UNIVERSITY OF ŽILINA
Faculty of Management Science
and Informatics

Presentation 7 – AWS M9 & AWS M10

AWS M9 - Cloud Architecture

AWS M10 - Automatic Scaling and Monitoring

aws academy

Vytvorené v rámci projektu **KEGA 026TUKE-4/2021**

Outline

- AWS Well-Architected Framework
- Reliability and high availability
- AWS Trusted Advisor

- Elastic Load Balancing
- Amazon CloudWatch
- Amazon EC2 Auto Scaling

Well-Architected Framework

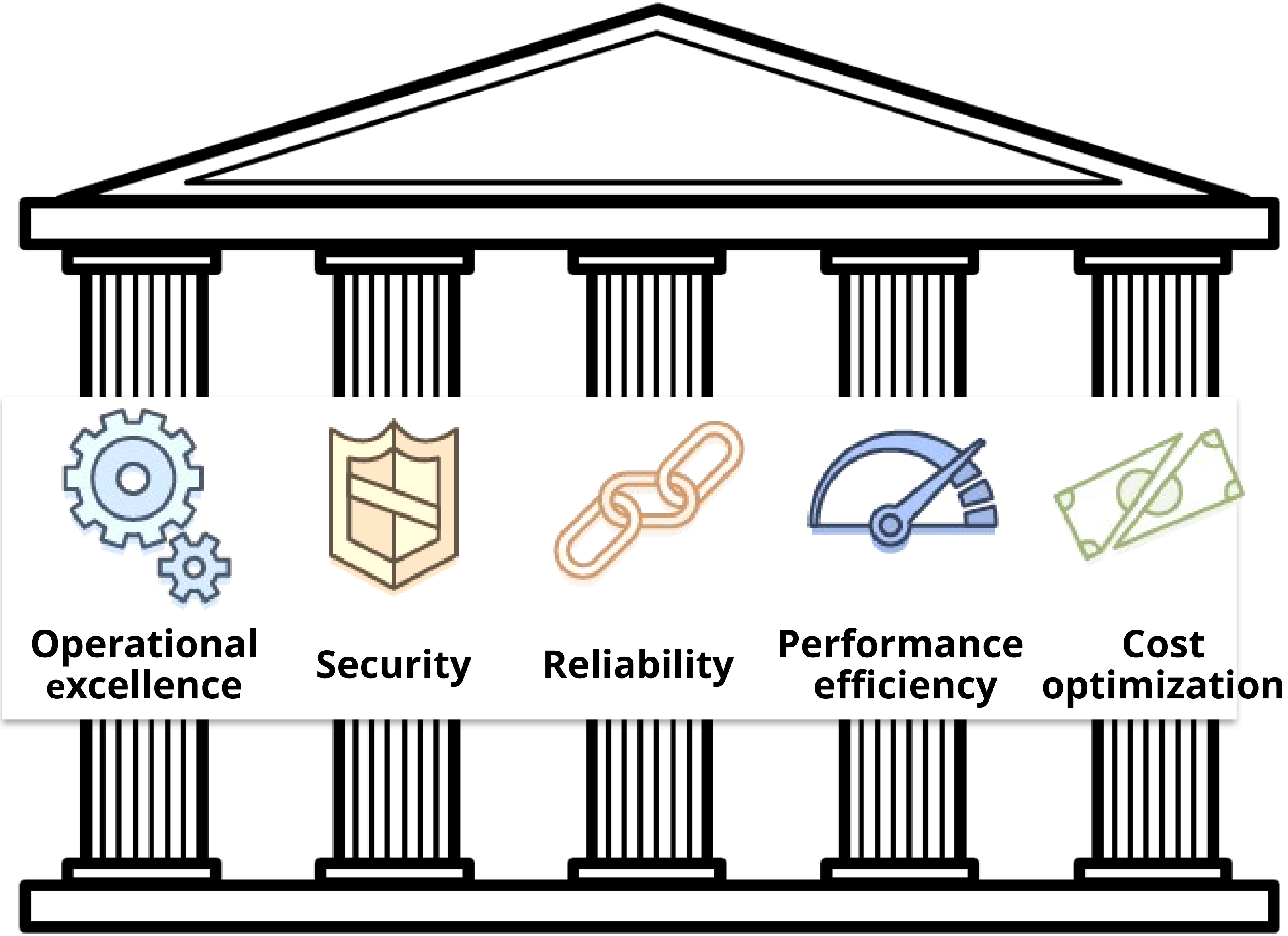
What is the AWS Well-Architected Framework?

- A guide for designing infrastructures that are:
  - Secure
  - High-performing
  - Resilient
  - Efficient

- A consistent approach to evaluating and implementing cloud architectures

- A way to provide best practices that were developed through lessons learned by reviewing customer architectures

Pillars of the AWS Well-Architected Framework



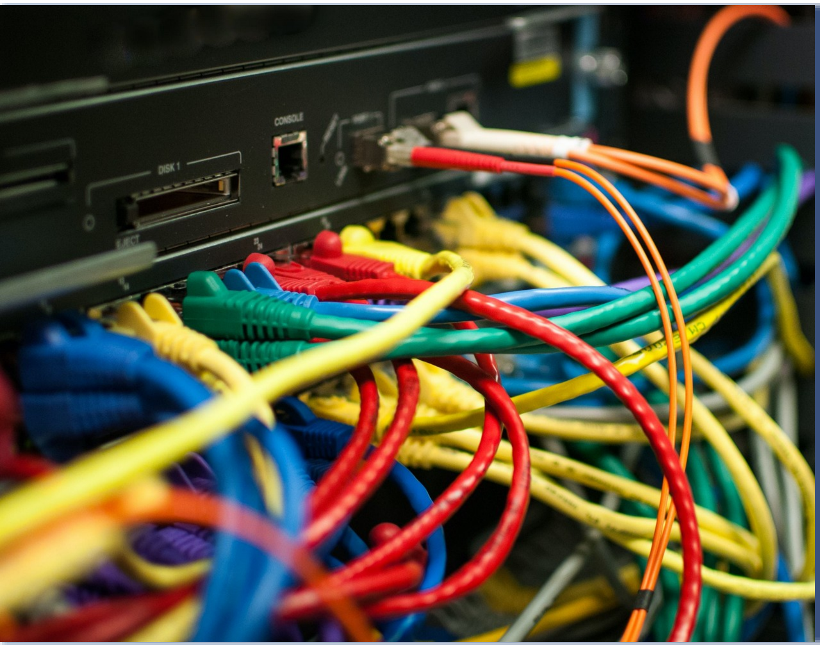
Operational excellence

Security

Reliability

Performance efficiency

Cost optimization



Operational Excellence pillar

Operational Excellence pillar

- Focus
  - Run and monitor systems to deliver business value, and to continually improve supporting processes and procedures.

- Key topics
  - Automating changes
  - Responding to events
  - Defining standards to manage daily operations

Operational Excellence pillar

Deliver business value

Operational excellence design principles

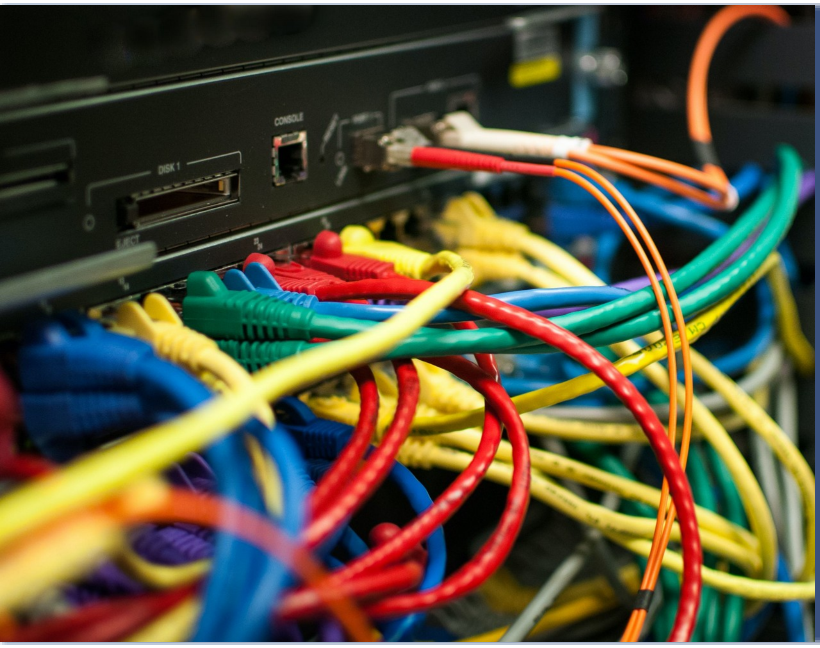
- Perform/define operations as code
- Make frequent, small, reversible changes
- Refine operations procedures frequently
- Anticipate failure
- Learn from all operational events and failures

Operational Excellence pillar

Deliver business value

Operational excellence questions

- Organization
  - How do you determine what your priorities are?
  - How do you structure your organization to support your business outcomes?
  - How does your organizational culture support your business outcomes?
- Prepare
  - How do you design your workload so that you can understand its state?
  - How do you reduce defects, ease remediation, and improve flow into production?
  - How do you mitigate deployment risks?
  - How do you know that you are ready to support a workload?
- Operate
  - How do you understand the health of your workload?
  - How do you understand the health of your operations?
  - How do you manage workload and operations events?
- Evolve
  - How do you evolve operations?



Security pillar

Security pillar

- Focus
  - Protect information, systems, and assets while delivering business value through risk assessments and mitigation strategies.

- Key topics
  - Protecting confidentiality and integrity of data
  - Identifying and managing who can do what
  - Protecting systems
  - Establishing controls to detect security events

Security pillar

Protect and monitor systems

Security pillar

- Implement a strong identity foundation
- Enable traceability
- Apply security at all layers
- Automate security best practices
- Protect data in transit and at rest
- Keep people away from data
- Prepare for security events

Security pillar

Protect and monitor systems

Security questions

- Security
  - How do you securely operate your workload?

- Identity and access management
  - How do you manage identities for people and machines?
  - How do you manage permissions for people and machines?

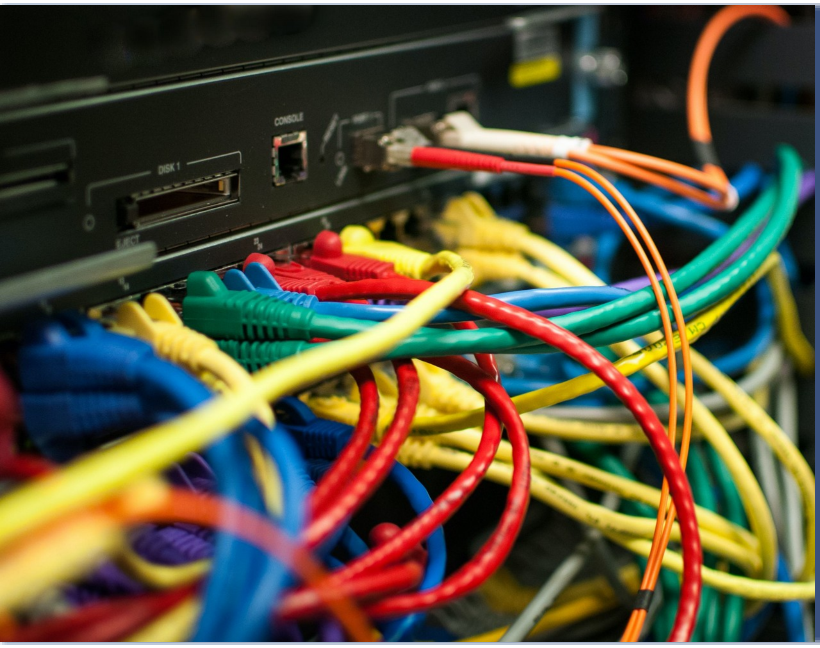
- Detection
  - How do you detect and investigate security events?

Security questions (2)

- Infrastructure protection
  - How do you protect your network resources?
  - How do you protect your compute resources?

- Data protection
  - How do you classify your data?
  - How do you protect your data at rest?
  - How do you protect your data in transit?

- Incident response
  - How do you anticipate, respond to, and recover from incidents?



Reliability pillar

Reliability pillar

- Focus
  - Ensure a workload performs its intended function correctly and consistently when it’s expected to.
- Key topics
  - Designing distributed systems
  - Recovery planning
  - Handling change

Reliability pillar

Recover from failure and mitigate disruption.

Reliability pillar

- Automatically recover from failure
- Test recovery procedures
- Scale horizontally to increase aggregate workload availability
- Stop guessing capacity
- Manage change in automation

Reliability pillar

Recover from failure and mitigate disruption.

Reliability questions

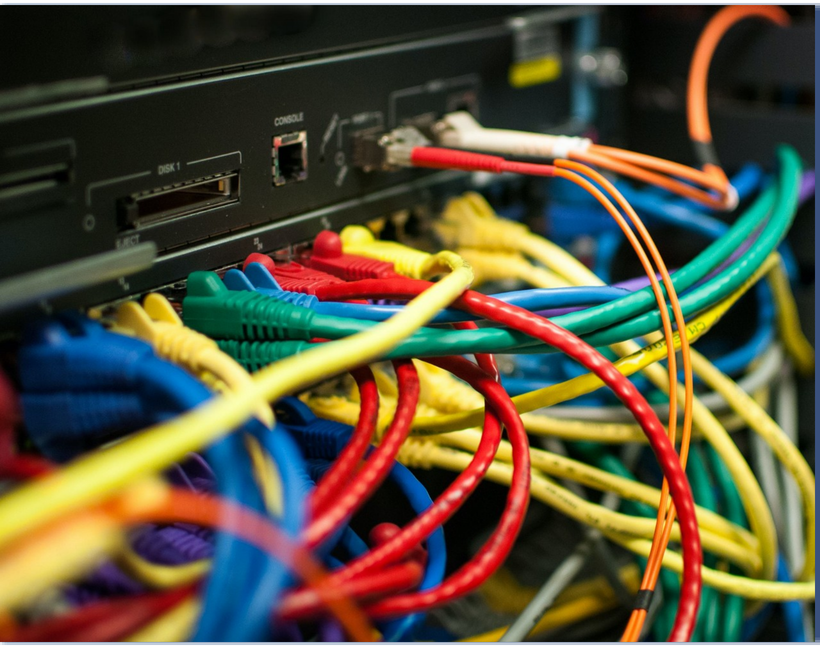
- Foundations
  - How do you manage service quotas and constraints?
  - How do you plan your network topology?

- Workload architecture
  - How do you design your workload service architecture?
  - How do you design interactions in a distributed system to prevent failure?
  - How do you design interactions in a distributed system to mitigate or withstand failures?

Reliability questions (2)

- Change management
  - How do you monitor workload resources?
  - How do you design your workload to adapt to changes in demand?
  - How do you implement change?

- Failure management
  - How do you back up data?
  - How do you use fault isolation to protect your workload?
  - How do you design your workload to withstand component failures?
  - How do you test reliability?
  - How do you plan for disaster recovery?



Performance Efficiency pillar

Performance Efficiency pillar

- Focus
  - Use IT and computing resources efficiently to meet system requirements and to maintain that efficiency as demand changes and technologies evolve.

- Key topics
  - Selecting the right resource types and sizes based on workload requirements
  - Monitoring performance
  - Making informed decisions to maintain efficiency as business needs evolve

Performance Efficiency pillar

Use resources sparingly.

Performance Efficiency pillar

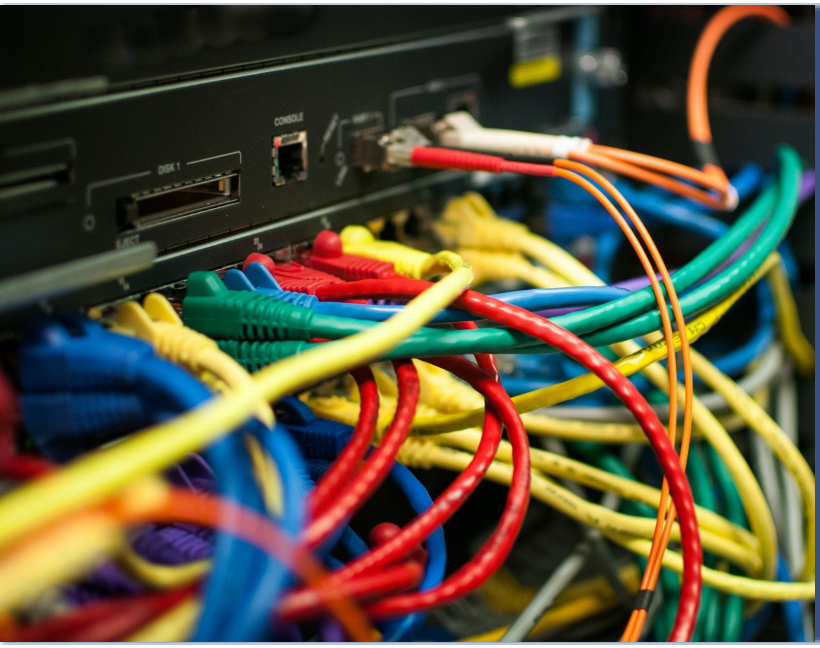
- Democratize advanced technologies
- Go global in minutes
- Use serverless architectures
- Experiment more often
- Consider mechanical sympathy

Performance Efficiency pillar

Use resources sparingly.

Performance efficiency questions

- Selection
  - How do you select the best performing architecture?
  - How do you select your compute solution?
  - How do you select your storage solution?
  - How do you select your database solution?
  - How do you configure your networking solution?
- Review
  - How do you evolve your workload to take advantage of new releases?
- Monitoring
  - How do you monitor your resources to ensure they are performing?
- Tradeoffs
  - How do you use tradeoffs to improve performance?



Cost Optimization pillar

Cost Optimization pillar

- Focus
  - Avoid unnecessary costs.

- Key topics
  - Understanding and controlling where money is being spent
  - Selecting the most appropriate and right number of resource types
  - Analyzing spend over time
  - Scaling to meeting business needs without overspending

Cost Optimization pillar

Eliminate unneeded expense.

Cost Optimization pillar

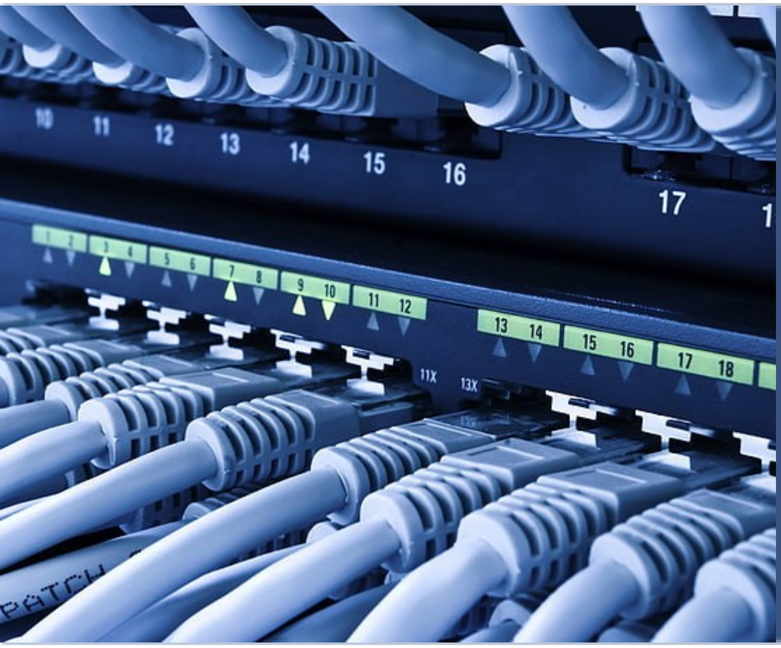
- Implement Cloud Financial Management
- Adopt a consumption model
- Measure overall efficiency
- Stop spending money on undifferentiated heavy lifting
- Analyze and attribute expenditure

Cost Optimization pillar

Eliminate unneeded expense.

Cost optimization questions

- Practice cloud financial management
  - How do you implement cloud financial management?
- Expenditure and usage awareness
  - How do you govern usage?
  - How do you monitor usage and cost?
  - How do you decommission resources?
- Cost-effective resources
  - How do you evaluate cost when you select services?
  - How do you meet cost targets when you select resource type, size, and number?
  - How do you use pricing models to reduce cost?
  - How do you plan for data transfer changes?
- Manage demand and supply resources
  - How do you manage demand and supply resources?
- Optimize over time
  - How do you evaluate new services?



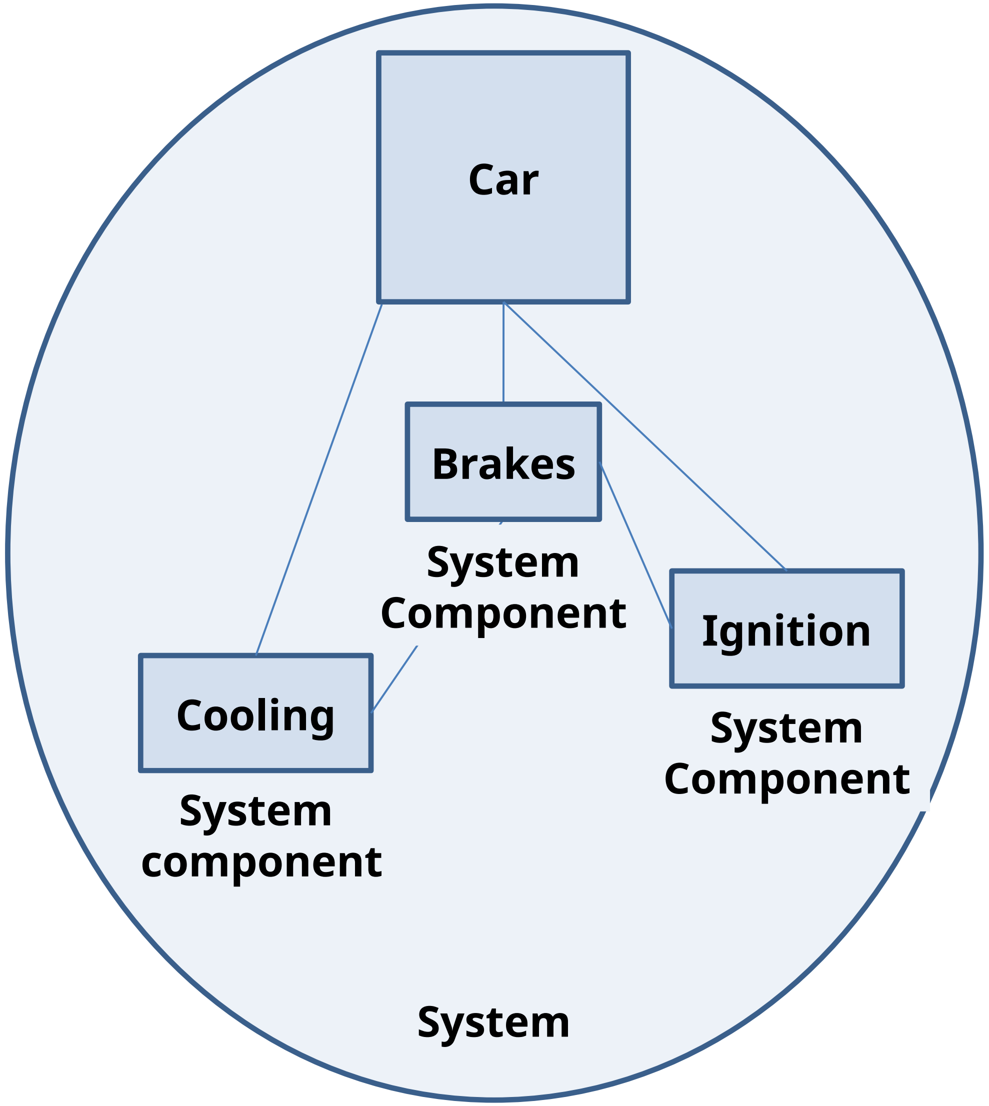
Reliability and availability

“Everything fails, all the time.”

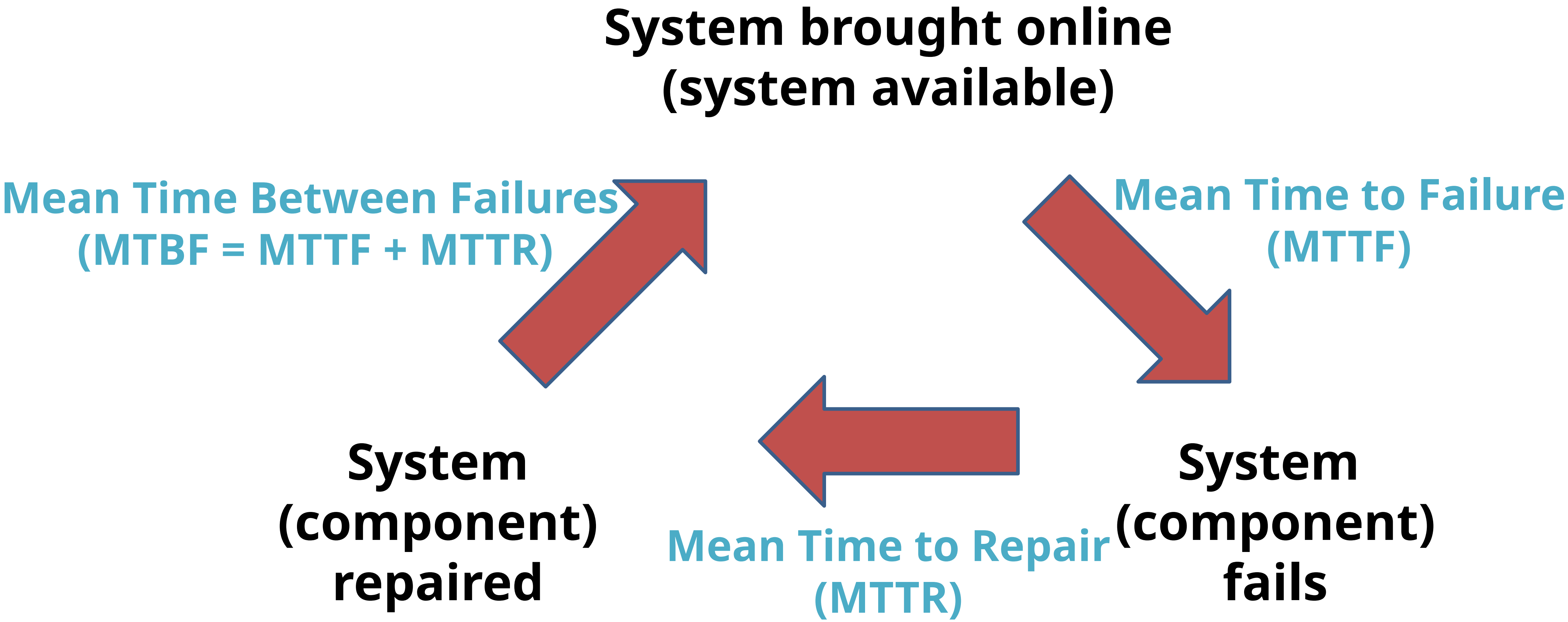
Werner Vogels, CTO, Amazon.com

Reliability

- A measure of your system’s ability to provide functionality when desired by the user.
- System includes all system components: hardware, firmware, and software.
- Probability that your entire system will function as intended for a specified period.
- Mean time between failures (MTBF) = total time in service/number of failures



Understanding reliability metrics



System brought online
(system available)

Mean Time to Failure
(MTTF)

System
(component)
fails

Mean Time to Repair
(MTTR)

System
(component)
repaired

Mean Time Between Failures
(MTBF = MTTF + MTTR)

Availability and High Availability (HA)

- Availavility
  - Normal operation time / total time
  - A percentage of uptime (for example, 99.9 percent) over time (for example 1 year)
  - Number of 9s – Five 9s means 99.999 percent availability

- High Availability
  - System can withstand some measure of degradation while still remaining available.
  - Downtime is minimized.
  - Minimal human intervention is required.

Availability tiers

Availability	Downtime/Year	Downtime/Month	Downtime/Week	Downtime/Day
90%	36,53 days	73,05 hours	16,8 hours	2,4 hours
99%	87,6 hours	7,3 hours	101,077 minutes	14,4 Minutes
99,9%	8,76 hours	43,8 minutes	10,108 minutes	1,44 minutes
99,99%	52,56 minutes	4,38 minutes	1,011 minutes	8,64 seconds
99,999%	5,256 minutes	26,28 seconds	6,06 seconds	0,84 seconds
99,9999%	31,56 seconds	2,63 seconds	6,4 miliseconds	86,4 miliseconds
99,99999%	3,16 seconds	262,98 miliseconds	60,48 miliseconds	8,64 miliseconds
99,999999%	315,58 miliseconds	26,3 miliseconds	60,5 miliseconds	864 µseconds
99,9999999%	31,56 miliseconds	2,63 miliseconds	604,8 µseconds	86 µseconds

Factors that influence availability

- Fault tolerance
  - The built-in redundancy of an application's components and its ability to remain operational.
- Scalability
  - The ability of an application to accommodate increases in capacity needs without changing design.
- Recoverability
  - The process, policies, and procedures that are related to restoring service after a catastrophic event.

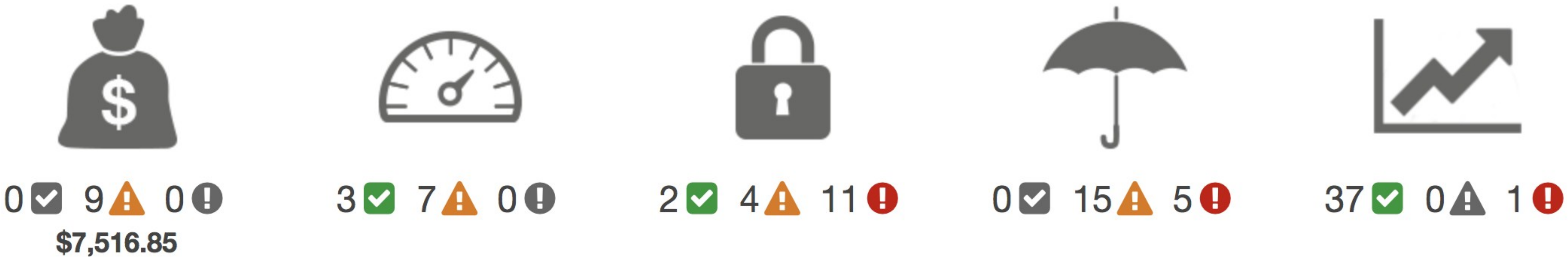
Ak chcete pridať obrázok, kliknite na ikonu

AWS Trusted Advisor

AWS Trusted Advisor

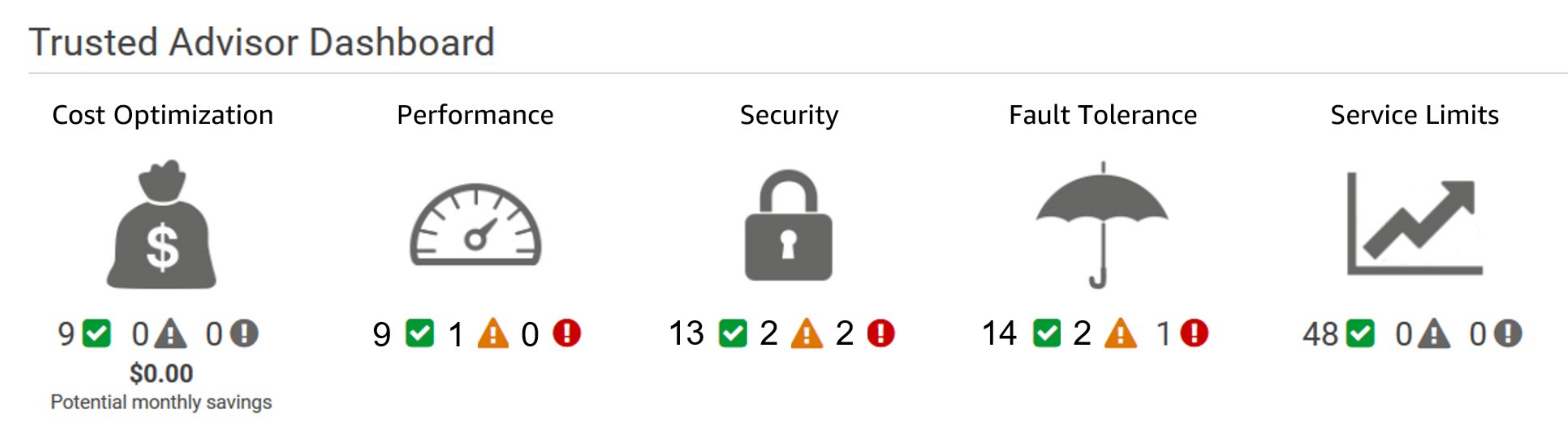
- Online tool that provides real-time guidance to help you provision your resources following AWS best practices.
- Looks at your entire AWS environment and gives you real-time recommendations in five categories.

Cost Optimization Performance Security Fault Tolerance Service Limits



Potential monthly savings

Activity: Interpret AWS Trusted Advisor recommendations



Trusted Advisor Dashboard

Cost Optimization	Performance	Security	Fault Tolerance	Service Limits
9 ✓ 0 ⚠ 0 ❗	9 ✓ 1 ⚠ 0 ❗	13 ✓ 2 ⚠ 2 ❗	14 ✓ 2 ⚠ 1 ❗	48 ✓ 0 ⚠ 0 ❗
$0.00 Potential monthly savings				

Activity: Recommendation #1

❗ **MFA on Root Account**

Description: Checks the root account and warns if multi-factor authentication (MFA) is not enabled. For increased security, we recommend that you protect your account by using MFA, which requires a user to enter a unique authentication code from their MFA hardware or virtual device when interacting with the AWS console and associated websites.

Alert Criteria: MFA is not enabled on the root account.

Recommended Action: Log in to your root account and activate an MFA device.

Activity: Recommendation #2

⚠ **IAM Password Policy**

Description: Checks the password policy for your account and warns when a password policy is not enabled, or if password content requirements have not been enabled. Password content requirements increase the overall security of your AWS environment by enforcing the creation of strong user passwords. When you create or change a password policy, the change is enforced immediately for new users but does not require existing users to change their passwords.

Alert Criteria: A password policy is enabled, but at least one content requirement is not enabled.

Recommended Action: If some content requirements are not enabled, consider enabling them. If no password policy is enabled, create and configure one. See Setting an Account Password Policy for IAM Users.

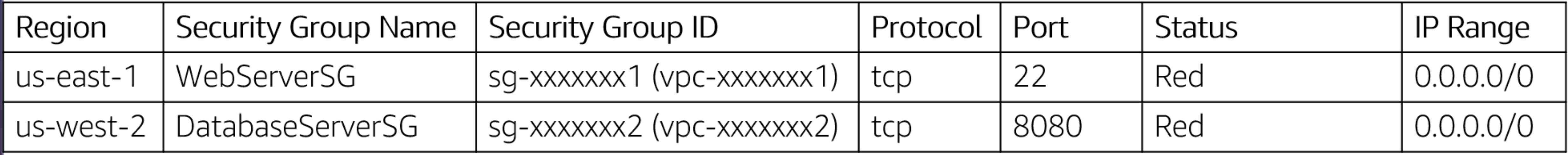
Activity: Recommendation #3

Security Groups – Unrestricted Access

Description: Checks security groups for rules that allow unrestricted access to a resource. Unrestricted access increases opportunities for malicious activity (hacking, denial-of-service attacks, loss of data).

Alert Criteria: A security group rule has a source IP address with a /0 suffix for ports other than 25, 80, or 443.)

Recommended Action: Restrict access to only those IP addresses that require it. To restrict access to a specific IP address, set the suffix to /32 (for example, 192.0.2.10/32). Be sure to delete overly permissive rules after creating rules that are more restrictive.



Region	Security Group Name	Security Group ID	Protocol	Port	Status	IP Range
us-east-1	WebServerSG	sg-xxxxxxx1 (vpc-xxxxxxx1)	tcp	22	Red	0.0.0.0/0
us-west-2	DatabaseServerSG	sg-xxxxxxx2 (vpc-xxxxxxx2)	tcp	8080	Red	0.0.0.0/0

Activity: Recommendation #4

❗ Amazon EBS Snapshots

Description: Checks the age of the snapshots for your Amazon Elastic Block Store (Amazon EBS) volumes (available or in-use). Even though Amazon EBS volumes are replicated, failures can occur. Snapshots are persisted to Amazon Simple Storage Service (Amazon S3) for durable storage and point-in-time recovery.

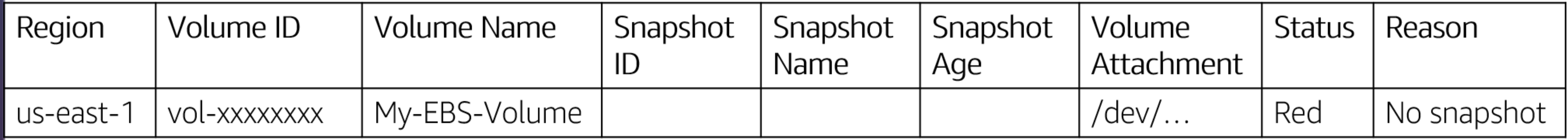
Alert Criteria:

Yellow: The most recent volume snapshot is between 7 and 30 days old.

Red: The most recent volume snapshot is more than 30 days old.

Red: The volume does not have a snapshot.

Recommended Action: Create weekly or monthly snapshots of your volumes



Region	Volume ID	Volume Name	Snapshot ID	Snapshot Name	Snapshot Age	Volume Attachment	Status	Reason
us-east-1	vol-xxxxxxxx	My-EBS-Volume				/dev/…	Red	No snapshot

Activity: Recommendation #5

⚠ **Amazon S3 Bucket Logging**

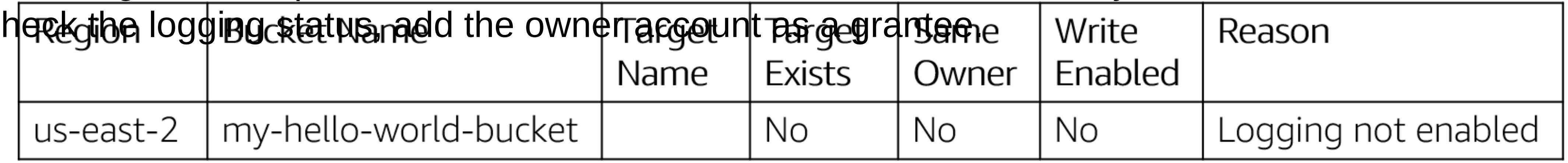
Description: Checks the logging configuration of Amazon Simple Storage Service (Amazon S3) buckets. When server access logging is enabled, detailed access logs are delivered hourly to a bucket that you choose. An access log record contains details about each request, such as the request type, the resources specified in the request, and the time and date the request was processed. By default, bucket logging is not enabled; you should enable logging if you want to perform security audits or learn more about users and usage patterns.

Alert Criteria:

Yellow: The bucket does not have server access logging enabled.
Yellow: The target bucket permissions do not include the owner account. Trusted Advisor cannot check it.

Recommended Action:
Enable bucket logging for most buckets.

If the target bucket permissions do not include the owner account and you want Trusted Advisor to check the logging status, add the owner account as a grantee.



Region	Bucket Name	Target Name	Target Exists	Same Owner	Write Enabled	Reason
us-east-2	my-hello-world-bucket		No	No	No	Logging not enabled

Sample exam question

A SysOps engineer working at a company wants to protect their data in transit and at rest. What services could they use to protect their data?

A. Elastic Load Balancing

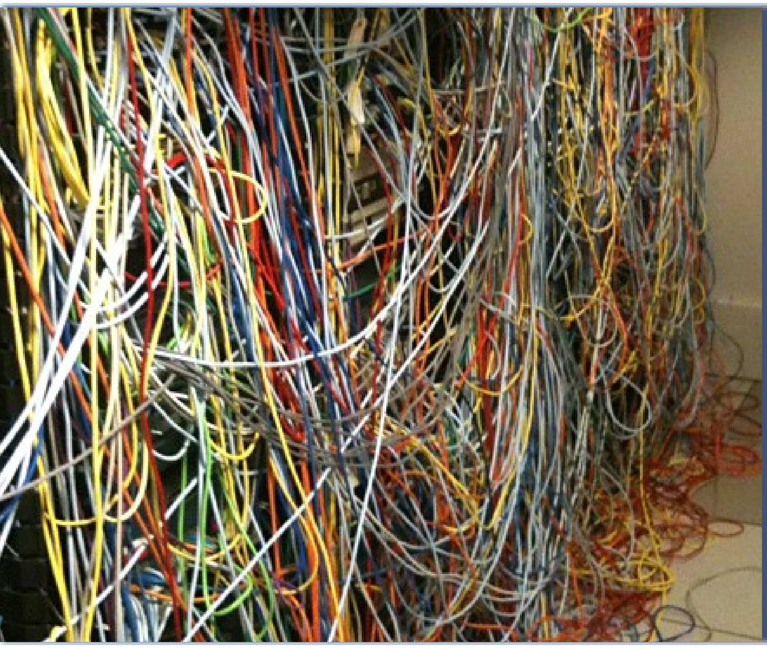
B. Amazon Elastic Block Store (Amazon EBS)

C. Amazon Simple Storage Service (Amazon S3)

[✓ D. All of the above]

Additional resoures

- AWS Well-Architected website
- AWS Well-Architected Framework whitepaper
- AWS Well-Architected Labs
- AWS Trusted Advisor Best Practice Checks



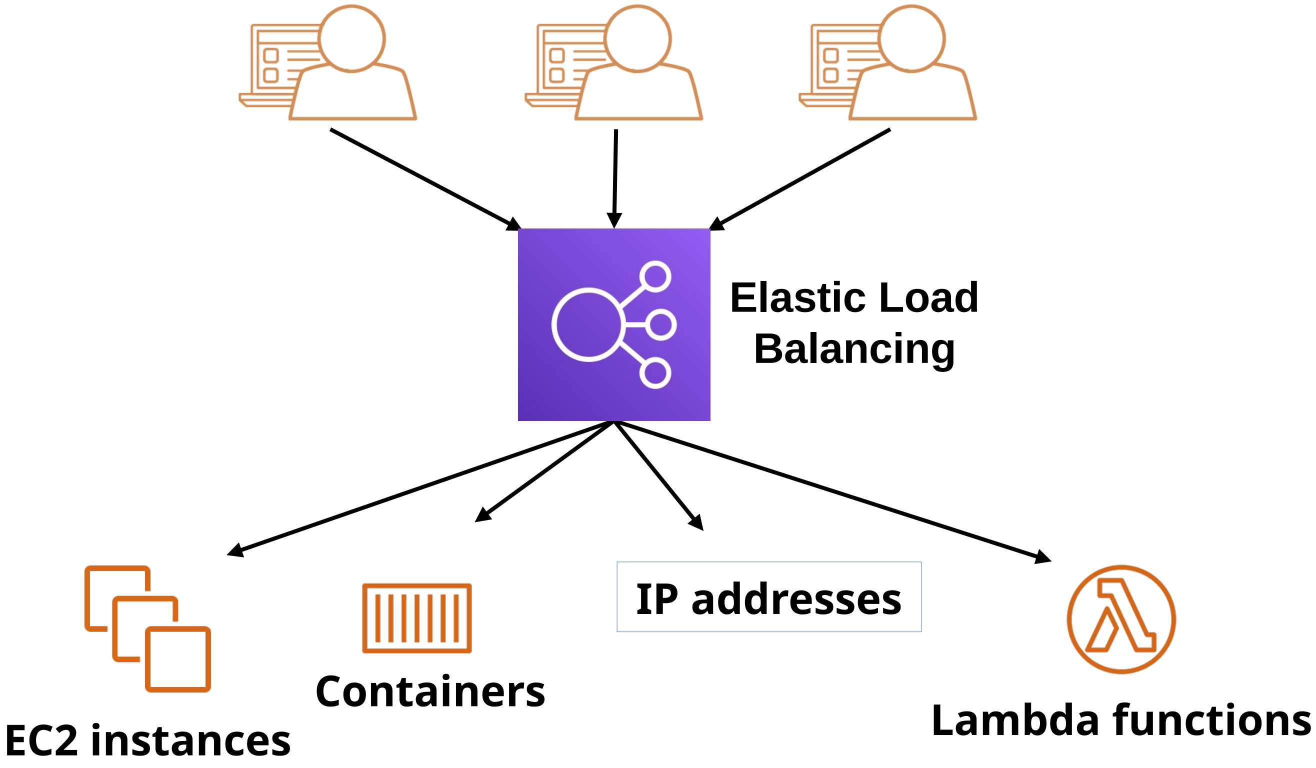
AWS M10 - Automatic Scaling and Monitoring

Elastic Load Balancing

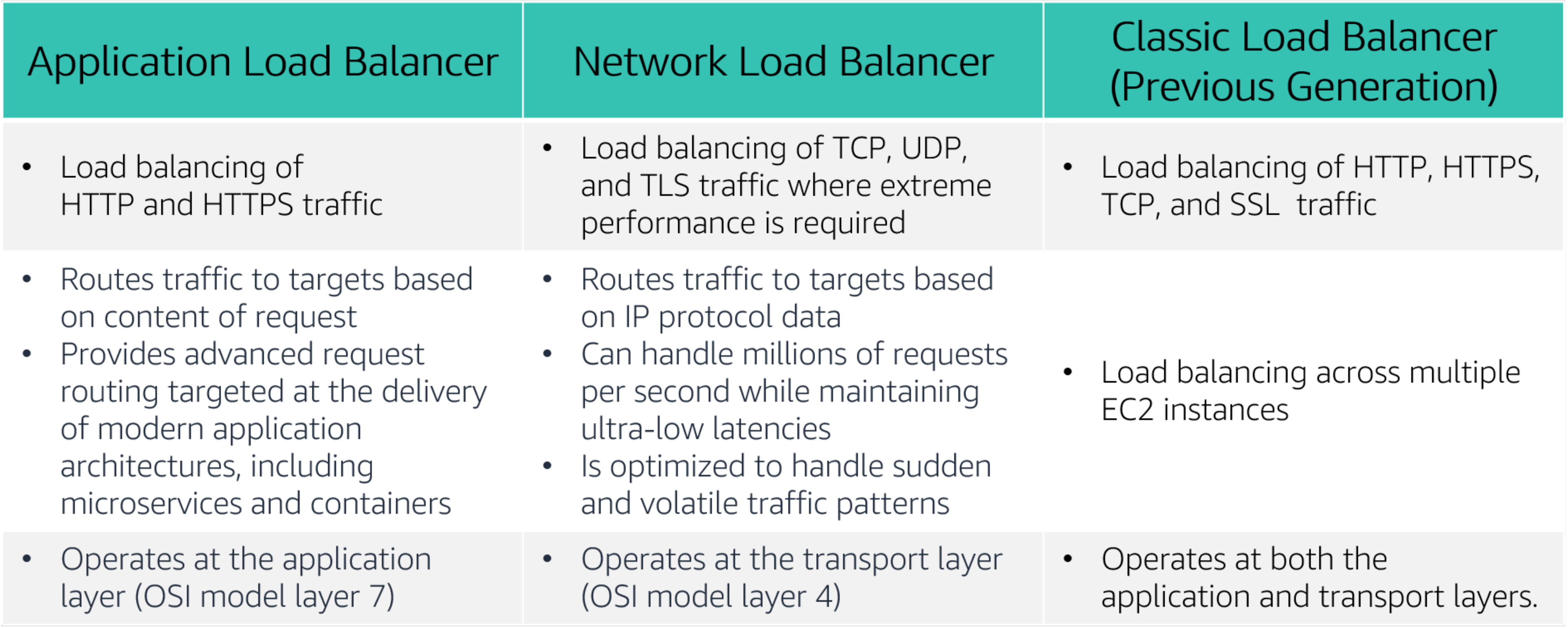
Elastic Load Balancing

- Distributes incoming application or network traffic across multiple targets in a single Availability Zone or across multiple Availability Zones.

- Scales your load balancer as traffic to your application changes over time.

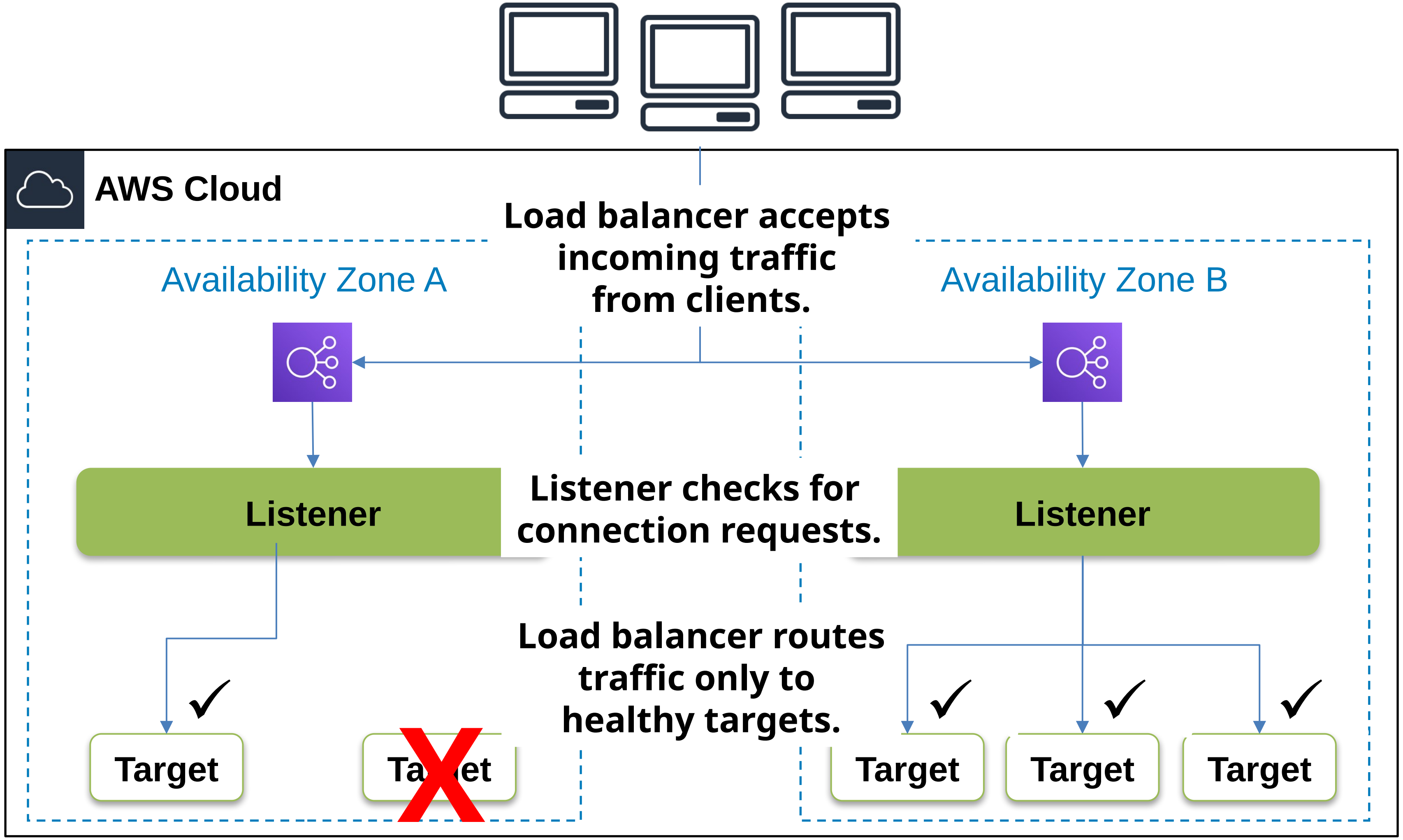


Types of load balancers



Application Load Balancer	Network Load Balancer	Classic Load Balancer (Previous Generation)
- Load balancing of HTTP and HTTPS traffic	- Load balancing of TCP, UDP, and TLS traffic where extreme performance is required	- Load balancing of HTTP, HTTPS, TCP, and SSL traffic
- Routes traffic to targets based on content of request - Provides advanced request routing targeted at the delivery of modern application architectures, including microservices and containers	- Routes traffic to targets based on IP protocol data - Can handle millions of requests per second while maintaining ultra-low latencies - Is optimized to handle sudden and volatile traffic patterns	- Load balancing across multiple EC2 instances
- Operates at the application layer (OSI model layer 7)	- Operates at the transport layer (OSI model layer 4)	- Operates at both the application and transport layers.

How Elastic Load Balancing works



Load balancer accepts incoming traffic from clients.

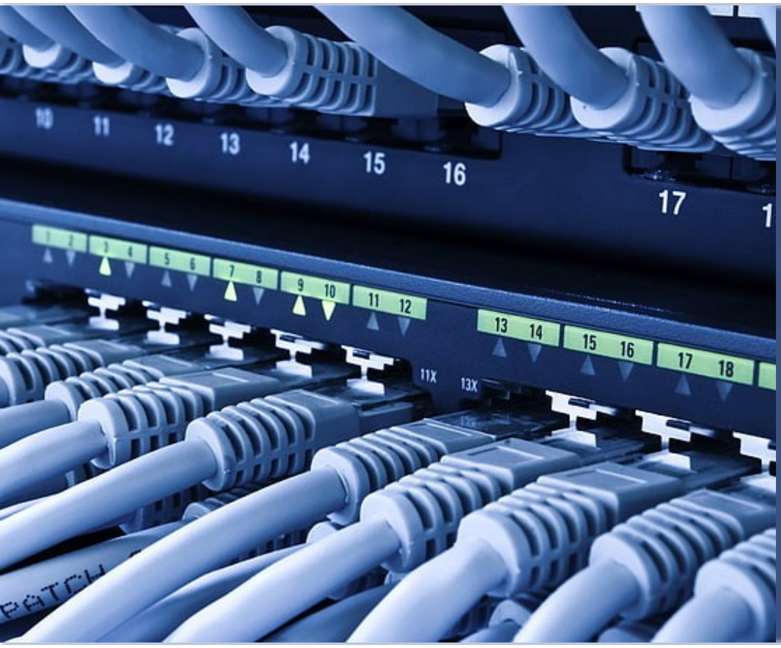
Listener checks for connection requests.

Load balancer routes traffic only to healthy targets.

Load balancer performs health checks to monitor health of registered targets.

Activity: Elastic Load Balancing

You must support traffic to a containerized application.	Application Load Balancer
You have extremely spiky and unpredictable TCP traffic.	Network Load Balancer
You need simple load balancing with multiple protocols.	Classic Load Balancer
You need to support a static or Elastic IP address, or an IP target outside a VPC.	Network Load Balancer
You need a load balancer that can handle millions of requests per second while maintaining low latencies.	Network Load Balancer
You must support HTTPS requests.	Application Load Balancer



Amazon CloudWatch

Monitoring AWS resources

To use AWS efficiently, you need insight into your AWS resources:

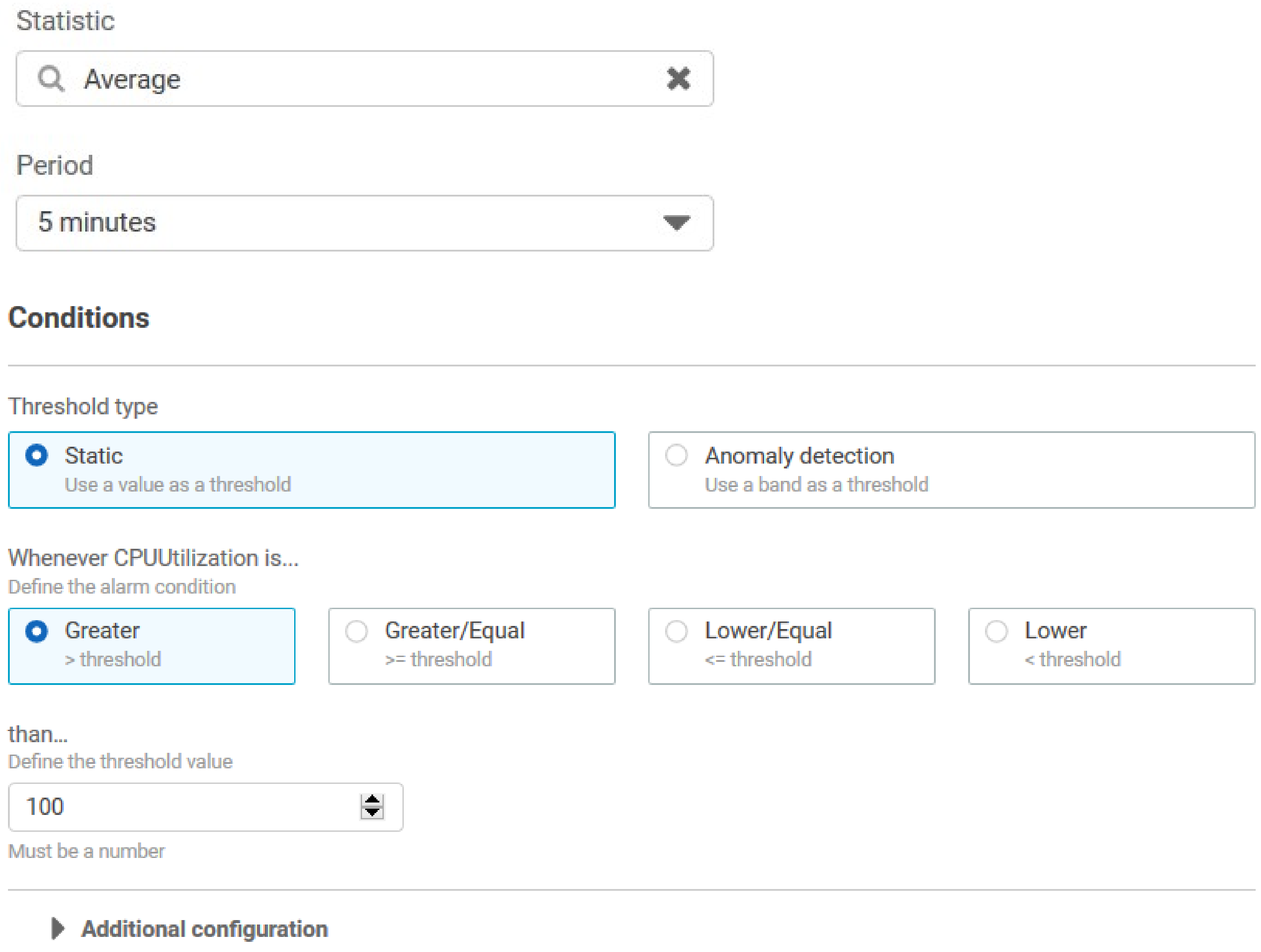
- How do you know when you should **launch more Amazon EC2 instances**?
- Is your **application's performance or availability** being affected by a lack of sufficient capacity?
- How much of your infrastructure is actually **being used**?

Amazon CloudWatch

- Monitors
  - AWS resources
  - Applications that run on AWS
- Collects and tracks
  - Standard metrics
  - Custom metrics
- Alarms
  - Send notifications to an Amazon SNS topic
  - Perform Amazon EC2 Auto Scaling or Amazon EC2 actions
- Events
  - Define rules to match changes in AWS environment and route these events to one or more target functions or streams for processing

CloudWatch alarms

- Create alarms based on
  - Static threshold
  - Anomaly detection
  - Metric math expression
- Specify
  - Namespace
  - Metric
  - Statistic
  - Period
  - Conditions
  - Additional configuration
  - Actions



Statistic

Average

Period

5 minutes

Conditions

Threshold type

● Static
Use a value as a threshold

○ Anomaly detection
Use a band as a threshold

Whenever CPUUtilization is...
Define the alarm condition

● Greater
> threshold

○ Greater/Equal
>= threshold

○ Lower/Equal
<= threshold

○ Lower
< threshold

than...
Define the threshold value

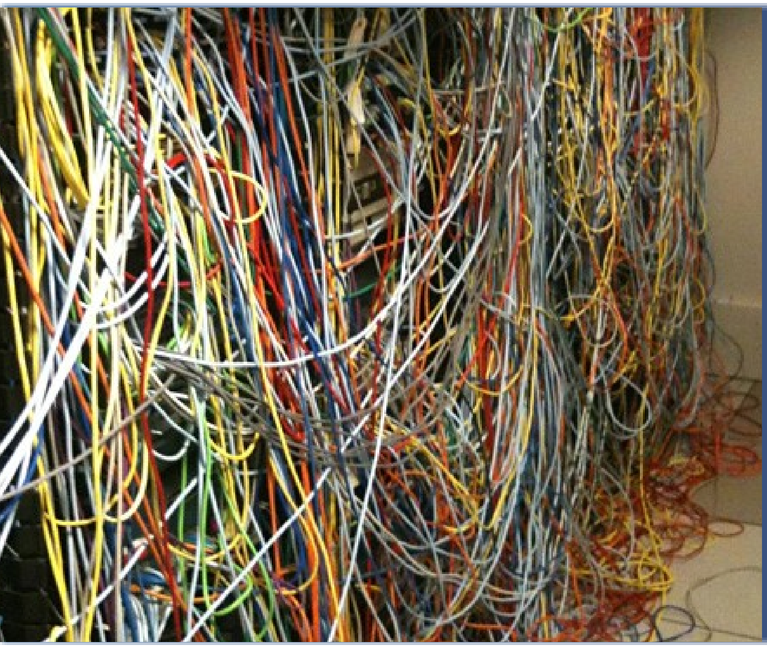
100

Must be a number

Additional configuration

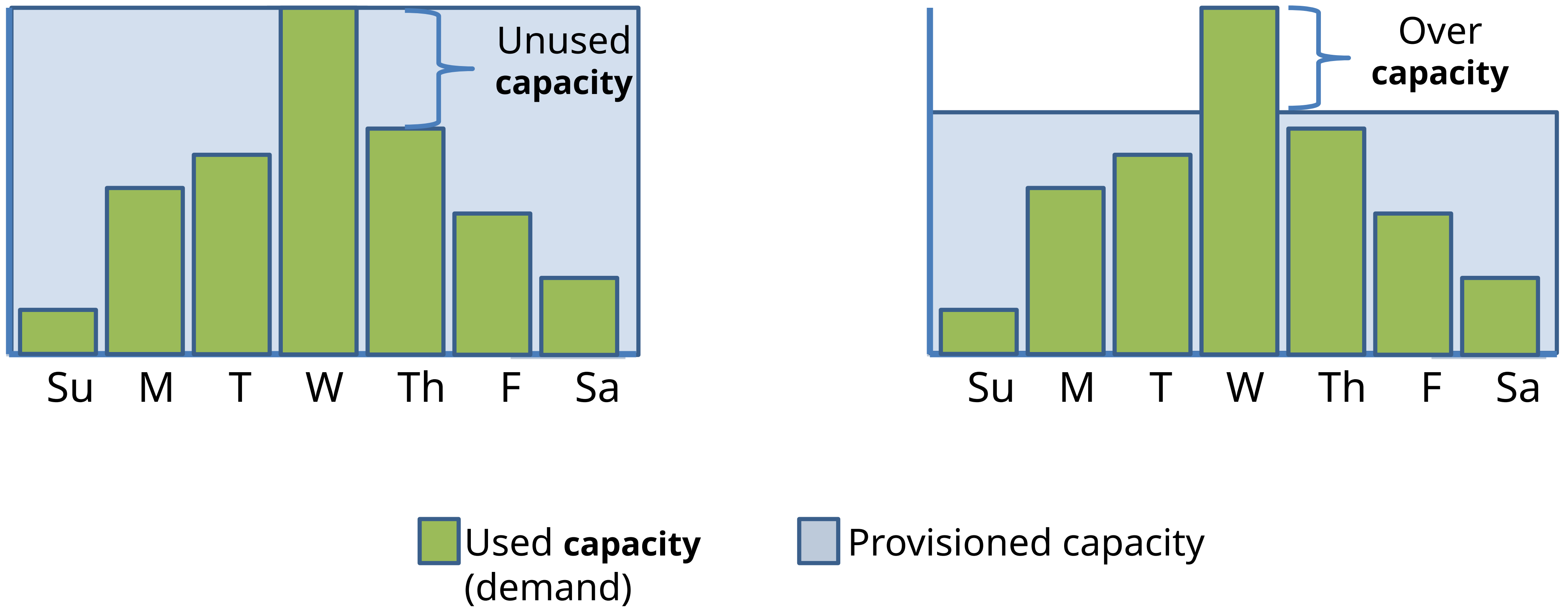
Activity: Amazon CloudWatch – What can I monitor?

Amazon EC2	If average CPU utilization is > 60% for 5 minutes…	Correct!
Amazon RDS	If the number of simultaneous connections is > 10 for 1 minute…	Correct!
Amazon S3	If the maximum bucket size in bytes is around 3 for 1 day…	Incorrect. *Around* is not a threshold option. You must specify a threshold of >, >=, <=, or <.
Elastic Load Balancing	If the number of healthy hosts is < 5 for 10 minutes…	Correct!
Amazon Elastic Block Store	If the volume of read operations is > 1,000 for 10 seconds…	Incorrect. You must specify a statistic (for example, *average volume*).

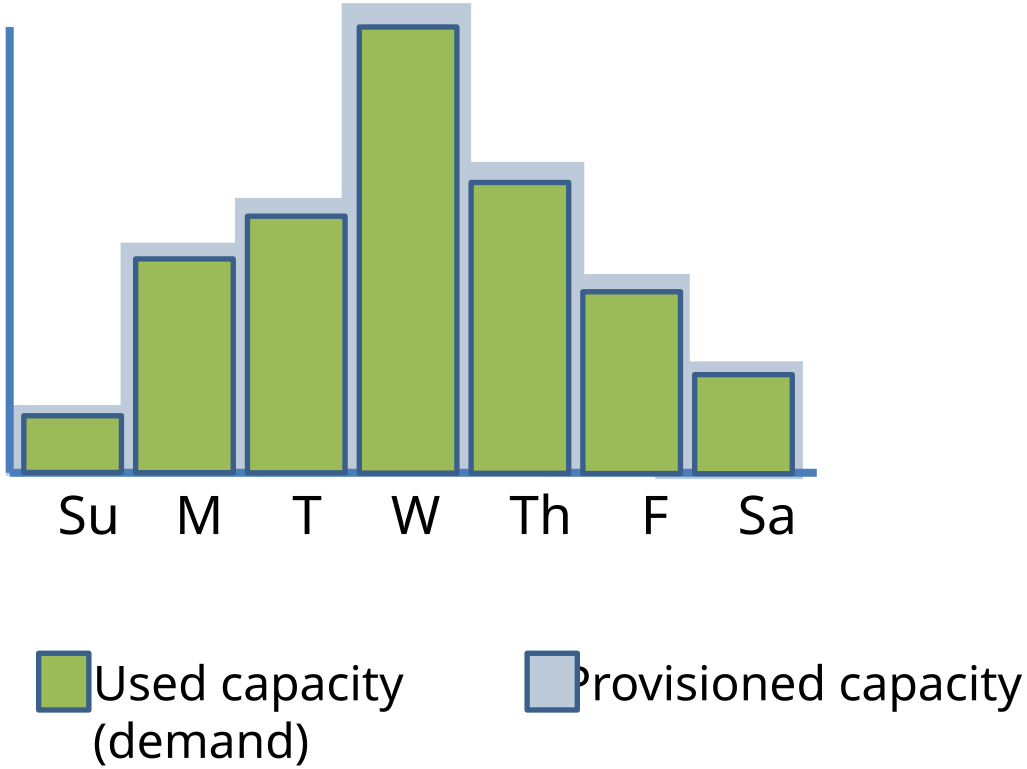


Amazon EC2 Auto Scaling

Why is scaling important?

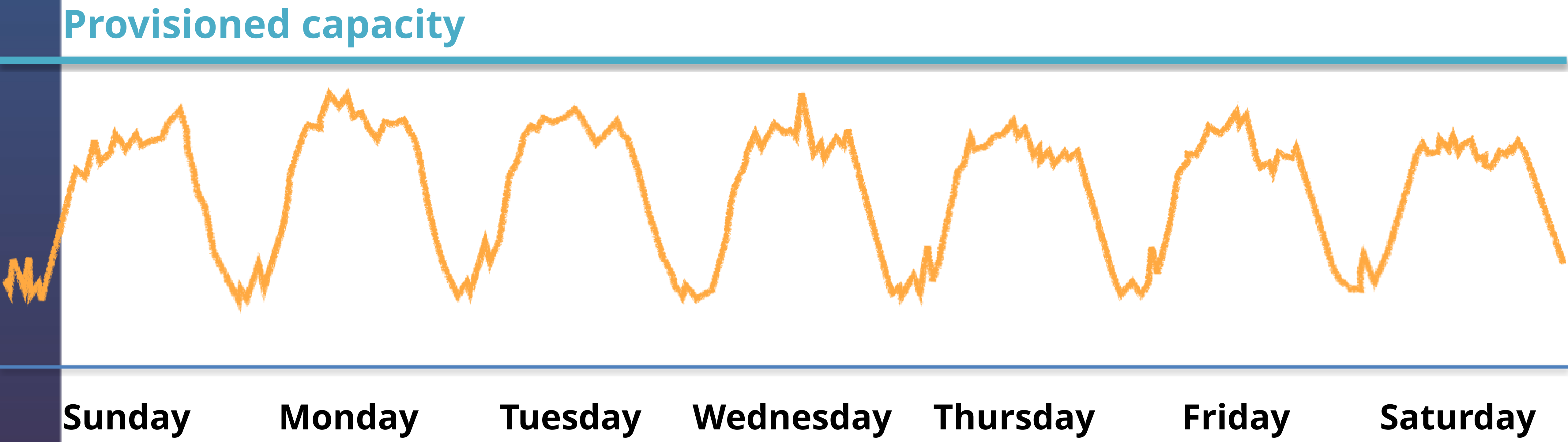


Amazon EC2 Auto Scaling

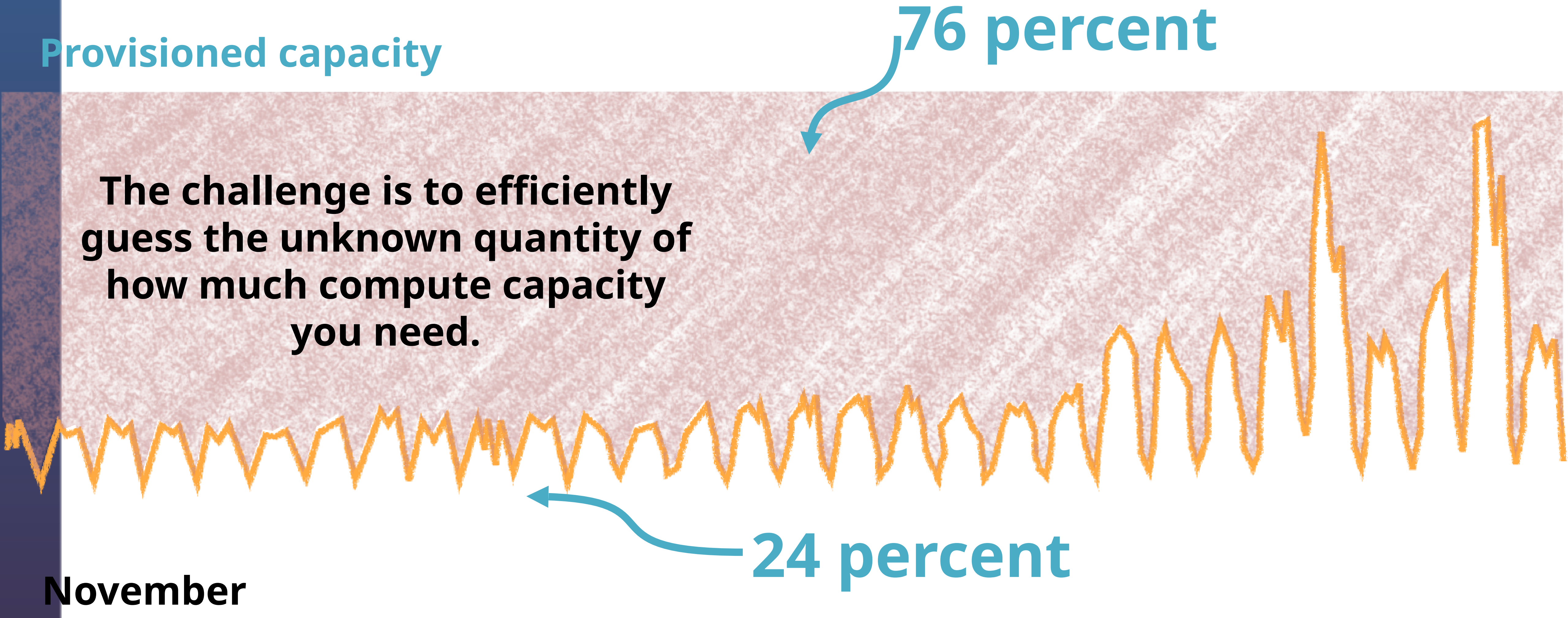


- Helps you maintain application availability
- Enables you to automatically add or remove EC2 instances according to conditions that you define
- Detects impaired EC2 instances and unhealthy applications, and replaces the instances without your intervention
- Provides several scaling options – Manual, scheduled, dynamic or on-demand, and predictive

Typical weekly traffic at Amazon.com

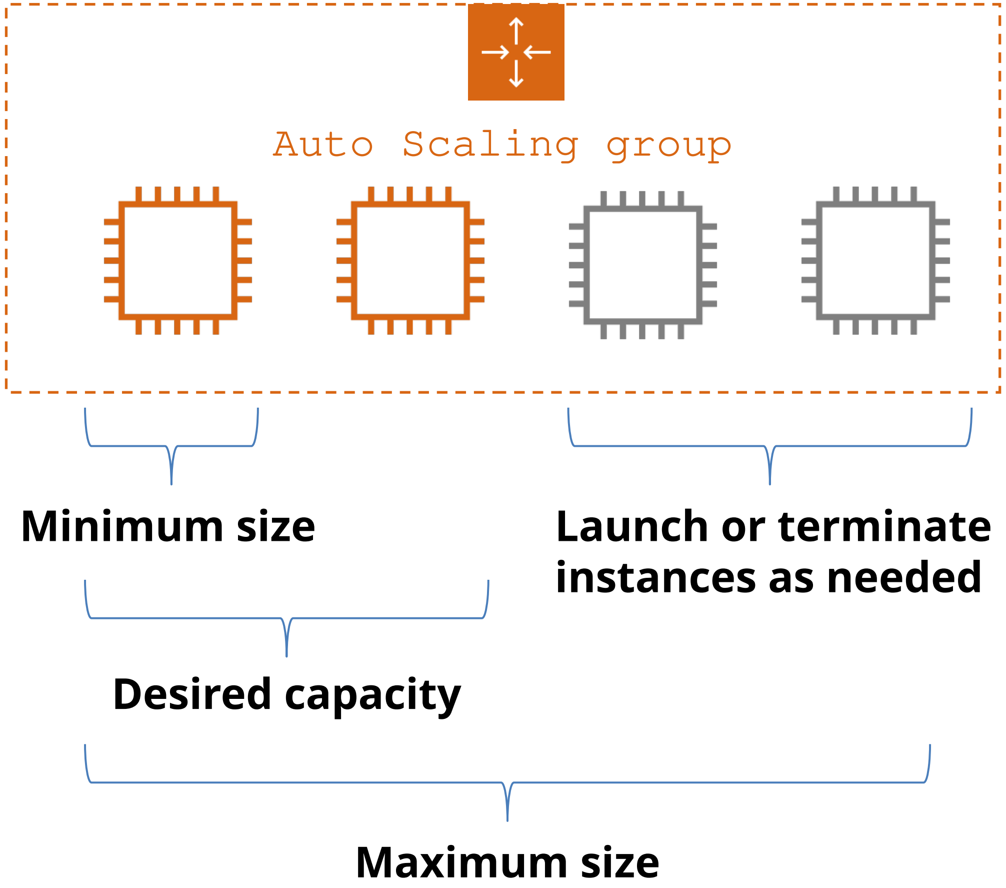


November traffic to Amazon.com



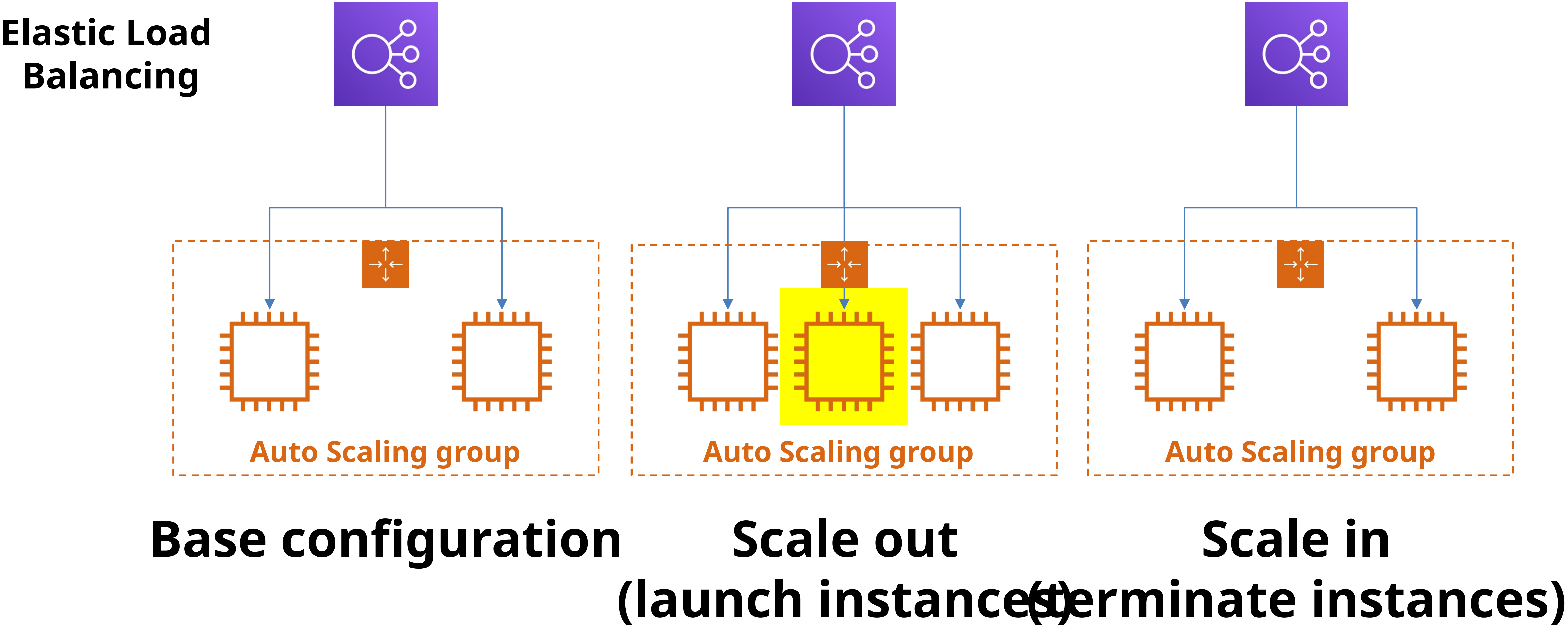
The challenge is to efficiently guess the unknown quantity of how much compute capacity you need.

Auto Scaling groups

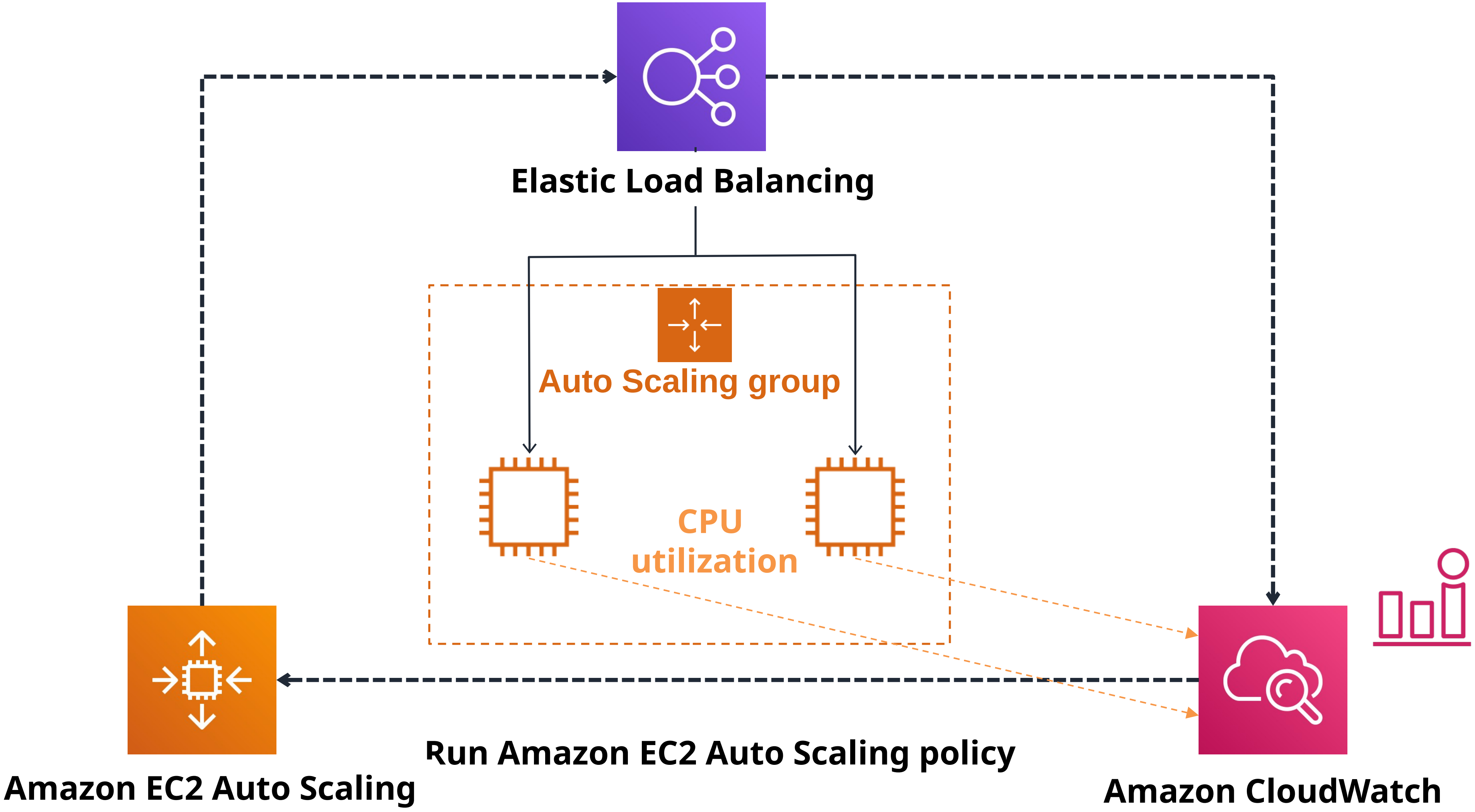


- An Auto Scaling group is a collection of EC2 instances that are treated as a logical grouping for the purposes of automatic scaling and management.

Scaling out versus scaling in



Implementing dynamic scaling



If average CPU utilization is > 60% for 5 minutes...

AWS Auto Scaling

- Monitors your applications and automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost
- Provides a simple, powerful user interface that enables you to build scaling plans for resources, including –
  - Amazon EC2 instances and Spot Fleets
  - Amazon Elastic Container Service (Amazon ECS) Tasks
  - Amazon DynamoDB tables and indexes
  - Amazon Aurora Replicas

Sample exam question

Which service would you use to send alerts based on Amazon CloudWatch alarms?

A. Amazon Simple Notification Service
B. AWS CloudTrail
C. AWS Trusted Advisor
D. Amazon Route 53

UNIVERSITY OF ŽILINA
Faculty of Management Science
and Informatics

MINISTERSTVO
ŠKOLSTVA, VEDY,
VÝSKUMU A ŠPORTU
SLOVENSKEJ REPUBLIKY

Thank you for your attention.

The content was chapter from AWS Foundations Module 9 - Cloud Architecture and AWS Foundations Module 10 - Automatic Scaling and Monitoring

Vytvorené v rámci projektu **KEGA 026TUKE-4/2021**