The IEEE802.16d Fixed WMAN
– a definitive description of the network to be simulated

Technical Report CS09-01-00

Paolo Pileggi, Giuseppe Iazeolla and Pieter Kritzinger *Senior Member, IEEE*

Abstract The purpose of this document is to understand the details of the IEEE 802.16d, or fixed WMAN standard. Specifically to understand the air interface and the general architecture of the protocol in order to represent these as accurately as needed in a simulation. The various assumptions that are needed to abstract the system are given in the conclusion.

Keywords: WLAN, WMAN, IEEE 802.11, IEEE 802.16, access protocols, broadband networks, radio access networks, MAC layer, QoS, broadband wireless access, quality of service support.

I. Introduction

IEEE 802.16 [2], [1] was designed from start to be connection-oriented in order to allow for better quality of service (QoS) management. Resource And Connection Management (RACM) attempts to maintain certain QoS-levels for the various Traffic Categories (TCs) specified by the IEEE 802.16d standard [1]. The RACM primarily consists of the connection admission controller (CAC) and scheduler components. The standard purposely does not specify these components in order to introduce some resourcefulness amongst vendors.

The objective of our project is to model a version of the IEEE 802.16 Wireless Metropolitan Area Network (WMAN), in fact 802.16d, and the RACM in particular. In order to do this we clearly need to understand exactly *what* the *customer* is and how that customer is served by the physical radio link. That is, how much of and which of the MAC Protocol Data Units (MPDUs) (as we shall see later) are transmitted in a Time Division Duplexing (TDD) frame, both in the up-link (UL) and the down-link (DL)? The literature also refers to this as the "service flow". Amongst others, we need to know the relationship between a Network PDU and a Transmission Control Sub-layer PDU (TCS-PDU, defined later), a frame and a slot and the size and duration of the latter two.

The purpose of this document is thus no more than to collect, in one place, all the information that is relevant for the purpose. In other words, this document is for internal use and we do not intend to publish it.

We shall focus on the more popular TDD version of the standard, rather than the Frequency Division Duplexing (FDD) version. In addition, we shall assume the Point-to-Multipoint (PMP) rather than the MESH mode of operation. Throughout, we shall use the IEEE 802.16-2004 standard document [1] as definitive.

Giuseppe Iazeolla {iazeolla@info.uniroma2.it} leads the Software Engineering Laboratory (SEL), University of Rome (Tor Vergata) while the other authors are with the Data Network Architectures Group, Computer Science Department, University of Cape Town, South Africa, Email: {psk,ppileggi@cs.uct.ac.za}

II. Network Topology

A PMP IEEE 802.16 network, such as that illustrated in Figure 1, consists of one Base Station (BS) and several Subscriber Stations (SSs). Stations (or workstations) are connected via an SS through an IEEE 802.11x Wireless Local Area Network (WLAN) interface for data transfer either to an internet server or for communicating with one another (via the SS and then the BS).

In the MESH mode SSs can also communicate directly with one another.

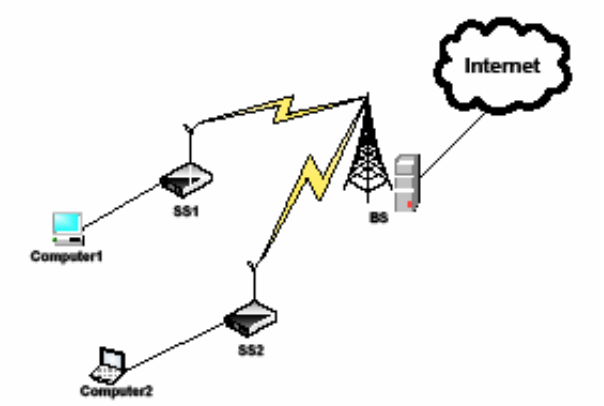

Fig. 1. Typical WMAN network. Notice the fixed line link from the BS to the Internet

Each SS communicates in the UL with the BS using a slot or slots assigned to it in an UL-MAP distributed by the BS in the previous frame. Because of the PMP architecture, the BS transmits a Time-division Multiplexing (TDM) signal in the DL direction, with individual SSs allocated time slots divided by time; the identity of an SS is contained in the data header and all SSs have to listen to all transmissions.

The services required by stations are varied in their nature and include legacy TDM voice and data, Internet Protocol (IP) connectivity, and packetized Voice over IP (VoIP) as well as synchronized video traffic. Although the standard is specific about the various traffic types catered for (see Section VI), nothing is said about the mapping of IEEE 802.11e traffic Access Classes (ACs) (there are 4 ACs) to the four IEEE 802.16d Traffic Categories (TCs).

III. Protocol Architecture

The functional entities (or the protocol architecture) needed to map an N-PDU to the transmitted radio frame, is illustrated in Figure 2. The Service-specific Convergence Sub-layer (CS)

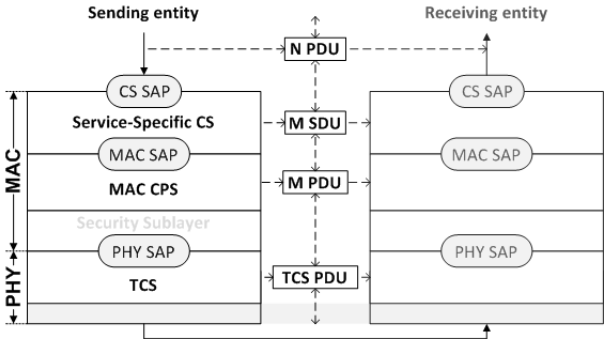

Fig. 2. WMAN protocol stack

maps the external Network Protocol Data Unit (N-PDU), or packet, arriving at the CS Service Access Point (SAP) into the MAC M-SDU. It also classifies external network data and associate them to proper MAC Service Flow Identifiers (SFIDs) and Connection IDs (CIDs). The M-SDU is then sent to the MAC CPS via the MAC SAP and mapped onto the M-PDU.

Note that one M-SDU may extend over more than one M-PDU (fragmentation) or more than one M-SDU may go into one M-PDU (packing or concatenation). The M-PDU is the data unit logically exchanged between the MAC layers of the BS and a remote SS.

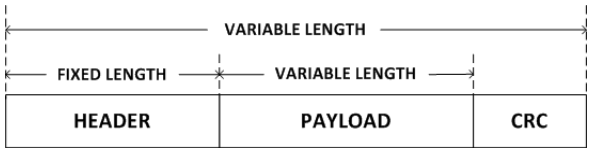

Fig. 3. M-PDU Format

An M-PDU itself, shown in Figure 3, consists of a fixed-length MAC header, a variable-length payload, and a cyclic redundancy check (CRC) which is optional in the case of the WirelessMAN-SC^{TM}interface (see Section IV) below. The length of the header is fixed at 48 bits and there are two types of MAC header: the generic MAC header for management messages and CS data, and the bandwidth request header used when requesting additional bandwidth.

Within the MAC Common Part Sub-layer, the following functions are provided:

1) Bandwidth allocation and connection establishment, in other words, connection admission control (CAC).
2) Subsequent maintenance of the service flow for a particular Connection ID (CID).
3) Building both the UL-MAP and DL-MAP using various scheduling schemes for real time, non-real time and best effort services.

RACM, the focus of our interest, is therefore located in this sub-layer.

Between the PHY and MAC is an *optional* Transmission Convergence Sub-layer (TCS). The TCS specifies how to fit the M-PDUs into PHY Forward Error Control (FEC) codewords. The TCS-PDU, illustrated in Figure 4, starts with a pointer indicating where the next M-PDU header begins within the FEC block. This enables the TC sub-layer to quickly recovery from receiving one or multiple uncorrectable codewords, at the cost of using one byte per FEC codeword as the pointer field. Such a quick recovery reduces lost M-PDUs and reduces ARQ retransmissions. The TC sub-layer can be applied to both the UL and DL. The Security Sub-layer is ignored for our purposes.

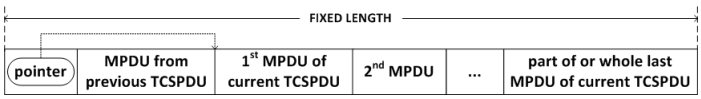

Fig. 4. TCS-PDU Format

IV. Physic Air Interface

The IEEE 802.16 Standard specifies 5 different air interfaces. These are summarized in Table I. In that table, Adaptive Antenna System (AAS) exploits more than one antenna to improve the coverage and the system capacity. Similarly, spacetime coding (STC) is a method employed to improve the reliability of data transmission in wireless communication systems using multiple transmit antennas. STCs rely on transmitting multiple, redundant copies of a data stream to the receiver in the hope that at least some of them may survive the physical path between transmission and reception in a good enough state to allow reliable decoding. Mesh refers to the network architecture and automatic repeat request (ARQ) has its usual meaning. Line of Sight (LOS) means just that.

For each air interface in Table I, the standard specifies a set of system profiles. Each profile specifies MAC and PHY parameters separately.

For the scope of our project, we will assume the WirelessMAN-SC^{TM}air interface. In that case, the basic packet MAC profile[1] specifies that fragmentation and packing/concatenation features are mandatory but may be turned off per connection. Table II gives the relevant mandatory PHY parameters for the 25MHz WirelessMAN-SC^{TM}PHY TDD profile[2].

Parameter	Value
Operation mode	TDD
Frame duration	1ms
DL modulation	QPSK and 16-QAM
UL modulation	QPSK
Roll-off factor	0.25
Symbol rate	20 MBaud
PS per frame	5000 PSs

TABLE II
Relevant mandatory profile (*profP1t*) parameters as specified by the standard

Transmission over the PHY[3] is framed in time as shown in Figure 5. The standard specifies a frame duration of 0.5, 1 or 2 milliseconds. Each TDD frame consists of n physical

[1] *profM2* in the standard
[2] *profP1t* in the standard
[3] Henceforth the PHY is assumed to refer to the WirelessMAN-SC^{TM}interface

Designation	Applicability	Options	Duplexing options	Operation
WirelessMAN-SC™	10 – 66GHz		TDD/FDD	LOS
WirelessMAN-SCa™	Below 11GHz licensed bands	AAS/ARQ/STC	TDD/FDD	NLOS
WirelessMAN-ODFM™	Below 11GHz licensed bands	AAS/ARQ/Mesh/STC	TDD/FDD	NLOS
WirelessMAN-OFDMA™	Below 11GHz licensed bands	AAS/ARQ/STC	TDD/FDD	NLOS
WirelessHUMAN™	Below 11GHz licensed exempt bands	AAS/ARQ/Mesh/STC	TDD	NLOS

TABLE I
AIR INTERFACE ALTERNATIVES

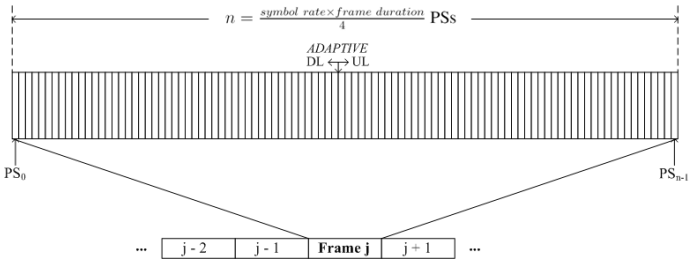

Fig. 5. TDD frame structure

Channel Bandwidth (MHz)	Baud rate (MHz)	Capacity (Mb/s) QSPK	Capacity (Mb/s) 16-QAM	Capacity (Mb/s) 64-QAM	Number of slots/frame
20	16	32	64	96	4000
25	20	40	80	120	5000
28	22.4	44.8	89.6	134.4	5600

TABLE III
BAUD RATES VS CHANNEL CAPACITY FOR A FRAME DURATION OF 1MS AND $\beta = 0.25$

slots (PS), where each PS consists of four modulation symbols (either QPSK, 16-QAM or 64-QAM). The value of n, given by Equation 1, is fixed for all frames during system operation and is a function of the symbol rate and frame duration chosen by the network operator. The symbol rate $S_R = \frac{1}{T}$, where T is the symbol-period of the communications system, is given by Equation 2 and is a function of the bandwidth (B_W) and the roll-off factor β.

$$n = \frac{\text{symbol rate} \times \text{frame duration}}{4} \tag{1}$$

$$S_R = \frac{B_W - 0.088}{1 + \beta} \tag{2}$$

The roll-off factor, β, is a measure of the excess bandwidth of the filter, i.e., the bandwidth occupied beyond the Nyquist bandwidth of $\frac{1}{2T}$. If we denote the excess bandwidth as Δf, then:

$$\beta = \frac{\Delta f}{\left(\frac{1}{2T}\right)} = \frac{\Delta f}{S_R/2} = 2T\Delta f \tag{3}$$

Table III, taken from the standard [1], shows baud rates and channel sizes using Nyquist square-root raised, cosine pulse shaping [3], with a roll-off factor of $\beta = 0.25$. The recommended frame duration is 1 ms and n is shown for various baud rate and modulation type combinations. Each modulation type represents the obvious number of bits per baud.

The modulation type and the symbol rate(the number of symbols per second) therefore play a role in determining system bandwidth. The specified frame durations 0.5, 1.0 and 2 milliseconds determine the number of physical slots in a frame. The frame duration thus does not influence the capacity of the channel, only the capacity of a frame. The standard specifies that, in the DL, TDM bursts may be transmitted with different robustness profiles, with QPSK being more robust than 16-QAM being more robust than 64-QAM. For example, data begins with QPSK modulation, followed by 16-QAM, followed by 64-QAM. In our model we shall assume one burst profile of 16-QAM for simplicity, since multiple burst profiles imply a variable bandwidth from frame to frame. If so, and arbitrarily choosing a channel bandwidth of 25MHz, and a frame duration of 1ms, we derive an overstated channel capacity of 80Mbps (see Table III).

Note that the standard specifies that the UP and the portion of the DL frame used for the MAPs must be sent using QPSK modulation for maximum likelihood of error-free reception. Moreover, there are preambles that must be sent at the beginning of each new SS transmission on the UL for synchronization. Also, the ULMAP should be larger for a larger number of SSs.

In the model we shall ignore the synchronization effects and simply assume that a frame lasts 1ms. Assuming the same duration for the UL and DL, the effective frame size is $\frac{(40+80)}{2} \times 10^6 \times 10^{-3}$ bits. That is, each frame is of maximum size 60Kb, which we will consider the "chunks" of data removed (or placed into) the MAC memory buffers by the physical transmitter/receiver at the rate of 10^3 per second.

V. LOGICAL AIR INTERFACE STRUCTURE

The TDD frame is logically divided into two sub-frames, for DL and UL transmissions, with an adaptive sub-frame boundary. DL bandwidth is defined with a granularity of one physical slot (PS) while the UL bandwidth is defined with the granularity of a mini-slot (MS), where one MS is 2^m, $0 \leq m \leq 7$ PSs long (see Table III).

In turn, the TDD sub-frames are grouped into logical portions, as shown in Figure 6. This logical organisation of the TDD frame shows the grouping of PSs. The components of interest are the

- preamble and MAP information,
- DL TDM burst,
- UL management and bandwidth request contention,
- UL Time Division Multiple Access (TDMA) burst, and
- TTG and RTG[4] periods.

The preamble, MAP information, DL TDM bursts and TTG constitute the DL sub-frame while the UL management and

[4]Both TTG and RTG are gaps between UL burst and the subsequent DL burst in a TDD transceiver. This time allows the BS to switch from receive to transmit mode and the SSs to switch from transmit to receive mode. Also see Figure 6.

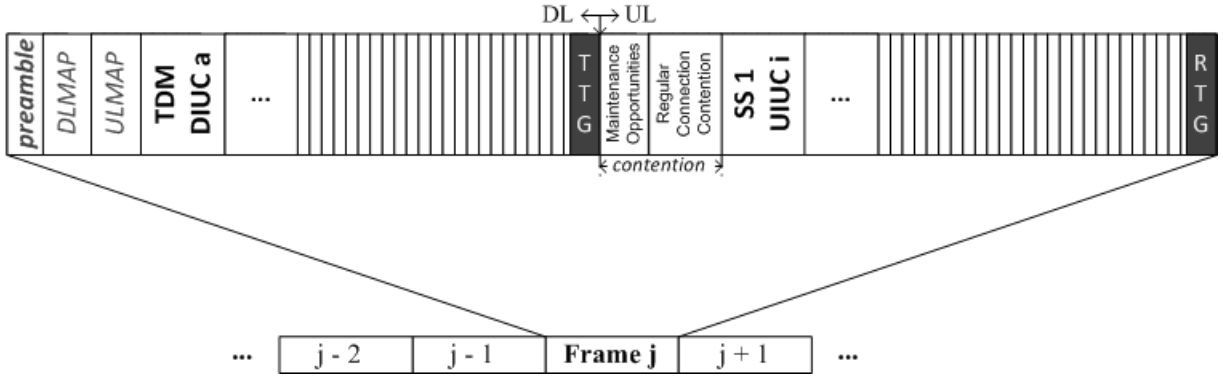

Fig. 6. The IEEE 802.16 PMP MAC frame structure showing the TDM and TDMA formats for the DL and UL respectively

contention, UL TDMA bursts and RTG constitute the UL sub-frame. The flexible frame structure of the TDD signal consists of an adaptive boundary between the DL and UL sub-frames. A short transition gap is placed between the DL and UL sub-frames and is called the transmit/receive transition gap (TTG). After the completion of the UL sub-frames, another short gap is added between this sub-frame and the next DL sub-frame. This gap is called the receiver/transmit transition gap (RTG). The time durations of the transition gaps are given in the standard and are a function of the channel bandwidth and the symbol time. As mentioned before, we shall not take this detail into consideration.

VI. Traffic Categories

While not part of the air-interface, one has to know what the five traffic categories in the specification are in order to understand certain parts of the interface, e.g., the contention interval during the uplink.

IEEE 802.16 [2] defines five types of Traffic Categories (TCs) to represent internet traffic, typically HTML, VoIP, Video Streaming, P2P and FTP in wireless networks:

1) *Unsolicited Grant Service (UGS)* is designed to support real-time, with strict delay requirements. These are applications that generate fixed-size data packets on a periodic basis, such as T1/E1 and Voice over IP.
2) *Real-time Polling Service (rtPS)* is designed to support real-time applications with less stringent delay requirements that generate variable size data packets on a periodic basis, such as moving pictures experts group (MPEG) streaming video.
3) *Non-Real-Time Polling Service (nrtPS)* is designed to support delay-tolerant, minimum rate requirement data streams. It is almost identical to rtPS except that connections have to utilize random access transmit opportunities for sending bandwidth requests. The nrtPS is suitable for Internet access with a minimum guaranteed rate and for ATM GFR connections.
4) *Best Effort (BE)* is designed to support data streams for which no minimum transmission rate is required and therefore may be handled on a space-available basis, such as HTTP. In the case of the BE service neither the throughput nor delay guarantees are provided. The SS sends requests for bandwidth in either random access slots or dedicated transmission opportunities.
5) *Extended real-time variable rate (ertPS)*, which was added in 802.16e-2005 (or Mobile-WMAN), that supports real-time applications where the applications require guaranteed data rate and delay. This service is for applications that would typically, in 802.16-2004, subscribe to the *rtPS* service even though they may behave similarly to *UGS* traffic at times, such as VoIP with silence suppression.

Typical applications of each of these TCs are shown in Table IV. Since ertPS applies to IEEE802.16e only, we shall not consider it for the system we model.

Category	Typical application
UGS	E1 transport, VoIP
ertPS	VoIP
rtPS	MPEG video
nrtPS	FTP with guaranteed minimum throughput
BE	HTTP

TABLE IV
Typical applications of each TG

VII. QoS parameters

We said before that the effect of RACM on the performance of the system is the central focus of our work. Therefor the QoS parameters specified in the standard are important. The standard addresses the following parameters[5]:

1) *Tolerated jitter* is defined as the maximum delay variation (jitter) of a connection. The value is 4ms.
2) *Maximum latency* defined the maximum latency between the reception of a packet by the BS or the SS on its network interface and the forwarding of the packet on its wireless(RF) interface. The value is 4ms.
3) *Maximum sustainable traffic rate.* This parameter defines the peak information rate of the service. The rate is expressed in bits per second and pertains to the SDUs at the input to the Convergence Sublayer. If this parameter is omitted or set to zero, then there is no explicitly mandated maximum rate. This field specifies only a bound, not a guarantee that the rate is available. For the WirelessMAN-SC™interface the maximum value of this parameter is 80Mbps.

[5]page 702 of [1]

4) *Minimum reserved traffic rate*. This parameter specifies the minimum rate reserved for this service flow. The rate is expressed in bits per second and specifies the minimum amount of data to be transported on behalf of the service flow when averaged over time. The specified rate shall only be honored when sufficient data is available for scheduling.
The BS and SS shall be able to transport traffic and satisfy bandwidth requests *for a service flow* up to its Minimum Reserved Traffic Rate. If less bandwidth than the its Minimum Reserved Traffic Rate is available requested for a service flow, the BS and SS may reallocate the excess reserved bandwidth for other purposes. The data for this parameter is measured at the input of the Convergence Sublayer. The aggregate Minimum Reserved Traffic Rate of all service flows may exceed the amount of available bandwidth. The value of this parameter is calculated from the byte following the MAC header HCS to the end of the MAC PDU payload. If this parameter is omitted, then it defaults to a value of 0 bits per second (i.e., no bandwidth is reserved for the flow).
Neither Version d [1] nor Version e [2] of the standard specifies a minimum rate for a service flow.

VIII. Service Flow Management

The Scheduler shares the network resources amongst the connections that have previously been admitted. Some researchers refer to the scheduling as "service flow management". As seen in Figure 7, the Scheduler is responsible for both the UL-MAP, which decides the order of transmission from the SS to the BS, and the DL-MAP, which transmits data downward to the SS. Note that we do not explicitly consider data traffic destined for an Internet fixed-line network, since we assume that these will be sent from the BS on a high bandwidth fixed line which does not involve the Scheduler. However, data arriving at the BS *from* the Internet environment, or new connection requests arriving from there, are treated as wireless traffic originating destined for an SS.

The IEEE 802.16 MAC accommodates two categories of SS, differentiated by their ability to accept bandwidth grants simply for a connection or for the SS as a whole. With the grant per connection (GPC) class of SS, bandwidth is granted explicitly to a connection, and the SS uses the grant only for that connection. With the grant per SS (GPSS) class, SSs are granted bandwidth aggregated into a single grant to the SS itself. In our case the network operates in GPSS mode and the BS allocates UL time per SS. Very importantly, this means that a separate (from the BS) scheduler at each SS must manage/schedule UL transmissions for the connections it manages on behalf of its stations.

The Scheduler receives connection updates and schedules these requests in the UL-MAP by executing the UL algorithm that uses the virtual queueing information and the QoS parameters. The DL-MAP is updated by consulting the MAC Memory Buffers and executing the DL algorithm to decide the allocation of the DL resource for these requests. Together, the UL-MAP and DL-MAP are built by the Scheduler and they coordinate the transmissions amongst the stations in the network. The Scheduler therefore has as output, the updated MAP.

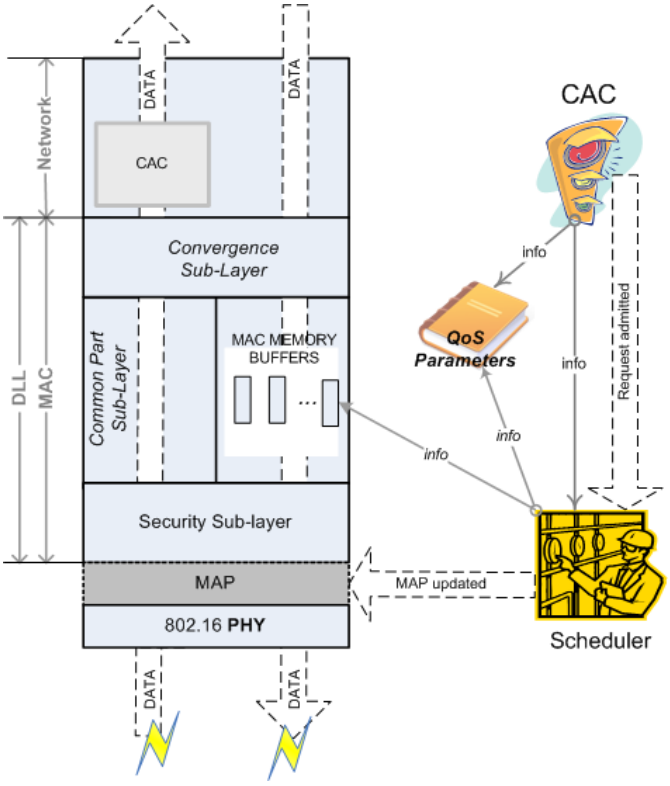

Fig. 7. Information flow between the 802.16d Scheduler and CAC and the data flow.

In the left-hand part of Figure 7, the protocol stack shows the data flow for UL and DL traffic arriving at and departing, respectively, from the 802.16 BS. On the right-hand side, the relationship between the Scheduler and CAC are shown as discussed before.

In the UL incoming data, or change requests, enter the PHY and the common MAC memory. Depending on the implementation, the M-PDUs are then passed through the Common Part Sub-Layer (CPS) and the Convergence Sub-Layer (CS), in that order. The individual SDU's with their individual identifying SFID ultimately enter the Network layer where the CAC would consider change requests, or route data packets as needed.

Referring again to Figure 7, on the DL, and again depending on the implementation, information entering the CS is queued in the MAC memory buffers. Information about these queues is used by the Scheduler to allocate or re-allocate resources. Transmission from the MAC memory buffers occurs in the PHY according to the DL-MAP and is transmitted over the wireless medium.

IX. Conclusion

The following is a summary of the assumptions and focus of the network illustrated in Figure 1 we make for the simulation model to follow.

1) The mode of network operation is point-to-multipoint (PMP).
2) Connections are granted per SS rather than per individual connection.
3) The Security Sub-layer is ignored for our purposes.
4) We will ignore the Transmission Convergence Sub-layer, including FEC and any overhead in the M-PDU.

5) Both packing and fragmentation will be allowed depending on the M-PDU size (assumed fixed) and that of the N-PDU's.
6) We will use the WirelessMAN-SC$^{\text{TM}}$air interface (see Section IV). Hence OFDM is not involved.
7) The frame duration is 1 millisecond.
8) The bandwidth is 25Mhz and the roll-off factor $\beta = 0.25$. This determines the symbol rate S_R from Eq. 2.
9) With this symbol rate and bandwidth and the assumptions mentioned before, the assumed capacity of a frame is 60Kbits and there are 5000 symbols in a frame.
10) The two quality of service parameters we shall consider are jitter, specified to have a maximum value of 4ms, and maximum latency, also 4ms.

REFERENCES

[1] 802.16d Task Group, "IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems (2004)," IEEE, IEEE802-16-2004 Version, 802.16d or Fixed WMAN, June 2004.

[2] 802.16e Task Group, "IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Mobile Broadband Wireless Access Systems," IEEE, Active International Standard, 802.11e or Mobile WMAX, approved 7 December 2005, December 2005.

[3] F. Amoroso, "On the Convolutional Square Root of a Nyquist Pulse," *Wireless Personal Communications*, vol. 1, pp. 287–290, 1995.

APPENDIX

Although not relevant for the single carrier network we have chosen to model, Orthogonal Frequency Division Multiplexing (ODFM)is central to, particularly IEEE802.16e, and unless one understands the basics of it, reading the WMAN literature can lead one astray. So the very basics of ODFM are repeated here as we copied it from Wikipedia.

Orthogonal Frequency Division Multiplexing (ODFM) is a frequency-division multiplexing (FDM) scheme utilized as a digital multi-carrier modulation method. A large number, N, of closely-spaced orthogonal sub-carriers are used to carry data. The data is divided into several parallel data streams or channels, one for each sub-carrier. Each sub-carrier is modulated with a conventional modulation scheme (such as quadrature amplitude modulation or phase shift keying) at a low symbol rate, maintaining total data rates similar to conventional single-carrier modulation schemes in the same bandwidth.

With reference to Figure 8, $s[n]$ is a serial stream of binary digits. By inverse multiplexing and assuming the number of sub-carriers is N, these are first de-multiplexed into N parallel streams, and each one mapped to a (possibly complex) symbol stream using some modulation constellation (4QAM, 8QAM, PSK, etc.). Note that the constellations may be different, so some streams may carry a higher bit-rate than others.

An inverse Fast Fourier Transform (FFT) is computed on each set of symbols, giving a set of complex time-domain samples. These samples are then quadrature-mixed to pass-band in the standard way. The real and imaginary components are first converted to the analogue domain using digital-to-analogue converters (DACs); the analogue signals are then used to modulate cosine and sine waves at the carrier frequency, f_c, respectively. These signals are then summed to give the transmission signal, $s(t)$.

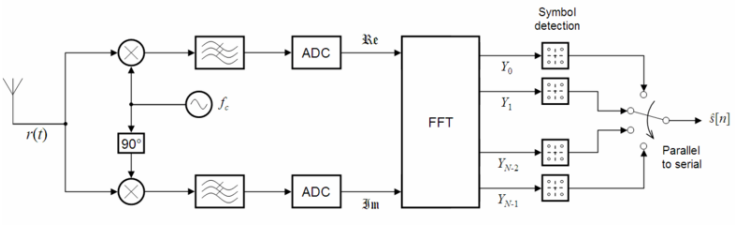

Fig. 8. Principles of OFDM: Transmitter

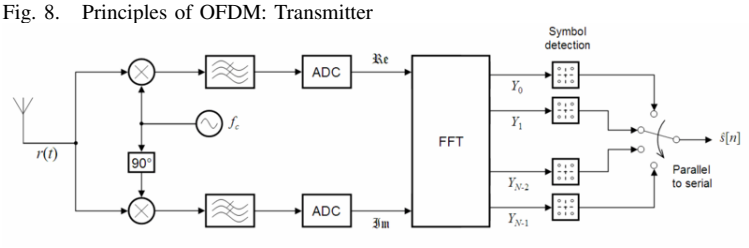

Fig. 9. Principles of OFDM: Receiver

The receiver picks up the signal $r(t)$ (which, ideally will be the same as $s(t)$) which is then quadrature-mixed down to baseband using cosine and sine waves at the carrier frequency. This also creates signals centered on $2f_c$, so low-pass filters are used to reject these. The baseband signals are then sampled and digitised using analogue-to-digital converters (ADCs), and a forward FFT is used to convert back to the frequency domain.

This returns N parallel streams, each of which is converted to a binary stream using an appropriate symbol detector. These streams are then re-combined into a serial stream, $\hat{s}[n]$, which is an estimate of the original binary stream at the transmitter.