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

CMP204

Build a cost-, energy-, and resource-efficient compute environment

Steffen Grunwald (he/him)

Principal Solutions Architect
AWS

Adam Boeglin (he/him)

Principal Specialist, Amazon EC2
AWS

Troy Gasaway (he/him)

VP of Infrastructure and Engineering
Arm

aws

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

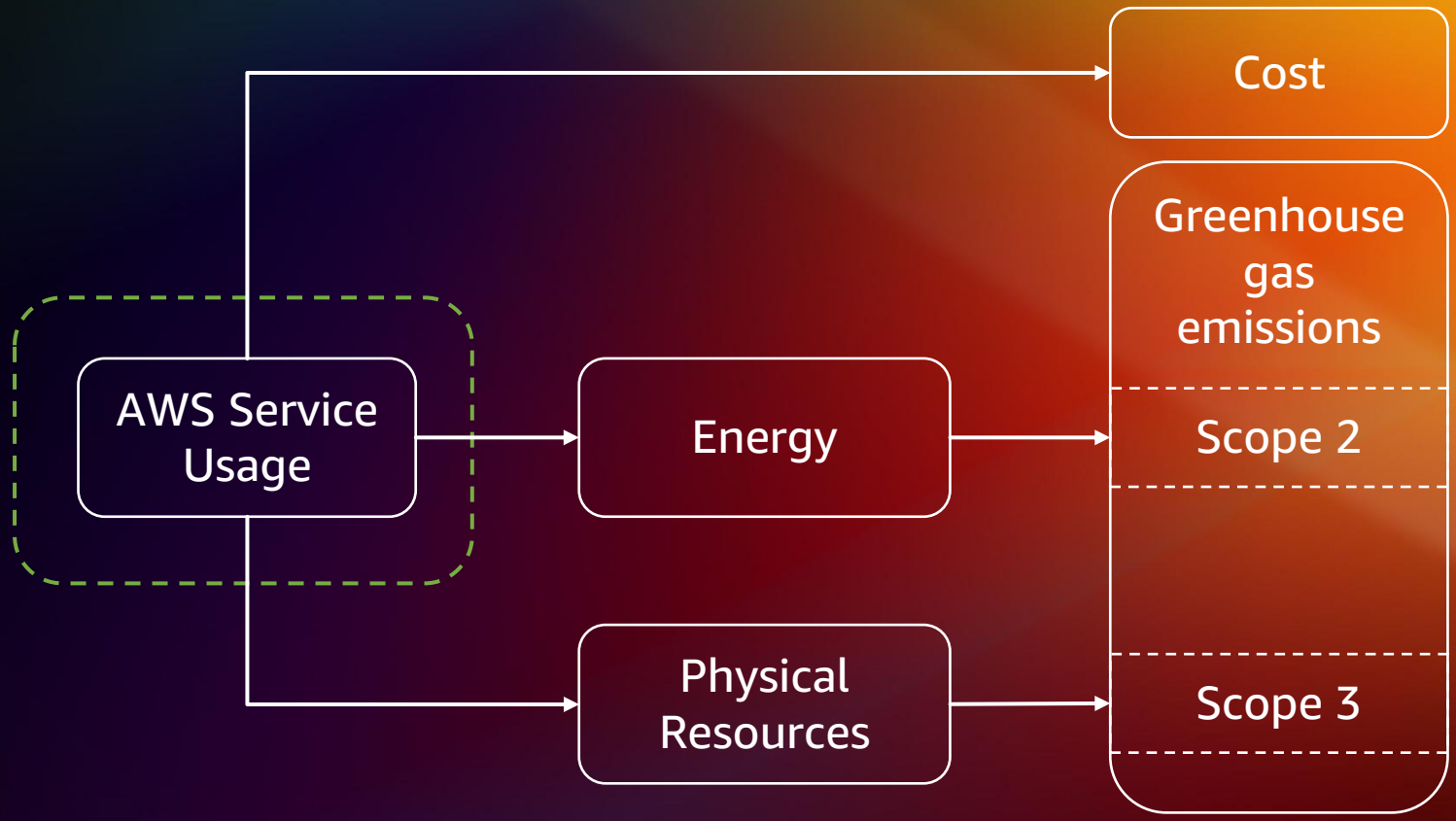
“I need to reduce my…”

Cost

Greenhouse gas emissions

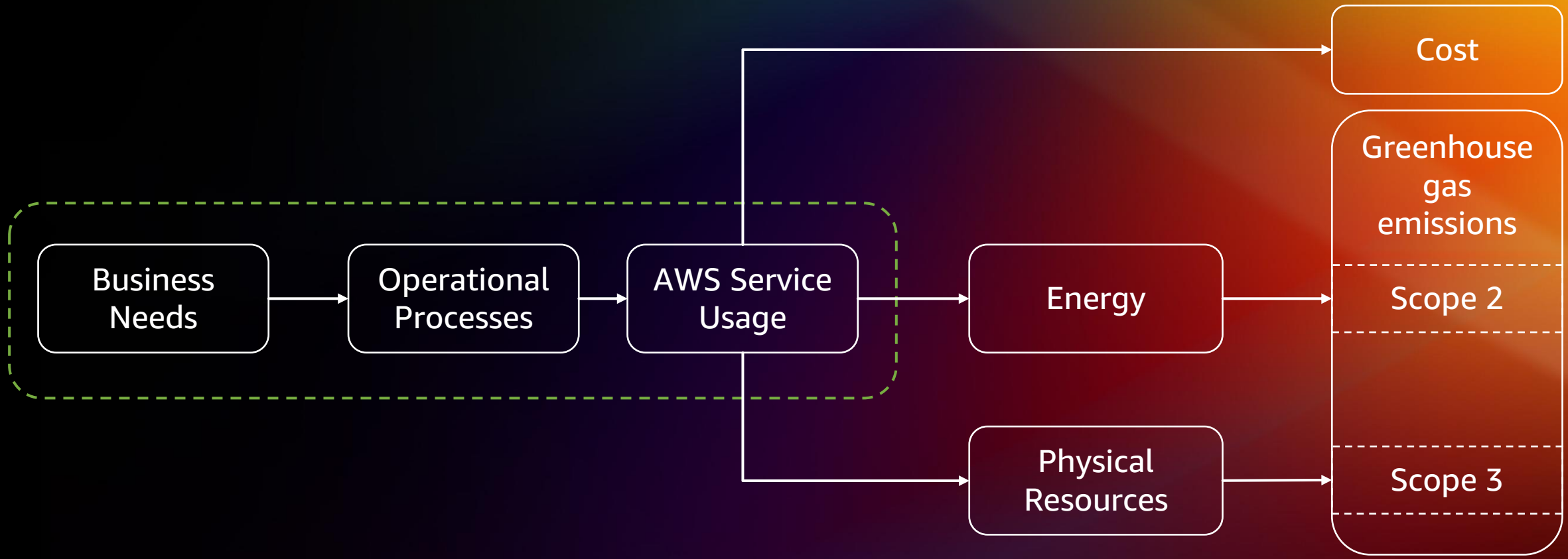
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Focus on your inputs



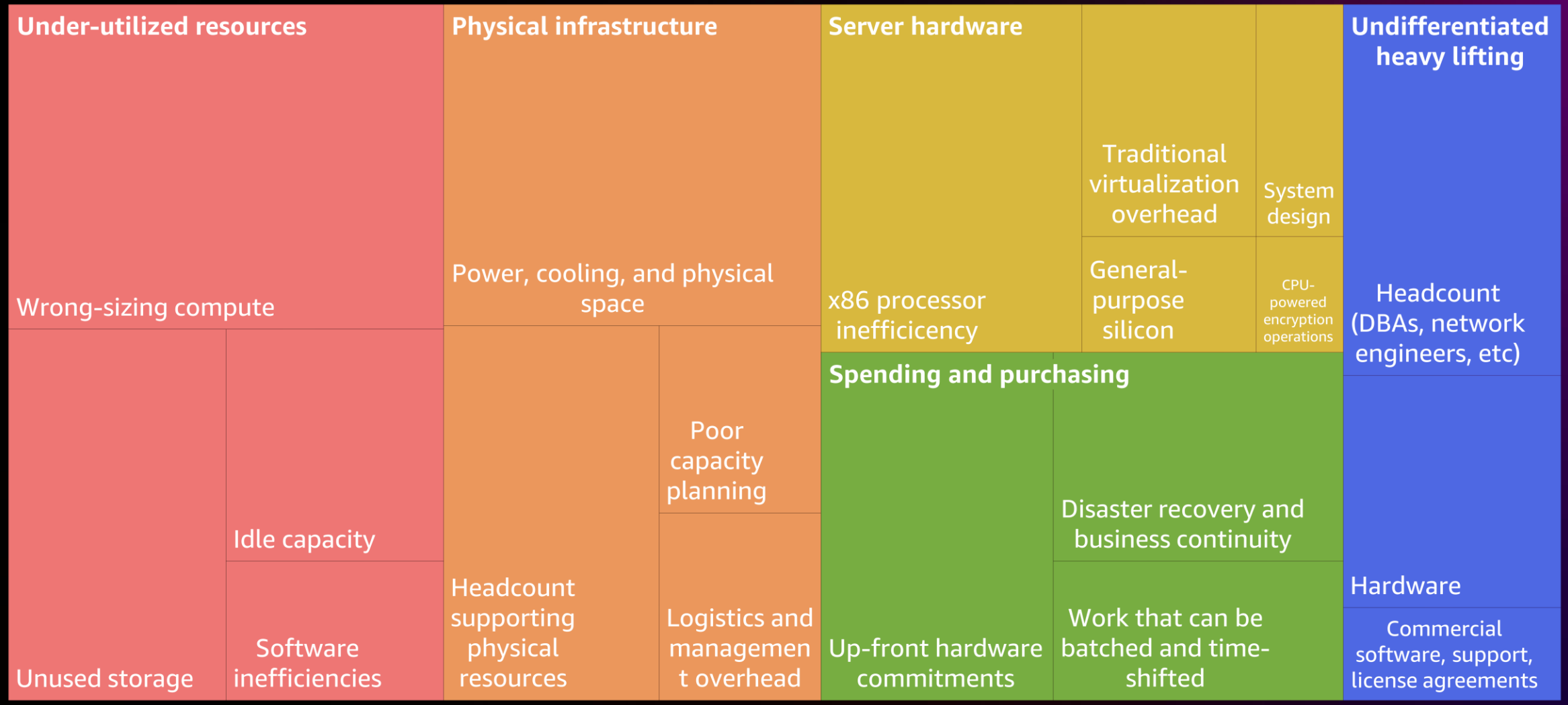
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Focus on your inputs



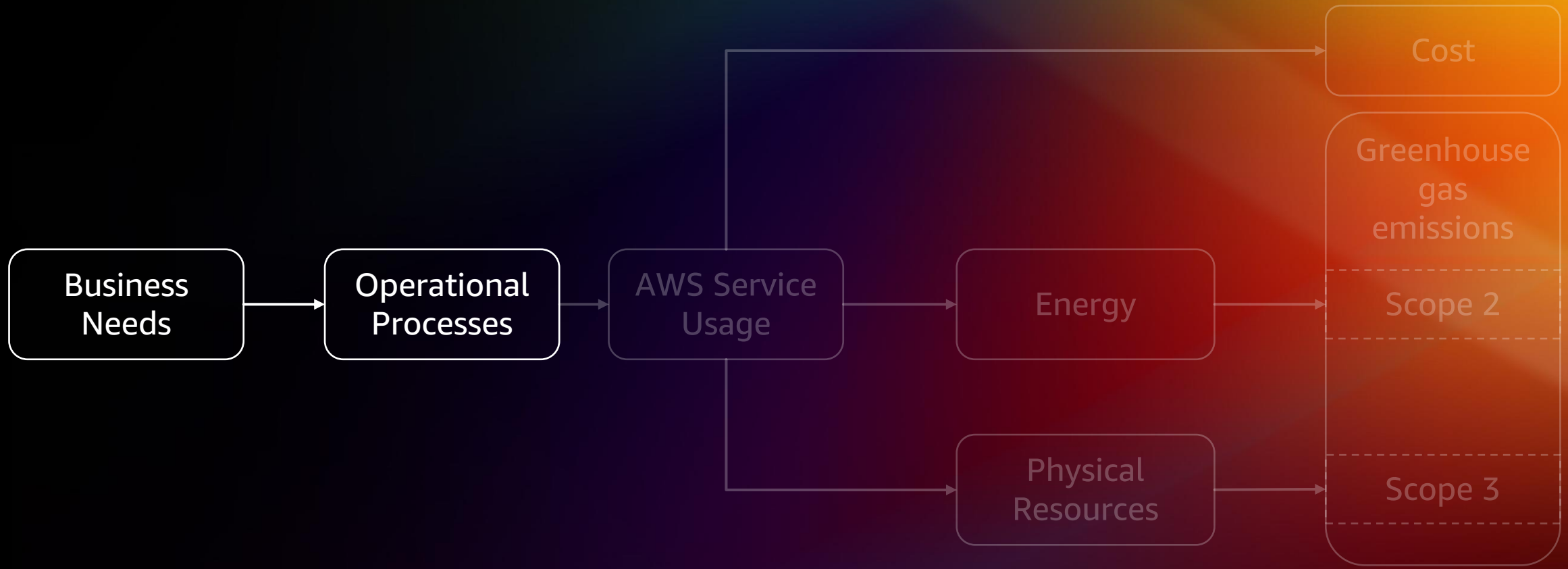
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Map your compute infrastructure inefficiencies

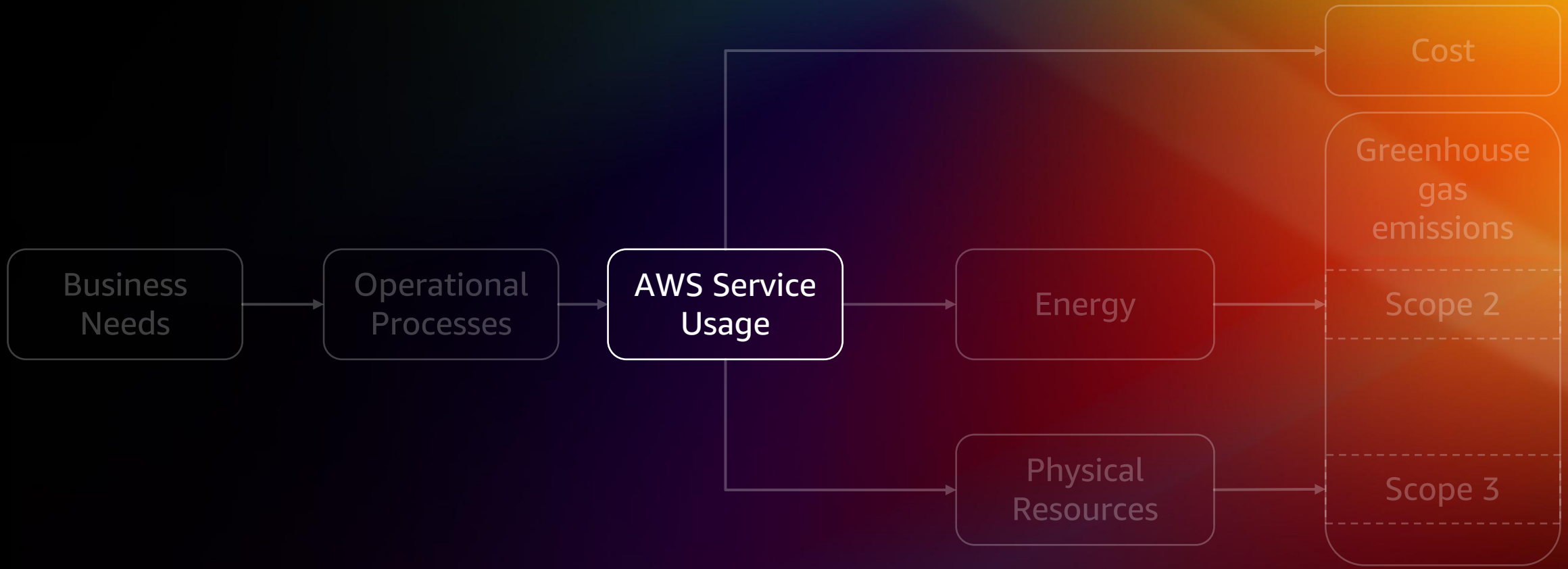


aws

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

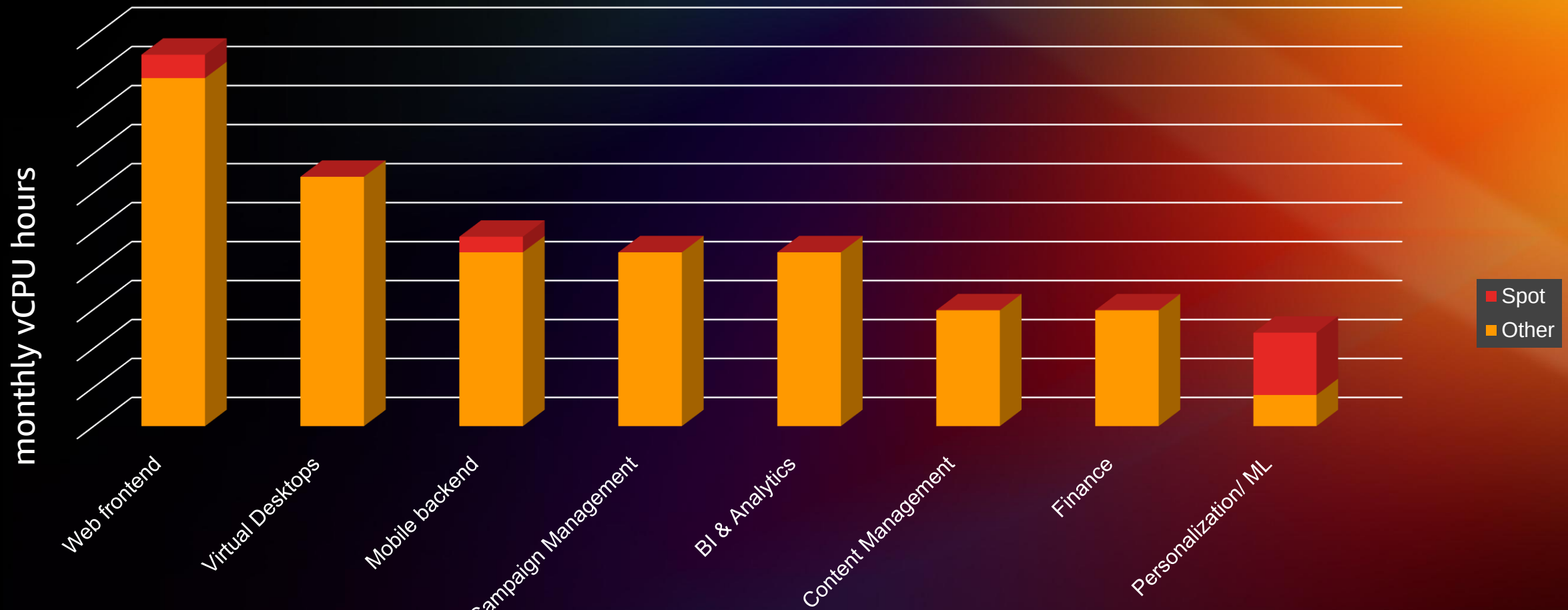


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



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

Example: Understand & track usage for compute



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

Examples for cost & usage as normalized KPIs

ESTABLISH NORMALIZED KEY PERFORMANCE INDICATORS (KPI) TO MAKE THEM COMPARABLE OVER TIME

Always normalize usage:

EC2 instance hours ➔ vCPU hours

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

Examples for cost & usage as normalized KPIs

ESTABLISH NORMALIZED KEY PERFORMANCE INDICATORS (KPI) TO MAKE THEM COMPARABLE OVER TIME

Always normalize usage: EC2 instance hours ➔ vCPU hours

Track desired usage in relation to total usage:

CPU utilization

$$\frac{\text{EC2 Spot vCPU hours}}{\text{all vCPU hours}}$$

$$\frac{\text{AWS Graviton AWS Lambda Functions}}{\text{all Lambda Functions}}$$

aws

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

Examples for cost & usage as normalized KPIs

ESTABLISH NORMALIZED KEY PERFORMANCE INDICATORS (KPI) TO MAKE THEM COMPARABLE OVER TIME

Always normalize usage:

EC2 instance hours ➔ vCPU hours

Track desired usage in relation to total usage:

CPU utilization

$$\frac{\text{EC2 Spot vCPU hours}}{\text{all vCPU hours}}$$

$$\frac{\text{AWS Graviton AWS Lambda Functions}}{\text{all Lambda Functions}}$$

Track resource efficiency in relation to your business:

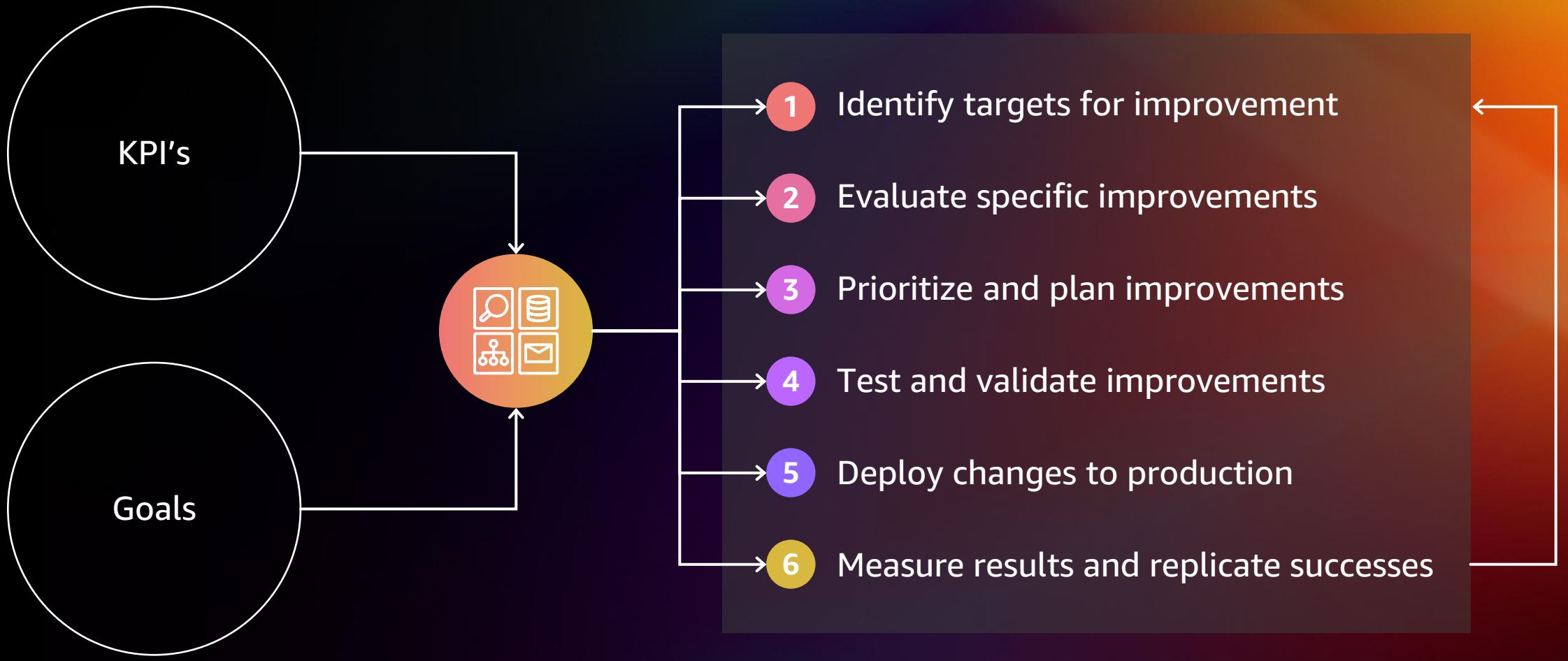
$$\frac{\text{cost}}{\text{business units of work}}$$

$$\frac{\text{vCPU hours}}{\text{business units of work}}$$

aws

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

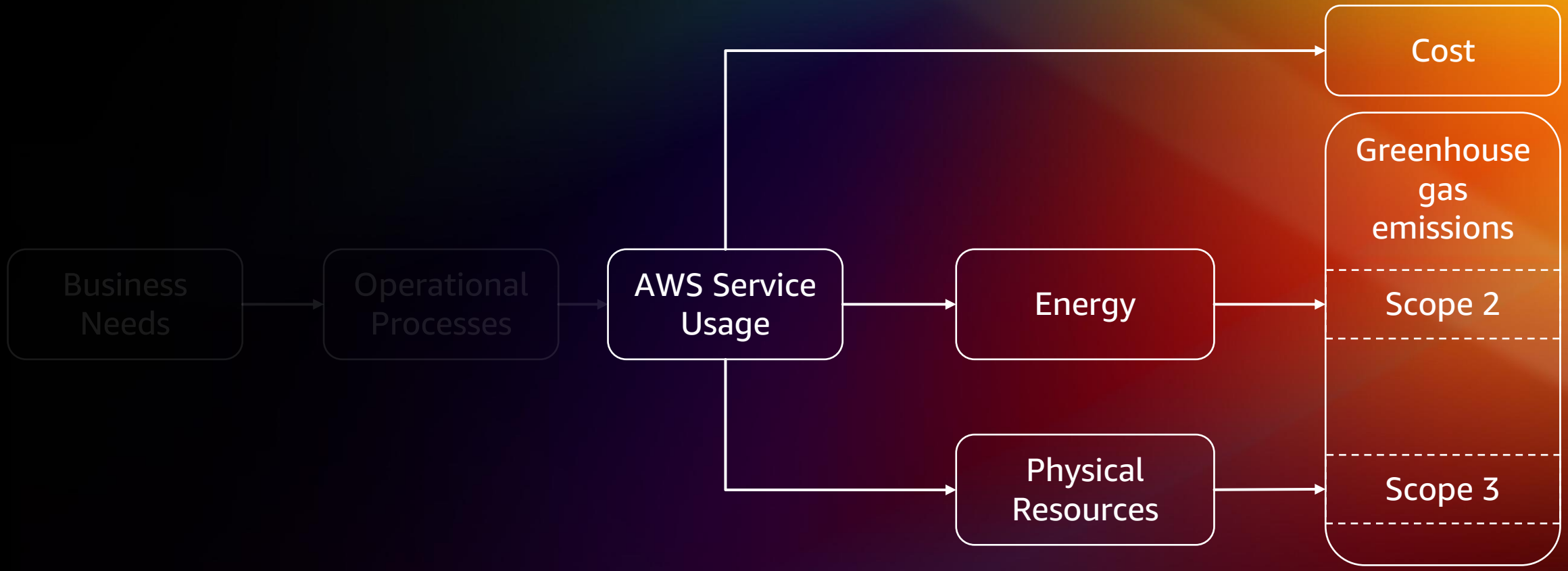
Improvement process



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

You don’t have to do it alone

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



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

Carbon reduction opportunity

AWS can lower the carbon footprint of average on-premises data center workloads by nearly 80% today and up to 96% once AWS is powered with 100% renewable energy





Europe
Up to 5 times more energy efficient

US
3.6 times more energy efficient

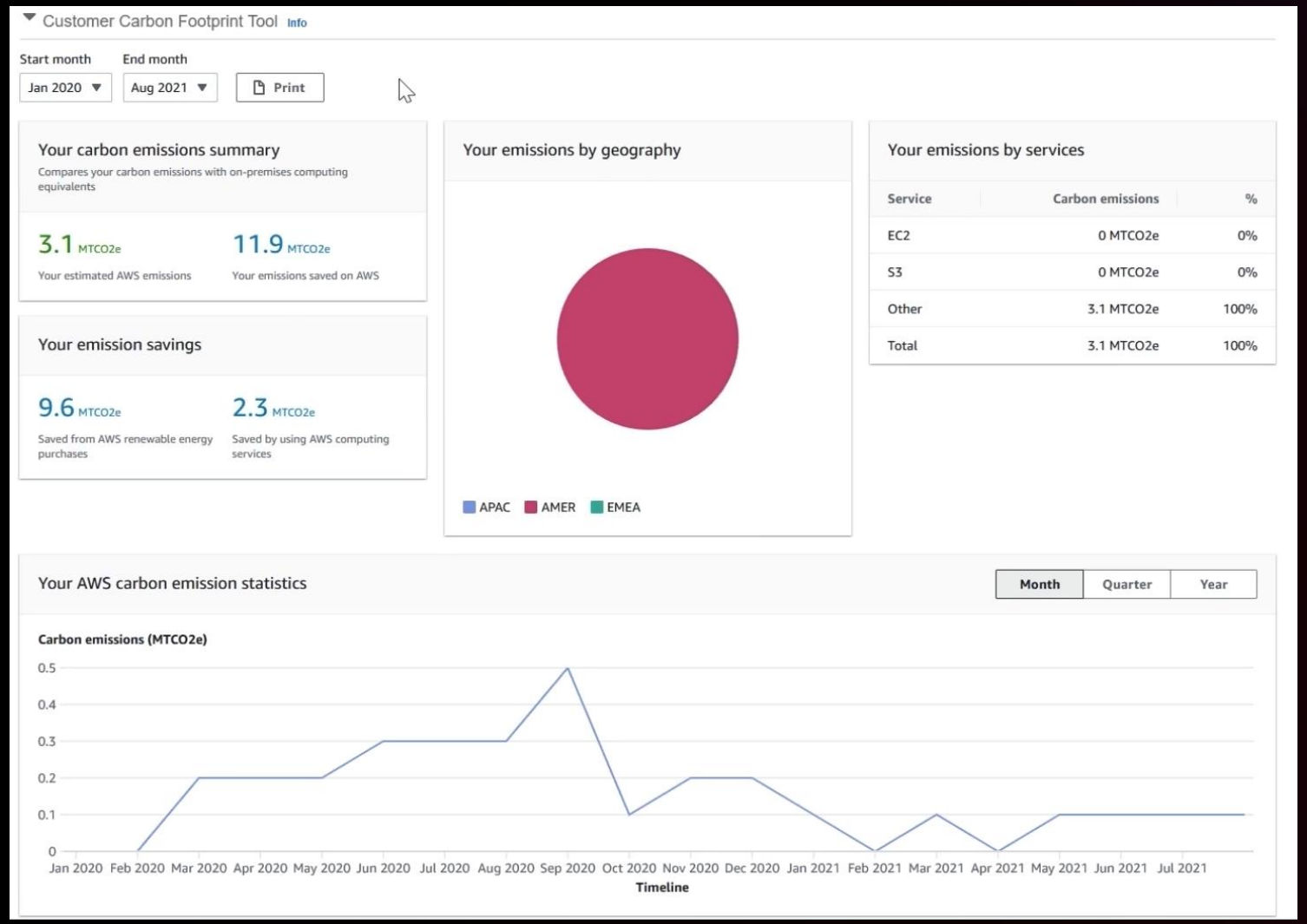


Asia Pacific
5 times more energy efficient

Find all the reports on aws.amazon.com/sustainability/resources/

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

AWS customer carbon footprint tool

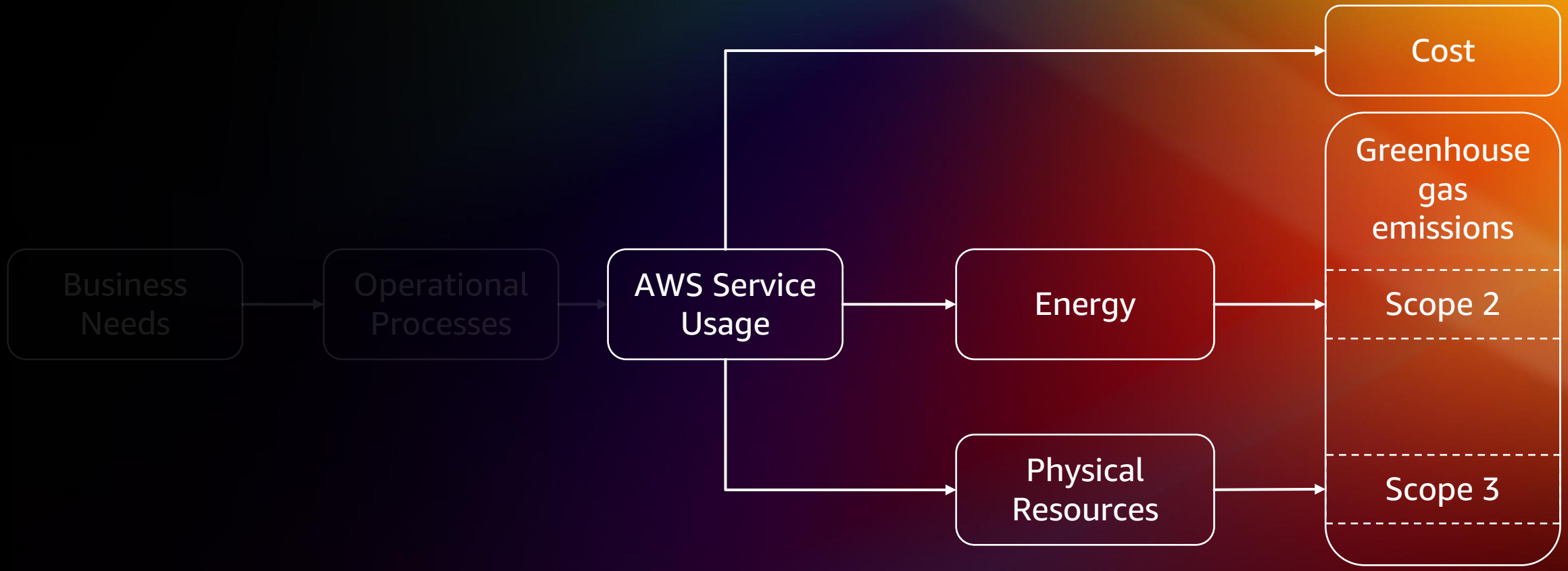


Calculate carbon emissions generated from your AWS workloads

Understand historical carbon footprint and review changes in emissions over time

Forecast changes as AWS moves forward on path to 100% renewable power in data centers by 2025

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



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

The instance configuration to exactly fit your needs

Categories

Burstable
(t3, t4)

General purpose
(m5, m6)

Compute optimized
(c5, c6)

Memory optimized
(r5, r6, x1, x2)

Storage optimized
(d3, i3, i4, im4, is4)

Accelerated
(p4, g4, f1, inf1)

Mac
(mac1)

+

Capabilities

Choice of processor
Intel (i), AMD (a),
AWS Graviton (g)

High frequency
up to 4.5+ GHz (z)

Extra memory
up to 16 TiB (e)

Instance storage
Local NVMe disks (d)

High networking
up to 400 Gbps (n)

Enhanced EBS performance
up to 60Gbps &
260,000 IOPS (b)

+

Instance size

Size
(.nano to .32xlarge)

Bare metal
(.metal)

+

Options

Linux, Windows, macOS

Amazon EBS

Amazon Elastic Inference

Amazon Elastic GPU

Elastic Fabric Adapter

=

500+ instance configurations

for virtually every business need

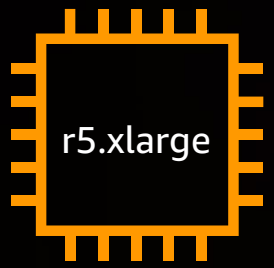
examples:
c6gn.8xlarge, m5zn.metal,
g4ad.16xlarge, r5b.large

aws

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

Attribute-based instance selection

Stop picking instances! Instead tell us what you actually need…





```
{
  "ArchitectureTypes": [ "x86_64" ],
  “VirtualizationTypes": [ "hvm" ],
  "InstanceRequirements":
   {
       "VCpuCount": { "Min": 4 },
       "MemoryMiB": { "Min": 32768 },
       "InstanceGenerations": [ "current" ]
   }
}
```

aws

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

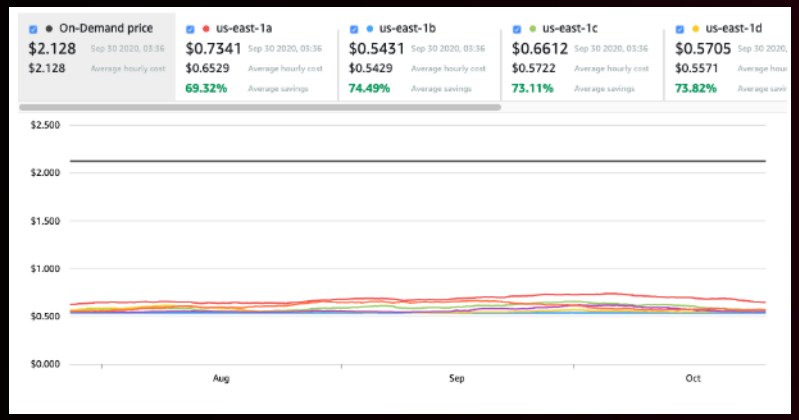
Amazon EC2 Spot Instances and Auto Scaling

Spot Instances are a spare EC2 capacity made available to customers as a purchase option with steep discounts



Spare EC2 capacity

Same infrastructure as on demand



Up to 90% less than on-demand pricing

Spot price is based on long-term supply and demand trends of EC2 instances (no bidding)



Interruptible

Spot Instances can be interrupted with a 2-minute warning if EC2 needs the capacity back

aws

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

Use instance and Availability Zone diversity to your advantage



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

Customers who have cost-optimized with Spot Instances

Saved **75% per month** on CI/CD workloads by changing four lines of code to incorporate Spot

Was able to **save 74%** on their K8s cluster by using EC2 Auto Scaling and Spot

Saved 40% over Reserved Instances and increased ETL **processing time 100%** with Spot Instances

Completed **2.5 million** tasks in 8 hours by spinning up an EC2 cluster with over **1 million vCPUs**

Reduced grid infrastructure **costs by 60%** with Spot Instances

Uses Spot Instances and auto scaling for its rendering-as-a-service workload to spend **less and scale more**

A containerized environment that would have taken more than a year to create previously only took **a few months.**

75% reduction in compute cost with Fleet

Reduced monthly compute **costs by 75%** while gaining more compute power

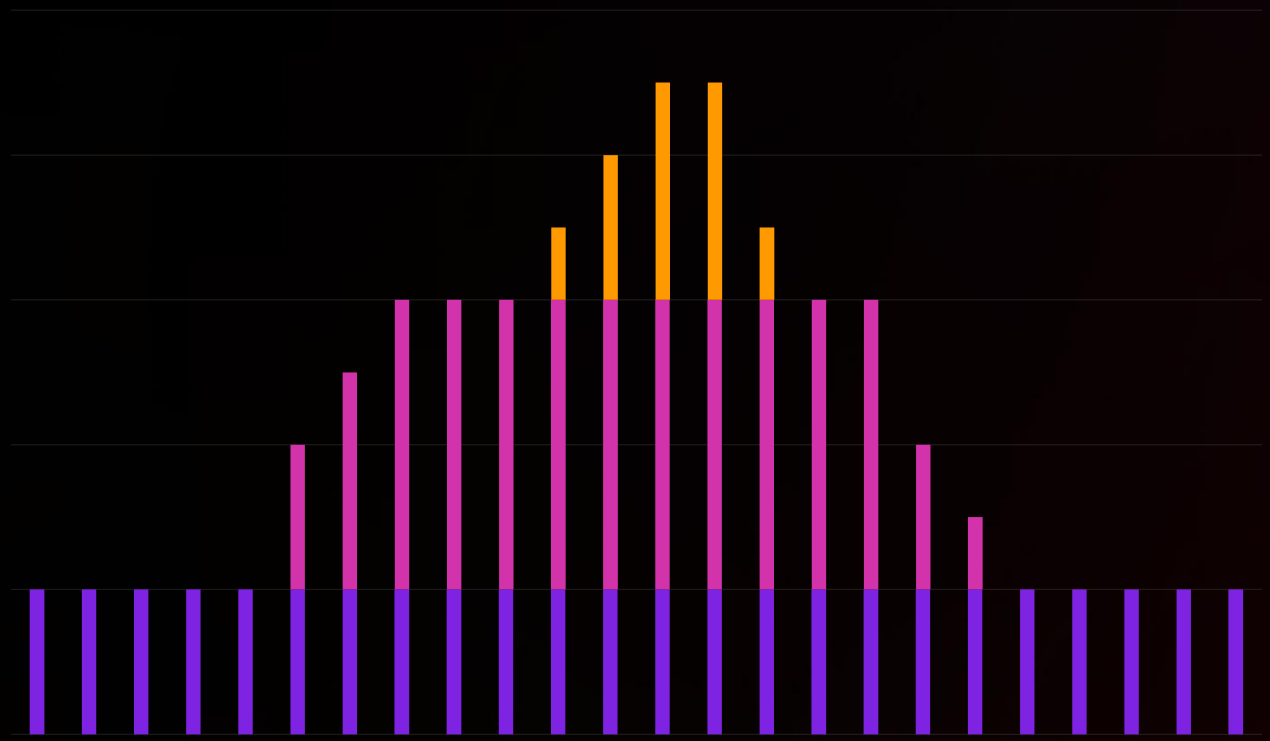
Processes over **100 billion** requests per day with an average response time of 90ms, saving over **$3M per year**

Decreased the time it took to analyze 10,000 biological samples from **7 years to 7 days** with Spot Instances

Saved 70% running Kubernetes workloads on AWS

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

Eliminating wasted spend by combining purchase options



Spot for fault-tolerant, flexible, stateless workloads

Discounts up to 90%

On-Demand, for new or stateful spiky workloads

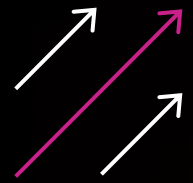
List price

Use **Savings Plans** for always-on instance hours

Discounts up to 72%

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

Migrate to Graviton: Up to 40% better price performance



Highest performance
in their instance families



20% lower cost
vs. same-sized
comparable instances



Up to 40% better price performance
vs. comparable instances

Best price performance within their instance families

aws

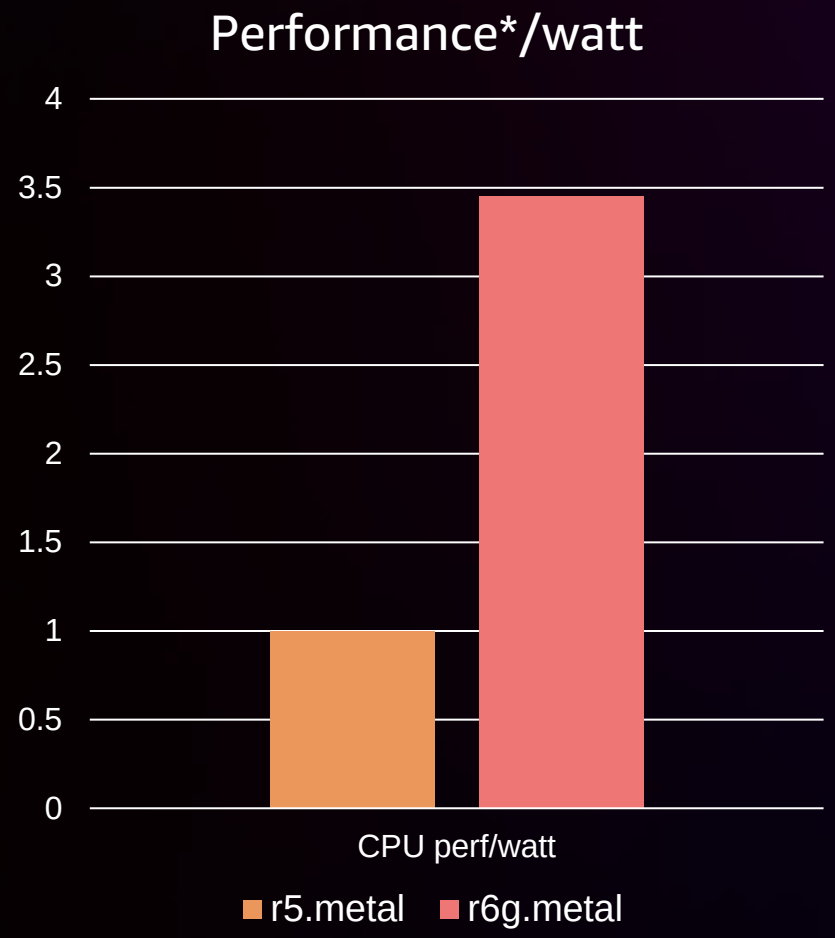
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Graviton CPU power efficiency

Graviton-based Amazon EC2 instances **use up to 60% less energy for the same performance** than comparable instances

Lower power

- Higher density
- Lower costs
- Lower carbon footprint



*Estimated SPECint2017

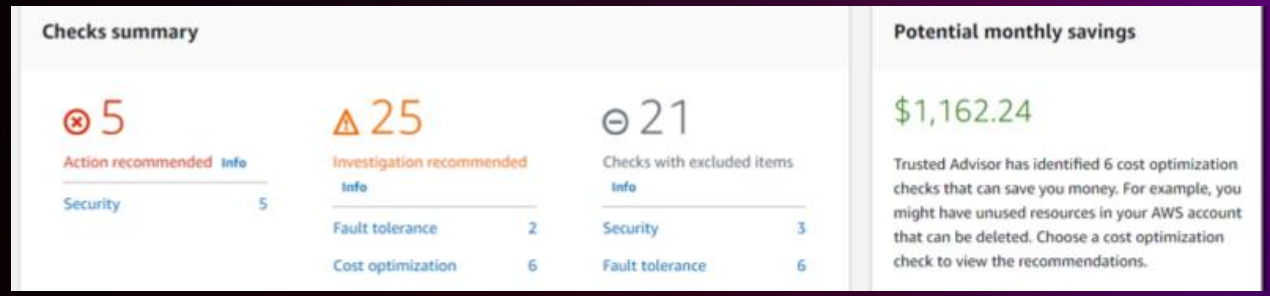
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

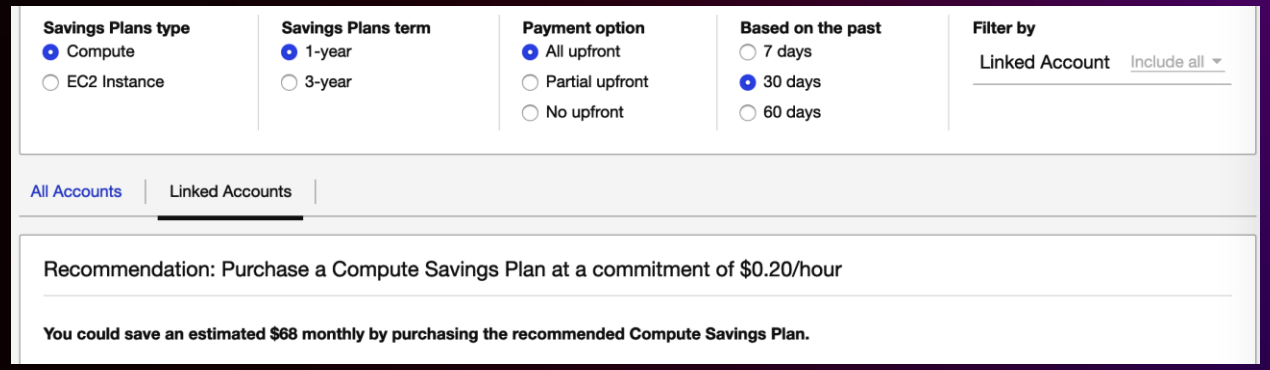
Automated recommendations

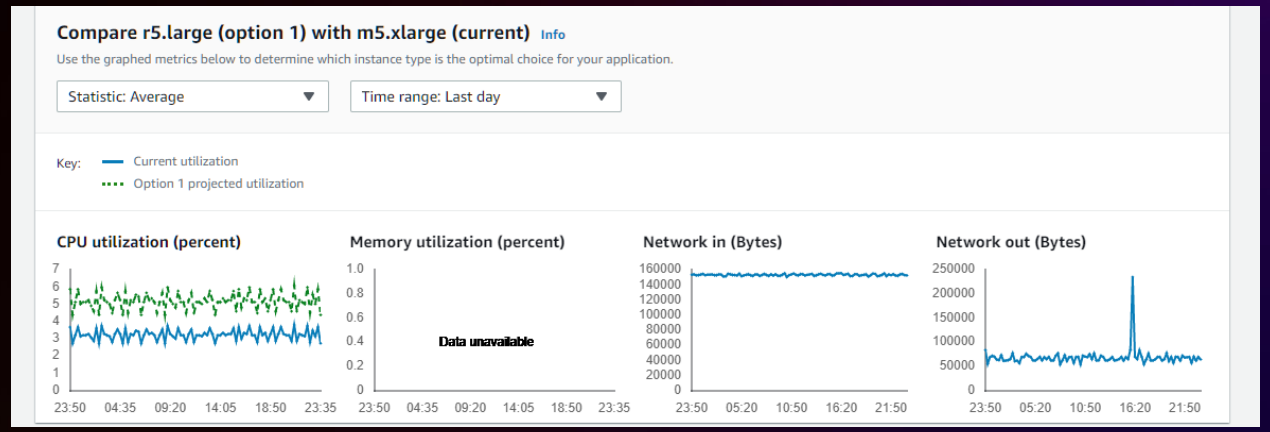
AWS Trusted Advisor: covers broad set of best practices including cost and utilization

AWS Cost Management: analyze savings plans and reservations, e.g., Amazon EC2, Amazon OpenSearch Service, Amazon ElastiCache, Amazon RDS

AWS Compute Optimizer: dive deep on recommendations based on metrics from EC2, Amazon EBS, AWS Lambda







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

Arm’s journey of optimizing for efficiency

Troy Gasaway (he/him/his)

VP of Infrastructure and Engineering
Arm

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

70% of the world's population use Arm processor technology

Semiconductor IP & Solutions Business

arm

240bn+
Arm-based chips shipped to-date

29bn+
Arm-based chips shipped in 2021

Mobile Revolution



To Cloud Disruption



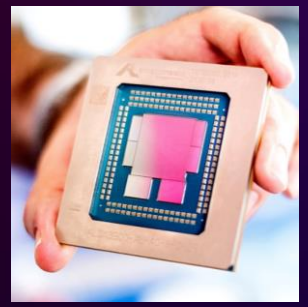
AWS (Annapurna Labs) is an Arm IP customer



AWS Nitro (Cortex-A72)



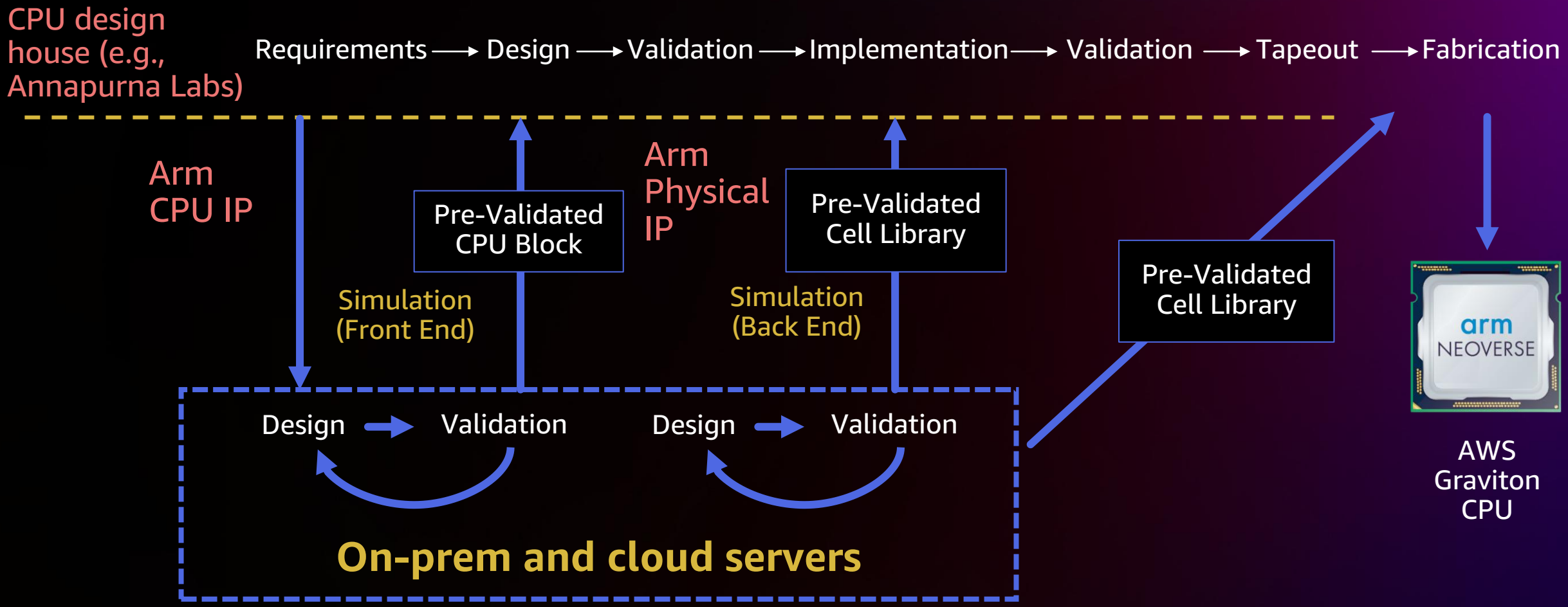
AWS Graviton2 (Neoverse N1)



AWS Graviton3 (Neoverse V1)

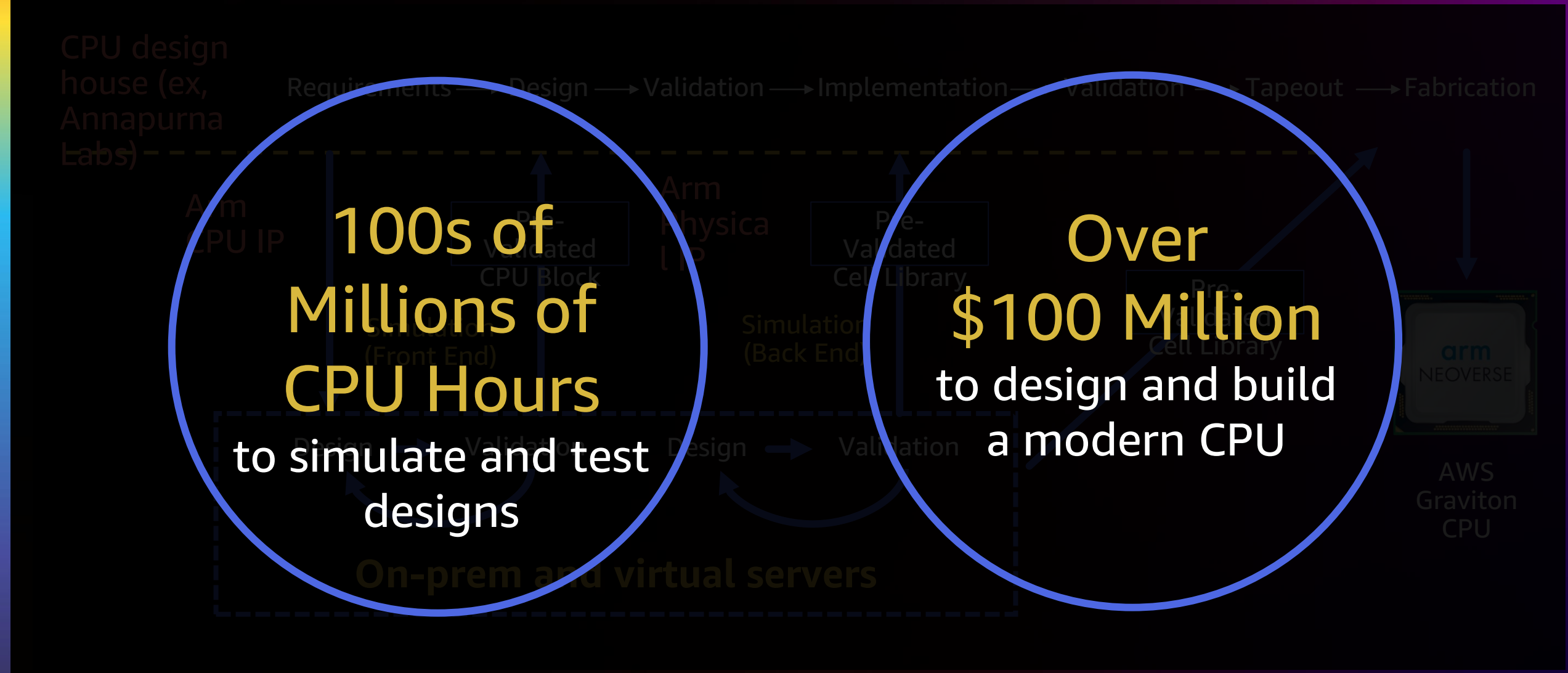
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Chip Creation



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

Chip Creation



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

Why Cloud for Arm IP Design and Verification?







Hardware efficiency

Scaling efficiency

Faster quality

Improve time-to-market

Maximize productivity

Reduce TCO

aws

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

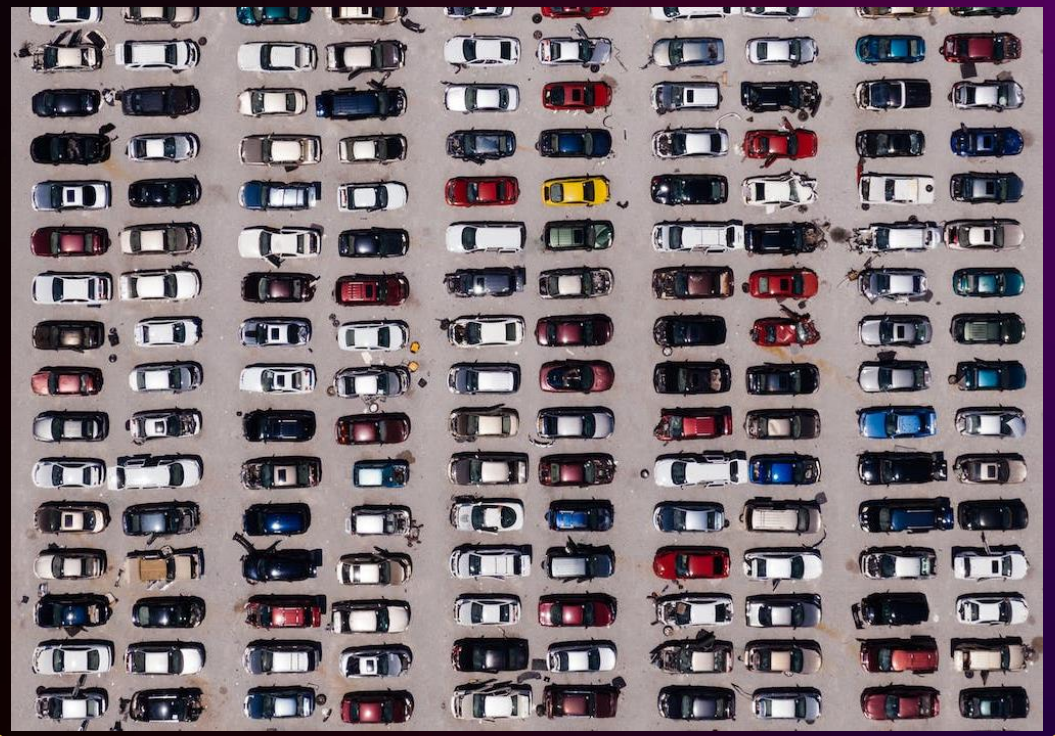
Cloud drives Arm's hardware efficiency

On-Prem





Cloud



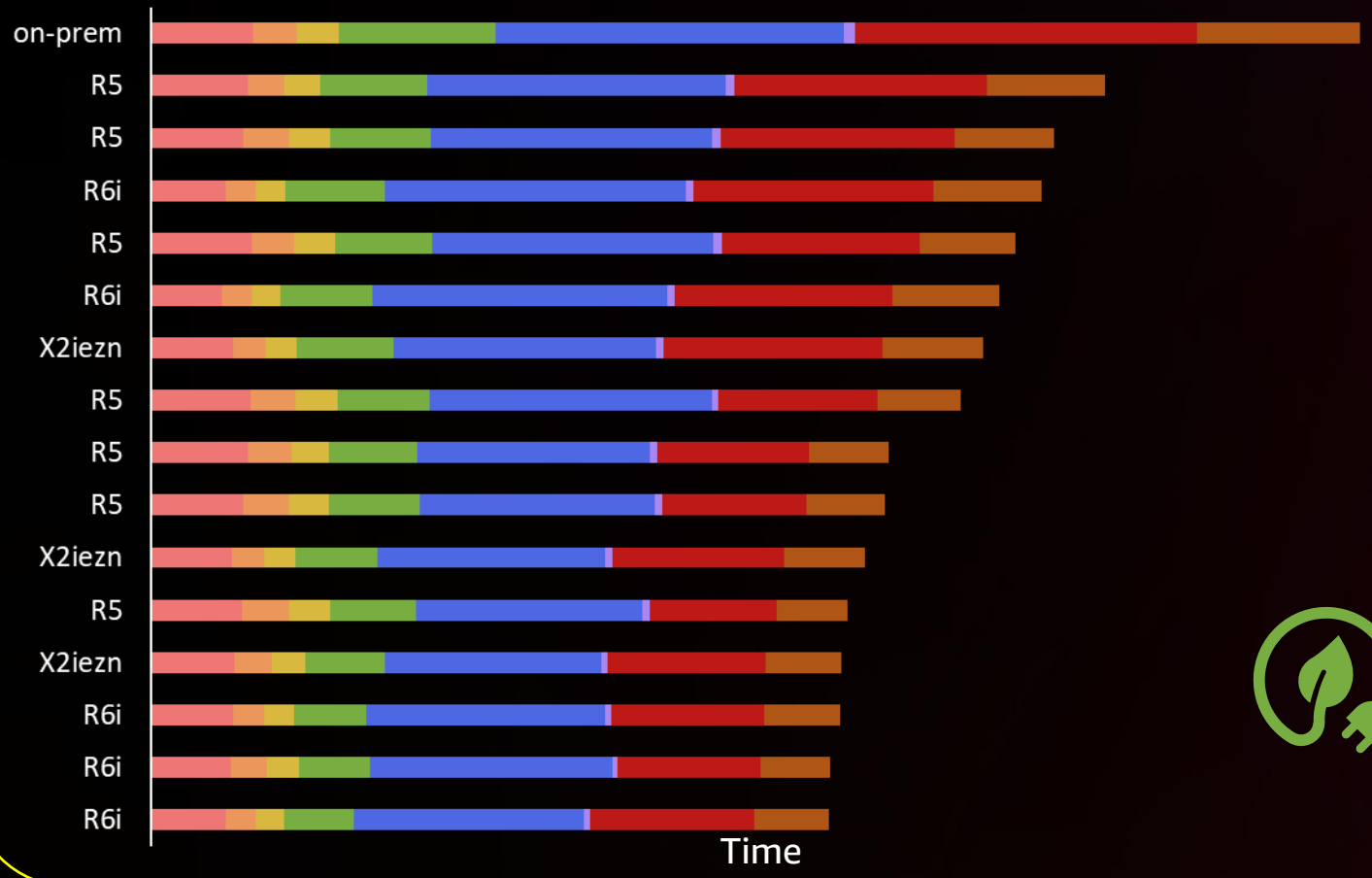
Instance is always right-sized for the job
No wasted resources

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

Choosing the best instance types

Runtime for different phases of an implementation workload

Lower is better



Graph is for a single EDA workload

Instances tested with different thread counts

All workloads are different – testing is crucial to identify the fastest configuration of application and instance type

Instance is always right-sized for the job

No wasted resources

aws

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

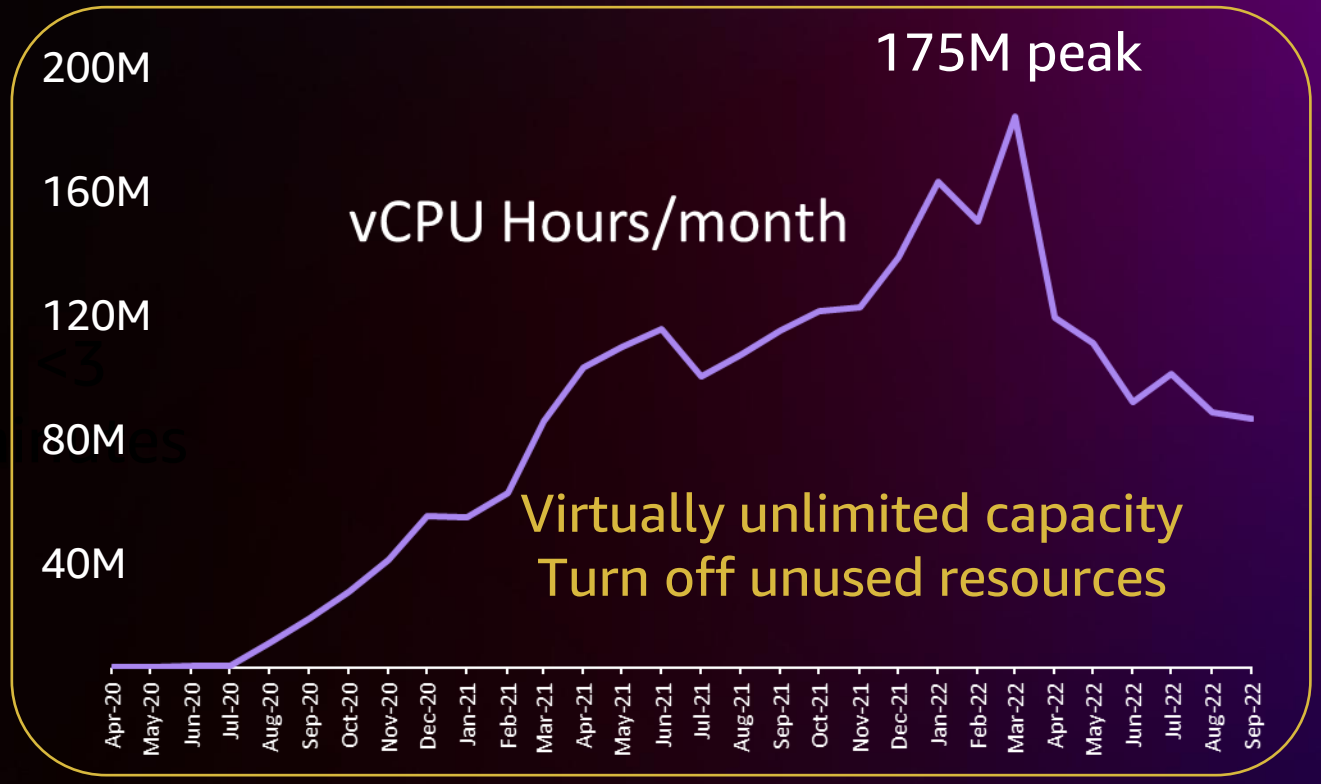
Cloud drives Arm's scaling efficiency

On-Prem



>60% of time wasted!

Cloud



No idle, unproductive resources
AWS Cloud datacenter efficiency >> Arm on-prem efficiency

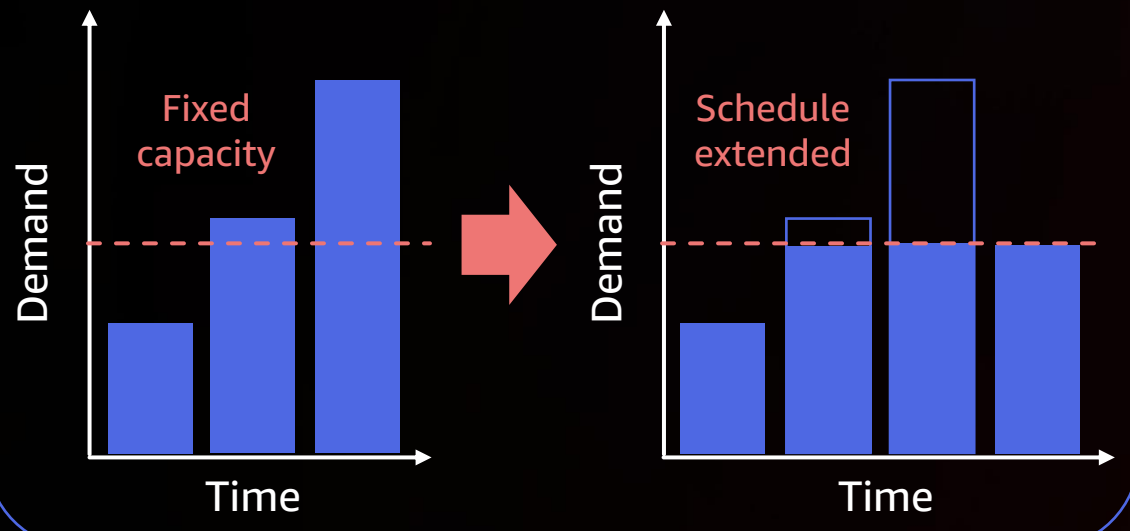
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cloud drives Arm's scaling efficiency, continued

On-Prem

Fixed capacity (125k "slots")

Had to "beat demand flat"



Cloud

Unlimited capacity (350k+ "slots")



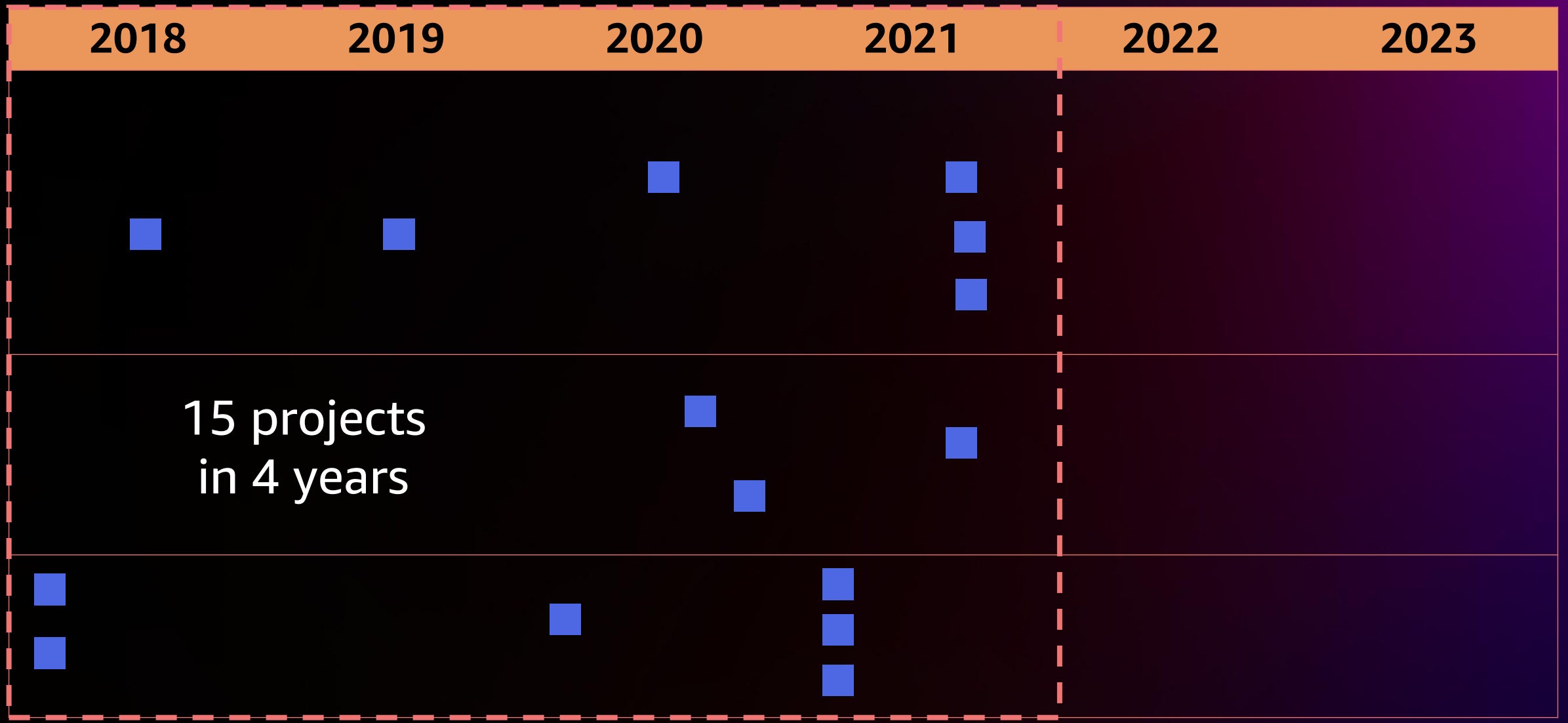
ML + Cloud scalability = faster quality

Makes for happier customers and faster time-to-revenue for all

Incremental jobs run more efficiently
AWS Cloud datacenter efficiency >> Arm on-prem efficiency

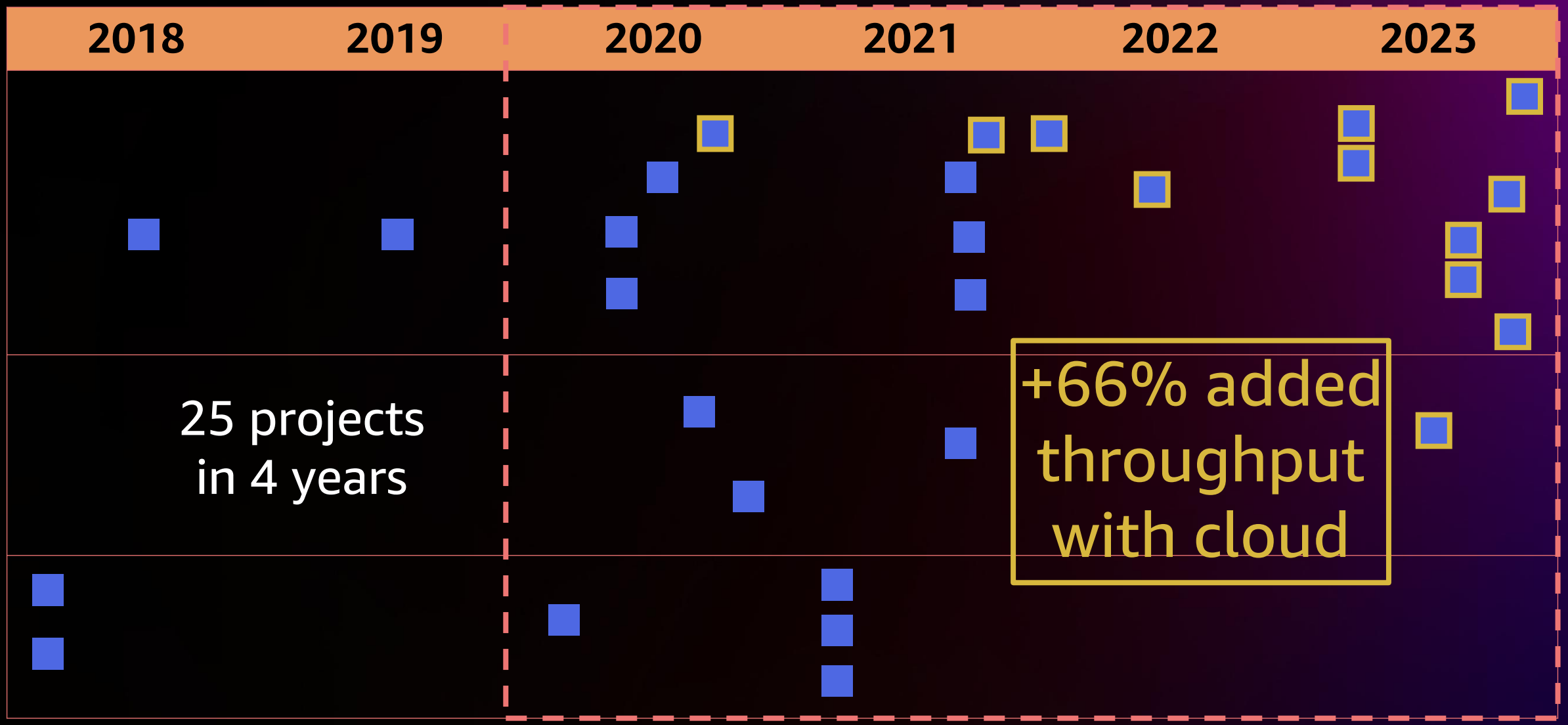
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cloud’s impact on Arm productivity (2018 roadmap)



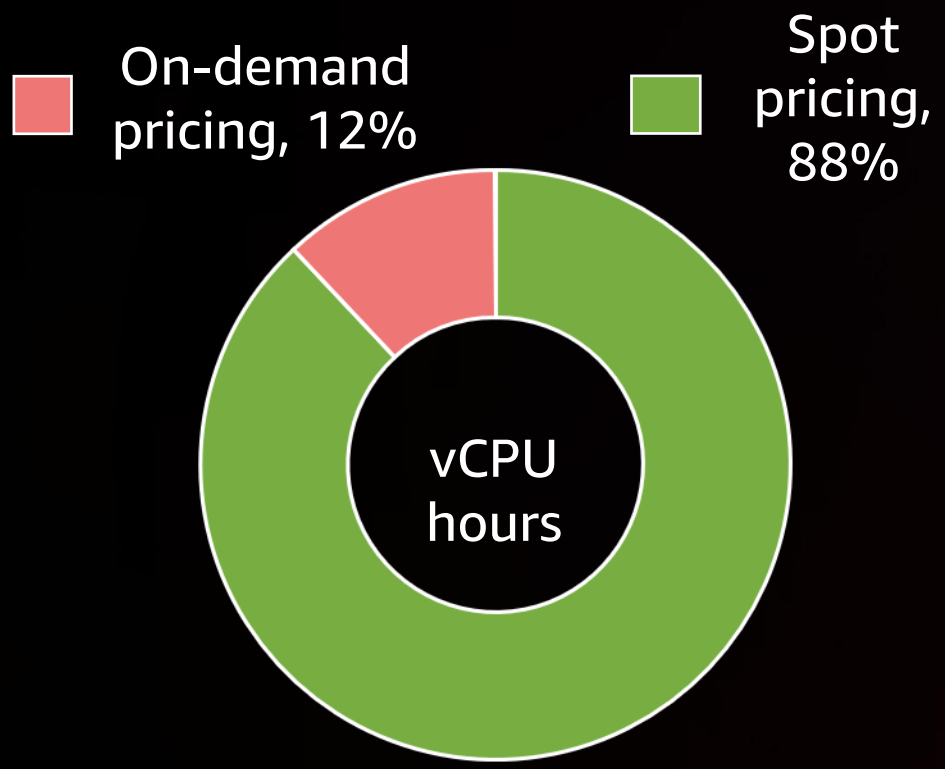
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

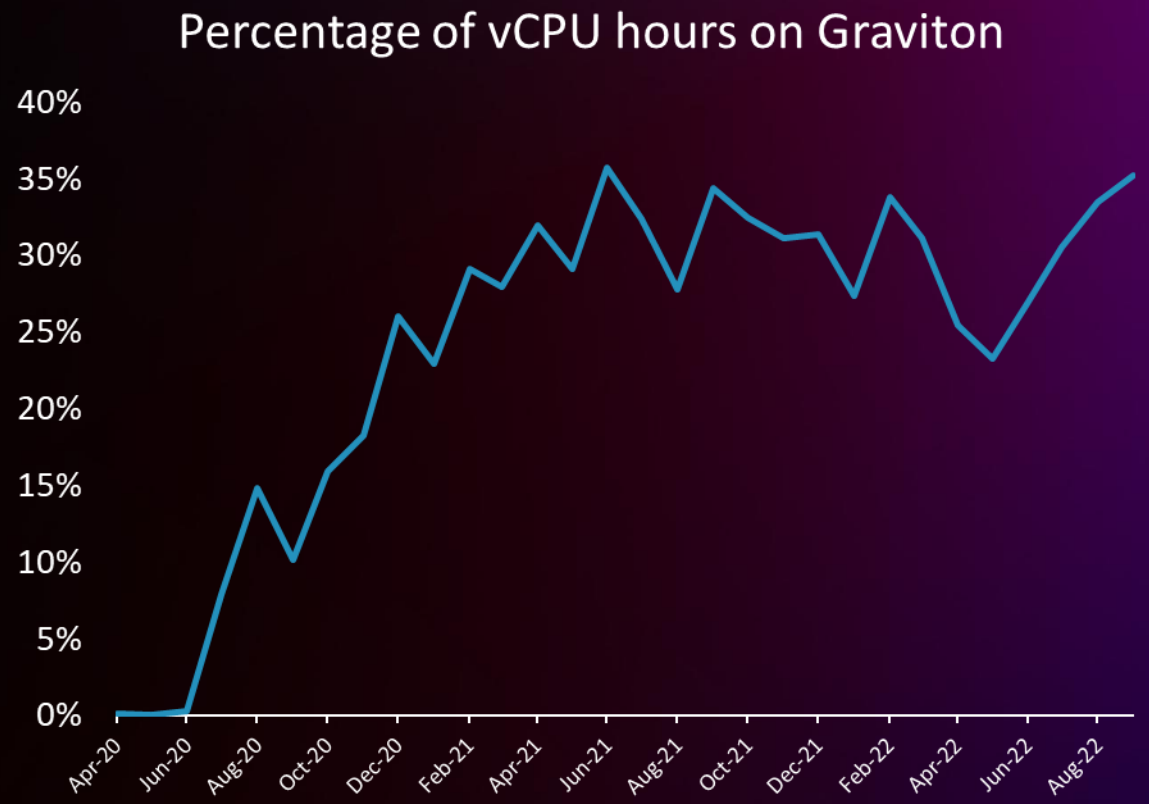
Cloud’s impact on Arm productivity (2022 roadmap)



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

Cloud drives Arm's cost efficiency

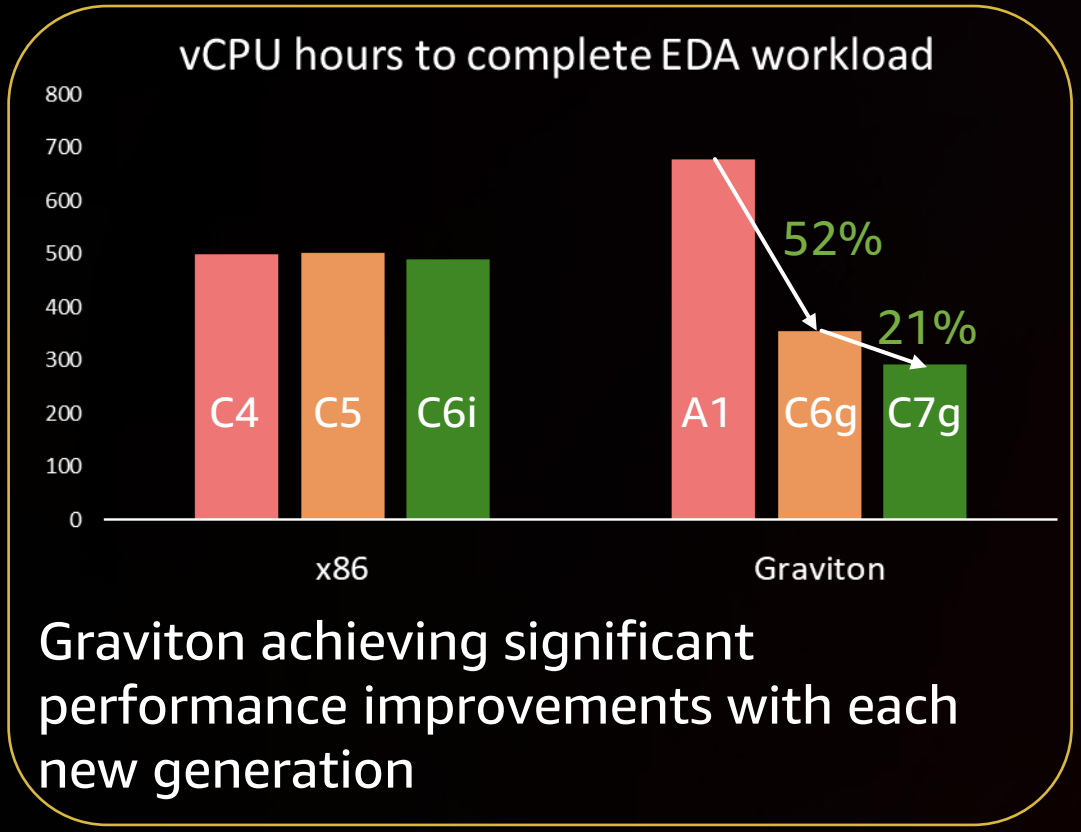




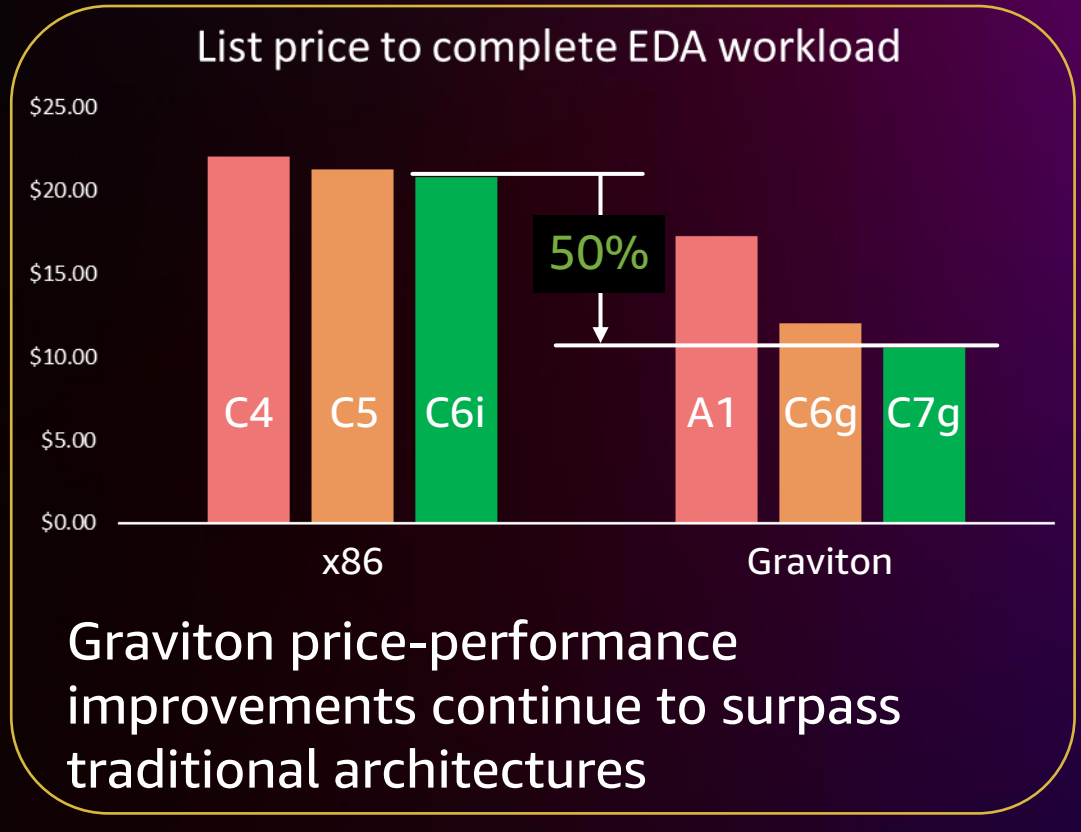
Spot turns idle CPU capacity into productive CPU capacity
AWS Graviton3 uses 60% less power per workload than equivalent x86-based instances

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

Price-performance gains with AWS Graviton



Graviton achieving significant performance improvements with each new generation



Graviton price-performance improvements continue to surpass traditional architectures

With AWS Graviton we can use fewer instances per workload, reducing our carbon footprint

Graviton offers up to 3.5x better performance/watt over comparable x86

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

But first, we had to enable a hybrid environment

On-Prem





Hybrid Enablement

CloudRunner







Cloud Foundation









Cloud

aws

Cloud B

Cloud C



CloudRunner and Cloud Foundation are the glue enabling Arm EDA workloads to run in the (multi) cloud and use AWS Graviton

aws

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

Summary



Capacity constrained to engineering constrained



Up to 6x improved throughput

1000x faster bug finding



Growing spot usage with AWS Graviton

Best perf/$ and perf/W

aws

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

Bringing it all together

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

Managed services increase utilization



Amazon EC2

Amazon Elastic Container Service
or
Amazon Elastic Kubernetes Service on EC2

Containers services on AWS Fargate

AWS Batch

AWS Lambda

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

Your journey for efficiency

Track your cost & usage as normalized KPIs, prioritize for impact

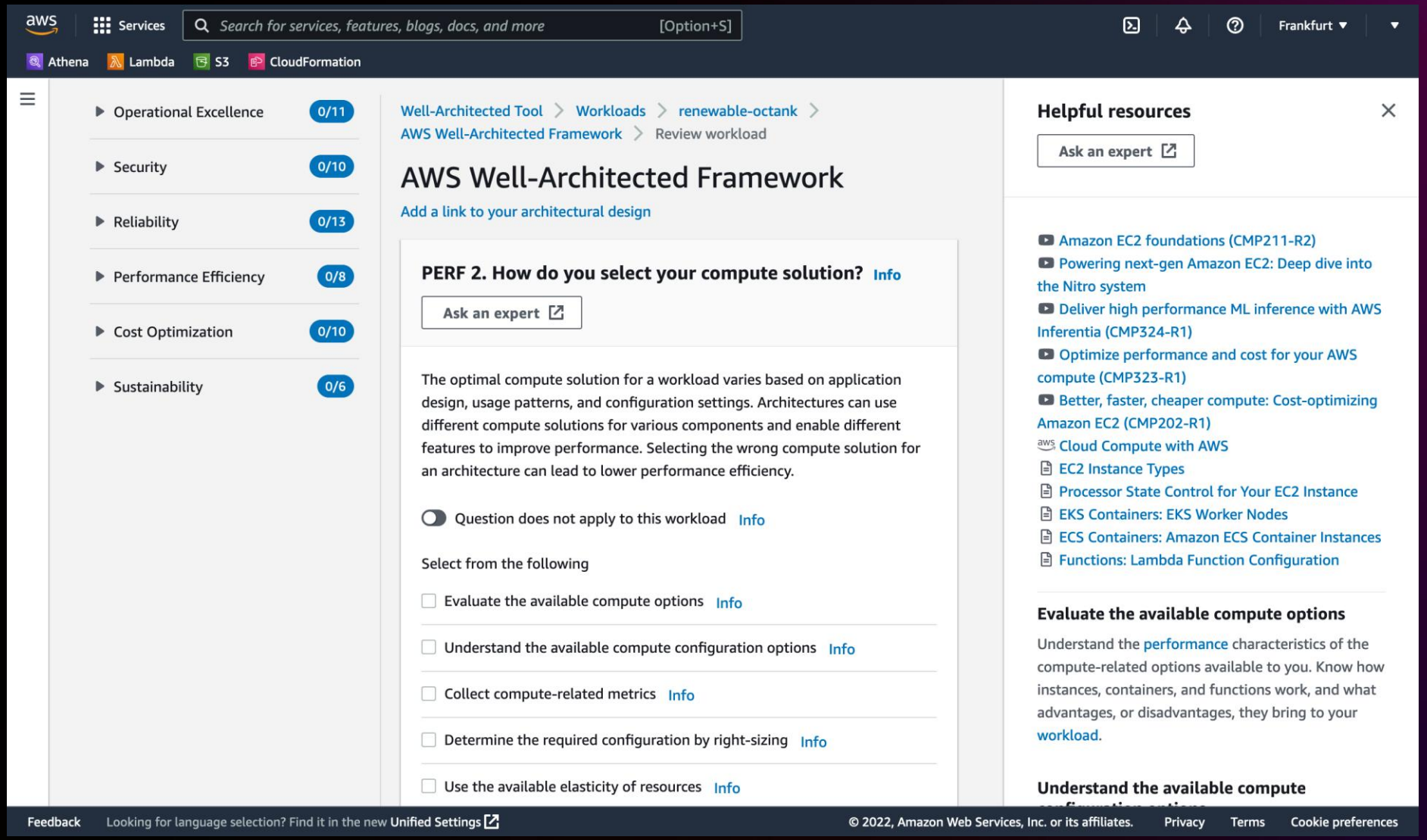
Optimization for efficiency is a shared responsibility, use higher level services to shift responsibilities to the cloud.

Leverage guides and tools to identify optimization potential for your accounts:

- AWS Compute Optimizer
- AWS Cost Management Console
- AWS Trusted Advisor
- AWS Well-Architected Tool & whitepapers

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

Dive deeper with AWS Well-Architected



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

Related compute & sustainability sessions

CMP328: Lower your IT Carbon Footprint with AWS Compute Services

CMP327: AWS Graviton deep dive

SUS204: Sustainability in AWS Global Infrastructure

SUS205: Architecting sustainably and reducing your AWS carbon footprint

SUS206: Sustainability and AWS Silicon

SUS301: Sustainable machine learning for protecting natural resources

SUS303: Delivering sustainable, high-performing architectures

SUS209: Optimizing architectures for sustainability

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

Thank you!

Adam Boeglin
boeglina@amazon.com

Steffen Grunwald
steffeng@amazon.de

Troy Gasaway
troy.gasaway@arm.com



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

aws

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