# 5.1t Transcript

What is QoS and QoE

Maintaining the quality of our networks has become increasingly difficult, Cloud computing, big data, the ubiquitous use of mobile devices on enterprise networks, and the increasing use of video streaming all contribute to the increasing issues we have ensuring satisfactory network performance. The various types of traffic that are produced differs and therefore will need to be handled differently.

If we look at what we can consider to be the three main types of traffic, Voice, video, and data traffic these can be differentiated by the various transmission characteristics:

For Voice—Traffic flows with a regular pattern at a constant rate that is sensitive to delay and delay variation. When compression techniques are in use, voice traffic is more sensitive to error than uncompressed voice.

For Video—Real-time video traffic has similar transmission characteristics to voice traffic, but also requires high bandwidth. When compression techniques are in use, video traffic is more sensitive to error than uncompressed video.

For Data—Traffic flows with an irregular pattern that is often called bursty because of its variability in rate and amount of traffic. Data traffic is not sensitive to delay or delay variation, but it is sensitive to error.

Prior to Ethernet becoming the dominant technology there were protocols that could deliver quality of service, one such protocol was ATM, Asynchronous Transfer Mode.

The ATM Traffic Management specifications define several service categories to group traffic according to different transmission characteristics and performance needs. Each ATM service category is qualified by certain traffic parameters and QoS parameters that define the desired network performance for the permanent virtual circuit (PVC) or switched virtual circuit (SVC) on the ATM network.

However, Ethernet has none of these service categories, it is by its design a best effort system. This doesn’t mean we don’t have to worry about quality. In fact, in modern networks there are Service Level Agreements, SLA’s between user and provider, to ensure network performance is maintained.

How network performance is achieved is by measuring it, the two-key metrics for this are Quality of Service, QoS and Quality of Experience, QoE.

Quality of service is the measurable end to end performance of the network and would be guaranteed in advance in the SLA.

 Quality of Experience is a subjective measure of performance as conveyed by the user. This is obviously opinion based, so we must be careful when using it.

QoS and QoE enable us to determine whether the network is meeting user needs and to diagnose problem areas that require adjustment either to network management or network traffic control, or possibly both.

I’ve mentioned the Ethernet is a best effort delivery service, let’s just quantify what this actually means.

Our networks attempt to allocate its resources with equal availability and priority to all traffic flows, this sounds like a sensible notion. But with no regard for application priorities, traffic patterns and load, and customer requirements. This was historically less of an issue, but if the finance department is working on year-end figures and the network is running very slow because someone is taking all the bandwidth watch cat videos on YouTube then we now have a problem.

To protect the network from congestion collapse and to guarantee that some flows do not crowd out other flows, congestion control mechanisms were introduced, which tended to throttle traffic that consumed excessive resources.

There have been attempts to resolve the issues of QoS in IP networks, remember back in week one we discussed the OSI, the Transport and session layers, the Session Layer Controls the conversations between different computers. Also responsible for session control of communications be on the transport layer.

Transport Layer is responsible for segmentation of packets into smaller chunks. Provides reliable (connection orientated) and unreliable transfer protocols (connectionless).

Basically, this is the TCP congestion control mechanism.

TCP congestion control has become increasingly complex and sophisticated, but it is worth discussing it.

For each TCP connection between two end systems across a network, in each direction, a concept known as sliding window is used. TCP segments on a connection are numbered sequentially, our sequence numbers.

The sending and receiving TCP nodes maintain a window that defines the range of sequence numbered segments that may be transmitted. As segments arrive and are processed by the receiver, the receiver returns an acknowledgment indicating which segments have been received and indicate to the transmitter that the window of sequence numbers has advanced to allow more segments to be sent.

Various algorithms are used by the sender to deduce the amount of congestion, one such algorithm is called Slow Start.

With TCP slow start, the window size will initially grow exponentially (window size doubles) but once a packet is dropped, the window size will be reduced to one segment. It will then grow exponentially again until the window size is half of what it was when the congestion occurred. At that moment, the window size will grow linearly instead of exponentially.

When an interface gets congested, it’s possible that all your TCP connections will experience TCP slow start. Packets will be dropped and then all TCP connections will have a small window size. This is called TCP global synchronization.

Although systems like Slow Start and other network congestion control techniques can reduce the risk of excessive congestion, these techniques do not directly address QoS requirements.

As the intensity and variety of traffic increased, various QoS mechanisms were developed, including Integrated Services Architecture (ISA) and differentiated services (DiffServ), we’ll discuss these in the next video.