Architecture of some Popular Applications

RACHAMALLA HINDUSREE-2103128
JANAGANI AMRUTHA SAI - 2103118
KASTHALA JOHN AUGUSTEEN-2103119

LINEUP

Architecture of

1. What'sapp
2. Youtube
3. Google meet


vodafone NL
13:04
Instellingen
Info en contact opnemen
WhatsApp
Versie 2.12.13

Have you ever wondered

how WhatsApp works?

Introduction

- WhatsApp is the most common application that almost all of us are using every day. WhatsApp helps us connect people across the world in a friendly and convenient manner.
- WhatsApp was founded by **Brian Acton** and **Jan Koum** in February 2009.
- There are over 2 billion WhatsApp users all over the globe.
- WhatsApp uses **Erlang**, a language that is built for **writing scalable** applications that are designed to **withstand errors.**
- WhatsApp is available in 60 different languages and more than 180 countries all over the world.

Let's see what are the different types of protocols used at different layers in WhatsApp

Protocols

- In WhatsApp application users use HTTP, and WebSockets to send and receive different types of media from the web Server.
- WebSocket Protocol is used to establish the connection between the client and server.
- The transport layer of WhatsApp for the reliable sending of messages and any attachments uses XMPP(Extensible Messaging and Presence Protocol) and TCP(Transmission Control Protocol).
- The network layer uses Internet Protocol(IP) to send messages from one device to another device.

- WhatsApp uses Voice over internet protocol(VoIP) for sending **voice messages.**
- The images, videos, and messages are first sent to the HTTP server and then encrypted using **SRTP**(Secure Real-Time Transport Protocol), **SPL,** and **GRLv3** protocols.

What is Voice over Internet Protocol ?

- VoIP is a technology that allows a person to make voice calls over a broadband Internet connection instead of a regular phone line.
- VoIP is different from common phone services, it has many more features that normal phone service doesn't have.
- VoIP services convert the voice into a **digital signal** from **audio signals** and send the data through the internet connection. If the other client is using a normal or regular signal then the converted signal is converted back to a telephone signal before it reaches the other client.

What is XMPP protocol?

XMPP is a protocol used for exchanging messages and information regarding the presence of the other person. This protocol is mainly used in WhatsApp in which we can send instant messages using it.

X - E**X**tensible, which says that this protocol can be extended according to the new versions.

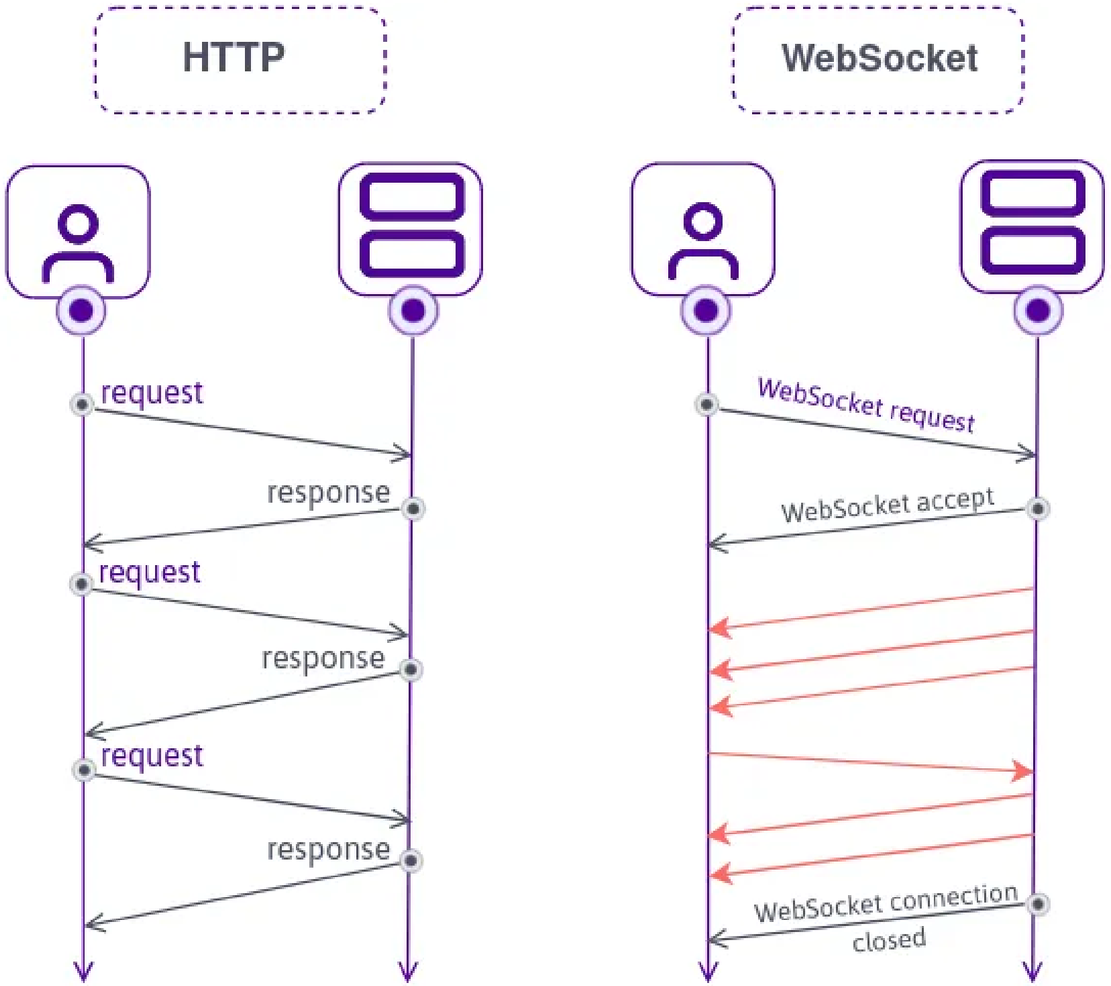
M - **M**essaging, indicates that this protocol is used to send messages in live. It is best among all other protocols.

P - Presence, which tells whether the person on the other side is online or offline.

P - Protocol, XMPP is a protocol that is useful in communication between two clients.

How are the end-points communicating?

WebSocket's



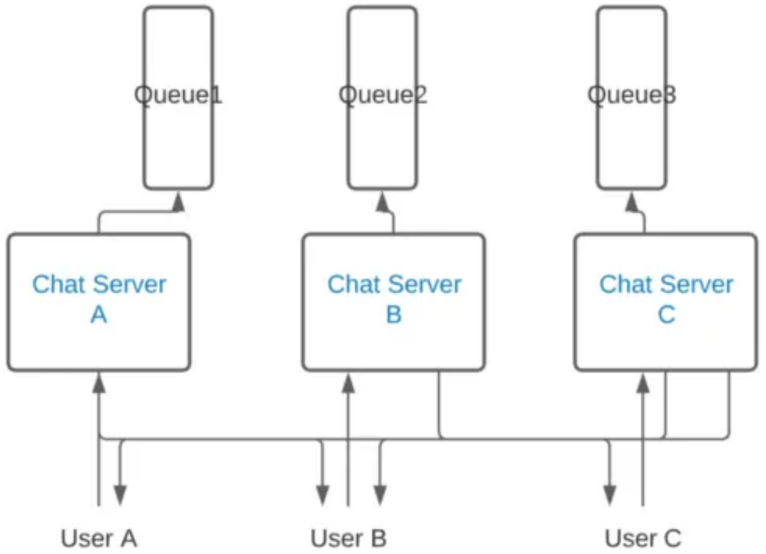
The connection is established between the server and the client using WebSockets. WebSockets are used instead of HTTP because once the response is received connection will be **closed**, this indicates that every time the client wants to send a message he should wait for the full process again. But, this is not the case with WebSockets, the connection in WebSockets is not closed immediately. WebSocket handler will be connected to the client, this handler keeps an open connection with all the online clients.

Chat service (online)

"User-A needs to send a message to User-B, which means that User-A does not need to know User-B's address because there is a server between them. Which provides the connection between them."

Here the chat server creates a new **thread or a process** for User-A and the same goes for User-B if User-B is online.

User Name or ID	PID
A	7229
B	7110



The server finds out the name of the receiver, then **fetches** the data from the chat data storage and finds out the process id (pid) for user B so that the message can be sent to user B.

- The queue here will make sure that there is no **failure** in sending messages and will also handle the **excessive load** of messages from sender

Transient service

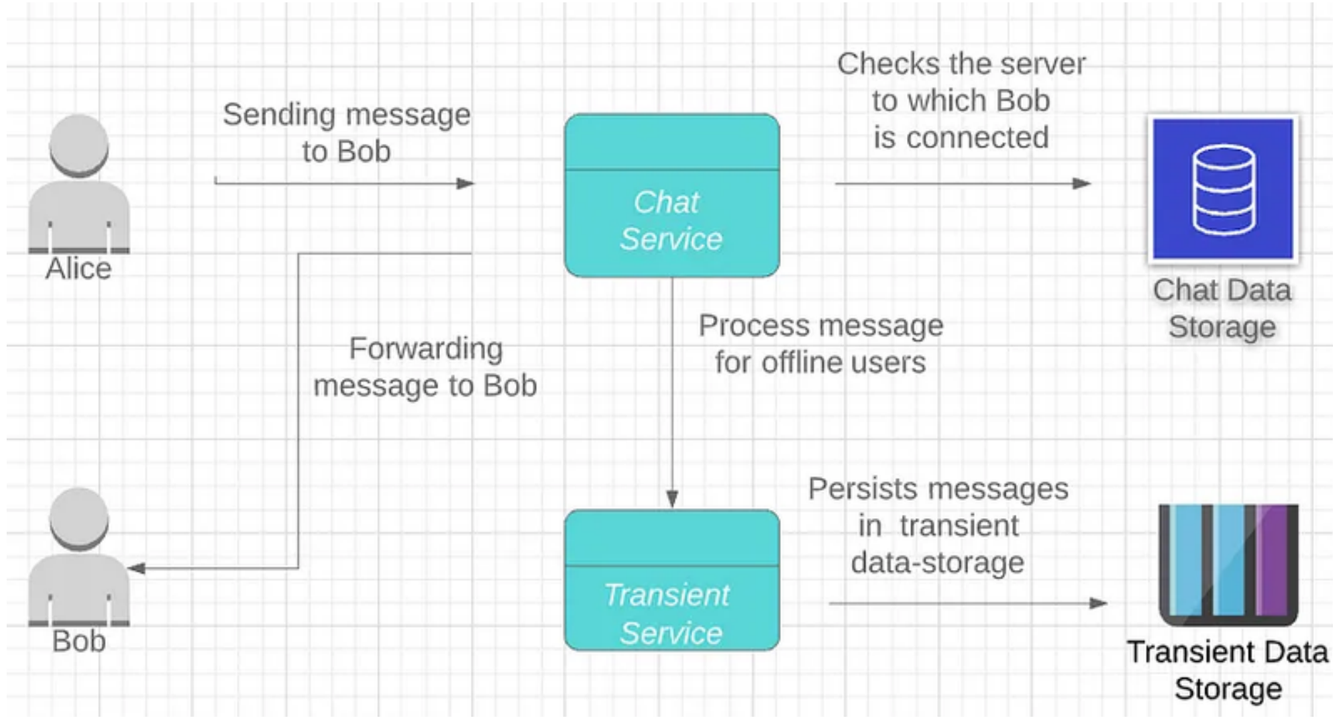
Before the User-A is connected to the server by the load balancer the messages are stored in the **local database**(SQL) of User-A

If User-B is offline then the messages will be stored in the transient database, after the User is online the chat service checks whether there are messages or not, if there are it **collects** them from the database and **sends** them to the respective receiver.

FACTS:

- When we are registered into the WhatsApp application, It creates a map with our mobile number as **phonenumber@s.whatsapp.net** usually known as WhatsApp Username.
- For the password, WhatsApp is using **Mobile's Wi-Fi MAC address**. But previously they have used the IMEI(International Mobile Equipment Identity) number of the mobile.

An Example



A simple example, Lets assume that Alice is sending a message to the Bob through WhatsApp.

First Alice send a message then it is received by chat server (which has components load balancer, messaging server).
Then chat server checks to which server/ ID Bob is connected to.
If Bob is offline (i.e. not connected to any server) then the message goes to transient service and get stored in transient data storage.

How do servers scale to so many users?

WhatsApp uses a programming language called Erlang. Erlang is a super fast programming language that supports many features like Hot Reload, Update on Fly, etc. Erlang has the concept of light weighted thread which makes it capable of handling millions of connections at a time.

- WhatsApp handles about 10 million connections on a single server.
- There are several servers of this kind.
- This is the reason Erlang is an ideal choice for WhatsApp, which makes the server scale to so many users.


YouTube

What is YouTube?

YouTube is a **video-sharing service** where users can watch, like, share, comment, and upload their own videos. The video service can be accessed on PCs, laptops, tablets, and via mobile phones.

Main Functions of You Tube :

- Users can search for and watch videos
- Like/Comment/share other YouTube videos
- Users can subscribe/follow other YouTube channels and users
- Create a personal YouTube channel
- Upload videos to your channel
- Create playlists to organize videos and group videos together

PROTOCOLS USED BY YOU TUBE AT DIFFERENT LAYERS

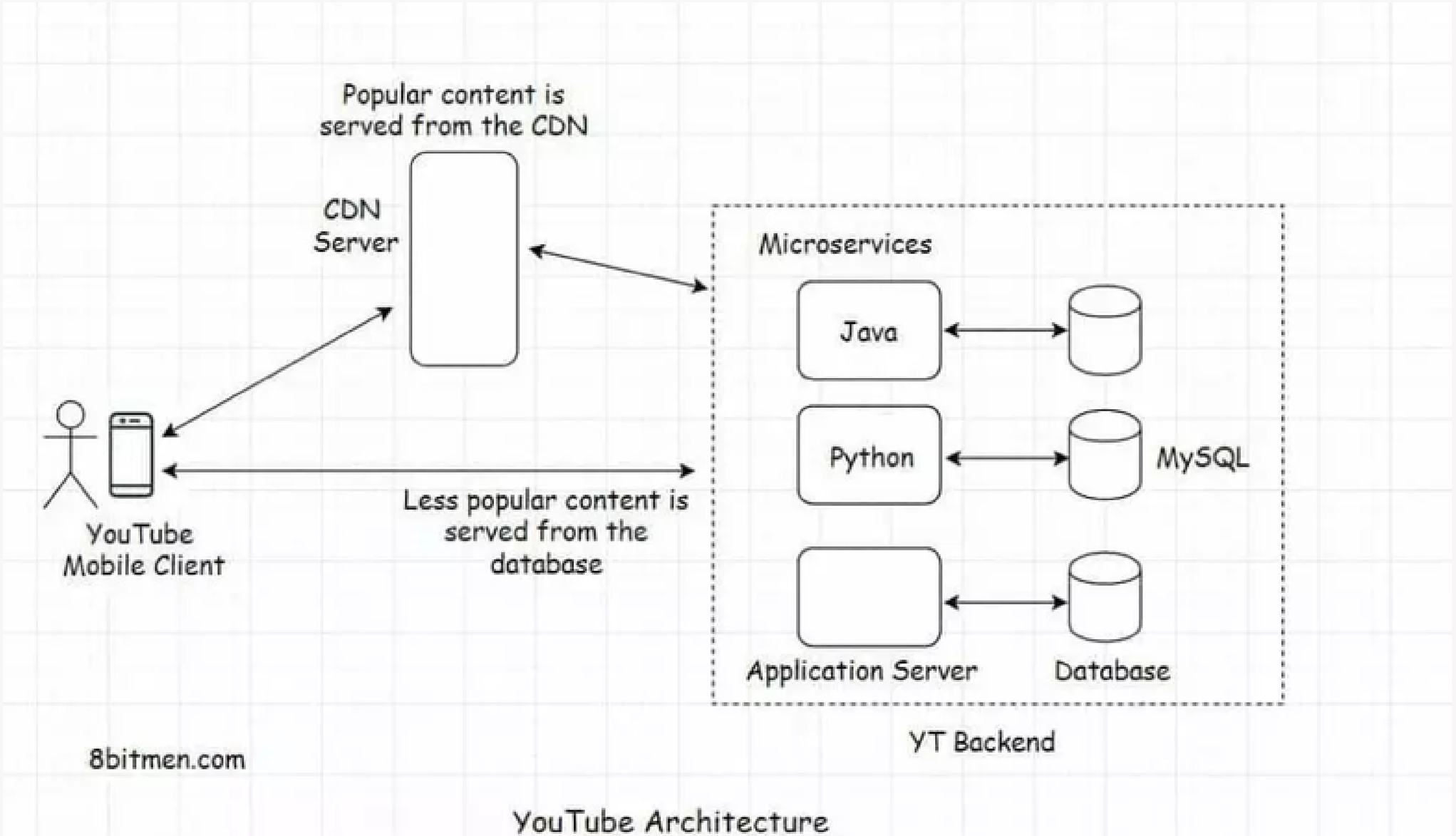
PROTOCOLS

- YouTube uses HTTP protocol at the Application Layer.
- QUIC(Quick UDP Internet Connection) and TCP(Transmission Control Protocol) protocols are used at the Transport layer of YouTube.
- RTMP is a widely-used protocol for **video streaming** in YouTube application.
- The Network Layer uses IP (Internet Protocol) protocol to transfer the data between the user and server.

What is **RTMP** protocol?

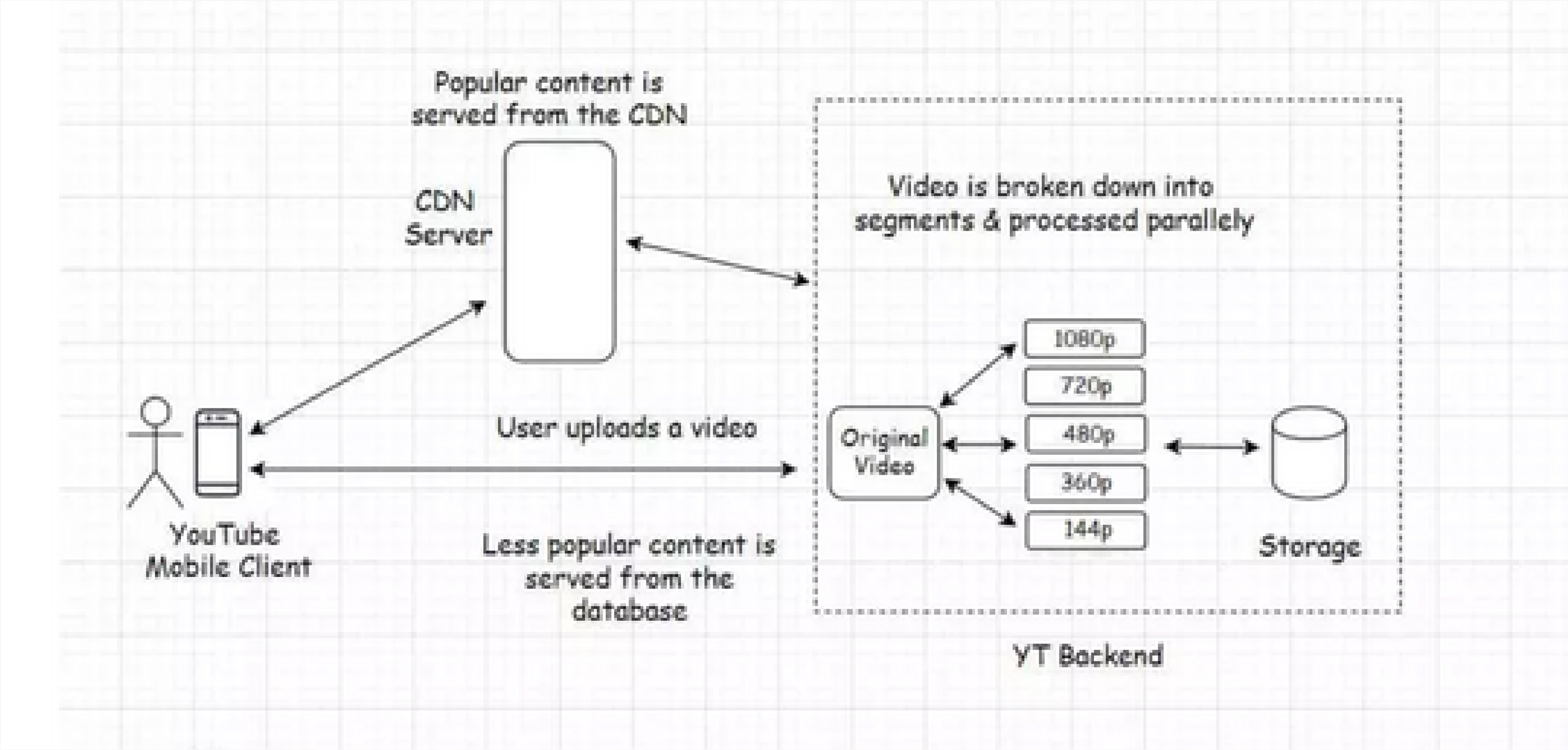
- Real-Time Messaging Protocol (RTMP) is a widely-used protocol for **video streaming** that YouTube Live has accepted since the service began.
- RTMPS is a secure extension to RTMP. RTMPS **benefits** both content creators and viewers by preventing **man-in-the-middle** attacks on the ingestion side of live streams.
- This ensures that all of a creator's live streaming data — including video, audio, and control signals — is **securely transmitted** to YouTube's servers, **protecting** it from tampering or interception in transit.

End-point Communication in YouTube



YouTube Architecture

YouTube's Video Delivery architecture



How does YouTube serve High-Quality video's ?

- A key element in the process of delivery of high-quality videos on YouTube is **video transcoding.**
- When a video is uploaded on YouTube it is first transcoded from its original format to a temporary intermediate format to facilitate the conversion of the content into different resolutions and formats.
- This enables YouTube to stream videos in different resolutions such as 144p, 240p, 360p, 480p, 720p, 1080p, and 4K.
- Delivery of content based on the network bandwidth and the device type of the end-user is known as **adaptive streaming**.
- Without adaptive streaming, there is no way users with low bandwidth networks can watch a high-resolution video.

- To overcome these kinds of issues YouTube uses Dynamic Adaptive Streaming over HTTP protocol.
- **Dynamic Adaptive Streaming** over HTTP protocol is used for the streaming of videos.
- It’s the adaptive bitrate streaming technique that enables high-quality streaming of videos over the web from conventional HTTP web servers.
- Via this technique, the content is made available to the viewer at different bit rates.
- YouTube client automatically adapts the video rendering as per the **internet connection** speed of the viewer thus cutting down the buffering as much as possible.

Have you ever thought how YouTube stores so much data?

- The videos are stored in hard drives in warehouse-scale Google data centers. The data is managed by the Google File System and BigTable.

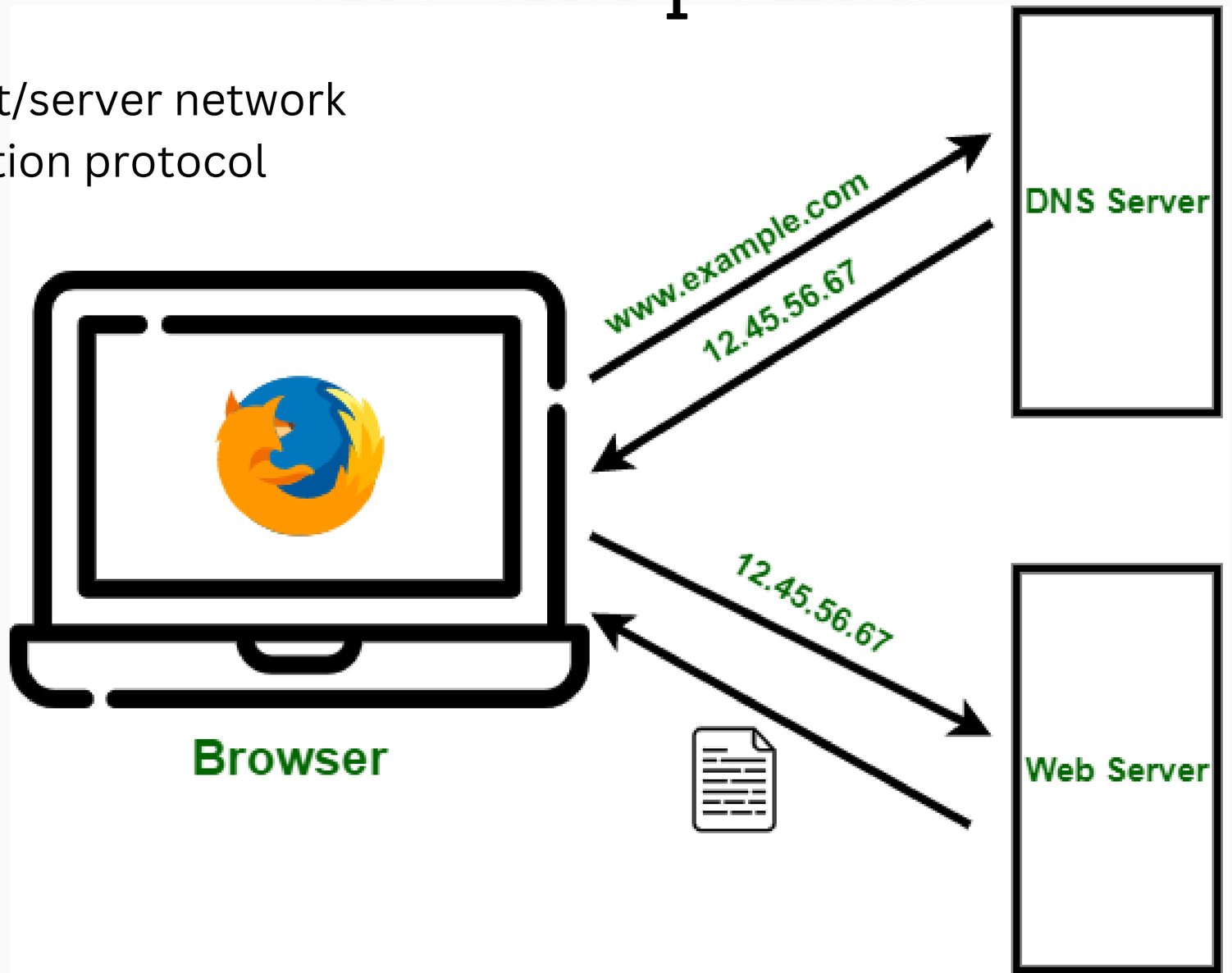
GFS Google File System is a distributed file system developed by Google to **manage large-scale data** in a distributed environment.

BigTable is a low latency distributed data storage system built on Google File System to deal with **petabyte-scale data** spread over thousands of machines. It's used by over 60 Google products.

https://scaleyourapp.com/

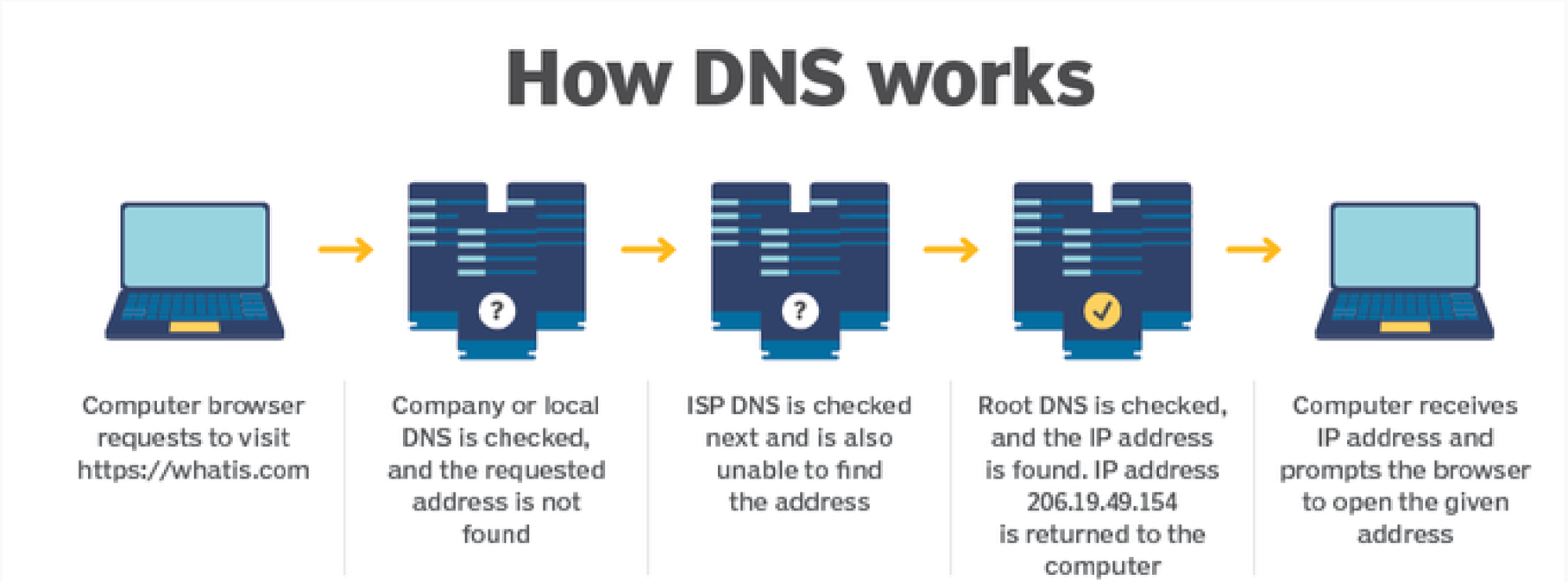
How does the server know the IP address of the end points

DNS is the client/server network communication protocol



What is Domain Name Server

- DNS is a service that translates the domain name into IP addresses
- This allows the users of networks to utilize user-friendly names when looking for other hosts instead of remembering the IP addresses.
- A query from the browser goes to the 'recursive server'(also called recursive resolver) If the recursive resolver has the address, it will return the address to the user, and the webpage will load.



Priority order to check the IP DNS -- root name servers, top-level domain (TLD) name servers and authoritative name servers.

https://www.techtarget.com/searchnetworking/definition/domain-name-system

What happen if there is a server breakdown?

The user data is backed up in the data centers located in different geographical zones across the world.

Having several data centers across the world also helped YouTube reduce the latency of the system as user requests were routed to the nearest data center as opposed to being routed to the origin server located in a different continent.

How does YouTube application scale to so many users?

Load Balancing: YouTube uses load balancing to distribute traffic across its servers to prevent any server from becoming overwhelmed and ensures that the requests are evenly distributed across the network.

Horizontal Scaling: YouTube scales horizontally by adding more servers to its infrastructure to handle increased traffic. This allows it to ensure high availability

Google Meet

Google Meet

Google Meet (formerly known as Hangouts Meet) is a **video conferencing** service that enables you to join virtual meetings via **audio**, **video**, **chat**, and **screen sharing** with up to 100 people with no time limits and developed by Google.

- The architecture of Google Meet is based on a combination of **Google's cloud infrastructure** and **Web Real-Time Communication (WebRTC)** technology.
- The Google Meet architecture also includes a range of features to ensure **security** and **privacy**, such as end-to-end encryption for video meetings, and strong authentication and access controls for meeting hosts.
- The architecture of Google Meet is designed to provide a **scalable**, **secure**, and user-friendly platform for **remote communication** and collaboration.

Features

1. **Two-way and multi-way** audio and video calls with a resolution of up to 720p
2. **Call encryption** between all users
3. Screen-sharing, browser tab sharing
4. Hosts are able to deny a user's entry, remove a user from, and control microphone and video access in a call.
5. Shared reactions, polls, voting, Q&A.
6. Integration with the Google ecosystem, e.g. **live streaming**

Let's see what the different types of protocols used at different layers in

- **GOOGLE MEET**

Here's a breakdown of the protocols used at each layer in Google Meet's architecture:

APPLICATION LAYER

- Google Meet uses a web application built with JavaScript, HTML, and CSS.
- The web application communicates with Google's servers over HTTPS (HTTP Secure) using the **WebSocket protocol** to establish a persistent two-way communication channel between the browser and the server.





TRANSPORT LAYER

- The transport layer of Google Meet's architecture uses the **Transmission Control Protocol (TCP)** to establish a reliable connection between the client and the server. This ensures that all packets are delivered in order without any loss

NETWORK LAYER

THE NETWORK LAYER USES THE **INTERNET PROTOCOL (IP)** TO ROUTE PACKETS BETWEEN DIFFERENT NETWORKS. GOOGLE MEET USES A GLOBAL NETWORK OF DATA CENTERS TO ENSURE **LOW LATENCY** AND **HIGH AVAILABILITY** FOR ALL USERS.

Data Link Layer

The data link layer provides a reliable connection between two adjacent nodes in a network. Google Meet uses the User Datagram Protocol (UDP) to send real-time video and audio data between clients and servers.

SOME OTHER PROTOCOLS

(RTP) and (UDP)

RTP is a protocol used for transmitting real-time data, such as audio and video, over a network. UDP is a transport protocol that does not require the establishment of a dedicated connection before data transmission. Google Meet uses RTP and UDP to transmit audio and video data between participants in real time.

Secure Real-time Transport Protocol (SRTP)

SRTP is an extension of RTP that provides encryption, authentication, and integrity checking to secure data transmissions. Google Meet uses SRTP to ensure that audio and video data is transmitted securely between participants.

Hypertext Transfer Protocol Secure (HTTPS)

HTTPS is a protocol used for secure communication over a network. Google Meet uses HTTPS to encrypt the transmission of data, including chat messages, between participants

Session Initiation Protocol (SIP)

SIP is a signaling protocol used to initiate and manage communication sessions. Google Meet uses SIP to establish and maintain connections between participants.

(WebRTC)
Web Real-Time Communication

WebRTC is an open-source project that enables real-time communication between browsers and mobile applications. Google Meet uses WebRTC to facilitate audio and video data transmission, as well as screen sharing and other collaborative features

Google Cloud Platform:

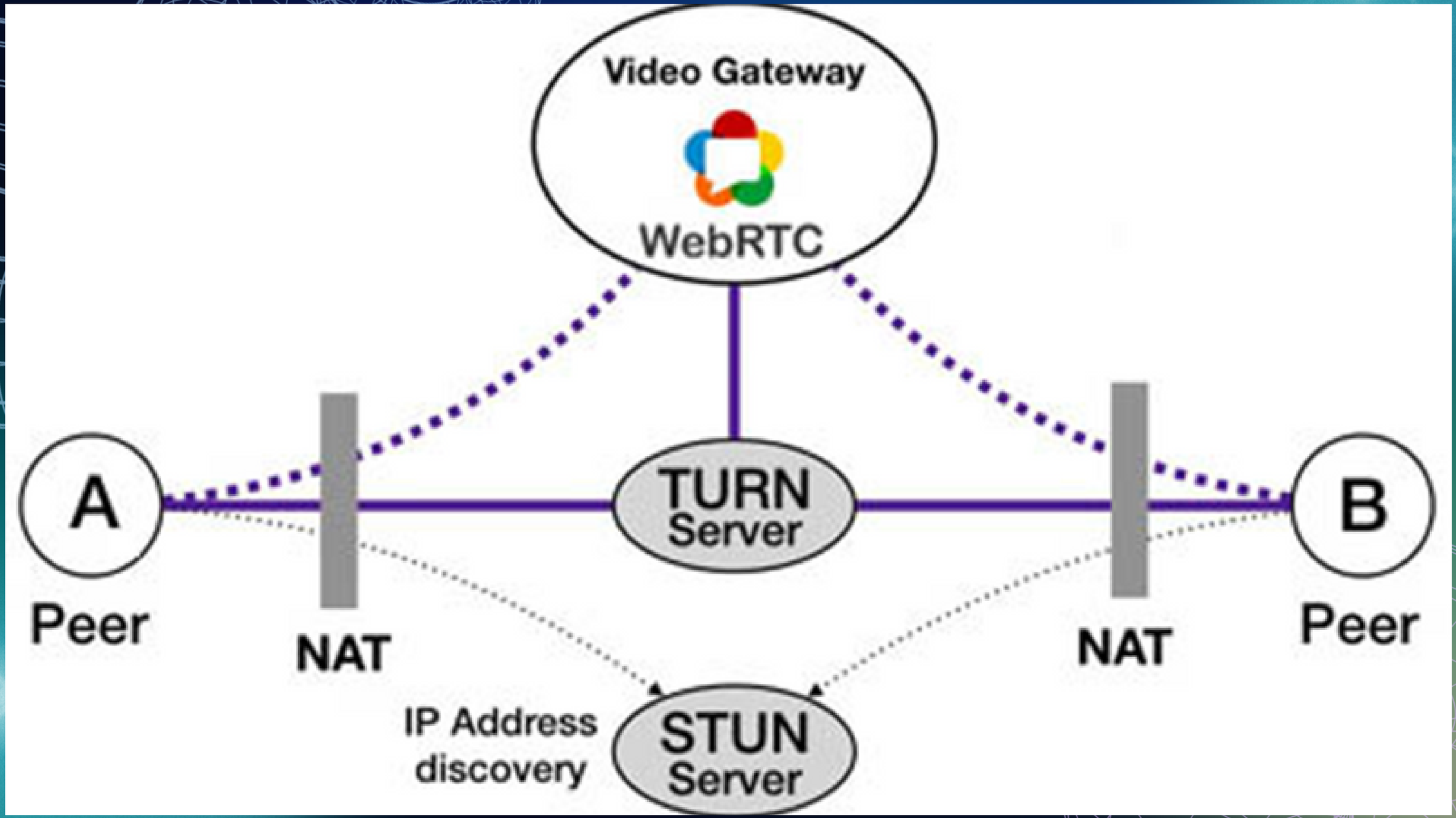
- the Google Cloud Platform to provide reliable and scalable infrastructure.
- This platform offers a highly available and distributed computing infrastructure, which ensures that the service is always available for users, regardless of their location.

Encryption and Security:

- Google Meet uses end-to-end encryption to protect the privacy and security of user data.
- All video and audio data is encrypted before transmission and decrypted only at the recipient's end.

How does each end-point know the address of the others

In terms of end-points, each participant in a Google Meet call communicates with Google's servers, which act as intermediaries. The servers handle all the data routing, media processing, and synchronization required to ensure a smooth video conferencing experience. The STUN server helps the endpoints to know the IP address of each other. For some restrictive networks, the TURN server is used.


Video Gateway
WebRTC
TURN
Server
A
Peer
NAT
NAT
B
Peer
IP Address
discovery
STUN
Server

how do these application servers scale to so many users

To scale to so many users, Google Meet uses a distributed architecture with a global network of data centers. Each data center is equipped with its own set of servers and networking infrastructure, allowing the platform to handle a large volume of traffic and ensure low latency for all users.

References:

- https://stackoverflow.com/questions/34983909/how-whatsapp-works
- https://interviewnoodle.com/whatsapp-system-architecture-8df0250d572f
- Wikipedia
- https://gohulan.medium.com/anatomy-of-whatsapp-messenger-60c876be4b14#:~:text=WhatsApp%20is%20a%20Instant%20messenger%20allows%20to%20sending%20of%20texts,IP%2C%20XMPP%20%2C%20TCP%20protocols.
- https://www.geeksforgeeks.org/xmpp-protocol/
- https://www.techtarget.com/searchunifiedcommunications/definition/VoIP
- https://medium.com/codingurukul/whatsapp-engineering-inside-2-bdd1ec354748#:~:text=%2D%3E%20Programming%20Language%3A%20Erlang,of%20connections%20at%20a%20time.