May 2023

# EDGE RELAY CASCADING TO FACILITATE REAL TIME COMMUNICATIONS

Mauro Sardara

Michele Papalini

# EDGE RELAY CASCADING TO FACILITATE REAL TIME COMMUNICATIONS

AUTHORS:
Mauro Sardara
Michele Papalini

## ABSTRACT

For networking scenarios in which real time traffic is exchanged between multiple participants for a given network session (e.g., a conference, a gaming session, etc.), such traffic often flows from one participant to another by passing through a server, such as a cloud server, which can be expensive and can introduce latency for the traffic. Presented herein are techniques through which mesh edge nodes can be utilized to facilitate real time communications in a manner that allows for the ability to save cloud costs.

## DETAILED DESCRIPTION

A conference or a multiplayer game session is typically composed of multiple participants exchanging information via a server (e.g., a media-bridge or a game server) running in the cloud. In such an environment, real time traffic flows from one participant to the other by passing through server, effectively creating a star topology.

In some instances, an edge node can be inserted in the path between a client and the server, thus providing support for re-transmissions and cost savings. In scenarios in which there are several edge nodes involved in a session (e.g., 'N' participants connected to 'M' edge nodes), additional benefits may be realized by trying to keep most of the traffic at the edge; the edge nodes can be connected together in such a manner as to prevent the traffic from passing through the cloud, which is typically more expensive, both in terms of money and latency. Figure 1, below, illustrates an example architecture for such a network environment involving interconnections among two clients (C), two edge devices (E), and a cloud server (S).

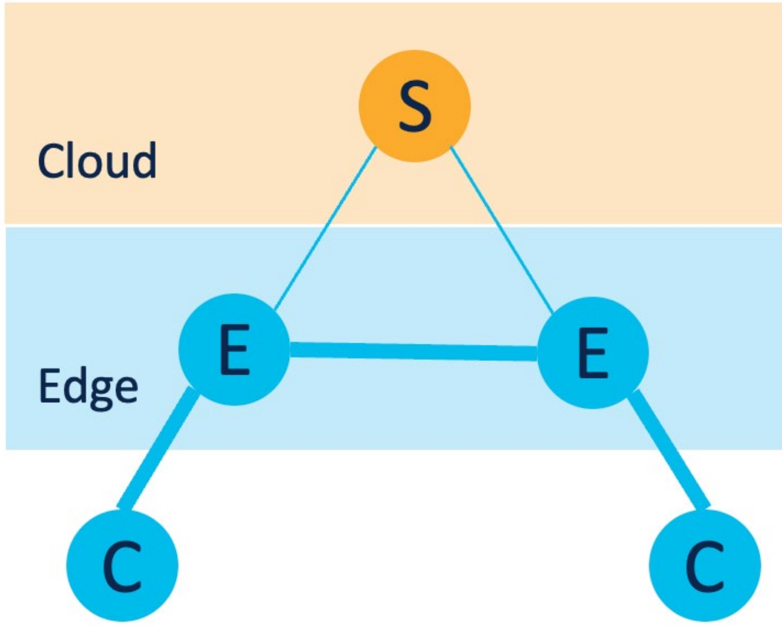


*Figure 1: Example Star Network Architecture*

In order to address such issues, techniques are presented herein through which mesh edge nodes can be utilized to facilitate real time communications in a manner that allows for the ability to save cloud costs.

Consider an example scenario for a video conference involve one video-bridge, typically running on the cloud, 'N' clients/participants, and 'M' edge nodes inserted in the path between the clients and the server.

During the conference, the cloud server can bound the conference traffic to flow between the edge nodes and the participants by dynamically re-arranging the connections and the routing between the edge nodes so that the traffic flows through the mesh of edge nodes and never reaches the cloud server. An objective of such bounding is to reduce the traffic in the cloud, which is typically more expensive, both in terms of money and latency. The server can act as the control plane in order to connect the edge nodes to create an edge topology, while not participating in the data plane communications (i.e., the cloud server does not forward conference traffic).

The edge topology can be created with the objective of minimizing latency between participants. As network conditions may vary over time, the server can operate to actively rearrange the topology in an optimal manner. To facilitate such operation by the cloud server, the edge nodes involved in a call can periodically send a report indicating the latency measured between each other and their current load so that the server can react if the network condition on a given connection fail to satisfy real time communication requirements or if a node is overloaded.

When a new participant arrives, the participant can be connected directly to an edge node if one exists that is close enough to the new participant. This decision regarding the connection can be provided by the server that has a full view of the network/call topology. However, if no edge node can be used (either because to ping is too high or all edges are overloaded), the new participant can be directly connected to the server. In this case the traffic, at least partially, flows also through the cloud.

Accordingly, techniques proposed herein can help to facilitate both cost savings (computing + network), as network traffic does not reach the cloud, and also latency reduction. Further, techniques proposed herein can help to prevent a single-point-of-failure issue, such as may be caused when traffic runs through a single, central (cloud) node.