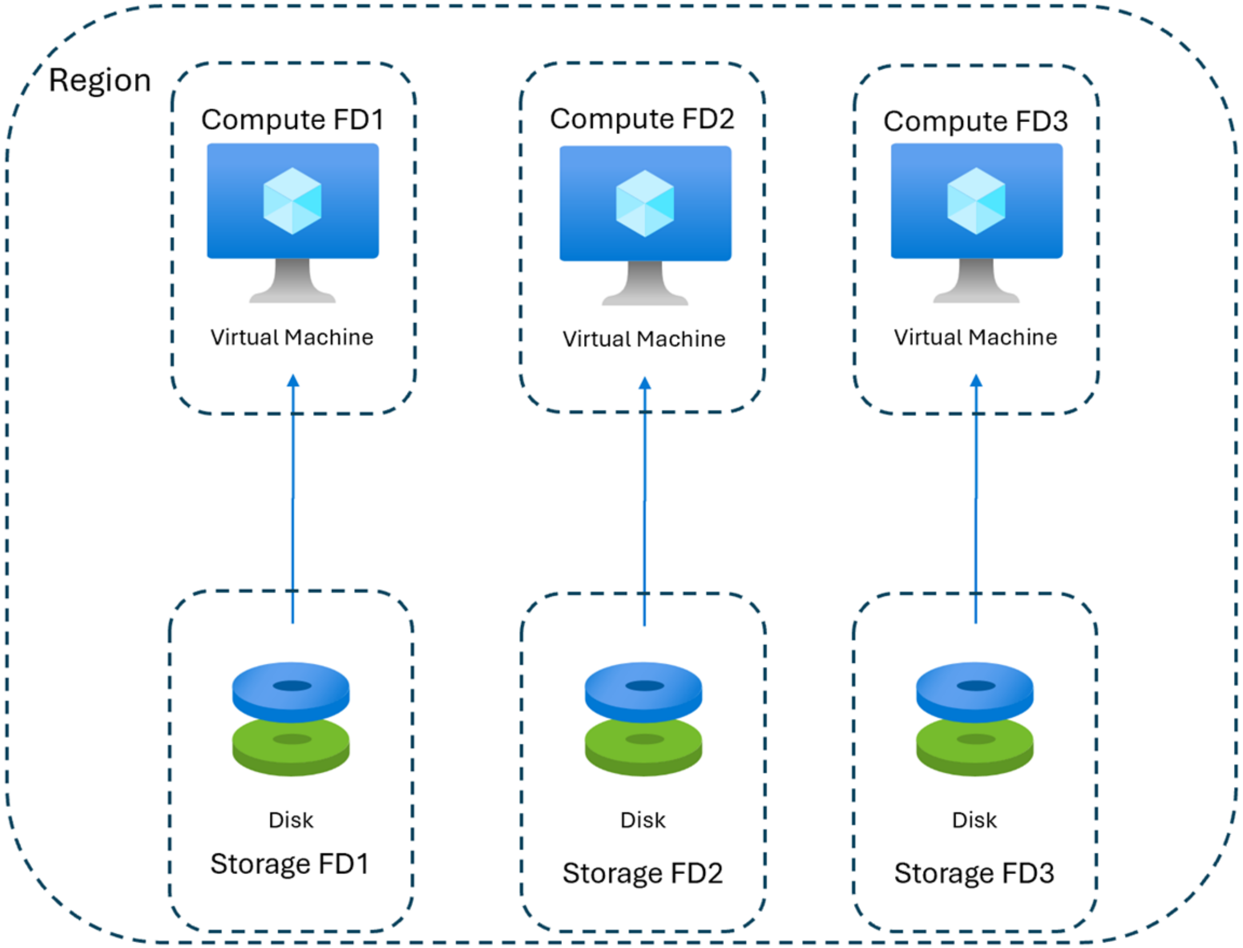
Availability Sets



VMs are placed on nodes in a rack

Failures can occur at node and rack levels

Maintenance is also required on hosts and VMs are **not** live migrated

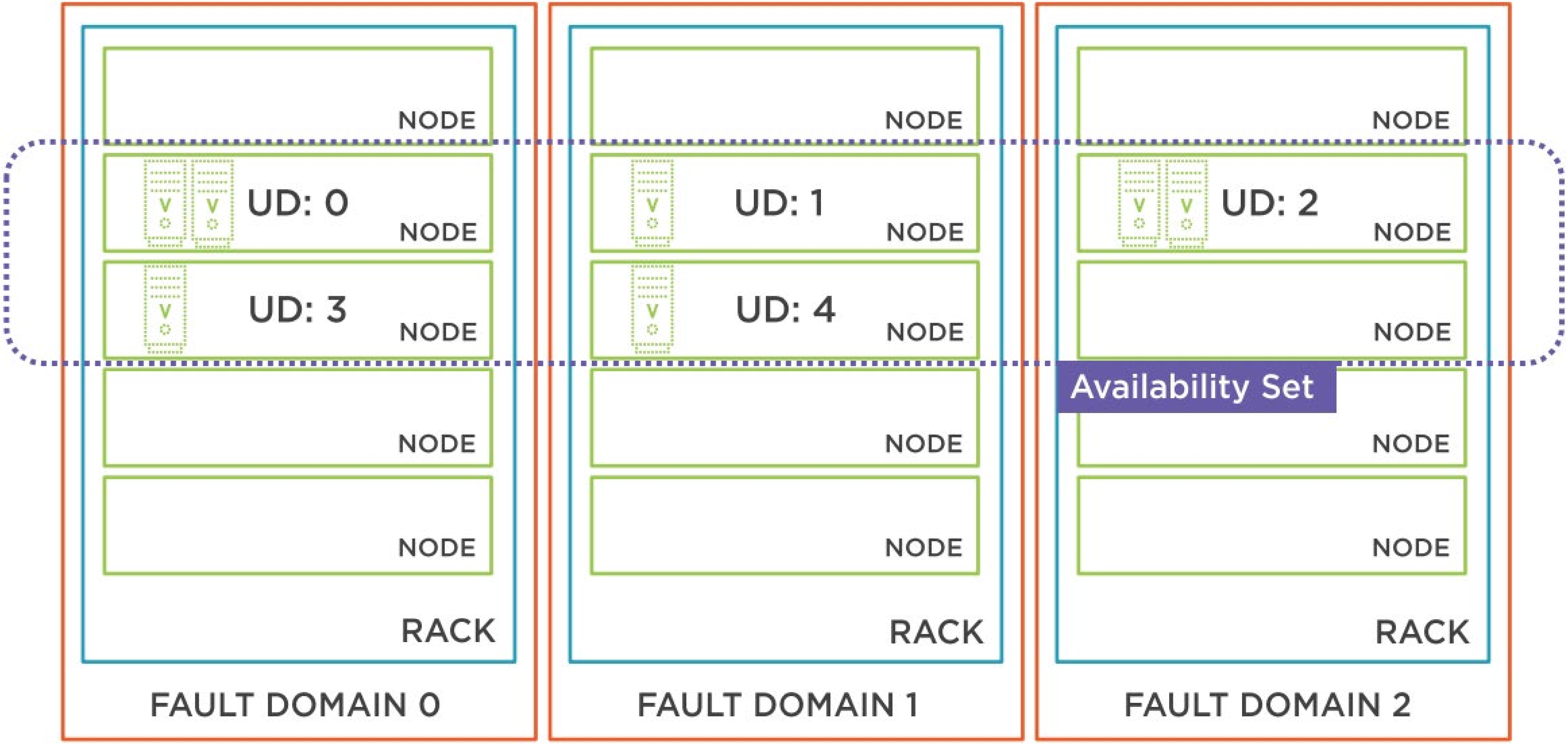
Availability Sets

To ensure availability of services, always deploy minimum 2 instances of any service and place in a unique availability set

This ensures VMs are spread over three fault domains (racks) and five (by default) update domains

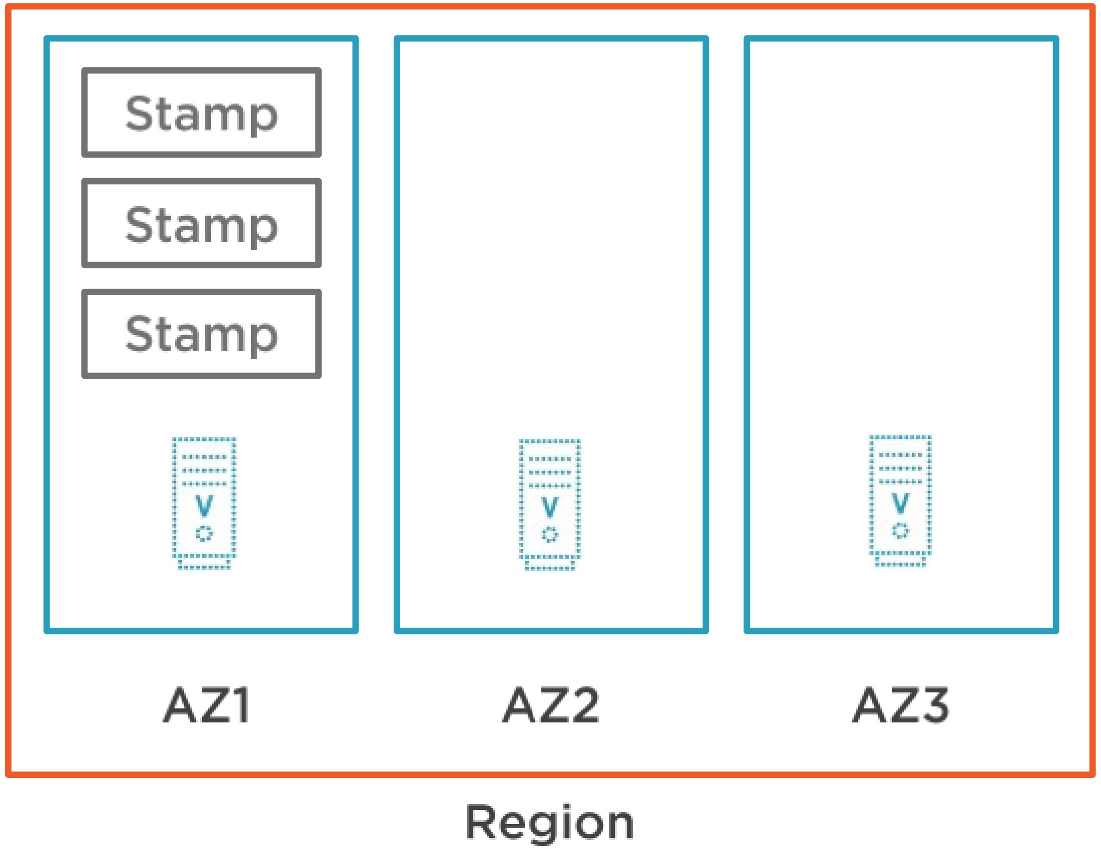
Must have minimum 2 VMs to receive SLA of 99.95%

Availability Sets

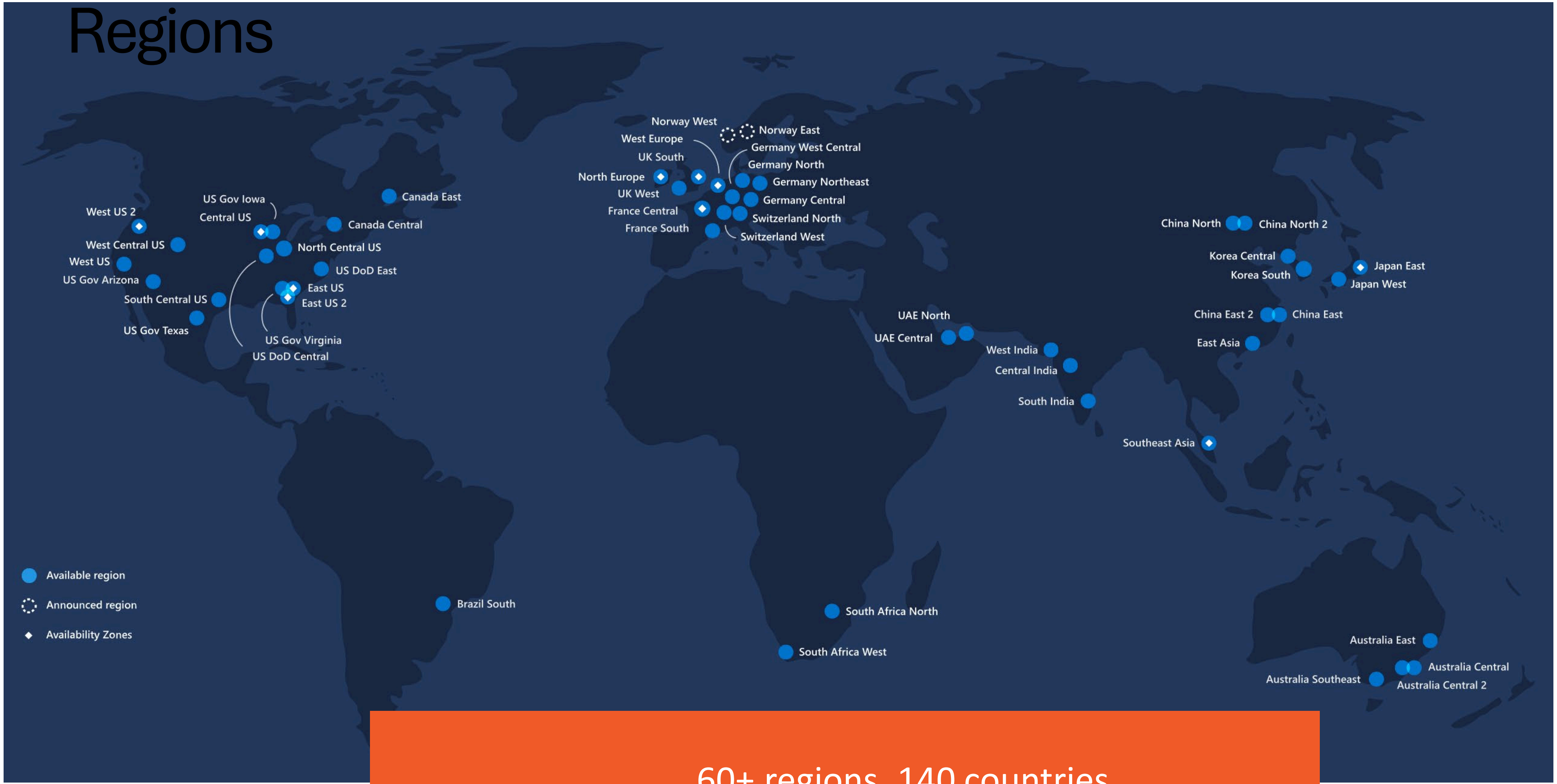


Availability zones

- Regions are broken up into physically separate AZs
- AZs have independent power, cooling and networking
- 3 AZs are exposed per subscription
- VMs spread over AZs receive 99.99% SLA
- Each AZ can be thought of as separate fault domain and update domain
- Virtual networks span AZs



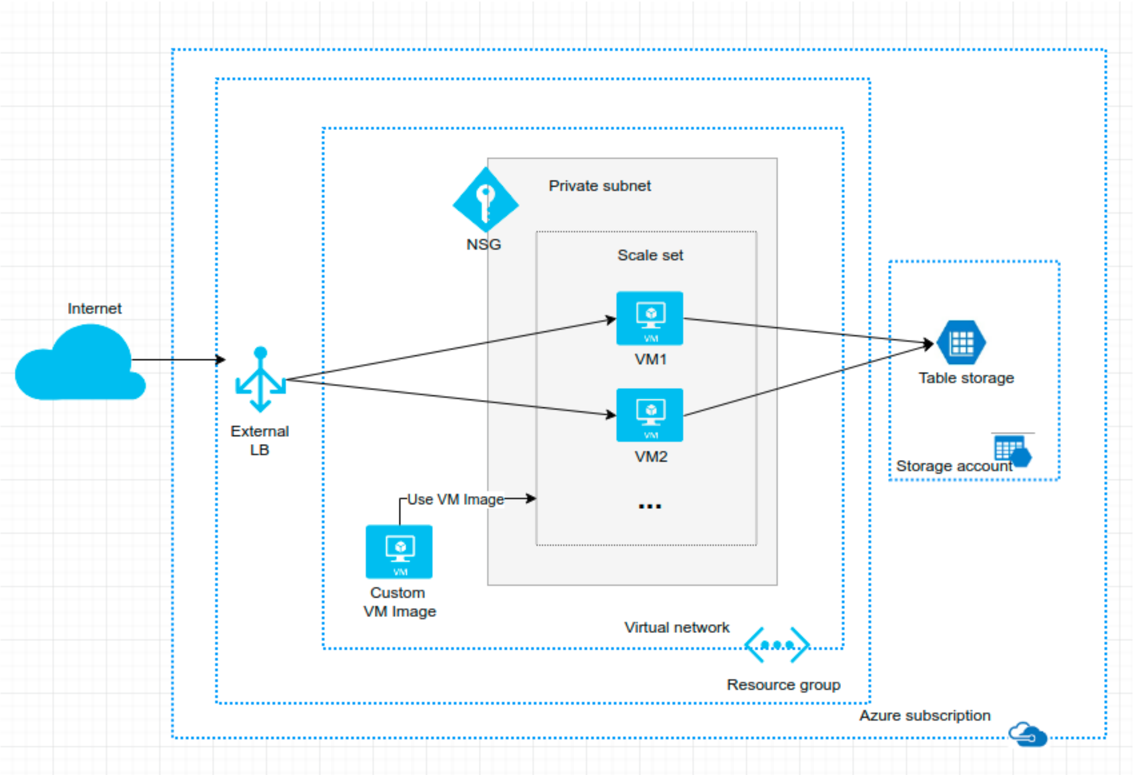
Regions



60+ regions, 140 countries

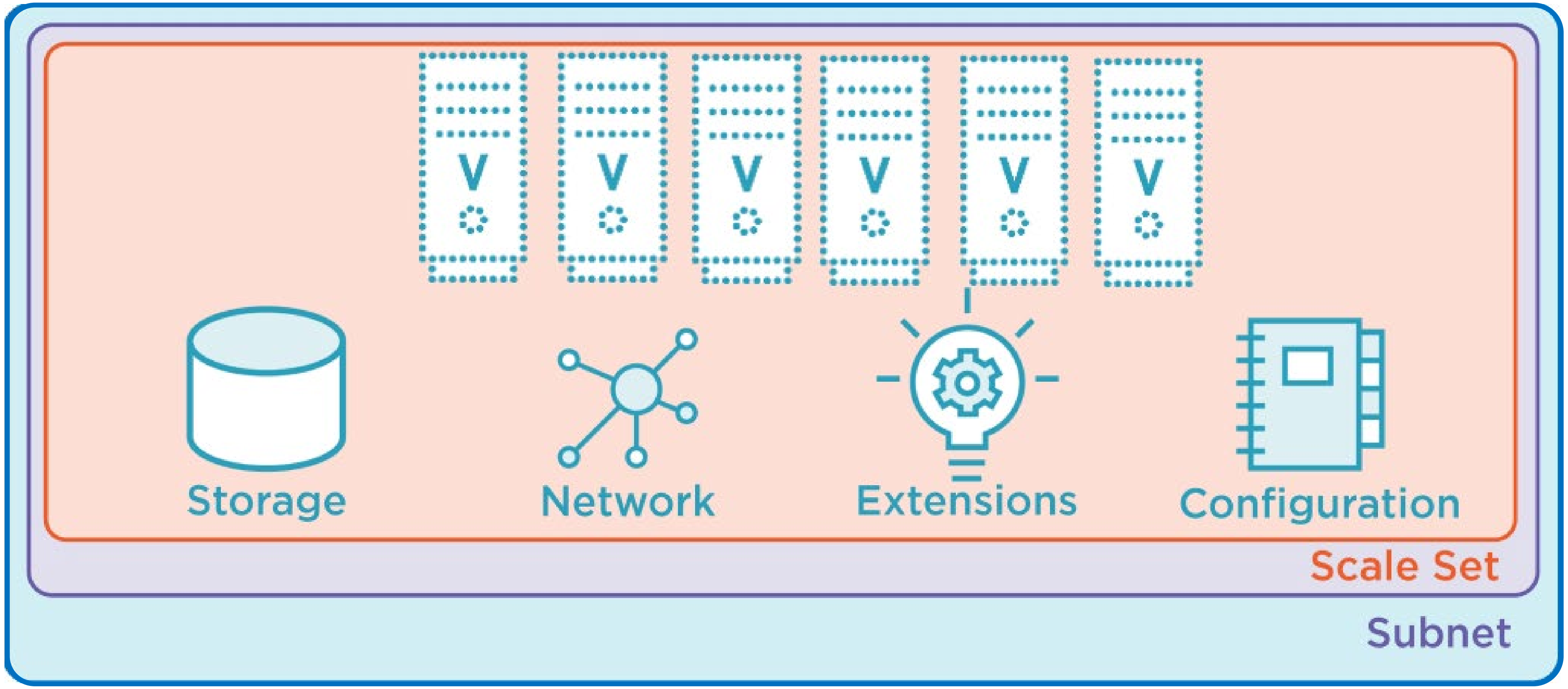
Scale sets

- Enables large scale deployment of VMs from single gold image with automatic configuration
- Supports auto-scale based on metrics and schedules
- Scale sets can be updated without taking down the entire set



Scale sets

- Azure VM Scale Sets enable the entire deployment to be managed in a simple fashion
- Enable rollout of updates without taking down the entire service.



Scale sets Limitations

- 1000 VMs per Scale Set maximum and use with Azure Standard Load Balancer for matching scale

- 2000 Scale Sets per region per subscription

Auto-scaling Scale Sets

Scheduled

Metric-Based

Scale Sets support many types of scale

- On a schedule both specific day and recurrence (enables scaling ahead of the load increase)
- Based on resource metrics
- Can combine (schedule rules take precedence over metric rules)

Pop quiz:

You need high availability in a single region and want to support 200 instances in a scale set. Which option should you choose?

A. Regional scale set (non-zonal) without placement groups
B. Zonal scale set with true Availability Zones
C. Regional scale set with placement groups
D. Single VM in an Availability Set

Pop quiz:

You need high availability in a single region and want to support 200 instances in a scale set. Which option should you choose?
A. Regional scale set (non-zonal) without placement groups
B. Zonal scale set with true Availability Zones
C. Regional scale set with placement groups
D. Single VM in an Availability Set