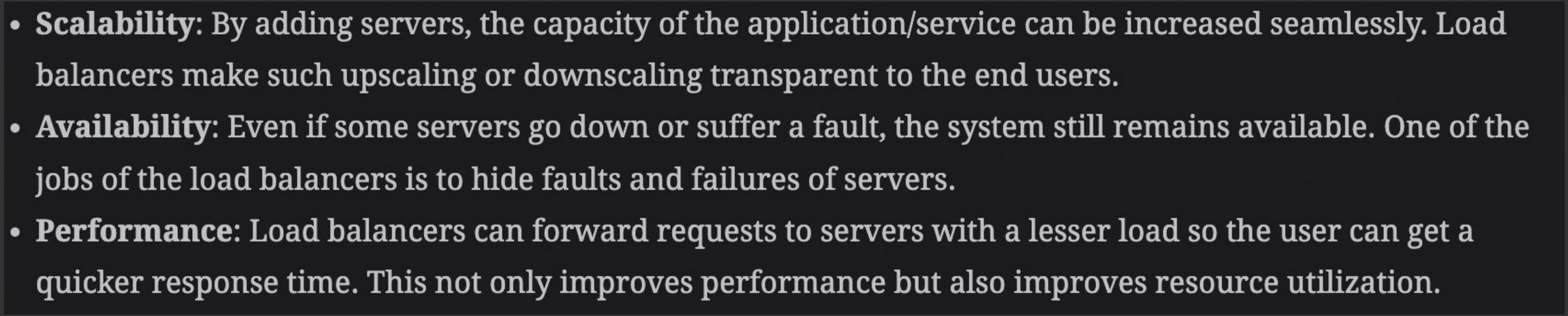
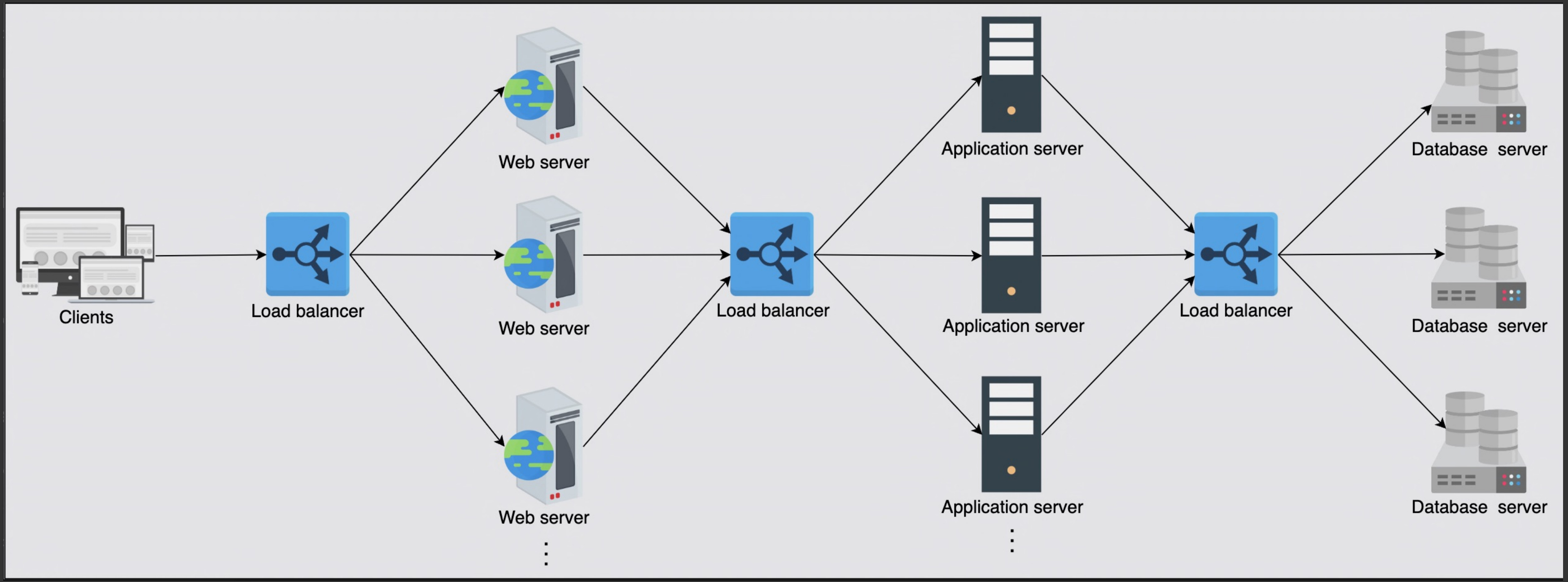
Introduction to Load Balancers

A load Balancers (LB) : is the answer to the question. The job of the LB is to fairly divide all Clints request among the pool of available Servers. Load balancers perform this job to avoid overloading or crashing servers.



- **Scalability**: By adding servers, the capacity of the application/service can be increased seamlessly. Load balancers make such upscaling or downscaling transparent to the end users.
- **Availability**: Even if some servers go down or suffer a fault, the system still remains available. One of the jobs of the load balancers is to hide faults and failures of servers.
- **Performance:** Load balancers can forward requests to servers with a lesser load so the user can get a quicker response time. This not only improves performance but also improves resource utilization.

Placing Load Balancers



services offered by Load Balancers ?

→ Health Checking

→ TLS Termination

→ Predictive analytics

→ Reduced human intervention

→ service discovery

→ security.

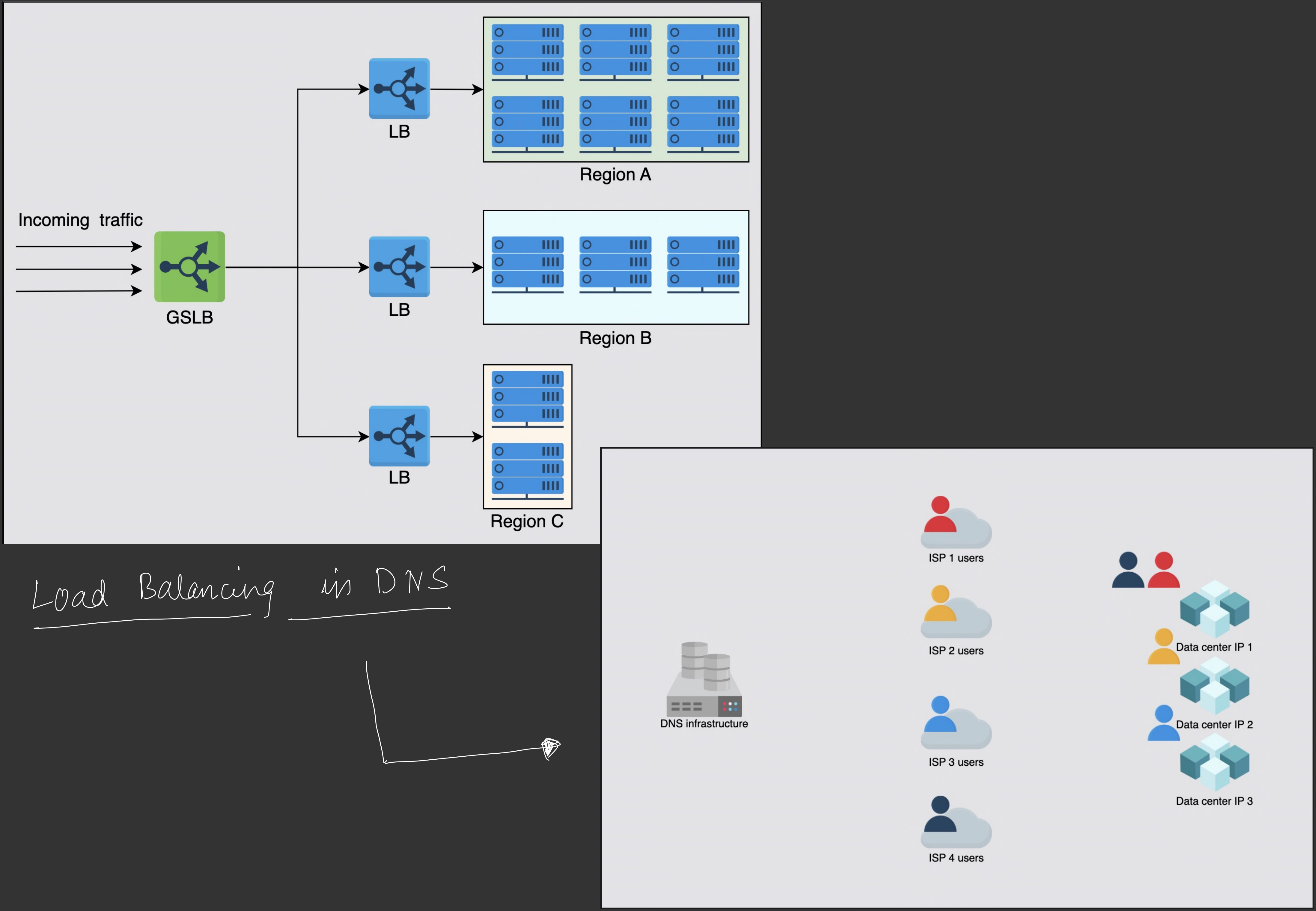
What if Load Balancers fail? Are they not a single point of failure (SPOF)?

LB are usually deployed in pairs as a means of disaster recovery. If one LB fail, and there's nothing to failover to, the overall service will go down. Generally, to maintain high availability, enterprise use clusters of load balancers that use heartbeat communication to check the health of load balancers at all times. On failure of Primary LB, the backup can take over. But if the entire cluster fails, manual rerouting can also be performed in case of emergencies.

Global & Local Load Balancing

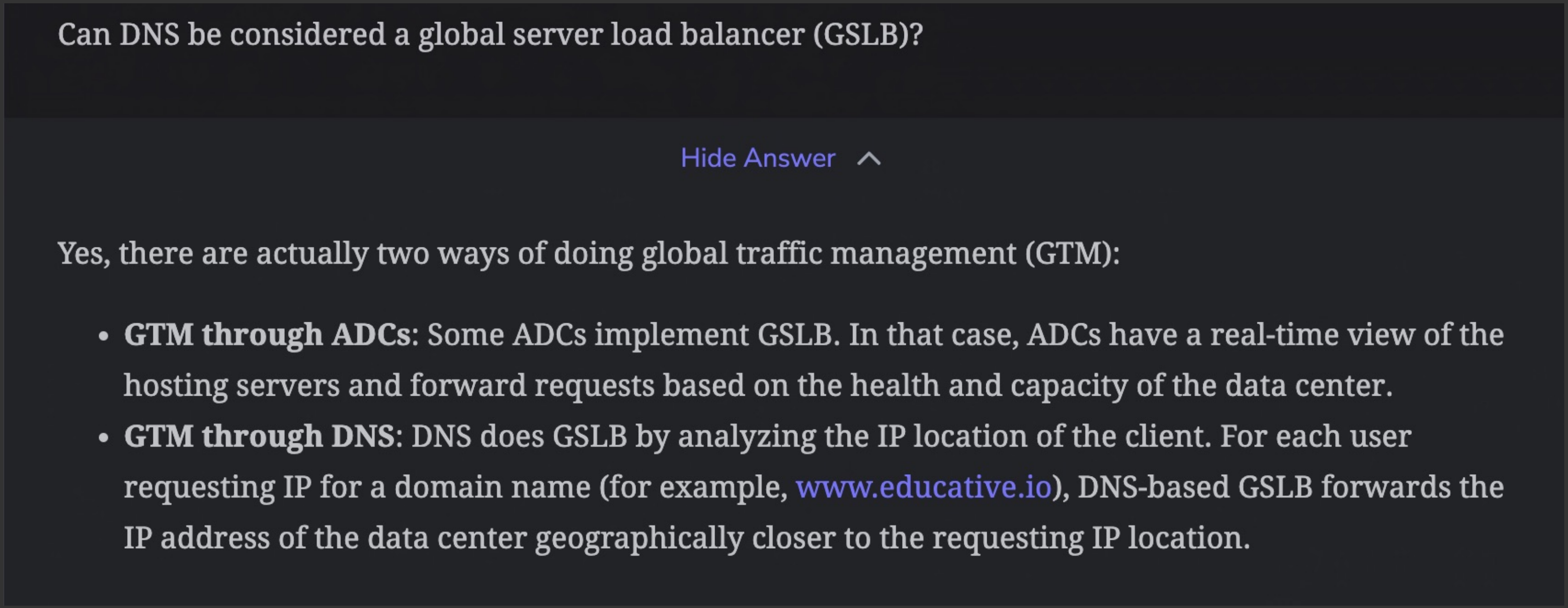
→ Global Server Load balancing (GSLB) : GSLB involves the distribution to traffic load across multiple geographical regions

→ Local Load balancing : This refers to load balancing achieved within a data center. This type of LB focuses on improving efficiency & better resource utilization of the hosting servers in a data center.



Load Balancing in DNS

Round Robin in DNS → data centers in a strict circular order.

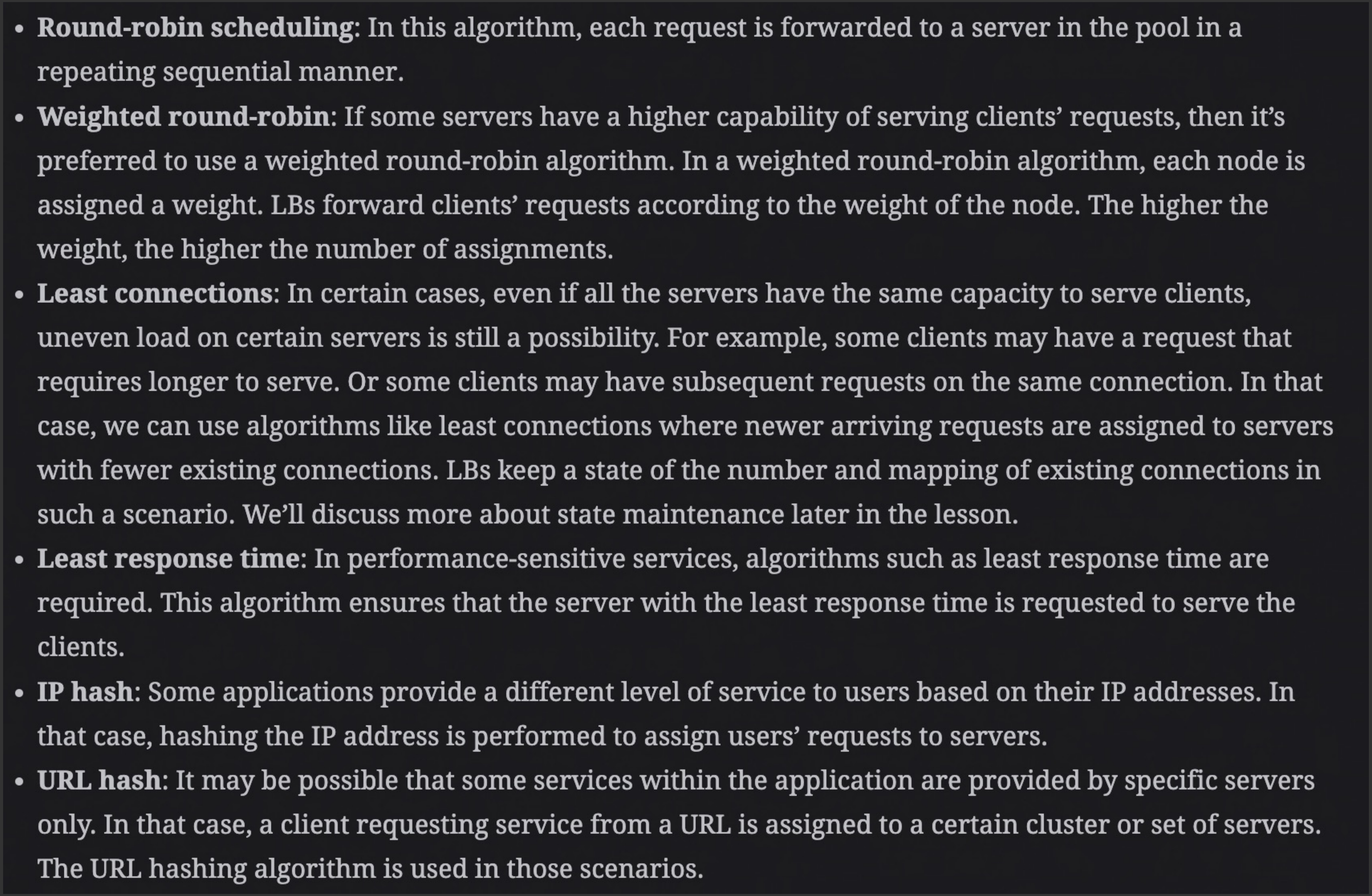


Can DNS be considered a global server load balancer (GSLB)?

Yes, there are actually two ways of doing global traffic management (GTM):

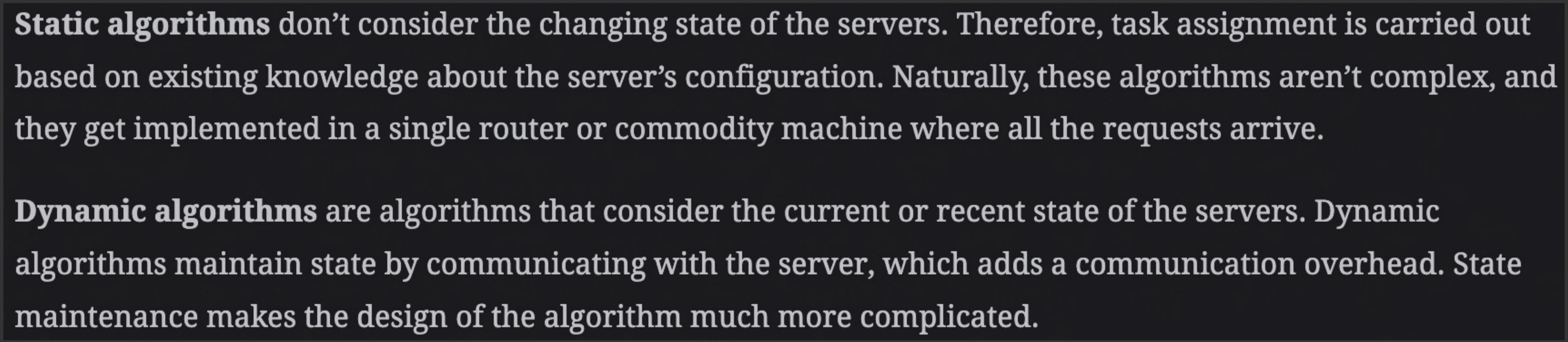
- **GTM through ADCs**: Some ADCs implement GSLB. In that case, ADCs have a real-time view of the hosting servers and forward requests based on the health and capacity of the data center.
- **GTM through DNS**: DNS does GSLB by analyzing the IP location of the client. For each user requesting IP for a domain name (for example, www.educative.io), DNS-based GSLB forwards the IP address of the data center geographically closer to the requesting IP location.

Algorithms of Load Balancers



- **Round-robin scheduling**: In this algorithm, each request is forwarded to a server in the pool in a repeating sequential manner.
- **Weighted round-robin**: If some servers have a higher capability of serving clients' requests, then it's preferred to use a weighted round-robin algorithm. In a weighted round-robin algorithm, each node is assigned a weight. LBs forward clients' requests according to the weight of the node. The higher the weight, the higher the number of assignments.
- **Least connections**: In certain cases, even if all the servers have the same capacity to serve clients, uneven load on certain servers is still a possibility. For example, some clients may have a request that requires longer to serve. Or some clients may have subsequent requests on the same connection. In that case, we can use algorithms like least connections where newer arriving requests are assigned to servers with fewer existing connections. LBs keep a state of the number and mapping of existing connections in such a scenario. We'll discuss more about state maintenance later in the lesson.
- **Least response time**: In performance-sensitive services, algorithms such as least response time are required. This algorithm ensures that the server with the least response time is requested to serve the clients.
- **IP hash**: Some applications provide a different level of service to users based on their IP addresses. In that case, hashing the IP address is performed to assign users' requests to servers.
- **URL hash**: It may be possible that some services within the application are provided by specific servers only. In that case, a client requesting service from a URL is assigned to a certain cluster or set of servers. The URL hashing algorithm is used in those scenarios.

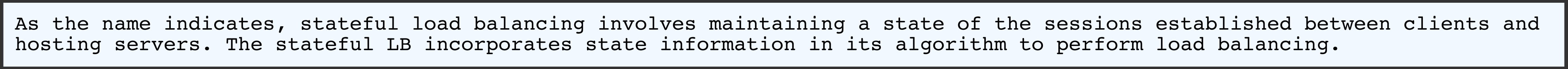
Static Vs Dynamic algorithms



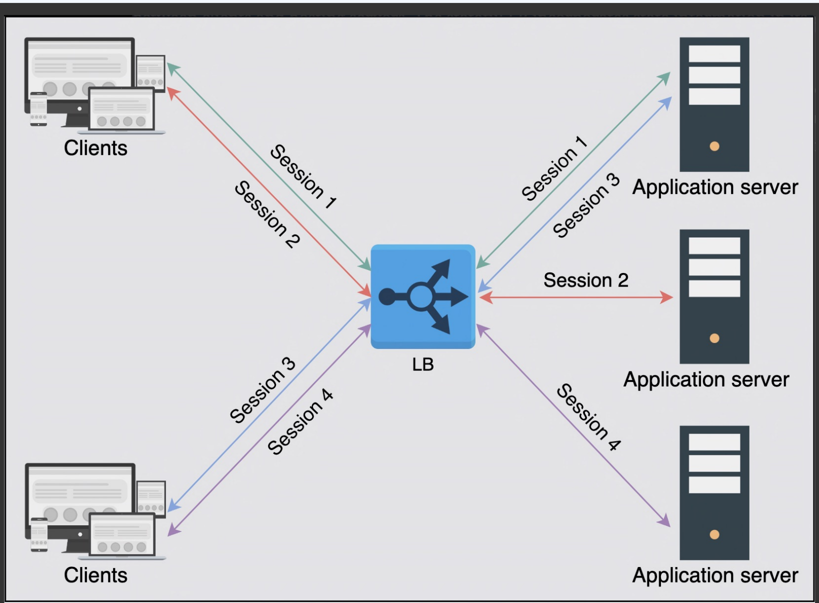
Static algorithms don't consider the changing state of the servers. Therefore, task assignment is carried out based on existing knowledge about the server's configuration. Naturally, these algorithms aren't complex, and they get implemented in a single router or commodity machine where all the requests arrive.

Dynamic algorithms are algorithms that consider the current or recent state of the servers. Dynamic algorithms maintain state by communicating with the server, which adds a communication overhead. State maintenance makes the design of the algorithm much more complicated.

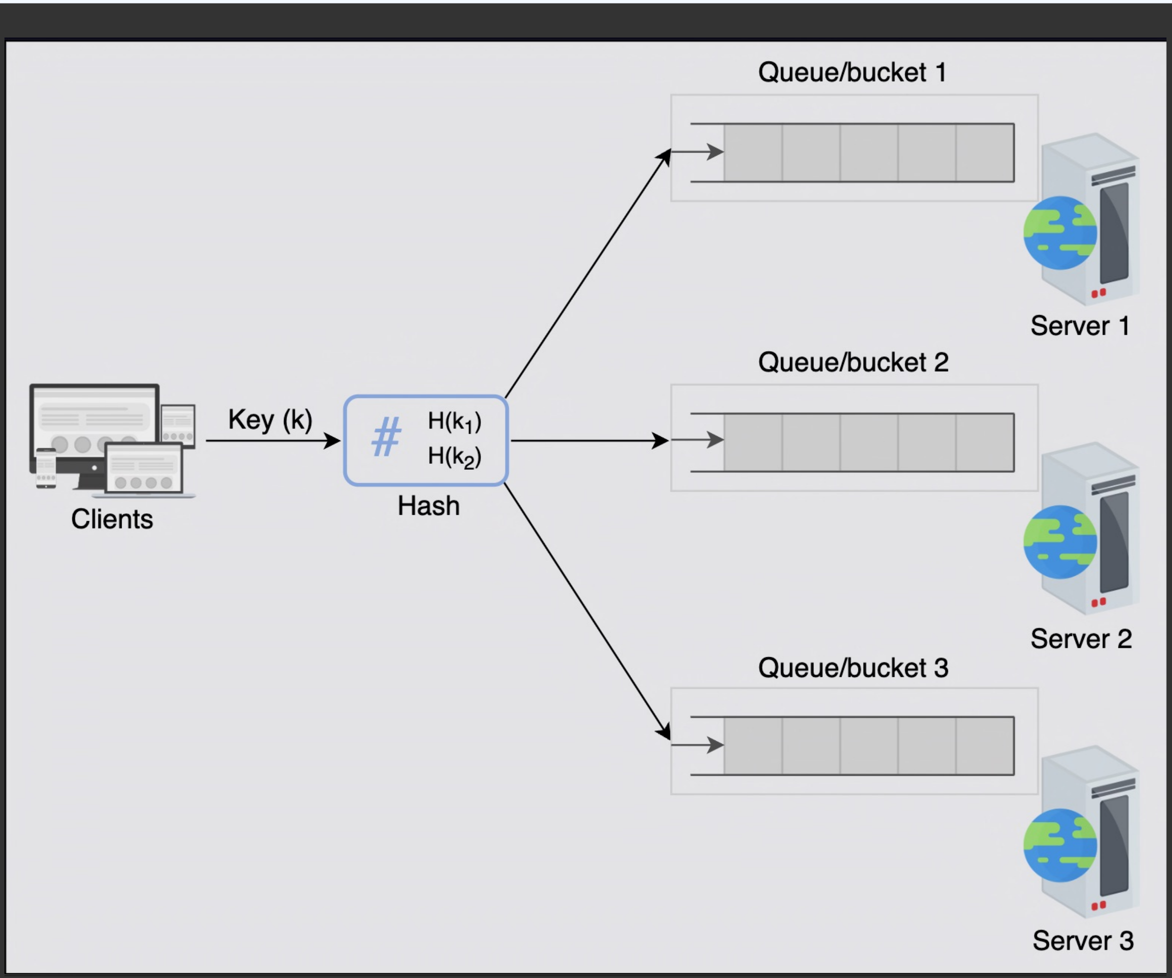
Stateful & Stateless LB's

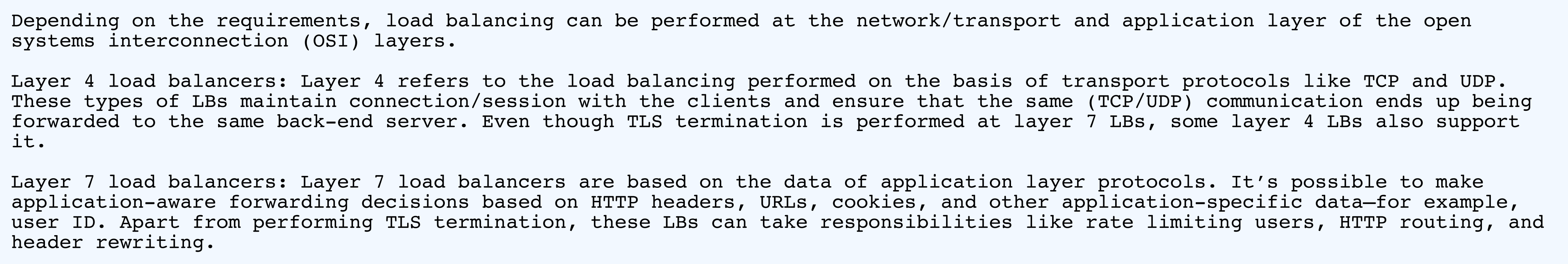


As the name indicates, stateful load balancing involves maintaining a state of the sessions established between clients and hosting servers. The stateful LB incorporates state information in its algorithm to perform load balancing.



Stateless load balancing maintains no state and is, therefore, faster and lightweight. Stateless LBs use consistent hashing to make forwarding decisions. However, if infrastructure changes (for example, a new application server joining), stateless LBs may not be as resilient as stateful LBs because consistent hashing alone isn’t enough to route a request to the correct application server.



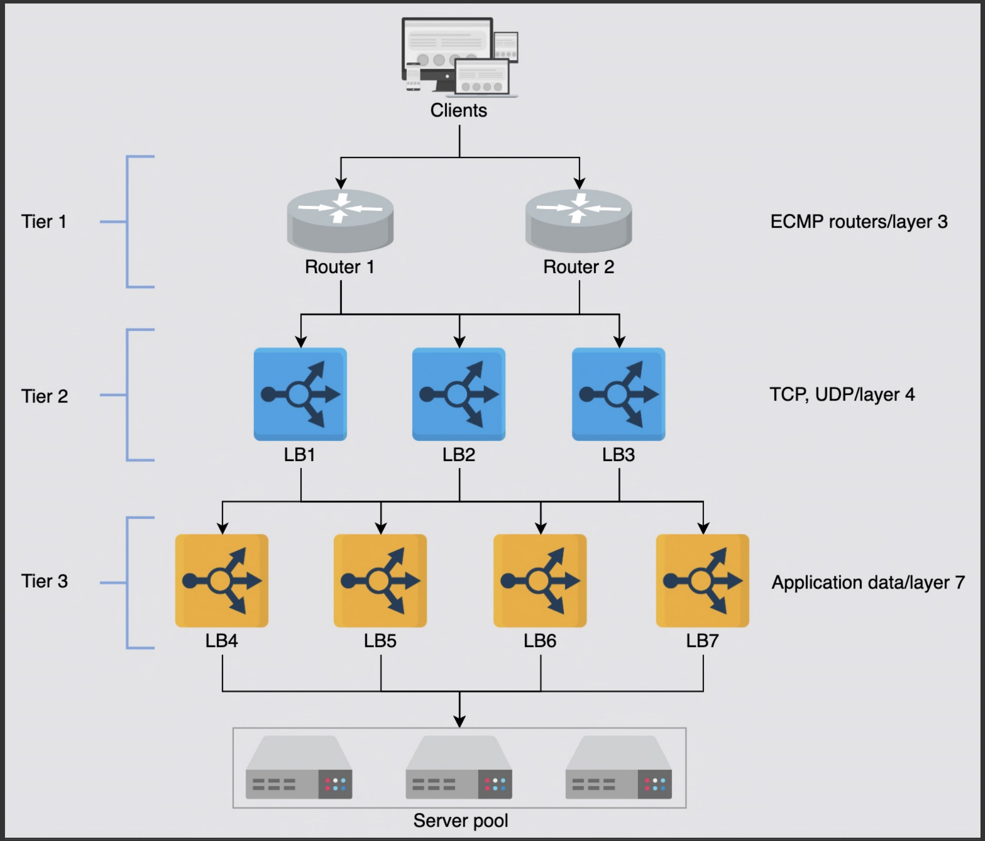


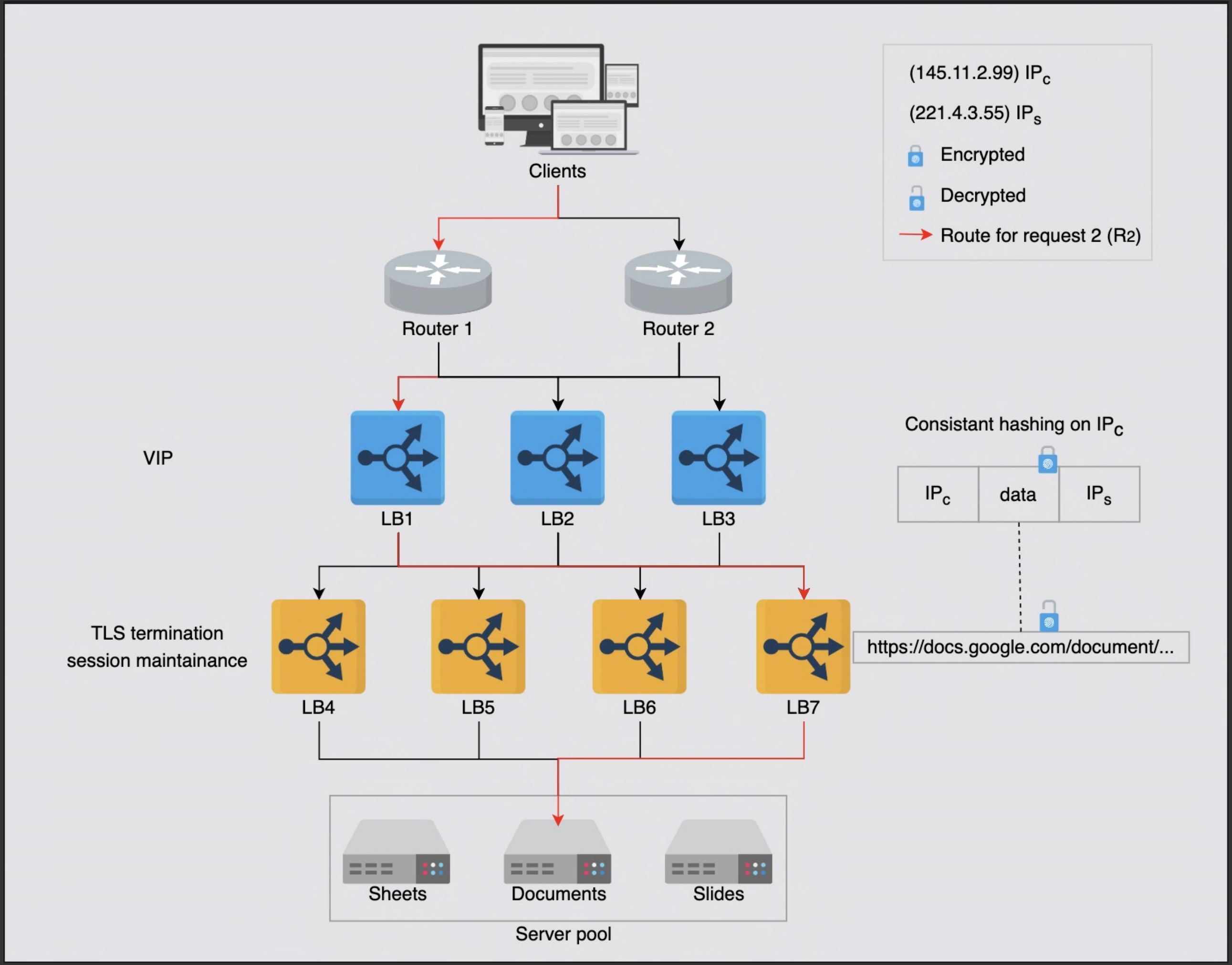
Depending on the requirements, load balancing can be performed at the network/transport and application layer of the open systems interconnection (OSI) layers.

Layer 4 load balancers: Layer 4 refers to the load balancing performed on the basis of transport protocols like TCP and UDP. These types of LBs maintain connection/session with the clients and ensure that the same (TCP/UDP) communication ends up being forwarded to the same back-end server. Even though TLS termination is performed at layer 7 LBs, some layer 4 LBs also support it.

Layer 7 load balancers: Layer 7 load balancers are based on the data of application layer protocols. It’s possible to make application-aware forwarding decisions based on HTTP headers, URLs, cookies, and other application-specific data—for example, user ID. Apart from performing TLS termination, these LBs can take responsibilities like rate limiting users, HTTP routing, and header rewriting.

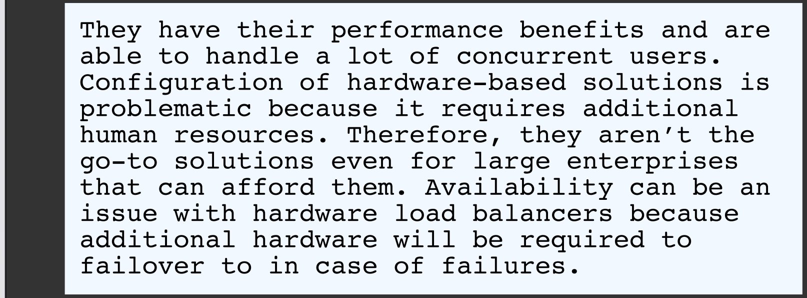
Load Balancer Deployment



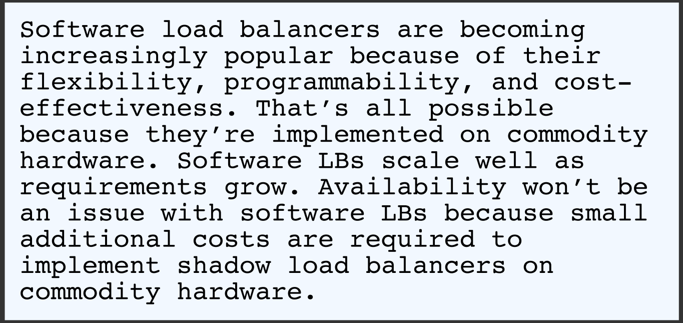


Implimentation of LB

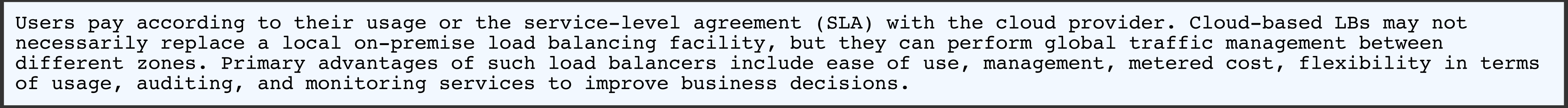
1. Hardware LB

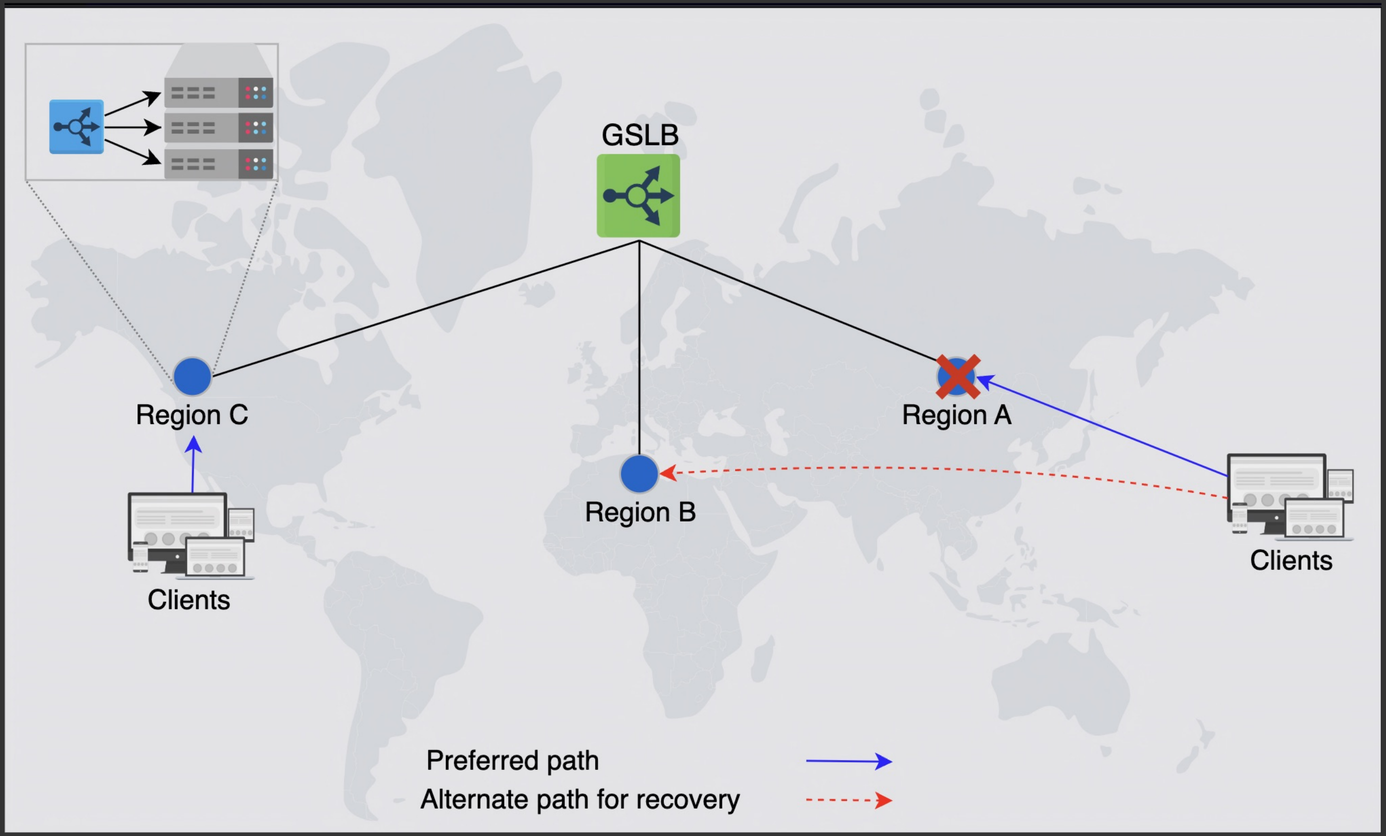

They have their performance benefits and are able to handle a lot of concurrent users. Configuration of hardware-based solutions is problematic because it requires additional human resources. Therefore, they aren’t the go-to solutions even for large enterprises that can afford them. Availability can be an issue with hardware load balancers because additional hardware will be required to failover to in case of failures.

2. Software Load balancers


Software load balancers are becoming increasingly popular because of their flexibility, programmability, and cost-effectiveness. That’s all possible because they’re implemented on commodity hardware. Software LBs scale well as requirements grow. Availability won’t be an issue with software LBs because small additional costs are required to implement shadow load balancers on commodity hardware.

3. Cloud Load balancers


Users pay according to their usage or the service-level agreement (SLA) with the cloud provider. Cloud-based LBs may not necessarily replace a local on-premise load balancing facility, but they can perform global traffic management between different zones. Primary advantages of such load balancers include ease of use, management, metered cost, flexibility in terms of usage, auditing, and monitoring services to improve business decisions.


GSLB
Region C
Region B
Region A
Clients
Clients
Preferred path
Alternate path for recovery