ESOA
EMEA SATELLITE OPERATORS ASSOCIATION

LATENCY IN COMMUNICATIONS NETWORKS

Overview

- The Quality of Experience (QoE) of users of communications networks is affected by a number of technical and commercial metrics.
- Aspects such as **coverage, data throughput rates, resilience, security, total cost of ownership, and latency** are important considerations. Hence, latency is one of several factors for determining the QoE of a communications network.
- In many cases, the optimum solution for the most robust network will involve hybridization: multiple transmission media working in concert. This allows technologies to complement one another to deliver the best and most appropriate user experience.
- Service providers should be able to choose the technology that best supports their needs for the application(s) they are delivering to their customers. Similarly, users should be able to choose the technology that best supports the applications they are using on a regular basis.
- Accordingly, regulators should adopt technology neutral regulatory regimes that enable users to choose the most appropriate technology for their need.

Introduction

The world of telecommunications has undergone a massive transformation in the last decade. Many new technologies have been introduced and the reliance on wireless networks to transmit data is staggering. The applications that drive this explosive growth vary greatly, and this has prompted the need to establish metrics to help determine the suitability of these technologies for the applications they are intended to serve. Some of the common metrics often used to characterize communication technologies are data rates (speed), contended data rate, availability, and latency, among others.

Latency is just one requirement against which a communications network should be evaluated. For example, making 5G a reality is likely to involve multiple network technologies with varying latency ranges, working together as a 'system of systems' to enable diverse applications for different uses. One thing for certain is that latency must be examined in the context of the end-to-end performance of a given use case.

Latency in any communications network is unavoidable. Certain applications are not affected by latency, while other applications can cope with latency by using mitigation techniques to overcome its effects. Yet, there are (notably real-time) applications that are intrinsically sensitive to the effects of latency and therefore are designed to operate on low-latency networks. As a result, each application must be examined individually to determine the best technologies to meet its required usage demands.

This paper explains latency, examines the evolution of mitigation techniques to reduce latency, and considers latency's impact on different applications.

Latency Defined

The latency of a communications network is defined as the time needed to transport information from a sender to a receiver. One of the most commonly used measures of latency is the so-called Round-Trip-Time (RTT), which is defined as the time taken for a packet of information to travel from the sender to the receiver and back again. Although we speak of latency as a finite term, it actually is an accumulation of delays which occur in different segments of a network. In other words, the RTT is the total time it takes for a packet of data to travel from the sender to the receiver, across multiple hops, plus the total length of time it takes for receiver to send an acknowledgment back to the sender, through multiple hops[1].

Latency in a communications network is caused by three different elements:

- **Propagation time:** the time needed for a packet of data to travel from one place to another, which is directly related to the physical distance between the two places.
- **Routing/Switching time:** in order to be transmitted from one place to another, data has to be routed/switched through multiple links. Handing off data at routers and network interconnections can incur delays due to buffering at these intermediate routers and switches.
- **Congestion:** when a network is congested, packets of data are queued and have to wait until there is space for them to be sent. For acknowledgement based protocols like TCP, this causes retransmissions and consequently delays in packets getting to the receiver. In the worst case, packets can be dropped if there is insufficient capacity to carry them.

[1] It is worth noting that the number of hops in a satellite network is often less than the number of hops in a terrestrial network.

Latency in Communications Networks

- **Fibre networks:** Transcontinental fibre connections can have latency that reaches several hundred milliseconds, especially if the destination is physically far or is accessed through and involves multiple hops or transmitting through several hubs. Fundamental computing can also add to latency.
- **Cable/ VDSL networks:** Latency on a local cable or VDSL connection is in the range of a few tens of milliseconds, but if heavily congested, users "at the end of the line" may have a poor experience. Both the number of hops and routing/switching time can add to the latency.
- **Satellite Networks:** Satellites have the ability to link together different locations on the surface of the earth irrespective of geographical distance, bypassing the many hops of terrestrial networks. The physics of the orbit used dictates latency (RTT). In addition, because satellite networks tend to be owned or controlled by a single operator, they are well-positioned to limit latency. Latency for a geostationary orbit (GEO) is approximately 500ms; latency for a medium-Earth orbit (MEO) network is around 125ms, and latency for Low-Earth orbit (LEO) networks could be as low as 20ms (sometimes, lower than a fibre connection).
- **WiFi Networks:** The typical latency of a public Wi-Fi network is in the range of a few tens of milliseconds, but Wi-Fi networks are subject to self-interference with multiple access points sharing the same unlicensed spectrum. Additionally, due to the low levels of power permissible on Wi-Fi networks, the range is quite small (less than 100m outdoors) and the bit error rate (BER) can be high for users at the edge. The high BER leads to packet loss and retransmissions.

Accordingly, each technology, whether terrestrial or space-based, has its own limitations with regard to latency. As latency is just one factor in any determination of technology choices, users should balance this factor against other requirements, such as data throughput rates, coverage, availability, security, quality of service, and total cost of ownership, among others.

How Latency Affects the User Experience

For some applications, latency has little impact on performance or experience. For others, the impact on user experience can be significant, or even render an application unusable. Some selected examples of consumer and industrial applications are below:

Fibre Cellular Public Wi-Fi Satellite

Application	Sensitivity to Latency	Mitigation Techniques	Likely Transmission Medium
Television	Low	Some	
SCADA, and other telematics applications	Low	Some	
Streaming services	Low	Many	
Over-the-Air (OTA) updates	Low	Some	
Internet browsing	Medium	Many, very effective	
Encrypted Internet Browsing	Medium	Few	
Voice & Videoconferencing	Medium	Few	
Cloud-computing & ERP	Medium - High	Some	
High-frequency trading	Extreme	Few	
V2V and V2X	Depends on application	Few	
Mobility connectivity (In-flight/maritime/Cars /train)	Depends on application	Many	
IoT	Depends on application	Very Few	

Applications with low sensitivity to latency can benefit from some common mitigation techniques in use today. For highly latency-sensitive applications, no amount of mitigation can enable the application to work effectively. This can be the case regardless of the transmission medium.

Mitigation Techniques

While it is physically impossible to reduce propagation-related latency in any network, in many cases it is possible to minimize the undesirable effects of latency through network optimization technologies. Each technology builds on these mitigation techniques to address latency requirements. Examples of these mitigation techniques include:

- Acceleration or "Spoofing": An optimisation device on the same low-latency network of the sender, for example at a satellite ground station, acknowledges each packet instantaneously, making the sender believe that the packet has been successfully acknowledged by the receiver. The sender then immediately sends the next packet. In order to preserve reliability, the optimization device keeps track of any lost packets and requests re-transmissions on behalf of the receiver as necessary.
- Caching: Each time a user browses a new website, a copy of the site is stored locally on a hard drive. Each subsequent request for the same content is transparently delivered from the local copy instead of retrieving it over the satellite link. In addition to eliminating the latency, this caching technique greatly reduces bandwidth consumption.
- Pre-fetching: In this approach, computer algorithms dynamically predict and retrieve content that users are likely to request in the future. Much like caching, the content is stored locally, all but eliminating latency.

Some types of Internet traffic, such as encrypted data, are either unassisted or only partially assisted by these mitigation techniques, and thus will be more susceptible to the transport network latency.

Conclusions

Several metrics can be used when assessing the suitability of a communication technology to a given application. Metrics such as **data throughput rates, coverage, resilience, security, latency** and **cost of ownership,** are examples of such metrics. Latency is an important but not the sole metric, and should always be considered in the context of the end-to-end link performance. In many cases, the optimum solution for the most robust network will involve hybridisation: multiple transmission media working in concert to ensure that applications are delivered over the most appropriate path. In such a way, technologies can complement one another to deliver the best and most appropriate user experience. Accordingly, service providers and users must have the flexibility to determine the best technology for the application(s) and services they are using. Accordingly, technology neutrality must govern any regulatory regime in order to ensure that users have the ability to address their individual network needs.