Engineering Quality of Experience:

A Brief Introduction

Neil Davies and Peter Thompson

November 2012

Connecting the quality of user experience to parameters a network operator can directly measure and control is a challenge, but one that is of fundamental importance to the successful, efficient and ultimately sustainable operation of packet-switched networks. This paper presents a simple, scientific, solution to this problem based on a quality, rather than bandwidth, centred approach. *Quality attenuation* is a property that can be directly measured end-to-end across a network, and can also be tied to application outcomes. This permits a systematic approach to delivering good QoE called AREA^{TM}, connecting: user/application Aspirations; network performance Requirements; traffic Execution; and outcome Assurance.

Contents

List of Figures

List of Tables

1 Introduction: Users, Applications and Networks

End users have no interest in networks: they are interested in experiences. These are provided by applications, implemented as a collection of clients and servers, connected by the network. Thus, for the user, the network is a price they must pay to obtain the experience; for the application, the network is a necessary evil that impedes its performance; and for the network operator, designing, dimensioning, configuring and operating their network to satisfy the needs of a plethora of ever-changing applications is a major challenge. The connection between parameters the operator can measure or control and the quality of experience for the end-user seems tenuous at best.

In this paper we show how this gap can be closed using the concept of *quality attenuation*, which captures the impact of the network on the traffic streams required by the application. Quality attenuation is a property of the end-to-end path between different components of the application (usually the client and server) that can be measured, and hence managed, by the operator. The sensitivity of an application to quality attenuation can also be determined, and while the range of applications is vast, the range of protocols they use to connect their distributed components is not. Thus, understanding how quality attenuation affects the performance of common protocols is sufficient to determine how network performance impacts end-user QoE.

2 Measuring Quality Attenuation

Since every network element introduces some delay in the delivery of packets, and some a probability of packet loss, neither of which can be undone, the overall effect of the network is to degrade the stream of packets between one component of the application and another. We call this *quality attenuation*, and it imposes a fundamental limit on what the network is able to deliver, just as noise does in an analogue system. No matter how well an application is constructed, there will be a level of quality attenuation above which it cannot function acceptably. The worst-case quality attenuation requirement of an application is fundamentally related to time, in particular the timescales over which it responds to the transport of its packets. This will vary from one application to another, so the only common denominator over *all* applications is the set of *instantaneous* transport characteristics that the network delivers, rather than any average; we call this 'ΔQ'. Measuring ΔQ requires more sophistication than typical network performance measures for several reasons:

1. Quality attenuation must be measured as a distribution, not an average;
2. Each direction of packet travel must be measured separately, not as a round trip;
3. It must be measured end-to-end, not just on individual network segments or elements;
4. It needs to be analysed into components that depend on the loading of the network and those that do not.

This is all achievable with appropriate skill and care, however, and doing so frequently reveals sub-optimal aspects of the network topology or configuration.

3 Customer Experience

Customers' applications can deliver "good experiences" only when they are provided with sufficient quality in sufficient quantity for their needs. Each application has a tolerance for loss and delay, and so has an end-to-end quality attenuation budget.

While the wide range of applications may seem daunting, in practice there are only a handful of different types of requirements, as illustrated in table 1. Each application will have a region of 'loss-delay space' in which it can meet its end-user expectations. Figure 1 is representative of the boundaries that, if crossed, start to have a customer-impacting effect. This gives bounds on measurable network properties to manage the network against – an instantaneous quality attenuation (ΔQ) budget. It is entirely feasible to run sections of the network against fixed portions of this end-to-end budget, as discussed in section 4.

Table 2 shows in general terms the impact of quality attenuation on the performance of TCP, and figure 2 shows how this impacts a typical http download.

Relating this to a direct end-user experience such as the time to open a typical Facebook page requires a model of the interaction between the client and the server, which in this case would be:

- The number of items to be downloaded (e.g. number of pictures on the page);
- The number of concurrent http streams supported (typically a function of the browser used).

Application	QoE Metric	Principle of Operation	Behaviour Group	End-to-end requirements
Media "streaming" (iPlayer, YouTube)	Time to first frame and probability of "pause" per hour	Block transfer using HTTP over TCP/IP – blocks transferred "just in time"	Stop-Start bulk transfer	iPlayer HD: Throughput > 5Mbps Loss < 0.5% Delay < 300ms
Web Browsing simple pages	Time to complete	Use HTTP/TCP/IP to transfer small data block. Slow Start dominates	Single-shot remote procedure call	Throughput >75kbps Delay < 200ms Loss < 1.5%
VoIP (Sip, UC, Skype, Google Talk)	Perceived call quality (c.f. MOS/PESQ)	Fixed flow of same sized packets, "inelastic" load	Inelastic media (loss-tolerant)	Throughput: 80kbps Loss <2.5% Jitter <60ms Delay <150ms
Gaming (multiplayer, real-time online)	Response time	Clients connected to servers, real time interaction	Non-TCP based realtime, lossy, protocols	Throughput >1Mbps Loss < 0.5-1% Jitter <5-10ms Delay < 40-80ms

Table 1: Example Requirements by Application Type

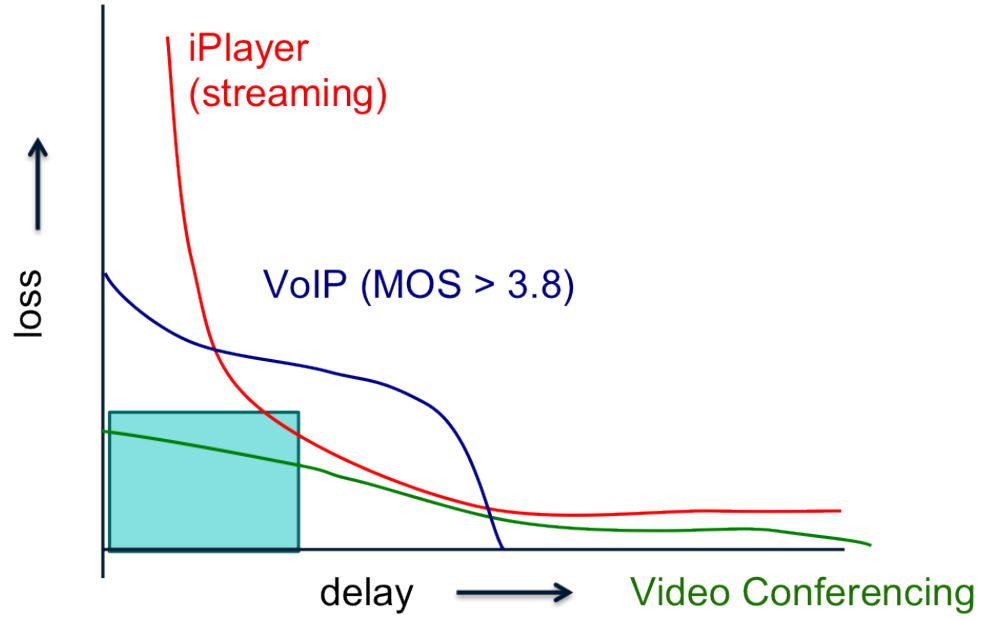

Figure 1: A Schematic Representation of Application Behaviour for Varying Loss and Delay

TCP State	*Effect of Loss*	*Effect of Delay*
Connection Establishment	Retransmission resulting in longer time to connect	Longer time to connect, in extreme cases cause retransmission
Slow Start	Single loss will just cause retransmission and hence delay; multiple losses will re-start the slow-start process	Will take longer for slow-start to complete; when streaming, will take longer to increase the sending rate
Congestion Avoidance	Single loss will cause delay and retransmission; multiple losses will cause slow-start to begin again.	If bulk transfer should not be noticeable, otherwise data takes longer to arrive
Connection Termination	Will take longer to close connection, retransmissions, may result in FIN_WAIT_2	Takes longer to close the connection down

Table 2: TCP Behaviour under Quality Attenuation

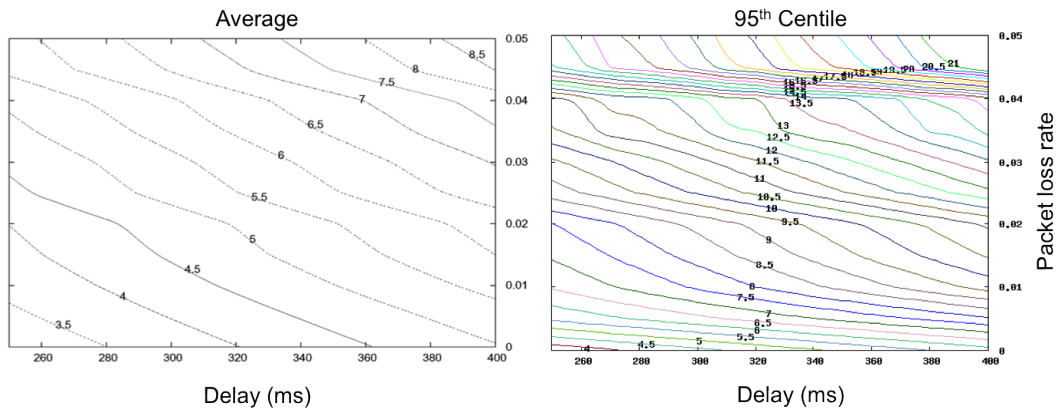

Figure 2: HTTP Completion Time as a Function of Quality Attenuation

These can then be combined with the response of http to quality attenuation in figure 2 to obtain the distribution of page load times. Working backwards from this enables the maximum quality attenuation that allows X% of page loads to complete within Y seconds to be calculated. Managing the quality attenuation to meet this target then guarantees the end-user experience up to factors beyond the network operator's control.

4 Managing Quality Attenuation

Since quality attenuation is the critical aspect of network performance for determining application outcomes, the key question for a network operator is how to ensure it remains within bounds, at least for those traffic streams deemed important. A typical approach is to dimension the network so that congestion is rare and so quality attenuation is bounded. Maintaining this situation as traffic loads rise requires an upgrade planning process; basing this on direct measurements of ΔQ rather than imprecise 'proxy factors' such as average loading enables the operator to target their investment where it will actually benefit end-user QoE.

Another method is to exploit QoS features of network equipment to manage or prioritise traffic streams so as to mitigate the impact of congestion on those streams, but in practice the benefit of doing this rarely justifies the cost and complexity of implementing it. This is because QoS mechanisms have no impact until congestion occurs, and in this case the 'protected' streams are liable to be congested themselves (and hence subject to quality attenuation) unless they comprise a very small proportion of the overall traffic. Thus, the standard approaches to QoS deployment are, typically, self-defeating, since they only create additional "value" for a set of conditions that are both rare and transient.

Other approaches exist, based on the concept of a Quality Transport Agreement as discussed in section 5.5 below. In delivering a QTA the operator can establish a budget for different sections of the network, as illustrated in figure 3.

5 Summary: The AREA Approach

What does it mean for a network to 'work'? What is it that the users of networks really want, and how can networks provide it? Moreover, how can the operator of a network demonstrate that they have met their users' needs in order to justify charging for this provision?

Using the concept of quality attenuation, these questions can be resolved within a consistent framework that distinguishes the users' *aspirations* from their *requirements* of the network, and the network's mechanisms for *ensuring* that these requirements are met from those that *assure* that they were; this is the AREA^{TM} approach.

5.1 Aspiration

The user's *aspiration* is for some application function to work correctly, typically within some time bound. This might range from loading a web page across the Internet to receiving GSM service; whatever it is,

the details of the network interactions that make it possible are of no interest, providing it works with a satisfactory level of performance.

5.2 Requirement

In order for an application to perform satisfactorily, the packet flows that it generates must be transported by the network with no more than a certain level of quality attenuation, which can be quantified either by mathematical analysis of the protocols involved or by running the application against a credible network simulator. So the user's aspiration for the application generates (often implicitly) a *requirement* from the application for some level of service from the network[1].

5.3 Execution/Ensurance

Networks can (and often do) make no particular provision to ensure that applications' requirements are met, which makes it all too likely that they won't be, depending on the collection of demands on the network at the same time. Clearly, ensuring that the varying requirements of a wide range of applications for a large population of users are all met simultaneously is no simple matter. A mechanism that attempts to make sure that the quality attenuation requirements of multiple applications are met we call network *ensurance*.

5.4 Assurance

Applications often fail for reasons that are nothing to do with whether the network meets the requirements placed upon it, often when client or server resources are inadequate. Network operators need to measure whether the quality attenuation they deliver meets requirements, both to check that the network and its ensurance mechanism are functioning correctly and to assist in locating user-affecting problems.

5.5 Quality Transport Agreements

A Quality Transport Agreement (QTA) provides a way for the network operator to satisfy the needs of multiple users/applications. It has two key aspects:

1. Specification of demand: what will be the offered load[2], and what constraints apply to the arrival pattern of the packets.
2. Specification of service: the minimum requirement of ΔQ.

Both parts are essential: without any constraints on the offered traffic, the network cannot guarantee to meet the requirements of even one user, let alone many; and without a specification of the service, no application can be sure to meet its aspirations. Provided the overall set of QTAs matches the physical

[1] These requirements may be unrealistic, which is one common reason why an application that works in the lab may fail in the field.

[2] Including the specification of the entry and exit points of the network between which it should be transported.

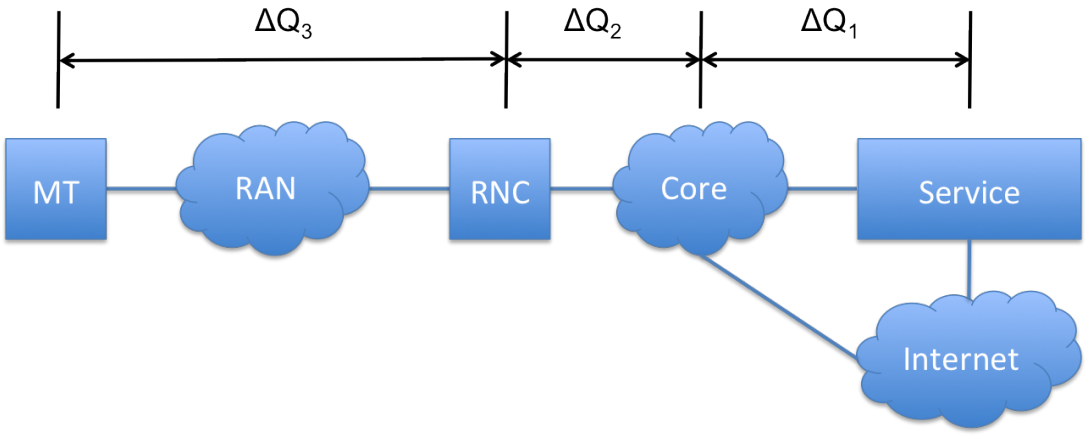

Figure 3: End to End Quality Attenuation Budgets

and technological capabilities of the network, an effective ensurance mechanism will deliver the expected ΔQ of every application that doesn't exceed its specified demand.

A complete contract between a network operator and a customer might include several QTAs for different applications and types of traffic, together with an SLA specifying how rarely the QTAs should be breached, circumstances in which they might be abrogated, the assurance processes in place, and so forth.

6 Conclusion

Connecting the quality of user experience to parameters a network operator can directly measure and control is of fundamental importance to the successful, efficient and ultimately sustainable operation of packet-switched networks. Fortunately, there is a simple, scientific, solution to this problem based on a quality, rather than bandwidth, centred approach. *Quality attenuation* is a property that can be directly measured end-to-end across a network, and can also be tied to application outcomes. This permits a systematic approach to delivering good QoE called AREA$^{\text{TM}}$, connecting: user/application Aspirations; network performance Requirements; traffic Execution; and outcome Assurance. Using this approach, operators can save unnecessary capacity upgrade costs while delivering superior service, and different parts of the communications delivery chain can establish clear contractual relationships using implementable Quality Transport Agreements, so that application outcomes can be assured at reasonable cost.