

Azure Redundancy and Availability

Azure provides many ways to architect for fault tolerance and redundancy.

Physical

Geography

At the highest level is the concept of Geography. Within a specific geography are one or more Regions. The geography boundary provides for known data residency. With our ExpressRoute location and default region of "East US" we will be in the "Americas, United States" geography. This includes the regions : Central US, East US 2, East US, North Central US, South Central US, West US 2, West Central US, and West US.

<https://azure.microsoft.com/en-us/global-infrastructure/geographies/>

Regions

A Region is made up of one or more Availability Zones. The datacenters within a Region are deployed within a latency-defined perimeter and connected through a dedicated regional low-latency network.

<https://azure.microsoft.com/en-us/global-infrastructure/regions/>

Region Pairs

Each Azure region is paired with another region within the same geography, together making a regional pair. Across the region pairs Azure serializes platform updates (planned maintenance), so that only one paired region is updated at a time. In the event of an outage affecting multiple regions, at least one region in each pair will be prioritized for recovery. For the majority of our work we will deploy in East US which is paired with West US.

<https://docs.microsoft.com/en-us/azure/best-practices-availability-paired-regions>

Availability Zones

Availability Zones are physically separate locations within an Azure region. Each availability zone is made up of one or more datacenters equipped with independent power, cooling, and networking.

<https://azure.microsoft.com/en-us/global-infrastructure/availability-zones/>

Compute

Availability Set

An Availability Set is a logical grouping capability for isolating VM resources from each other when they're deployed. Azure makes sure that the VMs you place within an Availability Set run across multiple physical servers, compute racks, storage units, and network switches. Each VM in your availability set is assigned an update domain and a fault domain.

<https://docs.microsoft.com/en-us/azure/virtual-machines/windows/manage-availability>

Update Domain

For a given availability set, five non-user-configurable update domains are assigned by default (this can be increased to 20). When more than five virtual machines are configured within a single availability set, the sixth virtual machine is placed into the same update domain as the first virtual machine, the seventh in the same update domain as the second virtual machine, and so on. The order of update domains being rebooted may not proceed sequentially during planned maintenance, but only one update domain is rebooted at a time. A rebooted update domain is given 30 minutes to recover before maintenance is initiated on a different update domain.

Fault Domain

Fault domains define the group of virtual machines that share a common power source and network switch. In our deployments, the virtual machines configured within your availability set are separated across up to three fault domains.

Opposite

Proximity Placement Groups

Let's say latency is an extreme concern and you would like the exact opposite of geographic redundancy. A Proximity Placement Group would meet that need. When you assign your virtual machines to a proximity placement group, the virtual machines are placed in the same data center, resulting in lower and deterministic latency for your applications.

<https://azure.microsoft.com/en-us/blog/introducing-proximity-placement-groups/>

<https://azure.microsoft.com/en-us/blog/announcing-the-general-availability-of-proximity-placement-groups/>

<https://docs.microsoft.com/en-us/azure/virtual-machines/linux/co-location>