Seven steps to lower costs while improving application performance

Boyd McGeachie (he/him)

Head of Go-To-Market Flexible Compute
AWS

Todays Agenda

1. Build and run your applications on AWS
2. Optimize costs and accelerate innovation with serverless computing
3. Choose the AWS compute instance type that matches your application needs
4. Select the compute purchase models that best fits your budget
5. Migrate to AWS Graviton for the best price performance for a broad set of applications
6. Optimize your workload price and performance with AWS Storage
7. Optimize your resource capacity to fit demand

> **"We don't want to make money from customers that aren't getting value from us… How many of your partners call you up and say 'stop spending money with us'?"**

Andy Jassy

CEO, Amazon

Build and run your applications on AWS

Why cloud infrastructure?

Security

Availability

Performance

Scalability

Flexibility

← Low Cost →

Optimize costs and accelerate innovation with serverless computing

What is Serverless?

No infrastructure provisioning, no management

Automatic scaling

Pay for value

Highly available and secure

Serverless spans many different categories of services

Compute

AWS Lambda

AWS Fargate

Data stores

Amazon S3

Amazon Aurora Serverless

Amazon DynamoDB

Integration

Amazon API Gateway

Amazon SQS

Amazon SNS

AWS Step Functions

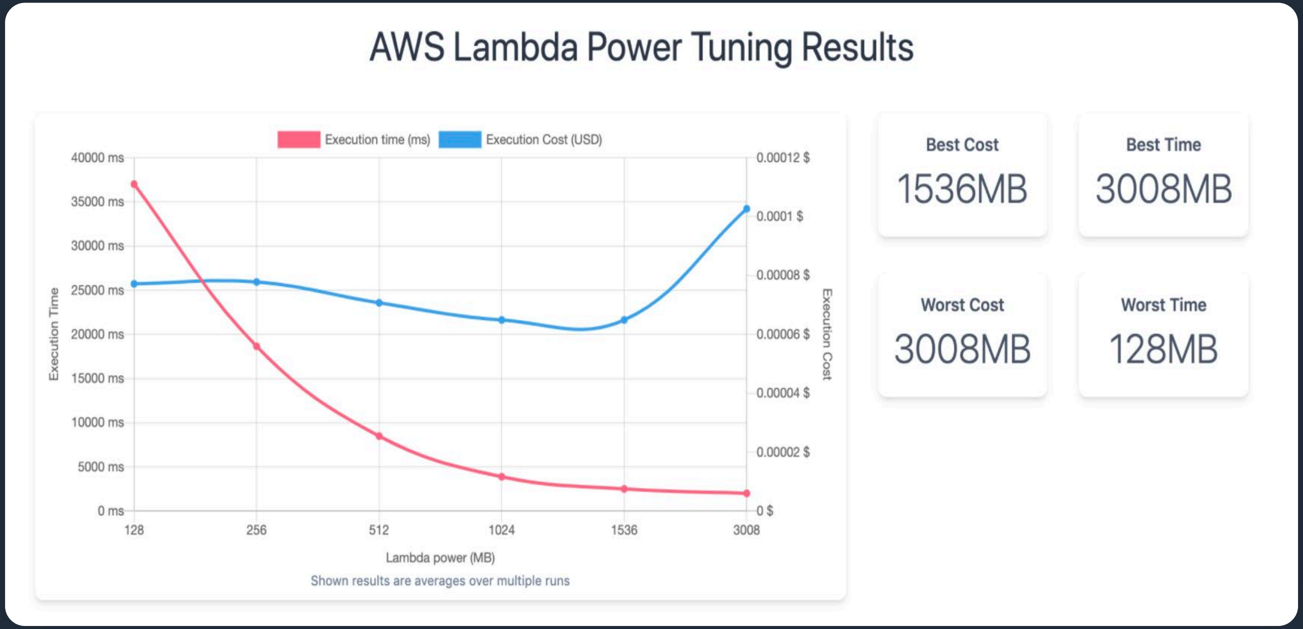
AWS AppSync

Lambda Power Tuner

Easily deploy from the Serverless Application Repository (SAR)

All you need to run is a sample payload

Reports on which function size is the best cost as well as best performance



AWS Lambda Power Tuning Results

Best Cost
1536MB

Best Time
3008MB

Worst Cost
3008MB

Worst Time
128MB

https://github.com/alexcasalboni/aws-lambda-power-tuning

NEW

AWS Lambda SnapStart

Up to 10x faster startup performance

‘Cold’ starts are fast with SnapStart!

	P50	P99.9
Without SnapStart	8 ms	5,114 ms
With SnapStart	8 ms	536 ms

> **“SnapStart requires little to no code changes and delivers significant cold start improvements in numerous Lambda use cases… this unlocks the value of serverless on workloads that were previously not suitable for Lambda ”**

Marty Andolino

Capital One Retail Bank Technology, VP of Engineering/Divisional Chief Architect

Choose the AWS compute instance type that matches your application needs

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

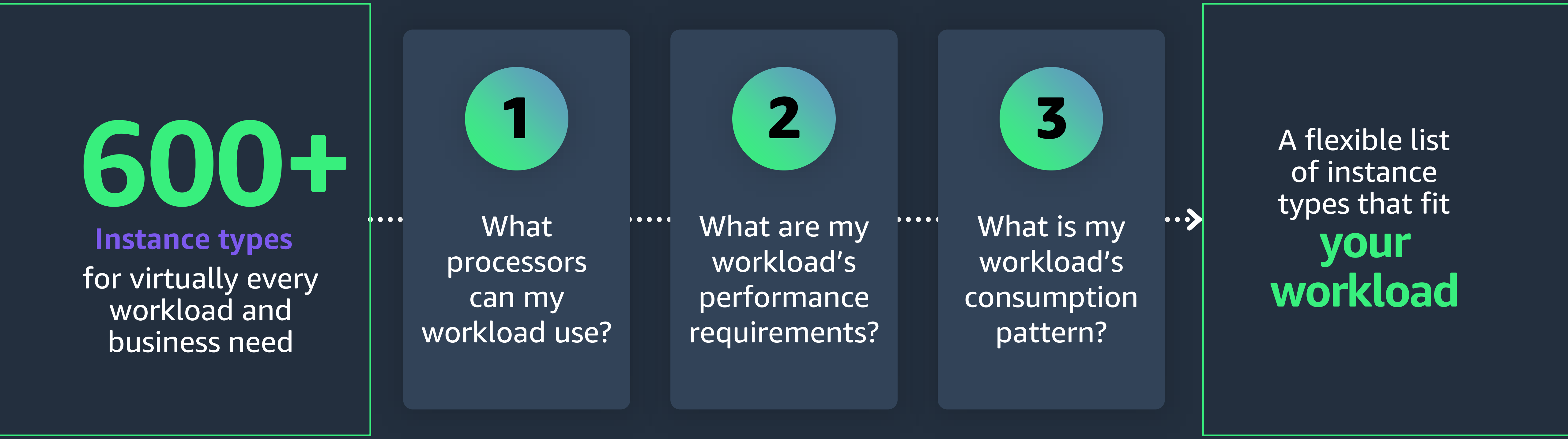
=

600+ instance configurations

for virtually every business need

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

Select the right compute options for your workload



600+

Instance types

for virtually every workload and business need

1

What processors can my workload use?

2

What are my workload's performance requirements?

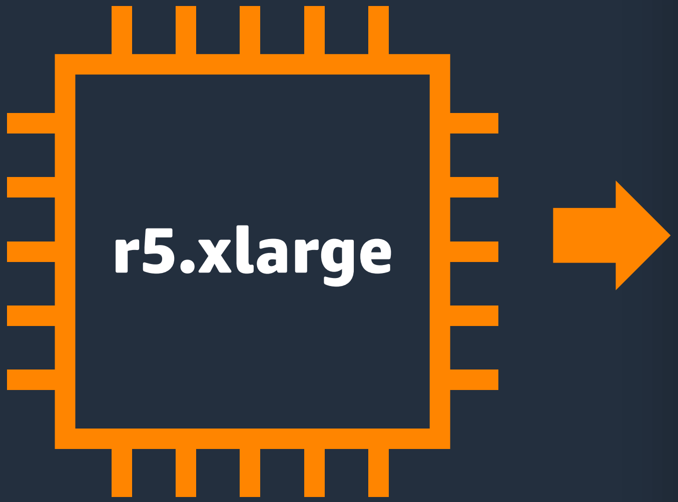
3

What is my workload's consumption pattern?

A flexible list of instance types that fit **your workload**

Attribute-based instance selection

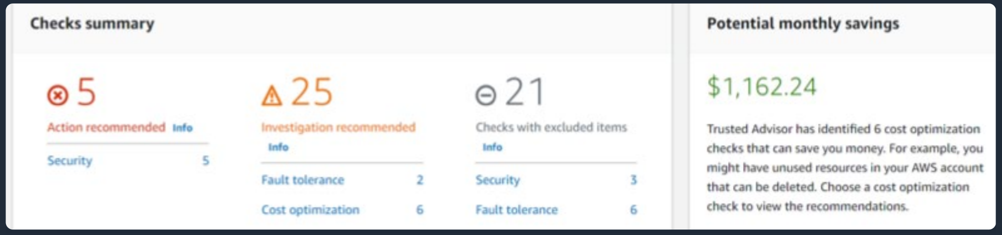
STOP PICKING INSTANCES! INSTEAD TELL US WHAT YOU ACTUALLY NEED…



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

Automated recommendations

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



Checks summary

⊗ 5 Action recommended

Security	5

△ 25 Investigation recommended

Fault tolerance	2
Cost optimization	6

⊖ 21 Checks with excluded items

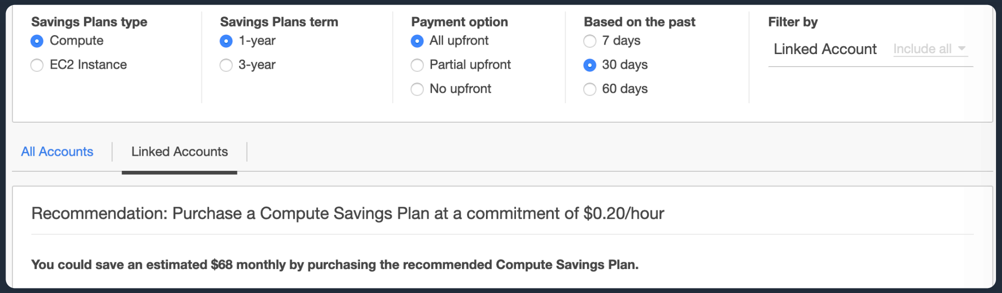
Security	3
Fault tolerance	6

Potential monthly savings

$1,162.24

Trusted Advisor has identified 6 cost optimization checks that can save you money. For example, you might have unused resources in your AWS account that can be deleted. Choose a cost optimization check to view the recommendations.

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



Savings Plans type: ● Compute ○ EC2 Instance

Savings Plans term: ● 1-year ○ 3-year

Payment option: ● All upfront ○ Partial upfront ○ No upfront

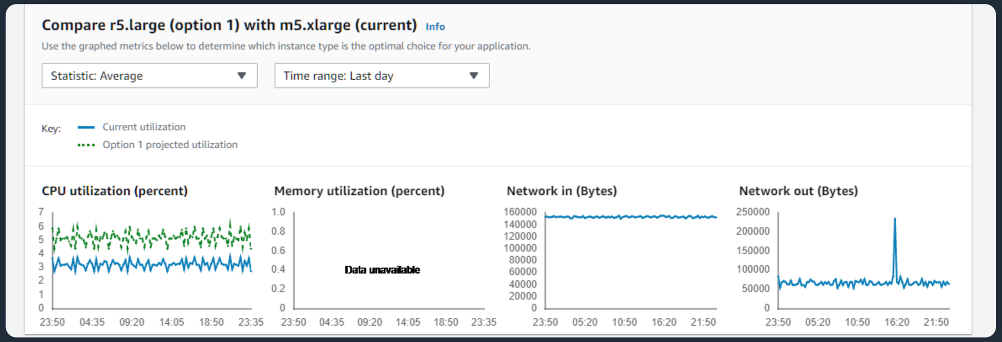
Based on the past: ○ 7 days ● 30 days ○ 60 days

Filter by: Linked Account Include all

Recommendation: Purchase a Compute Savings Plan at a commitment of $0.20/hour

You could save an estimated $68 monthly by purchasing the recommended Compute Savings Plan.

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



Compare r5.large (option 1) with m5.xlarge (current)

Use the graphed metrics below to determine which instance type is the optimal choice for your application.

Statistic: Average

Time range: Last day

Select the compute purchase models that best fits your budget

AWS has multiple purchase options to optimize Compute costs

On-Demand

Pay-for-what you use with no long-term commitments

Stateful spiky workloads

Savings Plans

Up to 72% savings for 1 or 3 year hourly usage commitments

Committed & steady-state usage

Spot

Spare capacity at up to 90% off On-Demand prices

Fault-tolerant, flexible, stateless workloads

The best practice is to combine all three purchase options

Types of Savings Plans

Compute Savings Plans

Offers the **greatest flexibility**

Discounts of up to **66%**

Automatically applied to any usage across:

- **Region**
- **Instance family**
- **Instance sizes**
- **Tenancy**
- **Operating system**
- **Compute service options**

EC2 Instance Savings Plans

Provides the **deepest savings**

Discounts of up to **72%**

Automatically applied to selected EC2 Instances & Regions across:

- **Instance sizes**
- **Operating system**
- **Tenancy**

SageMaker Savings Plans

Up to 64% off eligible SageMaker machine learning instance usage

Flexible across:

- **Sagemaker ML usages**
- **Instance family**
- **Size**
- **Region**

Why Savings Plans?

Cost savings

Benefit from significant cost savings of up to 72% compared to on-demand prices.

Easy to use

Easily reduce your bill as Savings Plans automatically and simultaneously apply to eligible AWS usage.

Flexible

Innovate faster by using the newest instance families, generations, and Regions while continuing to save.

Spot Instances for interruptible workloads

Same infrastructure

Spare Amazon EC2 capacity from the same infrastructure as on-demand

Capacity

AWS can reclaim with a 2-minute notice; interruptions happen when Amazon EC2 needs the spare capacity back

Workloads Spot is ideal for

- Fault-tolerant
- Flexible
- Loosely coupled
- Stateless

Migrate to AWS Graviton for the best price performance for a broad set of applications

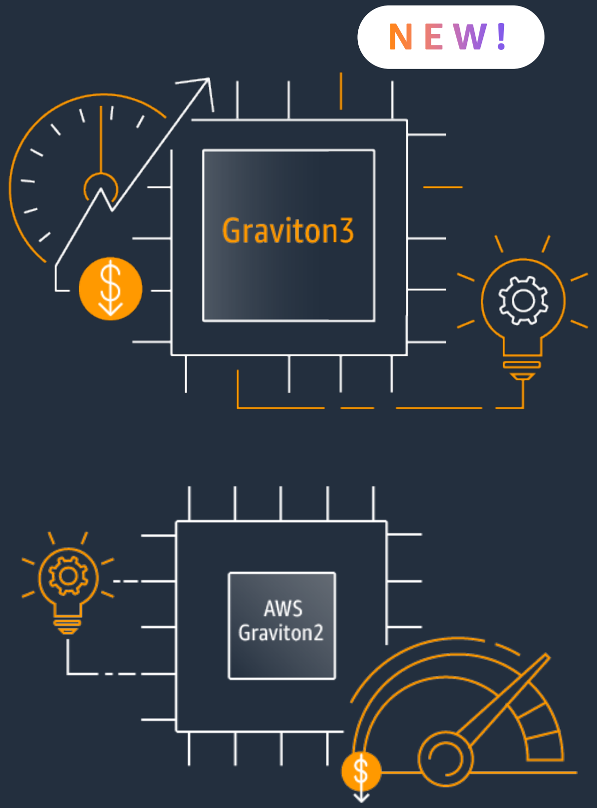
The latest choice of **processors**

intel AMD aws

and **accelerators**

NVIDIA habana XILINX

AWS Graviton processors



Custom AWS silicon with 64-bit Arm processor cores

Targeted optimizations for cloud-native workloads

Rapidly innovate, build, and iterate on behalf of customers

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 Graviton: broad workload applicability

Web and gaming servers

Open-source databases

High performance computing

In-memory caches

Media encoding

Electronic design automation

Analytics

Microservices

AWS managed services supporting Graviton

EXTENDING THE GRAVITON PRICE PERFORMANCE TO MANAGED SERVICES

Databases

- Amazon DocumentDB
- Amazon Aurora
- Amazon RDS
- Amazon Elasticache
- Amazon MemoryDB
- NEW! Amazon Neptune

AI/ML

- NEW! Amazon SageMaker

Analytics

- Amazon OpenSearch
- Amazon EMR

Compute

- AWS Lambda
- NEW! AWS Fargate
- NEW! AWS Elastic Beanstalk
- NEW! Amazon FSx for Lustre, Open ZFS

https://github.com/aws/aws-graviton-getting-started/blob/main/managed_services.md

What customers are saying about Amazon EC2 C7g instances

honeycomb.io

“We were able to run 30% fewer instances of C7g than C6g serving the same workload, and with 30% reduced latency.”

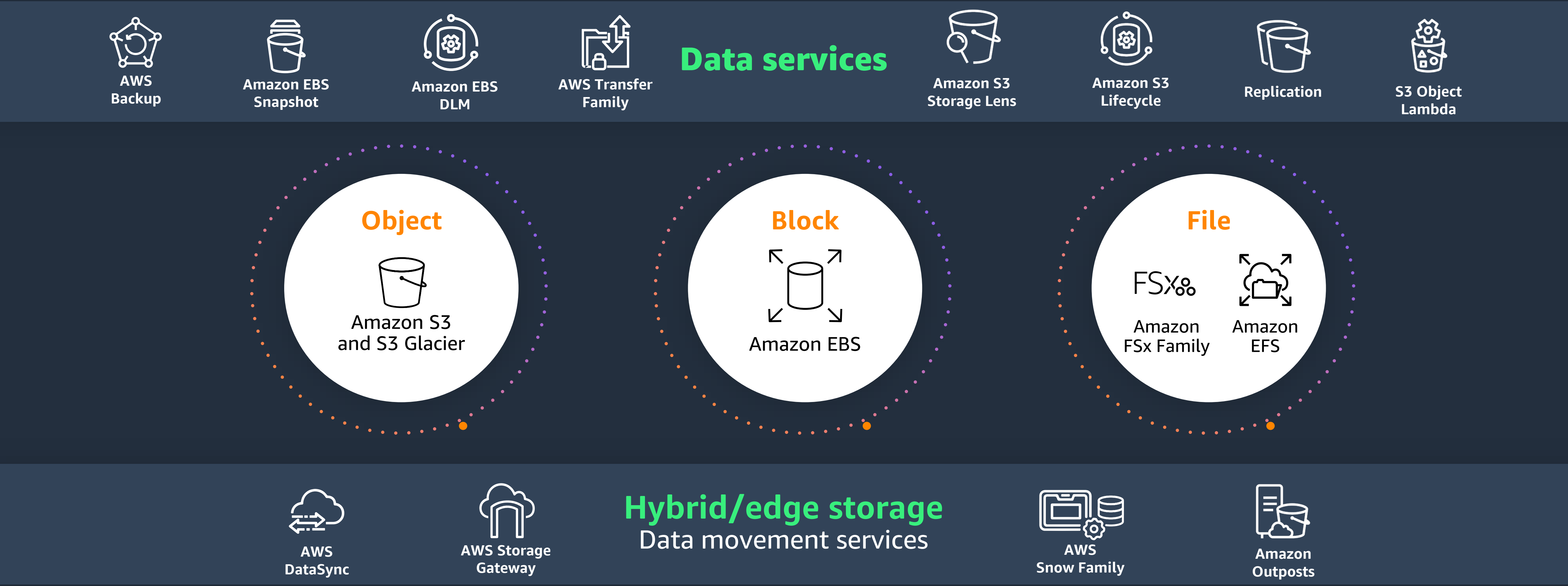
EPIC GAMES

“They are suitable for even the most demanding latency sensitive workloads while providing significant price performance benefits.”

“We have now found Graviton3 C7g instances to be 40% faster than the Graviton2 C6gn instances for those same simulations.“

Optimize your workload price and performance with AWS Storage

AWS Storage Portfolio



Data services

AWS Backup

Amazon EBS Snapshot

Amazon EBS DLM

AWS Transfer Family

Amazon S3 Storage Lens

Amazon S3 Lifecycle

Replication

S3 Object Lambda

Object

Amazon S3 and S3 Glacier

Block

Amazon EBS

File

Amazon FSx Family

Amazon EFS

Hybrid/edge storage

Data movement services

AWS DataSync

AWS Storage Gateway

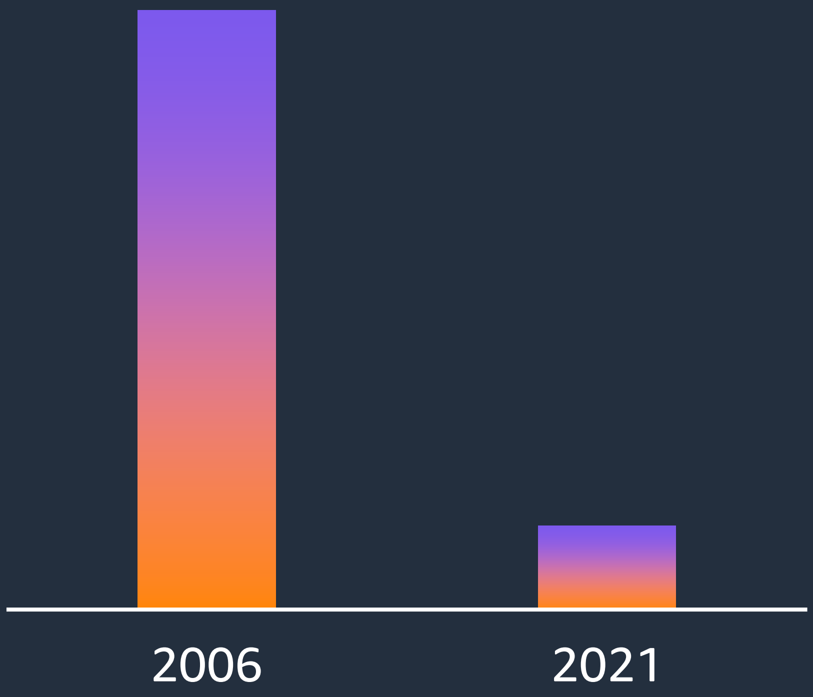
AWS Snow Family

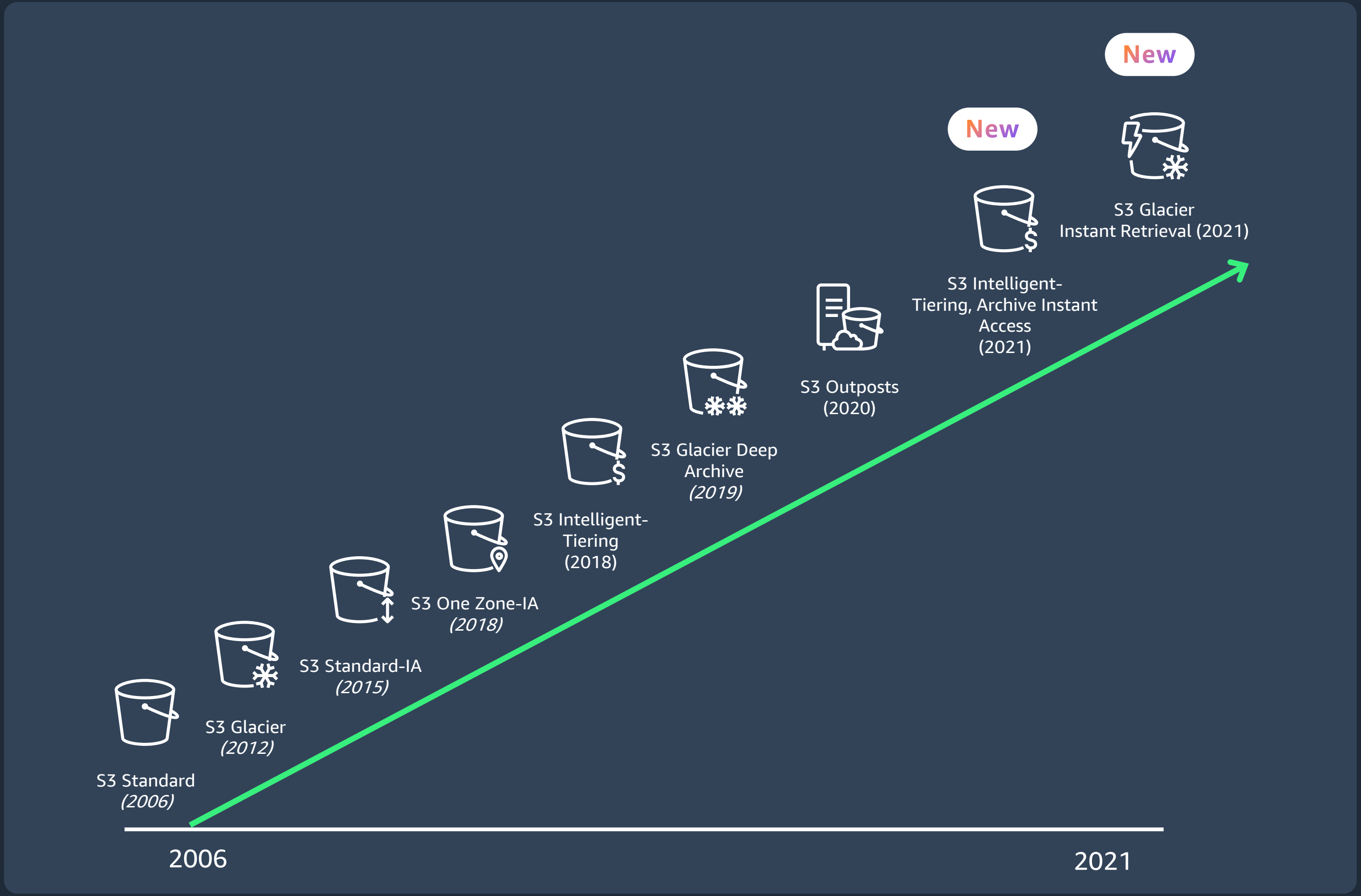
Amazon Outposts

Amazon S3 storage classes

OPTIMIZE YOUR STORAGE COST BY USING ALL AMAZON S3 STORAGE CLASSES

Decreasing storage prices




S3 Standard (2006)
S3 Glacier (2012)
S3 Standard-IA (2015)
S3 One Zone-IA (2018)
S3 Intelligent-Tiering (2018)
S3 Glacier Deep Archive (2019)
S3 Outposts (2020)
New
S3 Intelligent-Tiering, Archive Instant Access (2021)
New
S3 Glacier Instant Retrieval (2021)
2006
2021

Your choice of 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						**AWS AZ**	**AWS Outposts**
Data with changing access patterns	**Frequently accessed data**	**Infrequently accessed data**	**Rarely accessed data**	**Archive data**	**Long-term archive data**	**Re-creatable, less accessed data**	**On-premises 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	• Milliseconds access • Retrieval charge per-GB	• Milliseconds access • Retrieval charge per-GB

Since the launch of S3 Intelligent-Tiering, customers' storage cost savings now exceed

$750,000,000

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 **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 Elastic Block Store (EBS) is...

Scalable

Reduce deployment times from **months to minutes**

Address rapid data growth, purchase what you need now and grow capacity on-demand

Virtually unlimited capacity available for scaling

Simple to manage

Simplified provisioning of resources from compute to storage

Seamlessly migrate data to optimum storage tiers to address changing workload requirements

Built in security and recovery features that can be configured with 'a few clicks'

Optimized

Eliminate investment in infrastructure management and complex datacenter operations

Eliminate the "infrastructure lifecycle tax". No more business disruptions due to forklift upgrades and complex migration projects

...now available for your mission critical workloads.

...with SAN in the Cloud

Amazon EBS: Latest-Generation Volume types

	gp3 **General Purpose SSD**	**io2** **Provisioned IOPS SSD**	**io2** **Block Express**	**st1** **Throughput Optimized HDD**	**sc1** **Cold HDD**
Use-cases	Relational and non-relational databases, enterprise applications, containerized workloads, big data, file system, media workflows	Large database workloads, mission-critical business applications requiring sustained high performance	Critical applications and databases requiring sustained IOPS	Big data workloads, data warehouses, log processing, streaming workloads	Large volumes of infrequently accessed data, cost-sensitive workloads
Volume Size	1 GiB–16 TiB	4 GiB–16 TiB	4 GiB–64 TiB	125 GiB–16 TiB	125 GiB–16 TiB
Max IOPS per volume	16,000	64,000	256,000	500	250
Max Throughput per volume	1,000 MiB/s	1,000 MiB/s	4000 MiB/s	500 MiB/s	250 MiB/s
Pricing	**$0.08 per GB-month** of provisioned storage **3,000 IOPS free** and $0.005/provisioned IOPS-month over 3,000 **125 MB/s free** and $0.04/provisioned MB/s-month over 125	**$0.125 per GB-month** of provisioned storage **$0.065 per provisioned IOPS-month** up to 32,000 **$0.046 per provisioned IOPS-month** from 32,001 to 64,000	**$0.125 per GB-month** of provisioned storage **$0.065 per provisioned IOPS-month** up to 32,000 **$0.046 per provisioned IOPS-month** from 32,001 to 64,000 **$0.032 per provisioned IOPS-month** for greater than 64,000	**$0.045 per GB-month** of provisioned storage	**$0.015 per GB-month** of provisioned storage

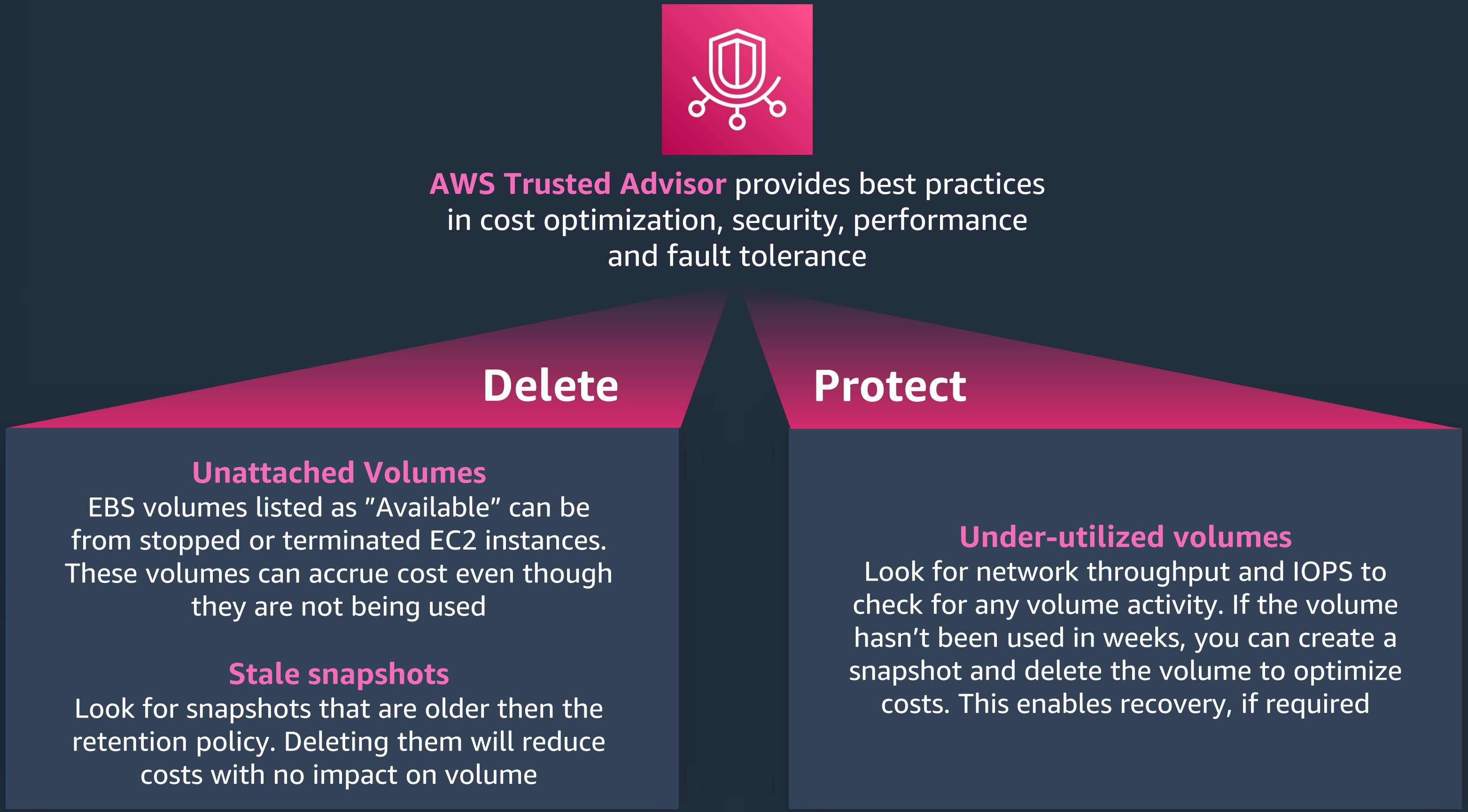
Amazon EBS pptimization

GAIN INSIGHTS INTO YOUR EBS DEPLOYMENT

Use **Amazon CloudWatch** metrics to gain insight into performance and utilization of EBS volumes

AWS Compute Optimizer provides optimization recommendations for EC2 instances and EBS volumes

Use **AWS Cost Explorer** to analyze EBS usage and cost to explore optimization options



AWS Trusted Advisor provides best practices in cost optimization, security, performance and fault tolerance

Delete

Unattached Volumes
EBS volumes listed as "Available" can be from stopped or terminated EC2 instances. These volumes can accrue cost even though they are not being used

Stale snapshots
Look for snapshots that are older then the retention policy. Deleting them will reduce costs with no impact on volume

Protect

Under-utilized volumes
Look for network throughput and IOPS to check for any volume activity. If the volume hasn't been used in weeks, you can create a snapshot and delete the volume to optimize costs. This enables recovery, if required

Optimize data placement

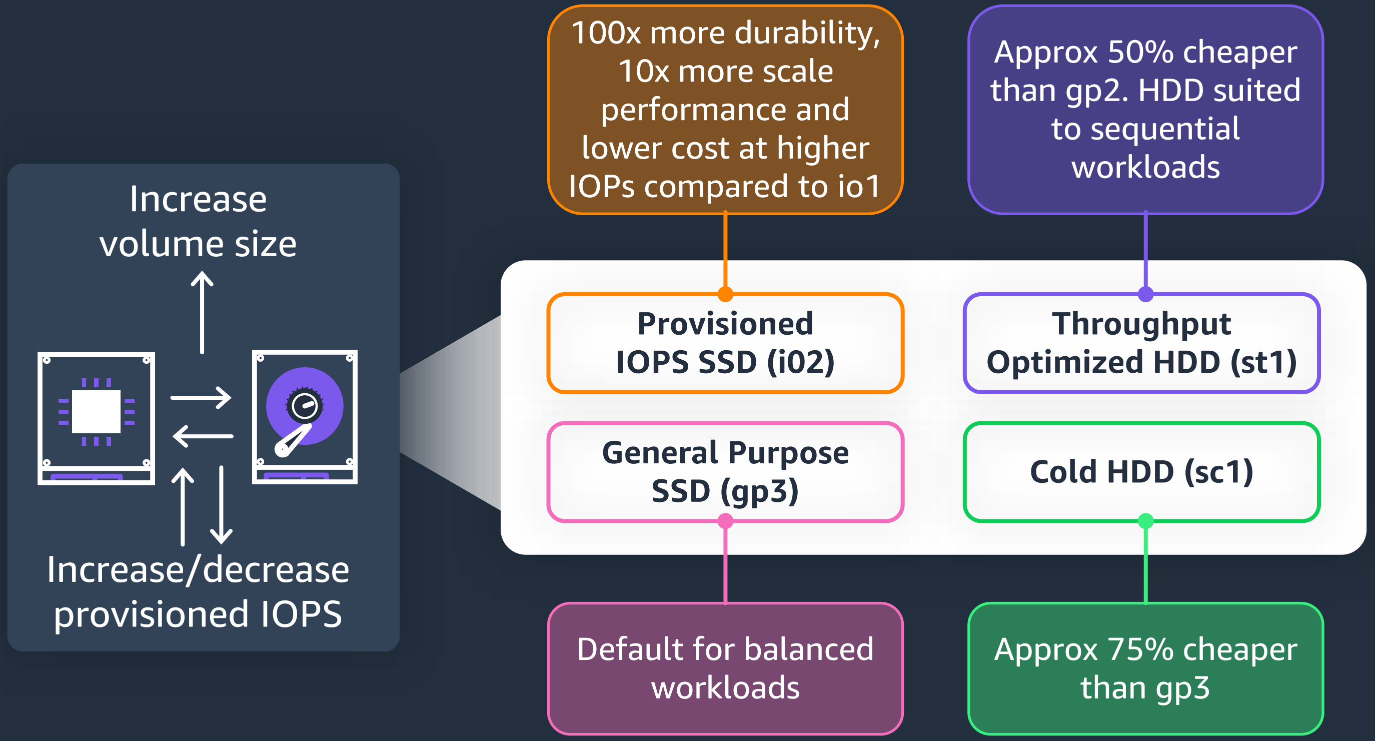
USE ELASTIC VOLUMES TO DYNAMICALLY CHANGE VOLUME FEATURES AND DATA PLACEMENT TO SUPPORT GROWTH AND COSTS

Provision minimum required capacity

Provision EBS for minimum required size and expand as needed. Maintenance is easy with zero downtime

Optimize data placement based on workload requirements

Migrate data non-disruptively across EBS volume types to align with changing application performance requirements



Increase volume size

Increase/decrease provisioned IOPS

100x more durability, 10x more scale performance and lower cost at higher IOPs compared to io1

Provisioned IOPS SSD (i02)

Approx 50% cheaper than gp2. HDD suited to sequential workloads

Throughput Optimized HDD (st1)

General Purpose SSD (gp3)

Default for balanced workloads

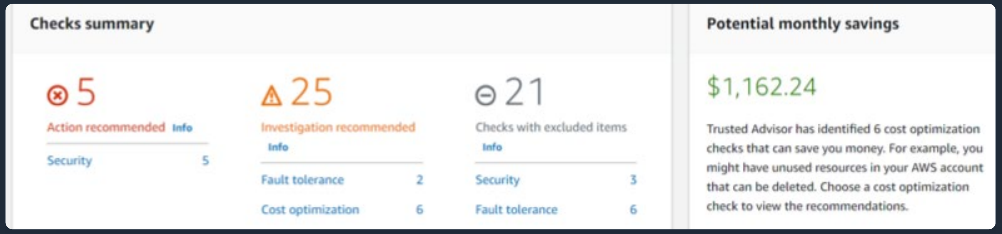
Cold HDD (sc1)

Approx 75% cheaper than gp3

Optimize your resource capacity to fit demand

Right size your resource size

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



Checks summary

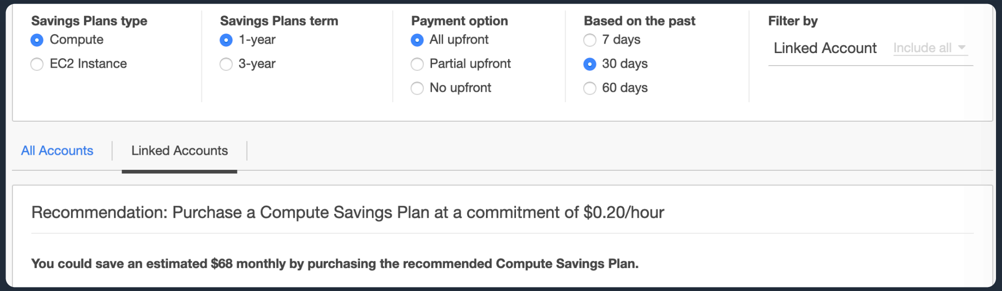
⊗ 5	⚠ 25	⊖ 21
Action recommended Info	Investigation recommended Info	Checks with excluded items Info
Security 5	Fault tolerance 2	Security 3
	Cost optimization 6	Fault tolerance 6

Potential monthly savings

$1,162.24

Trusted Advisor has identified 6 cost optimization checks that can save you money. For example, you might have unused resources in your AWS account that can be deleted. Choose a cost optimization check to view the recommendations.

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



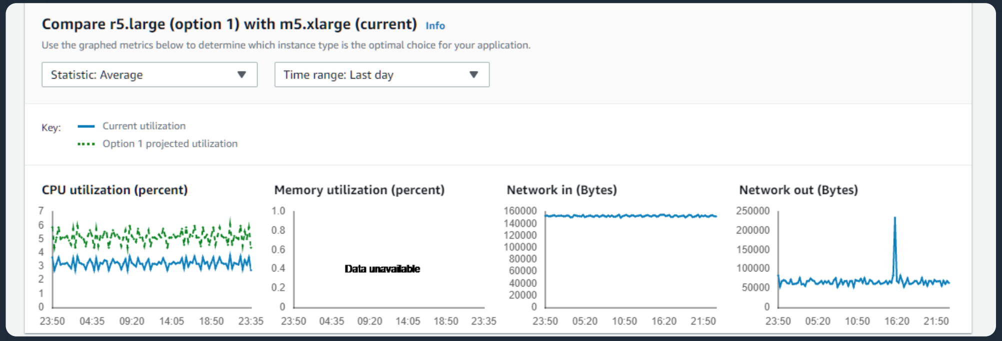
Savings Plans type	Savings Plans term	Payment option	Based on the past	Filter by
● Compute	● 1-year	● All upfront	○ 7 days	Linked Account Include all ▾
○ EC2 Instance	○ 3-year	○ Partial upfront	● 30 days	
		○ No upfront	○ 60 days	

All Accounts | Linked Accounts

Recommendation: Purchase a Compute Savings Plan at a commitment of $0.20/hour

You could save an estimated $68 monthly by purchasing the recommended Compute Savings Plan.

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



Compare r5.large (option 1) with m5.xlarge (current) Info

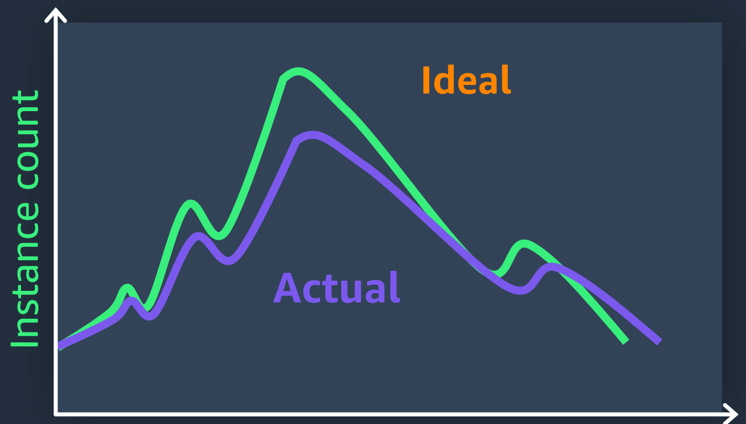
Use the graphed metrics below to determine which instance type is the optimal choice for your application.

Statistic: Average ▼

Time range: Last day ▼

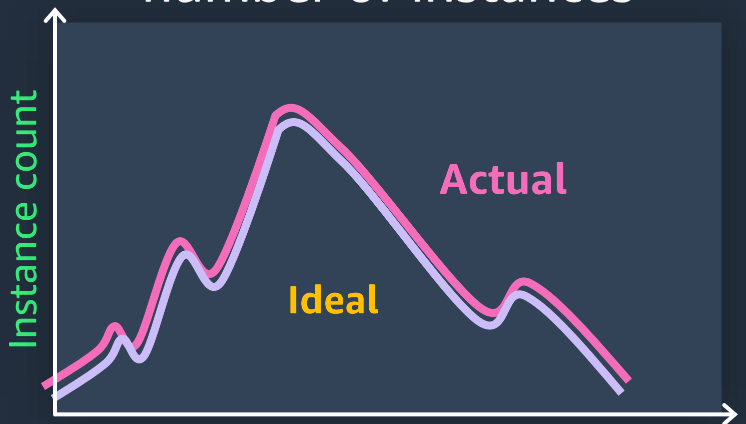
Right size your resource count

Too few instances

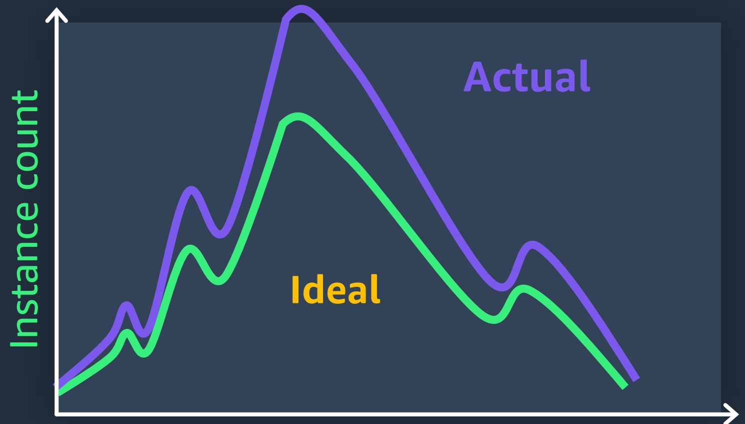


Under-provisioned;
poor experience

The **right**
number of instances



Too many instances



Over-provisioned;
wasteful

Scaling policies

Dynamic scaling

Simple/step scaling

Monitors metrics and adds/removes instances as per steps defined by the customers

Manually calculate capacity
Reactive in nature

Target tracking

Thermostat-like control mechanism that automatically adds or removes instances to maintain metrics at a customer defined target

Automated
Reactive in nature

Scheduled scaling

Launch/terminate instances as defined by customer on a schedule

Manually calculate capacity
Proactive in nature

Predictive scaling

Proactively launch capacity based on historic trends

Automated
Proactive in nature

Thank you!

Boyd McGeachie