MAS DSE 230
Scalable Analytics
Cloud Computing

Mai H. Nguyen

CLOUD ANALYTICS

- Introduction to Cloud Computing
- AWS Services
- Amazon EMR
- Amazon EMR Exercise
- Assignments

CLOUD ANALYTICS

- Introduction to Cloud Computing
- AWS Services
- Amazon EMR
- Amazon EMR Exercise
- Assignments

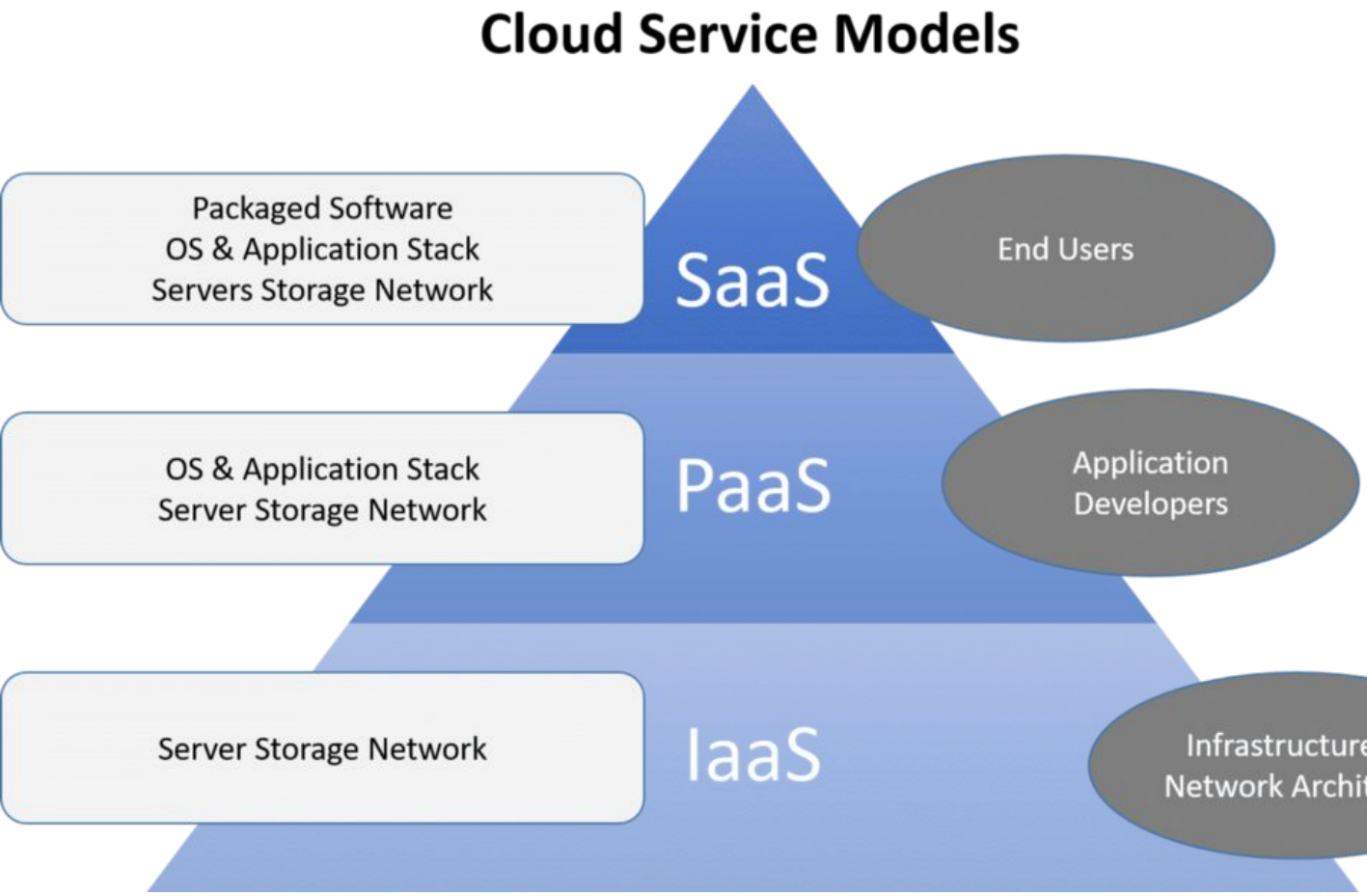
WHAT IS CLOUD COMPUTING?

- On-demand delivery of IT resources
  - Compute power, memory, database storage, operating systems, networking, applications
- How it works
  - Cloud services platform owns and maintains hardware and software needed for services
  - User requests HW/SW needed for application and uses them over web interface
- Compute, storage, memory, networking, etc. exist on remote servers
  - Made available through **virtualization**
  - Rented by application users
- Provides simple & flexible way to have access to computational resources

BENEFITS OF CLOUD COMPUTING

- No capital expenses to build up infrastructure
  - Don’t have to invest in servers, etc.
  - Costs include infrastructure research, build-up, integration, & maintenance
- Elasticity
  - Can dynamically add or reduce capacity based on actual workload’s demand. No capacity guessing
- Pay-as-you-go
  - Pay for actual usage, with granularity from seconds to years
- Ease of deployment
  - Fast deployment; can be globally if needed
- Users can focus on application/business
  - Infrastructure is managed by cloud provider

TYPES OF CLOUD COMPUTING



https://medium.com/@IDMdatasecurity/types-of-cloud-services-b54e5b574f6

End-user applications

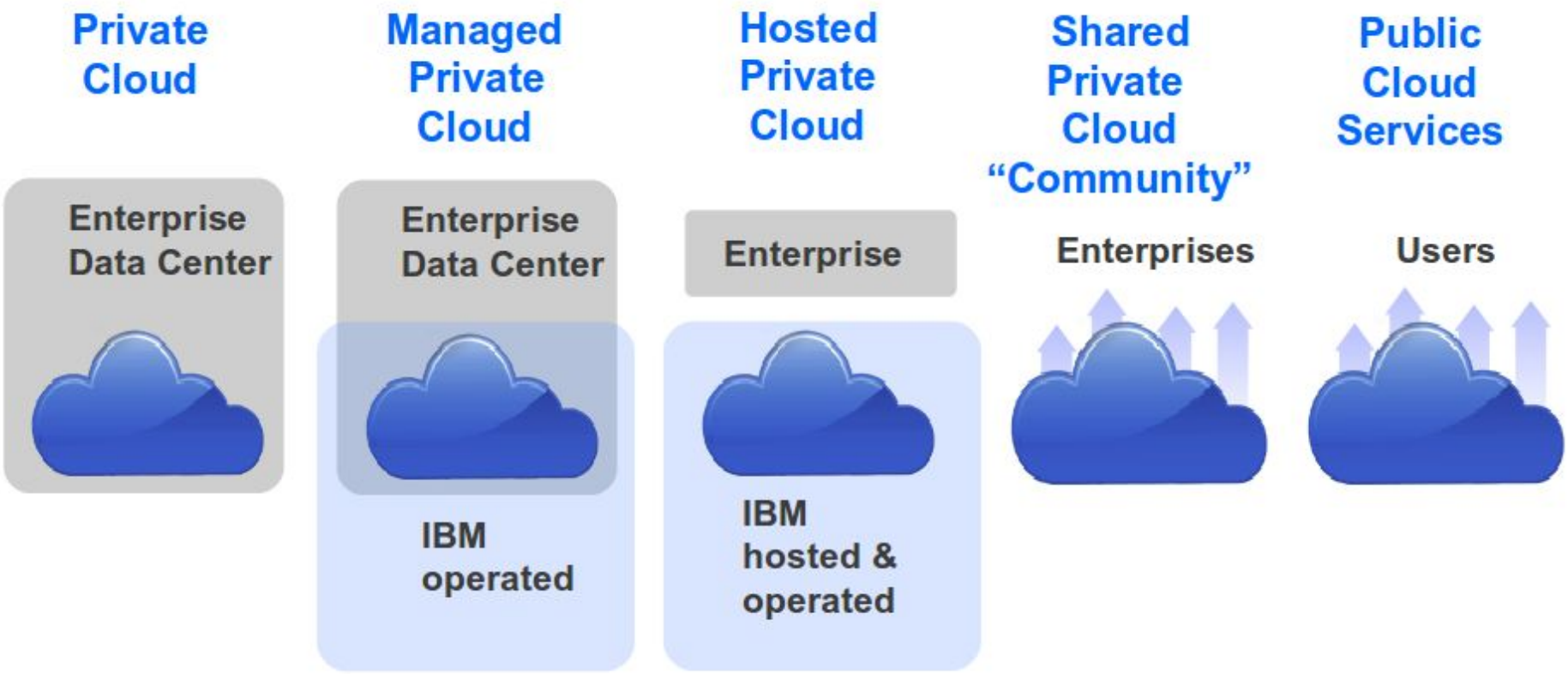
Computing platforms (operating system, execution environment, web server, …)

Computing infrastructure (servers, data storage space, networking, …)

COMPUTING DEPLOYMENT MODELS

- On-Premise
  - Resources are physically located at or close to company
  - Provides dedicated resources
- Cloud
  - Resources are provided and managed by cloud service provider
  - Resources can be scaled on-demand
- Hybrid
  - Connect existing resources with cloud-based resources
  - Provides extension to existing infrastructure

PRIVATE VS PUBLIC CLOUD SERVICES



https://emarcus.net/tag/integration/

CLOUD SERVICE PROVIDERS

- Amazon Web Services
- Microsoft Azure
- Google Cloud Platform
- IBM Cloud Services
- Alibaba Cloud
- Oracle
- Dell Technologies/VMWare
- Cisco Systems
- Others …

CLOUD ANALYTICS

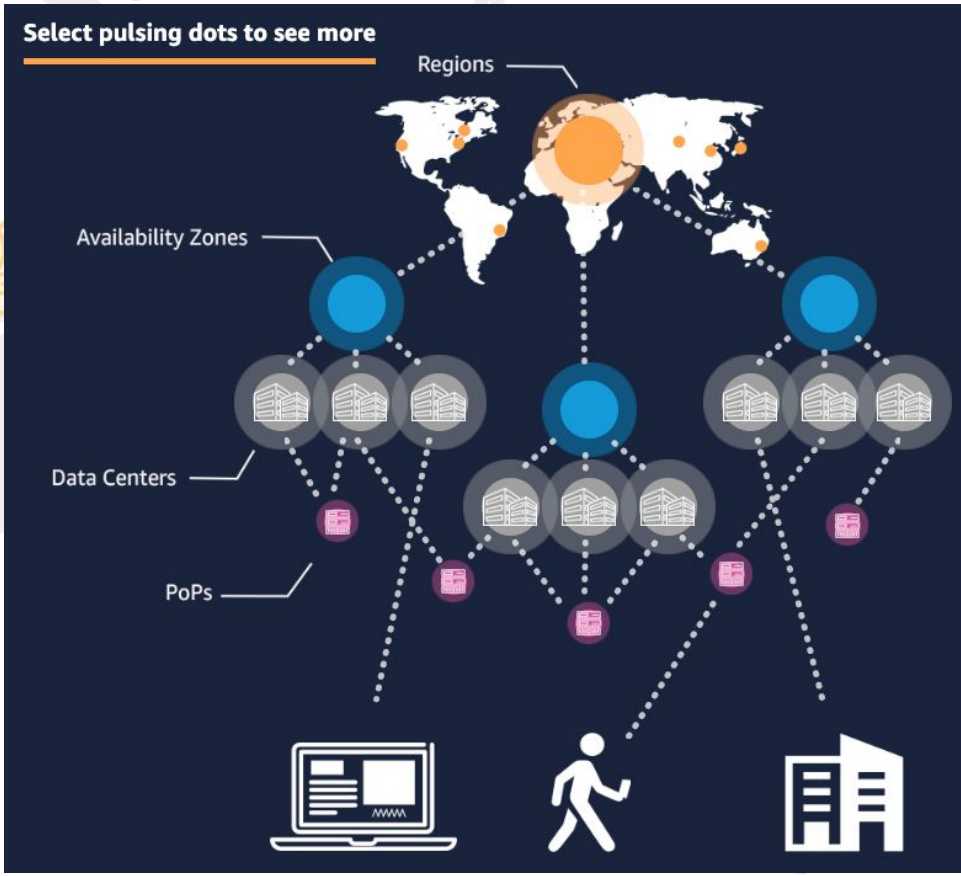
- Introduction to Cloud Computing
- AWS Services
- Amazon EMR
- Amazon EMR Exercise
- Assignments

AWS OVERVIEW

- History
  - Subsidiary of Amazon
  - Started in 2002. Launched in 2006.
- Offers cloud-based products
  - Compute (CPU/GPU), storage, databases, analytics, networking, mobile, developer tools, management tools, IoT, and security
- Customers
  - 100,000s of businesses
  - 1,000,000s of customers
  - 100s of countries

AWS GLOBAL INFRASTRUCTURE

- AWS Region
  - Geographical location
  - Has multiple Availability Zones
- Availability Zone
  - AZs isolated from each other
  - Has >=1 Data Centers
- Data Center
  - Has redundant power, dedicated connectivity
  - Housed in separate facilities
- Points of Presence
  - Edge Locations + Regional Edge Cache servers
  - Sites to cache data and reduce latency



AWS GLOBAL INFRASTRUCTURE

- AWS
  - Currently: 77 Availability Zones within 24 Regions worldwide
    - More are planned
  - 245 countries and territories
  - 220+ PoPs

AWS SERVICES

- Services
  - > 200
  - Most available through APIs
- Fee Structure
  - Based on resources used and required availability, redundancy, and security
  - Different pricing for each service
- Accessing Services
  - AWS Management Console
    - Web-based UI
    - AWS Console Mobile App for mobile devices
  - AWS Command Line Interface (CLI)
    - Access services from command line and scripts
  - Software Development Kits (SDK)
    - Use AWS services in applications via API

AWS SERVICES


Business Applications


Compute


Customer Engagement


Database


Developer Tools


End User Computing


Game Tech


Internet of Things


Machine Learning


Management & Governance


Media Services


Migration & Transfer


Mobile


Networking & Content Delivery


Robotics

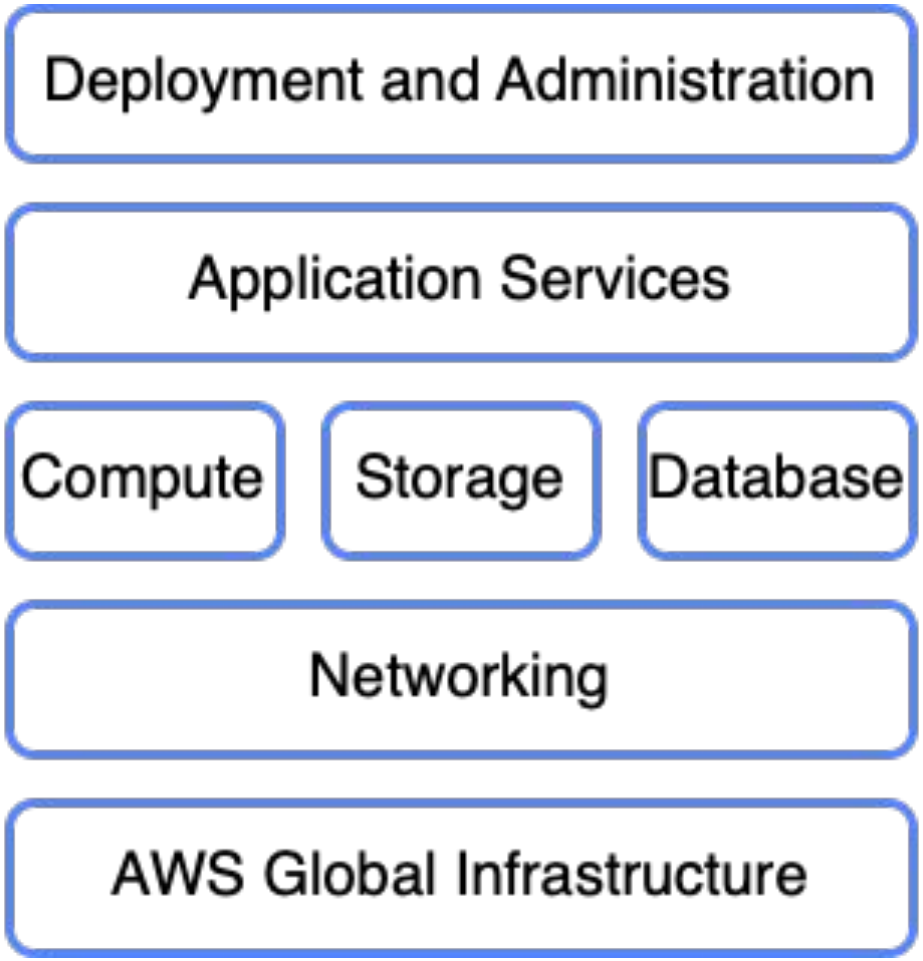

Satellite


Security, Identity & Compliance

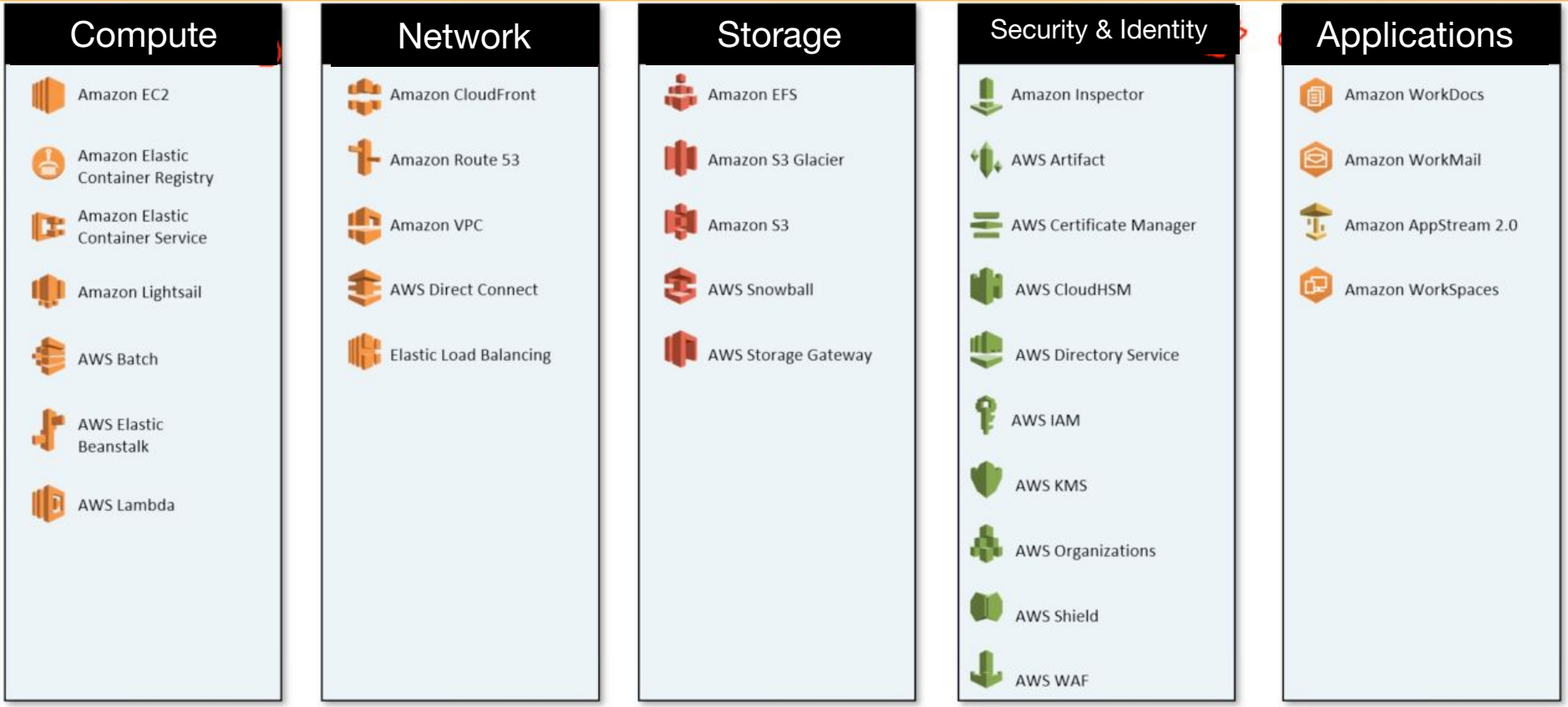

Storage


See All Products

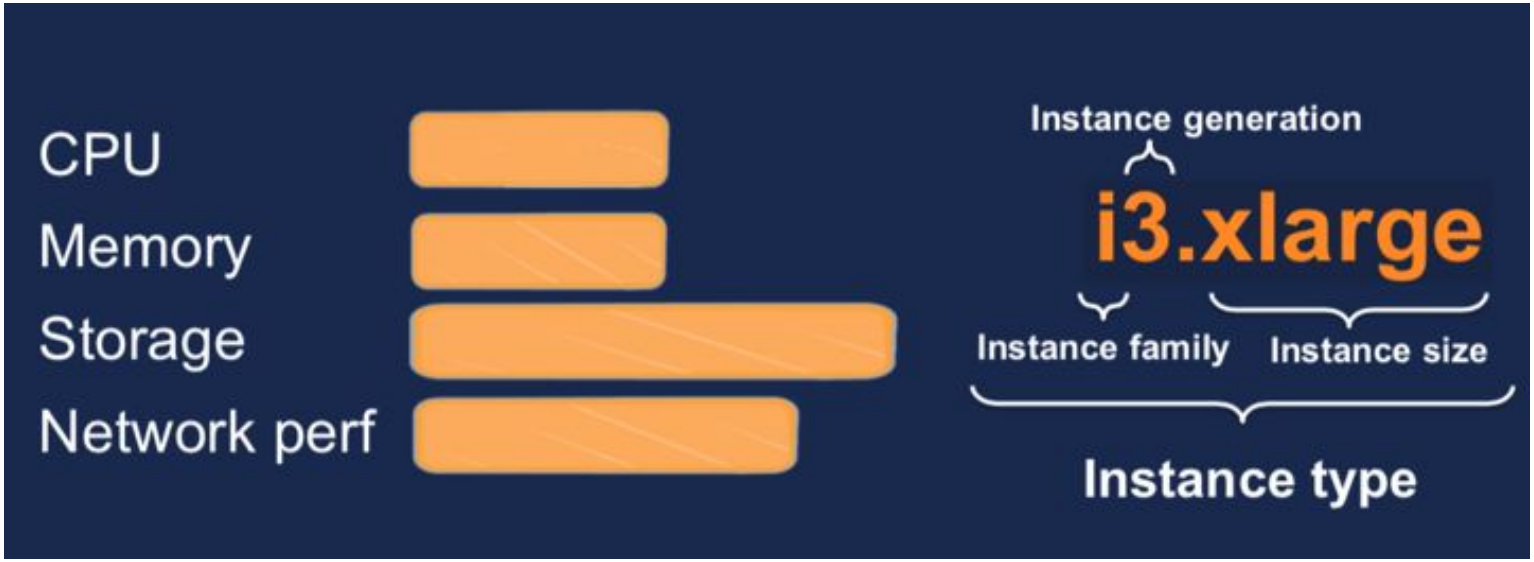
AWS SERVICES



AWS FOUNDATION SERVICES



AWS EC2 INSTANCE TYPES



- **EC2 = Elastic Compute Cloud**
  - Provides virtual servers in the cloud
- **Instance**
  - = virtual private server
  - Starts up with an Amazon Machine Image (AMI)
- **“Elastic”**
  - Configurable compute capacity

EC2 INSTANCE GENERATION & SIZE

C5n Instances

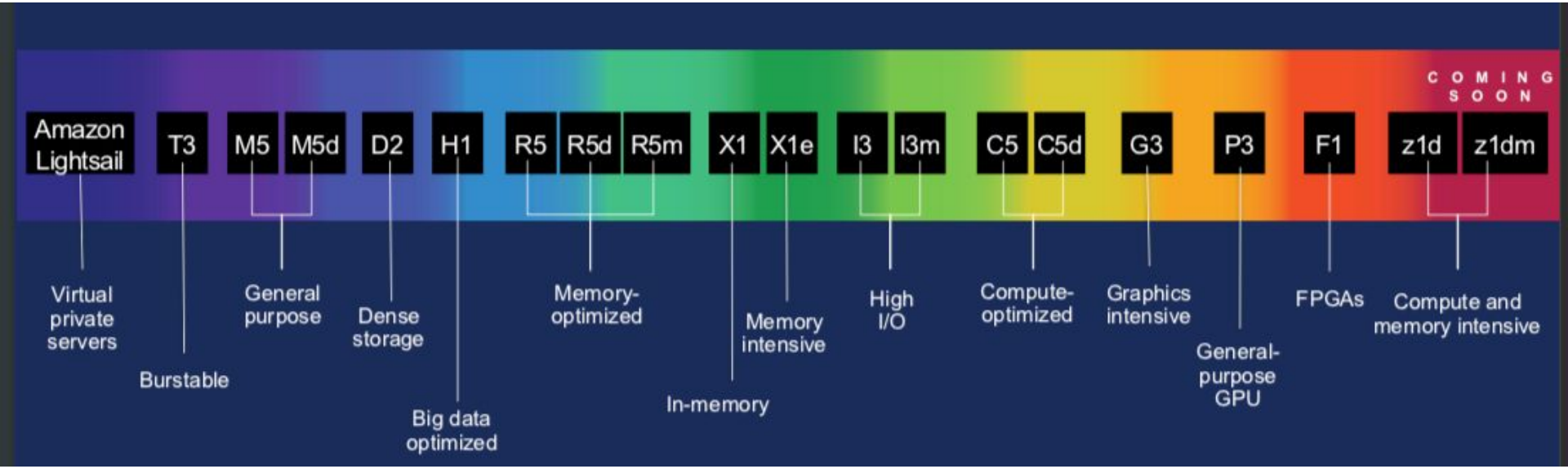
- 3.0 GHz Intel Xeon Platinum (Skylake) CPUs
- Up to 72 vCPUs
- Up to 192 GiB memory
- Up to 14k Mbps dedicated EBS bandwidth
- 25-100 Gbps network bandwidth

C4 Instances

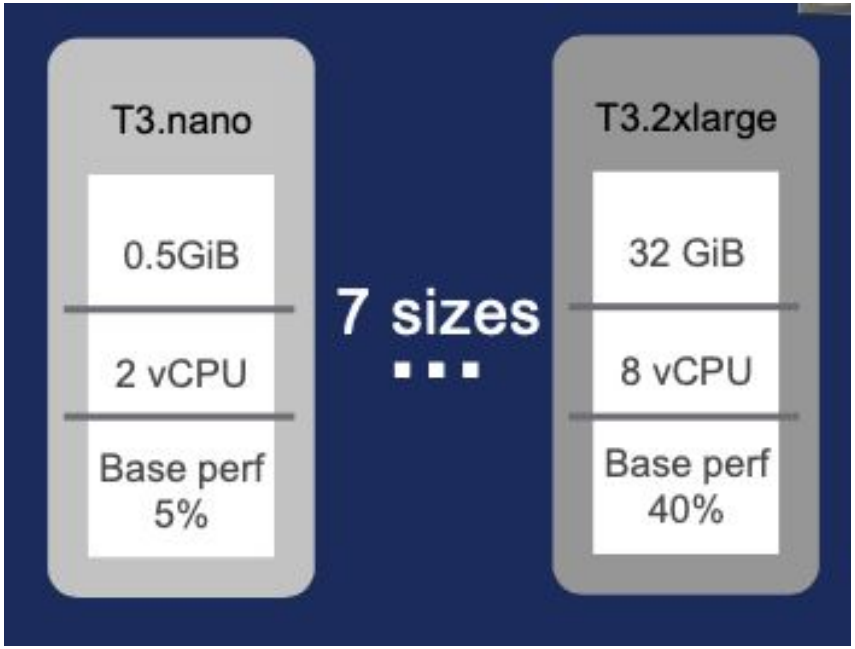
- 2.9 GHz Intel Xeon E5-2666 v3 (Haswell) CPUs
- Up to 36 vCPUs
- Up to 60 GiB memory
- Up to 4k Mbps dedicated EBS bandwidth
- Up to 10 Gbps network bandwidth

	vCPUs	Memory (GiB)	Network (Gbps)
large	2	5.25	Up to 25
xlarge	4	10.5	Up to 25
2xlarge	8	21	Up to 25
4xlarge	16	42	Up to 25
9xlarge	36	96	50
18xlarge	72	192	100

EC2 INSTANCE TYPES



T3: GENERAL PURPOSE BURSTABLE INSTANCES



- Provides guaranteed level of CPU performance with ability to burst for heavy loads
- Lowest-cost EC2 instance
  - $0.0052 per hour
- T3 Unlimited
  - Burst whenever needed for as long as needed

GPU INSTANCES

- General purpose GPU instances
- P3: NVIDIA Tesla V100
- Up to 1 Petaflop of computational performance
- Machine learning, artificial intelligence, high performance computing

Instance Size	GPUs	Accelerator (V100)	GPU Peer to Peer	GPU Memory (GB)	vCPUs	Memory (GB)	Network Bandwidth	EBS Bandwidth
P3.2xlarge	1	1	No	16	8	61	Up to 10Gbps	1.7Gbps
P3.8xlarge	4	4	NVLink	64	32	244	10Gbps	7Gbps
P3.16xlarge	8	8	NVLink	128	64	488	25Gbps	14Gbps

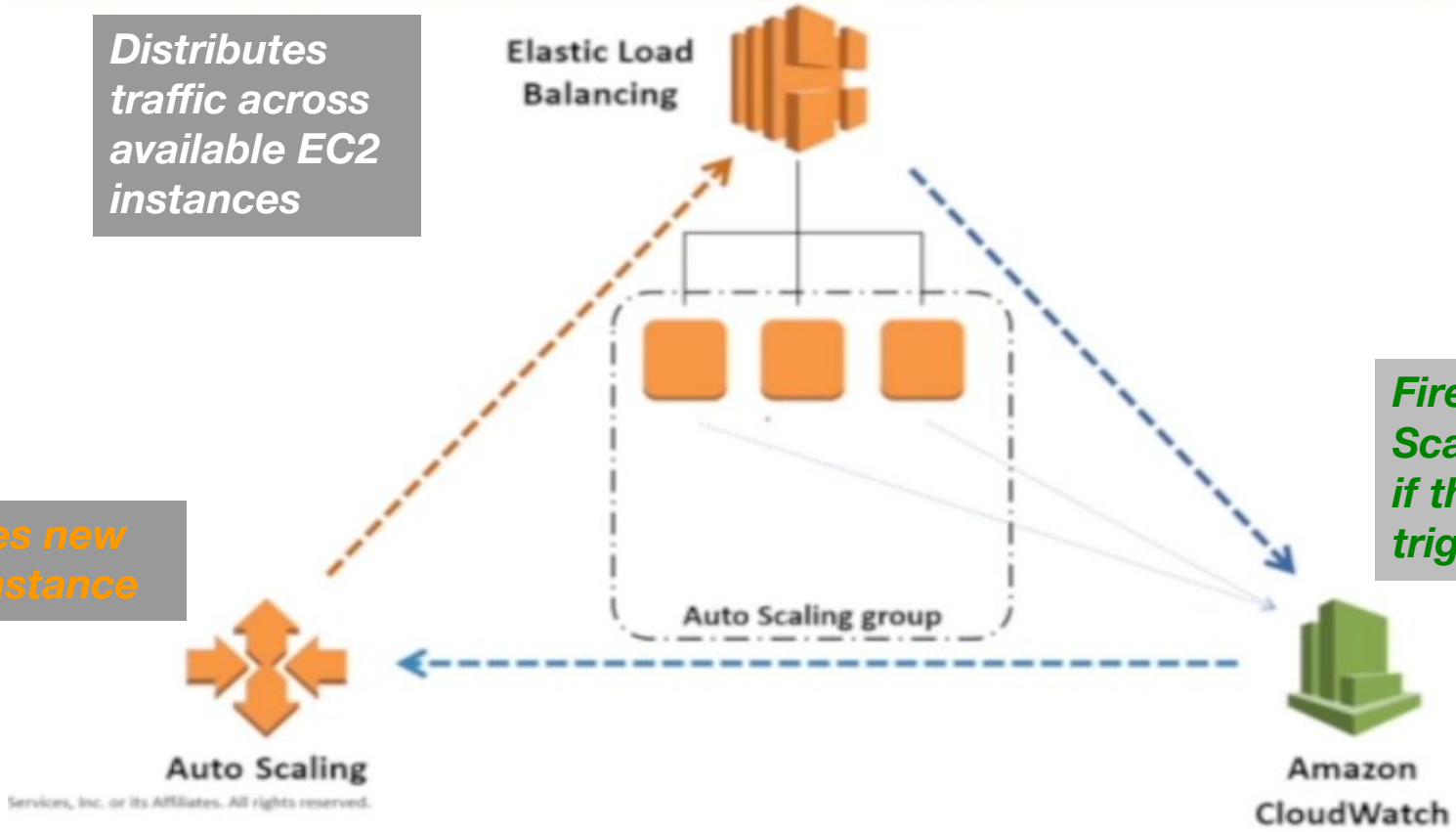
LAUNCHING EC2 INSTANCE

- Determine AWS Region
- Launch EC2 instance from AMI
- Choose instance type
- Configure
  - Network
  - IP address
  - Storage volume
  - Security

AMAZON MACHINE IMAGES (AMIs)

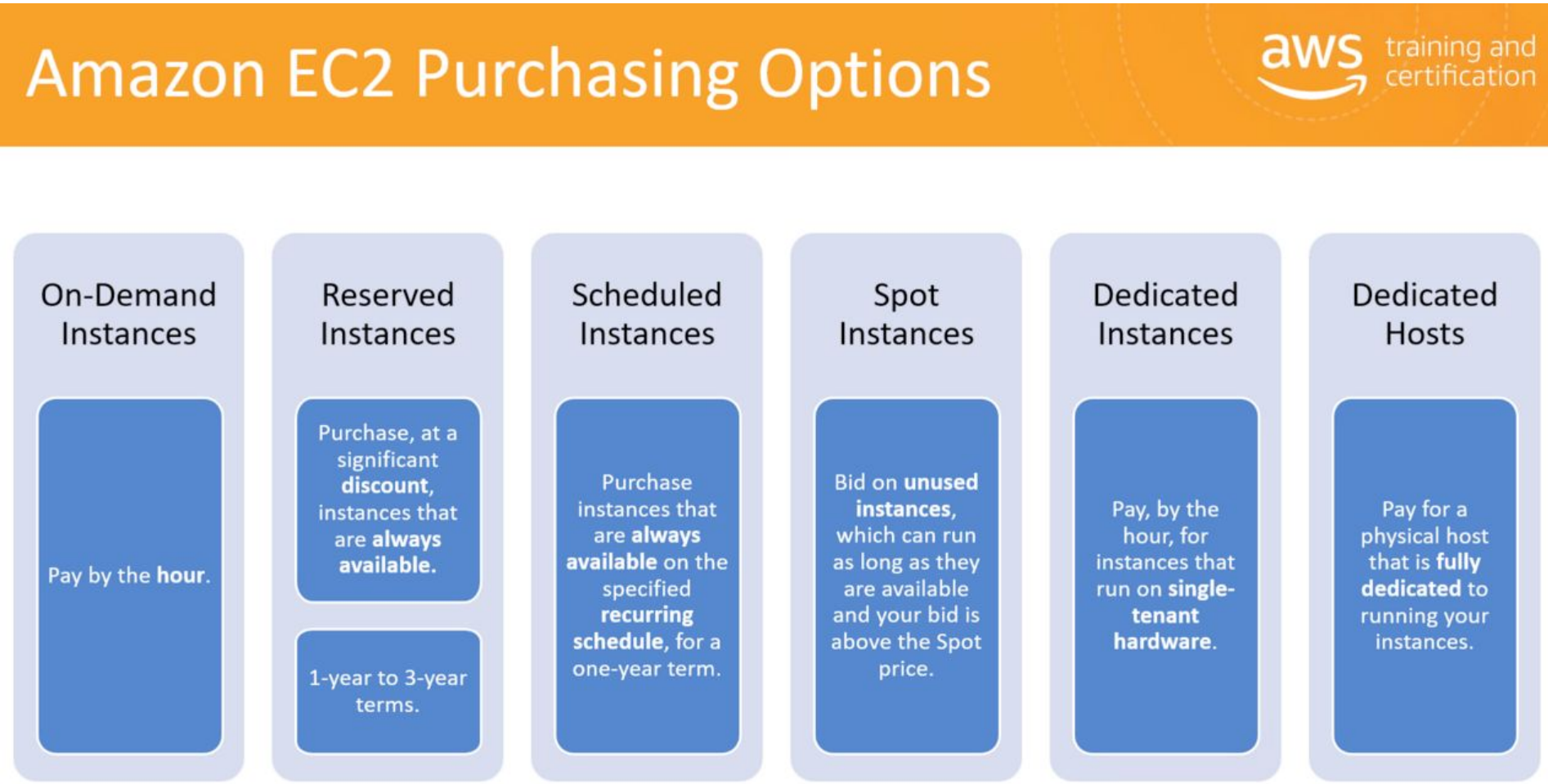
- AMI includes
  - Template for root volume (OS, app server, applications)
  - Permissions that control who can use AMI
  - Block device mapping to specify volumes to attach
- Amazon maintained
  - Kept up-to-date by AWS in each region
- Community maintained
  - Published by other AWS users
  - Managed and maintained by Marketplace partners
- User maintained
  - Created by users
  - Can be kept private or shared with other users

AUTO SCALING



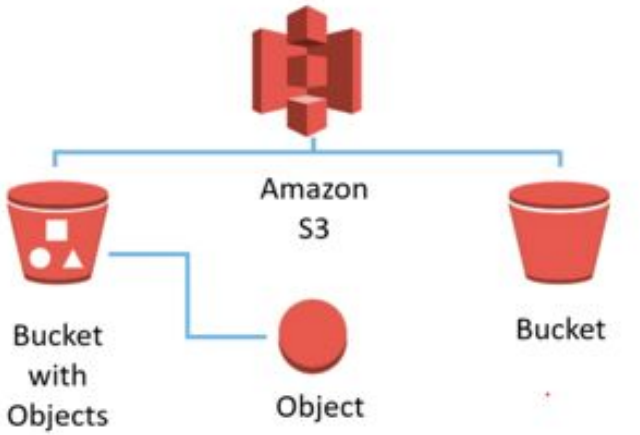
- EC2 Instance Rate Options
  - On-Demand: Specify capacity by the hour
  - Reserved: Reserve dedicated resources
  - Spot: Bid on spare computing capacity
- EC2 Auto Scaling
  - Automatically add/remove EC2 instances
  - Has predictive scaling features

EC2 Purchasing Options



AWS STORAGE SERVICES

S3: SIMPLE STORAGE SERVICE



- **S3**
  - Web-accessible object store with HTTP/S access
  - Data is stored as objects in buckets
  - SLA of 99.99999999999% for durability
- **Object**
  - file + optional metadata
- **Bucket**
  - Basic storage unit
  - Identified by unique, user-assigned key
  - Object can be up to 5 TB in size with 2KB of metadata
  - Can store unlimited number of objects
  - Can have up to 100 buckets in each account
  - User controls access to bucket and its objects

S3 VERSIONING & PRICING

S3 Versioning

- Protects from accidental overwrites and deletes
  - With no performance penalty
- Allows retrieval of deleted objects or roll back to previous version
- Versioning options
  - Un-versioned (default)
  - Versioning-enabled
  - Versioning-suspended

S3 Pricing

- Pay for what you use
- No minimum
- Based on location of bucket
- Pricing available as
  - Storage pricing
  - Request pricing
  - Data transfer pricing: data transferred out of S3

EBS: ELASTIC BLOCK STORAGE

- Persistent block level storage
  - HDD (hard disk drive): large workloads
  - SSD (solid state drive): transactional workloads
- Offers consistent and low-latency storage
  - Virtual hard drive for EC2 instance
- Built-in redundancy
  - Stored data is automatically replicated within its AZ
  - Snapshots can be created and are stored in S3
- Can be encrypted
- Size can be changed on-the-fly
- Pay for what you provision, not what you use
  - Pricing based on region

EBS USE CASES

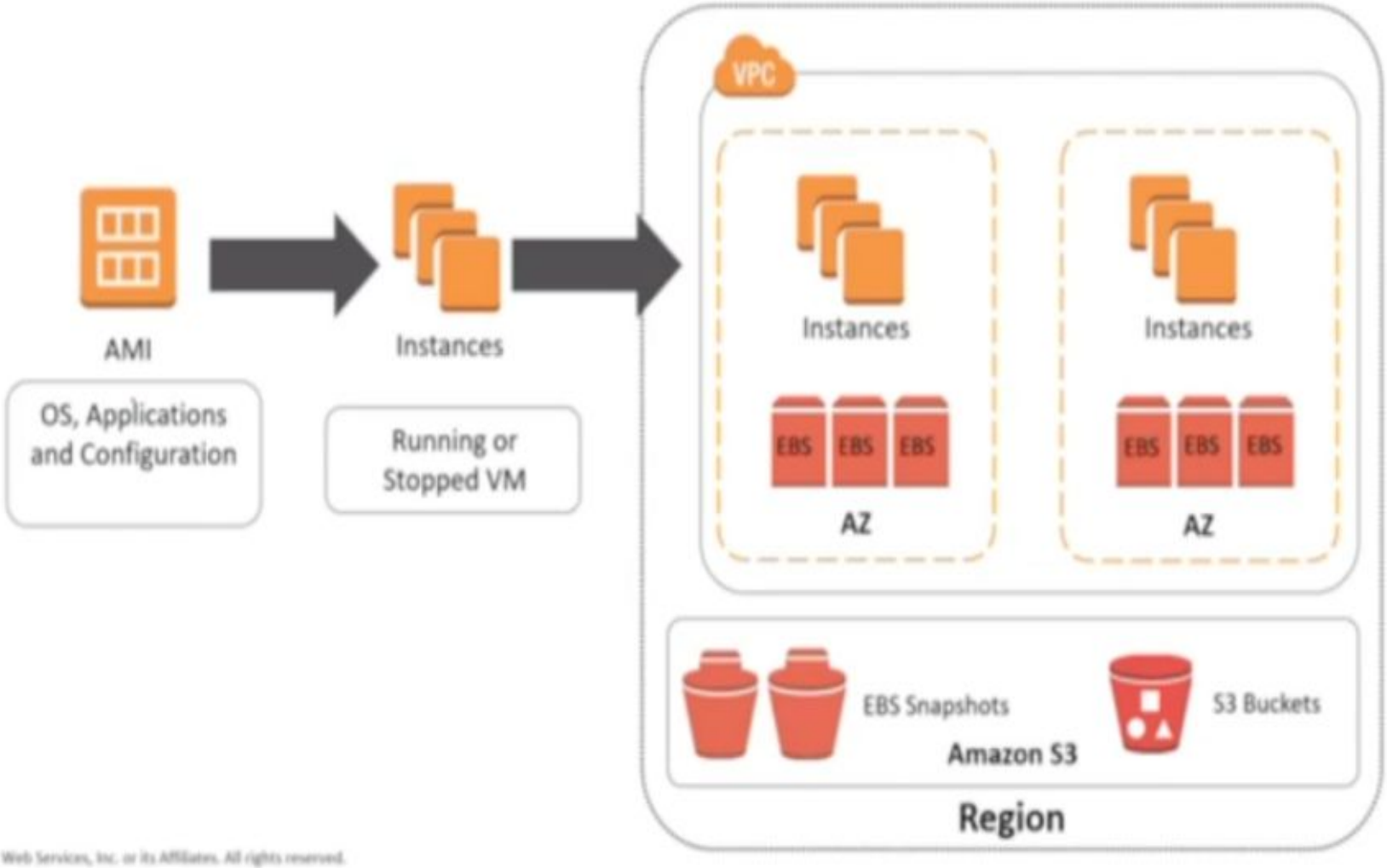
- Operating system
  - Boot volume
- Databases
  - Persistent data stores that can scale with app needs
- Enterprise applications
  - Reliable block storage for mission-critical applications
- Business continuity
  - Back data up using EBS Snapshots

EBS vs S3

	EBS	S3
Paradigm	Block storage with file system	Object store
Performance	Very fast	Fast
Redundancy	Across multiple servers in an Availability Zone	Across multiple facilities in a Region
Security	Encryption	Encryption
Access from Internet	No (1)	Yes (2)
Used as	Disk drive	Online storage

(1) Accessible from Internet if mounted to server and set up as FTP, etc.
(2) Only with proper credentials

EC2 with S3 and EBS



GLACIER

- Used for long-term archiving and long-term backup
- Not available for real-time access
- Optimal for infrequently accessed data
- Several options for access to archives: few minutes to several hours
- Less than $0.01 per GB/month
- Designed for 99.999999999% durability

AWS DATABASE SERVICES

Database type	Use cases	AWS service
Relational	Traditional applications, ERP, CRM, e-commerce	Amazon Aurora, Amazon RDS, Amazon Redshift
Key-value	High-traffic web apps, e-commerce systems, gaming applications	Amazon DynamoDB
In-memory	Caching, session management, gaming leaderboards, geospatial applications	Amazon ElastiCache for Memcached Amazon ElastiCache for Redis
Document	Content management, catalogs, user profiles	Amazon DocumentDB
Wide column	High scale industrial apps for equipment maintenance, fleet management, and route optimization	Amazon Managed Apache Cassandra Service
Graph	Fraud detection, social networking, recommendation engines	Amazon Neptune
Time series	IoT applications, DevOps, industrial telemetry	Amazon Timestream
Ledger	Systems of record, supply chain, registrations, banking transactions	Amazon QLDB

RDS

- Relational Database Service
  - Used to set up and operate relational DB in cloud
  - Resizable capacity
- Manages common DB administrative tasks
  - Updates, backups, etc.
- Works with common DB engines
- Several DB instance types
  - Optimized for memory, performance, I/O
- Database Migration Service
  - Migrate or replicate existing DB

DYNAMODB

- NoSQL database service
  - Key-value and document database
- Features
  - Encrypts all data by default
  - Built-in backup and restore
  - Allocates necessary resources to meet user-specified throughput capacity
  - Can handle > 10 trillion requests per day and support peaks of > 20 million requests per second

REDSHIFT

- Data warehouse service
- Features
  - Handles large-scale datasets (PBs)
  - Supports large-scale database migrations
  - Supports large-scale data analysis queries
  - Connects with ODBC and JDBC
- Architecture
  - Column-oriented
  - Based on PostgreSQL
  - Massively parallel query execution

AWS APPLICATION SERVICES


Business Applications


Blockchain


Customer Engagement


AR & VR


Developer Tools


End User Computing


Game Tech


Internet of Things


Machine Learning


Management & Governance


Media Services


Migration & Transfer


Mobile


Networking & Content Delivery


Quantum Technologies


Robotics


Satellite

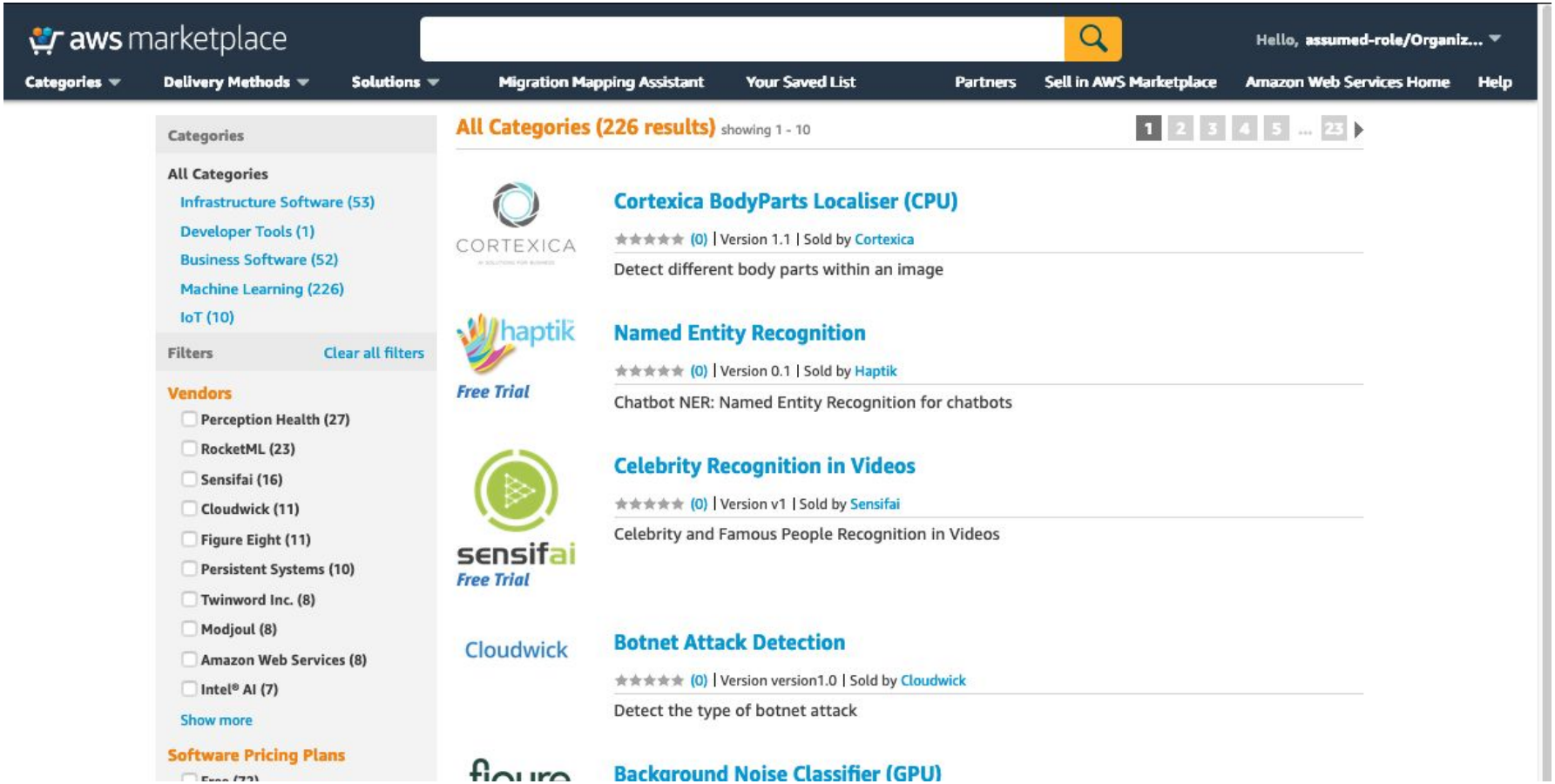

Security, Identity & Compliance


Analytics

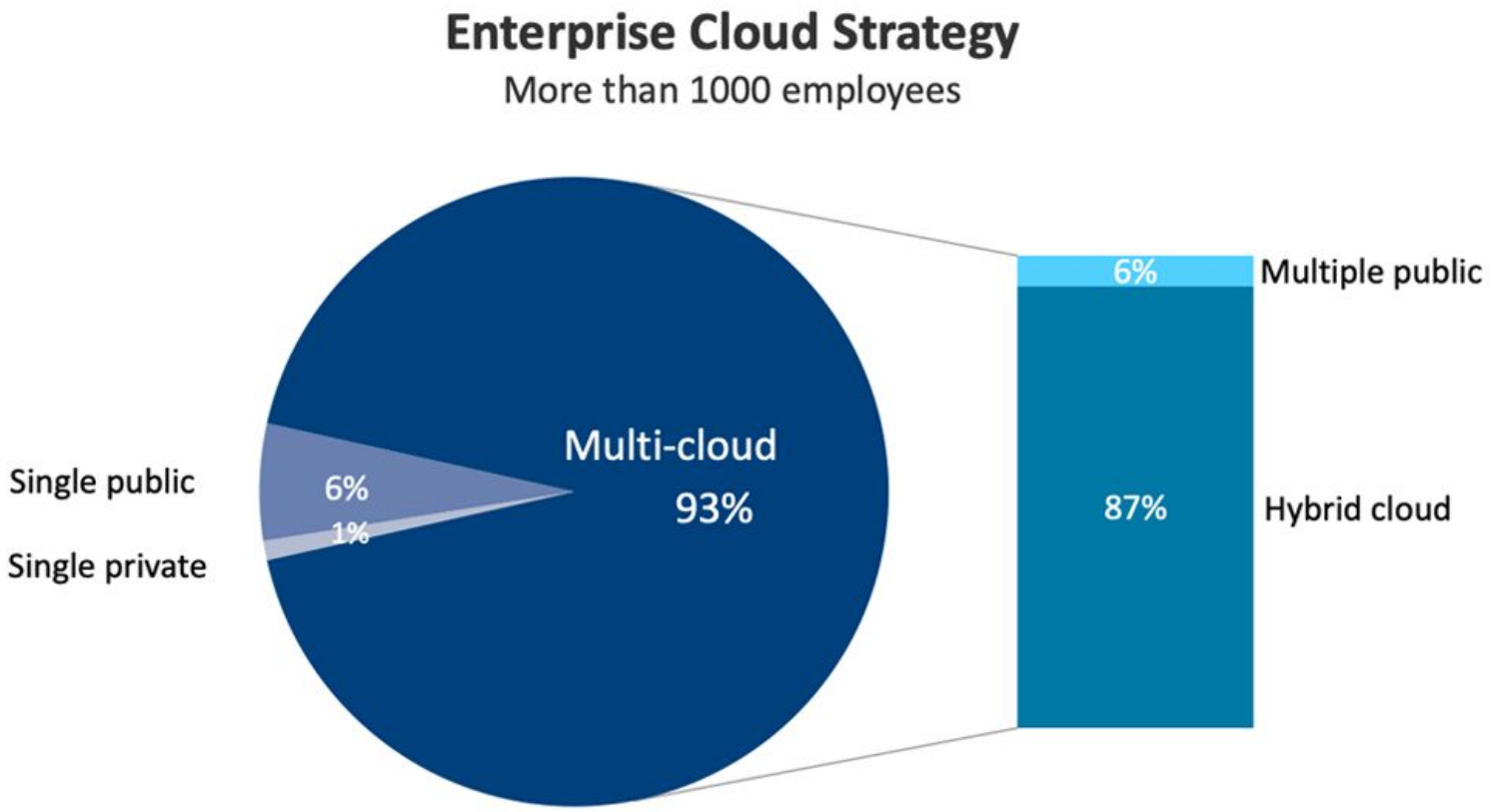
CATEGORIZATION OF AWS SERVICES

- IaaS:
  - Compute: EC2, ECS, Lambda
  - Storage: S3, EBS, EFS, Glacier
  - Networking: CloudFront, VPC
- PaaS:
  - Database/Analytics Systems: Aurora, Redshift, Neptune, ElastiCache, DynamoDB, Timestream, EMR, Athena
  - Blockchain: QLDB; IoT: Greengrass
- SaaS:
  - ML/AI: SageMaker, Elastic Inference, Lex, Polly, Translate, Transcribe, Textract, Rekognition, Ground Truth
  - Business Apps: Chime, WorkDocs, WorkMail

AWS MARKETPLACE

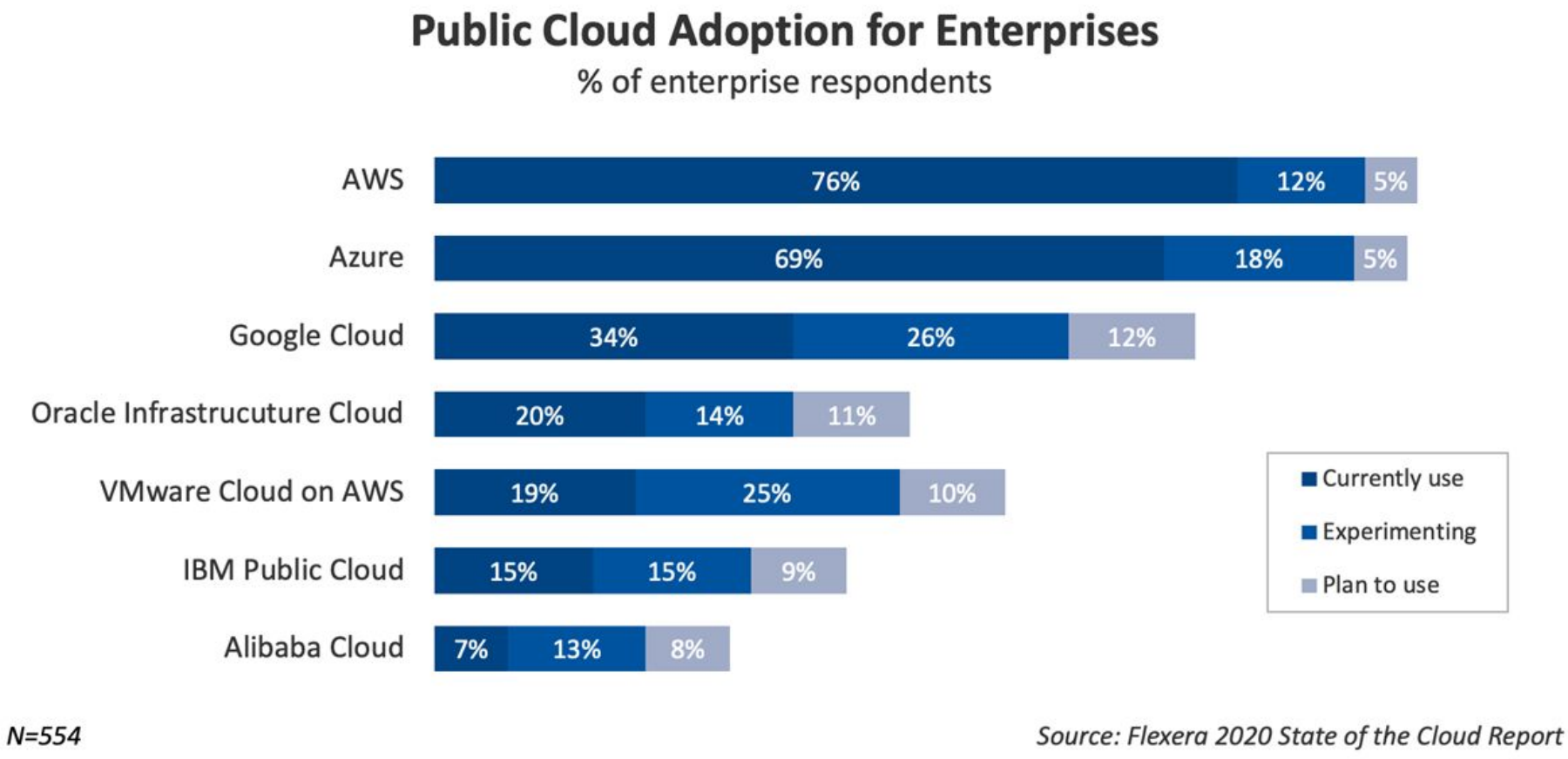


STATE OF THE CLOUD SURVEY



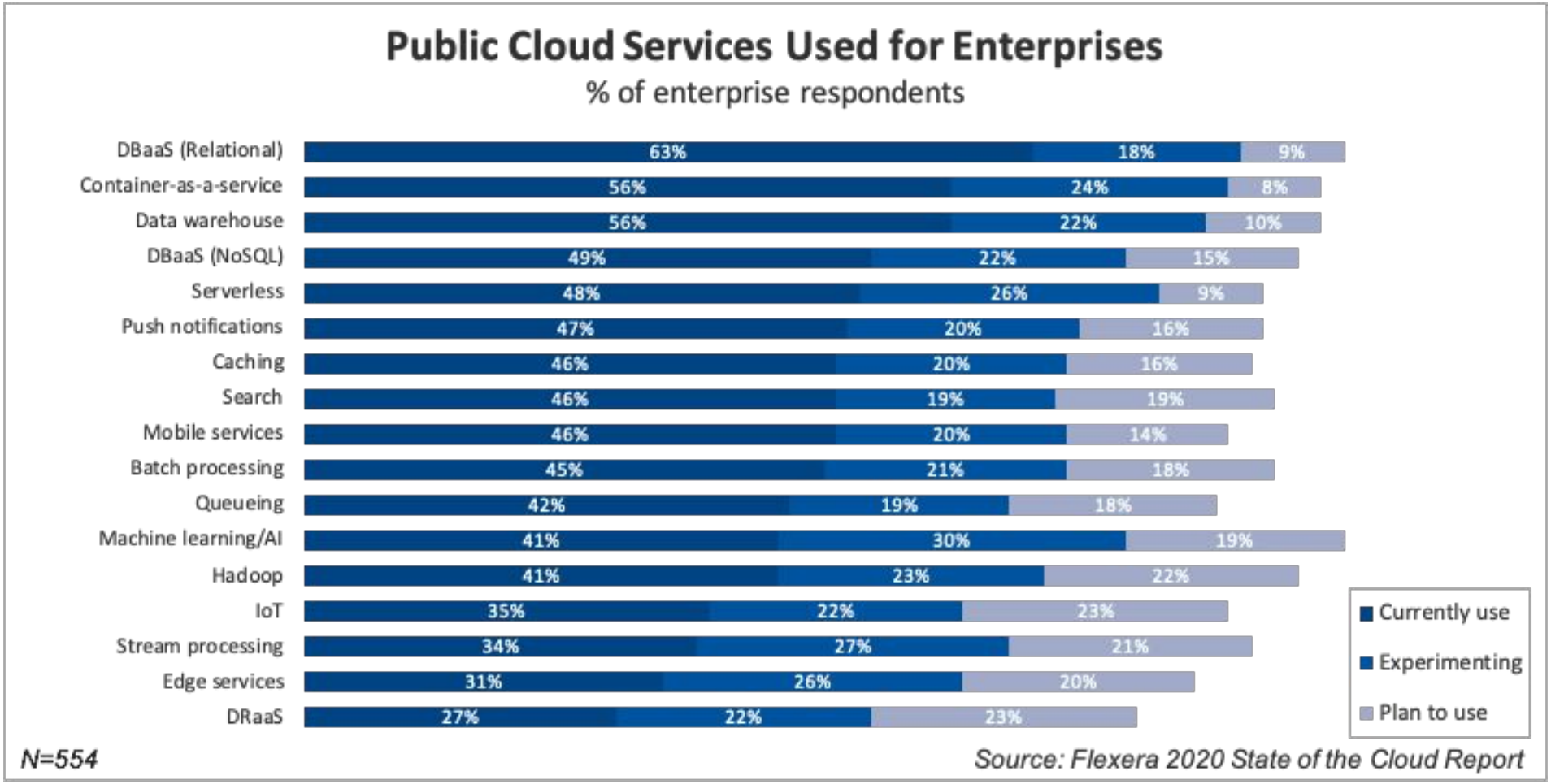
https://info.flexera.com/SLO-CM-REPORT-State-of-the-Cloud-2020

STATE OF THE CLOUD SURVEY



https://info.flexera.com/SLO-CM-REPORT-State-of-the-Cloud-2020

STATE OF THE CLOUD SURVEY



https://www.flexera.com/blog/industry-trends/trend-of-cloud-computing-2020/

STATE OF THE CLOUD SURVEY

https://www.flexera.com/blog/industry-trends/trend-of-cloud-computing-2020/

STATE OF THE CLOUD SURVEY

Top Growing Cloud PaaS Services

% of enterprise respondents

PLACE	SERVICE	2019	2020	GROWTH
1	IoT	29%	35%	21%
2	Container-as-a-service	48%	56%	17%
3	Machine learning/AI	35%	41%	17%
4	Data warehouse	50%	56%	12%
5	Serverless	43%	48%	12%

N=554

Source: Flexera 2020 State of the Cloud Report

https://www.flexera.com/blog/industry-trends/trend-of-cloud-computing-2020/

QUIZ

QUIZ

Which of the following is/are (a) major reason(s) for the increasing adoption of cloud computing?

A. Pay-as-you-go economics
B. Manageability
C. Elasticity
D. Both A & B
E. A, B, & C

QUIZ

What is AWS EC2?

A. Service that provides configurable storage capacity in the cloud
B. Service that provides automatic scaling of computing resources in the cloud
C. Service that provides computing servers in the cloud
D. Service that provides the information required to launch a compute instance in the cloud

QUIZ

What are the different types of cloud computing services?

A. servers, storage devices, networking services
B. SaaS, PaaS, IaaS
C. EC2, S3, EBS
D. IaaS, OaaS, UaaS
E. None of the above

QUIZ

Which of the following AWS service(s) can be considered as an offering of Infrastructure-as-a-Service?

A. S3
B. EC2
C. SageMaker
D. Redshift
E. Both A & B

QUIZ

What does this AWS specification mean: m5.xlarge?

A. 5 EC2 servers with extra large persistent storage
B. 5 S3 buckets with extra large storage capacity
C. EC2 instance with 5 xlarge S3 buckets
D. EC2 general-purpose instance, 5th generation, extra large instance size
E. None of the above

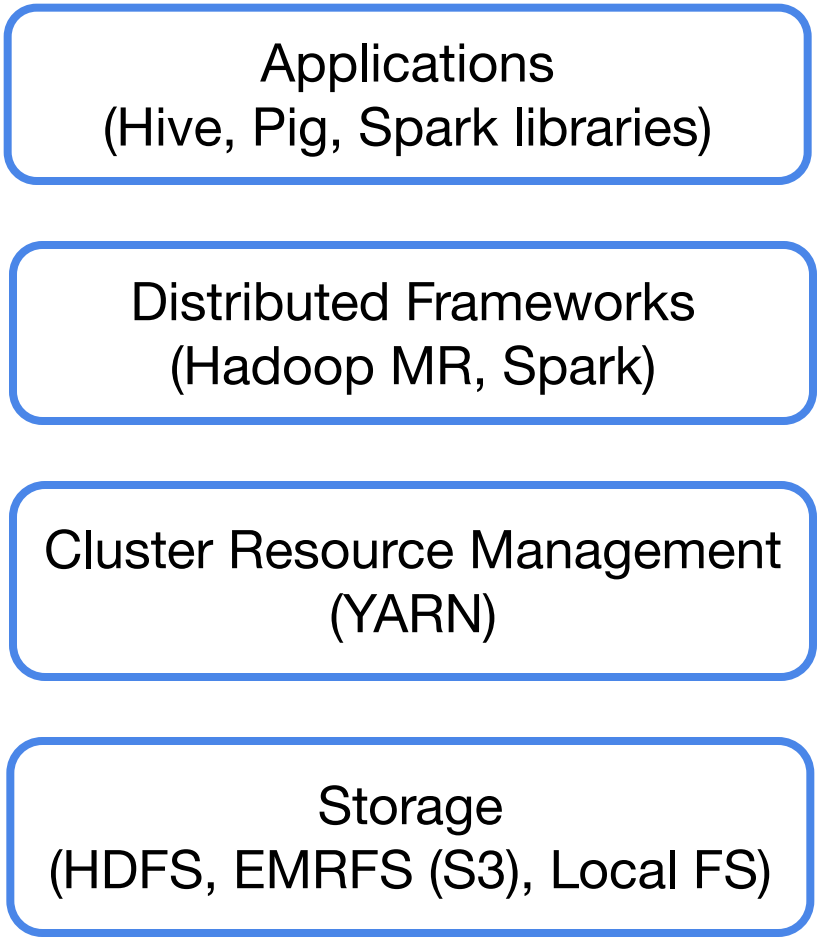
CLOUD ANALYTICS

- Introduction to Cloud Computing
- AWS Services
- Amazon EMR
- Amazon EMR Exercise
- Assignments

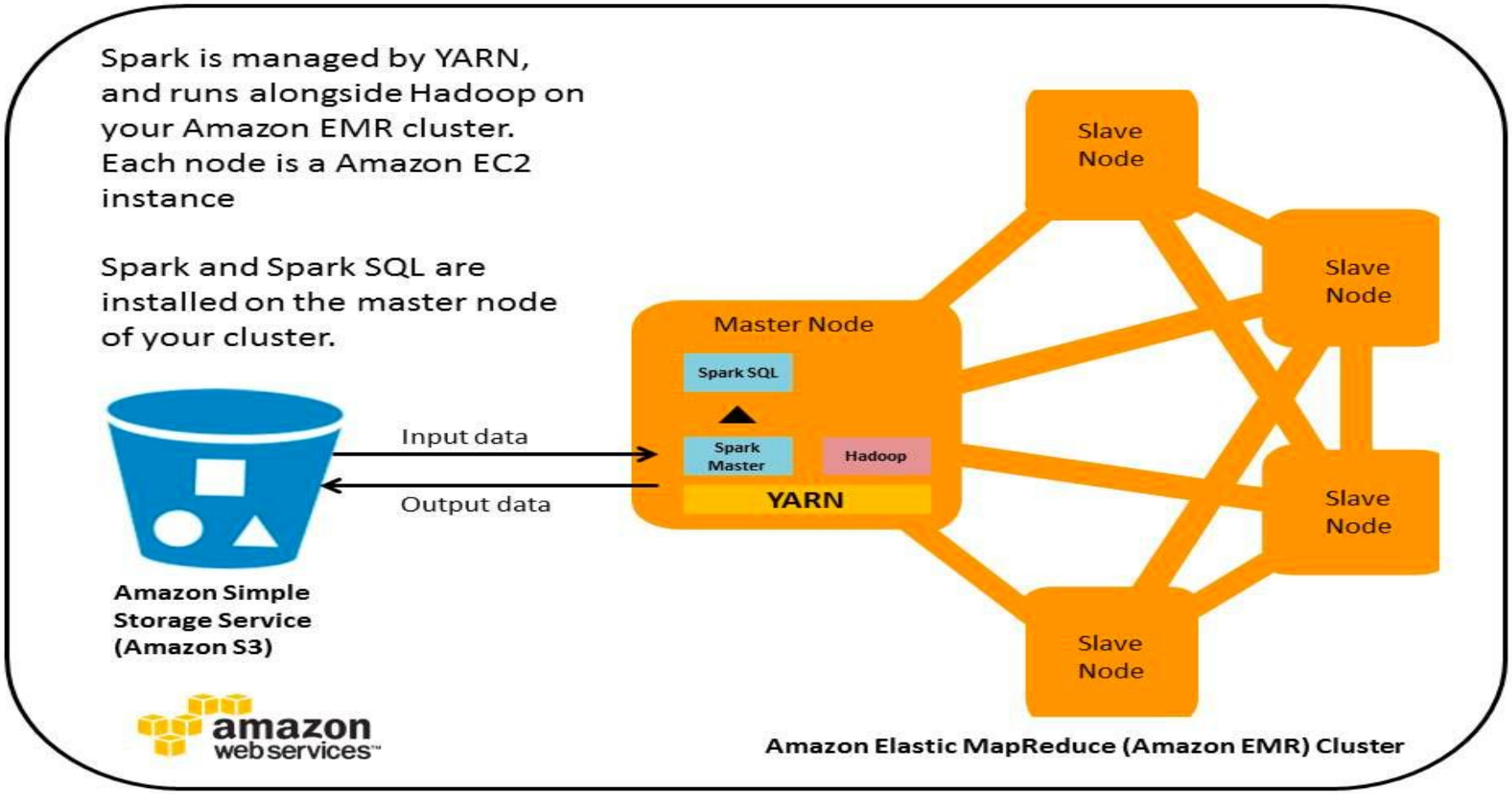
AMAZON EMR

- Elastic Map Reduce
  - Managed cluster platform for running big data applications
- Runs distributed frameworks
  - Hadoop, Spark, Hive, HBase, Presto, Flink
- Operates on top of EC2
  - Provides scalable computing
- Works with AWS data stores
  - S3, DynamoDB, etc.

EMR ARCHITECTURE



SPARK ON AMAZON EMR



https://aws.amazon.com/articles/run-spark-and-spark-sql-on-amazon-elastic-mapreduce/

AWS RESOURCES

- AWS Overview
  - https://aws.amazon.com/getting-started/fundamentals-overview/?e=gs2020&p=gsrc
- AWS Documentation
  - https://docs.aws.amazon.com/index.html
- EC2
  - https://docs.aws.amazon.com/ec2/?id=docs_gateway
- S3
  - https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html
- EMR
  - https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-overview.html
  - https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-spark.html
- Tutorials
  - https://aws.amazon.com/getting-started/hands-on/

CLOUD ANALYTICS

- Introduction to Cloud Computing
- AWS Services
- Amazon EMR
- Amazon EMR Exercise
- Assignments

CLOUD ANALYTICS

- Introduction to Cloud Computing
- AWS Services
- Amazon EMR
- Amazon EMR Exercise
- Assignments

SESSION 4 ASSIGNMENTS

- Overview
  - Dask PA: Cluster analysis on weather dataset
  - AWS PA: Wordcount on AWS EMR
  - Due Friday 2021-05-28 at 11:59pm Pacific Time

- Project: Continue to work on project tasks

DASK PA: CLUSTER ANALYSIS

- Dask
  - Data: Weather Dataset “minute_weather.csv”
  - Task: Perform k-means clustering using Dask
- Submit
  - Complete Jupyter notebook (.ipynb and .pdf)

AWS PA: WORDCOUNT ON EMR

- Word Count on Amazon EMR
  - Use PySpark DataFrame
  - Data files
    - BookReviews_1M.txt (1M reviews, 200 MB)
    - BookReviews_5M.txt (5M reviews, 1.2 GB)
  - Find average & standard deviation of execution times over 3 runs
    - BookReviews_1M.txt: 1 master + 1 worker
    - BookReviews_5M.txt: 1 master + 1 worker
    - BookReviews_5M.txt: 1 master + 3 workers
- Submit
    - Jupyter notebook printout (.pdf)
    - Jupyter notebook (.ipynb)
    - First 25 results for BookReviews_5M.txt (.csv)
    - Screenshot showing that all EMR clusters have been terminated

BOOK REVIEWS DATASETS

- BookReviews_1M.txt
  - 1M reviews
  - 209 MB
  - Available on Canvas
  - Upload to your S3 bucket
  - Read from your S3 bucket
- BookReviews_5M.txt
  - 5M reviews
  - 1.2 GB
  - Read from class S3 bucket (dse230-emr)

FINAL PROJECT

- Final Project Presentation
  - Presented on Saturday 2021-06-05
  - 15 minutes presentation + 5 minutes Q&A
  - To include in presentation:
    - Brief overview of problem
    - Brief description of data
    - Description of modeling approaches
    - Description of challenges with modeling and/or data and approaches to address challenges
    - Discussion of analysis results
    - Discussion of insights gained
    - Discussing of future work

PROJECT COMPONENTS

- Presentation (15 points)
  - 15 minutes presentation + 5 minutes Q&A
  - Presentation order will be provided
  - Submit by Friday 2021-06-04
- Code (5 points)
  - Dask or PySpark code
  - Using Binder is optional but highly recommended
  - Submit by Friday 2021-06-04
- Peer review of two other teams (4 points)
  - Each team will evaluate two other teams' presentations
  - Will also ask questions of presenting team during Q&A
  - Peer review template will be provided
  - Submit by Sunday 2021-06-06
- Team member evaluation (1 point)
  - Evaluate your team member
  - Evaluation template will be provided
  - Individual submission
  - Submit by Sunday 2021-06-06