Project	IEEE P802.16 Broadband Wireless Access Working Group		
Title	Services and Performance Requirements for Broadband Fixed Wireless Access		
Date Submitted	22 June, 1999		
Source	Imed Frigui Nortel Networks 14 Fultz Blvd Winnipeg MB	Voice: Fax: E-mail:	204-631-2320 204-631-2473 ifrigui@nortelnetworks.com
Re:	System Requirements: Call for contribution 802.16sc-99/12 (99/05/20).		
Abstract	This paper provides a framework for the different class services and their performance requirements for broadband wireless access. The frame suggests three services (i.e., voice, data, and multimedia). Based on the service, a framework for the performance requirements for both the network and connection level is addressed.		
Purpose	The purpose of this document is to help the system requirement group lay a foundation for IEEE 802.16 interoperability.		
Notice	This document has been prepared to assist the IEEE P802.16. It is offered as a basis for discussion and is not binding on the contributing individual(s) or organization(s). The material in this document is subject to change in form and content after further study. The contributor(s) reserve(s) the right to add, amend or withdraw material contained herein.		
Release	The contributor acknowledges and accepts that this contribution may be made publicly available by 802.16.		

Services, Performance, and Capacity for BWA

Imed Frigui
Nortel Networks

Scope

The purpose of this contribution is to provide some input to the system requirements group regarding the services, performance and capacity for Broadband Wireless Access (BWA).

Introduction

Broadband Wireless Access is a wireless system designed to deliver voice, data, and multimedia. BWA offers operators the capability to deliver data rates in the tens of Mbps without relying on existing infrastructure. Although BWA is wireless, it should be emphasized that it is different from the traditional cellular system in two significant aspects. Firstly, BWA does not support mobility and the associated complexity due to hand-over. Secondly, BWA operates at high frequency, which limits the cell size.

The purpose of BWA is to provide high speed, high throughput, and low delay over a wide area thereby operating like a metropolitan area network (MAN). BWA operates in a point to multipoint fashion with one base station (BST) and several customer premise equipment (CPE). The BST transmits traffic over a logical downstream channel and receives traffic from the CPE over a logical upstream channel.

Background and Reference Model

License holders for broadband wireless access are interested in providing a variety of services to the end user. These services include traditional services such as T1/E1, voice, Internet access and new services such as voice and video over IP. While it is possible to provide an inter-working function between the BWA system and the different existing systems, an encapsulation mechanism is better suited. There are two advantages for encapsulation. First, with inter-working we need to provide an inter-working function between BWA and each existing layer 2 protocol (ATM, Frame Relay, Ethernet, etc.). Second, the amount of work needed to provide inter-working and interoperability may result in missing the market opportunity. Figure 1 shows the proposed protocol stack. A BWA convergence (sub)-layer provides a mechanism to encapsulate all of layer 2 and 3 protocols over BWA MAC protocol. The function of this layer is to provide for a header, which among other functions, identifies which protocol is being carried and performs segmentation and reassembly when needed. The header of the convergence layer need not be defined for all the protocols for the first release of the standard. However, it does need to be flexible to enable future extensions to include other layer 2 or layer 3 protocols.

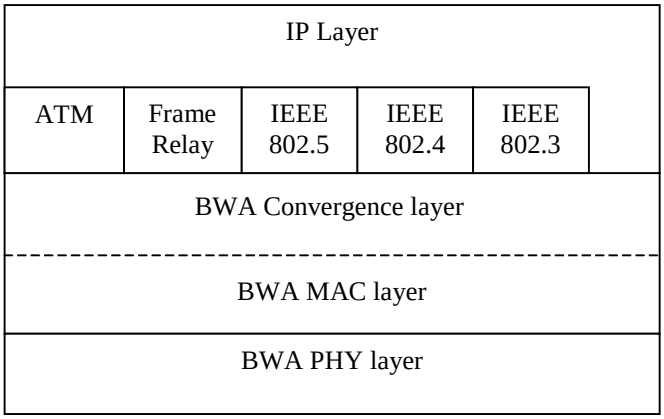

Figure 1: Protocol stack for BWA

Figure 1 represents only a subset of all the protocols that needs to be supported. For the first release we can focus on IP, IEEE 802.3, and ATM.

Classification of Services

This section addresses the possible applications and their performance requirements. We first consider the applications and classify the services, in general. Second, we present a description for each class of service.

The service bandwidth requirements can be divided in three categories: reverse, forward, and symmetric. The classification is based on the bit rate requirement for upstream and downstream directions.

1. Reverse channel bandwidth (e.g., telemetry) occurs when most of the traffic is sent in the upstream direction.
2. Forward channel bandwidth (e.g., Video-on-Demand) occurs when most of the traffic is sent in the downstream direction.
3. Symmetric channel bandwidth (e.g., POTS) occurs when the traffic requirements in the upstream and downstream directions are equal.

In general, services that require bandwidth primarily in one direction are referred to as asymmetric while those that demand comparable bandwidth in both directions are referred to as symmetric.

In addition to the bandwidth requirements in the forward and reverse direction, the services can be categorized as either requiring constant bit rate or variable bit rate. When the application requires a constant bit rate, the most important performance measures are the transfer delay and delay jitter. However, when the application requires a variable bit rate, the performance is measured mainly in terms of the transfer delay. Another dimension of performance that has to be considered is the bit error rate delivered to the application. While real time applications are mostly delay sensitive, non-real time application are bit error rate sensitive. Table 1 presents some typical services and their characteristics.

Table 1: Potential Services Addressed by Broadband Wireless Access

Bandwidth requirement	Characteristics	Services	Examples
Symmetrical	Real time	Voice	POTS

	Non-real time	Data	LAN-to-LAN
Asymmetrical	Real time	Multimedia	Inter-active video
	Non-real time	Messaging	Email

In elaborating further on the service requirements, we consider data, voice, and multimedia separately.

Data Traffic

Due to the Internet, data traffic is growing at a tremendous rate and an estimation of future requirements is only a guess-estimate. However, research done at AT&T labs for Internet users with no streaming video or data suggests that the average bit rate per user is 40 Kbps in the downstream direction, while the bit rate in the upstream direction is only 4 Kbps [9]. The limitations are mostly due to human inability to absorb information at a higher rate. In addition, research done in the early 90's for LANs suggests that the average throughput for a 10 Mbps LAN is in the neighborhood of 3 Mbps. The limitation is due to the characteristic of the CSMA/CD protocol used in most LANs. For 100 Mbps LAN, we could extrapolate that the maximum throughput would be in the order of 30 Mbps.

Based on the research mentioned above, we can estimate the bit rate requirement for data traffic for small office/home office (SOHO) and small to medium size enterprises (SME). If we assume that a SOHO has between 5 and 20 end users, then the average bit rate required per CPE is between 200 Kbps and 800 Kbps. The traffic for these users is asymmetric and follows the traditional 1:10 ratio of traffic between the upstream and the downstream. For the SME, more than 20 users, the average bit rate required can be assumed to be in the range of 3 Mbps and is more like LAN-to-LAN traffic.

In addition, to the bit rate requirements we need to be concerned about the performance requirements. Data traffic can be classified according to its delay tolerance. Delay tolerance varies from highly tolerant to delay sensitive depending on the application. For example, email is highly delay tolerant and web browsing is more delay sensitive. However, the research and standardization community is still looking at ways to quantify these delays. Two parallel groups, the IETF Internet Protocol Performance Measurement (IPPM) [1,2,3] and the ITU-T 35IP [4] are looking at ways to quantify the performance parameters to support QoS over the Internet [5].

Based on the above discussion we can conclude that data needs to be served in either predictive or best effort manner. Under the predictive service, the application is guaranteed a "committed packets rate" (CPR). Under the best effort, the system provides a reasonable quality of service for the connections. A third mechanism can be added to serve the in-between applications.

Voice

Traditional voice traffic is well understood and has very specific delay and jitter requirements as defined in ITU-G series, where G.114 is used for one way transmission time and G.176 for the integration of ATM to support voice. However, packet oriented voice service, whether it is over ATM, Frame Relay, or IP, has some ways to go before it can become widely used to service the end user. The BWA is most surely going to be packet (packet here can mean a cell as well) oriented thus, this section focuses on the requirements to deliver voice over ATM over BWA, voice over Frame Relay over BWA, and voice over IP over BWA.

Because of the number of layers involved and the associated headers/trailers (see Figure 3), the BWA protocol should be able to support silence suppression. The silence suppression feature provides for the ability to recover bandwidth during periods of silence (60% for voice for traditional voice) thereby improving the efficiency of the protocol.

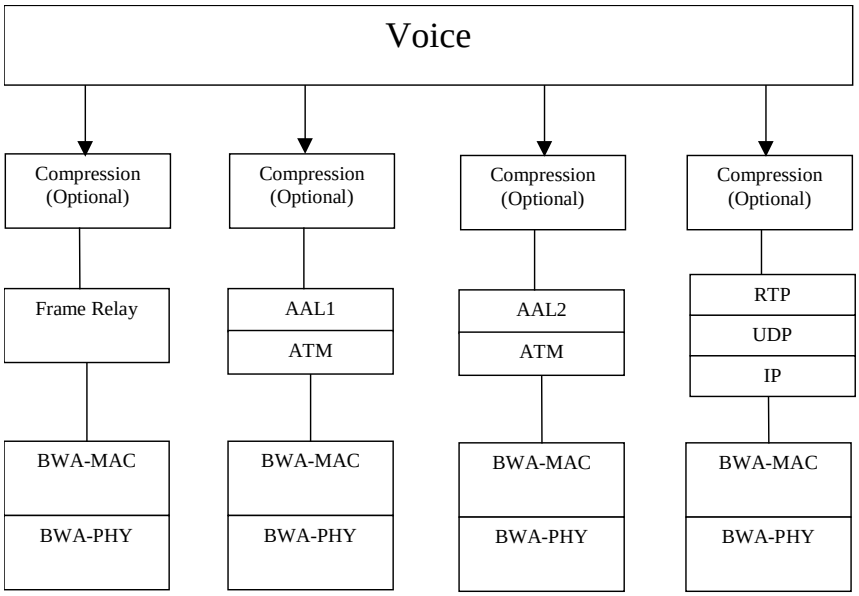

Figure 2: Protocol stack for voice over packet network (FR, ATM, IP)

Voice over ATM

Voice over ATM can be delivered either using AAL1 or AAL2. When voice traffic is carried using AAL1, it is characterized as CBR traffic. The ATM forum specifies how to carry voice, in general TDM traffic, in the circuit emulation standard af-vtoa-0078.000. In addition, voice can be carried as VBR using AAL2.

The BWA system should be able to deliver voice over ATM while maximizing resource utilization and maintaining an acceptable QoS. In order to deliver voice over ATM, a periodic slot allocation mechanism is needed. The grant period should match the delay and processing time to assemble an ATM cell. For example, the packetization delay to carry 64 Kbps PCM voice transmitted using AAL1 would be 47 bytes x 8(bits/byte) / 64 msec or 5.875 msec. In order to improve resource utilization, a mechanism for header compressing or inter-working is desirable to improve efficiency of the air interface.

Voice over IP

The Internet is based on the TCP/IP protocol suites. IP, the network layer protocol, is connectionless and does not provide for QoS. However, in order to provide voice services, the network needs to ensure a fixed delay. The end-to-end delay according to ITU G.114 recommendation should be less than 300 msec. In addition to the delay requirement, the packet loss rate should be less than 1% [9]. However, we should be careful about the packet loss rate specially when voice is compressed. In order to provide voice over IP (VoIP), and other services with QoS, some changes are needed for the Internet. Several initiatives at the IETF are taking place to provide for IP QoS. Among these initiatives are the Integrated Services (IntServ),

resource reservation protocol (RSVP), differentiated services (DiffServ), and multiprotocol label switching (MPLS). The common goal of these competing/complementing protocols is to deliver QoS. Based on the protocol chosen, the network operator objective is to deliver voice with quality that is compatible with the traditional standard rate of 64 Kbps PCM.

However, the operator objective is to maximize resource utilization while delivering a reasonable acceptable QoS. This can be achieved by using different coding schemes like G723.1 and G.729 with rates of 5.3 Kbps and 6.3 Kbps, and G.729 with rate 8 Kbps. While G.729 produces a toll quality voice, G.723.1 produces less than the desired quality due to packetization and processing delay [6].

The BWA system should be able to deliver VoIP while maximizing the bandwidth utilization. This can be achieved using a periodic grant mechanism. The periodic grant mechanism will guarantee an upper bound for the delay. The period for the grant mechanism would correspond to the packetization and processing delay. The bandwidth efficiency can be increased if a header compressing or an inter-working function is provided. The purpose of the header compression or inter-working function is to reduce the overhead due to the RTP/UDP/IP overhead.

Voice over Frame Relay

There is a small need to support voice over Frame Relay. During 1997, less than 1% of the Frame Relay traffic was voice [7]. This number is not expected to grow significantly since most service providers seem to be migrating toward the voice over IP solution. Nevertheless, voice over Frame Relay makes sense for the enterprise customer. Voice over frame relay is delivered using FRF.11 and BWA should deliver it without degradating its quality of service.

Again, similar to voice over IP over BWA or, voice over ATM over BWA, it is desirable to provide a mechanism for header compression or inter-working as such the overhead due to layer 2 protocols can be reduced, thereby, increasing the efficiency of the air interface utilization.

Multimedia Traffic

Several new services are multimedia oriented. The word multimedia means the combination of two or more media and delivering them to the user for example, graphics and text. However, the challenging problem is in delivering two or more continuous media for example, the delivery of audio and video. Examples of multimedia services include video conferencing, tele-education among many others. Depending on whether the service is provided as a single stream carrying voice and picture or two separate streams, one for voice and a second for picture, the performance requirements differ. The performance requirements are driven mainly by the synchronization needs between the different media streams. For example, during a videoconference, there is a need to synchronize between the audio and the video streams. In general, the synchronization needs are more complex when the voice and picture streams are delivered over separate flows, thereby, requiring a consistent quality of service for all connection belonging to the same class of service.

Multimedia services require very stringent end-to-end delay and delay jitter. For example, the end-to-end delay and delay jitter requirements for 64 Kbps videoconference is 300 msec and 130 msec, respectively [8]. In addition to delay and delay jitter, for multimedia services we need to define skew. Skew is a measure of the synchronization (or lack of) between the different flows.

For example, for an audio plus still picture, the skew has to be less than 1 sec. However, for an audio plus videoconference the skew has to be smaller than 200-msec [8].

Multimedia traffic is particularly sensitive to packet delay jitter. The protocol should aim for keeping the jitter as low as possible within the access network.

In addition, the BWA protocol needs to be able to support multimedia signaling protocol such as Session Initiation protocol (SIP), H323, Multimedia Gateway Control Protocol (MGCP), and Distributed Open Signaling Architecture (DOSA). Although, these are mainly control plane protocols, they may have delay requirements.

Traffic Performance Parameters

Because BWA is on the access side its performance parameters are a subset of the end-to-end performance parameters. In general, two sets of performance need to be defined. First, the performance parameters for a connection or the user-level quality and second for the network or the network quality of service and queuing. The connection performance parameters are defined in terms of delay, delay jitter, and bit error rate among many others. The network performance parameters addresses questions such as “what is the maximum capacity before degradation of service?” and “does the system provide for switched connection over the air interface?”. The latter is particularly important when the amount of resources over the air interface becomes scarce. These performance issues are addressed separately over the next two sections.

Connection Performance Parameters

The performance parameters depend mainly on the type of service (i.e., they are application depend). However, BWA intends to support other layer 2 protocols by encapsulation or interworking when such a mechanism is defined. When operating in an encapsulation mode the BWA air interface needs to provide quality of service that is at least as good as that of the layer 2 being encapsulated. For example, if BWA is carrying ATM then the air interface has to guarantee the ATM QoS requirements of the connection.

A connection has its own multi-objective quality of service requirements. In general, there are 4 QoS parameters and each connection may have all or a subset of them defined. For BWA, the QoS parameters, assuming it is packet-oriented, are defined next.

Packet loss ratio

Packet loss ratio is the ratio between the number of packet-transmitted error free and the total number of packet transmitted. This is not a BWA specific metric, however, due to the wireless link characteristic, special mechanisms such as forward error correction, must be in place to ensure that the packet error ratio is acceptable. In general, if the wireless packet size is p, and assuming that bit error occurs randomly, then the probability that there is no error in a packet is $(1\text{-BER})^p$. BER is the bit error rate. For example, if the packet size is 500 bits and the BER is 10^{-5}, then the probability that the packet has no error is $(1\text{-}10^{-5})^{500}$, which is equal to 0.9950. This implies that 0.5% of the packet may have a one bit error.

Packet delay

The packet delay is the delay between the customer premise equipment and the base station. It is composed of several elements. The propagation delay is defined for an RF link as 4 μsec/km

(G.114). For BWA systems this delay depends on the distance between the CPE and the BST. In general, it will be in the range of few μsec to 100 μsec. In addition to propagation delay, there is the transmission and potentially re-transmission delay if an automatic repeat request (ARQ) mechanism is implemented. The transmission delay is the time it takes for all the bits belonging to the same packet to arrive to the destination. For example, with a 10 Mbps channel and 500 bits packet the transmission delay is in the order of 50 μsec. The delay due to re-transmission can only be determined probabilistically since it depends on several factors including acknowledgment time, time-out and the ARQ mechanism implemented. In general, it is not recommended ARQ be used for real-time applications such as voice since the complexity introduced in the system and the delay incurred out weigh the benefits.

For the wireless link, a common technique to reduce the packet error rate is to implement interleaving with varying depth depending on the application. The delay, due to interleaving/de-interleaving, depends on the channel bit rate and the interleaving depth, and can be in the error of milliseconds.

Packet delay variation

Any packet delivery system subjects the packet belonging to the same application to jitter. Jitter is the amount of variance in the arrival of packet to their destination. BWA is envisioned to be packet oriented and therefore the jitter introduced has to be controlled. Packet delay variation in a BWA system may be significant because of its architecture. The base station controller has to inform the CPEs when they can transmit and for how long.

For BWA systems, packet delay variation is due to controller time slot assignment, buffering, and the arrival time of packet at the entry point of the system. However, the jitter contributors should not hinder the system from delivering time sensitive application. It is possible to overcome jitter by properly buffering the early packets. The jitter introduced by the bandwidth controller can be removed by transmitting the packets to the network at their expected transmission time. This is known as traffic shaping.

Burstiness

Traditional circuit switched networks define a connection only in terms of their peak bit rate, which may result in a lot of wasted resources, especially when the traffic source is bursty. Packet oriented networks couple burstiness with statistical multiplexing to increase resource utilization. Statistical multiplexing allows the network resources to be utilized by different users at different points in time. This is because each user can transmit a burst of data and then go into an idle state. The ratio between the peak and the average bit rate is defined as the burst ratio. The burst ratio varies substantially from application to application and can be as high as 200. For example, video telephony has a burst ratio between 2 and 5 and a burst length up to 10 Kbytes [10]. The BWA system should be able to handle bursty data without degradation of service.

Network Performance

In general, the network performance parameters depend on whether the network is connection-oriented or connectionless. BWA is intended to support both connection-oriented and connectionless networks. For connection-oriented networks, ITU-T Recommendation I.352 defines the average connection set-up delay for 64 Kbps for 27,500 Km (the longest connection

set-up delay) to be 4.5 sec. In addition to connection set-up time, connection-oriented networks define a blocking probability. However, unlike voice service, which has a long history of data and statistically well documented connections holding time is, new services (e.g., video conferencing) do not have a long history and estimating their holding time is quite difficult due to the lack of data. Nevertheless, several attempts have been made to characterize these services [11,12]. Table 2 gives an example of the number of busy hour call attempts for two typical services in a metropolitan area. The number in () is the number of users.

Table 2: Busy hour call attempts

Service	Residential (20,000)	Small Business (4700)	Medium Business (150)	Large Business (50)
Telephony	2.1	5	40	200
Fax	-	2.3	20	100

In addition to the blocking probability and connection set-up delay, the virtual channel address space available or the lack of bandwidth may limit accepting new session in a connection-oriented networks.

For connectionless networks, the performance is measured in terms of delay and availability. In general, the delay versus throughput curve is used to determine the threshold point beyond which a new connection should not be accepted, because it may increase the delay beyond a tolerable level. And the loss probability versus the utilization to determine the point beyond which the probability of packet loss is unacceptable. These two performance measure curves are used to scale the network.

The availability is a measure of the long-term average service time to unavailability time. The unavailability may be due to scheduled or unscheduled maintenance.

Reference

[1] G. Almes et. al. “A One-way Delay Metric for IPPM”. Internet Draft, May 1999.
[2] G. Almes et. al. “A Round-trip Delay Metric for IPPM”. Internet Draft, May 1999.
[3] G. Almes et. al. “A One-way Packet Loss Metric for IPPM”. Internet Draft, May 1999.
[4] ITU-T Recommendation I.35IP. Internet Protocol Data Communication Service – IP Packet Transfer Performance Parameters.
[5] C. Demichelis. “Packet Delay Variation: Comparison between ITU-T and IETF draft definitions”. Draft, carlo.demichelis@cselt.it.
[6] M. Hamdi et. al. “Voice Service Interworking for PSTN and IP Networks”. IEEE Comm. Magazine, Vol. 37 No. 5, May 1999.
[7] Voice over Frame Relay and ATM book.
[8] J. Russell. “Multimedia Networking Performance Requirements”. ATM Networks, I. Viniotis and R. O. Onvural, Eds., New York; Plenum, 1993, pp. 187-198.
[9] A. Dutta-Roy. “Cable it’s not just for TV”. IEEE Spectrum, Vol. 36 No. 5, May 1999.
[10] K. Dubose and H.S. Kim. “An Effective Bit Rate/Table Lookup Based Admission Control Algorithm for the ATM B-ISDN.” In: Proc of the 17th IEEE Conf. On Local Computer Networks, Sept. 1992.
[11] L. P. Bermejo et. al. “Service Characteristic and Traffic Models in Broadband ISDN”. Electrical Communication, Vol. 64-2/3, 1990, pp.132-138.

[12] G. Galassi et. al. “Resource Management and Dimensioning in ATM Networks”. IEEE Network Magazine, May 1990, pp. 8-17.