COMPUTING, ELASTICITY, AND SCALING

AUTO SCALING GROUPS

API

HYBRID CLOUDS

COMPUTING, ELASTICITY, AND SCALING

In this section, we are going to discuss various ways of scaling

Although *serverless* is promoted by clouds as the way

However, all scaling is based on the virtual machines

These VMs are coming from a pool

So first, let us review the virtual machine pools in the cloud

AUTO SCALING GROUPS

AUTO SCALING GROUPS
API
HYBRID CLOUDS

AUTO SCALING GROUP = ASG

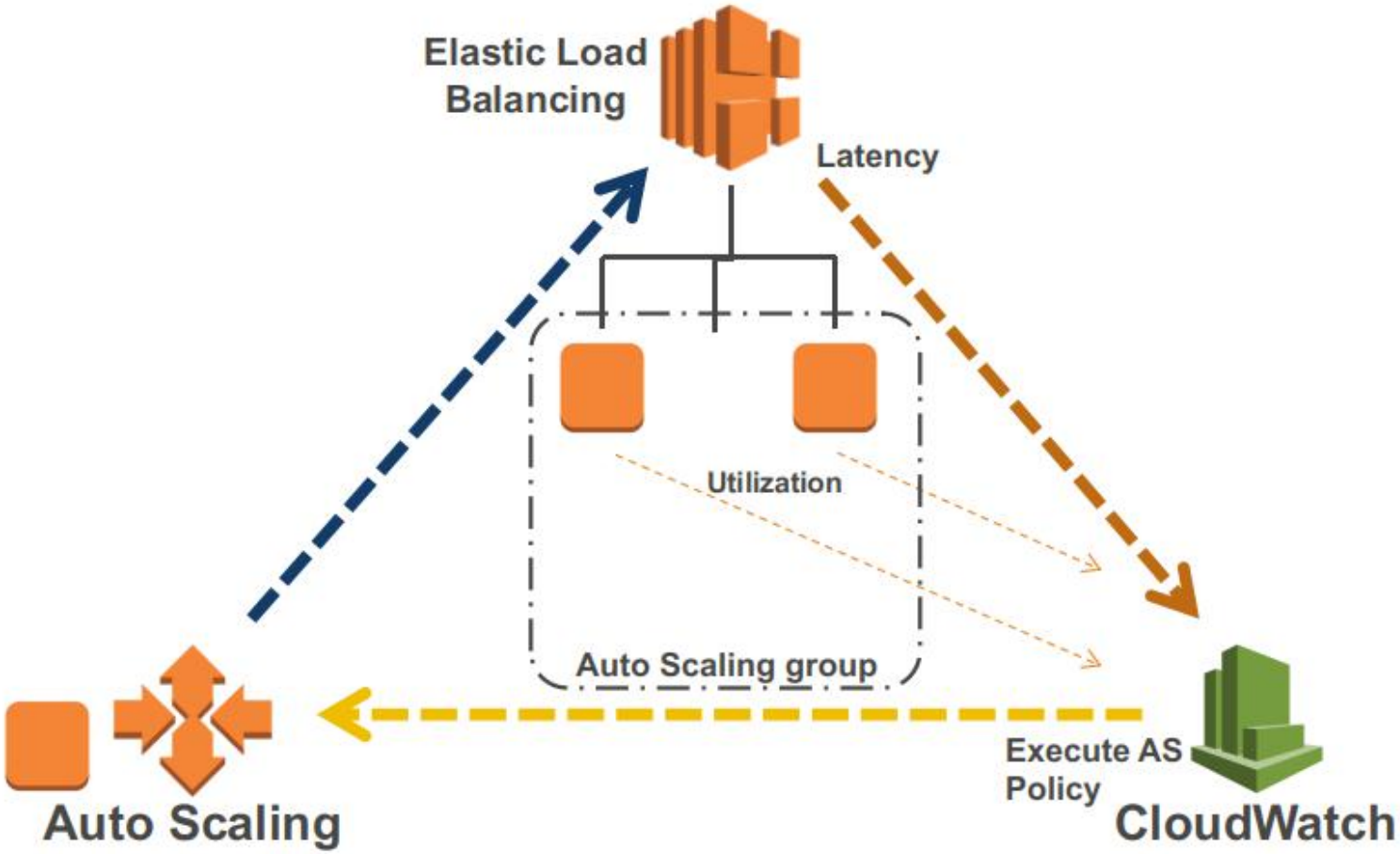
Auto scaling group = ASG

We will explain ASG first using AWS as example

Auto Scaling

- **Scale** your Amazon EC2 capacity **automatically**
- Well-suited for applications that experience **variability in usage**
- Available at no additional charge

TRIO OF SERVICES



AUTO SCALING BENEFITS

Better Fault Tolerance

Better Availability

Better Cost Management

LAUNCH CONFIGURATIONS

A launch configuration is a template that an Auto Scaling group uses to launch EC2 instances.

When you create a launch configuration, you can specify:

- AMI ID
- Instance type
- Key pair
- Security groups
- Block device mapping
- User data

LAUNCH CONFIGURATION EXAMPLE

EC2 > Launch configurations > Create launch configuration

Create launch configuration Info

Launch configuration name

Name

Amazon machine image (AMI) Info

AMI

timetrackeer backup 2020-04-08

Instance type Info

Instance type

Choose instance type

Additional configuration - *optional*

Purchasing option Info

Request Spot Instances

IAM instance profile Info

aws-elasticbeanstalk-ec2-role

Monitoring Info

Enable EC2 instance detailed monitoring within CloudWatch

Advanced details

Later, if you want to use a different launch configuration, you can create a new one and apply it to any Auto Scaling group. Existing launch configurations cannot be edited.

Storage (volumes) Info

EBS volumes

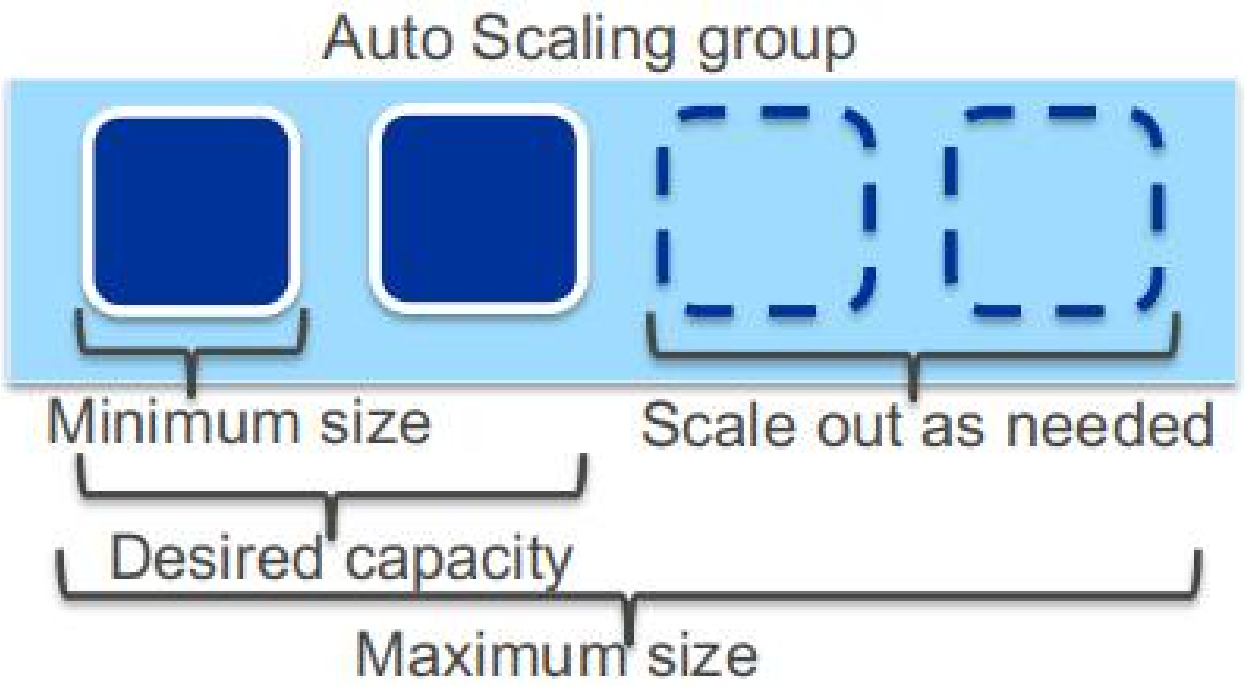
Remove

Volume type	Devices	Snapshot	Size (GiB)	Volume type
Root	/dev/sda1	snap-04308b13d91b27436	600	General purpose (SSD)

AUTO SCALING GROUPS

Contain a collection of EC2 instances that share similar characteristics.

Instances in an Auto Scaling group are treated as a logical grouping for the purpose of instance scaling and management.



DYNAMIC SCALING

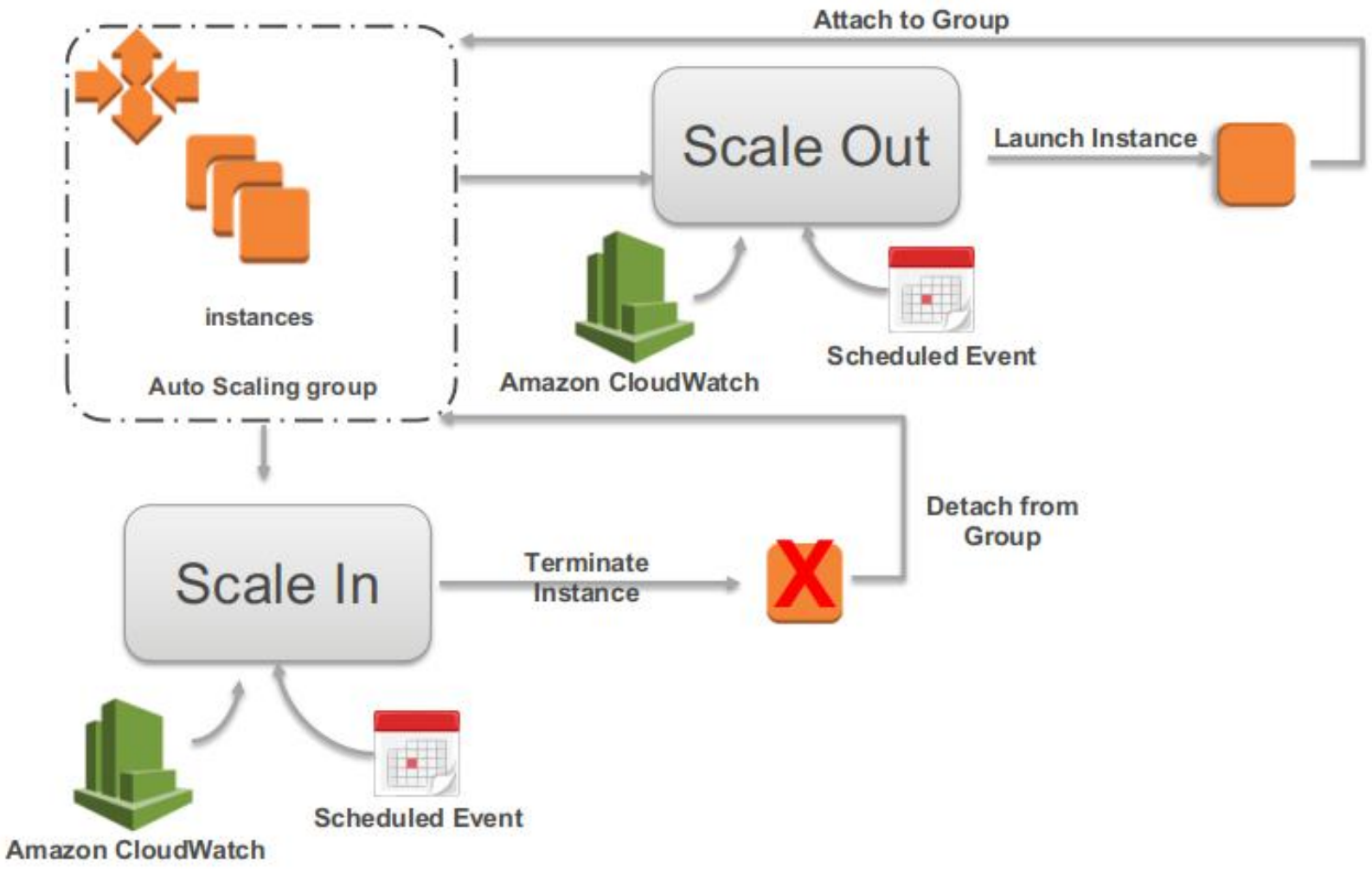
You can create a scaling policy that uses CloudWatch alarms to determine:

- When your Auto Scaling group should scale out.
- When your Auto Scaling group should scale in.

You can use alarms to monitor:

- Any of the metrics that AWS services send to Amazon CloudWatch.
- Your own custom metrics.

AUTO SCALING BASIC LIFECYCLE



QUIZ

Your architecture for an application currently consists of EC2 Instances sitting behind a classic ELB. The EC2 Instances are used to serve an application and are accessible through the internet. What can be done to improve this architecture in the event that the number of users accessing the application increases?

- A. Add another ELB to the architecture.
- B. Use Auto Scaling Groups.
- C. Use an Application Load Balancer instead.
- D. Use the Elastic Container Service.

QUIZ

You are told that a huge download is occurring on your instance. You have already set the Auto Scaling policy to increase the instance count when the network I/O increases beyond a certain limit. How can you ensure that this temporary event does not result in scaling?

- A. The policy cannot be set on the network I/O
- B. There is no way you can stop scaling as it is already configured
- C. The network I/O are not affected during data download
- D. You can suspend scaling temporarily

QUIZ

A user creates an Auto Scaling group from the Amazon AWS Console and assigned a tag with a key of "environment" and a value of "Prod". Can the user assign tags to instances launched in the Auto Scaling group, to organize and manage them?

- A. Yes, this is possible only if the tags are configured at the launch configuration with a maximum length of 300 characters.
- B. Yes
- C. Yes, this is possible only if the tags are in the same AZ and the tag names are uppercase.
- D. No

AZURE BATCH

Azure Batch enables large-scale parallel and high-performance computing (HPC) batch jobs with the ability to scale to tens, hundreds, or thousands of VMs.

When you're ready to run a job, Batch does the following:

- Starts a pool of compute VMs for you.
- Installs applications and staging data.
- Runs jobs with as many tasks as you have.
- Identifies failures.
- Requeues work.
- Scales down the pool as work completes.

There might be situations in which you need raw computing power or supercomputer-level compute power. Azure provides these capabilities.

AZURE AUTOSCALE

Dynamically scale apps to meet changing demand

Key scenarios:

- Maximize app responsiveness
- Scale by any metric
- Anticipate load with different schedules
- Save money by not wasting servers
- Dev-test at day, shut down at night

SCALE BY METRIC

Autoscale is a built-in feature of

- Cloud Services
- Mobile Services
- Virtual Machines
- Websites

Example

- Web app that handles millions of requests during the day and none at night.

Enable diagnostics in the portal

Collect logging data with the diagnostics library

Enable diagnostics in the Azure portal

EXAMINE THE CURRENT PRICING

webappplan - Scale up (App Service plan)
App Service plan

Search (Ctrl+/)

- Overview
- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems

Settings

- Apps
- File system storage
- Networking
- Scale up (App Service plan)
- Scale out (App Service plan)
- Resource explorer
- Properties
- Locks
- Export template

Monitoring

- Alerts
- Metrics

Support + troubleshooting

- Resource health
- New support request

Dev / Test
For less demanding workloads

Production
For most production workloads

Isolated
Advanced networking and scale

Recommended pricing tiers

S1
100 total ACU
1.75 GB memory
A-Series compute equivalent
55.43 GBP/Month (Estimated)

P1V2
210 total ACU
3.5 GB memory
Dv2-Series compute equivalent
110.86 GBP/Month (Estimated)

P2V2
420 total ACU
7 GB memory
Dv2-Series compute equivalent
221.79 GBP/Month (Estimated)

P3V2
840 total ACU
14 GB memory
Dv2-Series compute equivalent
443.57 GBP/Month (Estimated)

See additional options

Included features
Every app hosted on this App Service plan will have access to these features:

Custom domains / SSL
Configure and purchase custom domains with SNI and IP SSL bindings

Auto scale
Up to 10 instances. Subject to availability.

Staging slots
Up to 5 staging slots to use for testing and deployments before swapping them into production.

Daily backups
Backup your app 10 times daily.

Traffic manager

Included hardware
Every instance of your App Service plan will include the following hardware configuration:

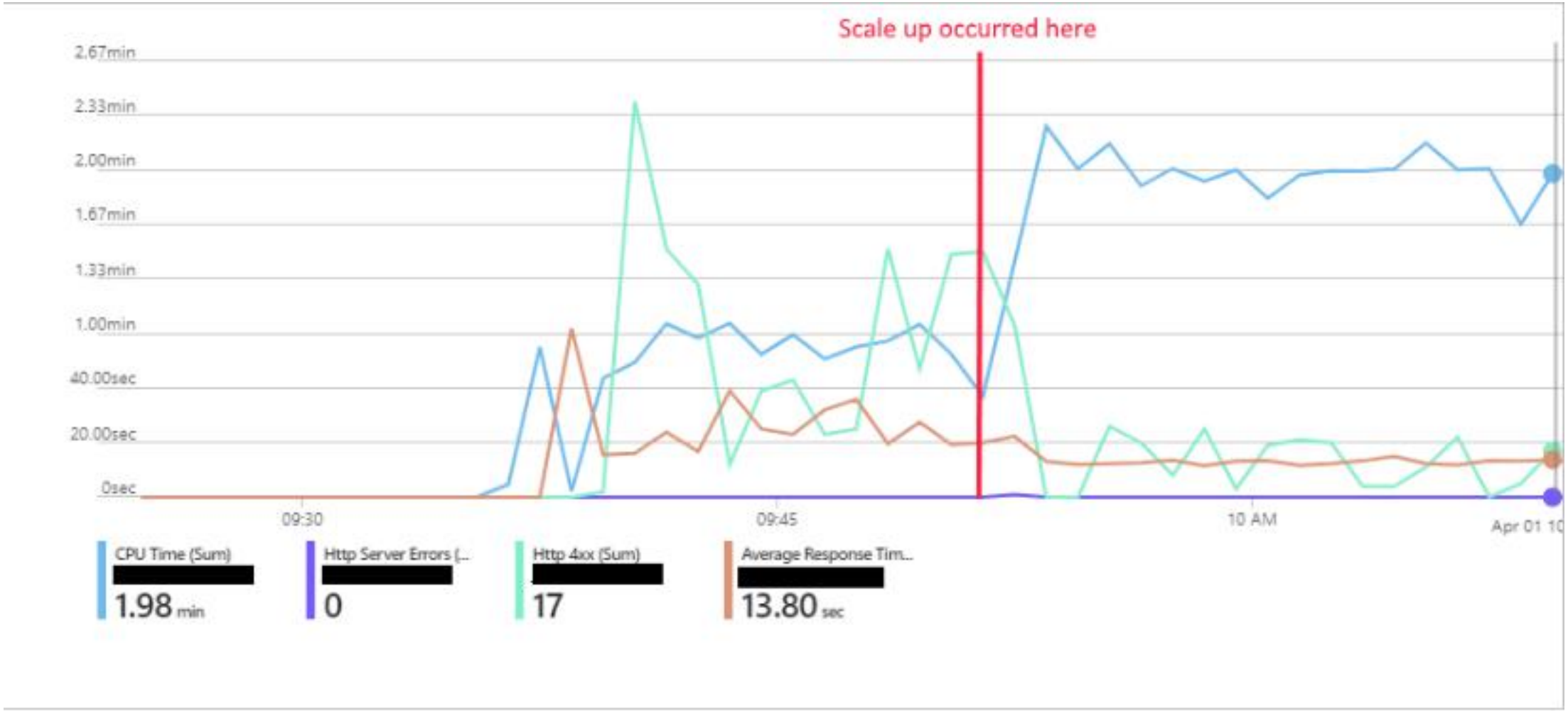
Azure Compute Units (ACU)
Dedicated compute resources used to run applications deployed in the App Service Plan. Learn more

Memory
Memory per instance available to run applications deployed and running in the App Service plan.

Storage
50 GB disk storage shared by all apps deployed in the App Service plan.

Apply

SCALING IN ACTION



QUIZ

Which Azure compute resource can be deployed to manage a set of identical virtual machines?

- A. Virtual machine availability sets
- B. Virtual machine availability zones
- C. Virtual machine scale sets

AUTOSCALING IN GCP

Available as part of the Compute Engine API

Used to automatically scale number of instances in a managed instance group based on workload

- Helps reduce costs by terminating instances when not required

Create one autoscaler per managed instance group

Autoscalers can be used with zone-based managed instance groups or regional managed instance groups

Autoscaler is fast, typically ~ 1 min moving window

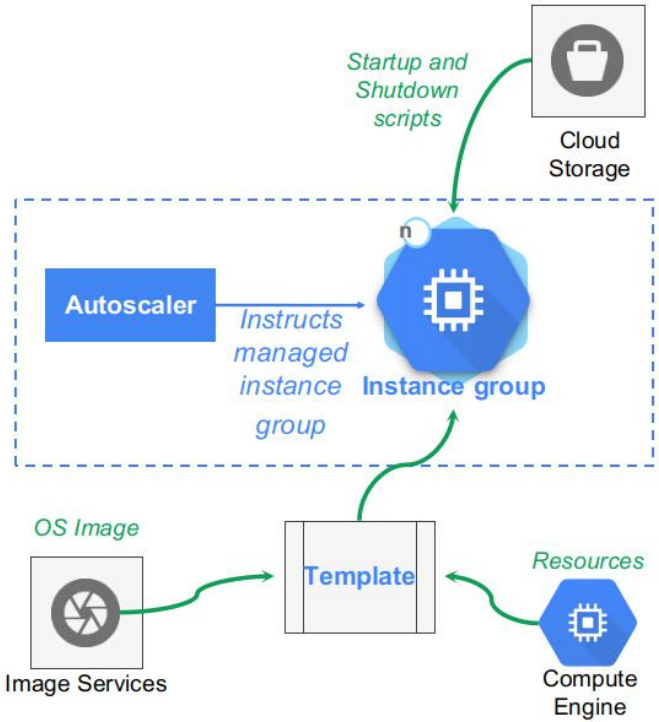
HOW AUTOSCALING WORKS

Autoscaler controls managed instance group

Adds, removes instances using policies

Policy includes number of replicas

- Max number
- Min number



POLICIES DETERMINE BEHAVIOR

Policy options

- Average CPU utilization
  - If average usage of total vCPU cores in instance group exceeds target, autoscaler adds more instances
- HTTP load balancing serving capacity (defined in the backend service)
  - Maximum CPU utilization
  - Maximum requests per second/instance
- Stackdriver standard and custom metrics
- Autoscaling

MULTIPLE POLICIES

Autoscaler allows multiple policies (up to 5)

Autoscaler handles multiple policies by calculating recommended number of virtual machines for each policy and picking policy that leaves the largest number of virtual machines in the group

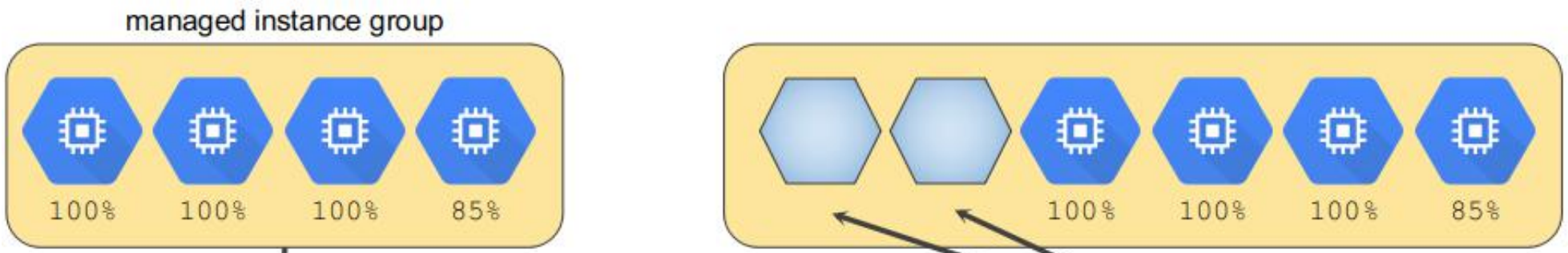
- Ensures enough virtual machines to handle application workloads and allows you to scale applications that have multiple possible bottlenecks

POLICY EXAMPLE: CPU UTILIZATION

Enable autoscaling for a managed instance group using CPU utilization

```
gcloud compute instance-groups managed \
      set-autoscaling example-managed-instance-group \
    --max-num-replicas 20 \
    --target-cpu-utilization 0.75 \
    --cool-down-period 90
```

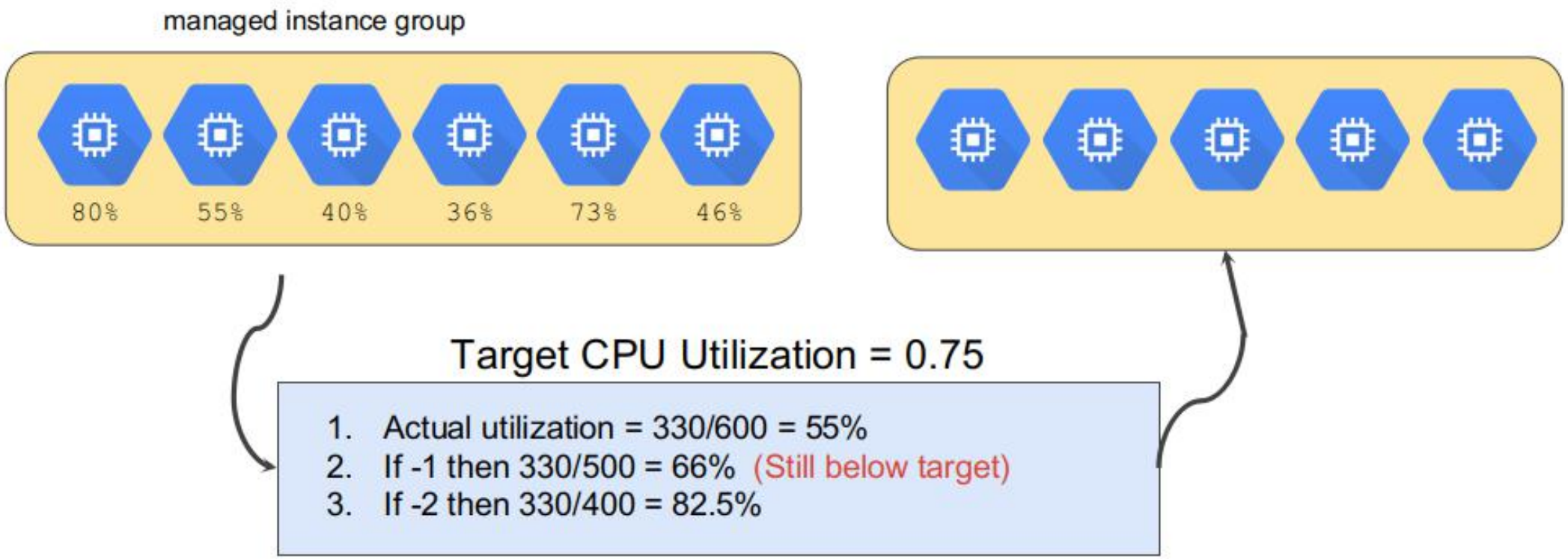
SCALE-OUT POLICY DECISION



Target CPU Utilization = 0.75

1. Actual utilization = 385/400 = 96.25%
2. If +1 then 385/500 = 77% (Still above target 75%)
3. If +2 then 385/600 = 64.16%

SCALE-IN POLICY DECISION



QUIZ

How does the autoscaler resolve conflicts between multiple scaling policies?

- A. First come, first served.
- B. It selects the one that recommends the most VMs, to ensure the application is supported.
- C. It selects the one with the fewest VMs to provide the lowest cost.
- D. It is based on priority, a value set in each policy that determines the precedence.

QUIZ

When autoscaling using Total CPU utilization, what is the difference on Total CPU utilization between adding the 4th VM to a group versus adding the 10th VM?

- A. The 4th VM adds 25% additional capacity, the 10th VM adds only 10% additional capacity.
- B. There is no difference, the VMs are identical and afford the same CPU capacity.
- C. The 4th VM uses a smaller CPU, so the 10th VM will provide 2.5 times more CPU capacity.
- D. The 4th VM adds 4% CPU capacity and the 10th VM adds 10% CPU capacity.

QUIZ

Which statement is true of autoscaling custom metrics?

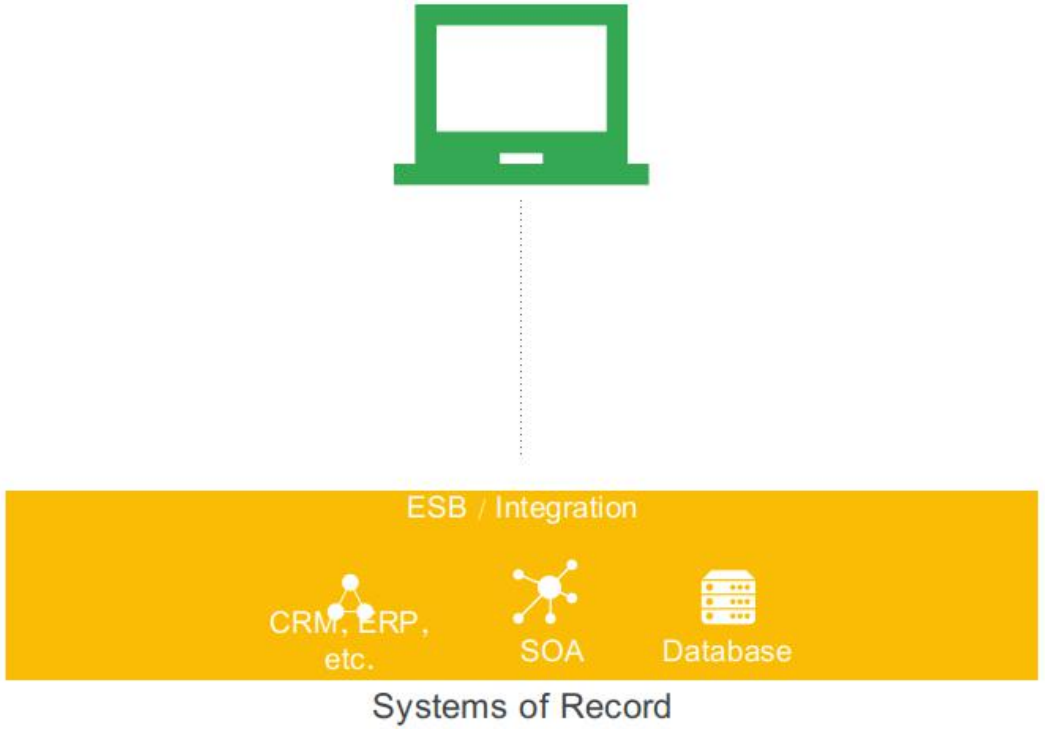
- A. Autoscaling does not support custom metrics.
- B. Custom metrics are much slower than native autoscaling metrics, so avoid using them.
- C. Stackdriver metrics can be used as custom metrics for autoscaling policies.
- D. Every custom metric includes a multiplier variable that you can use to adjust the input value range.

API

AUTO SCALING GROUPS
API
HYBRID CLOUDS

EARLY "API"

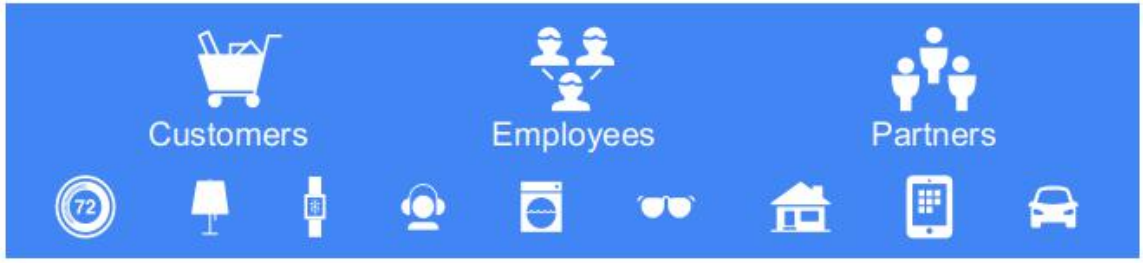
This is how you exposed your resources back then



THE GAP

How do you bridge the gap?

Modern best practice is to decouple, or "bridge"



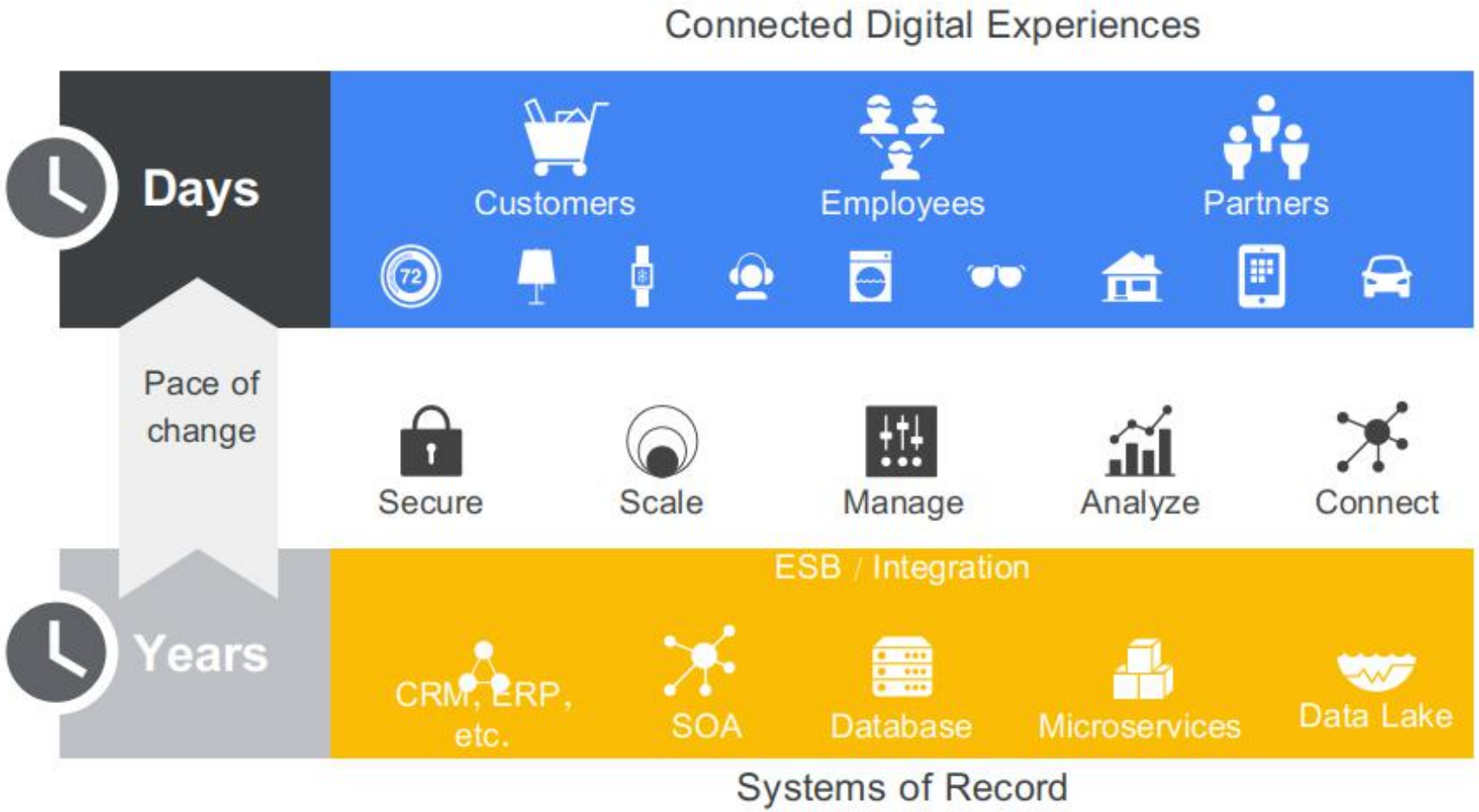
NEW ARCHITECTURE

This is a complete layer

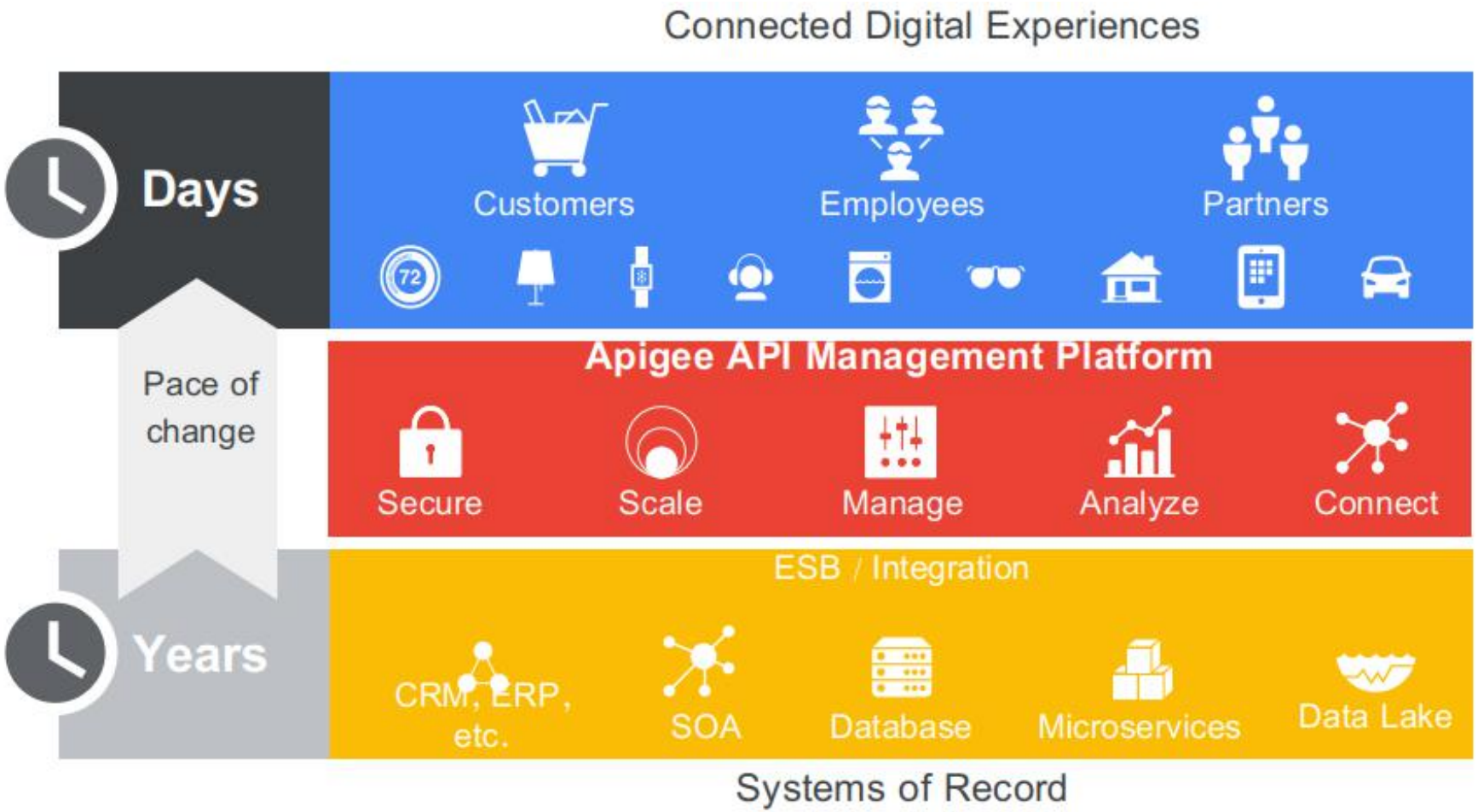
You may not need all of it, but it is a good list

PACE

The speed of development is increasing



APIGEE, THE "ALL-OUT" SOLUTION



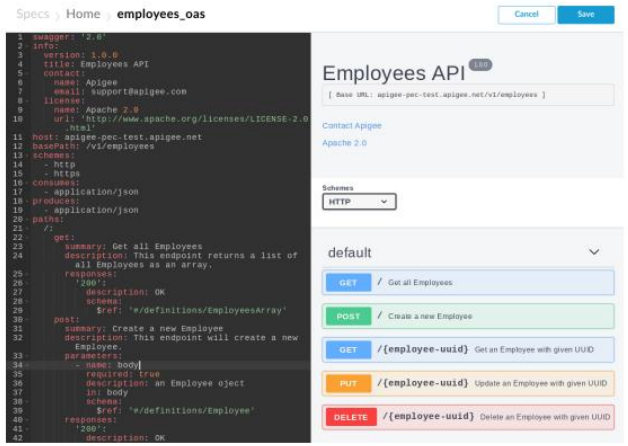
APIGEE IS ON GCP

OPENAPI IS THE COMMON LANGUAGE

You may or may not need to use API right now

But it is an upcoming architectural design pattern

So, it is good to know



QUIZ

Which versioning scheme follows Apigee's API design best practices?

- A. GET /customers/{customend}/v1
- B. GET /customers/v1/{customerid}
- C. GET /v1/customers/{customerid}
- D. GET /customers?customend={customerid}&version=v1

AMAZON API GATEWAY

API Types

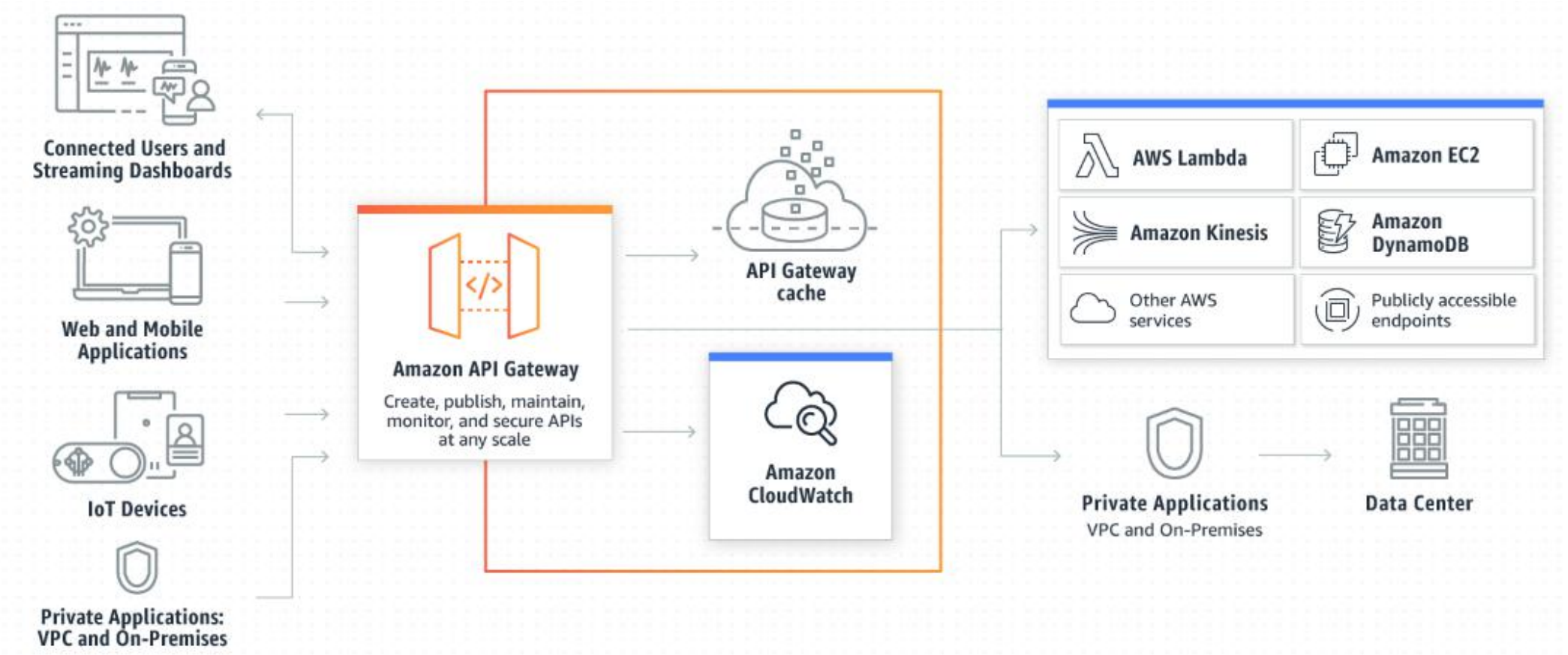
RESTful APIs

- Build RESTful APIs optimized for serverless workloads and HTTP backends using HTTP APIs.

WEBSOCKET APIs

- Build real-time two-way communication applications, such as chat apps and streaming dashboards

HOW AWS API GATEWAY WORKS



AWS API FEATURES

Private integrations with AWS ELB & AWS Cloud Map

- Route requests to private resources in your VPC, behind private ALBs, private NLBs

Resiliency

- Throttling based on number of requests per second

API creation

- AWS Lambda code in your account
- Start AWS Step Functions state machines
- Make calls to AWS Elastic Beanstalk, etc.

Monitoring

SDK

Lifecycle management

AZURE API GATEWAY

Azure API management service is between your APIs and the Internet

An Azure API gateway is an instance of the Azure API management service

Azure portal controls how particular APIs are exposed to consumers

AZURE API MANAGEMENT FEATURES

API documentation with Open API

Rate limiting access

Health monitoring

Modern formats like JSON

Analytics visualization

Security

- OAuth 2.0 user authorization
- Integration with Azure Active Directory.

BUILDING API ON AZURE

swagger Select a spec NorthWindShoes API V1

NorthWindShoes Products v1

[Base URL: shoecoapi61e0ec7c74.azurewebsites.net]
/swagger/v1/swagger.json

Inventory

GET /api/Inventory Retrieve the entire product inventory for the company.

GET /api/Inventory/{productid} Retrieve the number in stock for the specified product

Products

GET /api/Products Retrieve the details of every product sold

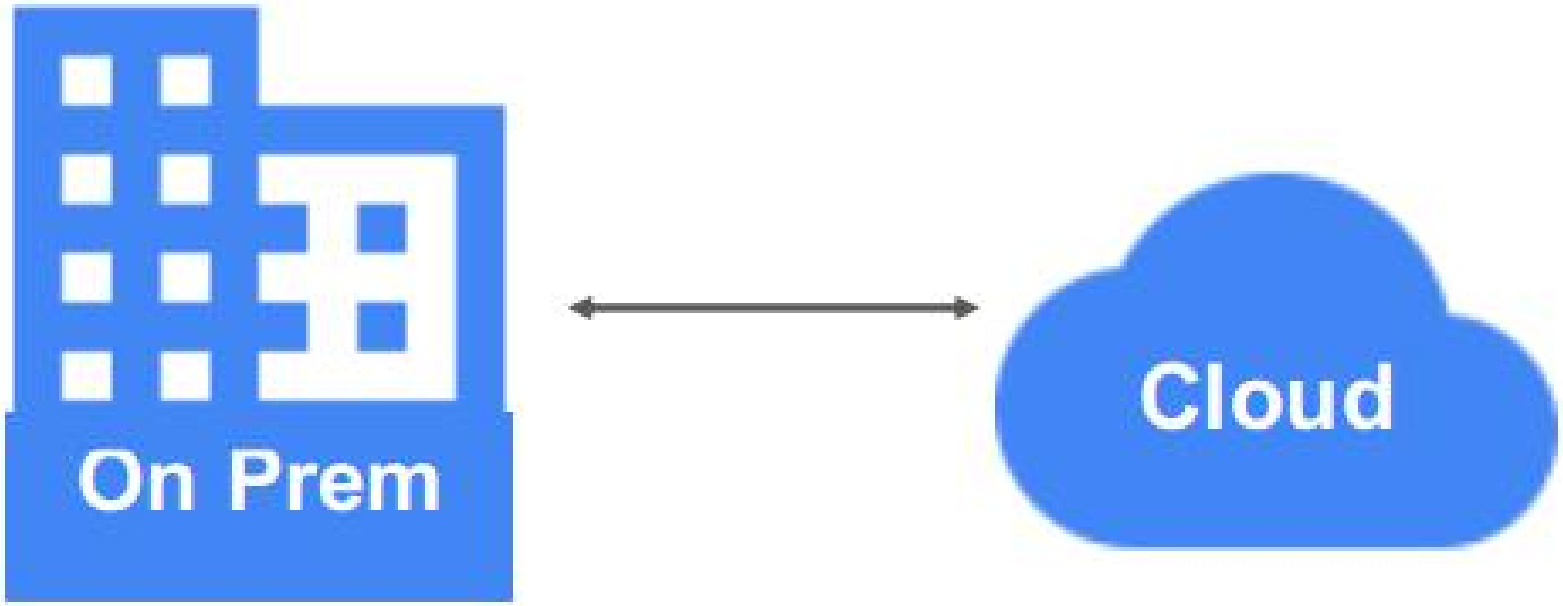
GET /api/Products/{productid} Find the details of the specified product

HYBRID CLOUDS

AUTO SCALING GROUPS
API
HYBRID CLOUDS

HYBRID CLOUDS

This, and more...

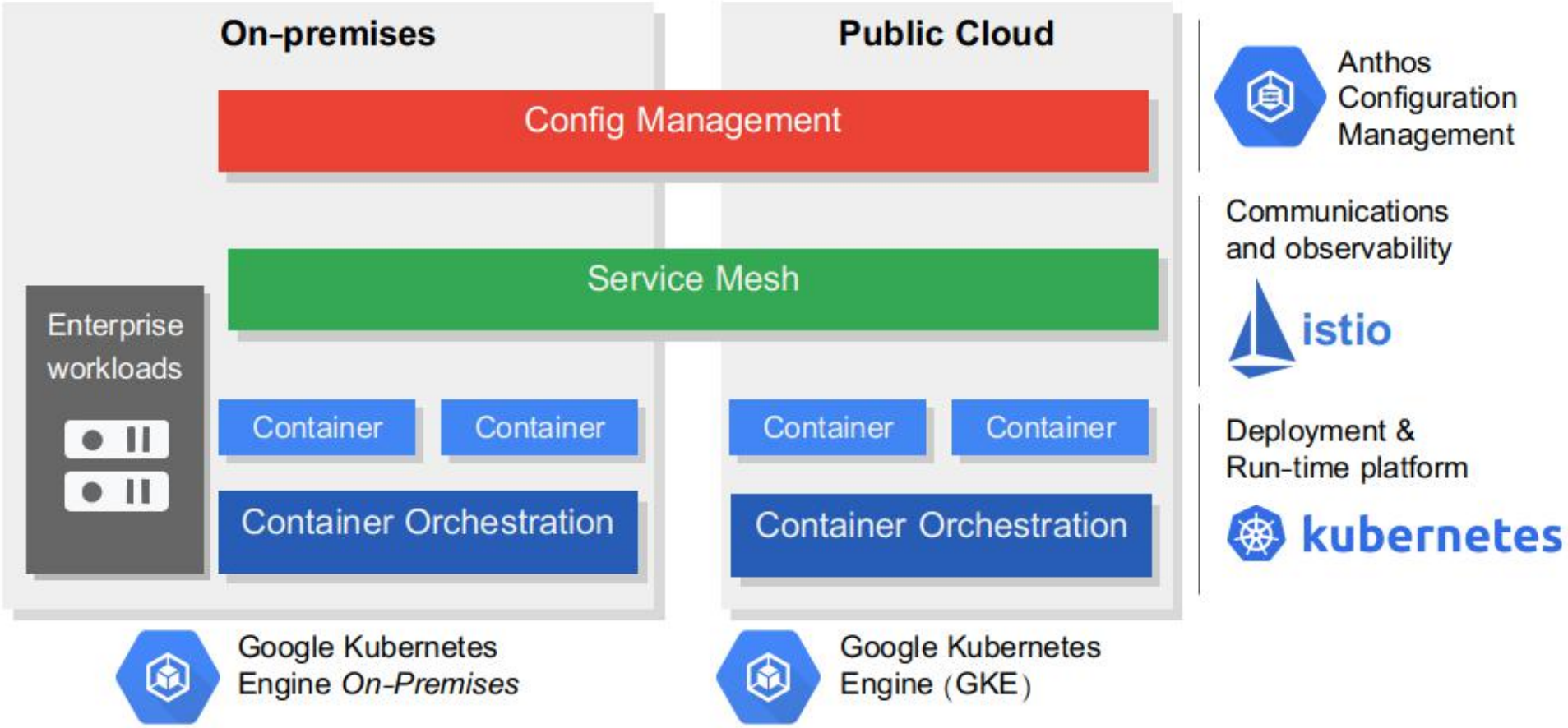


HYBRID ENVIRONMENT WISHLIST

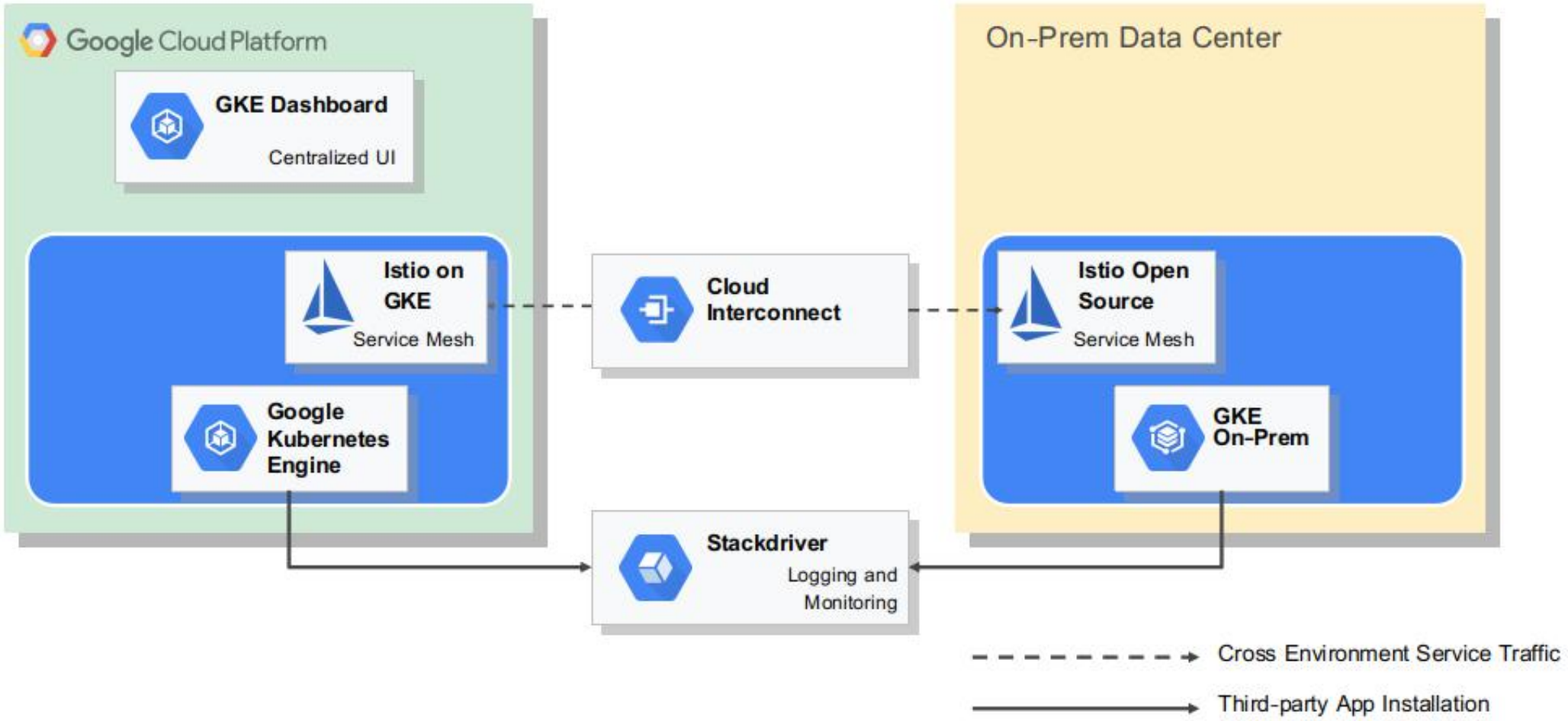
This may or may not be your goal though

- Write once, deploy in any cloud
- Accelerate developer velocity
- Consistency across environments
- Interoperability with legacy workloads
- Increased observability and SLO
- Decoupling across critical components
- Increased workload mobility
- Avoid vendor lock in

ANTHOS

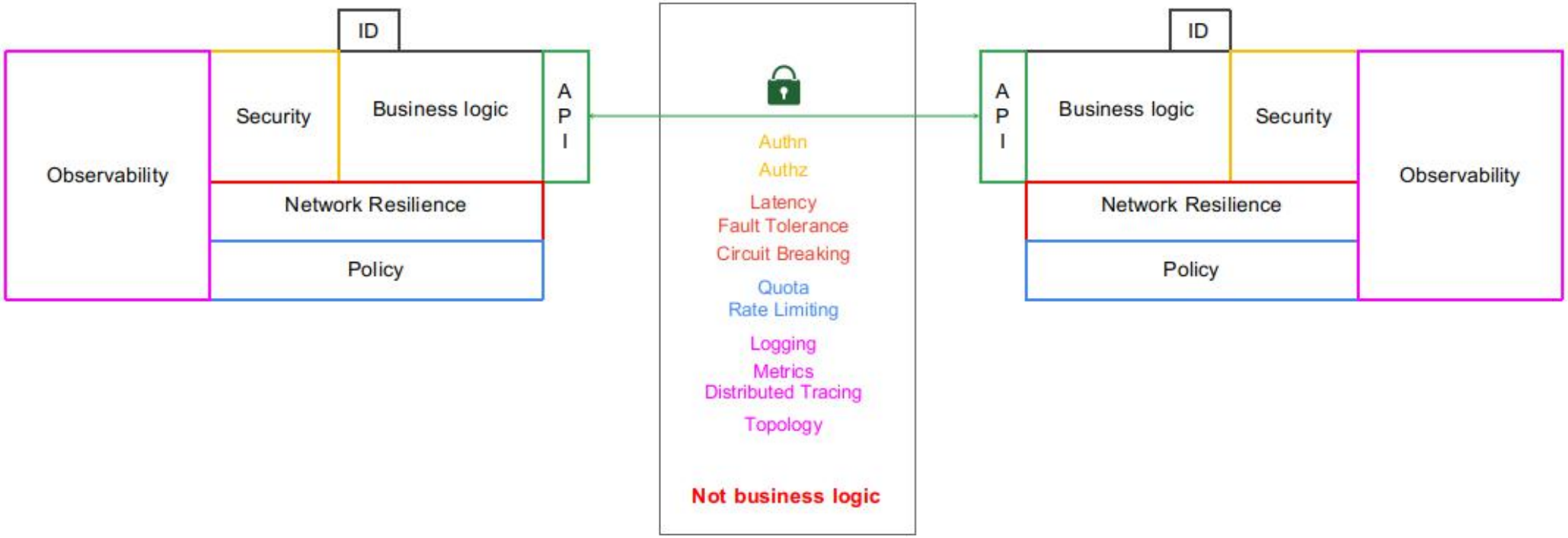


ANTHOS MORE DETAILS



ENABLER

Service mesh (ISTIO)



AZURE STACK

AZURE STACK OFFERINGS

CLOUD-MANAGED APPLIANCE

Azure Stack Edge

Run edge-computing workloads

Get rapid insights with an Azure managed appliance using compute and hardware-accelerated machine learning at edge locations for your Internet of Things (IoT) and AI workloads.

Use Azure Stack Edge for:

- Machine learning at the edge
- Edge and IoT solutions
- Network data transfer from edge to cloud

HYPERCONVERGED INFRASTRUCTURE

Azure Stack HCI

Modernize your datacenter

Refresh your virtualization host using a hybrid and hyperconverged solution integrated with Azure.

Use Azure Stack HCI for:

- Scalable virtualization and storage
- Modernizing on-premises architecture
- Remote branch offices
- High-performance workloads

CLOUD-NATIVE INTEGRATED SYSTEM

Azure Stack Hub

Use cloud services on-premises

Run your own private, autonomous cloud—connected or disconnected with cloud-native apps using consistent Azure services on-premises.

Use Azure Stack Hub for:

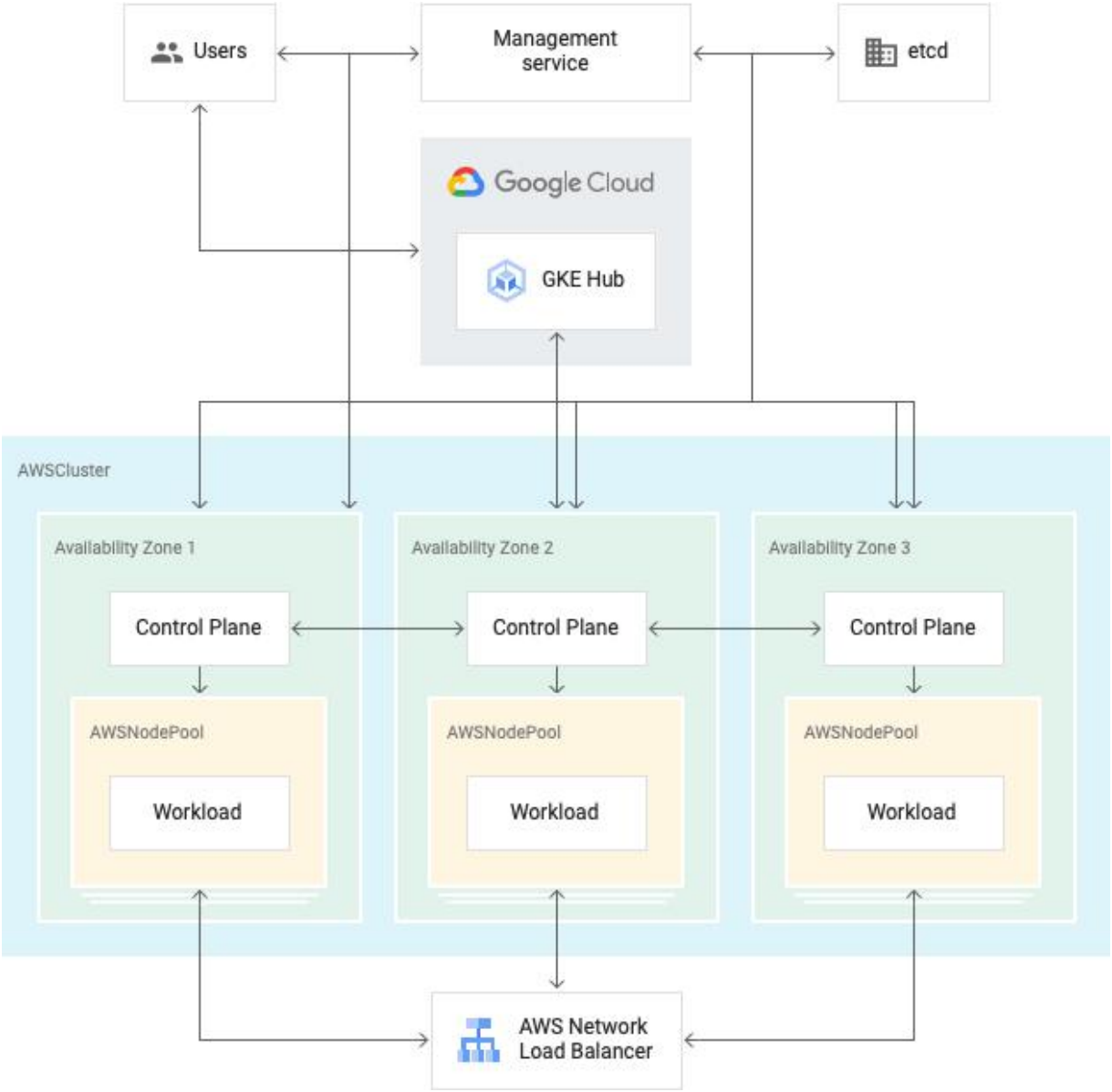
- Connected and disconnected scenarios
- Data sovereignty
- App modernization

AWS OUTPOSTS

Fully managed service

- Offers the same AWS infrastructure, AWS services, APIs, and tools to virtually any datacenter
- AWS compute, storage, database, etc. run locally on Outposts
- You can access the full range of AWS services available in the Region

ANTHOS ON AWS



AZURE ARC

Extends Azure management and services anywhere

Organize resources such as Windows and Linux Servers, Kubernetes clusters, and Azure data services.

Manage and govern resources at scale with powerful scripting tools, the Azure portal and APIs, and Azure Lighthouse.

Enforce organization standards and assess compliance at scale for all your resources, anywhere, with Azure Policy.

Modernize on-premises and multicloud operations through a plethora of Azure management and governance services.

CONGRATS ON COMPLETION