Quality of Service and Admission Control within the UMTS Terrestrial Radio Access Network

Rani Makké and Samir Tohmé
ENST-Paris, InfRes department
46, rue Barrault, 75634
Paris cedex 13, France
Tel: +33 1 45 81 75 39 / 78 61
Fax: +33 1 45 81 31 19
{makke, tohme}@enst.fr

Jean-Yves Cochennec
France Télécom R&D
2, av. Pierre Marzin,
22300 Lannion, France
Tel: +33 2 96 05 20 22
Fax: +33 2 96 05 18 52
jeanyves.cochennec@rd.francetelecom.fr

Sophie Pautonnier
Mitsubishi Electric ITE-TCL
80, av. des Buttes de Coemes,
35700 Rennes, France
Tel: +33 2 99 84 21 15
Pautonnier@tcl.ite.mee.com

Abstract: In this paper, we present some performance and dimensioning key issues of the AAL2 protocol in the context of the UTRAN. First, we propose to define three classes of service at AAL2 layer in order to provide the differentiation of services between different types of applications. We propose a connection admission control scheme for AAL2 channels and we present some important aspects related to the optimal Timer-CU value, the scheduling mechanism in the AAL2 multiplexer and the AAL2 switching technology. Simulation results are provided in order to validate our proposal.

Keywords: UMTS, UTRAN, AAL2/ATM, Quality of Service, Admission Control

1. Introduction

In 1997, ITU-T has standardized the ATM Adaptation Layer Type-2 (AAL2) [1] which allows multiplexing of several "mini-cells" in one ATM cell. The AAL2 is therefore well suited for transport of low bit rate real-time applications. Another advantage of AAL2 is its switching capability as each AAL2 mini-cell contains a header including an AAL2 channel identifier. In UMTS, AAL2 was chosen as a transport technology for the radio access network UTRAN in the release 99 of 3GPP standard. As delay requirements are very strict in the access network, it is important to evaluate accurately the performance of the AAL2 protocol especially the AAL2 switching. However, there were few studies so far to determine delays introduced by AAL2 switching. This paper presents the first results of the MINICEL project [2] in which a prototype including a UTRAN environment simulator and an AAL2 switch has been developed.

The AAL2 protocol has been standardized to transport variable low bit rate applications with real-time constraint (e.g. the compressed voice). The AAL2 was defined to solve the problem of the ATM cell packetization delay that becomes critical for the low bit rates (at 16kbps, its value is 24 ms). The solution is simple: when multiplexing several communication flows in the same ATM channel, the delay becomes reasonable for a given communication. The AAL2 protocol consists of variable length data units called mini-cells with a maximum payload length of 45 bytes (optionally 64 bytes). The AAL2 layer is divided into two sub-layers: the SSSC [3] and the CPS. The SSSC is divided into three sub-layers: SS-ADT, SS-TED and SS-SAR. Only the SS-SAR [4] is used in UTRAN. It segments higher-level data units exceeding 45 bytes into packets with a maximum length of 45 bytes (optionally 64 bytes). The CPS layer overloads each mini-cell by 3-bytes header and inserts it in the payload of the ATM cell. Then the Timer-CU is armed. The cell is then transmitted either when it is full or when the Timer-CU has expired. If the Timer-CU expires before the cell becomes full, padding is added and the cell is transmitted. Overlapping is used: one mini-cell inserted in the ATM cell can overlap onto the next cell. The structure of the AAL2 header (CPS-Header) is showed in Figure 1.

![CPS-Header format](image)

Figure 1: CPS-Header format

The CID field identifies the AAL2 connection. There are 256 possible values, 8 of them are reserved for signalling purposes and the rest can be used to identify 248 different AAL2 connections. The LI field determines the mini-cell payload length. By default, the maximum length is 45 bytes but it may be 64 bytes if there is an indication at the connection establishment procedure. The UUI field is assigned to the SSSC sub-layer. The HEC is used for error detection in the mini-cell header. The ATM header allows two levels of addressing (VPI and VCI). Thus, it is possible to set up several ATM PVCs between AAL2 endpoints and to allow them to use VCI and CID to create multiple native connections.
2. Performance key issues

In the UMTS specifications, four traffic classes are defined for different types of applications: Conversational, Streaming, Interactive and Background. Differentiation between classes is based on time constraints and transport reliability. Within the UTRAN, radio channels are extended between the user equipment and the RNC and they are transported on Iub and Iur interfaces where the AAL2 protocol is chosen as transport technology in the release 99. The AAL2 layer must offer the possibility to guarantee QoS requirements for UMTS traffic classes handled by radio channels. Since several classes are defined for UMTS traffic, AAL2 layer should provide differentiation between different types of flows. In the paper, we propose three classes of services at AAL2 level and we assign a certain number of QoS parameters to each class.

In order to guarantee QoS requirements, connection admission control is needed for AAL2 channels. In this paper, we propose a connection admission control scheme and a bandwidth management scheme in order to share the available bandwidth between different traffic classes. At the AAL2 layer, differentiation between different services is needed in order to guarantee QoS requirements for each type of traffic. We propose different scheduling mechanisms at AAL2 layer and we study the advantages and drawbacks of each one in order to choose the optimal algorithm.

2.1 QoS parameters at AAL2 level

In order to specify AAL2 classes, QoS parameters are needed at AAL2 layer. We define three parameters as follow:

- **MTD**: Mini-cell Transfer Delay. This is the transport delay of a packet between two AAL2 SAPs including the assembly delay of mini-cells in case of overlapping.
- **MDV**: Mini-cell Delay Variation. This is delay variation between two consecutive mini-cells.
- **MLR**: Mini-cell Loss Ratio. This is the ratio of lost mini-cells to total number of sent mini-cells.

2.2 Classes of service at AAL2 layer

**AAL2 classes**: In the 3GPP standard, four traffic classes are defined for UMTS applications: Conversational, Streaming, Interactive and Background [5]. This classification is based on the time constraint for each type of flows. The AAL2 protocol will be deployed on the Iub [6,7] and Iur [7] interfaces within the UTRAN. The services offered by the UMTS will use the services offered by the underlying AAL2/ATM layer. In order to differentiate between flows at AAL2 layer, we define three service classes:

- The **AAL2-Class 1** used for real-time applications with stringent time constraint (e.g. voice). For this class, stringent requirement are needed for MTD and MDV and the MLR should be acceptable.
- The **AAL2-Class 2** used for applications with tolerant time constraint and low loss ratio (e.g. Web browsing). For this class, MDV is not specified. MTD should be limited but it doesn't have stringent requirements. MLR should be minimized because this class has stringent requirements for packet loss.
- The **AAL2-Class 3** used for applications without any time constraint but with a low loss ratio (e.g. E-mail). MTD and MDV are not specified but the MLR should be minimized.

**Mapping between classes**: Mapping between UMTS traffic classes and AAL2 service classes is needed. Conversational and Streaming classes may be supported by AAL2-Class 1. Interactive class may be supported by AAL2-Class 2. Finally, Background class might be supported by AAL2-Class 3.

AAL2-Class 1 will be supported by the stringent class of the underlying ATM layer. AAL2-Class 2 may be supported by the stringent class or the tolerant class. AAL2-Class 3 may be supported by the tolerant class or the bi-level class of the ATM layer.

2.3 Connection Admission Control for AAL2 channels

An appropriate resource allocation scheme allows the implementation of a suitable CAC function that will be able to provide the quality of service required by each AAL2 connection. At AAL2 layer, the resource allocation scheme is based on the equivalent bandwidth of each connection, which is defined by the ratio between the average ATM bit rate of the VC and the number of simultaneous AAL2 connections handled by this VC. A new AAL2 connection is accepted in the VC by the CAC function if the total amount of equivalent bandwidths of all connections in this VC (including the new one) is less than a certain threshold (a percentage of the VC capacity); otherwise the connection is rejected.

2.4 Bandwidth management

The bandwidth offered to the AAL2 layer may be shared between different classes in several ways. Two schemes are proposed:

- Flows of each AAL2 class are aggregated into the same VC. This is the scheme of mono-service VC in which at least three VCs are needed for the different AAL2 classes. Differentiation between services is done at ATM layer.

- Flows of all AAL2 classes are aggregated into the same VC. This multi-service VC allows high multiplexing gain. Differentiation between services is done at AAL2 layer. In this case, scheduling mechanism is needed in the AAL2 multiplexer in order to differentiate between AAL2 classes.

2.5 Scheduling mechanism

In the scheme of mono-service VC, all connections in the same VC have the same class of service. Thus, they have the same priority and a fair scheduling algorithm is recommended. We propose the classic FCFS and the RR algorithms.
In the scheme of multi-service VC, different service classes are handled in the same VC. An appropriate scheduling algorithm should be implemented in the AAL2 multiplexer that selects the packet that should be inserted in the ATM cell. We should take into account a fair share of the bandwidth between different services and the time constraint of each service. Four scheduling algorithms are proposed: FCFS, Priority of real-time traffic over non-real-time traffic, WRR and EDF.

2.6 Impact of the Timer-CU
Transmission delays in UTRAN should be reduced as much as possible [5]. In fact, even non-real-time traffic coming from the core network becomes more or less "real-time" inside the UTRAN because of air interface shaping [8]. The Timer-CU value is a critical parameter in the case of low bit rate traffic. Small Timer-CU value leads to poor bandwidth utilization. High Timer-CU value maximizes the bandwidth utilization but leads to higher packetization delay and consequently to degradation in the QoS for real-time traffic. An optimal Timer-CU value should be chosen carefully in order to maximize the bandwidth utilization and keep an acceptable transfer delay.

2.7 AAL2 switching
Mini-cell transport as defined by the AAL2 protocol leads to the idea of AAL2 mini-cell switching. The main difference with ATM switching is that the switched entities have variable lengths. Within the first UTRAN networks, the choice between an ATM switch and an AAL2 switch will be an important issue. In this paper, we present a comparison between an AAL2 switch and an ATM switch in order to evaluate the advantages and the drawbacks of each switching technology.

3. Simulation model
In this paper, we consider two types of traffic sources: voice and data sources. The voice model [9] used is the AMR, which consists of an ON/OFF model with exponential distribution of ON and OFF periods with a mean of 3 seconds. In the ON period, packets are sent periodically each 20 ms. The packet size depends on the AMR type. In our simulations, we use the AMR 12.2 Kbps type. Data model [9] used in this paper is the UDD traffic, which represents a Web browsing session. It consists of a number of packet-calls separated by a period called reading-time. The number of packet-calls is geometrically distributed with a mean of 5. The reading-time period is geometrically distributed with a mean of 12 seconds. Each packet-call corresponds to the download of a file that has a Pareto distributed size with a minimal file size of 1858 bytes and a maximal file size of 5 000 000 bytes. The protocol stack described in Figure 2 shapes the UDD traffic. In fact, the RLC layer [10] segments packets into RLC-PDUs that will be sent to the MAC layer [11], which puts each RLC-PDU into one TB and sends a certain number of TBs in each TTI. The FP layer at the Node B reassembles all TBs sent in the same TTI for the same service into one FP-PDU that will be sent to the AAL2 layer.

![Figure 2: Protocol stack and CPS multiplexer](image)

The traffic entering in the AAL2 layer has the pattern showed in Figure 3.

![Figure 3: UDD traffic pattern at AAL2 level](image)

In our simulation scenario, we consider two classes of service: AAL2-Class 1 (AMR traffic) and AAL2-Class 2 (UDD traffic). In the case of multi-service VC, two queues are implemented as shown in Figure 2 in order to differentiate between services. In the case of mono-service VC, only one queue for all classes is implemented.

We consider an ATM link between a Node B and a RNC. In the Node B, an AAL2 multiplexer is implemented in order to aggregate several AAL2 connections into one ATM VC (Figure 2). Each AAL2 connection corresponds to one radio channel. The ATC used for all VC is the CBR. Other ATCs are conceivable but this subject is beyond the scope of this paper.

4. Simulation results
We studied by simulation the impact of the Timer-CU on traffic performance, different scheduling algorithms, resource allocation and connection admission control function and finally a comparison between AAL2 and ATM switching technologies. In this paper, we will present only a part of the simulation results because of the limited number of pages.

4.1 Impact of the Timer-CU
The Timer-CU is an important parameter of the AAL2 protocol especially in the case of low loaded VC. Its value may affect the performance of the traffic transported on an AAL2 connection. At low loaded VC, if the Timer-CU value is very large, it may lead to large packetization delay and consequently to quality of service degradation. Very low Timer-CU value may lead to poor bandwidth utilization. Its value should be chosen carefully to obtain a trade-off between delay and bandwidth utilization efficiency. In [12], we studied in details the impact of this parameter. We showed that the optimal Timer-CU value chosen in the case of mono-service VC is suitable for
multi-service VC. The transfer delay depends on the value of the Timer-CU regardless the PCR of the VC. On the other hand, the filling ratio depends on the Timer-CU value that is a function of the PCR. However, the delay is the more critical parameter in the UTRAN. Thus, a Timer-CU value chosen to satisfy the delay requirements is needed. In the case of low loaded VC between a Node B and a RNC, the Timer-CU may not have an important impact because we can choose a low value in order to guarantee the packetization delay and we do not care about the bandwidth efficiency because the VC is low loaded. But this is not true if there is a concentration point between the Node B and the RNC especially if this point is an ATM switch. In fact, if the VC’s entering to the ATM switch are low loaded and if we choose a low Timer-CU value in order to guarantee the packetization delay, the ATM cells will be partially filled. This is not a problem for a VC entering to the concentration point because it is low loaded but the output VC may be high loaded and since in the ATM switch there is no multiplexing at the AAL2 level, cells in the output VC will be partially filled and consequently lead to a bandwidth loss. In this case, we should carefully choose the value of this important timer. A Timer-CU value between 1 ms and 2 ms may be an optimal value.

4.2 Scheduling mechanism
In the case of mono-service VC, all AAL2 connections have the same priority. Thus, the FCFS policy is suitable. Another possible algorithm is the RR. It is a cyclic mechanism that serves periodically one packet from each connection. In [13] we showed that the FCFS scheduler is a good compromise between performance and complexity. In the case of multi-service VC, where voice channels and data channels are aggregated, the implementation of a scheduling algorithm in the CPS multiplexer is mandatory. In fact, data packets are more tolerant for delay than voice packets. If a voice packet and a data packet are presented at the CPS multiplexer, we can serve the voice packet first, then the data packet. The FCFS policy is not an appropriate mechanism to arbitrate between voice and data packets because data flow may disturb the voice flow and consequently leads to QoS degradation for real-time traffic (e.g. voice). A scheduling mechanism based on priority of real-time flow over non-real-time flow gives better performance for voice traffic but increases data delay. A trade-off between these two solutions is needed. Two mechanisms are proposed: the EDF based on transmission deadlines for all packets and the WRR based on weights according to the importance of each flow. These two alternatives may be a compromise if deadlines and weights are appropriately chosen. In [13] we studied different scheduling mechanisms in several situations and for different weights for WRR and deadlines for EDF in order to obtain general conclusions about the performance of each scheduling algorithm. In special cases, WRR gives better performance than EDF. This result is very clear when data flow has very bursty pattern. For example, when UDD 384 Kbps traffic is used, FP-PDU has a greater size and it will be segmented into a large number of mini-cells that have the same deadlines. When the CPS multiplexer gives the hand to the data queue, it will serve all data mini-cells before giving the hand to the voice queue. At this time, if a voice mini-cell arrives at the CPS multiplexer, it should wait for the end of service of all data mini-cells for which deadline was expired, and the waiting time may be very large depending on data source type and data packet length. WRR algorithm is a good solution in this case because it fairly shares the bandwidth between all flows. Another advantage of WRR mechanism is its implementation simplicity in comparison to EDF.

4.3 Connection Admission Control
Connection admission control is a key issue in order to provide a guaranteed QoS for network users. Thus, the simultaneous number of AAL2 connections in an ATM VC must be controlled to guarantee the required Quality of Service for the traffic supported by the AAL2 connections. Two main parameters are used to accept or reject a new AAL2 connection request as mentioned in paragraph 2.2: the equivalent bandwidth and the threshold (percentage of the PCR value). We should determine these two parameters for different PCR values and different traffic types in order to provide a suitable CAC function. The maximal utilization ratio is computed with the constraint of respecting the required delay for each type of traffic. For voice packets, the delay (MTD) must be less than 5 ms and for data packets it must be less than 50 ms.

Mono-service VC for voice traffic: Figure 4 represents the equivalent bandwidth for AMR12.2 connections and for different PCR values depending on the number of simultaneous connections. In the case of small number of simultaneous connections (less than 70), the multiplexing gain is not optimal and the filling ratio of ATM cells is less than 90%. Consequently, certain amount of the total ATM throughput is lost as padding bytes. The computed equivalent bandwidth of an AAL2 connection is larger than the effectively used bandwidth by this connection because of the lost bandwidth. In the other hand, when the number of simultaneous connections increases, the multiplexing gain becomes important (the filling ratio is about 98%) and consequently the equivalent bandwidth decreases. When the number of simultaneous connections becomes large, the equivalent bandwidth reaches a constant value of 9.7 Kbps.

![Figure 4: equivalent bandwidth for AMR 12.2](image-url)
The maximal utilization ratio of the VC is represented in Figure 5 depending on the PCR value. This parameter gives the threshold used to accept or reject a new AAL2 connection request. For small PCR values, the threshold is less than 75%. For a PCR value more than 1 Mbps, the utilization ratio may reach a constant value of 80%. For the CAC function, if the sum of equivalent bandwidths of all AAL2 connections (including the new one) is less than 80%, the new connection is accepted; otherwise it is rejected.

![Figure 5: maximal utilization ratio](image)

Mono-service VC for data traffic: In this paper, we present simulation results of the UDD64 Kbps type only. Figures 6 and 7 represent respectively the equivalent bandwidth and the maximal utilization ratio.

![Figure 6: equivalent bandwidth for UDD 64](image)

![Figure 7: maximal utilization ratio](image)

The equivalent bandwidth is about 9.3 Kbps for UDD 64 Kbps connections. The maximal utilization ratio is very small for small PCR values (about 25% for a PCR of 500 Kbps) but it increases with the PCR value and may reach a constant value of 63% for PCR values more than 2.5 Mbps. This difference is due to the very bursty pattern of UDD traffic.

Multi-service VC: Equivalent bandwidths computed in the previous sections are used to perform an appropriate CAC function in the case of multi-service VC supporting AMR12.2 flow and UDD64 flow. We determine the maximal utilization ratio in two cases:

- AMR12.2 is the majority traffic (80% AMR and 20% UDD): in this case, the maximal utilization ratio is about 72%.
- UDD64 is the majority traffic (80% UDD and 20% AMR): in this case, the maximal utilization ratio is about 61%.

It is clear that when UDD is the majority traffic, the maximal utilization ratio is smaller than the case where AMR is the majority traffic.

4.4 Comparison between AAL2 and ATM switching technologies

Performance of AAL2 switching technology is treated in this section in the case of concentration point between several Node Bs and a RNC. The concentration point is considered to be an AAL2 switch or an ATM switch to make comparison between these two switching alternatives. All CBR VPs coming from the Node B to the switch are aggregated in the same output CBR VP. PCR value of each VP entering to the switch is 2 Mbps. PCR value of the VP between the switch and the RNC is 2 Mbps. The load of a VP between the Node B and the switch is 13% (low loaded VP). We considered two types of traffic: AMR 12.2 Kbps and UDD 64 Kbps with a combination of 50% voice and 50% data. The scheduling mechanism used at the AAL2 multiplexer is the absolute priority of voice packets over data packets. We considered that links between the Node Bs and the switch are low loaded in order to study the impact of the AAL2 switching on the multiplexing gain. Figure 8 represents the 99.9-percentile delay for voice and data packets in the case of an AAL2 switch and an ATM switch. This is the delay between the SAP of the AAL2 layer in the Node B and the other SAP of the AAL2 layer in the RNC including switching delay.

![Figure 8: AAL2 and ATM switching](image)

We observe that in the case of an AAL2 switch, we can aggregate larger number of Node Bs than the case of an
ATM switch with guaranteed quality of service. In fact, when multiplexing at AAL2 layer, we can differentiate between different types of traffic and implement a scheduling mechanism in order to give priority for stringent delay packets (e.g. voice packets). Furthermore, in the case of an AAL2 switch, we may benefit from the CPS multiplexer to eliminate padding in the cells coming from low loaded VC and reduce the bit rate of the outgoing flow. In the case of an ATM switch, the bit rate of the outgoing flow is the sum of all incoming bit rates. If the entering VPs are high loaded, AAL2 switching will not have an important advantage and an ATM switch is recommended.

5. Conclusion and perspective

AAL2 transport technology was studied in this paper in order to evaluate its capabilities in the case of the particular constraints of the UTRAN. The first deployments of the UTRAN should use this technology because it is the only solution proposed by the release 99 of the 3GPP. In this paper, we proposed a QoS architecture for the AAL2 layer in order to provide adequate services for upper layers. We proposed a CAC function and we validated it by simulation. We studied also the scheduling mechanism at the AAL2 multiplexer in order to choose the optimal scheduler. The optimal Timer-CU value is treated in this paper and a comparison between AAL2 and ATM switching technologies is done. The study results may be used for dimensioning of AAL2 network within the UTRAN.

Within the framework of the release 5 of 3GPP, the IP solution has been selected as another transport technology within the UTRAN. Many manufacturers and operators defend this solution in the perspective of deployment of global IP networks. This technology becomes very attractive if the future data traffic will represent a significant part of the whole traffic transported by the network. The IP transport technology in UTRAN will be the subject of our future study.

References


Glossary

3GPP: Third Generation Partnership Project
AAL-2: ATM Adaptation Layer - Type 2
AMR: Adaptive Multi-Rate
ATM: Asynchronous Transfer Mode
CAC: Connection Admission Control
CBR: Constant Bit Rate
CID: Channel Identifier
CSS: Common Part Sublayer
EDF: Earliest Deadline First
FCFS: First Come First Served
FP: Framing Protocol
HEC: Header Error Control
IP: Internet Protocol
ITU-T: International Telecommunication Union - Telecommunication standardization sector
LI: Length Indicator
MAC: Medium Access Control
PCR: Peak Cell Rate
PDU: Packet Data Unit
PVC: Permanent Virtual Circuit
RLC: Radio Link Control
RNC: Radio Network Controller
RR: Round Robin
SAP: Service Access Point
SSCS: Service Specific Convergence Sublayer
SS-ADT: Service Specific - Assured Data Transfer
SS-SAR: Service Specific - Segmentation and Reassembly
SS-TED: Service Specific - Transmission Error Detection
TB: Transport Block
TTI: Transmission Time Interval
UDD: Unconstrained Delay Data
UMTS: Universal Mobile Telecommunication System
UTRAN: Universal Terrestrial Radio Access Network
UU: User-to-User Indication
VC: Virtual Circuit
VCI: Virtual Circuit Identifier
VP: Virtual Path
VPI: Virtual Path Identifier
WRR: Weighted Round Robin