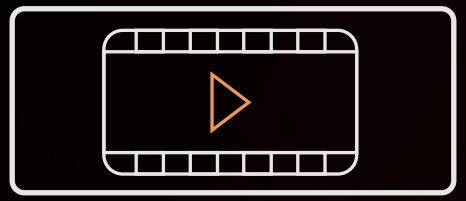
High performance computing

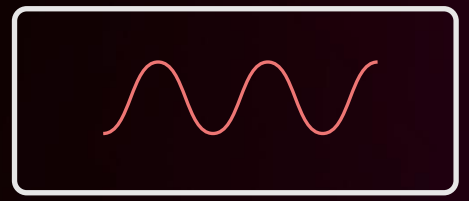




In-memory caches

Media encoding





Electronic design automation

Analytics





Microservices

aws

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

Amazon EC2 and Fargate Spot are how you can do more for less

Up to 90% off On-Demand



Same infrastructure as On-Demand

89+ security standards

300+ instance options

24+ Regions

Spot is more than cost savings

More compute + Faster results = Accelerated innovation







aws

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

Some nuances of EC2 Spot and Fargate Spot



EC2 Spot instance

Price up to 90% less than On-Demand

Can be **reclaimed** (with two-minute warning)

You choose instance pools



Fargate Spot

Price up to 70% less than standard Fargate

Can be **reclaimed** (with two-minute warning)

Automatic diversification

aws

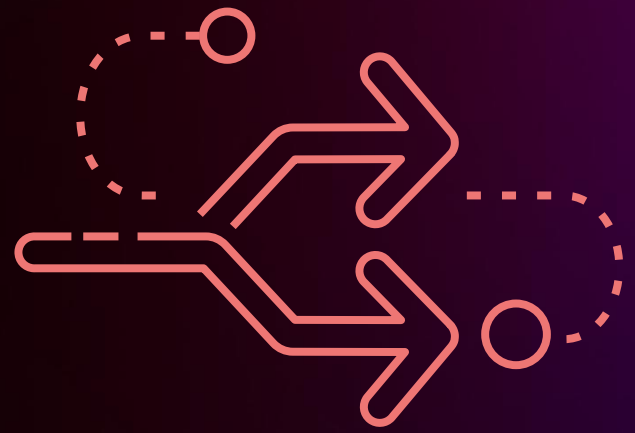
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon ECS makes it easier to use Spot by handling interruptions

Amazon ECS handles Spot interruptions automatically

Amazon ECS takes over the coordination of termination of tasks

Automatic graceful termination of Load Balancer connections



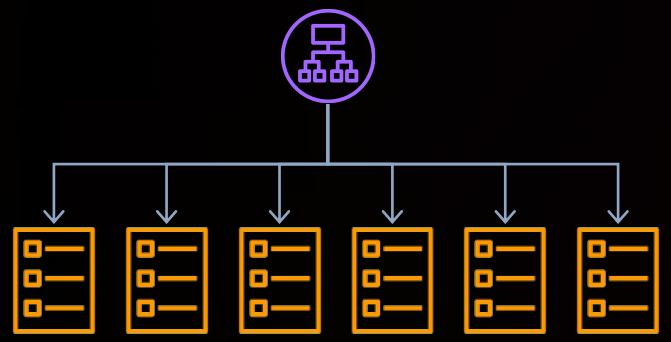
aws

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

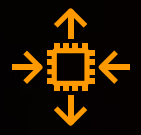
Mix On-Demand and Spot with capacity providers

Overprovision by 50%:
Reduce metric target value by 1/3

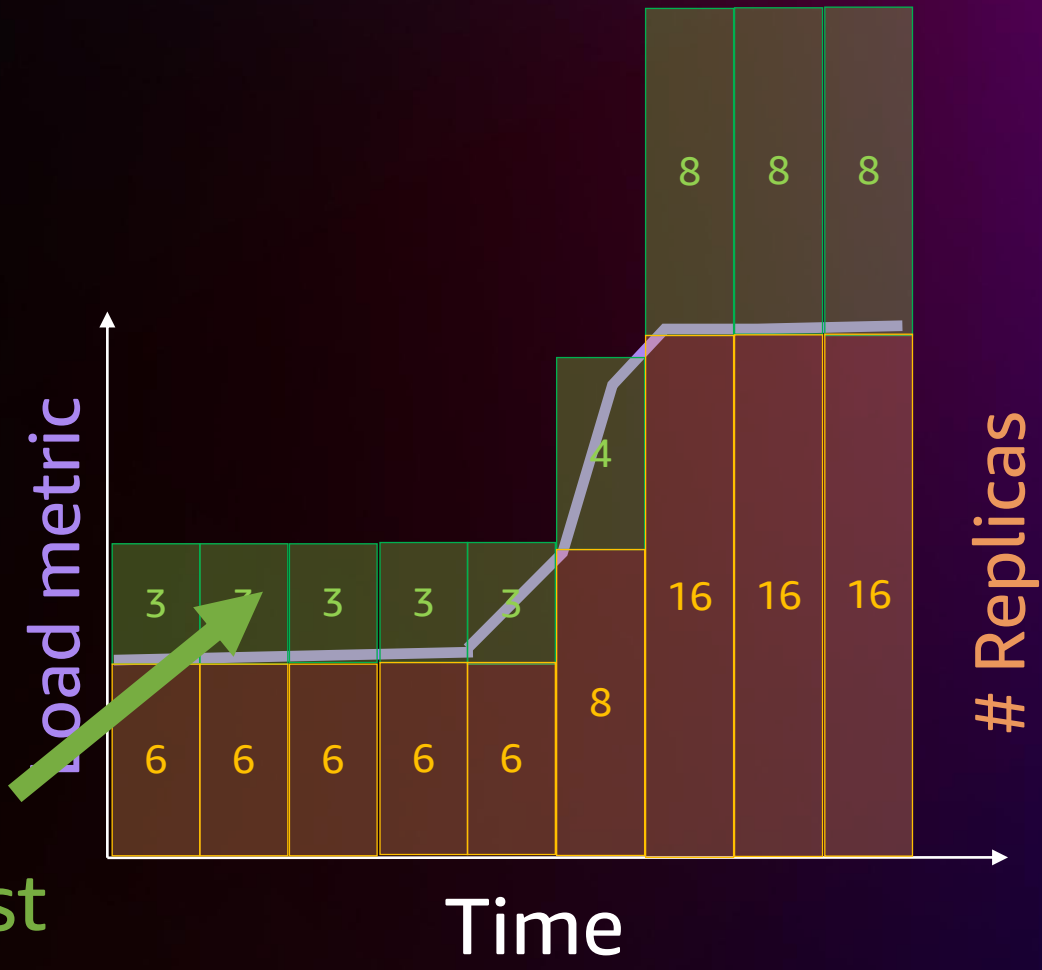
Run 2/3 On-Demand, 1/3 on Spot







+50% capacity
for +5–10% cost



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

Commitment discounts with Savings Plans



What is it?

A **flexible** pricing model that helps you save **up to 72%** for committed compute usage



How does it work?

Simply commit to an hourly rate and the savings will **automatically** be applied to your usage

aws

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

Savings Plans work automatically with Amazon ECS

Instance Saving Plans



Instance Savings Plans apply savings to selected instance usage across any tenancy or OS

Compute Savings Plans



- **Compute Savings Plans** discounts apply to **any** compute usage automatically
- **Compute Savings Plans** apply to **Fargate, AWS Lambda,** and **Amazon EC2**

Savings Plans offer the **same great savings** as Reserved Instances but offer **wider coverage** and dramatically **more flexibility,** and **minimize overhead** management

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

Fargate Savings Plan overview

Term length	Payment option	CPU discount	Memory discount
1 year	No upfront	15-20%	15-20%
1 year	Partial upfront	20-25%	20-25%
1 year	All upfront	22-27%	22-27%
3 years	No upfront	40-45%	40-45%
3 years	Partial upfront	45-50%	45-50%
3 years	All upfront	47-52%	47-52%

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

https://aws.amazon.com/savingsplans/pricing/

Optimizing throughput and costs with Amazon ECS and EC2 Spot

Challenge

athenahealth is a leading provider of network-enabled software and services for medical groups and health systems nationwide. **athenahealth's QMaaS** (Quality Management as a Service) application analyzes clients' clinical data to measure effectiveness of care and performances of healthcare providers.

QMaaS scores hundreds of millions of quality events each month, with peaks of nearly 6 billion events. QMaaS was incurring an average monthly compute (including Lambda and EC2) spend of close to $170,000 per month since launch in 2019, with a peak monthly spend of $440,000.

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

athenahealth

Solution and results

Deploying only one of the key Lambda functions into containers orchestrated by Amazon ECS with EC2 Spot capacity providers improved throughput and scalability, and simultaneously yielded cost reduction of 73% per month and overall savings of $1.3 million per year

> "Migrating the performance metrics computation from Lambda to a cluster of Docker containers, orchestrated by Amazon ECS and backed by Amazon EC2 Spot Instances improved throughput and scalability, and simultaneously yielded cost reduction of 73% per month and overall savings of $1.3 million per year."

Karthik Kalkur, Senior Architect at athenahealth

CleverTap’s journey to cost optimization

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

CleverTap

Making Mobile Retention a Piece of



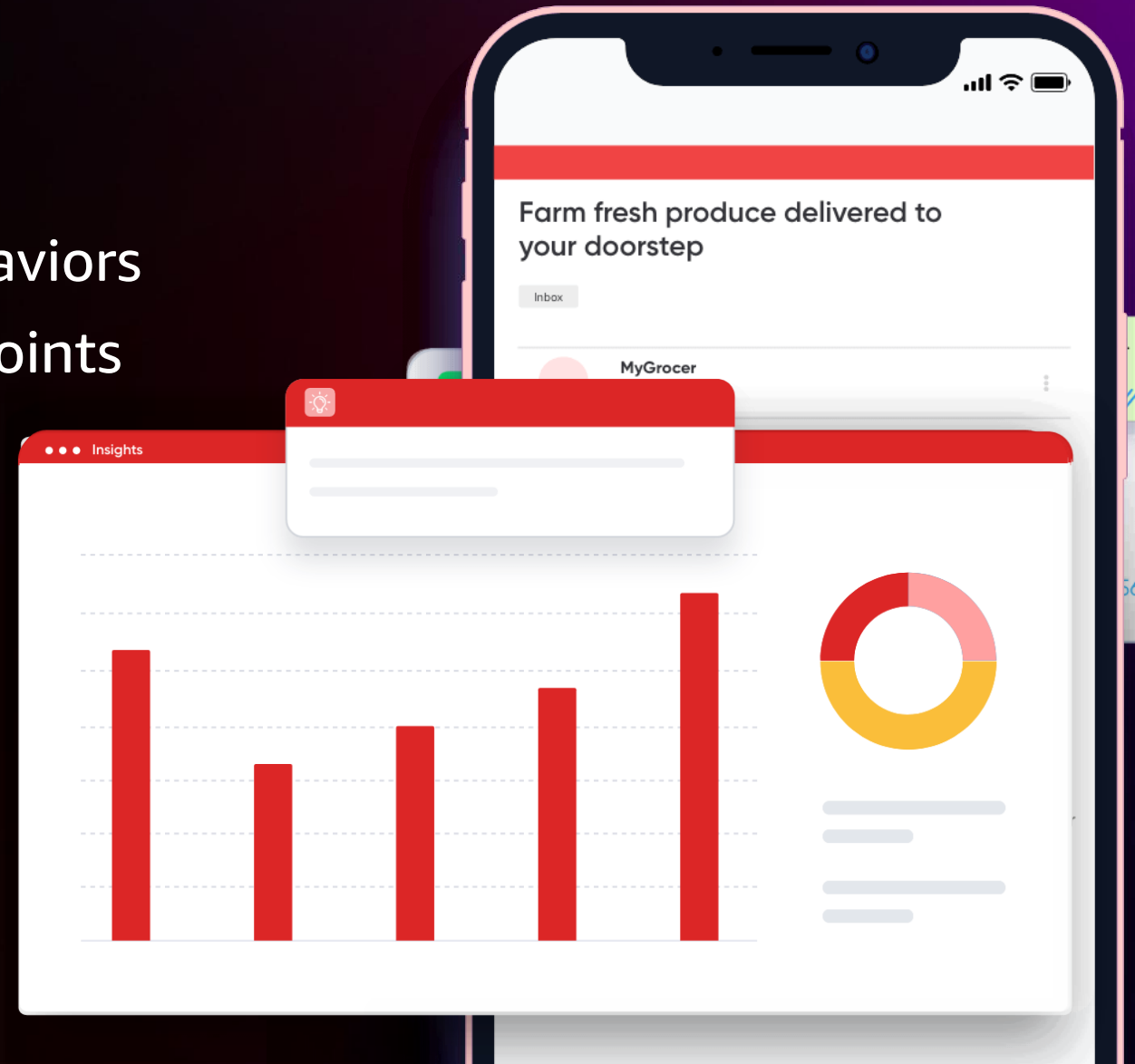
aws

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

CleverTap

OMNICHANNEL CUSTOMER ENGAGEMENT & USER RETENTION PLATFORM

1. Powered by TesseractDB™
2. Deep insights into user trends and behaviors
3. Engage users across messaging touchpoints



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

Hello.

- Francis Pereira, VP Infrastructure and Security Engineering
- Helping keep the lights on for the 16 years. Last 8 years @ CleverTap

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

A brief history

1. 2012 - 2016: AMI with env bootstrapped by Puppet
2. 2017: Container v1. Applications packaged in containers, Docker compose and code deploy for orchestration
3. 2018 – 2021: Applications packaged in containers, running on ECS/EC2
4. 2022: Stateless applications packaged in containers, running on ECS/Fargate



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

ECS/EC2 => ECS/Fargate

1. Focus time and energy on delivering business value
    1. Instead of Operating System vulnerability assessment
    2. Instead of Operating System hardening
2. Simplifies cluster resource management
3. All our stateless containers run on ECS/Fargate

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

Lessons learned operating ECS (EC2 and Fargate)

1. Cluster events are not available in CloudTrail
2. Tasks can’t be protected from health check failures
3. Limited KernelCapabilities on Fargate
4. Fargate requires using task IAM roles
5. Stateful workloads with EBS support can be orchestrated using container pining on ECS/EC2 but not on Fargate
6. EC2 cluster resources are decoupled from ECS service unless you use capacity providers. Choose Fargate if possible
7. ECS exec for dropping into the running container
8. Instance type updates are not supported

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

Running truly containerized

1. Resist the urge to rebuild upstream containers
2. ECS exec for break glass events
3. Sidecar deployment model

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

Scaling with ECS

1. Ship standard out to a remote logging service
2. Ship ECS agent logs and cluster events to a remote logging service
3. One cluster per microservice - clusters are free

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

A few tactical callouts

VPC endpoints with Amazon ECR for image pulls

[New] Amazon ECS Service Connect for service to service communication inside VPC

A single ALB can be shared across multiple Amazon ECS services

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

Key takeaways

Amazon ECS = Serverless orchestration + Built-in capabilities + AWS integrated

Fargate = No instance and capacity management

Use service auto-scaling and right-size tasks to improve efficiency

Amazon ECS + AWS Graviton for 40% better price performance

Mix on-demand + Spot with capacity providers

Savings plan for usage commitment discounts

… Some of this can depend

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

Resources

Amazon ECS + Fargate cost optimization checklist: https://go.aws/3ENR8XM

Amazon ECS scaling best practices: https://bit.ly/3FeqloJ

Using AWS Graviton to save costs with Amazon ECS: https://go.aws/3ueEL22

Save costs with Amazon ECS capacity providers and Spot: https://go.aws/3ufaq3u

aws

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

Thank you!

Vibhav Agarwal
agvibhav@amazon.com
@VibhavAgarwal22

Francis Pereira
@francispereira



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

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