MARKETING REPORT

MR-452.2

Use of ΔQ to manage customer SLA

Issue: 1
Issue Date: July 2021

Issue History

Issue Number	Approval Date	Release Date	Issue Editor	Changes
1	8 July 2021	8 July 2021	Jonathan Newton (Vodafone)	Original

Comments or questions about this Broadband Forum Marketing Report should be directed to info@broadband-forum.org.

Editor Jonathan Newton, Vodafone

Work Area Director(s) David Sinicrope, Ericsson

Contributors Jonathan Newton, Vodafone
Peter Thompson, Predictable Network Solutions Limited (PNSol)
Bjørn Ivar Teigen, Domos

Table of Contents

Executive Summary

The Broadband Forum has been instrumental in driving the Broadband industry to improve speeds, resulting in the increasingly widespread availability of Gigabit broadband connections, which underpin the 'Gigabit Society'. However, in today's Gigabit broadband era, we are now experiencing diminishing returns in terms of the ability of further speed increases to ensure an application's effective performance. MR452.1[1] introduces the concept of 'Quality Based Broadband', describing why quality should become a first- class objective in broadband delivery, and how the 'Quality Attenuation' measurement and analysis framework can enable this transition.

In today's networks, service providers and their customers can have a hard time identifying the underlying network performance required to deliver a predictable outcome for an application. This becomes even more complicated when we consider that there may be many applications in use simultaneously, and that any number of short-lived or ongoing artefacts of network performance may influence each of these applications in different ways.

In order to support the move to 'Quality-Based Broadband', we need a simple and clear way of identifying the required network quality for any application, and a way to prove whether a broadband service is delivering against this requirement. For this purpose, the Quality Attenuation framework defines the concept of a Quantitative Timeliness Agreement or 'QTA', which can be regarded as a set of requirements to relate the network and application domains.

By combining measurement of the Quality Attenuation with a QTA for each type of application in use, it is possible to create a service level agreement that relates network performance directly to the application outcomes, allowing customers and service providers to have a clear and provable agreement on the service requirements.

1 Introduction

It is quite reasonable for a customer to expect a certain level of service from any provider; this service level may be assumed (perhaps based upon past service experience) or may be explicitly specified in contractual terms. The likely end result of a service provider delivering below this level is customer dissatisfaction and either explicit or implicit consequences, such as an increased probability that the customer will leave and find a new provider. It is therefore in the interests of both parties that they have a common understanding of the minimum required service level. However, this is not easy to achieve for packet networks because the service provider is working at the packet layer whilst the end user 'sees' poor network performance through the impact that it has on the application that they are using.

The Broadband Forum initiative called Quality Experience Delivered (or QED) is documenting an approach known as Quality Attenuation (or ΔQ, pronounced “Delta Q”), which includes a language for defining these service levels. ΔQ provides a mathematically robust performance measurement and analysis technique that integrates both packet loss and packet delay instead of treating them separately. As described later in this document, the ΔQ framework also allows the specification of the minimum packet-level performance required for an end application to deliver a good experience. This specification is contained in the Quantitative Timeliness Agreement, or QTA.

For a QTA to be deliverable, there must also be a bound on the offered load from the customer, which will be included within the QTA specification, but may be implicitly derived from some other part of the customer service specification, such as the interface speed. Without such a bound, it would be impossible for a service provider to commit to delivering the ΔQ of the QTA.

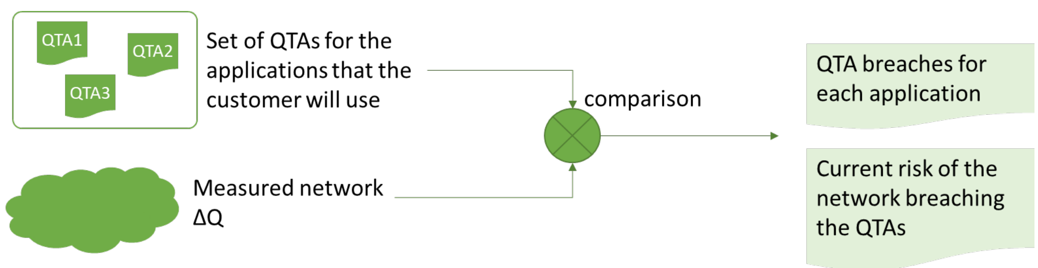

Figure 1 - Comparing a set of QTAs with measured network performance

As shown in Figure 1, once an application's requirements have been defined in the form of a QTA, it is then possible to understand whether or not the application will be negatively affected by the network performance purely from observations of the ΔQ at the packet level. Importantly, it is also possible to identify how close the network performance is to breaching the required QTA(s), which is invaluable information for a service provider who wants to ensure that their network is built at the right size to deliver consistent customer application performance.

A possible further step is to use a set of QTAs to derive packet scheduling policies that can ensure that the QTAs will be met as shown in Figure 2.

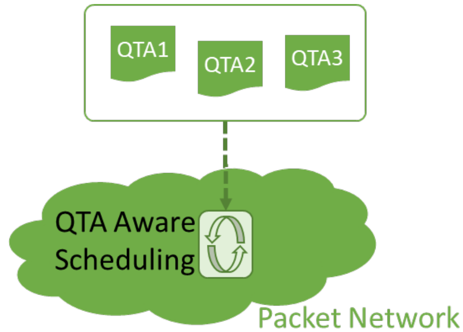

Figure 2 – QTA aware Scheduling

The ‘application’ that a QTA refers to may be an internal service, such as routing updates or other control traffic needed for network stability. It may be a subsidiary service such as a third-party bearer; such inter-operator QTAs provide a measurable criterion for satisfactory service, which the supplier can use to dimension and manage their service, and the consuming operator can use as a foundation for delivery of QTAs to their customers.

2 Representing Quality Attenuation as a Cumulative Distribution

'Performance' is typically considered as a positive attribute of a system. However, a 'perfect' system would be one that responds without error, failure or delay, whereas real systems always fall short of this ideal; we can therefore say that the quality is *attenuated* relative to the ideal. We denote quality attenuation by the symbol ΔQ and reformulate the problem of managing performance as one of maintaining suitable bounds on ΔQ. This is an important conceptual shift because 'performance' may seem like something that can be increased arbitrarily, whereas ΔQ (rather like noise) may be minimized but never eliminated.

On this basis, the performance of the network can be represented statistically as an improper Cumulative Distribution Function (CDF) showing the time taken to deliver a packet from point A to point B (improper because it generally never reaches 1 as some packets are not delivered due to packet loss or error). See the ΔQ in Figure 3.

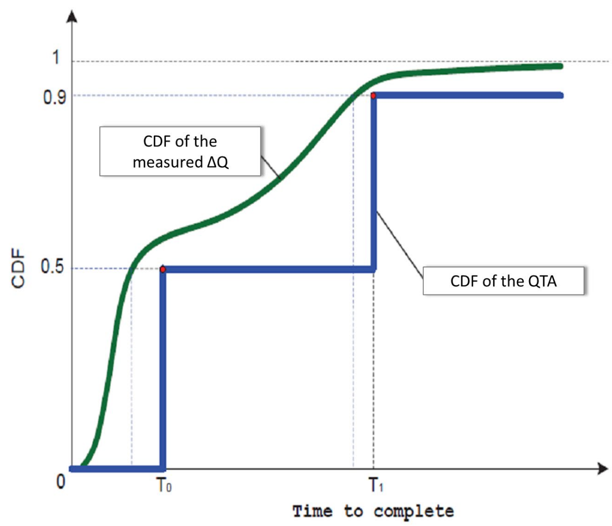

Figure 3 - Cumulative distribution to represent quality attenuation

3 The Quantitative Timeliness Agreement

A QTA can be regarded as a set of requirements or even as a two-sided ‘contract’ to relate the network and application domains. It is a way of expressing an application’s packet delay and loss requirements in a single integrated way (via a CDF) and relating this to measured network ΔQ performance. ΔQ effectively unifies a continuously variable quantity (e.g., packet delay) with discrete events (e.g., packet loss). Application performance requirements are then expressed as a bound on ΔQ for a given load, e.g., 50% of packets to arrive within 50ms, 95% within 75ms, etc. Every application has some level of ΔQ beyond which it will deliver a poor User Experience (UX). This varies significantly between applications.

The challenge is to quantify timeliness rather than simply rates. This is key for any individual end-user. As an example, let us suppose we could agree how long it’s OK to wait for the first frame of a chosen video to start, as described by the following video application specification:

- 50% of responses within 5 seconds
- 95% of responses within 10 seconds
- 99.9% of responses within 15s
- 0.1% packet loss ratio

We can represent this using an 'improper' CDF as shown in Figure 4, which represents the worst-case observable behavior that would still fulfill the requirement.

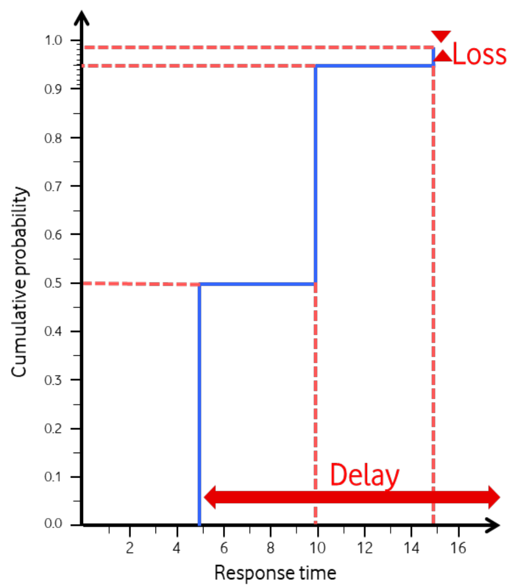

Figure 4: Example QTA

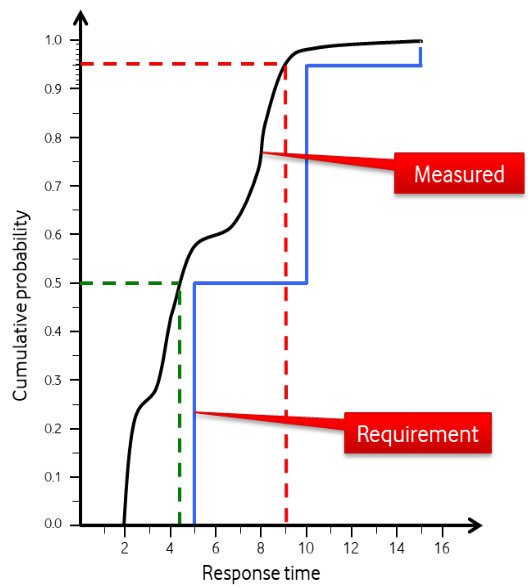

Figure 5: Satisfying a QTA

Having captured the application requirement as a QTA specification, we can then use it as a basis to compare with measured results as shown in Figure 5.

Provided the measured ΔQ CDF is at or above the QTA CDF, the network is meeting its QTA obligations. Furthermore, it is possible to evaluate how close the system is to breaching the QTA, or by how far it has breached the QTA by measuring the distance between the required and measured CDFs.

3.1 An example QTA for Voice over IP

Figure 3 in TR-452.1[2] provides a plot of VoIP quality as a function of jitter and packet loss. Toll quality voice corresponds to a loss of < 2% and jitter of < 20ms. Since VoIP packets are a uniform size, there is a straightforward relationship between jitter and $\Delta Q_{|V}$ (see TR-452.1[2] for details). If we combine this with a plausible one-way maximum delay of 150ms, this gives an overall QTA ΔQ part as shown in Figure 6.

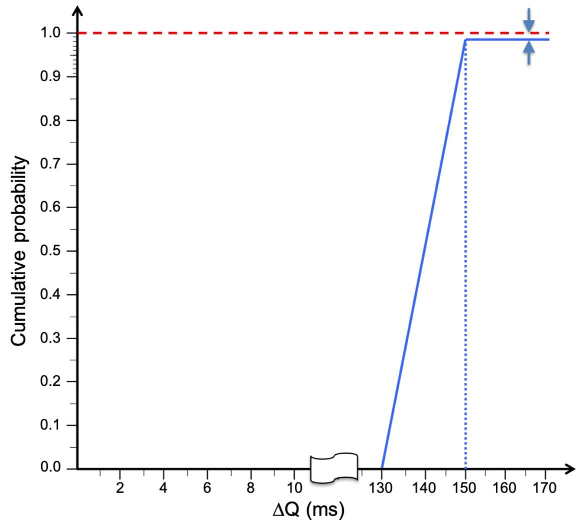

Figure 6 - Example QTA for VoIP Traffic

This shows that the maximum delay can vary between 130 and 150ms with a 2% loss probability.

The other side of the QTA is the load specification, which for a VoIP stream is typically very simple. For example, a fixed-sized packet is sent every 20ms.

4 Example Use Cases

As discussed above, if a network service delivers traffic within a QTA, this will ensure that the network component of delivered user experience is satisfactory. The following are examples where QTA budgets for segments of the end-to-end path can be used for design and assurance purposes.

4.1 QTA-based SLAs for Business Customers

We have already described how Quantitative Timeliness Agreements (QTAs) relate ΔQ budgets to application requirements and outcomes. This approach to expressing requirements could potentially be incorporated into a Service Level Agreement (SLA) for business customers. Characterising the options for customer connectivity technology may also help the service provider or network operator to understand what parameters in an SLA (e.g., for business services) may be safe to offer or which technologies may pose a risk to achieving certain SLAs.

4.2 QTA for Wholesale Fixed Access

This use case is between two network operators, one wholesaling connectivity to the other.
The analysis of typical wholesale product specifications shows the presence of network performance hazards, both for network control traffic (affecting system stability) and end-user traffic (affecting application outcomes). Mitigating these hazards is one reason for the deployment of traffic management in broadband infrastructure.

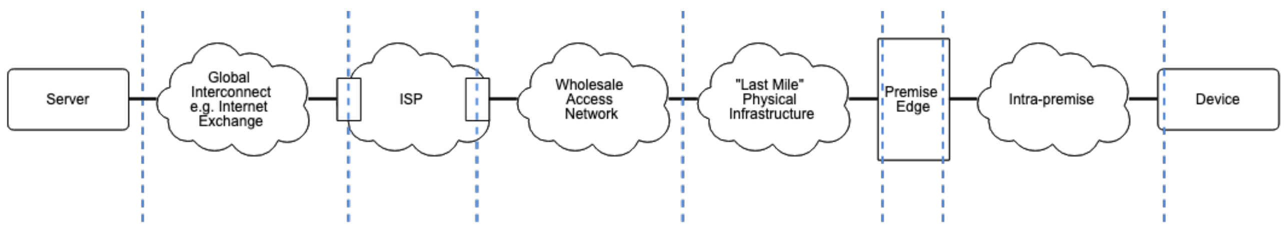

Figure 7: End to end digital delivery chain

Figure 7 shows how the end-to-end supply chain can be segmented and that QTAs can be applied to each segment. The end-to-end ΔQ is simply the sum of the ΔQ for each segment. Provided the QTAs for all segments are satisfied, then the overall QTA for the entire chain will also be satisfied.

5 Relating QTAs to an SLA

A Service Level Agreement (SLA) is a contract between a service provider and its customers. The SLA defines what service the service provider promises to deliver, and how reliably the service provider must deliver that service. As an example, a service provider may promise to deliver a 100 Mbit/s service 99.9% of the time. This means the average throughput measured over a second must be at least 100 Mbit/s 99.9% of the time.

Although typical SLAs might include delay and loss, this will generally be averaged over a substantial period. However, the short term (sub-second) delay and loss can have a significant impact upon the user experience.

If an SLA is defined only in terms of throughput and long-term average loss and delay, it is possible for a service provider to successfully deliver 100Mbit/s while the connection appears unusable from the perspective of the customer due to excessive short term latency, delay variation or loss. By contrast, a QTA quantifies how much delay or loss packets can experience at any point in time.

A QTA can apply to a specific subset of the traffic, for instance it is possible to have one QTA for video conferencing and another for less urgent video streaming. Together, a set of QTAs (including the peak customer traffic load) along with when they apply can form the foundation of an SLA.

6 Conclusions

The problem of how to specify a satisfactory service has not been fully solved since the introduction of statistically multiplexed packet networking.

The Quality Attenuation approach of the Broadband Forum's Quality Experience Delivered initiative provides a bridge between what can be measured at the network level and the application user experience (UX). Every distributed application implicitly imposes a limit on the Quality Attenuation (ΔQ) it can withstand between its components (such as a server and client) and still deliver acceptable UX. Making this limit explicit defines what service is sufficient at the network level for the application to deliver satisfactory UX. The fact that ΔQ 'adds up' also means this limit can be devolved into 'budgets' applying to different network segments, including third-party delivery chain suppliers.

Combining the limit on acceptable ΔQ with a bound on the load the application can submit to the network creates an enforceable 'contract' called a Quantitative Timeliness Agreement (QTA). A QTA is measurable by both the supplier and consumer of the network service and provides a way for operators to deliver a more valuable service, and for users to have confidence in the UX of their applications. Together, a set of QTAs, along with when they apply, can form the foundation of an SLA.

7 References

The following references are of relevance to this Marketing Report. At the time of publication, the editions indicated were valid. All references are subject to revision; users of this Marketing Report are therefore encouraged to investigate the possibility of applying the most recent edition of the references listed below.

A list of currently valid Broadband Forum Technical Reports is published at www.broadband-forum.org.

Document		Title	Source	Year
[1]	MR452.1	*Motivation for Quality Verified Broadband Services*	Broadband Forum	2019
[2]	TR452.1	*Quality Attenuation Measurement Architecture and Requirements*	Broadband Forum	2020

End of Broadband Forum Marketing Report MR-452.2