Load Balancers

Memi Lavi
www.memilavi.com

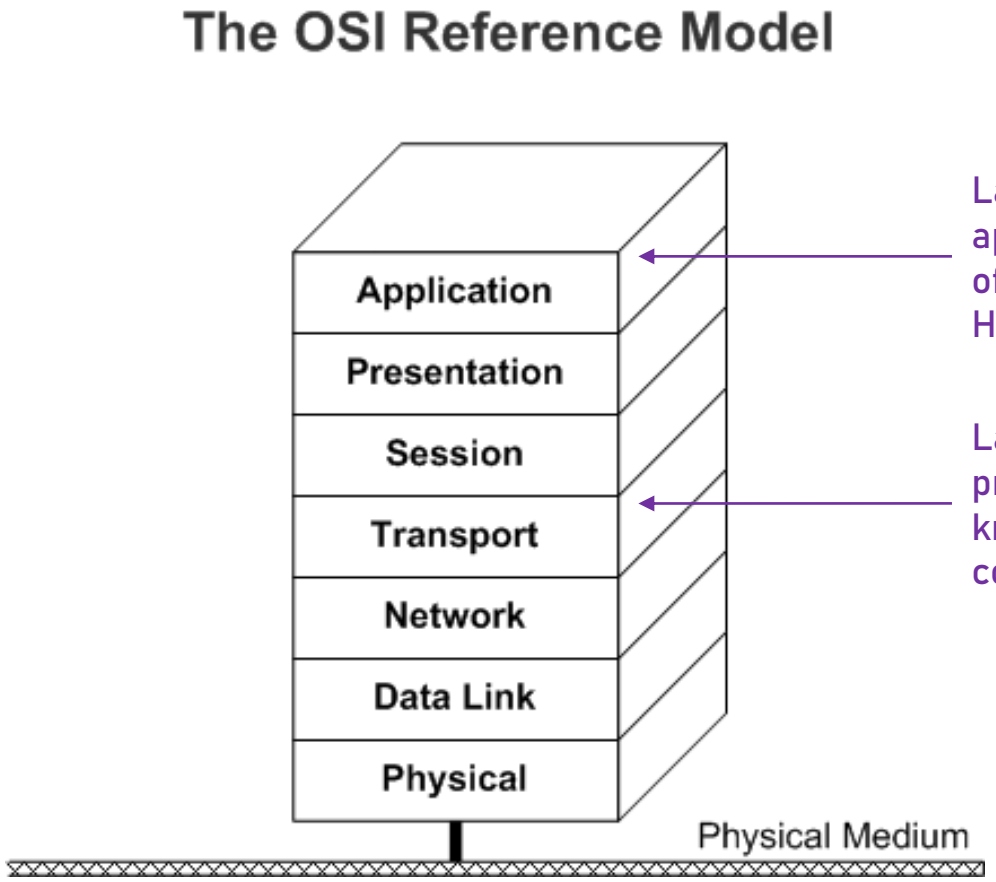
Load Balancers

- GCP offers various types of load balancers
- Target different scenarios
- We'll learn about the different types and when to use each
- We'll integrate some of them in our ReadIt app

Load Balancer Role

- Cloud service that distributes load and checks health of instances
- When an instance is not healthy – no traffic is directed to it
- Can work with various types of backends
- Can be external or internal
- Operates at layer 4 or 7 of the OSI model

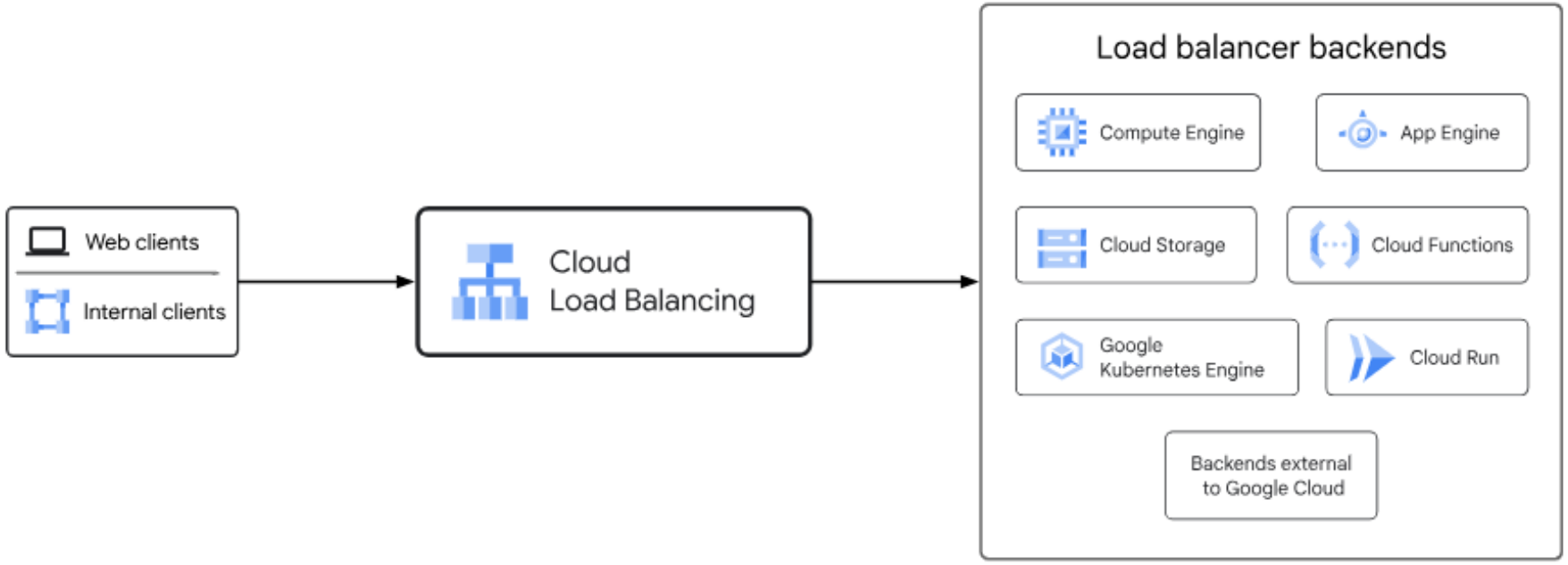
7 Layers Model



Layer 7: Interacts with the application. Can see the content of the transmission. Example: HTTP URL + path + params

Layer 4: Knows about IP, Port, protocol, TLS and more. Has no knowledge about the actual content

Cloud Load Balancer



Load Balancer Distribution Algorithm

- Depends on the load balancer type
- Can be based on 5 tuple hash:
  - Source IP
  - Source port
  - Destination IP
  - Destination port
  - Protocol type
- Same tuples used by Firewall Rules
- Can be based on backend utilization, weights and more

Load Balancer Rule of Thumb

In general, always use load balancer as the front of your app

Never expose VM / AE / GKE etc. directly to the internet

Load Balancer Types

- GCP offers various types of load balancers
- It's important to choose the right one for your scenario
- Differences in various factors:
  - Traffic type
  - External or internal
  - Deployment mode

Traffic Type

- What kind of traffic is going to go through the load balancer

HTTP/S	TCP	TCP, UDP, ICMP…
Application load balancer	Network load balancer	Passthrough load balancer
• Use as front to a web app • Looks at layer 7 • Can route based on HTTP path • Proxy-based	• Use as front to a TCP listener • Looks at layer 4 • Proxy-based	• Distributes traffic in multiple network protocols • Looks at layer 4 • Traffic is terminated at the backend, not at the load balancer

External or Internal

- Where does traffic comes from

External

- Traffic comes from the internet
- Example: public web site

Internal

- Traffic comes from the VPC the load balancer is in
- Example: backend services calling each other

Deployment Mode

- Where is the load balancer distributed

Global

- Load balancer is deployed in all regions
- Great for DR scenarios *
- Only in External load balancers

Cross-region

- Load balancer is deployed in multiple regions
- Resilient to regional outage
- Only in internal load balancers

Regional

- Load balancer is deployed in a single region
- Distributed across multiple zones
- Serves apps in the specific region

* We'll delve into DR later in this course

Load Balancer Types

Type	Deployment Mode	Comments
Application Load Balancer (HTTP/S)	Global external	
	Regional external	
	Regional internal	
	Cross-region internal	
Proxy Network Load Balancer (TCP)	Global external	Optional SSL offload
	Regional external	
	Internal	Always regional
Passthrough Network Load Balancer (TCP, UDP, ICMP…)	External	Always regional
	Internal	Always regional

Load Balancer Types

Type	Deployment Mode	Use For…
Application Load Balancer (HTTP/S)	Global external	Public global websites distributed in multiple regions
	Regional external	Public global websites distributed in a single region
	Regional internal	Intra-backend communication distributed in a single region
	Cross-region internal	Intra-backend communication distributed in multiple regions
Proxy Network Load Balancer (TCP)	Global external	Public TCP listener distributed in multiple regions
	Regional external	Public TCP listener distributed in a single region
	Internal	Internal TCP listener
Passthrough Network Load Balancer (TCP, UDP, ICMP…)	External	Public network listener
	Internal	Internal network listener

Load Balancer Pricing

- All load balancers EXCEPT internal application load balancer:

First 5 forwarding rules	$0.025 / hour
Additional forwarding rule	$0.01 / hour
Inbound data processed	$0.008 / GiB / month
Outbound data processed	$0.008 / GiB / month

Load Balancer Pricing

- Internal application load balancer (regional and cross-region):

Per instance	$0.025 / hour
Inbound data processed	$0.008 / GiB / month
Outbound data processed (only in cross-region LB)	$0.008 / GiB / month

Load Balancer and App Engine

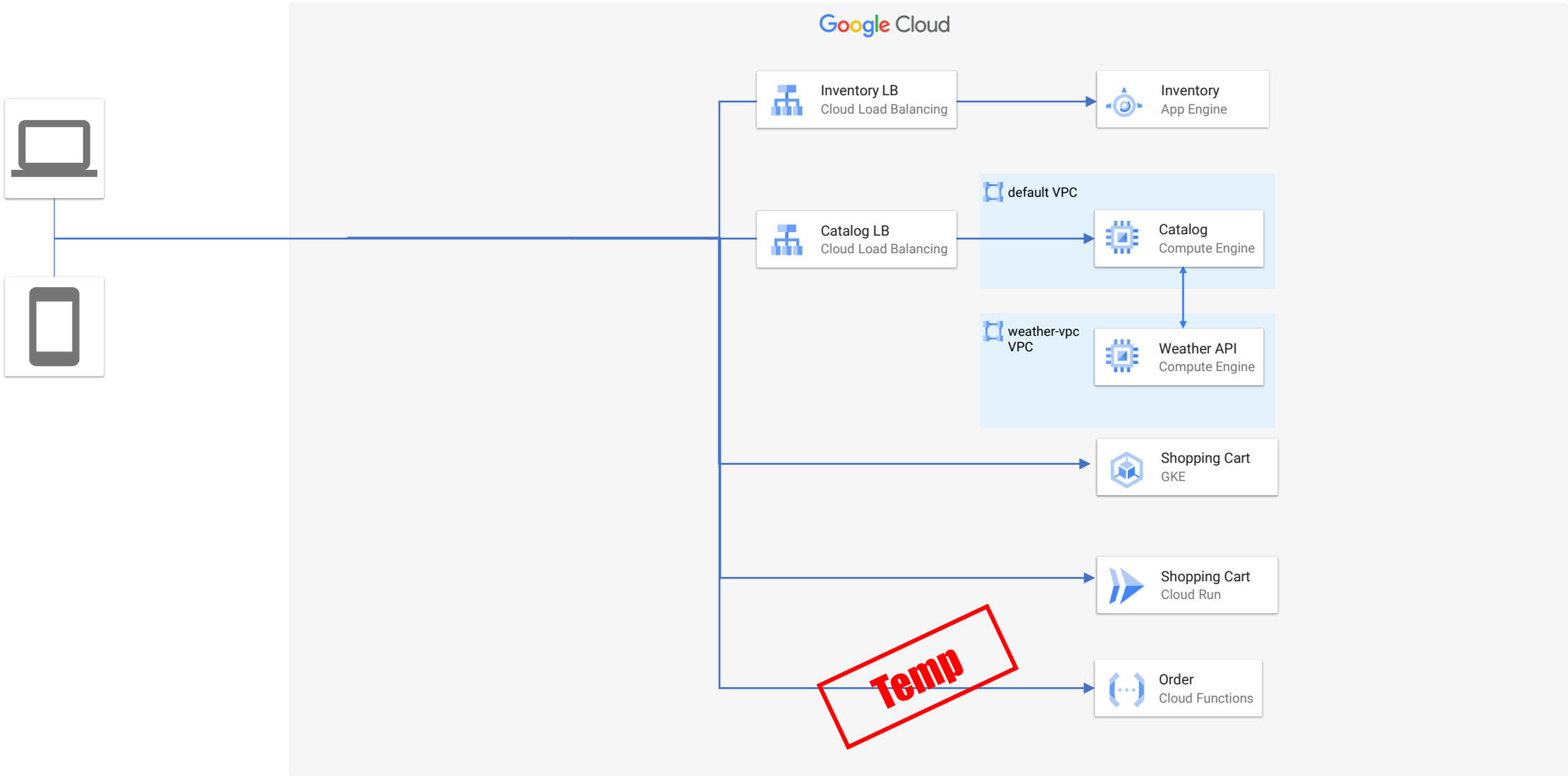
- App Engine has its own load balancer
- Provides basic load balancing capabilities with the autoscaling of App Engine
- Sometimes it's a good idea to still have Load Balancer on top of App Engine

Load Balancer and App Engine

- Reasons for that:
  - Advanced capabilities not found in the built-in load balancer
    - ie. CDN integration
  - Balancing between two or more App Engines in different projects

ReadIt!

Architecture: ReadIt Cloud System



Affinity

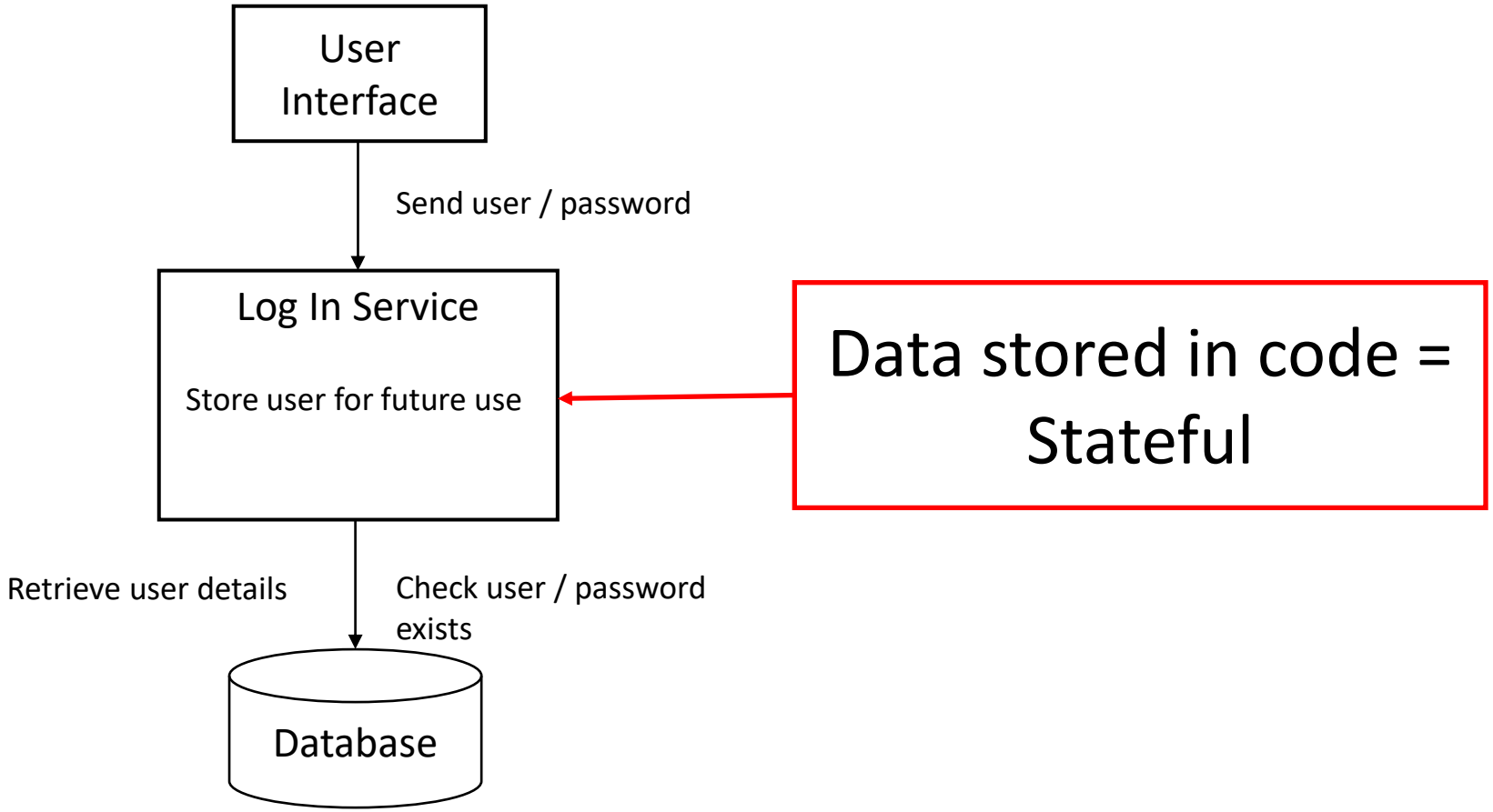
- Makes sure user will always be directed to the same instance (VM / App Engine) it began with
- Should be avoided when possible
- Usually required in Stateful apps
- Usually a sign of bad design
- Always try to design Stateless app

Stateless

The application's state is stored in only two places – the data store and the user interface

State = Application's Data

Stateless Example



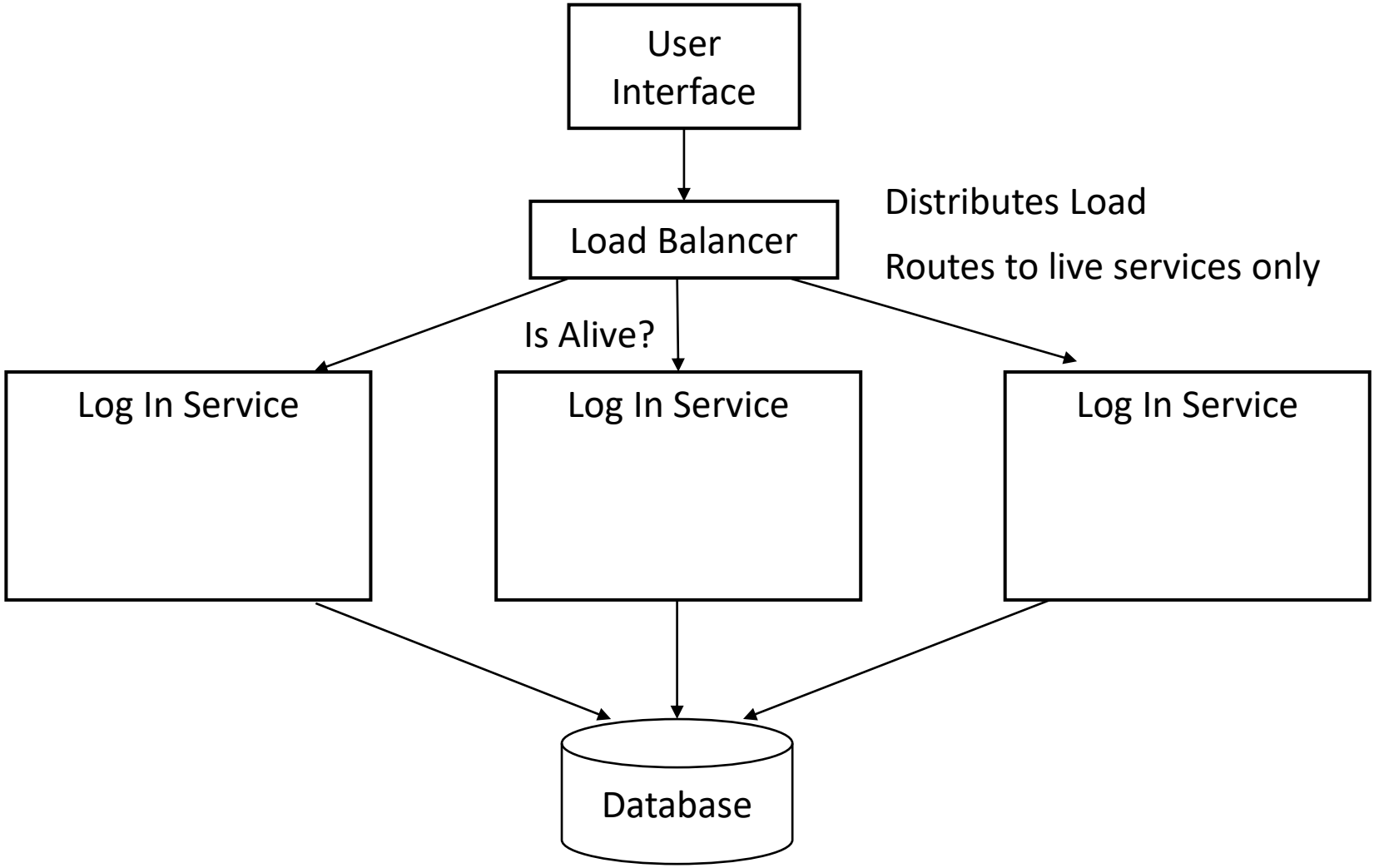
Scalability – A Reminder

- Grow and shrink as needed
- Scale Up vs Scale Out
- Scale Out is usually preferred

Redundancy - A Reminder

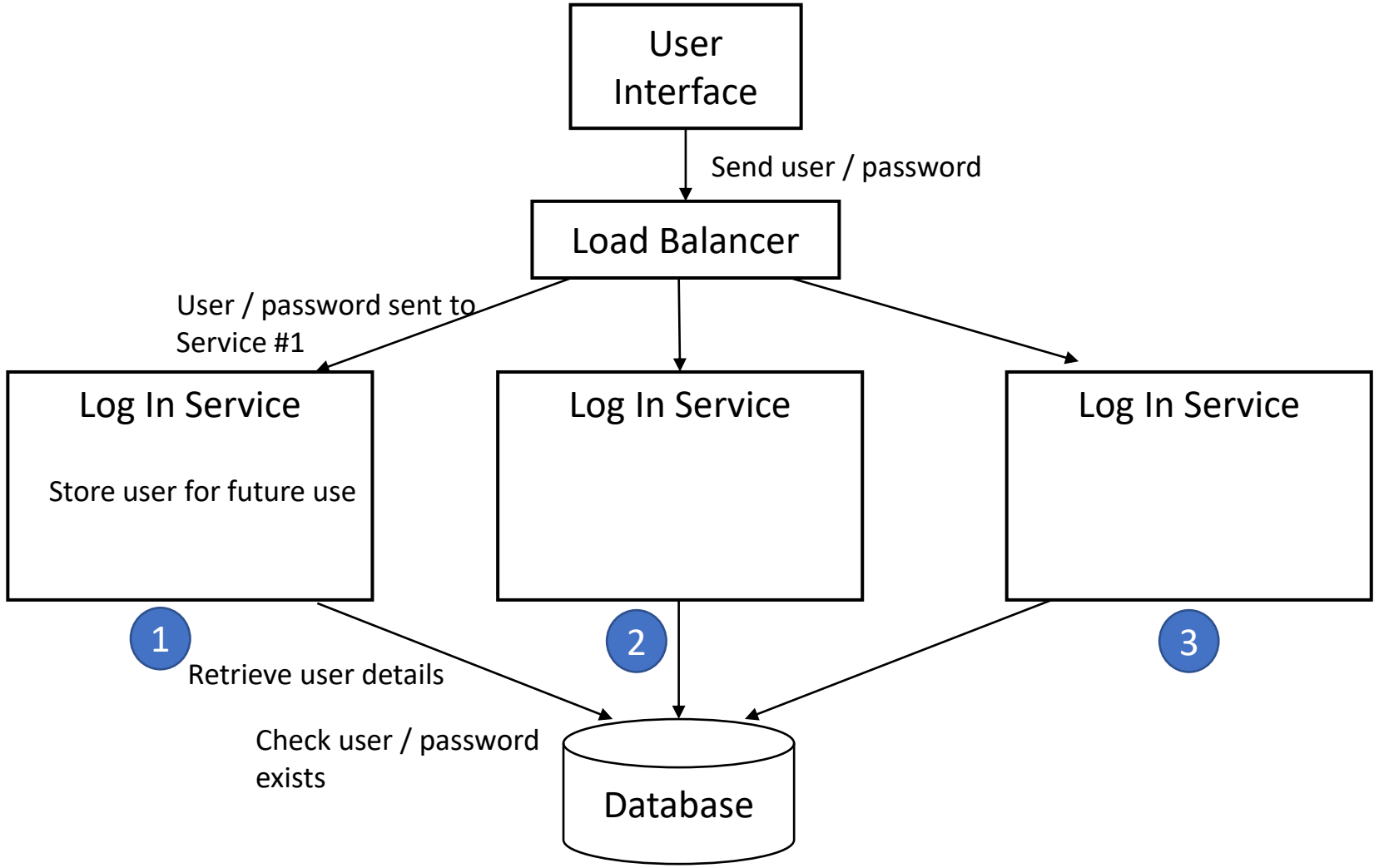
- Allows the system to function properly when resource is not working
- Example:
  - A system with more than one server
  - When a server goes down, the other continue working

Scalable & Redundant Architecture

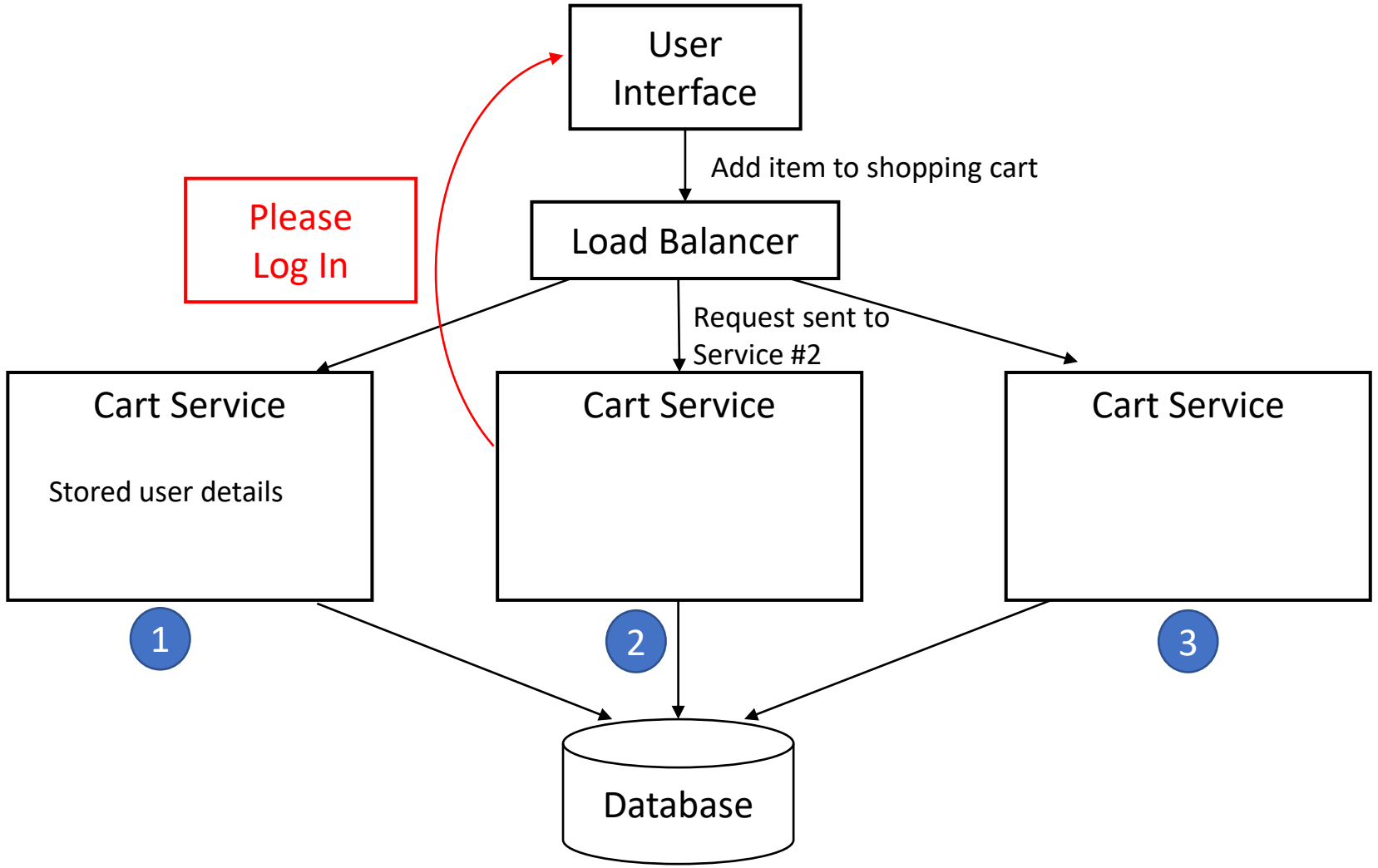


- Scalability
- Redundancy

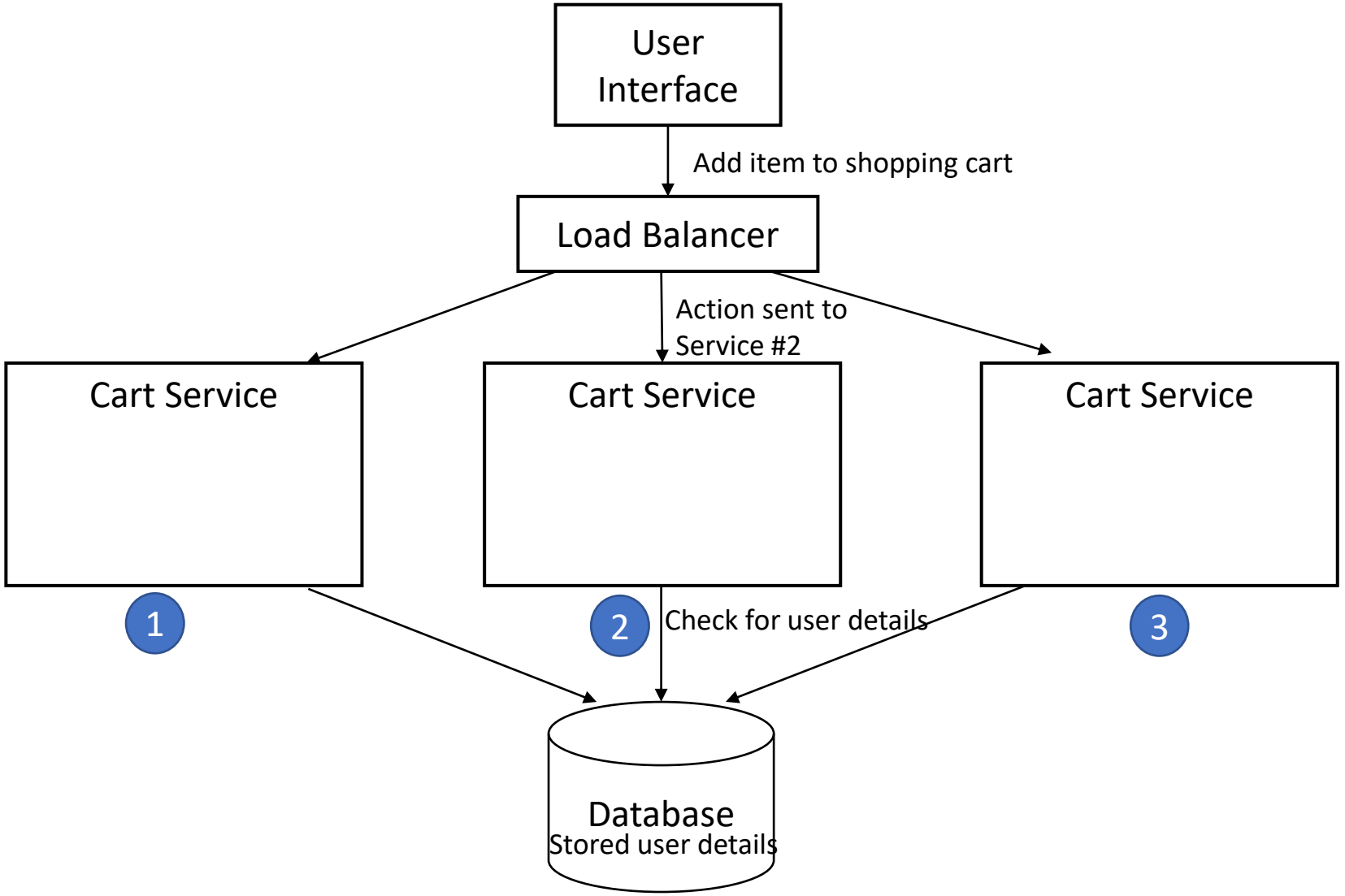
Stateful Example



Stateful Example



Stateless Example



Stateless

- Always use stateless architecture
- Supports Scalability and Redundancy

Secure Network Design