Cost Optimization for Startups

Actionable steps for immediate results

Natavit Rojcharoenpreeda

Startup Solutions Architect
Amazon Web Services

Agenda

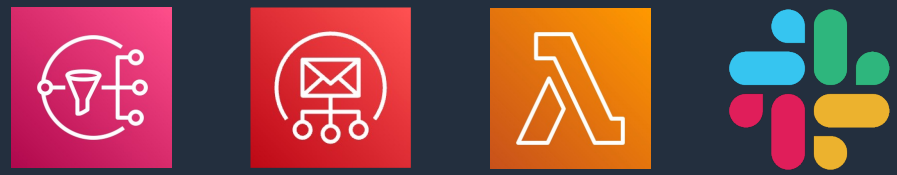
- Getting started
- Purchasing options
- Cost optimization: EC2, EBS, S3, Data Transfer
- Wrap up

Getting started with Cost Optimization

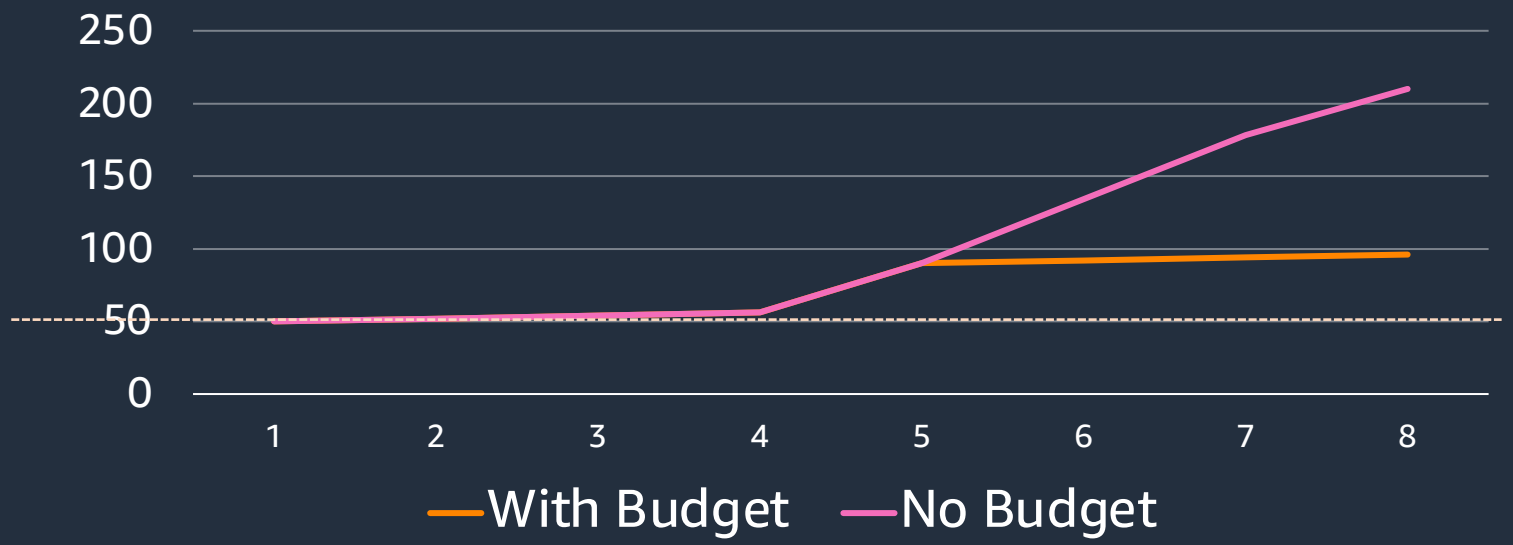
Start with the highest one

AWS Cost Explorer

AWS Budgets



Budget Alert Level

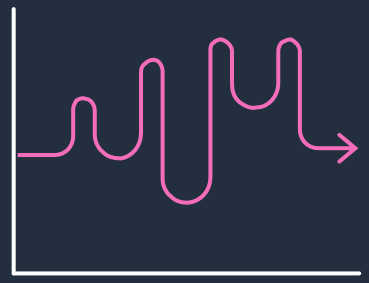


Purchasing Options

Purchasing Options

On-demand

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

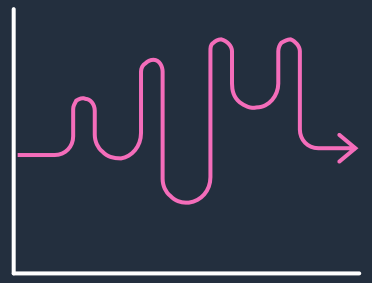


Spiky or fluctuating workloads

Purchasing Options

On-demand

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



Spiky or fluctuating workloads

Reserved

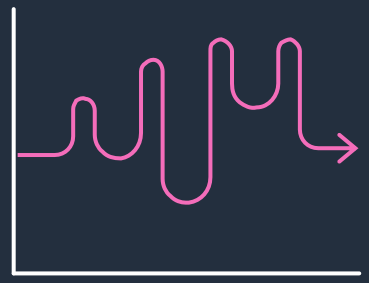
Significant discount compared to on-demand instance pricing

Steady state applications or predictable usage, databases

Purchasing Options

On-demand

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



Spiky or fluctuating workloads

Reserved

Significant discount compared to on-demand instance pricing

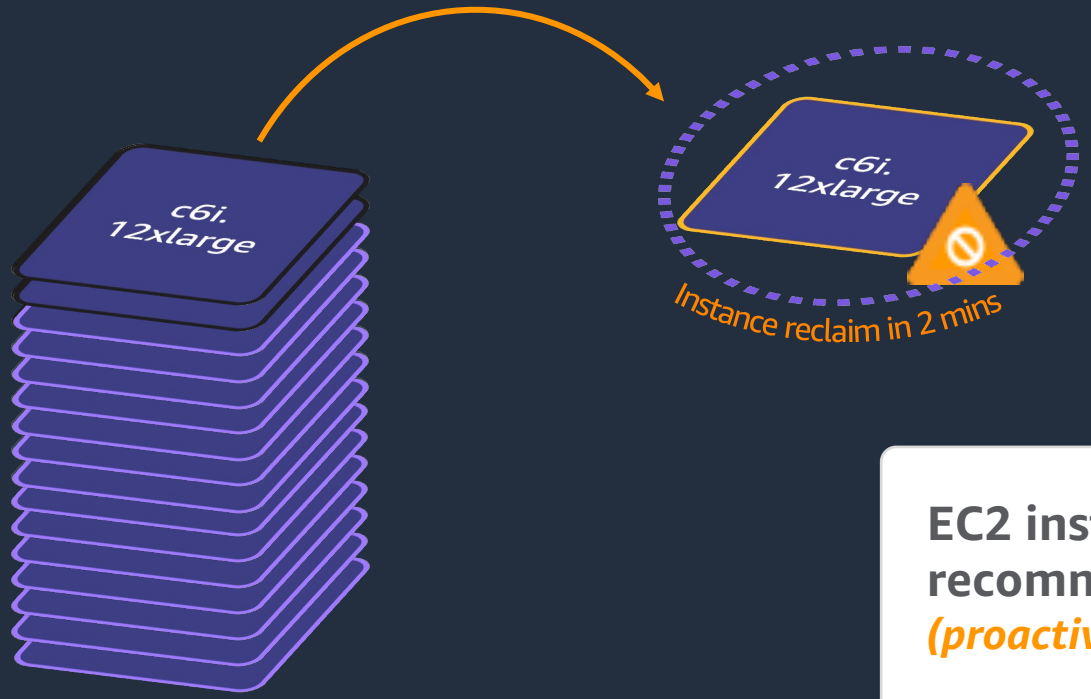
Steady state applications or predictable usage, databases

Spot

Spare Amazon EC2 capacity for **up to 90% off** the on-demand price

Fault tolerant, flexible, stateless workloads

Amazon EC2 – Spot Interruption



By the nature of Spot as spare-capacity, a Spot instance can be interrupted if the instance is needed by On-Demand.

AWS provides two types of notifications to enable you to handle the response in an automated way:

EC2 instance rebalance recommendation ***(proactive)***

- Spot instance is at elevated risk of interruption
- Built in support for AWS integrations such as EC2 Auto Scaling and EKS Managed Node Groups +

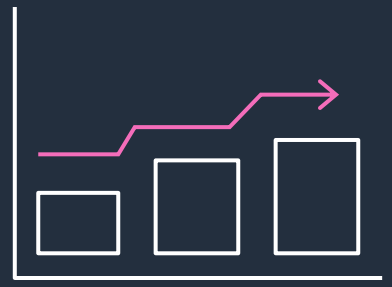
Spot instance termination notice ***(reactive)***

- Interruption of instance will happen in 2 minutes, adjust your workload appropriately
- Built in support for AWS integrations such as EC2 Auto Scaling and EKS Managed Node Groups +

Reserved Options

Reserved Instances (RIs)

Make a 1- or 3-year commitment and receive a **significant discount** on On-Demand prices



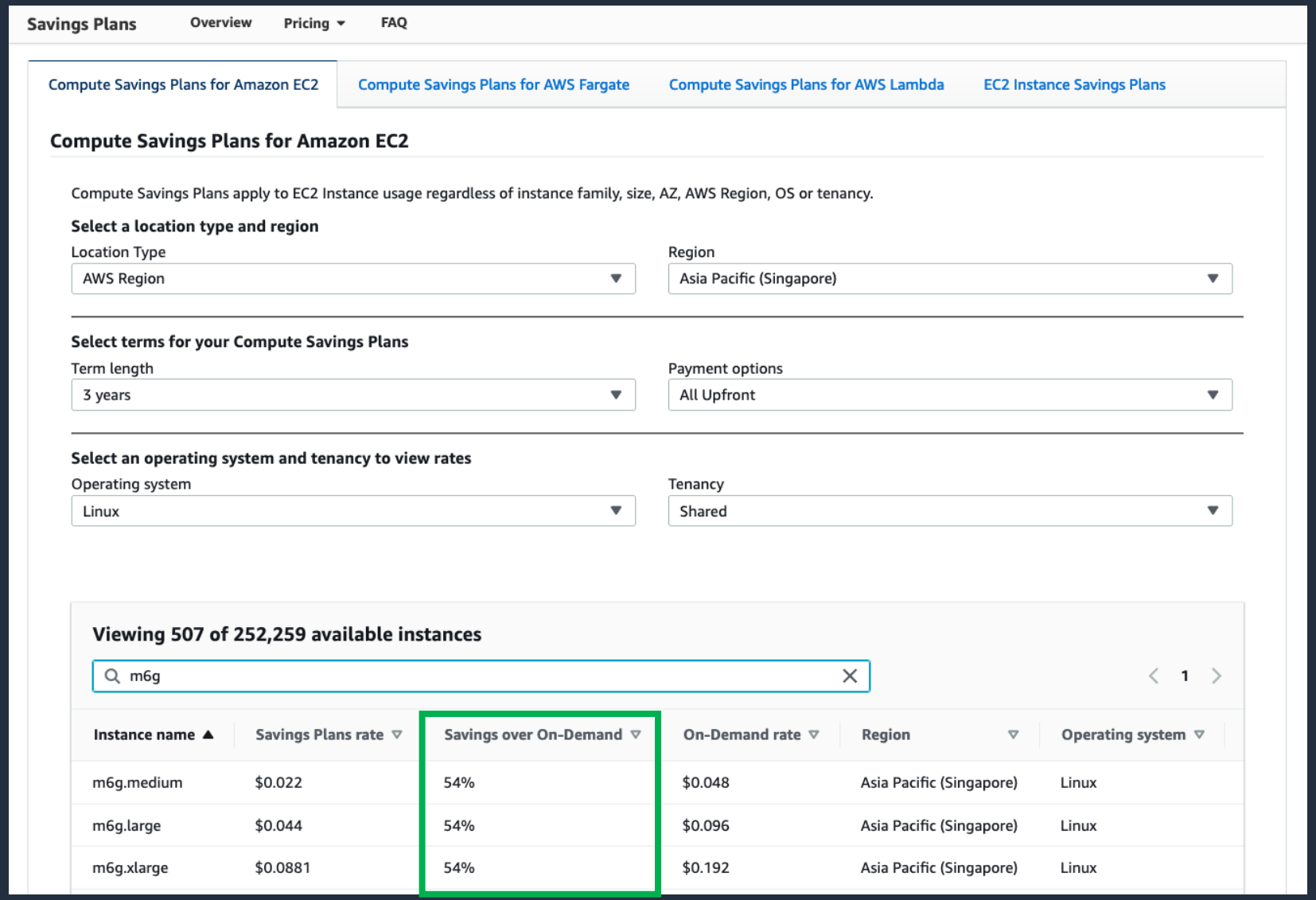
Committed and steady-state usage

Savings Plans

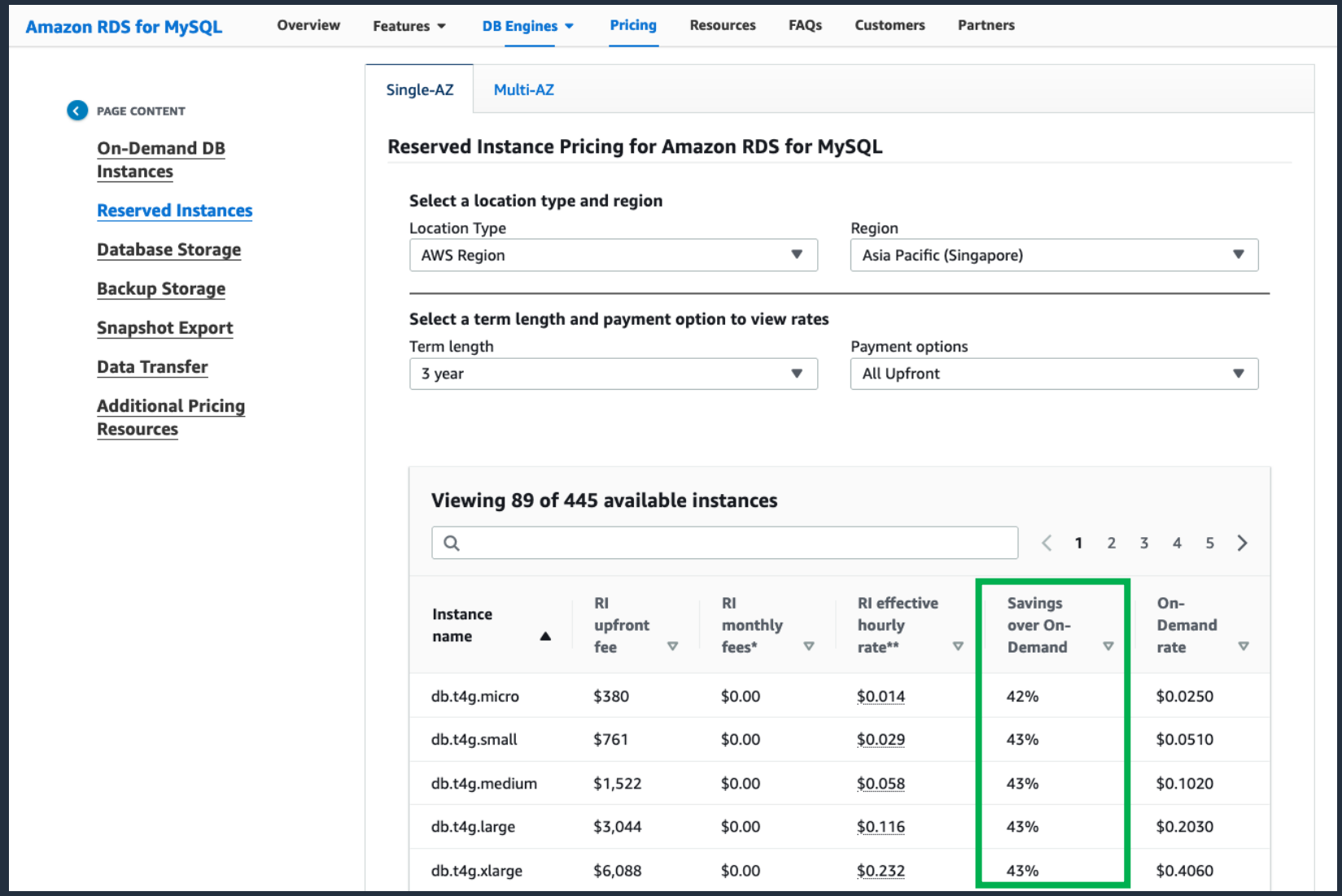
Same great discounts as Amazon EC2 RIs with **more flexibility**

Flexible access to compute

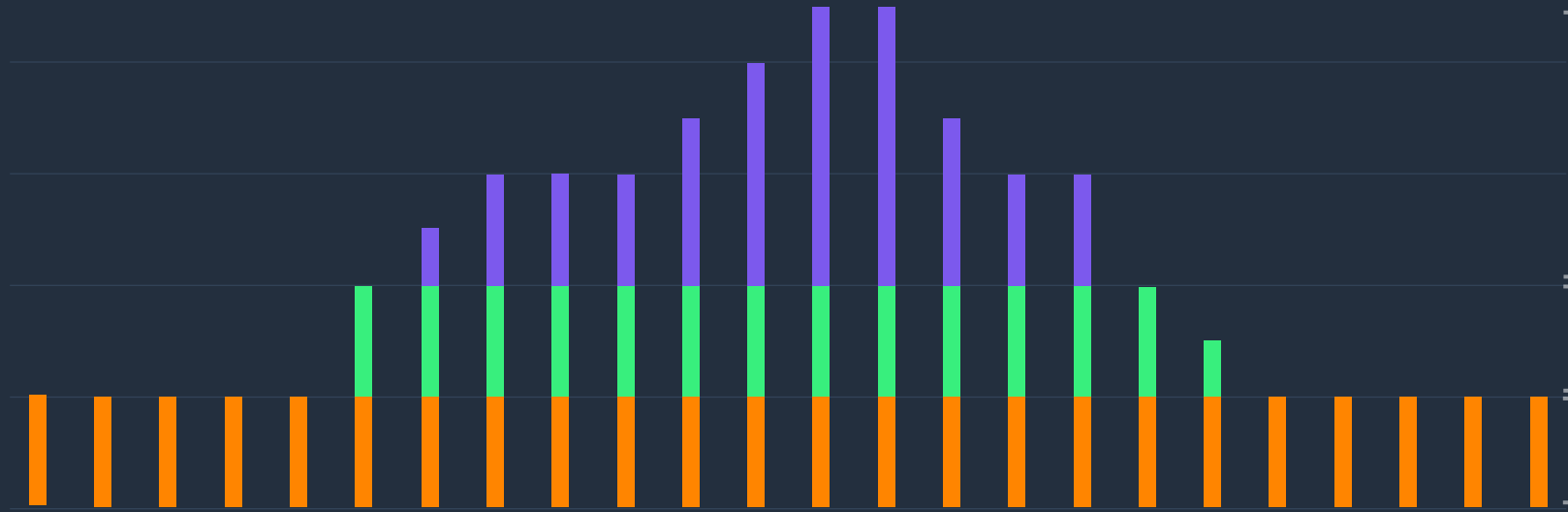
Savings Plans Pricing Page



RDS MySQL Pricing Page – Reserved Instances



Combining to achieve Optimization



Scale using **Spot** for fault-tolerant, flexible, stateless workloads

Use **On-Demand** for new or stateful spiky workloads

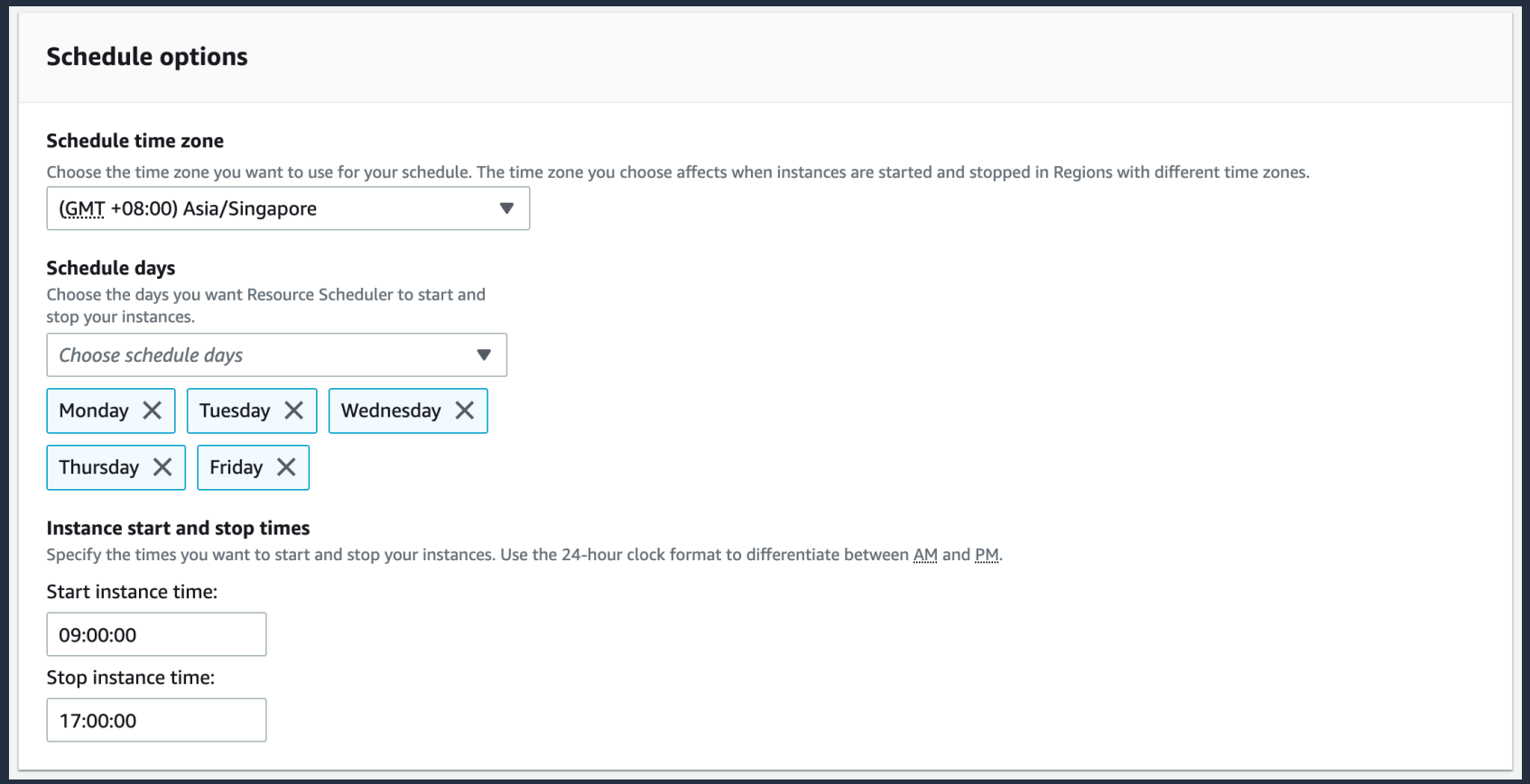
Use **RIs or a Savings Plan** for known, steady-state workloads

Amazon EC2 Cost Optimization

with potentially low effort

AWS Systems Manager – Resource Scheduler

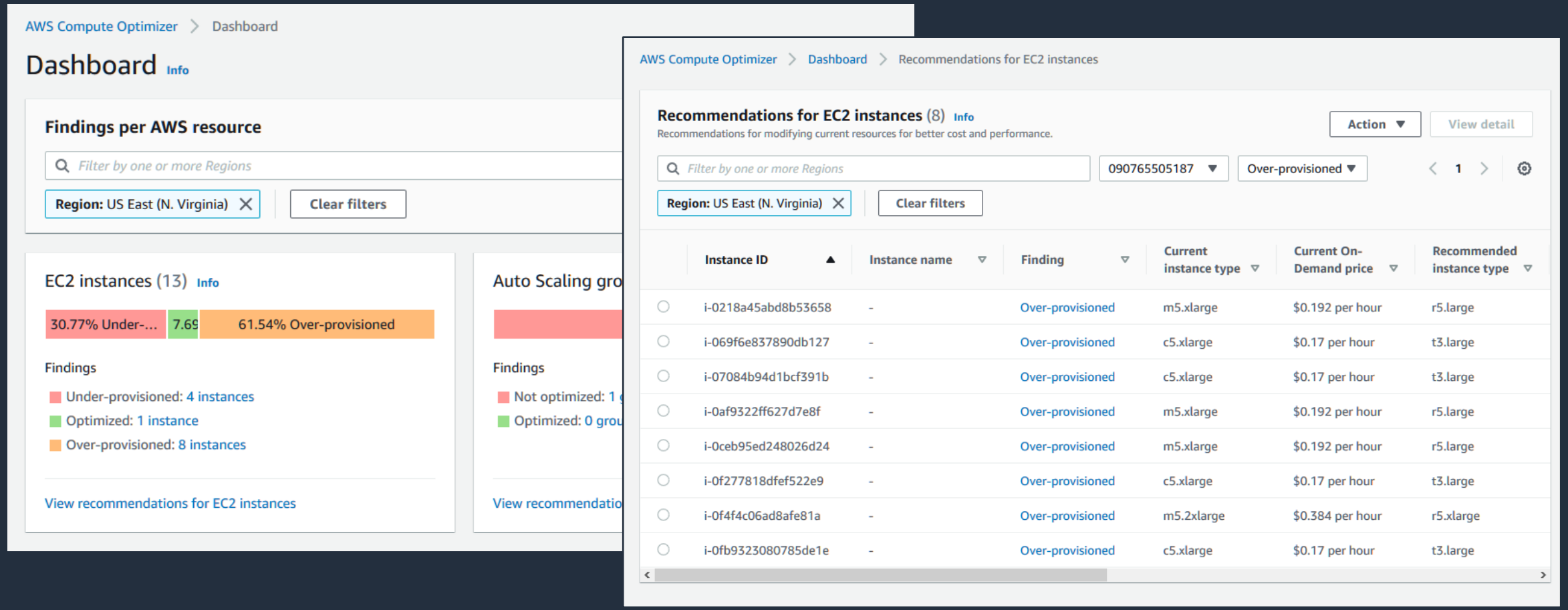
RUN EC2 INSTANCES ONLY WHEN NEEDED



EC2 Right-sizing – CloudWatch Metrics

EC2 Right-sizing - AWS Compute Optimizer

OPTIMIZE INSTANCES WITH RECOMMENDATIONS



EC2 – know your instance classes

EC2 instance classes

1. General Purpose: T (Burstable), M
2. Compute Optimized: C
3. Memory Optimized: R, X
4. Accelerated Computing (GPU): P, G,…

EC2 – know your processors

EC2 instance processors

1. Intel (x86): T2, T3, M5, C5, R5, M6i, C6i, R6i, …
2. AMD (x86): T3a, M6a, C6a, R6a, …
3. Graviton (ARM): T4g, M6g, C6g, R6g, …

EC2 – know your processors

EC2 instance processors

1. Intel (x86): T2, T3, M5, C5, R5, M6i, C6i, R6i, …
2. AMD (x86): T3a, M6a, C6a, R6a, …
3. Graviton (ARM): T4g, M6g, C6g, R6g, …

AMD: - 10% lower cost vs comparable x86 instances

EC2 – know your processors

EC2 instance processors

1. Intel (x86): T2, T3, M5, C5, R5, M6i, C6i, R6i, …
2. AMD (x86): T3a, M6a, C6a, R6a, …
3. Graviton (ARM): T4g, M6g, C6g, R6g, …

AMD: - 10% lower cost vs comparable x86 instances

Graviton: - Highest performance in their family

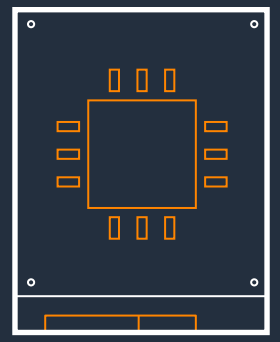
- 20% lower cost vs comparable x86 instances

- Up to 40% better price-performance

Amazon EBS Storage Optimization

with potentially low effort

EBS Block volume types

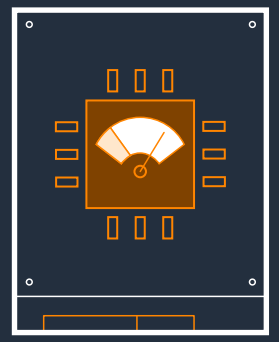


General-purpose SSD

NoSQL databases

Transactional workloads, low-latency applications

Cassandra, MongoDB, CouchDB

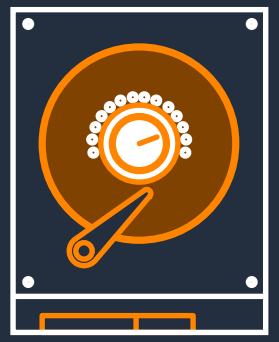


Provisioned IOPS SSD

Relational databases

I/O-intensive database applications

MySQL, SQL Server, PostgreSQL, SAP, Oracle



Throughput-optimized HDD

Big data, analytics

Large datasets and large I/O sizes

Kafka, Splunk, Hadoop, data warehousing

Cold HDD

File, media, reference

Less frequently accessed workloads with large, cold datasets

Transcoding, encoding, rendering, archive

gp3: New General Purpose SSD volume

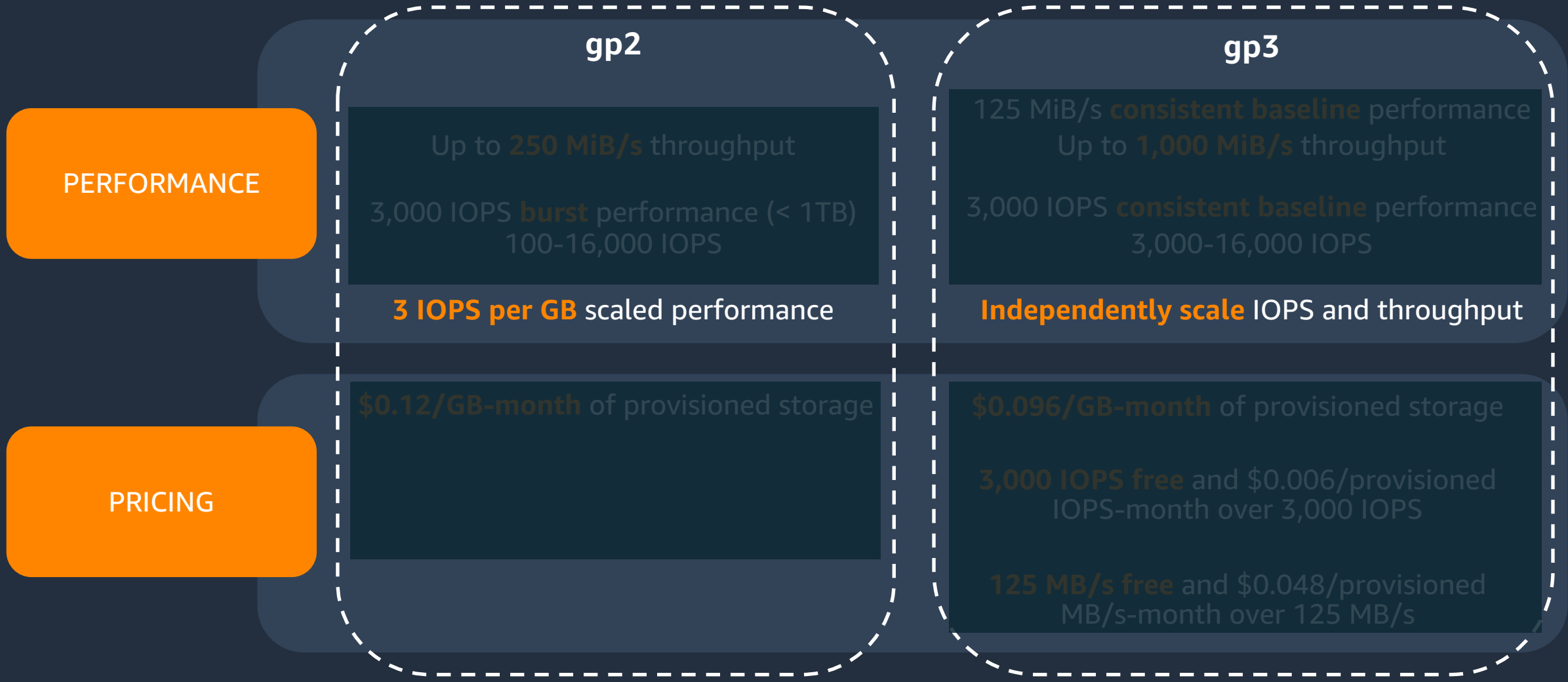
INDEPENDENTLY PROVISION IOPS AND THROUGHPUT AT UP TO 20% LOWER PRICE PER GB

	gp2	gp3
PERFORMANCE	Up to **250 MiB/s** throughput 3,000 IOPS **burst** performance (< 1TB) 100-16,000 IOPS **3 IOPS per GB** scaled performance	125 MiB/s **consistent baseline** performance Up to **1,000 MiB/s** throughput 3,000 IOPS **consistent baseline** performance 3,000-16,000 IOPS **Independently scale** IOPS and throughput
PRICING	**$0.12/GB-month** of provisioned storage	**$0.096/GB-month** of provisioned storage **3,000 IOPS free** and $0.006/provisioned IOPS-month over 3,000 IOPS **125 MB/s free** and $0.048/provisioned MB/s-month over 125 MB/s

Pricing for Singapore (March 2023)

gp3: New General Purpose SSD volume

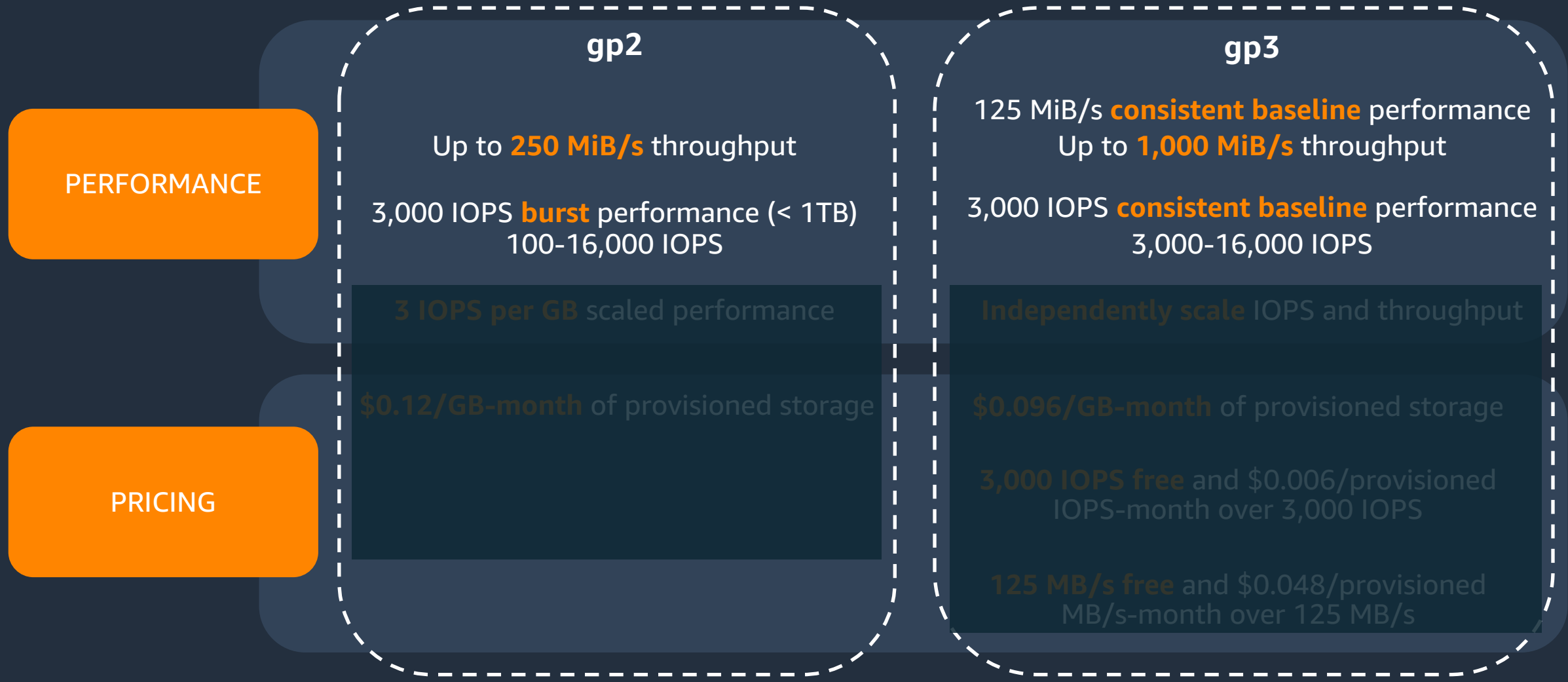
INDEPENDENTLY PROVISION IOPS AND THROUGHPUT AT UP TO 20% LOWER PRICE PER GB



Pricing for Singapore (March 2023)

gp3: New General Purpose SSD volume

INDEPENDENTLY PROVISION IOPS AND THROUGHPUT AT UP TO 20% LOWER PRICE PER GB



Pricing for Singapore (March 2023)

gp3: New General Purpose SSD volume

INDEPENDENTLY PROVISION IOPS AND THROUGHPUT AT UP TO 20% LOWER PRICE PER GB

	gp2	gp3
PERFORMANCE	Up to **250 MiB/s** throughput 3,000 IOPS **burst** performance (< 1TB) 100-16,000 IOPS **3 IOPS per GB** scaled performance	125 MiB/s **consistent baseline** performance Up to **1,000 MiB/s** throughput 3,000 IOPS **consistent baseline** performance 3,000-16,000 IOPS **Independently scale** IOPS and throughput
PRICING	**$0.12/GB-month** of provisioned storage	**$0.096/GB-month** of provisioned storage **3,000 IOPS free** and $0.006/provisioned IOPS-month over 3,000 IOPS **125 MB/s free** and $0.048/provisioned MB/s-month over 125 MB/s

Pricing for Singapore (March 2023)

Pricing Example: gp2 vs gp3 (ap-southeast-1, Singapore)

gp2				gp3				gp3 cost benefit	
Volume Size in GiB	Max IOPS	Throughput MiB/s	gp2 Cost (US$/Month)	IOPS		Throughput MiB/s		gp3Cost (US$/Month)	Cost reduction compared to GP2
				Baseline	Provisioned	Baseline	Provisioned		
30	3000	128	$3.60	3000	0	125	0	$2.88	20%
100	3000	128	$12.00	3000	0	125	0	$9.60	20%
500	3000	250	$60.00	3000	0	125	125	$54.00	10%
1000	3000	250	$120.00	3000	0	125	125	$102.00	15%
2000	6000	250	$240.00	3000	3000	125	125	$216.00	10%
6000	16000	250	$720.00	3000	13000	125	125	$660.00	8%

Pricing related to Singapore (March 2023)
Considering 1 volume for 730 hours/month

Pricing Example: gp2 vs gp3 (ap-southeast-1, Singapore)

gp2				gp3				gp3 cost benefit	
Volume Size in GiB	Max IOPS	Throughput MiB/s	gp2 Cost (US$/Month)	IOPS		Throughput MiB/s		gp3Cost (US$/Month)	Cost reduction compared to GP2
				Baseline	Provisioned	Baseline	Provisioned		
30	3000	128	$3.60	3000	0	125	0	$2.88	20%
100	3000	128	$12.00	3000	0	125	0	$9.60	20%
500	3000	250	$60.00	3000	0	125	125	$54.00	10%
1000	3000	250	$120.00	3000	0	125	125	$102.00	15%
2000	6000	250	$240.00	3000	3000	125	125	$216.00	10%
6000	16000	250	$720.00	3000	13000	125	125	$660.00	8%

Pricing related to Singapore (March 2023)
Considering 1 volume for 730 hours/month

Amazon S3 Storage Optimization

with potentially low effort

“Delete it if you don’t need it”

Storage class choice matters at scale

S3 Intelligent-Tiering	S3 Standard	S3 Standard-IA	S3 One Zone-IA	S3 Glacier Instant Retrieval	S3 Glacier Flexible Retrieval	S3 Glacier Deep Archive
						
Changing access patterns	Frequently accessed data	Infrequently accessed data	Re-creatable, less accessed data	Rarely accessed data	Archive data	Long-term archive data

Milliseconds access — Minutes to hours

What is Amazon S3 Intelligent-Tiering?

- **Delivers automatic storage cost savings**
- Moves objects between **three access tiers** for a small monthly monitoring and automation fee

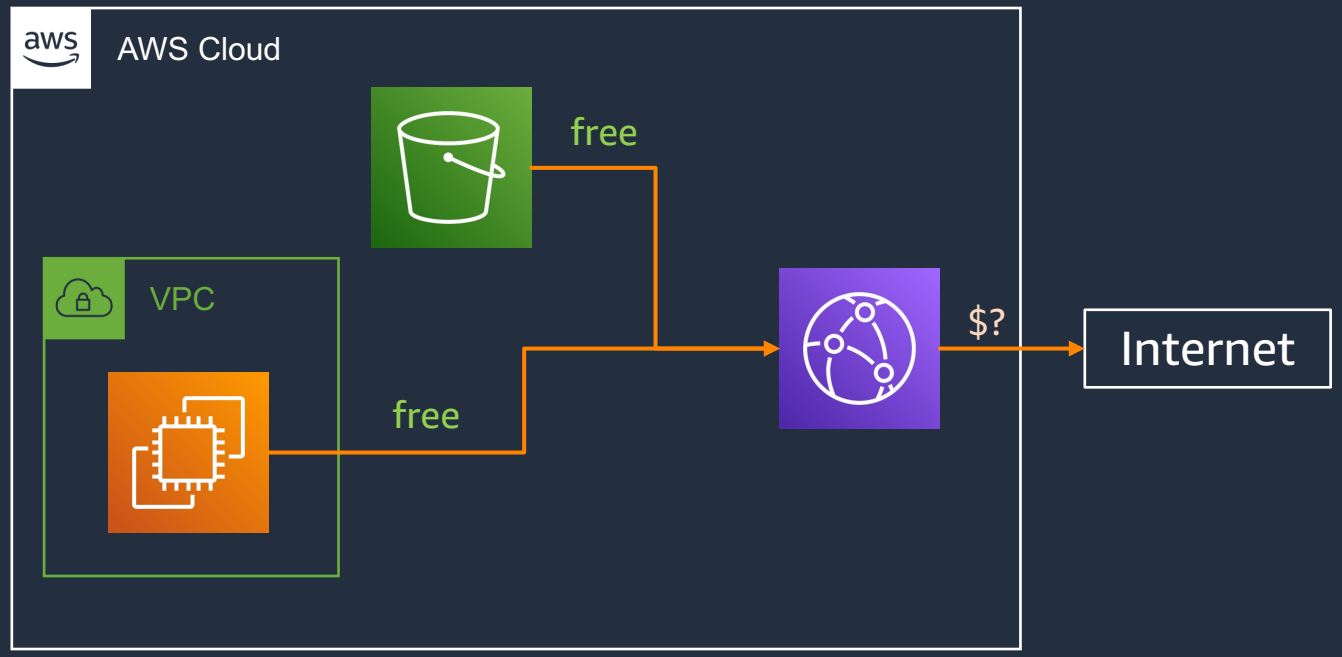
New

- **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

Data Transfer Optimization

with potentially low effort

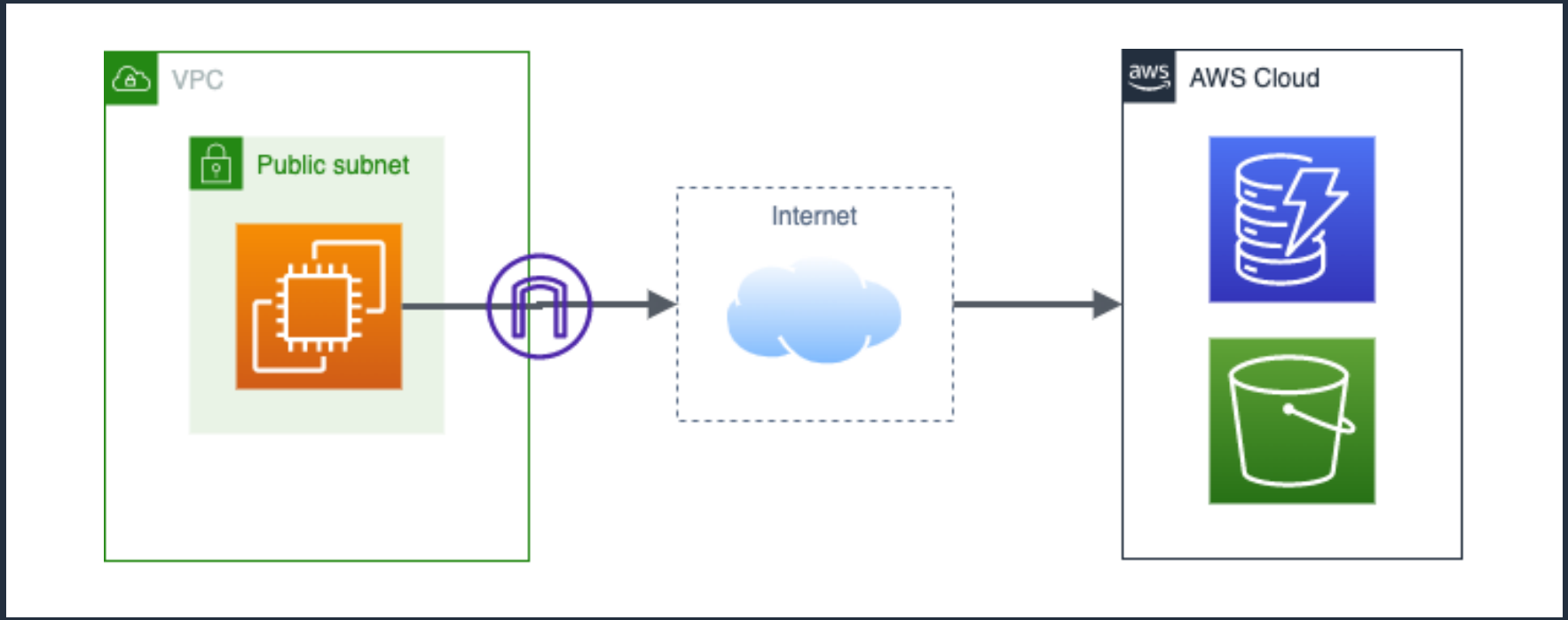
DTO – Adding CloudFront



- Enjoy Free Tier: first 1 TB free every month
- Benefited from volume-based discounts
- Better security & performance

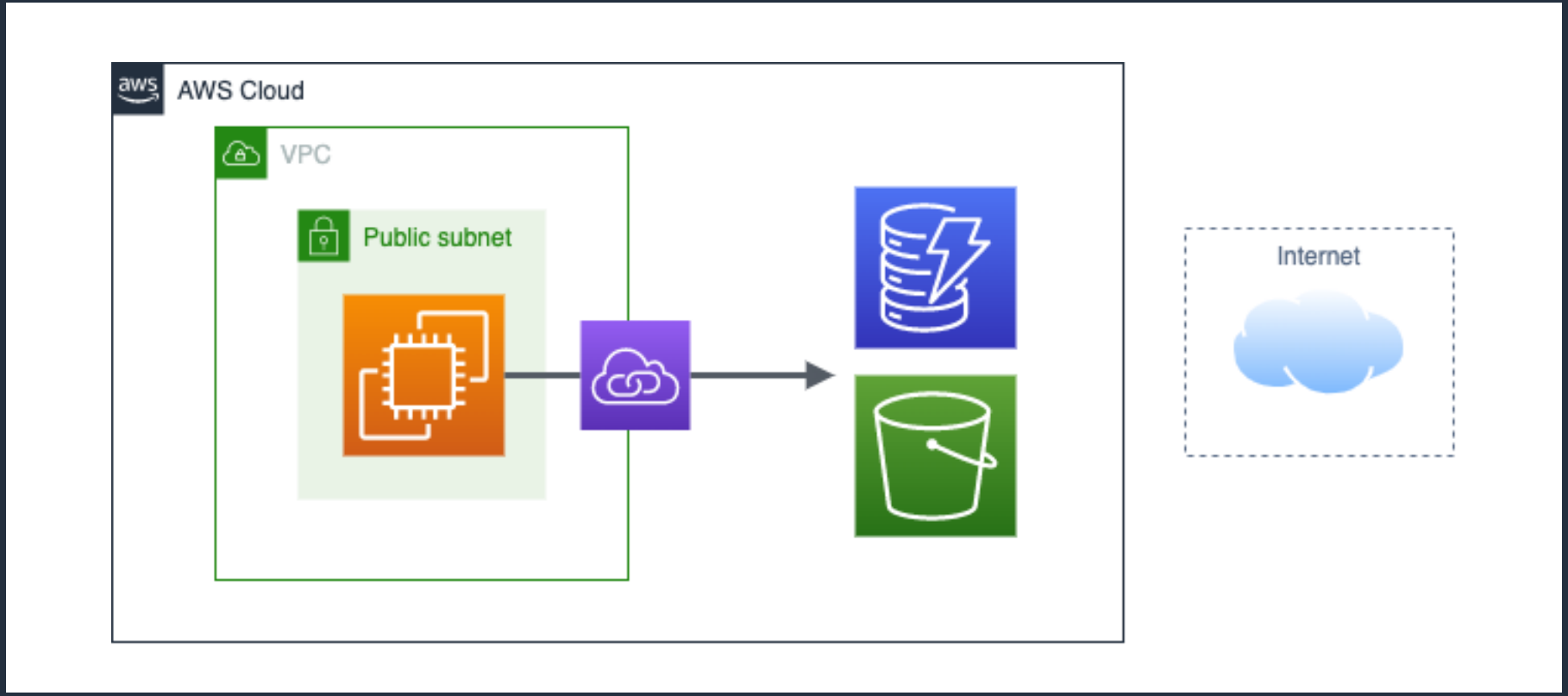
DTO – Anti-Pattern

EC2/Fargate/other services in VPC accessing S3/DynamoDB
over the internet and back into AWS



DTO – Best Practice

Solution: **VPC Gateway Endpoint**



Wrap Up: Key things to look for

- Purchasing options: Savings Plans, Reserved Instances, Spot
- EC2
  - Resource Scheduler
  - Right-Sizing: CloudWatch, Compute Optimizer
  - Graviton + New generation
- EBS GP2 → GP3
- S3 Intelligent-Tiering
- DTO: CloudFront, VPC Endpoint

Thank you!

Natavit Rojcharoenpreeda
Startup Solutions Architect
Amazon Web Services