agora

WHITE PAPER

Agora's Software Defined Real-Time Network™ Delivers Real-Time Internet Advantages Over Content Delivery Network

www.agora.io

Preface

The public internet as we know it has been evolving and growing exponentially in the past three decades. From the days of dial-up modem connection and small internet cafes in the early ‘90s to the social media-driven mobile networks of the late 2000s, the public internet has become a critical part of our daily lives. Like air and water, many of us can’t imagine living a day without internet connectivity. Despite its significant growth and many innovations, the public internet still has many limitations today that prevent it from fulfilling the high demands of live audio/video streaming.

In order to understand the challenges of live audio/video streaming over the internet, we must first understand the fundamental architecture of the internet.

The Internet Is Best Effort-Based

The internet was designed as a best effort system. In short, this means that although the internet prioritizes connectivity and scalability, there is no guarantee on delivery or quality of service. While this is inconvenient for live streaming, the public internet is designed this way for good reason. As of March 2021, there were an estimated 1.84 billion websites on the internet, which is eight times more than the number of websites measured in 2008 (Netcraft). Therefore, the top priority of the public internet is 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 any data across the internet, the data packets actually pass through different physical networks managed by various network operators. Since the public internet is like a huge complex spider web, there are many paths your data traffic could take going through the internet. Internet Service Providers (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. However, each ISP will write its own routing policies to determine which of these routes to use. These routing policies take many factors into consideration, including the cost of forwarding.

ISPs generally forward traffic in this descending order of preference:

1. Utilize 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 over incurring the cost of purchasing transit for your media packets.

AGORA'S SD-RTN™ PHILOSOPHY

Agora is a real-time engagement (RTE) platform-as-a-service (PaaS) company. As part of our core platform, Agora offers live interactive audio/video streaming powered by our Software-Defined Real-time Network (SD-RTN™) that is accessible and available anytime, anywhere in the world.

We built our SD-RTN™ with the philosophy that advancements in technology would allow us to deliver a higher quality experience than traditional carrier networks. The advantage of SD-RTN™ versus a traditional carrier network is that SD-RTN™ isn't confined by device, phone numbers, or a telecommunication provider's coverage area. Agora can deliver live user engagement experiences in form of real-time communication (RTC) with the following advantages:

1. Cost Effective

- Agora's SD-RTN is a software overlay that is much less costly than traditional telco.
- Communication realized through software modules on PC or mobile devices is inexpensive especially when compared with international phone calls, which still require dedicated hardware.
- Global connected network with no roaming cost regardless where a user is located or traveling from one area to another.

2. Adaptability

- Able to deliver 1 to 1, 1 to many and many to many engagement experiences.
- Adaptable to use cases that phone devices cannot accommodate. Experience can further extend into 360-degree video, AR/VR environment, and other interactive engagement experiences not possible with a phone network.
- Compatible with a wide variety of experience delivery devices (e.g., smart watch, VR headsets, smart TVs, future IoT devices…etc).
- Engagement experience is not tied to phone device, phone numbers, and country code limitations.
- Service is available anywhere, works indoor (internet/wi-fi) and outdoor (wireless).

3. Quality of Service

- SLA performance is measured in real-time and covers the entire global network that many other service providers cannot guarantee.
- Truely scalable, able to accommodate hundreds of thousands of users while maintaining service consistency and stability.

SD-RTN™ Quality of Experience

Agora's SD-RTN™ is built with the following characteristics in mind to deliver quality of experience (QoE) advantages over our competitors' solutions:

- Fast call connection establishment anywhere in the world when 100 – 200ms versus the ten-plus seconds of some major competitors
* Stable video and audio with no scrambling, blurriness, slow motion, freezing or other similar undesired experiences
* Smooth, natural, lifelike (voice, color…etc) experience

SD-RTN™ Quality of Service

Agora's SD-RTN™ is built with the following characteristics in mind to deliver quality of experience (QoE) advantages over our competitors' solutions

Scalable and Efficient

- Scale from several users to millions in a short amount of time without downtime
- Large Scale
  - Millions of Concurrent Users During Peak Time (PCU)
  - Millions of Concurrent Channels During Peak Time (PCC)
  - Synchronized experience regardless of Connection Speed
- High Bandwidth
  - Terabytes of total network bandwidth
  - 50MB+ single connection upload bandwidth possible (e.g., 4K UHD broadcast)
- Large Channel
  - Million-plus broadcast audience size in a single channel

Reliable and Available

- Zero system-wide downtime in company history
- Median global latency <76ms
- Uninterrupted network service

High Performance

- Carrier-grade SLA and service quality
- Business and Service Data Reporting
  - Online data report
  - Offline data report
  - Incident notification in real-time
- Standardized Connection Points
  - Support RTMP / HLS / OBS
  - Support RTP and WebRTC
- Advanced Network Monitoring
  - Monitor every user, every stream, every connection point
  - Automatic re-route in case of disaster recovery

No Maintenance

- Focus on what your business does best and leave the network to our team of experts

Cost Effective

- Low single-usage cost across any region(s) in the world

Reliability Improvement Over Public Internet

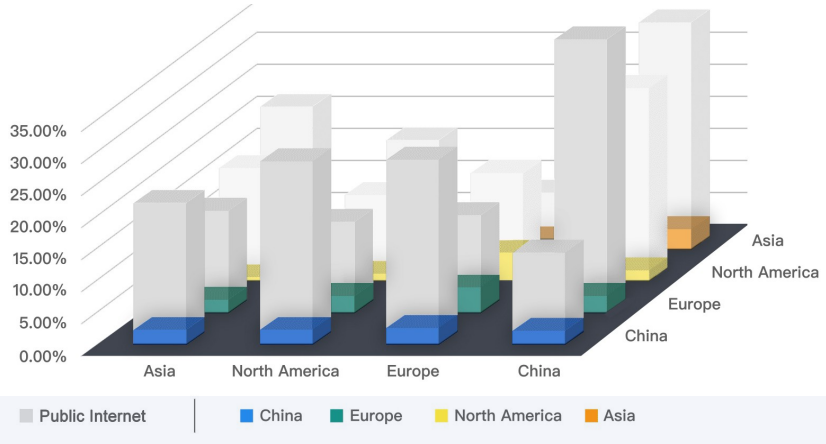
Delivering real-time experience over the internet is not just a milli-second number game, it's a challenge involving a combination of packet loss (random, serial, other), bandwidth, jitter and network delay which are all contributing factors to latency. Agora's network is designed with the highest standard of reliability to combat all of the above challenges in order to create the best ultra low-latency experience over our SD-RTN™.

Packet Loss UDP-based

Packet Loss characterizes the packet drops that occur between a defined network ingress point and a defined network egress point. We consider a packet lost if it does not arrive at the specified egress point within a defined time period. Network packet loss has three primarily causes,

- Network Congestion: When congestion occurs, queues build up and packet drops occur as the buffer becomes full.
- Lower-layer Errors: Bit errors, which might occur due to noise or attenuation in the transmission medium. Actual bit error rate may vary depending on the underlying technology used. For example, fiber optic links could have bit error rate around 1x(10^-13) while older ADSL service could have higher bit error rate around 0.001 (1x10^-3).
- Network Failures: While most networks are built to be highly resilient, from time to time network elements can still fail resulting in temporary loss of connectivity which causes packet loss.

Note: The following charts are based on a snapshot of internet data at a specific point in time. The internet is in a fluid state and always changing.



AGORA SD-RTN™	China	Europe	North America	Asia
China	low	low	low	low
Europe	low	low	low	very low
North America	low	low	very low	very low
Asia	low	very low	low	low

PUBLIC INTERNET	China	Europe	North America	Asia
China	medium	high	high	medium
Europe	very high	medium	medium	medium
North America	very high	medium	medium	medium
Asia	very high	medium	high	high

Figure 1: Table comparison of packet loss rate of public internet vs SD-RTN™

Intra-Region

Measurement Region	Minutes of loss ≥ 0.5% on Public Internet (24h)	Minutes of loss ≥ 0.5% on SD-RTN™ (24h)
China (Same ISP)	116.6.69	1.01
China (Cross ISP)	169.34	0.00
East Asia	15.98	
Southeast Asia	51.41	
North America	38.74	0.00
Europe	132.27	0.00

Inter-Region

Measurement Region	Minutes of loss ≥ 0.5% on Public Internet (24h)	Minutes of loss ≥ 0.5% on SD-RTN™ (24h)
China - East Asia	377.62	0.34
China - Southeast Asia	227.04	9.36
China - to North America	240.67	6.00
China - Europe	247.34	0.67
China - India	91.20	30.77
North America - East Asia	11.09	1.00
North America - Southeast Asia	46.08	0.00
North America - Europe	150.91	0.00
North America - India	25.34	11.95
Europe - East Asia	11.09	0.00
Europe - Southeast Asia	42.19	0.00
Europe - India	27.22	0.00

Figure 2: Table showing minutes of high data loss of public internet vs SD-RTN™

After years of development and fine tuning, Agora today measures our SD-RTN™ 's reliability using our very own very strict standard of ‘less than or equal 0.5% data packet loss per minute.’ Based on Agora’s test measurements, SD-RTN™ is ten to several hundred times more reliable in packet delivery than the best effort public internet.

Network Jitter

Network Jitter is the variation in network delay caused by factors such as fluctuations in queuing and scheduling. In general, jitter can be considered as one-way send delay of two or more consecutive packets. Network jitter typically results in momentary audio/video skip or out of sync.

Network Delay

Network Delay is the time different between an IP packet at a defined ingress point to a defined egress point. Network delays have five contributing factors:

1) propagation delay along the network path (roughly 5ms per 1,000km for fiber optic)
2) switching delay (typically a few microseconds)
3) queueing delay
4) ‘data’ serialization
5) software application end delay

In summary, the best live interactive audio/video streaming network is one that can deliver close to 100% of all data packets within the shortest time frame.

Edge-to-Edge

Ultra Low-Latency Global Performance

Agora's Software Defined Real-Time Network™

Agora provides the world's most advanced private RTC network in the world dedicated to voice, video and live streaming.

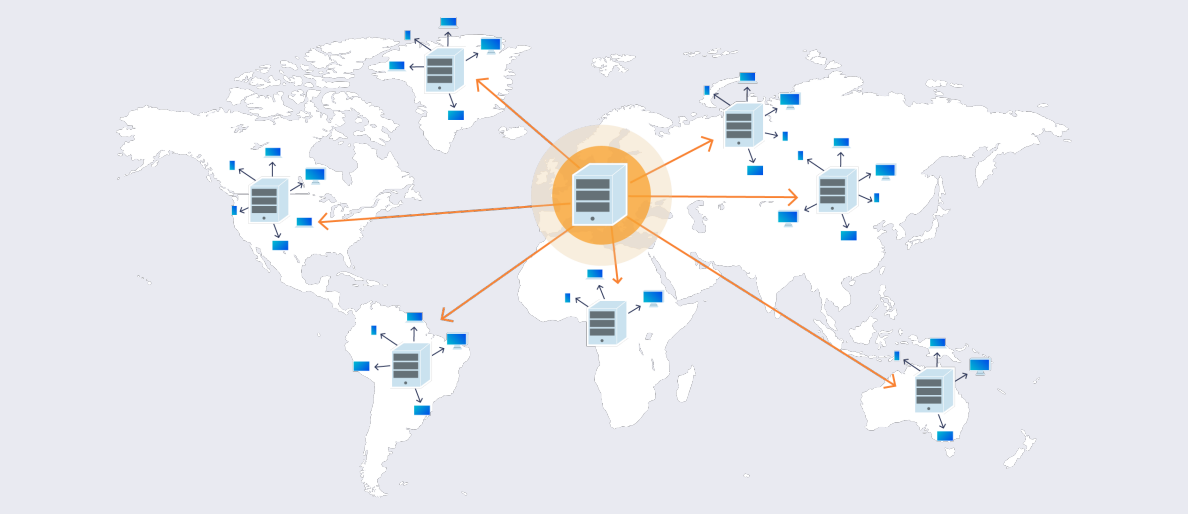
Agora's global SD-RTN™ is made up of data centers dedicated to sub-second latency and high-availability of real-time video and audio anywhere on the globe.

Agora's network is:

- Scalable and efficient
- Reliable and available
- Zero system-wide downtime in company history

**US West Coast to US East Coast*

How Agora handles large audience steaming differently than CDN



First, let's establish how a basic Content Delivery Network works. A CDN replicates content from a central origination source to local distribution centers. Then content is delivered from local distribution centers to each user in the local area based on their location proximities.

Agora's SD-RTN™ architecture uses a similar design to a CDN to support one million-plus users in large broadcast sessions (demonstrated in Figure 4 above). Starting from the broadcast source, the broadcast video stream is replicated to various continental network nodes (e.g., North America, Asia, Europe, etc.). From the continental level, the broadcast video stream is sent further to various local IDCs (internet data centers) and then finally to each individual connected audience.

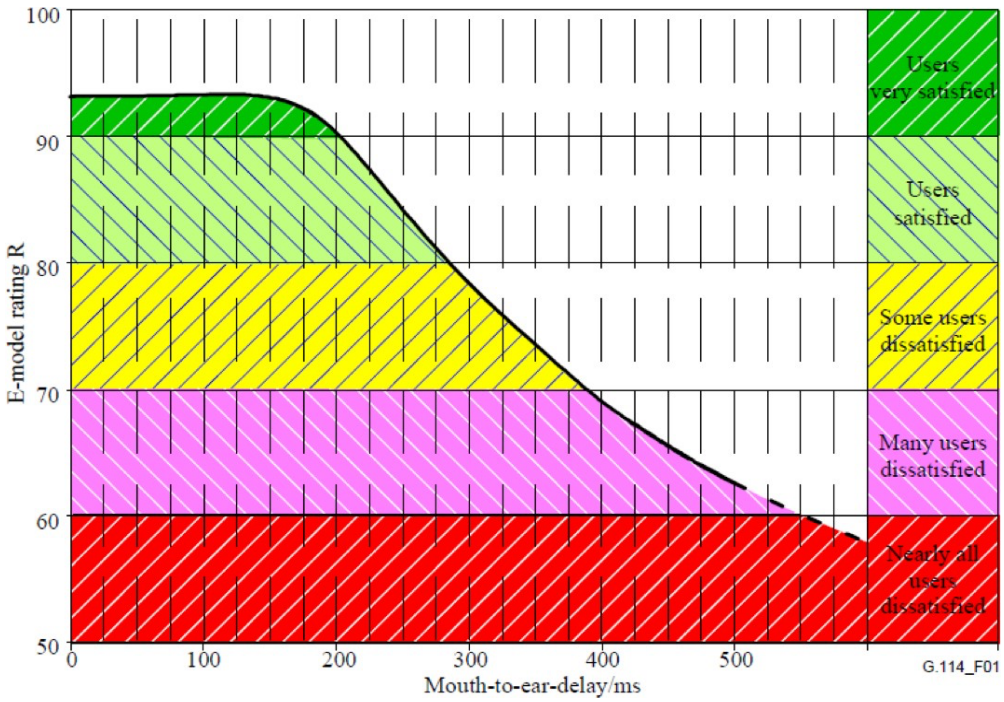
However, this is where the similarities between Agora's SD-RTN™ and CDN end.

SD-RTN™'s "Real-Time Internet" Advantage Over Traditional CDN and Network Services

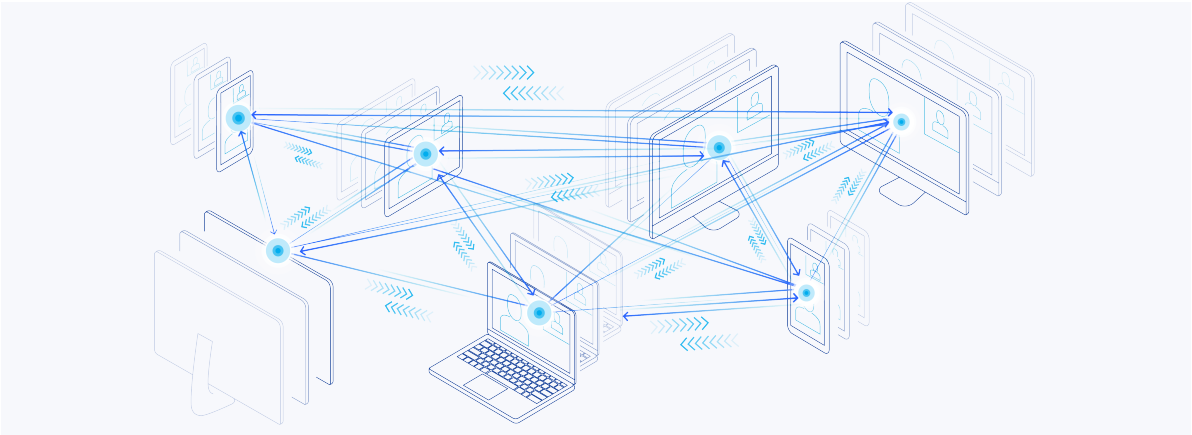
1 LOW LATENCY

The key to the 'real-time' element of live interaction streaming is very low latency. Agora's SD-RTN™ incorporates many network logic and algorithm enhancements in order to deliver the real-time experience needed in a live interactive video streaming session. Loosely referencing the International Telecommunication Union (ITU) standard (aka: MOS - Mean Opinion Score), latency of less than 200ms is considered ideal (more than 400ms would create obvious negative real-time experience impact to users). Latency is a particularly challenging issue in real-time communication. Even at minimal levels, latency can impede the natural flow of conversation.

At high levels, latency makes real-time communication nearly impossible.
Below figure extracted from ITU G.114 standard depicts telecommunication industry's study on voice latency vs user satisfaction quality.

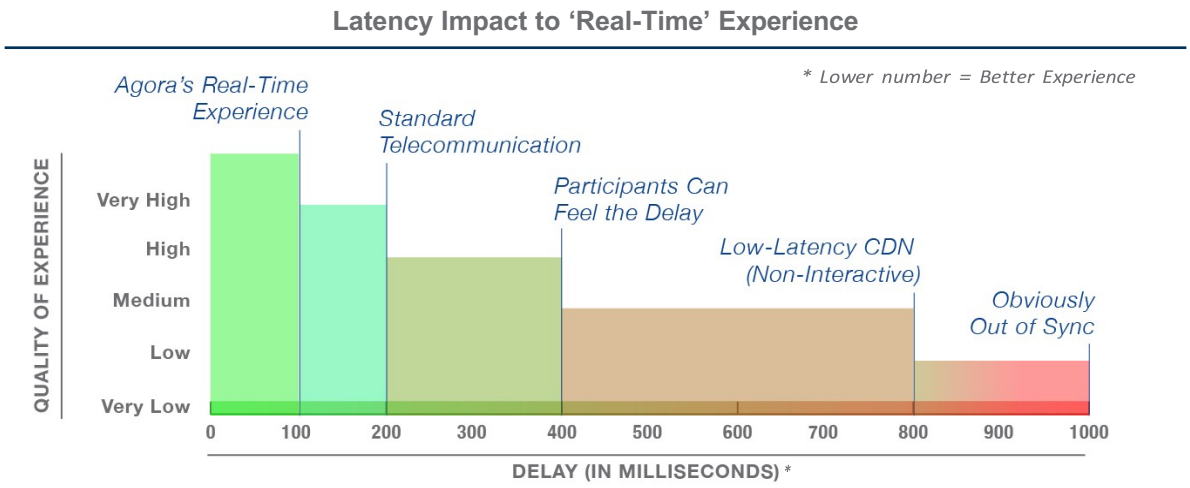


In summary, the lower the latency or delay, the more 'real time or live' a user's experience feels.



In the case of a traditional CDN, ultra low-latency is not required because a CDN only needs to worry about sending content to an end user in the highest quality possible. Usually this means that CDNs build in seconds of buffer and delays to achieve smooth high resolution video streaming.

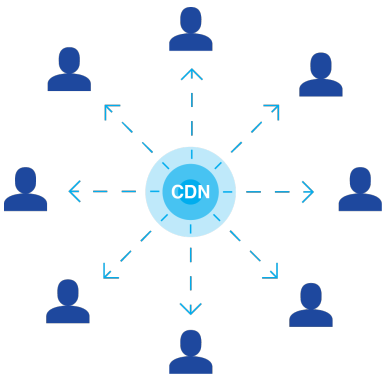
However, during a live interactive video streaming session, when the platform needs to create a natural interactive dialogue between participants, achieving ultra low-latency becomes a huge challenge.
The chart below shows the impact to 'real-time' experience as latency grows.



Agora's SD-RTN™ achieves a median latency less than 76ms globally. Above latency is sustainable from a small 2-4 person chat session all the way to one million users in a live interactive video streaming session. Traditional CDN cannot achieve Agora's level of ultra low-latency. This is the first 'Real-Time Internet' advantage SD-RTN™ delivers over competitors.

2 BEYOND UNIDIRECTIONAL STREAMING

In the case of content delivery, the CDN only needs to worry about delivering video content to each end user in the shortest amount of time possible. CDN achieves this with a local cache content node. What traditional CDN architecture doesn't need to take into consideration is the 'relationship' and bi-directional streaming between participants. For example, if we're broadcasting a concert globally, a CDN only needs to cache a local copy of the concert throughout all of its local content nodes. CDN does not need to worry about the millions of end users communicating back and forth with each other in real time or make each of its end users discoverable (aka: presence) by others in real time.



Traditional CDN :

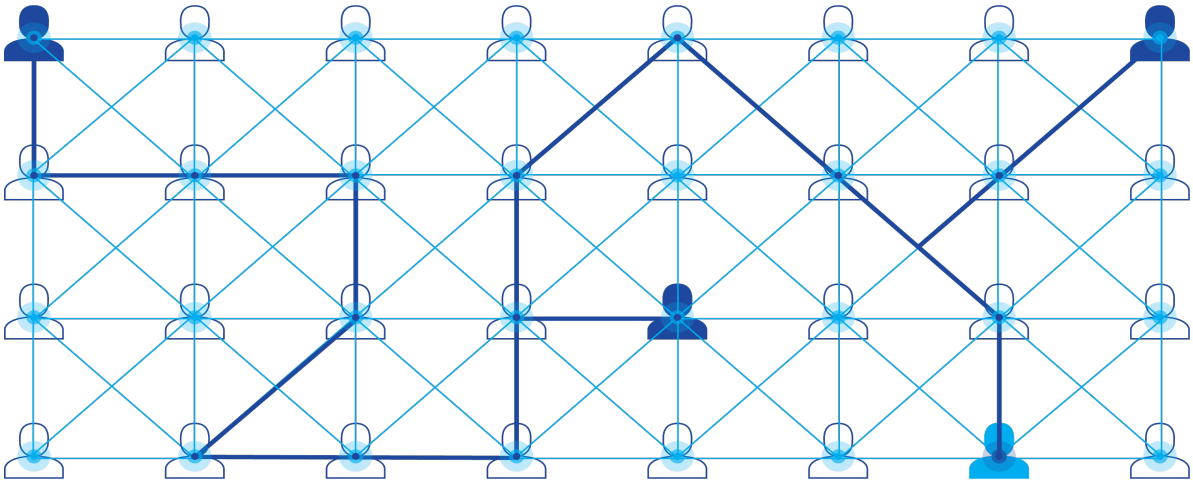
Single Connected Relationship

Agora's SD-RTN™ :

Multiple Connected Relationships

This is the second 'Real-Time Internet' advantage SD-RTN™ delivers over competitors.

3 LARGE CHANNEL



To further achieve the lowest possible latency during a large audience broadcast session, Agora incorporates a 'large channel' concept into SD-RTN™'s smart routing logic. To understand the concept behind large channel, first we'll review a point we discussed in a previous whitepaper. In order to minimize data packet loss, Agora's SD-RTN™ sends redundant data through the three most optimized network paths possible by default. Large channel builds upon the above concept that in a multi-participant live interactive video streaming session, Agora's SD-RTN™ will try to establish optimized routing paths that are shared by as many common participants as possible. Agora's large channel approach achieves a fine balance between ultra low-latency and audio/video experience synchronization between all participants during a live interactive video streaming session. This is the third 'Real-Time Internet' advantage SD-RTN™ delivers over competitors.

4 HIGH AVAILABILITY AND ACCESSIBILITY

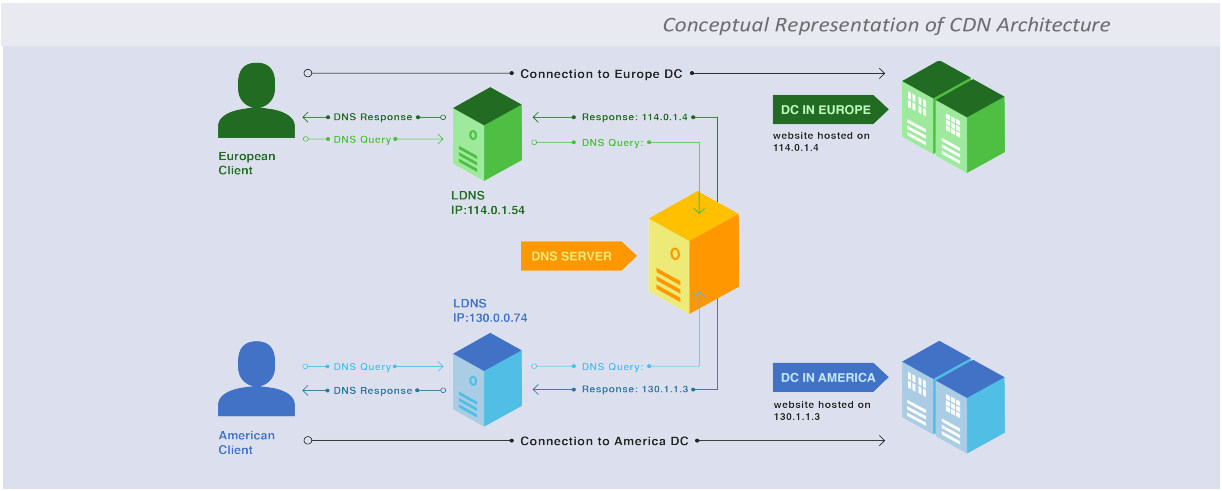
The best network design is useless if the service is not available or accessible. Anyone who has tried making a Skype or similar product call from a country outside of the service provider's primary network coverage knows the pain of waiting to connect just to get a 'Failure to Connect' error at the end of a 15 – 20 second connection attempt. To provide best in class network availability and accessibility, Agora's SD-RTN™ goes beyond standard disaster recovery and hardware redundancy practices.

To better appreciate Agora's high availability and high accessibility architecture, let's revisit how a CDN works. A traditional CDN typically offers a web entry point for client connection requests. Then the CDN service compares the client's IP address location, determines the closest CDN distribution node from the client, and then sends back the best matched IP address (of the closest distribution node) listed in the DNS server's list.

The first weakness in this traditional CDN design is that the CDN cannot reroute connection in case of a local service outage. When an outage occurs, the active client connection would drop and then the client would immediately attempt to reconnect by making a new web request. Depending on how quickly the CDN service provider recognizes there's an outage and how frequently the DNS lists get updated (usually every 10-15 minutes), there's no guarantee that the client won't get routed again to the same local distribution service that's experiencing an outage. This weakness also impacts new clients trying to connect during an initial outage prior to the DNS list getting updated. Regardless, a traditional CDN with a DNS-based design cannot achieve seamless cutover during a service outage/interruption.

The second weakness in traditional CDN design is that the DNS list is fixed and biased towards load balancing. This can result in a client on the West Coast of the U.S. getting routed over to a content distribution node on the East coast, especially during peak West Coast evening hours. Since the East Coast is three hours ahead, it would have much lighter network traffic.

Conceptual Representation of CDN Architecture



In comparison to a CDN's DNS-based design above, Agora's SD-RTN™ takes a completely different approach. Agora's SD-RTN™ uses an 'access point' (similar in concept to a universal load balancer) design. Rather than depending on a fixed DNS list, Agora's SD-RTN™ backend customizes a list of best access points for each connecting client upon each access request. In other words, Agora's network logic is dynamic and always looking for the best connection route based on actual network conditions in real-time. As a result, Agora's SD-RTN™ achieves the following benefits over our competitors' designs:

- Guaranteed first-time connection success from anywhere in the world
- An established connection within 1-2 seconds.
- Seamless recovery. If a server goes down, our network backend is able to automatically reroute and re-establish everyone in an active session to a different server with no perceptible interruption.

5 SCALABILITY

Last but not least, Agora's SD-RTN™ design is highly scalable. Our network architecture design allows us to quickly add server capacity at the local data center or connect more data centers at the continent level. Besides being able to accommodate business demand growth rapidly and flexibly across all of the global regions from a traditional hardware capacity ramping perspective, Agora's SD-RTN™ also delivers the same low latency live interactive video streaming experience whether there are a few or a million participants in the session. In other words, not only is our network capacity highly scalable, our quality of experience (QoE) is also highly scalable.

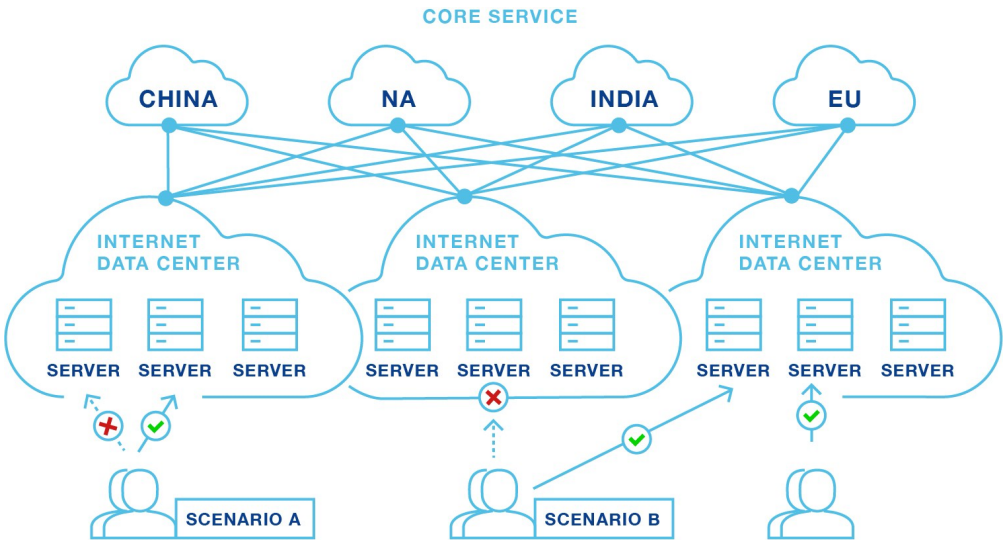


Figure 4: Agora's SD-RTN™ Availability and Scalability High Level Architecture

IN SUMMARY

Today, our SD-RTN™ achieves the following:

- 50+ Billion minutes per month, size equivalent of a small telco
- Ultra-low latency with a median global latency <76ms.
- Over 1,000,000 concurrent participants in a single large-channel broadcast session
- Scalability from a few users to hundreds of thousands without downtime
- Smart monitoring and route optimization. In the catastrophic event of a particular regional data center outage, the network will automatically reroute through our network in other regions