Cloud Applications Architecture

Course 6 - High Availability

Highly Available (HA) Systems

What Makes a System Highly Available?

- Hardware
- Software
- Data
- Network

Why is Availability Important?

Business continuity

Loss of revenue, customers, lives

SLAs (certain SLAs might allow “uncounted” downtime if the service recovers within a certain time)

High Availability vs Continuous Availability

Availability Formula

$$\text{Availability} = \frac{\text{Uptime}}{\text{Uptime + Downtime}}$$

https://uptime.is/

Availability %	Downtime per year[note 1]	Downtime per month	Downtime per week	Downtime per day
90% ("one nine")	**36.53 days**	**73.05 hours**	**16.80 hours**	**2.40 hours**
95% ("one and a half nines")	18.26 days	36.53 hours	8.40 hours	1.20 hours
97%	10.96 days	21.92 hours	5.04 hours	43.20 minutes
98%	7.31 days	14.61 hours	3.36 hours	28.80 minutes
99% ("two nines")	**3.65 days**	**7.31 hours**	**1.68 hours**	**14.40 minutes**
99.5% ("two and a half nines")	1.83 days	3.65 hours	50.40 minutes	7.20 minutes
99.8%	17.53 hours	87.66 minutes	20.16 minutes	2.88 minutes
99.9% ("three nines")	**8.77 hours**	**43.83 minutes**	**10.08 minutes**	**1.44 minutes**
99.95% ("three and a half nines")	4.38 hours	21.92 minutes	5.04 minutes	43.20 seconds
99.99% ("four nines")	**52.60 minutes**	**4.38 minutes**	**1.01 minutes**	**8.64 seconds**
99.995% ("four and a half nines")	26.30 minutes	2.19 minutes	30.24 seconds	4.32 seconds
99.999% ("five nines")	**5.26 minutes**	**26.30 seconds**	**6.05 seconds**	**864.00 milliseconds**
99.9999% ("six nines")	**31.56 seconds**	**2.63 seconds**	**604.80 milliseconds**	**86.40 milliseconds**
99.99999% ("seven nines")	**3.16 seconds**	**262.98 milliseconds**	**60.48 milliseconds**	**8.64 milliseconds**
99.999999% ("eight nines")	**315.58 milliseconds**	**26.30 milliseconds**	**6.05 milliseconds**	**864.00 microseconds**
99.9999999% ("nine nines")	**31.56 milliseconds**	**2.63 milliseconds**	**604.80 microseconds**	**86.40 microseconds**

Data from Wikipedia

Common Availability Tiers

95%

99%

99.9%

99.95%

99.99%

99.999%

Availability Concerns - **Nature**

Availability Concerns - Technical

Unplanned

Usually due to human error

Planned (maintenance)

You can define the maintenance window for certain services

Usually third-party APIs notify you

Techniques to Achieve **HA**

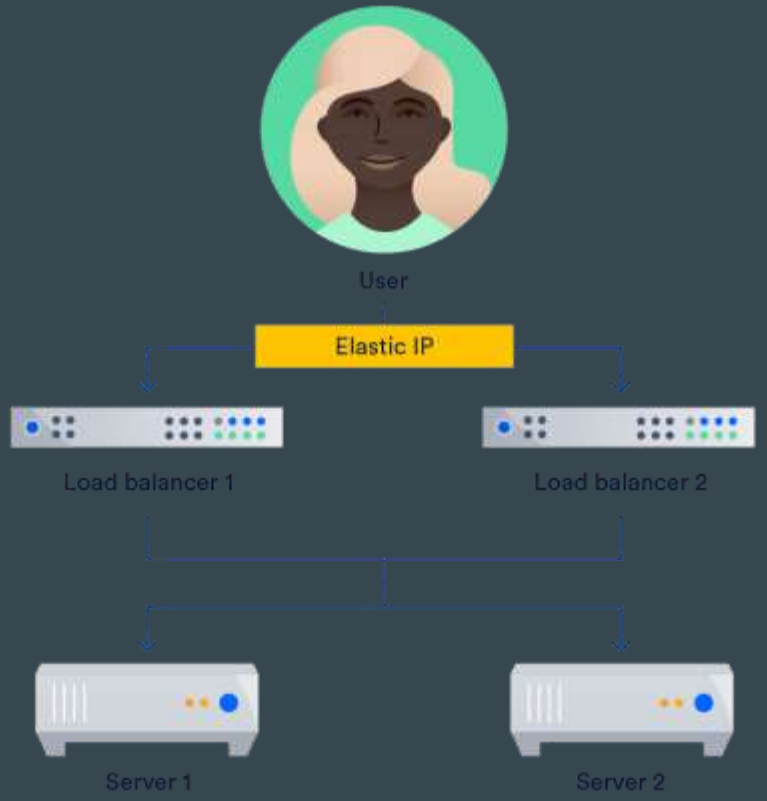
Techniques

Infrastructure level

Application level

Floating/Elastic IP

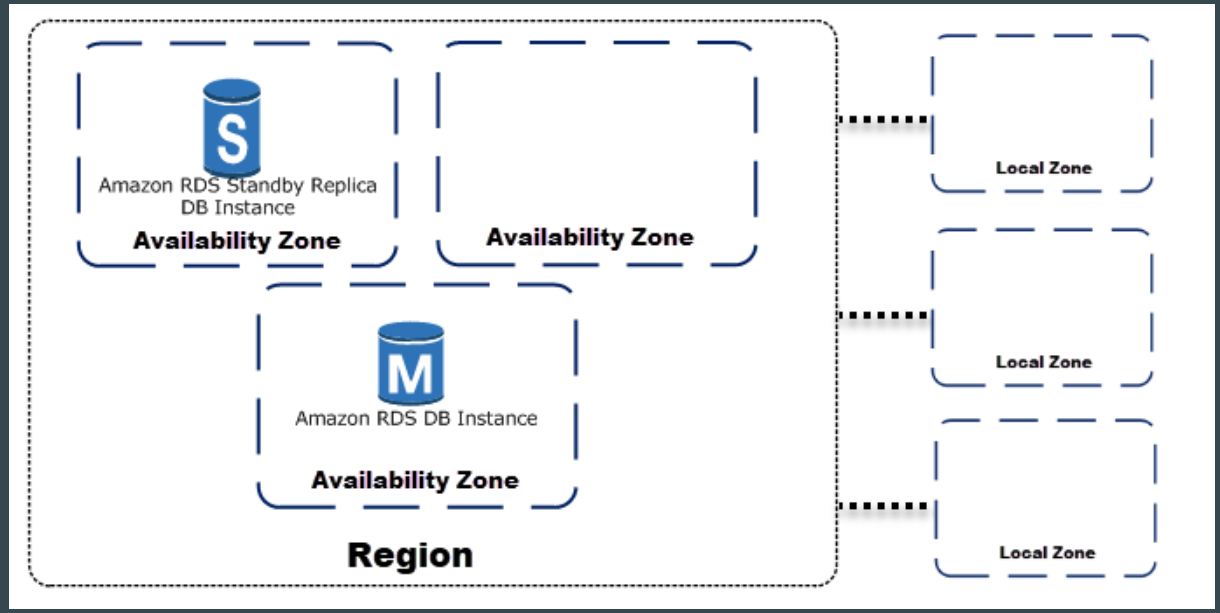
Eliminate Single Point of Failure



Multi-AZ

Many services support it natively

Can be hard to maintain otherwise



Multi-AZ

How to achieve it?

- Cluster-aware routing (load balancers, DNS)
  - Health checks/probes - natively supported by many services
- Data replication
  - Sync/Async
  - Available solutions
- Packet mirroring (Traffic duplication)

Multi-Region

Regions are usually entirely different clouds.

Netflix users didn’t even notice an entire AWS region went offline

Minimizing Impact Radius

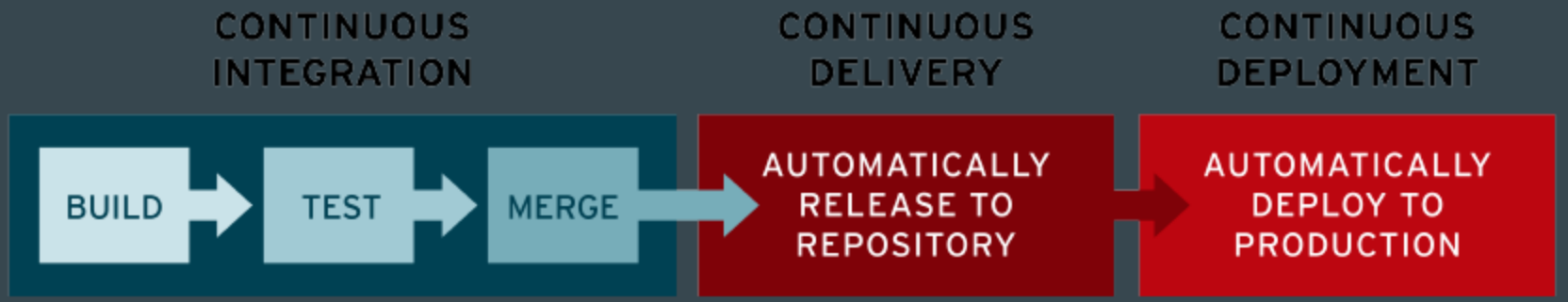
(Decoupled) Microservices

Proper Processes

I.e. try to avoid human errors

Code reviews (4-eye principle)

CI/CD



Proper Monitoring

React quickly

What to monitor:

- Database
- Website
- Virtual network
- Storage
- VM

“Embrace the Chaos”

Netflix Chaos Monkey

https://principlesofchaos.org

Resiliency

Resiliency

Capability of a system to remain functional/useful even if parts of it become unavailable.

No matter how perfect a system is, failure is certain.

Resilient system is:

- Adaptive
- Self healing
- Predictable

Requires

- Investment
- Automation
- Monitoring
- Simplicity

Disaster Recovery (DR)

Recovery Time Objective (**RTO**)

How long it takes to bring the system back.

Highly dependant on the **DR Strategy**.

Lower RTO usually means (considerably) increased cost.

Recovery Point Objective (**RPO**)

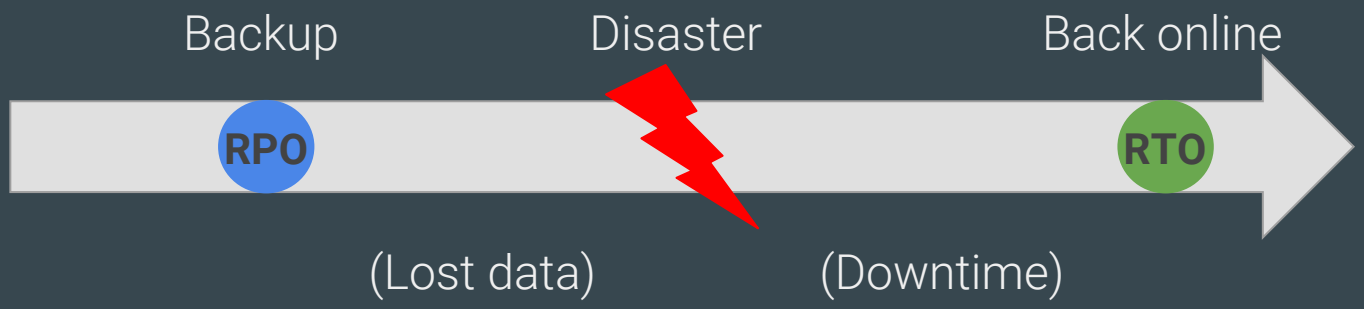
How much data was lost.

I.e. How much time has passed since the last backup.

Usually easier to improve:

- More frequent backups
- Leverage incremental backups to reduce costs

RTO & RPO



DR Strategies

Backup and Restore

- In case if disaster, restart/recreate everything based on the latest backup

Pilot Light

- Have the critical components prepared
- E.g. have a database replica ready for DR (but no compute)

Warm Standby

- Full system replica ready, reduced size (e.g. smaller VMs)

Hot Site

- Exact replica ready

Worse RTO, cheaper

Better/lower RTO, (much) more expensive

Summary