Design Zoom / Video Confrencing

Functional Requirement

1. Peer to Peer communication
2. Group Call
3. Audio / Video
4. Screen Share
5. Recording

Non Function Requirement

1. It should support vareity of devices
2. It shouble be consistent across devices
3. Low latency
4. Super Fast

Network Layer Protocol

1. TCP
2. UDP

TCP

1. Transimission control protocol
2. Connesction is established using 3-way handsake
3. Ensure all data receive in order and without duplication
4. Flow control and cnogestion control
5. Eg: Web browsering, Filter transfer etc

UDP

1. user Datagram protocol
2. No handsake just deliver the data
3. No gurantee on delivery order
4. Missing of packet possible
5. Fast Delivery
6. Eg: Video confrencing etc

NAT (Network adddress translation)

Device - 1

10.x.x.x

Google Server

4.4.4.4

1. Why can't we directly assign some public IP to device one

Restrcrition

IPV4

0.0.0.0

255:255:255:255

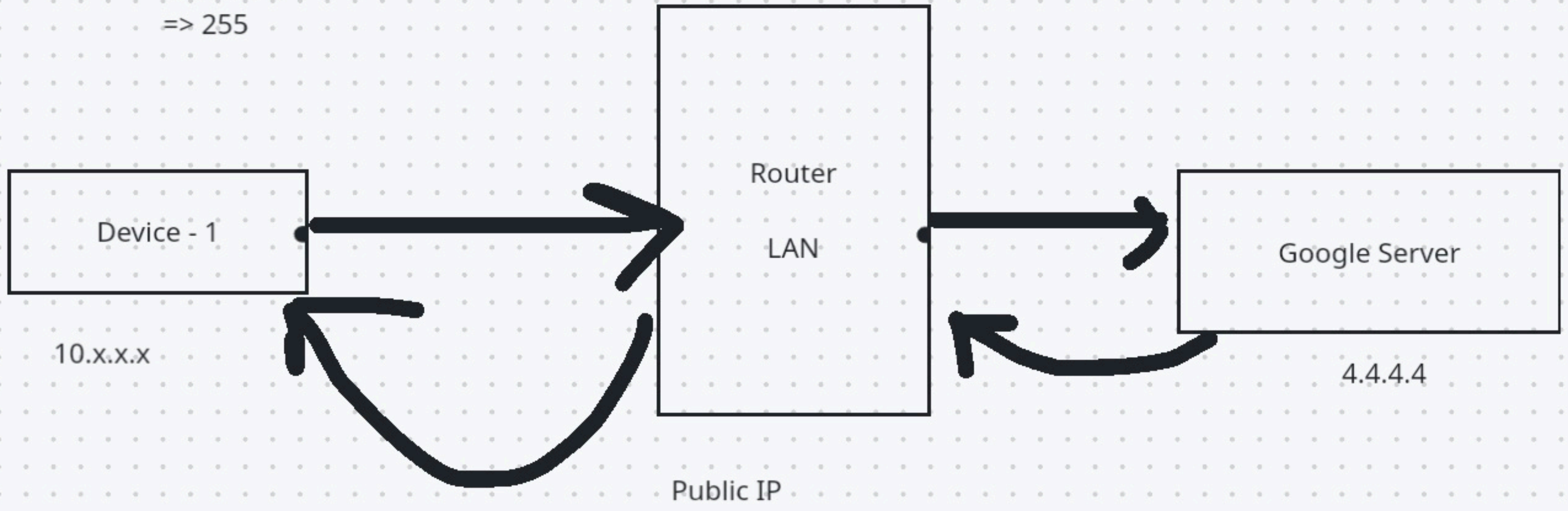
256*256*256*256 => 2Billion

IPV4 => 32 bits
8 Bits

XXXXXXXX => (0, 1)

00000000 => 0

11111111
=> 255



Device - 1

10.x.x.x

Router

LAN

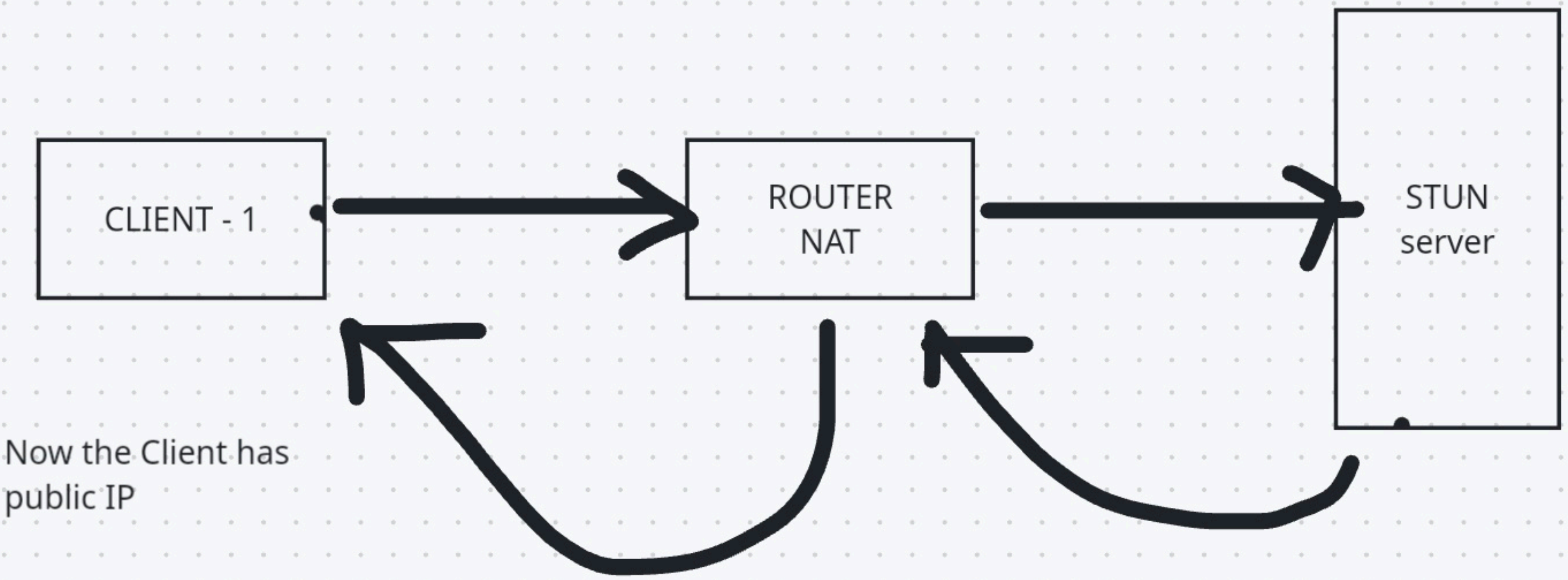
Public IP

Google Server

4.4.4.4

Why can't we directly get the public IP of user ?

For this we need STUN server

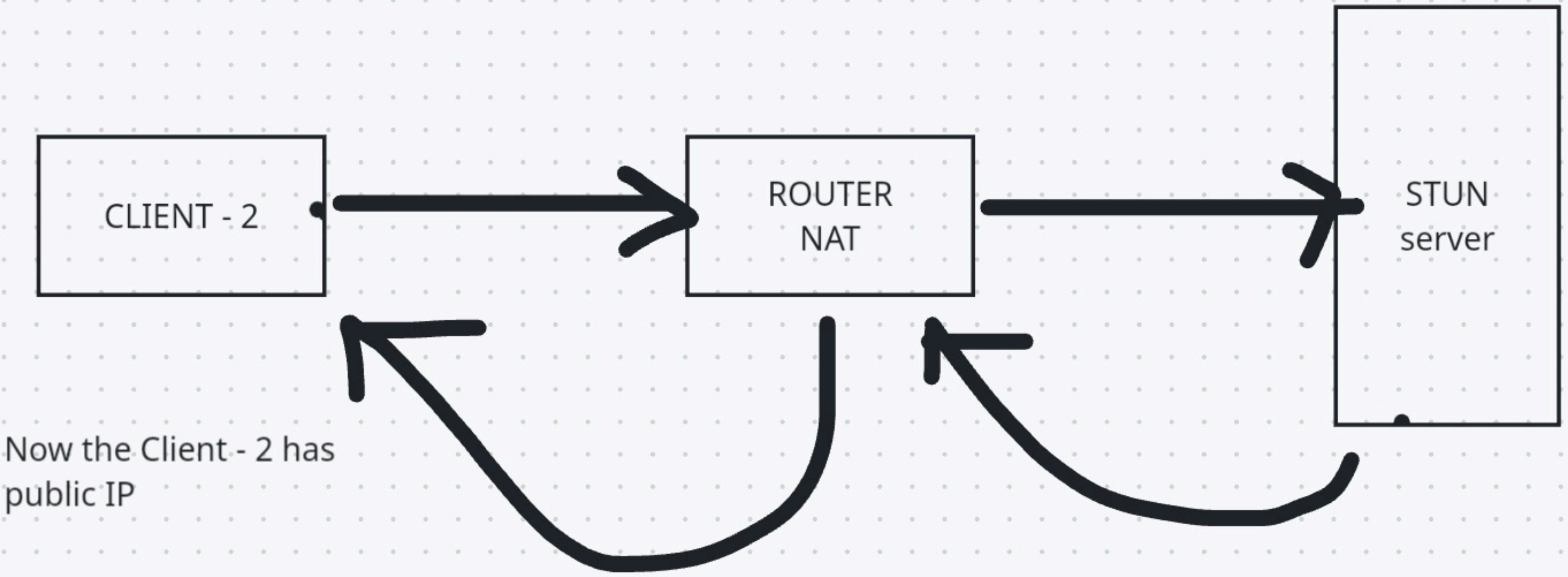


CLIENT - 1

ROUTER NAT

STUN server

Now the Client has public IP



CLIENT - 2

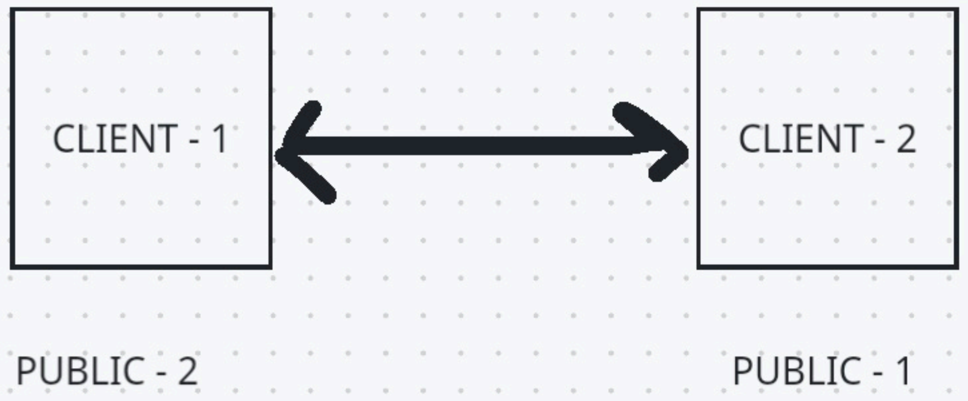
ROUTER NAT

STUN server

Now the Client - 2 has public IP

So, now both the user need to share the public IP and others details
like, Network , codec related details , session related details, and media capabilities

which is also called as SDP(Session description protocol)

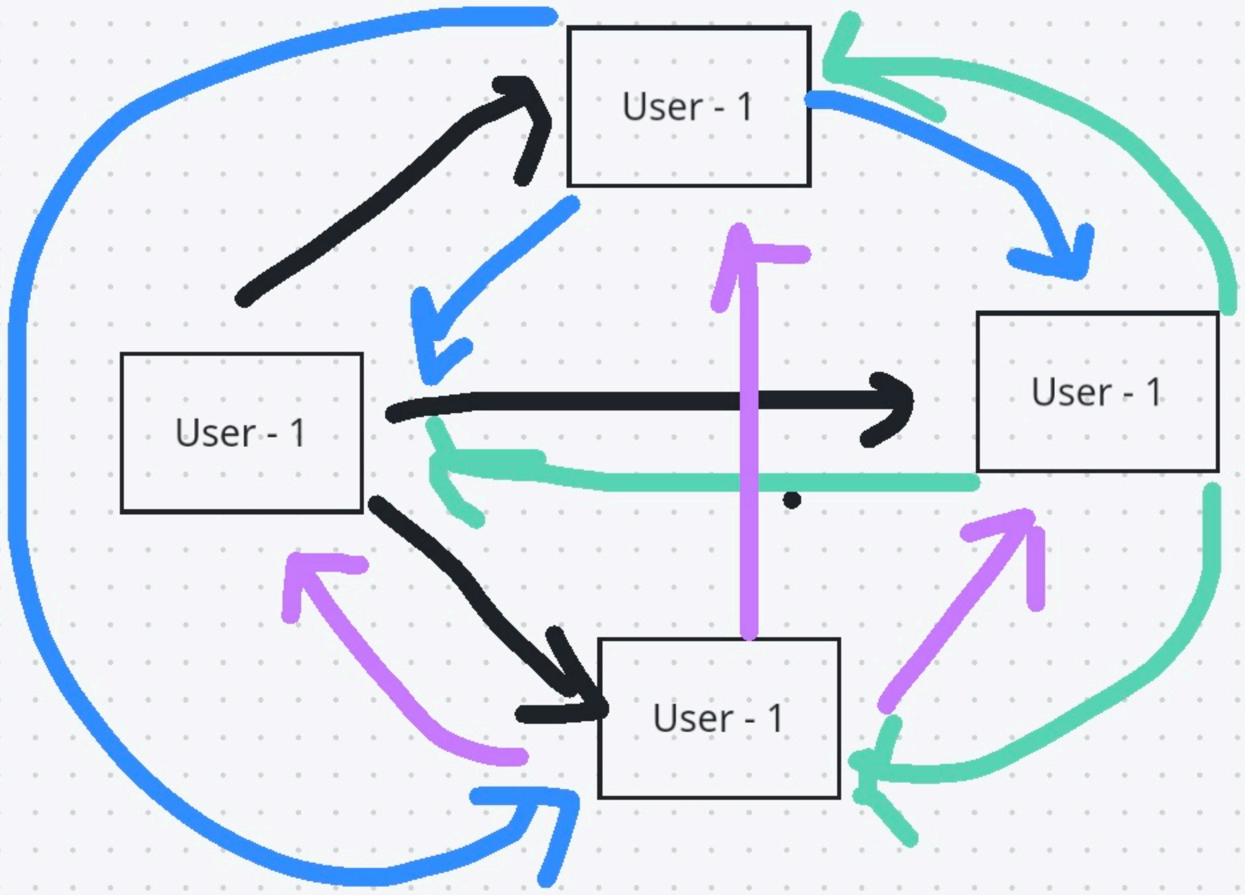


Now, we have successfully created the peer to peer connection

Now, we can do video call with one to one user

What if we have to do in group ?

1. Peer to Peer architecture



Now what is the problem here

for one video call we have so many connections

4 user
EVer user is connecte with the other 3 user

which is 4*3 => 12 Connection

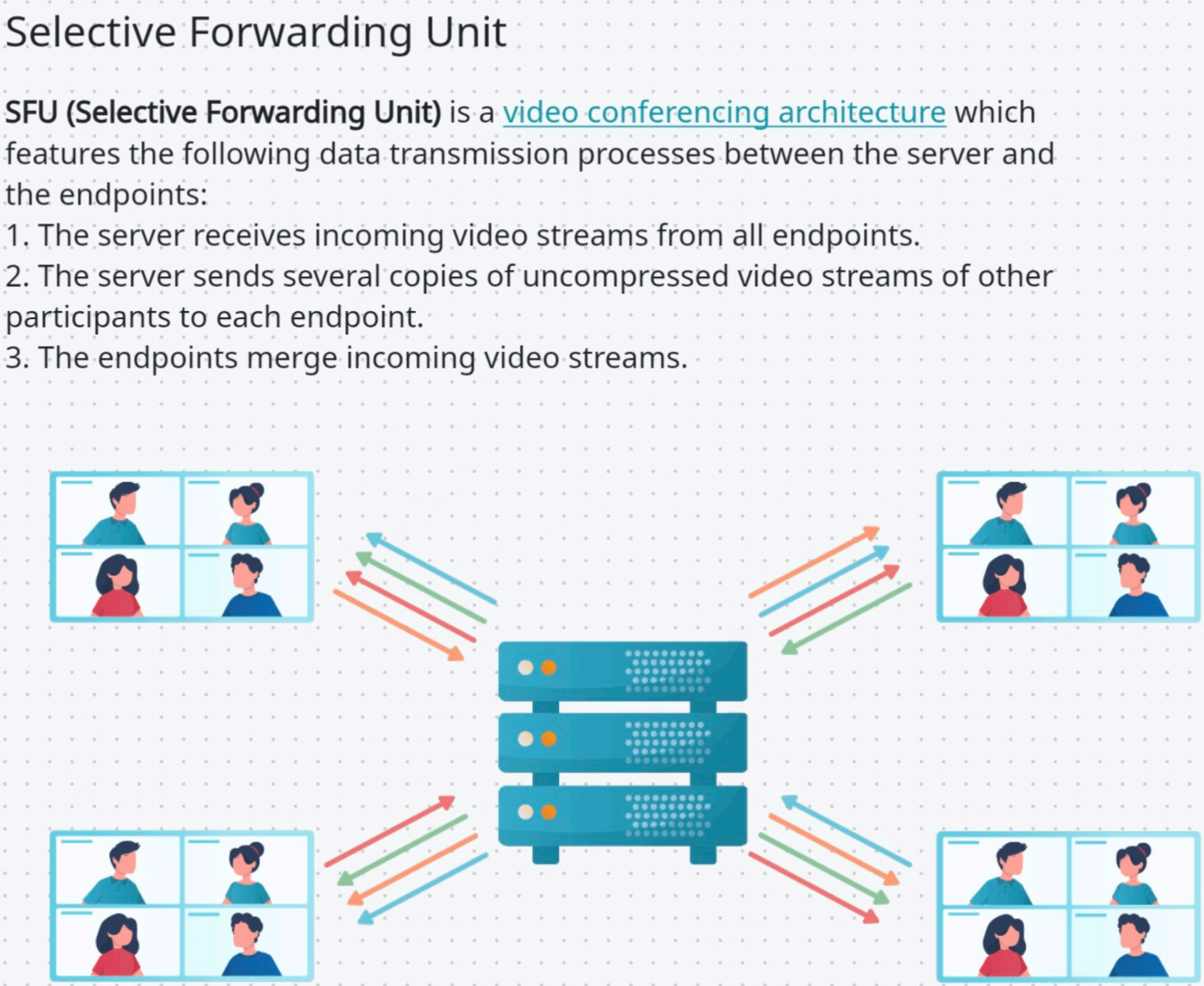
For general if n user
n-1 connection for every user

Total Connection = N*(N-1) => N^2 - N = ~N^2
which is pretty expensive

Suppose if we want to go ahead with this approach,
then how we will do Recording

1. Recod and every user computer then at the some how manage to merge all in async fashion is not possible and also not good approach, what If one user reload his browser, or remove all the cache then the data of the user will get lost
2. Other apporach we can use some centralized server

Centralized server architecture



Selective Forwarding Unit

SFU (Selective Forwarding Unit) is a video conferencing architecture which features the following data transmission processes between the server and the endpoints:
1. The server receives incoming video streams from all endpoints.
2. The server sends several copies of uncompressed video streams of other participants to each endpoint.
3. The endpoints merge incoming video streams.

Something like this