A Utility-based QoS Model for Emerging Multimedia Applications

Mu Mu, Andreas Mauthe  
Computing Department, Lancaster University  
Lancaster, United Kingdom  
{m.mu, andreas}@comp.lancs.ac.uk

Francisco Garcia  
Agilent Laboratories  
Edinburgh, United Kingdom  
frankie_garcia@agilent.com

Abstract

Existing network QoS models do not sufficiently reflect the challenges faced by high-throughput, always-on, inelastic multimedia applications. In this paper, a utility-based QoS model is proposed as a user layer extension to existing communication QoS models to better assess the requirements of multimedia applications and manage the QoS provisioning of multimedia flows. Network impairment utility functions are derived from user experiments and combined to application utility functions to evaluate the application quality. Simulation is used to demonstrate the validity of the proposed QoS model.

1. Introduction

For traditional internet applications, the quality of network delivery is not critical since applications like web browsing and email are elastic and can tolerate certain network impairments. However, with increasing bandwidth more high-throughput, always-on, inelastic multimedia applications are introduced to the IP infrastructure. Guaranteed quality of end user’s experience in multimedia services is a challenge to traditional network QoS measurement and management models.

This paper proposes a utility-based QoS model as an extension of existing network QoS mechanisms. This QoS model is driven by utility functions. Network impairment utility functions represent the network QoS dimensions (e.g. delay and packet loss). These empirical functions are derived from experiments. Network impairment utility functions are combined to form application utility functions to model the user experience in a specific application context. Our discussion mainly focuses on utility representation modeling user experiences and how these application utility functions can be used to support the evaluation of service quality.

The paper is structured as follows: In Section 2, existing network QoS models are discussed. The utility model and utility functions are introduced in Section 3. Subsequent experiments and an evaluation are presented in Section 4. Section 5 concludes the paper.

2. Related works

Previous studies on network QoS models are mainly concerned with resource reservation and allocation algorithms to guarantee the availabilities of resource for internet services.

2.1. Integrated Services

Integrated Services (IntServ) [1] uses a flow-based model coupled with a signalling protocol (i.e. RSVP [2]) along the packet path. The signalling protocol guarantees that adequate resources are reserved on the hops along the flow path. However, all the network nodes have to keep and exchange the state information of ongoing flows. Because information is required for each flow, the quantity of information increases proportionally with the number of flows.

2.2. Differentiated Services

Differentiated Services (DiffServ) [3, 4] uses packet label to classify priorities of each packet. Each packet is processed according to its label within the specified priority class. Because traffic is processed within a specific pre-defined class, DiffServ is more scalable than IntServ. The amount of state information required in network nodes is constrained by the number of classes. However, the performance of the model is affected by the quantity of predefined service classes. If only few service classes are defined, system complexity is reduced but more conflicts will appear within each class.

2.3. Parametric model

Partridge[5] suggests the “flow specification” data structure to describe the delivery requirements of each
flow. Network nodes on the transmission path can read the thresholds information (e.g. “Minimum Delay Noticed=0/10/100 [100 ms]”) from the specification signal to guarantee the requirement of each flow. For traditional IP services, the fine tuning between upper and lower threshold is not critical since most of the services are light-weighted and elastic. However, this scheme does not consider characteristics between thresholds nor does it allow for inter-flow optimization.

3. Utility model

The impact of the different QoS parameters (delay, jitter, packet loss and bandwidth limit) on different services have been widely studied. As described in section II, most of the results of this research only use discrete values as threshold parameters. Usually only two thresholds are defined. QoS models using threshold parameters are by nature binary (i.e. only making provisions for cases falling outside). However, variation properties between thresholds can have significant impact on multimedia services.

In order to provide a better model for the properties of multimedia flows, utility functions are introduced. Utility functions are grouped into two types. Network impairment utility functions model the impact of specific QoS dimensions on the user perception. Network impairment utility functions represent the utility between thresholds as well as the change of the utility at the transition points (i.e. above and below the upper and lower threshold). Application utility functions are introduced to evaluate the user experience by considering results from sub network impairment utility functions (details in 3.6). The application utility can be then normalized to the MOS scale to express the user experience of an application.

VoIP, video streaming and online gaming are three typical emerging rich applications. Different multimedia services can be modeled as combination of these three types. Utility functions of VoIP, video streaming and online gaming are explored in detail in the following sections. As mentioned in [6], traditional QoS dimensions can easily be defined by numerical utility functions whereas others, such as video codec and picture quality are often suitably expressed in a non-numeric, non-uniform way. We propose a mapping mechanism between the quality space and numeric utility rating space for the QoS dimensions which are not suitable to be modeled solely by mathematical function.

3.1. Thresholds of QoS dimensions

The following threshold values for VoIP, video streaming and online gaming applications are based on user studies in formal literature. The minimum and maximum thresholds are summarized in Table 1. Minimum threshold identifies the value from which the user will experience certain network impairment. The user experience is predicted to be unacceptable when the maximum threshold is reached. An empty entry indicates that this parameter has no impact on the application.

<table>
<thead>
<tr>
<th>Application</th>
<th>Delay (ms)</th>
<th>Jitter (ms)</th>
<th>Loss (%)</th>
<th>SRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VoIP</td>
<td>100</td>
<td>150</td>
<td>40</td>
<td>75</td>
</tr>
<tr>
<td>Video Streaming</td>
<td>200</td>
<td>1000</td>
<td>100</td>
<td>500</td>
</tr>
<tr>
<td>Online Gaming</td>
<td>50</td>
<td>150</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Thresholds for bandwidth are not listed in Table 1 because they are decided by the encoding rate or frame rate of the application. More details about these thresholds will be discussed in following sections and can also be found in [15].

3.2. Impairment Utility Function for VoIP

VoIP is a popular service in IP network, besides conversational voice, network radio and voice message are main instances of VoIP applications. Conversational voice is more vulnerable to delay, jitter and loss than other non-interactive voice services.

3.2.1. Influence of delay. Referring to the model defined in [16] for users with strong interactivity requirements, we derive the following generic utility function for audio flows. The function has three regions reflecting the average tolerance level of humans to distortions.

$$
\mu_{dc} = \begin{cases} 
100 - y_1 \chi & \chi < c_1 \\
\delta - y_2 \chi & \chi > c_2 \\
\beta \tanh(\beta(\chi - b_2)) + b_3, & c_1 \leq \chi \leq c_2 
\end{cases}
$$

(1)

In (1), the delay value $\chi$ is given in milliseconds, $y_1, b_1, y_2$ are parameters which represent the properties of quality degradation in each region. $c_1$ and $c_2$ are the lower threshold (the borderline of the first region and the second region) and the upper threshold (the borderline of the second and the third region) of the application. $b_1, b_2, b_3$ are tuning parameters to ensure the continuity between regions. With the empirical thresholds in Table 1, the delay utility function of a conversational VoIP application can be derived (2).
Regarding the effectiveness and efficiency of the network, the optimal delay for a VoIP audio flow should lay around the borderline of Region I and Region II in Figure 1(a).

3.2.2. Influence of jitter. The human ear is very intolerant to short-term delay variations (jitter) [11]. A hyperbolic tangent is chosen to model the jitter utility because the upper/lower thresholds and transition between them can be easily modeled. The generic jitter utility function is then defined as:

$$\mu_j = b_1 \tanh(\beta(x - b_2)) + b_3$$  \hspace{1cm} (3)

Considering the jitter thresholds of Table 1, an instance of delay utility function for G.711 VoIP flow is defined as:

$$\mu_j = -30 \tanh \left( \frac{x - 50}{15} \right) + 70$$  \hspace{1cm} (4)

3.2.3. Influence of loss. During a voice conversation one packet loss may make a word unrecognizable, and a single word can be the key word to the whole sentence. From [18-21], we derive the generic utility function of packet loss as:

$$\mu_l = b_1 + b_2 \log(b_3 + \beta x)$$  \hspace{1cm} (5)

In the generic function above, \( x \) stands for packet loss in percentage; \( \beta \) is the steepness of the quality degradation when packet loss increases. A loss utility function for an audio flow that fulfills the thresholds in Table 1 is present below:

$$\mu_l = 882 - 200 \log(50 + 3x)$$  \hspace{1cm} (6)

3.2.4. Influence of bandwidth. The available bandwidth determines the throughput which then decides the encoding scheme. The compression mechanism of the encoding scheme will then affect the quality of the encoded audio. Obtained from subjective tests in [19], the utility mapping of a VoIP flow are shown in Table 2 and Fig.1(d).

<table>
<thead>
<tr>
<th>Codec/Encoding rate</th>
<th>Utility Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.711/64Kbps</td>
<td>94</td>
</tr>
<tr>
<td>G.726/40Kbps</td>
<td>92</td>
</tr>
<tr>
<td>G.726/32Kbps</td>
<td>91</td>
</tr>
<tr>
<td>G.726/24Kbps</td>
<td>89.5</td>
</tr>
<tr>
<td>G.729/8Kbps</td>
<td>84</td>
</tr>
<tr>
<td>G.723.1/6.3Kbps</td>
<td>79</td>
</tr>
<tr>
<td>GSM/5.6Kbps</td>
<td>71</td>
</tr>
</tbody>
</table>

3.3. Impairment Utility Function for video streaming

Due to the common characteristics of multimedia streams, the generic network impairment utility functions for VoIP can be used to model video streaming applications.

3.3.1. Influence of delay. Equation (7) defines a delay utility function of an interactive video stream with delay thresholds at 200ms and 1s (Table 1).

$$\mu_d = \begin{cases} 
100 - 0.004x, & x < 200 \text{ms} \\
-25 \tanh(0.04(x - 205)) + 65, & 65 \leq x \leq 300 \text{ms} \\
45.5 - 0.016x, & x > 300 \text{ms} 
\end{cases}$$  \hspace{1cm} (2)

3.3.2. Influence of jitter. Typical set-top boxes can buffer 100-500ms of SDTV video, so network jitter must be within these limits [12]. From the generic jitter utility function (3) we can derive the jitter utility function of a video stream.

$$\mu_j = -20 \tanh(0.005x - 3) + 8$$  \hspace{1cm} (8)
3.3.3. Influence of loss. The loss thresholds for a common video stream are chosen from Table 1 for the entire video stream instead of individual frames. A loss utility function of video flow is specified from the generic loss utility function (5) as below:

\[ \mu_l = 250 - 50\ln(20 + 12.5\chi) \]  

(9)

3.3.4. Influence of bandwidth. Using application of video compression techniques can result in coding distortion such as blocking effect, blurring, colour bleeding, staircase effect, etc [22]. In [23] researchers compare different codecs through user tests. In [24] video quality considering the encoding rate with 3 different video samples and two codecs are compared. Figure 2(d) shows an instance of the bandwidth utility distribution considering the encoding rate.

![Table 3. Utility value mapping of video streaming](image)

<table>
<thead>
<tr>
<th>Codec</th>
<th>H.264</th>
<th>H.263</th>
<th>H.263/VBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoding Rate</td>
<td>960Mbit/s</td>
<td>256Kbit/s</td>
<td>QP: 5</td>
</tr>
<tr>
<td>Profile</td>
<td>High</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Resolution</td>
<td>4096x2048</td>
<td>176x144</td>
<td>176x144</td>
</tr>
<tr>
<td>Frame Rate</td>
<td>30</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Utility</td>
<td>95</td>
<td>75</td>
<td>80</td>
</tr>
</tbody>
</table>

The quality of a video stream is also decided by other factors such as resolution, frame rate, video profile etc. These factors are all affected by the available bandwidth. To determine the encoding impairment on a video flow is a multi-dimensional QoS task. Instead of deriving utility functions on all of these factors individually, we propose the mapping scheme as in section III.B.4. Table 3 gives the mapping from five encoding impairment dimensions to utility values.

3.4. Utility Function for Online Gaming

Real Time Strategy (RTS) and First Person Shooter (FPS) are two typical online games. Since FPS games have stronger requirements on network response time, they will be taken as example for deriving utility functions.

3.4.1. Influence of delay. Lag (network delay) has large impact on user’s performance in online gaming especially FPS type of games. Because of the strong sensitivity of gamers to network delay, the hyperbolic tangent (with no extended transitions) is suitable for modeling the delay impact. From Table 1 we can generate the delay utility function of a FPS online gaming (10).

\[ \mu_d = -45\tanh(0.015\chi - 1.35)) + 60 \]  

(10)

3.4.2. Influence of jitter. As indicated in [25], absolute value of network delay dominates the delay impact on online gaming application. Jitter has little impacts on gamers’ performance. As discussed before, it is doubtful that there is jitter in the “traditional” sense. Since traffic patterns are not necessarily periodic and delay variations are small enough to be filtered by the gaming platform so users do not experience jitter directly. The utility value of jitter is then set to be 100 (11).

\[ \mu_j = 100 \]  

(11)

3.4.3. Influence of loss. Loss of signalling is critical for gaming. From the subjective tests results (Table 1), we can generate the following loss utility for online gaming:

\[ \mu_l = 100 - 50\ln(1 + \chi) \]  

(12)

3.4.4. Influence of bandwidth.

Table 4. Utility value mapping for an online gaming

<table>
<thead>
<tr>
<th>Frame rate (fps)</th>
<th>Client to Server (Kbps)</th>
<th>Server to Client (Kbps)</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1.92</td>
<td>2.88</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>4.48</td>
<td>6.72</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>9.6</td>
<td>14.4</td>
<td>70</td>
</tr>
<tr>
<td>30</td>
<td>19.2</td>
<td>28.9</td>
<td>92</td>
</tr>
<tr>
<td>60</td>
<td>38.4</td>
<td>57.6</td>
<td>95</td>
</tr>
</tbody>
</table>

FPS games choose the appropriate frame rate regarding the available bandwidth. From the analysis of the experiment results in [26], the performance of players increase with an increasing frame rate. Though at 30fps, a saturation point seems to be reached. Table 4 gives the utility value mapping for a game with average client packet size of 80 bytes and average server packet size of 120 derived from the evaluation of “Counter Strike” from Faerber [27].

3.5. Influence between QoS dimensions

There are inter-dependencies between QoS dimensions. For example, for interactive applications, packets with jitter over certain threshold have the same impact on the application as packet loss. Hence, in this case the jitter utility should be modeled with the corre-
sponding loss utility function. Dependency Utility Functions can be introduced to model the dependency between QoS dimensions. However, in this initial study, network impairments are assumed to be independent in this paper. Dependencies between QoS dimensions are for future study.

3.6. Application Utility Function

Network impairment utility functions are introduced in the last few sections to reflect the correlation between individual network impairment and the end user’s perception. In reality, end users sense all impairments simultaneously. A combination methodology needs to be designed to aggregate the network impairment utility functions to the application quality utility. Product, Weighted Summation and Minimum Value are three types of combination methods.

The Minimum Value combination function simply takes the minimum value of all sub-utilities. The combined utility is dominated by the lowest utility value. This method is easy to perform but unable to prioritize between applications with the same minimum sub-utility.

In [15], Kangasharju et al. propose the Product type of combination utility function, which takes the product values of the sub-utilities (which are scaled to [0, 1]). This method is based on the assumption that sub-utilities have equal importance to the user’s perception. However, different applications may have specific requirements on certain sub-utilities.

The Weighted Summation method (13) considers all the sub-utility with weight coefficients. Weight coefficients are chosen based on application attributes and user preferences. Because of the nature of the weighted summation method, each QoS dimension contributes positively or negatively according to the weight coefficient that reflects its relevance for a specific application type.

\[ U_c = \delta_a \cdot \mu_a + \delta_j \cdot \mu_j + \delta_l \cdot \mu_l + \delta_{bw} \cdot \mu_{bw} \quad (13) \]

\( U_c \) represents the combination utility value for a specific target application. \( \mu \) is the impairment utility which ranges from 0 to 100. The degree of contributions from different impairments is balanced by the weight value \( \delta \) ranging from 0 to 1 with the restriction of \( (\delta_a + \delta_j + \delta_l + \delta_{bw}) = 1 \).

4. Evaluation

In order to complement and confirm the analytical investigations, experiments using NS2 simulation have been carried out. The goal is to demonstrate the utility based QoS measurement and management within different scenarios.

4.1. Simulation Testbed

The simulation is based on real networks configurations. Three homes and one ISP are connected through access networks and a backbone network (Figure 4). The analysis of IP network can be split into four segments: home network, home internet connection, access aggregation and backbone.

Several traffic generating application nodes are added to the home network (e.g. node 2, 3, 4, 5 of Home 1 in Figure 4). Besides the VoIP, video streaming and online gaming applications, FTP is also introduce in the testbed since this is another major traffic source that needs to be considered in a realistic set-up. All application nodes connect to the Home gateway (node 1 in each home). The network resource for intra-home communication is set to simulate transmission with no loss and no delay.

The home internet connection (e.g. links between node 1 and node a0) aggregates all the inbound and outbound traffic. In reality, network resource can be overloaded on this link. A link with 1.5 Mb and 2.5Mb of bandwidth are chosen for Home 1 and Home 2. The access aggregation segment (e.g. the link between a0 and b0) is one of the main network segments where network impairment happens since large quantities of multimedia flows will share the bandwidth. Considering the quantity and type of applications, we chose 6Mb/s of bandwidth with 5ms of delay for the link between a0 and b0 in experiments. The link from node 1 of ISP 1 to the backbone and the connection between node a1 to the backbone are set up with enough bandwidth and very low delay. Backbone (the link between b0 and b1) is responsible to transfer data with no packet loss. In the simulation, we setup the delay of backbone as 30ms to simulate domestic traffic.

4.2. Traffic Generator

4.2.1. VoIP. The Pareto On/Off object (packetSize: 160, burst_time: 500ms, idle_time: 50ms, rate: 68kb, shape: 1.5) simulates the traffic of a typical VoIP ap-
application. In the simulation, VoIP applications of Home 1 and Home 2 connect with VoIP application in Home 4 to simulate a two way voice conversation between Home 1/2 and Home 4.

4.2.2. Video Streaming. The trace file of the H.263 VBR encoded movie “Jurassic Park I” [28] is used to simulate video streaming traffic from the video sender (node 3) of ISP 1. The simulated video streaming is sent to video streaming receivers (IPTV) in Home 1 and Home 2.

4.2.3. Online Gaming. According to the research in [14], traffic generated by online gaming is nearly constant bit rate. Thus, constant bit rate stream with sending rate of 15kbit/s and packet size of 80 Bytes is chosen to simulate game traffic.

4.2.4. FTP. TCP flows are sent from an FTP Sender agent (windowsize: 20, windowInit: 2, packetSize: 1000, packetSize: 40) on node 5 of ISP 1 to a FTP Sink agent on node 5 of Home 1.

4.3. Simulation Scenario

Figure 5 shows the simulation scenario. Applications are enabled and disabled at certain time to study the network traffic.

4.4. Simulation Results

In Figure 6, utility of Home 1 VoIP is plotted. The quality utility value is generated by combination utility functions with inputs of four impairment utility values (delay utility, jitter utility, loss utility and bandwidth utility) with different weight coefficients. The weight coefficients are chosen to model different application characteristics.

According to the application utilities, there are quality degradations between 150s and 200s, between 250s and 320s and around 600s. From the analysis of the sub network impairment utilities, the root cause of application quality degradation can be found. The delay utility fluctuates between 92 and 94 (Figure 7) which means that the delay has little impact on the application quality. The Jitter utility drops to 80 around 180s, 280s and 600s (Figure 8) which degrades the application quality slightly.

From Figure 9, we can draw the conclusion that packet loss is the main source of application quality deterioration since the loss utility value drops under 70 around 150s, 250s and 600s. Furthermore, with the end-to-end monitoring mechanism, it can be concluded that the packet loss happened on Home Internet Connection segment and Access Aggregation segments. The packet loss from Access Aggregation segments is relatively small. Thus, in order to increase the quality of this application we need to mainly reduce the packet loss in the Home Internet Connection segment.

Considering the utility plots of the applications on the same network segment (Figure 10 and Figure 11),
it can be concluded that the quality of Home 1 VoIP is affected by other applications in Home 1. The packet loss of VoIP around 150s and 250s are results of the bit rate impulse of the video streaming application and the FTP traffic. The packet loss of VoIP around 600s is also a result of the bit rate impulse of video streaming.

Figure 10. Throughput Values of Home 1 IPTV

Figure 11. Throughput Values of Home 1 FTP

In the previous paragraphs it has been demonstrated how utility-based network QoS assessment models evaluates the user experience on multimedia applications. Root cause of application quality degradation can also be diagnosed by cross associating utility values monitored from different network sections on the delivery path.

Apart from the quality measurement, utility functions can also be used for network management. Figure 12(a) and (b) show the delay and loss utility (discrete values of accumulated time slot) distributions of the Home 1 Video Steaming application. It can be concluded from the figures that packets of this application are received with a delay below 200ms and some impulse packet losses with 1% and 2% of packet loss rate. Due to the loss utility curve, a packet loss rate over 1% is annoying to the end user. Routers on the delivery path may use traffic management or routing algorithms to reduce the packet loss rate of this application to below 1% to increase the application quality. What can also be concluded from the utility curves is that this application can tolerate up to 300ms of packet loss. Further, packet loss rate below 0.5% is unperceivable to end user. To optimize the utilization of network resource and potentially improve other applications on the same network, network management can be used for inter flow optimization by adapting certain QoS parameters. Figure 12(c) and (d) show the results of a possible new route for the Home 1 Video Streaming application. With extra network delay and less packet loss, the quality of Home 1 video streaming application can be improved. If computing resources are sufficient, routers or QoS servers can choose the optimal routes for certain applications in real time using network impairment utility functions and application utility functions.

Figure 12. Distribution of delay and loss utility on Home 1 Video Streaming

5. Conclusion

Most of the current QoS models are using media flow resource reservations or adaptation based models acting on pre-defined thresholds. With an increase of multimedia applications using the Internet, resource reservation for all the multimedia flows appears to be unpractical. Moreover, purely threshold based application classification may not be suitable for emerging multimedia applications which have various delivery requirements. Furthermore, properties of applications between thresholds as well as inter-flow optimisation are also essential for both QoS measurement and management.

In this paper the utility based QoS model is introduced to assess multimedia delivery considering end user experience. The empirical impairment utility functions are derived and summarized from subjective experiments found in literature. Moreover, combination methods are studied to aggregate sub utilities. The utility model can be used as a user layer extension of existing network QoS models. Utility functions are used in the assessment of the perceived quality and the subsequent management of multimedia flows considering system wide aspects. For instant, an application utility function can be applied to the existing QoS mechanisms such as queue management and routing techniques for optimised network or application control decisions. Based on the results of this work, future
work will also deal with utility based network management.

6. Acknowledgement

The work presented in this paper is supported by the European Commission (Network of Excellence - CONTENT) and Agilent Laboratories UK.

References