Architect for Savings on AWS

Cost Optimization Strategies

AWS Vietnam Solutions Architect (SA) team

© 2020, Amazon Web Services, Inc. or its Affiliates.

Objectives for today

To reduce the amount you spend on Amazon Web Services…

…for your existing workloads

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What we'll cover

1. Cost Optimization – AWS WAR
2. Cost-effective resources
3. Matching supply with demand
4. Usage & expenditure awareness

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

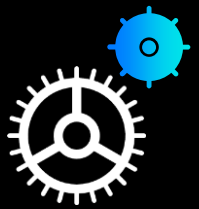
AWS **Well-Architected**

https://aws.amazon.com/well-architected/

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

aws

Well-Architected: The 5 pillars



Operational excellence

Security

Reliability

Performance efficiency

Cost optimization

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

General Design Principles

Stop guessing capacity needs

Test systems at production scale

Automate to make architectural experimentation easier

Allow for evolutionary architectures

Drive your architecture using data

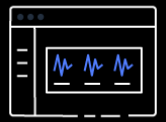
Improve through Game Days

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Well-Architected: Cost optimization

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

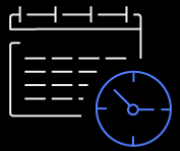
Well-Architected: Cost optimization areas of focus



Cost-effective resources

Matching supply with demand

Expenditure awareness



Optimizing over time

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cost-effective resources

1. Cost Optimization – AWS WAR
2. Cost-effective resources
   - a **Managed Services**
3. Matching supply with demand
4. Usage & expenditure awareness

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Managed services

Remove the burden of undifferentiated heavy lifting

- Focus on innovating rather than keeping the lights on
- Inherit AWS's approach to security, availability, performance

Managed services operate at cloud scale

- Can offer a lower cost per transaction or service

Help reduce or retire technical debt

- Move to services that are maintained by AWS
- Potential to remove or reduce license costs

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Managed services: Compute options

Amazon Elastic Compute Cloud (Amazon EC2)

AWS Elastic Beanstalk

Amazon Elastic Container Service

Amazon Elastic Container Service for Kubernetes

AWS Fargate

AWS Lambda

Unmanaged → Highly managed

Pay for Infrastructure → Pay per Transaction

aws

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Managed services: Relational database options

Amazon EC2

Amazon Relational Database Service (Amazon RDS)

Amazon Redshift

Amazon Aurora Serverless

Amazon Athena

Unmanaged → Highly managed

Pay for Infrastructure → Pay per Transaction

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Managed services: NoSQL database options

Amazon EC2

Amazon DocumentDB (with MongoDB compatibility)

Amazon DynamoDB

Unmanaged → Highly managed

Pay for Infrastructure → Pay per Transaction

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Managed services: Message queue options

Amazon EC2

Amazon MQ

Amazon Simple Queue Service (Amazon) SQS

Unmanaged → Highly managed

Pay for Infrastructure → Pay per Transaction

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Managed services: Takeaway

Managed Services can help you

- Remove the burden of undifferentiated heavy lifting.
- you can pay as you use and as you grow.

What services do you currently run (either on-premises or already in AWS) that could be replaced by an AWS managed service?

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cost-effective resources

1. Cost Optimization – AWS WAR
2. Cost-effective resources
   - a. **Managed Services**
   - b. **Appropriate provisioning**
3. Matching supply with demand
4. Usage & expenditure awareness

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Appropriate provisioning: Architectural choices

Steady-state vs burst workloads

- High throughput analytics workloads may benefit from constant compute capacity
- Bursting workloads often favour Serverless / Microservices architectural patterns

Consolidated vs separated workloads

- Can you combine multiple services (i.e., multiple databases on Amazon RDS)?
- Do workloads require isolation from other processes / data?

Cost vs performance trade-offs

- Is performance or cost optimisation the key business requirement?

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Appropriate provisioning: Server vs Serverless

Scenario 1: Constant, steady-state application workload

- Constant 100000 req/sec like system logging of calls (in telecom)
- 16 concurrent processes running constantly
- each process requiring 512 MB RAM

Scenario 2: Bursty HTTPS-based API service

- Airplane booking system
- ~6M requests per month
- 200ms and 128 MB used per request

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Appropriate provisioning: Cost-conscious design

Example: Should I use Amazon Simple Storage Service (Amazon S3) or Amazon DynamoDB?

AWS Simple Monthly Calculator

https://calculator.s3.amazonaws.com/index.html

aws

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Appropriate provisioning: Cost-conscious design

Scenario

- Webforms startup capturing form responses as JSON
- Application making high number of writes per second
- Need to understand the most cost-effective AWS storage service

Request rate (Writes/sec)	Avg Object size (Bytes)	Total size (GB/month)	Objects per month
300	2,048	1,483	777,600,000

aws

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon S3 or Amazon DynamoDB

Request rate (Writes/sec)	Object size (Bytes)	Total size (GB/month)	Objects per month
300	2,048	1,483	777,600,000

Amazon DynamoDB is a high performance n
is easy to set up, operate, and scale. It is de
problems of database management, perform
also provides predictable high performance

Indexed Data Storage:

Dataset Size: 1483 GB

Provisioned Throughput Capacity *:

Item Size (All attributes): 2 KB

Number of items read per second: 0 Reads/Second

Read Consistency: Strongly Consistent / Eventually Cons cheaper)

Number of items written per second: 300 Writes/Second

Amazon DynamoDB Service (US-East)		$ 644.30
Provisioned Throughput Capacity:	$ 261.69	
Indexed Data Storage:	$ 382.61	

Amazon S3 is storage for the Internet. It is designed to make web-scale computing easier for developers.

Storage:

Storage: 1483 GB

Reduced Redundancy Storage: 0 GB

Requests:

PUT/COPY/POST/LIST Requests: 77760000 Requests

GET and Other Requests: 0 Requests

Amazon S3 Service (US-East)		$ 3932.27
Storage:	$ 44.27	
Put/List Requests:	$ 3888.00	

Amazon S3 or Amazon DynamoDB

"…but what happens if I change the object size to 32 KB?"

Scenario

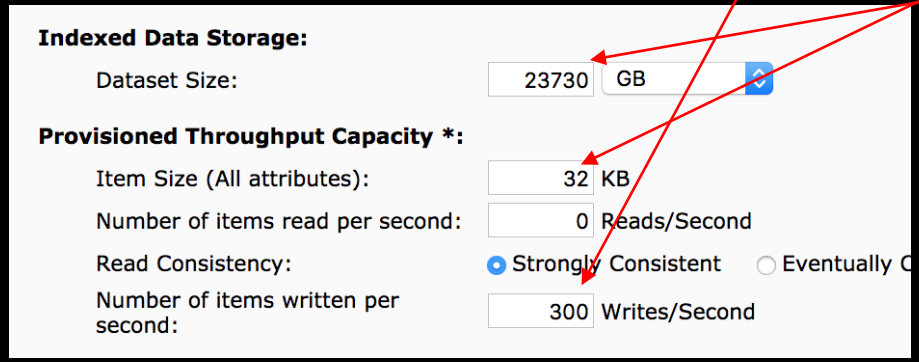
- Office claims receipts captured as PDFs

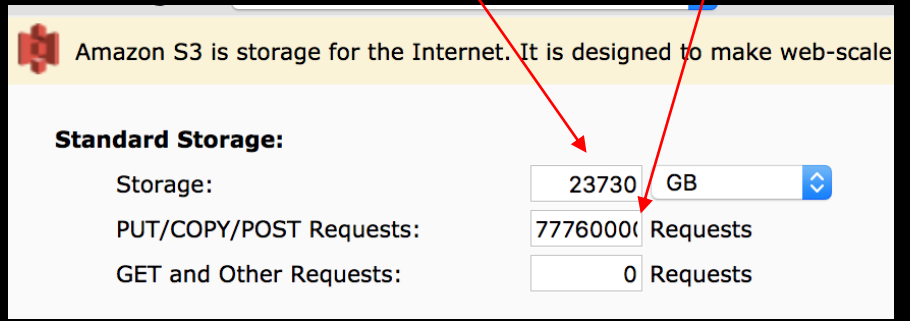
Request rate (Writes/sec)	Avg Object size (Bytes)	Total size (GB/month)	Objects per month
300	**32,768**	**23,730**	777,600,000

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon S3 or Amazon DynamoDB

Request rate (Writes/sec)	Object size (Bytes)	Total size (GB/month)	Objects per month
300	32,768	23,730	777,600,000





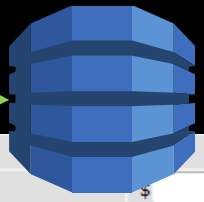
Amazon DynamoDB Service (US East (N. Virginia))		$ 10500.15
Provisioned Throughput Capacity:	$ 4555.79	
Indexed Data Storage:	$ 5944.36	

Amazon S3 Service (US East (N. Virginia))		$ 4433.79
Standard Storage:	$ 545.79	
Standard Put Requests:	$ 3888.00	

Amazon S3 or Amazon DynamoDB

	Request rate (Writes/sec)	Object size (Bytes)	Total size (GB/month)	Objects per month
Option 1	300	2,048	1,483	777,600,000
Option 2	300	32,768	23,730	777,600,000

use



Amazon S3 Service (US-East)		$ 3932.27
Storage:	$ 44.27	
Put/List Requests:	$ 3888.00	
Amazon DynamoDB Service (US-East)		$ 644.30
Provisioned Throughput Capacity:	$ 261.69	
Indexed Data Storage:	$ 382.61	
DynamoDB Streams:	$ 0.00	

use

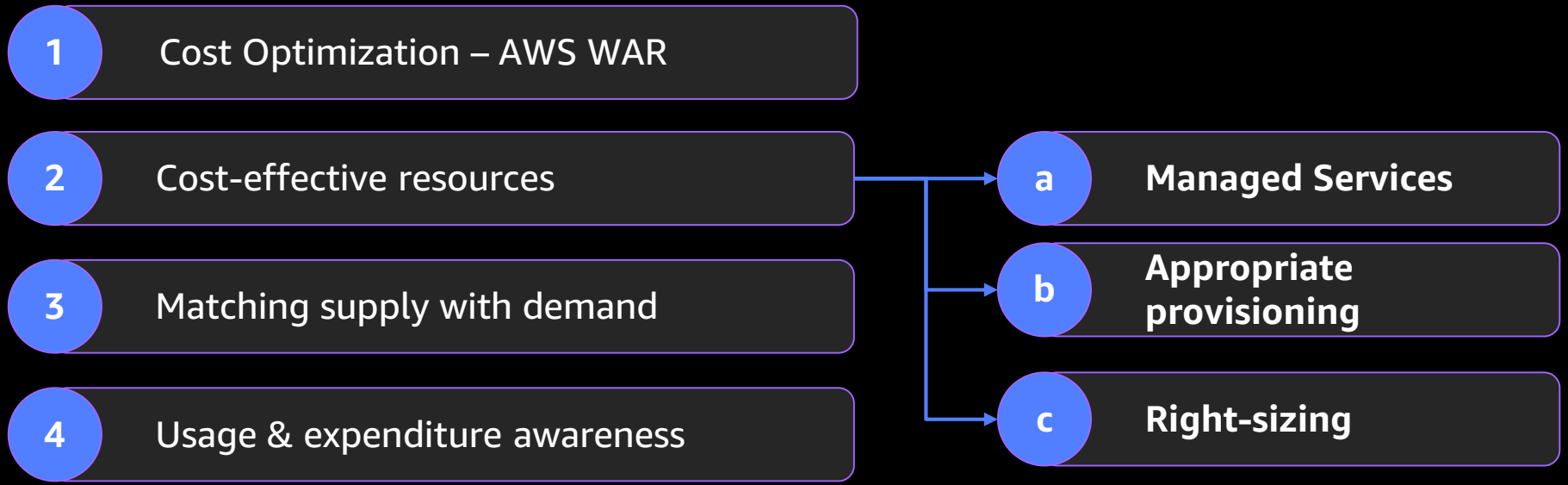
Amazon S3 Service (US East (N. Virginia))		$ 4433.79
Standard Storage:	$ 545.79	
Standard Put Requests:	$ 3888.00	
Amazon DynamoDB Service (US East (N. Virginia))		$ 10500.15
Provisioned Throughput Capacity:	$ 4555.79	
Indexed Data Storage:	$ 5944.36	

Appropriate provisioning: Takeaway

Have you selected the right architectures and associated AWS services needed to deliver each of your workloads?

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cost-effective resources



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Right-sizing

Use the lowest cost resources

- that meet the requirements of the specific workload

Iterate by adjusting the size of resources to optimize for costs

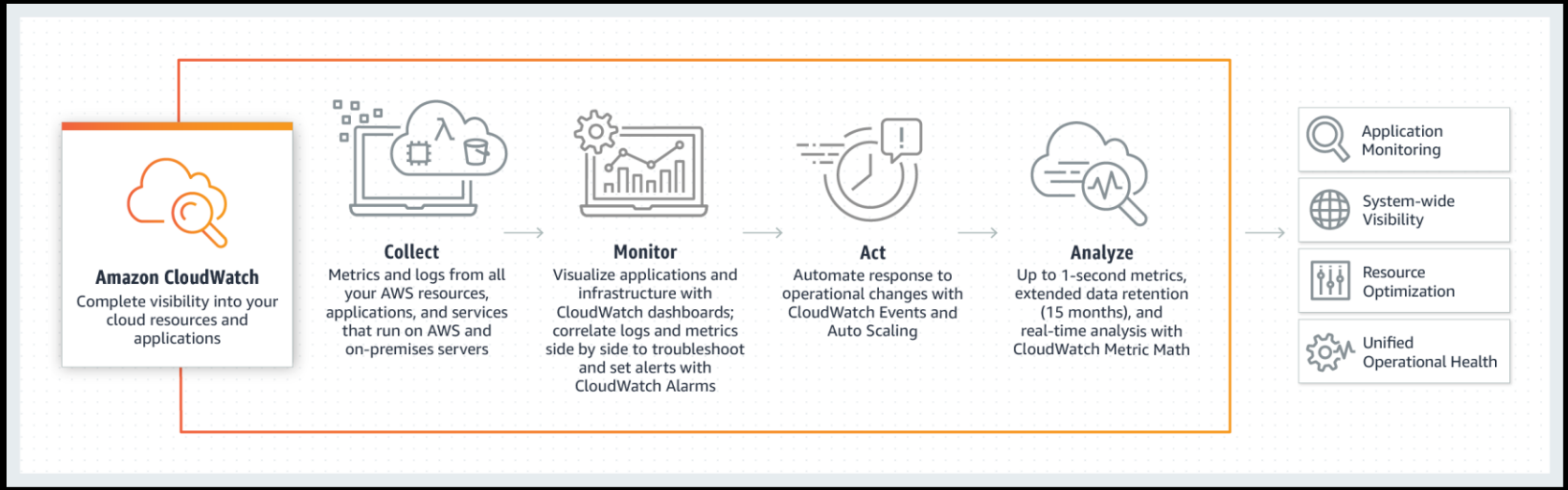
- Assess the cost of modification

Monitor resources and alarms to provide the data for right-sizing

- Monitoring should accurately reflect the end-user experience
- Select the correct granularity for the time period

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Right-sizing starts with monitoring



Get started quickly

Cloudwatch Agent

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Simplifying compute optimization

AWS Compute Optimizer

Identify optimal AWS Compute resources for your workloads

AWS resources metadata

CloudWatch Metrics

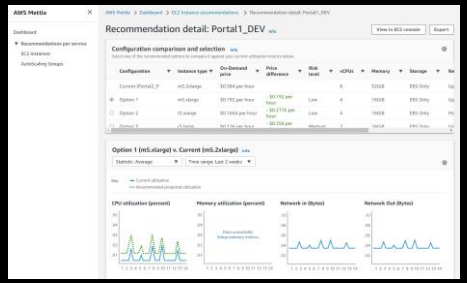
Mettle scans your AWS infrastructure and uses machine learning to automatically identify optimal AWS resources for your workloads

Identifies workload characteristics and profile based on the data gathered

Matches the resource requirements of your workloads to optimal AWS resources by recommending:

- EC2 Instance
- EC2 Auto Scaling Groups

Helps you visualize what-if scenarios based on the recommended resources



Lower costs

Optimize performance

Get started quickly

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Right-sizing: Amazon EC2 instances

Select the cheapest instance available

- Ensure you meet performance requirements
- Consider different instance families, not just sizes

Analyze using CloudWatch metrics

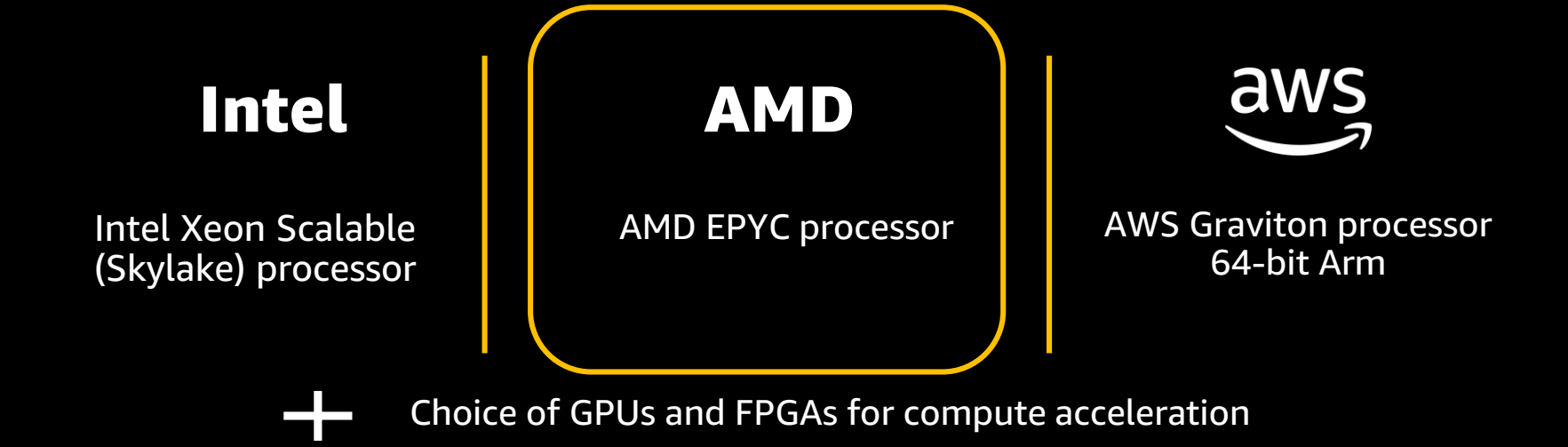
- Monitor CPU, RAM, storage, and network utilization
- Identify potential instances that can be downsized
- Set up custom metrics (i.e. RAM) where needed

Rule of thumb: Right-size first, then reserve

aws

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Choice of processors and architectures

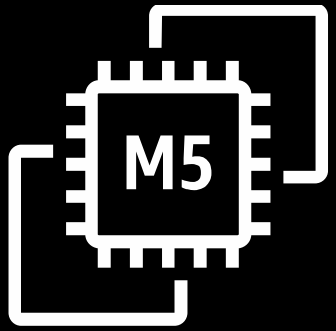


Right compute for each application and workload

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

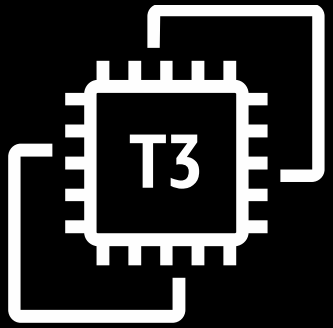
EC2 instance types with AMD EPYC™

General Purpose



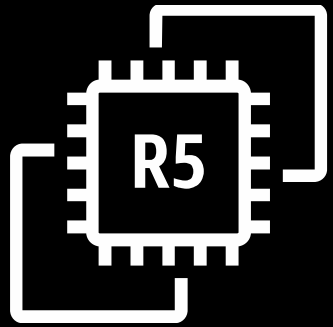
General
cloud server

Burstable



Cost
effective

Memory Optimized



Memory
intensive

A complete set of instances with custom AMD EPYC™ CPUs

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Choice of processors and architectures

Intel

Intel Xeon Scalable (Skylake) processor

AMD

AMD EPYC processor

aws

AWS Graviton processor 64-bit Arm

\+ Choice of GPUs and FPGAs for compute acceleration

Right compute for each application and workload

aws

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

First instance powered by AWS Graviton processor

Amazon EC2 A1

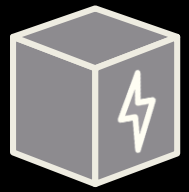
Run scale-out and Arm-based applications in the cloud

Up to 45% cost savings

AWS Graviton processor
64-bit Arm Neoverse cores and custom AWS silicon

Flexibility and choice for your workloads

Lower cost



Maximize resource efficiency with AWS Nitro System

aws

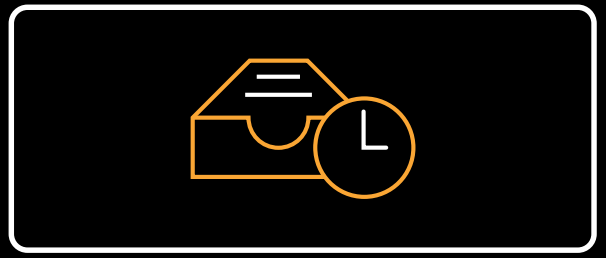
© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Targeted applications for Amazon EC2 A1

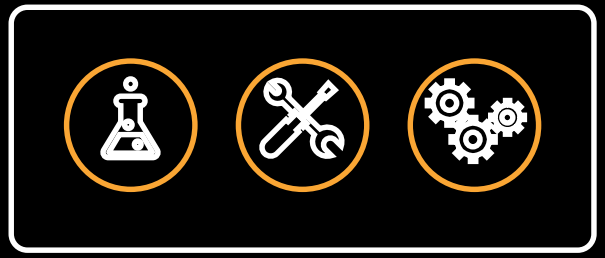
Web tier

Containerized microservices

Caching fleets



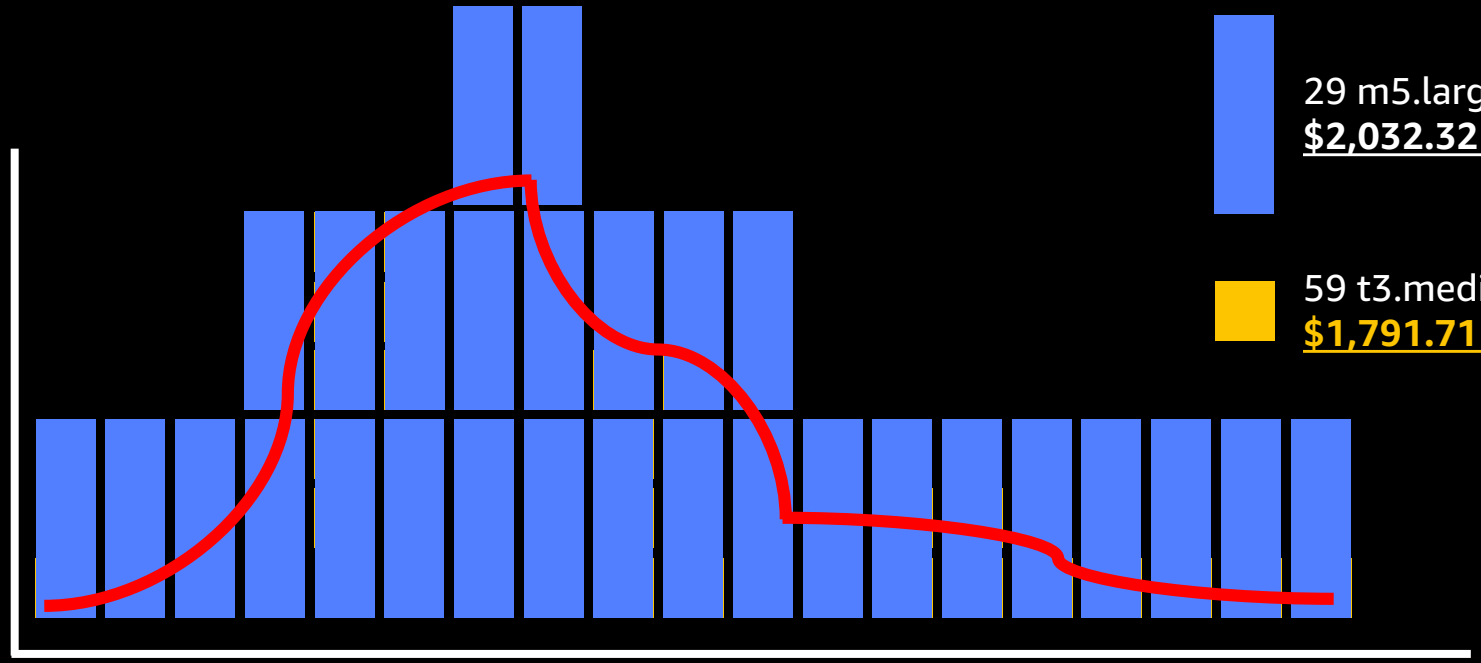
IoT, gaming, Arm workloads



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Right-sizing: Elasticity

More smaller instances vs. fewer larger instances



29 m5.large @ $0.096 /hr
$2,032.32 per month

59 t3.medium @ $0.0416/hr
$1,791.71 per month

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Right-sizing: AWS Lambda functions

Performance test your Lambda function

- CPU allocation is bound to memory. At 1,792 MB = 1 full vCPU
- Observe usage in AWS CloudWatch Logs

Identify if your function is memory-bound or cpu-bound

- Benchmarking (https://bit.ly/2VpeQmx)

Minimize your deployment package size and complexity

- simpler frameworks, smaller packages
- VPC if really needed
- Evaluate your language for cold start performance (https://bit.ly/2RB3L0x)
- Consider strategies to keep lambdas warm

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Lambda - CPU-bound example

"Compute **1,000 times** all prime numbers **<= 1M**"

128 MB	11.722 sec	$0.024628
256 MB	6.678 sec	$0.028035
512 MB	3.194 sec	$0.026830
1024 MB	1.465 sec	$0.024638

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Right-sizing: Amazon S3

Amazon S3 offers a range of storage classes:

- Standard
- Standard – Infrequent Access
- One Zone – Infrequent Access
- Glacier
- Glacier Deep Archive
- Reduced Redundancy (no longer recommended)

Key points to note

- Standard, Standard-IA and One Zone-IA are “real-time” storage tiers
- Glacier and Glacier Deep Archive are “near-line” storage tiers

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Right-sizing: Amazon S3 Storage Options

	Standard	Standard-IA	One Zone-IA	Glacier	Glacier DA
Storage Pricing (GB/mth)	$0.023	$0.0125	$0.01	$0.004	$0.00099
Request Pricing (per 1k GETs)	$0.0004	$0.001	$0.001	$0.05	$0.10
Retrieval Pricing (per GB)	N/A	$0.01	$0.01	$0.01	$0.02

Key points to note

- Storing data in IA tiers is about 50% cheaper than in Standard tier
- Request charges for IA tiers are about 60% more expensive than for Standard tier
- **IA tiers charge smaller objects as though they were 128 KB in size**
- IA tiers have a minimum storage duration of 30 days; Glacier 90 days, Glacier DA 180 days

aws

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Right-sizing: Amazon S3 Scenario

Assumptions

- 1,000 objects, each 1 GB in size, are stored in Amazon S3

Scenarios

- **Scenario 1:** every object is retrieved once per quarter (Shareholder reports)
- **Scenario 2:** every object is retrieved once per month (End of month bank statements)
- **Scenario 3:** every object is retrieved once per day (Shared todo list)

What storage class should we use for each scenario?

- We want to optimize for cost over a 12-month period?

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Right-sizing: Amazon S3 Scenarios

Cost components	Scenario 1 (Once/Quarter)	Scenario 2 (Once/Month)	Scenario 3 (Once/Day)
S3 Standard Storage Cost	1,000 x 0.023 x 12 $276.00	1,000 x 0.023 x 12 $276.00	1,000 x 0.023 x 12 $276.00
S3 Standard Request Cost	1,000 x 0.0004 x 4 $1.60	1,000 x 0.0004 x 12 $4.80	1,000 x 0.0004 x 365 $146
S3 Standard Annual TOTAL	**$277.60**	**$280.40**	**$422.00**
S3-IA Storage Cost	1,000 x 0.0125 x 12 $150.00	1,000 x 0.0125 x 12 $150.00	1,000 x 0.0125 x 12 $150.00
S3-IA Request Cost	1,000 x 0.001 x 4 $4.00	1,000 x 0.001 x 12 $12.00	1,000 x 0.001 x 365 $365.00
S3-IA Retrieval Cost	1,000 x 0.01 x 4 $40.00	1,000 x 0.01 x 12 $120.00	1,000 x 0.01 x 365 $3,650.00
S3-IA Annual TOTAL	**$194.00**	**$282.00**	**$4,165.00**

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Right-sizing: Amazon S3

Rule of thumb: if you're retrieving an object once per month or more, Standard is more cost effective storage class than Infrequent Access

- Remember, storage classes can be set on a per-object basis, not just per bucket
- Managing storage classes on a per-object basis can be complex and time-consuming

Consider using Amazon **S3 Intelligent-Tiering** if you have changing or unknown access patterns

- Automatically moves your data based on changing access patterns
- Moves data between Standard and Infrequent-Access tiers
- Additional management charge ($0.0025 per 1,000 objects per month)

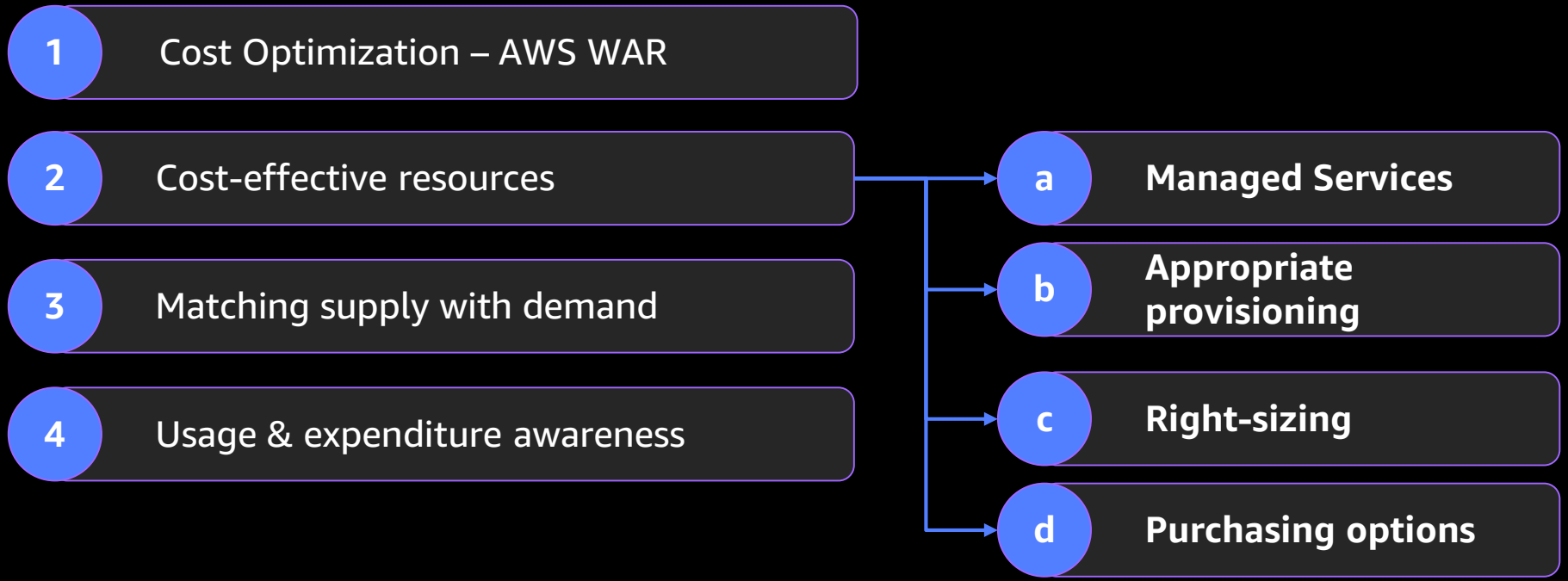
© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Right-sizing: Takeaway

When was the last time you reviewed your AWS infrastructure and looked for right-sizing opportunities?

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cost-effective resources



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Purchasing options

On-demand: pay per unit of capacity as used

- Examples: Amazon EC2, Amazon S3

Provisioned: pay per unit of capacity as provisioned

- Examples: Amazon DynamoDB, Amazon Kinesis Data Streams

Reserved: discounted pricing in return for a fixed-term commitment

- Examples: Amazon EC2, Amazon Elasticsearch Service

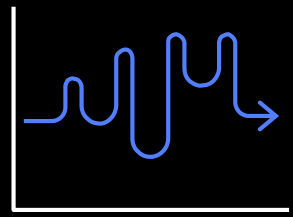
Many services support multiple billing options

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 purchase options

On-Demand

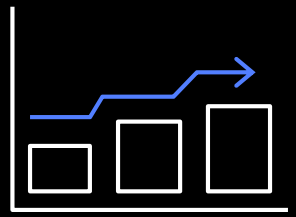
Pay for compute capacity by **the second** with no long-term commitments



Spiky workloads to define needs

Reserved Instances (RIs)

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



Committed and steady-state usage

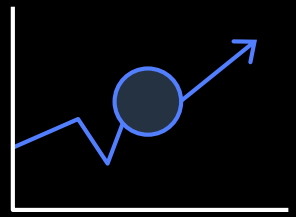
Savings Plans

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

Flexible access to compute

Spot Instances

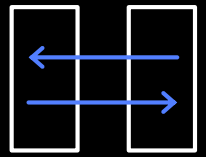
Spare Amazon EC2 capacity at **savings of up to 90%** on On-Demand prices



Fault-tolerant, flexible, stateless workloads

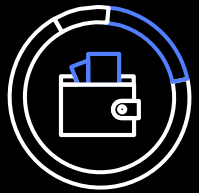
© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Save up to 90% using EC2 Spot Instances



Instances

Same infrastructure as On-Demand and RIs

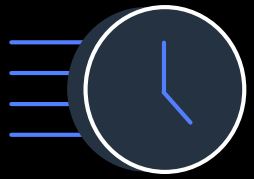


Pricing

Smooth, infrequent changes, more predictable

Usage

Choose different instance types, sizes, and AZs in a single fleet or EC2 Auto Scaling group



Capacity

Interruptions only happen if OD needs capacity

Pricing is based on long-term supply and demand trends; **no bidding!**

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Handling Spot interruptions

Less than 5% of Spot Instances were interrupted in the last 3 months

Minimal interruptions

Check for 2-minute interruption notification via instance metadata or Amazon CloudWatch events, and automate by

- ✓ Checkpointing
- ✓ Draining from ELB
- ✓ Using stop-start and hibernate to restart faster

Interruption handlers for Amazon ECS and Amazon EKS

Amazon Elastic Kubernetes Service (Amazon EKS)

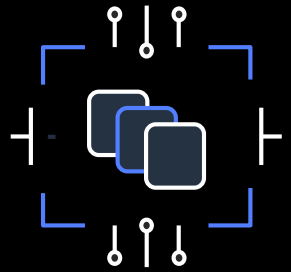
Amazon Elastic Container Service (Amazon ECS)

- ✓ Connection between termination requests from AWS infrastructure to nodes
- ✓ Tasks running on Spot Instances will automatically be triggered for shutdown before the instance terminates, and replacement tasks will be scheduled elsewhere on the cluster

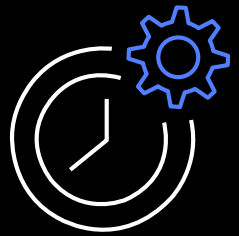
aws

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

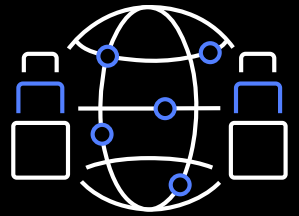
Flexibility is key to successful Spot usage



Instance flexible



Time flexible



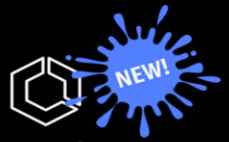
Region flexible

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EC2 Spot integrations - AWS Managed Services and ISVs

Amazon EC2 Auto Scaling

EC2 Fleet



Amazon Elastic Container Service (Amazon ECS)

Amazon Elastic Container Service for Kubernetes (Amazon EKS)

Amazon SageMaker

AWS Thinkbox

Amazon EMR

AWS CloudFormation

AWS Batch

AWS Elastic Beanstalk

databricks

Jenkins

docker

IBM Spectrum Symphony

Qubole

Bamboo

MESOS

cloudera

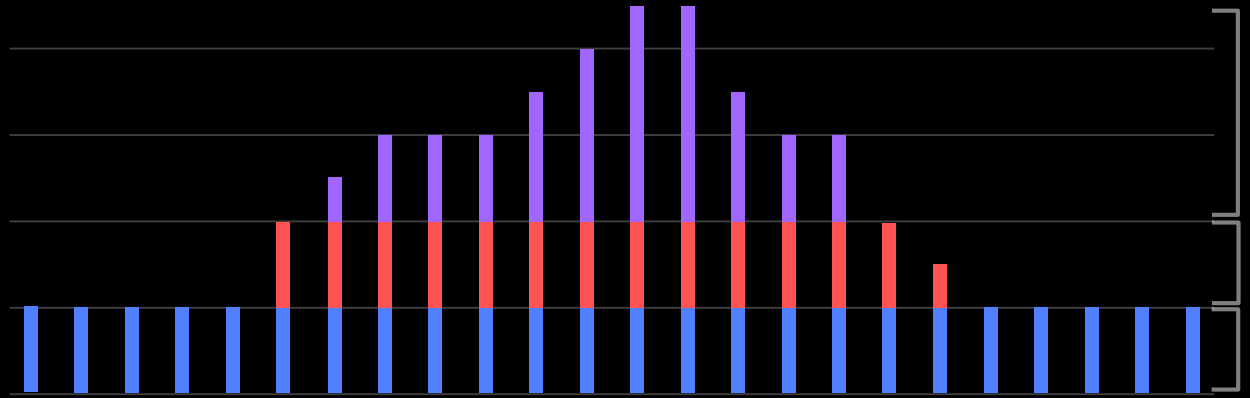
aws

HashiCorp Terraform

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

kubernetes

To optimize Amazon EC2, combine purchase options



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

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Purchasing options: Takeaway

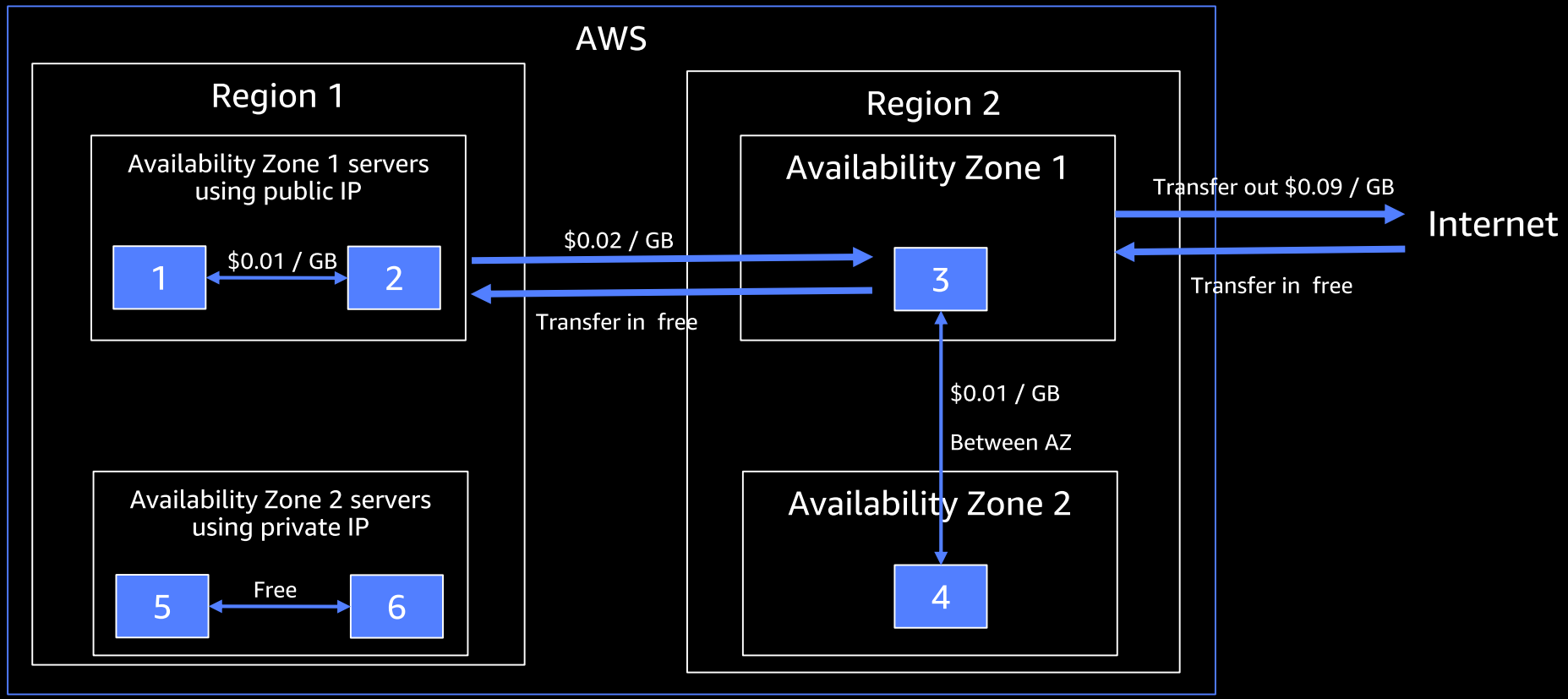
Look across your AWS infrastructure for opportunities to optimize the way in which you purchase services

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Transfer

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

Data Transfer Costs



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data Transfer Cost Optimization Tips

- Analyze Data Transfer Costs using Cost Explorer
- Use Private IP for Internal Services Communication
- Use VPC End-Points for S3 and DynamoDB
- Consider using CloudFront to reduce Data Transfer Out Costs

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Optimized data transfer: Takeaway

When Architecting your systems, balance your needs between availability, disaster recovery, security and cost

Architecting your systems to run within availability zone or regions will minimize data transfer costs

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Matching supply with demand

1. Cost Optimization – AWS WAR
2. Cost-effective resources
3. Matching supply with demand
4. Usage & expenditure awareness

a. **Demand based**

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demand-based

Leveraging the elasticity of the cloud to meet demand as it changes can provide significant cost savings

- programmatically vary the amount of cloud resources in your architecture dynamically
- increase the number of resources during demand spikes to maintain performance
- decrease capacity when demand subsides to reduce costs

Within AWS this is normally accomplished using Auto Scaling

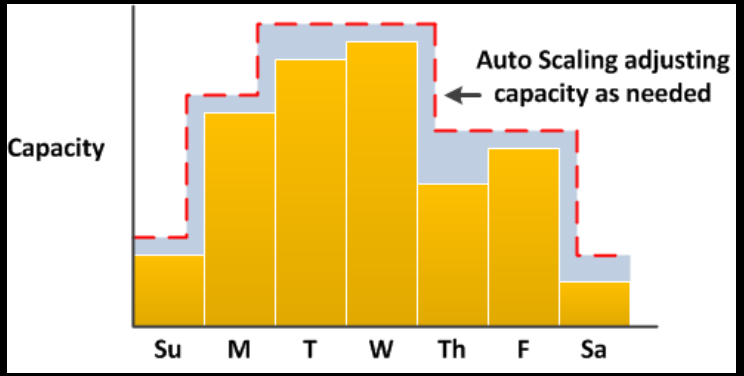
- EC2 Auto Scaling: Add or remove compute capacity to meet changes in demand
- AWS Auto Scaling: Configure and manage scaling for scalable AWS resources through a scaling plan

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demand-based: Amazon EC2 elastic provisioning

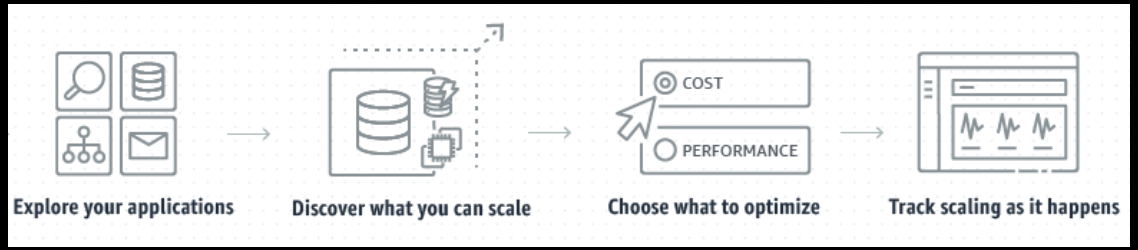
EC2 Auto Scaling allows you to:

- React dynamically to changes in load
- Schedule regular workloads
- Optimise your instance usage
- Reduce over-provisioning
- No cost service!



AWS Auto Scaling

- Unified scaling for your cloud applications



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demand-based: Takeaway

Look across your workloads and consider what metrics you could use with auto-scaling to ensure capacity is closely aligned with demand

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Matching supply with demand

1. Cost Optimization – AWS WAR
2. Cost-effective resources
3. Matching supply with demand
   - a. **Demand based**
   - b. **Time based**
4. Usage & expenditure awareness

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Time-based

A time-based approach aligns resource capacity to demand that is predictable or well defined by time

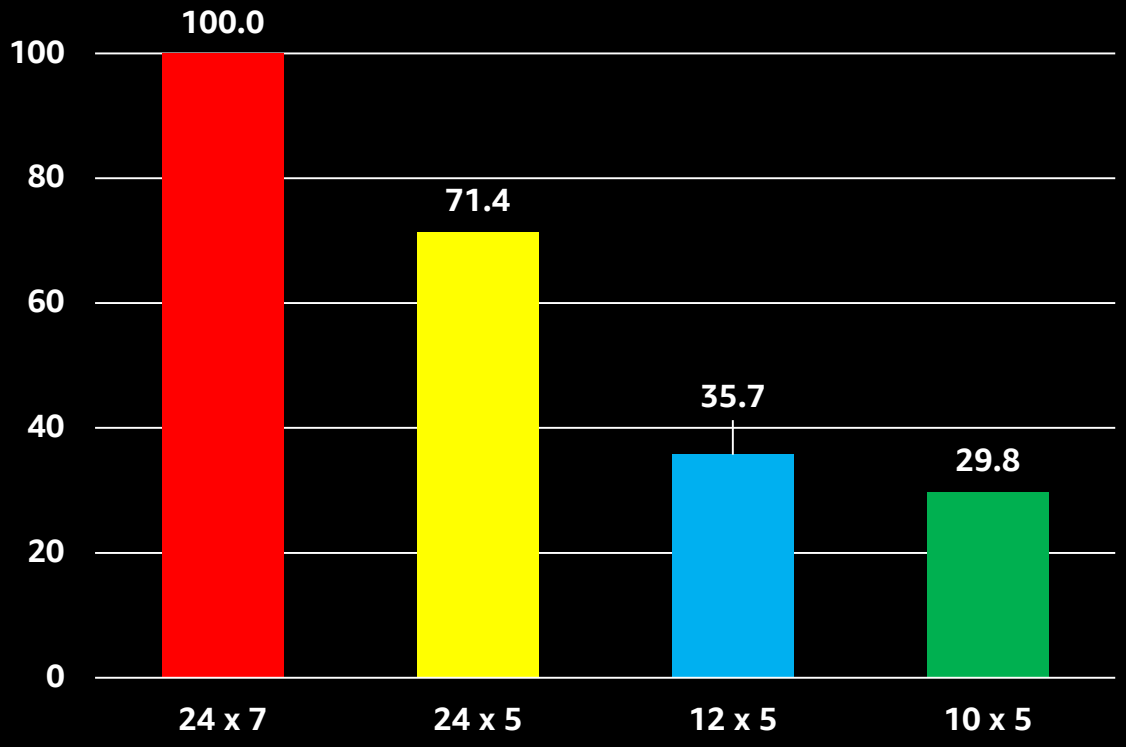
- typically not dependent upon utilization levels of the resources
- ensures that resources are available at the specific time they are required
- provided without any delays due to start-up procedures

Key considerations

- how consistent is the usage pattern?
- what is the impact if the pattern changes?

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Time-based: Workload scheduling



Up to **70%** savings for non-production workloads

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Instance Scheduler

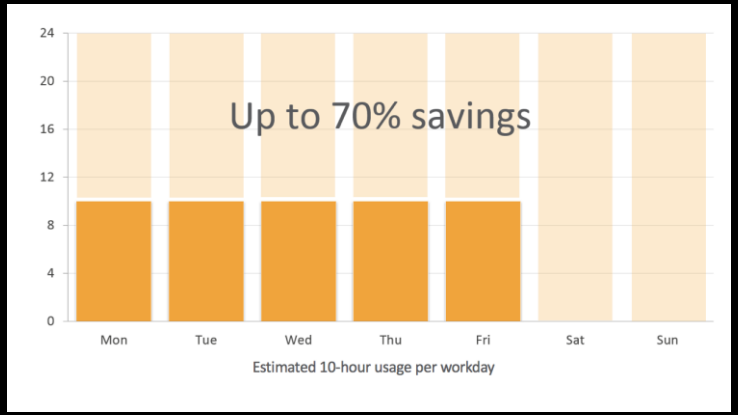
AWS-provided solution

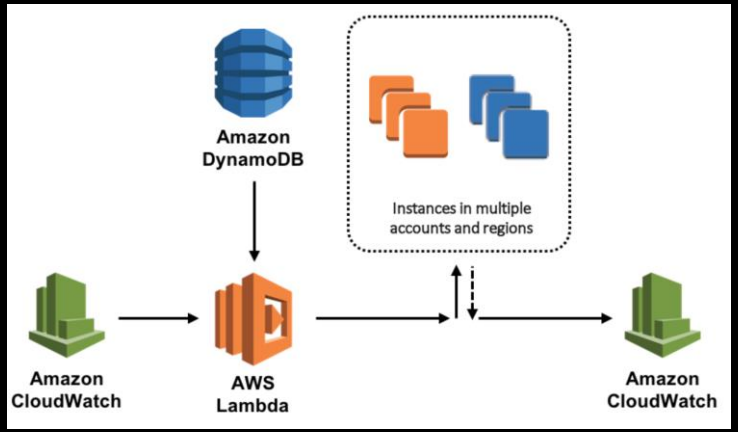
- Custom start & stop schedules
- Works with Amazon EC2 & Amazon RDS instances
- Deploy using AWS CloudFormation

Selectively tag instances to schedule

- Multiple schedules per instance
- 5-minute granularity

https://aws.amazon.com/answers/infrastructure-management/instance-scheduler/





© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Time-based: Takeaway

Identify workloads that don’t need to be running 24x7,
and start scheduling them appropriately

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Usage & expenditure awareness

1. Cost Optimization – AWS WAR
2. Cost-effective resources
3. Matching supply with demand
4. Usage & expenditure awareness
   - a. **Tagging**

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tagging

Tagging provides a number of benefits

- Automation (autoscaling, scheduling)
- Control & compliance (IAM policies)
- Cost allocation (reporting & chargebacks)

Not all resources support tagging

- Not all cost line-items support tagging (e.g., Data transfer charges)

Other tagging gotchas

- Maximum of 50 user-applied tags (not counting system tags)
- Values are optional
- Tags are case-sensitive

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

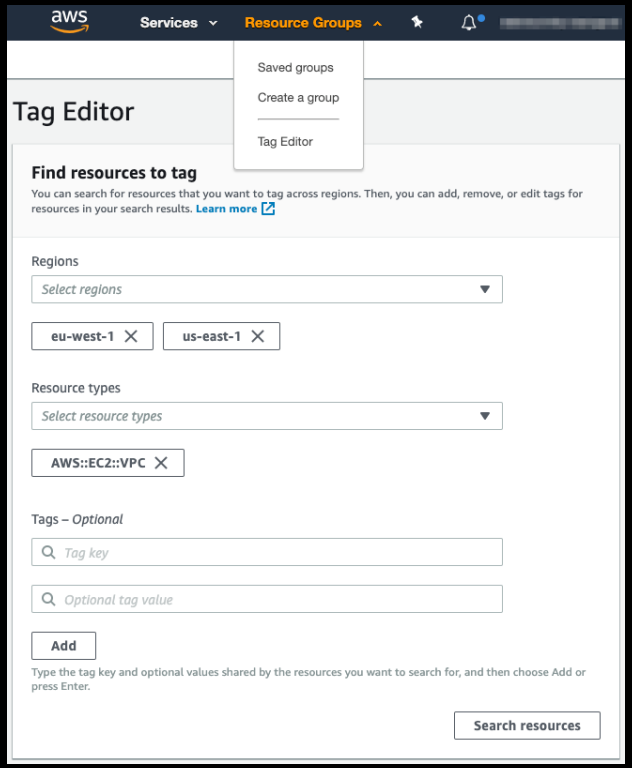
Tagging: Stakeholders & examples

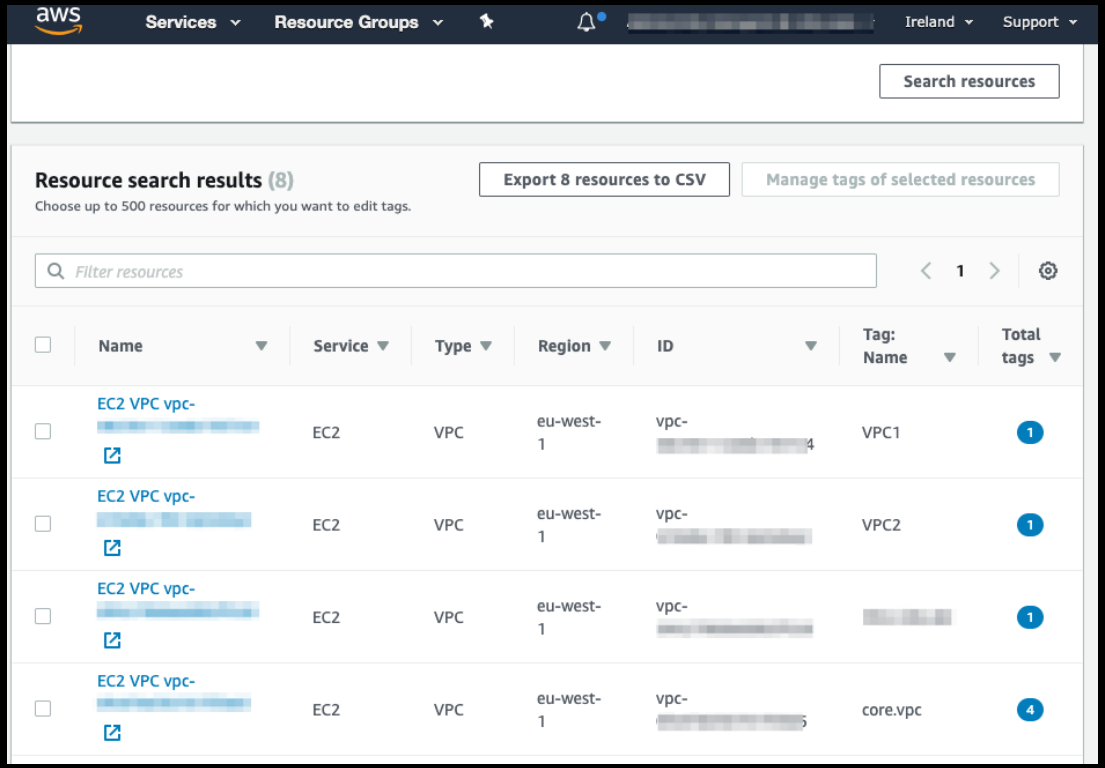
Stakeholder	Example Tag Key	Example Tag Value
Finance	CostCenter BudgetCode	Engineering EG-001
Engineering	Workload Codebase	Website Python
Line-of-business owners	Project User	SuperSecretProject Alice
IT	BackupRegime Environment	24x7 Production
Security	PatchStrategy AutomationSupport	Immutable True

https://aws.amazon.com/answers/account-management/aws-tagging-strategies/

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tagging: Using the tag editor





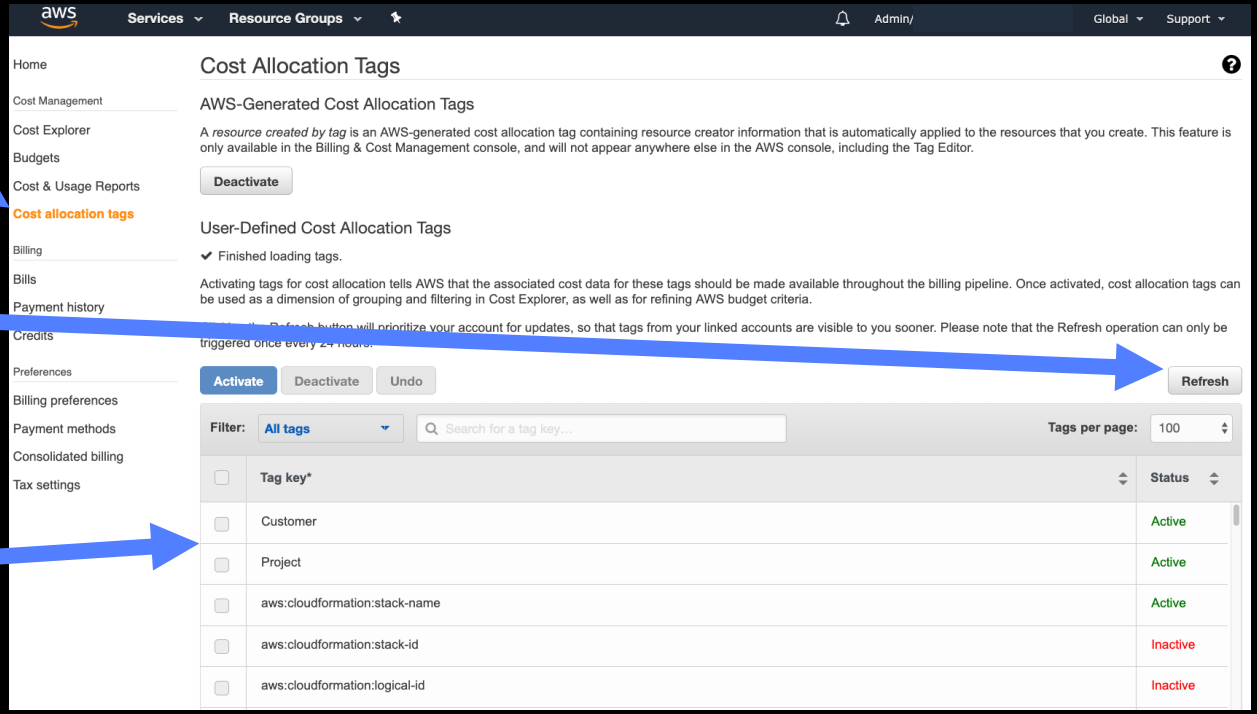
© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tagging: Enabling cost allocation tagging

Enable cost allocation tags

Refresh to get tag list from AWS Organizations

Choose the tags to activate



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tagging: Takeaway

Identify and tag all existing resources across your AWS accounts, and then implement AWS Identity and Access Management controls to enforce appropriate tagging policies

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Usage & expenditure awareness

1. Cost Optimization – AWS WAR
2. Cost-effective resources
3. Matching supply with demand
4. Usage & expenditure awareness
   - a. **Tagging**
   - b. **Visibility & governance**

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Visibility & governance

Detailed visibility into your AWS environment

- identify opportunities for savings

Cost optimization requires

- a granular understanding of the breakdown in spend
- Ability to model and forecast future spend
- Having sufficient mechanisms in place to align cost and usage to business objectives

AWS provides a suite of reports and tools

- estimate, monitor, plan, notify, report on, and analyze your AWS spend

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Visibility & governance: AWS Cost Explorer

Comprehensive dashboards

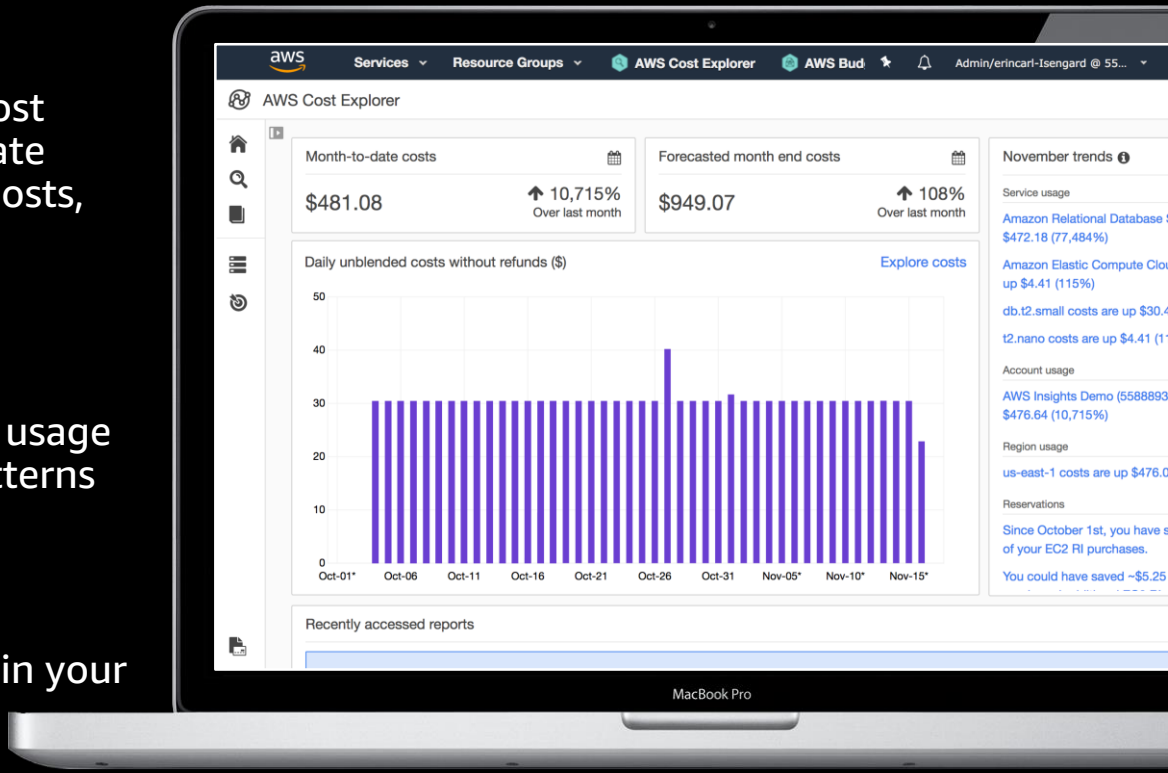
- Gain a summary view of key cost details, including month-to-date costs, month-end forecasted costs, and saved reports

Automated trend analysis

- Identifies anomalous cost and usage events, based on historical patterns

Optimized user experience

- Users of all levels of expertise in your organization



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Visibility & governance: RI recommendations

Automated purchase recommendations

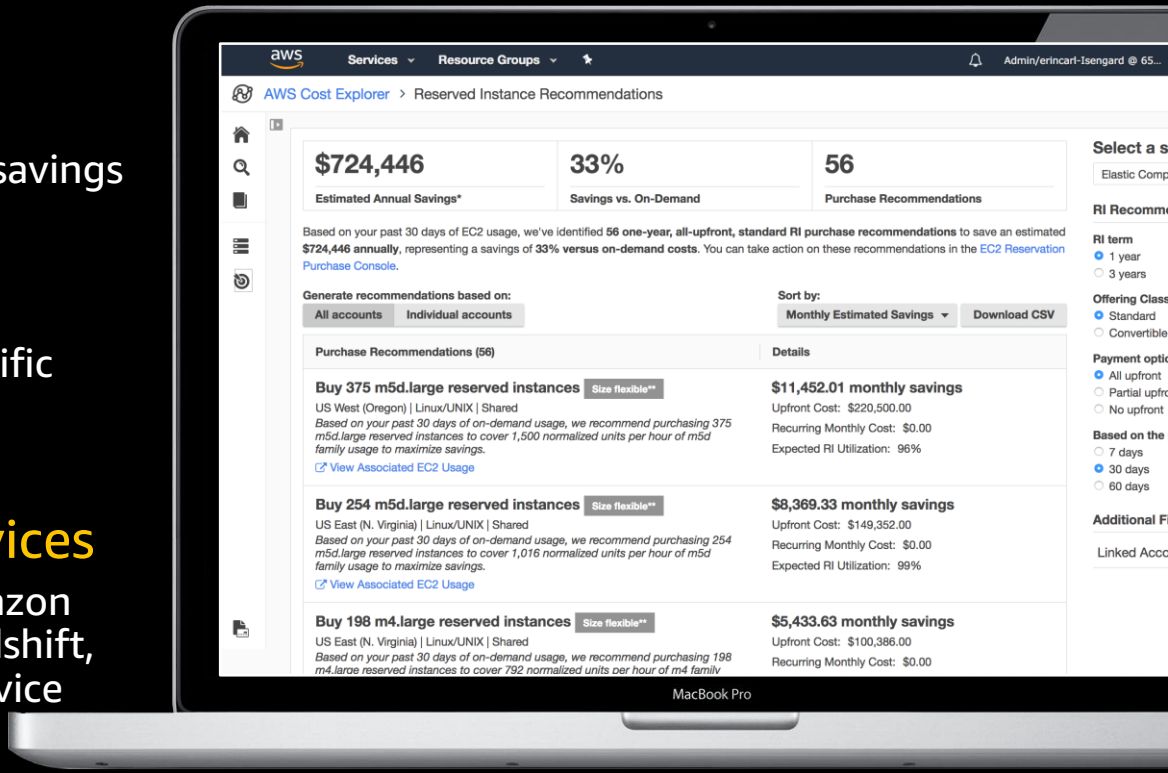
- Analyzes your historical usage patterns to identify potential savings

Customizable parameters

- Purchase RIs that fit your specific business requirements

Supports multiple AWS services

- Support for Amazon EC2, Amazon RDS, ElastiCache, Amazon Redshift, and Amazon Elasticsearch Service reservations



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Visibility & governance: AWS Budgets

User experience

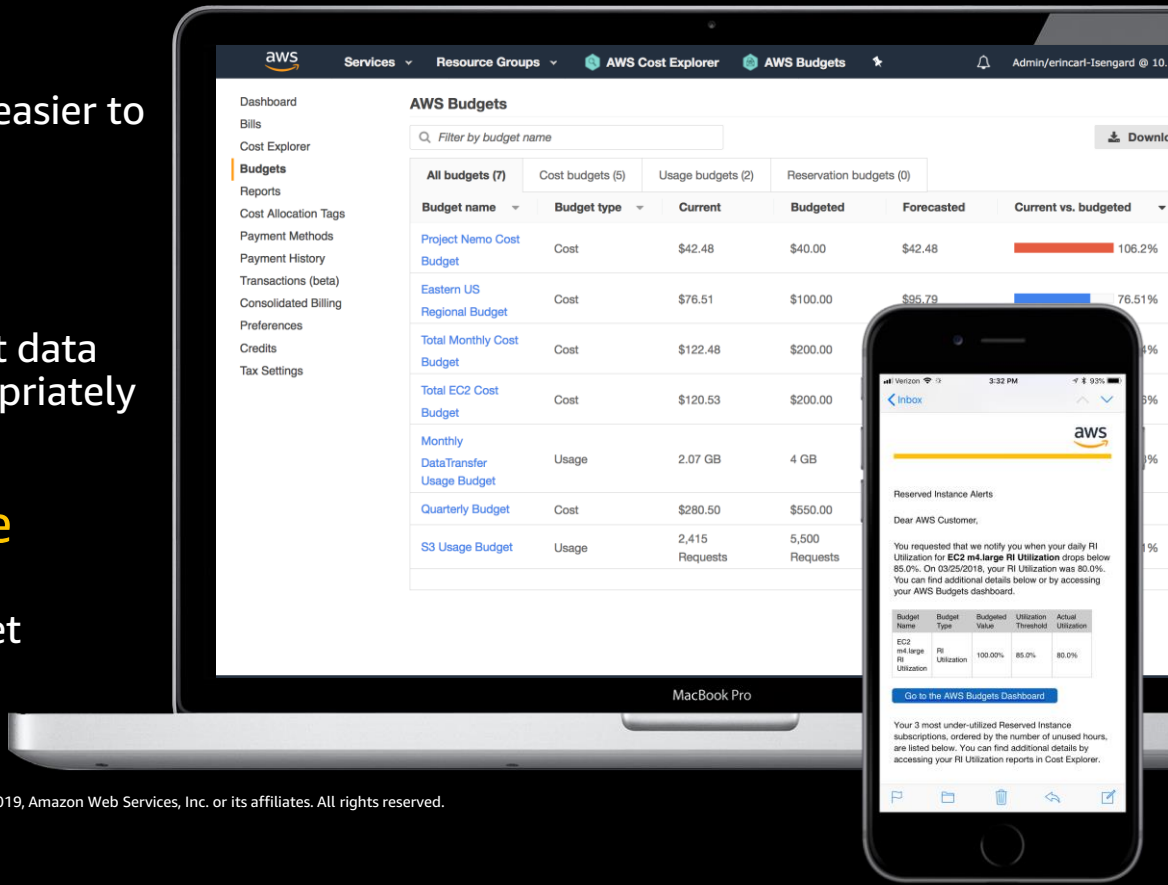
- Simplified workflows make it easier to create and manage budgets

Cost Explorer integration

- Provides contextually-relevant data to help you set budgets appropriately

Review budget performance

- View how your actuals have performed against your budget



© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Visibility & governance: AWS Trusted Advisor

Taking away the heavy lifting of monitoring best practices

- Trusted Advisor provides best practices (or checks)
- Only available for Business support customers

Red (action recommended)
Yellow (investigation recommended)
Green (no problem detected)

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Visibility & governance: Takeaway

Create reports, budgets, and alarms to track spend and alert when this deviates from expected norms

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What Next?

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark

AUTODESK

Uses Spot Instances and AWS Auto Scaling for its rendering-as-a-service workload to spend **less and scale more**

FRED HUTCH
CURES START HERE™

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

Standard Chartered

Reduced grid infrastructure **costs by 60%**

Western Digital.

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

skyscanner

Was able to **save 74%** on its Kubernetes cluster

matterport

Processes tens of thousands of 3D models daily; reduced compute costs by **70%,** savings **$1 million** yearly

NOVARTIS

What was originally estimated to take 39 years and $40 million took **9 hours and $4,232**

lyft

Saved **75% a month** by changing four lines of code

MOBILEYE

A job that took **weeks now takes hours,** thanks to great parallelism, at a very cost-efficient price

AdRoll

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

illumina

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

QBiC

Reduced **queue time by 50%** by using Spot Instances

aws

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Summary of takeaways (1)

Cost-effective resources

- Select the right Cloud-native architectures for each workload
  - Choose instance or serverless based on compute behaviour
  - Calculate costs on for storage based on data behaviour (S3
- Review your existing workloads and right-size as required
- Make use of the full range of AWS purchasing models

Matching supply and demand

- Make use of auto-scaling to closely match capacity with demand
- Schedule non-production workloads to run only when they are needed

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Summary of takeaways (2)

Usage & expenditure awareness

- Tag all of your resources to enable cost attribution
- Ensure business owners have visibility into their workload costs
- Create reports and budgets, and alert when they deviate from what's expected
- Most of these tools are freely

Optimise over time

- Give responsibility (and authority) for cost optimization within your organization

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Useful Resources

AWS Pricing

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

Online TCO Calculator:

- https://awstcocalculator.com/

AWS Well-Architected Framework

- https://aws.amazon.com/well-architected/

AWS Simple monthly calculator

- https://calculator.s3.amazonaws.com/index.html

AWS Pricing calculator:

- https://calculator.aws/

AWS Cloud Economics Center:

- https://aws.amazon.com/economics/

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you

My Nguyen
SUS Solution Architect
Vietnam
Linkedin - https://www.linkedin.com/in/my-nguyen-4b5378100/

Andy Tran
SUS Ecosystem Solution Architect
Linkedin - https://www.linkedin.com/in/ahtran/

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark