Charging, Accounting and Billing of Multimedia Streaming in 3G Mobile Networks

Tamás Jursonovics, Sándor Imre

Abstract—**In this paper, we describe a CAB model for 3G mobile networks, and we present a new streaming proxy based charging architecture for QoS-differentiated charging in 3G. Our goal is to point out the importance of bursty packet loss, which determines the quality of the multimedia streaming, in contrast with the average loss rate. We propose several methods for QoS-differentiated charging, which are in compliance with the recommendations of 3GPP. We describe three models, which can predict the QoS of a stream. The algorithms are evaluated on a real GPRS streaming flow.**

Index Terms—**3G, accounting, billing, CAB model, charging, mobile network, real time services, streaming, QoS**

I. Introduction

The evolution of mobile telecommunication technology has accelerated in the last 5 years. Users can access the Internet faster with the newly developed, IP based network technologies (i.e. EDGE & UMTS). The existing GPRS system allows only a limited range of services, like e-mail or WAP based Internet surfing, but the 3G offers the users to connect to a large number of Value Added Service Providers (VASPs). The VASPs can deploy their services (like real-time multimedia streaming, value added, location based services or audio/video conferencing) through a 3G network provider, therefore the 3G systems require new methods for charging and charging.

Many billing solutions are developed for 3G. L. Baugé looks into the key challenges of GPRS billing [1]. He seeks to show how solving these challenges will help operators succeed in the not so distance future of UMTS. Maria Koutsopoulou introduces a novel billing scheme for UMTS networks [2] which is capable of handling efficiently charging, accounting and billing for value added services that are provided either by the network operator or third trusted parties. Hitesh Tewari presents new AAA methods for Real-Time Payments of Mobile IP [3].

But the most solutions are not able to consider the special charging requirements of streaming. The visual quality of a streaming media depends on various network conditions (for example delay, delay jitter, loss distribution, etc…), therefore the 3G needs new methods for the QoS differentiated charging. Moreover if the customers connect to a 3rd party streaming providers through a network operator then the service which is used by the customer is on the side of the 3rd party, therefore only the 3rd party can charge this service usage, which must be shared with the NO, which must trust this information. Summing up, the 3G requires new methods for charging and billing.

This paper is organized as follows: Section II introduces the business model. Section III describes the streaming from mobile operator point of view. Section IV.A reviews the charging requirements in 3G networks. Sections IV.B and IV.C present a new architecture. Section IV.D explains mechanisms for QoS differentiated charging, Section IV.E and IV.F describe and evaluate some models, which can predict the QoS of a streaming flow. Finally Section V concludes the paper.

II. Business model

The customers are billed separately in the GSM business model: they have subscriptions to one mobile operator and to a large number of Value Added Service Providers (3rd parties), so they receive multiple bills: one for the network usage and several other ones from the VASPs. That makes the charging very uncomfortable and difficult, which is unacceptable for the 3G, because the customers probably have a great number of value added service subscriptions (streaming services, etc.). So, the 3G will require more than just the addition of a module to the existing GSM billing system. It must

--handle the enhanced functionality of the new system,

--preserve and enhance exiting defined processes.

The UMTS-Forum describes in its report [4] three new models: Network Operator/Content Aggregator/Content Provider (~VASP) Centric Business Model. In the immediate future, the implementation of the 3G will based on the 2G system, thus in the early age of 3G, the main concept of billing scheme will correspond with the present GSM concept, which is a very mobile operator centered. So, to describe the 3G billing from the view of the business in my paper, let's see the first (NO) model (Fig. 1).

There is a network operator (NO) in the center that provides mobile access to a telecommunication network. It makes its capability and added value increase by offering a mobile portal to the customers, so therefore manages its own content aggregator (CA) or possibly value added service provider (VASP) role.

This work was supported by ETIK and OTKA F042590.

T. Jursonovics is with the Budapest University of Technology and Economics, Department of Telecommunications, Budapest, Hungary (corresponding author to provide phone: +36-1-463-3227; fax: +36-1-463-3263; e-mail: jursonovics@mcl.hu).

S. Imre, Dr., is with the Budapest University of Technology and Economics, Department of Telecommunications, Budapest, Hungary (e-mail: imre@hit.bme.hu).

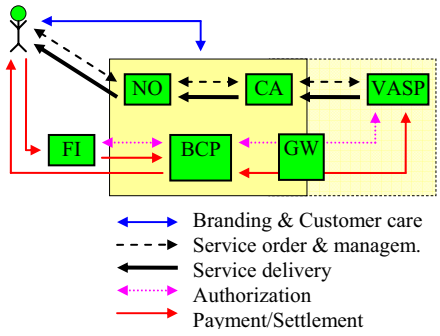

Fig. 1. Network operator centric business model

External parties involved may be VASP and financial institutions (FI). The NO has its own billing, collections and customer care infrastructure (BCP). The customers have a direct relationship with the NO, which sets the prices of the services, and handles the payments. The VASPs can receive the charging and authentication information of their services from the NO through a charging gateway (GW). This model has the following key characteristics:

--price of services defined by the network operator,

--billing and payment arranged by the NO,

--NO revenues gained as Airtime, Data volume, Message, Subscription, Advertising and Transaction / Event.

Concerning the discussions of streaming in 3G, this model is used, but firstly, in the next chapter, we will give a little review about it: "What the streaming is really?"

III. The Streaming from Mobile Operator Point of View

Streaming is a network technology to real time broadcast live or on demand multimedia contents (like audio and/or video) from one centralized streaming server to many clients on a wide network. This content delivery is a one way communication from the server to the clients (on the other hand, the video conferencing is a bidirectional communication among the users). This delivery is a unicast or multicast transport based on IP protocol. Certainly, the streaming uses bidirectional communication for the management between the clients and the server. By means of streaming, users can listen to the radio, watch TV or movie trailer on their mobile.

At present, there are many streaming solutions, such as: RealNetworks Helix, Windows Media Streaming, and Quick Time Streaming. Those use different protocols and formats, but the basic method of the technology is the same in either system. So, in this chapter, the expression "Streaming" is used without naming its real system.

Streaming has three, separate planes based on the same Network layer (Fig. 2) (in most cases, on IP): management, transport and quality control plane.

The Management plane allows the users to control the streaming, such as: connection opening and closing, getting information about the media, preparing the server for the broadcast and starting/pausing/stopping the delivery.

The transport plane makes the real time data delivery; it reserves the greatest part of available bandwidth. It must adjust oneself to the variable network parameters, so it cooperates with the quality control plane to avoid jitter and traffic congestion.

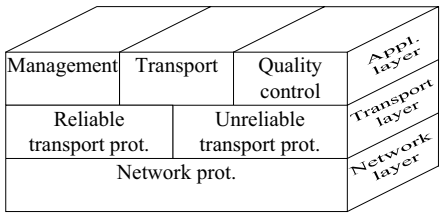

Fig. 2. Streaming protocols

The quality control plane measures the transport actual quality, and generates QoS reports for the both side.

IV. Charging of Streaming

In today's mobile networks do not exist universal, scalable, well uniform methods for charging of streaming. Most operators build their own charging procedures on the existing network usage measurement, so the streaming bills can not be separated from other IP services bills. In this case, the streaming "plays only a data volume generator's role" and the operators can not make an extra profit out of these services.

The GSM IP payment system (like NDS) is developed for WAP services, so it can handle the multimedia download, but the mobile equipment memory capacity limits the maximal size of the downloadable content. This solution is unsuitable for real time delivery, and applies only event based charging methods.

Streaming manufacturers carried out their own charging solutions (like Real SDS, see at [8]), but these are very platform dependent, and – in most cases – not centralized.

Solving the above problems, in this chapter, we present a proposal for CAB model in today's mobile networks, with the aim of creating a platform independent, centralized, well scalable architecture. So, in the following subsections, we offer a brief survey of charging requirements of streaming in 3G, and we describe our charging architecture and mechanisms.

A. Charging requirements

3GPP recommends three methods for charging of streaming in its technical specification [5]:

--charging by duration of session

--one-off set-up charge

--charging by volume of data, optionally QoS-differentiated

The great advantage of the duration based charging is that, it can be easy implemented, but it can not adjust itself to the temporary network parameters, like available bandwidth or packet loss. Two users with different access speeds pay for the same multimedia stream equally, but they have different visual experiences. Whereas, the volume based charging can adapt itself to the diverse network bandwidth, the users pay for, what they get. This is enough for HTTP or FTP services, which do not need delay sensitive protocols, but the streaming or real-time transfer requires permanent delay and reliable transport. If these

conditions are not fulfilled, the quality of streaming decreases, so QoS-differentiation must be used to charge these services.

The 3GPP describes in its report [see above] additional requirements, which we accentuate out of three:

--for network operators and 3rd parties to charge each other for the use of their resources

--to charge for different level of QoS applied for and/or allocated during a session for each type of medium or service used

--to charge using pre-pay, post-pay charging techniques.

B. *The architecture*

Over a classic 3G data session, the customer's communication begins with a PDP context establishment through the RNS, SGSN and GGSN [6]. The GSNs are able to separate the user's flows from one another with their PDP context, but they can not look inside the IP payload (clearly speaking, they do not decode the encapsulated protocol (UDP, TCP, RTP, etc…). So, the GSNs are unsuitable for a charging method, which is based on properties of these protocols. Therefore we introduce a new network element, which makes the existing network structure complete (Fig. 3).

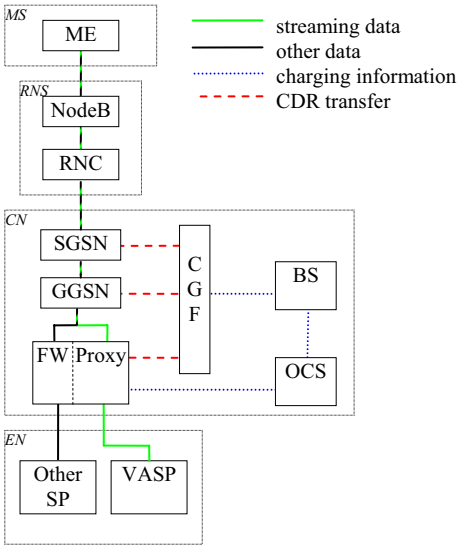

Fig. 3. Proxy based charging architecture

The customers can not connect directly to the VASPs after the PDP context activation, because we put a streaming proxy and a firewall near the GGSN, on its Internet side. The proxy monopolizes the management of streaming and it enables the users to connect through itself only to streaming providers (VASPs). Other services are accessible through the firewall, so its key function is only the access restriction (the proxy can be bypassed without it).

Offline charging information (for example: usage of radio interface, usage duration, destination and source) are collected in the Packet Switched domain network for each MS by the GSNs, and they are transferred in a CDR to the BS via the CGF. The generation of CDR depends on the charging characteristic profile, the GSNs are able to define separate trigger conditions, like data volume limit, time or maximum number of charging conditions [6]. In this architecture the proxy can measure the streaming flow due to its central position, so it has an own CDR generation function for the offline charging.

The pre-paid account is limited. If the amount of money runs out while a customer uses a service, the access must be denied for the customer. The offline charging mechanism can not provide this condition therefore the proxy uses account reservation for online charging. Firstly it reserves a piece of the customer's account. If the customers use up this amount of money for a value added service, then the proxy tries to reserve an other piece of the account. If the allocation of money is unavailable, the proxy restricts the access of service.

C. *Related mechanisms*

The radio interface between the RNC and MS is a really unreliable medium. It is possible to loss some data during a call session, so to protect from inaccurate charging, the 3G-SGSN will always instruct the RNC at RAB set-up to count the unsent downlink data toward the MS. At RAB release the RNC report [6] the unsent data to the SGSN, which send this value to the BS in the "RNC Unsent Downlink Volume" filed of a CDR.

In case of a streaming flow, a single frame loss (mostly, one frame represents one network packet) is probably imperceptible for the user, because the streaming server uses MPEG-4 (or similar) standard to encode the media stream. Since the available bandwidth in a mobile network is very limited and these algorithms archive high compression rate. Typically, the encoded frames have three different types (Fig. 4): "I-frames" are intra-coded images, coded independently of other frames. These are reference frames. "P-frames" are coded predictively from the closest previous reference frame, and "B-frames" are coded bi-directionally from the preceding and succeeding reference frames.

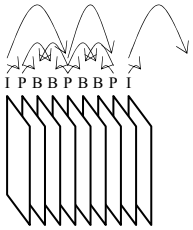

Fig. 4. Frame dependencies

The decoding of a multimedia stream depends on the successful receipt of the reference frames (I-frames). If a packet loss with an I-frame, then the decoding of dependent frames fails. This causes skewing and degrades heavy the quality. On the other hand, if a packet loss with a non I-frame, it causes a small, mostly imperceptible pixelisation on a small part of the display. If the loss distribution is bursty, then the longer the loss burst, the higher the probability of loss an I-frame. Moreover, the streaming technology can use a FEC algorithm, which can prevent only from the non-bursty packet loss. *So, one can realize that the loss distribution – and not the average loss rate – determines the quality of a multimedia stream.*

Today's charging mechanisms do not handle this problem, they measure only the summarized volume of unsent data (see above), and they do not use

QoS-differentiation. So, it is possible during two streaming calls that 10 single packet losses do not cause and 10 neighboring errors cause quality deterioration, however the two payments (the two numbers of successfully received packets) are equal.

We recognized this problem, and developed new mechanisms which are presented in Subsection D and E. These methods can determine the momentary loss distribution (e.g. the numbers of various length loss bursts), which is in proportion to the QoS of the stream. These values should be sent from time to time to the billing system in CDRs, which would be able to quantify the cash equivalent of the actual loss distribution. We give some advice in this decision in Subsection G.

D. Measuring the loss distribution

--*Measuring bursty packets at the RNC*: if the RNC measured not only the cumulative number of unsent data but also the occurrence of error runlengths, than it would send this extra information in a CDR to the billing system, which would determine the actual QoS class.

This is an exact method, but it measures only the QoS of the radio bearer. It can not consider the other side of the network, so it is usable, when the transmission between the RNC and VASP is "very reliable", namely the connection supports QoS (ATM, leased lines, etc…). In this case, the VASP must have a direct connection with the NO, so the 3rd party can not be charged with QoS-differentiation on the Internet.

--*Measuring bursty packets at the proxy*: most streaming protocols (like RTP/RTCP) have a built in QoS measurement procedure [9]: the receiver can report back periodically the actual QoS of the stream, for example: the cumulative number of the lost packets, the interval jitter, etc… The exact number of loss runlengths could not be calculated from these values, because the client does not send an acknowledgement after each packet, therefore these are time averages values. The RTP/RTCP standard enables to define new, profile dependent report elements, so if the receiver measured and stored the loss runlengths, it would be sent to the sender in a new report. Of course the proxy should understand this new type of reports, which should be converted to a CDR for the billing system.

This method is available to measure the loss bursts, when the mobile operator trusts the software of clients. But it is possible, if a hacker modified the media player, and if he (or she) sent back higher error number, than the real value, then he (she) would decrease the payment. To prevent from cheating, the NO has to use client authentication, but only those users can be served, whose media player supports this authentication method. This limits the number of marketable mobile equipments.

E. Estimating the loss distribution

Usually the QoS is specified in few classes, so mostly the exact number of lost packets is not needed to be known at charging, it is enough, when a few parameters of the streaming is given, which the loss distribution can be deduced or predicted from. So, we suggest that the continuous measuring of the packet loss is pointless, it enough to take samples of the packet transfer for time to time, and the loss runlengths should be estimated between two sampling periods from these samples. Clearly speaking, the client (or the proxy) should take 1000 samples from the stream; it should calculate some parameters from these, which the loss distribution can be calculated from for the next 10000 packets. To determine these parameters, we present three models for packet loss metric.

H. Sanneck and G. Carle use two Markov models in their paper [7] to predict temporal loss dependency. Firstly, we describe the two-state Markov model (known as Gilbert model) (see Fig. 5).

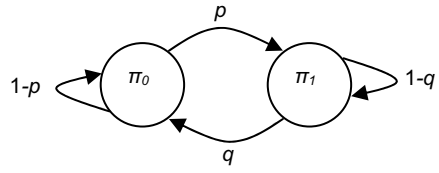

Fig. 5. Gilbert model

The Gilbert model has only two states with the following probabilities: π_0 represents the event that a packet is received successfully while π_1 refers to the opposite. p denotes the probability the packet is lost, provided that the next is not lost, q is the opposite. 1-q is called *conditional loss probability* (clp). If $p+q=1$ the Gilbert model reduced to the Bernoulli model.

The state probabilities can be calculated in the following way:

$$\pi_0 = \frac{q}{p+q},\ \pi_1 = \frac{p}{p+q} \tag{1}$$

p and q can be computed from the streaming data flow, using the loss length distribution statistics [7]:

$$p = \frac{\sum_{i=1}^{\infty} m_i}{m_0},\ 1-q = \frac{\sum_{i=1}^{\infty} m_i(i-1)}{\sum_{i=1}^{\infty} m_i i}, \tag{2}$$

where m_i, i=1…n-1denotes the number of loss burst having length i, and n-1 is the maximal length of bursts.

The probability distribution of loss runs (p_k) with length k has geometric distribution

$$p_k = (1-q)^{k-1} q, \tag{3}$$

where p (mean loss rate) and 1-q (conditional loss probability) can be used to describe the QoS.

The third model is the extended Gilbert model [7]. It remembers not only the last one but the last n-1 events, therefore it needs n states to describe these n events (Fig. 6), because only the past n consecutive loss events will affect the future, if the correlation between two loss bursts will be small.

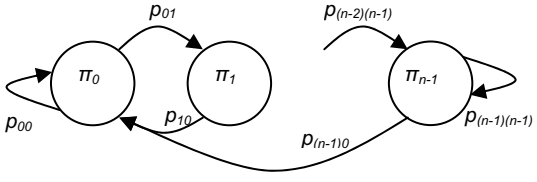

Fig. 6. Extended Gilbert model

The transition matrix of the extended Gilbert model is the following

$$\begin{bmatrix} 1-p_{01} & 1-p_{12} & 1-p_{23} & \cdots & 1-p_{(n-2)(n-1)} & 1 \\ p_{01} & 0 & 0 & \cdots & 0 & 0 \\ 0 & p_{12} & 0 & \cdots & 0 & 0 \\ 0 & 0 & p_{23} & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & p_{(n-2)(n-1)} & 0 \end{bmatrix} \quad (4)$$

We can compute the state probabilities like this

$$\underline{\underline{T}} \begin{bmatrix} \pi_0 \\ \pi_1 \\ \vdots \\ \pi_{(n-1)} \end{bmatrix} = \begin{bmatrix} \pi_0 \\ \pi_1 \\ \vdots \\ \pi_{(n-1)} \end{bmatrix}, \sum_{i=0}^{n-1} \pi_i = 1. \quad (5)$$

The probability distribution of loss runs (p_k)

$$p_{01} = \sum_{i=1}^{n-1} m_i \Big/ m_0 \,, p_{(k-1)(k)} = \sum_{i=1}^{n-1} m_i \Big/ \sum_{i=k-1}^{n-1} m_i \,. \quad (6)$$

The state probabilities ($\pi_k ... \pi_{n-1}$) can determine the QoS-class.

These methods enable to decrease the load of the measuring node and to give the capability to predict the fluctuation of the customer's account.

F. Model evaluation

To evaluate and compare the above described models, we captured a real streaming flow over GPRS calls with the following properties (TABLE I).We could not measure over 3G, because it was inaccessible for us.

TABLE I
STREAM PROPERTIES

Properties	*Value*
Encoder	MPEG-4
Bit rate control	CBR
Bit rate	23 kbps
Key frame period	5000 s
Error resilience	none
Size	QCIF
Max. packet size	400 byte

We compared how to fit the three probability distributions of loss runs to the occurrence of real loss bursts. So, we computed the state probabilities and state transitions oat the first 1000 packets with the equations (1), (2), (4), (5), and we computed the predictions (3), (6) for the next 10000 packets (the estimated occurrences of loss runlengths can be seen in TABLE II).

TABLE II
LOSS RUNLENGTHS OCCURENCE

Model	Loss burst length				
	1	2	3	4	5
Measured	408	46	8	5	1
Bernoulli	544,40	4,56	0,04	0,00	0,00
Gilbert	402,32	59,36	8,76	1,29	0,19
E. Gilbert	411,45	46,39	8,07	5,04	1,01

The Bernoulli model is a stateless model, so it can not predict the time-relations between errors; it overestimates the occurrence of single losses, and underestimates the multiple losses.

The Gilbert model approaches the real occurrence in a good way, when the maximal length of the error burst is smaller than 4, but it underestimates the multiple losses, which causes the quality deterioration of streaming (it has only 2 states for prediction). So this model is not usable for charging, because it does not predict well the significant error bursts.

The extended Gilbert model makes the closest prediction, but it requires as many states as the maximal loss runlengths is. Its state probability can make the base of the QoS-differentiation for charging.

G. Using the loss distribution for charging

We determined the loss distribution in the above, but afterwards an algorithm is needed which can classify the QoS from the probability of loss bursts. This can be based on limits or on a special function, which depends on subjective decisions (for example: on the marketing strategy of the network operator, on the content of the stream, etc…), therefore in this paper we do not discuss these methods.

V. CONCLUSION

We have presented a new, streaming proxy based charging architecture for QoS-differentiated charging in 3G networks. The recommended solution is available for online and offline charging. We have pointed out the importance of bursty packet loss, which determines the quality of streaming, in contrast with the average loss rate. We have proposed several methods for QoS-differentiated charging, which are in compliance with the recommendations of 3GPP. We have described three models, which could predict the QoS of a streaming flow. The algorithms were evaluated them on a real GPRS streaming flow.

In the future we would implement our charging method in a simulation environment and we plan make an experimental charging system, which can be used to evaluate the developed methods.

REFERENCES

[1] Lucas Baugé, "GPRS billing: getting ready for UMTS", SITCOM, 2002

[2] M. Koutsopoulou, E. Gazis, A. Kaloxylos, "A Novel Billing Scheme for UMTS Netowrks", International Symposium on 3rd Generation Infrastructure and Services, Athens, Greece, 2001

[3] H. Tewari, Donald O'Mahony, "Real-Time Payments for Mobile IP", Management of Next Generation Wireless Networks and Services, IEEE Communications Magazine 126-136, Feb. 2003

[4] "Charging, Billing and Payment Views on 3G Business Models," UMTS-Forum, 2002 v4.4a

[5] Service aspects; Charging and billing (Release 5), 3GPP technical specification TS 22.115 v5.2.0, 2002-03.

[6] Charging Management; Charging principles,, 3GPP technical specification TS 32.200 v5.7.0, 2004-06.

[7] H. Sanneck, G. Carle, "A Framework Model for Packet Loss Metrics Based on Loss Runlengths," presented at the SPIE/ACM SIGMM Multimedia Computing and Networking Conference, San Jose, CA, 2000

[8] Helix Service Delivery Suite www.realnetworks.com

[9] H. Schulzrine, RTP: A Transport Protocol for Real-Time Applications, www.rfc.org