ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.7, July 2009, pp.1921−1930
doi: 10.3724/SP.J.1001.2009.03297

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

保障连续媒体流用户层 QoS 的缓存控制*

邱 菡[1+], 李玉峰[1,2], 邬江兴[1]

[1](国家数字交换系统工程技术研究中心,河南 郑州 450002)

[2](防空兵指挥学院 信息控制系,河南 郑州 450052)

Buffer Control for Guaranteeing User-Level QoS of Continuous Media Flows

QIU Han[1+], LI Yu-Feng[1,2], WU Jiang-Xing[1]

[1](National Digital Switching System Engineering and Technological R&D Center, Zhengzhou 450002, China)

[2](Department of Information and Control, Air Defense Command College, Zhengzhou 450052, China)

+ Corresponding author: E-mail: qiuhan_ndsc@yahoo.com.cn

Qiu H, Li YF, Wu JX. Buffer control for guaranteeing user-level QoS of continuous media flows. *Journal of Software*, 2009,20(7):1921−1930. http://www.jos.org.cn/1000-9825/3297.htm

Abstract: A study on the effect of buffer control on user-level QoS of media flows is presented. In multimedia systems, playout buffer at the destination site is often adopted to compensate the delay jitter and improve the continuity of information playback. Buffer contol can reduce the effect of the delay jitter but increase the end-to-end delay. As delay and delay jitter are both the user-perceived QoS parameters, how does buffer control affect user-level QoS? By utilizing the former results on QoS mapping from application-level to user-level, and by investigating the relation between the buffer control parameter, end-to-end-level QoS parameters and application-level QoS parameters, the relationship between the buffer control parameter and user-level QoS parameter is obtained. The effect of buffer control on user-level QoS having been studied in-depth with theoretical analysis. The buffer size providing determinate delay and delay jitter guarantee is found and the buffer size providing the optimal user-level QoS for a certain network condition is demonstrated. Experimental results validate the analysis.

Key words: buffer control; delay; delay jitter; QoS mapping

摘 要: 研究了缓存控制对媒体流用户层 QoS 的影响.多媒体系统信宿端通常采用播放缓存来补偿时延抖动,提高媒体流播放的连续性.缓存控制虽然能够降低时延抖动的影响,却增加了端到端时延.时延或时延抖动是用户可感知的 QoS 参数,缓存控制对用户层 QoS 的影响究竟如何呢?利用已有的应用层向用户层 QoS 映射的研究结果,分析缓存控制参数与端到端 QoS 参数、应用层 QoS 参数的关系,获得了缓存控制参数与用户层 QoS 参数的关系.从理论上深入挖掘缓存控制对用户层 QoS 参数的作用,给出了提供确定时延和时延抖动保障的缓存容量值,论证了在网络环境一定时存在提供最佳用户层 QoS 的缓存容量值.实验结果验证了分析.

关键词: 缓存控制;时延;时延抖动;QoS 映射

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2005AA121210 (国家高技术研究发展计划(863)); the National Basic Research Program of China under Grant No.2007CB307102 (国家重点基础研究发展计划(973))

Received 2007-10-07; Accepted 2008-02-21

中图法分类号: TP393 文献标识码: A

近年来,随着高速接入网络和高性能终端的迅速发展,Internet 上的连续媒体应用越来越普及.连续媒体流区别于文本数据的最主要特征是具有时间结构特性.由于分组在网络上传输时所经历的时延不同,连续媒体流的时间结构可能被时延和时延抖动损害.多媒体系统通常在信宿端使用缓存对时延抖动进行补偿,称为抖动缓存或者补偿缓存.一方面,抖动缓存能够修复时延抖动对连续媒体流时间结构的损害;另一方面,抖动缓存的使用引入了额外的端到端时延,而时延的增加可能导致应用的主观质量下降.因此,作为可控的多媒体系统终端 QoS 控制的主要机制,适当的缓存控制对于提供较好的用户层 QoS 至关重要.

缓存控制对媒体流的时间结构修复是通过调整分组的缓存时间来实现的.缓存时间与抖动缓存的播放启动时间与缓存容量有关.播放启动时间是信宿端的客户播放程序开始播放的时间,若播放启动时间太短,则延迟到达的分组将因为过期而被丢弃;若播放启动时间太长,则将引入用户不能忍受的额外时延.对于缓存容量而言,如果过小,则将导致下述两类问题:其一,如果分组延迟到达接收端,则将被抖动缓存丢弃;其二,如果到达信宿端的突发分组超过了抖动缓存的容量,则将被丢弃.同样地,缓存容量过大将引入不可接受的额外时延,造成主观质量下降[1].由此,最佳的缓存控制是寻求时延抖动的补偿作用和时延增加作用上的平衡,于是出现了自适应式抖动缓存[2].自适应抖动缓存是一个智能过程,采用不同的策略修复将送往接收端的媒体流,虽然它能够根据网络情况动态地调节缓存,但这只是缓存控制的一种实现策略,并不能揭示缓存控制如何影响用户层 QoS.

时延和时延抖动是用户可感知的 QoS 参数,缓存控制对二者的影响反映在用户层 QoS 上存在着折衷,那么,缓存控制对用户层 QoS 的影响究竟如何呢?文献[3]中关于自适应缓存和确定容量缓存在不同网络时延标准差下的用户层 QoS 参数 MOS(mean opinion score)值的实验结果仅表明:在相同网络环境下,不同的缓存容量将造成不同的用户层 QoS.相关缓存控制对连续媒体流传输影响的研究也很多,但是它们只是针对时延抖动或者时延单独进行的[4−6].文献[7]综合分析了时延和时延抖动对用户层 QoS 的影响,文献[8]估计了缓存控制对用户层 QoS 的作用,并发现在实验环境下存在着最佳的初始缓存时间,这些研究都是基于实验进行的,受实验环境的局限,其结论缺乏可扩展性.目前,在相关方向上国内外并无成熟的理论分析.

本文从理论分析的角度,揭示了缓存控制对用户层 QoS 的影响,论证了在网络环境一定的情况下存在提供最佳用户层 QoS 的缓存容量值.虽然抖动缓存隶属于应用层,但是对端到端层 QoS 参数产生了直接影响.本文首先深入分析了缓存控制参数——缓存容量与端到端层 QoS 参数、应用层 QoS 参数之间的关系,基于以往的研究建立了应用层向用户层的 QoS 映射模型,实现了端到端层向用户层的 QoS 映射,最终建立了缓存容量与用户层 QoS 参数的关系.本文的理论分析结论能够与实验结果很好地吻合,验证了分析的正确性.

1 缓存控制与端到端 QoS 参数之间的关系

在多媒体系统中,每个媒体流由一系列有序的信息单元组成,称为媒体单元(media units,简称 MUs).一个媒体流的时间结构将受到其所经历的网络时延抖动的影响,因此,定义媒体流经历的网络时延 d 为包含媒体流任意 MU 的分组所经历的网络时延,媒体流经历的网络时延抖动 $J=d-E(d)$,其中 $E(d)$为平均网络时延.从统计角度来看,网络时延和时延抖动都是随机序列.在一定的网络环境下,$E(d)$是确定的,网络时延抖动的期望为 0,网络时延抖动的方差等于网络时延的方差,即 $D(J)=D(d)$,网络时延抖动分布由网络时延分布唯一确定.

播放启动时间和缓存容量决定了缓存控制对媒体流时间结构的作用,是缓存控制的重要参数.播放启动时间越大,可以吸收的正的时延抖动越大,而缓存容量越大,则可以补偿的负的时延抖动越大.当缓存容量与播放启动时间的差值较大时,抖动缓存可以吸收的正的时延抖动值和补偿的负的时延抖动值将严重不对称,造成丢弃事件或者空播事件的增加.因此,播放时间和缓存容量的合理选择对于控制缓存的上溢事件和下溢事件的发生都是至关重要的.一般地,在抖动缓存占用水平达到一半时开始播放媒体单元是很好的选择,此时,缓存容量或者播放启动时间可以作为缓存控制的唯一参数.假设抖动缓存的容量为 b(这里,缓存容量由时间来表示,而并非存储空间的数量),播放启动时间(即初始缓存时间)为 $b/2$,定义经过抖动缓存的媒体流的残留时延 d_r 和残留时

延抖动 J_r.那么,经过抖动缓存后,以信宿端客户的播放时刻为时间起点,媒体流的残余时延和残余时延抖动与网络时延和时延抖动的关系如下所示:

$$d_r=\begin{cases}d-b/2, & J>b/2\\ d-J, & |J|\le b/2\\ d+b/2, & J<-b/2\end{cases} \tag{1}$$

$$J_r=\begin{cases}J-b/2, & J>b/2\\ 0, & |J|\le b/2\\ J+b/2, & J<-b/2\end{cases} \tag{2}$$

由于缓存控制对媒体流时间结构的影响直接作用于端到端 QoS 参数,故选取端到端平均时延和时延方差作为端到端层 QoS 参数.定义端到端时延为 d_e,端到端时延抖动为 J_e,端到端平均时延为 $E(d_e)$,端到端时延方差为 $D(d_e)$.由此,以信宿端客户的请求播放时间为时间起点,端到端时延和时延抖动与缓存容量存在如下关系:

$$d_e=d_r+b/2=\begin{cases}E(d)+J, & J>b/2\\ E(d)+b/2, & |J|\le b/2\\ E(d)+J+b, & J<-b/2\end{cases} \tag{3}$$

$$J_e=J_r=\begin{cases}J-b/2, & J>b/2\\ 0, & |J|\le b/2\\ J+b/2 & J<-b/2\end{cases} \tag{4}$$

为了获得缓存容量与端到端层 QoS 参数的关系,假设网络时延抖动的概率密度函数为 $f(J)$,由此可计算出端到端时延和时延抖动的一阶和二阶统计特性:

$$E(d_e)=\int_{-\infty}^{\infty}E(d)f(J)\mathrm{d}J+\left[\int_{-\frac{b}{2}}^{\frac{b}{2}}\frac{b}{2}f(J)\mathrm{d}J+\int_{-\infty}^{-\frac{b}{2}}bf(J)\mathrm{d}J\right]+\left[\int_{\frac{b}{2}}^{\infty}Jf(J)\mathrm{d}J+\int_{-\infty}^{-\frac{b}{2}}Jf(J)\mathrm{d}J\right] \tag{5}$$

$$E(J_e)=\left(\int_{\frac{b}{2}}^{\infty}+\int_{-\infty}^{-\frac{b}{2}}\right)Jf(J)\mathrm{d}J+\left(-\int_{\frac{b}{2}}^{\infty}+\int_{-\infty}^{-\frac{b}{2}}\right)\frac{b}{2}f(J)\mathrm{d}J \tag{6}$$

$$D(d_e)=\int_{\frac{b}{2}}^{\infty}\left[E(d)+J\right]^2f(J)\mathrm{d}J+\int_{-\frac{b}{2}}^{\frac{b}{2}}\left[E(d)+b/2\right]^2f(J)\mathrm{d}J+\int_{-\infty}^{-\frac{b}{2}}\left[E(d)+J+b\right]^2f(J)\mathrm{d}J-E^2(d_e) \tag{7}$$

$$D(J_e)=\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)^2f(J)\mathrm{d}J+\int_{-\infty}^{-\frac{b}{2}}\left(J+\frac{b}{2}\right)^2f(J)\mathrm{d}J-E^2(J_e) \tag{8}$$

将网络看作一个完全随机的系统,网络时延抖动的概率密度函数 $f(J)$则是关于均值 0 的偶对称函数.许多相关研究证实了该假设.文献[9]分析得出时延抖动是均值为 0 的近似高斯分布.文献[10]提出,时延抖动在较大的时间尺度下的平稳过程可以由高斯白噪声过程完全地模拟.而文献[11]表明,正态分布或者对数正态分布是合适的 Internet 上整个分组时延的分布模型.由此,可求出端到端时延和时延抖动的均值:

$$E(d_e)=E(d)+\left(\int_{-\frac{b}{2}}^{\frac{b}{2}}+\int_{-\infty}^{-\frac{b}{2}}+\int_{\frac{b}{2}}^{\infty}\right)\frac{b}{2}f(J)\mathrm{d}J+\left[\int_{\frac{b}{2}}^{\infty}Jf(J)\mathrm{d}J+\int_{-\infty}^{-\frac{b}{2}}(-J)f(-J)\mathrm{d}(-J)\right]=E(d)+b/2 \tag{9}$$

$$E(J_e)=0 \tag{10}$$

结合式(3)和式(4),端到端时延和时延抖动的方差可表示如下:

$$D(d_e)=D(j_e)=2\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)^2f(J)\mathrm{d}J \tag{11}$$

式(9)表明,端到端的平均时延随着缓存容量的增加呈线性增长.该结果与文献[8]中的实验结果图 6 相吻合,端到端的平均时延对播放启动时间的斜率均为 1.由式(11)可以看出,缓存容量的增加将导致端到端时延方差的减小,这就意味着缓存吸收时延抖动能力的提高.

2 缓存控制与用户层 QoS 参数的关系

缓存控制直接作用于端到端层 QoS 参数,而端到端层并非用户层的毗邻,因此,研究缓存控制与用户层 QoS 参数之间的关系需要从研究缓存控制参数与端到端层 QoS 参数的关系入手,通过建立缓存控制与应用层 QoS 参数的关系,并利用已有的应用层向用户层 QoS 映射的研究结果来实现.

2.1 应用层向用户层的QoS映射

近年来,应用层向用户层 QoS 映射的研究日趋成熟,形成了以心理测量方法进行用户层 QoS 估计,以多元回归分析进行 QoS 映射等主要技术[7,8,12].选择适当的应用层 QoS 参数作为预测变量是进行多元回归分析的重要环节.研究表明,9 个应用层 QoS 参数可以表示媒体的同步质量[13],而同步质量与媒体流的时延抖动密切相关.通过实验研究端到端时延和时延抖动对用户层 QoS 影响的文献[7],研究缓存控制对用户层 QoS 作用的文献[8]都选取了 9 个应用层 QoS 参数中的 7 个与平均媒体时延组合,通过比较各组合的自由度,分别选择了(D_a,C_v)和(D_v,C_a)作为多元回归的预测变量,其中 D_a 和 D_v 分别表示音频和视频的平均媒体单元时延,而 C_a 和 C_v 分别表示音频和视频的输出间隔变化系数.在上述两种实验中,某种媒体的输出间隔变化系数和各媒体的平均媒体单元时延均有最高的组合值,而选择不同媒体的平均时延作为预测变量,如(D_a,C_v)和(D_v,C_a),是为了同时体现音视频流.因此,可以选择某种媒体流的平均媒体单元时延 D 为一个预测变量,而将该媒体流的输出间隔变化系数 C 作为另一个预测变量,如(D_a,C_a)和(D_v,C_v).

从哲学的观点来看,应用层 QoS 参数与用户层 QoS 参数具有本质的、确定的关系,不同的实验结果只是这一本质的不同反映.综合文献[7,8,12,13]中多元回归分析的结果,可以建立应用层向用户层映射的一般性模型.假设 S 是通过心理测量分析方法对用户层 QoS 参数的一个估计,可分为 5 类,“感觉不到”对应 4,“可觉察,但不讨厌”3,“些微地讨厌”2,“讨厌”1,“非常讨厌”0[7],那么应用层向用户层 QoS 映射的一般性模型可以表示如下:

$$S = S_0 - M_0 D - N_0 C,\ 0 < S_0 \le 4,\ M_0 \ge 0,\ N_0 \ge 0 \tag{12}$$

对于确定的一种媒体流,S_0,M_0 和 N_0 是确定的常数,而对于不同特征的媒体流,它们是不同的.每次实验结果都为 S_0,M_0 和 N_0 提供一个估计值,例如文献[7]中的式(4)给出了预测变量为(D_a,C_v)时的估计值,依次是 3.94, 2.750×10^{-3} 和 2.175.

2.2 缓存控制与应用层QoS参数的关系

对于应用层 QoS 参数平均媒体单元时延 D,在忽略解码时延和处理时延时,即为端到端时延:

$$D = E(d_e) \tag{13}$$

为了便于研究端到端 QoS 参数与 C 的关系,首先回顾 C 的定义,输出间隔变化系数,即媒体单元输出间隔的标准差与平均输出间隔的比值.由此,假设时间结构连续的两个 MU i 和$(i-1)$的输入间隔为 I_{0i},d_{ei} 和 $d_{e(i-1)}$分别表示两个连续的 MU i 和$(i-1)$所经历的端到端时延,那么,第 i 和第$(i-1)$个 MU 的输出间隔 I_i 为

$$I_i = I_{0i} + d_{ei} - d_{e(i-1)} \tag{14}$$

将网络看作一个随机系统,则 I_{0i} 和 I_i 是随机序列,d_{ei} 和 $d_{e(i-1)}$是独立同分布的.设 $E(I_0)$为该媒体流的 MU 平均输入间隔,$D(I_0)$为 MU 输入间隔方差,从而得出输出间隔均值 $E(I)$和输出间隔方差 $D(I)$为

$$E(I) = E(I_0) + E(d_{ei}) - E(d_{e(i-1)}) = E(I_0) \tag{15}$$

$$D(I) = D(I_0) + D(d_{ei}) + D(d_{e(i-1)}) = D(I_0) + 2D(d_e) \tag{16}$$

综合式(15)和式(16),输出间隔系数 C 可以表示为

$$C = \sqrt{D(I_0) + 2D(d_e)} / E(I_0) \tag{17}$$

将式(9)和式(11)分别代入式(13)和式(17),可以得到缓存控制与应用层 QoS 参数的关系:

$$D = E(d) + \frac{b}{2} \tag{18}$$

$$C=2\sqrt{D(I_0)+\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)^2 f(J)\mathrm{d}J}\Big/E(I_0),\ D(J)=D(d) \tag{19}$$

式(18)、式(19)表明,增大缓存容量(播放启动时间)将引起平均媒体单元时延增加,导致输出间隔变化系数减小.

依据网络层次功能特性,端到端层无法区分流特性,而应用层则可以识别流的不同特性,上述关于端到端 QoS 参数与应用层 QoS 参数关系的分析并不能包含代表所有特性的流.不难看出,上述分析是基于分组时延等于媒体单元时延的假设,而这一假设并不适用于大的媒体单元,原因在于,大的媒体单元在网络传输时将被分片组装在多个分组之中.测量结果表明,音频媒体单元远小于视频媒体单元,能够封装在一个分组中[14].因此,上述端到端 QoS 参数与应用层 QoS 参数的关系更适用于音频流.为了验证分析的正确性,图 1 和图 2 分别给出了同一环境下应用层 QoS 参数随播放启动时间变化的曲线,平均媒体单元帧速率为 20MU/s,平均网络时延为 50ms,时延的标准差为 20ms(该环境与文献[8]中的实验环境相同).

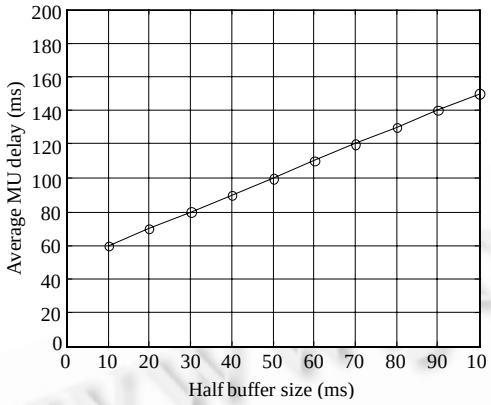

Fig.1 Average MU delay vs. playout startup time

图 1 平均 MU 时延和播放启动时间

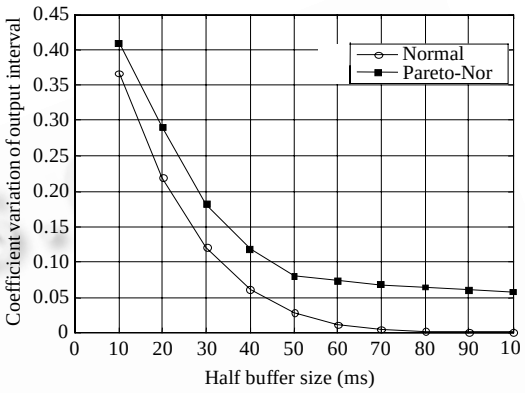

Fig.2 Coefficient of variation of output interval vs. playout startup time

图 2 输出间隔变化系数和播放启动时间

与文献[8]中图 6 相比,图 1 描述的平均媒体单元时延与播放启动时间关系与文献[8]中图 6 完全相同,唯一的区别在于平均媒体单元初始值不同,图 1 为 60ms,而文献[8]中图 6 为 120ms.这个区别源于本文在计算平均媒体单元时延时忽略了解码时延、传输时延和处理时延等,这些时延由多媒体系统(编码算法、传输距离等)所决定,为固定时延.因此,可引入附加时延 d_0 修正缓存控制与应用层 QoS 参数的关系:

$$D=d_0+E(d)+\frac{b}{2} \tag{20}$$

其中,在文献[8]的实验环境下,d_0 等于 60ms.从文献[8]中的图 6 还可以看出,音频流的平均媒体单元时延和视频流的平均媒体单元时延并没有显著区别.由此,式(20)能够正确地描述缓存控制与应用层 QoS 参数 D_a 或者 D_v 的关系,采用 D_a 或者 D_v 作为预测变量的映射结果是相同的.另外,式(20)还表明平均媒体单元时延与网络时延抖动或网络时延的分布特性无关.

输出间隔变化系数 C 与网络时延抖动的概率分布函数密切相关,依据本文给出的定义,网络时延抖动的概率分布仅由网络时延的概率分布唯一决定.研究表明,Pareto 分布是最合适的网络时延分布的拖尾部分模型,而正态分布则是合适的 Internet 上整个分组时延的分布模型[11].因此,图 2 分别采用正态分布和 Pareto-Normal(具有 Pareto 拖尾的正态分布)作为网络时延抖动的分布模型来计算 C.比较来看,图 2 中的 Pareto-Normal 曲线能够与文献[8] 中图 4 的音频系数曲线完全匹配.由于拖尾部分的分布不同,正态曲线和 Pareto-Normal 曲线在细节上存在着差别,但是播放启动时间对输出间隔变化系数 C 的作用趋势是相同的.图 2 还表明,忽略解码时延、传输时延和处理时延等对 C 并无影响,与文献[8]中图 4 吻合则说明式(19)能够准确地描述缓存控制和 C_a 关系.

2.3 缓存控制与用户层QoS参数的关系

经过上述两节的分析,缓存容量与用户层 QoS 参数的关系可以描述如下:

$$S = S_0 - M_0\left[d_0 + \frac{b}{2} + E(d)\right] - \frac{2N_0}{E(I_0)}\sqrt{D(I_0) + \int_{\frac{b}{2}}^{\infty}\left(J - \frac{b}{2}\right)^2 f(J)\mathrm{d}J} \tag{21}$$

其中,系数 S_0,M_0,N_0,d_0,$E(I_0)$和 $D(I_0)$的值可以通过在特定网络环境下针对特定媒体流进行实验测量评估得到.但是,缓存控制与用户层 QoS 参数的关系式(21)与网络时延抖动的概率密度函数仍然密切相关.由于网络的复杂性和不确定性,目前关于网络时延分布并没有统一的结论.研究认为,正态分布可以作为整个网络时延分布的合理模型[9−11],因此,假设网络时延服从正态分布是分析缓存容量与用户层 QoS 参数关系的一个可行方法.在此假设下,缓存容量与用户层 QoS 参数的关系可简化如下:

$$S = S_0 - M_0\left[d_0 + \frac{b}{2} + E(d)\right] - \frac{2N_0}{E(I_0)}\sqrt{D(I_0) + \left[D(d) + \frac{b^2}{4}\right]\cdot\left[1 - \Phi\left(\frac{b}{2\sqrt{D(d)}}\right)\right] - \frac{b}{2}\sqrt{\frac{D(d)}{2\pi}}\mathrm{e}^{-\frac{b^2}{8D(d)}}} \tag{22}$$

从式(21)可以看出,当缓存容量确定时,网络平均时延的增加或者网络时延方差的增加都将引起用户层 QoS 参数 S 的减少;当网络情况一定时,即网络平均时延和网络时延方差确定时,缓存容量的增加将引起应用层 QoS 参数 D 的增加,C 的减少,对用户层 QoS 参数 S 的作用并不明朗,可能是增加或者减少,甚至可能不变.下面将深入研究缓存容量对用户层 QoS 参数的作用,从提供最佳用户层 QoS 的角度论证最佳缓存容量值的存在性.

3 缓存控制对用户层 QoS 的作用

用户层 QoS 测量实验一方面可以估计用户层 QoS 参数,另一方面可以测量一些人类可感知的 QoS 参数,如时延和时延抖动.测量结果表明,时延抖动不超过一定值时,流看起来是“同步”的[6],如音频流允许的最大时延为 250ms,最大时延抖动为 10ms.从可感知的 QoS 角度来看,用户直接对端到端的时延和时延抖动提出了要求.那么,分析一定网络环境下提供确定时延和时延抖动保障的缓存容量值是研究缓存控制提供用户层 QoS 保障的重要基础.

3.1 保障媒体流同步的缓存容量分析

目前,关于保障媒体流同步的研究多是针对 QoS 要求的最大最小值来进行,分析的是最为极端的情况.对于时延而言,极端情况很少出现,极端情况下的分析并不适合发挥网络“尽力而为”的特长.另外,目前的媒体解码技术能够容忍一定的丢弃而不影响主观质量,如 VoIP(voice over IP)能够忍受的最大丢弃率为 1%[15].因此,以牺牲少量媒体单元为代价换取系统要求降低的概率保障是进行媒体流同步保障研究的最佳方法.

假设允许的最大时延为 $d_{\max}$,允许的最大时延抖动为 $J_{\max}$,则端到端时延和时延抖动应满足下述要求:

$$P(d_e > d_{\max} - d_0) \le \varepsilon_d,\ 0 < \varepsilon_d < 1 \tag{23}$$

$$P(|J_e| > J_{\max}) \le \varepsilon_j,\ 0 < \varepsilon_j < 1 \tag{24}$$

结合式(4)和式(24),对时延抖动的要求变为

$$P(|J_e| > J_{\max}) = P(J - b/2 > J_{\max}, J > b/2) + P(J + b/2 < -J_{\max}, J < -b/2) = P(|J| > J_{\max} + b/2) \le \varepsilon_j \tag{25}$$

对一个媒体系统来说,其平均网络时延与播放启动时延显然应小于最大允许时延,即

$$E(d) + b/2 < d_{\max} \tag{26}$$

综合式(3)、式(23)和式(25),对时延的要求变为

$$P(E(d) + J > d_{\max}) = P(J > d_{\max} - E(d)) \le \varepsilon_d \tag{27}$$

Chebyshev 不等式:假设 X 为一个随机变量,如果 X 的期望为 μ,即 $E(X)=\mu$,且 X 的方差为 δ^2,那么对于任意的正数 ε,下式成立: $P(|X-\mu| \ge \varepsilon) \le \frac{\delta^2}{\varepsilon^2}$.由上述分析可知,时延抖动 J 的期望值为 0,其方差为 $D(d)$,则依据 Chebyshev 不等式,式(25)变换如下:

$$P(|J|>J_{\max}+b/2)<\frac{D(d)}{(J_{\max}+b/2)^2}\le\varepsilon_j \tag{28}$$

将网络看作随机系统,网络时延抖动关于均值 0 呈偶对称分布.运用 Chebyshev 不等式,式(26)变为

$$P(J>d_{\max}-E(d))=\frac{1}{2}P(|J|>d_{\max}-E(d))<\frac{D(d)}{2[d_{\max}-E(d)-d_0]^2}\le\varepsilon_d \tag{29}$$

结合式(28)、式(29),在网络时延均值和方差已知的情况下,保障端到端时延和时延抖动对缓存容量的要求如下:

$$\begin{cases}2\left(\sqrt{\dfrac{D(d)}{\varepsilon_j}}-J_{\max}\right)\le b\le 2\sqrt{\dfrac{D(d)}{2\varepsilon_d}}, & \sqrt{\dfrac{D(d)}{\varepsilon_j}}-J_{\max}\le\sqrt{\dfrac{D(d)}{2\varepsilon_d}}\\ \max\left(2\left[\sqrt{\dfrac{D(d)}{\varepsilon_j}}-J_{\max}\right], 2\sqrt{\dfrac{D(d)}{2\varepsilon_d}}\right)<b\le 2[d_{\max}-E(d)-d_0] & \end{cases},$$

即

$$2\left(\sqrt{\frac{D(d)}{\varepsilon_j}}-J_{\max}\right)\le b\le 2[d_{\max}-E(d)-d_0] \tag{30}$$

式(30)的左边表达式与过去的研究结果一致[9],即提供确定时延抖动保障对缓存容量的要求.对式(30)进行变换,可以得到一个重要的结论,即网络提供确定的端到端时延和时延抖动保障的充分条件是

$$\sqrt{\frac{D(d)}{\varepsilon_j}}+E(d)+d_0\le d_{\max}+J_{\max} \tag{31}$$

上述分析虽然给出了保障确定时延和时延抖动要求的缓存容量范围,但是没有表明在要求范围内哪个值更佳.文献[3,8]的实验结果均表明,在相同网络环境下,不同的缓存容量对应于不同的用户层 QoS,因此,从提供最佳用户层 QoS 角度来看,在缓存容量要求范围内应该存在着最佳缓存容量值,下面将对此进行深入分析.

3.2 缓存控制对用户层QoS参数的作用

从缓存容量与用户层 QoS 参数的关系式(21)无法直接观察出缓存容量对 S 的影响,本节按照变限积分函数的求导方法[16]求解用户层 QoS 参数 S 关于播放启动时间 $b/2$ 的一阶导数 S'和二阶导数 S'',借助数学工具深入研究缓存容量对 S 的作用,得到定理 1.

定理 1. 当给定网络时延分布且一阶距、二阶距存在时,存在一个合适的缓存容量 b_0 使得用户层 QoS 参数 S 取得最大值 S_m.

证明详见附录.

定理 1 表明,确定的网络环境对应唯一的最佳用户层 QoS 参数.当最佳的用户层 QoS 不能满足用户要求时,无论怎样调节缓存控制都无法满足用户 QoS 要求,此时必须对网络进行优化.

由定理 1 的证明可知,缓存容量 b 由 0 增加至 b_0,用户层 QoS 参数随之增加,缓存容量 b 由 b_0 继续增加,用户层 QoS 参数将随之减少.结合第 3.1 节的结论式(30),提供确定时延和时延抖动的最佳缓存容量值如下:

$$b_{opt}=\begin{cases}2\left(\sqrt{D(d)/\varepsilon_j}-J_{\max}\right), & b_0<2\left[\sqrt{D(d)/\varepsilon_j}-J_{\max}\right]\\ b_0, & 2\left[\sqrt{D(d)/\varepsilon_j}-J_{\max}\right]\le b_0<2[d_{\max}-E(d)-d_0]\\ 2[d_{\max}-E(d)-d_0], & b_0>2[d_{\max}-E(d)]\end{cases} \tag{32}$$

为了直接展示缓存容量对用户层 QoS 参数的影响,下面分别采用正态分布和 Pareto-Normal 分布作为网络时延抖动的整体分布来描画用户层 QoS 参数随缓存容量变化的曲线.为了同时呈现出提供确定时延和时延抖动保障的缓存容量范围,设媒体流允许的最大时延为 250ms,最大时延抖动为 10ms.采用文献[8]中的结果 3.859,3.496×10^{-3} 和 9.145 作为系数 S_0,M_0 和 N_0 的估计.为了便于与实验结果比较,假定网络环境与文献[8]中实验的网络环境相同,附加时延为 60ms,平均网络时延为 50ms,网络时延标准差为 20ms,而平均媒体单元输入间隔

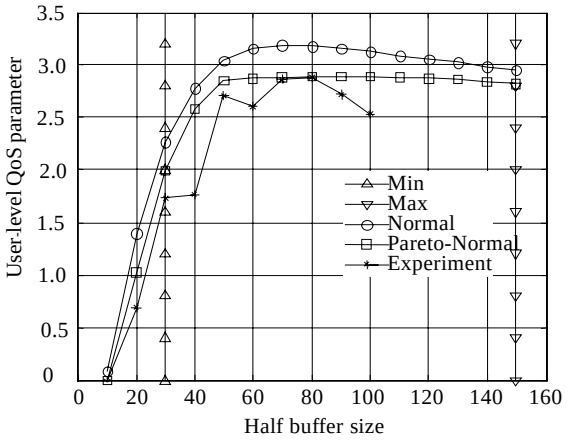

Fig.3 User-Level QoS parameter vs. playout startup time

图 3 用户层 QoS 参数和播放启动时间

为 20MU/s,输入间隔方差为 0.图 3 分别描述了网络时延服从正态分布时播放启动时间对用户层 QoS 参数的影响(Normal 线),网络时延服从 Pareto-Normal 分布时播放启动时间对用户层 QoS 参数的影响(Pareto-Normal 线),并依据式(30)绘出了缓存容量范围的最小值(min 线)和最大值(max 线).缓存容量的最小值与最大值曲线所夹的曲线能够保障端到端时延在 250ms 以内,时延抖动在 10ms 以内.从 Normal 线和 Pareto-Normal 线可以看出,用户层 QoS 参数随着缓存容量的增加先增加后减少.文献[8]中关于初始缓存时间对用户层 QoS 参数的影响(experiment 线)与 Pareto-Normal 线变化趋势相一致,且极大值点均为 80ms,验证了本文的理论分析结果.由于文献[8]中应用层向用户层的 QoS 映射并不完美,所以理论分析曲线(Pareto-Normal 线)并不能完全与实验结果曲线相匹配.由于正态分布的拖尾部分衰减快于 Pareto-Normal 分布,所以正态分布下最佳缓存容量半值为 72ms(Normal 线),略小于 Pareto-Normal 分布下的最佳缓存容量值.综上,从协助网络提供用户层 QoS 保障的角度来看,依据网络状态进行适当的缓存控制可以维持较高的用户层 QoS,即用户层 QoS 参数值在 3 以上.

4 总 结

众所周知,缓存控制虽然能够吸收时延抖动,但却增加了端到端时延.时延和时延抖动都是用户可感知的 QoS 参数,缓存控制对二者产生了截然相反的作用,缓存控制对用户层 QoS 的影响如何呢?为了回答这个问题,本文建立了缓存容量与用户层 QoS 参数的关系,分析了缓存控制对时延和时延抖动的相反作用反映在用户层 QoS 上的折衷,给出了在保障确定时延和时延抖动下提供最佳用户层 QoS 参数的缓存容量值,并论证了网络环境一定的情况下存在提供最佳用户层 QoS 的缓存容量值.文献[8]的实验结果验证了本文的理论分析结果.

本文的研究只是迈出了探索多媒体系统提供媒体流 QoS 保障征途的一小步.作为理论分析,还需要通过大量不同环境下的实验来验证结论.另外,研究合适的网络时延分布对于计算最佳缓存容量至关重要.最后,本文的结论对于实际的缓存控制设置提出了有益的理论指导,下一步将运用结论设计合适的缓存控制策略.

References:

[1] VoIP trouble shooter Website 2007. http://www.voiptroubleshooter.com/problems/jitterbuffer.html

[2] Ramjee R, Kurose J, Towsley D, Schulzrinne H. Adaptive playout mechanisms for packetized audio applications in wide-area networks. In: Proc. of the IEEE INFOCOM'94. New York: IEEE Press, 1994. 680−688. http://www.cs.columbia.edu/~hgs/papers/Ramj94_Adaptive.pdf

[3] Ribadeneira AF. An analysis of the MOS under conditions of delay, jitter and packet loss and an analysis of the impact of introduction piggybacking and reed Solomon for FEC FOR VOIP [Ph.D. Thesis]. Georgia State: College of Arts and Sciences Georgia State University, 2007.

[4] Int'l Telecommunication Union Telecommunication Standardization Sector (ITU-T) Recommendation G.114: Transmission Systems and Media, General Characteristics of International Telephone Connections and International Telephone Circuits, One. Way Transmission Time. 1996. http://www.itu.int/rec/T-REC-G.114-200305-I/en

[5] Kouvelas I, Hardman V, Watson A. Lip synchronization for use over the Internet: Analysis and implementation. In: Proc. of the IEEE GLOBECOM'96. New York: IEEE Press, 1996. 893−898. http://www-mice.cs.ucl.ac.uk/multimedia/publications/sync_globe.ps

[6] Steinmetz R. Human perception of jitter and media synchronization. IEEE Journal on Selected Areas in Communications, 1996,14(1):61−72.

[7] Ito Y, Tasaka S, Fukuta Y. Psychometric analysis of the effect of end-to-end delay on user-level QoS in live audio-video transmission. In: Proc. of the IEEE ICC 2004. New York: IEEE Press, 2004. 2214−2220. http://inl.elcom.nitech.ac.jp/PDFs/userlevel/icc2004y.pdf

[8] Ito Y, Tasaka S, Fukuta Y. Psychometric analysis of the effect of buffering control on user-level QoS in an interactive audio-visual application. In: Proc. of the 2004 ACM Workshop on Next-Generation Residential Broadband Challenges. New York: ACM Press, 2004. 2−10. http://inl.elcom.nitech.ac.jp/PDFs/userlevel/nrbc2004y.pdf

[9] Xu Y, Chang YL, Liu ZJ. Buffer design for P-QoS in multimedia synchronization systems. In: Proc. of the 2001 Int'l Conf. on Info-Tech and Info-Net. New York: IEEE Press, 2001. 486−491. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=983625

[10] Daniel E, White C, Teague K. An inter-arrival delay jitter model using multi-structure network delay characteristics for packet networks. In: Proc. of the 37th Asilomar Conf. on Signals, Systems, and Computers. New York: IEEE Press, 2003. 1738−1742. http://www.clsp.jhu.edu/~cwhite/papers/asilo_03_Jitter.pdf

[11] Fujimoto K, Ata S, Murata M. Statistical analysis of packet delays in the Internet and its application to playout control for streaming applications. IEICE Trans. on Communication, 2001,84(6):1504−1512.

[12] Ito Y, Tasaka S. Quantitative assessment of user-level QoS and its mapping. IEEE Trans. on Multimedia, 2005,7(3):572−584.

[13] Ito Y, Tasaka S. User level QoS assessment with psychometric methods and QoS mapping. In: Proc. of the IIIS Int'l Conf. on Computer, Communications and Control Technologies. 2003. 127−131. http://inl.elcom.nitech.ac.jp/PDFs/userlevel/ccct2003y.pdf

[14] Kuang T, Williamson C. A measurement study of realmedia audio/video streaming traffic. In: Proc. of the SPIE ITCOM. 2002. 68−79. http://pages.cpsc.ucalgary.ca/~carey/papers/2002/itcom02.pdf

[15] Amir Y, Danilov C, Goose S, Hedqvist D, Terzis A. 1-800-Overlays: Using overlay networks to improve VoIP quality. In: Proc. of the 15th Int'l Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV). New York: ACM Press, 2005. 51−56. http://www.dsn.jhu.edu/pub/papers/cnds-2004-2.pdf

[16] Lu Y, Li Y, Li Z, Sun S, Li Y. Study about the uncertain limit integral functions derivative methods. Journal of He'nan Institute of Education (Natural Science), 2004,13(1):4−6 (in Chinese with English abstract).

附中文参考文献:

[16] 卢亚丽,李艳华,李战国,孙书安,李晔.变限积分函数求导方法研究.河南教育学院学报(自然科学版),2004,13(1):4−6.

附录

定理 1. 当给定网络时延分布且一阶距、二阶距存在时,存在一个合适的缓存容量 b_0 使得用户层 QoS 参数 S 取得最大值 S_m.

证明:对式(21)进行求导,得出用户层 QoS 参数 S 关于 $b/2$ 的一阶导数 S' 为

$$S'=\frac{\mathrm{d}S}{\mathrm{d}\left(\frac{b}{2}\right)}=-M_0+\frac{2N_0}{E(I_0)}\frac{\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)f(J)\mathrm{d}J}{\sqrt{D(I_0)+\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)^2 f(J)\mathrm{d}J}} \tag{33}$$

由表达式(33)可知,S'是在缓存容量 b 的取值区间 $[0,\infty)$ 上的连续函数.按照变限积分函数的求导方法对表达式(33)进行求导,可得用户层 QoS 参数 S 的二阶导数 S'':

$$S''=\frac{\mathrm{d}S'}{\mathrm{d}\left(\frac{b}{2}\right)}=\frac{2N_0}{E(I_0)}\left\{\frac{\left[\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)f(J)\mathrm{d}J\right]^2}{\left[D(I_0)+\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)^2 f(J)\mathrm{d}J\right]^{\frac{3}{2}}}-\frac{\int_{\frac{b}{2}}^{\infty}f(J)\mathrm{d}J}{\left[D(I_0)+\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)^2 f(J)\mathrm{d}J\right]^{\frac{1}{2}}}\right\} \tag{34}$$

当给定网络时延分布且一阶矩、二阶矩存在时,依据积分域的 Cauchy 不等式,可得到 $\left[\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)f(J)\mathrm{d}J\right]^2\le$

$\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)^2 f(J)\mathrm{d}J \cdot \int_{\frac{b}{2}}^{\infty} f(J)\mathrm{d}J$,则由上述不等式可得

$$S''<0 \tag{35}$$

由此可知,S'在取值区间$[0,\infty)$上单调递减,b 为 0 时 S'取最大值,$b\to\infty$ 时 S'取最小值.当缓存容量 b 为 0 时,S'取最大值 $S'(0)$:

$$S'(0)=-M_0+\frac{2N_0}{E(I_0)}\sqrt{\frac{2}{2D(I_0)+D(d)}}\int_0^{\infty} Jf(J)\mathrm{d}J \tag{36}$$

当缓存容量 b 趋于无穷大时,利用 L'Hospital 准则可以求得其一阶导数 $S'(\infty)$:

$$\lim_{b\to\infty}S'=-M_0+\frac{2N_0}{E(I_0)}\lim_{b\to\infty}\left\{\frac{2\left[\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)f(J)\mathrm{d}J\right]\left[-\int_{\frac{b}{2}}^{\infty}f(J)\mathrm{d}J\right]}{-2\int_{\frac{b}{2}}^{\infty}\left(J-\frac{b}{2}\right)f(J)\mathrm{d}J}\right\}^{\frac{1}{2}}=-M_0<0 \tag{37}$$

不等式(37)给出的 $S'(0)$取值依赖于各参数的具体取值,下面分两种情况加以证明:

(1) $S'(0)\geq 0$

运用零点定理,$S'(0)\geq 0$, $S'(\infty)<0$,必然存在 $b_0\in[0,\infty)$ 使得 $S'(b_0)=0$,即有下列等式成立:

$$M_0E(I_0)\sqrt{D(I_0)+\int_{\frac{b_0}{2}}^{\infty}\left(J-\frac{b_0}{2}\right)^2 f(J)\mathrm{d}J}-2N_0\int_{\frac{b_0}{2}}^{\infty}\left(J-\frac{b_0}{2}\right)f(J)\mathrm{d}J=0 \tag{38}$$

依据判断极值点的第 2 充分条件:b 取 b_0 时,S'为 0 且 S''小于 0,那么 S 在 b_0 处取得极大值.由式(36)和式(39),b_0 为最大值点得证.

(2) $S'(0)<0$

由 $S'(0)<0$, $S''<0$ 可知,用户层 QoS 参数 S 在 b 的取值区间$[0,\infty)$上单调递减,在 0 时取得最大值,即 $b_0=0$. 此时,缓存容量 b 在区间$[0,\infty)$内增大将引起 S 的减少.

综上所述,当方程式(38)有解时,b_0 为方程式(38)的解,当方程式(38)无解时,b_0 为 0,总是存在 b_0 使得用户层 QoS 参数 S 取得最大值 S_m:

$$S_m=S_0-M_0\left[d_0+\frac{b_0}{2}+E(d)\right]-\frac{4}{M_0}\left[\frac{N_0}{E(I_0)}\right]^2\int_{\frac{b_0}{2}}^{\infty}\left(J-\frac{b_0}{2}\right)f(J)\mathrm{d}J \tag{39}$$

由此,定理 1 得证. □

邱菡(1981－),女,湖北随州人,博士,主要研究领域为宽带信息网络,流媒体技术.

李玉峰(1976－),男,博士,讲师,主要研究领域为宽带信息网络,高速路由器核心技术.

邬江兴(1953－),男,教授,博士生导师,中国工程院院士,主要研究领域为信息网络与交换.