
AWS
re:Invent

SVS206

Understanding your serverless application's TCO

Chhaya Pandey

Principal Product Manager, AWS Lambda

AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Module 1 **Modern application and TCO**

Module 2 **AWS Lambda pricing: Pick the right option**

Module 3: **Provisioned concurrency**

Module 4: **Savings Plans: How, what, why?**

Module 5: **What's next?**



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

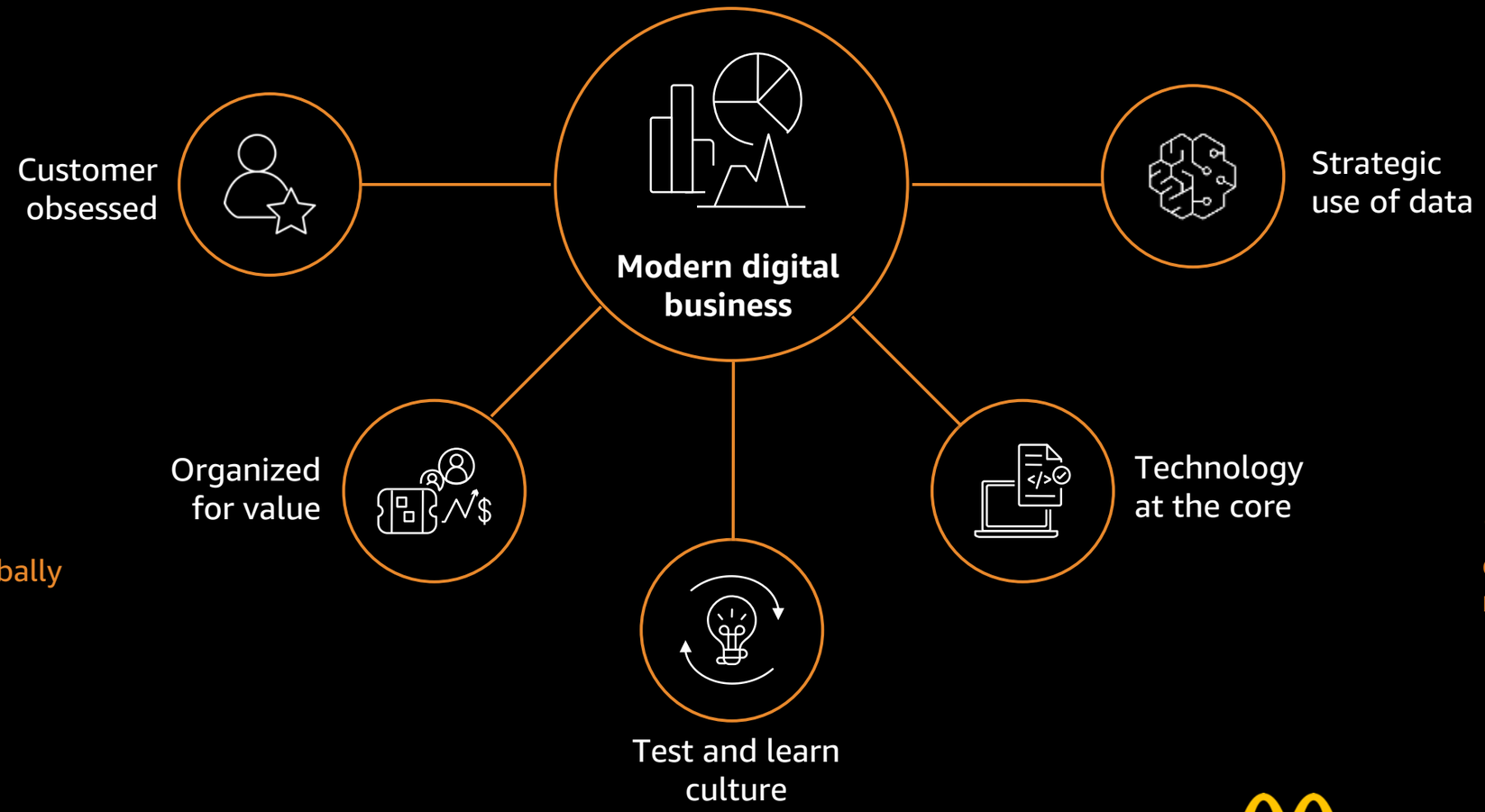
Module 1
Modern applications and TCO



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Characteristics of a modern digital business



Capital One

New features in weeks, enabling developers globally

FINRA

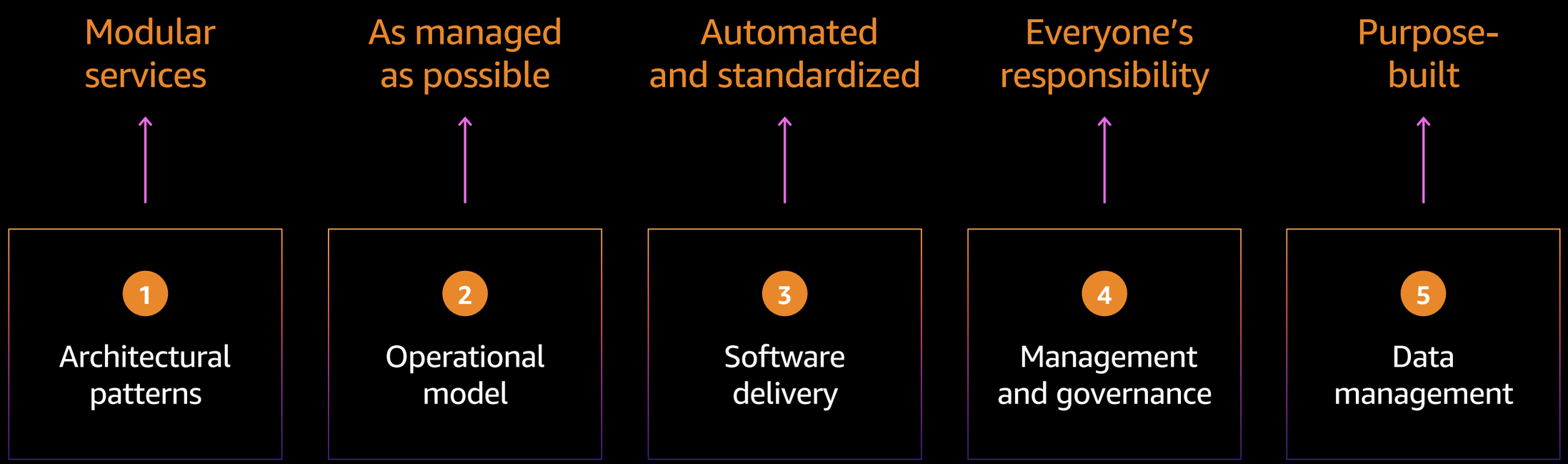
1,000x more market events, querying from minutes to seconds

Guardian

Invest, test, learn, and fail quickly – Retrained 2,500 employees to Agile

Home delivery rollout in 4 months, scalable to 1M orders / hour

Technology decisions drive your business strategy



A serverless business strategy

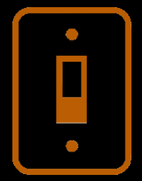


No infrastructure provisioning, no management



Automatic scaling

Pay for value



Highly available and secure



Serverless-first is the decision to opt for serverless technologies in your application as a first choice

Serverless TCO framework

Infrastructure cost

- Utilization on server-based architectures
- Peak-to-average ratio per function vs. overall workload

Maintenance cost

- Provisioning and scaling (weekly dev hours)
- Security implementation
- Patching and OD updates
- Ongoing application operations

Development cost

- Days to deploy
- Number of developers

Serverless TCO: Infrastructure cost savings

- Lambda configures resources on a per-function basis
- Functions can scale independently of each other, without needing a box big enough to meet the peak of your most bursty function
- As you only pay for hardware your function needs, utilization of resources on Lambda is higher than what it would be on a server-based architecture
- Even without the savings in developer time, the ability of serverless to increase utilization means it is cheaper than using VMs when the code is executed fewer than 500,000 times each month

Serverless TCO: Infrastructure cost savings

A customer moving from another cloud-based solution discussed the benefit of automatic scaling to demand for compute / storage

> "We've really benefited from automatic responses to capacity needs with AWS Lambda and the AWS Serverless Platform. Before, we had a capacity of 100 instances, whereas now we have the equivalent capacity of 2,000–3,000 instances per day."

Source: IDC 2018

Serverless TCO: Maintenance cost savings

	Without AWS serverless	With AWS serverless	Difference	Change % (based on avg.)	Change % (based on per org.)
Compute					
Time to deploy new compute (days)	19.3	0.1	19.2	100%	89%
Staff time to deploy new compute (hours)	75.8	0.1	75.7	100%	88%
Storage					
Time to deploy new storage (days)	14.2	0.5	13.7	96%	96%
Staff time to deploy new storage (hours)	80.4	0.1	80.3	100%	100%

Source: IDC 2018

Serverless TCO: Dev cost savings

- The TCO of serverless computing tends to be lower than VMs, according to 451 Research, even when the VM is hosting containers, because there is no need for developers to provision, configure, and manage infrastructure
- When a serverless function is active for just three quarters of the month, it only takes a 10-minute savings in operational overhead for serverless to beat VMs on TCO

Serverless TCO: Dev cost savings

	Without AWS serverless	With AWS serverless	Difference	Change %
Application developer productivity levels				
Productivity levels in terms of FTEs per organization	60.7	80.8	20.1	33%
Equivalent value of application developers per year per organization	$6.07 million	$8.08 million	$2.01 million	33%
New applications, new logic				
Number developed per year	20.5	24.2	3.7	18%
Development lifecycle (weeks)	21.9	5.2	16.6	76%
New features				
Number developed per year	187	402	215	115%
Development lifecycle (weeks)	5.2	1.8	3.4	66%

Source: IDC 2018

Serverless TCO: Dev cost savings

A customer moving from an on-premises environment cited the ease of deploying compute resources to support development

> "We deploy new software with the AWS Serverless Platform, but we don't have to manage compute deployment because it's automatic for us . . . We've gone from using internal tools and needing two days to deploy compute to it taking about two minutes with the AWS Serverless Platform."

Source: IDC 2018

Module 2
Lambda pricing



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Lambda pricing

Requests	$0.20 per 1M requests
Duration	$0.0000166667 for every GB-second

Module 3
Provisioned concurrency



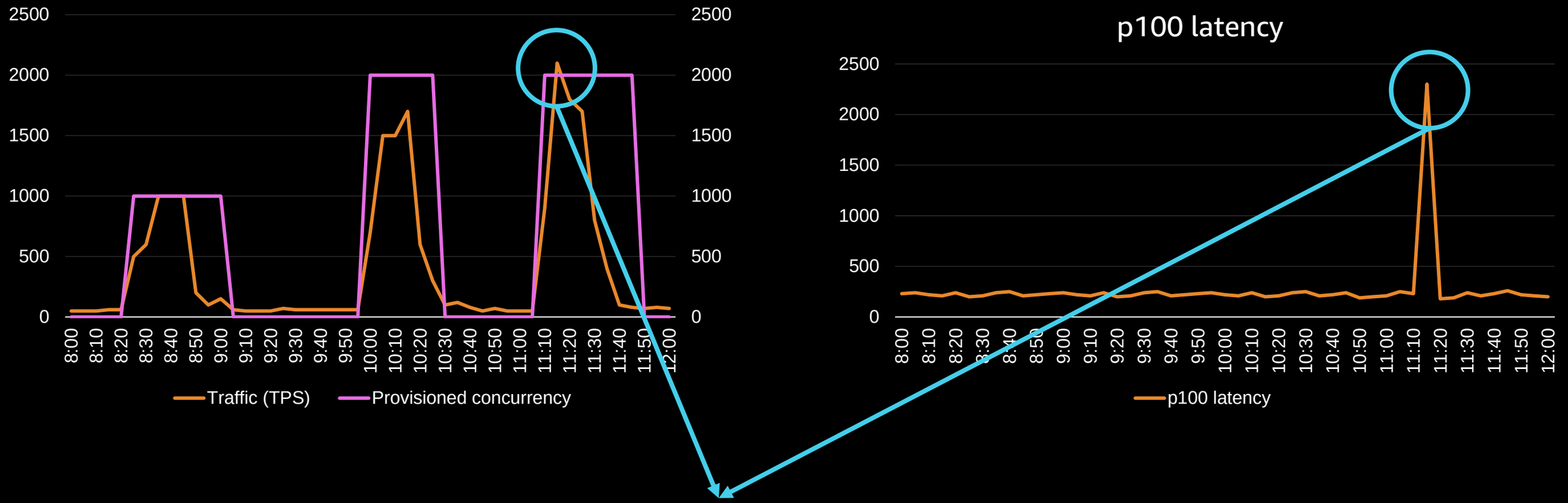
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Provisioned concurrency: Why?

1. Applications that have strict **latency SLAs**
2. Applications that support **high-velocity traffic bursts**

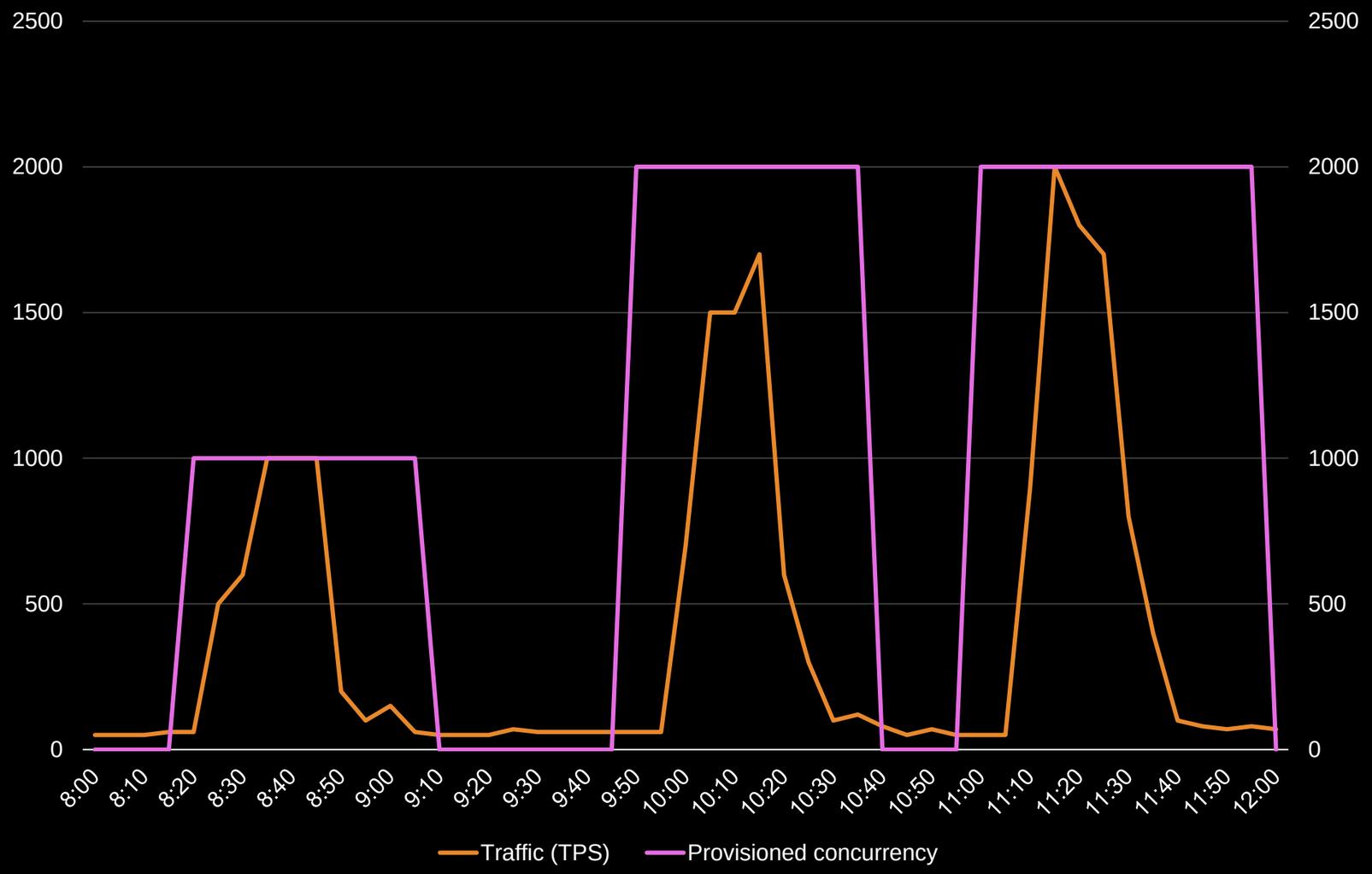
Provisioned concurrency: Why?



P100 latency spike for a regular cold start when we exceeded the provisioned concurrency

This behavior is **configurable**

Provisioned concurrency: How?



Provisioning **scheduled** in advance

Schedule in advance to allow for **scale-up**

Provision in increments of **5 minutes**

Invocations above the provisioned concurrency are handled by **on-demand** Lambda

Provisioned concurrency pricing

- Provisioned concurrency: **$0.015 for per GB-hour** (5-minute roundup)
- Duration: **$0.035 per GB-hour**
- To compare, on-demand Lambda costs **$0.060 per GB-hour**
- Requests: **$0.20 for 1M requests**

Module 4
Compute Savings Plans



AWS re:Invent

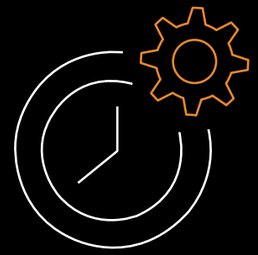
© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What are Savings Plans?



What are they?

A new flexible pricing model that helps you save up to 72% on Amazon EC2 and AWS Fargate usage and up to 17% on Lambda usage. You simply commit to a consistent amount of usage (e.g., $10/hour) over 1 or 3 years, and in exchange, you receive a discount for that usage.



How do they work?

Every type of compute usage has an On-Demand rate and a Savings Plans rate. If you commit to $10/hour of compute usage, then you get Savings Plans prices on all usage up to $10. Any usage beyond the commitment will be charged at regular On-Demand rates.

Types of Savings Plans



Compute Savings Plans

Offer the greatest flexibility, up to 66% discounts (same discounts 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 among EC2, Fargate, and Lambda



EC2 instance Savings Plans

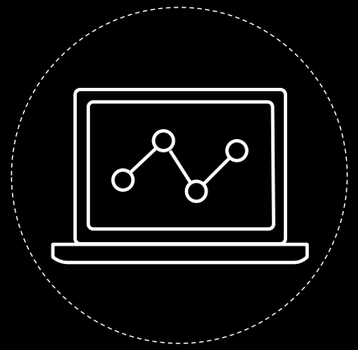
Provide the deepest discounts, up to 72% (same as Standard RIs) on the selected instance family (e.g., C5 or M5), in a specific AWS Region

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

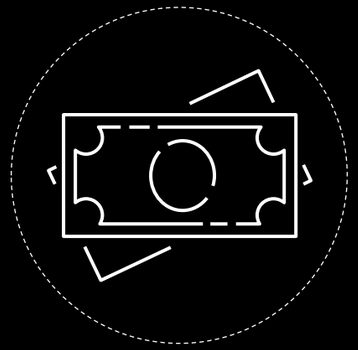
Why Savings Plans?

Flexible purchase option that offers up to 72% discounts on Amazon EC2, AWS Fargate, and AWS Lambda usage



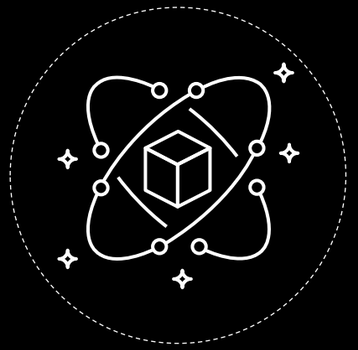
Easy to use

Receive discounted rates automatically in exchange for a monetary commitment



Significant discounts

Select from two types of savings plans to receive discounts of up to 66% on compute Savings Plans and 72% on EC2 instance plans



Flexible

Make a single commitment that applies across multiple AWS compute services, even as your requirements change

Same great discounts as EC2 RIs with more flexibility

Getting started with Savings Plans

AWS Cost Explorer will guide you through the purchasing process



Review your Savings Plans recommendations in the **AWS Cost Explorer**



Customize recommendations based on your needs (type of Savings Plan, payment option, term length)



Review hourly commitment (e.g., $10/hr) and add to cart



Eligible Amazon EC2 and AWS Fargate usage is charged at a discounted Savings Plans rate up to your commitment level

Just like RIs, you can purchase Savings Plans via the RI Operations Team as well

How do Savings Plans apply?

If you have purchased a compute Savings Plan with a $100 per hour commitment and below is what your usage looks like for Hour 1

	Billed amount by On-Demand rates	CSP discount (descending)	Billed amount by CSP discounted rate	Cumulative
Amazon EC2	$50	30%	$35	$35.0
AWS Fargate	$30	25%	$23	$57.5
AWS Lambda				
Duration	$20	15%	$17	$74.5
Provisioned concurrency	$20	15%	$17	$91.5
Provisioned concurrency duration	$10	15%	$9	$100.0
Request	$5	0%	$5	

How do Savings Plans apply?

If the usage for Hour 2 looks like this . . .

	Billed amount by On-Demand rates	CSP discount (descending)	Billed amount by CSP discounted rate	Cumulative
Amazon EC2	$50.0	30%	$35.0	$35.0
AWS Fargate	$30.0	25%	$22.5	$57.5
AWS Lambda				
Duration	$20.0	15%	$17.0	$74.5
Provisioned concurrency	$20.0	15%	$17.0	$91.5
Provisioned concurrency duration	$8.8	15%	$7.5	$99.0
Request	$5.0	0%	$5.0	

Module 5
What's next?



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

1 ms billing launch

- AWS Lambda reduced the billing granularity for Lambda function duration from 100 ms down to 1 ms
- This will lower the price for most Lambda functions, more so for short duration functions
- Their compute duration will be billed in 1 ms increments instead of being rounded up to the nearest 100 ms increment per invoke
- For example, a function that runs in 30 ms on average used to be billed for 100 ms. Now, it will be billed for 30 ms resulting in a 70% drop in its duration spend

The complete TCO picture



Infrastructure



Maintenance



Developer

How to think about serverless TCO and cost

"Our engineering team releases a new image version for servers on average 2 times per month. A typical deployment for applying patches allocates 1 hour per stage/region pair. Assuming we apply patches 2 times per month to all 6 of our stage/region pairs (Beta, Gamma, Prod) we should expect **12** developer hours per month of effort. At **$70** per developer hour, we would expect to pay **$840 per month** on host maintenance if we deployed in a server-based architecture. Since Lambda doesn't require us to maintain any hardware, we should expect a baseline savings of $840 per month in decreased operational workloads just for this one deployment – we have hundreds of similar processes."

Current customer

Serverless TCO: Maintenance cost savings

In other words . . .

When a serverless function is active for just three quarters of the month, it only takes a 10-minute savings in operational overhead for serverless to beat virtual machines on TCO

Serverless has a big impact on TCO

60%

Lower 5-year cost of operations

409%

5-year ROI

$20.7M

Higher revenue per organization

According to a study of 11 AWS customers conducted by IDC

Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.