[544] The Cloud

Tyler Caraza-Harter

Outline

Background

Resources

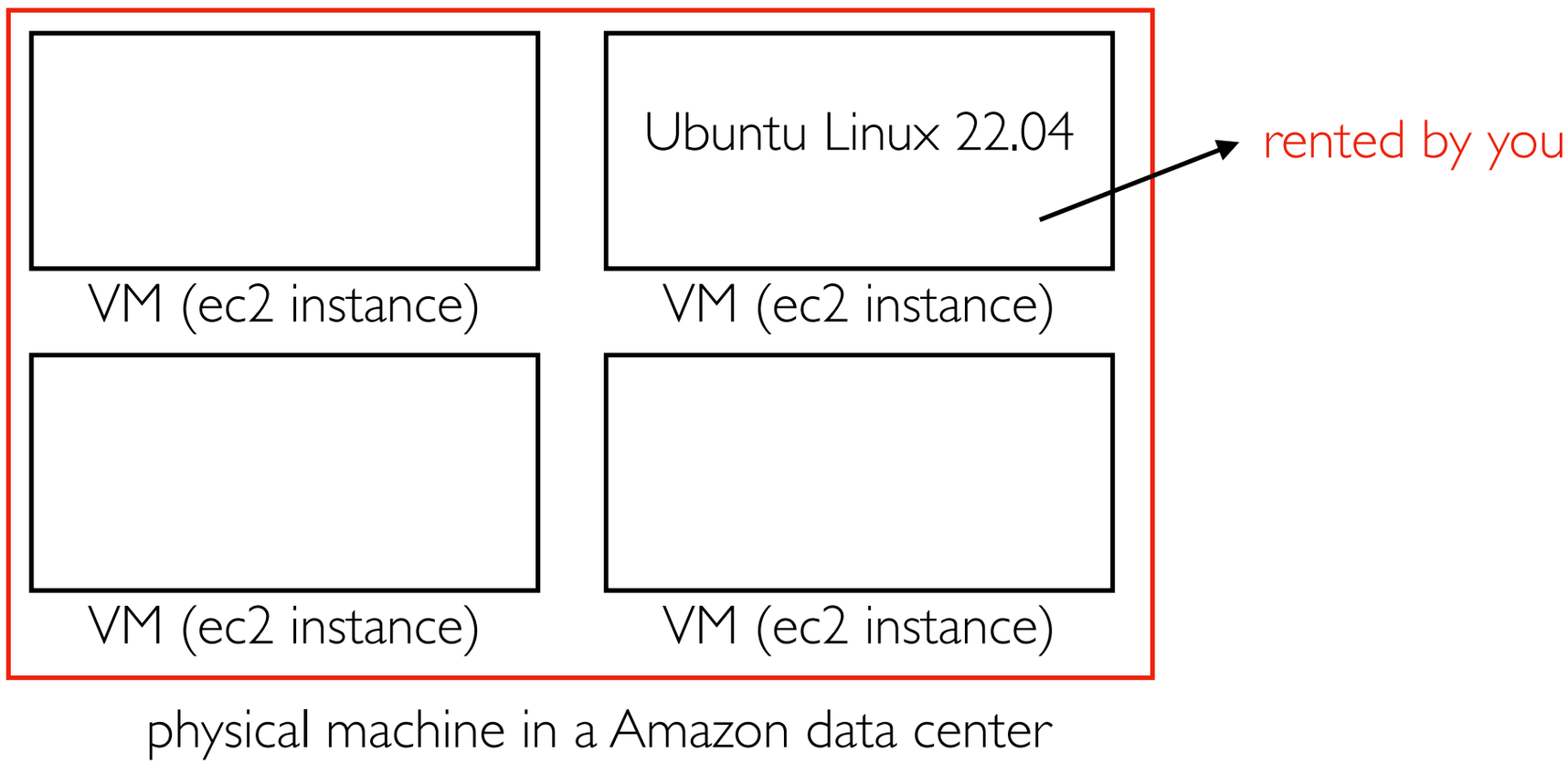
Billing Models

Platforms

The Beginning

Amazon Web Services (AWS)

- Elastic Compute Cloud (EC2), rented VMs, launched in 2006
- "Infrastructure as as Service" (IaaS) -- rent infrastructure (network, storage, compute) instead of owning the hardware yourself.



"Sometimes you need a lot of processing power, and sometimes you need just a little. Sometimes you need a lot, but you only need it for a limited amount of time."
~ Jeff Barr (https://aws.amazon.com/blogs/aws/amazon_ec2_beta/)

VM Hours

Pricing summary

Monthly estimate

$25.46

That's about $0.03 hourly

Pay for what you use: no upfront costs and per second billing

Item	Monthly estimate
2 vCPU + 4 GB memory	$24.46
10 GB balanced persistent disk	$1.00
Total	**$25.46**

Pricing comparison

- one VM for a month: about $25
- about 744 hours/month (31*24)
- 744 VMs for an hour: about $25
- same computation resources
- very different wait time

Other Cloud Services

AWS now has >200 services beyond EC2 (and growing).

IaaS (Infrastructure as a Service)

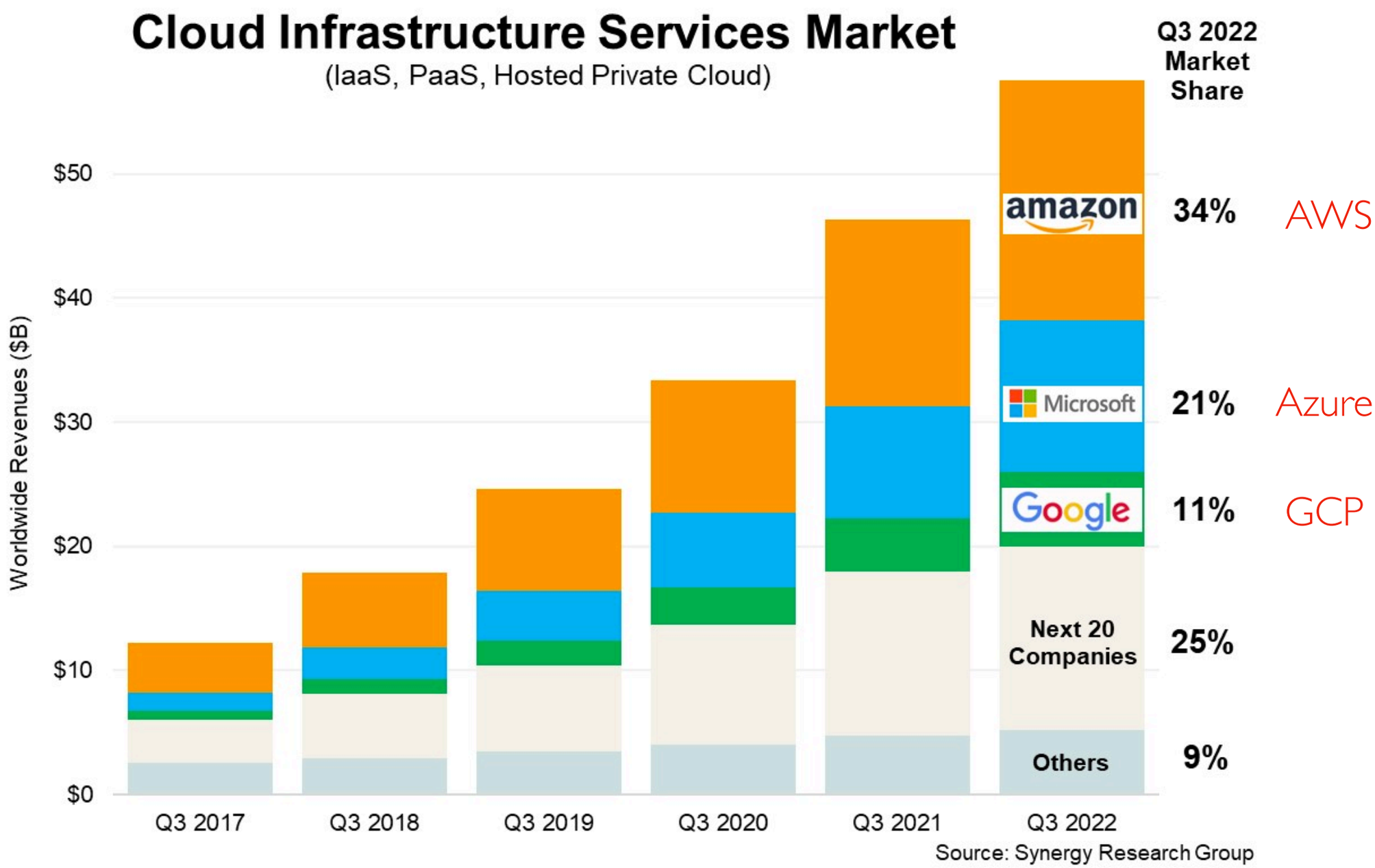
- EC2, other services that feel closer to raw hardware
- virtual disks, virtual network, some storage systems, etc.
- cheap+flexible -- you can deploy anything on it (Cassandra, Kafka, etc).

PaaS (Platform as as Service)

- Cloud provider has deployed systems on the infracture; you pay to use the deployed system
- databases, application framework/platforms, ML training/deployment systems
- less flexible, easier to use
- often more expensive (though not necessarily more than doing it yourself due to effiencies available to cloud provider but not you)

Line between IaaS vs. PaaS distinction is a bit subjective.

Major Cloud Providers Today



https://www.srgresearch.com/articles/q3-cloud-spending-up-over-11-billion-from-2021-despite-major-headwinds-google-increases-its-market-share

Numerous Regions Globally

https://cloud.google.com/about/locations#regions

Outline

Background

Resources

Billing Models

Platforms

Compute - Memory - Storage - Network

Machine configuration

General purpose | Compute optimized | Memory optimized | ✓ GPUs

Graphics processing units (GPUs) accelerate specific workloads on your instances such as machine learning and data processing. Learn More

GPU type
NVIDIA T4

Number of GPUs
2

can choose number and type of GPUs

☐ Enable Virtual Workstation (NVIDIA GRID)

Series

N1

Machine type
n1-standard-1 (1 vCPU, 3.75 GB memory)

vCPU	Memory
1	3.75 GB

can choose number of vCPUs

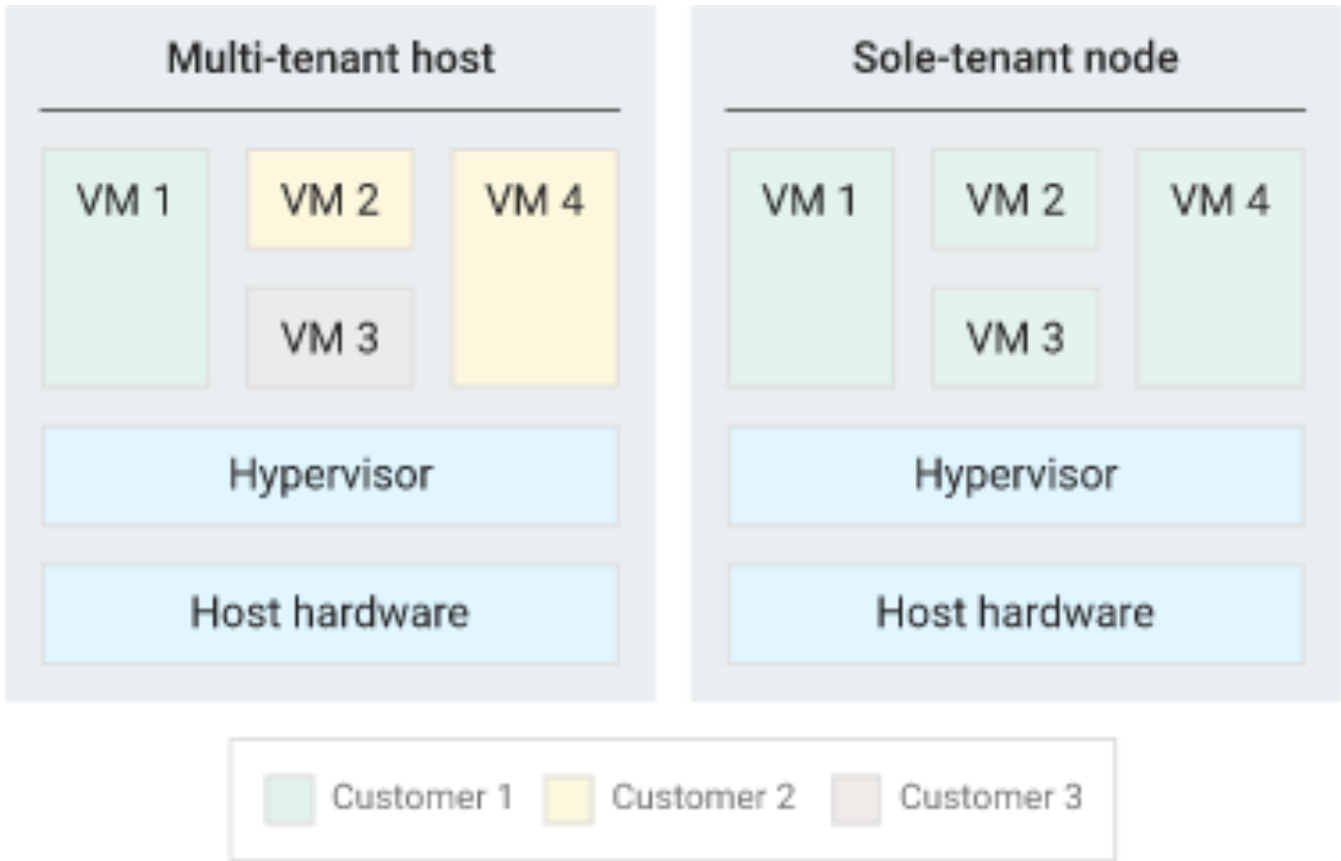
Google offers TPUs (tensor processing units) -- custom hardware for ML. Works with PyTorch and TensorFlow

this VM is ~$400/month (or $0.50/hour)

Compute - Memory - Storage - Network

Forms in which to buy compute

- VMs on multi-tenant hosts (typical case)
- VMs on sole-tenant hosts (better isolation/security, $1000s/month)
- Containers (Kubernetes Engine)
- Serverless Functions (functions run when events happen; pay by 1/10th of a second)



https://cloud.google.com/compute/docs/nodes/sole-tenant-nodes

Compute - **Memory** - Storage - Network

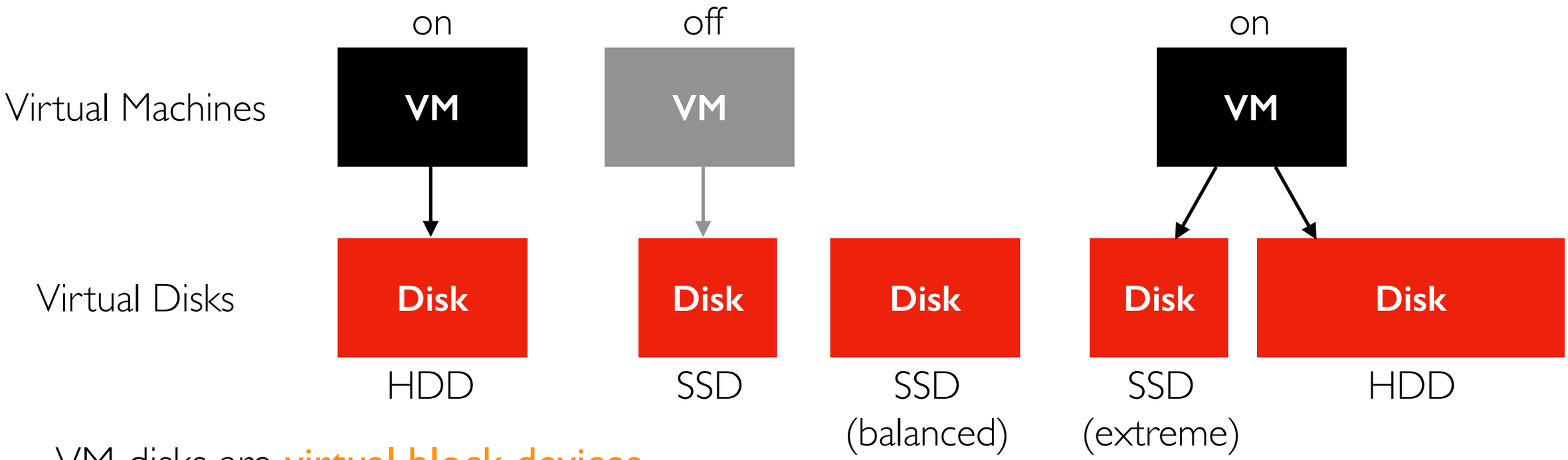
IaaS

- memory is often roughly proportional to CPU resources
- "memory optimized" VMs skew heavy on RAM
  (very expensive! at high end >10 TB)

PaaS: often open-sources platforms provided as a service. Examples:

- memcached (cache)
- redis (in-memory DB)

Compute - Memory - **Storage** - Network



VM disks are **virtual block devices**

- can be attached, detached, re-attached to VMs
- different disk types offer different performance/price tradeoffs
- HDD (standard); SSD (balanced, SSD, extreme)
- price depends on size and type

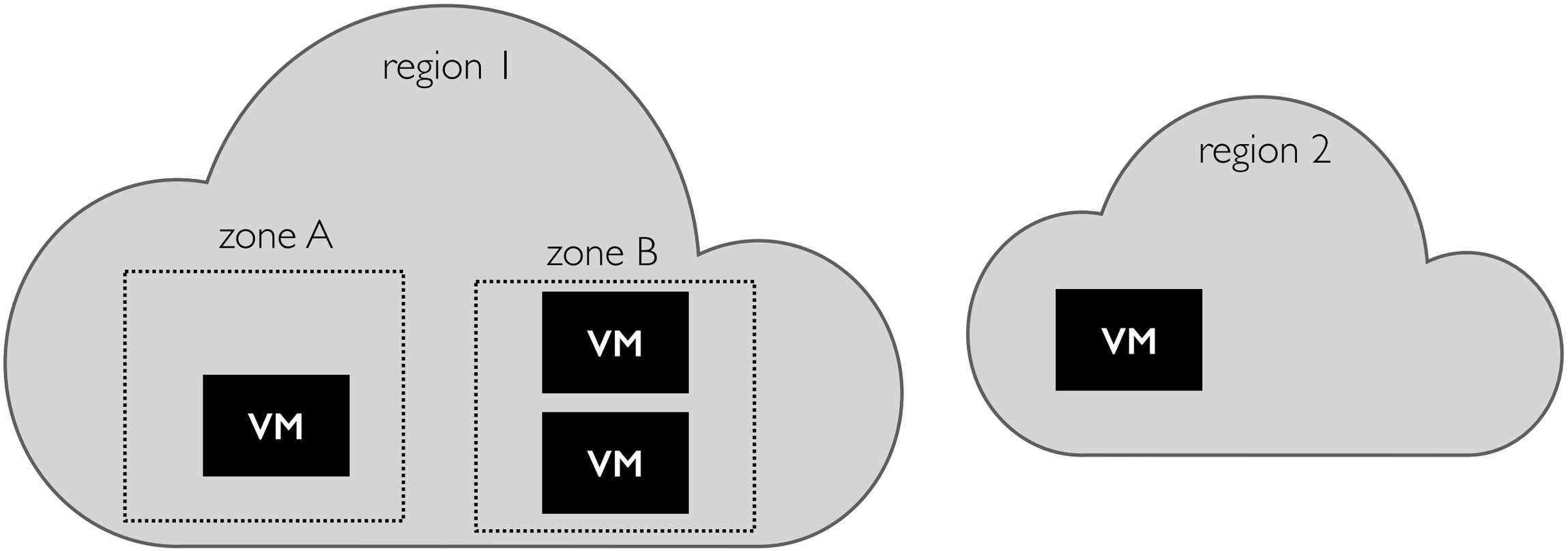
VM creation:

Item	Monthly estimate
2 vCPU + 4 GB memory	$24.46
10 GB balanced persistent disk	$1.00
Total	**$25.46**

cost when running

cost when off (or deleted)

Compute - Memory - Storage - Network



Cloud hierarchy

- continents (approximate)
- regions (data center consisting of 1 or more nearby buildings)
- zone (area of region with fast interconnect but usually common points of failure, like power, routers, etc)

Compute - Memory - Storage - **Network**

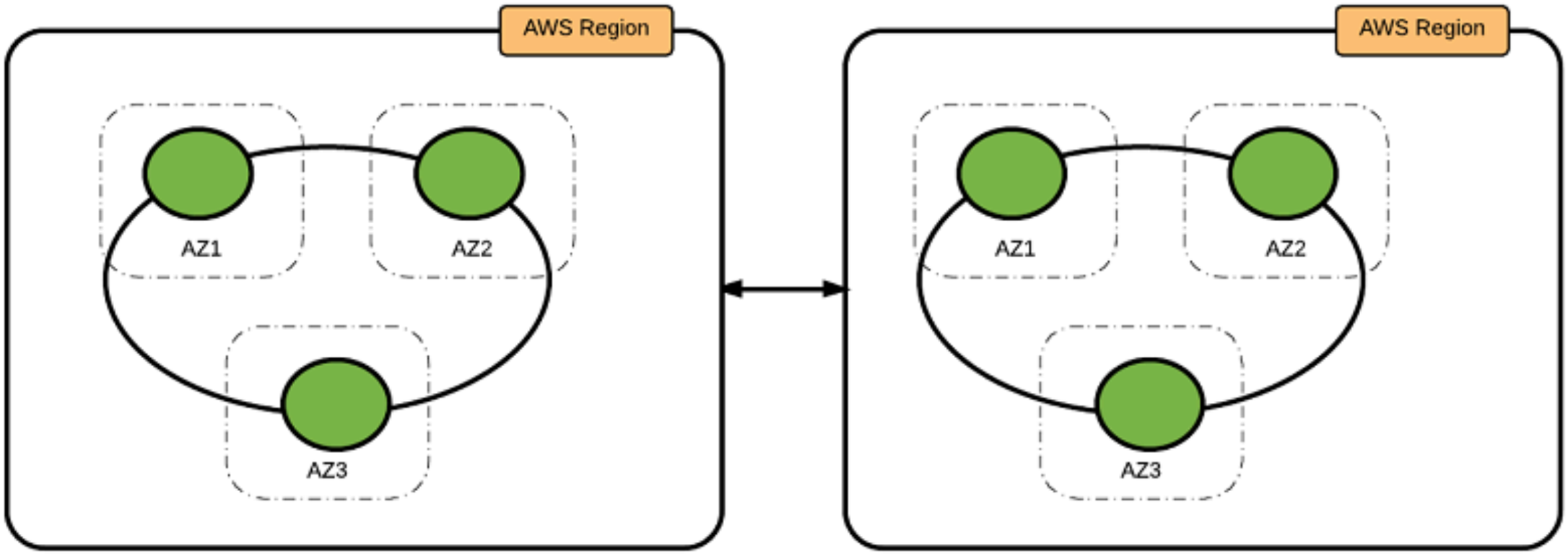
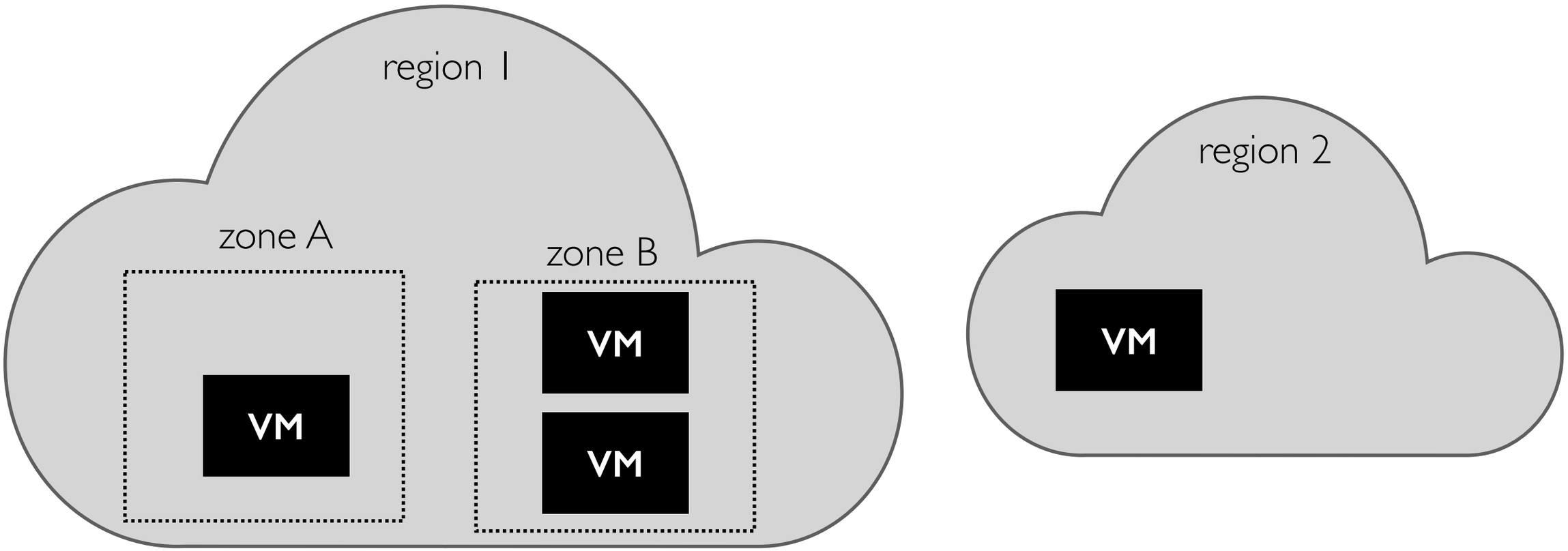


Image from *Best Practices for Running Apache Cassandra on Amazon EC2*
(https://aws.amazon.com/blogs/big-data/best-practices-for-running-apache-cassandra-on-amazon-ec2/)

Fault tolerance

- deploy under the assumption that nodes in the same zone may reasonably all go down together (e.g., due to power loss)
- being extra careful: assume a region can go down (e.g., tornado destroys couple buildings)

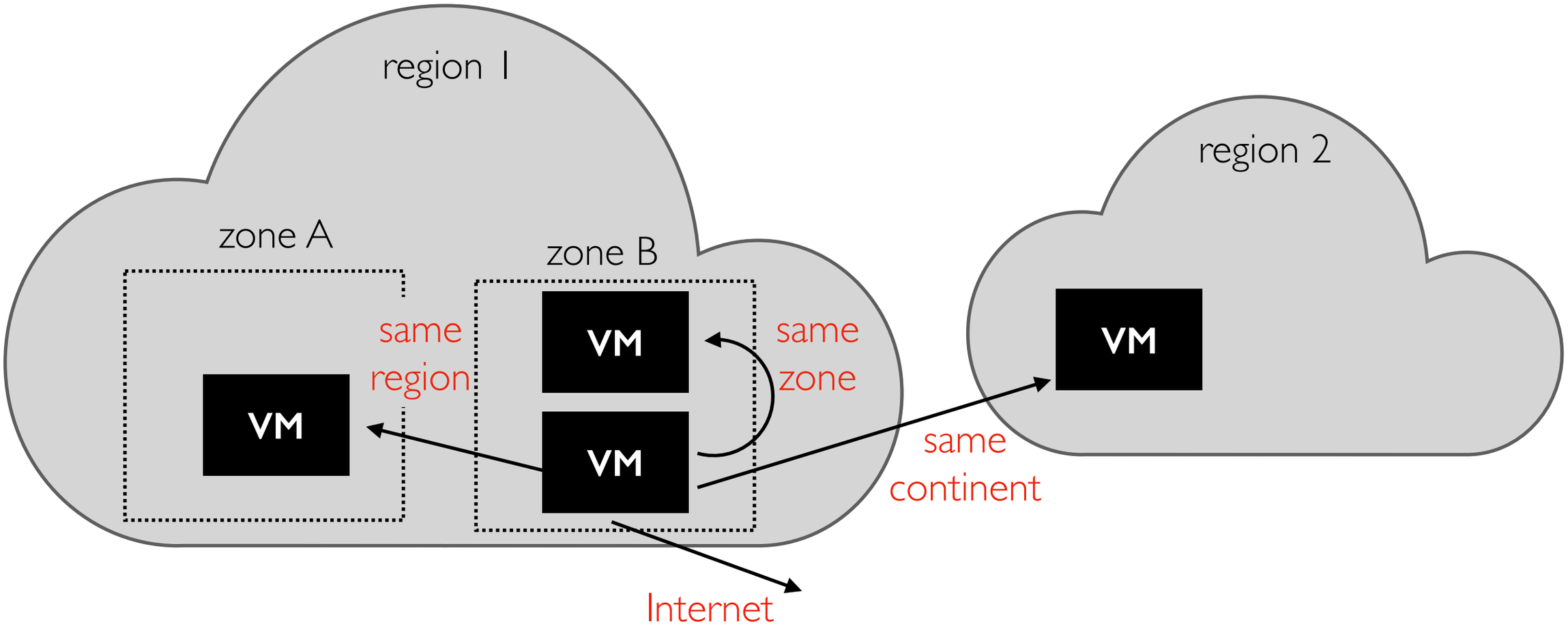
Compute - Memory - Storage - Network



Clouds generally bill per GB of network I/O

- ingress is usually free (incentivize you to start using the service, charge to move your data elsewhere
- egress rate is complicated (depends on many factors)

Compute - Memory - Storage - Network



Egress examples (ballpark for GCP, but very simplified):

- Internet: $0.085/GB
- Same continent: $0.05/GB (Asia)
- Same region: $0.01/GB
- Same zone: free

Outline

Background

Resources

Billing Models

Platforms

Free Tier, Economies of Scale (AWS Lambda Example)

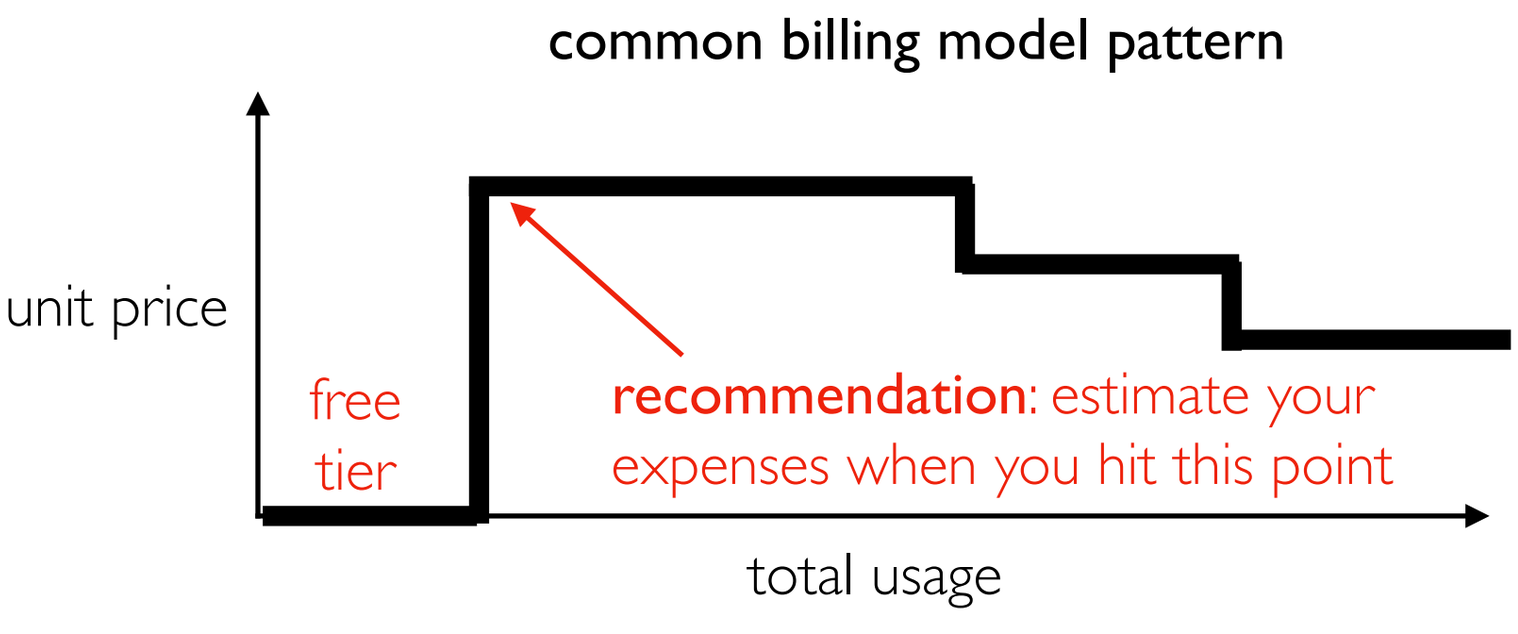
AWS Lambda Pricing

Region: US East (Ohio)

Architecture	Duration
x86 Price	
First 6 Billion GB-seconds / month	$0.0000166667 for every GB-second
Next 9 Billion GB-seconds / month	$0.000015 for every GB-second
Over 15 Billion GB-seconds / month	$0.0000133334 for every GB-second

"The AWS Lambda free tier includes one million free requests per month and 400,000 GB-seconds of compute time per month"

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



"Duration is calculated from the time your code begins executing until it returns or otherwise terminates, rounded up to the nearest 1 ms*"

recommendation: check if you have a large number of small ops getting rounded up

TODO

autoscaling fixed billing vs. pay as you go

spot instances

Outline

Background

Resources

Billing Models

Platforms

TODO

GFS, MapReduce, BigTable

HDFS, Spark, HBase+Cassandra, Kafka

Colossus, GCS, BigQuery, BigTable, Kafka (confluent)

Hadoop Ecosystem

Yahoo, Facebook, Cloudera, and others developed open-source Hadoop ecosystem, mirroring Google's systems

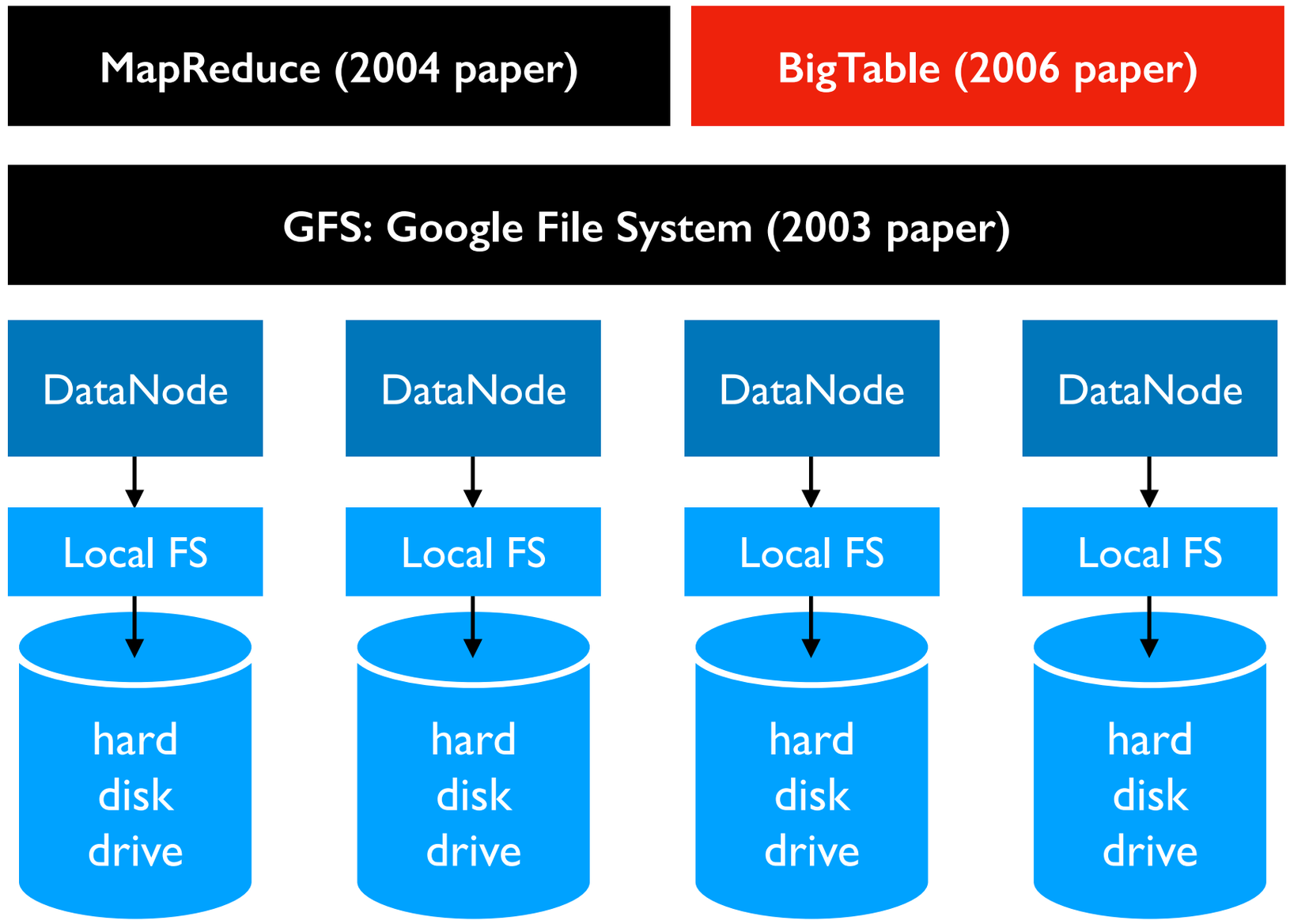
	Google (paper only)	Hadoop, 1st gen (open source)	Modern Hadoop
Distributed File System	GFS	HDFS	
Distributed Analytics	MapReduce	Hadoop MapReduce	Spark
Distributed Database	BigTable	HBase	Cassandra

Dynamo (Amazon) → Cassandra

Ecosystem: Ambari, Avro, Cassandra, Chukwa, HBase, Hive, Mahout, Ozone, Pig, Spark, Submarine, Tez, ZooKeeper

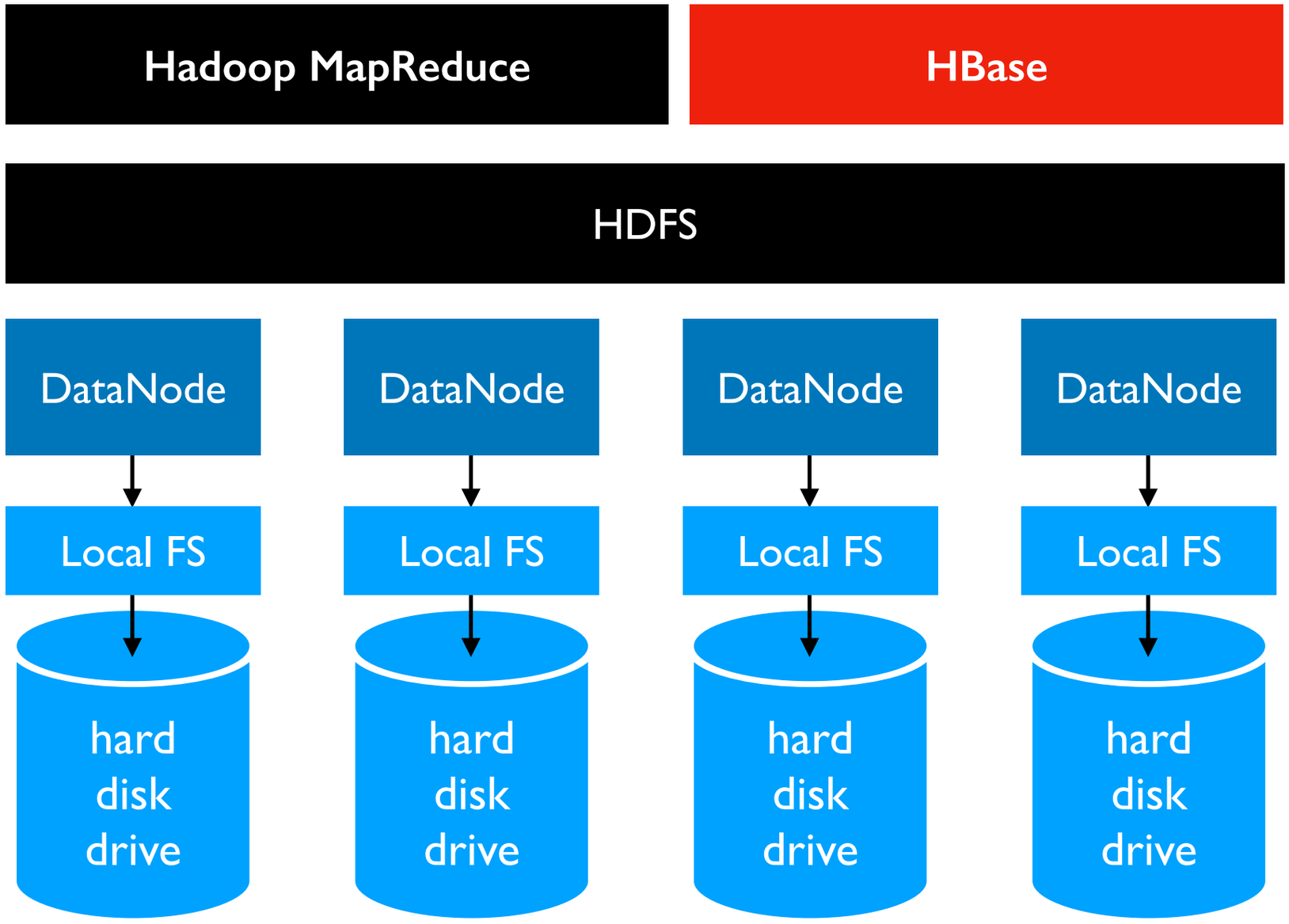
https://hadoop.apache.org/

Google Architecture



radical idea: base everything on lots of cheap, commodity hardware

Hadoop Ecosystem



Hadoop Ecosystem