Aggregation of VoIP Streams in a 3G Mobile Network: A Teletraffic Perspective

Olufemi Komolafe
Agilent Laboratories, South Queensferry, EH30 9TG, UK, o.komolafe@ieee.org

Robert Gardner
Agilent Laboratories, South Queensferry, EH30 9TG, UK, robert_gardner@agilent.com

Abstract
The ubiquity and convergence of mobile telephony and the Internet leads many to believe that future isochronous telephony data will be transported in the form of Voice-over-IP (VoIP) datagrams. This paper investigates the modelling and aggregation of multiple VoIP streams. The relationship between the number of VoIP sources, output link rate and certain teletraffic metrics is investigated. Thereafter, an exemplar network consisting of a small cluster of wireless cells is considered and a number of different handover strategies enacted. The frequency and duration of handovers are varied to investigate the effects on system performance when VoIP calls are ongoing. The findings in this paper could be extremely useful when dimensioning networks as they provide insight into the QoS obtainable for the individual streams. This is a critical issue for real-time applications such as VoIP.

1. Introduction
The proliferation of the Internet has been accompanied by a rapid growth in the uptake of mobile communication services. It is often argued that a consequence of the likely convergence of these two evolving areas will be that wireless networks will become the primary point of attachment to the Internet. In such a scenario, all data traffic, including real-time, streamed voice calls [1,2,3], will ultimately be transported in the form of packet-switched, Internet Protocol (IPv4 or IPv6) datagrams.

Specifications released by the 3rd Generation Project Partnership (3GPP), which directs and oversees the generation of GSM-based wireless cellular specifications, already describe an architecture for an ‘All-IP’ wireless core network [4], which would be expected to convey all data. However, the transport of largely isochronous telephony traffic is not commensurate with IP-based networks that have traditionally operated according to a best-effort service paradigm. The efficient delivery of high volumes of heterogeneous IP data traffic (including VoIP, MPEG-4 video, MP3 audio, web transactions, SMS etc.) will thus require substantial provisioning, configuration management and traffic engineering effort.

Accurate traffic forecasting, modelling and network simulation is a cost effective way of predicting network performance in advance of real-life infrastructure deployment. Of critical importance therein is the development of true and representative traffic source models. Furthermore, a major cause of packet loss in wireless cellular networks is the base station handover (or handoff) process. Hence, it is important to understand and quantify the impact of handovers on system performance [5].

This paper studies the modelling of multiple Voice-over-IP (VoIP) sources and considers the effects of such traffic on network performance, which in turn determines the end-to-end quality-of-service (QoS) of individual, uni-directional VoIP packet streams. The network consists of ‘cells’ containing sources of VoIP packets, with a single point of aggregation, corresponding to a base station in a wireless cellular network. The aggregate traffic from all the cells is in turn aggregated to give a much higher bit-rate, merged traffic stream.

The network performance is measured in terms of teletraffic parameters such as the buffer occupancy at base stations and mean path delay across the simulated radio access network. In order to investigate the effect of the handover process on network and VoIP stream performance, a number of different handover strategies are also enacted and compared.

Section 2 describes the VoIP source model and the aggregation process. The results of the aggregation of VoIP streams in a single cell are presented in Section 3. Section 4 introduces and compares three different handover processes. Section 5 concludes the paper.

2. Implementation
2.1 VoIP Source Model
Voice traffic models owe much to the seminal work conducted by Paul T Brady, investigating the statistical nature of telephone conversations [6]. A key finding was that a telephone conversation can be represented by ON/OFF patterns. ON periods correspond to a conversant talking and OFF periods are due to silences. Additionally, it was found that the ON and OFF periods may be approximated using exponential distributions, with the mean durations estimated to be 1.2s and 1.8s respectively [7] which are the values adopted in this work. Daigle and Langford built on Brady’s work, looking at the packetisation of voice traffic [8]. Voice sources were modelled using exponentially distributed ON/OFF sources. During the ON period, each voice source generates packets at a constant rate. Consequently, a VoIP source is typically modelled as an ON/OFF source with an exponentially distributed sojourn time in each state. When in the ON state, a VoIP source emits fixed-size IP packets at a constant rate. No packets are produced in the OFF period. The packet size and spacing depend on the codec used and particular protocol stack, typically RTP/UDP/IP [9].

In this work, during the ON periods, 64 byte packets are generated every 30ms. These values correspond to the ITU G.723.1 codec which generates a
24 byte frame every 30ms [1,9]. Thereafter, the 40 byte RTP/UDP/IPv4 header is added. Evidently, this encoding results in an inefficient use of network resources. The efficiency may be increased by including multiple codec frames in each packet (although a tradeoff is essential as the impact of a single lost packet is increased) and/or using header compression methods [2,3,10]. In this paper, it is assumed that header compression is not employed and that each codec frame is uniquely packetised.

2.2 Aggregation of VoIP Packet Streams
VoIP sources may be readily modelled using the above description. The outputs of these sources are aggregated and it is the characteristics of this packet aggregation that is studied in this paper. At the aggregation point, the VoIP packets are buffered, if necessary, and transmitted serially at a prescribed rate. Hence, metrics pertaining to the buffer occupancy or delay are indicative of the system performance.

Given the relatively straightforward and well-studied nature of a VoIP source and aggregator, the simulator developed for this work was tested and validated using known results such as the fluid model analysed by Anick et al [11]. It was found that the results matched those in the cited paper, engendering confidence in the VoIP system simulator developed for this work.

Two distinct scenarios are considered in this paper. Firstly, the aggregation of VoIP packet streams from multiple sources to a single buffering point is studied. This scenario corresponds to the simple case of a given number of VoIP handsets being simultaneously located within a single cell, hence all the VoIP packets are aggregated at the base station. Secondly, a number of adjacent cells exist and the impact of the handover strategy, frequency and duration on the system performance is investigated. In either case, the focus is on the uplink.

3. Single Cell
Using the VoIP source model described, the impact of the number of VoIP sources and the output line rate from the base station is investigated in this section. The line rate, \( C \), is defined relative to the effective output bit rate of a VoIP source during the ON period. Figure 1 illustrates the model used. The relationship between the number of sources and the mean buffer occupancy and mean delay was studied for different values of \( C \). Figure 2 presents the probability of different levels of buffer occupancy for the exemplar case in which \( C = 50 \). (The results in Figure 2, and all the other results presented in this paper, are the mean results of a number of simulation runs, each with different random number generation seeds.) It is evident that the buffer occupancy increases with the number of VoIP sources, as would be expected. Figures 3 and 4 plot the mean buffer occupancy and mean delay respectively, against the number of sources for different output rates, \( C \). Figures 3 and 4 are similar due to the fact that only the buffering delay is considered in this work. It can be seen that for each different output rate, the mean buffer occupancy and the mean delay rise with the number of sources. Additionally, for a given number of VoIP sources, the mean buffer occupancy and delay increase with a decreasing output rate. The fact that the trends in Figures 2, 3 and 4 are intuitive is encouraging. Results such as these are extremely useful when dimensioning networks as provide insight into the QoS obtainable for the individual streams.

4. Handover Between Adjacent Cells
4.1 Handover Strategies
A simple network of a cluster of three cells is considered. The VoIP conversants are free to roam among these cells randomly, initiating handovers when appropriate [5,12]. Three alternative handover techniques are compared in this work:

- **Hard handover (discard packets)**: the VoIP source does not transmit any packets during the handover period.
- **Hard handover (buffer packets)**: any packets generated during the handover period are buffered at the handset and transmitted to the new base station once the handover is completed.
- **Soft handover**: the handset sends packets to all three base stations during the handover period.

The relative impact of these three different handover strategies on buffer occupancy and delay at the base stations was evaluated as the handover time and user mobility were varied. Three exemplar handover times were considered in this work; 10ms, 100ms and 1s. The mobility of users is simulated by varying the probability of handover. The probability of handover is defined relative to the generation of packets during ON periods. Hence, the handover frequency may be readily computed from mean ON and OFF periods and the packet generation rate during the ON period.

In the simulation, there are initially 10 users in each cell. All the base stations are identical and are serviced at 10 times the effective bit rate during ON periods, i.e. \( C = 10 \). The output of these three base stations is further aggregated and buffered, if necessary, at a hierarchical node.

4.2 Hard Handover (Discard Packets)
Figure 5a plots the mean of the mean buffer occupancy at the three base stations for different handover times and probabilities for hard handover (discard packets). Figure 5b presents the mean delay experienced by packets for the same handover strategy. (The delay is measured from when the packet is generated at the handset to when it exits the base station.) Somewhat surprisingly, it appears that increasing the frequency and duration of handovers improves the system performance from a teletraffic perspective; the mean buffer occupancy and mean delay decrease. However, it may be readily deduced that since packets are discarded during the handover, the apparent improvement in the mean buffer occupancy and delay are actually because of the substantially large number of packets thrown away. (This hypothesis was verified by counting the number of lost packets.) Hence, there will be a significant degradation in the perceived quality of the telephone conversation. Figures 5a and
5b also show that for low probabilities of handover, the handover duration has a negligible impact on the results, suggesting that, when the handover probability is low, handovers occur so infrequently that their relative impact is insignificant.

4.3 Hard Handover (Buffer Packets)

Figures 6a and 6b respectively show the mean buffer occupancy and the mean delay for hard handover (buffer packets). In contrast to Figures 5a and 5b, significantly different trends may be observed for the mean buffer occupancy and the mean delay. The mean buffer occupancy falls with increasing handover durations and probability whereas in general, the mean delay rises. The reason for this discrepancy is that packets are buffered within handsets during handovers. Essentially, since packets are queued in the handset during handovers, longer or more frequent handovers do not increase the buffer occupancy at the base station. However, since the delay is computed from the time the packet is generated in the handset to when it exits the base station, the buffering of a packet within the handset contributes to and thus increases the delay as is evident in Figure 6b. Somewhat counter-intuitively, however, when the probability of handover is 0.1, the mean delay for a handover time of 1s is less than for that for 100ms. This anomalous result may be understood by realising that, in this specific case, handovers are being executed most of the time and, since nothing is transmitted by the handset during handovers, only a very low number of packets actually reach the base station, contributing to the measured delay. In order words, the queue in the handset will grow indefinitely; few packets will have reached the base station when the simulation ends.

4.4 Soft Handover

Finally, Figures 7a and 7b give the mean buffer occupancy and mean delay obtained for soft handover. For low handover probabilities, the handover has a negligible impact on the performance metrics considered. However, as a greater number of handovers occur, the mean buffer occupancy and mean delay both rise dramatically. Of all the handover strategies, Soft handover results in the greatest mean buffer occupancy. This is the expected result since all the base stations will be sent packets during handovers. Somewhat surprisingly, Hard handover (buffer packets) tends to produce much longer delays than soft handover even though the latter results in the generation of a greater number of packets because in the former, the handover packets are dealt with sequentially (within handsets), whereas in the latter, the duplicate handover packets are handled concurrently (at the different base stations).

5. Conclusions

The rapid growth and likely convergence of mobile communications and the Internet has given impetus to the development of services that straddle both areas. The transport of VoIP traffic is one such topic and is the focus of this paper. The aggregation of VoIP calls has been investigated and the nature of the relationship between the number of VoIP sources, output link rate and certain teletraffic metrics studied. Thereafter, a simple network consisting of a cluster of 3 cells was used to explore the impact of different handover strategies. Interesting results were obtained and explained. The results presented are a useful guide when dimensioning networks to carry QoS sensitive applications such as VoIP. Natural extensions to the work include investigating the perceived voice quality, studying the packet delay distribution, implementing a more sophisticated mobility model, considering more complex mobile networks and other real-time streaming services.

Acknowledgements

The authors acknowledge the contributions of David Harle (University of Strathclyde, UK) and Joseph Sventek and Francisco Garcia (Agilent Technologies, UK) to this research.

Bibliography

Figure 1: Model of aggregation of N VoIP sources

Figure 2: Buffer occupancy for different numbers of VoIP sources for $C = 50$

Figure 3: Mean buffer occupancy as number of VoIP sources varied

Figure 4: Mean delay as number of VoIP sources varied
Figure 5: Hard handover (discarding packets during handover)

Figure 6: Hard handover (buffering packets at handsets during handover)

Figure 7: Soft handover