Providing QoS in General Packet Radio Service

ZORAN PERISIC[(1)], ZORAN BOJKOVIC[(2)]

[(1)] Serbia & Montenegro army
Raska bb, 11000 Belgrade

[(2)] Faculty of Transport and Traffic Engineering
UNIVERSITY OF BELGRADE
Vojvode Stepe 305, 11000 Belgrade

Abstract- In this paper we proposed dynamic adaptive guaranteed quality-of-service (QoS) provisioning scheme over General Packet Radio Service (GPRS) wireless mobile links. The guaranteed QoS media access control (GQ-MAC) protocol and the accompanying adaptive prioritized-handoff call admission control (AP-CAC) protocol are used to maintain QoS guarantees. The effect of mobile handoffs are also taken into account.

Key Words: - General Packet Radio Service, Global System for Mobile Communication, protocols

1. Introduction

GPRS is a Global System for Mobile communications (GSM) service that provides mobile subscribers with performance guaranteed packet data services over GSM radio channels and external packet data networks. The GPRS wireless subsystem consists of mobile stations (MS's) contending for access to a base station (BS) in a radio cell, with traffic generated according to the negotiated QoS profiles, defined in terms of precedence, delay, reliability, mean and peak throughputs. QoS profiles can be classified as streaming, conversational, interactive and background, with attributive values described in Tabele 1 [7].

The wireless link is characterized by a broadcast mode in the downlink (BS to MS) and a multiple access mode in the uplink (MS to BS). A medium access control (MAC) protocol distributes packet transmission over the shared medium among all users [6]. The GPRS standard [1] specifies the FDD/TDMA multiple access with four radio access priorities, and gives reference guidelines on the resource sharing method. Resource sharing based on demand assignment can be employed to minimize wasted bandwidth due to underutilization with dedicated assignment and to collision with random access. With Packet Reservation Multiple Access (PRMA) [2] protocol, voice source uses slotted ALOHA (a type of TDMA transmission system developed by the University of Hawaii used for satellite and terrestrial radio links) to reserve the same slot position in future frames and data sources contend for a slot when they have packets to send.

Enhanced PRMA protocols (such as Centralized-PRMA [3] and Integrated-PRMA [4]) improve channel efficiency and provide some kind of service fairness for data sources. Since all these protocols suffer variable packet access delay, QoS with bounded delay could not be guaranteed. This paper proposes the Guaranteed QoS Medium Access Control (GQMAC) protocol to enable performance guarantees for the four defined GPRS QoS classes. The protocol supports per-session dedicated reservation for streaming traffic class and prioritized on-demand reservation for conversational and interactive traffic classes [5]. Traffic burstiness is counteracted with dynamic adaptive resource allocation with peak bandwidth allocation adapted to the current queue length.

Traffic Classes	Latency	Jitter	Loss	Throughput	Burstiness
Streaming	Bounded (<500 msec)	Stringent	Tolerable (<10^{-2})	Guaranteed	Low
Conversational	Bounded (<80 msec)	Stringent	Tolerable (<10^{-2})	N/A	High
Interactive	Less than Conversational	N/A	Sensitive (<10^{-5})	Guaranteed	Greater than Conversational
Background	N/A	N/A	N/A	N/A	N/A

Table 1. QoS profile

In a mobile wireless system, the call admission control (CAC) protocol should give prioritized admission to handoff requests to enable lower blocking probability relative to new calls, since forced terminations of ongoing call sessions, due to mobile handoff blocking, are generally more objectionable than new calls blocking from the user's perspective. The static guard channel scheme supports prioritized admission by assigning greater channel capacity limits to handoff requests over new calls. Dynamic guard channel scheme adapts the number of guard channels in a radio cell according to the current estimate of the handoff arrival rate derived from the current number of ongoing

calls in neighboring cells and the mobility pattern. The goal is to keep the handoff block probability close to the targeted objective while constraining the new call blocking probability to be below a given level.

In this paper dynamic guard channel scheme is extended for multiple traffic classes. The handoff blocking probabilities are minimized at any cost without considering the degradation of the new call blocking probability. We propose the Adaptive Prioritized handoff CAC (AP-CAC) protocol, which extends the multiple dynamic guard channel scheme to provide admission control for multiple traffic classes with the objective to minimize handoff blocking probabilities while minimizing the degradation to the new call blocking probability.

2. GQ-MAC protocol

2.1 . Channel Access Procedures

The GPRS packet data channels (PDCH) are classified as follows: (1) *Packet Random Access Channel (PRACH)* is the request access channel for uplink, consisting of two time multiplexed channels of Signaling PRACH (S-PRACH) and User-data PRACH (U-PRACH); (2) *Packet Access Grant Channel (PAGCH)* is the request acknowledgement channel for downlink, which is used by the BS to broadcast the request status information; and (3) *Packet Data Traffic Channels (PDTCH)* are the remaining PDCH's to carry the uplink/downlink payload. The request access priorities of the S-PRACH and U-PRACH are illustrated as follows:

Request Acces Type		Priority	PRACH
Signaling	New call	Low	S-PRACH
	Handoff	High	
User data	Conversational traffic	High	U-PRACH
	Interactive traffic	Low	

Table 2. Access priorities

Admitted streaming traffic sources enjoy per-session dedicated reservations, while admitted conversational and interactive traffic sources have to perform on-demand reservation by multi-accessing the U-PRACH. When collision occurs, only the conversational traffic sources are allowed to multi-access the U-PRACH via the tree limited contention access protocol. Consequently, conversational class is given a higher U-PRACH access priority than interactive class. The background traffic sources are allocated with unused PDTCH's in a round robin fashion. The access protocols of the GQ-MAC are described as follows.

2.1.1. Slotted ALOHA
This is used by the signaling requests of new calls and handoffs to access the S-PRACH. MS's with signaling packets transmit immediately on the first available slot. If a collision occurs, they transmit on the next slot with probability "Pa". Handoff requests are given higher priority by having higher "Pa" value.

2.1.2. Tree Protocol
This limited-contention access protocol is used for in session channel access request for conversational traffic because it provides a deterministic channel access time. By allowing only conversational traffic sources to participate in the contention resolution cycle, it is possible to guarantee a bounded delay on channel access. The contention cycle can be showed to have a bounded length of:
TDMA_FRAME_LENGTH $* (2^{(\log 2j + 1)} - 1)$; Where j is the number of conversational MS's, simultaneously trying to access the U-PRACH. Based on the on-off model for voice, with 40% voice activity, it can be shown that the probability of more than 5 MS's trying to access the U-PRACH simultaneously is very low, thus limiting the in-session channel access delay to 20 msec (assuming GSM frame size).

2.1.3. Modified Slotted ALOHA
This is used for U-PRACH access by interactive traffic sources. It is similar to slotted ALOHA protocol, except that when a collision occurs, a binary exponential back-off algorithm is used which reduces the transmission probability, "Pa", in the next slot by 0.5 (for 8 max. retries). When a contention resolution cycle is in progress, the value of "Pa" is reduced to "0", i.e. interactive MS's do not participate in the contention resolution cycle.

2.2. QoS Support

To initiate a new call, a Call Initiation (Call_Init) request is sent on the S-PRACH using slotted ALOHA. The Call_Init request contains one or more of the following: Desired service type (Conversational, Streaming, Interactive, or background) and Requested Data Rate (RDR). On successful reception of this request, the PH-CAC module determines if enough resources are available to support this new call, and if admissible, sets up two state variables for that session in the BS via RDR and ADR (Achieved Data Rate). A temporary buffer is also set up to hold packets for that session and a suitable MS identifier is generated, which is transmitted to the corresponding MS on the PAGCH. We now discuss how each traffic type can be supported in the system.

2.2.1. Streaming Traffic

The system offers streaming as a dedicated service in multiples of quantized data rates. Thus for a streaming rate 'X', one full PDTCH is allocated, while for rate '0.5X', only half of the PDTCH (a slot in every alternate frame) will be allocated and so on. Since resources are permanently allocated to a streaming call, it faces the problem of multi-access only during call set up. Thus a streaming call, once admitted, is guaranteed a bounded packet delay, constant inter-packet delay (i.e. minimal jitter) and a guaranteed throughput. By using a suitable FEC scheme, the packet loss due to corruption on the wireless link is limited.

2.2.2. Conversational Traffic

Upon admission, a conversational traffic source demands a channel resource only when it has data to send. The request is sent on the U-PRACH using the tree protocol. If resource is unavailable, the BS will reallocate resources allocated to other sources (except streaming) to this conversational source. If all the resources are currently allocated to other conversational sources, the BS rejects the resource request, and the MS has to send another request after discarding the first packet in the queue. When there is no data, an allocated resource is held for a channel holding time of 3 TDMA frames, following which an explicit release message is sent. Thus by using the tree protocol for channel access, packet delay for a conversational traffic is guaranteed to be bounded. Since resource is reserved till it is explicitly released, the inter-packet delay is guaranteed to be constant.

2.2.3. Interactive Traffic

For interactive traffic sources, a scheduling algorithm is required which can guarantee the required throughput. We propose a distributed scheduling algorithm for allocation of uplink PDTCH's. In the proposed algorithm, MS takes active part in the scheduling process on the uplink. For every interactive stream that is admitted into the cell, the BS and the corresponding MS maintain the following state variables: RDR which is sent in the Call-Init request packet, and ADR which is continuously updated by MS (BS) as it sends (gets) data packets. Every MS maintains a queue at its output interface having a finite length. After getting admitted into the cell, the MS sends a Rate Request Packet (RRP) on the U-PRACH requesting some number of PDTCH's, closely matching its RDR value. The BS attempts to allocate as many PDTCH's requested as available. For this, it might even pre-empt other similar sources, which have achieved $ADR \geq RDR$.

If the MS is allocated a rate, which is less than its packet generation rate, packets will start queuing at the output and the queue length increases. Depending on the queue length the MS will send another RRP, demanding higher rate, with access probability $P(x)$ as:

$$P(x) = e^{\frac{(x/L_U)-1}{(1-\alpha)}} \ \forall x\varepsilon\{1..L_U - 1\}; and P(x) = 1 \forall x > L_U \ (1)$$

'L_U' is a fixed upper threshold and factor 'α' is directly proportional to the ADR/RDR value. This mechanism allows dynamic resource allocation for bursty traffic with peak bandwidth allocation adapted to the current queue length.

2.2.4. Background Traffic

After getting admitted and allocated an identifier, the background traffic sources camp on the PAGCH to see which slots are allocated to them in the uplink frames. The BS allocates unused PDTCH's to background traffic sources in a round robin fashion.

2.3. Performance Analysis and Results

We simulated a single cell containing one BS and a number of MS's, with the following simulation parameters as it is shown in Table 3.:

Parameter	Value
No. of Uplink/Downlink carrier pair	1
TDMA frame duration	4.615 msec
No. of time slots in a TDMA frame	8
No. of traffic channels/carrier	7
Channel data rate	270 kbps
Average length of talkspurt	1 sec (216 frames)
Average length of silent periods	1.35 sec (292 frames)
Avg. msg. inter-arrival time	Varied as 2,3, and 5 msec
Data rate	56, 37. 33, 22.4 kbps
Average data message size	112 bits
Simulation time	10 minutes (13K frames)

Table 3. Simulation parameters for a single cell containing one BS and a number of MS`s.

Streaming and background traffic types were not simulated because streaming calls with dedicated reservations would only reduce the capacity of the system without affecting other results. Since no guarantees are given to background flows, their presence or absence do not affect the results.

Simulation experiments were carried out for conversational and interactive traffic types. An on-off model is used to simulate the conversational source. Voice packets are dropped if they exceed a limit of 60 msec. An ideal wireless channel is assumed, hence packet loss occurs only because of deadline violation. The interactive data users generate packets according to a Poisson distribution, with rates ranging from 22.4 to 56 Kbps. Since interactive traffic is not real-time, no packet is discarded due to excessive delay. Simulations for two different cases are described as follows.

2.3.1. Conversational traffic only
With only conversational traffic sources present in the cell, it was found that in order to get a good mean opinion score, the number of admitted users must be limited to 11, as 95% of the users experienced a packet dropping probability (PDP) of less than 2%. Thus a multiplexing gain of 1.6 (approx.) was obtained. Also the average channel access delay and average packet delay was found to be limited to 20 msec as predicted before.

2.3.2. Integrated Conversational and Interactive traffic
Fig. 1 shows the cumulative distribution function (CDF) of the average voice PDP, for 11 conversational MS's (denoted as 'Voice') and the interactive MS's (denoted as 'Data') are varied from 0 to 9. It can be seen that there is no significant change in the voice PDP. This is due to the tree protocol used by conversational mobiles for in-session channel access on

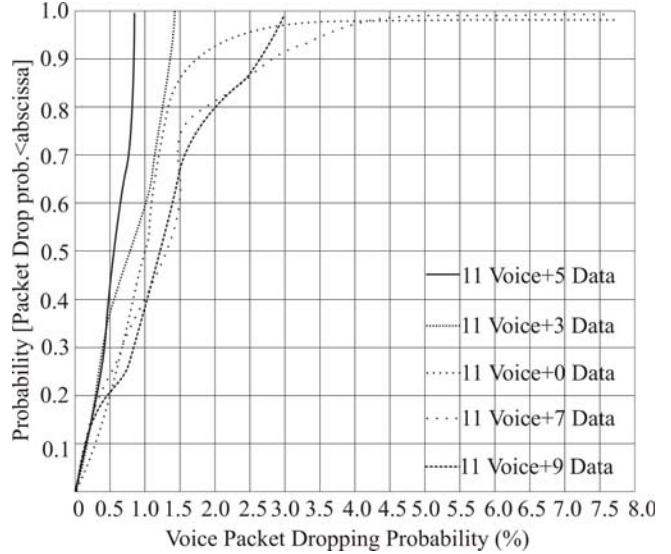the U-PRACH, which gives them sufficient isolation from others.

Fig. 1. CDF of Voice Packet Dropping Probability.

The CDF for the average access time is also shown in Fig. 2. This graph shows that the in-session channel access time for conversational MS's is relatively independent of the number of interactive mobiles present in the cell.
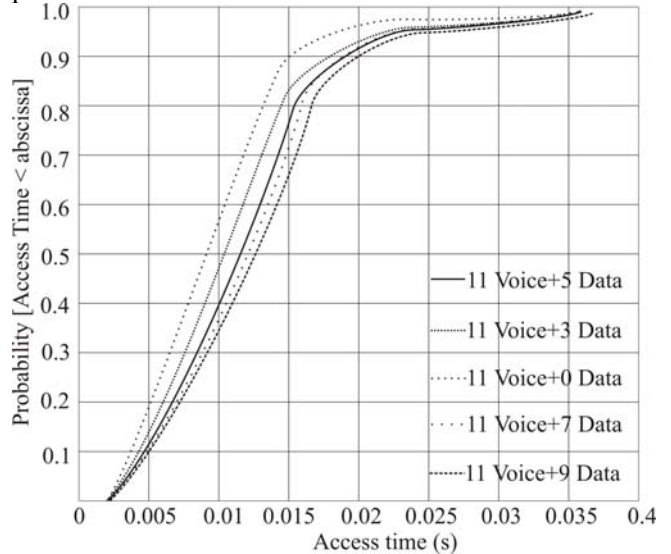

Fig. 2. CDF of average channel access time.

Fig. 3 illustrates the CDF of the average packet delay faced by conversational mobiles. As observed, this is also limited to 20 msec and is also independent of number of interactive mobiles present in the cell.
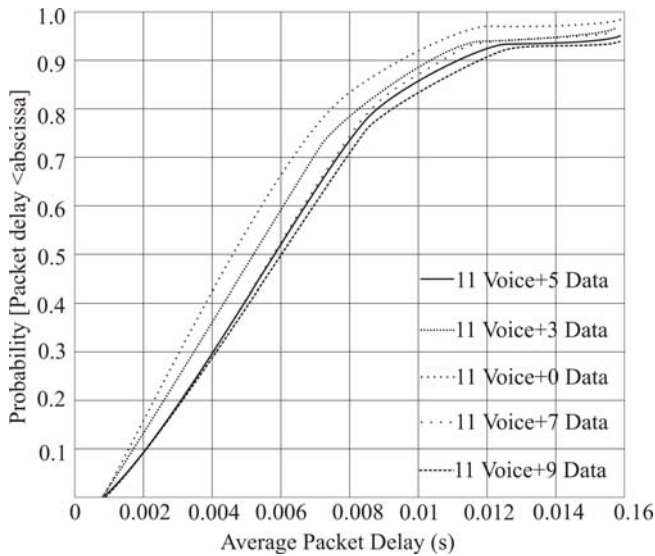
Fig. 3. CDF of average packet delay.

3. AP-CAC protocol

3.1. Operations and procedures

The AP-CAC protocol supports multiple dynamic admission priorities for handoff requests of multiple traffic QoS classes over new calls. Two levels of guard channels (NG1 level and NG2 level) are used to support a three-priority level admission scheme, with two premium priority levels for handoff requests over the base priority level for new calls. The three admission priority classes in ascending priority order are as follows: (1) class 'N' is associated with new calls, which are admitted only when the number of free channels exceeds NG1; (2) class 'H1' is associated with handoff requests of interactive, conversational or background traffic, which are admitted only when the number of free channels exceeds NG2; (3) class 'H2' is associated with handoff requests of streaming traffic, which are admitted whenever a free channel is available.

Each level of guard channels is continuously adapted according to the instantaneous estimate of the handoff request arrival rate of the corresponding traffic class, which depends on the number of active MS's with ongoing calls in the neighboring cells, the mobility patterns of the active MS's in terms of speed and direction during the estimation interval, the size of the cells currently resided by the active MS's and the remaining call duration of the ongoing calls.

For a given total number of channels NT, the ranges of the time-varying NG1(t) and NG2(t) are given as: $0 \leq$ NG2 (t) $\leq$ NG2max $= \beta2NT$; NG2(t) $\leq$ NG1(t) $\leq$ NG1max $= \beta1NT$; where, $0 < \beta2 \leq \beta1 \leq 1$. In deriving

the blocking probability for each of the three proposed admission priority classes, the arrival processes for H2 handoffs, H1 handoffs and new calls are assumed to be Poisson with time varying rates of λH1(t), λH2(t) and λN(t) respectively. The departure processes are also assumed to be Poisson with a constant rate of μ. The blocking probabilities BH2(t) and BH1(t) for H2 and H1 handoff calls, and BN(t) for new calls are given as follows:

$$B_N(t) = \sum_{j=N_T-N_{G1}(t)}^{N_T} P_j(t); \; B_{H1}(t) = \sum_{j=N_T-N_{G2}(t)}^{N_T} P_j(t); \; B_{H2}(t) = P_{N_T}(t)$$

$$let \; \lambda_\Delta(t) = \lambda_{H1}(t) + \lambda_{H2}(t) + \lambda_N(t); \; \lambda_\Omega(t) = \lambda_{H1}(t) + \lambda_{H2}(t);$$

$$N_\Delta(t) = N_T - N_{G1}(t); \; N_\Omega(t) = N_T - N_{G2}(t);$$

then $P_J(t)$

$$= \lambda_\Delta(t)^j P_0(t) / j!\mu^j \qquad \forall j \in \{1 \, to \, N_\Delta(t)\}$$

$$= \lambda_\Delta(t)^{N_\Delta(t)} \lambda_\Omega(t)^{j-N_\Delta(t)} P_0(t) / j!\mu^j \qquad \forall j \in \{N_\Delta(t)+1 \, to \, N_\Omega(t)\}$$

$$= \lambda_\Delta(t)^{N_\Delta(t)} \lambda_\Omega(t)^{N_{G1}(t)-N_{G2}(t)} \lambda_{H2}(t)^{j-N_\Omega(t)} P_0(t) / j!\mu^j \quad \forall j \in \{N_\Omega(t)+1 \, to \, N_T\}$$

$$with \; P_0(t) = \left[\sum_{i=0}^{N_\Delta(t)} \frac{\lambda_\Delta(t)^i}{i!\mu^i} + \sum_{i=N_\Delta(t)+1}^{N_\Omega(t)} \frac{\lambda_\Delta(t)^{N_\Delta(t)} \lambda_\Omega(t)^{i-N_\Delta(t)}}{i!\mu^i} + \sum_{i=N_\Omega(t)+1}^{N_T(t)} \frac{\lambda_\Delta(t)^{N_\Delta(t)} \lambda_\Omega(t)^{N_{G1}(t)-N_{G2}(t)} \lambda_{H2}(t)^{i-N_\Omega(t)}}{i!\mu^i} \right]^{-1}$$

3.2. Simulation and Results

The simulated cellular network model has concentric layers of rural, suburb and city cells over the central downtown core cell. The transit probability PX of an active MS from one cell to another depends on the time of the day.

For example during morning rush hours, PX is set as follows:

downtown-to-city 0.167, city-to-city 0.1, city-to-downtown 0.7, city-to-suburb 0.1, suburb-to-suburb 0.167, suburb-to city 0.5, suburb-to-rural 0.167, rural-to-rural 0.5, and rural-to suburb 0.167.

The calls of H1 and H2 classes are assumed to have 0.8 and 0.2 occurrence probabilities respectively. We assume their unencumbered call durations to be exponential with respective means of 150s and 300s, and the call holding times to be exponential with respective means of 120s and 240s.

New call arrival rates under stationary and non-stationary traffic conditions have the same long-term nominal rate of 0.475 call/second. Under non-stationary traffic condition, the average new call arrival rate during morning/afternoon rush hours increases to 0.8

call/second, and varies among 0.2, 0.3 and 0.6 call/second during other hours. The target objectives of BH2 = 0.003, BH1 = 0.05, and BN = 0.1 are set given that NT = 100.

The performances of the reference (multiple static guard channel) scheme and the AP-CAC (multiple dynamic guard channel) scheme are compared under non-stationary traffic condition. The reference scheme employs static guard channels of NG1 = 2 and NG2 = 1. The AP-CAC scheme would try to track the target objectives (allow 50% tolerance for the BH1 and BN soft targets, but no tolerance for the BH2 hard target) by adapting the guard channels dynamically. With the reference scheme, BN deviates from the target objective by 40% to 0.14, BH1 deviates from the target objective by 25% to 0.06, and BH2 deviates from the target objective by 300% to 0.05. With the APC-CAC scheme, BN deviates from the target objective by 70% to 0.17, but BH1 and BH2 meet the target objectives.

Under the non-stationary traffic condition, the results show that the AP-CAC scheme is able to maintain the handoff blocking probabilities at the target objectives, while the reference scheme fails to accomplish that. The AP-CAC scheme does cause the new call blocking probability to deviate greater from the target objective than that of the reference scheme. The reference scheme fails to achieve all the target objectives, while the AP-CAC scheme is able to meet the target objectives of the prioritized handoff blocking probabilities at the reasonable expense of the new call blocking probability.

4. Conclusion

Performance analysis of the GQ-MAC protocol shows that it is capable of providing guaranteed QoS performances for streaming, conversational, interactive and background traffic classes over GPRS wireless links while optimizing channel resource utilization. Performance analysis of the AP-CAC protocol shows that it is capable of maintaining QoS performance guarantees under the effect of mobile handoffs by providing dynamic adaptive prioritized admission control for multiple traffic classes via the multiple dynamic guard channel scheme, which dynamically adapts the capacity reserved for dealing with handoff requests based on the current number of ongoing calls in the neighboring radio cells and the mobility pattern. The results show that handoff blocking probabilities can be minimized while minimizing the degradation to the new call blocking probabilities for multiple traffic classes.

References:
[1] GSM 04.60: "Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Mobile Station (MS) – Base Station System (BSS) interface; Radio Link Control / Medium Access Control (RLC/MAC) protocol," version 7.4.1, Release 1998.
[2] D. J. Goodman, R.A. Valenzuela, K.T. Gayliard and B. Ramamurthi, "Packet Reservation Multiple Access for local wireless Communication", *IEEE Trans. on Comm.*, Vol 24, pp. 254-260, August 1989.
[3] G. Bianchi, F. Borgonovo, L. Fratta, L. Musumeci and M. Zorzi, "CPRAMA: A Centralized Packet Reservation Multiple Access for Local Wireless Communications", *IEEE Trans. on veh. Tech.*, vol. 46, pp: 422-436, May, 1997.
[4] P. Ramanathan, K. M. Sivalingam, P. Agrawal and Kishore, "Dynamic Resource Allocation Schemes During Handoff for Mobile Multimedia Wireless Networks," *IEEE JSAC,* vol.17, pp.1270-1283, July 1999.
[5] K. R. Rao, Z. S. Bojkovic, ``Packet video communications over ATM networks``, Prentice-Hall, PTR, 2000.
[6] K. R. Rao, Z. S. Bojkovic, D. A. Milovanovic, ``Multimedia communication systems, Prentice-Hall, PTR, 2002.
[7] K. R. Rao, Z. S. Bojkovic, D. A. Milovanovic, ``Introduction to multimedia communications: applications, middleware, networking``, Wiley 2005.