Chapter 3

Performance Analyses of Access Points In Cellular Wireless ATM Networks

A Marshall, F Garcia-Palacios, S Sezer, D Chieng

*School of Electrical and Electronic Engineering, The Queen's University of Belfast, Belfast
BT9 5 AH, Northern Ireland, U.K*

Key words: Quality of Service, Mobility, Wireless ATM, Access Point.

Abstract: The performance of Wireless Asynchronous Transfer Mode (WATM) networks is influenced by both cell level and call level issues. These can introduce delay and loss into the services that use the WATM Access Points (APs). At the cell or burst levels, buffering and schemes in the APs, and the instantaneous load applied to the radiocells can have an impact on the Quality of Service (QoS) offered. At the call level, the number of calls blocked at setup or handoff contributes to the Grade of Service (GoS) offered. The GoS and QoS must be considered together when assessing the performance of multimedia services in a WATM network. A novel modelling approach is presented here, whereby an integrated call level-cell level model allows realistic traffic patterns to be generated and applied at both levels. This approach permits analysis of a WATM AP architecture in terms of delay (at the cell level), and blocked calls (at the call level). Results presented show the performance of different buffering techniques in the WATM APs. Additionally, call level performance parameters allow the GoS offered to clients in the radio picocells to be adjusted across a range of traffic classes in order to optimise the QoS for time-sensitive applications.

1. INTRODUCTION

Wireless ATM (WATM) is an extension of fixed ATM networks, and provides mobile users with wireless access to broadband multimedia services, similar to those expected to be provided by fixed access ATM networks. Hence, WATM has to provide high bandwidth and support many services with different QoS objectives, and with traffic profiles similar to those encountered in fixed ATM networks. The importance of QoS has been recognised in recent years as critical for achieving predictable B-IDSN network performance. The provision and support for QoS was defined within the early standardisation of ATM [6] and its detail specified by the ATM-Forum [1], [2].

The poor transmission link quality in WATM, requires a much greater effort in order to achieve the target QoS for all connections. Besides the Physical radio layer, the ATM and the MAC layers may be viewed as two layers that also influence QoS. In the MAC layer, QoS guarantees may be violated by a poorly designed access scheme [7]. In the ATM layer one of the most important issues that impact QoS is buffer performance in the AP [13].

At the system level, call parameters such as arrival rate and session length can have a direct influence on the dynamic load at the AP (i.e. they dictate the dynamic load presented to the buffers). Traditional call level performance parameters such as the call blocking probability also give an idea of how well the system performs (i.e. the call level performance). This forms part of what may be termed the Grade of Service (GoS) provided by the WATM network. Current research considers both fields independently [9,14], while this paper considers that they are very much interrelated.

This paper describes the development of a novel approach for modelling WATM, which employs integrated call and cell level models of a tessellated picocellular WATM access network. The network supports four classes of multimedia traffic, Constant Bit Rate (CBR), Real-time Variable Bit Rate (rt-VBR), Non Real-time Variable Bit Rate (nrt-VBR) and Unspecified Bit Rate (UBR). These are termed Class 1, Class 2, Class 3 and Class 4 services respectively throughout this paper. The approach allows analysis of the dynamically changing, and multi-class traffic loads presented to the AP. The call level performance of the network is characterised by several call parameters, while the cell level analysis allows the evaluation of a number of buffering strategies considered for use in the APs.

2. PERFORMANCE OF A WATM SYSTEM

For a multi-class traffic network, two parameters must be considered: QoS, which concerns cell loss and delay, and GoS, defined here as the ratio between call time granted and call time requested for each traffic class. GoS and QoS should be considered together. If the system achieves very low blocking probabilities (i.e. high GoS) but the buffers are overloaded, then connections may suffer from considerable delay or cell loss (i.e. QoS not achieved). Hence there is a compromise between connections accepted and connections delivered with QoS.

At connection establishment, the network management must estimate the network resources required. It is necessary to provide information about the multimedia traffic characteristics, required bandwidth and QoS requirements, in order to guarantee a desired level of service. The ATM Forum has specified the traffic contract parameters (Traffic Descriptors) [1]. Negotiation of a connection is provided through the Connection Admission Control (CAC) in conjunction with a Wireless CAC [9] (WCAC) that manages the radio resources in WATM. A connection is progressed only when sufficient resources are available at each successive network element to establish the connection through the whole network, based on its service category, traffic contract, and QoS; and in order to maintain the agreed QoS for existing connections [3, 1, 6].

At connection establishment, the network management must estimate the network resources required. It is necessary to provide information about the multimedia traffic characteristics, required bandwidth and QoS requirements, in order to guarantee a desired level of service. The ATM Forum has specified the traffic contract parameters (Traffic Descriptors) [1]. Negotiation of a connection is provided through the Connection Admission Control (CAC) in conjunction with a Wireless CAC [9] (WCAC) that manages the radio resources in WATM. A connection is progressed only when sufficient resources are available at each successive network element to establish the connection through the whole network, based on its service category, traffic contract, and QoS; and in order to maintain the agreed QoS for existing connections [3, 1, 6].

From the traffic descriptors and QoS requirements, the CAC and WCAC estimate the amount of bandwidth needed to achieve the required QoS for an individual connection. There are many bandwidth estimation algorithms [3]. Allocation of bandwidth in our model is based on the Sustainable Cell Rate (SCR), and on an estimated rate midway between the SCR and the Peak Cell Rate (PCR), termed Modified Cell Rate (MfCR) [8]. This is obtained by:

$$\text{MfCR} = \text{SCR} + A (\text{PCR} - \text{SCR}) \quad (1)$$

Cell loss in the buffer is not considered here, however this has only a minor effect on QoS since large buffers are provided (100K cells) as is the trend in commercial switches [10]. The analysis presented here is mainly concerned with delay sensitive traffic. Cell delay and hence QoS can be improved by implementing different queuing strategies or by employing MfCR in the schedulers for delay sensitive connections.

At the call level, the parameters which influence GoS are: the number of calls serviced (normal calls), the number of blocked calls (at set up and handoff) and the “out of coverage area” calls. By minimising the number of calls blocked, the GoS will be improved. Minimisation of calls blocked at handoff is considered a critical issue, since more connections can proceed without their session being interrupted. This is beneficial even if it is done at the expense of a small number of extra calls being blocked at set-up, since they can be queued and serviced later. The number of handoffs performed for each traffic class is also important, since handoff implies a service disruption and by minimising this number, QoS can also be improved.

3. MODELLING AND SIMULATION

In order to undertake the analysis, a computer-based Mobile ATM Call level-Cell level system has been developed [5]. In this model various buffering schemes can be implemented at the AP. Call level performance parameters can also be measured individually for each traffic class. To achieve a realistic scenario, every mobile connection has an individual traffic source, which is modelled according to the class of the connection. Mobile terminals are created randomly in the coverage area and given call level parameters such as session length and speed. For each connection, the corresponding traffic source is activated at the cell level.

The call level parameters can dynamically vary the load offered to the AP. These are assigned independently according to the traffic class, for example a video connection (class 2 traffic) will have very low mobility and a long session length.

3.1 TRAFFIC MODEL (Cell Level)

Four traffic classes have been considered, with the following traffic characteristics:

- Class 1.- 64 Kbps voice connections with constant inter cell delay and burst size are used for constant bit-rate (CBR) traffic.
- Class 2.- MPEG2 Video Conferencing Source according to [12] is used for real-time variable bit-rate (rt-VBR) traffic.

- Class 3.- A nrt-VBR source based on a standard Ethernet traffic source model, which statistically models traffic by generating bursts from a bimodal distribution with exponential silence times between bursts.
- Class 4.- A random traffic source with exponentially distributed inter-burst delay and geometrically distributed burst size is used as an Unspecified Bit Rate (UBR) source (e.g. File transfer and email).

The objective of these traffic models is to create the superposition of bursts from individual mobile VC's, as occurs in practice [15].

3.2 THE INTERACTIVE CALL LEVEL/ CELL LEVEL MODELLING APPROACH

At the call level, the system emulates an indoor scenario where users move randomly through a tessellated cluster of WATM picocells. The model assumes that mobile users' arrivals into the coverage area is Poisson distributed. The attributes of the user, such as speed of motion, direction, and session length, are randomly generated from appropriate uniform distributions. Each user is assumed to move along a set of straight lines in different directions (updated at certain intervals) for the duration of a session. As users arrive in a cell, they request a certain bandwidth (depending on their connection class) from the AP. If the AP has the requested bandwidth available the connection is created, otherwise the call is blocked. As the user moves, the channel quality, measured as a function of distance to the AP, is periodically checked. If the quality is not acceptable the channel is released and another AP and channel are assigned to the mobile user (i.e. hard handoff is performed).

The handoffs and blocked calls in each cell can be monitored, and this in turn can be used to dynamically adjust the cell level traffic sources so that the load presented to the buffers in the APs will also vary dynamically. The Call level model passes 4 parameters (the number of users in each class) to the cell level model. These parameters are updated at call set up, call termination and during handoffs. The corresponding traffic sources are applied to the AP buffers, allowing the performance of each type of buffering strategy to be analysed.

4. THE CALL LEVEL MODEL - RESULTS

The simulations presented here consider a tessellated 7-cluster picocellular WATM network with 10m cell radii, covering an area of approximately 5000m² with 16 cells. The capacity of each picocell was set to

10 Mbps. Within specific picocells mobile populations will vary due to call establishment and termination, and handoffs. Calls can be blocked at either setup or handoff. Calls are terminated because of two reasons i) the session length is over (normal call), or ii) the user goes out of coverage area. The call parameters are given in Table 1 and are similar to those considered in AWACS ACTS project [4].

Table 1. Call Level Set Up Parameters

Call level Parameters	Class 1 (CBR)	Class 2 (rt-VBR)	Class 3 (nrt-VBR)	Class 4 (UBR)
Interarrival T.(calls/s)	15	20	10	20
MaxSpeed limit(m/s)	1.5	0.1	0.5	1
MinSpeed limit(m/s)	0.01	0.005	0.005	0.01
MaxSession length(s)	240	3600	1800	60
MinSession length(s)	15	1800	300	10
Capacity mean (bps)	64000	1300000	500000	500000

Calls may be displayed according to their traffic class, call type and blocking type as shown in Table 2. In these results a connection is accepted based on the SCR of the traffic class. The results are obtained for the total coverage area and one hour call time simulation. Alternatively any cell in the coverage area can be monitored individually.

Table 2. Call Statistics for GoS

Call Type	Class 1 (CBR)	Class 2 (rt-VBR)	Class 3 (nrt-VBR)	Class 4 (UBR)	Total by Type
Normal	115	29	107	157	408
Out of Region	119	2	114	11	246
Blocked at Set-up	0	19	14	9	42
Blocked at Hand-off	0	3	37	4	44
Total by class	234	53	272	181	740

The blocking ratios can be assessed from Table 2. These are high for classes 2 and 3, due to the high bandwidth requested and the long session times of those classes. Class 4 has a moderate blocking ratio due to its short session length, and class 1 (voice connections) performed extremely well considering that out of coverage area calls, which are high due to the high mobility of these users, could be covered by deploying more APs. The total number of handoffs, and handoff per class can also be monitored. As expected from a picocellular environment, the handoff rate is high (an average of 1 handoff every 40secs in a cell), although classes 2 and 4 have much lower handoff rates due to their low mobility or short session time. These results picture a realistic scenario in a multimedia indoors environment and provide the metrics to assess the GoS at the call level. This is computed as the percentage of the total time requested to the time granted.

Figure 1 shows the effect of increasing cell capacity (bandwidth) upon the GoS. In particular, the results show that the major beneficiaries of such bandwidth prioritisation are the class 3 users (nrt-VBR) who are highly mobile. An alternative strategy entails reserving existing bandwidth in a picocell specifically for class 3 users attempting handoff (figure 2). This significantly decreases the number of class 3 calls blocked at handoff, and the results show it is beneficial if only a small amount (5%) of the bandwidth is reserved. Reserving more than 15% rapidly increases the total number of class 1 calls blocked during handoff. This suggests that bandwidth prioritisation, or reallocation from neighbouring cells, especially for highly mobile and burst users, should be studied further.

Figure 1. Decrease of blocked calls due to Cell Bandwidth increase (16 cells, r=15m)

Figure 2. Decrease of Class 3 blocked at handoff due to Bandwidth reservation

At the call level, the model may also be used for capacity planning. The maximum number of users of each class can be set for a given blocked call probability or GoS, making sure that the QoS objectives are met at the cell level.

5. THE CELL LEVEL MODEL – RESULTS

At the cell level, the performance of several buffering techniques that can be used in the AP are analysed in terms of delay. The analysis considers downstream buffering, as this is the most critical path. That is, the capacity of the picocell is generally less than the capacity of the fixed link [13].

5.1 BUFFERING TECHNIQUES

Three buffering techniques are considered:-

The single queue approach: the cells arriving from the switch are queued in a FIFO manner. All traffic classes are queued into a single queue. While this model represents a very simplistic AP, it nevertheless serves as a reference against which the performances of more sophisticated models can be compared.

The traffic profile at the AP can be expected to be heterogeneous, therefore the bursty and variable characteristics of best-effort type traffic and VBR traffic can cause random cell transmission delays with associated jitter (CTD, CDV) for all of the connections at the access point. This unpredictable nature of the delay when using a single queue is the cause of QoS violation for CBR and rtVBR services [11].

Per-class queuing approach : In this approach one FIFO queue is used for each class of traffic. Connections of the same class are queued in one of the four corresponding queues. The number of cells NS_{Class} , which will be serviced for a particular class is determined by:

$$NS_{Class}(t) = \frac{B_{Class}(t) \cdot T_{SCycle}}{8 \cdot 53} \quad (2)$$

NS_{Class} : maximum number of serviced cells for a class during a scheduling cycle.

B_{Class} : Allocated bandwidth of the class in a link.

T_{SCycle} : Scheduler cycle time

If the serviced queue is empty before the allocated service time slot is finished, the next queue will be serviced, in order to utilise the remaining time-slots (i.e. work-conserving scheduling). Different scheduling algorithms may be considered to improve QoS at the AP, although they are outside the scope of this paper. $B_{Class}(t)$ will change over time as the number of users of a particular traffic class in a picocell vary. In our simulation this is achieved by the call level model passing the number of users per class to the cell level scheduler.

Per Virtual Connection queuing approach (Per-VC): In this approach, one queue is allocated per established connection. Queue servicing is performed according to the allocated bandwidth of each connection. Fair servicing over all queues is important in order to provide predictable delay. The scheduler allocates a number of service timeslots ST_i for each connection during each service period.

$$ST_i = \delta_i \cdot T_{MSP} \quad (3)$$

T_{MSP} = The main service period with a fixed time interval

δ_i = A certain percentage of T_{MSP} for this connection.

If a connection is not in use, the timeslots can be passed to the next connection. Since every connection is allocated an individual buffer, the average delay for a particular connection i can be obtained as a function of the utilisation of the allocated bandwidth for that connection. Interference from other connections is thus avoided [11].

5.2 CELL LEVEL RESULTS

The results in figures 3 to 7 show the performance of different buffering techniques. The results were measured at the centre AP in a cluster. The call time simulated was 30 minutes and the instantaneous total load is also shown. Figures 3 and 4 compare single queue against per-class performance. The major improvement here is with CBR traffic. It may be observed that for per-class queuing the mean delay for class 1 never exceeds 1ms, compared to the single queue system where there are periods when the delay approaches 1 second. It may also be observed however, that the single queue performs better for rt-VBR traffic. The reason for this is that the per-class scheduler allocates bandwidth according to the SCR of the source. However the peak rate of the rt-VBR traffic sources is nearly 5 times larger than its mean rate. This means that whenever a large traffic burst occurs from this type of source, it cannot be completely serviced within the schedule time allocated to the rt-VBR queue. In a single queue system however, all bursts are serviced at the full line rate and rt-VBR connections benefit. Nrt-VBR and UBR traffic delays are also shown for completeness, although delay is not significant for these classes.

In Figure 5 a new bandwidth allocation policy was implemented in the per class scheduler. Bandwidth is no longer allocated according to the SCR, instead it has been increased from 1,378,000 bit/s to a MfCR of 1,929,000 bit/s for rt-VBR sources and decreased from 500,000 bit/s to a MfCR of 256,000 bit/s for non delay sensitive traffic (nrt-VBR and UBR). It can be observed from figure 5, that the results are excellent for real time services (always below 0.1 sec.). When the load is high, nrt-VBR services have high delay, which could be decreased by increasing the bandwidth allocated in the scheduler if a specific application requires it.

Figures 6 and 7 show that a per-VC scheduling according to MfCRs performs only slightly better than scheduling according to SCR at the expense of extra bandwidth. Hence, an optimal technique would be per-VC scheduling according to a SCR allocation scheme.

Figure 3. Mean Delay Single Queuing

Figure 4. Mean Delay Per Class Queuing

Figure 5. Mean Delay Per Class with MfCR allocation

Figure 6. Mean Delay for Real Time Services (PVC with SCR and MfCR allocation)

Figure 7. Mean Delay for Non-Real Time Services (PVC with SCR and MfCR allocation)

6. CONCLUSION

This paper has shown that QoS in WATM networks is affected by the delay introduced in wireless access points. The paper has analysed these delays whenever a dynamically varying load is applied in a mobile multimedia environment. A novel approach to modelling the dynamic load presented to the access point has been developed by employing a two-level simulation. A call-level model of the mobile users and their applications is used to predict the dynamically changing load per class of traffic that is offered to the access point. This in turn is used by a cell-level model to analyse the performance of any buffering and scheduling strategies employed in the access point. Simulations have been undertaken for a tessellated picocellular WATM network providing multi-class services and three buffering approaches in the access point (single, per-class and per-VC queuing) have been examined.

The analysis shows that the access point buffer delay, is influenced by each connection when using a single queue. This causes unpredictable delays for all traffic when such a queue is used. This causes QoS violations for CBR and rtVBR services. An improvement is obtained by applying per-class and per-VC buffering. Also, scheduling according to rates above the SCR can improve delay significantly in a per class scheme. However per-VC scheduling above the SCR yield only slight improvement. The extra complexity required to implement this, may be traded against MfCR per-class scheduling, which is easier to implement and produces similar performance.

The call level – cell level model interaction is also very important, in the sense that for a given call blocking probability (GoS) the model can assess if QoS (in terms of delay) at the cell level is achieved. The effect of increasing picocell capacity (bandwidth) on the GoS for each traffic class was analysed with the call level model. This shows rt-VBR is the main beneficiary, which suggests that bandwidth reallocation from neighbouring cells, especially for class 2 traffic, should be considered.

Further studies at the cell level will investigate intelligent schedulers that adapt better to the traffic mix in order to improve delay sensitive traffic.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support and financial assistance provided by the UK EPSRC, under the CAMA project GR/L07406.

REFERENCES

- [1] ATM-Forum Traffic Management 4.0
- [2] ATM Forum Wireless ATM Working Group, BTD-WATM-01.03
- [3] M. Fontaine, 1996 "Bandwidth allocation and CAC in ATM networks" *Electronics & Comm. Eng. J.*, Aug. 1996.
- [4] Proceedings of the Wireless Broadband Communications Workshop, Brussels, Sept. 1997, AWACS project
- [5] BONES Alta Group, Simulation Tool, Designer3.6 1997
- [6] ITU-T Recommendation I.350, I.371
- [7] J. Sanchez, R. Martinez "A survey of MAC protocols proposed for Wireless ATM", *IEEE Network*, Nov/Dec 97, pp.52-62.
- [8] M. Ajmone Marsan, A. Bianco et Al. "A CAC Algorithm for ATM Networks Supporting Different QoS Classes", COST 257 project, www-info3.informatik.uni-wuerzburg.de/cost/
- [9] ACTS project AC085 Magic WAND, <http://www.tik.ee.ethz.ch/~wand/>.
- [10] S. Sezer, R. Woods, A. Marshall, "Architectural trends in Commercial ATM switches", 4th Communication Network Symposium, 1997, Manchester, pp.49-52.
- [11] S. Sezer, F.E. Garcia-Palacios, A. Marshall, R. Woods "Buffer Architectures for Predictable Quality of Service at the ATM layer", *Globecom'98*, Sydney, Nov. 1998.
- [12] D.P. Heyman, A. Tabatabai et al, 1992 "Statistical Analysis and Simulation Study of Video Teleconference Traffic in ATM Networks", *IEEE Trans on circuits and systems for video technology*, vol.2, No.1, pp.49-59
- [13] F.E. Garcia-Palacios, A. Marshall, S. Sezer, D. Chieng, "QoS Analysis of a Wireless ATM Network Access Point", *ICC'99*, Vancouver, June 1999.
- [14] M. Naghshineh, M. Schwartz, "Distributed call admission control in mobile/wireless networks", *IEEE JSAC*, vol 14, no.4, pp.711-717, May 1996.
- [15] William Stallings, "Viewpoint: Self-similarity upset data traffic", *IEEE Spectrum*, Jan 1997.