LLM Load Balancer

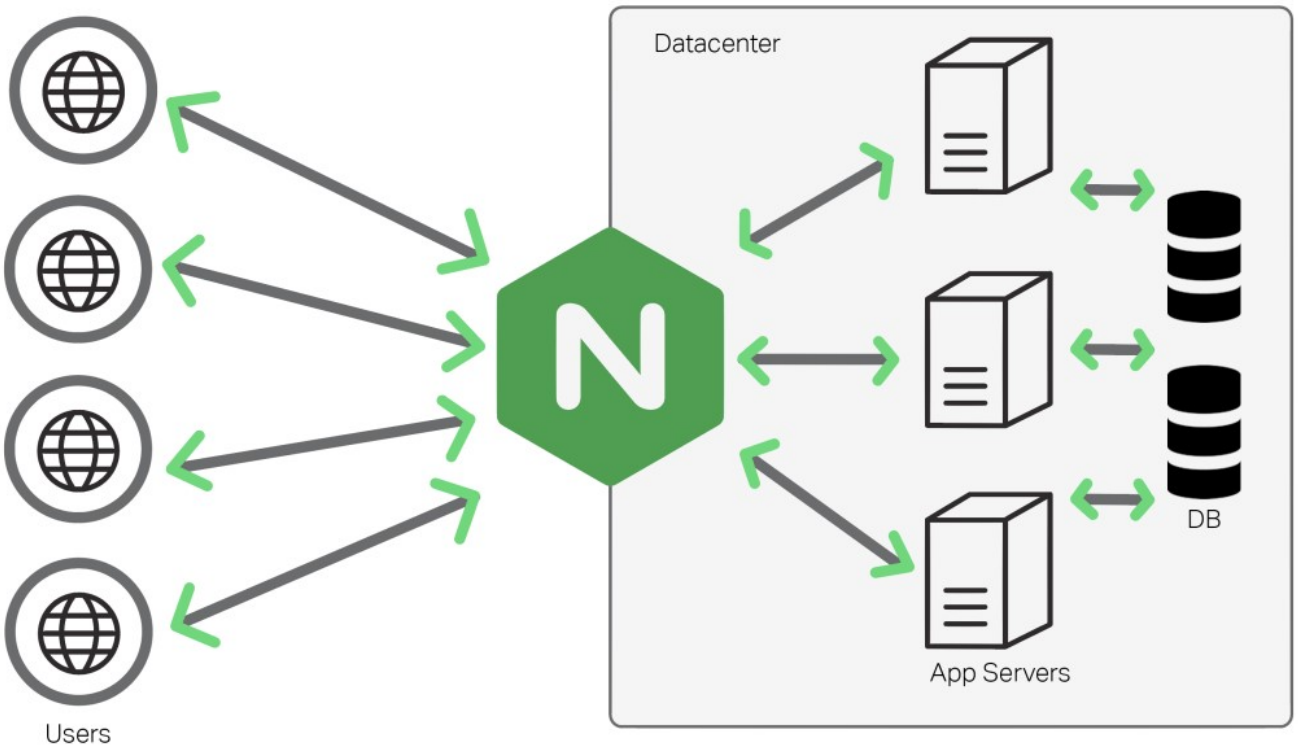
Blaise Swartwood

Jared Kagay

CSSE432

05/22/25

Load Balancer



Implementation

Implemented in Python

ChatGPT-2 backend model

Key features:

Round robin

Least Connections

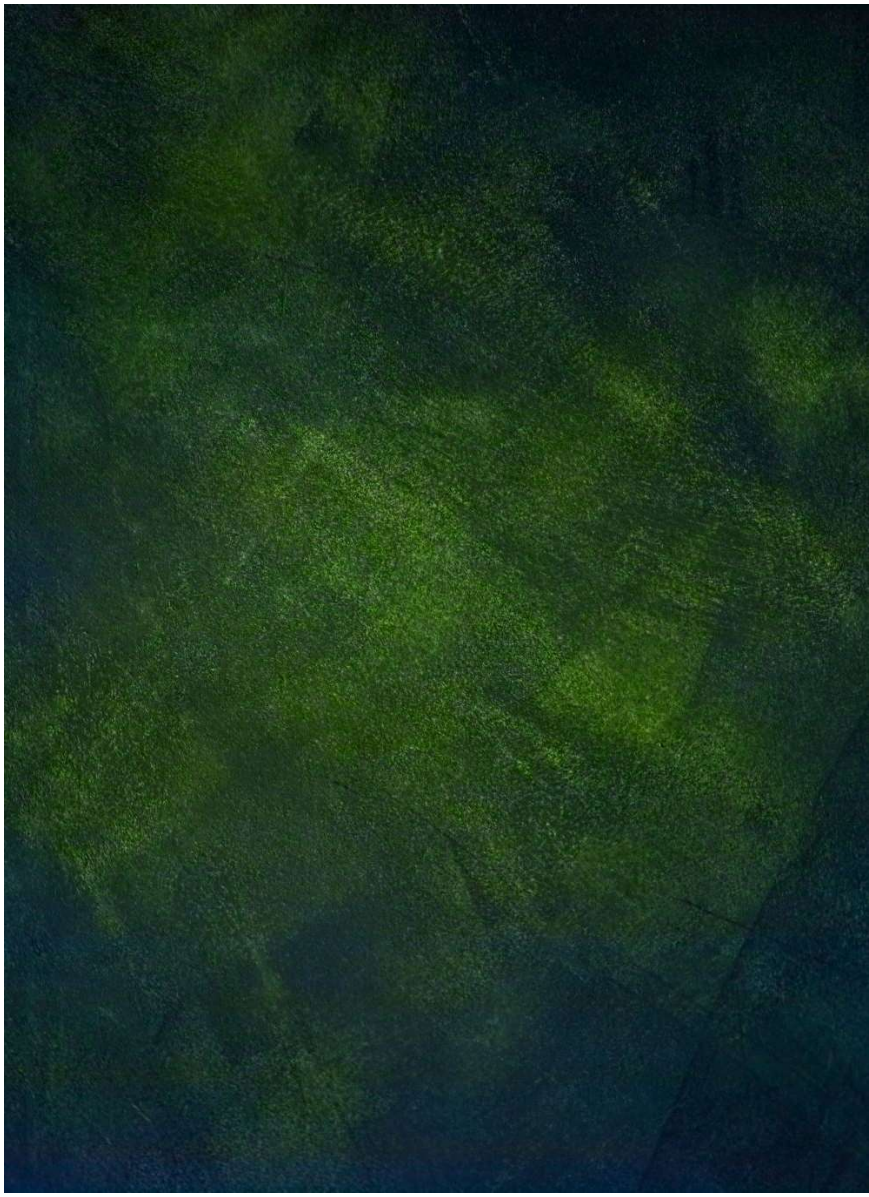
Semantic LRU Caching

Load Balancer Basics

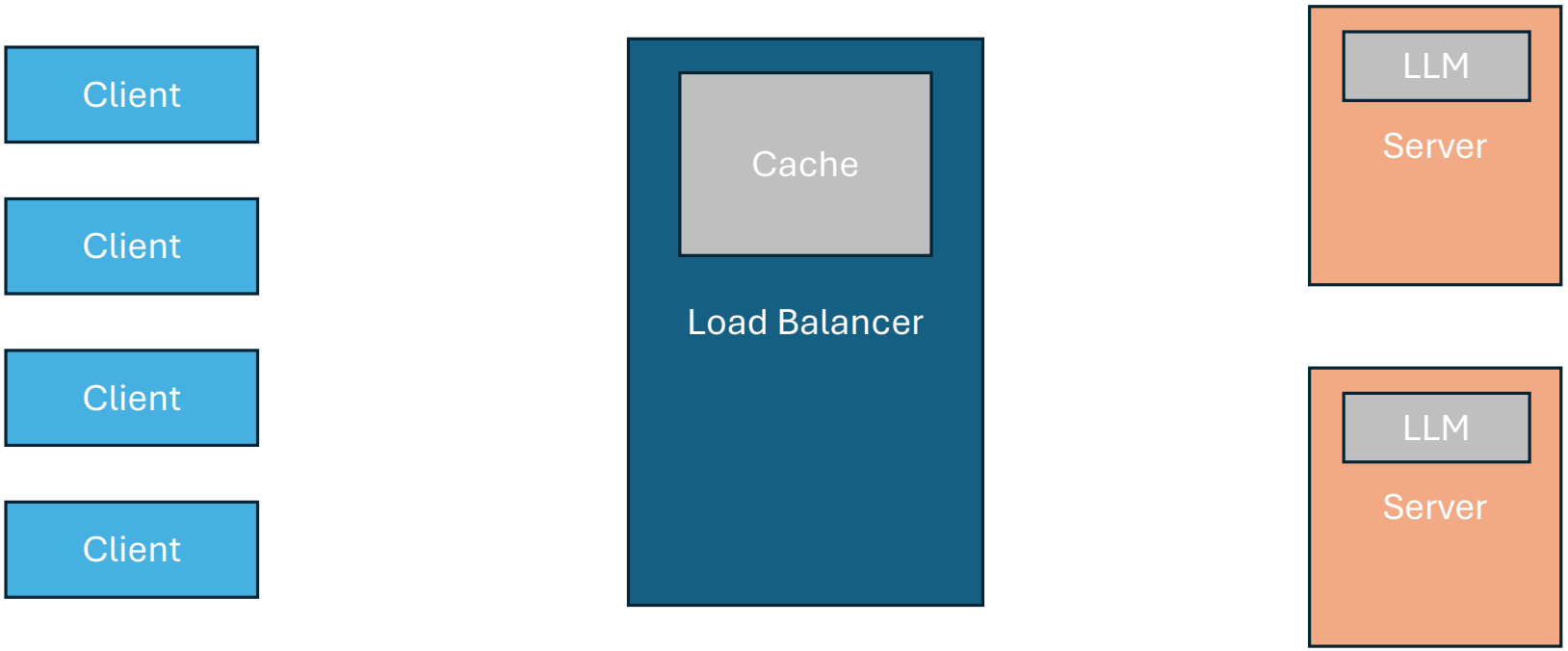
1. Set up simple server backend that is multithreaded to just echo back
2. Set up client to send data
3. Create load balancer server to forward requests
4. Load balancing algorithms implementation

Enhancements

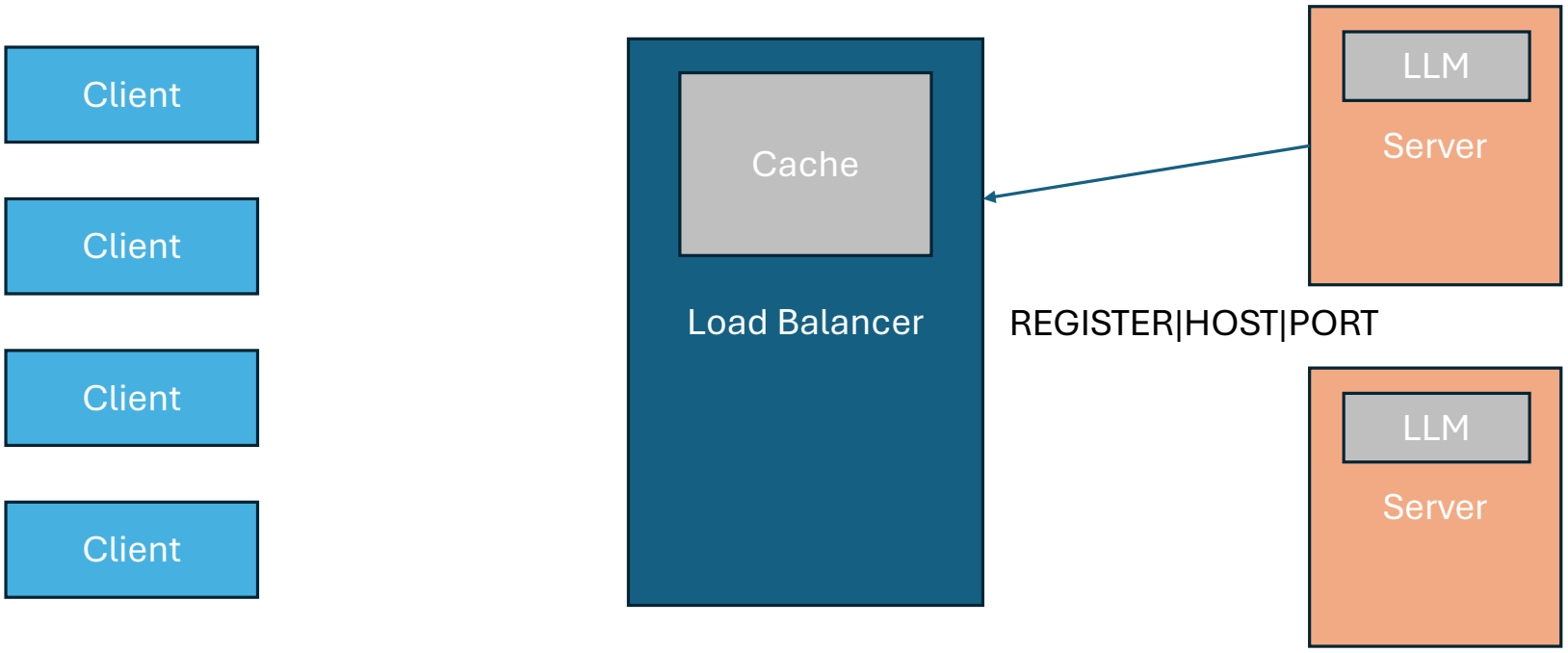
- Async/await implementation
- Server LLM hosting
- LLM Caching (basic, semantic, LRU)
- Dynamic Server Connection
- Health Checks
- Add super simple UI



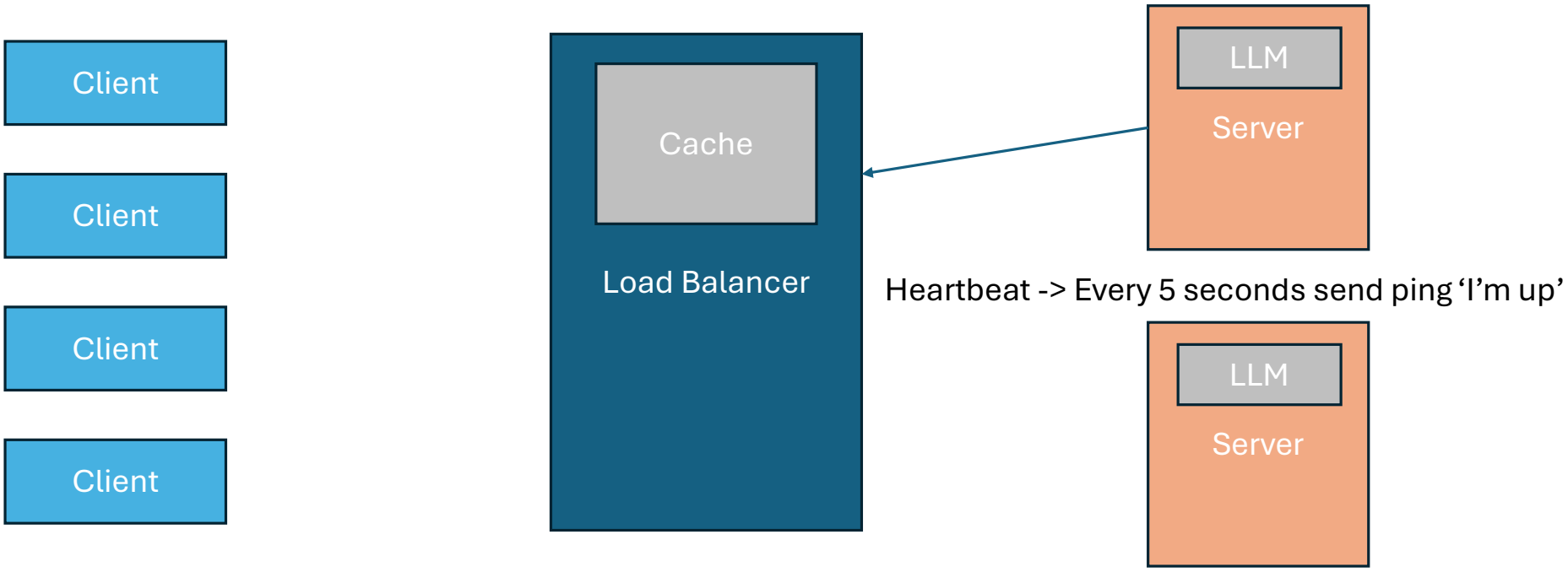
Protocols:


Client
Client
Client
Client
Cache
Load Balancer
LLM
Server
LLM
Server

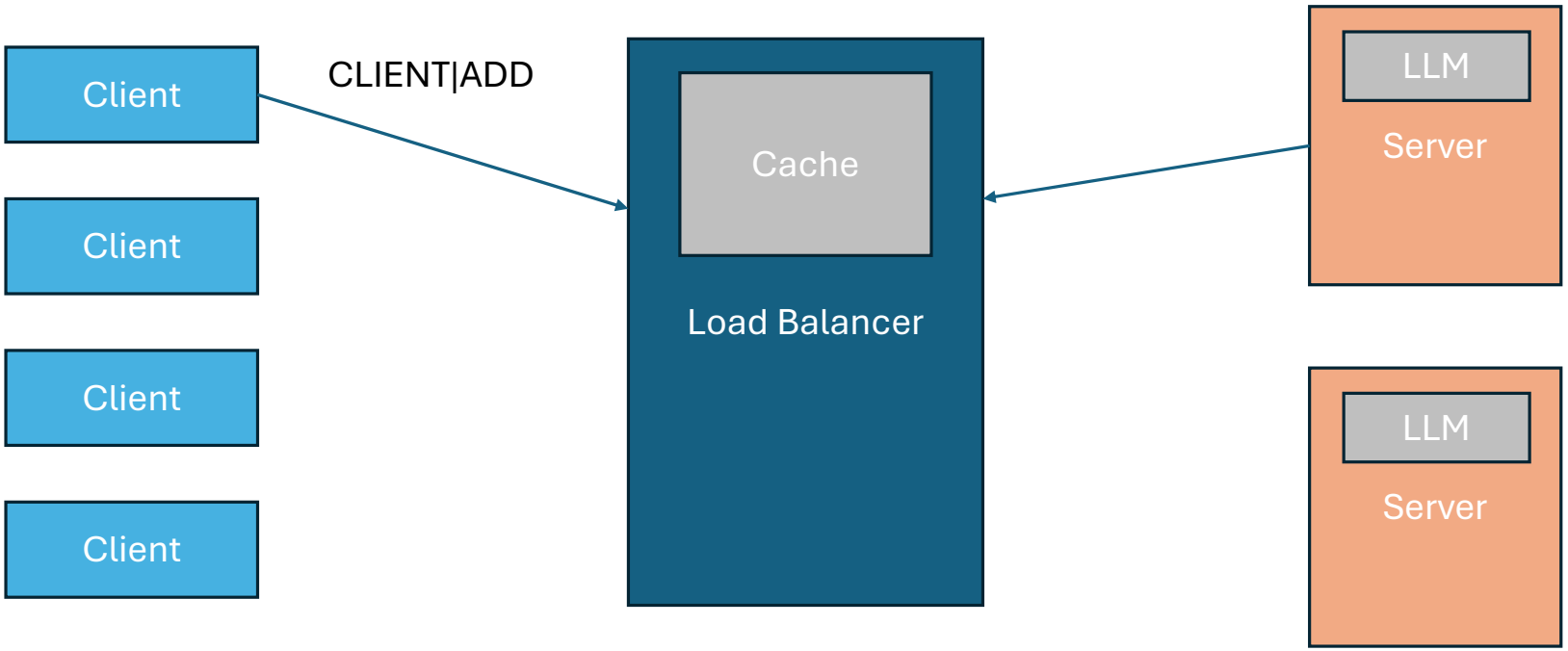
Protocols:


Client
Client
Client
Client
Cache
Load Balancer
LLM
Server
REGISTER|HOST|PORT
LLM
Server

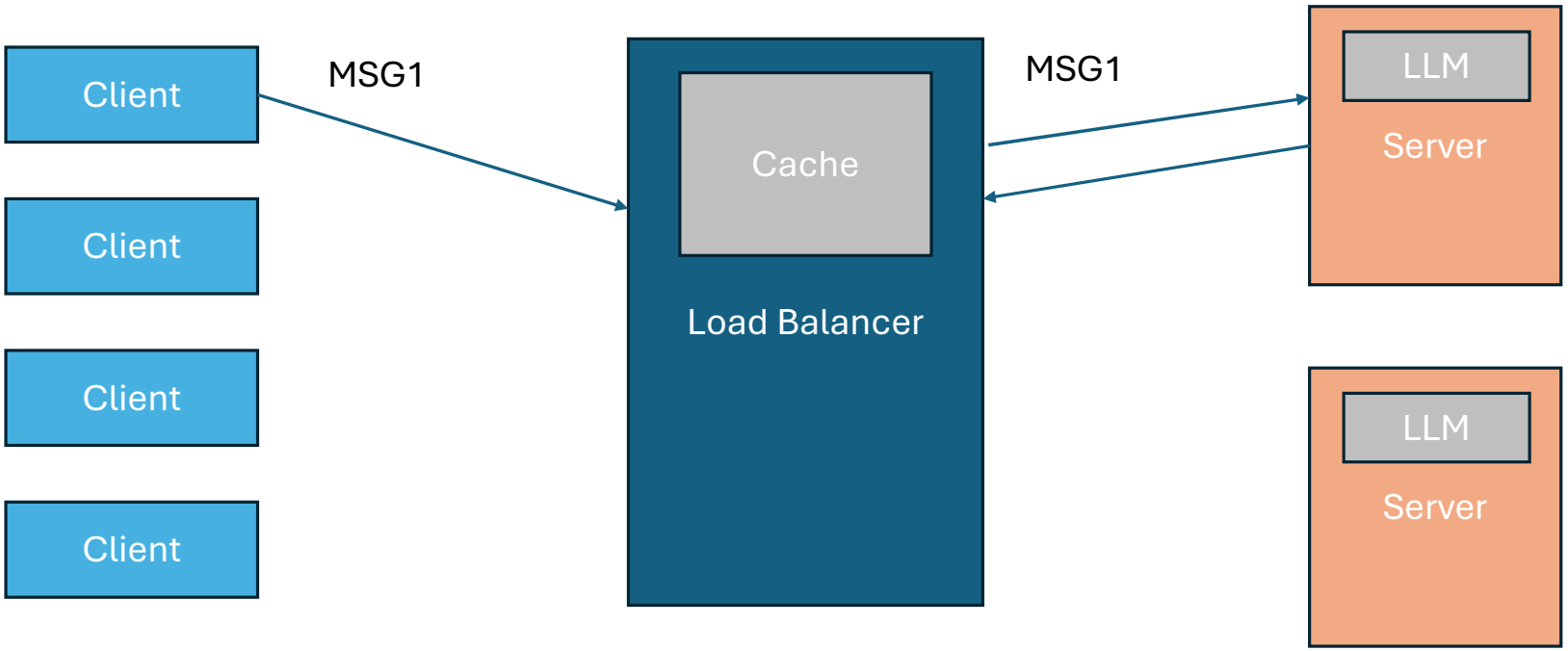
Protocols:


Client
Client
Client
Client
Cache
Load Balancer
LLM
Server
Heartbeat -> Every 5 seconds send ping ‘I’m up’
LLM
Server

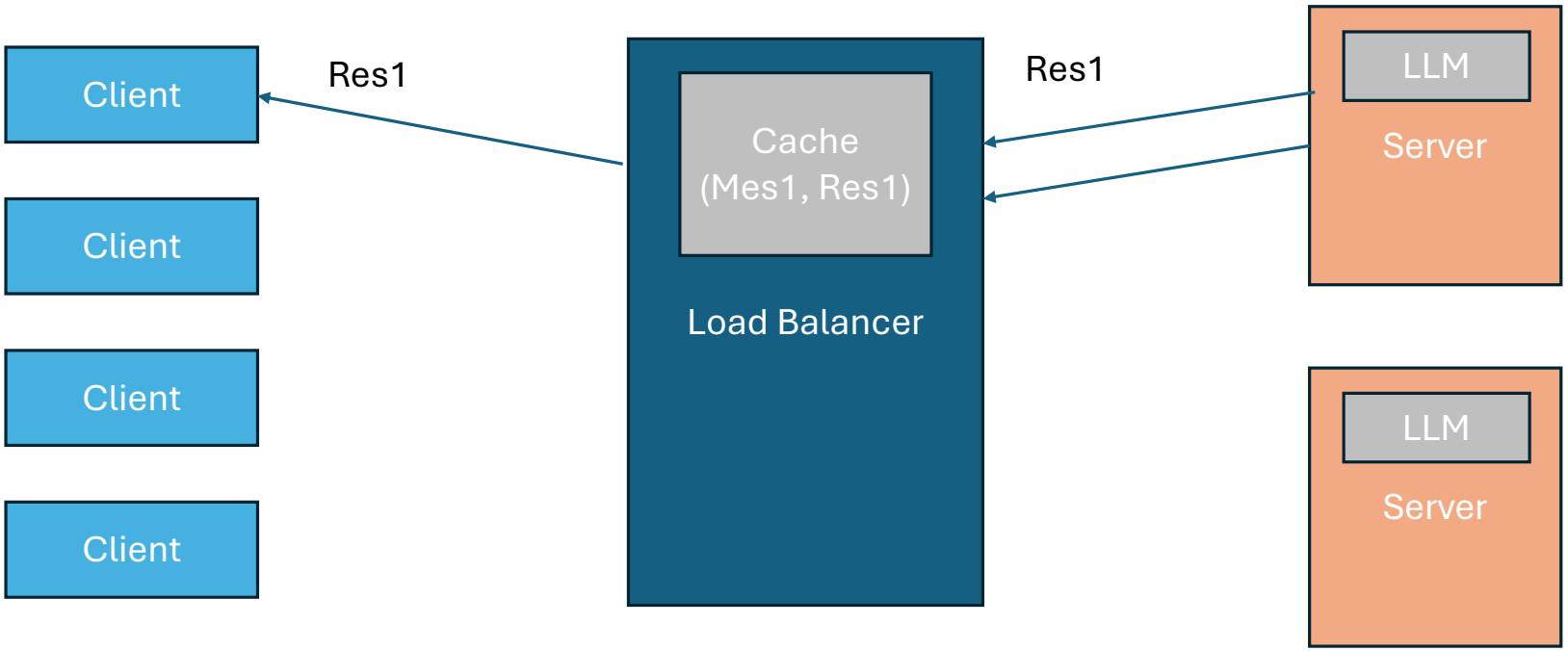
Protocols:


Client
Client
Client
Client
CLIENT|ADD
Cache
Load Balancer
LLM
Server
LLM
Server

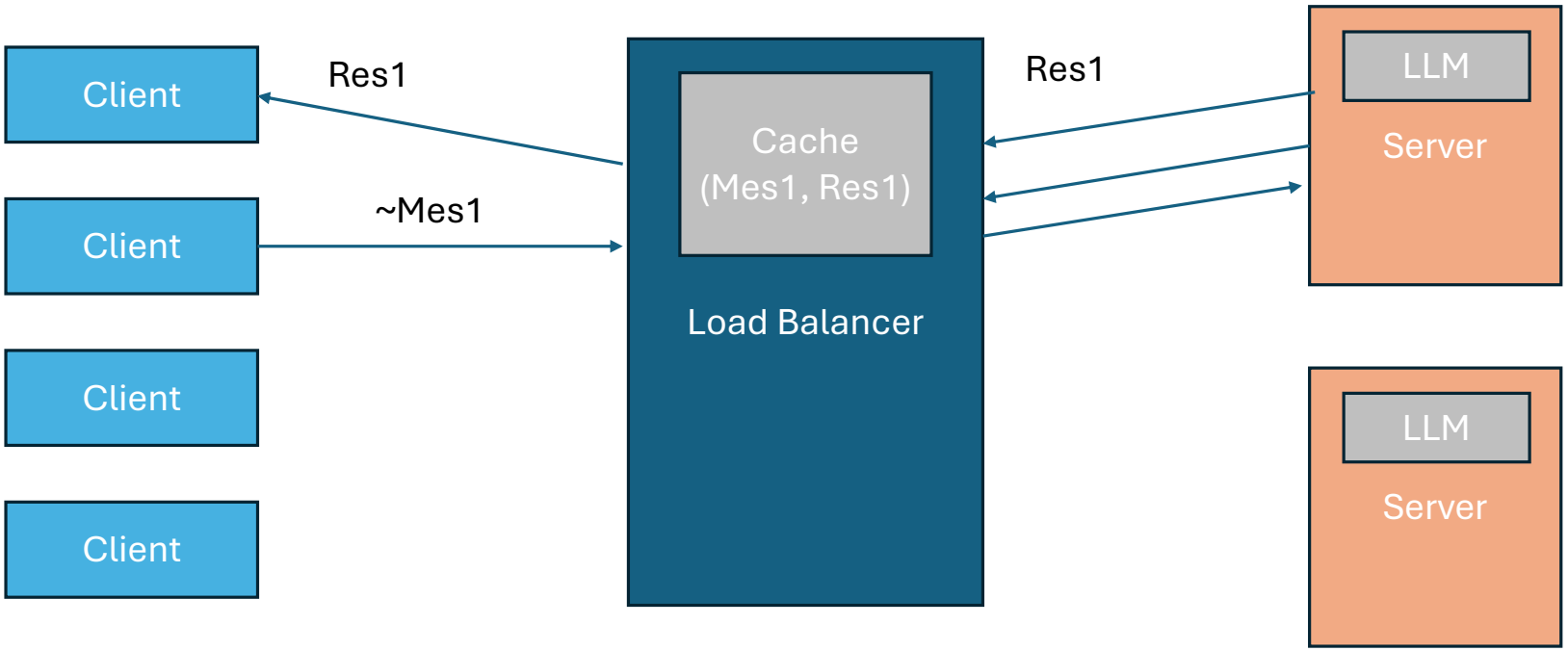
Protocols:


Client
Client
Client
Client
MSG1
Cache
Load Balancer
MSG1
LLM
Server
LLM
Server

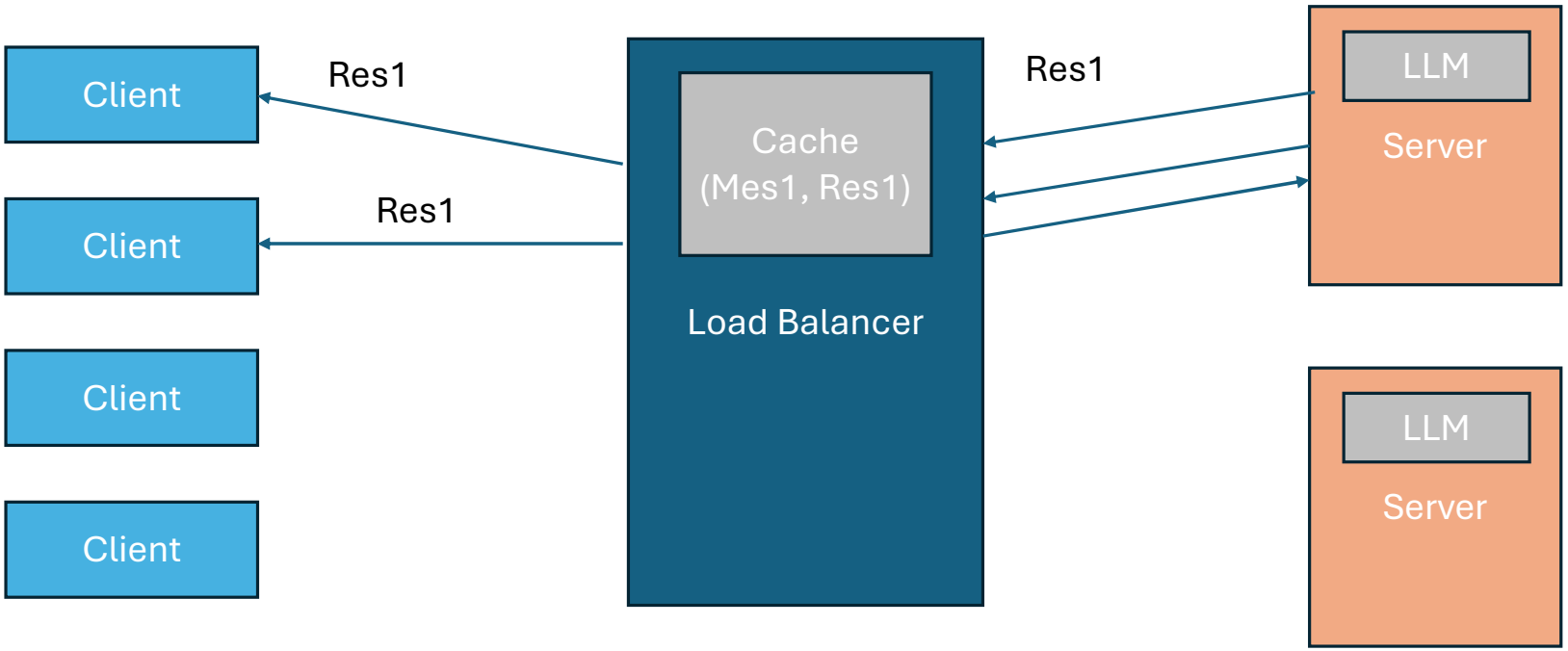
Protocols:


Client
Client
Client
Client
Res1
Cache
(Mes1, Res1)
Load Balancer
Res1
LLM
Server
LLM
Server

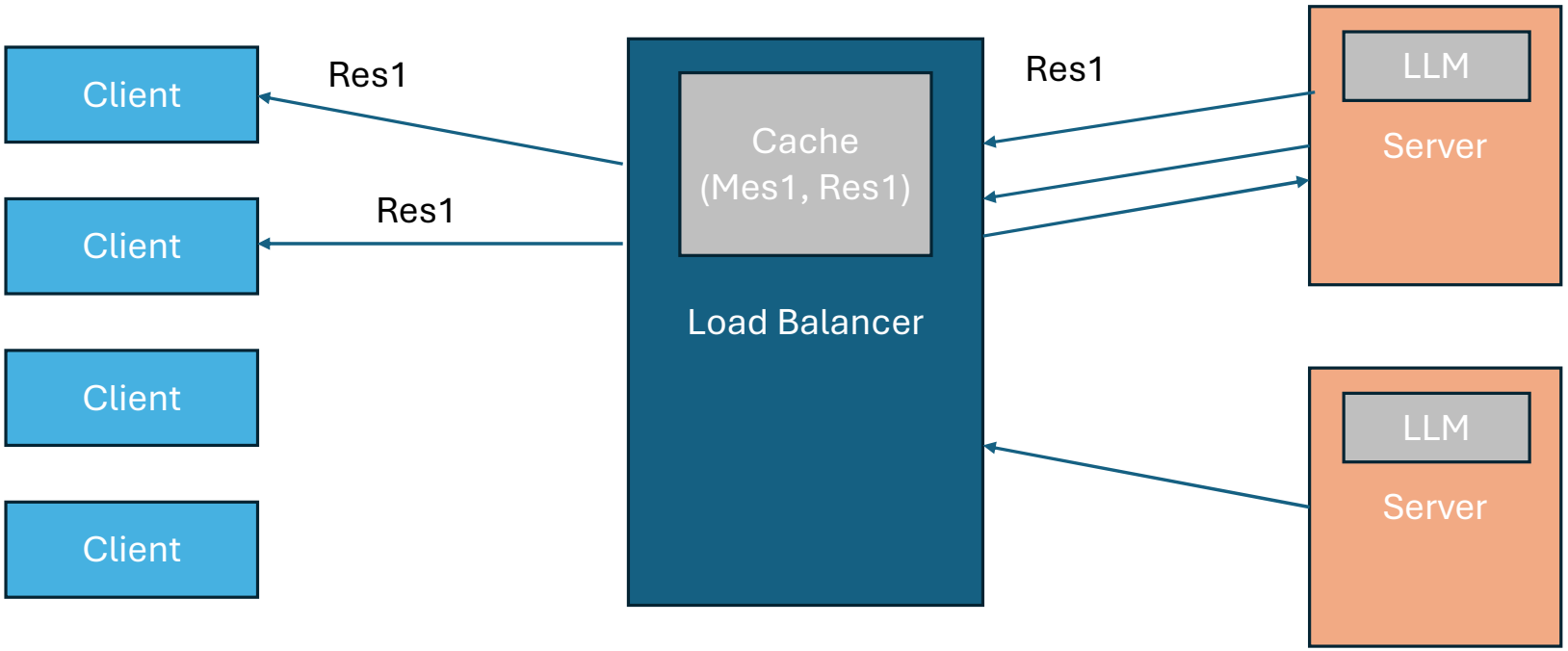
Protocols:


Client
Client
Client
Client
Res1
~Mes1
Cache
(Mes1, Res1)
Load Balancer
Res1
LLM
Server
LLM
Server

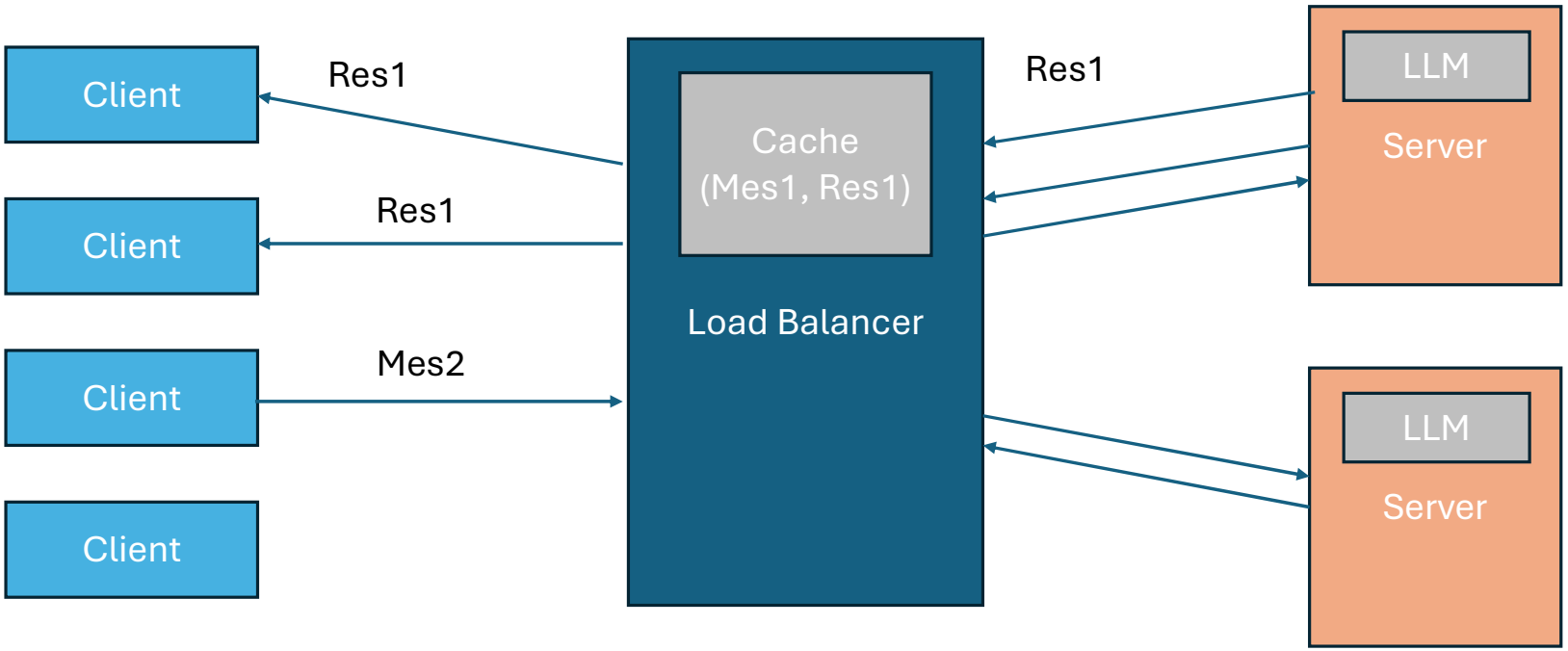
Protocols:

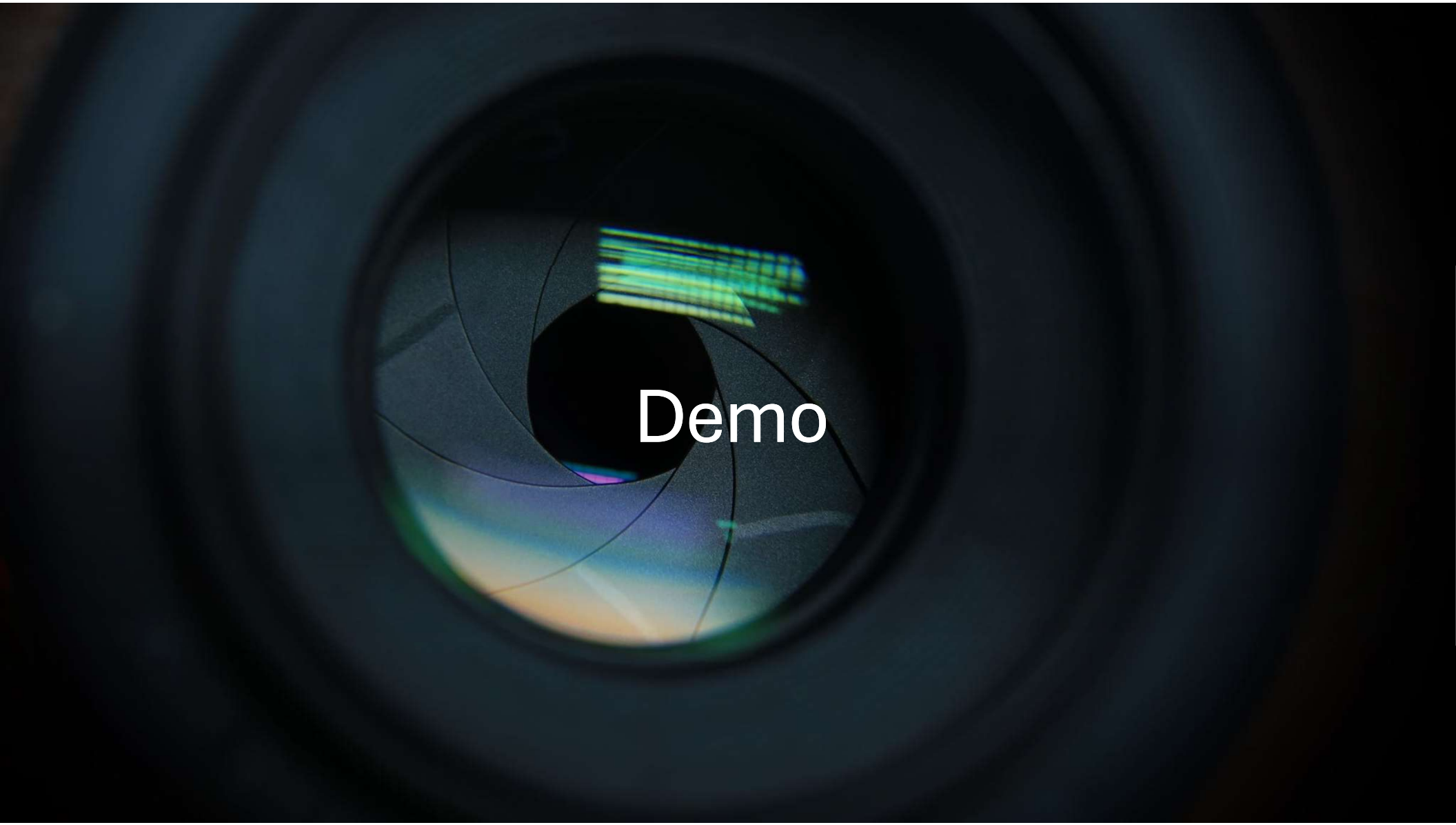

Client
Client
Client
Client
Res1
Res1
Cache
(Mes1, Res1)
Load Balancer
Res1
LLM
Server
LLM
Server

Protocols:


Client
Client
Client
Client
Res1
Res1
Cache
(Mes1, Res1)
Load Balancer
Res1
LLM
Server
LLM
Server

Protocols:


Client
Client
Client
Client
Res1
Res1
Mes2
Cache
(Mes1, Res1)
Load Balancer
Res1
LLM
Server
LLM
Server


Demo

References

Build You Own Load Balancer | Coding Challenges

How to design a load balancer from scratch?

Let's Build! A Simple Load Balancer with Golang