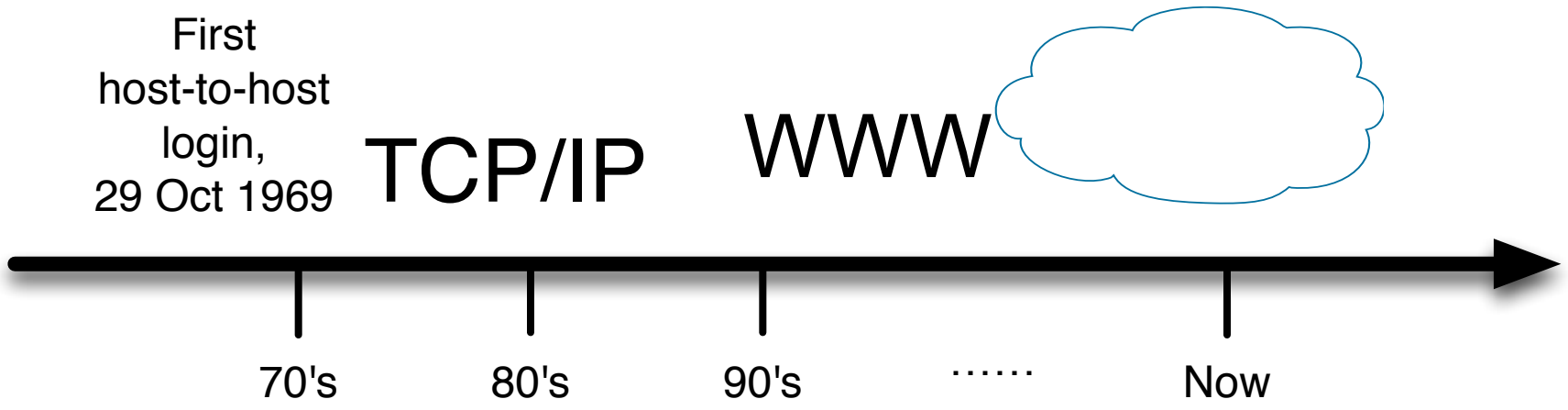
Server Load Balancing

Dr Posco Tso

p.tso@ljmu.ac.uk

http://www.poscotso.info

Very brief history of computer network



Very brief history of computer network

Today's Mega Data Centres

Today's Mega Data Centres

Today's Mega Data Centres

Today's Mega Data Centres

What is Load Balancing?

Load balancing is a computer networking methodology

- to distribute workload across multiple computers or a computer cluster, network links, central processing units, or other resources
  - To achieve optimal resource utilisation, maximise throughput, minimise response time, and avoid overload

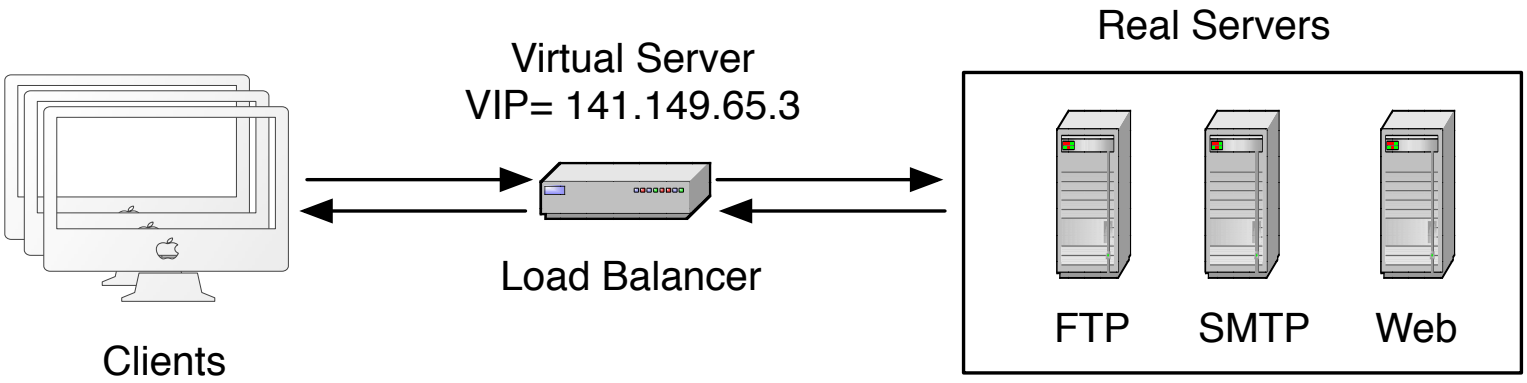
Why it is needed?

There are two dimensions that drive the need for load balancing: **servers** and **networks**.

- Must ensure scalability and high availability for all components starting from the edge routers that connect to the Internet, to the database servers in the back end.

What is load balancer?

Conceptually, server load balancers (SLBs) are the bridge between the servers and the network. They understand many higher-layer protocols and network protocols.



Major load balancer applications

Load balancers have at least four major applications:

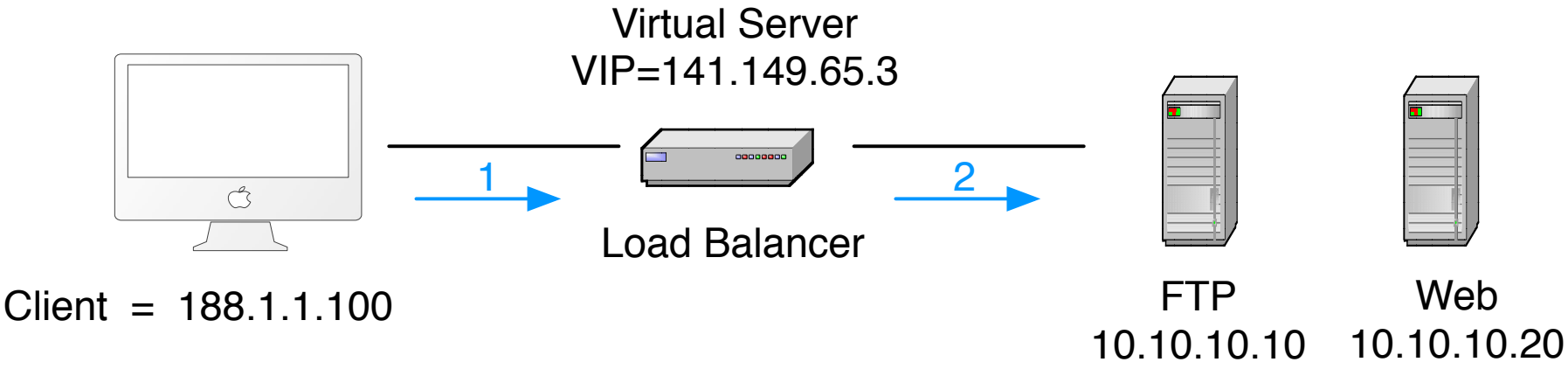
- Server load balancing deals with distributing the load across multiple servers
- Global server load balancing deals with directing users to different data center sites consisting of server farms
- Firewall load balancing distributes the load across multiple firewalls
- Transparent cache switching transparently directs traffic to caches

The benefits of server load balancing

By deploying the load balancer, we can immediately gain several benefits:

- Scalability. The collective processing capacity of the virtual server is far greater than the capacity of one server.
- Availability. The load balancer continuously monitors the health of the real servers and the applications running on them.
- Manageability. By deploying a load balancer, we can transparently take the server offline for maintenance without any downtime.
- Security. Load balancers are the front end to the server farm, they can protect the servers from malicious users.
- Quality of Service. Load balancers can be used to distinguish the users based on some information in the request packets to provide the desired class of service.

Basic packet flow in load balancing – an example



Load distribution methods

- *Stateless load balancing*: The load balancer uses some algorithms to distribute all the incoming traffic to available servers but **does not keep track** of any individual session.
- *Stateful load balancing*: The load balancer **keeps track** of state information for every session and makes load-balancing decision for each session.

Stateless load balancing

Stateless load balancing involves hashing algorithms on one or more of the following information in the IP packets: source IP, destination IP, source port and destination port.

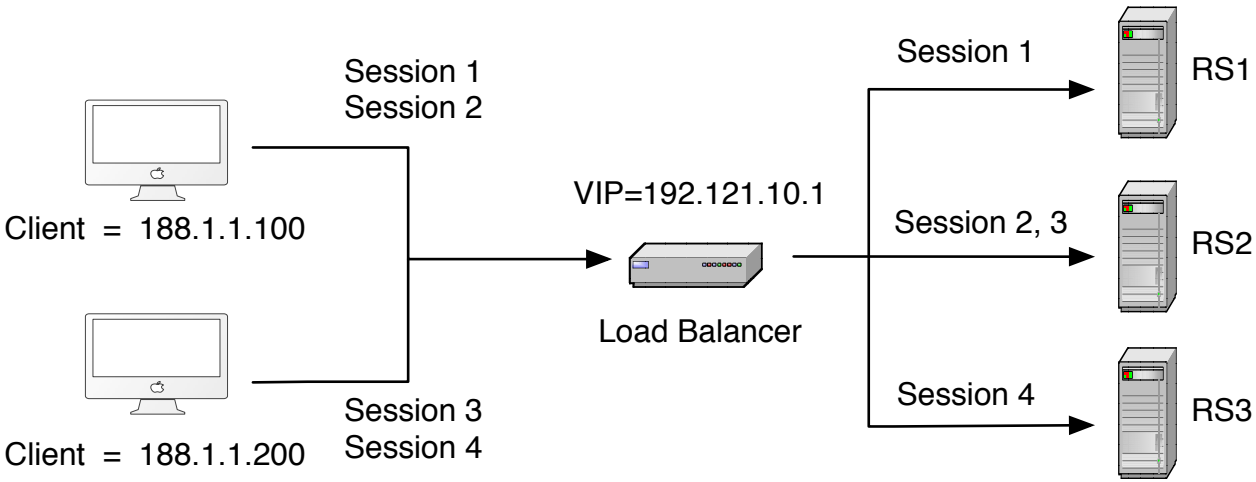
- A hashing computation will direct all packets belong to the same session to the same server
- Simple hashing is easy to implement, but is not robust to server failure.

1. md5(packet header) => hex value
2. hex value % n (# of paths) => 0 … n-1

Stateful load balancing

The load balancer users ***session table*** to keep track of sessions

Protocol	Source IP	Destination IP	Source Port	Destination Port	Server
TCP	188.1.1.100	192.121.10.1	2001	80	RS1
TCP	188.1.1.100	192.121.10.1	2002	80	RS2
TCP	188.1.1.200	192.121.10.1	4500	80	RS3
UDP	188.1.1.200	192.121.10.1	4501	6201	RS3

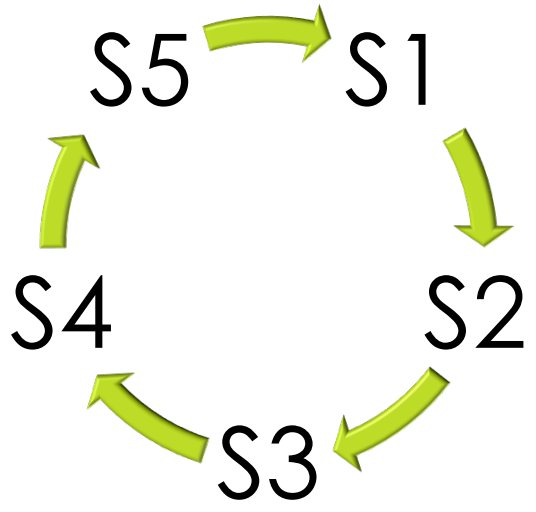


Stateful load balancing

Round-Robin

A load balancer give each connection to a server in a round-robin manner.

- Simple and lightweight
- uneven distribution of load

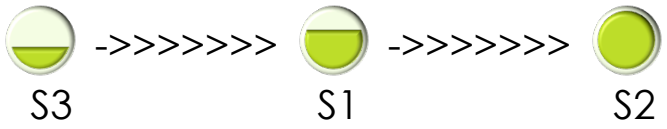


Stateful load balancing

Least Connections

A load balancer sends a new request to the server with the least number of concurrent connections.

- load balancer needs to keep track of the total number of concurrent active connections on each server
- it balances active connections, not actual computation resources.



Stateful load balancing

Weighted Distribution

A load balancer then assigns a new request to the server in accordance to the weights.

- Server administrators specify the relative capacity of each server by assigning a weight to each server
- it needs to be used with another load-balancing method, e.g. least connections.
- it is a great way to mix servers of different capabilities.

Stateful load balancing

Response Time

A load balancer measures and sends new requests to the server providing the fastest response time.

- It is good for services with performance constraints, e.g. to provide the best response time.
- Due to its complexity, it is not the best load-distribution method.

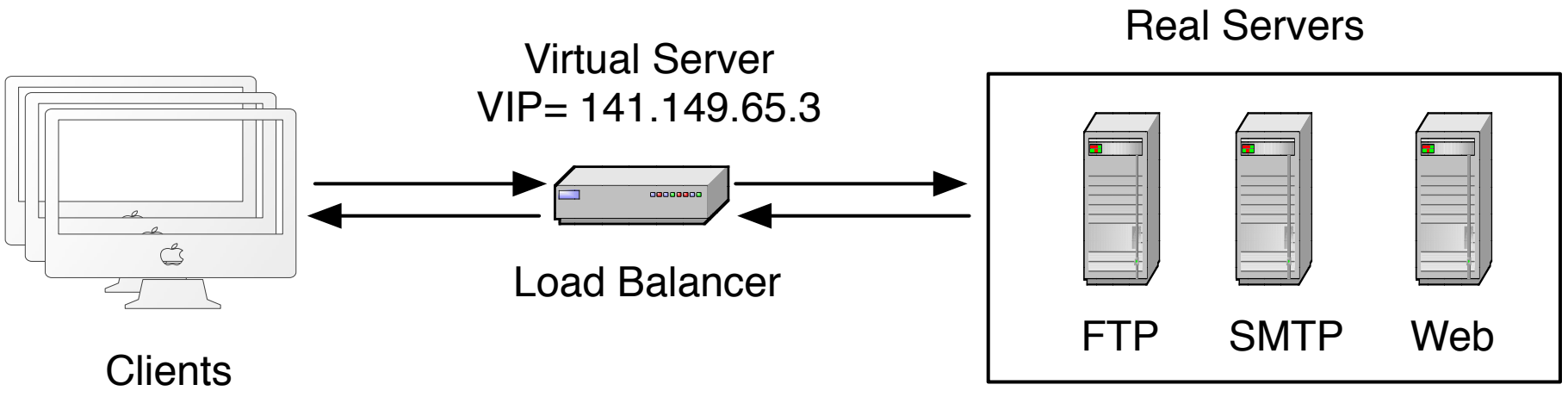
Stateful load balancing

Server Probes

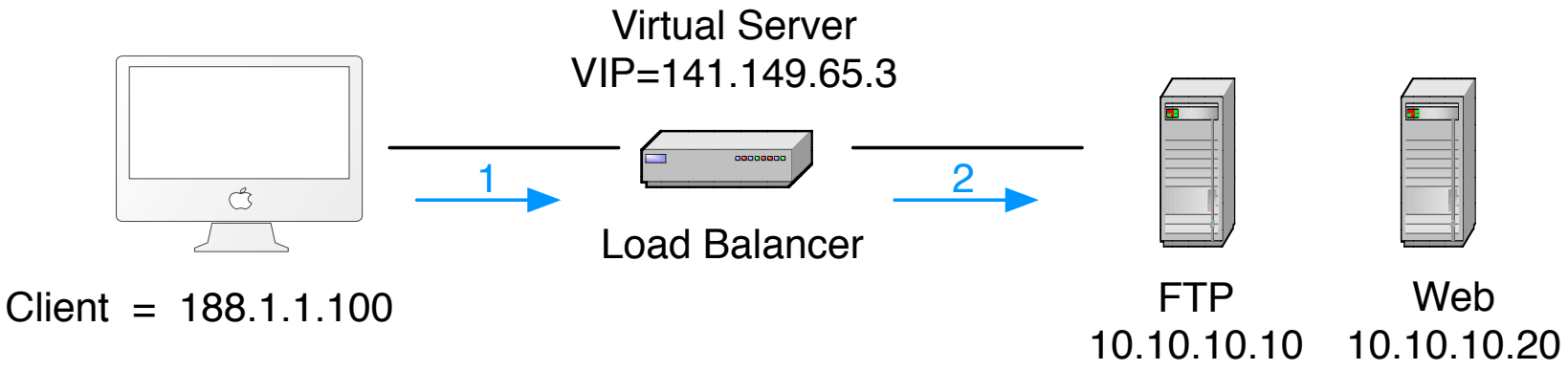
The load balancer detects the load conditions on the server at a very detailed level.

- It requires programs (known as server-sides agents) to run on each server
- It requires extra effort for agents maintenance.
- It is uncertain whether the agents will accurately reflect the load on the server

Network Address Translation (NAT)



Network Address Translation



Network Address Translation (NAT) is the process of modifying IP address information in IP packet headers while in transit across a traffic routing device.

Destination NAT

The process of changing the destination address in the packets is referred to as destination NAT.

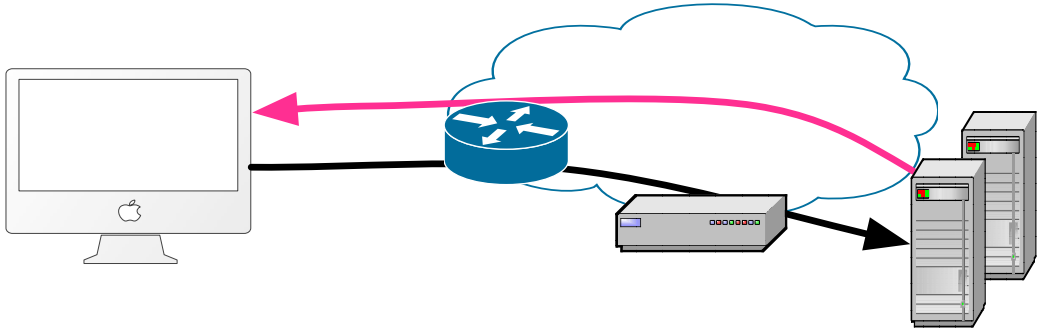
- Most load balancers perform destination NAT by default.
- Since destination NAT deals with changing only the destination address, it‘s also sometimes referred to as half-NAT.

Source IP	Source Port	Destination IP	Destination Port
188.1.1.100	80	141.149.65.3	80
188.1.1.100	80	10.10.10.20	8080

Source NAT

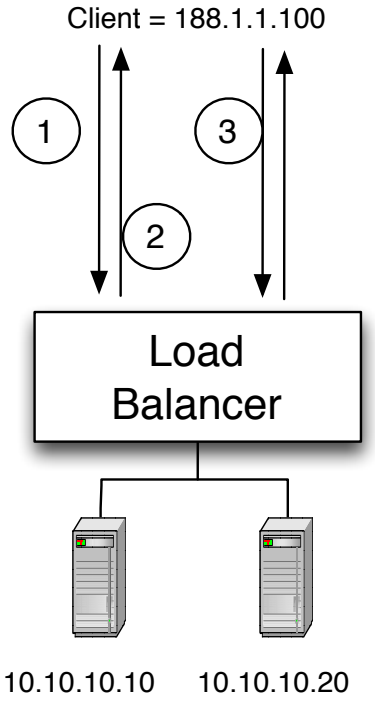
The load balancer changes the source IP address in the packets along with destination IP address translation.

- This is also sometimes referred to as full-NAT, as this involves translation of both source and destination addresses.
- Source NAT is generally not used unless there is a specific network topology that requires source NAT.



Enhanced NAT

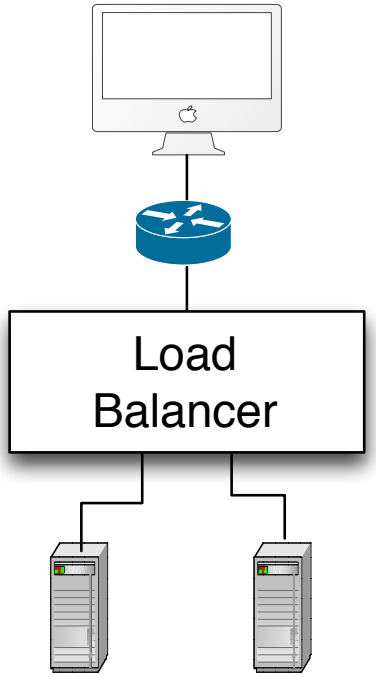
The NAT performed by the load balancer with protocol-specific knowledge in order to make certain protocols work with load balancing, e.g. streaming media protocols.

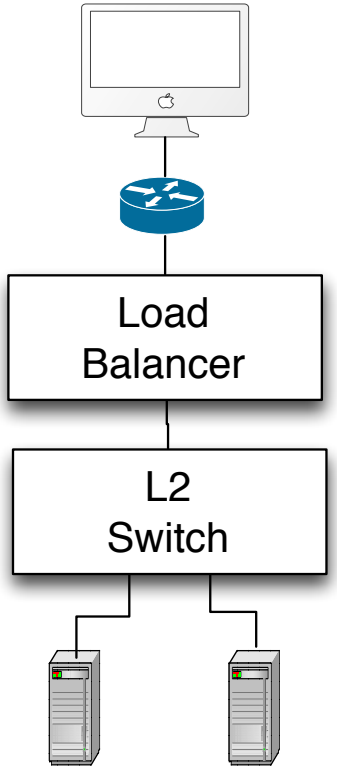


Step	Load Balancer's Action
1	Control connection received at the VIP for a well-know TCP port load balanced to a real server
2	Change any IP address or port information about the data connection
3	Assign the data connection to a server even though the destination UDP port is not specifically bound to a real server in the load balancer's configuration

Network Design with Load Balancers

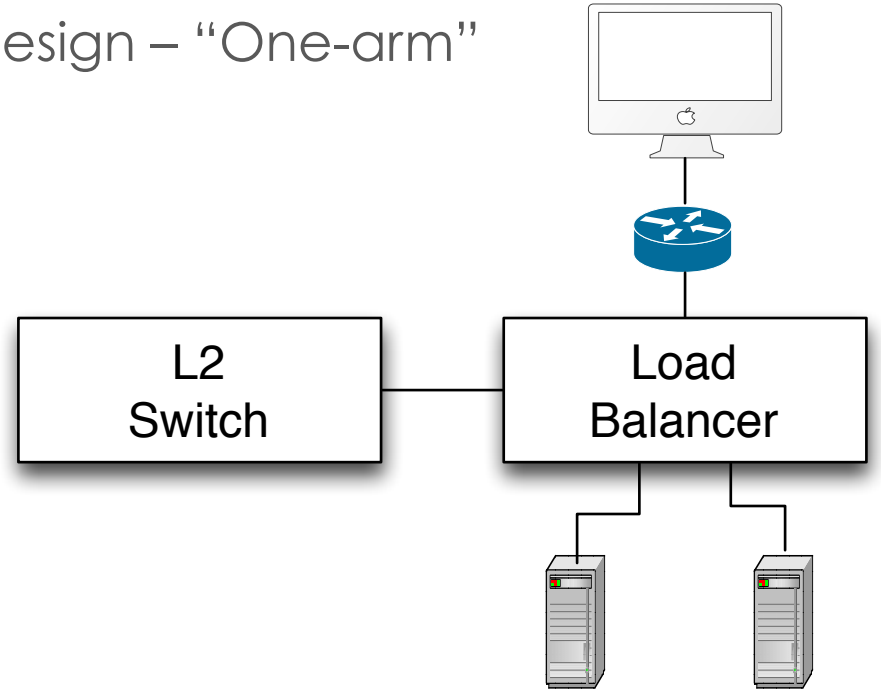
- Simple Design





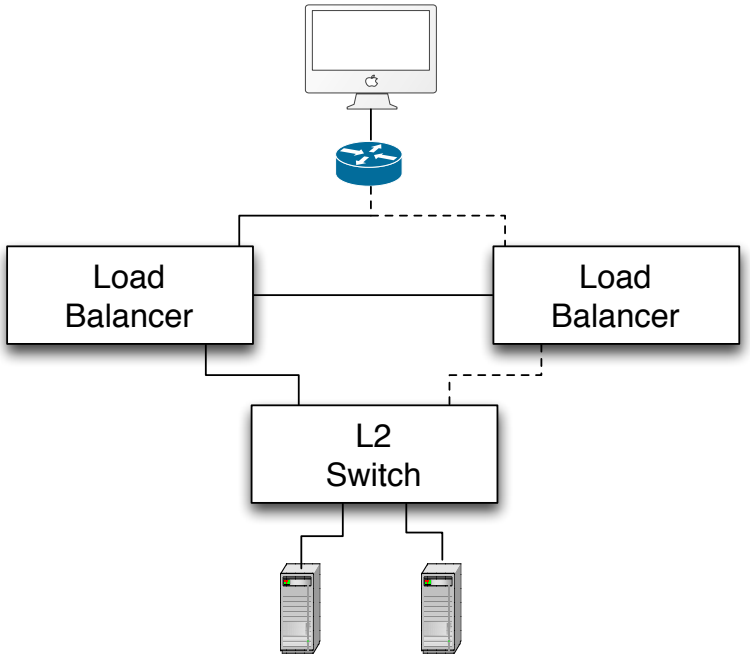
Network Design with Load Balancers

- Simple Design – "One-arm"



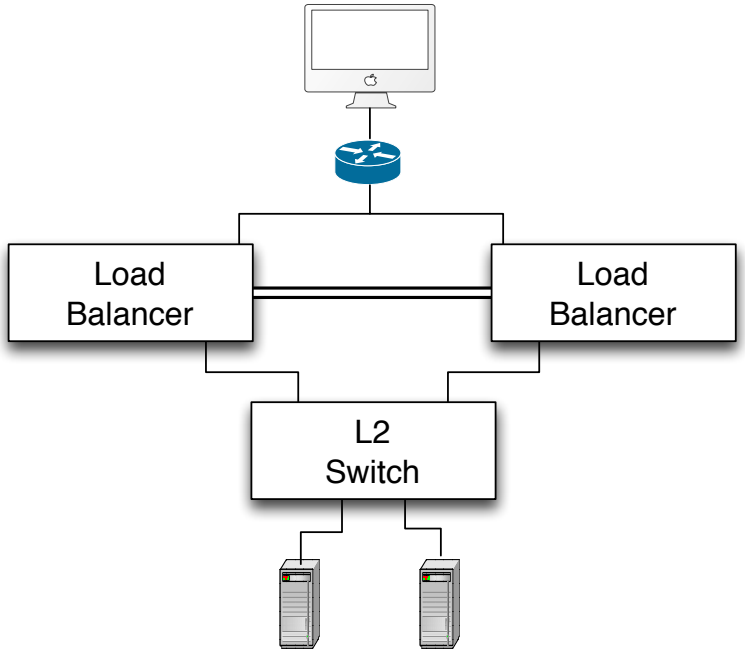
Network Design with Load Balancers

- High Availability – "Active-Standby"



Network Design with Load Balancers

- High Availability – "Active-Active"



Conclusion

- Load balancing/balancer
- load balancer applications
- The benefits of server load balancing
- Load distribution methods
- Network address translation
- Network design with load balancers