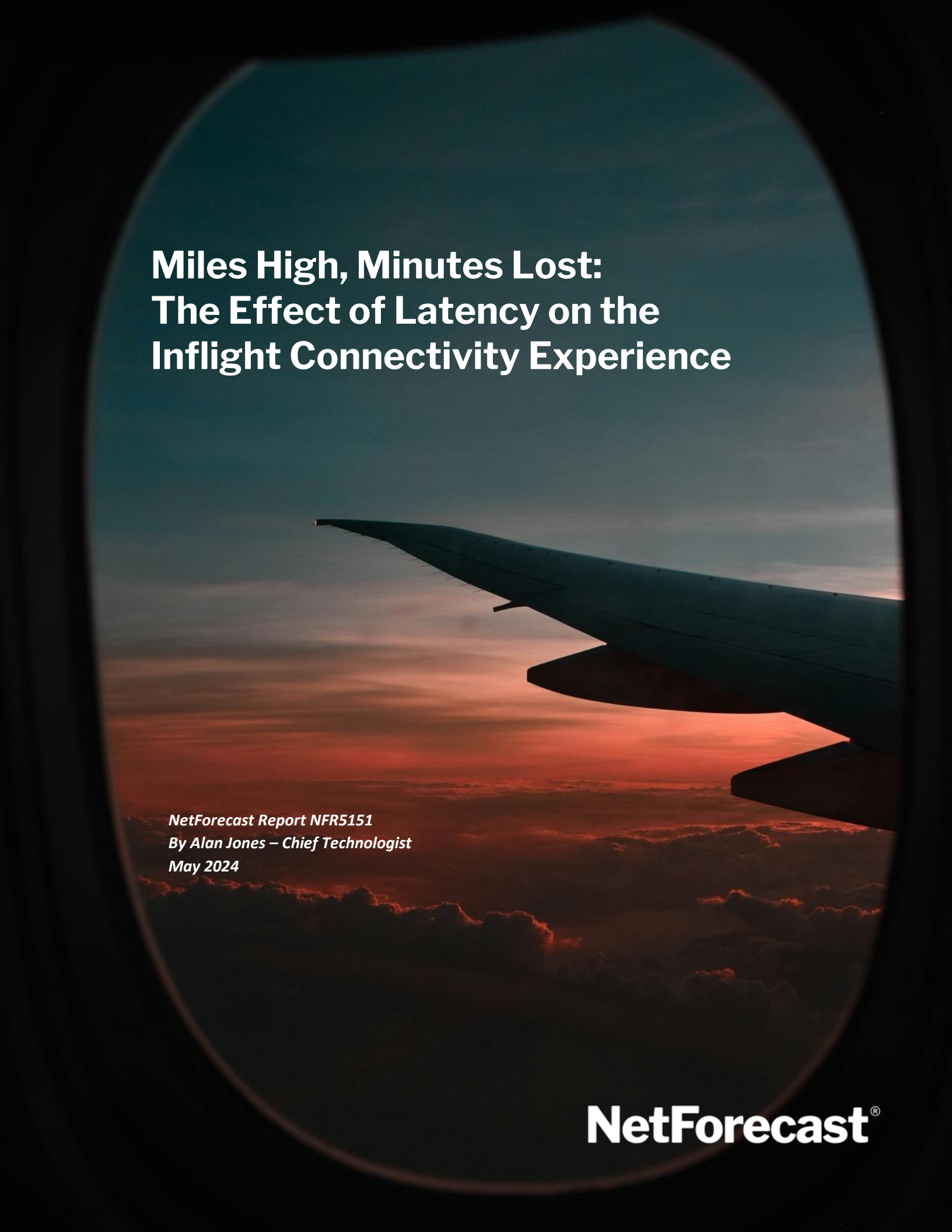
# Miles High, Minutes Lost: The Effect of Latency on the Inflight Connectivity Experience



*NetForecast Report NFR5151*
*By Alan Jones – Chief Technologist*
*May 2024*

NetForecast®

# Table of Contents



# List of Figures

# 1 Overview

Latency's role in influencing a user's Quality of Experience (QoE) is becoming increasingly important as the way we use an ever expanding library of applications interact with online content continues to evolve. Modern applications such as social media, real-time video chat, and gaming require low latency connections to online servers and content to provide a satisfactory experience. The growing phenomenon of AI-enabled assistants and text composition make the need for low latency even more acute. The relationship between latency and the user experience is painfully evident for airline passengers connecting to the internet via Inflight Connectivity (IFC) services. The technology airlines select to connect their passengers to ground-based services is pivotal in determining latency and thus influencing passenger satisfaction.

# 2 Understanding the Passenger Experience

## 2.1 The Changing Needs of the Connected Passenger

Historically, bandwidth has been the most frequently reported and analyzed metric—but bandwidth is only one of the factors that can impact/influence a passenger's inflight experience. Latency, DNS resolution time, loss, and jitter can also significantly affect the user experience.[(1)] Understanding the general way in which networks operate, and looking at the details of how typical applications interact with connected content brings into focus the increasing importance of latency in delivering satisfactory user experience.

This need for continuous, reliable and low latency connectivity applies to both business applications and personal use. Over 90% of business utilize cloud-based services for storage and applications. The same percentage of companies utilize or plan to utilize multiple cloud services.[(2)] Two thirds of small business workloads are now cloud based.[(3)] Business application interactions with enterprise data (e.g., web services, databases, and enterprise software) require frequent exchanges with the cloud-hosted services, and some applications exchange data many times per second.[(4)]

Connectivity for personal use is also driving the need for uninterrupted and low latency connectivity. This affects inflight connectivity users in multiple ways. Surveys indicate users strongly prefer their own devices over seat-back displays. Personal devices typically have a mix of social media apps, games, and music streaming services, which users (particularly 18-to-35-year-olds) strongly prefer over content provided on board the aircraft. Gaming is fast surpassing social media as the primary online activity during flight.[(5)]

## 2.2 Networking Basics

Most data over the internet uses Transport Control Protocol/Internet Protocol (TCP/IP). A TCP/IP transfer breaks the data into small packets and sends the packets individually. To ensure proper transfer, each packet is numbered to assure proper reassembly at the receiving end. The sending device gets an acknowledgment that each packet has been correctly transmitted. Senders transmit a limited number of packets before expecting an acknowledgement. If the delay in acknowledgment is large, then latency not bandwidth becomes the primary limitation on data transfer. Factors such as the number of routers each packet must traverse and how much buffer space individual routers have to hold packets in transit can add significant latency to network traffic.

## 2.3 All Passenger Experiences Are Not Created Equal

The variety and complexity of the applications passengers use during a flight is burgeoning. Calendars, photo organization/editing, entertainment, and social media are among the many application types increasingly essential to passengers' daily lives. Almost without exception, these applications require internet connectivity—and the user's experience of each application is uniquely affected by network metrics.[(6)] User applications can be categorized and analyzed as to how network behavior affects them.

### 2.3.1 Business Suites

In the past, business-centric applications like e-mail, document creation, and spreadsheets were typically considered offline activities and thus insensitive to network conditions. We already discussed how these applications are moving to cloud hosting, with the corresponding dependency on good connectivity. As a sign of the times, business-centric applications are also incorporating AI-powered assistants to boost productivity, and these enhancements are becoming integral to business suite functionality. The added computing power required means that local resources are inadequate, so applications like document editors and email require continuous low-latency connections to deliver a satisfactory experience.

For its Office 365 suite, Microsoft recommends latency under 150 milliseconds, and warns that latency above 200 milliseconds causes performance issues.[7][8] The following table outlines the recommended values:

*Table 1: Microsoft Office Latency Recommendations*

| Application | Latency Guidelines and Impact |
|---|---|
| Email | Above 100-200 milliseconds latency may cause delays in sending or receiving emails |
| Productivity apps (Word, Excel, PowerPoint) | Above 100-300 milliseconds latency may cause lag when opening/saving files |
| Real-time collaboration (Teams, SharePoint) | 50-150 millisecond latency is recommended for seamless collaboration |

### 2.3.2 Gaming

Gaming encompasses a wide variety of online activities. Word solution games and individual interactive games which preload challenges one at a time, have relatively modest network requirements to satisfy users. Conversely, multiplayer games that allow users to move through an environment in a first-person role require low latency to perform well. The most popular of the mobile device games (e.g. Monopoly Go!, Help Me: Tricky Story, Roblox, Match Frenzy: 1 Line Draw, Block Blast!), require low latency to support level change and interaction with friend groups. Poor network performance leads to issues such as awkward character movement and hit marker delays. Services like Twitch that offer the real-time streaming of game play to other users, require even more network resources to deliver a good experience. Cloudflare, a leader in low-latency content delivery services, has created a scoring system to assess the impact of activities like gaming. (https://speed.cloudflare.com) Cloudflare scores are significantly lower when latencies exceed 50 milliseconds.[1]

### 2.3.3 Web Browsing

Loading a web page involves a multitude of operations orchestrated by the browser during page loading and formatting. A typical website pulls data from many servers in many locations. Loading a web page involves not only downloading content, but also style and JavaScript libraries that drive page functionality. Figure 1 shows a breakdown of requests and bytes from a popular sports news website. Note that most of the data is JavaScript ('js') which supports page functionality.


*Figure 1: Breakdown of Request and Bytes (Source - WebPageTest.org)*

This page load involved 297 individual steps, and the page called over 90 unique domain locations. For each of those 90 domains the browser ran a Domain Name Service (DNS) resolution to change a domain name to an IP address. The large number of server interactions in loading a page such as this causes latency to significantly increase the page load time, adversely affecting the passenger experience.

### 2.3.4 Social Media (Infinite Scroll)

The predominant user interface for social media apps today is what is referred to as 'infinite scroll'. Infinite scroll populates content as the user scrolls through a news feed, for example. With sufficient bandwidth and a low latency/low loss connection, the user perceives that infinite content is available. In the past much social media content was text based, while today it more likely contains pictures or videos in individual postings. Other popular apps consist purely of video clips, which require low latency and significant bandwidth to deliver a satisfactory experience.

### 2.3.5 Video Streaming

Video streaming platforms use a variety of protocols to improve the user experience under a range of network conditions. RTMP (Real-time Messaging Protocol) or HLS (HTTP Live Streaming), along with advanced video codecs (e.g., H.264 and VP9), have greatly reduced the adverse effect of bandwidth and latency limitations. The main impact of latency is to slow startup time. Limited bandwidth usually manifests as buffering during playback. Figure 2 shows a YouTube video with the "stats for nerds" feature enabled. This provides real-time insights into the playback of a video and how things like buffer health, resolution, and frame rate are controlled to provide the best playback experience. Applying these technologies allows video streaming to be less sensitive (but not immune) to high latency.


*Figure 2: 'Stats for Nerds' Feature - Streaming Metrics Used to Control Playback (Source – YouTube.com)*

### 2.3.6 Video Conferencing

Unlike video streaming, video conferencing cannot use buffering to improve performance. Although video conferencing services use advanced video codecs to compress video, the separate encoding of video and audio and the sensitivity of most users to lag and imprecise synchronization of video and audio, make video conferencing very sensitive to latency issues.[9]

# 3 Defining Latency and Its Sources

## 3.1 Latency Basics

Ping is a common network utility used to measure the Round-Trip Time (RTT) of a small packet between a computer and a server. Ping is considered so important that it has been built into virtually every computer system since the earliest days of the internet. Unless a network is heavily loaded, the ping utility is an excellent way to determine the latency between a device and a server. Under good conditions, the primary contributor to RTT is the physical distance the packet travels to reach the server and return. Considering routing and other network overhead factors, packets travel through the internet at roughly 30% the speed of light. Routes with long physical distances between them increase this.

## 3.2 Routing and Hardware Sources of Latency

User devices on the internet cannot connect directly to online servers, so data packets must travel through multiple network paths connected by routers. At each router the packet is read into an input buffer and directed to the next router until it reaches its destination. This buffering is the primary reason packet speed is significantly less than the speed of light. During this process, packets typically pass through various networks owned by different entities.

Latency issues become significant when network loading increases to the point that the input buffer on routers along the packet path becomes full and can no longer accept input packets. Sophisticated congestion detection algorithms help ensure packets continue to flow even when input buffers are near capacity. The relationship between latency and loading explains why some speed tests measure and report the effect of loading on latency.[1]

## 3.3 Content Delivery Networks

The adverse effect of latency on the user experience has led to decentralized data centers called Content Delivery Networks (CDNs). CDNs are hosted in local internet service provider data centers or in private data centers in major localities. The considerable expense in using CDNs is worth it to content providers to improve the user experience—primarily through lowered latency.

# 4 Latency and Inflight Connectivity

The nature of a passenger's online experience depends on the aggregate performance of many interconnected network segments as Figure 3 shows. The latency a passenger experiences is governed by the sum of the latency on all the individual network segments in the data path. Bandwidth, on the other hand, is limited by the link with the lowest capacity, while loss is additive and can occur on any network segment along the path.

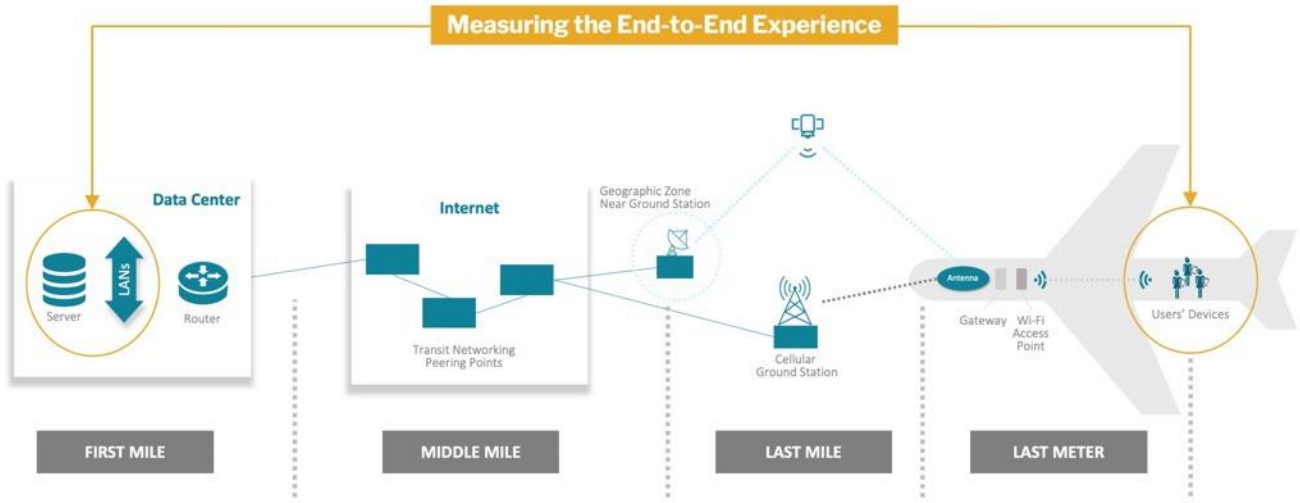



*Figure 3: Total Network Path - User Perspective*

## 4.1 Network Segment Details

The "First Mile" consists of servers delivering services and content to passengers. The first mile servers are hosted in large data centers with excellent connections to the internet backbone networks. These data centers have redundant routing and other safeguards to ensure high reliability and performance. The "Middle Mile" consists of a few, very high-capacity network segments referred to as the internet backbone that carry cross-continent and intercontinental traffic. Issues do occasionally occur at interconnection points between individual internet service providers, but these systems are the most reliable of the internet infrastructure. Latency contributed by the middle mile is primarily distance related. (10) (11)

Connecting an aircraft to the internet happens in the "Last Mile" along the network path. Connecting an aircraft to the internet poses significant technical challenges because it involves complex technology that must endure severe inflight conditions. The primary technologies currently used are geosynchronous satellites (GEOs), low earth orbit satellites (LEOs), and air-to-ground (ATG) cellular systems. The physics of GEO technology dictates that the latencies are an order of magnitude larger than for the other technologies, with latencies exceeding 500 milliseconds. ATG and LEO systems have latencies in the 50-to-100 millisecond range. A major limitation of ATG is the inability to connect over open water. GEOs have limited coverage at far North and South latitudes, while LEOs can provide complete global coverage.

In-cabin Wi-Fi ("Last Meter") faces some challenges; however, most airlines have deployed technology to improve the RF environment, including multiple access points and access point upgrades to newer Wi-Fi protocols.

## 4.2 Latency Implications for Inflight Connectivity Application Classes

### 4.2.1 Web Browsing

NetForecast has examined the relationship between latency and application performance in field observations and in controlled conditions. Figure 4 depicts web page load times for a well-known sports website (ESPN.com) as it was subjected to increasing network latency under controlled conditions.

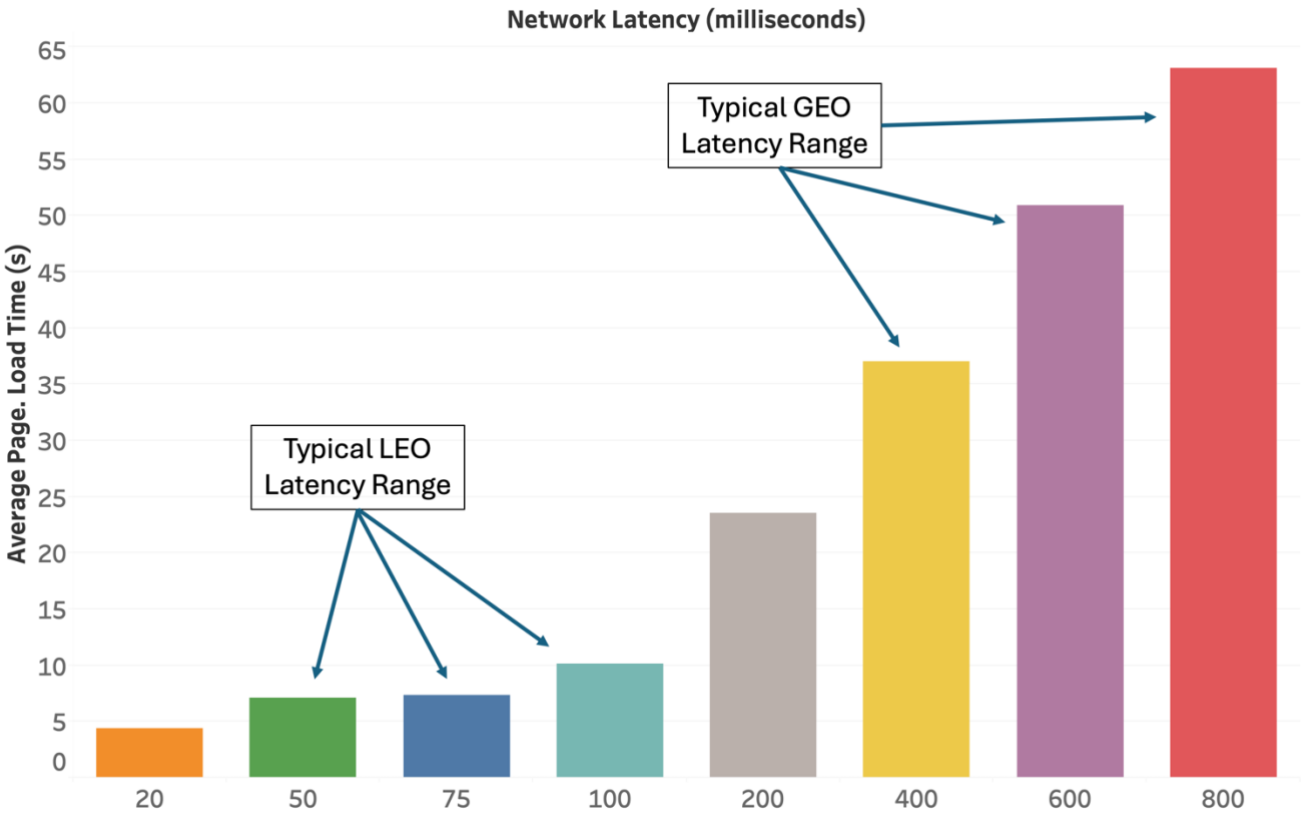



*Figure 4: Initial Page Load Time vs Latency*

As seen in Figure 4, the initial page load time increases significantly after latency exceeds 100 milliseconds. Page reloads with DNS and browser caching in place cut the load time in half, however, the same pattern of increasing page load time remains.

While Figure 4 shows the average page load time under different latencies, Figure 5 shows the probability distribution of page loads across time. The probability distribution, based on NetForecast models, indicates the likelihood that a web page is **completely** loaded at a specific point in elapsed time. Users would see partial formatting of data during the loading process, and refreshing the page would shorten load times based on cached information.

Figure 5 highlights page load time differences between a complex website such as ESPN.com, and a simple website such as Google's search homepage as experienced over a LEO and a GEO last mile connection.

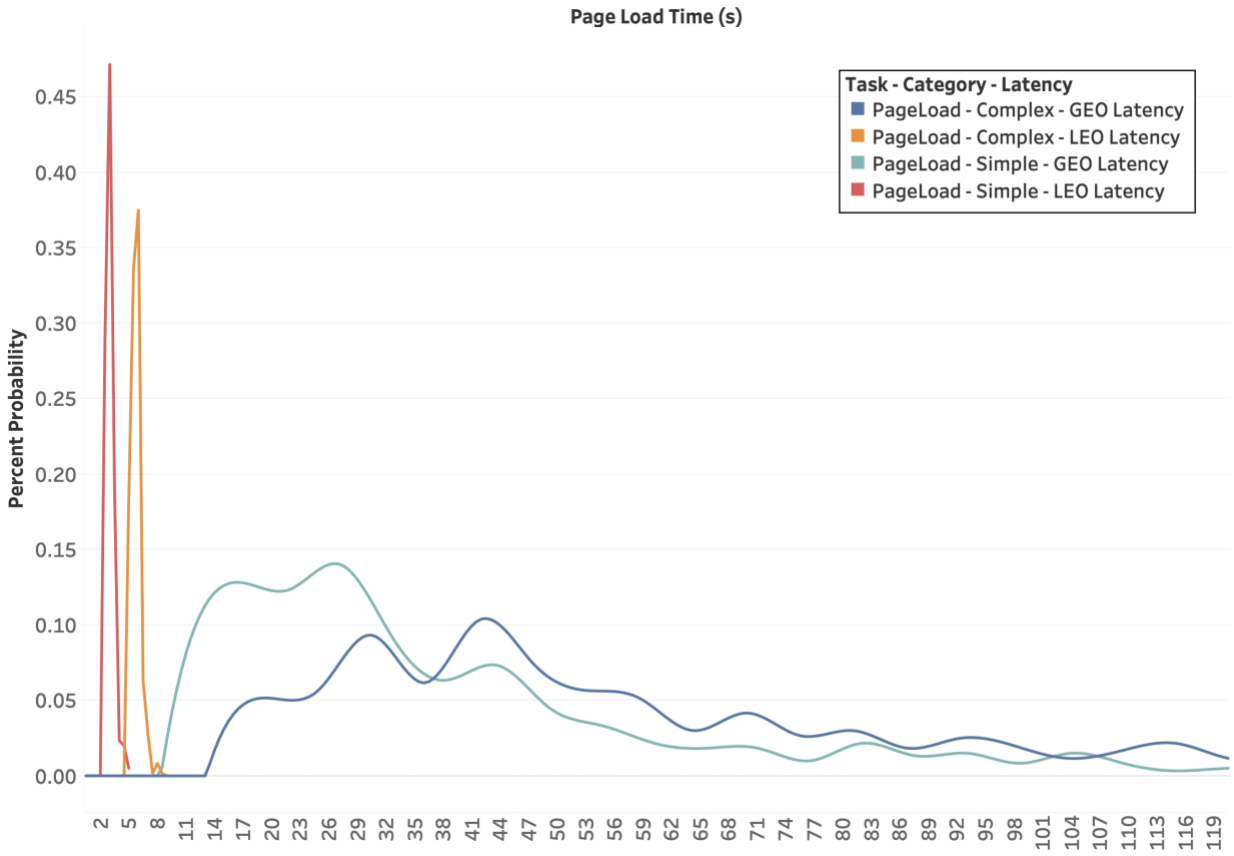



*Figure 5: Precent Probability Distribution of Webpage Load Time vs Latency*

### 4.2.2 Social Media Interactions

Interacting with social media using an infinite scroll interface can be modeled as a series of abbreviated web page loads separated by a time delay for the user to digest the content. This approach was used to model an interaction with an initial load, and 4 'swipes' (scrolls) with two seconds in between for the user to view the content. Figure 6 shows the distribution of each load time for the duration of the simulation, using typical LEO latency distribution.

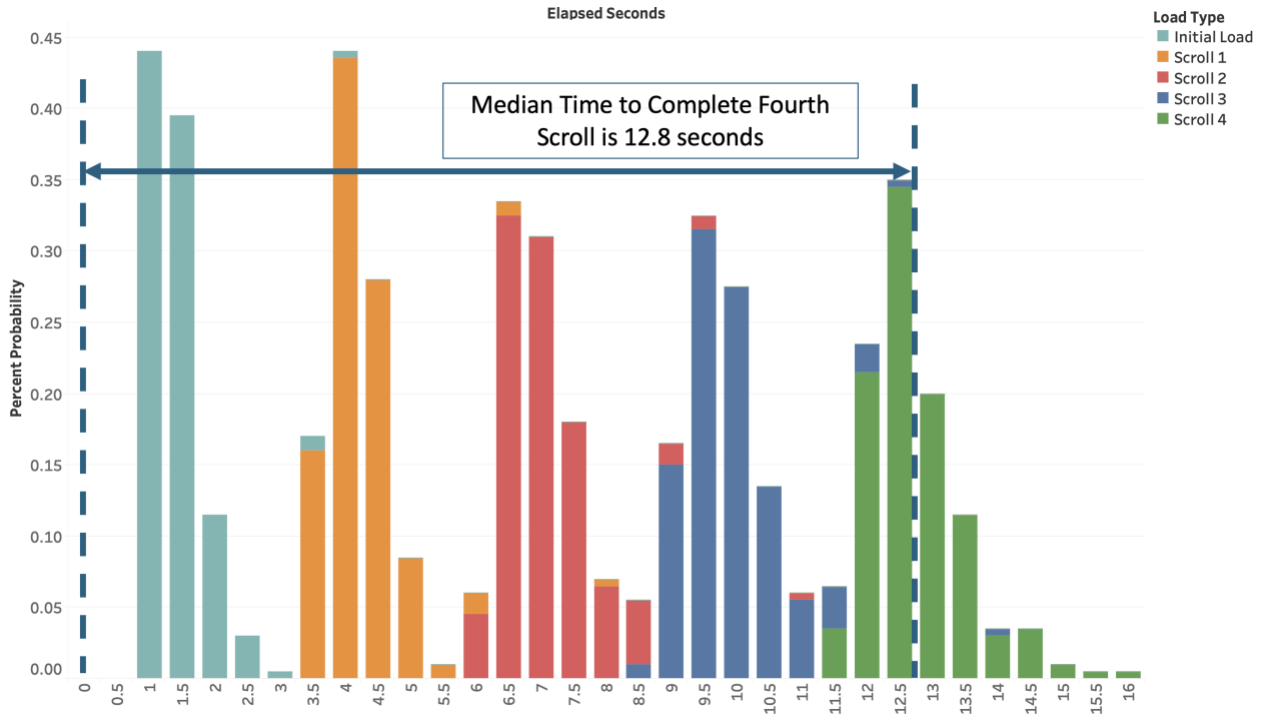



*Figure 6: Social Media Interaction with LEO Latency*

Figure 7 is based on the same model, but using latency from a GEO satellite.

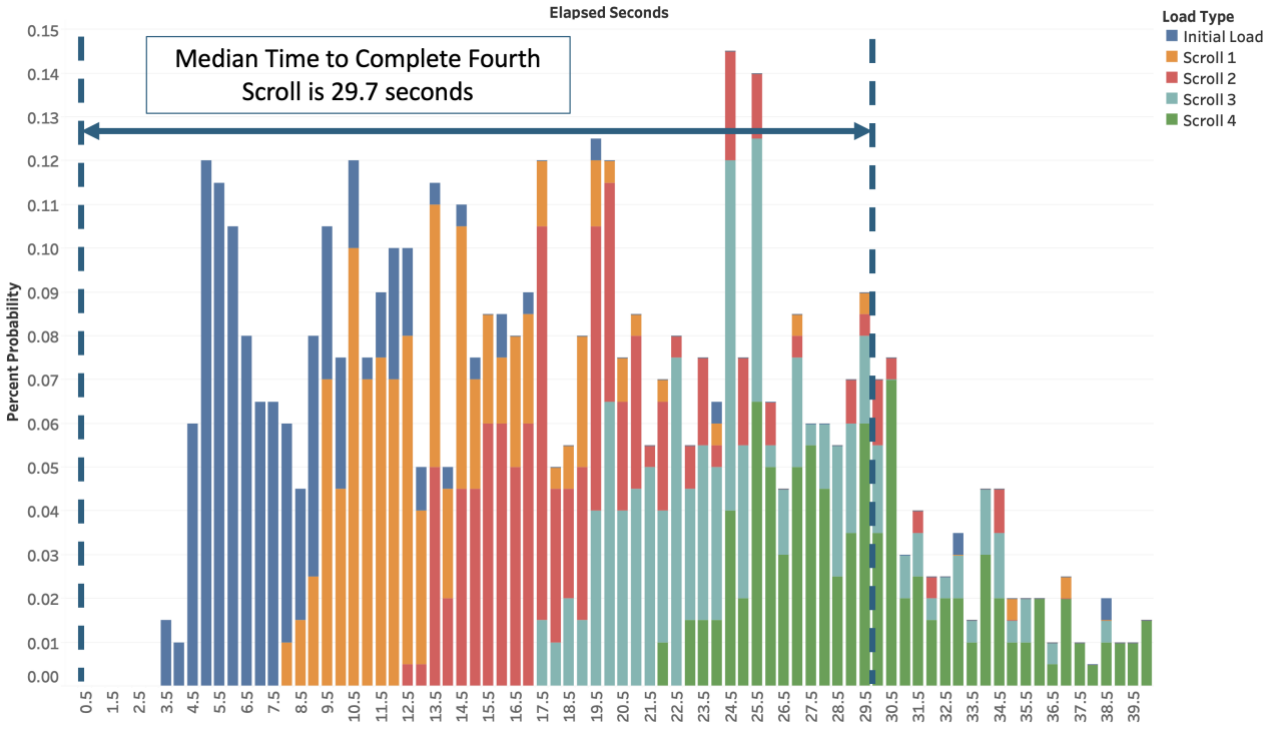



*Figure 7: Social Media Interaction with GEO Latency*

The increase in latency creates delays in the individual loads which degrades the user experience.

### 4.2.3 Other Application Classes

Testing of video stream shows that latency causes slow startup times and reduction of resolution for high-definition videos. Latencies above 100 milliseconds make online gaming difficult to impossible. Latencies above 100 milliseconds start to adversely affect business suite applications like video conferencing—and above 250 milliseconds, the performance of all business suite applications suffers.[7][8]

Although not directly related to the passenger connectivity experience, a wide-ranging set of airline-specific tasks occur in-flight also benefit from low latency.

- Flight Operations and Safety: Airlines increasingly rely on real-time data exchange between aircraft and ground operations for flight monitoring, weather updates, and other critical operational data. Low latency ensures timely communication and enhances flight safety and efficiency.
- Crew Communications: Cabin and cockpit personnel use connectivity for various operational communications, including passenger services, safety announcements, and coordination with ground staff. Low latency helps maintain clear and timely communication.
- IoT and Smart Aircraft Systems: Modern aircraft are increasingly equipped with IoT devices and smart systems for monitoring and maintenance. These systems often require real-time data transfer, which benefits from low-latency connectivity.
- Emergencies: In case of medical emergencies or other critical situations, low-latency connectivity allows for quick consultation with ground-based medical experts or emergency services, potentially saving lives.
- Marketing and Personalization: Airlines can use real-time data to provide personalized services and targeted advertising to passengers during the flight. Low latency helps deliver timely and relevant content.

# 5 Conclusions

Latency is becoming the primary driver of user experience in the air and on the ground, as applications increasingly require continuous interaction with online hosts. The need to deliver IFC services with lower latency will grow as passengers come to expect an online experience in the air that is consistent with their ground-based internet experience.

While geosynchronous satellites are a well-established and trusted source of internet connectivity, they are limited in the ultimate experience they can provide based on the inherent latency limitations of the technology. Air-to-ground cellular and the currently deploying low earth orbit satellite constellations offer vastly improved latency performance, and low earth orbit technologies offer the added benefit of worldwide coverage.

Conclusions by application type:

- Real-time Applications: Passengers increasingly wish to use real-time applications like video calls, VoIP services, and online gaming, which require low latency to function properly. High latency can cause delays, poor audio and video quality, and an overall unsatisfactory experience. Ideally, latency should be under 100 milliseconds for most of these applications, and 50 milliseconds or less for gaming. While these numbers provide a general guideline, the actual acceptable latency can vary based on user expectations and specific application requirements.
- Web Browsing and Streaming: Even for less interactive activities such as web browsing and streaming, low latency improves the user experience by reducing buffering times and making pages load faster. Latency of around 100-250 milliseconds is typically acceptable. This range allows for reasonably fast page loads and smooth email communication.

- Business Productivity: Many passengers use in-flight connectivity for work purposes, including accessing corporate networks, using cloud-based applications, and participating in online meetings. Low latency is essential for productivity and seamless communication. Latency in the 150-250 millisecond range ensures that users can work productively without risk of delays in data synchronization or interactions with online services.
- Video Streaming: Latency under 200-300 milliseconds is generally acceptable for streaming services like Netflix, YouTube, or Hulu. This range helps minimize slow startup and reductions in resolution during playback.

# 6 About the Author

**Alan Jones** is NetForecast's Chief Technologist. He architects and deploys cloud-based internet performance measurement systems. He has a long history leading teams developing products and internal infrastructure for some of the largest telecom companies in the world. After five years designing and testing cellular handsets, he spent over a decade designing and deploying test systems for mobile networks.

# 7 References


1. Ahmad et al. (2023). Measuring network quality to better understand the end-user experience [Blog post]. Cloudflare Blog. Retrieved from https://blog.cloudflare.com/aim-database-for-internet-quality
2. Zippia (2023). 25 Amazing Cloud Adoption Statistics: Cloud Migration, Computing, And More. [Zippia.com] Retrieved from https://www.zippia.com/advice/cloud-adoption-statistics/
3. Spacelift (2024) 55 Cloud Computing Statistics for 2024**.** Retrieved from https://spacelift.io/blog/cloud-computing-statistics
4. TechRadar (2022) The three most important metrics in cloud: latency, latency, latency. Retrieved from https://www.techradar.com/features/the-three-most-important-metrics-in-cloud-latency-latency-latency
5. BDAE. How Inflight Entertainment influences passengers in their travel decision-making. Retrieved from: https://www.bdae.com/en/magazine/3612-how-inflight-entertainment-influences-passengers-in-their-travel-decision-making
6. APEX committee paper (2019) Retrieved from https://www.netforecast.com/wp-content/uploads/APEX-Technical-Specification-0119_Passenger-Connectivity-v.2.pdf
7. Microsoft (2024) Microsoft 365 network assessment. Retrieved from https://learn.microsoft.com/en-us/microsoft-365/enterprise/office-365-network-mac-perf-score?view=o365-worldwide
8. Microsoft (2021) Media Quality and Network Connectivity Performance in Microsoft Teams. Retrieved from https://learn.microsoft.com/en-us/skypeforbusiness/optimizing-your-network/media-quality-and-network-connectivity-performance
9. Jones et al. (2021). Internet Connection Requirements for Video Conferencing at Home. Retrieved from https://www.netforecast.com/news/netforecasts-report-on-minimum-network-bandwidth-latency-loss-thresholds-for-acceptable-videoconferencing/
10. Vercel (2021). A DNS outage just took down a large chunk of the internet [TechCrunch]. Retrieved from https://techcrunch.com/2021/07/22/a-dns-outage-just-took-down-a-good-chunk-of-the-internet/
11. Clark (2021). What is BGP, and what role did it play in Facebook's massive outage. [The Verge] Retrieved from https://www.theverge.com/2021/10/4/22709260/what-is-bgp-border-gateway-protocol-explainer-internet-facebook-outage