Cloud Basics

28 May 2021 12:56

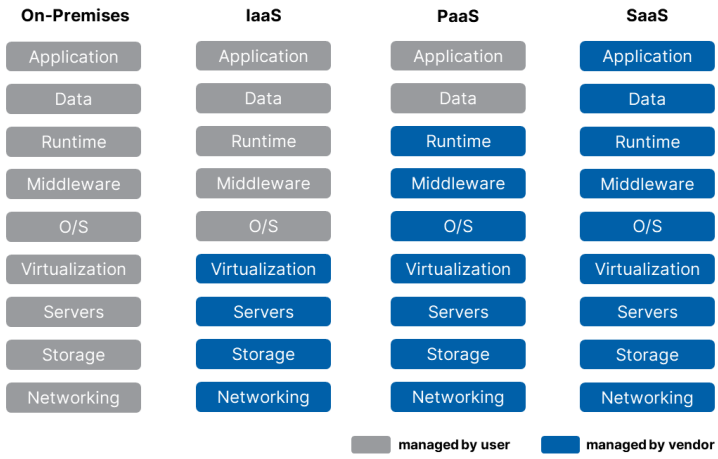
1. Cloud Benefits

- Cost-effective - PAYG
- Global
- Secure
- Reliable - data backup, disaster recovery
- Scalable - increase / decrease resource #
- Elastic - automat scaling during spikes and drop in demand
- Current - underlying hardware and software is patched, upgraded and replaced by cloud provider

2. Types of cloud computing

- SaaS - run & maintained by service provider e.g. Gmail, office365. customer-orientated
- PaaS - for ppl who building app, but dont worry about provisioning, configurating or understanding the hardware or OS e.g. Google app engine
- IaaS -provides networking / computing/ data storage features etc. e.g. Azure, AWS

3. Types of cloud computing responsibilities

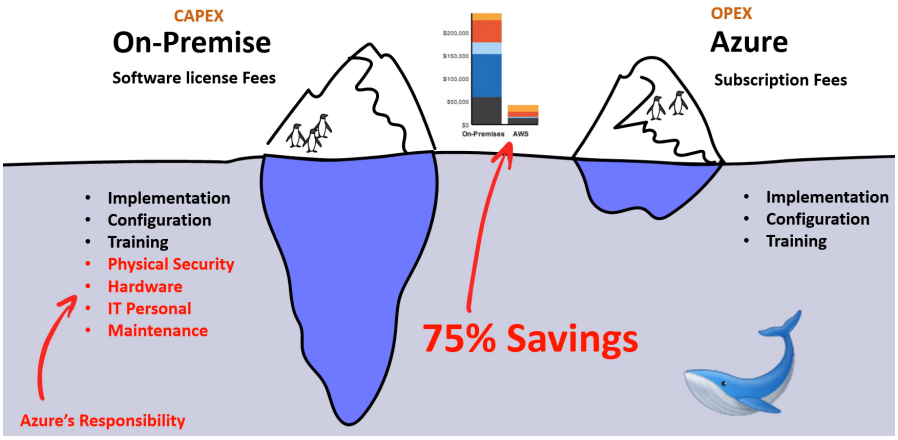


4. Azure Deployment models

- Public cloud
- Private cloud - on premise , everything built on company's datacentres
- Hybird - using both on premise and cloud service provider
- Cross cloud - using multiple cloud provides

	Cost	Security	Level of Configuration	Technical Knowledge
Public Cloud	👍 Most cost-effective	👍 Security Controls by Default 👎 Might not meet security requirements	👎 Limited based on what the Cloud Service Provider exposes to you.	👍 You don't need in-depth knowledge of underlying infrastructure
Private Cloud	👎 Most expensive	👎 no guarantee its secure 👍 can meet any security compliance requirement if you put in the work.	👍 You can configure the infrastructure however you like.	👎 You need to know in-depth how to configure all levels of your infrastructure
Hybrid	👍👎 Could be more cost-effective based on what you offload to the cloud.	👎 you now have to secure your connection to the cloud 👍 can meet all security requirements	👍 You get the best of both worlds.	👎 You need to know in-depth how to configure all levels of your infrastructure and know the CSPs services.

5. Total cost of ownership (TCO)



Capital Expenditure (CAPEX)

Spending money upfront on **physical infrastructure**
Deducting that expense from your tax bill over time.

- Server Costs (computers)
- Storage Costs (hard drives)
- Network Costs (Routers, Cables, Switches)
- Backup and Archive Costs
- Disaster Recovery Costs
- Datacenter Costs (Rent, Cooling, Physical Security)
- Technical Personal

With Capital Expenses **you have to guess upfront** what you plan to spend

Operational Expenditure (OPEX)

The costs associated with an on-premises datacenter that has shifted the cost to the service provider. The customer only has to be concerned with non-physical costs.

- Leasing Software and Customizing features
- Training Employees in Cloud Services
- Paying for Cloud Support
- Billing based on cloud metrics eg.
  - compute usage
  - storage usage

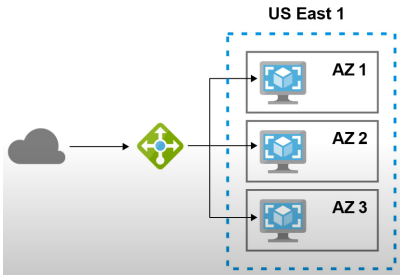
With Operation Expenses you can try a product or service **without investing in equipment**

6. Cloud architecture terminologies

- Availability - **High Availability (HA)**
- Scalability -ability to grow rapidly or unimpeded
- Elasticity - ability to shrink and grow to meet demand
- Fault tolerance - ability to prevent a failure
- Disaster recovery - ability to recover from failure - **Highly Durable (DR)**

High Availability (HA)

- Ability for your service to remain available by ensuring there's no single point of failure and /or ensure a certain level of performance
- Run your workload across multiple availability zones ensures if 1/2 datacentres failures your service still available



Azure Load Balancer

A load balancer allows you to evenly distribute traffic to multiple servers in one or datacenter. If a datacenter or server becomes unavailable (unhealthy) the load balancer will route the traffic to only available datacenters with servers.

High scalability

- Ability to increase your capacity based on the increase traffic, memory, and computing power

Vertical Scaling
Scaling **Up**

Upgrade to a bigger server

Horizonal Scaling
Scaling **Out**

Add more servers of the same size

High elasticity

- Ability to **automatically increase / decrease** capacity based on current traffic, memory, and computing power
- Horizonal sailing

Horizonal Scaling
Scaling **Out** — Add more servers of the same size
Scaling **In** — Removing more servers of the same size

Azure VM scale sets
- Automatically increase or decrease in response to demand or schedule

SQL Server Stretch Database
- Dynamically stretch warm and cold transactional data from Microsoft SQL server 2016 to Azure

Vertical Scaling is generally hard for traditional architecture so you'll usually only see horizontal scaling described with Elasticity.

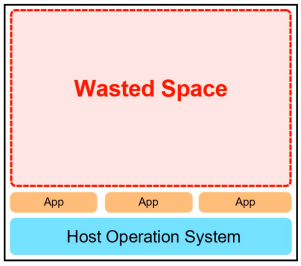
High durability

- Ability to recover from a disaster and to prevent the loss of data
- Solution that recover from a disaster is known as disaster recovery (DR)
  - Do you have a backup?
  - How fast can you restore that backup?
  - Does your backup still work?
  - How do you ensure current live data is not corrupt?

7. The evolution of computing

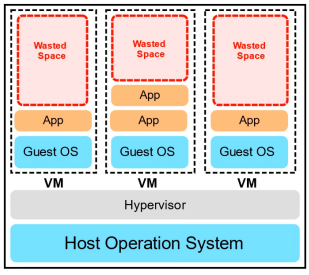
Dedicated server

*Dedicated 👉 VMs 👉 Containers 👉 Functions



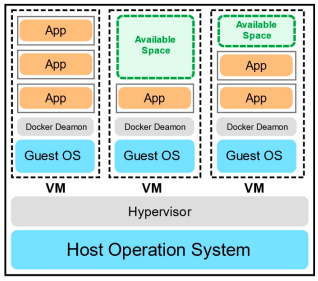
- A physical server **wholly utilized by a single customer**.
- You have to guess your capacity, you'll overpay for an underutilized server
- Upgrading beyond your capacity will be slow and expensive
- You are limited by your Operating System
- Multiple apps can result in conflicts in resource sharing
- You have a ***guarantee of security, privacy and full utility of underlying resources**

VMs



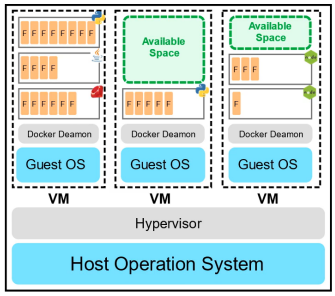
- You can run **multiple Virtual Machines on one machine**.
- **Hypervisor** is the software layer that lets you the VMs
- A physical server shared by multiple customers
- You are pay for a fraction of the server
- You'll overpay for an underutilized Virtual Machine
- You are limited by your Guest Operating System
- Multiple apps on a single Virtual Machine can result in conflicts in resource sharing

Containers



- Virtual Machine running multiple containers
- **Docker Deamon** is the name of the software layer that lets you run multiple containers.
- You can maximum the utilize the available capacity which is more cost-effective
- Your containers share the same underlying OS so containers are more efficient than multiple VMs
- Multiple apps can run side by side without being limited to the same OS requirements and will not cause conflicts during resource sharing

Functions

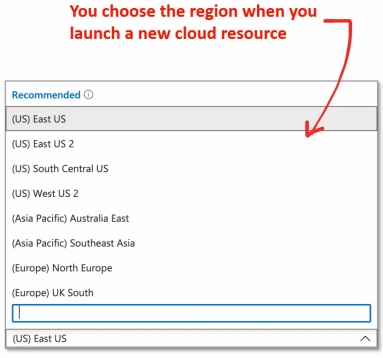


- A managed VMs running managed containers.
- Known as **Serverless Compute**
- You upload a piece of code choose the amount of memory and duration.
- Only responsible for code and data, nothing else
- Very cost-effective, only pay for the time code is running, VMs only run when there is code to be executed
- Cold Starts is a side-effect of this setup

8. Global infrastructure

regions and geographies

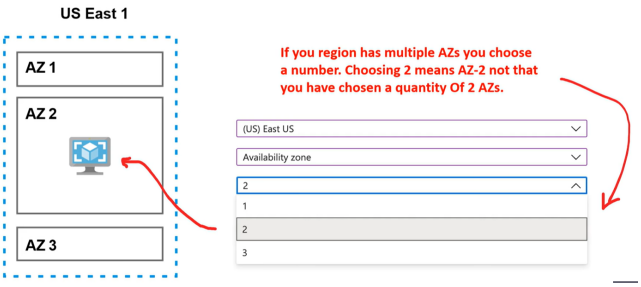
a. A region is a grouping of multiple datacentres (AZ)
   i. Azure has 58 regions available across 140 countries
   ii. E.g. West US, Canada Central, West Europe, Australia East, and Japan West



b. **Availability Zone (AZ)**

i. Physical separate your resrouces within in Azure **region,** made up of 1 or +1 data centers
ii. A region generally has 3 AZs, isolate from each other (diff buildings) but close enough to provide low latency
iii. Common practice to run workloads in 3 AZs, in case +1 datacentres fails (HA: high availability)

可以将**地域**理解为不同城市的机房，将**可用区**理解为同一个城市的不同机房



a. A geography is discreet market of two or more regions that preserves **data residency and compliance boundaries**

Azure Geographies
- United States
- Azure Government (US)
- Canada
- Brazil
- Mexico

Imagine you are in Canada and **you want a guarantee that data will remain within Canada**.
You would want to use Canada Azure Geographies

b. Paired regions
- Each region is paired with another 300 miles always
- Only one region updated at a time to ensure availability
- Some azure service rely on paired regions for disaster recovery
  - e.g. Azure geo-redundant storage (GRS), replicates data to 2nd region automatically, ensuing data durable when 1st region not recoverable

Canada	Canada Central	Canada East
North America	East US	West US
Germany	Germany Central	Germany Northeast

Region types and service availability

a. NOT ALL Azure cloud service available in every single region

Recommended region

A region that provides the broadest range of service capabilities and is **designed to support Availability Zones** now, or in the future.

Alternate (other) region

A region that extends Azure's footprint within a data residency boundary where a recommended region also exists. **Not designed to support AZs.**
These Regions are label as **Other** in the Azure Portal

b. general availability (GA) - when service is ready for everyone publicly, Azure service grouped in 3 categories
i. Foundational - GA, immediately or in 12 months in both RR and AR
ii. Mainstream - GA, immediately or in 12 months in RR, may available in AR based on demand
iii. specialised - based on customers demand

Special regions

a. Azure has SR to meet compliances and legal reasons e.g. US / China

- US DoD Central
- US Gov Virginia
- US Gov Iowa

***Three Azure Government secret locations undisclosed**

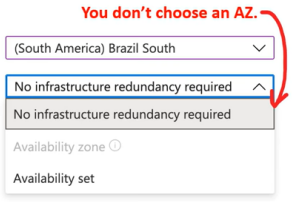
- China East
- China North

***Available through a unique partnership between Microsoft and 21Vianet. Microsoft does not directly maintain the datacenters.**

Availability Zone supported regions

NOT EVERY has supported AZs - Alternate or Other
Recommended Regions = have +3 Azs

The following Regions **have a minimum of 3 AZs**
- Central US
- East US 2
- West US 2
- West Europe
- France Central
- North Europe
- Southeast Asia

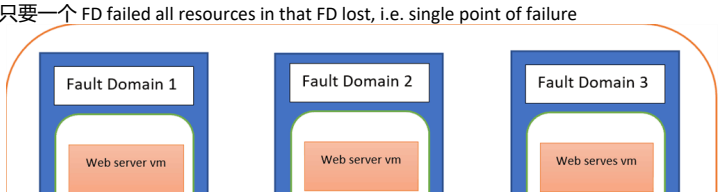


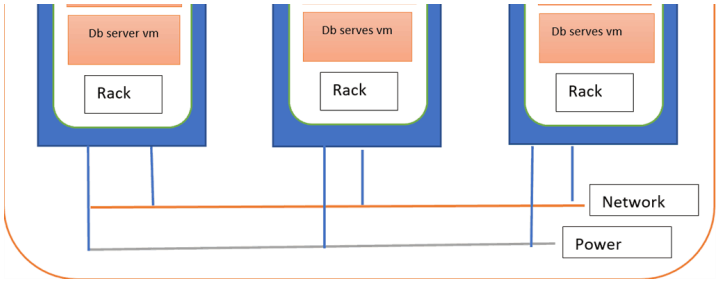
Althernate region

Fault and Update domains

Each VM in an **Availability Set (ASet)** is assigned with **a fault domain (FD)** and **a update domain (UD)**

a. Fault domain - group of VMs that share the same networking ad power supplies.
i. Azure有很多机架，每个机架上都有很多服务器，一个rack of servers 相当于一个 FD
ii. 所以只要一个 FD failed all resources in that FD lost, i.e. single point of failure





b. Update domain - VMs get UDs automatically once put them in the ASet. All VMs within that UD will reboot together
   i. Ensure your resources don't go offline
   ii. Only 1 UD can be updated at the time

a. Availability set - a logical grouping of VMs (FDs+UDs) to avoid downtime
   i. **Automatically distribute your VMs across multiple FDs,** thereby eliminating any single point of failure.
   ii. Recommend +2 VMs in an ASet for highly available application and to meet the 99.95% Azure SLA