WEBINAR

Cost Optimization

Achieving More, Spending Less

Cost Optimization Specialist

Alex Piric
apiric@amazon.com

Agenda

1. Introduction to AWS Cost Optimization
2. Right Sizing and Modernizing
3. Elasticity
4. Optimizing Storage
5. Leveraging the right Pricing Models
6. Controlling your costs with AWS Cost Explorer tools
7. Live Q&A

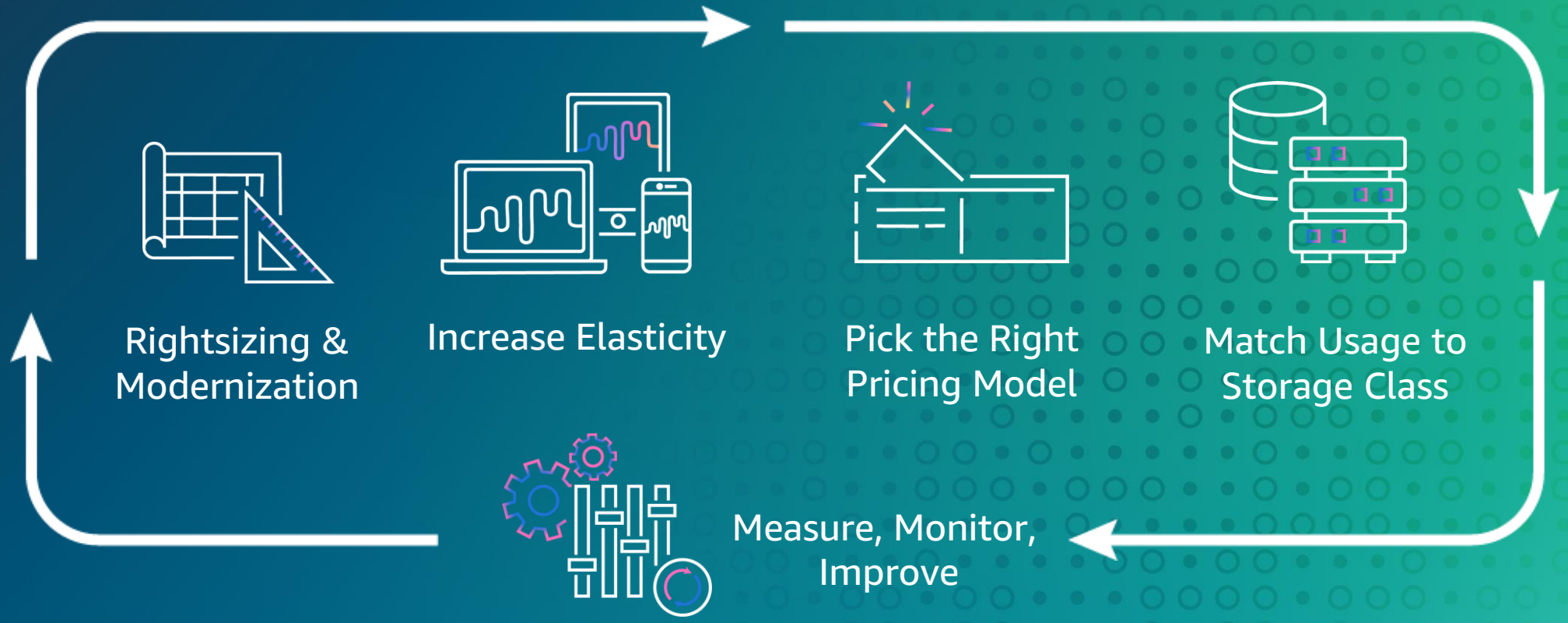
Why is cost optimization important?

In the chat…

Please rate the importance of Cost Optimization for you & your company on a scale from

1 (lowest) to 5 (highest)

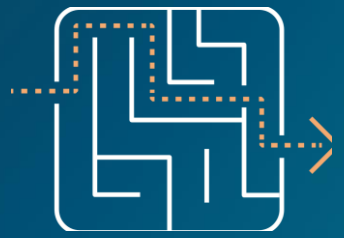
The 5 Pillars of Cost Optimization



Pillar 1:
Rightsizing & Modernization

Rightsizing

- Definition: The process of matching instance types and sizes to your workload performance and capacity requirements at the lowest possible cost.



Agility

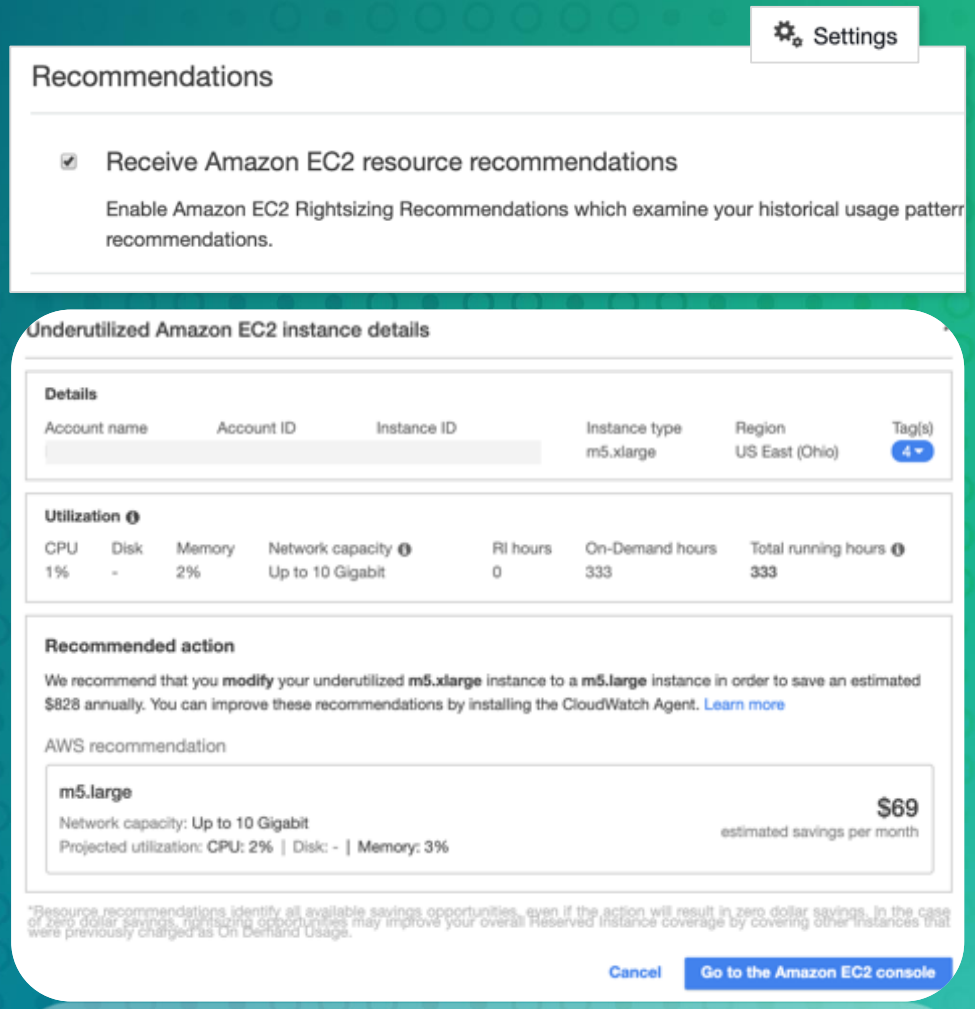
Lowest Possible Cost

Continuous Analysis

Rightsizing – Get Started

AWS Cost Explorer will provide you with EC2 rightsizing recommendations:

1. Enable rightsizing recommendations in the AWS Cost Explorer
2. Review individual rightsizing recommendations
3. Perform the EC2 instance size modifications



Rightsizing Tools

AWS Cost Explorer

Recommendations for:
EC2

Pricing:
- **Free**

AWS Compute Optimizer

Recommendations for:
EC2, ECS, Fargate, EBS, Lambda

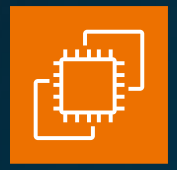
Pricing:
- *Default:* **Free**
- *Enhanced:* Costs per Resource and Hour

AWS Trusted Advisor

Recommendations for:
EC2, EBS, RDS, Lambda, Redshift

Pricing:
- **Included** (Cost Optimization Recommendations available with Business and Enterprise Support Plan)

Modernization



Graviton

Adopting Graviton Processors

Highest performance
in their instance families

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

Up to 40% better price-perormance
vs comparable x86-based instances

Best price-performance within their instance families

EBS

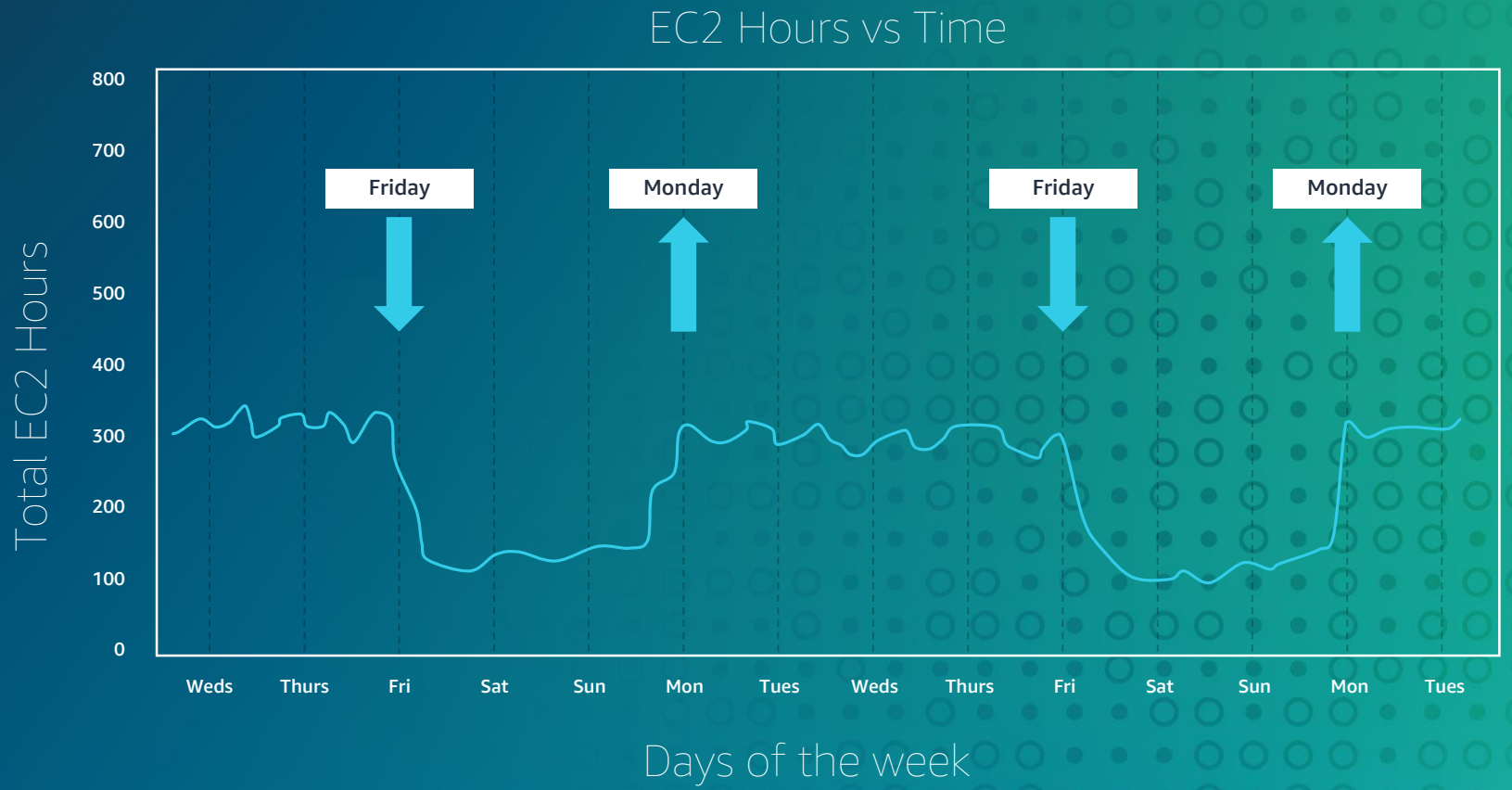
Adopting GP3

20% cheaper
with the same exact performance

Higher performance at lower cost
provision throughput and IOPS
separately from capacity

Pillar 2:
Increase Elasticity

Pay for what you need



Managing Elasticity

AWS Instance Scheduler

Functionalty:
Custom start & stop schedules
Multiple schedules per instance
Deployed via AWS CloudFormation

Supported Services:
EC2, RDS

Pricing:
Dependent on configuration

AWS Auto Scaling

Fuctionality:
React dynamically to changes in load
Schedule regular workload

Supported Services:
EC2, EC2 Spot Fleets, ECS, DynamoDB, Aurora

Pricing:
Free

Pillar 3:
Pick the Right Pricing Model

Pricing Models

On-Demand

Pay for compute capacity by the second with no long-term commitments

Spiky workloads, to define needs

Reserved Instances (RI)

Make a 1 or 3-year commitment and receive up to 72% discount off On-Demand prices

Committed & steady-state usage

Savings Plans

Commit to consistent usage (e.g. $10/hour) over 1-3 years and receive up to 72% discount off On-Demand prices

Savings like RI but more flexibility

Spot Instances

Use spare EC2 capacity at savings of up to 90% off On-Demand prices

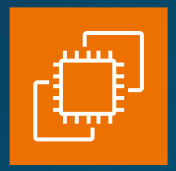
Fault-tolerant, flexible, stateless workloads

Payment Options: All Upfront | Partial Upfront | No Upfront

Default

Can be booked any time depending on workload and commitment

Reserved Instances



Amazon EC2

Standard **Covertible**

Region specific

AZ, instance size, networking type changeable

instance family, OS, tenancy changeable

EC2 Instance SP same savings more flexible

Compute SP same saving more flexible

Amazon RDS

Amazon Redshift

Amazon ElastiCache

Amazon OpenSearch

Region specific

Instance size flexibility*

No instance size flexibility

*Only for Amazon Aurora, PostgreSQL, MySQL, MariaDB,& BYOL Oracle DB

No Saving Plans available

Savings Plan - Save up to 72%

What is it?

A new flexible pricing model that helps you save up to 72% on Compute (EC2, Fargate, and Lambda) or Amazon SageMaker usage.

Customers simply commit to a consistent amount of usage (e.g. $10/hour) over 1 or 3 years, and in exchange, they will receive a lower price for that usage.

How does it work?

* The hourly commitment is valid for the duration of the plan term

Types of Savings Plans

Compute Savings Plans

Offer the greatest flexibility, up to 66% off (same prices as Convertible RIs)

FLEXIBLE ACROSS

- ✓ Instance family: E.g. Move from C5 to M5
- ✓ Region: E.g. change from EU (Ireland) to EU (London)
- ✓ OS: E.g. Windows to Linux
- ✓ Tenancy: E.g. switch Dedicated tenancy to Default tenancy
- ✓ Compute options: E.g. move from EC2 to Fargate, Lambda

EC2 Instance Savings Plans

Provide the lowest prices, up to 72% off (same as Standard RIs) on the selected instance family (e.g. C5 or M5), in a specific AWS region

FLEXIBLE ACROSS

- ✓ Size: E.g. move from m5.xl to m5.4xl
- ✓ OS: E.g. change from m5.xl Windows to m5.xl Linux
- ✓ Tenancy: E.g. modify m5.xl Dedicated to m5.xl Default tenancy

EC2 Spot instance

Fault-tolerant

Flexible

Loosely coupled

Stateless

Containers

Web Services

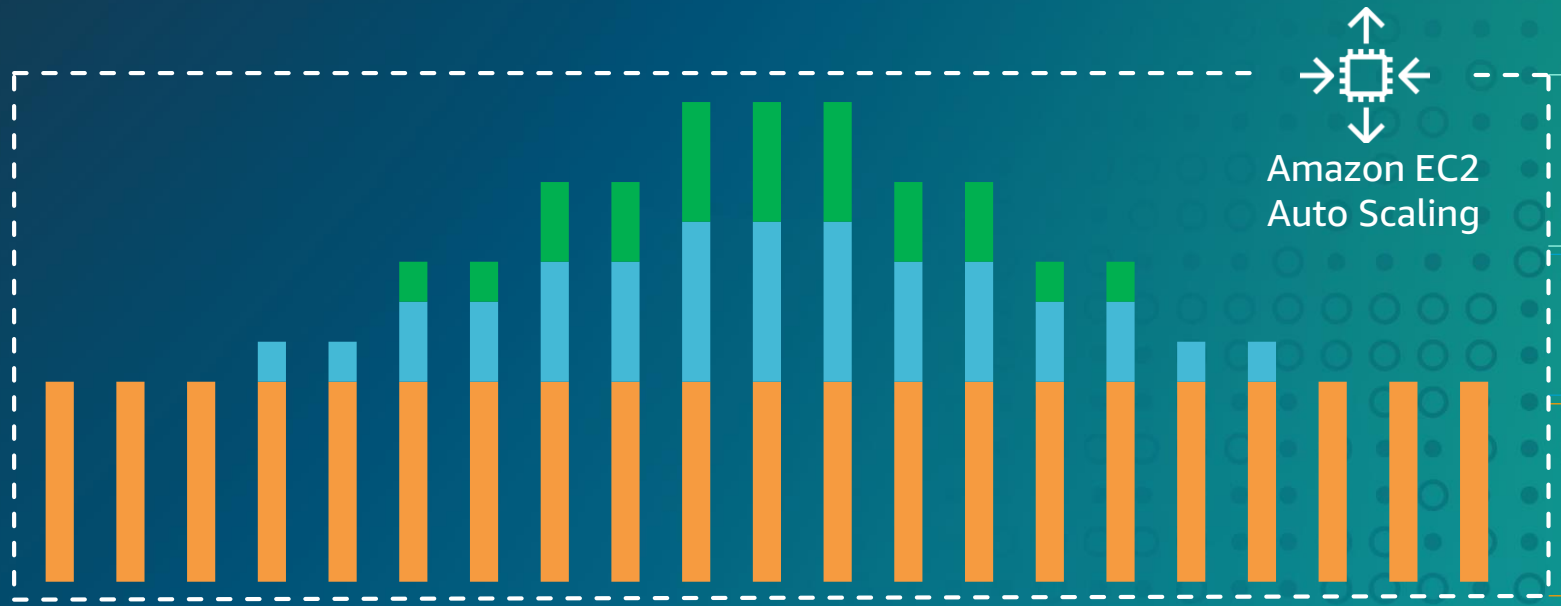
High Performance Compute (HPC) + Batch

CI/CD

Big Data

AI/ML

Mix & Match - Purchase models



Scale using EC2 Spot for short-lived, fault-tolerant, stateless workloads (up to 90% savings)

Scale using On-Demand for short-lived, stateful workloads

Use Savings Plans/ Reserved Instances for known/steady-state workloads (up to 72% savings)

Pillar 4:
Match Usage to Storage Class

Amazon S3 Storage Classes

S3 Intelligent-Tiering

S3 Standard

S3 Standard-IA

New

S3 Glacier Instant Retrieval

S3 Glacier Flexible Retrieval (formerly S3 Glacier)

S3 Glacier Deep Archive

S3 One Zone-IA

S3 Outposts

AWS Region ≥ 3 Availability Zones

Data with changing access patterns	Frequently accessed data	Infrequently accessed data	Rarely accessed data	Archive data	Long-term archive data
• Milliseconds access • No retrieval charge • Object monitoring charge • **Archive Instant Access tier** New • Opt-in Async Archive tiers	• Milliseconds access	• Milliseconds access • Retrieval charge per-GB	• **Milliseconds access** • **Minimum storage duration** • Retrieval charge per-GB	• Retrieval options from minutes to hours • **Free bulk retrievals** New • Retrieval charge per-GB	• Retrieval in hours • Retrieval charge per-GB

AWS AZ

Re-creatable, less accessed data

- Milliseconds access
- Retrieval charge per-GB

AWS Outposts

On-premises data

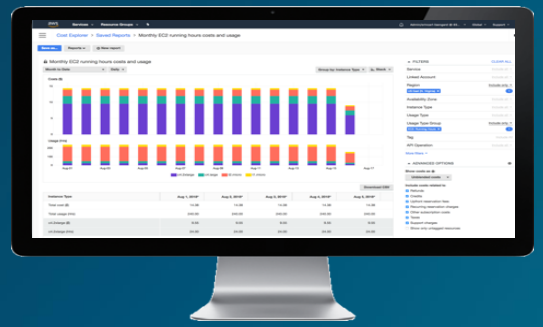
- Milliseconds access
- Retrieval charge per-GB

Pillar 5:
Measure, Monitor, Improve

AWS Cost Explorer

View top cost trends and group and filter data to establish a uniform view of your cost and usage.

- Measure, Monitor, Improve



Enabling AWS Cost Explorer: Access via Cost Management console. For first-time users, it may take up to 24 hours for cost and usage data to load.

Comprehensive dashboards

View month-to-date costs, month-end forecasted costs, and saved reports.

Automated trend analysis

Identify anomalous cost and usage events across your account(s) based on historical patterns.

Friendly user experience and programmatic access

Access a user-friendly interface for users of all expertise levels, and use the AWS Cost Explorer API for custom applications to address specific cost management needs.

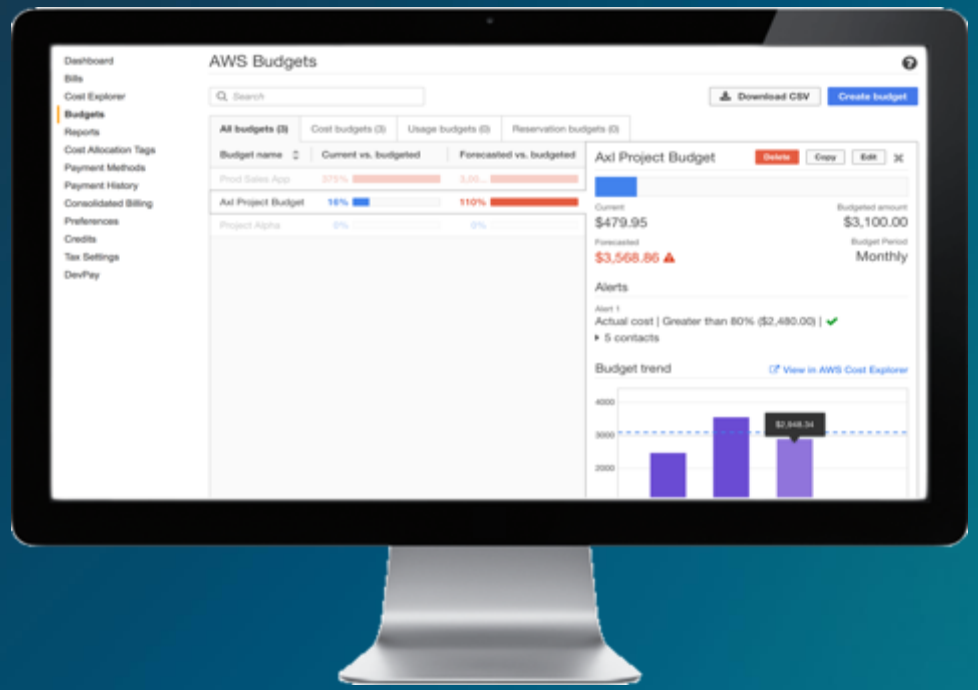
Tagging

- **Technical:**
- Application ID – Identify resources that are related to a specific application
- Environment – Distinguish between development, test, and production resources

- **Business:**
- Project – Identify projects that the resource supports
- Owner – Identify who is responsible for the resource

- **Security:**
- Confidentiality – Identify the specific data confidentiality level for a given resource

Budget Alerts



Customizable budgets

- Set custom budgets for cost and usage thresholds
- Set utilization and coverage targets for your Reserved Instances and Savings Plans

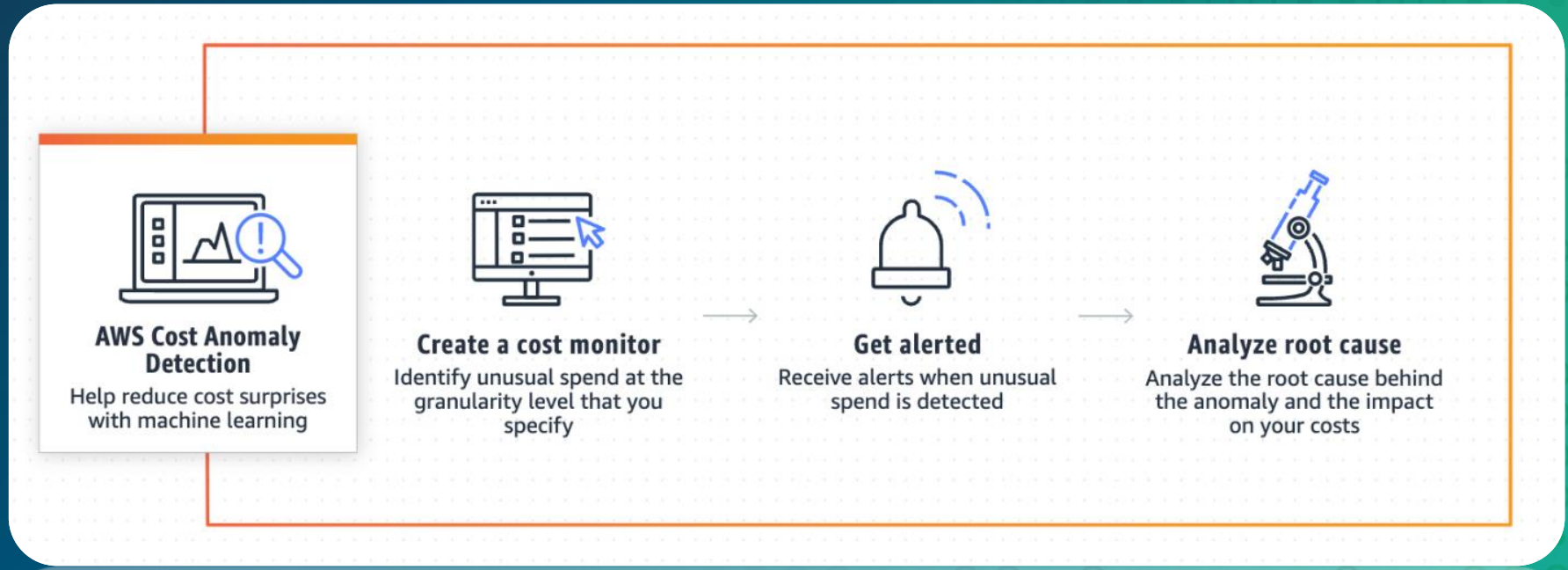
Ongoing monitoring and alerting

- Receive regular budget reports
- Be alerted when cost and usage exceed (or forecasted to exceed) budgeted amount

Flexible automation

- Execute actions when a budget exceeds its threshold (actual or forecasted amounts)

AWS Cost Anomaly Detection



- Benefit from automated cost anomaly detection and root cause analysis
- Define custom anomaly thresholds and receive alerts, either individually or on a daily or weekly cadence

Summary

- Rightsizing & Modernization

- Enable Rightsizing Recommendation in Cost Explorer
- Save with Modernization: e.g. Adopting Graviton/Latest Processors and GP3

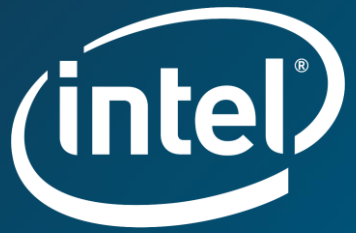
- Leveraging Different Pricing Models

- Leverage Saving Plans for your EC2 usage
- Leverage Reserved Instances for your OpenSearch, RDS and ElastiCache usage

Thank you!

Back Up
DeepDive: Modernization

AWS Instances

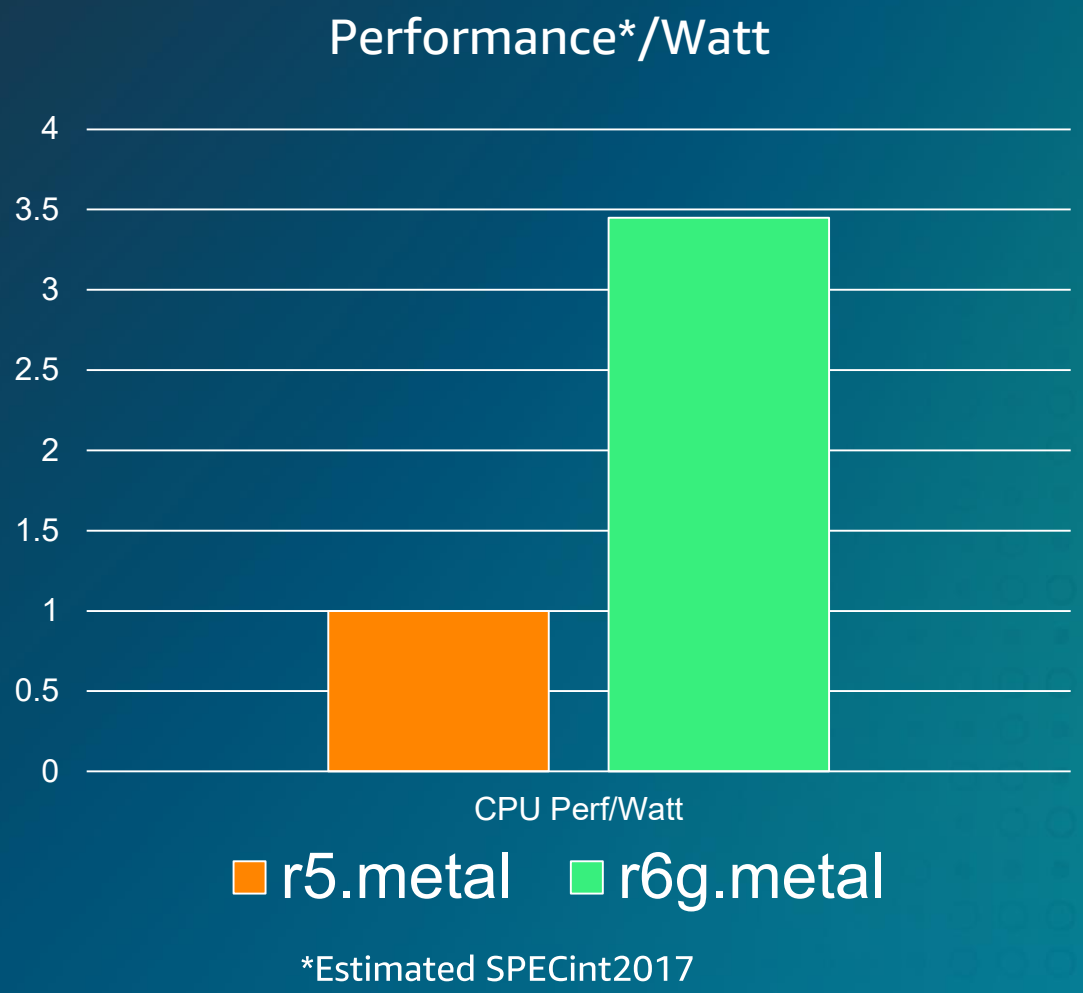


Intel® Xeon Scalable processors

AMD EPYC processors

AWS Graviton processors

AWS Graviton2 Processor



*Estimated SPECint2017

Power Efficiency

Lower power
→ Higher density
→ Lower costs
→ Lower carbon footprint

More Sustainable

AWS Graviton2-based Amazon EC2

Up to 40% better price-performance over COMPARABLE x86-based instances

M6g, M6gd

General purpose workloads

T4g

Burstable general purpose workloads

C6g, C6gd, C6gn

Compute-intensive workloads

R6g, R6gd, X2gd

Memory-intensive workloads

Im4gn, Is4gen

NEW!

Storage-intensive workloads

G5g

NEW!

GPU-based graphics and machine learning workloads

AVAILABLE ACROSS 23 AWS REGIONS GLOBALLY*

* Not all instances are available in all regions

aws

AWS managed services supporting Graviton2

- EXTENDING THE GRAVITON2 PRICE PERFORMANCE TO Managed SERVICES

RDS:	Graviton2 instances provide up to 35% performance improvement and up to 52% price/performance improvement for open source databases depending on database engine, version, and workload.
Aurora:	Graviton2 instances provide up to 20% performance improvement and up to 35% price/performance improvement for Aurora, depending on database size.
EMR:	Amazon EMR provides up to 35% lower cost and up to 15% improved performance for Spark workloads on Graviton2-based instances versus previous generation instances.
Elasticache:	Up to a 45% price/performance improvement over previous generation instances. Graviton2 instances are now the default choice.
OpenSearch:	Get up to 38% improvement in indexing throughput, 50% reduction in indexing latency, and 30% improvement in query performance when compared to the corresponding x86-based instances from the current generation (M5, C5, R5).
Lambda:	Run your functions on ARM and get up to 34% better price performance with Graviton2.
DocumentDB:	Achieve up to 30% better performance with Amazon DocumentDB (with MongoDB compatibility) using new Graviton2 instances.

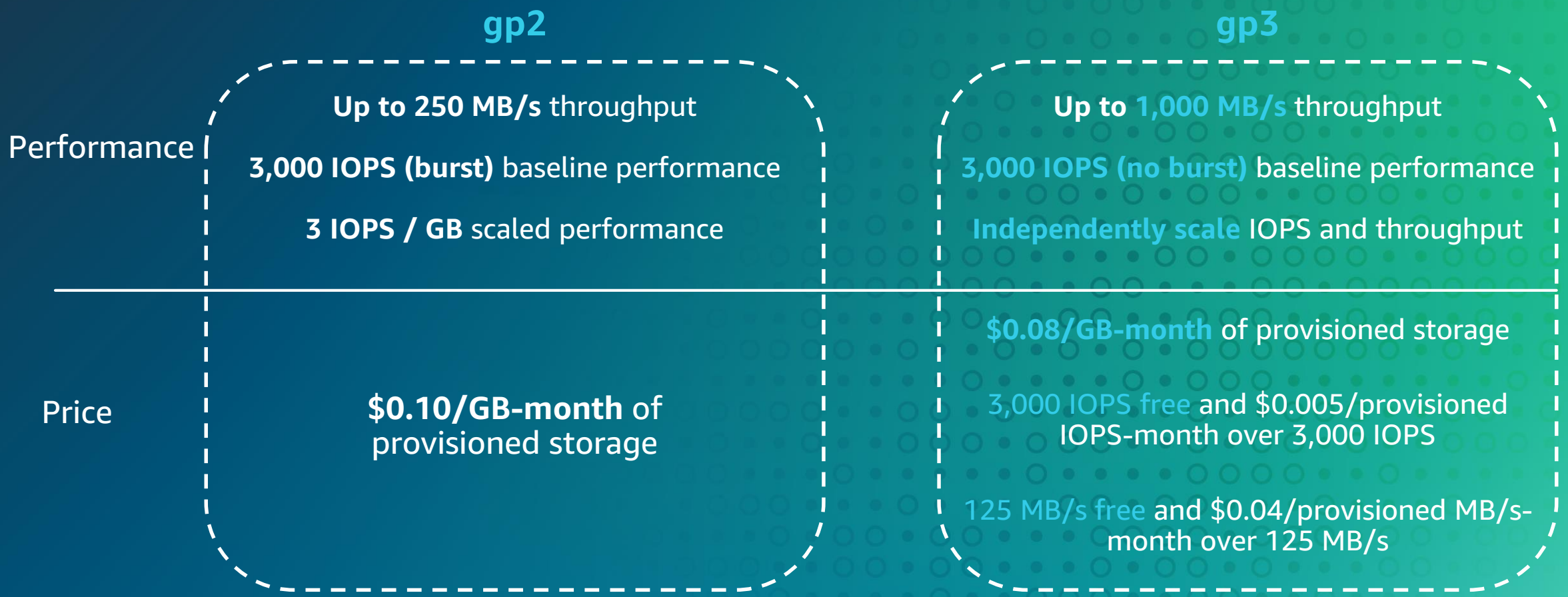
AWS Graviton- Ease of Adoption

AS A RULE, THE MORE CURRENT YOUR SOFTWARE STACK, THE BETTER

Difficulty	Workload	Actions
Virtually no effort	RDS, Aurora, ElastiCache, OpenSearch, MemoryDB, & Neptune	Upgrade to latest and enjoy
Super easy	EMR	Typically, just works
Pretty easy	AWS Lambda	Typically, just works with Lambda managed runtimes or base images.
Quite easy	Linux – portable programming languages (e.g., Java, PHP, Node.js)	Select Arm64 AMI and Install Bonus if containerized
More involved	Linux – Compiled languages (e.g., C/C++, Python, Go)	Select Arm64 AMI and compile
Some work, high reward	Microsoft Windows – .NET	Migrate to Linux + .NET core on Arm64
Not yet possible	Microsoft Windows	Microsoft Windows Server not yet available for Arm64

GP3: Optimized volume

Independently provision IOPS and throughput at up to 20% lower price per GB



Use Amazon EBS volumes to migrate from gp2 to gp3 without application impact

aws

Back Up
DeepDive: Pricing Models

RI & SP Similarities

Reserved Instances (RIs)

- Billing subscription (“discount coupon” that’s applied hourly)
- Commitment for 1yr or 3yr term
- Payment options: All/Partial/No Upfront
- By default discount is shared between all accounts in AWS Organizations family
- In AWS Organizations account that purchased RI or SP benefits from the discount first
- Amazon EC2: both RIs and SPs offer same* discount for same usage types

Savings Plans (SPs)

* Exception: SUSE platform. RIs offer greater discount for SUSE instances compared to SPs

Difference between Savings Plans & Reserved Instances?

Reserved Instances

Discounts apply to usage with an exact match to the EC2 Instance configuration

Have modest flexibility for configuration modification

Configuration modifications are manually managed

Savings Plans

Compute Savings Plans discounts apply to ***any*** compute usage automatically

Have modest flexibility for configuration modification

Configuration modifications are manually managed

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

Savings Plans for Amazon SageMaker

SageMaker Savings Plans

Offer up to 64% off on-demand ML instances

- ✓ Studio Notebook
- ✓ On-Demand Notebook
- ✓ Processing
- ✓ Data Wrangler
- ✓ Training
- ✓ Real-Time Inference
- ✓ Batch Transform

FLEXIBLE ACROSS

- ✓ Instance family: E.g. Move from C5 to M5
- ✓ Instance size: E.g. Move from p3.2x to p3.8x
- ✓ Region: E.g. change from EU (Ireland) to EU (London)

SageMaker components NOT covered by SPs:

Non On-Demand ML Instances

- SPOT ML Instance
- Feature Store
- Ground Truth (Tired pricing available)
- Edge Manager (Tired pricing available)
- ML Storage
- Inference Acceleration (ml.eia)
- Data Processing (Data Transfer)

Save up to 30% with

CloudFront Security Savings Bundles

Savings up to 30% off the On-Demand price

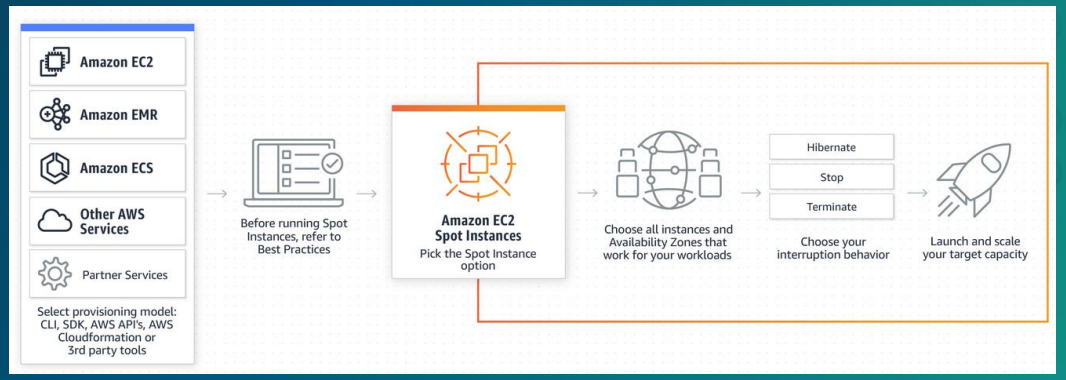
1-year term

Stable workloads

How it works:

- You commit to pay a consistent monthly amount (dollars per month) for CloudFront for one year.
- CloudFront automatically applies credits to your AWS bill that automatically offset CloudFront charges on your AWS bill resulting in up to 30% discount.
- In addition to the CloudFront credits, you get credits to offset the per-request charges for using AWS WAF. The amount of the AWS WAF credits is up to 10% of the amount of the monthly CloudFront commitment.

Save up to 90% with EC2 Spot Instances



- With Spot Instances , you pay the Spot price that's in effect for the time period your instances are running. These prices are set by Amazon EC2 and adjust gradually based on long-term trends in supply and demand for Spot Instance capacity. You can use Spot Instances for various stateless, fault-tolerant, or flexible applications such as big data, containerized workloads, CI/CD, web servers, high-performance computing (HPC), and test & development workloads.

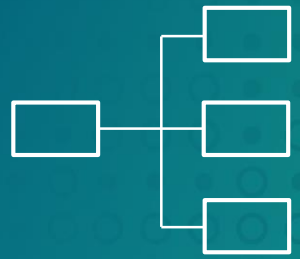
Back Up
DeepDive: S3 Storage Classes

Optimizing storage using Amazon S3

1

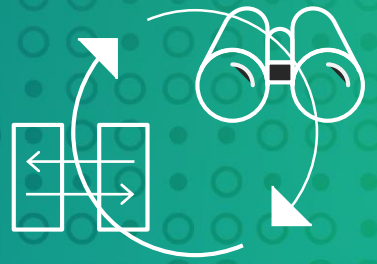
Define your workload requirements

2



Organize your data

3



Analyze, act, and measure

Analyze your Amazon S3 Storage

S3 Storage Lens

Organization-wide S3 usage and activity dashboard for cost optimization

Storage Class Analysis

Bucket-level analysis of retrievals for predictable workloads

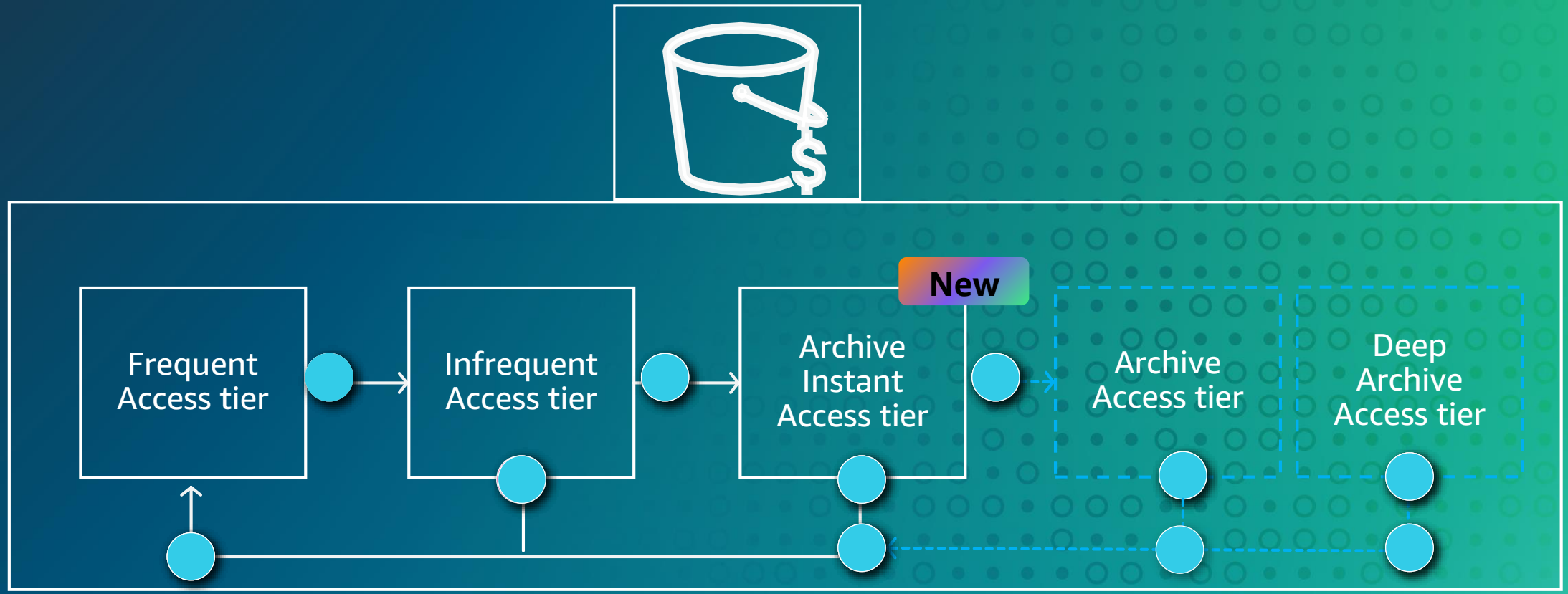
S3 Inventory

Object-level analysis for analytics and auditing

What is Amazon S3 Intelligent-Tiering?

- Only cloud storage that delivers automatic storage cost savings
- Moves objects between three access tiers for a small monthly monitoring and automation fee
- New Archive Instant Access tier delivers up to 68% lower cost, without any impact on performance
- No operational overhead, no lifecycle fees, and no retrieval fees
- Designed for 99.9% availability and 99.999999999% (11 nines) durability

Amazon S3 Intelligent-Tiering for data with unknown/changing access patterns



Amazon S3 Intelligent-Tiering Pricing Enhancements

New

Small objects not monitored or auto-tiered

Use S3 Intelligent-Tiering without analyzing object size distributions

Optimizing for short-lived objects

Use S3 Intelligent-Tiering without analyzing the average life of objects