Chapter 4

Infrastructure groups

The major "full service" cloud providers separate their infrastructure geographically. This is done for two reasons. Firstly it lets cloud providers separate services in different geographical regions from each other for legal and business reasons. Secondly it facilitates redundancy in case of failures.

4.1 Regions

Regions are **designed to separate data and resources** for business and legal reasons.

- Regions function independently of each other. Essentially they are different instances of the cloud provider's infrastructure.
- Regions usually chosen such that legal environment is similar across them: e.g. varied FOI, GDPR, HIPAA, COPPA, data siphoning regulations.
- Provider may introduce and update services to different regions at different times.
- Pricing may vary across regions for the same services.
- Most cloud providers identify each region with both a human readable name like "EU (Ireland)" and a code name like `eu-west-1`. We will in class use only the code names for the most part.
- **Must consider region when deploying most services!**

Also need to consider:

Non-geographic regions are purpose-specific rather than geographical, like AWS GovCloud.

Geography is a region-grouping that Azure use. AWS doesn't have a multi-region grouping concept.

- Legal environment is normally consistent across all regions in a Geography.

4.1.1 Region listing

A list of regions can be accessed in the CLI using

```
aws ec2 describe-regions
```

4.2 Availability zones

Availability zones are a **technical construct** to tolerate failures in physical infrastructure.

- Each AZ normally consists of ≥ 1 data centres, each internally redundant. The data centres in an AZ are connected to each other by low-latency highly-redundant network links.
- AZs are identified only by a code name derived from their region code name. For the `eu-west-1` region the corresponding AZs are `eu-west-1a`, `eu-west-1b` and `eu-west-1c`.
- Note that the AZ code `a, b, c` mappings to physical AZs *are not* consistent across different AWS accounts. This is to balance usage across AZs. (Many people will just choose the `a` AZ since it appears first in the list.)
- **In general, we need to consider AZs when deploying IaaS services but NOT PaaS.**

4.2.1 AZ listing

A list of availability zones can be accessed in the command-line interface:

```
aws ec2 describe-availability-zones
```

4.3 Edge locations

Edge locations are an entirely parallel structure provididing access points and edge caching for users of customer applications.

- These are normally a separate set of locations to the provider's regions/AZs for customer use.
- Only encountered when using certain services (not this semester).

4.4 Infrastructure maps

The providers all do a better job on constantly updating their global infrastructure maps than any lecture notes will:

- **AWS:** `https://aws.amazon.com/about-aws/global-infrastructure/`
- **Google Cloud:** has a similar breakdown of regions and AZs to AWS. `https://cloud.google.com/about/locations`
- **Azure:** `https://azure.microsoft.com/en-us/global-infrastructure/`