COMPUTING INSTANCES

VIRTUAL MACHINES

LAMBDA

VIRTUAL MACHINES

VIRTUAL MACHINES
LAMBDA

AMAZON EC2

Resizable compute capacity

Complete control of your computing resources

Reduces the time required to obtain and boot new server instances to minutes

AMAZON EC2 FACTS

Scale capacity as your computing requirements change

Pay only for capacity that you actually use

Choose Linux or Windows

Deploy across AWS Regions and Availability Zones for reliability

EC2 INSTANCE VIA THE WEB CONSOLE

Determine the AWS Region in which you want to launch the Amazon EC2 instance.

Launch an Amazon EC2 instance from a pre-configured Amazon Machine Image (AMI).

Choose an instance type based on CPU, memory, storage, and network requirements.

Configure network, IP address, security groups, storage volume, tags, and key pair.

AMI DETAILS

An AMI includes the following:

- A template for the root volume for the instance (for example, an operating system, an application server, and applications).
- Launch permissions that control which AWS accounts can use the AMI to launch instances.
- A block device mapping that specifies the volumes to attach to the instance when it's launched.

INSTANCES AND AMIS

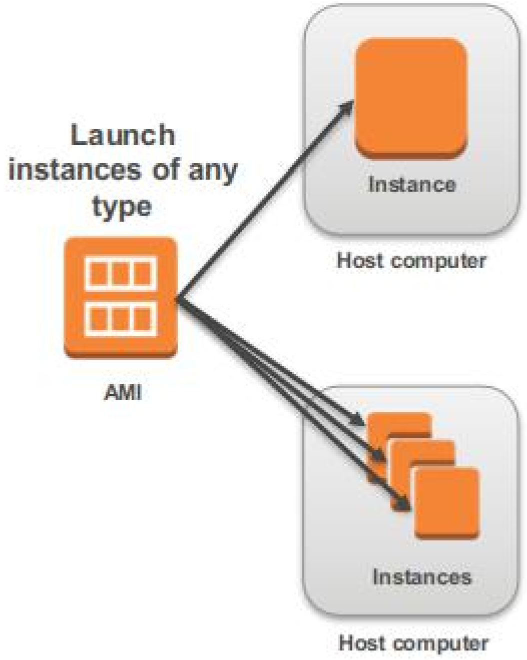
Region

Operating system

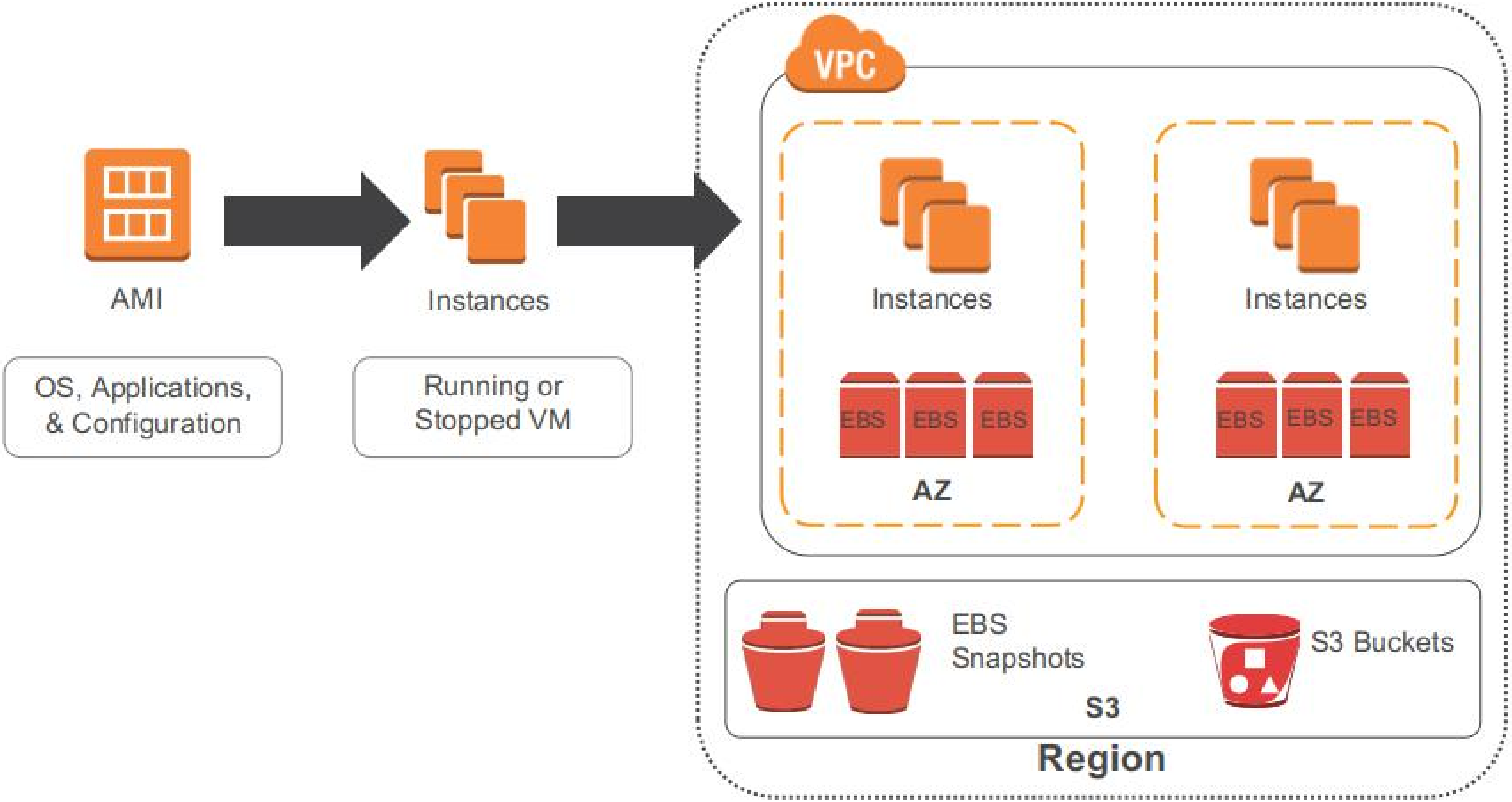
Architecture (32-bit or 64-bit)

Launch permissions

Storage for the root device



AMAZON EC2 INSTANCES



EBS VS. EC2 INSTANCE STORE

Amazon EBS

- Data stored on an Amazon EBS volume can persist independently of the life of the instance. Storage is persistent.

Amazon EC2 Instance Store

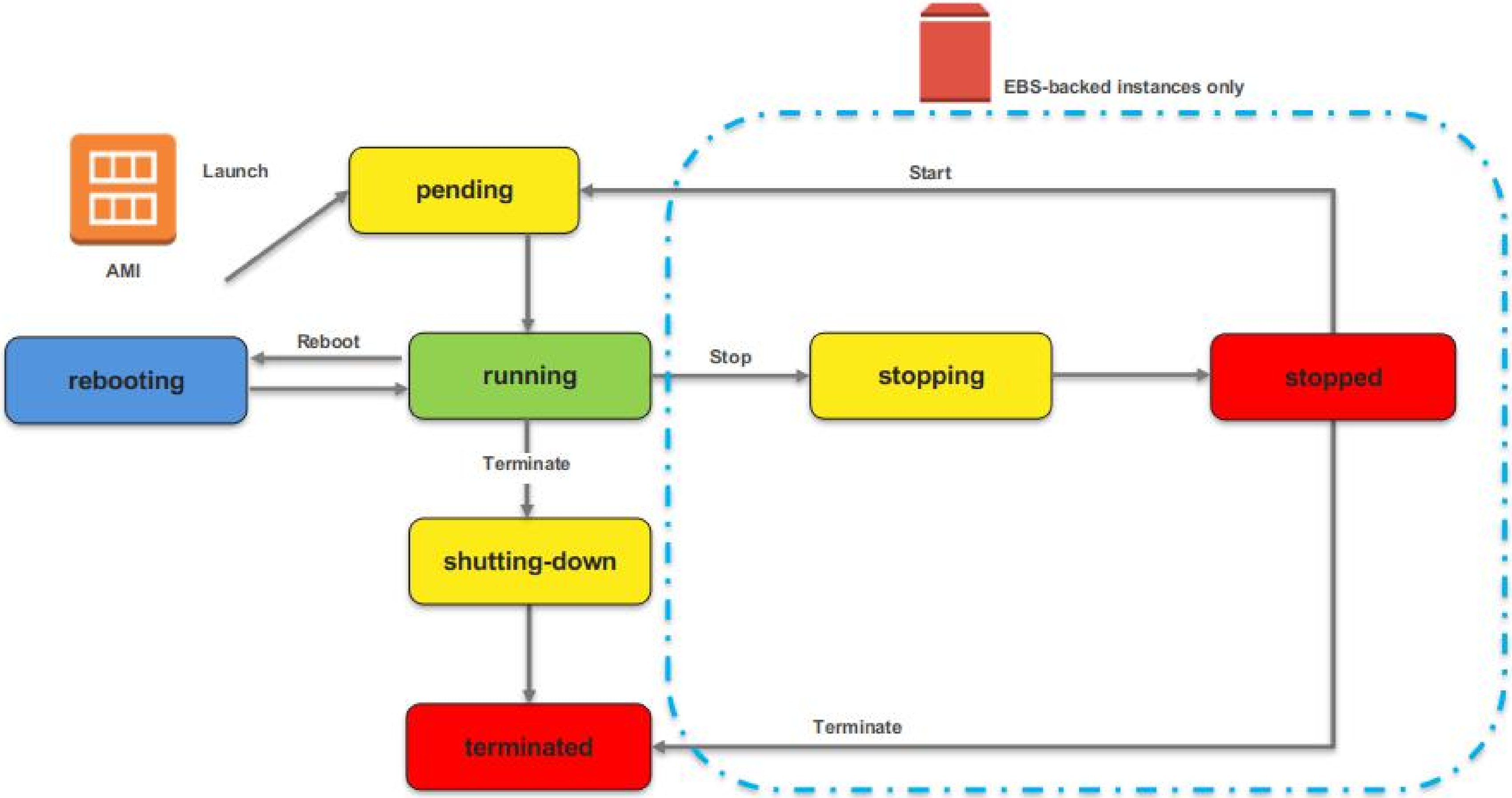
- Data stored on a local instance store persists only as long as the instance is alive. Storage is ephemeral.

Today, stateless systems are advocated as best practice, so let's discuss this

AMI TYPES - STORAGE FOR THE ROOT DEVICE

Characteristic	Amazon EBS-Backed	Amazon Instance Store-Backed
Boot time	Usually < 1 minute	Usually < 5 minutes
Size limit	16 TiB	10 GiB
Data persistence	The root volume is deleted when the instance terminates. Data on any other Amazon EBS volumes persists after instance termination.	Data on any instance store volumes persists only during the life of the instance.
Charges	Instance usage, Amazon EBS volume usage, and storing your AMI as an Amazon EBS snapshot.	Instance usage and storing your AMI in Amazon S3.
Stopped state	Can be stopped.	Cannot be stopped.

INSTANCE LIFECYCLE



CHOOSING THE RIGHT AMAZON EC2 INSTANCE

EC2 instance types are optimized for different use cases and come in multiple sizes. This allows you to optimally scale resources to your workload requirements.

AWS uses Intel® Xeon® processors for EC2 instances, providing customers with high performance and value.

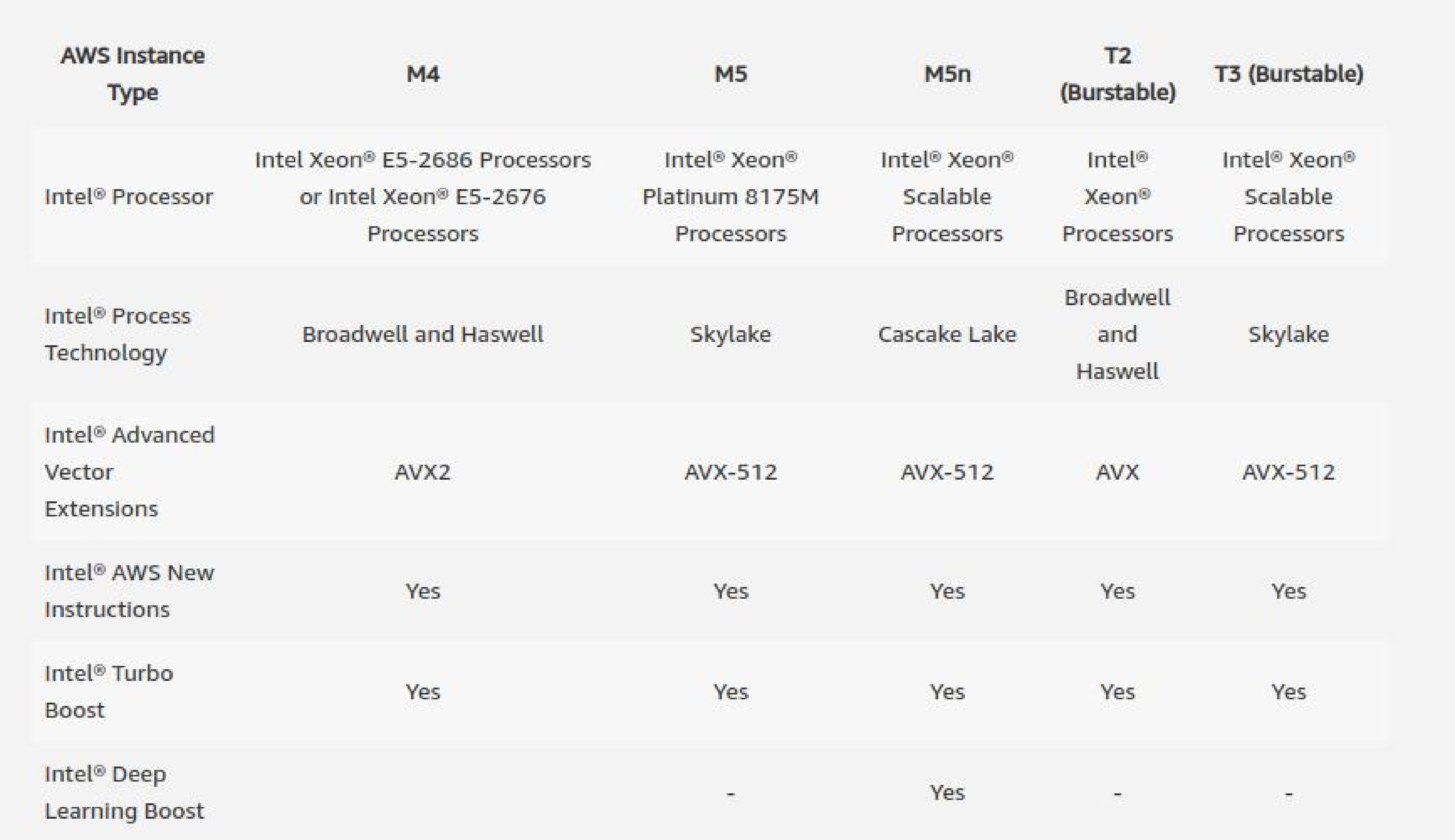
- **Note** AWS quotes instances with virtual cores (hyperthreading), which are not real cores if the instance is intended to do serious compute work. Need to divide the vCores by 2 to get real cores for compute-intensive tasks

Consider the following when choosing your instances: Core count, memory size, storage size and type, network performance, and CPU technologies.

Hurry Up and Go Idle - A larger compute instance can save you time and money, therefore paying more per hour for a shorter amount of time can be less expensive.

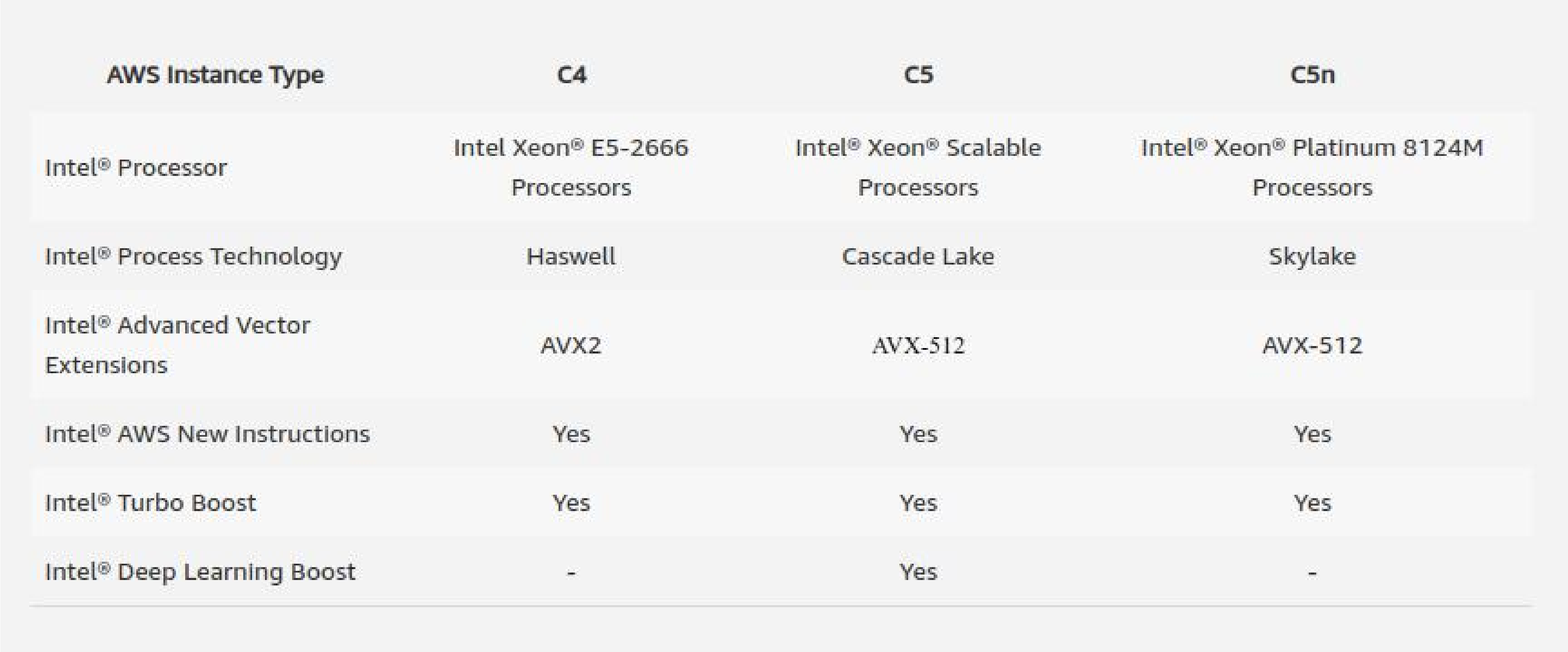
In other cases, horizontal scalability in clusters is the way to go

GENERAL PURPOSE



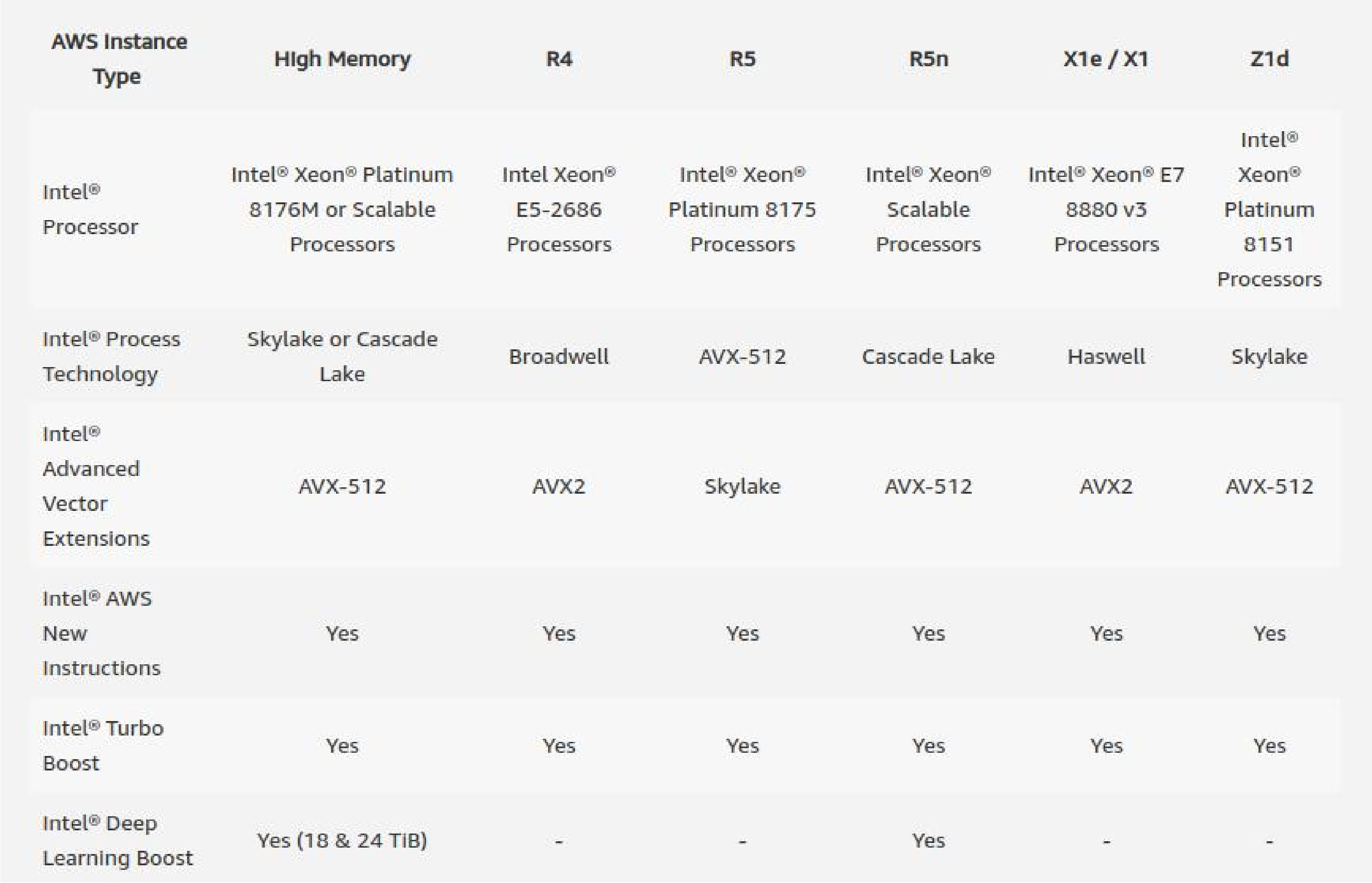
AWS Instance Type	M4	M5	M5n	T2 (Burstable)	T3 (Burstable)
Intel® Processor	Intel Xeon® E5-2686 Processors or Intel Xeon® E5-2676 Processors	Intel® Xeon® Platinum 8175M Processors	Intel® Xeon® Scalable Processors	Intel® Xeon® Processors	Intel® Xeon® Scalable Processors
Intel® Process Technology	Broadwell and Haswell	Skylake	Cascake Lake	Broadwell and Haswell	Skylake
Intel® Advanced Vector Extensions	AVX2	AVX-512	AVX-512	AVX	AVX-512
Intel® AWS New Instructions	Yes	Yes	Yes	Yes	Yes
Intel® Turbo Boost	Yes	Yes	Yes	Yes	Yes
Intel® Deep Learning Boost		-	Yes	-	-

COMPUTE OPTIMIZED



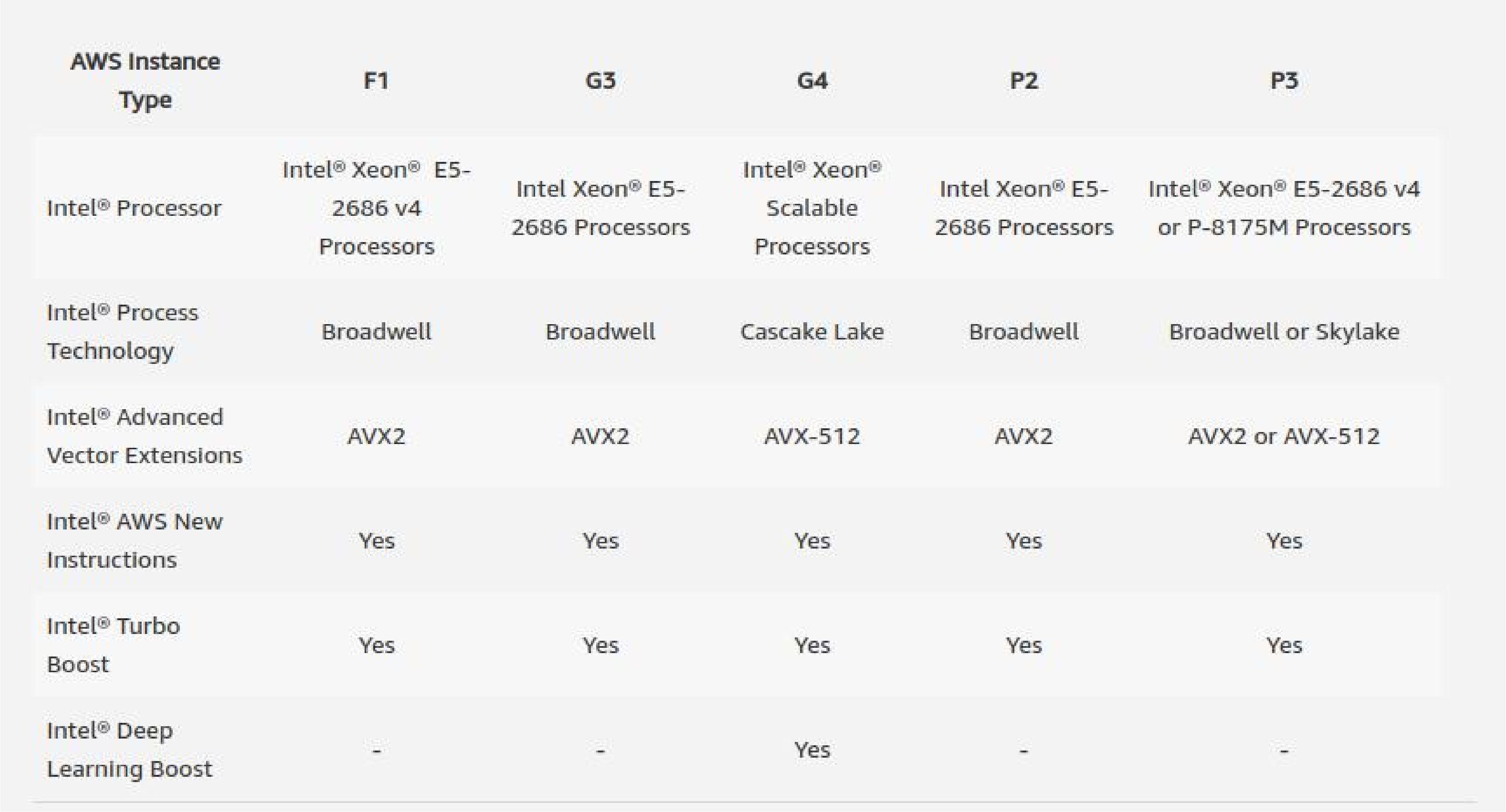
AWS Instance Type	C4	C5	C5n
Intel® Processor	Intel Xeon® E5-2666 Processors	Intel® Xeon® Scalable Processors	Intel® Xeon® Platinum 8124M Processors
Intel® Process Technology	Haswell	Cascade Lake	Skylake
Intel® Advanced Vector Extensions	AVX2	AVX-512	AVX-512
Intel® AWS New Instructions	Yes	Yes	Yes
Intel® Turbo Boost	Yes	Yes	Yes
Intel® Deep Learning Boost	-	Yes	-

MEMORY OPTIMIZED



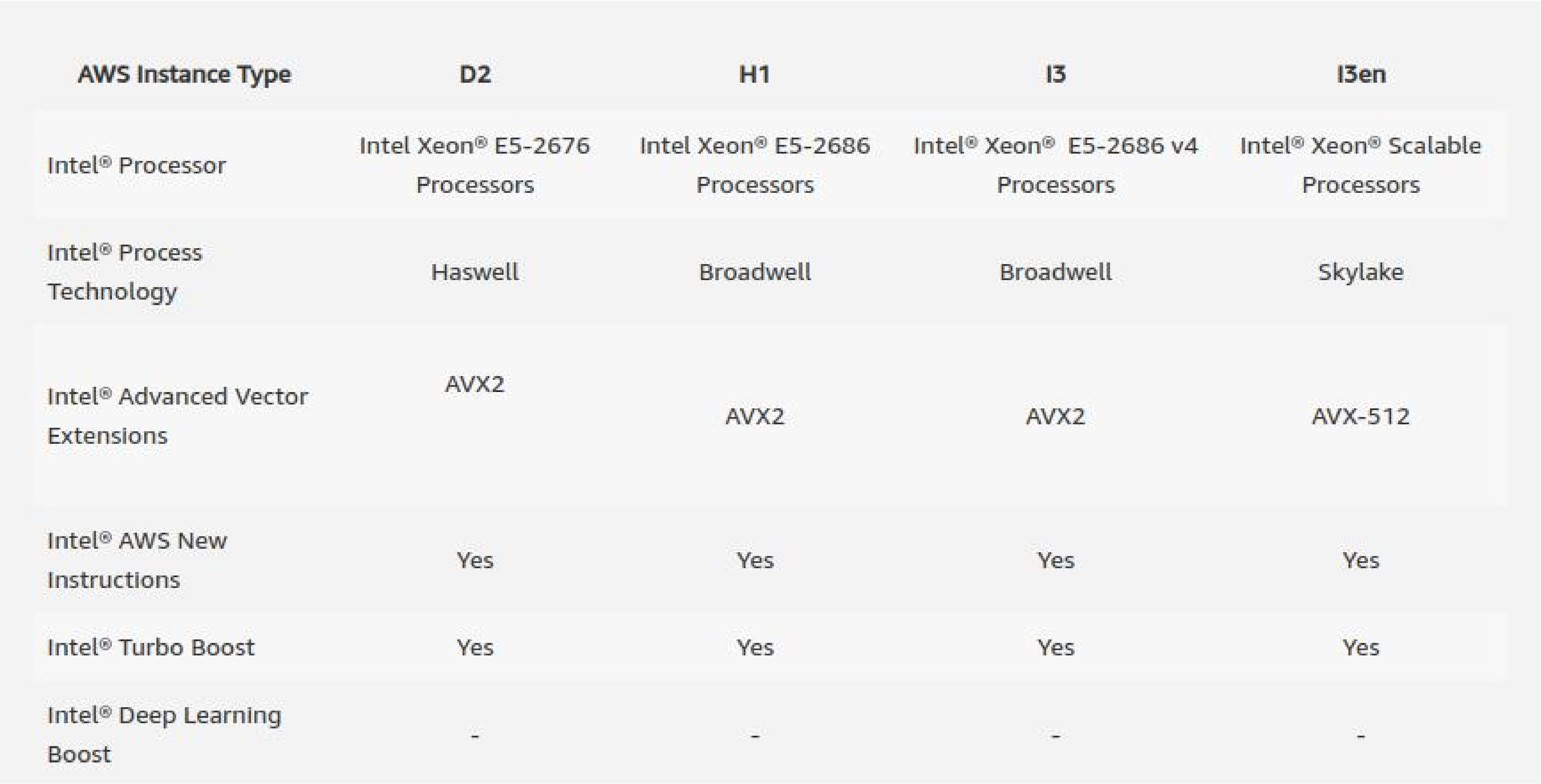
AWS Instance Type	High Memory	R4	R5	R5n	X1e / X1	Z1d
Intel® Processor	Intel® Xeon® Platinum 8176M or Scalable Processors	Intel Xeon® E5-2686 Processors	Intel® Xeon® Platinum 8175 Processors	Intel® Xeon® Scalable Processors	Intel® Xeon® E7 8880 v3 Processors	Intel® Xeon® Platinum 8151 Processors
Intel® Process Technology	Skylake or Cascade Lake	Broadwell	AVX-512	Cascade Lake	Haswell	Skylake
Intel® Advanced Vector Extensions	AVX-512	AVX2	Skylake	AVX-512	AVX2	AVX-512
Intel® AWS New Instructions	Yes	Yes	Yes	Yes	Yes	Yes
Intel® Turbo Boost	Yes	Yes	Yes	Yes	Yes	Yes
Intel® Deep Learning Boost	Yes (18 & 24 TiB)	-	-	Yes	-	-

ACCELERATED COMPUTING



AWS Instance Type	F1	G3	G4	P2	P3
Intel® Processor	Intel® Xeon® E5-2686 v4 Processors	Intel Xeon® E5-2686 Processors	Intel® Xeon® Scalable Processors	Intel Xeon® E5-2686 Processors	Intel® Xeon® E5-2686 v4 or P-8175M Processors
Intel® Process Technology	Broadwell	Broadwell	Cascake Lake	Broadwell	Broadwell or Skylake
Intel® Advanced Vector Extensions	AVX2	AVX2	AVX-512	AVX2	AVX2 or AVX-512
Intel® AWS New Instructions	Yes	Yes	Yes	Yes	Yes
Intel® Turbo Boost	Yes	Yes	Yes	Yes	Yes
Intel® Deep Learning Boost	-	-	Yes	-	-

STORAGE COMPUTING



AWS Instance Type	D2	H1	I3	I3en
Intel® Processor	Intel Xeon® E5-2676 Processors	Intel Xeon® E5-2686 Processors	Intel® Xeon® E5-2686 v4 Processors	Intel® Xeon® Scalable Processors
Intel® Process Technology	Haswell	Broadwell	Broadwell	Skylake
Intel® Advanced Vector Extensions	AVX2	AVX2	AVX2	AVX-512
Intel® AWS New Instructions	Yes	Yes	Yes	Yes
Intel® Turbo Boost	Yes	Yes	Yes	Yes
Intel® Deep Learning Boost	-	-	-	-

CURRENT GENERATION INSTANCES

Instance Family	Some Use Cases
General purpose (t2, m4, m3)	Low-traffic websites and web applications Small databases and mid-size databases
Compute optimized (c4, c3)	High performance front-end fleets Video-encoding
Memory optimized (r3)	High performance databases Distributed memory caches
Storage optimized (i2, d2)	Data warehousing Log or data-processing applications
GPU instances (g2)	3D application streaming Machine learning

INSTANCE METADATA & USER DATA

Instance Metadata:

- Is data about your instance.
- Can be used to configure or manage a running instance.

Instance User Data:

- Can be passed to the instance at launch.
- Can be used to perform common automated configuration tasks.
- Runs scripts after the instance starts.

RETRIEVING INSTANCE METADATA

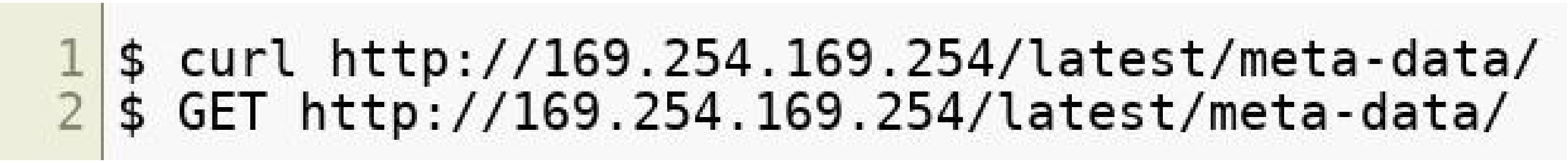
To view all categories of instance metadata from within a running instance, use the following URI: http://169.254.169.254/latest/meta-data/

On a Linux instance, you can use:

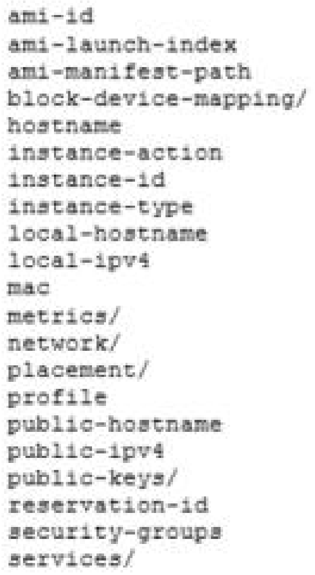
Please note that this means that the metadata is open to all users on that machine

- Is this the behavior you would expect?

All metadata is returned as text (content type text/plain).



```
$ curl http://169.254.169.254/latest/meta-data/
$ GET http://169.254.169.254/latest/meta-data/
```



```
ami-id
ami-launch-index
ami-manifest-path
block-device-mapping/
hostname
instance-action
instance-id
instance-type
local-hostname
local-ipv4
mac
metrics/
network/
placement/
profile
public-hostname
public-ipv4
public-keys/
reservation-id
security-groups
services/
```

ADDING USER DATA

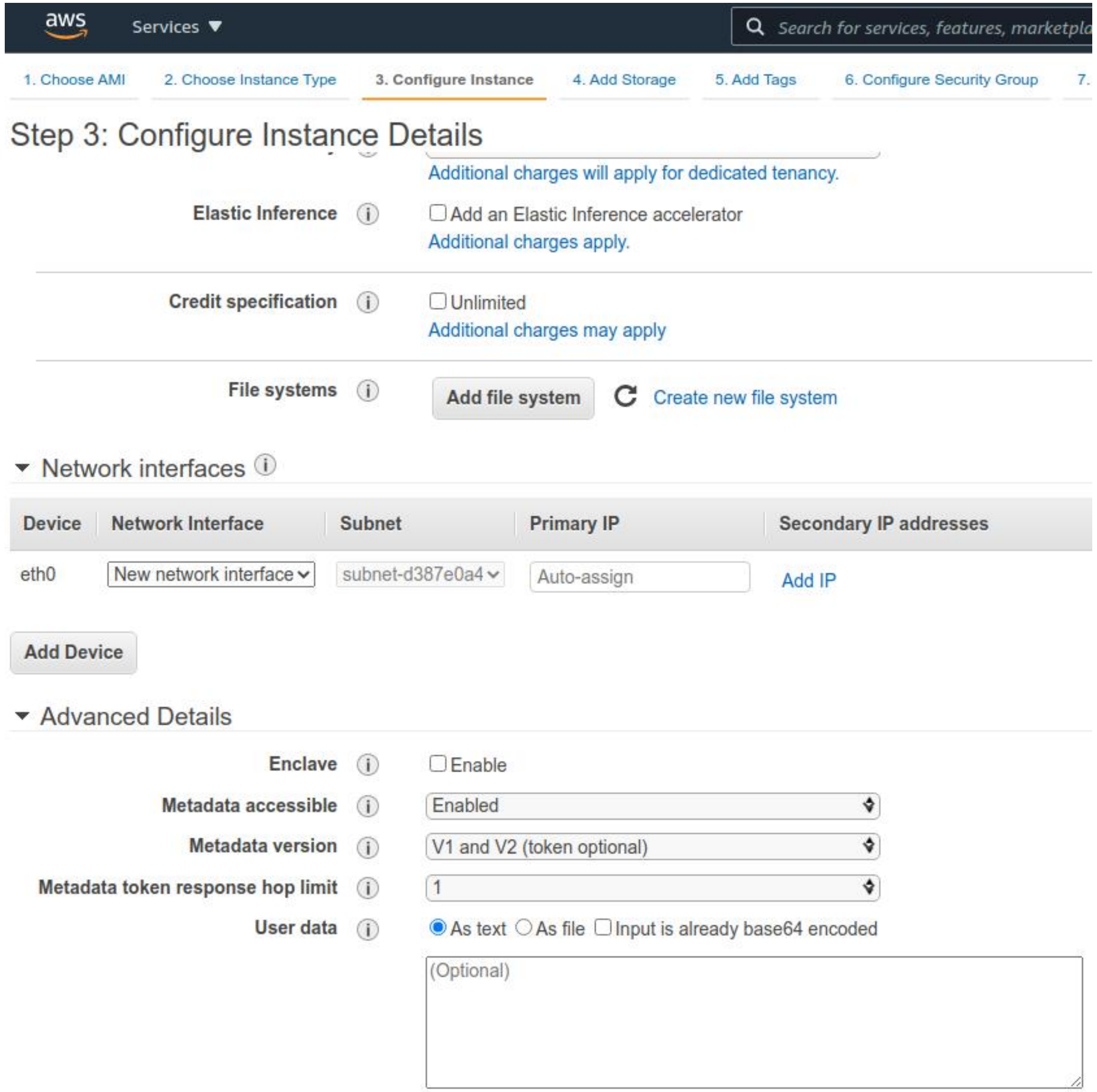
You can specify user data when launching an instance.

User data can be:

- Linux script – executed by cloud-init
- Windows batch or PowerShell scripts – executed by EC2Config service

User data scripts run once per instance-id by default.

ADDING USER DATA



Step 3: Configure Instance Details

Additional charges will apply for dedicated tenancy.

Elastic Inference ⓘ □ Add an Elastic Inference accelerator
Additional charges apply.

Credit specification ⓘ □ Unlimited
Additional charges may apply

File systems ⓘ

▾ Network interfaces ⓘ

Device	Network Interface	Subnet	Primary IP	Secondary IP addresses
eth0	New network interface	subnet-d387e0a4	Auto-assign	Add IP

▾ Advanced Details

Enclave ⓘ □ Enable

Metadata accessible ⓘ Enabled

Metadata version ⓘ V1 and V2 (token optional)

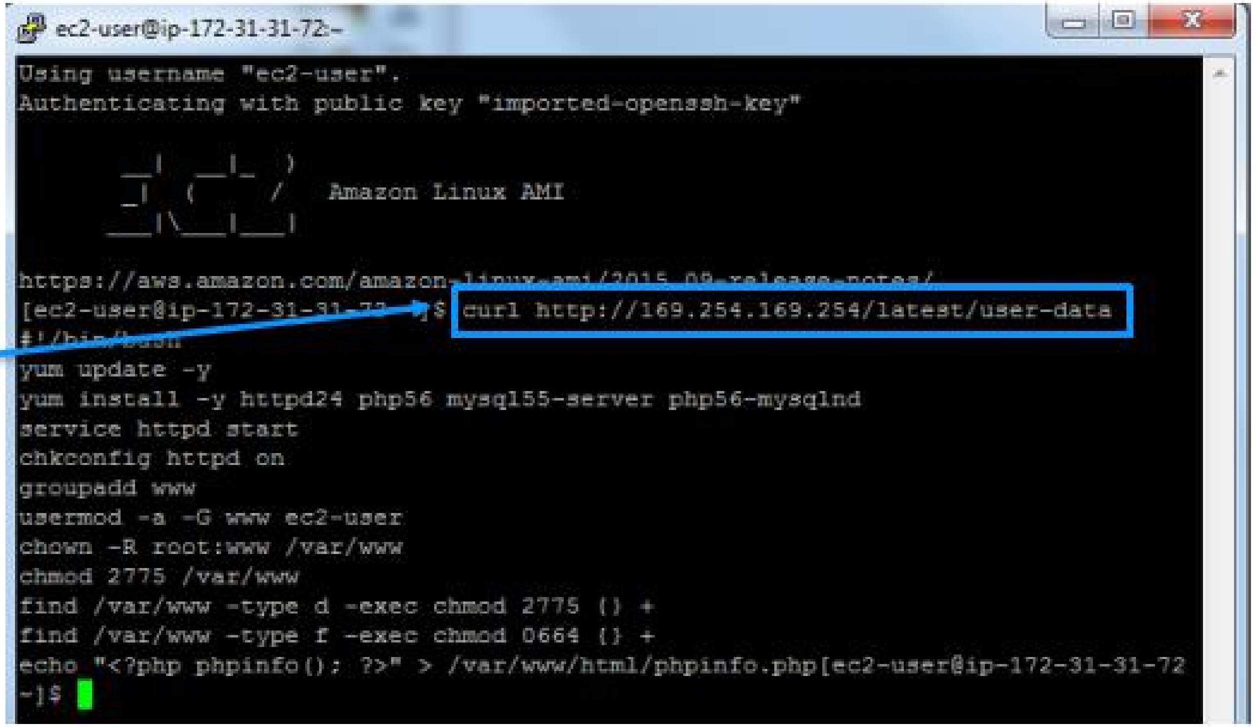
Metadata token response hop limit ⓘ 1

User data ⓘ ● As text ○ As file □ Input is already base64 encoded

(Optional)

RETRIEVING USER DATA

- To retrieve user data, use the following URI: http://169.254.169.254/latest/user-data
- On a Linux instance, you can use:
  - `$ curl http://169.254.169.254/latest/user-data/`
  - `$ GET http://169.254.169.254/latest/user-data/`

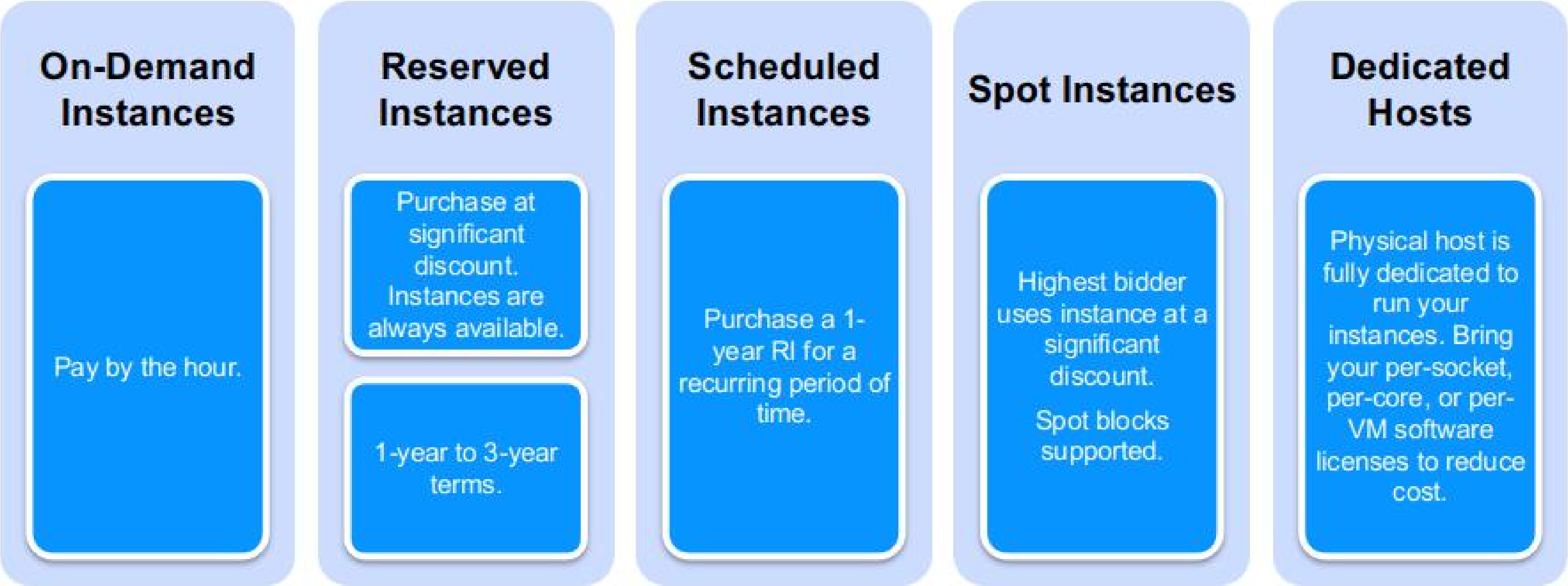


```
Using username "ec2-user".
Authenticating with public key "imported-openssh-key"

       __|  __|_  )
       _|  (     /   Amazon Linux AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-ami/2015.09-release-notes/
[ec2-user@ip-172-31-31-72 ~]$ curl http://169.254.169.254/latest/user-data
#!/bin/bash
yum update -y
yum install -y httpd24 php56 mysql55-server php56-mysqlnd
service httpd start
chkconfig httpd on
groupadd www
usermod -a -G www ec2-user
chown -R root:www /var/www
chmod 2775 /var/www
find /var/www -type d -exec chmod 2775 {} +
find /var/www -type f -exec chmod 0664 {} +
echo "<?php phpinfo(); ?>" > /var/www/html/phpinfo.php[ec2-user@ip-172-31-31-72
~]$
```

AMAZON EC2 PURCHASING OPTIONS



On-Demand Instances

Pay by the hour.

Reserved Instances

Purchase at significant discount. Instances are always available.

1-year to 3-year terms.

Scheduled Instances

Purchase a 1-year RI for a recurring period of time.

Spot Instances

Highest bidder uses instance at a significant discount.

Spot blocks supported.

Dedicated Hosts

Physical host is fully dedicated to run your instances. Bring your per-socket, per-core, or per-VM software licenses to reduce cost.

Pricing is by the hour or by the second depending on instance type (Linux is per-second)

https://aws.amazon.com/ec2/pricing/

QUIZ

When creating a new security group, all inbound traffic is allowed by default.

- A. True
- B. False

QUIZ

To help you manage your Amazon EC2 instances, you can assign your own metadata in the form of

- A. Wildcards
- B. Certificates
- C. Tags
- D. Notes

QUIZ

Can I move a reserved instance from one region to another?

- A. Yes
- B. No
- C. It depends on the region
- D. Only in the U.S.

QUIZ

You need to know both the private IP address and public IP address of your EC2 instance. You should

- A. Run `ipconfig` in Windows or `ifconfig' in Linux
- B. Retrieve the instance metadata from http://169.254.169.254/latest/meta-data
- C. Retrieve the User Data from http://169.254.169.254/latest/meta-data
- D. Run the following command: `aws ec2 display-ip`

QUIZ

Individual instances are provisioned

- A. In regions
- B. In availability zones
- C. Globally

VIRTUAL MACHINES ON AZURE

Start with the network

- Segregate
- Secure

Name the VM

Decide the location for the VM

Determine the size of the VM

Understanding the pricing model

Storage for the VM

Select an operating system

NAMING THE VM

Element	Example	Notes
Environment	dev, prod, QA	Identifies the environment for the resource
Location	uw (US West), ue (US East)	Identifies the region into which the resource is deployed
Instance	01, 02	For resources that have more than one named instance (web servers, etc.)
Product or Service	service	Identifies the product, application, or service that the resource supports
Role	sql, web, messaging	Identifies the role of the associated resource

NAMING THE VM EXAMPLE

Example:

- `devusc-webvm01`
- to represent the first development web server hosted in the US South Central location.

AZURE VM SIZE

Option	Description
General purpose	General-purpose VMs are designed to have a balanced CPU-to-memory ratio. Ideal for testing and development, small to medium databases, and low to medium traffic web servers.
Compute optimized	Compute optimized VMs are designed to have a high CPU-to-memory ratio. Suitable for medium traffic web servers, network appliances, batch processes, and application servers.
Memory optimized	Memory optimized VMs are designed to have a high memory-to-CPU ratio. Great for relational database servers, medium to large caches, and in-memory analytics.
Storage optimized	Storage optimized VMs are designed to have high disk throughput and IO. Ideal for VMs running databases.
GPU	GPU VMs are specialized virtual machines targeted for heavy graphics rendering and video editing. These VMs are ideal options for model training and inferencing with deep learning.
High performance computes	High performance compute is the fastest and most powerful CPU virtual machines with optional high-throughput network interfaces.

AZURE VM PRICES

Option	Description
Pay as you go	With the pay-as-you-go option, you pay for compute capacity by the second, with no long-term commitment or upfront payments. You're able to increase or decrease compute capacity on demand as well as start or stop at any time. Prefer this option if you run applications with short-term or unpredictable workloads that cannot be interrupted. For example, if you are doing a quick test, or developing an app in a VM, this would be the appropriate option.
Reserved Virtual Machine Instances	The Reserved Virtual Machine Instances (RI) option is an advance purchase of a virtual machine for one or three years in a specified region. The commitment is made up front, and in return, you get up to 72% price savings compared to pay-as-you-go pricing. RIs are flexible and can easily be exchanged or returned for an early termination fee. Prefer this option if the VM has to run continuously, or you need budget predictability, and you can commit to using the VM for at least a year.

OPTIONS TO CREATE VM

Azure Resource Manager

Azure PowerShell

Azure CLI

Azure REST API

Azure Client SDK

Azure VM Extensions

Azure Automation Services

QUIZ

Suppose you want to run a network appliance on a virtual machine. Which workload option should you choose?

A. General purpose

B. Compute optimized

C. Memory optimized

D. Storage optimized

QUIZ

True or false: Resource Manager templates are JSON files?

- A. True
- B. False

VM ON GCP

GCE - Google Compute Engine

- High CPU, high memory, standard and shared-core machine types
- Persistent disks
- Standard, SSD, local SSD
- Snapshots
- Resize disks with no downtime
- Instance metadata and startup scripts

GCE PRICING

Compute Engine offers customer friendly pricing

- Per-second billing, sustained use discounts, committed use, discounts
- Preemptible instances
- High throughput to storage at no extra cost
- Custom machine types: Only pay for the hardware you need

SCALING WITH COMPUTE ENGINE

Scale up or scale out with Compute Engine

Use big VMs for memory- and compute-intensive applications

Use Autoscaling for resilient, scalable applications

COMPUTE ENGINE DISKS

Feature	Amazon EBS	Compute Engine
Volume types	EBS Provisioned IOPS SSD, EBS General Purpose SSD, Throughput Optimized HDD, Cold HDD	Zonal standard persistent disks (HDD), regional persistent disks, zonal SSD persistent disks, regional SSD persistent disks
Volume locality rules	Must be in same zone as instance to which it is attached	Must be in same zone as instance to which it is attached
Volume attachment	Can be attached to only one instance at a time	Read-write volumes: Can be attached to only one instance at a time Read-only volumes: Can be attached to multiple instances
Attached volumes per instance	Up to 40	Up to 128
Maximum volume size	16TiB	64TB
Redundancy	Zonal	Zonal or multi-zonal depending on volume type
Snapshotting	Yes	Yes
Snapshot locality	Regional	Global

GOOGLE LOCAL SSD

Feature	Amazon EC2	Compute Engine
Service name	Instance store (also known as ephemeral store)	Local SSD
Volume attachment	Tied to instance type	Can be attached to any non-shared-core instance
Device type	Varies by instance type	SSD
Attached volumes per instance	Varies by instance type	Up to 8
Storage capacity	Varies by instance type	375 GB per volume
Live migration	No	Yes
Redundancy	None	None

QUIZ

Data on local SSDs persists through live migration events

- A. True
- B. False

QUIZ

What size should your boot disks be to ensure the best performance when using persistent disks

- A. Larger than 50GB
- B. Larger than 100GB
- C. Smaller than 50GB
- D. Larger than 200GB

QUIZ

The data that you store on a local SSD persists only until you stop or delete the instance

- A. True
- B. False

QUIZ

What size are local SSD devices?

- A. 375GB
- B. 250GB
- C. 500GB
- D. 1TB

LAMBDA

VIRTUAL MACHINES
LAMBDA

WHAT IS AWS LAMBDA?

Compute service that runs your functions in response to event.

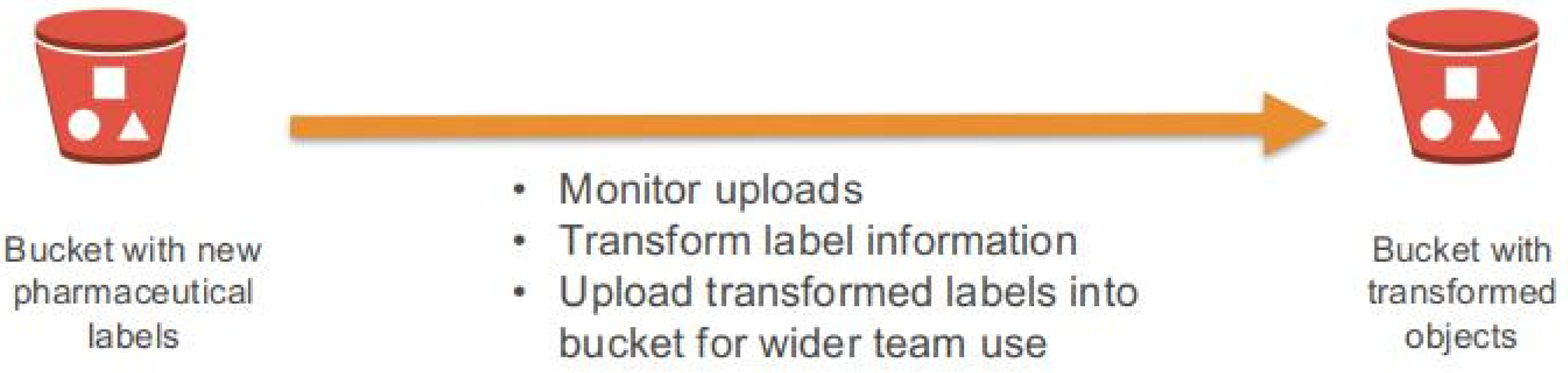
Automatically manages the compute resources for you.

Requires zero administration.

WHY AWS LAMBDA?

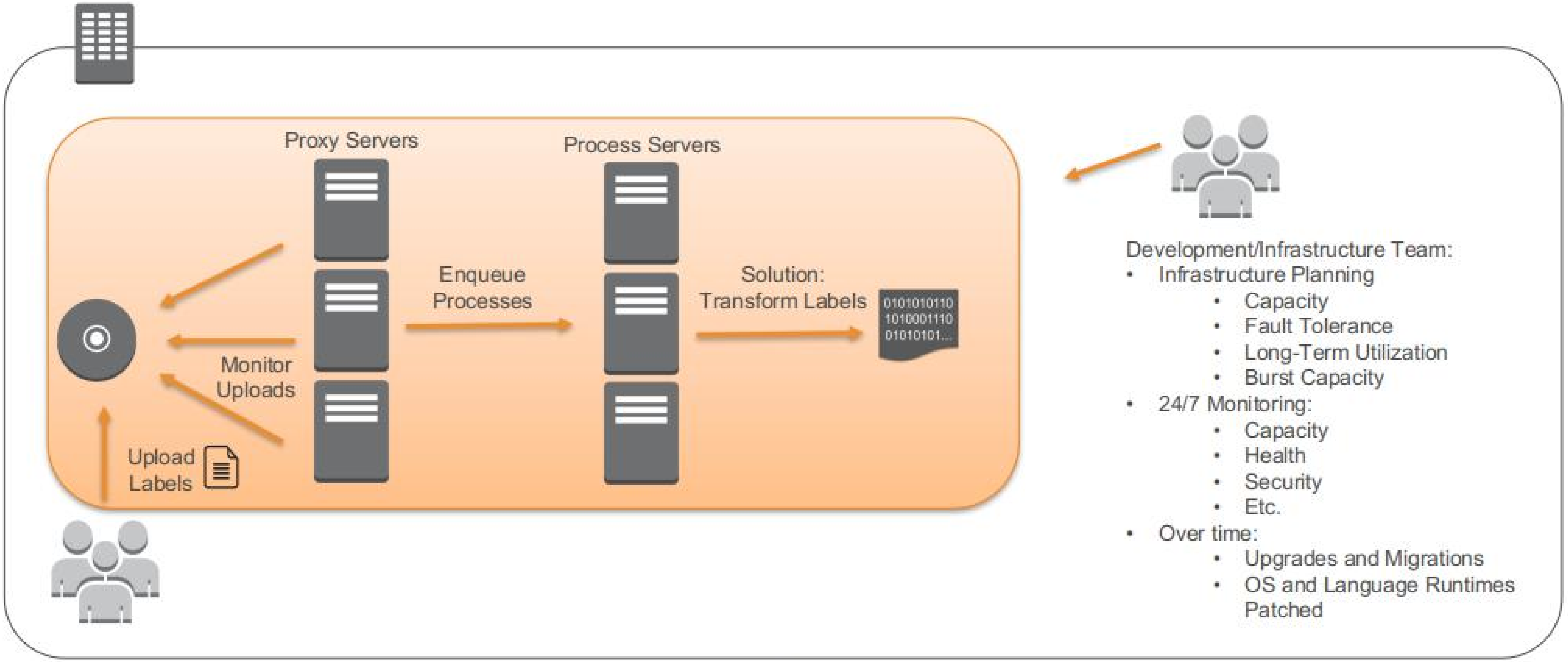
Simple problem:

- GoGreen Healthcare creates pharmaceutical labels when new products are released in compliance with the Food & Drug Administration’s (FDA) Structured Product Labeling (SPL).
- Transform label data into a format to be used in trend analysis by other teams.


Bucket with new pharmaceutical labels
Bucket with transformed objects

- Monitor uploads
- Transform label information
- Upload transformed labels into bucket for wider team use

WHY AWS LAMBDA?



Development/Infrastructure Team:

- Infrastructure Planning
    - Capacity
    - Fault Tolerance
    - Long-Term Utilization
    - Burst Capacity
- 24/7 Monitoring:
    - Capacity
    - Health
    - Security
    - Etc.
- Over time:
    - Upgrades and Migrations
    - OS and Language Runtimes Patched

corporate data center

WHY AWS LAMBDA?

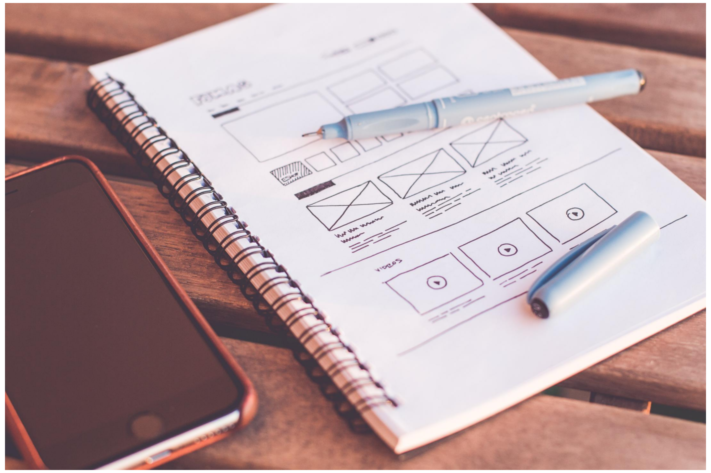
AWS Lambda is easy to:

- Author, deploy, manage, maintain
- Scale and monitor

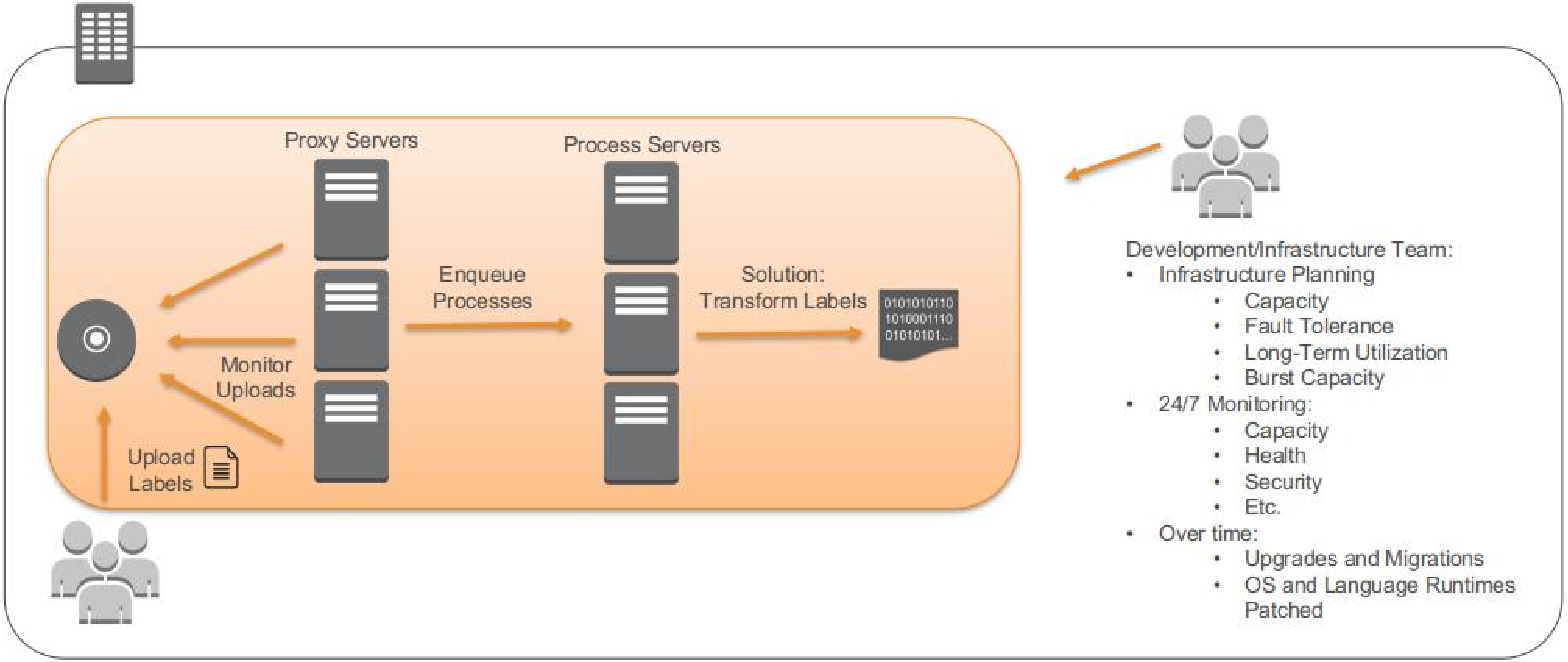
Design pattern applicable to every cloud

Why NOT Lambda?

- Max time is in the minutes, default is in the seconds
- Payload limited to 6 GB, memory to 3 GB
- Rules of scaling, cold start



WHY AWS LAMBDA?



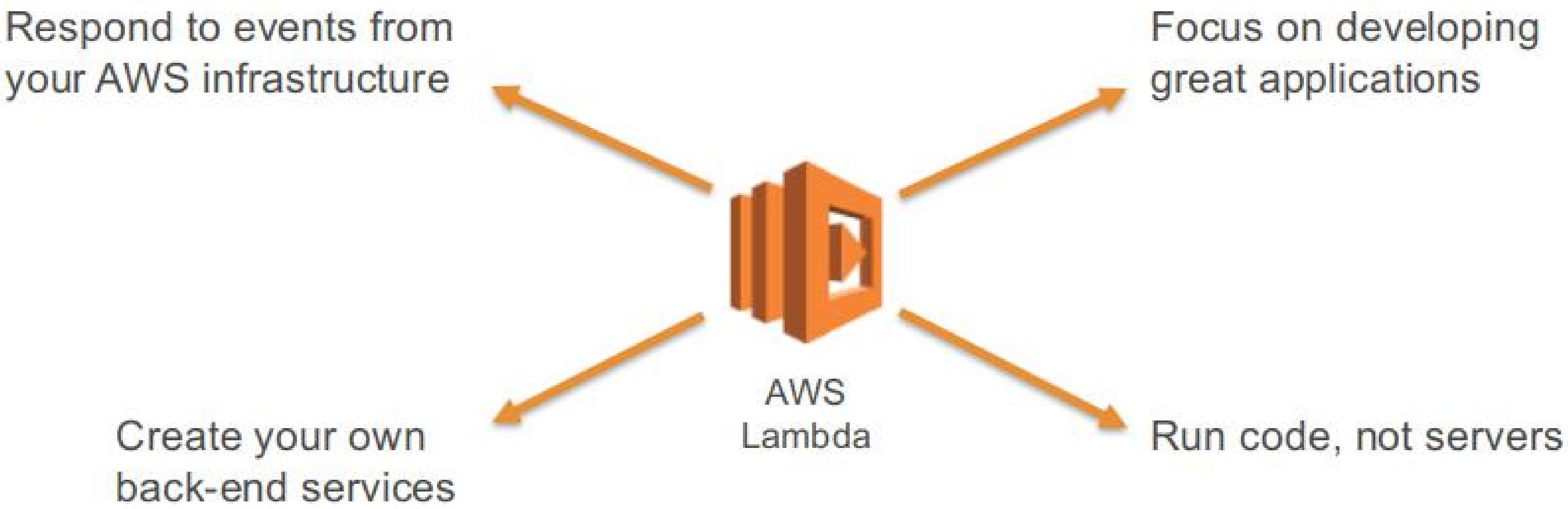
Development/Infrastructure Team:

- Infrastructure Planning
  - Capacity
  - Fault Tolerance
  - Long-Term Utilization
  - Burst Capacity
- 24/7 Monitoring:
  - Capacity
  - Health
  - Security
  - Etc.
- Over time:
  - Upgrades and Migrations
  - OS and Language Runtimes Patched

corporate data center

AWS LAMBDA: OVERVIEW

AWS Lambda: connective tissue for AWS services



Respond to events from your AWS infrastructure

Focus on developing great applications

AWS Lambda

Create your own back-end services

Run code, not servers

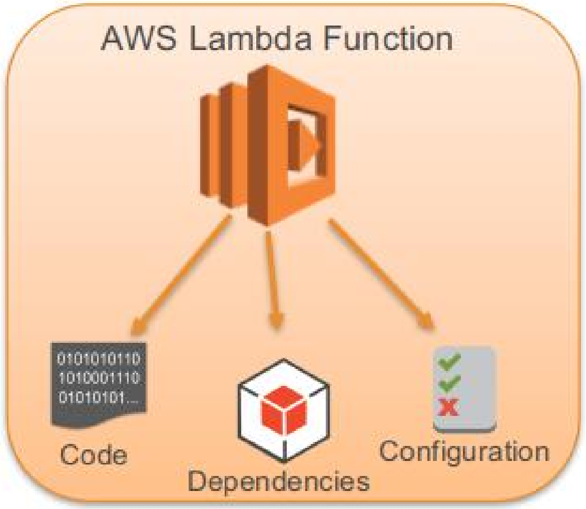
AWS LAMBDA – HOW IT WORKS

Function is invoked by:

- The event source (Push model)
- AWS Lambda (Pull model)
- Direct invocation (RequestResponse model)

Authored in:

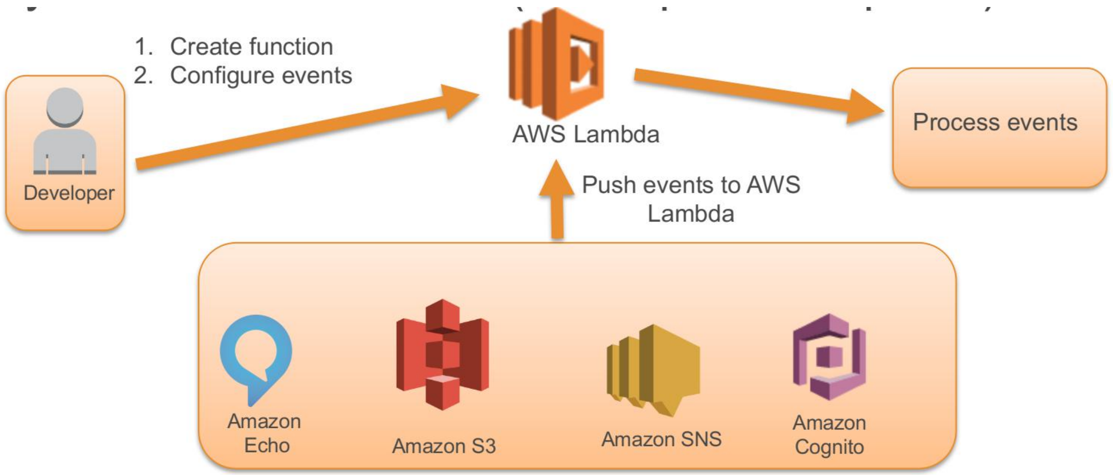
- Java, Go, PowerShell, Node.js, C#, Python, and Ruby
- Runtime API for additional languages



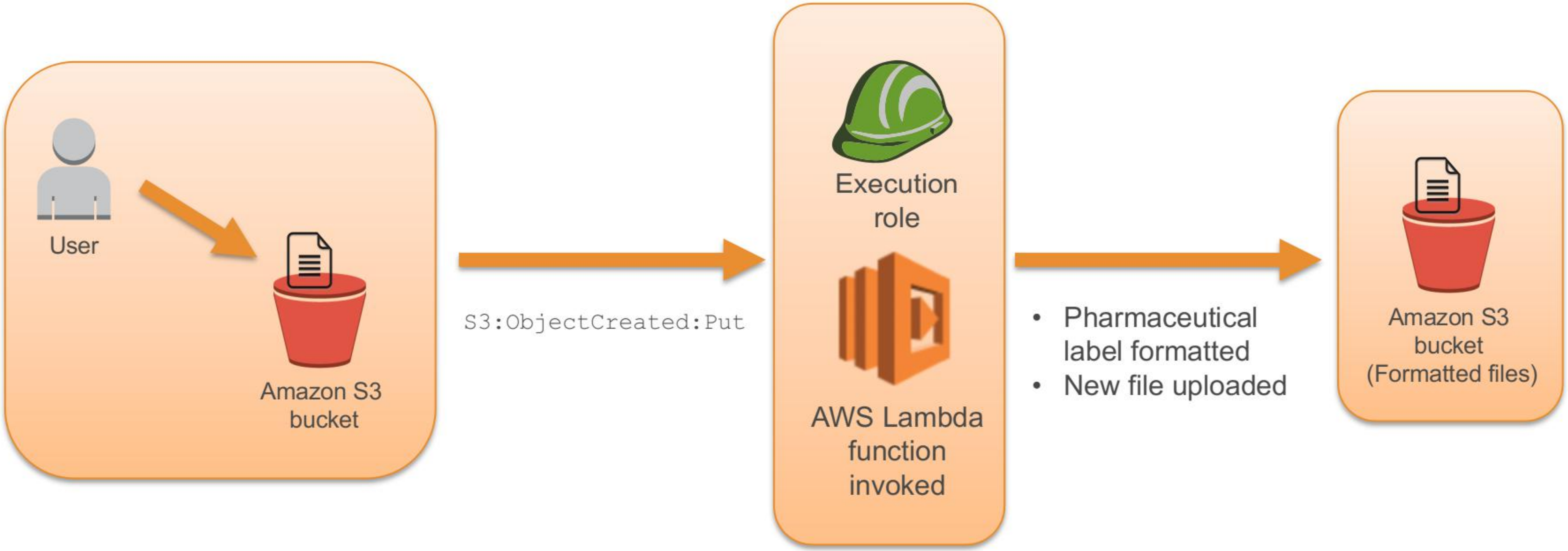
PUSH EVENT MODEL

Event-based invocation where the event source invokes the Lambda function

Asynchronous execution (no response required)



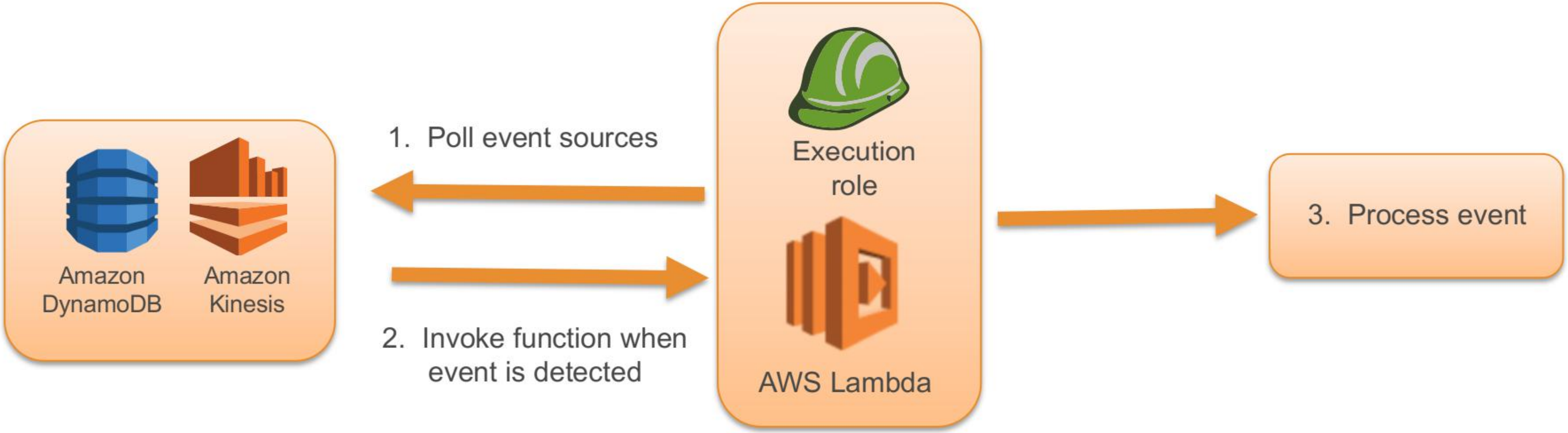
AWS LAMBDA PUSH EVENT MODEL EXAMPLE


User
Amazon S3 bucket
S3:ObjectCreated:Put
Execution role
AWS Lambda function invoked
• Pharmaceutical label formatted
• New file uploaded
Amazon S3 bucket (Formatted files)

PULL EVENT MODEL

Event-based invocation where AWS Lambda polls the event source

AWS Lambda invokes your Lambda function when it detects an event



QUIZ

You have created a simple serverless website using S3, Lambda, API Gateway and DynamoDB. Your website will process the contact details of your customers, predict an expected delivery date of their order and store their order in DynamoDB. You test the website before deploying it into production and you notice that although the page executes, and the lambda function is triggered, it is unable to write to DynamoDB. What could be the cause of this issue?

- A. The availability zone where the DynamoDB is hosted is down
- B. The availability zone where the Lambda is hosted is down
- C. Your lambda function does not have sufficient Identity Access Management (IAM) permissions to write to DynamoDB
- D. You have written your function in Python which is not supported in the runtime of Lambda

QUIZ

In which direction(s) does Lambda scale automatically?

- A. Up
- B. Up and out
- C. Out
- D. None - Lambda does not scale automatically

QUIZ

What AWS service can be used to help resolve an issue with a lambda function?

- A. API Gateway
- B. CloudTrail
- C. AWS X-Ray
- D. DynamoDB

QUIZ

You have created a serverless application to add metadata to images that are uploaded to a specific S3 bucket. To do this, your lambda function is configured to trigger whenever a new image is created in the bucket. What will happen when multiple users upload multiple different images at the same time?

- A. Multiple instances of lambda function will be triggered, one for each image
- B. A single lambda functions will be triggered, which will process all images at the same time
- C. Multiple lambda functions will trigger, one after the others, until all images are processed
- D. A single lambda function will be triggered, that will process all images that have finished uploading one at a time

QUIZ

As a DevOps engineer you are told to prepare complete solution to run a piece of code that required multi-threaded processing. The code has been running on an old custom-built server based around a 4 core Intel Xeon processor. Which of these best describe the AWS compute services that could be used?

- A. EC2, ECS, and Lambda
- B. ECS and EC2
- C. None of the above
- D. Only and EC2 'bare metal' server

AZURE FUNCTIONS

Abstraction of servers:

- Serverless computing abstracts the servers you run on.

Event-driven scale:

- Serverless computing fits for workloads that respond to incoming events.
  - Timers, for example, if a function needs to run every day at 10:00 AM UTC.
  - HTTP, for example, API and webhook scenarios.
  - Queues, for example, with order processing.
  - And more

Micro-billing

- Pay only for the time their code runs

Functions:

- you write code to complete each step.

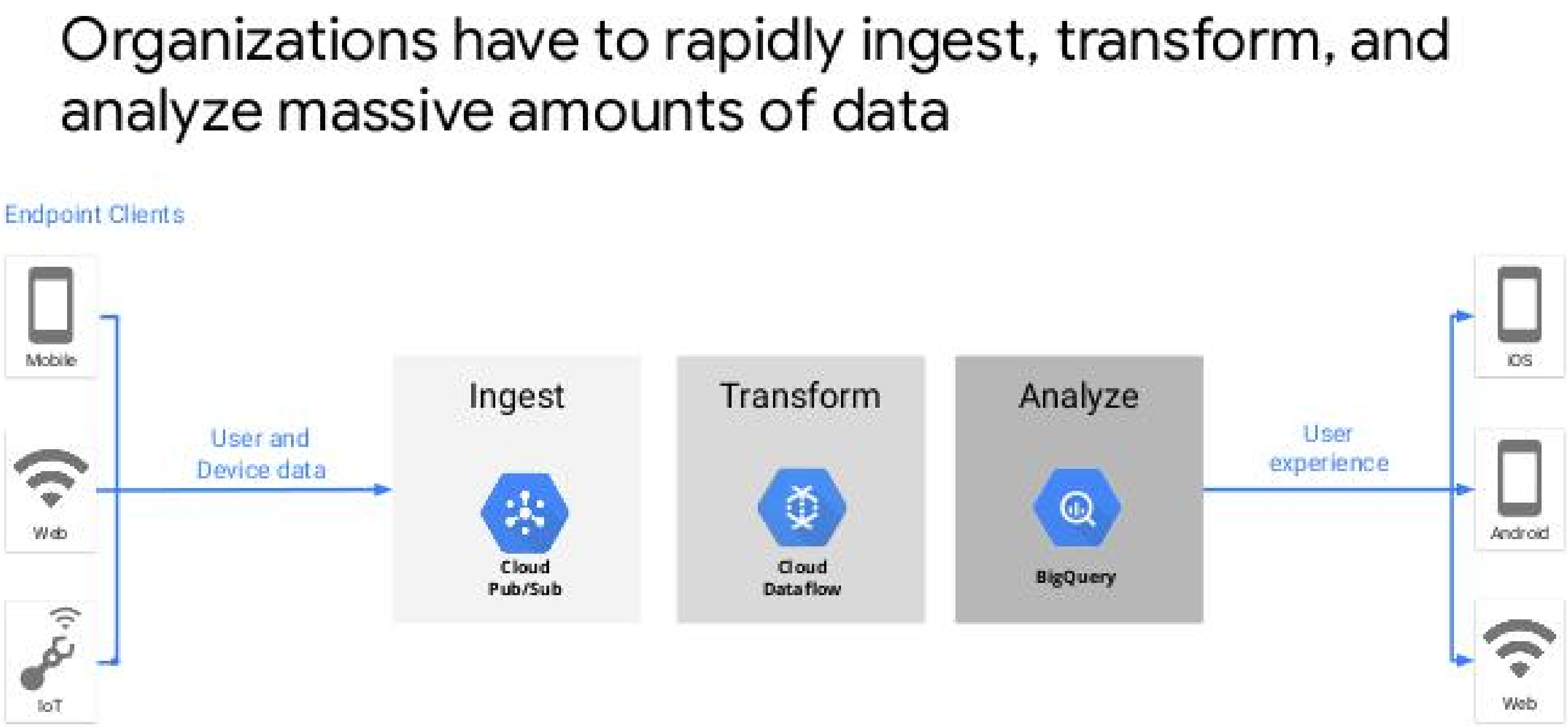
Logic Apps

- you use a GUI to define the actions and how they relate to one another.

FUNCTIONS VS LOGIC APPS

	Functions	Logic Apps
State	Normally stateless, but Durable Functions provide state.	Stateful.
Development	Code-first (imperative).	Designer-first (declarative).
Connectivity	About a dozen built-in binding types. Write code for custom bindings.	Large collection of connectors. Enterprise Integration Pack for B2B scenarios. Build custom connectors.
Actions	Each activity is an Azure function. Write code for activity functions.	Large collection of ready-made actions.
Monitoring	Azure Application Insights.	Azure portal, Log Analytics.
Management	REST API, Visual Studio.	Azure portal, REST API, PowerShell, Visual Studio.
Execution context	Can run locally or in the cloud.	Runs only in the cloud.

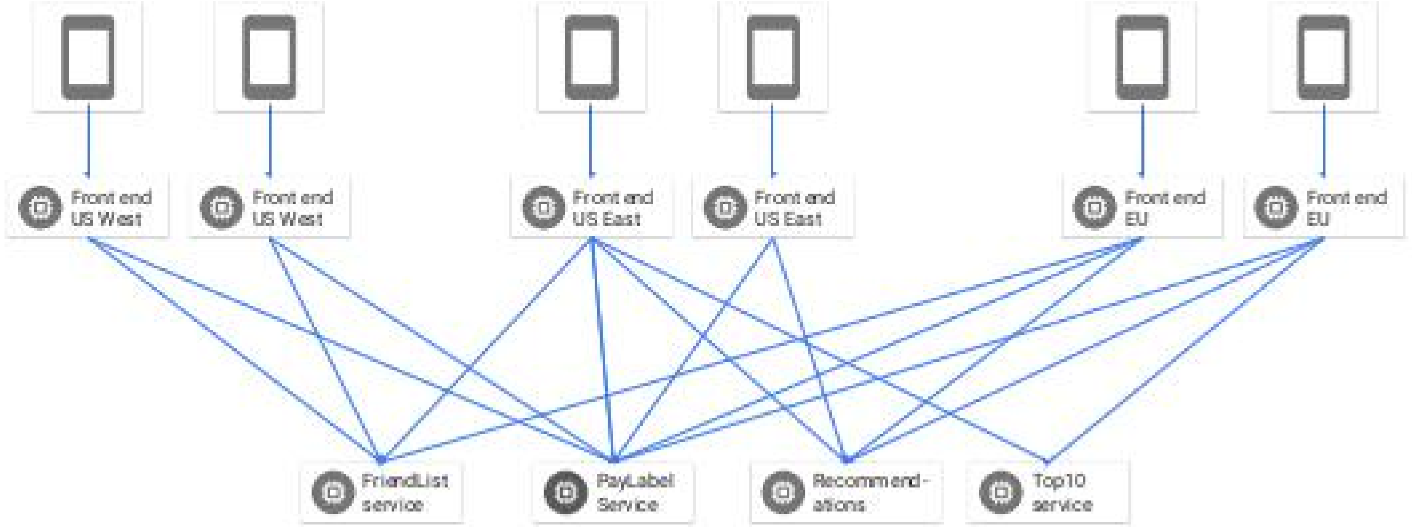
GCP CLOUD FUNCTIONS



Organizations have to rapidly ingest, transform, and analyze massive amounts of data

CLOUD FUNCTIONS USE CASES

Organizations have to orchestrate complex business processes



CONGRATS ON COMPLETION