What Happens When a Meteor Takes Out My Data Centre?

Peter Jakowetz - PrivSec Consulting Limited

So what do you do if a disaster occurs, and your service is suddenly unavailable?

Disasters?

- A disaster is an unexpected problem resulting in a slowdown, interruption, or network outage in an IT system. Outages come in many forms, including the following examples:
  - An earthquake or fire
  - Technology failures
  - System incompatibilities
  - Simple human error
  - Intentional unauthorised access by third parties
- These disasters disrupt business operations, cause customer service problems, and result in revenue loss. A disaster recovery plan helps organisations respond promptly to disruptive events and provides key benefits.

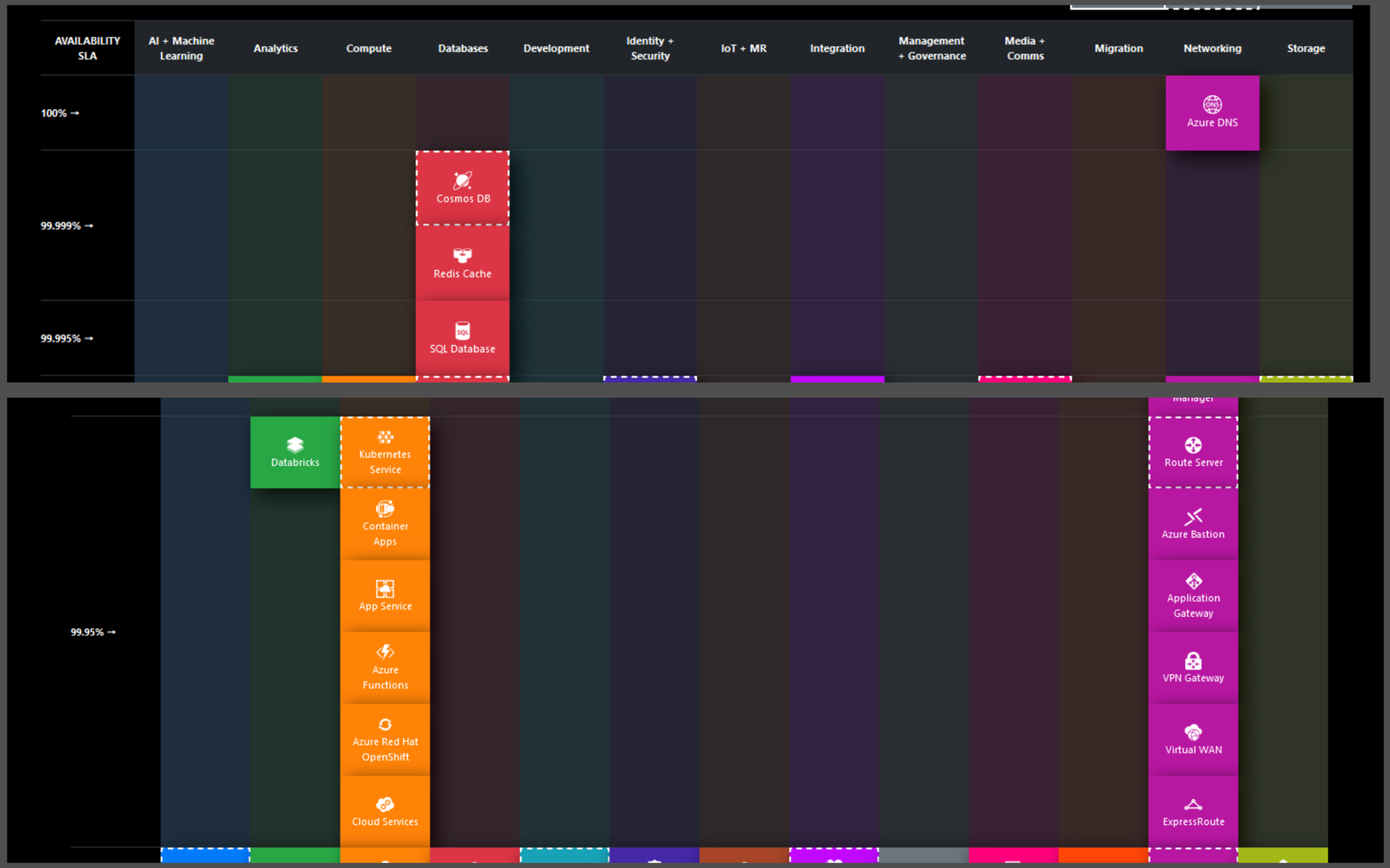
https://aws.amazon.com/what-is/disaster-recovery/

What is the problem here?

- We generally have a high confidence that our hosting providers will always be available
- The likelihood of a meteor strike might be low - but of a geographic outage for other reasons... a bit higher. The likelihood of a technical failure due to human error - much higher.

- What happens if the entire data center is unavailable?
- What about one key service?
- What if one of your servers fail?
- What about that as-a-service you didn't realise was key to your web application?
- Do you know exactly what services are reliant on other services within our cloud providers?

But I'm using AWS/ MS Azure. It's always available right?


AVAILABILITY SLA
AI + Machine Learning
Analytics
Compute
Databases
Development
Identity + Security
IoT + MR
Integration
Management + Governance
Media + Comms
Migration
Networking
Storage
100% →
99.999% →
99.995% →
99.95% →
Azure DNS
Cosmos DB
Redis Cache
SQL Database
Databricks
Kubernetes Service
Container Apps
App Service
Azure Functions
Azure Red Hat OpenShift
Cloud Services
Route Server
Azure Bastion
Application Gateway
VPN Gateway
Virtual WAN
ExpressRoute

Uptime Calculations

- 99.9% -->
  - Daily: 1m 26s
  - Weekly: 10m 4.8s
  - Monthly: 43m 28s
  - Quarterly: 2h 10m 24s
  - Yearly: 8h 41m 38s

- 99.95%
  - Daily: 43s
  - Weekly: 5m 2.4s
  - Monthly: 21m 44s
  - Quarterly: 1h 5m 12s
  - Yearly: 4h 20m 49s

https://uptime.is/

March 2023

23 **Post Incident Review (PIR) – Azure Resource Manager – West Europe** Tracking ID: RNQ2-NC8

6 **Post Incident Review (PIR) - Azure Storage - West Europe** Tracking ID: R_36-P80

1 **Azure Active Directory - AAD Authentication Issues** Tracking ID: XKW3-5T0

February 2023

Post Incident Review (PIR) - Multi-service outage – Asia-Pacific Area Tracking ID: VN11-JD8

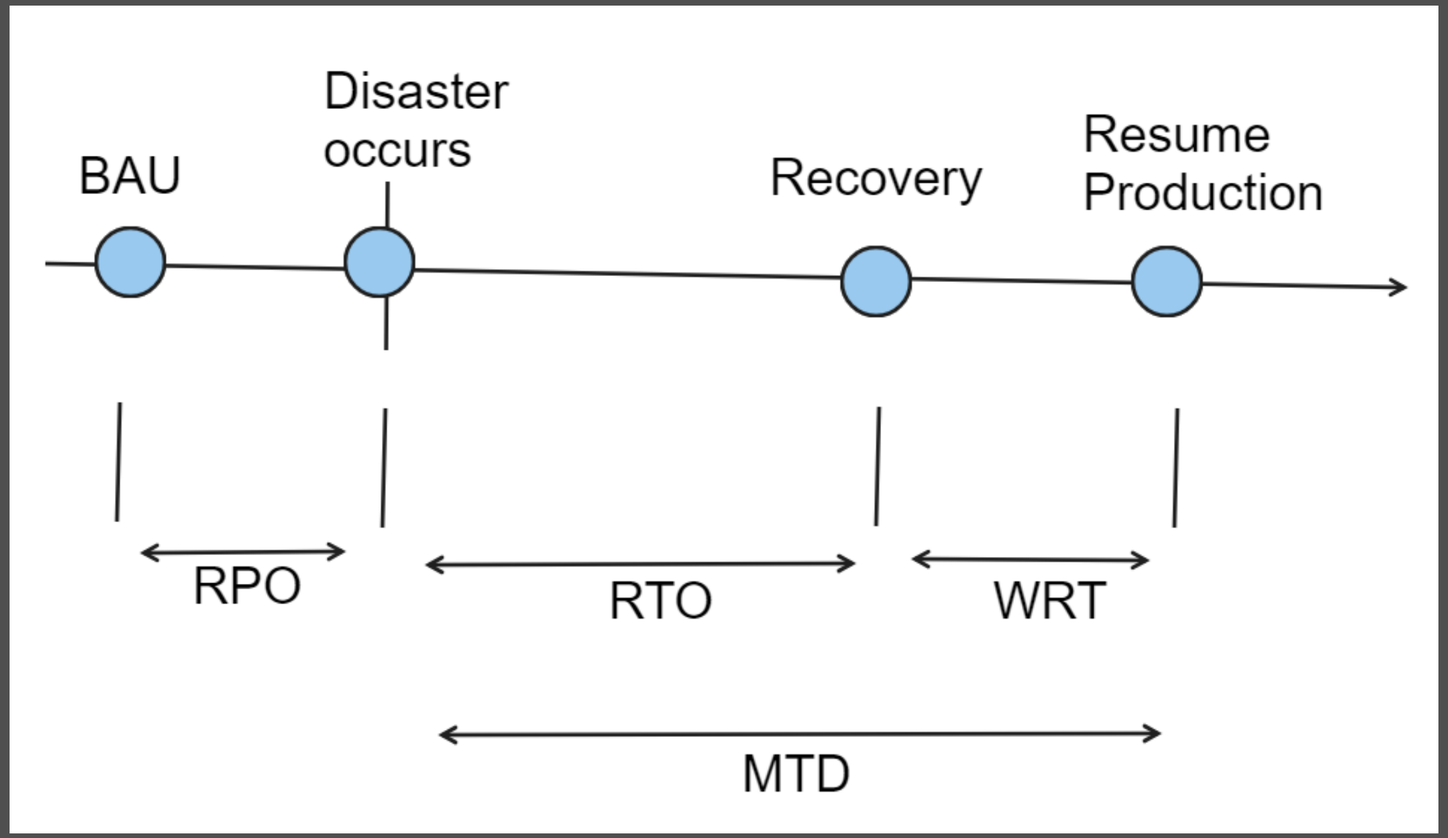
What happened?

Between 20:19 UTC on 7 February 2023 and 04:30 UTC on 9 February 2023, a subset of customers with workloads hosted in the Southeast Asia and East Asia regions experienced difficulties accessing and managing resources deployed in these regions.

One Availability Zone (AZ) in Southeast Asia experienced cooling failure. Infrastructure in that zone was shutdown to protect data and infrastructure. This led to failures in accessing resources and services hosted in that zone. However, this zone failure also resulted in two further unexpected failures. Firstly, regional degradation for some services and secondly, services designed to support failover to other regions or zones did not work reliably.

Terminology

- **RPO: Recovery Point Objective -** Max acceptable data loss measured in time
- **RTO: Recovery Time Objective -** Max time needed to bring systems back online
- **WRT: Work Recovery Time -** The max tolerable amount of time to verify a system and data integrity
- **MTD: Maximum Tolerable Downtime -** the total amount of time a business process can be disrupted without causing an unacceptable consequence


BAU
Disaster occurs
Recovery
Resume Production
RPO
RTO
WRT
MTD

What matters when?

- Minimise RPO - Where integrity of the information is critical
  - i.e. banking transactions/ banking systems
  - Replicas to read only DBs
  - Regular backups
- Minimise RTO - Where availability of the information is critical
  - i.e. Live broadcast, Call center, 111 service
  - May leverage redundant systems
  - High availability services
- They may be equally important
  - i.e. voting system for an election
- RPO can be greater or less than RTO - depending on the situation

Implications of an outage

- Loss of $ - Eftpos service is unavailable
- Loss of productivity - Email down for an org
- Reputational - Public website unavailable
- Health and Safety - Duress system unavailable
- Death...? - 111 call center doesn't work
- Destruction of records/ artefacts/ servers? – HVAC/ environmental systems

Business Continuity

- How do you continue to deliver your business service in case of an outage or service disruption
- What are your contingency plans in case of a business disruption?
- A disaster recovery plan is part of this, but it goes wider
- What else do you have to enact while you are bringing the technical service back up?
- Do you have manual processes that you can enact?
- Can a MVP product be stood up in the meantime?
- In the case of a call center - can people use their mobiles while the IP phones are unavailable?
- For a digitised system - can paper records be used?

Disaster Recovery

- Disaster recovery is the process of maintaining or reestablishing vital infrastructure and systems following a natural or human-induced disaster, such as a storm or battle.
- It employs policies, tools, and procedures. Disaster recovery focuses on information technology (IT) or technology systems supporting critical business functions.
- What do you have to do technically to restore a system/ service
- Can you fail over to an alternative location... ie. a DR location?
- Can you stand up a cold site and restore service?

Disaster Recovery Sites

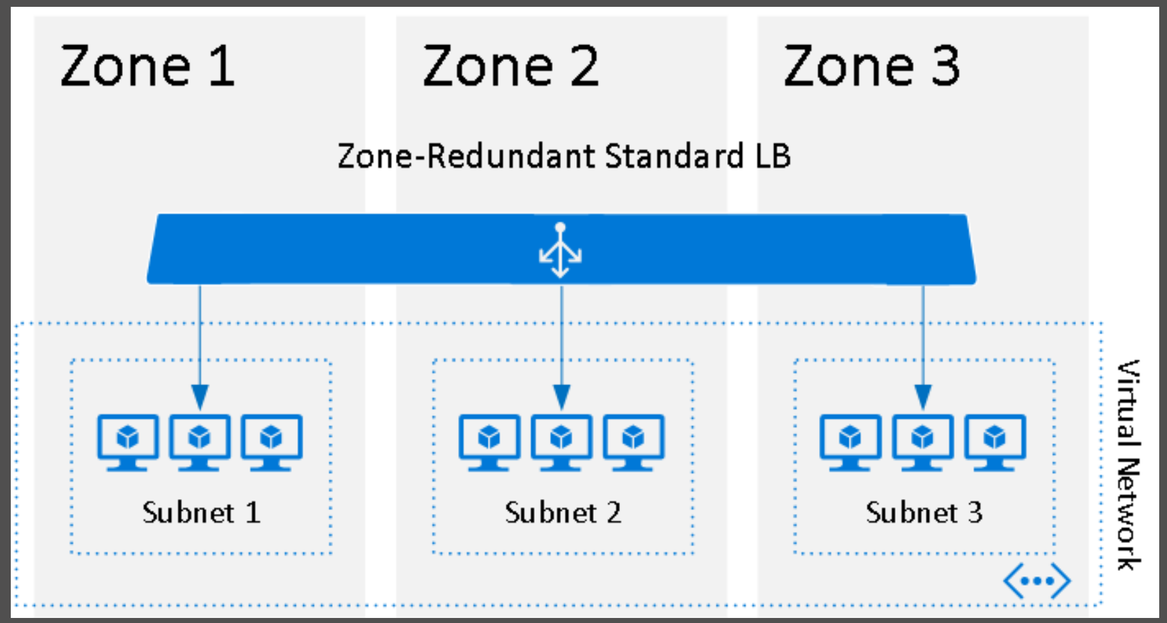
- **Hot Site:** Mirrors the capabilities of the primary location as fully as possible. Great for resiliency. Bad for budget.
- **Warm Site:** Includes key equipment and functionality, but takes longer to set up and may not have the complete data and functionally offered by a hot site.
- **Cold Site:** A 'barebones' implementation. May require migration of capability. Will require increased time to restore to. Beneficial for cost, but has a slower restoration time.

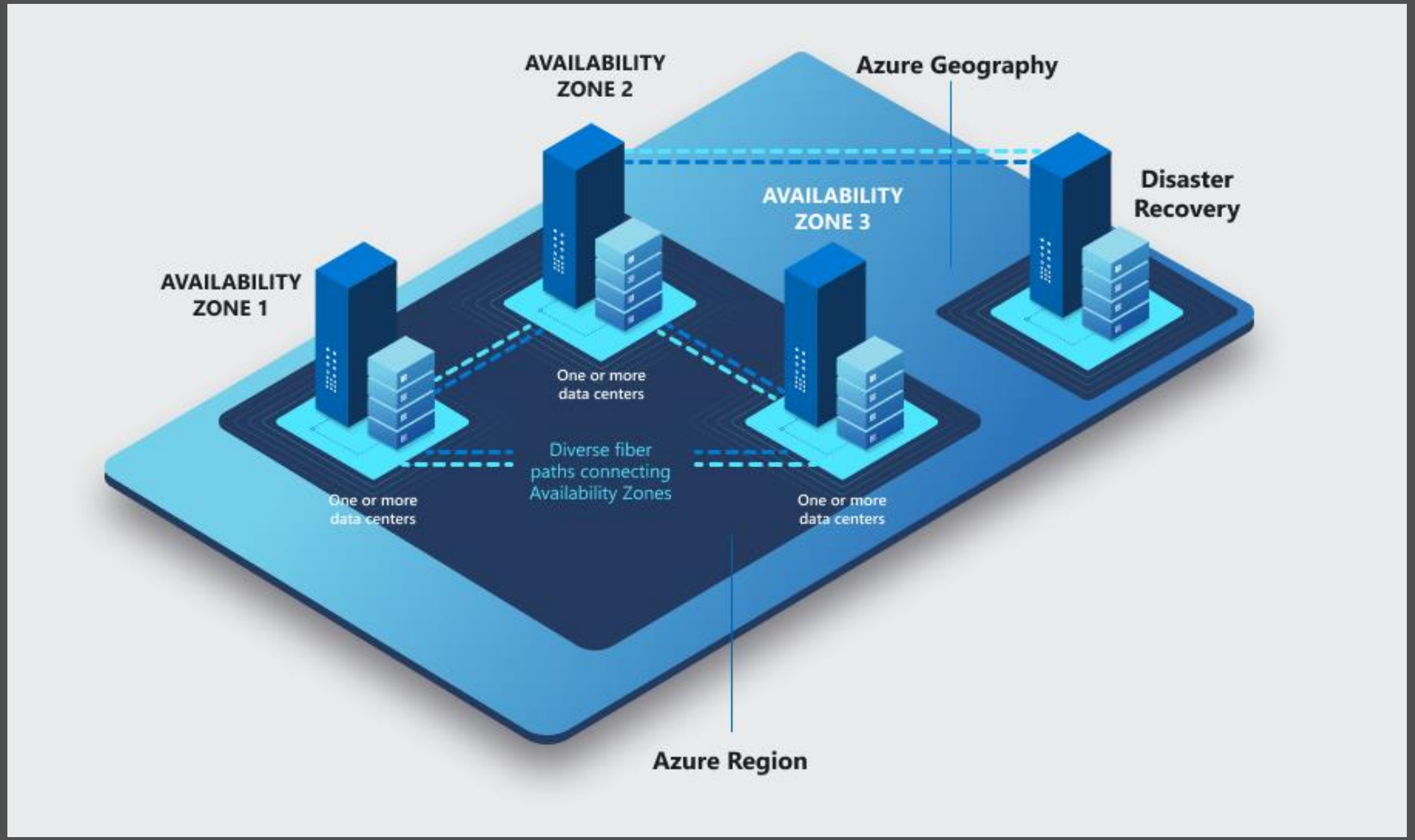
How does DR support BCP?

- A disaster recovery plan prompts the quick restart of systems so that operations can continue as scheduled
- Disaster recovery is a subset of business continuity
- Your BCP may heavily reference business process and alternative measures to keep operations occurring while technical recovery is occurring

High Availability

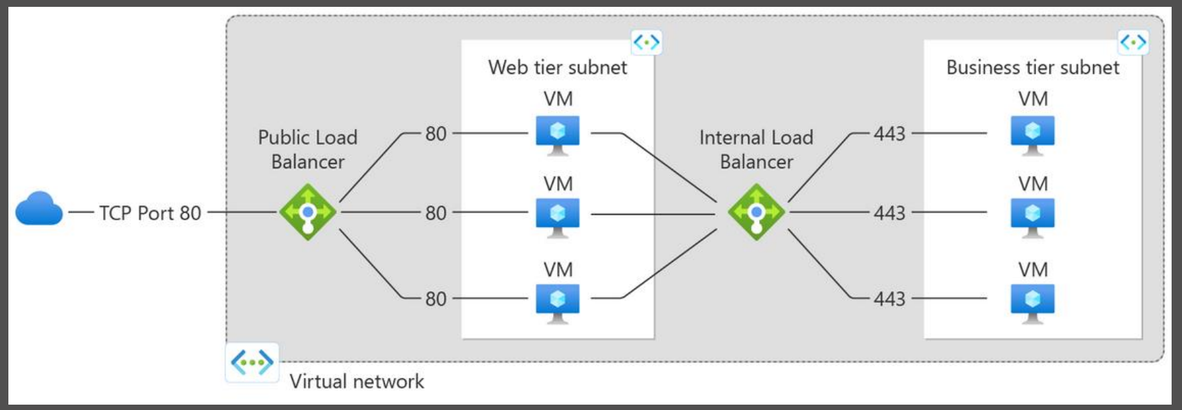
- Might be part of a DR strategy
- Be considerate of if a failure replicates from one node to another - this may affect your ability to recover
- Backups are still required!




AVAILABILITY ZONE 2
Azure Geography
AVAILABILITY ZONE 3
Disaster Recovery
AVAILABILITY ZONE 1
One or more data centers
Diverse fiber paths connecting Availability Zones
One or more data centers
One or more data centers
Azure Region

Load Balancing

- A load balancer can balance traffic over a number of nodes to reduce the impact if one fails
- But what happens if the load balancer fails - is there resilience at that layer?
- Consider what the best method is for your load balancing - round robin, least connections, least response time, least bandwidth...



Backups

- In the Essential 8, Cert top 10... every controls list you'll ever see
- Back up regularly
- Have different types of backups
- Have offline backups
- Regularly test back ups
- Protect your backed up data

- Have three copies of the data, keep them on two different formats, store one offsite, and store one offline.

- Do you know how long it takes to restore from backup?
- How long would it take to retrieve an offline backup?
- Do you have logging enabled if your backups fail?
- Is the data stored by your service as important as the configuration? Or vice versa? What needs to be backed up?

Monitoring

- Consider where is appropriate to monitor
  - Inside the environment
    - Monitor components
    - Failures
    - Connectivity
    - Error codes
  - Outside the environment (i.e. connectivity to)
    - Ping tests
    - Can all regions/ instances be reached
    - What's being hit on the load balancers?
  - Do you know what 'normal' is?

Testing

- Resilience without testing is a bad time
- Misconfigurations tend to slip in, especially after changes.
- Is all your routing in place, and configured correctly?
- Did you get all the DNS entries entered correctly?
- Are your TTL's configured appropriately?
- Testing your resilience, gives confidence that if something does go wrong, you can actually come back from the issue

- Table top testing - can give you confidence your human processes will work - do people know what they should do? Tease out if your RPO and RTO are appropriate?
- Full exercise - Actually exercising your plan - testing your resilience. What happens if you loose the ASE region?

Chaos Monkey is responsible for randomly terminating instances in production to ensure that engineers implement their services to be resilient to instance failures.

See how to deploy for instructions on how to get up and running with Chaos Monkey.

Once you're up and running, see configuring behavior via Spinnaker for how users can customize the behavior of Chaos Monkey for their apps.

https://netflix.github.io/chaosmonkey/

Summary

- There are lots of ways to implement resiliancy in your service
- DR != BCP
- Consider your business requirements
- Understand your operating environment
- Determine your RPO, RTO, MTD to determine what the best options are for you
- Weigh up the cost vs. your requirements
- Implement the correct level of resiliancy for your solution
- Implement monitoring so you know if your system is available or not
- Test, test and test

Thanks for your time

peter@privsec.nz