LOAD BALANCING IN CLOUD COMPUTING

Dr Hitesh Mohapatra

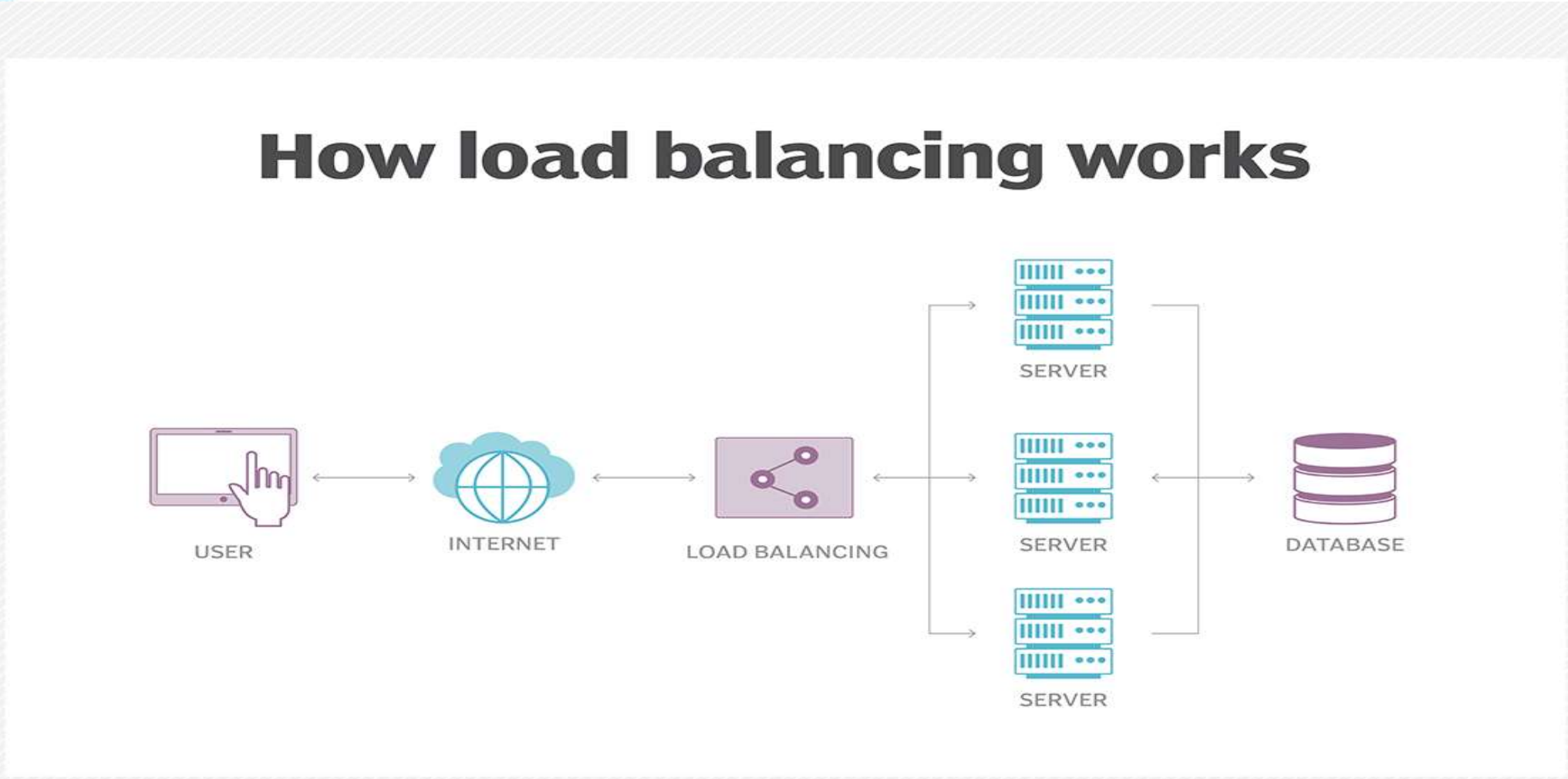
WHAT IS CLOUD LOAD BALANCING?

- Cloud load balancing is the process of distributing workloads across computing resources in a cloud computing environment and carefully balancing the network traffic accessing those resources.
- Load balancing enables organizations to meet workload demands by routing incoming traffic to multiple servers, networks or other resources while improving performance and protecting against disruptions in services.
- Load balancing also makes it possible to distribute workloads across two or more geographic regions.

CONT…

- Cloud load balancing helps enterprises achieve high-performance levels for potentially lower costs than traditional on-premises load balancing technology.
- Cloud load balancing takes advantage of the cloud's scalability and agility to meet the demands of distributed workloads with high numbers of client connections. It also improves overall availability, increases throughput and reduces latency.
- In addition to workload and traffic distribution, cloud load balancing services typically offer other features, such as application health checks, automatic scaling and failover and integrated certificate management.

HOW DOES IT WORK?



CONT.

- Cloud load balancing takes a software-based approach to distribute network traffic across resources, as opposed to hardware-based load balancing, which is more common in enterprise data centres.
- A load balancer receives incoming traffic and routes those requests to active targets based on a configured policy.
- A load-balancing service also monitors the health of the individual targets to ensure that those resources are fully operational.

HOW?

- In cloud computing, load-balancing technology resides between back-end servers and client devices.
- As requests are received, the load balancer distributes them to available servers using an algorithm that may take into account a variety of criteria, including geographical distance and server load.

WHAT ARE CLOUD LOAD-BALANCING TECHNIQUES AND ALGORITHMS?

- Load balancing in the cloud is usually achieved by using an algorithm to determine how to direct traffic.
- These algorithms typically fall into two categories: static and dynamic.

STATIC ALGORITHMS

- **Round Robin** forwards requests to each cloud server in a simple, repeating order.
- **Weighted Round Robin** assigns servers with greater capacity a higher “weight,” enabling them to receive more incoming application traffic.
- **IP Hash** performs a hash, or a mathematical computation, on the client IP address, converting it to a number and mapping it to individual servers.

DYNAMIC ALGORITHMS

- **Least Connections** distributes traffic to the servers with the fewest active connections.
- **Least Response** Time directs traffic to the servers with the lowest average response time and fewest connections.
- **Least Bandwidth** sends requests to the servers consuming the lowest amount of bandwidth during a recent period of time.

WHAT ARE THE BENEFITS OF CLOUD LOAD BALANCING?

- **Improved performance.** By automatically distributing workloads across multiple resources, load balancing enables applications running in the cloud to handle traffic spikes more easily.
- **Greater reliability.** Hosting applications at multiple cloud hubs enables organizations to route traffic around outages.
- **Reduced costs**. With software-based load balancing in the cloud, organizations can eliminate the cost of installing, housing, configuring, and maintaining on-premises load-balancing appliances.

CONT.

- **Improved flexibility.** By routing traffic to alternative servers, cloud load balancing supports the needs of development teams when performing updates, applying patches, remediating issues with servers, or conducting tests in production environments.
- **Better security.** Cloud load balancing improves defences against distributed denial-of-service (DDoS) attacks by spreading traffic across multiple servers and rerouting traffic away from overloaded servers.
- **Seamless scalability.** Cloud load balancing solutions can help scale applications automatically and efficiently to manage fluctuations in workloads.
- **Health checks.** Cloud DNS load balancers automatically perform periodic checks to monitor the health of upstream servers.

WHAT IS CLOUD LOAD BALANCING VS. TRADITIONAL LOAD BALANCING?

- Traditional load balancing technology is hardware-based, requiring IT teams to install, manage, and maintain proprietary hardware within a data center.
- In contrast, cloud load balancing tends to be a software-based technology, as most cloud vendors will not allow customer hardware to run within their environment.
- Software-based load balancers can run in any location or environment, and they are more affordable for smaller businesses.

CONT.

- **Reduced latency.** Load balancing minimizes response time for application users by spreading cloud workloads evenly across available resources.
- **Easier automation.** Cloud load balancing improves automation by enabling organizations to deliver insight into applications in near-real time and use predictive analytics to identify potential bottlenecks in advance.
- **Faster recovery.** During network emergencies or natural disasters, providers offering cloud load balancing can redirect traffic to other regions to ensure continuity and availability.

WHAT IS CLOUD LOAD BALANCING AS A SERVICE (LBAAS)?

- Load balancing as a service provides cloud load balancing on an as-needed basis, replacing on-premises, dedicated appliances.
- LBaaS provides greater scalability, as load balancing in the cloud can accommodate traffic spikes without needing to reconfigure physical infrastructure.
- Greater availability is ensured by connecting to the closest servers, and LBaaS can also reduce the cost of investment and maintenance when compared to hardware-based appliances.

WHAT ARE DIFFERENT TYPES OF LOAD BALANCING?

Load balancing in the cloud falls into four broad categories.

- **Application load balancing** redirects traffic by looking at the content of a request — for example, HTTP headers or SSL session IDs.
- **Network load balancing** technology considers IP addresses and other network information when redirecting traffic to an optimal resource.
- **Global server load balancing** redirects traffic to destinations that are geographically closest to the client to minimize latency.
- **DNS load balancing** configures a domain to route network requests across a collection of resources within the domain.

WHAT ARE THE TYPES OF LOAD BALANCING TECHNOLOGY?

Load balancers are one of two types: hardware load balancers and software load balancers.

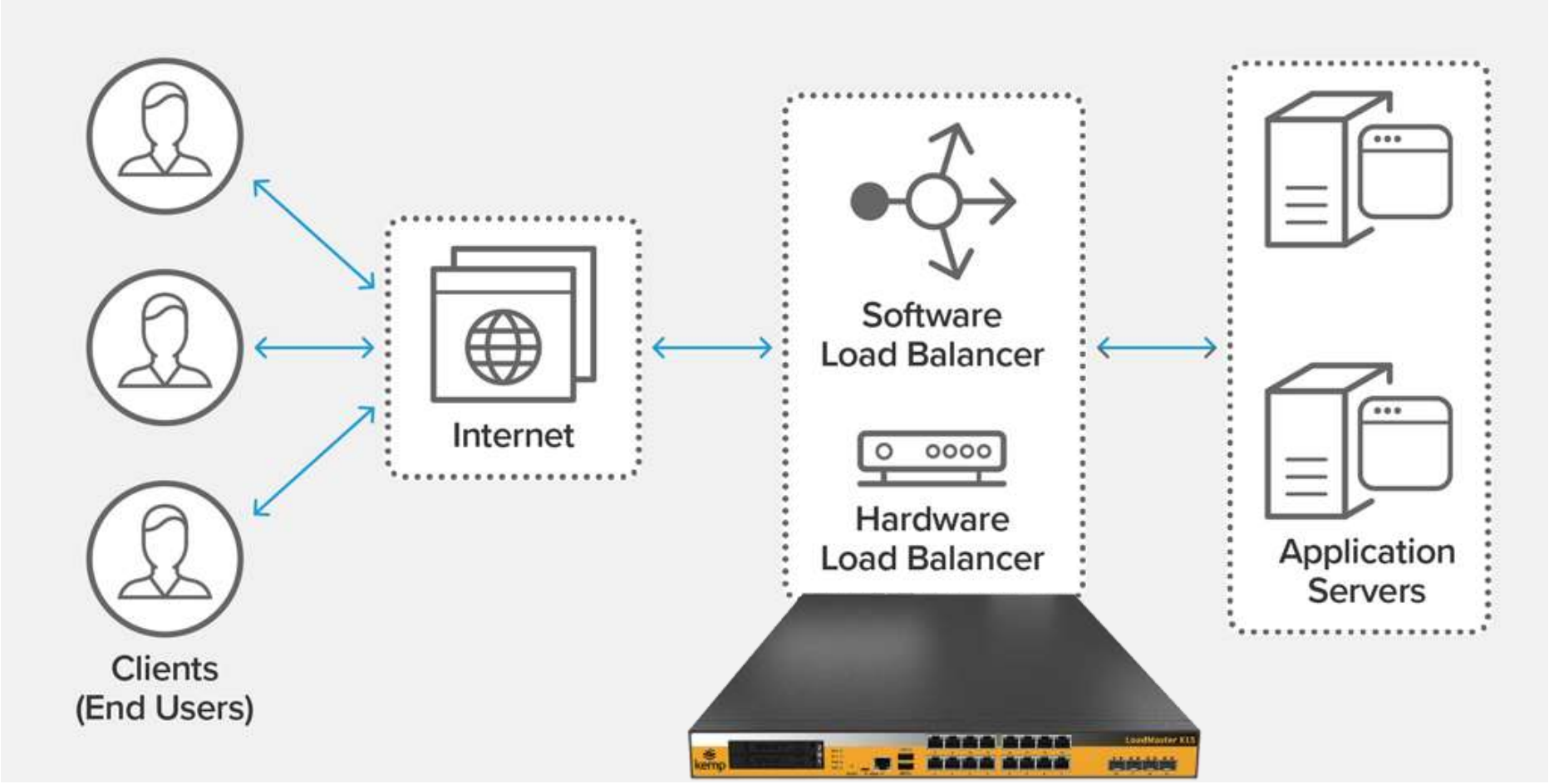
Hardware load balancers

- A hardware-based load balancer is a hardware appliance that can securely process and redirect gigabytes of traffic to hundreds of different servers. You can store it in your data centres and use virtualization to create multiple digital or virtual load balancers that you can centrally manage.
- Hardware-based load balancers are dedicated boxes which include Application Specific Integrated Circuits (ASICs) adapted for a particular use. ASICs allow high-speed promotion of network traffic and are frequently used for transport-level load balancing because hardware-based load balancing is faster in comparison to a software solution.

Software load balancers

- Software-based load balancers are applications that perform all load-balancing functions. You can install them on any server or access them as a fully managed third-party service.
- Software-based load balancers run on standard hardware (desktop, PCs) and standard operating systems.

TITLE AND CONTENT LAYOUT WITH CHART



EXAMPLES: AMAZON WEB SERVICES (AWS)

- Amazon Web Services (AWS) Elastic Load Balancing distributes incoming client traffic and routes it to registered targets such as EC2 instances. Elastic Load balancing supports four types of load balancers: Application, Network, Gateway and Classic. The load balancers differ in the features offered, the network layers at which they operate and supported communication protocols.

GOOGLE CLOUD PLATFORM

- The Cloud Load Balancing service available on Google Cloud Platform is built on the same front-end server infrastructure that powers Google. The service offers a range of load balancers that vary depending on whether the customer needs external or internal load balancing, global or regional load balancing, Premium or Standard network service tiers, proxy or pass-through services, among other factors.

MICROSOFT AZURE

- Microsoft Azure offers four load balancing services. Azure Traffic Manager is a (OSI model) layer 7 DNS-based traffic load balancer for delivering services across global Azure regions. Azure Load Balancer is a layer 4 network load balancer for routing traffic between VMs. Azure Application Gateway is a layer 7 delivery controller for regional applications. Azure Front Door is a highly secure, layer 7 global load balancer for microservice

THANK YOU

Any questions?