Load Balancer – A Candidate Solution for CAN

Shraddha Hegde

CAN BoF, IETF-113

CAN Problem Statement

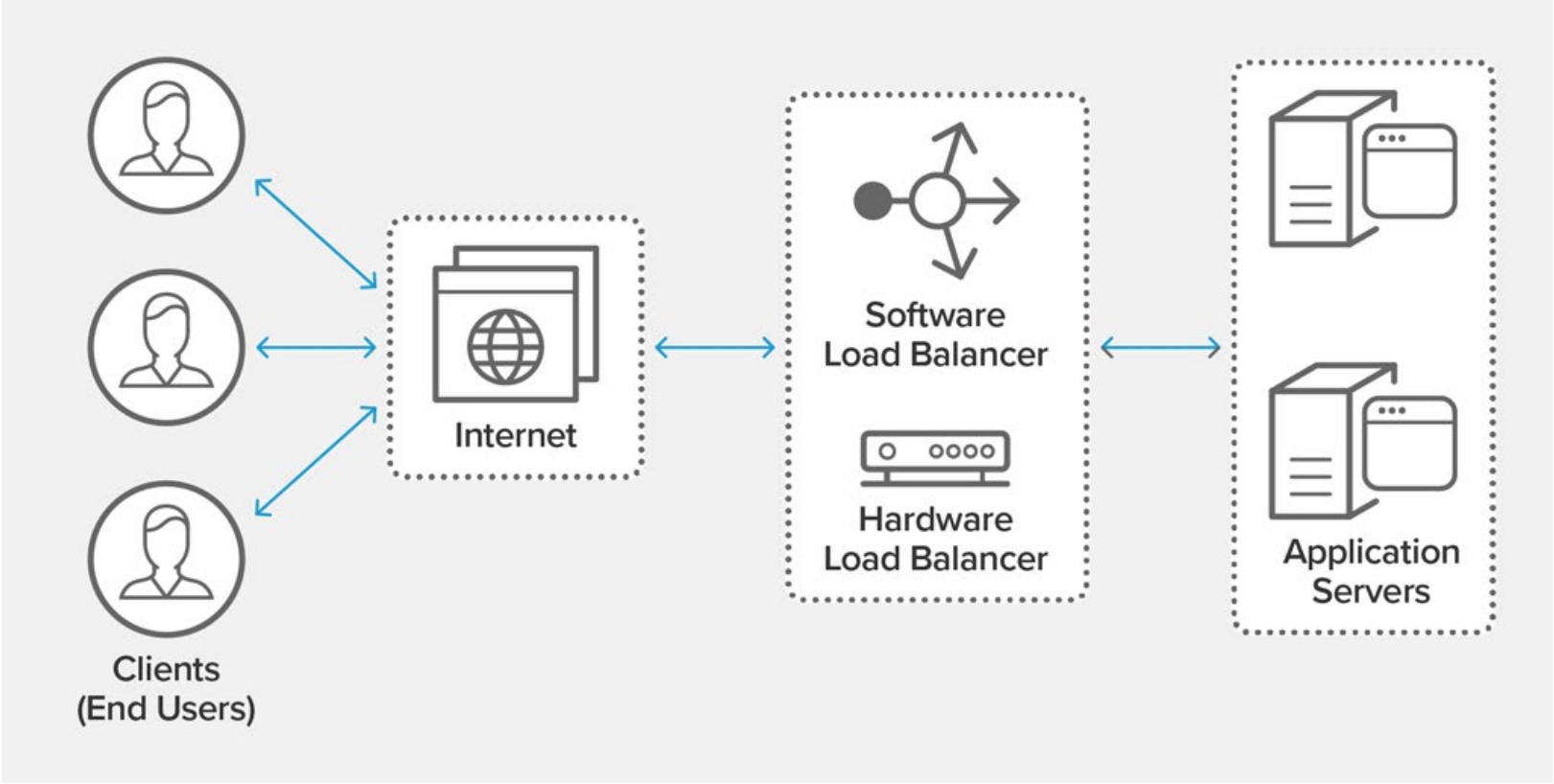
- Steer Traffic to computing resources with considerations of:
  - Computing resource load metric
  - Network metric
  - Service Affinity
    - Same computing resource SHOULD always be used
      - Even when a mobile client moves from one anchoring UPF to another
- Sounds very familiar …

Load Balancer (LB)

- A random Google search led to the definition

A load balancer acts as the “traffic cop” sitting in front of your servers and routing client requests across all servers capable of fulfilling those requests in a manner that maximizes speed and capacity utilization and ensures that no one server is overworked, which could degrade performance. If a single server goes down, the load balancer redirects traffic to the remaining online servers. When a new server is added to the server group, the load balancer automatically starts to send requests to it.

- Distributes client requests or network load efficiently across multiple servers
- Ensures high availability and reliability by sending requests only to servers that are online
- Provides the flexibility to add or subtract servers as demand dictates



Session Persistence

…it is essential that all requests from a client are sent to the same server for the duration of the session. This is known as session persistence.

The best load balancers can handle session persistence as needed. Another use case for session persistence is when an upstream server stores information requested by a user in its cache to boost performance. Switching servers would cause that information to be fetched for the second time, creating performance inefficiencies.

Modern load-balancers

- **Service Discovery**

  Abstract the service instance location and their names/addresses

- **Health checking**

  Find the availability of a service instance

  Route traffic around overloaded service instances

- **Load-balancing**

  Intelligent load balancing algorithms

  Keep the traffic within Zones

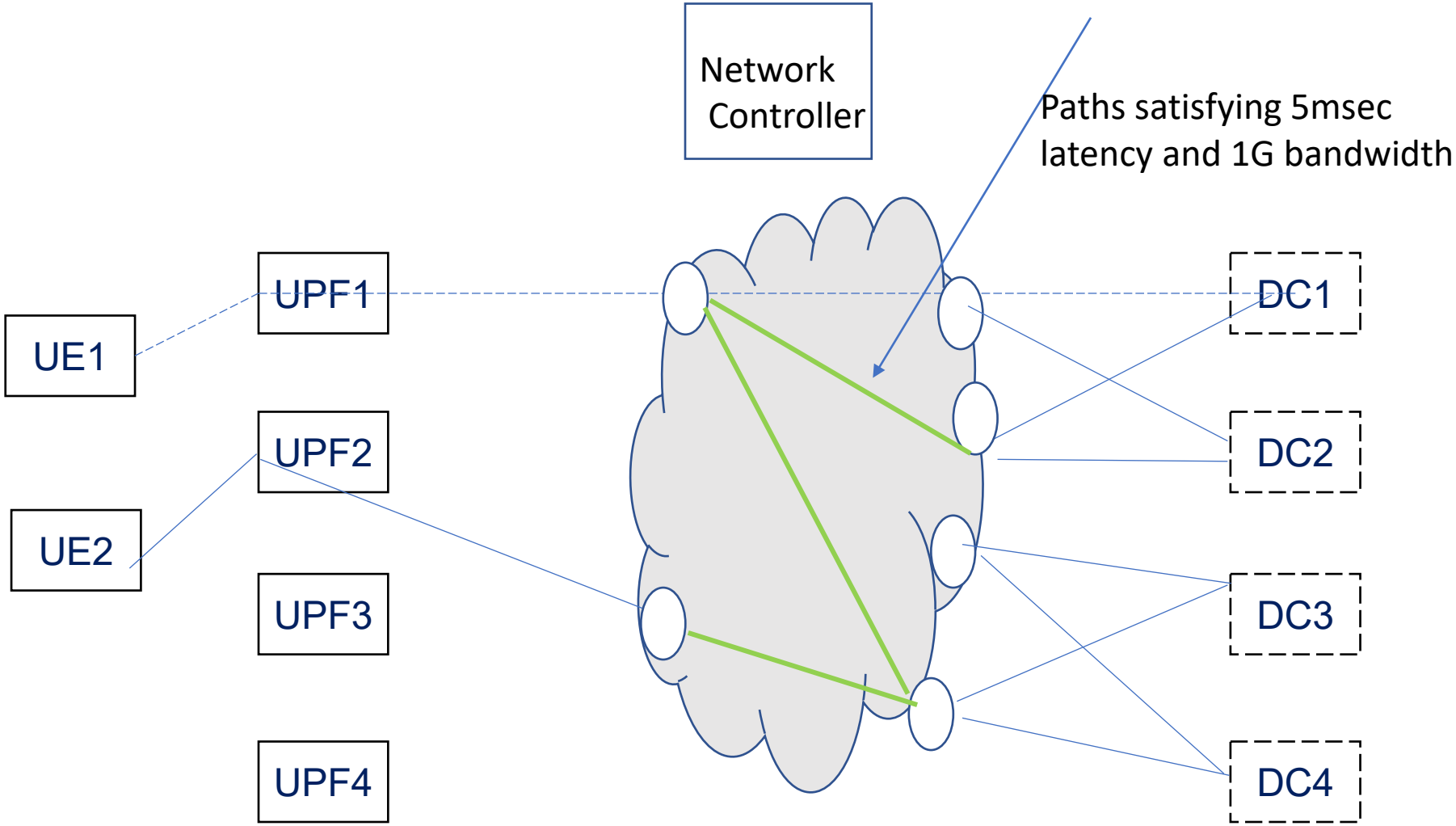
PS:Information collected from google search. Most information comes from official blogs of commercial load-balancer products

Types of Load balancers

- L4 Load Balancers
  - Load-balance based on TCP/UDP headers
  - High scalability/Performance
  - Suffer in efficiency due to application layer invisibility
- L7 Load balancers
  - Look into application layer information
  - Most efficient load-balancing
  - May be slower due to deep packet inspection
- Combination of L4/L7 load-balancers
  - Hybrid model
  - Benefits of both L4 and L7 load-balancers

Network layer with strict SLA guarantees

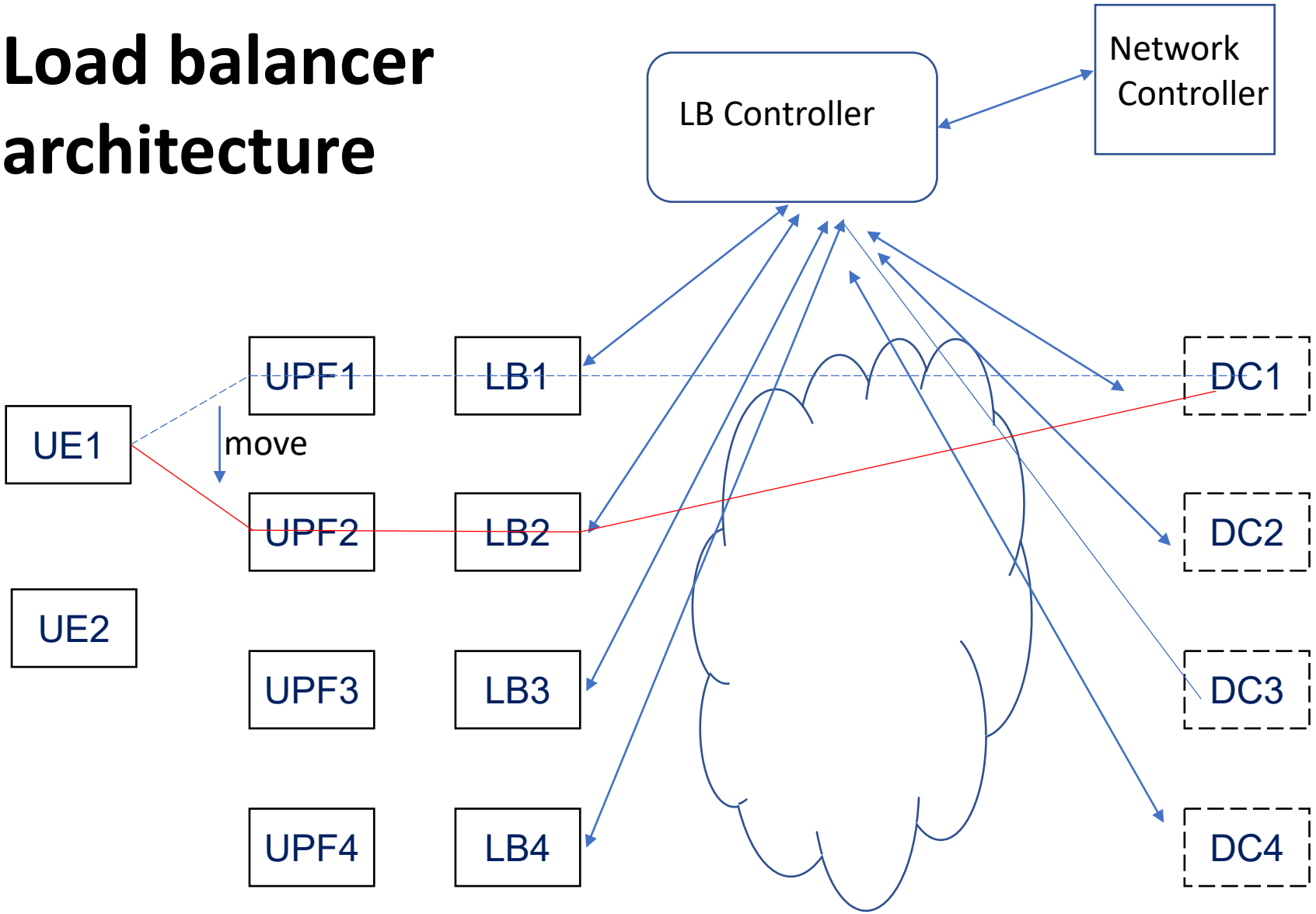
- Network layer responsible for delivering strict SLAs
  - Traffic engineering
    - Bandwidth guaranteed paths
    - Latency bounded paths
  - Network slicing
    - Reserved resources
    - Pack prioritization and queueing


Network Controller
Paths satisfying 5msec latency and 1G bandwidth
UE1
UE2
UPF1
UPF2
UPF3
UPF4
DC1
DC2
DC3
DC4

Load Balancer for CAN

- LBs and servers at “overlay”
  - Could be at different locations
  - LBs turn anycast server address in traffic to individual non-anycast addresses
    - Via tunneling or direct change of DST address in original packets
  - Exchange resource load metric
  - Learning network metric
    - Those exchanged by routing protocols
    - Actual delay/whatever measurement
- Multiple LBs with an LB-controller
  - E.g. one LB attached to each edge UPF; LBs could be VNFs
  - Exchange flow session information
    - required for service affinity when clients move
    - This may be better handled at the application/overlay level vs. complicating routing

Load balancer architecture



> **Network controllers**
> > creation stringent SLA paths
> > Management and monitoring of SLA paths

> Load-balancers
>On-demand Network information collection through APIs

Wireline/wireless usecases for load-balancers

> Need to take care of UE movement for wireless

Summary

- CAN requirements involve selecting service instance based on current state of the service instance
- Modern day load-balancers perform most sophisticated functions of choosing the right service instance
- Load-balancers running as overlay abstract the applications from the network
- Intelligent load-balancers with smart network layer may be able to meet the CAN requirements