Microsoft Azure

Business Continuity and Disaster Recovery in Azure

Making sure your BCDR plan prepares your organization to thrive even in times of crisis

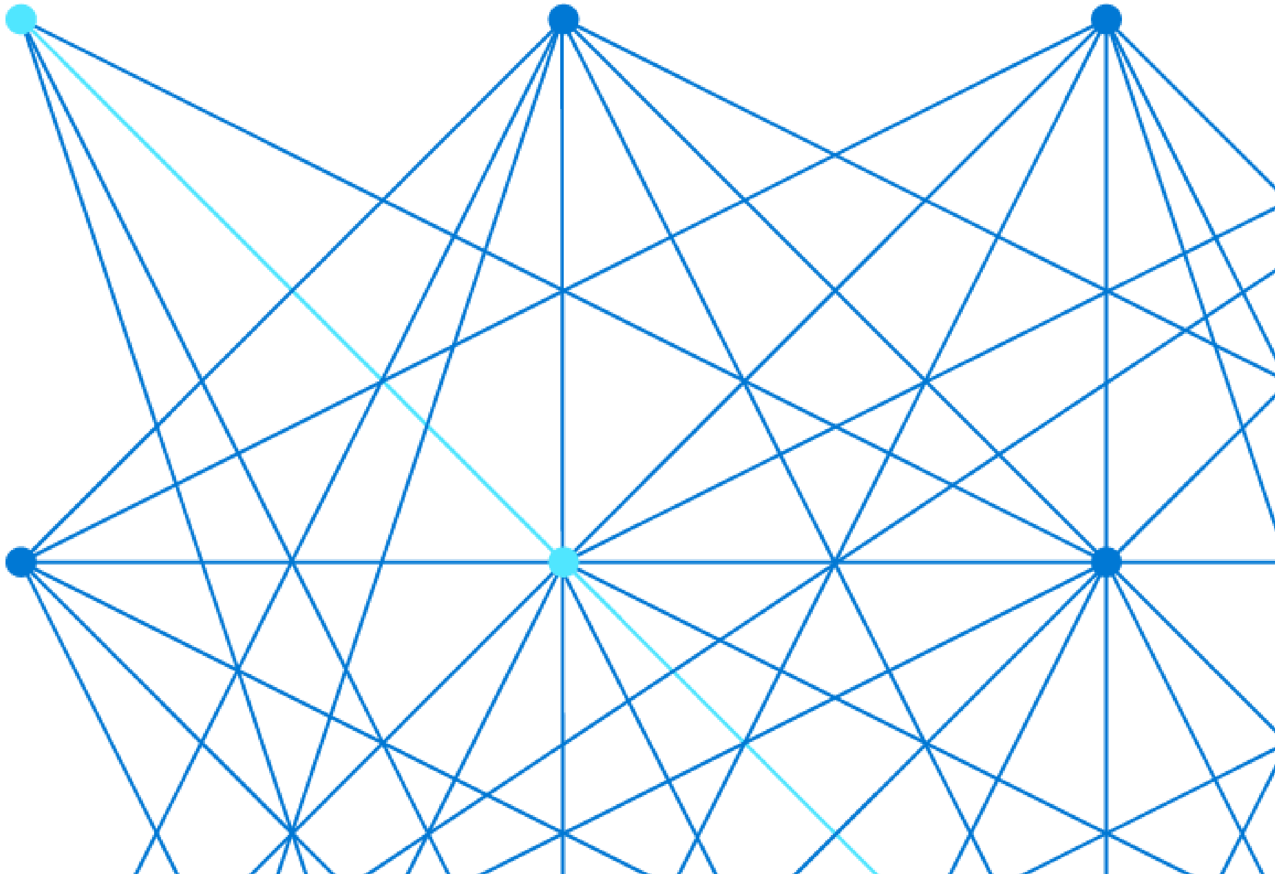

Table of Contents

Introduction

Catastrophic events are not just a possibility, but they are an eventuality that companies need to prepare for. Businesses are supported by a wide range of applications that run their day-to-day operations. Some of these applications may be critical to the business while others may not. Every minute of downtime and lost data can negatively impact the business and customer trust. To ensure that services can continue to operate during maintenance, failure, and even large-scale outages, most companies implement Business Continuity and Disaster Recovery plans (BCDR).

This whitepaper serves as a guide to help you craft a BCDR framework that fits your company's needs. It defines Business Continuity and Disaster Recovery, introduces common BCDR scenarios in Azure, and provides methodologies and solutions. This guide can help you make informed decisions regarding the Disaster Recovery services you choose and architecture your apps need for Business Continuity.

Understanding Business Continuity, Disaster Recovery, and other related functions

Most BCDR plans provide strategies for High Availability, Disaster Recovery, and Business Continuity. Each is an aspect of contingency planning that is vital to preparing a company for failures and outages. Since they are strongly connected to one another, many organizations lump them all under the same definition. As a result, conversations can run the risk of tackling one part of BCDR while neglecting another. This can lead to an insufficient BCDR plan that isn't prepared to mitigate most types of failure or to quickly recover with minimal impact to business operations. Each BCDR concept is described below so that you can more precisely identify your business needs and the specific requirements of your BCDR plan.

High Availability

High Availability is a design principle in which a given application or workload is distributed across redundant infrastructure, be it physical or virtual. The purpose of High Availability is to provide sufficient redundancy so that an application can tolerate the failure of one or more infrastructure components without impacting uptime. This is different from Disaster Recovery, which is intended to protect against large-scale outages that could involve the entire region where an application is deployed.

Disaster Recovery

Disaster Recovery is a combination of tools, backups, replications, and asynchronous processes used to recover workloads from large-scale outages that may affect an entire geographical region. Disaster Recovery is similar to High Availability, but it protects workloads across regions rather than in a single region.

Business Continuity

Business Continuity describes how an organization responds and maintains operations during a disaster. A Business Continuity system includes the processes, tools, and architectures built on top of DR to ensure that the most critical applications can continue to run in the event of a disaster. By comparison, DR ensures that you have sufficient infrastructure and processes to allow you to recover your applications in the event of a disaster. This is typically reserved for the workloads that can tolerate the least amount of downtime or data loss.

Backup

Backup solutions allow applications to back up their data and servers from a previous version to prevent data loss. This also allows applications to recover more quickly from local failures. Backup is typically used for operational restoration and data loss only and not long-term retention. Long-term retention of data should be done through a data archival service or tool. While backup is not specifically aligned with Disaster Recovery, there can be some overlap between the two functions. If the backup is replicated to another geographic region, it can be used as a solution depending on your RTO/RPO needs for specific applications. Typically, geo-replicated backups don’t meet the DR requirements for every application in an organization.

Disaster Recovery planning

All versions of additional layers of protection for an application increase consumption cost, but not every application may require the same level of protection based on how important that application is to the business.

Most organizations reduce costs and management complexity by standardizing protection groups based on how much downtime and data loss applications in this group can tolerate. This is an alternative to always paying premium protection costs or managing a different protection plan for each application. Protection groups can be visualized by mapping each application to a Disaster Recovery tier.

DR tiers are organized by a set of predefined recovery time objective (RTO) and recovery point objective (RPO). Applications that need to have lower RTO and RPO are mapped to more critical DR tiers.

- **Recovery time objective (RTO):** The maximum length of time before a lack of functionality severely impacts the business.
- **Recovery point objective (RPO):** The point in time to which lost data must be restored. It is also known as maximum acceptable data loss.

These will differ based on the company and its business requirements.

The following is an example of what a DR tier table may look like for a given organization.

The lowest DR tiers (often 0 or 1) represent the most business-critical applications that require the highest fidelity solutions. Fidelity refers to the level of detail and thoroughness of the solution, which can make high fidelity solutions costly but resilient. The DR tiers with the largest number (typically between 5 and 10) represent the least critical applications that require less protection and can experience extended outages or data loss with little to no business impact.

DR tier	RTO	RPO
Tier 0 – Most critical	5 minutes	5 minutes
Tier 1	4 hours	6 hours
Tier 2	12 hours	24 hours
Tier 3	24 hours	2 days
Tier 4 – Least critical	7 days	7 days

Business Continuity and Disaster Recovery scenarios

Understanding what should be considered in your BCDR plan can be challenging when there are so many options and factors. Azure has numerous tools to support BCDR plans. The situations that most plans need to address for protection are data and applications hosted in the cloud or data and apps hosted by on-premises datacenters.

The following are two different scenarios based on where the application or IT asset that needs protection is hosted:

- **Protect in Azure:** If you are migrating on-premises workloads to Azure or you already run workloads in Azure and you need to protect them against disasters, you can use one or more additional Azure regions as your BCDR failover locations.

- **Protect to Azure:** If you run your workloads on-premises and you need to protect them against disasters, you can use Azure as your BCDR failover location instead of another self-hosted datacenter.

Scenario 1 - Protect in Azure

When Azure is the primary location for your application, you can utilize Azure's offerings to set up a secondary Azure region for failover. The method used to achieve your desired BCDR goals may include a combination of Azure services, application architecture, and operational processes. The best choice for you will depend on the RTO/RPO requirements for each workload you want to protect.

- Most solutions will require a passive region. Part of RTO considerations stem from how quickly the passive region can be activated. Using On-Demand Capacity Reservation allows you to reserve compute capacity for passive regions so that the capacity to run VMs is guaranteed to be available. You can also combine existing or new Azure Reserved VM Instances with on-demand capacity to lower total cost of ownership. When you aren't using the capacity reserved for DR, you can run additional workloads.

For illustrative purposes, this document groups a few different levels of fidelity that a given application may require into the following groups:

- **DR tier for highly business-critical applications (lowest RTO/RPO):**
  The highest priority/mission critical applications running in Azure that can tolerate minimal downtime or data loss before negatively impacting the business.
- **DR tiers for moderately business-critical applications (moderate RTO/RPO):**
  Applications that are important to the business but can tolerate some downtime or data loss without negatively impacting the business.
- **DR tiers for non-business-critical applications (high RTO/RPO allowances):**
  Applications that are non-critical and can tolerate hours (if not days) of downtime and data loss without significantly affecting the business. They still require DR to ensure that they are not permanently lost in the event of a disaster.

Highly business-critical DR tier options

Applications and services that need the most protection need to either run in multiple regions at once or failover to another region with minimal delay after failure. Managing traffic for both regions' public-facing services requires additional Azure tools. Azure Front Door or Traffic Manager can support geo-load balancing while Azure Load balancer or Application Gateway can help with local load balancing. For more details on these tools, check the Azure services and capabilities list.

The following DR options allow an application to operate with a minimal amount of downtime or data loss in the event of a disaster. Note that to achieve the lowest

RTO/RPOs possible, the ability to run in multiple geographic regions must be part of the application's architecture. This option needs more than just an outside tool or service. The following two solutions describe application architectural design patterns that can be followed to achieve extremely low RTO/RPOs.

Active/Active

In an Active/Active model, an application is running in multiple regions and traffic is distributed across them. Each region can handle the added load from the other region should it go offline. This option can provide the best RTO and RPO which can achieve near real-time RTO and RPO. If one region becomes unavailable, the other region will take over.

However, the Active/Active process only works for applications that can serve requests out of multiple regions at once and perform bi-directional replication between the regions. Bi-directional replication is not necessarily supported by every database engine and multi-region application architecture can be difficult to implement.

Active/Passive Hot Standby with Automated Failover

In an Active/Passive Hot Standby with Automated Failover DR model, the primary region handles all the traffic for the application. The secondary region is built out to the same specifications as the primary and left running, so it's ready to take over should the primary region fail.

This option provides a very low RTO of near real-time after a disaster is declared. RPO can also achieve near real-time, but this will be dependent on the synchronization technology employed by the application's data layer; Azure technology builds the regions, but the application architecture needs to be designed to support this scenario. Failover will be executed through automation based on application monitoring. This option is much easier to implement for applications than the Active/Active model.

Moderately business-critical DR tier options

As the RTO/RPO requirements loosen for less and less critical applications, so too does the requirement for applications to be specifically designed to run in multiple geographic locations. Instead, applications can rely on tools in Azure that are designed to duplicate VMs from one region to another.

Options for applications with built-in DR support

Built-in DR support has some lower cost solutions. Using careful VM management and architectural patterns in passive regions can save on consumptions costs for moderately business-critical applications.

For instance, the Azure service Virtual Machine Scale Sets in flexible orchestration mode can manage VMs in passive regions based on the requirements of an application at any time for automated failover. Flexible orchestration mode enables automatic distribution of VMs across multiple Availability Zones for High Availability. For more details on Virtual Machine Scale Sets flexible orchestration mode, check the Azure services and capabilities list.

In addition to properly managing VMs, DR support for this tier option requires intentional design. Using a passive region that consumes less capacity can save on costs and be much easier to design compared to the Active/Active model. However, these options are still based entirely on application architectures and require the application to be capable of being deployed to multiple geographic regions. The following options describe architectural patterns for moderately business-critical applications:

Active/Passive Warm Standby

In Active/Passive Warm Standby, the workloads are identically deployed to both regions, but only one region is active at a time. the standby region is “pre-warmed.” This means that the VMs in the secondary region are deployed at a lower capacity and are ready to take transactions.

Depending on how the application is architected, this may involve running fewer instances of specific VMs than the primary region (scale out), the same number of VMs but as smaller, less expensive VM sizes (scale up), or some combination of both. This solution is not designed to take the full production load until the event of failover occurs. Otherwise, it is fully functional. The scaling can occur as part of a manual process or as a part of an automated failover process (e.g., Virtual Machine Scale Sets can scale out automatically based on load).

This model may be used for services designed to fail over to the secondary region automatically (as a result of monitoring, like in the previous scenario) or through a manual decision to fail over. However, an automated failover process would be responsible, by necessity, for rehydrating the warm standby environment to full capacity as part of that failover. Virtual Machine Scale Sets can support the failover process by scaling the VMs manually or automatically as your situation may require.

Active/Passive Cold Standby

In an Active/Passive with Cold Standby DR model, a secondary region is configured for a given application but the resources in that region are either turned off (thereby not incurring compute consumption costs) or not yet deployed. The standby region would turn on resources when the failover occurs. Once all the services are up and running, the region then takes over the load from the primary region. This failover mechanism is cost-effective but provides moderate RTO (typically measured in hours).

The data layer of an application would need to be running continuously to receive updates, or a replication technology such as Azure Site Recovery would need to be used to keep the secondary region's data in sync. Stateful data that needs to exist in the secondary region will be charged for, but the intent of this tier is to minimize the compute costs necessary for DR.

Options for applications without built-in DR support

The following options use Azure services or capabilities to duplicate VMs running in Azure between two different Azure regions. VM cloning and restoring VMs from backups across regional boundaries offer additional ways to protect workloads in Azure without the need for application specific architecture supporting DR. These options are suitable for commercial off the shelf software or other applications where the organization doesn't control the architecture.

Azure Site Recovery

Azure Site Recovery is a VM replication technology within Azure that is used for "lift and shift" migrations and for Disaster Recovery. Through source and destination configuration, Azure Site Recovery can replicate data from any Azure region to another. For more information on Azure Site Recovery, check the Azure services and capabilities list.

An application doesn't need to be multi-region capable, or have a mature deployment process for Azure Site Recovery to function. Azure Site Recovery can be deployed across all the VMs in a given application, including the data tier, provided the databases are not too large or have a high volume of change.

Azure Backup

Azure Backup is a service to perform operational backups of VMs and other storage-related services and database engines. It provides the following benefits:

- Fabric-level snapshots of VMs
- Highly efficient data storage that only charges consumers for the difference in data
- Short- and long-term retention of backup data
- A native Azure solution that can easily scale to meet your needs

When backing up VMs, Azure Backup offers a capability called Cross Region Restore which allows backups to be replicated to a second region, this technology leverages Geo-Replicated Storage and can only replicate to the Azure-designated BCDR paired region. Any regions outside of the pair require a manual migration of VMs to the region which adds additional time to DR calculations. For more info on Azure Backup and its Cross Region Restore feature, check the Azure services and capabilities list.

Both Azure Site Recovery and Azure Backup are only capable of replicating VMs from one region to another, which means they don't protect platform as a service (PaaS) products across regions although many PaaS services offer their own native multi-region capabilities. Nonetheless, Azure Backup can support the following services:

- Azure VMs or individual managed disks
- Azure Files (Storage Accounts)
- Azure Blob Storage (Storage Accounts)
- SQL Server running on Azure VMs
- SAP HANA databases running on Azure VMs

Azure Backup also supports the following capabilities

- **SQL Server in Azure VMs:** Backs up SQL Server databases running on Azure VMs and restores them to the Azure paired region.
- **SAP HANA databases in Azure VMs:** Backs up SAP HANA databases running on Azure VMs and restores them to the Azure paired region.

Independent Software Vendor (ISV) options

Some Independent Software Vendor options that run on Azure can handle backup in similar ways to Azure Site Recovery or geo-replicated backup services. In some cases they may also provide features relevant to your BCDR needs. Here are a few factors to consider when evaluating the use of ISV options:

- Your organization may already use one or more of these services in your existing datacenters. Extending the existing ISV infrastructure may provide a solution that brings less operational overhead than learning and managing a new set of tools
- Some ISV solutions may offer data deduplication capabilities. While native services like Azure Recovery Vaults offer very efficient methods of data storage, specific features like data deduplication may result in lower overall storage costs depending on the application.
- ISV tools may offer features like application or database consistent backups for products. ISVs and Azure Backup together offer a versatile set of options.

The following is a non-exhaustive list of ISV options that integrate with Azure:

- Commvault
- Dell
- Rubrik
- Veeam
- Veritas
- Zerto

Non-Business-critical DR tier options

Azure Backup

Azure Backup offers a capability called Cross Region Restore, which allows backups to be replicated to a second region. This technology leverages Geo-Replicated Storage (GRS) and can only replicate to the Azure-designated pair region. If the Azure pairs are not being used for a given application, then the time required to restore the VM to the Azure-paired region and the time required to manually migrate that VM to the Customer paired region must be factored into any DR calculations.

Both Azure Site Recovery and Azure Backup are only capable of replicating VMs from one region to another, which means they don’t protect PaaS products across regions.

Redeployment of application to secondary region

This method involves replicating data from the primary region to a secondary region, but other components of the application (such as the middle or web tiers) are not deployed in the secondary region at all. The only components that are running are the ones that are required for data replication.

At the time of failover, the rest of the application components are deployed into the second region. This is typically done through automated build pipelines that deploy the infrastructure as code along with the application components that then run on that infrastructure.

This provides one of the least expensive DR plans because it deploys a bare minimum amount of infrastructure and services to the secondary region when not in use, but it requires a more sophisticated deployment model for the applications and can take longer than warm/cold standby since new services need to be deployed as part of the failover event.

Scenario 2 - Protect to Azure

Azure can be used as a secondary, highly-resilient failover location to run workloads or store data in your existing datacenters. This comes with a few advantages to consider before using an additional on-premises datacenter as a failover location:

- **Consumption-based cost**
  - Unlike on-premises datacenters where you pay all the costs up front (facilities, hardware, electricity, cooling, etc), Azure only charges you for what you are using or reserving at any given moment. If you need to reserve compute capacity (e.g., VMs) ahead of time, you can purchase On-Demand Capacity Reservations that provide SLA guarantees.
- **Built-in resilience**
  - The Azure fabric is far more resilient than the majority of on-premises datacenters.
- **Low- to no-impact native options**
  - Azure offers multiple native workload protection options by replicating live servers or server backups to Azure regions—with limited or no impact on production workloads.
- **Integration with top proprietary data protection**
  - Azure can support a number of on-premises data protection ISVs at various levels of depth and complexity.

Protect to Azure options

On-premises applications can use Azure an alternate datacenter location to protect against disasters. Through flexible Azure tools and certain ISVs, said applications can back up data to Azure and then have it replicated to another Azure region that functions, in essence, as an alternate datacenter location.

Azure Site Recovery

Azure Site Recovery is an agent-based tool that allows you to clone live servers into Azure as storage data or virtual disks. At the time of failover, Azure Site Recovery creates new VMs and attaches them to the virtual disks to create Azure-hosted clones of your VMs. For more information on Azure Site Recovery, check the Azure services and capabilities list.

The benefit of this method is that cloud providers charge based on consumption while with Azure Site Recovery the only consumption paid for prior to failover is the amount of data that is being stored in Azure plus a flat, per VM protection fee. There is no charge for data ingress into an Azure region, nor do you pay for compute consumption on virtual disks until the VMs are provisioned. Azure Site Recovery is an excellent option for protecting on-premises applications by keeping a duplicate of your workload in Azure. Another benefit of Azure Site Recovery is that since the VMs themselves are being cloned, the application that's being protected doesn't have to be specifically designed for DR, or in other words designed to run in multiple regions.

Independent Software Vendor options

Azure's Independent Software Vendor (ISV) options can function like Azure Site Recovery or geo-replicated backup services, but they also provide advantages that you may want to consider depending on your BCDR needs. Here are a few factors to consider when evaluating ISV options:

- Your organization may already use one or more of these services in your existing datacenters. Extending the existing ISV infrastructure may provide a solution that brings less operational overhead than learning and managing a new set of tools.
- Some ISV solutions may offer data deduplication capabilities. While native services like Azure Recovery Vaults offer very efficient methods of data storage, specific features like data deduplication may result in lower overall storage costs depending on the application.

ISV tools may offer features like application or database-consistent backups for certain third-party applications. ISVs and Azure Backup together offer a versatile set of options.

The following is an alphabetical but non-exhaustive list of ISV options that integrate with Azure to provide backup services:

- Commvault
- Dell
- Rubrik
- Veeam
- Veritas
- Zerto

Azure services and capabilities

The following list contains tools, products and services from Azure that can help implement BCDR plans:

High Availability services and capabilities

The following services and capabilities can increase the availability of applications:

- **Locally Redundant Storage**: Managed disk storage option that protects against failure by replicating data 3 times within your data center. This is the minimum level of protection Azure provides for every application.
- **Availability Sets**: Owner-created VM groupings that describe your application’s redundancy and availability framework to Azure. Virtual Machine Scale Sets in flexible orchestration mode offer this feature for no additional cost.
- **Availability Zones**: Similar to Availability Sets, but it creates more physical distance by spreading VMs and platform as a service (PaaS) services across physical buildings in an Azure region. Virtual Machine Scale Sets in flexible orchestration mode offers this feature at no additional cost.
- **Virtual Machine Scale Sets**: A service that updates and configures the number of balanced VMs based on demand to increase High Availability and scalability. With flexible orchestration mode, Virtual Machine Scale Sets can manage as many as 1000 VMs. Flexible orchestration provides the same High Availability as Availability Zones and Availability Sets. It can enable automatic distribution of VMs across multiple Availability Zones. It can also distribute VMs across multiple Fault Domains in a single Availability Zone.
- **Application Gateway**: Application-level service that routes web traffic to servers based on URLs. It works at Layer 7 (HTTP/HTTPS traffic) for public or internal facing services. Application Gateway presents a single endpoint for multiple VMs and distributes incoming requests across them to balance the load. It can automatically exclude failed VMs from load balancing based on health probes. The newly released V2 of Application Gateway adds new features like zone redundancy and autoscaling.

Azure Disaster Recovery fabric features

The following features support Disaster Recovery for application architectures:

- **Azure BCDR Regional Pairs**: Azure-designated groups of regions that share the same services, although service availability can differ. Disaster Recovery solutions that replicate and store applications in these pairs gain higher availability across multiple geographic zones.
- **Global VNet Peering**: Virtual networks can share resources and traffic with each other through Global VNet Peering. This differs from local VNet peering because the two VNets are in different Azure regions. Global VNet Peering has a higher data transfer rate than local VNet peering, but it is still the recommended method to connect software-defined networks in two regions. It helps with Disaster Recovery replication traffic. This isn't required for Geo-Redundant Storage the replication occurs at the Azure fabric level and does not traverse software-defined networks for the application.
- **Geo-Redundant Storage**: Storage option that replicates your data and stores it in multiple regions to protect against regional failure. This is typically replicated into the Azure Regional Pair. Three copies are written in the primary region while 3 copies are written into the secondary region. Data storage charges or Geo-Redundant Storage are roughly double that of Locally-Redundant Storage and accrue additional replication charges.

Global Load Balancing services

The following services can manage traffic across different levels of your applications:

- **Azure Load Balancer**: Cloud fabric level service for public preview that evenly balances traffic to VMs in an Availability Set. It balances loads by offering a single endpoint for multiple VMs and distributes incoming requests across them. Azure Load Balancer operates at Layer 4 of the OSI model (TCP/IP traffic). Azure Geo Load Balancers only work for public-facing Azure Load Balancers.
- **Traffic Manager**: DNS level service balances traffic for public-facing applications across multiple regions. It operates by returning CNAME (Canonical Name) redirects instructing the client to connect to the DNS name of one of the infrastructure components in the load balancing pool. Traffic manager is a non-regional Azure service, so it is not dependent on any specific Azure region.

- **Azure Front Door**: Service that balances traffic from HTTP requests for public-facing services only. It is a global content delivery network that can act as the “front door” for applications in any Azure region. Azure Front Door has 118 global endpoints distributed across 100 different metropolitan areas around the world. It is built-in layer 3 and 4 with distributed denial of service (DDoS) protection and Web Application Firewall capabilities.

Disaster Recovery products

The following are additional services that can be used to implement DR on a per-application basis:

- **Azure Site Recovery:** Agent-based tool for “lift and shift” migrations and Disaster Recovery. It clones workload data for replication into other regions. These copies are stored as virtual disks instead of virtual machines until the application needs to failover. At the time of failover, the Azure Site Recovery service will create new VMs and attach them to the replicated disks to create Azure-hosted clones of your VMs. Virtual disks don’t consume compute costs unlike VMs, so you only pay for storage until failover.
- **Azure Backup with Cross Region Restore:** Creates an operational restoration of workloads. Azure Backup’s Cross Region Restore feature is only available in Azure-designated pairs. The restore uses geo-replicated storage to replicate backups to the paired secondary region. Replicating to other regions requires a manual migration of VMs.
- **On-Demand Capacity Reservation:** Azure’s On-Demand Capacity Reservation for Azure Virtual Machines is a feature to reserve compute capacity ahead of actual VM deployments. When failover happens, the paid amount of reserved compute capacity is guaranteed to be available. On-Demand Capacity Reservation can be combined with Azure Reserved VM Instances to lower costs.

The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.

This white paper is for informational purposes only. Microsoft makes no warranties, express or implied, in this document.