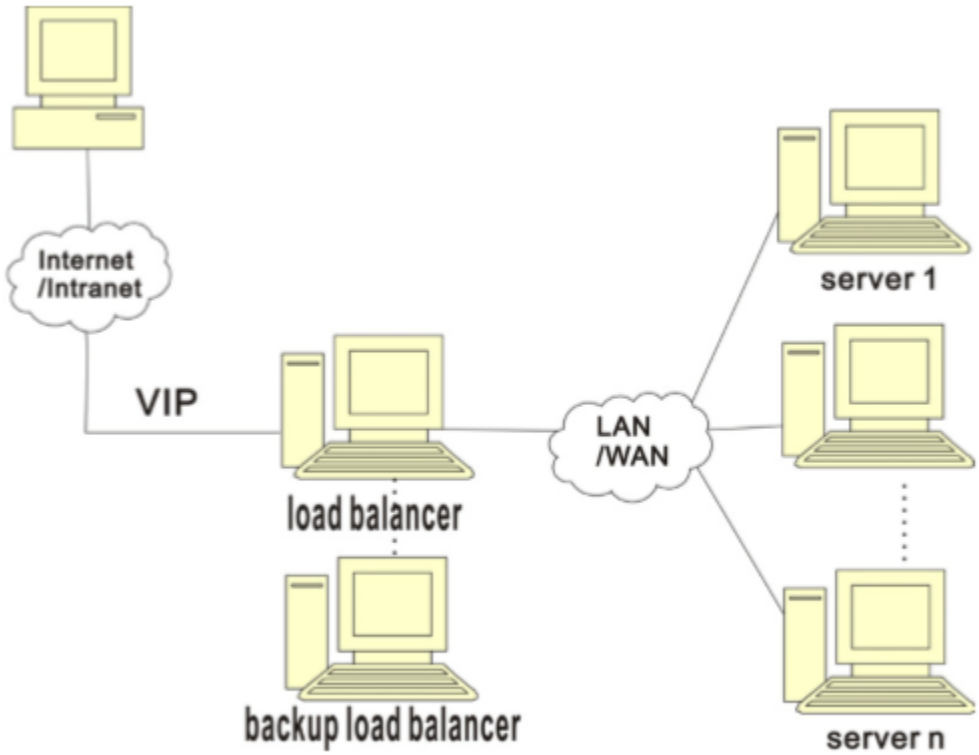
System Design, Chapter 3: Load Balancing

Load balancing

Load balancing is to distribute a large number of requests to different servers, to ease the burden of a single server. Load-balancing technology can balance conflicting factors such as cost, performance, and scalability, through a relatively low total cost of the computer cluster to achieve a strong performance that can not be achieved by stand-alone system. As a result of the introduction of load balancing, network and resources can be best made use of.

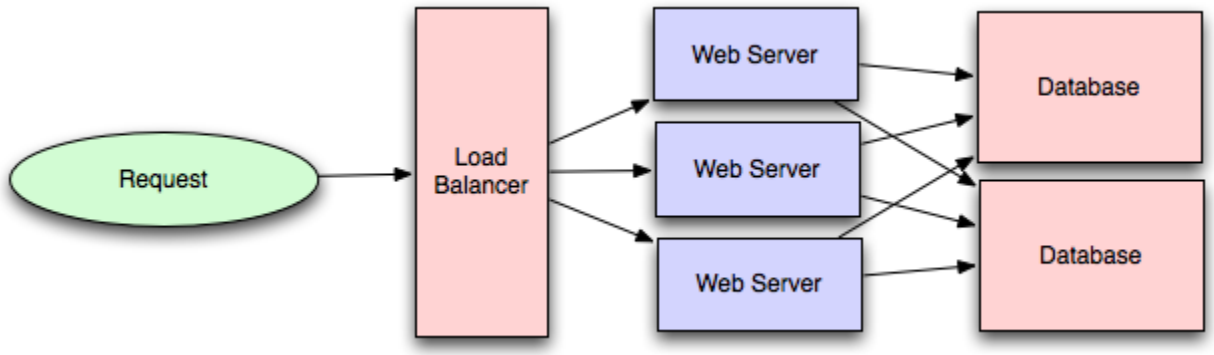


Generally speaking, load balancers fall into three categories:

- DNS Round Robin (rarely used): clients get a randomly-ordered list of IP addresses.
  pros: easy to implement and free
  cons: hard to control and not responsive, since DNS cache needs time to expire

- L3/L4 Load Balancer: traffic is routed by IP address and port. L3 is network layer (IP). L4 is session layer (TCP).
  pros: better granularity, simple, responsive
- L7 Load Balancer: traffic is routed by what is inside the HTTP protocol. L7 is application layer (HTTP).

It is good enough to talk in this level of detail on this topic, but in case the interviewer wants more, we can suggest exact algorithms like round robin, weighted round robin, least loaded, least loaded with slow start, utilization limit, latency, cascade, etc. and for L4/L7 load balancer, please read below.



Load Balancing

L4 Load Balancer

“Layer 4 load balancing” most commonly refers to a deployment where the load balancer’s IP address is the one advertised to clients for a web site or service (via DNS, for example). As a result, clients record the load balancer’s address as the destination IP address in their requests.

When the Layer 4 load balancer receives a request and makes the load balancing decision, it also performs Network Address Translation (NAT) on the request packet, changing the recorded destination IP address from its own to that of the content server it has chosen on the internal network. Similarly, before forwarding server responses to clients, the load balancer changes the source address recorded in the packet header from the server’s IP address to its own. (The destination and source TCP

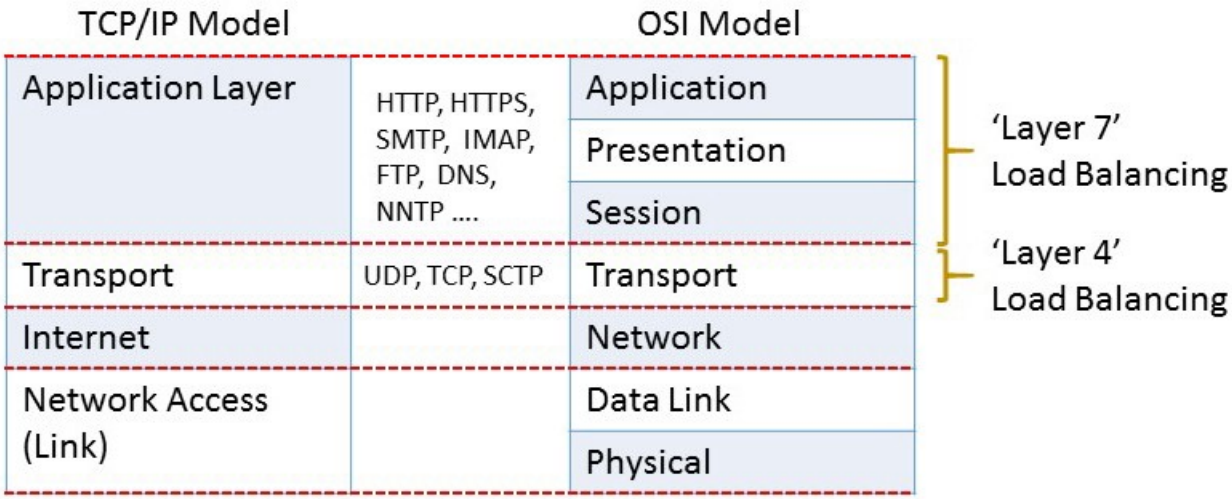
port numbers recorded in the packets are sometimes also changed in a similar way.)

Layer 4 load balancers make their routing decisions based on address information extracted from the first few packets in the TCP stream, and do not inspect packet content. A Layer 4 load balancer is often a dedicated hardware device supplied by a vendor and runs proprietary load-balancing software, and the NAT operations might be performed by specialized chips rather than in software.

Layer 4 load balancing was a popular architectural approach to traffic handling when commodity hardware was not as powerful as it is now, and the interaction between clients and application servers was much less complex. It requires less computation than more sophisticated load balancing methods (such as Layer 7), but CPU and memory are now sufficiently fast and cheap that the performance advantage for Layer 4 load balancing has become negligible or irrelevant in most situations.

L7 Load Balancer

Layer 7 load balancing enables ADC (Application Delivery Controllers) to redirect traffic more intelligently by inspecting content to gain deeper context on the application request. This additional context allows the ADC to not only optimize load balancing but to also rewrite content, perform security inspections and to implement access controls.



An Example of Layer 7 Load Balancing

Let's look at a simple example. A user visits a high-traffic website. Over the course of the user's session, he or she might request static content such as images or video, dynamic content such as a news feed, and even transactional information such as order status. Layer 7 load balancing allows the load balancer to route a request based on information in the request itself, such as what kind of content is being requested. So now a request for an image or video can be routed to the servers that store it and are highly optimized to serve up multimedia content. Requests for transactional information such as a discounted price can be routed to the application server responsible for managing pricing. With Layer 7 load balancing, network and application architects can create a highly tuned and optimized server infrastructure or application delivery network that is both reliable and efficiently scales to meet demand.

Knowledge is Power!

Hope you like it and looking forward for next chapters. I will add Goto Next Chapter link here soon.