agora

WHITE PAPER

Agora Network Performance

agora.io

Table of Contents

Introduction 1

The Public Internet 1

Internet Topology 2

- Tiers 2
- Peering and Transit 3

Routing 4

Internet in Real-Time (Routing) 5

- Dynamic Routing Conditions 5

Agora SD-RTN™ 6

- Goals 6
- Design 6
- Architecture 7-8

SD-RTN™ Performance 9

- Latency (ping times) in MS 9
  - Intra-Region 9
  - Inter-Region 9

Conclusion 10

Introduction

The delivery of reliable live audio/video streaming on the internet is met with a number of challenges. This paper describes these challenges are, the traditional solutions that technology providers have historically offered, and how Agora takes a different approach to solving the problem.

The Public Internet

The internet was designed as a best effort system. This, in short, means that although the internet prioritizes connectivity and scalability, there is no guarantee on delivery or quality of service. This is inconvenient for live streaming, but the public internet is designed this way for a specific reason. As of March 2021, there is an estimate of over 1.84 billion websites on the internet; this is a staggering 8x increase relative to the same measure from 2008. Therefore, the top priority of the public internet has been to sustain this rate of hypergrowth and to ensure that each and every website on the internet is both searchable and reachable. As a result, key quality elements for live audio/video streaming like user experience, reliability, and low latency are not priorities of the public internet.

When you send media across the internet, the data packets pass through a number of segments of the internet managed by various network operators. As we'll soon learn, how these data packets are transmitted has a great impact on live streaming's quality of experience (QoE) and quality of service (QoS). In the next section, we'll first explore what happens in order for data to travel from one end of the public internet to the other.

Internet Topology

Diagram 1 below provides a high-level overview of the topology of the internet. The internet is comprised of a number of interconnected IP networks operated by a large number of businesses, educational institutions, and government entities. The majority of these entities are internet service providers of differing sizes and scopes. These providers are categorized as tier 1, tier 2, and tier 3 providers.

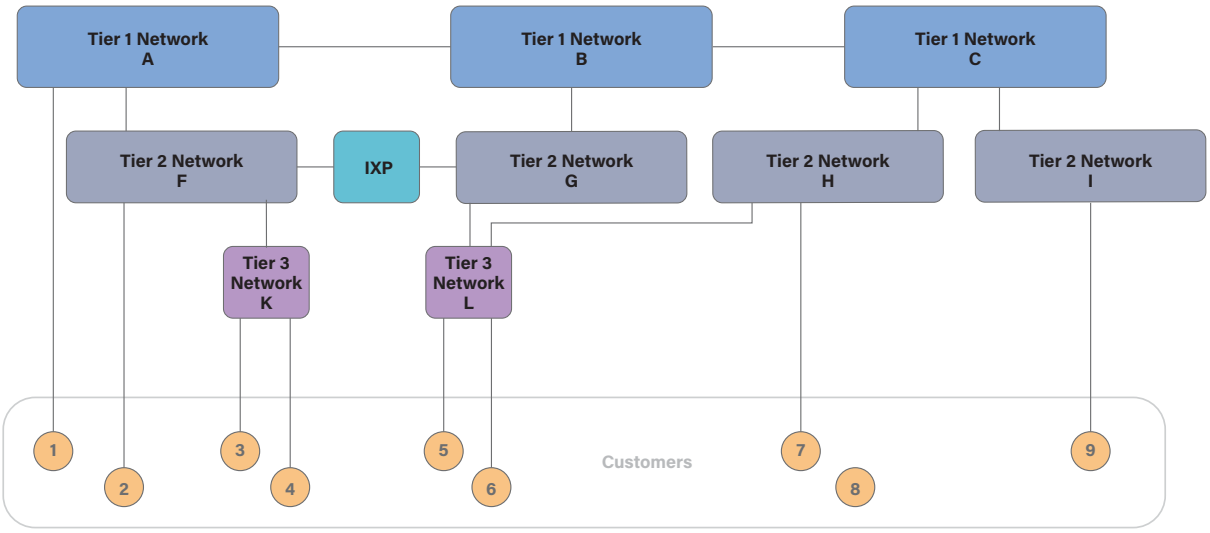

Diagram 1: High Level Topology of Public Internet

Internet Service Provider (ISP) Tiers

Internet Service Providers (ISPs) are typically divided into three tiers. Tier 3 ISPs typically serve as entities that provide internet access to individual customers, focusing on getting as many consumers as possible connected to the public internet. In other words, Tier 3 ISPs are mainly in a volume operating model. Examples include mid-sized home cable broadband providers such as Mediacom Cable, RCN, and Cable One.

Tier 2 ISPs are larger providers with a broader direct reach that serve a diverse customer base, including the provision of transport to tier 3 ISPs. Examples include Comcast, British Telecom, Vodafone, China Telecom, and British Telecom.

Tier 1 providers are often global in scale. Examples of Tier 1 ISPs include AT&T, Lumen Technologies, PCCW Global, NTT, and Tata Communications.

Peering and Transit

If the destination is not directly connected to the ISP, the traffic can be forwarded towards the destination either through 'Peering' or by paying for 'Transit.'

ISPs enter into peering agreements to carry each other's traffic reciprocally through an interconnection at no cost to either ISP.

However, tier 2 and tier 3 networks usually do not have a peering agreement span sufficient to reach all destinations on the internet. Therefore, in order to reach destinations beyond peering range, ISPs must purchase 'transit' (a toll for interconnection between a lower tier ISP and a higher tier ISP) from a higher-level ISP. Some tier 2 and tier 3 providers connect to more than one higher-tiered provider. In this case, they're said to be "multi-homed".

In short, when a lower-tier network does not have enough network reach to send or connect data to a distant point, they pay a higher-tier, 'larger' network provider to complete the connection.

Routing

As you can imagine, this all means that are many paths your real-time media traffic could take traversing through the internet.

ISPs use BGP (Border Gateway Protocol) to create a map of forwarding routes for the destinations on the internet. This map may contain a number of options (see Diagram 2).

However, each ISP will write its own routing policies to determine which of these routes to use. These routing policies take into consideration many factors, including the cost of forwarding.

ISPs generally forward traffic in this descending order of preference:

1. Use ISP's own network
2. Forward to peer ISP for free
3. Pay for transport over higher-tier ISP

In many instances, an ISP will choose lower-quality delivery, because it's free, to avoid incurring the cost of purchasing higher-quality transit for your data packets.

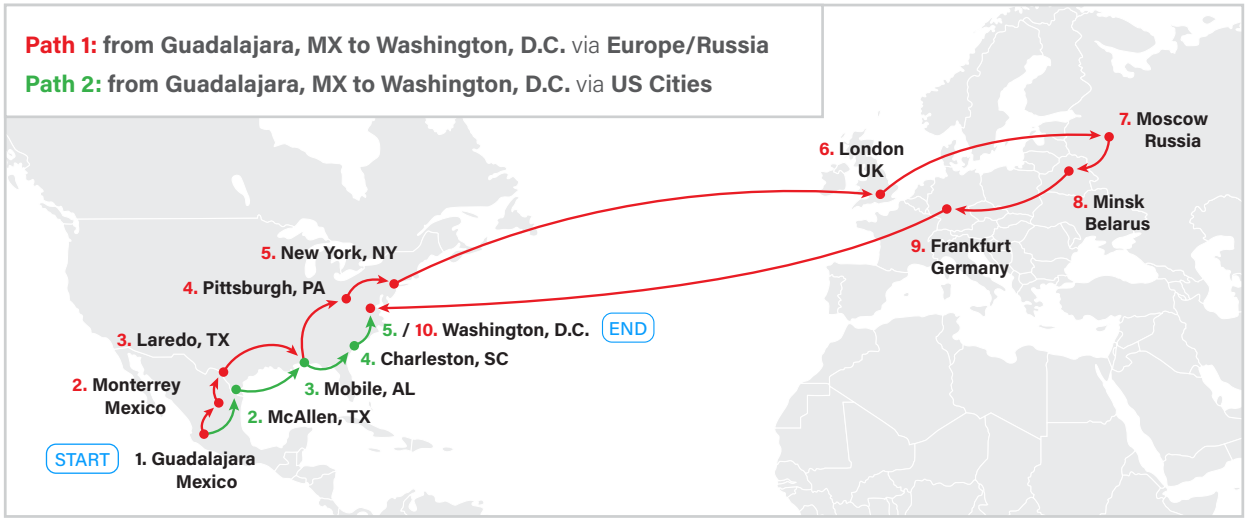

Diagram 2: Examples of two routing paths during high congestion (Path 1 - red) and low congestion (Path 2 - green) scenarios.

Internet in Real-Time (Routing)

Dynamic Routing Conditions

In real-time conditions, the best path across the internet will often fluctuate very rapidly depending on the time of day, number of connected users and data flowing through the network, as well as dependencies with other peer networks.

The routing maps maintained by ISPs through use of BGP often do not keep up with these rapid fluctuations. Therefore, stale route information results in data traffic being routed through points of congestion, which causes packet loss and latency (see Diagram 3).

New protocols for media transport work to minimize this impact by improving the responsiveness of media retransmission when these routing failures occur. A much better approach, though, is to avoid the routing failures altogether.

In the next section, we'll explore how Agora's Software Defined-Real-Time Network™ (SD-RTN™) mitigates above conditions.

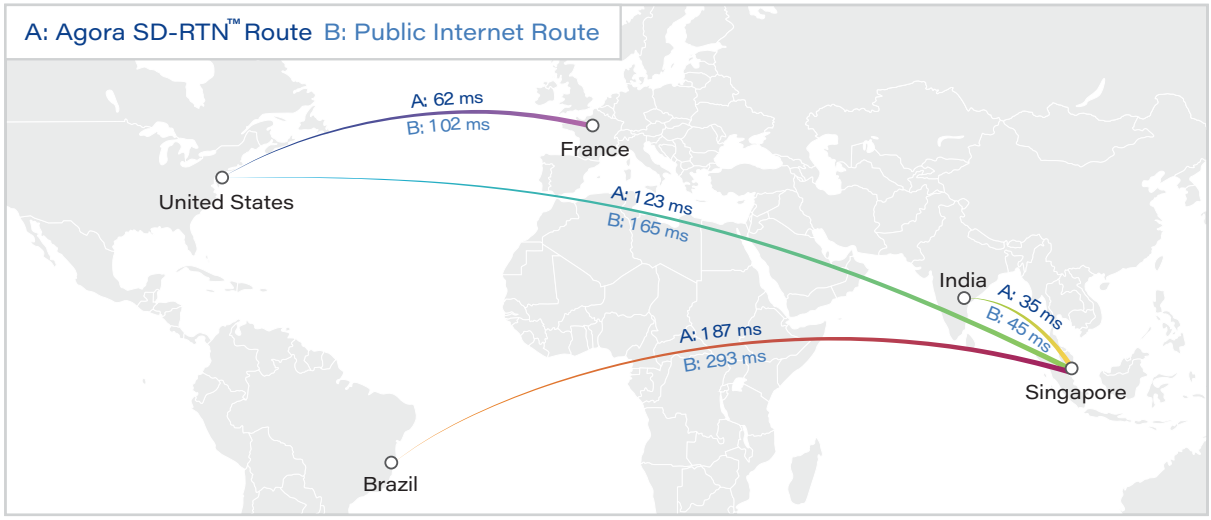

Diagram 3: Comparison of latency in real-time (measured in milliseconds) between Agora's SD-RTN™ routes (A) vs. Public Internet routes (B).

Agora Software Defined-Real-Time Network™ (SD-RTN™)

Goals

The Agora SD-RTN™ was designed to provide the same level of performance as the best-designed enterprise network architecture while leveraging the public internet.

SD-RTN™ aims to achieve two main goals:

- Deliver high-quality live audio/video streaming performance when sender, receiver, or both are on the public internet, anytime and anywhere in the world.
- Provide this service economically by utilizing the public internet rather than relying on expensive enterprise networking technology.

Design

SD-RTN™ has more than 200 data center POPs (Points of Presence) across the world (see Diagram 4).

Each of these POPs serve two purposes:

- Each is an access point to the SD-RTN™. Every Agora customer in the world has a nearby access point for fast access.
- Each is a node in our network, managing traffic flow around any problem points on the internet.

Diagram 4: SD-RTN™ overlay network, conceptual model; as a private network, SD-RTN™ enables smart routing with more direct paths than the public internet.

Architecture

The locations of SD-RTN™ POPs are chosen for maximum ability to improve and control routing across the internet. Agora lays our SD-RTN™ on top of the public internet with many POP data center entry points spread throughout the world for close-proximity network access (see Diagram 5).

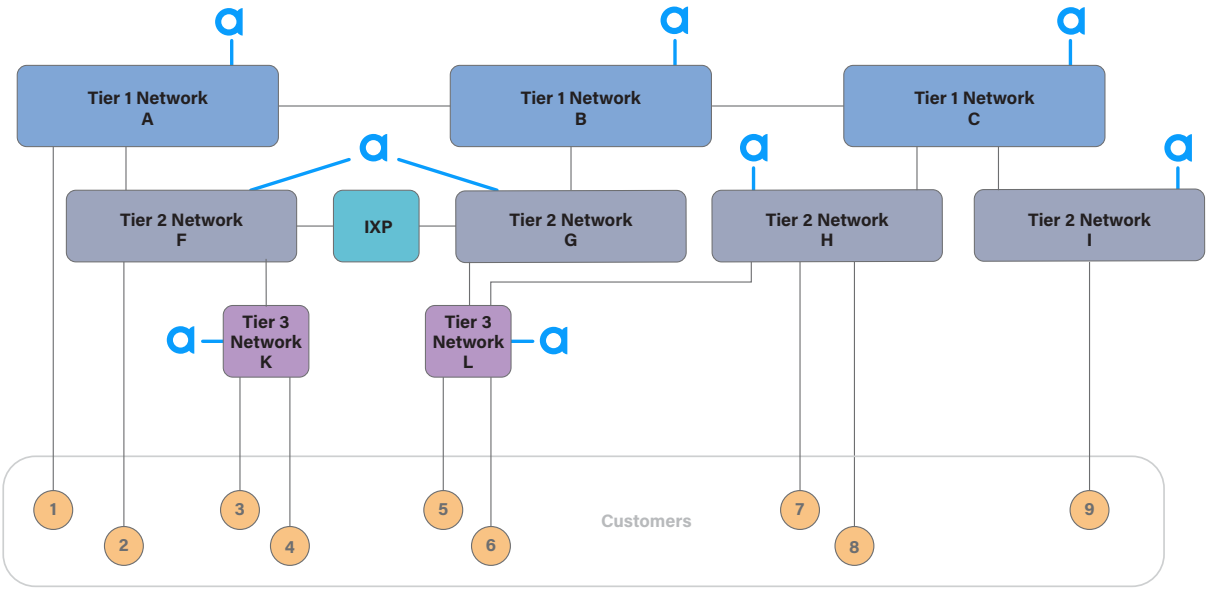

Diagram 5: SD-RTN™ POP locations overlay on top of existing ISP providers.

The POPs of the SD-RTN™ work in full mesh communication with each other. This means that each POP is continually measuring the performance of every possible path through the global network (see Diagram 6).

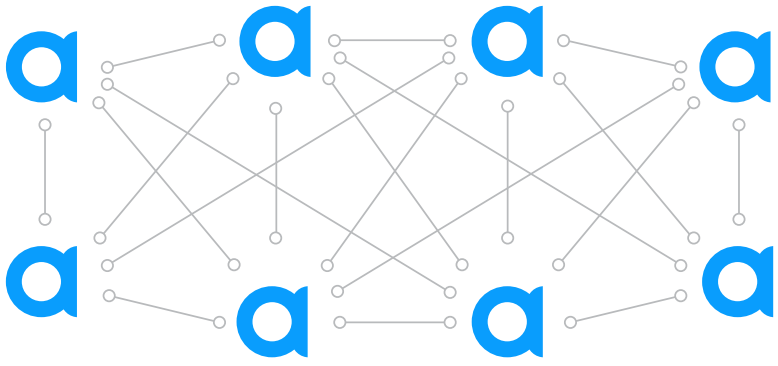

Diagram 6: Agora's SD-RTN™ nodes are connected with one another for maximum routing options.

This constant communication and measurement creates a virtual overlay network on top of the internet with superior routing capabilities. But beyond just routing, Agora's SD-RTN™ also intelligently sends redundant data packets through several separate 'optimized' paths to ensure high data packet delivery success rate within the smallest time window. The data packet that arrives first is used while any lost or late data packet would be ignored.

While what's been described above happens within SD-RTN™, Agora's SDK also handles anti-data packet loss during last mile journey over the public internet.

SD-RTN™ Performance

The routing performance of the internet is dynamic and needs to be considered in terms of an outcome distribution curve.

To understand outcomes, we can select four points on this curve, representative of the latency experienced by selected percentages of users, (see Diagrams 7-8). The tables below describe latency performance for various geographies based on a snapshot of internet data at a specific point in time (note: the internet is in a fluid state and always changing). For example, in the first row of Diagram 7, we show 44ms latency for internet routing within North America under the 50% column. This means that 50% of the time, two points in North America will experience no worse than 44ms latency.

At the far right, on the same row of the same diagram (under 95%), we have the value 94ms. This means that 95% of the time, two points in North America will experience no worse than 94ms latency.

Latency (ping times) in MS

Geography	Routing		Percentage of Users			
			50%	**70%**	**90%**	**95%**
North America	Public Internet	Latency (ms)	44	63	82	94
	SD-RTN™		32	32	32	33
Europe	Public Internet		6	44	72	165
	SD-RTN™		5	6	7	9
Asia	Public Internet		38	55	85	128
	SD-RTN™		30	38	42	62
South America	Public Internet		47	92	127	203
	SD-RTN™		38	56	63	82

Diagram 7: Intra-region outcome distribution table, assessing latency experienced by SD-RTN™ users. Example: Above chart shows 50% of SD-RTN™ users experience 32ms or less latency within the continent of North America.

Inter-Region

Geography	Routing		Percentage of Users			
			50%	**70%**	**90%**	**95%**
North America - Europe	Public Internet	Latency (ms)	102	159	217	273
	SD-RTN™		62	80	81	83
Asia - North America	Public Internet		165	260	343	420
	SD-RTN™		103	115	119	124
Asia - China	Public Internet		45	83	163	236
	SD-RTN™		35	51	63	92
South America - China	Public Internet		293	304	364	472
	SD-RTN™		187	209	219	227

Diagram 8: Inter-region outcome distribution table, assessing latency experienced by SD-RTN™ users. Example: Above chart shows 50% of SD-RTN™ users experience 62ms or less latency when transmitting data between the continents of North America and Europe.

While this data doesn't specifically include packet loss, it generally results from the same congestion phenomenon. Thus, we can extrapolate that packet loss will demonstrate the same trends as latency described above.

Conclusion

Congestion points on the internet are serious barriers to the real-time performance needs of live audio/video streaming. They result in packet loss, latency, and low bandwidth. In turn, this causes dropped calls, inability to connect calls, scrambled audio/video, blurry/low-motion video, frozen video, bad quality audio, and unnatural pauses in the conversation.

The Agora SD-RTN™ is able to route around congestion on the internet, providing true real-time performance for easy and fast connections, stutter-free video, high-quality audio, and low-latency, allowing for more natural communications.

For more details on Agora's SD-RTN advantages, please read our whitepaper "Agora SD-RTN™ Delivers Real-Time Internet Advantages".