2018

Global Azure BOOTCAMP

High Availability and Scalability

Balu Kalepu

Balu.Kalepu@ValueMomentum.com

VALUEMOMENTUM

Deliver Value. Drive Momentum.

Corporate Overview – ValueMomentum Software Services

- Software & Services Firm
- **Financial Services & Insurance focused**
- Established in 2000 with HQ in NJ, USA
- 150+ dedicated R&D team
- Executive Leadership and Practice Heads based in the US
- Offshore centers are SSAE 16 SOC 2 certified. Clean Rooms for several clients offshore

23%
Compound Annual Growth Rate since 2000

4
Analysts covering ValueMomentum Software & Services

>65
Clients Served in North America

1,850+
Global employee strength

Top 15
IT Services Vendor for North American P&C Carriers by # of customers*

14
>5 Year Customer Relationships Average ~8 years

BUSINESS FOCUS

- Banking & Lending
- Capital Markets

- Property & Casualty
- Healthcare
- Life & Annuities

Session Agenda

- What is HA?
- Azure Resiliency
- Availability Sets
- Availability Zones
- What is Scalability?
- Azure Compliance
- Azure Trust Center
- Questions

High Availability

Design Highly Available Applications on Azure

What is High-Availability

HA ensures System is operational without interruption

Availability is often measured as a percentage

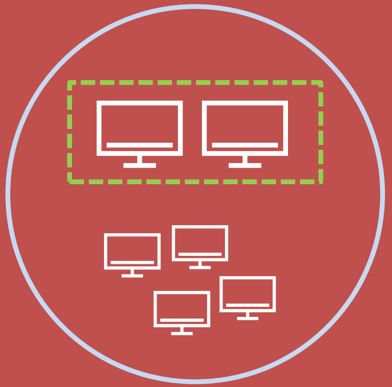
HA provides maximum uptime to achieve SLA

Eliminates Single Point of Failure

Functions as failure response mechanism

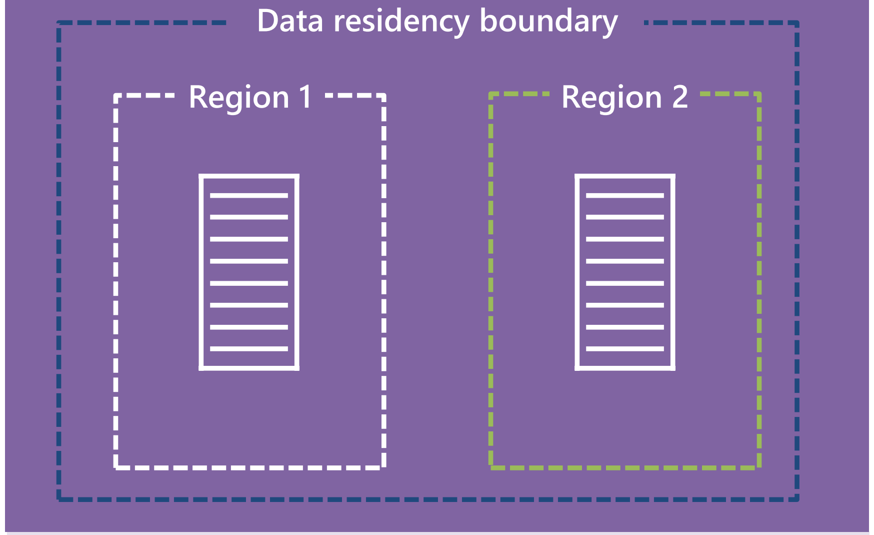
SLA is the key Metric to design High Availability

Azure Resiliency Today



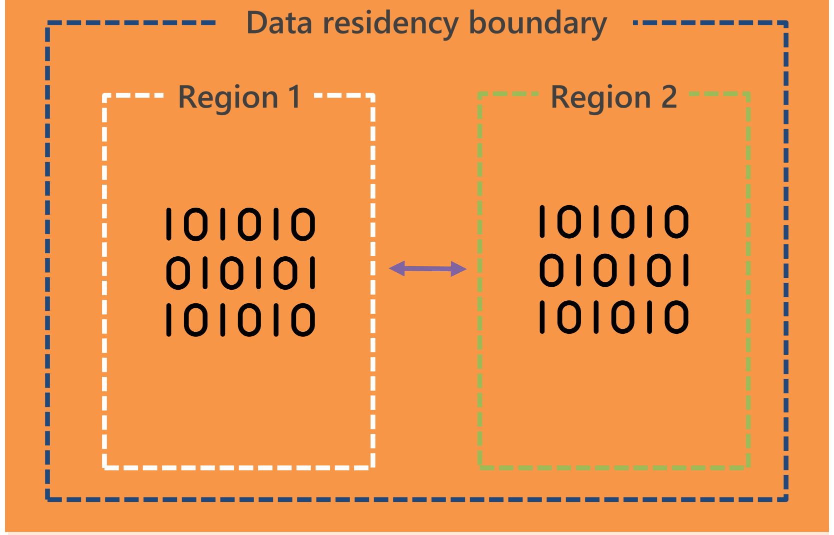
High availability

High Availability using Availability Sets for protection from hardware failures in a datacenter.



Disaster recovery

Replication from one region to another, with standby VMs in the other region. Azure offers protection between regions within data residency boundaries.



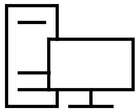
Backup

Data is asynchronously replicated and stored for redundancy purposes with data residency options.

Most Comprehensive Resiliency

INDUSTRY-ONLY

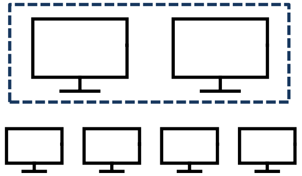
VM SLA
99.9%



SINGLE VM
Protection with Premium Storage

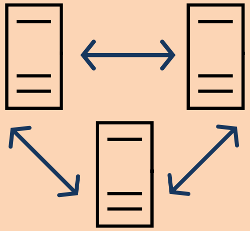
INDUSTRY-LEADING HIGH AVAILABILITY SLA

VM SLA
99.95%



AVAILABILITY SETS
Protection against failures within datacenters

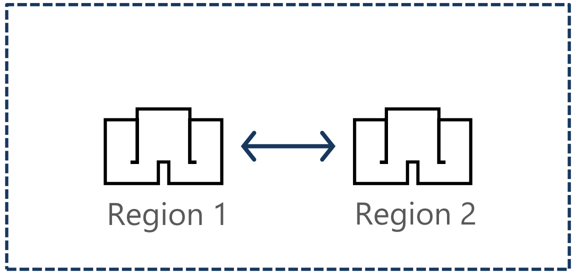
VM SLA
99.99%



AVAILABILITY ZONES (Preview)
Protection from entire datacenter failures

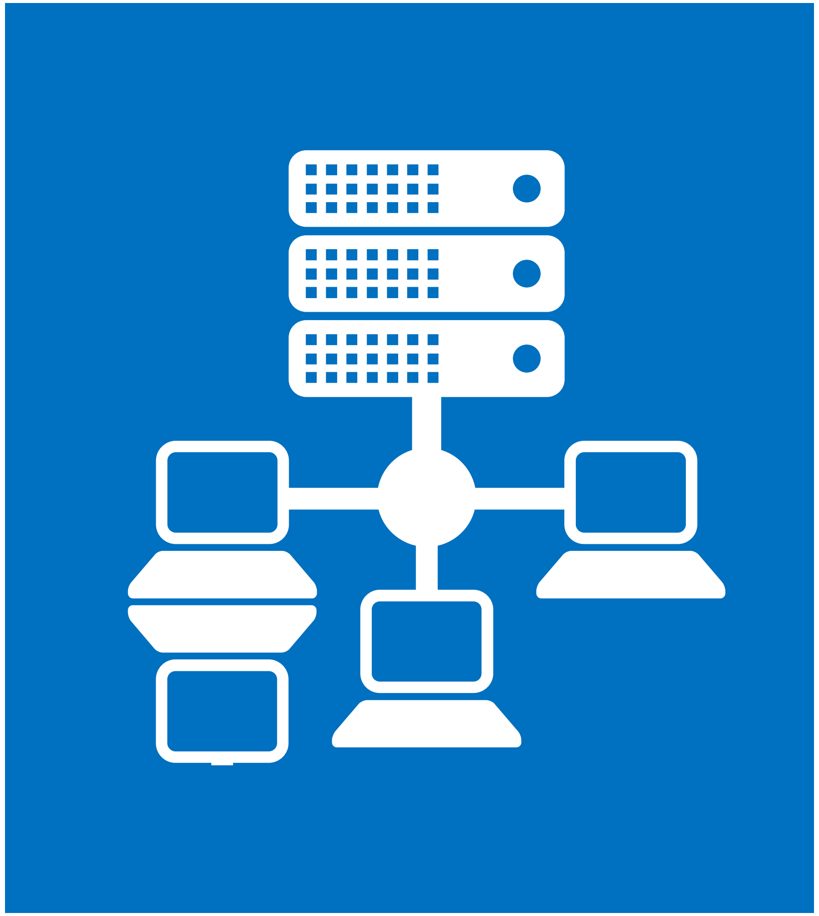
INDUSTRY-LEADING DISASTER RECOVERY

REGIONS
42



REGION PAIRS
Protection from disaster with Data Residency compliance

SLA – What it means?

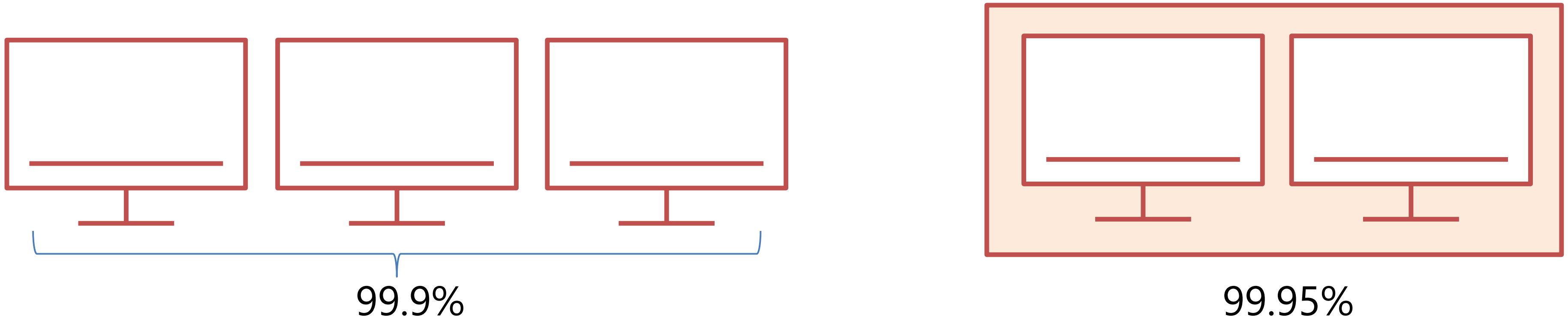


SLA	Monthly Downtime	Yearly Downtime
99.0 (Two Nines)	7h 18m 17.5s	3d 15h 39m 29.5s
99.5%	3h 39m 8.7s	1d 19h 49m 44.8s
99.9% (Three Nines)	43m 49.7s	8h 45m 57.0s
99.95%	21m 54.9s	4h 22m 58.5s
99.99% (Four Nines)	4m 23.0s	52m 35.7s
99.999% (Five Nines)	26.3s	5m 15.6s

Usually 99.99 (Four Nines) is considered as a Industry leading uptime

Availability Set

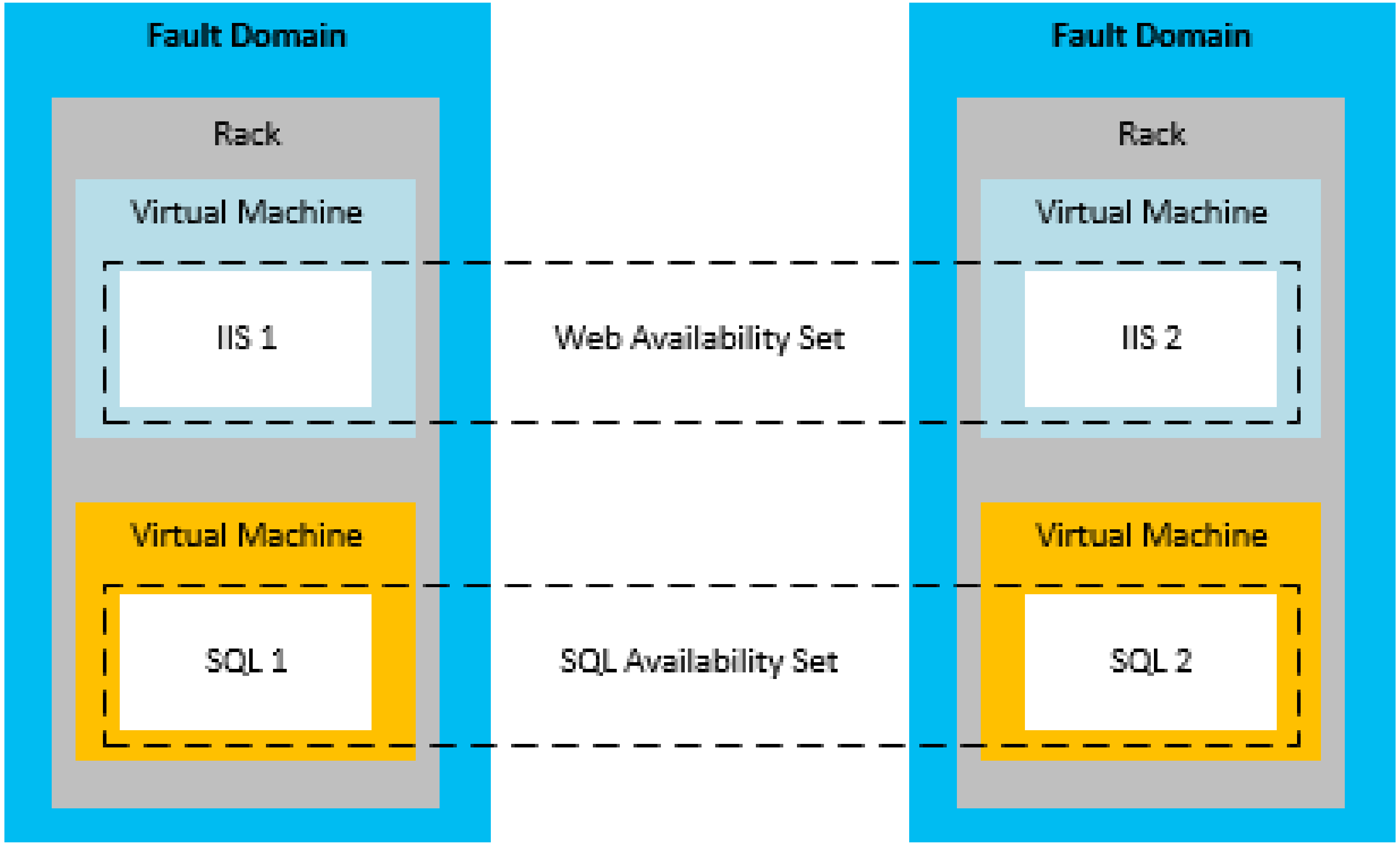
- Ensures one instance will be online all the time
- Span across Fault Domain and Update Domains
- Mitigates the risk of Unplanned and Planned downtimes
- Guarantees 99.95% SLA If two or more VMs are deployed in same AS
- Azure Offers Single VM SLA as 99.9%



Fault Domain and Update Domains

- Fault domains:
  - Represent groups of resources anticipated to fail together, i.e. same rack, same server, same switch
  - The number of fault domains is controlled by the Azure Fabric
  - 3 fault domains by default
- Update domains:
  - Represents groups of resources that will be updated together
  - Host OS updates honour service update domains
  - Default of five (up to 5)

About Fault Domain



Reference Architecture

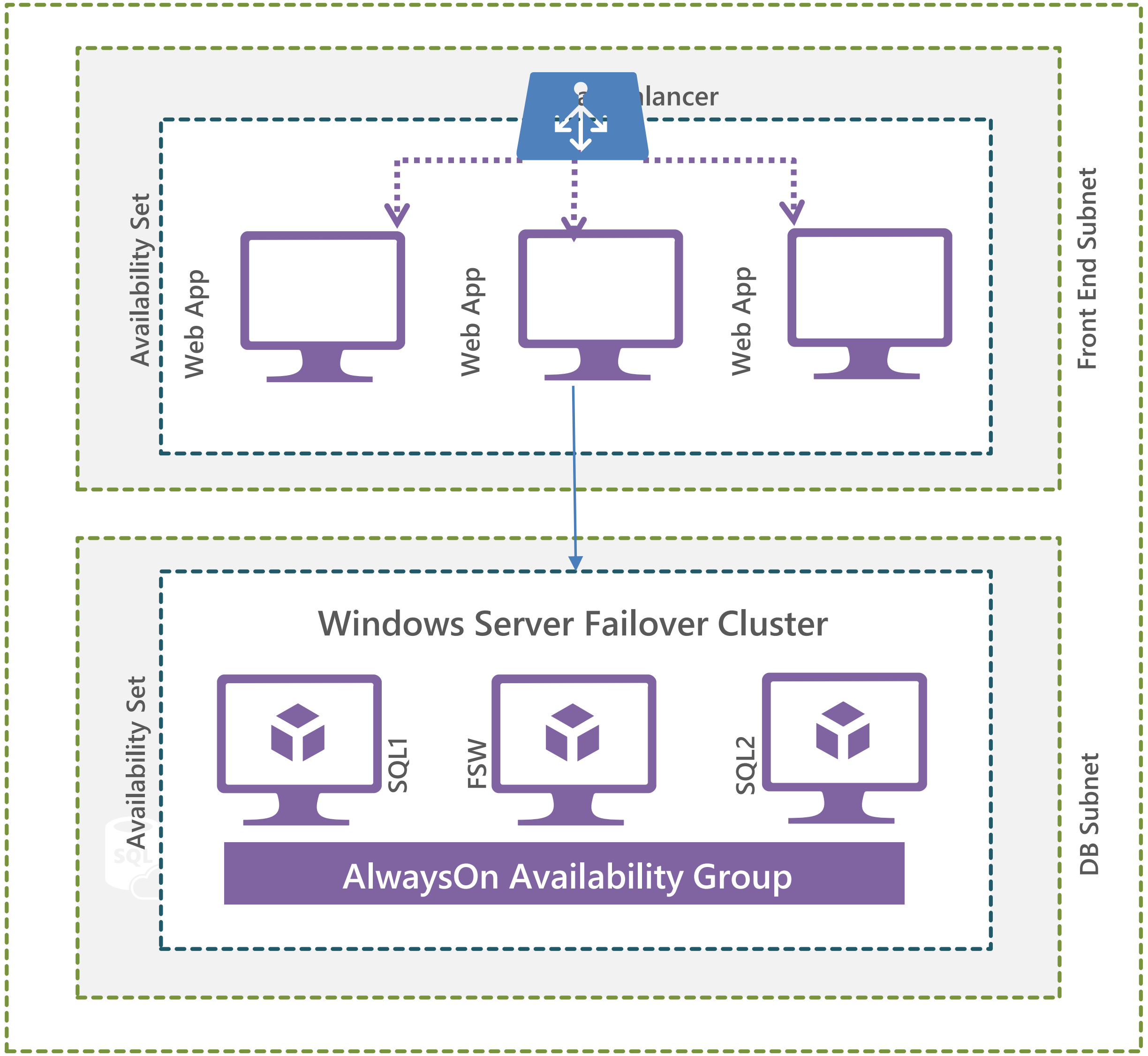
Web frontend in an Availability Set

Load balanced across the VMs within Availability Set

Data layer redundant in AlwaysOn Availability Set

Windows Server Failover Cluster

SQL AlwaysOn Availability Group



Availability Zones

Part of Azure's native HA/DR solutions

Provides protection from Datacenter failure

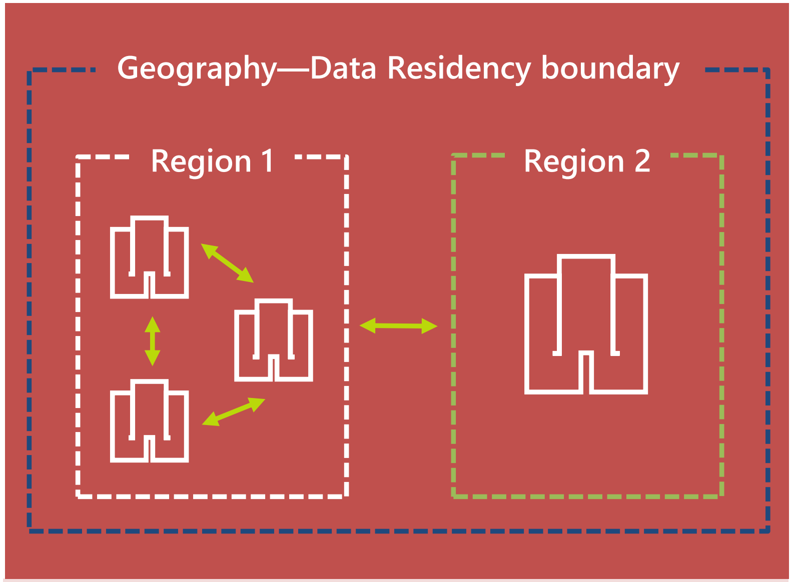
Currently Supports VM, VMSS, Managed Disks, IPs and Load Balancers

GA in US Central and France Central

PP in East US 2, West Europe and Southeast Asia

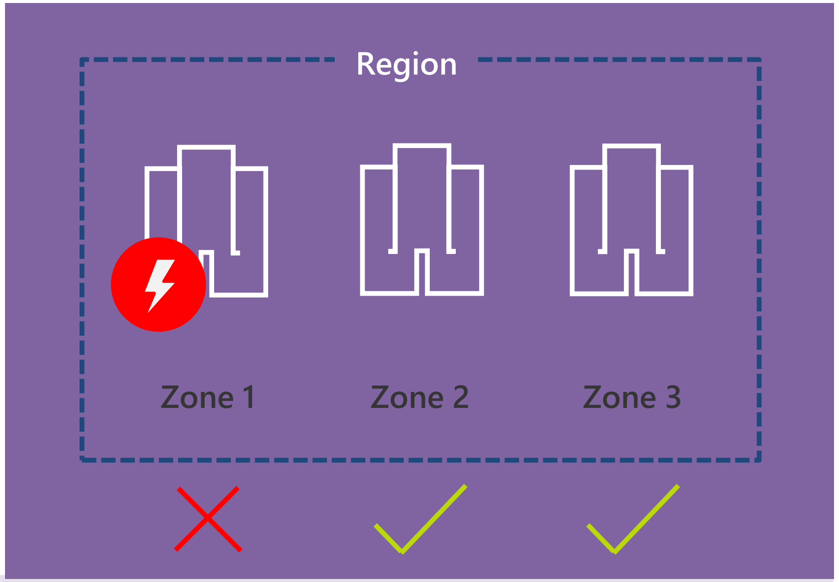
More regions and services coming soon

Availability Zones



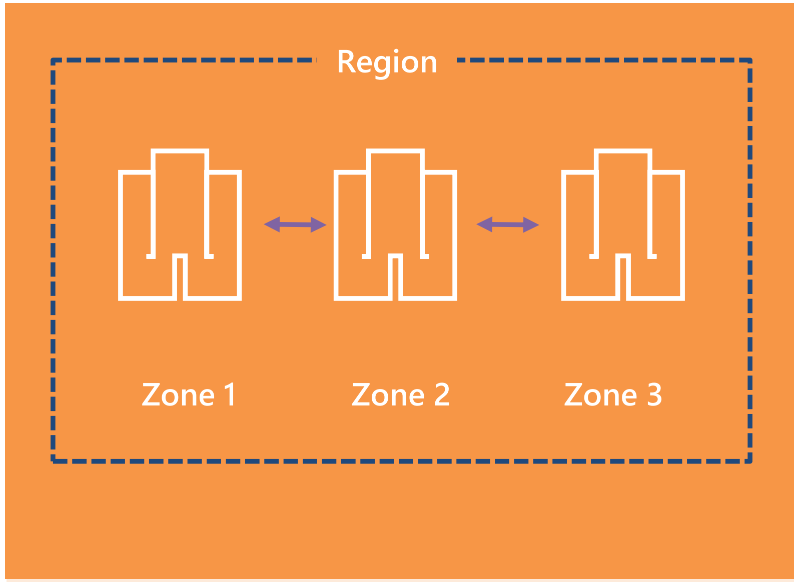
Achieve full resiliency with Data Residency

Availability Zones and a paired region within the same data residency boundary provides high availability, disaster recovery, and backup.



Protect against entire datacenter loss

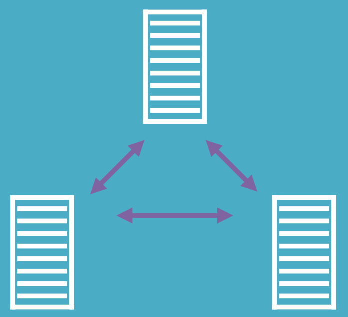
Each zone is physically separated with independent power, network, and cooling and logically separated through zone-isolated services.



Run mission-critical apps with 99.99% SLA at GA

High Availability supported with industry best SLA when VMs are running in two or more Availability Zones in the same region.

Protection through Redundancy

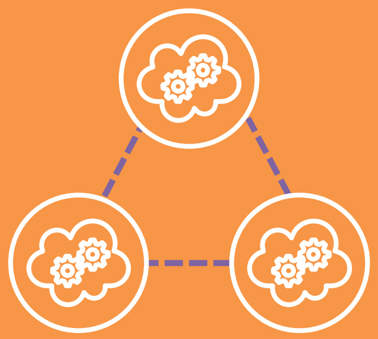


Minimum of three physically separated locations

Three zones to support Quorum based workloads like SQL, Service Fabric, Cassandra, MongoDb.

Independent power, cooling, network

A facility level failure or single fiber path failure will affect only a single AZ.

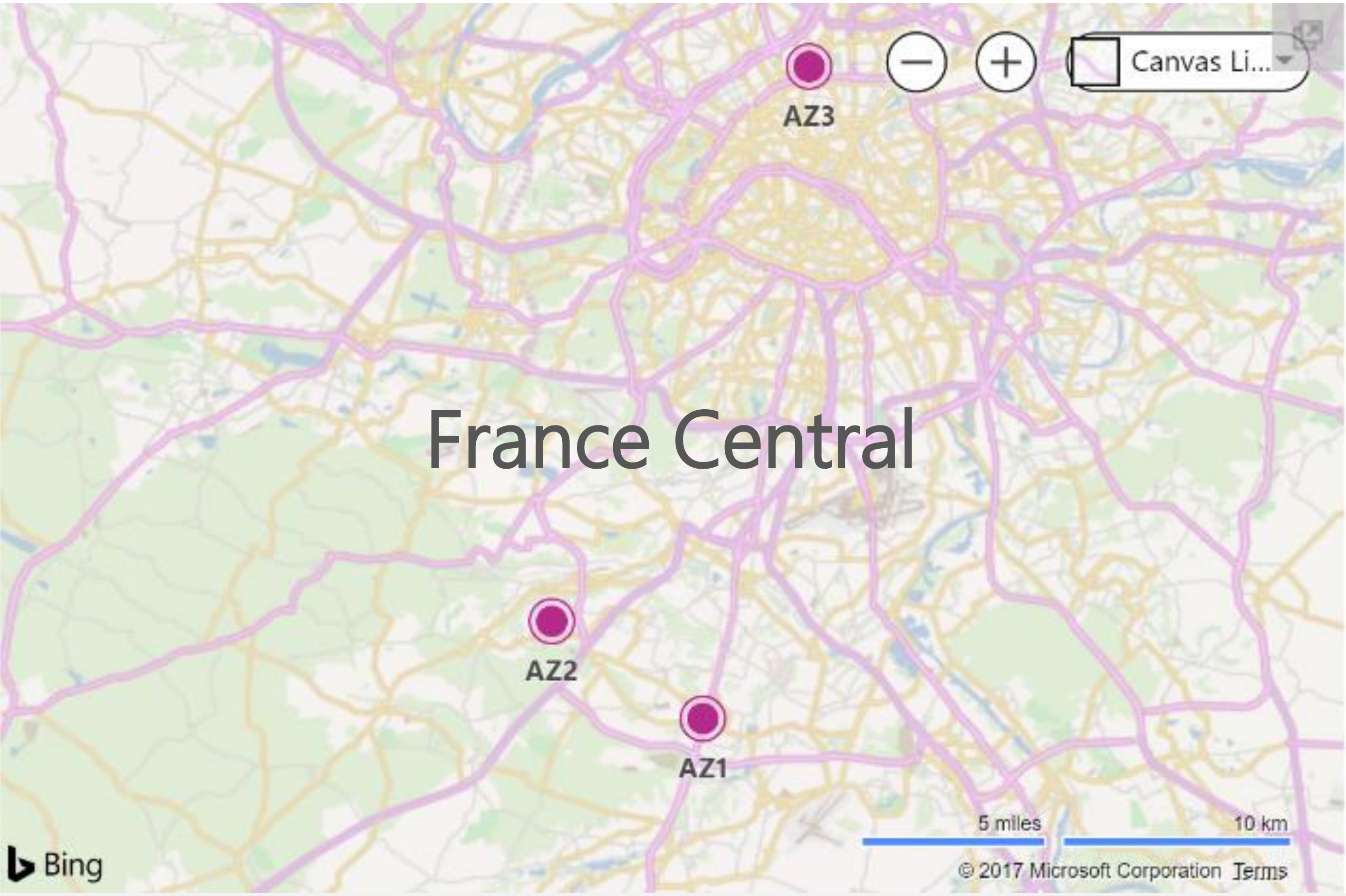


Azure management services replicated across zones

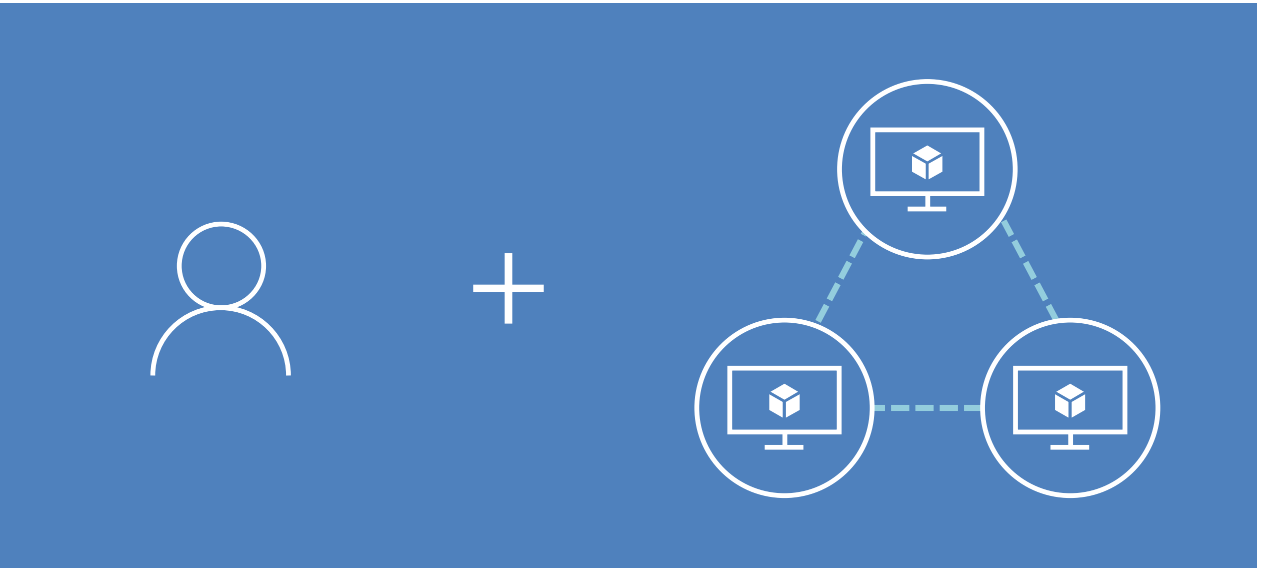
The management services are redundant and single zone failure will not affect availability. Updates orchestrated zone-by-zone.

Region Topology

- Fault-isolated locations within an Azure region
- Independent power, network, and cooling
- Protection against physical and logical failures



Zone Aware Services



Zonal Services - Customer pin to AZs

VMs

VM Scale Sets

Managed disks

VIPs

Zone Redundant Services replicate across 3 AZs

SQL DB

Cosmos DB

Web Apps

Application Gateway

...and more

HA of PaaS Services

Azure PaaS services are HA Enabled

Offered with Independent Service SLA

Azure SQL 99.99%

Azure CosmosDB 99.99%

Azure WebApps 99.95%

Azure Function Apps 99.95%

Reference Architecture

Web frontend across 3 AZs

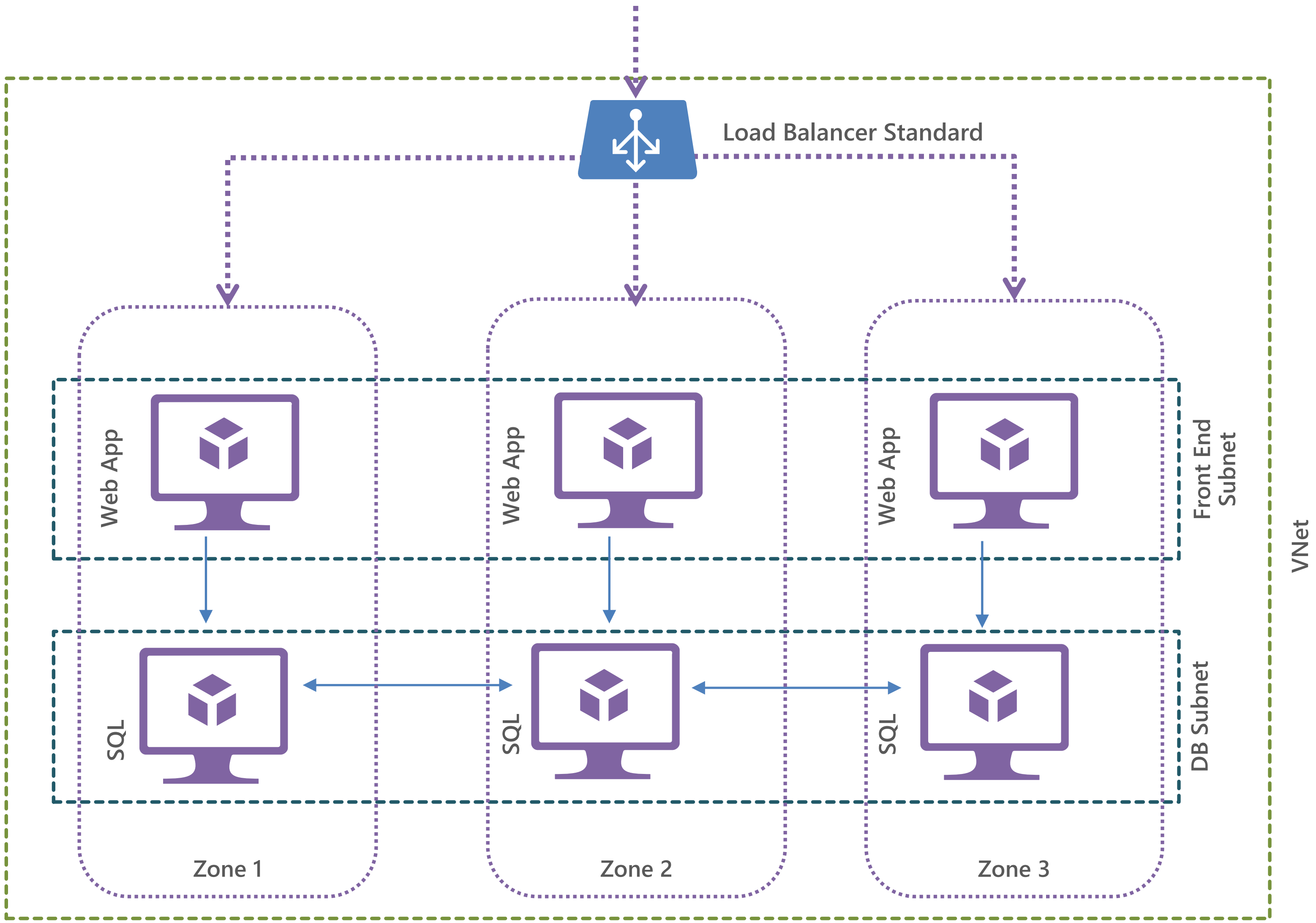
Load balanced across VIPs using Standard Load Balancer

Data layer redundant across 3 AZs

SQL on IaaS

SQL Azure or CosmosDB

NoSQL (Cassandra, MongoDB)



Demo

Let’s see it in action

VALUEMOMENTUM

Deliver Value. Drive Momentum.

Scalability

What is Scalability?

Scalability is a term used to describe how the application will handled increased loads of traffic volume

Also provides Application Resiliency

Allows to optimize the Azure cost

Application performance will not be disturbed

Called as Elasticity and Works at large-scale

Vertical Vs Horizontal

Vertical

“Scale-Up” and “Scale-Down”

- Increasing the Hardware Resources without changing the number of nodes
- Simple to Implement
- Finite Limit
- Required down-time

Horizontal

“Scale-Out” and “Scale-In”

- Increasing the Number of Nodes, Distribution of load through a Load Balancer
- Moderate efforts to implement
- Almost Infinite Limit
- No down time require

Scalability on Azure

- Almost all of Azure services supports “Scalability”
- Virtual Machine supports Auto Scale though VM Scale Sets
- Azure App Service Supports Scale-Up and Scale-Out
- Horizontal and Vertical Scaling can be achieved on AzureSQL

Scalability - Azure Scale Sets

- ❖ Dynamically Increase or decrease the number of VMs based on resource consumption
- ❖ Automatically creates other required resources such as Load Balancers, Networks and so on
- ❖ Scale Sets also works with Availability Zones
- ❖ A single scale set may contain up to 1,000 instances. However, there are few restrictions - Reference
- ❖ Custom Images can be used as Base Image
- ❖ Supports CustomExtentions to install or configure the VM during Startup
- ❖ Configurable "Cool Down" Timers
- ❖ Uses a technique called "flapping" while performing Scale-In

Demo

Let’s see it in action

Questions?

VALUEMOMENTUM

Deliver Value. Drive Momentum.(SM)

Thank You

Balu Kalepu

Technical Architect

Balu.Kalepu@valuemomentum.com